Dissertation
submitted to the
Combined Faculty of Natural Sciences and Mathematics
of the Ruperto Carola University Heidelberg, Germany
for the degree of
Doctor of Natural Sciences

Presented by

Umut Toprak, MSc.
Born in 27.06.1986 Üsküdar, Turkey

Oral Examination 18.10.2019

INTEGRATIVE ANALYSIS OF OMICS DATASETS

Referees:
Examiner: Prof. Dr. Benedikt Brors
Co-examiner: Prof. Dr. Christoph Plass

We lost my dear friend Kenan Düz on 05.06.2018.
I loved our long conversations about every aspect of life and learned much from him. In a world full of tension, conflict and high ambition, he lived by impeccable moral values and had the most gregarious nature and calm demeanour. He perfectly balanced his intellectual and enlightened mind with an ability to find enjoyment from the simplest things in life, sharing this at every opportunity. He was an exemplary family man.
Proudly, I dedicate my thesis to his memory. I know he would have been proud of me today.
We all miss you very much, even with your memories that are always with us.

ACKNOWLEDGEMENTS

I would like to thank the following people who were an important part of my life at the DKFZ both from a scientific and personal perspective:

Roland Eils for accepting me in his group after I decided to resign from an experimental PhD and wanted to come back to computational science. He gave me the chance to prove myself as a scientist again and I hope that I managed to deserve this opportunity.

Matthias Schlesner for his scientific supervision, nice conversations, positive attitude, accessibility and friendliness. As I tell him and others time and again, he is among the most gifted people I have had the privilege of meeting. He is a brilliant scientist in interpreting observations, seeing patterns beyond others' perception. Even at times when we had our differences, I never stopped thinking that I was very lucky to have learned from him.

Frank Westermann for accepting me into his group in a senior capacity, giving me a longer term horizon and always being so supportive in every aspect of our working life. His never-changing positive attitude and calm nature is helping everybody around him. I hope to be a successful component of his team going forward towards an ambitious future for our group.

Christoph Plass for highly respecting my scientific input and ideas in all the projects we have been working on. As a young scientist, I always felt honoured to have his trust and respect.

Kai-Oliver Henrich for being a great friend, office-mate and mentor. I am looking forward to more fun years with him. I especially hope I can get him back into birding!

Edward Oakeley of Novartis and **Peter Lichter** for supporting me in my thesis committee in their busy schedules.

Benedikt Brors, **Christoph Plass**, **Şevin Turcan** and **Frederik Graw** for agreeing to read and grade my dissertation, and helping me take the last step towards my doctorate.

My collaborators **Marcel Kool**, **Stefan Pfister** and **Dominik Sturm** for a great team effort in our fun, exciting, and successful PNET project that introduced me cancer omics research. I am very happy that our work will continue in the new KiTZ institute.

My collaborators **Justyna Wierzbinska**, **Aurore Touzart**, and **Daniel Lipka** for our exciting ongoing collaboration. I am proud to have been able to find good results for our project that I could write here and am very much looking forward to having out paper together!

My collaborators **Felix Sahm**, **Andreas von Deimling** and **Philipp Sievers** for the very pleasant work on various brain tumours and how supporting they have been over the years.

My collaborators **Nicola Giesen**, **Marc-Steffen Raab** and **Jing Xu** for allowing me to join their interesting project on multiple myeloma, which I sincerely hope we can follow up and combine with our neuroblastoma research.

My collaborators **Sabine Hartlieb**, **Moritz Gartlgruber**, **Daniel Dreidax** and **Selina Jansky** for their great work in our projects on neuroblastoma. I am very happy to have now joined them and become their colleagues.

My colleagues from the DKFZ bioinformatics groups **Naveed Ishaque**, **Pavlo Lutsik**, **Ivo Buchhalter**, **Daniel Hübschmann**, **Kortine Kleinheinz**, **Nagarajan Paramasivam**, **Lena Weiser**, **Ingrid Scholz**, **Michael Heinold**, **Philip Kensche**. Thanks to you, the DKFZ has an

excellent bioinformatics community both with regards to scientific development and infrastructure. I feel privileged to have worked with you.

The scientific community of the DKFZ, especially cancer genome sequencing researchers who freely shared their data in the most helpful manner for supporting the SOPHIA project.

My dear friends **Justyna Wierzbinska** and **Bouchra Tawk**. I had great times with both of you over the years, surely these were the best times that I enjoyed in Heidelberg. I am sure you both have great careers and personal lives ahead of you and I wish you all the best of luck in every challenge ahead.

My old friends from Turkey, **Deniz Özalp**, **Erkan Buğdaycı**, **Emre Oğuzman**, **Cem Cansever**, **Bahacan Aktaş**, **Çağlar Erdoğan**, **Emir Bölen**, **Aycan Hacıoğlu**, **Enes Kayaalp**. You have always kept in touch with me and put up with me and my stress during some of the difficult times over the many years I have been abroad. I truly appreciate your friendship and hope you all the very best in your lives.

My old friends from my Zürich years **Ludovic Gillet**, **Pedro Navarro** and **Stefania Vaga** for keeping in touch after my MSc and being so caring and supportive.

Mehmet Tuncel, **Marianne Herr** and **Karolin Tuncel**. You have been like a family away from home for me. Thank you so very much for seeing me as one of your own.

Zeki Kırımcı, **Suna Kırımcı**, **Sevgi Zorbozan**. It was a great stroke of luck that I got to know you just as I was preparing to move to Mannheim. I am so happy when I visit you and witness your joyful family life. Thank you so very much for making me feel at home when I am visiting you and being so supportive during and after my move.

Somnur Merdman Kurt and **Fahir Merdman** for being so important and stalwart members of the Istanbul Erkek Lisesi community in Germany and **Aslıhan Zöllner** for being such a loyal, old and great friend.

Sema Algedik and **Nurettin Algedik** for allowing me to use their house in a most beautiful and peaceful region of Turkey while writing a major part of my dissertation.

My cousin **Diren Toprak**. You were always an interested and caring friend. You were so helpful to me in many occasions. I hope I can likewise help you in many ways in our future.

Jing Yang and **Huangshan Chen**. You are such a truly beautiful family and great friends to me. I am looking forward to watching Xingzhi grow up, confident that you will raise him as a great person.

My dear friend **Bojana Kriznik**. You are truly a great person with so many qualities. Whenever I talk to you, I remain in awe of your intellect and how you can combine this with your fun character. May you always be happy!

My dear friend **Sevinç Gücüm**. As your name goes, it was a joy for me to have met you and you deserve your meaningful name much more than I do mine. I am really very happy to see you deservedly both happy in your personal and research life. I hope you can always be in my life and listen to my newest jokes.

My dear friend **İsmail İlkan Ceylan**. You have been as close to a brother to me as anybody could have ever been. We supported each other through the best of times and some of the worst of misfortunes. You have my love as a brother and utmost respect both as a human being and a brilliant scientist.

My cousin **Yaşar Bayri** and **Koray Özduman** who I met thanks to him. You have both been truly like elder brothers to me, always lending an ear when I wanted to talk and being supportive and full of interest for my progress here. Though I still do and always will lament that I could not fully join you in your profession, I am always extra motivated when I can raise your interest with my newest research here.

My father **Emin Toprak** and sister **Güneş Toprak**. Thanks to you, I personally learned how important it is to have a good family with a peaceful environment. With me abroad, life has separated us for now, but no matter where we are on earth and in our lives, this and our memories together will live on within us.

My mother **Dilek Toprak**. I can fill entire books with what you mean to me and all you have done for me. I love you with all my heart.

TABLE OF CONTENTS

SUMMARY

Cancer is a disease of aberrant cell proliferation and tumour growth arising from the perturbation of the epigenetically defined, regulated and maintained cell identity by genetic mutations. It is a leading cause of death worldwide and most cancer types remain incurable. Omics technologies are quantitative analytical assays that allow high-quality and high-throughput measurements of different aspects of cellular regulation including genomics, transcriptomics, epigenomics, proteomics and metabolomics. These high-throughput technologies transformed the way cancer research is done, leading to tremendous advances in our understanding of cancer biology and modern targeted therapies.

Integrative analysis of multi-omics datasets in cancer research requires use of dedicated algorithms, data analysis and visualization tools. These are developed and applied in interdisciplinary teams of scientists and clinicians working on collaborative projects. Both the technical complexities of data analysis and their integration, and the efficient independent exploration of the observations by all project partners are contemporary research challenges. This dissertation presents results addressing a broad spectrum of these questions.

Chapter 1, Replacing the CNS-PNET Superentity with Four Novel Molecularly Defined Entities Driven by Structural Variants: Central nervous system primitive neuroectodermal tumours (CNS-PNETs) were a heterogeneous family of paediatric brain tumours with no histopathological markers, challenging diagnosis and poor prognosis. My work as a computational biologist contributed to the comprehensive description of this entity. In this study, we applied an integrative omics data analysis of methylomes, transcriptomes and genomes revealing that CNS-PNETs are a combination of a large group of misdiagnosed cases from other entities and four novel molecularly defined entities. I showed that these novel entities are driven by distinct and recurrent molecular drivers altered by different mechanisms of structural variants: the *FOXR2* oncogene and *MN1*, *CIC* and *BCOR* tumour suppressors. Our results contributed to the elimination of CNS-PNETs as an officially recognized cancer entity and the recognition of four novel paediatric brain tumour entities in the World Health Organization classification of brain tumours.

Chapter 2, SOPHIA, Structural Rearrangement Detection Based on Supplementary Alignments and a Population Background Model: Building on my work on structural variation in our study of CNS-PNETs, I developed the **SOPHIA** algorithm for detecting SVs in cancer genomes based on a large population background database and a corresponding bioinformatics tool written allowing fast detection of SVs from short read cancer genome sequencing datasets. SOPHIA later became the standard tool for structural variant detection in the DKFZ's cancer genome analysis workflow.

Chapter 3, EPISTEME, an Interactive and Integrative Platform for Analysing, Interpreting and Sharing Multi-Omics Data: During the development of SOPHIA and my research in projects analysing and interpreting structural variant data, I developed experiences analysing structural variant data detected by SOPHIA, integrating them with different omics layers such as gene expressions, interpreting, visualizing and sharing them with collaborators who were not computational scientists. Based on these experiences and using modern tools

of interactive data visualization, I developed an interactive platform for integrative omics data analysis and visualization named **EPISTEME**, with the aim of facilitating omics data analysis by scientists with conceptual knowledge of cancer omics but no programming skills. EPISTEME is a comprehensive tool integrating genome, transcriptome, methylome and proteome data with clinical metadata in a user-friendly web-based system with in-browser statistical analyses and publication-quality vector graphics output.

Chapter 4, SOPHIA-EPISTEME integration in DKFZ Cancer Genomics Projects Reveals Novel Disease Subtypes and Insights Across Cancer Types: With the integration of SOPHIA and EPISTEME in an integrative omics data analysis setting, my work identified novel oncogenes activated by enhancer hijacking and revealed novel molecularly defined subtypes in refractory multiple myeloma (*MYCN* enhancer hijacking via immunoglobulin rearrangements as a *MYC* replacement), adult acute myeloid leukaemia (*MNX1* activation via enhancer hijacking putatively acting as a differentiation block mechanism) and paediatric neuroblastoma (*ATOH1* activation via enhancer hijacking putatively acting as a *MYCN* replacement) in projects supported by the DKFZ Heidelberg Center for Personalized Oncology (DKFZ-HIPO) and the German Society for Paediatric Oncology and Haematology (GPOH) cancer research programmes.

ZUSAMMENFASSUNG

Krebs entsteht infolge von deregulierter Zellteilung, starker Expansion der betroffenen Zellpopulationen und dem resultierendem Wachstum von Tumoren. Dies ist auf genetische Mutationen zurückzuführen, welche die epigenetisch definierte, regulierte Zellidentität stören. Krebs gehört zu den weltweit führenden Todesursachen und die meiste Krebsarten sind bisher unheilbar. *Omics*-Technologien sind quantitative analytische Untersuchungsmethoden, die eine Analyse verschiedener Aspekte der zellulären Steuerung mit hohem Durchsatz und hoher Qualität ermöglichen. Hierzu gehören die Bereiche Genomik, Transkriptomik, Proteomik und Metabolomik. Diese Hochdurchsatztechnologien haben die Methodik der Krebsforschung grundlegend gewandelt, und zu enormen Fortschritten im Verständnis der Krebsbiologie und der modernen gezielten Krebstherapien geführt. Die Integrative Analyse von *Multi-Omics*-Datensätzen in der Krebsforschung benötigt maßgeschneiderte Algorithmen und Programme zur Datenanalyse und Datenvisualisierung. Diese Werkzeuge werden in interdisziplinären kooperierenden Forscherteams entwickelt und eingesetzt, die sowohl aus Wissenschaftlern als auch aus Klinikern bestehen. Sowohl die technische Komplexitäten der Datenanalyse von *Omics*-Datensätzen und deren Integration, als auch die Möglichkeit zur effizienten und unabhängigen Erkundung der Datensätzen von allen Projektpartnern sind aktuelle Forschungsfragen. Die vorliegende Dissertation adressiert eine umfangreiche Auswahl dieser Fragen zur integrativen Analyse von *Omics*-Datensätzen.

Kapitel 1, Die Supraentität ZNS-PNET wird durch vier neue molekular definierte Krebsarten ersetzt, die durch strukturelle Varianten getrieben werden: Primitive neuroektodermale Tumoren des Zentralnervensystems (**ZNS-PNETs**) waren eine heterogene Familie von pädiatrischen Hirntumoren mit schlechter Prognose, ohne definierte histopathologischen Marker und somit mit herausfordernder Diagnose. Meine bioinformatischen Arbeiten haben zur umfassenden Beschreibung dieser Supraentität beigetragen. In dieser Studie haben wir die integrative *Omics*-Datenanalyse vom Methylomen, Transkriptomen und Genomen durchgeführt und gezeigt, dass ZNS-PNETs neben einer großen Gruppe von fehldiagnostizierten anderen Krebsentitäten, vier neue molekular definierte Krebsarten umfassen. Ich habe entdeckt dass diese neuen Krebsarten von ausgeprägten rekurrenten strukturellen Genveränderungen getrieben werden die das *FOXR2* Onkogen und die *MN1 CIC* und *BCOR* Tumorsuppressorgene betreffen. Unsere Resultate haben dazu beigetragen, dass ZNS-PNETs von der World Health Organization nicht mehr offiziell als eine Krebsentität anerkannt werden.

Kapitel 2, SOPHIA, Erkennung struktureller Varianten durch Supplementary Alignments und ein Populationshintergrundmodell: Meine Forschung zu strukturellen Varianten innerhalb der ZNS-PNETs Studie habe ich fortgesetzt indem ich den **SOPHIA** Algorithmus zur Erkennung von strukturellen Veränderungen in Krebsgenomen basierend auf einem großen Populationshintergrundmodell und eine dazugehörige Bioinformatik-Software entwickelt habe. Diese Werkzeuge ermöglichen eine schnelle Erkennung von strukturellen Veränderungen in short read Krebsgenomdatensätzen. SOPHIA ist derzeit die Standardsoftware zur Erkennung von strukturellen Veränderungen in den Pipelines zur Analyse von Hochdurchsatzkrebsgenomen des Deutschen Krebsforschungszentrums in Heidelberg.

Kapitel 3, EPISTEME, eine interaktive und integrative *Omics*-Datenanalyseplattform

für die Analyse, Auswertung und zum Austausch von *Multi-Omics*-Datensätzen: Im Laufe der Entwicklung von SOPHIA habe ich zusätzlich zur Analyse und Auswertung von strukturellen Varianten andere *Omics*-Datensätze wie Genexpression integriert, die entsprechenden Resultate ausgewertet, visualisiert und an Kooperationspartner vermittelt, die keinen bioinformatischen Forschungshintergrund haben. Auf diese Erfahrungen beruhend und mittels moderner Software für interaktive Datenvisualisierung habe ich eine interaktive *Omics*-Datenanalyseplattform für die integrative Analyse und Visualisierung namens **EPISTEME** entwickelt. Der Zweck dieser Entwicklung war die Möglichkeit zur Analyse der *Omics*-Datensätze für Wissenschaftler mit Kenntnissen über *Krebs-Omics*-Daten aber ohne Programmierkenntnisse. EPISTEME ist eine umfangreiche *Omics*-Datenanalyseplattform und integriert Genom-, Transkriptom-, Methylom- und Proteom-Daten mit klinischen Metadaten in einem benutzerfreundlichen, Web-basierten System. EPISTEME bietet die Durchführung von statistischen Analysen und Erstellung von publikationsreifen Vektorgrafiken.

Kapitel 4, SOPHIA-EPISTEME Integration in DKFZ Krebsgenomanalyseprojekten zeigt neue Krebsuntergruppen und ermöglicht neue Einblicke für verschiedene Krebsarten: Durch die Integration von **SOPHIA** und **EPISTEME** als Teil einer integrativen *Omics*-Datenanalyse, hat meine Arbeit neue, durch enhancer hijacking aktivierte Onkogene identifiziert und ermöglicht somit die Charakterisierung neuer molekularer Untergruppen von refraktären multiplen Myelomen (*MYCN* enhancer hijacking via Immunoglobulintranslokationen), akuter myeloischer Leukämie (*MNX1* enhancer hijacking als mutmaßlicher Mechanismus von Dedifferenzierung) und pädiatrischen Neuroblastomen (*ATOH1* enhancer hijacking als mutmaßlicher Ersatz für *MYCN*) in Forschungsprojekten unterstützt vom DKFZ Heidelberg Center for Personalized Oncology (DKFZ-HIPO) und der Deutschen Gesellschaft für Pädiatrische Onkologie und Hämatologie (GPOH).

INTRODUCTION

Life is a complex and interlinked network of information flow. The simplest form of this information flow, the passing of genetic information, genetic inheritance, was known by humankind since pre-historic times: consanguinity has been avoided in complex mating networks [1] and later via social and religious norms [2], plants and animals were domesticated selecting for desired traits [3], showing some level of recognition of heredity by humankind. A systematic description of the processes governing genetic inheritance of traits was first presented by Gregor Mendel's work [4], contemporary to the first model of evolution via natural selection by Charles Darwin [5].

These phenotypic observation based explanations for genetic inheritance and evolution were expanded by a molecular understanding of the building blocks of genetic information, deoxyribonucleic acid (DNA) [6]. Identification of the structure of the DNA as a double-helix of deoxy-ribonucleotides which can replicate across cell divisions [7] followed by the hypothetical proposal and experimental verification of the "messenger" compounds ribonucleic acid (RNA) [8] [9], which led to an understanding of the way proteins were synthesized in life's cellular processes. The relationship and information flow from the DNA to RNA and to proteins was a proposed model by James Watson named the *Central Dogma* of molecular biology [10], where the following "general transfers" were named without detailed mechanistic models for all stages: DNA is replicated and confers the hereditary characteristics to cells [11], RNA is transcribed using DNA as a template from genes [12] [13], peptide chains consisting of amino acids are translated using messenger RNA (mRNA) as a template and transfer RNA (tRNA) as part of the machinery. Some "special transfers" such as RNA tanscribing DNA hypothesized by James Watson in his Central Dogma article were later experimentally confirmed [14] [15].

The Central Dogma proposes a model for the information flow in individual cells of an organism starting from their underlying genetics and how this translates to differences in their protein usage, which determine both the structure and chemical/enzymatic control of all biochemical processes. The underlying genetic differences trivially account for the differences between organisms. What the central dogma does not cover is the differences between cells of a multicellular organism: by the process of DNA replication, genetic material (DNA sequences) is, ideally faithfully ([16], [17], [18]), distributed to daughter cells which make up the organism, so the question remains as to how different cells of a multicellular organism get their different identity and function in this higher order organization despite sharing near-equal genetic material. Epigenetics is the study of cell identity and non-genetic regulation thereof [19]. Cell identity is determined by a combination of DNA methylation and histone marks, which are inherited across cell divisions.

DNA methylation is a non-genetic, reversible modification of DNA structure, but not DNA sequence. First shown in prokaryotes [20], DNA methylation is largely confined to Cytosine bases followed by Guanine (CpG) in mammals [21] except for the brief developmental state of embryonal stem cells which also shows CpH (Cytosine followed by non-guanine) methylation [22]. During embryonal development, the genome is reaches a demethylated state in a wave of

demethylation prior to differentiation [23]. DNA is then methylated in tissue specific patterns [24] [21]. An important property of DNA methylation is that it is stable across cell divisions: DNA methylation is transferred to elongating replicating DNA via DNA methyltransferases [25], which contributes to establishment of mature cell identity across cell divisions in tissue development starting from the globally demethylated embryonal state [26]. DNA Methylation is associated with closed, inactive chromatin regions [27] and repression of promoter regions of genes and consequently gene expression [28].

The second major mechanism of epigenetic regulation is histone marks: DNA is organized in higher order physical structures within a hierarchy named the nucleosome [29]. DNA, starting from a simple double-helix chain of nucleotides is packed by a family of DNA binding proteins called histones [30], and creates distinct patterns of open chromatin (also called euchromatin) [31] and constitutively condensed chromatin (heterochromatin), together with regions of non-constitutive openness defining chromatin accessibility [32]. On a larger scale, DNA is organized into local domains named TADs [33] [34] which carry co-regulated genes and are stable across tissues/cell types but are differentially activated across different cell types. Beyond the local-level organization in TADs, DNA chromatin forms loops [35] enabling long-range interactions of DNA domains. Chromatin states are determined by post translational modifications of histones [36] [37], and their binding patterns on DNA are called histone marks. The most important and well studied of these are acetlyation of lysine residues of the H3 and H4 histone subunits [38] and methylation of lysine residues of the H3 histone subunit [39]. On the most basic level, histone acetylation is associated with open chromatin [40] and histone methylation can either be associated with open chromatin or closed chromatin depending on the amino acid residues and the number of methyl groups added to the histone protein [41].

The relationship between DNA methylation and histone modifications are complex: while DNA methylation and histone acetylation suggest a reciprocal relationship, predicting a 1:1 mapping of histone mark states from DNA methylation levels is currently not possible, even though some limited relationships between DNA methylation machinery and histone modifications have been established [42], [43]. Recently, experimentally induced demethylation was shown to lead to increased chromatin accessibility in only a very limited subset of the genome [44], suggesting at least a transient mechanism to maintain histone-DNA binding patterns across cell divisions even in the absence of maintained methylation patterns.

Cell identity is established during development starting from embryonal development and continuing throughout infancy, childhood, adolescence and even later in life using the described mechanisms of epigenetic programming [45] [46] [47]. Starting from pluripotent or multipotent states [48], stem cells undergo cell divisions, losing their differentiation potency and yielding daughter cells which mature into stable states called the cell fate [49]. Establishment of cell identity happens in steps following a cell lineage [50] [51]. Previous work increased estimated number of (final) cell types from around 200 [52] to 1058 [53]. These estimates are expected to increase with single cell sequencing technologies showing an untapped diversity of cell types in mouse [54], which will be followed by the work of the *Human Cell Atlas* consortium [55]. Even though the number of cell types shows a large diversity, the number of cell lineages are comparatively small and these are governed by core transcription factors that confer lineage

commitment, which can be experimentally reprogrammed by manipulating core transcription factor activity [56].

In their mature states, cells perform defined functions following a defined and regulated transcriptomic and metabolic programme [57]. Organs of higher multicellular organisms are developed as well-defined compartments using transcriptional regulatory mechanisms [58]. In adult humans, most cell types are terminally differentiated, meaning that they are in their mature state having reached their cell fate and do not undergo further cell divisions as well as further differentiation [59]. Some non-quiescent cell types which are replenished include epithelial lining of the small intestine and colon [60] [61], epithelial lining of the breast duct [62], epihelial lining of the airways (e.g. lung) [63], epithelial lining of the the prostate [64], the haematopoietic system [65] and epidermal cells of the skin [66]. Cell differentiation through a lineage and tissue organization is tightly regulated by programmed cell death with diverse mechanisms [67], of which the most common and important is apoptosis [68]. Dysregulation of cell division and bypassing of programmed cell death leads to formation of tumours (or neoplasia), which can develop into cancer if differentiation and established cell identity are also dysregulated in a process called malignant transformation.

Non-communicable diseases including cancer that affect older individuals are assuming an increasing role as a cause of death [69]. Cancer is currently the second leading cause of death worldwide [70] and the US [71]. Cancer can arise from the wide spectrum of paediatric cell lineages of the developing body and adult cell lineages undergoing renewal [72]. By far the most common adult cancer types, and cancer cases overall, arise from epithelial linings of diverse organs, collectively named carcinomas [71]. If left untreated, cancer is lethal almost without exception, with diverse causes of death such as local tumour effects like brain herniation due to increased intracranial pressure [73], bleeding predominantly seen in haematological malignancies [74], electrolyte abnormalities such as hypercalcemia seen across cancer types [75], infection commonly seen across cancer types intrinsically or in a treatment associated manner [76]. The earliest known cancer surgery dates to ancient Egypt [77], but the greatest advances in surgery followed the development of anaesthetics [78]. Advances in surgery were followed by the introduction of radiation therapy into clinical practice [79]. Cancer treatment by medication, chemotherapy, started with successes from hormonal therapies and alkylating agents [80]. Chemotherapy drugs target biological processes that are more active in cancer, or specific weaknesses of cancer: alkylating agents (Cyclophosphamide, Temozolomide, Cisplatin, Oxaliplatin etc.) induce DNA damage that cancer cells cannot repair [81], topoisomerase inhibitors (Irinotecan, Topotecan, Etoposide, Doxorubicin, etc.) target the DNA replication process in cell division [82] [83], mitotic inhibitors (Paclitaxel, Vinblastine, etc.) which disrupt microtubule formation necessary for cell division [84], antimetabolites (5-Fluorouracil, Gemcitabine etc.) block usage of metabolites used in DNA production [85] [86], hormonal therapy (Everolimus, Letrozole, Leuprorelin, Tamoxifen, Flutamide, etc.) which target hormone dependent cancers such as some subtypes of breast cancer [87] and prostate cancer [88]. As chemotherapy targets dividing/replenishing cell populations with the intent of killing proliferating cancer cells, it also targets healthy cells with regulated proliferation such as the haematopoietic system or intestinal epithelial tissue. Therefore chemotherapy has been known

to be cause of severe side effects such as diarrhoea, nausea, potentially lethal neutropenia, potentially lethal bleeding, among others. Despite advances in the understanding of cancer biology and improvements of treatment strategies that followed, outcomes for most cancer types remain poor [89], reinforcing the great societal and research interest on the biology and clinical management of cancer.

Cancer is governed by the overarching biological concepts formally and extensively discussed in the seminal publication series Hallmarks of Cancer by Douglas Hanahan and Robert Weinberg [90] [91]. These reviews were written after a near century of molecular biological studies of cancer including high-throughput technologies, investigating in great detail the aetiology, biological mechanisms and treatment strategies of cancer. Hanahan and Weinberg proposed 6 hallmarks and later extended them by 4 new hallmarks termed enabling characteristics (*):

1. Evading growth suppressors: cancer cells bypass molecular signals that are part of normal cell lineages commanding cells to stop proliferation.

2. Sustaining proliferative signalling: cancer cells have the ability to control their own cell division (mitogenic) signalling, and ignoring the homoeostatic tissue regulation.

3. Resisting cell death: cancer cells bypass programmed cell death mechanisms of apoptosis and autophagy, and rather die using the necrosis type of cell death, promoting tumour growth.

4. Enabling replicative immortality: cancer cells can replenish or maintain their telomeres via different mechanisms which allow them to replicate indefinitely without DNA damage due to lack of telomeric protection.

5. Inducing angiogenesis: tumours promote the formation of blood vessels that ensure the availability of biomaterials to sustain their growth.

6. Activating invasion and metastasis: cancer cells invade into healthy tissue and tumours release cancer cells into the bloodstream or the lymphatic system which can seed into distant locations growing new tumours called metastases.

7. Genome instability and mutation *: cancer cells develop and accumulate somatic mutations and other genomic alterations altering gene function, activity and regulation, enabling the other hallmarks listed here.

8. Deregulating cellular energetics *: cancer cells adapt their metabolism to hypoxic conditions that arise due to deregulated rapid and dense growth, first described by Otto Warburg [92].

9. Tumour promoting inflammation *: tumours have an inflammatory microenvironment and are infiltrated by immune cells where the wound healing and dead cell and cellular debris removal functions of innate immune system are hijacked to promote and sustain their growth.

10. Avoiding immune destruction *: through the accumulation of somatic mutations through mutagens or DNA damage repair deficiencies (cancer neoantigens), cancer cells assume a genetic makeup different enough for the immune system to recognize them as foreign organisms and targets of immune response. The immune surveillance system consisting of T cells, nacrophages and natural killer cells both pre-emptively kill transformed cells and infiltrate established tumours as part of an anti-tumour immune response. Cancer cells evade this immune response in a process called immune escape.

An aspect of cancer biology, partially covered under "genome instability and mutation" but not explicitly discussed as a hallmark of cancer by Hanahan and Weinberg is the concept of somatic or clonal cancer evolution: cancer cells, once free of regulated tissue homoeostasis enter an independent evolutionary programme following the general principles of evolution by natural selection. This evolutionary process allows tumours to develop traits to optimize their growth characteristics or to evade treatment due to resistant cell subpopulations [93] [94] [95].

Cancer starts from somatically mutated precursor cells that continue to accumulate mutations over the course of the tumour evolution. Diverse gene types recurrently undergo somatic mutations across different cancer types, promoting functions that sustain their proliferative program through pressures like oxidative stress and cancer treatment. The following is a list of main classes of genes that are frequently mutated in different cancer types, along mentions of key genes and reviews.

- Pathways promoting growth such as RAS or NOTCH can be constitutively activated by mutations of key genes: *KRAS*, *HRAS*, *NRAS* [96], *BRAF* [97], *NOTCH1* [98] [99].
- Genes encoding growth factors can be aberrantly activated or amplified to promote constitutive growth: *IGF2* [100], *EGFR* [101], *ERBB2* [102], *FGFR1* and *FGFR3* [103].
- Genes encoding cyclin dependent kinases and cyclins can be amplified like *CDK4 CDK6* [104] *CCND1* [105], or cyclin dependent kinase inhibitors can be lost like *CDKN2A CDKN2B* [106] to disrupt the G1/S mitotic chekpoint leading to constitutive growth.
- DNA damage response genes blocking cell division are also targets of inactivating mutations: *TP53* [107], *CHEK2* [108].
- Suppression of growth regulation signals can be achieved by loss of function mutations or homozygous losses: *RB1* [109], *PTEN* [110], *PTCH1* [111], *NF1* [112], *NF2* [113], *SMAD4* [114].
- Evasion of apoptosis can be achieved by loss of function *APC* [115], *BAX* [116], *TP53* [107], or aberrant activation of apoptosis evasion genes *BCL2* [117], *YAP1* [118].
- Epigenetic dysregulation can be achieved by diverse mutations of histone demethylases *KDM6A* [119], histone methyltransferases *KMT2A* [120], subunits of the polycomb repressive complex 2 with histone methyltransferase activity *EZH2* [121], *SUZ12* [122], DNA methyltransferases *DNMT3A* [123], histone subunits *H3F3A* [124] or SWI/SWF complex chromatin modelling genes *SMARCB1*, *SMARCA4*, *ARID1A* [125] [126].

- Basic helix loop helix (BHLH) transcription factors can be aberrantly activated and promote growth and ”stemness”: *MYC MYCN MYCL* [127] *TAL1* [128]. The same effect can be created by inactivating mutations of *MAX* [129].
- Aberrant activation of homeobox transcription factors can dysregulate cell identity: *MNX1* [130] *NKX2-1* [131].
- Transcriptional (co-)repressors can be inactivated like *BCOR* and *CIC* or aberrantly activated like *GFI1B* [132] leading to dysregulation of gene expression.
- DNA damage repair pathways can be dysregulated by deactivation of key genes allowing further accumulation of tumour promoting mutations and chromosomal alterations: *BRCA1* and *BRCA2* [133] [134], *ATM* [135], *MLH1*, *POLE* [136].
- Immune evasion can be supported by amplification of *CD274* (*PD-L1*) [137] or loss of HLA class I antigen genes [138].
- Protein kinases are frequent targets of mutations with diverse functional consequences: *ALK* [139], *FLT3* [140], *NTRK1* [141], *KIT* [142], *PIK3CA* [143], *JAK2* [144].
- Telomere maintenance and consequent cell immortalization can be achieved by *TERT* activation [145] or *ATRX* [146] inactivation.
- Ubiquitin ligases that regulate protein degradation can either act as tumour suppressors when inactivated, like *VHL* [147] or oncogenes when aberrantly activated, like *MDM2* [148] depending on downstream targets.
- Mutations of metabolic enzymes can lead to aberrant production of growth promoting metabolites and develop resistance against the oxidative stress conditions arising due to rapid tumour growth: *IDH1* and *IDH2* [149].
- RNA processing and silencing genes have also been implicated as mutational drivers of cancer: POLR2A [150], DICER1 [151].

Identifying the importance of mutations promoting different aspects of tumour biology allowed the development of diverse modern therapies of cancer, that directly target a mutated/activated oncogene or oncogenic pathway or a pathway that emerges as a weakness due to loss of a tumour suppressor or via exploiting immune characteristics of the tumour [152]. These biomarker-based targeted therapies are promising with regards to reduction of side effects by targeting cancer cells not with respect to proliferation characteristics but molecularly defined targets.

- The *RAS* oncogene family is currently not directly targetable but MEK inhibition is a promising avenue to indirectly target RAS-mutant cancers [153]. BRAF [154] and *NOTCH* family genes [155] are targeted by clinically approved inhibitors.
- Growth factors can be targeted by inhibitor or antibody based strategies: *IGF* (experimental) [156], *EGFR* (cetuximab, antibody) [157] (lapatinib, inhibitor) [158], *ERBB2* (lapatinib) [159] (trastuzumab, antibody) [160].

- Cyclin dependent kinases are targetable by selective kinase inhibitors: *CDK4 CDK6* [161].
- DNA damage response genes blocking cell division are currently not directly targetable such as *TP53* and *CHEK2*.
- Loss of growth signal suppressing genes remains mostly not directly actionable with genes such as *RB1*, *PTEN*, *NF1*, *NF2*, *SMAD4* lost in cancers without a direct rationale for therapeutic molecular targeting. *PTCH1* or *SMO* mutations and consequent activation of the sonic hedgehog pathway can be targeted by SMO or GLI inhibitors [162].
- Evasion of apoptosis is also generally not targetable even though exceptions exist such as BCL2 inhibition in haematological cancers [163].
- Epigenetic dysregulation is a broadly targetable process with drugs both available and in advanced trials that target aberrant histone deacetylation (HDAC inhibitors) or aberrant cancer methylomes arising due to DNA methylase mutations (demethylating agents) [164], [165].
- Aberrant activation of Basic helix loop helix (BHLH) or homeobox transcription factors is currently only indirectly targetable, with direct TF inhibition generally currently not possible [166] [167].
- Loss of transcriptional (co-)repressors is currently not a targetable process in cancer.
- Deficiencies in DNA damage repair pathways is a well-targetable process with double strand break repair deficiency (e.g. caused by *BRCA* mutations) treated by PARP inhibitors [168]. Mismatch repair deficiency leading to microsatellite instability and hypermutation arising due to *MLH1* or *POLE* mutations can be rationale for checkpoint blockade immunotherapy [169] [170].
- Immune evasion is an actionable biological property of cancers and is a promising treatment avenue of great clinical and research interest, most notably via checkpoint blockade by PD1 PD-L1 and CTLA-4 inhibition via monoclonal antibodies [171] [172]. Nevertheless, loss of HLA class I antigen genes remains not actionable, and is even a resistance mechanism against checkpoint blockade immunotherapy [173].
- Kinase inhibitors [174], with approved inhibitor drugs for all of the hitherto discussed protein kinases such as the NTRK family [175] ALK [139] KIT [176], JAK family [177], PIK3CA [178] and FLT3 [179].
- Dysregulated telomere maintenance via *TERT* activation and the ALT phenotype as a consequence of *ATRX* mutations is currently not actionable. Availability of inhibitors and *in vitro* results [180] did not translate to a clinically approved strategy of TERT inhibition.
- Aberrations of ubiquitin ligases have currently no approved therapies but MDM2 inhibition is a subject of study both on the research and clinical trial level [181].

- Mutations of metabolic enzymes can be biomarkers for targeted treatment with both IDH1 [182] and IDH2 [183] as targets of selective and clinically approved inhibitors.
- Among the RNA processing and silencing genes implicated in cancer POLR2A targeting is a subject of pre-clinical research [184] [185], while DICER1 is not targetable.

Overall, diverse families of genes with broad biological functions are mutated in cancer, of which many are actionable with modern targeted treatments. A general pattern is that gene activation or activating mutations can be candidates for direct targeting, whereas loss of function mutations or loss of genes can only be targeted indirectly by pathway-level strategies or via synthetic lethality. Biomarkers can make not only therapy recommendations but rather rule out therapies: for instance, *RAS* family or *BRAF* mutant colorectal cancer are resistant to anti-EGFR therapy [186], which constitutes a negative recommendation. Even with no directly actionable consequences, particular biomarkers can affect therapy decisions like germline TP53 mutations ruling out neoadjuvant radiotherapy [187].

Similar to targets of somatic mutations, the mechanisms of somatic mutations also show diversity in cancer biology:

- Point mutations, (also named single nucleotide variants, SNVs) or small insertions or deletions (indels) can constitutively activate proto-oncogenes in gain of function events by inducing precise changes in protein structure, for instance in the activating oncogenic mutations of the *RAS* family.
- SNVs or small indels can deactivate tumour suppressors by precise amino acid substitutions (missense mutations, in-frame indels) or large-scale changes in protein structure (nonsense stopgain/stoploss SNVs, frameshift indels) [188].
- Splice site mutations can lead to loss of regulated splicing patterns and aberrations in spliced transcripts including chain terminations [189].
- Mutations in promoter regions can aberrantly activate oncogenes like *TERT* [190].

For inactivating mutations of tumour suppressors, mutations are often paired with a concomitant loss of the second allele of the gene by a prior or following copy number loss, a prior copy neutral loss of heterozygosity (LOH) event, or a direct mutation of the second allele possibly starting from the germline. Chromosome arm level losses via chromosome missegregation during cell division followed by apoptosis are common across cancer types and frequently happen on chromosome arms carrying core tumour suppressors like *TP53* or *VHL* [191] [192].

A large class of somatic alterations in cancer are collectively named structural variation. These are large-scale changes in chromosome structure such as the deletion, duplication, inversion of large chromosomal segments possibly leading to copy number changes and creating complex patterns of chromosomal alterations such as chromothripsis or chromoplexty [193].

Structural variants can have significant oncogenic effects in almost all cancer types with a variety of mechanisms. They can,

- create chimeric oncogenic fusion genes with functions combining those of the fusion partners or deleting tumour-suppressing domains of one or both partners such as *BCR-ABL1* [194] *PML-RARA* [195], *CBFB-MYH11* [196], and *TMPRSS2-ERG* [197],

- alter protein structure by targeted insertion or deletion of sequences as part of in-frame indels such as *FLT3* [140] and *BCOR* [198],
- amplify genes for oncogene overexpression and oncogenic transformation such as *ERBB2* [199], *MYCN* [200] and *EGFR* [201],
- (focally) delete tumour suppressor genes such as *TP53* [202] *VHL* [203] *RB1* [204] *CDKN2A/B* locus [205] and *PTEN* [206] which sometimes happens in a two-hit setting [207] or in a haploinsufficient setting [208],
- activate oncogenes by hijacking of active enhancers as part of normal haematological process of V(D)J or CSR recombination in B-cells [209] and T-cells [210],
- activate oncogenes by hijacking of active enhancers in a more general context on genes such as *GFI1B* [132] *PRDM6* [211] *TERT* [212] *EVI1* [213] and *IGF2* and *IRS4* [100],
- activate oncogenes by deletion of insulator regions such as *PDGFRA* [214],
- truncate tumour suppressor genes by intragenic duplications or deletions spanning multiple exons such as *MLH1* [215] and *ATRX* [146],
- truncate tumour suppressor genes by interchromosomal structural variants hitting the gene body such as *TP53* in Osteosarcoma [216] or *RB1* in Multiple Myeloma [217],
- activate oncogenes by insertion of mitochondrial sequences to gene promoters such as *FOXR2* [218].

Discovery of cancer related genes followed two separate methodological trajectories for tumour suppressor genes that are lost or deactivated and oncogenes that are amplified or activated by gain of function mutations.

Tumour suppressor genes were first hypothesized to exist due to the familial nature of the retinoblastoma disease, where Alfred Knudson observed the discrepancy in age of onset for familial and sporadic cases of retinoblastoma and postulated that two hits on the genetic material of a patient is required for onset of retinoblastoma [207]. This observation was later confirmed by observations suggesting that the dual deletion of the *RB1* gene is observed in retinoblastoma patients [219] [220].

Oncogenes were first discovered following studies of transforming animal oncoviruses that induced tumours upon infection of a host [221]. These viruses were found to harbour oncogenes that were observed to have homologs in humans as part of their normal, non-viral genetic material [222] [223]. Key genes that confer transforming properties were described and hypothesized to be altered by mutations conferring [224] aberrant activation. These hypotheses were shortly thereafter confirmed by the discovery of activating mutations of the *HRAS* [225], and *KRAS* [226] oncogenes using DNA cloning followed by sequencing [223]. Later, chromosomal translocations were shown to be a second oncogenic activation mechanism in B cell lymphoma [227]. This was followed by gene amplifications as a third oncogenic activation mechanism with the *MYCN* oncogene shown to be amplified in neuroblastoma cell lines [228] and tumours [200] indicating a poor prognosis.

Early studies on oncogenes and tumour suppressors led Kenneth Kinzler and Bert Vogelstein to postulate a progression of cancer from preneoplastic lesions to malignant tumours with accumulation of mutations in multiple steps [229]. These pioneering studies on molecular genetics of cancer development used low throughput techniques such as linkage analysis, Giemsa banding, gene cloning, DNA sequencing also called "Sanger sequencing" and *in situ* chromosome hybridization allowing simultaneous study of only single genes or loci, limiting their scope of analysis. Of these, Sanger sequencing [230] was instrumental in facilitating the assembly of the genomes of viruses phi X174 [231], SV40 [232], followed by the human mitochondrial genome [233]. The assembly of the human genome itself was a considerably tougher challenge costing 2.7 billion USD (1991 values) [234] and was greatly facilitated by the method of shotgun sequencing [235]. Genomes of model organisms were similarly assembled, supporting experimental studies of oncogenic processes and treatment modalities [236] [237]. While the first human reference genome was based on few individuals, the "1000 Genomes Project" expanded our knowledge of population-based variation in human genomes from 2504 individuals across 26 populations using the methods of whole genome sequencing and genotyping by single nucleotide polymorphism (SNP) arrays [238].

Development of high-throughput analytical techniques and the availability of reference genomes enabled broader and more advanced analyses of the cancer genome, starting the era of cancer omics. As the genome (DNA), transcriptome (RNA), epigenome (DNA methylation and histone marks), proteome (proteins) and metabolome (metabolites) are all dysregulated in cancer, cancer research uses methods of analytical chemistry and computational analysis to measure and analyse all of these biological processes. A typical cancer genomics project focuses on one or more of these processes and data types in a cohort of cases representing a disease of interest. This is followed by more detailed investigation of findings such as novel oncogenes or dysregulated pathways, which can form the basis of preclinical drug development in an area of research named translational genomics [239] [240].

RNA microarrays based on oligonucleotide probes of complementary DNA (cDNA) were the first high-throughput transcriptome analysis technology [241]. Using selected probes in the order of tens of thousands representing genes, enables quantitative molecular profiling of tumours based on gene expressions. Gene expression based molecular profiling and classification was applied to gastrointestinal cancers [242], lung adenocarcinoma [243] and breast cancer [244]. On the example of breast cancer, gene expression profiling made a major clinical impact: Breast cancer subtypes based on gene expression profiling have first been introduced by hierarchical clustering [245] and the field's consensus converged to a set of 5 intrinsic expression subtypes with strong predictive powers for prognosis [246]. Later, these expression subtypes were used to train a classifier based on 50 genes [247], which did not need microarray or RNA-Sequencing data and could be run on paraffin embedded tissue using a cost-effective quantitative polymerase chain reaction (qPCR) assay [248]. This classifier was named Prediction Analysis of Microarray 50 (PAM50), later commercialized as the Prosigna Breast Cancer Gene Signature Assay established in clinical practice [249].

RNA Sequencing extended the transcriptome analysis capacities offered by microarrays [250], [251]: sequencing of individual transcripts enables the detection of unannotated tran-

scripts [252] and *de novo* transcriptome assembly [253], alternative splicing and alternative transcript start site usage [254], detection of genomic variants on the transcriptome level [255], detection of gene fusions [256]. RNA-Seq has therefore largely replaced microarrays in cancer research. The Genotype-Tissue Expression (GTEx) project was an international effort to study various tissue-specific aspects (gene expression, alternative transcript usage, protein truncation by variants) of the transcriptome in healthy individuals using RNA sequencing [257].

DNA analysis in cancer, "cancer genomics" improved greatly under evolving technologies:

- Already in 2004, a first census of 291 human cancer genes was published based on established knowledge from low throughput technologies before results from larger cancer genome sequencing studies were available [258]. Using such known gene lists, 188 lung cancer cases were sequenced with targeted sequencing of 623 polymerase chain reaction (PCR) amplified genes in 2008 [259]. A mass spectrometry based large-scale analysis of 1000 tumours across 17 cancer types investigated 238 targeted cancer genes in 2007, reporting a diverse set of mutations [260], but mass spectrometry was not widely used thereafter in cancer genomics.

- The first (2006) large-scale sequencing of genes in cancer was on breast and colorectal cancer genomes, where 13023 genes were sequenced by PCR amplification followed by Sanger sequencing [261]. This study served as a pilot for future whole exome and whole genome sequencing projects with more cost-effective technologies, with study design recommendations presented here guiding future projects.

- A series of very expensive and pioneering cancer genome sequencing projects at a near whole-exome level with the cost of around 100000 USD per case are published using PCR amplification followed by Sanger sequencing: breast and colorectal cancer 11 cases each (2007) [262], pancreatic cancer 24 cases (2008) [263], glioblastoma multiforme 22 cases (2008) [264].

- A series of new sequencing technologies called next-generation sequencing (NGS) or short-read sequencing drastically lowered sequencing costs with the tradeoff of higher computational complexity of data analysis and increased difficulty of recovering higher order genome structure (SV detection, genome assembly) [265] [266] [267] [268] [269]. NGS allowed the expansion of the scope of cancer genome sequencing by enabling whole genome sequencing (WGS) and making the whole genome sequencing (WES) of larger cohorts possible.

- The first cancer WGS projects are published with limited cohort sizes, revealing mutations of the non-coding genome and structural variation outside of the scope of WES: breast cancer 1 case (2009) [270], prostate cancer 7 cases (2011) [271], multiple myeloma 23 cases (2011) [272], colorectal cancer 9 cases (2011) [273].

Following this very early era of cancer genomics pre- and post- NGS [274], the United States' Cancer Genome Atlas (TCGA) [275] and the International Cancer Genome Consortium (ICGC) [276] acquired, analysed and presented integrative genomics, transcriptomics, epigenomics and to a limited extent proteomics data from a broad range of cancers, sharing

them openly with the public, cancer research and clinical communities [277] [278]. Genome sequencing costs have maintained their downward trend to allow WGS analysis of very large cohorts with paired normals such as 500 breast cancer genomes [279] or 491 medulloblastoma genomes [211], yielding insights on mutational processes in cancer [280] and enabling use of WGS in a clinical setting [281]. Thanks to large-scale cancer WES and WGS studies, the accumulated knowledge of cancer related genes and their mutations expanded dramatically [282], with the curated Catalogue Of Somatic Mutations In Cancer (COSMIC) presenting 719 cancer genes with "almost 6 million coding mutations across 1.4 million tumour samples, curated from over 26000 publications" as of 2018 [283], underlining the current scope of cancer genomics. The article titled "Comprehensive Characterization of Cancer Driver Genes and Mutations" published by Bailey and colleagues in 2018 as part of the TCGA Pan-Cancer Atlas project concludes with the statement that the field of cancer omics research is expected to enter a new era, moving beyond gene-centric analysis of driver mutations and towards integration of other data sources such as the tumour microenvironment [282]. Based on the numerous large-scale studies on most cancer types characterizing their mutational landscapes, this statement might be true, with less opportunities to reach insights with a genomics-only strategy. However, the TCGA's analysis was mostly based on whole exome sequencing data and constrained to the coding genome. An unpublished preprint from the PCAWG structural variation working group states that the accumulated WGS dataset analysed in the largest WGS pan-cancer analysis to date is still insufficient to address the diversity and the complexity of the structurally altered cancer genome, recommending both larger WGS studies and use of newer technologies to improve SV research in cancer [284].

The analysis of the cancer epigenome encompasses the methylome, histone marks, chromatin accessibility and chromatin interactions with dedicated methods for each data type:

- The cancer methylome can be analysed by dense methylation arrays [285] or the much denser whole genome bisulfite sequencing (WGBS) technology measuring every CpG, first applied to colorectal cancer [21]. Methylation array analysis was a central technique in the TCGA projects with 9759 of 11286 analysed cases across 33 cancer types having available methylation array data [275]. Methylation arrays can be used for cell type determination, promoter methylation status and copy number analysis, while WGBS can additionally reveal partially methylated domains, variably methylated regions, high-resolution methylation profiles of individual genes as shown on medulloblastoma [286] and B cell lymphoma [287].

- Chromatin immunoprecipitation (ChIP) sequencing is the capture of DNA binding proteins such as histones or transcription factors with post-translational modification-specific antibodies, followed by the isolation and sequencing of the DNA sequences bound to the captured target proteins. This technology has been instrumental in defining chromatin states of healthy tissues [288] and cancer types like medulloblastoma [289].

- Chromatin accessibility was first assessed by DNase-I hypersensitive sites [32]. Development of the rapid and cost-effective Assay for Transposase-Accessible Chromatin using sequencing (ATAC-Seq) assay [290] and improvements allowing its use in frozen

tissues [291] led to its wide-spread application on cancer genomes with data released on 410 tumours from 23 cancer types by the TCGA [292].

- Chromatin interactions can be analysed by chromosome conformation capture techniques with different scopes addressing different biological questions: 4-C can be used to investigate interactions of a selected regulatory region with all other genomic regions it is interacting with in a targeted manner [293], often used as a validation in SV studies. On the other hand, Hi-C is used to analyse genome-wide chromatin interactions [294], enabling the discovery of the TAD concept and definition of TAD boundaries [33]. Hi-C can be combined with chromatin immunoprecipitation to study both chromatin interactions and a target protein such as a H3K27ac for defining active enhancers or [295]

Using these analytical techniques, broad and high-quality repositories of epigenome data were made available to the scientific community by international consortia such as the Encyclopedia of DNA elements (ENCODE) [296] [297] and the NIH Epigenome Roadmap [298].

(Methods of cancer proteomics and metabolomics will not be introduced here as they are outside of the scope of this dissertation.)

A central theme in omics research is integration of different layers of omics data [299]. Integrative omics goes beyond the identification of genomic variants or dysregulation of gene expression or disease subtypes defined by methylomes, but rather integrates these data types arising from interdependent biological processes. For instance, the link between oncogene amplifications such as *MYCN* or *EGFR* and the corresponding increases in gene expressions between amplified and non-amplified subpopulations in a cohort integrate genomic and transcriptomic data. Taking this one step further, the "enhancer hijacking", i.e. activation of the proto-oncogene *GFI1*/*GFI1B* by a structural variant in the Group 3 and Group 4 subtypes of medulloblastoma integrates methylome-based classification, oncogenic overexpression, genomic structural variant information and histone marks indicating chromatin states [132]. Deeper analysis of this example not explicitly discussed in [132] could yield insights such as lack of mutations co-occurring with *GFI1*/*GFI1B*, suggesting them to be sole driver mutations in another level of genomic integration. Similarly, description of transcriptomic changes driven by *GFI1*/*GFI1B* activation in medulloblastoma would constitute another level of transcriptomic integration. Integrative omics presents a challenge both in terms of data analysis, visualization and sharing.

High-throughput data generation enabled by omics technologies co-evolved with corresponding algorithms to address the data analysis challenges posed by its nature. Cancer genomics, transcriptomics and epigenomics all make use of advanced bioinformatics algorithms to process raw high-throuput data output of omics assays to results that can be analysed towards dissecting the complex biology of cancer samples and cohorts. In a typical cancer omics project analysing multiple omics data types, a diverse set of bioinformatics tools and algorithms are used. Vice versa, large genome sequencing projects spur the development and optimization of bioinformatics tools and workflows.

In genomics data analysis, the bioinformatics workflow starts from the alignment of raw short reads produced by the sequencer to the used reference genome. Sequence alignment algorithms evolved from the BLAST algorithm [300] to the fast BWA algorithm [301] based

on the Burrows-Wheeler transform [302], which established itself as the de facto standard for the alignment of genome sequences obtained by NGS. DNA alignment is followed by variant calling for various genomic variant types. Single nucleotide variants (SNVs) can be called by the mpileup algorithm under the Samtools suite [303] or the CaVEMan algorithm [304]. Algorithms like DeconstructSigs are used to investigate mutational signatures taking the SNV output as their input [305]. An algorithm like Platypus is used to call small insertions and deletions [306]. SNV and indel calling is followed by variant annotation on criteria such as genomic context, predicted impact, novelty or lack of novelty, clinical actionability using a tool like VEP [307]. Larger scale variants such as SVs and copy number variants (CNVs) are called by other dedicated tools like Delly [308] and ACEseq [309], respectively. Annotation and interpretation of impact of SVs is a task requiring more progress. Tools like CESAM [100] are promising regarding the study of SV impact with a multi-omics strategy combining genome transcriptome and chromatin interaction data.

Transcriptomics by RNA (cDNA) microarrays requires normalization to regress out variation due to hidden technical variables like dye imbalances [310]. Normalized gene expressions obtained by RNA microarrays can be compared between groups using the limma package [311]. RNA Sequencing also requires an alignment step similar to DNA sequencing, but the aligner needs to be "splice-aware", meaning that reads obtained from spliced transcripts should be correctly aligned to exons that are separated by introns with gapped alignments. STAR [312] performs well in its two-pass mode in benchmarks [313] and is a commonly used RNA aligner. Counts of reads aligned to the reference genome in a splice-aware manner are counted by a tool like featureCounts [314] which are then normalized by a normalization algorithm like TMM [315]. Normalized transcript counts of RNA-Seq data can be analysed in differential comparison using DESeq which can address confounders in the data by integrating metadata information [316]. Alternative splicing and alternative transcript usage is typically analysed by the DEXSeq algorithm [317]. RNA-Seq enables gene fusion detection by chimeric transcript alignments in STAR or dedicated algorithms like deFuse [318].

Methylome array analysis can be done using Minfi [319] or RnBeads [320]. Copy number profiles of tumour samples can be extracted using methylome array data with the conumee [321] or ChAMP algorithms [322]. *(WGBS algorithms and tools are omitted as they are outside of the scope of this dissertation)*

Multi-omics data integration via statistical algorithms is a recent development in cancer omics, and is used for the identification of latent variables spanning multiple omics layers such as copy number profiles, transcriptome and methylome. MOFA [323] and iCluster [324] are among the tools to cluster cases using data from multiple omics layers.

General statistical analysis of high dimensional data such as transcriptome or metyhlome analysis encompasses dimensionality reduction algorithms based on manifold learning like t-SNE [325], clustering by a diverse selection of algorithms which are selected based on the expected structure of the data with HDBSCAN being one example unifying hierarchical and density-based approaches [326]. Dimensionality reduction, clustering and further supervised and unsupervised machine learning algorithms are available in the Scikit-learn library [327].

Processed, analysed and integrated multi-omics data is presented in internal meetings or

with external audiences using data visualization tools. General-purpose tools such as ggplot2 [328] produce static images based on a descriptive grammar of data visualization. Modern data visualization approaches prioritize interactivity with D3.js being the most commonly used tool for interactive graphics [329]. The data analysis requirements of the field of (cancer-) omics have motivated the development of dedicated data visualization tools such as ggbio specializing in the display of local genomic regions [330], ComplexHeatmap for advanced heatmaps and oncoplots[331], and Circos for interactions of genomic loci in SVs or as part of chromatin interactions[332].

This prevalence of computational tools in cancer omics research requires expertise in their development and use in collaborative projects. As science becomes more interdisciplinary [333], cancer research is no exception to this trend [334] [335]. With the collaborative work of medical doctors, biologists, bioinformaticians and computational biologists, a contemporary cancer omics project presents challenges in data sharing and communication between experts from different scientific backgrounds. This is partially addressed by availability of omics data visualization portals such as the cBioPortal [336] and R2 [337] facilitating access of non-computational scientists to high-throughput multi-omics data.

With the advanced and cost-effective analytical chemistry methods yielding high-throughput omics data, which are analysed by dedicated bioinformatics algorithms, the state of the art in cancer omics research offers the availability of large and high quality multi-omics datasets. The current aims of omics technology-driven cancer research are broad and cover both the fundamental biology of cancer and clinical applications. Some of them are:

- Determining mutational drivers and biological mechanistic properties of rare cancers and rare subtypes of common cancers.
- Analysing the similarity of cancer types in pan-cancer analyses [275] for rational design of basket clinical trials [338].
- Determining molecular biomarkers for prognosis, treatment response and resistance mechanisms for targeted treatments, establishing a knowledge basis for personalized medicine [339] [340] [281].
- Developing new and improved statistical methods and bioinformatics tools for computational biology, broadly supporting each field of cancer omics. This includes more accurate or faster sequence alignment algorithms, more sensitive or specific variant calling algorithms, statistical methods to integrate high-dimensional data across data layers or to reduce them to lower dimensions, better and more intuitive and interactive ways to visualize the complex datasets obtained in multi-omics.

In this context, my doctoral research as a bioinformatician and computational biologist aimed to address the following research questions, each presented in individual chapters of this dissertation titled *Integrative Analysis of Omics Datasets*:

I. **Replacing the CNS-PNET Superentity with Four Novel Molecularly Defined Entities Driven by Structural Variants**: My work as a computational biologist contributed to the comprehensive description of central nervous system primitive neuroectodermal

tumours (CNS-PNETs). In this study we applied integrative omics data analysis of methylomes, genomes and transcriptomes revealing that CNS-PNETs are a combination of a large group of misdiagnosed cases from other entities and four novel molecularly defined entities. I showed that these novel entities are driven by distinct and recurrent molecular drivers altered by different mechanisms of structural variants: the *FOXR2* oncogene and *MN1*, *CIC* and *BCOR* tumour suppressors. Our results contributed to the elimination of CNS-PNETs as an officially recognized cancer entity and the recognition of four novel paediatric brain tumour entities in the WHO classification of brain tumours.

II. **SOPHIA, Structural Rearrangement Detection Based on Supplementary Alignments and a Population Background Model**: Building on my work on structural variation in our study of CNS-PNETs, I developed the **SOPHIA** algorithm for detecting SVs in cancer genomes based on a large population background database and a corresponding bioinformatics tool written allowing fast detection of SVs from short read cancer genome sequencing datasets. SOPHIA later became the standard tool for SV detection in the DKFZ's cancer genome analysis workflow.

III. **EPISTEME, an Interactive and Integrative Platform for Analysing, Interpreting and Sharing Multi-Omics Data**: During the development of SOPHIA and my research in projects analysing and interpreting structural variant data, I developed experiences analysing SV data detected by SOPHIA, integrating them with different omics layers such as gene expressions, interpreting, visualizing and sharing them with collaborators who were not computational scientists. Based on these experiences and using modern tools of interactive data visualization, I developed an interactive platform for integrative omics data analysis and visualization named **EPISTEME**, with the aim of facilitating omics data analysis by scientists with conceptual knowledge of cancer omics but no programming skills. EPISTEME is a comprehensive tool integrating genome, transcriptome, methylome and proteome data with clinical metadata in a user-friendly web-based system with in-browser statistical analyses and publication-quality vector graphics output.

IV. **SOPHIA-EPISTEME integration in DKFZ Cancer Genomics Projects Reveals Novel Disease Subtypes and Insights Across Cancer Types**: With the integration of SOPHIA and EPISTEME in an integrative omics data analysis setting, my work identified novel oncogenes activated by enhancer hijacking and revealed novel molecularly defined subtypes in refractory multiple myeloma (*MYCN* enhancer hijacking via immunoglobulin rearrangements as a *MYC* replacement), adult acute myeloid leukaemia (*MNX1* activation via enhancer hijacking putatively acting as a differentiation block mechanism) and paediatric neuroblastoma (*ATOH1* activation via enhancer hijacking putatively acting as a *MYCN* replacement) in projects supported by the DKFZ-HIPO and GPOH cancer research programmes.

(The source code used in the generation of the data presented in all parts of this dissertation is available from the repository `https://github.com/umut-h-toprak/PhD_`

`Dissertation_codebase`. *Externally used repositories are cited and documented as appropriate.)*

CHAPTER 1
REPLACING THE CNS-PNET SUPERENTITY WITH FOUR NOVEL MOLECULARLY DEFINED ENTITIES DRIVEN BY STRUCTURAL VARIANTS

1.1 Introduction

In human cancers, the cell of origin is a strong indicator of disease progression, sensitivity to treatment and is consequently an integral part of establishing the correct diagnosis and prognosis [275]. The cell of origin also indirectly, via chromatin states, influence the mutational profile of a tumour [341]. Before the quantitative tools of omics- or array-based technologies measuring the transcriptome or the methylome of a sample were available for molecular pathology, the only means of establishing a tumour classification was a combination of radiology, classical histopathology and the experience of physicians. This approach works well for tumour types that have distinct localization patterns, macroscopic manifestations, histopathological signatures of cell shapes or established molecular markers with available antibodies [342]. For example, histopathological classification of invasive breast carcinoma follows a two-pronged strategy involving tumour morphology (lobular, tubular, cribriform, medullary-like, micropapillary, papillary, metaplastic, no-special-type), and marker status (oestrogen receptor positive, progesterone receptor positive, epidermal growth factor receptor *HER2*/*ERBB2* positive, triple-negative) [343]. Thanks to the low diversity of cell types in mammary tissue, there is no ambiguity in the diagnosis of breast carcinoma by classical histopathology. However, tumours with dedifferentiated/anaplastic/high-grade or primitive cells of origin present a more difficult challenge to pathologists.

Tissues harbouring diverse cell types such as the brain or mesenchymal tissue from which sarcomas arise are another challenging diagnosis and research question to pathologists. Especially soft tissue sarcoma has a considerable variety of subtypes (and implicitly cells of origin) with more than 100 described in the latest WHO classification [344]. Similarly, the human brain both in its adult and developing form, harbours a diverse set of cell types with distinct transcriptomic (and implicitly epigenomic) profiles [257] which have recently been described with further precision using the new technology of single-cell RNA sequencing [345] [346]. This diversity of cell types also reflects in the diversity of tumours originating from the central nervous system [347].

Frequently, methods of classical histopathology had been found to be insufficient in addressing the diagnostics needs for the diverse landscape of brain tumours. For instance, in the case of Medulloblastoma, a paediatric tumour type with great clinical significance, age and resection/metastasis status based risk estimation supporting classical histopathology was considered insufficient [348] [349], and was supported by methods such as transcriptome profiling [350], mutation detection [351]. Methylome profiling is based on the hypothesis that the cell of origin determines the type and correct diagnosis of a tumour even in the absence of appropriate histopathological markers, and that cell of origin is imprinted in the methylome which can be quantified by methylome array of whole genome bisulfite sequencing (WGBS) technologies.

Methylome profiling has been successfully applied as a critical component of ICGC-Pedbrain for classification of tumours into molecularly defined entities starting with molecular classification of medulloblastoma[352]. The method has been iteratively improved culminating in a recent landmark study describing the molecular classification of "approximately 100 known tumour types" of the central nervous system [353]. The study describes a mature, comprehensive and accessible set of methods and a web portal for molecularly classifying tumours of the central nervous system with the claim of outperforming classical histopathology, with 12% of analysed cases correctly reclassified thanks to molecular classification analysis.

A key milestone in the progression to this mature molecular classification methodology was the detailed dissection of the CNS-PNET superentity. *CNS-PNETs* is an abbreviation for Central Nervous System Primitive Neuroectodermal Tumours. As the name implies, the cell of origin is "primitive" cells of neuroectodermal lineage. This group of embryonal tumours originally encompassed the medulloblastoma entity, previously named as PNET of the Cerebellum [354]. Cerebral PNETs are considerably rarer at 20% incidence and generated controversy in the history of pathological diagnosis [355]. With unknown driver genes, a difficult pathological profile making them prone to misdiagnosis and aggressive clinical characteristics, dissecting the biological background of this superentity was of great clinical and research interest. Equipped by the methylome-based classification method developed and implemented during ICGC-PedBrain, an international collaboration was set up involving numerous centres with large biospecimen banks worldwide (Germany: main coordinator groups Pfister & Kool at DKFZ and Korshunov at Uni. Heidelberg, USA: main coordinator group Ellison at St. Jude Children's Research Hospital, Canada, UK, Australia among others) in order to collect rare CNS PNET specimens with sufficient quality for molecular assays and sufficient quantity for letting subtypes emerge.

This chapter of my dissertation describes this study where we described a tumour superentity, CNS-PNETs to be a combination of a large group of misdiagnosed cases from other entities and four novel molecularly defined entities with distinct molecular drivers and presented our results to the community in the journal *Cell* [218]. In this study, I was the leading bioinformatics contributor in the effort of dissecting the molecular mutational drivers of the four novel brain tumour entities. For each of these four novel tumour entities, I managed to present a convincing driver gene and mechanism(s): *MN1* fusions, *CIC* Fusions, *BCOR* Internal Tandem Duplications, *FOXR2* activation by Structural Rearrangements. My contributions to this study encompass the methods and results Sections 1.2.7, 1.2.11, 1.2.12, 1.2.13, 1.3.2, 1.3.3, 1.3.4, 1.3.5 where the remaining sections are included to ensure the completeness of presentation in my dissertation. Parts of the study and the resulting article that are not directly related to my work as prerequisites or consequences are omitted for the sake of brevity and more clearly outlining my own contributions

1.2 Methods

1.2.1 Study Design

Participating centres provided specimens of cases diagnosed as CNS-PNETs from their pathological archives. Considering the high rarity of the disease, FFPE (Formalin-fixed, Paraffin-embedded) specimens were collected in addition to higher quality fresh-frozen specimens. Specimens that were unable to provide sufficient quality DNA for a methylome analysis were excluded from the study.

The aims of the study were to investigate if the CNS-PNET diagnosis truly has a high misdiagnosis rate, if it encompasses a new or multiple novel tumour entities and how histopathological diagnoses are in agreement with an expert panel review (Section 1.2.2) and with molecular classification and to describe the mutational drivers of any new tumour entities if applicable.

All collected samples were subjected to methylome analysis for classification and copy number profiling (Section 1.2.3) and gene expression analysis via Affymetrix microarrays for secondary classification and determination of candidate genes via differential expression analysis (Section 1.2.6) where sufficient RNA quality was available. After each sample was classified, samples belonging to novel tumour classifications with specimens where fresh-frozen material was available, were also submitted for sequencing by whole genome sequencing (along with matching blood controls) and RNA sequencing for determination of mutational drivers (Section 1.2.8).

1.2.2 Histopathological Review

For establishing a high quality histopathological diagnosis, each sample accepted in the study was reviewed by an experienced panel of experts that participated in the study (Brent A. Orr, David Capper, David W. Ellison, Andrey Korshunov). The results were classified as i) Histologically matching PNET: *Classic PNET*, ii) Small-cell tumours with increased nuclear content and without specific markers for definite differentiation between HGG (high grade glioma) and PNET: *PNET/HGNET*, iii) Cases where the PNET diagnosis is considered questionable upon expert review: *HGNET with different diagnosis favored*, iv) Cases where the PNET diagnosis is considered inaccurate upon expert review.

No samples were excluded from the study including cases classified as questionable or inaccurate upon expert review: while the study had the privilege of having access to the services of world-class neuropathologists, the study is designed to help reduce or solve the controversy around the CNS-PNET pathology worldwide, where the availability of such an expert panel review though desirable cannot be expected. Hence, the study investigates the full spectrum of CNS-PNETs based on institutional diagnosis.

In forming the consensus expert panel opinion, each case was systematically investigated for the following morphological signatures, which were then presented in the manuscript: 1) hemorrhage, 2) small cell, 3) cell size (small/intermediate/large), 4) neurophil, 5) ependymoblastic rosettes, 6) perivascular rosettes, 7) vasculo-centric, 8) Homer Wright rosettes, 9) mitoses per 10 hpf, 10) apoptosis, 11) apoptosis score (low/intermediate/abundant), 12) fibrillar processes, 13) infiltration, 14) necrosis, 15) secondary structures, 16) ganglion cells, 17) neu-

rocytes, 18) pallisades, 19) vascular proliferation, 20) capillary network, 21) hyaline, 22) calcification, 23) papillary growth, 24) pseudopapillary growth, 25) epithelial surfaces, 26) tissue amount (scant/low/intermediate/abundant), 27) myxoid material, 28) nucleoli, 29) rhabdoid cells, 30) giant cells, 31) inflammation, 32) anaplasia, 33) macrophages.

1.2.3 DNA Methylation Array Processing

DNA methylation profiling was performed using the Infinium HumanMethylation450 BeadChip array (450k array) following manufacturer instructions (Illumina, San Diego, USA) as described in [352]. The following participating centres contributed to the joint metyhlome profiling effort i) DKFZ Genomics and Proteomics Core Facility (Heidelberg, Germany), ii) St. Jude Children's Research Hospital (Memphis, USA), iii) NYU Langone Medical Center (New York, USA), iv) McGill University and Génome Québec Innovation Centre (Montreal, Canada).

For most fresh-frozen samples, $> 500ng$ of DNA was submitted for methylome analysis. $250ng$ of DNA was used for most FFPE tissues. Quality control was done by checking on-chip quality metrics and unexpected genotype matches by pairwise comparison of the 65 genotyping probes on the 450k array.

Data analysis was performed in R version 3.2.0 (R Development Core Team, 2015). Raw signal intensities were obtained from IDAT-files using the minfi Bioconductor package version 1.14.0 [319]. Samples were individually normalized by a background correction (shifting of the 5 % percentile of negative control probe intensities to 0) and a dye-bias correction (scaling of the mean of normalization control probe intensities to 10000)for both color channels. This approach was tested against the functional normalization method [356], and determined to perform similarly. Furthermore, probes were filtered/removed from the analysis with the following criteria for more accurate clustering i) Probes mapping to the X and Y chromosomes ($n = 11551$), ii) Probes containing a single-nucleotide polymorphism (dbSNP132 Common) within five base pairs of and including the targeted CpG-site ($n = 24536$), iii) Probes not mapping uniquely to the human reference genome (hg19) allowing for one mismatch ($n = 9993$). 438370 probes were kept for analysis following this filtering step.

1.2.4 DNA Methylation Based Clustering

DNA Methylation Based Clustering was performed in the following configurations: i) samples diagnosed as PNETs ($n = 323$), ii) reference samples where the histopathological and molecular analysis defines the tumour entity without ambiguity including some non-neoplastic brain tissue samples ($n = 211$), iii) the combination of PNETs and reference samples ($n = 534$). The PNET-reference combination configuration was performed with the aim of exploring the hypothesis that PNETs commonly include misdiagnosed cases because misdiagnosed PNETs would cluster with the reference samples where applicable, and pure clusters of PNET cases would indicate novel entities for which no reference samples exist. Following this step, the PNET-only clustering would reveal how cases diagnosed as PNETs are distributed between novel entities and known entities. The reference-only clustering is performed with the motivation of controlling if the chosen reference samples act as an unbiased reference.

Unsupervised hierarchical clustering of samples was performed on the 10000 most variably

methylated probes, with 1-Pearson correlation coefficient as the distance measure. Average linkage was used to generate dendrograms. The same distance matrix was used to generate t-SNE visualizations (t-Distributed Stochastic Neighbour Embedding [325], Rtsne package version 0.11) with the following non-default parameters: theta=0, isdistance=T, pca=F, maxiter=10000. Individual clustering of CNS-PNET samples, reference samples and additional CNS tumor samples was performed with a similar approach. Specimen type (i.e. FFPE or fresh-frozen) was checked and determined not to influence on unsupervised clustering in the form of enrichments of a specimen type within clusters.

The Kruskal-Wallis test was used to compare CpG methylation values between the four new CNS tumor entities. P-values were adjusted for multiple testing using the Benjamini-Hochberg procedure, followed by Dunn's test for post-hoc pairwise comparisons. CpG sites were reported entity-specifically methylated if for all pairwise comparisons $p < 0.001$ and pairwise mean methylation difference $> 1/3$.

Following the identification of the four novel tumour entities, a larger repository of CNS tumour methylomes ($n > 10000$) was used in a correlation analysis to identify further cases where the initial diagnosis was not a PNET but the molecular pathology indicates an incidence of one of the four novel entities. A new set of reference samples including the four new entities was formed ($n = 159$) and was subjected to methylome clustering as described with candidates from the master repository matching the four new entities ($n = 59$). This was done with the motivation of maximizing the number of captured cases of novel molecular groups, including previously "missed" cases.

1.2.5 DNA Methylation Based Copy-Number Variation Analysis

Copy-number variation (CNV) analysis was performed using the conumee Bioconductor package version 1.0.0. Two sets of 50 control samples displaying a balanced copy-number profile from both male and female donors were used. Scoring of focal amplifications and deletions and chromosomal gains and losses was performed by manual inspection of each profile.

1.2.6 Gene Expression Array Processing

Samples for which RNA of sufficient quantity and quality was available were analysed on the Affymetrix GeneChip Human Genome U133 Plus 2.0 Array at the Microarray Department of the University of Amsterdam, the Netherlands. Sample library preparation, hybridization, and quality control were performed according to manufacturer's protocols. Expression data were normalized using the MAS5.0 algorithm of the GCOS program (Affymetrix, Santa Clara, USA). Processed data was deposited in the R2 platform in a private session totalling 2273 brain tumours and healthy brain tissue controls to facilitate analysis [337].

1.2.7 Differential Gene Expression Analysis for Candidate Gene Discovery

We used the R2 platform to perform differential expression analysis and subsequent gene expression visualizations [337]. The statistical test ANOVA was applied to the microarray dataset normalized as explained in Section 1.2.6. Results were filtered to contain p-values below 0.01 and corrected for multiple testing using the Benjamini-Hochberg procedure.

Each of the four novel tumour entities with unknown candidate driver genes was subjected to a differential expression analysis in a "1 vs 3" configuration with the intrinsic assumption that each group is embryonal and the inter-entity differences would allow a specific differential expression analysis revealing tumour pathways rather than only strongly enriching cell of origin related differences.

Following the identification of candidate overexpressed/underexpressed genes, candidate genes were plotted in a box-dotplot configuration with all tumour entities included in the $n = 2273$ R2 brain tumour repository. These include 35 tumour types and 1 group of normal brain tissue. The aim here was to indicate if results from the "1 vs 3" comparisons also held in a more general analysis without pooling a large number of tumours from very different brain cell lineages. Each outlier candidate was thus visually confirmed for novel group specificity. The shortlisted candidates were prioritized for further cross-omics-layer analysis in terms of dysregulation by gene fusions or enhancer hijacking events using next generation sequencing.

1.2.8 Next Generation Sequencing

As described in [357], paired-end (PE) DNA library preparation was carried out using Illumina, Inc. v2 protocols. In brief, $1 - 5\mu g$ of genomic DNA were fragmented to ~300 bp (PE) insert-size with a Covaris device, followed by size selection through agarose gel excision. Deep sequencing was carried out with the HiSeq2000 instrument.

Paired-end RNAseq libraries were prepared with purified polyA+ RNA fractions using methods preserving the strand specificity, following the dUTP-based protocol as described in [358], featuring cDNA fragmentation after mRNA priming with a mixture of random hexamers (dN)6 and oligo (dT) primers. A fraction of the libraries was constructed with a modified protocol where the polyA+ RNA fraction was fragmented at $70°C$ for 5 minutes using RNA fragmentation reagents (Ambion, Cat. #AM8740),according to the manufacturer's instructions; first strand synthesis was then performed with random hexamers (dN)6 only (and the cDNA fragmentation step was omitted).

1.2.9 Next Generation Sequencing Analysis: Alignment of DNA sequences

For each sequencing lane, read pairs were aligned against human reference genome including decoy sequences (hs37d5) using BWA mem [359] version 0.7.8 with default parameters and -T=0 . Single lane bam files were post-processed using biobambam [360] (version 0.0.148): the lanes were sorted by bamsort and were merged with duplicates marked using bammarkduplicates. This workflow is also known as the PCAWG workflow because it was used in the *Pan-Cancer Analysis of Whole Genomes* project as the uniform alignment workflow for all participating centres. This workflow forms the backbone of our whole-genome sequencing data processing in the DKFZ and is used for all data presented in this dissertation, including the other chapters. The workflow is described in [361] and is available from `https://github.com/ICGC-TCGA-PanCancer/Seqware-BWA-Workflow`

1.2.9.1 SNV Calling

As described in [357], single nucleotide variant (SNV) detection integrates publicly available tools with custom in-house software and applies several filtering and annotation steps. SNV calling is based on samtools mpileup [303] and bcftools [362] (version 0.1.17), using parameter adjustments to allow calling of somatic variants. Initial SNV candidates were identified by using samtools mpileup for each tumour sample considering only reads with a minimum mapping quality of 30 and bases with a minimum base quality of 13, after application of the extended base alignment quality (BAQ) model. BAQ is the Phred-scaled probability of a read base being misaligned [363], and it is designed to reduce false SNV calls caused by misalignments. After the pile up of high quality bases at each position of the input BAM file, bcftools applies the prior and performs the actual SNV calling. We changed the default probability of calling a variant if $P(ref\|D) < 0.5$ to 1.0, which results in all positions containing at least one high quality non-reference base to be reported as a variant. This initial SNV call set, rich in false positives, is further filtered: SNVs covered by fewer than three reads in the tumour and control sample, with somatic allele frequency $< 5\%$, or with only one read supporting the variant were excluded. If the variant call was supported by reads from only one strand, the ± 10 bases around the SNV were automatically screened for Illumina specific error profiles [364] and excluded if a profile was matched.

For all tumour SNV calls the pipeline generates a pileup of the bases in the normal sample considering only uniquely mapping reads. SNV calls were categorized as germline or somatic according to whether there was evidence for the same event at the same locus in the BAM file of the tumour-matched control sample.

This workflow described in [357] did not yield candidate driver genes (recurrently mutated) for the analysed novel tumour entities in the presented study. The analysis was performed by Dr. Ivo Buchhalter, was not included in the final manuscript and is presented here for the sake of completion

1.2.10 Next Generation Sequencing Analysis: InDel Calling

As described in [357], small insertions and deletions were identified with samtools [303] and bcftools [362]. The InDel discovery pipeline is similar to the SNV pipeline (as described above), but using default bcftools parameters, to reduce the known high false positive rate ($\sim 60\%$) associated with InDel detection methods for deep sequencing data. To call an indel a germline event, we only required one InDel supporting read in the matching normal sample, again to reduce the high fraction of false positive somatic InDel calls. Calls overlapping simple repeat or microsatellite regions were excluded as such regions are commonly observed to yield false positive calls. Annotation of InDels was identical to SNV annotation. All coding somatic InDel calls were manually reviewed using the Integrative Genomics Viewer (IGV) [365].

This workflow described in [357] did not yield candidate driver genes (recurrently mutated) for the analysed novel tumour entities in the presented study. The analysis was performed by Dr. Ivo Buchhalter, was not included in the final manuscript and is presented here for the sake of completion

1.2.11 Next Generation Sequencing Analysis: Analysis of Chimeric Fusion Transcripts

I detected chimeric fusion transcripts from RNA-Seq FastQ files by de novo annotation of fusion transcripts using the deFuse algorithm, version 0.6.1 [318] with default parameters using a pipeline developed by Dr. Zuguang Gu. Results were filtered to exclude chimera arising from differential splicing and chimera of adjacent genes. DNA breakpoints of the structural rearrangements leading to the chimeric fusions were then detected as described in Section 1.2.12.

1.2.12 Next Generation Sequencing Analysis: Analysis of Structural Rearrangements

Our in-house experiences with CREST [366], LUMPY [367] the publicly available version of Delly [308] at the time of the PNET study were not straightforward or positive (*personal communication*). Generally, Delly, in the hands of dedicated experts, was known to perform better with better specificity, but Delly was unable to capture mid-sized indels ($\sim 50 < l < \sim 1000 bps$) at the time this study was running. The analysis described in Section 1.2.7 revealed a small number of key candidate overexpressed genes and two of the four novel entities (Sections 1.3.2 & 1.3.3) were explained by the procedure described in Section 1.2.11. Thus, for the remaining two groups we decided from the ground-up to favour a manual inspection approach which had normally been applied in a validation setting as described in Section 1.2.10.

I used the Integrative Genomics Viewer (IGV) [365] to manually inspect the genomic neighbourhoods of the candidate genes obtained from the procedure described in Section 1.2.7. As recurrently mutated genes via small variants (SNVs and small InDels) were shown to be unlikely by the analysis described in Sections 1.2.9.1 and 1.2.10, I prioritized my search for structural rearrangements. After opening both the tumour and control alignments in IGV with the non-default alignment display parameters i) disable downsampling, ii) show soft-clipped bases, iii) do not filter supplementary alignments, iv) do not filter secondary alignments, I checked each gene's gene body and 5' and 3' neighbourhoods up to 2 megabases away. I specifically looked in the tumour alignments for features of soft-clipped reads with matching hard-clipped supplementary alignments falling on consistent genomic positions, with high quality bases in the split reads' overhang sequences and absence of such features in the control alignments. Each detected finding was validated by cross-checking other tumours in the same molecular type and also chemically validated using PCR or FISH (Dr. Dominik Sturm and colleagues) where sufficient quality DNA remained from the initial tumour specimen.

1.2.13 Data Availability

The data generated in this study was provided to the community in three sets, where I was responsible for setting up the controlled-access release of the next-generation sequencing data:

Dataset (*controlled)	Data Repository	Accession Number
Methylome Arrays	NCBI Gene Expression Omnibus	GSE73801
Gene Expression Arrays	NCBI Gene Expression Omnibus	GSE73038
NGS Data*	European Genome-phenome Archive	EGAS00001001632

1.3 Results

1.3.1 Methylome Clustering of Reference Paediatric Central Nervous System Tumours and CNS-PNETs Reveals a High Rate of Misdiagnosis and Novel Molecular Subgroups for CNS-PNETs

The merged clustering approach described in Section 1.2.4 revealed a striking pattern of embryonal tumours misdiagnosed as PNETs. More than half of the cases $196/323(61\%)$ clustered with other reference tumour entities, indicating a major clinical problem that should be addressed (Section 1.1). Upon further investigation of the entities most prone to misdiagnosis, we observed as the most frequent sources of misdiagnosis Embryonal Tumours with Multilayered Rosettes-ETMRs $(36/323, 11\%)$ [368] [369], *MYCN*-amplified high-grade gliomas-HGG$_{MYCN}$ $(28/323, 9\%)$ and *IDH*/*H3F3A* wild-type HGG from receptor tyrosine kinase (RTK) subgroups-HGG$_{RTK}$ $(28/323, 9\%)$ [370]. This is followed by entities that are less prone to misdiagnosis as PNET: *IDH*-mutant HGG-HGG$_{IDH}$ $(17/323, 5\%)$, *H3F3A* G34-mutant HGG-HGG$_{G34}$ $(17/323, 5\%)$, supratentorial ependymomas-EPN $(15/323, 5\%)$, AT/RTs $(14/323, 4\%)$, *H3F3A* K27-mutant diffuse midline gliomas-HGG$_{K27}$ $(10/323, 3\%)$, pineal tumors-PIN $(8/323, 2\%)$, Ewing sarcomas-EWS $(5/323, 2\%)$, choroid plexus carcinomas-CPC $(2/323, 1\%)$, pleomorphic xanthoastrocytomas-PXA $(1/323, < 1\%)$, or meningiomas-MNG $(1/323, < 1\%)$.

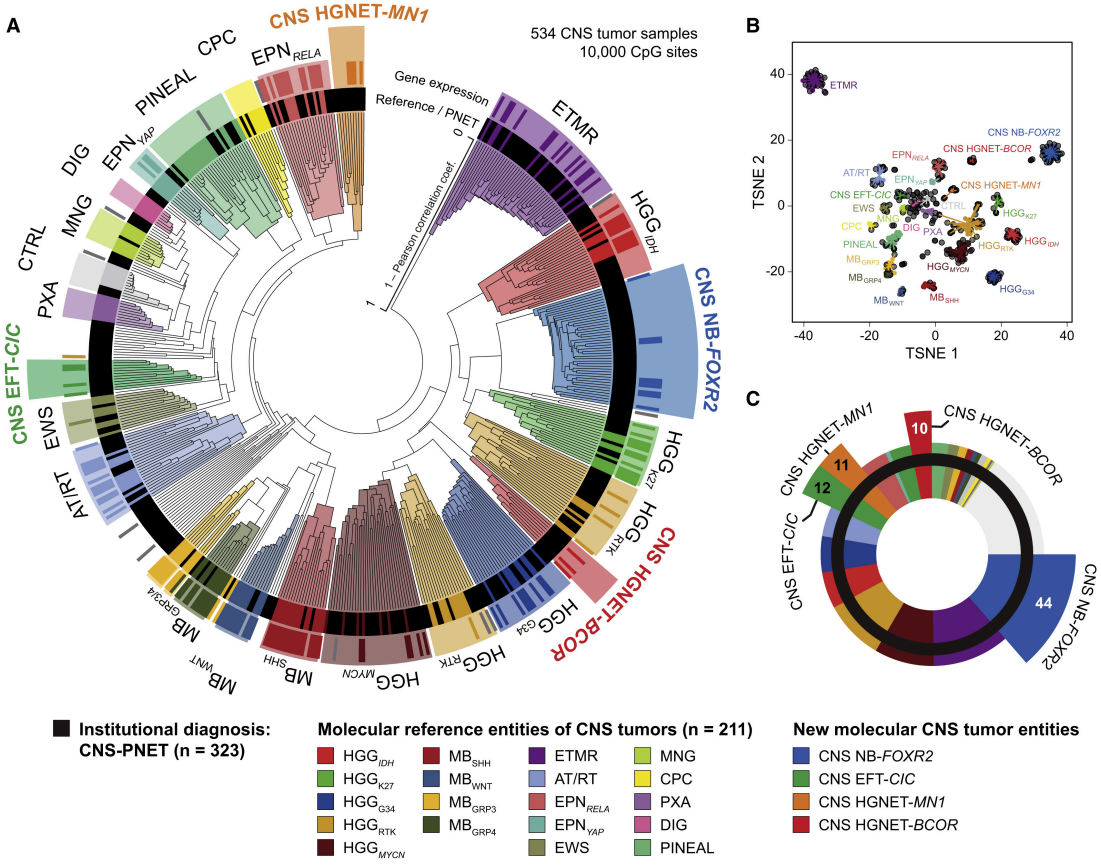

Figure 1.1: Merged clustering of case diagnosed as PNETs and reference tumours using methylation and gene expression profiles (Figure 1 of [218]). A: Methylome hierarchical clustering of 534 CNS tumour samples of which 323 are institutional PNET diagnoses (black bars) and 211 are reference specimens of confidently molecularly diagnosed reference instances of other CNS tumour entities (coloured bars). PNET-diagnosed cases disperse to diverse non-PNET entities, 4 novel molecularly defined entities and unspecified clusters. B:Methylome tSNE of the analysed CNS tumour cohort shows the 4 novel clusters as distinct clusters. C: The numbers of cases that were assigned to one of the 4 novel clusters from the original collection of 323 institutional PNET diagnoses.

Having established the previously hypothesized but not molecularly shown a high misdiagnosis rate involving PNETs, we searched for *bona fide "PNETs"* belonging to previously unknown entities. We observed that such cases exist and fall into five categories: i) small, inhomogeneous clusters (< 5 tumours) or distant outliers which failed to group with each other or any of the reference tumour entities, possibly representing exceedingly rare entities $(50/323, 15\%)$, and ii) four separate and homogeneous clusters clearly distinct from reference entities $(77/323, 24\%)$. At that point, the observation that there is not one but four "real" PNET clusters with distinct cells of origins and likely distinct drivers further emphasized the difficulties of PNETs for histopathological diagnosis.

The four new tumor entities were provisionally and finally named

i) PNET with chr1q gains and chr16q losses, PNET 1q-16q $\rightarrow$ "CNS neuroblastoma with

FOXR2 activation" CNS NB-*FOXR2* $(44/323, 14\%)$,

ii) Ewing Sarcoma-like PNET, PNET-EWS-like $\rightarrow$ "CNS Ewing Sarcoma family tumor with *CIC* alteration" CNS EFT-*CIC* $(12/323, 4\%)$,

iii) PNET with chr16q losses and chr22q losses, PNET 16q-22q $\rightarrow$ "CNS high-grade neuroepithelial tumor with *MN1* alteration" CNS HGNET-*MN1* $(11/323, 3\%)$, and

iv) PNET with WNT pathway activation, PNET-WNT $\rightarrow$ "CNS high-grade neuroepithelial tumor with *BCOR* alteration" CNS HGNET-*BCOR* $(10/323, 3\%)$.

These provisional names were proposed by copy number variation analysis and pathway enrichment analysis which will not be discussed here (details available in [218], Supplementary Figure 5, Figure 7 and Supplementary Figure 7). The provisional names were used during the analysis of our results leading to the presented study, before their respective driver genes and mechanisms were discovered (Sections 1.3.2, 1.3.3, 1.3.4 and 1.3.5) and their final names were determined.

Our next step was to confirm that the PNET diagnoses that did not fit the new four entities but rather other reference entities were indeed a correct match to those entities not only in terms of methylome (cell of origin) profiles but also with respect to known hallmark genomic alterations (Figure 1.2).

i) Cases of the ETMR cluster were checked following [369] for C19MC amplifications $(33/36, 92\%, p < 0.001)$ and high LIN28A protein expression $(17/17, 100\%;\ p < 0.001)$,

ii) Cases of the AT/RT cluster were checked following [371] for *SMARCB1* mutations and/or deletions $(14/14, 100\%;\ p < 0.001)$ and loss of the *SMARCB1* protein product INI-1 $(5/5, 100\%;\ p < 0.001)$,

iii) Cases of the HGG$_{IDH}$ cluster were checked for *IDH1* mutations using targeted sequencing $(15/15100\%;\ p < 0.001)$,

iv) Cases of the HGG$_{G34}$ cluster were checked for *H3F3A* G34 mutations using targeted sequencing $(17/17100\%;\ p < 0.001)$,

v) Cases of the HGG$_{K27}$ cluster were checked for *H3F3A* K27 mutations using targeted sequencing $(4/757\%;\ p < 0.001)$,

vi) Cases of the HGG$_{MYCN}$ cluster were checked for *MYCN* amplifications revealing $(20/2871\%;\ p < 0.001)$ and *MYCN-ID2* co-amplifications $(12/2843\%;\ p < 0.001)$. The latter was a novel observation at the time of this study

vii) Cases of the HGG$_{RTK}$ cluster showed copy-number alterations, and half $(14/28, 50\%)$ harboured focal amplifications and/or deletions of known oncogenes and/or tumor suppressor genes (*MDM2*(4), *CDK4*(3), *PDGFRA*(2), *MYCN*(2), *MYC*(1), *CDKN2A*(4), *PTEN*(2), *RB1*(1)),

viii) Cases of the EWS cluster were confirmed to harbour *EWSR1* rearrangements using FISH.

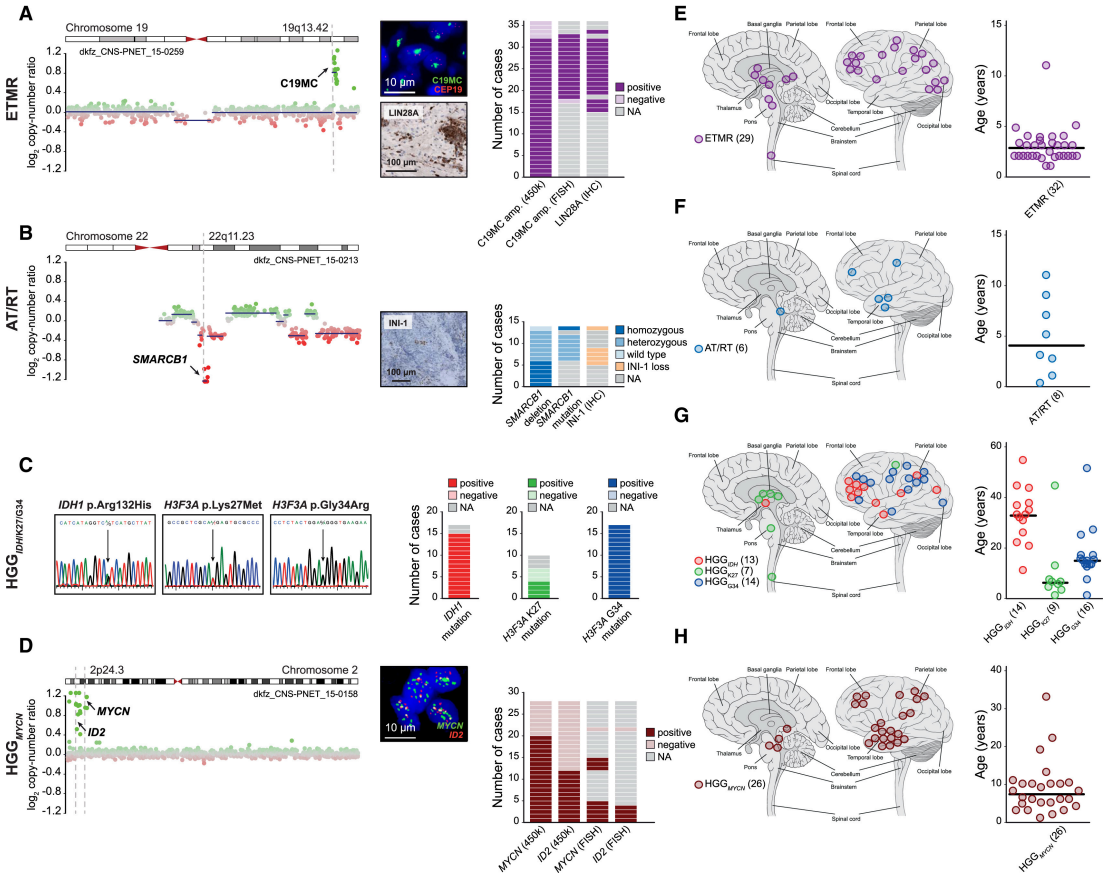

Figure 1.2: Confirmation of misdiagnosis of PNETs and accurate diagnosis into known non-PNET tumour entities (Figure 2 of [218]). A: PNET-diagnosed cases reclassified as embryonal tumours with multilayered rosettes (ETMR) harbour *C19MC* amplifications as shown by metyhlome based CNV analysis and FISH and express LIN28A as shown by immunohistochemistry (IHC), hallmarks of ETMR. B: PNET-diagnosed cases reclassified as atypical teratoid rhabtoid tumours (ATRT) harbour *SMARCB1* deletions as shown by metyhlome based CNV analysis and show loss of INI-1 expression (IHC) as hallmarks of ATRT. C: PNET-diagnosed cases reclassified as *IDH1* mutant or *H3F3A* mutant high grade glioma (HGG) harbour the respective hallmark mutations as confirmed by Sanger sequencing. D: PNET-diagnosed cases reclassified as *MYCN*-amplified HGG harbour *MYCN* amplifications as shown by metyhlome based CNV analysis and FISH. E-H: the reclassified cases show spatial and age-related distributions that are in line with established knowledge representing the entities.

Having thus confirmed the prevalence of misdiagnosis and the existence of four new tumour entities, we later used a larger methylome array repository encompassing all available brain tumours ($n = 10000$) to look for new cases that belong to the four new entities but were previously not diagnosed as PNETs and were thus not part of the initial candidate case list ($n = 323$) as described in the last paragraph of Section 1.2.4. We found the following 77 new cases with this approach (Figure 1.3)

i) PNET with chr1q gains and chr16q losses, PNET 1q-16q → ”CNS neuroblastoma with FOXR2 activation” CNS NB-*FOXR2*: 2 new cases,

ii) Ewing Sarcoma-like PNET, PNET-EWS-like → ”CNS Ewing sarcoma family tumor with CIC alteration” CNS EFT-*CIC*: 3 new cases,

iii) PNET with chr16q losses and chr22q losses, PNET 16q-22q → ”CNS high-grade neuroepithelial tumor with *MN1* alteration” CNS HGNET-MN1: 30 new cases,

iv) PNET with WNT pathway activation, PNET-WNT → ”CNS high-grade neuroepithelial tumor with *BCOR* alteration” CNS HGNET-*BCOR*: 24 new cases.

Both these counts and the presented results in Figure 1.3 A-B show an imbalance between the numbers of new recovered cases across different entities of previous diagnosis. This imbalance will be explained in the corresponding sections of the new entities (Sections 1.3.2, 1.3.3, 1.3.4 and 1.3.5).

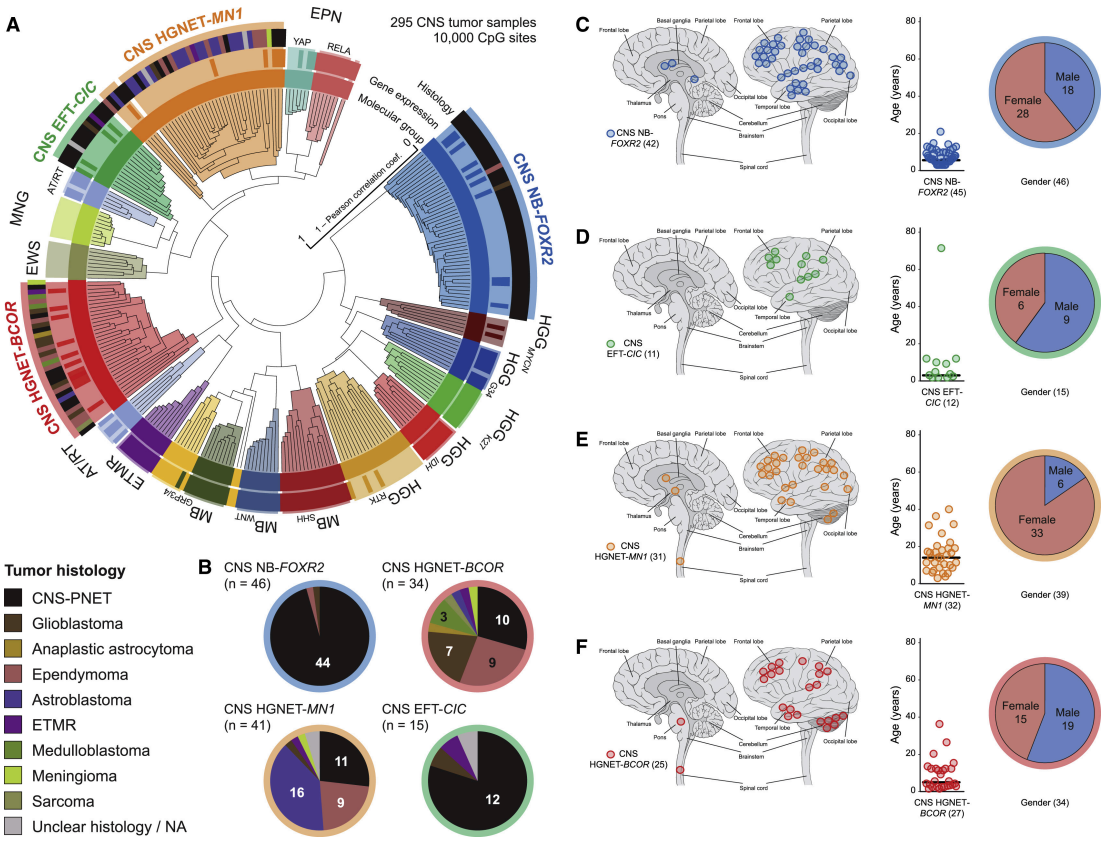

Figure 1.3: Recovery of additional cases belonging to the 4 novel entities from initial diagnoses of non-PNET entities (Figure 3 of [218]
). A: hierarchical clustering of the recovered 77 new cases, 159 reference samples and 59 additional samples reveals the recovery of additional cases to the 4 novel entities. B: The 4 novel entities represent different misdiagnosis incidence profiles with the CNS NB-*FOXR2* and CNS EFT-*CIC* groups representing the highest *bona fide* PNET diagnosis likelihood and the CNS HGNET-*MN1* group receiving a large number of cases from the astroblastoma entity. C-F: The spatial, age and gender distribution characteristics of the four entities following establishment of final cohorts from initial PNET diagnoses and later recovery from other entities as initially diagnosed.)

All new entities have homogeneous localization in the brain (Figure 1.3C-F). As per study design, we required a PNET diagnosis and cerebellar PNETs have historically been histopathologically diagnosed as Medulloblastoma and not as PNET. The inclusion of new cases by this analysis extended the entities from purely cerebral to cerebral and cerebellar. Overall, following the initial and secondary clustering, we obtained the following numbers.

Entity (Provisional)	Entity (Final)	N_{Total}	N_{WGS}	N_{RNAseq}	N_{both}
PNET 16q-22q	CNS HGNET-*MN1*	41	1	4	0
PNET EWS-like	CNS EFT-*CIC*	15	0	2	0
PNET WNT	CNS HGNET-*BCOR*	34	3	8	0
PNET 1q-16q	CNS NB-*FOXR2*	46	5	4	1

1.3.2 MN1 Fusions, mainly MN1-BEND2, Drive a Subgroup of CNS-PNETs

I ran a gene fusion analysis on the PNET 16q-22q subgroup as described in Section 1.2.11, which revealed *MN1* chimeric fusions in all samples with available RNA-Seq data. In $(3/4, 75\%)$ of the cases the partner was the *BEND2* gene, whereas the remaining case had a *MN1-CXXC5* fusion (Representative fusions in Figure 1.4). In the absence of other recurrent genomic alterations and considering the 100% recurrence of *MN1* fusions, we designated *MN1* as a primary candidate gene of interest for this entity.

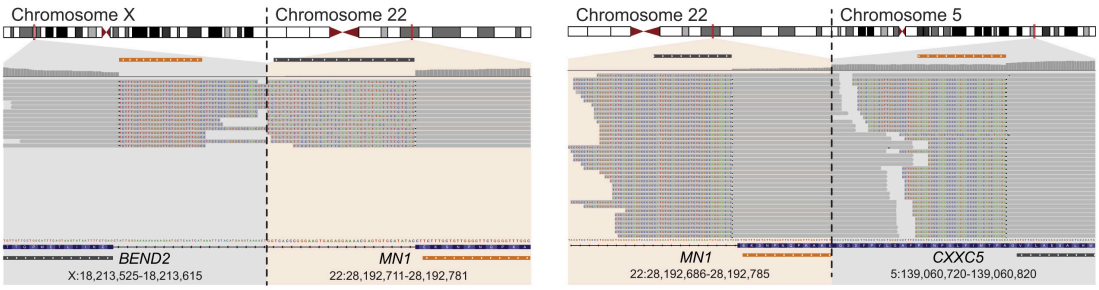

Figure 1.4: Representative *MN1* fusions *MN1-BEND2* and *MN1-CXXC5* as displayed on IGV. Extracted from Supplementary Figure 6 of [218]. Left: a *MN1-BEND2* fusion with breakpoints denoted by sharp coverage drops and gains and split read mapping to the partner site marked by orange and black coloured guide rectangles. Right: a *MN1-CXXC5* fusion case with the same visualization principles.

I first investigated if the *MN1* fusions lead to overexpression or suppression of *MN1* in this entity compared to other brain tumour entities and observed that *MN1* in the PNET 16q-22q group does not have a strikingly different expression profile (Figure 1.5).

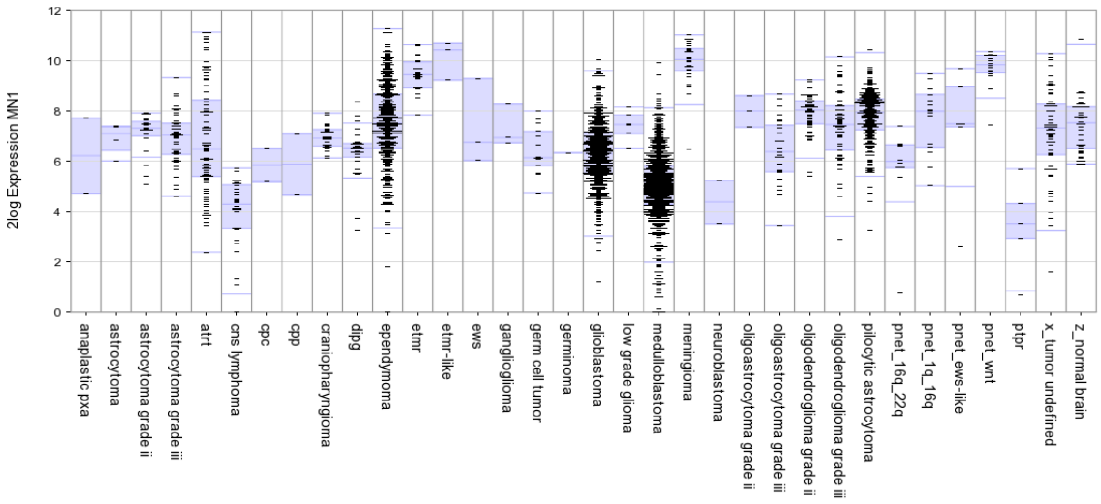

Figure 1.5: *MN1* expression across brain tumour entities shows no outlier characteristics in the candidate pnet-16q-22q subcohort (dataset ps-mkheidel-mkdkfz2273-u133p2 in R2 [337])

Next, I investigated if the frequent translocation partner of *MN1*, *BEND2* has a significantly different/altered expression profile and observed that *BEND2* is uniquely and recurrently activated in the PNET 16q-22q subgroup (Figures 1.6, and 1.7F). The cohort where this analysis was run is a microarray-based gene expression profiling cohort for which RNA-Seq data is not available apart from the aforementioned 4 cases. However, we observed that *BEND2* was activated in $(7/887.5\%)$ cases whereas the *CXXC5* fusion detected from RNA-Seq data remained the only exception also in the larger cohort.

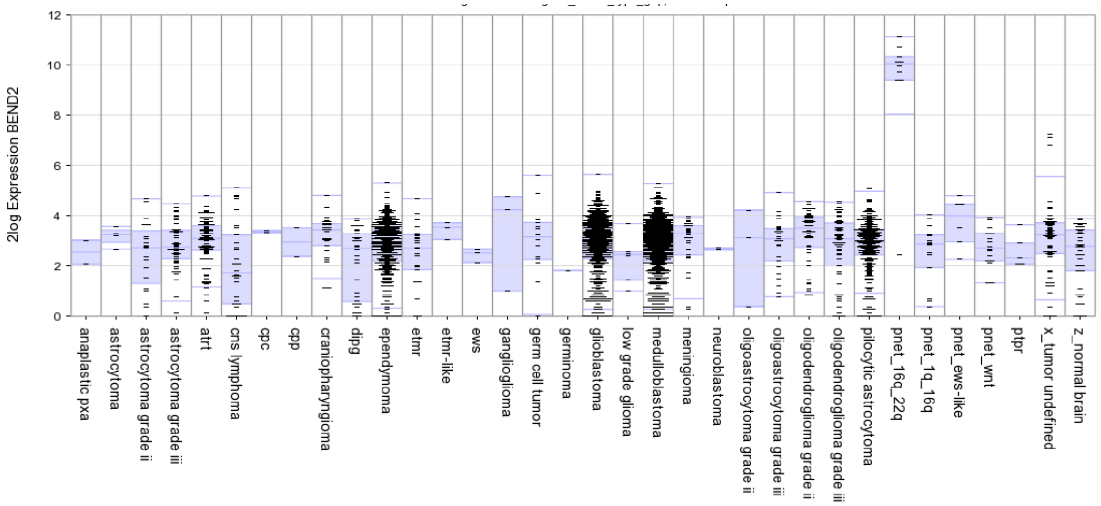

Figure 1.6: *BEND2* expression across brain tumour entities shows a unique and significant upregulation in the candidate pnet-16q-22q subcohort (dataset ps-mkheidel-mkdkfz2273-u133p2 in R2 [337])

Having established *MN1* as the chief partner of recurrent gene fusion events with the significantly preferred secondary partner *BEND2*, we named this new entity Central nervous system

high-grade neuroepithelial tumour with *MN1* alteration (CNS HGNET-*MN1*) (Figure 1.7).

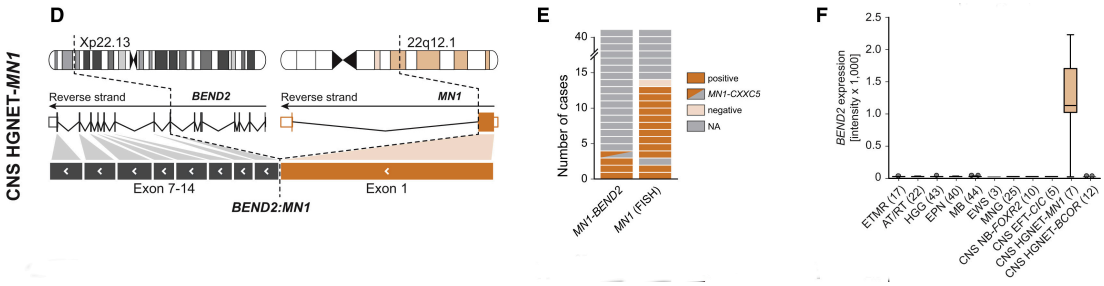

Figure 1.7: Description of the CNS HGNET-*MN1* entity. Extracted from Figure 6 of [218]. D: *MN1-BEND2* fusions shown on an example of typical breakpoint locations. E: RNA-sequencing and FISH analyses confirm *MN1* fusions and *MN1* in the entity. We validated our findings in the initial larger cohort from which the methylome analysis was run. In the 13 specimens with sufficient material, we ran FISH assays targeting *MN1* with break-apart probes. 3 of those cases had matching RNA-Seq data for establishing concordance between RNA-sequencing and FISH. We confirmed all *MN1* fusions detected by RNA-Sequencing and observed that all but one of the cases with sufficient material were positive for *MN1* breaks F: *BEND2* activation is a unique and recurrent characteristic of the entity.

CNS HGNET-*MN1* showed a high rate of new case recovery from a large group of tumours initially not diagnosed as CNS-PNETs (Figure 1.3). We observed that the brain tumour entity astroblastoma [372] [373] accounted for $16/41$ of our CNS HGNET-*MN1* cohort and that these 16 cases were $16/23$ of our entire astroblastoma collection. We thus postulated that the entity previously known as Astroblastoma was not a *bona fide* brain tumour entity but rather a collection of CNS HGNET-*MN1* and a heterogeneous set of tumours from other entities. Our assessment was approved by the WHO classification of Tumours of the Central Nervous System, 2016 [374] where astroblastoma was removed as an entity. Under histopathological investigation, CNS HGNET-*MN1* revealed itself as a heterogeneous entity where only some cases had an Astroblastoma-like histopathology.

We then investigated if the gene *MN1* is a candidate oncogene for tumourigenesis in this new proposed entity with either of the two detected fusion partners *BEND2* and *CXXC5*. *MN1* is disrupted in balanced translocations in meningioma [375] and is part of the *MN1-ETV6* chimeric fusion oncogene in myeloproliferative disorders such as myeloid leukemia [376]. The *MN1-ETV6* oncogene was described as an oncogenic transcription factor [377] with a dominant negative effect on the wild-type allele of *MN1* [378]. The *MN1-BEND2* fusion in the CNS HGNET-*MN1* entity fuses the transactivating domains of *MN1* with the BEN domains of *BEND2*, previously suggested to mediate protein-DNA and protein-protein interactions during chromatin organization and transcription [379]. As *BEND2*, *ETV6* and *CXXC5* are all DNA binding proteins, we hypothesized that *MN1-BEND2* and other *MN1* fusions such as *MN1-CXXC5* have similar oncogenic mechanisms to *MN1-ETV6*. In the absence of a viable cell line or other models, we were unable to further test this hypothesis in this study.

1.3.3 CIC Fusions, mainly CIC-NUTM1, Drive a Subgroup of CNS-PNETs

I ran a gene fusion analysis on the PNET EWS-like subgroup as described in Section 1.2.11, which revealed *CIC* chimeric fusions in $2/3$ of samples with available RNA-Seq data. In both the cases the partner was the *NUTM1* gene, whereas the negative case had a frameshift deletion on *CIC*. In both detected fusion events, exon 16 of *CIC* was fused in-frame to exon 4 of *NUTM1*, retaining the DNA-binding high mobility group (HMG) box domain of *CIC* (Figure 1.8). In the absence of other recurrent genomic alterations and considering the 100% involvement of the *CIC* gene, we designated *CIC* as a primary candidate gene of interest for this subgroup.

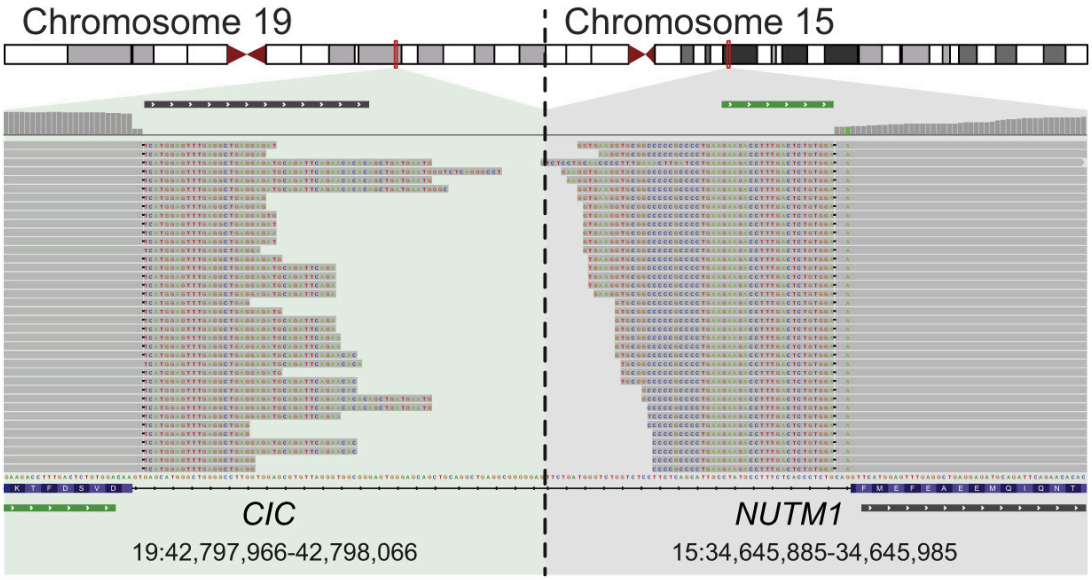

Figure 1.8: A representative *CIC-NUTM1* fusion as displayed on IGV. Extracted from Supplementary Figure 6 of [218]

I first investigated if *CIC* fusions lead to overexpression or suppression of *CIC* in this entity compared to other brain tumour entities (Figure 1.9).

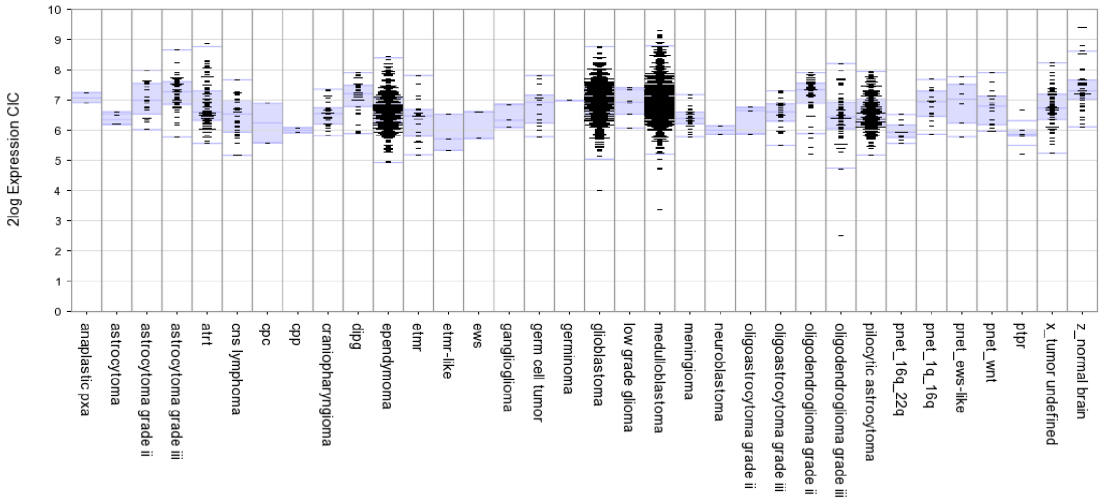

Figure 1.9: *CIC* expression across brain tumour entities: *CIC* in the PNET EWS-like group does not have a strikingly different expression profile (dataset ps-mkheidel-mkdkfz2273-u133p2 in R2 [337])

Next, I investigated if the frequent translocation partner of *CIC*, *NUTM1* has a significantly different/altered expression profile (Figures 1.10 and 1.11C).

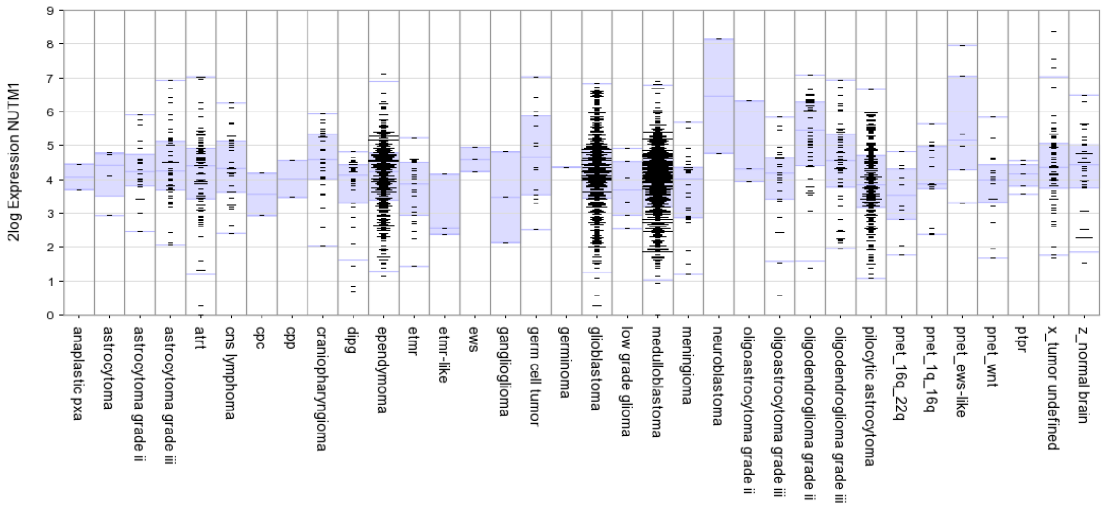

Figure 1.10: *NUTM1* expression across brain tumour entities: *NUTM1* is recurrently activated in the PNET EWS-like subgroup, leading to median and upper quartile values significantly higher than most other CNS tumour entities (dataset ps-mkheidel-mkdkfz2273-u133p2 in R2 [337])

Having established *CIC* as the recurrently implicated gene in this entity, we named it CNS Ewing Sarcoma Family Tumor with *CIC* Alteration (CNS EFT-*CIC*) (Figure 1.11).

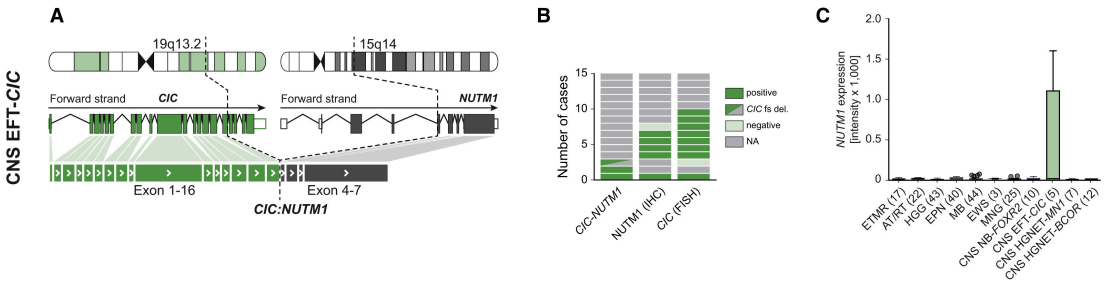

Figure 1.11: Description of the CNS EFT-*CIC* entity. Extracted from Figure 6 of [218]. A: exon 16 of *CIC* is fused in-frame to exon 4 of *NUTM1*, retaining the DNA-binding high mobility group (HMG) box domain of *CIC*. B: We validated our findings in the initial larger cohort from which the methylome analysis was run. In the 9 specimens with sufficient material, we ran Fluorescence in Situ Hybridization assays targeting *CIC* with break-apart probes. 2 of those cases had matching RNA-Seq data for establishing concordance between RNA-sequencing and FISH, including the case with no *CIC-NUTM1* fusion with the *CIC* frameshift deletion. All tested cases apart from the mentioned case were positive for *CIC* breaks. Furthermore, we also tested stained the specimens with sufficient material with the *NUTM1* antibody in an immunohistochemistry (IHC) procedure. All but one tested case tested positive for the *NUTM1* gene, hinting at the recurrent nature of *CIC-NUTM1* fusions even in cases where assays for fusion detection cannot be run. C: *NUTM1* is recurrently activated in the PNET EWS-like subgroup.

Under histopathological investigation, CNS EFT-*CIC* was characterized by a small-cell phenotype but with variable histology. The tumour architecture included both alveolar and fascicular patterns of growth. Although tumors were uniformly high grade, the CNS EFT-*CIC* entity lacked defining histological features and failed to express markers of differentiation, reinforcing the challenges it poses to classical histopathology. However, this entity, along with CNS NB-*FOXR2* had the most consistent clinical PNET diagnoses, and had relatively few gains of new cases from the extension analysis described in Section 1.2.4 (Figure 1.3).

As *CIC-DUX* fusions were previously described in a subgroup of pediatric primitive round cell sarcomas [380] and shown to have a distinct transcriptional signature [381], we analyzed CNS EFT-*CIC* tumors for similar gene expression patterns, confirming transcriptional changes (Figure 1.12).

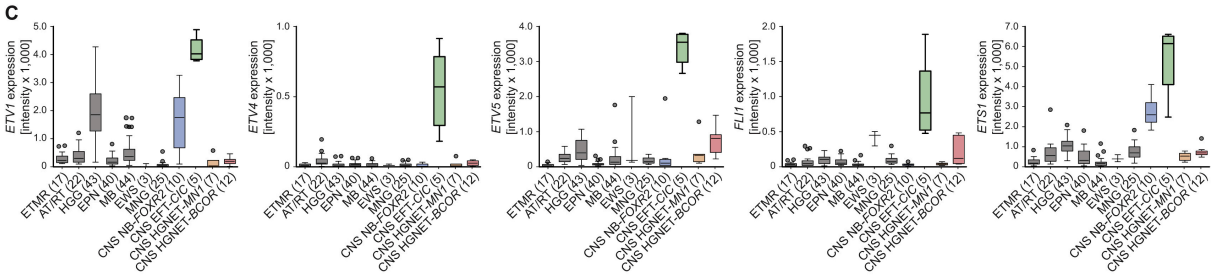

Figure 1.12: Selected targets of the *CIC-NUTM1* oncogene in CNS EFT-*CIC* (Extracted from Supplementary Figure 6 of [218]). Members of the *ETS* transcription factor family, including *ETV1*, *ETV4*, *ETV5*, *FLI1*, and *ETS1* were specifically upregulated in this group as also observed in peripheral EWS.

We later investigated if the gene *CIC* is a candidate oncogene for tumourigenesis in this new proposed entity as part of the *CIC-NUTM1* chimeric fusion gene. Oncogenic re-arrangements of *NUTM1* are known to be the main driver of NUT midline carcinomas, most frequently involving *BRD4* [382]. Considering the preferred location of *CIC-NUTM1* fusions on an exonic level, we hypothesized a molecular mode of action of *CIC-NUTM1* fusions in which specific *CIC* target genes are transcriptionally activated by the *NUTM1* moiety via the recruitment of histone acetyl transferases, similar to a model of how *BRD4-NUTM1* might block differentiation in NMC [382]. In the absence of a viable cell line or other models, we were unable to further test this hypothesis in this study.

1.3.4 BCOR Internal In-Frame Tandem Duplications Drive a Subgroup of CNS-PNETs

In the investigation of the PNET-WNT subgroup, neither small mutation and SV calling from 3 WGS specimens, nor a gene fusion analysis from 8 cases led to a recurrent candidate gene for further investigation. I thus used the 1-vs-3 comparison approach described in Section 1.2.7 and looked for overexpressed oncogene candidates with potentially unknown or unrecoverable activation mechanisms. This approach yielded the candidate gene *BCOR*, which is recurrently and significantly upregulated in the PNET-WNT group compared to all brain tumour entities with sufficient sample sizes (Figure 1.13).

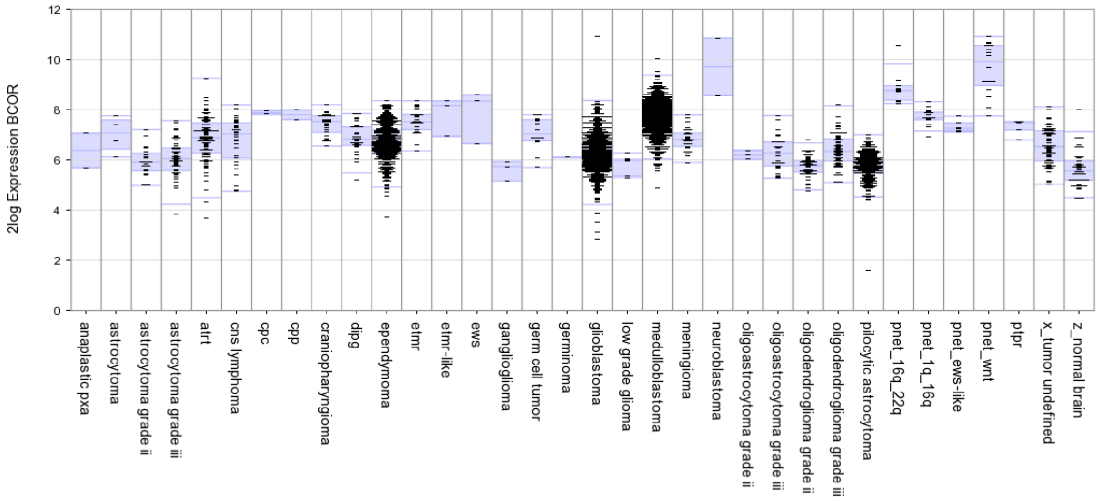

Figure 1.13: *BCOR* expression across brain tumour entities: *BCOR* is recurrently and significantly upregulated in the pnet-wnt group (dataset ps-mkheidel-mkdkfz2273-u133p2 in R2 [337])

I ran a manual inspection of *BCOR* both on RNA-Seq and WGS data in IGV as described in Section 1.2.12, which revealed recurrent and very similar in-frame internal tandem duplications (ITDs) on the exon 15 of *BCOR* $(10/13)$. Two cases had frameshift mutations detected by WES in an extension cohort (data not discussed) $(2/13)$. One exceptional case where we no *BCOR* overexpression was observed had an in-frame deletion between the exons 14 and 15 of *BCOR*, directly targeting the frequently duplicated domain on exon 15 $(1/3)$.

