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

Etienne Sollier, M.Sc.

Born in Paris, France

Oral Examination: 2nd December 2024

Enhancer hijacking in acute myeloid leukemia with a complex karyotype

Referees

Prof. Dr. Benedikt Brors

Prof. Dr. Christoph Plass

SUMMARY

Acute myeloid leukemia (AML) is a cancer of blood cells, in which hematopoietic progenitor cells have their differentiation impaired and proliferate excessively. This is a heterogeneous disease with many subtypes. While new treatments have been developed for several subtypes, which have led to great improvements in survival, one subtype which still retains a dismal prognosis is AML with a complex karyotype (ckAML). It is defined by the presence of at least three cytogenetic abnormalities and is still poorly understood at the molecular level. Genomic studies have identified that copy number alterations (CNAs) are very frequent in ckAML, especially deletions which occur predominantly in the chromosome arms 5q, 7q, 17p and 12p. However, for the most part, it is still unknown how these deletions might drive the disease.

Here, I hypothesized that the numerous genomic rearrangements in ckAML might, in addition to the CNAs, lead to the aberrant expression of oncogenes by repositioning them in the vicinity of active enhancers. Such "enhancer hijacking" events are known to occur in many cancer types, but in AML only a few genes have been reported to be activated by this mechanism. Thirty-nine ckAML samples were profiled with whole genome sequencing (WGS) and RNA-seq. I analyzed the somatic alterations and the transcriptomes of these samples, and confirmed the high frequency of *TP53* mutations, deletions and chromothripsis events. I developed a tool called pyjacker to systematically search for enhancer hijacking events. This led to the identification of 19 genes activated by structural rearrangements, including known genes like *MECOM*, *BCL11B* and *MNX1*, as well as novel genes like *EPO* and *GSX2*. I further analyzed nanopore sequencing data from a subset of these samples, and found that enhancer hijacking can result in allele-specific alterations in the DNA methylation profile, although these changes are modest.

Taken together, these results show that enhancer hijacking plays a more important role in ckAML than previously thought. Studying the genes activated by this mechanism could lead to novel targeted therapies. Nevertheless, enhancer hijacking is not as recurrent in ckAML as the most common deletions like del(5q) and del(7q), so studying how these deletions drive ckAML might be the most promising route towards improving therapies for ckAML patients.

ZUSAMMENFASSUNG

Akute myeloische Leukämie (AML) ist eine Krebserkrankung der Blutzellen, bei der die hämatopoetischen Vorläuferzellen in ihrer Differenzierung gestört sind und sich übermäßig vermehren. Es handelt sich um eine heterogene Krankheit mit vielen Subtypen. Während für mehrere Subtypen neue Therapien entwickelt wurden, die zu einer erheblichen Verbesserung der Überlebensrate geführt haben, ist AML mit komplexem Karyotyp (ckAML) ein Subtyp, der nach wie vor von einer schlechten Prognose betroffen ist. CkAML ist durch das Vorhandensein von mindestens drei zytogenetischen Anomalien definiert und auf molekularer Ebene noch wenig verstanden. Genomische Studien haben ergeben, dass Kopienzahlveränderungen (CNAs) bei ckAML sehr häufig sind, insbesondere Deletionen, die vorwiegend in den Chromosomenarmen 5q, 7q, 17p und 12p auftreten. Es ist jedoch noch weitgehend unbekannt, wie diese Deletionen die Krankheit auslösen können.

Hier stellte ich die Hypothese auf, dass die zahlreichen Strukturvarianten bei ckAML zusätzlich zu den CNAs zu einer erhöhten Expression von Onkogenen führen könnten, indem sie in der Nähe von aktiven Enhancern neu positioniert werden. Es ist schon bekannt, dass solche "Enhancer-Hijacking" Ereignisse in vielen Krebsarten auftreten, aber in AML sind nur wenige Gene bekannt, die durch diesen Mechanismus aktiviert werden. 39 ckAML-Proben wurden mittels Ganzgenomsequenzierung (WGS) und RNA-seq profiliert. Ich analysierte die somatischen Veränderungen und das Transkriptom dieser Proben und bestätigte die hohe Häufigkeit von *TP53*-Mutationen, Deletionen und Chromothripsis. Ich habe eine Software, "pyjacker", entwickelt, um systematisch nach "Enhancer-Hijacking" Ereignissen zu suchen. Dies führte zur Identifizierung von 19 Genen, die durch Strukturvarianten aktiviert werden, darunter bekannte Gene wie *MECOM*, *BCL11B* oder *MNX1*, aber auch neue Gene wie *EPO* oder *GSX2*. Ich analysierte außerdem Nanopore Sequenzierungsdaten von einigen dieser Proben und stellte fest, dass Enhancer Hijacking zu allelspezifischen Veränderungen im DNA-Methylierungsprofil führen kann, obwohl diese Veränderungen nur geringfügig sind.

Insgesamt zeigen diese Ergebnisse, dass Enhancer Hijacking bei ckAML eine wichtigere Rolle spielt als bisher angenommen. Die Untersuchung der durch diesen Mechanismus aktivierten Gene könnte zu neuen zielgerichteten Therapien führen. Dennoch tritt Enhancer Hijacking bei ckAML nicht so häufig auf wie die häufigsten Deletionen wie del(5q) und del(7q), so dass die Untersuchung, wie diese Deletionen

ckAML antreiben, der vielversprechendste Weg zur Verbesserung der Therapien für ckAML-Patienten sein könnte.

CONTENTS

CHAPTER 1

INTRODUCTION

The goal of this thesis is to investigate to what extent enhancer hijacking drives acute myeloid leukemia with a complex karyotype (ckAML). In the introduction, I will first outline how epigenetic variation allows the numerous cells of the human body to cooperate by serving various roles (Section 1.1), and highlight the role of enhancers in this respect. I will then describe how this multicellular cooperation can break down in cancer (Section 1.2), with various types of somatic alterations that can drive aberrant proliferation. I will then define enhancer hijacking and show its importance in cancer (Section 1.3). Finally, I will summarize the current knowledge on AML (Section 1.4), and in particular on the subtype ckAML that is the focus of this thesis.

1.1 Epigenomics

All cells of a multicellular organism are derived from a single cell which divided numerous times. This initial cell had genetic information encoded as DNA sequences of nucleotide bases A, C, G and T. The human genome is made up of three billion bases (3Gb) split into 23 pairs of chromosomes. DNA replication is very faithful, so all cells of an organism will share the same genome of the initial cell, apart from some somatic alterations to the genome, which can lead to cancer. In spite of this shared genome, different cells of an organism can play very different roles and express different sets of genes. This is possible because they differ in their epigenome: chemical modifications that are added on top of the genome, and that regulate which genes are expressed in a particular cell. Epigenomic marks are inherited through cell division, but they are more plastic than the genome and can change as cells undergo differentiation into specialized cell types. The epigenome is made up of several layers: DNA methylation, chromatin accessibility, histone modifications and 3D chromatin interactions

regulated by insulators. The combination of these epigenetic layers results in an epigenomic profile for each genomic region, which can be characteristic for functional units like enhancers and promoters.

1.1.1 DNA methylation

Cytosine methylation is an important epigenetic mark. In vertebrates, it is only found at cytosines in a CpG context (a C followed by a G). There are 28 million such CpG sites in the human genome, amounting to less than 1% of the genome. Considering that the percentage of G and C bases in the human genome, also known as its GC content, is 40% [1], one would expect the proportion of CpG dinucleotides to be $0.2 * 0.2 = 4\%$, much higher than the actual value. This CpG depletion is likely due to the tendency of methylated cytosines to spontaneously deaminate into thymines [2]. Even though most of the human genome is CpG-poor, CpG islands are an exception: they are regions of about 1kb with a high concentration of CpG sites [3]. They are particularly enriched at gene promoters: 50-70% of promoters have a CpG island, and housekeeping genes (which are expressed in many cell types) are particularly likely to have a CpG island at their promoter [4]. 70-80% of the CpG sites are methylated in the human genome, but most CpG islands are unmethylated, which protects them from deamination.

A first clue concerning the functional role of DNA methylation came from a study by McGhee and Ginder in 1979, where they showed that in chickens, the promoter of the beta-globin gene is only unmethylated in cell types which express it, and methylated in other cell types, indicating that promoter DNA methylation could silence a gene [5]. This was confirmed by later studies, also in humans, which showed that promoter methylation is strongly associated with a silencing of the corresponding gene [6, 7]. Further studies showed that DNA methylation at promoters is really causative for silencing and not merely associated with it, since targeted methylation can silence a gene [8], and DNA demethylation can reactivate genes [9]. Even though promoter methylation is strongly associated with silencing of the corresponding gene, the converse is not true: an unmethylated promoter does not imply that the gene is expressed [6, 7] (Figure 1A). In fact, most promoters are always unmethylated, regardless of their transcriptional state.

DNA methylation is involved in several normal processes, including development, silencing of transposable elements, X-chromosome inactivation and genomic imprinting [10]. For example, the inactive X chromosome shows hypermethylation at pro-

moters, which can take part in the silencing of this chromosome. However, this hypermethylation occurs after the X chromosome has already been inactivated [11], so it cannot be directly responsible for the inactivation. In fact, it appears that in most cases, DNA methylation does not silence an active gene, but merely locks an already inactive gene in a repressed state [12]. Interestingly, DNA methylation within gene bodies is not associated with silencing, and transcribed genes have on average higher methylation within their body [6]. The inactive X chromosome is less methylated than the active one, despite the hypermethylation at promoters, and this difference mainly comes from gene bodies being less methylated in the inactive X chromosome [13].

DNA methylation also has a functional impact outside of genes. For example, it can prevent binding of CTCF, thus altering large-scale 3D chromatin structure (Figure 1B). *H19* and *IGF2* are two oppositely imprinted genes: *IGF2* is expressed from the paternal allele and *H19* from the maternal allele. The two genes can be activated by the same enhancer, but a CTCF binding site lies between them. If the CTCF binding site is methylated, CTCF cannot bind, and the enhancer activates *IGF2* [14, 15]. However, if it is unmethylated, CTCF binds, and the same enhancer instead activates *H19*. More generally, DNA methylation can alter transcription factor binding, which could also play a role at enhancers. Some enhancers appear to be sensitive to DNA methylation, although most are not [16].

Several enzymes are involved in methylating and demethylating DNA [10]. De novo DNA methylation is carried out by DNMT3A and DNMT3B (Figure 1C). DNMT1 is responsible for DNA methylation maintenance, by methylating hemi-methylated CpG sites after DNA duplication (Figure 1D). TET enzymes (TET1, TET2, TET3) can demethylate CpG sites by successively converting methylated cytosines (5mC) to hydroxymethylated (5hmC), formylated (5fC), carboxylated (5caC), and finally unmodified cytosines (Figure 1C). Decitabine and 5-azacytidine are drugs that can lower methylation levels genome-wide by acting as DNMT inhibitors [17]. They are cytidine analogs that can be integrated into DNA (and RNA for 5-azacytidine) in place of cytosines, bind DNMTs and trap them, leading to global DNMT depletion, and methylation loss upon DNA replication.

Several technologies can be used to profile DNA methylation. Methylation arrays profile only a subset of CpGs (850,000 for EPIC arrays) but are rather cheap and can be applied to large cohorts. Whole genome bisulfite sequencing, where the DNA is first treated with bisulfite to convert unmethylated cytosines to uracil, used to be the gold standard for genome-wide DNA methylation profiling. However, the bisulfite treatment results in an important loss of the input material. Nanopore sequencing is be-

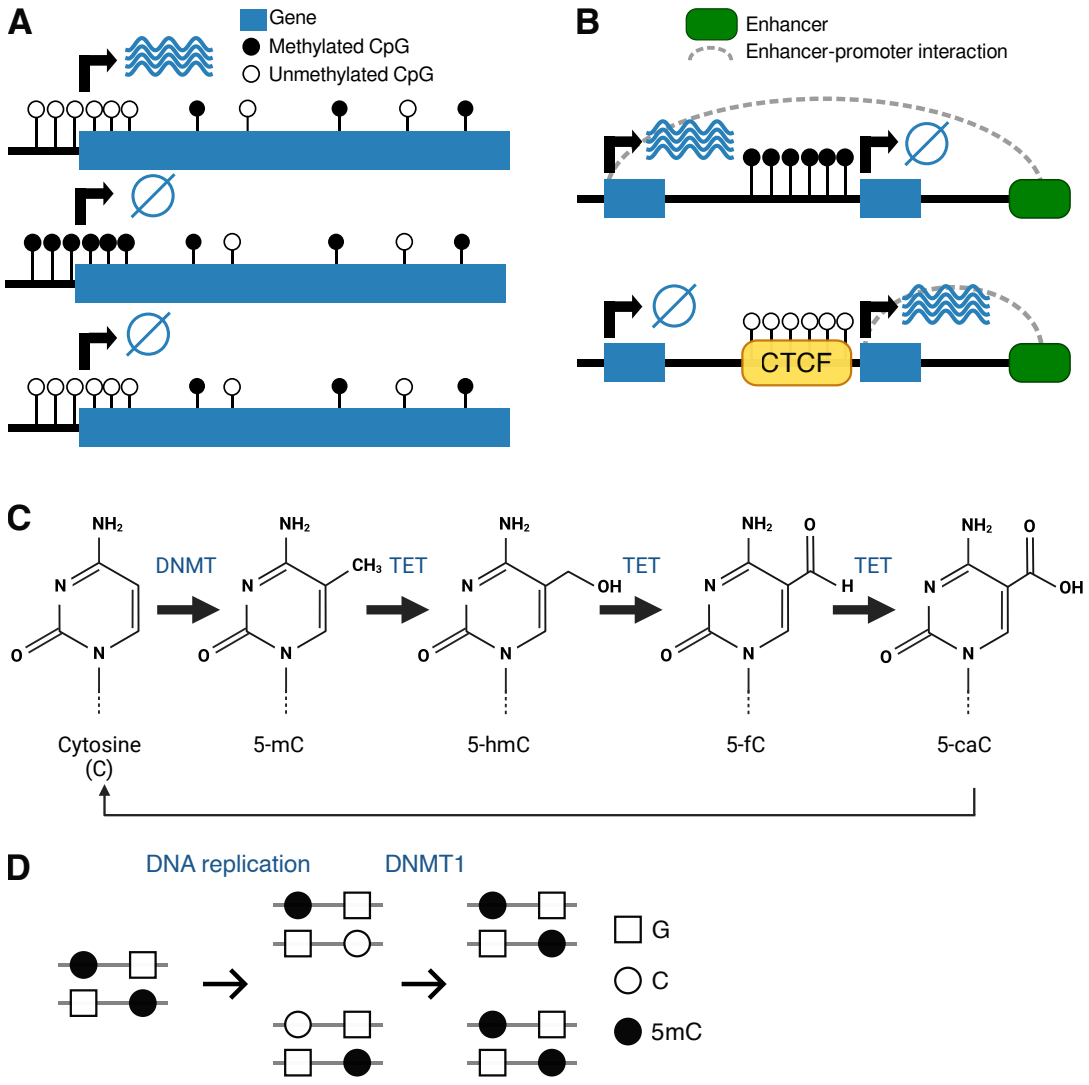

Figure 1: Schematic representation of DNA methylation. **A**. Schematic representation of the role of CpG methylation at promoters: methylation can silence a promoter, but when a promoter is unmethylated, it can be either active or inactive. **B**. Schematic representation of the role of CpG methylation at CTCF binding sites: methylation prevents binding of CTCF, thus altering long range chromatin interactions, and potentially gene expression. **C**. Chemical structure of cytosine and its modified forms, with arrows indicating how one form can be generated from another. **D**. Schematic representation of DNA methylation maintenance during DNA replication with DNMT1.

coming an attractive approach to profile DNA methylation: an electrical current is applied to a nanopore and, as DNA fragments go through the pore, they alter the electric signal. This signal alteration depends on the nucleotides present in the pore, and also on their modifications. Therefore, it can be used to measure DNA methylation without bisulfite treatment [18]. Additionally, nanopore sequencing provides long reads, which often cover several single-nucleotide polymorphisms (SNPs) and can therefore be phased to each of the two parental haplotypes, thus providing allele-specific infor-

mation [19]. Each long read typically covers many CpGs, so this also provides information about intra-read methylation heterogeneity. Remora, the deep learning model used to detect base modification from nanopore sequencing, was trained using DNA amplified by PCR to remove modifications, and optionally treated with the methyltransferase M.Sss1 to add methylation at CpG sites.

1.1.2 Nucleosomes: histone marks and chromatin accessibility

Each human cell has a total linear DNA length of about 2m, which has to fit inside a nucleus with a diameter of 10 µm. It therefore needs to be densely packaged, which is made difficult by the fact that it is negatively charged, so repulses itself. The solution found by nature is to wrap the DNA around positively charged proteins, thus negating the charge [20]. 147bp of DNA are wrapped around a histone octamer (two of each of the four canonical core histones: H2A, H2B, H3 and H4), which forms the nucleosome core, and linker DNA separates two nucleosome cores. The canonical histone proteins can be replaced by histone variants, which give them specific functions [21]. For example, CENPA is an H3 variant found at centromeres which interacts with the kinetochore, enabling the segregation of chromosomes during mitosis [22]. In addition, histones frequently carry post-translational modifications, specifically at their N-terminal tail which protrude out of the nucleosome core and can interact with other nucleosomes, thus regulating chromatin structure [23]. The most studied histone modifications are acetylation and methylation, although many more exist. Acetylation is found on lysines and neutralizes their positive charge. It is added by histone acetyltransferases (HATs) and removed by histone deacetylases (HDACs). Histone methylation is found on lysines and arginines, and can be found as either monomethylation, dimethylation or trimethylation. Different histone modifications are associated with various functions. For example, H3K27ac is found at active promoters and enhancers, H3K4me1 at active enhancers, H3K4me3 at promoters, and H3K27me3 is associated with gene silencing. These histone marks can be profiled with ChIP-seq (chromatin immunoprecipitation followed by sequencing) using an antibody against the specific histone modification [24]. Other methods for profiling histone marks include CUT&RUN [25], CUT&TAG [26] or ACT-seq [27], which require fewer input cells than ChIP-seq.

The density of nucleosomes, their modifications, and the binding of transcription factors determine how accessible the chromatin is in a particular genomic region [28]. Typically, inactive regions are densely packed and inaccessible, while active promoters and enhancers are accessible. Chromatin accessibility is very cell type specific, and

may be better suited than RNA-seq to distinguish cell types. For example, *TET2* is expressed in all hematopoietic cell types, but is activated by several enhancers, each of which is only accessible in some cell types [29]. This chromatin accessibility can be measured by various assays, the most popular one currently being the assay for transposase accessible chromatin with high-throughput sequencing (ATAC-seq) [30]. An alternative method is NOMe-seq [31], which uses a GpC methyltransferase to methylate accessible GpC sites (which are not normally methylated), which allows the simultaneous profiling of accessibility and endogeneous CpG methylation in the same DNA molecules. NOMe can be combined with nanopore sequencing (nanoNOMe), which adds a new layer of information to nanopore data [32].

1.1.3 3D genome organization

The chromosomes that make up our genome consist of linear DNA sequences, but the DNA is folded in the nucleus of the cells, meaning that DNA sequences which are far apart in the linear genome can interact with each other. As a consequence, the expression of a gene can be regulated by regions far away from its promoter, provided that they are able to interact with the promoter.

Chromosome Conformation Capture (3C) can be used to quantify interaction frequencies between two regions [33]. Formaldehyde is used to cross-link segments of DNA which are in spatial proximity. Then, DNA is digested, and re-ligated at low DNA concentration, which favours the ligation of DNA segments which were cross-linked. In 3C, the ligated products are then quantified by PCR, using primers specific to the two regions of interest. 4C allows the quantification of all interactions with one particular region [34]. It can be used to investigate which enhancers interact with a particular promoter, and reciprocally, which promoters interact with a particular enhancer. Hi-C captures all interactions of any region with any other region in the genome, providing the most comprehensive picture of DNA-DNA interactions [35, 36]. Hi-C contact maps can be visualised as heatmaps, where the color of position (x,y) indicates the frequencies of contacts between region x and region y. When looking at whole chromosomes with a resolution of 1Mb, Hi-C data reveals a checker-board pattern, indicating that the genome is divided into two compartments (A and B), and that regions from one compartment preferentially interact with other regions from the same compartment [35] (Figure 2A). When looking at a higher resolution, with bin sizes below 100kb, squares (or triangles) begin to appear, which correspond to topologically associating domains (TADs) [37] (Figure 2B). TADs are megabase-sized self-interacting regions, which are

flanked by CTCF binding sites and are rather conserved across cell types. The current model explaining TADs is that DNA loops are extruded by cohesin until it reaches converging CTCF binding sites. Enhancer-promoter interactions only occur within a TAD. If a high sequencing depth is used, peaks can be detected in Hi-C data. They correspond to one pixel in the heatmap with stronger interactions than in the surrounding area, and are thought to occur because of a loop whose anchors are the two interacting regions corresponding to the pixel [36]. Loops often occur between the two ends of a TAD, and they also often occur between a promoter and its enhancers. CTCF, as well as the cohesin subunits RAD21 and SMC3, are often found at loop anchors, and the CTCF motifs are in convergent orientation [36]. CTCF and cohesin have been shown to be required for loop formation at TAD boundaries, as their depletion results in a loss of TADs [38, 39]. However, active and inactive compartments were not lost upon CTCF or cohesin depletion, and the transcriptional changes were rather small considering the complete loss of TADs.

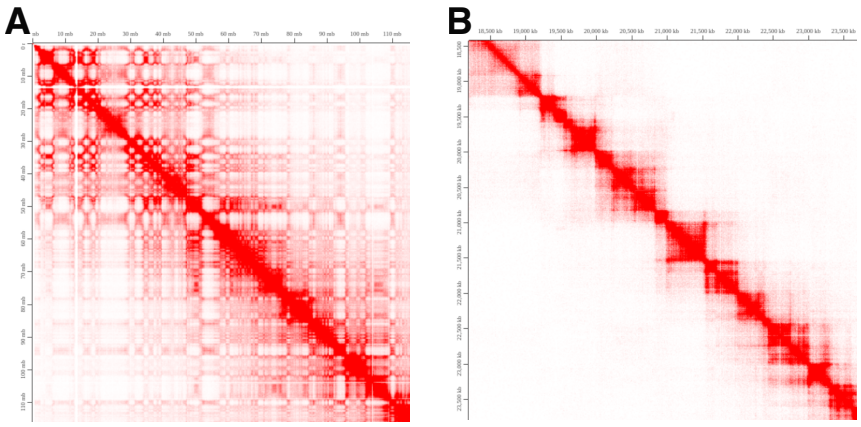

Figure 2: Hi-C data visualizased as heatmap. **A**. "Checkerboard" pattern observed in Hi-C data at a large scale, here the first 120Mb of chr1, for IMR90 data from Rao et al. [36] visualized with the Juicebox web app. **B**. TADs observed in Hi-C data at a smaller scale, here chr1:18-23.5Mb for IMR90 data from Rao et al. [36] visualized with the Juicebox web app.

While Hi-C is the most comprehensive method, it requires a very high sequencing depth. If one is only interested in DNA contacts involving regions marked by a specific protein, it is possible to use HiChIP: contact libraries are created similarly to Hi-C, and ChIP-seq is then performed to select the long range interactions which also interact with a protein of interest [40]. For example, it is possible to use HiChIP with H3K27ac in order to look for enhancer-promoter interactions, since both of these elements harbour H3K27ac [41]. HiChIP against H3K27ac requires less than 10% of the sequencing depth of Hi-C to get similar resolution, which makes it very efficient.

1.1.4 Enhancers

Enhancers are DNA sequences that can increase the transcription of target genes in *cis* (located nearby on the same DNA molecule) [42]. They contain DNA binding motifs for transcription factors. Enhancers can be located far away from their target genes, upstream or downstream, in either orientation. The only requirement is that they must be able to physically interact with their target genes, so they should be located in the same TAD. The first enhancer to be discovered was a 72bp sequence from the genome of the virus SV40 [43], which was found to increase the transcription of the beta-globin gene by 200-fold. Many other enhancers were discovered later, not only in viruses but also in cells, and they vastly outnumber genes: the human genome is estimated to contain hundreds of thousands of enhancers [42]. However, enhancers are very cell type specific, and only a small subset is active in a particular cell type.

Enhancers are defined based on their ability to increase gene expression. This requires experimental validation, by inserting or removing a putative enhancer, in order to really prove that a sequence can increase expression of the associated gene. However, it is possible to predict putative enhancers based on epigenomic profiling. Active promoters reside in open chromatin (e.g. profiled with ATAC-seq), harbour H3K4me1 and H3K27ac marks, and are bound by P300 [42]. Since H3K4me1 can also be found at primed enhancers, which do not harbour H3K27ac and are not yet active, H3K27ac is a popular mark to identify active enhancers. It is also found at promoters, but these can be distinguished from enhancers based on gene annotations, or with H3K4me1/3. P300 is a histone acetyltransferase, which deposits H3K27ac, and was found to be a very good predictor of enhancer activity in ChIP-seq data [44].

A gene is typically regulated by several enhancers. Enhancers do not always activate the gene closest to them. It is unclear why an enhancer would activate a gene and not another one. It was hypothesized that there could be enhancer-promoter compatibilities, but no evidence was found for it [45]. Instead, most enhancers can activate most genes, and they act multiplicatively.

Enhancers are often found in clusters, leading to the definition of super-enhancers. Unlike typical enhancers, which are defined biologically based on their ability to enhance transcription, super-enhancers are defined bioinformatically [46, 47]: peaks are called in ChIP-seq data against H3K27ac, peaks within 12.5kb of each other are stitched together, and they are ranked according to their total H3K27ac signal. When plotting these values, the point where the curve has a slope of 1 is identified, and enhancers to the right of that point are classified as super-enhancers, while the other ones are typical

enhancers. The hierarchy within a super-enhancer is not always clear: are all components independent or do they act synergistically? Are some components essential? For the alpha-globin super-enhancer, the different enhancers seem to act independently, and deletion of a single enhancer does not result in significant downregulation of the gene [48]. For the *Wap* gene, the deletion of a single enhancer in a super-enhancer resulted in a 90% expression reduction, while a deletion of other enhancers had a smaller impact [49]. This indicates that there is an enhancer hierarchy, where some components of a super-enhancer are more important than others, although this cannot be predicted based on ChIP-seq data alone.

1.2 Cancer

1.2.1 A historical perspective: oncogenes and tumor suppressors

In a multi-cellular organism, cells must cooperate for the good of the whole, to the point where a cell should not divide if this is not useful for the organism, and even undergo apoptosis if it is not needed anymore, if it is infected by a virus, or if it failed to repair damage to its DNA. However, if a cell stops cooperating and instead decides to proliferate without restraints, if this rebel behaviour is inherited by its daughter cells, and if the immune system fails to eliminate it, a neoplasm will arise. This aberrant cell proliferation may be benign, but it will become cancerous if it spreads and invades other tissues. For this uncontrolled proliferation to be inheritable through cell division, it must be encoded genetically, as a result of a somatic genomic alteration. Theodor Boveri was the first to boldly postulate in 1914 that cancer may be caused by alterations to the chromosome [50]. At the time, it was still unknown that chromosomes carried hereditary information, but by observing chromosomes through a microscope, he noticed that cancer cells had scrambled chromosomes, leading to his visionary idea. However, it took many years for the proof to come and for this idea to be accepted. Indeed, this theory was at odds with another hypothesis which was gaining traction at the same time: that cancer was caused by viruses, and not by something internal to the cell. In 1910, Peyton Rous discovered that a virus, now known as the Rous sarcoma virus (RSV), could cause cancer in chickens [51]. In hindsight, this viral cause of cancer was somewhat of a false lead, since only a small minority of cancers are actually caused by viruses, but the RSV still led to the identification of the first oncogene, and cancer-causing viruses in general were instrumental in many discoveries in cancer research. By studying the RSV, researchers found that it is a retrovirus, and that

its genome contains a gene, *v-Src*, which is reponsible for tumor induction [52], and was therefore called an oncogene. It was later found that *v-Src* had high homology to a gene found in the genome of normal chicken cells, called *c-Src* for cellular Src [53]. *c-Src* is a tyrosine kinase [54] which integrates various signals in the cell, and which can, when activated, trigger cell proliferation. *v-Src* is actually a mutated version of *c-Src*, which is constitutively active and always stimulates proliferation, regardless of external signals. This encouraged the field to come back to Boveri's idea: that cancer could be caused by alterations to the chromosomes. c-Src is a proto-oncogene: a gene found in the genome of normal cells which, when mutated or overexpressed, can become an oncogene and drive cancer. Since then, many other oncogenes have been discovered, like *MYC* and *RAS*, which all start as proto-oncogenes in the normal genome, and which get mutated in cancer, resulting in a fitness advantage for the cells having the mutation. However, the activation of a single oncogene is not sufficient to drive cancer, and a multi-step process involving several somatic events is necessary for malignant progression [55].

In 1971, Knudson postulated that retinoblastomas could be initiated by only two mutational events [56]. This prediction was based on statistical modelling, starting from the observation that retinoblastomas come in two distinct forms: familial and sporadic cases. In familial cases, several individuals from the same family have the disease and they develop it early in life, often with several independent tumors. Conversely, sporadic cases are developed much later in life, with a single tumor arising in each individual. This can be explained by the fact that familial cases already inherited one mutation, and only need one further somatic event to initiate tumor growth, while sporadic cases can only occur if two somatic events occur in the same cell lineage, which is much rarer and takes more time. Further studies identified that the two "hits" postulated by Knudson actually occur on the two alleles of the same gene *RB1*, and completely inactivate it [57]. The normal role of the retinoblastoma protein is to halt the cell cycle at the G1 checkpoint until the cell is ready to divide, and when both copies are no longer present, this control is lost. *RB1* was the first tumor suppressor gene (TSG) to be identified, and since then others have been discovered, which must also be biallelically inactivated in order to drive cancer, like *TP53* or *APC*. *RB1* is mutated in several cancer types, not just retinoblastomas, but *TP53* is the most commonly mutated gene across all cancer types [58]. TP53 was first identified as a 53KDa-protein which binds to an antigen of the simian virus SV40 (another tumor-inducing virus) [59, 60]. TP53 was also found in high amounts in other cancer cells, but absent or very lowly expressed in normal cells, and it was shown to induce cancer formation [61], leading to the inital hypothesis that it was an oncogene. However, sequencing of the gene revealed that

the TP53 found in cancer cells and which could induce transformation was a mutated version of the wild-type gene [58]. In 1989, the wild-type TP53 was shown to inhibit tumor formation [62], and it was also found that both *TP53* alleles were inactivated in colorectal cancer [63], typically through a combination of a deletion and a mutation. This made it clear that *TP53* is a TSG, which protects against cancer, but can be biallelically inactivated in cancer.

1.2.2 Cancer in the sequencing era

It is now well established that cancer is a multi-step process which requires several somatic mutational events. Some events are seen across many cancer types, like *TP53* or *KRAS*, but many are specific to certain tumor entities. With next generation sequencing, we now have a clear picture of the somatic alterations that occur in cancer. Several types of alterations can be detected, and each of these types of events can drive cancer in various ways.

1.2.2.1 Single nucleotide variants and indels

The most simple type of somatic alteration to the genome is a single nucleotide variant (SNV), which is when one nucleotide is changed to one of the three other bases. Slightly more complex, indels are small insertions or deletions of less than 1kb. The number of somatic SNVs found in tumor cells is very variable, both within a cancer type and across cancer types, ranging from less than 1 per megabase in leukemias to more than 10 per megabase in lung and skin cancers [64]. Some mutagens create specific types of mutations, and it is possible to computationally infer the mutational signatures which resulted in a particular set of mutations [64]. For example, tobacco smoking results in a high rate of C>A mutations and exposure to UV light leads to C>T mutations (specifically CC>TT). Some SNVs will be driver events, for example by activating an oncogene or inactivating a TSG, but the majority of them will be passenger events with a neutral effect. An SNV can be coding, if it occurs in a coding portion of a gene, or non-coding otherwise. Coding SNVs are more studied since they are more likely to be driver events. Among coding SNVs, some are synonymous, meaning that they change the nucleotide sequence but leave the amino acid sequence unchanged, and they are likely passenger events. Potential drivers are missense mutations which change an amino acid, nonsense mutations which lead to an early stop, and variants which affect splice sites. The pattern of SNVs differs between oncogenes and TSGs: since there are many more ways to inactivate a protein than to grant it a new function,

mutations in oncogenes are typically only found at "hotspots", where the exact same mutation occurs in many samples, whereas SNVs in TSGs are typically spread throughout the gene, and may be nonsense or frameshifts (when the number of deleted or added bases in an indel is not a multiple of three). Mutations outside of genes are more difficult to study. Many of them are likely passenger events, but some have been identified as driver events, for example those which result in creation or removal of transcription factor binding sites. A well-known example is the activation of *TERT* by mutations in its promoter [65]. Similarly, in T-ALL, micro-insertions upstream of *TAL-1* can create a neo-enhancer, which drives aberrant expression of *TAL-1* [66].

1.2.2.2 Structural variants and copy number alterations

Structural variants alter the structure of chromosomes. They include balanced translocations, inversions, deletions and duplications (Figure 3A). They can be detected from whole-genome sequencing (WGS) data with split-reads, where two different segments of a read are aligned to different regions of the reference genome. In the case of paired-end sequencing, they can also be detected with split pairs, where two reads from a pair map to different regions. Several bioinformatic tools can detect SVs from WGS data, including manta [67], GRIDSS [68] and DELLY [69]. Short-read data is sufficient to detect most SVs, but some regions contain repetitive sequences where the mapping of short reads is difficult, and SVs occurring in these regions can be missed. Long reads can help to identify such SVs, and new SV callers such as Sniffles2 [70] are being developed specifically for long-read data.

Another type of genomic aberration, which often co-occurs with structural variants, is a copy number alteration (CNA). This is when a genomic region is lost or duplicated, leading to an altered number of copies for this region. CNAs can affect whole chromosomes (e.g. monosomy, trisomy) or smaller portions of the genome. In WGS data, CNAs are detected based on the coverage of genomic regions: the coverage is expected to be uniform across the genome, and regions with lower (resp. higher) coverage have a lower (resp. higher) copy number. The allelic imbalance can also be taken into account, since a deletion would for example lead to loss of heterozygosity. WGS is best suited for the detection of CNAs, but methylation arrays, SNP arrays or whole exome sequencing can also be used for this purpose, albeit with a lower resolution. Unlike for SVs, short reads are actually better than long reads for CNA calling, because the precision depends on the number of reads rather than on the coverage. Tools to detect CNAs from short-read WGS data include Control-FREEC [71] and the HMF pipeline which uses AMBER, COBALT and PURPLE (`https://github.com/hartwigmedica`

`l/hmftools`). ASCAT can detect allele-specific CNAs from SNP array data [72], and has been extended to sequencing data. Conumee calls CNAs from methylation array data [73].

Structural variants can drive cancer through different mechanisms. Breakpoints can create oncogenic fusion proteins, or activate silent oncogenes by enhancer hijacking. Deletions can contribute to the inactivation of a TSG (usually in complement to a mutation, but occasionally through biallelic deletions), or to haploinsufficiency if only one copy is lost and if the cells cannot compensate by increasing the transcription from the remaining allele. Finally, amplifications can lead to increased expression of oncogenes.

1.2.2.3 Chromothripsis

SVs can be simple, but they may also be part of very complex events with a high number of breakpoints. The most extreme example is chromothripsis, which is a single catastrophic event where one or several chromosomes are shattered into tens, hundreds, or even thousands of fragments, and are then randomly pieced back together, with some parts being lost and the others completely reshuffled (Figure 3B). This was first discovered in chronic lymphocytic leukemia [74], but was later found to occur in most cancer types, with a prevalence of about 50% across all cancers [75, 76]. Chromothripsis is thought to be caused by DNA damage in micronuclei, or by shattering of a dicentric chromosome. Several criteria have to be met to ensure that the genomic rearrangements observed were caused by a single chromothriptic event [77]. The main criterion is the presence of copy number oscillations, which result from the loss of some genomic regions not integrated into the derivative chromosome. The exact threshold is somewhat arbitrary, but typically a minimum of ten copy number oscillations is used. This criterion can be assessed with many data types, including SNP arrays, and is sometimes used as only criterion for chromothripsis when other criteria cannot be assessed. This, however, can result in false positives, in cases where many deletions happened successively. Other criteria, which require WGS data, include the clustering of breakpoints and the randomness of fragment joins. Shatterseek can be used to assess these criteria [75].

