Inaugural dissertation
for
obtaining the doctoral degree
of the
Combined Faculty of Mathematics, Engineering and Natural Sciences
of the
Ruprecht-Karls-University
Heidelberg

Presented by
Lukas Madenach, M.Sc.
Born in Nürnberg, Germany
Oral examination: 16.12.2024

Computational multi-omics analysis of pediatric DADDR patients and synthetic lethality prediction in K27M mutated pediatric high-grade glioma

Referees:
Prof. Dr. Michael Boutros
Prof. Dr. Stefan Pfister

To Silke, Oskar and my parents

Declaration

Herewith I declare that I have written and submitted this dissertation myself and in the process, have not used any other sources or tools than indicated. The results presented in this thesis were obtained at the DKFZ Heidelberg under the supervision of Dr. Natalie Jäger, Dr. Robert Autry and Prof. Dr. Stefan Pfister from 2020 till 2024. Parts of the results, methods and discussion related to the analysis of methylation of DADDR patients are used for a manuscript currently in preparation.

I hereby declare that I have not applied to be examined at any other institution, nor have I used the dissertation in this or any other form at any other institution as an examination paper, nor submitted it to any other faculty as a dissertation.

Lukas Madenach

1 Abstract

Pediatric cancer is a conglomerate of complicated diseases remaining one of the leading causes of death worldwide in patients aged 1 – 19. In recent years, major sequencing and precision oncology programs were launched and aggregated unprecedented amounts of data giving detailed insights into pediatric cancer at multiple omics levels.
One possible avenue for improved treatment strategies, leverages synthetic lethality (SL) interactions between genes, potentially resulting in clinically relevant combination treatments. Despite the great interest, discovery of gene pairs with synthetically lethal interaction is very challenging and resource consuming because of the sheer size of the combinatorial space that needs to be covered. Advancements in *in-silico* prediction methods, utilizing multi-omics data, for SL interactions to narrow down the scope of investigation have gained popularity over the last years. In the first part of this study, I present and extensively evaluate a computational approach for the prediction of interacting pairs of genes exhibiting synthetic lethality. I apply my approach to two dedicated dataset I curated from multi-omics data of pediatric high-grade glioma (pedHGG) patients and compare the results. Finally, I describe a set of predicted SL pairs specific for pedHGG K27M, a particularly challenging childhood brain tumor, which includes multiple drug targets, targetable for example by HDAC inhibitors, that might serve as a guide for future investigations.

Mutational signatures as a proxy for underlying mutational processes can be used as a biomarker for example for the detection of homologous recombination (HR) deficiency in adult cancer patients. However, biomarkers identified for adult cancer cannot be used for pediatric patients without further investigation and validation because of the different disease types and biological characteristics of pediatric cancer. In the second part of this study, I present an investigation into the active mutational processes in pediatric patients with disorders with abnormal DNA damage response. Using the latest set of mutational signatures and comparing two extraction algorithms, I thoroughly investigate associations of mutational signatures with clinically relevant characteristics and present a comprehensive overview of the mutational landscape in pediatric patients with abnormal DNA damage response. I was able to confirm reported associations of mutational signatures and my results further refine previous results especially with regard to differences between patients with germline *TP53* mutation and patients with wildtype *TP53*. Further, I compare differences, reflected in mutational signatures, between patients with germline mutation in *MSH6, MSH2, MLH1* or *PMS2* that belong to the mismatch repair deficiency syndrome.

Not just mutational signatures but also methylation can be used as a reliable biomarker e.g. to classify pediatric tumors with high accuracy into subgroups beyond what is possible via morphological differences. This classification method poises challenges for patients with hereditary abnormal DNA damage response and currently there is only limited knowledge about methylation patterns in such patients. In the third part of this study, I present new insights investigating methylation patterns in patients with abnormal DNA damage response. Using a molecularly characterized control cohort I assembled and accounting for immune cell infiltration, I was able to identify methylation signatures specific for defective DNA damage response across different tumor types. After detailed characterization of the discriminatory power of the identified methylation signature, that indicated achieving 90% precision is possible, I further investigated the biological function of identified methylation sites. This revealed association with biological functions including the RISC complex, RNAi and DNA damage response pathways such as base excision repair and nucleotide excision repair. Finally, I validated the presented methylation signatures in an additional internal and one external patient cohort consisting of liquid biopsy samples, demonstrating the broader applicability and highlighting a potential clinical application of the methylation signatures.

2 Zusammenfassung

Krebs in Kindern ist eine komplizierte Ansammlung von Krankheiten und bleibt weltweit eine der häufigsten Todesursachen bei Patienten im Alter von 1 bis 19 Jahren. In den letzten Jahren wurden große Sequenzierungs- und Präzisionsonkologieprogramme gestartet, die beispiellose Datenmengen zusammengetragen haben, die detaillierte Einblicke in Kinderkrebs auf mehreren omics-Ebenen liefern.

Ein möglicher Weg für eine verbesserte Behandlung besteht darin, die synthetische Letalitätsinteraktion zwischen Genen zu nutzen, was zu klinisch relevanten kombinatorischen Behandlungsstrategien führen kann. Trotz des großen Interesses ist die Entdeckung von Genpaaren mit synthetisch tödlicher Interaktion aufgrund der schieren Größe des kombinatorischen Raums, der abgedeckt werden muss, eine große Herausforderung und ressourcenintensiv. Fortschritte bei *in-silico*-Vorhersagemethoden unter Verwendung von Multi-omics-Daten für SL-Interaktionen zur Eingrenzung der Untersuchung haben in den letzten Jahren an Popularität gewonnen. Im ersten Teil dieser Studie stelle ich einen verbesserten Ansatz zur Vorhersage interagierender Genpaare welche synthetische Letalität aufweisen vor und evaluiere ihn ausführlich. Ich wende meine Methode auf zwei Datensätze an, die ich aus Multi-omics-Daten von pedHGG-Patienten kuratiert habe und vergleiche die Ergebnisse. Abschließend beschreibe ich eine Reihe vorhergesagter SL-Paare, die spezifisch für pedHGG K27M, einem besonders herausfordernden Hinrtumor, sind und Ziele mehrerer Medikamente enthalten, beispielsweise HDAC-Inhibitoren.

Mutationssignaturen als Proxy für zugrundeliegende Mutationsprozesse können und wurden als Biomarker, beispielsweise für die Erkennung von HR-Defekten bei Erwachsenen, verwendet. Allerdings können für Krebs bei Erwachsenen identifizierte Biomarker, aufgrund der unterschiedlichen Krankheitsbilder und biologischen Merkmalen von Krebs bei Kindern, nicht ohne Weiteres für pädiatrische Patienten verwendet werden. Im zweiten Teil dieser Studie präsentiere ich eine Untersuchung der aktiven Mutationsprozesse bei pädiatrischen Patienten mit abnormaler DNA-Schadensreaktion. Mithilfe des neuesten Satzes von Mutationssignaturen und dem Vergleich zweier Extraktionsalgorithmen untersuche ich gründlich die Zusammenhänge von Mutationssignaturen mit klinisch relevanten Merkmalen und präsentiere einen umfassenden Überblick über die Mutationslandschaft bei pädiatrischen Patienten mit abnormaler DNA-Schadensreaktion. Ich konnte die bekannten Assoziationen von Mutationssignaturen bestätigen und meine Ergebnisse verfeinern frühere Ergebnisse weiter, insbesondere im Hinblick auf Unterschiede zwischen Patienten mit Keimbahn-*TP53*-Mutation und Patienten mit Wildtyp-*TP53*. Darüber hinaus vergleiche ich Unterschiede, die sich in Mutationssignaturen widerspiegeln, zwischen Patienten mit Keimbahnmutationen in *MSH6, MSH2, MLH1* oder *PMS2*, die zum Mismatch-Repair-Defizienz-Syndrom gehören.

Nicht nur Mutationssignaturen, sondern auch die Methylierung können als zuverlässiger Biomarker verwendet werden, z.B. um pädiatrische Tumoren mit einer hohen Präzision in Untergruppen zu klassifizieren, die über das hinausgehen, was durch morphologische Unterschiede möglich ist. Diese Klassifizierungsmethode funktioniert bei Patienten mit abnormaler DNA-Schadensreaktion weniger zuverlässig und es liegen derzeit nur begrenzte Kenntnisse über Methylierungsmuster bei solchen Patienten vor. Im dritten Teil dieser Studie präsentiere ich neue Erkenntnisse zur Untersuchung von Methylierungsmustern bei Patienten mit abnormaler DNA-Schadensreaktion. Mithilfe einer molekular charakterisierten Kontrollkohorte, die ich zusammengestellt habe und unter Berücksichtigung der Immunzellen in Tumoren, identifizierte ich spezifische Methylierungssignaturen. Nach einer detaillierten Charakterisierung der Unterscheidungskraft der identifizierten Methylierungssignatur, die zeigte, dass eine Genauigkeit von 90 % möglich ist, untersuchte ich weiter die biologische

Funktion der identifizierten Methylierungsprobes. Dies zeigte einen Zusammenhang mit biologischen Funktionen, einschließlich des RISC-Komplexes, RNAi und DNA-Schadensreaktionswegen wie der Base-excision-repair und der Nucleotide-excision-repair. Abschließend validierte ich die präsentierten Methylierungssignaturen in einer internen und einer externen Patientenkohorte welche aus Liquid-biopsy-Proben besteht, um die breitere Anwendbarkeit zu demonstrieren und eine potenzielle klinische Anwendung der Methylierungssignatur hervorzuheben.

3 Acronyms

SL	Synthetic lethality
RF	Random forest
KNN	K-nearest neighbors
ABC	AdaBoost classifier
CMFW	Collective matrix factorization weighted
CMF	Collective matrix factorization
RNA	Ribonucleic acid
DNA	Deoxyribonucleic acid
KG	Knowledge graph
HGG	High-grade glioma
ped (prefix)	Pediatric
WES	Whole exome sequencing
WGS	Whole genome sequencing
lc (prefix)	Low coverage
CNV	Copy number variation
SNV	Single nucleotide variation
INDEL	Insertion or Deletion
CV	Crossvalidation
NER	Nucleotide excision repair
HR	Homologous recombination
NHEJ	Non-homologous end joining
RNAi	RNA interference
NGS	Next generation sequencing
PXA	Pleomorphic xanthoastrocytoma
NMF	Non-negative matrix factorization
WHO	World health organization
DADDR	Disorders with abnormal DNA damage response
DMR	Differentially methylated region
VMR	Variably methylated region

4 Contents

5 Introduction

5.1 Cancer

Cancer is a complex disease characterized by the uncontrolled proliferation of cells. Locally limited, tissue invasive clusters of such cells are called tumors, which can stay in their place of origin or spread via the bloodstream or the lymphatic system in which case they are referred to as metastatic tumors. Cancer is one of the leading causes of disease related deaths worldwide, causing around 10 million premature deaths in 2020 and is one of the major barriers preventing higher life expectancies globally [1]. There are many distinct types of cancer, often arising from different locations in the body, these can further be divided into subtypes based on intrinsic or extrinsic characteristics that can require a range of different treatment modalities with different clinical outcome and prognosis. Overall the lifetime risk of being diagnosed with cancer is 40.9% for males and 39.1% for females [2, 3]. The causes for this marginal difference are not clear but are generally attributed to environmental influences (e.g. smoking habits, occupation, etc.) and endogenous properties. Next to gender, other factors such as age or race and ethnicity play an important role in the risk of being diagnosed with a certain type of cancer and outcome. For example black males in the USA have a 2 to 4 fold increased risk of dying from prostate cancer than any other ethnic group, although this is a purely descriptive statement [2]. Normal cells transform into cancer cells when acquired damage to the DNA, alterations to the epigenome or other biological components renders the cell unable to regulate this damage with appropriate responses through either apoptosis or repair mechanisms. Researchers have identified six overarching characteristic capabilities developed by cancer cells that enable uncontrolled proliferation and metastasis across all distinct subtypes of cancer. These characteristics are referred to as hallmarks of cancer: (1) sustaining proliferative signaling, (2) evading growth suppressors, (3) enabling replicative immortality, (4) activating invasion and metastasis, (5) inducing angiogenesis, (6) resisting cell death [4].

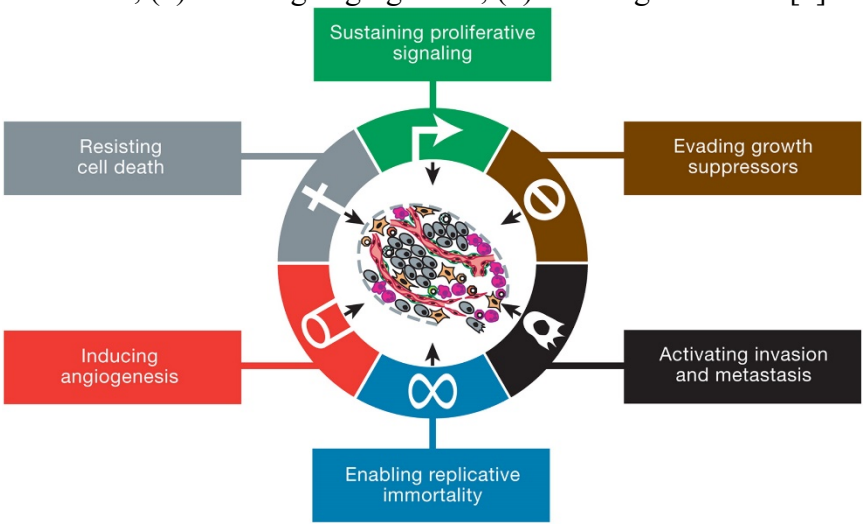


Figure 1: The 6 original hallmarks of cancer (sustaining proliferative signaling, evading growth suppressors, enabling replicative immortality, activating invasion and metastasis, inducing angiogenesis, resisting cell death). Graphic adapted from Hanahan 2011 [4].

More recently two additional hallmarks, (7) deregulating cellular metabolism and (8) avoiding immune destruction, were described as well as other enabling factors and emerging hallmarks [5].

5.1.1 Pediatric cancer

One of the major factors influencing the risk of cancer is age with the risk of developing cancer rising from 3.5% in the 0-49 years age bracket to 34% in the 70+ age bracket for males and from 5.8% to 27.2% in the respective age brackets for females [2]. This increased risk with rising age is linked to prolonged exposure to mutational processes and decreased fitness of the immune system. As such pediatric cancer is fundamentally different from adult cancers across multiple characteristics including but not limited to cancer types, cellular origins, driver mutations and underlying processes [6]. One major difference between adult and pediatric cancers is the lower overall mutational burden in pediatric cancers as a direct results from the shorter exposure time [7]. Further, pediatric tumors often lack the immune response typical for their adult counterparts, a phenomenon referred to as immune cold [8]. Overall, the treatment of pediatric cancers harbors distinct challenges for example the different ways children metabolize drugs and the larger emphasize that has to be placed on long-term consequences of applied treatments [9, 10]. Other challenges include insufficient economic incentives to develop therapies for pediatric oncology [11]. The most common types of cancer among children (age 0-14 years) are acute lymphocytic leukemia (26%), CNS tumors (21%), neuroblastoma (7%) and Non-Hodgkin lymphoma (6%). Among adolescents (age 15-19 years), the most common entities are Hodgkin lymphoma (15%), thyroid carcinoma (11%), CNS tumors (10%) and testicular germ cell tumors (8%). Since 1975, incidence rates for pediatric cancer have increased slightly at 0.6% per year (Figure 2). The exact causes for this rise are unknown, however a change in environmental factors as well as more advanced diagnosis are speculated to contribute [12].

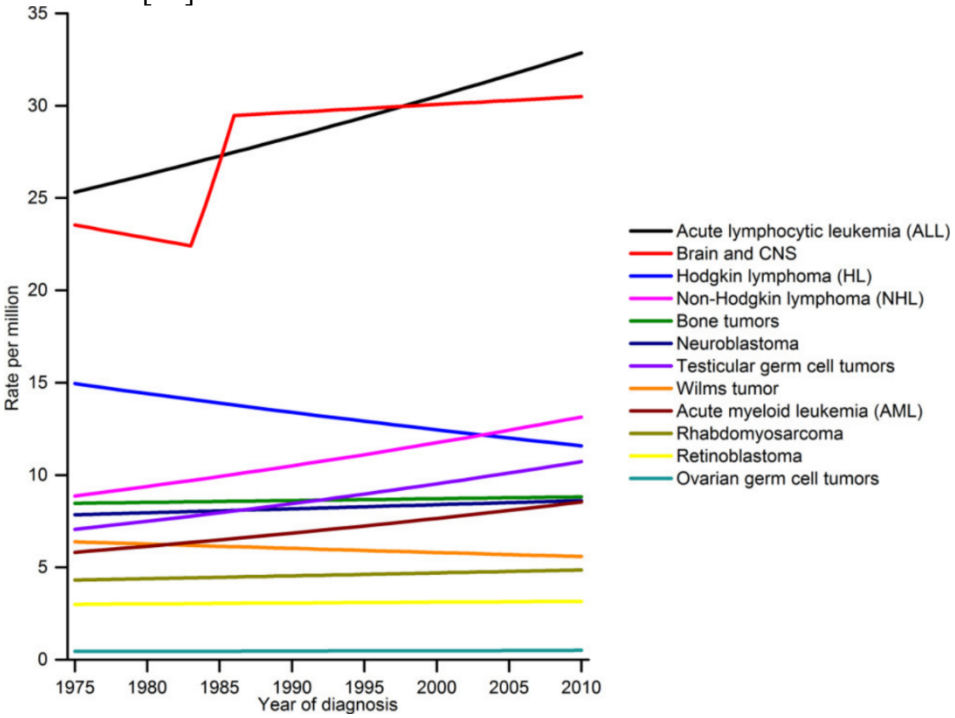


Figure 2: Incidence rates of different types of pediatric cancer in the USA. Ages 0-19 years. Graphic adapted from Ward et al. [12].

With this ever rising incidence rate and the decidedly different treatment requirements compared to adult cancer, there is considerable interest in addressing the specific needs of diagnosing and treating pediatric cancer, because just like in adults cancer remains the leading disease related cause of death in children from 1 to 19 globally [13-15]. Additionally, patients treated for pediatric cancer exhibited a significantly higher risk to develop cancer again later in life compared to the general population, thus urgently asking for more research into better treatment options [16]. It is not entirely understood, yet, whether the main factors influencing the increased risk are treatment or hereditary predisposition associated [17, 18]. For a better

understanding and more effective treatment of pediatric cancer while minimizing risk later in life, multiple international projects in the area of sequencing and precision medicine leverage the latest multi-omics techniques generating vast amounts of data and valuable insights leading to advances in the treatment and diagnosis of pediatric cancer. Among them there are the International Cancer Genome Consortium (ICGC), the pediatric cancer genome project (PCGP) and the INdividualized Therapy For Relapsed Malignancies (INFORM) to name but a few [19-23].

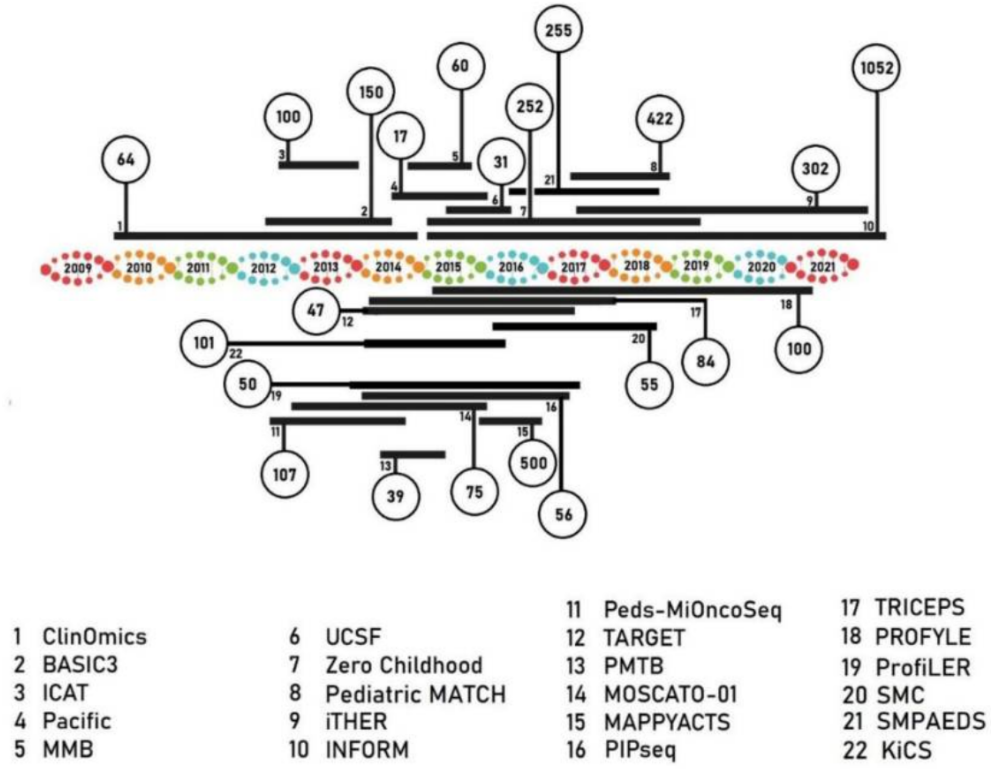


Figure 3: Precision medicine programs in pediatric oncology. Graphic adapted from Langenberg et al. [23].

The survival rate for pediatric cancer has risen considerably by 50% on average since 1975. While there are types of cancer with greater than average improvement of survival rates such as various leukemias, for other types there is still a lot more room for improvement [12]. The more recent efforts made in the sequencing or precision medicine programs mentioned above and other similar projects, however, had mixed effects [23]. In particular it is challenging to assess and compare the overall clinical benefit of these programs because there is no standardized patient selection plan, no standardized sample and data processing and no standardized treatment recommendations. For example some programs recruit relapsed patients or patients with refractory cancer (e.g. ClinOmics, MOSCATO-01, TRICEPS) while other programs focus on primary high-risk patients or rare tumors (e.g. TARGET, KiCS, BASIC3) [24-29]. There is also considerable variance in the applied NGS techniques ranging from only WES or targeted gene panel sequencing (e.g. BASIC3, ProfiLER) to more extensive protocols using WES, lcWGS, RNAseq and methylation arrays (e.g. INFORM, iTHER) [21, 22, 27, 30, 31]. Consequently the tumor boards of each program, responsible for translating molecular findings into clinically relevant recommendations, have to base their decisions on different sets of information while taking into account the rapidly changing landscape of biomarkers and treatments. Another challenge faced by these programs are legal hurdles concerning the sharing of patient data, increasing the difficulty for cooperation.

5.1.2 The importance of omics

On the molecular level, many cancers have been shown to be caused by alterations in critical locations to the genome or the epigenome. These alterations influence multiple processes across the cell that are working in a dynamic, connected manner. To capture a complete picture of all interactions caused by a set of alterations it is necessary to interrogate multiple layers of omics [32-34]. In this study, I will focus mainly on DNA sequencing and methylation.

WES and WGS

DNA sequencing used to be a very time consuming and laborious task with the human genome project taking roughly a decade to sequence one complete human genome [35]. In a large step towards high-throughput DNA sequencing the advent of commercially available next generation sequencing since 2005 dramatically changed the applicability of DNA sequencing in research paving the way for whole exome sequencing (WES) and whole genome sequencing (WGS). Today, NGS can be used to capture not only point mutations, insertions and deletions but also events like amplifications, translocations, inversions and gene fusions [36, 37]. WES refers to the targeted sequencing of only the exomes, so the protein coding part of the human genome. While this technique is very valuable, it misses about 99 % of the human genome. WGS on the other hand covers the whole human genome, at least what is technically feasible, enabling researchers to additionally investigate the non-coding regions of the genome which are equally important in understanding cancer. While WGS has been traditionally more expensive to conduct in the laboratory and requires more resources when saving and processing the much larger generated data, the additional insights are likely to add great benefit [38, 39]. Association of features in the non-coding part of the genome like promoters and other regulatory regions with cancer has highlighted the need to study the whole genome [40, 41]. Importantly, the cost of whole genome sequencing is becoming more and more affordable allowing for its more frequent use. From the bioinformaticians perspective, WES or WGS typically are the starting point for investigations, enabling routine things like mutation calling but also more advanced techniques like identification of mutational signatures.

DNA methylation

Methylation of the DNA is most commonly investigated in regards to the transfer of a methyl group to the C5 position of cytosine forming 5-methylcytosine. Such a methyl transfer most often takes place on cytosines located before a guanine (in 5’ to 3’ direction) referred to as CpG sites. DNA methylation plays a very important role in regulating gene expression, influencing the repression of proteins or the binding of transcription factors [42]. Methylation patterns can lead to different phenotypes even if the underlying genome is identical and can be highly tissue specific. Especially during tumorigenesis, DNA methylation and demethylation has been shown to play an important role in the initiation, maintenance and progression of cancer [43]. DNA methylation has become a proven biomarker in cancer diagnosis and classification leading to new insights enabling researchers to distinguish tumor subtypes not previously distinguishable by histology alone [44]. Microarrays are a cost effective and fast way to measure the methylation level at defined CpG sites throughout the whole genome, however the affordability has been decreasing in recent years. Methylation microarrays use a series of treatments with methylation sensitive restriction enzymes to measure the level of DNA methylation at each defined CpG site [45]. The most widely used commercially available microarray platform for methylation analysis is the Illumina BeadChip, primarily the 450K , EPIC and most recently the EPICv2 versions [46].

5.2 Unveiling genetic dependencies

In the search of more effective treatment for different types of cancers while simultaneously minimizing adverse effects to healthy cells, identification and leveraging of specific genetic vulnerabilities has been a promising avenue of investigation [47]. Especially the concept of synthetic lethality (SL), which was first described in fruit flies and yeast, has been used for treatment [48, 49]. Briefly, a synthetic lethality pair are two genes where disturbance of either does not have negative effects on cell survival but simultaneous disturbance of both leads to cell death. In a cancer cell that is dependent on the disturbance of one gene of an SL pair, the other gene of the SL pair becomes a suitable target for drugs that will then selectively kill only cancer cells (Figure 4). An example for the successful usage of this phenomenon is the well known SL relationship between BRCA1/2 and PARP1/2 which lead to effective treatment with PARP inhibitors in different cancer types affected by disturbance of BRCA1/2 such as breast cancer or ovarian cancer [50-52].

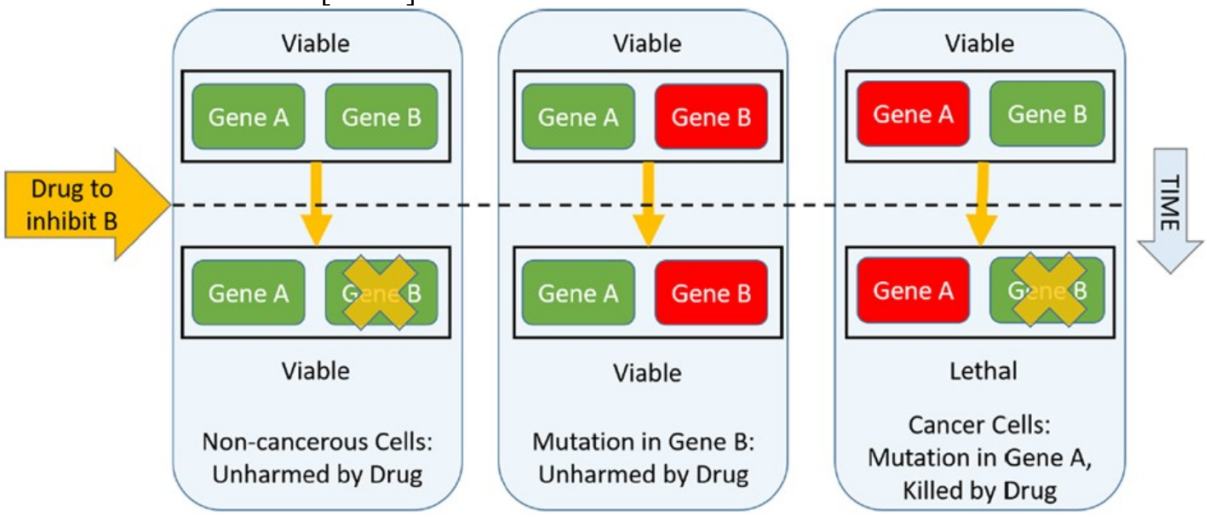


Figure 4: General concept of synthetic lethality. Two genes A and B don't lead to cell death if only one of them is disturbed. If gene B is disturbed, the cell can survive with no negative side effects. If gene A is disturbed and if that event is characteristic for a certain cancer cell type, these cancer cells can then selectively be targeted by a drug targeting gene B (and/or its products). Graphic adapted from Liany et al. [53].

Identification of such SL pairs which offer potential vulnerabilities useful for treatment is usually done with high-throughput approaches such as RNAi or CRISPR screens, leading to valuable insights into genetic dependencies [54-60]. However, such approaches are very resource and time consuming, because of the size of the combinatorial space that needs to be covered on top of the influence of the genetic background for each cell type that needs to be taken into consideration. This lead to increased interest in computational predictions of SL pairs being used to narrow down the scope of investigations to more tractable and likely candidates. So far, previously proposed prediction methods fall into two broad categories, knowledge based predictions and machine learning (ML) methods, the latter can in turn be split into classic ML and advanced ML methods. The knowledge-based methods use expert knowledge and assumptions about specific biological entities, e.g. copy-number changes or specific mutations, together with statistical methods to generate their predictions. A prime example for this method is DAISY, which leverages copy number profiles, expression profiles and shRNA screens to make SL predictions [61]. This makes the formulation of such a knowledge based prediction framework difficult for real world applicability, because a lot of entity specific expert knowledge is required up front. Another drawback of such knowledge based methods is that the integration of additional layers of data is very challenging since the assumptions behind the original model are very specific for a certain type of input. This issue can be addressed by using ML methods for SL predictions. On the more classical side of these methods there are approaches such as by Paladugu et al. where they extracted features from protein-protein

interaction networks and fed them into a trained support vector machine classifier to predict SL pairs [62]. Another approach by De Kegel et al. focused on SL relationships between paralog pairs using a mixture of features including neighbourhood, sequence and expression features in combination with a RF classifier to make predictions [63]. Common for classic ML approaches is the integration of features generated from multiple heterogeneous types of omics data, highlighting the ability to overcome the problem faced by knowledge based approaches [53]. It is important to note that these classic ML methods allow for robust performance even if the data used for training is somewhat limited in cohort size as is often the case with SL pairs. Promising even better performance are more advanced ML models which leverage graph neural network approaches for their predictions such as GCATSL, DDGCN or KG4SL [64-66]. These methods are reliant on substantial datasets for training, sometimes using entire online databases as input for example KG4SL. A major issue currently for advanced ML methods is that confirmed SL pairs are very limited and online databases often include unconfirmed *in-silico* predictions [67]. Additionally, classic ML methods produced predictions which could be later confirmed in separate experiments, demonstrating their usefulness selecting potential SL pairs [63].

5.3 Cancer predisposition and DADDR patients

Because of the earlier age of onset and lower mutational burden than adult cancers, pediatric cancers are not thought to be commonly related to environmental exposures because of lower exposure time. Rather at least 10% of pediatric cancer patients harbor an underlying cancer predisposition syndrome (CPS) [68]. CPS are often defined by alterations in genes already known to be associated with cancer from adult cancer, such as *TP53*, but recent efforts expanded that list to include *ELP1* which is associated with pediatric medulloblastoma [69]. That number is suspected to increase in the coming years when more cancer driver genes are identified and linked to a specific cancer type. For the treating physician, it is important to know about the presence of a cancer predisposition syndrome to avoid syndrome-specific increased toxicity, for example avoiding DNA damaging agents in patients who are more predisposed to DNA damage, or resistance to certain treatments.

Even before cancer is diagnosed, it may be important for an individual to know about their CPS syndrome, which can have great impact for a patient and their relatives. Individuals with a CPS have a significantly higher risk of being diagnosed with cancer throughout their lifetime and it is generally accepted that earlier detection of cancer, when the tumors have not had time to progress or metastasize, improves overall outcome [70]. Harmonization of recommendations for surveillance programs targeting patients with CPS syndromes resulted in well accepted AACR guidelines but simultaneously is ongoing work and will remain dynamic as new information becomes available [71]. One study proposed an easy to use questionnaire to determine if a patient might benefit from further genetic assessments based on five characteristics associated with cancer predisposition: family history, specific malignancies, multiple primary cancers, specific features and excessive toxicity [72]. Another study surveilling individuals with *TP53* mutation via biochemical analysis and imaging lead to a 88.8% 5-year-survival rate for the group under surveillance while the control group who declined surveillance had a 59.6% 5-year-survival rate, highlighting the importance of early detection and surveillance [73, 74]. The effectiveness of surveillance programs suffers from the tendency of patients to not participate any longer or even start if the protocol used for surveillance is too complex and/or time consuming [75].

Not all cancer predisposition syndromes are equal. Some types of cancer correlate with a certain CPS. Choroid plexus carcinoma for example has been linked to germline *TP53* mutation and colorectal cancer which is more commonly found in adults has been linked to mismatch repair deficiency (MMRD) having biallelic germline mutation *MLH1, PMS2, MSH2, MSH6*, acknowledging the fact that the reversal of these associations is not necessarily true [76, 77]. Li-Fraumeni syndrome and MMRD can be pooled with other syndromes under the umbrella

term “disorders with abnormal DNA damage response” (DADDR). Among these DADDR syndromes are those that affect single strand break repair pathways like Xeroderma pigmentosum or the already mentioned MMR deficiency. Other DADDR related syndromes affect pathways that are responsible for the repair of double strand breaks like Nijmegen breakage syndrome or Fanconi anemia [78].

This study mainly focuses on two DADDR syndromes: mismatch repair deficiency (MMR) and Li-Fraumeni syndrome (LFS). MMR is characterized on the molecular level by a germline mutation in one of the genes involved in human DNA mismatch repair system: *PMS2, MLH1, MSH2, MSH6*. In healthy cells the DNA mismatch repair system controls for single-base mismatches and INDEL loops that may arise during replication. Defects in these genes and therefore decreased possibility for the cell to control this kind of DNA damage leads to genomic instability and an increased risk of cancer (for a review see Peltomäki 2003) [79]. When talking about defects in the MMR genes one often differentiates whether they are heterozygous (Lynch syndrome) or homozygous/biallelic (constitutional MMR-deficiency syndrome CMMRD) [77]. The presence of Lynch or CMMRD syndrome also has influence on the management of patients with a reported higher response to cytotoxic drugs or a decrease in response to alkylating agents [80, 81]. Li-Fraumeni syndrome, characterized on the molecular level by the presence of a *TP53* germline mutation, is not directly related to DNA damage response. Rather *TP53* serves as a tumor suppressor gene regulating various other pathways in the cellular stress response. The detection of DNA damage and subsequent activation of the *TP53* pathway can result in (among other things) apoptosis, cell-cycle arrest or activation of DNA repair mechanisms [82]. Alterations affecting the function of the *TP53* pathway like germline mutations in *TP53* itself or alterations of upstream regulators are linked to early development of cancer as well as unusually early onset of certain cancer types for example breast cancer in women before menopause [83]. Overall, Li-Fraumeni syndrome is also linked to a higher overall genomic instability. Much like with Lynch/CMMRD syndrome, the management of Li-Fraumeni patients requires special attention and patients have been reported to benefit from regular surveillance measures [75].

5.3.1 Mutational signatures

During the lifetime of any person, their genome, on a cell individual basis, acquires thousands of somatic mutations, most of them without any effect on tumorigenesis. While some of these mutations happen at random, some can be attributed to a specific source or a specific mutational process. Comparable to a microphone in a crowded room, recording multiple conversations at once, the microphone would be the genome of an individual and the conversations are the mutational processes active simultaneously. There is considerable interest in deconvolution algorithms capable of identifying the mutational source processes from the mutational landscape of a genome, enabling researchers to leverage this knowledge for preventative care or treatment. In the comparison above, this would be an algorithm capable of restoring the individual conversations from the recording of the entire room. Pioneering work regarding deconvolution algorithms capable of obtaining mutational signatures used to describe these mutational source processes was done by Alexandrov et al. using non-negative matrix factorization (NMF) [84, 85]. The initial set of 21 distinct mutational signatures extracted from 7042 cancer cases has since been expanded to over 70 single base substitution (SBS) signatures and over 20 small insertions or deletions (ID) signatures in the latest release of the COSMIC database (cancer.sanger.ac.uk) [86]. Analyzing large cohorts of cancer samples revealed that some mutational signatures are ubiquitous across different cancer types while others are very specific for certain entities. Efforts linking the obtained mutational signatures to inherent characteristics of the samples or outside influences showed varying success. Some signatures, especially the ubiquitous, are speculated to represent clock-like processes and are heavily

correlated with the age of a patient [87]. Mutational signatures often active in lung cancer are reported to be linked to tobacco smoking which causes overproduction of C>A transversions [88]. Another example is signatures most active in skin cancer that were linked to UV light exposure [89]. Retroactively investigating and deducing what the sources causing mutations in certain cancers were is helpful for avoiding them in healthy individuals, but has little use for individuals already diagnosed with cancer.

Leveraging mutational signatures

As mentioned above, the vast majority of somatic mutations acquired in a genome throughout a lifetime are not related to tumorigenesis and are often referred to as passenger mutations. The set of mutations that are directly related to tumorigenesis and that are often very specific for a certain cancer type are called driver mutations [90]. These driver mutations can serve as reliable biomarkers for a decision about targeted therapy increasing the success of the treatment [91, 92]. While driver mutations are very useful, not all cancer types are linked to a driver mutation (yet). The number of known driver mutations is expected to go up in the future with further investigations. Mutational signatures can serve as a proxy biomarker in cases where no driver mutation is known since they are directly linked to the underlying mutational process and can give insights into the history of mutational processes that have occurred before tumorigenesis [84, 93]. A prominent example where mutational signatures can be used as biomarker is in tumors with defective homologous repair pathways. Mutations in this pathway force cancer cells to compensate by using other repair pathways such as non-homologous end joining (NHEJ). These alternative repair pathways are themselves not error free and produce a mutational pattern that is very characteristic [94-97]. Identification of samples with HR deficiency via mutational signatures is possible with dedicated tools (a prominent example is HRDetect by Davies et al.) and paves the way for recommendations on the targeted use of PARP inhibitors for such patients [98]. In their investigation, Davies et al. propose a regression model for the prediction of HR deficiency, taking various mutational signatures as input with a reported sensitivity of 98.7%. This model was not only able to detect samples with mutations in BRCA1/2, the genes causing HR deficiency, but was also able to find additional samples without mutations in these genes but with functional inactivation. This analysis revealed that the proportion of breast cancer patients with HR deficiency is roughly 4 times as high as previously thought, meaning that more patients could benefit from treatment with PARP inhibitors, showcasing a direct benefit for the patient from the analysis of mutational signatures.

Mutational signatures in pediatric cancer

Since pediatric cancers are different from their adult counterparts regarding multiple characteristics as discussed above, the mutational signatures active in them need to be studied independently. In independent pediatric pan-cancer studies different sets of driver mutations compared to adult type cancers were identified and the activity of mutational signatures was determined to be specific for distinct types of cancer [99, 100]. Gröbner et al. investigated 24 distinct types of pediatric cancer identifying multiple driver genes and actionable drug targets. In particular, they point out the importance of hereditary cancer predisposition in pediatric cancer with 7.6% of pediatric cancer patients being affected on average and for some types as many as 50% of patients carry such a germline mutation [101]. Regarding mutational signatures, Gröbner et al. identified signatures active across the entire cohort such as the clock-like signature Signature 1, signatures linked to genomic instability and *TP53* mutation (Signature 3,8 and 13) and signatures more active in certain types of cancer. For example, signature 16 was reported to be linked to pilocytic astrocytomas and signature 18 is linked to neuroblastoma, while the aetiology for both these signatures is unknown. They also describe a novel signature (signature P1) particularly active in the SHH subgroup of ATRT tumors, highlighting the fact that more in-depth investigations into mutational signatures active in

pediatric cancer is needed. Building on the work done by Gröbner et al, Thatikonda et al. focused more on the investigation of mutational signatures and used the latest set of algorithms and known mutational signatures for their analysis. In their work, Thatikonda et al. showed that pediatric cancer is affected by a relatively smaller number of mutational processes compared to adult cancers, and that much of the present somatic mutations (45.4% of SBS and 93.2% of ID mutations) are attributed to clock-like signatures SBS1, SBS5, ID1 and ID2. Due to using the latest set of known mutational signatures, they were able to refine the previously reported novel signature P1 and reported a novel signature themselves, IDN appearing to be active only in pediatric leukemia, further highlighting the need to study mutational signatures in pediatric cancers. Thatikonda et al. also briefly touched upon signatures linked to HR deficiency and their predictive power. They postulate that since in pediatric cancers the overall mutational burden is much lower compared to adult tumors multiple orthogonal signatures should be considered for assessment of HR deficiency. Currently, there is ongoing investigation towards developing a pediatric equivalent to the HRDetect tool (BRCAddict https://transcan.eu/output-results/funded-projects/brcaddict.kl).

5.3.2 Cancer methylome

The methylome of a cell, the methylation status of each cytosine located before a guanine in 5' to 3' direction (5'-CpG-3') throughout the genome, is affected just like the genome itself by somatically acquired DNA methylation changes and methylation patterns characteristic for the cell of origin. Taking the methylome of cells into account is very important since it has large effects on development, differentiation and phenotype of different cells even if the underlying genotype is the same [102, 103]. Furthermore, it was shown that methylation plays a key role in tumorigenesis, meaning that it is necessary to interrogate the methylome to understand the full context of alterations in a cancer cell [104]. The preserved methylation pattern allowing to trace the origins of a cell is especially useful in the classification of cancer that is not distinguishable by histological characteristics alone or to characterize metastasized cancer cells [105, 106]. Even if classification via histology is possible, classification via methylome still offers advantages because it removes the influence of human error. From a practical viewpoint, methylation exhibits higher stability than transcriptomic data or proteomics and it essentially shows a bimodal distribution of methylation values making it very suitable for use as a biomarker [107, 108]. Another key factor making the methylome so attractive for investigations into biomarkers is the fact that one does not need much sample material and that processing at different laboratories yields reliable results [109].

Leveraging the methylome

The methylome of cancer cells, mainly from solid tumors, has led to novel sub classifications of tumors previously thought to be homogenous entities [110-112]. To address the interobserverational variability in hisopathological diagnosis shown in multiple cancer subtypes, Capper et al. demonstrated in a landmark study, that the cancer cell methylome can be used to reliably distinguish between a large cohort of diverse cancer subtypes, in this case brain tumors [44, 110, 113, 114]. In total, they considered 91 by methylation distinguishable classes of brain cancer, some of which were equivalent to known WHO classes while others represented novel but distinct subclasses. This was achieved by training a random forest classifier followed by logistic regression on the methylation data of the training cohort leading to highly robust prediction of the correct class with small error rates. In cross validation, an AUCROC of 0.99 and an error rate of 4.28% were achieved. Testing the prediction method in practice resulted in reclassification for 12% of total cases with subsequent amendment of the histopathological assessment in favor of the predicted tumor class. With their work, Capper et al. demonstrated the valuable impact that classification based on methylation can have on diagnosis although in their study they mainly focused on CNS tumors and used a mixture of

adult and pediatric samples. In a follow-up study focusing on pediatric patients with CNS tumors, Sturm et al. confirmed the applicability of methylation based tumor classifier leading to a change in treatment protocol for 5% of samples with a very high proportion in tumors that were histologically diagnosed as high-grade gliomas, but molecularly appeared to be low-grade [115]. Simultaneously Sturm et al. also point out the prevalence of cancer predisposition syndrome in pediatric brain tumors and the fact that for those samples the methylation based classifier often gives unreliable predictions, showing the need for further investigation into such samples with the main suggestions being the inclusion of more such samples in the discovery cohort or even a dedicated cohort.

Methylation patterns in cancer predisposition patients

A common cancer predisposition syndrome in pediatric cancers is Li-Fraumeni syndrome characterized canonically by germline mutation in the *TP53* gene and often clinically diagnosed by Chompret criteria [116]. Samuel et al. investigated the differential methylation in blood leucocytes across multiple cancer types between patients with and without germline *TP53* mutation [117]. They identified hypomethylation of microRNA miR-34A to be relevant, however their statistical analysis did not take the different tumor types in the cohort into account. Another study by Subasri et al. investigated the associations of characteristic methylation patterns with germline *TP53* mutation and other mutations based on blood leukocytes [118]. They found that multiple genes linked to cancer are mutated in LFS patients and concluded that LFS should not be viewed as only being driven by *TP53* status. Unfortunately, only a small subset of their methylation samples was complimented by either WES or WGS giving insights about further somatic mutations of the patient limiting the study in that regard and the overall age distribution of the cohort was rather large ranging from 0-70.4 years of age with the older patients being mainly breast cancer patients. Because changes in methylation are strongly correlated with age, investigating a discovery cohort with a drastically different age structure compared to the target demographic, in this case pediatric patients, might be detrimental to translational efforts [119-121]. Taking a step towards early detection of Li-Fraumeni syndrome, a study on the application of liquid biopsy as a potential complement for traditional screening methods was conducted by Wong et al. [122]. With their blood samples and a curated set of methylation cancer markers described by Vrba et al., Wong et al. described a pan-cancer, LFS specific methylation pattern [123]. While the methylation signal accessible via liquid biopsy yields some valuable insights, analysis of tumor samples accompanied by WGS data is still needed to refine previous results.

5.4 Aims

With this thesis, I pursued three overall goals. First, I aimed to generate predictions of synthetically lethal gene pairs in pediatric high grad gliomas. The two subsequent aims were directed towards the investigation of samples from patients with cancer predisposition syndrome that took into consideration both genomic and epi-genomic data.

1) Prediction of synthetically lethal gene pairs in pedHGG K27M

There is considerable interest in the prediction of synthetically lethal gene pairs to narrow down the scope of investigation needed to be realistically covered in laboratory experiments for the identification of new specific vulnerabilities in tumors. With this study, I aimed to evaluate different data preparation techniques for the integration of heterogeneous sources of data and the performance of different prediction models for the prediction of SL gene pairs. Specifically, I focused on pediatric high-grade gliomas and the multi-omics dataset I curated from the INFORM cohort. First, I compared the performance of different models and different data preparation techniques between a dataset prepared from pedHGG K27M patients and another dataset prepared from other pedHGG patients. Culminating in the prediction of SL pairs specific for pedHGG K27M tumors.

2) Association of mutational signatures with different DADDR CPS syndromes

Mutational signatures can be a valuable tool for detecting underlying genomic features. In this study, I aimed to investigate the association of extracted mutational signatures with features of interest focusing specifically on the presence of germline mutations characteristic for DADDR patients. For this purpose, I called mutational signatures for a set of DADDR patients with high coverage WGS. Further, I compared two mutational signature calling algorithms and compared results to other studies conducted on pediatric cancer patients. Finally, I investigated the association of the identified mutational signatures with LFS and MMR DADDR syndrome and searched for possible intra-DADDR-syndrome variability.

3) Methylation patterns in DADDR patients across multiple cancer types

Capper et al. showed the ability of the methylome to serve as the basis for categorization of pediatric tumors into distinct clinically relevant classes. However, they noted that the performance of such categorization might deteriorate in the presence of a pathogenic germline mutation, which hinted at an influence on the methylome. With this study, I aimed to discover specific methylation patterns associated with germline mutations leading to a DADDR syndrome and investigated if this pattern is present across multiple tumor types. First, I assembled a dataset from the INFORM cohort and used the associated WGS data to make sure that there was no somatic mutation in any DADDR related gene possibly disturbing the analysis. Further, I matched tumor classes in my control cohort to the DADDR cohort, which made it possible to account for tumor type specific methylation. I analyzed this dataset with various statistical methods and ML models.

6 Materials and methods

Paragraphs in quotes were written by other authors than myself.

6.1 Prediction of synthetic lethality in pedHGG K27M

The following sections describe the extraction and preparation of data and the used methods for the prediction of synthetically lethal gene pairs.

6.1.1 Selection of known synthetic lethality pairs

For the selection of known synthetically lethal gene pairs to be used as training data I considered SynLethDB, a comprehensive database of synthetically lethal gene sets from various sources including experiments and *in-silico* predictions, as a prime source [67]. From SynLethDB I created a set of criteria that allowed for the identification and selection of high confidence SL pairs. Specifically, a pair was considered high confidence if it was derived from analysis of functional experiments such as CRISPR screens or other similar functional analysis and not *in-silico* predictions. I manually added known SL interactions from consultation with experts. Further, I annotated the selected known SL pairs with loss-loss, gain-loss or gain-gain. This annotation refers to the change in expression level/function both SL partners need to undergo to exhibit synthetic lethality. Loss-loss, meaning that both genes exhibit a drop in expression or lose their function, is the most common case for which synthetic lethality is described [48, 124, 125]. Induction of synthetic lethality if one or both genes are overexpressed, also known as dosage lethality, is less common and is annotated as gain-loss or gain-gain [126-128]. Overall I prepared ~1200 known SL pairs, the vast majority of which are annotated with loss-loss (~1000), because the vast majorities of studies these SL pairs are taken from were loss of function studies. For training of the prediction models I only used the loss-loss SL pairs because the experimental techniques for potential verification only allowed to mimic a loss-loss SL interaction.

6.1.2 Patient cohort, available data and data processing

Patients from which I extracted and aggregated data were selected from the INFORM cohort and are listed in the appendix. Briefly the INFORM pipeline takes samples from tumor and nonmalignant “germline” tissue (usually from a blood draw) as input which are then subjected to WES, WGS, lcWGS, RNAseq and other molecular profiling techniques. Raw sequencing data is further processed with bioinformatics tools for calling SNVS, INDELS as well as CNV. Further details on the INFORM processing pipeline are described in the original publications [21, 22]. From the selected INFORM patients, I assembled two datasets: a nonK27M dataset containing pedHGG cases classified as pedRTK1, PXA or OTHER and a K27M dataset containing exclusively pedHGG K27M cases. All tumor types used were generated in the context of the INFORM pipeline via the molecular classification algorithm and associated expert panels.

From each patient in these datasets, I extracted SNV, INDEL, RNAseq and CNV data for further processing with the steps outlined below to finally generate a specific matrix format which could be used as input for collective matrix factorization and subsequent machine learning models. In total, 1 matrix with known SL pairs and 8 data matrices were used for the predictions. Each data matrix exist as two version, one constructed from the nonK27M dataset and one constructed from the K27M dataset.

Coexpression matrix

The first data matrix is called the coexpression matrix. For the coexpression matrix I extracted expression values normalized to TPM from the RNAseq data and arranged them in a matrix

with patients as columns and genes as rows. From this matrix I calculated the pearson correlation coefficients for all possible pairs of genes across all patients in the matrix. After I calculated the absolute values of the correlation coefficients, I introduced a binary label where correlation coefficients equal to or greater than 0.8 were replaced by 1 while those <0.8 were replaced with 0.

INDEL matrices

The second and third data matrix contains information about recurrence and mutual exclusivity of INDELS. For the INDEL matrices, I counted INDELS and arranged them in a matrix with patients as columns and genes as rows that will subsequently be referred to as raw INDEL matrix. Only protein coding genes were considered in the downstream analyses. From this raw INDEL matrix, I prepared a recurrence matrix with genes both as columns and rows. This recurrence matrix was labeled with a 1 on the diagonal if a particular gene had more than 1 INDEL and 0 everywhere else. To calculate mutual exclusivity with the R package Rediscover (version 0.3.2), I prepared the raw INDEL matrix as binary matrix where any count equal or greater than 1 was replaced by 1 else 0 and used that as input for Rediscover [129]. I used Rediscover as described in its manual to obtain a matrix with information about mutual exclusivity.

SNV matrices

The fourth and fifth data matrix contains information about recurrence and mutual exclusivity of SNVs. For the SNV matrices, I counted SNVS and arranged them in a matrix with patients as columns and genes as rows called raw SNV matrix. Only protein coding genes were considered. From this raw SNV matrix I prepared a recurrence matrix with genes both as columns and rows. This recurrence matrix was labeled 1 on the diagonal if a particular gene had more than 10 SNVs and 0 everywhere else. To calculate mutual exclusivity with the R package Rediscover, I prepared the raw SNV matrix as binary matrix where any count equal or greater than 1 was replaced by 1 else 0 and used that as input for Rediscover. I used Rediscover as described in its manual to obtain a matrix with information about mutual exclusivity.

CNV matrix

The sixth data matrix contains information on CNVs. For the CNV matrix, I extracted the $\log_2$ fold change for each protein coding gene from the data produced by the INFORM pipeline. From this I calculated a binary matrix which was labelled 1 on the diagonal if a gene had a $\log_2$ fold change of smaller than -1 else 0.

Protein matrix

The seventh data matrix contains information on protein co-occurrence. I prepared the protein co-occurrence matrix from data downloaded from CORUM database (version 3.0) [130-132]. With this data, I calculated a binary matrix with genes as both columns and rows with a 1 in a cell if two genes contributed to the same protein complex and 0 otherwise.

Pathway matrix

The eighth matrix contains information on pathway co-occurrence. For the pathway co-occurrence matrix I downloaded data from MSigDB, specifically the HALLMARK pathways dataset (version v2023.2) [133-135]. With this data I calculated a binary matrix with genes as both columns and rows with a 1 in a cell if two genes occurred in the same pathway and 0 otherwise.

Integration of data layers

The whole data preparation and integration process is visualized in Figure 5. The preparation of matrices from the dataset together with external sources MSigDB and CORUM to compile, together with the known SL pairs, the input for CMFW model.

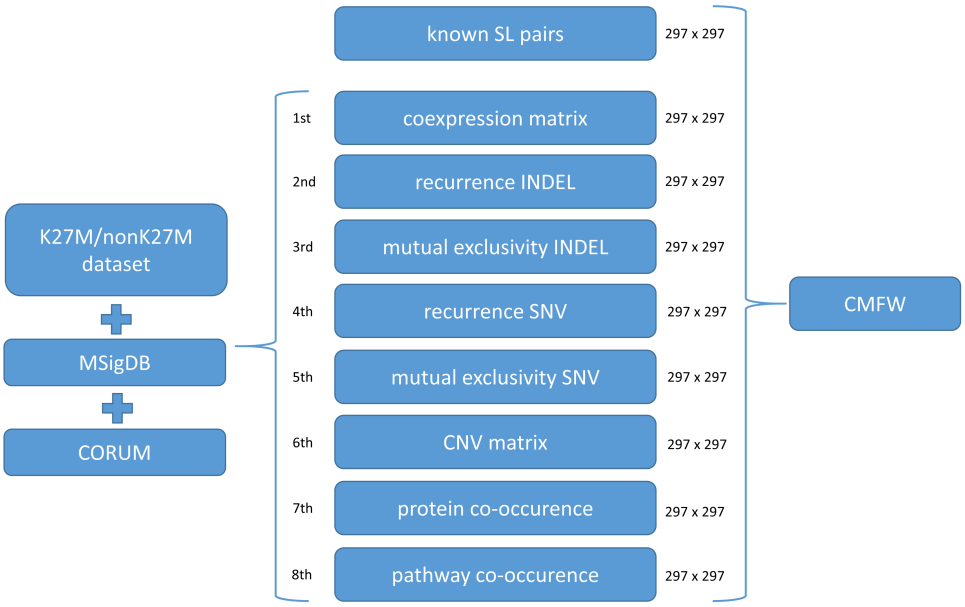


Figure 5: Data preparation for input into CMFW model. From either the K27M or nonK27M dataset, which I prepared as described above, 8 layers of data were extracted and saved in 297 x 297 matrices with genes both as rows and columns. In addition to these 8 data layers, I compiled a 297 x 297 matrix with known SL pairs, as described above. I used these in total 9 matrices as input for training the CMFW model.

6.1.3 Weighted collective matrix factorization model

Collective matrix factorization as proposed by Singh and Gordon aims to obtain low-rank representations of arbitrary collections of matrices, each matrix representing relations between two distinct entities used as row and column labels, by solving an optimization problem for which solutions can be obtained with stochastic gradient descent [136, 137]. One limitation of CMF is that only matrices relating two distinct entities, such as genes and patients, can be used but no matrices with the same entities in both rows and columns, e.g. genes in both rows and in columns. To overcome this limitation Liany et al. proposed CMFW, where a matrix specific weight is added representing the transformation in each input source responsible for different values and data types [53].

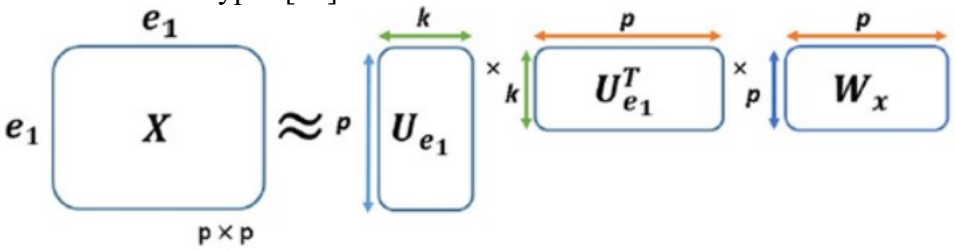


Figure 6: CMFW general concept. The matrix X with input values and same entities e_1 for both rows and columns is decomposed. The decomposition results in the matrices U_{e1} and U_{e1}^T with latent dimensions k (later referred to as KMAX), known from CMF, and the matrix specific weight W_X characteristic for CMFW. Graphic adapted from Liany et al. [53].