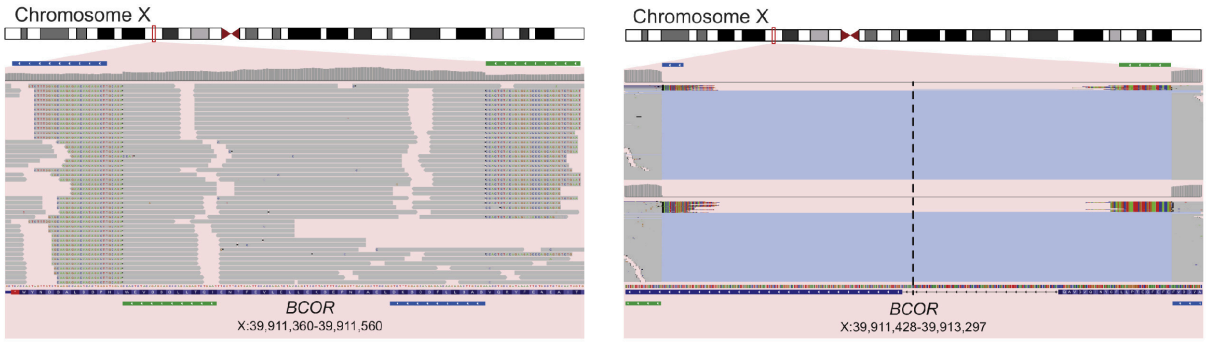

Figure 1.14: A representative *BCOR* ITD and the exceptional *BCOR*-deletion case as displayed on IGV on RNA-Seq data (extracted from Supplementary Figure 6 of [218]). Left: the prototypical ITD events on *BCOR* exon 15 as denoted by the coverage increase and split read mapping indicating the tandem duplication, with the in-frame nature of the duplication confirmed in other analyses. Right: one case had an in-frame deletion of *BCOR* and no concomitant overexpression, with the deletion shown by the coverage drop and split reads mapping to the sequences flanking the deletion breakpoints.

I then investigated if the recurrently duplicated region on *BCOR* has a conserved sequence

on the protein level. I reconstructed the translated protein sequence of BCOR following the duplications for each case where an ITD was detected (Figure 1.15). This analysis confirmed that the ITDs are indeed similar, and revealed that there are two main classes of duplicated sequences in our cohort: i) Subsequences of VSASLLFCSKDLEAFNPESKELLDLVEFT-NEIQTLL, and ii) Subsequences of SASLLFCSKDLEAFNPESKELLDLVEFTNEIQTLLGSSVEW. All cases had the minimally duplicated sequence DLVEFTNEIQTLL.

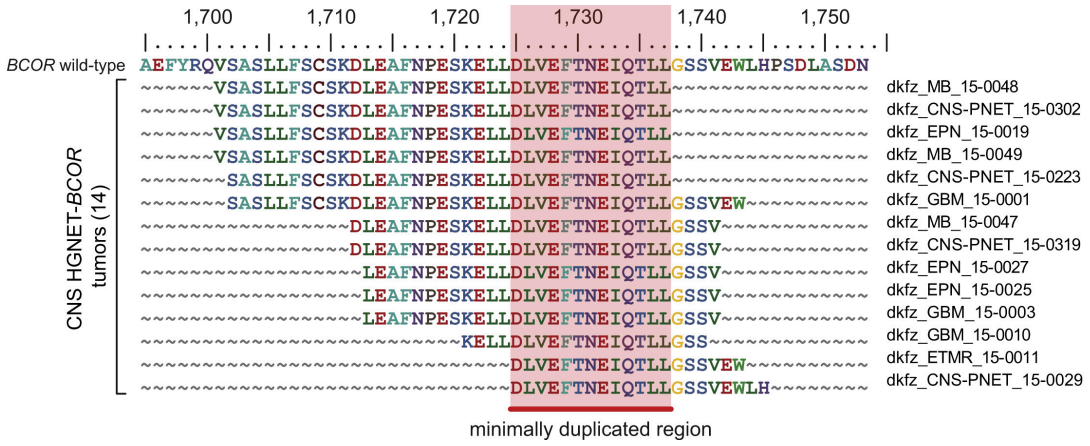

Figure 1.15: Conserved duplicated peptide sequences following ITDs on *BCOR* exon 15. (extracted from Supplementary Figure 6 of [218])

Having established *BCOR* as a target of recurrent target of in-frame ITDs, we named this new entity Central nervous system high-grade neuroepithelial tumour with *BCOR* alteration (CNS HGNET-*BCOR*) (Figure 1.16G).

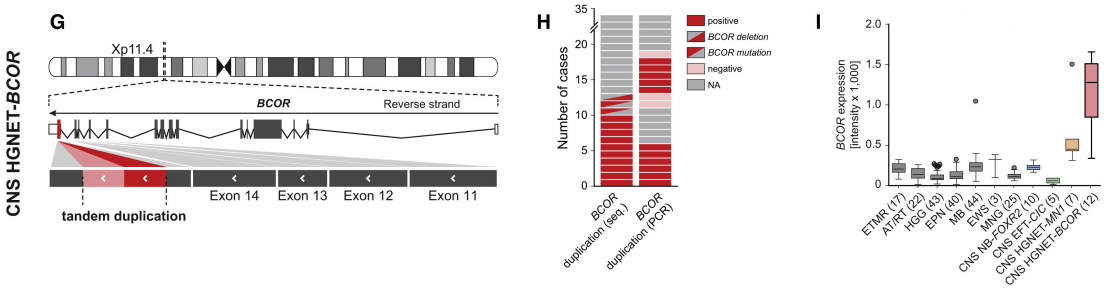

Figure 1.16: Description of the CNS HGNET-*BCOR* entity (extracted from Figure 6 of [218]). G: *BCOR* Exon 15 ITDs as the hallmark event of the CNS HGNET-*BCOR* entity. H: validation of *BCOR* ITDs with a PCR designed using our knowledge of the conserved ITDs. $(11/14)$ of the cases with available material were found to be harbouring ITDs of the described type on *BCOR*. I: *BCOR* is recurrently and significantly upregulated in the CNS HGNET-*BCOR* entity.

A protein motif search using the Motif search tool of GenomeNet [383] yielded hits from NCBI-CDD [384] and Pfam [385], indicating a duplication affecting the PCGF Ub-like fold discriminator of the BCOR protein which binds the RAWUL (RING finger and WD40-associated

ubiquitin-like) domain of the polycomb-group RING finger homologs PCGF1 and PCGF3 [386].

Next, our pathologist colleagues investigated the histopathological patterns of CNS HGNET-*BCOR*. This entity consisted of relatively compact tumours containing a combination of cells with shapes ranging from spindle to oval. Tumours were observed to often exhibit perivascular pseudorosettes, giving them an ependymoma-like appearance. Tumours frequently showed fibrillary processes, typically observed in glial differentiation, and only rarely exhibited a true embryonal pattern. This diversity of histopathological patterns is reflected in the high rate of new case recovery from a large group of tumours initially not diagnosed as CNS-PNETs (Figure 1.3).

Shortly before the submission and eventual acceptance of our study, a similar finding on *BCOR* ITDs was published by a different group in Japan, on a different paediatric disease: Clear Cell Sarcoma of the Kidney (CCSK) [198] (Figure 1.17). Their results were in a remarkably significant agreement with ours in terms of the location and minimum conserved peptide sequence of the ITDs, with an almost total match of the ITD sequences. These results later led to differing viewpoints on entity classification where one study claimed that CNS HGNET-*BCOR* and CCSK-*BCOR* were local variants of the same entity [387] whereas another study focusing on the comparison of CNS HGNET-*BCOR* and CCSK-*BCOR* [388] claimed that the former is of neuroepithelial origin and the latter is of mesenchymal origin and should be considered as distinctly different entities.

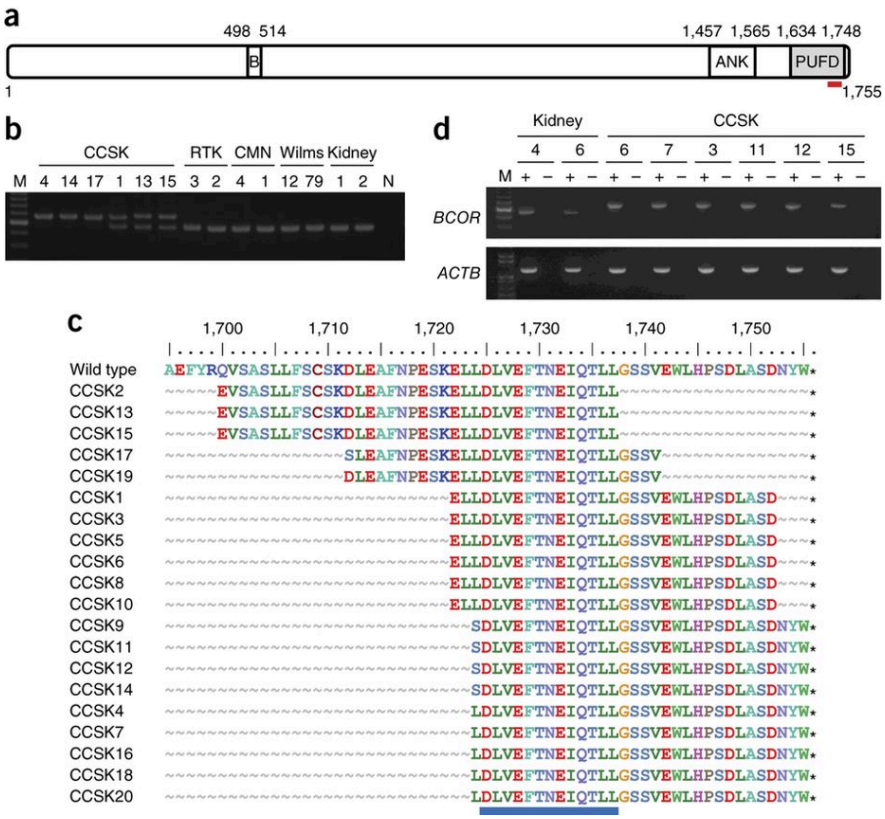

Figure 1.17: Conserved duplicated peptide sequences following ITDs on *BCOR* exon 15 in CCSK (Figure 1 of [198])

The most comprehensive analysis of *BCOR*-ITD in CCSK to date was published in 2015 following the initial report on CCSK [389], again before our study. Recently another entity, high-grade uterine sarcoma was found to have a rare subgroup affecting young adults (average age 24) [390] carrying *BCOR*-ITDs with the exact same conserved sequence of duplications. Neither of these three studies and ours managed to dissect the functional effect of *BCOR*-ITD in either of the three entities, and were mainly limited to a descriptive treatment of the subject using omics technologies and histopathology.

1.3.5 FOXR2 Activation via Diverse Mechanisms Drive a Subgroup of CNS-PNETs

In the investigation of the PNET 1q-16q subgroup, neither small mutation and algorithmic SV detection (with Delly [308] by Jan Korbel's team, EMBL, results not shown) from 5 specimens with available WGS data, nor a gene fusion analysis from 4 cases led to a recurrently altered candidate gene for further investigation. I thus used the 1-vs-3 comparison approach described in Section 1.2.7 and looked for overexpressed oncogene candidates with potentially unknown or unrecoverable activation mechanisms, yielding the candidate gene *FOXR2* (Figure 1.18).

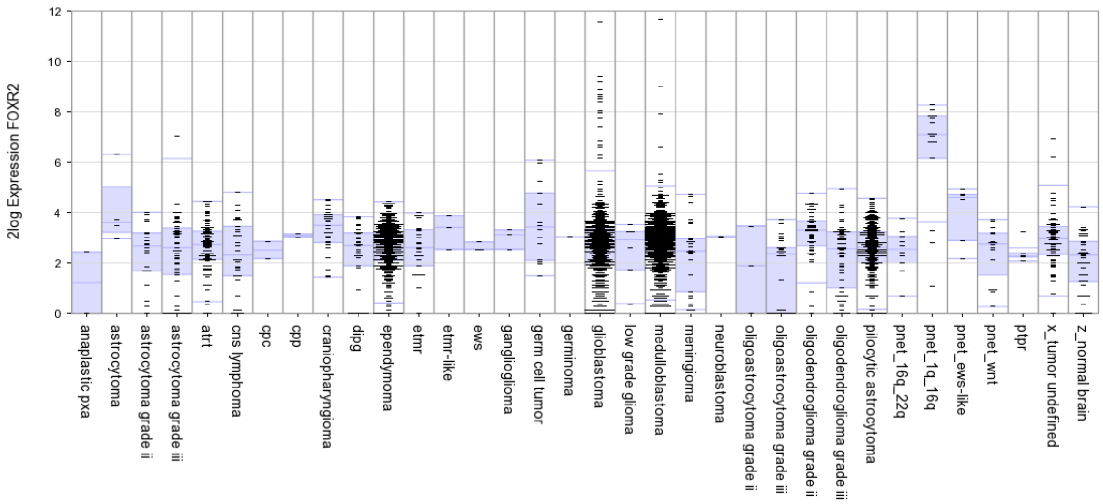

Figure 1.18: *FOXR2* expression across brain tumour entities: *FOXR2* is recurrently and significantly upregulated in the PNET 1q-16q group compared to all brain tumour entities

FOXR2 is in normal tissues only expressed in the testis (Figure 1.19) according to RNA-Sequencing data from the GTEx Consortium [257].

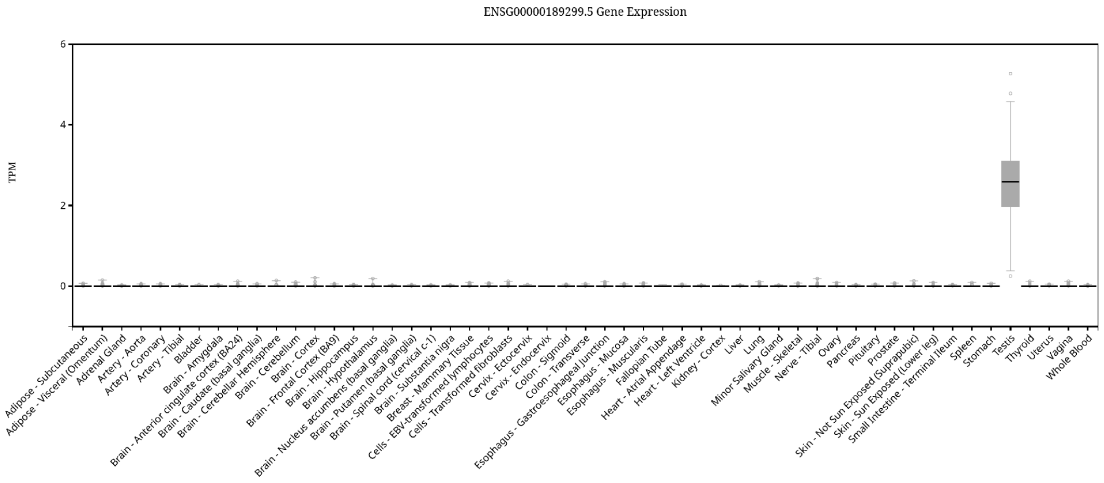

Figure 1.19: *FOXR2* expression across healthy human tissues is confined to the testis (GTEx Consortium)

I ran a manual inspection of *FOXR2* both on RNA-Seq and WGS data in IGV as described in Section 1.2.12 revealing the activation of a novel transcript of the gene (Figure 1.20). This novel transcript was first predicted in [391] as a putative long transcript and is an entry named *FOXR2, alternative variant aAug10* in AceView [392].

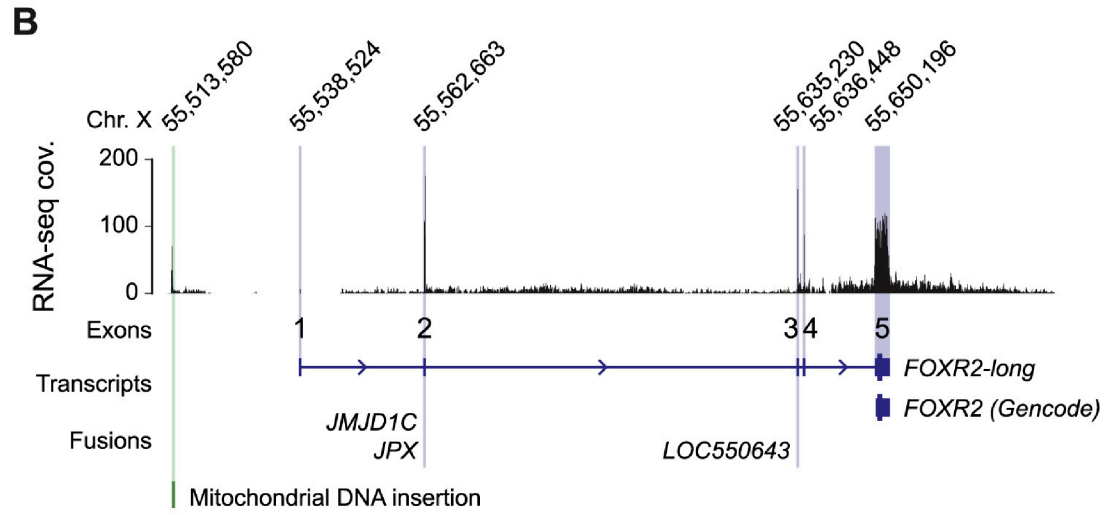

Figure 1.20: The canonical and novel (long) isoforms of *FOXR2*, together with locations of a subset of detected rearrangements (extracted from Supplementary Figure 5 of [218]). The GENCODE annotation of *FOXR2* does not accurately represent the isoform detected to be activated in the CNS NB-*FOXR2* entity, where a diverse set of recurrent structural rearrangements target a previously unknown promoter of a novel long isoform transcript.

Contrary to the other three entities discussed here, mechanisms of *FOXR2* activation showed a great level of diversity. We observed by WGS and RNA-Seq data analysis:

i) 2 intrachromosomal deletions (~35kb) fusing the *MAGEH1* promoter with the *USP51*

gene body upstream of long-FOXR2 (Figure 1.21C-D),

ii) 2 tandem duplications fusing long-*FOXR2* with the *JPX* 5'UTR and LOC550643 5'UTR (Figure 1.21C-D),

iii) 1 interchromosomal translocation t(10,X) fusing the 5'UTR of *JMJD1C* with the promoter of long-*FOXR2* (Figure 1.21C-D),

iv) 1 mitochondrial insertion on the *USP51* gene body upstream of long-*FOXR2* leading to a novel promoter and another novel *FOXR2* transcript (Figure 1.21C-D-E).

Furthermore, analysis of copy number variation patterns from methylation arrays revealed a number of other patterns of structural rearrangements in cases where sequencing data was not available (Figure 1.21D):

i) 4 intrachromosomal deletions (~500kb) connecting the *MAGED2* locus with the *FOXR2* locus,

ii) 2 intrachromosomal deletions (~3 Mb) targeting the *FOXR2* locus,

iii) 1 intrachromosomal deletion (~8 Mb) targeting the *FOXR2* locus,

iv) 1 case of chromothripsis [393] on chromosome X with *FOXR2* on one of the amplified loci.

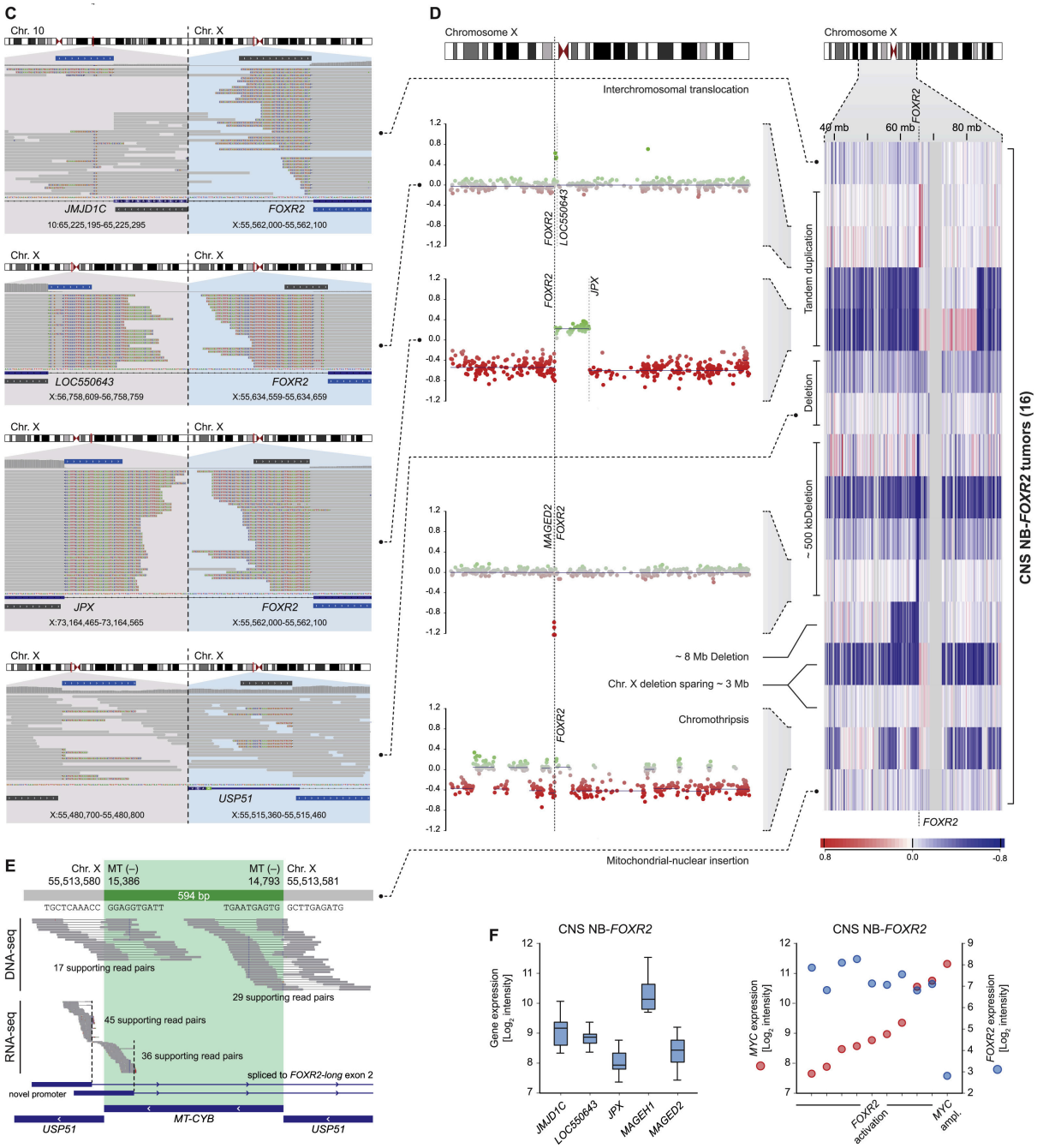

Figure 1.21: Diverse activation mechanisms of *FOXR2* (extracted from Supplementary Figure 5 from [218]).

With all presented mechanisms a common pattern was the activation of a putative oncogene by structural rearrangements connecting actively expressed genomic regions to the inactive oncogene. This mode of activation is known as enhancer hijacking [394] was previously shown to be the main driver of subsets of Group 3 and Group 4 Medulloblastoma [132] [211] by activation of the *GFI1B* and *PRDM6* genes. The *FOXR2* activating partner genes are transcriptionally active (Figure 1.21F).

There were only two exceptions to the recurrent pattern of *FOXR2* activation via SVs in this

novel entity (Figure 1.22B): i) One case had no *FOXR2* activation on the transcriptome level, but carried a focal amplification of *MYC* (Figure 1.21G, and ii) One case had *FOXR2* activation determined by RNA microarray analysis, no SV detected by very detailed manual inspection $\pm$ 4MB starting from *FOXR2*, and no CNVs to explain the mode of *FOXR2* activation. Due to lack of RNA-Sequencing data we also couldn't determine what kind of *FOXR2* isoform was transcribed.

Having established *FOXR2* activating rearrangements are the main driver of this novel entity and due to its neuroblastoma-like histopathological characteristics, we named it central nervous system neuroblastoma with *FOXR2* activation (CNS NB-*FOXR2*) (Figure 1.22A).

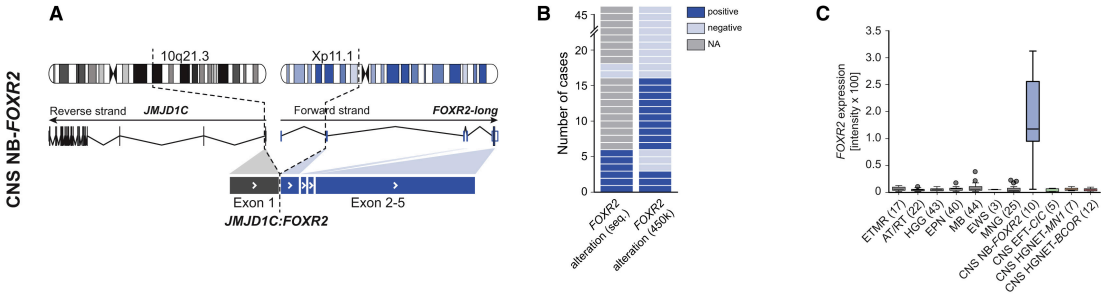

Figure 1.22: Description of the CNS NB-*FOXR2* entity (Figure 5 from [395]). A: A representative *FOXR2*-activating rearrangement showing the long *FOXR2* transcript. B: *FOXR2* alterations detected by WGS (6/8) and methylome arrays (13/46) C: *FOXR2* is recurrently and significantly upregulated in the CNS NB-*FOXR2* entity

Next, our pathologist colleagues investigated the histopathological patterns of CNS NB-*FOXR2*, noting embryonal architectural and cytological features with a small-cell phenotype, frequently with areas of differentiation in the form of neuropil, neurocytic cells, or ganglion cells. Some specimens showed frequent perivascular anuclear zones ("vascular pseudorosettes"), nuclear palisades, and Homer Wright rosettes. Tumours of this entity nearly uniformly expressed OLIG2 and the neuronal antigen synaptophysin. Overall, CNS NB-*FOXR2* showed a histopathological profile most closely matching the classical definition of CNS-PNETs, which is reflected in the lowest rate of new case recovery from a large group of tumours initially not diagnosed as CNS-PNETs (Figure 1.3).

Following our study, it has been published that FOXR2 acts with MYC in a stable complex FOXR2-MYC-MAX complex to promote MYC transcriptional activities in adult cancers [396]. This fits our data that one exceptional case of the CNS NB-*FOXR2* entity had a *MYC* amplification instead of *FOXR2* activation: If the downstream targets and activated pathways are identical for both oncogenes, the same cell of origin that requires the activation of this pathway for oncogenic transformation, can in principle use either mechanism. However, it is not known why *FOXR2* is preferentially activated in the CNS NB-*FOXR2* entity with such a strong bias as suggested by our data whereas *MYC* amplifications are not common in this cell of origin. As with the other entities introduced here, there exist to date no cell lines of CNS NB-*FOXR2* that allow a precise study of the role of *FOXR2* in this cell type.

1.4 Discussion

1.4.1 Interpretation of our findings and their impact on the field of paediatric neurooncology

Our study had a significant impact on the field of paediatric neurooncology and indirectly, cancer research as a whole. We showed that CNS-PNETs are not a monolithic entity and the CNS-PNET diagnosis coming from classical histopathological methods are error-prone. Our findings both emphasized the limitations of classical histopathology and proved molecular pathology via methylome profiling as a viable method of overcoming one of the most challenging diagnostic questions in the paediatric neurooncology. Our finding that CNS-PNETs are a mixture of other paediatric brain tumour entities, including those with dismal prognosis such as ETMRs and HGGs led to a reassessment of the similarly dismal expected prognosis of CNS-PNETs: a clinical trial on CNS-PNETs was closed for recruitment following our study, the participating patients were molecularly profiled as presented in our study and it was observed that CNS-PNETs do not universally have a dismal prognosis when misdiagnosed cases are withdrawn from analysis [397]. Even though we currently cannot propose molecularly designed therapies for the four entities discussed in our study, this observation alone has an appreciable clinical impact: through accurate and molecularly defined diagnosis, patients can avoid intensive treatments with severe side effects. Indeed, this statement can be extended to all brain tumour entities and became one of the critical milestones towards the landmark methylome profiling-based molecular classification of brain tumours [353].

We presented four novel entities of paediatric brain tumours with distinct oncogenes with a diverse set of activation mechanisms. Thanks to the high quality of our data and the well-executed classification of the cases, we managed to describe all of these four new entities. Our findings emphasized the power of using whole genome sequencing and RNA-sequencing for discovery of driver oncogenes and their mechanisms: Without WGS, the CNS NB-*FOXR2* entity would have remained unexplained, without RNA-Seq both CNS HGNET-*MN1* and CNS EFT-*CIC* entities would have remained unexplained. Our approach was successful in great part due to the multi-omics & array approach adopted here.

Following the publication of our study in *Cell*, the WHO classification of Tumours of the Central Nervous System, 2016 [374] was updated to remove CNS-PNETs (replaced by the four new entities and CNS HGNET-NOS), astroblastoma (replaced by CNS HGNET-*MN1*), CNS neuroblastoma (replaced by CNS NB-*FOXR2*), CNS ganglioneuroblastoma (replaced by CNS NB-*FOXR2*).

Our study also led to a number of unpublished follow-up projects: Currently, there is a large clinical follow-up study investigating the survival characteristics of the four new entities, as we did not have sufficient sample sizes and availability of survival data for this purpose in the study presented here. Also, in order to be able to do functional genomics analysis on the four new entities, mouse models are being developed with considerable success: 3/4 of the new entities (except for CNS HGNET-*BCOR*) now have mouse models of the tumours where viable growth is observed. These models will be used to test treatments *in vivo*. Worldwide, it is already leading to impact in personalized medicine with new treatment protocols being tested in case studies [398].

1.4.2 Impact of this study on CNS-PNETs on my PhD research

In this study, we successfully integrated methylome, genome and transcriptome data from a disease type with undefined histopathological characteristics. We applied integrative omics data analysis methods to classify novel disease subgroups, found aberrantly overexpressed genes and their underlying genomic mechanisms of dysregulation, identifying recurrent somatic structural variants as the drivers of the novel entities.

Our successful application of integrative omics data analysis strongly influenced the rest of my PhD: I did not continue towards a specialization on omics data analysis of paediatric brain tumours but rather on algorithm development for systematic detection of structural variants from whole genome sequencing data. I was most interested and impressed by the diversity of the structural rearrangements observed in our study: gene fusions, in-frame ITDs, in-frame deletions, enhancer hijacking via duplications & deletions of diverse sizes as well as interchromosomal translocations, and an entirely novel case of an oncogene activating mitochondrial promoter gene insertion. In a time of transition where increasingly more studies with large cohorts had access to the WGS assay, structural rearrangements had great potential to identify novel oncogenic drivers. There were also practical and organizational concerns: at the time of this study, the DKFZ's centralized omics data analysis platform and practices did not have the tools to call structural variants with high sensitivity, specificity, independent of external collaborators. This motivated me to pursue algorithm development for detection of structural variants as my next PhD research subject.

CHAPTER 2
SOPHIA: STRUCTURAL REARRANGEMENT DETECTION BASED ON SUPPLEMENTARY ALIGNMENTS AND A POPULATION BACKGROUND MODEL

2.1 Introduction

Structural variants can have significant oncogenic effects in almost all cancer types with a variety of mechanisms and are subject of great and broad interest in cancer research as discussed in the general introduction of this dissertation and shown on the results on my first doctoral research project (Chapter 1).

Simple structural variants can broadly be classified into i) Deletions, ii) Duplications, iii) Inversions, iv) Translocations and arise from a variety of mechanisms and as part of a variety of higher order complex patterns including [399] [193] i) Double-strand break repair defects in either homologous recombination or non-homologous end joining, ii) Microhomology mediated break-induced replication, iii) Breakage-fusion-break cycles, iv) L1 retrotransposition, v) Double minute chromosomes and neochromosomes, vi) Regional amplifications via HPV insertion.

Technologically, structural variants were first detected in low resolution using "chromosomal banding" with Giesma Staining [400] which yielded the final karyotype of cells, whether rearranged or not. Later, the Fluorescence in situ Hybridization (FISH) technology enabled the detection of the simple building blocks of structural variants by showing the proximity/pairing of targeted sites in high resolution [401]. Indeed, it was the FISH assay that allowed the discovery of most of the oncogenic structural variant examples of highest significance listed above starting from the presentation of the *BCR-ABL1* oncogene and went on to become a standard clinical assay [402].

With the development of genome sequencing techniques, it became possible to detect structural variants across the whole genome in an unbiased manner. With genome sequencing, it also became possible to systematically study the quantitative burden of structural variation [403] as well as the higher order structures or signatures of structural variants in cancer genomes [193] [404] [284].

Detection of structural variants by paired-end genome sequencing relies on discordancy of the mate reads and split reads [366] [308]. Briefly, amplified DNA is fragmented into oligonucleotides of pre-defined lengths (insert size), which are then sequenced from both ends. During the sequence alignment process, these individual read pairs are mapped to the (human) genome. If these both ends map to different chromosomes, are unexpectedly distant from or close to each other, or finally they have the same strand orientation as opposed to different strands (as dictated by the sequencing approach from both ends), this indicates a discordant mapping. As discordant mate based structural variant detection relies on differences between mate reads, structural variants smaller than a size range are unable to be detected with this approach. This size does not have a clear cutoff and is dependent on sequencing technology and library preparation [405] [406]. Also, a given read in a read-pair can span a breakpoint in which case split

reads are generated that map to the original site and the aberrant partner site generated by the structural variant. Such split reads are also capable of detecting small structural variants such as deletions, duplications or small inversions that are outside of the range of discordant mate based structural variant detection.

As the current state of genome sequencing technology, "next generation sequencing" relies on short reads (151 base pairs as of late 2018, typical previous values included 50 and 100), and because of the prevalence of repetitive sequences in the human genome, [407], [408], [409], [410], [411], [412], [413], [414], reads frequently map non-specifically during the sequence alignment process [415]. This lack of mapping specificity directly clashes with the described principles of structural variant detection and causes a high prevalence of false positives: a short repetitive read can align to any location in the reference genome where the repeat is included, whereas its mate can map properly to a nonrepetitive site. In such cases, the aligner cannot decide if the fragment arises from a normal sequence flanking a repeat or if the normal sequence should map a distant repeat in a discordant configuration, and outputs all possible solutions with ambiguous mapping scores (MAPQ 0) [359]. Long repeats can generate entire fragments consisting of repetitive sequences. These yield two repetitive mate reads which, in turn, also have the described discordancy issue. There are also large homologous loci that are repeated in the human genome such as the pseudoautosomal regions [416]. These regions can also yield alignments with low quality mapping scores and discordant read pairs even in the absence of structural variants.

While these described issues can be mitigated using long-read sequencing, the high cost and low base-level accuracy of the current long-read sequencing technologies do not allow them to be a desirable tool in cancer genomics: i) the high cost would prevent sequencing with high coverage and consequently a proper investigation of the clonal heterogeneity of a tumour specimen, ii) the low base-level accuracy can lead to ambiguities between subclonal single nucleotide variants and base calling errors. Thus, it is a technological and practical necessity to develop structural variant detection approaches and algorithms which can work with the limitations of short-read based DNA sequencing. Due to the strategic importance of the goal of structural variant detection and the difficulties presented by the employed technology, a number of different algorithms have been developed which employ different structural variant detection and filtering strategies. A non-exhaustive list of some prominent tools is as follows:

i) BreakDancer (2009) [417] collects discordant read pairs generated by the aligner and uses a probabilistic model that compares the discordant read load of a given candidate region with the expected background discordant read generation parameters. A pool of samples can be used for the generation of this probabilistic model, which would account for the artefact-rich repeats in the human genome. BreakDancer does not use the concept of split reads which leads to sensitivity issues [366].

ii) CREST (2011) [366] uses soft clipped reads and assembles reads supporting candidate breakpoints using CAP3 [418] and aligns to the reference genome using BLAT [419] to annotate the structural variant. Additional filtering steps are not applied to account for the effects of genomic repeats, which leads to a loss of specificity. CREST does not use the concept of discordant read pairs which precludes SVs without aligned split reads

from being detected.

iii) Delly (2012) [308] combines the concepts of split read and discordant pairs based structural variant detection and filters variants based on evidence level using a k-mer based approach. Delly starts from candidate breakpoints proposed by discordant read pairs and adds split read evidence to strengthen structural variant calls and to make them specific to the base-pair level. The Delly manuscript does not explicitly discuss further filtering based on genomic regions prone to the generation of artefacts, which was likely to have contributed to the specificity issues discussed in [367]. To account for this shortcoming, we were informed in personal communications that Delly is usually used with in-house filters that are not part of the official software package. Delly has been a major contributor to a number of successful studies on the cancer genome including [132] [211] [100] [420].

iv) Lumpy (2014) [367] uses a generalized probabilistic model for combining all types evidence pertaining to the existence of a structural variant including split reads, discordant read pairs, copy-number profiles and known structural variants. It was published with a simulation based benchmark comparing it to other established structural variant detection tools of the time and claimed superiority from a theoretical standpoint. In practice, Lumpy has since its publication in 2014, only been cited in two publications of cancer cohort studies [421] [422] as of December 2018, with most of its citations coming from non-human studies, human non-cancer studies, or other theoretical publications. While this does not indicate an indisputable weakness *per se*, it is currently not widely adopted in the cancer genomics community.

v) Manta (2016) uses a graph algorithm based on ”Breakend Graphs”. It combines split reads and discordant mate pairs. It uses custom-designed additional filters such as eliminating very-high coverage regions in the control sample in somatic analysis, high ratio of MAPQ0 (unspecific mapping in exactly duplicated genomic regions) reads in the call, large structural variants only with split read support and no discordant mate support, as well as other internal scores developed with respect to the core graph algorithm of Manta. Manta has a strong focus on cancer genome analysis and has been used in a number of cancer genomics publications, including large-scale ones [423] [424].

vi) novoBreak (2017) [425] uses ”local assembly”, it generates k-mers of reads that have common short sections that do not properly map to the reference genome including discordant mate pairs and split reads, creating assemblies from each k-mer set. The scoring for each k-mer’s local assembly uses a statistical likelihood model based on the beta-binomial distribution, where low quality read-ends are trimmed. There is no separate treatment of repetitive or otherwise artefact-rich regions.

vii) SvABA (2018) [426] similarly uses ”local assembly”, also including gapped reads covering very small insertions and deletions less than 50 base-pairs. It also has additional features for identifying short templated sequence insertions in the final form of the modelled structural variant following the local assembly procedure, which was presented in a

large panel of cancer cohorts as a frequent biological process. Interestingly while novoBreak claims excellent sensitivity in low-coverage regions, SvABA admits the opposite despite using very similar design principles, citing Lumpy and Delly as more sensitive tools in the larger structural variant range especially in lower-coverage use cases. Citing this, the publication recommends SvABA as a structural variant and indel detection algorithm covering a broad range of events in genomics workflows possibly as a key component in a multi-tool consensus approach.

viii) BRASS (*no publication as of July 2019*, `https://github.com/cancerit/BRASS`) is an unpublished tool with no open documentation available for its design principles except for the fact that it uses "local assembly". It is included in this list due to its participation in the PCAWG Consortium as one of its structural variant detection algorithms.

While the field seems to be converging towards "local assembly" based approaches, the primary benefit is likely to come from combining multiple types of breakpoint evidence as opposed to the earlier tools considering only split reads or mate discordancy. Supporting this point, the recently pre-released results of the PCAWG Consortium's SV working group [284] shows that Delly, performs similarly to SvABA, which does use local assembly.

What is not discussed in the majority of the publications are difficulties regarding the analysis of cancer datasets. Detection of structural variants requires sensitivity for subclonality arising due to the tissue impurity or inherent clonal heterogeneity, specificity required for dealing with genome artefacts as well as the ability to distinguish somatic and germline structural variants. Furthermore, significant hallmark Studies, probably for concerns regarding data availability and controllability, methods publications focused on freely available genomes [427], or simulations generating diverse types and size of structural variants [428] [429] [430]. Some of the discussed tools did discuss applications in cancer genomes in a limited number of cases (novoBreak: 1 case, CREST: 5 cases, Manta: 1 case, SvABA: comprehensive analysis across multiple cohorts).

None of the articles reviewed here discuss the aspects of lower quality samples, structural variants of particular detection difficulties, or speed & memory considerations for particularly challenging inputs. Our institutional experience at the DKFZ (mainly with CREST and Delly) taught us that all of these aspects are significant practical considerations in a large-scale sequencing centre. As of 2015, we had accumulated a massive number of tumour and control whole genome sequencing runs and accordingly, diverse experiences on the quality control of whole genome sequencing datasets [431]. This includes experiences on the detection of structural variants across different cancer cohorts with various algorithms. Frequently, we observed that runs would entirely fail in some challenging samples or take up to weeks of processing time. Similarly, it was a common occurrence that CREST or Delly output would contain massive amounts of false positive calls. These issues were exacerbated in samples with lower quality sequencing data, which is dependent on both input material and the quality of sequencing itself. Also, Delly was unable to detect mid-sized indels (50-~1000 bps) until recently, which covers an important class of structural variants such as *BCOR* or *FLT3* ITDs.

Considering the state of the art at the time, and the necessity to improve detection for a biologically very important class of structural variation mid-sized indels, we wanted to develop

an structural variant detecting approach that took full advantage of our rich repository of whole genome sequencing datasets. Our aims were to achieve

i) ability to capture mid-sized indels

ii) excellent sensitivity capturing disease hallmark structural variants with a negligible frequency of misses and while retaining specificity especially on challenging genomic regions

iii) excellent data processing speed and memory efficiency

iv) excellent robustness with respect to input sample quality, without unreasonable loss of sensitivity or specificity in cases with low tumour purity or low proper pair ratios

.

With the SOPHIA algorithm presented here, we have reached all of these goals using a fast and efficient algorithm developed in C++17 without the complexity introduced by the modern local assembly approaches.

2.2 Methods

2.2.1 Study Design

SOPHIA uses matched or single alignments of whole genome sequencing data for detection of structural variants. One particular feature of SOPHIA is to not need a realignment or assembly step which has great benefits for speed and memory usage as discussed in Section 2.3.4. Instead of building an assembly for each breakpoint by collecting candidate reads from all over the genome, SOPHIA reads alignments in a linear stream in a single pass, storing only the currently read region in memory.

This fast single-pass low memory approach is possible thanks to the already calculated ”supplementary alignments” provided by the aligner BWA-MEM [359]. Supplementary alignments propose for split reads one or multiple alternative sites of mapping in the genome. Without a consensus building approach via modern local assembly approaches, or without probabilistic model as in Lumpy, such estimates based only on around less than the half of a short read length are highly error prone due to the inherent issues of genomic repeats and sequencing quality. Nevertheless, they are a valuable source of information because they contain all the (split-read mappable) candidate structural variants albeit with a massive load of false positives. Discordant mate information is similarly error-prone due to sequencing quality and genomic repeats.

In order to benefit from the integration of pre-calculated supplementary alignments and discordant mate information while accounting for this inherent and expected high rate of errors, SOPHIA increases specificity by an integration of i) clinical standard highly sensitive and specific but targeted FISH data, ii) expert knowledge of biologists in interpreting FISH output, iii) a background database of control (healthy tissue, most often blood, from donors in cancer studies), iv) and expert knowledge of bioinformaticians in training a decision tree based on these criteria (Figure 2.1).

This integrative approach on both during the candidate proposal stage and the filtering stage combines to present a fast, sensitive and specific algorithm for the detection of structural variants.

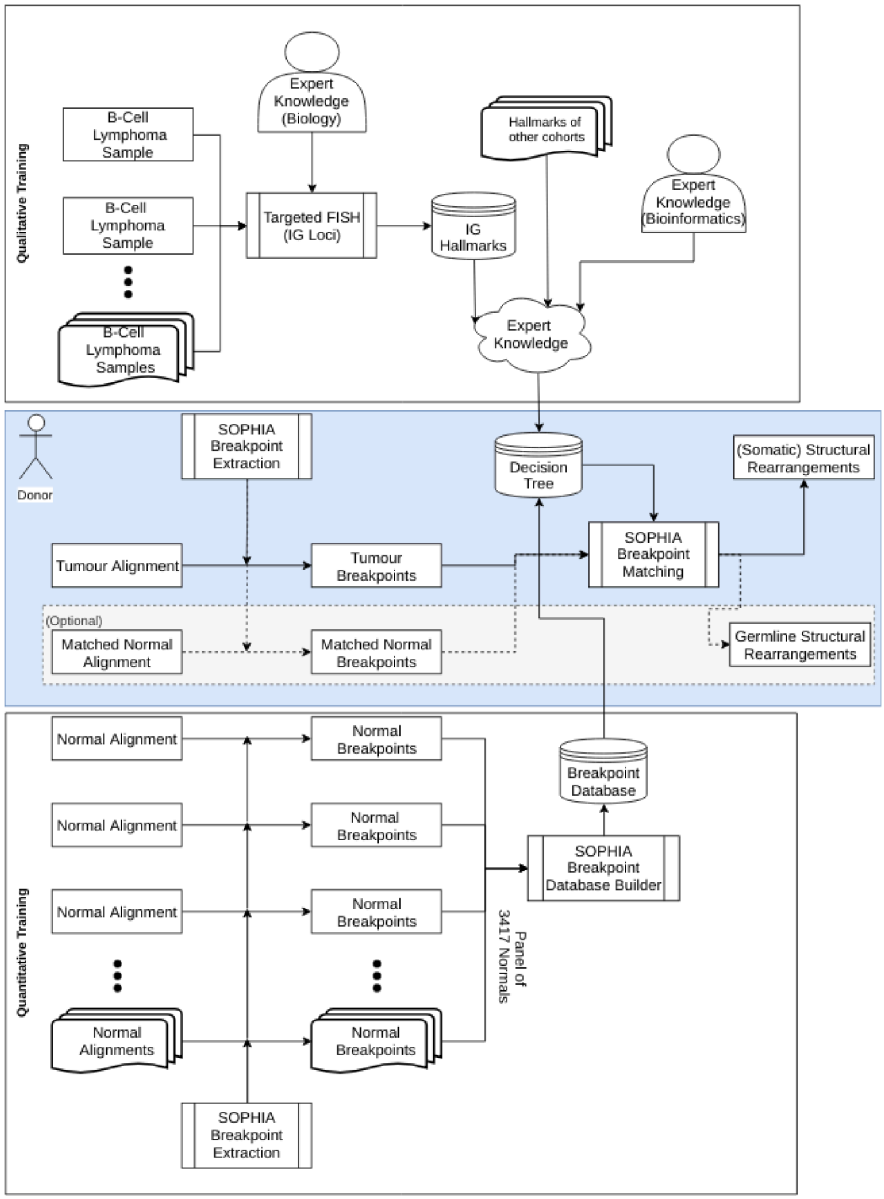

Figure 2.1: The general SOPHIA workflow integrating multiple sources of knowledge for detection of structural variants.

SOPHIA is implemented as a C++17 program using no external libraries apart from strtk (strtk.hpp, Arash Partow `http://www.partow.net/programming/strtk/index.html`) for string tokenization and Boost (v1.6.9) for command line input argument parsing. All remaining code including the SAM parser is written using the C++ standard library.

2.2.2 Classification of Aligned Reads into Quality Categories

The SOPHIA workflow starts by a linear stream of uncompressed alignment output i.e. BAM files converted to the SAM format, where each line corresponds to an aligned read, generated

by the view command of the Samtools toolkit (version 1.2 or above). In a decision tree, each read is either discarded or classified into one of eight read categories. This procedure is based on a decision tree (Figure 2.2).

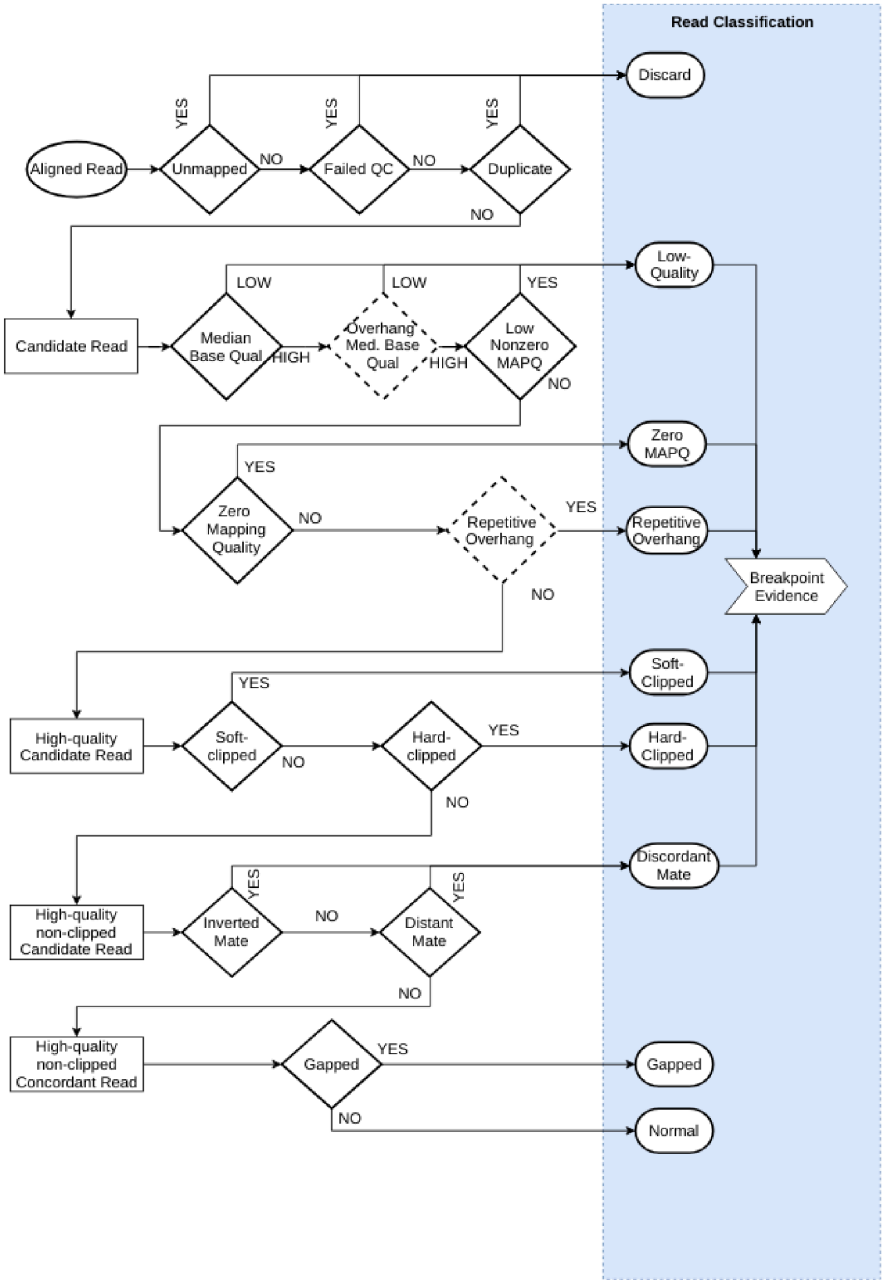

Figure 2.2: Read classification in SOPHIA

Criteria for the decisions are as follows:

i) The discarded reads are decided according their SAM flags and pre-defined cutoffs (Figure 2.2).

ii) A full-length read is required to have the median base quality 23.

iii) A read is considered to be ”clipped” if it has a split-read overhang of at least 10 bases.

iv) A soft-clipped read's overhang is required to have the median base quality 23 with no more than 4 consecutive bases below base quality 12 at the clip's end otherwise the read is considered to be low quality.

v) A mapping quality below 13 is considered low, where reads with zero mapping quality and otherwise high read quality (according to the criteria (i-iv) are classified in a separate category which allow them to contribute to breakpoint detection without the low-quality read penalties.

vi) A read is considered discordant if it is on the same strand as its mate as it is inverted.

vii) A read is considered discordant if it is more than 5 standard deviations away from the median insert size of the merged alignment (bimodal or otherwise multimodal insert size distributions are not given special consideration) where this value is capped at 4000 base pairs.

viii) A read is considered gapped if it contains an I or D in its CIGAR string indicating an insertion or deletion of one or multiple bases.

All read categories except for discarded, gapped and normal are used in the definition of breakpoints (Section 2.2.3). Breakpoints are subsequently paired to form structural variants (Section 2.2.5), where the presence of gapped and normal reads are used in some of the filters.

2.2.3 Definition of Breakpoints as Precursors of Structural Variants

SOPHIA collects evidence for a structural variant candidate breakpoint in a single-ended fashion where split read and discordant mate evidence for one breakpoint is gathered on-the-fly during the linear SAM stream without influence from the candidate partner site(s).

Technically, the algorithm collects any read that is classified as "discordant" in a pool of discordant reads during the line-by-line processing of the SAM stream. Each of these reads has a mate that can be an evidence for a particular imprecise structural variant. For any split read, soft-clipped (primary, outgoing evidence) or hard-clipped (secondary, incoming evidence) a new breakpoint is formed if necessary. Any further reads that exactly support this breakpoint are added as evidence to the previously initialized breakpoint. Right-sided clipped reads are supported by discordant reads to the left/upstream of the breakpoint, whereas the left-sided clipped reads are supported by discordant reads to the right/downstream of the breakpoint. We observed that the range where discordant reads supporting a structural variant are for the vast majority of true positive cases three times the default read length used by the sequencing technology (101 bps to 151 bps for our study). Following this guideline, there is a check during the processing of each SAM line, that breakpoints more than $3xDefaultReadLength$ away in the upstream direction (less on the coordinate space) than the aligned start position of the current read are prepared for "finalization". During finalization, the split read evidence is combined with discordant mate evidence for a given breakpoint. Split reads propose candidate structural variant target sites by their "Supplementary Alignments", whereas discordant mates propose candidate structural variant target sites by their mate coordinates. These target positions are matched by a fuzzy coordinate matching function allowing coordinate mismatches of up to

100 bases for split read target coordinate matching (required because breakpoint mapping can be imprecise when the breakpoint location and the split read overhang sequence share a similar repetitive pattern or if the breakpoint is on a repeat) and $2.5xDefaultReadLength$ for discordant mate based coordinate matching (required because of the inherent imprecise nature of discordant mate based breakpoint estimation). Despite this fuzzy comparison approach, in general, SOPHIA reports breakpoints with base-pair resolution provided that at least one split read is available.

There is a second check during the processing of each SAM line, that the discordant mate pool with reads at aligned start positions more than $6xDefaultReadLength$ are flushed, which corresponds to the theoretical maximum distance needed to keep reads in the discordant read pool to ensure availability in breakpoint evidence collection as described. This dynamic flushing ensures a minimal use of memory and an efficient operation by keeping the discordant mate pool, and hence the search space for mate-evidence small.

The ideal evidence for a single breakpoint is depicted in Figure 2.3, with the relevant read classes annotated (Figure 2.1): i) Soft clipped split reads clipped at a consistent breakpoint location with varying overhang base lengths indicating primary (outgoing) split read evidence, ii) Hard clipped split reads clipped at a consistent breakpoint location with varying base lengths indicating secondary (incoming) split read evidence, iii) Discordant reads with mate mapping locations consistent with each other and the target locations proposed by the split read evidence, iv) A low number of discordant reads with mate mapping locations that are inconsistent with the main proposed target, v) A low number of low quality reads.

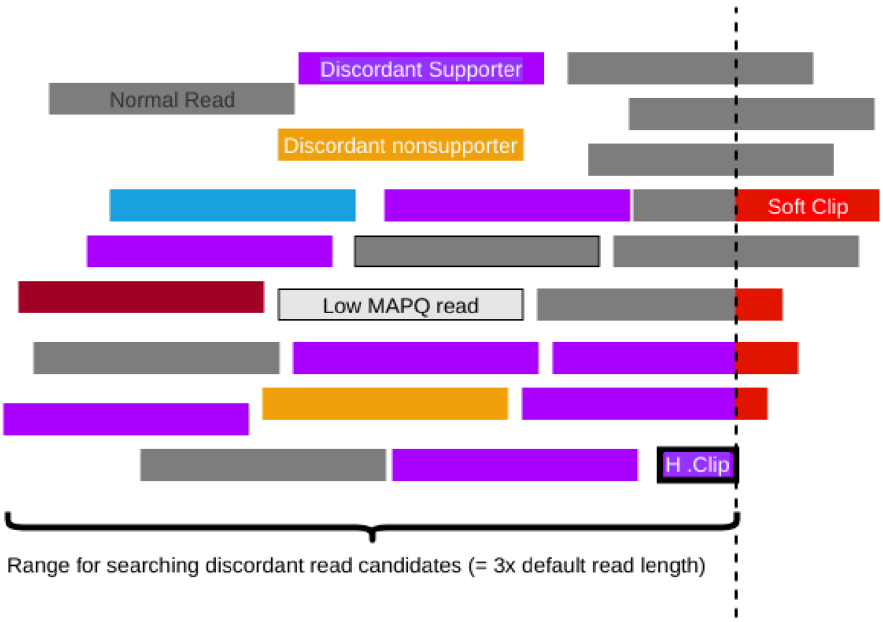

Figure 2.3: Extraction of a high-quality breakpoint from sequencing data according to the SOPHIA read categories. Each coloured bar corresponds to an aligned read. Gray reads are reads normally aligning to the shown genomic region. Reads with hollow interior have low mapping quality. Purple reads are discordant supporting the SV. Orange, red and blue reads are discordant reads not supporting the SV. The dashed line shows the position of the breakpoint. Red partial blocks correspond to split read sequences that match to the distant breakpoint partner of the shown breakpoint.

Split reads can lead to multiple alternative supplementary alignments, where both multiple solutions are possible per read and per breakpoint. SOPHIA builds consensus overhang sequences starting from the longest split read overhang as a seed and matching each shorter (or equal) read is matched to all seeds (allowing up to 2 mismatching bases), and generating new seeds as necessary. Each supplementary alignment proposed by a read belonging to an overhang consensus is assumed to be supported by all other reads that contribute to the consensus. While this assumption is not theoretically correct and leads to a number of proposed false positive targets, these false positives are easily eliminated because they would not be supported by the discordant read evidence and would be eliminated by the filters introduced in Section 2.2.6. While full-size clips i.e. the longest overhang seeds that are observed for a breakpoint, can intuitively be expected to propose the most correct target location information, complex structural variants can lead to insertion or deletion of bases at the target site, which can make the short split read mapping by BWA-mem more difficult. Thus, overall, this procedure significantly increases sensitivity thanks to the consideration of full-size clips as well as partial clips.

Frequently, a breakpoint can be located at a genomic location which does not allow the precise mapping of split reads, such as repeat regions. In particularly difficult instances of these cases, it can be that not even a single supporting split read is generated by the alignment of the rearranged region. SOPHIA can in such cases use unrelated split reads to initialize a breakpoint and propose a discordant-mate-only solution as (Figure 2.4). Unfortunately, due to

current algorithmic implementation limitations, clusters of discordant reads in regions without a single supporting or non-supporting split read cannot be reported. Such structural variants can still be called in a single-ended fashion if the partner site has sufficient evidence.

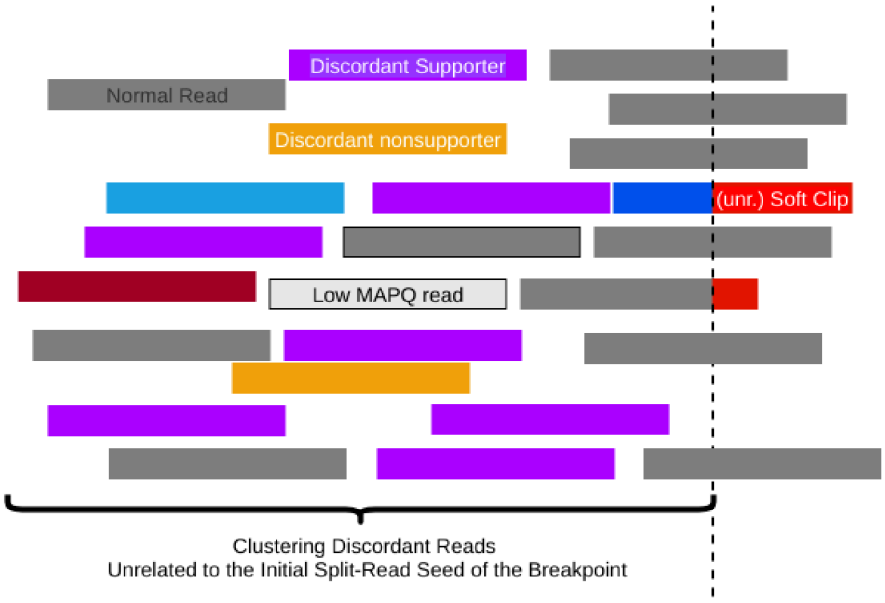

Figure 2.4: Extraction of an indirectly detected breakpoint from sequencing data according to the SOPHIA read categories. SOPHIA can use breakpoints with unrelated soft-clipped read sequences or entirely unrelated breakpoints for estimating inexact breakpoint locations in discordant read-based breakpoint detection.

Artefact regions generate a number of artefact signatures (Figure 2.5, with the relevant read classes annotated as in Figure 2.1): i) Inconsistent clipping locations across a repetitive region, ii) Split read overhangs with low quality base calls, iii) Overall a large ratio of low quality reads both with respect to mapping quality and base quality, iv) Dispersion of discordant mates to many different regions in the genome without a consistent clustering that proposes a single target site, v) Piling up of a large number of low quality reads that leads to an artefactual high coverage at the artefact breakpoint. These patterns are detected with a variety of mechanisms as presented in Sections 2.2.4 and 2.2.6 without relying on a theoretical and categorical exclusion of classes of repeat regions.

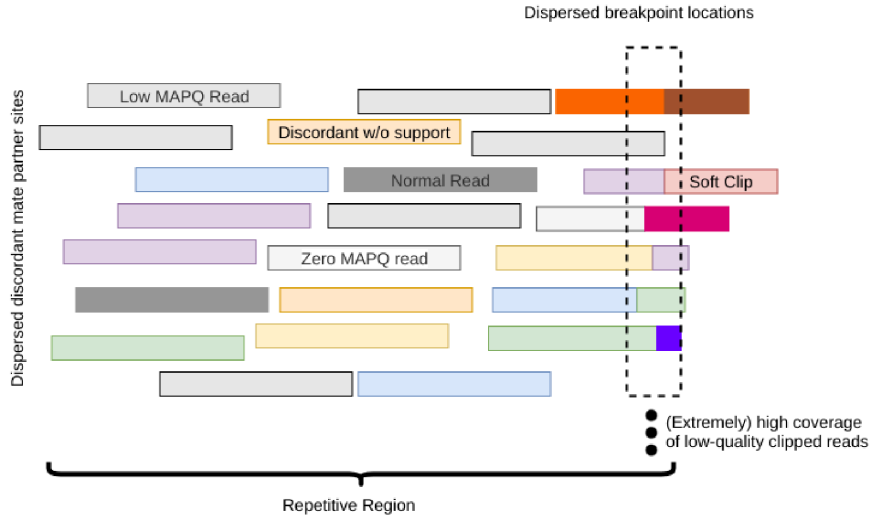

Figure 2.5: Extraction of an artefact breakpoint from sequencing data according to the SOPHIA read categories. The faded colours of the reads in the diagram correspond to the IGV convention of using hollow reads for non-unique mapping. The variety of colours indicates a dispersion of mate read mapping locations across different chromosomes.

A final class of structural variant breakpoints that has a particular evidence signature and different detection considerations is mid-sized duplications and deletions (Figure 2.6). Mid-sized implies that the event is not small enough to be supported by gapped reads and is not large enough to generate read pairs with discordant mate characteristics (size ¡ 5 standard deviations away from the median insert size of the merged alignment (bimodal or otherwise multimodal insert size distributions are not given special consideration) where this value is capped at 4000 base pairs). Because such structural variants have to entirely rely on split read evidence, i.e. a single class of evidence, sensitivity is overall lower, and specificity is ensured by different filters than the larger structural variants, not making use of discordant mate information.

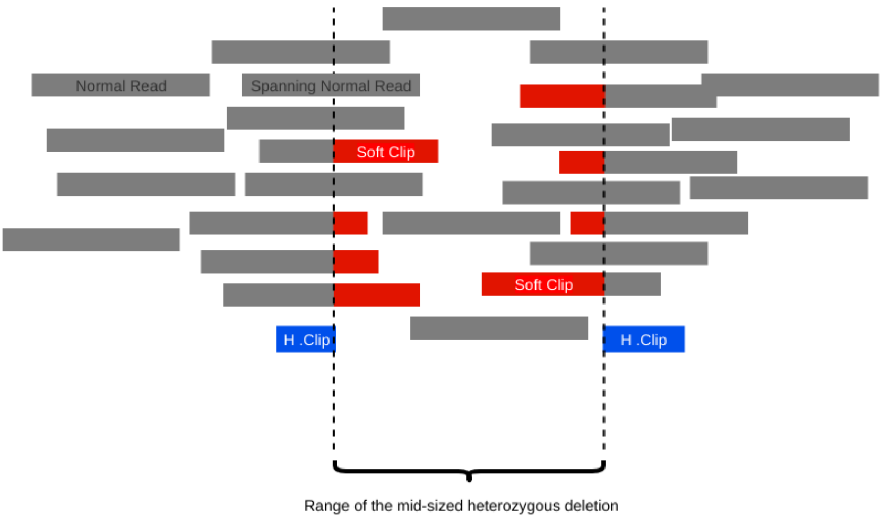

Figure 2.6: Extraction of the breakpoints of a heterozygous mid-sized deletion from sequencing data according to the SOPHIA read categories. A mid sized indel generates split reads and coverage changes (except in the case of inversions) but no discordant mates.

Currently SOPHIA does not incorporate gapped read evidence in its breakpoint detection approaches. Consequently, small insertions and deletions are not detected by the SOPHIA algorithm. More appropriate and fast tools such as Platypus [306] can be used for detecting such indels. However, it is planned to include gapped read evidence into the evidence collection for indels exactly at the border of the gapped-read domain and the split-read domain. Currently, SOPHIA could be losing some sensitivity in this indel size range with events that have combined gapped and split-read evidence.

2.2.4 Generation of a Population Background Database as a Quality Control Tool for Detection of Structural Variants

One of the cornerstones of the SOPHIA approach to structural variant detection is the training of a background database of breakpoints from whole genome sequencing data of "normal" blood tissue from a large number of diseases and ethnicities. We postulated based on previous experience with (inspecting) NGS data that: i) Artefacts most often emerge in repetitive regions at imprecise locations on and flanking the repeat, ii) The human genome is also rich in gaps and common breakpoints of structural variation which should be filtered out especially in the detection of structural variants without available paired normal data, iii) A common structural variant in the germline can be misclassified as somatic in regions with low depth of coverage and thus candidate somatic calls should be compared to a population background rather than only the available matched normal, iv) An algorithm trained on a large dataset should intuitively be much stronger than a paired analysis between a tumour sample and its matched normal.

To design a filter based on these expectations, we devised a strategy to collect data from "normal" samples in "paired normal" analyses in cancer whole genome sequencing projects. Here, we took advantage of two particular strengths of the DKFZ: i) Our participation in the Pan-Cancer Analysis of Whole Genomes (PCAWG) project which allowed us to capture a worldwide diversity of diseases and ethnicities, ii) Our recent acquisition of an Illumina X-Ten sequencing system which allowed us to build and a panel of artefact and real structural variant breakpoints from both older and newer sequencers and read lengths. We built a database of breakpoints from 3417 control samples, of which 2694 were sequenced with a 101bp sequencer (Illumina HiSeq 2000/2500 family) and 723 were sequenced with a 151bp sequencer (Illumina HiSeq X-Ten) covering a diverse range of participating countries and malignancies. A full list of all contributing projects and countries are available in the appendix of this dissertation.

The chosen samples were processed with the SOPHIA breakpoint extraction algorithm as described in Section 2.2.3. As the breakpoint database is expected to filter both artefactual and true breakpoints, breakpoints carrying "low-quality" read evidence (Section 2.2.2) and no proposed specific structural variant targets are also considered for inclusion in the database. However, breakpoints with very low numbers of either low- or high-quality reads are spurious and lead to noise in the database and can reduce sensitivity by making the database-based filters unnecessarily stricter. In order to avoid these issues, breakpoints were chosen to contain

i) At least 10 split reads with low or high quality, OR

ii) At least 3 gapped reads, OR

iii) At least 5 split reads AND an estimated clonality of 0.3, OR

iv) At least 5 split reads AND an estimated clonality of 0.1 AND at least 3 soft clipped split reads with at least 10 bases of overhangs that match with each other with at most 2 mismatching bases

. In addition to this information, the proposed high quality structural variant partners of each breakpoint in the database is also stored. This information is later used to capture hits in the database which do not match at the single base pair level, but rather due to the imprecise matching of the proposed structural variant partners. A crucial example for the importance of this additional information is genomic regions with high sequence homologies: In such regions, mapping quality is often zero and breaks can be assigned across the region on a non-deterministic basis by the aligner. This non-deterministic procedure leads to genes with homologies (such as genes to their pseudogenes, or closely related genes) being misassigned as rearranged with each other in some samples and not in others.

For each breakpoint in the database, an "artefact ratio" score is stored which is calculated using $\frac{N_{artefactReads}}{N_{artefactReads}+N_{highQualityBreakpointReads}+N_{normalSpanningReads}}$, showing the overall (lack of) quality of the stored breakpoint. This score is high for breakpoints originating from poorly mapped regions such as centromeres and telomeres but can be low for gaps in the human genome which are clean breaks. As such, the score does not necessarily indicate a common breakpoint to be definitely an artefact or a real common variant but it is a useful measure to investigate in the context of repeats. Therefore, the *ArtefactRatio* score is not used in the filtering procedure but rather in the analysis of the breakpoint database described in Section 2.3.1.

SOPHIA assigns a hit score to each breakpoint using a fast binary search algorithm during the annotation stage. The database is searched first for the closest existing breakpoint position to the exact position of the searched breakpoint. A broad search window of $6 * L_{default}$ bps (depending on the technology used for sequencing) is then used to find a matching existing structural variant that supports the proposed variant (hence, this operation is applied only in annotation, and not during the initial breakpoint definition as a breakpoint can "propose" multiple different variants). If there are multiple existing breakpoints that propose a given variant, the highest (worst-case) hit score is taken for the sake of higher specificity. Breakpoints are additionally searched within a narrower window (5 bps) compared to the standard search ($6 * L_{default}$ bps), which is then taken as the solution if the exact same variant is not known in the database. As before, if there are multiple breakpoint in the searched ± 5 bps window in the database, the highest (worst-case) solution is taken.

Thanks to this approach, SOPHIA can be used for somatic structural variant detection without paired controls or with low-quality paired controls (e.g. due to DNA degradation).

2.2.5 Pairing of Breakpoints as Candidates for Structural Variants

Breakpoints defined and characterized using the approach described in 2.2.3 need to be paired for defining simple structural variants. Even though more complex patterns from more than two breakpoints can emerge, these can be described as combinations of simple structural variants.

We define simple structural variants as described in [366] and [308] based on split read orientation. One challenge that should be addressed when breakpoints are only supported by discordant reads and not split reads is to manage pairings of imprecise locations. Such imprecise candidate SVs are mapped to a list of breakpoints by a fuzzy matching approach where an imprecise match with up to $2.5 * DefaultReadLength$ is accepted to constitute a match. For mid-sized SVs, imprecise matches are not accepted. A precise matching is allowed a smaller acceptable error margin of $0.5 * DefaultReadLength$.

2.2.6 Filtering Criteria for Structural Variant Candidates

A pair of two breakpoints connected contains a number of data points of evidence which overall determine the quality of a candidate structural variant. The expert model is built on these criteria, where quality cutoffs are set based on gold-standard structural variant information (Figure 2.1).

Evidence	Data Type	Evidence Type	Comments
Mid-sized SV	Boolean	Neutral	Mid-sized SVs incorporate no discordant mate info and require different filtering criteria
Inverted orientation of the two breakpoints	Boolean	Neutral	Mid-sized inversions are more artefact-prone
Decoy contig breakpoint-1	Boolean	Negative	Breakpoints emerging from decoy contigs are more artefact-prone
Decoy contig breakpoint-2	Boolean	Negative	...
MAPQ-0 only evidence for breakpoint-1	Boolean	Negative	Breakpoints emerging from entirely nonnspecifically mapped regions require stronger filters
MAPQ-0 only evidence for breakpoint-2	Boolean	Negative	...
Imprecise structural variant mapping proposed by breakpoint-1	Boolean	Negative	Imprecise variants are more artefact-prone
Imprecise structural variant mapping proposed by breakpoint-2	Boolean	Negative	...

Hits in the population background database for breakpoint-1	Integer	Negative	A breakpoint over-represented in the breakpoint background database is likely to be an artefact or a common polymorphism rather than a real rare structural variant
Hits in the population background database for breakpoint-2	Integer	Negative	...
Soft-clipped reads supporting breakpoint-1	Integer	Positive	Soft-clipped reads constitute important primary evidence for real structural variants
Soft-clipped reads supporting breakpoint-2	Integer	Positive	...
Hard-clipped reads supporting breakpoint-1	Integer	Positive	Hard-clipped reads constitute secondary evidence for real structural variants
Hard-clipped reads supporting breakpoint-2	Integer	Positive	...
Normal reads spanning breakpoint-1	Integer	Negative	Extremely subclonal breakpoints are more artefact-prone
Normal reads spanning breakpoint-2	Integer	Negative	...
Discordant-mate reads supporting breakpoint-1	Integer	Positive	Discordant-mate reads constitute important evidence for real structural variants
Discordant-mate reads supporting breakpoint-2	Integer	Positive	...

Ratio of discordant-mate reads supporting/(supporting+not supporting) breakpoint-1	Real number bounded $[0, 1]$	Positive	Dispersion of discordant mates to multiple unrelated locations indicates artefacts
Ratio of discordant-mate reads supporting/(supporting+not supporting) breakpoint-2	Real number bounded $[0, 1]$	Positive	...
Estimated germline clonality of breakpoint-1	Real number bounded $[0, 1]$	Negative	(only for somatic analysis), reads supporting the breakpoint in the germline indicate a likely artefact or germline structural variant
Estimated germline clonality of breakpoint-2	Real number bounded $[0, 1]$	Negative	...

These parameters are assembled in a complex decision-tree. A text format as in this dissertation is not optimal for showing each branch in this decision tree. However, the source code for SOPHIA is available under `https://bitbucket.org/utoprak/sophia/src` where the filters discussed in this section reside in the file *SOPHIA/src/SvEvent.cpp*. Following this decision tree, variants are ranked by a score between 1-5 and scores 3-5 are accepted as filtered structural variant candidates.

2.2.7 Tuning SOPHIA Structural Variant Detection Parameters using FISH Data as a Gold Standard

One special class of SVs is those created by aberrant actions of the haematological system. B-cells generate natural rearrangements in the Immunoglobulin loci IGH, IGK and IGλ (IGL) in processes called V(D)J recombination [432] and Class Switch Recombination (CSR) [433]. This process which has the original purpose of generating and extending antigen repertoires, can lead to malignancies such as B-cell lymphoma, multiple myeloma, chronic lymphocytic leukemia if it aberrantly targets oncogenic partner loci. T-cells similarly undergo the V(D)J recombination process in T-Cell receptor loci, whose aberrant action can lead to T-cell malignancies such as T-cell leukemia or T-cell lymphoma.

B-Cell lymphoma frequently harbours rearrangements of the immunoglobluin locus to hallmark oncogenes and less frequently to sporadic targets. Of the three immunoglobulin loci IGH, IGK and IGL, IGH rearrangements are the most common. We analysed hallmark rearrangement target oncogenes of B-Cell Lymphoma *MYC*, *BCL2* and *BCL6* as well as the rearrange-

ments of the IGH locus with break-apart FISH probes acquired by the expert Dr. Cristina López Gonzalez, who previously ran this assay in Chronic Lymphocytic Leukaemia (CLL) in a large-scale project [434]. FISH on interphase nuclei was performed on frozen tissue sections applying the specific probes LSI *BCL6*, LSI *MYC*, LSI IGH/*MYC*, CEP8 Tricolor, LSI IGH, and LSI *BCL2* as described in [435]. We used the oncogene probes *BCL2*, *BCL6* and *MYC* for parameter training, but not the IGH break-apart probes.

The parameters listed in Section 2.2.6 were assembled in a decision tree using high quality FISH data. FISH on Immunoglobulin translocations were selected because of existing clinical-grade workflows offering gold standard reference data and the intrinsic difficulty involving the immunoglobulin loci: in immunoglobulin (IG) loci, the genomic complexity is high, which is exacerbated by the combination of internal rearrangements with oncogenic rearrangements variants [436]. Furthermore, immunoglobulin variants are with few exceptions balanced translocations which lead to two, single-sided breakpoints per variant, providing more training data, as well as taking out the contribution of coverage differences, which increases the difficulty scope covered by the training data.

Gold standard IG rearrangements detected by FISH were manually inspected for determining performance of parameters presented in Section 2.2.6 in an iterative manner. Parameters were optimized to capture the known variants, while keeping track of the emergence of subclonal variants (more frequently of repetitive regions) to ensure that specificity is not being unduly lost. Successively, parameters were optimized to capture more and more subclonal and difficult-to-detect IG rearrangements with progressing SOPHIA versions.

As part of the collaboration agreement of the ICGC-MMML sequencing consortium, the Korbel group affiliated with EMBL, Heidelberg provided the consortium with SV calls originating from their algorithm Delly (called by Stéphanie Sungalee, using v0.5.9 as described in [435]). We did not use Delly calls for parameter optimization purposes, i.e we did not use SVs captured by Delly and missed by FISH or SOPHIA for further parameter optimization.

In addition to IG rearrangements, we also used a more global analysis of large variants using the M-FISH assay [437] which is not a targeted technique unlike FISH, but still is limited to large variants and offers limited resolution. Nevertheless, it does not suffer from the issues of short-read sequencing around repetitive regions, and can act as a gold standard provided that it is executed by an experienced expert. We used Dr. Larisa Savelyeva's work on the neuroblastoma cell line NB-69 and further improved SOPHIA filtering parameters especially in complex genomic regions.

2.2.8 Custom Filters Based on Known Artefact Structural Variants

Systematically established filters are unfortunately insufficient in ensuring a perfect rate of specificity despite the power of SOPHIA's breakpoint database. Some of the recurrent artefacts that we frequently observed in SOPHIA results and needed to develop additional filters for are as follows:

i) t(X,Y) translocations that emerge due to pseudoautosomal regions,

ii) rearrangements between coding genes and their pseudogenes,

iii) artefactual mitochondrial insertions into nuclear pseudogenes of mitochondrial DNA,

iv) translocations between *NOTCH2* and *NOTCH2NL*

These filters are all designed to eliminate artefacts arising from sequence homology. While it was our aim to ensure that the background breakpoint database would account for this type of error, some additional filtering proved to be necessary based on detailed examination of SOPHIA results from a large number of projects.

2.2.8.1 Accounting for Sequencing Quality Issues that Lead to Lowered Proper Pairing, Excessive Mate Dispersion and Loss of Sensitivity and Specificity

We observed lowered proper pairing in a subset of samples across multiple projects, sequencers and sequencing read lengths. This manifested as a high load of read pairs with incorrect orientation suggesting artefactual inversions, and dispersed mapping of mate reads to diverse genomic regions and chromosomes suggesting artefactual intrachromosomal and interchromosomal translocations. Both of these observations can lead to a high rate of false positives, i.e. lowered specificity. Compensating for these false positives consequently leads to a high rate of false positives, i.e. lowered sensitivity. Recognizing this, we nevertheless developed a method to account for lowered sequencing quality, addressing both of the described types of error. Our effort to attempt to salvage such samples was motivated by the preciousness of starting material in human cancer studies.

We modified the SOPHIA workflow in two stages:

i) If a sample has a "proper pair ratio" as calculated by samtools flagstats [303] lower than 90%, the read assignment procedure to breakpoints described in Section 2.2.3 is modified to expect a background error rate of $(100 - PP_{ratio})\%$. Breakpoints with partners beyond the mid-sized structural variant range with $N_{mateSupport}/N_{totalDiscordant} \leq (100 - PP_{ratio})/100$ are discarded from further analysis, with the assumption that the suggested structural variant is below the error/noise level. As the formula suggests, the effect gets progressively stronger as the sequencing quality is decreased, consequently very low quality samples can be expected to have a large number of false negatives, encouraging the removal of the sample from the study or resequencing it if material and funding is available.

ii) A secondary fix is made after the breakpoints are paired and preliminary filtered structural variants are obtained: If the total count of preliminary (filtered) structural variant candidates are above 300 and the ratio of candidates with inverted pairing are over 0.7, a first clean-up stage is applied where SVs with missing classes of evidence (split reads and discordant mates for both breakpoints) are removed from further analysis for SVs larger than mid-sized. During the same clean-up stage, mid-sized SVs with less than 5 supporting reads for either side are also removed. This is followed by a second clean-up stage which is applied, if more than 200 candidate SVs remain with the ratio of candidates with inverted pairing over 0.7, which has even stricter filters: Inverted SVs with imprecise mapping are removed no matter how many reads support them, SVs with

$N_{supportingDiscordantMates}/N_{expectedDiscordantMates} < 0.6$ for both breakpoints are removed. Overall, these procedures rescue many clonal variants with strong evidence and filter out massive numbers of artefact SVs. However, when sufficient material is available, samples where these filtering levels are applied should be resequenced, especially if the issues affect the tumour sample. Issues affecting the blood sample only have the effect of increasing the number of misclassified germline SVs, which is compensated by the use of the background breakpoint database.

2.2.9 Designation of Structural Variants as Somatic or Germline

For samples where a paired normal sample is available, we started by building a database of germline breakpoints by processing the paired normal alignment using the procedures described in Section 2.2.3. Following the establishment of this database, we used the exact same procedure as described in Section 2.2.4 for searching for tumour breakpoints, this time in the paired germline breakpoint database as opposed to the background population breakpoint database. In the event that one side is a germline breakpoint, and the other is a somatic breakpoint, we designated the structural variant as a germline variant, for the sake of protecting specificity.

During our work on SOPHIA, we encountered two situations which raised the need to address the specificity of germline-somatic designation:

2.2.9.1 Tumour in Normal (TiN) contamination in Plasma Cell Leukaemic Multiple Myeloma and MYCN amplified Neuroblastoma have somatic structural variants misclassified as germline

We observed in an analysis of two different disease types, a prevalence of tumour cells in blood leading to subclonal evidence in the paired control samples suggesting the existence of a germline structural variant, which is expected to be a clonal somatic structural variant according to established knowledge.

The first example of this type of artefactual observation was made in the HIPO-067 refractory multiple myeloma project. In the late stages of this disease, plasma cell leukaemia [438] where plasma cells circulate in peripheral blood. As the circulating plasma cells are transformed, and carry the clonal structural variants that led to the neoplasm or evolved with the neoplasm, sequencing results from a peripheral blood sample used as a matching control would carry evidence for the somatic structural variant. This would interfere with the correct classification of these structural variants as somatic. In order to compensate for such artefacts, users should manually revert to the no-control mode of the SOPHIA workflow. This can be decided by a manual inspection of the results searching for hallmark somatic structural variants suggested as germline variants, or with a quantitative approach using single nucleotide variants (SNVs) [439] proving a tumour-in-normal contamination.

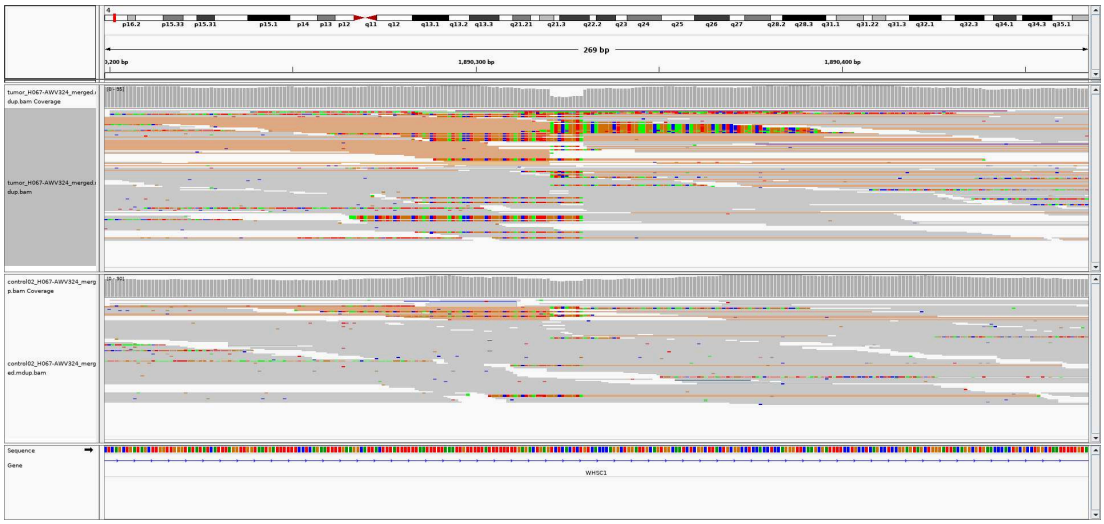

Figure 2.7: A Tumour-in-Normal contamination example involving a prototypical balanced IGH-NSD2 rearrangement in refractory multiple myeloma. There is strong support for the rearrangement in the tumour data and weaker, but still significant support in the normal data. Orange reads indicate reads on the *NSD2* locus on chromosome 4 whose mates map to chromosome 14, on the IGH locus. The upper subfigure indicates the tumour sample, whereas the lower subfigure indicates the matched normal sample with tumour material contamination.

The second observation we made was the existence of *MYCN* amplifications in the germline in our analysis of the GPOH Neuroblastoma cohort, as suggested by structural variant calling using a paired tumour-blood approach. It is known that *MYCN* amplifications are typically over 10-fold from the baseline state [440] and even a small concentration of tumour cells in blood can generate enough evidence suggesting *MYCN* structural variants in blood. Interestingly, we did not observe this in *EGFR* or *MYEOV* amplifications in adult cancers such as Glioblastoma Multiforme or Head and Neck Cancer, suggesting the higher order amplifications in Neuroblastoma to be the main reason of this observation. In order to compensate for such artefacts, we imposed a condition that evidence suggesting high order amplification in the tumour sample ($>$ 200 split reads in support of the rearrangement for both breakpoints) constitute somatic variants.

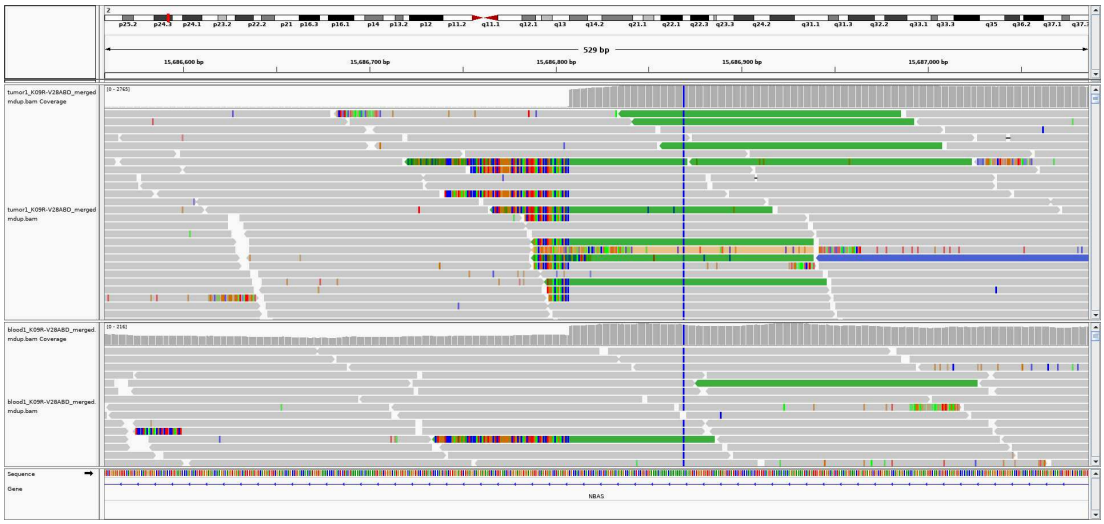

Figure 2.8: A Tumour-in-Normal contamination example involving a prototypical high order *MYCN* amplification in neuroblastoma. There is strong support for the amplification in the tumour data and remarkably strong support in the normal data. The upper subfigure indicates the tumour sample, whereas the lower subfigure indicates the matched normal sample with tumour material contamination. The high coverage increase and the green reads showing distant mate mapping indicate the amplification event, for which strong evidence exists also in the matched normal data.

2.2.9.2 Massive germline load of retrotransposons in patients from ethnicities underrepresented in the SOPHIA breakpoint database

We observed a massive load of germline interchromosomal translocations in a small number of cases across diverse projects. Manual inspection of candidate interchromosomal translocations suggested them to be not artefacts, but rather retrotransposons where short sequences jump between chromosomes, which is a normal evolutionary process in mammals [441].

The following representative cases fit this description:

- The case pseudonymized as PCSI_0101 from the Canadian ICGC Pancreatic Cancer project, originating from Kuwait (personal communication, Dr. Lincoln Stein, OICR)
- The case pseudonymized as 4154480 in the DKFZ *RCC1-IRF4* lymphoma project, originating from South-East Nigeria (personal communication, Cristina López Gonzalez, Uni. Ulm)
- The case pseudonymized as XI102_AML-3 in the XI102 DKFZ Acute Myeloid leukemia project, with an unknown ethnic origin

For the last two cases paired normal specimens were not available, which made it impossible to filter out rare ethnicity related transposons from somatic structural variants. For such cases, there is no currently available solution as our trials with the larger background breakpoint database obtained with Lumpy [442] also failed to filter out most of the misclassified variants on the Nigerian case.

Samples with the described characteristics should be carefully manually inspected, and be excluded from germline analysis if paired normal data is available, or be excluded from the study if only no-control analysis can be run.

A representative example of such a transposable element is a t(2,7)(p22.3;q36.3) from the Kuwaiti case PCSI_0101, (Figures 2.9 and 2.10).

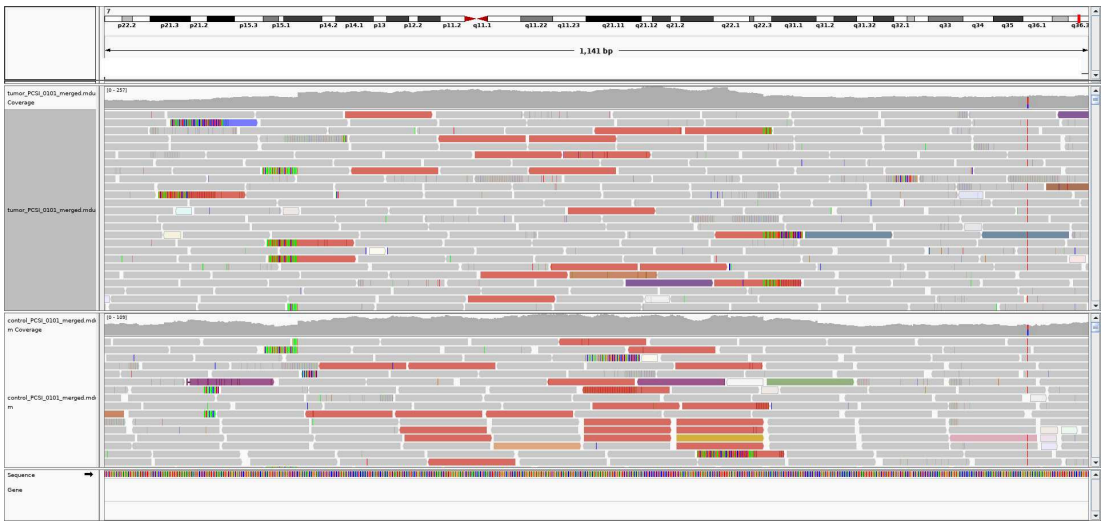

Figure 2.9: The jumping transposable element originating from chr7q36.3. Exact base-pair positions are masked. The light-red reads are reads whose mates map to chromosome 2, specifically chr2p22.3. The coverage increase indicates an extra copy of the sequence being created before the sequence jumping event. The upper subfigure indicates the tumour sample, whereas the lower subfigure indicates the matched normal sample indicating the presence of the jumping sequence in the germline with equal clonality.

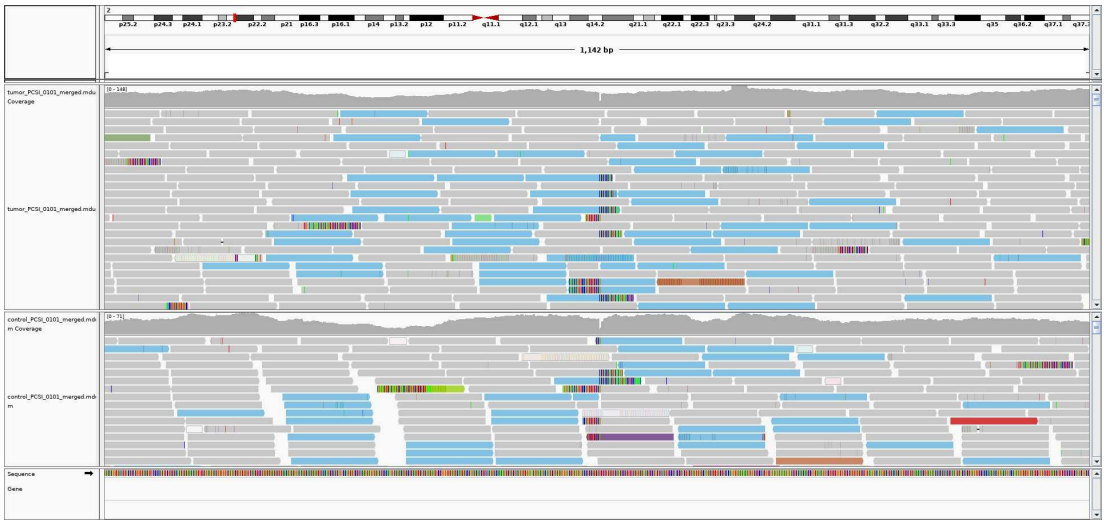

Figure 2.10: Insertion target of the transposed sequence on chr2p22.3. Exact base-pair positions are masked. The blue reads are reads whose mates map to chromosome 7, specifically chr7q36.3. The sharp and narrow coverage fall indicates the insertion site for the sequence jumping event. The upper subfigure indicates the tumour sample, whereas the lower subfigure indicates the matched normal sample indicating the presence of the jumping sequence in the germline with equal clonality.

These observations are difficult or impossible to be distinguished from a normal balanced translocation such as those observed on immunoglobulin loci and their targeted oncogene translocation partners. Thus, we currently do not have automated methods to specifically exclude this class of structural variants from further analysis be it in the germline or in the no-control setting.

2.2.10 Annotations for Structural Variants called by SOPHIA

Structural variants can have significant effects on chromatin conformation and lead to gene dysregulation even if they are not directly on gene bodies. Thus, the interpretation of structural variant calls requires an approach to map breakpoint locations to genes. The concept of Topologically Associating Domains (TADs) [33] introduces a systematic model for the co-regulation of genes in close proximity. TADs offer a data-driven model of cis-regulation of genes, and a more advanced approach than fixed windows for the interpretation of breakpoint effects on genomic regions.

2.2.10.1 Remapping the Human Decoy Chromosome hs37d5

We started the SV effect annotation procedure by remapping breakpoints that initially map to the human decoy chromosome hs37d5, *when appropriate*, to other chromosomes or smaller contigs. We used the definitions of the hs37d5 decoy chromosome provided by Dr. Heng Li's repository `https://github.com/lh3/misc/tree/master/seq/novoseq`. We used the mapping files hs37d5cs.bed and hs37d5ss.info to localize the hs37d5 contig segments to

other genomic regions. We skipped all segments mapping to chrY. We also skipped all segments with unlocalized anchored 5', 3' ends as well as and unlocalized top hit to the GRCh37 primary assembly. As a result we remapped $4673/4715$ of the decoy segments to other chromosomes or unlocalized (GL***) sequences.

While this procedure is error-prone, we were motivated by the existence of 9 decoy contigs mapping to the IGH locus on chr14q32.33. As IGH translocations are important for haematological malignancies, it is important to characterize them sensitively. Taking this as the starting motivation, we expanded the correction to other loci, recognizing the possible inaccuracies introduced by this remapping. Hence, the original locations of the structural variant candidates on the hs37d5 chromosome are kept in SOPHIA outputs.

2.2.10.2 Definition of Consensus TADs from Public Datasets

We first used the list of TADs published in [294] (13 cell lines), [443] (6 cell lines), [444] (5 cell lines), [445] (4 tissues), and [446] (9 tissues) obtained from the chromatin-capture database processed and provided by the Feng Yue Lab (`http://promoter.bx.psu.edu/hi-c/downloads/hg19.TADs.zip`, obtained 07.01.2018) to build a consensus list of TADs. As discussed in [446], TADs show a remarkable similarity between different cell types, confirming previous assumptions of stability across tissues [294], with their level of activation differing between each tissue depending on its epigenetic development. We took this assumption to be true and consequently took differences between the TAD boundary measurements coming from these studies to be due to experimental and technical factors, justifying a consensus approach.

For creating a consensus between 37 datasets with TAD range data, we first converted TADs to TAD boundaries, which are due to the nature of the Hi-C assay and its data processing digitized in 40kb windows. Next, we assembled TAD boundaries in clusters where the data from the 37 datasets is sorted by genomic coordinates, and each TAD boundary is added to growing clusters if its starting position is at most 120kb away from the current cluster. Clusters of TAD boundaries are then "compressed" into consensus TAD boundaries. These TAD boundaries are then converted to overlapping TADs. (3246 TADs from chromosomes 1-X)

Due to lack of available data chromosome Y and other contigs were not assembled into TADs derived from experimental data:

1. There is no chromatin capture-based data for the TADs of chromosome Y in the used data sources, so we used the existing cytoband definitions for ChrY as a rough replacement for TADs. (12 TADs from chromosome Y with a cytoband approximation)

2. The mitochondrial chromosome (MT) was not segmented into TADs or cytobands due to lack of available data. (1 TAD representing the mitochondrial genome)

3. The Epstein-Barr Virus contig (NC_007605) was not segmented into TADs or cytobands due to lack of available data. (1 TAD representing the EBV contig)

4. The human unlocalized sequences (GL****) were not segmented into TADs or cytobands due to lack of available data. (59 TADs each representing one unlocalized sequence)

5. The human decoy chromosome hs37d5 is a collection of independent and unrelated contigs that are not mappable to the normal chromosomes, so each of these contigs were considered independent TADs and not processed using the approach above. (4715 TADs representing contigs of the hs37d5 chromosome)

Overall, we generated using this procedure 8201 consensus TADs, of which 3246 stem from 37 experimentally acquired, Hi-C based chromatin capture datasets.

2.2.10.3 Assignment of Structural Variants to Consensus TADs

For chromosomes with TADs obtained from available chromatin capture data, structural variants were classified into three groups for the purpose of assigning to ”seed” TADs:

i) Interchromosomal translocations: Interchromosomal translocations are considered as a union of two breakpoints. Breakpoints that are on a TAD boundary are considered to affect both TADs separated by the boundary. Otherwise, the ”seed” TAD is the TAD which is directly hit for a given breakpoint.

ii) Intrachromosomal structural variants within discordant read supported range: Intrachromosomal SVs are considered in the same manner as interchromosomal translocations. In addition to the described procedure, the TADs between the two breakpoints are also considered as affected if the following conditions are met:

   - Both breakpoints are on the same chromosome arm
   - The spanned genomic range is shorter than 10MB
   - There are less than 4 TADs between the smallest and largest ”seed” TAD

   This procedure is important for some focal deletions such as deletions of the CDKN2A/B locus.

iii) Mid-sized structural variants with no discordant read support:

   - Mid-sized SVs that are intergenic, and not spanning a TAD boundary are discarded from further analysis.
   - Mid-sized SVs that are on a TAD boundary are considered to affect both TADs separated by the boundary, regardless of gene hitting status.
   - Mid-sized SVs that are on gene bodies, but intronic and not directly hitting ROADMAP enhancers are discarded from further analysis.
   - Mid-sized SVs that overlap transcribed regions are considered to affect the TAD which is directly hit.

For all SVs apart from Mid-sized SVs, a TAD offset extension procedure is applied for investigating possibly affected genes across longer ranges. Following the definition of the initial ”seed” TADs, extensions at 1, 2, and 3 TAD offsets are calculated in both directions. For right-sided extensions, the starting position of the extended TAD and for left-sided extension, the ending position of the extended TAD is tested for closeness to the position of the breakpoint.

If the distance exceeds 5MB, the extension is cancelled. This procedure allows the estimate for the influence of an SV on TADs that it is not directly hitting, within a sensitivity limit set by the user.

The results are reported as part of the annotated SOPHIA output both as the indices of annotated TADs and as lists of the affected genes for each TAD offset level.

For the benchmark presented in Figure 2.3.3, the highest permissible TAD-offset is 1, i.e. a hallmark gene is considered affected if a filtered SV is estimated to affect the gene's TAD (or one of its TADs if the gene spans multiple TADs) or one of its neighbouring TADs in either direction.

2.2.10.4 Gene Definitions

We used the ENCODE consortium's gene reference annotation GENCODE [447], version 27 (released 08.2017), lifted over to the GRCh37 genome assembly.

The following gene types were included in the reference used by SOPHIA: IG_C_gene, IG_C_pseudogene, IG_D_gene, IG_J_gene, IG_J_pseudogene, IG_pseudogene, IG_V_gene, IG_V_pseudogene, lincRNA, macro_lncRNA, miRNA, polymorphic_pseudogene, processed_pseudogene, unprocessed_pseudogene, protein_coding, transcribed_processed_pseudogene, transcribed_unitary_pseudogene, transcribed_unprocessed_pseudogene, translated_processed_pseudogene, TR_C_gene, TR_D_gene, TR_J_gene, TR_J_pseudogene, TR_V_gene, TR_V_pseudogene while the following were discarded from the reference: 3prime_overlapping_ncRNA, bidirectional_promoter_lncRNA, misc_RNA, Mt_rRNA, Mt_tRNA, non_coding, retained_intron, processed_transcript, rRNA, scRNA, sense_intronic, sense_overlapping, snoRNA, snRNA, TEC, unitary_pseudogene, unprocessed_pseudogene, vaultRNA

As genes can have multiple alternative transcripts, we attempted to reduce the gene set as far as possible to the most canonical transcripts in order to facilitate further analysis. To this end, we ranked isoforms based on their APPRIS [448] scores, in order of precedence: appris_principal, appris_principal_1, appris_principal_2, appris_principal_3, appris_principal_4, appris_principal_5, appris_candidate_longest, appris_candidate, appris_alternative_1, appris_alternative_2, (not available).

Where multiple transcripts exist for a gene and APPRIS scores are not sufficient for tie-breaking, we used the "transcript_support_level" entry in GENCODE in order of precedence: 1 (all splice junctions of the transcript are supported by at least one non-suspect mRNA), 2 (the best supporting mRNA is flagged as suspect or the support is from multiple ESTs), 3 (the only support is from a single EST), 4 (the best supporting EST is flagged as suspect), 5 (no single transcript supports the model structure), NA (the transcript was not analyzed).

Where multiple transcripts exist for a gene and APPRIS scores and GENCODE "transcript_support_level" scores are not sufficient for tie-breaking, we used the "level" entry in GENCODE in order or precedence: 1 (verified loci), 2 (manually annotated loci), 3 (automatically annotated loci).

Where multiple transcripts exist for a gene and APPRIS scores and GENCODE "transcript_support_level" and GENCODE "level" scores are not sufficient for tie-breaking, we used the exon counts of the alternative transcripts as a tie-breaker, taking the transcript with the highest number of exons as the canonical transcript for the gene model.

The SOPHIA workflow uses BEDTOOLS [449] for annotating direct gene hits as well as the nearest genes upstream and downstream of the breakpoint for each of the two breakpoints that make up an SV.

2.2.11 Gene Expression Data Processing

There are two major technologies for quantitative analysis for gene expression data: array based technologies, and sequencing based technologies. RNASeq has a number of advantages over RNA microarrays such as the ability do detect novel transcripts and gene fusion events, do genotyping, avoiding transcript probe based artefacts. Despite these advantages, RNASeq has its own challenges such as biases within samples due to transcript length and biases between samples or batches. Hence, normalization of read counts from RNASeq data is an important step in ensuring comparability between different genes or samples in a study and is an active research question with different benchmarking studies and tools on this subject [450] [315] [451] [452].

Choosing a RNASeq count normalization method depends on the desired application. Comparison of a gene across multiple cohorts requires different approaches from the comparison of a gene across donors in a single cohort. In the cited benchmarks, TMM normalization [315] offered by the edgeR Bioconductor package [453] was consistently ranked as a top-class normalization algorithm along with the DESeq2 [454] approach. For the purpose of benchmarking SOPHIA (Section 2.3.3), we used the TMM normalization in the edgeR package within each TCGA cohort: Raw read counts were obtained from the Genomics Data Commons (GDC) mirror of UCSC Xena [455] and pre-normalized by the Counts Per Million (CPM) calculation. Genes with less than 1 CPM for all samples across a given cohort were discarded from further analysis. Then edgeR normalization was applied with default parameters on the initial gene counts of the filtered gene set, followed by another application of CPM and $log_2(n+1)$ normalization.

For the GPOH-NB project, gene expression read count values were obtained using the DKFZ RNA-Seq pipeline [456] and normalized as described.

For Medulloblastoma cohort under ICGC-PedBrain, we used the RNA microarray data instead of RNA-Seq data because of the better coverage of the cohort [211]. Results were downloaded from the R2: Genomics Analysis and Visualization Platform [337], obtained using the Affymetrix u133p2 array and normalized with the MAS5.0 algorithm [457].

Finally, normalized gene expression values were visualized and inspected for ”breaks” indicating bimodality at known hallmark genes with oncogenic activation via structural variants, where the existence of an affecting SV is estimated using the approach described in Section 2.2.10.3.

2.3 Results

2.3.1 Analysis of the SOPHIA Background Breakpoint Database

We analysed the SOPHIA background breakpoint database for behaviour around repeat classes, both with respect to breakpoint counts and breakpoint quality. For filtering of structural vari-

ants, we used the merged complete database, while for this analysis, we separated the database into two components based on the used sequencing technology: 2694 from a 101bp technology (Illumina HiSeq 2000/2500 family), and 723 from a 151bp technology (Illumina HiSeq X-Ten).

First, we investigated how repeat families differ in terms of attracting breakpoints either due to real germline structural variation or due to sequencing or mapping artefacts (Figure 2.11).

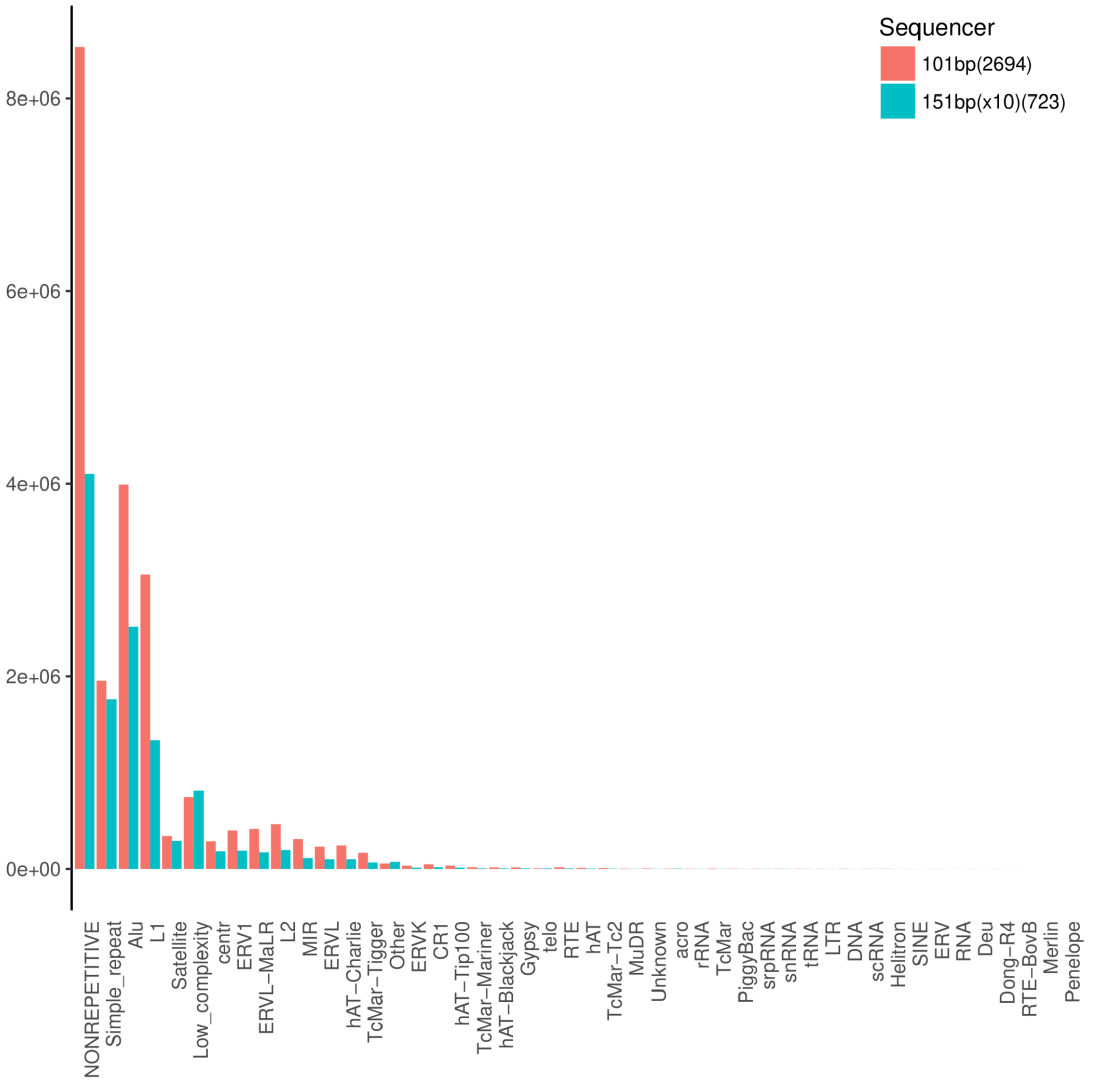

Figure 2.11: Distribution of database breakpoints across all repeat families and nonrepetitive sequences. The y-axis corresponds to the number of breakpoints that map to a genomic location belonging to a given repeat family.

Due to the strong contributions of the top 14 repeat families, we decided to combine the rest into an ”other” category, and focus on the top 14 families (Figure 2.12).

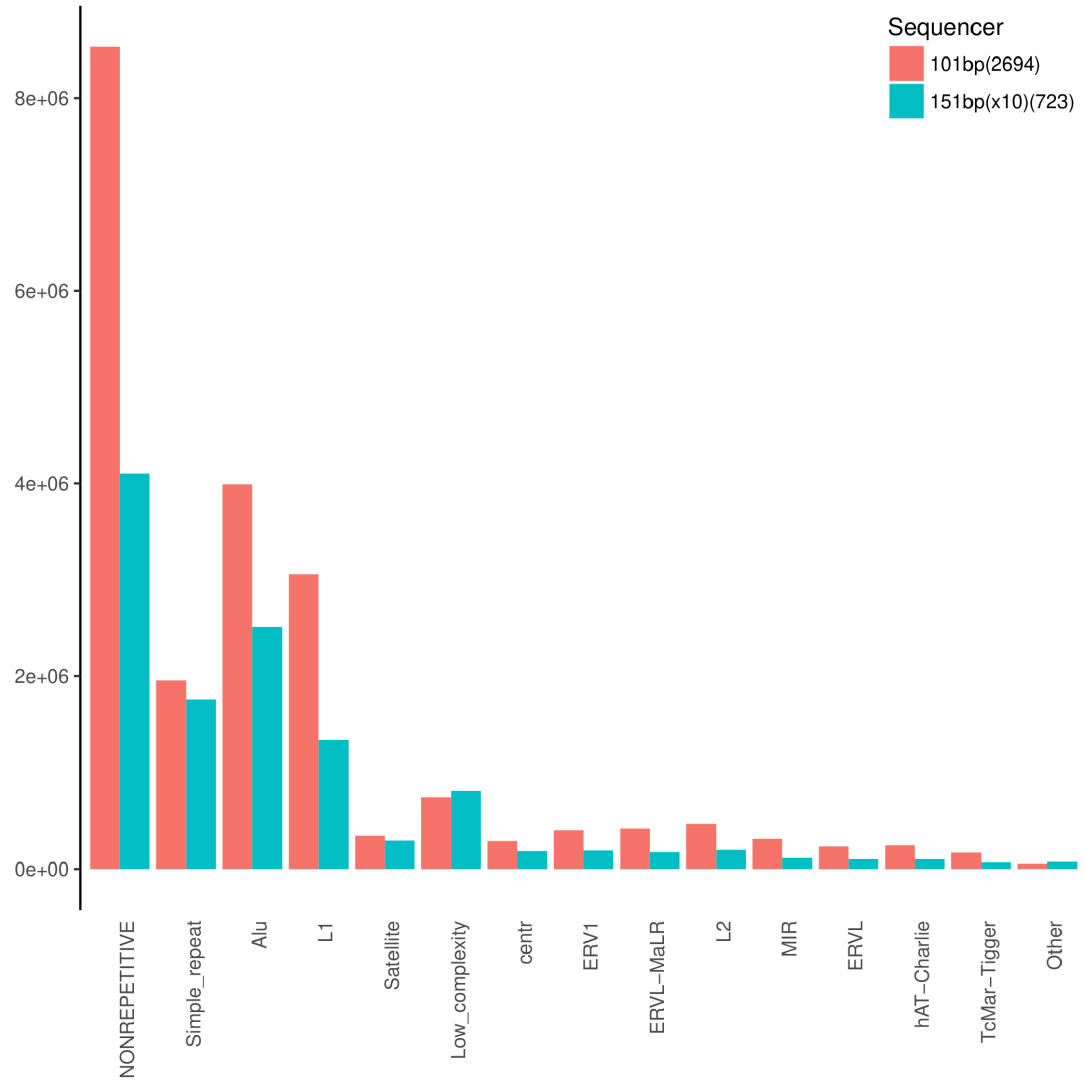

Figure 2.12: Distribution of database breakpoints across top repeat families with respect to total numbers of breakpoints from both technologies.

The 151bp breakpoint database consists of data from the DKFZ, which can be assumed to lead to a significant enrichment for German donors even though we did not run a genotyping analysis on the cohort. Thus, the 101bp breakpoint database has a greater ethnic diversity as well as a larger sample size. Consequently, for a given randomly selected set of genomic positions, it is expected that the larger and more diverse 101bp breakpoint database would have higher counts of matching breakpoints. In our analysis of repeat families (Figures 2.11 and 2.12), we see this expectation to be confirmed with a few notable exceptions: the 151bp breakpoint database outperforms its normal trend (of lower breakpoint counts per repeat family) for the repeat families *simple repeats*, *satellite repeats* and *low complexity repeats*. This observation suggests that these repeat families could be less ”accessible” with sequencing using shorter reads.

We then investigated the two breakpoint databases from the perspective of breakpoint quality for the top repeat families (Figure 2.13).

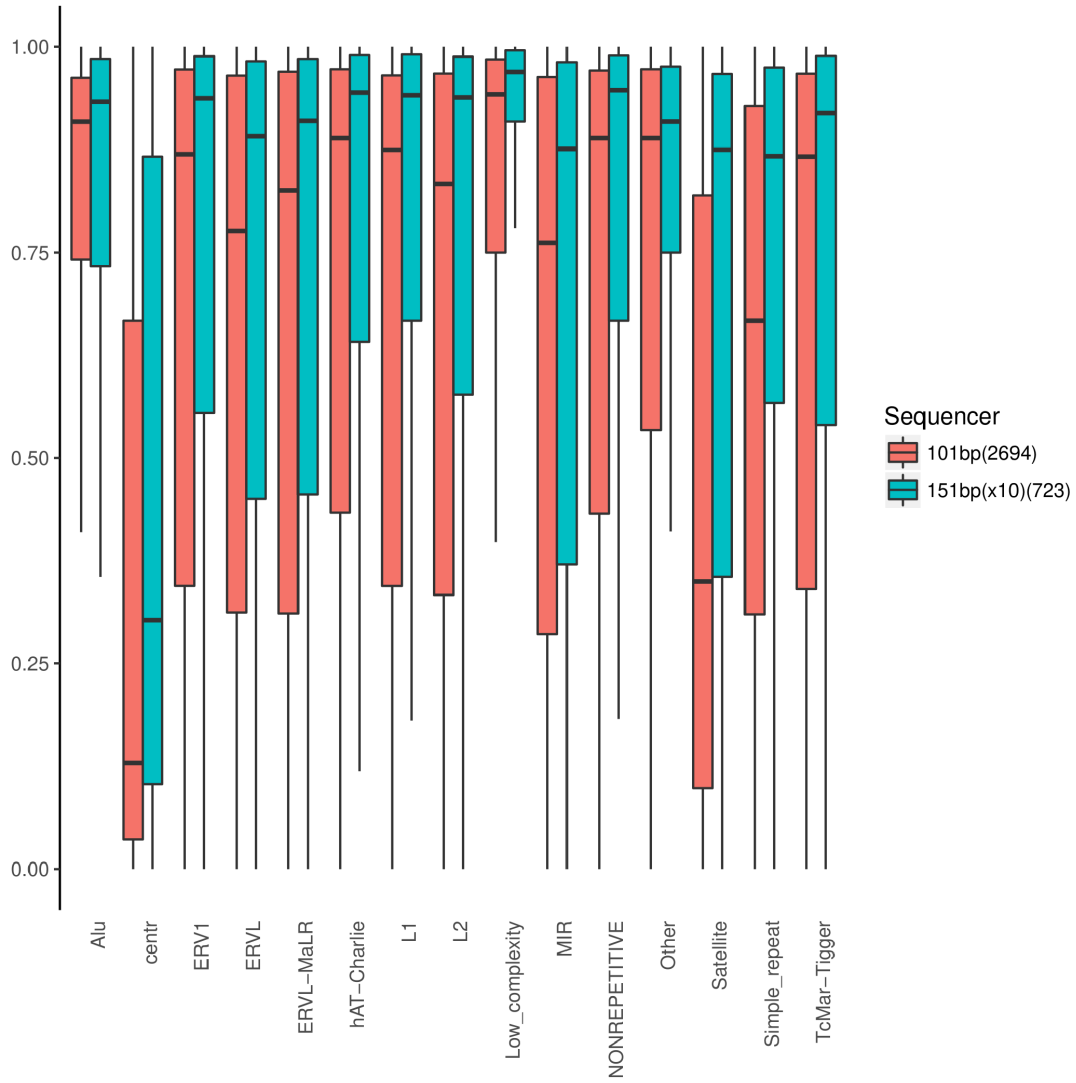

Figure 2.13: ArtefactRatio score comparison across the two background population breakpoint databases. The ArtefactRatio score indicates a lack of breakpoint quality/mappability where higher values are likely to yield regions with artefactual breakpoints and candidates for false positive SVs.

The X-Ten sequencer overall yields breakpoints with a higher *ArtefactRatio* score for most repeat families (Figure 2.13). Strikingly, the difference between the two databases seems to be strongest where the 151bp database captures more breakpoints: *simple repeats*, *satellite repeats* and *low complexity repeats*. Another important observation is that only *low complexity* and *Alu* repeat families are consistently generating predominantly low quality breakpoints, whilst most other repeat families and also nonrepetitive sequences generate both high and low quality breakpoints. This observation shows issues with a repeat family-based blacklisting of structural variant results: filtering SV results based on one of the two breakpoint mapping to blacklisted repeat family types would yield both false positives and negatives. Though we do not have the data to conclusively prove this, it could be that the 151bp reads reach some repeat regions that 101bp reads cannot, while still being unable to cleanly resolve them. It will be interesting to

observe if read lengths will continue to grow as sequencing technology advances, and at which point the mapping issues on repetitive regions, and especially those of lower complexity, will be fully addressed.

2.3.2 SOPHIA Detects Hallmark Immunoglobulin Rearrangements in B-Cell Lymphoma with High Sensitivity

We ran a three-way discrepancy analysis on each of the four FISH break-apart assays (*MYC*, *BCL2*, *BCL6*, IGH). Because the first three assays were used for parameter optimization purposes, the following three figures 2.14, 2.19 and 2.21 are not intended as proper benchmarks of SOPHIA's performance.

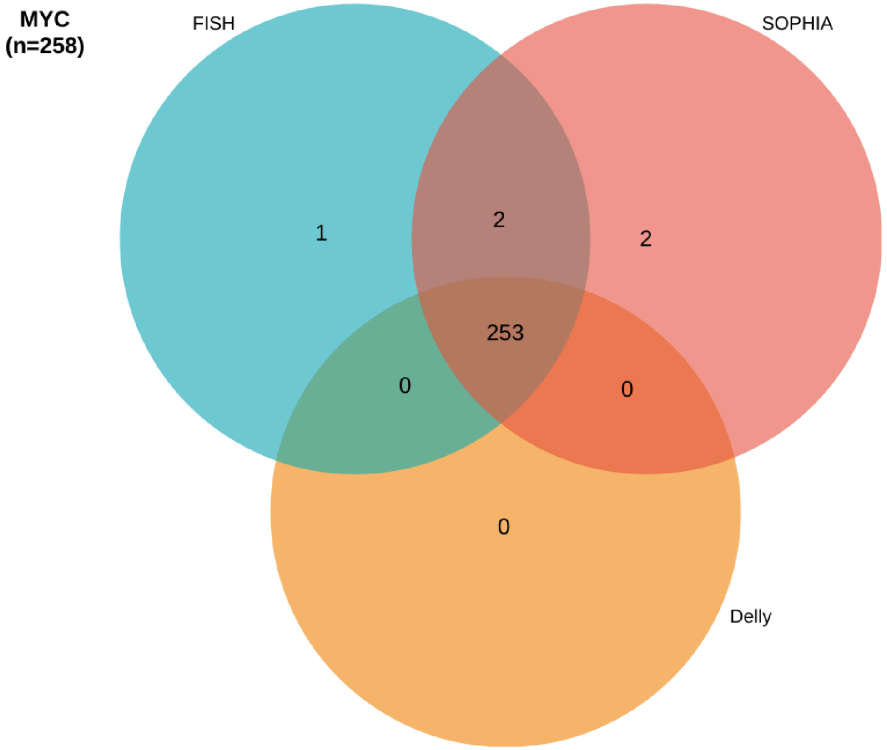

Figure 2.14: Discrepancy analysis of *MYC* SV detection between FISH-SOPHIA-Delly

In our analysis of *MYC* breaks, we observed two IGL-*MYC* rearrangements to be co-detected by FISH and SOPHIA and being missed by Delly (Figure 2.15). We investigated and established these false negatives to be due to lack of mappability on the IGL side Figure 2.16. This was our first indication that MAPQ-0 regions were of potential significance in the detection of biologically important rearrangements.

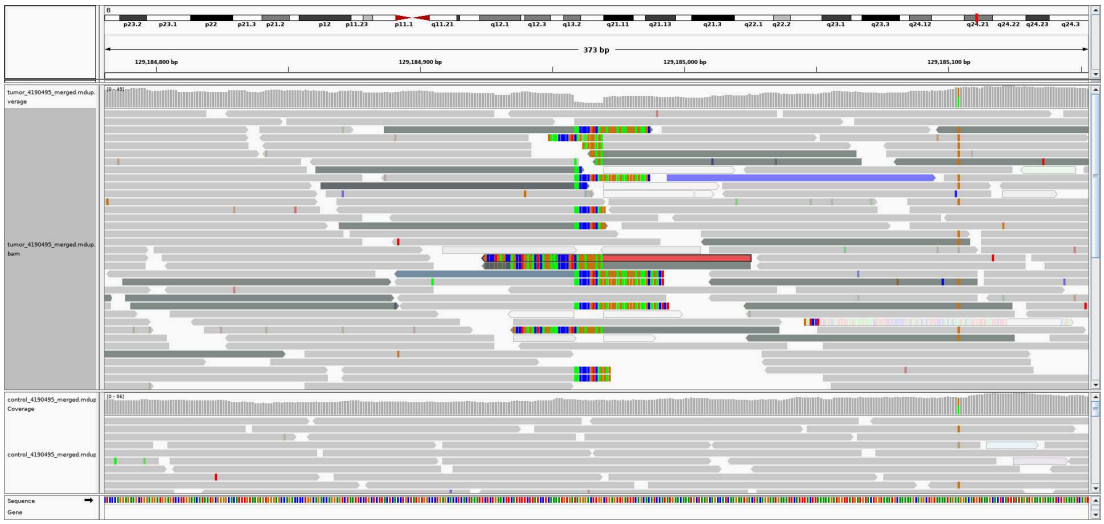

Figure 2.15: A discrepant call (detected by SOPHIA and FISH, missed by delly) for an IGL-*MYC* rearrangement, on the MYC-side breakpoint. The dark green reads have mates mapping to chr22, specifically to the IGL locus.

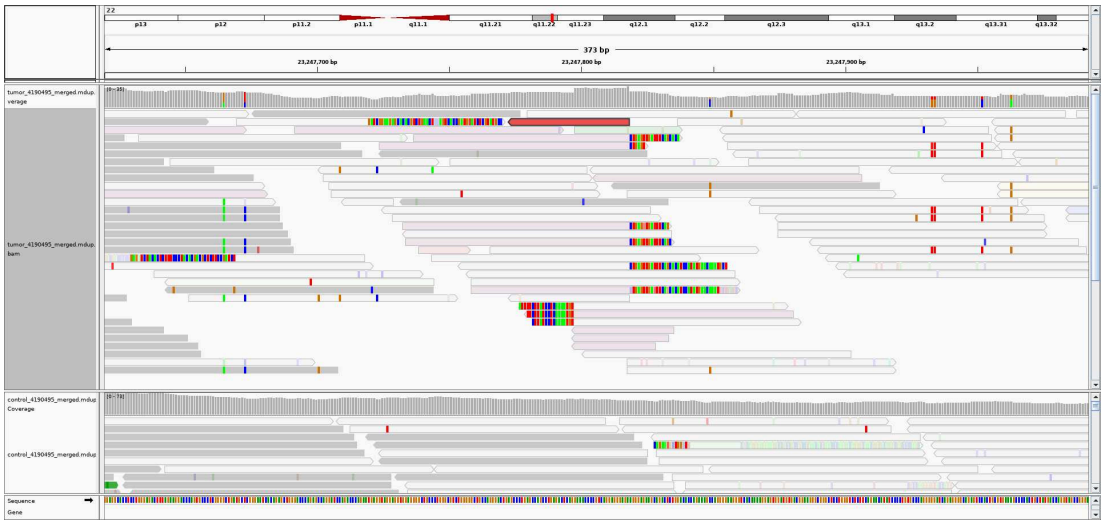

Figure 2.16: A discrepant call (detected by SOPHIA and FISH, missed by delly) for an IGL-*MYC* rearrangement, on the IGL-side breakpoint showing lack of mappability with predominantly hollow reads with 0 MAPQ.