Chromothripsis directly leads to deletions, for the parts that are not integrated into the reshuffled chromosome, but may also indirectly promote amplifications. One mechanism for amplification is through extrachromosomal circular DNA (eccDNA). After chromothripsis, some DNA fragments might, instead of joining the reshuffled chromosome, be fused as a small circle of less than a few megabases. These eccDNA lack a cen-

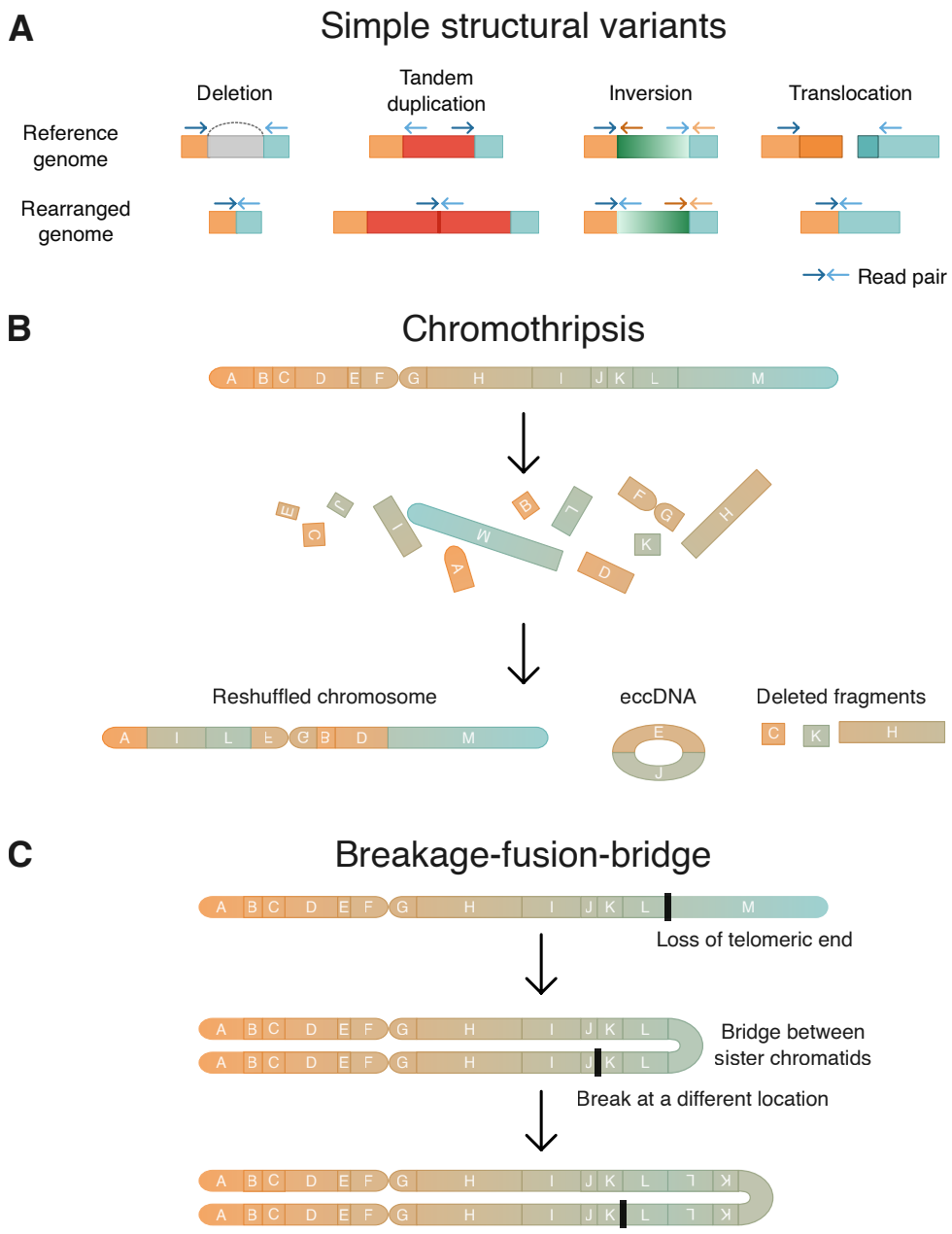

Figure 3: Schematic representation of simple structural variants (**A**), chromothripsis (**B**), and BFB cycles (**C**).

tromere, and are randomly assigned to each of the two daughter cells during mitosis. If a higher copy number of the eccDNA provides a fitness advantage, then the daughter cell which inherits a higher copy number will be selected, which can progressively lead to massive amplifications. The eccDNA can either stay extrachromosomal, or be reintegrated into the genome. Another mechanism is breakage-fusion-bridge (BFB) cycle. When a chromosome loses its telomeric end because of a DNA break, there is a

chance that, during mitosis, the two sister chromatids which lack the telomere might fuse together. This is because cells generally try to repair DNA breaks, unless they are protected by a telomere. When the two sister chromatids are pulled towards opposite poles for cell division, this bridge will break, but not necessarily at the exact position where the two chromatids had fused. Consequently, one cell will inherit a larger part of the chromosome, with a duplicated fragment, while the other cell will have a deletion (Figure 3C). If the duplication provides a fitness advantage to the cell, it will be selected for. The cell which inherited the longer chromosome still lacks a telomere for this chromosome, so other BFB cycles can occur successively, leading to massive amplifications. BFB amplification can be detected in sequencing data with foldback inversions: at amplification boundaries, a read goes in one direction, and then "turns" to go back in the other direction. BFB can occur after chromothripsis, if the reshuffled chromosome lacks a centromere, but it can also trigger chromothripsis, if the chromosomes are shattered when they are pulled apart. The relative order of chromothripsis and BFB amplifications can potentially be reconstructed [78]: if chromothripsis occurs on a normal chromosome, the breakpoints link regions which have the same copy number. If BFB amplifications occur later, regions which had been linked together by chromothripsis will still have the same copy number. On the other hand, if chromothripsis occurs after the amplifications, the breakpoints will link segments with a different copy number. Hence, looking at copy number jumps across breakpoints can inform the ordering of events.

1.2.2.4 Epigenetic alterations in cancer

In addition to genomic alterations, cancer cells exhibit epigenetic alterations in comparison to their healthy counterparts, for example at the methylation level. In 1983, researchers already observed that cancer cells had a global hypomethylation [79]. This global hypomethylation is now mostly attributed to a methylation loss in partially methylated domains (PMDs) [80]. These are megabase-sized regions which are already partially methylated in normal cells, and which lose methylation in cancer cells. They cover close to 50% of the genome, mostly in gene-sparse, lamina-associated, late-replicating regions [81]. The hypomethylation of PMDs is thought to be due to imperfect methylation maintenance, and it is unclear whether it is a passenger or a driver event.

Another type of epigenetic alteration in cancer for which the role is much clearer is hypermethylation of gene promoters. This can silence a TSG, potentially in combination to another "hit" on the other allele. This was first observed in renal carcinoma, where

the promoter of the TSG *VHL* was methylated on one allele in some samples, with the other allele being deleted or mutated [82]. Since then, many TSG have been found to be hypermethylated in some cancer samples, including *CDKN2A* and *BRCA1* [83]. However, it is not clear whether hypermethylation is causal, or whether the gene is silenced by a different mechanism and the methylation simply occurs later to lock the gene in the inactive state. An important thing to note is that promoter hypermethylation is much more pronounced in cancer cell lines that in primary cancer samples [84, 85]. This must be taken into account when analyzing DNA methylation from cell lines, as their methylation may not faithfully represent the methylation patterns of the cancers from which they were derived.

1.3 Enhancer hijacking

When a structural rearrangement brings an active enhancer close to a gene that is not normally expressed in this tissue, the gene can "hijack" this enhancer, leading to its aberrant expression (Figure 4). This was first reported in Burkitt's lymphoma, where a translocation t(8;14) brings the *IGH* enhancer close to the *MYC* oncogene and activates it [86]. Since then, many oncogenes have been reported to be activated by enhancer hijacking in various cancer types, including *GFI1* in medulloblastoma [87], *TERT* in neuroblastoma [88], and *IRS4* in lung cancer and sarcomas [89]. In AML, the most recurrent enhancer hijacking event is *MECOM* activation, typically though an inv(3) or a t(3;3) [90], but other genes have been reported to be activated in this manner, such as *BCL11B* [91] and *MNX1* [92].

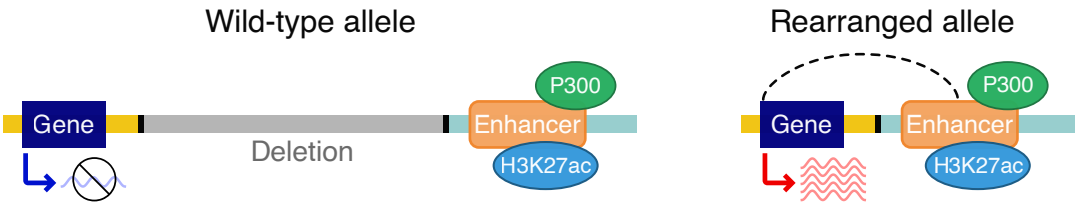

Figure 4: Schematic representation of an enhancer hijacking event.

Several tools have been developed to detect enhancer hijacking events. Cis-Expression Structural Alteration Mapping (CESAM) performs a linear regression on gene expression based on the presence of a breakpoint nearby [89]. The breakpoints are inferred from SNP array data, which has the advantage of being a data type widely available from large cohorts, but cannot detect balanced translocations and inversions which do

not result in CNAs. CESAM successfully identified oncogenes activated by enhancer hijacking, like *IRS4* and *IGF2*, but an important limitation is that it can only detect events which are very recurrent, and not genes activated by enhancer hijacking in a single sample. HYENA [93] is similar to CESAM, in that it uses a regression of the gene expression based on the presence of breakpoints and can only detect recurrent events. Cis-X follows a different approach: it identifies *cis*-activated genes based on overexpression and monoallelic expression, using WGS and RNA-seq data [94]. For the overexpression, cis-X requires gene expression from a reference cohort of the same cell type, where only samples with biallelic expression are included, and tests for overexpression by comparing the gene expression of a tested sample to this reference cohort. Most genes are expected to be biallelically expressed, so the read counts for each allele should be similar, whereas if a gene is activated by a somatic event in *cis*, only one allele should be expressed. Cis-X compares the observed allelic read counts to a binomial distribution where the probability of observing each allele is the same. A major advantage of cis-X is that it can detect genes activated in a single sample. In addition, cis-X does not only detect genes activated by structural rearrangements, but it also analyses somatic SNVs and indels to detect new transcriptor factor binding sites, which might also lead to the activation of a gene in *cis*. While this somatic analysis of SNVs and indels might capture more events than an analysis only including SVs, this step is not optional in cis-X and it requires matched normal samples, which are rarely available for AML samples, thus reducing the applicability of this method. NeoLoopFinder is a completely different method, which takes as input Hi-C data instead of expression and breakpoints [95]. It is based on the detection of neo-loops, which are peaks in Hi-C data between regions brought together by a structural rearrangement. An important limitation is that Hi-C data is not as widely available as RNA-seq and WGS, thus precluding large-scale screens. In addition, the absence of gene expression in the method may result in many false positives, with neo-loops not being functionally relevant.

1.4 Acute myeloid leukemia

1.4.1 AML as a hematological malignancy

The blood is made up of many different cell types. Hematopoietic stem cells (HSCs), which can self-renew or give rise to more differentiated progenitors. HSCs can give rise to cells of the two main hematopoietic lineages: myeloid and lymphoid. Myeloid cells include monocytes, red blood cells and megakaryocytes, while the lymphoid lineage is

made up of natural killer (NK) cells, T-cells and B-cells (Figure 5). Several types of cancers can develop from myeloid cells. The most threatening is acute myeloid leukemia (AML), which is defined clinically by the presence of more than 20% of blasts (undifferentiated progenitors) in the blood or bone marrow. AML is very heterogeneous, with different subtypes associated with very different prognoses. Myelodisplastic syndrome (MDS) is milder than AML, defined by the presence of less than 20% blasts in the bone marrow, but can progress to AML. Myeloproliferative neoplasms (MPN) correspond to an abnormal proliferation of myeloid cells, but without block in differentiation, often associated with somatic mutations in *JAK2*, *CALR*, *TET2* or *MPL*. Chronic myeloid leukemia (CML) is now considered as an MPN. It is defined genetically by the presence of the Philadelphia chromosome: a translocation t(9;22) resulting in the BCR-ABL1 fusion protein [96], a constitutively active tyrosine kinase which drives proliferation. Today, CML can be well treated with tyrosine kinase inhibitors [97]. Recently, sequencing of large cohorts has uncovered that clonal evolution can occur in HSCs even in the absence of symptoms: HSCs acquire somatic mutations which grant them a fitness advantage, leading to the growth of a clone with this mutation. This is known as clonal hematopoiesis of indeterminate potential (CHIP), is more frequent in older individuals, is associated with an increased risk of leukemia, and is often driven by mutations in *DNMT3A*, *TET2*, or *ASXL1* [98].

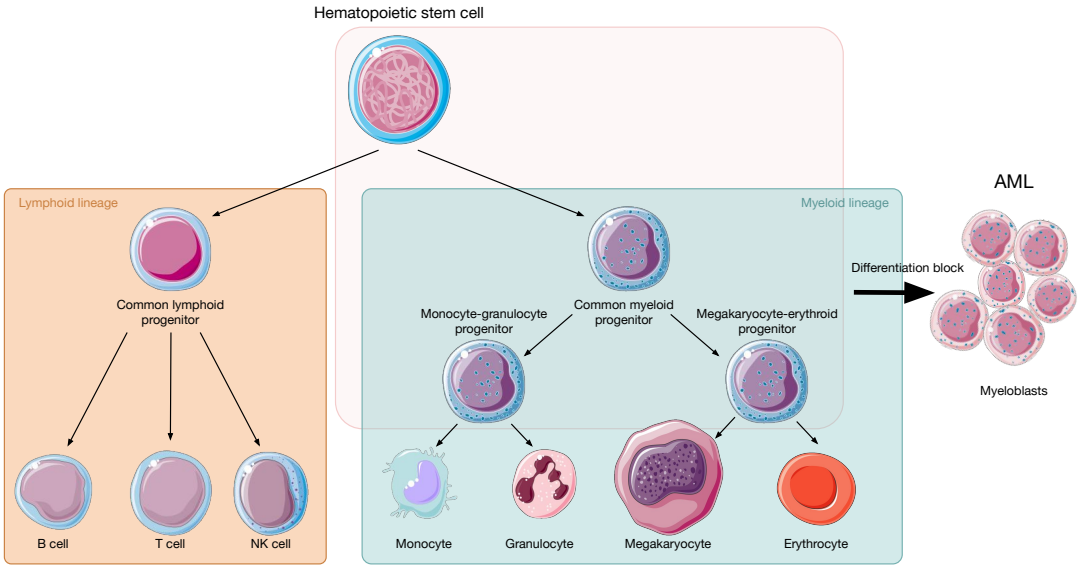

Figure 5: Simplified diagram of hematopoiesis showing the main cell types, with possible block in differentiation leading to AML. Cell type images were downloaded from bioicons.com, and had originally been made by Servier.

1.4.2 Genomic landscape of AML

The clinical classification in AML used to be based on the French-American-British (FAB) classification, which relied on the determination of the cell of origin with immunophenotyping [99]. More recently, the World Health Organization (WHO) provided a new classification, which was last updated in 2022 and which is based on genetic alterations, or on differentiation if no genetic classification is available [100]. The genetic classification is based on recurrent translocations and mutations: t(8;21) with RUNX1::RUNX1T1, inv(16) with CBFB::MYH11, t(15;17) with PML::RARA, inv(3)(q21q26.2) or t(3;3)(q21;q26.2) with MECOM expression, *NPM1* mutation, *RUNX1* mutation, etc. These different subgroups have very different prognoses, e.g. inv(16) and t(15;17) have rather good survival while patients with *MECOM* rearrangements have much poorer prognosis.

Recent sequencing endeavours have attempted to refine these subgroups by analyzing more somatic alterations [101, 102]. AML has fewer somatic alterations than most other cancer types, and the most commonly mutated genes are *FLT3*, *NPM1*, *DNMT3A*, *NRAS* and *KRAS*, *TET2*, *IDH1/2*, *CEBPA*, *PTPN11*, *TP53* and *SRSF2*. There are patterns of co-occurrence and mutual exclusivity between the frequent somatic alterations, for example mutations in *FLT3*, *NPM1*, *DNMT3A* tend to co-occur together. *TP53* mutations co-occur with many copy number alterations, and this corresponds to the subtype of AML with a complex karyotype.

1.4.3 AML with a complex karyotype

AML with a complex karyotype (ckAML) is defined by the presence of at least three cytogenetically detectable alterations, in the absence of other class-defining genetic alterations. It accounts for 10-15% of all AML cases, is more frequent among older patients and carries a very poor prognosis with a 3-year survival of only 12% [103]. 60% of ckAML samples harbour mutations in *TP53*, almost always with biallelic inactivation, and this is associated with a higher number of CNAs and a worse survival [103]. Chromothripsis is found in 35% of all ckAML samples, almost always co-occurring with *TP53* inactivation [104].

Losses are more common than gains in ckAML, and the most common deletions occur in 5q, 7q, 17p and 12p [105]. Deletions in cancer typically lead to the complete inactivation of a TSG, in combination with a mutation on the remaining allele. Such TSGs are traditionally identified by mapping the minimally deleted regions across many sam-

ples, and screening for mutations in genes located in these minimally deleted regions. *TP53* is almost always mutated in cases with del(17p), leading to biallelic inactivation [103]. However, the other frequently deleted regions like 5q and 7q do not harbour such clear TSGs.

Del(5q) is very frequent in ckAML where it co-occurs with many CNAs, but it is also very common in MDS, typically as a sole abnormality, in which case it is associated with a good prognosis [106]. Some genes in the minimally deleted region are sporadically mutated on the remaining allele, like *CSNK1A1*, *KDM3B* or *G3BP1*, but only in a small percentage of del(5q) samples [107, 108, 109]. Since no gene is frequently biallelically inactivated in del(5q), the prevailing hypothesis is that del(5q) is selected for because it leads to haploinsufficiency of some genes in the deleted region. Partial loss of function of *RPS14*, located within the minimally deleted region, was shown to recapitulate the del(5q) phenotype with impaired erythroid differentation [110]. Haploinsufficiency of *EGR1*, also located in the minimally deleted region, was shown to favour leukemia development in mice [111].

Del(7q) is also very frequent in AML, both in ckAML or as a sole abnormality [112]. Several genes in the minimally deleted region are mutated in some del(7q) cases: *KMT2C* in 16% of cases, *EZH2* in 10% and *CUX1* in 5% [113]. In total, approximately one third of del(7q) cases have a concomitant mutation in at least one gene of the deleted region. However, since the majority of del(7q) cases do not harbour any mutations in the deleted region, haploinsufficiency of some genes in 7q might already be sufficient to drive leukemogenesis, and this effect might be increased by mutations.

The molecular mechanisms driving ckAML remain therefore unclear, although the most likely explanation for the CNAs is that they drive the disease by reducing or increasing the expression of genes in the deleted or gained regions.

CHAPTER 2

AIMS

Complex karyotype AML has a very poor prognosis and is still poorly understood at the molecular level. Until now, it had mainly been studied through cytogenetics, SNP arrays and targeted sequencing, which do not provide a complete picture of the rearrangements occurring in these samples. The general aim of this thesis is to investigate how more advanced sequencing assays like WGS, RNA-seq and nanopore sequencing might provide more information and shed light on the mechanisms driving the disease. The current prevailing paradigm in ckAML is that the disease is driven by CNAs, mainly deletions, which can be detected through SNP arrays. However, the SVs may also have a driver effect through the elements that are joined together, rather than through regions being gained or lost, for example with enhancer hijacking.

2.1 Comprehensively analyzing molecular alterations in ckAML

Thirty-nine ckAML samples were profiled with WGS and RNA-seq, and I aimed to use these data to answer the following questions:

- What are the recurrent genomic alterations in ckAML?
- Do the genomic alterations in ckAML differ between younger and older patients?
- Are there patterns of mutual exclusivity or co-occurrence among the recurrent alterations?
- Can ckAML be subdivided into several subtypes, based on genomics or transcriptomics?

- Can WGS provide novel insights into chromothripsis in ckAML?
- How relevant are fusion transcripts in ckAML?

2.2 Identifying novel oncogenes activated by enhancer hijacking in ckAML

The division of Cancer Epigenomics at DKFZ recently identified that *MNX1* was a gene activated by enhancer hijacking in some AML cases with del(7q), which led to the hypothesis that some of the rearrangements in ckAML could lead to enhancer hijacking, in addition to haploinsufficiency of genes in the deleted regions. Therefore, I aimed to:

- Develop a pipeline to systematically search for enhancer hijacking events, including rare ones.
- Apply this pipeline to a ckAML cohort.
- Analyze the most interesting events and understand how they are activated and how they may drive the disease.

2.3 Investigating allele-specific methylation with nanopore sequencing

Nanopore sequencing is a third generation sequencing technology which provides long reads with methylation information, thus enabling the study of allele-specific methylation, which could be relevant for enhancer hijacking. Here, I aimed to answer the following questions:

- Can nanopore sequencing be used to study allele-specific methylation?
- Does enhancer hijacking lead to allele-specific methylation, and if yes, can it be used as a criterion to detect enhancer hijacking?

CHAPTER 3

RESULTS

3.1 Genomic and transcriptomic landscape of ckAML

CkAML is characterized by a strong genomic complexity, with many CNAs, most of which are deletions, and frequent *TP53* mutations and chromothripsis events. To date, ckAML has mainly been studied with SNP arrays, which provide information about CNAs, but not about breakpoints, so WGS data could provide new information. In this project, 39 ckAML samples were profiled with WGS and RNA-seq (sequencing performed by Anna Riedel). They were part of the ASTRAL-1 clinical trial [114, 115] (see methods for details). I analyzed this data in order to get a more complete picture of the genomic and transcriptomic landscape of ckAML. In addition, I collected SNP and methylation array data from several ckAML cohorts, resulting in the largest dataset of CNAs in ckAML ever generated, which I leveraged to gain new insights into how these CNAs may drive ckAML.

3.1.1 Genomic alterations in 39 ckAML samples profiled with WGS

I processed the WGS data and identified somatic alterations. Since no matched normal samples were available for this cohort, I only considered SNVs in genes known to be mutated in AML, and used a panel of normal samples to filter out germline SVs (see Section 5.2 for details about the processing of the WGS data). The landscape of somatic genomic alterations in this cohort matched what had previously been reported in ckAML [103] with frequent deletions in 5q (69.2%, N=27) and 7q (66.7%, N=26) and *TP53* mutations (61.5%, N=24) (Figure 6A). *TP53* was by far the most frequently mutated gene, followed by *PTPN11*, *SRSF2* and *RUNX1* which were each mutated in less than 20% of samples (Figure 6B). Apart from one sample, all samples with a *TP53* mu-

tation also had a second inactivating event: 9 with a deletion of the other allele, 8 with a copy number neutral loss of heterozygosity (CNLOH), 5 with a second mutation, and one with a breakpoint within *TP53* (Figure 6A). Interestingly, one sample (16PB3075) had no mutation in *TP53*, but two breakpoints within the gene, which probably resulted from a single chromoplexy event involving several other chromosomes. Samples with *TP53* inactivation had more CNAs (Figure 6C), in agreement with previous reports [103].

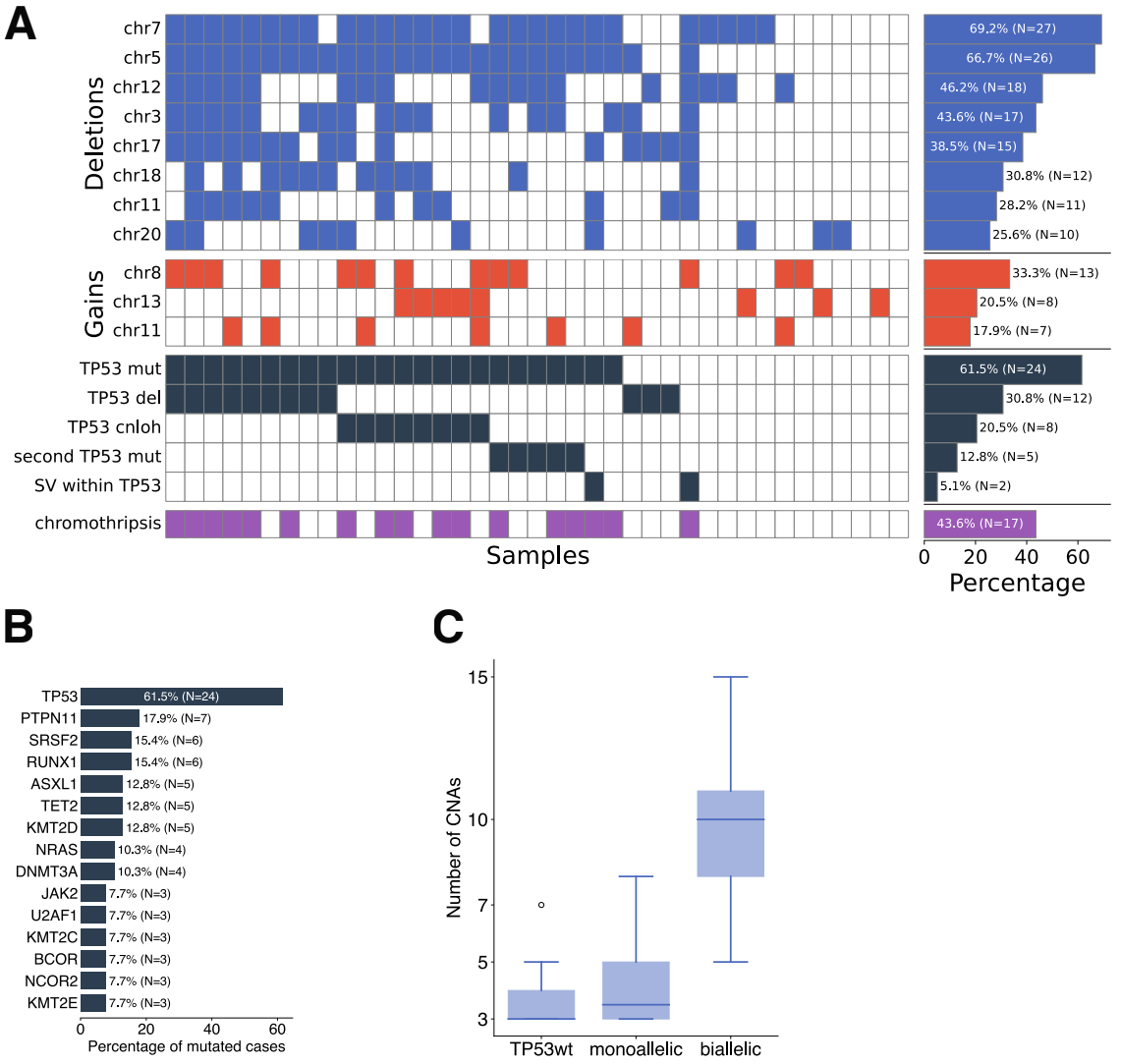

Figure 6: Summary of the somatic alterations in the ckAML cohort. **A**. Oncoplot showing the most frequent somatic alterations in each sample. **B**. Frequency of SNVs among the 39 ckAML samples. **C**. Number of CNAs in samples without *TP53* alterations (TP53wt), with one alteration (monoallelic), or with both alleles disrupted (biallelic).

Chromothripsis

I detected chromothripsis events in 17/39 samples (see Section 5.2 for the criteria used), all of which had biallelic *TP53* inactivation. Chromothripsis affected various chromosomes: 13/24 chromosomes were affected by chromothripsis in at least one sample (Figure 7A), and chr3 was most commonly affected (7 samples), followed by chr12 (4 samples) and chr17 (3 samples). Rücker et al. had also found that chromothripsis could affect various chromosomes, although in their cohort chr7 was the chromosome most commonly affected by chromothripsis [104]. Some samples have several chromosomes affected by chromothripsis. SNP arrays only provide CNA information but no SVs, so they cannot be used to distinguish whether one chromothripsis event affected several chromosomes (resulting in many breakpoints between the chromosomes), or whether several chromothripsis events occurred independently. With WGS data, I found that only a single sample had a chromothriptic chromosome with only intra-chromosomal breakpoints (chromosome 3 in sample 15PB19457). All other chromothriptic chromosomes had breakpoints leading to other chromosomes, although these other chromosomes did not necessarily harbour massive rearrangements too. I observed very complex chromothriptic events involving several chromosomes (Figure 7B). Several samples had several independent chromothripsis events, some of which were subclonal (Figure 7C-E).

Chromothripsis can result in amplifications through eccDNA or BFB cycles. In this cohort, I found several samples with copy numbers >=4 in some regions, especially on chr11, chr13 and chr21 (Figure 7C-D for example). However, only one sample had a region with a copy number greater than 10: sample 15KM18875 had complex rearrangements on chr19 with massive amplifications, up to a copy number of 30 around *EPOR* (Figure 8A). This sample contained foldback inversions at amplification boundaries (Figure 8B), so these amplifications were likely due to BFB cycles. Analysis of copy number jumps at breakpoints revealed that many breakpoints joined segments with different copy numbers (Figure 8C), indicating that chromothripsis probably occurred after the BFB amplifications, since if chromothripsis had occurred first, both ends of most breakpoints should have the same copy number. However, it is possible that the chromosome was shattered several times, both before and after the amplifications.

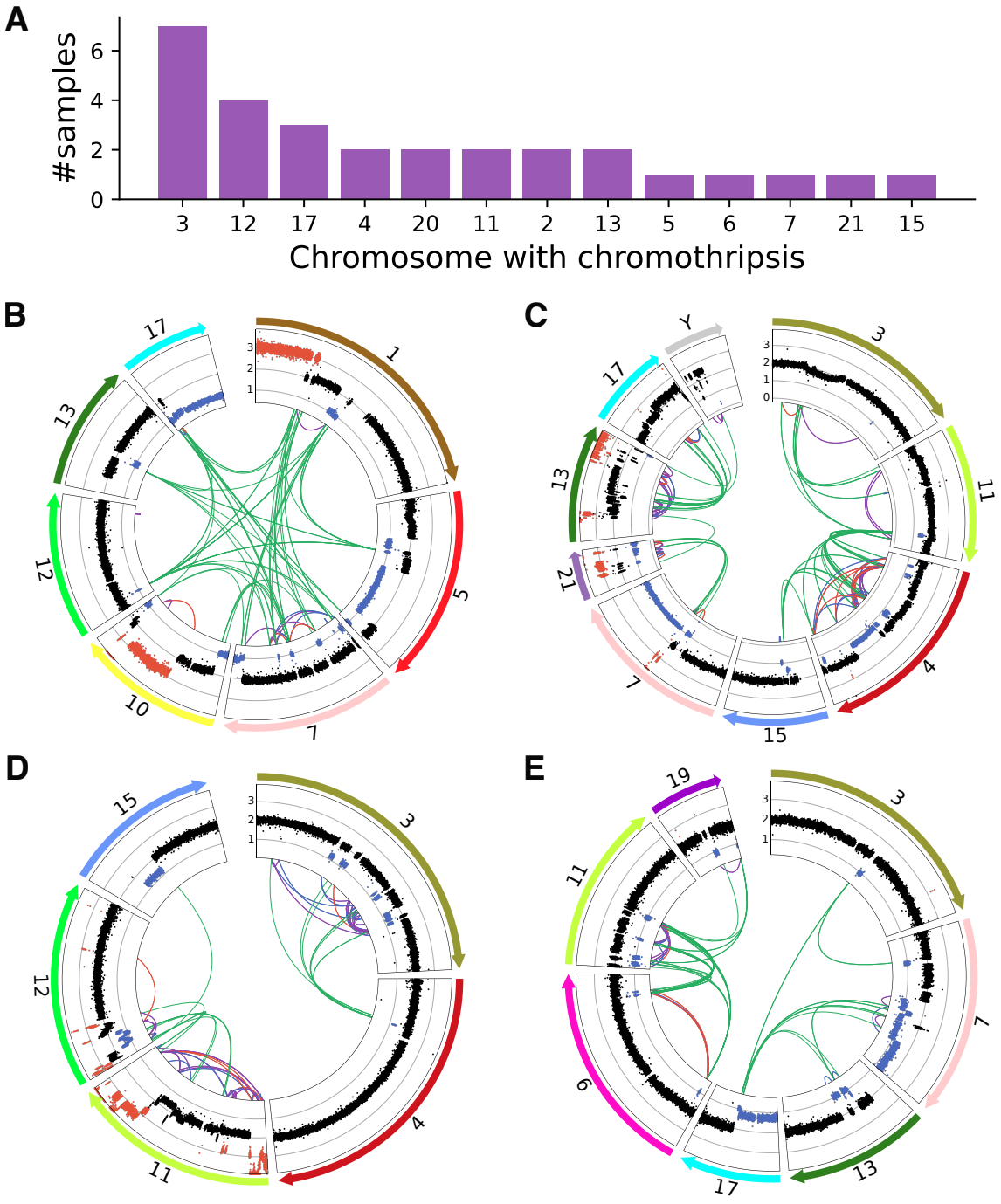

Figure 7: Chromothripsis in the ckAML cohort. **A**. Number of samples with chromothripsis on each chromosome. **B**. Circos plot showing structural rearrangements for sample 16PB1441, where only the chromosomes involving in a large chromothripsis event are shown. Shatterseek only called chromothripsis on chr7, but there were breakpoints leading to other chromosomes. **C**-**E**. Circos plots showing structural rearrangements in samples 16KM4020 (**A**), 15PB9630 (**B**) and 16KM16320 (**C**). Only chromosomes involved in chromothriptic events are shown, and they were reordered so that chromosomes involved in the same chromothripsis event are grouped together.

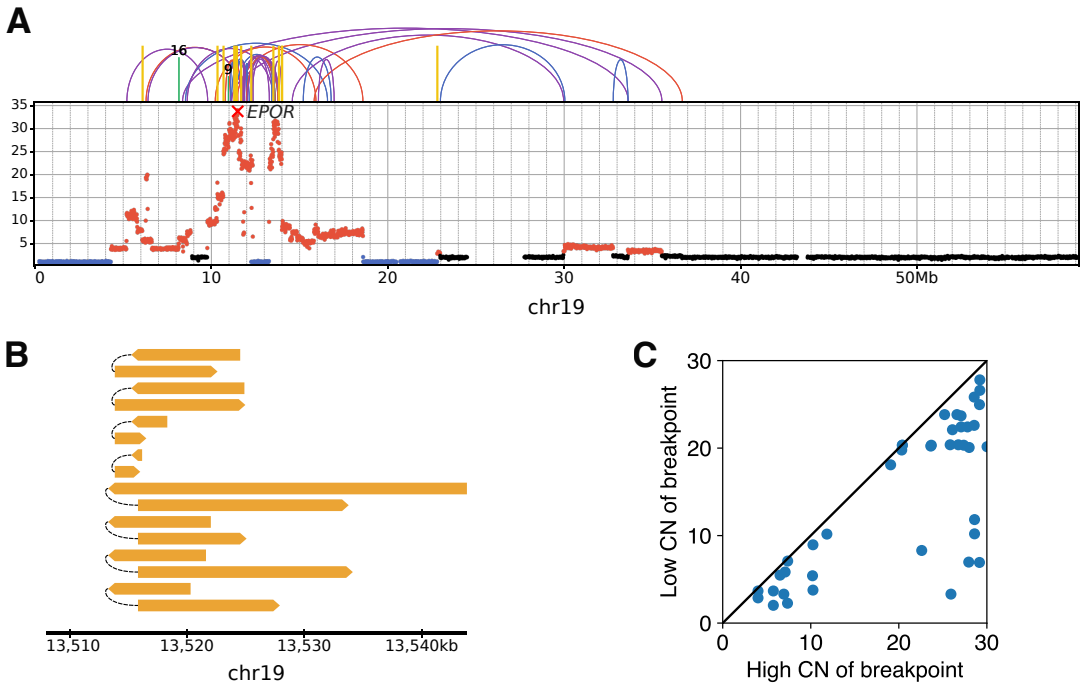

Figure 8: Amplification of *EPOR* with foldback inversions. **A**. Copy numbers and SVs for chr19 in sample 15KM18875. Vertical yellow lines indicate foldback inversions. **B**. Example foldback inversions on chr19 in 15KM18875: here nanopore reads are shown in orange, with black dashed lines linking the two alignments of a split read. BFB results in such foldback inversions, where a read goes in one direction, and then back in the other direction. **C**. Copy numbers at each end of a breakpoint, for SVs on chr19 in sample 15KM18875. Most points are away from the diagonal, indicating that most breakpoints join segments with different copy numbers.