I implemented CMFW as described by Liany et al. with tensorflow (version 2.13.1) and used the Adam optimizer to solve the optimization problem [138, 139]. The input data matrices and the known SL pairs for training were prepared as described above. It is important to note that

CMFW was intended as and is used for matrix completion in this study, filling unknown entries in the matrices based on known data, not predicting interactions outside the matrices.

6.1.4 Data transformation with knowledge graphs

I also used the nonK27M dataset, the K27M dataset and the data layers extracted from them as described above, for making SL predictions based on classic machine learning models described below.

In addition to the described matrices themselves, I processed the data with graph based methods to engineer additional features attempting to better capture the interwoven nature of the different layers of data used in this investigation. For this purpose, an undirected graph was constructed with genes as vertices. The data matrices were interpreted as adjacency matrices and labeled edges between vertices were added if in a data matrix a 1 was present. This results in 8 types of edges being possible in the graph, one for each input matrix. Multiple edges between vertices are possible but only one per type. From the resulting graph, I extracted a set of features of both vertices and communities. The graph construction and processing was done with igraph for python (version 0.10.8), in braces are the actual igraph functions and their parameters [140]. Specifically, I extracted information on betweenness (betweenness(cutoff = 10)), pagerank (pagerank()), harmonic centrality (harmonic_centrality(cutoff = 10)), eigenvector centrality (eigenvector_centrality()), authority score (authority_score()), closeness (closeness(cutoff = 10)), coreness (coreness()), degree (degree(loops = False)), neighborhood size (neighborhood_size(order = 1)) , walktrap community (community_walktrap()), infomap community (community_infomap()), label propagation community (community_label_propagation()), leading eigenvector community (community_leading_eigenvector()), leiden community (community_leiden()) and multilevel community (community_multilevel()). Since these features are at vertex level, I moved them to edge level either by matrix product, resulting in a single feature, or by outer vector product, resulting in a matrix where features of the vertices are directly translated into features of the edges. All results presented in this study are obtained by processing with outer vector transformation unless explicitly stated otherwise. The features extracted from the graph were merged with the input features (the data layers used in CMFW) in a long format and used as input for the classic ML models. This process is visualized in Figure 7. Starting from the layers of data directly extracted from the nonK27M/K27M datasets as described above, the KG transformation (steps Ia, IIa and IIIa) takes place in parallel to the transformation into long format of the original data (step Ib) before concatenation and feeding into the classic ML models together with the known SL pairs (steps Ic and IIc).

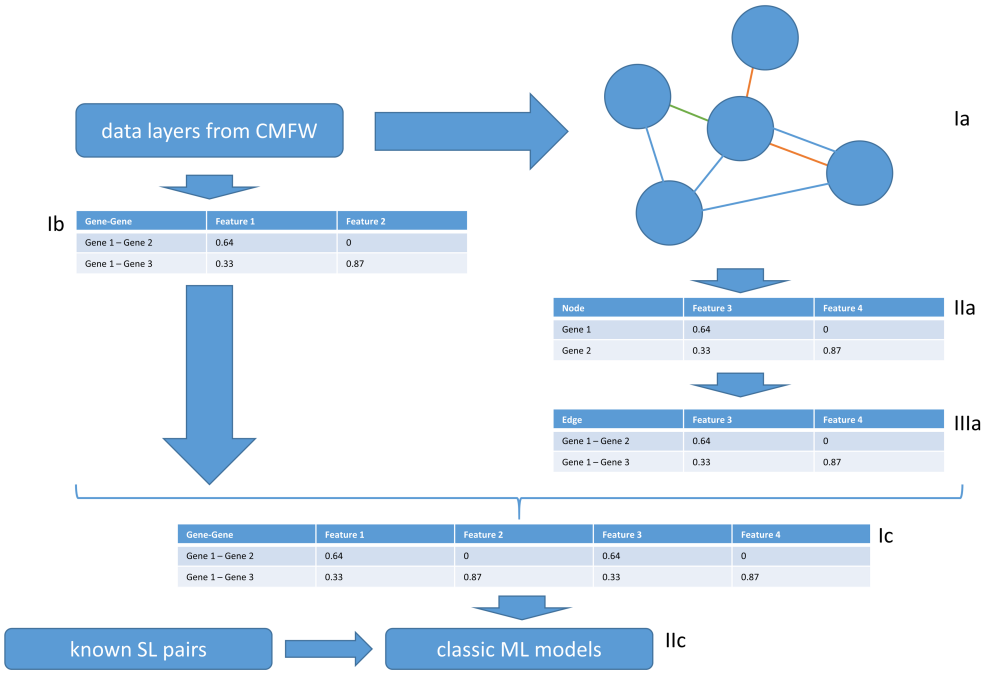


Figure 7: Processing via KG. The 8 data layers used as input into CMFW are interpreted as adjacency matrices and transformed into an undirected graph with genes as nodes (Ia). Multiple types of edges possible are possible with one type of edge for each layer of input data. From this graph features are extracted at node level (IIa). These features are then transformed to edge level (IIIa) either via outer product or dot product and then scaled. Simultaneously the 8 data layers are transformed into a long format (Ib). The data from IIIa and Ib are concatenated (Ic) and used as input together with the known SL pairs (also transformed into long format) for the classic ML models.

6.1.5 Classic ML models and evaluation

Three machine learning models, which I refer to as classic machine learning models, were used with the input data generated via knowledge graph processing. Specifically, from scikit-learn (version 1.2.2) the models RF (ensemble.RandomForestClassifier), ABC (ensemble.AdaBoostClassifier) and KNN (neighbors.KNeighborsClassifier) were used. Especially the RF method was already successfully used by De Kegel et al. for prediction of SL interactions among paralogs [63].
I conducted several test detailed below with these classic ML models as well as with the CMFW model to evaluate their performance and to investigate the connection from input data to model performance.

Dot product vs outer vector product

I calculated the evaluation metrics described below with only the original data (no KG processing), original data plus the aggregated KG features (dot product) and the original data plus the outer KG features (outer vector product). This test gives insights about which data preparation technique leads to better performance.

Shuffle test

Further, I conducted a shuffle test, where the labels of the learning data (here the known SL pairs) is randomly shuffled and the effect on performance is observed. The shuffle test was carried out with 5x cross validation. This test gives insights into the causal connection between input data and model performance.

Permutation test
On top of the shuffle test I calculated a p-value to assign significance to the observed differences in metrics via permutations (n = 1000).

Downsampling test
Lastly, I conducted a downsampling test, also with 5x cross validation, were I calculated the evaluation metrics with different levels of downsampling of either the negative learning class (nonSL pairs) or both classes.

Making predictions
With performance evaluation of the models done, I decided to move forward with the RF model to make the actual predictions. For this purpose, input data was prepared in the same way as described above, using the outer vector product method on KG features. The trained model was applied to all possible gene pairs where both partners showed a minimum mean expression of 20 TPM within each dataset across all patients in the K27M dataset. The obtained SL predictions were investigated via GO enrichment analysis with clusterProfiler (version 4.8.2).

6.1.6 Validation metrics
For the evaluation of the models with the tests described above, I chose three evaluation metrics. First, I used the well-known precision recall and the associated area under curve metric (PRAUC), which is very well suited for situations where the learning dataset is unbalanced [141, 142]. Second, I used the receiver operating characteristic and the associated area under curve metric (ROCAUC) [143]. Lastly, the Probability-at-N metric, giving information about the probability of how many true positive predictions are in the top N predictions. This metric is especially useful in this context, because while evidence for positive SL pairs is strong, evidence for negative SL pairs is weaker, meaning one could interpret negative SL pairs as unlabeled, making metrics from the area of positive-unlabeled training relevant such as probability-at-N [53, 144].

6.2 Multi-omics analysis of DADDR patients
The following sections describe the used data, processing of the data and investigation methods used for the multi-omics analysis of DADDR patients.

6.2.1 Patient cohort and available data
WGS
In preparation for the calling of mutational signatures, tumor samples and matching nonmalignant tissue were recruited from DADDR patients in the context of the ADDRess project (http://www.krebs-praedisposition.de/register/address/). These samples were supplemented with suitable DADDR patients from INFORM for which enough tumor material was still available to carry out additional high coverage WGS on top of the usual INFORM processing (samples listed in appendix). The WGS was carried out with a minimum coverage of 30x for germline control and 60-90x tumor coverage. The WGS was carried out with Illumina NovaSeq 6000 S4 by the DKFZ Genomics and Proteomics Core Facility (GPCF) and downstream processing was performed by the omics and Data Core Facility with the well-established and validated OTP workflow. Specifically, the OTP SNVCallingWorkflow (https://github.com/DKFZ-ODCF/SNVCallingWorkflow) was used for calling of SNVS and the OTP IndelCallingWorkflow (https://github.com/DKFZ-ODCF/IndelCallingWorkflow) was used for the calling of INDELS. These two pipelines were already successfully used in two other studies investigating mutational signatures [100, 145].

Filtering of the generated VCF files before calling of mutational signatures was done with Nagarajan's blacklist, a DKFZ in-house list used for filtering out common artefacts produced by the downstream pipeline that was constructed as follows: "A variant frequency > 1% in the DKFZ local control database consisting of 4879 WGS and 1198 WES samples was used to remove common artefacts and single-nucleotide polymorphisms from the somatic SNVs and indels." [146]. Tumor types were determined by INFORM molecular classification algorithm and associated expert panels.

Methylome data
Samples (listed in appendix) were subjected to methylation analysis with Illumina Infinium HumanMethylation450 BeadChip or Illumina Infinium MethylationEPIC to be processed and saved in the MNP database as described in Capper et al. [44]. Batch effect correction was applied with limma's removeBatcheffect linear model taking into account preparation of the tissue (frozen vs FFPE) and the Illumina chip (450 vs EPIC). CpG probes were filtered to contain the intersection of CpG probes from the 450K the EPIC and EPICv2 Illumina array and on top the 450K filter as described in Zhou et al. was used to remove certain CpG probes [147]. Tumor types were determined by INFORM molecular classification algorithm and associated expert panels.

6.2.2 Mutational signature calling

Preparation of mutational catalogues
I aggregated the VCF files with SNV and INDEL information generated by the in-house DKFZ sequencing and processing service as described above into several subgroups. From each subgroup I prepared a mutational catalogue, one for SNV and one for INDEL as described previously [85, 96]. Briefly, for the SNV mutational catalogue the 6 classes of base substitutions (C>A, C>G, C>T, T>A, T>C, T>G) are recorded as well as their immediate 5' and 3' bases resulting in 96 substitution classes. These 96 classes serve as row names in a matrix called mutational catalogue. As column names in a mutational catalogue, serve the samples contributing to the catalogue. The matrix cells are filled with counts how often a substitution class can be found in a sample.
For the INDEL mutational catalogues, I applied a similar process. 83 classes of INDEL described previously are used as row names [87]. As column names the samples are used again and the cells are filled with counts how often a given INDEL class occurs in a sample.

Reference mutational signature catalogue
As curated reference for mutational signature calling I downloaded the COSMIC SBS96 and ID83 signature catalogue (v3) [86].

Mutational signature calling with SigProfiler
Using the mutational catalogues as input, I did *de novo* extraction of mutational signatures with SigProfiler (SigProfilerExtractor version 1.1.22) [148].
The core concept at the heart of SigProfiler is a matrix decomposition given by the following equation.

$$M \approx S \, x \, E$$

In this equation, M is the mutational catalogue with mutation classes as rows and samples as columns and cells with counts how often a certain mutation was present in each sample. S, the signature matrix, is a matrix with mutation classes as rows and signatures as columns and cells containing the probability that a mutation is produced by a signature. E, the exposure matrix, is a matrix with signatures as rows and samples as columns and cells containing the contribution,

or exposure, of a signature to a sample. For a given mutational catalogue, SigProfiler estimates both S and E via non-negative matrix factorization (NMF). Overall, SigProfiler searches for an optimal number of novel signatures between 1 and 25 by minimizing the generalized Kullback-Leibler distance constrained for non-negativity. The selection for the number of novel signatures is based on the average stability of the decomposition and the error of the reconstruction. For each number of novel signatures, SigProfiler performs as default 100 independent NMF runs with the matrix M being resampled for every run. Clustering of the results from these 100 NMF runs is used to determine the most stable solutions.
The final selection of *de novo* extracted signatures is compared to known signatures, such as the COSMIC signatures, by calculation of cosine distance between signatures. *De novo* signatures are deconstructed into the known signatures if they pass a certain threshold of similarity. Completely novel signatures are reported as such if they cannot be reconstructed with known signatures.

Mutational signature calling with SIGNAL
In contrast to SigProfiler, I did not use SIGNAL (signature.tools.lib version 2.4.3) for *de novo* signature extraction [149]. Instead, I used it to assign the contributions of a known set of signatures. With regard to the matrix decomposition, this means that the matrix M is given as well as the matrix S describing the set of known signatures.

$$M \approx S \, x \, E$$

The aim is to obtain the exposure matrix E using NMF. This is implemented in the "FitMS" function in the mentioned R library. This function takes multiple steps for the extraction of mutational signatures. In the first step the most common mutational signatures are assigned before more rare signatures are assigned in a second step but only if they improve the total error above a certain threshold.

Splitting cohort for analysis
While I applied both SigProfiler and SIGNAL to the whole cohort, I also split the cohort into several subgroups and extracted mutational signatures again from these subgroups. The rationale behind this is to avoid bleed-over effects that can occur when analyzing a heterogeneous dataset [150]. The problem here is that the extraction algorithms assume that all samples share a somewhat similar mutational signature landscape and try to assign signatures accordingly. If a subset of samples is influenced by a different set of mutational signatures than the rest, this assumption is violated. For this reason, I split the cohort in two ways. Once by cancer types and once based on whether or not CPS was present. The exact group composition is given in the appendix. On top of splitting the cohort in the described way, if a subgroup contained hypermutators, I reanalyzed that group without the hypermutators since they can disproportionally impact the signature assignments.

POL* mutations
In this study POL* should be read as wildcard notation and refers to any of the following genes: *POLR2A, POLR3G, POLQ, POLD3, MIPOL1, POLG2, POLDIP3, POLA1, PAPOLG, POLE4, POLR1A, POLR1B, POLN, POLR2B, POLK, POLH, POLR3D, POLB, POLR1E, POLR3A, POLA2, APOLD1, POLR3B, POLE, POLE2, PAPOLA, POLR2M, POLI, POLRMT, POLD1, APOL5, APOL3, POLR2F, POLR3H, POLR3GL, POLR1D, POLR3F, POLR3K, PRIMPOL, POLR3C, POLD2, POLE3, POLR3E, APOL6, APOL1, POLR2E, POLR2D, POLR2J4, POLR2G, POLG, APOL2, APOL4, PAPOLB, POLM, POLR2K, POLR2C, POLDIP2, POLR2J, POLD4*

6.2.3 Statistical analysis of mutational signatures

For further statistical analysis of the extracted mutational signatures, I used R (version 4.3.0). I normalized the activity per sample to the total activity in that sample to enable comparison among samples. To obtain activity per megabase, I divided the activity by 2800, the effective size of the human genome in megabases that can be accessed by WGS.

To determine correlation of mutational signatures with age I used spearman correlation. I compared means of signature activity between groups with Wilcoxon test as previously described [151]. I investigated the association of traits with specific mutational signatures with linear regression models followed by ANOVA [152, 153]. P-values were adjusted for multiple testing via Benjamini-Hochberg method [154].

6.2.4 Purification of methylation signal

Tumor samples are a mixture of immune cells, normal cells and cancer cells. This results in a mixed signal when analyzing the methylation of a given sample. To extract the methylation signal specific for the cancer cells, I used an in-house tool called PROMISCE to purify the mixed methylation signal (publication pending). Briefly, the cell type composition of a given samples was estimated with EpiDISH [155]. These estimates for the composition of different cell types are the base for the purification of the methylation values. Using a reference database build from methylation profiles of very pure cell cultures of cells expected in a pediatric tumor sample, together with the estimates of composition, an expectation maximization algorithm determines the methylation signal specific for cancer cells.

6.2.5 Statistical analysis of methylation

I conducted the statistical analysis of the methylation data with R (version 4.3.0). First, I identified differentially and variably methylated probes and regions with linear models. Second, I assigned permutational importance to each probe using a random forest to select the most important probes. After these steps, I analyzed the identified probes and regions for functional enrichment, generated UMAP embeddings and calculated metrics to quantify the specificity for DADDR syndromes of the identified probes. I applied further downstream processing by embedding identified probes in correlated probe clusters via networks analysis to remove noise followed by additional enrichment analysis. Lastly, I used an internal and an external sample cohort for validation. P-values were adjusted for multiple testing via Benjamini-Hochberg method [154]. M values, which I use because of their continuous nature and because they offer better performance for differential analysis, are defined by the following equation [156]:

$$M = \log_2\left(\frac{Beta}{1 - Beta}\right)$$

Selection of reference samples for methylation analysis

For the statistical analysis of the methylation data I selected matching samples of the same tumor type without germline mutation from the INFORM methylation database to serve as control. The absence of somatic or even germline mutations in the control cases was confirmed by molecular profiling done in the context of INFORM. Additionally, I curated a set of samples from the INFORM dataset, harboring somatic mutations in at least one of the genes of interest.

Linear contrasts

Differentially methylated probes were identified with linear contrasts built on top of linear models using the Limma R package(version 3.56.2) [157]. The linear contrasts and models used are detailed in Figure 8. First, I used a model which does not account for tumor type, only for whether there is a germline mutation or not. With this model, I used a linear contrast referred to as contrast I which contrasts methylation between samples with germline mutation (CPS)

and non-germline mutation (NCPS). This model was both applied to the original M values and to the purified M values.

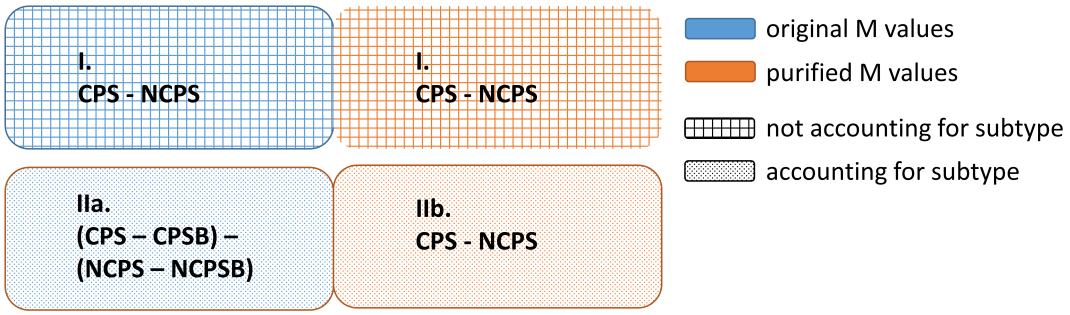


Figure 8: Linear models and linear contrasts used for identification of differentially and variably methylated probes and regions. CPS = cancer predisposition syndrome, NCPS = non CPS, CPSB = CPS blood, NCPSB = non CPS blood

On the second level, I used two different linear models accounting for the tumor type on top of the presence of germline mutation. Linear model IIa accounts for presence of germline mutation, tumor type and tissue (blood or tumor). For model IIa, the matching nature of the blood and tumor methylation data is incorporated by modelling patients as a random effect. The linear contrast for IIa investigates differential methylation regardless of tumor type while controlling for methylation in non-malignent tissue. Linear model IIb accounts for both presence of germline mutation and tumor type, with the linear contrast IIb being the same as for I. For the identification of differentially methylated regions, I used the same linear models and linear contrasts as described above for analysis with DMRcate, which allows for region based combination of clustered methylation sites (version 2.14.1) [158, 159]. For the identification of differentially methylated probes and differentially methylated regions, I used M values.

Variable methylation

Since not only differences in mean methylation are important, I also investigated the differences in variance of methylation between groups. For the identification of these variably methylated probes, I used the same linear models and linear contrasts as described above together with the varfit function from the missMethyl R package(version 1.34.0) [160]. For the identification of variably methylated regions, I used the same linear models and linear contrasts with DMRcate. For the identification of variably methylated probes and variably methylated regions, I used M values.

Pathway enrichment

For gene ontology (GO) enrichment, I used the function “gsameth” for differentially/variably methylated probes and the “gsaregion” for differentially/variably methylated regions, both from the MissMethyl package [160-163]. Contrary to a traditional over representation analysis, these functions take into account how many CpG probes map to each gene.

I extracted the GOTERMS from the org.Hs.eg.db package (version 3.17.0), specifically org.Hs.egGO2ALLEGS and filtered them for a minimum length of 5 genes and a maximum length of 2000 genes similar to MSigDB C5 dataset [133, 134]. Further I downloaded DNA

damage response pathways prepared by Pearl et al. and used them for enrichment analysis on top of the GO pathways [164].

Permutation importance

Next to the analysis with limma and DMRcate, I identified probes which contributed most to the differentiation between germline and non-germline cases by calculating their permutational importance with a RF. The filtered methylation data gets split into 36 equal chunks, each containing ~10000 probes. To each chunk, a RF is fitted, considering 600 features for splitting. After fitting the RF, I calculated the permutation importance for all probes. Afterwards, I concatenated the results from the 36 chunks and calculated the rank of each probe according to the permutation importance metric.

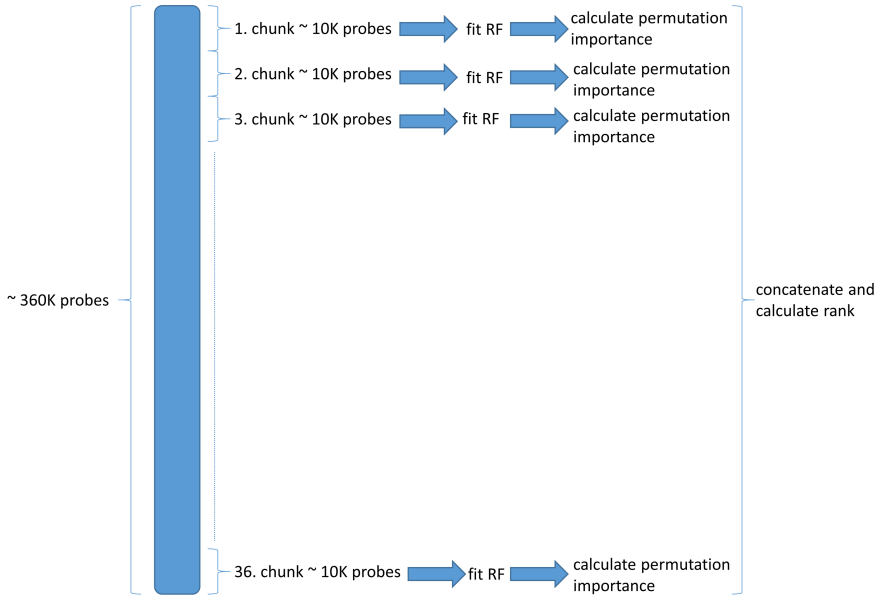


Figure 9: One round of calculations for permutation importance and according rank for each probe.

In Figure 9 I show the process of division into chunks and calculation of permutation importance. I repeated this process 100 times, each time shuffling the order of probes, before calculating the median rank of each probe. Selection of the most important probes with such an algorithm was already successfully applied by Capper et al. [44].

WGCNA

I performed weighted correlation network analysis using the WGCNA R package (version 1.72) [165, 166]. As mentioned above this method was applied to obtain clusters of correlated probes with the goal to reduce the noise inside these clusters.

GRAPH analysis

From a matrix of beta values with samples as columns and CpG probes as rows, I calculated the Pearson correlation across samples for all probes. The cells in the resulting correlation matrix were set to 0 if the correlation was smaller than 0.8 except for those cells which represent probes which were manually specified, for example probes identified as significantly differentially enriched. This updated correlation matrix was interpreted as adjacency matrix from which I constructed an undirected graph with probes as vertices. Edges were weighted with the absolute Pearson correlation as weight and colored according to the sign of the Pearson correlation. Loops and multiple edges were removed as well as vertices with no edges.

Communities were detected with the fast greedy algorithm. Graph computations were implemented with igraph (version 1.5.1) [140]. This procedure was applied with the same goal as WGCNA, to obtain clusters of correlated probes with reduced noise inside each cluster.

UMAP

I calculated the low dimensional embeddings of samples with UMAP, implemented in the R package umap (version 0.2.10.0) [167].

External control cohort

As external control, I downloaded the methylation data from LFS patients from the european genome-phenome archive (EGAD00010002461). To obtain the beta values I processed the data similar to what is described in Subasri et al. [118]. I used the ssNoob normalization from the minfi package, removed probes located on the sex chromosomes and removed all patients older than 21 years from the cohort and filtered for tumor types present in the discovery cohort investigated in this study. Afterwards I removed all probes not present in the discovery cohort. Finally I applied batch correction with Harmony [168].

7 Results

7.1 Prediction of synthetic lethality in pedHGG K27M

The aim of predicting synthetically lethal gene pairs was to narrow down the scope of investigation for a potential follow-up screen in order to identify specific vulnerabilities in pediatric high-grade glioma. To assess and compare the performance of different data preparation and prediction methods, I prepared two datasets containing different types of pedHGG cases. From the INFORM patients, I identified a subset of 149 patients diagnosed as pedHGG by the INFORM pipeline (methods 6.1.2). 70/149 pedHGG patients were classified as molecular type K27M and used for the K27M dataset. The rest of the pedHGG patients (79 patients) were classified as pedRTK1, PXA or OTHER and were subsequently used for the nonK27M dataset (Figure 10).

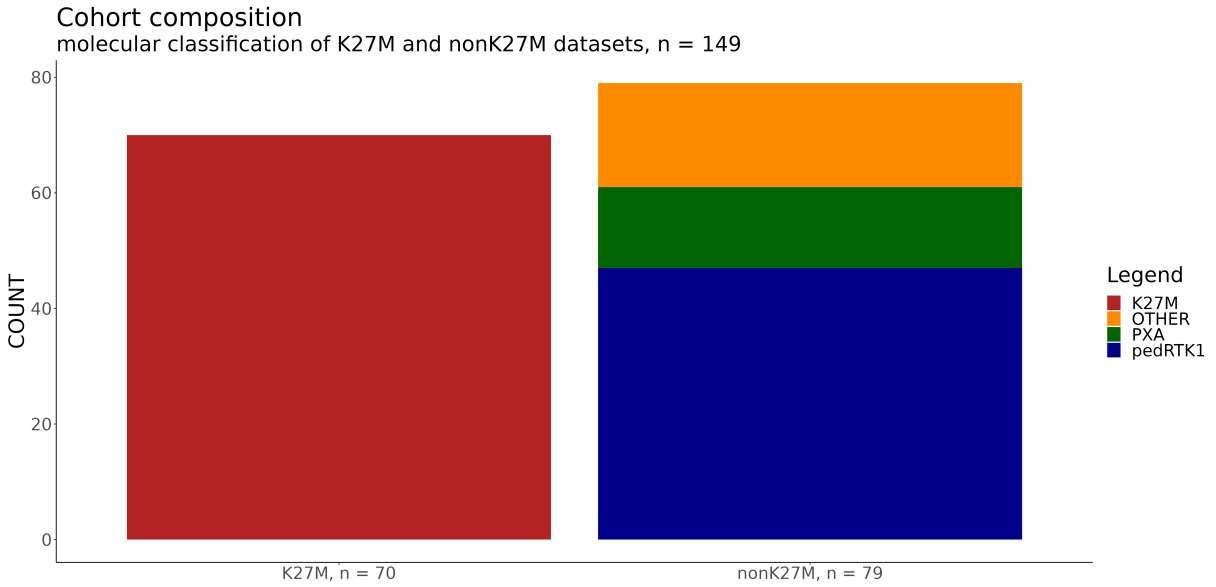


Figure 10: Composition of K27M and nonK27M datasets used for prediction of SL pairs. The K27M dataset is solely comprised of patients classified as molecular type K27M while the nonK27M dataset is comprised of patients classified as pedRTK1, PXA and OTHER.

With both datasets of roughly equivalent size, I evaluated the effects of different data preparation methods on predictive performance, specifically unprocessed data or processed via knowledge graphs. Further, I evaluated the predictive performance of different models, specifically collective matrix factorization, random forest classifier, k-nearest neighbours classifier or AdaBoost classifier.

7.1.1 Predictions with CMFW model

With both the K27M and nonK27M datasets as input, I performed detailed performance evaluation of the CMFW model. First, I investigated the influence of the KMAX parameter (methods 6.1.3). This parameter decided the dimensions of the components of the matrix decomposition. This dimension represented latent factors describing the most important features in the original data. For performance evaluation I mainly used area-under-curve (AUC, theoretical maximum indicating perfect performance at AUC = 1) values for the precision recall curve (PR) and receiver operating characteristic curve (ROC) that relates true and false positive rates. Below (Figure 11) it was immediately visible that, for both the K27M dataset (panel A) and the nonK27M dataset (panel C), there was a sharp initial rise in both ROCAUC and PRAUC from very low values of KMAX till a plateau was reached at about KMAX = 30 for the ROCAUC metric and at KMAX = 60 for the PRAUC metric.

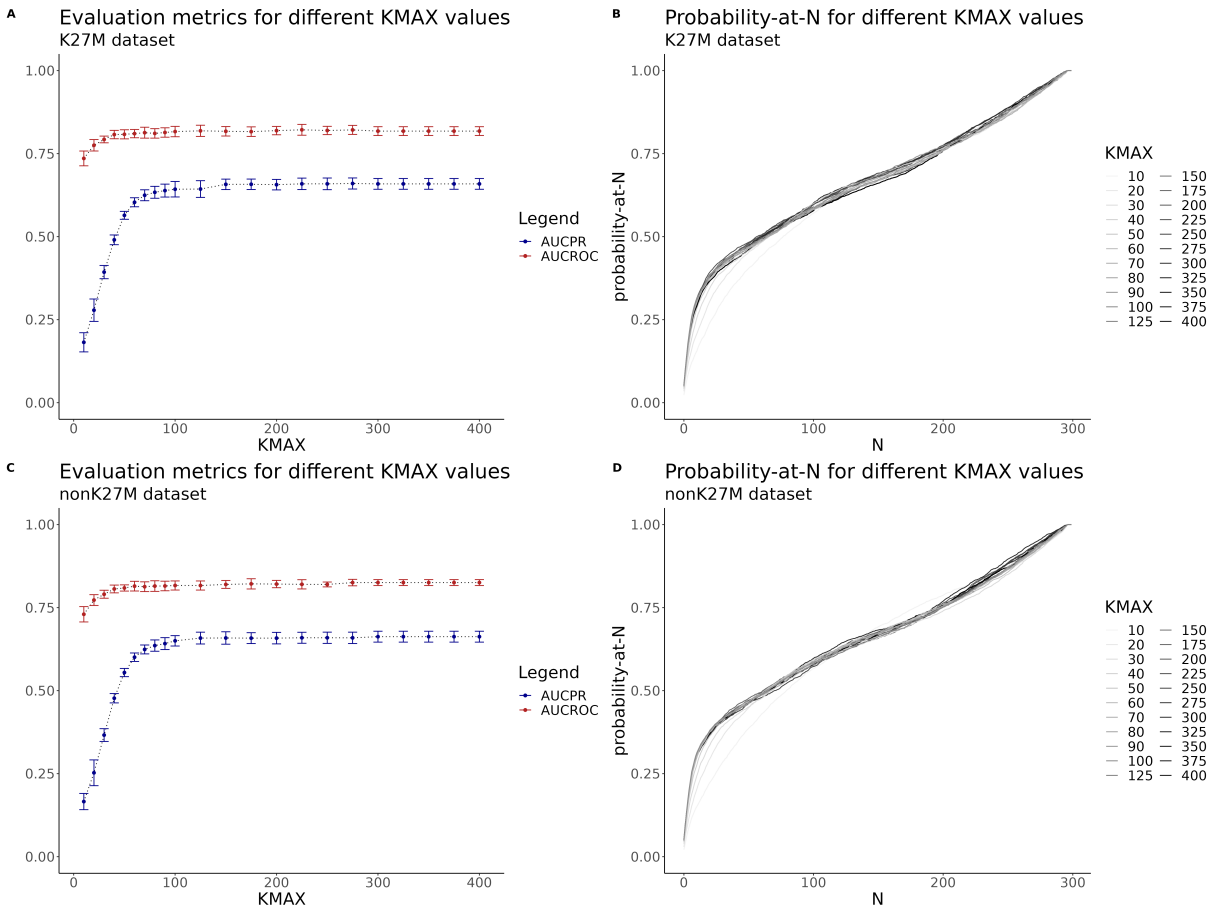


Figure 11: Performance evaluation of the CMFW model for selection of KMAX parameter, given the number of latent dimensions represented by KMAX. A) AUCPR and AUCROC for different values of KMAX with K27M dataset as input. Error bars calculated from 3x CV. B) Probability-at-N curves for different values of KMAX with K27M dataset as input. C) AUCPR and AUCROC for different values of KMAX with nonK27M dataset as input. Error bars calculated from 3x CV. D) Probability-at-N curves for different values of KMAX with nonK27M dataset as input.

I observed similar behaviour by the probability-at-N curves, desired behaviour of which was sharp initial rise to high values (methods 6.1.6), for both the K27M dataset (panel B) and the nonK27M dataset (panel D). For very low KMAX values, there was improvement, visible by a sharper initial rise in the curve, but starting from KMAX = 40 a plateau was reached with no further improvement in performance. Because of this observation, I decided to use KMAX = 60 for all further testing of the CMFW model. Next, I investigated differences in model performance depending on the composition of the input data to better understand the information offered by each incorporated layer and possibly remove redundant input data to reduce noise. To investigate this, I applied two strategies: first, I iteratively removed one by one layer from the input, second I used one source layer of data as the only input. For the K27M dataset (Figure 12 A and B) there were no significant differences visible for the tested compositions of input data in both the ROCAUC and PRAUC evaluations metrics. Similarly, there was no apparent improvement or decline in performance when looking at the probability-at-N curves. The same held true for the nonK27M dataset (Figure 12 C and D) with no improvement or decline in performance observed due to input data composition. The only deviation for both datasets was that there appeared to be a dip in performance of ~10% points as measured by AUCPR when using pathways as the only source layer of input.

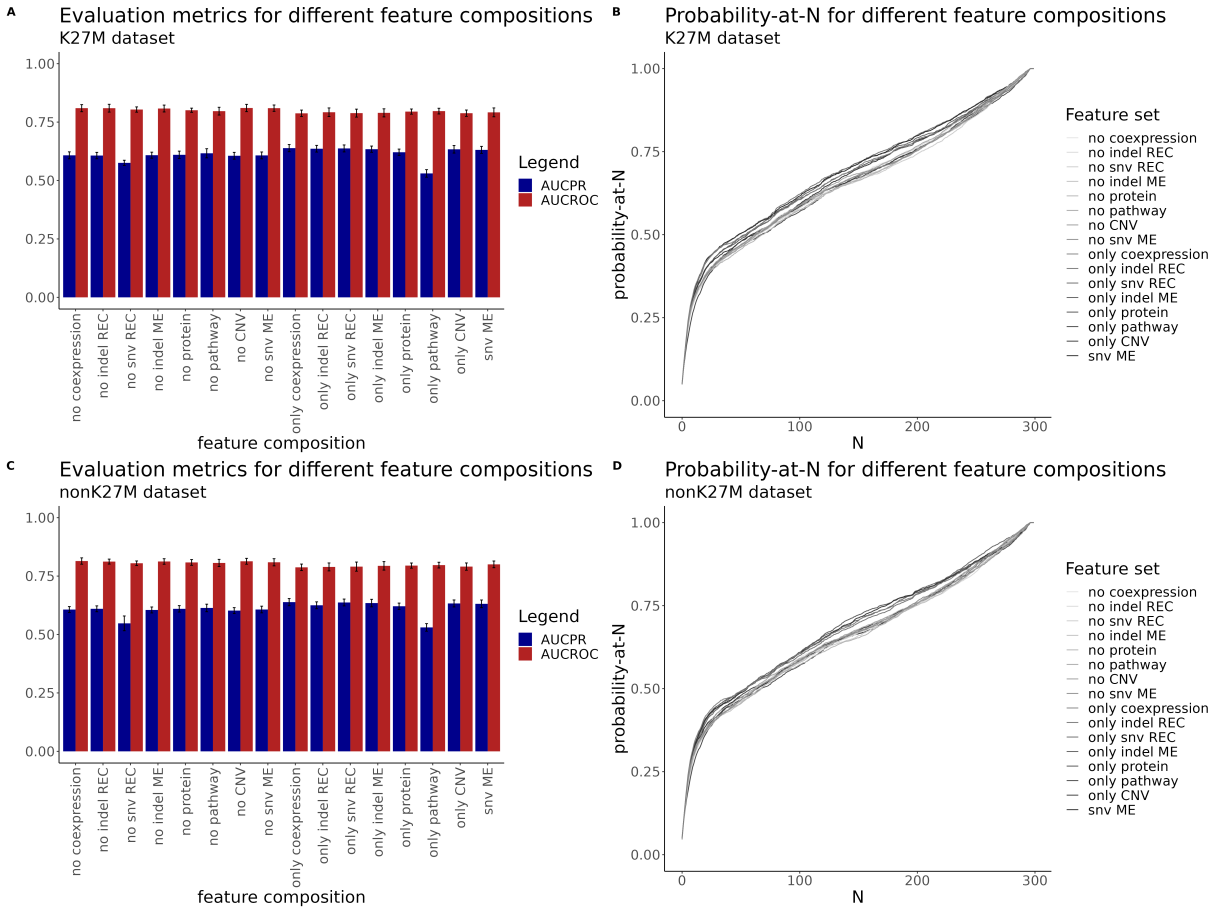


Figure 12: Performance evaluation of the CMFW model for selection of different composition of input data. A) AUCPR and AUCROC for different input value compositions based on K27M dataset. Error bars calculated from 3x CV. B) Probability-at-N curves different input value compositions based on K27M dataset. C) AUCPR and AUCROC for different input value compositions based on nonK27M dataset. Error bars calculated from 3x CV. D) Probability-at-N curves for different input value compositions based on nonK27M dataset.

This apparent lack of variation across all cases, except “only pathway”, regarding the composition of source layers of data was not expected. Especially because using only one source layer as input achieved comparable performance to the leave-one-out cases seemed counterintuitive. This prompted further evaluation of the CMFW model by investigating the causal connection between input data and performance. For this purpose, I shuffled the known SL pairs before training and calculating evaluation metrics again (methods 6.1.5). For the K27M dataset, (Figure 13 A and B) I observed improved performance across all metrics and against expectations for the shuffled data. The same behaviour was observed when using the nonK27M dataset (Figure 13 C and D). While not the desired behaviour, this was in line with the previously described behaviour of apparent indifference to composition of input data.

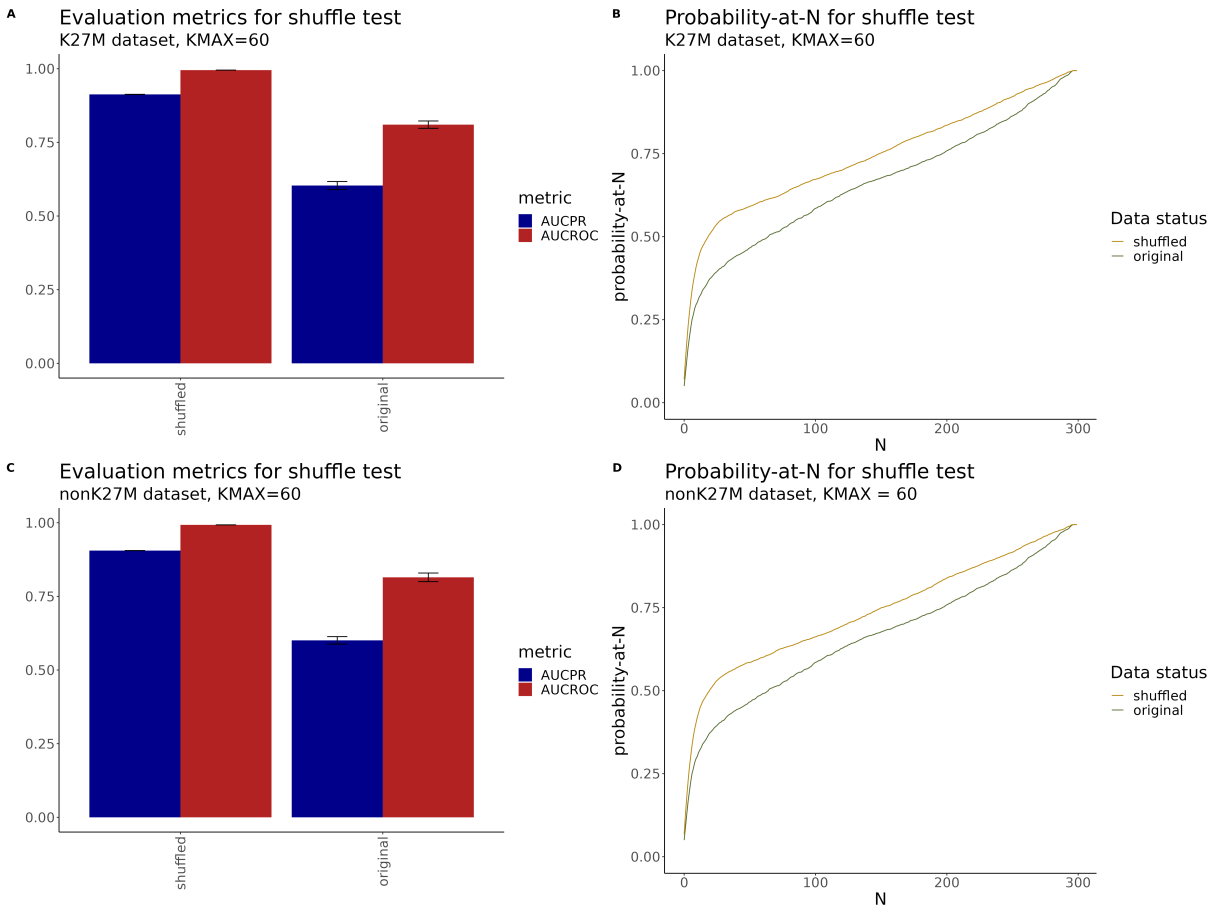


Figure 13: Performance evaluation how model behaves with shuffled or original input data. A) AUCPR and AUCROC calculated for original and shuffled input data from K27M dataset. Error bars from 3x CV. B) Probability-at-N curves for original and shuffled data from K27M dataset. C) AUCPR and AUCROC calculated for original and shuffled input data from nonK27M dataset. Error bars from 3x CV. D) Probability-at-N curves for original and shuffled data from nonK27M dataset.

The better performance when using shuffled data indicated that there was no causal connection between input data and model performance that was detectable by CMFW. The differences in performance measured were likely based on mathematical and technical artefacts, e.g. the specific parameters used for training. These results demonstrated that CMFW was not fit for my purposes and that using the known SL pairs and input data that was available to me with this method would not yield reliable results. These findings prompted me to explore other avenues for this analysis.

7.1.2 Predictions with classic ML models

Turning to the evaluation of the classic ML models, I focused mainly on ROCAUC and PRAUC as evaluation metrics because comparisons between these metrics are more widely used and studied compared to probability-at-N curves [169, 170]. In addition to the investigation of multiple ML models, I also looked at further data processing before feeding the data into a model. Biological systems are very complex and interrogation of only a single aspect, e.g. via RNAseq, makes it difficult to capture this complexity. Integration of multiple heterogeneous data sources via the early integration approach can offer benefits for capturing the complexity [171-173]. I hypothesized that such data integration might improve the performance because it captures the interwoven nature of the data better and mimics the implicit sharing of information across layers by the CMFW approach in a more explicit way. In particular, I decided to use a graph based approach for further data processing, since matrices are easily interpretable in the context of a graph and because graphs were used previously with great success in the area of SL prediction [63, 66]. To assess the feasibility of classic ML models, I used an AdaBoost

classifier (ABC), a k-nearest neighbours classifier (KNN) and a random forest classifier (RF). To set a benchmark for comparisons, I calculated the evaluation metrics for the investigated models with data before further processing via knowledge graphs (KG). In Figure 14 the metrics for ABC, KNN and RF models for both K27M dataset (panel A) and nonK27M dataset (panel B) are shown. The AUCROC was consistently around 0.5 for both datasets across all models. The values for PRAUC hovered around 0.45 with particularly large error bars in comparison to the ROCAUC values. The best PRAUC value was reached at 0.5 for K27M dataset with KNN and the worst value was at 0.3 for RF with nonK27M dataset.

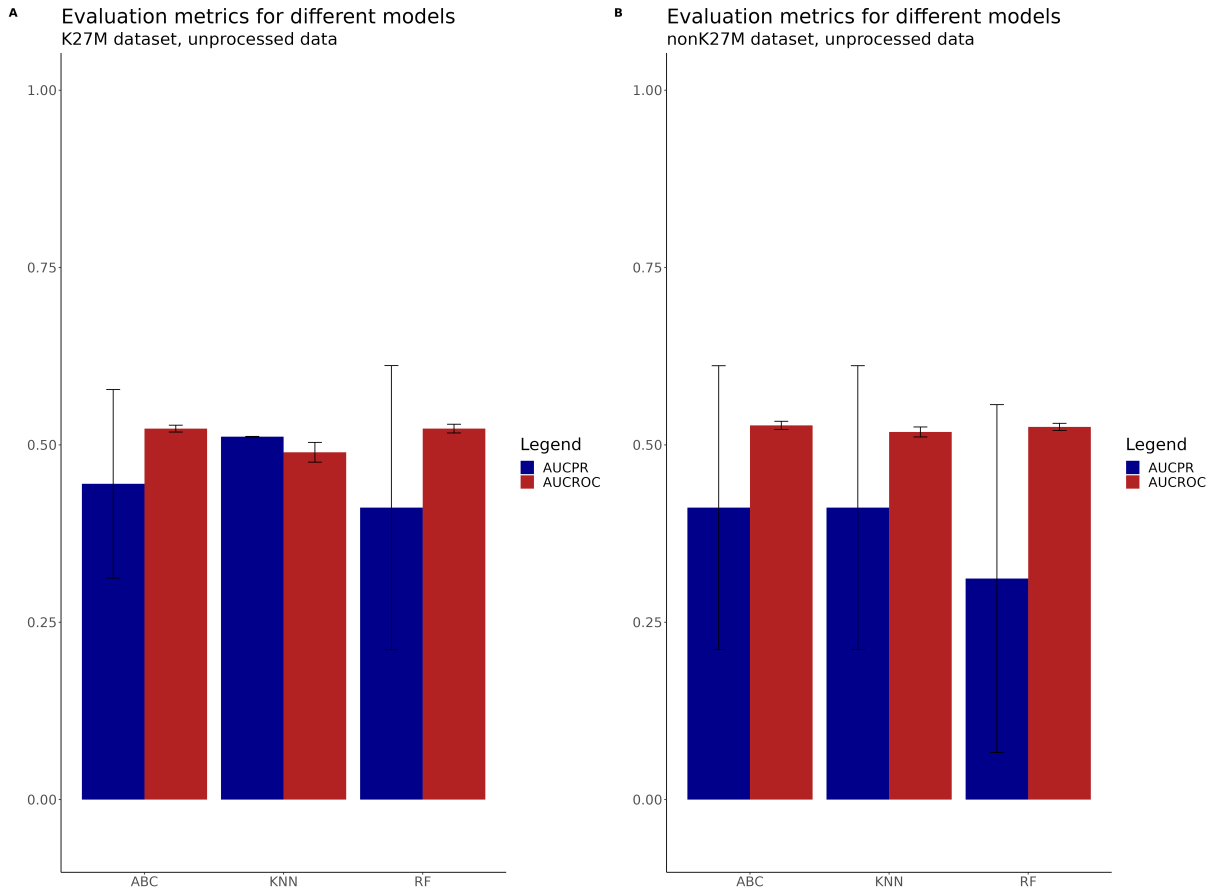


Figure 14: Evaluation metrics for the three classic ML models using unprocessed data..A) 5x cross validation of AUCPR and AUCROC for all three classic ML models with K27M dataset. B) 5x cross validation of AUCPR and AUCROC for all three classic ML models with nonK27M dataset.

As a next step, with the results from the CMFW model in mind, I proceeded to evaluate the causal connection between input data and performance again by shuffle test. In contrast to the CMFW model the expected behaviour was exhibited here. Specifically, I measured a decline in performance when calculating the evaluation metrics for the models with shuffled data. In Figure 15 panel A are the results from the shuffle test when using the K27M dataset. All three evaluated models showed a decline in performance for both metrics when comparing original and shuffled data. The most severe drop was exhibited by the RF model with a drop in AUCROC from 0.83 ± 0.01 to 0.5 ± 0.01 and a drop in AUCPR from 0.54 ± 0.02 to 0.035 ± 0.01. When using the nonK27M dataset (panel B) a slightly different behaviour was exhibited by the ABC and KNN models compared to using the K27M dataset. The ABC model had a drop in AUCROC as expected but the AUCPR showed hardly any difference between original and shuffled data. The KNN model showed a drop in AUCROC as well that was expected but the AUCPR, contrary to expectations, increased from original to shuffled data. Meanwhile the RF model also showed good performance with the nonK27M dataset for both metrics with the

AUCROC dropping form 0.87 ± 0.01 to 0.49 ± 0.01 and AUCPR dropping from 0.55 ± 0.01 to 0.02 ± 0.01.

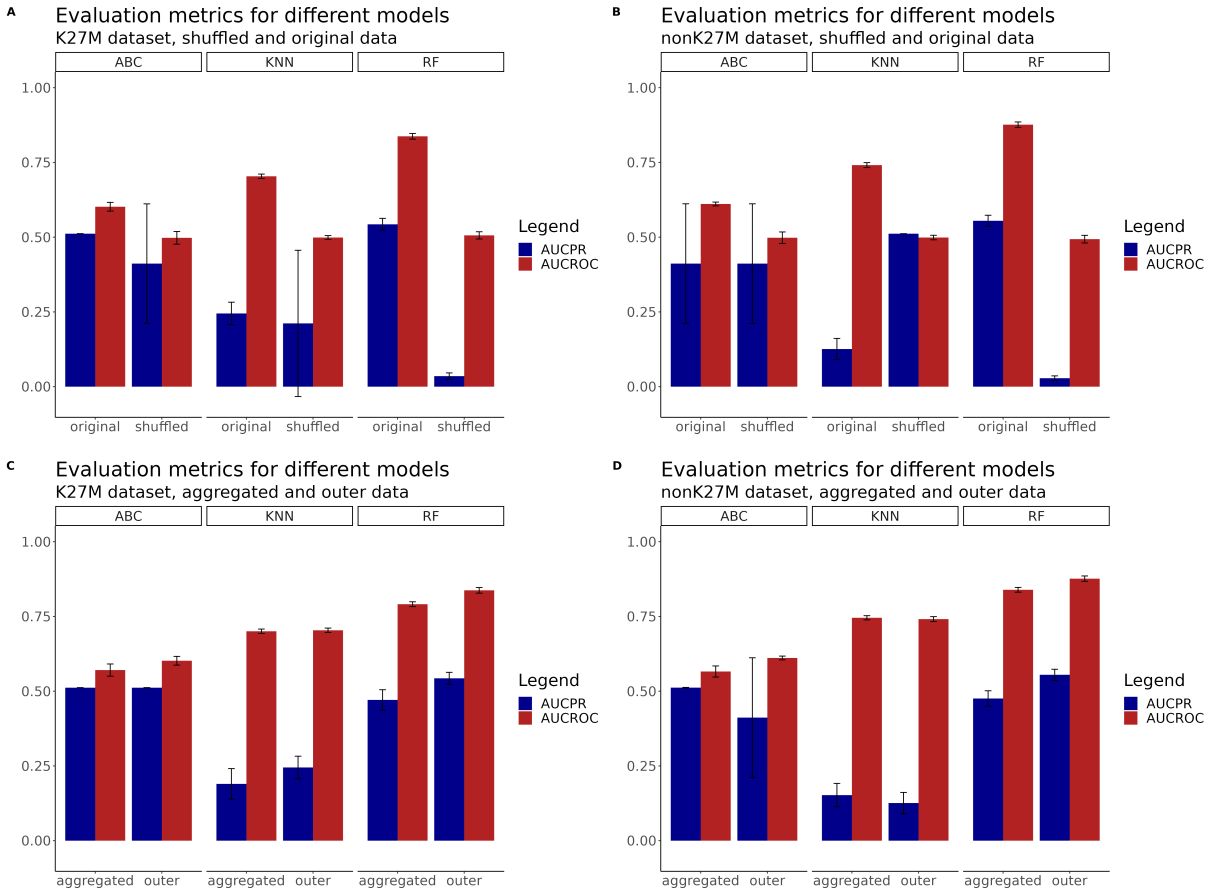


Figure 15: Evaluation metrics for shuffled or original data and for aggregated or outer transformation. A) Comparison of evaluation metrics for shuffled or original data from K27M dataset processed with KG and outer product across all three classic ML models. All metrics calculated as 5x cross validation. B) Comparison of evaluation metrics for shuffled or original data from nonK27M dataset processed with KG and outer product across all three classic ML models. All metrics calculated as 5x cross validation. C) Comparison of evaluation metrics for data preparation via dot product (aggregated) or outer vector product (outer) from K27M dataset across all three classic ML models. All metrics calculated as 5x cross validation. D) Comparison of evaluation metrics for data preparation via dot product (aggregated) or outer vector product (outer) from nonK27M dataset across all three classic ML models. All metrics calculated as 5x cross validation.

After confirming that there was a causal connection between input data and performance, I turned towards evaluation of data processing technique. The evaluation of the data processing technique showed a clear trend towards preparation via outer product, suggesting that this method effectively prepared data in a way that was beneficial for subsequent analysis. Figure 15 (panel C for K27M and panel D for nonK27M) demonstrates a clear improvement in performance for the RF model regardless of the used dataset. The ABC and KNN models showed an improvement or no change in performance between preparation techniques, but this was dependent on dataset input. Going forward I always used the data processing technique via outer product. Finally, I calculated p-values for the differences between metrics calculated from original and shuffled data. In Figure 16 (panels A and B) are the AUCPR and AUCROC values shown when I used the unshuffled data. Asterisks (***) above each column indicate if there was a p-value < 0.05 when compared to shuffled data. For the AUCROC value all models showed significant differences between shuffled and unshuffled data. For the AUCPR value only the RF method showed significant differences for both K27M and nonK27M dataset.