Two further cases were only positive by SOPHIA. One of these cases is particularly interesting as it involves two sources of errors: First, we observed that there is a two-sided balanced somatic translocation on the *MYC* locus as expected (Figure 2.17), what was peculiar was the concomitant involvement of the chr15q11.2 locus, which harbours a number of inactive pseudogenes of the immunoglobulin heavy chain genes. Due to the homology between the canonical IGHV genes and their inactive pseudogene counterparts (IGH orphons), mapping can be unspecific and interfere with SV detection. Second, we observed that the IGH-side

of this rearrangement has the effects of a concomitant somatic hypermutation (Figure 2.18), which reduces the quality of mapping to this region. Though both can in different ways allow us to speculate for the issues Delly had in detecting this SV, we do not know why FISH failed to detect this rearrangement.

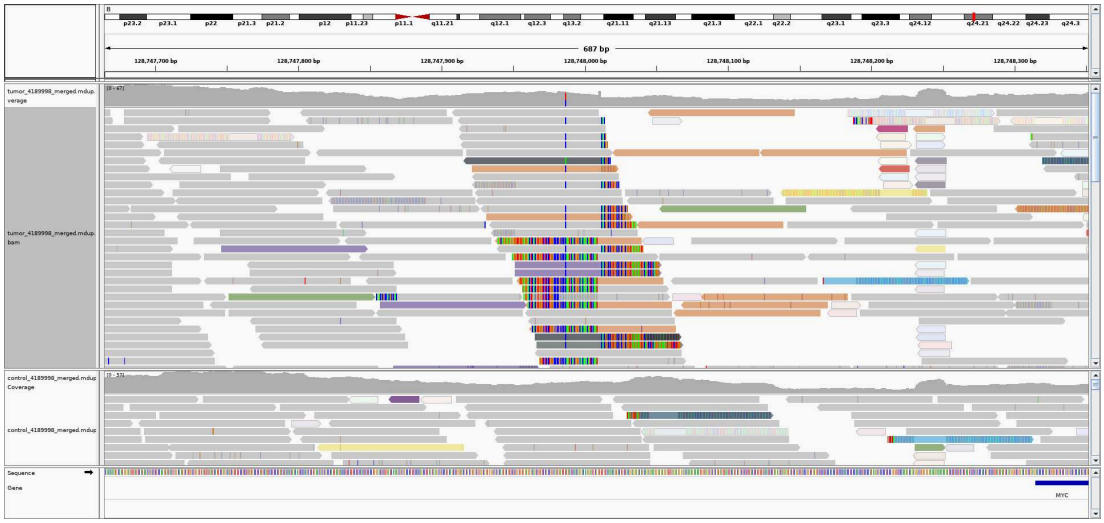

Figure 2.17: A discrepant call (detected by SOPHIA, missed by FISH and delly) for an IGH-*MYC* rearrangement, on the MYC-side breakpoint, showing concomitant mapping to the chr14q33.32 (IGH locus, orange reads) and chr15q11.2 loci harbouring IGH orphon genes (purple reads). The upper subfigure indicates the tumour sample, whereas the lower subfigure indicates the matched normal sample with no evidence for the rearrangement, indicating a somatic rearrangement.

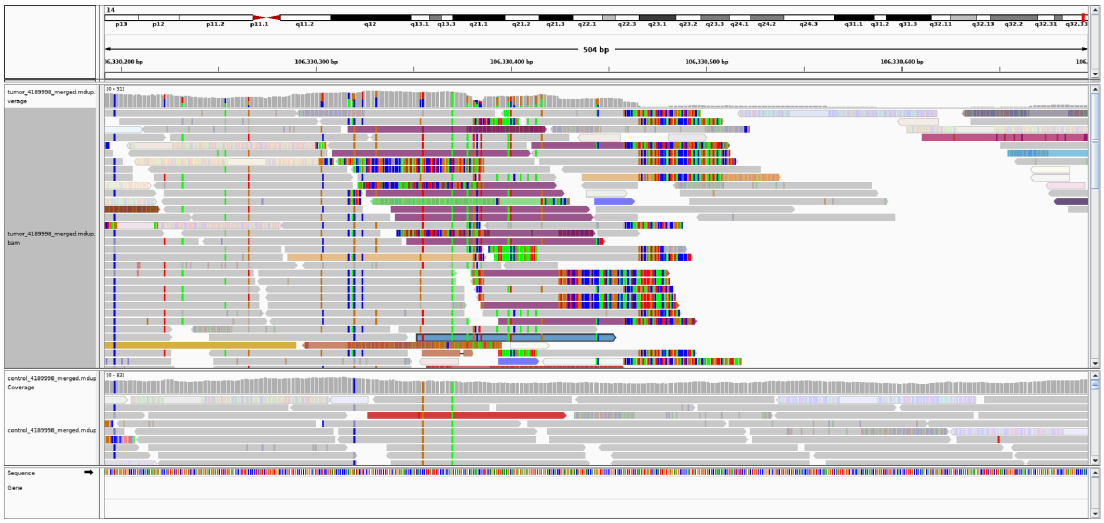

Figure 2.18: A discrepant call (detected by SOPHIA, missed by FISH and delly) for an IGH-*MYC* rearrangement, on the IGH-side breakpoint showing mate mapping to chr8q24.21 (*MYC*, purple reads) concomitant with somatic hypermutation (coloured bars on the upper coverage layer and individual reads showing base mismatches). The upper subfigure indicates the tumour sample, whereas the lower subfigure indicates the matched normal sample with no evidence for the rearrangement, indicating a somatic rearrangement.

The analysis of *BCL2* SV detection yielded two types of discrepancies (Figure 2.19): SVs co-detected by SOPHIA and Delly but not by FISH, and SVs detected only by FISH.

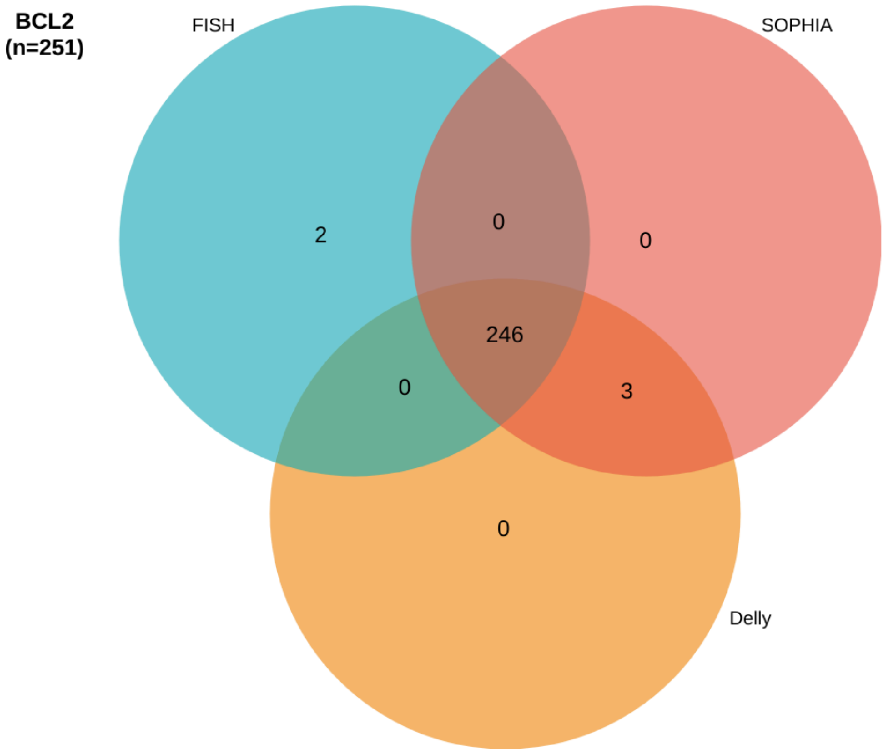

Figure 2.19: Discrepancy analysis of *BCL2* SV detection between FISH-SOPHIA-Delly

For one of the 3 cases detected by sequencing and confidently not detected by FISH, we observed a complex pattern (Figure 2.20): The left-side of the balanced translocation has no specific partner on chr14 (lack of orange discordant reads on the left side of the breakpoint). While the event is well-supported on the right-side of the breakpoint showing a clear mapping to the IGH locus, the left-side is unspecific, with only a short unspecific consensus split read overhang and no discordant mate information, hinting at the low complexity of the partner region. We cannot speculate if this played a part in the lack of detection by FISH.

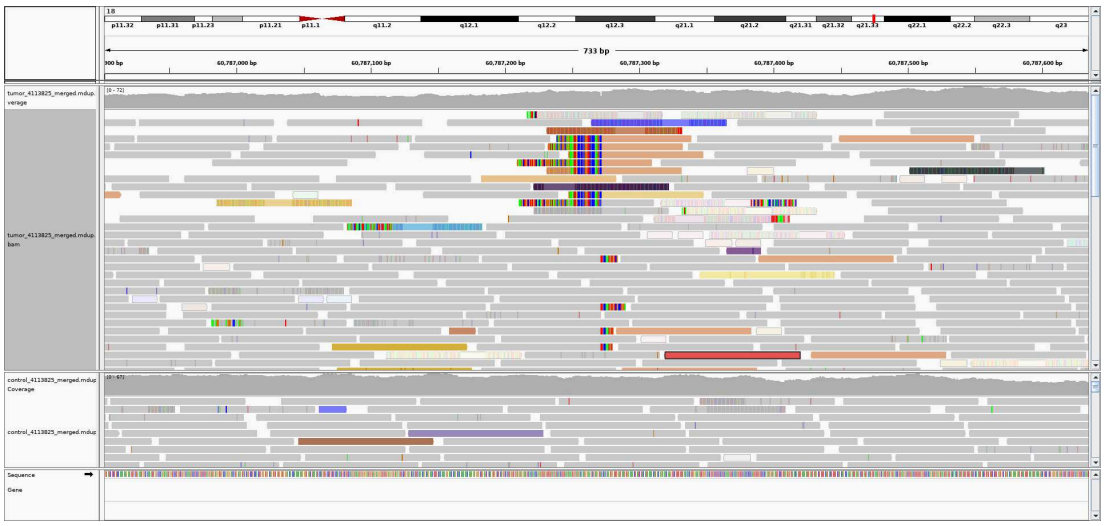

Figure 2.20: A discrepant call (detected by SOPHIA and delly, missed by FISH) for an IGH-*BCL2* rearrangement. The rearrangement is balanced and has both left and right-sided split reads, but the right-sided split reads are few, short, and do not have discordant mate support mapping to the IGH locus (orange reads). The upper subfigure indicates the tumour sample, whereas the lower subfigure indicates the matched normal sample with no evidence for the rearrangement, indicating a somatic rearrangement.

Next, we investigated *BCL6* breaks for discrepancies between the three approaches. We observed a diverse spectrum of discrepancies. 2/2 FISH-only calls had low tumour content (as estimated by ACEseq [309] using WGS data), again suggesting that FISH can be a more sensitive method because of its access to single-cell level information.

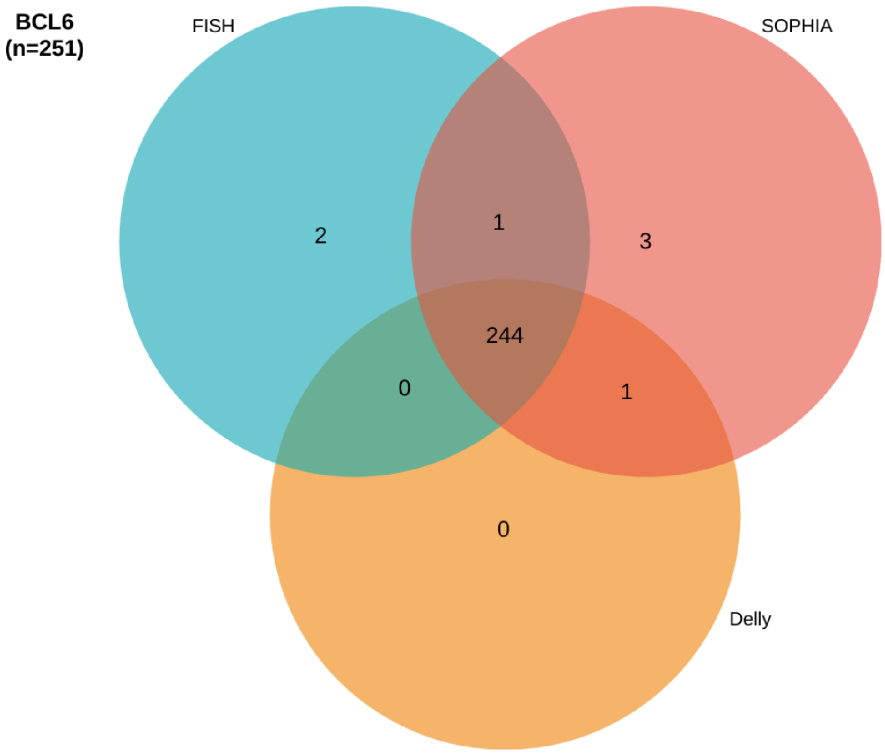

Figure 2.21: Discrepancy analysis of *BCL6* SV detection between FISH-SOPHIA-Delly

We observed a similar case to the previously discussed IGL-*MYC* case, which was this time also outside of the detection range of FISH (Figures 2.22, 2.23).

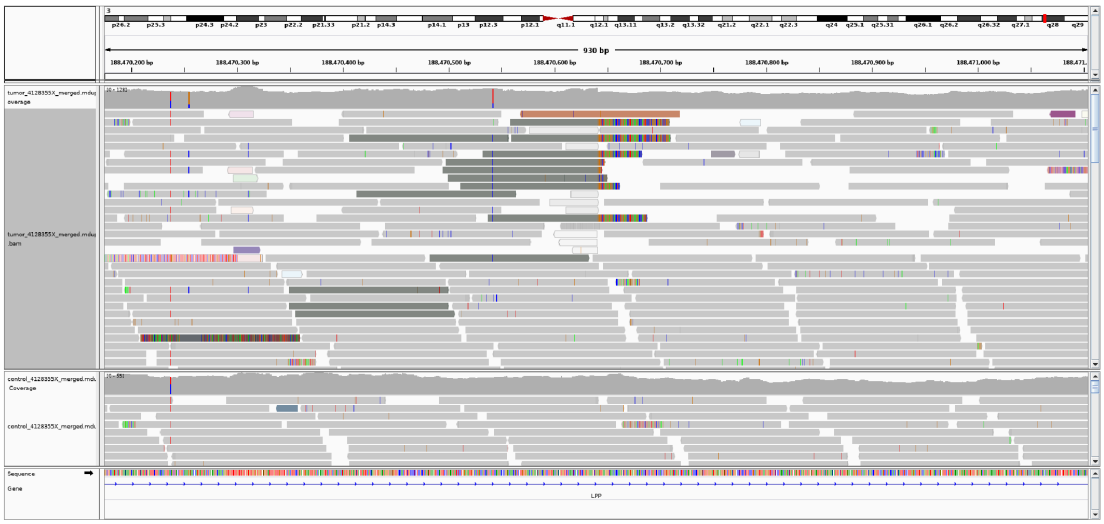

Figure 2.22: A discrepant call (detected by SOPHIA and FISH, missed by delly) for an IGL-*BCL6* rearrangement, on the *BCL6*-side breakpoint, with extensive discordant mate support mapping to the IGL locus (dark green reads). The upper subfigure indicates the tumour sample, whereas the lower subfigure indicates the matched normal sample with no evidence for the rearrangement, indicating a somatic rearrangement.

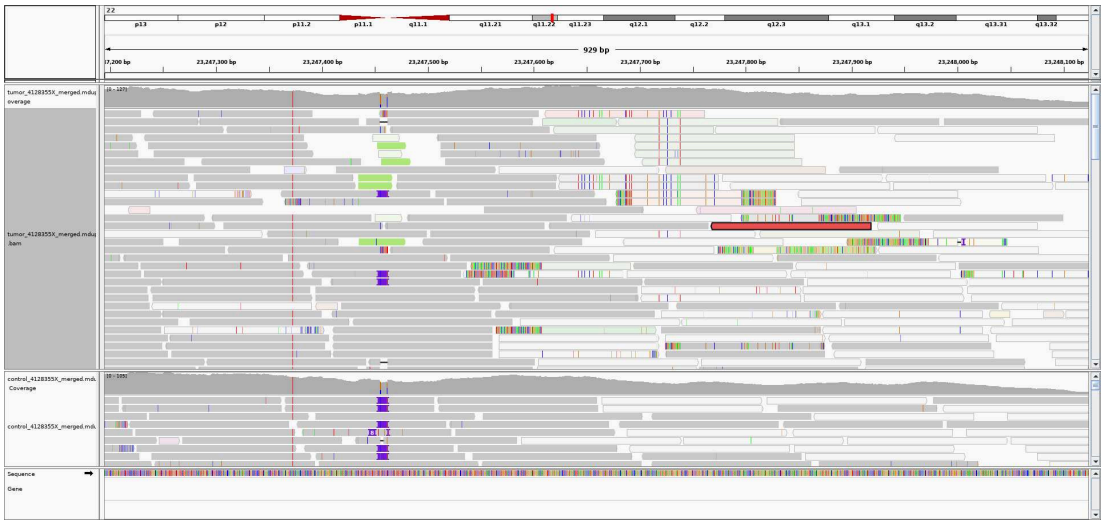

Figure 2.23: A discrepant call (detected by SOPHIA and FISH, missed by delly) for an IGL-*BCL6* rearrangement, on the IGL-side breakpoint showing lack of mappability (reads depicted as empty bars), still with discordant mate support mapping to the IGL locus (reads depicted as faded light green coloured empty bars). The upper subfigure indicates the tumour sample, whereas the lower subfigure indicates the matched normal sample with no evidence for the rearrangement, indicating a somatic rearrangement.

A second discrepant case only detected by SOPHIA was a rare instance of a T-cell Receptorα - locus to *BCL6* translocation that was missed both by FISH and Delly. As we are dealing with B-cell lymphoma, this finding was a surprise, but it was not entirely novel [458], [459]. Interestingly, TCRa locus which is intrinsically complex in a similar manner to the IG loci (with internal rearrangements as part of its normal function), did not pose the issue on this case, it was rather a GAn simple repeat that made the breakpoint on the *BCL6* locus poorly mappable. This result again reinforces the importance of considering SVs even when only one of the breakpoints is strongly supported.

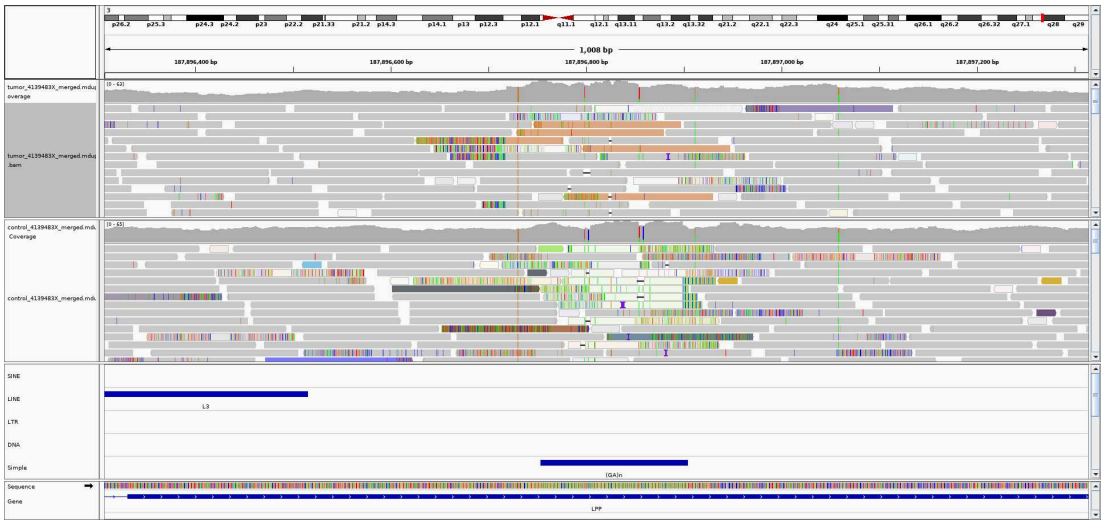

Figure 2.24: A discrepant call (detected by SOPHIA, missed by FISH and delly) for an TCR-*BCL6* rearrangement, on the *BCL6*-side breakpoint showing a (GA)n repeat interfering with proper mapping.

Figure 2.25: A discrepant call (detected by SOPHIA, missed by FISH and delly) for an TCR-*BCL6* rearrangement, on the TCR-side breakpoint showing a clean break

The fourth FISH assay, namely IGH breaks, were not used in SOPHIA parameter optimization. We observed a number of discrepant cases between the three assays, (Figure 2.26).

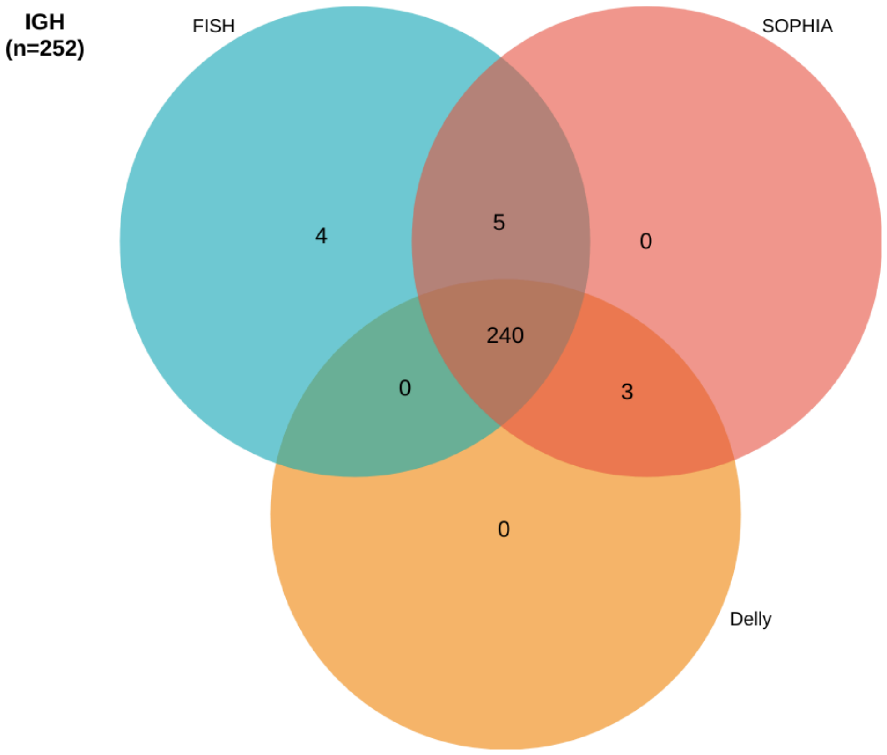

Figure 2.26: Discrepancy analysis of IGH SV detection between FISH-SOPHIA-Delly

3/5 cases co-detected by FISH and SOPHIA and missed by Delly were important oncogenes: 2 cases belonged to a IGH-*IRF4* rearrangement positive subtype previously presented in [460] (Figure 2.27). 1 case had an IGH-*PRDM6* rearrangement, which is novel for the entity, but is a recognized oncogene in other settings [211]. On the other hand, 3/4 of the FISH+/SOPHIA-/Delly- cases were marked for low tumour cell content of the starting specimen, suggesting a possible explanation for the calls missed by SOPHIA. One further interesting observation was the three cases where sequencing based SV detection succeeded and FISH failed to yield a positive result. One of these cases was a putative insertions of *BCL2 inside* the IGH locus rather than a canonical rearrangement. The other two were regarded as FISH false positives with regards to IGH breaks in the ICGC MMML project.

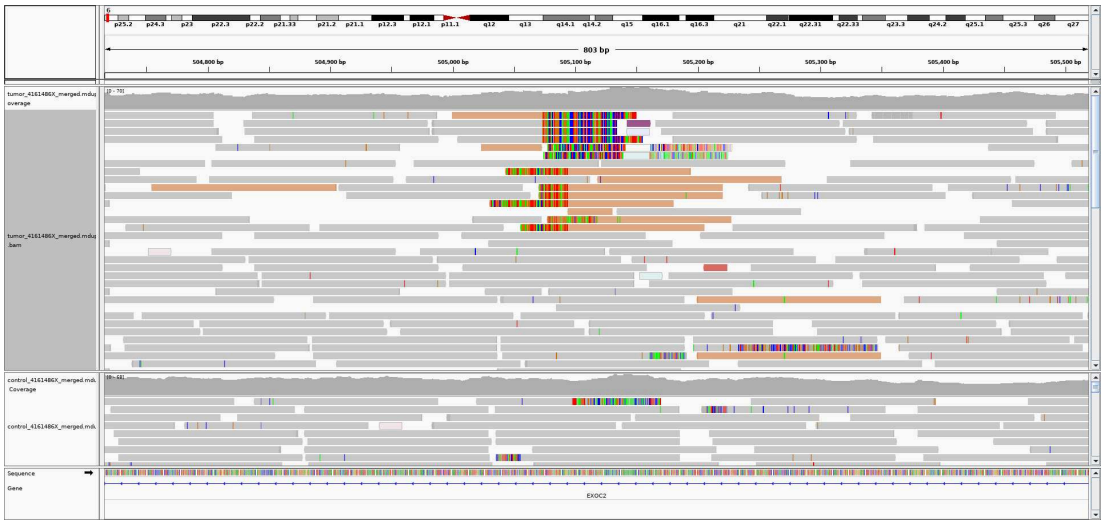

Figure 2.27: A discrepant call (detected by SOPHIA and FISH, missed by delly) for an IGH-*IRF4* rearrangement, with strong discordant mate support mapping to the IGH locus (orange reads) for a two-sided balanced rearrangement. The upper subfigure indicates the tumour sample, whereas the lower subfigure indicates the matched normal sample with no evidence for the rearrangement, indicating a somatic rearrangement.

2.3.3 SOPHIA Detects Hallmark Structural Variants with High Sensitivity Across Cancer Types Expression Data

Next, we benchmarked SOPHIA across a diverse set of human cancers with publicly available data from the The Cancer Genome Atlas (TCGA) Consortium with available whole genome sequencing and matched RNA Sequencing data and published DKFZ projects. We looked for ”expected” structural variants based on differentially expressed cancer hallmark genes with known dysregulation mechanisms.

While these hallmarks are by nature not necessarily difficult to detect unlike the immunoglobulin translocations that were used in the parameter training stage, they nevertheless provide an important benchmarking opportunity on clinically relevant events from real datasets and an SV detection algorithm is expected to show high sensitivity in this analysis.

Study	Gene	Expected Type	#Detected (by TAD SV)	#Expected (by expr.)	Notes
BLCA-US	CDKN2A	Deletion	13	13	
BLCA-US	CDKN2B	Deletion	12	14	(1)
BLCA-US	EGFR	Amplification	2	2	
BRCA-US	ERBB2	Amplification	20	20	
BRCA-US	PTEN	Deletion	15	15	
BRCA-US	IGF2	Activation	2	2	
GBM-US	EGFR	Amplification	18	19	
GBM-US	CDKN2A	Deletion	18	18	
GBM-US	CDKN2B	Deletion	17	17	
GBM-US	CDK4	Amplification	8	8	
GBM-US	MDM2	Amplification	4	5	(2)
HNSC-US	FADD	Amplification	15	15	
HNSC-US	CDKN2A	Deletion	10	10	
HNSC-US	CDKN2B	Deletion	9	9	
KICH-US	TERT	Activation	5	5	
LUSC-US	CDKN2A	Deletion	9	11	
LUSC-US	CDKN2B	Deletion	10	13	
LUSC-US	MYEOV	Amplification	9	12	(3)
MB-Group3-DKFZ	GFI1B	Activation	6	6	
MB-Group4-DKFZ	GFI1B	Activation	3	3	
MB-Group4-DKFZ	MYCN	Amplification	9	9	
MB-Group4-DKFZ	PRDM6	Activation	12	14	
NB-GPOH	MYCN	Amplification	53	53	
NB(MYCNwt)-GPOH	TERT	Activation	21	27	(4)
SARC-US	CDK4	Amplification	18	18	
SARC-US	MDM2	Amplification	20	20	
SARC-US	CDKN2A	Deletion	9	10	
SARC-US	CDKN2B	Deletion	8	9	
SKCM-US	CDKN2A	Deletion	8	10	(5)
STAD-US	IGF2	Activation	3	3	
UCEC-US	ERBB2	Amplification	3	3	
COADREAD-US	IGF2	Activation	9	10	(6)
NB-GPOH	FOXR1	Activation	4	4	(7)

Figure 2.28: A comprehensive SV TAD assignment and gene expression bimodality based benchmark of SOPHIA's sensitivity. In TCGA and DKFZ studies of diverse cancer types, hallmark oncogenes or tumour suppressors known to be dysregulated by somatic structural variants are investigated for gene expression bimodality, and a corresponding SV detection by SOPHIA, up to 1 TAD away from the gene of interest. The expected number of cases are defined by the counts of cases belonging to the higher or lower of the modes in the bimodal gene expression distribution, depending on the expected direction of dysregulation.

Notes: (1) 1/2 ”negative” cases also shows no copy number differences, (2) The ”negative” case also shows no copy number differences, (3) The ”negative” cases also show no copy number differences, (4) It is known that a hitherto unknown mechanism of *TERT* activation and promoter mutations exists apart from rearrangements [212], (5) The ”negative” cases also show no copy number differences, (6) 2 positive cases were used in parameter optimization, (7) One positive case was used in parameter optimization

Overall SOPHIA shows an excellent sensitivity for detection of oncogenic hallmark structural variants. The only recurrent source of false negatives were putative *CDKN2A/B* deletions. Following these benchmarks which were base on parameters trained on FISH and M-FISH data, we only did further revisions based on two oncogenes, namely undetected *IGF2* rearrangements in 2 colorectal cancer cases, and one undetected novel interchromosomal *FOXR1* rearrangement in neuroblastoma. These secondary optimizations did not further affect results on hallmark structural variants, but likely improved the overall sensitivity of SOPHIA.

2.3.3.1 Important Structural Variants such as Tandem Duplications upstream of IGF2 can Have a Breakpoint on Repetitive Regions

Tandem duplications near the *IGF2* locus have been reported [100] to cause an increased activation of the *IGF2* oncogene, suggesting it to be a hallmark structural variant in a pan-cancer setting.

We observed in the TCGA colorectal cancer cohort three cases with increased *IGF2* expression and no nearby rearrangements. In the absence of a known secondary activation mechanism, we assumed these observations to be putative false negatives. We observed that a hotspot site for the first of the breakpoints of the duplication involving *IGF2* is on a (TGGA)_n simple repeat and that this leads to two types of issues: i) the site on the repeat sequence is frequently encountered as a common artefact breakpoint in the background breakpoint database (Figure 2.29), ii) the partner site downstream of *IGF2* suffers from a mate read dispersion where multiple distant locations on the (TGGA)_n simple repeat are proposed as the partner breakpoint and do not form a coherent structural variant with consistent support (Figure 2.30).

Figure 2.29: The left-hand side of the *IGF2* locus duplication on the (TGGA)_n simple repeat, donor TCGA-AG-3593 of the TCGA-READ project: the upper subfigure indicates the tumour sample with green reads indicating discordant reads supporting the duplication, whereas the middle subfigure indicates the matched normal sample with no evidence for the rearrangement, indicating a somatic rearrangement. The lowermost subfigure shows multiple (TGGA)_n repeats covering the region. A precise breakpoint yielding split reads is not mapped by the aligner.

Figure 2.30: The right-hand side of the duplication on the partner site downstream of *IGF2*, donor TCGA-AG-3593 of the TCGA-READ project: the lowermost subfigure shows a lack of repeats and the partner breakpoint to be located on the *ASCL2* gene with a precise breakpoint yielding split reads.

We could address the first issue by relaxing the background breakpoint database hits thresholds when the partner site is clearly somatic and shows strong evidence. However, we could

not fully address the second issue, for rare cases where the multiple solutions are all weakly supported and distant from each other.

The representative figure is from the case TCGA-AG-3593 in the TCGA Rectal Cancer Project (READ-US). The given structural variant has been misclassified as low-quality with score 2 by SOPHIA even after further tuning, and constitutes the only one of the 15 high *IGF2* expressor cases in the entire TCGA (breast cancer, colorectal cancer, stomach cancer) pan-cancer WGS cohort where an *IGF2* duplication has been missed by SOPHIA, following a parameter reoptimization taking into account the observations made on *IGF2* described here.

2.3.3.2 A novel interchromosomal FOXR1 activating translocation in Neuroblastoma maps nonspecifically to multiple partner sites

Intrachromosomal rearrangements have previously been shown to activate the *FOXR1* oncogene [461] as sole known driver in a small subset of neuroblastoma cases, and *FOXR1* can thus be considered as a hallmark rearrangement of neuroblastoma. In our larger German Paediatric Oncology and Hematology (GPOH) Neuroblastoma cohort, we observed that one (out of 4) FOXR1 high expressor cases had a somatic interchromosomal rearrangement of *FOXR1* t(11,17)(q23.3;p11.2) with unspecific mapping on chr17p11.2: We identified two identical regions flanking the gene *GRAPL* where precise mapping with the current short read sequencing technology is not possible. The two possible breakpoints that could be considered to activate *FOXR1* were 17:19015973 at 17:19093582. Interestingly, these two regions of full homology were not annotated as repeats in RepeatMasker. As the two candidate regions are identical, mapping quality of every (high base-quality) read mapping to these regions is 0. Due to this important and representative example, we decided to support such regions in SOPHIA, with counterbalancing caveats. Briefly, we impose the conditions that at least one of the two paired breakpoints must not be in a MAPQ0 region, and if one of the two breakpoints is in a MAPQ0 region, the supporting read count must be higher than the normally used thresholds.

This representative case GPOH-NB-13264 that led to the described observations (Figures 2.31, 2.32 and 2.33) and allowed us to improve SOPHIA's detection sensitivity on MAPQ0 regions.

Figure 2.31: The *FOXR1*(chr11)-side of the *FOXR1* rearrangement. The purple reads indicate reads whose mates map to chr17. With large numbers of supporting discordant reads, availability of split reads and a high overall sequencing quality in this locus, this breakpoint has the signatures of a true somatic SV candidate.

Figure 2.32: The first of the two possible solutions for the chr17-side of the *FOXR1* rearrangement. This region is entirely a MAPQ0 region where mapping is not unique (likely same sequence as the second alternative breakpoint site). Discordant read pairs whose mates map to chr11 on the *FOXR1* locus are visible along with a precise breakpoint generating clean split reads.

Figure 2.33: The second of the two possible solutions for the chr17-side of the *FOXR1* rearrangement. This region is also entirely a MAPQ0 region where mapping is not unique (likely same sequence as the first alternative breakpoint site). Discordant read pairs whose mates map to chr11 on the *FOXR1* locus are visible along with a precise breakpoint generating clean split reads, albeit with less support than the first of the alternative solutions.

MAPQ0 regions are often discarded by structural variant detection algorithms as a source of systematic artefacts, as is also suggested by the results presented in Section 2.3.2. We tested the same case with SvABA and confirmed that neither of the two solutions were proposed by their local assembly based approach. This reinforced our opinion that it is justified to study MAPQ-0 regions for the purpose of detecting structural variants, though additional care is warranted.

2.3.4 SOPHIA Structural Variant Detection Speed

Due to the single-pass and minimally buffered evidence collection approach of SOPHIA, it has a number of significant performance advantages. First, the linear single-pass approach is I/O friendly and allows a fast and efficient parsing of the decompressed alignment in a data stream. Also, flushing of collected evidence on a per-breakpoint basis rather than a per-SV basis minimizes RAM usage for the first, and longest stage of the SOPHIA operation.

We ran a benchmark for a large number of SOPHIA runs across diverse projects, the two different sequencing technologies, tumours and controls, going beyond the usual standards provided in SV detection algorithm publications. Our aim was to show that SOPHIA runtimes vary in a robust manner in a narrow range, not influenced by sample type, quality or technology. As the SOPHIA workflow has two major parts (with separate executables), namely breakpoint evidence collection (Section 2.2.3) and breakpoint pairing and filtering (Sections 2.2.5 and 2.2.6), we ran benchmarks for these two parts separately. For both parts, we measured CPU time rather than absolute runtimes (wall time), because it is not sensitive to fluctuations in computing cluster data I/O performance. SOPHIA uses two cores, expected real runtimes are roughly a half of the benchmarked CPU time durations in the absence of technical I/O

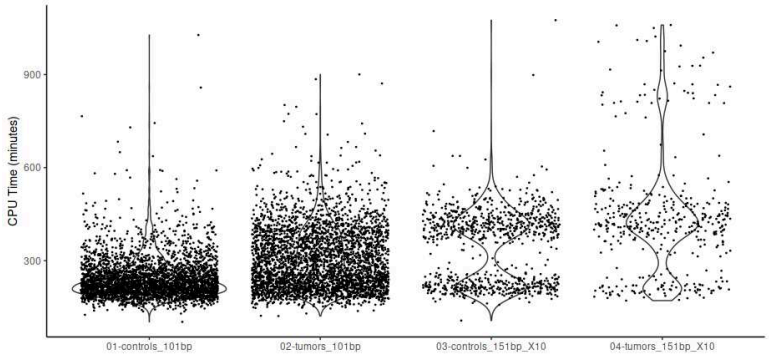

Figure 2.34: Benchmark of SOPHIA breakpoint evidence collection speed across sample types and sequencing technologies. This analysis corresponds to the "SOPHIA breakpoint extraction" step presented in Figure 2.1, which is run once per the tumour and control bam file each. Control and tumour alignments as well as those from different read length sequencing technologies / alignment workflows are analysed in separate categories.

bottlenecks.

For the breakpoint evidence selection stage, the results showed a strikingly narrow variance for all analysed groups (Figure 2.34). The bimodalities observed in the Hiseq X-Ten system were due to the 1-lane vs 2-lane choice made by different DKFZ projects depending on their needs and budgets.The lowest mode represents the $35x$ coverage level attained by single lane operation while the second represents the $70x$ coverage level attained by double lane operation. The highest runtimes were recorded for tumour samples from the BMBF eMED SYS-GLIO project, which used higher sequencing depths for tumour evolution modelling purposes as described in [462]. The much larger 101bp datasets were harder to interpret due to the diversity of countries and sequencing centres that provided cases. Overall, the benchmarked runtimes suggest that breakpoint evidence collection performance in SOPHIA is largely dependent on sequencing depth and not on other factors such as sample type, read length (within the constraint of 101bps vs 151bps) or sample quality. For the breakpoint evidence collection stage, the RAM usage is held at a 2GB via the tool called mbuffer. However, in exceptional cases involving large viral loads such as in gastric cancer and the Epstein-Barr Virus, RAM consumption can temporarily spike during processing of viral integration sites or the EBV chromosome itself. Such cases are exceptional, and it can safely be assumed that a normal SOPHIA breakpoint evidence collection run consistently consumes 2GB RAM per sample.

For the breakpoint pairing and filtering stage, we measured both runtime and memory usage characteristics for both paired and no-control operation, with a single core operation where the real runtime is approximately equal to the CPU time. The pairing and filtering stage is a very fast process, with few exceptions (Figures 2.35 and 2.36): In our measurement of 5779 SOPHIA SV pairing and filtering and analyses (of which 5415 corresponded to paired analysis of tumours and matched normals), we found that only 52 exceeded 10 minutes of operation, of which only 16 exceeded 20 minutes, with the highest recorded runtime at 133 minutes. The median runtime for no-control operation was 1.03 minutes, whereas the median runtime for paired analysis was 2 minutes.

We measured peak memory usage in a similar manner to speed (Figures 2.37, 2.38). In

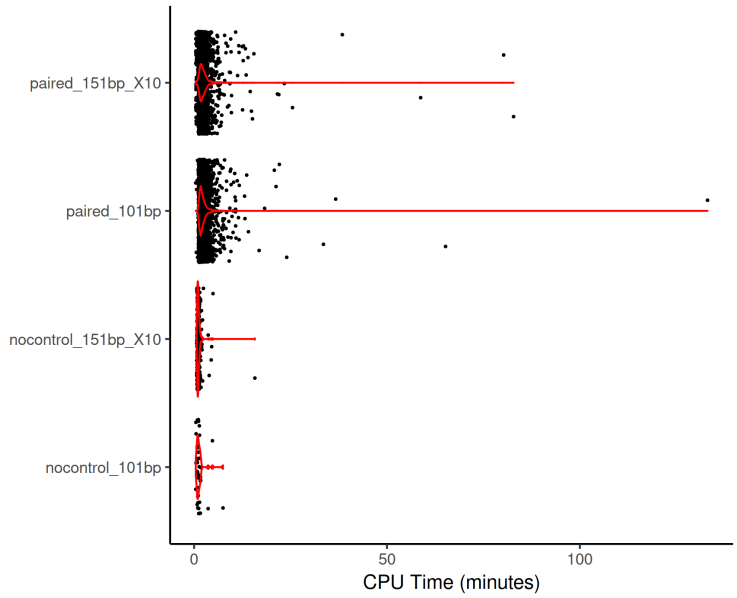

Figure 2.35: Benchmark of SOPHIA breakpoint pairing and filtering speed across analysis types and sequencing technologies. This analysis corresponds to the ”SOPHIA breakpoint matching” step in Figure 2.1, which is run once per paired or no-control analysis on the results on the ”SOPHIA breakpoint extraction speed”. Paired and no-control workflows as well as runs from different read length sequencing technologies / alignment workflows are analysed in separate categories.

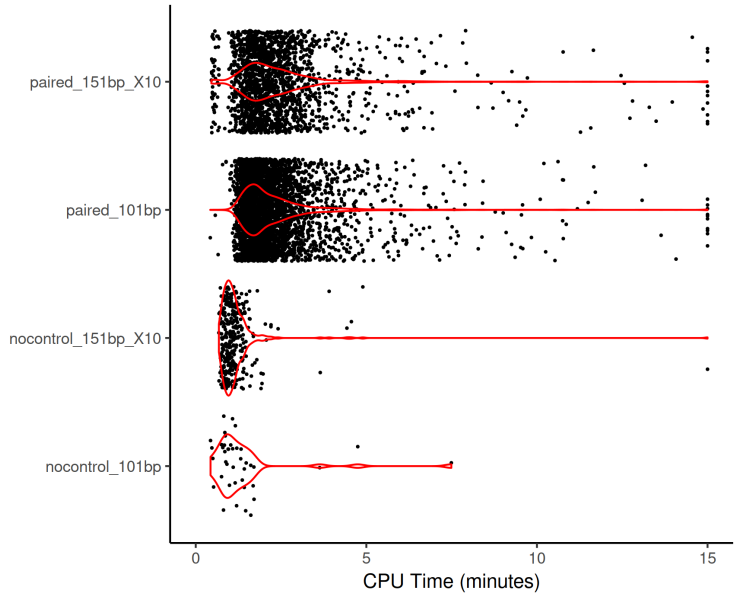

Figure 2.36: Benchmark of SOPHIA breakpoint pairing and filtering speed across analysis types and sequencing technologies, limited to 15 minutes for better visibility

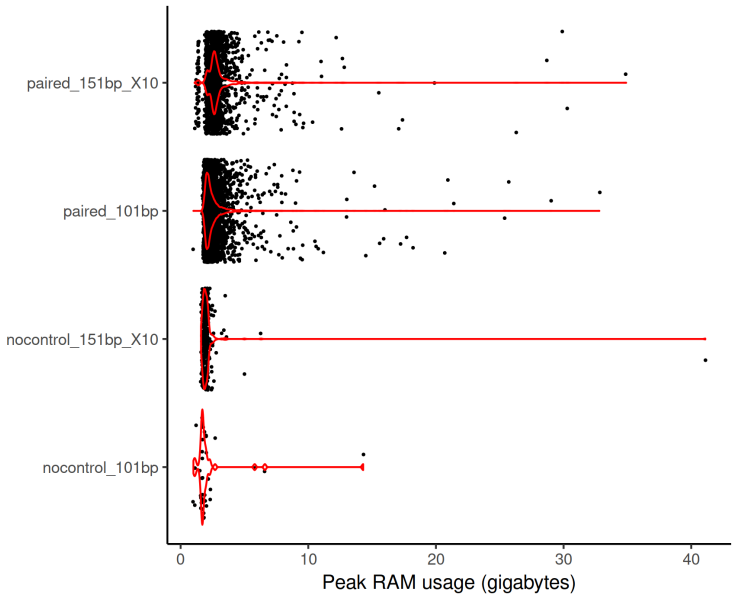

Figure 2.37: Benchmark of SOPHIA breakpoint pairing and filtering peak RAM usage across analysis types and sequencing technologies

our measurement of 5779 SOPHIA analyses (5415 paired), we found that only 68 exceeded 8 Gigabytes of peak memory usage, of which only 19 exceeded 16 Gigabytes, with the highest recorded memory usage at 41.1 Gigabytes. The median peak memory usage for no-control operation was 1.9 Gigabytes, whereas the median peak memory usage for paired analysis was 5 Gigabytes.

Finally we investigated the relationship between RAM usage and speed (Figure 2.39). We observed that high RAM usage does not generally follow extremely high runtimes.

Poor performance characteristics are often caused by samples that are later excluded from studies due to poor material quality. SOPHIA deals with such outlier cases without crashing or needing extremely long runtimes that can take up to days or weeks with other algorithms. Nevertheless, further reduction of peak RAM usage would be helpful in a cluster environment and should be a future development goal.

2.4 Discussion

2.4.1 Advantages and Novelties of SOPHIA

SOPHIA, to the best of our knowledge, is the first structural variant detection algorithm that is based on an expert model combining a rich and diverse set of training data sources such as FISH, RNA-Seq and known hallmark structural variants. Thanks to its powerful filtering features, we managed to explore complex regions and complex rearrangements without sacrifice of overall specificity. Also thanks to our single-sided evidence collection, we managed to keep RAM requirements to a minimum while maintaining a linear, single-pass operation for the majority of the workflow, yielding a fast, lightweight and effective tool for structural variant detection.

In early stages of SOPHIA's development we started from a split-read only approach in the

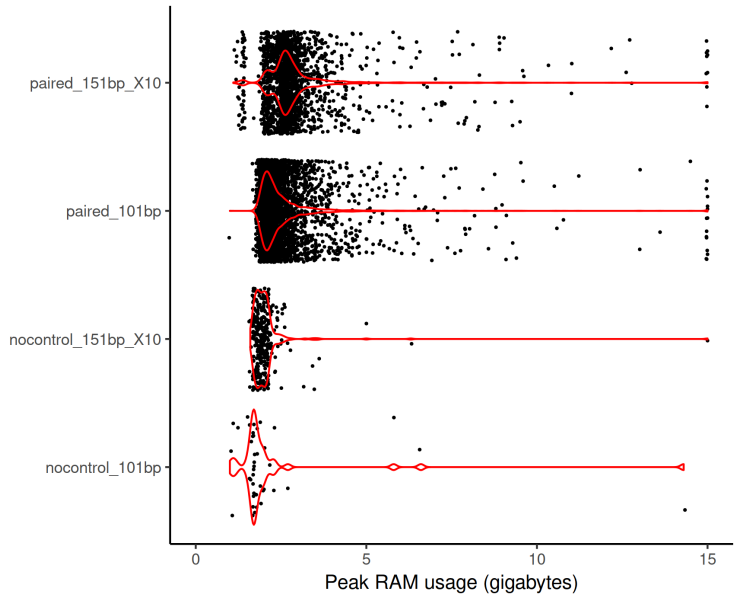

Figure 2.38: Benchmark of SOPHIA breakpoint pairing and filtering peak RAM usage across analysis types and sequencing technologies, clamped to 15 Gigabytes for better visibility

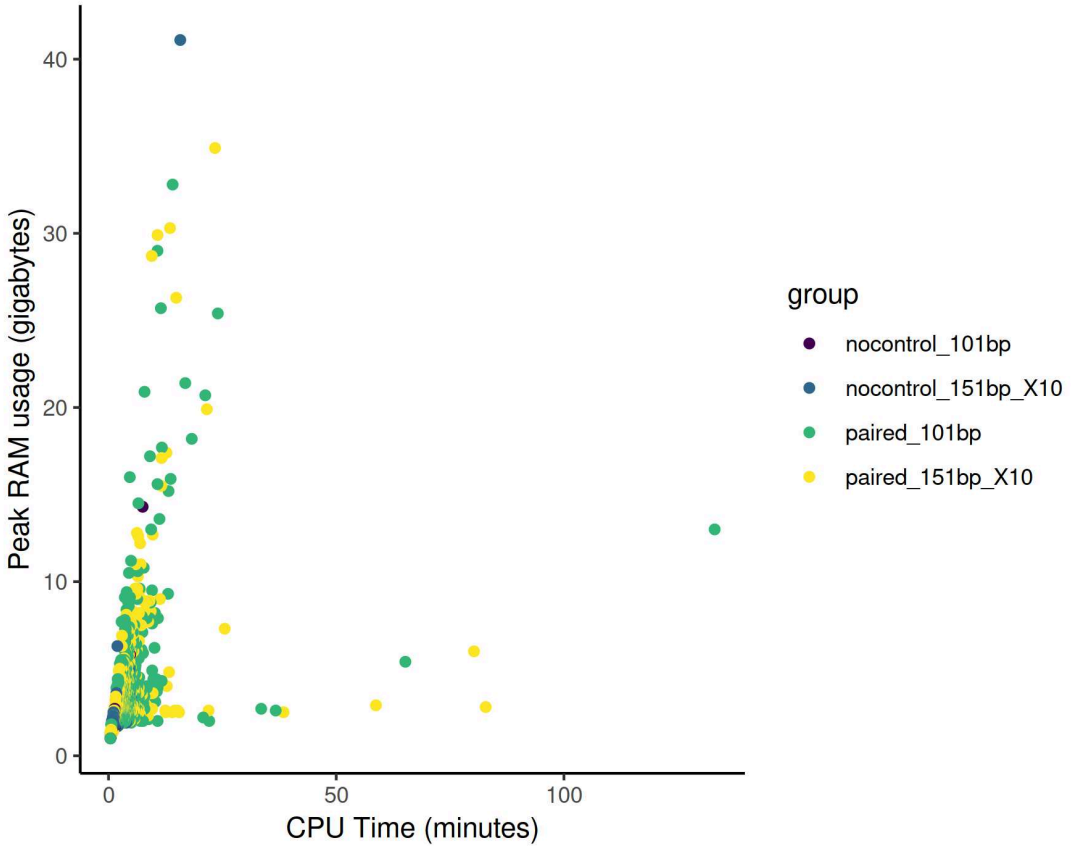

Figure 2.39: Benchmark of SOPHIA breakpoint pairing and filtering peak RAM usage vs speed

first proof of concept prototypes of the linear, single-pass operation. The discordant mate information was added in later versions, supporting SVs first called by split read information. The first version of SOPHIA used in a publication with my joint-first author contribution used this split read as primary, discordant mate information as secondary level evidence approach [463]. Our later experiences with a more extensive evaluation of ICGC-MMML FISH and NB69 cell line M-FISH results prompted us to improve our strategy where breakpoint detection can start from either split read evidence or discordant mate based evidence, improving the sensitivity of SOPHIA. This improvement was used in [464] (the follow-up article to [463]) along with other recent publications [435] and the case study [465], where SOPHIA was used to detect a cryptic IGH-*MYC* translocation where *MYC* is inserted into the IGH locus in a Burkitt leukaemia case which was not detectable by FISH. Our experiences in the evolution of the SOPHIA algorithm mimic the evolution in the field of SV detection where the first tools were either discordant-mate-based like BreakDancer [466] or split-read-based like CREST [366]. Delly became a standard tool in SV calling by combining these two approaches [308] and similar to the SOPHIA's population database-based filtering delly uses data from the 1000 genomes project in its filtering [132], though these specific filters were not openly published. During the development of SOPHIA, we generated a breakpoint repository including real and artefactual breakpoints along with their commonly detected partners obtained from normal tissue of donors in cancer genomics studies. This database, which will be released with SOPHIA, could potentially support researchers using other tools for SV detection because the sites of common artefacts and germline variation is a generally useful resource. As part of the effort of the development of this database, we compared sequencing data from two generations of sequencers using the same underlying technology. Our results show slight differences between the behaviour of these two generations of sequencers with respect to alignment performance around repeat regions: while the X-Ten system can resolve more breakpoints on/around the repeat families *simple repeats*, *satellite repeats* and *low complexity repeats*, this is at the expense of lower quality breakpoints with regards to base and mapping quality of reads. This suggests that more breakpoints can be reached with the slightly longer reads produced by the X-Ten sequencer, while even longer reads are likely to be necessary to get clean signals from these regions.

The current trend in SV calling approaches seems to be pointing at local assembly where novoBreak [425] and SvABA [426] both successfully reported results with this approach. We have not been able to test both approaches simultaneously, but SOPHIA's breakpoint evidence collection strategy could be described as a half-local-assembly with each breakpoint being processed separately, and later unified as an SV candidate finally being evaluated with SOPHIA's expert model-based filters. Evaluating these algorithms in a comprehensive benchmark would allow us to evaluate if the added computational and memory costs of the full local assembly process is an acceptable compromise with regards to possible improvements in sensitivity or specificity.

We showed SOPHIA's runtime and memory usage performance across a large cohort, with different sample and sequencing quality characteristics and sequencing depths, where our results indicated SOPHIA to be robust against sample quality control issues. We have discussed

here typical quality control issues such as contamination of normal material and low proper pairing statistics, suggesting strategies to deal with these common issues, an important practical concern which is not discussed in the discussed SV algorithms' articles. Overall, SOPHIA is a successful solution for structural variant detection as of 2019, quickly, robustly and reproducibly running across tens of cancer genome sequencing projects at the DKFZ with the DKFZ SV workflow `https://github.com/DKFZ-ODCF/SophiaWorkflow`. It has successfully ended the DKFZ's reliance on external tools and collaborations for the analysis of structural variants and already allowed us to present some novel findings across different cancer types in Chapter 4.

In concluding our work on SOPHIA, we will discuss potential avenues for improvement.

2.4.2 Shortcomings of SOPHIA and Suggestions for Potential Improvements

2.4.2.1 Lack of a formal specificity analysis

Our current knowledge and assessment of the specificity performance of SOPHIA are based on our and our collaborators' anecdotal institutional experiences with other algorithms such as delly and CREST. This is a shortcoming of this presented work we would like to improve. Based on a number of discussions, the following options emerged for a specificity analysis as preparation for the publication of the SOPHIA algorithm:

1. Using simulated tumour-normal pairs starting from available cell line sequencing data: A recent work (Nov. 2018) [467] presented a comprehensive simulation based benchmark of a broad selection of SV calling algorithms. Using the same strategy could be a feasible goal to assess both the specificity and the sensitivity of SOPHIA.

2. Using the Pan-Cancer Analysis of Whole Genomes Project's consensus SV dataset: The PCAWG consortium generated a consensus dataset of somatic SVs from 2693 adult cancer cases using four different SV callers BRASS, delly, dRanger and SvABA [284]. This dataset, when released, could be a very valuable tool to again assess both the specificity and the sensitivity of SOPHIA. This approach could also be an opportunity to assess the feasibility of running a consensus-based SV calling approach in the DKFZ's sequencing data analysis workflow, similar to the PCAWG consortium's strategy. To this end, it would be important to identify which tool would best complement SOPHIA by offering an expansion of true results with the least amount of redundant overlap possible.

2.4.2.2 Classes of structural variants missed by SOPHIA

In its current design, SOPHIA cannot detect structural variants involving unmapped reads. This might reduce sensitivity in identifying exact breakpoints of viral integration sites where one breakpoint correctly maps to the human genome, whereas the second breakpoint doesn't. We did not specifically address this question because we did not have a specific project where these issues posed a detectable problem motivating an immediate improvement of this aspect in SOPHIA. We would nevertheless like to address this during the transition of SOPHIA to the GRCh38 reference human genome.

SOPHIA can also not detect structural variants where both breakpoints of the variant fall on repetitive regions and consequently common breakpoints in the breakpoint database used for filtering. As described in Section 2.3.3 on the example of IGF2, strong and reliable evidence from one of the two breakpoints can be used to overrule the negative effect of strong database support for the other but it is possible that structural variants can be located in repetitive and unmappable regions on both ends. SOPHIA currently does not have a solution for such rare cases. The statement that such cases are rare are based on our extensive studies of disease hallmark variants across close to a hundred projects where we were unable to determine a systematically missed class/family of structural variants of this type.

SOPHIA also cannot detect structural variants where BWA-mem cannot propose either a split read or discordant mate based supplementary alignment as a candidate variant. This is sometimes the case with medium-sized SVs where the rearranged sequence also contains other small mutations leading to too many differences in the short split read sequence for BWA-mem to align it accurately. In such situations, SOPHIA would have no information to start from.

Finally, due to the usage of a population-based filtering approach, SOPHIA is not an appropriate tool for studying common germline structural variation. Therefore, its scope is focused on cancer genomics data analysis of (ideally) whole genome sequencing data where the focus is to find rare germline structural variation of somatic structural variants from mid-sized (roughly 20bps-1000bps) to interchromosomal. As this is the defined design scope of SOPHIA, and this covers our main use cases in cancer genomics projects, this is strictly speaking not a shortcoming.

2.4.2.3 Shortcomings of the current breakpoint database

As described in Section 2.2.5, we observed cases where the patient's ethnicity posed issues in the performance of the breakpoint database-based filtering. While such cases are relatively easy to detect by a prevalence of large numbers of germline structural variants, especially interchromosomal variants as manifestations of transposons, they are still useful to underline a weakness of this breakpoint database based filtering approach. However, as SOPHIA's performance is excellent in many other regards such as speed, sensitivity and specificity, efforts should probably focus on ensuring a better diversity in the background database rather than fundamentally changing its concept.

During my thesis project, a similar and very strong effort has been launched [442] to study the diversity of structural variants in the human genome where 17795 individuals were comprehensively characterized to this end. Though this large-scale effort has a distinct similarity to the SOPHIA background model approach, it diverges on two important aspects:

i) SOPHIA aims to characterize both the artefactual and real breakpoints in control samples for the purpose of filtering structural variants in cancer genomes whereas the approach in [442] characterizes only the real germline variants and their effects,

ii) The [442] study presents results of structural variant analysis based on an existing analysis tool, Lumpy, whereas SOPHIA is strongly based on a background breakpoint database for its function.

.

The latter point reveals an interesting conundrum: SOPHIA's performance strongly relies on its background database of breakpoints, whereas the quality of the breakpoints presented in [442] relies on the performance of the Lumpy algorithm that they employed. We believe that the current performance of SOPHIA is encouraging and the best solution may be to generate a larger background database of breakpoints covering more ethnicities.

2.4.2.4 Reliance on a very specifically defined set of input prerequisites

SOPHIA relies on alignments from the standard BWA-mem based PCAWG workflow [361] on the GRCh37 human genome. This alignment methodology has been used in many thousands of human samples including the PCAWG project encompassing the TCGA, most of the ICGC project and many published and unpublished DKFZ projects. Thus, it was not illogical to develop a variant detection approach strongly relying on the conventions and the data from this workflow. Nevertheless, the workflow is not robust to a change in any parameter: the input alignments have to be coming from this specific workflow and no changes from this standard are supported. Any change in alignment parameters would possibly lead to changes in where artefactual breakpoints would occur, how supplementary alignments are assigned and scored, among many other changes that cannot be fully described *a priori*.

Moreover, the background database of breakpoints are obtained from 3417 control samples coming from this exact same workflow, and have been chosen with meticulous quality checks discarding cases with excessive numbers of artefactual breakpoints in the germline, mainly due to DNA degradation in stored blood. Different users outside of large cancer genome research centres would likely find it difficult to obtain a sufficient number of control samples to build a satisfactory database of breakpoints.

Thus, SOPHIA is not a universal or easily adaptable solution for the detection of structural variants. While it is fully conceivable that the fundamental concepts behind SOPHIA would be applicable in different analysis settings, such as even non-human research, the current SOPHIA workflow is only available for alignments generated with a specific alignment workflow on a specific version of the human genome.

Nevertheless, SOPHIA offers a complete and strong solution for detection of structural variants in the cancer genome with the BWA-mem based PCAWG workflow on the GRCh37 human genome, a standard used in hundreds of projects, and many thousands of samples. We are already working on offering a similarly complete and strong solution based on the GRCh38 build of the human genome, an investment which will be valid for many years to come. We thus hope to compensate for the rigidity of SOPHIA's input expectations by covering a larger spectrum of use cases.

2.4.3 Outlook for SOPHIA's Future Development

At the end of this chapter describing SOPHIA, I would hereby like to suggest the following goals for its development as the developer and heaviest user of SOPHIA:

i) Transition of the SOPHIA workflow to the GRCh38 human genome along with the release of an updated breakpoint database,

ii) Developing a method to distinguish mobile sequences such as retrotransposons from balanced structural variants,

iii) Improvements to systematic weaknesses discovered by the detailed analysis comparing it to SvABA and other local assembly based algorithms,

iv) More advanced analysis of SV clonality by including more information regarding the coverage states,

v) Development of a complementary SV-signature [404] pipeline allowing automated detection of SV-signatures,

vi) Development of advanced annotations including predicted effects of intra-exonic variants such as internal tandem duplications and annotation of ENSEMBL regulatory regions.

An interesting direction to explore would be parameter learning via an automated process, i.e. machine learning rather than the expert-controlled parameter optimization for SV filtering presented in Section 2.2.6. Methods based on deep learning have been developed for the detection of point mutations [468] and deletions as an SV class [469]. These methods had the common concepts of working on images of visualized alignment data, and needing large training datasets due to the nature of deep learning. Our approach on identification of key parameters followed by their optimization was also partially built on visual analysis based on the experiences made in Chapter 1 and close inspection of IGV plots such as those showed in Section 2.3.2 was instrumental in our workflow. However, following the identification of parameters, we switched to a quantitative workflow based on read evidence from different SV evidence categories rather than relying on images. Furthermore, we used only a small training dataset but were still able to develop well-performing filtering criteria from FISH results. Following these points, a more appropriate tool could be random forests which automates the decision tree approach we used [470].

Any change in the SOPHIA algorithm should strive to maintain the standards of excellent performance and sensitivity set here. Even though this aim is not simply solvable by unit-tests, future developers of SOPHIA will have access to diverse and large international and DKFZ datasets for testing its performance following iterative improvements. Speed and memory usage are easily measurable and strict standards should be adhered to, rejecting any improvement that inflates runtimes or memory usage considerably over the current standard. Sensitivity should be checked by gold standard cohorts, such as the ICGC-MMML for known, hard-to-detect IG translocations, or ICGC-EOPC [420] for *ERG* fusions and GPOH-NB for *TERT* fusions [212] or the currently unreleased consensus SV dataset from the PCAWG consortium [284]. Specificity is harder to test in a systematic manner, but developers should pay extra attention to subclonal structural variants and ensure that these are not spurious, low quality observations that make it past filters such as those listed in Section 2.2.6.

CHAPTER 3
EPISTEME: AN INTERACTIVE AND INTEGRATIVE PLATFORM FOR ANALYSING, INTERPRETING AND SHARING MULTI-OMICS DATA

3.1 Introduction

Cancer omics is a highly collaborative field where medical doctors, biologists, bioinformaticians and computational biologists work together to study causes, classifications, mechanisms and possible treatment avenues for different types of cancer [276], [471]. The role of the bioinformatician in this context is broad: they develop statistical methods and fast and efficient software implementations for analysis of data, run analyses and visualize results sharing them preferably in an accessible and intuitive manner with the doctor and biologist partners of the research project, ending with a joint interpretation of the obtained and shared results.

Frequently, individual tasks of the computational biologist on the last two steps, namely analysis and data visualization are amenable to automation allowing the computational biologist to focus on more advanced tasks of method development and implementation [456]. This is accomplished by using modern techniques of interactive data visualization allowing dynamic execution and representations of complex data analysis tasks without the necessity to express these tasks in a programmatic manner. This field of interactive data visualization is currently gaining significant interest [472] with modern web technologies allowing rich interactions with complex datasets. As of mid-2019, there are both available specialized solutions catering to a broad set of bioinformatics needs as well as specialized tools that focus on single data types such as mutations or transcriptome analysis, with both types of tools being a research subject of great interest [473] [474] and a significant research effort in international cancer genome analysis consortia [475].

i) R2 (2008) [337] is among the most mature and feature-rich interactive data portals in the field of omics data visualization. While R2 started mostly as a tool for microarray data analysis and visualization, today it offers features on diverse fields such as SV visualization, ChIP-Seq data visualization and survival analysis. Recently, it served as the official data portal for the Pediatric PanCancer project [476], where it was used to visualize both genomics and transcriptomics data. To date, it has not been published in a peer-reviewed journal.

ii) cBioPortal (2012) [477] is a well-established data portal with an excellent oncoplot (OncoPrint) feature as well as analysis features on copy numbers, gene expressions, mutation mutual exclusivity and co-occurence, pathway enrichment etc. As such, it is a mainstay in data analysis in the field of cancer omics data analysis.

iii) iCanPlot (2012) [478] is an interactive HTML5 Canvas plotting library which offers fast and interactive plots of high-dimensional datasets on scatter plots. At the time, it was a modern implementation of a new web technology and deserves mention as a technical accomplishment.

iv) canEvolve (2012) [479] was another early effort in integrative omics data analysis and visualization in the web. The analysis features were limited, and relied on storage of pre-computed values. The visualization features also encompassed a limited number of basic plots, but integrative data analysis between omics data layers was available. Unfortunately, its user interface was not intuitive and with its analysis features limited, it did not get adopted by a large audience.

v) MAGI (2015) [480] was a tool dedicated to visualization of genomic aberrations using oncoprints for mutation incidences across a cohort, "lollipop plots" for mutations on transcripts, heatmaps for gene expressions, and an UCSC Genome Browser-like genomic region viewer for copy number variations. MAGI was a tool strictly dedicated to visualization and offered no user-controlled data analysis features. As it shared a large domain with cBioPortal, it was not adopted by a large audience.

vi) TumorMap (2017) [481] was the tool to showcase the large-scale sample classification efforts in the TCGA Pan-Cancer analysis consortia, both in its second generation publications [275]. It is a dedicated tool for interactive visualization dimensionality reduction analyses, most often used in pan-cancer analysis. It offers features on user-controlled custom selections, filtering, pathway enrichment analysis. In its very specific usage domain, it is a very good tool albeit with limited performance likely due to the complexity of the datasets involved.

vii) OncoScape (2018) [482] is a comprehensive integrative omics data analysis and visualization portal. It offers both a broad range of data analysis and visualization features including PCA, Survival analysis, visualization of mutations among others. It is a relatively new tool, and its adoption in our community is yet difficult to assess.

viii) Vizome (2018) [483] The Beat AML study's data portal with extensive visualization features for genomic variants, gene expression, tumour evolution, protein alterating mutation summarization, drug response, clinical metadata and gene set enrichment. It has excellent features for subcohort definitions but lacks features for user-controlled data analysis and integrative omics data analysis. Interestingly, it seems to be a prototype with no information regarding its developer team or its own publication.

The state of the literature for omics data analysis and visualization platforms clearly showed a lack of development in the visualization of structural variants. This is partially due to the excellent availability of Whole-Exome Sequencing (WES) data with Whole-Genome Sequencing (WGS) data lagging behind [275]. Of the tools presented thus far, only R2 and OncoScape have an interactive Circos plot feature, even though this might change in the future with better availability of WGS data and libraries to plot Circos plots in the browser such as BioCircos.js [484].

In addition to the (omics) domain-specific visualization libraries and data portals discussed here, Jeffrey Heer's (University of Washington) work deserves a mention as his group has been pioneering in the field of modern interactive data visualization: The Protovis toolkit [485], was the predecessor of the seminal D3.js data visualization library [329], which has since established itself as a standard tool for data visualization in the web. One of the recent innovations

from this group includes the Vega-Lite algorithm which facilitates automatic interactive plot generation from simple, declarative, non-programmatical expressions [486], underlining the powerful capabilities of web-based visualizations. D3.js formed the technical backbone of this dissertation chapter's work.

Having developed a well-performing structural variant(SV) detection algorithm in SOPHIA, and further tools for follow-up analysis studying the impact of SVs on potential target genes, I turned my attention to the open question of data integration, visualization, and sharing for multi-omics datasets. One of the primary motivations of investigating SVs is studying their effects on phenotypes like survival or gene expression, such as gene activations via enhancer hijacking and amplifications or loss of expression via homozygous deletions or loss of proximity to regulatory activators such as enhancers. Other such interesting processes include "double-hits" involving the loss of one allele by a copy number loss or a SV and the other via a deactivating small variant, or "multi-mechanism" activations of a gene involving amplifications or small variants in different patients which are often mutually exclusive. Of course, the eventual aim of such investigations is to reach biological insights, which is only possible in the context of established biological knowledge in databases of measurements or publications.

The result of my work was a comprehensive omics data analysis and visualization portal named EPISTEME. EPISTEME addresses the variant-to-variant, variant-to-phenotype, and observation-to-database integrations in an accessible manner empowering generators of omics data to rapidly reach and share biological insights without extensive consultation to bioinformatics experts for most steps. Indeed, EPISTEME was to a great extent motivated by my own experiences in collaborative projects: Following the development of SOPHIA, I had the opportunity as an "SV-Expert" to extensively interact with scientists who are not from a computational background, therefore depending on computational biologists to access their data. This access is enabled by the computational biologist providing them with visualizations on comprehensively pre-processed, quality-controlled, processed and analysed data. While the pre-processing, quality-control, processing and some advanced types of data analysis are most logically executed by a computational expert, most types of data analysis and data visualization can in principle be executed by any scientist with sufficient domain knowledge. This is partly due to the great exposure of biologists and medical doctors to the explosive growth in omics data-driven knowledge, which introduced a broad audience to omics data types and common visualization approaches [276], [471]. However, in daily practice, execution of simple analyses such as differential gene expression analysis, preparation of simple visualization such as Circos plots, scatter plots or volcano plots most frequently is the responsibility of the computational biologists. Simple questions like "Which patients have a structural variant on or near the *MYC* gene in my cohort?", "Which patients overexpress the *CCND1* gene, and how many of those cases are *TP53* mutant?", "What are the genome-wide SVs observed in these 3 cases, displayed on a Circos plot?" can rapidly accumulate over the course of a collaborative consortium project, taking valuable time away from computational biologists which could better be spent on advanced method development and programming of cutting-edge data analysis algorithms with the capacity of a bioinformatician. On the other hand, the dependence of the biologists and medical doctors on other team members for very simple tasks can lead to delays

in a project's progression. Based on these observations and my personal experiences in terms of which analysis tasks are the most common and how amenable they are to computational biologist-independent execution, I developed EPISTEME.

3.2 Methods

3.2.1 Study Design

EPISTEME is designed to integrate genomic variant data with omics phenotype data and clinical metadata. In its current form, it uses as variant data: SNVs, small indels (defined as short enough insertions and deletions to map correctly as single gapped reads and not create split reads in alignment), copy number variants, SVs, and cohort-wide recurrence profiles of these variant classes. As for phenotype data, it currently uses overall survival data and gene expression profiles from RNA-sequencing or RNA microarrays as well as beta values from methylation array assays. Figure 3.1 summarizes the process that EPISTEME uses to generate integrative cohort-wide and per-patient analyses from genomic variant and multi-omics phenotype and clinical (meta-)data, producing interactive, customizable and, publication-quality visualizations.

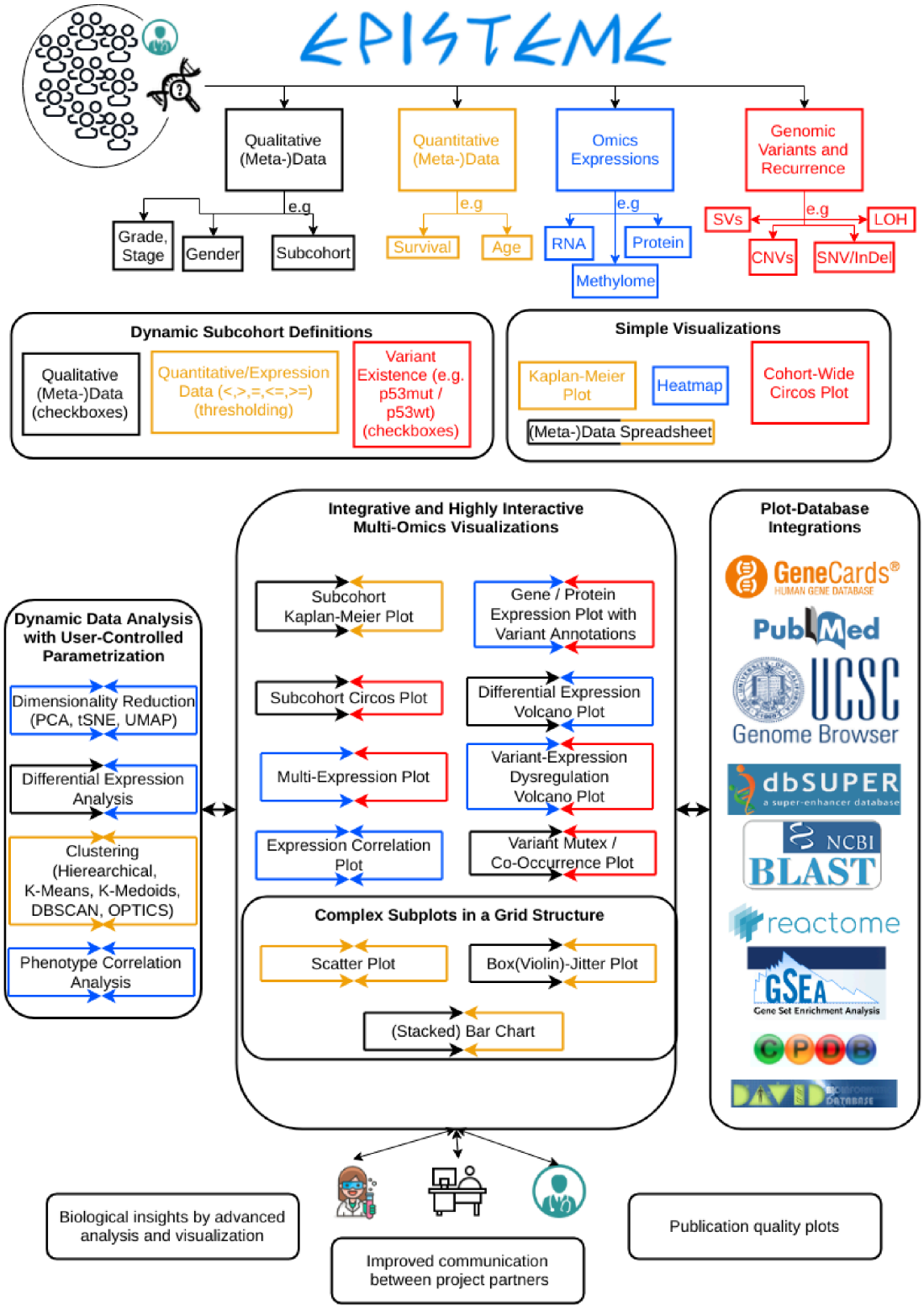

Figure 3.1: Study Design of EPISTEME

3.2.2 Data Sources

EPISTEME processes and visualizes somatic small variants including single nucleotide variations (SNVs) and small insertions and deletions, Copy Number Variants(CNVs) and Loss of Heterozygosity (LOH), and Structural Variants (SVs).

For the detection of somatic small variants, the methods follow those as presented in [211], using mpileup [362] for SNVs and Platypus for [306] small indels. Mutational signatures are analysed by DeconstructSigs [305] with standard parameters using "genome" for WGS and "exome2genome" for WES data.

Small variant annotations are done by ANNOVAR [487] as outlined in the studies cited for describing the method of detection. The following classes of SNVs were considered functional: nonsynonymous SNVs, stoploss and stopgain variants, frameshift and non-frameshift indels and splice site variants. In addition, "upstream", "downstream" and UTR3/5 variants as annotated by ANNOVAR were included in the variants database for further analysis.

Copy number variants and LOH are called using the ACEseq algorithm [309] developed by Kortine Kleinheinz in DKFZ Heidelberg using the default DKFZ workflow [309]. The ACESeq workflow yields segments along with their estimated copy number values and LOH status. As pre-processing for EPISTEME, homozygously deleted segments smaller than 1000 base-pairs and Segments on chrY are excluded from the results. The estimated tumour copy number for each segment is rounded and compared to the rounded estimated base copy number of the tumour. If $TCN_{segment} > TCN_{base}$, the segment is considered gained whereas $TCN_{segment} > 3 * TCN_{base}$, the segment is considered amplified. Similarly, if $TCN_{segment} < TCN_{base}$, the segment is considered lost, whereas if $TCN_{segment} = 0$, the segment is considered homozygously deleted.

Structural variants are called using the SOPHIA algorithm described in the Chapter 2. Recurrence of SVs based on TAD hits is analysed as described in Section 2.2.10.

RNA-Seq and RNA Microarray data is processed as described in Section 2.2.11. For dimensionality reduction, top N most variable genes were determined while excluding gonosomal genes.

Methylome data from methylation arrays and Reverse Phase Protein Lysate Microarray (RPPA) data for TCGA projects are obtained from the Genomics Data Commons (GDC) mirror of UCSC Xena [455] as normalized values without further transformations. Probes were filtered following the suggestions in [488], filtering out common SNPs, gonosome probes and underperforming probes as well as non CpG probes.

Genes and TADs are defined as described in Section 2.2.10.3.

3.2.3 Data Storage Backend

EPISTEME efficiently uses a SQL database (MariaDB) for minimizing the persistent data storage inside a given browser session. High-dimensional omics data is fetched dynamically, on an as-needed basis (Figure 3.2).

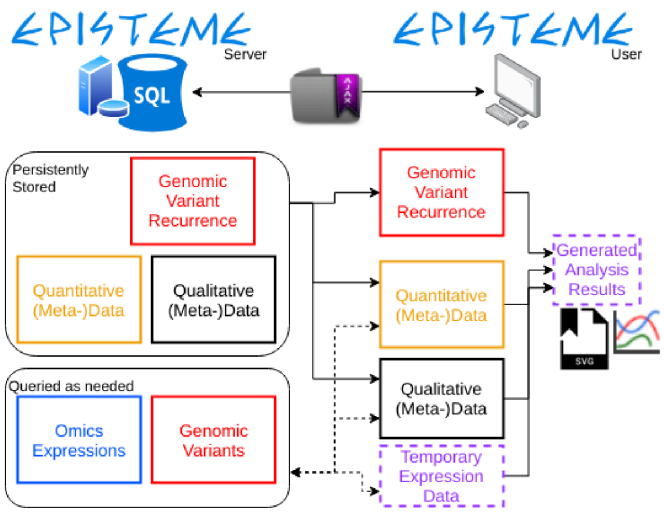

Figure 3.2: Data transfers and processing relations in EPISTEME

Briefly, EPISTEME stores low-dimensional data such as clinical metadata, genomic variant recurrence and SVs in persistent memory for quick access during a cohort's initialization while keeping high-dimensional data such as genomic variant status of each gene, cohort-wide transcriptome and methylome values on the server and fetching the data as needed. This dynamic fetching can be executed in three forms:

- a targeted fetching of a single gene's genomic variants or gene/protein expression values leads to the permanent addition of the fetched data as a quantitative or categorical cohort metadata category, giving users full power to using the data in scatter plots, box-violin-jitter plots or in subcohort assignments
- a high-dimensional fetching of expression values of *top N* most variable genes/probes for clustering purposes yields temporary data which are deleted after dimensionality reduction/clustering is run
- a genome wide fetching of genomic variant status for all genes for the purpose of showing variant recurrence in a (sub)cohort-wide Circos plot also yields temporary data which are deleted after the Circos plot is plotted

. This overall lowers memory usage with the tradeoff of occasional SQL queries via AJAX. As these queries are read-only, the concurrent activity of multiple users do not pose technical challenges for the database.

3.2.4 Data Visualization on the Frontend

EPISTEME is organized in a modular structure where data is received from the server, processed with in-browser algorithms and visualized in interactive visualization, allowing rapid development of new features (Figure 3.3).

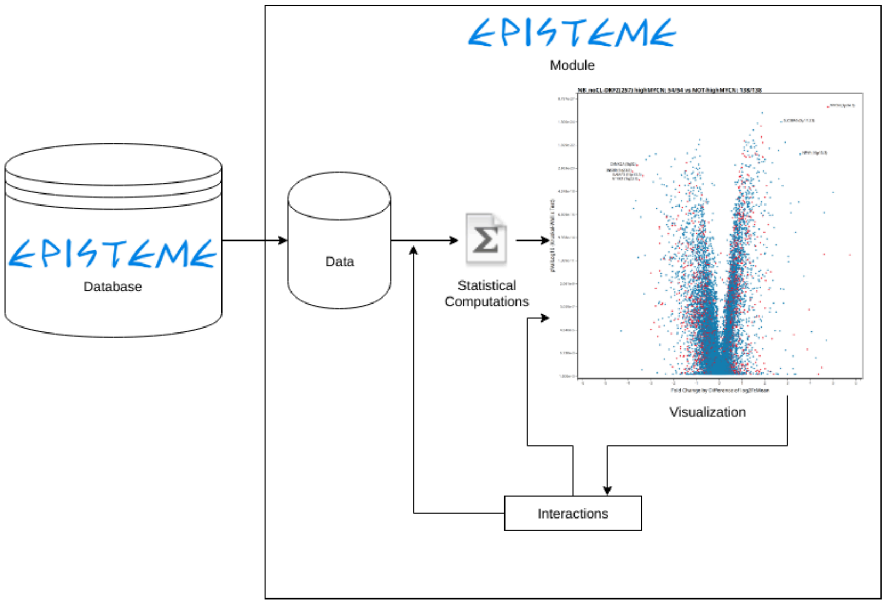

Figure 3.3: The general module structure in EPISTEME

EPISTEME generates vector graphics using the Web-Standard SVG format. Interactive visualizations are produced by the JavaScript library D3.js and exported using the svgsaver package. The only other external visualization library used in this project is venn.js providing Venn diagrams for the visualization of comparative variant frequency across two subcohorts.

The jQuery-QueryBuilder package is used to give users the ability to generate complex subcohort selections starting from simple subcohort selections (Section 3.3.8.5).

The user interface is built using Twitter Bootstrap library. jQuery is used to manage the user interactions.

Analysis Type	Repository	Version
D3.js	github.com /d3/d3	4.13.0
jquery	github.com /jquery/jquery	3.3.1
svgsaver	github.com /Hypercubed/svgsaver	0.9.0
Bootstrap	github.com /twbs/bootstrap	3.3.7
Bootstrap-Slider	github.com /seiyria/bootstrap-slider	10.0.2
QueryBuilder	github.com /mistic100/jQuery-QueryBuilder	2.5.2
Venn Diagrams	github.com /benfred/venn.js	0.2.20

3.2.5 Visualization of Genomic Variants and Genomic Variant Recurrence

EPISTEME's first screen when a cohort is loaded is a cohort-wide Circos plot summarizing the mutational landscape of a group of patients. It shows variant recurrence for chosen variant types. Circos plots have become a widely adopted standard for visualizing genome-wide recurrence information [332] and the genomics community has a strong familiarity with this visualization. While circular representations lead to a distortion of quantities across the r-axis as more layers of information are added, there doesn't exist to date a better class of diagrams for

visualizing correspondences between chromosomes such as chromatin interactions or translocations. Thus, EPISTEME maintains this standard in visualizing the mutational landscape of a cancer cohort.

EPISTEME organizes the genome in a hierarchy:

- Chromosomes, used only for visualization and zooming
- Chromosome arms, used only for visualization and zooming
- Cytobands, used for visualization, zooming and annotations
- TADs, used for visualization, annotations, variant recurrence visualization

EPISTEME maps genomic coordinates to Θ angles in a way allowing a dynamic and rapid switching between individually controllable magnification levels for each cytoband. Emphasizing particular chromosomes, chromosome arms or cytobands and showing regions rich in important SVs in higher detail is a trivial action in EPISTEME. Because TAD boundaries define overlapping TADs, the smallest non-overlapping segments of the genome are cytobands. Each cytoband has a coefficient, which is increased to emphasize, or decreased to de-emphasize the contributions of the given cytoband to the overall calculation of Θ.

Each genomic location is defined as a pairing of chromosome and position on the chromosome. This pairing can be converted to an "absolute position" using the following formulas.

Given the genomic location defined on chromosome with the index K, and on position pos, this position is on the Q^{th} cytoband of chromosome K, separated from the next chromosome $K+1$ by a padded gap Gap_i.

$$\Theta(chr_K, pos) = \frac{L_{elapsed}(chr_K, pos)}{TotalGenomeLengthWithGaps}\pi + \Theta_{offset}$$

$$\begin{aligned} L_{elapsed}(chr_k, pos) = & \sum_{i=1}^{K-1}(L_i + Gap_i) + \\ & PosOnChr_K BeforeCytoband_Q + \\ & CoeffCytoband_Q * LenCytoband_Q \end{aligned}$$

$$L_i = \sum_{p=1}^{NumCytobands_i} (CoeffCytoband_p * LenCytoband_p)$$

$$PosOnChr_K BeforeCytoband_Q = \sum_{p=1}^{Q}(CoeffCytoband_p * LenCytoband_p)$$

To make this calculation much quicker, EPISTEME stores the $L_{elapsed}$ for all cytobands for a given set of zooming coefficient values. Anytime the user changes a coefficients, this lookup table is updated. The lookup table simplifies the equation dramatically:

$$\Theta(chr_K, pos) = \frac{L_{elapsed}(chr_K, pos)}{TotalGenomeLengthWithGaps}\pi + \Theta_{offset}$$

$$L_{elapsed}(chr_k, pos) = L_{ElapsedBeforeChromosomeKCytobandQ} + CoeffCytoband_Q * LenCytoband_Q$$

This very fast memoization of $L_{ElapsedBeforeChromosomeKCytobandQ}$ for rapidly calculating genomic theta values makes the plotting zooming and rotation, of the Circos plots much faster. The user can interact with the Θ_{offset} parameter or any of the zooming coefficients $CoeffCytoband$ to modify the rotation and zooming characteristics of the plot. Zooming in or out on a chromosome, or chromosome arm sets the coefficients for each cytoband on the chromosome (arm) to the set value. Finer grained control over zooming is established by modifying the coefficient for individual cytobands.

As in the Circos package, EPISTEME uses quadratic Bézier curves to represent individual SVs or small variants. As standard tools in computer graphics, quadratic Bézier curves have an implementation in the SVG standard and are straightforward to use.

A Bézier curve in 2D space consists of 6 polar coordinates:

- R_{start},Θ_{start}
- $R_{control}$,$\Theta_{control}$
- R_{end},Θ_{end}

,with $R = R_{start} = R_{end}$.

Θ_{start} and Θ_{end} are determined by genomic positions as described. The control points are determined by the type and size of the displayed genomic variant: $R_{control} = (1 - h_{variant}) * R$ $\Theta_{control} = 0.5 * (\Theta_{start} + \Theta_{end})$

$$h_{variant} = \begin{cases} 0.15, & SmallVariant \\ 0.35, & MediumVariant \\ 1, & LargeVariant \end{cases}$$

With small variants are intrachromosomal variants spanning less than 9 MB, medium variants are intrachromosomal variants spanning less than 18 MB and large variants are any other variants.

These definitions allow users to visually distinguish between genomic variants of different sizes.

Genomic variant recurrence is plotted as polar arc spanning a TAD or gene borders. A polar arc is defined with four coordinates in 2D space:

- R_{start},Θ_{start}
- R_{end},Θ_{end}

Θ_{start} and Θ_{end} are defined using the genomic coordinate to Θ mapping approach as described. For TADs, the results are used as calculated, while for genes, if the $\Theta_{end} - \Theta_{start} < 0.01(rad)$, it is expanded to as $\Theta_{startNew} = 0.5 * (\Theta_{start} + \Theta_{end}) - 0.005$ and $\Theta_{endNew} = 0.5 * (\Theta_{start} + \Theta_{end}) + 0.005$.

R_{start} is determined by the starting point of the arc wheel on the Circos plot, which will be explained in Section 3.3.1. R_{end} is determined by the formula $R_{end} = min(R_{wheel}, R_{start} + (\frac{recurrence_{TAD/Gene}}{recurrence_{maxCohort}})(R_{wheel} - R_{start}))$. $recurrence_{maxCohort}$ is a user-controllable parameter allowing the modulation of the heights of the recurrence of interest, used in the interactions of Section 3.3.1.

3.2.6 Data Analysis Features

EPISTEME provides a number of data analysis tools, of which most originate from open source JavaScript repositories. Some useful tools were hand-converted to the web-compatible language JavaScript if they were not otherwise available as JavaScript code.

Analysis Type	Repository	Version	Comments
tSNE	github.com /karpathy/tsnejs	-	Hand-optimized and added support for late-exaggeration
UMAP	github.com /lmcinnes/umap	0.3.8	Hand-converted to JavaScript
Eigenvalue Decomposition (for UMAP)	github.com /mljs/matrix	5.2.1	
PCA	github.com /mljs/pca	2.1.0	
Hierarchical Clustering	github.com /tayden/clusterfck	0.7.0	
K-means Clustering	github.com /Philmod/node-kmeans	1.1.8	
K-medoids Clustering	github.com /stewartr/k-medoids	1.0.4	
Fuzzy DBSCAN Clustering	github.com /schulzch/fuzzy-dbscan-js	1.0.1	
OPTICS Clustering	github.com /uhho/density-clustering	1.3.0	
Concave Hull fitting	github.com /mapbox/concaveman	1.1.1	
Kernel Density Estimation	github.com /Planeshifter/kernel-smooth	0.2.3	
Sheather-Jones Bandwidth (KDE)	github.com /Neojume/pythonABC	Hand-converted to JavaScript	
Kolmogorov-Smirnov Test	github.com /pieterprovoost/jerzy	0.2.1	

T-Test	github.com /pieter-provoost/jerzy	0.2.1	
Kruskal-Wallis Test	lib.stat.cmu.edu /apstat/245	-	Hand-converted to JavaScript
Fisher's Exact Test	bioinfo.iric.ca /a-javascript-implementation-of-the-non-central-version-of-fishers-exact-test/	-	

All of these features/analyses run on the client-side with no processing load on the EPISTEME server.

3.2.7 Database Integrations

EPISTEME offers extensive integrations to genomic databases for interactive analysis, which are tailored to the needs of each visualization type.

Gene names are linked to *Genecards* [489] [490] regardless of visualization type. Gene names are additionally linked to *PubMed* [491] in "Variant-Expression Dysregulation Volcano Plot" and "Gene / Protein Expression Plot with Variant Annotations" visualizations to check the novelty of dysregulated genes.

Genomic variant data annotations are linked to *UCSC Genome Browser* [492]. SV data annotations are additionally linked to *dbSUPER* [493] for superenhancer annotations, and *NCBI BLAST* [494] for split read overhang annotations.

Visualizations including gene lists such as Volcano or 1-vs-All Correlation plots are linked to a number of options for pathway enrichment / gene set analysis: *DAVID* [495], *Reactome* [496], *GSEA* [497], ConsensusPathDB *CPDB* [498].

3.2.8 Differential Expression Analysis

Differential expression analysis combines the measures "statistical significance" showing the consistency of a difference across two groups and "fold change" showing the magnitude of a difference between two groups. Fold change and statistical significance information can be depicted in a "volcano plot", a standard visualization technique for differential gene expression analysis across the whole measured transcriptome: [499]. Each underlying data point in a volcano plot represents a gene (or some other single omics phenotype measurement such as protein expression or metabolite concentration).

EPISTEME approximates statistical significance by four methods:

i) modified Kruskal-Wallis test,

ii) Kolmogorov-Smirnov test,

iii) Student's T-Test,

iv) Fisher's Exact Test on Jenks-Optimized Breaks (only in variant-integrated mode)

For the modified Kruskal-Wallis test, no tie-correction is applied, but rather the input expressions for both groups are shuffled $2 * Ties_{Group1}$, or $2 * Ties_{Group2}$ whichever is larger up to a maximum number of 100 iterations. This procedure is inspired from permutation testing and is adopted because of some practical issues regarding the tie-correction on some heavily repeated samples. In its current implementation in EPISTEME, the Kolmogorov-Smirnov test is robust against this issue and produces similar results, but is slower for in-browser calculations.

EPISTEME measures fold change by trimean, a robust estimator ($= \frac{Q_{25}+2*Q_{50}+Q_{75}}{4}$); or mean, a quantity sensitive to outliers. Robustness is usually a desirable property in statistical comparisons, but in the context of genomic variant effects on gene expression, rare variants with strong effects on gene expression leading to outlier observations in a cohort can be biologically very interesting: rare enhancer hijacking events can be important oncogenic events defining rare cancer subtypes [132]. In calculating fold change, because the underlying phenotype quantities are normalized as $\log_2(x+1)$, mapping them back to the initial domain and keeping the $+1$ increment ensures that each group has nonzero mean or trimean values for a fold change comparison: $FC = \log_2(\frac{2^{Val_1}}{2^{Val_2}} + 1)$. This quantity ensures that events like the activation of a gene from 0 expression can be properly quantified.

EPISTEME offers two types of differential expression analysis for distinctly different purposes:

1. A differential expression analysis comparing two static groups chosen with arbitrary criteria (expression of an anchor gene, existence of a particular variant, tumour grade, gender etc.) In this analysis, the group sizes remain constant and results can be filtered for significance or fold change. In this mode, EPISTEME uses the trimean as the default fold change measure. This is a common type of bioinformatics data analysis.

2. A differential expression analysis for each gene with dynamic sample sizes based on variant status, with the aim of studying gene dysregulation. For each gene, the EPISTEME divides a cohort into a *Variant*$_+$ and a *Variant*$_-$ subcohort, which changes for any gene based on the existence status of the investigated variant types. Consequently, the sample sizes for each result data point will differ and the results cannot be filtered by significance or fold change and serve a purely exploratory purpose for discovering dyresulated genes due to genomic variants. In this mode, EPISTEME uses the mean as the default fold change measure. This is a non-canonical analysis and a novel approach offered by EPISTEME.

3.2.8.1 Special Properties of Variant-Integrated Phenotype Dysregulation Analysis

EPISTEME's variant-integrated phenotype (gene/protein) dysregulation analysis, is exploratory: EPISTEME makes no attempts to reduce the whole transcriptome to a list of candidate genes of assumed significance. This has a few underlying reasons, discussed on the particular example of SVs as variants: i) as discussed in the unpaired test consideration for the choice of the statistical test, the sample sizes for each group are different for each gene. This excludes the setting of a single p-value cutoff point, ii) the methods used for estimating the potential effect of an SV

on gene expression are not an exact indicator for SV-gene associations due to inaccuracies of TAD boundaries, nested TAD structures, potential impact of long-range interactions that affect genes on distant TADs and finally SVs that fall on a TAD but have no regulatory effect on a gene of interest residing on the TAD or one of its neighbours. Hence, there can be false positives and negatives due to inaccuracies in SV-gene associations, iii) finally, the SV detection algorithm used for this analysis can have false positive and false negative results leading to noise in the estimated SV-gene-expression associations. If the aim of this analysis is to identify gene activation or inactivation due to different genomic variants, considering the points discussed above, the problem statement for variant-gene expression association analysis can be changed from a simple rank-based comparison of $Variant_+$ vs. $Variant_-$ groups to the question "are patients carrying variants of a given class of genomic variants more likely to have higher or lower gene expression values based on rank?". In order to address this specific question of enrichment of gene over/underexpression for a variant class, EPISTEME uses a sweeping application of the statistical testing which attempts to maximize approximated statistical significance by selectively eliminating $Variant_+$ data points, which are estimated to be potentially due to noise. The following steps are applied: i) the robust estimator, trimean is used to determine the side of sample eliminations, ii) there are at least two patients with $Variant_+$ status, iii) at most 1/4 of samples with $Variant_+$ status from the initial state are eliminated, iv) eliminations stop if the estimated significance is not increased. Within the constraints of exploratory analysis, this procedure de-emphasizes erroneous, passenger or otherwise noise-related variants from $Variant_+$ cases leading to the opposite effect on gene expression for a gene of interest. This approach, while making it visually much easier to identify outlier gene candidates in the volcano plot, would not be mathematically valid in hypothesis testing but is useful especially in the identification of rare events of enhancer hijacking. Due to the complexity of this described procedure and the large number of combinatorial possibilities for allowed variant type combinations, this analysis is currently implemented as an upstream and precomputed type of analysis, where EPISTEME fetches results for each affected gene/protein if a variant-expression association volcano is requested by the user. Hence, it is limited to a visualization on a whole-cohort and not flexible applications to subsets of patients. Currently, this is the only analysis type with this weakness, and should be addressed, making this analysis dynamically runnable in the browser.

In addition to this important non-standard approach, Variant-Integrated Phenotype Dysregulation Analysis has a statistical test option named *Fisher's Exact Test on Jenks-Optimized Breaks* which divides a cohort into two not based on variant status but rather "breaks" in phenotype quantities. Jenks' natural break optimization algorithm was initially described for optimized categorizations in cartography data [500]. While the algorithm is designed to accept an arbitrary number of expected breaks, for this analysis it is applied with 1 expected breakpoint, signifying a bimodal distribution. Using this breakpoint, the cohort is separated into two subcohorts, $Pheno_{high}$ & $Pheno_{low}$ and compared these two groups using fold change measures as described. For statistical significance, the groups $Pheno_{high}$ & $Pheno_{low}$ are analysed together with the groups $Variant_+$ & $Variant_+$ as a contingency table followed by the application of Fisher's Exact Test. This test allows the detection of dysregulated genes with explained (variant-caused) and unexplained variance/bimodality.

3.3 Results

In the following sections, all major and most minor data visualization modules of EPISTEME will be presented, describing their design principles, visualization strategies, data analysis features, integrations to other visualizations, user interaction features and database integrations. The sections are organized in roughly chronological order of feature development, wherever this does not disturb a logical flow in the presentation. This design should therefore reflect the evolution of EPISTEME's features and concepts. The overarching concept in the introduction to EPISTEME is the usage of "pilot cohorts": while introducing each data analysis feature or data visualization approach "pilot cohorts" of well-studied diseases and established biological knowledge will be used.

3.3.1 A Cohort-Wide Circos Plot for Visualization of Mutational Landscapes

The cohort-wide Circos plot is the starting figure for any cohort in EPISTEME and serves the important purpose of summarizing the genomic variant landscape of a cohort. It displays recurrence frequencies of flexibly selectable genomic variant types as well as the variants themselves in an interactive visualization with extensive integrations to databases.

To showcase the features of the cohort-wide Circos plots, the TCGA Glioblastoma Multiforme study [501] will serve as the demonstration cohort, focusing on the subset of cases which were sequenced using Whole Genome Sequencing and analysed in the Pan-Cancer Analysis of Whole Genomes Project. The motivation in selecting this disease and project were as follows:

i) The TCGA-GBM cohort is a medium-sized (41) cohort with a sufficient size to showcase its genomic variant landscape features,

ii) The TCGA-GBM cohort is a heterogenous cohort with subpopulations such as $EGFR_{amp}$, $TP53_{mut}$, $CDK4_{amp}$,$PDGFRA_{amp}$,

iii) The TCGA-GBM cohort has both chromosome, chromosome arm and focal CNV and LOH events,

iv) The TCGA-GBM cohort does not have a generally high load of SVs (unlike BRCA) or SNVs (unlike SKCM or COAD), leading to results being easier to emphasize

Figure 3.4 displays the result of this section's concepts on visualizing mutational landscapes with Circos plots.

Figure 3.4: The mutational landscape of the TCGA-GBM (WGS) study displayed by EPISTEME's Circos plot module. Outermost variant recurrence layer: functional small variants on genes, middle variant recurrence layer: CNV or LOH (as loss) on TADs, innermost recurrence layer: SVs on TADs. Inner circle: SVs. Colour code for SVs; green: translocations, blue: deletions, red: duplications, black: inversions. The following user interactions are showcased: plot rotation (17°clockwise from default) and chromosome-arm level zooming (chr12q, 4x), cytoband level zooming (chr7p11.2 carrying *EGFR*, 16x), cytoband and gene labelling.

The Circos plot module implements two types of visualizations: variants and variant recurrence. Variant recurrence is displayed in recurrence wheels with radial bar charts and variants and genomic variants are displayed using Bézier Curves as described in Section 3.2.5. The recurrence wheels can flexibly be selected to show TAD-centric or gene-centric variant recurrence in 3 possible layers, selectable in an easy user interface (Figure 3.5).

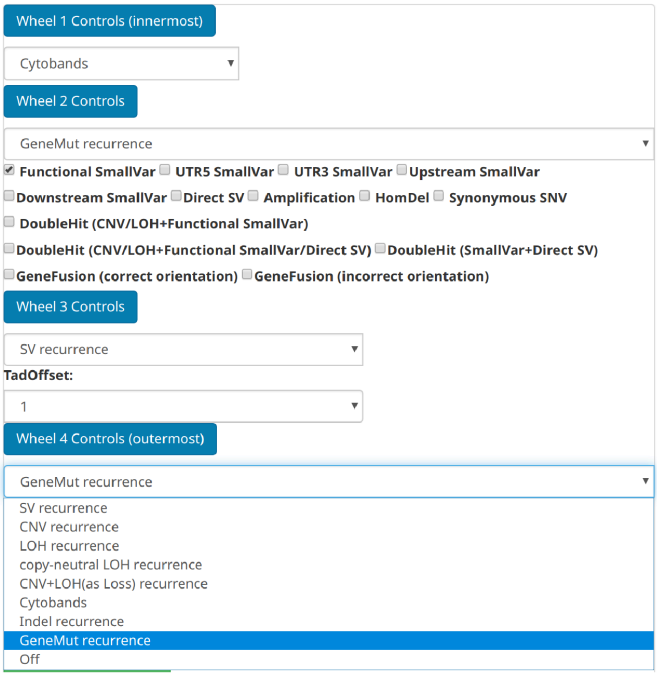

Figure 3.5: Flexible data selection for variant recurrence representation in EPISTEME

The displayed data can be modified and labelled through flexible user interactions with controls (Figure 3.6).

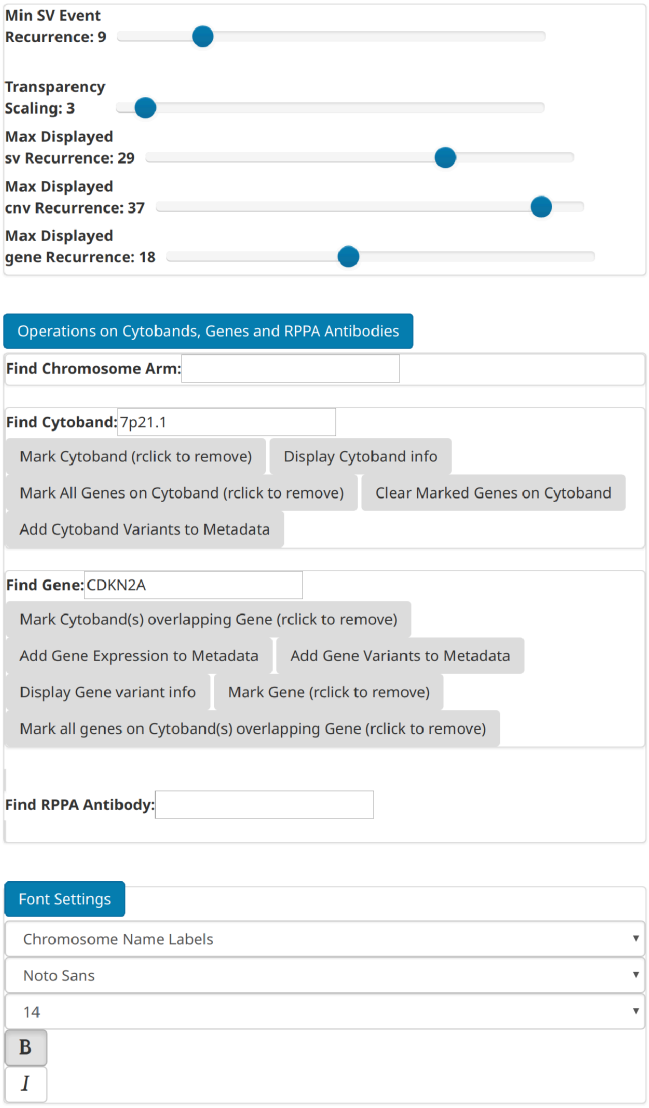

Figure 3.6: Flexible controls for data display and labelling in EPISTEME

The following subsections will discuss the individual concepts used in reaching this final result. The following data types will be discussed:

i) Structural variants,

ii) SV recurrence,

iii) CNV/LOH recurrence,

iv) Gene-centric recurrence.

3.3.1.1 Interactive Visualization and Annotation of Genomic Variants

As of the time of this dissertation, EPISTEME displays by default all Structural Variants, whereas small variants (indels and SNVs) are displayed on-demand, following a focus into a specific TAD or cytoband. This design decision was motivated by the generally lower counts

of SVs over a cohort compared to small variants. Furthermore, it is of interest to show what types of SVs are more common in the cohort in terms of SV type and breakpoint partners.

Overlaying all SVs with predicted functional impact in the TCGA-GBM cohort yields a visually unappealing and uninformative representation (Figure 3.7, left). EPISTEME uses a method to reduce the complexity of this visualization by omitting SVs that do not hit recurrently affected TADs (below a user-determined threshold) and by applying an alpha transparency as a CSS colour modification. This procedure yields an informative result emphasizing the high load of SVs on loci important for GBM biology (Figure 3.7, right). Both of these features that can be adjusted in the user interface (Figure 3.6), serving to reduce visualization complexity in different ways.

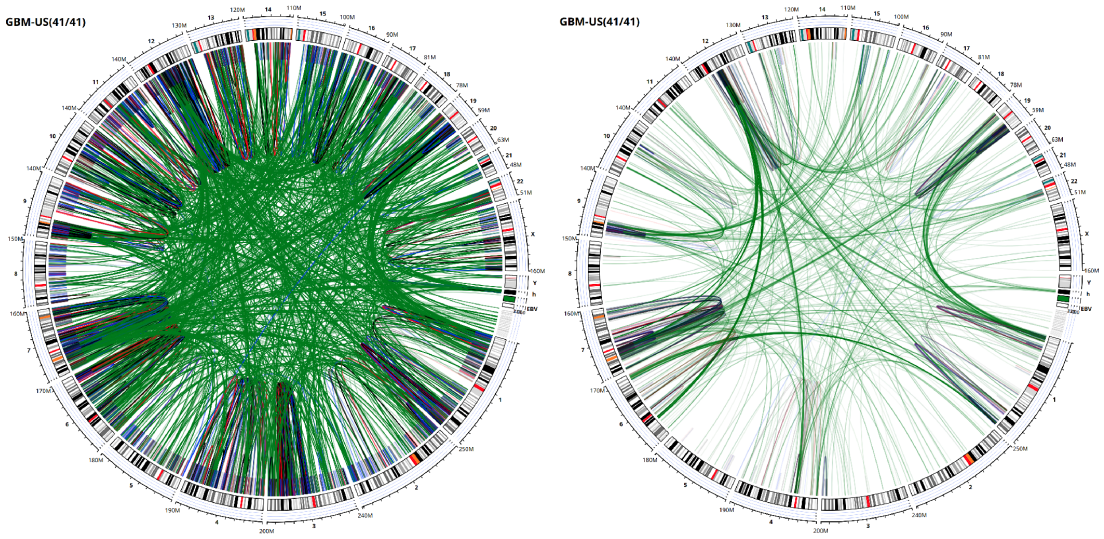

Figure 3.7: All SVs with predicted functional impact in the TCGA-GBM (WGS) study with no transparency scaling (left) and a transparency scaling of 3, filtered to show only SVs hitting TADs with more than 8 recurrence (right). Colour code for SVs; green: translocations, blue: deletions, red: duplications, black: inversions.

In Circos plots, SVs have the following colour code:

- Red: Duplication,
- Green: Translocation
- Blue: Deletion
- Black: Inversion.

The colours were selected with the consideration that they will be transparency-modulated and overlaid. This would rule out combinations such as Red-Blue-Purple. However, the current convention is unfortunately not safe for colour-blind users.

Users can focus on individual cytobands to show only SVs that hit the selected cytoband. Focusing on the cytoband chr9p21.3 carrying the *CDKN2A/B* genes shows detailed cytoband

annotations with integrations to Genecards and UCSC Genome Browser along with the ability to fetch small variants on this cytoband (Figure 3.8). The cytoband has also been marked manually to emphasize its selection. Users can navigate through the cytobands by dedicated buttons or by clicking on them on the Circos plot.

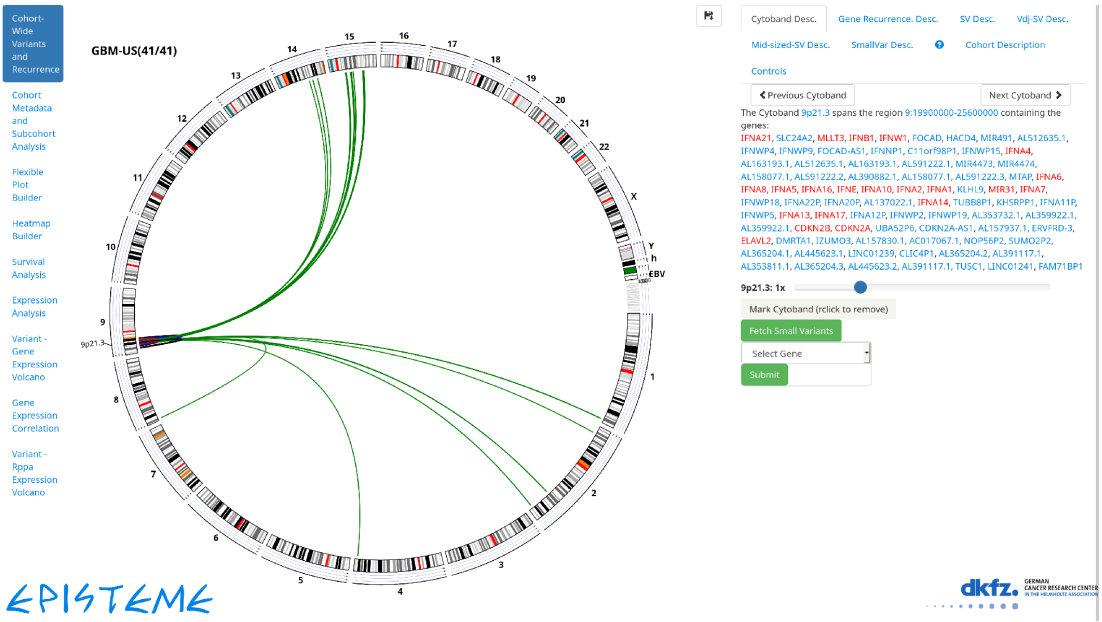

Figure 3.8: Focusing on the cytoband chr9p21.3 carrying the *CDKN2A/B* locus, with corresponding cytoband annotations and navigation features. The right hand side shows the list of genes on the manually selected cytoband with cancer-genes marked in red. User interactions include a zooming in-out slider as well as a small variant fetching function that recovers small variant data from the SQL backend on-demand while SV data is kept in persistent memory.

Zooming 256x into the cytoband chr9p21.3 labels all the genes that are on the cytoband as well as clearly showing the high load of SVs that affect this important locus in this cohort(Figure 3.9).

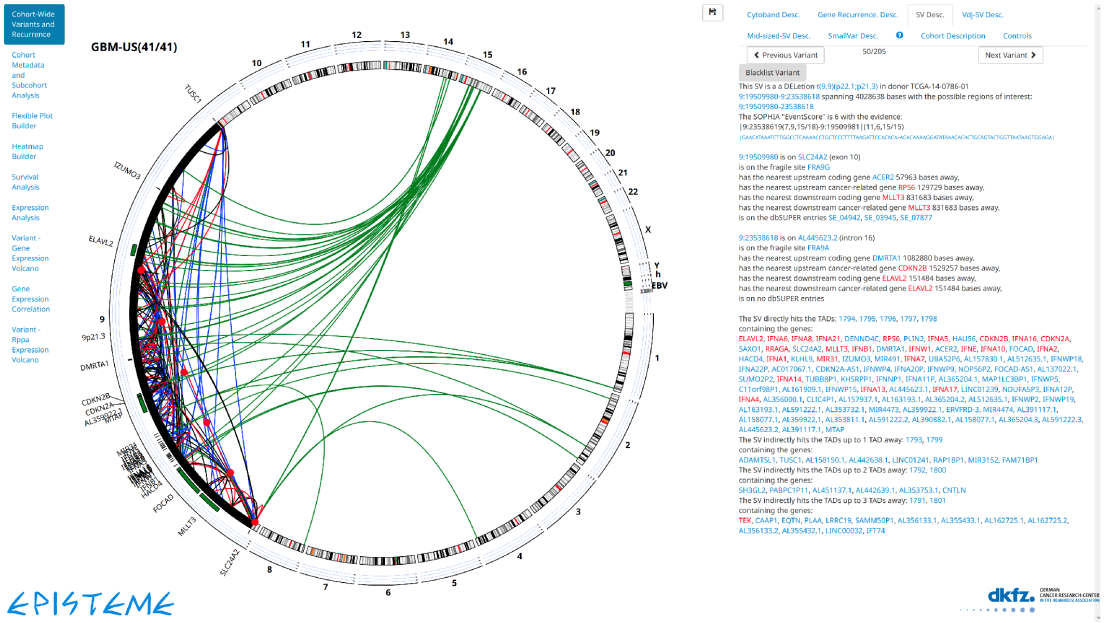

Figure 3.9: Focusing on the cytoband chr9p21.3 carrying the *CDKN2A/B* locus, with a 256x zoom, individual SV selections, annotations and navigation features: Users can navigate through the SVs of the selected cytoband by using dedicated buttons or clicking on the curves that represent the SVs. Clicking on one of the focal deletions launches a detailed and interactive annotation of the SV with integrations to Genecards, BLAST, UCSC Genome Browser, dbSUPER and PubMed. Here, one of the structural variants is selected with annotations (right).

In particular, the UCSC Genome Browser integration (Figure 3.10) is useful for more detailed annotations of SVs that contain more data sources than currently offered in EPISTEME.

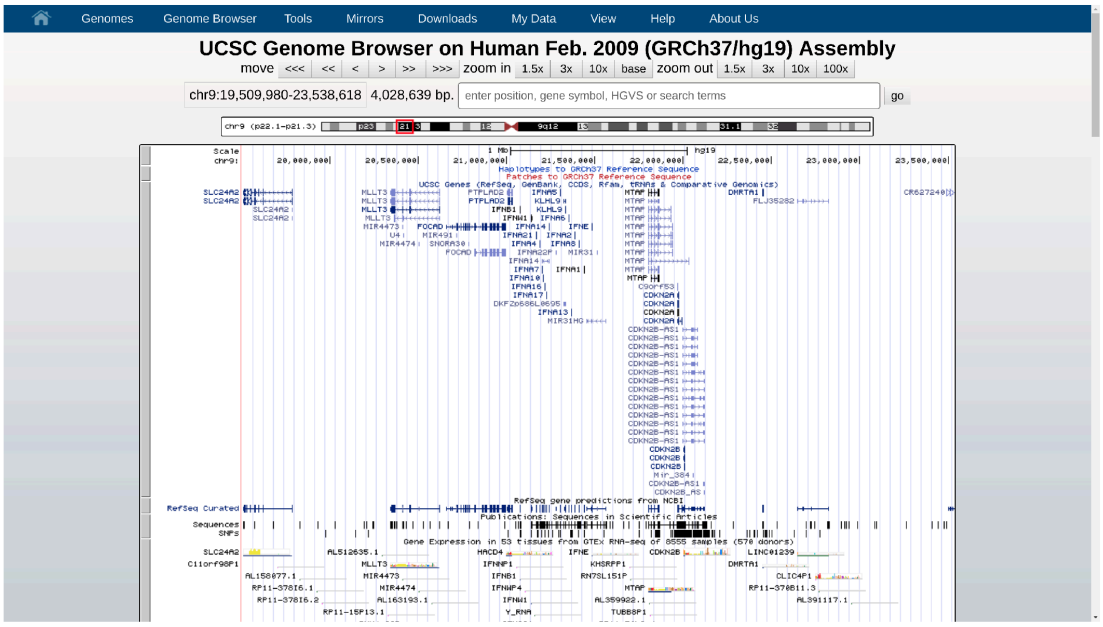

Figure 3.10: The UCSC Genome browser annotation of a *CDKN2A/B* focal deletion region in the GBM patient TCGA-14-0786.

Finally, EPISTEME has a dedicated mode which identifies and labels V(D)J / CSR rearrangements (introduced in Section 2.2.7) and their suggested target genes among many candidates which can be obtained from a TAD-based analysis. This feature is demonstrated in Section 4.3.3.1 on the pilot cohort of Multiple Myeloma.

3.3.1.2 Interactive Visualization and Annotation of Structural Variant Recurrence

The method of distributing detected SVs to pre-defined TADs described in Section 2.2.10.3 can be used to visualize the significance of genomic loci with regards to being frequent targets of SVs. As the displayed information is the number of patients, for which a given TAD is predicted to be affected by an SV of high functional potential, a radial bar chart is an appropriate visualization (Figure 3.11).

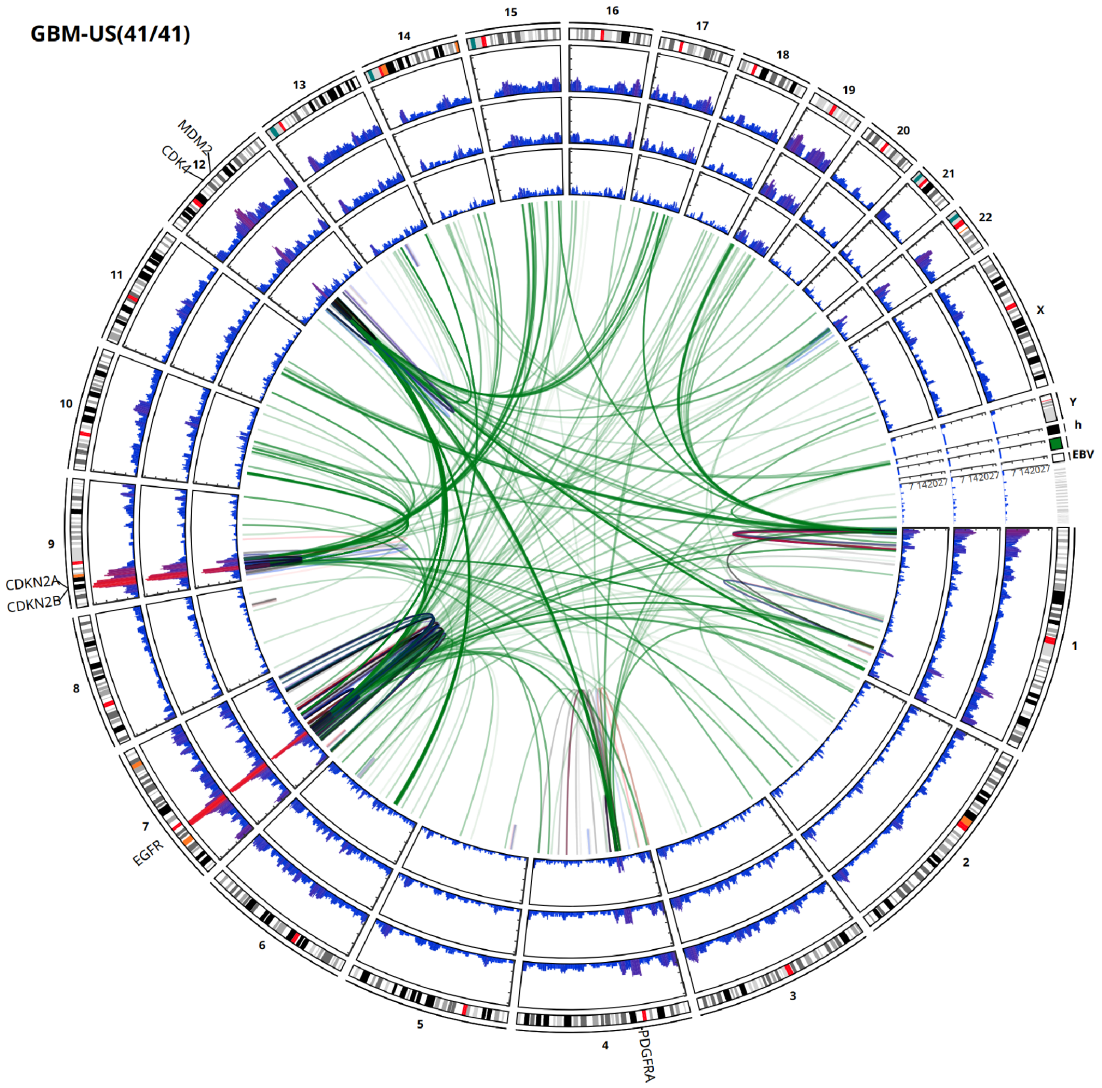

Figure 3.11: Structural rearrangement recurrence landscape of the TCGA-GBM (WGS) study across TADs at different assignment sensitivities (innermost: 1-offset, middle: 2-offset, outermost: 3-offset). Some frequently targeted loci relevant in this disease type have manually been selected and labelled: the *EGFR* oncogene locus, a frequent target of amplifications across different disease types is the most prominent peak closely followed by the *CDKN2A/B* locus, a frequent target of homozygous deletions.

By default, EPISTEME normalizes the height of each TAD-bar to the maximum occurrence in the cohort (for Figure 3.11, e.g 28/41 on the *EGFR* locus at allowed TAD offset=1), but this normalization can easily be altered if it is desired to emphasize the lack of rearrangements in a cohort or to suppress a locus that dominates over the rest of the genome and make more rarely hit regions stand out (Figure 3.6).

As with Cytobands and SVs, interactions are provided to investigate SV recurrence. Upon being clicked, TAD-bars take EPISTEME into a mode which annotates the selected TAD using the databases GeneCards and UCSC Genome Browser (figure omitted due to similarity to

cytoband annotations).

3.3.1.3 Interactive Visualization and Annotation of Copy Number Variant and LOH Recurrence

CNVs are represented in a very similar way to SVs using a TAD-centric approach. The main difference here is that CNVs can both be in a ”gain” or ”loss” direction. Hence, deviations from a baseline in each direction are visualized. A given locus (TAD) can be lost in some patients and gained in others and vice versa.

Figure 3.12 represents the Copy-Neutral-LOH (cnLOH), CNV and combined CNV-LOH recurrence landscape of TCGA-GBM. The cnLOH contributions to the combined CNV-LOH recurrence calculation is made on the Copy-Loss side. The results indicate the prevalence of cnLOH events affecting chr17p containing the master tumour suppressor gene *TP53*. The canonical mechanism of the two-hit loss on *TP53* is via a copy number loss affecting a single copy and a point mutation affecting the other. TCGA-GBM shows here a different mechanism, useful for showcasing this visualization in EPISTEME.

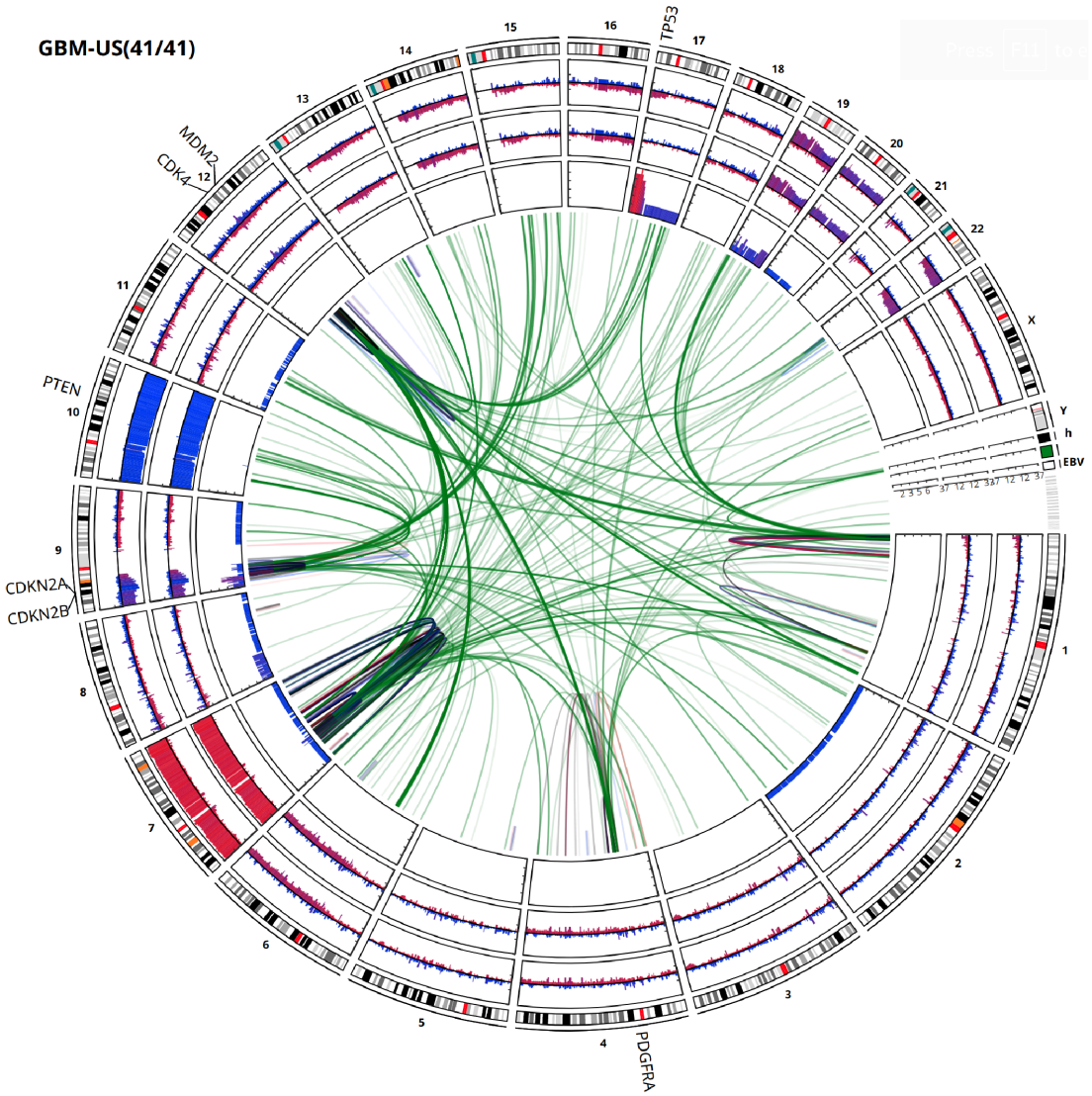

Figure 3.12: Genome-wide per TAD Copy-Neutral-LOH recurrence, Copy Number Variant recurrence, and combined CNV-LOH(as loss) recurrence characteristics of the TCGA-GBM (WGS) study, in order from inside to outside

3.3.1.4 Interactive Visualization and Annotation of Gene Mutation Recurrence

EPISTEME has a feature for picking different types of genomic variant types for recurrence analysis (Figure 3.5). The selected types of mutations are combined in an *OR* relationship and the result is displayed in a similar manner to TAD-based recurrence, with the exception that the displayed objects are genes rather than TADs.

The TCGA-GBM cohort contains a diverse selection of genomic variant types. *EGFR*, *TP53* and *ATRX* are frequently mutated by small mutations whereas *TERT* is mutated by promoter mutations, and finally as previously discussed *EGFR*, *MDM2*, *CDK4* and *PDGFRA* are frequently amplified during tumour development. Figure 3.13 shows these distinctly different genomic variant types across the three gene mutation recurrence analysis layers: Functional

small variant, UTR5/Upstream(promoter) variant, gene amplification in order from inside to outside.