3.1.2 Transcriptomic analysis

In order to see if the cohort consisted of different subgroups, I used the RNA-seq data and clustered the samples based on correlations of the 5000 most variable genes (Figure 9A). This revealed two main groups, but they were not biologically relevant, since they corresponded to whether the sample came from the peripheral blood (PB) or the bone marrow (BM). I performed a differential expression analysis between samples from BM and PB and found 1316 genes upregulated in the BM (log2 fold change>1, FDR<0.01), including chemokine ligands *CXCL12* and *CCL14*, and 378 genes upregulated in the PB. Apart from this PB/BM difference, no strong clustering was observed, even when this PB/BM difference was removed using ComBat [116]. One small subgroup corresponded to samples with a high erythrocyte enrichment computed with xCell [117], and they probably correspond to samples with acute erythroid leukemia (AEL), a rare AML subtype accounting for 5% of all AML cases, but strongly enriched for complex karyotypes. Samples did not cluster according to *TP53* mutation status. I a performed differential expression analysis between *TP53*-mutated (*TP53*-mut) and wild-type (*TP53*-wt) samples, and I identified 178 upregulated genes in *TP53*-mut samples and 129 downregulated genes (Figure 9B). *ZNF560* was the most upregulated gene in *TP53*-mut samples, and this gene was also reported to be the most differentially expressed in *TP53*-mut AML in another cohort [118]. I performed gene set enrichment analysis (GSEA) with gseapy. The most significant gene sets were a downregulation of the interferon alpha and gamma responses in *TP53*-mut samples (Figure 9C-D). This is in line with the role of *TP53* in innate immunity [119]. There was also an enrichment for the hematopoietic stem cell (HSC) gene set, indicating that *TP53*-mut samples might be more stem cell-like. The LSC17 score is a score for leukemic stemness based on the expression of 17 genes [120]. In this dataset, I observed a higher LSC17 score for *TP53*-mut cases (Figure 9E), confirming that *TP53*-mut samples are less differentiated, which might explain their poor prognosis.

Fusion transcripts

Many AML cases are driven by recurrent fusion proteins like RUNX1::RUNX1T1 or PML::RARA, but ckAML generally do not harbour those recurrent translocations. Nevertheless, I hypothesized that there could be many rare fusions in ckAML. Among the 39 ckAML samples, I detected 147 fusion transcripts, but none of them were recurrent. *RUNX1* is known to be involved in many fusions [121], and in this cohort I detected 4 fusions involving *RUNX1*: RUNX1::RCAN1, SMIM11::RUNX1, RUNX1::RWDD2B and

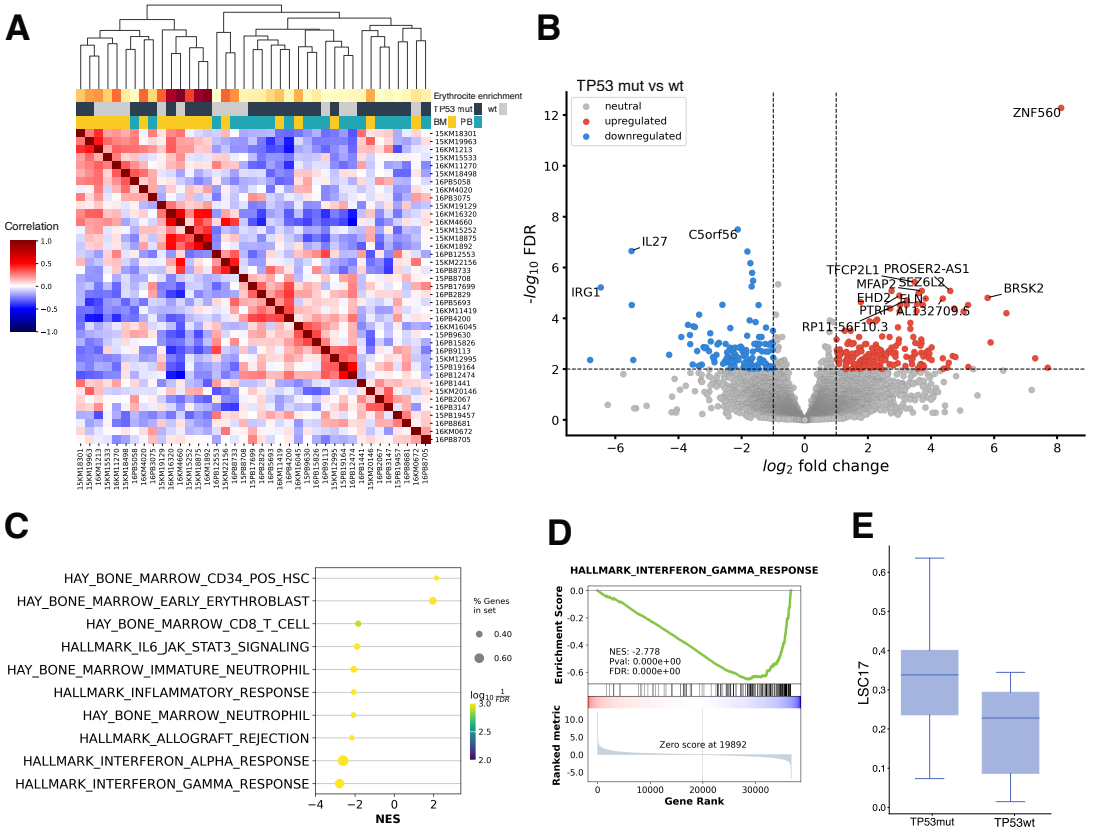

Figure 9: RNA-seq data analysis of the ckAML cohort. **A**. Clustering of the 39 RNA-seq samples based on correlations, using the 5000 most variable genes. **B**. Volcano plot showing differentially expressed genes between *TP53*-mut and wt samples. **C**. GSEA results for gene sets with FDR<1%. **D**. GSEA result for the interferon gamma response gene set. **E**. LSC17 score between *TP53*-mut and *TP53*-wt samples.

LTN1::RUNX1. Of those, only RUNX1::RCAN1 was mentioned in the literature [121]. The role of the other fusions is unclear, but many of them are likely passenger events. For example, there were two fusions involving *TP53* (C11orf48::TP53 and TP53::FBXL18). The fusion transcripts themselves are probably irrelevant, but the breakpoints were likely selected for because they disrupt *TP53*.

3.1.3 CNAs across several ckAML cohorts

Although WGS provides a higher resolution, CNAs can be detected with other assays, like SNP arrays or methylation arrays. I collected data from several ckAML cohorts profiled with different technologies (Table 1), resulting in a total of 418 ckAML samples with CNA information. Such a large dataset could be used to gain new insights into CNAs in ckAML.

Table 1: List of cohorts included in the CNA analysis, with the number of complex ckAML samples in each, the type of assay, whether *TP53* mutation information is available, and the mean age at diagnosis with standard deviation. For the BEAT-AML cohort, *TP53* status and age information was only available for a subset of the samples. The data for the OSUckAML cohort was provided by Ann-Kathrin Eisfeld and Christopher Walker from Ohio State University.

Cohort	**#ckAML**	**Assay**	**TP53**	**Age (years)**
ASTRAL1-WGS	41	WGS	YES	75.2 ± 5.3
ASTRAL1-EPIC	65	EPIC array	YES	76.0 ± 6.0
BEAT-AML [122]	59	EPIC array	Partial	54.6 ± 18.9
TCGA [101]	25	SNP array	YES	54.9 ± 16.1
Rücker2012 [103] (GSE34542)	82	SNP array	YES	58.6 ± 14.1
Parkin2015 [123] (GSE61323)	36	SNP array	YES	60.3 ± 14.4
OSUckAML	110	SNP array	NO	Unavailable
Total	**418**			

Regions frequently deleted or gained

I started by looking at the most common CNAs in ckAML, and the minimally deleted or gained regions (Figure 10). The most common CNA is del(5q), which is often large, resulting in haploinsufficiency of many genes, including the histone demethylase *KDM3B*, *EGR1* whose haploinsufficiency leads to increased rates of hematological malignancies in mice [111], *RPS14* whose downregulation recapitulates del(5q) syndrome [110], and *CSNK1A1* which is sometimes mutated in del(5q) cases [108]. Del(7q) is the second most common CNA, which also deletes a very large region containing many genes including *EZH2*. Del(12p) typically deletes a small region around *ETV6*. Chromosome 17 has two deletion peaks, the main one in 17p around *TP53* and a secondary peak in 17q around *NF1*, which is involved in the RAS pathway. Trisomy 8 is the most common gain, and some samples have small amplifications around *MYC*. Chromosomes 11 and 21 frequently harbour amplifications, involving numerous members of the ETS family of transcription factors like *ETS1*, *ETS2*, *FLI1* and *ERG*. Copy neutral loss of heterozygosity (CNLOH) frequently affected 17p, but not the other commonly deleted regions. CNLOH of 17p can inactivate the second *TP53* allele after the first one is mutated. The absence of CNLOH in the other commonly deleted regions is another argument against the presence of TSGs in these regions, and in favour of the hypothesis that these deletions drive leukemia through gene dosage alterations.

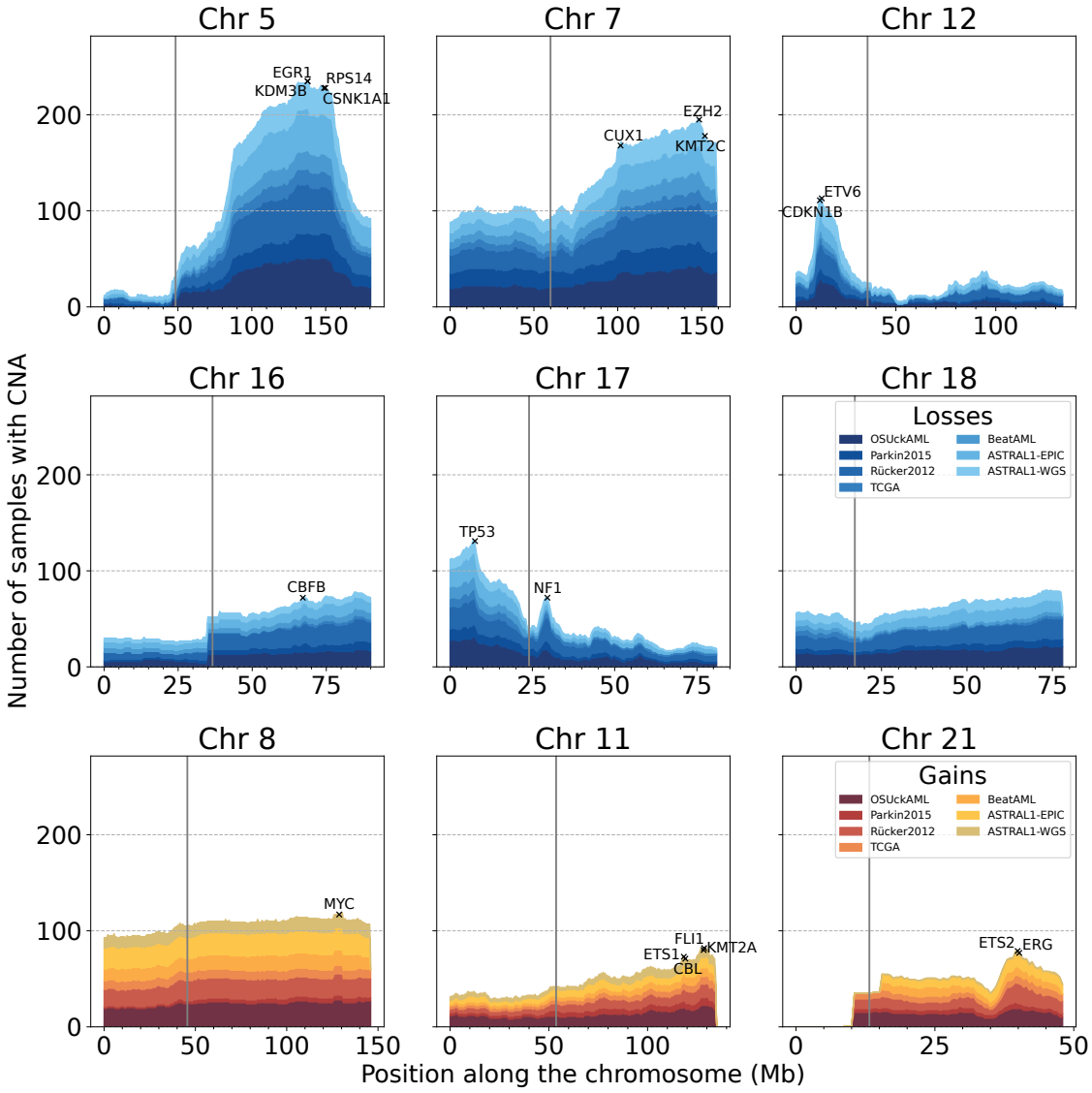

Figure 10: Recurrently deleted (first two rows) and gained (last row) regions across several ckAML cohorts. Vertical gray lines indicate the centromere of each chromosome. Somes genes, which have been proposed as putative drivers for these CNAs, are highlighted.

3.1.4 Haploinsufficiency of genes in the deleted regions

Apart from *TP53* in 17p, the deletions in ckAML do not contribute to the bialellic inactivation of a TSG, so the most likely explanation for their positive selection is that they result in haploinsuffiency of the deleted genes. If one copy of a gene is lost in all cells, we would expect the expression to be 50% of what it normally is in cells diploid for this gene, unless feedback mechanisms compensate for it by increasing transcription from the remaining allele. For three cohorts (ASTRAL1-WGS, TCGA-LAML and BEAT-AML), RNA-seq data was available in addition to copy number data, which allowed me to test whether the genes have a lower expression in samples in which they are

deleted. I computed for each gene the ratio of the mean expression (in TPM) of samples with a deletion divided by the mean expression of samples without a deletion for this gene. I performed this for each cohort (provided at least 5 samples had deletions in this cohort), and averaged the results for all cohorts. The median expression ratio was 0.6 (Figure 11A), which is the expected expression ratio if the tumor purity is 80% ($0.8 * \frac{1}{2} + 0.2 * \frac{2}{2} = 0.6$). In addition, 97.5% of genes had a lower expression when they are deleted than when they are not. This strong impact of deletions on gene expression suggests that gene dosage effects might play an important role in these deletions. However, since most deleted genes have their expression reduced, this information cannot be used to identify which genes in the minimally deleted regions are relevant. For example, almost all genes in the minimally deleted regions of del(5q) and del(7q) have their expression strongly reduced in samples with deletions (Figure 11B-C). One possible strategy to identify genes whose haploinsufficiency provides a fitness advantage would be to perform a CRISPR screen with a positive selection [124], but this is outside the scope of this thesis.

TP53 mutations are strongly associated with del(5q)

For some cohorts, *TP53* mutation status was also available, which enables the detection of associations between *TP53* mutations and CNAs. When I performed Fisher's exact test with the number of CNAs in each chromosome arm depending on *TP53* status, I found that many CNAs were positively associated with *TP53* status (20 chromosome arm-level CNAs with FDR<5%, Table 2), in line with previous reports [103]. However, since *TP53* mutations are also associated with the total number of CNAs in a sample, it could be that these associations between *TP53* mutation and specific CNAs are simply due to the fact that *TP53*-mut samples have more CNAs. In order to remove the impact of the total number of CNAs and to see which CNAs are specifically associated with *TP53* mutation, I computed empirical p-values by generating a null distribution, where I randomly assigned CNAs to samples, while keeping constant the number of CNAs in each sample and the number of occurrences of each CNA. This led to a much lower number of CNAs associated with *TP53*: del(5q) was by far the CNA most strongly associated with *TP53*, followed by CNLOH of 17p, del(18q) and del(16p). Association between *TP53* and CNLOH of 17p is not surprising, since this CNLOH provides a fitness advantage by inactivating the second *TP53* copy after the first one is mutated. Del(18q) and del(16q) are not as common as del(5q) and del(7q) and are not as well studied, but this analysis reveals that they could be very important secondary events in *TP53*-mutated ckAML. The association of del(5q) and *TP53* is well known, but this

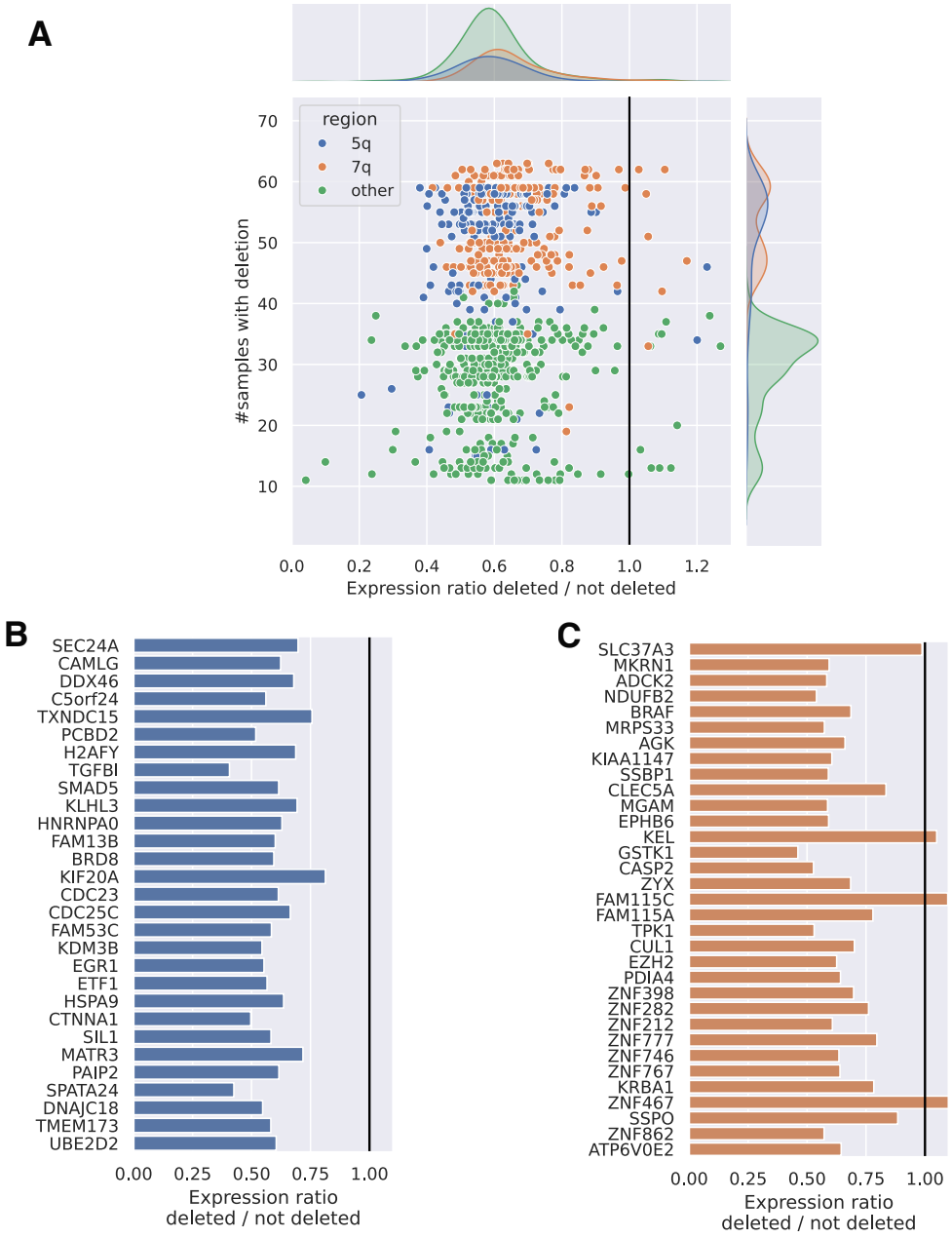

Figure 11: Reduced expression for deleted genes in AML. **A**. Scatter plot showing for each gene, the number of samples in which it is deleted (y-axis) and the ratio of expression between samples with a deletion compared to samples without deletion (x-axis), with marginal densities. **B**. Ratio of expression between samples with and without deletion of the gene, for genes in the minimally deleted region of 5q. **C**. Ratio of expression between samples with and without deletion of the gene, for genes in the minimally deleted region of 7q.

analysis shows that del(5q) is really a critical event in *TP53*-mut ckAML. Del(7q), the other very common deletion in ckAML, is not as strongly associated with *TP53* because, even though it is very frequent in *TP53*-mut samples, it is also common in *TP53*-wt cases, often as a monosomy 7. The association of del(5q) with *TP53* is interesting be-

Table 2: FDR of associations between *TP53* mutations and CNAs, either using Fisher's exact test or by performing a test where the number of CNAs in each sample was kept constant.

CNA	FDR Fisher's exact test	FDR constant number of CNAs
5q_loss	1.476E-25	**0.00001**
18q_loss	0.000009086	**0.01907**
17p_CNLOH	0.00001277	**0.0022**
16p_loss	0.0001061	**0.0264**
7q_loss	0.0005735	0.5594
18p_loss	0.001123	0.2394
17p_loss	0.001675	0.5802
16q_loss	0.002702	0.5594
12q_loss	0.005376	0.5594
7p_loss	0.005376	0.6975
3p_loss	0.008523	0.5802
3q_loss	0.01112	0.5802
20q_loss	0.01212	0.5802
5p_loss	0.01841	0.5802
17q_loss	0.03061	0.8913
21q_gain	0.03286	1
21q_loss	0.03286	0.5802
11q_gain	0.03299	0.9002
2q_loss	0.04676	0.7183
11p_gain	0.04676	0.8456

cause *TP53* mutations are associated with a complex karyotype and a poor prognosis, but outside of ckAML, del(5q) is often seen as a sole abnormality in MDS, and in this case it is associated with a good prognosis [106]. Even though the commonly deleted region is similar between MDS with isolated del(5q) and ckAML, a striking difference is that some regions to the left and to the right of the deleted region are always retained in isolated del(5q), whereas in AML the deletions can be larger and encompass these commonly retained regions (CRRs) [125]. I processed public SNP and methylation array data for MDS with isolated del(5q) [126, 127] and verified the presence of these CRRs (Figure 12). These CRRs might contain genes which, when lost, precipitate the progression to AML.

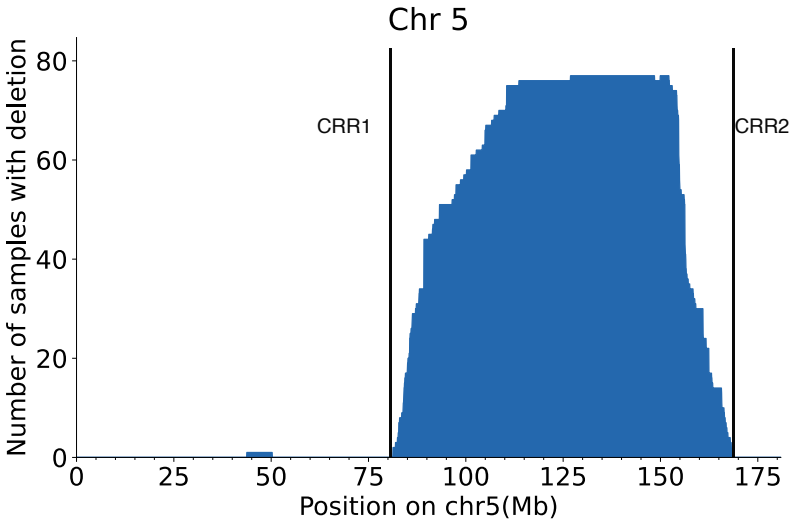

Figure 12: MDS with isolated del(5q). Number of samples having deletions at each position of chr5, for MDS samples having del(5q) as a sole abnormality. Vertical lines separate the commonly retained regions from the region with deletions.

Del(20q) is more frequent in older patients, but does not preferentially affect a particular allele

CkAML is more common among older patients, but can still occur among younger patients, and one open question is whether older and younger ckAML patients show the same CNA profiles. I tested associations between CNAs and age by comparing CNAs between the 33% younger patients and the 33% older patients. This revealed that for the most part, younger and older ckAML cases show the same frequency of CNAs, except for del(20q), which is more common among older patients (Fisher's test two-sided statistic 5.1, p-value 0.01, FDR 0.03; Figure 13A). This result is not very surprising, since mosaic del(20q) has been reported to be common in blood cells of healthy older individuals [128]. *ASXL1*, which together with *DNMT3A* and *TET2* is commonly mutated in clonal hematopoiesis, lies close to the boundary of the commonly deleted region, but is deleted only in a minority of samples, in accordance with previous reports [129]. Since *ASXL1* mutations are gain of function and not loss of function [130], it is not surprising that *ASXL1* is not in the minimally deleted region.

I performed a differential expression analysis between samples with del(20q) and others and found that several genes in the deleted region were downregulated with log2 fold changes lower than -1, as would have been expected if one of the two alleles is lost (Figure 13B). The deleted region in 20q contains several imprinted genes, so losing the active allele could be sufficient to completely abrogate the expression of the gene. In 2002, Kuerbitz et al. found that *NNAT*, a gene located in the commonly deleted

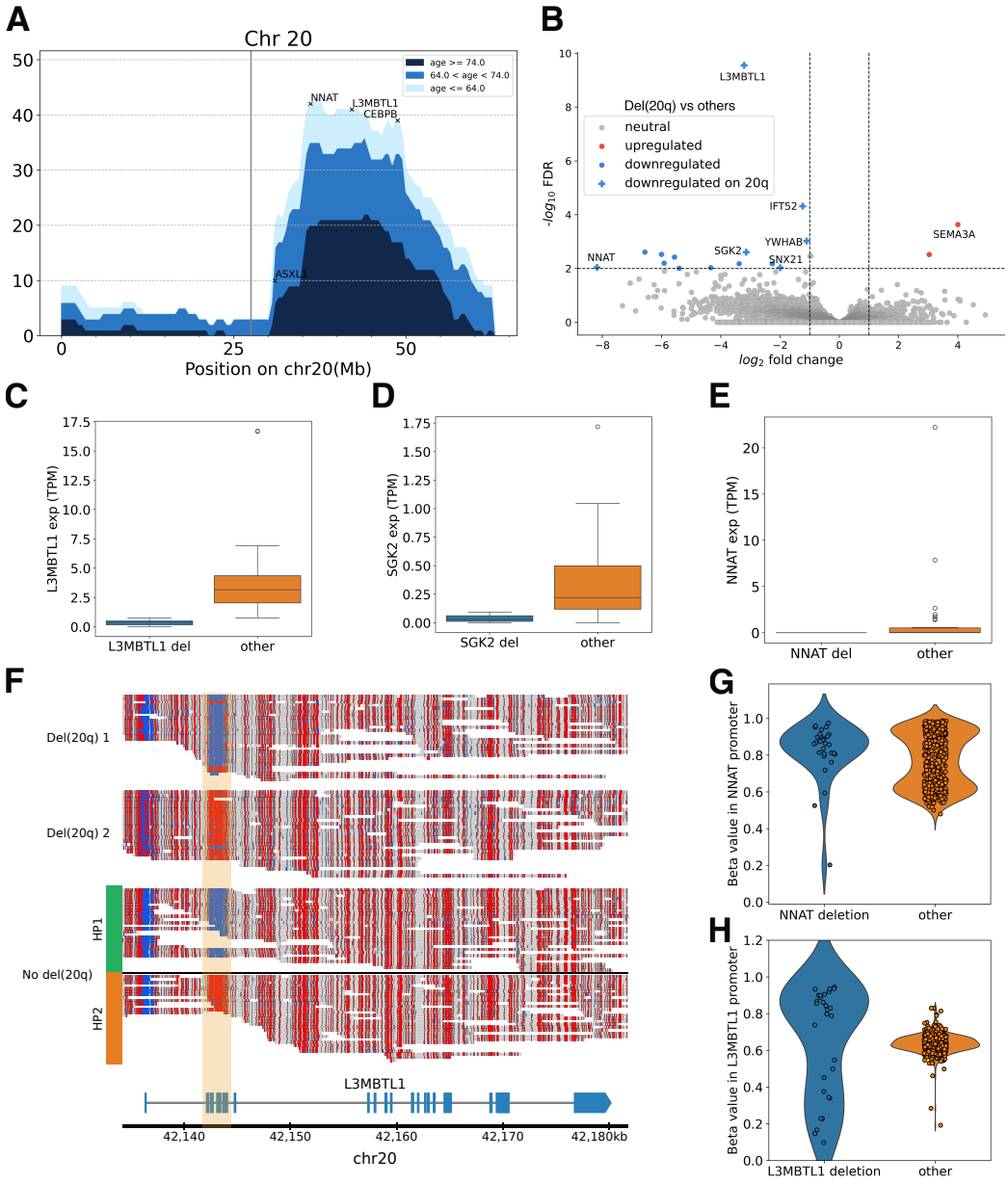

Figure 13: Deletion 20q. **A**. Number of samples with deletions at each position on chr20, split by age. There are as many samples in each age group. **B**. Differential expression between samples with del(20q) and others. Downregulated genes which lie on 20q are indicated by a cross. **C**-**E**. Boxplots for the expression of *L3MBTL1*, *SGK2* and *NNAT*, split between del(20q) and other samples. **H**. Nanopore sequencing data around *L3MBTL1* colored by methylation (red: methylated, blue: unmethylated) for two samples with del(20q), and one sample without del(20q) with reads split by haplotype. **G**-**H**. Average methylation values in the promoters of *NNAT* and *L3MBTL1* measured by EPIC array, split between samples with deletion and others.

region, was imprinted, and was also hypermerthylated in some AML cases without del(20q) [131]. In 2004, Li et al. found that *L3MBTL1* was another imprinted gene in this region [132]. However, among four samples with del(20q), they found two samples which retained the methylated allele and two which retained the unmethylated one,

so it was not always the inactive allele which was lost. In 2013, Aziz et al. found that samples with del(20q) lost expression of *L3MBTL1* and of its neighbour *SGK2* [133]. Out of 6 samples with del(20q), 5 had lost the unmethylated allele, which suggested that the unmethylated allele was preferentially lost, leading to complete inactivation of *L3MBTL1*. In addition, they showed that downregulation of *L3MBTL1* and *SGK2* could activate *MYC*. I verified that these genes were downregulated in the del(20q) samples of the ASTRAL-1 cohort profiled with RNA-seq (Figure 13C-E). I also verified in nanopore data that in samples without del(20q), there was a differential methylation at the *L3MBTL1* promoter between the two haplotypes (Figure 13F), and that only the allele unmethylated at the promoter was expressed in RNA-seq data. However, I found that both the methylated and the unmethylated allele could be retained, both in nanopore data and in EPIC array data (Figure 13F-H). There were slightly more samples hypermethylated (N=18) than hypomethylated (N=11) among the samples with del(20q), but this was not statistically significant (binomial test: p-value=0.26). Among the 11 del(20q) samples hypomethylated at the *L3MBTL1* promoter, only one had been profiled with RNA-seq, so it is possible that the other samples which lost the inactive *L3MBTL1* allele might still express it. For *NNAT*, I observed in samples without del(20q) a bimodal distribution of the methylation values: some samples had 50% methylation, typical of an imprinted gene, while others had higher methylation, which would agree with the observation that *NNAT* can become hypermethylated even without deletion. The majority of samples with del(20q) had high methylation at the *NNAT* promoter, but this can be explained by the fact that some samples lost the methylated allele, and the unmethylated allele then underwent hypermethylation. In conclusion, it appears that del(20q) does not have a strong bias for a particular allele, contrary to some reports [133]. *L3MBTL1* downregulation might be relevant, but it is likely not the only driver event in del(20q), otherwise a stronger allelic bias would have been expected.

Co-occurrence and mutual exclusivity of CNAs

The large number of ckAML samples with CNA information allowed me to also look at patterns of co-occurrence and mutual exclusivity between CNAs. In order to avoid seeing mainly patterns of co-occurrence driven by the fact that some samples harbour more CNAs than others, I again computed p-values by using a null distribution, where I randomly assigned CNAs to samples, keeping the number of CNAs in each sample and the total number of occurrences of each CNA constant (Figure 14). 8p and 8q gains appeared mutually exclusive with most CNAs, which reflects the fact that trisomy 8 is often seen as a sole abnormality, although it is also common in ckAML. Del(5q) co-

occurred frequently with deletions in chr7, in chr16 and in chr18. The strongest mutual exclusivity was between del(12p) and gain of 21q. This might be because 12p deletions are selected for because they result in haploinsuffiency of *ETV6*, while 21q gains amplify *ETS2* and *ERG*. These three genes are transcription factors of the ETS family, but they have opposite effects: *ETV6* is a silencer while *ETS2* and *ERG* are activators. Consequently, del(12p) might result in a similar effect as amplifications of 21q, which might be why these two events rarely co-occur. 1p, 11p and 22q gains often co-occur together. This might be because these gains are often caused by trisomies, and some samples harbour a large number of trisomies.

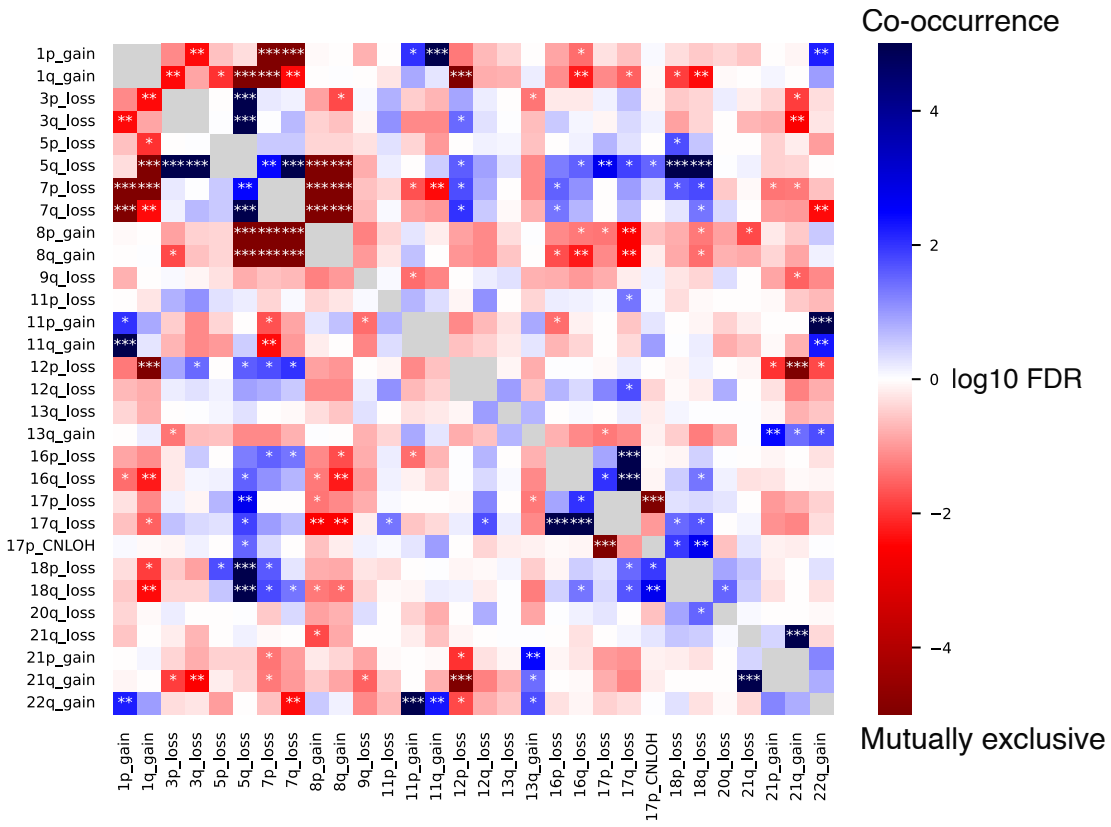

Figure 14: Co-occurrence and mutual exclusivity of chromosome arm-level CNAs in ckAML. Red colors indicate mutual exclusivity, blue co-occurrence, and grey indicates pairs which were not tested because they correspond to the same CNA type on the same chromosome. One asterisk indicates a FDR <0.05, 2 asterisks a FDR<0.01 and 3 asterisks a FDR<0.001.