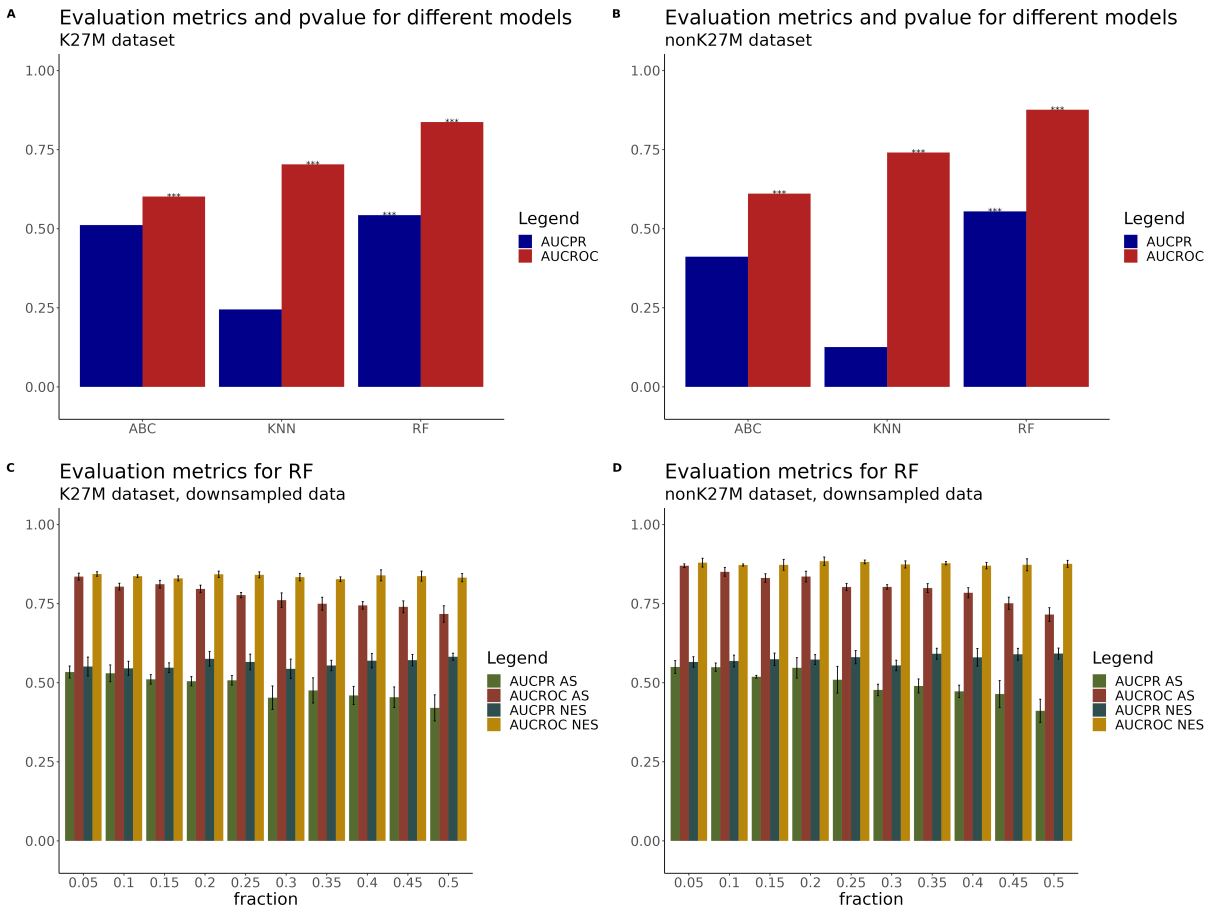


*Figure 16: Evaluation metrics from permutation testing and down sampling tests. A) AUCPR and AUCROC using the unshuffled, with KG and outer product transformed K27M dataset across all three classic ML models. *** indicates pvalue < 0.05 in comparison to same metric calculated with shuffled data as calculated by permutation testing. B) AUCPR and AUCROC using the unshuffled, with KG and outer product transformed nonK27M dataset across all three classic ML models. *** indicate pvalue < 0.05 in comparison to same metric calculated with shuffled data as calculated by permutation testing. C) 5x cross validation of AUCPR and AUCROC metrics for different downsampling strategies. AS means downsampling of both negative and positive SL pairs from training data. NES means downsampling of only negative SL pairs from training data. The RF model was used together with the unshuffled K27M dataset after KG and outer product transformation. D) 5x cross validation of AUCPR and AUCROC metrics for different downsampling strategies. AS means downsampling of both negative and positive SL pairs from training data. NES means downsampling of only negative SL pairs from training data. The RF model was used together with the unshuffled nonK27M dataset after KG and outer product transformation.*

With the expected behavior exhibited during the shuffle test and after selection of a data processing technique, I wanted to investigate the performance behavior for several downsampling techniques that further probed the causal connection between input data and model performance. In Figure 16 panels C and D show the evaluation metrics for the downsample tests. For both datasets, if only the nonSL pairs were downsampled for training there was no change in model performance for both metrics. This was expected since the nonSL pairs greatly outnumbered the SLpairs in the training data. The drop in performance at higher down sampling levels for the nonSL pairs could be explained with the processing via graphs. When too many vertices from the graph were removed, the graph became disconnected which had detrimental effects on the ability of the generated features to capture the remaining connectivity. When both the nonSL and SL pairs were downsampled equally, one could immediately observe a drop in performance for both datasets, with the AUCPR showing the effect even before the AUCROC metric. This again suggested a causal connection between input data and model performance. Based on these test results, I decided to go ahead using the RF algorithm, which showed the most robust performance for both datasets, together with the outer product preparation technique to make the actual predictions using both the K27M dataset and nonK27M dataset. For both datasets, I obtained 98 novel predicted SL pairs and 37527778 negative predictions, which was < 0.0001% positive prediction rate. While there was an overlap

(n = 69) between predicted SL pairs for both datasets, there were also such predictions that were unique for either dataset. Interestingly, the predictions of both datasets were made from the same 89 unique genes, which occurred in different combinations. Running GO term enrichment analysis on the 89 unique genes revealed a significant overrepresentation in a handful of pathways. Figure 17 panel A shows a cnetplot of the enriched GO terms, where the majority of genes fell into the mitochondrion GO term, which overlapped with the organelle envelope GO term. Another cluster was related to the Golgi apparatus, specifically the Golgi stack and Golgi cisterna. Two genes clustered away from these two GO terms. These were AP3S2 and AP3S1, two subunits of the same complex which fall inside the AP-3 adaptor complex GO term.

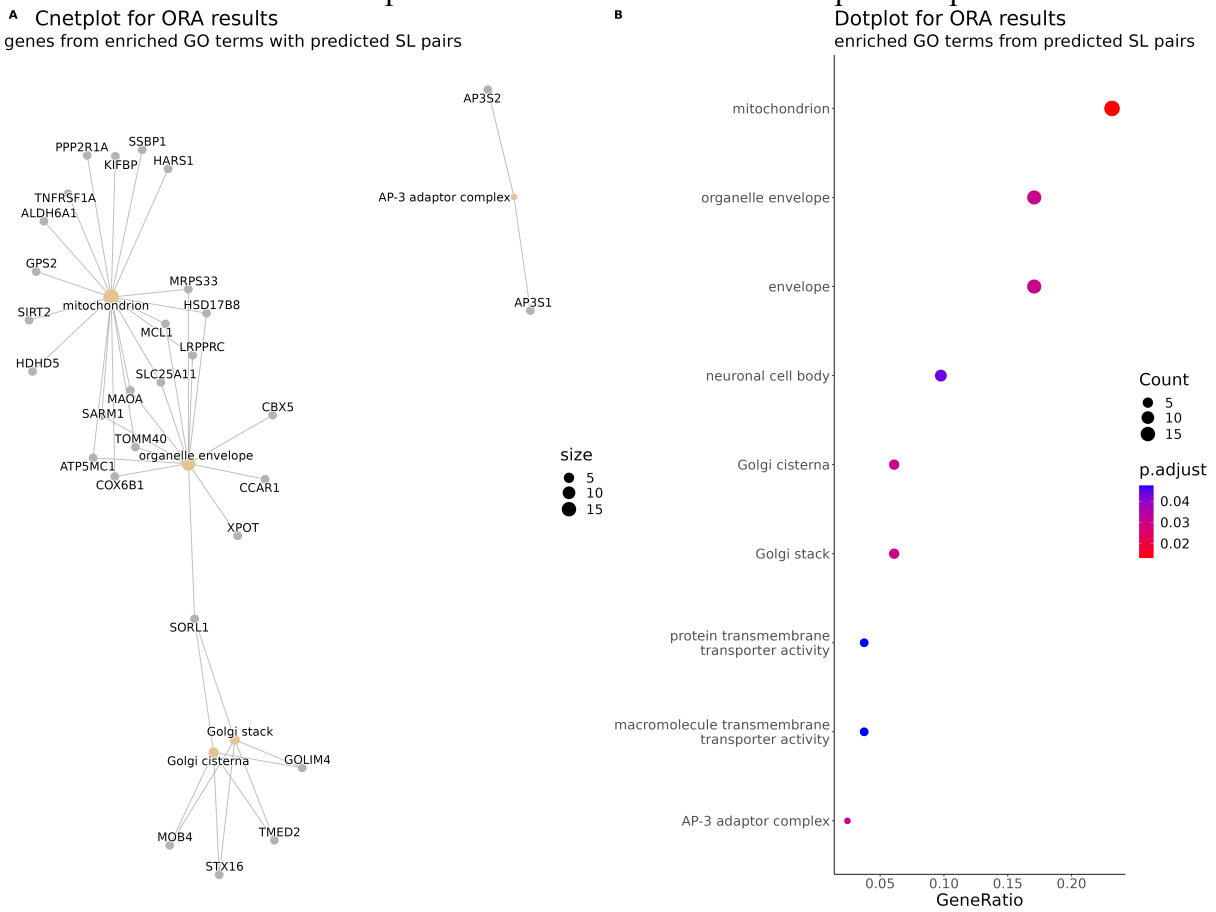


Figure 17: A) Cnetplot for unique genes from predicted SL pairs in significantly enriched GO terms. B) Adjusted pvalues and gene ratio for enriched GO terms using unique genes from positive SL predictions.

In panel B are the adjusted p-values of the enrichment analysis shown in more detail. I further filtered down the predictions and removed genes which were predicted to form an SL pair with themselves and removed duplicates that were predicted as SL in both orientations. This left me with the same 29 predicted SL pairs for both K27M dataset and nonK27M datasets, which were made of 55 unique genes. One concern with these predicted SL pairs was selection bias introduced via the training data [174, 175]. However as Seale et al. pointed out, topology based ML predictions methods are more affected by gene selection bias compared to feature based ML methods, which was the kind of technique used in this study to make the predictions. Interestingly the CMFW model, that was determined to be unsuitable here, is a topology based model which might be another reason its performance was lacking in this study. Another measure recommended to estimate the effect of selection bias is application of the prediction algorithm to multiple datasets, which I did by comparing nonK27M and K27M predictions, noting the good agreement, not only for the predictions themselves but also across the evaluation tests shown above. Further robustness to selection bias was introduced by my usage of context-free features, such as interaction network affiliation, and context-specific features,

derived from omics data, as recommended by Tepeli et al. [175]. After I filtered the predicted SL pairs, I repeated the GO enrichment that revealed that the mitochondrion GO term contained most of the predicted SL pairs.

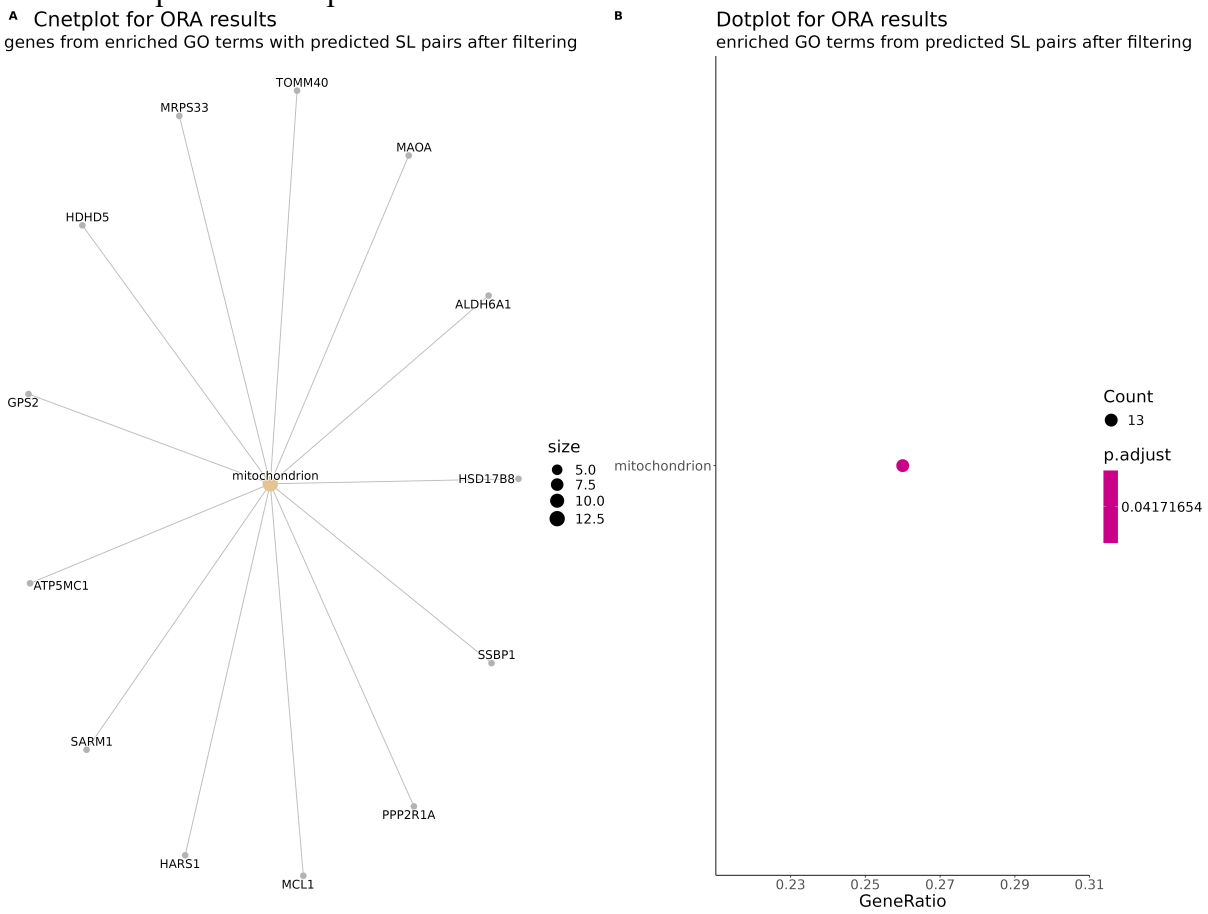


Figure 18: A) Cnetplot for unique genes from filtered predicted SL pairs in significantly enriched GO terms. B) Adjusted pvalues and gene ratio for enriched GO terms using unique genes from filtered positive SL predictions.

Figure 18 shows the GO term enrichment results for the filtered SL predictions and Table 1 shows the raw predictions, which include pairs not in the mitochondrion GO term.

Table 1: Filtered predictions made from both the K27M and nonK27M datasets.

Gene 1	Gene 2	Symbol 1	Symbol 2
ENSG00000268332	ENSG00000116675	-	DNAJC6
ENSG00000131238	ENSG00000168310	PPT1	IRF2
ENSG00000168710	ENSG00000178104	AHCYL1	PDE4DIP
ENSG00000143384	ENSG00000128016	MCL1	ZFP36
ENSG00000143727	ENSG00000115540	ACP1	MOB4
ENSG00000163930	ENSG00000161547	BAP1	SRSF2
ENSG00000144908	ENSG00000204228	ALDH1L1	HSD17B8
ENSG00000083896	ENSG00000269955	YTHDC1	FMC1-LUC7L2
ENSG00000170445	ENSG00000171720	HARS1	HDAC3
ENSG00000090263	ENSG00000106028	MRPS33	SSBP1
ENSG00000179388	ENSG00000138135	EGR3	CH25H
ENSG00000168003	ENSG00000087076	SLC3A2	HSD17B14
ENSG00000255639	ENSG00000111652		COPS7A
ENSG00000139613	ENSG00000179912	SMARCC2	R3HDM2
ENSG00000172458	ENSG00000266964	IL17D	FXYD1
ENSG00000165389	ENSG00000168175	SPTSSA	MAPK1IP1L
ENSG00000139990	ENSG00000176903	DCAF5	PNMA1
ENSG00000119711	ENSG00000189221	ALDH6A1	MAOA

ENSG00000157823	ENSG00000242498	AP3S2	ARPIN
ENSG00000167720	ENSG00000132382	SRR	MYBBP1A
ENSG00000185722	ENSG00000029725	ANKFY1	RABEP1
ENSG00000161920	ENSG00000132522	MED11	GPS2
ENSG00000004139	ENSG00000076351	SARM1	SLC46A1
ENSG00000141741	ENSG00000159199	MIEN1	ATP5MC1
ENSG00000267303	ENSG00000011451	-	WIZ
ENSG00000130175	ENSG00000105700	PRKCSH	KXD1
ENSG00000130204	ENSG00000105568	TOMM40	PPP2R1A
ENSG00000130204	ENSG00000069998	TOMM40	HDHD5
ENSG00000105568	ENSG00000069998	PPP2R1A	HDHD5

This table of predicted SL pairs includes multiple drug targets, making it a suitable resource for planning further investigations for example via combinatorial drug screens or CRISPR screens. These drugs include aldehyde dehydrogenase inhibitors, MAO inhibitors, HDAC inhibitors or BAP1 inhibitors [176-179]. Alternatively because the aimed at reduction in combinatorial space that needs to be investigated was substantial, low-throughput experiments of selected pairs might already be sufficient and bring novel insights. With this study I was able to further underline the suitability of random forest models for the prediction of synthetically lethal interactions, a result in agreement with previous studies [63, 174, 175]. I tested the presented method that integrates context-specific as well as context-free features to achieve higher robustness towards selection bias in training data on two datasets derived from pedHGG patients. The behavior across multiple tests, which demonstrated the ability of my method to capture causal connections between the training and input data, and regarding the predictions themselves were in good agreement for both datasets, an important characteristics hinting at robustness towards selection bias in training data. The context-specific features used for this method not only included the usual features like expression or mutation features but further included topology based features. Previous models for SL prediction that leveraged topology delivered good performance but severely suffered from a susceptibility to selection bias [174]. However in this study I was able to set up the model in a way that enabled use of topology based features while maintaining robustness. Another important aspect of this study was that in contrast to the majority of literature on SL prediction that uses publicly available data which is mainly derived from adult patients, here I used a dedicated in-house dataset from exclusively pediatric patients.

7.2 Mutational signature analysis of DADDR patients

With this analysis, I investigated active mutational processes in samples from patients with cancer predisposition syndrome (methods 6.2.1). In particular, I investigated mutational signature associations with Li-Fraumeni syndrome, MMRD syndrome and tumor type. Further, I compared the differences in mutational signatures relative to the respective gene(s) commonly associated with MMRD syndrome. I also investigating the connection of mutational processes to mutations in POL* genes, treatment and between patients with germline mutation vs somatic mutation or wild type.

7.2.1 Mutational burden

After I obtained the mutational catalogues, matrices containing information about which sample contained which mutations (SBS96 and ID83 context, methods 6.2.2) and how many, I checked whether the extracted mutations matched expectations regarding their number, correlation and distribution across samples before proceeding with the analysis of the extracted mutational signatures. Unexpected behavior regarding the number of mutations and the correlation between SNVs and INDELs could indicate a problem with the sequencing workflow or the mutation-calling algorithm. The correlation between SNV (or SBS) mutations and INDEL (or

ID) mutations by spearman correlation was as expected. These two kinds of mutations correlated significantly across the cohort with a correlation coefficient of R = 0.87 (Figure 19A). Some samples showed a hypermutator (> 10 mutations/megabase (MB)) behaviour, some even exhibited ultramutator (>100 mutations/MB) behaviour, where the hyper- and ultramutators appeared to be mainly associated with MMR related CPS, in this case *PMS2, MSH2, MSH6* and *MLH1*, a result in line with previous investigations [100]. This ultra-/hypermutator behavior was visible more clearly when I looked into the mutational burden produced by both the SBS and ID mutations across the whole cohort split by the different DADDR syndromes (Figure 19B). As expected, the cases with *BRCA1* and *BRCA2* mutations together with the cases with TP53 mutation exhibited a relatively lower mutational burden compared to the cases which carried mutations in *MSH2, MSH6, MLH1* or *PMS2*. This was clearly visible when observing the ID mutation results, while for the SBS mutations the division was less clear but still present since the hyper- and ultramutators were mainly present in the *MSH2, MSH6, MLH1* and *PMS2* groups.

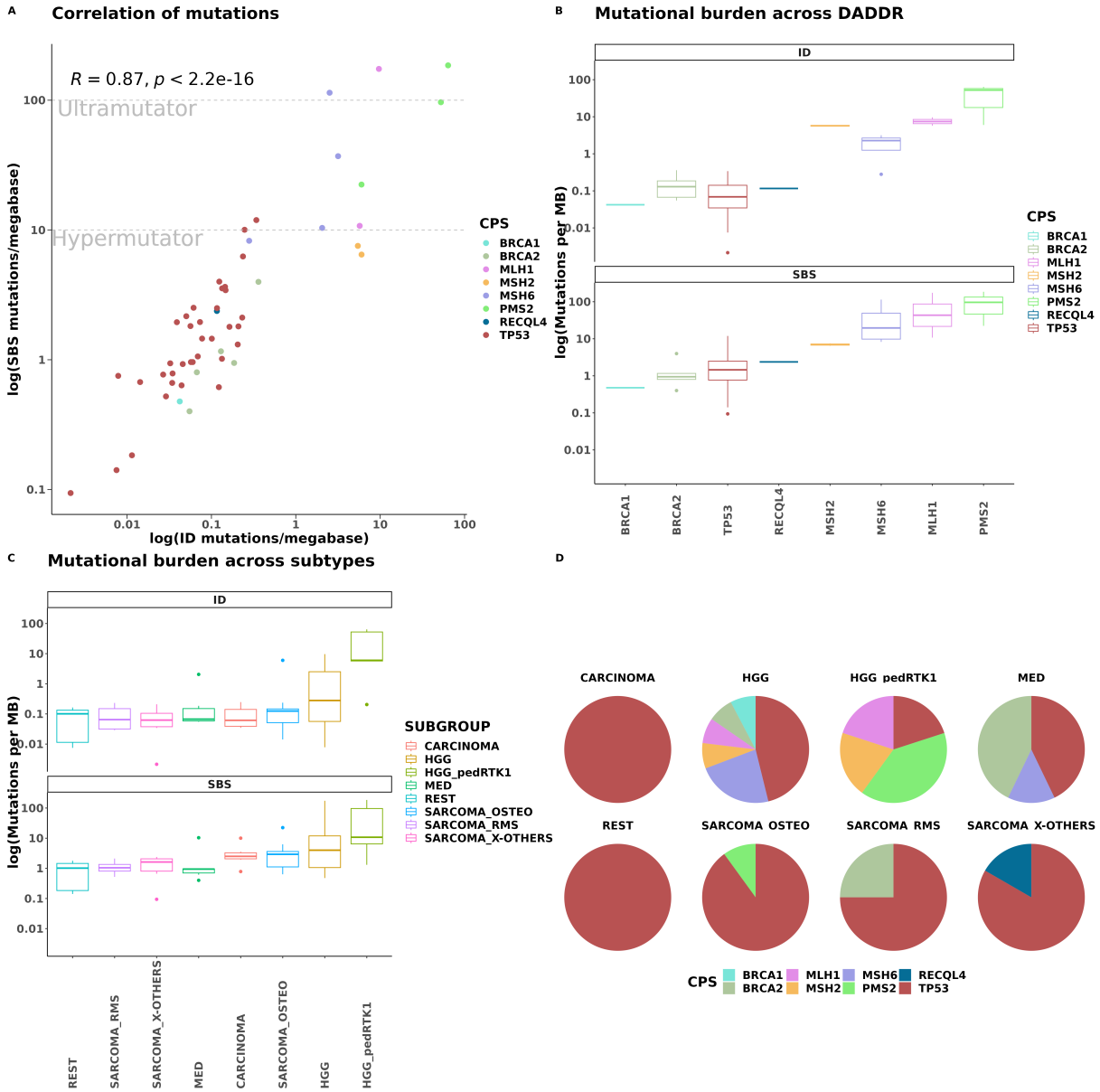


Figure 19: log = natural logarithm A) ID mutations per megabase over SBS mutations per megabase from all samples on a logarithmic scale. Spearman correlation coefficient and p-value are shown. Every dot is one patient. B) Mutational burden as mutations per megabase shown on a logarithmic scale across the different CPS present in the cohort for both ID mutations and SBS mutations. C) Mutational burden as mutations per megabase shown on a logarithmic scale across the different subgroups used in the analysis process for both ID mutations and SBS mutations. Pie charts show present CPS in each subgroup.

In panel C I show the mutational burden across the different cancer entities in the cohort. In case of ID mutations, all subgroups had a very similar mutational burden except for HGG and specifically the pedRTK1 HGG subgroup, which exhibited a slightly increased mutational burden. This was expected because these two subgroups majorly consisted of patients with mutations in MMR related genes that were associated with a higher mutational burden. A similar trend could be observed in the SBS mutations, where the HGG and pedRTK1 HGG subgroups had the highest mutational burden and all other subgroups again exhibited a roughly similar mutational burden. Statistical testing revealed that only the difference between the MMR and non-MMR subgroup was significant for both SBS and ID while the apparent difference of the HGG and pedRTK1 HGG was not statistically significant (graphic in appendix). Overall, all samples exhibited the expected behavior concerning their mutational burden, for both SBS and ID mutations, that reflected the CPS. Further, the mutational burden identified for the different cancer types was a reflection of the underlying CPS in the samples, as expected. This indicated that the workflow for sequencing and downstream processing did not introduce any artefacts or miss mutations, which was expected since it was already successfully used in multiple studies that investigated mutational signatures in pediatric cancers and consequently raised no concerns for further analysis in this study [100, 145].

7.2.2 Extracted SBS and ID signatures

Turning towards mutational signatures, in total I extracted 31 different SBS signatures and 12 different ID signatures with SigProfiler. Concurrently, I extracted 9 different SBS signatures with SIGNAL all of which were also extracted by SigProfiler. An overview of the extracted signatures and their contribution and proportion in the different groups is given in Figure 20 for SBS signatures and Figure 21 for ID signatures. As expected the clock-like mutational signatures SBS1 and SBS5 contributed to 93 % and 96 % of samples respectively when analysed with SigProfiler applied to all samples simultaneously. When analysed with SIGNAL, SBS1 and SBS5 contributed to 42 % and 84 % of samples respectively. Another prominent signature was SBS40, which contributed to 38 % of samples when analysed with Sigprofiler and was not assigned at all when analysed with SIGNAL. Instead SIGNAL heavily assigned SBS3 which has a high cosine similarity of 0.88 to SBS40 and was associated with mutations in *BRCA1* or *BRCA2*. That could indicate the assignment of SBS3 was a potential false assignment by SIGNAL. One reason could be inherent in the algorithm used in SIGNAL where in a first step a set of common signatures was fitted (which included SBS3) and in a second step a set of rare signatures (which included SBS40) was iteratively fitted but only kept if the overall error improved beyond a certain threshold. In contrast SigProfiler used simultaneous assignment of an iteratively increasing number of signatures and selected the best solution by overall stability. Because SBS3 and SBS40 had a high cosine similarity, additional fitting SBS40 could be judged by SIGNAL to not improve the error enough and therefore SIGNAL dropped SBS40 assignment altogether.

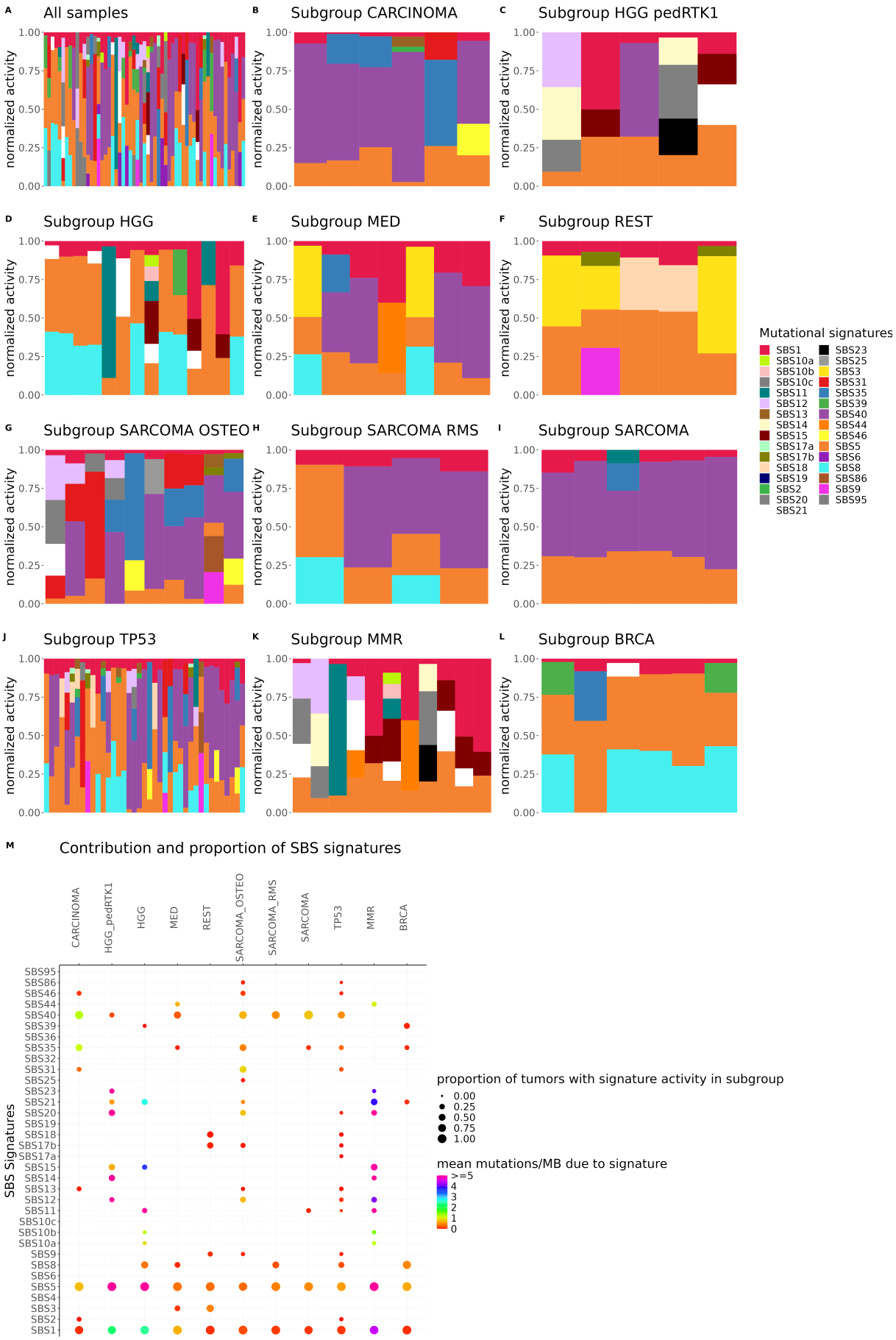


Figure 20: A - L) Mutational SBS signatures extracted by SigProfiler for different the different groups investigated. M) Contribution and proportion of each signature to the groups.

SBS8, which was proposed to be linked to HR deficiency and NER deficiency, was another signature that prominently contributed to 35 % of samples when analysed with SigProfiler and 98 % of samples when analysed with SIGNAL. SigProfiler most prominently assigned this signature in the BRCA subgroup, but also in other subgroups which contained samples with *BRCA* CPS such as HGG, MED and SARCOMA RMS. This was in line with previous results, that showed SBS8 contributes more to BRCA mutated cancers [180, 181]. SIGNAL on the other hand probably assigned SBS8 too much as a result of the different algorithm. As discussed above SBS8 was included in the first pass for assignment of common signatures. The signature linked to defective homologous repair and *BRCA1* or *BRCA2* mutation, SBS3, was assigned to almost all samples with SIGNAL (85 %) while to no samples with SigProfiler using cross-cohort analysis. Especially in the BRCA subgroup, were assignment of SBS3 would be expected, not a single sample was assigned SBS3 by SigProfiler, instead SBS8 and SBS2 were assigned. Perhaps this was due to the fact that assignment of SBS3 was reported to be linked with biallelic inactivation *BRCA1/2* mutation and not all *BRCA1/2* mutations [145]. SBS2 and SBS13, signatures associated with APOBEC activity, were assigned to 5% and 7 % respectively with SigProfiler, most prominently in the *TP53* subgroup, a result in line with previous studies [145, 182, 183]. As previously described by Thatikonda et al. not all tumors with germline *TP53* mutation were assigned SBS2 and/or SBS13, underlining a potential link to tissue specificity of these signatures. Looking at panels J, K, L and the respective columns in panel M it was immediately clear that the MMR samples exhibited a different mutational pattern from the *TP53* and *BRCA* CPS samples. This was mainly due to the extraction of MMR related signatures SBS14, SBS15, SBS20, SBS21, SBS23 and SBS44, which contributed as expected most prominently in the MMR subgroup [84, 87]. Due to bleeding effects (explained in detail section 6.2.2 splitting cohort), their contribution was diluted in other subgroups containing MMR related germline mutations although they were still present, most prominently in those subgroups harboring multiple MMR samples such as HGG and pedRTK1 HGG subgroup. On the other hand, *TP53* and *BRCA* CPS were not obviously different from each other in terms of extracted signatures and all differences still identifiable could be due to the different cancers inside each group. SBS31 and SBS35, signatures linked to treatment with platinum based drugs, were assigned to some samples which were treated with such drugs but not all of them without preference to a particular subgroup.

Regarding the ID signatures, I extracted a total of 12 different signatures with SigProfiler. Most prominent in cross-cohort analyses were the clock-like signatures associated with slippage during DNA replication, ID1 and ID2 assigned to 87 % and 62 % of samples respectively. As reported previously, these two ID signatures also contributed most strongly to samples with MMR deficiency in which most mutations/MB were attributed to them. In the MMR subgroup, they were accounting for all present ID mutations. Other prominent ID signatures that contributed across the cohort were ID5, ID8 and ID9 with 23 %, 35 % and 53 % of samples respectively. Particularly ID8, which was linked to repair of DNA via NHEJ mechanisms, was featured in every subgroup. ID3, which was supposedly linked to tobacco smoking, was present mainly in the *TP53* and osteosarcoma subgroup. Although passive smoking could not be ruled out, tobacco use of the patients with ID3 appeared unlikely since these were pediatric patients, which suggested another mechanism producing the mutational signature. Other ID signatures were rarer and appeared not to be linked to any investigated genomic feature.

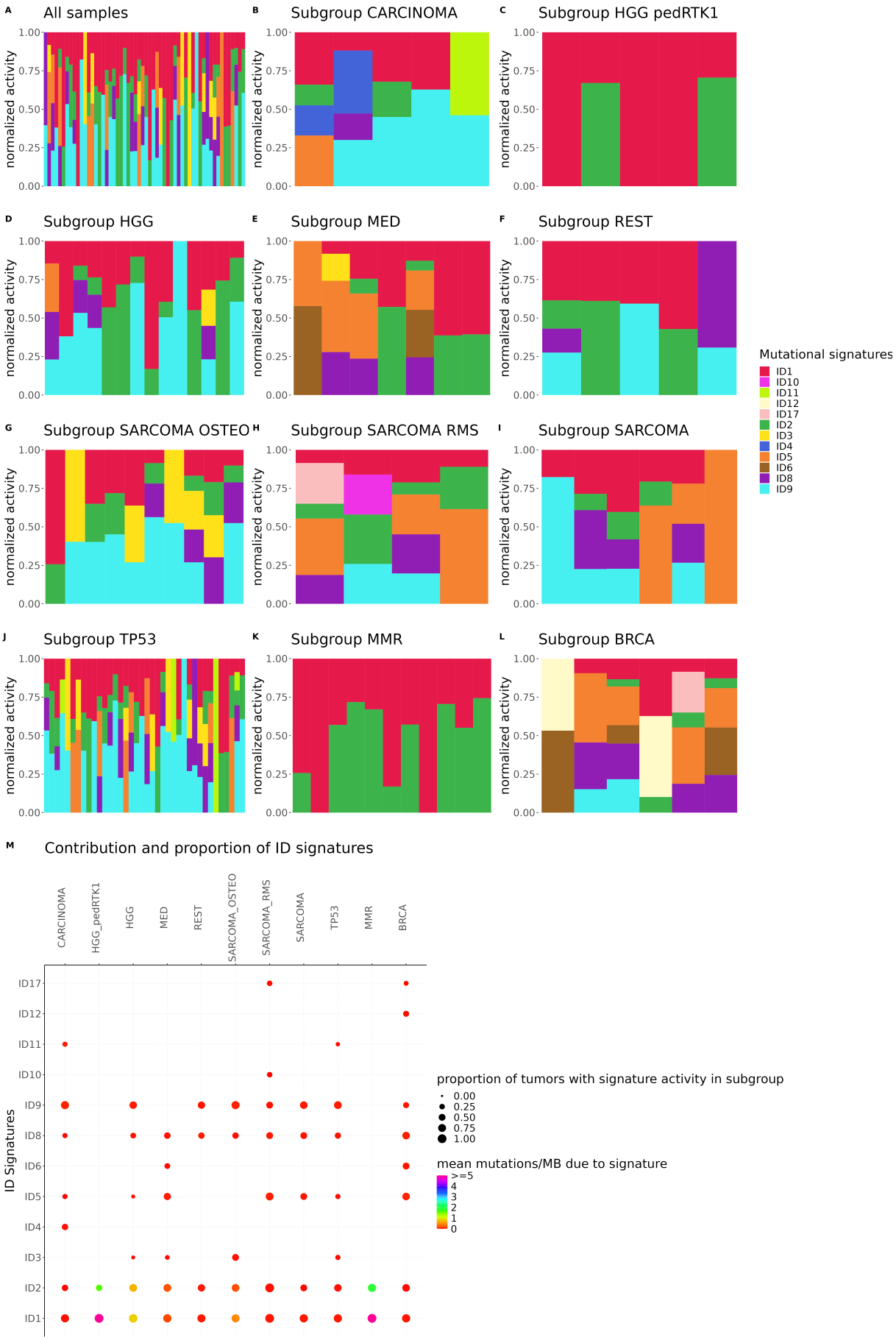


Figure 21: A - L) Mutational ID signatures extracted by SigProfiler for different the different groups investigated. M) Contribution and proportion of each signature to the groups.

In contrasts to the SBS signatures, the MMR subgroups did not exhibit a different profile of ID signatures compared to *TP53* and *BRCA* subgroup. The only difference was in mutational burden, but this was due to the presence of hypermutators in the MMR subgroup.
With this analysis I was able to further confirm previously made associations of signatures with certain features for pediatric cancer. For example the major contributions across samples of clock-like signatures SBS1 and SBS5 as well as ID1 and ID2 and the prevalence of SBS40. The assignment of SBS40 also revealed problems with the analysis via SIGNAL, which appeared to be less sensitive due to differences in the assignment algorithm. The assignment of SBS2 and /or SBS13 to some but not all *TP53* mutated samples further hinted at previous suggestions of influence of tissues for these signatures [145]. The well-known association of SBS14, SBS15, SBS20, SBS21, SBS23 and SBS44 with MMR mutations could be confirmed again and was further analyzed below. Interesting ID signatures that contributed across the cohort were ID5, ID8 and ID9. For ID8 the suggested aetiology is DNA repair while the other two signatures have no known aetiology. Finally the suggested aetiology of ID3 as tobacco smoking appeared more unlikely for pediatric cancer after this analysis instead other suggested links to treatment induced DNA damage or an unknown mechanism appeared more likely, especially considering the correlation analysis below [145].

7.2.3 Correlation between signatures

Next, I investigated the correlations between ID and SBS signatures extracted by SigProfiler. In Figure 22 I show the spearman correlation among SBS and ID signatures sorted by hierarchical clustering. One could observe several clusters of correlated signatures along the diagonal. In the upper left corner, there are two clusters visible. First, SBS10a, SBS10b and SBS10c which all showed correlation with each other. These signatures are linked to mutations in the polymerase epsilon exonuclease domains or defective *POLD1* proofreading and are known to majorly contribute to the high amount of mutations found in hypermutators [84, 87]. The next cluster of signatures that exhibited elevated correlation among each other was made of SBS14, SBS20 and SBS23 all signatures linked to MMR deficiency and as such expected to correlate. Also correlated were SBS15, SBS6 and SBS44, which were also linked to defective mismatch repair. It is noteworthy that these three signatures were separate from the other MMR related signatures, in particular SBS15 showed negative correlation with SBS14, SBS20 and SBS23.

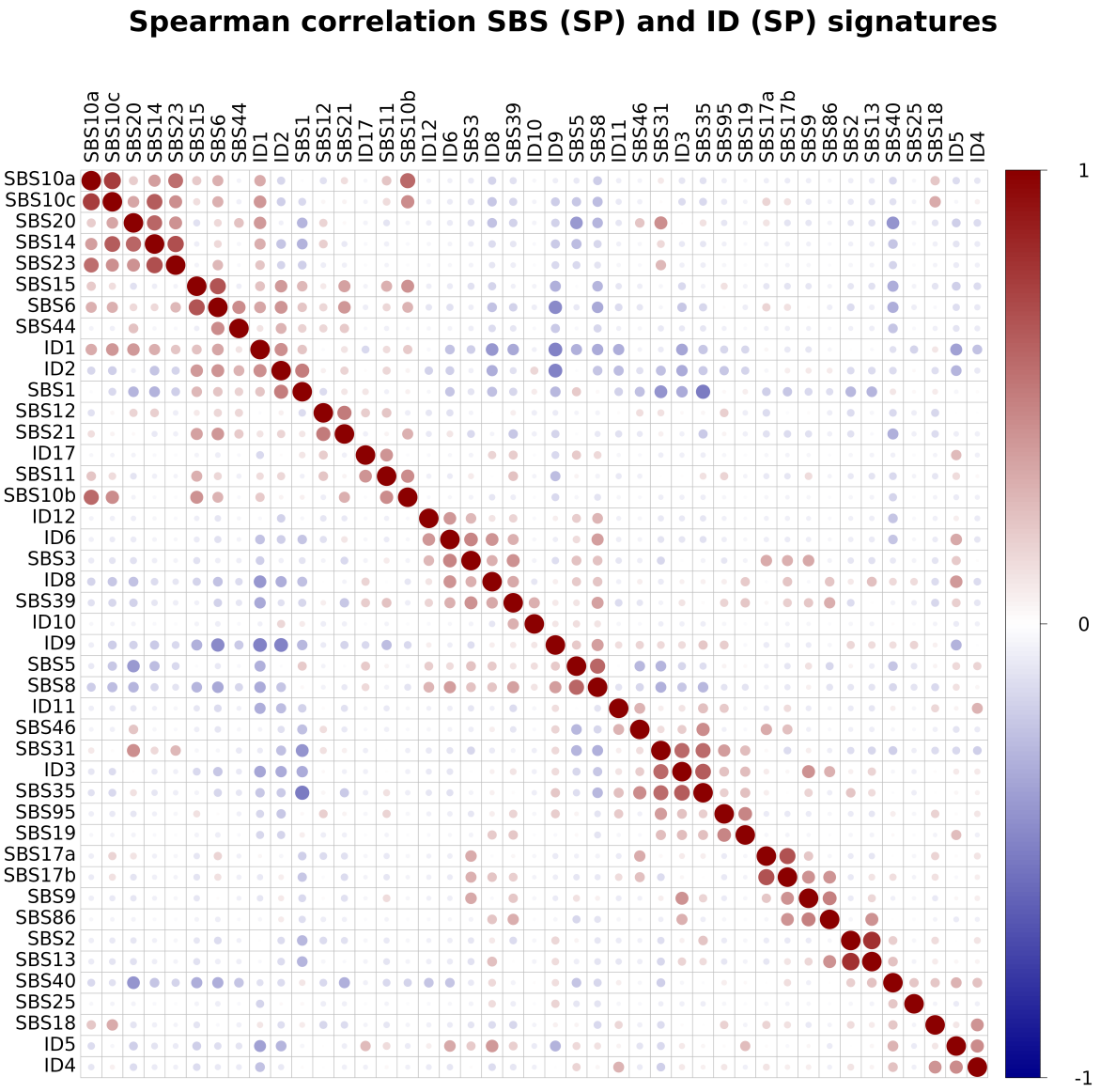


Figure 22: Spearman correlation of SBS and ID signatures extracted with SigProfiler (SP). Columns and rows are rearranged by hierarchical clustering.

The clock-like signatures SBS1, ID1 and ID2 exhibited a positive correlation between them and were clustered together while the other clock-like signature SBS5 was negatively correlated. The next bigger cluster of correlated signatures contained ID6, ID8, SBS3 and SBS39. While SBS3 and ID6 were linked to defective homologous recombination DNA damage repair previously, ID8 was linked to repair by NHEJ or more recently late replication error [184]. SBS39 is currently of unknown aetiology but this correlation suggested a link to DNA damage repair. Another strong cluster of correlation consisted of SBS31, SBS35 and ID3. SBS31 and SBS35 were linked to treatment with platinum based drugs while ID3 was supposedly linked to tobacco smoking. As mentioned above, a link to tobacco smoking was unlikely and suggested another mechanism behind the mutations represented by ID3. Interestingly these 3 signatures were negatively correlated with the clock-like signature SBS1, which might have been due to misassignment since SBS1 and SBS31 have similar patterns in C>T channels. SBS2 and SBS13 exhibited a very strong correlation, which was expected since they were both linked to APOBEC activity.

This analysis showed correlated blocks of signatures. While some correlations were expected like the clock-like signatures others suggested novel associations like SBS39 with DNA repair.

An interesting observation was the split among the MMR related signatures into two groups, which I hypothesized occured due to an influence of the particular MMR gene and further investigated below. The suspected alternative association of ID3 away from tobacco smoking towards treatment induced effects was further supported because of the observed correlation with SBS31 and SBS35, two signatures associated with platinum based drugs.

7.2.4 Correlation with age and differences among MMR mutations

After I evaluated the correlation of signatures among each other, I investigated the correlation of signatures with patient age. In Figure 23 I show all significant correlations with age I could identify in this study. The expected significant correlation with age of the clock-like signatures SBS1 (panel A), SBS5 (panel D and E) as well as ID1 (panel I) and ID2 (panel H) were all confirmed. Next to these I was also able to identify significant correlation for SBS8 (panel B and panel C), SBS40 (panel F) and SBS3 (panel G). Concerning the correlation of SBS3 with age, it has to be mentioned that it was no longer significant after removal of hypermutators, while for all other signatures the correlation remained significant when tested without hypermutators. Another caveat here is that there was only significant correlation of SBS3 with age when using SIGNAL as algorithm, the apparent inferior algorithm for this dataset. This could also be related to the false over-assignment of SBS3 by SIGNAL for technical reasons discussed above.

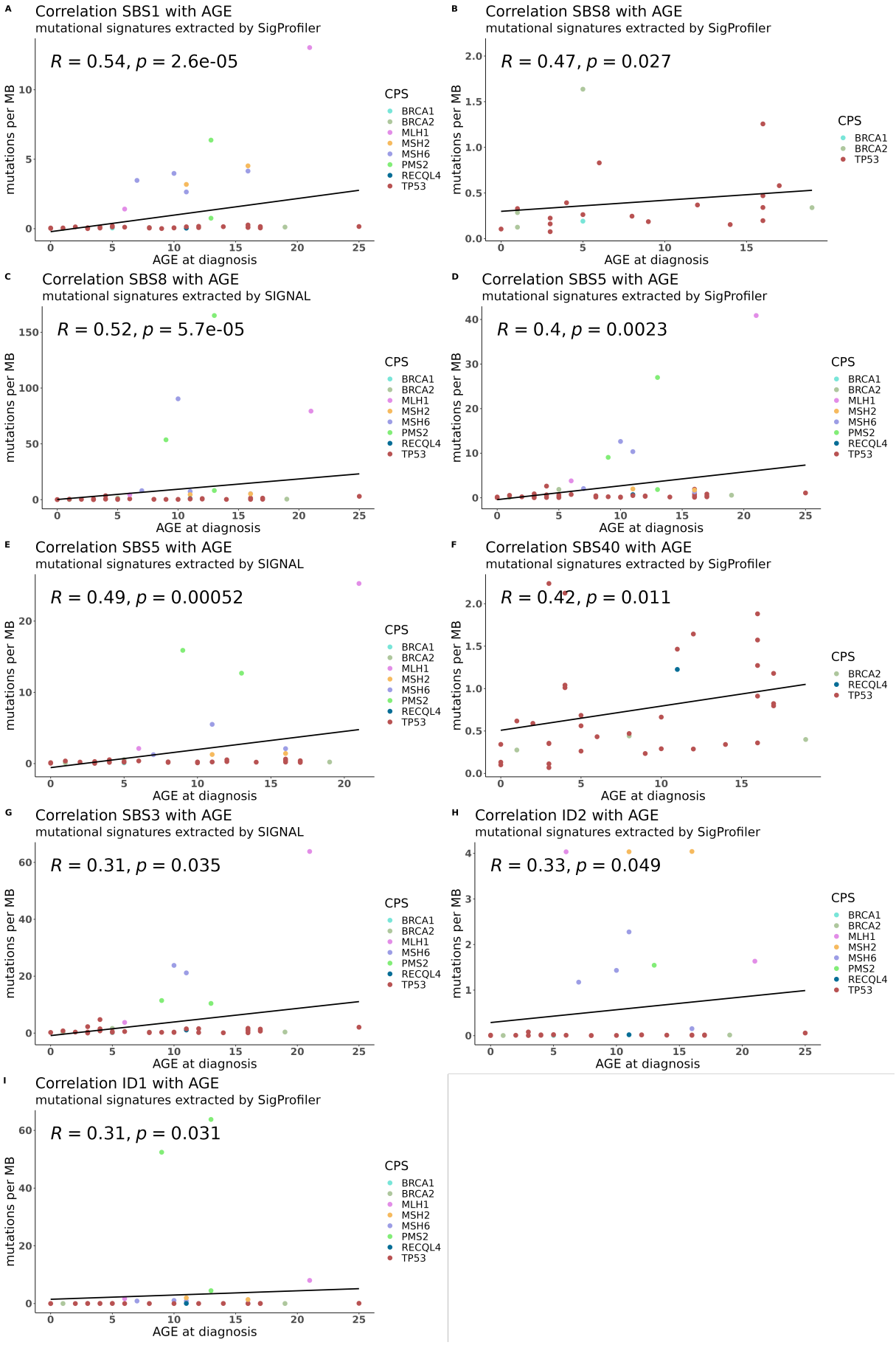


Figure 23: Spearman correlation (correlation coefficient = R) of mutational signatures with age. I also checked correlation without hypermutators (mutations/MB > 10), and it is still significant ($p < 0.05$) for all cases except the correlation of SBS3 extracted by SIGNAL.

SBS8 is reported to be linked with HR and/or NER deficiency or more recently to late replication errors, not with age so the identified correlation may not have a causal connection. However SBS40 has been reported to exhibit clock-like behavior in pediatric cancer before, which I could confirm in the context of my investigation [145].
After I investigated correlation with age I turned towards the MMR subgroup, wanting to learn if the different mutations present (*MSH6,MSH2,PMS2* and *MLH1*) were reflected in the extracted mutational signatures as hinted at by the correlation analysis. In Figure 24 I show the identified differences which were only statistically significant between *PMS2* germline mutated cases and the other (*MSH6, MSH2* and *PMS2*). For some signatures which were associated with age, specifically ID1 (panel A), ID2 (panel B) and SBS1 (panel E), the apparent differences could be influenced by the difference in mean age between the *PMS2* and non*PMS2* group which was 11,7 and 12.2 years respectively although this age differences was not statistically significant.

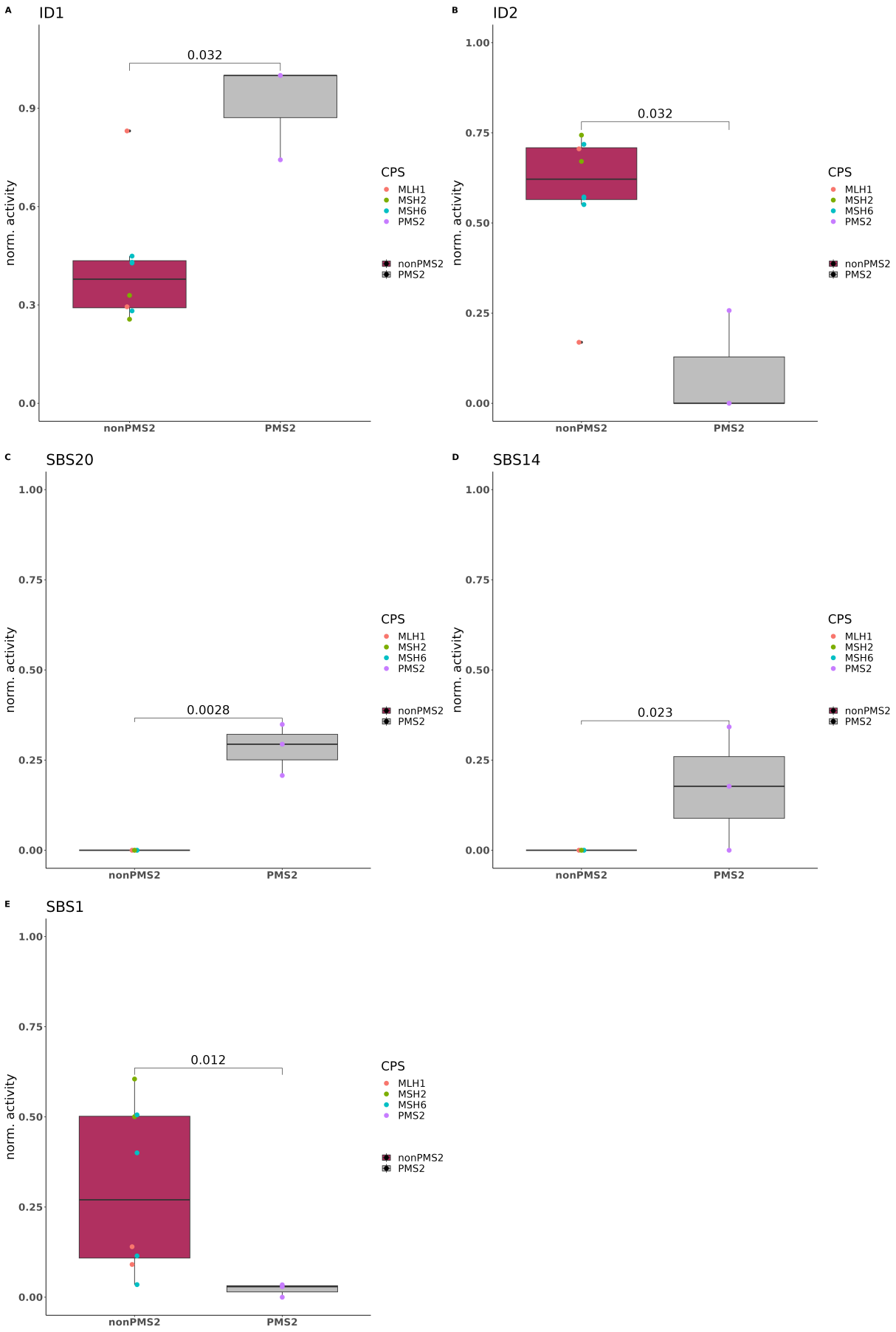


Figure 24: Significant differences between the CPS present in the MMR subgroup reflected in the mutational signatures. Pvalues calculated by Wilcoxon test.

The higher contribution of SBS20 (panel C) and SBS14 (panel D), both signatures mostly active in the C>A channel unlike the other MMR linked signatures, to *PMS2* mutated cases than to other MMR cases might be specific to this particular mutation. Although this assessment was limited by the number of samples inside the MMR subgroup given in the following table.

Table 2: Number of CPS present in the MMR subgroup.

PMS2	**MLH1**	**MSH2**	**MSH6**
3	2	2	4

Further, I investigated within the MMR subgroup if there were differences depending on the heterozygous or homozygous nature of the mutation. The only statistically significant difference I was able to identify in this dataset was the contribution of SBS5, a clock-like signature, although this could be influenced by the different mean patient age in the respective subgroups (graphic in appendix).
With this analysis I was able to confirm again the correlation with age for multiple clock-like signatures like SBS1 and SBS5. But also the previously described clock-like role of SBS40 in pediatric cancer was further confirmed [145]. SBS8, which was also identified as correlated with age, is usually not linked to age. One factor here could be the similarity SBS8 possesses to SBS5, which is a clock-like signature. Of note was also the elevated contribution of SBS14 and SBS20 to MMR cases with underlying *PMS2* mutation compared to the other MMR cases. SBS14 and SBS20 were also identified to correlate among each other as described above. This could be interpreted to be specific to PM2, although due to the low sample size of MMR cases in this study this statement should be treated carefully. On the other hand a difference in mutational signatures between *PMS2* and other MMR related genes has been described previously [185, 186].

7.2.5 Association with POL* mutations and treatment

As next steps, I investigated the association of the extracted mutational signatures with the presence of *POL** mutations (methods 6.2.2), chromothripsis, kataegis and different treatments as far as information was available to me with the help of linear models. I determined the presence of chromothripsis or kataegis manually by considering copy number variation (CNV) plots of the samples which were generated inside the INFORM pipeline. The linear models used the normalized signature exposure as response and the feature under investigation (mutation status, treatment, chromothripsis or kataegis), the CPS, the cancer type and the age as explanatory variables. In the typical R notation, the linear models were given by the following equation, with the leading 0 setting the intercept to 0, making interpretation of the other coefficients more straightforward:

$$SIGNATURE \sim 0 + FEATURE + CPS + CANCERTYPE + AGE$$

In the equation “SIGNATURE” was defined as the normalized exposure of the signature under investigation. “FEATURE” was a one-hot encoded variable giving information about the presence of a characteristic, here the presence of a *POL** mutation each investigated in turn, the presence of chromothripsis or kataegis and whether a treatment was administered. “CPS” was a categorical variable with information about the germline mutation in a sample. “CANCERTYPE” was also a categorical variable encoding the specific type of cancer in a patient. “AGE” was a continuous variable. Since there were categorical variables, some of which had multiple levels, after the fitting of the linear models I applied ANOVA (type II due to the unbalanced nature of the dataset) to estimate the overall effect of the explanatory variables.

Concerning the analysis for association of signatures with *POL** mutations I only considered *POL** genes which had at least 3 cases both in the mutated and wildtype state of the gene in question. For SBS signatures, both extracted by SIGNAL and SigProfiler, and for ID signatures I identified multiple statistically significant associations with *POL** mutations. Interestingly despite the linear model accounting for patient age, there were some significant associations made with *POL** mutation status and mutational signatures usually linked to age. Specifically with mutational signatures extracted by SIGNAL: SBS8, which exhibited correlation with age in this study, showed association with *POLR3D*. Other mutational signatures extracted by SIGNAL, which exhibited association with *POL** mutation were: SBS17b with *POLB* and *POLQ* and SBS17a with *APOL1*, *PAPOLB, POLM, POLR2E, POLR3D* and *POLR3K*. The associations with mutational signatures extracted by SIGNAL should be interpreted with caution due to the apparently less sensitive algorithm and a tendency to over-assign certain signatures as discussed above. There were also numerous associations of mutational signatures extracted with SigProfiler with *POL** mutation status. Among them association of MMR related signatures like SBS14, SBS15, SBS20 and SBS23. With these mutational signatures there appeared to be an alternating pattern of association, where *POL** mutation that were associated with SBS14 and SBS15 were not associated with SBS20 and SBS23 and vice versa. Previously concurrent loss of polymerase proof reading and defective mismatch repair function gave rise to mutational patterns resulting in assignment of SBS14 and SBS20 [187]. Of the clock-like signatures only SBS1, SBS5 and SBS8 were associated with *POL** mutations. While SBS1 was associated with 13 *POL** mutations, SBS5 was associated with *POLR1D* and SBS8 was only associated with *POLR2J4* and *PRIMPOL*. Overall in the associations of signatures extracted by SigProfiler with *POL** mutations, one could see a reflection of the correlation of the mutational signatures. Clusters of signatures which exhibited high correlation as shown above, e.g. SBS10a, SBS10b and SBS10c or SBS9 and SBS86, also tended to be associated with the same *POL** mutations. ID signatures showed fewer significant associations with *POL** mutations. The only statistically significant associations I identified was that of ID4 with *POLRMT*.
Regarding the association of signatures with different treatment there were fewer significant findings than in regard to *POL** status. For SIGNAL, the only significant association identified was SBS17b with the treatment with steroids. Association with steroids was also the only significant finding using the mutational signatures extracted with SigProfiler, which identified SBS9 and SBS86 to be associated. The ID signatures showed more associations with treatment: ID5 and ID10 were associated with peptide antibiotics, ID11 was associated with steroid treatment and ID5 again was associated with treatment with vinca alkaloids. Applying these linear models, I could not identify any association of any mutational signature, regardless of extraction algorithm, with chromothripsis or kataegis in this study.