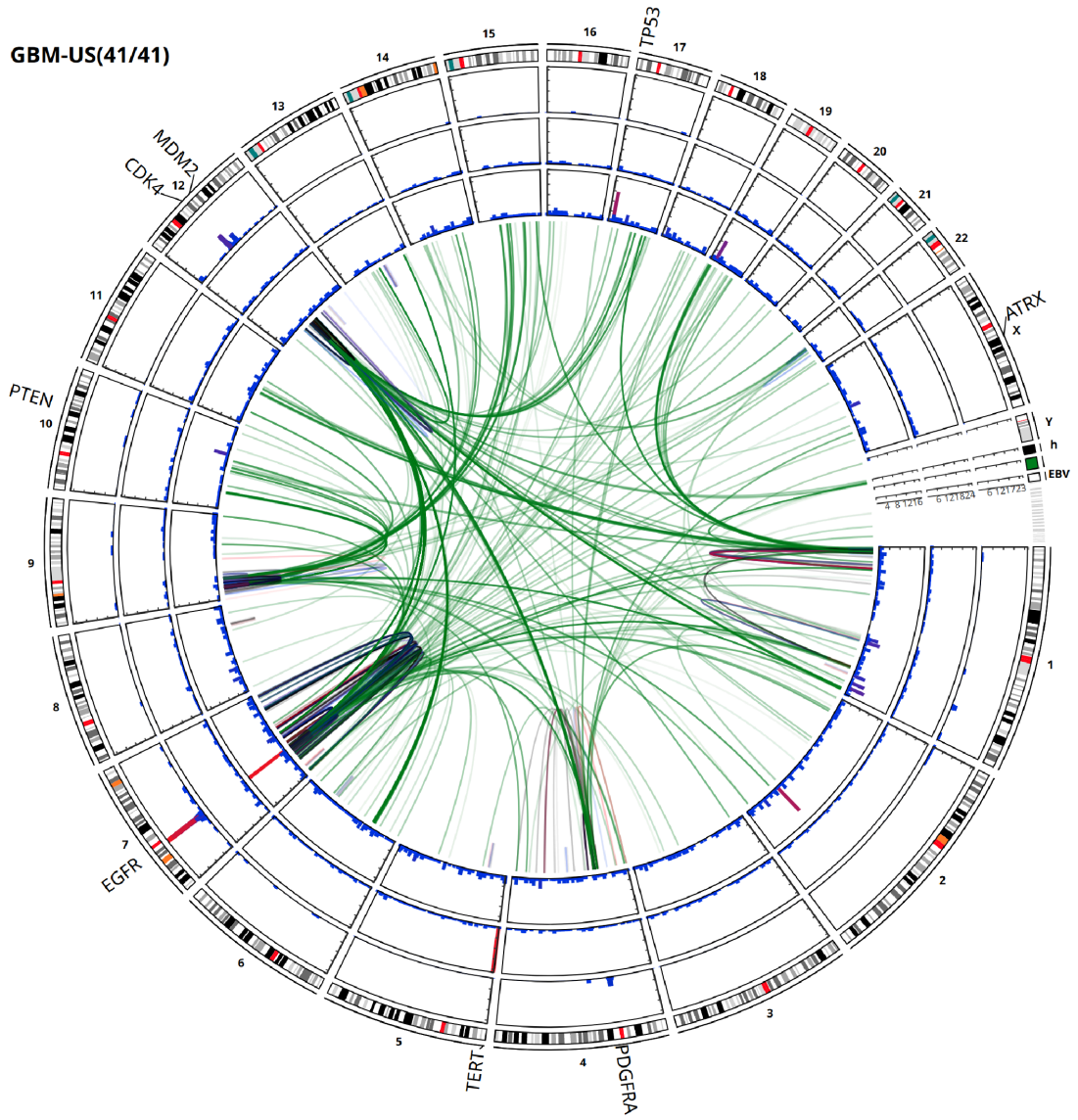

Figure 3.13: Per-gene functional small variant, UTR5/Upstream(promoter) variant, gene amplification recurrence analysis of the TCGA-GBM (WGS) study, in order from inside to outside.

Ultimately, EPISTEME cohort-wide genomic variant landscape plots can be a leading figure in a publication summarizing the main genomic variant classes of a cohort revealing its most important loci of genomic variation. Users can follow this summary and start investigating their datasets across other omics data layers such as the transcriptome, proteome, metabolome and methylome and study possible causal relations between mutational landscapes and cell of origin or transcriptomic dysregulation.

3.3.2 "Single-Phenotype Analysis Plots"

EPISTEME has a number of features for analysing and visualizing phenotype data, such as quantitative omics data. Gene expression is a crucial source of biological information combining influences from the cell type and the oncogenic dysregulation that comes on top of the cell of origin such as homozygous deletions, amplifications or enhancer hijacking events. Thus, studying the transcriptome motivated a number of both dedicated and generalizable visualizations of EPISTEME. This section will show a dedicated representation of the transcriptome and proteome (RPPA) data on the basis of single phenotypic quantities (single genes or RPPA antibodies), integrated with genomic variant information.

The gene expression for *ALK* gene in the GPOH-NB study will serve as a pilot for the demonstration of single-phenotype analysis because it shows both a diversity of genomic variant types and expression levels. *ALK* is an oncogene in multiple cancer types including lymphoma and non-small cell lung cancer, with the remarkable property that its product protein is targetable by small inhibitor molecules [139]. In particular, its role in neuroblastoma is of great interest due to the general lack of targetable alterations in that entity [502].

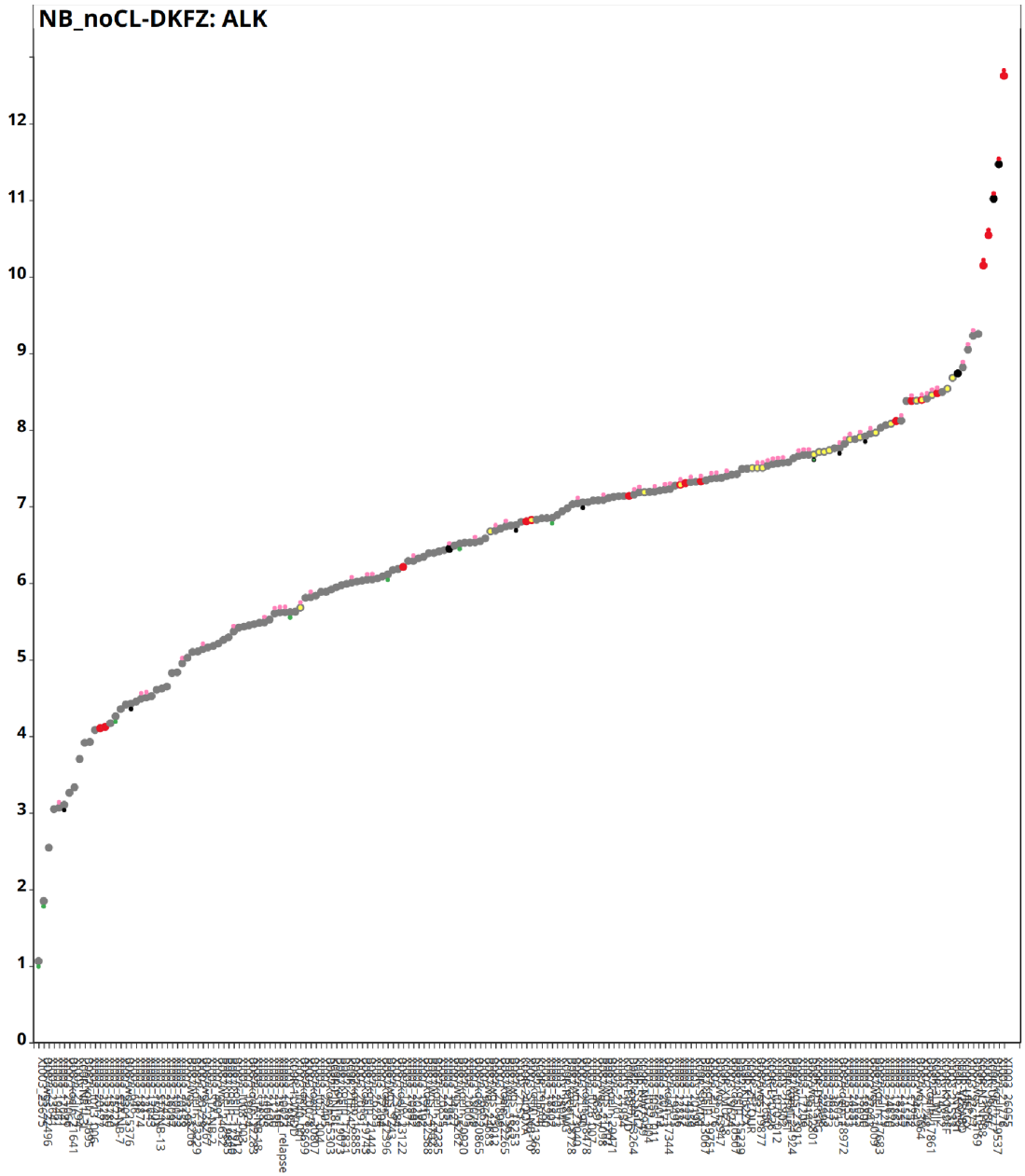

Figure 3.14: Genomic-transcriptomic integrated analysis of the *ALK* gene in the GPOH-NB study. Each symbol corresponds to a donor's gene expression data for the *ALK* gene, where the symbol shapes and colours encode the genomic variant status of *ALK* for the given donor. The Y axis encodes TMM normalized log2 gene expression values for the *ALK* gene.

For each donor, the "x" coordinate is determined by the expression level ("y" coordinate) of the gene of interest, where the ordering is ascending. The genomic variant information is encoded with circles:

- A central circle encoding the expression level and SV status (gray-circle: none,gray-X: no variant data available, black: direct SV hit on the gene body, red: SV hits not directly hitting the gene body, teal: off-gene small indel hit),

- Small central circles encoding different types of small variants (yellow: nonsynonymous SNV, orange: nontruncating/nonframeshift indel, purple: truncating/stopgain/stoploss SNV, pink: splice-site small variant, pink: synonymous SNV),
- Small left-sided circles encoding upstream (red) and 5'UTR (black) variants,
- Small left-sided circles encoding downstream (yellow) and 3'UTR (black) variants,
- Small upper circles encoding CN gain status (pink: low-order gain, red: amplification),
- Small lower circles encoding CN loss status (green: low-order loss, blue: homozygous loss, black: LOH),

. This described colouring scheme allows the packing of a large number of variant classes in an intuitively understandable visualization, where users get a quick overview of the diverse types of variants that affect a particular gene for each patient in a cohort and associate it visually with expression changes. Figure 3.14 shows 5 *ALK*-amplified cases and 19 cases with nonsynonymous SNVs with no overlap. Remarkably, both of these variant types are also associated with higher expression of the gene, which is a nontrivial observation for nonsynonymous SNVs in contrast to amplifications, where such an upregulation is expected and stronger.

As with other EPISTEME visualizations, single-phenotype plots are also interactive. Users can enlarge the default sizes of the circles (risking the emergence of hard to read overlaps between patients), hover on variant circles to find out which patient is showing which type of genomic variant or click on the different variant circles to get detailed annotations on the underlying variant call data that led to the *variant existence* calls (Figure 3.15).

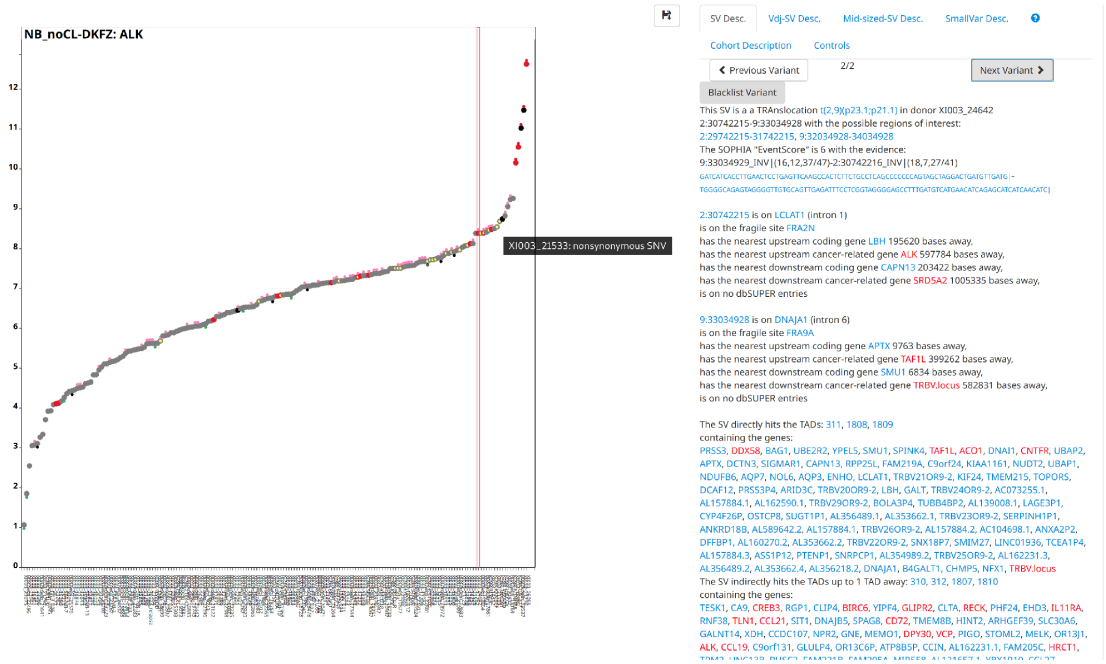

Figure 3.15: Genomic-transcriptomic integrated analysis of the *ALK* gene in the GPOH-NB study with interactions and detailed annotations: a 1.33x enlargement of the default variant circle size is applied, detailed annotations of the SVs of the donor with the ID 24642 and hovering on the yellow circle of the donor with the ID 21533 indicates the yellow circle to be representing a nonsynonymous SNV

Another example that underlines the utility of the described visualization conventions is to investigate two-hit mutations, first shown as a validation of the Knudson hypothesis on the example of Retinoblastoma and the *RB1* gene [207]. The two-hit hypothesis postulates that tumour suppressors require two hits on each allele for inactivation leading to tumour development. Even though this concept has since been partially supplanted by more advanced ideas [503], it still holds true for a number of very important genes such as *TP53* [504] across a wide range of cancer types and *VHL* [505] in renal cell carcinoma.

The gene-centric visualization in EPISTEME is ideal for showing examples of such two-hit processes on a single gene, which we show here on the frequent co-occurrence of copy number losses and nonsynonymous SNVs on *TP53* in Bladder Cancer (Figure 3.16).

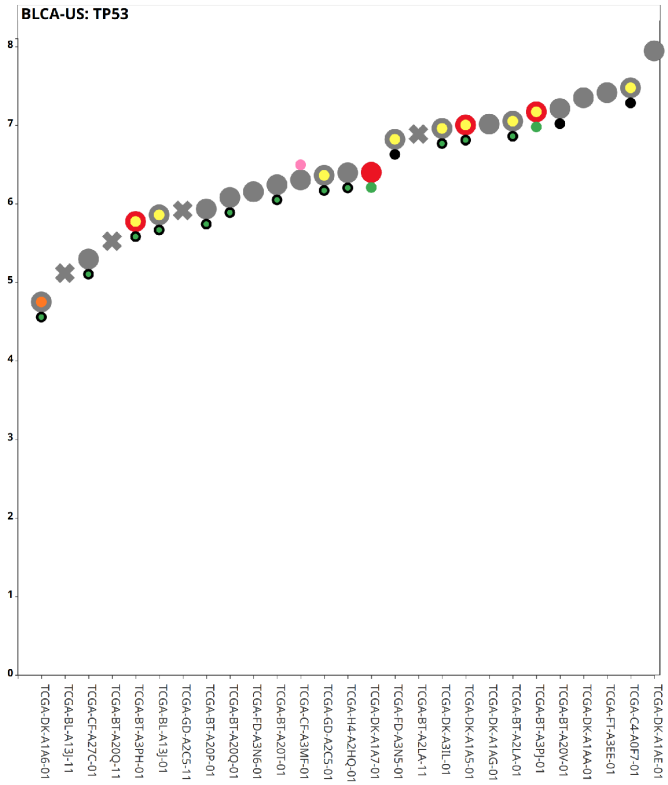

Figure 3.16: Double-hits of copy number losses, LOH and small variants deactivate *TP53* in the TCGA-BLCA (WGS) study. Small lower circles encoding CN loss status (green: low-order loss, black: LOH) co-occur with small central circles encoding different types of small variants (yellow: nonsynonymous SNV, orange: nontruncating/nonframeshift indel), leading to a two-hit deactivation of *TP53*

”Single-Phenotype Analysis Plots” are a fundamentally important visualization in EPISTEME, and are used as auxiliary plots in other visualizations such as the ”Variant-Expression Dysregulation Volcano Plots” (Section 3.3.3) and ”Differential Expression Volcano Plots” (Section 3.3.9.4).

3.3.2.1 ”Two-Phenotype Analysis Plots”

A special case for ”Single-Phenotype Analysis Plots” is the comparison of the expressions of two entities such as two genes. Taking the X-axis again as the anchor for different donors, EPISTEME adds a second Y-Axis to accommodate for the addition of a second gene. This allows the direct comparison of the expression profiles of two genes for each case, showing co-regulation or anticorrelation.

For this visualization, the different activation mechanisms of the *TERT* gene in neuroblastoma are an appropriate pilot showcase. *TERT* is normally suppressed in neuronal cells, including the cell of origin of neuroblastoma as part of neuronal differentiation [395]. In a subset of neuroblastoma cases, *TERT* is known to be upregulated by *MYCN*’s transcription factor activity and/or an enhancer hijacking process with diverse structural rearrangements aberrantly activating the gene [212]. This dual mode of *TERT* activation has been discussed as a major

telomere activation mechanism in high-risk neuroblastoma [506].

We analysed the expressions of the *MYCN* and *TERT* genes in the GPOH Neuroblastoma study (Figure 3.17). This study encompasses low-risk, intermediate-risk and high-risk neuroblastoma, presenting a broad range of disease progression characteristics and disease phenotypes. "Two-Phenotype Analysis Plots" in EPISTEME use one gene as an "anchor" and ranks patients in ascending order with respect to the anchor gene's expression. Here we took *MYCN* as the anchor gene, showing a clearly bimodal characteristic with the *MYCN* amplified subtype with significantly higher expression values. The *MYCN* amplifications and their corresponding SVs are observable, strongly suggesting them to be the underlying cause of the gene's significant overexpression. Overlaying *TERT* expressions using the anchor gene *MYCN*'s expressions defining the order of the donors, shows four modes of *TERT* expression:

- Low *TERT*, exclusively seen in cases with low *MYCN* expression
- *TERT* activation via structural rearrangements (red or black main variant circles), almost exclusively seen in cases with low *MYCN* expression,
- *MYCN*-mediated *TERT* activation,
- non-*MYCN*-mediated, low-level *TERT* activation with an unknown mechanism (gray main variant circles) (as discussed in [212]).

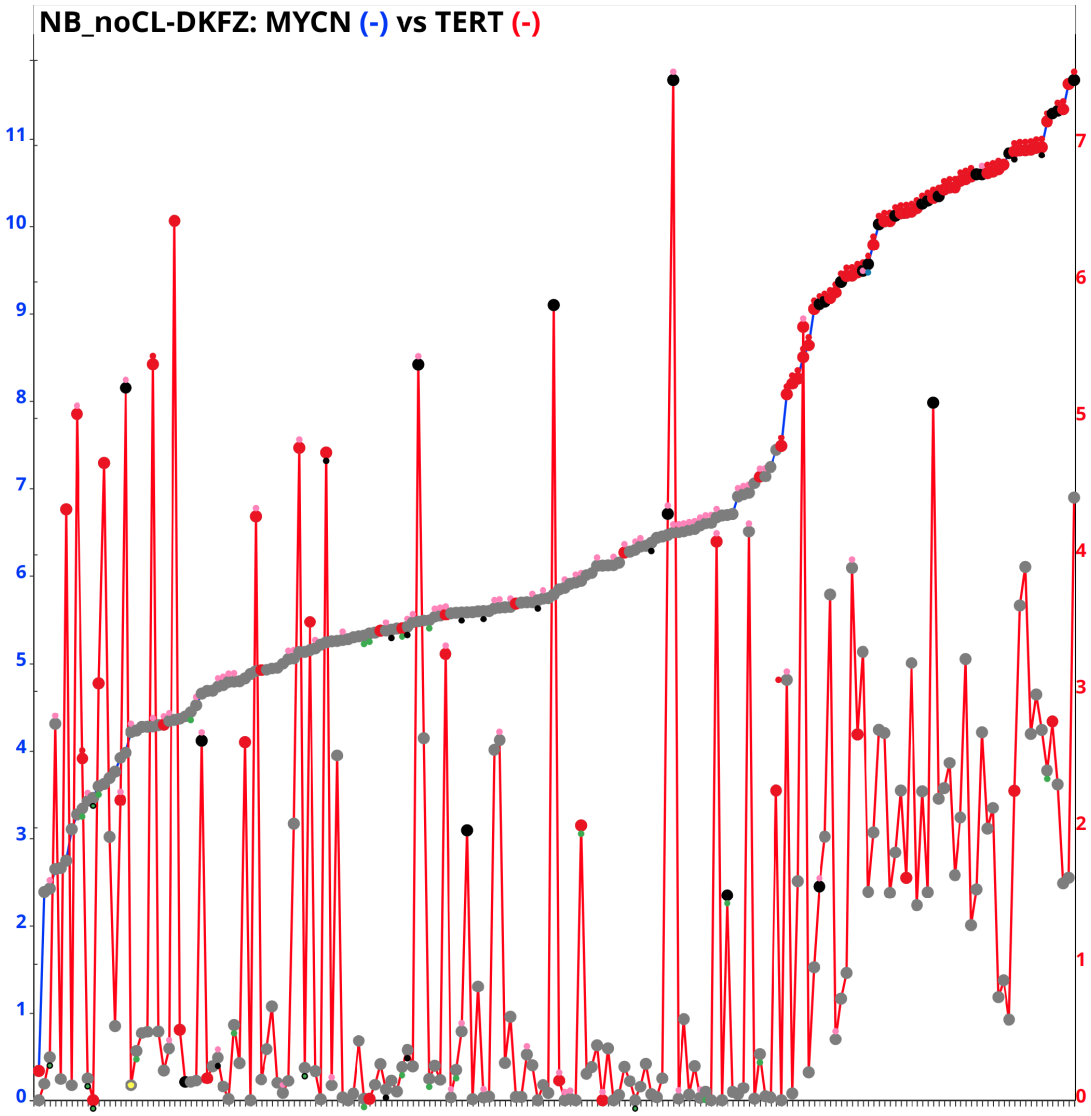

Figure 3.17: Expressions of the *TERT* and *MYCN* genes in the GPOH Neuroblastoma study. Circles representing *MYCN* expressions are connected by blue lines and scaled by the blue axis on the left, whereas circles representing *TERT* expressions are connected by red lines and scaled by the red axis on the right. The anchor gene *MYCN*'s expressions define the order of the donors and for each case, a second point is added to represent TERT expression for the same donor.

"Two-Phenotype Analysis Plots" offer a quick way to compare the per-patient characteristics of gene expressions across a cohort for two chosen genes of interest. They are used as auxiliary plots in visualization of global correlation patterns (Section 3.3.10).

3.3.3 "Variant-Expression Dysregulation Volcano Plots" in EPISTEME

The non-standard differential gene expression analysis strategy described in Section 3.2.8 can be used to identify genes (or proteins) that are differentially expressed due to the existence of selected types of genomic variant classes. The selection of the variant classes is flexible as well

as the type of the applied statistical significance and fold change measures (Figure 3.18). The default settings correspond to the most likely sources of genomic variant related dysregulation, SVs that are up to 1 TAD away from the gene of interest, amplifications and homozygous deletions. The user, however, can decide to add other variant types such as promoter mutations or focus solely on results coming from promoter mutations. The checkboxes here are combined in with the OR logical relation, meaning that a patient with either an SV hit on a gene or an amplification would be considered part of the "variant-positive" group.

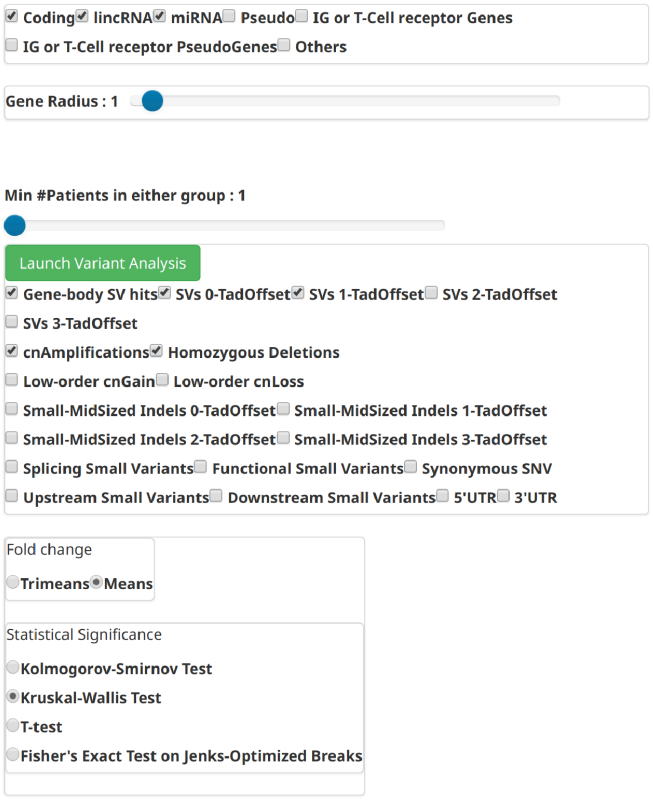

Figure 3.18: Flexible controls for Variant-Expression Dysregulation Volcano Plots in EPISTEME

To showcase this visualization, the TCGA-GBM (WGS) cohort will again serve as a pilot cohort for a proof-of-concept demonstration. As explained in Section 3.2.5, glioblastoma multiforme has amplicons upregulating oncogene expression, namely the *EGFR*, *MDM4*, *CDK4* and *PDGFRA* loci.

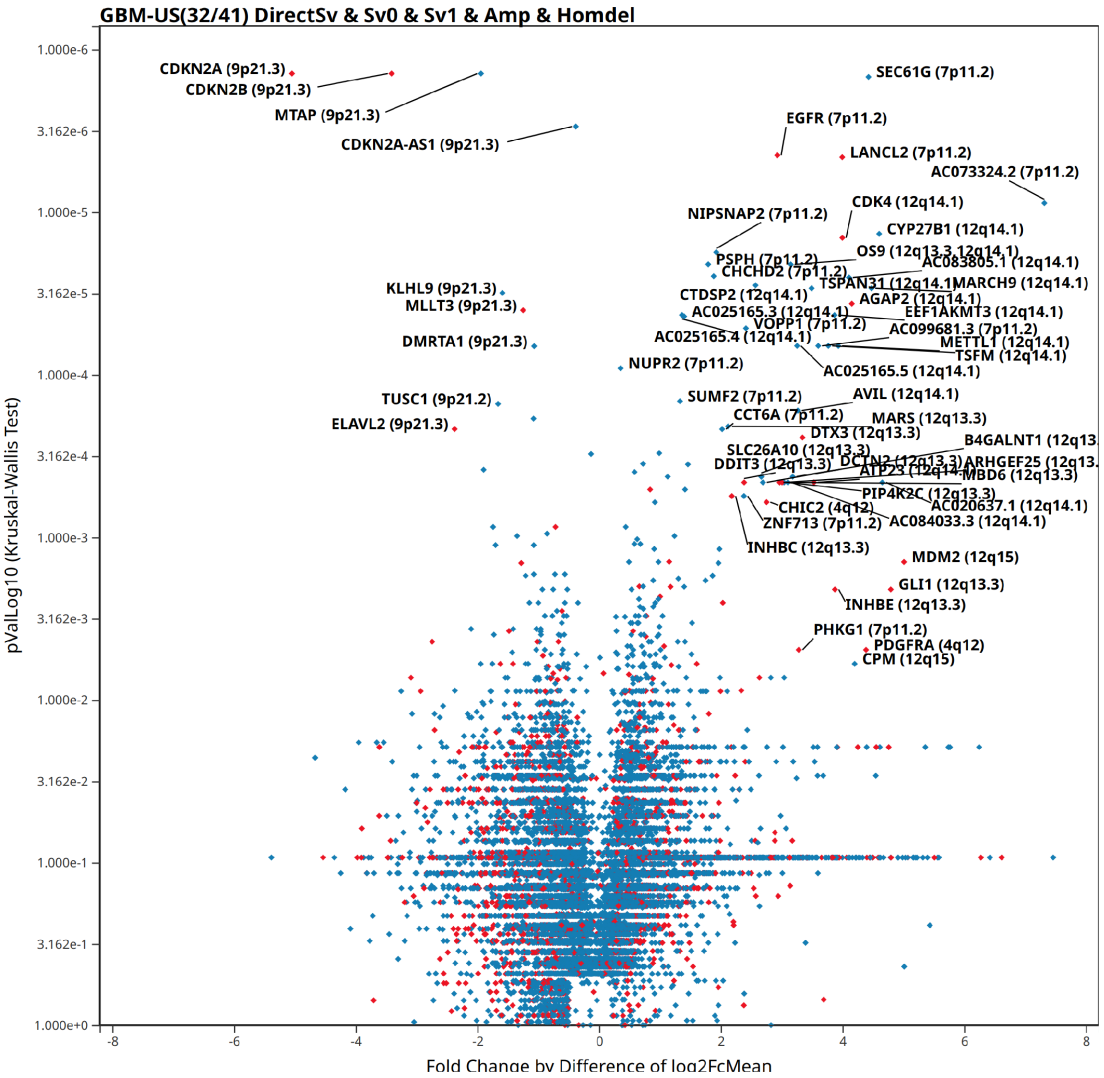

Figure 3.19: The transcriptomic dysregulation landscape of the TCGA-GBM (WGS) study: each symbol corresponds to a gene with genes estimated to be cancer-related coloured red. For each gene, the fold change (difference of log2 of means) and statistical significance (-log10 of Kruskal-Wallis test p Value) measure the differential expression between the cases positive for the selected genomic variant types and negative/wild-type for the selected genomic variant types. The upper-left quadrant shows the genes strongly and significantly downregulated by the allowed genomic variant types. The upper right quadrant likewise shows the genes strongly and significantly upregulated by the allowed genomic variant types

The upper-left quadrant of the plot in Figure 3.19 shows the genes strongly and significantly downregulated by the allowed genomic variant types. As expected, the strongest hits are the *CDKN2A/B* genes and their neighbours. The upper right quadrant likewise shows the genes strongly and significantly upregulated by the allowed genomic variant types. As suggested by the gene labels extended by cytoband information, they cluster remarkably around the expected amplicons *EGFR*, *MDM2* and *CDK4*. In order to facilitate the investigation of rarer and potentially novel gene dysregulation evens, EPISTEME has features to suppress the

display of the genes on user-determined cytobands. Suppressing the display of the genes on the cytobands carrying *EGFR*, *CDKN2A/B*, *MDM2* and *CDK4*, a simplified plot emerges (Figure 3.20). This simplification allows the easier determination of rarer gene activations such as the *PDGFRA* locus and confirms the dominance of the GBM gene activation landscape by common amplicons.

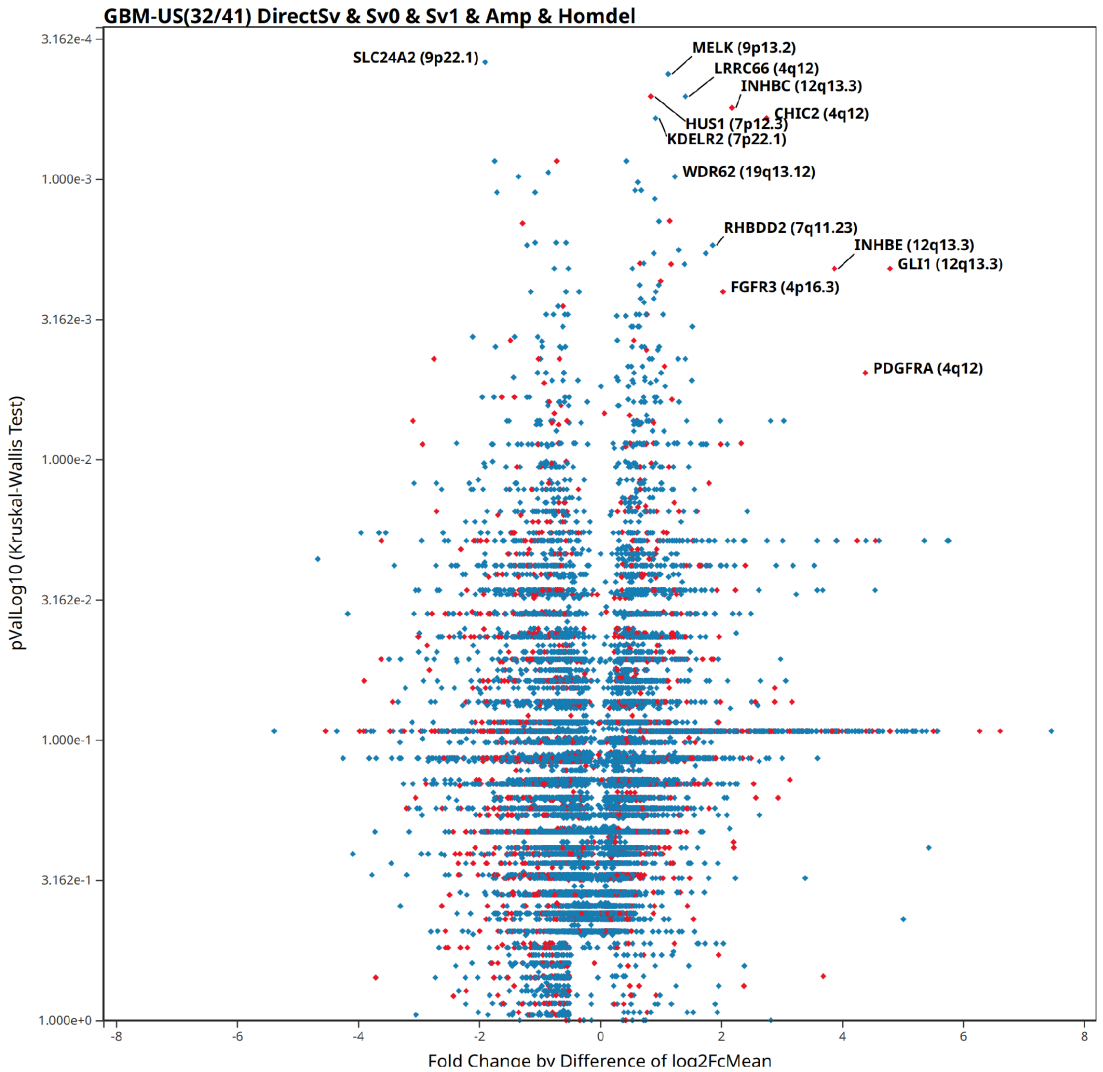

Figure 3.20: The transcriptomic dysregulation landscape of the TCGA-GBM (WGS) study after filtering of the most frequent amplicons (*EGFR*, *CDK4*, *MDM2*) and the *CDKN2A/B* locus followed by a rescaling of the p-value axis yields a highly simplified set of genes indicating the dominance of the hallmark cytobands on the previous genomic variant based transcriptomic dysregulation results.

Non-transcriptomic data such as proteomics data can also be analysed and visualized in the same manner. The TCGA acquired Reverse Phase Protein Array (RPPA) for a large number of cases where material of sufficient quality was available. EPISTEME uses this data in the same manner as transcriptomic data, mapping genomic variants to normalized RPPA readouts

for given proteins (Figure 3.21).

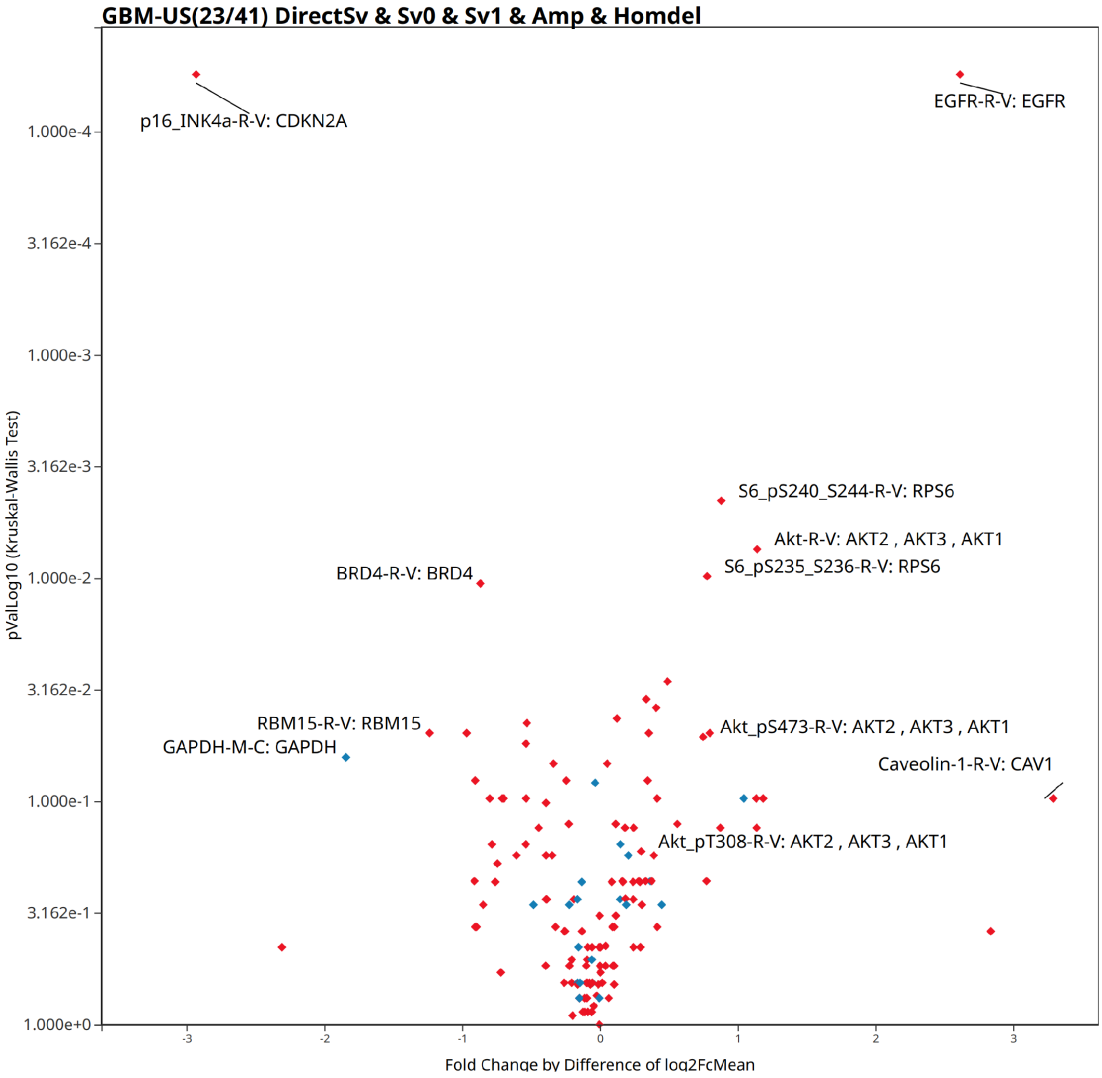

Figure 3.21: The proteomic (RPPA) dysregulation landscape of the TCGA-GBM (WGS) study cohort: each symbol corresponds to a measured antibody in the RPPA array. A strong dysregulation of the probes corresponding to the CDKN2A and EGFR proteins is observed as also expected from transcriptomic data

3.3.3.1 Known and Putative Novel Targets of Enhancer Hijacking in Group 4 Medulloblastoma

We then investigated the capacity of EPISTEME to identify enhancer hijacking events on non-amplified genomic loci in the MB-Group4-DKFZ-PEDBRAIN cohort. In two landmark studies, the genes *GFI1B* [132] *PRDM6* [211] were previously shown to be targets of recurrent activating rearrangements in Group 4 medulloblastoma. The analysis in Figure 3.22 confirms these findings as also presented in Section 2.3.3, showing that the combination of the SOPHIA detection sensitivity and the EPISTEME exploratory statistical procedures described here cor-

rectly reveal *MYCN*(9 cases, co-activated with *DDX1*), *PRDM6*(12 cases, co-activated with *PPIC* and *SNX24*) and *GFI1B* (3 cases). Furthermore, our analysis suggests novel, recurrent, but rare activations including *PKDCC* (3 cases), *INHBA* (3 cases) and *SOX8* (1 case). In particular, *INHBA* of the Inhibin gene family [507] [508] and *SOX8* of the Sox gene family [509] [510] have important roles in cancer development, including in cancers of nervous system cell of origin, hence they are of great interest in the context of medulloblastoma. Remarkably, the *INHBC* and *INHBE* genes of the inhibin family located close to the *CDK4* amplicon are also upregulated in GBM suggesting a pan-cancer significance of these gene activations.

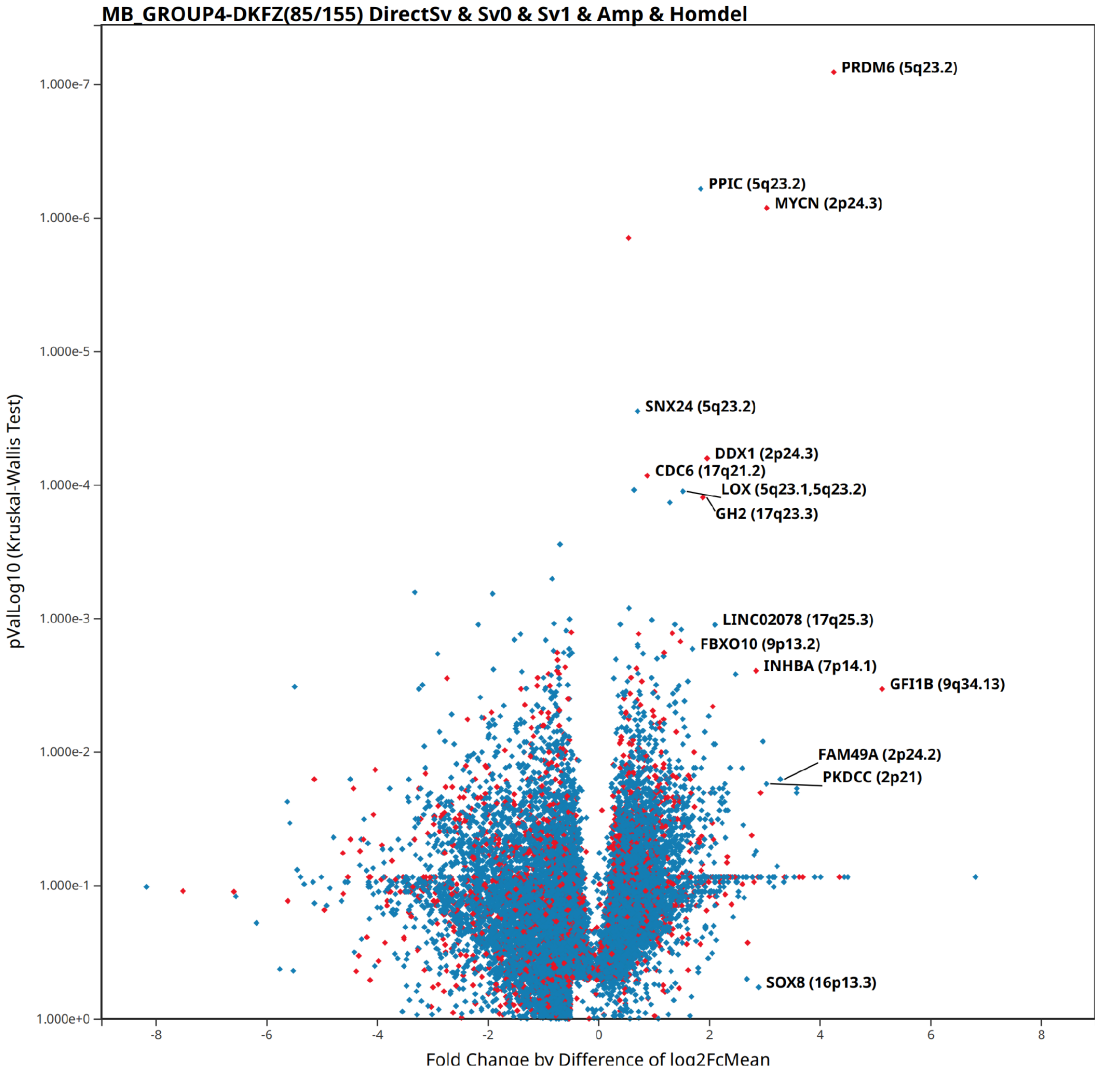

Figure 3.22: The genomic variant-caused transcriptomic dysregulation landscape of the MB-Group4-DKFZ-PEDBRAIN study.

3.3.3.2 The Landscape of Transcriptomic Dysregulation in Gastric Adenocarcinoma Reveals Novel Candidates for Enhancer Hijacking

After inspection of all TCGA cohorts with available WGS and RNA-Seq data, we decided to use the TCGA-STAD study [511] for demonstrating the discovery of novel candidates for

enhancer hijacking in EPISTEME.

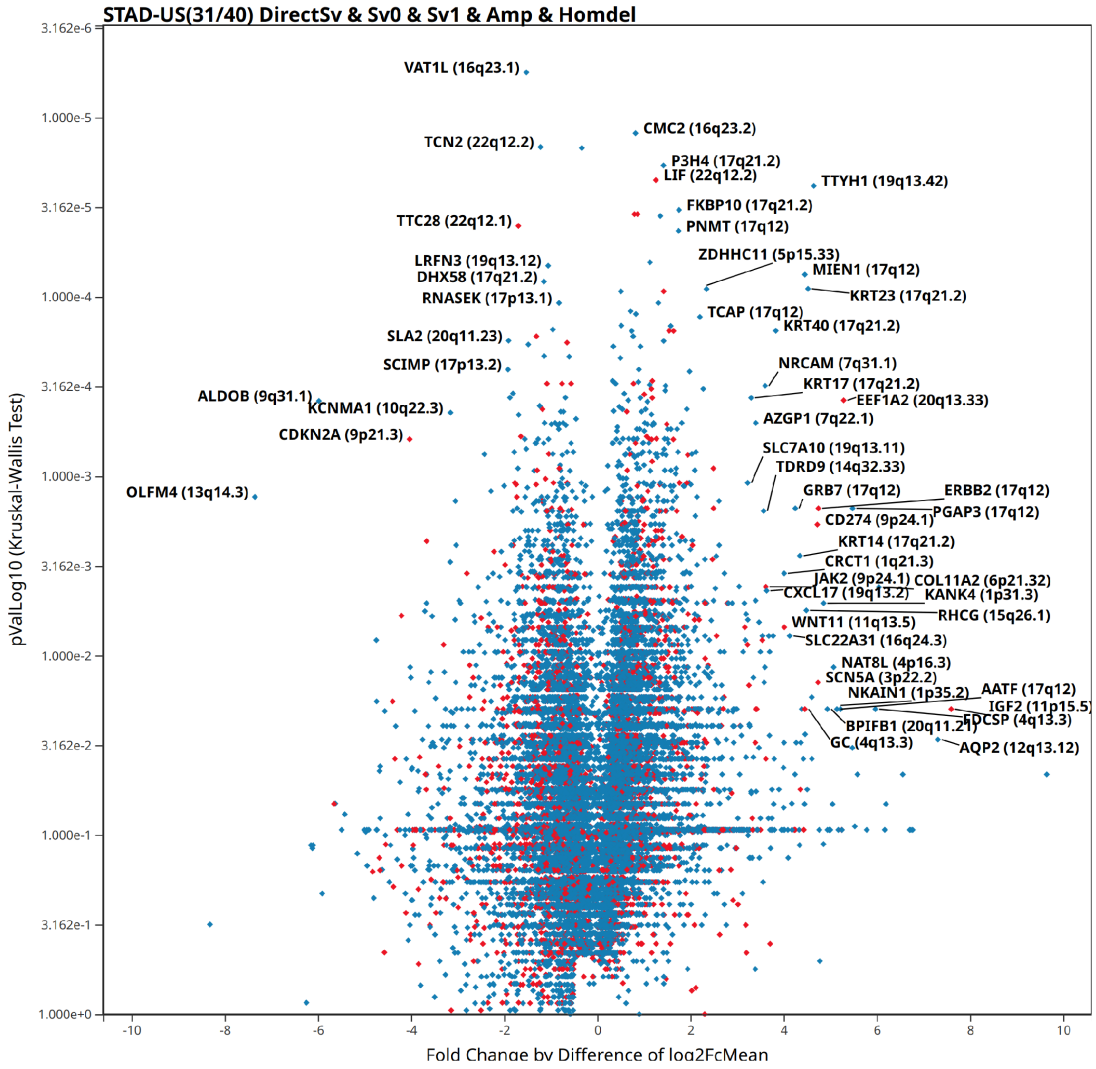

Figure 3.23: The transcriptomic dysregulation landscape of the TCGA-STAD (WGS) study

The transcriptomic dysregulation landscape of the TCGA-STAD cohort shows an interesting diversity of non-amplicon driven targets of upregulation as (Figure 3.23). The sole exception to this general trend is the *ERBB2* gene on chr17q12, previously shown to be a target of recurrent amplifications in this entity among other cancer types [512]. We identified the oncogenes *IGF2*, *JAK2* and *CD274* (*PD-L1*) to be activated by amplifications and non-amplified rearrangements in 3 cases each, where the *JAK2* and *CD274* activations correspond to co-activations in (Epstein-Barr Virus) EBV+ gastric adenocarcinoma.

Novel genes showing upregulation co-occurring with non-amplification rearrangements included FKBP10 (12 cases upregulated, 10 with detected SVs: 10/12), *AZGP1* (5/7), *WNT11* (4/4), *BPIFB1* (3/3), *KANK4* (3/3), *GC* (2/2), *TDRD9* (2/2) and *NKAIN1* (2/2). Of these, *WNT11* has clear potential for biological significance due to the roles of WNT signalling in Gastric Adenocarcinoma.

A particularly strong outlier worthy of mention was the *TTYH1* gene which is a neuronal adhesion molecule previously shown to be driving tumor microtube (TM)-mediated brain colonization by glioma cells [513]. *TTYH1* was upregulated in 6 cases where all (6/6) co-occurred with non-amplification rearrangements. Remarkably, we also identified 7 cases with upregulated *NRCAM*, a gene with similar functions in the context of neuronal adhesion, of which 6 up-regulations co-occurred non-amplified rearrangements (6/7). The activation profiles of the *TTYH1*, *NRCAM*, *KANK4* and *TDRD9* genes are shown in Figure 3.24.

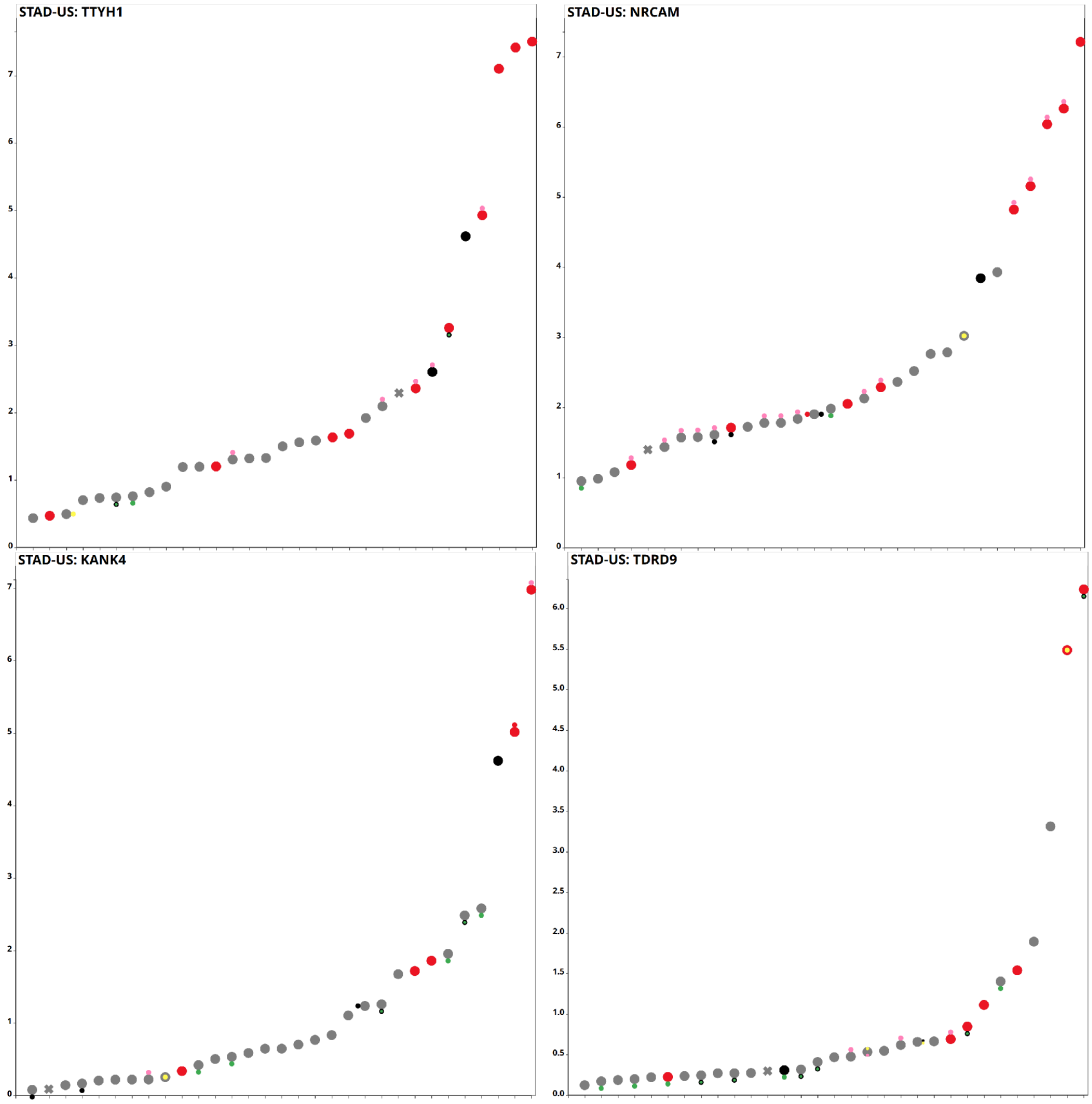

Figure 3.24: Activation of the *TTYH1*, *NRCAM*, *KANK4* and *TDRD9* genes in the TCGA-STAD (WGS) study by structural variants (red and black symbols indicating offset SV hits and direct SV hits on gene bodies, respectively) leading to significant overexpression of the genes compared to the cases without SVs (gray symbols)

NRCAM was previously claimed to be correlated with poor prognosis in colorectal adenocarcinoma [514] and gastric adenocarcinoma [515]. Remarkably, the 13 cases only had one overlapping case with both genes upregulated, suggesting a mutual exclusivity pattern. Al-

though a ”neuronal subtype” in the sense of the bladder cancer subtype with the same name [516] does not exist in gastric adenocarcinoma, the concept of neuroendocrine differentiation does exist [517] [518]. Furthermore, it was recently shown that cancer cell-nerve interactions in the gastric tumour microenvironment promote tumourigenesis by upregulation of WNT signalling [519]. Interestingly, the latter study followed earlier work demonstrating the tumour suppressing effects of denervation in gastric cancer [520]. While we cannot yet propose a model connecting these concepts to the upregulation of *TTYH1* or *NRCAM*, potentially due to non-amplified rearrangements, EPISTEME provides valuable starting points for hypothesis building in an easily accessible manner.

3.3.3.3 Interactive features of Volcano Plots in EPISTEME

As with other visualizations, ”variant-expression dysregulation volcano plots” are highly interactive, offering modifications of the visualization, and integrations to databases to facilitate the exploration of the results.

1. The users have the free choice of which variant types to include in the analysis, which gene types to show, which statistical significance and fold change measures to use (Figure 3.18)

2. Individual genes or groups of genes on user-selected cytobands can be hidden from view to facilitate discovery of novel candidate genes by suppressing contributions from known loci

3. EPISTEME allows users to label genes with draggable labels to improve readability (Figures 3.19, 3.20, 3.22 and 3.23). Gene labels can be clicked on to go to GeneCards.

4. The list of genes labelled at any time, can be quickly sent to external resources for pathway enrichment / gene set analysis (*DAVID*, *Reactome*, *GSEA*, *CPDB*).

5. Genes can be labelled by clicking on the data points on the volcano plot or explicitly naming the gene to label in a gene selector. The explicit naming of a gene is useful for finding a gene of interest, particularly if it is not a strong outlier data point.

6. Multiple genes on a cytoband can be labelled by choosing any gene on the cytoband of interest or explicitly naming the cytoband carrying the genes to label in a cytoband selector.

7. Multiple genes can also be labelled by dragging a selection box on the volcano plot

8. Clicking on a single data point, in addition to labelling the selected gene (or protein etc.), launches an auxiliary instance of the Single-Phenotype analysis plot described in Section 3.3.2, (Figure 3.25). This allows the users to get a quick overview on the underlying data for a given data point in the volcano plot. One can thus investigate why a particular point is an outlier or not, and which variant types are responsible for a possible dysregulation.

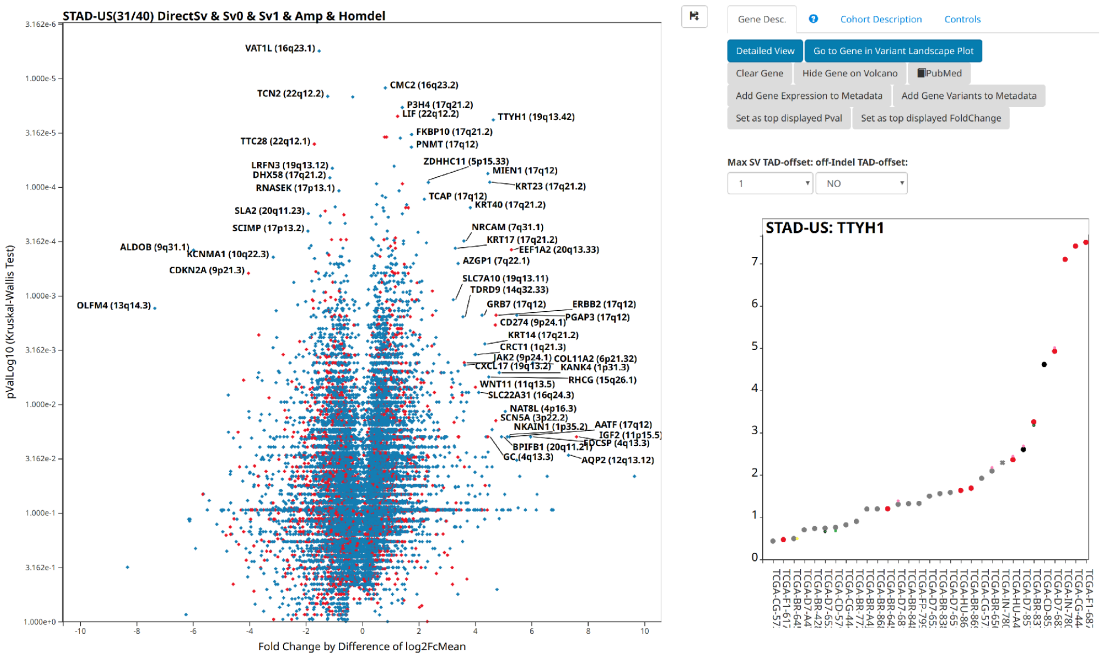

Figure 3.25: The transcriptomic dysregulation landscape of the TCGA-STAD (WGS) study extended by the an auxiliary Single-Phenotype analysis plot for *TTYH1*. Clicking on a gene on a volcano plot launches an auxiliary single-phenotype analysis plot, showing the underlying gene expression data for the gene of interest.

Auxiliary Single-Phenotype analysis plots are integrated to PubMed for checking the novelty of a candidate gene in the context of the studied disease. Users also have the option to restrict the main volcano plot's axis maxima based on the X or Y coordinates of the selected data point. Finally, auxiliary Single-Phenotype analysis plots can be directly updated to a full Single-Phenotype analysis plot with full interactions described in Section 3.3.2 including the integration of the underlying genomic variants behind the data point. This completes the connections from the global overview of dysregulation to the dysregulation profile of a single data point and finally the genomic variants leading to the observed dysregulation.

9. In plots where RPPA values are visualized, clicking on a given RPPA, labels all different RPPA readouts for the gene that the RPPA belongs to (accounting for different protein isoforms or post-translational modifications). Despite this multi-labelling, EPISTEME only launches the auxiliary single-phenotype analysis plot for the selected RPPA, while offering the other possible RPPAs for the same gene in a drop-down menu.

Interactive volcano plots are another central visualization in EPISTEME, and are used for other types of analysis than genomic variant-driven transcriptomic/proteomic dysregulation such as "Differential Expression Volcano Plots" (Section 3.3.9.4), "Variant Mutex/Co-Occurrence Plots" (Section 3.3.9.2) and "Expression Correlation Plots" (Section 3.3.10).

3.3.4 Managing and Classifying Quantitative and Categorical Data in a Cohort Study in EPISTEME

EPISTEME parses, processes, automatically categorizes, and flexibly visualizes quantitative and categorical data in a cohort study. The input data is a spreadsheet where each row corresponds to a patient/donor and each column corresponds to a variable. During cohort initialization, each donor-variable data point is tested for being a number or not, also taking into account typical missing value strings ”NAN”,”NaN”,”NA”,”N.A” and ””(blank). Variables with only numeric or missing values are considered to constitute quantitative data, whereas variables that do not fulfil this condition constitute categorical data. There is a third, internal, variable type in EPISTEME, called ”multi-categorical”. The semicolon character ”;” is considered a special character in EPISTEME separating different values for a given donor and variable. For instance, the value ”MYC;CCND1” in a ”multi-categorical” ”V(D)J target” column for a given patient would indicate that both *MYC* and *CCND1* genes are targeted by detected V(D)J / CSR rearrangements for that patient. This allows the consideration of a case to belong to multiple categories.

The quantitative or categorical nature of the underlying data determines the subcohort selection and data visualization features offered in EPISTEME. The appropriate routine is selected by EPISTEME in the background with no user involvement, and will be described in Section 3.3.5 for visualizations and Section 3.3.8 for selections.

EPISTEME gives users an overview of the quantitative and categorical metadata with a sortable and searchable spreadsheet with a ”frozen” first column and row, to facilitate the preliminary exploration of the data. The spreadsheet (Figure 3.26) can be extended by user-selected data such as chromosome arm variants, cytoband variants, gene expressions, gene variants and RPPA expressions, which leads to the automatic classification of the added data as quantitative or categorical and an update of the spreadsheet (Figure 3.27). These user-selected data can then be used for all subcohort selection and data visualization features of EPISTEME in the same manner as the original metadata of the cohort.

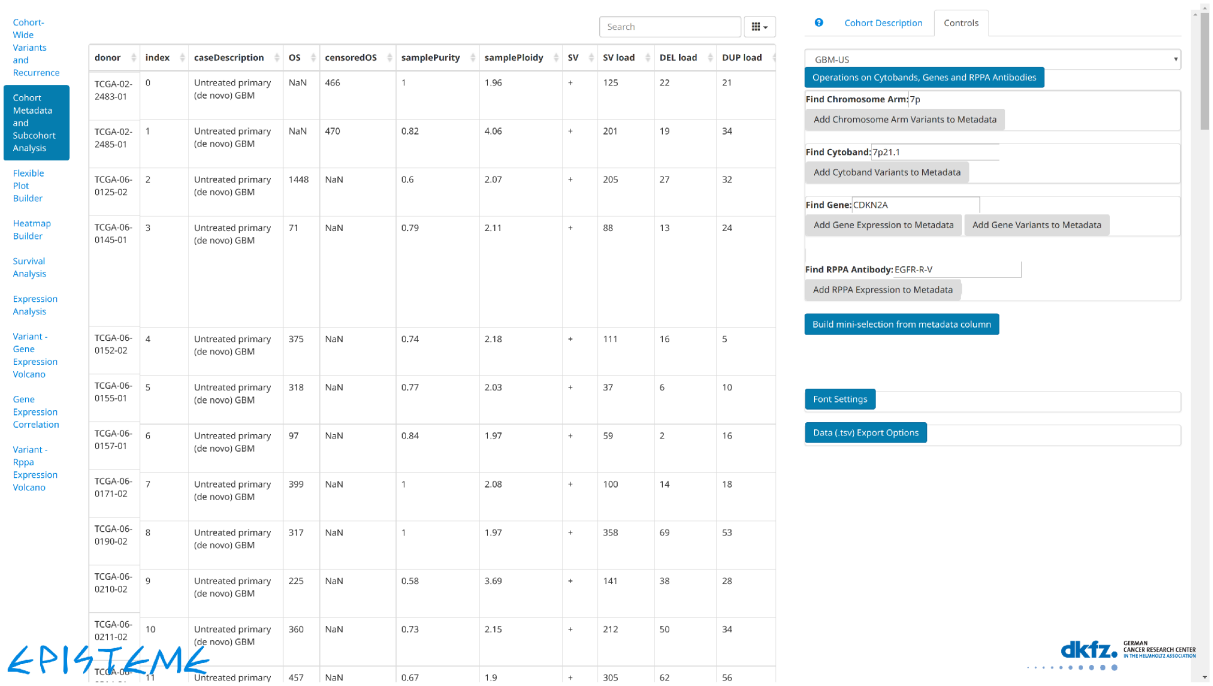

Figure 3.26: The metadata spreadsheet of the TCGA-GBM (WGS) study with the demonstration of the possible extension features: chromosome arm level copy number changes, cytoband level copy number changes or SV hits, gene expressions, gene variant status, RPPA probe quantities can be added on-demand to the cohort metadata table.

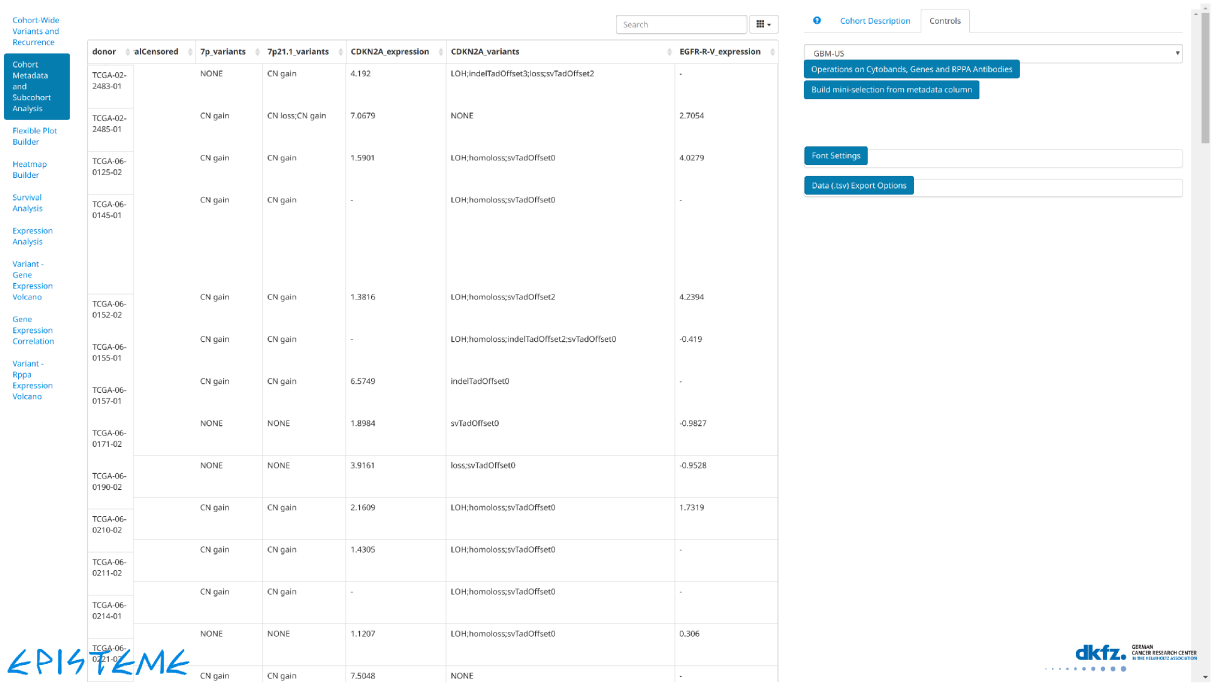

Figure 3.27: The metadata spreadsheet of the TCGA-GBM (WGS) study with the manual extensions for chr7p variants, chr7p21.1 variants, *CDKN2A* expression, *CDKN2A* variants, and EGFR-R-V RPPA expression

Categorical data is often used to designate disease subtypes, dividing a main cohort into individual, non-overlapping, subcohorts. Subcohorts can be defined based on methylation profiles, gene expression profiles, mutation status or tumour histology. To showcase EPISTEME's features on the management, analysis and visualization of categorical data as well as the presentation of its dimensionality reduction, and subcohort-specific features, the PAM50 classification of Breast cancer is an excellent example due to a number of reasons:

1. Breast cancer is among the most common malignancies, and constitute therefore a significant and increasing clinical challenge affecting western populations [521]. With this motivation, the TCGA's breast cancer project has been designed as a large study with excellent availability of mutation, gene expression, methylation and clinical metadata [522]

2. Breast cancer is known arise from two main cell types with the hallmark of oestrogen receptor expression (Luminal ER+, and Basal-like ER-) with gene expression profiling yielding a finer classification of 4 main subtypes [247] (Luminal A, Luminal B, Her2, Basal-like) and a secondary gene signature of normal tissues and impure tumour specimens (Normal-like).

3. None of the main 4 subtypes are exceedingly rare, which gives high statistical power to any analysis conducted with these designations.

3.3.5 Flexible 2D Plots in EPISTEME

EPISTEME offers a 2D plot grid where 2D subplots of different sizes can be created with individual parameters. The individual subplots encode different data types in the X-axis, Y-axis, colour, symbol and radius, flexibly chosen by the user:

1. X and Y axis data are mandatory and can either be quantitative or categorical. A donor with missing data on either, will not appear on the plot as a data point.
2. Colour-assigned data can either be quantitative or categorical. If quantitative, a donor with missing data on the column used for the colour encoding, will not appear on the plot as a data point. If categorical, missing values will be assigned their own colour.
3. Symbol data must be categorical. Missing values will be assigned their own symbol. There are up to 7 available symbols, so categorical variables with more than 7 different values are not encodable by symbols.
4. Radius data must be quantitative. A donor with missing data on the column used for the radius encoding, will not appear on the plot as a data point.

For figures with encoded colour, symbol or radius data, EPISTEME generates a legend that describes the data assigned to each encoding. Figure 3.28 is a demonstration of a multiplot including the following features on the TCGA-BRCA (WGS) cohort:

- Multiplots with different sizes
- X,Y,colour,symbol,radius encoding
- Legend for colour, symbol , radius encodings
- Box-KDE-Jitter plots categorical-quantitative data in two orientations
- Stacked bar chart for categorical-quantitative data for the special case where the ”donor” entry corresponds to the categorical data selection
- Scatter plot for quantitative-quantitative data

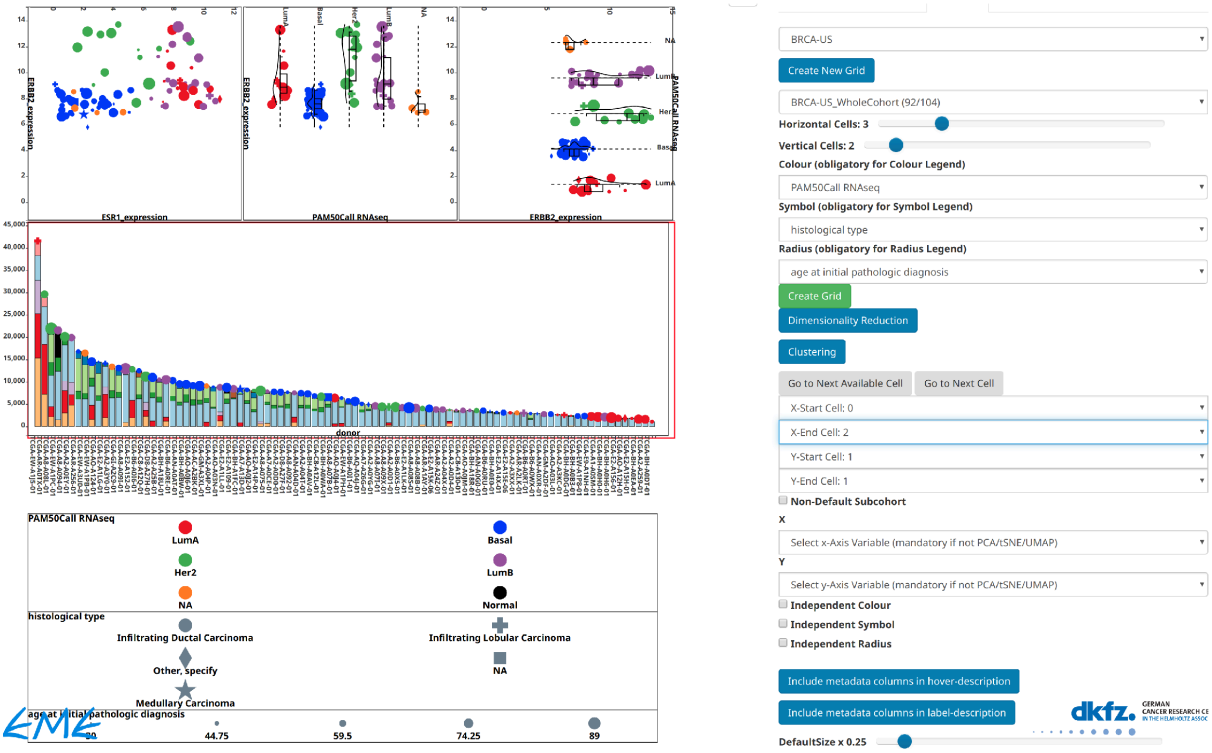

Figure 3.28: Demonstration of box-KDE-jitter, stacked bar chart, and scatter plots in a nulti-plot grid on the TCGA-BRCA (WGS) cohort in EPISTEME. Right panel: settings for the setup of the grid with default colour, symbol shape and symbol size variables. Left panel, top-left: scatter plot example showing *ESR1* expression vs *ERBB2* expression. Left panel, top-middle: vertical box-KDE-jitter plot example showing *ERBB2* expression across PAM50 subtypes. Left panel, top-right: horizontal box-KDE-jitter plot example showing *ESR1* expression across PAM50 subtypes. Left panel, middle row: absolute estimated mutational signature contributions for COSMIC mutational signatures. Left panel, bottom row: legend for the colour, symbol shape and symbol size variables.

The following sections describe and demonstrate the different plot types offered in this flexible 2D data visualization concept.

3.3.5.1 Box-KDE-Jitter Plots Visualizing Categorical-Quantitative Data in EPISTEME

EPISTEME visualizes the relation of categorical data to quantitative data using a combination of standard plots: 1. Half-box plots showing the 10th, 25th, 50th, 75th and 90th quantiles of each group, 2. Kernel Density Estimation (KDE) plots showing the distribution of the values across each group, 3. Jitter plots showing the individual data points across each group,. The dual approach for summarizing data allows both the observation of the quantile values as well as the investigation of possible multimodalities in the groups thanks to the KDE plot.

This visualization can be created in either horizontal or vertical orientation (Figure 3.28, Figure 3.29 showing the vertical alternative on the *ERBB2* gene expression across PAM50 subtypes in the TCGA-BRCA study).

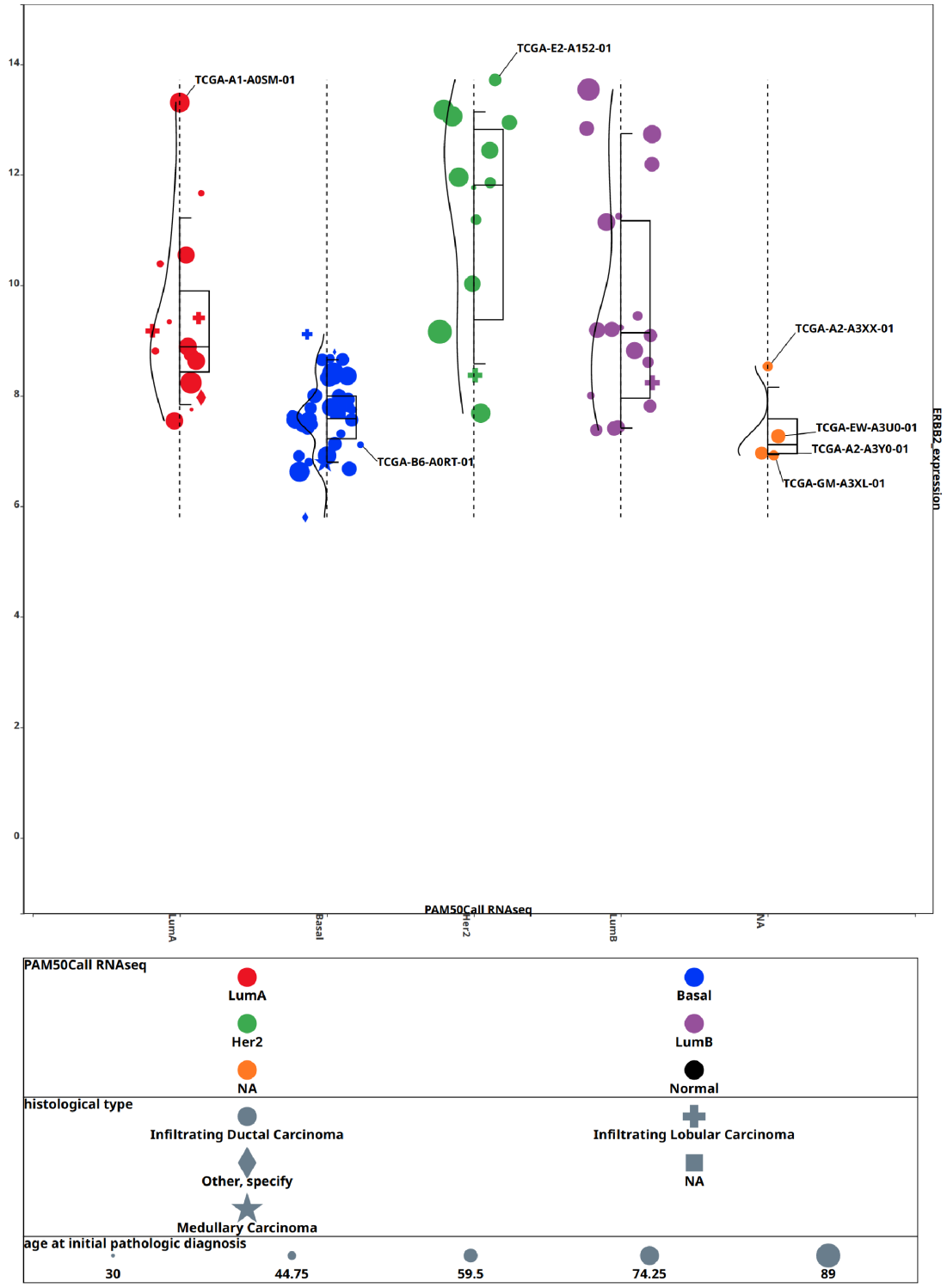

Figure 3.29: Vertical Box-KDE-Jitter Plot showing the expression of the *ERBB2* gene across PAM50 subtypes in the TCGA-BRCA study

As expected from the established knowledge on breast cancer, the Luminal A and B subtypes express high *ESR1*, the Her2 subtype can either express low or high *ESR1* and the Basal-like subtype expresses low *ESR1*.

The *ERBB2* gene shows a bimodal characteristic in the Luminal B but interestingly also in the Her2 subtype. The low-expressor Her2 cases suggest a similar cell of origin that led to the Her2 PAM50 classification but low *ERBB2* expression due to low purity or another influence. As with the *ESR1* gene, it is invariably expressed with low levels in the Basal-like subtype.

For both genes, the cases with available RNA data but no available PAM50 classification (due to lack of metadata) seem to cluster with the Basal-like subtype.

Flexible 2D Plots in EPISTEME are highly interactive like with its other features. Each data point corresponds to a donor, and donors can be marked using draggable labels either by clicking or explicit donor selection (Figure 3.29). Multiple donors can be marked to define, multi-mark or shade a subgroup, which will be discussed in Section 3.3.8.

3.3.5.2 Stacked Bar Charts Visualizing Categorical-Quantitative Data in EPISTEME

A special case in the analysis of categorical-vs-quantitative data arises when the categorical data column corresponds to the ”donor” data, meaning that for each donor, there can be only one data point and each donor thus corresponds to a category/group in the categorical variable. This special configuration can be used to visualize more than one quantitative variable using ”stacked bar charts”. Upon the selection of either the X or Y axis field as ”donor”, EPISTEME switches to the stacked bar chart setup mode and offers the users to select one or more quantitative data fields for visualization on the stacked bar chart (Figure 3.30). The stacked bar chart mode recognizes all quantitative data fields as possible inputs and offers special checkboxes for mass-selection of mutational signatures for either relative contributions or absolute contributions, so that the users do not have to click on each signature contribution up to 31 times (30 COSMIC signatures + unknown/other contributions).

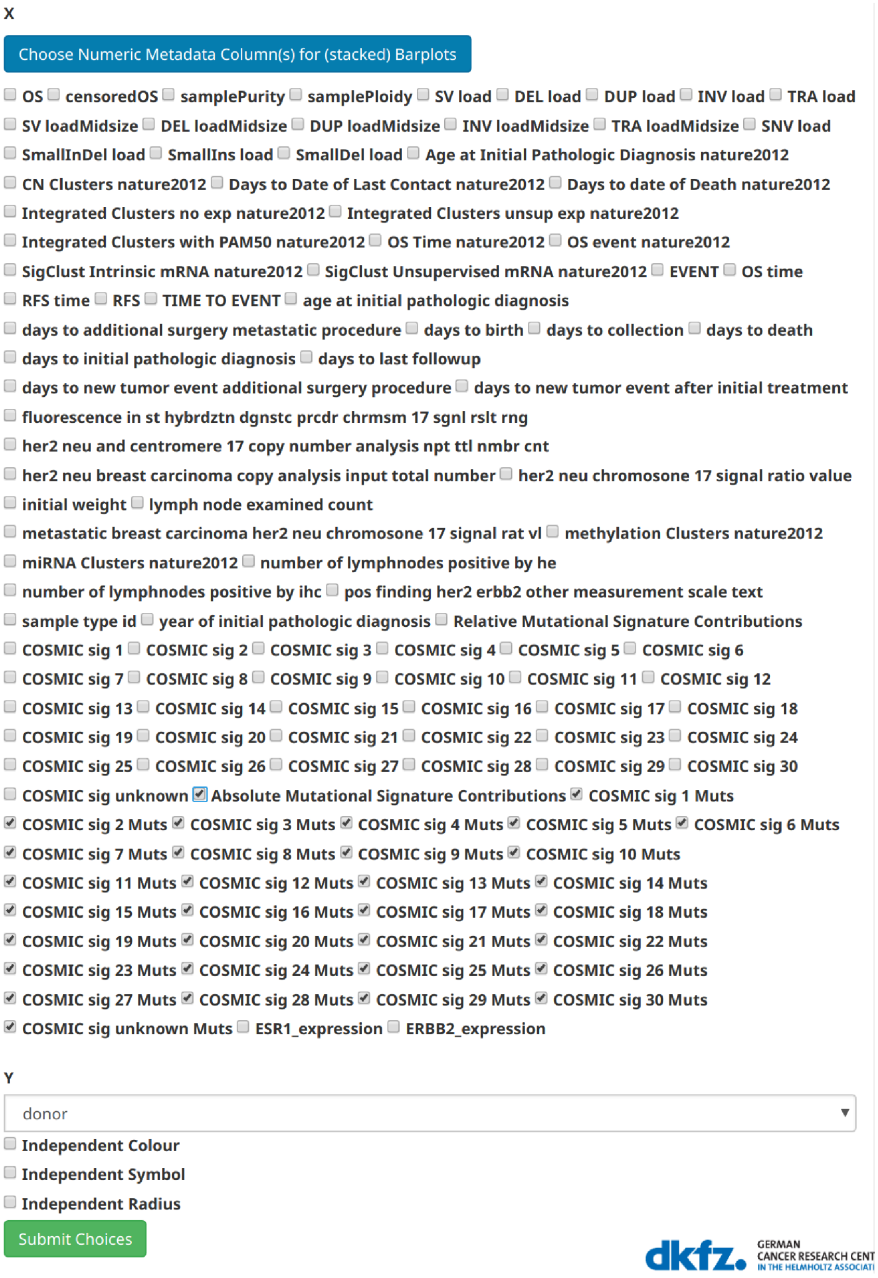

Figure 3.30: Horizontal stacked bar chart setup for the visualization of COSMIC signature contributions for each donor in the TCGA-BRCA (WGS) study.

Figure 3.31 shows a stacked bar chart in horizontal orientation representing the estimated absolute contributions of each COSMIC mutational signature in the TCGA-BRCA (WGS) study. What is non-standard compared to normal stacked bar charts is that each bar stacked is capped by a symbol with modulated shape, size and colour as with the other EPISTEME flexible 2D plot types. Thus, each case can be labelled and the data encoding features in terms of symbol type, symbol size and symbol colour are fully taken advantage of.

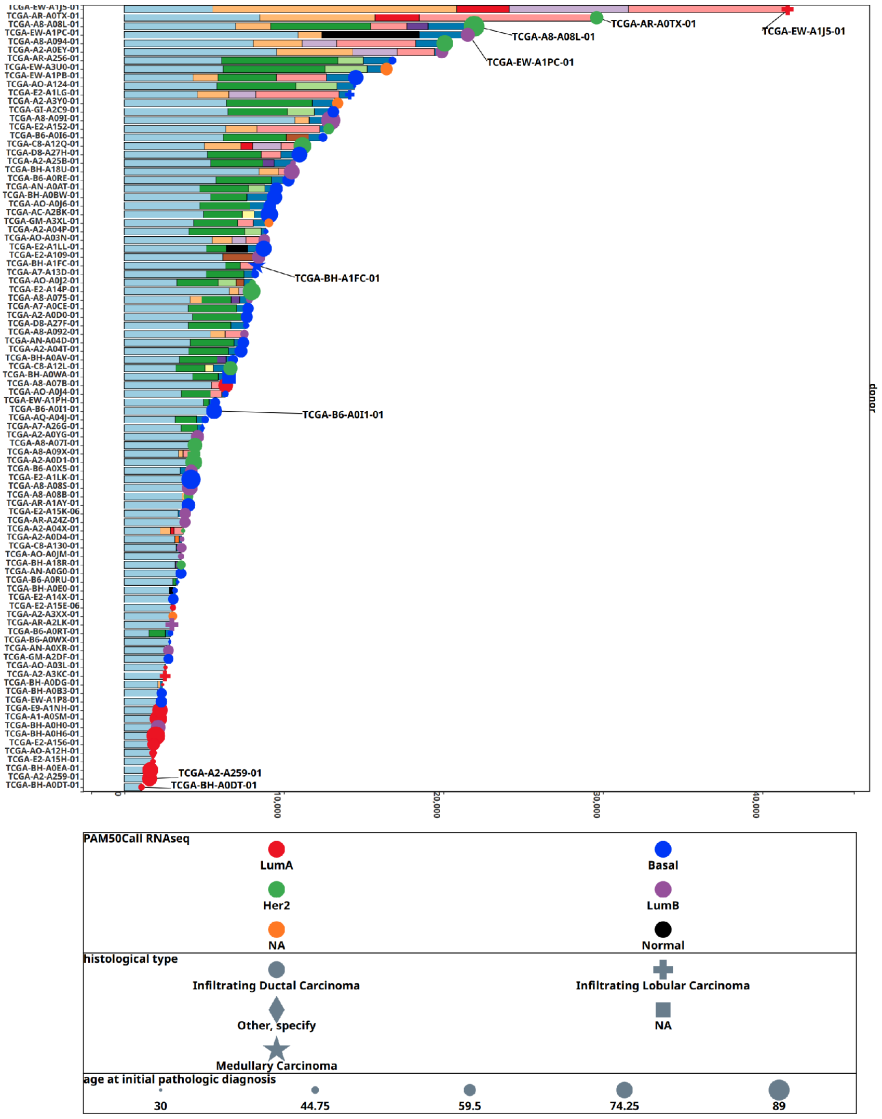

Figure 3.31: Horizontal stacked bar charts showing the COSMIC signature contributions for each donor in the TCGA-BRCA (WGS) study. The heights of the bars correspond to estimated mutation counts for a given COSMIC signature. Overall, all cases show the ageing-related signature AC1 (light blue), followed by DNA double-strand break repair defect signature AC3 (green) for cases with intermediate-level and above SNV counts. A strong contribution of the APOBEC signatures AC2 (light orange) and AC13 (salmon) is observed among the top cases with respect to somatic SNV load. The top case also has the AC10 (light purple) signature which is associated with *POLE* mutations [280], and might explain the high mutational load of this case.

3.3.5.3 Scatter Plots Visualizing Quantitative-Quantitative Data in EPISTEME

Relation of quantitative data to quantitative data is most naturally visualized as scatter plots. The upper left subplot of Figure 3.28 shows the expressions of the *ESR1* gene vs the *ERBB2* gene in breast cancer, whereas Figure 3.32 shows the visualization of this relation in a more detailed manner outside of a multiplot setting.

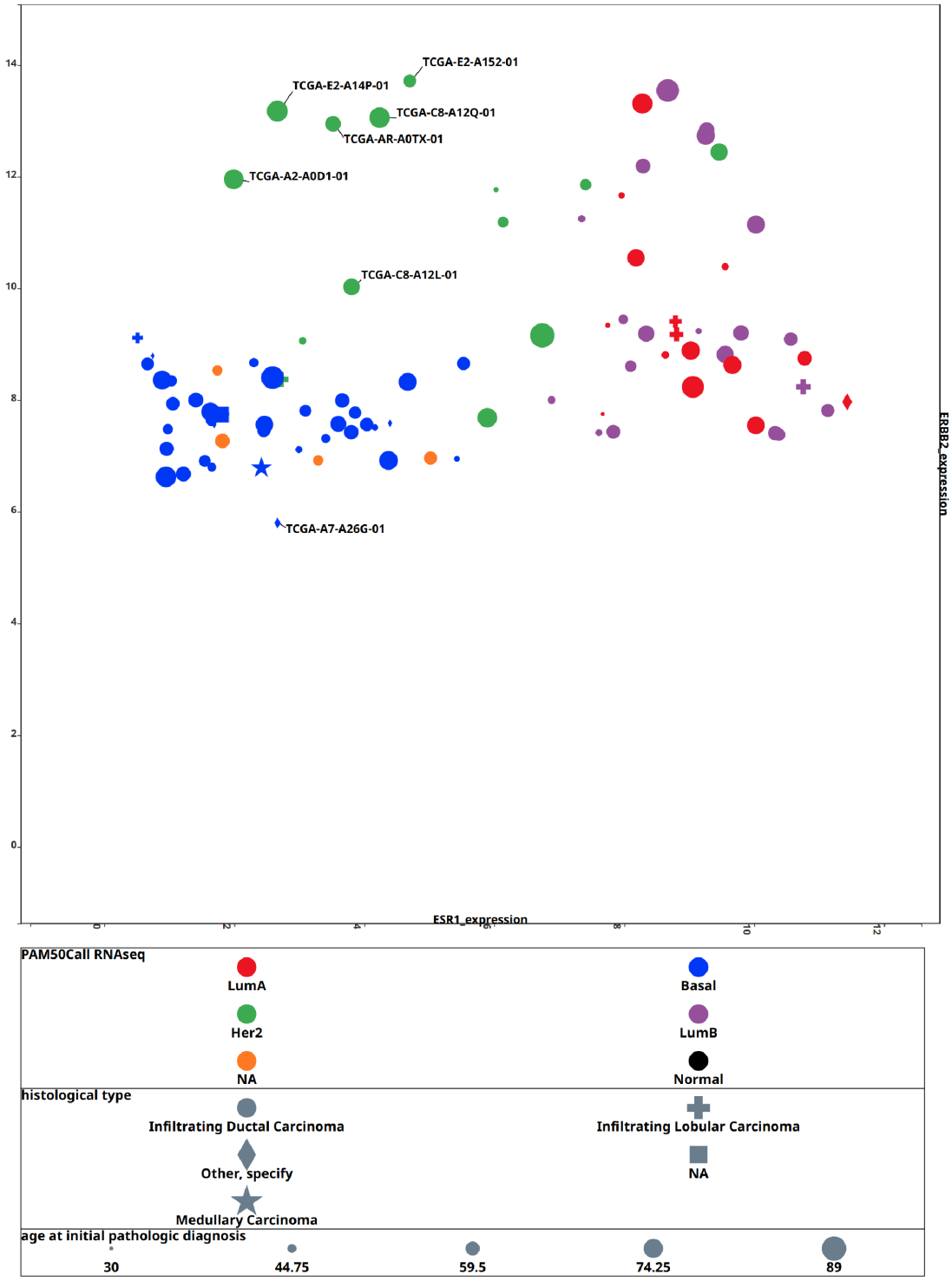

Figure 3.32: *ESR1* vs *ERBB2* gene expression in TCGA-BRCA, with labelling applied to selected donors in the TCGA-BRCA study.

Thanks to the colour encoding of the gene expression-based PAM50 subgrouping and the X-Y representation of *ESR1* and *ERBB2* expressions, users can quickly draw the following conclusions: 1. High *ERBB2* expressing cases that do not express high *ESR1* are rare and ex-

clusively cluster with the Her2 PAM50 subtype (Her2 is an alternative name for the *ERBB2* gene) 2. Low *ERBB2* and low *ESR1* expressing cases exclusively cluster in the Basal and "unknown" PAM50 subtypes 3. *ESR1* expression does not clearly separate the Luminal A and Luminal B subtypes.. While these observations do not constitute novel findings, they serve as a good demonstration of the interpretation of data in a scatter plot in EPISTEME with the support of its legend feature. The following sections will discuss the dimensionality reduction and clustering features building on the scatter plot feature of EPISTEME.

3.3.6 Dimensionality Reduction of High-Dimensional Omics Data in EPISTEME

Dimensionality reduction is a central concept in the analysis of high-dimensional omics datasets. An omics assay such as RNASeq can quantify all genes in a cohort (order of 10000 data points) or a methylation probe array assay can quantify the methylation states of even more (order of 100000 or 1000000 data points). Such high dimensional data poses challenges such as problems in the usage of distance functions, overfitting of classifiers, computational difficulties due to combinatorial explosion of variables. In addition, the human visual perception is limited to three dimensions and is most efficient in two dimension when interpreting data visualizations on a computer screen or a paper medium. Thus, dimensionality reduction is useful also for visualization purposes.

The variance in a dataset is usually captured by the identification of hidden variables that combine information from multiple dimensions (such as multiple genes, probes etc.) and representing them as a single pseudo-dimension. For example, in the usage of the Principle Component Analysis (PCA) technique, these pseudo-dimensions are called principle components. By summarizing the high-dimensional data of omics experiments in low-dimensional spaces, users get benefits both in analysis and visualization, making such techniques a key component of successful large cancer omics projects [218] [353], which serves as an important motivator in making these features available in EPISTEME.

EPISTEME currently offers dimensionality reduction on gene expressions, methylation probes, TAD-based SV recurrence, TAD-based CNV recurrence and user-selected sets of quantitative metadata columns (such as mutational signatures presented in Section 3.3.5.2).

The following sections will describe the different approaches of dimensionality reduction offered in EPISTEME and how they are visualized using the scatter plots previously described. For these discussions, the entire cohort of the TCGA-BRCA project will be used as opposed to the WGS subset. With 1233 donors, of which 1101 correspond to tumours and 132 correspond to normal tissue that was surgically resected from the same patient along with the tumour, this cohort is the largest in the whole TCGA study and has multiple subgroups to study, also according to available classifiers. Thus, the TCGA-BRCA study is an ideal cohort to study different approaches of dimensionality reduction and EPISTEME's features in this field of data analysis.

3.3.6.1 PCA in EPISTEME

EPISTEME currently offers and runs a dense, non-incremental PCA algorithm and automatically assigns the top 10 principle components with respect to explained variance in the cohort

metadata, flexibly runnable for up to a user-determined top-N dimensions to reduce.

Upon the calculation of the PCA, EPISTEME plots a scatter plot, by default showing the first two principal components (Figure 3.33, on the example of gene expressions in the TCGA-BRCA study).

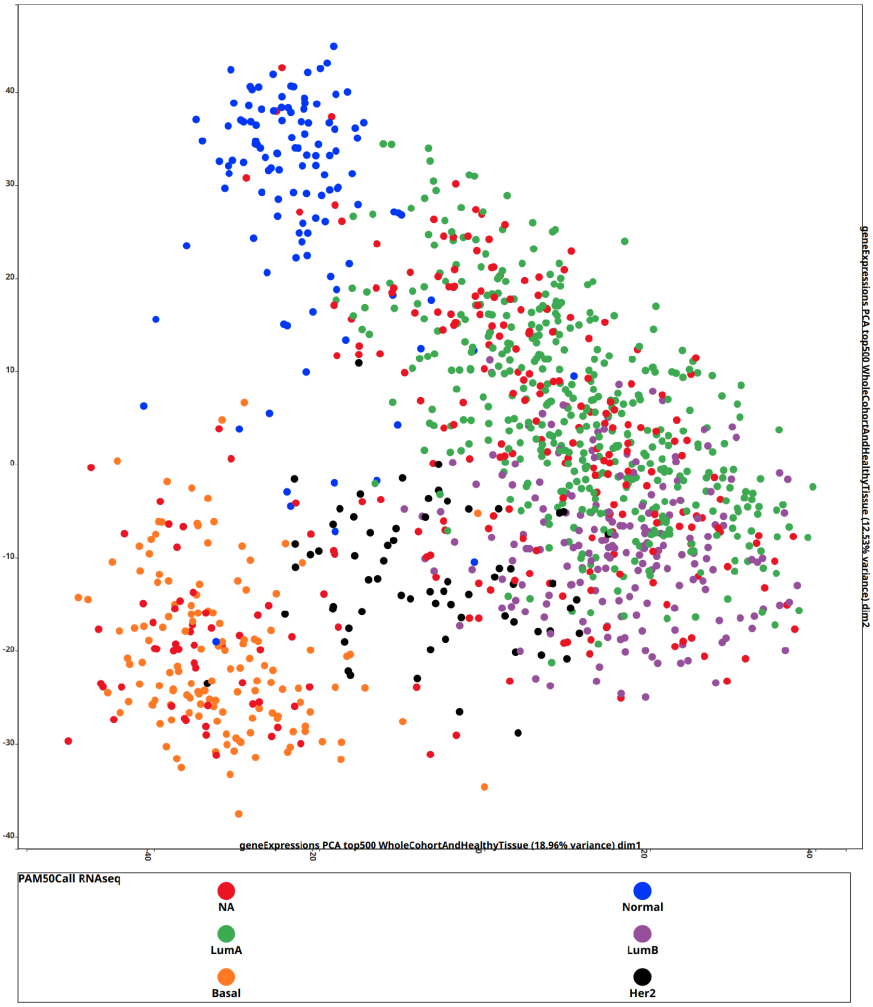

Figure 3.33: PCA of gene expressions with top 500 variance in the TCGA-BRCA study, colour-coded for the PAM50 classification showing the first two principal components

Frequently, users are interested in investigating higher order principal components that explain less of the variance in the data. Thanks to its grid-based multiplotting capabilities, this is possible in EPISTEME (Figure 3.34). On this plot, the combined PCA analysis of both the transcriptome and the methylome is shown, clearly indicating the first PCA component of the transcriptome to successfully separate the Basal and Her2 groups from the Luminal and Normal groups, while the Normal and Luminal groups cannot be separated by PCA analysis. The PCA analysis of the methylome yields poorer results with the Basal and Normal groups clustering away from the Luminal and Her2 groups, but not showing any separation.

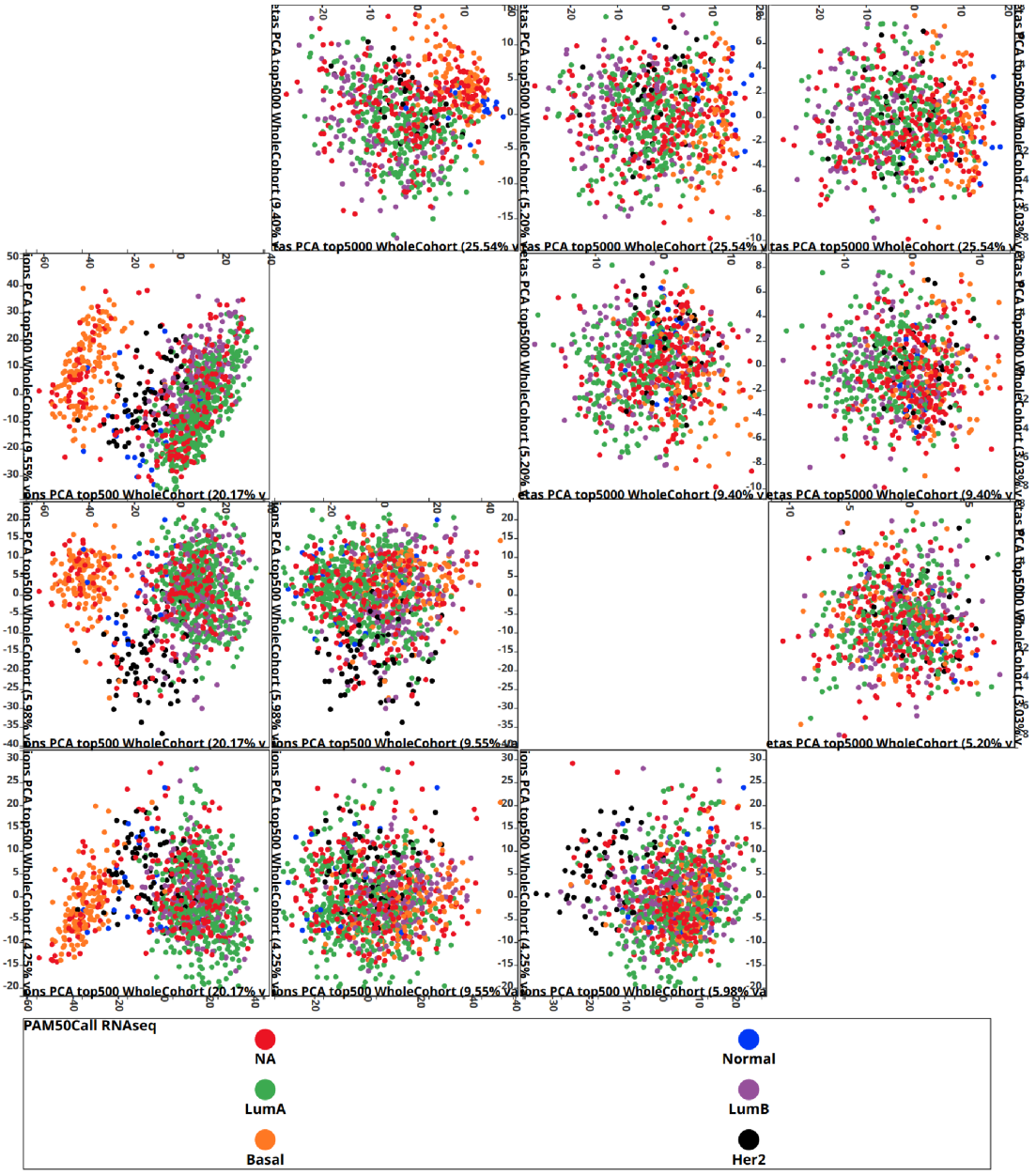

Figure 3.34: PCA matrices of gene expressions and methylome betas in the TCGA-BRCA study, colour-coded for the PAM50 classification. Subplots above the main diagonal are methylome PCAs (top 5000 most variable probes), whereas the subplots below are transcriptome PCAs (top 500 most variable genes). Subplot (i, j) corresponds to the scatter plot of the i$^{\text{th}}$ vs j$^{\text{th}}$ PCA for the analysed phenotype.

It is by now established that the PCA is not an ideal dimensionality reduction technique in terms of visualizing data [523], and it was largely supplanted by more modern and non-linear dimensionality techniques such as tSNE and UMAP, which will be discussed in the following sections.

3.3.6.2 *tSNE in EPISTEME*

t-Distributed Stochastic Neighbor Embedding (tSNE) [325] is a non-linear dimensionality reduction technique which has successfully been applied in the dimensionality reduction and subsequent classification of large cancer omics datasets [218], [353]. tSNE is a highly customizable algorithm, but its parameter selection requires great care because it is sensitive to changes in perplexity and exaggeration strategies [524].

EPISTEME implements flexible controls (with default parameters in parentheses) for top-N (in terms of variance) items to analyse (500), early exaggeration (12), iterations for early exaggeration (250), total iterations (1000), late exaggeration (1.5), iterations for late exaggeration (100), heavy-tailed kernel coefficient alpha (1) [525], the optional step of applying a preliminary PCA on the dataset with up to 50 principal components (not applied by default), and finally and perhaps most importantly perplexity (5). These settings are modifiable in an intuitive user interface (Figure 3.35). Following parameter selection, tSNE will rapidly be calculated on the client-side with no load on the EPISTEME server, and the results will be output as a scatter plot.

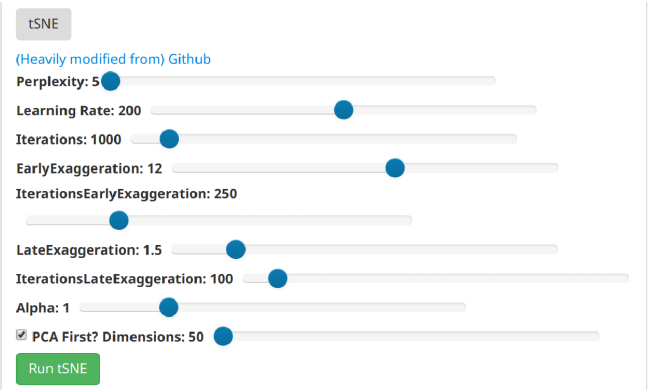

Figure 3.35: tSNE settings in EPISTEME

The application of tSNE with appropriate parameters leads to an excellent separation of the expected gene expression subtypes of breast cancer (Figure 3.36), except for the separation of the Luminal A and Luminal B groups, where Luminal B cases mostly co-cluster, but do not separate from the rest of the Luminal group.

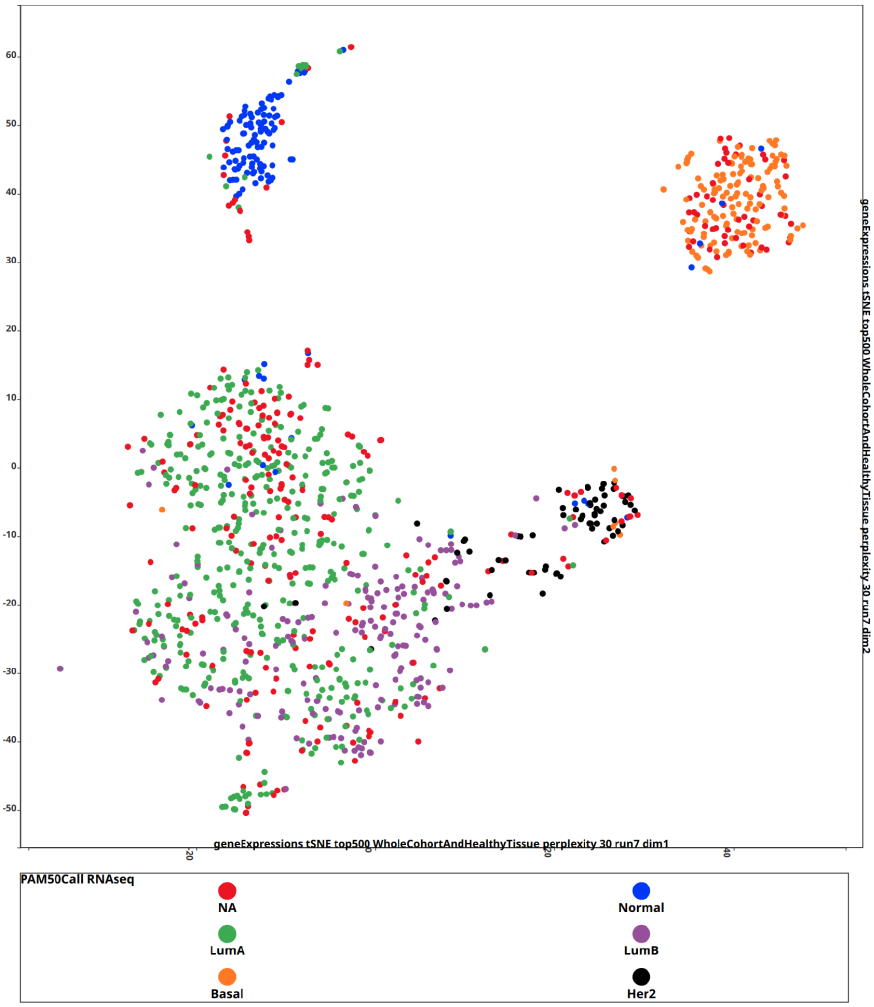

Figure 3.36: tSNE on the top 500 most variable gene expressions from the TCGA-BRCA study with a preliminary PCA reducing the dimensionality to 50, followed by an application of tSNE with perplexity 30 and late exaggeration 1.2

This well-performing set of parameters was chosen by a tSNE parameter sweep, taking advantage of the multiplot grid offered in EPISTEME. The procedure was repeated both with a preliminary PCA (Figure 3.37) and without (data not shown due to high similarity with a slight superiority observed in favour of a pre-PCA). Finally, the parameters perplexity 30 and late exaggeration 1.2 were selected.

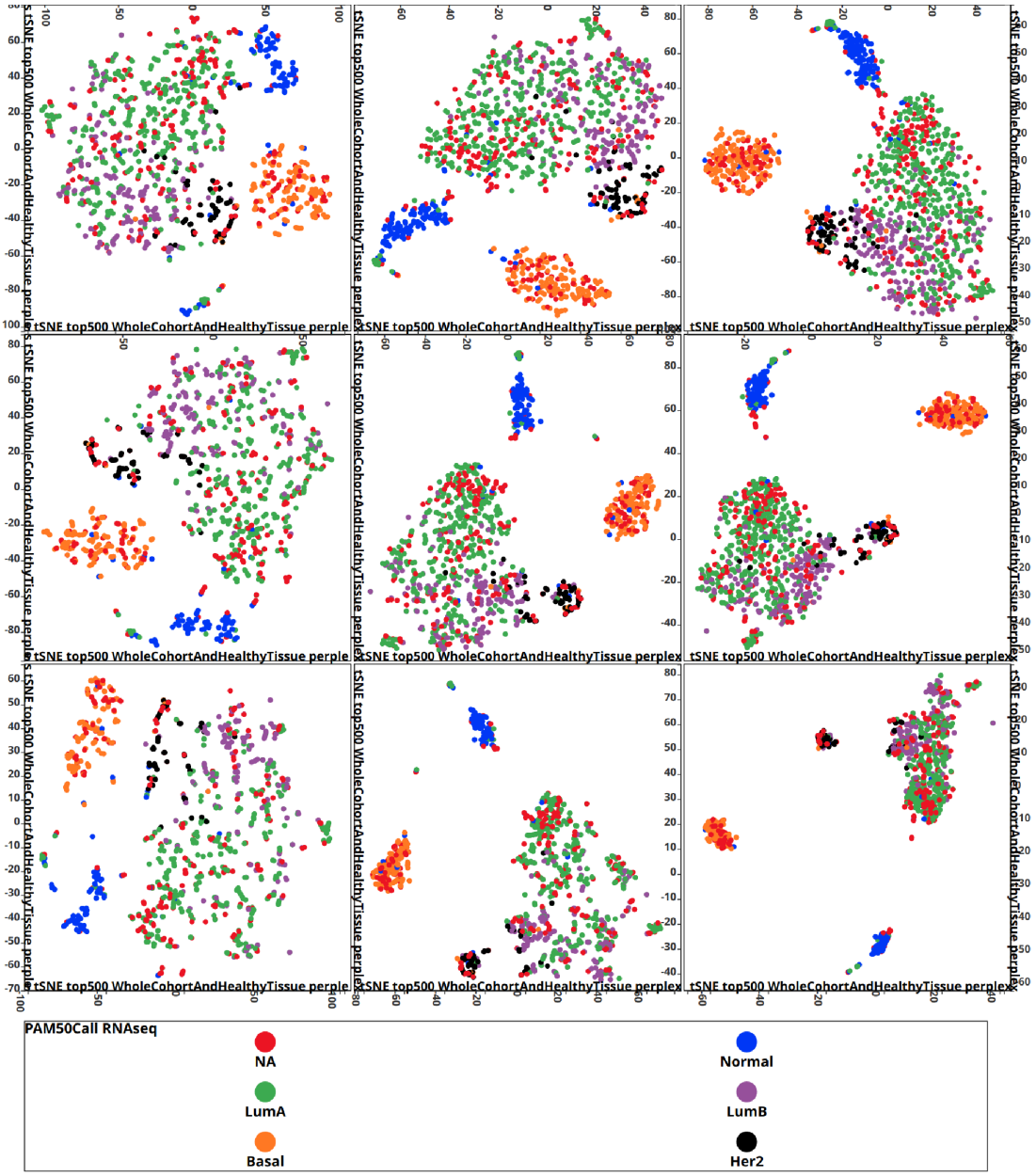

Figure 3.37: tSNE parameter sweep on perplexity and late exaggeration on the top 500 most variable gene expressions from the TCGA-BRCA study with a preliminary PCA reducing the dimensionality to 50. The perplexity parameter was swept across the X-axis starting from the left with the steps 5, 15, 30 and the late exaggeration parameter was swept across the Y-axis starting from the top with the steps 1, 1.2, 1.5.

In particular, late exaggeration seems have the strongest effect despite the small steps used in its application. While applications with much larger (and often sparse) datasets in the order of 10^4 dimensions such as scRNA sequencing are amenable to higher perplexities and higher late exaggeration values, applications on bulk tumour samples with sample sizes in the order of 10^3 dimensions with dense data seem to require more modest values. Notably, a combination of low perplexities, and even moderately high late exaggeration (5-vs-1.2, 5-vs-1.5, 15-vs-

1.5) leads to an over-fragmentation of the Luminal cluster. While it can be argued that the 15-vs-1.5 configuration has the desirable consequence of separating the LumB cases from the LumA cases, this separation is also flawed due to the emergence of smaller clusters, hinting at overfitting.

tSNE can similarly be applied to methylome beta values (obtained by a methylation array or WGBS) or other data matrices. The application on the methylome will be presented in Section 3.3.6.4 in comparison with the UMAP algorithm.

3.3.6.3 UMAP in EPISTEME

Uniform Manifold Approximation and Projection (UMAP) is a recently introduced algorithm for dimensionality reduction [526], which has desirable properties such as quick runtimes, preservation of global structure and embeddings usable for clustering purposes. A recent study claimed superiority to tSNE in the field of single-cell RNA sequencing data analysis [527] most notably in the preservation of global structure and higher reproducibility. It is therefore of great interest in the cancer omics community. Driven by this interest, it was a great motivation to implement a user-runnable UMAP algorithm in EPISTEME. Due to the lack of available JavaScript libraries for this task, the UMAP module of EPISTEME was written from scratch in JavaScript based on its Python implementation.

EPISTEME offers a highly customizable set of parameters akin to its tSNE features for running UMAP with defaults as (Figure 3.38). Notably, internal testing showed setting the parameter set-op-mix-ratio to 0.5 rather than the UMAP default 1 to have desirable effects in terms of equally weighing global and local structure, leading to tighter and well-separated clusters.

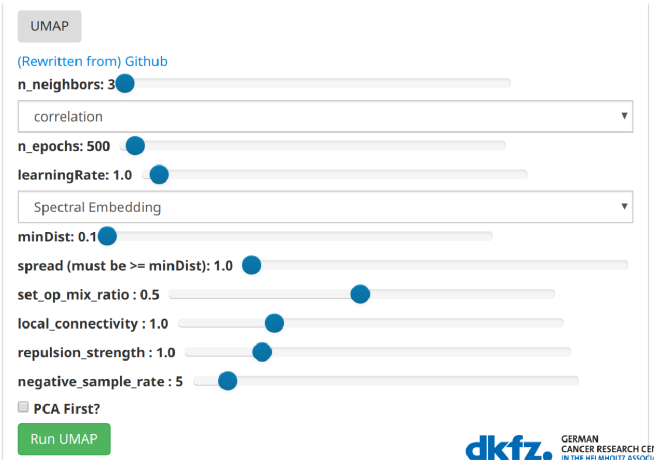

Figure 3.38: UMAP settings in EPISTEME

Similar to the application of tSNE, UMAP with appropriate parameters leads to an excellent separation of the expected gene expression subtypes of breast cancer (Figure 3.39), except for the separation of the Luminal A and Luminal B groups, where Luminal B cases mostly co-cluster, but do not separate from the rest of the Luminal group. The Her2 group seems to be ”emerging” from the LumB-rich subset of the Luminal group, whereas the normal cases seem

to be closer a LumA-rich region of the Luminal cluster.

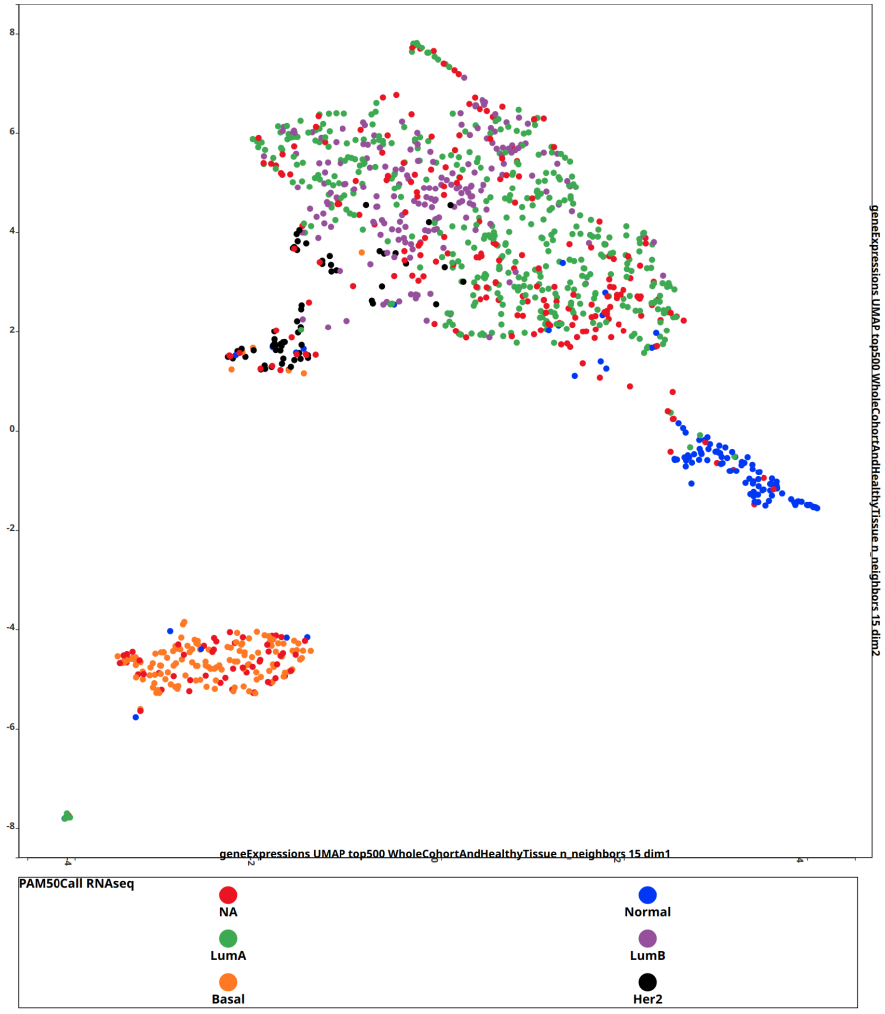

Figure 3.39: UMAP on the top 500 most variable gene expressions from the TCGA-BRCA study, with Euclidean Distance, nNeighbors 15 and minDist 0.05

Again, this well-performing set of parameters was chosen by a UMAP parameter sweep, taking advantage of the multiplot grid offered in EPISTEME. The procedure was repeated both with the distance measure Euclidean Distance (Figure 3.40) and Correlation Distance (data not shown due to very high similarity). Finally, the parameters nNeighbors 15 and minDist 0.05 were selected.

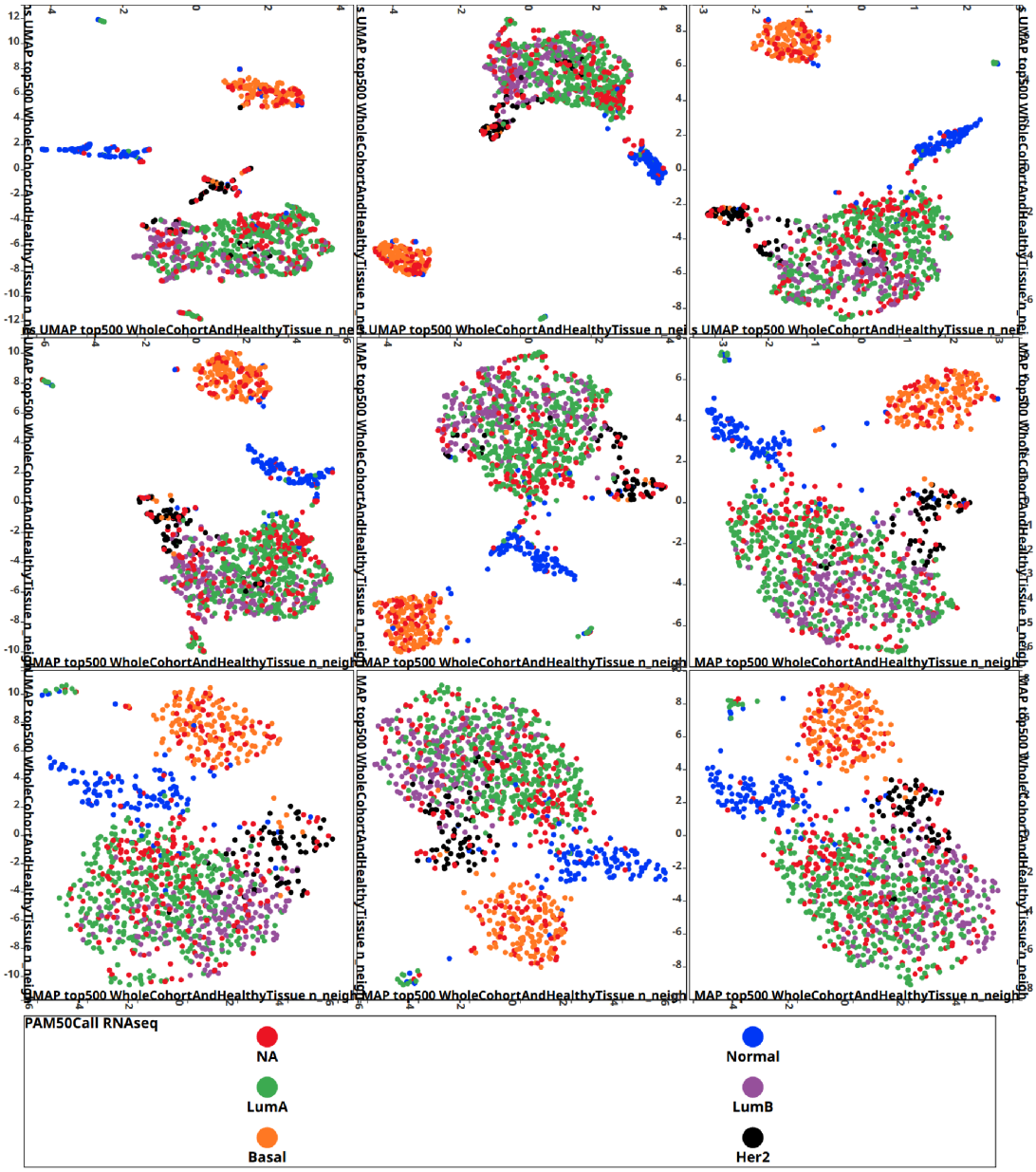

Figure 3.40: UMAP parameter sweep on nNeighbors and minDist on the top 500 most variable gene expressions from the TCGA-BRCA study. The nNeighbors parameter was swept across the X-axis starting from the left with the steps 5, 15, 30 and the minDist parameter was swept across the Y-axis starting from the top with the steps 0.05, 0.25, 0.75.

The minDist parameter seems to have the strongest effect on the global behaviour of the embeddings, with smaller values leading to significantly tighter clusters. As opposed to the strong effects of perplexity in the tSNE algorithm, the nNeighbors parameter doesn't seem to lead to drastic changes. This can be considered a desirable attribute because it leads to less assumptions on the expected cluster sizes in the application of the algorithm.

UMAP can also similarly be applied to methylome beta values or other data matrices. The application on the methylome will be presented in the following Section 3.3.6.4 in comparison

with the tSNE algorithm.

3.3.6.4 Comparison of Dimensionality Reduction Techniques using a Multi-plot in EPISTEME

Following testing and optimization of parameters for transcriptome analysis by tSNE and UMAP, it is of interest to compare these two algorithms including their applications to methylome data. In a 2-by-2 multiplot grid, the following configurations are compared using the optimal parameters obtained from transcriptome analysis:

i) Upper-Left: tSNE-transcriptome (top 500 genes, perplexity 30 and late exaggeration 1.2),

ii) Upper-Right: tSNE-methylome (top 5000 probes, perplexity 30 and late exaggeration 1.2),

iii) Lower-Left: UMAP-transcriptome (top 500 genes, nNeighbors 15 and minDist 0.05),

iv) Lower-Right: UMAP-methylome (top 5000 probes, nNeighbors 15 and minDist 0.05),

.

Figure 3.41 summarizes the results from this comparison. The general structures are remarkably similar between the two dimensionality reduction approaches, whereas the transcriptome and methylome analyses yield slightly different results. The Basal-like and Normal-like groups show excellent separation from the Luminal group in all configurations. The Luminal B group's subtle co-clustering in transcriptome analysis is not maintained in methylome analysis, suggesting a shared cell of origin with slightly different transcriptomic programmes between the Luminal A and Luminal B groups. Without the colour-coding guiding this process, clearly Luminal A and B groups would not emerge as distinct subtypes from any of the chosen dimensionality techniques.