3.2 Detection of enhancer hijacking events

I showed in the previous section that ckAML samples have very complex genomes, which in particular harbour many deletions. In this section, I investigated whether enhancer hijacking events could play an important role in ckAML, as well as in other cancer types. I developed a computational method called pyjacker to detect enhancer hijacking events using WGS and RNA-seq, and I applied it to a ckAML cohort, as well as to sarcoma and prostate cancer samples.

3.2.1 Pyjacker

An enhancer hijacking event can occur when an SV brings an active enhancer close to an inactive gene, which can lead to the aberrant expression of this gene and potentially drive cancer. Although several methods already exist for the detection of enhancer hijacking events, none of them were directly applicable to our ckAML data (Table 3). CESAM [89] and HYENA [93] can only detect recurrent events, at least present in four samples of a cohort, whereas in this project I aimed at discovering rare events, including those present in only a single sample. Cis-X [94] can detect genes activated by SVs in single samples, but it requires matched normals to be provided, and is therefore not applicable to our cohort, for which only leukemic samples were available. I developed pyjacker, a computational method to detect enhancer hijacking events using WGS, RNA-seq and enhancer information, even in single samples, without the need for matched normals.

For each gene, the first step in pyjacker is the identification of "candidate samples" which have a breakpoint in the same TAD as the gene (Figure 15A). Then, pyjacker will assign a score to each candidate sample, which reflects how likely this gene is to be expressed in this sample because of the SV. This score consists of an overexpression score, a monoallelic expression score, and an enhancer score, which are combined with custom weights (see methods section). Yhe scores for the same gene from all candidate samples are aggregated, resulting in a single score for each gene. In order to get a more interpretable FDR, I compute empirical p-values by generating a null distribution of scores in the absence of enhancer hijacking. For each gene, I ignore the true candidate samples with breakpoints nearby, and I randomly assign some of the reference samples (without breakpoint) to be candidate samples. I then compute the scores for these "false" candidate samples, resulting in a null distribution. I can then compute an empirical p-value by counting the proportion of null scores higher than a particular

Table 3: Comparison of tools to detect enhancer hijacking. Features that are problematic for this study are colored in red, others in blue.

Method	Required data	Can be run without matched normals	Can detect enhancer hijacking events in single samples	Uses expression level	Uses monoallelic expression	Uses enhancers
CESAM	breakpoints + expression	YES	NO	YES	NO	NO
SVXpress	WGS + RNA-seq	YES	NO	YES	NO	NO
HYENA	WGS + RNA-seq	YES	NO	YES	NO	NO
cis-X	WGS + RNA-seq	NO	YES	YES	YES	YES
NeoLoop Finder	HiC	YES	YES	NO	NO	NO
pyjacker	WGS + RNA-seq	YES	YES	YES	YES	YES

score. Finally, I correct for multiple testing of all genes with the Benjamini-Hochberg method. This yields a ranked list of genes putatively activated by structural rearrangements, with a false discovery rate for each (Figure 15B). In addition to enhancer hijacking, an SV can cause monoallelic overexpression of a gene by creating a fusion transcript. In order to differentiate between genes activated by a fusion and enhancer hijacking, I detected gene fusions with STAR-Fusion [134], and annotated the results of pyjacker with the potential presence of fusion transcripts. Pyjacker is implemented in python and available on GitHub at `https://github.com/CompEpigen/pyjacker`.

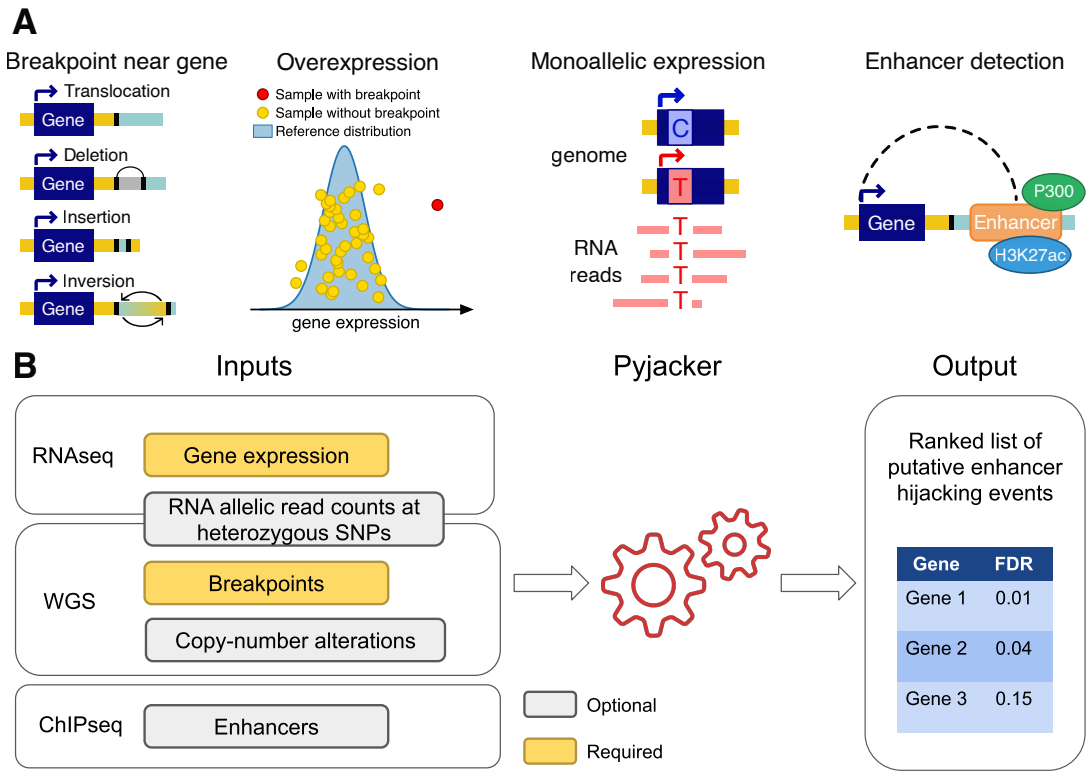

Figure 15: Pyjacker overview. **A**. Schematic representation of the main sources of information used by pyjacker: breakpoints, overexpression, monoallelic expression, enhancers. **B**. Diagram of pyjacker's inputs and outputs.

3.2.2 Pyjacker applied to ckAML

3.2.2.1 Results overview

I applied pyjacker to the cohort of 39 ckAML samples, which resulted in 19 genes activated by SVs with FDR<20%, 9 of which were not involved in fusions and are therefore likely activated by enhancer hijacking (Figure 16). Among those events were genes which were known to be activated by enhancer hijacking in AML: *MECOM* [90] was overexpressed as a result of a SV in two samples, and *BCL11B* [91] and *MNX1* [92] in one sample each. These positive controls show that pyjacker identifies previously reported enhancer hijacking events. In addition, pyjacker found several novel interesting genes. Most of those genes were recurrently expressed in the TCGA-LAML [101], BEAT-AML [122] and TARGET-AML [135] cohorts, albeit at low frequencies (Figure 16B). *BCL11B* and *CLEC10A* are expressed in normal T cells and dendritic cells, respectively, and are therefore expressed in a high proportion of AML samples, probably due to low tumor purity.

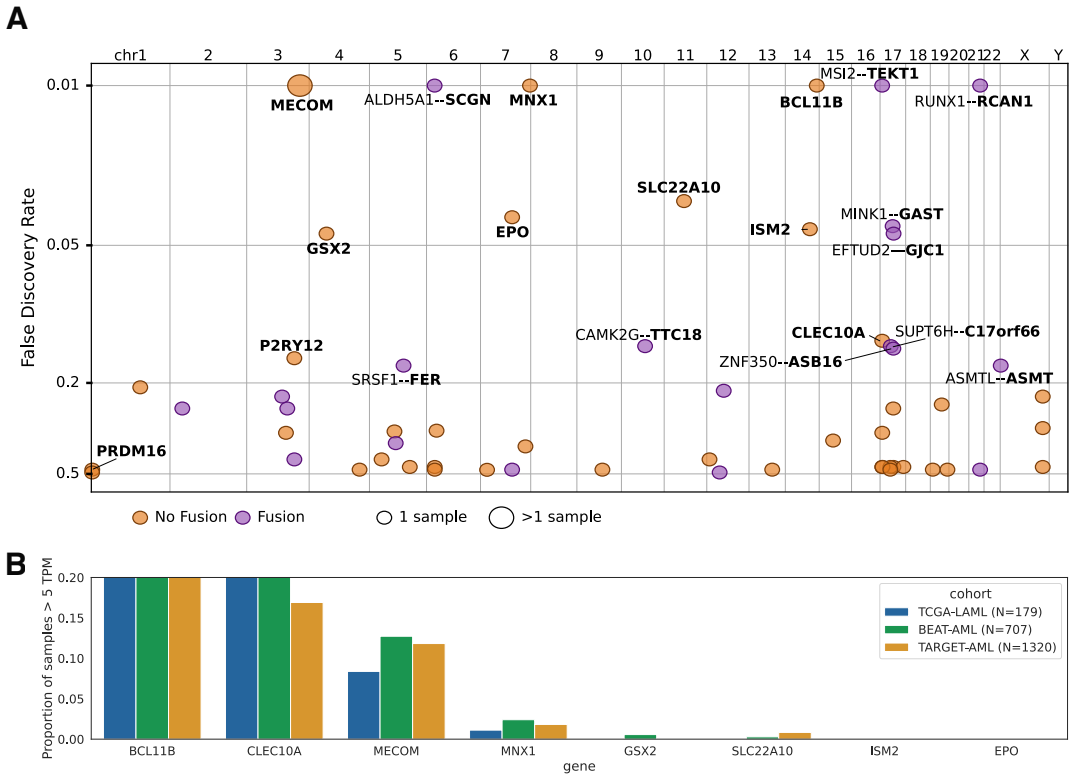

Figure 16: Putative enhancer hijacking events and fusion transcripts detected in 39 ckAML samples. **A**. Scatter plots of genes identified by pyjacker as being potentially activated by genomic rearrangements in one or more samples, where the x-axis shows the genomic location of the genes and the y-axis shows the FDR. Gene names for the enhancer hijacking candidates are written in bold, and if a fusion transcript was detected, the fusion partner is named. **B**. Proportion of samples expressing the top candidate genes, for three AML cohorts profiled with RNA-seq: TCGA-LAML [101], BEAT-AML [122] and TARGET-AML [135].

3.2.2.2 *MECOM* and its homolog *PRDM16* activated by the *GATA2* enhancer

The only gene identified by pyjacker in more than one sample from this ckAML cohort was *MECOM*, found in two samples (15PB19457 and 15KM20146), where I found both overexpression and monoallelic expression (Figure 17A-C). In both cases, the rearrangements were more complex than t(3;3) or inv(3) which are the most frequent rearrangements responsible for *MECOM* activation. Sample 15PB19457 had chromothripsis on chromosome 3 (Figure 17D), while sample 15KM20146 had several breakpoints between chr3 and chr14 (Figure 17F). Even though these rearrangements were complex, they still resulted in the *GATA2* enhancer (next to *RPN1*) coming close to *MECOM* (Figure 17E), which is the same enhancer that activates *MECOM* in the more common t(3;3) and inv(3) [90]. However, the complexity of the rearrangements prevented the identification of the 3q26 region to be identified through cytogenetics.

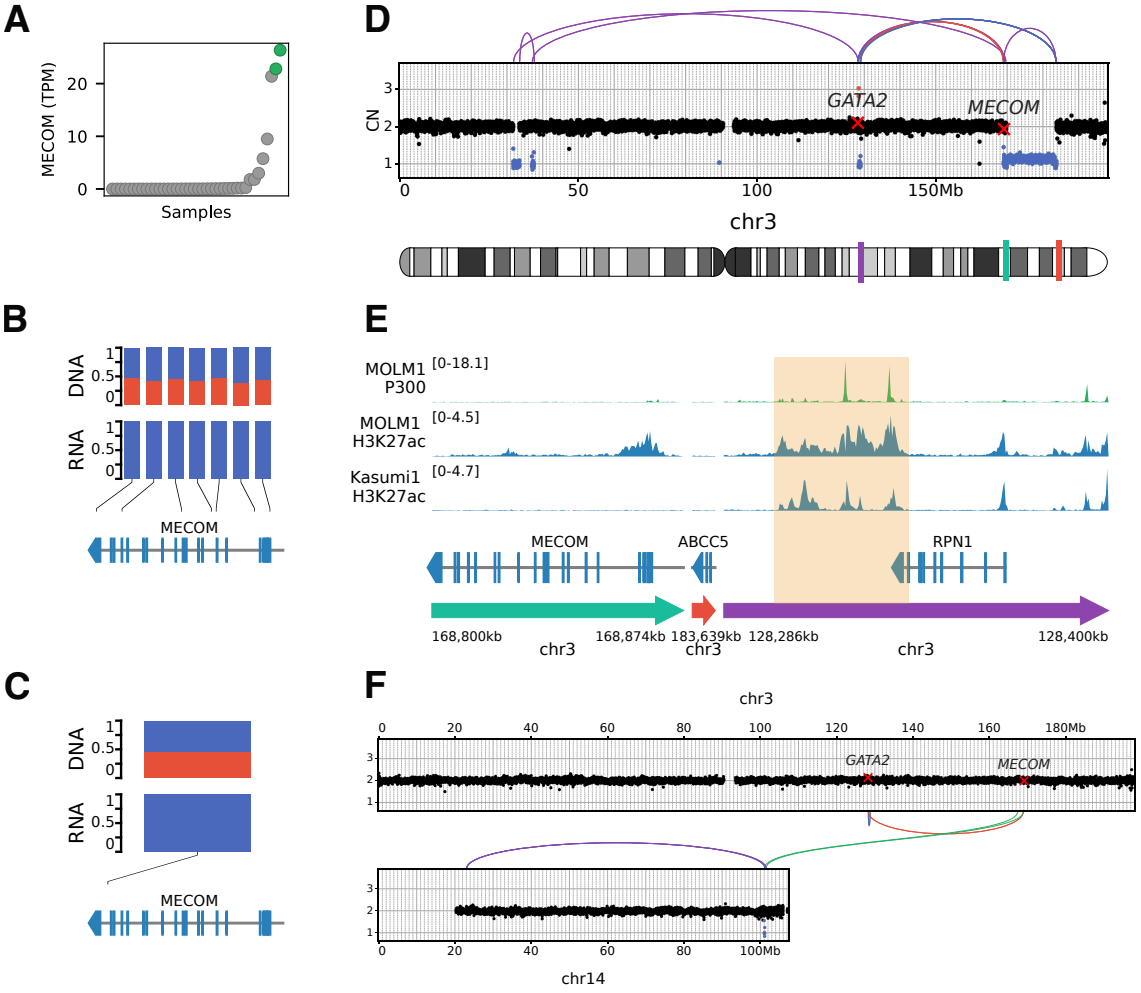

Figure 17: Rearrangements leading to *MECOM* activation. **A**. Expression of *MECOM* in all samples, ranked by expression, where the two samples 15PB19457 and 15KM20146 with *MECOM* activation by SVs are highlighted in green. **B,C**. Variant allele frequencies in DNA and RNA for SNPs in *MECOM*, for samples 15PB19457 and 15KM20146 respectively. **D**. Copy numbers and SVs on chr3 for sample 15PB19457. **E**. ChIP-seq tracks for P300 and H3K27ac in the myeloid cell lines MOLM-1 and Kasumi-1 in the region around *MECOM* for the rearranged chromosome of sample 15PB19457. **F**. Copy numbers and SVs on chr3 and chr14 for sample 15KM20146.

The *GATA2* enhancer was also reported by pyjacker to activate, in a different sample (16KM11270), *PRDM16*, which is a homolog of *MECOM* [136]. In this sample, a t(1;3)(p36;q21) translocation juxtaposed *PRDM16* next to the *GATA2* enhancer (Figure 18C). This translocation t(1;3) has been reported in the literature as a rare event [136]. Even though the expression was monoallelic (Figure 18B), which is a strong indicator of activation by enhancer hijacking, the FDR reported by pyjacker was high (47%) because eight samples without breakpoints near *PRDM16* had a higher expression than in this sample (Figure 18A). *MECOM* is also expressed in samples without breakpoints nearby, although to a lesser extent, which indicates that there must be additional activation mechanisms for these two genes in addition to enhancer hijacking.

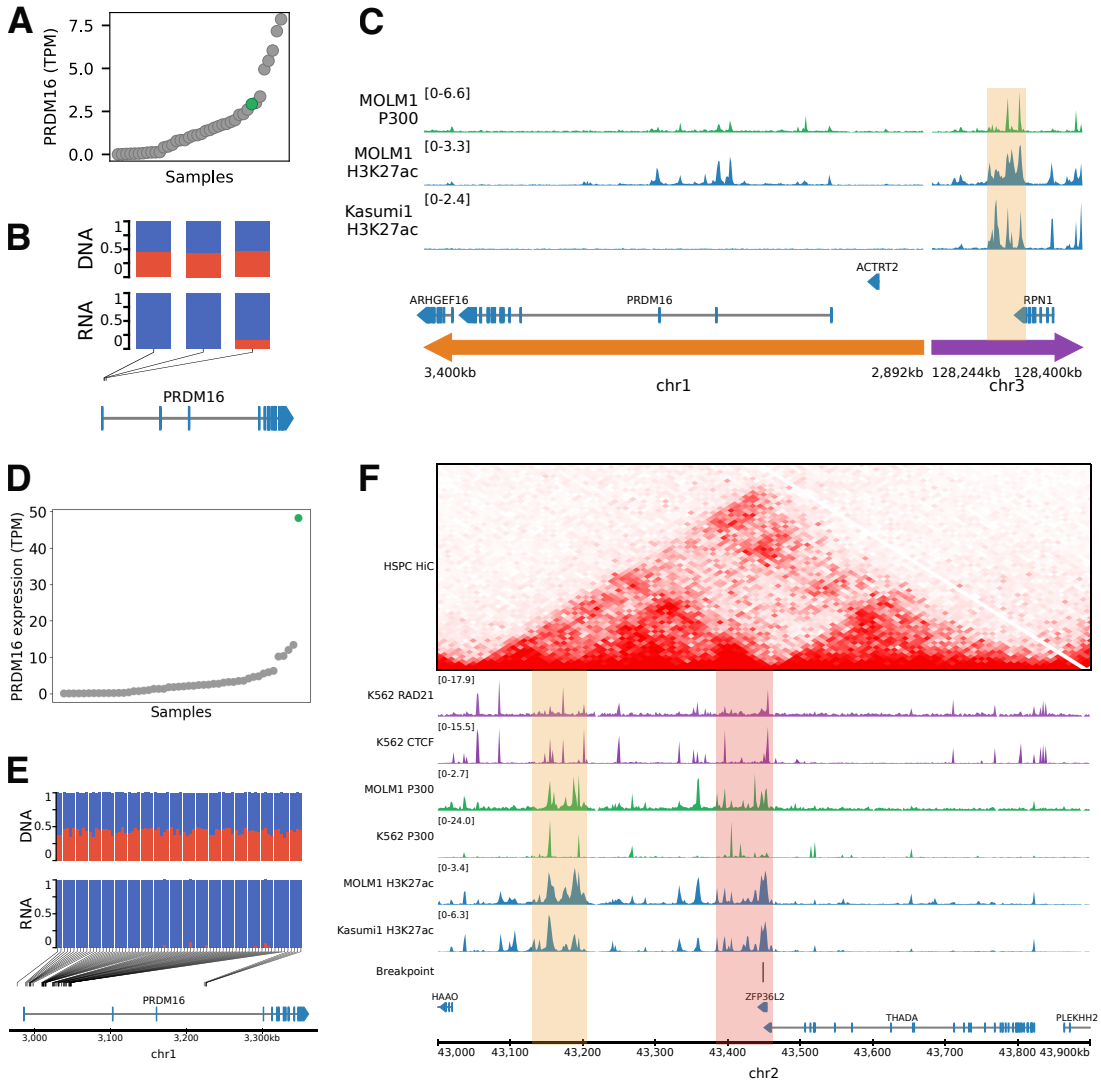

Figure 18: *PRDM16* enhancer hijacking. **A**. Expression of *PRDM16* in all ckAML samples, ranked by expression, where sample 16KM11270 with *PRDM16* activation by SVs is highlighted in green. **B**. Variant allele frequencies in DNA and RNA for SNPs in *PRDM16*, for sample 16KM11270. **C**. ChIP-seq tracks for P300 and H3K27ac in the myeloid cell lines MOLM-1 and Kasumi-1 in the region around *PRDM16* for the rearranged chromosome of sample 16KM11270. **D**. Expression of *PRDM16* in all ckAML samples (grey) and the t(1;2) sample (green). **E**. Variant allele frequencies in DNA and RNA for SNPs in *PRDM16*, for the t(1;2) sample. **F**. Enhancer marks and CTCF/RAD21 binding near the breakpoint in the t(1;2) sample, with Hi-C data from HSPCs. The two super-enhancers are highlighted in orange.

Outside of this ckAML cohort, I also analyzed an AML sample with diabetes insipidus, which harboured a t(1;2) translocation leading to *PRDM16* expression. Compared to the ckAML sample with t(1;3) which had lower *PRDM16* expression than some samples without rearrangements (Figure 18A), this t(1;2) sample showed much higher *PRDM16* expression than all other samples (Figure 18D). This expression was also monoallelic (Figure 18E). The t(1;2) translocation has been reported before, and the expression of *PRDM16* was hypothesized to be due to a juxtaposition to the *THADA* pro-

moter [137]. However, in the sample analyzed here, the *THADA* promoter was not in the region translocated to *PRDM16*, but two strong hematopoietic super-enhancers were brought close to *PRDM16* (Figure 18F). Therefore, it is likely that *PRDM16* expression in this sample is driven by the hijacking of these two enhancers. Although these two enhancers were ranked by ROSE [46, 47] among the strongest hematopoietic super-enhancers, their role in normal hematopoiesis is unclear. Considering that they are in the same TAD as *ZFP36L2*, a gene with high expression in hematopoietic cells, it is likely that these enhancers normally activate this gene. These enhancers on chr2 have also been reported to activate *MECOM* in atypical rearrangements [138].

3.2.2.3 *MNX1* and *GSX2* can be activated by atypical mechanisms

Among the top pyjacker hits were two homeobox genes, *MNX1* and *GSX2*. Homeobox genes are often upregulated in AML [139], so activation of homeobox genes by enhancer hijacking could be a driver event. Both *MNX1* and *GSX2* are known to be activated by rare but recurrent translocations to the *ETV6* locus: *MNX1* is activated by t(7;12)(q36;p13) in pediatric AML [92], and *GSX2* becomes activated by t(4;12)(q11-q12;p13) in adult AML [140]. However, here I found these two genes activated by atypical mechanisms. Sample 15PB8708 had outlier high and monoallelic expression of *MNX1* (Figure 19A-B). A 200kb region in the *CDK6* region on chr7, containing two putative enhancers, was duplicated and inserted next to *MNX1* (Figure 19D-E). This hematopoietic super-enhancer has already been reported to be involved in enhancer hijacking events in AML, activating *BCL11B* [91] or *MECOM* [138].

Sample 16PB5693 has outlier high *GSX2* expression (Figure 19C) and harbours a chromothripsis event involving multiple chromosomes, with several parts amplified, including *GSX2*. (Figure 19F). The putative enhancer is located less than 1Mb away from *GSX2* in the wild-type state, but in a different TAD (Figure 19G). A deletion removed the TAD boundary, which likely allowed *GSX2* to interact with the enhancer. *GSX2* is usually expressed as a result of the t(4;12) translocation, which also frequently leads to *PDGFRA* activation and to an *ETV6-CHIC2* fusion transcript in addition to *GSX2* expression [141]. Here, I found only *GSX2* expression without *PDGFRA* expression and without fusion transcript, indicating that *GSX2* expression is likely the driving event.

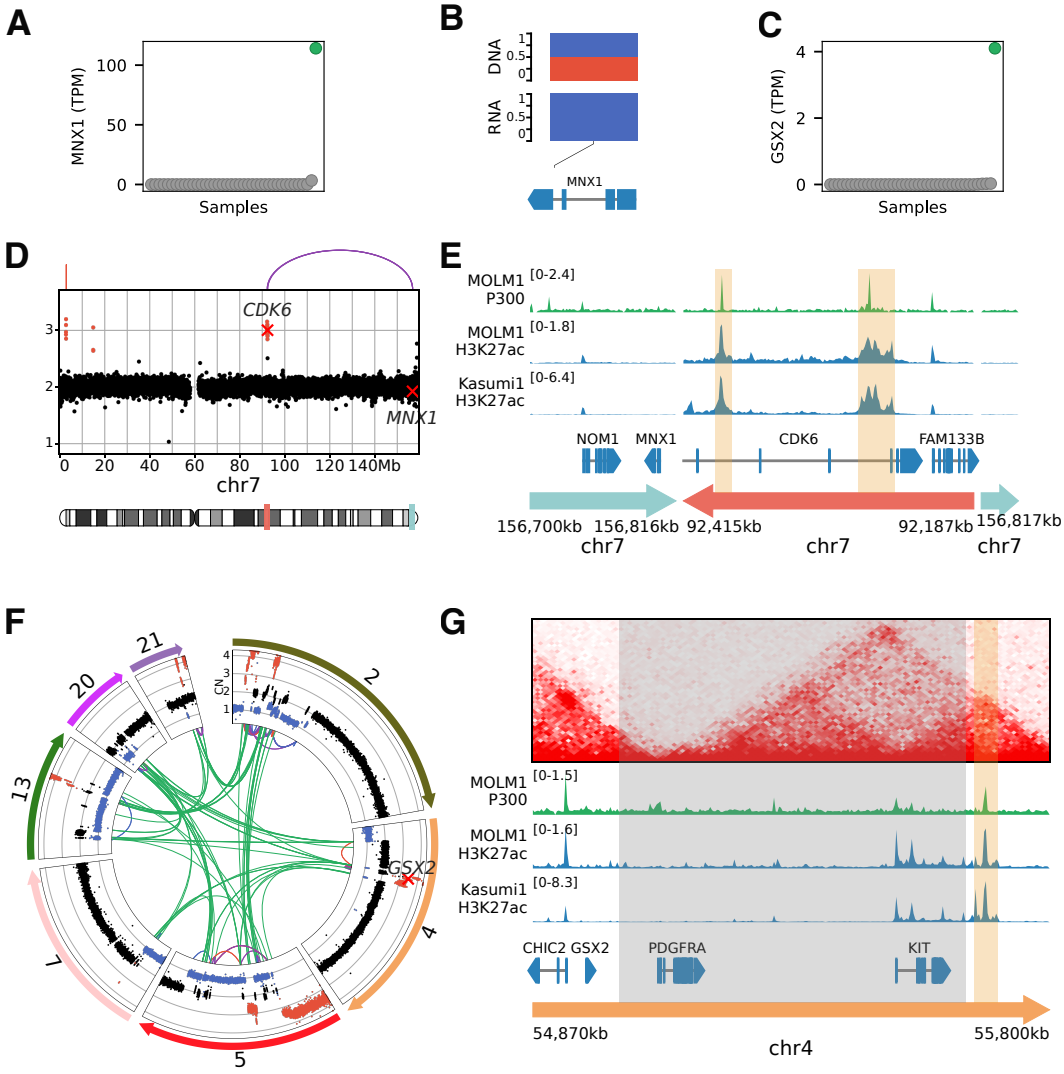

Figure 19: Activation of *MNX1* and *GSX2* by atypical mechanisms. **A**. *MNX1* expression in all samples, with the sample 15PB8708 with *MNX1* overexpression highlighted in green. **B**. Variant allele frequencies in DNA and RNA for a SNP in *MNX1* in sample 15PB8708. **C**. *GSX2* expression in all samples, with the sample 16PB5693 with *GSX2* expression highlighted in green. **D**. Copy numbers and breakpoints on chr7 for sample 15PB8708. **E**. ChIP-seq tracks for P300 and H3K27ac in myeloid cell lines MOLM-1 and Kasumi-1 in the region around *MNX1*, on the rearranged chromosome of sample 15PB8708. Enhancers of the *CDK6* region are highlighted in yellow. **F**. Circos plot showing CNAs and SVs in sample 16PB5693, for the chromosomes involved in a chromothripsis event. **G**. HiC data from HSPCs and ChIP-seq data from myeloid cell lines in the region around *GSX2*. The putative enhancer is highlighted in yellow and the region in grey is deleted in sample 16PB5693.

3.2.2.4 Aberrant *EPO* expression cooperates with *EPOR* amplification to drive AEL

EPO is a novel gene identified by pyjacker that has never been reported to be activated by enhancer hijacking in human leukemias. However, *EPO* can be overexpressed because of genomic rearrangements in a mouse model of erythroleukemia, providing growth factor independence [142, 143]. *EPO* is normally expressed in the kidneys

when oxygen levels in the blood are low, and it stimulates the proliferation of erythroid progenitor cells by binding to its receptor (EPOR) and activating the JAK/STAT pathway [144]. If EPO can promote the growth and survival of erythroid progenitor cells, it is likely that it could drive acute erythroleukemia (AEL), which is a rare subtype of AML enriched for complex karyotypes. AS-E2 is an AEL cell line which requires EPO for survival, which further highlights the importance of EPO for AEL cells [145]. In our ckAML cohort, the AEL sample 15KM18875 had high *EPO* expression (Figure 20A). Although no samples from the TCGA-LAML, BEAT-AML and TARGET-AML cohorts expressed EPO, I found that among three AEL cohorts profiled with RNA-seq [146, 147, 148], one sample from each cohort expressed *EPO* (Figure 20B), indicating that *EPO* expression is a rare but recurrent event in AEL. In sample 15KM18875, a 100kb region on chr7 around EPO was duplicated, with breakpoints leading to a 200kb duplicated region on chr11 (Figure 20C). The breakpoints indicated that these two pieces of DNA from chr7 and chr11 likely formed a circle (Figure 20D). Extrachromosomal circular DNA (eccDNA) are rather common in cancer, but they are often amplified, whereas in this sample I found that the average copy number of this circle was less than 1. This eccDNA is therefore subclonal, but it is unclear whether most cells have one copy, or whether a small percentage of cells contain numerous copies. The region on chr11 which comes close to *EPO* contains a putative enhancer with P300 and H3K27ac peaks in the erythroid cell line K562, and this putative enhancer is likely responsible for the activation of *EPO*. In addition to high *EPO* expression, I also observed very high *EPOR* expression in this sample (Figure 20E), which was due to a massive amplification of *EPOR* on chr19 (Figure 20F). Chr19 harboured patterns of chromothripsis, as well as foldback inversions, indicating that the amplifications were likely due to breakage-fusion-bridge cycles [149]. Amplification of *EPOR* has recently been reported as a recurrent driver event in AEL [148], but *EPO* overexpression was not mentioned. High *EPOR* expression could make the cells very sensitive to *EPO*, thus increasing the fitness advantage provided by endogenous *EPO* expression by the leukemic cells. In both the Iacobucci 2019 and Fagnan 2020 cohorts, the sample with EPO expression also had outlier *EPOR* overexpression, indicating that *EPO* is recurrently overexpressed together with *EPOR*.

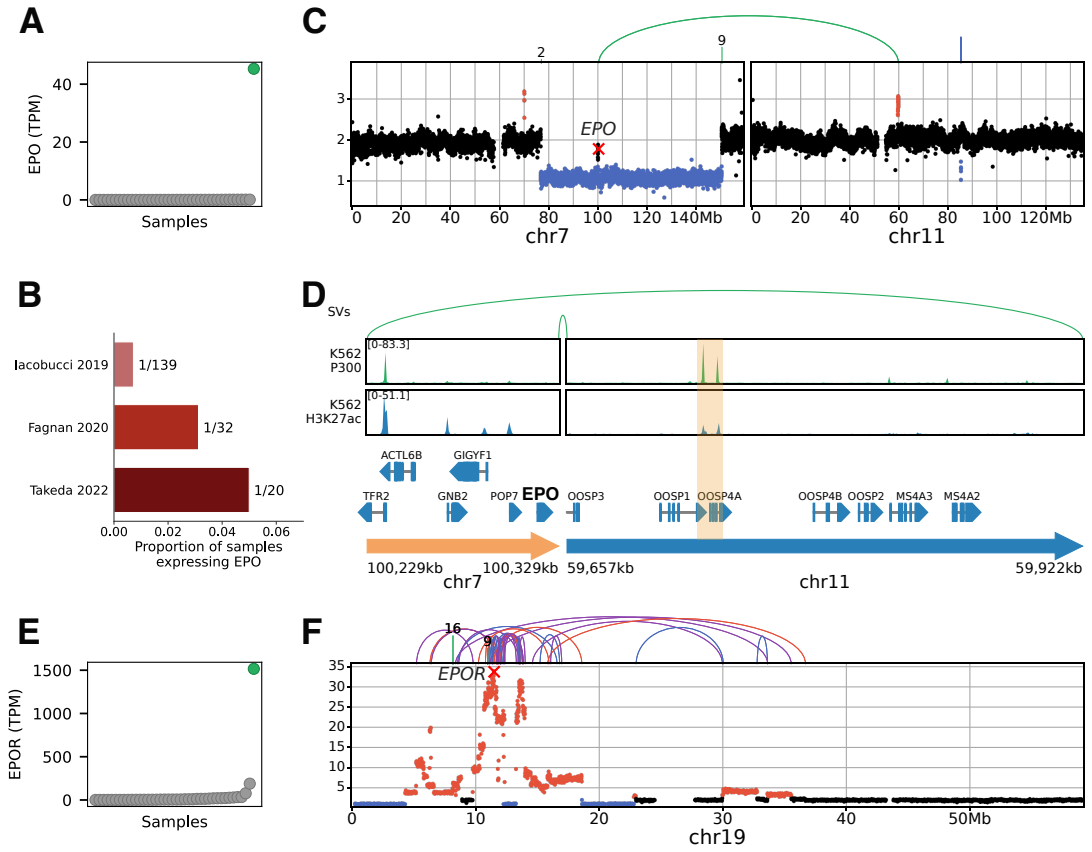

Figure 20: Aberrant *EPO* expression cooperates with *EPOR* amplification to drive AEL. **A**. *EPO* expression in all samples, with the sample 15KM18875 with *EPO* overexpression highlighted in green. **B**. Proportion of samples with *EPO* expression in three AEL cohorts profiled with RNA-seq. **C**. Copy numbers and SVs on chr7 (containing *EPO*) and chr11 in sample 15KM18875. **D**. 300kb circular piece of DNA containing *EPO* and a putative enhancer (highlighted in yellow), with P300 and H3K27ac peaks in the erythroid cell line K562. **E**. *EPOR* expression in all samples, with sample 15KM18875 highlighted in green. **F**. Copy numbers and SVs on chr19 for sample 15KM18875.

3.2.2.5 Rare genes overexpressed as a result of a complex rearrangement

Among the genes identified by pyjacker, some were not found overexpressed in other cohorts (Figure 16). This can be either because they are false positives, because the activation of these genes is a very rare driver event, or because the activation of these genes was a passenger event and was selected for because it was part of a complex genomic rearrangement which contained other driver events. For example, the activations of *TEKT1* (in 16PB3075) and of *SLC22A10* (in 15KM20146) were due to complex rearrangements which also contained SVs within TP53 (Figure 21), so these rearrangements might have been selected for because of the *TP53* disruption rather than *TEKT1* or *SLC22A10* activation.