7.2.6 Differences among germline, somatic and wildtype samples

Next, I turned my attention towards the subgroup of samples with germline *TP53* mutation, aiming to learn whether there was a difference, reflected in the mutational signatures, to samples that carried a somatic *TP53* mutation. For this purpose, I utilized the mutational signatures extracted by Thatikonda et al. to use in the comparison [145]. Since Thatikonda et al. used SigProfiler for extraction of mutational signatures, I only used the signatures I extracted with SigProfiler for these comparisons. In total, I compared 66 samples with somatic *TP53* mutation to 36 samples with germline *TP53* mutation analysed in this study, which were complimented by 17 samples from Thatikonda et al. that also carried germline *TP53* mutation. Overall, 8 signatures showed significant differences between somatic and germline mutated *TP53* cases, shown in Figure 25. Among these differences were signatures usually associated with age (SBS1, SBS5, ID1 and ID2). SBS40, which was associated with age in this study and by Thatikonda et al. was also identified with a significant difference between somatic and germline *TP53* samples.

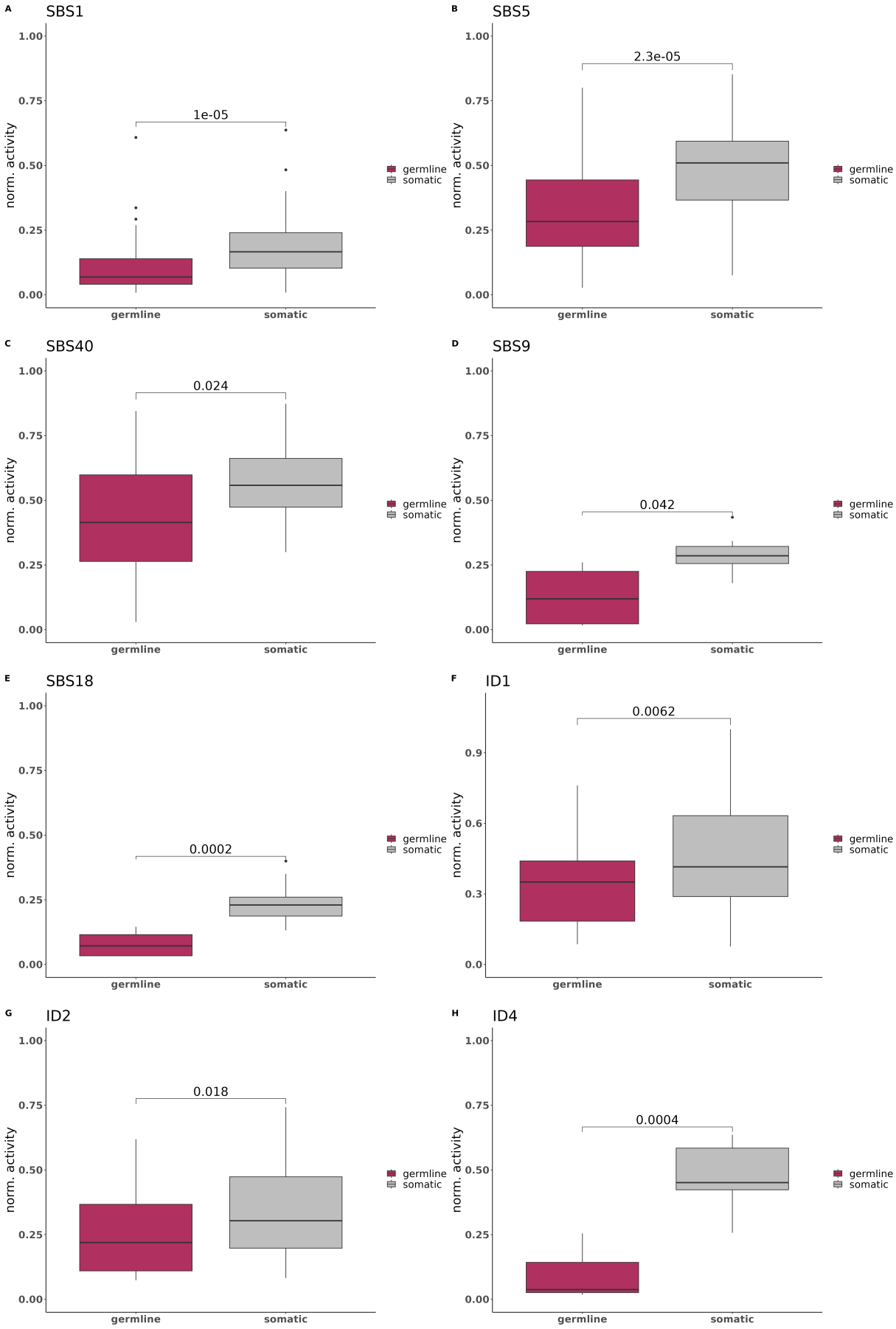


Figure 25: Significant differences between somatic and germline mutated TP53 cases. Pvalues calculated by Wilcoxon test.

The other signatures that exhibited a significant difference between somatic and germline mutated *TP53* samples were SBS9, SBS18 and ID4. ID4 showed a positive correlation with SBS18, while SBS18 and SBS9 exhibited a negative correlation to each other. While ID4 does not have a known aetiology, SBS18 was shown to be related to reactive oxygen species and was previously observed in germline *TP53* samples [188, 189]. However ID4 was recently suggested to be linked to transcription associated process [190]. Specifically it was linked to *TOP1* which in turn was shown to interact with *TP53* [191]. SBS9 on the other hand has not been linked to Li-Fraumeni syndrome before to the best of my knowledge but was primarily observed in leukemia. While there were three leukemia cases investigated, all of them had zero contribution from SBS9. The proposed mutational process underlying SBS9 is currently polymerase eta somatic hypermutation.

Finally, wanting to learn about the association of mutational signatures with the CPS of DADDR patients I employed linear models and used control samples from Thatikonda et al. The linear model is described by the following equation:

$$SIGNATURE \sim 0 + CPS + CANCERTYPE + AGE$$

The response variable "SIGNATURE" was the normalized exposure of the signature under investigation. "CANCERTYPE" was a categorical variable encoding the specific cancer type. "AGE" was a continuous variable with patient age. "CPS" was a categorical variable encoding whether a sample belongs to the *BRCA*, MMR or *TP53* group. The leading 0 was again setting the intercept to 0, making interpretation of the other coefficients more straightforward. Since there were categorical variables, some of which had multiple levels, I applied ANOVA (type II due to the unbalanced nature of the dataset) to estimate the overall effect of the explanatory variables. In total, I compared 659 control samples to 57 samples with germline *TP53, BRCA* or MMR mutation analysed in this study, which additionally included 25 samples from Thatikonda et al. that carried germline mutations in the gene of interest.
Overall, 20 SBS signatures were significantly associated with germline mutation status, while only 6 among them were not simultaneously significantly associated with cancer type (Table 3). For signatures like SBS1, SBS5 and SBS40, which are known to (and did in this study as well) exhibit clock-like behaviour, it was interesting that they were associated with CPS despite accounting for age in the linear model. Another false association with CPS could be SBS35 and SBS31, both signatures linked to treatment with platinum based drugs. Treatment was not accounted for with this linear model since the information was not available for all samples. Signatures linked to possible sequencing artefacts, SBS46 and SBS95 may also be ruled out to be linked with CPS.

Table 3: FDR for association of SBS signatures with CPS and cancer type. Lines in bold are only significantly associated with CPS.

SBS signature	CPS	cancer type
SBS6	3.1e-40	1.2e-26
SBS40	1.3e-30	2.9e-34
SBS15	1.2e-26	4.8e-02
SBS21	1.1e-22	1.5e-02
SBS13	7.8e-22	1.5e-22
SBS2	5.9e-21	1.2e-24
SBS12	**3.4e-12**	**6.7e-02**
SBS19	**5.2e-10**	**2.6e-01**
SBS35	1.1e-09	2.2e-04
SBS11	**2.5e-09**	**7.8e-02**

SBS10a	**1.4e-07**	**2.1e-01**
SBS10b	**4.6e-07**	**2.4e-01**
SBS20	7.1e-07	1.3e-04
SBS5	8.1e-07	6.5e-26
SBS8	2.7e-06	5.6e-03
SBS31	4.0e-04	1.2e-05
SBS95	8.6e-04	6.2e-03
SBS46	**1.8e-03**	**1.9e-01**
SBS1	7.9e-03	9.5e-07
SBS39	4.3e-02	8.7e-07

Other signatures have been known to be associated with certain molecular characteristics in the CPS group. For example, both SBS2 and SBS13 have been known to be associated with APOBEC activity and *TP53* germline mutation. Their simultaneous significant association with cancer type underlines a potential tissue specificity as discussed above already and pointed out previously by Thatikonda et al. SBS6, SBS15, SBS20 and SBS21 are known to be associated with DNA mismatch repair deficiency so their association with the samples in the MMR group was not surprising. A similar picture presented itself when considering the significant association of ID signatures with CPS status, although all ID signatures were simultaneously significantly associated with CPS and cancer type. Just as with SBS signatures, some ID signatures are well known for their association with other characteristics. For example ID1 and ID2 are known to be associated with age, as was also confirmed in this study. Another signature reported to have clock-like behavior is ID5, although I was not able to confirm this in this study. Despite accounting for age in the linear model, these 3 signatures still showed significant association with CPS. ID6 and ID8 are linked to DNA damage repair and it was not surprising that they were linked to the MMR samples. ID3 is linked to tobacco smoking.

Table 4: FDR for association of ID signatures with CPS and cancer type.

ID signature	CPS	cancer type
ID9	1.9e-16	9.1e-12
ID6	2.3e-14	7.9e-20
ID1	3.7e-10	6.1e-15
ID3	3.2e-08	3.7e-05
ID8	3.2e-08	6.9e-17
ID2	1.9e-06	3.7e-05
ID11	5.2e-05	1.5e-02
ID12	6.2e-05	1.7e-14
ID5	1.7e-03	1.8e-02

For the signatures not linked to any known characteristic or without any known aetiology, I considered the coefficients from the linear models to see whether the overall association was skewed by one of the CPS present in this cohort or if they were similar across CPS (Figure 26).

Figure 26: Linear model coefficients quantifying contribution of signatures to the CPS present and WT. Scale shows high or low association.

The majority of coefficients were close to 0, which indicated that the level of activity for a signature was hardly influenced by the CPS status. Some signatures showed the same direction of association with the different CPS and WT only the absolute value was different. Specifically SBS5, SBS40, SBS35, SBS2, SBS13, ID9 and ID1 all showed positive coefficients across CPS status. As discussed above some of these signatures are known to be linked to certain characteristics. SBS5, SBS40 and ID1 are linked to age, while SBS35 is linked to treatment with platinum based drugs. SBS2 and SBS13 are linked to APOBEC activity, but a possible tissue specificity was not accounted for in the linear model. ID9 has an unknown aetiology currently and was most active in the *BRCA* and *TP53* subgroups. While still positively associated with MMR and WT, the contribution of ID9 was notably smaller in those subgroups. Next to those signatures there were also those that showed positive and negative association with certain subgroups. I identified SBS8, SBS1, ID8, ID5 and ID2 that showed this behaviour. SBS8, suspected to be linked to HR or NER deficiency, was negatively associated with MMR and WT, while positively associated with *BRCA* and *TP53*. SBS1 and ID1, known to exhibit clock-like behaviour, were negatively associated with *BRCA* and positively associated with all other subgroups. ID5, also linked to age, and ID8, linked to NHEJ deficiency, were positively associated with *BRCA*, negatively associated with MMR while they hardly played any role for *TP53* or WT. Overall the differences between MMR and the other three subgroups were more pronounced than *BRCA* vs *TP53* or any of the 3 CPS compared to WT. This trend was already visible from the activity and contributions of each signature shown earlier, but was confirmed with this analysis.

Overall this analysis confirmed several known associations of mutational signatures made for pediatric cancer. For example the previously pointed out clock-like nature of SBS40 was confirmed [145]. In the latest release of the COSMIC database (v3.4), SBS40 was split into

SBS40a, SBS40c and SBS40b all of which have unknown aetiology. Another example would be the association of MMR signatures with samples that carried a germline mutation in one of the MMR genes. I reported in this study that there appeared two correlating blocks of MMR signatures appeared and hypothesized that these blocks were the result of the specific underlying germline mutation. I was able to identify SBS14 and SBS20 to contribute significantly more to samples with *PMS2* germline mutation compared to all other MMR cases. Through the correlation analysis, I could also report that SBS39, a signature with unknown aetiology, clustered with several other signatures known to be linked with DNA damage repair suggesting such a link for SBS39 as well. Another observed correlation was that of ID3 with SBS31 and SBS35. The latter two are associated with platinum based drugs treatment while ID3 was linked to tobacco smoking. As discussed above and in agreement with previous reports, this analysis suggested another link to treatment caused DNA damage or something novel. Concerning the differences between germline and somatic *TP53* mutated samples, several signatures were identified with significantly different contributions (that were higher in the somatic samples), among them clock-like signatures as well as SBS9, SBS18 and ID4. Estimations for association of signatures with CPS via linear model revealed several SBS signatures. SBS10a, SBS10b, SBS11, SBS12 and SBS19 all were significantly associated with CPS and have potential to be leveraged for classification purposes based on their contribution. Other signatures significantly associated with CPS were simultaneously associated with tumor type, which may be a hint at tissue specific processes, but could also be leveraged for classification.

7.3 Methylome analysis of DADDR patients

After the analysis of the mutational signature landscape, I turned to the impact of the different DADDR syndromes on the methylome. I hypothesized that there might be an influence of DADDR on the methylome because the in-house methylation based tumor classifier (as initially published by Capper et al.) exhibited bad performance with samples later identified with a DADDR germline mutation [44]. For this purpose, I analyzed 62 samples from patients with germline *TP53* mutation, among them 28 had methylation data available from both the tumor and the blood from the same patient, and 20 MMR patients, among which there were 10 with both tumor and blood methylation available from the same patient (cohort details methods 6.2.1). While I had 6 samples from patients with *BRCA* mutation available, unfortunately for none of them a matching blood control sample was available. For this reason and because of the small number of cases, patients with *BRCA* germline mutation were excluded from further analysis.

7.3.1 Assembling the control cohort

Greatly influencing an investigation like this, were the control samples used for statistical testing. It was important to make sure that the samples used as control did not carry any germline mutation in the genes in question but also to make sure that there was no somatic mutation acquired in the CPS genes as well. To achieve this I utilized the INFORM dataset, which offers in depth molecular characterization that allowed me to select with high confidence a suitable control cohort. I aggregated 811 samples without germline or somatic mutations in the genes of interest (here *MSH6, MSH2, MLH1, PMS2, TP53, BRCA1* or *BRCA2*). The top 10 genes with the most alterations in the control cohort are shown in Figure 27.

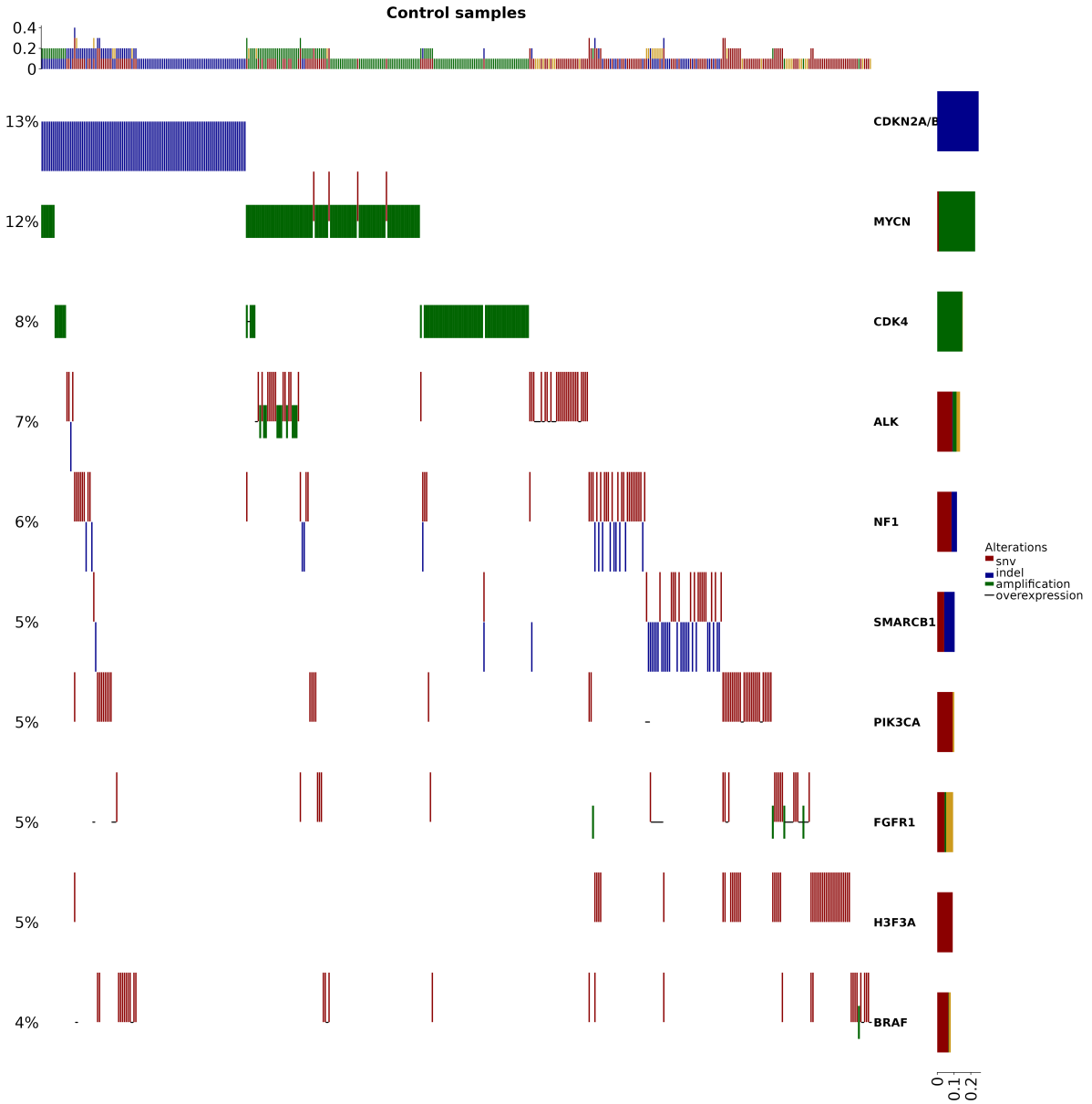


Figure 27: Oncoprint of the control samples (n = 811) used in this analysis. Shown here are only the top 10 genes with the most alterations. In the barplots overexpression is represented by golden color. Barplots at the sides show fractions.

13% of control samples had an alteration in *CDKN2A/B*, next *MYCN* with 12% alterations and then *CDK4* with 8% followed by *ALK* alterations with 7%.
For later evaluation of the results, I also curated a set of samples (n = 364) that carried somatic mutations in the genes of interest but no germline mutation. As expected, the somatic samples (methods 6.2.5) had a different set of most prominent mutations compared to control cases. 83% had a somatic mutation in *TP53*, 25% in *H3F3A* followed by 17% and 13 % in *ATRX* and *CDKN2A/B* respectively (Figure 28). The high prevalence of somatic *TP53* mutations was expected and is already well described for human cancers in literature, although it is usually lower in pediatric patients compared to adults.

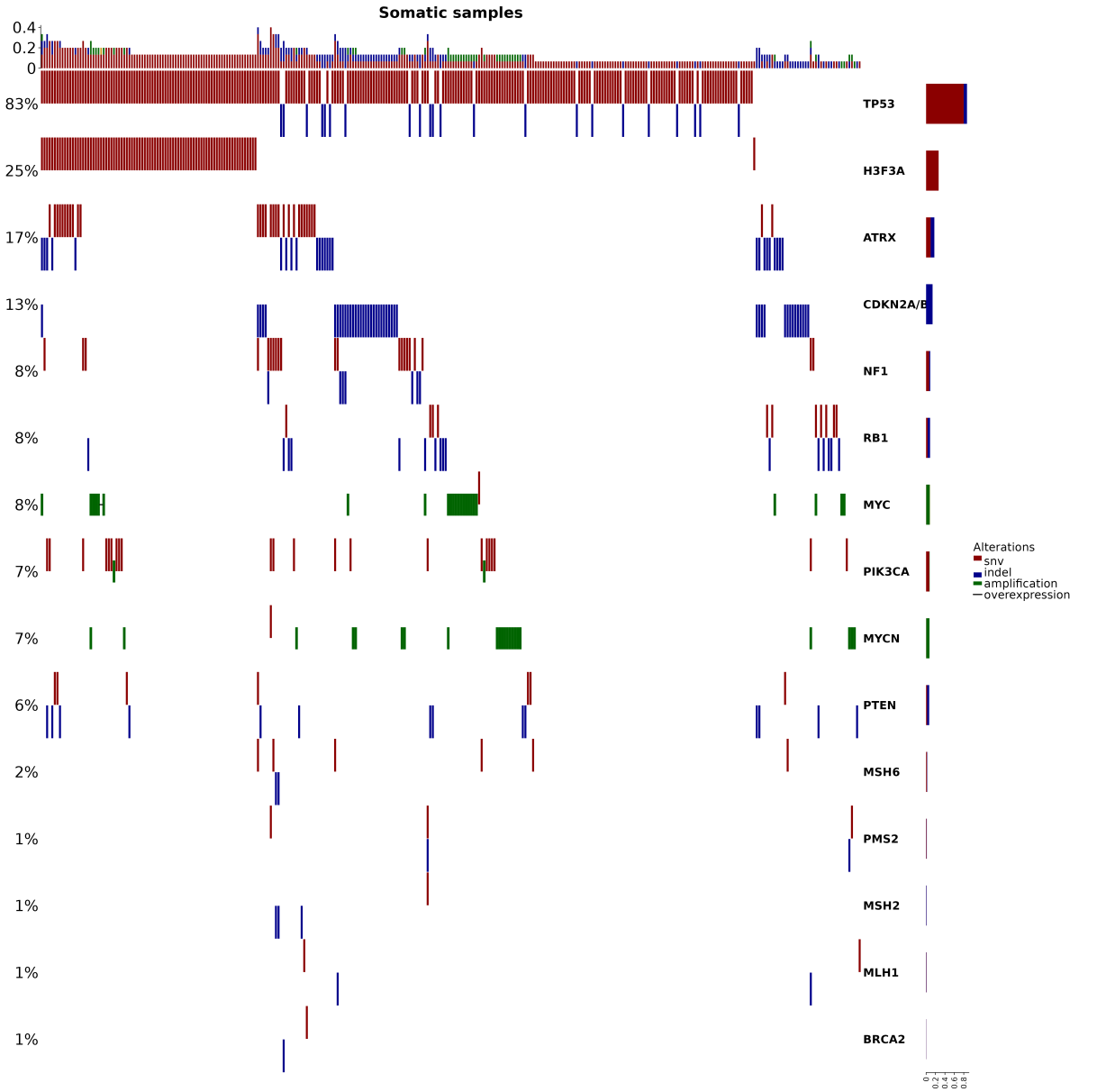
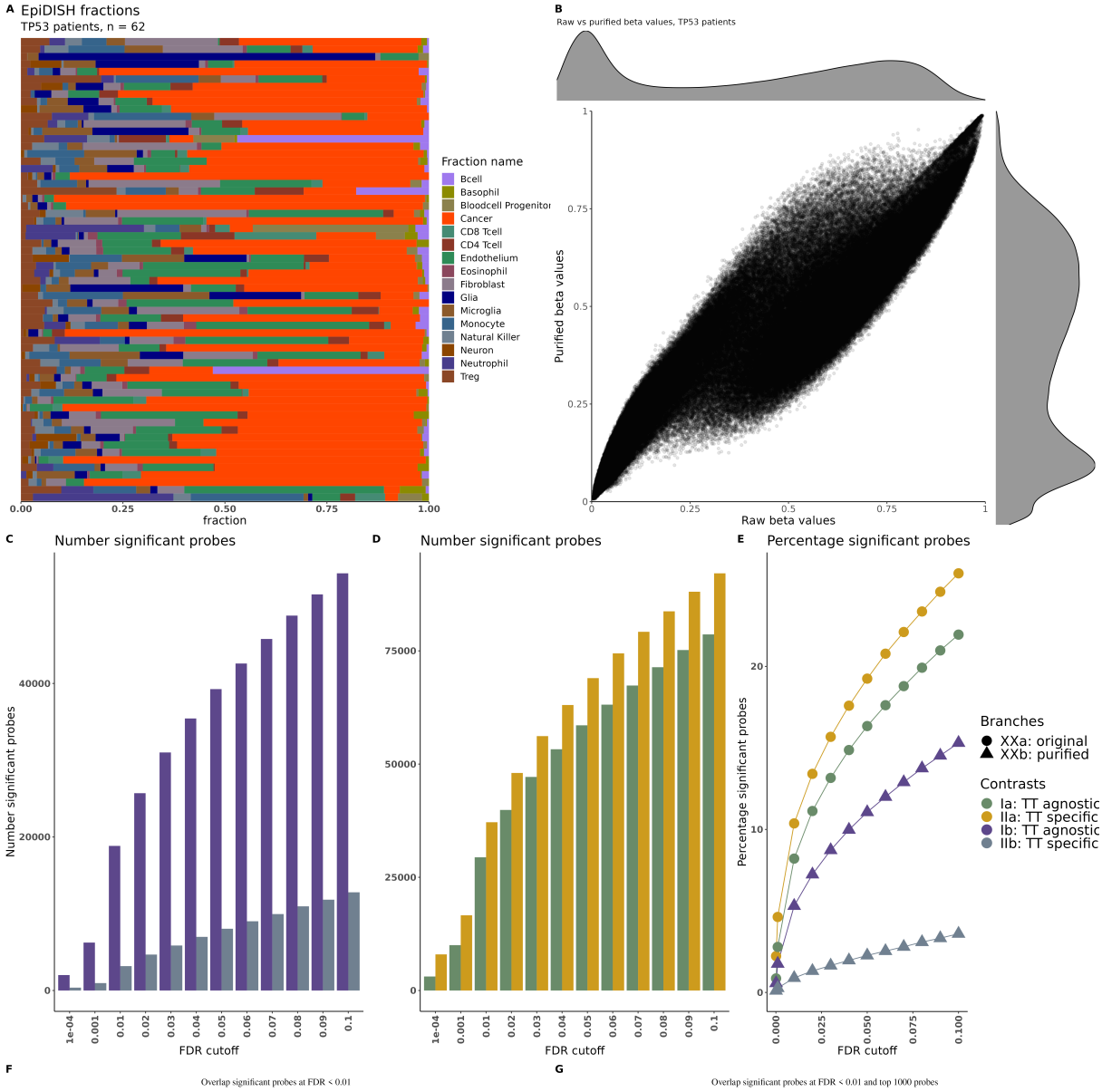


Figure 28: Oncoprint of the somatic samples (n = 364) used in this analysis. Shown here are is the intersection of the genes of interest (MSH6, MSH2, MLH1, PMS2, TP53, BRCA1 and BRCA2) and the top 10 genes with the most alterations. In the barplots overexpression is represented by golden color. Barplots at the sides show fractions.

7.3.2 Methylation landscape overview

Before going into the finer details of the analysis results, I present an overview of the cell type composition, the influence of the purification process and the differentially methylated probes. In Figure 29 I show a summary of the results from the analysis of the germline *TP53* mutated samples. In panel A is the estimated composition of different cell types in each tumor sample, which shows that as expected none of the samples consisted of pure cancer cells based on methylation data and EPIDISH estimation, rather they were a mixture of various brain cells, immune cells and cancer cells. Cancer cell fractions ranged from 0.03 to 0.9 (mean = 0.45 ± 0.25) and one sample showed no cancer cells and subsequently was excluded in the analysis when I used purified beta values, but kept for the analysis with raw values because I could not determine with absolute certainty that there were no tumor cells present. Comparisons of cell type fractions of tumor samples between the LFS germline and control cohort while taking into account tumor type, revealed significant differences for all investigated cell types except cancer cells (Figure 48). Further investigation into the fractions of cell types showed influences from

certain tumor types or from the interaction term “germline:tumor type” on the fractions. For example, ALL and AML types showed a higher fraction of blood cell progenitors and a lower fraction of glia cells. The interaction of germline status with HGG pedRTK1 or ATRT SHH status showed increased association with basophile cells, maybe a hint at an increased immune response. This heterogeneity in non-cancer cell composition and its interaction with germline mutation prompted me to purify the methylation values, based on reference methylation profiles of immune cells derived from pediatric tumors, using the estimated fractions of immune cells [192]. In panel B I show raw methylation beta values compared with the purified methylation beta values. The purification was based on the estimation of cell type fractions shown in panel A and aimed to obtain only the cancer related methylation signal (methods 6.2.4). The histograms on the side of panel B reflect an overall shift in the corrected methylation values when comparing non-purified values to purified values. This is visually observed as the peaks at the fully methylated and un-methylated ends of the purified beta values that increased in size but were shifted away from the maxima and minima and formed wider and slightly less focused peaks. The interpretation here was that the cancer related methylation signal had more emphasize on beta values closer to the middle than before purification. Beta values closer to the middle could mean a mixture of methylated and un-methylated probes which prompted me to include the analysis of variable methylation. Panel C and D show the number of differentially methylated probes identified with contrasts Ia, Ib, IIa and IIb at different FDR cutoff values. Contrast Ib (using purified beta values) returned less differentially enriched probes than its counterpart Ia that used raw methylation values, and the same was true about contrasts IIa and IIb. However, contrast IIa (raw methylation values) returned more values than Ia, the opposite situation could be observed for contrasts Ib (purified methylation values) and IIb. This is clearly visible in panel E where the percentages of differentially enriched probes for each of the 4 contrasts are shown. A lower number of selected methylation probes was expected when purified methylation values were analysed since the purification algorithm was used to remove all non-cancer methylation signal. This removed the influence of immune cell infiltration, which could be highly different between different cancer types or even between different tumors of the same type, or other impurities that resulted from the sampling process. Consequently, less noise was present so less probes were falsely returned as significantly differentially enriched between CPS and non-CPS probes.



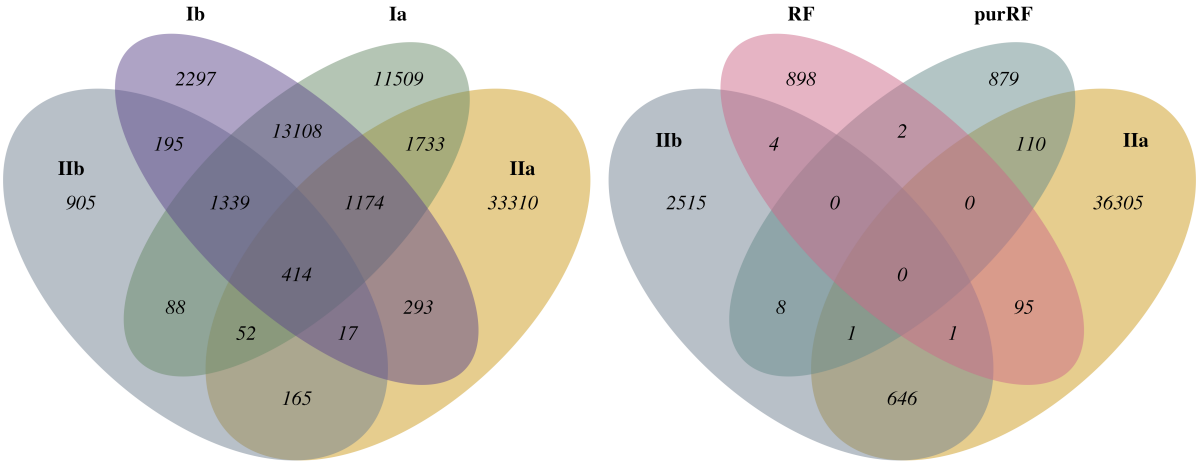


Figure 29: Overview of results from analysis of samples with TP53 germline mutation. A) Estimation of present fractions in of different cell types in the tumor samples. B) Original beta values vs purified beta values based on the estimation of cell type fractions. C-D) Number and percentages of significant differentially methylated probes at different FDR cutoffs for contrasts Ia, Ib, IIa and IIb. F) Overlap of differentially enriched probes by the different contrasts at FDR < 0.01 G) Overlap of

differentially enriched probes from contrasts IIa and IIb at FDR < 0.01 with top 1000 most important probes identified by RF method. TT = tumor type, purRF = random forest method with purified methylation values

Panel F visualizes the overlap of probes identified as differentially enriched between contrasts at FDR < 0.01. For the contrasts Ia and Ib, there was substantial (>50%) overlap in identified probes while for IIa and IIb the overlap regarding significant probes was more limited. The overlap between all contrasts was rather small with only 414 probes compared to the total number of probes in contrasts Ia, Ib and IIa but might be explained by the limited number of probes identified with contrast IIb which only produced 3175 differentially enriched probes. The largest overlap was between contrasts Ia and Ib with ~13000 probes. Looking at contrasts IIa and IIb and the probes identified with random forest method (panel G) a different picture emerged. Most differentially enriched probes were unique to their method of identification and there was no overlap among all 4 methods (contrasts IIa and IIb and random forest on original and purified data) shown in this graphic and only 2 probes (located in *SEMA4C* and *GPN3*) overlapped between both RF methods. The analysis of the MMR samples showed comparable results (graphic in appendix). All the tumor samples were estimated to contain a mixture of cells very similar to the composition identified for the *TP53* samples with cancer cell fraction ranging from 0.06 to 0.89 (mean = 0.55 ± 0.27). Although with the MMR samples there was no sample with 0 contribution from cancer cells. The comparison of the raw and purified beta values revealed a similar behaviour for the MMR samples. The raw beta values had a very similar distribution to the *TP53* samples, the purified beta values showed peaks that were shifted away from the maxima and minima and formed wider peaks, that again hinted at the importance to investigate variable methylation. Another explanation for this behaviour might also be inherent in the purification algorithm and in particular with the prior distribution (a uniform distribution) assumed for the cancer methylation signal. The overall difference in cell type estimation to the *TP53* samples might be explained by the different identified tumor type composition in these cohorts. Regarding the number of probes identified as differentially methylated, the contrasts Ia and Ib returned more probes than IIa and IIb, while Ia and IIa returned more probes than their respective counterparts using purified methylation values. The absolute numbers of probes was comparable to the situation with the *TP53* cases but with slightly more overlap between contrasts. This might be explained again by less tumor types present in the MMR cohort. For the non-type adjusted contrasts Ia and Ib, there were more identified probes shared between them (~24000) than were unique to either (Ia ~10000, Ib ~21000). Contrasts IIa and IIb showed a similar behaviour compared to the *TP53* analysis. The overlap between results by random forest method (for both raw and purified values) and contrasts IIa and IIb remained low.

With this analysis I was able to show that not only was there a very diverse set of tumor microenvironment (TME) present in the samples but there was influence of the tumor type on the composition of the TME and there also was influence from the interaction of germline status with tumor type. Comparison of the germline mutated samples with the control samples revealed that there were significant differences in all considered cell types except cancer cells, and that the germline mutated samples usually exhibited higher levels of non-cancer cell invasion. This prompted me to account for the influence of the different TME by applying a purification method. Purely judged on the amount of identified probes, the purification method appeared to work as intended because it removed noise that was otherwise falsely identified as differentially methylated. Likewise the investigation methods that aimed to remove the influence of tumor type appeared to work, again judged purely on the number of identified probes. Another important observation here was that the RF methods appeared to be more sensitive to the influence of the purification and overall less stable. It is important to note that these observations held true for the MMR samples as well.

7.3.3 Quantifying the differentiation power

The applied linear models, linear contrasts built on top of them and random forest permutation importance calculations aimed at the identification of a methylation signature specific for *TP53* or MMR germline mutation regardless of tumor type and obtained partially overlapping sets of probes with potential clinical application for diagnosis that needed further investigation. Having gained a general overview of the results, next I turned towards quantification of how good a differentiation between germline and non-germline samples, regardless of tumor type, was possible using only the identified differentially methylated probes. For this purpose I used a random forest classifier and the raw beta values. I trained the classifier on a subset of samples and calculated evaluation metrics on the remaining held out samples not used for training. Applying cross validation, I re-calculated the evaluation metrics 3 times and made sure that the split into training and test dataset was stratified to ensure that the percentages of germline and control classes in each remained the same compared to the complete dataset. This was important since the overall dataset was rather unbalanced with few CPS cases compared to non-CPS cases in the control set. First, I set a baseline using all available probes resulting in ROCAUC of 0.8 ± 0.07 and PRAUC of 0.59 ± 0.06 for *TP53* germline cases.

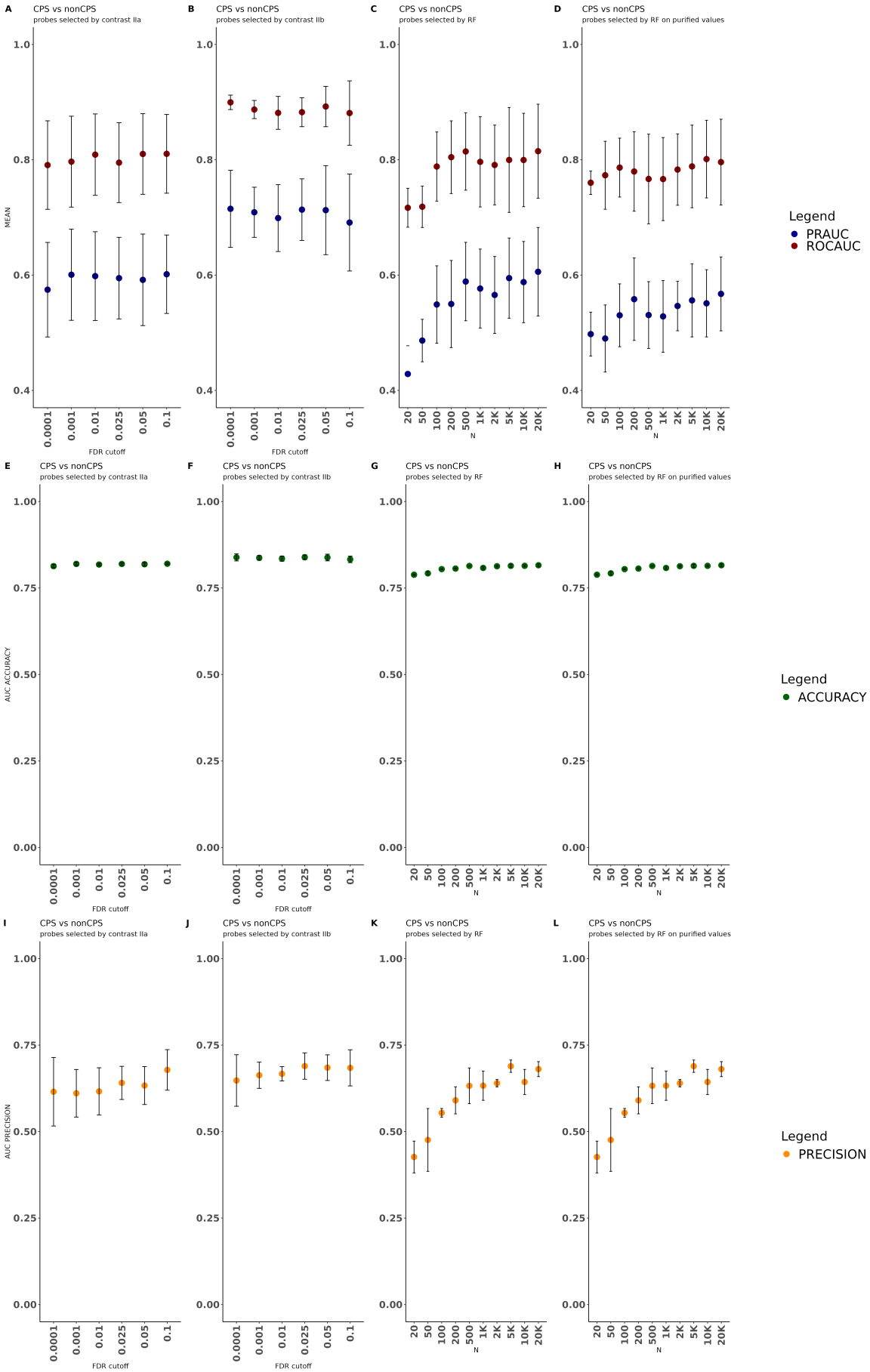


Figure 30: 3x cross validation for differentiation between TP53 vs nonTP53 germline mutated tumors using probes identified as differentially enriched by different methods. A – D shown PRAUC and ROCAUC over different thresholds. E – F show AUC accuracy over different thresholds. I – L show AUC precision over different thresholds. A,E and I: probes identified by contrast

IIa . B,G and K: probes identified b contrast IIb. C,G and L: probes identified by RF on raw methylation values. D,H and L: probes identified by RF on purified methylation values.

Next, I calculated these metrics, as well as accuracy and precision, using only the probes identified by contrasts IIa and IIb at different FDR cutoff values or using the top N probes ranked by the permutation importance calculated with raw and purified beta values (Figure 30). Across both linear contrast methods of selecting probes (panel A and B), the ROCAUC value did not show much movement regardless of FDR cutoff, however an increase was visible in panel B above the baseline of 0.8 close to 0.9 compared to panel A were it stayed just around 0.8. In contrast, when using probes selected by RF permutation importance, the ROCAUC value increased before it reached a plateau around 0.8 after selecting the top 250 probes especially when raw beta values were used (panel C). In regard to the ROCAUC values, contrasts IIb appeared to give the best performance of the 4 methods. However to judge the performance of theses probe selection methods the PRAUC value was more informative since this is a "needle in a haystack" situation. While for contrast IIa the PRAUC value did not change much, only the variance across the different FDR thresholds, contrast IIb showed an increased performance up to PRAUC close to 0.7 above baseline with performance dropping of only at higher FDR thresholds. The maximum of PRAUC = 0.71 ± 0.05 was achieved at FDR cutoff 0.025 for contrast IIb. Turning towards the probes selected by ranking by permutation importance calculated from raw beta values (panel C) and purified beta values (panel D), performance as measured by PRAUC values started below baseline and improved only with the inclusion of more probes but never increased above baseline. Comparing the permutation importance methods to contrasts IIa and IIb, they appeared to give worse results when measured by PRAUC but comparable results when judged by ROCAUC except for contrast IIb that resulted in better performance. To assess accuracy and precision which are two important metrics for clinical classification, I also calculated the AUC of these methods. In this case, AUC refers to the area-under-curve resulting from plotting decision threshold on the x-axis and accuracy or precision at a given decision threshold on the y-axis. The accuracy (panels E - H) hardly changed across FDR levels for contrasts IIa and IIb and only minimally increased with N for the RF methods. Between methods, the linear contrasts performed slightly better than the RF methods as already shown with the PRAUC and ROCAUC metrics. The apparent indifference to changing thresholds compared to PRAUC and ROCAUC could be explained by the imbalanced nature of the dataset where the nonCPS cases far outweigh the CPS cases. The differences in performance between methods and applied thresholds again became visible when considering the AUC precision (panels I - L). Both RF methods again started with a lower precision and increased with N, but never achieved a better performance than the linear contrasts. Between contrasts IIa and IIb, IIb achieved a slightly better performance hovering around 0.7 paired with a smaller error. Looking at the precision curves themselves revealed that with both set of probes, precision above 90% was achievable (data not shown). Overall, the linear contrasts outperformed both RF methods and between the contrasts, IIb offerd slightly better performance than IIa across all metrics investigated here.

Running the same evaluation for the MMR samples, a similar behaviour could be observed (graphic in appendix). The baseline using all available probes resulted in ROCAUC of 0.97 ± 0.02 and PRAUC of 0.92 ± 0.04 for MMR germline cases. Both contrasts IIa and IIb gave good differentiation performance, especially the most selective FDR cutoff values with contrast IIb improved the PRAUC the most, up to 0.99 ± 0.006. The RF methods of probe selection interestingly offered decreased performance not only compared to the linear contrasts methods but also compared to baseline. The ROCAUC values did not offer much insight, hovering just under the theoretical maximum of 1.0 for contrasts IIa and IIb while for the RF methods an improvement with the inclusion of more probes was observable. The PRAUC values started below the baseline for the RF methods and approached baseline with the increase in cutoff. Regarding the AUC accuracy the same trend was visible, with the linear contrasts offering the

best performance, especially contrast IIb. The same was true for the AUC precision, especially the most restrictive cutoff value for contrast IIb gave the best results. In summary, in this investigation the tumor type adjusted linear contrasts offered better differentiation between germline mutated cases and control cases for both the *TP53* and MMR syndromes compared to probes identified by RF methods and improved performance above the baseline. High precision values of 90% or better were achievable, a good starting position for further evaluation and potential clinical application. The better differentiation was observed with a drastically more limited set of probes than the baseline that used all available probes, hinting that the more limited selection was indeed a methylation pattern linked to the respective syndromes. In any case, further investigation into the selected probes and contextualization for biological interpretability was needed for a deeper understanding.

7.3.4 Cancer type specificity

After picking cutoff values for each of the identification methods with consideration to the evaluation metrics, I investigated the qualitative specificity of the identified probes for the respective germline mutation across cancer types. For this purpose, I plotted the tumor samples with germline mutation and the control tumor cases in a UMAP (methods 6.2.5) plot, using only the probes identified by the respective selection methods. In Figure 31, I show UMAP plots where I used the probes identified by contrasts IIa, IIb and by ranking via permutational importance calculated from raw and purified beta values. In general, there were clusters of tumors representing the different tumor types in this analysis, sometimes more distinct (e.g. NBL samples cluster tightly together) and sometimes with more fluid borders, for example the different sarcoma or HGG types. Using the probes returned by contrast IIa (panel A) there was no clear separation of germline mutated cases from their control counterparts without such a mutation. However, there were two agglomerations of germline mutated samples, one in the vicinity of sarcoma control samples the other near HGG control samples. A similar picture emerged when using contrast IIb (panel B). The two agglomerations of germline mutated samples were still visible but in contrast to IIa they were at the fringes and not in the middle, with a smaller third cluster forming. The larger two still mainly consisted of sarcoma types and HGG types, notably the HGG cluster now also included the present SHH medulloblastoma samples. Also worth noting, is that the NBL samples that previously clustered away from all the other samples, now moved towards the sarcoma dominated cluster so that there no longer was such obvious separation. Looking at the probes selected by random forest from raw methylation beta values (panel C) a very similar picture to the selection made by contrast IIa presented itself. However, there no longer was one sarcoma dominated cluster and one HGG dominated cluster, instead only one sarcoma dominated cluster was visible and the HGG cases clustered with their respective control samples. The probes selected by random forest from the purified methylation values (panel D) resulted in a very similar picture to the RF method on raw values. One thing immediately noticeable when comparing the linear model methods with the RF methods was that for the RF methods the individual tumor type clusters were more preserved while the borders were much more fluid with the linear models. Comparing these 4 methods of selecting probes, again hinted at the influence of methylation signal from immune or other non-cancer cells since the linear contrast using purified beta values exhibited better separation. Together with the evaluation metrics discussed above, the contrast IIb appeared to achieve the desired results best followed by contrast IIa.

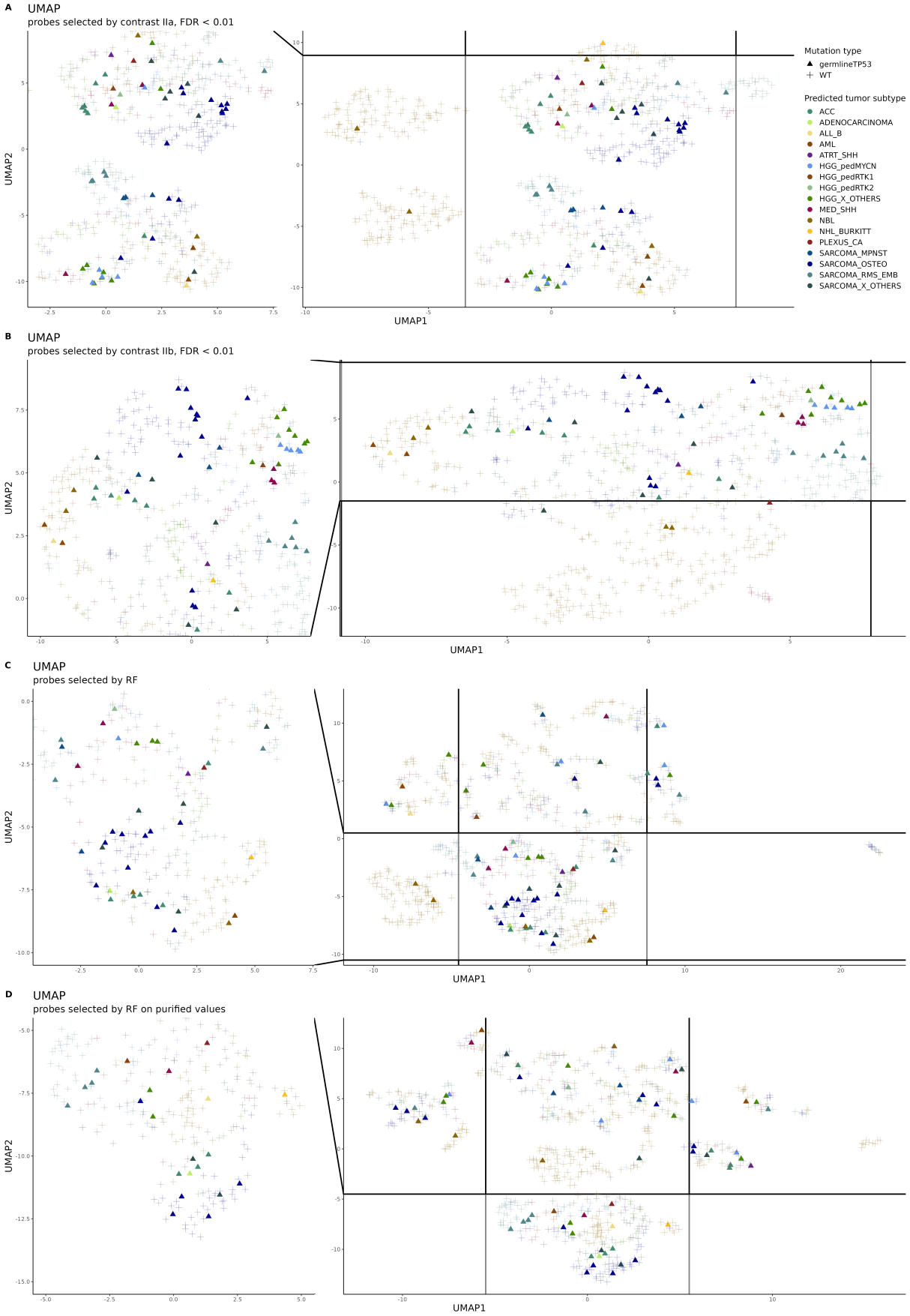


Figure 31: Umap plots using tumor samples with TP53 germline mutation and tumor samples from control cohort. A) UMAP plot using only significantly differentially methylated probes identified by contrast IIa with cutoff FDR < 0.01. B) UMAP plot using only significantly differentially methylated probes identified by contrast IIb with cutoff FDR < 0.01. C) UMAP plot using only top 1000 probes ranked by permutation importance calculated with raw beta values. D) UMAP plot using only top 1000 probes ranked by permutation importance calculated with purified beta values.

Of course, one concern with this analysis was that for each tumor type, there were not enough samples available, especially germline mutated samples, to accurately infer the methylation pattern associated with the syndrome and therefore correct for it. A comparison with contrasts Ia and Ib, which did not take tumor type into account, helped to contextualize the problem. In the figure below the UMAP plots using only probes selected by contrasts Ia (panel A) and Ib (panel B) are shown. If one directly compared contrast Ia to its type adjusted counterpart IIa (Figure 31 panel A) one could see that contrast Ia better preserved the individual type clusters while with contrast IIa borders between types were more fluid.

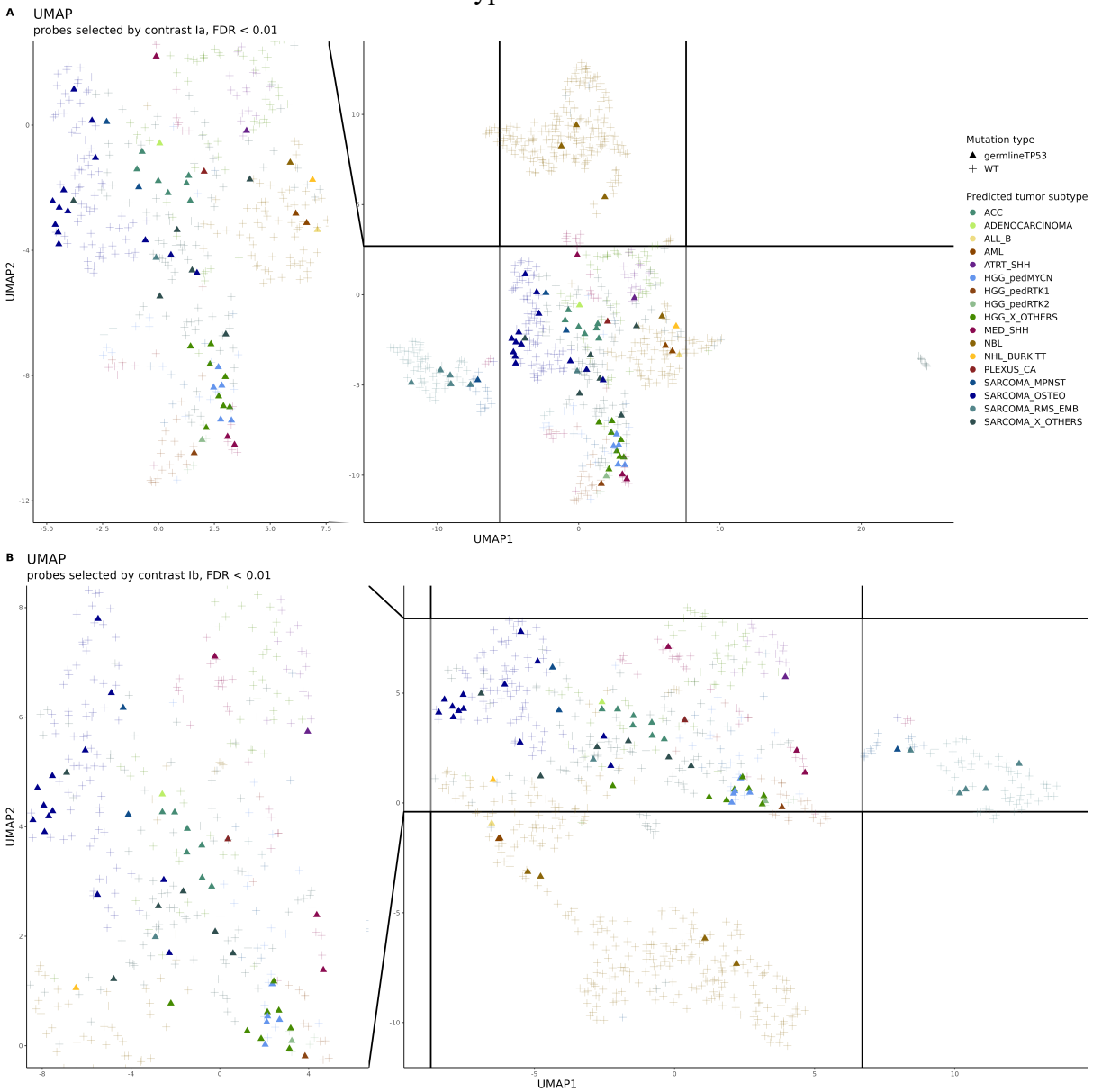


Figure 32: Umap plots using tumor samples with TP53 germline mutation and tumor samples from control cohort. A) UMAP plot using only significantly differentially methylated probes identified by contrast Ia with cutoff FDR < 0.01. B) UMAP plot using only significantly differentially methylated probes identified by contrast Ib with cutoff FDR < 0.01

The same could be observed when comparing contrast Ib (Figure 32b) to its type adjusted counterpart IIb (Figure 31 panel B). The individual tumor types were more clearly organized in clusters using contrast Ib while there were more fluid borders using contrast IIb. For both contrasts Ia and Ib the germline samples did not form distinct clusters away from all other probes, but a tendency to cluster near each other was clearly visible, again with two major agglomerations dominated by sarcoma and HGG tumors visible. Overall, the contrasts taking into account the tumor type appeared to identify CPS specific methylation patterns as desired,

giving insights into the methylation pattern associated with the underlying germline mutation regardless of tumor type. The calculation of permutational importance achieved a worse performance than linear contrasts and occasionally even below baseline performance set by making no selection at all. Further, the RF methods performed worse than linear contrasts with the identification of a tumor type agnostic methylation signature, visible by the still present type clusters (Figure 31 C and D). As in almost all statistical analysis, more germline samples for each of the tumor types in question would lead to higher power of the analysis. Generating the same plots for the MMR cohort (graphics in appendix) revealed the same behaviour as observed with the *TP53* cohort. Using probes identified by contrast IIb resulted in the most distinct cluster of germline cases, an observation in line with the calculated metrics, which hinted again at the influence TME. The other methods of investigation resulted in less clear clusters of germline cases, however the separation was still better than what was achieved in the *TP53* cohort. This was mainly due to the fewer cancer types present in the MMR cohort. As a general trend, the samples with germline mutation investigated here were not drastically different from their control counterparts and a tendency to cluster together if only the identified probes were used was observed. Further there was no batch effect or other technical artefact that led to false results.