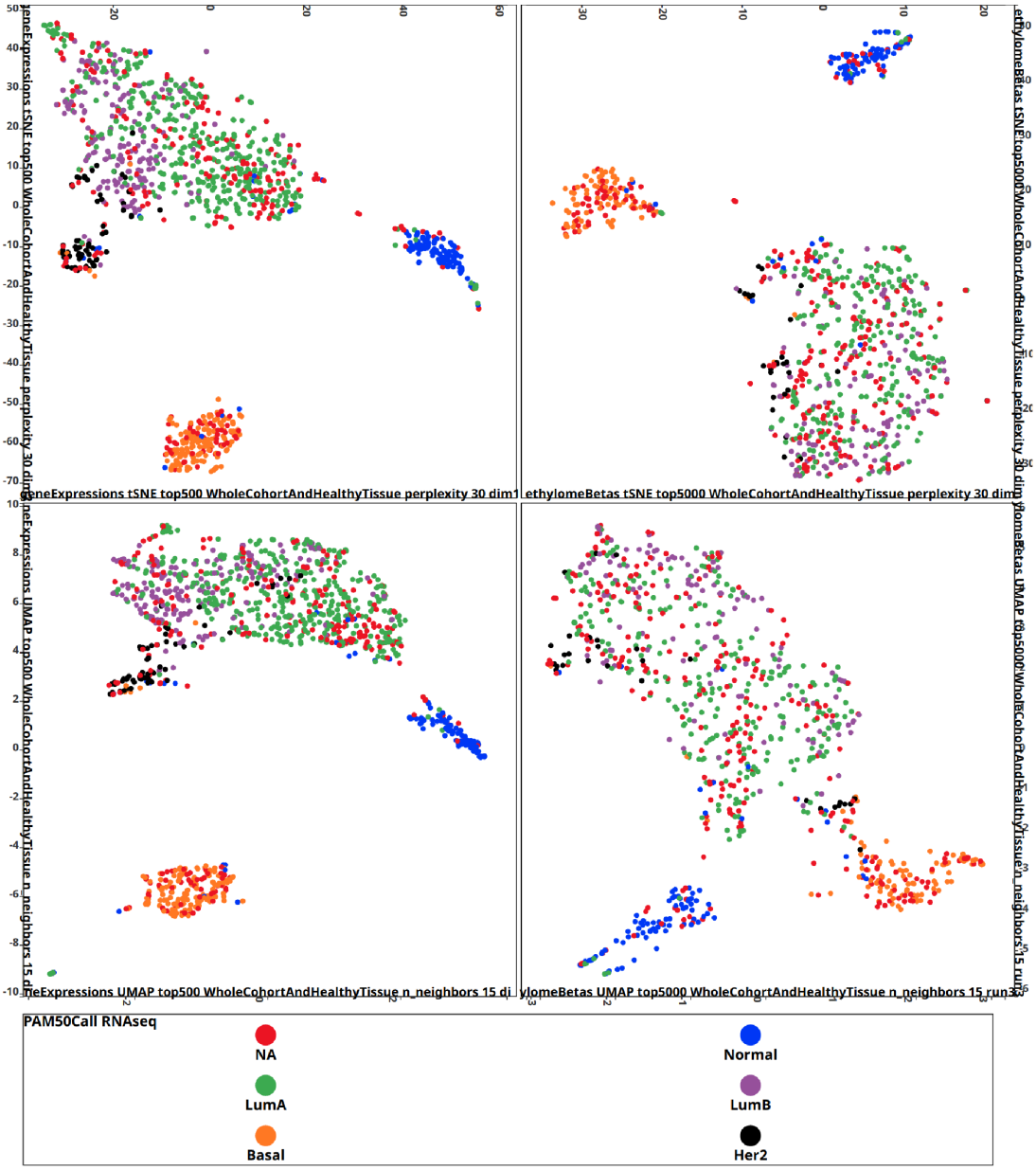

Figure 3.41: tSNE and UMAP dimensionality reduction techniques applied to the transcriptome (top 500 most variable genes) and methylome (top 5000 most variable probes) datasets from the TCGA-BRCA study with the previously chosen best parameters. Upper-left: tSNE on the transcriptome. Upper-right: tSNE on the methylome. Lower-left: UMAP on the transcriptome. Lower-right: UMAP on the methylome.

Interestingly, the Her2 group seems to get broken into two clusters in methylome analysis, confirmed by both dimensionality reduction approaches, suggesting two different possible cells of origin yielding a similar and distinct transcriptomic profile due to the *ERBB2* gene amplification. This is best observed in Figure 3.42 where the well-separated Her2-rich cluster in the Upper-left tSNE-transcriptome subplot was user-selected, prompting a parallel selection of the same cases across all 4 subplots. This feature can be used for any type of subplots be it

scatter plots, stacked bar charts or "Box-KDE-Jitter Plots" where the behaviour of a single case or a group of cases can be tracked. This facilitates applications such as the test for stability of clustering approaches, similar to the results presented here.

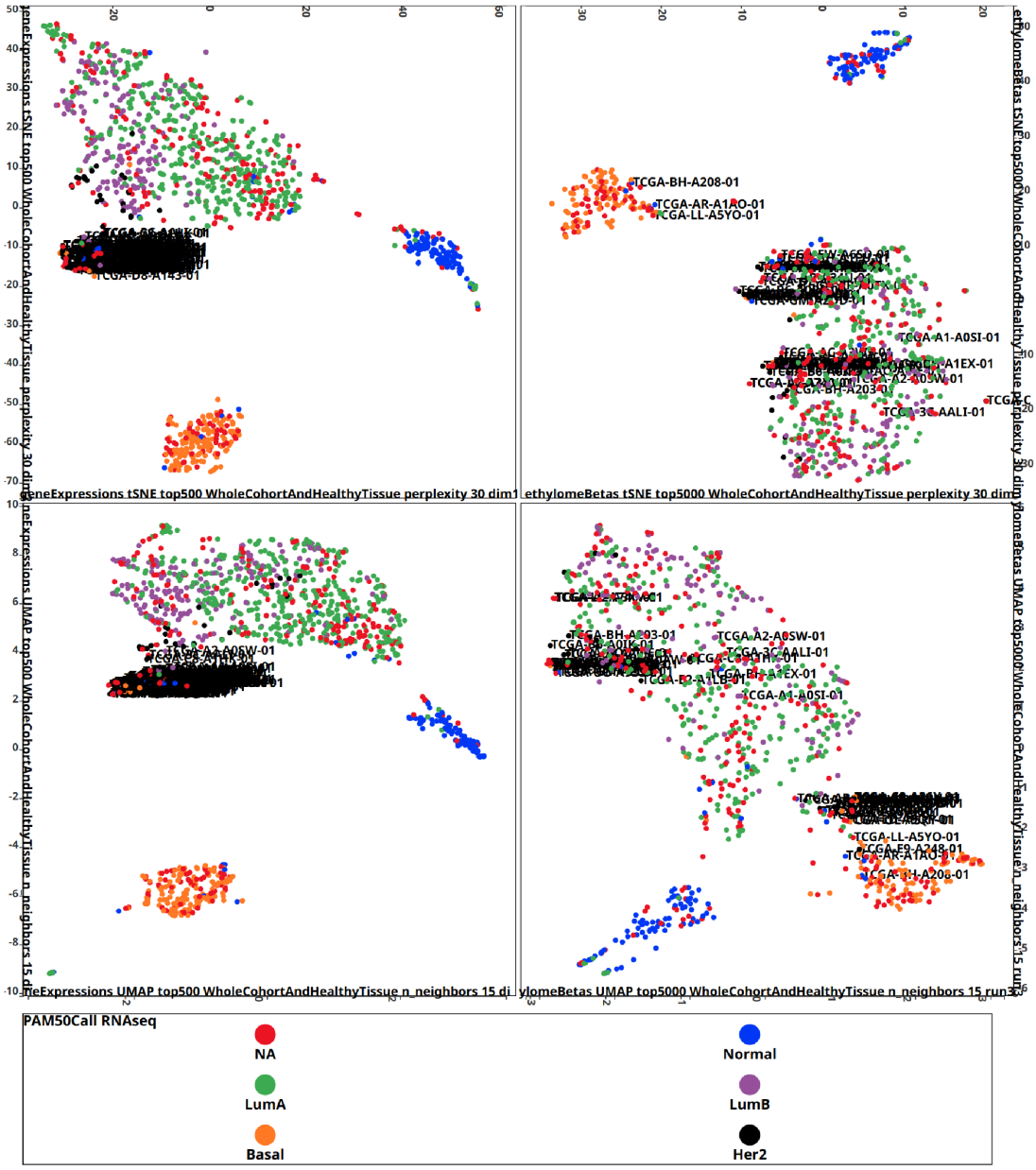

Figure 3.42: tSNE and UMAP dimensionality reduction techniques applied to the transcriptome and methylome datasets from the TCGA-BRCA study, with the the user selection of the well-separated Her2-rich cluster in the Upper-left tSNE-transcriptome subplot.

In summary, EPISTEME offers an extensive and highly customizable and interactive set of features to apply dimensionality reduction on complex high dimensional datasets and make them accessible to users without programming knowledge. The dimensionality reduction module of EPISTEME along with its grid-based plots, facilitates parameter optimization and guides

investigations of the existence of cell of origin or transcriptomic regulation based subtypes in a cohort.

3.3.7 Clustering in EPISTEME

Clustering is a central task in data analysis, giving users the power to detect patterns in unlabelled data in an unsupervised manner. Clustering algorithms are based on different cluster models in terms of criteria that determine co-clustering such as distance connectivity, local density, centroids of a group of data points etc. EPISTEME implements clustering algorithms from three strategies:

1. Connectivity-based, hierarchical clustering: agglomerative hierarchical clustering is implicitly used in ordering rows and columns in the heatmap feature of EPISTEME (in prototpye stage) and will not be discussed here
2. Centroid-based clustering: the K-means and K-medoids algorithms
3. Density-based clustering: the fuzzy DBSCAN and OPTICS algorithms

While the K-means and K-medoids algorithms are straightforward in terms of usage, apart from the choice of an appropriate distance function and a K parameter indicating the expected number of centroids, the fuzzy DBSCAN and OPTICS algorithms offer more parameters (Figure 3.43).

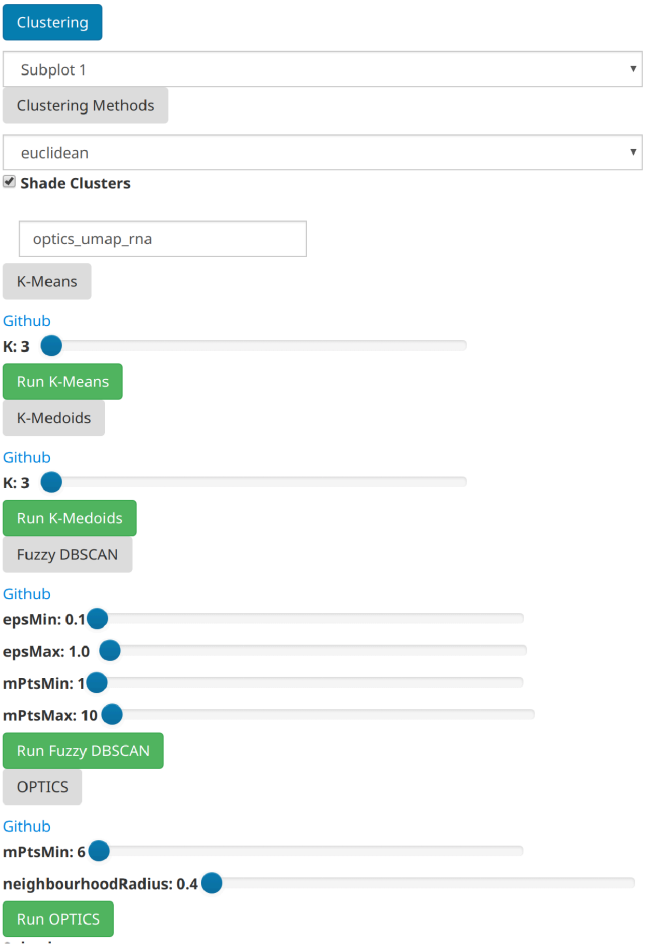

Figure 3.43: Clustering settings offered in EPISTEME

Figure 3.44 summarizes the findings from the K-Means, K-Medoids and OPTICS clustering algorithms on the TCGA-BRCA transcriptome data following UMAP application with previously optimized parameters. The UMAP algorithm was chosen for its preservation of global structure and distances. For the centroid-based algorithms, 5 turned out to be an appropriate number of clusters. Interestingly, both centroid-based approaches successfully recaptured LuminalB-rich clusters that are well separated from a clusters that consist almost purely of LuminalA cases. While the K-medoids algorithm was more successful at separating the Her2 cluster from the LuminalB-rich cluster, the bridge between the Her2 cluster and the Luminal cluster was challenging for all algorithms.

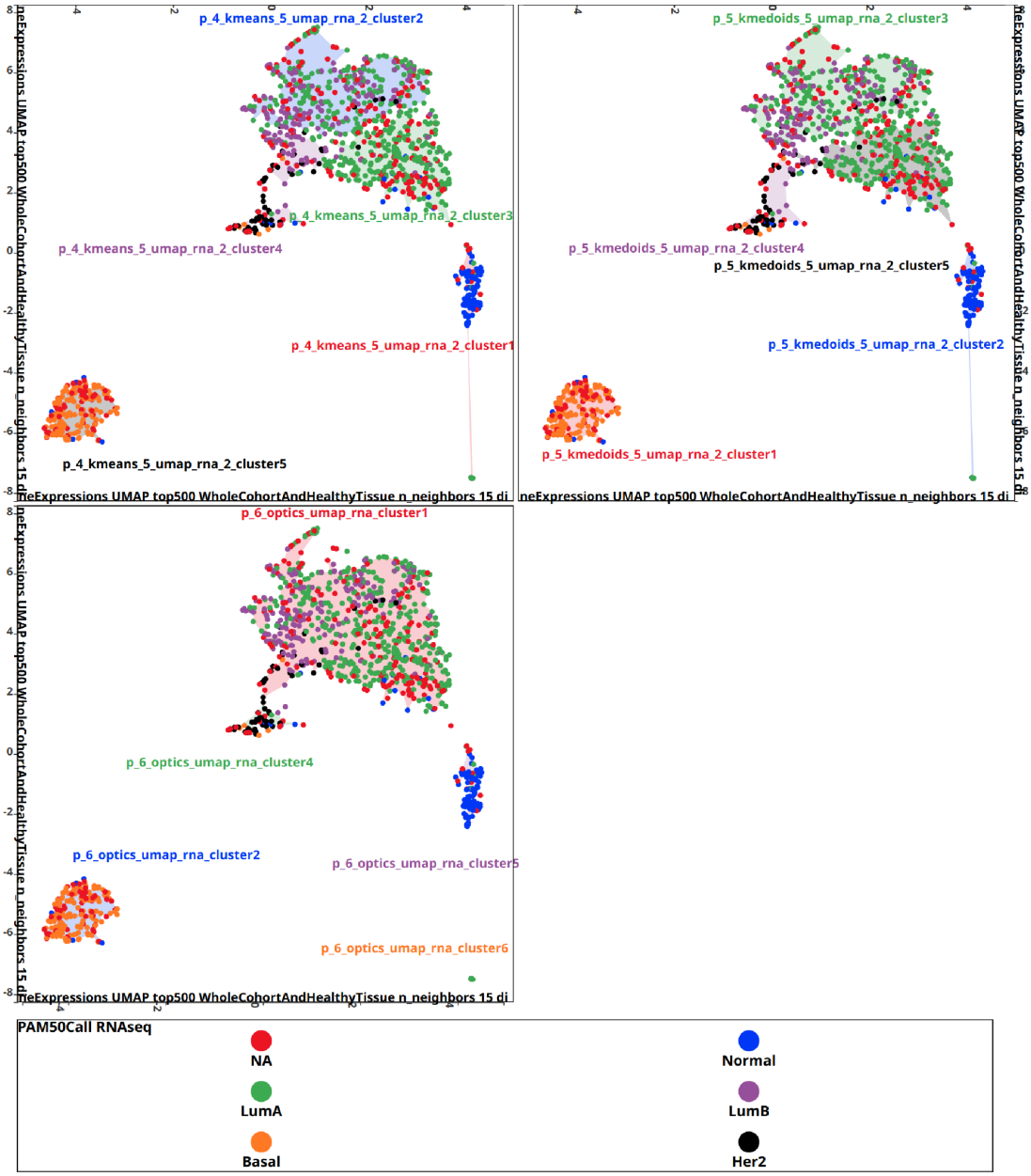

Figure 3.44: Comparison of the K-Means, K-Medoids and OPTICS clustering algorithms on the TCGA-BRCA transcriptome data following UMAP application with optimized parameters. Upper-left: K-means,k=5. Upper-right: K-medoids,k=5. Lower-left: OPTICS.

The DBSCAN algorithm [528], due to its requirement of uniform densities across the clusters, did not perform well in the clustering of UMAP-reduced gene expression data from the TCGA-BRCA study, and has therefore been omitted from this figure.

Overall, EPISTEME offers a selection of clustering algorithms with user control over distance measures and other parameters. They can be applied to any kind of scatter plot and are particularly useful when applied to UMAP outputs in order to detect subgroups in an unsupervised manner in dimensionality-reduced data.

3.3.8 Subcohort Selections in EPISTEME

In any data analysis setting, discovery and analysis of unknown and known subgroups of the data is a crucial concept. This is also true for cohort analysis in omics data analysis projects, where identification and designation of subcohorts opens up avenues for advanced comparative data analysis approaches and leads to biological insights. EPISTEME offers powerful features for designation of subcohorts from a number of its standard visualizations.

Subcohort selections are offered both for categorical and quantitative data and are tightly integrated to previously introduced data visualization modules such as Circos plots and flexible 2D plots. The following comprehensively describe the subcohort designation features of EPISTEME, in the order of previous introduction of the aforementioned data visualization features.

3.3.8.1 Subcohort Selection from Variant Recurrence Selections in Circos Plots

Any recurrence item in a Circos plot discussed in Sections 3.3.1.2, 3.3.1.3, 3.3.1.4 effectively shows a selection of cases that fulfil a condition while excluding others. This information can be used to define subcohorts (Figure 3.45), where the gene mutation recurrence layer (Figure 3.4, outermost arc) is used to extract the *TP53* mutant cases in this cohort. EPISTEME thus gives its users the ability to rapidly isolate the cases that contribute to any type of genomic variant recurrence.

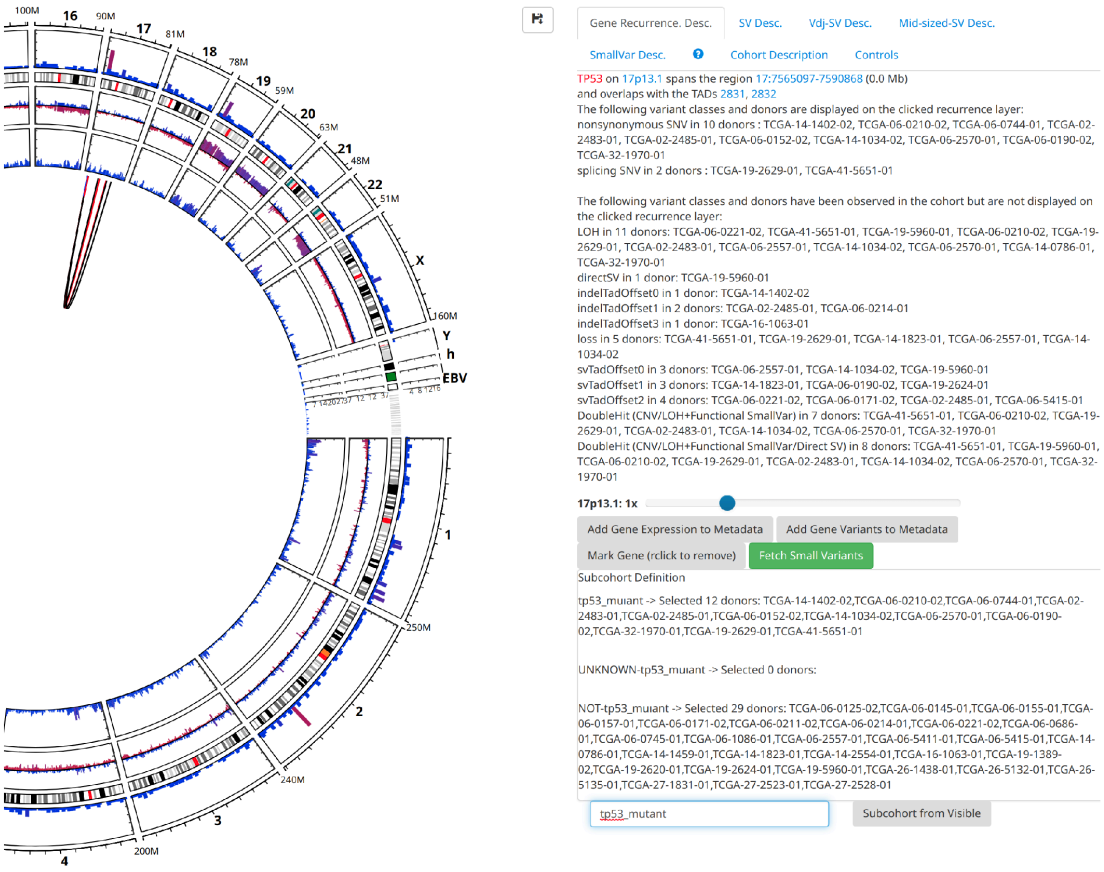

Figure 3.45: Subcohort selection from a recurrence item in the Circos plot feature of EPISTEME: selecting for *TP53* functional mutant cases in the TCGA-GBM (WGS) study.

EPISTEME parses the selected cases, and defines a positive selection, namely the *TP53* mutant cases (12), unknown cases (0, as all cases have mutational data) and negative cases (29), assigning them to subcohorts named by the user (tp53-mutant) with appropriate prefixes. This automatic definition of selections as well as their complements in a manner that takes data availability into consideration is a key feature facilitating the subcohort operations that are described in Section 3.3.9.

3.3.8.2 Subcohort Selection from Quantitative Metadata Variables

Quantitative metadata variables defined in Section 3.3.4 can be used for subcohort definitions with dedicated tools in EPISTEME. For quantitative variables, this is accomplished by a thresholding tool where the quantitative variable of interest is displayed in a helper plot with all distinct values encountered in the cohort. This helper plot allows rational selection of cutoffs with relations less-than, less-than-or-equal-to, equal-to, between, greater-than and greater-than-or-equal-to.

Figures 3.46 and 3.47 showcase this feature both for thresholding based on a low-cutoff and high-cutoff, on the genes *EGFR* and *CDKN2A* respectively. EPISTEME automatically assigns the cases with unavailable RNA-Seq data into UNKNOWN categories to facilitate accurate

processing of subcohort-based analyses.

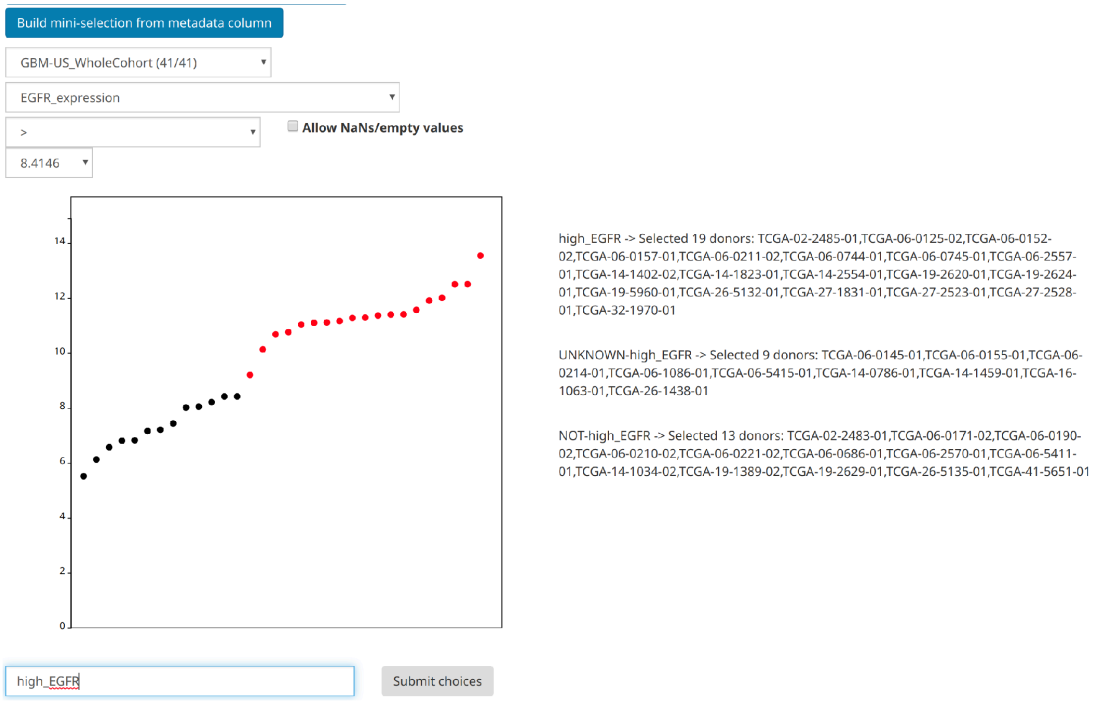

Figure 3.46: Subcohort selection by thresholding a quantitative variable in EPISTEME: selecting for high *EGFR* expression in the TCGA-GBM (WGS) study.

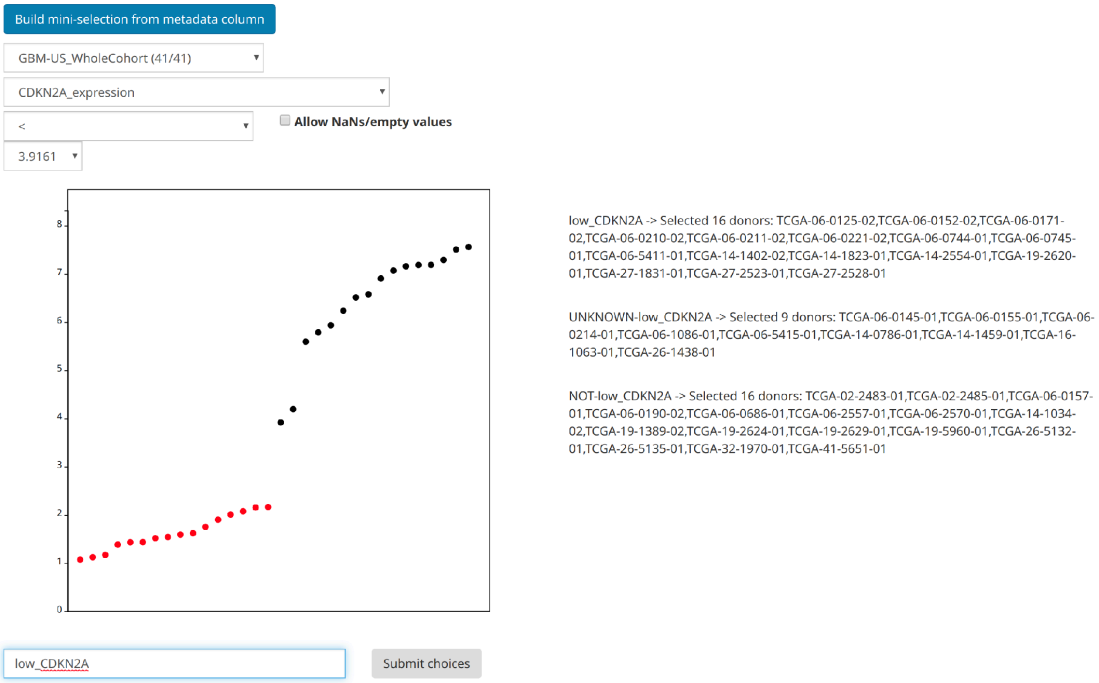

Figure 3.47: Subcohort selection by thresholding a quantitative variable in EPISTEME: selecting for low *CDKN2A* expression in the TCGA-GBM (WGS) study

With this feature, EPISTEME's users can flexibly designate subcohorts on any quantitative variable such as age, gene expression and patient survival among others.

3.3.8.3 Subcohort Selection from Categorical Metadata Variables

Categorical metadata variables are best represented as checkboxes for subcohort filtering purposes. Figure 3.48 shows a checkbox-based subcohort designation setup based on the example of *TP53* mutations in the TCGA-GBM (WGS) cohort. All encountered values for the categorical variable are summarized, with the individual checked items combined in an "OR" relationship.

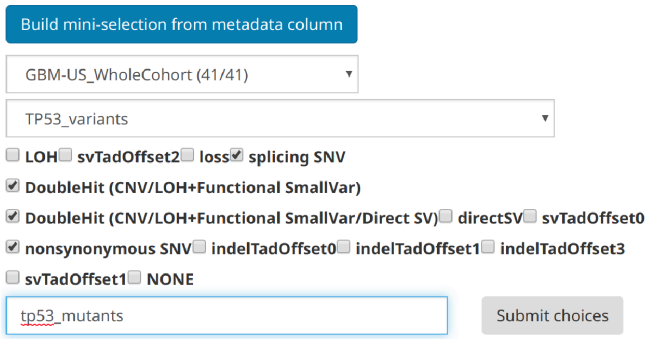

Figure 3.48: Subcohort selection by checkboxes on a categorical variable in EPISTEME: selecting for functional *TP53* mutations in the TCGA-GBM (WGS) study

This feature also finds a good use for individually selecting cases using the standard "Donor" column with a flexibility beyond data-driven criteria (Figure 3.49).

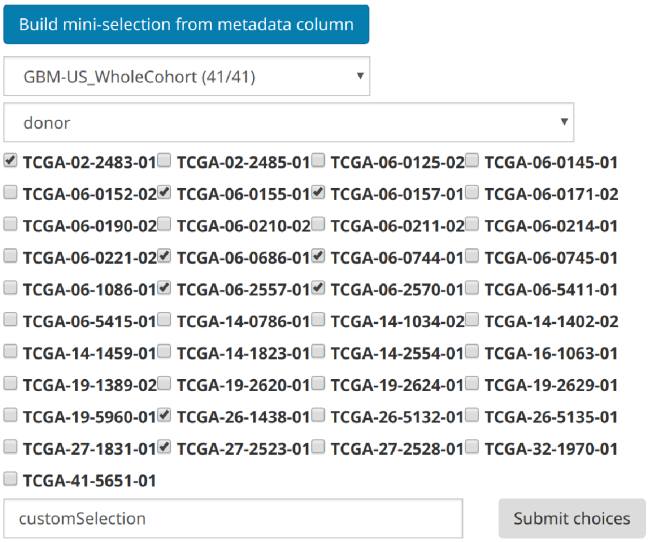

Figure 3.49: Subcohort selection by checkboxes on a categorical variable in EPISTEME: selecting individual donors in a custom selection in the TCGA-GBM (WGS) study

With this feature, EPISTEME's users can flexibly designate subcohorts on any categorical variable such as patient gender, gene mutation status and disease histological type among others.

3.3.8.4 Subcohort Selection from Flexible 2D Plots

The donor labelling feature of scatter plots in EPISTEME presented in Section 3.3.5.3 lends itself very well to subcohort designations. Figure 3.50 demonstrates this procedure on the example of tSNE plots described in Section 3.3.6.2. In this analysis, the normal tissue samples have been excluded, leading to only bona fide tumour specimens being considered, and assigned to mostly Basal, Her2, Luminal-A and Luminal-B PAM50 subtypes with a very small number of cases in the Normal PAM50 subtype (likely to be specimens with low tumour cell content). Manual selection of the Basal-rich cluster yields 187 positive, 904 negative and 142 unknown cases, where the unknown cases both correspond to the excluded normal tissue specimens and cases with no available RNA-Seq data.

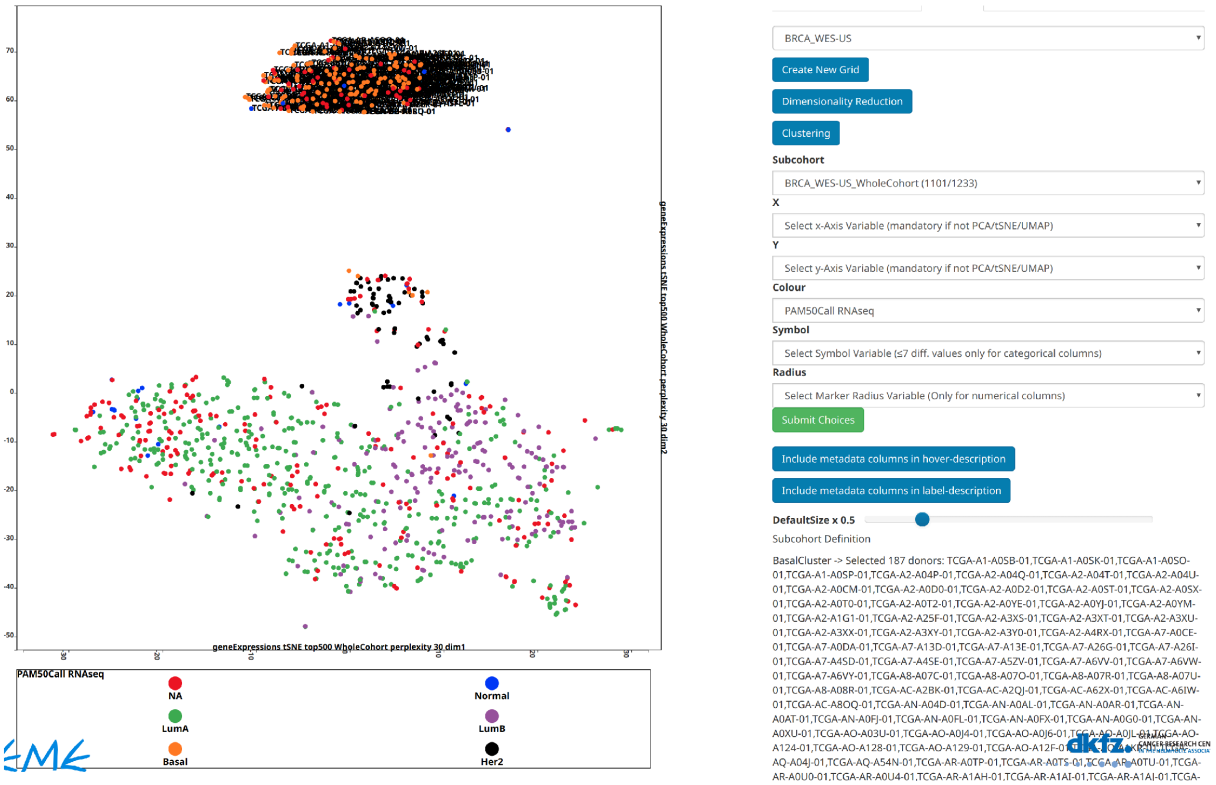

Figure 3.50: Subcohort selection from flexible 2D plots in EPISTEME: selecting the Basal-rich cluster based on PAM50 classification from a RNA tSNE plot of the TCGA-BRCA study, excluding normal tissues

The clustering techniques described in Section 3.3.7 automatically generate subcohort definitions from the calculated clusters. For instance, all shaded clusters depicted in Figure 3.44 have their corresponding subcohorts generated following the clustering calculation, both for the positive, unknown and negative prefixes.

In practice, any plot type in the Flexible 2D plot module of EPISTEME can be used for subcohort designations. This includes bar charts where the capping symbols are used for donor labelling as well as subcohort definitions (Figure 3.31).

With this feature, EPISTEME's users can flexibly designate subcohorts from any scatter plot, kde-box-jitter plot, stacked bar chart, including multiplots. Users can therefore manually select outlier cases with respect to any user-defined visualization criteria and explore the causes and consequences of the outlier characteristics in downstream analysis features.

3.3.8.5 Definition of Complex Subcohort Selections Using Logical Expressions on Simple Subcohort Selections

The hitherto described subcohort definition features share one common, central property: they describe one condition and one condition only. While the specific implementation of subcohort selections differ for each underlying data type (thresholding, checkboxes, manual selections from plot data), there is always one specific selection criterion. EPISTEME's user interface

calls these selections mini-selections and offers a feature to build complex selections using these mini-selections using flexible, nested combinations of the fundamental logical expressions AND, OR and NOT.

EPISTEME offers a flexible tool to build complex logical expressions from individual subcohort selections with an intuitive user interface. Multiple nesting levels are possible. Each mini-selection is defined along with "UNKNOWN" cases with missing information such as missing gene expressions. Any mini-selection used in this complex selection building tool contributes their corresponding "UNKNOWN" cases as also "UNKNOWN" to the resulting complex selection (Figure 3.51).

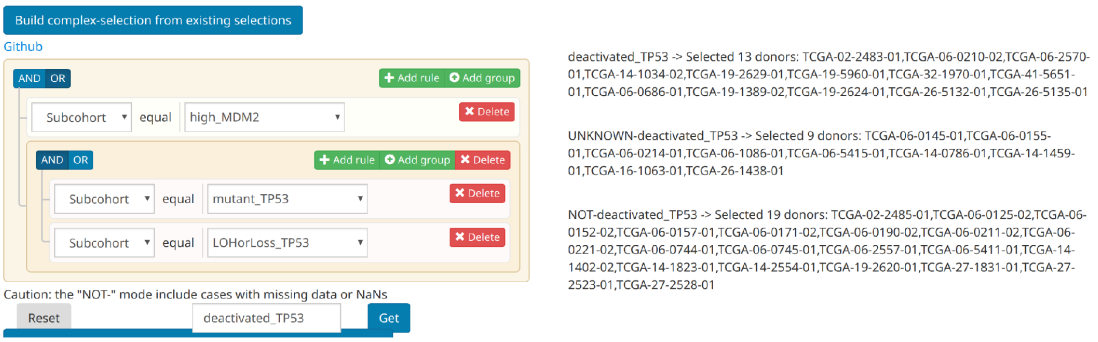

Figure 3.51: Definition of p53 deactivation as a logical expression combining statuses of *MDM2* expression, *TP53* mutation and *TP53* copy number and heterozygosity in the TCGA-GBM (WGS) study: p53 deactivation is thus defined as either a double-hit on the *TP53* gene (AND expression) or an amplification of the *MDM2* gene, which degrades p53 when expressed in high levels [529]. The resulting selection correctly clasifies 9 cases as unclassifiable.

3.3.8.6 Transformation of Multiple Subcohort Selections to Categorical Variables

Each subcohort selection is, by definition, a categorical variable of boolean type (true/false). Extending this concept, one can combine multiple subcohort selections into a single categorical variable. N selections with 2 possible values (true/false) each, yield 2^N possible combinations for each case, not considering combinations that do not occur within the given dataset. If the search space is narrowed by mutual exclusivities, the number of possible combinations can drop dramatically. Figure 3.52 shows the combination of the possible states of *EGFR* and *CDKN2A* expression in the TCGA-GBM study, where high-low-unknown states of both genes are merged into a categorical variable. Upon displaying the gene expressions, the unknown cases are excluded due to lack of data, while the mutually exclusive positive-negative states do not lead to valid combinations and do not yield any results. This effectively reduces the combinations from $2^6 = 64$ to $2^2 = 4$. The results suggest the lack of mutual exclusivity or exclusive co-occurrence between *EGFR* amplification and *CDKN2A* deactivation in this cohort.

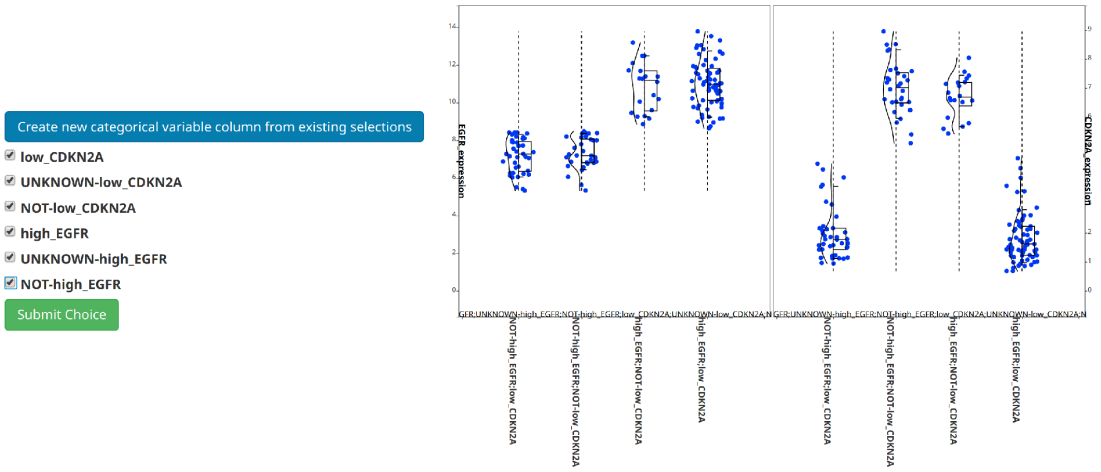

Figure 3.52: Definition of a categorical variable from *EGFR* and *CDKN2A* expression status in the TCGA-GBM study. Left: selection of categorical variables to combine using the "Create new categorical variable column from existing selections". Right: *EGFR* (left subplot) and *CDKN2A* (right subplot) expressions for the created categorical variables.

As the comparison of of a subcohort selection to its complement is a fundamentally important operation in cohort data analysis, each new subcohort selection by default creates a categorical variable from its positive, negative and unknown counterparts. For instance, definition of the low-CDKN2A group would automatically create the lowCDKN2A-UNKNOWNlowCDKN2A-NOTlowCDKN2A categorical variable for further use.

This combination approach can also be applied in the reverse manner as a decomposition of categorical variables into individual mini-selections. For instance, the PAM50 category of the TCGA-BRCA study can be, with one user command transformed into the positive, negative and unknown states of the possible categories Luminal A, Luminal B, Basal, Normal and Her2 (Figure 3.53).

Figure 3.53: Decomposing the PAM50 subtype variable in the TCGA-BRCA (WGS) study into individual selections with corresponding inverse and unknown selections. Left: selection of the PAM50 subtype as the categorical variable to decompose into selections. Right: the resulting selections generated by the decomposition operation applied to the PAM50 subtype variable.

3.3.9 Operations on Subcohorts in EPISTEME

The powerful and flexible subcohort definition features of EPISTEME are used in a number of specialized advanced data visualizations visualizing data from subcohorts and comparing user-selected subcohorts in a pairwise manner. These features offer tightly integrated flexible data analysis, visualization and analysis features. The following sections explain the rationale, design decisions and features of each subcohort-based data analysis and visualization feature of EPISTEME.

For all features, the focus will be on the comparison of the Basal-rich cluster obtained from tSNE analysis of RNA-Seq data from the TCGA-BRCA study. For features involving genomic variants, the WGS cohort is used, whereas all other sections use the full cohort with available RNA-Seq data. In the reduced, WGS cohort, the Basal-rich cluster consists of 43 cases whereas the ”Not-Basal-rich” cluster corresponding to Luminal A, Luminal B and Her2 cases consists of 49 cases. In the full cohort, the case numbers are 188 and 903, respectively. The choice of this comparison is due to the diversity of variant landscapes, gene expression profiles and survival characteristics between these two subtypes of breast cancer.

3.3.9.1 ”Subcohort-wide Circos plots”

Cohort-wide Circos plots which were extensively discussed in Section 3.3.1 can intuitively be extended for the representation of subcohorts. Instead of the full cohort with all available in the EPISTEME instance of a cohort, only the genomic variants and genomic variant recurrence

data from a subcohort of interest is displayed.

Figure 3.54 shows the application of the described concept to the pilot comparison between the Basal-rich and NOT-Basal-Rich subcohorts of the TCGA-BRCA (WGS) study. From outside to inside, the recurrence layers correspond to: functional small variant recurrence, TAD-based copy number recurrence and TAD-based structural variant recurrence. For both cohorts and all recurrence layers, the recurrence axis maximum values are manually set to 49, the size of the larger of the two subcohorts instead of the default practice of normalization. The higher prevalence of chr17q copy number gains and structural variants as well as a higher frequency of *PIK3CA* mutations define the foremost obvious characteristics of the NOT-Basal-rich subcohort. The Basal-rich subcohort, on the other hand presents a higher frequency of SVs converging on the *PTEN* locus and mutations on the *TP53* gene.

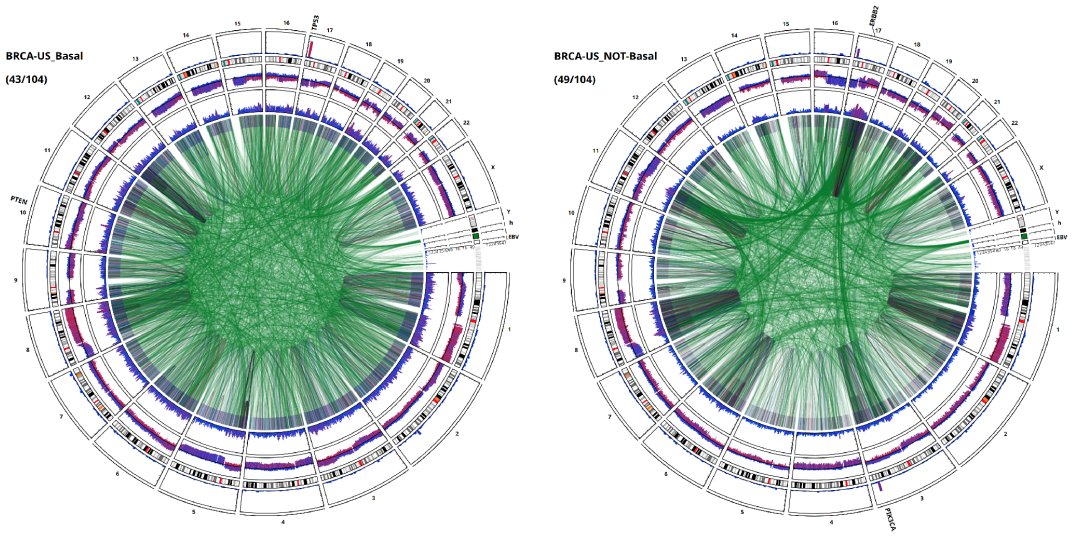

Figure 3.54: Genomic variant and variant recurrence landscapes of the Basal-rich (left) and NOT-Basal-Rich (right) subcohorts of the TCGA-BRCA (WGS) study

To make differences between subcohorts more apparent in an easy albeit simplistic manner, EPISTEME offers the possibility of "subtracting" subcohorts in Circos plots as set up in Figure 3.55 and displayed on 3.56. In this visualization, TAD or gene-based recurrences are subtracted in the user-defined directions and minimum resulting recurrence values are clamped at zero as "negative recurrence" is not defined. copy number recurrences are treated independently in the loss and gain direction. The results make the previously described differences much more apparent as well as showing differential copy number profiles involving chr5q, chr14q and chr15q which are enriched for losses in the Basal-rich subcohort.

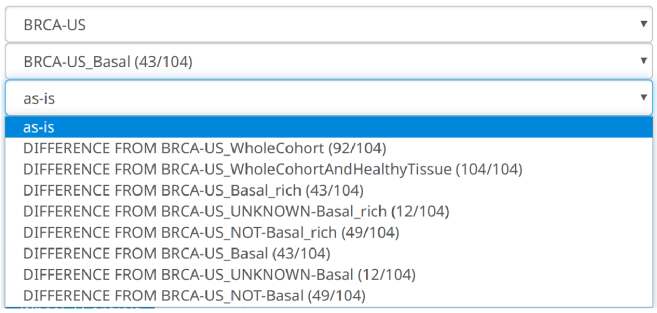

Figure 3.55: Subcohort Circos plot settings in EPISTEME

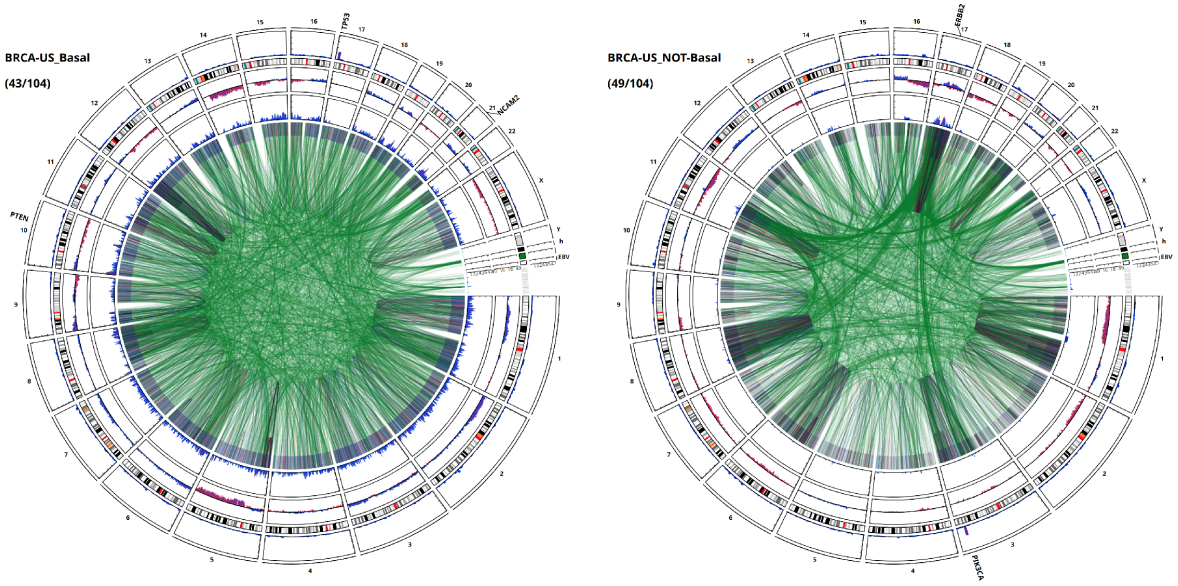

Figure 3.56: Differential subcohort Circos visualization reveals global differences between the genomic variant and variant recurrence landscapes of the Basal-rich and NOT-Basal-Rich subcohorts of the TCGA-BRCA (WGS) study

The subcohort subtraction procedure described here is a purely visual aid and works best when the subcohort sizes are comparable. In strongly imbalanced configurations, especially if the primary subcohort from which the secondary is subtracted is considerably larger, the results do not show differential behaviour in a convincing manner.

3.3.9.2 "Variant Mutex/Co-Occurrence Plots"

In order to account for the shortcomings of visual subcohort differentiation described in the previous section, EPISTEME implements a novel type of subcohort genomic variant landscape comparison analysis named "Variant Mutex/Co-Occurrence Plot". For this analysis, valid variant types and subcohorts to compare are selected in a flexible user selection (Figure 3.57). Selections of valid variant types are possible in levels from individual genes to TADs, cyto-

bands and chromosome arms. Thus, a comprehensive and well-controlled comparison of two subcohorts is facilitated.

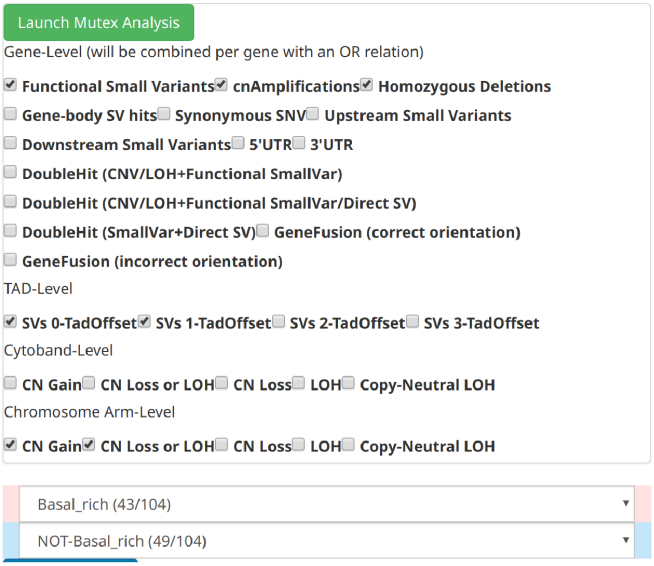

Figure 3.57: Variant Mutex/Co-Occurrence Plot settings in EPISTEME

The method relies on comparing variant frequencies both in terms of differences as in Section 3.3.9.1 and in terms of statistical significance. To this end, EPISTEME creates a contingency table with four conditions i) Positive cases in subcohort 1 ii) Negative cases in subcohort 2 iii) Positive cases in subcohort 2 iv) Negative cases in subcohort 2. Fisher's exact test is applied to this comparison table estimating the statistical significance of each variant frequency comparison. Figure 3.58 shows the result of this analysis, displayed on a volcano-like plot. The observations previously described for the differences between Basal-rich and not-basal-rich subcohorts are confirmed with *TP53* (37/43 vs 21/49), *PIK3CA* (3/43 vs 22/49), chr17q (carrying *ERBB2* and a number of other co-amplified genes such as *CDK12* (0/43 vs 20/49), the *PTEN* locus (25/43 vs 8/49), and the chromosomes 5,14 and 15. Moreover, recurrent SVs on chr21q21 (carrying *NCAM2*, 26/43 vs 5/49), chr3p14.2 (carrying *FHIT*, 29/43 vs 3/49), chr10p15.1 (27/43 vs 6/49) and chr19q13.2 (25/43 vs 5/49) present themselves as enriched in the Basal-rich subcohort.

The Basal-rich and non-basal-rich cohorts show different preferences for PI3K/Akt pathway activation with *PIK3CA* activating mutations and *PTEN* losses as described in the literature [530]. The enrichment of various chromosome arm level losses on chromosomes such as chr14 and 15 in the Basal-rich subcohort has been discussed in [531]. The lack of *ERBB2* amplifications is a well known hallmark of the Basal subtype of breast cancer [532] [522], with a strong enrichment of lack of chr17q12 amplifications in the Basal-rich subcohort consequently being an expected result. *FHIT* losses were previously shown to be enriched in oestrogen and progesterone receptor negative breast cancer [533]. In summary, EPISTEME's results shown here capture the established knowledge regarding Basal-like breast cancer and its distinct genomic alteration landscape compared to Luminal breast cancer.

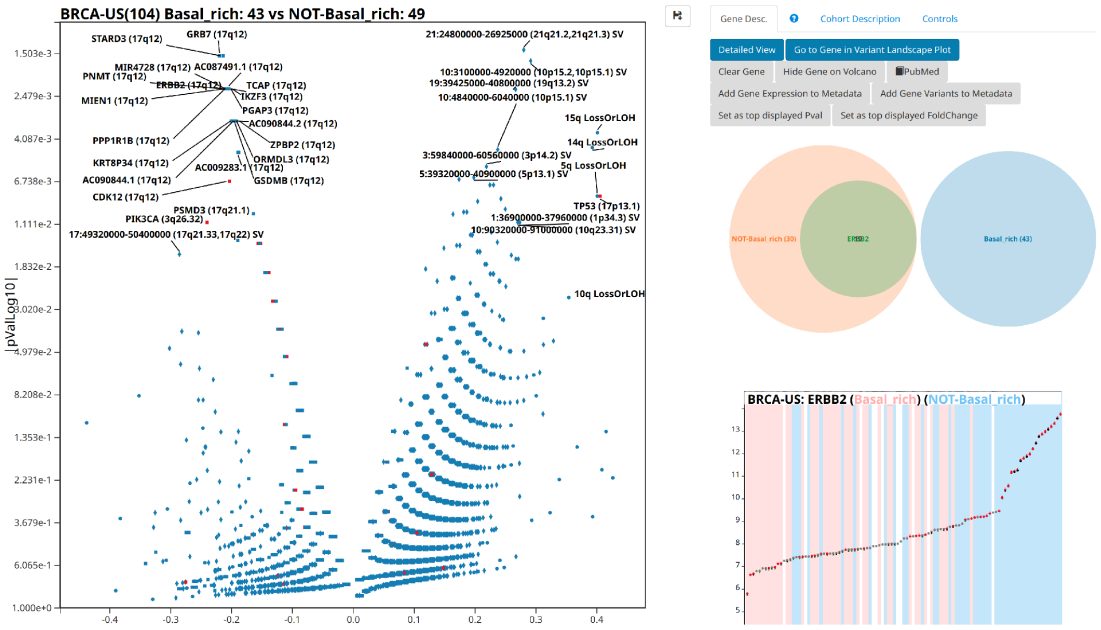

Figure 3.58: Comparative genomic variant landscapes of the Basal-rich subcohort versus the non-Basal-rich subcohort in the TCGA-BRCA (WGS) study. Left: Variant Mutex/Co-Occurrence Plot comparing the two subcohorts, where each data point corresponds to a compared gene, TAD or chromosome arm with the data points in the upper-right quadrant of the plot showing variants enriched in the Basal-rich subcohort and upper-left showing the enrichment in the non-Basal-rich subcohort. Right-middle: Venn diagram showing the exclusivity of *ERBB2* alterations to a subset of the non-Basal-rich cases. Right-bottom: Auxiliary gene expression plot for the *ERBB2* gene.

In a similar manner to "Variant-Expression Dysregulation Volcano Plots" described in Section 3.3.3, "Variant Mutex/Co-Occurrence Plots" are also a highly interactive visualization. Upon clicking on a data point, EPISTEME provides an auxiliary Venn diagram summarizing the contingency table that led to the displayed result. For genes, an auxiliary "Single-Phenotype Analysis Plots" is launched, which also displays the gene expression and (where available) RPPA quantities for the clicked gene of interest, with the subcohort information encoded as background shadings. The pilot example used for this auxiliary information feature is the *ERBB2* gene, with its very strong enrichment in the not-Basal-rich subcohort (Figure 3.58).

"Variant Mutex/Co-Occurrence Plots" formalize and make systematic the comparison of the genomic variant recurrence landscape of two user-selected subcohorts with user-selected genomic variant types in consideration. This visualization facilitates the exploration of rare variants and slight enrichments, moving beyond what differential Circos plots can offer.

3.3.9.3 Kaplan-Meier Plots

EPISTEME offers a simple survival visualization tool which supports single, double and higher order comparative visualizations of user-defined subcohorts with currently no statistical signif-

icance analysis features. To showcase the preparation of Kaplan-Meier curves, Figures 3.59 and 3.60 compare the RNA-Seq tSNE based Basal-rich vs rest subcohorts and the four PAM50 subcohorts (excluding the Normal-like subtype), respectively.

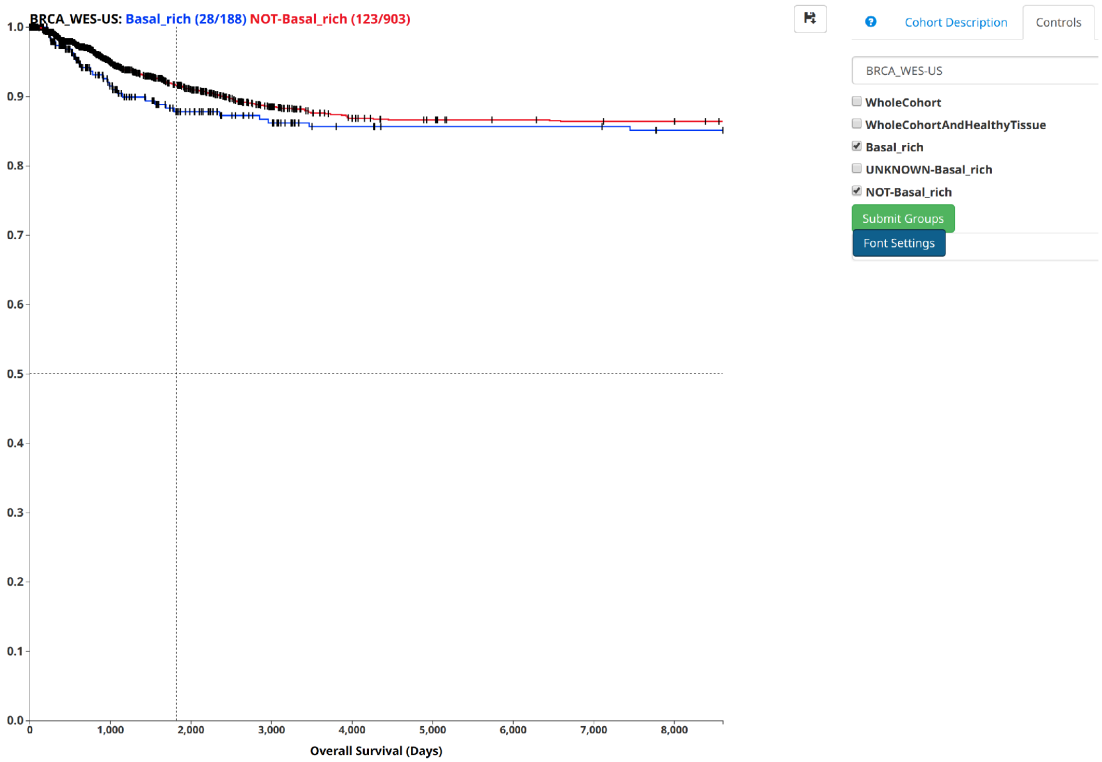

Figure 3.59: Comparative survival analysis of the Basal-rich subcohort versus the non-Basal-rich subcohort in the TCGA-BRCA study. Each included subcohort is automatically assigned a colour, whereas the censoring events are marked as black. The 5-year time-point is marked with a dashed line and the 50-percent mark in terms of recorded death events is also marked with a dashed line to guide interpretations of the displayed data.

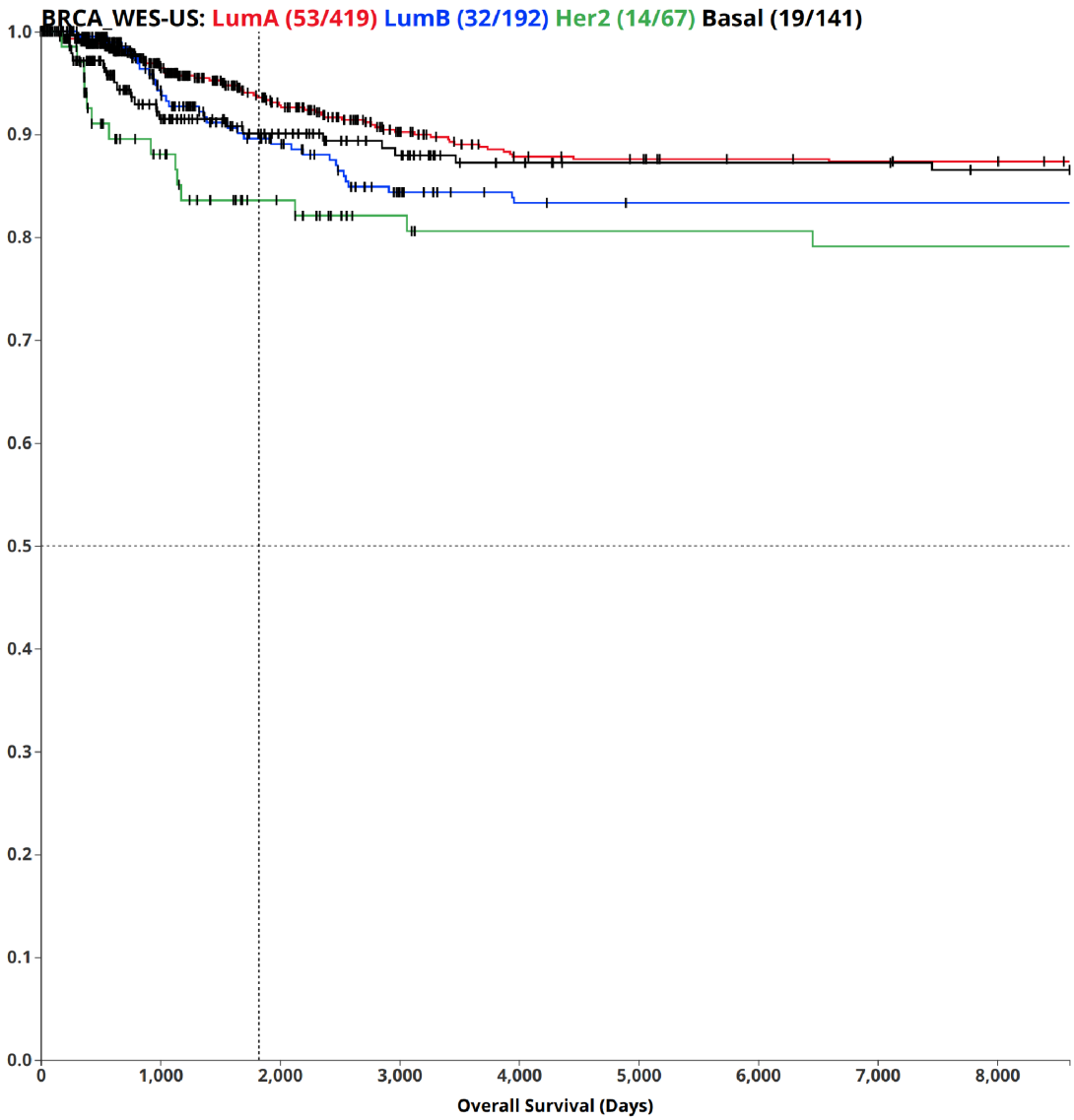

Figure 3.60: Comparative survival analysis of the PAM50 subtypes in the TCGA-BRCA study. Each included subcohort is automatically assigned a colour, whereas the censoring events are marked as black. The 5-year time-point is marked with a dashed line and the 50-percent mark in terms of recorded death events is also marked with a dashed line to guide interpretations of the displayed data.

Taking the 5-year mark as the anchor point for comparisons, the results show a similar overall pattern to those presented in ([534], Figure 2), where the demonstration was made on oestrogen receptor status instead of the Basal-rich tSNE RNA-Seq based subcohort used here: The Basal-rich subcohort shows a poorer overall survival. This observation seems to be driven by the prominence of the largest non-Basal subcohort, Luminal A, which shows a favourable overall survival characteristic. On the other hand, Luminal B and Her2 subcohorts show poorer survival than the Luminal A subcohort, where the Her2 subcohort showed the poorest survival characteristics in this dataset.

3.3.9.4 *"Differential Expression Volcano Plots"*

Comparing the global expression characteristics of RNA transcripts (or peptide quantities) between two sample groups is a central task in computational biology and cancer omics. Identifying the overexpressed and underexpressed genes between two groups or conditions allows the assignment of cell identity, identification of the transcriptomic programme used in the analysed sample groups, determination of the dysregulated genes between two conditions, making this feature essential to implement in a user-accessible manner in an interactive data analysis system. EPISTEME implements a module for differential gene expression analysis with user-defined settings (Figure 3.61).

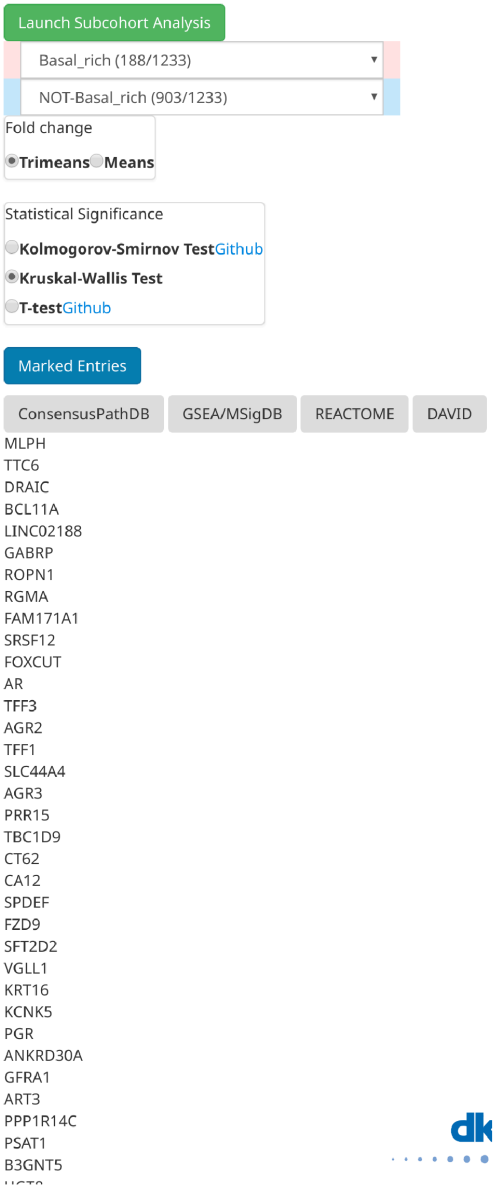

Figure 3.61: Differential expression and pathway enrichment analysis settings in EPISTEME

The Basal-rich and non-Basal-rich subcohorts of the TCGA-BRCA study as identified by tSNE based RNA clustering on the top 500 most variable genes in the cohort is the pilot com-

parison for showcasing differential gene expression analysis. Due to the strong separation of these two subcohorts already with a top 500 gene based dimensionality reduction, they are expected to be significantly different in terms of global gene expression profiles. The two groups also show strongly different methylation profiles in top-5000 most variable probe based dimensionality reduction analyses with all tested parameters of the two used dimensionality approaches tSNE and UMAP (Figures 3.37 and 3.40) and consequently are very likely to stem from a different methylation state arising from a different starting cell type. Therefore, different transcriptional programmes are expected to govern these two cell types with a large number of differentially expressed genes owing purely to gene identity. This hypothesis is confirmed with a large number of significantly differentially regulated genes (Figure 3.62). Top outliers include a number of well known genes such as *ESR1*, *AR1*, *ERBB2*, *PGR*, *FOXC1*, *FOXA1* and *GATA3*. Dissecting further how strongly and with which profile these genes are differentially expressed is possible with auxiliary "Single-Phenotype Analysis Plots", where the auxiliary plot in Figure 3.62 shows the gene expression profile of the *FOXA1* gene, with shading colours guiding subcohort identification. For this gene of interest, *FOXA1*, the observed profile shows a strong and significant suppression of the two compared subcohorts with near perfect separation.

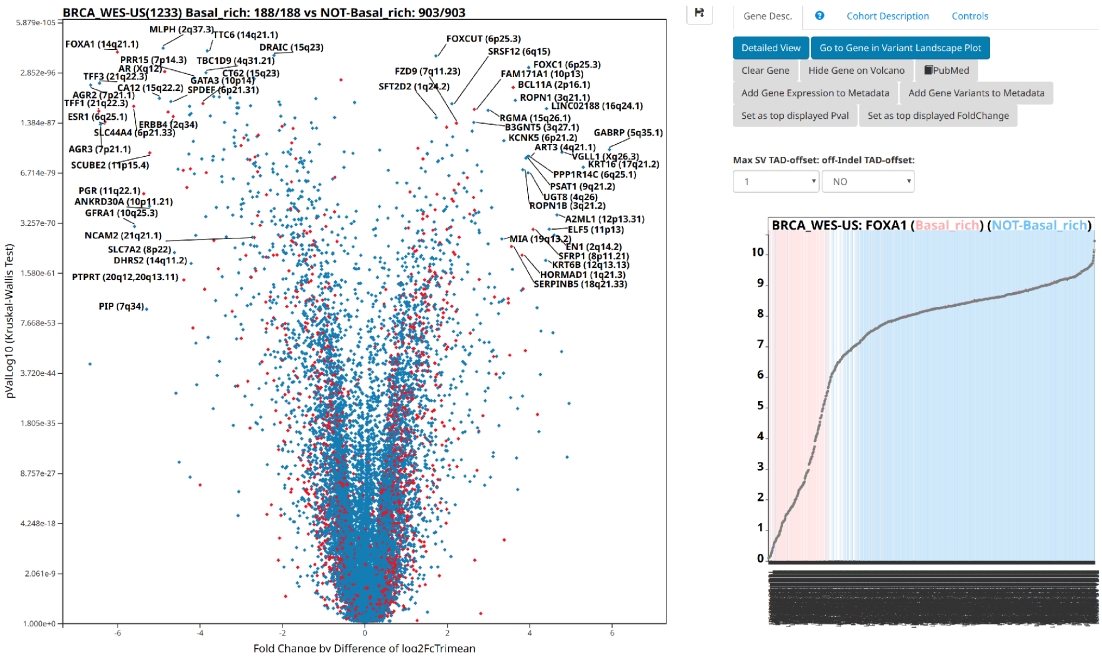

Figure 3.62: Differential gene expression analysis between the Basal-rich and non-Basal-rich subcohorts in the TCGA-BRCA study. Left: Differential gene expression volcano plot. Right-top: user interaction features for a selected gene.Right-bottom: Auxiliary gene expression plot for the *FOXA1* gene, where donors with pink-shaded backgrounds are in the Basal-rich subcohort whereas the donors with light blue-shaded backgrounds are in the non-Basal-rich subcohort.

Investigating the chosen genes *ERBB2*, *ESR1*, *PGR* and *FOXC1* (Figure 3.63) reveals the following characteristics:

- *ESR1* expression follows a likely cell type dependent profile very similar to the previously discussed *FOXA1* gene with near perfect separation of the two compared subcohorts.
- *PGR* expression follows a likely cell type dependent profile, but not a cell-identity defining profile for these two subcohorts. Its suppression only in a subset of cases in the Basal-rich subcohort is in support of the established knowledge that triple (*ERBB2*, *ESR1*, *PGR*) negative breast cancer is distinct from the Basal-like subtype of breast cancer [535].
- *ERBB2* expression follows a bimodal profile with high expressors exclusively observed in *ERBB2*-amplified cases in the non-Basal-rich subcohort, suggesting a variant dependent profile.
- *FOXC1* expression follows a near-perfect separation similar to that observed in the *FOXA1* gene, in the opposite direction. Its expression is exclusively, strongly and significantly high in the Basal-rich subcohort. Due to its function as a developmental transcription factor, it can be hypothesized to be a master regulator in the Basal-rich subcohort's cell of origin, which is in line with established knowledge regarding this gene and cell type [536].

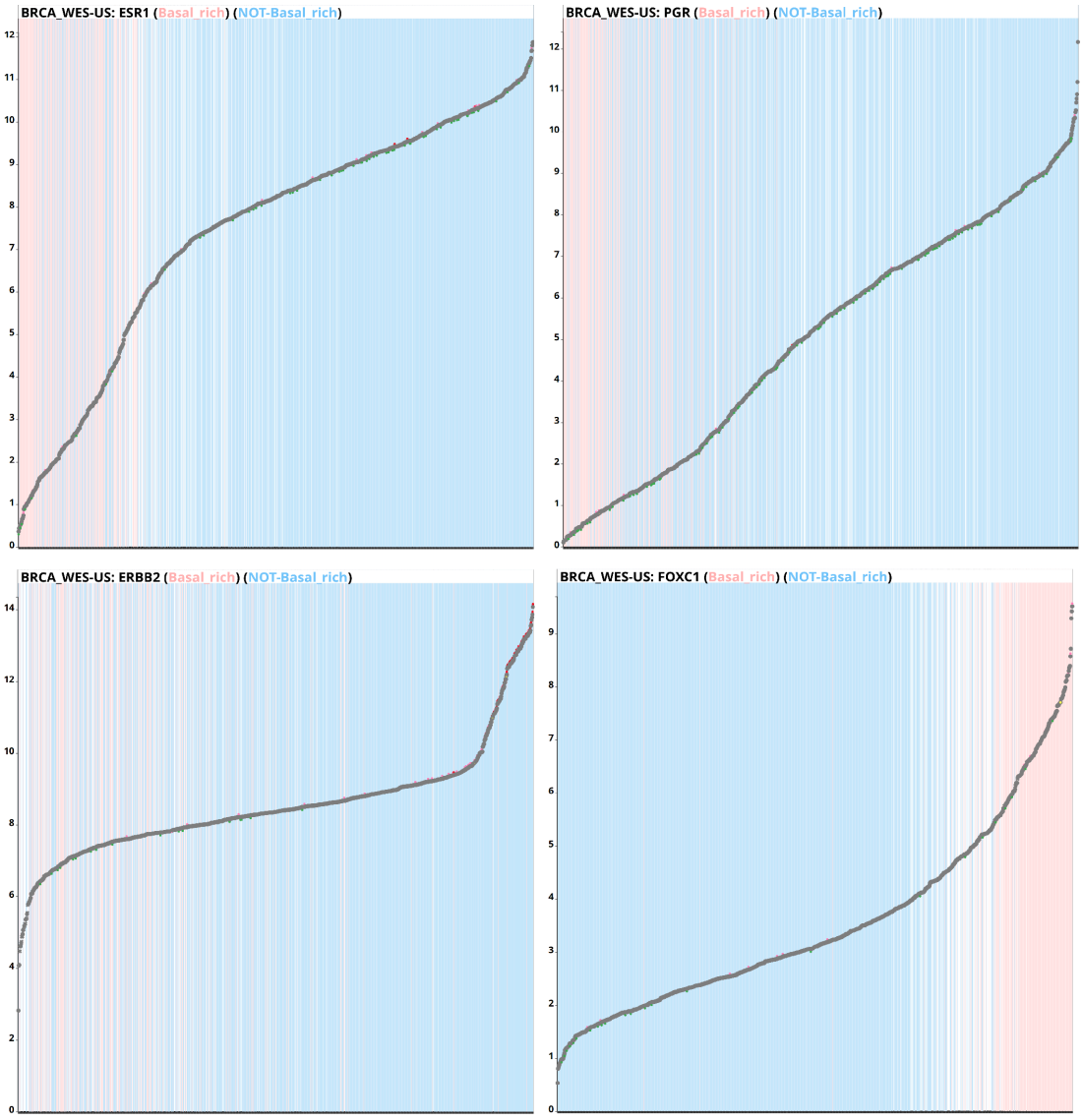

Figure 3.63: *ESR1*, *PGR*, *ERBB2* and *FOXC1* expressions across the Basal-rich and the non-Basal-rich subcohorts in the TCGA-BRCA study, where donors with pink-shaded backgrounds are in the Basal-rich subcohort whereas the donors with light blue-shaded backgrounds are in the non-Basal-rich subcohort.

Selecting all genes showing between -3 and +3 log2-Trimean fold change in this comparison and using EPISTEME's integration to pathway enrichment analysis tools on the example of MSigDB (Figure 3.64) shows the differentially expressed genes to represent known gene sets capturing difference of Breast cancer types.

Converted 336 submitted identifiers into 289 entrez genes. click here for details.

Collections	# Overlaps Shown	# Gene Sets in Collections	# Genes in Comparison (n)	# Genes in Universe (N)
C1, C2, C3, C4, C5, C6, C7, H	10	17810	289	45956

Click the gene set name to see the gene set page. Click the number of genes [in brackets] to download the list of genes.

Color bar shading from light green to black, where lighter colors indicate more significant FDR q-values (< 0.05) and black indicates less significant FDR q-values (>= 0.05).

Save to: Excel | GenomeSpace

Gene Set Name [# Genes (K)]	Description	# Genes in Overlap (k)	k/K	p-value	FDR q-value
SMID_BREAST_CANCER_BASAL_UP [648]	Genes up-regulated in basal subtype of breast cancer samles.	101		$3.07\ e^{-112}$	$5.46\ e^{-108}$
FARMER_BREAST_CANCER_BASAL_VS_LULMINAL [330]	Genes which best discriminated between two groups of breast cancer according to the status of ESR1 and AR [GeneID=2099;367]: basal (ESR1- AR-) and luminal (ESR1+ AR+).	81		$2.55\ e^{-106}$	$2.27\ e^{-102}$
DOANE_BREAST_CANCER_ESR1_UP [112]	Genes up-regulated in breast cancer samples positive for ESR1 [GeneID=2099] compared to the ESR1 negative tumors.	60		$2.65\ e^{-103}$	$1.57\ e^{-99}$
SMID_BREAST_CANCER_RELAPSE_IN_BONE_DN [315]	Genes down-regulated in bone relapse of breast cancer.	78		$1.41\ e^{-102}$	$6.3\ e^{-99}$
SMID_BREAST_CANCER_LUMINAL_B_DN [564]	Genes down-regulated in the luminal B subtype of breast cancer.	88		$4.98\ e^{-97}$	$1.77\ e^{-93}$
SMID_BREAST_CANCER_BASAL_DN [701]	Genes down-regulated in basal subtype of breast cancer samles.	89		$7.41\ e^{-90}$	$2.2\ e^{-86}$
DOANE_BREAST_CANCER_ESR1_DN [48]	Genes down-regulated in breast cancer samples positive for ESR1 [GeneID=2099] compared to the ESR1 negative tumors.	37		$6.85\ e^{-73}$	$1.74\ e^{-69}$
SMID_BREAST_CANCER_RELAPSE_IN_BONE_UP [97]	Genes up-regulated in bone relapse of breast cancer.	38		$1.71\ e^{-58}$	$3.8\ e^{-55}$
SMID_BREAST_CANCER_LUMINAL_B_UP [172]	Genes up-regulated in the luminal B subtype of breast cancer.	44		$4.72\ e^{-58}$	$9.34\ e^{-55}$
LIEN_BREAST_CARCINOMA_METAPLASTIC_VS_DS_DUCTAL_DN [114]	Genes down-regulated between two breast carcinoma subtypes: metaplastic (MCB) and ductal (DCB).	34		$1.48\ e^{-47}$	$2.64\ e^{-44}$

Figure 3.64: Pathway enrichment analysis of differentially expressed genes between the Basal-rich and non-Basal-rich subcohorts in the TCGA-BRCA study

EPISTEME offers a flexible differential expression analysis tool where users can determine the statistical significance estimation strategy as well as the use of a outlier-robust or outlier-sensitive fold change method. The results of this analysis are tightly coupled to intuitive user interactions, which facilitates further exploration of outlier data points by auxiliary single-phenotype plots or with integrations to external pathway enrichment analysis tools. Users can go move from auxiliary single-phenotype plots to fully-featured single-phenotype plots in order to view single genes with full detail including specific observations on detected variants such as SVs.

For many computational biology or cancer omics researchers, differential expression analysis is probably the first application of subcohort comparisons that comes to mind. Here, it is presented as the last of EPISTEME's subcohort analysis features. The reason for this is the connection to the next section, namely correlation analysis, which is a closely related and complementary analysis to differential expression analysis.

3.3.10 Correlation Analysis in EPISTEME

Differential gene expression analysis identifies genes that are differentially expressed between two conditions or subcohorts. A natural question that arises upon identifying genes that are

differentially expressed if and how they are co-regulated. For instance, considering genes *A* and *B* and subcohorts I and II, *A* and *B* can both be strongly differentially expressed between the two subcohorts with both being exclusively highly expressed in I and vice versa in II. This does not necessarily mean that all cases within subcohort I that express *A* medium-high, high, or very high levels also express *B* with a similar pattern and vice versa. The existence of such an agreement / correlation between two genes, regardless of a subcohort relationship, suggests co-regulation where one of the genes might be activating the other or that they are co-activated by a common upstream gene. The existence of an opposite relationship might suggest a suppression of one by the other.

Such relationships can be studied by correlation analysis. EPISTEME analyses two measures for correlation analysis, visualizing them in a volcano-like plot: Linear correlation (Pearson Correlation), which estimates the accuracy of a linear model fit for the expression profiles of two genes and rank-based correlation (Spearman Correlation), which estimates the consistency of magnitude ranks for the same cases' gene expressions for two genes of interest. Linear correlation quantifies the strength of the co-regulation, and can be considered to be similar to fold change. Rank-based correlation quantifies a rank-based consistency, and can be considered to be similar to rank-based statistical significance. By default, EPISTEME visualizes correlations as $x = |\rho_{Pearson}|$ and $y = \rho_{Spearman}$, which creates a volcano-like plot analogous to differential gene expression analysis volcano plots.

EPISTEME analyses correlations of quantitative variables versus gene expressions in a cohort or confined to a selected subcohort of interest. The quantitative variables which are to be used as correlation anchors can themselves be gene expressions or any other quantitative metadata variables such as patient age or survival. Rank-based correlation analysis starts to become unreliable when there are a large number of ties leading to tied ranks. EPISTEME uses a shuffling based calculation of Spearman Correlations to alleviate this problem, where the input expressions for the anchor gene are shuffled $2 * Ties$ times up to a maximum number of 100 iterations.

3.3.10.1 1-vs-all Correlation Analysis based on Gene Expression

The pilot analysis to showcase the correlation analysis features of EPISTEME is the correlation of the gene *FOXA1* versus all other genes in the TCGA-BRCA study, confined to tumour samples excluding normal tissue specimens. *FOXA1* constitutes an ideal showcase gene because its expression is nonzero for all samples and as a transcription factor, its expression is expected to be highly correlated with its targets and regulators. Previously shown results in the context of differential gene expression analysis in Section 3.3.9.4 suggest that the *FOXC1* and *FOXA1* genes might be potent master transcriptional regulators of the Basal-like and Luminal cell types of breast cancer, respectively. In order to investigate if these two genes are indeed mutually exclusive in terms of transcript factor usage and what genes are co-regulated by *FOXA1*, a correlation analysis is the appropriate tool.

Figure 3.65 shows very strong linear and ranked-based correlation and anti-correlation scores in the global correlation profile analysis of the *FOXA1* gene, suggesting a role in direct transcriptional regulatory activity. *ESR1*, *GATA3*, *TTC6*, *TTC8*, *AR* are among the co-regulated

genes, with *PGR* showing modest positive correlation. On the other hand, *FOXC1*, *HAPLN3*, *BCL11A* and *FOXCUT* are strongly anti-correlated.

The auxiliary helper plots for volcano-like correlation plots in EPISTEME are "Two-Phenotype Analysis Plots" described in Section 3.3.2.1, which provide an intuitive visualization of the two correlated or anti-correlated gene expression profiles (Figure 3.65, where *TTC6* shows a strong positive correlation with *FOXA1* expression).

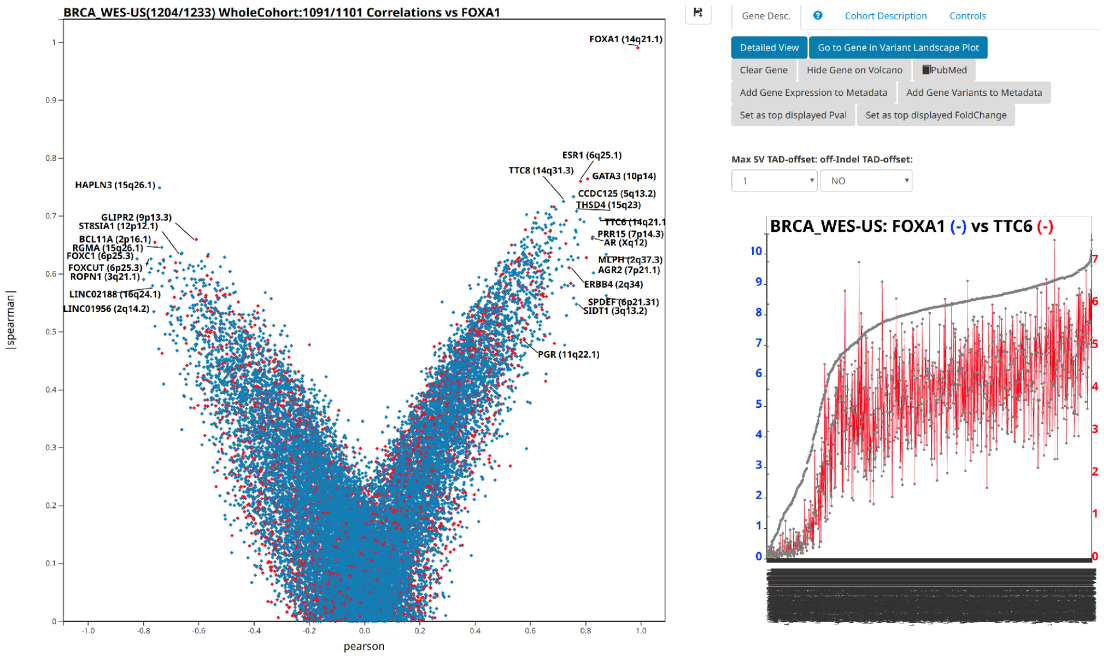

Figure 3.65: Spearman and Pearson Correlations of all genes vs *FOXA1* expression in the TCGA-BRCA study. Left: Correlation plot for genome-wide correlations of gene expressions vs *FOXA1* expression, where he x-Axis shows the Pearson correlation coefficient whereas the y-Axis shows the absolute value of the Spearman correlation coefficient. Right-bottom: auxiliary two-phenotype correlation plot with *FOXA1* as the anchor gene and *TTC6* as a representative gene showing positive correlation.

Figure 3.66 shows the investigation of selected top correlated and anti-correlated genes with "Two-Phenotype Analysis Plots". The selected genes for strong positive correlation characteristics are *ESR1* and *GATA3* which show an high correlation with *FOXA1* but only where *FOXA1* is high, and not for *FOXA1*-low cases which are enriched in the Basal-like subtype. The same is true for the *FOXC1* and *HAPLN3* genes where the anti-correlation characteristics are only observed when *FOXA1* is active.

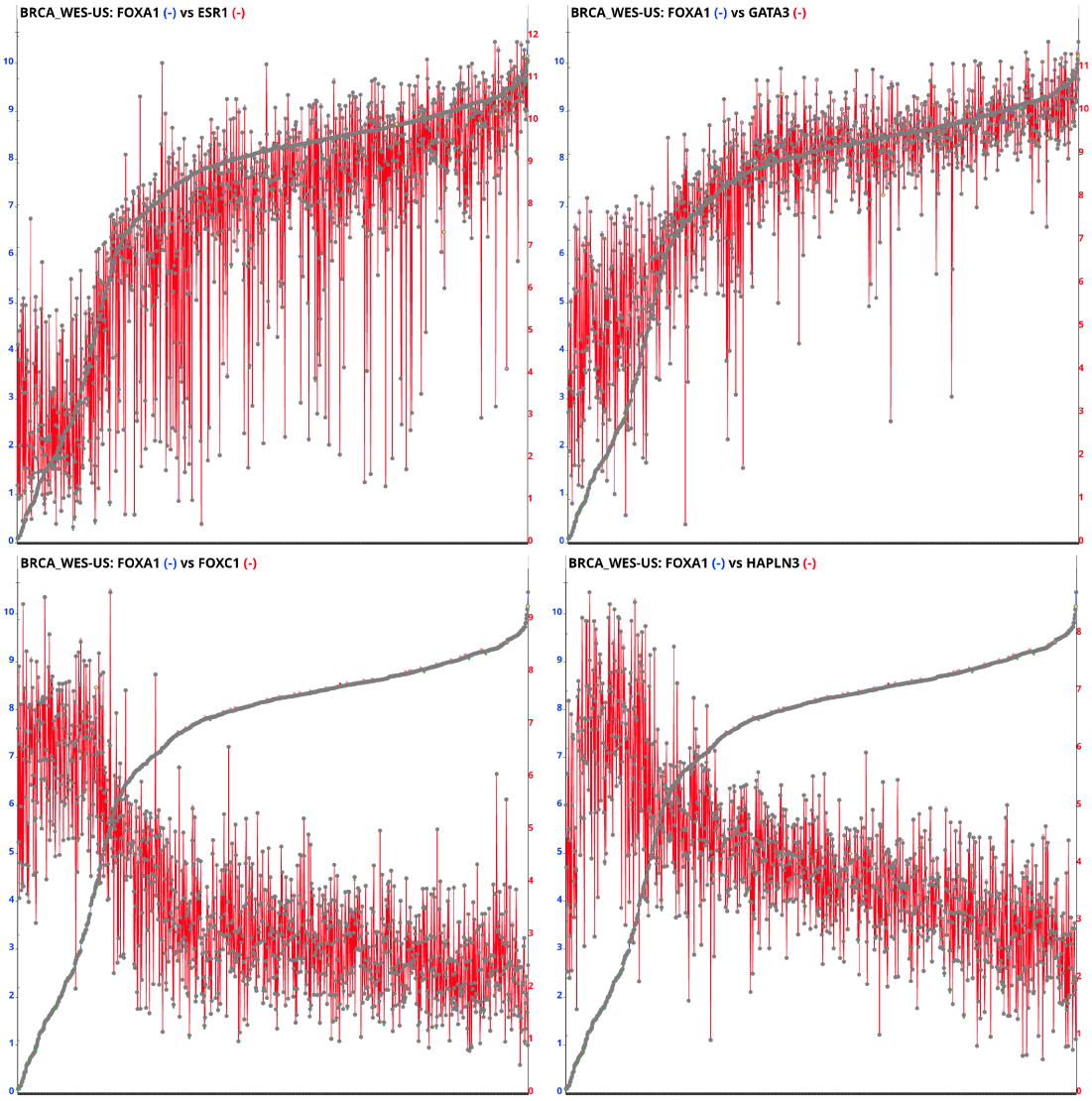

Figure 3.66: Correlations of the *FOXA1* gene expression versus the expressions of the *ESR1*, *GATA3*, *FOXC1* and *HAPLN3* genes in the TCGA-BRCA study.

Overall, these results confirm the role of the *FOXA1* as core transcription factor which gives the Luminal reast cancer cell of origin its gene identity [537] [538] [539]. The anticorrelation between *FOXC1* and *GATA3* and *ESR1* has been investigated with molecular biological assays [540], where the mechanism of action has been shown to be competition for *GATA3* binding sites. Furthermore *GATA3* has been shown to act upstream of *FOXA1* [541], suggesting that it could be the bridge that drives the anticorrelation between *FOXA1* and *FOXC1*.

3.3.10.2 1-vs-all Correlation Analysis based on Arbitrary Quantitative Data Fields

Similar to gene expressions, any quantitative metadata field can be used as an ”anchor” in a ”1-vs-all” correlation analysis. The only restriction is that all quantities of the selected quantitative metadata field should be non-negative. Figure 3.67 shows the ”1-vs-all” correlation analysis where patient at diagnosis age is compared to all gene expressions in this cohort. Interestingly,

ESR1 shows the highest correlation in terms of ranks (Spearman) and is very close to be the absolute top in terms of non-rank-based correlation (Pearson), where only the *DBX2* gene shows a higher anticorrelation coefficient. It should be noted that all correlations are rather low, indicating a modest effect of patient age at diagnosis on gene expressions. The high correlation with *ESR1* likely shows the effects of the younger age of Basal-like breast cancer patients, which do have low *ESR1* expression. [542]

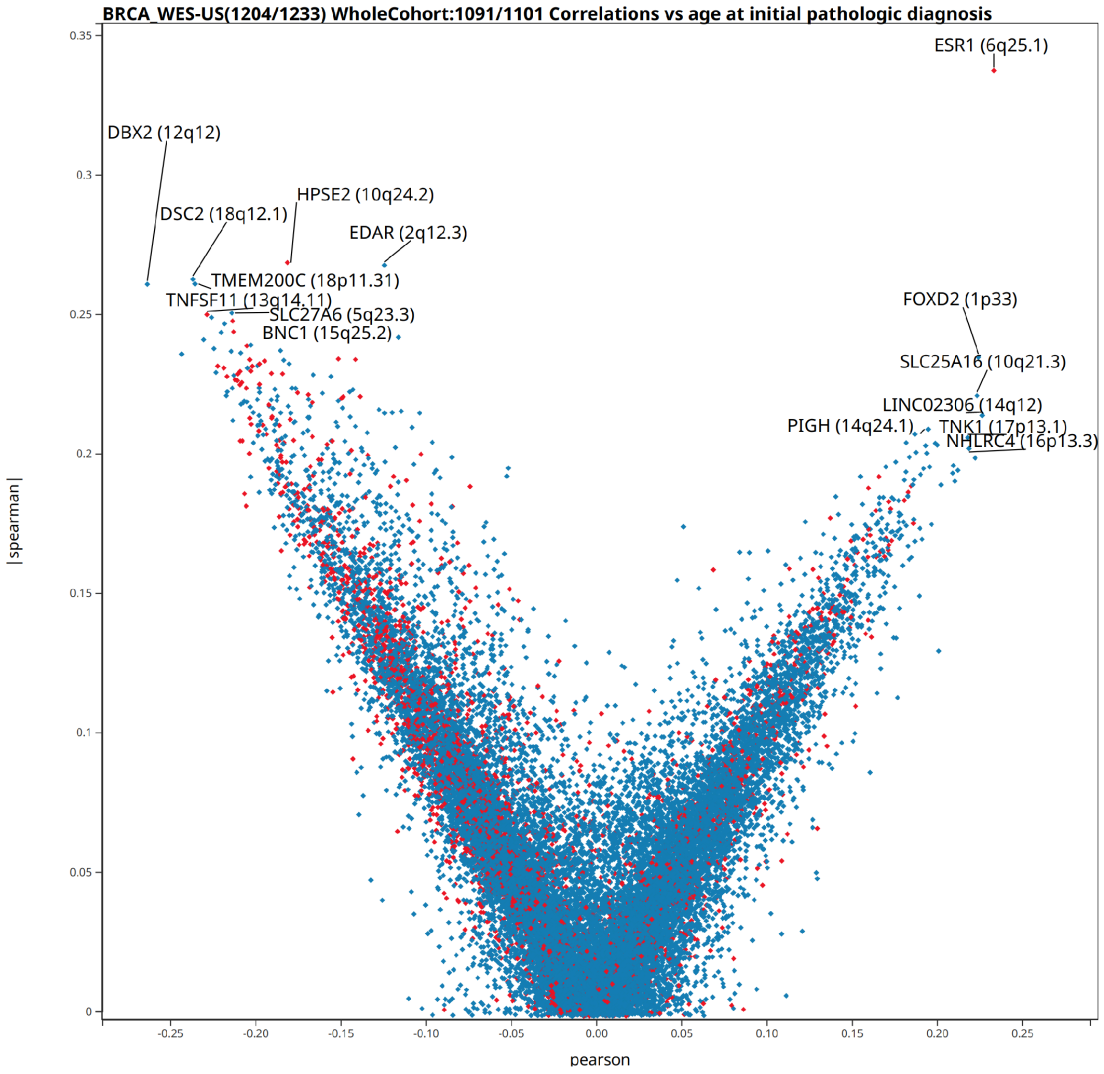

Figure 3.67: Spearman and Pearson Correlations of all genes vs age at diagnosis in the TCGA-BRCA study. Age at diagnosis is taken as a quantitative variable analogous to a gene's expression values.

The correlation features of EPISTEME facilitate the study of mutually exclusive gene activities or targets of transcriptional activators or repressors. It is a useful feature following results of "Variant-Expression Dysregulation Volcano" analysis where aberrantly activated or suppressed genes can be investigated for direct downstream effects. It is also a useful feature following results of "Differential Expression Volcano" analysis as presented here.

3.4 Discussion

EPISTEME is a comprehensive omics data analysis and visualization tool offering a broad range of visualization types building on simple, general visualizations to highly integrative visualizations with rich user interactions bringing together different omics data layers. It is aimed to improve the analysis and sharing of omics data, facilitating the communication between computational biologists/bioinformaticians and experts from biological backgrounds, by making advanced data analysis tools easily accessible to a broad audience without programming knowledge required. It runs in the web-browser with all analysis features running on the client-side with high performance, which removes the need for intensive server-side computations. Its interactions are intuitive and the create publication-quality, vector-based visualizations with modifiable font settings.

Comparing EPISTEME in a self-assessed manner to its closest competitors (Figure 3.68), one can say that it can still grow in terms of data analysis features (R2 has more extensive survival analysis features, OncoScape has more diverse dimensionality reduction features and a more interactive spreadsheet), visualizations(R2 and OncoScape have more advanced Heatmaps, cBioPortal has offers Oncoprints) and data scope (R2 has support for epigenetics data analysis). However, its integrative analysis features across omics data layers, and the breadth of its interactive features such as highly interactive volcano plots with auxiliary plots are currently not matched by any of its competitors. In particular, some of EPISTEME's integrative analysis features such as "Variant-Expression Dysregulation Volcano Plots" and "Variant Mutex/Co-Occurrence Plots" are novel ways to visualize important omics data analysis concepts.

Tool	R2	OncoScape	Vizome	cBioPortal	EPISTEME
Data analysis features	xxxx	xxxx			xxx
Data visualization features	xxx	xxx	xxx	xx	xxxx
Integrative omics analysis	x	x	x		xxxx
Client-side computing	xxx				xxx
Subcohort definitions	x	x	xx		xxxx
Rich user interactions	x	xx	x	x	xxxx
Support for user-provided data	x	x		x	
Epigenetics data visualization	x				
Genome Browser	x	x	xx	xxx	

Figure 3.68: A comparison of EPISTEME's features to its closest competitors

In addition to EPISTEME's ability to capture established biological knowledge, its use as an accessible tool to reach novel biological insights in the context of diverse disease types

and hypotheses has not only been shown in the novel enhancer hijacking candidates in gastric adenocarcinoma in this chapter, but also in a number of DKFZ projects which will be presented in the next and final Chapter 4 of this dissertation.

3.4.1 Development Roadmap for EPISTEME

EPISTEME is a work under constant development. Thanks to the excellent availability of genomics data from international consortia and the DKFZ Heidelberg's wide spectrum of projects, the flexibility of the D3.js framework under JavaScript, and the growing power of SVG rendering and interactivity, it is a technically challenging but feasible and rewarding endeavour to implement new data types, visualizations and interactions into EPISTEME. The following is a summary of planned near-term features and possible strategies for EPISTEME in ascending order of estimated technical challenge of implementation. In general, for each planned feature, the design effort starts with the classification of the data type as genomic variant, metadata information, or phenotype and deciding on an appropriate mode of visualization.

- Gene fusion data: Including gene fusion data in EPISTEME is an almost trivial task: fusion genes are clearly genomic variants and each chimeric fusion transcript called from RNASeq data can be represented as an SV and the gene labelling features shown in Section 4.3.3.1 for V(D)J rearrangements can be used to clearly mark the candidate fusion partners. Similarly, the fusion partner of a gene can be added in the single-gene visualization presented in Section 3.3.2. This is considered a very near-term goal to add as a data-layer to EPISTEME.
- Dynamic variant-expression dysregulation volcano plots: In their current implementation, the calculations in Sections 3.2.8.1 and 3.3.3 have a significant technical shortcoming: due to the intensive computation required to run the statistical tests especially considering the presented sweep-based approach, EPISTEME displays pre-computed values for the statistical significance and fold change values. Unfortunately, this design decision takes away the potential ability of dynamically generating different volcano plots based on subcohorts, a central concept in EPISTEME. This feature will need to be extended in the short term to address this shortcoming.
- Proteomics data: Though its pre-processing steps such as normalization is not comparable to RNAseq with regards to the requirements and algorithms [543], protein abundance data in its post-processed state, is in principle not different from gene expression data from a technical perspective. The same SQL-based data storage and single-gene or cohort-wide-volcano plot visualizations can be directly adopted to be used for proteomics data as phenotype data. The main issue here is the availability of data: none of the ICGC or DKFZ cohorts used in the development of EPISTEME had a satisfactory proteomics component apart from some TCGA cohorts with a very limited number of proteins measured by reverse phase protein lysate microarrays, a targeted technique which is not suitable for discovery analysis. Once a cohort is available with a sufficiently rich proteomics component in addition to genomics and transcriptomics data tracks, it will be straightforward to integrate proteomics data into EPISTEME with a simple set of

features focusing on protein abundance. Such plans are already under way in the DKFZ Neuroblastoma Genomics research programme.

- Germline variants: The challenge of including germline data in EPISTEME is not a technical challenge within EPISTEME itself but rather upstream, in the definition of such variants. In general, germline analysis is technically more challenging because it is more difficult to filter variant calling artefacts due to the absence of a paired "normal/control" data, and the difficulty of deciding what is a rare germline variant of likely significant biological impact. Given the availability of "clean" germline variant data, the existing variant visualization and analysis of infrastructure can be adapted to cover germline variants. Currently, there are efforts under way to improve the variant calling algorithms used in [211] and in this version of EPISTEME to call rare germline small variants of likely functional impact, after which their results will be ported into EPISTEME giving users to run somatic-only, germline-only and somatic-germline-merged integrative variant analysis.
- Methylation data: Inclusion of methylation data in EPISTEME is an interesting challenge due to the definition of the data class and the diversity of data sources: First, methylation data can be treated both as genomic variants as in Variably Methylated Regions(VMRs) and as phenotypes such as the methylation value of a single probe of a methylation array. Second, methylation data can originate either from targeted methylation arrays highly enriched for regulatory regions and gene promoters, or the genome-wide assay Whole-Genome Bisulfite Sequencing (WGBS).

  For cohorts where only methylation array data is available, EPISTEME could adopt a phenotype-centric approach, and treat each probe as a gene in a genomic variant-phenotype integrated analysis, where the probe methylation for a given patient would be the phenotype akin to gene expression for a given patient. However, as there are many more probes (450.000 to 1.000.000 depending on the technology) than genes, it will be a challenge to maintain the performance of the SQL backend of EPISTEME.

  For cohorts where only WGBS data is available, VMR calls can be sourced externally as with any other class of genomic variants, and its recurrence can be analysed on a TAD-basis. For WGBS, in the absence of specific probes, deciding on what values to use as a phenotypic readout for each gene is an open question.

  With adding methylation data into EPISTEME, the challenges are mostly conceptual and technical as data availability is already good and is getting better with the increasing availability of WGBS assays. For data visualization, the previously published tool MethCNA could offer design ideas [544].
- Improved analysis of survival data: In its current state, EPISTEME's survival analysis features are fairly rudimentary, and better statistical tests and visualizations are needed. Correlation of patient survival with genomic variants is prone to the influence of confounders: i) age of diagnosis, ii) disease stage at diagnosis, iii) disease subtype, iv) ethnic background, v) treating centre and treatment strategy.. The implemented survival analysis features should use appropriate methods to account for such confounders. R2's advanced survival features could offer design ideas.

- Analysis of user-provided cohorts: Ultimately EPISTEME could serve the cancer genomics community not only by hosting public datasets but also by providing a service for processing and visualizing user-provided private datasets. Due to time and manpower constraints, this interesting but difficult challenge has not been explored and is not a short or mid-term priority as accepting user input requires i) the implementation of strong security measures to protect the data on the EPISTEME server and the user's private data from leaking, ii) sanitizing the user input based on a threshold of maximum acceptable cohort size or variant data size, iii) providing parsers for any realistic combination of variant callers rather than only mpileup, platypus, ACESeq and SOPHIA, iv) either having substantial server-side processing capacity dedicated to processing user inputs or a huge effort to implement the whole stack of cohort processing in JavaScript.

CHAPTER 4
SOPHIA-EPISTEME INTEGRATION IN DKFZ CANCER GENOMICS PROJECTS REVEALS NOVEL DISEASE SUBTYPES AND INSIGHTS ACROSS CANCER TYPES

4.1 Introduction

We introduced and reported results from a fast, efficient and sensitive Structural Variant(SV) detection algorithm SOPHIA (Chapter 2), and a comprehensive integrative and interactive omics data analysis and visualization tool EPISTEME (Chapter 3). The focus in describing each tool was their design motivations, design principles, major features, unique advantages and potential avenues for improvement. In both chapters, each SV detection, omics data analysis or omics data visualization concept was described on established cohorts, or individual observations where the concept was already established. We thus aimed to reduce the complexity of concepts' presentation by not introducing new concepts and new biological findings simultaneously. Furthermore, the biologically established ("pilot") observations also served as a confirmation of the validity of the used analysis and visualization approaches.

This chapter presents novel findings using the SOPHIA algorithm and the EPISTEME platform from unpublished projects of the DKFZ cancer omics research programme with an integrative omics data analysis strategy. Making use of the well-described concepts of SVs, enhancer hijacking, EPISTEME's volcano plots, differential gene expression analysis, gene correlation analysis, variant mutex analysis, dimensionality reduction (e.g. tSNE) features, a step-by-step dissection of three different diseases is shown, yielding novel subtypes with potential implications on disease biology and treatment. The projects are organized in three case studies:

1. Late-stage, multi-refractory multiple myeloma, a haematological adult malignancy
2. Acute Myeloid Leukaemia (AML) with Chromosome 7q-monosomy (7q-AML), a haematological adult malignancy
3. Neuroblastoma (NB), a solid paediatric peripheral nervous system malignancy.

In each of these projects, I served as a leading bioinformatician contributor, having responsibilities in data pre-processing, quality control, processing, interpretation and presentation. In this dissertation, the sole focus is on the key findings made with the SOPHIA-EPISTEME integration, which allowed us to introduce new disease subtypes and put forward hypothesis on their development.

4.2 Common Methods

All projects described in this section used Whole Genome Sequencing (WGS) and RNA Sequencing (RNA-Seq) protocols based on the Illumina HiSeq X-Ten System. The sequencing

protocols, variant detection pipelines, gene expression counting methods have recently been described in [464]. All protocols are run on the DKFZ's One Touch Pipeline (OTP) data processing platform [464]. Normalization of RNA-Seq data was executed as described in Section 2.2.11.

4.3 Case Study 1: Late-Refractory Multiple Myeloma has a Diverse Immunoglobulin and Oncogene Rearrangement Landscape

4.3.1 Introduction

Multiple Myeloma (MM) is a haematological malignancy of older adults defined by the clonal proliferation of plasma cells of the bone marrow [545]. Plasma cells are a type of immune (white blood) cells originating from the bone marrow which specializes in producing large amounts of antibodies. They are differentiated from Memory B Cells, a type of immune cell which recognize, internalize and store foreign antigens to define an immune response [546]. In their de-differentiated, "plasmablastic" state, plasma cells divide rapidly, but later mature into differentiated plasma cells. Dysregulation of this process can lead to a spectrum of malignancies: MM starts with the asymptomatic phases Monoclonal Gammopathy of Undetermined Significance (MGUS) [547] and Smouldering (asymptomatic) Multiple Myeloma (SMM) [548] [549] before progressing into a full-blown MM and possibly extramedullary, soft tissue plasmacytoma [550] and Plasma Cell Leukaemia (PCL) [551] observed in late-stage patients.

In its progressed, MM form, this family of plasma cell malignancies are lethal [552], and pose a clinical challenge despite their rarity. This has fuelled a great interest in investigating the molecular mutations and mechanisms in MM development, revealing a complex and diverse set of driver mechanisms and mutations [553], [554], [555], [556]. These omics-based studies, along with preceding work, contributed significantly to the understanding of MM, where we now know the main molecular hallmarks of this disease:

- Immunoglobulin rearrangements activating the *CCND1*, *MAF*, *NSD2*(*MMSET*), *MYC* oncogenes [557]
- "Hyperdiploidy" (recurrent trisomies of chromosomes 3, 5, 7, 9, 11, 15 and 19) [558]
- Somatic rearrangements activating the *MYC* oncogene [559]
- Activating point mutations on the *RAS* oncogene family [560]
- Activating point mutations on the *BRAF* oncogene [561]
- Deactivating mutations of *TP53* [562]
- Loss of *FAM46C* as a tumour-suppressor, often concomitant with *MYC* activation as part of a rearrangement [563]
- Homozygous Losses of *RB1*, *FAF1* and *TRAF3* tumour suppressors [217] [564]

Along with a better understanding of MM, more advanced treatment strategies (including targeted treatments) evolved such as proteasome inhibition (Bortezomib [565], Carfilzomib [566]), immunomodulation (Thalidomide [567], Lenalidomide [568], Pomalidomide [569]), CD38 Antibodies [570], and autologous stem cell transplantation [571]. These improved therapies statistically significantly improved survival [572] [573]. Despite these improvements, the current clinical management and prognosis of MM remains as an incurable disease [574]. The current treatment options inevitably all lead to relapses and the condition called Relapsed/Refractory Multiple Myeloma [575]. Understanding the mechanisms of relapse in RMM will be key to further improving survival in MM and potentially making it a curable or chronic but non-lethal disease. With this motivation, we investigated a cohort of patients which were refractory to multiple treatment approaches using genomics and transcriptomics assays.

4.3.2 Study Design and Methods

The RMM sequencing project funded as the Heidelberg Institute for Personalized Oncology (HIPO) HIPO-067 Project, is coordinated by Prof. Marc-Steffen Raab, Dr. Nicola Giesen (née Lehners) and Dr. Matthias Schlesner. In this project, we collected a cohort of MM patients that were refractory to at least two regimens of immunomodulatory agents and proteasome inhibitors or CD38 antibody based immunotherapy. The cohort was selected with or without prior autologous stem cell transplantation.

A cohort of 44 patients fulfilling these conditions were selected, and sequenced for WGS. 39 Patients with sufficient DNA quality and high quality sequencing data were included in the study, of which 37 had sufficient RNA quality for RNA-Seq. Just for this dissertation, this cohort was later expanded by one *MYCN* expressing patient which did not fulfil the study conditions (died shortly after initial diagnosis during initial therapy before relapses rather than being multi-refractory).

The aims of the study were to determine recurrent (novel) drivers and mutational processes of (R/)RMM. To this end, we also obtained and processed the SMM cohort presented in [549] for comparisons underlining the differences of early and late stage MM. These comparative analyses are outside the scope of this dissertation.

4.3.3 Results

4.3.3.1 Immunoglobulin and Oncogene Rearrangement Landscape of Multiple Myeloma

We first investigated the mutational landscape of RMM using a EPISTEME Circos plot with default settings (SV recurrence, CNV recurrence with no cnLOH, Functional Small variant Recurrence). As shown in Figure 4.1, the complex mutational landscape of RMM encompasses the diverse spectrum of variants described in Section 4.3.1.

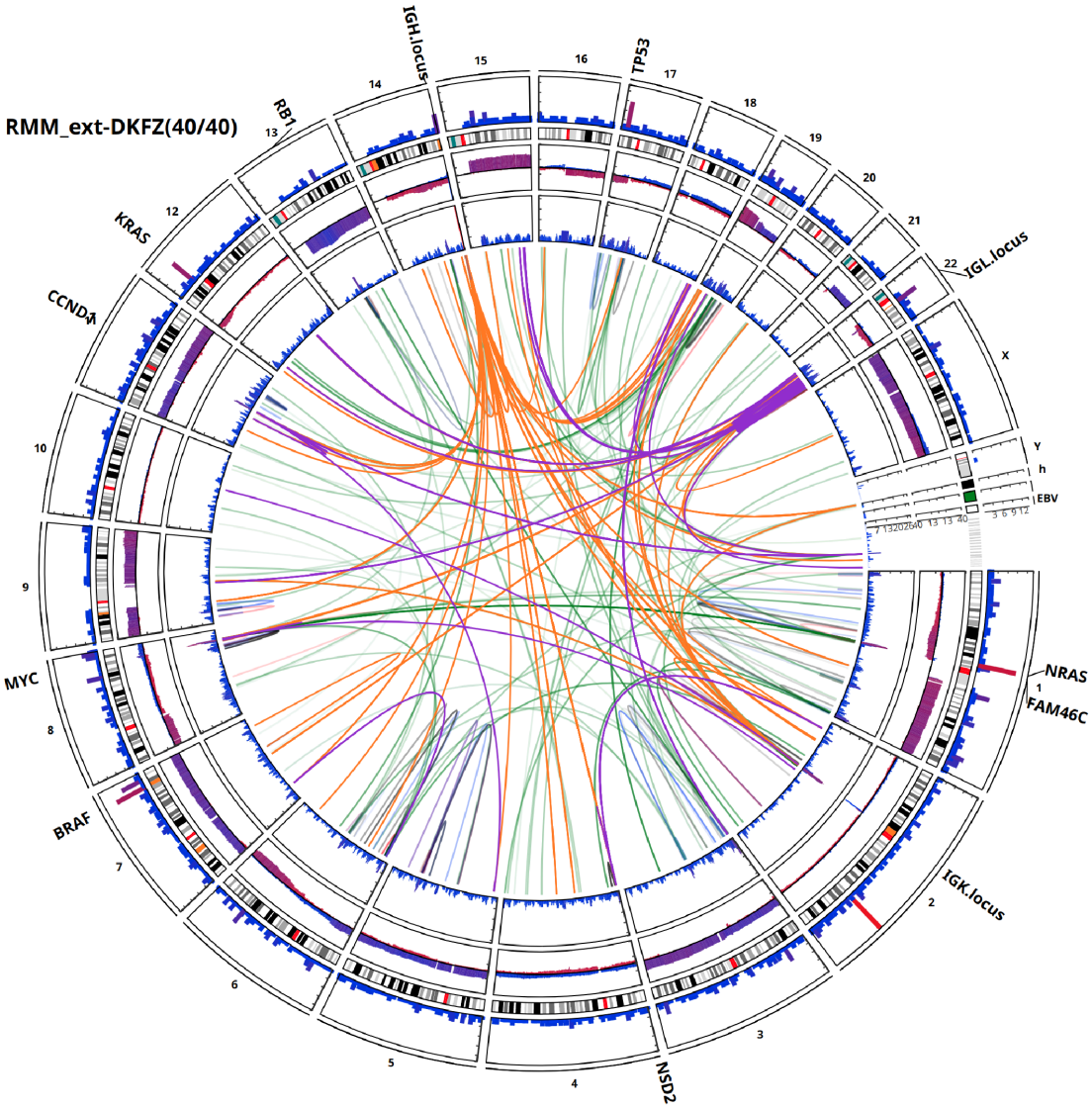

Figure 4.1: Mutational landscape of the HIPO Refractory Multiple Myeloma study. Recurrence layers from outer to inner correspond to (the default settings) gene-based functional small variant recurrence, TAD-based copy number variant recurrence, TAD-based SV recurrence (1-TAD offset).

A hallmark of many haematological malignancies if aberrant, oncogenic V(D)J recombinations, which correspond to Immunoglobulin locus translocations in B-Cell malignancies like RMM. In RMM, *CCND1*, *MAF*, *NSD2*, *MYC* are recurrent and known hallmark targets of immunoglobulin translocations. Figure 4.2 shows the immunoglobulin rearrangement landscape of RMM, where primary rearrangements are marked in orange and secondary rearrangements are marked in purple. In our RMM study, we observed recurrent immunoglobulin rearrangements targeting *CCND1*, *NSD2*, *MYC*, but not *MAF*. Notably, we observed recurrent immunoglobulin rearrangements on the *MYCN* oncogene which were 2/2 concomitant with IGH-*NSD2* translocations, of which one was definitely a two-step translocation happening after the initial *NSD2* event. Identifying the importance of this atypical finding, we added one

IGK-*MYCN* (direct, without *NSD2* involvement) case from an independent MM study which does not fulfil the conditions of the RMM study, yielding overall 3 cases where *MYCN* is targeted by immunoglobulin rearrangements. Remarkably, all IG rearrangement targets except for *CCND1*(8), *NSD2*(6), *MYC*(3) and *MYCN*(3) were sporadic, with only single patients. Nevertheless, these singletons included relevant genes for haematology and tumour development like *PAX5*(activating), *NFKB1*(activating), *TAF8*(deactivating), *EZH1*(truncating), *SLAMF1*(truncating).

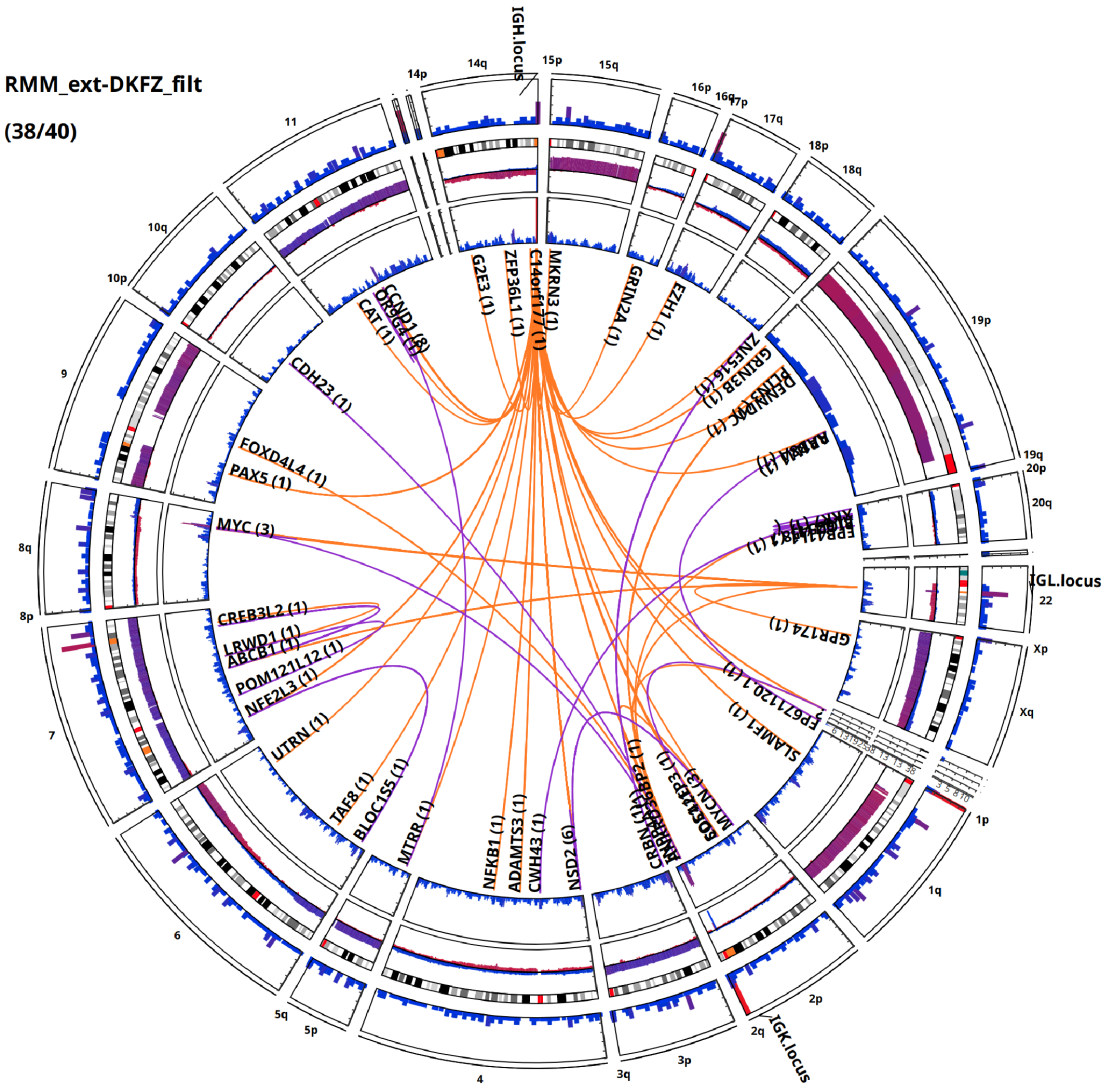

Figure 4.2: Immunoglobulin rearrangement landscape of the HIPO Refractory Multiple Myeloma study (excluding 2 cases with complex rearrangements of the IG loci). IG rearrangements are shown as orange Bézier curves for direct IG translocations and purple Bézier curves for indirect IG translocations.

Figure 4.2 is prepared with the omission of two cases which showed complex rearrangements involving the IGL and IGK loci, shown in Figure 4.3. In such cases, it becomes difficult to assign potential targets of the rearrangement, and to predict downstream effects of the IG

superenhancer's translocation.

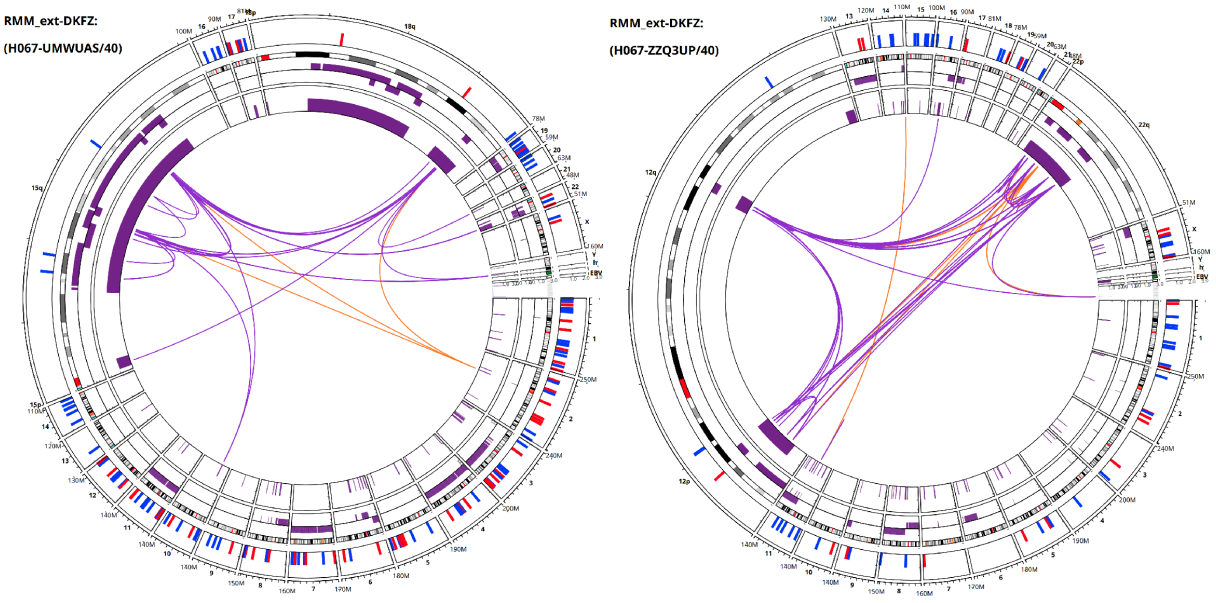

Figure 4.3: Two cases with complex rearrangements involving the IGL and IGK loci in the HIPO Refractory Multiple Myeloma study. It was not possible to determine the original, direct target of the initial catastrophic event involving the respective IG loci.

4.3.3.2 Enhancer Hijacking Events in in Multiple Myeloma

Next, we investigated the transcriptomic changes observed due to genomic variant existence, with EPISTEME's default settings (SVs are assumed to affect genes up to 1 TAD away from the initial hit, amplifications, homozygous deletions), with results in Figure 4.4. We observed strong and recurrent upregulation of the hallmark oncogenes *MYC*, *NSD2*, *FGFR3*, *CCND1* and *MYCN* (Figure 4.5, *MYCN* not shown). Cases with IGL and IGH translocations were observed to recurrently suppress *IGLL5* and *FAM30A*, respectively. We also observed a recurrent downregulation of the *TRAF3*, *RB1* and *HLA-DRA* and *HLA-DRB1* genes. Our analysis yielded no recurrent targets of enhancer hijacking apart from the listed hallmark immunoglobulin rearrangement partners. Sporadic cases of enhancer hijacking includes genes like *IGF2* and *FOXR1*. *IGF2* was shown to be a recurrent target of enhancer hijacking across multiple cancer types [100], whereas *FOXR1* was shown to be a recurrent target of enhancer hijacking in paediatric neuroblastoma [461] and B-cell lymphoma [576].