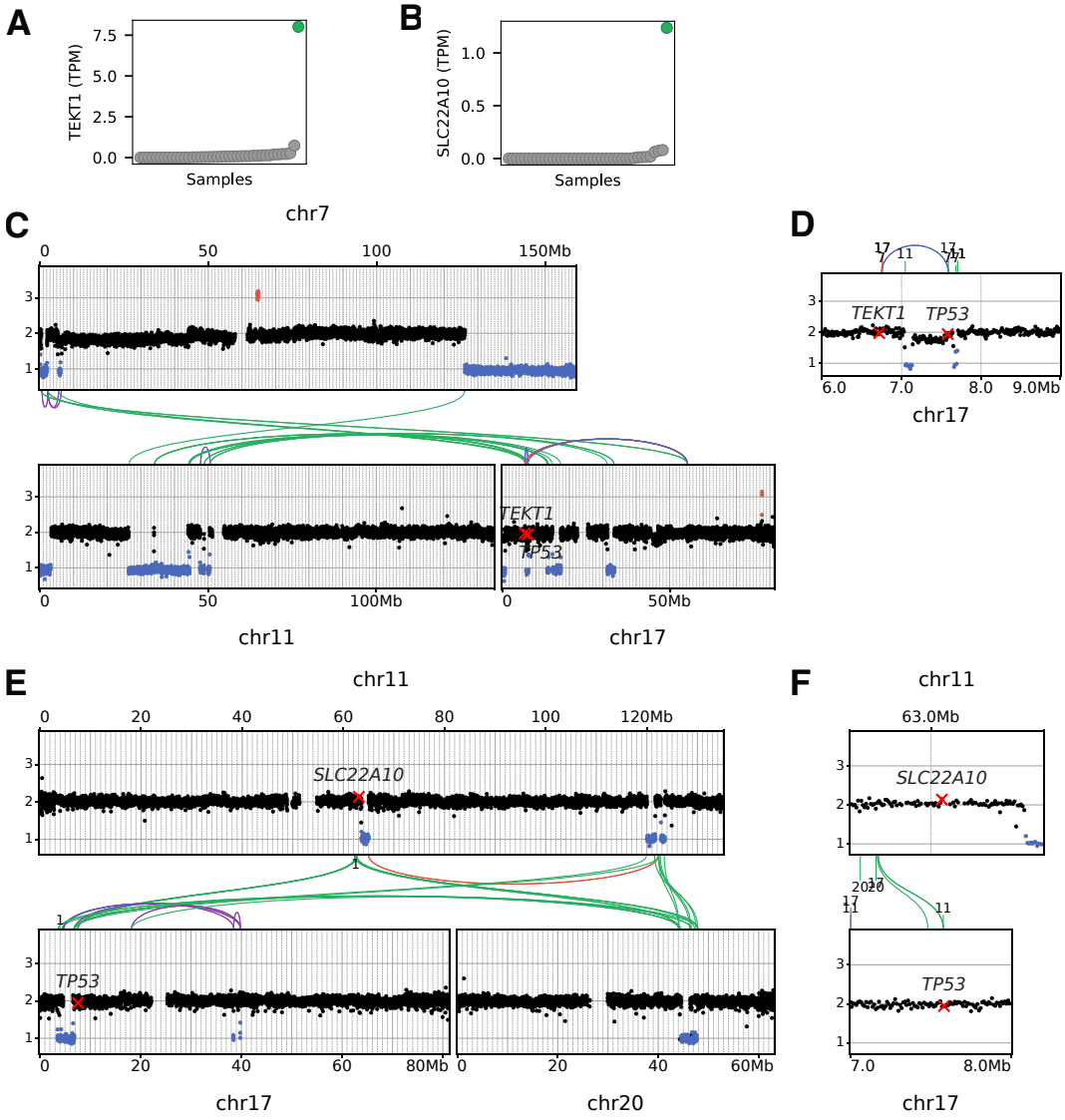

Figure 21: Example rearrangements leading to gene activation and *TP53* inactivation. **A**. *TEKT1* expression in all samples, with sample 16PB3075 (with breakpoint close to *TEKT1*) highlighted in green. **B**. *SLC22A10* expression in all samples, with sample 15KM20146 (with breakpoint close to *SLC22A10*) highlighted in green. **C**. Copy numbers and SVs on chromosomes 7, 11 and 17 in sample 16PB3075. **D**. Copy numbers and SVs around *TEKT1* and *TP53* in sample 16PB3075. **E**. Copy numbers and SVs on chromosomes 11, 17 and 20 in sample 15KM20146. **F**. Copy numbers and SVs around *SLC22A10* and *TP53* in sample 15KM20146.

3.2.3 Mapping of enhancer elements

Once an enhancer hijacking is predicted, whether by pyjacker or by any other method, the next step is to precisely identify the enhancer responsible for the activation of this gene. This is ultimately achieved by inserting the predicted enhancer element next to the target gene, and measuring if the gene becomes expressed, which would unequivocally prove that the predicted enhancer can activate the gene. However, this re-

quires knowledge of the most likely enhancer element. Enhancer marks (e.g. H3K27ac, H3K4me1, P300 measured by ChIP-seq, and ATAC-seq signal) in the correct cell type are very useful to predict active enhancers, but this is not always sufficient. Sometimes, several regions are equally good candidates based on enhancer marks, but only one of them can lead to gene activation. Another complementary method consists in analyzing multiple samples where the same enhancer is involved in enhancer hijacking, and mapping the minimal region which is always brought to the target gene. For the *GATA2* enhancer activating *MECOM*, Gröschel et al. analyzed 41 AML samples with t(3;3) or inv(3) and found that the breakpoint in the *GATA2* region were always located between *GATA2* (on the left) and *RPN1* (on the right) [90]. All breakpoints occurred at least 18kb to the left of *RPN1*. This 18kb region contained two putative enhancers, which were equally strong candidate based on enhancer marks. However, the fact that the breakpoint always occurred to the left of the leftmost enhancer in 41 samples was a strong indication that this enhancer was important for gene activation. Luciferase assays proved that this leftmost putative enhancer was able to activate genes in myeloid cell lines, while the rightmost enhancer was not. Here, I applied a similar method to precisely map the enhancer element in the *ETV6* and *CDK6* regions, which are recurrently involved in enhancer hijacking.

3.2.3.1 *ETV6* enhancer

The *ETV6* region is involved in multiple enhancer hijacking events. It can activate *MNX1* with the t(7;12) translocation in pediatric AML [92], *GSX2* with the t(4;12) translocation in adult AML [140], *BCL11B* with a t(12;14) [91] or *MECOM* with a t(3;12) [150].

In all samples analyzed, the breakpoints always occur within the TAD of *ETV6*, which extends up to *BCL2L14* (Figure 22, bottom). Within these regions, there are multiple H3K27ac peaks present in several myeloid cell lines. Two of them coincide with P300 peaks, and are therefore strong enhancer candidates: the first one is within intron 2 of *ETV6* at chr12:11,952,000 (hg19), and the second one is located close to *BCL2L14* at chr 12:12,165,000.

I analyzed 6 samples with t(7;12) and 9 samples with t(4;12) and in all cases, the breakpoint was in intron 1 or in intron 2 of *ETV6*, and the part to the right of the breakpoint was translocated to *MNX1* or *GSX2*. Among all these samples, the rightmost breakpoint was located at chr12:11,948,616 in sample T4 with a t(7;12) (Figure 22). The fact that all of these breakpoints are located to the left of the leftmost enhancer is already a very

strong indication that this enhancer might be relevant. Otherwise, some breakpoints could have occurred to its right, although the small number of samples does not allow to exclude the possibility of breakpoints occurring to the right.

I also analyzed one sample with a t(3;12) and one with a t(12;14). In these cases, the other side of the breakpoint gets into contact with the target gene (*MECOM* or *BCL11B*). Assuming that the same enhancer is hijacked with these translocations, this allows for a mapping of the enhancer from two sides, which drastically reduces the size of the region where the enhancer could be located: sample SJAML040681 with a t(12;14) [91] has a breakpoint located at chr12:11,956,435, so the enhancer should be located in the 8kb region chr12:11,948,616-11,956,435 (hg19). This region contains the leftmost P300 peak, which is therefore the most likely enhancer responsible for the activation of *MNX1*, *GSX2*, *MECOM* or *BCL11B* in these translocations.

In an iPSC/HSPC model, HSPCs with the translocation t(7;12) express *MNX1* while wild-type HSPCs do not [151]. The deletion of a 200kb region containing the two P300 peaks resulted in the loss of *MNX1* expression in HSPCs [92]. Based on my prediction that the leftmost P300 peak is likely the enhancer responsible for *MNX1* activation, Anna Riedel deleted this enhancer in the t(7;12) iPSCs, and observed that this abolished *MNX1* expression in HSPCs, confirming that this enhancer is necessary for *MNX1* activation. However, to date, this enhancer has not been inserted on its own, so it remains unknown whether this enhancer is sufficient for *MNX1* activation, or whether other elements in this region are required as well.

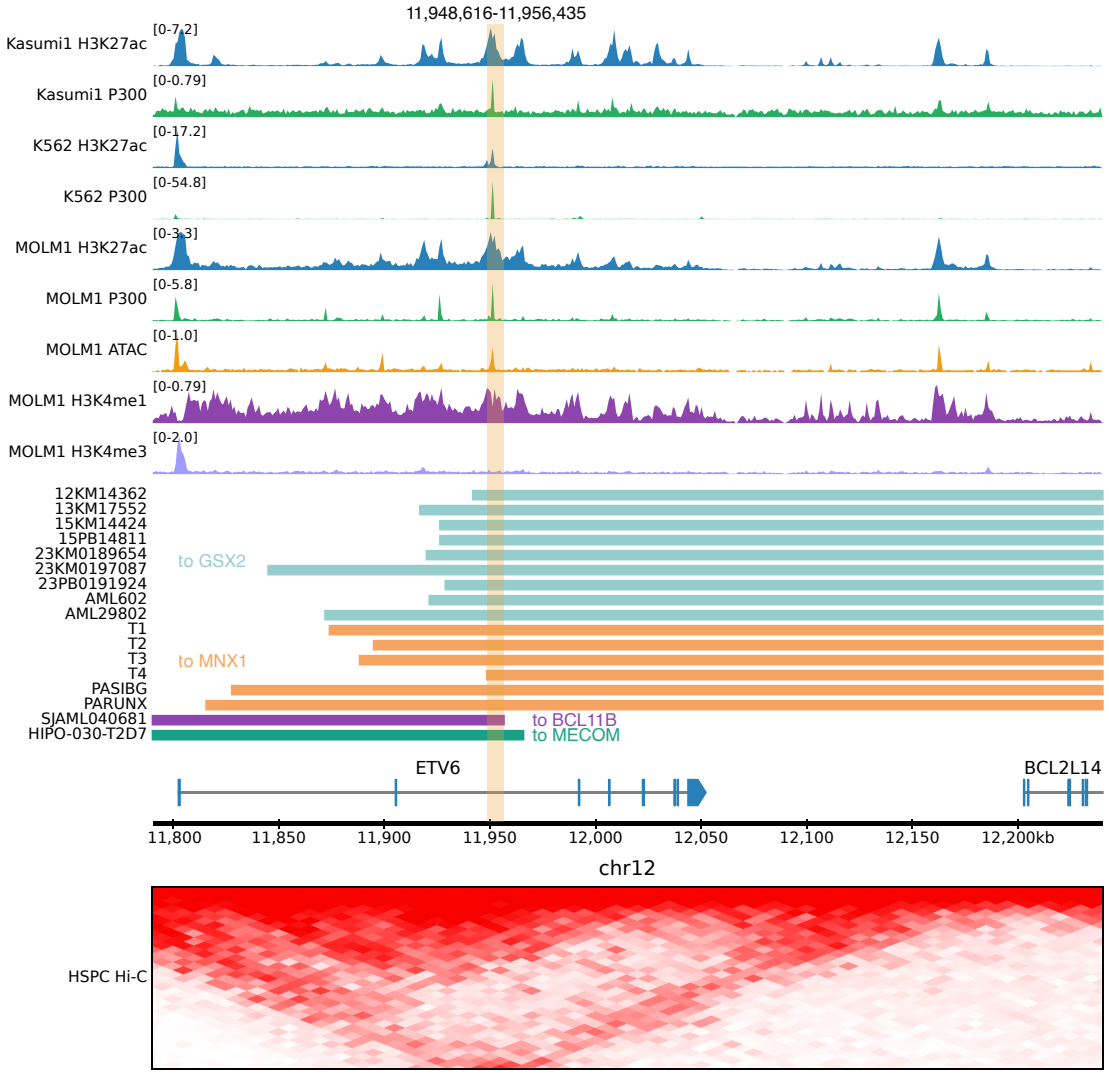

Figure 22: Enhancer mapping in the *ETV6* region. ChIP-seq and ATAC-seq tracks in myeloid cell lines are shown at the top. Regions translocated in samples with t(4;12), t(7;12), t(12;14) and t(3;12) are shown in the middle, where the color indicates the gene activated by the enhancer: turquoise for GSX2, orange for *MNX1*, purple for *BCL11B* and green for *MECOM*. Hi-C interactions in HSPCs in the region around *ETV6* are shown at the bottom. The region highlighted in orange is always translocated to the activated gene, and contains a putative enhancer based on enhancer marks.

3.2.3.2 *CDK6* enhancer

The *CDK6* region is also involved in several enhancer hijacking events, with different target genes: *MNX1* in del(7q) [152] or with the enhancer duplication observed in this ckAML cohort, *BCL11B* [91], and *MECOM* [138]. Sample 15PB8708 from the ASTRAL-1 ckAML cohort has a duplication of the region chr7:92187729-92415065 which is inserted next to *MNX1*. This already restricts the search for the enhancer to a 200kb region containing two P300 peaks: at chr7:92,268,000 and at chr7:92,384,500. I analyzed thirteen del(7q) samples with *MNX1* expression (either sequenced at the DKFZ or at

the MLL), as well as two t(3;7) samples with *MECOM* expression [138] and four t(7;14) samples [91], three of which had *BCL11B* expression confirmed. One t(7;14) sample (SJMPAL068288) was only profiled with WGS and not with RNA-seq, so *BCL11B* expression could not be confirmed. If this sample is excluded, then both putative enhancers are always brought close to the target gene, and several samples have a breakpoint very close to the rightmost enhancer, indicating that this enhancer could be required (Figure 23). However, the sample for which *BCL11B* expression was not assessed had a breakpoint located to the left of the rightmost enhancer. If this sample does express *BCL11B*, then this would imply that the rightmost enhancer is not required. In conclusion, for *CDK6*, I could not reliably identify a small enhancer element that is required for *CDK6* expression. It may also be that none of the two P300 peaks are sufficient on their own to drive the expression of the target gene, and that both are required.

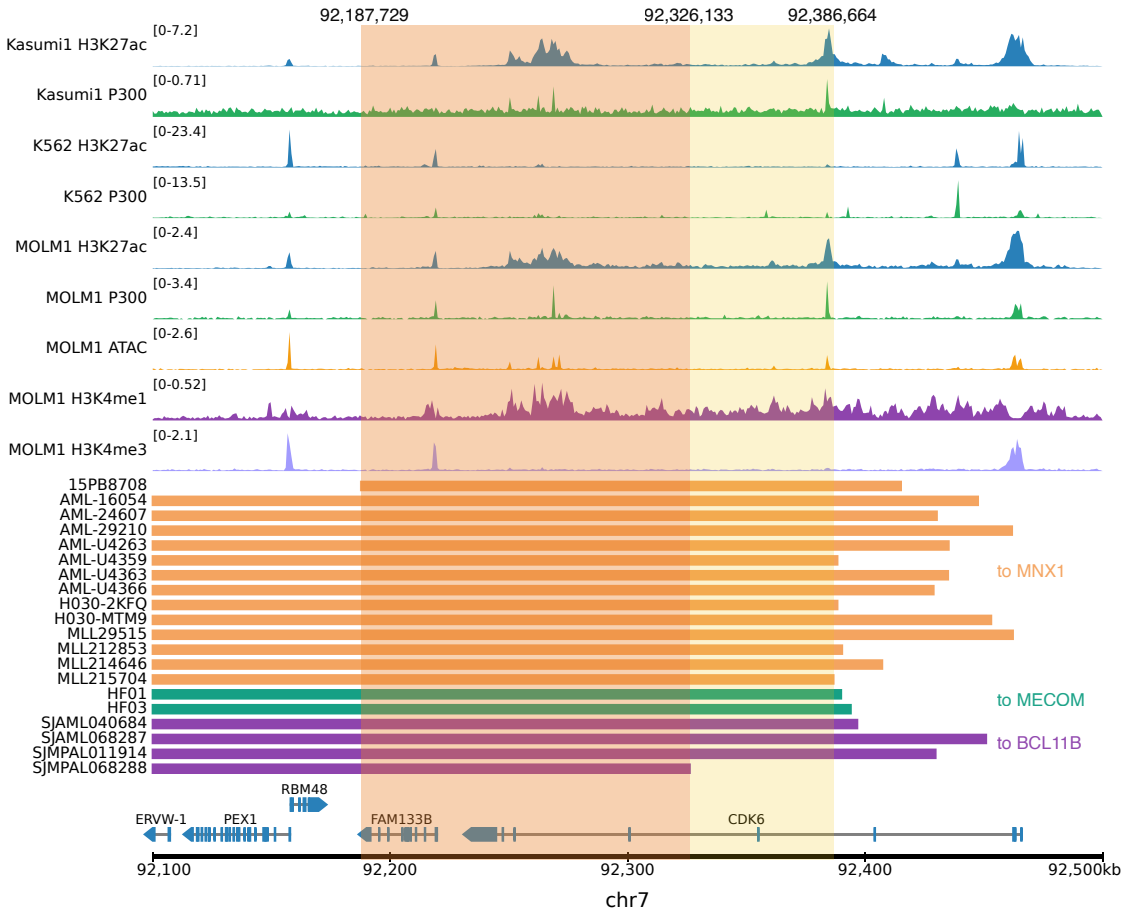

Figure 23: Enhancer mapping in the *CDK6* region. ChIP-seq and ATAC-seq tracks in myeloid cell lines, regions translocated in t(7;12), t(3;12) and t(12;14) in the region around *CDK6*. The region highlighted in red is always translocated to the activated gene, and the region in yellow is translocated to the activated gene in all but one sample.

3.2.4 Pyjacker applied to sarcoma

Although I developed pyjacker with the intent to apply it to ckAML samples, the method is general and can be applied to any cancer type. I collaborated with Simon Linder from the group of Stefan Fröhling at DKFZ to apply pyjacker to a large cohort of sarcoma samples (MASTER cohort). 639 samples were profiled with WGS and RNA-seq, from various sarcoma entities (Table 4). Pyjacker must be run on a homogeneous dataset where all tumor samples have a similar cell type as cell of origin. Otherwise, if different cancer types are combined, some events might be missed, if one gene is normally expressed in a particular cancer type, but only expressed in another cancer type when it is activated by enhancer hijacking. As a consequence, I did not run pyjacker on the whole sarcoma dataset at once, but I instead ran pyjacker on several homogeneous subsets. Since some tumor entities are similar, and since some entities had too few samples, I grouped similar tumor entities, resulting in 20 groups (Table 4). The very large number of samples resulted in a high number of pyjacker hits: 154 with FDR<5% and 1744 with FDR<20% across all 20 entities. To focus on interesting candidates, I only displayed genes activated in at least two samples (Figure 24). Similar to the ckAML dataset, this revealed both known and novel genes.

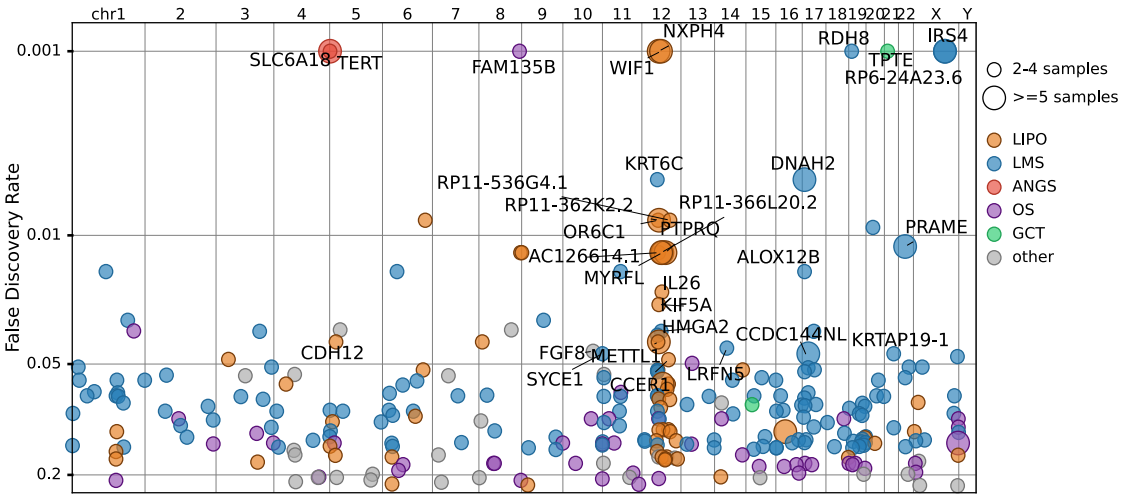

Figure 24: Overview of the genes identified by pyjacker for the sarcoma dataset.

TERT activated in several sarcoma entities

TERT (telomerase reverse transcriptase) was the gene found overexpressed in the highest number of samples: four angiosarcomas (ANGS), three liposarcomas (LIPO), one

Table 4: List of sarcoma groups to which pyjacker was applied.

group	entities	#samples
Leiomyosarcoma (LMS)	Leiomyosarcoma (LMS), Uterine leiomyosarcoma (ULMS)	113
Synovial sarcoma (SYNS)	Synovial sarcoma (SYNS)	36
Liposarcoma (LIPO)	Liposarcoma (LIPO), dedifferentiated liposarcoma (DDLS), well-differentiated liposarcoma (WDLS), pleomorphic piposarcoma (PLLS), myxoid/round-cell liposarcoma (MRLS)	60
Chordoma (CHDM)	Chordoma (CHDM)	51
Solitary fibrous tumor (SFT)	Solitary fibrous tumor (SFT)	24
Ewing sarcoma (ES)	Ewing sarcoma (ES), Ewing sarcoma of soft tissue (ESST)	44
Malignant Fibrous Histiocytoma (MFH)	Malignant Fibrous Histiocytoma (MFH)	22
Spindle cell sarcoma (SCSARC)	Spindle cell sarcoma (SCSARC)	21
Chondrosarcoma (CHS)	Chondrosarcoma (CHS), dedifferentiated chondrosarcoma (DDCHS), Myxoid chondrosarcoma (MYCHS), Mesenchymal chondrosarcoma (MCHS)	26
Osteosarcoma (OS)	Osteosarcoma (OS), osteoblastic osteosarcoma (OSOS), chondroblastic osteosarcoma (CHOS), high-grade surface osteosarcoma (HGSOS)	39
Angiosarcoma (ANGS)	Angiosarcoma (ANGS), breast angiosarcona (BA)	23
Low-Grade Fibromyxoid Sarcoma (LGFMS)	Low-Grade Fibromyxoid Sarcoma (LGFMS)	15
Rhabdomyosarcoma (RMS)	Low-Grade Fibromyxoid Sarcoma (PLRMS), alveolar rhabdomyosarcoma (ARMS), rhabdomyosarcoma (RMS), embryonal rhabdomyosarcoma (ERMS), spindle cell rhabdomyosarcoma (SCRMS)	30
Uterine sarcoma (USARC)	Endometrial stromal sarcoma (ESS), uterine leiomyoma (ULM)	15
Giant cell tumor (GCT)	giant cell tumor not otherwise specified (GCTNOS), giant cell tumor of bone (GCTB)	16
Alveolar soft part sarcoma (ASPS)	Alveolar soft part sarcoma (ASPS)	10
Epithelioid sarcoma (EPIS)	Epithelioid sarcoma (EPIS)	10
Desmoplastic small-round-cell tumor (DSRCT)	Desmoplastic small-round-cell Tumor (DSRCT)	9
Desmoid/aggressive fibromatosis (DES)	Desmoid/aggressive fibromatosis (DES)	8
Sarcoma not otherwise specified (SARCNOS)	Sarcoma not otherwise specified (SARCNOS), epithelioid hemangioendothelioma (EHAE), fibrosarcoma (FIBS), clear cell sarcoma (CCS), malignant phyllodes tumor of the breast (MPT), inflammatory myofibroblastic tumor (IMT)	67

leiomyosarcoma (LMS), one osteosarcoma (OS), one synovial sarcoma (SYNS), and one germinal cell tumor (GCT). This is in line with *TERT* being overexpressed in many cancer types, through various mechanisms. Here, I did not find a recurrent SV leading to *TERT* overexpression, but instead various types of rearrangements. In angiosarcomas, *SLC6A18*, which lies directly next to *TERT* on 5p, was overexpressed together with *TERT* when breakpoints were located near *TERT*.

IRS4 activated in LMS

IRS4 was found activated in eight leiomyosarcomas (LMS, including six uterine leiomyosarcomas) (Figure 25A), as well as one liposarcoma (LIPO) and one rhabdomyosarcoma (RMS). This gene had already been reported to be activated by enhancer hijacking in sarcomas [89]. All eight LMS samples with *IRS4* rearrangements had a deletion which overlapped the promoters of *COL4A5* and *COL4A6* (Figure 25B-D), which probably allowed the *IRS4* promoter to interact with their enhancers. Due to the lack of epigenetic data for the relevant cell types, it was not possible to identify putative enhancers.

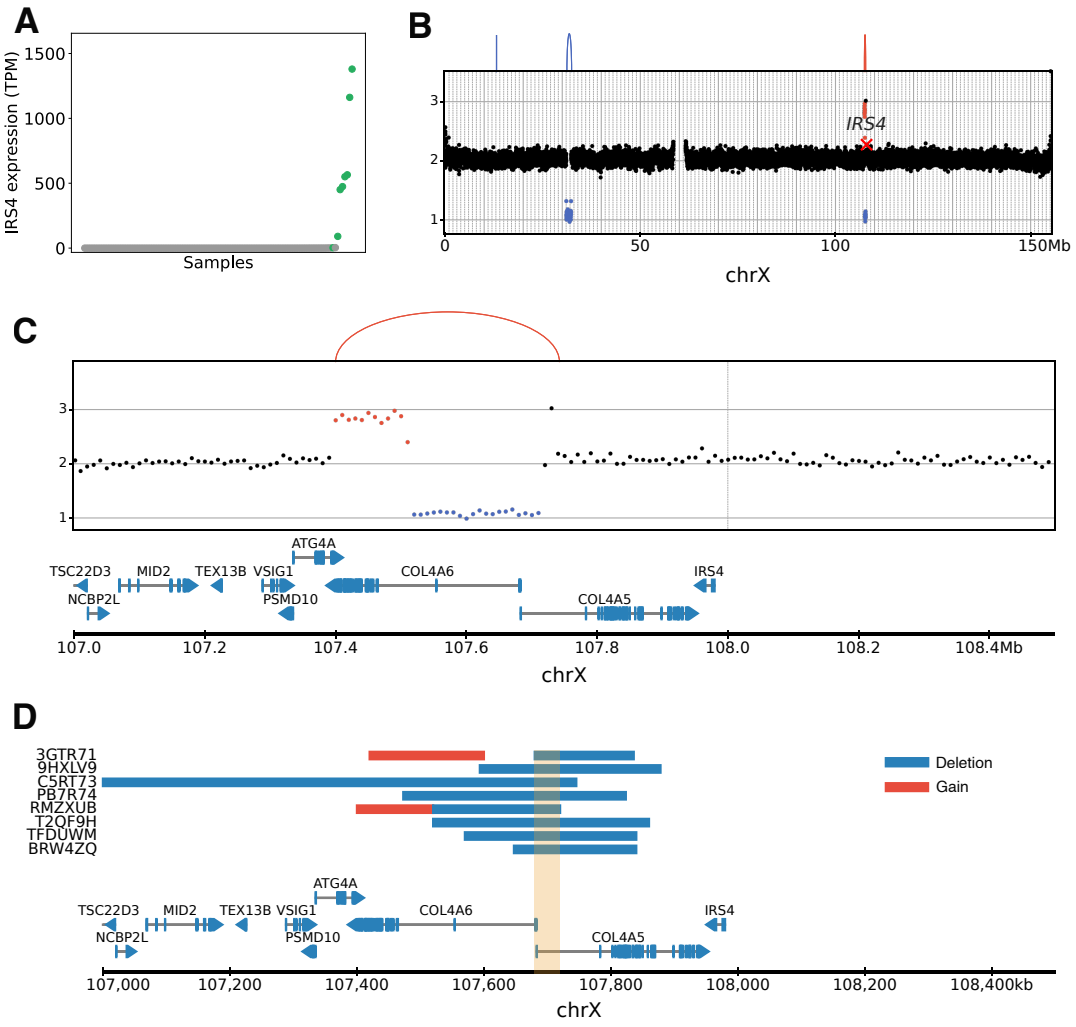

Figure 25: *IRS4* enhancer hijacking in leiomyosarcoma. **A**. *IRS4* expression in the leiomyosarcoma cohort, with the 8 rearranged samples highlighted in green. **B**. Copy numbers and SVs on chrX for sample RMZXUB, with *IRS4* rearrangements. **C**. Copy numbers and SVs around *IRS4* for sample RMZXUB. **D**. Summary of the rearrangements leading to *IRS4* overexpression in the LMS cohort, with the minimally deleted region highlighted in orange.

FGF8 activated in LGFMS and SCSARC

A novel interesting gene identified by pyjacker is fibroblast growth factor 8 (*FGF8*). This gene is overexpressed in several cancer types, including alveolar rhabdomyosarcoma (ARMS) [153], but has not been reported to be activated by enhancer hijacking. I found *FGF8* activated by an imbalanced t(1;10) translocation in four samples (three low-grade fibromyxoid sarcomas (LGFMS) and one spindle cell sarcoma (SCSARC)) (Figure 26). Again, the lack of epigenetic data in the relevant cell types prevented me from identifying the putative enhancers. I also confirmed that *FGF8* is overexpressed in ARMS (data not shown), but without SVs, indicating that different mechanisms can lead to *FGF8*

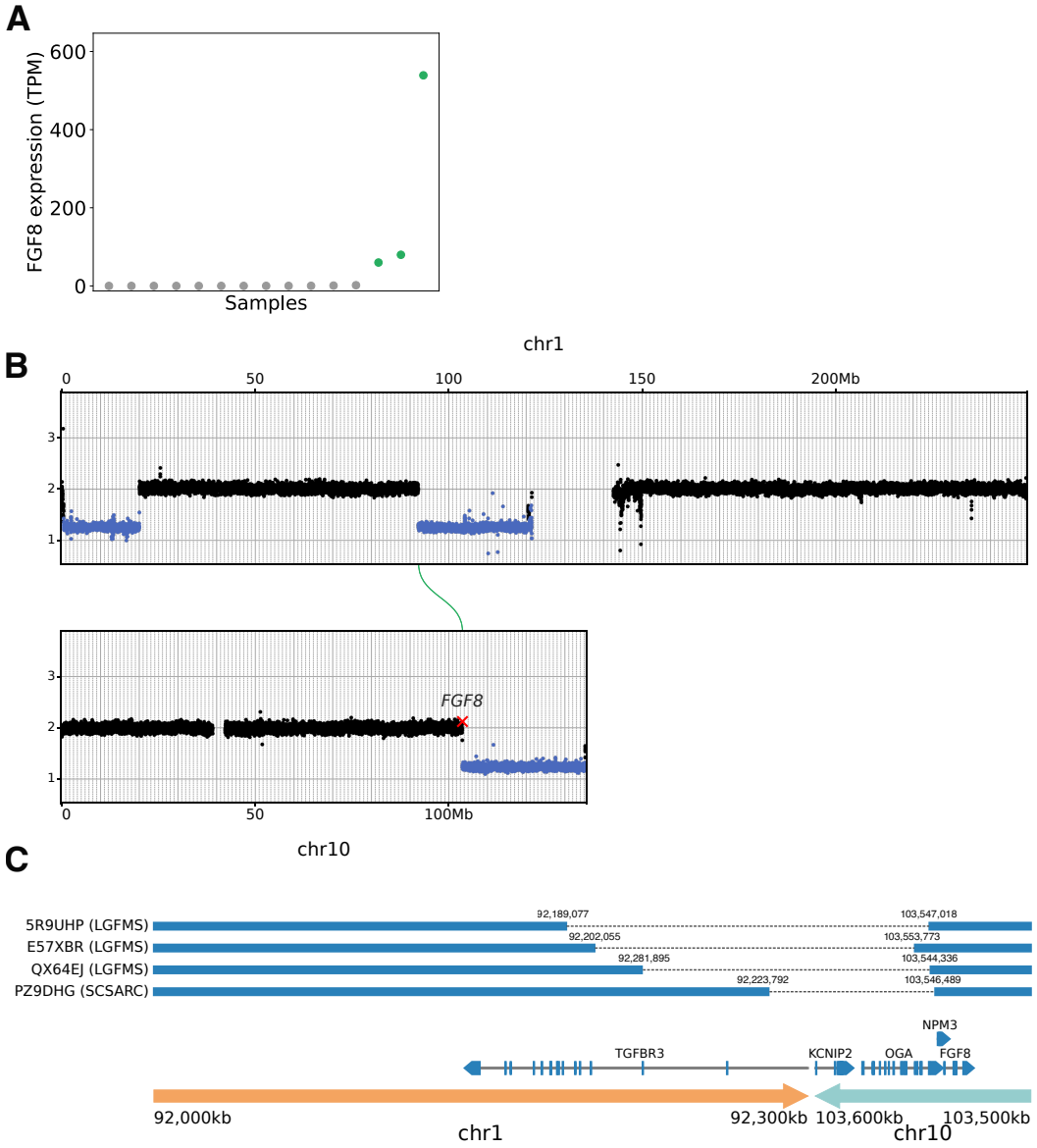

Figure 26: *FGF8* enhancer hijacking in sarcoma. **A**. *FGF8* expression in the low-grade fibromyxoid sarcoma (LFGMS) cohort, with the 3 rearranged samples highlighted in green. **B**. Copy numbers and SVs on chr1 and chr10 for sample 5R9UHP, with t(1;10). **C**. Summary of the rearrangements leading to *FGF8* overexpression in the 3 LGFMS samples and the SCSARC sample with t(1;10).

overexpression.

Chromothripsis on chr12 in liposarcomas lead to the activation of various genes

Pyjacker identified many putative enhancer hijacking events on chr12 in liposarcomas (Figure 24), including *GLI1*, *NXPH4*, *WIF1*. Almost all liposarcoma samples have a chromothriptic chr12 with massive amplifications, and often other chromosomes are

involved in the complex rearrangements (Figure 27A-C). This high frequency of chromothripsis on chr12 in liposarcomas has been reported before [154]. The main driver event that is selected for with these amplifications is thought to be the overexpression of *MDM2*, a negative regulator of *TP53* [155]. Other genes are often frequently co-amplified together with *MDM2*, including *CDK4* and *HMGA2*. *CDK4* phosphorylates the retinoblastoma protein, thus allowing progressing into the S phase of the cell cycle [156], so its overexpression may driver cancer. Pyjacker corrects the gene expression for amplifications in order to specifically select enhancer hijacking events, and not amplifications. Consequently, *MDM2* and *CDK4* were not identified by pyjacker, since these genes are expressed in all samples, and that their overexpression is explained by the higher copy number. However, *HMGA2* was identified by pyjacker, indicating that the SVs in its vicinity lead to an increased expression beyond what would be expected from the increased copy number. Pyjacker identified several genes which are activated more rarely and that may have been overlooked before. *NXPH4* was among the top pyjacker hits and is overexpressed in various cancer types, including lung cancer [157], but through different mechanisms than enhancer hijacking. *WIF1* (Wnt inhibitory factor 1) is a more surprising candidate since it inhibits Wnt, which is usually upregulated in cancer, so *WIF1* is generally considered to be a TSG in many cancer types [158]. Therefore the overexpression of *WIF1* may be a passenger event, or this gene may have a unique role in liposarcomas. *GLI1* was overexpressed only in 3 samples, but with a very strong overexpression (Figure 27D). It is involved in the Sonic Hedegehog pathway and has been hypothesized to play a role in some cancer types [159].