7.3.5 Pathway enrichment

After I investigated the discriminatory power and looked at the specificity for germline mutation regardless of tumor type, I further wanted to know about the biological function associated with the identified sets of methylation probes. For this purpose, I analysed the enrichment in selected DNA damage response (DDR) pathways and gene ontology (GO) pathways, taking into account the nature of the CpG probes, specifically that they can map to multiple genes. Starting with contrasts Ia and Ib (Figure 33) I analysed pathway enrichment for both differentially and variably methylated probes and regions. The first thing to note is that no pathways were significantly enriched by the variably methylated probes despite roughly 1400 probes being identified as significantly variably methylated by contrast Ia and roughly 3700 by contrast Ib. This might be a hint that variably methylated probes were not associated with a meaningful biological pattern captured by the pathways used here or that the cutoff values were not restrictive enough so too much noise was still included. Likewise, the contrasts applied to identify significant differences in mean methylation did not lead to enriched pathways at FDR < 0.05. Contrast Ia returned roughly 29000 significant probes at FDR < 0.01 and Ib returned roughly 18000 at FDR < 0.01. Again, this might be a sign that too much noise was still included. I investigated the differentially and variably methylated regions (DMRs and VMRs) within these contrasts and contrasts Ia and Ib (panel A and B respectively) both produced significantly enriched pathways. For both contrasts, the VMRs lead to enriched pathways in all three GO categories (molecular function, cellular composition and biological process), even the number of affected genes in these pathways was very similar with 771 for VMR Ia and 410 for VMR Ib. However only DMRs from the purified contrast Ib lead to enriched pathways, again across all three GO categories, and with significantly more genes affected inside these pathways (11458 affected genes). This difference in the set of enriched pathways when looking at regions instead of individual probes was the result of different cutoff values used to decide which probes are taken into further consideration for identification of regions.

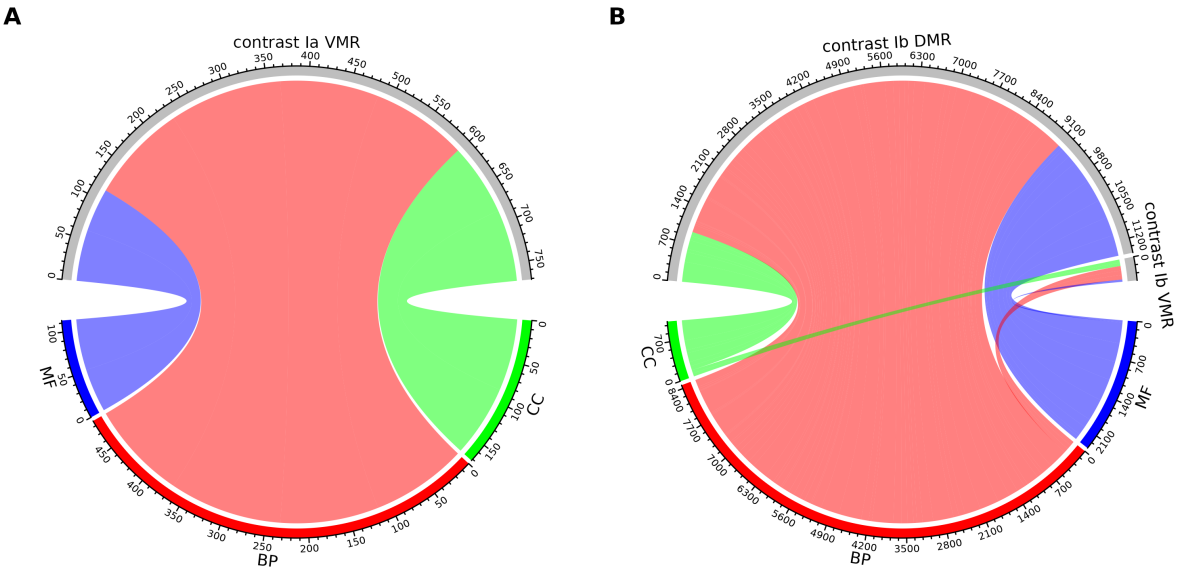


Figure 33: Enrichment of GO paths (FDR < 0.05) for probes and regions identified by contrasts Ia and Ib. The numbers on the axes show the amount of differentially methylated genes in the pathways. CC = cellular component pathway, BP = biological process pathway, MF = molecular function pathway

Looking at the enriched pathways directly, not only at the amount and composition, showed a trend towards certain functions (Figure 34). The enriched GO terms for VMR identified by contrast Ia (panel A) were related to RNAi effector complex, RISC complex, post-transcriptional gene silencing and translation repressor activity among others. The VMRs identified by contrast Ib (panel C) were a subset of the pathways identified by VMR Ia. The enriched GO terms from the DMRs identified by contrast Ib (panel B) also included the same pathways related to RNAi interference, RISC complex and transcriptional regulation. This convergence on similar biological functions with varying degrees of noise was reassuring, indicating there was a common biological function being picked up by the contrasts. On top of that, effects on pathways related to transcriptional regulation were not unexpected since *TP53* has been previously shown to effect polymerase activity and transcriptional regulation [193]. Beyond pathways related to transcriptional regulation, the DMRs Ib also enriched pathways related to morphogenesis and embryonic skeletal system development pathways. Two of the more known affected genes previously linked to cancer inside the skeletal development pathway were *FGFR2* and *FOXC2* which was linked to metastasis [194, 195].

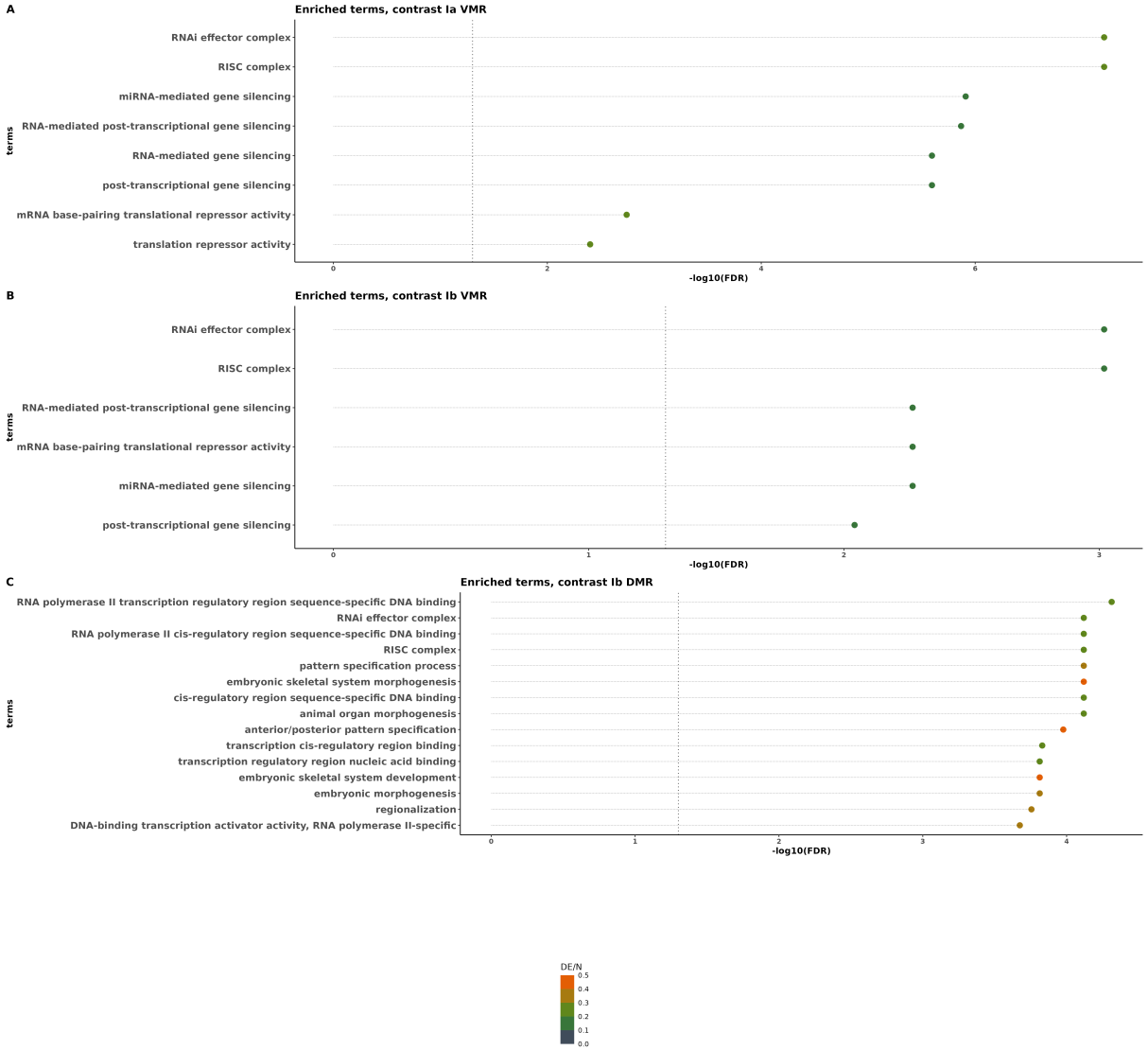


Figure 34: Top significantly enriched GO terms (FDR < 0.05) sorted by FDR. Only top 15 shown in case there were more. Color codes the ratio of differentially methylated genes (DE) in a term to the number of total genes (N) in that term. Dotted line indicated FDR = 0.05

The shared set of pathways between the three enrichment analysis for VMR Ia, VMR Ib and DMR Ib were pathways related to transcriptional regulation, miRNA and RISC complex. Effects of *TP53* mutations on posttranscriptional gene expression regulation have been identified in previous studies and the important role of certain miRNA coding genes have been highlighted [117, 196]. These pathways in particular could indicate that germline *TP53* mutation has an influence in the RNAi pathway steps involving the RISC complex.

Running the enrichment analysis for contrasts IIa, IIb and the RF methods of probe identification revealed that the different contrasts and methods caused distinct behaviour between them in the enrichment analysis (Figure 35). The first thing to point out is that no significantly enriched pathways were identified for contrast IIa variable methylation, none for contrast IIa VMRs, none for contrast IIb differentially methylated as well as none for either RF method. For contrast IIa differential methylation, significantly enriched pathways exclusively from the molecular function category were identified (panel A). For contrast IIa DMR (panel C) significantly enriched pathways from molecular function, cellular composition, biological

processes and DDR were found, although molecular function and cellular composition clearly dominated. For both IIb differential methylation, DMR IIb and VMR IIb only pathways from biological processes were identified (panel B and D).

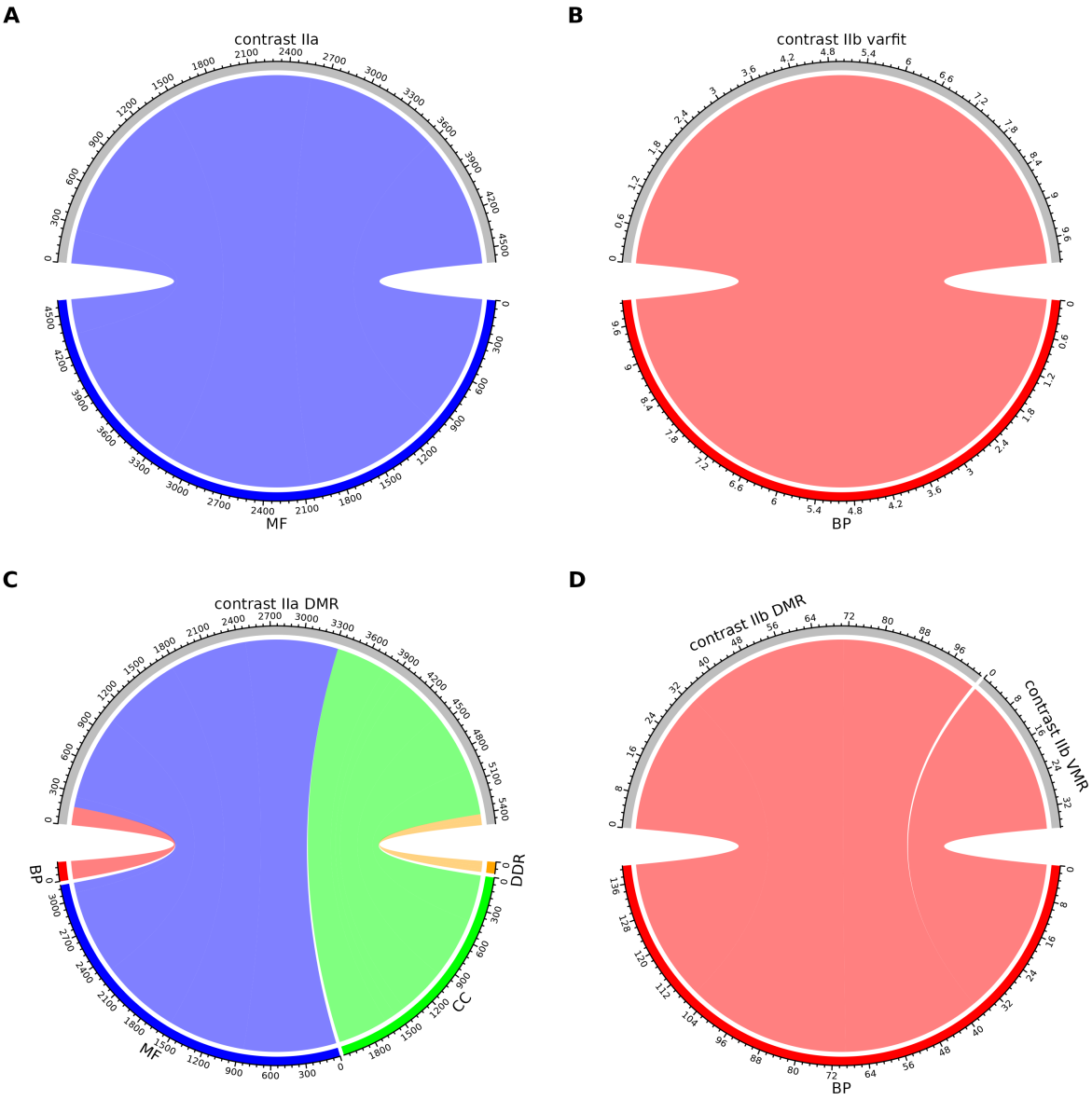


Figure 35: Enrichment of GO paths for different contrasts (FDR < 0.05). The numbers on the axes show the amount of differentially methylated genes in the pathways. CC = cellular component pathway, BP = biological process pathway, MF = molecular function pathway, DDR = DNA damage response pathway

Looking at the enriched pathways directly (Figure 36) revealed a slightly different focus compared to contrasts Ia and Ib. For contrast IIa, the focus was mainly on pathways involved in the energy metabolism of the cell e.g. the ATP binding pathway. The pathways enriched by IIa DMR also included energy metabolism related pathways but also DDR specific pathways were affected, specifically CPF, NER and BER. Contrasts IIb DMR and IIb VMR enriched pathways related to cell-cell adhesion and pathways with functions in plasma membrane cell adhesion.

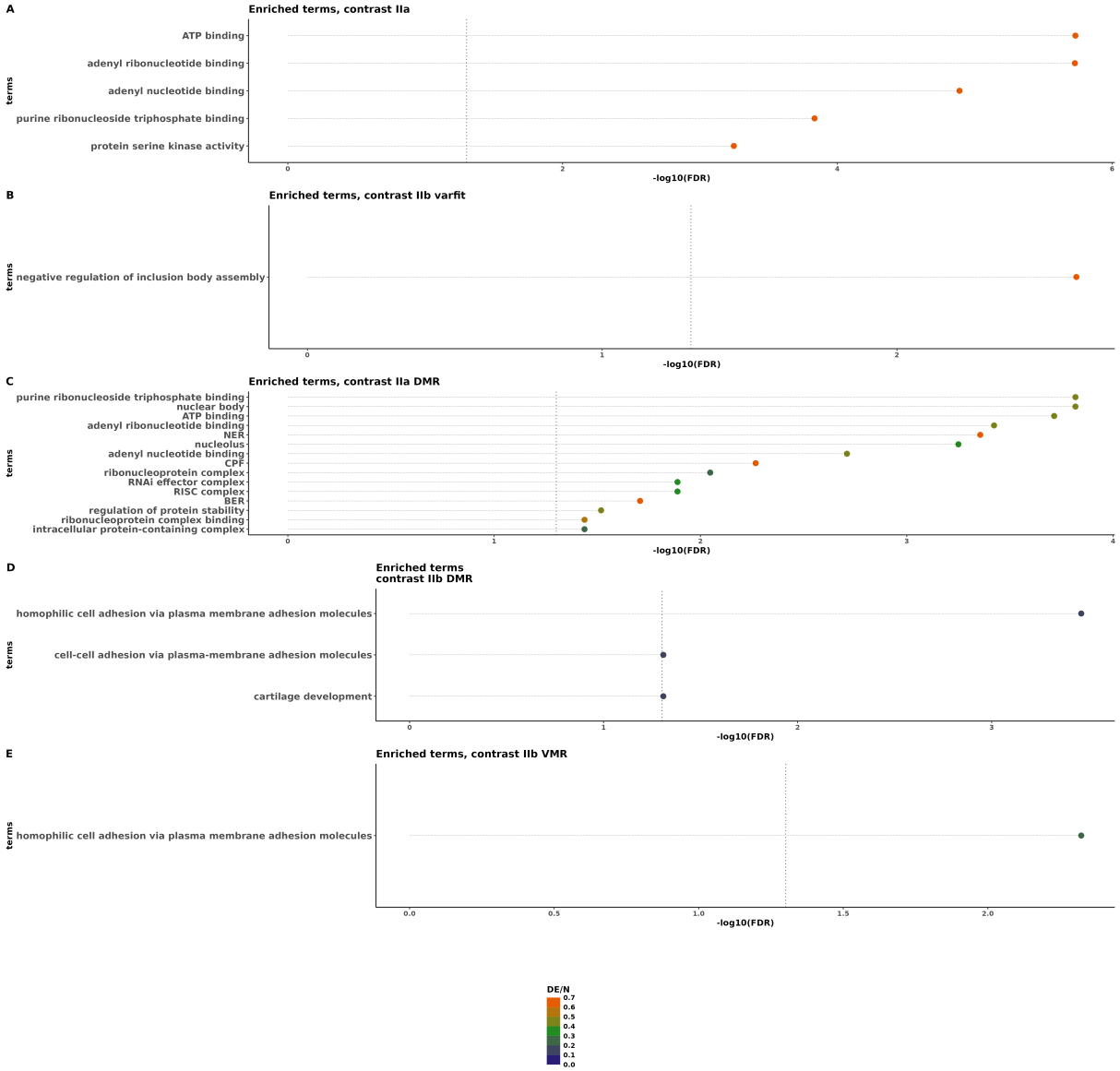


Figure 36: Top significantly enriched GO terms (FDR < 0.05) sorted by FDR. Only top 15 shown in case there were more. Color codes the ratio of differentially methylated genes (DE) in a term to the number of total genes (N) in a term.

Probes identified by contrast IIb variable methylation enriched one pathway involved in the regulation of inclusion body formation. Inclusion bodies are aggregations of faulty proteins and known to be correlated to inactive or mutated p53 protein [197].
While this enrichment analysis yielded some insights, especially for contrasts Ia, Ib and IIa DMR, the other methods hardly led to enriched pathways. Just from the number of identified probes and regions and the success of these probes in the classification shown above a different result was expected. The most interesting identified pathways revolved around RNAi effects, RISC complex and DNA damage response and are further discussed in chapter 8.

7.3.6 Network analysis and enrichment

As alluded to above, despite the often large number of probes identified as significantly differentially or variably methylated across all contrasts, the number of significantly enriched pathways identified with them was lower than expected. I hypothesised that this was mostly due to noise carried over from the differential methylation analysis that subsequently disturbed the enrichment analysis. To obtain a clearer picture of the processes associated with the identified probes and to gain more undisturbed insights into the biology behind the identified

probes from each contrast I applied network analysis methods followed by additional enrichment analysis. The rationale behind this was to obtain clusters of probes that correlated regarding their methylation pattern. These clusters should be larger and more pure in function with less noise, ideally leading to a more insightful enrichment analysis. As a first step, I extracted the top 10% of probes with the highest variance in methylation signal. These were combined with the probes identified by each method which resulted in three datasets: top 10% of probes plus probes identified by contrasts Ia and Ib (networkd dataset 1, ND1), top 10% of probes plus probes identified by contrasts IIa and IIb (ND2) and top 10% of probes plus probes identified by RF method on raw and purified methylation values (ND3). Next, I processed these three datasets with WGCNA and GRAPH method to identify clusters of correlated probes. For each of these identified clusters of correlated probes I ran pathway enrichment analysis, correlation analysis with traits of interest and tested for significant contribution of probes identified via contrasts or RF to the clusters. Traits of interest in this context meant gender (male or female), disease status (primary, progression or relapse), sarcoma (SA) or non-sarcoma (NSA) and PCA cluster (Cluster1 or Cluster2). PCA clusters refers to the two cluster detected via PCA coupled with KNN clustering ran on the same germline tumor data used for the UMAP plots above (PCA not shown), which I hypothesized were a proxy for sarcoma and non-sarcoma status. The results are shown in the following 3 graphics. First, Figure 37 shows the analysis of top 10% of probes plus probes identified by contrasts Ia and Ib. Overall WGCNA detected more clusters than GRAPH, which was true for the analysis of all three datasets mentioned above (ND1, ND2 and ND3), indicating that WGCNA was able to achieve a finer resolution. From the clusters identified by WGCNA, 15 showed significant enrichment with probes identified by contrasts Ia and Ib while only two clusters identified by GRAPH were significantly enriched. Interestingly, not every cluster that was significantly enriched showed significant correlation with a trait of interest and vice versa, an observation that held true for the network analysis of IIa and IIb as well as RF methods. For example, clusters MElightgreen and MEsalmon were significantly enriched but showed only weak correlation let alone significant correlation with any trait. In contrast, MEgreen was not significantly enriched but showed strong and significant correlation with the sarcoma (SA) and nonsarcoma (NSA) trait and their associated clusters found in PCA (Cluster1 and Cluster2).

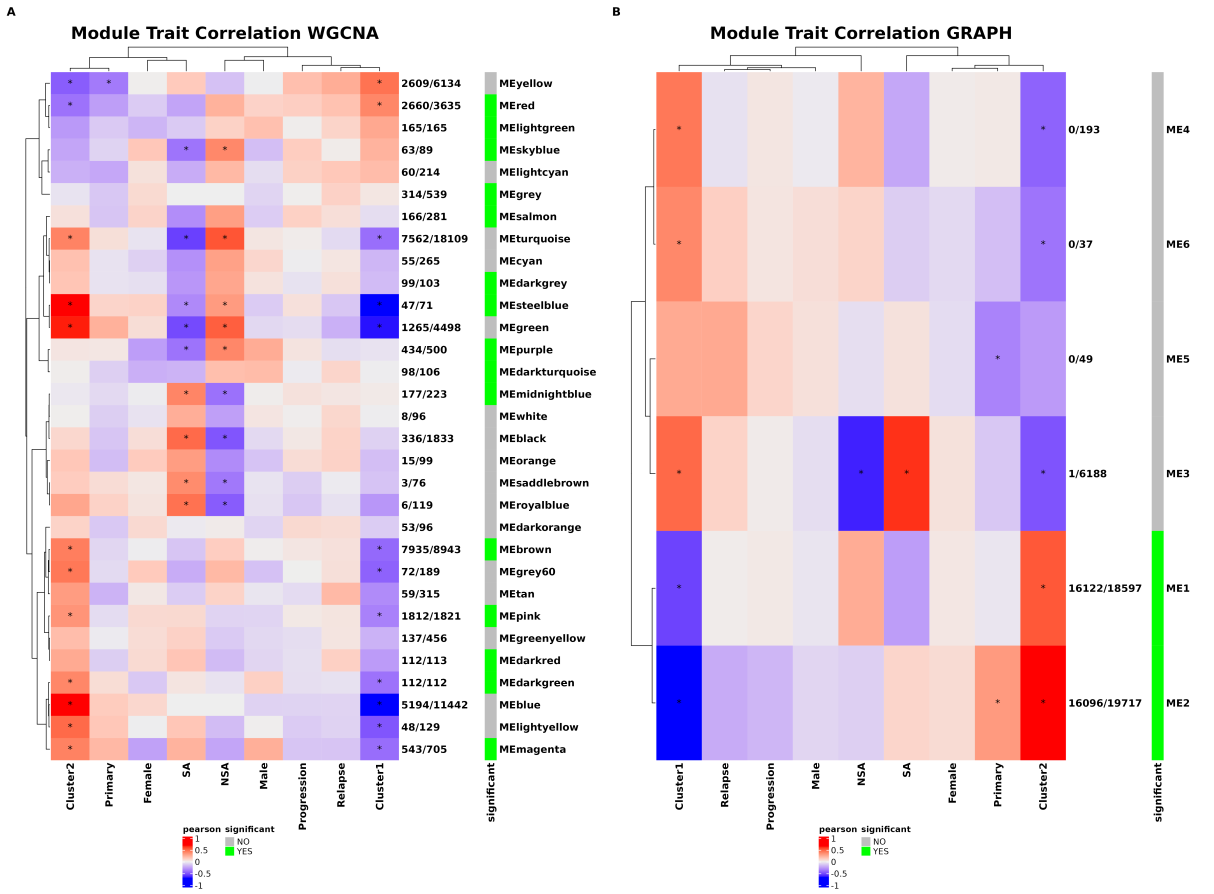


*Figure 37: Results from network analysis with WGCNA (A) and GRAPH (B) for probes identified with contrasts Ia and Ib pooled with top 10% most variably methylated probes. The heatmap shows pearson correlation coefficients with each column a trait of interest. * inside a cell indicates significant correlation with FDR < 0.05. Row annotation "significant" indicates if an identified module is significantly enriched with probes identified by contrast Ia or Ib. Row annotation numbers show how many probes are in a given module and how many of those were identified with contrast Ia or Ib.*

There were also those clusters that were significantly enriched with probes and also exhibited significant correlation with a trait of interest, for example ME1 (Figure 37b). The general trend for both WGCNA and GRAPH analysis was that most significant correlations were related to the SA and NSA traits and their associated PCA clusters. This observation was in line with previous observations made during tumor type classification based on methylation, which resulted in a dedicated sarcoma classifier next to a brain classifier (www.molecularneuropathology.org). Focusing on the clusters significantly enriched with probes identified by Ia and Ib that resulted in significantly enriched pathways (FDR < 0.05), MEbrown, MEdarkred and MEmidnightblue, the identified pathways were associated with different processes. MEbrown, which was significantly correlated with traits Cluster1 and Cluster2, enriched pathways associated with ATP and nucleotide binding as well as catalytic activity acting on nucleic acid. MEdarkred enriched pathways with functions in nucleotide metabolic processes. MEmidnightblue, which was significantly correlated with the SA and NSA trait, enriched pathways related to embryonic skeletal development, morphogenesis as well as polymerase transcription regulation already known from analysis on contrasts Ia and Ib without network analysis shown above. The clusters not significantly enriched with probes were associated with various function like transcriptional regulation and chromatin organisation (MEblack, MEblue, MEgrey60), cytoskeletal function (MEcyan), nucleosome organization (MEorange), embryonic skeletal development and morphogenesis (MEsaddlebrown) as well as dendritic tree functions (MEyellow).

An association with the SA and NSA traits and their associated PCA clusters was also visible from the analysis of contrasts IIa and IIb, underlining previous findings. In Figure 38, there is one cluster identified by WGCNA (panel A) which exhibited a correlation with the progression trait and was significantly enriched with probes identified by IIa and IIb, MEplum1, which unfortunately did not significantly enrich any pathways. In contrast, the clusters identified with GRAPH only exhibited correlation with SA or NSA and their respective PCA clusters. In total 31 clusters identified by WGCNA and 2 by GRAPH showed significant enrichment with probes identified by contrasts IIa and IIb.

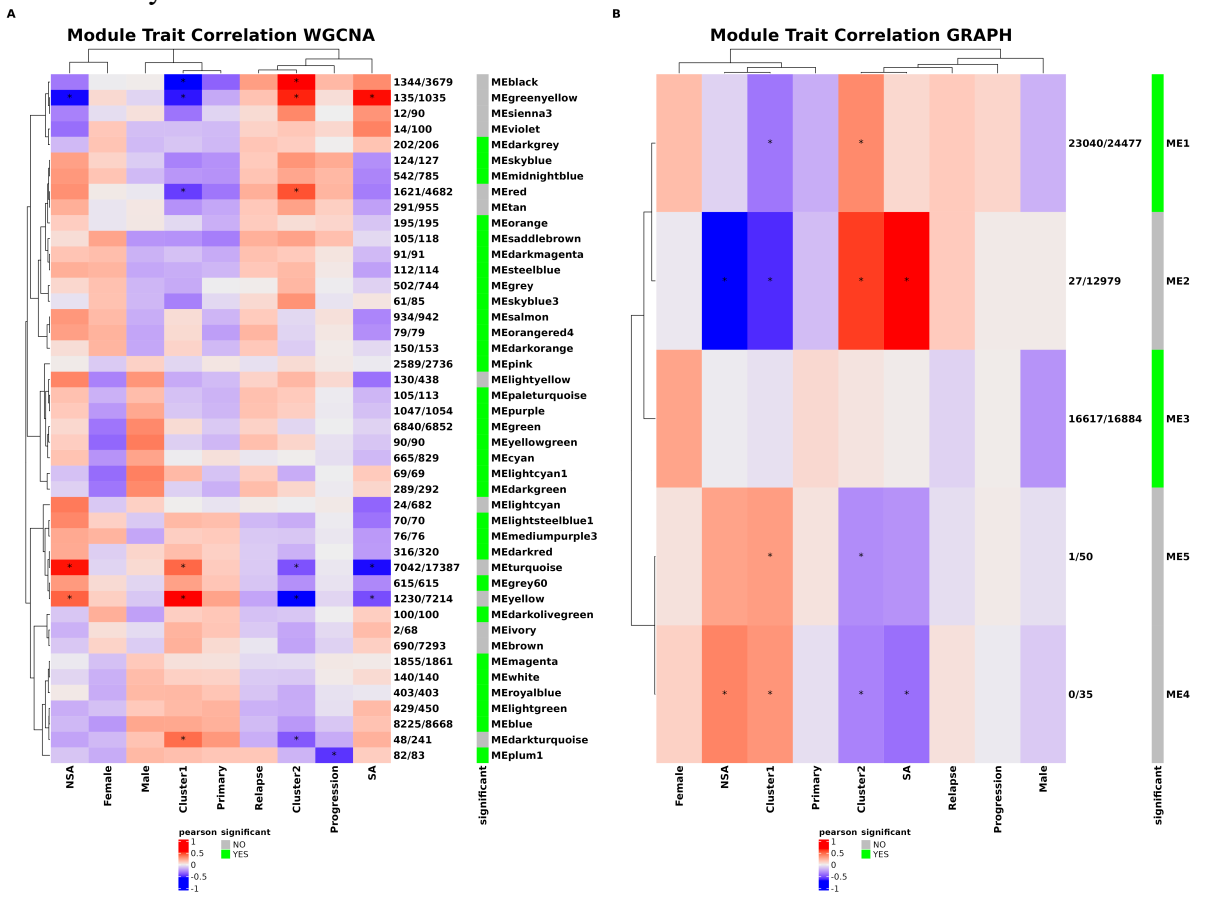


*Figure 38: Results from network analysis with WGCNA (A) and GRAPH (B) for probes identified with contrasts IIa and IIb pooled with top 10% most variably methylated probes. The heatmap shows pearson correlation coefficients with each column a trait of interest. * inside a cell indicates significant correlation with FDR < 0.05. Row annotation "significant" indicates if an identified module is significantly enriched with probes identified by contrast IIa or IIb. Row annotation numbers show how many probes are in a given module and how many of those were identified with contrast IIa or IIb.*

Focusing on the clusters from WGCNA which were significantly enriched by probes and lead to enriched pathways, MEgreen, MElightgreen, MEpink and MEsalmon, again each cluster was associated with different functions. MEgreen was associated with functions related to chromosomal organization, cellular response to DNA damage, mitotic cell cycle and mRNA processes. MElightgreen was associated with extracellular matrix structure while MEpink was associated with protein modification. Finally, MEsalmon was also associated with cellular response to DNA damage, regulation of cell cycle and DNA repair. The clusters which lead to pathway enrichment that were not enriched in probes identified by IIa and IIb were associated with function in transcriptional regulation and chromatin organization (MEbrown, MEgreenyellow), embryonic skeletal development and morphogenesis (MEgreenyellow), DNA packaging and nucleosome organization (MEviolet) or immune response (MEyellow).

Finally turning towards the network and enrichment analysis of probes identified by RF methods (Figure 39), WGCNA (panel A) identified 4 clusters significantly enriched with probes and GRAPH (panel B) identified only one cluster.

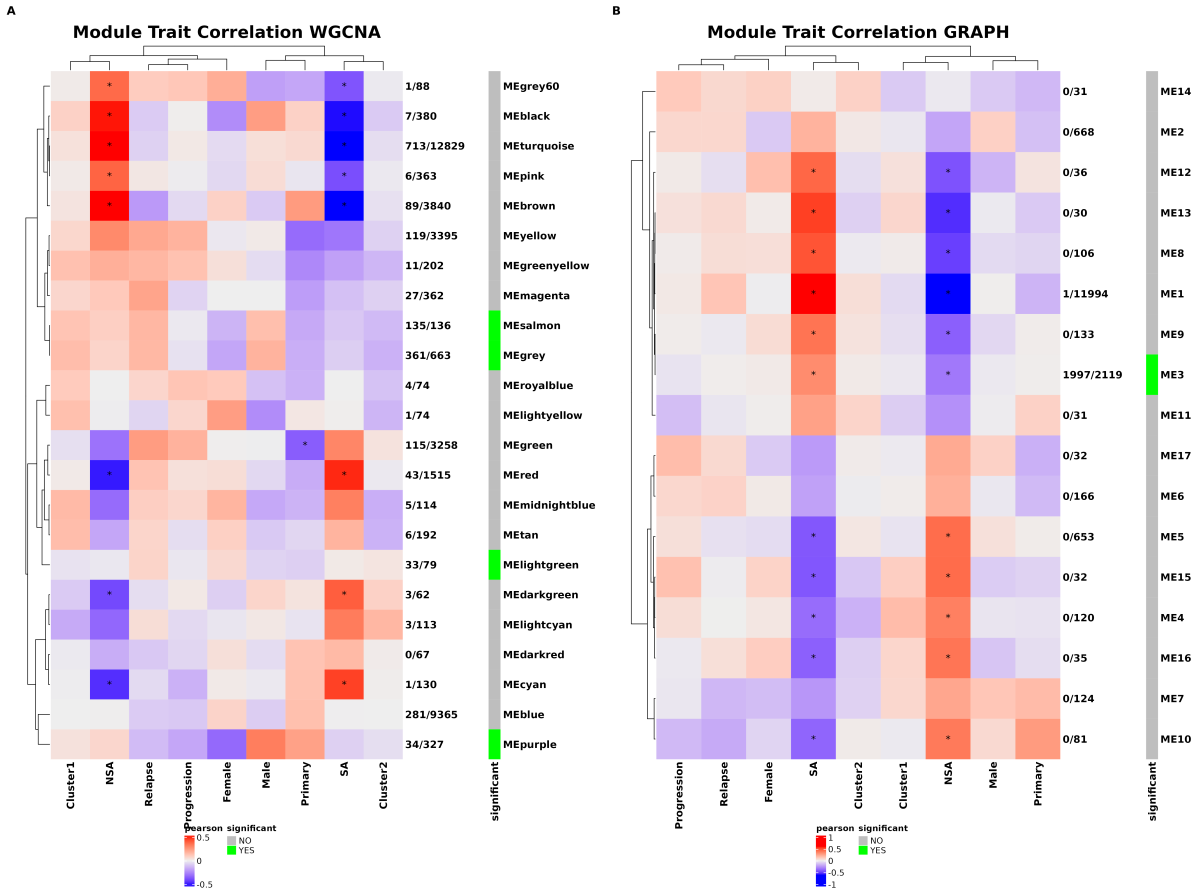


*Figure 39: Results from network analysis with WGCNA (A) and GRAPH (B) for probes identified with contrasts RF on original and purified methylation values pooled with top 10% most variably methylated probes. The heatmap shows pearson correlation coefficients with each column a trait of interest. * inside a cell indicates significant correlation with FDR < 0.05. Row annotation "significant" indicates if an identified module is significantly enriched with probes identified by RF method Row annotation numbers show how many probes are in a given module and how many of those were identified with RF method.*

Of the clusters enriched with probes, none lead to significant enrichment in pathways. However multiple other clusters did produce enriched pathways already known from network analysis of the linear contrasts. Some pathways were associated with transcriptional regulation and chromatin organization (MEblue, MEred, ME4, ME5), embryonic skeletal development and morphogenesis (ME4), RNAi and RISC complex (ME13) or DNA packaging and telomere organization (MEmidnightblue). Interestingly one cluster, MEgreen, was significantly correlated with the primary trait, but unfortunately no significantly enriched pathways were identified for this cluster. Further details on the enriched pathways are given in the supplements. While looking at all the clusters and their associated enriched pathways gave a good oversight, I further focused on clusters identified by WGCNA used on ND1 and ND2 because the linear contrasts outperformed the RF methods as shown above and because they showed more agreement among each other regarding identified probes than the RF methods. In particular I was interested in clusters that were enriched with identified probes from their respective contrasts. From ND1 that were the clusters MEbrown, MEmidnightblue and MEdarkred, all other enriched clusters unfortunately did not lead to enriched pathways. From each cluster I selected the top 10 enriched pathways (sorted by FDR) and inside each pathway I selected the top 10 affected genes (sorted by FDR of the matching methylation probes) and used this data for a network plot (Figure 50). Cluster MEmidnightblue revolved around embryonic skeletal

development, RNA polymerase II regulation or more general cell differentiation functions. Of the affected genes inside each pathway *HOXA3* was ranked first for all pathways. Other HOX genes were also affected as well as *RUNX3*, a development regulator gene, involved in cell cycle regulation among other things that was previously associated with metastasis and *MYC* interaction [198-200]. For network cluster MEdarkred, pathways related to the turnover of polyphosphates were enriched. Specifically the two paralogs *NUDT4* and *NUDT4B* were affected, that were proposed to be involved in signal transduction [201, 202]. Network cluster MEbrown revolved around nucleotide binding, ATP binding and catalytic activity. *KSR1*, a kinase involved in downstream signalling of RAS and positive regulation of MAPK cascade, was ranked first for three pathways [203, 204]. Another gene present in three pathways was *EIF4A3*, a helicase which is involved in RNA processing via the spliceosome and has been linked to glioblastoma growth in adults [205, 206]. These three clusters did not share any enriched pathway nor gene. One reason for this could be that MEbrown and MEmidnightblue were significantly associated with the NSA/SA trait or the Cluster1/Cluster2 trait, which could mean they were be more linked to tumor type than germline TP53 mutation. The same network plot for WGCNA analysis of ND2 revealed more shared pathways and genes among MEpink, MEsalmon, MElightgreen and MEgreen. Cluster MEpink lead to two enriched pathways with function in protein modification by protein conjugation. The two genes ranked first and second for both pathways were *UNKL*, which was suspected to be involved in Rac signalling, and *AMBRA1* which regulates autophagy, is involved in cell cycle control and acts as tumor suppressor [207-212]. Another interesting gene in this cluster was *HDAC4*, which codes a histone deactetylase, with HDACs recently being investigated as drug targets in pediatric brain cancer [176]. Network cluster MElightgreen enriched three pathways involved in extracullular matrix components. Most prominently in all pathways were genes *TNXB* and *LAMA2*, which were ranked first and second respectively, involved in signalling and organization of cells during embryonic development. MEsalmon mainly enriched pathways with function in cell cycle control, response to DNA damage and DNA repair. Genes ranked first place for multiple pathways were *RPTOR* and *EP400*. *RPTOR* is a vital component of the mTOR pathway that regulates cell growth and that has been linked to emergence of cancer [213-215]. *EP400* codes for a member of an acetyltransferase complex and an important paralog is *SMARCA4*, which was also ranked in the top 10 for each pathways where *EP400* was ranked first [216]. Further functions of *EP400* include cell cycle regulation and DNA repair [217-219]. Another pair of paralogs that appeared in the top 10 affected genes were *EHMT2* and *EHMT1*. These two histone methyltransferase coding genes play a critical role in methylation status of histone H3 and interact with HDACs [220]. Another identified differentially methylated gene, directly linked to the p53 pathway and other cell cycle regulation functions, was *E4F1* [221]. Lastly, MEgreen also enriched pathways with function in cell cycle control and response to DNA damage but also pathways involved in chromosomal organization and mRNA metabolic processes. Some genes already described in the context of other network clusters also occured here for example *RPTOR, HDAC4* and *EHMT2*. Other genes included *FOXO1*, a transcription factor that has been linked to tumor growth, *DDB1*, part of a DNA binding complex involved in nucleotide excision repair, or *TNKS1BP1*, a member of the PARP superfamily involved in double-strand break repair [222-224]. Looking at the shared enriched pathways and genes between the 4 clusters (Figure 51) revealed a focus on mitotic cell cycle pathway and cellular response to DNA damage pathway shared between MEgreen and MEsalmon. Genes shared between MEgreen and MEsalmon included DEAD/DEAH-box helicases and methyltransferases such as *RUVBL1, PPP2R2D, RPTOR, TFIP11, DHX30* and *EHMT2*. Between MEgreen and MEpink no pathways were shared but two genes: *TADA2B* and *UBE2D4*. *TADA2B* and its paralog *TADA2A* have previously been reported to influence p53 stability and facilitate apoptosis as response to DNA damage [225]. Likewise *UBE2D4* and its paralog *UBE2D2* are known to be involved in regulation of p53 via ubiquitination [226]. Of

note was *HDAC4*, which was shared between MEgreen, MEsalmon and MEpink. MElightgreen, focused on extracellular functions, did not show overlap with the other three clusters.
The enrichment analysis without previous network analysis demonstrated a link to RNAi related pathways as well as DNA damage repair pathways and other functions. However not all contrasts led to enriched pathways, let alone to a comparable number of enriched pathways if they did. Especially the discrepancy between which investigation methods led to good classification performance and which investigation method led to enriched pathways was interesting. While one could speculate this was due to the differential methylation analysis identifying mainly random noise that happened to lead to good classification power in this dataset I hypothesized that additional filtering was needed. With the applied network analysis methods and subsequent enrichment analysis of the network clusters I was able to identify biological functions mainly related to transcriptional regulation, cell cycle control, chromosomal organization or response to DNA damage. In particular the network plot derived from WGCNA analysis of ND2 revealed a focus on cellular response to DNA damage and mitotic cell cycle control. The affected genes linked to this function included pairs of paralogs and focused on a few shared genes. Especially *HDAC4* was shared among most clusters in the analysis.

7.3.7 Validation

Finally, after having quantified the power of differentiation offered by the identified probes via cross validation above, I further tested the performance with two additional datasets. One dataset consisted of in-house tumor samples from patients with somatic mutations in *TP53* (somatic dataset), the other was prepared from the methylation data generated from liquid biopsy samples collected from Li-Fraumeni patients by Subasri et al. (Subasri dataset) [118].
When applying the classification process to the somatic dataset, the desired behaviour was that none of the somatic *TP53* mutated samples would be classified as germline *TP53* mutated, so the correct prediction would be negative. To measure this behaviour, I calculated the negative predictive value, which could be interpreted as precision for negative predictions. In Figure 40, I show the AUC negative predictive values across multiple thresholds for contrasts IIa and IIb (panel A and B) and both RF methods (panel C and D). Across all methods and thresholds, the negative predictive value was very high, around 0.95, and close to the theoretical maximum of 1.0. For both tumor type adjusted contrasts IIa and IIb the negative predictive value hardly changed across thresholds and contrast IIb obtained slightly higher values than IIa. This indicated that false positive predictions for germline mutated *TP53* cases were unlikely even for very selective thresholds. For the RF methods, one could observe a slight increase in performance with the inclusion of additional probes and a minimally worse performance compared to contrast IIb. This demonstrated a classification behaviour exactly as desired, which would be important for possible future application in a clinical setting.

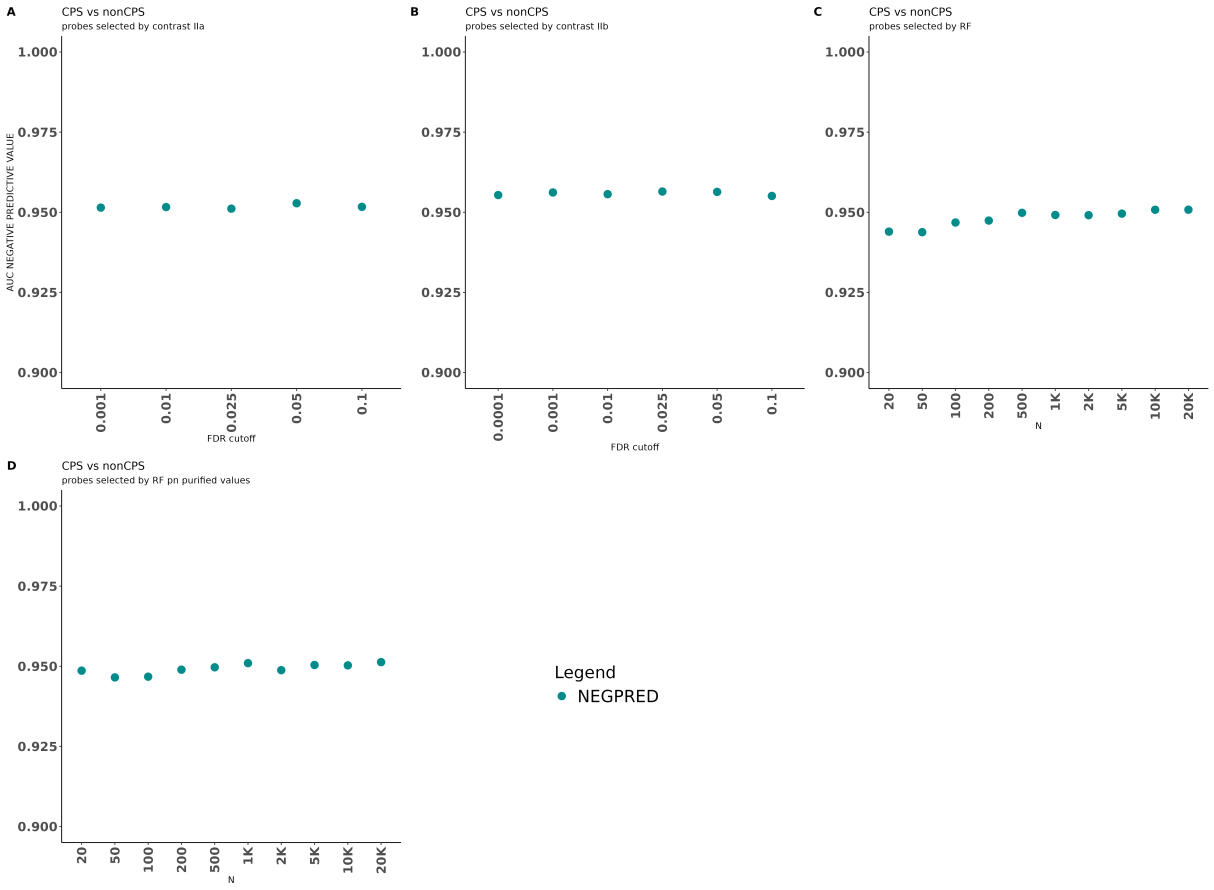


Figure 40: Validation for differentiation between TP53 vs nonTP53 germline mutated tumors using probes identified as differentially enriched by different methods on somatic validation cohort.

Next, I moved on to test the ability to correctly classify samples from germline mutated *TP53* patients that were not in the original discovery dataset. Although the cross validation gave a first insight into this with encouraging results as shown above, I wanted to further test the performance with external data. The Subasri dataset offered a unique opportunity in this regard. Although these results have to be interpreted with care since not only were the Subasri samples processed with a different protocol than the discovery dataset, they were also derived from blood samples in contrast to the tissue samples available in the discovery cohort. Another issue with the Subasri dataset in the context of this investigation was its cell type composition which was vastly different from the samples in the discovery cohort. As shown in the supplements (Figure 49) they mainly were made of neutrophil type cells, CD8 T-cells and B-cells. Only two samples were estimated to contain any cancer cells at all. Nevertheless, in Figure 41 I show the validation metrics, using the Subasri cohort for testing, across multiple thresholds for contrasts IIa, IIb and both RF methods.

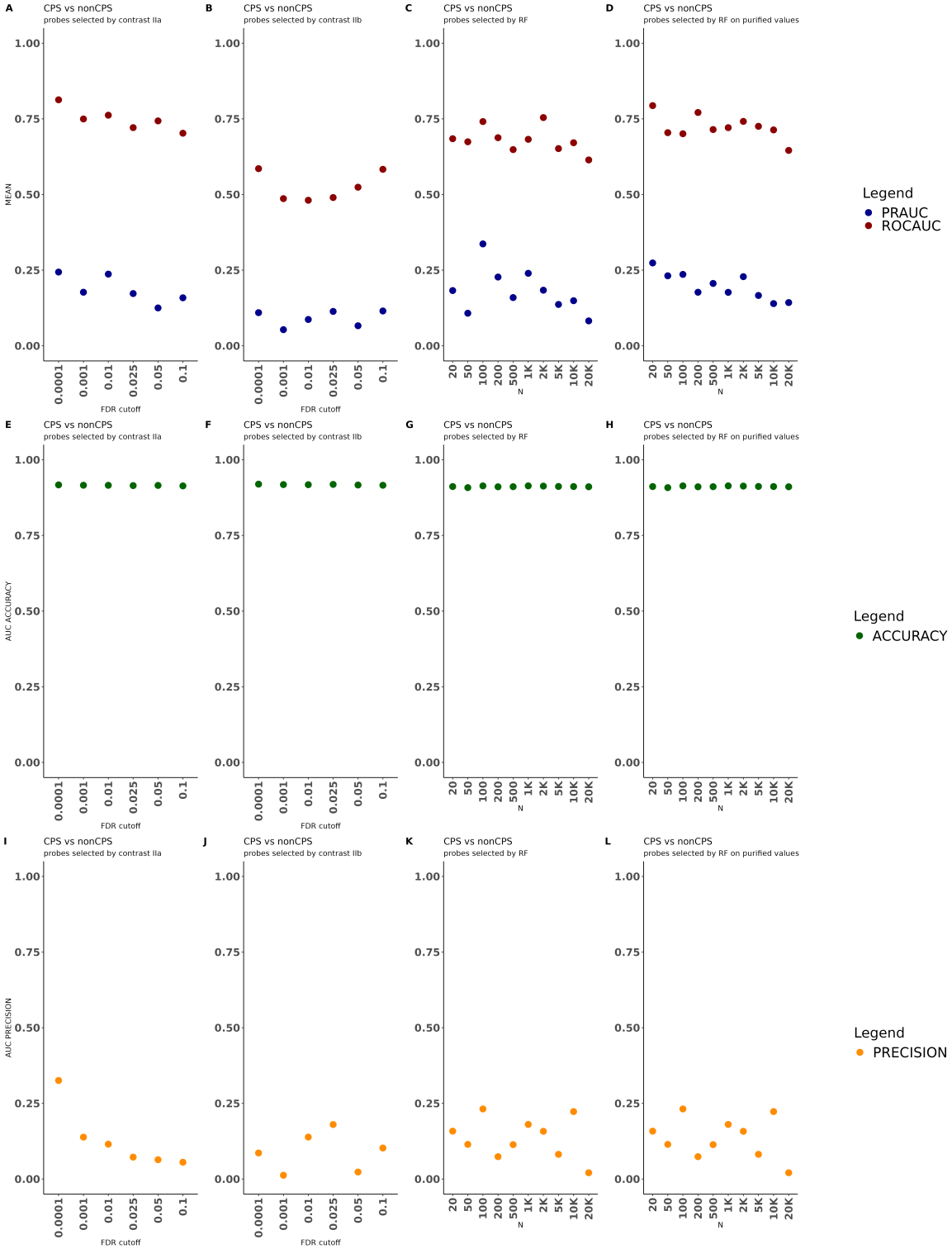


Figure 41: Validation for differentiation between TP53 vs nonTP53 germline mutated tumors using probes identified as differentially enriched by different methods on external validation cohort. A – D shown PRAUC and ROCAUC over different thresholds. E – F show AUC accuracy over different thresholds. I – L show AUC precision over different thresholds.

The ROCAUC value was the highest for contrast IIa and both RF methods, hovering around or just above 0.75, while for contrast IIb it started at slightly above 0.5 before it dipped and then increased again with less restrictive thresholds. While with contrast IIa the ROCAUC did not exhibit major changes, with the RF methods there appeared to be more instability, with sudden spikes in performance at certain thresholds. Also noteworthy was the fact that for RF method on purified methylation values (panel D) the performance appeared to decrease with less restrictive thresholds. The PRAUC value behaved similar for both RF methods, it indicated

spikes in performance and decreased with less restrictive values. Among the linear contrasts, contrast IIa came out on top with a maximum PRAUC value of 0.25, while IIb hovered around 0.15 across thresholds. Looking at the AUC precision both RF methods showed the same behaviour as with the PRAUC and ROCAUC values, indicating spikes in performance at seemingly arbitrary values, although this time spikes appeared at the same values for N for both methods. Among the contrasts, IIa lead to the highest AUC precision values at 0.32 for the most restrictive FDR cutoff. In general, while encouraging, the performance for testing with the Subasri dataset was not as good as the cross validation but that was expected. Ultimately, if the decrease in performance came from the different type of sample (tumor tissue vs liquid biopsy) or the different processing, that could not be fully corrected via the applied batch effect correction methods, or a mixture of both remained unclear.

Next to testing with these two validation cohorts, I compared the probes and regions identified in this study with previous results of studies that investigated the methylation landscape of LFS patients. Wong et al. investigated a liquid biopsy cohort containing a mixture of adult and pediatric Li-Fraumeni patients and identified one methylation signature derived from differential methylation analysis of LFS patients (LFS-signature) and one pan-cancer signature by mapping probes they identified as hypermethylated onto a cancer marker set described by Vrba et al. (pan-cancer-signature) [122, 123]. It must be noted that I could only identify overlaps with either signature with probes or regions identified by contrasts Ia DMR, Ib DMR, IIa DMR, Ia, Ib, IIa and both RF on raw and purified methylation values. The most overlaps were identified for Ia DMR with 53 overlaps (29 from the LFS-signature and 24 from the pan-cancer-signature, notation 29:24) followed by DMR IIa (32:13) and DMR Ib (22:19) with 45 and 41 overlaps respectively. Contrasts Ia, IIa and Ib had 19 (16:3), 18 (13:5) and 10 (10:0) overlaps respectively while both RF methods had 3 overlaps (3:0 for both). Of the 18 overlaps from contrast IIa, 7 were labelled as promoter associated and linked to genes *KLF3*, *PIK3R5*, *TPM4* and *CDK5R1* among others. The overlaps from DMR IIa included 406 unique methylation probes, 165 of which were promoter associated while the rest was labelled as unclassified or had no label. The affected genes include *PIK3R5* again as well as *HOXA1*. Overlap of the probes and regions identified in this study with the regions identified by Wong et al. was encouraging given that they analysed liquid biopsy samples. At the same time, the limited overlap was also expected because Wong et al. investigated a cohort made of 26 pediatric individuals mixed with 63 adult individuals representing a different set of tumor types than in this study and the samples were liquid biopsy. Testing the pan-cancer signature reported by Wong et al. for classification on the discovery cohort resulted in evaluation metrics that described performance below what was achieved with the signatures presented in this study (ROCAUC = 0.74 ± 0.06, PRAUC = 0.43 ± 0.01). The results presented in this study can be interpreted as a refinement of an LFS specific methylation signature towards pediatric cancer patients without the preselection bias introduced by considering only probes described by Vrba et al. The focus of the enrichment analysis on functions related to cell cycle control, RNAi, transcriptional regulation or the RISC complex went beyond the LFS signature described by Wong et al. Additionally, here I presented methylation patterns specific for MMR, but unfortunately comparisons to external cohorts were not possible in the same fashion as with the LFS specific patterns because knowledge is currently much more limited.