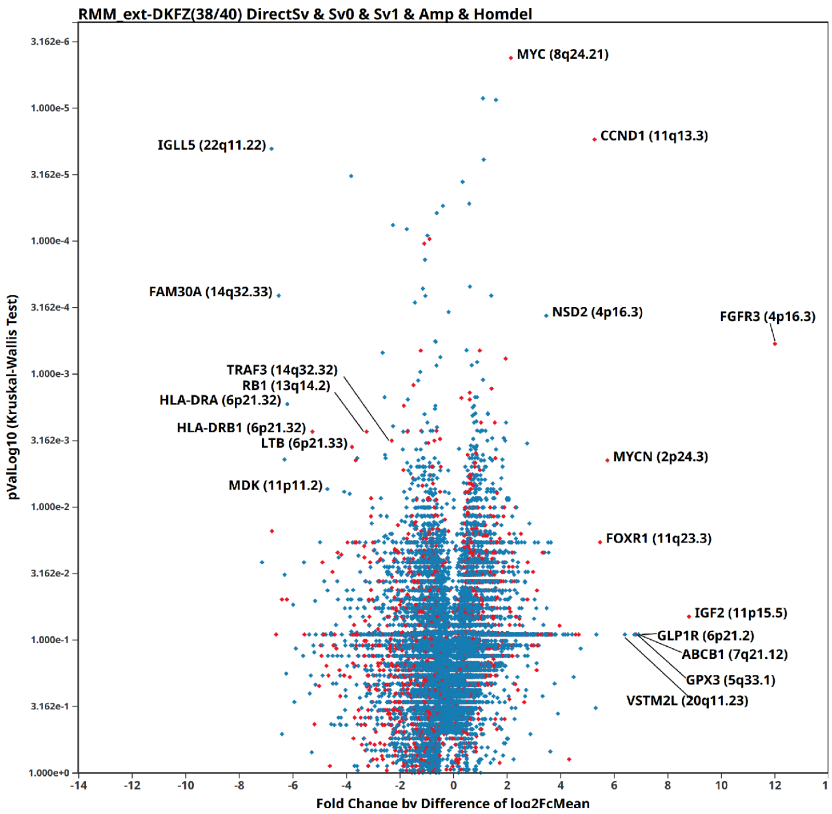

Figure 4.4: Genomic variant driven transcriptomic dysregulation landscape of the HIPO Refractory Multiple Myeloma study: Strong and recurrent upregulation of the hallmark oncogenes *MYC*, *NSD2*, *FGFR3*, *CCND1* and *MYCN*. Recurrent downregulation of the *TRAF3*, *RB1* and *HLA-DRA* and *HLA-DRB1* genes

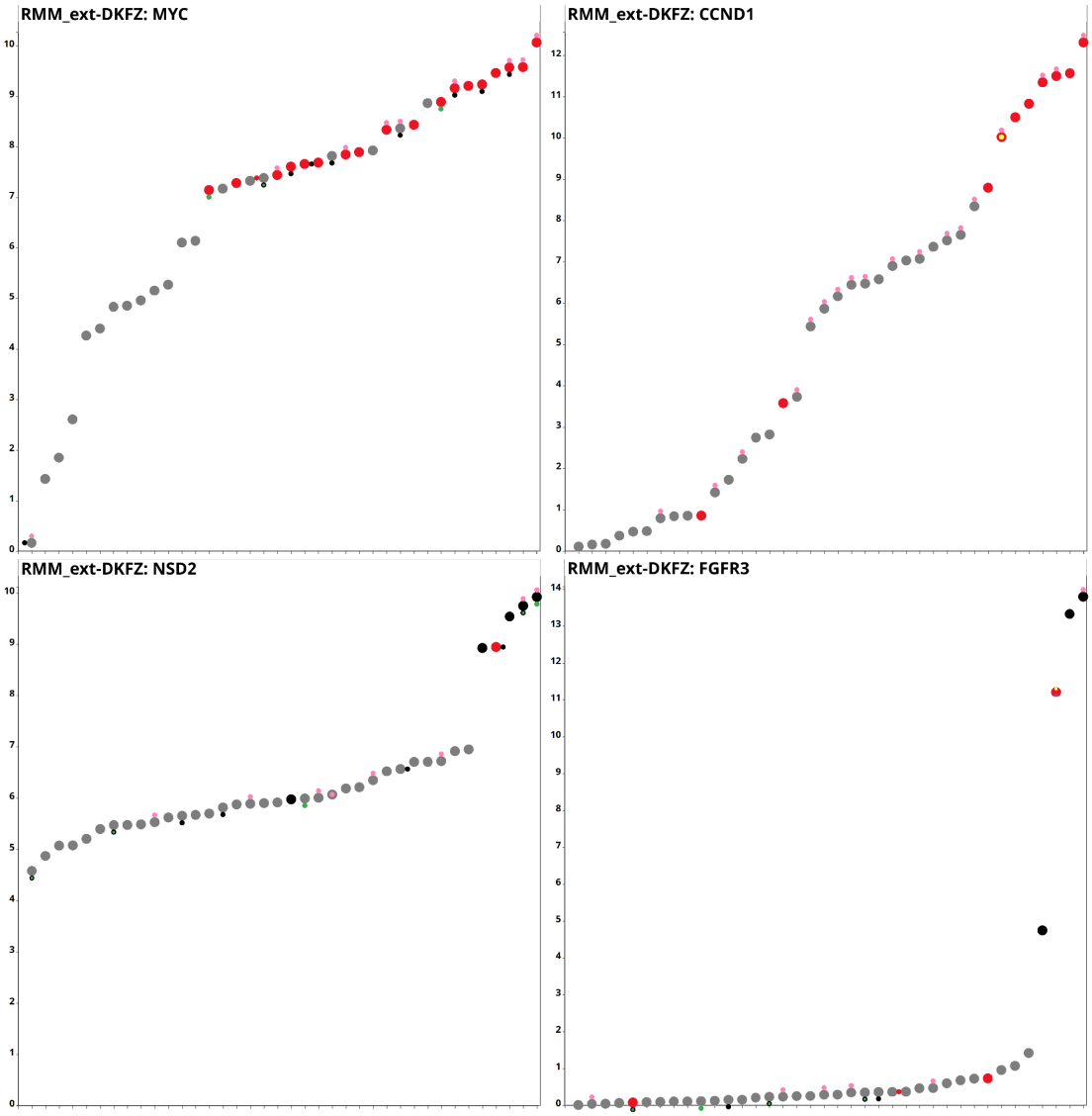

Figure 4.5: Hallmark immunoglobulin rearrangement targets *MYC*, *NSD2*, *FGFR3* and *CCND1* as targets of enhancer hijacking in the HIPO Refractory Multiple Myeloma study. SV-positive cases are displayed with red (off-gene SV hits) or black(gene-body SV hits) symbols, whereas SV-negative cases are coloured gray.

4.3.3.3 MYCN is an Oncogene in Multiple Myeloma, Potentially Defining a Subgroup with Extramedullary Manifestations and Dismal Prognosis

Having identified recurrent immunoglobulin translocations targeting the *MYCN* oncogene and having established that these rearrangements lead to the overexpression/activation of *MYCN*, we investigated the correlation between *MYC* and *MYCN* expression. Figure 4.6 shows three levels of *MYCN* expression and three levels of *MYC* expression: high-*MYCN* expressors are the three IG translocated cases, low-*MYCN* expression seems to be the normal state of this cell type, and there is a third, unexplained group of 2 intermediate-level *MYCN* expressors. As for *MYC*, there is a low-expressor group, intermediate-expressor group and a high-expressor

group which is enriched for somatic structural variants. The correlation of *MYC* and *MYCN* reveals a strong co-occurrence pattern between high *MYCN* expressors and low *MYC* expressors, strongly suggesting that these genes not only replace but also strongly suppress the other. A similar characteristic is observed in paediatric neuroblastoma but not medulloblastoma group 4 with *MYCN* amplification (Figure 4.7). This behaviour does not extend to the intermediate-level expressors, suggesting that *MYCN* is not active, hence not replacing *MYC*, in these cases.

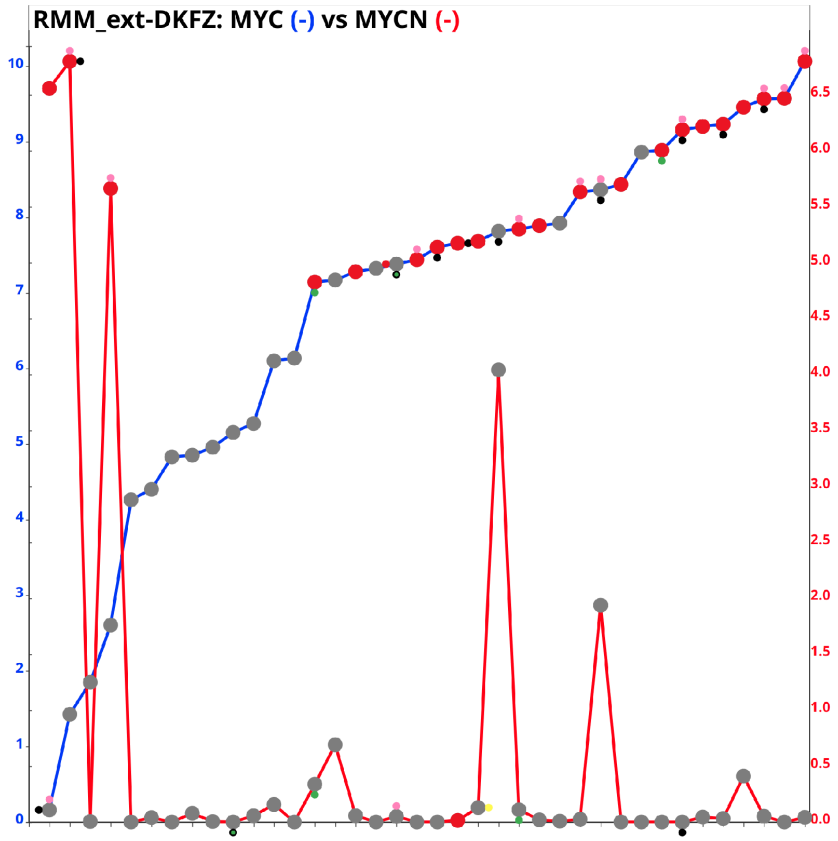

Figure 4.6: The mutually exclusive regulation of *MYC* and *MYCN* activity in the HIPO Refractory Multiple Myeloma study.

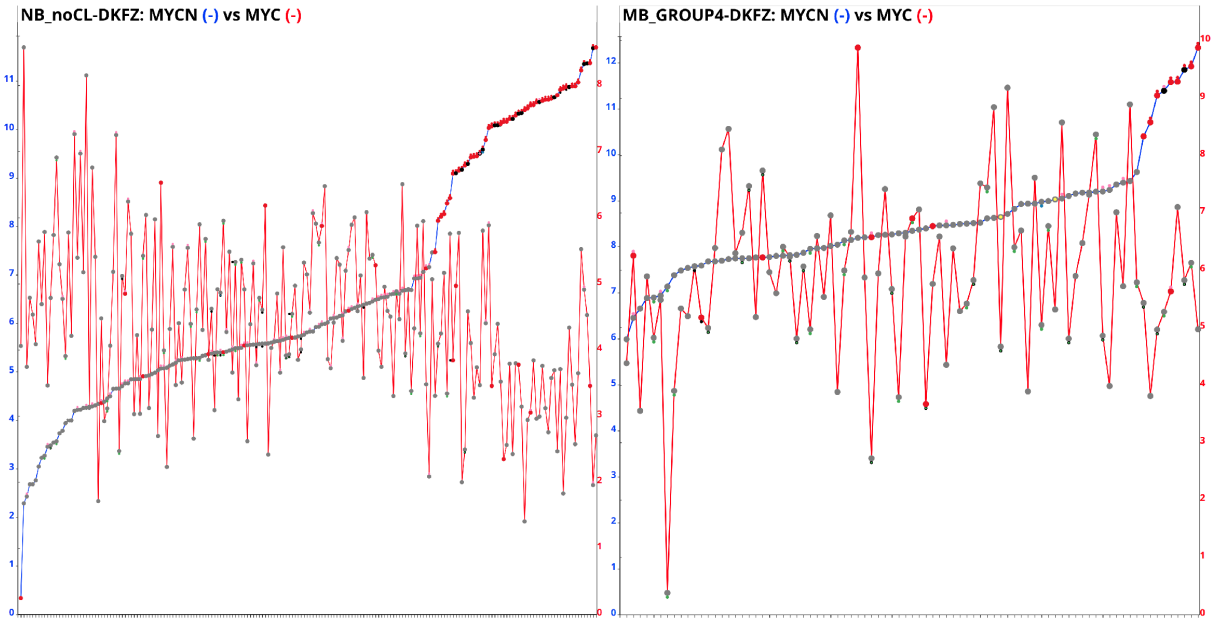

Figure 4.7: Transcriptomic correlation of the *MYC* and *MYCN* oncogenes in paediatric neuroblastoma (GPOH) and medulloblastoma group 4 (ICGC). In neuroblastoma, *MYCN* and *MYC* are in an anticorrelated relationship but not in medulloblastoma group 4.

The only case expressing low levels of *MYC* without high-level *MYCN* activation in Figure 4.6 has a *MAX* mutation. *MAX* forms a heterodimer with both the MYC and MYCN oncogenes and their mutations have previously been shown to lead to low MYC expression in MM [577].

Next, we analysed the transcriptomic effects of *MYCN* activation in RMM using both rank-based Kruskal-Wallis test and Student's T-test (Figure 4.8). The differential gene expression analysis was performed while ignoring the single *MAX* mutant case and the two intermediate *MYCN* expressors, with the motivation of purely comparing the main High*MYCN*-Low*MYC* and Low*MYCN*-High*MYC* states. The results suggest modest changes with respect to significantly and strongly differentially expressed genes.

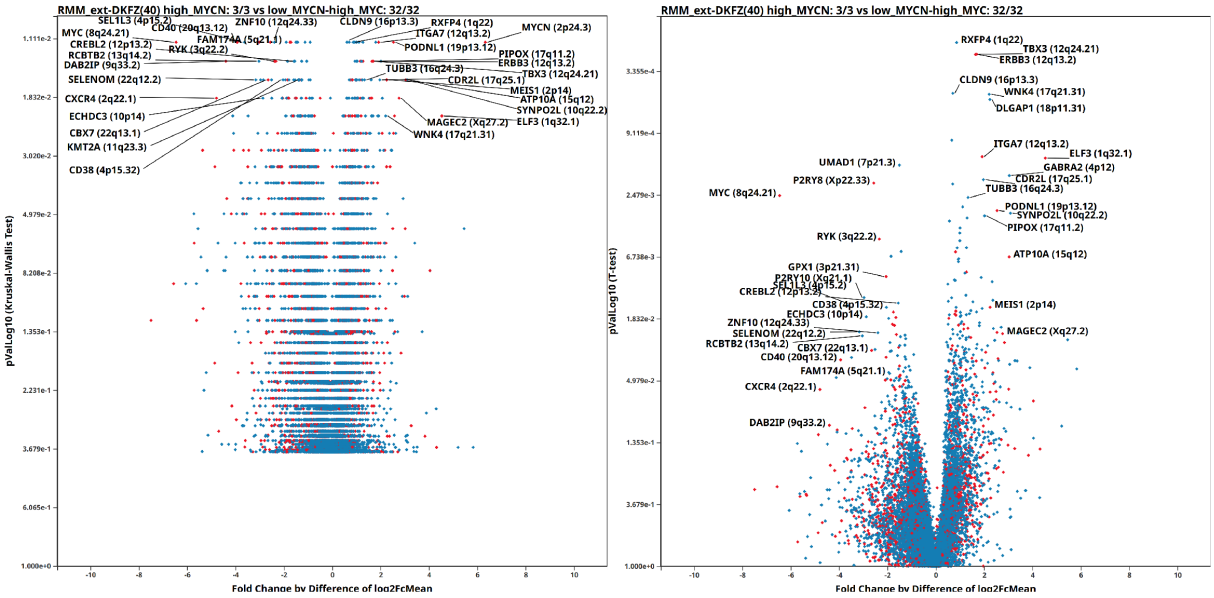

Figure 4.8: Transcriptomic effects of *MYCN* activation in the HIPO Refractory Multiple Myeloma study, with Kruskal-Wallis Test (left) and T-test(right). Comparison of the 3 *MYCN*-activated cases to the rest of the cohort excluding the 2 *MYCN*-intermediate cases and the 1 *MAX*-mutant case, thereby focusing on *MYC*-high and *MYCN*-low cases. The entry for *MYCN* gene is masked from the right panel to improve readability.

Figure 4.9 shows the correlation analysis of *MYCN* expression with selected genes showing strong differential expression characteristics.

CD40 is a B-cell associated antigen highly expressed in MM, with potential functions in cell-cell interactions [578] and cell migration [579]. Remarkably, because of its high expression levels in MM, it was recently presented as a potential immunotherapeutic target [580] via oncolytic virotherapy. If *MYCN*-activated MM is indeed suppressing *CD40* expression, this therapeutic strategy would meet a predictable resistance mechanism.

FAM174A showed a strong and consistent downregulation upon *MYCN* activation, but as a gene it remains understudied, apart from a recent proposal as a gene fusion partner in early onset colorectal cancer [581].

ITGA7 was previously shown to be a cancer stem cell marker in [582] in oesophageal squamous cell carcinoma and its higher expression in Glioma was shown to be correlated with poorer survival [583].

ATP10A is a plasma membrane protein which flips the phospholipid bilayer of plasma membranes and promotes increased endocytosis and cell bending [584], where the authors also discussed potential implications in cancer cell invasion. In our RMM cohort, its upregulation correlates with structural variants, low-order copy number gains, nonsynonymous mutations and most strongly *MYCN* expression, potentially showing a diverse range of activation mechanisms. As both high-level *MYCN* expressors in our RMM study had extramedullary plasmacytoma, it is possible to suggest that the increased migratory capacity conferred by *ATP10A* upregulation can facilitate extramedullary invasion of malignant plasma cells.

The genes presented here were chosen with consideration of the *MAX* mutant case in order not to bias the analysis towards effects of low *MYC* expression without influence from the *MYCN* gene itself.

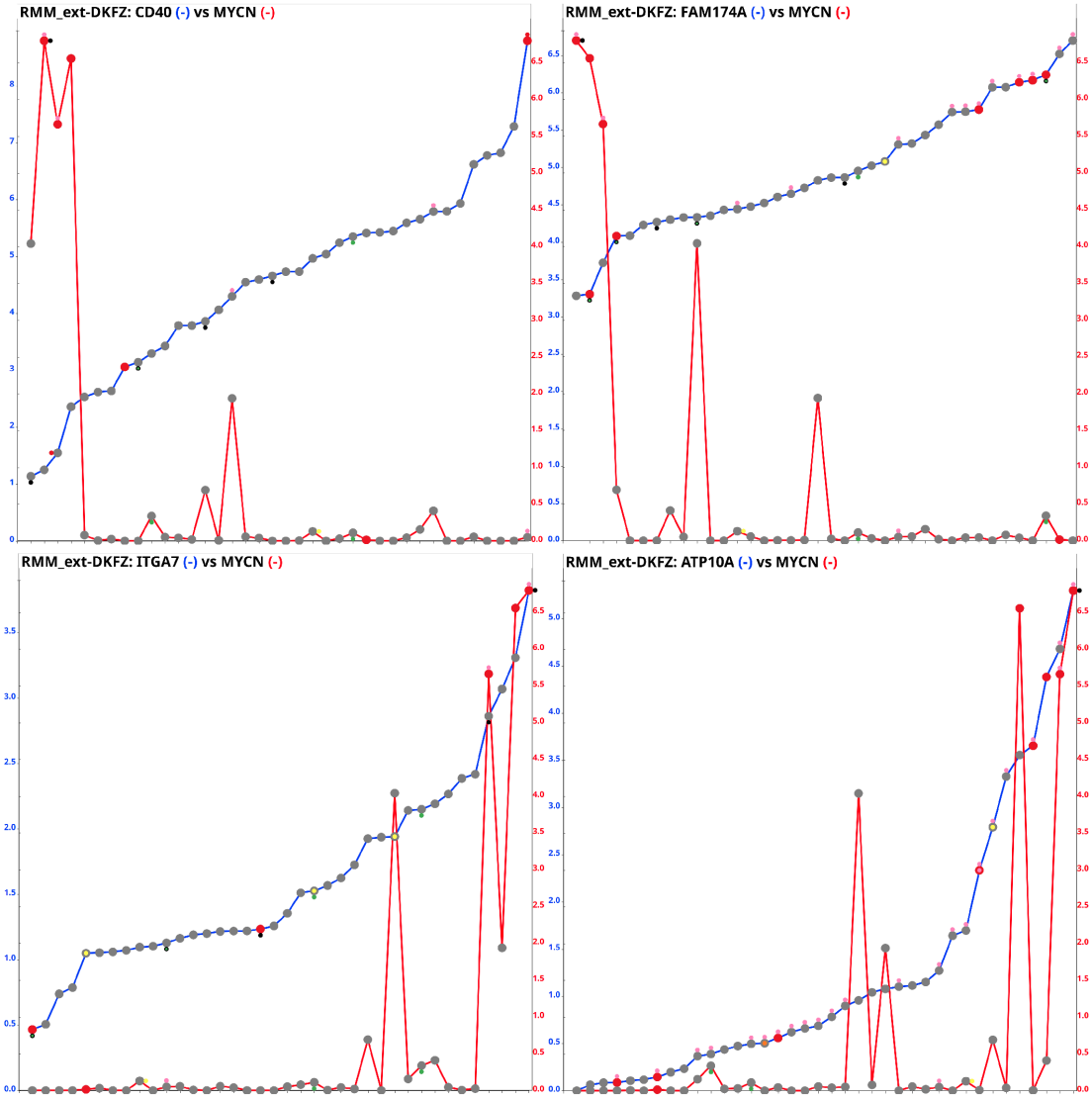

Figure 4.9: *CD40*, *FAM174A*, *ITGA7* and *ATP10A* as putative targets of *MYCN* in RMM shown on two-phenotype correlation plots with the *MYCN* gene as the anchor.

Next, we investigated if *MYCN* activation can be observed in a larger cohort of early-stage MM, and if it has an influence on survival in the context of early-stage disease. To this end, we used the R2 instance of the dataset desribed in [585], where authors did a RNA microarray analysis on a cohort of 542 cases. The results on Figure 4.10 both confirmed the existence of a small subset of *MYCN* overexpressing cases in this larger study (8/542 vs 2/39 in our RMM study), and that *MYCN* expression correlates with significantly lower survival, without any finer classification of the non-*MYCN* expressor cases.

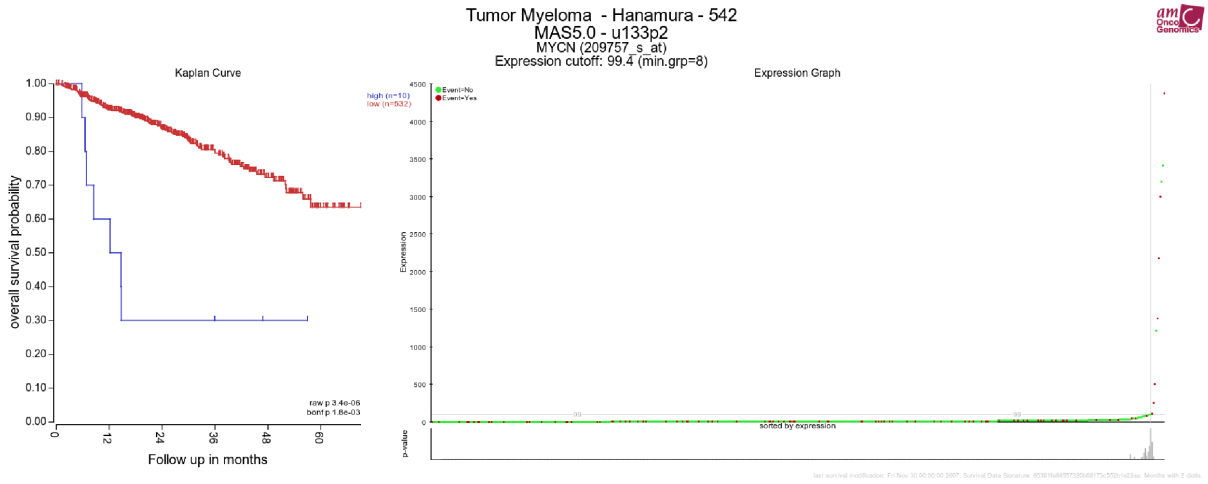

Figure 4.10: *MYCN* activation leads to significantly poorer survival in an independent and larger Multiple Myeloma project [585].

Finally, we investigated a potential explanation of mechanism for the 2 intermediate *MYCN* expressor cases. We visualized the 2-patient subcohort as a Circos plot followed by an application of an EPISTEME mutex analysis comparing these two cases to the rest of the cohort (38 cases). Both cases showed small mutations on *HOXA4* (1 frameshift deletion, 1 nonsynonymous mutation), a gene which was not mutated in any other patient in this study. There is, however, as of 07.2019 no published connection between the *HOXA4* and *MYCN* genes suggesting co-regulation or binding.

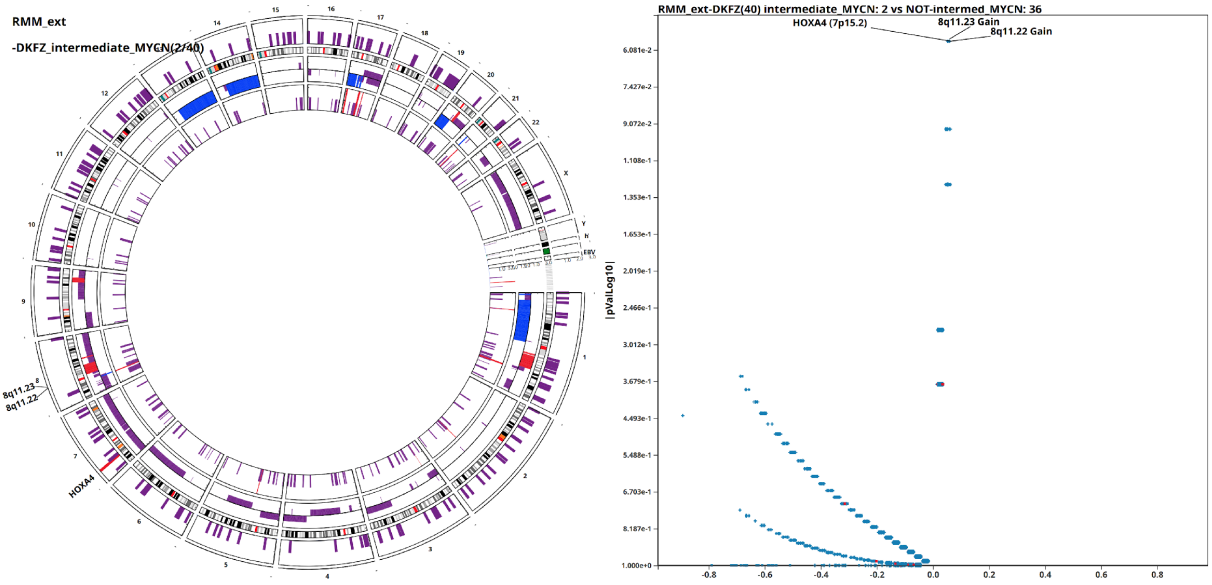

Figure 4.11: Intermediate-level *MYCN* expressing and not *MYC*-suppressing cases exclusively have *HOXA4* mutations in the HIPO Refractory Multiple Myeloma study.

4.3.4 Discussion

We investigated using EPISTEME, genomic and transcriptomic sequencing data from a medium-sized cohort of highly-selected and highly-refractory MM patients, who underwent disease relapse under at least 2 treatment strategies. We showed a diverse immunoglobulin translocation landscape with translocation partners of relevant function for MM. We showed a lack of rare, but recurrent enhancer hijacking events, instead emphasizing the established hallmark oncogenes and tumour suppressors as the only recurrently dysregulated genes with somatic genomic variants as the dysregulation mechanism.

We presented *MYCN* as a rare target of IG rearrangements and consequent enhancer hijacking. *MYCN* has only on one instance been discussed in the literature as a potential oncogene in MM: In [586], the authors discovered rare rearrangements of IG loci targeting the *MYCN* locus, mentioning a cell line named PE-2 as a prototypical example of such rearrangements. As of 07.2019, PE-2 is not commercially available and no other studies have been published using this cell line. The authors also used the dataset described in [585] to underline the rarity of *MYCN* expression in MM, but did not proceed to do further analyses on its putative targets.

More samples are needed to dissect the true influence of *MYCN* and to what extent the transcriptomic dysregulation is due to *MYC* downregulation and to *MYCN* activation. The analysis presented here suffered from low sample sizes (2 in the publication cohort, 3 in this dissertation), and a comparison with a larger *MAX*-mutant control cohort with low *MYC* expression and low *MYCN* expression would have been very valuable. Another issue in our setting is the existence of IGH-*NSD2*/*MMSET* rearrangements in 2/3 of the *MYCN* expressors, which adds another confounder to any analyses investigating *MYCN*'s true effects in MM. Our study coordinator colleagues created a cell line from one of the *MYCN* expressor patients, generation of other such cell lines would undoubtedly be an excellent first step in developing a mechanistic understanding of *MYCN* function in MM, in the long-run allowing the rational design of treatment strategies [587].

4.4 Case Study 2: The MNX1 Oncogene is Activated by Recurrent CDK6-NOM1 Rearrangements in chr7q-Monosomy Acute Myeloid Leukemia

Acute Myeloid Leukaemia (AML) is a haematological malignancy of myeloid cells. Mature myeloid cells encompass granulocytes and monocytes, which are differentiated from a common myeloid precursor cell type [588]. AML arises from these myeloid precursor (or stem) cells, where they are blocked from maturation and clonally expand and proliferate due to diverse somatic mutations.

The established World Health Organization (WHO) classification of Adult AML [589]

1. AML with certain genetic abnormalities,
2. AML with myelodysplasia-related changes,
3. AML related to previous chemotherapy or radiation and
4. AML not otherwise specified (NOS).

Of these, only the NOS "subtype" is defined by histopathological/morphological features. AML currently has the following recognized somatic alterations in the classification of AML:

1. AML with a translocation between chromosomes 8 and 21 t(8;21) involving *RUNX1-RUNX1T1* fusions [590],
2. AML with t(16;16) or inv(16) involving *CBFB-MYH11* fusions [196],
3. Acute Promyelocytic Leukaemia (APL) involving *PML-RARA* fusions [195],
4. AML with t(9;11) involving *KMT2A-MLLT3* fusions [591],
5. AML with t(6:9) involving *DEK-NUP214* (formerly *DEK-CAN*) fusions [592],
6. AML with a t(3;3) or inv(3) involving *RPN1-MECOM* (formerly *RPN1-EVI1*) fusions [593],
7. Megakaryoblastic AML (AMKL) t(1:22) involving *RBM15-MKL1* fusions ,
8. (Provisional) AML with *BCR-ABL1* (formerly *BCR-ABL*) fusions [594],
9. (Provisional) AML with *NPM1* mutations [595],
10. (Provisional) AML with biallelic *CEBPA* mutations [596],
11. (Provisional) AML with mutated *RUNX1* gene [597].

Apart from these recognized, molecular subtype defining somatic genomic alterations, AML has a diverse landscape of genomic variants as investigated using genome sequencing in [598] and [483]. Cytogenetic and mutational classification as well as mutational status of significant genes are now accepted to be a rational predictor of treatment responses [599], where it was suggested to incorporate the mutational status of *NPM1*, *FLT3*, *CEBPA*, *TP53*, *SRSF2*, *ASXL1*, *DNMT3A* and *IDH2* into prognostic guidelines.

Another aspect of AML biology is chromosomal imbalances, and karyotype complexity. Chromosome arm level or focal losses involving known tumour suppressors is a central theme in tumour biology, where haematological malignancies are no exception [600]. Common chromosome arm losses such as 5q, 7q have been associated with secondary Myelodysplastic Syndrome or AML caused by prior treatment of other conditions such as lymphoma [601], [602], [603]. Among these, chromosome 7q losses are observed both in de novo and treatment induced AML [604] and have been of particular interest due to their commonness [605],[604],[606] and a lack of a strong and validated tumour suppressor candidate, with multiple having been proposed such as *CUX1* and *EZH2* [607], [608], [609]. Recently, successful responses to demetyhlating agents were reported in AML with chr7q loss or chr7 total loss [610], prompting interest in the roles of key epigenetic genes such as *EZH2* and *METTL2B* on commonly deleted segments in AML with chr7q-monosomies.

According to the WHO classification of AML, most subtypes of AML are defined by balanced structural rearrangements leading to fusion genes. The WHO classification recognizes only the most commonly encountered fusion genes, while there are many more rarer balanced

translocations and gene fusions observed in larger cohorts [611]. A notable subset of such rare translocations are those encountered in paediatric AML, which are rare in childhood AML and very rare to non-existent in adult AML and are not recognized as bona fide subtypes of AML. One of these rare translocations enriched in paediatric AML is the t(7;12)(q36.3;p13.2) *ETV6-MNX1* fusion gene. This translocation is observed in a rare subtype of infant AML, with implications of poor prognosis [612]. Motor Neuron Homeobox 1(*MNX1*) is a homeobox gene of the Antennapedia (ANTP) class and Hox-Like (HOXL) family, while originally renamed *HLXB9*/*HB9*, it was later renamed to *MNX1* after being classified as the sole member of a gene family [613]. Its primary function is as a key player in motor neuron differentiation [614], [615] [616]. In the haematological setting, the roles of *MNX1* overexpression in t(7;22) infant AML have been investigated in multiple studies, putting forward mechanisms on altered cell-cell interactions [617] and blockade of haematopoietic differentiation [618].

In this study, we addressed the open question of how interstitial chromosome 7q losses contribute to the development of AML. Starting with the hypothesis of an elusive tumour suppressor gene residing on chr7q, we designed a project to use the power of WGS on a larger cohort of chr7q-monosomy patients to study SVs in addition to the coding regions analysed by the more commonly established WES. Our investigation of this question did not reveal a recurrently mutated tumour suppressor in a double-hit process, but rather yielded the *MNX1* oncogene known from paediatric AML as the putative driver of a subset of AML cases with chromosome 7q losses via an enhancer hijacking process. With a novel mechanism of enhancer hijacking via *CDK6-NOM1* rearrangements, the neighbour gene to *MNX1*, without the direct involvement of the *MNX1* gene body, *MNX1* is activated by the enhancer of the constitutively active *CDK6* gene. We show that *MNX1* expression is tightly regulated and repressed in AML and only activated by recurrent structural variants targeting chr7q36.3, yielding to significant and global changes in gene expression patterns, in line with a differentiation block hypothesis, but with a novel list of key components of haematopoietic regulation being dysregulated.

4.4.1 Study Design and Methods

The chr7q-monosomy AML sequencing project funded as the HIPO-030 Project, is coordinated by Prof. Christoph Plass, Dr. Daniel Lipka and Prof. Konstanze Döhner (University of Ulm). 19 adult AML cases were collected with matching normal blood samples and sequenced with WGS and RNA-Seq with the following characteristics: 5 normal karyotype controls, 2 isodicentric 7p cases, 2 case with a balanced translocation on chr7q leading to an aberrant karyotype and 10 cases with partial chr7q losses. Cases with centromere-to-telomere full losses of the chromosome 7q arm were not collected in this project. RNA Sequencing was run on 25 cases, including 6 cases without sufficient DNA quality for WGS.

Following the identification of the *MNX1* oncogene as a putative key driver in a subset of partial 7q-monosomy cases additional AML cases expressing *MNX1* were collected: 2 paediatric t(7;12)(q36.3;p13.2) *ETV6-MNX1* cases, the GDM1 adult AML cell line with a *MYB-MNX1* rearrangement, 1 *MNX1* expressing AML case with claimed normal karyotype, 2 cases with *MNX1* expression and cytogenetically detected chr7q losses. Unfortunately none of the cases in the extension cohort have matching normal blood available. Of these cases GDM1 was

not submitted for RNA-Sequencing due to concerns of lack of comparability with patient data, 1 case is currently being RNA-sequenced, the remaining 4 did not have sufficient RNA quality for RNA sequencing.

4.4.2 Results

4.4.2.1 Commonly Deleted Segments on the chr7q do not Reveal a Clear Tumour Suppressor Gene Candidate for AML

We first investigated the somatic genomic variant landscape of the described AML cohort using EPISTEME's cohort-wide Circos plot feature (Figure 4.12), which showed recurrently deleted segments on chr7q, but no recurrent functional small mutations on any genes residing on the commonly deleted segments. Expanding the analysis to further mutation types such as direct gene hits revealed the *CDK6*, *NOM1* and *CNTNAP2* genes as the only candidates for direct involvement along with the likely artefacts *MUC17* and *TECPR1*. The literature and our experience with the *CNTNAP2* gene from different diseases suggests that it is a very large gene prone to passenger mutations and fragile site loss based SVs [619]. Because *CDK6* and *NOM1* form deletion boundaries, they do not qualify as "second-hit" candidates for a tumour suppressor gene.

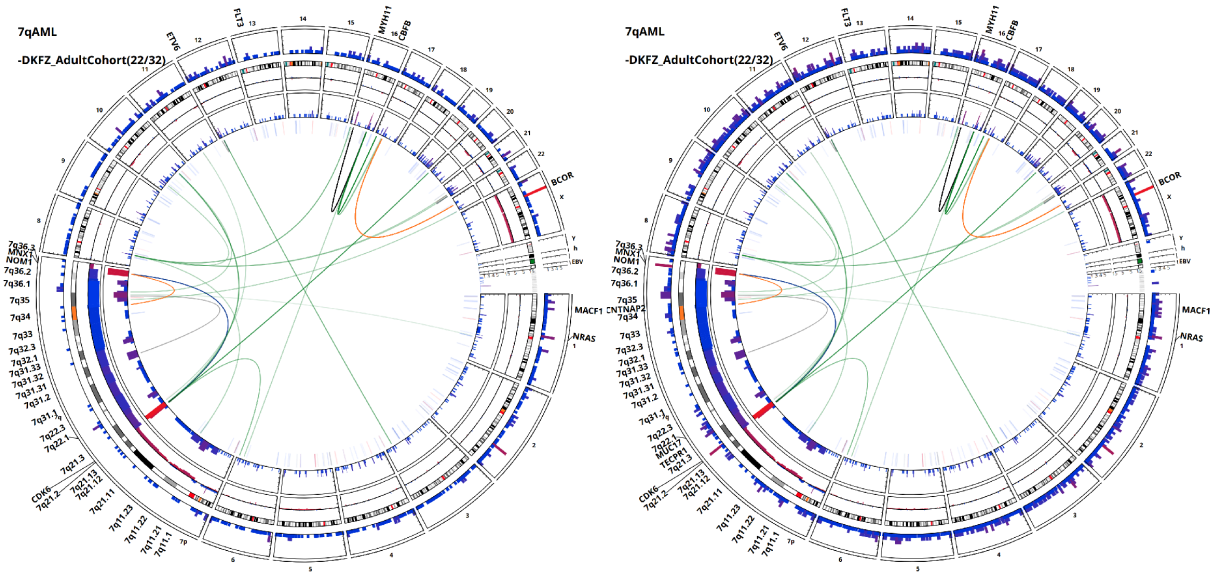

Figure 4.12: Mutational landscape of the HIPO 7q-AML study with 8x enlargement of chr7q. Left: only functional small variants on the outermost variant recurrence layer, Right: functional small variants, UTR variants, upstream/downstream gene variants and direct SV hits on genes. For both plots the middle recurrence layer denotes TAD-based CNV recurrences and the inner recurrence layer denotes TAD-based SV recurrences.

Based on these observations, we concluded that there is likely no common tumour suppressor akin to *TP53* or *VHL* that is being deactivated in a biallelic manner on the q arm of chromosome 7 in AML, while not ruling out a role for haploinsufficiency for any genes on the

commonly lost chr7q segments.

4.4.2.2 *MNX1 activation by recurrent CDK6-NOM1 rearrangements Defines a Subset of chr7q-Monosomy AML*

We then investigated in EPISTEME the genomic variant driven transcriptomic dysregulation landscape of our AML cohort with two strategies, including and excluding the effects of lower-order copy number gains and losses. Apart from known targets of dysregulation such as *MYH11*, *CBFB*, *BMI1*, *SPAG2* (*MLLT10* neighbour), this analysis revealed *MNX1* as a recurrent target of strong upregulation, which was upon closer investigation revealed to be a total activation, suggesting total repression in normal conditions with activity entirely dependent on distal effects of SVs (Figure 4.14). The *MNX1* activating SVs were in all cases large deletions fusing different introns of the *CDK6* and *NOM1* genes in noninverted orientation. As the *CDK6* and *NOM1* genes reside on opposite strands, this gene fusion does not yield a viable chimeric fusion, as expected.

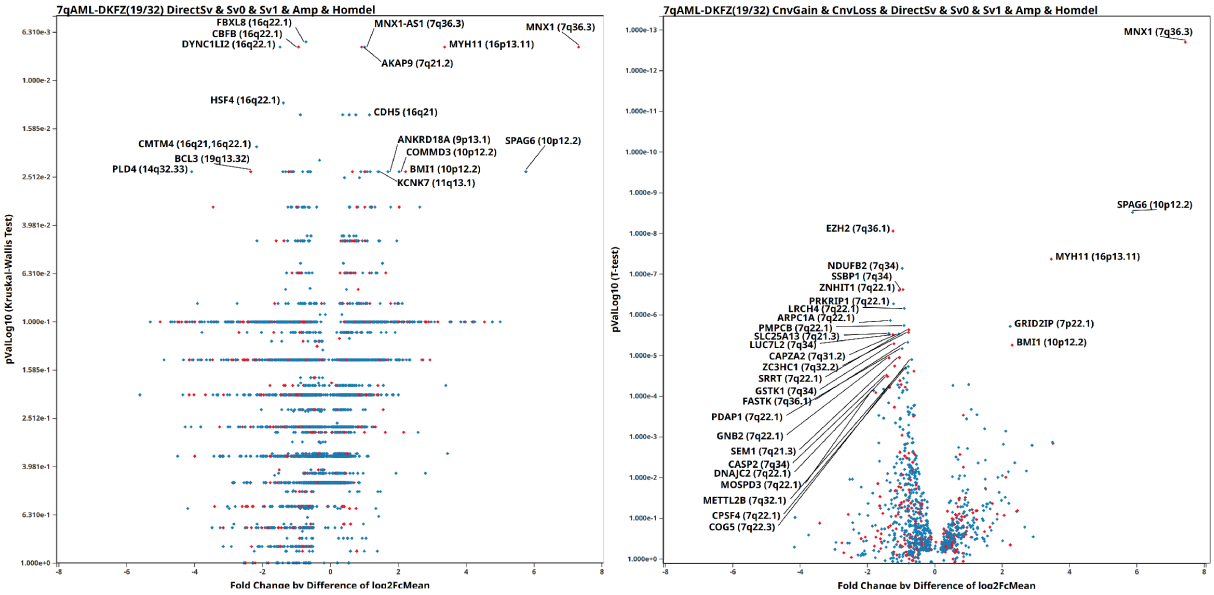

Figure 4.13: Genomic variant driven transcriptomic dysregulation landscape of the HIPO 7q-AML study. Left: analysis only with direct or nearby SV effects, amplifications and homozygous deletions with the rank-based Kruskal-Wallis test, right: analysis including lower-order copy number changes with the T-test providing better separation of the genes on chr7q.

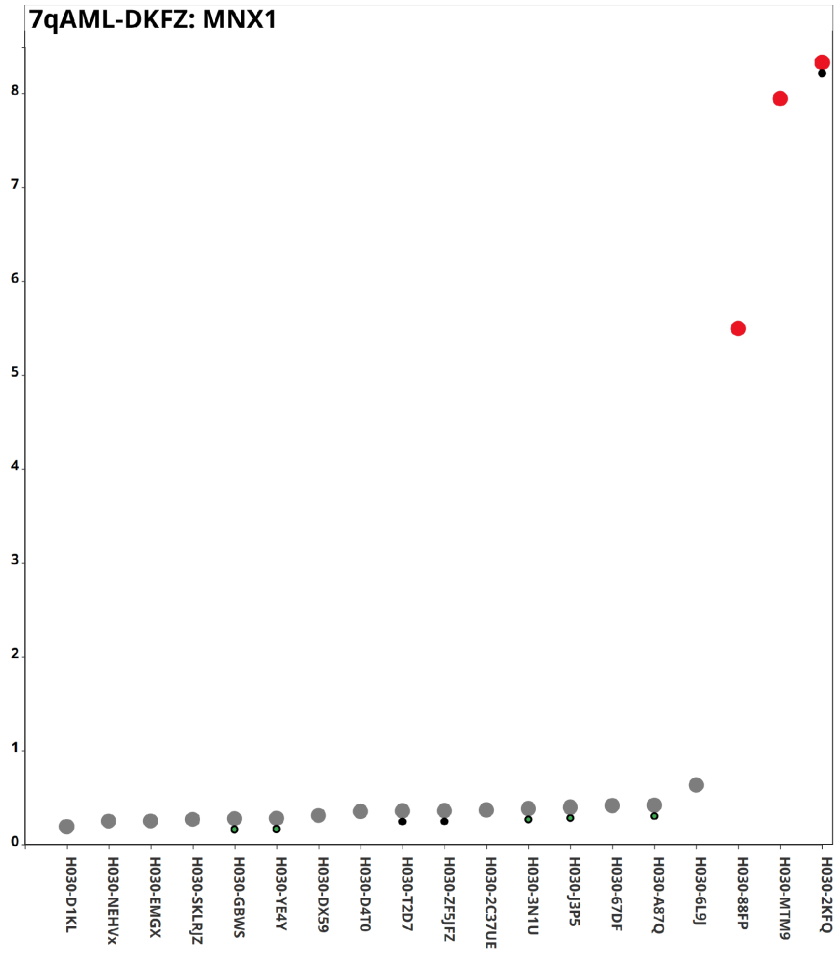

Figure 4.14: *MNX1* Activation in the HIPO 7q-AML study. The normally repressed expression of *MNX1* is recurrently activated in three cases with off-gene SV hits, denoted by red circles.

Further analysis of lower-order copy number alterations revealed multiple genes with strong separation of the chr7q-monosomy and chr7q-wild-type cohorts with similar patterns with expression levels significantly falling but not leading to total suppression. Among these outliers, *EZH2*, *FASTK*, *ZNHIT1* and *METTL2B* are shown in Figure 4.15 due to their functions in tumour development and epigenetic regulation.

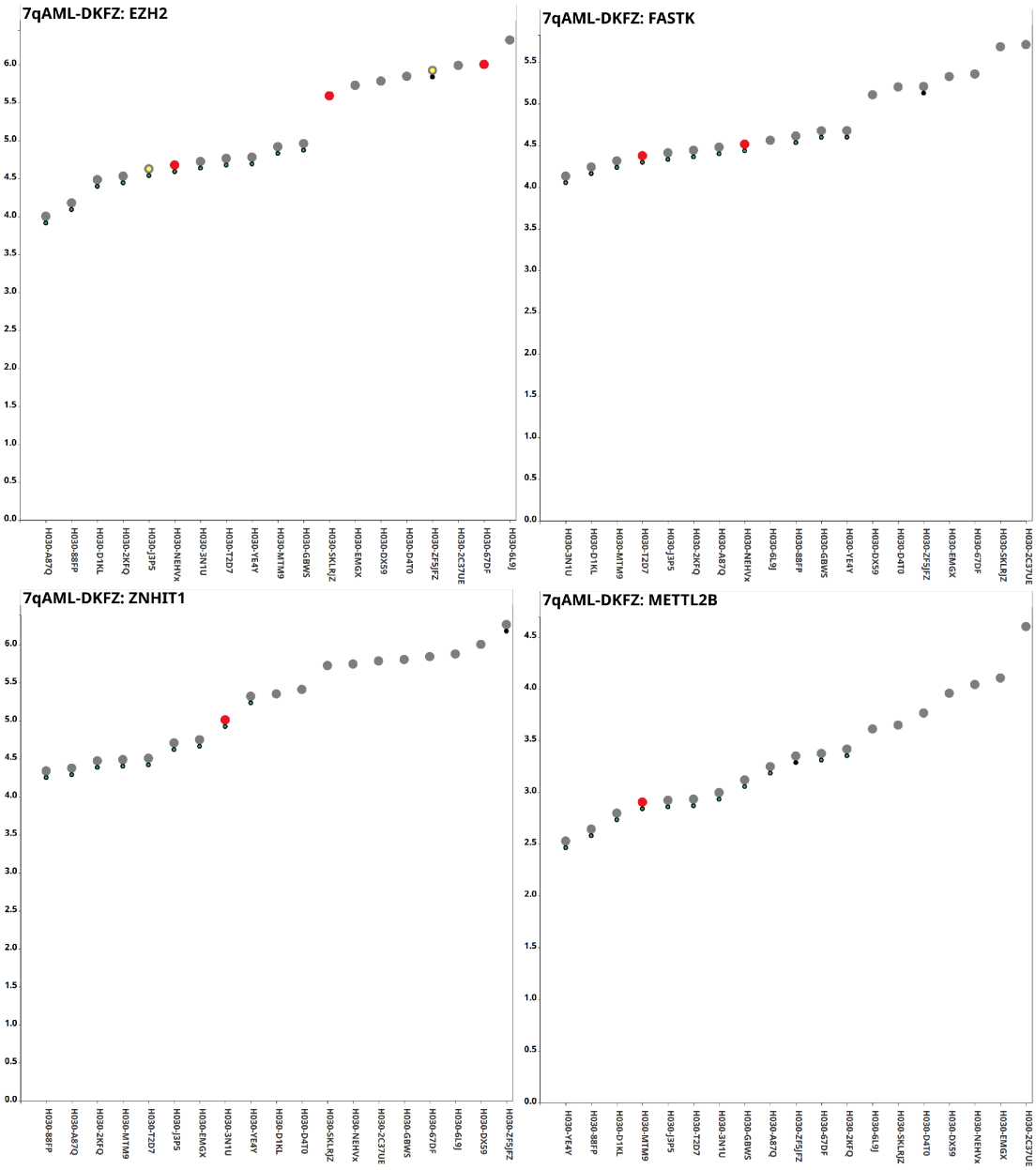

Figure 4.15: Lower-order copy number losses (green lower circles) leading to downregulation of *EZH2*, *FASTK*, *ZNHIT1* and *METTL2B* in the HIPO 7q-AML study

Next, we used the described extension cohort, expanding the cohort of *MNX1* expressing cases from 3 to 9. Figure 4.16 shows the diverse activation mechanisms of the *MNX1* oncogene in paediatric and adult AML: we hitherto collected 5 cases using the CDK6 enhancer of which 4 cases have the *NOM1* gene as deletion partner, 2 cases using the *ETV6* enhancer (paediatric t(7;12) AML), 1 case using the *MYB* enhancer (GDM1 adult AML cell line), and 1 particularly interesting case using the T-cell receptor beta locus as an enhancer. This last case was initially annotated as a ”normal karyotype AML” according to clinical cytogenetic analysis because the loss of the segments between chr7q34 and chr7q36.3 were too subtle for classical cytogenetics. Furthermore, the use of the T-cell receptor beta locus was unexpected in itself due to its known

rarity in AML [620].

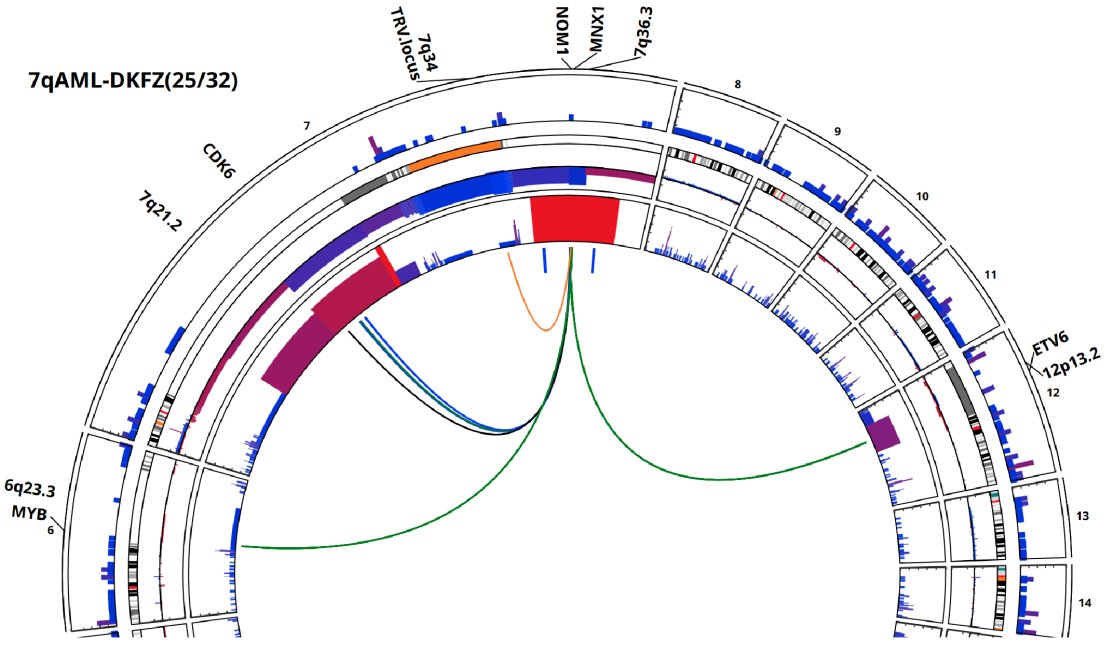

Figure 4.16: *MNX1* activation by diverse enhancer partners and SV mechanisms in the HIPO 7q-AML study. We identified 4 distinct enhancers that are hijacked by *MNX1*, in order of frequency: *CDK6* (5 deletions) *ETV6* (2 translocations, paediatric), *MYB* (1 translocation), T-cell receptor beta (1 deletion)

We then investigated if *MNX1* expressing tumours require or prefer secondary mutations, focusing on cancer-related genes and especially those described in AML. The analysis revealed a particularly diverse set of secondary mutations including previously annotated drivers of AML.

- 2KFQ: *TET2*, *ASH1L*
- 88FP: *DDX41* [621], *HDAC4*, *RARB*
- MTM9: (high mutational load) *BCOR* [622], *DNMT3A* [623], *IDH1* [624], *EIF4E*, *NACA*
- 6AQ4F8: (high mutational load) *BCOR*, *ETV6*, *RUNX1* [625], *FGF6*, *DISC1*
- T1242: (high mutational load, nocontrol case) *BCOR*, *DNMT3A*, *ABL2*, *BCR*, *CUX1*, *HECW1*, *PHF6*, *ULK1*, *TET2*
- 16054: (high mutational load, nocontrol case) *ATF2*, *NRAS*, *IRF6*, *BCOR*, *PTCH2*, *RUNX1*, *KRAS*, *PLK1*, *MLKL*, *MSH6*, *NACA*, *DISC1*
- paediatric-T1: (high mutational load, nocontrol case) *ALK* [626], *AFF1*, *PRDM2*, *MN1*, *HLTF*, *FOXR1*, *WRN*, *ULK1*

- paediatric-T2: (high mutational load, nocontrol case) *BRCA2*, *RELA*, *BCL9L*, *POT1*, *ZHX2*, *APC*, *PIK3R4*, *MN1*, *IRF3*, *KLF2*, *KDM6B*, *SP1*, *ABRAXAS1*, *PIK3R4*, *ULK1*
- GDM1 cell line: *ERG*, *NFKB1*, *PRDM2*

Overall, the recurrent secondary hits excluding the GDM1 cell line were *BCOR*(4/8), *ULK1*(3/8), *DNMT3A*(2/8), *RUNX1*(2/8), *TET2*(2/8), *NACA*(2/8) and *MN1*(2/8, paedatric cases). As no secondary mutations are shared between the adult and paediatric cases, considering only adult cases the enrichment of *BCOR*(4/6), *DNMT3A*(2/6), *RUNX1*(2/6), *TET2*(2/6) and *NACA*(2/6) become stronger. One weakness of this analysis is the lack of availability of matched normals for 5/9 of the *MNX1* expressor cases. In particular, the strongest candidate gene *BCOR* is only somatically mutated in 2/4 of the cases with matched normals, but remains as a strong candidate for secondary mutations due to its well described importance in AML.

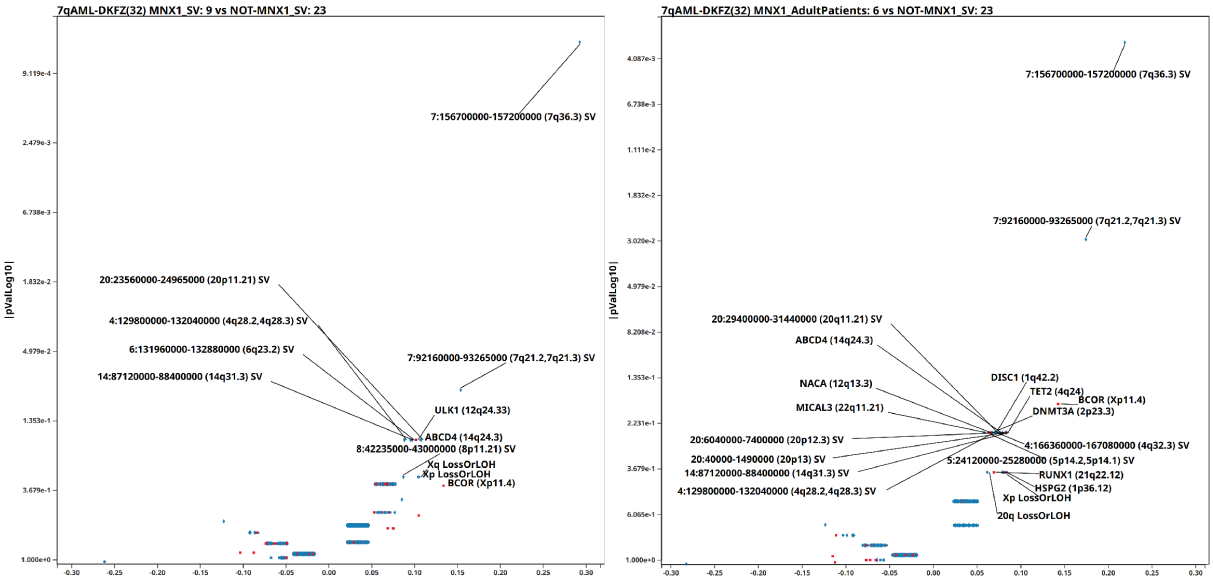

Figure 4.17: Co-mutations enriched in *MNX1*-Activated AML. Left: all *MNX1*-Activated case, Right: Adult *MNX1*-activated patients.

Having established the activation of *MNX1* in a subset of AML cases with chr7q-monosomy, we investigated if *MNX1* activation takes place in a distinct transcriptomic or methylome subtype of AML. Figure 4.18 shows that *MNX1* cases cluster similarly but do not form a distinct group in both analyses.

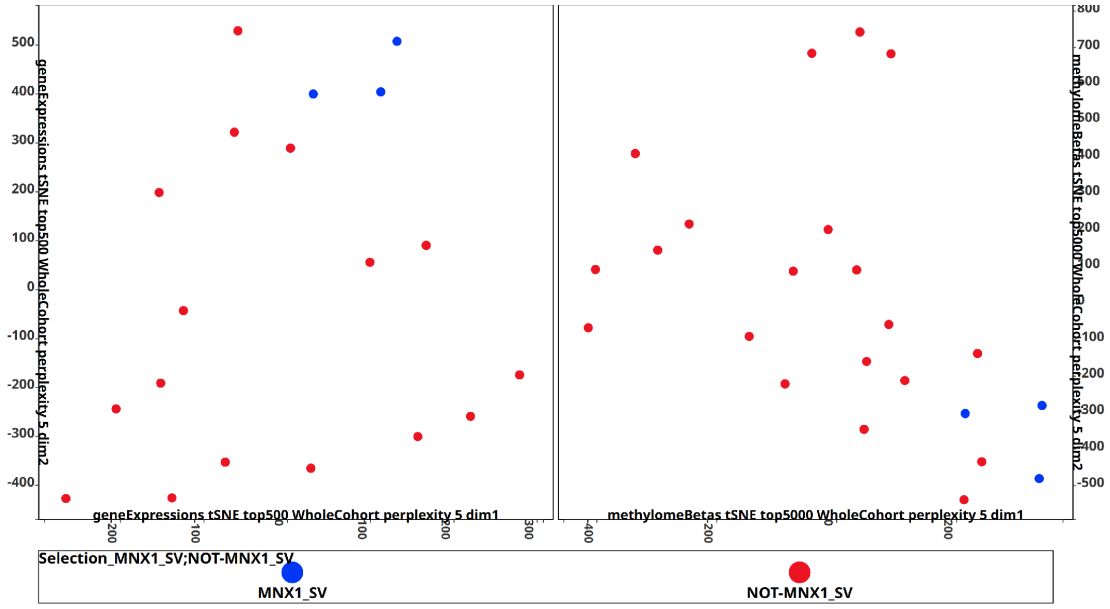

Figure 4.18: *MNX1* Activated cases (blue) co-cluster but do not form a distinct transcriptome (left, tSNE top 500 most variable genes, perplexity 5, late-exaggeration 1.1) or methylome (right, tSNE top 5000 most variable methylation probes, perplexity 5, late-exaggeration 1.1) subtype in the HIPO 7q-AML study

4.4.2.3 MNX1 Activation Leads to Comprehensive Transcriptomic Changes in chr7q-Monosomy AML

Recognizing the role of *MNX1* as a key developmental transcription factor, we investigated its effects on the transcriptome of adult AML upon activation. Using EPISTEME's differential expression (Figure 4.19, using the T-test) and correlation analysis (Figure 4.20) features, we identified massive transcriptomic changes, where large numbers of genes were both strongly dysregulated and strongly correlated with *MNX1* expression.

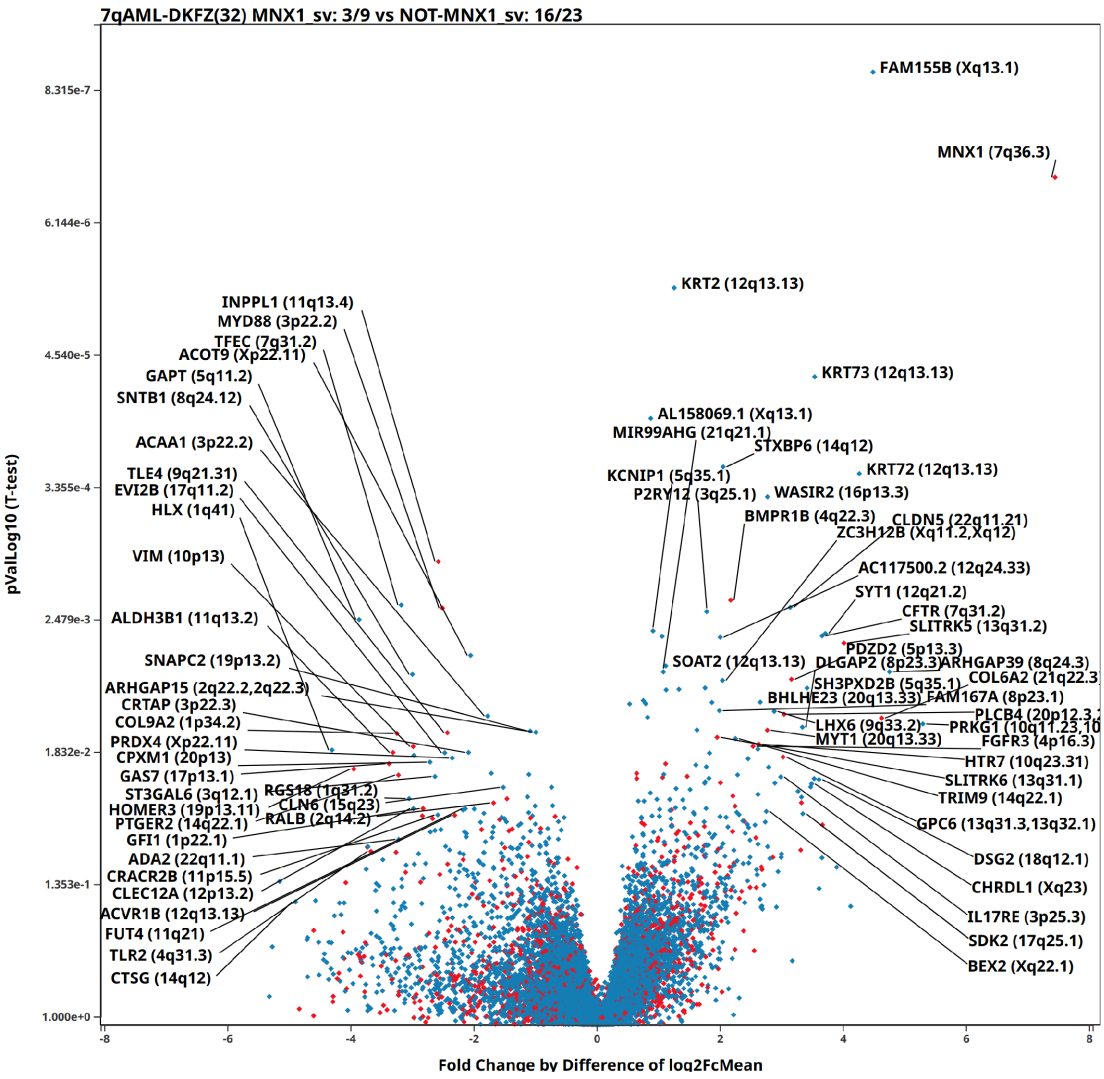

Figure 4.19: *MNX1* Activation Leads to comprehensive transcriptomic changes in the HIPO 7q-AML study

In the double Pearson-vs-Spearman correlation analysis, we ignored the Spearman correlation results (Figure 4.20) because of the misleading nature of rank-based statistics in the presence of large numbers of ties, which was the case in hand due to the near-total suppression of AML in most patients in our cohort.

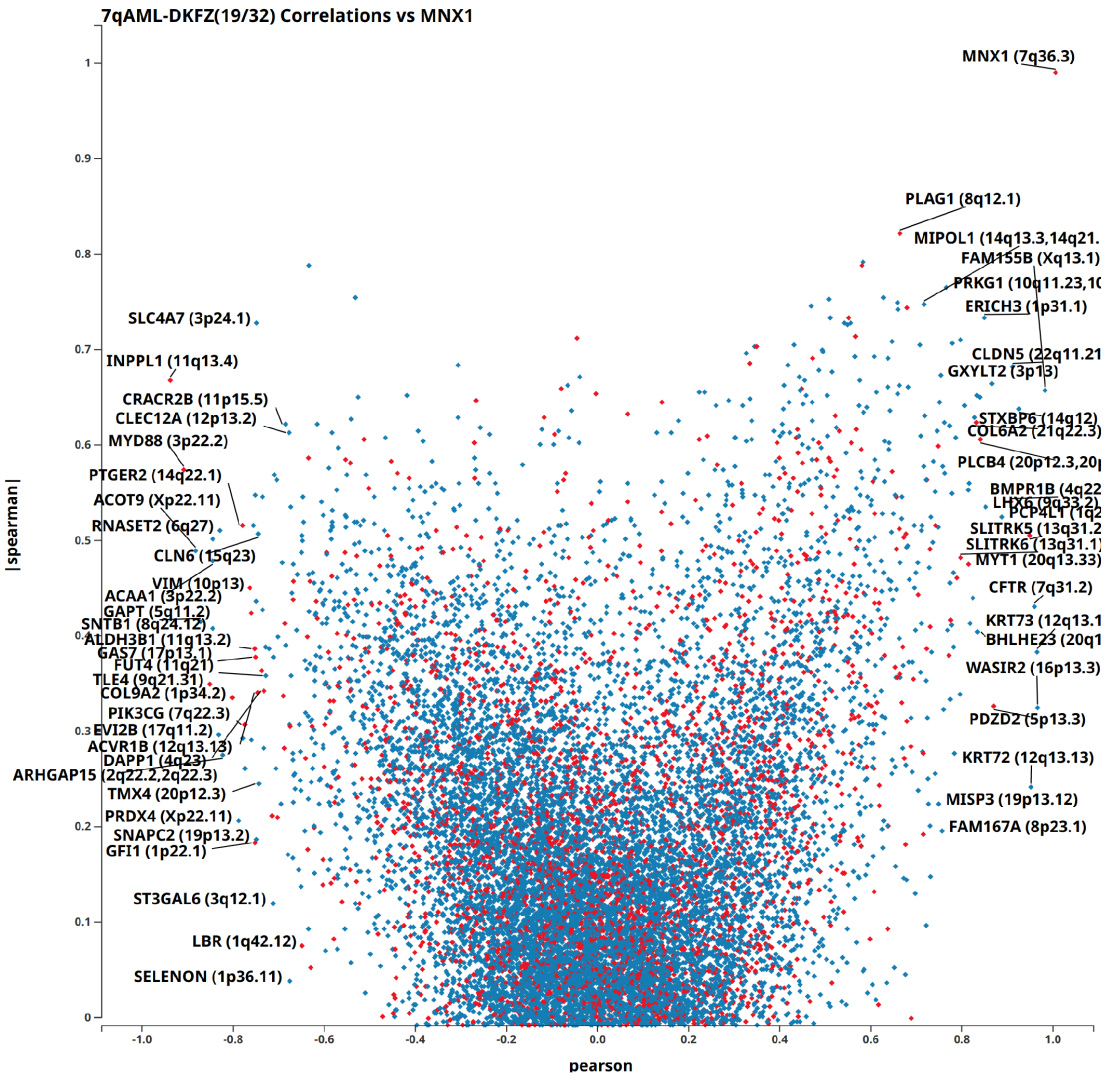

Figure 4.20: Correlation analysis for *MNX1* expression in the HIPO 7q-AML study

Next, we dissected the differential expression characteristics following *MNX1* activation, focusing on the strongest expression changes due to the low available sample sizes (3-vs-16). We established the following 40 genes as upregulated: *MNX1*, *FAM155B*, *KRT73*, *KRT72*, *WASIR2*, *STXBP6*, *KRT2*, *AL158069.1*, *BMPR1B*, *CLDN5*, *SYT1*, *CFTR*, *SLITRK5*, *ARHGAP39*, *PRKG1*, *COL6A2*, *SH3PXD2B*, *PDZD2*, *P2RY12*, *AC117500.2*, *DLGAP2*, *PLCB4*, *LHX6*, *BHLHE23*, *DSG2*, *CHRDL1*, *KCNIP1*, *FAM167A*, *MYT1*, *SOAT2*, *ZC3H12B*, *GPC6*, *BEX2*, *SDK2*, *IL17RE*, *HTR7*, *FGFR3*, *TRIM9*, *SLITRK6*, *MIR99AHG* And the following 34 genes as downregulated: *INPPL1*, *TFEC*, *GAPT*, *SNTB1*, *ACOT9*, *MYD88*, *HLX*, *COL9A2*, *TLE4*, *HOMER3*, *GAS7*, *ALDH3B1*, *VIM*, *PTGER2*, *CPXM1*, *ACAA1*, *CRTAP*, *PRDX4*, *ST3GAL6*, *RGS18*, *GFI1*, *CLEC12A*, *CTSG*, *PIK3CG*, *FUT4*, *ACVR1B*, *CRACR2B*, *ADA2*, *TLR2*, *RALB*, *CLN6*, *SNAPC2*, *ARHGAP15*, *EVI2B*.

Using ConsensusPathDB [498] we investigated if the upregulated and downregulated gene sets are enriched in known pathways and gene ontologies.

Upregulated genes largely followed the expected characteristics from a gene defined as a

motor neuron homeobox development gene (Figures 4.21 and 4.22).

pathway name	set size	candidates contained	p-value	q-value	pathway source
Protein-protein interactions at synapses	88	4 (4.5%)	3.06e-05	0.00196	Reactome
Keratinization	129	4 (3.1%)	0.000136	0.00436	Reactome
Receptor-type tyrosine-protein phosphatases	20	2 (10.0%)	0.000772	0.0136	Reactome
BMP Signalling Pathway	21	2 (9.5%)	0.000852	0.0136	HumanCyc
Signaling by BMP	26	2 (7.7%)	0.00131	0.0139	Reactome
Developmental Biology	620	6 (1.0%)	0.00143	0.0139	Reactome
Selective Serotonin Reuptake Inhibitor Pathway, Pharmacodynamics	28	2 (7.1%)	0.00152	0.0139	PharmGKB
Platelet activation - Homo sapiens (human)	123	3 (2.4%)	0.00205	0.0164	KEGG
BMP receptor signaling	41	2 (4.9%)	0.00324	0.0228	PID
Syndecan-1-mediated signaling events	43	2 (4.7%)	0.00356	0.0228	PID
Neurexins and neuroligins	57	2 (3.5%)	0.00619	0.0273	Reactome
IL-7 signaling	185	3 (1.6%)	0.00636	0.0273	INOH
EPO signaling	186	3 (1.6%)	0.00646	0.0273	INOH
Neuronal System	368	4 (1.1%)	0.00657	0.0273	Reactome
VEGF	188	3 (1.6%)	0.00666	0.0273	INOH
Long-term depression - Homo sapiens (human)	60	2 (3.3%)	0.00683	0.0273	KEGG
Cell-cell junction organization	64	2 (3.2%)	0.00751	0.0283	Reactome

Figure 4.21: Pathways enriched for involvement of genes upregulated by *MNX1* activation in the HIPO 7q-AML study.

gene ontology term	category, level	set size	candidates contained	p-value	q-value
GO:0030154 cell differentiation	BP 3	4089	22 (0.5%)	2.32e-07	2.48e-05
GO:0048869 cellular developmental process	BP 2	4275	22 (0.5%)	5.28e-07	3.06e-05
GO:0048513 animal organ development	BP 3	3452	19 (0.6%)	2.23e-06	0.000119
GO:0048731 system development	BP 3	4713	20 (0.4%)	6.01e-05	0.00168
GO:0048468 cell development	BP 3	2110	13 (0.6%)	6.29e-05	0.00168
GO:0048856 anatomical structure development	BP 2	5790	22 (0.4%)	0.000105	0.00305
GO:0007275 multicellular organism development	BP 2	5289	20 (0.4%)	0.000337	0.00651
GO:0009887 animal organ morphogenesis	BP 3	981	8 (0.8%)	0.000367	0.00786
GO:0035988 chondrocyte proliferation	BP 2	17	2 (11.8%)	0.000454	0.00658
GO:0009888 tissue development	BP 3	1954	11 (0.6%)	0.000613	0.0109
GO:0005886 plasma membrane	CC 2	5614	20 (0.4%)	0.000788	0.0215
GO:0031424 keratinization	BP 2	227	4 (1.8%)	0.000797	0.00925
GO:0048589 developmental growth	BP 2	631	6 (1.0%)	0.00102	0.00984
GO:0071944 cell periphery	CC 2	5729	20 (0.4%)	0.00105	0.0215
GO:0099106 ion channel regulator activity	MF 3	123	3 (2.4%)	0.00156	0.0405
GO:0044087 regulation of cellular component biogenesis	BP 3	959	7 (0.7%)	0.00172	0.0251
GO:0099536 synaptic signaling	BP 3	711	6 (0.8%)	0.00187	0.0251
GO:0007267 cell-cell signaling	BP 2	1594	9 (0.6%)	0.00215	0.0178
GO:0060322 head development	BP 3	772	6 (0.8%)	0.00276	0.0328
GO:0016247 channel regulator activity	MF 2	151	3 (2.0%)	0.00279	0.0489
GO:0098868 bone growth	BP 3	47	2 (4.3%)	0.00348	0.0367
GO:0007155 cell adhesion	BP 2	1389	8 (0.6%)	0.00349	0.0253
GO:0060076 excitatory synapse	CC 2	48	2 (4.2%)	0.00363	0.0368
GO:0098609 cell-cell adhesion	BP 3	819	6 (0.7%)	0.00377	0.0367
GO:0098632 cell-cell adhesion mediator activity	MF 3	50	2 (4.0%)	0.00393	0.0511
GO:0098978 glutamatergic synapse	CC 2	351	4 (1.1%)	0.00403	0.0368
GO:0044459 plasma membrane part	CC 2	2876	12 (0.4%)	0.00448	0.0368
GO:0001503 ossification	BP 2	373	4 (1.1%)	0.00495	0.0316
GO:0030900 forebrain development	BP 3	381	4 (1.1%)	0.00529	0.0472
GO:0098631 cell adhesion mediator activity	MF 2	59	2 (3.4%)	0.00544	0.0489
GO:0009653 anatomical structure morphogenesis	BP 2	2560	11 (0.4%)	0.00546	0.0316
GO:0050808 synapse organization	BP 3	391	4 (1.0%)	0.00585	0.0481
GO:0050432 catecholamine secretion	BP 3	66	2 (3.1%)	0.00656	0.0502
GO:0001533 cornified envelope	CC 3	65	2 (3.1%)	0.00656	0.189
GO:0005911 cell-cell junction	CC 2	446	4 (0.9%)	0.00931	0.0636

Figure 4.22: Gene Ontology (GO) terms enriched for involvement of genes upregulated by *MNX1* activation in the HIPO 7q-AML study

Downregulated genes, rather showed characteristics suggesting roles in haematological development (Figures 4.23 and 4.24).

pathway name	set size	candidates contained	p-value	q-value	pathway source
Neutrophil degranulation	490	7 (1.4%)	9.32e-05	0.0069	Reactome
Simplified Depiction of MYD88 Distinct Input-Output Pathway	17	2 (11.8%)	0.000686	0.0131	Wikipathways
LTF danger signal response pathway	19	2 (10.5%)	0.00086	0.0131	Wikipathways
ER-Phagosome pathway	24	2 (8.3%)	0.00138	0.0131	Reactome
IL1 and megakaryocytes in obesity	24	2 (8.3%)	0.00138	0.0131	Wikipathways
Relationship between inflammation, COX-2 and EGFR	25	2 (8.0%)	0.0015	0.0131	Wikipathways
Endogenous TLR signaling	26	2 (7.7%)	0.00162	0.0131	PID
Toll-like Receptor Signaling Pathway	102	3 (2.9%)	0.00163	0.0131	Wikipathways
Glycosphingolipid biosynthesis - lacto and neolacto series - Homo sapiens (human)	27	2 (7.4%)	0.00175	0.0131	KEGG
nfkb activation by nontypeable hemophilus influenzae	29	2 (6.9%)	0.00201	0.0131	BioCarta
Toxoplasmosis - Homo sapiens (human)	113	3 (2.7%)	0.00219	0.0131	KEGG
Innate Immune System	1077	8 (0.8%)	0.00227	0.0131	Reactome
Toll-like Receptor Signaling	31	2 (6.5%)	0.0023	0.0131	Wikipathways
toll-like receptor pathway	35	2 (5.7%)	0.00292	0.0155	BioCarta
Inositol phosphate metabolism	41	2 (4.9%)	0.004	0.0197	INOH
Regulation of toll-like receptor signaling pathway	143	3 (2.1%)	0.00426	0.0197	Wikipathways
Class I PI3K signaling events	45	2 (4.4%)	0.0048	0.0209	PID
Malaria - Homo sapiens (human)	49	2 (4.1%)	0.00567	0.0221	KEGG
Antigen processing-Cross presentation	49	2 (4.1%)	0.00567	0.0221	Reactome
Synthesis of PIPs at the plasma membrane	53	2 (3.8%)	0.0066	0.0244	Reactome
Legionellosis - Homo sapiens (human)	55	2 (3.6%)	0.00709	0.025	KEGG
TYROBP Causal Network	60	2 (3.3%)	0.00839	0.0282	Wikipathways

Figure 4.23: Pathways enriched for involvement of genes downregulated by *MNX1* activation in the HIPO 7q-AML study

gene ontology term	category, level	set size	candidates contained	p-value	q-value
GO:0002263 cell activation involved in immune response	BP 3	709	10 (1.4%)	2.46e-07	3.1e-05
GO:0043299 leukocyte degranulation	BP 3	539	8 (1.5%)	3.32e-06	0.000209
GO:0002443 leukocyte mediated immunity	BP 3	901	9 (1.0%)	1.85e-05	0.000619
GO:0006887 exocytosis	BP 3	909	9 (1.0%)	1.96e-05	0.000619
GO:0045321 leukocyte activation	BP 2	1284	10 (0.8%)	4.92e-05	0.00131
GO:0002252 immune effector process	BP 2	1287	10 (0.8%)	5.02e-05	0.00131
GO:0006955 immune response	BP 2	2260	13 (0.6%)	6.34e-05	0.00131
GO:0044281 small molecule metabolic process	BP 2	2011	12 (0.6%)	9.67e-05	0.0015
GO:0070887 cellular response to chemical stimulus	BP 3	3148	15 (0.5%)	0.000123	0.0031
GO:0001775 cell activation	BP 2	1435	10 (0.7%)	0.000126	0.00156
GO:0071216 cellular response to biotic stimulus	BP 3	227	4 (1.8%)	0.000661	0.0139
GO:0032940 secretion by cell	BP 2	1493	9 (0.6%)	0.000858	0.00887
GO:0009617 response to bacterium	BP 3	712	6 (0.8%)	0.00134	0.0241
GO:1901615 organic hydroxy compound metabolic process	BP 3	517	5 (1.0%)	0.00198	0.0312
GO:0031983 vesicle lumen	CC 3	340	4 (1.2%)	0.00291	0.0552
GO:0006066 alcohol metabolic process	BP 3	351	4 (1.1%)	0.00323	0.0448
GO:0048519 negative regulation of biological process	BP 3	5219	17 (0.3%)	0.00356	0.0448
GO:0044433 cytoplasmic vesicle part	CC 3	1498	8 (0.5%)	0.00381	0.0552
GO:0065008 regulation of biological quality	BP 2	3912	14 (0.4%)	0.00442	0.0391
GO:0010033 response to organic substance	BP 3	3173	12 (0.4%)	0.00598	0.0685
GO:0044282 small molecule catabolic process	BP 3	437	4 (0.9%)	0.00702	0.0694
GO:0032637 interleukin-8 production	BP 3	72	2 (2.8%)	0.00716	0.0694
GO:0042445 hormone metabolic process	BP 2	227	3 (1.3%)	0.00739	0.051
GO:0005539 glycosaminoglycan binding	MF 3	230	3 (1.3%)	0.00766	0.0788
GO:0019904 protein domain specific binding	MF 3	710	5 (0.7%)	0.0077	0.0788
GO:0051707 response to other organism	BP 2	1020	6 (0.6%)	0.00796	0.051
GO:0043207 response to external biotic stimulus	BP 3	1022	6 (0.6%)	0.00803	0.0719
GO:0042221 response to chemical	BP 2	4656	15 (0.3%)	0.00823	0.051
GO:0042802 identical protein binding	MF 3	1733	8 (0.5%)	0.0091	0.0788
GO:0009607 response to biotic stimulus	BP 2	1053	6 (0.6%)	0.00925	0.0521
GO:1901698 response to nitrogen compound	BP 3	1051	6 (0.6%)	0.00925	0.0719

Figure 4.24: Gene Ontology (GO) terms enriched for involvement of genes downregulated by *MNX1* activation in the HIPO 7q-AML study

Due to the large number of differentially expressed genes including key haematological regulators, we attempted to determine the direct targets of *MNX1* with the help of TF motif binding database information, where we used Motifmap [627], Transcription factor target gene database [628] with SELEX [629] and TRANSFAC [630] and finally TF2DNA [631] with SELEX [629], with the databases in great disagreement between each other (Figure 4.25). These results, and the lack of representation for individually dissected genes showing a great correlation with *MNX1* expression led us think to that these databases could be incomplete, especially with regards to the myeloid tissue relevant for our project.

Gene	Motifmap	TFBSDB (TRANSFAC)	TFBSDB (SELEX)	TF2DNA (SELEX)
TLE4 (down)	x	x	x	
VIM (down)			x	
EVI2B (down)			x	
ACOT9 (down)			x	
RALB (down)			x	
TFEC(down)				x
GPC6 (up)			x	
SOAT2 (up)			x	
DSG2(up)			x	
PLCB4 (up)		x	x	
IL17RE (up)			x	
MYT1 (up)		x		
BMPR1B (up)				x

Figure 4.25: Putative direct targets of *MNX1* binding among the dysregulated genes in the HIPO 7q-AML study

Among the upregulated genes, we selected a shortlist with strong differential expression and high correlation as representative upregulated genes in *MNX1*-activated adult AML: *FAM155B*, *KRT72*, *STXBP6* and *IL17RE*. *FAM155B* is an uncharacterized gene and is activated from zero expression in our *MNX1*+ cohort. *KRT72* is a member of the large human keratin gene family [632] and is activated along with *KRT2* and *KRT73* in our MNX1+ cohort, AML is as a rule keratin-negative entity and keratinization in AML subtypes has not been shown in the literature. *STXBP6* is a regulator of SNARE proteins, which serve in mediating vesicle fusion such as synaptic vesicles with the presynaptic membrane in neurons [633]. *IL17RE* is a functional receptor for *IL17C* and has functions in the development of immune responses to infection [634].

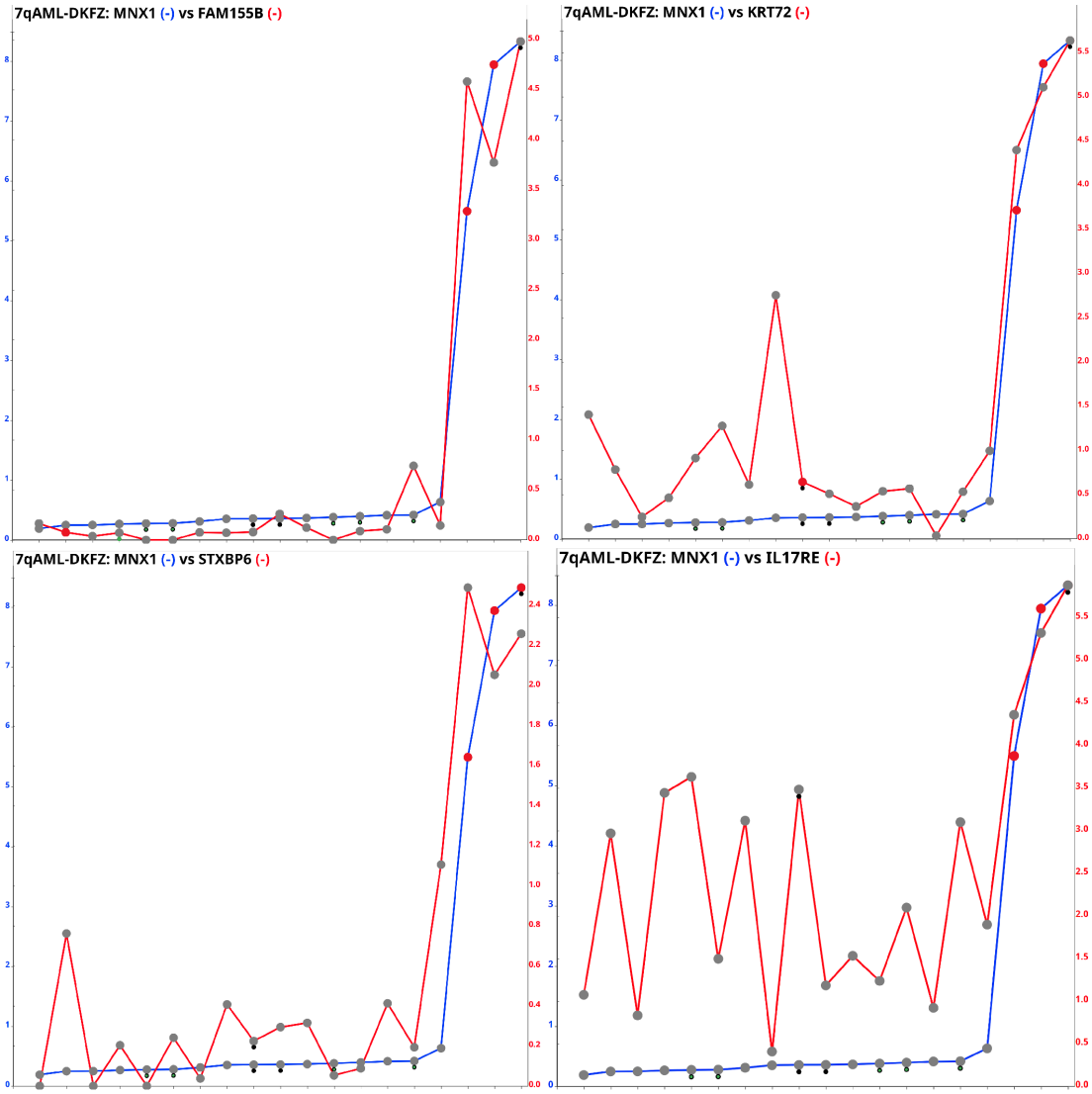

Figure 4.26: MNX1 activates/upregulates *FAM155B*, *KRT72*, *STXBP6* and *IL17RE** in the HIPO 7q-AML study (* predicted direct binding)

To cross-validate our findings in a different cohort with a different, paediatric, setting, we used the dataset from [635] and generated the Figure 4.27 in R2 [337]. We observed that the upregulation of *KRT72* and *STXBP6* are maintained in the paediatric t(7;12) AML, while *FAM155B* and *IL17RE* are not upregulated in this different age group. This could be due to the biological differences in infant AML, as *IL17RE* is indeed a predicted direct target of MNX1 according to TF database data.

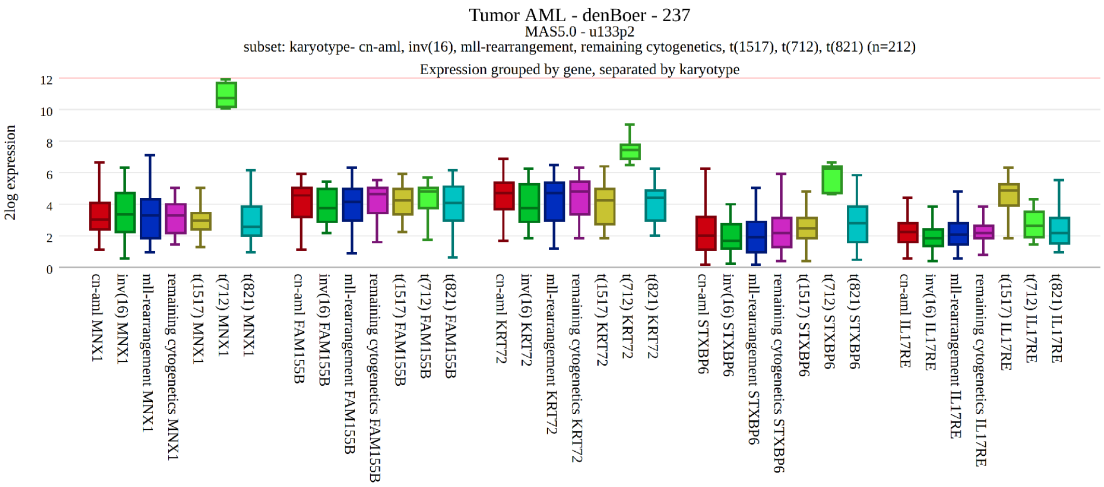

Figure 4.27: MNX1 activates/upregulates *KRT72* and *STXBP6* but not *FAM155B* and *IL17RE** in t(7;12) paediatric AML (* predicted direct binding)

Next, we shortlisted 8/34 genes among the downregulated gene list, pritoritizing direct binding predictions from databases and strong downregulation and inverse correlation characteristics (Figure 4.28). *TLE4* has been identified as a tumour suppressor in AML, with antiproliferative and apoptotic functions validated in *in vitro* experiments [636]. *GAPT* is a regulator of B-cell proliferation [637], its aberrant expression or mutations have not been described to date in cancer research. *EVI2B* is a downstream target of *CEBPA* and a regulator of granulocytic differentiation of myeloid cells [638]. INPPL1 (formerly *SHIP2*) is a phosphoinositol phosphatase regulating the PI3K/Akt pathway [639], and along with its homolog *INPP5D* (formerly *SHIP1*), were shown to lead to haematopoietic perturbations in myeloid cell development in mice [640], its role in CML but not AML have also been discussed [641].

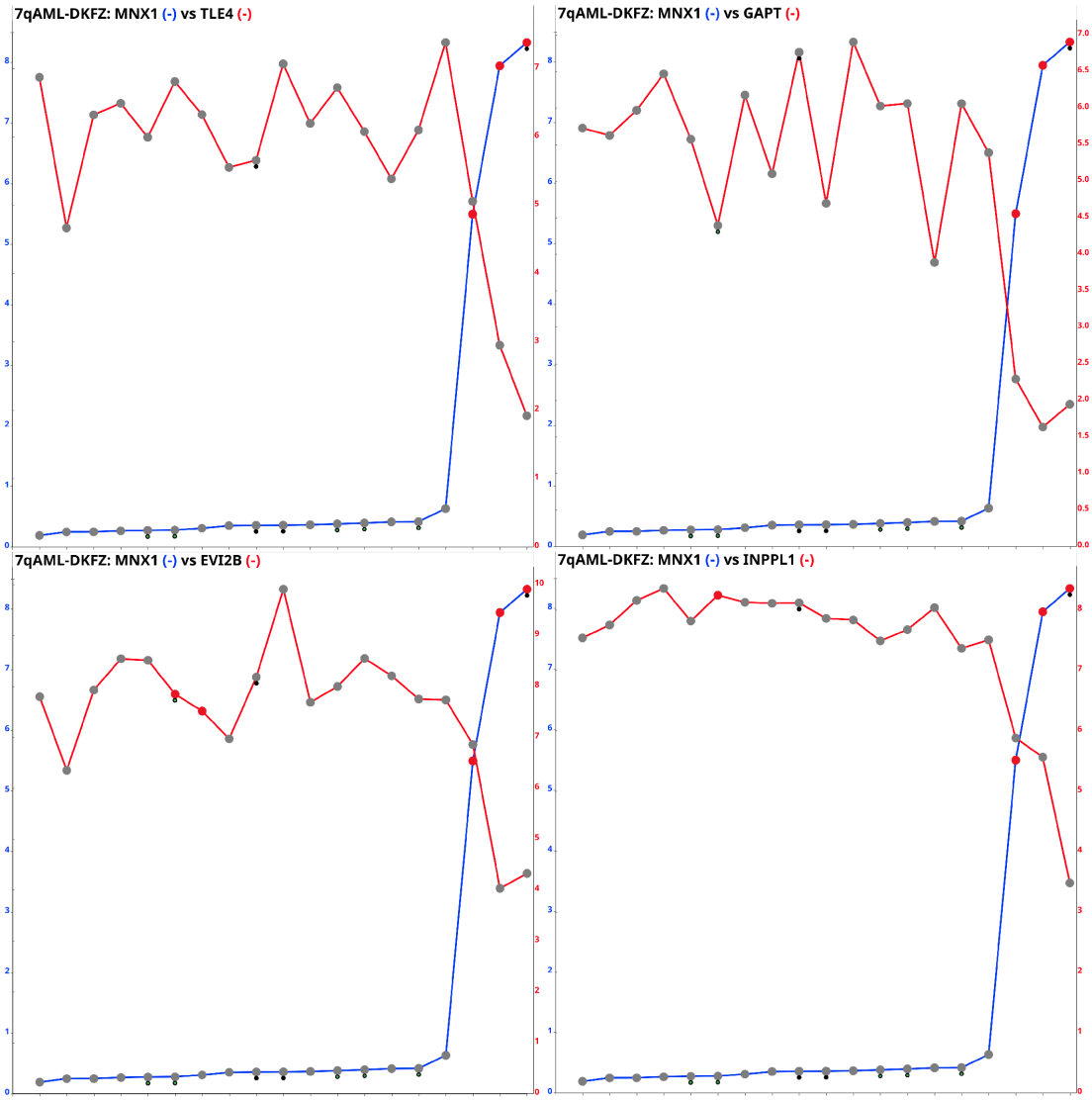

Figure 4.28: MNX1 suppresses *TLE4**, *GAPT*, *EVI2B** and *INPPL1* in the HIPO 7q-AML study (* predicted direct binding)

In cross-validating our results, we established that all four genes *TLE4*, *GAPT*, *EVI2B* and *INPPL1* are also significantly and strongly suppressed in t(7;12) AML [635] (Figure 4.29).

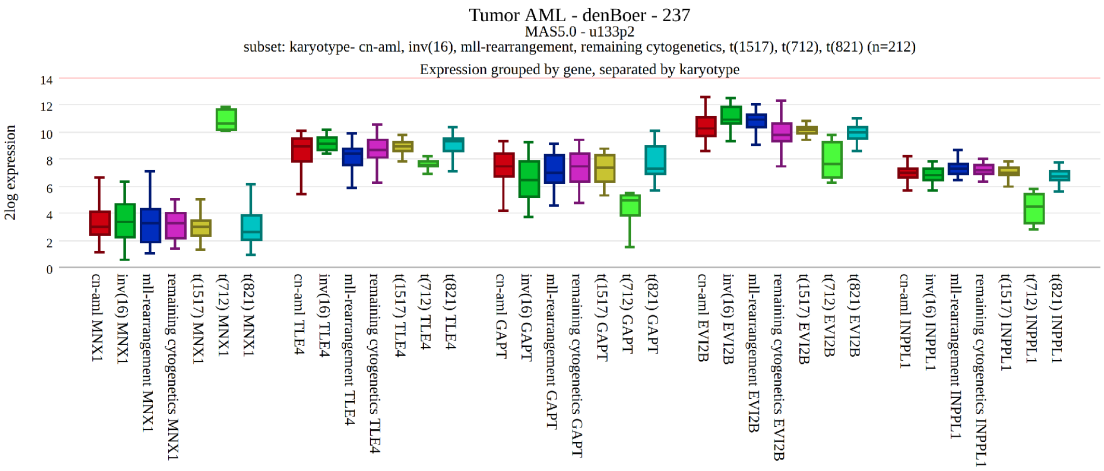

Figure 4.29: *MNX1* suppresses *TLE4**, *GAPT**, *EVI2B** and *INPPL1* in t(7,12) paediatric AML (* predicted direct binding)

We then investigated the next 4 of the shortlisted 8/34 downregulated genes (Figure 4.30). *HLX* is a homeobox transcription factor, previously reported as an oncogene in AML where its overexpression was postulated to lead to a myeloid differentiation block [642] [643], contrasting our results showing a strong suppression. *MYD88* is a myeloid differentiation antigen which is mutated with gain of function mutations in Chronic Lymphocytic Leukaemia (CLL) [644] and B Cell lymphoma [645], again in contrast to our suppression observation. Vimentin (*VIM*) is a well-established histopathological marker of mesenchymally-derived cells in human cancers, and its suppression in neuronal cells is to be expected, even though its expression in cells of haematological lineage are also not high [646]. *TFEC* is an E-box basic helix-oop-helix transcription factor that acts as a transcriptional repressor of *TFE3*-based transcription activation [647] and *TFE3* has known oncogenic functions in juvenile renal cell carcinoma and acts in coordination with Leukaemia Inhibitory Factor (*LIF*) [648], in our cohort it is among the most strongly suppressed genes and is a predicted direct target of *MNX1*.

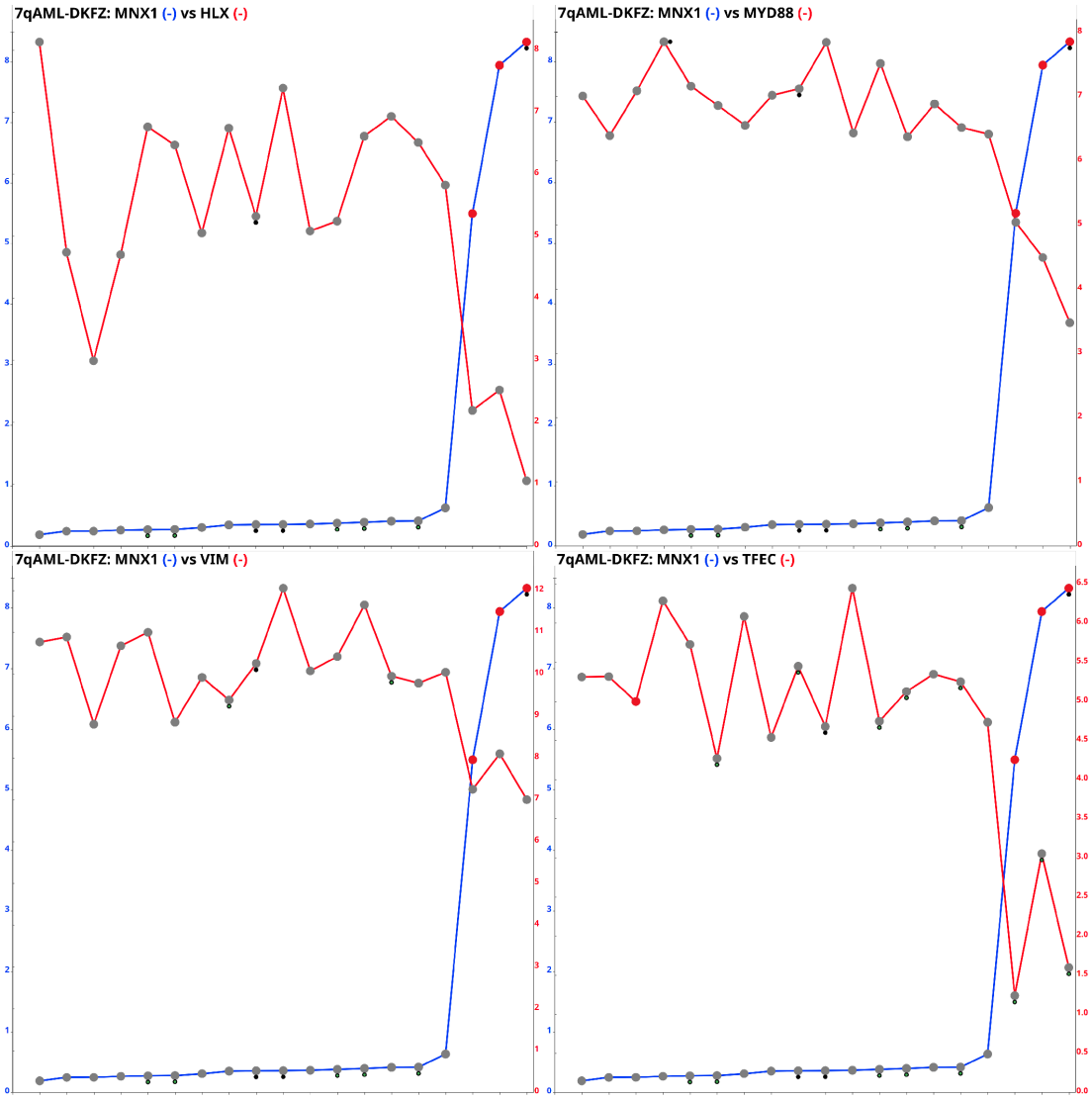

Figure 4.30: MNX1 suppresses *HLX*, *MYD88*, *VIM** and *TFEC** in the HIPO 7q-AML study (* predicted direct binding)

As with *TLE4*, *GAPT*, *EVI2B* and *INPPL1*, this second investigated set of genes *HLX*, *MYD88*, *VIM* and *TFEC* are also significantly and strongly suppressed in paediatric t(7;12) AML [635] (Figure 4.31).

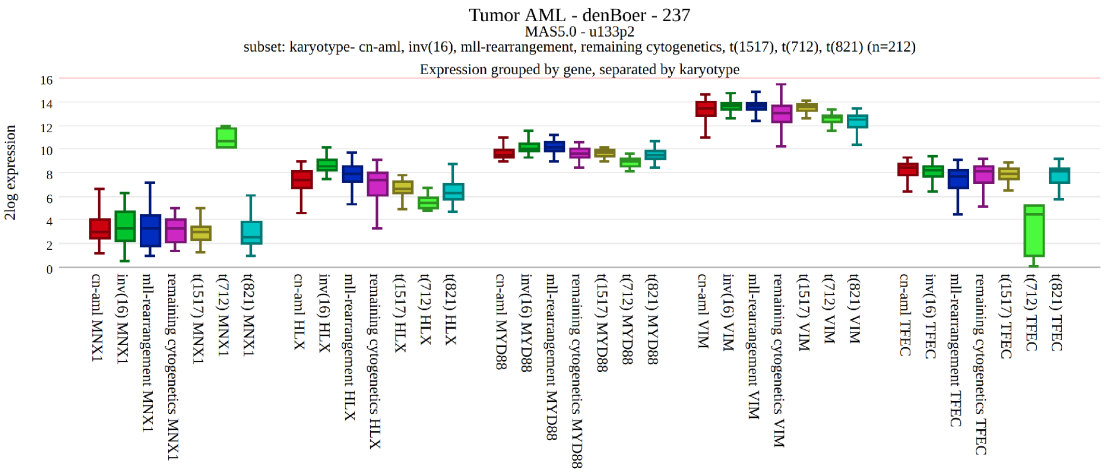

Figure 4.31: *MNX1* suppresses *HLX*, *MYD88*, *VIM** and *TFEC** in t(7,12) paediatric AML (* predicted direct binding)

4.4.3 Discussion

We investigated using EPISTEME, genomic and transcriptomic sequencing data from a set of patients of AML with monosomy on chromosome 7q (7q-AML). We identified *CDK6-NOM1* rearrangements to be a strong activator of the *MNX1* oncogene. Further analysis on an extension cohort strongly suggests that *MNX1* activation in AML is dependent on structural variants targeting *MNX1* and activating it via enhancer hijacking.

The classical two-hit model of AML development requires somatic mutations increasing proliferation functioning in tandem with somatic alterations blocking myeloid cell differentiation. We investigated in great detail the dysregulated transcriptome of 7q-AML, revealing a strong dysregulation of tumour suppressors, myeloid differentiation genes, homoebox genes. Our results suggest that *MNX1* activation fulfils the myeloid differentiation block axis of AML development, whereas the proliferation axis could be coming from secondary mutations such as *BCOR* and *TET2* in our cohort.

One of the most interesting aspects of *MNX1* function seems to be its dual effects on the myeloid stem cell transcriptome: neuronal development genes are upregulated and myeloid development genes are downregulated in a process establishing cell identity. In its normal function as a motor neuron homeobox gene, *MNX1* thus seems to suppress developmental pathways that confer cells a haematological identity. We postulate that this natural function of *MNX1* is hijacked in AML in establishing a differentiation block.

Our results show no similarity to recently published, experimentally obtained in vitro data using aberrant activation of *MNX1* in CD34+ haematopoietic stem cells [618]. While the complete lack of any commonly differentially genes in our patient and our study's experimental data would normally be a cause for concern, our cross-validation analyses with a patient-based study of t(7;12) infant AML [635] largely confirmed our strongest results. Remarkably, we found key players strongly downregulated upon *MNX1* activation such as *EVI2B*, *TLE4*, *INPPL1*(*SHIP2*), *HLX*, *TFEC*, *GAPT* and *MYD88*, and to our surprise saw that none of these key regulators of

myeloid differentiation and proliferation have been discussed in the context of *MNX1* despite availability of data [635]. In particular, our results are novel in suggesting new mechanisms in AML dedifferentiation, introducing a direct, TF-promoter-interaction-based suppression of *TLE4* without copy number losses and *EVI2B* without *CEBPA* mutations respectively. ChIP-Seq on *MNX1* will help determine direct targets beyond motif-based approaches in databases, which proved unsatisfactory. In parallel, we are going to add 4 more cases of *MNX1*-activated adult AML to our study, which would drastically improve the quality of our transcriptomic analysis.

As *MNX1* as a homeobox transcription factor, is not a targetable gene, our transcriptome data could be useful in designing treatment concepts: We observed a strong suppression of *INPPL1* (*SHIP2*) and *PIK3CG*, which could lead to a treatment strategy targeting the PI3K pathway. Unfortunately the strongly activated genes *FAM155B* and *KRT2*,*KRT72*, *KRT73* are not appropriate as cancer antigen based immunotherapeutic targeting due to their high expression in the heart and skin, respectively.

Recently another study postulated that *CDK6-MNX1* rearrangements could be activating *MNX1* due to nuclear reorganization of *MNX1* location, without discussing the role of *CDK6* or other partner enhancers or investigating the mechanistic downstream effects or co-occurring mutations with *MNX1* activation [649]. We believe we have a more correct and complete model of *MNX1*-mediated *AML* oncogenesis with our current data.

Finally, neither of the two major recent publications on AML recognize *MNX1*-activated cases as an entity [483] [599], and the WHO classification does not recognize the paediatric *MNX1*-activated AML as a *bona fide* subtype of AML due to its rarity. It remains to be seen if our case collection will promote *MNX1*-translocated AML as an officially recognized subtype.

4.5 Case Study 3: ATOH1 is a Novel Target of Enhancer Hijacking in MYCN-Negative High-Risk Neuroblastoma

Neuroblastoma (NB) is a malignancy of neural crest stem cells. Neural crest cells are transient and multipotent precursors of a wide variety of neural and non-neuroanal cell types including sympathoadrenal cells during development. Sympathoadrenal cells encompass sympathetic neurons and chromaffin cells and constitute the lineage of cells from which NB develops [650]. NB is a childhood disease because neural crest cells are fully differentiated, and do not exist and consequently cannot generate malignant tumours during adulthood (adult NB is an exceedingly rare condition with an unclear cell of origin [651] and is outside the scope of this study). With methods of epigenetic profiling [652] and single cell transcriptomics [653], we are in the process of delineating the specific cellular origins in NB, which remains an open field similar to most solid tumour types.

Since early whole exome sequencing based results indicating the low mutational load of NB [654], it has consistently appeared among the cancer types with the lowest number of somatic point mutations [655] [476]. The driver mutational processes in NB are known to prefer alterations based on copy number changes, gene amplifications, in a group of alterations which can be summarized using the structural variation umbrella term. Following the landmark discovery of the amplification-based activation of the *MYCN* (formerly and also known as *N-MYC*) onco-

gene [656] [200], great progress has been made on the further dissection of this disease with a risk classification from clinical parameters stage, age, chr1p deletion and *MYCN* amplification status [657] effectively guiding clinical decision making. This risk classification was further improved with a support vector machine based classifier using gene expression profiles [658], accurately predicting clinical outcome. Interestingly, despite the availability of large-scale and genome-wide transcriptome datasets [654], [659] [660], there have so far been no efforts to use modern methods of transcriptomic classification of NB into molecular subtypes. Our knowledge on the NB mutational landscape and biology greatly improved with advances in genome sequencing technologies and RNA sequencing together revealing mutational drivers of NB. NB arises due to dysregulation of differentiation of proliferating neural crest stem cells during infant/childhood development and consequently does not have a direct influence of mutagens and does not carry a high load of somatic mutations including protein coding mutations. Genomic variant driver hallmarks of NB are in rough order of frequency:

1. Chromosome 17q gains [661]
2. *MYCN* amplifications [656] [200],
3. Chromosome 1p arm losses or loss of heterozygosity [662],
4. Chromosome 11q losses [663]
5. *TERT* structural variants leading to activation via enhancer hijacking [212]
6. *ATRX* truncating deletions and deactivating small mutations [664], [654]
7. *ALK* co-amplifications [665] [666]
8. *CDK4* and *MDM2* amplifications [667]
9. *FOXR1* structural variants leading to activation via enhancer hijacking [461]
10. *LIN28B* amplifications [668].

A common theme in the mutational drivers of NB is the concept of telomere maintenance mechanisms: to maintain proliferative potential, high-risk NB cells need to maintain their telomeres. The first telomere maintenance mechanism in NB is via *TERT* expression either activated by *MYCN* amplification or by SVs. The second is alternative lengthening of telomeres (ALT) [669], which happens in around half of the cases via *ATRX* deactivating mutations and in the other half via an unknown mechanism and is detected by an assay named ”C-circle”. The recently established and current consensus on NB clinical risk characteristics is based on a combination of *telomere maintenance* status and *RAS* mutation status: cases with no telomere maintenance mechanism are considered low-risk, cases with a telomere maintenance mechanism are high-risk, and cases with both a telomere maintenance mechanism and RAS pathway mutations are considered very-high-risk [506]. The role of RAS-MAPK pathway genes and their mutations were previously discussed in a study where they were found to be enriched in relapsed NB [670], with the current results validating this assessment.

Neuroblastoma remains to date a clinical challenge: NB is the most common paediatric cancer type after leukaemia and central nervous system malignancies (such as medulloblastoma and PNETs as discussed in Chapter 1). Overall, NB accounts for 8% of all paediatric cancer cases, shows a heterogeneous clinical course [671] and with cases classified as "high-risk" remaining largely incurable [672] [506]. Historically, survival rates improved thanks to aggressive treatment protocols [673] [674]. However, the lack of effective targeted treatments for *MYCN* amplification, *TERT* activation or the ALT mechanism and a lack of a mutation/neoantigen load preventing checkpoint blockade based immunotherapeutic strategies, together lead to a poor prognosis for all cases that fall under the high-risk classification. A better understanding of the NB biology is therefore urgently required. To address this need, we designed a project to characterize the molecular subtypes of NB with advanced transcriptome and methylome based clustering using modern statistical approaches in a larger cohort compared to previous studies such as [675] [676]. We also aimed to use EPISTEME's genomic variant-based transcriptomic dysregulation methods to look for novel targets of enhancer hijacking events beyond the established prototypical examples of this mechanism, *TERT* and *FOXR1*.

In this study, we describe the mutational landscape of paediatric NB and discuss a potential role for a focal-SV-based truncation of tumour suppressor genes *ANKS1B*, *ZFHX3*, *DLG2*, *CNTNAP2*, *TENM3*, *AGBL3* and *PTPRD*, in line with recent findings from other groups. We identify for the first time the Basic-Helix-Loop-Helix (BHLH) transcription factor *ATOH1* as an oncogene in NB showing that it is recurrently activated by an enhancer hijacking process, predominantly with the *HAND2* enhancer. We propose first steps towards a novel transcriptome-based subgroups of NB and show their agreement with methylome-based subgroups. We introduce a novel transcriptome and methylome subtype of ALT NB encompassing two sub-subtypes with defined driver genes *ATOH1* among younger patients and ALK among older patients. Finally, we show results postulating that *ATOH1* could act as a less potent MYCN replacement in its function as a BHLH gene, with common targets *NHLH2* and *DLL3* between *MYCN* and *ATOH1*.

4.5.1 Study Design and Methods

In the DKFZ Division of Neuroblastoma Genomics (Dr. Frank Westermann), we ran Whole Genome Sequencing (WGS) on 246 NB cases from the German GPOH NB Study's central sample collection, of which 191 cases had sufficient quality RNA extracted for RNA-Seq and 121 cases were profiled with methylome arrays using the standard DKFZ protocols [353]. 3 specimens (2 patients with rare brain metastasis of NB of which 2 relapses were sampled from a single case) had RNA-Seq data and no WGS or methylome array data.

Case collection was not made in a prospective or unbiased manner, i.e. patient selection was enriched for characteristics of interest such as high-risk cases, expressors of rare oncogenes, ALT cases and ALT cases without *ATRX* mutations. Therefore, the results presented here are interpreted and discussed without assumptions on mutation/phenotype frequencies and epidemiological characteristics.

4.5.2 Results

4.5.2.1 Mutational Landscape of Neuroblastoma

We first analysed with an EPISTEME cohort-wide Circos plot the mutational landscape of NB as shown in Figure 4.32. Due to the prevalence and significance of SV-based mutational and biological mechanisms in NB, the mutation class "Direct SV hit on gene body" is included in the gene-based mutation recurrence layer (outermost layer on the figure). SV recurrence analysis (innermost layer on the figure) results here largely confirmed the established mutational drivers in NB such as *MYCN*, *ALK*, *ATRX*, *TERT*, *CDK4*, *MDM4*, and focally converging SVs on sites of chr17q gains and chr11q losses. We found with the gene-hit analysis highly recurrent SVs and mutations hitting and truncating the genes (SV cases + small variant cases, gene size): *ANKS1B* (15+1 cases, 1.3 Mb), *ZFHX3* (12+2 cases, 0.4 Mb), *DLG2* (12+1 cases, 2.2 Mb), *CNTNAP2* (23+2 cases, 2.3 Mb), *TENM3* (13 cases, 0.7Mb), *AGBL3* (12 cases, 1.5Mb), *PTPRD* (44 cases, 2.3Mb), *SHANK2* (11+1 cases, 0.6 Mb), *DMD* (12+3 cases, 2.2Mb), *EYS* (15+1, cases 2.0Mb).

Figure 4.32: Mutational landscape of the GPOH NB study. Recurrence layers from outer to inner correspond to gene-based functional small variant + direct (gene-body) SV hit recurrence, TAD-based copy number variant recurrence, TAD-based SV recurrence (1-TAD offset).

Structural variation preferentially affecting larger genes [619] and open chromatin / highly expressed genes [677] as fragile sites is well-known. Our results here are also in line with pre-

vious findings on neurite development genes being directly hit by SVs and acting as deactivated tumour suppressors in NB [678], [679].

4.5.2.2 Discovery of novel targets of Enhancer Hijacking in Neuroblastoma

Next, we investigated the genomic variant driven transcriptomic dysregulation landscape of the NB cohort (Figure 4.33). In an unbiased analysis, we found the dominating effect of the hallmark loci such as *MYCN*, *TERT*, *ALK*, *CDK4* and *MDM2* where the amplifications and other structural rearrangements lead to a broader local activation of neighbour genes. Suppressing these common loci of genomic variant driven transcriptomic dysregulation, emphasizes or reveals less common genes such as *LIN28B* and *FOXR1* while *ATOH1* emerges as a potentially novel candidate of recurrent enhancer hijacking events among others such as *FGF19*, *C1QL1*, *S100B* and *TP73*.

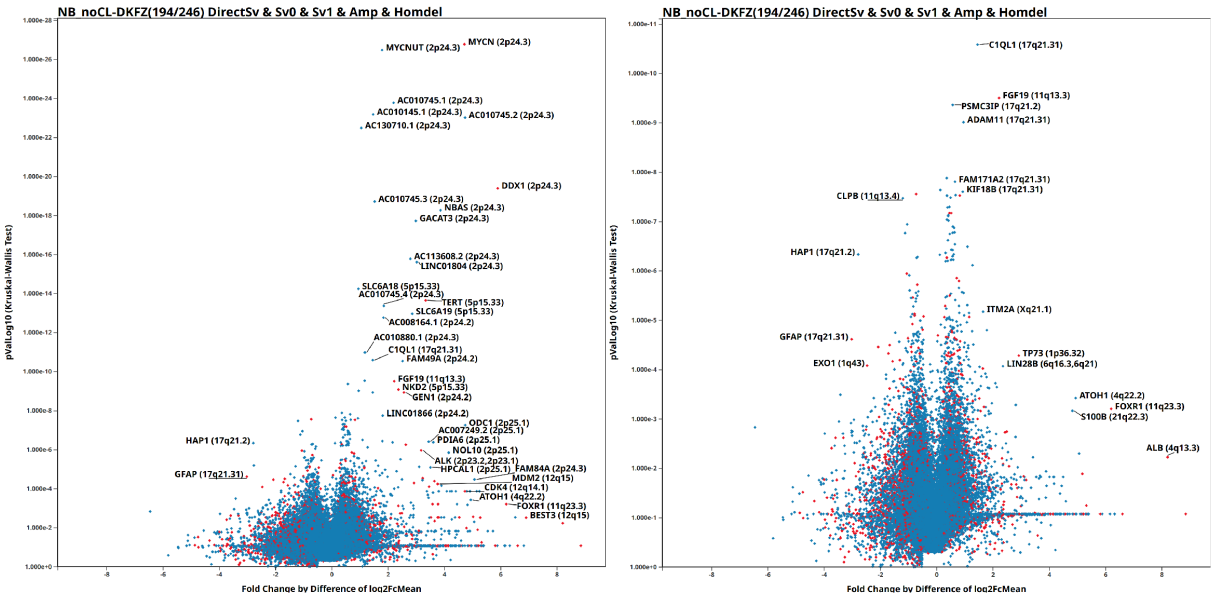

Figure 4.33: Genomic variant driven transcriptomic dysregulation landscape of the GPOH NB study. Left: without restrictions, Right: masking commonly dysregulated cytobands (*MYCN*, chr2p24.2, chr2p25.1, *TERT*, *ALK*, *MDM2*, *CDK4*)

We observed *ATOH1* to follow a similar profile to *FOXR1* and *TERT* activation where the gene is initially strongly suppressed and activated by nearby SVs in 7/8 cases where it shows high expression. It is therefore a candidate for a novel hallmark gene. Figure 4.34 shows the amplification / activation characteristics of the *MYCN*, *TERT*, *FOXR1* (5/5) and *ATOH1* (7/8) genes.

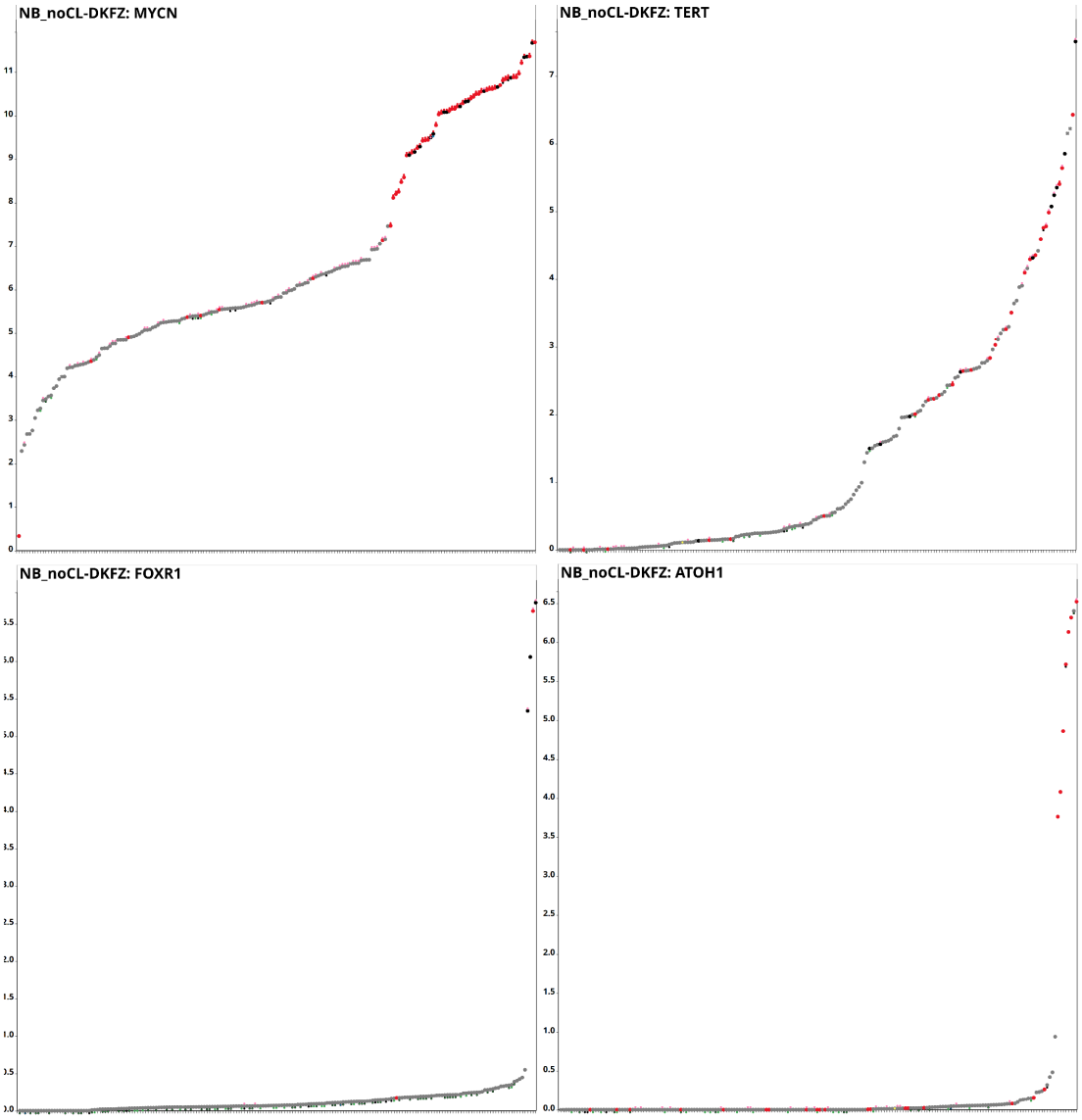

Figure 4.34: *MYCN*, *TERT*, *FOXR1* and *ATOH1* as hallmark genes and recurrent targets of enhancer hijacking in neuroblastoma. Red and black coloured cases respectively have off-gene and gene-body SV hits activating/upregulating their expression. Amplifications are denoted by small upper red circles.

It is established that *MYCN* and *TERT* expression follow a common and partially interdependent profile (Section 3.3.2.1), but *FOXR1* expression is assumed to be mutually exclusive to both *MYCN* and *TERT* based on the early results from [461]. We investigated the hypothesis that *ATOH1*, due to its precise activation pattern could be part of a broader mutual exclusivity relationship with *MYCN*, *TERT* and *FOXR1*. Using EPISTEME, we set cutoff points for high and low *MYCN*, *TERT*, *FOXR1* and *ATOH1* expression and created high and low expressor groups of patients. We then created a categorical variable combining the membership to these binary groups. The results shown on Figure 4.35 show an imperfect but very strong mutual exclusivity pattern involving these hallmark genes:

- Out of 8 *ATOH1* expressor cases, 1 is *MYCN* amplified and highly expresses *MYCN* and *TERT*, 2 highly express *TERT*, and 5 are in an ATOH1-only group.
- Out of 5 *FOXR1* expressor cases, 1 highly expresses *TERT*, and 4 are in an *FOXR1*-only group.
- *MYCN* and *TERT* are in an expected interdependent relationship where *MYCN* amplification upregulates *TERT* and a subset of cases have *TERT* rearrangements and no *MYCN* expression.

. Overall, this result establishes a strong likelihood of a main driver candidacy for *ATOH1*.

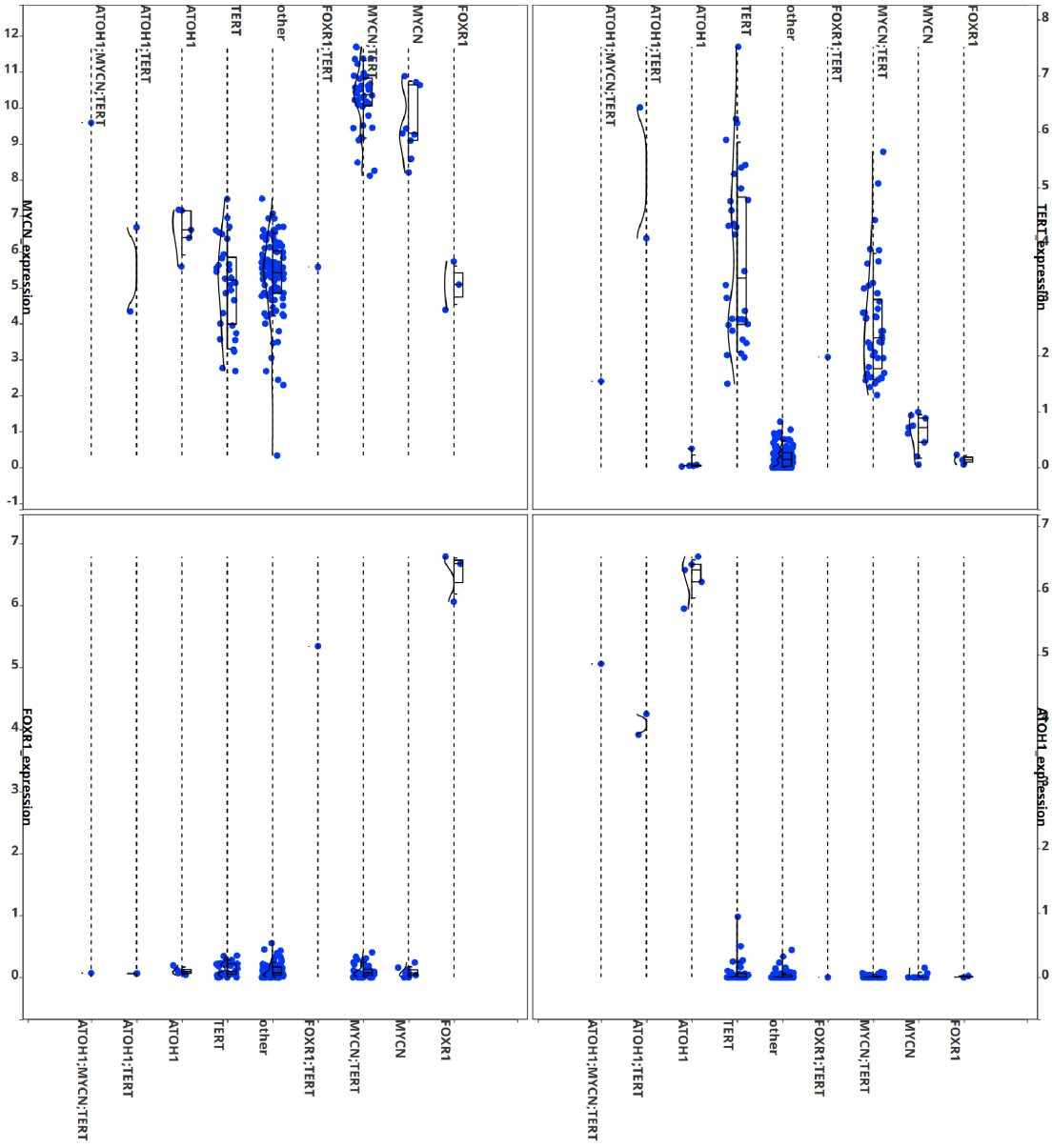

Figure 4.35: *MYCN* (upper left), *TERT* (upper right), *FOXR1* (lower left) and *ATOH1* (lower right) expressions show strong mutual exclusivity patterns as hallmark genes in neuroblastoma. Blue symbols show expressor cases for each gene.