Potential false positives because of metastasis when allele-specific expression is not available

Among the pyjacker hits for which no SNP information was available (and for which the FDR was therefore rather high), I noticed that several false positives could be due to contamination from normal cells at the metastatic site. Indeed, if the sample comes from a metastasis to a rare site and contains normal cells from this site, it will have a high expression for the genes normally expressed in this tissue. If some of these genes happen to be near a breakpoint, pyjacker will detect high outlier expression near a breakpoint, and identify this as a candidate enhancer hijacking event. In such cases, the expression should be biallelic, but if no heterozygous SNPs are present, pyjacker will still report it, albeit with a high FDR. For example, *REG3A* was reported as a putative enhancer hijacking event in one leiomyosarcoma sample, based on a high expression near a breakpoint, but without SNP information. This gene is normally ex-

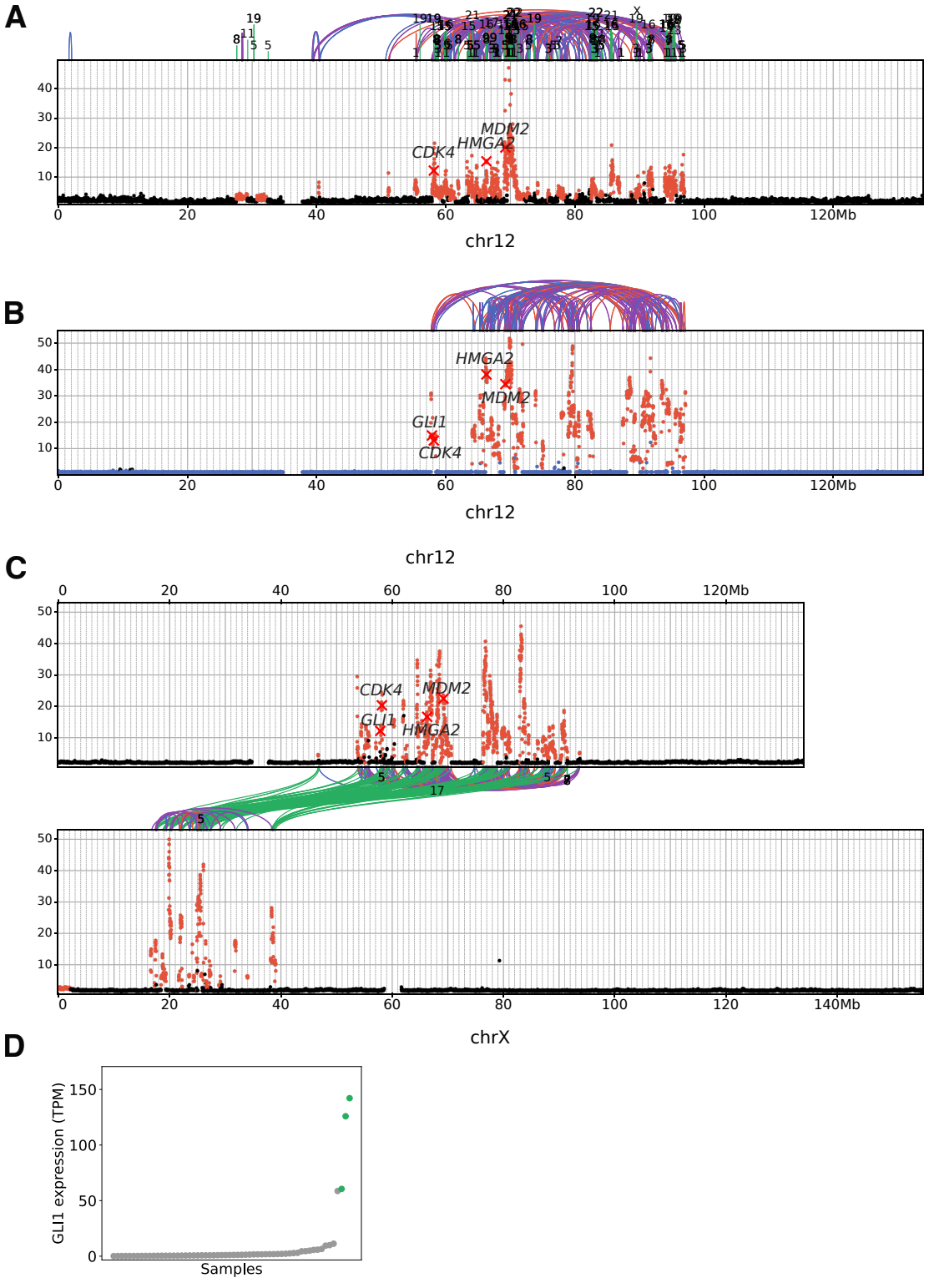

Figure 27: Chr12 rearrangements in liposarcoma. **A-C**. Copy numbers and SVs on chr12 in three liposarcoma samples: 1GLAY, CB5S4F and QFC8A8. **D**. *GLI1* expression among the liposarcoma samples, with pyjacker hits highlighted in green.

pressed in the pancreas and intestines. In this sample, I found a very high expression of insulin and glucagon (highly expressed in the beta and alpha cells of the pancreas, respectively), indicating that this sample is probably a metastasis to the pancreas, and that the *REG3A* expression likely comes from the pancreas cells and not from the tumor cells. In the case of pancreas, there are some clear marker genes (insulin and glucagon), but contamination from other tissues may not always be so easy to detect. Hence, I did not implement an automatic detection of aberrant expression because of contamination from cells from the metastatic site. This potential source of false positives should be taken into account when analyzing pyjacker results from metastatic samples. A possible future improvement of pyjacker would be to perform cell type deconvolution of RNA-seq data to account for this contamination.

3.2.5 Pyjacker applied to prostate cancer

I applied pyjacker to a cohort of 63 prostate cancer samples profiled with WGS and RNA-seq. This resulted in 18 genes activated by structural rearrangements with FDR<5% and 64 with FDR<20% (Figure 28A). However, the large majority were fusion genes, most of which are already known. By far the most recurrent event was a *TMPRSS2*::*ERG* fusion (leading to *ERG* upregulation), which was found in 26 samples, often with a simple deletion between *ERG* and *TMPRSS2*, and sometimes with more complex events. This *TMPRSS2*::*ERG* fusion is indeed known to be the most common somatic event in prostate cancer, occurring in approximately half of all prostate cancer cases [160]. One sample had an *ERG* fusion with *FANCC* instead of *PRSS2*, and three other members of the ETS family of transcription factors, ETV1, ETV4 and ETV5, were among the top pyjacker hits and were upregulated through fusions in 4, 3, and 1 samples respectively.

If I exclude fusions, the list of candidate genes identified by pyjacker becomes much smaller, without any recurring event. One interesting candidate is *NRP2*, found upregulated in a single sample, with monoallelic expression (Figure 28B-C). This gene has been reported in relation to prostate cancer, because it may favor bone metastasis [161] and may also be involved in neuroendocrine-like prostate cancer [162]. Thus, its activation by enhancer hijacking may be a driver event in prostate cancer. In this sample ICGC_PCA161, a breakpoint brought *NRP2* close to a putative enhancer within *CSGALNACT1* on chr8 (Figure 28D).

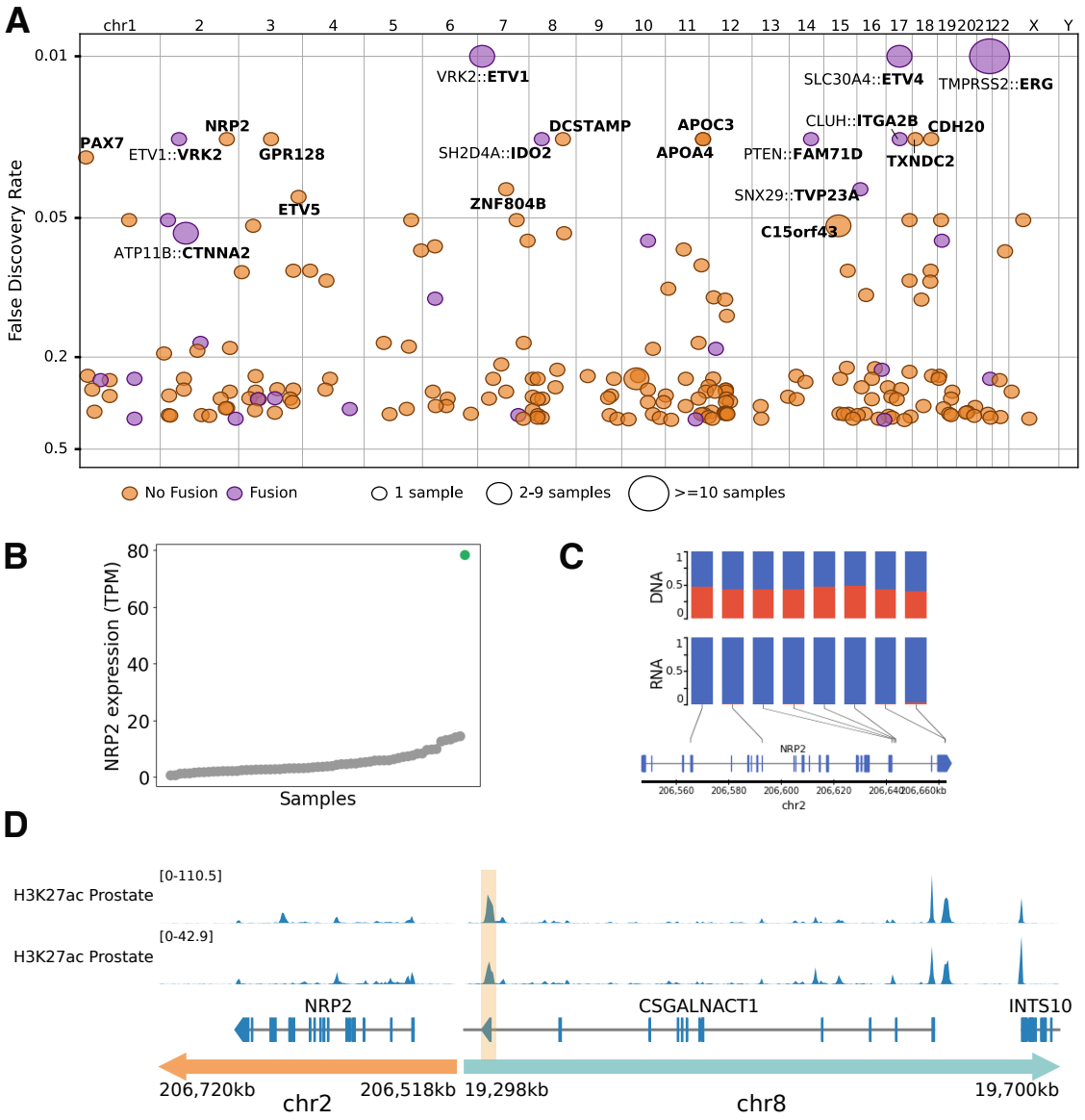

Figure 28: Enhancer hijacking in 63 prostate cancer samples. **A**. Overview of the top pyjacker hits in the cohort of 64 prostate cancer samples. **B**. *NRP2* expression in the 63 prostate cancer samples, with sample ICGC_PCA161 with a breakpoint near *NRP2* highlighted in green. **C**. Variant allele frequency of SNPs in *NRP2* in DNA and RNA, for sample ICGC_PCA161. **D**. Breakpoint linking *NRP2* to a putative enhancer within *CSGALNACT1* on chr8. ChIP-seq against H3K27ac for two prostate cancer samples from the ENCODE project were used (file IDs: ENCFF648XQW and ENCFF364WFY).

3.3 Allele-specific methylation with nanopore sequencing

Nanopore sequencing is a third generation sequencing technology which provides methylation information as well as long reads that can be phased to each of the two parental haplotypes. In this section, I investigated what information can be gained from this allele-specific methylation (ASM) information, and in particular whether it could help to detect enhancer hijacking events.

3.3.1 Methylation detection with nanopore sequencing

15 AML samples and two AML cell lines were profiled with nanopore sequencing (sequencing performed by Jessica Heilmann). Each sample was sequenced on one PromethION flow cell, except two (15KM12995 and 15KM20146) which were sequenced on two different flow cells, with two different size selection protocols. Most samples (17/19) had an N50 higher than 15kb, meaning that more than 50% of their coverage comes from reads longer than 15kb (Figure 29), which is sufficient to phase most reads. Most samples also had a good coverage > 25x, except one (16KM16045) which had a coverage of only 11x and was excluded from the subsequent analyses. When the size selection was more stringent (SRE kit), the N50 was generally higher, although it also varied a lot from sample to sample. For the two samples that were sequenced twice with different size selection kits, the more stringent size selection resulted, as expected, in longer reads. There is usually a negative correlation between coverage and read length with nanopore data, but this was not observed here because a narrow range of read lengths was used; the coverage might drop if the N50 went beyond 30kb. The median sequence identity was higher than 99% (99.2% on average), meaning that the sequencing error rate is lower than 1%.

Nanopore sequencing directly provides base modification information, without the need for bisulfite treatment. Since the 15 AML samples profiled with nanopore sequencing had also been profiled with EPIC array, I verified that the nanopore data had good correlation with EPIC array, for the CpGs targeted by the EPIC array. The correlation was higher than 0.83 for all samples, and higher than 0.90 for most samples (Figure 30A-B). The genome-wide methylation level was about 80% for most samples, but only about 22% in CpG islands (Figure 30C), in agreement with previous reports [3]. Using the RNA-seq data, I defined expressed genes as genes with expression greater

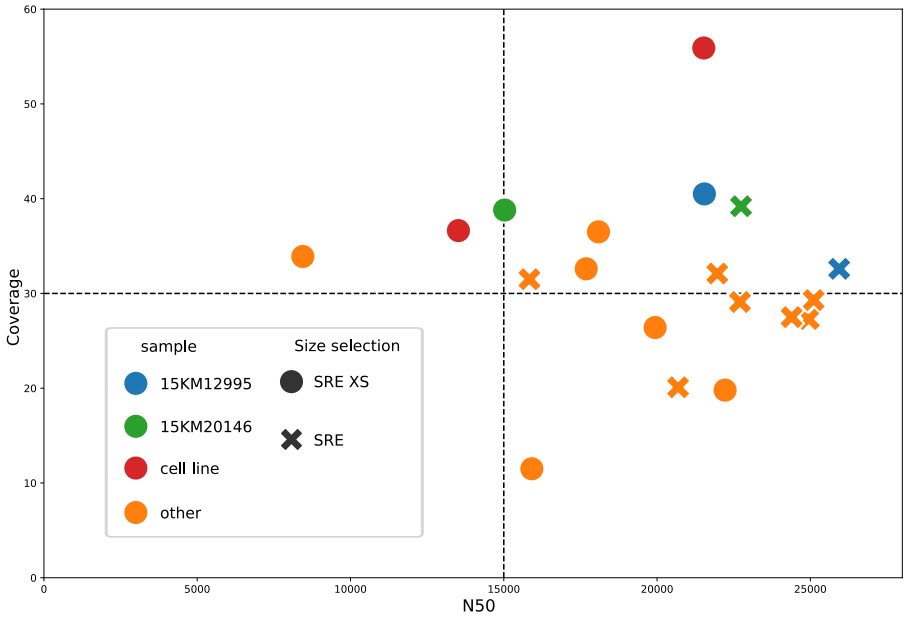

Figure 29: Coverage and N50 of the nanopore runs. The shape of the dots indicate the size selection kit that was used (SRE is more stringent than SRE XS).

than 5 TPM in all samples, and unexpressed genes as genes with expression lower than 0.2 TPM in all samples. When restricting CpG islands to those that overlapped promoters of expressed genes, the methylation level was even lower (mean: 2.4%), but promoters of unexpressed genes had similar methylation levels as other CpG islands. Nanopore sequencing also provides hydroxymethylation information, but this base modifications is much rarer (1.7% genome-wide).

The long reads provided by nanopore sequencing typically cover several germline heterozygous SNPs, so can be phased to each of the two parental haplotypes. Combined with methylation, this provides allele-specific methylation information. I first verified that I could observe allele-specific methylation at imprinted genes, for example *H19* (Figure 31A). Female cells inactivate one X chromosome, and this inactivation is inherited through cell division, so all cells from the same tumor should have the same allele inactivated. When looking at genes on the X chromosome, for example *IGBP1*, I could indeed see that, for female samples, one allele was unmethylated at the promoter and the other allele was methylated (Figure 31B). The active allele had higher methylation levels within the gene body, consistent with the fact that the active X chromosome has on average higher methylation [13].

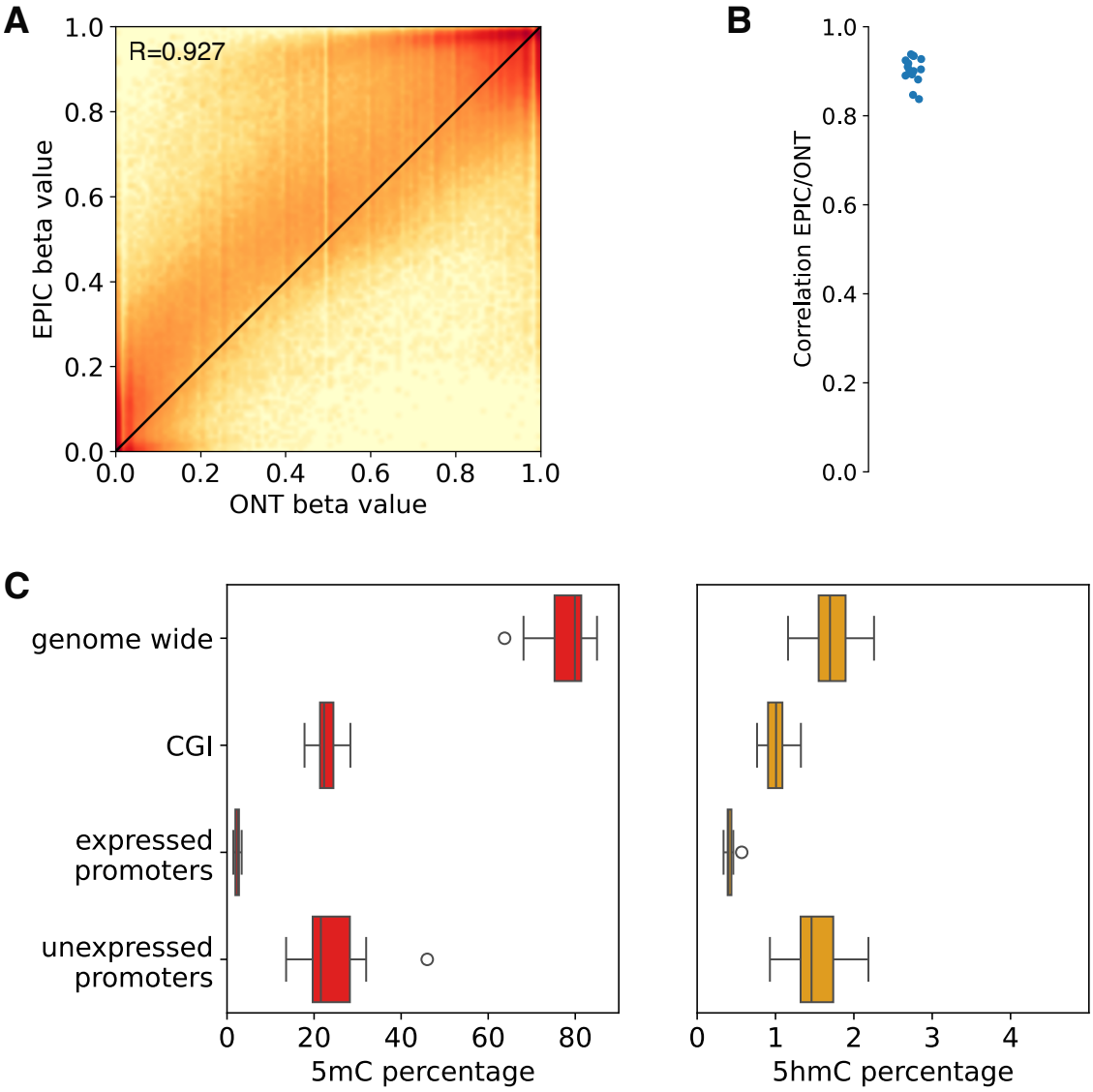

Figure 30: Overview of methylation data from nanopore sequencing data. **A**. Heatmap showing the correlations between the beta values measured from EPIC array and the beta values measured with nanopore sequencing, for sample 15PB8708 and all CpGs covered with the EPIC array and covered by at least 6 nanopore reads. Red indicates higher density. **B**. Correlations between ONT and EPIC array data, for all 15 samples profiled with nanopore sequencing. **C**. Average methylation (5mC) and hydroxymethylation (5hmC) for each sample, by considering different types of regions: all CpGs (genome wide), only CpGs in CpG islands (CGI), only CpGs located in CpG islands at promoters of expressed genes, or only CpGs located in CpGs islands at promoters of unexpressed genes.

3.3.2 Within-sample methylation heterogeneity

At a particular location, a read is either methylated or unmethylated, but in a bulk sample there might be heterogeneity across reads, leading to methylation frequencies between 0 and 1. This can have several reasons, the main one being a mixture of different cell types, and several metrics have been defined to measure this heterogeneity [163]. For simplicity, I only considered CpG islands on autosomes, and I computed, for each

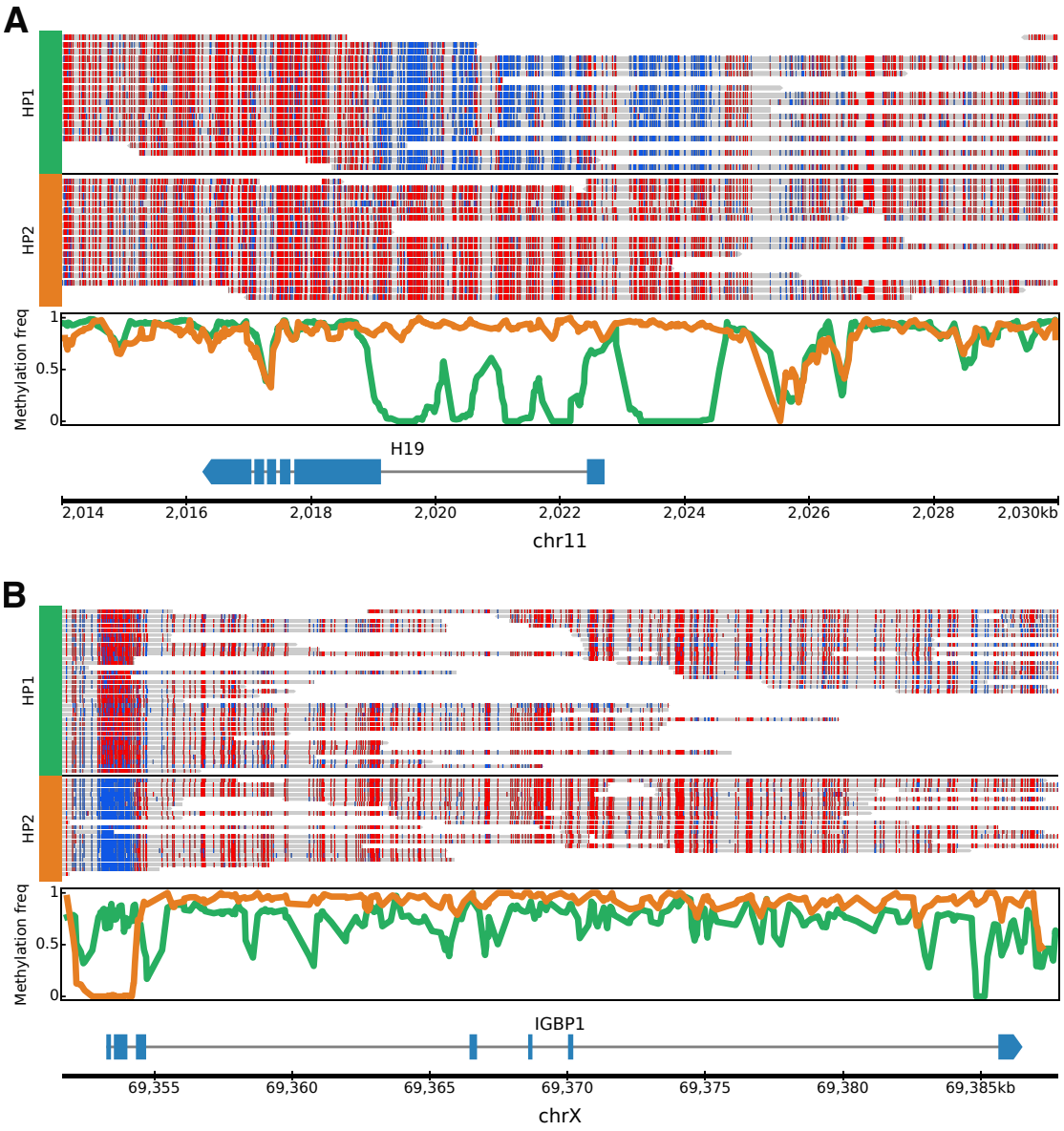

Figure 31: Allele-specific methylation with nanopore sequencing. **A**. Alignments around *H19* for sample 15PB8708, grouped by haplotype and colored by methylation (blue: unmethylated, red: methylated). **B**. Alignments around *IGBP1* for sample 16KM16320, grouped by haplotype and colored by methylation (blue: unmethylated, red: methylated).

read, its average methylation level within the CGI. I then computed the standard deviation across all reads, and I defined a CGI to have a high heterogeneity in a sample if this standard deviation is greater than 0.3 (when methylation values are between 0 and 1). I also defined an allele-specific heterogeneity, by computing separately the standard deviations from reads of each haplotype, and then averaging the heterogeneities within each haplotype. This is exemplified in Figure 32, where sample 15KM12995 does not have heterogeneity in the region displayed, sample 15KM15252 has heterogeneity, but only across and not within haplotypes, and sample 15PB9630 is heterogeneous, even

within haplotypes.

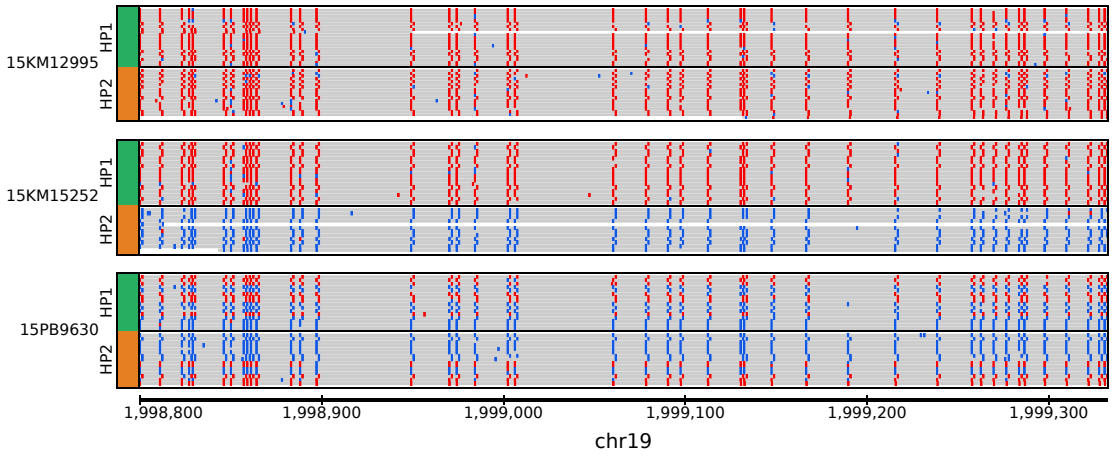

Figure 32: Example of within-sample methylation heterogeneity. Alignments colored by methylation and split by haplotype for three samples, in a region with methylation heterogeneity in some samples.

Approximately 1.5% of CGI had a high within-sample heterogeneity (Figure 33A). When looking at heterogeneity within haplotypes, the two cell lines GDM-1 and MUTZ-3 were not heterogeneous anymore, as expected for a homogeneous population of cells. For most patient samples, the methylation heterogeneity also disappeared when looking within haplotypes. Sample 15KM19129 was a strong outlier, since it showed heterogeneity for 8% of CGIs, and remained heterogeneous when looking within haplotypes. This high methylation heterogeneity in one sample could be due to a low tumor purity, but based on CNAs the purity was 89% in this sample, similar to most samples. Sample 15KM19129 has an enhancer hijacking of *BCL11B*, which is associated with mixed-phenotypes leukemia, so the heterogeneity could be due to some cells having a myeloid phenotype while other cells being more lymphoid. However, I did not observe differential methylation between myeloid and lymphoid progenitors at regions with allele-specific methylation in sample 15KM19129, using data from Farlik et al. [164]. In conclusion, for most samples the methylation heterogeneity at CGIs is largely allele-specific, and I could not find an explanation for the very high methylation heterogeneity in sample 15KM19129.

I then focused on allele-specific methylation. For each sample and each CGI, I considered that there was allele-specific methylation if the absolute difference in average methylation between the two haplotypes was greater than 0.5. For each CGI, I counted the proportion of samples with allele-specific methylation. The large majority of CGIs did not show allele-specific methylation (Figure 33B). 18 CGIs, corresponding to 13 genes, had allele-specific methylation in more than 80% of the samples. I checked in databases of imprinted genes [165, 166], and found that all of those genes, except

SNU13, had been reported as imprinted genes in at least one database (Table 5). For example *H19* is a well known imprinted gene [167]. Many CGIs showed allele-specific methylation in 20-50% of samples. This might be because the methylation state is influenced by nearby SNPs, but I could not test this because the number of samples was low, and SNPs might potentially impact methylation over long distances, leading to a very high number of tests.

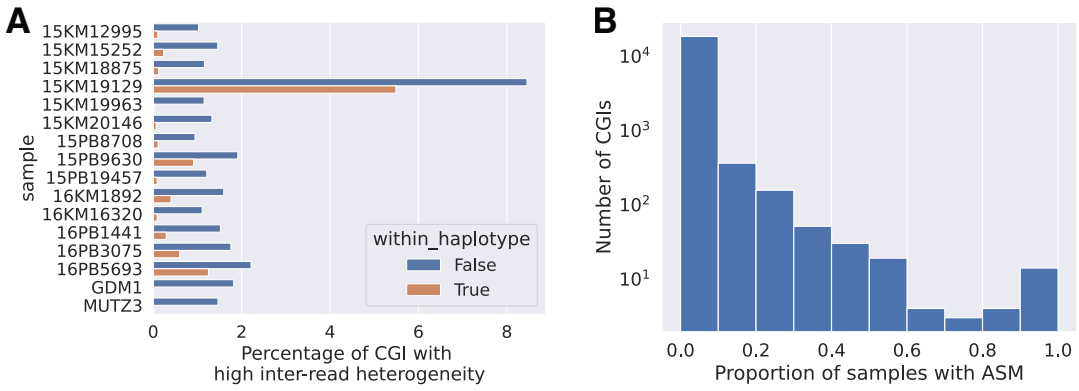

Figure 33: **A**. Proportion of CGIs having a high methylation heterogeneity (std>0.3) in each sample, either by considering heterogeneity across all reads (blue) or by considering heterogeneity only within haplotypes (orange). **B**. Number of CGIs showing allele-specific methylation (ASM) in different proportions of samples, with a logarithmic scale for the y-axis.

3.3.3 Allele-specific methylation in cases of enhancer hijacking

Enhancer hijacking leads to the activation of a single allele, and I wondered whether this would lead to methylation differences between the two alleles. I first started by looking at cell lines with known enhancer hijacking events: GDM-1 has a translocation t(6;7) leading to *MNX1* activation [168], and MUTZ-3 has an inv(3) leading to *MECOM* activation [90] as well as a t(12;22) leading to *MN1* activation [169]. In these three cases, I observed clear hypomethylation at the promoter of the gene activated by enhancer hijacking in the rearranged allele, compared to the wild-type allele (Figure 34). This was very promising, because such allele-specific methylation could be used as a criterion to detect enhancer hijacking, especially when no SNPs are covered in the RNA-seq data and allele-specific expression cannot be assessed.

However, the methylation state of cell lines often differs from primary cancer samples, and in particular cell lines often have the promoters of their unexpressed genes hypermethylated [84, 85]. When I analyzed AML samples, I realized that in patient samples most promoters are unmethylated, even when the gene is not expressed, so there cannot be a strong hypomethylation in cases of enhancer hijacking. Neverthe-

Table 5: List of CGIs with allele-specific methylation in more than 80% of samples. The column "% ASM" indicates the percentage of samples with allele-specific methylation, the column "geneimprint" indicates whether the gene was listed as imprinted in geneimprint.com [165], and the column "Court2014" indicates whether the CGI was reported as imprinted in Court et al. [166].

CGI (hg19)	Gene	% ASM	geneimprint	Court2014
11:2019565-2019863	H19	100	YES	YES
19:54040812-54041857	ZNF331	100	NO	YES
19:54057414-54058254	ZNF331	100	NO	YES
15:25200035-25201054	SNURF	100	YES	YES
7:50849752-50850871	GRB10	100	YES	YES
20:57429024-57431239	GNAS	100	YES	YES
22:42078060-42078549	SNU13	100	NO	NO
8:141107837-141110984	PEG13	100	YES	YES
11:2720410-2722087	KNCQ1	100	YES	YES
20:42143210-42143591	L3MBTL1	100	YES	YES
14:101292043-101292709	MEG3	93	YES	YES
6:144328916-144329847	PLAGL1	92	YES	YES
19:57351283-57351995	PEG3	92	YES	YES
16:3493098-3493569	ZNF597	91	YES	YES
20:57415135-57417153	GNAS	89	YES	YES
20:57426729-57427047	GNAS	89	YES	YES
19:57349997-57350470	PEG3	85	YES	YES
15:25018174-25018533	SNURF	82	YES	NO

less, for most of the genes identified by pyjacker for which the corresponding sample had been profiled with nanopore, I observed a slight hypomethylation on the rearranged allele (Figure 35). The WT allele is already mostly unmethylated, but some CpG sites are methylated, whereas the rearranged allele is fully unmethylated. In addition, the unmethylated region was wider on the rearranged allele, extending beyond the promoter. Nevertheless, the methylation profile was only mildly altered by the enhancer hijacking, and the effects were not systematic, so I did not integrate the allele-specific methylation into the enhancer hijacking detection with pyjacker.

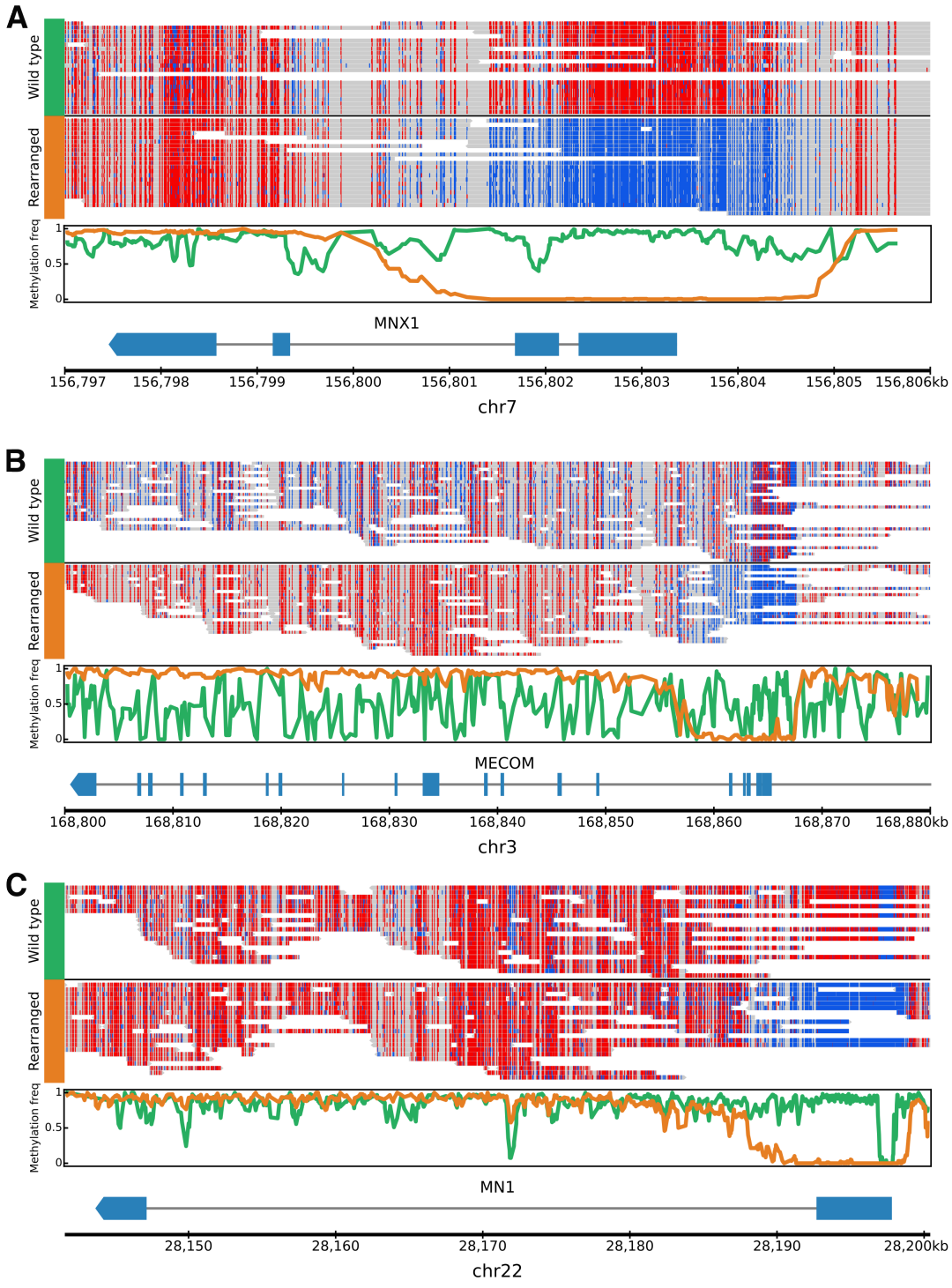

Figure 34: Allele-specific methylation for cell lines with enhancer hijacking. Nanopore reads colored by methylation and split by haplotype, around *MNX1* in GDM-1 (**A**), around *MECOM* in MUTZ-3 (**B**), and around *MN1* in MUTZ-3 (**C**).