8 Discussion

Pediatric cancer is a complex disease and remains one of the leading causes of death worldwide in patients aged 1 – 19. Since 1975, the average survival rate for pediatric cancer has risen by 50%. In recent years, major sequencing and precision oncology programs were launched and aggregated unprecedented amounts of data giving detailed insights into pediatric cancer at multiple omics levels. Despite the major resource investments, the beneficial effects for patients through such programs were mixed. Thus, there still is need for more investigations regarding treatment and diagnosis.

Prediction of synthetic lethality in pedHGG K27M

Treatment of BRCA1/2 deficient cancers with PARP1 inhibitors, leveraging the synthetic lethality interaction between these genes, resulted in clinically relevant improvements of progression free survival e.g. for ovarian and breast cancer [227]. Despite the great interest, discovery of gene pairs with synthetically lethal interaction is very challenging and resource consuming even with high-throughput methods because of the sheer size of the combinatorial space that needs to be covered while simultaneously taking into account the genetic background of a given tumor type. Advancements in *in-silico* prediction methods for SL interactions to narrow down the scope of investigation have gained popularity over the last years. In the first part of this study, I presented a computational approach for the prediction of interacting pairs of genes exhibiting synthetic lethality.

The first challenge faced was the integration of heterogeneous data sources. While non-trivial, integration of multiple information layers derived from omics data, interaction databases and other sources is highly needed since synthetic lethality is a very complex phenomenon and one needs to interrogate multiple data sources to capture a complete picture [228]. From a purely technical standpoint biological data arrives in multiple formats but can almost always be represented in a matrix format. The matrix format has several advantages: it easily allows to describe relations between entities (in the SL context two genes), it is directly accessible and interpretable for manual inspection, it is easily interpreted in other contexts (e.g. as graph adjacency matrix). One strength of the described method was that it used exclusively matrices as input, a useful feature as already pointed out earlier [53]. Other published models for predicting SL interaction need more complex input for example: a genome-scale model of metabolism and an objective function for IDLE, a systems biology markup language model of an organism for Fast-SL, LINCS L1000 expression profiles for EXP2SL or a SL graph for DDGCN, making it much harder for researchers to prepare custom datasets or use additional data to be used with those models [65, 229-231]. Another advantage of using exclusively matrices as input was that any additional data that may be available to researches could be easily integrated. Concerning the actual data sources, there is no consensus currently on what data is needed to sufficiently describe SL interaction. In general the rational was to use data sources that have a causal connection to synthetic lethality phenomenon, analog to efforts in protein-protein-interaction predictions [232]. Some approaches use exclusively protein sequences, exclusively information on SL interaction, GO and KEGG pathways or a mixture of omics-data [65, 233-235]. Another strength of my method was the integration of omics-data as well as publicly available data, which brought advantages discussed in more detail below. Another advantage of the selected set of omics-data from which the input was prepared was that they are widely used in large scale sequencing projects, resulting in a very large pool of potential input data. While more specialized input for example LINCS L1000 expression profiles could offer more performance, generating this kind of data is more complex and often not part of sequencing or precision oncology programs [23, 231, 236].

As hinted at earlier, machine learning techniques were favorable for the available data (see Wang et al. for a review) because of size limitations of the data and because of their ability to integrate multiple data sources without much upfront expertise needed [237]. At first glance, my application of CMFW, previously reported to produce robust results, appeared to confirm these reports [53]. As indicated by multiple metrics, the performance was decent and it reacted as expected to hyperparameter tuning. However, further investigations into the influence of input data composition on performance led to doubts about the suitability of CMFW in this study. Further probing the connection of input data with performance via a shuffle test confirmed my suspicions about CMFW not being able to capture a connection between the available input data and training data. This result stands in contrast to the original publication by Liany et al. where CMFW achieved good performance [53]. However they did not perform a shuffle test. Overall the presented results should not be interpreted as proof that CMFW is unsuitable for SL predictions in general, just that it did not perform well with this dataset, highlighting the need to carefully evaluate the chosen ML model when making SL predictions. Why CMFW failed in this study did not become clear.

Afterwards I pursued other methods for this analysis and in contrast to CMFW, I was able to show that my method captured a connection between the input data and the training data. The ability to capture a connection between input data and training data with my approach was demonstrated via shuffle tests and downsample tests. In separate tests, I was able to show that my strategy of processing data with graphs for feature generation improved the performance above what was achievable with the unprocessed data. Further, I compared the performance of three ML classifiers and was able to confirm that the random forest classifier delivered the most robust performance for the datasets in this study, in good agreement with previous results [63, 174, 175]. This suggested that a random forest classifier has the potential for more general application to other datasets for prediction of SL interaction. However this statement has to be taken with a slight caveat because it is not clear, from this study or literature, if the good performance of random forests is because such algorithms are good in general with tabular data or if there is an aspect making them uniquely suitable for SL prediction [238]. One strength of the data preparation technique presented here was the usage of both context-specific (derived from omics data) and context-free (derived from pathways, databases etc.) features. This improved the robustness against selection bias in the training data [175]. Another strength was the usage of context-specific features derived from topology while the prediction model was feature based. Feature based models are less influenced by selection bias but topology based models are top performers across multiple metrics [174, 175]. This observation might also be a hint at why CMFW failed, because that is a topology based model.

Finally I applied my approach, which was not limited to predictions on a subset of genes in contrast to methods like CMFW, to a dedicated dataset I curated from multi-omics data of pedHGG patients. I presented in this study the first prediction of SL pairs for pediatric high-grade glioma using a cohort of 149 patients to generate the input data that was integrated with other context-free data. In particular, I made predictions based on data derived from a cohort containing only K27M type (n = 70) and predictions based on data derived from nonK27M types (n = 79). This focus on pediatric cancer was another unique feature of the presented results. Because large sequencing projects and other online databases mainly focus on adults, consequently a lot of the SL prediction literature which leverages this data also focuses on adults [237, 239]. Noteworthy about the made predictions was the overlap between those based on the K27M and nonK27M datasets. The same set of genes were predicted for both datasets, although the predicted interaction among them appeared in slightly different combinations for both datasets. This could indicate a smaller influence of selection bias, as application to multiple datasets was a suggested method to estimate this [174]. Overall, the amount of positive SL

predictions was small with only 98 pairs in total, which was both expected and desired because the goal was to limit the scope of a potential downstream investigation and because SL interaction is a rare phenomenon anyway. This overlap may also indicate that the input data was not sufficient for the classifier to pick up the biological differences present. Another interpretation could be that while there are molecular differences between pedHGG types important for diagnostic and treatment purposes, they do not influence the present SL interactions. Whether the problem lies primarily with the input training data that was not specific enough for HGG or with the patient derived omics data remains unclear.
Investigating the made predictions with gene set enrichment of GO terms showed that the involved genes mainly had functions related to the mitochondrion, the Golgi apparatus or cell membrane functions. Further filtering of the made predictions resulted in 55 unique genes predicted to interact in various ways. Taking into account the hallmarks of cancer, especially the characteristic that cancer cells deregulate their energy metabolism, might make it worthwhile to further investigate the predictions related to the mitochondrion. However it must be admitted that despite taking measures to prevent too much selection bias, among the predictions there were most likely false positives. To name one example that might be interesting to follow up on with low-throughput experiments is the predicted SL interaction between HARS1 and HDAC3. The potential benefit of follow up experiments regarding the HARS1-HDAC3 interaction would be that there is ongoing investigation into the therapeutic potential of HDAC inhibitors in pediatric brain cancer with special attention to K27M mutated DIPGs, making a potential SL partner to one HDAC gene an interesting target for combination treatment [176]. Next to this pair, multiple other predicted SL pairs included genes for which drugs are available, including aldehyde dehydrogenase inhibitors, MAO inhibitors, HDAC inhibitors and BAP1 inhibitors, making such pairs also interesting for further investigation. Overall, these predictions may serve as guidance for future research investigating SL interactions specific for pedHGG.

As already alluded to, a general problem with this study was that I trained this classifier on an imperfect set of known SL interaction pairs likely to contain both false positives and false negatives. Further, these SL pairs were not specific for pediatric HGG tumors. Unfortunately, today there is no feasible alternative to using this dataset since there are few confirmed SL interactions available and the evidence for the absence of SL interaction is lacking as well. Likewise the omics data and features engineered from it used in this study were a reflection of current trends in this research area and don’t represent a definitive set of data and features needed to describe synthetic lethality.
One of the biggest challenges when predicting synthetic lethality was how to account for the context specificity of the interactions, which in this study was done by using data derived from patients diagnosed with the tumor in question and comparing it to results produced from data of closely related tumors.
Another point towards preparation of known SL pairs for training purposes is that while there is overlap in identified SL pairs across studies, most SL pairs appear to be specific to the study they were identified in. A harmonized analysis method across studies could improve the situation and lead to better training data. Interesting in this context is a publication by Zhao et al., which described sets of 100-300 genes where interrogation via CRISPR enables loss-of-function prediction across 18000 genes with only minor information loss [240]. This not only suggested that cell line specificity of SL interactions is describable by measuring a very limited set of features, but that it can also be leveraged to make predictions. Integrating such cell line characterization with a prediction model for synthetic lethality in a two-branch model, similar to what is already known from drug response prediction models, could lead to further refinement of SL predictions [241]. Overall this study presented an easily extendible,

performant and widely applicable method for SL predictions and the first prediction of potential SL pairs for pedHGG patients that might guide future research.

Mutational signatures in DADDR patients

Mutational signatures, as a proxy for underlying mutational processes, can be used as a biomarker. A prominent example would be the detection of HR deficiency via the HRDetect tool [98]. However, biomarkers identified for adult cancer cannot be used with pediatric patients without further investigation and validation because of the different characteristics of pediatric cancer [6, 100]. Consequently, there is a need for dedicated investigation into potential biomarkers specific for pediatric cancer. To extend the knowledge of mutational processes related to cancer predisposition syndromes and possibly identify potential biomarkers, an investigation into the association of mutational signatures with CPS syndromes is needed. Gröbner et al. and Thatikonda et al. described the mutational signature landscape for different types of pediatric cancer [100, 145]. Both these studies noted the importance of CPS in the context of pediatric cancer and hinted at the differences these CPS samples exhibited with regard to mutational signatures. While Gröbner et al. worked with the older COSMIC v2 signatures, Thatikonda et al. already used the latest COSMIC v3 signatures and identified a higher contribution of SBS2 and SBS13 to samples with germline *TP53* mutation compared to wildtype samples. In this study, I presented a dedicated investigation into the mutational processes that are active in DADDR patients and identified mutational signatures that showed significant differences in their activity in these patients.

Investigations into the amount of mutations, both in the SBS96 and ID83 context and their distribution revealed the expected behavior across samples with a Pearson correlation coefficient at $R = 0.87$ and a p-value < 0.05. A high correlation between SBS and ID mutations was expected since a higher overall mutational burden leads to both types of mutations, indicating that the processing and downstream mutation calling did not introduce artefacts or miss present mutations [145]. Looking at the individual samples, some exhibited a hypermutator behaviour (> 10 mutations/MB) and some even ultramutator behaviour (>100 mutations/MB). In line with previous results, the ultra- and hypermutator samples were mainly associated with the MMR CPS [242-245].

Proceeding with the evaluation of assigned signatures by SigProfiler and SIGNAL showed some similarities between the two methods but also disagreement. As expected the clock-like mutational signatures SBS1 and SBS5 were most abundant. Another prominent signature was SBS40, which contributed to 38 % of samples when analysed with SigProfiler but was not assigned at all when analysed with SIGNAL. Instead, SIGNAL probably made false assignments of SBS3, a behavior that has already been discussed previously for pediatric cancer [100, 145]. Overall SIGNAL appeared to be less sensitive compared to SigProfiler which I hypothesized was due to the inherent differences in the algorithm, as suggested by previous evaluations [148]. For various signatures I was able to confirm a correlation with age among them SBS40 as described previously [145]. Very recently, SBS40 was proposed to be split into three separate signatures, SBS40a, SBS40b and SBS40c [246]. A subsequent study on lung cancer in non-smokers identified SBS40a to contribute to the majority of analyzed adenocarcinoma samples, which often carried *TP53* driver mutations [247]. This observation could be a hint that the mutational process resulting in assignment of SBS40 (or its variations) is linked to *TP53* mutation status although further research is needed. Some additional signatures I was able to identify with SigProfiler were linked to HR and NER deficiency or APOBEC activity like SBS8 or SBS2 and SBS13 respectively, which was reported previously [145]. While some of the *TP53* mutated samples exhibited SBS2 and SBS13 activity, in line

with previous results linking *TP53* and APOBEC activity, not all of them did [182, 183]. A previously suspected tissues specificity of signatures SBS2 and SBS13 appeared more likely with the presented results in mind although a link to the small intestine as suggested by Wang et al. could not be made [145, 248]. Signatures specific to MMR CPS samples, SBS14, SBS15, SBS20, SBS21, SBS23 and SBS44 were identified as well, appearing exclusively in such samples, as was expected [87, 187]. Correlation analysis of the signatures revealed two clusters of MMR signatures, which I hypothesized were due to the different underlying mutated genes. Among the genes inside the MMR subgroup (*MSH2, MSH6, MLH1, PMS2*), I was able to identify significant differences between the PMS2 germline mutated samples and all other MMR samples regarding assignment of signatures SBS14 and SBS20 as well as other signatures usually suspected to be clock-like. Different mutational patterns resulting from *PMS2* mutation were described earlier and could potentially be leveraged for pediatric patients as well, comparable to the MMRdetect tool [185, 186]. However the identified differences inside the MMR subgroup have to be treated with care because of the very limited set of patients. The correlation analysis revealed another cluster of correlating signatures. Specifically SBS39, which has unknown aetiology, correlated with signatures linked to DNA repair function. Currently SBS39 is most often identified in breast cancer samples and has been reported to correlate with APOBEC activity [153].
Regarding ID signatures, twelve different signatures were identified with SigProfiler most prominently, as expected, the clock-like signatures ID1 and ID2. Next to those two signatures, other ID signatures linked to NHEJ repair mechanisms such as ID8 were extracted, interestingly only in samples from patients with underlying germline mutation in *TP53* or *BRCA1/2*, not in MMR patients. Co-occurrence of SBS3 and ID6, which has previously been reported for samples with HR deficiency, was not observed in samples with *BRCA* germline mutation, a result in line with findings by Thatikonda et al. [100, 145]. Another identified signature, ID3, previously linked to tobacco smoking, was correlated in this analysis with signatures SBS31 and SBS35, which are linked to treatment with platinum based drugs. This further supported previous suggestions where ID3 could be linked to treatment induced DNA damage or another unknown mutational process [145].

Next, I tested the differences of signature activity between germline and somatic *TP53* mutations. Among the significant differences identified were clock-like signatures like SBS1, SBS5, SBS40, ID1 and ID2 despite the investigation method accounting for patient age. This could indicate that the level of activity of these signatures is different between germline and somatic patients. In case of SBS40 there could be a more direct link to TP53 mutation status as discussed above. Re-analysis with the split signature (SBS40a, SBS40b and SBS40c) could be beneficial. Other signatures identified with a significantly higher contribution to somatic cases, that were previously linked to other characteristics or without aetiology, were SBS9, SBS18 and ID4. SBS9 has been previously linked to DNA damage. SBS18 shares a similar profile with SBS36, that was associated with defective base excision repair, which could lead to false assignment. The identified signatures could be leveraged via their activity for future efforts towards classification. Unique assignment of a signature to only germline or somatic TP53 cases was not observed nor a novel signature.
Finally, I investigated the association of mutational signatures with the different germline mutations in this cohort by comparing germline mutated samples with wildtype samples. This identified, among others, SBS10a, SBS10b, SBS11, SBS12, SBS19, ID6 and ID9 to be significantly influenced by germline mutation status. SBS10a and SBS10b were linked to polymerase epsilon status and usually contributed to a high mutational load resulting in hypermutators which were inside the MMR group in this study [249]. ID6 was linked to defective homologous DNA damage repair, usually linked to BRCA mutation [87]. Signatures SBS12, SBS19 and ID9 currently have unknown aetiology. These signatures could serve as

potential biomarker to detect germline mutation status similar to what has been proposed with MMRdetect, however more research and careful evaluation is needed [185]. Other signatures were associated with germline mutation and simultaneously with tumor type. This could be due to correlation among germline mutation and tumor types or hint at a tissue specific influence with those signatures. However further investigations into the quantity of mutations assigned did not reveal an obvious pattern regarding specificity for one of the germline mutations. The original reason to use SIGNAL along with SigProfiler was the tissue specific signatures the former offers. Unfortunately these could not be leveraged as hoped to separate tissue specific effects from germline related effects but future investigations might include such considerations. Overall no obvious separation by a single signature was identified. Instead the different levels of activity could be leveraged for a classification.

Despite the limited size of this cohort, especially in the MMR (n = 11) and BRCA (n = 6) groups, the presented results could serve as refinement of the mutational landscape in pediatric cancer with an emphasis on patients with germline mutations. Inclusion of more patients across a more diverse set of tumor types could improve the power of the analysis, but recruitment of additional pediatric DADDR samples is slow and is hampered in any case due to the fact that certain tumor types are more common for certain CPS. Further investigation into structural variants might also be beneficial in the context of DADDR patients, at least for LFS they play a role [250, 251]. The strength of SIGNAL, the tissue specific signatures, could not prevail. In future studies with more information on the tissue of origin, in the SIGNAL sense, consideration of tissue specific signatures could lead to more insights.

Methylation patterns in DADDR patients

The methylome of cancer cells allows not only tracing of the origin of cells but also is a powerful tool for classification of tumors and for use in identifying biomarkers by differential analysis. Classification via the methylome of tumors into categories not distinguishable by morphological characteristics is possible as demonstrated by Capper et al. [44]. Currently there is only limited knowledge about methylation patterns associated with CPS syndromes in pediatric cancer. While the classification tool presented by Capper et al. gives robust results for most tumor samples, a drop in performance was observed when applying this technique in-house to samples that harbor a germline mutation, suggesting a distinct methylation landscape in these samples. Research into methylation patterns specific for Li-Fraumeni patients was done previously using liquid biopsy samples and identified differentially enriched methylation sites, although those investigations often included adult patients, influencing the applicability of the results on pediatric patients [117, 118, 122].

In this study, I presented results from the investigation of a cohort of exclusively pediatric patients with different CPS, including Li-Fraumeni patients, using matched tumor tissue and blood samples, giving a deeper insight into the associated methylation pattern and refining previous results based on liquid biopsy samples from mixed cohorts. Focusing on samples with *TP53* germline mutation (n = 62) and MMR syndrome (n = 20), I assembled a control cohort of samples from the same tumor type, making sure that no germline or somatic mutation in *TP53, PMS2, MSH2, MSH6* or *MLH1* was present in the control samples. Further filtering of control samples with several strategies including based on correlation or distance in 2D projections to CPS samples did not improve results (data not shown) but could be further investigated in the future.

Since tumor tissue samples in general do not consist of cancer cells exclusively, I estimated the cell type composition of the tumor tissue samples. As expected, all samples contained a mixture of various cell types, ranging from immune cells to blood cells. Further statistical testing revealed significant differences in cell type composition between germline *TP53* and control

cases and showed the different influences of tumor types, which justified the application of the purification algorithm. The TME of pediatric cancers is an area of ongoing research. While most pediatric solid tumors were reported to be immune-cold some of them showed invasion with immune cells but those lacked activation. However these studies did not take into consideration the presence of possible cancer predisposition syndromes despite a not insignificant amount of affected pediatric patients [252-254]. The identified higher invasion of certain immune cells like natural killer cells in CPS tumors could suggest a possible benefit from immune therapy for those patients, although further investigations would be needed because this study only considered the TME insofar as to correct for it.

Based on the composition estimations, I applied a purification algorithm to obtain the methylation signal specific for the cancer cells. In a two-pronged approach, I applied statistical methods on both the original raw methylation data as well as the purified methylation data for the identification of differentially and variably methylated probes and regions. The number of identified differentially methylated probes for all applied statistical methods exhibited the expected behavior concerning the number of identified probes. This indicated that the methods accounting for tumor type identified less probes as significant compared to the methods that only accounted for presence of CPS. Similarly, analysis using the purified methylation values resulted in less significant probes compared to their counterparts using the raw methylation values. This was expected because the purification process removed noise stemming from non-cancer cells that could mistakenly be picked up by the differential analysis. One strength of this study was that the purification process actually obtained methylation beta values that could be further analyzed, not just estimations of cell type composition that were integrated as factors in the analysis. While there can be certain benefits to this as demonstrated by Lee et al., working with actual methylation beta values allowed more flexibility regarding the downstream analysis methods [255].

One caveat with the applied purification algorithm was a somewhat uncharacteristic distribution of methylation values after purification. The resulting beta value profiles with less defined peaks at both ends of the scale appeared unfamiliar. Inspection of the influence of the purification process on the control cohort revealed a similar picture although less extreme with peaks shifted less from their original position. One explanation could be in the technical details of the applied algorithm, which leveraged Bayesian statistics and assumed a uniform prior for cancer specific methylation, which could have contributed to the resulting less defined distribution of the purified beta values. Another explanation could be that the patients in the LFS and MMR groups actually had a more than usually shifted methylation landscape. At the same time, one has to acknowledge that the results obtained by analysis of purified methylation values resulted in good differentiation performance, at times even better compared to the raw methylation value analysis, lending some credibility to the purification algorithm. Looking at the overlap of identified probes between analysis methods, the most agreement was between the linear contrasts regardless of raw or purified methylation values. A larger effect of the purified beta values on the analysis with random forest was observable compared to the linear models with much less overlap in identified probes between random forest methods. This apparent instability of the random forest approach could be a result of the limited dataset available and may be improved with inclusion of more patients. The capabilities of random forests for identification of methylation biomarker has been demonstrated by Capper et al. so in the future with more data available the random forest analysis method could be applied again for further refinement of results [44].

Proceeding with the quantification of classification power offered by the sets of identified probes, I was able to show that the desired behavior was achieved. A quality of classification, as quantified by PRAUC, ROCAUC and other metrics, above baseline was achieved with a much more limited set of probes. Especially the probes identified via linear models achieved particular good performance while the random forests methods could not reach a performance

above baseline. Looking at the achieved precision values, a metric particularly important in a diagnostic setting, the probes identified via linear model applied to purified methylation values achieved the best performance although the probes identified via linear model on raw methylation values were not far off. This again underlined the issues that were apparent with the random forest analysis in this study. Further evaluation revealed a performance above what was achievable with a previously suggested LFS pan-cancer signature. The previously suggested LFS pan-cancer signature was obtained by analysis of liquid biopsy samples from a mixed cohort of pediatric and adult patients. Consequently the presented signatures in this study can be interpreted as a refinement of previous signatures with special attention towards pediatric patients because I investigated only pediatric patients and did not preselect methylation probes. Fearing that the achieved performance was only a technical artefact veiled by the encouraging metrics, I turned towards a more qualitative inspection of the achieved separation between CPS and nonCPS cases. UMAP plots confirmed the assessment that separation was achieved and that the linear contrasts delivered better performance compared to random forest application. Further, it was visible that the measured performance was not merely a technical artefact but there were clusters of CPS cases. One thing immediately noticeable was that there appeared to be two clusters of CPS cases, one dominated by sarcoma tumors, the other by nonsarcomas. These two clusters were further confirmed by application of PCA and KNN clustering. Such a behavior of sarcoma cases needing separate consideration was already known and resulted in two tumor type classifier in previous studies, one dedicated for sarcoma types, one dedicated to central nervous system tumors (www.molecularneuropathology.org). In this study whether the shown behavior of forming two clusters was a result of underlying biology or was a reflection of the used dataset remained unclear. Splitting the cohort further appeared unfeasible because of the already limited availability of CPS samples. Re-analysis while integrating a variable in the linear models accounting for the observed clusters could be an option for future research. Comparing the qualitative differentiation power between linear contrasts that took tumor type into account and those who only accounted for CPS showed the expected behavior. The tumor type cluster were more clearly separated with sharper borders with the linear contrasts that only took CPS into account while the contrasts that accounted for tumor type produced more fluid borders and a more compact clustering across tumor types. This was an encouraging observation that could mean that accounting for tumor type worked as intended. Of course, better results could be obtained with more samples, especially a more uniform distribution of samples across tumor types. However, since recruitment of patients with CPS tumors is slow, the dataset analyzed in this study represents, to the best of my knowledge, the largest of its kind. In any case the bias towards certain tumor types could most likely never be rectified completely because certain tumor types are associated more frequently with certain CPS. The fact that the probes identified via linear models gave better performance than those identified via random forest could be the result of the ability to encode the matching nature of samples, the different types of tissue and the tumor groups explicitly in the linear model. As already mentioned with increased sample size a revisit to the random forest analysis could lead to better performance eventually.

Turning towards the biological interpretability of the identified sets of differentially and variably methylated probes and regions, I presented the significantly enriched GO terms and DDR pathways. Using linear models only accounting for CPS (models Ia/Ib), most enriched pathways were identified using the differentially and variably methylated regions. Across all three major GO categories, enriched pathways were related to the RISC complex, RNA interference, miRNA regulated gene silencing, transcriptional regulation, embryonic skeletal development and organ morphogenesis. Genes inside the RISC complex in which differentially methylated probes were located included, next to a number miRNA coding genes, *DHX9, DCP2* and *DICER1*. Especially the association of germline TP53 mutation with miRNA functionality has been pointed out earlier [117]. Comparing the effects of mutations in *TP53* on miRNAs for

adult cancer and the presented results may offer new insights into the miRNA landscape in pediatric patients [256]. The linear models accounting for tumor type (model IIa/IIb) enriched pathways across all three major GO categories but also some DDR pathways. The function of the enriched pathways were related to energy metabolism, cell-cell adhesion and again the RISC complex and RNAi. Further, DDR pathways related to CPF, BER and NER were significantly enriched. A link between LFS and miRNA and the RISC complex and its importance for tumorigenesis has been discussed earlier [257]. In particular it was pointed out that the interferon signaling pathway (IFN) was disrupted by methylation [258, 259]. Further the authors pointed out that treatment with demethylating drugs restored some function of IFN which led to senescence [260]. Currently there are multiple studies testing drugs that influence the methylome in pediatric cancer and additional special attention towards patients with DADDR syndrome could lead to novel insights [261]. Other pathways implicated in the study by Fridman et al. included cell cycle pathways and cytoskeletal pathways that were also significantly enriched in this study [260]. Cytoskeletal components are under active research as drug targets and special attention towards pediatric patients with CPS could bring benefits [262]. While the cited study was conducted on immortalized cells from an adult male patient, with the presented results a link to pediatric cancer has been made.

Comparing the number of probes and regions identified as differentially or variably methylated with the number of enriched pathways resulting from them, I hypothesized that noise carried over from the differential analysis disturbed the enrichment analysis. To deal with the noise, I applied network analysis methods to obtain clusters of probes with highly correlated methylation patterns. Correlation analysis of these clusters with traits of interest underlined the different behaviour exhibited by sarcoma and nonsarcoma samples observed earlier in the UMAP plots. Some of the identified network clusters (n = 5) correlated with disease state (primary, progression or relapse) but unfortunately for those no enriched pathways could be identified at the applied FDR threshold. For example, cluster MEplum1, identified using WGCNA with IIa and IIb probes (Figure 38), correlated with progression state. Affected genes in the MEplum1 cluster included *HDAC4* and *PRMT5*. HDAC inhibitors have been studied earlier as treatment in pediatric brain cancer while *PRMT5* has been discussed as potential target in medulloblastomas and treatment relevant links to *TP53* mutation status have been identified [176, 263, 264]. Looking at which cluster contained a significant amount of probes identified as differentially methylated revealed a subset of clusters. Running enrichment analysis on all clusters obtained via network analysis but focusing on those that contained a significant amount of probes identified as differentially enriched revealed further insight into the biology behind the differentially enriched probes. The pathways identified in those cases were associated with functions in energy metabolism, nucleotide binding, embryonic skeletal development, chromatin organization, polymerase transcription regulation, chromosomal organization, cellular response to DNA damage and mitotic cell cycle to name but a few. Concentrating at the most significant enriched pathways and the most affected genes inside them revealed a focus on certain biological functions and genes. Pathways related to mitotic cell cycle and cellular response to DNA damage appeared to be most important. Further *HDAC4* was implicated most often. In a recent study that investigated pediatric CNS tumors and accounted for cell type composition, *HDAC4* was also one of the most important genes identified [255]. Previously *HDAC4* has been associated with multiple cancer types as well as unfavourable disease progression, an observation in line with the above mentioned MEplum1 that was associated with progression [265-268]. Although targeting of HDACs alone or as part of a combinatorial treatment isn't a novel idea, further evaluation for CPS patients could be beneficial [269]. Another thing to note were the affected paralog pairs identified. There could be potential for a treatment if the differential methylation of both paralogs led to some form of synthetic sickness. In general, the network analysis offered the possibility to investigate smaller correlated subsets

of differentially methylated probes that otherwise might have been drowned by the overall noise. Considering the genes affected by differential methylation, some of which were highlighted above, it appeared unlikely that the identified sets of differentially methylated probes were coincidental or based on technical artefacts but rather were related to underlying biology.

Finally, I turned towards additional validation that complemented the already encouraging results obtained via cross validation. Investigating the ability of the identified probes to distinguish somatic from germline *TP53* mutated samples by using a dedicated dataset containing only samples with somatic mutations demonstrated robust performance. Across all applied thresholds and for all methods of investigation a very high negative predictive value at or above 0.95 was measured. Negative predictive value can be interpreted as precision for negative predictions, meaning that the identified probes result in a low chance for false positive predictions erroneously labeling a somatic case as a germline case. This knowledge in itself may provide a benefit for treatment decisions because knowing that a given patient is not affected by MMR or LFS is important information. Overall such a behavior is very important and highly desirable in case of possible clinical application with patients in the future. Taking one step further, I prepared a dataset based on liquid biopsy samples from Li-Fraumeni patients to test a more generalized applicability. The best performance labeling the germline mutated *TP53* cases was achieved with the probes identified by a contrast taking into account the matching nature of the blood and tumor tissue samples in the discovery cohort. Even though the performance achieved during testing on the liquid biopsy samples was far below what was achieved in cross validation, this result is encouraging. Precision of 90% was still achievable, although at cost for recall but that might be acceptable in a clinical setting. The contrast fully leveraging the matching blood and tumor tissue samples performed the best, indicating that the performed analysis works as intended and currently the bottleneck is mainly the small sample size. Especially considering the fact that not all germline samples in the discovery cohort had matching blood samples available. Further, the processing of the samples in the liquid biopsy cohort was different from the sample processing of the discovery cohort. Even though I was able to correct for this to some extent with different batch effect correction methods, a standardized sample processing would benefit the performance. Another major issue with the liquid biopsy samples was revealed by cell type estimation. As discussed only two samples were identified to contain any cancer cells at all. Of course these estimations have to be taken with a grain of salt, but overall the cancer cell fraction in the liquid biopsy samples was very low. Comparing the probes identified in this study with methylation signatures identified by Wong et al. showed some overlap [122]. Taking into account that Wong et al. investigated liquid biopsy samples, that their cohort was a mixture of adult and pediatric patients and that they preselected probes for their pan-cancer signature, this study offered refinement of the methylation patterns in germline mutated TP53 patients because it focused exclusively on pediatric patients and did not introduce a preselection bias for the probes. This refinement lead to the demonstrated increased performance achieved with the methylation signatures presented in this study beyond the one described by Wong et al. The investigation of tumor tissue samples instead of liquid biopsy samples was a great strength of this analysis because liquid biopsy does have its drawbacks, in particular when sensitivity is concerned [270]. In addition to the presented results for germline *TP53* mutated cases, I presented results for patients with germline mutations in MMR related genes. The analysis for this cohort had to be interpreted with care however because the cohort was only 1/3 of the *TP53* cohort and much more biased towards one tumor type. Also in the analysis of the MMR cohort, I did not take into account if a Lynch syndrome of CMMRD was present which might have an influence [77, 271]. Unfortunately, for the MMR cases there was no external validation cohort available, so further confirmation beyond cross validation was not possible. Judged by the performance of the analysis methods

with the *TP53* cohort, the results produced from the analysis of the MMR cohort had a certain credibility.

In summary, I further expanded on the latest computational methods for prediction of synthetic lethality by combining several of the best aspects of previous methods. Especially the combination of techniques that offer robustness towards selection bias with topology based techniques that are known to offer good performance stands out. I integrated multiple sources of omics input prepared from a dedicated set of pediatric patients to generate the first SL predictions for pedHGG tumors. My predictions may serve as a basis to decide on future targets when investigating vulnerabilities in pedHGG. The computational method for predicting SL pairs can be easily used and extended by researchers and is set up in such a way that seamless integration of additional data is possible. Further, I applied state of the art bioinformatics methods to broaden the knowledge on mutational processes and the methylome in DADDR patients. I applied and compared two cutting-edge mutational signature calling algorithms, described the mutational signature landscape of DADDR patients which further confirmed several observations made previously and led to novel suggestions for signature aetiology. Additionally, I compared intra MMR syndrome variability and described mutational signatures associated with underlying germline mutations. Concerning the differences between wildtype, germline or somatic mutation I identified several signatures with significantly different contributions that may serve as basis for a classification tool. Concerning the methylome of DADDR patients, I identified methylation signatures capable of identifying such patients with high precision which may improve molecular diagnostic tools that previously struggled with such patients. The presented signatures can be viewed as a refinement of previously published signatures, offering improved performance and specificity for pediatric patients. In the process of my analysis I also discovered interesting properties of the immune cell composition of the investigated samples that may offer possibilities for future research avenues. Finally, with my work I was able to show that DADDR syndromes influence the methylome in a way that several key biological functions are implicated, for example functions related to *HDAC4* or cytoskeletal functions. Some of these functions offer possibilities for treatment or are currently under investigation for treatment and could benefit from special attention towards their impact on DADDR patients.

8.1 Outlook

The presented results offered insights into potential novel avenues for future research. The predicted SL interactions may serve as guideline for future high-throughput interaction screens in pedHGG cell lines, especially the predictions involving HDAC could be interesting for pedHGG K27M. Until that time, a manually selected subset of the predicted SL pairs could be investigated in low-throughput experiments for evaluation. Leveraging one of the major advantages of the presented ML model, the ability to integrate additional data sources, future research could include additional layers of data or more patients as they become available. For example more context-free features such as protein sequence or context-specific features such as methylation could be useful. While major online databases still focus on adults, in recent years data for pediatric patients became available which could be integrated as well. Especially useful for SL prediction could be the pediatric DepMap that includes CRISPR screens, which were leveraged to great effect in other SL prediction projects [272]. Computational aspects of the presented methods could also be further investigated, especially the generation of topology based features and evaluation of other tree-based ML models could yield better performance. Another area future research could focus on, are the much needed improvements of cell type specificity. The identification of small subsets of genes whose loss-of-function behavior enables predictions for the loss-of-function behavior of the remaining genes, could be extended and applied to two way interaction predictions as presented in this study [240]. The necessary two branch prediction models, were one arm is responsible for SL prediction and the other is responsible for cell line specificity, could be adapted from drug response prediction models in use today. In the long run, standardized detection of SL pairs from high-throughput screen could improve the needed training data and consequently the predictions.

The current cohort of DADDR patients used for analysis of mutational signatures was biased towards certain tumor types, especially sarcomas and high-grade gliomas. Inclusion of more samples from underrepresented tumor types could improve and generalize the analysis, although this will be difficult since these CPS are enriched in certain tumor types. Nevertheless the *de novo* extraction of mutational signatures could benefit from more input data. At the same time, bleeding effects in the assignment of signatures between tumor types could be reduced, giving a clearer picture of the mutational landscape. Expanded annotation including treatment response and clinical outcome for samples could add beneficial insights, especially for the presented considerations towards ID3. Also the application of tissue specific signatures to distinguish tissue effects from CPS effects could drastically improve the analysis. Translational relevance of the signatures identified to be associated with the CPS does require further investigation into their discriminatory power. One interesting aspect was SBS40 and the latest signatures based on it (SBS40a, SBS40b and SBS40c) which could be linked to *TP53* germline mutation but more research is needed. Other possible avenues include further investigation into the identified signatures contributing more towards *PMS2* affected samples than to other MMR samples which may enable a higher mutational signature based resolution of MMR mutations.

Like the dataset used for mutational signature calling, the discovery dataset for methylation patterns associated with CPS was biased towards sarcomas and high-grade gliomas. Further inclusion of samples of underrepresented tumor types should improve the power of the applied analysis although the inherent bias towards certain tumor types is an issue here as well. Crucially, further inclusion of paired samples for which both blood and tumor tissue methylation data is available is of utmost importance. This will benefit the generalization of the identified CPS associated patterns and further improve the diagnostic power as hinted at with the evaluation of the external liquid biopsy cohort. Especially a more thorough evaluation of the application of the methylation signatures on liquid biopsy samples is advisable with a clinical application in mind to fully reap the benefits of liquid biopsy samples as a surveillance

measure. Stepping towards clinical application for example with the MNP project, further tuning of the algorithm for classification is advisable before eventual dissemination. Specifically the application of gradient boosting algorithms is promising in this area because of their suitability for tabular data and their capability to deal with missing values. Another exciting avenue for future research could be further refinement of the methylation purification algorithm. In its current state the distribution of the purified values is somewhat unusual. Further work is required to better understand the reasons behind this and ideally amend the purification algorithm. As discussed there were influences of the tumor type and its interaction with CPS status on the TME. Immunotherapy is a promising field and further research could help to better understand its applicability in pediatric patients with CPS. Other possible avenues of research to improve the overall outcome for patients could include further investigation into HDAC targeting drugs or demethylating drugs or possible combinatorial treatment involving them.

9 Acknowledgements

I want to thank **Dr. Natalie Jäger**, my first supervisor, for giving me this great opportunity to join the Clinical Bioinformatics group, for her support and guidance throughout.

A big thank you to **Dr. Robert Autry**, my second supervisor, for the smooth transition, for his support, his advice, his guidance, helpful discussions and encouragements. He played a major role in the successful completion of this thesis.

I also want to thank **Prof. Dr. Stefan Pfister**, for giving me the opportunity to join his division, for his advice in the TAC meetings, for the helpful discussions, for pointing in the right direction when I was stuck, for his guidance and his open ear.

Further, I want to thank **Prof. Dr. Michael Boutros** for his advice and his input on the TAC committee and for acting as referee and chair of the defense committee.

To my group, a big thank you for your support and fun moments and going together through difficult COVID lockdowns and extended periods of remote work. **Pengbo** and **Apurva**, who were there from day one, **Elias**, **Dina**, **Luisa**, **Devishi** and **Martha** a big thank you. A big thank you to **Christopher** and **Prakash** for helping me find my way around INFORM data. And a big thank you to **Martin** for his help with all things methylation. And a big thank you to **Rolf** for his help with computational problems.

I want to thank my **parents**, for supporting me from day one and providing inspiration to pursue science.

Last but not least, the biggest thank you goes to **Silke** and **Oskar**. Without your support through uncertain and hard times, this would not have been possible.

10 Appendix

Table 5: INFORM patients used in the predictions of SL pairs for both CMFW and classic ML methods. Patients are annotated with their tumor type. Here both K27M and nonK27M patients are listed.

TUMOR TYPE	INFORM_PUBLIC_PID
HGG_K27M	INF_R_140_primary
HGG_K27M	INF_R_031_relapse1
HGG_K27M	INF_R_1261_relapse1
HGG_K27M	INF_R_005_primary
HGG_K27M	INF_R_006_primary
HGG_K27M	INF_R_007_primary
HGG_K27M	INF_R_043_primary
HGG_K27M	INF_R_047_primary
HGG_K27M	INF_R_089_primary
HGG_K27M	INF_R_091_primary
HGG_K27M	INF_R_1130_primary
HGG_K27M	INF_R_1153_primary
HGG_K27M	INF_R_1155_primary
HGG_K27M	INF_R_1186_primary
HGG_K27M	INF_R_1194_primary
HGG_K27M	INF_R_1199_primary
HGG_K27M	INF_R_1209_primary
HGG_K27M	INF_R_1286_primary
HGG_K27M	INF_R_1339_primary
HGG_K27M	INF_R_1350_primary
HGG_K27M	INF_R_1387_primary
HGG_K27M	INF_R_1446_primary
HGG_K27M	INF_R_145_primary
HGG_K27M	INF_R_176_primary
HGG_K27M	INF_R_199_primary
HGG_K27M	INF_R_201_primary
HGG_K27M	INF_R_206_primary
HGG_K27M	INF_R_253_primary
HGG_K27M	INF_R_257_primary
HGG_K27M	INF_R_281_primary
HGG_K27M	INF_R_293_primary
HGG_K27M	INF_R_331_primary
HGG_K27M	INF_R_449_primary
HGG_K27M	INF_R_472_primary
HGG_K27M	INF_R_510_primary
HGG_K27M	INF_R_529_primary
HGG_K27M	INF_R_533_primary
HGG_K27M	INF_R_538_primary
HGG_K27M	INF_R_539_primary
HGG_K27M	INF_R_574_primary
HGG_K27M	INF_R_614_primary
HGG_K27M	INF_R_617_primary
HGG_K27M	INF_R_665_primary
HGG_K27M	INF_R_717_primary
HGG_K27M	INF_R_772_primary
HGG_K27M	INF_R_813_primary
HGG_K27M	INF_R_815_primary
HGG_K27M	INF_R_835_primary
HGG_K27M	INF_R_843_primary
HGG_K27M	INF_R_845_primary
HGG_K27M	INF_R_854_primary
HGG_K27M	INF_R_898_primary
HGG_K27M	INF_R_961_primary
HGG_K27M	INF_R_989_primary
HGG_K27M	INF_R_928_progression
HGG_K27M	INF_R_954_progression
HGG_K27M	INF_R_955_progression
HGG_K27M	INF_R_1068_progression
HGG_K27M	INF_R_1092_progression
HGG_K27M	INF_R_782_relapse1
HGG_K27M	INF_R_878_relapse1
HGG_K27M	INF_R_1329_relapse1
HGG_K27M	INF_R_1362_relapse1
HGG_K27M	INF_R_1376_relapse1
HGG_K27M	INF_R_1403_relapse1
HGG_K27M	INF_R_375_relapse2
HGG_K27M	INF_R_548_relapse2

HGG_K27M	INF_R_709_relapse2
HGG_K27M	INF_R_1254_relapse2
HGG_K27M	INF_R_542_relapse6
HGG_X-OTHERS	INF_R_273_primary
HGG_X-OTHERS	INF_R_013_primary
HGG_X-OTHERS	INF_R_1451_primary
HGG_X-OTHERS	INF_R_223_primary
HGG_X-OTHERS	INF_R_1060_primary
HGG_X-OTHERS	INF_R_1082_primary
HGG_X-OTHERS	INF_R_1264_primary
HGG_X-OTHERS	INF_R_859_primary
HGG_X-OTHERS	INF_R_1320_relapse1
HGG_X-OTHERS	INF_R_042_primary
HGG_X-OTHERS	INF_R_1123_primary
HGG_X-OTHERS	INF_R_864_primary
HGG_X-OTHERS	INF_R_114_primary
HGG_PXA	INF_R_1061_primary
HGG_pedRTK1	INF_R_729_relapse1
HGG_pedRTK1	INF_R_118_primary
HGG_pedRTK1	INF_R_148_primary
HGG_pedRTK1	INF_R_431_primary
HGG_pedRTK1	INF_R_639_primary
HGG_pedRTK1	INF_R_1164_primary
HGG_pedRTK1	INF_R_1204_primary
HGG_pedRTK1	INF_R_1223_primary
HGG_pedRTK1	INF_R_1341_primary
HGG_pedRTK1	INF_R_968_primary
HGG_X-OTHERS	INF_R_237_relapse1
HGG_pedRTK1	INF_R_305_progression
HGG_pedRTK1	INF_R_1078_progression
HGG_PXA	INF_R_384_progression
HGG_X-OTHERS	INF_R_667_progression
HGG_X-OTHERS	INF_R_1369_progression
HGG_X-OTHERS	INF_R_609_relapse1
HGG_pedRTK1	INF_R_212_relapse1
HGG_pedRTK1	INF_R_756_relapse1
HGG_pedRTK1	INF_R_940_relapse1
HGG_pedRTK1	INF_R_1180_relapse1
HGG_PXA	INF_R_063_relapse1
HGG_PXA	INF_R_727_relapse1
HGG_PXA	INF_R_1189_relapse1
HGG_PXA	INF_R_1234_relapse1
HGG_X-OTHERS	INF_R_026_relapse1
HGG_X-OTHERS	INF_R_195_relapse1
HGG_X-OTHERS	INF_R_594_relapse1
HGG_X-OTHERS	INF_R_1063_relapse1
HGG_X-OTHERS	INF_R_1281_relapse1
HGG_X-OTHERS	INF_R_072_relapse1
HGG_X-OTHERS	INF_R_127_relapse1
HGG_X-OTHERS	INF_R_320_relapse1
HGG_X-OTHERS	INF_R_085_relapse1
HGG_X-OTHERS	INF_R_1275_relapse1
HGG_X-OTHERS	INF_R_057_relapse1
HGG_X-OTHERS	INF_R_292_relapse1
HGG_X-OTHERS	INF_R_1160_relapse1
HGG_X-OTHERS	INF_R_1358_relapse1
HGG_X-OTHERS	INF_R_1058_relapse1
HGG_X-OTHERS	INF_R_1357_relapse1
HGG_X-OTHERS	INF_R_1448_relapse1
HGG_X-OTHERS	INF_R_220_relapse1
HGG_pedRTK1	INF_R_1073_relapse1a
HGG_pedRTK1	INF_R_1176_relapse2
HGG_PXA	INF_R_707_relapse1
HGG_PXA	INF_R_065_relapse2
HGG_PXA	INF_R_241_relapse2
HGG_PXA	INF_R_302_relapse2
HGG_PXA	INF_R_490_relapse2
HGG_PXA	INF_R_1188_relapse2
HGG_X-OTHERS	INF_R_025_relapse2
HGG_X-OTHERS	INF_R_1246_relapse2
HGG_X-OTHERS	INF_R_505_relapse2
HGG_X-OTHERS	INF_R_638_relapse2
HGG_X-OTHERS	INF_R_806_relapse2

HGG_X-OTHERS	INF_R_893_relapse2
HGG_PXA	INF_R_1319_relapse3
HGG_X-OTHERS	INF_R_1048_relapse3
HGG_X-OTHERS	INF_R_1301_relapse3
HGG_X-OTHERS	INF_R_686_relapse3
HGG_X-OTHERS	INF_R_073_relapse3
HGG_PXA	INF_R_513_relapse4
HGG_X-OTHERS	INF_R_569_relapse4
HGG_X-OTHERS	INF_R_275_relapse4b

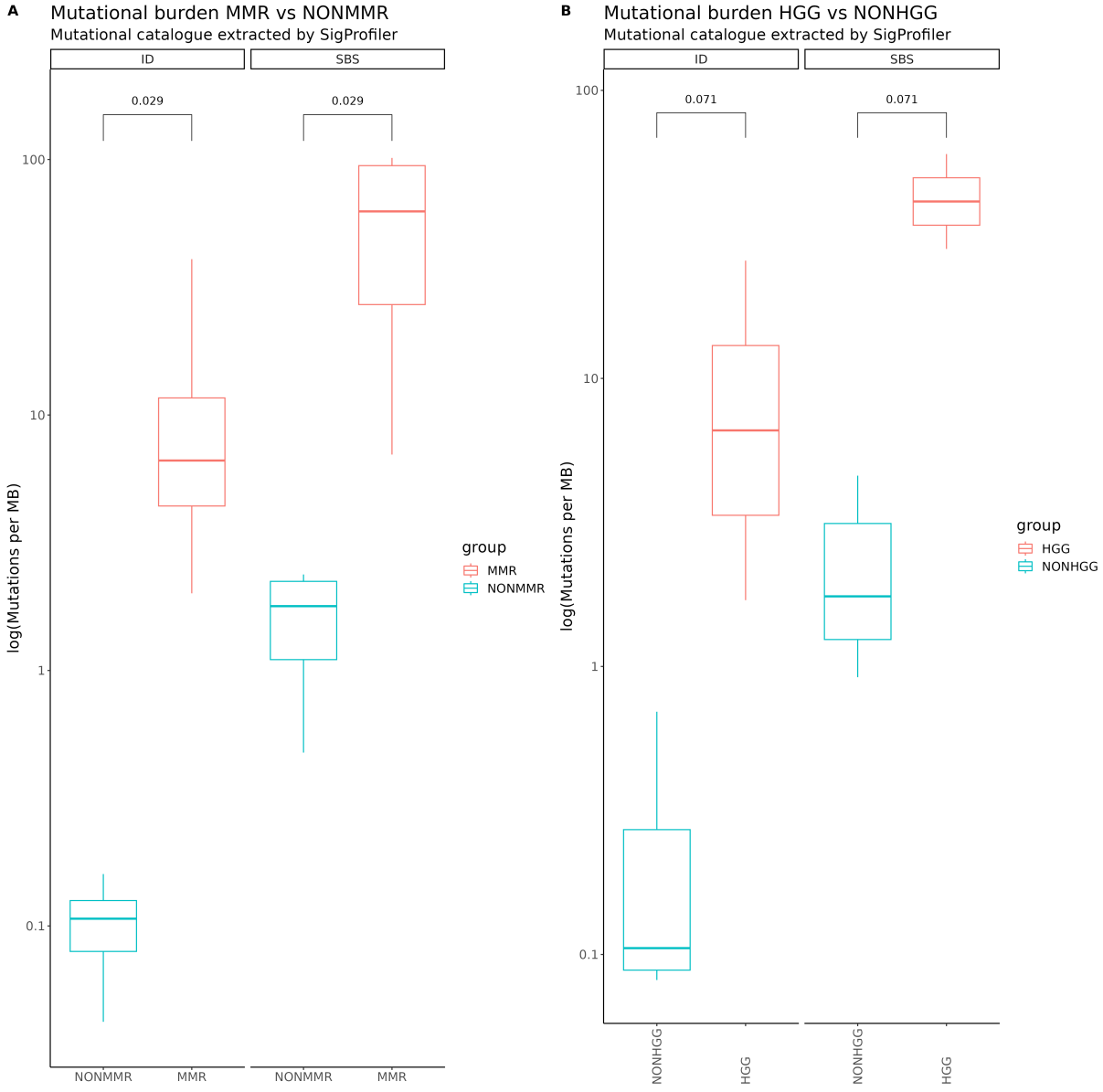


Figure 42: A) Comparison of mutational burden between MMR and non-MMR for both ID and SBS. B) Comparison of mutational burden between HGG and pedRTK1 HGG and other tumor types for both ID and SBS. All p-values calculated with the Wilcoxon test.

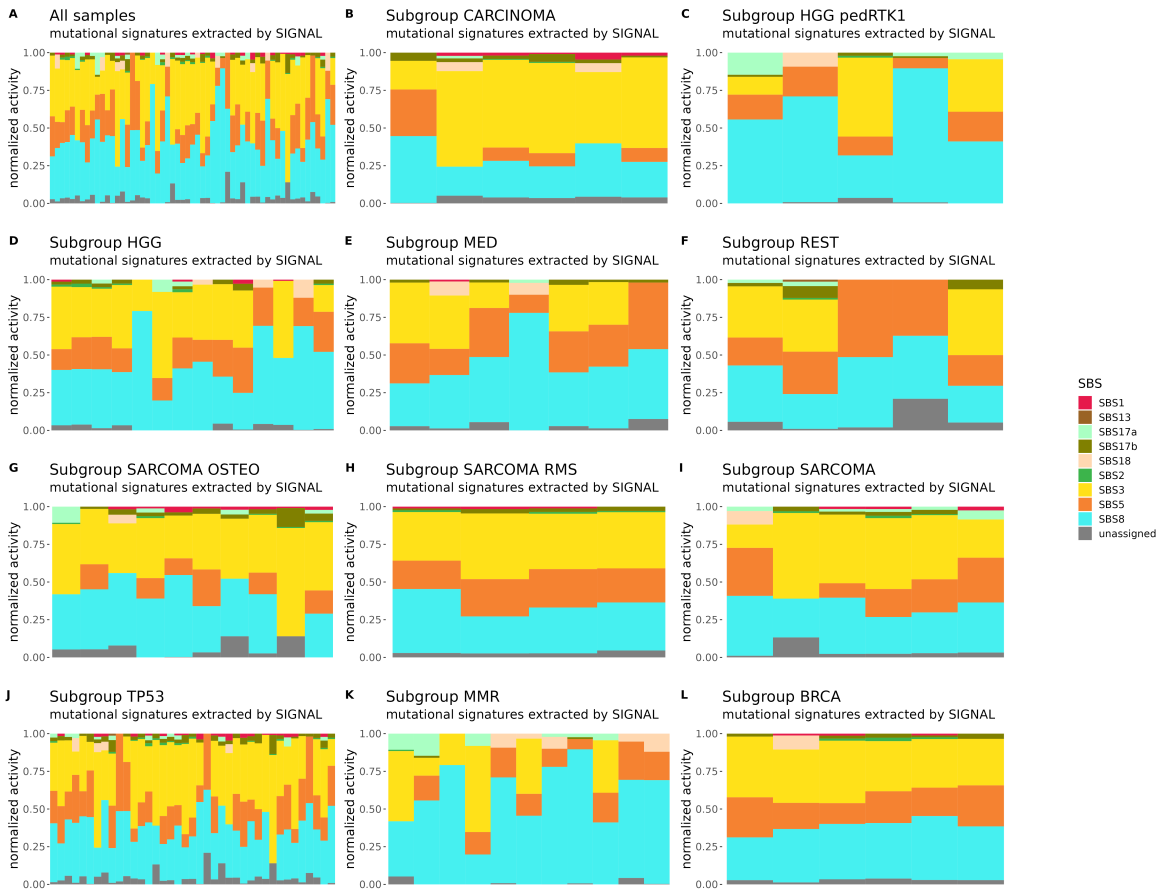


Figure 43: Mutational signatures assigned by SIGNAL.

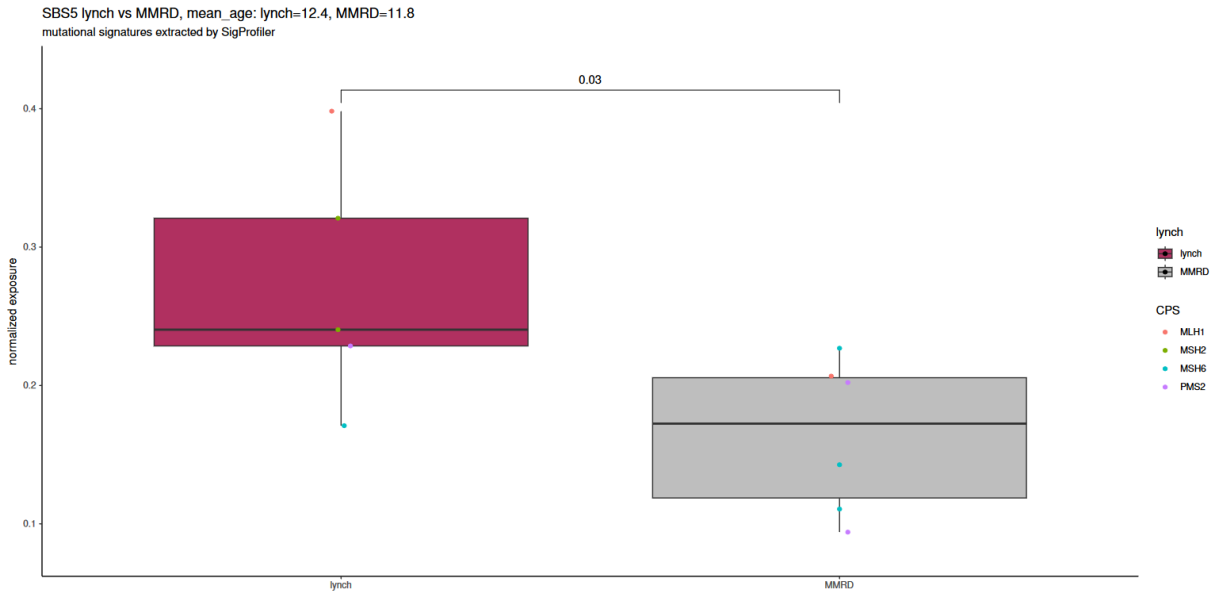


Figure 44: Differences between lynch syndrom (heterozygous) or cMMRD (homozygous) mutations inside MMR subgroup.