Further analysis of all other outlier data points in the genomic variant driven transcriptomic dysregulation landscape plot revealed no other recurrent candidates for SV-based total (from near-0 expression) activation. We observed *FGF19* on chr11 and *C1Q1L* on chr17, each near frequently translocated loci associated with copy number changes, to show significant, but subtle overexpression patterns (Figure 4.36). We also observed diverse genes with rare overexpressions that correlate with existence of SVs in the proximity of the gene. Two selected genes *TP73* and *S100B* are shown on Figure 4.36.

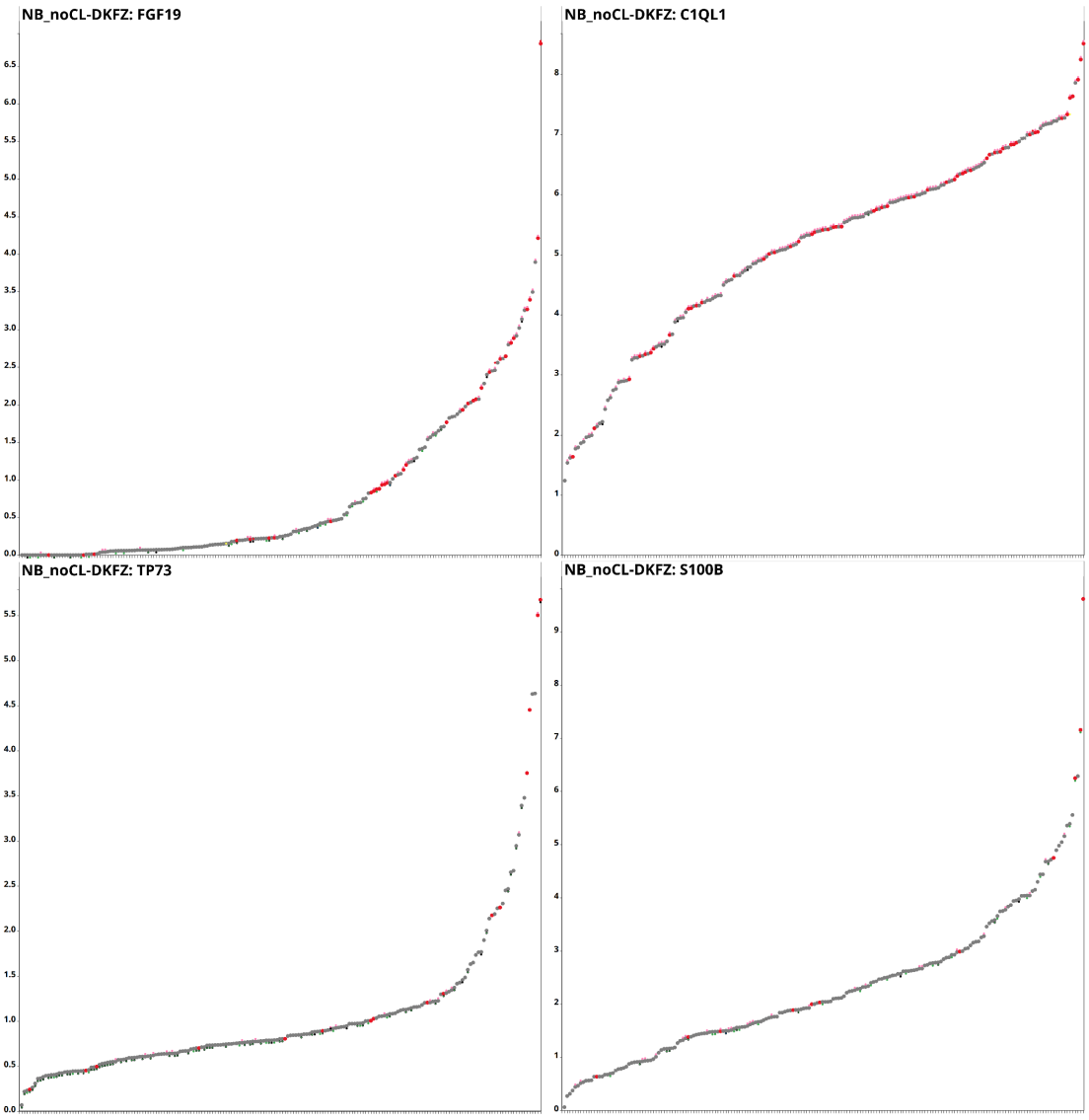

Figure 4.36: *FGF19*, *C1Q1L*, *TP73* and *S100B* are putative novel targets of enhancer hijacking in neuroblastoma. Red and black coloured cases respectively have off-gene and gene-body SV hits activating/upregulating their expression.

Our results reveal *ATOH1* (formerly called *MATH1*) as the by far strongest new target of enhancer hijacking in NB. *ATOH1* is a Basic Helix-Loop-Helix (BHLH) transcription factor and a key regulator of proneural development of the cerebellum [680]. As a gene in the same

family of BHLH genes as *MYCN*, it could be of great biological and mechanistic importance in NB. In the rest of this study, we focused on *ATOH1* as a novel oncogene in NB development, characterizing its role across NB subtypes.

4.5.2.3 Transcriptome and Methylome Based Subtyping of Neuroblastoma

To characterize the transcriptomic (cell identity / metabolic state) and methylome (cell identity) profiles of NB, we ran tSNE dimensionality reduction analyses on the top 500 genes and top 5000 CpG probes in terms of intra-cohort variance. The results in Figure 4.37 revealed heterogeneity in the cohort both across the transcriptome and methylome. Older patients (shown in lower panels with larger symbols) co-occurred with the ALT (black symbols) phenotype and were clustered in distinct methylome clusters.

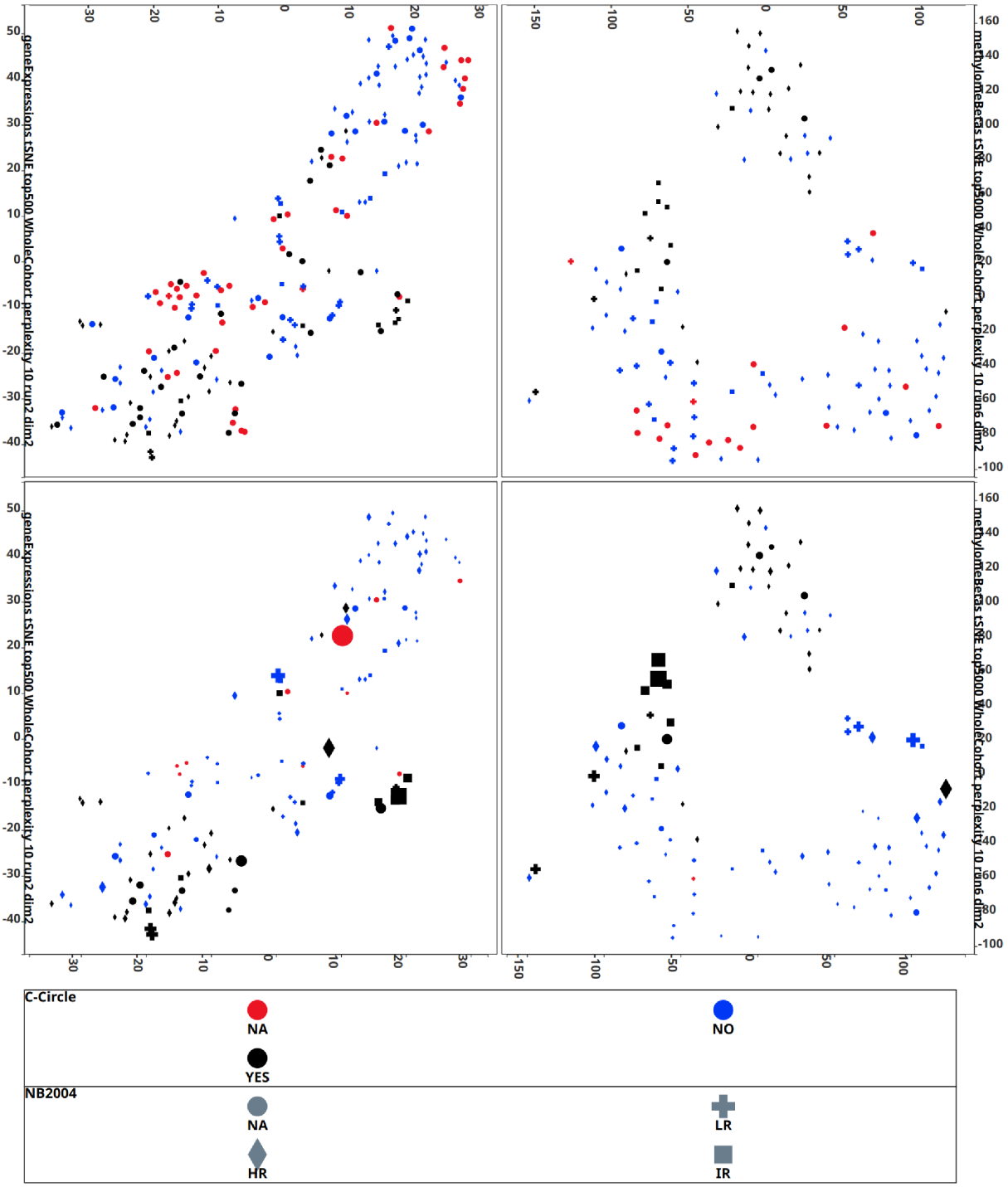

Figure 4.37: Transcriptome (left tSNE top 500 genes, perplexity 10, late-exaggeration 1.1) and Methylome (tSNE top 5000 CpG probes, perplexity 10, late-exaggeration 1.1) based dimensionality reduction reveals distinct molecularly defined subtypes of neuroblastoma. The lower two panels modulate the data point sizes with the ”age at diagnosis”.

Next, we mapped the activation status of the hallmark oncogenic drivers *MYCN*, *TERT*, *FOXR1* and *ATOH1* to the transcriptome tSNE map of the NB cohort. Figure 4.38 shows strong co-clustering patterns for *MYCN*, *TERT*, *ATOH1* but not for *FOXR1*. *MYCN* is active

almost uniquely in the cases on the upper-right of the tSNE map, and in most cases co-activates *TERT* downstream. *TERT* is in addition activated in another subset of cases in the lower-left of the tSNE map, and co-clusters with ALT (black) cases (Figure 4.37). *ATOH1* is highly enriched in a distinct ALT cluster of 4/8 cases in the lower-right of the tSNE map.

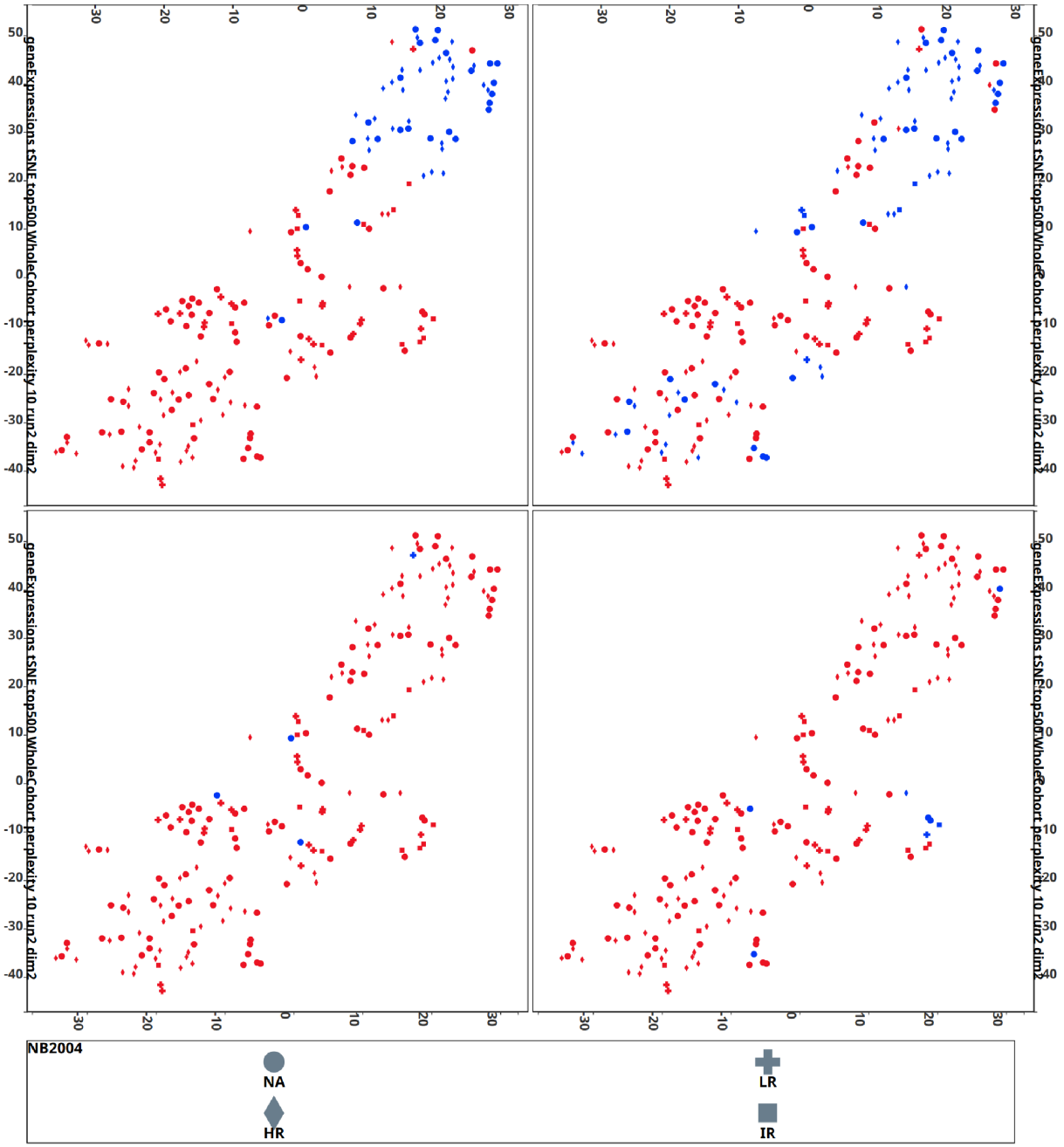

Figure 4.38: Hallmark genes *MYCN* (upper left), *TERT* (upper right), *ATOH1* (lower right) but not *FOXR1* (lower left) follow strong co-clustering patterns across the transcriptomic subtypes of neuroblastoma.

Using manual case selection and differential gene expression analysis between the cohorts, we identified 8 tentative clusters for NB transcriptome profiles.

4 represent larger clusters:

1. high*MYCN*-low*NTRK1*: These cases constitute the most aggressive *MYCN*-amplified subtype of NB and suppress *NTRK1* and activate *BAMBI* as the strongest hallmarks

2. het*MYCN*: This is a heterogeneous cluster with an enrichment for *MYCN* cases, but also including ALT and intermediate-risk cases

3. hetLowRisk-rich: This is a cluster enriched for LR-NB, but includes some *TERT* and ALT cases. Further dissection of this cluster could yield finer groups, and this cluster would most likely not be stable under a tSNE parameter sweep.

4. *TERT*-ALT-rich: This heterogeneous cluster contains *TERT* (rearranged) and ALT (both *ATRX* mutant and wild-type) cases in roughly equal ratios. Interestingly, the ALT and *TERT* groups could not be separated neither in the transcriptome nor in the methylome.

4 of these are smaller clusters:

1. outlierRNA (8 cases): This is a novel cluster of ALT+, *ATRX*wt cases and consists of two distinct subgroups with defined mutational profiles: half of them are older *ATOH1*-rearranged patients and the other half are 3/4 *ALK* mutant, even older patients.

2. low*ABR*-migratory (6 samples, 5 unique cases): The two rare brain metastasis cases (3 samples) co-clustered with 3 primary tumours of NB. They suppress *PLD2* [681] and *ABR* [682], suggesting some role for dysregulated cytoskeletal organization in migratory/metastatic processes.

3. impureRich (5 cases): We identified a small cluster of cases suppressing *TTBK1*, *MYEF2*, *CACNA1B*, strongly overexpressing *SOX17* and *RSPO3*. These cases have low computationally estimated tumour purities. As they are not the only impure tumours in our NB cohort, more work needs to be done to identify the reason for them to co-cluster and to identify possible lineages of infiltrating cells.

4. ALT-slow (4 cases): This is a novel cluster of ALT+ *ATRX*wt cases with suppressed *EZH2*, very low mitotic gene expression such as *MKI67*, *TOP2A*, *MYBL2*, *POLE* implying very slow growth, in line with some previous observations [683].

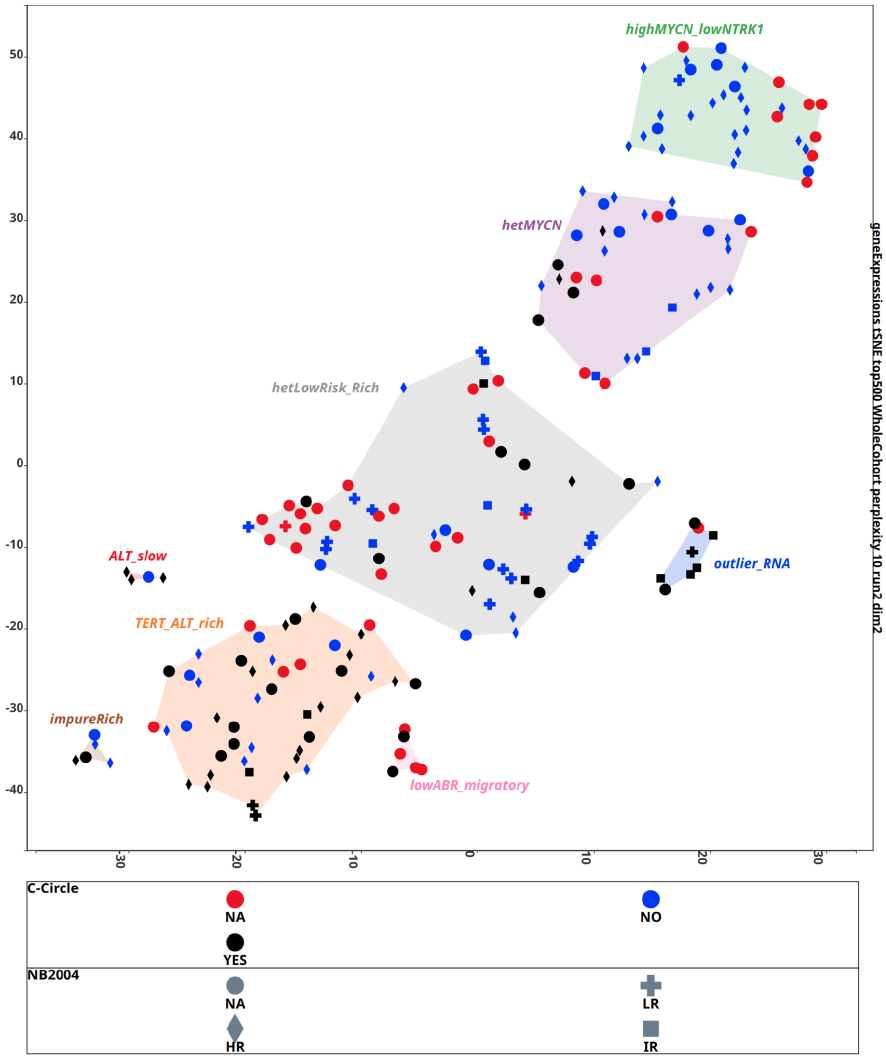

Figure 4.39: Proposed, tentative transcriptome-based clusters of NB

Next, we identified 5 tentative clusters for NB methylome profiles as shown on Figure 4.40.

1. meth*MYCN*-rich: a cluster of cases consisting almost uniquely of *MYCN* amplified cases,
2. methALT-*TERT*: a heterogeneous cluster similar to the *TERT-ALT-rich* cluster identified by transcriptome profiling but with a lower incidence of *TERT* cases,
3. methTERT-LR-rich: a heterogeneous cluster consisting largely of low-risk cases where some *TERT* cases co-cluster,
4. methOld-LR: a small cluster consisting almost uniquely of older low-risk cases
5. methOutlierRNA-rich: a cluster similar to the *outlierRNA* cluster identified by transcriptome profiling, with a less clear separation from the *methTERT-LR-rich* cluster

. Overall, due to the smaller number of cases with available methylome data, a direct comparative analysis is difficult: none of the cases identified as *ALT-slow* were included in our

methylome array series. However, the general patterns seem to be in agreement with two exceptions: 1. The *MYCN* cluster does not separate into two subclusters. 2. a significant number of *TERT* cases co-cluster with LR cases.

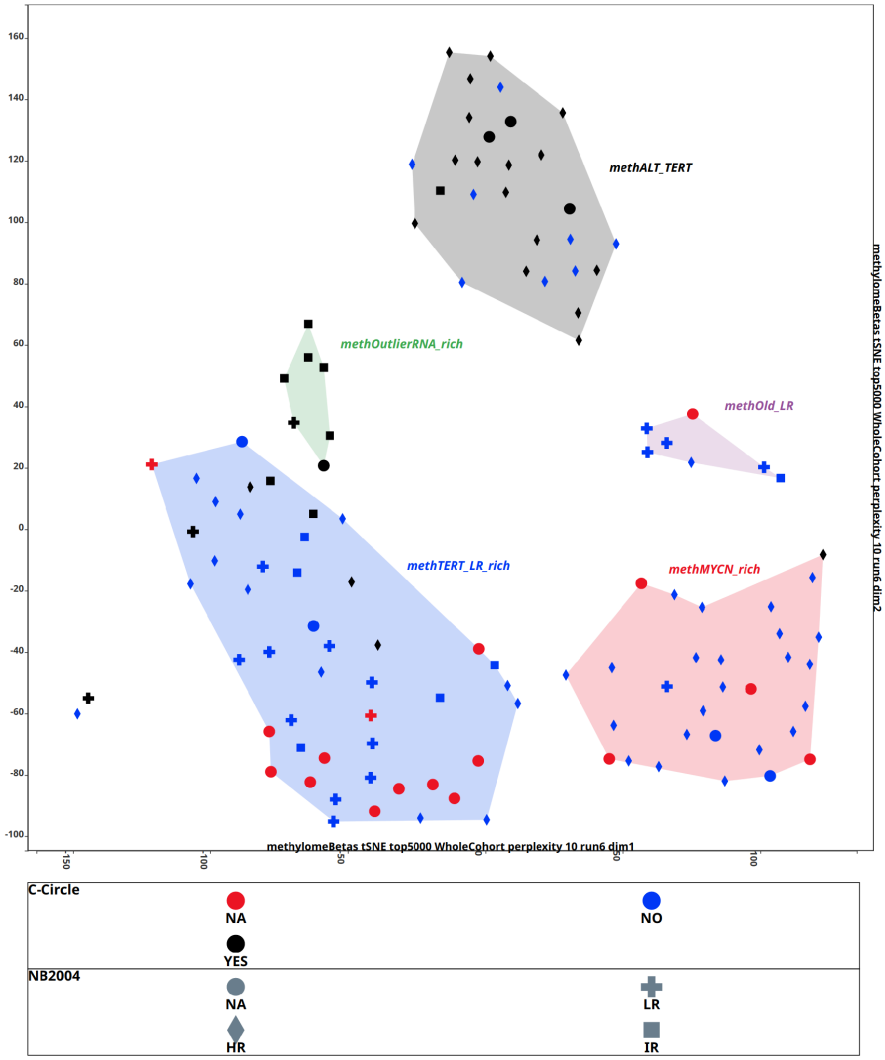

Figure 4.40: Proposed, tentative methylome-based clusters of NB

Due to the expression pattern of the novel enhancer-hijacked oncogene *ATOH1* and its distinct expression pattern clearly different from all other NB subtypes across risk groups and drivers, we focused on the novel transcriptome-based cluster ”outlierRNA” in the rest of this study.

4.5.2.4 *A Novel and Distinct Gene Expression and Methylation Cluster in Neuroblastoma with ATOH1 Enhancer Hijacking and ALK Point Mutations as Driver Events*

The ”outlierRNA” cluster is a transcriptionally defined cluster of 8 cases, which also co-cluster and form a less distinct but observable subtype in methylome analysis (Figure 4.41).

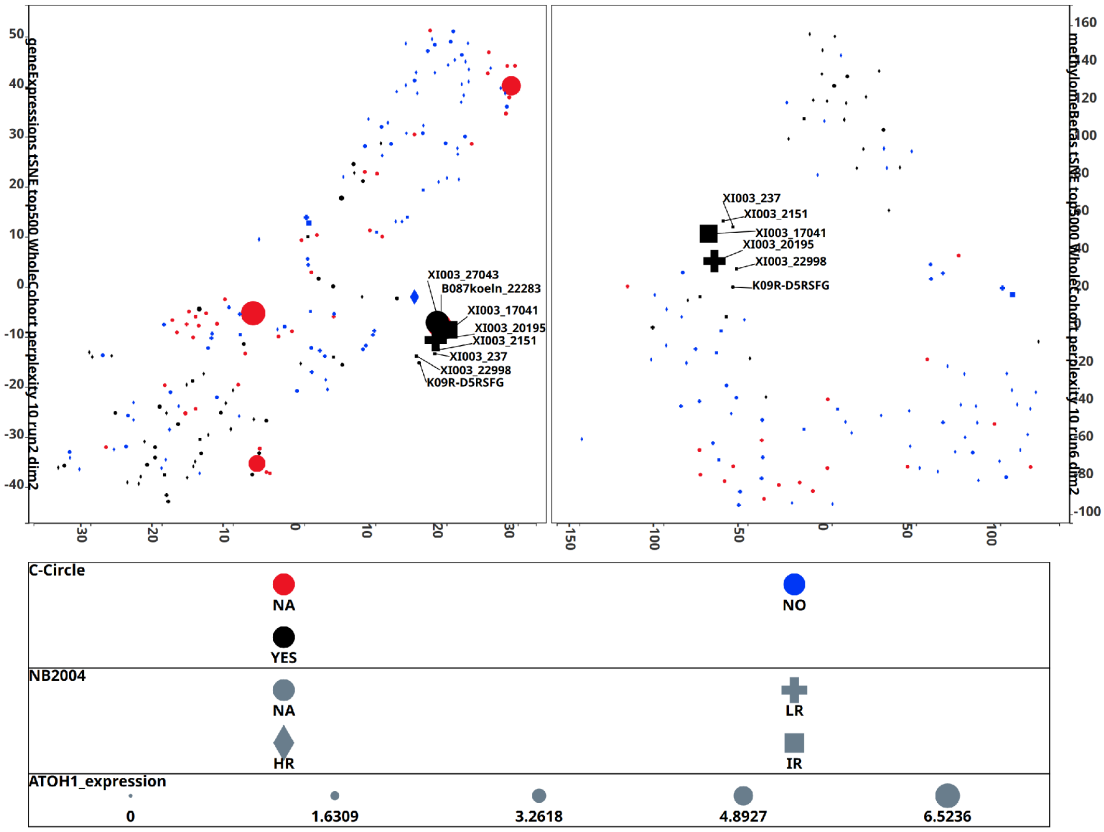

Figure 4.41: A novel and distinct gene expression and methylation cluster in NB is enriched for *ATOH1* expression (encoded in symbol sizes).

In line with the definition of this novel group of cases by transcriptome analysis, they show a distinct transcriptomic profile with high expressions of *NEUROD6* (formerly called *ATOH2*), *KCNH7*, *KCNH5*, *KCNB2* (Figure 4.42, left), hinting at a more mature neuronal developmental state compared to the other NB groups such as *MYCN* and *TERT*/ALT. The most strongly upregulated gene for this cluster *NEUROD6*, and the driver gene relevant for half of the cluster *ATOH1* are shown on Figure 4.42, right.

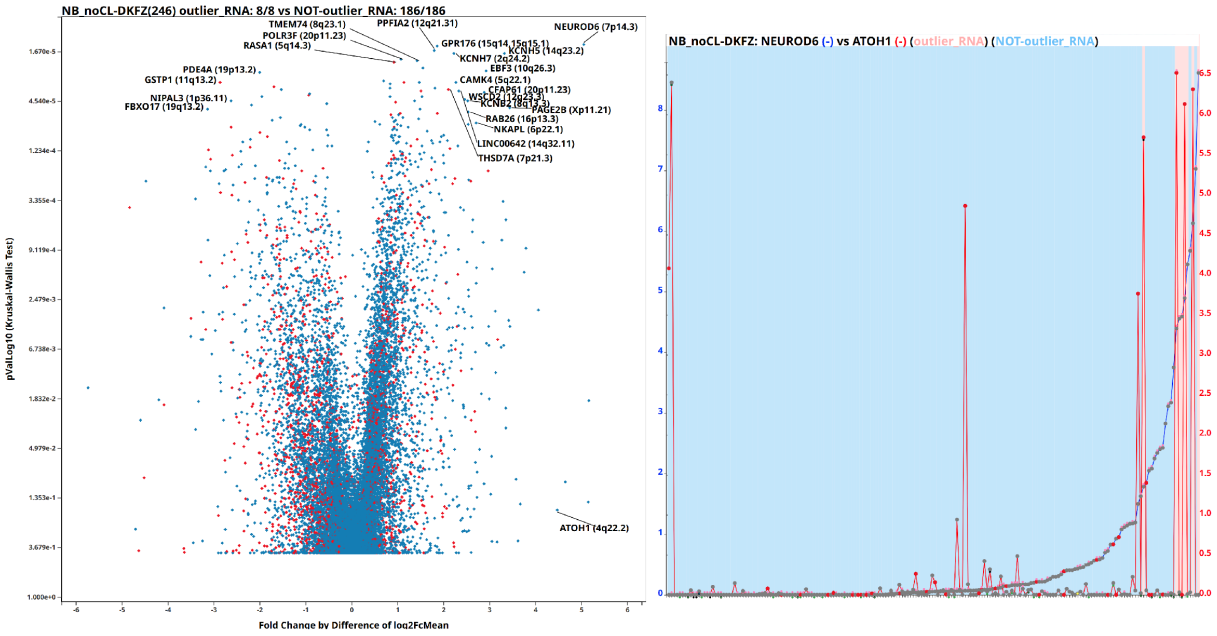

Figure 4.42: Differential gene expression profile of the novel NB subtype.

Comparing the *ATOH1* expressors to the rest of the novel cluster with regards to transcriptomic profiles (Figure 4.43, left) revealed the overexpression of the *NHLH2*, *DOK5* gene in the *ATOH1* group along with a downregulation of the *RET* and *VGF* genes. ALK mutations are likely to lead to the activation of the *RET* and *VGF* oncogenes, along with the downregulation of the *DOK5* gene compared to the *ATOH1*-activated group [684]. *NHLH2*, on the other hand, is a BHLH gene and is upregulated by *ATOH1*, which will be shown in the next section. As this intra-group comparison is expected to regress out the influence of the common cell of origin, the top differentially regulated genes are those directly dysregulated by somatic variants and their downstream effects. From this perspective, the top results are in line with expectations. A systematic comparison of the genomic variant landscapes of the two subgroups(Figure 4.43, right) reveals the discussed prevalence of the *ALK* somatic mutations (3/4) in the *ATOH1*-negative group. *ATOH1* activation is concomitant with chr6 and chr12 and chr20q gains.

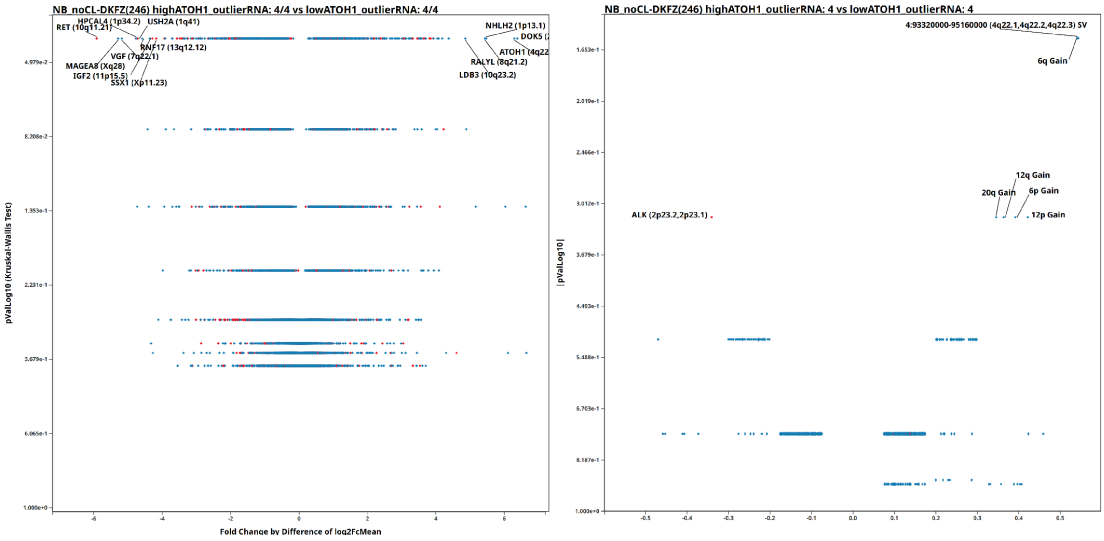

Figure 4.43: Systematic comparison of the gene expression profiles (left) and variant landscapes (right) of *ATOH1*+ and *ATOH1*- subchohorts of the novel NB subtype

Half of the cases express *ATOH1* via recurrent enhancer hijacking events along with the (Figure 4.44, left). The other half have recurrent *ALK* mutations (Figure 4.44, right). Remarkably, 2/4 of the *ATOH1*-negative cases have chr4q rearrangements not activating ATOH1, suggesting the structural fragility of the chr4q arm in this cell state.

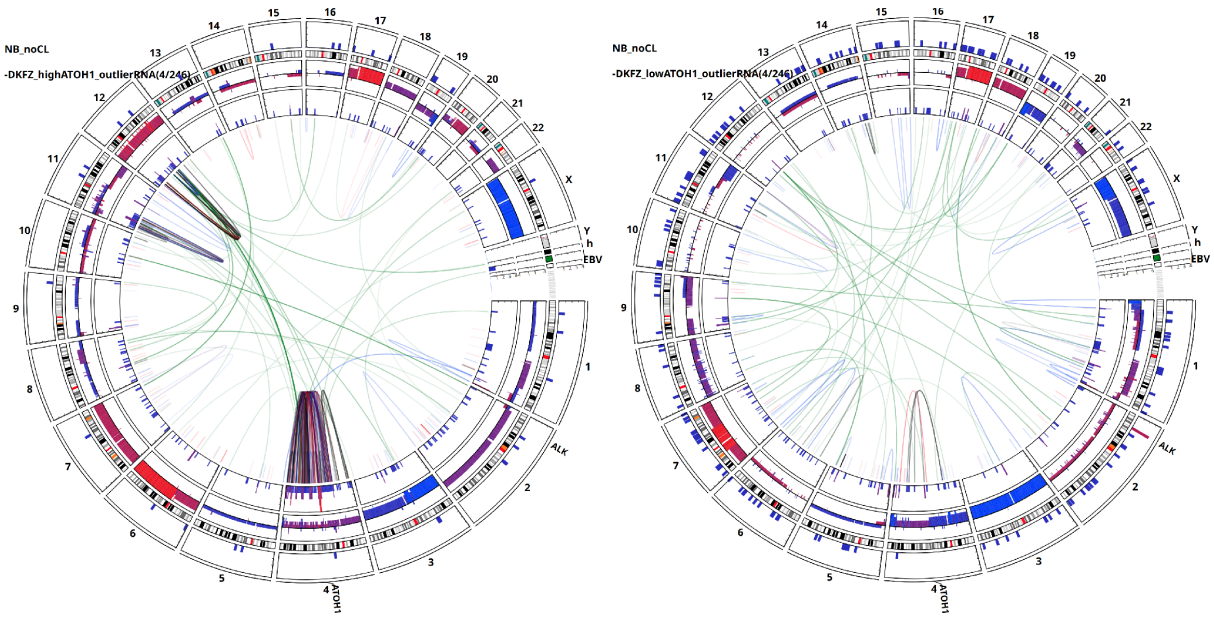

Figure 4.44: Variant landscapes of *ATOH1*+ and *ATOH1*- subchohorts of the novel NB subtype

As the initial observation on Figure 4.41 suggests, the two subgroups of this novel NB subtype show a bimodality with regards to age distribution: Diagnosis of the *ATOH1*-activated

cases was observed between 2.5 and 12 years of age whereas, *ATOH1*-negative cases consisted of 3 cases with *ALK* mutations between 10 and 14 years and one 26 year-old case with no clear driver event. Overall, almost all of these patients can be considered as old in the context of neuroblastoma (Figure 4.45), the physiopathology and clinical course of adolescent and adult NB was discussed in [685] and the slow-growing, indolent course of disease described there is in line with the ALT phenotype observed in this novel subtype of NB.

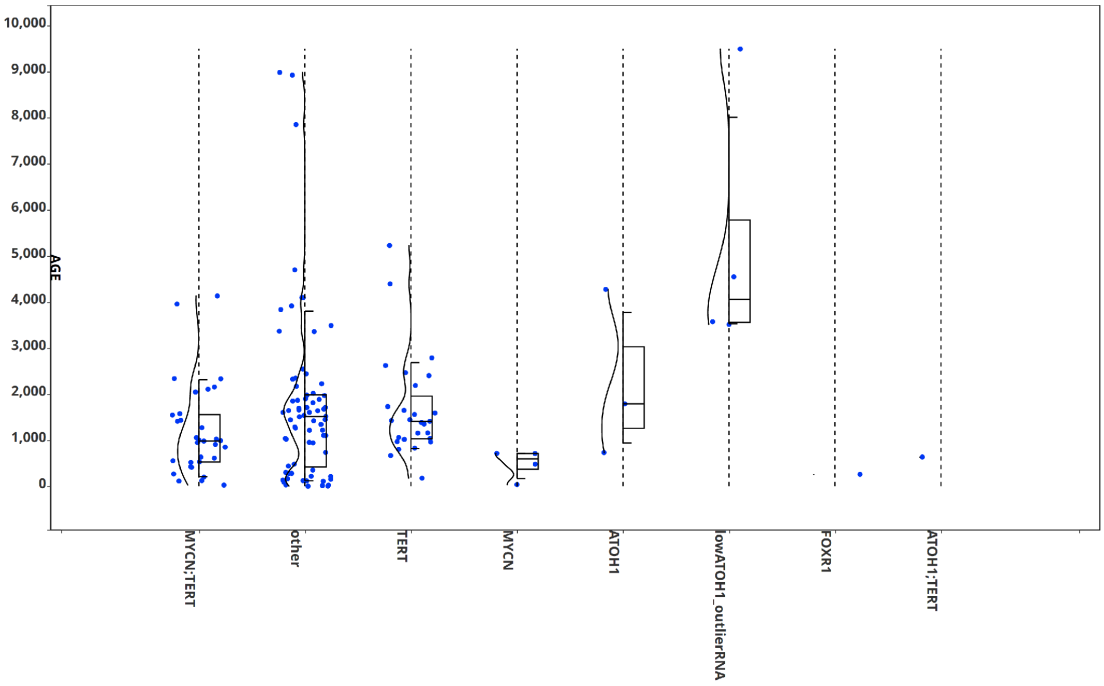

Figure 4.45: The *ATOH1*- subcohort of the novel NB subtype is exclusively seen in older patients

4.5.2.5 Transcriptomic Downstream Targets of ATOH1 in Neuroblastoma

ATOH1 is a BHLH gene, and a transcriptional activator in its normal developmental function [680], [686]. Therefore, the description of its direct targets are of interest in the NB setting, where it is aberrantly activated due to somatic structural variants.

BHLH transcription factors are embryonal developmental genes and are organized in 6 phylogenetic classses driving distinct transcriptional programmes [687]. *ATOH1* and *NEUROD6* (formerly called *ATOH2*) are class A BHLH TFs, whereas *MYCN* and its heterodimer partner *MAX* are class B BHLH TFs. Due to the observation that *ATOH1* activating rearrangements are predominantly seen in *MYCN*-negative and *NEUROD6*-expressing cases, we postulated that *ATOH1* might be replacing *MYCN*, thereby activating some of *MYCN*'s downstream targets upon aberrant activation.

In order to investigate the direct targets of *ATOH1*, and to address the validity of the hypothesis for *ATOH1* as a *MYCN* replacement, we did a differential gene expression analysis with two configurations:

1. *ATOH1*-expressor cases compared to all *ATOH1*-negative cases in the cohort (Figure 4.46, left): this analysis yielded a surprisingly small number of overexpressed genes such as *GNAT3*, *AC022893.2*, *KCNB2*, *CNGA3* and *MEIS1*. Except for *MEIS1*, whose upregulation was modest, these top candidates have not been shown to have direct roles in cancer development or to have cancer-related functions.

2. *ATOH1*-expressor cases compared to *ATOH1*-negative cases that are not *MYCN* amplified, and not in the novel described NB RNA subtype (Figure 4.46, right): This analysis, postulating that *ATOH1* and *MYCN* share targets, reveals further genes such as *KCTD8*, *DLL3*, *HPCAL4* (downregulated) and *HMX2*.

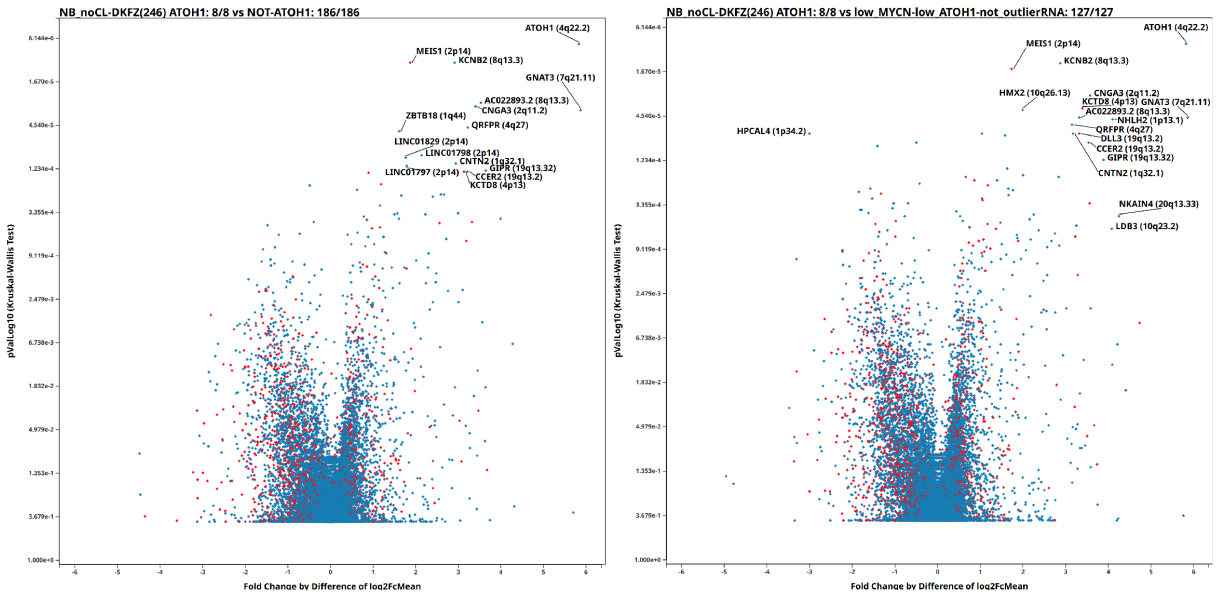

Figure 4.46: Transcriptomic changes in neuroblastoma upon *ATOH1* activation (left: *ATOH1*-positive vs *ATOH1*-negative, right: *ATOH1*-positive vs *ATOH1*-negative, AND *MYCN*-negative, AND NOT in the novel NB subtype)

We observed some of the strongest upregulation characteristics upon *ATOH1*-activation on the *GNAT3*, *AC022893.2*, *KCNB2* and *CNGA3* genes (Figure 4.47): *GNAT3* is a G-protein coupled receptor with functions in taste sensing and no prior reports of involvement in cancer development, it shows a strong and specific activation co-occurring with *ATOH1* activation. *AC022893.2* is an uncharacterized RNA gene. *KCNB2* is an understudied potassium voltage gated channel with no prior reports of involvement in cancer development. *CNGA3* is a cone photoreceptor cyclic nucleotide-gated channel which is implicated in loss of colour vision but not cancer [688].

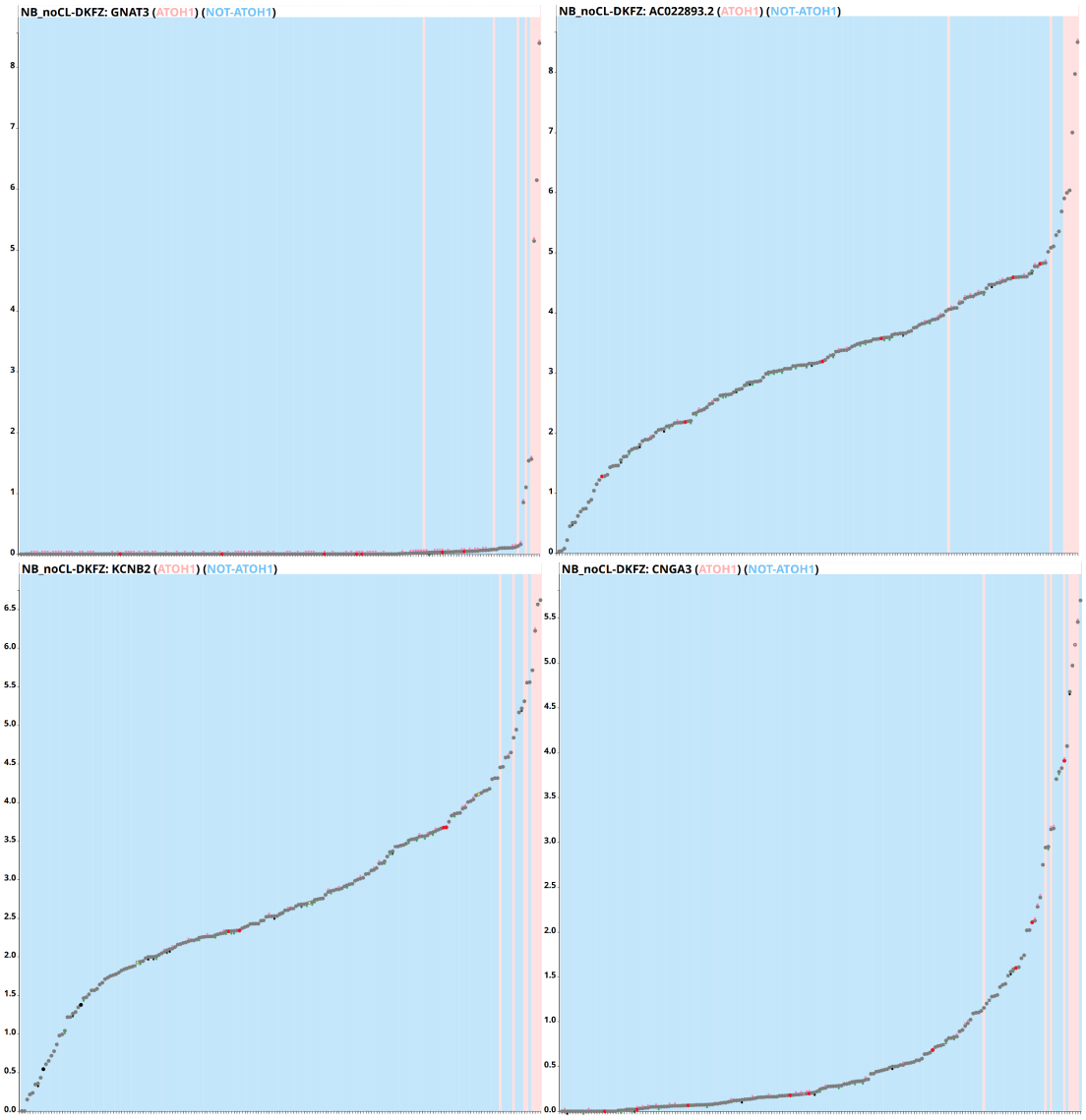

Figure 4.47: *GNAT3*, *AC022893.2*, *KCNB2* and *CNGA3* as putative targets of the *ATOH1* transcription factor. Cases with pink-shaded backgrounds are in the *ATOH1*+ group, whereas cases with blue-shaded backgrounds are in the *ATOH1*- group.

Introducing the additional constraint of *MYCN*-negative status outside of the novel NB RNA subtype revealed *NHLH2*, *KCTD8*, *DLL3* and *HPCAL4* (downregulated) as dysregulated upon *ATOH1* activation (Figure 4.48), where *ATOH1*-expressor NB cases show similar profiles to *MYCN*-amplified NB cases: *NHLH2* (formerly known as *HEN2*) has been described as a known target of *MYCN* [689], an oncogene in NB [690], and a target of *ATOH1* [691]. *KCTD8* is an understudied potassium voltage gated channel with no prior reports of involvement in cancer development. *DLL3*, *NOTCH1* ligand and a known *MYCN* target in promoting neurogenesis in brain development [692], and a recently discussed cell surface marker [693] in NB. The *DLL* family of Notch ligands are on the same pathways as *ATOH1* in cochlear sensory hair cell development [694]. *HPCAL4* is an understudied neuron-specific calcium binding protein

[695] with unknown functions in the context of cancer.

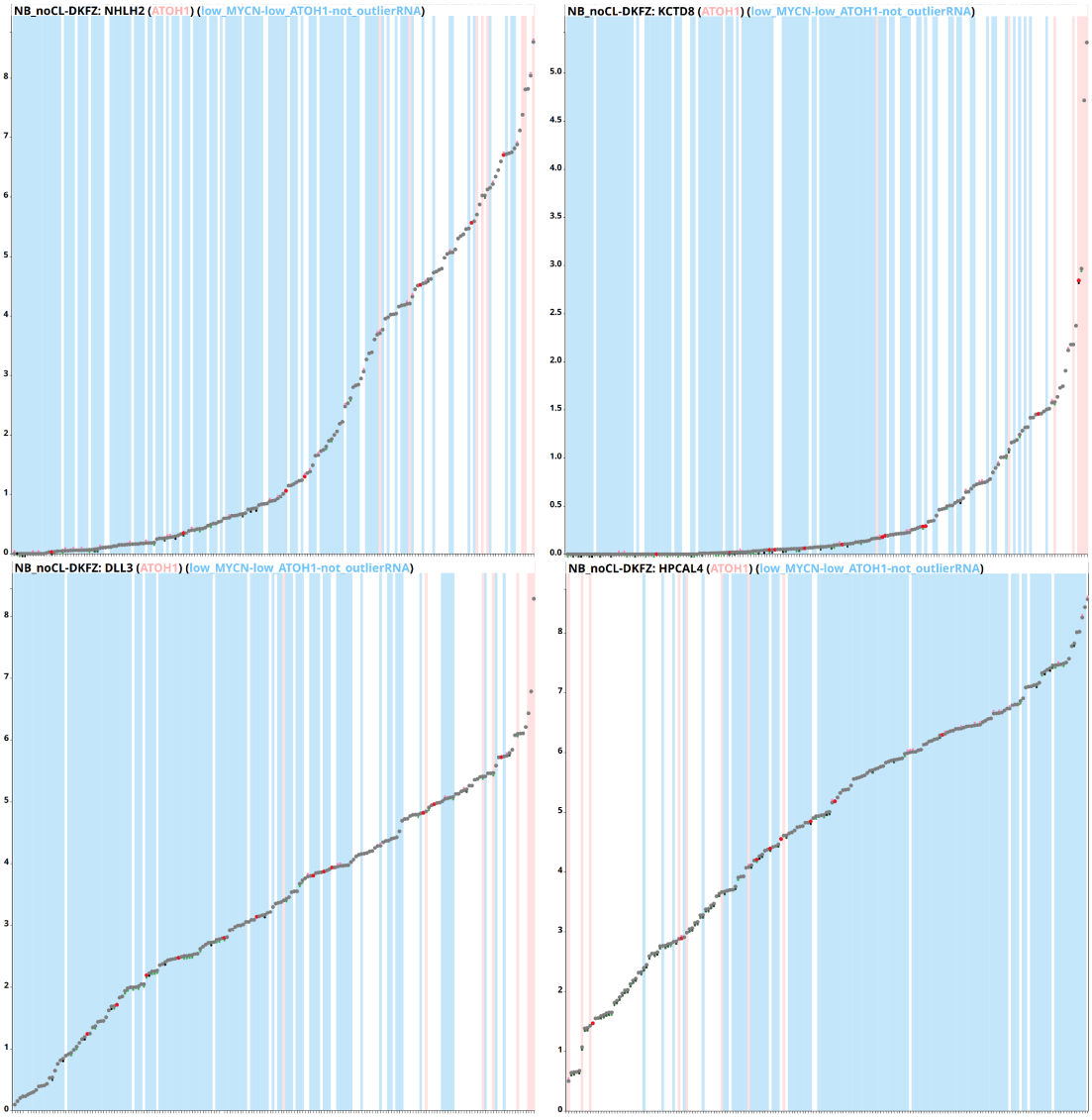

Figure 4.48: *NHLH2*, *KCTD8*, *DLL3*, and *HPCAL4* as targets of the *ATOH1* transcription factor, where *ATOH1* function potentially replaces *MYCN* amplification. *MYCN* amplified cases and cases of the novel NB RNA subtype are on white backgrounds and are not used in differential gene expression analysis. Cases with pink-shaded backgrounds are in the *ATOH1*+ group, whereas cases with blue-shaded backgrounds are in the *ATOH1*- group.

Overall the results obtained here strongly support a hypothetical role for *ATOH1* activation as a replacement mechanism for *MYCN* activation in NB. The co-upregulated genes with known functions in NB *NHLH2* (*HEN2*) and *DLL3* were also shown to be direct interactors of *ATOH1* in a large-scale screen, which remarkably also returned *NEUROD6* as an *ATOH1* interacting protein [686].

4.5.3 Discussion

We investigated the genomic variant landscape of NB and established that it has a rich SV recurrence profile, directly hitting and truncating neuronal differentiation genes. This aspect of SVs driving NB development was first explored in [678], followed by a report pre-published in March 2019 with similar findings to our broader work, [679]. A differentiation block is a crucial component of tumourigenesis of malignancies with dedifferentiated/undifferentiated/primitive cells of origin. Interestingly, all three case studies (MM, AML, NB) in this dissertation chapter shared this concept with recurrent aberrant activations of transcription factors conferring cell identity. In the context of NB, the differentiation block is not permanent, and this is used as part of the standard treatment strategy with retinoic acid [696]. In particular, in *MYCN*-nonamplified NB, the *TERT* and ALT mechanisms ensure telomere maintenance and the proliferative nature of the NB cells come from the stem cell identity of the neural crest cells. This leaves open questions regarding the origin of the differentiation block in NB, and we here postulate based on our and others' results that recurrent SVs truncating neurite differentiation genes could be contributing to the maintenance of the undifferentiated state of NB cells. Comprehensive experimental work will be needed to characterize candidate genes such as *ANKS1B*, *ZFHX3*, *DLG2*, *CNTNAP2*, *TENM3*, *AGBL3*, *PTPRD*, *SHANK2*, *DMD*, *EYS*.

We then investigated the landscape of enhancer hijacking in NB. Our comprehensive analysis with high-coverage WGS and high-quality RNA-Seq in a large cohort confirmed established genes such as *MYCN*, *TERT*, *FOXR1* and *LIN28B* and revealed only one novel candidate: the BHLH transcription *ATOH1*, with a strong activation profile. Subtle upregulation of genes such as *FGF19* and *C1Q1L* on commonly rearranged loci requires a more stringent analysis supported by lab assays such as 4-C to validate, therefore we did not focus on them in this case study. We also did not focus on sporadic cases of enhancer hijacking. In a personalized medicine setting, these might be relevant, and EPISTEME offers an easy tool for clinical bioinformaticians to recover such events.

We characterized the transcriptome and methylome subtypes of NB with a modern tSNE based manifold learning strategy, repeatedly used with great success in the paediatric brain tumour research community [218] [353]. Previous transcriptomic characterization efforts on NB were based on hierarchical clustering, revealing only the major subtypes of this heterogenous disease [675] [676]. With our large cohort, we identified finely grained subtypes of NB with distinct biological and phenotypic characteristics such as very slow growth and increased migratory potential for crossing the blood-brain barrier. Our work revealed a proposed novel subtype of *ATRX* wild-type ALT NB with the well-defined drivers *ATOH1* and *ALK*, showing a distinct age profile. This novel subtype of NB could be stemming from a more mature step in the neural crest cell lineage, and uses the *NEUROD6* BHLH TF in contrast to most other cases of NB using the *MYCN-MAX* axis. As there are currently no cell lines of this lineage, further work to characterize this new entity will be challenging. A clinical screening programme for *NEUROD6* expression in NB histopathological examination or transcriptome analysis could help prospectively identify new cases, from which new *in vitro* models can be developed.

Our analysis of *ATOH1* in the context of NB revealed a role for ATOH1 beyond its normal roles in sensory hair cell and cerebellar development [697] [698]. In the context of NB, *ATOH1*

partners with *NEUROD6* (*ATOH2*) and upregulates the transcription factor oncogene *NHLH2* (*HEN2*) [691], and the *DLL3* Notch ligand in a shared oncogenic mechanism with *MYCN* amplification [689], [692]. To the best of our knowledge, our discovery of the *ATOH1* oncogene in NB is its first example for somatic activation via genomic alterations, with previous reports discussing its overexpression without a direct genomic activation [699]. Also to the best of our knowledge, our discovery also represents the first instance of a somatic genomic activation of a Class A BHLH TF in NB. *ATOH1*-activated NB cases in the novel NB subtype were previously characterized as intermediate-risk (NB2004). With their ALT status, they would, with our current understanding, be characterized as high-risk NB and considered a clinical challenge [506]. It is our hope that with powerful *in vitro* models, the unique clinical challenge posed by the *ATOH1* gene would be better addressed.

CONCLUSION AND OUTLOOK

My doctoral research on multi-omics data integration with an especially strong focus on SV analysis from WGS data was on a subject of contemporary interest. With multi-omics data analysis strategies and method development efforts in the SOPHIA and EPISTEME projects presented here, I have had the privilege of making major or leading(*) contributions to research projects on paediatric brain tumours [218]*, meningioma [463]*, paediatric Burkitt lymphoma [435] [465], multiple myeloma (ongoing)*,acute myeloid leukaemia (ongoing)*, and paediatric neuroblastoma (ongoing)*. In these concluding remarks, I will go through the major findings presented in this dissertation and outline my expectations on how these can be further developed.

First, I discovered in a large international collaboration recurrent mutational drivers in four novel molecularly defined paediatric brain tumour entities formerly belonging to a histopathologically poorly defined group of diseased named CNS-PNETs [218]. Our strategy of first grouping the entities by cell type using the methylome, followed by candidate gene identification using the transcriptome, finally discovering the mutational drivers using the genome in a three-pronged multi-omics approach proved fruitful and impacted established clinical diagnostic practices with longer term prospects in changes of treatment strategies. Nevertheless, it must be recognized that CNS-PNETs were overall a rare entity and similar new findings will become more and more difficult with each new publication on other rare diseases such as the recent finding on infant soft tissue tumours showing *EGFR* and *BRAF* ITD events [700]. Nevertheless, the impact of such studies goes beyond epidemiological considerations and clinical interest: we and the scientific community have not yet characterized the molecular and biological function of *MN1* fusions, *CIC* fusions, *BCOR* ITDs as well as the more recently published *EGFR* and *BRAF* ITDs. Molecular functions of *FOXR2* also remain unclear apart from preliminary findings on MYC stabilization functions. As these molecularly defined entities enter clinical practice as recognized tumour entities, their biological basis will drive clinical trials based on rational targeted approaches. Further work on rare and uncharacterised tumour types may reveal novel mutational mechanisms and shed light on functions of understudied genes such as *FOXR2* or *NUTM1*.

In my second project, I developed an SV detection algorithm named SOPHIA. My priorities were fast and efficient execution and achieving high sensitivity, suitable for clinical projects. I reached these design goals and SOPHIA established itself as a part of DKFZ's standard bioinformatics workflow including its biggest personalized medicine projects in the HIPO framework [281]. In addition to its main task of SV detection, it is indirectly being used in CNV analysis and RNA fusion analysis workflows as a supporting tool. Nevertheless, more work is needed to address the need for a systematic specificity benchmark for SOPHIA and in its transition to the hg38 human reference genome. More broadly, research on SVs has significant room for improvement: there are still classes of SVs that are not accessible to short-read sequencing based technologies requiring combinations of different technologies to resolve [701]. One particularly attractive goal would be to reconstruct cancer karyotypes using a com-

bination of long-read sequencing, optical mapping and short-read sequencing, addressing the full spectrum of genomic alterations. Such a strategy is routinely used in genome assembly of different species [702], but applications in cancer remain uncommon [703], especially on primary tumour material [704]. Reconstructing a correct karyotype would allow a more precise understanding of chromatin interactions rather than relying on TADs obtained from cell lines or tissues without rearrangements. In this context, [705] is a landmark study combining optical mapping and Hi-C offering simultaneous chromatin interaction, SV calling and karyotyping, even though it might be difficult for this strategy to find wide-spread adoption in clinical diagnostics or even cancer omics research projects due to the material requirements dictating a need for *in vitro* material generation such as xenografts, organoids or cell lines. Therefore, improving SV detection from short-read sequencing data will remain as an important bioinformatics task. In this context, balancing sensitivity and specificity is the biggest challenge. To this end, a consensus approach taking advantage of different algorithms could be a rational strategy as adopted by the PCAWG consortium [284]. However, a key concern here is the general trend of local assembly based algorithms becoming more prevalent would make gains from a consensus building approach more limited. As SOPHIA is not based on local assembly, this could be turned into an advantage by pairing it with SvABA [426] or NovoBreak [425]. We are currently in the process of obtaining optical mapping (Bionano) based SV data as part of the GPOH NB genomics project and over the next years will have the opportunity to assess and improve SOPHIA both based on data from orthogonal technologies, similar to what we did with FISH, and as part of a consensus strategy with other tools.

My third project was the development of the EPISTEME integrative omics data analysis and interactive visualization tool, aiming to make high-throughput and complex cancer multi-omics datasets accessible to a broader community of users. With high-throughput sequencing methods becoming cost-effective and published datasets made freely available, we are now in a period where we have a molecular classification and analysis-based understanding of cancer and advanced, data-driven concepts of targeted treatments that are part of clinical practices. Therefore, there is a growing demand in the scientific and medical community to access cancer multi-omics datasets without the need for programming skills. Though there are a number of cancer data portals addressing this demand, EPISTEME has particularly strong features in multi-omics data integration and user interactions interconnecting the different data layers, which enables discovery and further characterization of enhancer hijacking candidates or disease molecular subtypes, as shown on pilot examples from the TCGA Chapter 3 and unpublished DKFZ datasets in Chapter 4. The next challenges for EPISTEME will be in three main directions:

- Adding support for further layers of omics data: with the cancer proteome [706] [707] and metabolome [708] [709] attracting interest, they would be powerful additions to the established features on genomics transcriptomics and methylome data analysis offered in EPISTEME.

- Increasing the technical scalability of EPISTEME: EPISTEME currently supports hundreds of cases from WGS studies. This should be further scaled up to thousands due to the requirements of pan-cancer analysis projects such as PCAWG including close to

2700 WGS samples.

- Developing features for single-cell omics data analysis: In line with anticipated improvements in input data scaling, single cell omics data will need accommodation of tens of thousands of cells with sparse data structures, and require implementation of sparse data analysis algorithms.

The first of these goals depends on the availability of data, but will also require new ways managing omics data due to the dynamic nature of metabolomics data in assays such as flux. The latter two will need technical improvements in the data management and processing implementation of EPISTEME. These include switching from SVG vector graphics to the faster WebGL technology, switching from JavaScript to Web Assembly using ”web workers” for computationally heavy algorithms. Along with ongoing work on developing epigenetics features such as visualization fof chromatin state and interaction data, these technical tasks will keep us active on further development of EPISTEME for the next years.

In the final chapter of this dissertation, I presented ongoing progress in DKFZ projects that was made possible by using SOPHIA and EPISTEME together as part of an integrative omics data analysis strategy. By only using the currently available feature set of EPISTEME and structural variant calls from SOPHIA, we managed to identify *ATOH1* as a novel candidate proto-oncogene in paediatric neuroblastoma and revealed the hitherto unstudied functions of aberrantly activated *MNX1* in acute myeloid leukaemia and *MYCN* in multiple myeloma; characterizing their putative functions in their aberrantly activated state:

- In multiple myeloma, our work could lead to dedicated studies on *MYCN*-activated multiple myeloma, studying how the *MYC*-to-*MYCN* switch alters the cellular metabolism of these tumours and if this confers a survival or proliferation advantage, given the poor survival of *MYCN*-activated multiple myeloma cases. More participants will need to be recruited in a targeted manner for further work on this subject.

- In acute myeloid leukaemia, the dual role of the *MNX1* homeobox as a transcriptomic reprogrammer activating motor neuron development genes and suppressing myeloid differentiation genes is a novel finding. Though the *MNX1* gene is of broad interest due to its known roles in multiple cancer types including infant AML, its likely molecular function in leukaemogenesis had not been recognized before our findings.

- In neuroblastoma, our work revealed a likely partial *MYCN*-replacement role for *ATOH1* as an oncogene, which could impact our understanding of BHLH transcription factors. Furthermore, we identified three novel transcriptomic clusters of NB with distinct biological characteristics. Both *ATOH1* functions as a proto-oncogene and transcriptomic classification of NB will require and attract further work.

Appendices

APPENDIX A
DONORS IN THE SOPHIA POPULATION BACKGROUND DATABASE

Hiseq family sequencers (101bp) Database

BLCA-US (TCGA) 20 Donors

TCGA-BL-A13J TCGA-BT-A20P TCGA-BT-A20Q TCGA-BT-A20T TCGA-BT-A20V TCGA-BT-A2LA TCGA-BT-A3PH TCGA-BT-A3PJ TCGA-C4-A0F7 TCGA-CF-A27C TCGA-CF-A3MF TCGA-DK-A1A5 TCGA-DK-A1A6 TCGA-DK-A1A7 TCGA-DK-A3IL TCGA-FD-A3N5 TCGA-FD-A3N6 TCGA-FT-A3EE TCGA-GD-A2C5 TCGA-H4-A2HQ

BOCA-UK (IGCC) 61 Donors

CGP_donor_1397077 CGP_donor_1397083 CGP_donor_1397084 CGP_donor_1437403 CGP_donor_1437405 CGP_donor_1437406 CGP_donor_1437407 CGP_donor_1437408 CGP_donor_1437409 CGP_donor_1437411 CGP_donor_1437412 CGP_donor_1437413 CGP_donor_1437414 CGP_donor_1437415 CGP_donor_1437416 CGP_donor_1437417 CGP_donor_1437418 CGP_donor_1437419 CGP_donor_1437420 CGP_donor_1437423 CGP_donor_1437424 CGP_donor_1437425 CGP_donor_1475256 CGP_donor_1490914 CGP_donor_1528364 CGP_donor_1528371 CGP_donor_1528374 CGP_donor_1528381 CGP_donor_1602516 CGP_donor_1602529 CGP_donor_1691121 CGP_donor_1691124 CGP_donor_1691131 CGP_donor_1691132 CGP_donor_1691133 CGP_donor_1691135 CGP_donor_1691139 CGP_donor_1691143 CGP_donor_1691145 CGP_donor_1691147 CGP_donor_1691148 CGP_donor_1691149 CGP_donor_1691150 CGP_donor_1691151 CGP_donor_1691152 CGP_donor_1691153 CGP_donor_1691154 CGP_donor_1691205 CGP_donor_1691206 CGP_donor_1691207 CGP_donor_1691208 CGP_donor_1691209 CGP_donor_1691210 CGP_donor_1691211 CGP_donor_1691212 CGP_donor_1691213 CGP_donor_1691214 CGP_donor_1691215 CGP_donor_1691216 CGP_donor_1691217 CGP_donor_1841267

BRCA-EU (IGCC) 75 Donors

CGP_donor_1163904 CGP_donor_1186987 CGP_donor_1186990 CGP_donor_1187025 CGP_donor_1230754 CGP_donor_1230755 CGP_donor_1230796 CGP_donor_1230797 CGP_donor_1232859 CGP_donor_1234120 CGP_donor_1234121 CGP_donor_1234122 CGP_donor_1234123 CGP_donor_1234124 CGP_donor_1234129 CGP_donor_1333047 CGP_donor_1333048 CGP_donor_1337214 CGP_donor_1337217 CGP_donor_1337218 CGP_donor_1337220 CGP_donor_1337222 CGP_donor_1337223 CGP_donor_1337225 CGP_donor_1337226 CGP_donor_1337231 CGP_donor_1337236 CGP_donor_1337237 CGP_donor_1337238 CGP_donor_1337240 CGP_donor_1337241 CGP_donor_1347723 CGP_donor_1347731 CGP_donor_1347742 CGP_donor_1347751 CGP_donor_1347756 CGP_donor_1353426 CGP_donor_1353427 CGP_donor_1353428 CGP_donor_1353429 CGP_donor_1353434 CGP_donor_1364028 CGP_donor_1364029 CGP_donor_1364033 CGP_donor_1374617 CGP_donor_1374618 CGP_donor_1397086 CGP_donor_1397088 CGP_donor_1397260 CGP_donor_1397261 CGP_donor_1397262 CGP_donor_1397263 CGP_donor_1397264 CGP_donor_1397266 CGP_donor_1397277 CGP_donor_1397278 CGP_donor_1397279 CGP_donor_1397281 CGP_donor_1397282 CGP_donor_1397284 CGP_donor_1451422 CGP_donor_1451426 CGP_donor_1451427 CGP_donor_1475201 CGP_donor_1475202 CGP_donor_1503014 CGP_donor_1503016 CGP_donor_1503017 CGP_donor_1503019 CGP_donor_1503020 CGP_donor_1503021 CGP_donor_1503128 CGP_donor_1503140 CGP_donor_1503150 CGP_donor_1503156

BRCA-UK (IGCC) 39 Donors

CGP_donor_1069291 CGP_donor_1114881 CGP_donor_1114929 CGP_donor_1167078 CGP_donor_1167080 CGP_donor_1187030 CGP_donor_1187031 CGP_donor_1187033 CGP_donor_1199129 CGP_donor_1199137 CGP_donor_1199138 CGP_donor_1212361 CGP_donor_1230722 CGP_donor_1230724 CGP_donor_1230728 CGP_donor_1230729 CGP_donor_1230785 CGP_donor_1309223 CGP_donor_1310131 CGP_donor_1337235 CGP_donor_1347720 CGP_donor_1347737 CGP_donor_1347739 CGP_donor_1347745 CGP_donor_1347813 CGP_donor_1353431 CGP_donor_1353432 CGP_donor_1363963 CGP_donor_1363965 CGP_donor_1363969 CGP_donor_1410205 CGP_donor_1410210 CGP_donor_1456607 CGP_donor_1472394 CGP_donor_1472395 CGP_donor_1503143 CGP_donor_1606179 CGP_donor_1654385 CGP_donor_1701345

BRCA-US (TCGA) 88 Donors

TCGA-A1-A0SM TCGA-A2-A04P TCGA-A2-A04T TCGA-A2-A04X TCGA-A2-A0D0 TCGA-A2-A0D1 TCGA-A2-A0D4 TCGA-A2-A0EY TCGA-A2-A0YG TCGA-A2-A259 TCGA-A2-A25B TCGA-A2-A3KC TCGA-A2-A3XX TCGA-A2-A3Y0 TCGA-A7-A0CE TCGA-A7-A13D TCGA-A7-A26G TCGA-A8-A075 TCGA-A8-A07B TCGA-A8-A07I TCGA-A8-A08B TCGA-A8-A08L TCGA-A8-A08S TCGA-A8-A092 TCGA-A8-A094 TCGA-A8-A09X TCGA-AC-A2BK TCGA-AN-A04D TCGA-AN-A0AT TCGA-AN-A0G0 TCGA-AN-A0XR TCGA-AO-A03L TCGA-AO-A03N TCGA-AO-A0J2 TCGA-AO-A0J4 TCGA-AO-A0J6 TCGA-AO-A0JM TCGA-AO-A124 TCGA-AO-A12H TCGA-AQ-A04J TCGA-AR-A0TX TCGA-AR-A1AY TCGA-AR-A24Z TCGA-AR-A256 TCGA-AR-A2LK TCGA-B6-A0I1 TCGA-B6-A0I6 TCGA-B6-A0RT TCGA-B6-A0RU TCGA-B6-A0WX TCGA-B6-A0X5 TCGA-BH-A0AV TCGA-BH-A0BW TCGA-BH-A0DG TCGA-BH-A0DT TCGA-BH-A0E0 TCGA-BH-A0H0 TCGA-BH-A0H6 TCGA-BH-A0WA TCGA-BH-A18R TCGA-BH-A18U TCGA-BH-A1FC TCGA-C8-A12L TCGA-C8-A12Q TCGA-C8-A130 TCGA-D8-A27F TCGA-D8-A27H TCGA-E2-A109 TCGA-E2-A14P TCGA-E2-A14X TCGA-E2-A152 TCGA-E2-A156 TCGA-E2-A15E TCGA-E2-A15H TCGA-E2-A15K TCGA-E2-A1LG TCGA-E2-A1LK TCGA-E2-A1LL TCGA-E9-A1NH TCGA-EW-A1J5 TCGA-EW-A1P8 TCGA-EW-A1PB TCGA-EW-A1PC TCGA-EW-A1PH TCGA-EW-A3U0 TCGA-GI-A2C9 TCGA-GM-A2DF TCGA-GM-A3XL

BTCA-SG (ICGC) 12 Donors

BTCA_donor_27 BTCA_donor_A035 BTCA_donor_A096 BTCA_donor_A153 BTCA_donor_B070 BTCA_donor_B083

BTCA_donor_C080 BTCA_donor_R149 BTCA_donor_Y008 BTCA_donor_Y065 BTCA_donor_Y140 BTCA_donor_Z2403

CESC-US (TCGA) 16 Donors

TCGA-C5-A0TN TCGA-C5-A1BF TCGA-C5-A1BN TCGA-C5-A1M9 TCGA-C5-A1MI TCGA-C5-A1ML TCGA-C5-A1MQ TCGA-C5-A2LT TCGA-C5-A2LV TCGA-C5-A2LY TCGA-DG-A2KJ TCGA-DS-A0VL TCGA-EK-A2PK TCGA-EK-A2R9 TCGA-EK-A2RM TCGA-EX-A1H5

CLLE-ES (ICGC) 95 Donors

10 115 122 125 128 12 134 137 138 139 141 145 148 151 157 15 166 16 176 177 179 181 188 192 199 1 20 23 244 25 26 277 278 27 282 283 290 296 2 306 308 30 318 32 33 342 343 356 358 367 371 372 386 393 39 3 435 442 44 467 473 477 48 4 519 523 56 577 58 594 5 628 63 64 654 661 677 684 6 723 749 761 776 783 785 795 802 803 824 82 832 83 84 90 9

CMDI-UK (IGCC) 15 Donors

CGP_donor_1605283_1 CGP_donor_1338575_1 CGP_donor_1364401_1 CGP_donor_1364416_1 CGP_donor_1364444_1 CGP_donor_1364465_1 CGP_donor_1364470_1 CGP_donor_1405305_1 CGP_donor_1405314_1 CGP_donor_1463315_1 CGP_donor_1500976_1 CGP_donor_1500978_1 CGP_donor_1600817_1 CGP_donor_1600818_1 CGP_donor_1733896_1

COAD-US (TCGA) 40 Donors

TCGA-A6-2680 TCGA-A6-2681 TCGA-A6-3807 TCGA-A6-6141 TCGA-A6-6781 TCGA-A6-A565 TCGA-A6-A566 TCGA-A6-A567 TCGA-A6-A56B TCGA-AA-3514 TCGA-AA-3518 TCGA-AA-3529 TCGA-AA-3534 TCGA-AA-3555 TCGA-AA-3664 TCGA-AA-3666 TCGA-AA-3685 TCGA-AA-3956 TCGA-AA-3977 TCGA-AA-3994 TCGA-AA-A01S TCGA-AA-A01T TCGA-AA-A01V TCGA-AA-A01X TCGA-AA-A02O TCGA-AA-A02Y TCGA-AD-6964 TCGA-AD-A5EJ TCGA-AD-A5EK TCGA-AY-A54L TCGA-CA-6717 TCGA-CA-6718 TCGA-D5-6540 TCGA-NH-A50T TCGA-NH-A50V TCGA-QG-A5YV TCGA-QG-A5YW TCGA-QG-A5YX TCGA-QG-A5Z1 TCGA-QG-A5Z2

DLBC-US (TCGA) 7 Donors

TCGA-FF-8041 TCGA-FF-8042 TCGA-FF-8043 TCGA-FF-8046 TCGA-FF-8047 TCGA-FF-8061 TCGA-FF-8062

EOPC-DE (ICGC) 40 Donors

EOPC-010 EOPC-011 EOPC-017 EOPC-018 EOPC-019 EOPC-01 EOPC-021 EOPC-022 EOPC-023 EOPC-024 EOPC-025 EOPC-026 EOPC-029 EOPC-02 EOPC-030 EOPC-031 EOPC-032 EOPC-033 EOPC-034_1 EOPC-035_1 EOPC-036_1 EOPC-037_1 EOPC-03 EOPC-040_1 EOPC-041 EOPC-048 EOPC-049 EOPC-04 EOPC-051 EOPC-052 EOPC-053 EOPC-054 EOPC-056 EOPC-057 EOPC-058 EOPC-05 EOPC-06 EOPC-07 EOPC-08 EOPC-09

ESAD-UK (IGCC) 98 Donors

OCCAMS-AH-011 OCCAMS-AH-014 OCCAMS-AH-021 OCCAMS-AH-036 OCCAMS-AH-039 OCCAMS-AH-042 OCCAMS-AH-046 OCCAMS-AH-047 OCCAMS-AH-048 OCCAMS-AH-061 OCCAMS-AH-062 OCCAMS-AH-063 OCCAMS-AH-064 OCCAMS-AH-071 OCCAMS-AH-077 OCCAMS-AH-082 OCCAMS-AH-085 OCCAMS-AH-086 OCCAMS-AH-088 OCCAMS-AH-091 OCCAMS-AH-096 OCCAMS-AH-108 OCCAMS-AH-112 OCCAMS-AH-120 OCCAMS-AH-127 OCCAMS-AH-131 OCCAMS-AH-133 OCCAMS-AH-135 OCCAMS-AH-136 OCCAMS-AH-139 OCCAMS-AH-140 OCCAMS-AH-143 OCCAMS-AH-146 OCCAMS-AH-155 OCCAMS-AH-160 OCCAMS-AH-167 OCCAMS-AH-173 OCCAMS-AH-174 OCCAMS-AH-182 OCCAMS-AH-183 OCCAMS-AH-196 OCCAMS-AH-197 OCCAMS-AH-213 OCCAMS-ED-003 OCCAMS-ED-007 OCCAMS-ED-036 OCCAMS-ED-041 OCCAMS-GS-002 OCCAMS-PS-001 OCCAMS-PS-002 OCCAMS-PS-008 OCCAMS-PS-012 OCCAMS-PS-013 OCCAMS-PS-014 OCCAMS-QE-095 OCCAMS-RS-006 OCCAMS-RS-007 OCCAMS-RS-008 OCCAMS-RS-010 OCCAMS-RS-014 OCCAMS-RS-022 OCCAMS-RS-024 OCCAMS-RS-027 OCCAMS-RS-028 OCCAMS-RS-029 OCCAMS-RS-031 OCCAMS-RS-032 OCCAMS-RS-035 OCCAMS-RS-036 OCCAMS-RS-047 OCCAMS-SH-003 OCCAMS-SH-020 OCCAMS-SH-024 OCCAMS-SH-038 OCCAMS-SH-051 OCCAMS-SH-071 OCCAMS-ST-020 OCCAMS-ST-023 OCCAMS-ST-029 OCCAMS-ST-030 OCCAMS-ST-033 OCCAMS-ST-035 OCCAMS-ST-036 OCCAMS-ST-037 OCCAMS-ST-041 OCCAMS-ST-043 OCCAMS-WG-001 OCCAMS-WG-002 OCCAMS-WG-005 OCCAMS-WG-006 OCCAMS-WG-008 OCCAMS-WG-009 OCCAMS-WG-019 OCCAMS-ZZ-004 OCCAMS-ZZ-009 OCCAMS-ZZ-011 OCCAMS-ZZ-016 OCCAMS-ZZ-019

GACA-CN 39 Donors

CGP_donor_GC00001 CGP_donor_GC00002 CGP_donor_GC00003 CGP_donor_GC00004 CGP_donor_GC00005 CGP_donor_GC00007 CGP_donor_GC00008 CGP_donor_GC00013 CGP_donor_GC00014 CGP_donor_GC00015 CGP_donor_GC00016 CGP_donor_GC00017 CGP_donor_GC00018 CGP_donor_GC00019 CGP_donor_GC00020 CGP_donor_GC00021 CGP_donor_GC00022 CGP_donor_GC00026 CGP_donor_GC00027 CGP_donor_GC00028 CGP_donor_GC00029 CGP_donor_GC00030 CGP_donor_GC00031 CGP_donor_GC00032 CGP_donor_GC00034 CGP_donor_GC00035 CGP_donor_GC00037 CGP_donor_GC00038 CGP_donor_GC00039 CGP_donor_GC00040 CGP_donor_GC00046 CGP_donor_GC00047 CGP_donor_GC00048 CGP_donor_GC00049 CGP_donor_GC00050 CGP_donor_GC00051 CGP_donor_GC00052 CGP_donor_GC00053 CGP_donor_GC00054

GBM-US (TCGA) 32 Donors

TCGA-02-2483 TCGA-02-2485 TCGA-06-0155 TCGA-06-0157 TCGA-06-0190 TCGA-06-0211 TCGA-06-0214 TCGA-06-0221 TCGA-06-0686 TCGA-06-0744 TCGA-06-0745 TCGA-06-1086 TCGA-06-2557 TCGA-06-2570 TCGA-06-5411 TCGA-06-5415 TCGA-14-0786 TCGA-14-1402 TCGA-14-1823 TCGA-14-2554 TCGA-16-1063 TCGA-19-1389 TCGA-19-2620 TCGA-19-2624 TCGA-19-2629 TCGA-19-5960 TCGA-26-5132 TCGA-26-5135

TCGA-27-1831 TCGA-27-2523 TCGA-27-2528 TCGA-32-1970

HNSC-US (TCGA) 42 Donors

TCGA-BA-4076 TCGA-BA-5149 TCGA-BA-5556 TCGA-BA-6869 TCGA-BA-6872 TCGA-BA-6873 TCGA-BA-A4IH TCGA-CN-4737 TCGA-CN-5365 TCGA-CN-5374 TCGA-CN-6011 TCGA-CN-6989 TCGA-CN-6994 TCGA-CQ-6228 TCGA-CR-5249 TCGA-CR-5250 TCGA-CR-6467 TCGA-CR-6470 TCGA-CR-6472 TCGA-CR-6480 TCGA-CR-6482 TCGA-CR-6487 TCGA-CR-6491 TCGA-CR-7382 TCGA-CR-7385 TCGA-CR-7391 TCGA-CR-7404 TCGA-CV-5431 TCGA-CV-5432 TCGA-CV-5442 TCGA-CV-5443 TCGA-CV-5973 TCGA-CV-6433 TCGA-CV-6956 TCGA-CV-7090 TCGA-CV-7100 TCGA-CV-7180 TCGA-CV-7255 TCGA-CV-7432 TCGA-CX-7086 TCGA-DQ-5625 TCGA-HD-7753

KICH-US (TCGA) 43 Donors

TCGA-KL-8323 TCGA-KL-8325 TCGA-KL-8326 TCGA-KL-8328 TCGA-KL-8330 TCGA-KL-8331 TCGA-KL-8332 TCGA-KL-8333 TCGA-KL-8334 TCGA-KL-8340 TCGA-KL-8341 TCGA-KL-8342 TCGA-KL-8343 TCGA-KL-8344 TCGA-KL-8346 TCGA-KM-8438 TCGA-KM-8439 TCGA-KM-8440 TCGA-KM-8441 TCGA-KM-8442 TCGA-KM-8443 TCGA-KM-8476 TCGA-KM-8477 TCGA-KM-8639 TCGA-KN-8418 TCGA-KN-8419 TCGA-KN-8421 TCGA-KN-8422 TCGA-KN-8424 TCGA-KN-8425 TCGA-KN-8426 TCGA-KN-8427 TCGA-KN-8428 TCGA-KN-8429 TCGA-KN-8431 TCGA-KN-8432 TCGA-KN-8434 TCGA-KN-8435 TCGA-KN-8437 TCGA-KO-8405 TCGA-KO-8406 TCGA-KO-8407 TCGA-KO-8411

KIRC-US (TCGA) 32 Donors

TCGA-A3-3308 TCGA-A3-3363 TCGA-A3-3372 TCGA-A3-3387 TCGA-AK-3454 TCGA-AK-3455 TCGA-B0-5094 TCGA-B0-5693 TCGA-B0-5695 TCGA-B2-4099 TCGA-B2-4101 TCGA-B2-4102 TCGA-BP-4326 TCGA-BP-4756 TCGA-BP-4807 TCGA-BP-4968 TCGA-BP-5010 TCGA-BP-5168 TCGA-CJ-4639 TCGA-CJ-4870 TCGA-CJ-4878 TCGA-CJ-4899 TCGA-CJ-5681 TCGA-CJ-5682 TCGA-CJ-6033 TCGA-CW-5585 TCGA-CW-6087 TCGA-CW-6093 TCGA-CZ-5453 TCGA-CZ-5454 TCGA-CZ-5987 TCGA-DV-5566

KIRP-US (TCGA) 33 Donors

TCGA-A4-A48D TCGA-A4-A4ZT TCGA-A4-A57E TCGA-AL-3466 TCGA-AL-3468 TCGA-AL-3472 TCGA-AL-3473 TCGA-AL-A5DJ TCGA-B1-A47M TCGA-B1-A47N TCGA-B1-A47O TCGA-B3-3925 TCGA-B3-3926 TCGA-B9-4113 TCGA-B9-4114 TCGA-B9-4115 TCGA-B9-4116 TCGA-B9-4117 TCGA-B9-4617 TCGA-B9-A44B TCGA-GL-A4EM TCGA-GL-A59R TCGA-HE-A5NF TCGA-HE-A5NH TCGA-HE-A5NJ TCGA-HE-A5NL TCGA-IA-A40X TCGA-IA-A40Y TCGA-MH-A55W TCGA-MH-A55Z TCGA-MH-A560 TCGA-MH-A561 TCGA-MH-A562

LAML-KR 5 Donors

SNU_WGS_01 SNU_WGS_05 SNU_WGS_09 SNU_WGS_10 SNU_WGS_12

LAML-US (TCGA) 9 Donors

TCGA-AB-2906 TCGA-AB-2976 TCGA-AB-2980 TCGA-AB-2983 TCGA-AB-2987 TCGA-AB-2989 TCGA-AB-2991 TCGA-AB-2993 TCGA-AB-2998

LGG-US (TCGA) 17 Donors

TCGA-CS-5395 TCGA-CS-6668 TCGA-DB-5278 TCGA-DU-5874 TCGA-DU-6401 TCGA-DU-6407 TCGA-DU-7009 TCGA-DU-7301 TCGA-E1-5318 TCGA-E1-5319 TCGA-EZ-7264 TCGA-FG-5964 TCGA-FG-8182 TCGA-HT-7602 TCGA-HT-7695 TCGA-HW-7487 TCGA-IK-7675

LICA-FR 4 Donors

CHC205 CHC320 CHC322 CHC433

LIHC-US (TCGA) 53 Donors

TCGA-BC-A10Q TCGA-BC-A216 TCGA-BC-A217 TCGA-BW-A5NO TCGA-BW-A5NP TCGA-BW-A5NQ TCGA-CC-5260 TCGA-CC-5261 TCGA-CC-5262 TCGA-CC-A1HT TCGA-DD-A1E9 TCGA-DD-A1EB TCGA-DD-A1ED TCGA-DD-A1EG TCGA-DD-A1EH TCGA-DD-A1EI TCGA-DD-A1EJ TCGA-DD-A1EL TCGA-DD-A3A6 TCGA-DD-A3A7 TCGA-DD-A3A8 TCGA-DD-A3A9 TCGA-DD-A4NA TCGA-DD-A4NB TCGA-DD-A4ND TCGA-DD-A4NE TCGA-DD-A4NG TCGA-ED-A459 TCGA-ED-A4XI TCGA-EP-A26S TCGA-EP-A2KA TCGA-EP-A2KB TCGA-EP-A3RK TCGA-ES-A2HS TCGA-ES-A2HT TCGA-FV-A23B TCGA-FV-A2QQ TCGA-FV-A3I0 TCGA-FV-A3I1 TCGA-FV-A3R2 TCGA-FV-A3R3 TCGA-FV-A495 TCGA-FV-A496 TCGA-FV-A4ZQ TCGA-G3-A25S TCGA-G3-A25T TCGA-G3-A25V TCGA-G3-A25W TCGA-G3-A25Y TCGA-G3-A3CK TCGA-HP-A5MZ TCGA-MR-A520 TCGA-PD-A5DF

LINC-JP 21 Donors

HX18 HX20 HX21 HX22 HX23 HX24 HX25 HX26 HX27 HX28 HX29 HX30 HX31 HX32 HX33 HX34 HX35 HX36 HX37 HX5 HX9

LIRI-JP 218 Donors

RK001 RK002 RK003 RK004 RK005 RK007 RK010 RK012 RK014 RK015 RK016 RK018 RK019 RK020 RK021 RK022 RK023 RK024 RK026 RK027 RK028 RK029 RK030 RK032 RK033 RK034 RK035 RK037 RK038 RK041 RK042 RK043 RK044 RK048 RK049 RK051 RK052 RK053 RK054 RK055 RK056 RK057 RK058 RK059 RK060 RK061 RK062 RK063 RK064 RK065 RK066 RK067 RK068 RK070 RK071 RK072 RK073 RK074 RK075 RK076 RK077 RK079 RK080 RK081 RK082 RK083 RK084 RK085 RK086 RK088 RK089 RK090 RK092 RK093 RK095 RK096 RK098 RK100 RK102 RK104 RK105 RK106 RK107 RK111 RK113 RK120 RK121 RK122 RK124 RK125 RK126 RK128 RK130 RK133 RK135 RK136 RK137 RK139 RK140 RK142 RK143 RK144 RK146 RK148 RK150 RK151 RK152 RK153 RK154 RK155 RK156 RK163

RK164 RK166 RK167 RK169 RK170 RK172 RK175 RK176 RK177 RK178 RK181 RK183 RK184 RK185 RK186 RK188 RK190 RK191 RK193 RK194 RK195 RK196 RK197 RK198 RK199 RK200 RK201 RK202 RK205 RK206 RK207 RK208 RK210 RK211 RK213 RK215 RK216 RK217 RK220 RK221 RK222 RK223 RK224 RK225 RK226 RK227 RK228 RK229 RK230 RK232 RK234 RK235 RK236 RK237 RK241 RK243 RK244 RK245 RK254 RK256 RK257 RK258 RK259 RK260 RK262 RK263 RK264 RK265 RK266 RK267 RK268 RK269 RK270 RK272 RK275 RK277 RK278 RK279 RK280 RK282 RK284 RK285 RK297 RK298 RK303 RK304 RK305 RK306 RK307 RK308 RK309 RK310 RK312 RK316 RK317 RK326 RK337 RK338 RK006_1 RK036_1 RK046_1 RK145_1 RK180_1 RK261_1 RK287_1 RK289_1

LUAD-US (TCGA) 34 Donors

TCGA-05-4389 TCGA-05-4395 TCGA-05-4396 TCGA-05-4397 TCGA-05-4398 TCGA-05-4420 TCGA-05-5429 TCGA-38-4628 TCGA-44-2659 TCGA-44-6148 TCGA-49-4486 TCGA-49-4512 TCGA-49-6742 TCGA-50-5066 TCGA-50-5930 TCGA-50-5932 TCGA-50-6591 TCGA-50-6597 TCGA-55-6972 TCGA-55-6982 TCGA-55-6984 TCGA-55-6986 TCGA-55-7281 TCGA-55-8299 TCGA-64-1678 TCGA-64-1680 TCGA-73-4659 TCGA-75-5147 TCGA-75-6203 TCGA-75-7030 TCGA-78-7158 TCGA-91-6840 TCGA-91-6847 TCGA-97-8171

LUSC-US (TCGA) 35 Donors

TCGA-18-3408 TCGA-18-3415 TCGA-18-4721 TCGA-21-5782 TCGA-22-5477 TCGA-22-5485 TCGA-22-5492 TCGA-33-4586 TCGA-34-2596 TCGA-34-2600 TCGA-34-5240 TCGA-37-4135 TCGA-43-3394 TCGA-43-3920 TCGA-43-5670 TCGA-52-7812 TCGA-56-1622 TCGA-56-7582 TCGA-60-2698 TCGA-60-2711 TCGA-60-2719 TCGA-66-2756 TCGA-66-2789 TCGA-66-2793 TCGA-66-2795 TCGA-68-7755 TCGA-68-8250 TCGA-77-6843 TCGA-77-7139 TCGA-85-8052 TCGA-85-8277 TCGA-92-8064 TCGA-94-7943 TCGA-96-7545 TCGA-98-8022

MB-DKFZ (ICGC) 201 Donors

ICGC_LFS_MB1 ICGC_MB101 ICGC_MB102 ICGC_MB104 ICGC_MB106 ICGC_MB108 ICGC_MB110 ICGC_MB111 ICGC_MB112 ICGC_MB113 ICGC_MB114 ICGC_MB115 ICGC_MB117 ICGC_MB118 ICGC_MB119 ICGC_MB121 ICGC_MB122 ICGC_MB124 ICGC_MB125 ICGC_MB126 ICGC_MB128 ICGC_MB129 ICGC_MB12 ICGC_MB130 ICGC_MB131 ICGC_MB132 ICGC_MB134 ICGC_MB136 ICGC_MB139 ICGC_MB140 ICGC_MB141 ICGC_MB144 ICGC_MB145 ICGC_MB146 ICGC_MB151 ICGC_MB152 ICGC_MB154 ICGC_MB157 ICGC_MB159 ICGC_MB15 ICGC_MB160 ICGC_MB161 ICGC_MB163 ICGC_MB164 ICGC_MB165 ICGC_MB166 ICGC_MB168 ICGC_MB16 ICGC_MB170 ICGC_MB171 ICGC_MB174 ICGC_MB175 ICGC_MB176 ICGC_MB177 ICGC_MB178 ICGC_MB179 ICGC_MB180 ICGC_MB181 ICGC_MB183 ICGC_MB184 ICGC_MB185 ICGC_MB188 ICGC_MB189 ICGC_MB18 ICGC_MB193 ICGC_MB194 ICGC_MB198 ICGC_MB199 ICGC_MB19 ICGC_MB1 ICGC_MB204 ICGC_MB205 ICGC_MB206 ICGC_MB20 ICGC_MB213 ICGC_MB214 ICGC_MB216 ICGC_MB217 ICGC_MB21 ICGC_MB224 ICGC_MB225 ICGC_MB226 ICGC_MB227 ICGC_MB228 ICGC_MB229 ICGC_MB230 ICGC_MB232 ICGC_MB233 ICGC_MB234 ICGC_MB235 ICGC_MB236 ICGC_MB237 ICGC_MB239 ICGC_MB23 ICGC_MB240 ICGC_MB241 ICGC_MB242 ICGC_MB243 ICGC_MB244 ICGC_MB246 ICGC_MB247 ICGC_MB248 ICGC_MB249 ICGC_MB24 ICGC_MB250 ICGC_MB256 ICGC_MB260 ICGC_MB261 ICGC_MB262 ICGC_MB264 ICGC_MB265 ICGC_MB266 ICGC_MB268 ICGC_MB269 ICGC_MB26 ICGC_MB270 ICGC_MB272 ICGC_MB274 ICGC_MB275 ICGC_MB276 ICGC_MB277 ICGC_MB278 ICGC_MB279 ICGC_MB280 ICGC_MB281 ICGC_MB282 ICGC_MB284 ICGC_MB285 ICGC_MB286 ICGC_MB287 ICGC_MB288 ICGC_MB289 ICGC_MB28 ICGC_MB290 ICGC_MB291 ICGC_MB292 ICGC_MB295 ICGC_MB297 ICGC_MB299 ICGC_MB2 ICGC_MB302 ICGC_MB307 ICGC_MB31 ICGC_MB32 ICGC_MB35 ICGC_MB36 ICGC_MB37 ICGC_MB39 ICGC_MB3 ICGC_MB40 ICGC_MB45 ICGC_MB46 ICGC_MB49 ICGC_MB50 ICGC_MB518 ICGC_MB51 ICGC_MB53 ICGC_MB54 ICGC_MB56 ICGC_MB57 ICGC_MB58 ICGC_MB59 ICGC_MB5 ICGC_MB60 ICGC_MB612 ICGC_MB61 ICGC_MB62 ICGC_MB63 ICGC_MB64 ICGC_MB66 ICGC_MB6 ICGC_MB75 ICGC_MB76 ICGC_MB78 ICGC_MB7 ICGC_MB800 ICGC_MB81 ICGC_MB82 ICGC_MB83 ICGC_MB84 ICGC_MB85 ICGC_MB86 ICGC_MB88 ICGC_MB89 ICGC_MB8 ICGC_MB90 ICGC_MB91 ICGC_MB92 ICGC_MB94 ICGC_MB95 ICGC_MB96 ICGC_MB98 ICGC_MB99 ICGC_MB9 MBRep_T27 MBRep_T36 MBRep_T40 MBRep_T41 MBRep_T54 MBRep_T70 MBRep_T79

MELA-AU (IGCC) 66 Donors

MELA-0001 MELA-0002 MELA-0003 MELA-0005 MELA-0007 MELA-0008 MELA-0009 MELA-0011 MELA-0012 MELA-0015 MELA-0022 MELA-0034 MELA-0037 MELA-0043 MELA-0046 MELA-0048 MELA-0050 MELA-0051 MELA-0053 MELA-0055 MELA-0056 MELA-0060 MELA-0061 MELA-0064 MELA-0066 MELA-0067 MELA-0069 MELA-0070 MELA-0075 MELA-0076 MELA-0160 MELA-0161 MELA-0167 MELA-0168 MELA-0169 MELA-0170 MELA-0173 MELA-0174 MELA-0179 MELA-0180 MELA-0183 MELA-0184 MELA-0185 MELA-0187 MELA-0190 MELA-0192 MELA-0193 MELA-0196 MELA-0197 MELA-0200 MELA-0202 MELA-0203 MELA-0205 MELA-0213 MELA-0228 MELA-0229 MELA-0230 MELA-0231 MELA-0234 MELA-0236 MELA-0237 MELA-0238 MELA-0239 MELA-0256 MELA-0257 MELA-0259

MMML-DKFZ (ICGC) 196 Donors

4100314 4100636 4101316 4101626 4101815 4102009 4103141 4103434 4103570 4104105 4104893 4105105 4105746 4107137 4107559 4107990 4108101 4108588 4108988 4108992 4109142 4109808 4109956 4110120 4110378 4110498 4110996 4111326 4111337 4112447 4112512 4113140 4113191 4113211 4113825 4113971 4115001 4116268 4116738 4117030 4119027 4119279 4119463 4120157 4120193 4120468 4121263 4121361 4122063 4123945 4124188 4124542 4124795 4125240 4126473 4127766 4128477 4128852 4128970 4130003 4130051 4130194 4131095 4131257 4131738 4131744 4132318 4132950 4133263 4133511 4133863 4134005 4134434 4135099 4135278 4135350 4136702 4138527 4138652 4138885 4139696 4140531 4141476 4142267

4142605 4142761 4144131 4144366 4144633 4144951 4145056 4145177 4145391 4145528 4146136 4146289
4147081 4147968 4148261 4148771 4149246 4150895 4151028 4152036 4156551 4157186 4158268 4158483
4158726 4158769 4158933 4159170 4160069 4160100 4160468 4160810 4161696 4161781 4162154 4162611
4163297 4163639 4163741 4164943 4165379 4166151 4166503 4166706 4166940 4167381 4168738 4169012
4170577 4170686 4170844 4171706 4171810 4172511 4173863 4174742 4174905 4175837 4176133 4176325
4177175 4177376 4177406 4177434 4177601 4177639 4177810 4177856 4177987 4178310 4178345 4178518
4178655 4179894 4180106 4181037 4181460 4182393 4183136 4183924 4184011 4184094 4187640 4188398
4188800 4188879 4188900 4189035 4189200 4189998 4190231 4190316 4190495 4190784 4191799 4192483
4193278 4193435 4193638 4193646 4194218 4194891 4196654 4196670 4197155 4197438 4198478 4198519
4198542 4199714 4199848 4199996

NB-DKFZ (GPOH) 33 Donors

B087koeln_13264 B087koeln_15239 B087koeln_15303 B087koeln_15403 B087koeln_16885 B087koeln_1695 B087koeln_17240
B087koeln_17344 B087koeln_17612 B087koeln_17683 B087koeln_17861 B087koeln_18478 B087koeln_18699 B087koeln_18728
B087koeln_18972 B087koeln_19537 B087koeln_19751 B087koeln_19885 B087koeln_20471 B087koeln_20507 B087koeln_20807
B087koeln_20865 B087koeln_20920 B087koeln_21368 B087koeln_21442 B087koeln_21641 B087koeln_21776 B087koeln_21924
B087koeln_23067 B087koeln_23122 B087koeln_23229 B087wgs_13169 B087wgs_18253

ORCA-IN (ICGC) 13 Donors

OSCC-GB_011301 OSCC-GB_011601 OSCC-GB_011701 OSCC-GB_011801 OSCC-GB_011901 OSCC-GB_012001
OSCC-GB_012101 OSCC-GB_012201 OSCC-GB_012301 OSCC-GB_012401 OSCC-GB_012501 OSCC-GB_012601
OSCC-GB_012701

OV-AU (IGCC) 70 Donors

AOCS-001 AOCS-004 AOCS-005 AOCS-034 AOCS-055 AOCS-056 AOCS-058 AOCS-059 AOCS-060 AOCS-
061 AOCS-063 AOCS-064 AOCS-065 AOCS-075 AOCS-077 AOCS-078 AOCS-079 AOCS-080 AOCS-081 AOCS-
083 AOCS-084 AOCS-085 AOCS-086 AOCS-088 AOCS-090 AOCS-091 AOCS-092 AOCS-093 AOCS-094 AOCS-
095 AOCS-096 AOCS-097 AOCS-104 AOCS-106 AOCS-107 AOCS-108 AOCS-109 AOCS-111 AOCS-112 AOCS-
113 AOCS-114 AOCS-115 AOCS-116 AOCS-117 AOCS-119 AOCS-120 AOCS-128 AOCS-134 AOCS-138 AOCS-
139 AOCS-141 AOCS-142 AOCS-150 AOCS-153 AOCS-155 AOCS-157 AOCS-158 AOCS-159 AOCS-160 AOCS-
161 AOCS-162 AOCS-163 AOCS-164 AOCS-165 AOCS-166 AOCS-167 AOCS-168 AOCS-169 AOCS-170 AOCS-
171

OV-US (TCGA) 38 Donors

TCGA-04-1331 TCGA-04-1347 TCGA-04-1349 TCGA-04-1367 TCGA-04-1514 TCGA-04-1542 TCGA-09-1666
TCGA-09-2045 TCGA-09-2050 TCGA-10-0934 TCGA-10-0937 TCGA-10-0938 TCGA-13-0727 TCGA-13-0906
TCGA-13-0912 TCGA-13-1477 TCGA-13-1487 TCGA-13-1491 TCGA-23-1110 TCGA-23-1118 TCGA-23-1124
TCGA-24-1419 TCGA-24-1466 TCGA-24-1544 TCGA-24-1548 TCGA-24-1552 TCGA-24-1557 TCGA-24-1558
TCGA-24-1562 TCGA-24-1614 TCGA-24-2024 TCGA-24-2290 TCGA-25-1632 TCGA-25-1634 TCGA-25-2391
TCGA-25-2400 TCGA-36-1574 TCGA-61-2000

PACA-AU (IGCC) 94 Donors

ICGC_0006 ICGC_0007 ICGC_0009 ICGC_0020 ICGC_0021 ICGC_0025 ICGC_0026 ICGC_0031 ICGC_0033
ICGC_0037 ICGC_0048 ICGC_0051 ICGC_0052 ICGC_0053 ICGC_0054 ICGC_0059 ICGC_0061 ICGC_0063
ICGC_0066 ICGC_0067 ICGC_0069 ICGC_0075 ICGC_0087 ICGC_0088 ICGC_0099 ICGC_0103 ICGC_0105
ICGC_0108 ICGC_0109 ICGC_0114 ICGC_0115 ICGC_0124 ICGC_0134 ICGC_0135 ICGC_0139 ICGC_0140
ICGC_0141 ICGC_0143 ICGC_0144 ICGC_0146 ICGC_0149 ICGC_0150 ICGC_0153 ICGC_0169 ICGC_0185
ICGC_0192 ICGC_0199 ICGC_0201 ICGC_0205 ICGC_0206 ICGC_0207 ICGC_0212 ICGC_0214 ICGC_0215
ICGC_0217 ICGC_0223 ICGC_0224 ICGC_0227 ICGC_0230 ICGC_0295 ICGC_0296 ICGC_0300 ICGC_0301
ICGC_0303 ICGC_0304 ICGC_0309 ICGC_0312 ICGC_0313 ICGC_0315 ICGC_0321 ICGC_0326 ICGC_0338
ICGC_0354 ICGC_0365 ICGC_0391 ICGC_0392 ICGC_0393 ICGC_0395 ICGC_0406 ICGC_0412 ICGC_0417
ICGC_0419 ICGC_0420 ICGC_0486 ICGC_0502 ICGC_0507 ICGC_0518 ICGC_0521 ICGC_0522 ICGC_0526
ICGC_0533 ICGC_0535 ICGC_0536 ICGC_0543

PACA-CA (ICGC) 107 Donors

PCSI_0001 PCSI_0002 PCSI_0004 PCSI_0015_1 PCSI_0024 PCSI_0074 PCSI_0077 PCSI_0078 PCSI_0080 PCSI_0081
PCSI_0082 PCSI_0083_1 PCSI_0084 PCSI_0096 PCSI_0101 PCSI_0103 PCSI_0105 PCSI_0108 PCSI_0111 PCSI_0132
PCSI_0142 PCSI_0145 PCSI_0146 PCSI_0161 PCSI_0162 PCSI_0164 PCSI_0170 PCSI_0171 PCSI_0173 PCSI_0174
PCSI_0175 PCSI_0208 PCSI_0210 PCSI_0217 PCSI_0226 PCSI_0227 PCSI_0228 PCSI_0230 PCSI_0233 PCSI_0235
PCSI_0239 PCSI_0240 PCSI_0248 PCSI_0250 PCSI_0253 PCSI_0256 PCSI_0264 PCSI_0268 PCSI_0269 PCSI_0274
PCSI_0279 PCSI_0280 PCSI_0281 PCSI_0283 PCSI_0284 PCSI_0285 PCSI_0286 PCSI_0287 PCSI_0290 PCSI_0292
PCSI_0294 PCSI_0297 PCSI_0300 PCSI_0302 PCSI_0324 PCSI_0325 PCSI_0326 PCSI_0328 PCSI_0334 PCSI_0337
PCSI_0338 PCSI_0340 PCSI_0341 PCSI_0345 PCSI_0351 PCSI_0353 PCSI_0375 PCSI_0392 PCSI_0404 PCSI_0406
PCSI_0413 PCSI_0450 PCSI_0451 PCSI_0456 PCSI_0457 PCSI_0463 PCSI_0465 PCSI_0466 PCSI_0467 PCSI_0468
PCSI_0469 PCSI_0472 PCSI_0473 PCSI_0476 PCSI_0477 PCSI_0492 PCSI_0504 PCSI_0506 PCSI_0508 PCSI_0509
PCSI_0527 PCSI_0528 PCSI_0531 PCSI_0537 PCSI_0547 PCSI_0572 PCSI_0352_1

PA-DKFZ (ICGC) 81 Donors

ICGC_PA100 ICGC_PA102 ICGC_PA103 ICGC_PA107 ICGC_PA108 ICGC_PA109 ICGC_PA10 ICGC_PA110
ICGC_PA112 ICGC_PA116 ICGC_PA117 ICGC_PA11 ICGC_PA126 ICGC_PA12 ICGC_PA131 ICGC_PA134 ICGC_PA135

ICGC_PA136 ICGC_PA138 ICGC_PA140 ICGC_PA143 ICGC_PA144 ICGC_PA145 ICGC_PA147 ICGC_PA148 ICGC_PA14 ICGC_PA150 ICGC_PA158 ICGC_PA159 ICGC_PA162 ICGC_PA163 ICGC_PA165 ICGC_PA17 ICGC_PA20 ICGC_PA21 ICGC_PA22 ICGC_PA24 ICGC_PA25 ICGC_PA29 ICGC_PA30 ICGC_PA33 ICGC_PA34 ICGC_PA36 ICGC_PA37 ICGC_PA3 ICGC_PA41 ICGC_PA42 ICGC_PA43 ICGC_PA46 ICGC_PA48 ICGC_PA4 ICGC_PA50 ICGC_PA53 ICGC_PA55 ICGC_PA56 ICGC_PA58 ICGC_PA59 ICGC_PA5 ICGC_PA64 ICGC_PA65 ICGC_PA69 ICGC_PA70 ICGC_PA71 ICGC_PA73 ICGC_PA75 ICGC_PA79 ICGC_PA81 ICGC_PA82 ICGC_PA83 ICGC_PA86 ICGC_PA88 ICGC_PA89 ICGC_PA8 ICGC_PA91 ICGC_PA92 ICGC_PA93 ICGC_PA94 ICGC_PA95 ICGC_PA96 ICGC_PA99 ICGC_PA9

PAEN-AU (IGCC) 102 Donors

ICGC_0425 ICGC_0427 ICGC_0428 ICGC_0431 ICGC_0432 ICGC_0433 ICGC_0434 ICGC_0435 ICGC_0436 ICGC_0437 ICGC_0438 ICGC_0439 ICGC_0440 ICGC_0441 ICGC_0443 ICGC_0446 ICGC_0447 ICGC_0449 ICGC_0452 ICGC_0453 ICGC_0455 ICGC_0456 ICGC_0457 ICGC_0459 ICGC_0489 ICGC_0491 ICGC_0492 ICGC_0497 ICGC_0498 ICGC_0500 ICGC_0501 ITNET-0026 ITNET-0028 ITNET-0052 ITNET-0087 ITNET-0100 ITNET-0107 ITNET-0118 ITNET-0128 ITNET-0134 ITNET-0144 ITNET-0148 ITNET-0151 ITNET-0152 ITNET-0673 ITNET-0681 ITNET-0695 ITNET-0700 ITNET-0783 ITNET-0797 ITNET-0809 ITNET-0813 ITNET-0833 ITNET-0850 ITNET-0900 ITNET-0911 ITNET-0935 ITNET-0938 ITNET-0941 ITNET-0962 ITNET-0968 ITNET-0993 ITNET-1000 ITNET-1001 ITNET-1027 ITNET-1044 ITNET-1047 ITNET-1050 ITNET-1053 ITNET-1081 ITNET-1257 ITNET-1265 ITNET-1266 ITNET-1270 ITNET-1273 ITNET-1286 ITNET-1288 ITNET-1293 ITNET-1301 ITNET-1304 ITNET-1308 ITNET-1309 ITNET-1312 ITNET-1314 ITNET-1317 ITNET-1320 NE-0009 NE-0010 NE-0012 NE-0017 NE-0018 NE-0020 NE-0021 NE-0023 NE-0025 NE-0026 NE-0027 NE-0028 NE-0029 NE-0032 NE-0033 NE-0038

PCNSL-DKFZ (H050 A050 & XD013) 11 Donors

H050-0GUK H050-46JU H050-6K3Z H050-D7C3 H050-D8YC H050-JVA9 H050-K5AJ H050-SECM H050-T0SR H050-TY1U H050-W01L

PGBM-DKFZ 41 Donors

ICGC_GBM11 ICGC_GBM16 ICGC_GBM17 ICGC_GBM18 ICGC_GBM19 ICGC_GBM1 ICGC_GBM22 ICGC_GBM23 ICGC_GBM24 ICGC_GBM25 ICGC_GBM2 ICGC_GBM42 ICGC_GBM43 ICGC_GBM44 ICGC_GBM45 ICGC_GBM48 ICGC_GBM52 ICGC_GBM53 ICGC_GBM54 ICGC_GBM55 ICGC_GBM56 ICGC_GBM57 ICGC_GBM58 ICGC_GBM59 ICGC_GBM5 ICGC_GBM60 ICGC_GBM62 ICGC_GBM63 ICGC_GBM65 ICGC_GBM67 ICGC_GBM6 ICGC_GBM79 ICGC_GBM7 ICGC_GBM82 ICGC_GBM83 ICGC_GBM85 ICGC_GBM86 ICGC_GBM96 ICGC_GBM97 ICGC_GBM98 ICGC_GBM9

PNET-DKFZ 7 Donors

ICGC_MB172 ICGC_MB182 ICGC_PNET01 ICGC_PNET02 ICGC_PNET03 ICGC_PNET04 ICGC_PNET05

PRAD-CA (ICGC) 108 Donors

CPCG0001 CPCG0003 CPCG0020 CPCG0040 CPCG0046 CPCG0047 CPCG0048 CPCG0057 CPCG0063 CPCG0073 CPCG0078 CPCG0081 CPCG0083 CPCG0087 CPCG0094 CPCG0095 CPCG0098 CPCG0099 CPCG0102 CPCG0121 CPCG0123 CPCG0127 CPCG0128 CPCG0154 CPCG0158 CPCG0166 CPCG0182 CPCG0184 CPCG0185 CPCG0189 CPCG0190 CPCG0191 CPCG0196 CPCG0199 CPCG0201 CPCG0206 CPCG0208 CPCG0210 CPCG0211 CPCG0213 CPCG0217 CPCG0232 CPCG0233 CPCG0234 CPCG0236 CPCG0238 CPCG0241 CPCG0242 CPCG0246 CPCG0248 CPCG0249 CPCG0250 CPCG0251 CPCG0255 CPCG0256 CPCG0258 CPCG0259 CPCG0262 CPCG0263 CPCG0265 CPCG0266 CPCG0267 CPCG0268 CPCG0269 CPCG0324 CPCG0331 CPCG0334 CPCG0336 CPCG0339 CPCG0340 CPCG0341 CPCG0342 CPCG0344 CPCG0345 CPCG0346 CPCG0348 CPCG0350 CPCG0352 CPCG0357 CPCG0360 CPCG0361 CPCG0362 CPCG0364 CPCG0365 CPCG0366 CPCG0368 CPCG0369 CPCG0371 CPCG0372 CPCG0373 CPCG0374 CPCG0375 CPCG0378 CPCG0379 CPCG0380 CPCG0387 CPCG0388 CPCG0391 CPCG0392 CPCG0401 CPCG0404 CPCG0407 CPCG0409 CPCG0410 CPCG0411 CPCG0412 CPCG0413 CPCG0414

PRAD-UK (IGCC) 33 Donors

0056_CRUK_PC_0056 0064_CRUK_PC_0064 0065_CRUK_PC_0065 0067_CRUK_PC_0067 0069_CRUK_PC_0069 0070_CRUK_PC_0070 0071_CRUK_PC_0071 0072_CRUK_PC_0072 0075_CRUK_PC_0075 0077_CRUK_PC_0077 0078_CRUK_PC_0078 0080_CRUK_PC_0080 0082_CRUK_PC_0082 0084_CRUK_PC_0084 0086_CRUK_PC_0086 0089_CRUK_PC_0089 0090_CRUK_PC_0090 0091_CRUK_PC_0091 0093_CRUK_PC_0093 0094_CRUK_PC_0094 A10-0015_CRUK_PC_0015_3 A12-0020_CRUK_PC_0020_1 A17-0095_CRUK_PC_0095_1 A21-0096_CRUK_PC_0096_1 A22-0016_CRUK_PC_0016_1 A24-0021_CRUK_PC_0021_1 A29-0017_CRUK_PC_0017_1 A31-0018_CRUK_PC_0018_1 A32-0019_CRUK_PC_0019_1 A34-0022_CRUK_PC_0022_1 0006_CRUK_PC_0006_1 0007_CRUK_PC_0007_1 0008_CRUK_PC_0008

PRAD-US (TCGA) 17 Donors

TCGA-CH-5750 TCGA-CH-5763 TCGA-CH-5771 TCGA-CH-5788 TCGA-CH-5789 TCGA-EJ-5503 TCGA-EJ-5506 TCGA-EJ-7791 TCGA-G9-6336 TCGA-G9-6365 TCGA-G9-6370 TCGA-G9-7522 TCGA-HC-7075 TCGA-HC-7233 TCGA-HC-7737 TCGA-HC-8258 TCGA-HI-7169

READ-US (TCGA) 14 Donors

TCGA-AF-2689 TCGA-AF-2691 TCGA-AG-3593 TCGA-AG-3727 TCGA-AG-3885 TCGA-AG-3890 TCGA-AG-3896 TCGA-AG-3901 TCGA-AG-4007 TCGA-AG-4008 TCGA-AG-4015 TCGA-AG-A032 TCGA-EI-6917 TCGA-F5-6814

RECA-EU (IGCC) 71 Donors

C0004 C0005 C0006 C0008 C0009 C0011 C0012 C0013 C0014 C0015 C0016 C0018 C0019 C0020 C0021 C0022

C0023 C0024 C0025 C0026 C0027 C0028 C0031 C0033 C0034 C0035 C0037 C0038 C0039 C0040 C0041 C0042 C0043 C0045 C0046 C0047 C0048 C0049 C0050 C0051 C0052 C0054 C0055 C0056 C0057 C0060 C0062 C0063 C0064 C0065 C0066 C0068 C0070 C0071 C0073 C0074 C0075 C0077 C0079 C0080 C0081 C0082 C0084 C0086 C0088 C0091 C0092 C0094 C0098 C0099 C0100

SARC-US (TCGA) 33 Donors

TCGA-DX-A1KU TCGA-DX-A1KW TCGA-DX-A1L0 TCGA-DX-A1L2 TCGA-DX-A1L3 TCGA-DX-A23R TCGA-DX-A240 TCGA-DX-A2IZ TCGA-DX-A2J0 TCGA-DX-A2J4 TCGA-DX-A3LS TCGA-DX-A3LT TCGA-DX-A3LU TCGA-DX-A3LW TCGA-DX-A3LY TCGA-DX-A3M1 TCGA-DX-A3U5 TCGA-DX-A3U6 TCGA-DX-A3U7 TCGA-DX-A3U8 TCGA-FX-A2QS TCGA-FX-A3NJ TCGA-FX-A3RE TCGA-FX-A48G TCGA-HB-A5W3 TCGA-IE-A4EI TCGA-IE-A4EK TCGA-IF-A4AJ TCGA-IS-A3K7 TCGA-IS-A3KA TCGA-IW-A3M4 TCGA-IW-A3M5 TCGA-MO-A47R

SKCM-US (TCGA) 36 Donors

TCGA-D3-A1Q1 TCGA-D3-A1Q5 TCGA-D3-A3MO TCGA-DA-A1HV TCGA-DA-A1HW TCGA-DA-A1HY TCGA-DA-A1I0 TCGA-DA-A1I2 TCGA-DA-A1I8 TCGA-DA-A3F3 TCGA-DA-A3F5 TCGA-DA-A3F8 TCGA-EB-A24D TCGA-EE-A185 TCGA-EE-A29B TCGA-EE-A2A0 TCGA-EE-A2GT TCGA-EE-A2M5 TCGA-EE-A2MI TCGA-EE-A3J5 TCGA-EE-A3JI TCGA-ER-A19D TCGA-ER-A19E TCGA-ER-A19J TCGA-ER-A19L TCGA-ER-A19T TCGA-ER-A2NF TCGA-FS-A1ZD TCGA-FS-A1ZK TCGA-FS-A1ZP TCGA-FS-A1ZU TCGA-GN-A262 TCGA-GN-A264 TCGA-GN-A266 TCGA-GN-A26A TCGA-GN-A26C

STAD-US (TCGA) 24 Donors

TCGA-BR-4255 TCGA-BR-4280 TCGA-BR-6452 TCGA-BR-6564 TCGA-BR-7722 TCGA-BR-8373 TCGA-BR-8486 TCGA-BR-8682 TCGA-BR-8690 TCGA-CD-8529 TCGA-CG-4442 TCGA-CG-4443 TCGA-CG-4474 TCGA-CG-5723 TCGA-D7-6519 TCGA-D7-6527 TCGA-D7-6528 TCGA-D7-6815 TCGA-D7-6822 TCGA-F1-6177 TCGA-F1-6875 TCGA-HF-7136 TCGA-HU-8245 TCGA-IN-7806

THCA-US (TCGA) 31 Donors

TCGA-BJ-A191 TCGA-BJ-A45K TCGA-DE-A2OL TCGA-DE-A3KN TCGA-DJ-A13R TCGA-DJ-A13W TCGA-DJ-A1QL TCGA-DJ-A2Q1 TCGA-DJ-A2Q2 TCGA-DJ-A2Q8 TCGA-DJ-A3US TCGA-EL-A3CV TCGA-EL-A3CX TCGA-EL-A3MY TCGA-EL-A3T0 TCGA-EL-A3T9 TCGA-EL-A3TB TCGA-EM-A2CN TCGA-EM-A2CP TCGA-EM-A2OW TCGA-EM-A3AL TCGA-EM-A3AQ TCGA-EM-A3FL TCGA-EM-A3FQ TCGA-ET-A3DV TCGA-FE-A22Z TCGA-FE-A233 TCGA-FE-A3PD TCGA-FK-A3SD TCGA-FK-A3SE TCGA-L6-A4ET

UCEC-US (TCGA) 44 Donors

TCGA-A5-A0G9 TCGA-A5-A0GE TCGA-A5-A0GG TCGA-A5-A0GJ TCGA-AJ-A23M TCGA-AP-A051 TCGA-AP-A052 TCGA-AP-A053 TCGA-AP-A054 TCGA-AP-A05A TCGA-AP-A0L8 TCGA-AP-A0L9 TCGA-AP-A0LD TCGA-AP-A0LE TCGA-AP-A0LF TCGA-AP-A0LH TCGA-AP-A0LI TCGA-AP-A0LL TCGA-AP-A0LO TCGA-AX-A0J1 TCGA-AX-A1CI TCGA-AX-A2H5 TCGA-B5-A0JN TCGA-B5-A0K8 TCGA-B5-A11G TCGA-B5-A11H TCGA-B5-A11I TCGA-B5-A1MY TCGA-BG-A18C TCGA-BK-A0CC TCGA-BK-A139 TCGA-BS-A0TC TCGA-BS-A0TD TCGA-BS-A0TE TCGA-BS-A0U9 TCGA-BS-A0V8 TCGA-D1-A16G TCGA-D1-A17K TCGA-D1-A1NU TCGA-DI-A1NN TCGA-E6-A1LZ TCGA-EO-A1Y8 TCGA-EY-A1GS TCGA-EY-A1GW

7qAML-DKFZ (H030) 18 Donors

H030-2C37UE, H030-2KFQ, H030-3N1U, H030-67DF, H030-6L9J, H030-88FP, H030-A87Q, H030-D1KL, H030-DX59, H030-EMGX, H030-GBWS, H030-J3P5, H030-MTM9, H030-NEHVx, H030-SKLRJZ, H030-T2D7, H030-YE4Y, H030-ZF5JFZ

AML-DKTK (XD001) 35 Donors

XD001_P001, XD001_P002, XD001_P003, XD001_P004, XD001_P005, XD001_P006, XD001_P007, XD001_P008, XD001_P009, XD001_P010, XD001_P011, XD001_P012, XD001_P013, XD001_P014, XD001_P015, XD001_P016, XD001_P017, XD001_P018, XD001_P019, XD001_P101, XD001_P102, XD001_P103, XD001_P104, XD001_P105, XD001_P108, XD001_P109, XD001_P110, XD001_P111, XD001_P112, XD001_P113, XD001_P114, XD001_P115, XD001_P116, XD001_P117, XD001_P118

ATRT-DKFZ (XI010) 3 Donors

XI010_ATRT_E1075-13, XI010_NCH3602, XI010_NCH3786

COLORECTAL_family-DKFZ (XI006) 63 Donors

XI006_F11_S3, XI006_F12_S1, XI006_F12_S2, XI006_F12_S3, XI006_F12_S4, XI006_F12_S5, XI006_F13_S11, XI006_F13_S1, XI006_F13_S2, XI006_F13_S3, XI006_F13_S6, XI006_F14_S1, XI006_F14_S4, XI006_F15_S1, XI006_F15_S2, XI006_F15_S3, XI006_F15_S4, XI006_F16_S2, XI006_F16_S3, XI006_F16_S8, XI006_F17_S1, XI006_F17_S2, XI006_F17_S3, XI006_F17_S4, XI006_F17_S5, XI006_F18_S1, XI006_F18_S2, XI006_F18_S3, XI006_F18_S4, XI006_F18_S5, XI006_F19_S1, XI006_F19_S2, XI006_F19_S3, XI006_F20_S1, XI006_F20_S2, XI006_F20_S3, XI006_F20_S4, XI006_F20_S5, XI006_F20_S6, XI006_F21_S1, XI006_F22_S1, XI006_F22_S2, XI006_F22_S3, XI006_F7_S1, XI006_F7_S3, XI006_F7_S4, XI006_F8_S1, XI006_F8_S2, XI006_F8_S3, XI006_F8_S5, XI006_F9_S1, XI006_F9_S2, XI006_F9_S3, XI006_F9_S4, XI039_CRC_L1_S1, XI039_CRC_L1_S2, XI039_CRC_L1_S3, XI039_CRC_L1_S4, XI039_CRC_L2_S1, XI039_CRC_L2_S2, XI039_CRC_L2_S3, XI039_CRC_L2_S4, XI039_CRC_L2_S5

EPN-DKFZ (XI049 & XI061) 41 Donors

XI049_11EP25, XI049_11EP26, XI049_11EP6, XI049_4EP24, XI049_4EP29, XI049_7EP10, XI049_7EP31, XI049_7EP48, XI049_9EP19, XI049_9EP1, XI049_9EP33, XI049_9EP7, XI049_9EP9, XI049_NCH2053, XI061_15EP1, XI061_15EP2,

XI061_15EP7, XI061_15EP8, XI061_15EP9, XI061_16EP10, XI061_16EP5, XI061_4EP17, XI061_4EP19, XI061_4EP28, XI061_4EP32, XI061_4EP33, XI061_4EP35, XI061_4EP5, XI061_4EP8, XI061_7EP11, XI061_7EP18, XI061_7EP23, XI061_7EP34, XI061_7EP3, XI061_7EP4, XI061_7EP53, XI061_9EP14, XI061_9EP15, XI061_9EP29, XI061_9EP4, XI061_9EP8

GCTB-DKFZ (XI041) 9 Donors

XI041_05, XI041_06, XI041_07, XI041_AP, XI041_AQ, XI041_AR, XI041_AW, XI041_AX, XI041_AZ

HODGKIN_family-DKFZ (XI019) 16 Donors

XI019_HL1_S1, XI019_HL1_S2, XI019_HL1_S3, XI019_HL1_S4, XI019_HL2_S1, XI019_HL2_S2, XI019_HL2_S3, XI019_HL2_S4, XI019_HL2_S5, XI019_HL2_S6, XI019_HL2_S7, XI019_HL2_S8, XI019_HL2_S9, XI040_HL3_S1, XI040_HL3_S2, XI040_HL3_S3

MASTER-DKFZ (H021) 254 Donors

H021-149FSY, H021-3NLUM3, H021-3W9ZP5, H021-418T28, H021-6181DL, H021-69KU1M, H021-6WQ2QS, H021-92XTB6, H021-93UM8E, H021-98T9DN, H021-A4M4YB, H021-AL8AJM, H021-ANN8D1, H021-ASP4P2, H021-C52P, H021-F3AMK8, H021-F7VR2D, H021-FQVVMY, H021-G2U6LS, H021-G83612, H021-G9Y735, H021-H7D3JF, H021-J3DFTN, H021-JTCU87, H021-KLFE4P, H021-KYULH4, H021-L4M2BM, H021-M4HLEN, H021-NKKQGS, H021-PW8E2Y, H021-PWTUGA, H021-QKQDKB, H021-QLJDYA, H021-RPLGMN, H021-TQ37KJ, H021-VD9X43, H021-VJWKMM, H021-W2LFQY, H021-W77W46, H021-WS1A21, H021-WY8KBR, H021-X2UVYS, H021-X4TX9T, H021-X5TC7W, H021-YQY48T, H021-YWHWPQ, H021-1MTCZ9, H021-39SEUY, H021-3AFKXB, H021-526FUD, H021-7U9BYK, H021-86G9QA, H021-C41MJC, H021-CBCJVV, H021-CW1ZAY, H021-DES7ZD, H021-DNMH, H021-DQM8YM, H021-E56A4U, H021-F4WCGV, H021-G3SKZ9, H021-HBJXZR, H021-J5LPWZ, H021-J8TBFU, H021-JK6EGS, H021-JY3G3X, H021-L47PRH, H021-L84HD5, H021-MBCV6L, H021-NEGW81, H021-NLUJR7, H021-P6WZNK, H021-RCCY57, H021-RH38ZJ, H021-S1RRD7, H021-T8XN8C, H021-TKDN9D, H021-TQDQ1W, H021-VAL68P, H021-VRTGDW, H021-VWTSTS, H021-WPZAQG, H021-XGJCTR, H021-XXZ32Z, H021-Y24BUL, H021-Y2A8X6, H021-YUQ155, H021-36H8XH, H021-3W6BBQ, H021-7VRVFG, H021-ATYVKC, H021-DW9VUH, H021-H1DNZG, H021-JCYXC8, H021-JSB14V, H021-ME44PK, H021-N5Y1YY, H021-NJQ3P2, H021-NZADYV, H021-PNKYGC, H021-PU9YWU, H021-QH6MPU, H021-QMVRLC, H021-SBTCJ6, H021-SM96YS, H021-VNE8G3, H021-YG488C, H021-0WXB, H021-11YR8P, H021-13JVXT, H021-19D5EB, H021-1K9WSG, H021-1MA1EB, H021-1S82CX, H021-25QTMN, H021-2EHFSJ, H021-2NNFEM, H021-2RTMCP, H021-33UKSM, H021-3AT3LR, H021-3GTR71, H021-3LSLZ1, H021-4QWVAG, H021-546PMF, H021-56G7K1, H021-5FFPX1, H021-5MM8HX, H021-5PS3DS, H021-64DVV7, H021-6B84TJ, H021-6MDRB5, H021-77DLK4, H021-79PAJQ, H021-7HFAWC, H021-7NPSJW, H021-7SZUP1, H021-7Z95ZH, H021-85D55Z, H021-8AAMZH, H021-8BS2HW, H021-8EFPXA, H021-8GEBK9, H021-8KF22F, H021-8NFHC7, H021-99G9EH, H021-9BRWLP, H021-9WNHUM, H021-A5F2C6, H021-ARFSS6, H021-B9JGVS, H021-BMXPBG, H021-BYAX6U, H021-C8HAJF, H021-CAWW7F, H021-CD28AE, H021-CEFU8J, H021-CK4ZJD, H021-CW6Z4Z, H021-DFFFNC, H021-E26QTY, H021-E54T3F, H021-E57XBR, H021-EGPXE1, H021-ERB96F, H021-EU5YD1, H021-F5HFJQ, H021-FBVAAW, H021-FY5ZQW, H021-J5DYDL, H021-JJ3WAR, H021-JL8KLN, H021-K5991L, H021-KEHA71, H021-KFPLSM, H021-KJC5NU, H021-KQPCJ2, H021-L27TZN, H021-LDUYDE, H021-LKS5UH, H021-LNGXFG, H021-LQDPEM, H021-MALFTE, H021-MDPEHB, H021-MRF5FT, H021-MST6JK, H021-MTK64L, H021-N5YRV7, H021-N8HWFF, H021-NDTATK, H021-NGLL19, H021-NRURPH, H021-NRXWGD, H021-NUVEYH, H021-NVGQD4, H021-P7EW6X, H021-PF81SF, H021-PLH861, H021-PMJNX8, H021-PSBRHM, H021-Q7RL, H021-QFC8A8, H021-QL4GEW, H021-QM4LSB, H021-QVNAQT, H021-QY2Q95, H021-R82FD7, H021-RFEX31, H021-RFH9U7, H021-RG2E96, H021-S7ZXC4, H021-S971AK, H021-SSBE3A, H021-T2J7Z8, H021-T62214, H021-TLYJGN, H021-TPPC3H, H021-V4X4A3, H021-V6CGPQ, H021-V7PYRH, H021-V9S4DY, H021-VMWW8J, H021-W1TY38, H021-WDZGV8, H021-WFRX55, H021-X44M88, H021-X8KH1U, H021-XM4N, H021-XPPAA6, H021-XXTJSH, H021-Y1KHHA, H021-Y68SSB, H021-Y799BH, H021-Y9D9WJ, H021-YCUT9Q, H021-YLAXVK, H021-YLNJYE, H021-YW3Y2R, H021-YY7PXK, H021-Z3254X, H021-Z7JY6A, H021-ZHGGJC, H021-ZJ59W8, H021-ZLVJZ3, H021-ZT3X9F, H021-ZY1A8Q, H021-7D43, H021-8LPS87, H021-8UULQ8, H021-9VNPZ3, H021-AMRWA3, H021-GYRGG6, H021-MC1F4S, H021-RYHU45, H021-TPL9

MMML-DKFZ (ICGC) 52 Donors

4100049, 4101392, 4101669, 4103593, 4103627, 4104119, 4105782, 4107597, 4112817, 4114033, 4115022, 4118156, 4119702, 4120879, 4121621, 4121974, 4124432, 4126692, 4128355, 4128435, 4128849, 4130865, 4131213, 4131750, 4135813, 4136095, 4137230, 4138059, 4138629, 4139212, 4139483, 4140544, 4146301, 4152611, 4161288, 4161486, 4164330, 4167925, 4171586, 4171946, 4175941, 4176046, 4176584, 4177842, 4178243, 4178605, 4179976, 4182605, 4184437, 4186613, 4186812, 4190929

MNG-DKFZ (H033 & A033) 35 Donors

A033-1BZ58C, A033-1U9DCT, A033-2CP3XF, A033-6UE3XW, A033-823NUE, A033-B1C6W3, A033-B4JH92, A033-FGGLS9, A033-J31ASL, A033-JSKEM7, A033-MHA1P2, A033-MY5SG9, A033-N4QFCQ, A033-P2ZRLK, A033-PQP37P, A033-TS1K41, A033-VGSY3R, A033-XLW5FQ, DKFZ-Mrad_10, DKFZ-Mrad_11, DKFZ-Mrad_12, DKFZ-Mrad_13, DKFZ-Mrad_14, DKFZ-Mrad_16, DKFZ-Mrad_18, DKFZ-Mrad_1, DKFZ-Mrad_20, DKFZ-Mrad_2, DKFZ-Mrad_3, DKFZ-Mrad_4, DKFZ-Mrad_5, DKFZ-Mrad_6, DKFZ-Mrad_7, DKFZ-Mrad_8, DKFZ-Mrad_9

MPNST-DKFZ (XI086) 6 Donors

XI086_T1320, XI086_T1340, XI086_T1507, XI086_T1794, XI086_T2300B, XI086_T2302,

NB-DKFZ (GPOH) 67 Donors

XI003_10023, XI003_14312, XI003_14359, XI003_14527, XI003_15015, XI003_15403, XI003_15836, XI003_15885, XI003_16663, XI003_17041, XI003_17209, XI003_17752, XI003_17777, XI003_17863, XI003_17871, XI003_18533, XI003_18800, XI003_18802, XI003_18874, XI003_19079, XI003_19461, XI003_19493, XI003_19624, XI003_19924, XI003_19986, XI003_20005, XI003_20153, XI003_20173, XI003_20195, XI003_20273, XI003_20289, XI003_20391, XI003_20513, XI003_20855, XI003_20992, XI003_21014, XI003_21129, XI003_21248, XI003_21260, XI003_21390, XI003_2151, XI003_21525, XI003_21533, XI003_21776, XI003_21809, XI003_21954, XI003_21976, XI003_21984, XI003_22650, XI003_22677, XI003_22981, XI003_22998, XI003_23011, XI003_231, XI003_23298, XI003_23484, XI003_24642, XI003_25509, XI003_3623, XI003_4091, XI003_91257, XI003_NB-AUT-2, XI003_NB-AUT-3, XI003_STA-NB-2, XI003_STA-NB-3, XI003_STA-NB-4, XI003_STA-NB-6

PANCREAS_family-DKFZ (XI001) 25 Donors

XI001a_25-4-46-1020304, XI001a_25-4-46-10203, XI001a_25-4-46-1020403, XI001a_25-4-46-10206, XI001a_25-4-46-102, XI001a_25-4-46-1, XI001a_25-9-44-202, XI001a_25-9-44-204, XI001a_25-9-44-20504, XI001a_25-9-44-206, XI001a_25-9-44-208, XI001a_25-9-44-23, XI001a_25-9-44-2, XI001b_02-5-0382-1, XI001b_09-2-0552-2, XI001b_25-1-000091-109, XI001b_25-2-000014-206, XI001b_25-2-000129-1, XI001b_25-4-000088-206, XI001b_25-4-000106-204, XI001b_25-6-000078-1, XI001b_25-7-000100-1, XI001b_25-7-000170-1, XI001b_25-8-000168-2, XI001b_25-9-000113-1

PCNSL-DKFZ (H050, A050 & XD013) 18 Donors

A050-4FXC, A050-96KS, A050-FXXS, A050-GXBF, A050-LHDE, A050-LHT0, A050-V7L7, XD013-9J3NEC, XD013-A1PYDP, XD013-BDA9BY, XD013-FABSPL, XD013-JM25UE, XD013-MT9TKD, XD013-N4HJJC, XD013-NVLQJ1, XD013-U8U5KA, XD013-VHXW1M, XD013-YFQAL6

RMM-DKFZ (H067) 14 Donors

H067-37GLKW, H067-4M6LA9, H067-BUJCEE, H067-VZTYYN, H067-XJPW7S, H067-2YCH24, H067-6D6RLV, H067-BUGVTY, H067-EC4C3W, H067-GJVKDV, H067-PL3LWG, H067-SRERYH, H067-UEUCJL, H067-ZMWZPS

SYSGLIO-DKFZ (H043) 42 Donors

H043-28GK, H043-2JY3, H043-3P8XJY, H043-4PGF, H043-5FM91P, H043-5VWP, H043-63R6, H043-6F91, H043-6FRXV9, H043-B7R7, H043-BU96, H043-CQM6QY, H043-CWJQEW, H043-D9MRCY, H043-DSX2, H043-F14UH1, H043-FEJ7C8, H043-GESMJV, H043-GG5L52, H043-GK5VYT, H043-GKS176, H043-KE3H42, H043-KZWS, H043-LNWEGT, H043-LQDGD4, H043-MXE7Y8, H043-N7LCPV, H043-NAFCCV, H043-PLM1, H043-PWC258, H043-PYL6FY, H043-QHGXQQ, H043-U65X, H043-ULLV, H043-URZNJE, H043-W99H5K, H043-WJ851A, H043-X33N8G, H043-XACH, H043-XG4KY2, H043-ZK2PKS, H043-ZMHY

THYROID_family-DKFZ (XI002) 25 Donors

XI002_102_1, XI002_102_2, XI002_102_3, XI002_102_6, XI002_102_7, XI002_102_8, XI002_146_1, XI002_146_2, XI002_146_3, XI002_188_2, XI002_188_3, XI002_188_4, XI002_224_1, XI002_224_2, XI002_224_3, XI002_224_4, XI002_224_5, XI002_224_6, XI002_224_7, XI002_224_8, XI002_75_12, XI002_75_13, XI002_75_28, XI002_75_2, XI002_75_8

REFERENCES

1. Sikora, M. *et al.* Ancient genomes show social and reproductive behavior of early Upper Paleolithic foragers. *Science* **358,** 659–662 (Nov. 2017).
2. Goody, J. *The development of the family and marriage in Europe* (Cambridge University Press, 1983).
3. Darwin, C. *The variation of animals and plants under domestication* (John Murray, 1868).
4. Mendel, G. Versuche ueber Pflanzen-Hybriden. *Verhandlungen des naturforschenden Vereines in Bruenn* (1866).
5. Darwin, C. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life* (John Murray, 1859).
6. Avery, O. T., Macleod, C. M. & McCarty, M. STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES : INDUCTION OF TRANSFORMATION BY A DESOXYRIBONUCLEIC ACID FRACTION ISOLATED FROM PNEUMOCOCCUS TYPE III. *J. Exp. Med.* **79,** 137–158 (1944).
7. WATSON, J. D. & CRICK, F. H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171,** 737–738 (1953).
8. JACOB, F. & MONOD, J. Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* **3,** 318–356 (1961).
9. NIRENBERG, M. W. & MATTHAEI, J. H. The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides. *Proc. Natl. Acad. Sci. U.S.A.* **47,** 1588–1602 (1961).
10. Crick, F. Central dogma of molecular biology. *Nature* **227,** 561–563 (1970).
11. Huberman, J. A. & Riggs, A. D. On the mechanism of DNA replication in mammalian chromosomes. *J. Mol. Biol.* **32,** 327–341 (1968).
12. Burgess, R. R., Travers, A. A., Dunn, J. J. & Bautz, E. K. Factor stimulating transcription by RNA polymerase. *Nature* **221,** 43–46 (1969).
13. Roeder, R. G. & Rutter, W. J. Multiple forms of DNA-dependent RNA polymerase in eukaryotic organisms. *Nature* **224,** 234–237 (1969).
14. Temin, H. M. & Mizutani, S. RNA-dependent DNA polymerase in virions of Rous sarcoma virus. *Nature* **226,** 1211–1213 (1970).
15. Baltimore, D. RNA-dependent DNA polymerase in virions of RNA tumour viruses. *Nature* **226,** 1209–1211 (1970).
16. Marteijn, J. A., Lans, H., Vermeulen, W. & Hoeijmakers, J. H. Understanding nucleotide excision repair and its roles in cancer and ageing. *Nat. Rev. Mol. Cell Biol.* **15,** 465–481 (2014).
17. Chang, H. H. Y., Pannunzio, N. R., Adachi, N. & Lieber, M. R. Non-homologous DNA end joining and alternative pathways to double-strand break repair. *Nat. Rev. Mol. Cell Biol.* **18,** 495–506 (Aug. 2017).
18. Renkawitz, J., Lademann, C. A. & Jentsch, S. Mechanisms and principles of homology search during recombination. *Nat. Rev. Mol. Cell Biol.* **15,** 369–383 (2014).