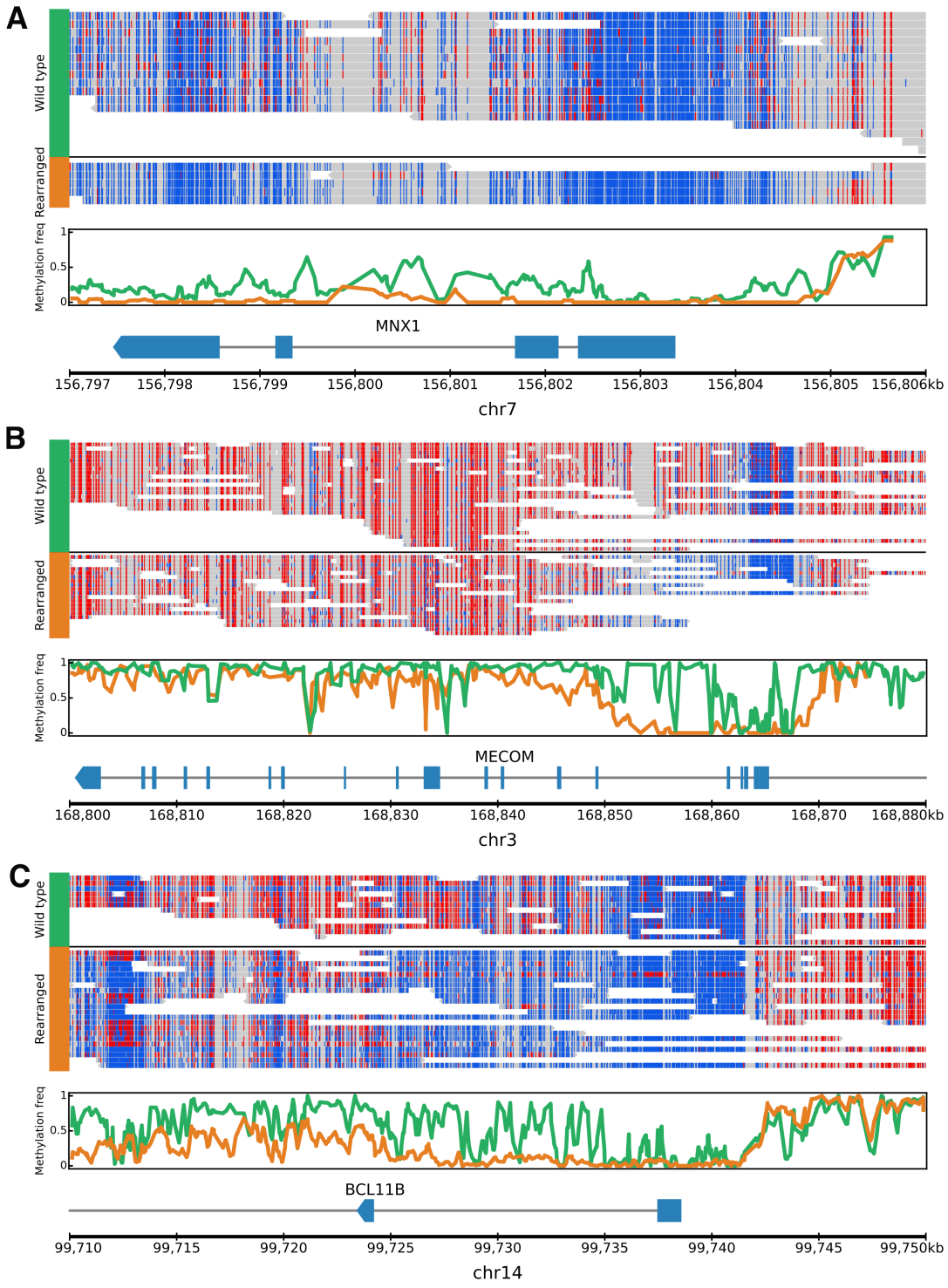

Figure 35: Allele-specific methylation for AML samples with enhancer hijacking. Nanopore reads are colored by methylation and split by haplotype, around *MNX1* in 15PB8708, *MECOM* in 15PB19457 and *BCL11B* in 15KM19129.

CHAPTER 4

DISCUSSION

4.1 Somatic alterations in ckAML

My analysis into the somatic alterations in ckAML mainly confirmed existing knowledge, but provided some novel insights thanks to the use of WGS and to the large number of samples analyzed. CkAML samples frequently harbour *TP53* mutations, in which case they typically have more CNAs and more complex rearrangements, including chromothripsis. The detection of breakpoints with WGS allowed me to show that in most cases, chromothripsis events involve multiple chromosomes, and some samples have several independent chromothripsis events involving different chromosomes. On the one hand, ckAML is very heterogeneous since each sample has a unique combination of rearrangements. On the other hand, ckAML is somewhat homogeneous because some events are very recurrent and found in almost all samples, like *TP53* mutations, del(5q) and del(7q). Although there are some patterns of co-occurrence and mutual exclusivity between CNAs, there are no clear subgroups within ckAML, neither at the genomic nor at the transcriptomic level. Deletions are more common than gains, and, apart from *TP53* in 17p, they do not lead to the biallelic inactivation of a TSG. Using RNA-seq data, I showed that the large majority of genes have a reduced expression when they are deleted. The haploinsufficiency of the genes in the deleted regions is therefore the likely reason for the positive selection of these deletions. However, it is very difficult to study how these deletions may drive cancer because they are often very large, especially in 5q and 7q, and therefore encompass many genes. Typically, minimally deleted regions across a large number of samples are identified, and this is used to map the candidate genes whose haploinsufficiency may drive AML [170]. However, for 5q and 7q, the deletions are larger than 50Mb for 90% of samples, so it is likely that many genes, even outside the minimally deleted regions, are relevant. Del(5q) is interesting because on the one hand, it is the CNA most strongly co-occurring with *TP53*

mutations, which are associated with a poor prognosis. On the other hand, it is frequently seen as a sole abnormality in MDS, in which case it is associated with a good prognosis. Deletions in 5q target a similar region in ckAML and MDS, but in MDS two commonly retained regions are always spared from the deletions. These retained regions could be important for the progression from MDS to AML.

4.2 Enhancer hijacking

I developed pyjacker, a computational method to detect enhancer hijacking events using WGS and RNA-seq, and applied it to 39 ckAML samples. Pyjacker identified genes known to be activated by enhancer hijacking in AML, like *MECOM*, *BCL11B*, and *MNX1*, as well as novel genes that could be interesting to study further. *GSX2* is a homeobox gene that was overexpressed in one sample, with a breakpoint nearby. This gene was expressed in other cohorts as a result of a rare translocation t(4;12) which brings an *ETV6* enhancer close to *GSX2*. This translocation typically does not only lead to *GSX2* overexpression, but also to *PDGFRA* overexpression and to an *ETV6*::*CHIC2* fusion [141]. Here, the ckAML sample overexpressing *GSX2* had a different genomic rearrangement which did not lead to *CHIC2* overexpression nor to an *ETV6*::*CHIC2* fusion, which argues in favor of *GSX2* activation being the driver event. Another novel gene is *EPO* which was found overexpressed in an AEL sample. EPO drives proliferation of red blood cells by binding to its receptor (EPOR), thus activating the JAK/STAT pathway [144]. This event could cooperate with *EPOR* amplification, a phenomenon recently described in AEL [148]. Ruxolitinib, a JAK inhibitor, was shown in a PDX model to be effective against *EPOR*-amplified samples, so it could also be a good option in samples overexpressing *EPO*.

An important question in the field of enhancer hijacking is whether there are any compatibility rules between promoters and enhancers, or whether a gene can be activated by any enhancer. On the one hand, several super-enhancers like *ETV6*, *CDK6* or *MYC* have been reported to be hijacked by several genes, so any strong hematopoietic enhancer might be able to activate any gene. On the other hand, some enhancer-gene pairs are more frequent. For example, although *MECOM* can be activated by various enhancers [138], it is most frequently activated by the *GATA2* enhancer. This could have various explanations: the colocalization of *MECOM* and of the *GATA2* enhancer on chr3 could make it more likely for SVs to bring them together, or the haploinsufficiency of *GATA2* resulting from the loss of its enhancer could cooperate with *MECOM* activation. The cooperation hypothesis might be more likely, considering that *PRDM16* is

frequently activated by the *GATA2* enhancer, even though it lies on a different chromosome, and that *PRDM16* and *MECOM* overexpression may result in similar phenotypes.

More generally, it is still very difficult to predict whether a particular enhancer sequence can activate a gene, and whether the location of the enhancer with respect to the target gene is relevant. Enhancer marks like H3K27ac, H3K4me1, P300 or open chromatin can help identify putative enhancers. However, the analysis of the *ETV6* enhancer shows that several regions might be equally good candidates based on these marks, but only one is required for activation of the target gene. Consequently, enhancer marks are not sufficient, and ultimately one needs to insert or delete a sequence to prove that it is an enhancer capable of activating a gene. However, this has a very low throughput, and very few enhancers can be tested in this manner. In the wake of the great results achieved by large language models in natural language processing, a lot of research is now being spent on applying them to DNA sequences. For example, Enformer can predict gene expression based on the surrounding 200kb sequence [171]. However, it still poorly leverages effects from distal enhancers [172]. I expect that the performance of such models will improve in the future, and that they might be able to accurately predict whether a particular enhancer sequence can activate a gene. The ability to test enhancer function in-silico would save a lot of time, although some in-vitro validation would still be required.

In this thesis, I focused on ckAML, but also applied pyjacker to sarcoma and prostate cancer. This led to the identification of numerous genes putatively activated by enhancer hijacking in these other cancer types, with some already known like *IRS4* or *TERT* and some new candidates like *FGF8* in low-grade fibromyxoid sarcoma. However, this also revealed that impurity in solid tumors can lead to false positives, in particular for metastatic samples. In case pyjacker is to be applied further to solid tumors, the detection of contamination from other cell types could be an area of improvement.

Since pyjacker relies on samples without genomic rearrangements near a gene to estimate a reference distribution for the expression of this gene, it requires a cohort of at least ten samples profiled with WGS and RNA-seq. A higher number of samples would be beneficial, because it would lead to a high number of candidate genes. In addition, more samples lead to lower FDR and more confident predictions by pyjacker. In ckAML, applying pyjacker to a larger cohort would likely lead to the identification of some additional genes, but this would not dramatically change the results.

4.3 Allele-specific methylation

With pyjacker, I test for monoallelic expression, which is a strong indicator that a gene is expressed because of enhancer hijacking. I initially planned to do the same with allele-specific methylation, where ASM could be an indication for enhancer hijacking, if this information is available. I did observe that in most cases, enhancer hijacking resulted in an altered DNA methylation profile on the rearranged allele, but the changes were rather minimal. This is because most gene promoters are unmethylated, even when the gene is not expressed. With the resolution provided by nanopore sequencing, I observed that, typically, promoters of unexpressed genes still have some methylated CpG sites, and this methylation is completely removed when an allele is activated by enhancer hijacking. In addition, the unmethylated region can become wider. However, these changes are mild and not systematic, so I was not able to include allele-specific methylation in pyjacker. However, this allele-specific DNA methylation information might still be very useful for other purposes. For example, fluctuating CpG sites, which randomly switch between a methylated and an unmethylated state, have been used to infer clonal dynamics, like the number of stem cells in colon crypts or in blood [173]. Briefly, if the population is clonal, the methylation percentage at a CpG site should be 0% (both alleles unmethylated), 50% (one allele methylated, one allele unmethylated), or 100% (both alleles methylated), while intermediate values are expected if the population is more heterogeneous. Having allele-specific information would improve the model, because if the methylation level is 50% in the bulk sample, one could distinguish between whether all reads from one allele are methylated and all reads from the other allele are unmethylated, or whether there is heterogeneity within each allele. The long reads provided by nanopore sequencing can also be used to look at within-read methylation heterogeneity. Typically, there is a strong correlation between the methylation state of neighbouring CpGs, but some regions or some samples might show high within-read heterogeneity.

4.4 Conclusion and outlook

In this thesis, I showed that the numerous genomic rearrangements in ckAML can lead to enhancer hijacking, and I discovered new genes activated by this mechanism. Among 39 ckAML samples, I found 19 genes activated by structural rearrangements, 10 of which were fusions and 9 were likely enhancer hijacking events. Enhancer hijacking is therefore an important event in ckAML, although it is not as recurrent as the

common deletions like 5q and 7q. The next step would be to investigate the function of these genes, to understand how they drive the disease, in order to ultimately find targeted therapies. However, the large number of candidate genes and their low recurrence makes it difficult to select genes which would be worth investigating. *MECOM* is the gene most frequently activated by enhancer hijacking in AML and is therefore the most studied. *MECOM* impairs differentiation of hematopoietic cells by interfering with transcription factors, but so far no targeted therapy has been identified [174]. BET inhibitors can reduce the activity of super-enhancers and have been proposed as a general treatment against oncogenes activated by super-enhancers [47, 90], but they have effects genome-wide and are therefore not targeted to a specific gene.

Considering that CNAs, and especially deletions, are a lot more recurrent in ckAML than enhancer hijacking, they remain the most probable main driver in ckAML. Oncogenes activated by enhancer hijacking might provide additional driver events, but they are not required in ckAML. Understanding how CNAs drive the disease would be important, but this is very difficult considering the large number of genes whose dosage is affected by the CNAs. A possible strategy would be to study in detail the difference between MDS with isolated del(5q) and ckAML with del(5q), in order to understand why the former has good survival while the latter is associated with a dismal prognosis.

CHAPTER 5

METHODS

5.1 ckAML samples from the ASTRAL-1 cohort

39 ckAML samples were profiled by WGS and RNA-seq (sequencing done by Anna Riedel). They are diagnostic samples from the ASTRAL-1 phase 3 clinical trial, which evaluated the efficacy and safety of guadecitabine, a new hypomethylating agent [114, 115]. The ASTRAL-1 cohort included previously untreated AML patients who were not eligible to intensive chemotherapy, either because they were too old (>75 years) or had co-morbidities. 480 samples from this cohort had been profiled with EPIC array. Anna Riedel selected 42 of them which had at least three CNAs (based on the EPIC array data) and for which sufficient material was available for further profiling, and performed WGS and RNA-seq on those samples. RNA-seq failed for three samples, resulting in a total of 39 ckAML samples with WGS and RNA-seq.

5.2 WGS data processing

I developed a nextflow workflow to process WGS data: `https://github.com/CompEpigen/wf_WGS`. The main motivation for this new workflow was to detect somatic variants from WGS data, even when no matched normals are available.

CNAS were called CNAs with Control-FREEC [71]. I initially ran Control-FREEC on healthy samples from the Simons Genome Diversity Project [175], which allowed me to identify regions recurrently affected by CNAs across several healthy samples, which might be due to germline CNVs or repetitive regions. I then excluded these regions from the CNA inference when running Control-FREEC on tumor samples. I also only considered CNAs greater than 40kb.

I called SVs with manta [67]. I used a panel of normals provided by the Hartwig Medical Foundation (`https://github.com/hartwigmedical/hmftools`) to filter out putative germline SVs (due to retrotransposons, and potential mapping errors in repetitive regions, there are many SVs, including interchromosomal ones, in healthy samples). I also excluded SVs smaller than 40kb, as well as long-distance SVs which, when combined to another SV, appeared to result in a very small insertion.

As an alternative workflow to Control-FREEC and manta, I also called SVs and CNAs with the HMF pipeline (`https://github.com/hartwigmedical/hmftools`): SVs called with GRIDSS [68] and CNAs with amber and cobalt, and SVs and CNAs are then merged with purple. This workflow is slower, because GRIDSS is slower than manta, but has the advantage to combine CNA and SV calls, which can be used to get the copy number at each end of a breakpoint.

SNVs were called with mutect2 [176]. Since germline SNPs vastly outnumber somatic SNVs (about 1000 times more in AML), I only considered variants in a list of 52 genes known to be recurrently mutated in AML, or which were of particular interest: *DNMT3A*, *NPM1*, *FLT3*, *RUNX1*, *SF3B1*, *SRSF2*, *U2AF1*, *NF1*, *JAK2*, *TP53*, *IDH1*, *IDH2*, *NRAS*, *KRAS*, *KIT*, *TET1*, *TET2*, *CEBPA*, *WT1*, *PTPN11*, *ASXL1*, *ASXL2*, *EZH2*, *KMT2A*, *KMT2C*, *KMT2D*, *KMT2E*, *CREBBP*, *KDM6A*, *KAT6A*, *DNMT3B*, *NSD1*, *SUZ12*, *JARID2*, *ETV6*, *KDM3B*, *RB1*, *CEBPG*, *NCOR1*, *NCOR2*, *BCOR*, *GATA2*, *NOTCH1*, *NOTCH2*, *ZRSR2*, *PHF6*, *MED12*, *SMARCA2*, *SMARCA4*, *SMC1A*, *SMC3*, *STAG2*. In addition, I filtered out non-coding and synonymous variants, as well as variants found in the gnomAD database [177], except for variants in *DNMT3A*, *TET2* and *ASXL1* and *TP53* for which I allowed them to have a small frequency in the population, since these variants in these genes are common in clonal hematopoiesis, and gnomAD can contain somatic variants of these genes.

In order to detect allele-specific expression, I called germline SNPs in genes in the WGS data using HaplotypeCaller [178], selected heterozygous ones (VAF between 0.28 and 0.72), and used ASEReadCounter [178] to count the allelic read counts in the RNA-seq data at these heterozygous SNPs.

Chromothripsis events were detected using shatterseek [75], where I used as criteria:

- At least ten copy number switches in one chromosome.
- At least six SVs in one chromosome.
- Clustered breakpoints: p-value>=0.05 for an exponential distribution of the distance between breakpoints, which would be expected if the breakpoints were uniformly distributed.

- A random orientation of breakpoints: p-value<0.05 for a multinomial distribution of SV types with equal weights for each of the four SV types (deletion, duplication, tail-to-tail inversion, head-to-head inversion).

5.3 Detection of CNAs from SNP arrays and methylation arrays

For methylation arrays, CNAs were called using conumee [73]. By comparing to WGS (for samples profiled both with EPIC array and WGS), I noticed that many small CNAs called by conumee were false positives. Therefore, I post-processed the CNA calls by removing CNAs smaller than 200kb, and those for which the signal was low.

For SNP array data, I used PennCNV-Affy to generate ASCAT's inputs starting from CEL files (`https://github.com/VanLoo-lab/ascat/blob/master/ExampleData/ASCAT_fromCELfiles.R`), and I then called CNAs with ASCAT [72]. I also called CNAs with ASCAT using samples with a normal karyotype (without CNAs) and used those samples to create a panel of normals, which contains CNAs found in normal samples and are therefore artefacts. I used this panel of normals to filter out CNA calls from ASCAT.

For testing associations between CNAs and categorical variables (*TP53* mutation status, age group), I considered all chromosome arm level CNAs (deletion, gain or CN-LOH) which occurred in at least 30 samples of the cohort. I performed Fisher's exact test, and for *TP53* associations, I also tested for specific CNA associations, while keeping the total number of CNAs constant. This was done by randomly assigning CNAs to samples, while keeping the number of occurrences of each CNA and the number of CNAs in each sample constant. With these random assignments, I computed p-values from Fisher's exact test, and used those p-values as a null distribution, to test for enrichment of specific CNAs.

5.4 RNA-seq processing

Differential expression analysis was performed using pydeseq2 [179], a python implementation of DESeq2 [180].

Gene set enrichment analysis (GSEA) was performed using gseapy [181]. I used the hallmark gene sets [182], as well as gene sets derived from bone marrow cell types [183].

I identified fusion transcripts using STAR-Fusion [134], and I filtered the results using the WGS data, by only keeping fusions supported by a breakpoint in the WGS data.

5.5 Identification and scoring of enhancers

In order to identify and score enhancers, I used public ChIP-seq data of H3K27ac and P300 from three myeloid cell lines: K562 (data from the ENCODE project [184]), MOLM-1 (data from array expression accession E-MTAB-2224 [90]), and Kasumi-1 (data from GEO accession GSE167163 [185]). I used ROSE [46, 47] to score and rank enhancers in each of these six datasets, where I excluded transcription start sites. Since the ranking was quite variable depending on the dataset being used, I then averaged the ROSE enrichment scores from all six datasets, yielding a list of enhancers scored based on the enrichment of active enhancer marks in myeloid cell lines.

5.6 Pyjacker details

For each gene, pyjacker splits samples between "candidate samples" that have a breakpoint in the same TAD as the gene, and might therefore be activated by a structural rearrangement, and "reference samples" which do not. Each pair of (gene, candidate sample) is then scored to prioritize events. This score is made up of three parts: an overexpression score, a monoallelic expression score, and an enhancer score.

Overexpression score The expression values in TPM (transcript per million) are first log-transformed: $E = \log(0.5 + E_{\text{TPM}})$. Then, the mean μ_{ref} and standard deviation σ_{ref} of these log-transformed expression are computed for the reference samples, which do not have breakpoints in the vicinity of the gene. For each candidate sample, I count the number of standard deviations away from the mean its expression lies: $t = (E - \mu_{\text{ref}}) / (\sigma_{\text{ref}} + 0.3)$, where I add 0.3 to the standard deviation to account for situations where all reference samples have the same expression. I then compute the overexpression score as follows: if $t > 2$, then $S_{\text{overexpression}} = \log(t - 1)$, else $S_{\text{overexpression}} = -2\log(3 - t)$. This results in the overexpression score being positive if the expression is more than two standard deviations above the mean, and negative otherwise. The log transformation ensures that the score does not get too extreme.

Allele-specific expression score For each gene and each sample, I identify heterozygous SNPs within this gene in the WGS data, and if there is coverage in the RNA-seq data, I count the number of reads corresponding to each allele. For each SNP i, I compute the log likelihood ratio llr_i between monoallelic and biallelic expression. For biallelic expression, I assume that the allelic read counts follows a beta-binomial distribution centered at 50%, and for monoallelic expression, I use a mixture of two beta-binomial distributions centered at 2% and 98% (to account for possible low expression from the other allele). This log-likelihood ratio is positive if the expression is more likely to be monoallelic, which would support an expression due to enhancer hijacking, and negative otherwise. If a gene contains several SNPs, they are combined, in a way that gives a higher score if several SNPs are present, but still reaches a threshold when a very large number of SNPs are present: $S_{\text{ase}} = (\sum_{i=0}^{n} \text{llr}_i)/(n+2)$, where n is the number of SNPs in the gene. This score is positive if the expression is more likely to be monoallelic, negative if it is more likely to be biallelic, and null if no SNP is present (or if the allele-specific expression is unclear). Imprinted genes and genes on the X chromosome (apart from the pseudoautosomal region) have their allele-specific expression score set to 0 because their expression is expected to be monoallelic.

Enhancer score Using the breakpoint and enhancer information (see Section 5.5), I identify enhancers which are brought into the same TAD as the gene by the breakpoints. For this, I take the orientation of the breakpoints into account, but I did not attempt to assemble multiple SVs together. Thus, the correct enhancers might be missed in case of complex rearrangements, but should be correctly identified for simple rearrangements. The enhancer score is a weighted sum of all ROSE [46, 47] scores, putting more weight on the strongest enhancers: $S_{\text{enhancer}} = \sum_{j=1}^{m} R_j/10000/j$, where R_j is the ROSE score of the j-th strongest enhancer coming close to the gene and m is the total number of enhancers coming close to the gene.

Combined score The final score for each putative enhancer hijacking event is a weighted sum of the overexpression, monoallelic expression and enhancer scores:

$$S = \omega_{\text{overexpression}} S_{\text{overexpression}} + \omega_{\text{ase}} S_{\text{ase}} + \omega_{\text{enhancer}} S_{\text{enhancer}}$$

where the weights ω can be chosen by the user, and were set by default to $\omega_{\text{overexpression}} = 4$, $\omega_{\text{ase}} = 2$, and $\omega_{\text{enhancer}} = 1$.

Gene score The score defined above is for each pair of gene and candidate sample. I then aggregate the scores for each gene, by giving higher scores to genes activated in

multiple samples, but with a threshold in case a large number of candidate samples are present:

$$S_{\text{gene}} = 5/(n+4) \sum_{i=1}^{n} S_{\text{gene},i}$$

where $S_{\text{gene},i}$ is the gene score in the i-th candidate sample, where only positive scores are considered and n is the number of candidate samples for the gene.

False discovery rate The score defined above reflects how likely a gene is expressed because of a structural rearrangement, but is somewhat arbitrary. In order to get a more meaningful FDR, I compute empirical p-values by first generating a null distribution of scores. For each gene, I ignore the true candidate samples which have breakpoints nearby, and I randomly assign some reference samples to the candidate samples. Here, I choose a random number between one and three, and randomly select this number of reference samples to be considered as candidate samples. I then compute the scores for these "false" candidate samples, which results in a null distribution of scores in the absence of enhancer hijacking. In order to get a sufficient number of null scores, I iterate several times (50 by default) through all genes, each time selecting different samples to be considered as candidate samples. I can then count the proportion of null scores which are higher than a particular score, which results in an empirical p-value corresponding to this score. I then correct for multiple testing with the Benjamini-Hochberg method, resulting in an FDR.

5.7 Nanopore sequencing and data processing

Nanopore sequencing was performed by Jessica Heilmann. DNA was extracted using the QIAamp DNA micro kit. Size selection was performed with either the PacBio SRE or the PacBio SRE XS kit. Libraries were prepared using the kit SQK-LSK114 and were sequenced 96h or 120h on one PromethION flow cell, with at least one wash and reload of the flow cell.

I performed basecalling with dorado (`https://github.com/nanoporetech/dorado`), using the model `dna_r10.4.1_e8.2_400bps_sup@v4.2.0` and the remora model for base modifications `dna_r10.4.1_e8.2_400bps_sup@v4.2.0_5mCG_5hmCG@v2`. In order to phase reads, I used the epi2me human variation workflow v1.9.0 (`https://github.com/epi2me-labs/wf-human-variation`). This calls SNPs using clair3 [186] and phases them with WhatsHap [187], thus generating a haplotagged bam file, where

the HP tag indicates the haplotype of the read. I used modkit to generate bedmethyl files.

5.8 Figure generation with figeno

A large number of the figures of this thesis were generated with figeno [188], a visualization tool that I developed. It can display various types of sequencing data, including ChIP-seq or ATAC-seq in bigwig format, Hi-C, nanopore data with base modifications, and WGS with copy numbers and breakpoints. Figeno can display several regions simultaneously, which can for example be used to show interactions across breakpoints in Hi-C data, or SVs linking several regions in WGS data. Figeno also provides several layouts, for example the circular layout can be used to generate circos plots for WGS data. Figeno is implemented in python, and provides a graphical user interface made with javascript and the React framework. It is available on GitHub at `https://github.com/CompEpigen/figeno`.

ACRONYMS

AEL: Acute erythroid leukemia

AML: Acute myeloid leukemia

ASM: Allele-specific methylation

ATAC-seq: Assay for transposase-accessible chromatin with sequencing

BM: Bone marrow

BFB: Breakage-fusion-bridge

CGI: CpG island

ChIP-seq: Chromatin immunoprecipitation followed by sequencing

ckAML: Acute myeloid leukemia with a complex karyotype

CNA: Copy number alteration

DNA: Deoxyribonucleic acid

FDR: False discovery rate

HSC: Hematopoietic stem cell

HSPCs: Hematopoietic stem and progenitor cells

PB: Peripheral blood

PCR: Polymerase chain reaction

PDX: Patient-derived xenograft

PMD: Partially methylated domain

RNA: Ribonucleic acid

RNA-seq: RNA sequencing

SNP: Single nucleotide polymorphism

SNV: Single nucleotide variant

SV: Structural variant

TAD: Topologically associating domain

TSG: Tumor suppressor gene

WGS: Whole genome sequencing

WT: Wild-type

List of Figures

List of Tables

BIBLIOGRAPHY

[1] Eric Lander, Lauren Linton, Bruce Birren, Chad Nusbaum, Michael Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William Fitzhugh, Roel Funke, Diane Gaige, Katrina Harris, Andrew Heaford, John Howland, Lisa Kann, Jessica Lehoczky, Rosie LeVine, Paul McEwan, and Eda Koculi. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 02 2001.

[2] Adrian Bird. Dna methylation and the frequency of cpg in animal dna. *Nucleic acids research*, 8:1499–504, 05 1980.

[3] Adrian Bird, Mary Taggart, Marianne Frommer, Orlando Miller, and Donald Macleod. A fraction of the mouse genome that is derived from islands of nonmethylated, cpg-rich dna. *Cell*, 40:91–9, 02 1985.

[4] Serge Saxonov, Paul Berg, and Douglas Brutlag. A genome-wide analysis of cpg dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proceedings of the National Academy of Sciences of the United States of America*, 103:1412–7, 02 2006.

[5] James McGhee and Gordon Ginder. Specific dna methylation sites in the vicinity of the chicken beta-globin genes. *Nature*, 280:419–20, 09 1979.

[6] Miho Suzuki and Adrian Bird. Dna methylation landscapes: provocative insights from epigenomics. *Nature Reviews Genetics*, 9:465–76, 07 2008.

[7] Peter Jones. Functions of dna methylation: Islands, start sites, gene bodies and beyond. *Nature Reviews Genetics*, 13:484–92, 05 2012.

[8] R Stein, Aharon Razin, and Howard Cedar. In vitro methylation of the hamster adenine phosphoribosyltransferase gene inhibits its expression in mouse l cells. *Proceedings of the National Academy of Sciences*, 79(11):3418–3422, 1982.

[9] R Scott Hansen and Stanley M Gartler. 5-azacytidine-induced reactivation of the human x chromosome-linked pgk1 gene is associated with a large region of cytosine demethylation in the 5'cpg island. *Proceedings of the National Academy of Sciences*, 87(11):4174–4178, 1990.

[10] Maxim VC Greenberg and Deborah Bourc'his. The diverse roles of dna methylation in mammalian development and disease. *Nature reviews Molecular cell biology*, 20(10):590–607, 2019.

[11] Leslie F Lock, Nobuo Takagi, and Gail R Martin. Methylation of the hprt gene on the inactive x occurs after chromosome inactivation. *Cell*, 48(1):39–46, 1987.

[12] Adrian Bird. Dna methylation patterns and epigenetic memory. *Genes & development*, 16(1):6–21, 2002.

[13] Asaf Hellman and Andrew Chess. Gene body-specific methylation on the active x chromosome. *science*, 315(5815):1141–1143, 2007.

[14] Adam C Bell and Gary Felsenfeld. Methylation of a ctcf-dependent boundary controls imprinted expression of the igf2 gene. *Nature*, 405(6785):482–485, 2000.

[15] Amy T Hark, Christopher J Schoenherr, David J Katz, Robert S Ingram, John M Levorse, and Shirley M Tilghman. Ctcf mediates methylation-sensitive enhancer-blocking activity at the h19/igf2 locus. *Nature*, 405(6785):486–489, 2000.

[16] Elisa Kreibich, Rozemarijn Kleinendorst, Guido Barzaghi, Sarah Kaspar, and Arnaud R Krebs. Single-molecule footprinting identifies context-dependent regulation of enhancers by dna methylation. *Molecular Cell*, 83(5):787–802, 2023.

[17] F Creusot, G Acs, and JK Christman. Inhibition of dna methyltransferase and induction of friend erythroleukemia cell differentiation by 5-azacytidine and 5-aza-2'-deoxycytidine. *Journal of Biological Chemistry*, 257(4):2041–2048, 1982.

[18] Jared T Simpson, Rachael E Workman, PC Zuzarte, Matei David, LJ Dursi, and Winston Timp. Detecting dna cytosine methylation using nanopore sequencing. *Nature methods*, 14(4):407–410, 2017.

[19] Vahid Akbari, Jean-Michel Garant, Kieran O'Neill, Pawan Pandoh, Richard Moore, Marco A Marra, Martin Hirst, and Steven JM Jones. Megabase-scale methylation phasing using nanopore long reads and nanomethphase. *Genome biology*, 22:1–21, 2021.

[20] Cizhong Jiang and B Franklin Pugh. Nucleosome positioning and gene regulation: advances through genomics. *Nature Reviews Genetics*, 10(3):161–172, 2009.

[21] Sara Martire and Laura A Banaszynski. The roles of histone variants in fine-tuning chromatin organization and function. *Nature reviews Molecular cell biology*, 21(9):522–541, 2020.

[22] Ben E Black and Emily A Bassett. The histone variant cenp-a and centromere specification. *Current opinion in cell biology*, 20(1):91–100, 2008.

[23] Andrew J Bannister and Tony Kouzarides. Regulation of chromatin by histone modifications. *Cell research*, 21(3):381–395, 2011.

[24] David S Johnson, Ali Mortazavi, Richard M Myers, and Barbara Wold. Genome-wide mapping of in vivo protein-dna interactions. *Science*, 316(5830):1497–1502, 2007.

[25] Peter J Skene, Jorja G Henikoff, and Steven Henikoff. Targeted in situ genome-wide profiling with high efficiency for low cell numbers. *Nature protocols*, 13(5):1006–1019, 2018.

[26] Hatice S Kaya-Okur, Steven J Wu, Christine A Codomo, Erica S Pledger, Terri D Bryson, Jorja G Henikoff, Kami Ahmad, and Steven Henikoff. Cut&tag for efficient epigenomic profiling of small samples and single cells. *Nature communications*, 10(1):1930, 2019.

[27] Benjamin Carter, Wai Lim Ku, Jee Youn Kang, Gangqing Hu, Jonathan Perrie, Qingsong Tang, and Keji Zhao. Mapping histone modifications in low cell number and single cells using antibody-guided chromatin tagmentation (act-seq). *Nature communications*, 10(1):3747, 2019.

[28] Sandy L Klemm, Zohar Shipony, and William J Greenleaf. Chromatin accessibility and the regulatory epigenome. *Nature Reviews Genetics*, 20(4):207–220, 2019.

[29] M Ryan Corces, Jason D Buenrostro, Beijing Wu, Peyton G Greenside, Steven M Chan, Julie L Koenig, Michael P Snyder, Jonathan K Pritchard, Anshul Kundaje, William J Greenleaf, et al. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nature Genetics*, 48(10):1193–1203, 2016.

[30] Jason D Buenrostro, Beijing Wu, Howard Y Chang, and William J Greenleaf. Atac-seq: a method for assaying chromatin accessibility genome-wide. *Current protocols in molecular biology*, 109(1):21–29, 2015.

[31] Theresa K Kelly, Yaping Liu, Fides D Lay, Gangning Liang, Benjamin P Berman, and Peter A Jones. Genome-wide mapping of nucleosome positioning and dna methylation within individual dna molecules. *Genome research*, 22(12):2497–2506, 2012.

[32] Isac Lee, Roham Razaghi, Timothy Gilpatrick, Michael Molnar, Ariel Gershman, Norah Sadowski, Fritz J Sedlazeck, Kasper D Hansen, Jared T Simpson, and Winston Timp. Simultaneous profil-

ing of chromatin accessibility and methylation on human cell lines with nanopore sequencing. *Nature methods*, 17(12):1191–1199, 2020.