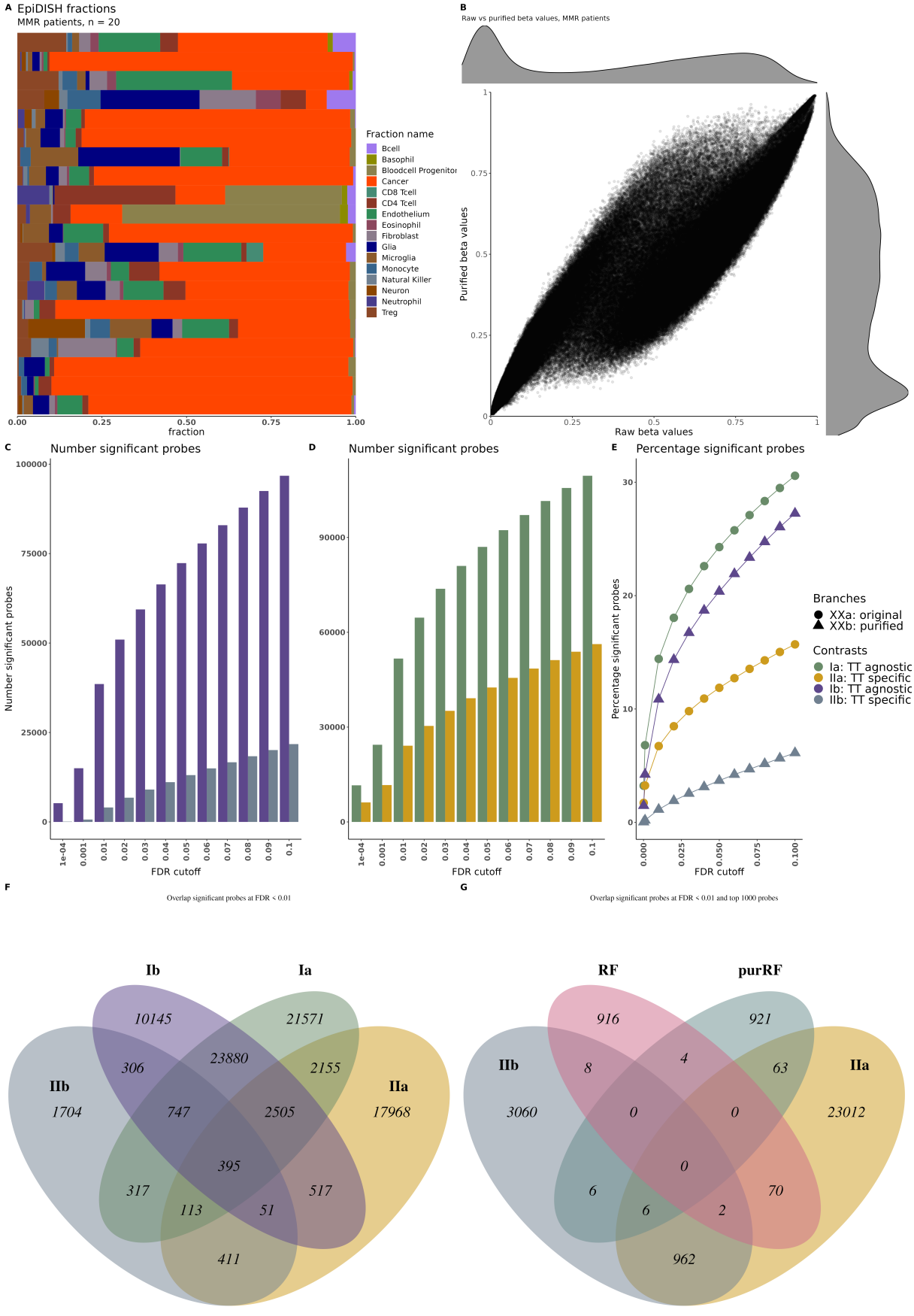


Figure 45: Overview of results from analysis of samples with MMR germline mutation. A) Estimation of present fractions in of different cell types in the tumor samples. B) Original beta values vs purified beta values based on the estimation of cell type fractions. C-D) Number and percentages of significant differentially methylated probes at different FDR cutoffs for contrasts Ia, Ib, IIa and IIb. F) Overlap of differentially enriched probes by the different contrasts at FDR < 0.01 G) Overlap of

differentially enriched probes from contrasts IIa and IIb at FDR < 0.01 with top 1000 most important probes identified by RF method.

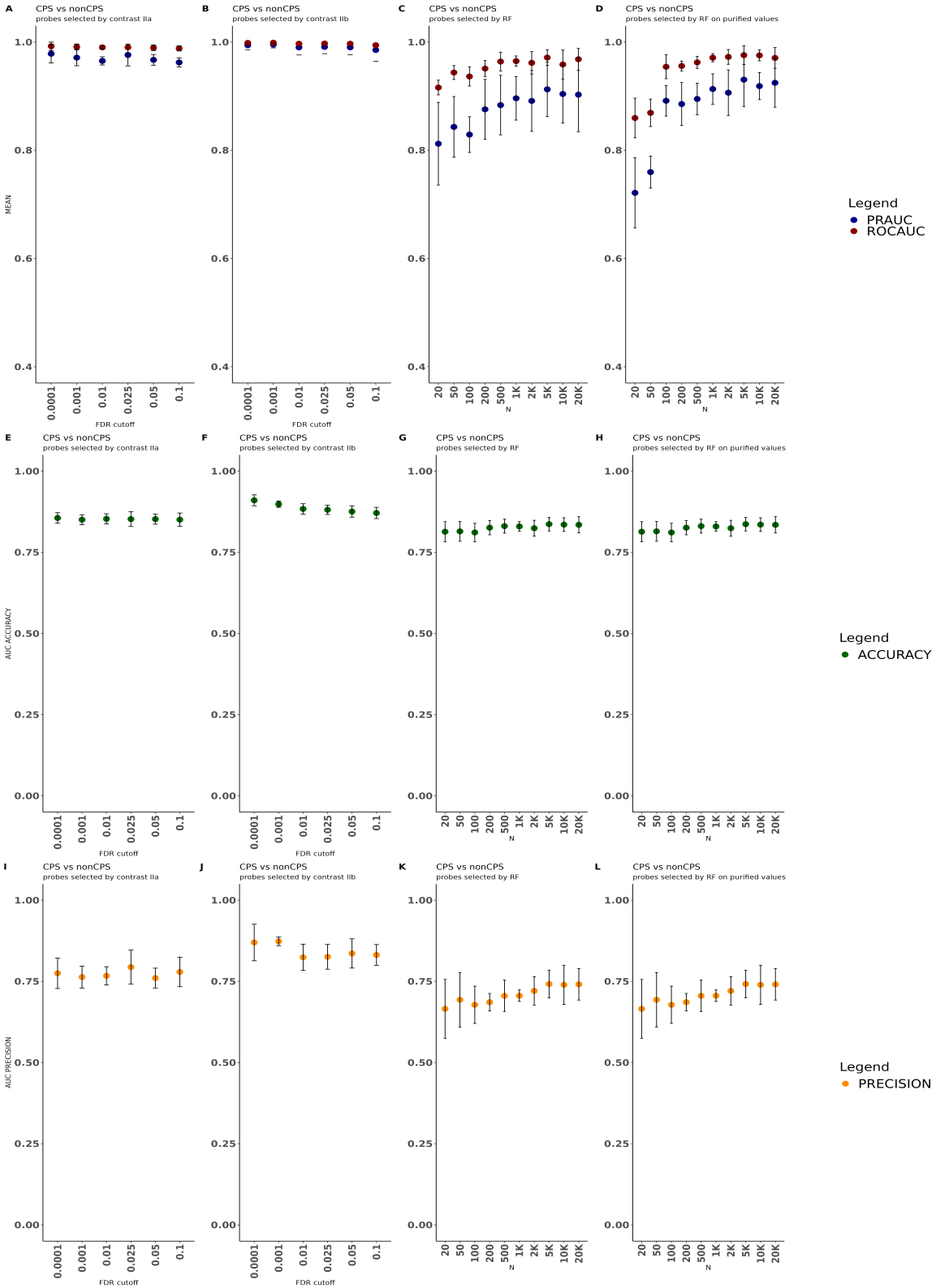


Figure 46: 3x cross validation for differentiation between MMR vs nonMMR germline mutated tumors using probes identified as differentially enriched by different methods.

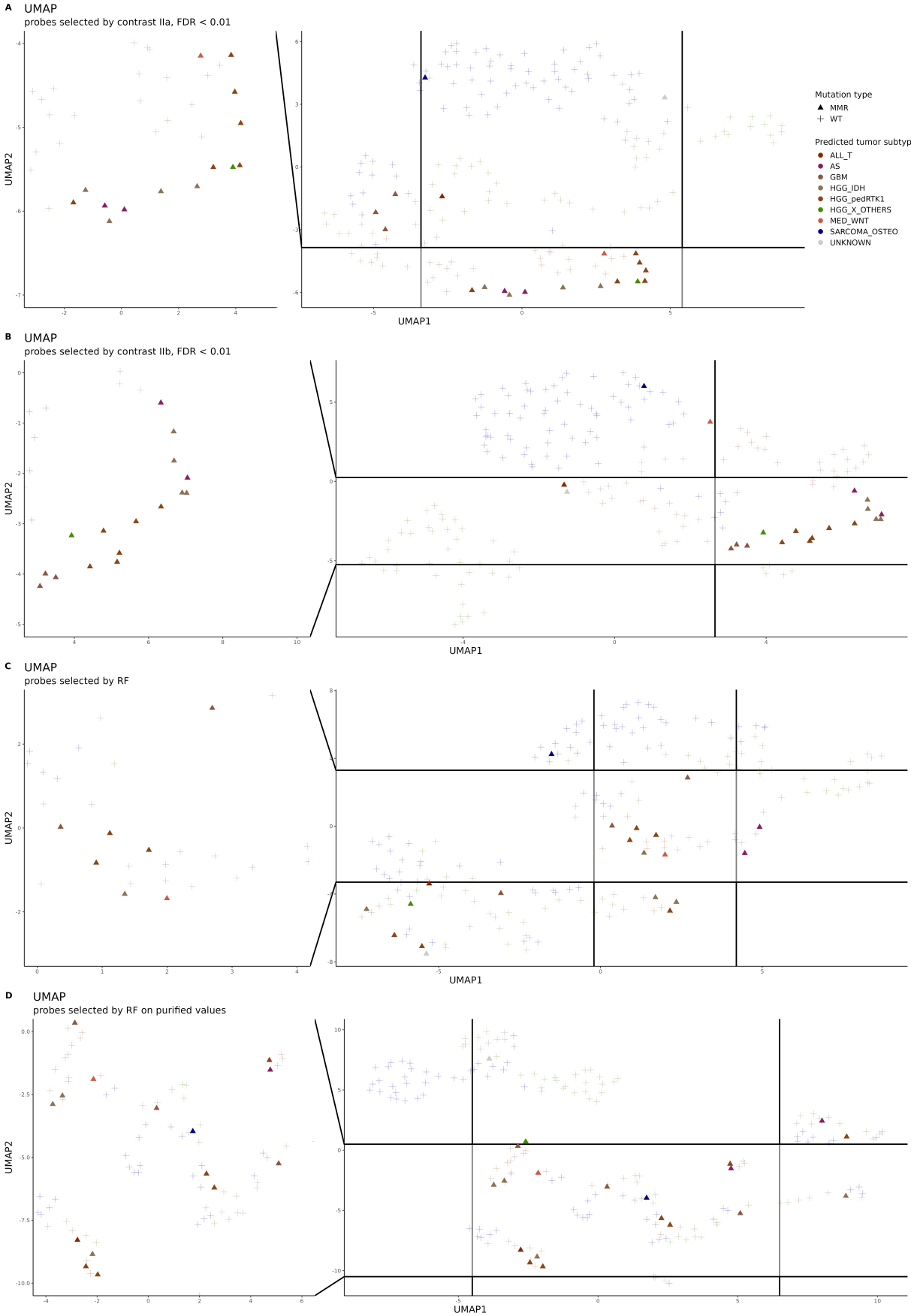


Figure 47: Umap plots using tumor samples with MMR germline mutation and tumor samples from control cohort. A) UMAP plot using only significantly differentially methylated probes identified by contrast IIa with cutoff FDR < 0.01. B) UMAP plot using only significantly differentially methylated probes identified by contrast IIb with cutoff FDR < 0.01. C) UMAP plot using only top 1000 probes ranked by permutation importance calculated with raw beta values. D) UMAP plot using only top 1000 probes ranked by permutation importance calculated with purified beta values.

Table 6: Pathway enrichment of network clusters identified by WGCNA and GRAPH. For probes identified by contrasts Ia and Ib pooled with the top 10% most variably methylated probes in the cohort. Per cluster top 10 pathways sorted by FDR listed. N = number of genes in pathway, DE = number of differentially methylated genes in pathway

GOID	N	DE	FDR	Cluster	TERM
GO:0000977	1118	93	1.14141E-06	black	RNA polymerase II transcription regulatory region sequence-specific DNA binding
GO:0000976	1276	96	1.42714E-06	black	transcription cis-regulatory region binding
GO:0001067	1277	96	1.42714E-06	black	transcription regulatory region nucleic acid binding
GO:0000785	1083	92	2.48706E-06	black	chromatin
GO:0001228	381	49	1.8301E-05	black	DNA-binding transcription activator activity, RNA polymerase II-specific
GO:0001216	388	49	2.01103E-05	black	DNA-binding transcription activator activity
GO:0007389	429	52	5.53019E-05	black	pattern specification process
GO:0000978	946	76	0.000233464	black	RNA polymerase II cis-regulatory region sequence-specific DNA binding
GO:0000987	981	77	0.000292115	black	cis-regulatory region sequence-specific DNA binding
GO:0045944	1123	81	0.000399313	black	positive regulation of transcription by RNA polymerase II
GO:0000977	1118	369	1.3509E-07	blue	RNA polymerase II transcription regulatory region sequence-specific DNA binding
GO:0000978	946	325	1.31876E-06	blue	RNA polymerase II cis-regulatory region sequence-specific DNA binding
GO:0000987	981	331	1.59527E-06	blue	cis-regulatory region sequence-specific DNA binding
GO:0000976	1276	386	1.59527E-06	blue	transcription cis-regulatory region binding
GO:0001067	1277	386	1.59527E-06	blue	transcription regulatory region nucleic acid binding
GO:0000785	1083	358	4.18812E-05	blue	chromatin
GO:0001216	388	164	4.88933E-05	blue	DNA-binding transcription activator activity
GO:0001228	381	162	5.9168E-05	blue	DNA-binding transcription activator activity, RNA polymerase II-specific
GO:0045165	272	121	0.0004734	blue	cell fate commitment
GO:0048568	386	165	0.000482029	blue	embryonic organ development
GO:0016604	550	230	0.000421903	brown	nuclear body
GO:0005524	996	410	0.002587673	brown	ATP binding
GO:0140640	611	159	0.006198366	brown	catalytic activity, acting on a nucleic acid
GO:0032559	1051	422	0.012433786	brown	adenyl ribonucleotide binding
GO:0030554	1126	446	0.012433786	brown	adenyl nucleotide binding
GO:0022613	310	101	0.031792372	brown	ribonucleoprotein complex biogenesis
GO:0005882	141	17	2.13394E-14	cyan	intermediate filament
GO:0045111	179	17	6.18161E-13	cyan	intermediate filament cytoskeleton
GO:0099513	565	17	9.32365E-05	cyan	polymeric cytoskeletal fiber
GO:0099512	858	17	0.006262936	cyan	supramolecular fiber
GO:0099081	865	17	0.006262936	cyan	supramolecular polymer
GO:0071543	2	2	0.047908991	darkred	diphosphoinositol polyphosphate metabolic process
GO:1901906	2	2	0.047908991	darkred	diadenosine pentaphosphate metabolic process
GO:1901907	2	2	0.047908991	darkred	diadenosine pentaphosphate catabolic process
GO:1901908	2	2	0.047908991	darkred	diadenosine hexaphosphate metabolic process
GO:1901909	2	2	0.047908991	darkred	diadenosine hexaphosphate catabolic process
GO:1901910	2	2	0.047908991	darkred	adenosine 5'-(hexahydrogen pentaphosphate) metabolic process
GO:1901911	2	2	0.047908991	darkred	adenosine 5'-(hexahydrogen pentaphosphate) catabolic process
GO:0000298	2	2	0.047908991	darkred	endopolyphosphatase activity
GO:0034431	2	2	0.047908991	darkred	bis(5'-adenosyl)-hexaphosphatase activity
GO:0034432	2	2	0.047908991	darkred	bis(5'-adenosyl)-pentaphosphatase activity
GO:0000977	1118	20	0.000581253	grey60	RNA polymerase II transcription regulatory region sequence-specific DNA binding
GO:0000976	1276	20	0.000581253	grey60	transcription cis-regulatory region binding
GO:0001067	1277	20	0.000581253	grey60	transcription regulatory region nucleic acid binding
GO:0000785	1083	19	0.001167419	grey60	chromatin
GO:0045944	1123	18	0.002020783	grey60	positive regulation of transcription by RNA polymerase II
GO:0000987	981	16	0.018397545	grey60	cis-regulatory region sequence-specific DNA binding
GO:0048704	88	13	4.85245E-11	midnightblue	embryonic skeletal system morphogenesis
GO:0009952	179	16	4.74616E-08	midnightblue	anterior/posterior pattern specification
GO:0048706	117	14	1.10189E-07	midnightblue	embryonic skeletal system development
GO:0048705	205	14	9.17678E-06	midnightblue	skeletal system morphogenesis
GO:0000987	981	24	9.27499E-06	midnightblue	cis-regulatory region sequence-specific DNA binding
GO:0000978	946	23	2.53401E-05	midnightblue	RNA polymerase II cis-regulatory region sequence-specific DNA binding

GO:0000977	1118	24	4.22981E-05	midnightblue	RNA polymerase II transcription regulatory region sequence-specific DNA binding
GO:0003002	388	16	5.26054E-05	midnightblue	regionalization
GO:0048562	249	14	6.33569E-05	midnightblue	embryonic organ morphogenesis
GO:0000976	1276	24	9.40471E-05	midnightblue	transcription cis-regulatory region binding
GO:0030527	29	9	9.31869E-14	orange	structural constituent of chromatin
GO:0000786	62	9	8.97923E-11	orange	nucleosome
GO:0044815	97	9	3.99955E-09	orange	DNA packaging complex
GO:0006334	63	7	1.24558E-08	orange	nucleosome assembly
GO:0034728	74	7	5.24295E-08	orange	nucleosome organization
GO:0032993	161	9	2.02545E-07	orange	protein-DNA complex
GO:0046982	202	9	8.37182E-07	orange	protein heterodimerization activity
GO:0032200	178	7	1.3033E-06	orange	telomere organization
GO:0065004	134	7	2.1468E-06	orange	protein-DNA complex assembly
GO:0071824	148	7	3.06379E-06	orange	protein-DNA complex subunit organization
GO:0007156	144	10	6.12012E-05	royalblue	homophilic cell adhesion via plasma membrane adhesion molecules
GO:0098742	288	11	0.000204127	royalblue	cell-cell adhesion via plasma-membrane adhesion molecules
GO:0048706	117	9	1.69768E-12	saddlebrown	embryonic skeletal system development
GO:0009952	179	9	4.29451E-11	saddlebrown	anterior/posterior pattern specification
GO:0001501	497	10	1.87714E-07	saddlebrown	skeletal system development
GO:0003002	388	9	1.44926E-06	saddlebrown	regionalization
GO:0048705	205	8	1.44926E-06	saddlebrown	skeletal system morphogenesis
GO:0048704	88	7	1.44926E-06	saddlebrown	embryonic skeletal system morphogenesis
GO:0007389	429	9	3.45022E-06	saddlebrown	pattern specification process
GO:0043009	514	9	7.24239E-06	saddlebrown	chordate embryonic development
GO:0009792	533	9	7.73028E-06	saddlebrown	embryo development ending in birth or egg hatching
GO:0000785	1083	10	4.45492E-05	saddlebrown	chromatin
GO:0007218	85	9	0.000568528	tan	neuropeptide signaling pathway
GO:0030594	149	9	0.000568528	tan	neurotransmitter receptor activity
GO:0004930	687	17	0.000568528	tan	G protein-coupled receptor activity
GO:0043005	1277	30	0.001276593	tan	neuron projection
GO:0007268	808	22	0.001331685	tan	chemical synaptic transmission
GO:0098916	808	22	0.001331685	tan	anterograde trans-synaptic signaling
GO:0099537	818	22	0.001379045	tan	trans-synaptic signaling
GO:0050877	1232	27	0.001572629	tan	nervous system process
GO:0099536	846	22	0.001671389	tan	synaptic signaling
GO:0007187	43	6	0.003323795	tan	G protein-coupled receptor signaling pathway, coupled to cyclic nucleotide second messenger
GO:0050911	200	183	1.64857E-50	turquoise	detection of chemical stimulus involved in sensory perception of smell
GO:0050907	221	194	6.92201E-48	turquoise	detection of chemical stimulus involved in sensory perception
GO:0009593	254	207	4.12063E-40	turquoise	detection of chemical stimulus
GO:0050906	277	222	1.79773E-39	turquoise	detection of stimulus involved in sensory perception
GO:0007608	264	192	2.34577E-34	turquoise	sensory perception of smell
GO:0004984	171	144	2.15531E-32	turquoise	olfactory receptor activity
GO:0007606	316	215	2.46578E-32	turquoise	sensory perception of chemical stimulus
GO:0051606	406	266	1.89392E-25	turquoise	detection of stimulus
GO:0004930	687	326	1.39455E-15	turquoise	G protein-coupled receptor activity
GO:0007600	696	370	8.22573E-14	turquoise	sensory perception
GO:0097447	520	155	0.001209259	yellow	dendritic tree
GO:0030425	517	154	0.001209259	yellow	dendrite
GO:0036477	747	201	0.011451819	yellow	somatodendritic compartment
GO:0015631	315	83	0.025860454	yellow	tubulin binding
GO:0050907	187	120	2.30336E-47	ME3	detection of chemical stimulus involved in sensory perception
GO:0050911	172	114	5.05609E-47	ME3	detection of chemical stimulus involved in sensory perception of smell
GO:0050906	235	135	7.41253E-45	ME3	detection of stimulus involved in sensory perception
GO:0009593	216	125	1.26503E-42	ME3	detection of chemical stimulus
GO:0007608	229	119	7.53271E-42	ME3	sensory perception of smell
GO:0007606	268	129	8.60949E-41	ME3	sensory perception of chemical stimulus
GO:0004984	152	91	8.47609E-34	ME3	olfactory receptor activity
GO:0051606	343	149	5.17034E-33	ME3	detection of stimulus
GO:0007600	586	205	9.46384E-32	ME3	sensory perception
GO:0050877	1042	302	6.60779E-31	ME3	nervous system process

Table 7: Pathway enrichment of network clusters identified by WGCNA and GRAPH. For probes identified by contrasts IIa and IIb pooled with the top 10% most variably methylated probes in the cohort. Per cluster top 10 pathways sorted by FDR listed. N = number of genes in pathway, DE = number of differentially methylated genes in pathway

GOID	N	DE	FDR	from	TERM
GO:0000977	1293	273	3.09956E-33	brown	RNA polymerase II transcription regulatory region sequence-specific DNA binding
GO:0000976	1488	280	2.24557E-29	brown	transcription cis-regulatory region binding
GO:0001067	1490	280	2.2464E-29	brown	transcription regulatory region nucleic acid binding
GO:0000987	1127	241	3.1417E-27	brown	cis-regulatory region sequence-specific DNA binding
GO:0000978	1087	236	3.35569E-27	brown	RNA polymerase II cis-regulatory region sequence-specific DNA binding
GO:0007389	489	148	9.92853E-24	brown	pattern specification process
GO:0003002	444	134	9.77876E-22	brown	regionalization
GO:0000785	1399	257	2.40328E-21	brown	chromatin
GO:0009887	1107	238	2.9571E-21	brown	animal organ morphogenesis
GO:0001216	442	125	1.09848E-17	brown	DNA-binding transcription activator activity
GO:0140535	1338	395	3.16206E-17	green	intracellular protein-containing complex
GO:1990234	1164	362	1.68174E-16	green	transferase complex
GO:0016604	747	331	2.77071E-12	green	nuclear body
GO:0098687	428	182	2.72524E-11	green	chromosomal region
GO:0006886	870	336	4.74508E-11	green	intracellular protein transport
GO:0006974	1213	357	6.04111E-11	green	cellular response to DNA damage stimulus
GO:0051276	778	265	6.04111E-11	green	chromosome organization
GO:0022613	501	191	3.01034E-10	green	ribonucleoprotein complex biogenesis
GO:0016071	1057	301	4.19337E-10	green	mRNA metabolic process
GO:0000278	1074	373	4.52586E-10	green	mitotic cell cycle
GO:0048562	292	28	0.001222693	greenyellow	embryonic organ morphogenesis
GO:0009887	1107	55	0.001222693	greenyellow	animal organ morphogenesis
GO:0048598	626	41	0.001222693	greenyellow	embryonic morphogenesis
GO:0048568	450	33	0.00149977	greenyellow	embryonic organ development
GO:0000977	1293	53	0.004179328	greenyellow	RNA polymerase II transcription regulatory region sequence-specific DNA binding
GO:0000976	1488	56	0.004179328	greenyellow	transcription cis-regulatory region binding
GO:0001067	1490	56	0.004179328	greenyellow	transcription regulatory region nucleic acid binding
GO:0048705	224	22	0.004179328	greenyellow	skeletal system morphogenesis
GO:0048736	186	19	0.013665574	greenyellow	appendage development
GO:0060173	186	19	0.013665574	greenyellow	limb development
GO:0050907	202	51	1.32472E-41	lightcyan	detection of chemical stimulus involved in sensory perception
GO:0050911	175	48	2.29657E-40	lightcyan	detection of chemical stimulus involved in sensory perception of smell
GO:0009593	238	52	6.72096E-40	lightcyan	detection of chemical stimulus
GO:0007608	239	49	6.34173E-39	lightcyan	sensory perception of smell
GO:0007606	296	52	1.28237E-38	lightcyan	sensory perception of chemical stimulus
GO:0050906	261	53	1.47016E-38	lightcyan	detection of stimulus involved in sensory perception
GO:0051606	400	57	6.23922E-35	lightcyan	detection of stimulus
GO:0004930	665	57	3.30786E-28	lightcyan	G protein-coupled receptor activity
GO:0007600	710	63	6.99199E-28	lightcyan	sensory perception
GO:0004984	148	34	3.19732E-26	lightcyan	olfactory receptor activity
GO:0005201	234	12	0.049532568	lightgreen	extracellular matrix structural constituent

GO:0031012	677	21	0.049532568	lightgreen	extracellular matrix
GO:0030312	678	21	0.049532568	lightgreen	external encapsulating structure
GO:0070647	1332	135	0.011651134	pink	protein modification by small protein conjugation or removal
GO:0032446	1103	123	0.011651134	pink	protein modification by small protein conjugation
GO:0022402	1542	85	0.000915321	salmon	cell cycle process
GO:0140097	359	24	0.013436774	salmon	catalytic activity, acting on DNA
GO:0010564	781	52	0.029522203	salmon	regulation of cell cycle process
GO:0006974	1213	57	0.029522203	salmon	cellular response to DNA damage stimulus
GO:0000278	1074	62	0.029522203	salmon	mitotic cell cycle
GO:0051726	1174	69	0.029522203	salmon	regulation of cell cycle
GO:0005819	426	32	0.029522203	salmon	spindle
GO:0006281	819	40	0.029522203	salmon	DNA repair
GO:0006289	96	13	0.029522203	salmon	nucleotide-excision repair
GO:0018205	461	31	0.029522203	salmon	peptidyl-lysine modification
GO:0050911	175	123	6.16367E-27	turquoise	detection of chemical stimulus involved in sensory perception of smell
GO:0050907	202	133	5.33821E-26	turquoise	detection of chemical stimulus involved in sensory perception
GO:0050906	261	162	6.09419E-24	turquoise	detection of stimulus involved in sensory perception
GO:0009593	238	147	1.11915E-23	turquoise	detection of chemical stimulus
GO:0004984	148	102	5.41335E-23	turquoise	olfactory receptor activity
GO:0007608	239	130	9.90816E-21	turquoise	sensory perception of smell
GO:0007606	296	151	1.89993E-20	turquoise	sensory perception of chemical stimulus
GO:0051606	400	209	9.47072E-20	turquoise	detection of stimulus
GO:0004930	665	264	5.69524E-17	turquoise	G protein-coupled receptor activity
GO:0050877	1289	525	1.32526E-15	turquoise	nervous system process
GO:0030527	47	9	1.92267E-11	violet	structural constituent of chromatin
GO:0000786	100	9	1.22827E-08	violet	nucleosome
GO:0032993	229	9	2.96842E-07	violet	protein-DNA complex
GO:0006334	93	7	2.96842E-07	violet	nucleosome assembly
GO:0044815	157	9	2.96842E-07	violet	DNA packaging complex
GO:0034728	112	7	1.10234E-06	violet	nucleosome organization
GO:0046982	241	9	7.30536E-06	violet	protein heterodimerization activity
GO:0032200	245	7	2.84863E-05	violet	telomere organization
GO:0065004	205	7	3.78651E-05	violet	protein-DNA complex assembly
GO:0071824	228	7	6.33017E-05	violet	protein-DNA complex subunit organization
GO:0045321	960	271	2.81228E-18	yellow	leukocyte activation
GO:0001775	1132	304	4.32102E-18	yellow	cell activation
GO:0046649	777	225	2.40304E-15	yellow	lymphocyte activation
GO:0042110	544	170	2.3904E-13	yellow	T cell activation
GO:0098609	1045	282	1.49713E-12	yellow	cell-cell adhesion
GO:0002684	1122	253	4.95342E-12	yellow	positive regulation of immune system process
GO:0050865	627	185	5.5599E-12	yellow	regulation of cell activation
GO:0002694	568	171	8.87431E-12	yellow	regulation of leukocyte activation
GO:0007159	401	132	9.13421E-12	yellow	leukocyte cell-cell adhesion
GO:0032101	1148	269	9.13421E-12	yellow	regulation of response to external stimulus
GO:0050911	138	123	2.59902E-59	ME2	detection of chemical stimulus involved in sensory perception of smell
GO:0050907	158	129	7.13992E-54	ME2	detection of chemical stimulus involved in sensory perception
GO:0009593	189	140	2.55732E-49	ME2	detection of chemical stimulus
GO:0007608	193	130	9.66444E-47	ME2	sensory perception of smell
GO:0050906	209	147	2.99867E-43	ME2	detection of stimulus involved in sensory perception
GO:0004984	120	99	7.88861E-43	ME2	olfactory receptor activity
GO:0007606	240	142	7.67864E-41	ME2	sensory perception of chemical stimulus
GO:0051606	330	181	9.48045E-35	ME2	detection of stimulus
GO:0007600	602	264	5.68443E-28	ME2	sensory perception
GO:0050877	1119	419	1.74165E-27	ME2	nervous system process
GO:0048706	106	4	0.002232034	ME4	embryonic skeletal system development

GO:0009952	173	4	0.002232034	ME4	anterior/posterior pattern specification
GO:0003002	380	4	0.011270886	ME4	regionalization
GO:0007389	419	4	0.013903905	ME4	pattern specification process
GO:0001501	485	4	0.026877435	ME4	skeletal system development
GO:0043009	553	4	0.03244181	ME4	chordate embryonic development
GO:0009792	574	4	0.03244181	ME4	embryo development ending in birth or egg hatching
GO:0048704	79	3	0.036578664	ME4	embryonic skeletal system morphogenesis

Table 8: Pathway enrichment of network clusters identified by WGCNA and GRAPH. For probes identified by random forest applied to raw and purified methylation values pooled with the top 10% most variably methylated probes in the cohort. Per cluster top 10 pathways sorted by FDR listed. N = number of genes in pathway, DE = number of differentially methylated genes in pathway

GOID	N	DE	FDR	from	TERM
GO:0045296	175	17	0.024521268	black	cadherin binding
GO:0000977	860	310	1.14391E-07	blue	RNA polymerase II transcription regulatory region sequence-specific DNA binding
GO:0000987	746	282	1.53214E-07	blue	cis-regulatory region sequence-specific DNA binding
GO:0000978	720	276	1.53214E-07	blue	RNA polymerase II cis-regulatory region sequence-specific DNA binding
GO:0000976	978	324	6.28136E-07	blue	transcription cis-regulatory region binding
GO:0001067	979	324	6.47739E-07	blue	transcription regulatory region nucleic acid binding
GO:0000785	796	299	6.67606E-06	blue	chromatin
GO:0001216	316	140	0.000240684	blue	DNA-binding transcription activator activity
GO:0007156	127	72	0.000264759	blue	homophilic cell adhesion via plasma membrane adhesion molecules
GO:0001228	314	138	0.000380824	blue	DNA-binding transcription activator activity, RNA polymerase II-specific
GO:0009887	846	294	0.001394889	blue	animal organ morphogenesis
GO:0045321	704	154	0.011872905	brown	leukocyte activation
GO:0001775	838	174	0.022218155	brown	cell activation
GO:0046649	573	128	0.035868894	brown	lymphocyte activation
GO:0098742	259	11	0.01305654	cyan	cell-cell adhesion via plasma-membrane adhesion molecules
GO:0007156	127	9	0.01305654	cyan	homophilic cell adhesion via plasma membrane adhesion molecules
GO:0007218	65	11	0.000519596	magenta	neuropeptide signaling pathway
GO:0004930	565	23	0.001899352	magenta	G protein-coupled receptor activity
GO:0030594	129	11	0.001899352	magenta	neurotransmitter receptor activity
GO:0008528	130	11	0.002668911	magenta	G protein-coupled peptide receptor activity
GO:0001653	134	11	0.002668911	magenta	peptide receptor activity
GO:0008188	43	7	0.004077594	magenta	neuropeptide receptor activity
GO:0007187	36	7	0.015442893	magenta	G protein-coupled receptor signaling pathway, coupled to cyclic nucleotide second messenger
GO:0031045	30	6	0.015442893	magenta	dense core granule
GO:0042923	16	5	0.015442893	magenta	neuropeptide binding
GO:0007186	940	28	0.015442893	magenta	G protein-coupled receptor signaling pathway
GO:0030527	25	7	2.51689E-06	midnightblue	structural constituent of chromatin
GO:0000786	51	7	5.06484E-06	midnightblue	nucleosome
GO:0044815	77	7	1.82541E-05	midnightblue	DNA packaging complex
GO:0032200	112	6	0.000380069	midnightblue	telomere organization
GO:0032993	114	7	0.00052576	midnightblue	protein-DNA complex
GO:0006334	49	5	0.002423917	midnightblue	nucleosome assembly
GO:0034728	53	5	0.002822227	midnightblue	nucleosome organization
GO:0045638	64	6	0.005811888	midnightblue	negative regulation of myeloid cell differentiation
GO:0046982	148	7	0.007647995	midnightblue	protein heterodimerization activity
GO:0061644	6	3	0.010726579	midnightblue	protein localization to CENP-A containing chromatin
GO:0000977	860	88	5.0081E-06	red	RNA polymerase II transcription regulatory region sequence-specific DNA binding
GO:0000976	978	92	5.0081E-06	red	transcription cis-regulatory region binding
GO:0001067	979	92	5.0081E-06	red	transcription regulatory region nucleic acid binding
GO:0000785	796	88	5.70992E-06	red	chromatin
GO:0051253	843	83	7.4743E-06	red	negative regulation of RNA metabolic process
GO:0000978	720	76	9.56223E-05	red	RNA polymerase II cis-regulatory region sequence-specific DNA binding
GO:0000987	746	77	9.56223E-05	red	cis-regulatory region sequence-specific DNA binding
GO:0045892	776	75	9.56223E-05	red	negative regulation of DNA-templated transcription

GO:1903507	778	75	9.56223E-05	red	negative regulation of nucleic acid-templated transcription
GO:1902679	784	75	9.60598E-05	red	negative regulation of RNA biosynthetic process
GO:0050911	173	156	7.59683E-33	turquoise	detection of chemical stimulus involved in sensory perception of smell
GO:0050907	194	166	9.0795E-30	turquoise	detection of chemical stimulus involved in sensory perception
GO:0009593	222	180	5.96689E-27	turquoise	detection of chemical stimulus
GO:0050906	240	192	3.38374E-26	turquoise	detection of stimulus involved in sensory perception
GO:0004984	145	121	2.07678E-20	turquoise	olfactory receptor activity
GO:0007608	231	163	2.51275E-20	turquoise	sensory perception of smell
GO:0007606	276	183	5.41772E-19	turquoise	sensory perception of chemical stimulus
GO:0051606	347	224	2.98917E-14	turquoise	detection of stimulus
GO:0005549	59	55	5.14487E-12	turquoise	odorant binding
GO:0007600	583	318	1.52904E-08	turquoise	sensory perception
GO:0050911	129	118	0.000290613	ME1	detection of chemical stimulus involved in sensory perception of smell
GO:0050907	139	123	0.007091223	ME1	detection of chemical stimulus involved in sensory perception
GO:0050906	166	143	0.024045676	ME1	detection of stimulus involved in sensory perception
GO:0031424	22	10	2.31554E-16	ME11	keratinization
GO:0030216	59	10	1.05929E-11	ME11	keratinocyte differentiation
GO:0009913	85	10	3.37099E-10	ME11	epidermal cell differentiation
GO:0008544	137	11	7.13674E-10	ME11	epidermis development
GO:0001533	24	7	2.40726E-09	ME11	cornified envelope
GO:0043588	112	10	2.91633E-09	ME11	skin development
GO:0030855	265	10	3.18151E-06	ME11	epithelial cell differentiation
GO:0060429	504	10	0.001071013	ME11	epithelium development
GO:0018149	19	3	0.049991115	ME11	peptide cross-linking
GO:0016442	55	7	0.008974025	ME13	RISC complex
GO:0031332	55	7	0.008974025	ME13	RNAi effector complex
GO:0035195	118	7	0.030726048	ME13	miRNA-mediated gene silencing
GO:0035194	125	7	0.030726048	ME13	RNA-mediated post-transcriptional gene silencing
GO:0016441	128	7	0.030726048	ME13	post-transcriptional gene silencing
GO:0031047	155	7	0.036941431	ME13	RNA-mediated gene silencing
GO:0048706	66	4	0.022649717	ME16	embryonic skeletal system development
GO:0009952	92	4	0.022649717	ME16	anterior/posterior pattern specification
GO:0048706	66	8	2.85679E-10	ME4	embryonic skeletal system development
GO:0009952	92	8	3.28292E-07	ME4	anterior/posterior pattern specification
GO:0001501	277	9	1.03505E-06	ME4	skeletal system development
GO:0003002	190	8	9.65161E-06	ME4	regionalization
GO:0007389	210	8	1.50251E-05	ME4	pattern specification process
GO:0048705	117	7	1.50251E-05	ME4	skeletal system morphogenesis
GO:0043009	248	8	2.18489E-05	ME4	chordate embryonic development
GO:0048704	49	6	2.26588E-05	ME4	embryonic skeletal system morphogenesis
GO:0009792	258	8	2.38691E-05	ME4	embryo development ending in birth or egg hatching
GO:0000785	492	9	3.85579E-05	ME4	chromatin
GO:0000977	545	26	5.12091E-06	ME5	RNA polymerase II transcription regulatory region sequence-specific DNA binding
GO:0000976	621	26	5.39002E-06	ME5	transcription cis-regulatory region binding
GO:0001067	622	26	5.39002E-06	ME5	transcription regulatory region nucleic acid binding
GO:0000785	492	23	9.80468E-05	ME5	chromatin
GO:0009790	472	21	0.000499401	ME5	embryo development
GO:0043009	248	16	0.000739359	ME5	chordate embryonic development
GO:0007389	210	15	0.000739359	ME5	pattern specification process
GO:0009792	258	16	0.001049634	ME5	embryo development ending in birth or egg hatching
GO:0009952	92	11	0.001482511	ME5	anterior/posterior pattern specification
GO:0045944	567	20	0.002578197	ME5	positive regulation of transcription by RNA polymerase II

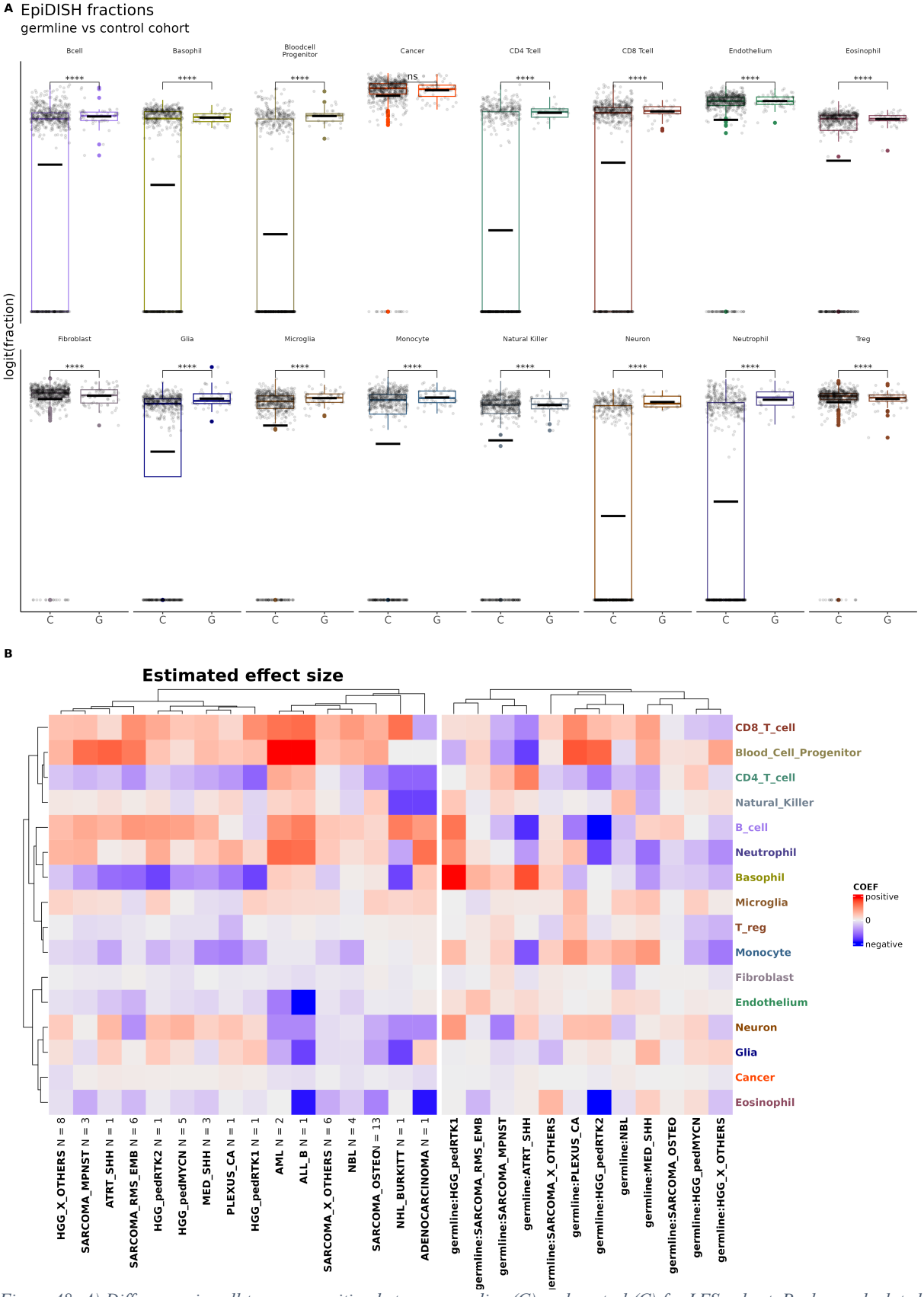


Figure 48: A) Differences in cell type composition between germline (G) and control (C) for LFS cohort. Pvalues calculated while accounting for tumor type. Bold black line = mean B) Estimation of effect size of tumor type and other on cell type composition by linear model with logit transformed cell type fraction as response variable.

Table 9: Germline mutated tumor samples of the TP53 and MMR cohort used for methylation analysis.

INFORM_PUBLIC_PID	SENTRIXID	CPS	TUMOR TYPE
INF_R_1121_primary	203723180078_R02C01	MMR	MED_WNT
INF_R_1164_primary	203726680028_R01C01	MMR	HGG_pedRTK1
INF_R_1180_relapse1	203946830162_R05C01	MMR	HGG_pedRTK1
INF_R_1246_relapse2	203949840087_R04C01	MMR	HGG_X_OTHERS
INF_R_1622_relapse1	204391650168_R02C01	MMR	HGG_IDH
INF_R_1733_primary	204792770073_R01C01	MMR	HGG_IDH
INF_R_1899_primary	205023290135_R06C01	MMR	HGG_pedRTK1
INF_R_305_progression	201332340037_R05C01	MMR	HGG_pedRTK1
INF_R_686_relapse3	202259350090_R07C01	MMR	HGG_IDH
INF_R_991_relapse2	203430580030_R05C01	MMR	SARCOMA_OSTEO
INF_R_1320_relapse1	203985930014_R02C01	MMR	HGG_IDH
INF_R_1575_relapse1	204391650033_R07C01	MMR	ALL_T
INF_R_1078_progression	203537580046_R07C01	MMR	HGG_pedRTK1
INF_R_756_relapse1	203033880113_R08C01	MMR	HGG_pedRTK1
INF_R_1472_relapse1	203990170064_R02C01	MMR	UNKNOWN
140F46SD	205799790171_R01C01	MMR	AS
140F42S1D	205814880070_R08C01	MMR	GBM
140F42S2D	205799790171_R03C01	MMR	GBM
140F45SD	205814880020_R02C01	MMR	GBM
140F57SD	205814880020_R03C01	MMR	AS
INF_R_050_relapse2	3998523055_R05C02	TP53	SARCOMA_OSTEO
INF_R_1116_relapse4	203717910104_R02C01	TP53	ACC
INF_R_1159_relapse1	203726680028_R08C01	TP53	MED_SHH
INF_R_1273_relapse2	203949840122_R02C01	TP53	SARCOMA_X_OTHERS
INF_R_1281_relapse1	203960200125_R04C01	TP53	HGG_X_OTHERS
INF_R_1357_relapse1	203986500025_R05C01	TP53	HGG_pedRTK2
INF_R_1397_relapse1	203986510105_R06C01	TP53	SARCOMA_RMS_EMB
INF_R_1407_relapse1	203989100011_R05C01	TP53	ACC
INF_R_1412_relapse1	203989100025_R03C01	TP53	SARCOMA_X_OTHERS
INF_R_1676_progression	204391650167_R05C01	TP53	AML
INF_R_168_relapse1	200397860036_R06C02	TP53	ADENOCARCINOMA
INF_R_1681_relapse1	204379160024_R05C01	TP53	HGG_X_OTHERS
INF_R_1691_primary	204390590044_R05C01	TP53	SARCOMA_RMS_EMB
INF_R_177_relapse1	200397860074_R01C02	TP53	SARCOMA_OSTEO
INF_R_1854_progression	205023290128_R01C01	TP53	ACC
INF_R_273_primary	201194000072_R01C01	TP53	HGG_X_OTHERS
INF_R_307_relapse1	201364900076_R05C01	TP53	ATRT_SHH
INF_R_354_relapse1	201465970003_R03C01	TP53	ACC
INF_R_401_progression	201530470018_R08C01	TP53	MED_SHH
INF_R_431_primary	201533480032_R04C01	TP53	HGG_pedRTK1
INF_R_473_relapse1	201869690154_R05C01	TP53	SARCOMA_OSTEO
INF_R_769_relapse1	203034110090_R02C01	TP53	SARCOMA_X_OTHERS
INF_R_790_relapse1	203049640025_R08C01	TP53	ACC
INF_R_914_primary	203197470212_R02C01	TP53	SARCOMA_X_OTHERS
INF_R_924_relapse3	203197470212_R04C01	TP53	PLEXUS_CA
INF_R_994_progression	203723190083_R02C01	TP53	SARCOMA_OSTEO
INF_R_743_primary	202292320097_R07C01	TP53	NBL
INF_R_1079_primary	203537580023_R08C01	TP53	SARCOMA_OSTEO
INF_R_1023_relapse1	203504440004_R07C01	TP53	SARCOMA_OSTEO
INF_R_1030_relapse3	203537580023_R03C01	TP53	SARCOMA_OSTEO
INF_R_823_relapse1	203049640066_R03C01	TP53	SARCOMA_RMS_EMB
INF_R_993_relapse2	203430580030_R04C01	TP53	NHL_BURKITT
INF_R_265_progression	201247480007_R02C01	TP53	SARCOMA_X_OTHERS
INF_R_1879_primary	205049780038_R08C01	TP53	SARCOMA_OSTEO
INF_R_1123_primary	203723180077_R08C01	TP53	HGG_pedMYCN
INF_R_852_relapse7	203949840163_R07C01	TP53	ACC
INF_R_1677_relapse1	204391650167_R06C01	TP53	AML
INF_R_130_relapse7	201869690154_R03C01	TP53	SARCOMA_OSTEO
INF_R_117_relapse1	200325530125_R04C01	TP53	NBL
INF_R_1662_relapse5	204391650167_R02C01	TP53	NBL
INF_R_1904_relapse1	205049780056_R03C01	TP53	HGG_X_OTHERS
INF_R_1908_relapse1	205049780056_R01C01	TP53	HGG_X_OTHERS
INF_R_1962_relapse2	205049780056_R07C01	TP53	NBL

INF_R_1998_primary	205055470084_R06C01	TP53	SARCOMA_MPNST
INF_R_2031_relapse1	205059630016_R03C01	TP53	HGG_X_OTHERS
INF_R_2050_relapse	205058490051_R01C01	TP53	HGG_pedMYCN
INF_R_2091_relapse1	205555380045_R05C01	TP53	SARCOMA_OSTEO
INF_R_2103_primary	205555380180_R07C01	TP53	HGG_pedMYCN
INF_R_2203_relapse2	205624630031_R01C01	TP53	SARCOMA_OSTEO
INF_R_2228_relapse2	205624630173_R08C01	TP53	ALL_B
INF_R_2301_primary	205799790061_R01C01	TP53	HGG_pedMYCN
INF_R_2345_primary	205799790159_R06C01	TP53	SARCOMA_MPNST
INF_R_2374_relapse3	205854140020_R07C01	TP53	SARCOMA_MPNST
INF_R_2418_primary	206129780103_R01C01	TP53	HGG_pedMYCN
INF_R_2560_relapse1	206462450069_R01C01	TP53	SARCOMA_RMS_EMB
INF_R_2564_primary	206466470167_R08C01	TP53	SARCOMA_RMS_EMB
INF_R_2570_relapse1	206466470172_R01C01	TP53	MED_SHH
INF_R_2601_relapse1	206467010109_R06C01	TP53	SARCOMA_X_OTHERS
INF_R_2687_relapse1	207131890002_R04C01	TP53	HGG_X_OTHERS
INF_R_2717_primary	207131890001_R06C01	TP53	HGG_X_OTHERS
INF_R_2728_relapse1	207131890002_R06C01	TP53	SARCOMA_OSTEO
INF_R_2252_relapse4	207127950039_R07C01	TP53	SARCOMA_RMS_EMB

Table 10: Samples used for the mutational signature analysis.

SUBGROUP	**INFORM_PUBLIC_PID**
NBL	INF_R_743_primary
SARCOMA_OSTEO	INF_R_050_relapse2
HGG_X-OTHERS	INF_R_1281_relapse1
PLEXUS_CA	INF_R_924_relapse3
HGG_X-OTHERS	INF_R_1733_primary
MED_SHH	INF_R_1702_primary
SARCOMA_OSTEO	INF_R_1079_primary
SARCOMA_OSTEO	INF_R_991_relapse2
SARCOMA_X-OTHERS	INF_R_1031_relapse1
SARCOMA_OSTEO	INF_R_1023_relapse1
MED_WNT	INF_R_1121_primary
HGG_X-OTHERS	INF_R_1622_relapse1
ACC	INF_R_1116_relapse4
SARCOMA_X-OTHERS	INF_R_1412_relapse1
ADENOCARCINOMA	INF_R_168_relapse1
ACC	INF_R_790_relapse1
MED_SHH	INF_R_401_progression
MED	INF_R_742_relapse2
SARCOMA_OSTEO	INF_R_1030_relapse3
HGG_X-OTHERS	INF_R_1681_relapse1
HGG_pedRTK1	INF_R_1164_primary
HGG_pedRTK1	INF_R_305_progression
ACC	INF_R_1407_relapse1
MED_SHH	INF_R_1159_relapse1
DIPG	INF_R_574_primary
SARCOMA_RMS_EMB	INF_R_823_relapse1
HGG_pedRTK1	INF_R_431_primary
NHL_BURKITT	INF_R_993_relapse2
SARCOMA_X-OTHERS	INF_R_265_progression
SARCOMA_X-OTHERS	INF_R_1273_relapse2
HGG_X-OTHERS	INF_R_686_relapse3
SARCOMA_RMS_ALV	INF_R_1007_relapse1
SARCOMA_OSTEO	INF_R_994_progression
SARCOMA_OSTEO	INF_R_473_relapse1
AML	INF_R_1676_progression
HGG_pedRTK1	INF_R_1180_relapse1
SARCOMA_RMS_EMB	INF_R_1397_relapse1
SARCOMA_RMS_EMB	INF_R_1691_primary
ATRT	INF_R_307_relapse1
HGG_X-OTHERS	INF_R_273_primary
HGG_pedRTK1	INF_R_1899_primary
HGG_X-OTHERS	INF_R_1320_relapse1

ACC	INF_R_354_relapse1
HGG_X-OTHERS	INF_R_1246_relapse2
SARCOMA_OSTEO	INF_R_177_relapse1
SARCOMA_OSTEO	INF_R_1879_primary
SARCOMA_X-OTHERS	INF_R_769_relapse1
SARCOMA_X-OTHERS	INF_R_914_primary
ACC	INF_R_1854_progression
HGG_X-OTHERS	INF_R_1357_relapse1
MED_SHH	INF_R_1076_primary
HGG_X-OTHERS	INF_R_025_relapse2
HGG_X-OTHERS	INF_R_1123_primary
SARCOMA_OSTEO	INF_R_130_relapse6
X-OTHERS_BRAIN	INF_R_852_relapse6

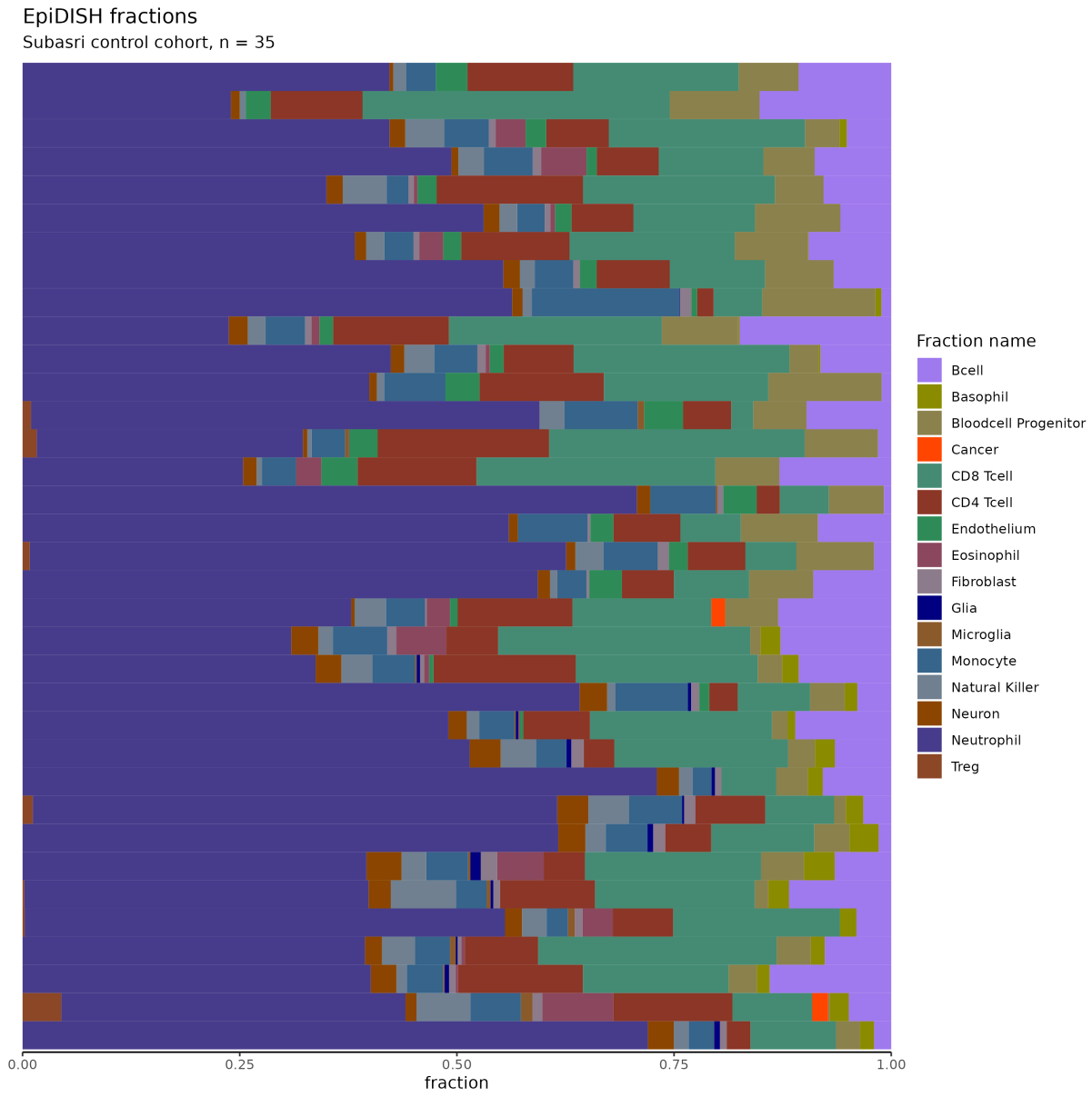


Figure 49: EpiDISH cell type composition estimations of liquid biopsy samples from Subasri cohort.