19. Allis, C. D. & Jenuwein, T. The molecular hallmarks of epigenetic control. *Nat. Rev. Genet.* **17,** 487–500 (Aug. 2016).

20. GOLD, M., HURWITZ, J. & ANDERS, M. The enzymatic methylation of RNA and DNA. *Biochem. Biophys. Res. Commun.* **11,** 107–114 (1963).

21. Lister, R. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462,** 315–322 (2009).

22. Ziller, M. J. *et al.* Genomic distribution and inter-sample variation of non-CpG methylation across human cell types. *PLoS Genet.* **7,** e1002389 (2011).

23. Reik, W., Dean, W. & Walter, J. Epigenetic reprogramming in mammalian development. *Science* **293,** 1089–1093 (2001).

24. Meissner, A. *et al.* Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454,** 766–770 (2008).

25. Bestor, T. H. The DNA methyltransferases of mammals. *Hum. Mol. Genet.* **9,** 2395–2402 (2000).

26. Nazor, K. L. *et al.* Recurrent variations in DNA methylation in human pluripotent stem cells and their differentiated derivatives. *Cell Stem Cell* **10,** 620–634 (2012).

27. Miller, O. J., Schnedl, W., Allen, J. & Erlanger, B. F. 5-Methylcytosine localised in mammalian constitutive heterochromatin. *Nature* **251,** 636–637 (1974).

28. Yang, X. *et al.* Gene body methylation can alter gene expression and is a therapeutic target in cancer. *Cancer Cell* **26,** 577–590 (2014).

29. Pombo, A. & Dillon, N. Three-dimensional genome architecture: players and mechanisms. *Nat. Rev. Mol. Cell Biol.* **16,** 245–257 (2015).

30. Hentschel, C. C. & Birnstiel, M. L. The organization and expression of histone gene families. *Cell* **25,** 301–313 (1981).

31. Weintraub, H. & Groudine, M. Chromosomal subunits in active genes have an altered conformation. *Science* **193,** 848–856 (1976).

32. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489,** 75–82 (2012).

33. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485,** 376–380 (2012).

34. Sexton, T. *et al.* Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell* **148,** 458–472 (2012).

35. Worcel, A. & Benyajati, C. Higher order coiling of DNA in chromatin. *Cell* **12,** 83–100 (1977).

36. Kouzarides, T. Chromatin modifications and their function. *Cell* **128,** 693–705 (2007).

37. Zhao, Y. & Garcia, B. A. Comprehensive Catalog of Currently Documented Histone Modifications. *Cold Spring Harb Perspect Biol* **7,** a025064 (2015).

38. Garcia-Ramirez, M., Rocchini, C. & Ausio, J. Modulation of chromatin folding by histone acetylation. *J. Biol. Chem.* **270,** 17923–17928 (1995).

39. Tan, M. *et al.* Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification. *Cell* **146,** 1016–1028 (2011).

40. Hebbes, T. R., Thorne, A. W. & Crane-Robinson, C. A direct link between core histone acetylation and transcriptionally active chromatin. *EMBO J.* **7,** 1395–1402 (1988).

41. Greer, E. L. & Shi, Y. Histone methylation: a dynamic mark in health, disease and inheritance. *Nat. Rev. Genet.* **13,** 343–357 (2012).

42. Espada, J. *et al.* Human DNA methyltransferase 1 is required for maintenance of the histone H3 modification pattern. *J. Biol. Chem.* **279,** 37175–37184 (2004).

43. Rose, N. R. & Klose, R. J. Understanding the relationship between DNA methylation and histone lysine methylation. *Biochim. Biophys. Acta* **1839,** 1362–1372 (2014).

44. Pandiyan, K. *et al.* Functional DNA demethylation is accompanied by chromatin accessibility. *Nucleic Acids Res.* **41,** 3973–3985 (2013).

45. Rossant, J. Genetic Control of Early Cell Lineages in the Mammalian Embryo. *Annu. Rev. Genet.* **52,** 185–201 (Nov. 2018).

46. Goodell, M. A., Nguyen, H. & Shroyer, N. Somatic stem cell heterogeneity: diversity in the blood, skin and intestinal stem cell compartments. *Nat. Rev. Mol. Cell Biol.* **16,** 299–309 (2015).

47. Avgustinova, A. & Benitah, S. A. Epigenetic control of adult stem cell function. *Nat. Rev. Mol. Cell Biol.* **17,** 643–658 (Oct. 2016).

48. Morgani, S. M. & Brickman, J. M. The molecular underpinnings of totipotency. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* **369** (2014).

49. Balazsi, G., van Oudenaarden, A. & Collins, J. J. Cellular decision making and biological noise: from microbes to mammals. *Cell* **144,** 910–925 (2011).

50. Arnold, S. J. & Robertson, E. J. Making a commitment: cell lineage allocation and axis patterning in the early mouse embryo. *Nat. Rev. Mol. Cell Biol.* **10,** 91–103 (2009).

51. Boroviak, T. *et al.* Lineage-Specific Profiling Delineates the Emergence and Progression of Naive Pluripotency in Mammalian Embryogenesis. *Dev. Cell* **35,** 366–382 (2015).

52. Valentine, J. W., Collins, A. G. & Meyer, C. P. Morphological complexity increase in metazoans. *Paleobiology* **20,** 131–142 (1994).

53. Stachelscheid, H. *et al.* CellFinder: a cell data repository. *Nucleic Acids Res.* **42,** D950–958 (2014).

54. Schaum, N. *et al.* Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562,** 367–372 (Oct. 2018).

55. Regev, A. *et al.* The Human Cell Atlas. *Elife* **6** (Dec. 2017).

56. Graf, T. & Enver, T. Forcing cells to change lineages. *Nature* **462,** 587–594 (2009).

57. Sonawane, A. R. *et al.* Understanding Tissue-Specific Gene Regulation. *Cell Rep* **21,** 1077–1088 (2017).

58. Sivakumar, A. & Kurpios, N. A. Transcriptional regulation of cell shape during organ morphogenesis. *J. Cell Biol.* **217,** 2987–3005 (2018).

59. Iyama, T. & Wilson, D. M. DNA repair mechanisms in dividing and non-dividing cells. *DNA Repair (Amst.)* **12,** 620–636 (2013).

60. Batlle, E. *et al.* Beta-catenin and TCF mediate cell positioning in the intestinal epithelium by controlling the expression of EphB/ephrinB. *Cell* **111,** 251–263 (2002).

61. Van de Wetering, M. *et al.* The beta-catenin/TCF-4 complex imposes a crypt progenitor phenotype on colorectal cancer cells. *Cell* **111,** 241–250 (2002).

62. Chepko, G. & Smith, G. H. Three division-competent, structurally-distinct cell populations contribute to murine mammary epithelial renewal. *Tissue Cell* **29,** 239–253 (1997).

63. Liu, X. & Engelhardt, J. F. The glandular stem/progenitor cell niche in airway development and repair. *Proc Am Thorac Soc* **5,** 682–688 (2008).

64. Tokar, E. J., Ancrile, B. B., Cunha, G. R. & Webber, M. M. Stem/progenitor and intermediate cell types and the origin of human prostate cancer. *Differentiation* **73,** 463–473 (2005).

65. Hofer, T. & Rodewald, H. R. Differentiation-based model of hematopoietic stem cell functions and lineage pathways. *Blood* **132,** 1106–1113 (Sept. 2018).

66. Cotsarelis, G., Sun, T. T. & Lavker, R. M. Label-retaining cells reside in the bulge area of pilosebaceous unit: implications for follicular stem cells, hair cycle, and skin carcinogenesis. *Cell* **61,** 1329–1337 (1990).

67. Tang, D., Kang, R., Berghe, T. V., Vandenabeele, P. & Kroemer, G. The molecular machinery of regulated cell death. *Cell Res.* **29,** 347–364 (2019).

68. Morin, P. J., Vogelstein, B. & Kinzler, K. W. Apoptosis and APC in colorectal tumorigenesis. *Proc. Natl. Acad. Sci. U.S.A.* **93,** 7950–7954 (1996).

69. Naghavi, M. *et al.* Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet* **385,** 117–171 (2015).

70. Fitzmaurice, C. *et al.* The Global Burden of Cancer 2013. *JAMA Oncol* **1,** 505–527 (2015).

71. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2019. *CA Cancer J Clin* **69,** 7–34 (2019).

72. Berman, J. J. Tumor taxonomy for the developmental lineage classification of neoplasms. *BMC Cancer* **4,** 88 (2004).

73. Parizel, P. M. *et al.* Brainstem hemorrhage in descending transtentorial herniation (Duret hemorrhage). *Intensive Care Med* **28,** 85–88 (2002).

74. Pereira, J. & Phan, T. Management of bleeding in patients with advanced cancer. *Oncologist* **9,** 561–570 (2004).

75. Ramos, R. E. O. *et al.* Malignancy-Related Hypercalcemia in Advanced Solid Tumors: Survival Outcomes. *J Glob Oncol* **3,** 728–733 (2017).

76. Rapoport, B. L. Management of the cancer patient with infection and neutropenia. *Semin. Oncol.* **38,** 424–430 (2011).

77. Breasted, J. H. *The Edwin Smith Surgical Papyrus* (University of Chicago Press, 1930).

78. Lawrence, W. in *Surgery* 1889–1900 (Springer, 2008).

79. Connell, P. P. & Hellman, S. Advances in radiotherapy and implications for the next century: a historical perspective. *Cancer Res.* **69,** 383–392 (2009).

80. DeVita, V. T. & Chu, E. A history of cancer chemotherapy. *Cancer Res.* **68,** 8643–8653 (2008).

81. Fu, D., Calvo, J. A. & Samson, L. D. Balancing repair and tolerance of DNA damage caused by alkylating agents. *Nat. Rev. Cancer* **12,** 104–120 (2012).

82. Pommier, Y. Topoisomerase I inhibitors: camptothecins and beyond. *Nat. Rev. Cancer* **6,** 789–802 (2006).

83. Nitiss, J. L. Targeting DNA topoisomerase II in cancer chemotherapy. *Nat. Rev. Cancer* **9,** 338–350 (2009).

84. Chan, K. S., Koh, C. G. & Li, H. Y. Mitosis-targeted anti-cancer therapies: where they stand. *Cell Death Dis* **3,** e411 (2012).

85. Kaye, S. B. New antimetabolites in cancer chemotherapy and their clinical impact. *Br. J. Cancer* **78 Suppl 3,** 1–7 (1998).

86. Luengo, A., Gui, D. Y. & Vander Heiden, M. G. Targeting Metabolism for Cancer Therapy. *Cell Chem Biol* **24,** 1161–1180 (2017).

87. Baselga, J. *et al.* Everolimus in postmenopausal hormone-receptor-positive advanced breast cancer. *N. Engl. J. Med.* **366,** 520–529 (2012).

88. Hellerstedt, B. A. & Pienta, K. J. The current state of hormonal therapy for prostate cancer. *CA Cancer J Clin* **52,** 154–179 (2002).

89. Bray, F. *et al.* Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* **68,** 394–424 (Nov. 2018).

90. Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100,** 57–70 (2000).

91. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144,** 646–674 (2011).

92. Warburg, O. H. *The metabolism of tumours: investigations from the Kaiser Wilhelm Institute for Biology, Berlin-Dahlem* (Constable & Company Limited, 1930).

93. Gerlinger, M. *et al.* Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* **366,** 883–892 (2012).

94. Greaves, M. & Maley, C. C. Clonal evolution in cancer. *Nature* **481,** 306–313 (2012).

95. McGranahan, N. & Swanton, C. Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. *Cell* **168,** 613–628 (Feb. 2017).

96. Prior, I. A., Lewis, P. D. & Mattos, C. A comprehensive survey of Ras mutations in cancer. *Cancer Res.* **72,** 2457–2467 (2012).

97. Garnett, M. J. & Marais, R. Guilty as charged: B-RAF is a human oncogene. *Cancer Cell* **6,** 313–319 (2004).

98. Aster, J. C., Pear, W. S. & Blacklow, S. C. The Varied Roles of Notch in Cancer. *Annu Rev Pathol* **12,** 245–275 (2017).

99. Choi, S. H. *et al.* The common oncogenomic program of NOTCH1 and NOTCH3 signaling in T-cell acute lymphoblastic leukemia. *PLoS ONE* **12,** e0185762 (2017).

100. Weischenfeldt, J. *et al.* Pan-cancer analysis of somatic copy-number alterations implicates IRS4 and IGF2 in enhancer hijacking. *Nat. Genet.* **49,** 65–74 (Jan. 2017).

101. Normanno, N. *et al.* Epidermal growth factor receptor (EGFR) signaling in cancer. *Gene* **366,** 2–16 (2006).

102. Normanno, N. *et al.* The ErbB receptors and their ligands in cancer: an overview. *Curr Drug Targets* **6,** 243–257 (2005).

103. Dienstmann, R. *et al.* Genomic aberrations in the FGFR pathway: opportunities for targeted therapies in solid tumors. *Ann. Oncol.* **25,** 552–563 (2014).

104. Hamilton, E. & Infante, J. R. Targeting CDK4/6 in patients with cancer. *Cancer Treat. Rev.* **45,** 129–138 (2016).

105. Musgrove, E. A., Caldon, C. E., Barraclough, J., Stone, A. & Sutherland, R. L. Cyclin D as a therapeutic target in cancer. *Nat. Rev. Cancer* **11,** 558–572 (2011).

106. Zhao, R., Choi, B. Y., Lee, M. H., Bode, A. M. & Dong, Z. Implications of Genetic and Epigenetic Alterations of CDKN2A (p16(INK4a)) in Cancer. *EBioMedicine* **8,** 30–39 (2016).

107. Muller, P. A. & Vousden, K. H. p53 mutations in cancer. *Nat. Cell Biol.* **15,** 2–8 (2013).

108. Cybulski, C. *et al.* CHEK2 is a multiorgan cancer susceptibility gene. *Am. J. Hum. Genet.* **75,** 1131–1135 (2004).

109. Di Fiore, R., D'Anneo, A., Tesoriere, G. & Vento, R. RB1 in cancer: different mechanisms of RB1 inactivation and alterations of pRb pathway in tumorigenesis. *J. Cell. Physiol.* **228,** 1676–1687 (2013).

110. Hollander, M. C., Blumenthal, G. M. & Dennis, P. A. PTEN loss in the continuum of common cancers, rare syndromes and mouse models. *Nat. Rev. Cancer* **11,** 289–301 (2011).

111. Reinders, M. G. *et al.* New mutations and an updated database for the patched-1 (PTCH1) gene. *Mol Genet Genomic Med* **6,** 409–415 (May 2018).

112. Philpott, C., Tovell, H., Frayling, I. M., Cooper, D. N. & Upadhyaya, M. The NF1 somatic mutational landscape in sporadic human cancers. *Hum. Genomics* **11,** 13 (June 2017).

113. Petrilli, A. M. & Fernandez-Valle, C. Role of Merlin/NF2 inactivation in tumor biology. *Oncogene* **35,** 537–548 (2016).

114. Miyaki, M. & Kuroki, T. Role of Smad4 (DPC4) inactivation in human cancer. *Biochem. Biophys. Res. Commun.* **306,** 799–804 (2003).

115. Kwong, L. N. & Dove, W. F. APC and its modifiers in colon cancer. *Adv. Exp. Med. Biol.* **656,** 85–106 (2009).

116. Ionov, Y., Yamamoto, H., Krajewski, S., Reed, J. C. & Perucho, M. Mutational inactivation of the proapoptotic gene BAX confers selective advantage during tumor clonal evolution. *Proc. Natl. Acad. Sci. U.S.A.* **97,** 10872–10877 (2000).

117. Schuetz, J. M. *et al.* BCL2 mutations in diffuse large B-cell lymphoma. *Leukemia* **26,** 1383–1390 (2012).

118. Lorenzetto, E. *et al.* YAP1 acts as oncogenic target of 11q22 amplification in multiple cancer subtypes. *Oncotarget* **5,** 2608–2621 (2014).

119. Schulz, W. A., Lang, A., Koch, J. & Greife, A. The histone demethylase UTX/KDM6A in cancer: Progress and puzzles. *Int. J. Cancer* **145,** 614–620 (2019).

120. Kudithipudi, S. & Jeltsch, A. Role of somatic cancer mutations in human protein lysine methyltransferases. *Biochim. Biophys. Acta* **1846,** 366–379 (2014).

121. Sauvageau, M. & Sauvageau, G. Polycomb group proteins: multi-faceted regulators of somatic stem cells and cancer. *Cell Stem Cell* **7,** 299–313 (2010).

122. Koontz, J. I. *et al.* Frequent fusion of the JAZF1 and JJAZ1 genes in endometrial stromal tumors. *Proc. Natl. Acad. Sci. U.S.A.* **98,** 6348–6353 (2001).

123. Zhang, W. & Xu, J. DNA methyltransferases and their roles in tumorigenesis. *Biomark Res* **5,** 1 (2017).

124. Nacev, B. A. *et al.* The expanding landscape of 'oncohistone' mutations in human cancers. *Nature* **567,** 473–478 (2019).

125. Mathur, R. ARID1A loss in cancer: Towards a mechanistic understanding. *Pharmacol. Ther.* **190,** 15–23 (Oct. 2018).

126. Helming, K. C., Wang, X. & Roberts, C. W. M. Vulnerabilities of mutant SWI/SNF complexes in cancer. *Cancer Cell* **26,** 309–317 (2014).

127. Mathsyaraja, H. & Eisenman, R. N. Parsing Myc Paralogs in Oncogenesis. *Cancer Cell* **29,** 1–2 (2016).

128. Lecuyer, E. & Hoang, T. SCL: from the origin of hematopoiesis to stem cells and leukemia. *Exp. Hematol.* **32,** 11–24 (2004).

129. Burnichon, N. *et al.* MAX mutations cause hereditary and sporadic pheochromocytoma and paraganglioma. *Clin. Cancer Res.* **18,** 2828–2837 (2012).

130. Von Bergh, A. R. *et al.* High incidence of t(7;12)(q36;p13) in infant AML but not in infant ALL, with a dismal outcome and ectopic expression of HLXB9. *Genes Chromosomes Cancer* **45,** 731–739 (2006).

131. Matsubara, D. *et al.* Inactivating mutations and hypermethylation of the NKX2-1/TTF-1 gene in non-terminal respiratory unit-type lung adenocarcinomas. *Cancer Sci.* **108,** 1888–1896 (2017).

132. Northcott, P. A. *et al.* Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. *Nature* **511,** 428–434 (2014).

133. O'Donovan, P. J. & Livingston, D. M. BRCA1 and BRCA2: breast/ovarian cancer susceptibility gene products and participants in DNA double-strand break repair. *Carcinogenesis* **31,** 961–967 (2010).

134. Mersch, J. *et al.* Cancers associated with BRCA1 and BRCA2 mutations other than breast and ovarian. *Cancer* **121,** 269–275 (2015).

135. Choi, M., Kipps, T. & Kurzrock, R. ATM Mutations in Cancer: Therapeutic Implications. *Mol. Cancer Ther.* **15,** 1781–1791 (Aug. 2016).

136. Rayner, E. *et al.* A panoply of errors: polymerase proofreading domain mutations in cancer. *Nat. Rev. Cancer* **16,** 71–81 (2016).

137. Roemer, M. G. *et al.* PD-L1 and PD-L2 Genetic Alterations Define Classical Hodgkin Lymphoma and Predict Outcome. *J. Clin. Oncol.* **34,** 2690–2697 (Aug. 2016).

138. McGranahan, N. *et al.* Allele-Specific HLA Loss and Immune Escape in Lung Cancer Evolution. *Cell* **171,** 1259–1271 (2017).

139. Holla, V. R. *et al.* ALK: a tyrosine kinase target for cancer therapy. *Cold Spring Harb Mol Case Stud* **3,** a001115 (2017).

140. Levis, M. & Small, D. FLT3: ITDoes matter in leukemia. *Leukemia* **17,** 1738–1752 (2003).

141. Okamura, R. *et al.* Analysis of NTRK Alterations in Pan-Cancer Adult and Pediatric Malignancies: Implications for NTRK-Targeted Therapeutics. *JCO Precis Oncol* **2018** (2018).

142. Fletcher, J. A. KIT Oncogenic Mutations: Biologic Insights, Therapeutic Advances, and Future Directions. *Cancer Res.* **76,** 6140–6142 (Nov. 2016).

143. Samuels, Y. & Waldman, T. Oncogenic mutations of PIK3CA in human cancers. *Curr. Top. Microbiol. Immunol.* **347,** 21–41 (2010).

144. Kralovics, R. *et al.* A gain-of-function mutation of JAK2 in myeloproliferative disorders. *N. Engl. J. Med.* **352,** 1779–1790 (2005).

145. Leao, R. *et al.* Mechanisms of human telomerase reverse transcriptase (hTERT) regulation: clinical impacts in cancer. *J. Biomed. Sci.* **25,** 22 (2018).

146. Heaphy, C. M. *et al.* Altered telomeres in tumors with ATRX and DAXX mutations. *Science* **333,** 425 (2011).

147. Gossage, L., Eisen, T. & Maher, E. R. VHL, the story of a tumour suppressor gene. *Nat. Rev. Cancer* **15,** 55–64 (2015).

148. Oliner, J. D., Saiki, A. Y. & Caenepeel, S. The Role of MDM2 Amplification and Overexpression in Tumorigenesis. *Cold Spring Harb Perspect Med* **6** (June 2016).

149. Yan, H. *et al.* IDH1 and IDH2 mutations in gliomas. *N. Engl. J. Med.* **360,** 765–773 (2009).

150. Clark, V. E. *et al.* Recurrent somatic mutations in POLR2A define a distinct subset of meningiomas. *Nat. Genet.* **48,** 1253–1259 (Oct. 2016).

151. Robertson, J. C., Jorcyk, C. L. & Oxford, J. T. DICER1 Syndrome: DICER1 Mutations in Rare Cancers. *Cancers (Basel)* **10** (2018).

152. Baudino, T. A. Targeted Cancer Therapy: The Next Generation of Cancer Treatment. *Curr Drug Discov Technol* **12,** 3–20 (2015).

153. Ryan, M. B., Der, C. J., Wang-Gillam, A. & Cox, A. D. Targeting RAS-mutant cancers: is ERK the key? *Trends Cancer* **1,** 183–198 (2015).

154. Hertzman Johansson, C. & Egyhazi Brage, S. BRAF inhibitors in cancer therapy. *Pharmacol. Ther.* **142,** 176–182 (2014).

155. Espinoza, I. & Miele, L. Notch inhibitors for cancer treatment. *Pharmacol. Ther.* **139,** 95–110 (2013).

156. Sanderson, M. P. *et al.* The IGF1R/INSR Inhibitor BI 885578 Selectively Inhibits Growth of IGF2-Overexpressing Colorectal Cancer Tumors and Potentiates the Efficacy of Anti-VEGF Therapy. *Mol. Cancer Ther.* **16,** 2223–2233 (Oct. 2017).

157. Van Cutsem, E. *et al.* Cetuximab and chemotherapy as initial treatment for metastatic colorectal cancer. *N. Engl. J. Med.* **360,** 1408–1417 (2009).

158. Wood, E. R. *et al.* A unique structure for epidermal growth factor receptor bound to GW572016 (Lapatinib): relationships among protein conformation, inhibitor off-rate, and receptor activity in tumor cells. *Cancer Res.* **64,** 6652–6659 (2004).

159. Rusnak, D. W. *et al.* The effects of the novel, reversible epidermal growth factor receptor/ErbB-2 tyrosine kinase inhibitor, GW2016, on the growth of human normal and tumor-derived cell lines in vitro and in vivo. *Mol. Cancer Ther.* **1,** 85–94 (2001).

160. Vogel, C. L. *et al.* Efficacy and safety of trastuzumab as a single agent in first-line treatment of HER2-overexpressing metastatic breast cancer. *J. Clin. Oncol.* **20,** 719–726 (2002).

161. Choo, J. R. & Lee, S. C. CDK4-6 inhibitors in breast cancer: current status and future development. *Expert Opin Drug Metab Toxicol* **14,** 1123–1138 (2018).

162. Rimkus, T. K., Carpenter, R. L., Qasem, S., Chan, M. & Lo, H. W. Targeting the Sonic Hedgehog Signaling Pathway: Review of Smoothened and GLI Inhibitors. *Cancers (Basel)* **8** (2016).

163. Zaman, S., Wang, R. & Gandhi, V. Targeting the apoptosis pathway in hematologic malignancies. *Leuk. Lymphoma* **55,** 1980–1992 (2014).

164. Williams, M. J., Singleton, W. G., Lowis, S. P., Malik, K. & Kurian, K. M. Therapeutic Targeting of Histone Modifications in Adult and Pediatric High-Grade Glioma. *Front Oncol* **7,** 45 (2017).

165. Ahuja, N., Sharma, A. R. & Baylin, S. B. Epigenetic Therapeutics: A New Weapon in the War Against Cancer. *Annu. Rev. Med.* **67,** 73–89 (2016).

166. Lambert, M., Jambon, S., Depauw, S. & David-Cordonnier, M. H. Targeting Transcription Factors for Cancer Treatment. *Molecules* **23** (2018).

167. Chen, H., Liu, H. & Qing, G. Targeting oncogenic Myc as a strategy for cancer treatment. *Signal Transduct Target Ther* **3,** 5 (2018).

168. Farmer, H. *et al.* Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. *Nature* **434,** 917–921 (2005).

169. Le, D. T. *et al.* Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. *Science* **357,** 409–413 (July 2017).

170. Nebot-Bral, L. *et al.* Hypermutated tumours in the era of immunotherapy: The paradigm of personalised medicine. *Eur. J. Cancer* **84,** 290–303 (Oct. 2017).

171. Pardoll, D. M. The blockade of immune checkpoints in cancer immunotherapy. *Nat. Rev. Cancer* **12,** 252–264 (2012).

172. Beavis, P. A. *et al.* Dual PD-1 and CTLA-4 Checkpoint Blockade Promotes Antitumor Immune Responses through CD4+Foxp3- Cell-Mediated Modulation of CD103+ Dendritic Cells. *Cancer Immunol Res* **6,** 1069–1081 (Sept. 2018).

173. Paulson, K. G. *et al.* Acquired cancer resistance to combination immunotherapy from transcriptional loss of class I HLA. *Nat Commun* **9,** 3868 (Sept. 2018).

174. Zhang, J., Yang, P. L. & Gray, N. S. Targeting cancer with small molecule kinase inhibitors. *Nat. Rev. Cancer* **9,** 28–39 (2009).

175. Lange, A. M. & Lo, H. W. Inhibiting TRK Proteins in Clinical Cancer Therapy. *Cancers (Basel)* **10** (2018).

176. Abbaspour Babaei, M., Kamalidehghan, B., Saleem, M., Huri, H. Z. & Ahmadipour, F. Receptor tyrosine kinase (c-Kit) inhibitors: a potential therapeutic target in cancer cells. *Drug Des Devel Ther* **10,** 2443–2459 (2016).

177. Vainchenker, W. *et al.* JAK inhibitors for the treatment of myeloproliferative neoplasms and other disorders. *F1000Res* **7,** 82 (2018).

178. Massacesi, C. *et al.* PI3K inhibitors as new cancer therapeutics: implications for clinical trial design. *Onco Targets Ther* **9,** 203–210 (2016).

179. Stone, R. M. *et al.* Midostaurin plus Chemotherapy for Acute Myeloid Leukemia with a FLT3 Mutation. *N. Engl. J. Med.* **377,** 454–464 (Aug. 2017).

180. Ward, R. J. & Autexier, C. Pharmacological telomerase inhibition can sensitize drug-resistant and drug-sensitive cells to chemotherapeutic treatment. *Mol. Pharmacol.* **68,** 779–786 (2005).

181. Burgess, A. *et al.* Clinical Overview of MDM2/X-Targeted Therapies. *Front Oncol* **6,** 7 (2016).

182. DiNardo, C. D. *et al.* Durable Remissions with Ivosidenib in IDH1-Mutated Relapsed or Refractory AML. *N. Engl. J. Med.* **378,** 2386–2398 (2018).

183. Myers, R. A., Wirth, S., Williams, S. & Kiel, P. J. Enasidenib: An Oral IDH2 Inhibitor for the Treatment of Acute Myeloid Leukemia. *J Adv Pract Oncol* **9,** 435–440 (2018).

184. Liu, Y. *et al.* TP53 loss creates therapeutic vulnerability in colorectal cancer. *Nature* **520,** 697–701 (2015).

185. Xu, J. *et al.* Precise targeting of POLR2A as a therapeutic strategy for human triple negative breast cancer. *Nat Nanotechnol* **14,** 388–397 (Apr. 2019).

186. Lievre, A. *et al.* KRAS mutation status is predictive of response to cetuximab therapy in colorectal cancer. *Cancer Res.* **66,** 3992–3995 (2006).

187. Heymann, S. *et al.* Radio-induced malignancies after breast cancer postoperative radiotherapy in patients with Li-Fraumeni syndrome. *Radiat Oncol* **5,** 104 (2010).

188. MacArthur, D. G. *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335,** 823–828 (2012).

189. Pagani, F. & Baralle, F. E. Genomic variants in exons and introns: identifying the splicing spoilers. *Nat. Rev. Genet.* **5,** 389–396 (2004).

190. Vinagre, J. *et al.* Frequency of TERT promoter mutations in human cancers. *Nat Commun* **4,** 2185 (2013).

191. Baker, S. J. *et al.* Chromosome 17 deletions and p53 gene mutations in colorectal carcinomas. *Science* **244,** 217–221 (1989).

192. Knuutila, S. *et al.* DNA copy number losses in human neoplasms. *Am. J. Pathol.* **155,** 683–694 (1999).

193. Yi, K. & Ju, Y. S. Patterns and mechanisms of structural variations in human cancer. *Exp. Mol. Med.* **50,** 98 (Aug. 2018).

194. Bartram, C. R. *et al.* Translocation of c-ab1 oncogene correlates with the presence of a Philadelphia chromosome in chronic myelocytic leukaemia. *Nature* **306,** 277–280 (1983).

195. Kakizuka, A *et al.* Chromosomal translocation t (15; 17) in human acute promyelocytic leukemia fuses RARA with a novel putative transcription factor, PML. *Cell* **66,** 663–674 (1991).

196. Liu, P., Hajra, A, Wijmenga, C & Collins, F. Molecular pathogenesis of the chromosome 16 inversion in the M4Eo subtype of acute myeloid leukemia. *Blood* **85,** 2289–2302 (1995).

197. Tomlins, S. A. *et al.* Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* **310,** 644–648 (2005).

198. Ueno-Yokohata, H. *et al.* Consistent in-frame internal tandem duplications of BCOR characterize clear cell sarcoma of the kidney. *Nat. Genet.* **47,** 861–863 (2015).

199. Slamon, D. J. *et al.* Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science* **235,** 177–182 (1987).

200. Brodeur, G. M., Seeger, R. C., Schwab, M., Varmus, H. E. & Bishop, J. M. Amplification of N-myc in untreated human neuroblastomas correlates with advanced disease stage. *Science* **224,** 1121–1124 (1984).

201. Schrock, E. *et al.* Comparative genomic hybridization of human malignant gliomas reveals multiple amplification sites and nonrandom chromosomal gains and losses. *Am. J. Pathol.* **144,** 1203–1218 (1994).

202. Miller, C. *et al.* Human p53 gene localized to short arm of chromosome 17. *Nature* **319,** 783–784 (1986).

203. Latif, F. *et al.* Identification of the von Hippel-Lindau disease tumor suppressor gene. *Science* **260,** 1317–1320 (1993).

204. Friend, S. H. *et al.* A human DNA segment with properties of the gene that predisposes to retinoblastoma and osteosarcoma. *Nature* **323,** 643–646 (1986).

205. Kamb, A. *et al.* A cell cycle regulator potentially involved in genesis of many tumor types. *Science* **264,** 436–440 (1994).

206. Li, J. *et al.* PTEN, a putative protein tyrosine phosphatase gene mutated in human brain, breast, and prostate cancer. *Science* **275,** 1943–1947 (1997).

207. Knudson, A. G. Mutation and cancer: statistical study of retinoblastoma. *Proc. Natl. Acad. Sci. U.S.A.* **68,** 820–823 (1971).

208. Berger, A. H., Knudson, A. G. & Pandolfi, P. P. A continuum model for tumour suppression. *Nature* **476,** 163–169 (2011).

209. Kuppers, R. & Dalla-Favera, R. Mechanisms of chromosomal translocations in B cell lymphomas. *Oncogene* **20,** 5580–5594 (2001).

210. Rabbitts, T. H. *et al.* The chromosomal location of T-cell receptor genes and a T cell rearranging gene: possible correlation with specific translocations in human T cell leukaemia. *EMBO J.* **4,** 1461–1465 (1985).

211. Northcott, P. A. *et al.* The whole-genome landscape of medulloblastoma subtypes. *Nature* **547,** 311–317 (July 2017).

212. Peifer, M. *et al.* Telomerase activation by genomic rearrangements in high-risk neuroblastoma. *Nature* **526,** 700–704 (2015).

213. Groschel, S. *et al.* A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in leukemia. *Cell* **157,** 369–381 (2014).

214. Flavahan, W. A. *et al.* Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature* **529,** 110–114 (2016).

215. Alonso-Espinaco, V. *et al.* Novel MLH1 duplication identified in Colombian families with Lynch syndrome. *Genet. Med.* **13,** 155–160 (2011).

216. Lorenz, S. *et al.* Unscrambling the genomic chaos of osteosarcoma reveals extensive transcript fusion, recurrent rearrangements and frequent novel TP53 aberrations. *Oncotarget* **7,** 5273–5288 (2016).

217. Chavan, S. S. *et al.* Bi-allelic inactivation is more prevalent at relapse in multiple myeloma, identifying RB1 as an independent prognostic marker. *Blood Cancer J* **7,** e535 (Feb. 2017).

218. Sturm, D. *et al.* New Brain Tumor Entities Emerge from Molecular Classification of CNS-PNETs. *Cell* **164,** 1060–1072 (2016).

219. Sparkes, R. S. *et al.* Gene for hereditary retinoblastoma assigned to human chromosome 13 by linkage to esterase D. *Science* **219,** 971–973 (1983).

220. Benedict, W. F. *et al.* Patient with 13 chromosome deletion: evidence that the retinoblastoma gene is a recessive cancer gene. *Science* **219,** 973–975 (1983).

221. Coffin, J. M. *et al.* Proposal for naming host cell-derived inserts in retrovirus genomes. *J. Virol.* **40,** 953–957 (1981).

222. Bishop, J. M. Enemies within: the genesis of retrovirus oncogenes. *Cell* **23,** 5–6 (1981).

223. Ellis, R. W. *et al.* The p21 src genes of Harvey and Kirsten sarcoma viruses originate from divergent members of a family of normal vertebrate genes. *Nature* **292,** 506–511 (1981).

224. Cooper, G. M. Cellular transforming genes. *Science* **217,** 801–806 (1982).

225. Reddy, E. P., Reynolds, R. K., Santos, E. & Barbacid, M. A point mutation is responsible for the acquisition of transforming properties by the T24 human bladder carcinoma oncogene. *Nature* **300,** 149–152 (1982).

226. Tabin, C. J. *et al.* Mechanism of activation of a human oncogene. *Nature* **300,** 143–149 (1982).

227. Kirsch, I. R., Morton, C. C., Nakahara, K. & Leder, P. Human immunoglobulin heavy chain genes map to a region of translocations in malignant B lymphocytes. *Science* **216,** 301–303 (1982).

228. Schwab, M. *et al.* Amplified DNA with limited homology to myc cellular oncogene is shared by human neuroblastoma cell lines and a neuroblastoma tumour. *Nature* **305,** 245–248 (1983).

229. Vogelstein, B. & Kinzler, K. W. The multistep nature of cancer. *Trends Genet.* **9,** 138–141 (1993).

230. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U.S.A.* **74,** 5463–5467 (1977).

231. Sanger, F. *et al.* Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* **265,** 687–695 (1977).

232. Fiers, W. *et al.* Complete nucleotide sequence of SV40 DNA. *Nature* **273,** 113–120 (1978).

233. Anderson, S. *et al.* Sequence and organization of the human mitochondrial genome. *Nature* **290,** 457–465 (1981).

234. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).

235. Adams, M. D. *et al.* Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* **252,** 1651–1656 (1991).

236. Waterston, R. H. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420,** 520–562 (2002).

237. Adams, M. D. *et al.* The genome sequence of Drosophila melanogaster. *Science* **287,** 2185–2195 (2000).

238. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).

239. Brooks, J. D. Translational genomics: the challenge of developing cancer biomarkers. *Genome Res.* **22,** 183–187 (2012).

240. Pappa, V. I. *et al.* Use of the polymerase chain reaction and direct sequencing analysis to detect cells with the t(14;18) in autologous bone marrow from patients with follicular lymphoma, before and after in vitro treatment. *Bone Marrow Transplant.* **22,** 553–558 (1998).

241. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270,** 467–470 (1995).

242. Zhang, L. *et al.* Gene expression profiles in normal and cancer cells. *Science* **276,** 1268–1272 (1997).

243. Bhattacharjee, A. *et al.* Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl. Acad. Sci. U.S.A.* **98,** 13790–13795 (2001).

244. Van 't Veer, L. J. *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415,** 530–536 (2002).

245. Parker, J. S. *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* **27,** 1160–1167 (2009).

246. Hu, Z. *et al.* The molecular portraits of breast tumors are conserved across microarray platforms. *BMC Genomics* **7,** 96 (2006).

247. Soerlie, T. *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. U.S.A.* **98,** 10869–10874 (2001).

248. Liu, M. C. *et al.* PAM50 gene signatures and breast cancer prognosis with adjuvant anthracycline- and taxane-based chemotherapy: correlative analysis of C9741 (Alliance). *NPJ Breast Cancer* **2** (2016).

249. Wallden, B. *et al.* Development and verification of the PAM50-based Prosigna breast cancer gene signature assay. *BMC Med Genomics* **8,** 54 (2015).

250. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5,** 621–628 (2008).

251. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10,** 57–63 (2009).

252. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28,** 511–515 (2010).

253. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29,** 644–652 (2011).

254. Kahles, A. *et al.* Comprehensive Analysis of Alternative Splicing Across Tumors from 8,705 Patients. *Cancer Cell* **34,** 211–224 (Aug. 2018).

255. Piskol, R., Ramaswami, G. & Li, J. B. Reliable identification of genomic variants from RNA-seq data. *Am. J. Hum. Genet.* **93,** 641–651 (2013).

256. Edgren, H. *et al.* Identification of fusion genes in breast cancer by paired-end RNA-sequencing. *Genome Biol.* **12,** R6 (2011).

257. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45,** 580–585 (2013).

258. Futreal, P. A. *et al.* A census of human cancer genes. *Nat. Rev. Cancer* **4,** 177–183 (2004).

259. Ding, L. *et al.* Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* **455,** 1069–1075 (2008).

260. Thomas, R. K. *et al.* High-throughput oncogene mutation profiling in human cancer. *Nat. Genet.* **39,** 347–351 (2007).

261. Sjoblom, T. *et al.* The consensus coding sequences of human breast and colorectal cancers. *Science* **314,** 268–274 (2006).

262. Wood, L. D. *et al.* The genomic landscapes of human breast and colorectal cancers. *Science* **318,** 1108–1113 (2007).

263. Jones, S. *et al.* Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science* **321,** 1801–1806 (2008).

264. Parsons, D. W. *et al.* An integrated genomic analysis of human glioblastoma multiforme. *Science* **321,** 1807–1812 (2008).

265. Ronaghi, M., Uhlen, M. & Nyren, P. A sequencing method based on real-time pyrophosphate. *Science* **281,** 363, 365 (1998).

266. Brenner, S. *et al.* Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.* **18,** 630–634 (2000).

267. Kling, J. Ultrafast DNA sequencing. *Nat. Biotechnol.* **21,** 1425–1427 (2003).

268. Shendure, J., Mitra, R. D., Varma, C. & Church, G. M. Advanced sequencing technologies: methods and goals. *Nat. Rev. Genet.* **5,** 335–344 (2004).

269. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437,** 376–380 (2005).

270. Shah, S. P. *et al.* Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature* **461,** 809–813 (2009).

271. Berger, M. F. *et al.* The genomic complexity of primary human prostate cancer. *Nature* **470,** 214–220 (2011).

272. Chapman, M. A. *et al.* Initial genome sequencing and analysis of multiple myeloma. *Nature* **471,** 467–472 (2011).

273. Bass, A. J. *et al.* Genomic sequencing of colorectal adenocarcinomas identifies a recurrent VTI1A-TCF7L2 fusion. *Nat. Genet.* **43,** 964–968 (2011).

274. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339,** 1546–1558 (2013).

275. Hoadley, K. A. *et al.* Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell* **173,** 291–304 (Apr. 2018).

276. Hudson, T. J. *et al.* International network of cancer genome projects. *Nature* **464,** 993–998 (2010).

277. Grossman, R. L. *et al.* Toward a Shared Vision for Cancer Genomic Data. *N. Engl. J. Med.* **375,** 1109–1112 (2016).

278. Jensen, M. A., Ferretti, V., Grossman, R. L. & Staudt, L. M. The NCI Genomic Data Commons as an engine for precision medicine. *Blood* **130,** 453–459 (July 2017).

279. Nik-Zainal, S. *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534,** 47–54 (June 2016).

280. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500,** 415–421 (2013).

281. Horak, P. *et al.* Precision oncology based on omics data: The NCT Heidelberg experience. *Int. J. Cancer* **141,** 877–886 (Sept. 2017).

282. Bailey, M. H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **173,** 371–385 (Apr. 2018).

283. Forbes, S. A. *et al.* COSMIC: High-Resolution Cancer Genetics Using the Catalogue of Somatic Mutations in Cancer. *Curr Protoc Hum Genet* **91,** 1–10 (Oct. 2016).

284. Wala, J. A. *et al.* Selective and mechanistic sources of recurrent rearrangements across the cancer genome. *bioRxiv,* 187609 (2017).

285. Bibikova, M. *et al.* High density DNA methylation array with single CpG site resolution. *Genomics* **98,** 288–295 (2011).

286. Berman, B. P. *et al.* Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. *Nat. Genet.* **44,** 40–46 (2011).

287. Kretzmer, H. *et al.* DNA methylome analysis in Burkitt and follicular lymphomas identifies differentially methylated regions linked to somatic mutation and transcriptional control. *Nat. Genet.* **47,** 1316–1325 (2015).

288. Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448,** 553–560 (2007).

289. Lin, C. Y. *et al.* Active medulloblastoma enhancers reveal subgroup-specific cellular origins. *Nature* **530,** 57–62 (2016).

290. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10,** 1213–1218 (2013).

291. Corces, M. R. *et al.* An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat. Methods* **14,** 959–962 (2017).

292. Corces, M. R. *et al.* The chromatin accessibility landscape of primary human cancers. *Science* **362** (Oct. 2018).

293. Simonis, M. *et al.* Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat. Genet.* **38,** 1348–1354 (2006).

294. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326,** 289–293 (2009).

295. Mumbach, M. R. *et al.* HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat. Methods* **13,** 919–922 (2016).

296. Birney, E. *et al.* Identification and analysis of functional elements in 1of the human genome by the ENCODE pilot project. *Nature* **447,** 799–816 (2007).

297. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).

298. Skipper, M. *et al.* Presenting the epigenome roadmap. *Nature* **518,** 313 (2015).

299. Karczewski, K. J. & Snyder, M. P. Integrative omics for health and disease. *Nat. Rev. Genet.* **19,** 299–310 (May 2018).

300. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).

301. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).

302. Burrows, M. & Wheeler, D. J. A block-sorting lossless data compression algorithm. *Technical report* **124** (1994).

303. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).

304. Jones, D. *et al.* cgpCaVEManWrapper: Simple Execution of CaVEMan in Order to Detect Somatic Single Nucleotide Variants in NGS Data. *Curr Protoc Bioinformatics* **56,** 1–15 (Dec. 2016).

305. Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B. S. & Swanton, C. DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol.* **17,** 31 (2016).

306. Rimmer, A. *et al.* Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.* **46,** 912–918 (2014).

307. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17,** 122 (June 2016).

308. Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28,** i333–i339 (2012).

309. Kleinheinz, K. *et al.* ACEseq allele specific copy number estimation from whole genome sequencing. *bioRxiv,* 210807 (2017).

310. Smyth, G. K. & Speed, T. Normalization of cDNA microarray data. *Methods* **31,** 265–273 (2003).

311. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43,** e47 (2015).

312. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).

313. Baruzzo, G. *et al.* Simulation-based comprehensive benchmarking of RNA-seq aligners. *Nat. Methods* **14,** 135–139 (Feb. 2017).

314. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30,** 923–930 (2014).

315. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11,** R25 (2010).

316. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11,** R106 (2010).

317. Anders, S., Reyes, A. & Huber, W. Detecting differential usage of exons from RNA-seq data. *Genome Res.* **22,** 2008–2017 (2012).

318. McPherson, A. *et al.* deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. *PLoS Comput. Biol.* **7,** e1001138 (2011).

319. Aryee, M. J. *et al.* Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30,** 1363–1369 (2014).

320. Assenov, Y. *et al.* Comprehensive analysis of DNA methylation data with RnBeads. *Nat. Methods* **11,** 1138–1140 (2014).

321. Hovestadt, V. & Zapatka, M. *conumee: Enhanced copy-number variation analysis using Illumina DNA methylation arrays* R package version 1.9.0 (Division of Molecular Genetics German Cancer Research Center (DKFZ) Heidelberg, Germany).

322. Feber, A. *et al.* Using high-density DNA methylation arrays to profile copy number alterations. *Genome Biol.* **15,** R30 (2014).

323. Argelaguet, R. *et al.* Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets. *Mol. Syst. Biol.* **14,** e8124 (June 2018).

324. Shen, R., Olshen, A. B. & Ladanyi, M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* **25,** 2906–2912 (2009).

325. Maaten, L. v. d. & Hinton, G. Visualizing data using t-SNE. *Journal of machine learning research* **9,** 2579–2605 (2008).

326. Campello, R. J., Moulavi, D. & Sander, J. *Density-based clustering based on hierarchical density estimates* in *Pacific-Asia conference on knowledge discovery and data mining* (2013), 160–172.

327. Buitinck, L. *et al. API design for machine learning software: experiences from the scikit-learn project* in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning* (2013), 108–122.

328. Wickham, H. *ggplot2: elegant graphics for data analysis* (Springer, 2016).

329. Bostock, M., Ogievetsky, V. & Heer, J. D3: Data-Driven Documents. *IEEE Trans. Visualization and Comp. Graphics (Proc. InfoVis)* (2011).

330. Yin, T., Cook, D. & Lawrence, M. ggbio: an R package for extending the grammar of graphics for genomic data. *Genome Biol.* **13,** R77 (2012).

331. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32,** 2847–2849 (Sept. 2016).

332. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res.* **19,** 1639–1645 (2009).

333. Wuchty, S., Jones, B. F. & Uzzi, B. The increasing dominance of teams in production of knowledge. *Science* **316,** 1036–1039 (2007).

334. Burgio, M. R. *et al.* Collaborative cancer epidemiology in the 21st century: the model of cancer consortia. *Cancer Epidemiol. Biomarkers Prev.* **22,** 2148–2160 (2013).

335. Savage, N. Collaboration is the key to cancer research. *Nature* **556,** S1–S3 (Apr. 2018).

336. Gao, J. *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal* **6,** pl1 (2013).

337. Koster, J. & Versteeg, R. R2: Genomics analysis and visualization platform. *Available at* `hhts://r2.amc.nl` *Accessed July 2019* **11,** 2017 (2008).

338. Redig, A. J. & Janne, P. A. Basket trials and the evolution of clinical trial design in an era of genomic medicine. *J. Clin. Oncol.* **33,** 975–977 (2015).

339. Roychowdhury, S. *et al.* Personalized oncology through integrative high-throughput sequencing: a pilot study. *Sci Transl Med* **3,** 111ra121 (2011).

340. Robinson, D. R. *et al.* Integrative clinical genomics of metastatic cancer. *Nature* **548,** 297–303 (Aug. 2017).

341. Polak, P. *et al.* Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature* **518,** 360–364 (2015).

342. Fletcher, C. D. *Diagnostic Histopathology of Tumors* (Elsevier Health Sciences, 2007).

343. Ellis, I. *et al.* Pathology reporting of breast disease in surgical excision specimens incorporating the dataset for histological reporting of breast cancer, June 2016. *Published by The Royal College of Pathologists* (2016).

344. Jo, V. Y. & Fletcher, C. D. WHO classification of soft tissue tumours: an update based on the 2013 (4th) edition. *Pathology* **46,** 95–104 (2014).

345. Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C. & Teichmann, S. A. The technology and biology of single-cell RNA sequencing. *Mol. Cell* **58,** 610–620 (2015).

346. Darmanis, S. *et al.* A survey of human brain transcriptome diversity at the single cell level. *Proc. Natl. Acad. Sci. U.S.A.* **112,** 7285–7290 (2015).

347. Louis, D. N. *et al.* The 2016 World Health Organization classification of tumors of the central nervous system: a summary. *Acta Neuropathologica* **131,** 803–820 (2016).

348. Taylor, M. D. *et al.* Molecular subgroups of medulloblastoma: the current consensus. *Acta Neuropathol.* **123,** 465–472 (2012).

349. Thompson, M. C. *et al.* Genomics identifies medulloblastoma subgroups that are enriched for specific genetic alterations. *J. Clin. Oncol.* **24,** 1924–1931 (2006).

350. Northcott, P. A. *et al.* Medulloblastoma comprises four distinct molecular variants. *J. Clin. Oncol.* **29,** 1408–1414 (2011).

351. Kool, M. *et al.* Integrated genomics identifies five medulloblastoma subtypes with distinct genetic profiles, pathway signatures and clinicopathological features. *PLoS ONE* **3,** e3088 (2008).

352. Hovestadt, V. *et al.* Robust molecular subgrouping and copy-number profiling of medulloblastoma from small amounts of archival tumour material using high-density DNA methylation arrays. *Acta Neuropathol.* **125,** 913–916 (2013).

353. Capper, D. *et al.* DNA methylation-based classification of central nervous system tumours. *Nature* **555,** 469–474 (Mar. 2018).

354. Rorke, L. B. The cerebellar medulloblastoma and its relationship to primitive neuroectodermal tumors. *J. Neuropathol. Exp. Neurol.* **42,** 1–15 (1983).

355. Rorke, L. B. *et al.* Primitive neuroectodermal tumors of the central nervous system. *Brain Pathol.* **7,** 765–784 (1997).

356. Fortin, J. P. *et al.* Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome Biol.* **15,** 503 (2014).

357. Jones, D. T. *et al.* Dissecting the genomic complexity underlying medulloblastoma. *Nature* **488,** 100–105 (2012).

358. Parkhomchuk, D. *et al.* Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Res.* **37,** e123 (2009).

359. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997* (2013).

360. Tischler, G. & Leonard, S. biobambam: tools for read pair collation based algorithms on BAM files. *Source Code for Biology and Medicine* **9,** 13 (2014).

361. Yung, C. K. *et al.* Large-Scale Uniform Analysis of Cancer Whole Genomes in Multiple Computing Environments. *bioRxiv* (2017).

362. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27,** 2987–2993 (2011).

363. Li, H. Improving SNP discovery by base alignment quality. *Bioinformatics* **27,** 1157–1158 (2011).

364. Nakamura, K. *et al.* Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res.* **39,** e90 (2011).

365. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29,** 24–26 (2011).

366. Wang, J. *et al.* CREST maps somatic structural variation in cancer genomes with base-pair resolution. *Nat. Methods* **8,** 652–654 (2011).

367. Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.* **15,** R84 (2014).

368. Korshunov, A. *et al.* Embryonal tumor with abundant neuropil and true rosettes (ETANTR), ependymoblastoma, and medulloepithelioma share molecular similarity and comprise a single clinicopathological entity. *Acta Neuropathol.* **128,** 279–289 (2014).

369. Spence, T. *et al.* CNS-PNETs with C19MC amplification and/or LIN28 expression comprise a distinct histogenetic diagnostic and therapeutic entity. *Acta Neuropathol.* **128,** 291–303 (2014).

370. Korshunov, A. *et al.* H3-/IDH-wild type pediatric glioblastoma is comprised of molecularly and prognostically distinct subtypes with associated oncogenic drivers. *Acta Neuropathol.* **134,** 507–516 (2017).

371. Biegel, J. A. *et al.* The role of INI1 and the SWI/SNF complex in the development of rhabdoid tumors: meeting summary from the workshop on childhood atypical teratoid/rhabdoid tumors. *Cancer Res.* **62,** 323–328 (2002).

372. Bailey, P. & Bucy, P. C. ASTROBLASTOMAS OF THE BRAIN. *Acta Psychiatrica Scandinavica* **5,** 439–461 (1930).

373. Navarro, R. *et al.* Astroblastoma in childhood: pathological and clinical analysis. *Childs Nerv Syst* **21,** 211–220 (2005).

374. Komori, T. The 2016 WHO Classification of Tumours of the Central Nervous System: The Major Points of Revision. *Neurol. Med. Chir. (Tokyo)* **57,** 301–311 (2017).

375. Lekanne Deprez, R. H. *et al.* Cloning and characterization of MN1, a gene from chromosome 22q11, which is disrupted by a balanced translocation in a meningioma. *Oncogene* **10,** 1521–1528 (1995).

376. Buijs, A. *et al.* Translocation (12;22) (p13;q11) in myeloproliferative disorders results in fusion of the ETS-like TEL gene on 12p13 to the MN1 gene on 22q11. *Oncogene* **10,** 1511–1519 (1995).

377. Buijs, A. *et al.* The MN1-TEL fusion protein, encoded by the translocation (12;22)(p13;q11) in myeloid leukemia, is a transcription factor with transforming activity. *Mol. Cell. Biol.* **20,** 9281–9293 (2000).

378. Van Wely, K. H. *et al.* The MN1-TEL myeloid leukemia-associated fusion protein has a dominant-negative effect on RAR-RXR-mediated transcription. *Oncogene* **26,** 5733–5740 (2007).

379. Abhiman, S., Iyer, L. M. & Aravind, L. BEN: a novel domain in chromatin factors and DNA viral proteins. *Bioinformatics* **24,** 458–461 (2008).

380. Haidar, A., Arekapudi, S., DeMattia, F., Abu-Isa, E. & Kraut, M. High-grade undifferentiated small round cell sarcoma with t(4;19)(q35;q13.1) CIC-DUX4 fusion: emerging entities of soft tissue tumors with unique histopathologic features–a case report and literature review. *Am J Case Rep* **16,** 87–94 (2015).

381. Specht, K. *et al.* Distinct transcriptional signature and immunoprofile of CIC-DUX4 fusion-positive round cell tumors compared to EWSR1-rearranged Ewing sarcomas: further evidence toward distinct pathologic entities. *Genes Chromosomes Cancer* **53,** 622–633 (2014).

382. French, C. NUT midline carcinoma. *Nat. Rev. Cancer* **14,** 149–150 (Mar. 2014).

383. Kanehisa, M., Goto, S., Kawashima, S. & Nakaya, A. The KEGG databases at GenomeNet. *Nucleic Acids Res.* **30,** 42–46 (2002).

384. Marchler-Bauer, A. *et al.* CDD: conserved domains and protein three-dimensional structure. *Nucleic Acids Res.* **41,** D348–352 (2013).

385. Finn, R. D. *et al.* Pfam: the protein families database. *Nucleic Acids Res.* **42,** D222–230 (2014).

386. Junco, S. E. *et al.* Structure of the polycomb group protein PCGF1 in complex with BCOR reveals basis for binding selectivity of PCGF homologs. *Structure* **21,** 665–671 (2013).

387. Appay, R. *et al.* HGNET-BCOR Tumors of the Cerebellum: Clinicopathologic and Molecular Characterization of 3 Cases. *Am. J. Surg. Pathol.* **41,** 1254–1260 (2017).

388. Yoshida, Y. *et al.* CNS high-grade neuroepithelial tumor with BCOR internal tandem duplication: a comparison with its counterparts in the kidney and soft tissue. *Brain Pathol.* **28,** 710–720 (Sept. 2018).

389. Roy, A. *et al.* Recurrent internal tandem duplications of BCOR in clear cell sarcoma of the kidney. *Nat Commun* **6,** 8891 (2015).

390. Marino-Enriquez, A. *et al.* BCOR Internal Tandem Duplication in High-grade Uterine Sarcomas. *Am. J. Surg. Pathol.* **42,** 335–341 (Mar. 2018).

391. Kimura, K. *et al.* Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes. *Genome Res.* **16,** 55–65 (2006).

392. Thierry-Mieg, D. & Thierry-Mieg, J. AceView: a comprehensive cDNA-supported gene and transcripts annotation. *Genome Biol.* **7 Suppl 1,** 1–14 (2006).

393. Korbel, J. O. & Campbell, P. J. Criteria for inference of chromothripsis in cancer genomes. *Cell* **152,** 1226–1236 (2013).

394. Zhang, Y. *et al.* A Pan-Cancer Compendium of Genes Deregulated by Somatic Genomic Rearrangement across More Than 1,400 Cases. *Cell Rep* **24,** 515–527 (2018).

395. Fu, W., Begley, J. G., Killen, M. W. & Mattson, M. P. Anti-apoptotic role of telomerase in pheochromocytoma cells. *J. Biol. Chem.* **274,** 7264–7271 (1999).

396. Li, X. *et al.* FOXR2 Interacts with MYC to Promote Its Transcriptional Activities and Tumorigenesis. *Cell Rep* **16,** 487–497 (July 2016).

397. Hwang, E. I. *et al.* Extensive Molecular and Clinical Heterogeneity in Patients With Histologically Diagnosed CNS-PNET Treated as a Single Entity: A Report From the Children's Oncology Group Randomized ACNS0332 Trial. *J. Clin. Oncol.* JCO2017764720 (2018).

398. Paret, C. *et al.* Personalized therapy: CNS HGNET-BCOR responsiveness to arsenic trioxide combined with radiotherapy. *Oncotarget* **8,** 114210–114225 (2017).

399. Currall, B. B., Chiang, C., Talkowski, M. E. & Morton, C. C. Mechanisms for Structural Variation in the Human Genome. *Curr Genet Med Rep* **1,** 81–90 (2013).

400. Kearney, L. Molecular cytogenetics. *Best Pract Res Clin Haematol* **14,** 645–669 (2001).

401. Pinkel, D. *et al.* Fluorescence in situ hybridization with human chromosome-specific libraries: detection of trisomy 21 and translocations of chromosome 4. *Proc. Natl. Acad. Sci. U.S.A.* **85,** 9138–9142 (1988).

402. Speicher, M. R. & Carter, N. P. The new cytogenetics: blurring the boundaries with molecular biology. *Nat. Rev. Genet.* **6,** 782–792 (2005).

403. Beroukhim, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463,** 899–905 (2010).

404. Li, Y. *et al.* Patterns of structural variation in human cancer. *bioRxiv,* 181339 (2017).

405. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456,** 53–59 (2008).

406. Williams, L. J. *et al.* Paired-end sequencing of Fosmid libraries by Illumina. *Genome Res.* **22,** 2241–2249 (2012).

407. Richard, G. F., Kerrest, A. & Dujon, B. Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. *Microbiol. Mol. Biol. Rev.* **72,** 686–727 (2008).

408. Batzer, M. A. & Deininger, P. L. Alu repeats and human genomic diversity. *Nat. Rev. Genet.* **3,** 370–379 (2002).

409. Subramanian, S., Mishra, R. K. & Singh, L. Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. *Genome Biol.* **4,** R13 (2003).

410. Dumbovic, G., Forcales, S. V. & Perucho, M. Emerging roles of macrosatellite repeats in genome organization and disease development. *Epigenetics* **12,** 515–526 (July 2017).

411. Warburton, P. E. *et al.* Analysis of the largest tandemly repeated DNA families in the human genome. *BMC Genomics* **9,** 533 (2008).

412. Usdin, K. The biological effects of simple tandem repeats: lessons from the repeat expansion diseases. *Genome Res.* **18,** 1011–1019 (2008).

413. Djian, P. Evolution of simple repeats in DNA and their relation to human disease. *Cell* **94,** 155–160 (1998).

414. Britten, R. J. Transposable element insertions have strongly affected human evolution. *Proc. Natl. Acad. Sci. U.S.A.* **107,** 19945–19948 (2010).

415. Treangen, T. J. & Salzberg, S. L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet.* **13,** 36–46 (2011).

416. Helena Mangs, A. & Morris, B. J. The Human Pseudoautosomal Region (PAR): Origin, Function and Future. *Curr. Genomics* **8,** 129–136 (2007).

417. Chen, K. *et al.* BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods* **6,** 677–681 (2009).

418. Huang, X. & Madan, A. CAP3: A DNA sequence assembly program. *Genome Res.* **9,** 868–877 (1999).

419. Kent, W. J. BLAT–the BLAST-like alignment tool. *Genome Res.* **12,** 656–664 (2002).

420. Gerhauser, C. *et al.* Molecular Evolution of Early-Onset Prostate Cancer Identifies Molecular Risk Markers and Clinical Trajectories. *Cancer Cell* **34,** 996–1011 (2018).

421. Wu, Y. M. *et al.* Inactivation of CDK12 Delineates a Distinct Immunogenic Class of Advanced Prostate Cancer. *Cell* **173,** 1770–1782 (June 2018).

422. Gunnarsson, R. *et al.* Mutation, methylation, and gene expression profiles in dup(1q)-positive pediatric B-cell precursor acute lymphoblastic leukemia. *Leukemia* **32,** 2117–2125 (Oct. 2018).

423. Quigley, D. A. *et al.* Genomic Hallmarks and Structural Variation in Metastatic Prostate Cancer. *Cell* **174,** 758–769 (July 2018).

424. Arthur, S. E. *et al.* Genome-wide discovery of somatic regulatory variants in diffuse large B-cell lymphoma. *Nat Commun* **9,** 4001 (Oct. 2018).

425. Chong, Z. *et al.* novoBreak: local assembly for breakpoint detection in cancer genomes. *Nat. Methods* **14,** 65–67 (Jan. 2017).

426. Wala, J. A. *et al.* SvABA: genome-wide detection of structural variants and indels by local assembly. *Genome Res.* **28,** 581–591 (Apr. 2018).

427. Zook, J. M. *et al.* Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci Data* **3,** 160025 (2016).

428. Xia, Y., Liu, Y., Deng, M. & Xi, R. Pysim-sv: a package for simulating structural variation data with GC-biases. *BMC Bioinformatics* **18,** 53 (2017).

429. Xia, L. C. *et al.* SVEngine: an efficient and versatile simulator of genome structural variations with features of cancer clonal evolution. *Gigascience* **7** (July 2018).

430. Jeffares, D. C. *et al.* Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat Commun* **8,** 14061 (Jan. 2017).

431. Whalley, J. P. *et al.* Framework for quality assessment of whole genome, cancer sequences. *bioRxiv,* 140921 (2017).

432. Bassing, C. H., Swat, W. & Alt, F. W. The mechanism and regulation of chromosomal V(D)J recombination. *Cell* **109 Suppl,** 45–55 (2002).

433. Stavnezer, J., Guikema, J. E. & Schrader, C. E. Mechanism and regulation of class switch recombination. *Annu. Rev. Immunol.* **26,** 261–292 (2008).

434. Delgado, J. *et al.* Genomic complexity and IGHV mutational status are key predictors of outcome of chronic lymphocytic leukemia patients with TP53 disruption. *Haematologica* **99,** e231–234 (2014).

435. Lopez, C. *et al.* Genomic and transcriptomic changes complement each other in the pathogenesis of sporadic Burkitt lymphoma. *Nat Commun* **10,** 1459 (Mar. 2019).

436. Alt, F. W. *et al.* VDJ recombination. *Immunol. Today* **13,** 306–314 (1992).

437. Mackinnon, R. N. & Chudoba, I. The use of M-FISH and M-BAND to define chromosome abnormalities. *Methods Mol. Biol.* **730,** 203–218 (2011).

438. Fernandez de Larrea, C. *et al.* Plasma cell leukemia: consensus statement on diagnostic requirements, response criteria and treatment recommendations by the International Myeloma Working Group. *Leukemia* **27,** 780–791 (2013).

439. Taylor-Weiner, A. *et al.* DeTiN: overcoming tumor-in-normal contamination. *Nat. Methods* **15,** 531–534 (July 2018).

440. Iehara, T. *et al.* MYCN gene amplification is a powerful prognostic factor even in infantile neuroblastoma detected by mass screening. *Br. J. Cancer* **94,** 1510–1515 (2006).

441. Buckley, R. M., Kortschak, R. D., Raison, J. M. & Adelson, D. L. Similar Evolutionary Trajectories for Retrotransposon Accumulation in Mammals. *Genome Biol Evol* **9,** 2336–2353 (Sept. 2017).

442. Abel, H. J. *et al.* Mapping and characterization of structural variation in 17,795 deeply sequenced human genomes. *bioRxiv,* 508515 (2018).

443. Rao, S. S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159,** 1665–1680 (2014).

444. Dixon, J. R. *et al.* Chromatin architecture reorganization during stem cell differentiation. *Nature* **518,** 331–336 (2015).

445. Leung, D. *et al.* Integrative analysis of haplotype-resolved epigenomes across human tissues. *Nature* **518,** 350–354 (2015).

446. Schmitt, A. D. *et al.* A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. *Cell Rep* **17,** 2042–2059 (Nov. 2016).

447. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22,** 1760–1774 (2012).

448. Rodriguez, J. M. *et al.* APPRIS: annotation of principal and alternative splice isoforms. *Nucleic Acids Res.* **41,** D110–117 (2013).

449. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).

450. Germain, P. L. *et al.* RNAontheBENCH: computational and empirical resources for benchmarking RNAseq quantification and differential expression methods. *Nucleic Acids Res.* **44,** 5054–5067 (June 2016).

451. Lee, S. *et al.* Accurate quantification of transcriptome from RNA-Seq data by effective length normalization. *Nucleic Acids Res.* **39,** e9 (2011).

452. Evans, C., Hardin, J. & Stoebel, D. M. Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Brief. Bioinformatics* **19,** 776–792 (Sept. 2018).

453. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–140 (2010).

454. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15,** 550 (2014).

455. Goldman, M *et al.* The UCSC Xena platform for public and private cancer genomics data visualization and interpretation. *bioRxiv* (2019).

456. Reisinger, E. *et al.* OTP: An automatized system for managing and processing NGS data. *J. Biotechnol.* **261,** 53–62 (2017).

457. Hubbell, E., Liu, W. M. & Mei, R. Robust estimators for expression analysis. *Bioinformatics* **18,** 1585–1592 (2002).

458. Vergier, B. *et al.* Combined analysis of T cell receptor gamma and immunoglobulin heavy chain gene rearrangements at the single-cell level in lymphomas with dual genotype. *J. Pathol.* **198,** 171–180 (2002).

459. Leber, B. F., Amlot, P., Hoffbrand, A. V. & Norton, J. D. T-cell receptor gene rearrangement in B-cell non-Hodgkin's lymphoma: correlation with methylation and expression. *Leuk. Res.* **13,** 473–481 (1989).

460. Salaverria, I. *et al.* Translocations activating IRF4 identify a subtype of germinal center-derived B-cell lymphoma affecting predominantly children and young adults. *Blood* **118,** 139–147 (2011).

461. Santo, E. E. *et al.* Oncogenic activation of FOXR1 by 11q23 intrachromosomal deletion-fusions in neuroblastoma. *Oncogene* **31,** 1571–1581 (2012).

462. Korber, V. *et al.* Evolutionary Trajectories of IDHWT Glioblastomas Reveal a Common Path of Early Tumorigenesis Instigated Years ahead of Initial Diagnosis. *Cancer Cell* **35,** 692–704 (2019).

463. Sahm, F. *et al.* Meningiomas induced by low-dose radiation carry structural variants of NF2 and a distinct mutational signature. *Acta Neuropathol.* **134,** 155–158 (July 2017).

464. Paramasivam, N. *et al.* Mutational patterns and regulatory networks in epigenetic subgroups of meningioma. *Acta Neuropathol.* (2019).

465. Wagener, R. *et al.* Cryptic insertion of MYC exons 2 and 3 into the IGH locus detected by whole genome sequencing in a case of MYC-negative Burkitt lymphoma. *Haematologica* (2019).

466. Fan, X., Abbott, T. E., Larson, D. & Chen, K. BreakDancer: Identification of Genomic Structural Variation from Paired-End Read Mapping. *Curr Protoc Bioinformatics* **45,** 1–11 (2014).

467. Lee, A. Y. *et al.* Combining accurate tumor genome simulation with crowdsourcing to benchmark somatic structural variant detection. *Genome Biol.* **19,** 188 (Nov. 2018).

468. Poplin, R. *et al.* A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **36,** 983–987 (Nov. 2018).

469. Cai, L., Wu, Y. & Gao, J. DeepSV: Accurate calling of genomic deletions from high throughput sequencing data using deep convolutional neural network. *bioRxiv* (2019).

470. Breiman, L. Random Forests. *Machine Learning* **45,** 5–32 (2001).

471. Weinstein, J. N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45,** 1113–1120 (2013).

472. Perkel, J. M. Data visualization tools drive interactivity and reproducibility in online publishing. *Nature* **554,** 133–134 (Feb. 2018).

473. Klonowska, K. *et al.* Oncogenomic portals for the visualization and analysis of genome-wide cancer data. *Oncotarget* **7,** 176–192 (2016).

474. Schroeder, M. P., Gonzalez-Perez, A. & Lopez-Bigas, N. Visualizing multidimensional cancer genomics data. *Genome Med* **5,** 9 (2013).

475. Goldman, M. *et al.* Online resources for PCAWG data exploration, visualization, and discovery. *bioRxiv,* 163907 (2018).

476. Grobner, S. N. *et al.* The landscape of genomic alterations across childhood cancers. *Nature* **555,** 321–327 (Mar. 2018).

477. Cerami, E. *et al.* The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov* **2,** 401–404 (2012).

478. Sinha, A. U. & Armstrong, S. A. iCanPlot: visual exploration of high-throughput omics data using interactive Canvas plotting. *PLoS ONE* **7,** e31690 (2012).

479. Samur, M. K. *et al.* canEvolve: a web portal for integrative oncogenomics. *PLoS ONE* **8,** e56228 (2013).

480. Leiserson, M. D. *et al.* MAGI: visualization and collaborative annotation of genomic aberrations. *Nat. Methods* **12,** 483–484 (2015).

481. Newton, Y. *et al.* TumorMap: Exploring the Molecular Similarities of Cancer Samples in an Interactive Portal. *Cancer Res.* **77,** e111–e114 (Nov. 2017).

482. McFerrin, L. G. *et al.* Analysis and visualization of linked molecular and clinical cancer data by using Oncoscape. *Nat. Genet.* **50,** 1203–1204 (Sept. 2018).

483. Tyner, J. W. *et al.* Functional genomic landscape of acute myeloid leukaemia. *Nature* **562,** 526–531 (Oct. 2018).

484. Cui, Y. *et al.* BioCircos.js: an interactive Circos JavaScript library for biological data visualization on web applications. *Bioinformatics* **32,** 1740–1742 (June 2016).

485. Bostock, M. & Heer, J. Protovis: a Graphical Toolkit for Visualization. *IEEE Trans. Visualization and Comp. Graphics (Proc. InfoVis)* (2009).

486. Satyanarayan, A., Moritz, D., Wongsuphasawat, K. & Heer, J. Vega-Lite: a Grammar of Interactive Graphics. *IEEE Trans. Visualization and Comp. Graphics (Proc. InfoVis)* (2017).

487. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38,** e164 (2010).

488. Zhou, W., Laird, P. W. & Shen, H. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. *Nucleic Acids Res.* **45,** e22 (Feb. 2017).

489. Rebhan, M., Chalifa-Caspi, V., Prilusky, J. & Lancet, D. GeneCards: integrating information about genes, proteins and diseases. *Trends Genet.* **13,** 163 (1997).

490. Safran, M. *et al.* GeneCards Version 3: the human gene integrator. *Database (Oxford)* **2010,** baq020 (2010).

491. Sayers, E. W. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **38,** 5–16 (2010).

492. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12,** 996–1006 (2002).

493. Khan, A. & Zhang, X. dbSUPER: a database of super-enhancers in mouse and human genome. *Nucleic Acids Res.* **44,** D164–171 (2016).

494. McGinnis, S. & Madden, T. L. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res.* **32,** W20–25 (2004).

495. Huang, D. W. *et al.* The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biol.* **8,** R183 (2007).

496. Fabregat, A. *et al.* The Reactome pathway Knowledgebase. *Nucleic Acids Res.* **44,** D481–487 (2016).

497. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* **102,** 15545–15550 (2005).

498. Kamburov, A., Wierling, C., Lehrach, H. & Herwig, R. ConsensusPathDB–a database for integrating human functional interaction networks. *Nucleic Acids Res.* **37,** D623–628 (2009).

499. Li, W. Volcano plots in analyzing differential expressions with mRNA microarrays. *J Bioinform Comput Biol* **10,** 1231003 (2012).

500. Jenks, G. The Data Model Concept in Statistical Mapping. *International Yearbook of Cartography* **7,** 186–190 (Jan. 1967).

501. Verhaak, R. G. *et al.* Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* **17,** 98–110 (2010).

502. Chen, Y. *et al.* Oncogenic mutations of ALK kinase in neuroblastoma. *Nature* **455,** 971–974 (2008).

503. Paige, A. J. Redefining tumour suppressor genes: exceptions to the two-hit hypothesis. *Cell. Mol. Life Sci.* **60,** 2147–2163 (2003).

504. Rivlin, N., Brosh, R., Oren, M. & Rotter, V. Mutations in the p53 Tumor Suppressor Gene: Important Milestones at the Various Steps of Tumorigenesis. *Genes Cancer* **2,** 466–474 (2011).

505. Cowey, C. L. & Rathmell, W. K. VHL gene mutations in renal cell carcinoma: role as a biomarker of disease outcome and drug efficacy. *Curr Oncol Rep* **11,** 94–101 (2009).

506. Ackermann, S. *et al.* A mechanistic classification of clinical phenotypes in neuroblastoma. *Science* **362,** 1165–1170 (Dec. 2018).

507. Schubert, D. *et al.* Activin is a nerve cell survival molecule. *Nature* **344,** 868–870 (1990).

508. Makanji, Y. *et al.* Inhibin at 90: from discovery to clinical application, a historical review. *Endocr. Rev.* **35,** 747–794 (2014).

509. Dong, C., Wilhelm, D. & Koopman, P. Sox genes and cancer. *Cytogenet. Genome Res.* **105,** 442–447 (2004).

510. Castillo, S. D. & Sanchez-Cespedes, M. The SOX family of genes in cancer development: biological relevance and opportunities for therapy. *Expert Opin. Ther. Targets* **16,** 903–919 (2012).

511. Bass, A. J. *et al.* Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* **513,** 202–209 (2014).

512. Yan, M. *et al.* HER2 expression status in diverse cancers: review of results from 37,992 patients. *Cancer Metastasis Rev.* **34,** 157–164 (2015).

513. Jung, E. *et al.* Tweety-Homolog 1 Drives Brain Colonization of Gliomas. *J. Neurosci.* **37,** 6837–6850 (July 2017).

514. Chan, J. Y., Ong, C. W. & Salto-Tellez, M. Overexpression of neurone glial-related cell adhesion molecule is an independent predictor of poor prognosis in advanced colorectal cancer. *Cancer Sci.* **102,** 1855–1861 (2011).

515. Liu, J. B. *et al.* Chemo-resistant Gastric Cancer Associated Gene Expression Signature: Bioinformatics Analysis Based on Gene Expression Omnibus. *Anticancer Res.* **39,** 1689–1698 (2019).

516. Robertson, A. G. *et al.* Comprehensive Molecular Characterization of Muscle-Invasive Bladder Cancer. *Cell* **171,** 540–556 (2017).

517. Qvigstad, G., Sandvik, A. K., Brenna, E., Aase, S. & Waldum, H. L. Detection of chromogranin A in human gastric adenocarcinomas using a sensitive immunohistochemical technique. *Histochem. J.* **32,** 551–556 (2000).

518. Zhang, T. *et al.* Prognostic role of neuroendocrine cell differentiation in human gastric carcinoma. *Int J Clin Exp Med* **8,** 7837–7842 (2015).

519. Hayakawa, Y. *et al.* Nerve Growth Factor Promotes Gastric Tumorigenesis through Aberrant Cholinergic Signaling. *Cancer Cell* **31,** 21–34 (Jan. 2017).

520. Zhao, C. M. *et al.* Denervation suppresses gastric tumorigenesis. *Sci Transl Med* **6,** 250ra115 (2014).

521. Anderson, W. F., Katki, H. A. & Rosenberg, P. S. Incidence of breast cancer in the United States: current and future trends. *J. Natl. Cancer Inst.* **103,** 1397–1402 (2011).

522. Koboldt, D. C. *et al.* Comprehensive molecular portraits of human breast tumours. *Nature* **490,** 61–70 (2012).

523. Platzer, A. Visualization of SNPs with t-SNE. *PLoS ONE* **8,** e56883 (2013).

524. Wattenberg, M., Viegas, F. & Johnson, I. How to use t-SNE effectively. *Distill* **1,** e2 (2016).

525. Kobak, D., Linderman, G., Steinerberger, S., Kluger, Y. & Berens, P. Heavy-tailed kernels reveal a finer cluster structure in t-SNE visualisations. *arXiv preprint arXiv:1902.05804* (2019).

526. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

527. Becht, E. *et al.* Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* (2018).

528. Ester, M., Kriegel, H.-P., Sander, J., Xu, X., *et al. A density-based algorithm for discovering clusters in large spatial databases with noise.* in *Kdd* **96** (1996), 226–231.

529. Wu, L. & Levine, A. J. Differential regulation of the p21/WAF-1 and mdm2 genes after high-dose UV irradiation: p53-dependent and p53-independent regulation of the mdm2 gene. *Mol. Med.* **3,** 441–451 (1997).

530. Jones, N. *et al.* Comprehensive analysis of PTEN status in breast carcinomas. *Int. J. Cancer* **133,** 323–334 (2013).

531. Yu, W., Kanaan, Y., Bae, Y. K., Baed, Y. K. & Gabrielson, E. Chromosomal changes in aggressive breast cancers with basal-like features. *Cancer Genet. Cytogenet.* **193,** 29–37 (2009).

532. Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature* **406,** 747–752 (2000).

533. Guler, G. *et al.* The fragile genes FHIT and WWOX are inactivated coordinately in invasive breast carcinoma. *Cancer* **100,** 1605–1614 (2004).

534. Liu, J. *et al.* An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell* **173,** 400–416 (Apr. 2018).

535. Seal, M. D. & Chia, S. K. What is the difference between triple-negative and basal breast cancers? *Cancer J* **16,** 12–16 (2010).

536. Ray, P. S. *et al.* FOXC1 is a potential prognostic biomarker with functional significance in basal-like breast cancer. *Cancer Res.* **70,** 3870–3876 (2010).

537. Carroll, J. S. *et al.* Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1. *Cell* **122,** 33–43 (2005).

538. Badve, S. *et al.* FOXA1 expression in breast cancer–correlation with luminal subtype A and survival. *Clin. Cancer Res.* **13,** 4415–4421 (2007).

539. Bernardo, G. M. *et al.* FOXA1 represses the molecular phenotype of basal breast cancer cells. *Oncogene* **32,** 554–563 (2013).

540. Yu-Rice, Y. *et al.* FOXC1 is involved in ERÎ± silencing by counteracting GATA3 binding and is implicated in endocrine resistance. *Oncogene* **35,** 5400–5411 (Oct. 2016).

541. Theodorou, V., Stark, R., Menon, S. & Carroll, J. S. GATA3 acts upstream of FOXA1 in mediating ESR1 binding by shaping enhancer accessibility. *Genome Res.* **23,** 12–22 (2013).

542. Sweeney, C. *et al.* Intrinsic subtypes from PAM50 gene expression assay in a population-based breast cancer cohort: differences by age, race, and tumor characteristics. *Cancer Epidemiol. Biomarkers Prev.* **23,** 714–724 (2014).

543. Griffin, N. M. *et al.* Label-free, normalized quantification of complex mass spectrometry data for proteomic analysis. *Nat. Biotechnol.* **28,** 83–89 (2010).

544. Deng, G., Yang, J., Zhang, Q., Xiao, Z. X. & Cai, H. MethCNA: a database for integrating genomic and epigenomic data in human cancer. *BMC Genomics* **19,** 138 (Feb. 2018).

545. Kyle, R. A. & Rajkumar, S. V. Multiple myeloma. *N. Engl. J. Med.* **351,** 1860–1873 (2004).

546. Roth, K. *et al.* Tracking plasma cell differentiation and survival. *Cytometry A* **85,** 15–24 (2014).

547. Landgren, O. *et al.* Monoclonal gammopathy of undetermined significance (MGUS) consistently precedes multiple myeloma: a prospective study. *Blood* **113,** 5412–5417 (2009).

548. Kyle, R. A. *et al.* Clinical course and prognosis of smoldering (asymptomatic) multiple myeloma. *N. Engl. J. Med.* **356,** 2582–2590 (2007).

549. Bolli, N. *et al.* Genomic patterns of progression in smoldering multiple myeloma. *Nat Commun* **9,** 3363 (Aug. 2018).

550. Blade, J. *et al.* Soft-tissue plasmacytomas in multiple myeloma: incidence, mechanisms of extramedullary spread, and treatment approach. *J. Clin. Oncol.* **29,** 3805–3812 (2011).

551. Albarracin, F. & Fonseca, R. Plasma cell leukemia. *Blood Rev.* **25,** 107–112 (2011).

552. Palumbo, A. & Anderson, K. Multiple myeloma. *N. Engl. J. Med.* **364,** 1046–1060 (2011).

553. Bergsagel, P. L. & Kuehl, W. M. Molecular pathogenesis and a consequent classification of multiple myeloma. *J. Clin. Oncol.* **23,** 6333–6338 (2005).

554. Morgan, G. J., Walker, B. A. & Davies, F. E. The genetic architecture of multiple myeloma. *Nat. Rev. Cancer* **12,** 335–348 (2012).

555. Manier, S. *et al.* Genomic complexity of multiple myeloma and its clinical implications. *Nat Rev Clin Oncol* **14,** 100–113 (Feb. 2017).

556. Walker, B. A. *et al.* Identification of novel mutational drivers reveals oncogene dependencies in multiple myeloma. *Blood* **132,** 587–597 (Aug. 2018).

557. Bergsagel, P. L. *et al.* Promiscuous translocations into immunoglobulin heavy chain switch regions in multiple myeloma. *Proc. Natl. Acad. Sci. U.S.A.* **93,** 13931–13936 (1996).

558. Smadja, N. V. *et al.* Chromosomal analysis in multiple myeloma: cytogenetic evidence of two different diseases. *Leukemia* **12,** 960–969 (1998).

559. Shou, Y. *et al.* Diverse karyotypic abnormalities of the c-myc locus associated with c-myc dysregulation and tumor progression in multiple myeloma. *Proc. Natl. Acad. Sci. U.S.A.* **97,** 228–233 (2000).

560. Neri, A. *et al.* Ras oncogene mutation in multiple myeloma. *J. Exp. Med.* **170,** 1715–1725 (1989).

561. Andrulis, M. *et al.* Targeting the BRAF V600E mutation in multiple myeloma. *Cancer Discov* **3,** 862–869 (2013).

562. Chng, W. J. *et al.* Clinical significance of TP53 mutation in myeloma. *Leukemia* **21,** 582–584 (2007).

563. Zhu, Y. X. *et al.* Loss of FAM46C Promotes Cell Survival in Myeloma. *Cancer Res.* **77,** 4317–4327 (Aug. 2017).

564. Annunziata, C. M. *et al.* Frequent engagement of the classical and alternative NF-kappaB pathways by diverse genetic abnormalities in multiple myeloma. *Cancer Cell* **12,** 115–130 (2007).

565. San Miguel, J. F. *et al.* Bortezomib plus melphalan and prednisone for initial treatment of multiple myeloma. *N. Engl. J. Med.* **359,** 906–917 (2008).

566. Siegel, D. S. *et al.* A phase 2 study of single-agent carfilzomib (PX-171-003-A1) in patients with relapsed and refractory multiple myeloma. *Blood* **120,** 2817–2825 (2012).

567. Singhal, S. *et al.* Antitumor activity of thalidomide in refractory multiple myeloma. *N. Engl. J. Med.* **341,** 1565–1571 (1999).

568. Palumbo, A. *et al.* Continuous lenalidomide treatment for newly diagnosed multiple myeloma. *N. Engl. J. Med.* **366,** 1759–1769 (2012).

569. Miguel, J. S. *et al.* Pomalidomide plus low-dose dexamethasone versus high-dose dexamethasone alone for patients with relapsed and refractory multiple myeloma (MM-003): a randomised, open-label, phase 3 trial. *Lancet Oncol.* **14,** 1055–1066 (2013).

570. De Weers, M. *et al.* Daratumumab, a novel therapeutic human CD38 monoclonal antibody, induces killing of multiple myeloma and other hematological tumors. *J. Immunol.* **186,** 1840–1848 (2011).

571. Palumbo, A. *et al.* Autologous transplantation and maintenance therapy in multiple myeloma. *N. Engl. J. Med.* **371,** 895–905 (2014).

572. Kumar, S. K. *et al.* Improved survival in multiple myeloma and the impact of novel therapies. *Blood* **111,** 2516–2520 (2008).

573. Costa, L. J. *et al.* Recent trends in multiple myeloma incidence and survival by age, race, and ethnicity in the United States. *Blood Adv* **1,** 282–287 (2017).

574. Bodker, J. S. *et al.* A multiple myeloma classification system that associates normal B-cell subset phenotypes with prognosis. *Blood Adv* **2,** 2400–2411 (Sept. 2018).

575. Nooka, A. K., Kastritis, E., Dimopoulos, M. A. & Lonial, S. Treatment options for relapsed and refractory multiple myeloma. *Blood* **125,** 3085–3099 (2015).

576. Pommerenke, C. *et al.* Chromosome 11q23 aberrations activating FOXR1 in B-cell lymphoma. *Blood Cancer J* **6,** e433 (June 2016).

577. Wang, D. *et al.* MAX is an epigenetic sensor of 5-carboxylcytosine and is altered in multiple myeloma. *Nucleic Acids Res.* **45,** 2396–2407 (Mar. 2017).

578. Tong, A. W. & Stone, M. J. CD40 and the effect of anti-CD40-binding on human multiple myeloma clonogenicity. *Leuk. Lymphoma* **21,** 1–8 (1996).

579. Tai, Y. T. *et al.* CD40 induces human multiple myeloma cell migration via phosphatidylinositol 3-kinase/AKT/NF-kappa B signaling. *Blood* **101,** 2762–2769 (2003).

580. Wenthe, J, Naseri, S, Hellstroem, A., Eriksson, E & Loskog, A. 1176P Oncolytic virotherapy for multiple myeloma targeting CD40, 41BB and/or IL6R. *Annals of Oncology* **29,** mdy288–049 (2018).

581. Kim, S. C. *et al.* Identification of a Novel Fusion Gene, FAM174A-WWC1, in Early-Onset Colorectal Cancer: Establishment and Characterization of Four Human Cancer Cell Lines from Early-Onset Colorectal Cancers. *Transl Oncol* **12,** 1185–1195 (2019).

582. Ming, X. Y. *et al.* Integrin Î±7 is a functional cancer stem cell surface marker in oesophageal squamous cell carcinoma. *Nat Commun* **7,** 13568 (Dec. 2016).

583. Haas, T. L. *et al.* Integrin Î±7 Is a Functional Marker and Potential Therapeutic Target in Glioblastoma. *Cell Stem Cell* **21,** 35–50 (July 2017).

584. Takada, N. *et al.* Phospholipid-flipping activity of P4-ATPase drives membrane curvature. *EMBO J.* **37** (May 2018).

585. Hanamura, I., Huang, Y., Zhan, F., Barlogie, B. & Shaughnessy, J. Prognostic value of cyclin D2 mRNA expression in newly diagnosed multiple myeloma treated with high-dose chemotherapy and tandem autologous stem cell transplantations. *Leukemia* **20,** 1288–1290 (2006).

586. Dib, A., Gabrea, A., Glebov, O. K., Bergsagel, P. L. & Kuehl, W. M. Characterization of MYC translocations in multiple myeloma cell lines. *J. Natl. Cancer Inst. Monographs,* 25–31 (2008).

587. Ghandi, M. *et al.* Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* **569,** 503–508 (2019).

588. Kawamoto, H. & Minato, N. Myeloid cells. *Int. J. Biochem. Cell Biol.* **36,** 1374–1379 (2004).

589. Arber, D. A. *et al.* The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. *Blood* **127,** 2391–2405 (May 2016).

590. Rowley, J. D. Identificaton of a translocation with quinacrine fluorescence in a patient with acute leukemia. *Ann Genet* **16,** 109–112 (1973).

591. Albain, K. S., Le Beau, M. M., Ullirsch, R. & Schumacher, H. Implication of prior treatment with drug combinations including inhibitors of topoisomerase II in therapy-related monocytic leukemia with a 9; 11 translocation. *Genes, Chromosomes and Cancer* **2,** 53–58 (1990).

592. Von Lindern, M *et al.* The translocation (6; 9), associated with a specific subtype of acute myeloid leukemia, results in the fusion of two genes, dek and can, and the expression of a chimeric, leukemia-specific dek-can mRNA. *Molecular and Cellular Biology* **12,** 1687–1697 (1992).

593. Rowley, J. D. & Potter, D. Chromosomal banding patterns in acute nonlymphocytic leukemia. *Blood* **47,** 705–721 (1976).

594. Soupir, C. P. *et al.* Philadelphia chromosome-positive acute myeloid leukemia: a rare aggressive leukemia with clinicopathologic features distinct from chronic myeloid leukemia in myeloid blast crisis. *American Journal of Clinical Pathology* **127,** 642–650 (2007).

595. Falini, B. *et al.* Cytoplasmic nucleophosmin in acute myelogenous leukemia with a normal karyotype. *New England Journal of Medicine* **352,** 254–266 (2005).

596. Van Doorn, S. B. V. W. *et al.* Biallelic mutations in the CEBPA gene and low CEBPA expression levels as prognostic markers in intermediate-risk AML. *Hematology* **4,** 31–40 (2003).

597. Schnittger, S. *et al.* RUNX1 mutations are frequent in de novo AML with noncomplex karyotype and confer an unfavorable prognosis. *Blood* **117,** 2348–2357 (2011).

598. Ley, T. J. *et al.* Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N. Engl. J. Med.* **368,** 2059–2074 (May 2013).

599. Papaemmanuil, E. *et al.* Genomic Classification and Prognosis in Acute Myeloid Leukemia. *N. Engl. J. Med.* **374,** 2209–2221 (2016).

600. Johansson, B., Mertens, F. & Mitelman, F. Cytogenetic deletion maps of hematologic neoplasms: circumstantial evidence for tumor suppressor loci. *Genes Chromosomes Cancer* **8,** 205–218 (1993).

601. Klampfl, T. *et al.* Genome integrity of myeloproliferative neoplasms in chronic phase and during disease progression. *Blood* **118,** 167–176 (2011).

602. Bloomfield, C. D. Chromosome abnormalities in secondary myelodysplastic syndromes. *Scand J Haematol Suppl* **45,** 82–90 (1986).

603. Rowley, J. D., Golomb, H. M. & Vardiman, J. Acute leukemia after treatment of lymphoma. *N. Engl. J. Med.* **297,** 1013 (1977).

604. Neuman, W. L. *et al.* Chromosomal loss and deletion are the most common mechanisms for loss of heterozygosity from chromosomes 5 and 7 in malignant myeloid disorders. *Blood* **79,** 1501–1510 (1992).

605. Bernstein, R, Philip, P & Ueshima, Y. Fourth International Workshop on Chromosomes in Leukemia 1982: Abnormalities of chromosome 7 resulting in monosomy 7 or in deletion of the long arm (7q-): review of translocations, breakpoints, and associated abnormalities. *Cancer Genetics and Cytogenetics* **11,** 300–303 (1984).

606. Hasle, H. *et al.* Myelodysplastic syndrome, juvenile myelomonocytic leukemia, and acute myeloid leukemia associated with complete or partial monosomy 7. European Working Group on MDS in Childhood (EWOG-MDS). *Leukemia* **13,** 376–385 (1999).

607. Le Beau, M. M. *et al.* Cytogenetic and molecular delineation of a region of chromosome 7 commonly deleted in malignant myeloid diseases. *Blood* **88,** 1930–1935 (1996).

608. Jerez, A. *et al.* Loss of heterozygosity in 7q myeloid disorders: clinical associations and genomic pathogenesis. *Blood* **119,** 6109–6117 (2012).

609. Honda, H., Nagamachi, A. & Inaba, T. -7/7q- syndrome in myeloid-lineage hematopoietic malignancies: attempts to understand this complex disease entity. *Oncogene* **34,** 2413–2425 (2015).

610. Schanz, J. *et al. Therapy with demethylating agents significantly improves overall-and AML-free survival in patients with MDS classified as high-risk by IPSS or very high risk by IPSS-R and partial or total monosomy 7- results from a German Multicenter Study* 2013.

611. Grimwade, D. *et al.* Refinement of cytogenetic classification in acute myeloid leukemia: determination of prognostic significance of rare recurring chromosomal abnormalities among 5876 younger adult patients treated in the United Kingdom Medical Research Council trials. *Blood* **116,** 354–365 (2010).

612. Beverloo, H. B. *et al.* Fusion of the homeobox gene HLXB9 and the ETV6 gene in infant acute myeloid leukemias with the t(7;12)(q36;p13). *Cancer Res.* **61,** 5374–5377 (2001).

613. Holland, P. W., Booth, H. A. & Bruford, E. A. Classification and nomenclature of all human homeobox genes. *BMC Biol.* **5,** 47 (2007).

614. Ferrier, D. E., Brooke, N. M., Panopoulou, G. & Holland, P. W. The Mnx homeobox gene class defined by HB9, MNR2 and amphioxus AmphiMnx. *Dev. Genes Evol.* **211,** 103–107 (2001).

615. Tanabe, Y., William, C. & Jessell, T. M. Specification of motor neuron identity by the MNR2 homeodomain protein. *Cell* **95,** 67–80 (1998).

616. Arber, S. *et al.* Requirement for the homeobox gene Hb9 in the consolidation of motor neuron identity. *Neuron* **23,** 659–674 (1999).

617. Wildenhain, S. *et al.* Expression of cell-cell interacting genes distinguishes HLXB9/TEL from MLL-positive childhood acute myeloid leukemia. *Leukemia* **24,** 1657–1660 (2010).

618. Ingenhag, D. *et al.* The homeobox transcription factor HB9 induces senescence and blocks differentiation in hematopoietic stem and progenitor cells. *Haematologica* **104,** 35–46 (2019).

619. Helmrich, A., Ballarino, M. & Tora, L. Collisions between replication and transcription complexes cause common fragile site instability at the longest human genes. *Mol. Cell* **44,** 966–977 (2011).

620. Leone, R. *et al.* Immunoglobulin heavy chain and T-cell receptor beta chain gene rearrangements in acute non lymphoid leukemia. *Haematologica* **75,** 125–128 (1990).

621. Polprasert, C. *et al.* Inherited and Somatic Defects in DDX41 in Myeloid Neoplasms. *Cancer Cell* **27,** 658–670 (2015).

622. Grossmann, V. *et al.* Whole-exome sequencing identifies somatic mutations of BCOR in acute myeloid leukemia with normal karyotype. *Blood* **118,** 6153–6163 (2011).

623. Ley, T. J. *et al.* DNMT3A mutations in acute myeloid leukemia. *N. Engl. J. Med.* **363,** 2424–2433 (2010).

624. Schnittger, S. *et al.* IDH1 mutations are detected in 6.6patients and are associated with intermediate risk karyotype and unfavorable prognosis in adults younger than 60 years and unmutated NPM1 status. *Blood* **116,** 5486–5496 (2010).

625. Gaidzik, V. I. *et al.* RUNX1 mutations in acute myeloid leukemia are associated with distinct clinico-pathologic and genetic features. *Leukemia* **30,** 2160–2168 (Nov. 2016).

626. Maxson, J. E. *et al.* Therapeutically Targetable ALK Mutations in Leukemia. *Cancer Res.* **75,** 2146–2150 (2015).

627. Daily, K., Patel, V. R., Rigor, P., Xie, X. & Baldi, P. MotifMap: integrative genome-wide maps of regulatory motif sites for model species. *BMC Bioinformatics* **12,** 495 (2011).

628. Plaisier, C. L. *et al.* Causal Mechanistic Regulatory Network for Glioblastoma Deciphered Using Systems Genetics Network Analysis. *Cell Syst* **3,** 172–186 (Aug. 2016).

629. Jolma, A. *et al.* DNA-binding specificities of human transcription factors. *Cell* **152,** 327–339 (2013).

630. Matys, V. *et al.* TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* **34,** D108–110 (2006).

631. Pujato, M., Kieken, F., Skiles, A. A., Tapinos, N. & Fiser, A. Prediction of DNA binding motifs from 3D models of transcription factors; identifying TLX3 regulated genes. *Nucleic Acids Res.* **42,** 13500–13512 (2014).

632. Schweizer, J. *et al.* New consensus nomenclature for mammalian keratins. *J. Cell Biol.* **174,** 169–174 (2006).

633. Scales, S. J., Hesser, B. A., Masuda, E. S. & Scheller, R. H. Amisyn, a novel syntaxin-binding protein that may regulate SNARE complex assembly. *J. Biol. Chem.* **277,** 28271–28279 (2002).

634. Chang, S. H. *et al.* Interleukin-17C promotes Th17 cell responses and autoimmune disease via interleukin-17 receptor E. *Immunity* **35,** 611–621 (2011).

635. Balgobind, B. V. *et al.* Evaluation of gene expression signatures predictive of cytogenetic and molecular subtypes of pediatric acute myeloid leukemia. *Haematologica* **96,** 221–230 (2011).

636. Dayyani, F. *et al.* Loss of TLE1 and TLE4 from the del(9q) commonly deleted region in AML cooperates with AML1-ETO to affect myeloid cell proliferation and survival. *Blood* **111,** 4338–4347 (2008).

637. Liu, Y. & Zhang, W. Identification of a new transmembrane adaptor protein that constitutively binds Grb2 in B cells. *J. Leukoc. Biol.* **84,** 842–851 (2008).

638. Zjablovskaja, P. *et al.* EVI2B is a C/EBPÎ± target gene required for granulocytic differentiation and functionality of hematopoietic progenitors. *Cell Death Differ.* **24,** 705–716 (Apr. 2017).

639. Le Coq, J. *et al.* Structural basis for interdomain communication in SHIP2 providing high phosphatase activity. *Elife* **6** (Aug. 2017).

640. Helgason, C. D. *et al.* Targeted disruption of SHIP leads to hemopoietic perturbations, lung pathology, and a shortened life span. *Genes Dev.* **12,** 1610–1620 (1998).

641. Giuriato, S. *et al.* SHIP2 overexpression strongly reduces the proliferation rate of K562 erythroleukemia cell line. *Biochem. Biophys. Res. Commun.* **296,** 106–110 (2002).

642. Kawahara, M. *et al.* H2.0-like homeobox regulates early hematopoiesis and promotes acute myeloid leukemia. *Cancer Cell* **22,** 194–208 (2012).

643. Piragyte, I. *et al.* A metabolic interplay coordinated by HLX regulates myeloid differentiation and AML through partly overlapping pathways. *Nat Commun* **9,** 3090 (Aug. 2018).

644. Martinez-Trillos, A. *et al.* Mutations in TLR/MYD88 pathway identify a subset of young chronic lymphocytic leukemia patients with favorable outcome. *Blood* **123,** 3790–3796 (2014).

645. Ngo, V. N. *et al.* Oncogenically active MYD88 mutations in human lymphoma. *Nature* **470,** 115–119 (2011).

646. Mele, M. *et al.* Human genomics. The human transcriptome across tissues and individuals. *Science* **348,** 660–665 (2015).

647. Zhao, G. Q., Zhao, Q., Zhou, X., Mattei, M. G. & de Crombrugghe, B. TFEC, a basic helix-loop-helix protein, forms heterodimers with TFE3 and inhibits TFE3-dependent transcription activation. *Mol. Cell. Biol.* **13,** 4505–4512 (1993).

648. Huan, C., Sashital, D., Hailemariam, T., Kelly, M. L. & Roman, C. A. Renal carcinoma-associated transcription factors TFE3 and TFEB are leukemia inhibitory factor-responsive transcription activators of E-cadherin. *J. Biol. Chem.* **280,** 30225–30235 (2005).

649. Federico, C. *et al.* Deletions of Chromosome 7q Affect Nuclear Organization and HLXB9Gene Expression in Hematological Disorders. *Cancers (Basel)* **11** (2019).

650. Huber, K. The sympathoadrenal cell lineage: specification, diversification, and new perspectives. *Dev. Biol.* **298,** 335–343 (2006).

651. Esiashvili, N., Goodman, M., Ward, K., Marcus, R. B. & Johnstone, P. A. Neuroblastoma in adults: Incidence and survival analysis based on SEER data. *Pediatr Blood Cancer* **49,** 41–46 (2007).

652. Van Groningen, T. *et al.* Neuroblastoma is composed of two super-enhancer-associated differentiation states. *Nat. Genet.* **49,** 1261–1266 (2017).

653. Furlan, A. *et al.* Multipotent peripheral glial cells generate neuroendocrine cells of the adrenal medulla. *Science* **357** (July 2017).

654. Pugh, T. J. *et al.* The genetic landscape of high-risk neuroblastoma. *Nat. Genet.* **45,** 279–284 (2013).

655. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499,** 214–218 (2013).

656. Schwab, M. *et al.* Chromosome localization in normal human cells and neuroblastomas of a gene related to c-myc. *Nature* **308,** 288–291 (1984).

657. Simon, T., Spitz, R., Faldum, A., Hero, B. & Berthold, F. New definition of low-risk neuroblastoma using stage, age, and 1p and MYCN status. *J. Pediatr. Hematol. Oncol.* **26,** 791–796 (2004).

658. Oberthuer, A. *et al.* Revised risk estimation and treatment stratification of low- and intermediate-risk neuroblastoma patients by integrating clinical and molecular prognostic markers. *Clin. Cancer Res.* **21,** 1904–1915 (2015).

659. Kocak, H. *et al.* Hox-C9 activates the intrinsic pathway of apoptosis and is associated with spontaneous regression in neuroblastoma. *Cell Death Dis* **4,** e586 (2013).

660. Su, Z. *et al.* A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat. Biotechnol.* **32,** 903–914 (2014).

661. Lastowska, M. *et al.* Promiscuous translocations of chromosome arm 17q in human neuroblastomas. *Genes Chromosomes Cancer* **19,** 143–149 (1997).

662. Maris, J. M. *et al.* Significance of chromosome 1p loss of heterozygosity in neuroblastoma. *Cancer Res.* **55,** 4664–4669 (1995).

663. Caren, H. *et al.* High-risk neuroblastoma tumors with 11q-deletion display a poor prognostic, chromosome instability phenotype with later onset. *Proc. Natl. Acad. Sci. U.S.A.* **107,** 4323–4328 (2010).

664. Cheung, N. K. *et al.* Association of age at diagnosis and genetic mutations in patients with neuroblastoma. *JAMA* **307,** 1062–1071 (2012).

665. Bresler, S. C. *et al.* ALK mutations confer differential oncogenic activation and sensitivity to ALK inhibition therapy in neuroblastoma. *Cancer Cell* **26,** 682–694 (2014).

666. Zhu, S. *et al.* Activated ALK collaborates with MYCN in neuroblastoma pathogenesis. *Cancer Cell* **21,** 362–373 (2012).

667. Su, W. T. *et al.* Positional gene expression analysis identifies 12q overexpression and amplification in a subset of neuroblastomas. *Cancer Genet. Cytogenet.* **154,** 131–137 (2004).

668. Diskin, S. J. *et al.* Common variation at 6q16 within HACE1 and LIN28B influences susceptibility to neuroblastoma. *Nat. Genet.* **44,** 1126–1130 (2012).

669. Cesare, A. J. & Reddel, R. R. Alternative lengthening of telomeres: models, mechanisms and implications. *Nat. Rev. Genet.* **11,** 319–330 (2010).

670. Eleveld, T. F. *et al.* Relapsed neuroblastomas show frequent RAS-MAPK pathway mutations. *Nat. Genet.* **47,** 864–871 (Aug. 2015).

671. Maris, J. M., Hogarty, M. D., Bagatell, R. & Cohn, S. L. Neuroblastoma. *Lancet* **369,** 2106–2120 (2007).

672. Whittle, S. B. *et al.* Overview and recent advances in the treatment of neuroblastoma. *Expert Rev Anticancer Ther* **17,** 369–386 (2017).

673. Matthay, K. K. *et al.* Treatment of high-risk neuroblastoma with intensive chemotherapy, radiotherapy, autologous bone marrow transplantation, and 13-cis-retinoic acid. Children's Cancer Group. *N. Engl. J. Med.* **341,** 1165–1173 (1999).

674. Smith, M. A., Altekruse, S. F., Adamson, P. C., Reaman, G. H. & Seibel, N. L. Declining childhood and adolescent cancer mortality. *Cancer* **120,** 2497–2506 (2014).

675. Wang, Q. *et al.* Integrative genomics identifies distinct molecular classes of neuroblastoma and shows that multiple genes are targeted by regional alterations in DNA copy number. *Cancer Res.* **66,** 6050–6062 (2006).

676. Abel, F. *et al.* A 6-gene signature identifies four molecular subgroups of neuroblastoma. *Cancer Cell Int.* **11,** 9 (2011).

677. Abyzov, A. *et al.* Analysis of deletion breakpoints from 1,092 humans reveals details of mutation mechanisms. *Nat Commun* **6,** 7256 (2015).

678. Molenaar, J. J. *et al.* Sequencing of neuroblastoma identifies chromothripsis and defects in neuritogenesis genes. *Nature* **483,** 589–593 (2012).

679. Lopez, G. *et al.* Structural variation targets neurodevelopmental genes and identifies SHANK2 as a tumor suppressor in neuroblastoma. *bioRxiv* (2019).

680. Akazawa, C., Ishibashi, M., Shimizu, C., Nakanishi, S. & Kageyama, R. A mammalian helix-loop-helix factor structurally related to the product of Drosophila proneural gene atonal is a positive transcriptional regulator expressed in the developing nervous system. *J. Biol. Chem.* **270,** 8730–8738 (1995).

681. Knoepp, S. M. *et al.* Effects of active and inactive phospholipase D2 on signal transduction, adhesion, migration, invasion, and metastasis in EL4 lymphoma cells. *Mol. Pharmacol.* **74,** 574–584 (2008).

682. Chuang, T. H. *et al.* Abr and Bcr are multifunctional regulators of the Rho GTP-binding protein family. *Proc. Natl. Acad. Sci. U.S.A.* **92,** 10282–10286 (1995).

683. Louis, C. U. & Shohet, J. M. Neuroblastoma: molecular pathogenesis and therapy. *Annu. Rev. Med.* **66,** 49–63 (2015).

684. Cazes, A. *et al.* Activated Alk triggers prolonged neurogenesis and Ret upregulation providing a therapeutic target in ALK-mutated neuroblastoma. *Oncotarget* **5,** 2688–2702 (2014).

685. Franks, L. M., Bollen, A., Seeger, R. C., Stram, D. O. & Matthay, K. K. Neuroblastoma in adults and adolescents: an indolent course with poor survival. *Cancer* **79,** 2028–2035 (1997).

686. Klisch, T. J. *et al.* In vivo Atoh1 targetome reveals how a proneural transcription factor regulates cerebellar development. *Proc. Natl. Acad. Sci. U.S.A.* **108,** 3288–3293 (2011).

687. Jones, S. An overview of the basic helix-loop-helix proteins. *Genome Biol.* **5,** 226 (2004).

688. Zelinger, L. *et al.* Genetics and Disease Expression in the CNGA3 Form of Achromatopsia: Steps on the Path to Gene Therapy. *Ophthalmology* **122,** 997–1007 (2015).

689. Olsen, R. R. *et al.* MYCN induces neuroblastoma in primary neural crest cells. *Oncogene* **36,** 5075–5082 (Aug. 2017).

690. Brown, L., Espinosa, R., Le Beau, M. M., Siciliano, M. J. & Baer, R. HEN1 and HEN2: a subgroup of basic helix-loop-helix genes that are coexpressed in a human neuroblastoma. *Proc. Natl. Acad. Sci. U.S.A.* **89,** 8492–8496 (1992).

691. Fritzsch, B., Eberl, D. F. & Beisel, K. W. The role of bHLH genes in ear development and evolution: revisiting a 10-year-old hypothesis. *Cell. Mol. Life Sci.* **67,** 3089–3099 (2010).

692. Zhao, X. *et al.* The N-Myc-DLL3 cascade is suppressed by the ubiquitin ligase Huwe1 to inhibit proliferation and promote neurogenesis in the developing brain. *Dev. Cell* **17,** 210–221 (2009).

693. Sano, R. *et al.* Abstract LB-136: Pediatric Preclinical Testing Consortium evaluation of a DLL3-targeted antibody drug conjugate rovalpituzumab tesirine, in neuroblastoma. *Cancer Research* **78,** LB–136–LB–136 (2018).

694. Lee, S., Jeong, H. S. & Cho, H. H. Atoh1 as a Coordinator of Sensory Hair Cell Development and Regeneration in the Cochlea. *Chonnam Med J* **53,** 37–46 (2017).

695. Kobayashi, M., Sakai, E., Furuta, Y. & Takamatsu, K. Isolation of two human cDNAs, HLP3 and HLP4, homologous to the neuron-specific calcium-binding protein genes. *DNA Seq.* **9,** 171–176 (1998).

696. Wainwright, L. J., Lasorella, A. & Iavarone, A. Distinct mechanisms of cell cycle arrest control the decision between differentiation and senescence in human neuroblastoma cells. *Proc. Natl. Acad. Sci. U.S.A.* **98,** 9396–9400 (2001).

697. Jahan, I., Pan, N. & Fritzsch, B. Opportunities and limits of the one gene approach: the ability of Atoh1 to differentiate and maintain hair cells depends on the molecular context. *Front Cell Neurosci* **9,** 26 (2015).

698. Zhong, C., Fu, Y., Pan, W., Yu, J. & Wang, J. Atoh1 and other related key regulators in the development of auditory sensory epithelium in the mammalian inner ear: function and interplay. *Dev. Biol.* **446,** 133–141 (Feb. 2019).

699. Ayrault, O. *et al.* Atoh1 inhibits neuronal differentiation and collaborates with Gli1 to generate medulloblastoma-initiating cells. *Cancer Res.* **70,** 5618–5627 (2010).

700. Wegert, J. *et al.* Recurrent intragenic rearrangements of EGFR and BRAF in soft tissue tumors of infants. *Nat Commun* **9,** 2378 (June 2018).

701. Porubsky, D. *et al.* Dense and accurate whole-chromosome haplotyping of individual genomes. *Nat Commun* **8,** 1293 (Nov. 2017).

702. Jiao, W. B. *et al.* Improving and correcting the contiguity of long-read genome assemblies of three plant species using optical mapping and chromosome conformation capture data. *Genome Res.* **27,** 778–786 (May 2017).

703. Chan, E. K. F. *et al.* Optical mapping reveals a higher level of genomic architecture of chained fusions in cancer. *Genome Res.* **28,** 726–738 (May 2018).

704. Jaratlerdsiri, W. *et al.* Next generation mapping reveals novel large genomic rearrangements in prostate cancer. *Oncotarget* **8,** 23588–23602 (2017).

705. Dixon, J. R. *et al.* Integrative detection and analysis of structural variation in cancer genomes. *Nat. Genet.* **50,** 1388–1398 (Oct. 2018).

706. Mun, D. G. *et al.* Proteogenomic Characterization of Human Early-Onset Gastric Cancer. *Cancer Cell* **35,** 111–124 (2019).

707. Vasaikar, S. *et al.* Proteogenomic Analysis of Human Colon Cancer Reveals New Therapeutic Opportunities. *Cell* **177,** 1035–1049 (2019).

708. Vriens, K. *et al.* Evidence for an alternative fatty acid desaturation pathway increasing cancer plasticity. *Nature* **566,** 403–406 (Feb. 2019).

709. Courtney, K. D. *et al.* Isotope Tracing of Human Clear Cell Renal Cell Carcinomas Demonstrates Suppressed Glucose Oxidation In Vivo. *Cell Metab.* **28,** 793–800 (2018).