Network graph
MEmidnightblue, MEbrown and MEdarkred

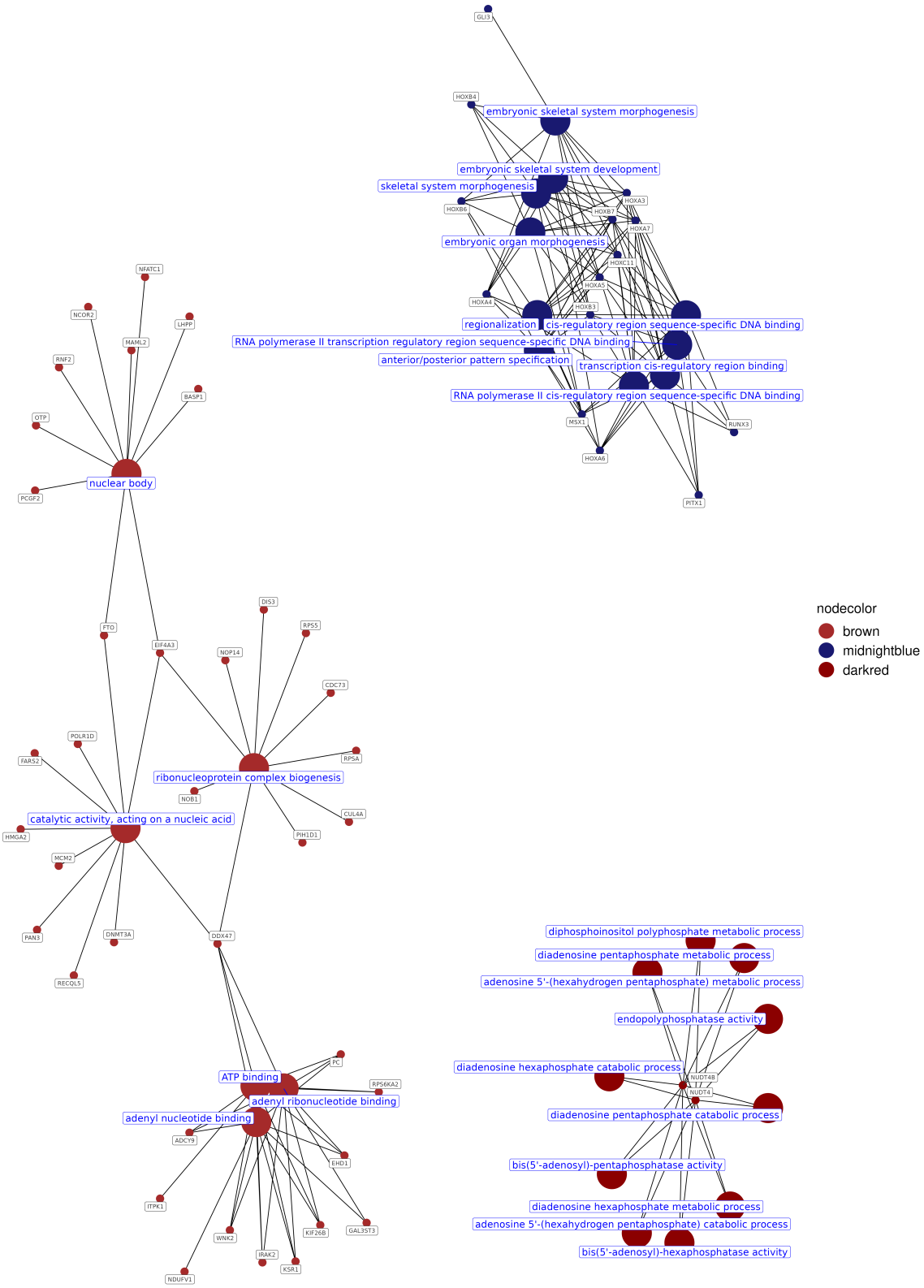


Figure 50: Networkplot generated by considering clusters identified via WGCNA on ND1 that were significantly enriched with probes identified by contrasts Ia or Ib. For each cluster the top 10 enriched pathways were selected (sorted by FDR) and inside each pathway the top 10 genes were selected (sorted by FDR of the matched methylation probe).

Network graph
MEgreen, MEpink, MElightgreen and MEsalmon

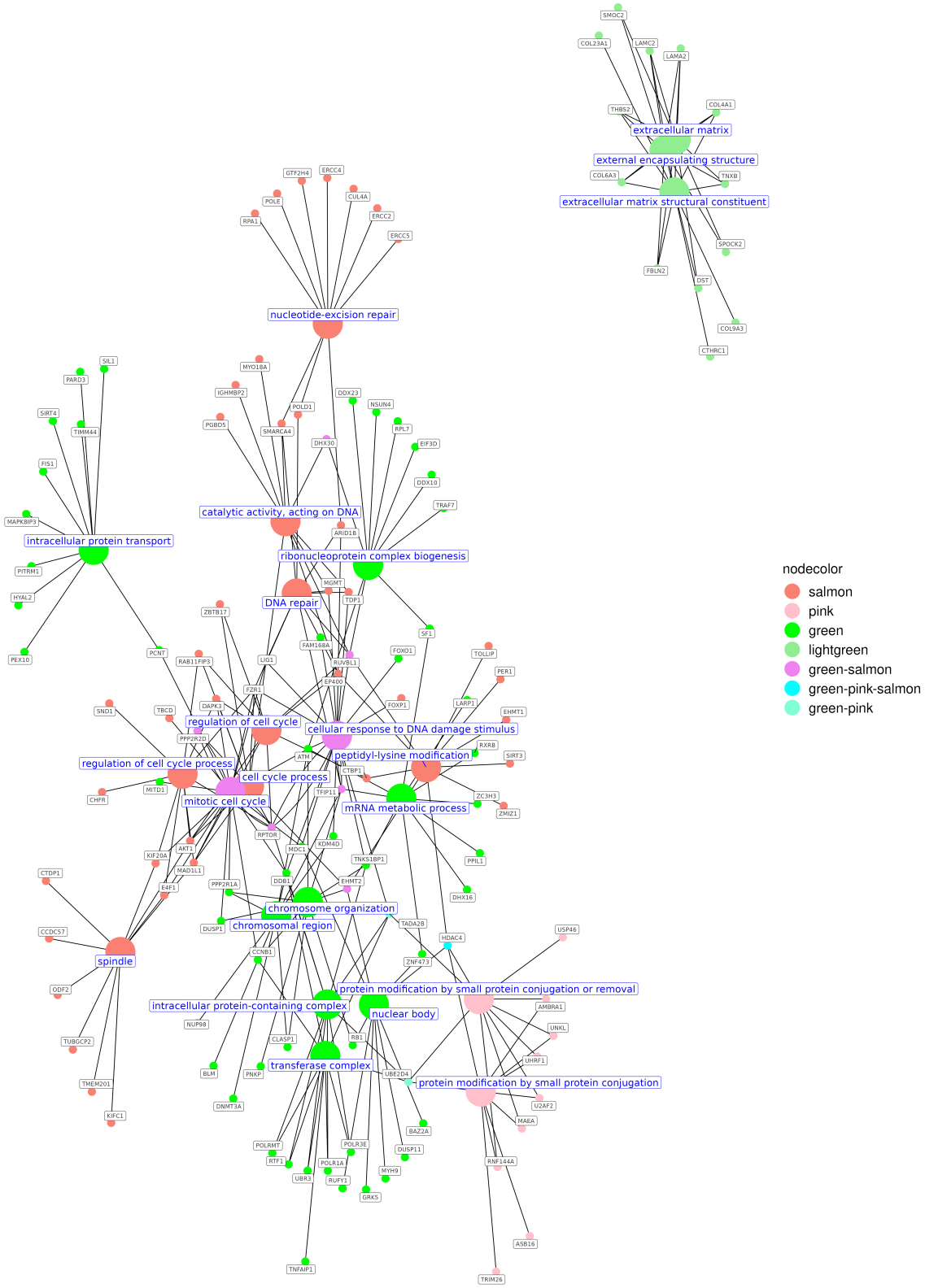


Figure 51: Networkplot generated by considering clusters identified via WGCNA on ND2 that were significantly enriched with probes identified by contrasts IIa or IIb. For each cluster the top 10 enriched pathways were selected (sorted by FDR) and inside each pathway the top 10 genes were selected (sorted by FDR of the matched methylation probe).

11 Bibliography

1. Sung, H., et al., *Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries.* CA Cancer J Clin, 2021. **71**(3): p. 209-249.
2. Siegel, R.L., et al., *Cancer statistics, 2023.* CA Cancer J Clin, 2023. **73**(1): p. 17-48.
3. Siegel, R.L., et al., *Cancer statistics, 2022.* CA Cancer J Clin, 2022. **72**(1): p. 7-33.
4. Hanahan, D. and R.A. Weinberg, *Hallmarks of cancer: the next generation.* Cell, 2011. **144**(5): p. 646-74.
5. Hanahan, D., *Hallmarks of Cancer: New Dimensions.* Cancer Discov, 2022. **12**(1): p. 31-46.
6. Jones, D.T.W., et al., *Molecular characteristics and therapeutic vulnerabilities across paediatric solid tumours.* Nat Rev Cancer, 2019. **19**(8): p. 420-438.
7. Rahal, Z., et al., *Genomics of adult and pediatric solid tumors.* Am J Cancer Res, 2018. **8**(8): p. 1356-1386.
8. Terry, R.L., et al., *Immune profiling of pediatric solid tumors.* The Journal of Clinical Investigation, 2020. **130**(7): p. 3391-3402.
9. Kattner, P., et al., *Compare and contrast: pediatric cancer versus adult malignancies.* Cancer Metastasis Rev, 2019. **38**(4): p. 673-682.
10. Ginsberg, G., et al., *Evaluation of Child/Adult Pharmacokinetic Differences from a Database Derived from the Therapeutic Drug Literature.* Toxicological Sciences, 2002. **66**(2): p. 185-200.
11. de Rojas, T., et al., *Changing incentives to ACCELERATE drug development for paediatric cancer.* Cancer Med, 2023. **12**(7): p. 8825-8837.
12. Ward, E., et al., *Childhood and adolescent cancer statistics, 2014.* CA Cancer J Clin, 2014. **64**(2): p. 83-103.
13. Siegel, D.A., et al., *Counts, incidence rates, and trends of pediatric cancer in the United States, 2003-2019.* J Natl Cancer Inst, 2023. **115**(11): p. 1337-1354.
14. Goldstick, J.E., R.M. Cunningham, and P.M. Carter, *Current causes of death in children and adolescents in the United States.* New England journal of medicine, 2022. **386**(20): p. 1955-1956.
15. Steliarova-Foucher, E., et al., *International incidence of childhood cancer, 2001-10: a population-based registry study.* Lancet Oncol, 2017. **18**(6): p. 719-731.
16. Inskip, P.D. and R.E. Curtis, *New malignancies following childhood cancer in the United States, 1973-2002.* Int J Cancer, 2007. **121**(10): p. 2233-40.
17. Meadows, A.T., et al., *Second neoplasms in survivors of childhood cancer: findings from the Childhood Cancer Survivor Study cohort.* J Clin Oncol, 2009. **27**(14): p. 2356-62.
18. Scholz-Kreisel, P., et al., *Second Malignancies Following Childhood Cancer Treatment in Germany From 1980 to 2014.* Dtsch Arztebl Int, 2018. **115**(23): p. 385-392.
19. Hudson, T.J., et al., *International network of cancer genome projects.* Nature, 2010. **464**(7291): p. 993-8.
20. Downing, J.R., et al., *The Pediatric Cancer Genome Project.* Nat Genet, 2012. **44**(6): p. 619-22.
21. van Tilburg, C.M., et al., *The Pediatric Precision Oncology INFORM Registry: Clinical Outcome and Benefit for Patients with Very High-Evidence Targets.* Cancer Discovery, 2021. **11**(11): p. 2764-2779.
22. Worst, B.C., et al., *Next-generation personalised medicine for high-risk paediatric cancer patients - The INFORM pilot study.* Eur J Cancer, 2016. **65**: p. 91-101.
23. Langenberg, K.P.S., E.J. Looze, and J.J. Molenaar, *The Landscape of Pediatric Precision Oncology: Program Design, Actionable Alterations, and Clinical Trial Development.* Cancers (Basel), 2021. **13**(17).
24. Chang, W., et al., *MultiDimensional ClinOmics for Precision Therapy of Children and Adolescent Young Adults with Relapsed and Refractory Cancer: A Report from the Center for Cancer Research.* Clin Cancer Res, 2016. **22**(15): p. 3810-20.

25. Harttrampf, A.C., et al., *Molecular Screening for Cancer Treatment Optimization (MOSCATO-01) in Pediatric Patients: A Single-Institutional Prospective Molecular Stratification Trial.* Clin Cancer Res, 2017. **23**(20): p. 6101-6112.
26. Khater, F., et al., *Molecular Profiling of Hard-to-Treat Childhood and Adolescent Cancers.* JAMA Netw Open, 2019. **2**(4): p. e192906.
27. Parsons, D.W., et al., *Diagnostic Yield of Clinical Tumor and Germline Whole-Exome Sequencing for Children With Solid Tumors.* JAMA Oncol, 2016. **2**(5): p. 616-624.
28. Lau, L., et al., *Pilot study of a comprehensive precision medicine platform for children with high-risk cancer*. 2017, American Society of Clinical Oncology.
29. Villani, A., et al., *The clinical utility of integrative genomics in childhood cancer extends beyond targetable mutations.* Nature Cancer, 2023. **4**(2): p. 203-221.
30. Benezech, S., et al., *Tumor Molecular Profiling: Pediatric Results of the ProfiLER Study.* JCO Precis Oncol, 2020. **4**: p. 785-795.
31. Langenberg, K., E. Dolman, and J. Molenaar, *Abstract A40: Integration of high-throughput drug screening on patient-derived organdies into pediatric precision medicine programs: The future is now!* Cancer Research, 2020. **80**(14_Supplement): p. A40-A40.
32. Gentles, A.J. and D. Gallahan, *Systems biology: confronting the complexity of cancer.* Cancer Res, 2011. **71**(18): p. 5961-4.
33. Koboldt, D.C., et al., *Comprehensive molecular portraits of human breast tumours.* Nature, 2012. **490**(7418): p. 61-70.
34. Heo, Y.J., et al., *Integrative Multi-Omics Approaches in Cancer Research: From Biological Networks to Clinical Subtypes.* Mol Cells, 2021. **44**(7): p. 433-443.
35. Hood, L. and L. Rowen, *The Human Genome Project: big science transforms biology and medicine.* Genome Medicine, 2013. **5**(9): p. 79.
36. Mardis, E.R., *A decade's perspective on DNA sequencing technology.* Nature, 2011. **470**(7333): p. 198-203.
37. Slatko, B.E., A.F. Gardner, and F.M. Ausubel, *Overview of Next-Generation Sequencing Technologies.* Curr Protoc Mol Biol, 2018. **122**(1): p. e59.
38. Khurana, E., et al., *Role of non-coding sequence variants in cancer.* Nature Reviews Genetics, 2016. **17**(2): p. 93-108.
39. Bailey, M.H., et al., *Retrospective evaluation of whole exome and genome mutation calls in 746 cancer samples.* Nature Communications, 2020. **11**(1): p. 4748.
40. Maurano, M.T., et al., *Systematic localization of common disease-associated variation in regulatory DNA.* Science, 2012. **337**(6099): p. 1190-5.
41. Chen, C.-y., et al., *On the identification of potential regulatory variants within genome wide association candidate SNP sets.* BMC Medical Genomics, 2014. **7**(1): p. 34.
42. Moore, L.D., T. Le, and G. Fan, *DNA Methylation and Its Basic Function.* Neuropsychopharmacology, 2013. **38**(1): p. 23-38.
43. Lakshminarasimhan, R. and G. Liang, *The Role of DNA Methylation in Cancer.* Adv Exp Med Biol, 2016. **945**: p. 151-172.
44. Capper, D., et al., *DNA methylation-based classification of central nervous system tumours.* Nature, 2018. **555**(7697): p. 469-474.
45. Schumacher, A., et al., *Microarray-based DNA methylation profiling: technology and applications.* Nucleic Acids Res, 2006. **34**(2): p. 528-42.
46. Bibikova, M., et al., *High density DNA methylation array with single CpG site resolution.* Genomics, 2011. **98**(4): p. 288-295.
47. O'Neil, N.J., M.L. Bailey, and P. Hieter, *Synthetic lethality and cancer.* Nature Reviews Genetics, 2017. **18**(10): p. 613-623.
48. Lucchesi, J.C., *Synthetic lethality and semi-lethality among functionally related mutants of Drosophila melanogaster.* Genetics, 1968. **59**(1): p. 37.
49. Dobzhansky, T., *Genetics of natural populations. XIII. Recombination and variability in populations of Drosophila pseudoobscura.* Genetics, 1946. **31**(3): p. 269.
50. Tutt, A., et al., *Phase II trial of the oral PARP inhibitor olaparib in BRCA-deficient advanced breast cancer.* Journal of clinical oncology, 2009. **27**(18_suppl): p. CRA501-CRA501.

51. Audeh, M.W., et al., *Oral poly (ADP-ribose) polymerase inhibitor olaparib in patients with BRCA1 or BRCA2 mutations and recurrent ovarian cancer: a proof-of-concept trial.* The lancet, 2010. **376**(9737): p. 245-251.
52. Bryant, H.E., et al., *Specific killing of BRCA2-deficient tumours with inhibitors of poly (ADP-ribose) polymerase.* Nature, 2005. **434**(7035): p. 913-917.
53. Liany, H., A. Jeyasekharan, and V. Rajan, *Predicting synthetic lethal interactions using heterogeneous data sources.* Bioinformatics, 2020. **36**(7): p. 2209-2216.
54. Luo, J., et al., *A genome-wide RNAi screen identifies multiple synthetic lethal interactions with the Ras oncogene.* Cell, 2009. **137**(5): p. 835-848.
55. Du, D., et al., *Genetic interaction mapping in mammalian cells using CRISPR interference.* Nature methods, 2017. **14**(6): p. 577-580.
56. Zhou, P., et al., *A Three-Way Combinatorial CRISPR Screen for Analyzing Interactions among Druggable Targets.* Cell Reports, 2020. **32**(6): p. 108020.
57. Wang, T., et al., *Genetic screens in human cells using the CRISPR-Cas9 system.* Science, 2014. **343**(6166): p. 80-84.
58. Shalem, O., et al., *Genome-scale CRISPR-Cas9 knockout screening in human cells.* Science, 2014. **343**(6166): p. 84-87.
59. Hart, T., et al., *High-resolution CRISPR screens reveal fitness genes and genotype-specific cancer liabilities.* Cell, 2015. **163**(6): p. 1515-1526.
60. Chen, S., et al., *Genome-wide CRISPR screen in a mouse model of tumor growth and metastasis.* Cell, 2015. **160**(6): p. 1246-1260.
61. Jerby-Arnon, L., et al., *Predicting Cancer-Specific Vulnerability via Data-Driven Detection of Synthetic Lethality.* Cell, 2014. **158**(5): p. 1199-1209.
62. Paladugu, S.R., et al., *Mining protein networks for synthetic genetic interactions.* Bmc Bioinformatics, 2008. **9**(1): p. 1-14.
63. De Kegel, B., et al., *Comprehensive prediction of robust synthetic lethality between paralog pairs in cancer cell lines.* Cell Systems, 2021. **12**(12): p. 1144-1159.e6.
64. Long, Y., et al., *Graph contextualized attention network for predicting synthetic lethality in human cancers.* Bioinformatics, 2021. **37**(16): p. 2432-2440.
65. Cai, R., et al., *Dual-dropout graph convolutional network for predicting synthetic lethality in human cancers.* Bioinformatics, 2020. **36**(16): p. 4458-4465.
66. Wang, S., et al., *KG4SL: knowledge graph neural network for synthetic lethality prediction in human cancers.* Bioinformatics, 2021. **37**(Suppl_1): p. i418-i425.
67. Wang, J., et al., *SynLethDB 2.0: a web-based knowledge graph database on synthetic lethality for novel anticancer drug discovery.* Database (Oxford), 2022. **2022**.
68. Kratz, C.P., et al., *Predisposition to cancer in children and adolescents.* Lancet Child Adolesc Health, 2021. **5**(2): p. 142-154.
69. Waszak, S.M., et al., *Germline Elongator mutations in Sonic Hedgehog medulloblastoma.* Nature, 2020. **580**(7803): p. 396-401.
70. Villani, A., et al., *Impact of early detection strategies on cancer mortality in germlineTP53 mutation carriers in Li-Fraumeni syndrome.* Journal of Clinical Oncology, 2010. **28**(15_suppl): p. 1509-1509.
71. Brodeur, G.M., et al., *Pediatric Cancer Predisposition and Surveillance: An Overview, and a Tribute to Alfred G. Knudson Jr.* Clin Cancer Res, 2017. **23**(11): p. e1-e5.
72. Jongmans, M.C., et al., *Recognition of genetic predisposition in pediatric cancer patients: An easy-to-use selection tool.* Eur J Med Genet, 2016. **59**(3): p. 116-25.
73. Villani, A., et al., *Biochemical and imaging surveillance in germline TP53 mutation carriers with Li-Fraumeni syndrome: 11 year follow-up of a prospective observational study.* The Lancet Oncology, 2016. **17**(9): p. 1295-1305.
74. Villani, A., et al., *Biochemical and imaging surveillance in germline TP53 mutation carriers with Li-Fraumeni syndrome: a prospective observational study.* Lancet Oncol, 2011. **12**(6): p. 559-67.
75. Lammens, C.R., et al., *Regular surveillance for Li-Fraumeni Syndrome: advice, adherence and perceived benefits.* Fam Cancer, 2010. **9**(4): p. 647-54.

76. Krutilkova, V., et al., *Identification of five new families strengthens the link between childhood choroid plexus carcinoma and germline TP53 mutations.* European journal of cancer, 2005. **41**(11): p. 1597-1603.
77. Wimmer, K. and J. Etzler, *Constitutional mismatch repair-deficiency syndrome: have we so far seen only the tip of an iceberg?* Human genetics, 2008. **124**: p. 105-122.
78. Sharma, R., S. Lewis, and M.W. Wlodarski, *DNA Repair Syndromes and Cancer: Insights Into Genetics and Phenotype Patterns.* Front Pediatr, 2020. **8**: p. 570084.
79. Peltomäki, P., *Role of DNA mismatch repair defects in the pathogenesis of human cancer.* J Clin Oncol, 2003. **21**(6): p. 1174-9.
80. Lukish, J.R., et al., *Prognostic significance of DNA replication errors in young patients with colorectal cancer.* Ann Surg, 1998. **227**(1): p. 51-6.
81. Karran, P. and R. Hampson, *Genomic instability and tolerance to alkylating agents.* Cancer Surv, 1996. **28**: p. 69-85.
82. Aubrey, B.J., A. Strasser, and G.L. Kelly, *Tumor-Suppressor Functions of the TP53 Pathway.* Cold Spring Harb Perspect Med, 2016. **6**(5).
83. Pinto, E.M., et al., *TP53-Associated Pediatric Malignancies.* Genes Cancer, 2011. **2**(4): p. 485-90.
84. Alexandrov, L.B., et al., *Signatures of mutational processes in human cancer.* Nature, 2013. **500**(7463): p. 415-21.
85. Alexandrov, Ludmil B., et al., *Deciphering Signatures of Mutational Processes Operative in Human Cancer.* Cell Reports, 2013. **3**(1): p. 246-259.
86. Tate, J.G., et al., *COSMIC: the Catalogue Of Somatic Mutations In Cancer.* Nucleic Acids Research, 2018. **47**(D1): p. D941-D947.
87. Alexandrov, L.B., et al., *The repertoire of mutational signatures in human cancer.* Nature, 2020. **578**(7793): p. 94-101.
88. Alexandrov, L.B., et al., *Mutational signatures associated with tobacco smoking in human cancer.* Science, 2016. **354**(6312): p. 618-622.
89. Hayward, N.K., et al., *Whole-genome landscapes of major melanoma subtypes.* Nature, 2017. **545**(7653): p. 175-180.
90. Vogelstein, B., et al., *Cancer genome landscapes.* Science, 2013. **339**(6127): p. 1546-58.
91. Saadeh, C., D. Bright, and D. Rustem, *Precision Medicine in Oncology Pharmacy Practice.* Acta Med Acad, 2019. **48**(1): p. 90-104.
92. Zsákai, L., et al., *Targeted drug combination therapy design based on driver genes.* Oncotarget, 2019. **10**(51): p. 5255-5266.
93. Van Hoeck, A., et al., *Portrait of a cancer: mutational signature analyses for cancer diagnostics.* BMC Cancer, 2019. **19**(1): p. 457.
94. Vanderstichele, A., et al., *Genomic signatures as predictive biomarkers of homologous recombination deficiency in ovarian cancer.* Eur J Cancer, 2017. **86**: p. 5-14.
95. Polak, P., et al., *A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer.* Nat Genet, 2017. **49**(10): p. 1476-1486.
96. Nik-Zainal, S., et al., *Mutational processes molding the genomes of 21 breast cancers.* Cell, 2012. **149**(5): p. 979-93.
97. Helleday, T., S. Eshtad, and S. Nik-Zainal, *Mechanisms underlying mutational signatures in human cancers.* Nat Rev Genet, 2014. **15**(9): p. 585-98.
98. Davies, H., et al., *HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures.* Nat Med, 2017. **23**(4): p. 517-525.
99. Ma, X., et al., *Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and solid tumours.* Nature, 2018. **555**(7696): p. 371-376.
100. Gröbner, S.N., et al., *The landscape of genomic alterations across childhood cancers.* Nature, 2018. **555**(7696): p. 321-327.
101. Zhang, J., et al., *Germline Mutations in Predisposition Genes in Pediatric Cancer.* N Engl J Med, 2015. **373**(24): p. 2336-2346.
102. Kaminsky, Z.A., et al., *DNA methylation profiles in monozygotic and dizygotic twins.* Nature genetics, 2009. **41**(2): p. 240-245.
103. Fraga, M.F., et al., *Epigenetic differences arise during the lifetime of monozygotic twins.* Proceedings of the National Academy of Sciences, 2005. **102**(30): p. 10604-10609.

104. Shen, H. and P.W. Laird, *Interplay between the cancer genome and epigenome.* Cell, 2013. **153**(1): p. 38-55.
105. Moran, S., et al., *Epigenetic profiling to classify cancer of unknown primary: a multicentre, retrospective analysis.* Lancet Oncol, 2016. **17**(10): p. 1386-1395.
106. Fernandez, A.F., et al., *A DNA methylation fingerprint of 1628 human samples.* Genome Res, 2012. **22**(2): p. 407-19.
107. Paziewska, A., et al., *DNA methylation status is more reliable than gene expression at detecting cancer in prostate biopsy.* Br J Cancer, 2014. **111**(4): p. 781-9.
108. Heikkinen, A., S. Bollepalli, and M. Ollikainen, *The potential of DNA methylation as a biomarker for obesity and smoking.* J Intern Med, 2022. **292**(3): p. 390-408.
109. 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 neuropathologica, 2013. **125**: p. 913-916.
110. Sturm, D., et al., *New Brain Tumor Entities Emerge from Molecular Classification of CNS-PNETs.* Cell, 2016. **164**(5): p. 1060-1072.
111. Sturm, D., et al., *Hotspot mutations in H3F3A and IDH1 define distinct epigenetic and biological subgroups of glioblastoma.* Cancer Cell, 2012. **22**(4): p. 425-37.
112. Wiestler, B., et al., *Integrated DNA methylation and copy-number profiling identify three clinically and biologically relevant groups of anaplastic glioma.* Acta Neuropathol, 2014. **128**(4): p. 561-71.
113. van den Bent, M.J., *Interobserver variation of the histopathological diagnosis in clinical trials on glioma: a clinician's perspective.* Acta Neuropathol, 2010. **120**(3): p. 297-304.
114. Ellison, D.W., et al., *Histopathological grading of pediatric ependymoma: reproducibility and clinical relevance in European trial cohorts.* J Negat Results Biomed, 2011. **10**: p. 7.
115. Sturm, D., et al., *Multiomic neuropathology improves diagnostic accuracy in pediatric neuro-oncology.* Nature Medicine, 2023. **29**(4): p. 917-926.
116. Matzenbacher Bittar, C., et al., *Clinical and molecular characterization of patients fulfilling Chompret criteria for Li-Fraumeni syndrome in Southern Brazil.* PLoS One, 2021. **16**(9): p. e0251639.
117. Samuel, N., et al., *Genome-Wide DNA Methylation Analysis Reveals Epigenetic Dysregulation of MicroRNA-34A in TP53-Associated Cancer Susceptibility.* J Clin Oncol, 2016. **34**(30): p. 3697-3704.
118. Subasri, V., et al., *Multiple Germline Events Contribute to Cancer Development in Patients with Li-Fraumeni Syndrome.* Cancer Res Commun, 2023. **3**(5): p. 738-754.
119. Sedivy, J.M., G. Banumathy, and P.D. Adams, *Aging by epigenetics—a consequence of chromatin damage?* Experimental cell research, 2008. **314**(9): p. 1909-1917.
120. Johnson, A.A., et al., *The role of DNA methylation in aging, rejuvenation, and age-related disease.* Rejuvenation Res, 2012. **15**(5): p. 483-94.
121. Jones, M.J., S.J. Goodman, and M.S. Kobor, *DNA methylation and healthy human aging.* Aging Cell, 2015. **14**(6): p. 924-32.
122. Wong, D., et al., *Early Cancer Detection in Li-Fraumeni Syndrome with Cell-Free DNA.* Cancer Discov, 2024. **14**(1): p. 104-119.
123. Vrba, L. and B.W. Futscher, *A suite of DNA methylation markers that can detect most common human cancers.* Epigenetics, 2018. **13**(1): p. 61-72.
124. Guarente, L., *Synthetic enhancement in gene interaction: a genetic tool come of age.* Trends in Genetics, 1993. **9**(10): p. 362-366.
125. Hartman IV, J.L., B. Garvik, and L. Hartwell, *Principles for the buffering of genetic variation.* Science, 2001. **291**(5506): p. 1001-1004.
126. Kroll, E.S., et al., *Establishing genetic interactions by a synthetic dosage lethality phenotype.* Genetics, 1996. **143**(1): p. 95-102.
127. Measday, V. and P. Hieter, *Synthetic dosage lethality*, in *Methods in enzymology*. 2002, Elsevier. p. 316-326.
128. Li, J.J. and I. Herskowitz, *Isolation of ORC6, a component of the yeast origin recognition complex by a one-hybrid system.* Science, 1993. **262**(5141): p. 1870-1874.
129. Ferrer-Bonsoms, J.A., L. Jareno, and A. Rubio, *Rediscover: an R package to identify mutually exclusive mutations.* Bioinformatics, 2022. **38**(3): p. 844-845.

130. Ruepp, A., et al., *CORUM: the comprehensive resource of mammalian protein complexes.* Nucleic Acids Res, 2008. **36**(Database issue): p. D646-50.
131. Ruepp, A., et al., *CORUM: the comprehensive resource of mammalian protein complexes--2009.* Nucleic Acids Res, 2010. **38**(Database issue): p. D497-501.
132. Giurgiu, M., et al., *CORUM: the comprehensive resource of mammalian protein complexes-2019.* Nucleic Acids Res, 2019. **47**(D1): p. D559-d563.
133. Liberzon, A., et al., *Molecular signatures database (MSigDB) 3.0.* Bioinformatics, 2011. **27**(12): p. 1739-1740.
134. Liberzon, A., et al., *The Molecular Signatures Database (MSigDB) hallmark gene set collection.* Cell Syst, 2015. **1**(6): p. 417-425.
135. Subramanian, A., et al., *Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles.* Proceedings of the National Academy of Sciences, 2005. **102**(43): p. 15545-15550.
136. Singh, A.P. and G.J. Gordon. *Relational learning via collective matrix factorization*. in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2008.
137. Bouchard, G., D. Yin, and S. Guo. *Convex collective matrix factorization*. in *Artificial intelligence and statistics*. 2013. PMLR.
138. Developers, T., *TensorFlow.* Zenodo, 2022.
139. Kingma, D.P. and J. Ba, *Adam: A method for stochastic optimization.* arXiv preprint arXiv:1412.6980, 2014.
140. Csardi, G. and T. Nepusz, *The igraph software package for complex network research.* InterJournal, complex systems, 2006. **1695**(5): p. 1-9.
141. Manning, C.D. and H. Schütze, *Foundations of Statistical Natural Language Processing*. 1999, Cambridge, Massachusetts: The MIT Press.
142. Raghavan, V., P. Bollmann, and G.S. Jung, *A critical investigation of recall and precision as measures of retrieval system performance.* ACM Trans. Inf. Syst., 1989. **7**(3): p. 205–229.
143. Peterson, W.W., T.G. Birdsall, and W.C. Fox, *The theory of signal detectability.* Trans. IRE Prof. Group Inf. Theory, 1954. **4**: p. 171-212.
144. Natarajan, N. and I.S. Dhillon, *Inductive matrix completion for predicting gene–disease associations.* Bioinformatics, 2014. **30**(12): p. i60-i68.
145. Thatikonda, V., et al., *Comprehensive analysis of mutational signatures reveals distinct patterns and molecular processes across 27 pediatric cancers.* Nature Cancer, 2023. **4**(2): p. 276-289.
146. Paramasivam, N., *Personal email conversation*, L. Madenach, Editor. 2023.
147. Zhou, W., P.W. Laird, and H. Shen, *Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes.* Nucleic Acids Res, 2017. **45**(4): p. e22.
148. Islam, S.M.A., et al., *Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor.* bioRxiv, 2022: p. 2020.12.13.422570.
149. Degasperi, A., et al., *A practical framework and online tool for mutational signature analyses show inter-tissue variation and driver dependencies.* Nat Cancer, 2020. **1**(2): p. 249-263.
150. Maura, F., et al., *A practical guide for mutational signature analysis in hematological malignancies.* Nat Commun, 2019. **10**(1): p. 2969.
151. Bauer, D.F., *Constructing Confidence Sets Using Rank Statistics.* Journal of the American Statistical Association, 1972. **67**(339): p. 687-690.
152. Pich, O., et al., *The mutational footprints of cancer therapies.* Nature Genetics, 2019. **51**(12): p. 1732-1740.
153. Wong, J.K.L., et al., *Association of mutation signature effectuating processes with mutation hotspots in driver genes and non-coding regions.* Nat Commun, 2022. **13**(1): p. 178.
154. Benjamini, Y. and Y. Hochberg, *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.* Journal of the Royal Statistical Society: Series B (Methodological), 1995. **57**(1): p. 289-300.
155. Teschendorff, A.E., et al., *A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies.* BMC Bioinformatics, 2017. **18**(1): p. 105.

156. Du, P., et al., *Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis.* BMC Bioinformatics, 2010. **11**(1): p. 587.
157. Ritchie, M.E., et al., *limma powers differential expression analyses for RNA-sequencing and microarray studies.* Nucleic Acids Research, 2015. **43**(7): p. e47-e47.
158. Peters, Timothy J., et al., *Calling differentially methylated regions from whole genome bisulphite sequencing with DMRcate.* Nucleic Acids Research, 2021. **49**(19): p. e109-e109.
159. Peters, T.J., et al., *De novo identification of differentially methylated regions in the human genome.* Epigenetics & Chromatin, 2015. **8**(1): p. 6.
160. Phipson, B., J. Maksimovic, and A. Oshlack, *missMethyl: an R package for analyzing data from Illumina's HumanMethylation450 platform.* Bioinformatics, 2016. **32**(2): p. 286-8.
161. Phipson, B. and A. Oshlack, *DiffVar: a new method for detecting differential variability with application to methylation in cancer and aging.* Genome Biol, 2014. **15**(9): p. 465.
162. Maksimovic, J., L. Gordon, and A. Oshlack, *SWAN: Subset-quantile Within Array Normalization for Illumina Infinium HumanMethylation450 BeadChips.* Genome Biology, 2012. **13**(6): p. R44.
163. Maksimovic, J., et al., *Removing unwanted variation in a differential methylation analysis of Illumina HumanMethylation450 array data.* Nucleic Acids Res, 2015. **43**(16): p. e106.
164. Pearl, L.H., et al., *Therapeutic opportunities within the DNA damage response.* Nature Reviews Cancer, 2015. **15**(3): p. 166-180.
165. Langfelder, P. and S. Horvath, *WGCNA: an R package for weighted correlation network analysis.* BMC Bioinformatics, 2008. **9**(1): p. 559.
166. Langfelder, P. and S. Horvath, *Fast R functions for robust correlations and hierarchical clustering.* Journal of statistical software, 2012. **46**(11).
167. McInnes, L., J. Healy, and J. Melville, *Umap: Uniform manifold approximation and projection for dimension reduction.* arXiv preprint arXiv:1802.03426, 2018.
168. Korsunsky, I., et al., *Fast, sensitive and accurate integration of single-cell data with Harmony.* Nat Methods, 2019. **16**(12): p. 1289-1296.
169. Rainio, O., J. Teuho, and R. Klén, *Evaluation metrics and statistical tests for machine learning.* Scientific Reports, 2024. **14**(1): p. 6086.
170. Davis, J. and M. Goadrich, *The relationship between Precision-Recall and ROC curves*, in *Proceedings of the 23rd international conference on Machine learning*. 2006, Association for Computing Machinery: Pittsburgh, Pennsylvania, USA. p. 233–240.
171. Zitnik, M., et al., *Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities.* Information Fusion, 2019. **50**: p. 71-91.
172. Žitnik, M. and B. Zupan, *Data Fusion by Matrix Factorization.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015. **37**(1): p. 41-53.
173. Ritchie, M.D., et al., *Methods of integrating data to uncover genotype–phenotype interactions.* Nature Reviews Genetics, 2015. **16**(2): p. 85-97.
174. Seale, C., Y. Tepeli, and J.P. Gonçalves, *Overcoming selection bias in synthetic lethality prediction.* Bioinformatics, 2022. **38**(18): p. 4360-4368.
175. Tepeli, Y.I., C. Seale, and J.P. Gonçalves, *ELISL: early–late integrated synthetic lethality prediction in cancer.* Bioinformatics, 2023. **40**(1).
176. Perla, A., et al., *Histone Deacetylase Inhibitors in Pediatric Brain Cancers: Biological Activities and Therapeutic Potential.* Front Cell Dev Biol, 2020. **8**: p. 546.
177. Kang, M., et al., *Targeting BAP1 with small compound inhibitor for colon cancer treatment.* Scientific Reports, 2023. **13**(1): p. 2264.
178. Ostadkarampour, M. and E.E. Putnins, *Monoamine Oxidase Inhibitors: A Review of Their Anti-Inflammatory Therapeutic Potential and Mechanisms of Action.* Front Pharmacol, 2021. **12**: p. 676239.
179. Koppaka, V., et al., *Aldehyde dehydrogenase inhibitors: a comprehensive review of the pharmacology, mechanism of action, substrate specificity, and clinical application.* Pharmacol Rev, 2012. **64**(3): p. 520-39.
180. Nik-Zainal, S. and S. Morganella, *Mutational signatures in breast cancer: the problem at the DNA level.* Clinical Cancer Research, 2017. **23**(11): p. 2617-2629.
181. Singh, V.K., et al., *Mutational signature SBS8 predominantly arises due to late replication errors in cancer.* Communications Biology, 2020. **3**(1): p. 421.

182. Periyasamy, M., et al., *p53 controls expression of the DNA deaminase APOBEC3B to limit its potential mutagenic activity in cancer cells.* Nucleic Acids Res, 2017. **45**(19): p. 11056-11069.
183. Wang, S., et al., *APOBEC3B and APOBEC mutational signature as potential predictive markers for immunotherapy response in non-small cell lung cancer.* Oncogene, 2018. **37**(29): p. 3924-3936.
184. Hwang, T., et al., *Defining the mutation signatures of DNA polymerase θ in cancer genomes.* NAR Cancer, 2020. **2**(3): p. zcaa017.
185. Zou, X., et al., *A systematic CRISPR screen defines mutational mechanisms underpinning signatures caused by replication errors and endogenous DNA damage.* Nat Cancer, 2021. **2**(6): p. 643-657.
186. Giner-Calabuig, M., et al., *Mutational signature profiling classifies subtypes of clinically different mismatch-repair-deficient tumours with a differential immunogenic response potential.* Br J Cancer, 2022. **126**(11): p. 1595-1603.
187. Haradhvala, N., et al., *Distinct mutational signatures characterize concurrent loss of polymerase proofreading and mismatch repair.* Nature communications, 2018. **9**(1): p. 1746.
188. Kucab, J.E., et al., *A Compendium of Mutational Signatures of Environmental Agents.* Cell, 2019. **177**(4): p. 821-836.e16.
189. Light, N., et al., *Germline TP53 mutations undergo copy number gain years prior to tumor diagnosis.* Nat Commun, 2023. **14**(1): p. 77.
190. Reijns, M.A., et al., *Signatures of TOP1 transcription-associated mutagenesis in cancer and germline.* Nature, 2022. **602**(7898): p. 623-631.
191. Søe, K. and F. Grosse, *p53 stimulates human topoisomerase I activity by modulating its DNA binding.* Nucleic Acids Res, 2003. **31**(22): p. 6585-92.
192. Grabovska, Y., et al., *Pediatric pan-central nervous system tumor analysis of immune-cell infiltration identifies correlates of antitumor immunity.* Nature Communications, 2020. **11**(1): p. 4324.
193. Stein, T., et al., *RNA polymerase III transcription can be derepressed by oncogenes or mutations that compromise p53 function in tumours and Li-Fraumeni syndrome.* Oncogene, 2002. **21**(19): p. 2961-2970.
194. Katoh, M. and H. Nakagama, *FGF Receptors: Cancer Biology and Therapeutics.* Medicinal Research Reviews, 2014. **34**(2): p. 280-300.
195. Mani, S.A., et al., *Mesenchyme Forkhead 1 (FOXC2) plays a key role in metastasis and is associated with aggressive basal-like breast cancers.* Proc Natl Acad Sci U S A, 2007. **104**(24): p. 10069-74.
196. Krell, J., et al., *TP53 regulates miRNA association with AGO2 to remodel the miRNA-mRNA interaction network.* Genome Res, 2016. **26**(3): p. 331-41.
197. De Smet, F., et al., *Nuclear inclusion bodies of mutant and wild-type p53 in cancer: a hallmark of p53 inactivation and proteostasis remodelling by p53 aggregation.* J Pathol, 2017. **242**(1): p. 24-38.
198. Kim, B.R., et al., *RUNX3 suppresses metastasis and stemness by inhibiting Hedgehog signaling in colorectal cancer.* Cell Death & Differentiation, 2020. **27**(2): p. 676-694.
199. Oei, V., et al., *RUNX3 inactivates oncogenic MYC through disruption of MYC/MAX complex and subsequent recruitment of GSK3β-FBXW7 cascade.* Communications Biology, 2023. **6**(1): p. 689.
200. Chuang, L.S.H., et al., *RUNX3 in Stem Cell and Cancer Biology.* Cells, 2023. **12**(3).
201. Caffrey, J.J., et al., *Discovery of molecular and catalytic diversity among human diphosphoinositol-polyphosphate phosphohydrolases. An expanding Nudt family.* J Biol Chem, 2000. **275**(17): p. 12730-6.
202. Fisher, D.I., et al., *Nudix hydrolases that degrade dinucleoside and diphosphoinositol polyphosphates also have 5-phosphoribosyl 1-pyrophosphate (PRPP) pyrophosphatase activity that generates the glycolytic activator ribose 1,5-bisphosphate.* J Biol Chem, 2002. **277**(49): p. 47313-7.
203. Lavoie, H., et al., *MEK drives BRAF activation through allosteric control of KSR proteins.* Nature, 2018. **554**(7693): p. 549-553.
204. Paniagua, G., et al., *KSR induces RAS-independent MAPK pathway activation and modulates the efficacy of KRAS inhibitors.* Molecular Oncology, 2022. **16**(17): p. 3066-3081.

205. Wei, L., et al., *Eukaryotic initiation factor 4 A-3 promotes glioblastoma growth and invasion through the Notch1-dependent pathway.* BMC Cancer, 2023. **23**(1): p. 550.
206. Barbosa, I., et al., *Human CWC22 escorts the helicase eIF4AIII to spliceosomes and promotes exon junction complex assembly.* Nat Struct Mol Biol, 2012. **19**(10): p. 983-90.
207. Lorès, P., et al., *The SWI/SNF protein BAF60b is ubiquitinated through a signalling process involving Rac GTPase and the RING finger protein Unkempt.* Febs j, 2010. **277**(6): p. 1453-64.
208. Di Bartolomeo, S., et al., *The dynamic interaction of AMBRA1 with the dynein motor complex regulates mammalian autophagy.* J Cell Biol, 2010. **191**(1): p. 155-68.
209. Simoneschi, D., et al., *CRL4(AMBRA1) is a master regulator of D-type cyclins.* Nature, 2021. **592**(7856): p. 789-793.
210. Maiani, E., et al., *AMBRA1 regulates cyclin D to guard S-phase entry and genomic integrity.* Nature, 2021. **592**(7856): p. 799-803.
211. Gu, W., et al., *Ambra1 is an essential regulator of autophagy and apoptosis in SW620 cells: pro-survival role of Ambra1.* PLoS One, 2014. **9**(2): p. e90151.
212. Chaikovsky, A.C., et al., *The AMBRA1 E3 ligase adaptor regulates the stability of cyclin D.* Nature, 2021. **592**(7856): p. 794-798.
213. Wang, L., et al., *Mammalian target of rapamycin complex 1 (mTORC1) activity is associated with phosphorylation of raptor by mTOR.* J Biol Chem, 2009. **284**(22): p. 14693-7.
214. Hara, K., et al., *Raptor, a binding partner of target of rapamycin (TOR), mediates TOR action.* Cell, 2002. **110**(2): p. 177-89.
215. Aleksandrova, K.V., M.L. Vorobev, and I.I. Suvorova, *mTOR pathway occupies a central role in the emergence of latent cancer cells.* Cell Death & Disease, 2024. **15**(2): p. 176.
216. Pradhan, S.K., et al., *EP400 Deposits H3.3 into Promoters and Enhancers during Gene Activation.* Mol Cell, 2016. **61**(1): p. 27-38.
217. Xu, Y., et al., *Histone H2A. Z controls a critical chromatin remodeling step required for DNA double-strand break repair.* Molecular cell, 2012. **48**(5): p. 723-733.
218. Chan, H.M., et al., *The p400 E1A-associated protein is a novel component of the p53→ p21 senescence pathway.* Genes & development, 2005. **19**(2): p. 196-201.
219. Tyteca, S., et al., *Tip60 and p400 are both required for UV-induced apoptosis but play antagonistic roles in cell cycle progression.* The EMBO journal, 2006. **25**(8): p. 1680-1689.
220. Padeken, J., S.P. Methot, and S.M. Gasser, *Establishment of H3K9-methylated heterochromatin and its functions in tissue differentiation and maintenance.* Nat Rev Mol Cell Biol, 2022. **23**(9): p. 623-640.
221. Le Cam, L., et al., *E4F1 is an atypical ubiquitin ligase that modulates p53 effector functions independently of degradation.* Cell, 2006. **127**(4): p. 775-88.
222. Zou, L.H., et al., *TNKS1BP1 functions in DNA double-strand break repair though facilitating DNA-PKcs autophosphorylation dependent on PARP-1.* Oncotarget, 2015. **6**(9): p. 7011-22.
223. Tomiyasu, H., et al., *FOXO1 promotes cancer cell growth through MDM2-mediated p53 degradation.* J Biol Chem, 2024. **300**(4): p. 107209.
224. Hu, J., et al., *Targeted ubiquitination of CDT1 by the DDB1-CUL4A-ROC1 ligase in response to DNA damage.* Nat Cell Biol, 2004. **6**(10): p. 1003-9.
225. Huang, J., et al., *CCDC134 interacts with hADA2a and functions as a regulator of hADA2a in acetyltransferase activity, DNA damage-induced apoptosis and cell cycle arrest.* Histochem Cell Biol, 2012. **138**(1): p. 41-55.
226. Saville, M.K., et al., *Regulation of p53 by the ubiquitin-conjugating enzymes UbcH5B/C in vivo.* J Biol Chem, 2004. **279**(40): p. 42169-81.
227. Ragupathi, A., et al., *Targeting the BRCA1/2 deficient cancer with PARP inhibitors: Clinical outcomes and mechanistic insights.* Front Cell Dev Biol, 2023. **11**: p. 1133472.
228. Senft, D., et al., *Precision Oncology: The Road Ahead.* Trends in Molecular Medicine, 2017. **23**(10): p. 874-898.
229. Megchelenbrink, W., et al., *Synthetic dosage lethality in the human metabolic network is highly predictive of tumor growth and cancer patient survival.* Proc Natl Acad Sci U S A, 2015. **112**(39): p. 12217-22.
230. Pratapa, A., S. Balachandran, and K. Raman, *Fast-SL: an efficient algorithm to identify synthetic lethal sets in metabolic networks.* Bioinformatics, 2015. **31**(20): p. 3299-3305.

231. Wan, F., et al., *EXP2SL: A Machine Learning Framework for Cell-Line-Specific Synthetic Lethality Prediction.* Front Pharmacol, 2020. **11**: p. 112.
232. Hu, L., et al., *A survey on computational models for predicting protein–protein interactions.* Briefings in Bioinformatics, 2021. **22**(5).
233. Zhang, Y., et al. *Predicting synthetic lethal genetic interactions in Saccharomyces cerevisiae using short polypeptide clusters*. in *Proteome Science*. 2012. Springer.
234. Li, J., et al., *Identification of synthetic lethality based on a functional network by using machine learning algorithms.* Journal of cellular biochemistry, 2019. **120**(1): p. 405-416.
235. Das, S., et al., *DiscoverSL: an R package for multi-omic data driven prediction of synthetic lethality in cancers.* Bioinformatics, 2019. **35**(4): p. 701-702.
236. Subramanian, A., et al., *A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles.* Cell, 2017. **171**(6): p. 1437-1452.e17.
237. Wang, J., et al., *Computational methods, databases and tools for synthetic lethality prediction.* Brief Bioinform, 2022. **23**(3).
238. Grinsztajn, L., E. Oyallon, and G. Varoquaux, *Why do tree-based models still outperform deep learning on typical tabular data?* Advances in neural information processing systems, 2022. **35**: p. 507-520.
239. Behan, F.M., et al., *Prioritization of cancer therapeutic targets using CRISPR-Cas9 screens.* Nature, 2019. **568**(7753): p. 511-516.
240. Zhao, B., et al., *A pan-CRISPR analysis of mammalian cell specificity identifies ultra-compact sgRNA subsets for genome-scale experiments.* Nat Commun, 2022. **13**(1): p. 625.
241. Zhu, Y., et al., *TGSA: protein–protein association-based twin graph neural networks for drug response prediction with similarity augmentation.* Bioinformatics, 2021. **38**(2): p. 461-468.
242. Ramchander, N., et al., *Homozygous germ-line mutation of the PMS2 mismatch repair gene: a unique case report of constitutional mismatch repair deficiency (CMMRD).* BMC medical genetics, 2017. **18**: p. 1-8.
243. Pritchard, C.C., et al., *Complex MSH2 and MSH6 mutations in hypermutated microsatellite unstable advanced prostate cancer.* Nature communications, 2014. **5**(1): p. 4988.
244. Martomo, S.A., W.W. Yang, and P.J. Gearhart, *A role for Msh6 but not Msh3 in somatic hypermutation and class switch recombination.* The Journal of experimental medicine, 2004. **200**(1): p. 61-68.
245. Hunter, C., et al., *A hypermutation phenotype and somatic MSH6 mutations in recurrent human malignant gliomas after alkylator chemotherapy.* Cancer research, 2006. **66**(8): p. 3987-3991.
246. Senkin, S., et al., *Geographic variation of mutagenic exposures in kidney cancer genomes.* Nature, 2024. **629**(8013): p. 910-918.
247. Díaz-Gay, M., et al., *The mutagenic forces shaping the genomic landscape of lung cancer in never smokers.* medRxiv, 2024.
248. Wang, Y., et al., *APOBEC mutagenesis is a common process in normal human small intestine.* Nature genetics, 2023. **55**(2): p. 246-254.
249. Li, H.-D., et al., *Polymerase-mediated ultramutagenesis in mice produces diverse cancers with high mutational load.* The Journal of Clinical Investigation, 2018. **128**(9): p. 4179-4191.
250. Jones, D.T., et al., *Dissecting the genomic complexity underlying medulloblastoma.* Nature, 2012. **488**(7409): p. 100-105.
251. Pinto, E.M., et al., *Genomic landscape of paediatric adrenocortical tumours.* Nature communications, 2015. **6**(1): p. 6302.
252. Pathania, A.S., *Immune Microenvironment in Childhood Cancers: Characteristics and Therapeutic Challenges.* Cancers, 2024. **16**(12): p. 2201.
253. Thakur, M.D., et al., *Immune contexture of paediatric cancers.* European Journal of Cancer, 2022. **170**: p. 179-193.
254. Lieberman, N.A., et al., *Characterization of the immune microenvironment of diffuse intrinsic pontine glioma: implications for development of immunotherapy.* Neuro-oncology, 2019. **21**(1): p. 83-94.
255. Lee, M.K., et al., *Associations in cell type-specific hydroxymethylation and transcriptional alterations of pediatric central nervous system tumors.* Nature Communications, 2024. **15**(1): p. 3635.

256. Donehower, L.A., et al., *Integrated Analysis of TP53 Gene and Pathway Alterations in The Cancer Genome Atlas.* Cell Rep, 2019. **28**(5): p. 1370-1384.e5.
257. Li, Q. and M.A. Tainsky, *Higher miRNA tolerance in immortal Li-Fraumeni fibroblasts with abrogated interferon signaling pathway.* Cancer Res, 2011. **71**(1): p. 255-65.
258. Li, Q., et al., *Interferon regulatory factors IRF5 and IRF7 inhibit growth and induce senescence in immortal Li-Fraumeni fibroblasts.* Molecular Cancer Research, 2008. **6**(5): p. 770-784.
259. Kulaeva, O.I., et al., *Epigenetic silencing of multiple interferon pathway genes after cellular immortalization.* Oncogene, 2003. **22**(26): p. 4118-4127.
260. Fridman, A.L., et al., *Expression profiling identifies three pathways altered in cellular immortalization: interferon, cell cycle, and cytoskeleton.* The Journals of Gerontology Series A: Biological Sciences and Medical Sciences, 2006. **61**(9): p. 879-889.
261. Skouras, P., et al., *Advances on Epigenetic Drugs for Pediatric Brain Tumors.* Curr Neuropharmacol, 2023. **21**(7): p. 1519-1535.
262. Zottel, A., et al., *Cytoskeletal proteins as glioblastoma biomarkers and targets for therapy: A systematic review.* Crit Rev Oncol Hematol, 2021. **160**: p. 103283.
263. Kumar, D., et al., *PRMT5 as a Potential Therapeutic Target in MYC-Amplified Medulloblastoma.* Cancers (Basel), 2023. **15**(24).
264. Erazo, T., et al., *TP53 mutations and RNA-binding protein MUSASHI-2 drive resistance to PRMT5-targeted therapy in B-cell lymphoma.* Nature Communications, 2022. **13**(1): p. 5676.
265. Zeng, L.-S., et al., *Overexpressed HDAC4 is associated with poor survival and promotes tumor progression in esophageal carcinoma.* Aging (Albany NY), 2016. **8**(6): p. 1236.
266. Cai, J.-Y., et al., *Histone deacetylase HDAC4 promotes the proliferation and invasion of glioma cells.* International journal of oncology, 2018. **53**(6): p. 2758-2768.
267. Cheng, C., et al., *HDAC4 promotes nasopharyngeal carcinoma progression and serves as a therapeutic target.* Cell death & disease, 2021. **12**(2): p. 137.
268. Wilson, A.J., et al., *HDAC4 promotes growth of colon cancer cells via repression of p21.* Molecular biology of the cell, 2008. **19**(10): p. 4062-4075.
269. Ecker, J., O. Witt, and T. Milde, *Targeting of histone deacetylases in brain tumors.* CNS oncology, 2013. **2**(4): p. 359-376.
270. Bronkhorst, A.J., V. Ungerer, and S. Holdenrieder, *Early detection of cancer using circulating tumor DNA: biological, physiological and analytical considerations.* Crit Rev Clin Lab Sci, 2019. **57**(4): p. 253-269.
271. Wimmer, K., et al., *Diagnostic criteria for constitutional mismatch repair deficiency syndrome: suggestions of the European consortium 'Care for CMMRD' (C4CMMRD).* Journal of Medical Genetics, 2014. **51**(6): p. 355-365.
272. Dharia, N.V., et al., *A first-generation pediatric cancer dependency map.* Nat Genet, 2021. **53**(4): p. 529-538.