[33] Job Dekker, Karsten Rippe, Martijn Dekker, and Nancy Kleckner. Capturing chromosome conformation. *Science*, 295:1306–11, 03 2002.

[34] Marieke Simonis, Petra Klous, Erik Splinter, Yuri Moshkin, Rob Willemsen, Elzo Wit, Bas Steensel, and Wouter Laat. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4c). *Nature Genetics*, 38:1348–54, 12 2006.

[35] Erez Lieberman-Aiden, Nynke Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan Lajoie, Peter Sabo, Michael Dorschner, Richard Sandstrom, Bradley Bernstein, MA Bender, Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, Leonid Mirny, Eric Lander, and Job Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science (New York, N.Y.)*, 326:289–93, 10 2009.

[36] S.S.P. Rao, M.H. Huntley, Neva Durand, E.K. Stamenova, Ivan Bochkov, J.T. Robinson, A.L. Sanborn, I. Machol, Arina Omer, E.S. Lander, and E.L. Aiden. A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159:1665–1680, 12 2014.

[37] Jesse Dixon, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun Liu, and Bing Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485:376–80, 04 2012.

[38] Elphège P Nora, Anton Goloborodko, Anne-Laure Valton, Johan H Gibcus, Alec Uebersohn, Nezar Abdennur, Job Dekker, Leonid A Mirny, and Benoit G Bruneau. Targeted degradation of ctcf decouples local insulation of chromosome domains from genomic compartmentalization. *Cell*, 169(5):930–944, 2017.

[39] Wibke Schwarzer, Nezar Abdennur, Anton Goloborodko, Aleksandra Pekowska, Geoffrey Fudenberg, Yann Loe-Mie, Nuno A Fonseca, Wolfgang Huber, Christian H Haering, Leonid Mirny, et al. Two independent modes of chromatin organization revealed by cohesin removal. *Nature*, 551(7678):51–56, 2017.

[40] Maxwell Mumbach, Adam Rubin, Ryan Flynn, Chao Dai, Paul Khavari, William Greenleaf, and Howard Chang. Hichip: efficient and sensitive analysis of protein-directed genome architecture. *Nature Methods*, 13:919–922, 11 2016.

[41] Maxwell Mumbach, Ansuman Satpathy, Evan Boyle, Chao Dai, Benjamin Gowen, Seung Woo Cho, Michelle Nguyen, Adam Rubin, Jeffrey Granja, Katelynn Kazane, Yuning Wei, Trieu Nguyen, Peyton Greenside, M. Corces, Josh Tycko, Dimitre Simeonov, Nabeela Suliman, Rui Li, Jin Xu, and Howard Chang. Enhancer connectome in primary human cells reveals target genes of disease-associated dna elements. *Nature Genetics*, 49:1602–1612, 09 2017.

[42] Roger Mulet-Lazaro and Ruud Delwel. From genotype to phenotype: How enhancers control gene expression and cell identity in hematopoiesis. *HemaSphere*, 7(11):e969, 2023.

[43] Julian Banerji, Sandro Rusconi, and Walter Schaffner. Expression of a β-globin gene is enhanced by remote sv40 dna sequences. *Cell*, 27(2):299–308, 1981.

[44] Axel Visel, Matthew J Blow, Zirong Li, Tao Zhang, Jennifer A Akiyama, Amy Holt, Ingrid Plajzer-Frick, Malak Shoukry, Crystal Wright, Feng Chen, et al. Chip-seq accurately predicts tissue-specific activity of enhancers. *Nature*, 457(7231):854–858, 2009.

[45] Drew T Bergman, Thouis R Jones, Vincent Liu, Judhajeet Ray, Evelyn Jagoda, Layla Siraj, Helen Y Kang, Joseph Nasser, Michael Kane, Antonio Rios, et al. Compatibility rules of human enhancer and promoter sequences. *Nature*, 607(7917):176–184, 2022.

[46] Warren A Whyte, David A Orlando, Denes Hnisz, Brian J Abraham, Charles Y Lin, Michael H Kagey, Peter B Rahl, Tong Ihn Lee, and Richard A Young. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, 153(2):307–319, 2013.

[47] Jakob Lovén, Heather A Hoke, Charles Y Lin, Ashley Lau, David A Orlando, Christopher R Vakoc, James E Bradner, Tong Ihn Lee, and Richard A Young. Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell*, 153(2):320–334, 2013.

[48] Deborah Hay, Jim R Hughes, Christian Babbs, James OJ Davies, Bryony J Graham, Lars LP Hanssen, Mira T Kassouf, A Marieke Oudelaar, Jacqueline A Sharpe, Maria C Suciu, et al. Genetic dissection of the α-globin super-enhancer in vivo. *Nature Genetics*, 48(8):895–903, 2016.

[49] Ha Youn Shin, Michaela Willi, Kyung Hyun Yoo, Xianke Zeng, Chaochen Wang, Gil Metser, and Lothar Hennighausen. Hierarchy within the mammary stat5-driven wap super-enhancer. *Nature Genetics*, 48(8):904–911, 2016.

[50] Theodor Boveri. Concerning the origin of malignant tumours by theodor boveri. translated and annotated by henry harris. *Journal of cell science*, 121(Supplement_1):1–84, 2008.

[51] Peyton Rous. A transmissible avian neoplasm.(sarcoma of the common fowl.). *The Journal of experimental medicine*, 12(5):696–705, 1910.

[52] G Steven Martin. The hunting of the src. *Nature reviews Molecular cell biology*, 2(6):467–475, 2001.

[53] Dominique Stehelin, Harold E Varmus, J Michael Bishop, and Peter K Vogt. Dna related to the transforming gene (s) of avian sarcoma viruses is present in normal avian dna. *Nature*, 260(5547):170–173, 1976.

[54] Marc S Collett and RL Erikson. Protein kinase activity associated with the avian sarcoma virus src gene product. *Proceedings of the National Academy of Sciences*, 75(4):2021–2024, 1978.

[55] Bert Vogelstein and Kenneth W Kinzler. The multistep nature of cancer. *Trends in genetics*, 9(4):138–141, 1993.

[56] Alfred G Knudson Jr. Mutation and cancer: statistical study of retinoblastoma. *Proceedings of the National Academy of Sciences*, 68(4):820–823, 1971.

[57] Sibylle Mittnacht. The retinoblastoma protein—from bench to bedside. *European journal of cell biology*, 84(2-3):97–107, 2005.

[58] Arnold J Levine and Moshe Oren. The first 30 years of p53: growing ever more complex. *Nature reviews cancer*, 9(10):749–758, 2009.

[59] David P Lane and Lionel V Crawford. T antigen is bound to a host protein in sy40-transformed cells. *Nature*, 278(5701):261–263, 1979.

[60] Daniel IH Linzer and Arnold J Levine. Characterization of a 54k dalton cellular sv40 tumor antigen present in sv40-transformed cells and uninfected embryonal carcinoma cells. *Cell*, 17(1):43–52, 1979.

[61] Daniel Eliyahu, Avraham Raz, Peter Gruss, David Givol, and Moshe Oren. Participation of p53 cellular tumour antigen in transformation of normal embryonic cells. *Nature*, 312(5995):646–649, 1984.

[62] Daniel Eliyahu, Dan Michalovitz, Siona Eliyahu, Orit Pinhasi-Kimhi, and Moshe Oren. Wild-type p53 can inhibit oncogene-mediated focus formation. *Proceedings of the National Academy of Sciences*, 86(22):8763–8767, 1989.

[63] Suzanne J Baker, Eric R Fearon, Janice M Nigro, Stanley R Hamilton, Ann C Preisinger, J Milburn Jessup, Peter VanTuinen, David H Ledbetter, David F Barker, Yusuke Nakamura, et al. Chromo-

some 17 deletions and p53 gene mutations in colorectal carcinomas. *Science*, 244(4901):217–221, 1989.

[64] Ludmil B Alexandrov, Serena Nik-Zainal, David C Wedge, Samuel AJR Aparicio, Sam Behjati, Andrew V Biankin, Graham R Bignell, Niccolo Bolli, Ake Borg, Anne-Lise Børresen-Dale, et al. Signatures of mutational processes in human cancer. *Nature*, 500(7463):415–421, 2013.

[65] Susanne Horn, Adina Figl, P Sivaramakrishna Rachakonda, Christine Fischer, Antje Sucker, Andreas Gast, Stephanie Kadel, Iris Moll, Eduardo Nagore, Kari Hemminki, et al. Tert promoter mutations in familial and sporadic melanoma. *Science*, 339(6122):959–961, 2013.

[66] Charlotte Smith, Ashish Goyal, Dieter Weichenhan, Eric Allemand, Anand Mayakonda, Umut Toprak, Anna Riedel, Estelle Balducci, Manisha Manojkumar, Anastasija Pejkovska, et al. Tal1 activation in t-cell acute lymphoblastic leukemia: a novel oncogenic 3'neo-enhancer. *Haematologica*, 108(5):1259, 2023.

[67] Xiaoyu Chen, Ole Schulz-Trieglaff, Richard Shaw, Bret Barnes, Felix Schlesinger, Morten Källberg, Anthony J Cox, Semyon Kruglyak, and Christopher T Saunders. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*, 32(8):1220–1222, 2016.

[68] Daniel L Cameron, Jonathan Baber, Charles Shale, Jose Espejo Valle-Inclan, Nicolle Besselink, Arne van Hoeck, Roel Janssen, Edwin Cuppen, Peter Priestley, and Anthony T Papenfuss. Gridss2: comprehensive characterisation of somatic structural variation using single breakend variants and structural variant phasing. *Genome biology*, 22:1–25, 2021.

[69] Tobias Rausch, Thomas Zichner, Andreas Schlattl, Adrian M Stütz, Vladimir Benes, and Jan O Korbel. Delly: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, 28(18):i333–i339, 2012.

[70] Moritz Smolka, Luis F Paulin, Christopher M Grochowski, Dominic W Horner, Medhat Mahmoud, Sairam Behera, Ester Kalef-Ezra, Mira Gandhi, Karl Hong, Davut Pehlivan, et al. Detection of mosaic and population-level structural variants with sniffles2. *Nature biotechnology*, pages 1–10, 2024.

[71] Valentina Boeva, Tatiana Popova, Kevin Bleakley, Pierre Chiche, Julie Cappo, Gudrun Schleiermacher, Isabelle Janoueix-Lerosey, Olivier Delattre, and Emmanuel Barillot. Control-freec: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics*, 28(3):423–425, 2012.

[72] Peter Van Loo, Silje H Nordgard, Ole Christian Lingjærde, Hege G Russnes, Inga H Rye, Wei Sun, Victor J Weigman, Peter Marynen, Anders Zetterberg, Bjørn Naume, et al. Allele-specific copy number analysis of tumors. *Proceedings of the National Academy of Sciences*, 107(39):16910–16915, 2010.

[73] Volker Hovestadt, Marc Remke, Marcel Kool, Torsten Pietsch, Paul A Northcott, Roger Fischer, Florence MG Cavalli, Vijay Ramaswamy, Marc Zapatka, Guido Reifenberger, 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*, 125:913–916, 2013.

[74] Philip J Stephens, Chris D Greenman, Beiyuan Fu, Fengtang Yang, Graham R Bignell, Laura J Mudie, Erin D Pleasance, King Wai Lau, David Beare, Lucy A Stebbings, et al. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell*, 144(1):27–40, 2011.

[75] Isidro Cortés-Ciriano, Jake June-Koo Lee, Ruibin Xi, Dhawal Jain, Youngsook L Jung, Lixing Yang, Dmitry Gordenin, Leszek J Klimczak, Cheng-Zhong Zhang, David S Pellman, et al. Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. *Nature Genetics*, 52(3):331–341, 2020.

[76] Natalia Voronina, John KL Wong, Daniel Hübschmann, Mario Hlevnjak, Sebastian Uhrig, Christoph E Heilig, Peter Horak, Simon Kreutzfeldt, Andreas Mock, Albrecht Stenzinger, et al. The landscape of chromothripsis across adult cancer types. *Nature communications*, 11(1):2320, 2020.

[77] Jan O Korbel and Peter J Campbell. Criteria for inference of chromothripsis in cancer genomes. *Cell*, 152(6):1226–1236, 2013.

[78] Yilong Li, Claire Schwab, Sarra L Ryan, Elli Papaemmanuil, Hazel M Robinson, Patricia Jacobs, Anthony V Moorman, Sara Dyer, Julian Borrow, Mike Griffiths, et al. Constitutional and somatic rearrangement of chromosome 21 in acute lymphoblastic leukaemia. *Nature*, 508(7494):98–102, 2014.

[79] Andrew P Feinberg and Bert Vogelstein. Hypomethylation distinguishes genes of some human cancers from their normal counterparts. *Nature*, 301(5895):89–92, 1983.

[80] Kasper Daniel Hansen, Winston Timp, Héctor Corrada Bravo, Sarven Sabunciyan, Benjamin Langmead, Oliver G McDonald, Bo Wen, Hao Wu, Yun Liu, Dinh Diep, et al. Increased methylation variation in epigenetic domains across cancer types. *Nature Genetics*, 43(8):768–775, 2011.

[81] Benjamin E Decato, Jianghan Qu, Xiaojing Ji, Elvin Wagenblast, Simon RV Knott, Gregory J Hannon, and Andrew D Smith. Characterization of universal features of partially methylated domains across tissues and species. *Epigenetics & chromatin*, 13:1–14, 2020.

[82] James G Herman, Farida Latif, Yongkai Weng, Michael I Lerman, Berton Zbar, Sue Liu, Dvorit Samid, DS Duan, James R Gnarra, and W Marston Linehan. Silencing of the vhl tumor-suppressor gene by dna methylation in renal carcinoma. *Proceedings of the National Academy of Sciences*, 91(21):9700–9704, 1994.

[83] Stephen B Baylin and James G Herman. Dna hypermethylation in tumorigenesis: epigenetics joins genetics. *Trends in genetics*, 16(4):168–174, 2000.

[84] Francisco Antequera, Joan Boyes, and Adrian Bird. High levels of de novo methylation and altered chromatin structure at cpg islands in cell lines. *Cell*, 62(3):503–514, 1990.

[85] Dominic J Smiraglia, Laura J Rush, Michael C Frühwald, Zunyan Dai, William A Held, Joseph F Costello, James C Lang, Charis Eng, Bin Li, Fred A Wright, et al. Excessive cpg island hypermethylation in cancer cell lines versus primary human malignancies. *Human molecular genetics*, 10(13):1413–1419, 2001.

[86] Adrian C Hayday, Stephen D Gillies, Haruo Saito, Charles Wood, Klas Wiman, William S Hayward, and Susumu Tonegawa. Activation of a translocated human c-myc gene by an enhancer in the immunoglobulin heavy-chain locus. *Nature*, 307(5949):334–340, 1984.

[87] Paul A Northcott, Catherine Lee, Thomas Zichner, Adrian M Stütz, Serap Erkek, Daisuke Kawauchi, David JH Shih, Volker Hovestadt, Marc Zapatka, Dominik Sturm, et al. Enhancer hijacking activates gfi1 family oncogenes in medulloblastoma. *Nature*, 511(7510):428–434, 2014.

[88] Martin Peifer, Falk Hertwig, Frederik Roels, Daniel Dreidax, Moritz Gartlgruber, Roopika Menon, Andrea Krämer, Justin L Roncaioli, Frederik Sand, Johannes M Heuckmann, et al. Telomerase activation by genomic rearrangements in high-risk neuroblastoma. *Nature*, 526(7575):700–704, 2015.

[89] Joachim Weischenfeldt, Taronish Dubash, Alexandros P Drainas, Balca R Mardin, Yuanyuan Chen, Adrian M Stütz, Sebastian M Waszak, Graziella Bosco, Ann Rita Halvorsen, Benjamin Raeder, et al. Pan-cancer analysis of somatic copy-number alterations implicates irs4 and igf2 in enhancer hijacking. *Nature Genetics*, 49(1):65–74, 2017.

[90] Stefan Gröschel, Mathijs A Sanders, Remco Hoogenboezem, Elzo de Wit, Britta AM Bouwman, Claudia Erpelinck, Vincent HJ van der Velden, Marije Havermans, Roberto Avellino, Kirsten van Lom, et al. A single oncogenic enhancer rearrangement causes concomitant evi1 and gata2 deregulation in leukemia. *Cell*, 157(2):369–381, 2014.

[91] Lindsey E Montefiori, Sonja Bendig, Zhaohui Gu, Xiaolong Chen, Petri Pölönen, Xiaotu Ma, Alex Murison, Andy Zeng, Laura Garcia-Prat, Kirsten Dickerson, et al. Enhancer hijacking drives oncogenic bcl11b expression in lineage-ambiguous stem cell leukemia. *Cancer discovery*, 11(11):2846–2867, 2021.

[92] Dieter Weichenhan, Anna Riedel, Etienne Sollier, Umut H Toprak, Joschka Hey, Kersten Breuer, Justyna A Wierzbinska, Aurore Touzart, Pavlo Lutsik, Marion Bähr, et al. Altered enhancer-promoter interaction leads to mnx1 expression in pediatric acute myeloid leukemia with t(7;12)(q36;p13). *Blood advances*, 2024.

[93] Anqi Yu, Ali E Yesilkanal, Ashish Thakur, Fan Wang, Yang Yang, William Phillips, Xiaoyang Wu, Alexander Muir, Xin He, Francois Spitz, et al. Hyena detects oncogenes activated by distal enhancers in cancer. *Biorxiv*, pages 2023–01, 2023.

[94] Yu Liu, Chunliang Li, Shuhong Shen, Xiaolong Chen, Karol Szlachta, Michael N Edmonson, Ying Shao, Xiaotu Ma, Judith Hyle, Shaela Wright, et al. Discovery of regulatory noncoding variants in individual cancer genomes by using cis-x. *Nature Genetics*, 52(8):811–818, 2020.

[95] Xiaotao Wang, Jie Xu, Baozhen Zhang, Ye Hou, Fan Song, Huijue Lyu, and Feng Yue. Genome-wide detection of enhancer-hijacking events from chromatin interaction data in rearranged genomes. *Nature methods*, 18(6):661–668, 2021.

[96] Emma Shtivelman, Batia Lifshitz, Robert P Gale, and Eli Canaani. Fused transcript of abl and bcr genes in chronic myelogenous leukaemia. *Nature*, 315(6020):550–554, 1985.

[97] Koji Sasaki, Sara S Strom, Susan O'Brien, Elias Jabbour, Farhad Ravandi, Marina Konopleva, Gautam Borthakur, Naveen Pemmaraju, Naval Daver, Preetesh Jain, et al. Relative survival in patients with chronic-phase chronic myeloid leukaemia in the tyrosine-kinase inhibitor era: analysis of patient data from six prospective clinical trials. *The Lancet Haematology*, 2(5):e186–e193, 2015.

[98] Siddhartha Jaiswal, Pierre Fontanillas, Jason Flannick, Alisa Manning, Peter V Grauman, Brenton G Mar, R Coleman Lindsley, Craig H Mermel, Noel Burtt, Alejandro Chavez, et al. Age-related clonal hematopoiesis associated with adverse outcomes. *New England Journal of Medicine*, 371(26):2488–2498, 2014.

[99] John M Bennett, Daniel Catovsky, Marie-Theregse Daniel, George Flandrin, David AG Galton, Harvey R Gralnick, and Claude Sultan. Proposals for the classification of the acute leukaemias french-american-british (fab) co-operative group. *British journal of haematology*, 33(4):451–458, 1976.

[100] Joseph D Khoury, Eric Solary, Oussama Abla, Yassmine Akkari, Rita Alaggio, Jane F Apperley, Rafael Bejar, Emilio Berti, Lambert Busque, John KC Chan, et al. The 5th edition of the world health organization classification of haematolymphoid tumours: myeloid and histiocytic/dendritic neoplasms. *Leukemia*, 36(7):1703–1719, 2022.

[101] Cancer Genome Atlas Research Network. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *New England Journal of Medicine*, 368(22):2059–2074, 2013.

[102] Elli Papaemmanuil, Moritz Gerstung, Lars Bullinger, Verena I Gaidzik, Peter Paschka, Nicola D Roberts, Nicola E Potter, Michael Heuser, Felicitas Thol, Niccolo Bolli, et al. Genomic classification and prognosis in acute myeloid leukemia. *New England Journal of Medicine*, 374(23):2209–2221, 2016.

[103] Frank G Rücker, Richard F Schlenk, Lars Bullinger, Sabine Kayser, Veronica Teleanu, Helena Kett, Marianne Habdank, Carla-Maria Kugler, Karlheinz Holzmann, Verena I Gaidzik, et al. Tp53 alterations in acute myeloid leukemia with complex karyotype correlate with specific copy number alterations, monosomal karyotype, and dismal outcome. *Blood, The Journal of the American Society of Hematology*, 119(9):2114–2121, 2012.

[104] Frank G Rücker, Anna Dolnik, Tamara J Blätte, Veronica Teleanu, Aurélie Ernst, Felicitas Thol, Michael Heuser, Arnold Ganser, Hartmut Döhner, Konstanze Döhner, et al. Chromothripsis is linked to tp53 alteration, cell cycle impairment, and dismal outcome in acute myeloid leukemia with complex karyotype. *Haematologica*, 103(1):e17, 2018.

[105] Claudia Schoch, Wolfgang Kern, Alexander Kohlmann, Wolfgang Hiddemann, Susanne Schnittger, and Torsten Haferlach. Acute myeloid leukemia with a complex aberrant karyotype is a distinct biological entity characterized by genomic imbalances and a specific gene expression profile. *Genes, Chromosomes and Cancer*, 43(3):227–238, 2005.

[106] Peter L Greenberg, Heinz Tuechler, Julie Schanz, Guillermo Sanz, Guillermo Garcia-Manero, Francesc Solé, John M Bennett, David Bowen, Pierre Fenaux, Francois Dreyfus, et al. Revised international prognostic scoring system for myelodysplastic syndromes. *Blood, The Journal of the American Society of Hematology*, 120(12):2454–2465, 2012.

[107] Naoko Hosono, Hideki Makishima, Reda Mahfouz, Bartlomiej Przychodzen, Kenichi Yoshida, Andres Jerez, Thomas LaFramboise, Chantana Polprasert, Michael J Clemente, Yuichi Shiraishi, et al. Recurrent genetic defects on chromosome 5q in myeloid neoplasms. *Oncotarget*, 8(4):6483, 2017.

[108] Vera Adema, Laura Palomo, Wencke Walter, Mar Mallo, Stephan Hutter, Thomas La Framboise, Leonor Arenillas, Manja Meggendorfer, Tomas Radivoyevitch, Blanca Xicoy, et al. Pathophysiologic and clinical implications of molecular profiles resultant from deletion 5q. *EBioMedicine*, 80, 2022.

[109] Alexander E Smith, Austin G Kulasekararaj, Jie Jiang, Syed Mian, Azim Mohamedali, Joop Gaken, Robin Ireland, Barbara Czepulkowski, Steven Best, and Ghulam J Mufti. Csnk1a1 mutations and isolated del (5q) abnormality in myelodysplastic syndrome: a retrospective mutational analysis. *The Lancet Haematology*, 2(5):e212–e221, 2015.

[110] Benjamin L Ebert, Jennifer Pretz, Jocelyn Bosco, Cindy Y Chang, Pablo Tamayo, Naomi Galili, Azra Raza, David E Root, Eyal Attar, Steven R Ellis, et al. Identification of rps14 as a 5q-syndrome gene by rna interference screen. *Nature*, 451(7176):335–339, 2008.

[111] John M Joslin, Anthony A Fernald, Thelma R Tennant, Elizabeth M Davis, Scott C Kogan, John Anastasi, John D Crispino, and Michelle M Le Beau. Haploinsufficiency of egr1, a candidate gene in the del (5q), leads to the development of myeloid disorders. *Blood, The Journal of the American Society of Hematology*, 110(2):719–726, 2007.

[112] Toshiya Inaba, Hiroaki Honda, and Hirotaka Matsui. The enigma of monosomy 7. *Blood, The Journal of the American Society of Hematology*, 131(26):2891–2898, 2018.

[113] Adriane Halik, Marlon Tilgner, Patricia Silva, Natalia Estrada, Robert Altwasser, Ekaterina Jahn, Michael Heuser, Hsin-An Hou, Marta Pratcorona, Robert K Hills, et al. Genomic characterization of aml with aberrations of chromosome 7: a multinational cohort of 519 patients. *Journal of hematology & oncology*, 17(1):70, 2024.

[114] Pierre Fenaux, Marco Gobbi, Patricia L Kropf, Jean-Pierre J Issa, Gail J Roboz, Jiri Mayer, Jürgen Krauter, Tadeusz Robak, Hagop Kantarjian, Jan Novak, et al. Guadecitabine vs treatment choice in newly diagnosed acute myeloid leukemia: a global phase 3 randomized study. *Blood Advances*, 7(17):5027–5037, 2023.

[115] Ekaterina Jahn, Maral Saadati, Pierre Fenaux, Marco Gobbi, Gail J Roboz, Lars Bullinger, Pavlo Lutsik, Anna Riedel, Christoph Plass, Nikolaus Jahn, et al. Clinical impact of the genomic landscape and leukemogenic trajectories in non-intensively treated elderly acute myeloid leukemia patients. *Leukemia*, 37(11):2187–2196, 2023.

[116] Abdelkader Behdenna, Maximilien Colange, Julien Haziza, Aryo Gema, Guillaume Appé, Chloé-Agathe Azencott, and Akpéli Nordor. pycombat, a python tool for batch effects correction in high-throughput molecular data using empirical bayes methods. *BMC bioinformatics*, 24(1):459, 2023.

[117] Dvir Aran, Zicheng Hu, and Atul J Butte. xcell: digitally portraying the tissue cellular heterogeneity landscape. *Genome biology*, 18:1–14, 2017.

[118] Jan Philipp Bewersdorf, Vanessa Hasle, Rory Michael Shallis, Ethan G Thompson, Daniel Lopes de Menezes, Shelonitda Rose, Isaac W Boss, Stephanie Halene, Torsten Haferlach, Brian Fox, et al. Molecular, epigenetic, and immune landscape of tp53-mutated (tp53-m) acute myeloid leukemia (aml) and higher risk myelodysplastic syndromes (hr-mds). *Blood*, 140(Supplement 1):6247–6249, 2022.

[119] César Muñoz-Fontela, Salvador Macip, Luis Martínez-Sobrido, Lauren Brown, Joseph Ashour, Adolfo García-Sastre, Sam W Lee, and Stuart A Aaronson. Transcriptional role of p53 in interferon-mediated antiviral immunity. *The Journal of experimental medicine*, 205(8):1929–1938, 2008.

[120] Stanley WK Ng, Amanda Mitchell, James A Kennedy, Weihsu C Chen, Jessica McLeod, Narmin Ibrahimova, Andrea Arruda, Andreea Popescu, Vikas Gupta, Aaron D Schimmer, et al. A 17-gene stemness score for rapid determination of risk in acute leukaemia. *Nature*, 540(7633):433–437, 2016.

[121] Umut Aypar, Jinjuan Yao, Dory M Londono, Rose Khoobyar, Angela Scalise, Maria E Arcila, Mikhail Roshal, Wenbin Xiao, and Yanming Zhang. Rare and novel runx1 fusions in myeloid neoplasms: A single-institute experience. *Genes, Chromosomes and Cancer*, 60(2):100–107, 2021.

[122] Daniel Bottomly, Nicola Long, Anna Reister Schultz, Stephen E Kurtz, Cristina E Tognon, Kara Johnson, Melissa Abel, Anupriya Agarwal, Sammantha Avaylon, Erik Benton, et al. Integrative analysis of drug response and clinical outcome in acute myeloid leukemia. *Cancer cell*, 40(8):850–864, 2022.

[123] Brian Parkin, Peter Ouillette, Mehmet Yildiz, Kamlai Saiya-Cork, Kerby Shedden, and Sami N Malek. Integrated genomic profiling, therapy response, and survival in adult acute myelogenous leukemia. *Clinical Cancer Research*, 21(9):2045–2056, 2015.

[124] Ophir Shalem, Neville E Sanjana, Ella Hartenian, Xi Shi, David A Scott, Tarjei S Mikkelsen, Dirk Heckl, Benjamin L Ebert, David E Root, John G Doench, et al. Genome-scale crispr-cas9 knockout screening in human cells. *Science*, 343(6166):84–87, 2014.

[125] Andres Jerez, Lukasz P Gondek, Anna M Jankowska, Hideki Makishima, Bartlomiej Przychodzen, Ramon V Tiu, Christine L O'Keefe, Azim M Mohamedali, Denise Batista, Mikkael A Sekeres, et al.

Topography, clinical, and genomic correlates of 5q myeloid malignancies revisited. *Journal of clinical oncology*, 30(12):1343, 2012.

[126] Mar Mallo, Mónica Del Rey, Mariam Ibanez, Mª José Calasanz, Leonor Arenillas, Mª José Larráyoz, Carmen Pedro, Andres Jerez, Jaroslaw Maciejewski, Dolors Costa, et al. Response to lenalidomide in myelodysplastic syndromes with del (5q): influence of cytogenetics and mutations. *British journal of haematology*, 162(1):74–86, 2013.

[127] Anna Hecht, Julia A Meyer, Johann-Christoph Jann, Katja Sockel, Aristoteles Giagounidis, Katharina S Götze, Anne Letsch, Detlef Haase, Richard F Schlenk, Torsten Haferlach, et al. Genome-wide dna methylation analysis pre-and post-lenalidomide treatment in patients with myelodysplastic syndrome with isolated deletion (5q). *Annals of Hematology*, 100:1463–1471, 2021.

[128] Mitchell J Machiela, Weiyin Zhou, Neil Caporaso, Michael Dean, Susan M Gapstur, Lynn Goldin, Nathaniel Rothman, Victoria L Stevens, Meredith Yeager, and Stephen J Chanock. Mosaic chromosome 20q deletions are more frequent in the aging population. *Blood advances*, 1(6):380–385, 2017.

[129] Ulrike Bacher, Torsten Haferlach, Susanne Schnittger, Melanie Zenger, Manja Meggendorfer, Sabine Jeromin, Andreas Roller, Vera Grossmann, Maria-Theresa Krauth, Tamara Alpermann, et al. Investigation of 305 patients with myelodysplastic syndromes and 20q deletion for associated cytogenetic and molecular genetic lesions and their prognostic impact. *British journal of haematology*, 164(6):822–833, 2014.

[130] Hui Yang, Stefan Kurtenbach, Ying Guo, Ines Lohse, Michael A Durante, Jianping Li, Zhaomin Li, Hassan Al-Ali, Lingxiao Li, Zizhen Chen, et al. Gain of function of asxl1 truncating protein in the pathogenesis of myeloid malignancies. *Blood, The Journal of the American Society of Hematology*, 131(3):328–341, 2018.

[131] Steven J Kuerbitz, Joshua Pahys, Alison Wilson, Nicole Compitello, and Todd A Gray. Hypermethylation of the imprinted nnat locus occurs frequently in pediatric acute leukemia. *Carcinogenesis*, 23(4):559–564, 2002.

[132] Juan Li, Anthony J Bench, George S Vassiliou, Nasios Fourouclas, Anne C Ferguson-Smith, and Anthony R Green. Imprinting of the human l3mbtl gene, a polycomb family member located in a region of chromosome 20 deleted in human myeloid malignancies. *Proceedings of the National Academy of Sciences*, 101(19):7341–7346, 2004.

[133] Athar Aziz, E Joanna Baxter, Carol Edwards, Clara Yujing Cheong, Mitsuteru Ito, Anthony Bench, Rebecca Kelley, Yvonne Silber, Philip A Beer, Keefe Chng, et al. Cooperativity of imprinted genes inactivated by acquired chromosome 20q deletions. *The Journal of clinical investigation*, 123(5):2169–2182, 2013.

[134] Brian J Haas, Alexander Dobin, Bo Li, Nicolas Stransky, Nathalie Pochet, and Aviv Regev. Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. *Genome biology*, 20:1–16, 2019.

[135] Nicole A McNeer, John Philip, Heather Geiger, Rhonda E Ries, Vincent-Philippe Lavallee, Michael Walsh, Minita Shah, Kanika Arora, Anne-Katrin Emde, Nicolas Robine, et al. Genetic mechanisms of primary chemotherapy resistance in pediatric acute myeloid leukemia. *Leukemia*, 33(8):1934–1943, 2019.

[136] Naomi Mochizuki, Seiichi Shimizu, Toshiro Nagasawa, Hideo Tanaka, Masafumi Taniwaki, Jun Yokota, and Kazuhiro Morishita. A novel gene, mel1, mapped to 1p36. 3 is highly homologous to

the mds1/evi1 gene and is transcriptionally activated in t (1; 3)(p36; q21)-positive leukemia cells. *Blood, The Journal of the American Society of Hematology*, 96(9):3209–3214, 2000.

[137] Prasuna Muppa, Daniel L Van Dyke, Michelle K Bianco, Beth A Pitel, Stephanie A Smoley, George Vasmatzis, James B Smadbeck, William R Sukov, Patricia T Greipp, Rhett P Ketterling, et al. Characterization of at (1; 2)(p36; p21) involving the prdm16 gene region by mate-pair sequencing (mpseq) in a patient with newly diagnosed acute myeloid leukemia with myelodysplasia-related changes. *Journal of Hematopathology*, 12:85–90, 2019.

[138] Sophie Ottema, Roger Mulet-Lazaro, H Berna Beverloo, Claudia Erpelinck, Stanley van Herk, Robert van der Helm, Marije Havermans, Tim Grob, Peter JM Valk, Eric Bindels, et al. Atypical 3q26/mecom rearrangements genocopy inv (3)/t (3; 3) in acute myeloid leukemia. *Blood, The Journal of the American Society of Hematology*, 136(2):224–234, 2020.

[139] Irum Khan, Mohammed A Amin, Elizabeth A Eklund, and Andrei L Gartel. Regulation of hox gene expression in aml. *Blood cancer journal*, 14(1):42, 2024.

[140] Jan Cools, Nicole Mentens, Maria D Odero, Pieter Peeters, Iwona Wlodarska, Michel Delforge, Anne Hagemeijer, and Peter Marynen. Evidence for position effects as a variant etv6-mediated leukemogenic mechanism in myeloid leukemias with at (4; 12)(q11-q12; p13) or t (5; 12)(q31; p13). *Blood, The Journal of the American Society of Hematology*, 99(5):1776–1784, 2002.

[141] Angelika Müller-Jochim, Manja Meggendorfer, Wencke Walter, Torsten Haferlach, Wolfgang Kern, and Claudia Haferlach. Aml with t (4; 12)(q12; p13): A detailed genomic and transcriptomic analysis reveals genomic breakpoint heterogeneity, absence of pdgfra fusion transcripts and presence of pdgfra overexpression in a subset of cases. *Blood*, 142:6014, 2023.

[142] Jeff C Howard, Lloyd Berger, Maria Rosa Bani, Robert G Hawley, and Yaacov Ben-David. Activation of the erythropoietin gene in the majority of f-mulv-induced erythroleukemias results in growth factor independence and enhanced tumorigenicity. *Oncogene*, 12(7):1405–1416, 1996.

[143] Stany Chrétien, Véronique Duprez, Leila Maouche, Sylvie Gisselbrecht, Patrick Mayeux, and Catherine Lacombe. Abnormal erythropoietin (epo) gene expression in the murine erythroleukemia iw32 cells results from a rearrangement between the g-protein $\beta 2$ subunit gene and the epo gene. *Oncogene*, 15(16):1995–1999, 1997.

[144] Ursula Klingmüller, Svetlana Bergelson, Jonathan G Hsiao, and Harvey F Lodish. Multiple tyrosine residues in the cytosolic domain of the erythropoietin receptor promote activation of stat5. *Proceedings of the National Academy of Sciences*, 93(16):8324–8328, 1996.

[145] Y Miyazaki, K Kuriyama, M Higuchi, H Tsushima, H Sohda, N Imai, M Saito, T Kondo, and M Tomonaga. Establishment and characterization of a new erythropoietin-dependent acute myeloid leukemia cell line, as-e2. *Leukemia*, 11(11):1941–1949, 1997.

[146] Ilaria Iacobucci, Ji Wen, Manja Meggendorfer, John K Choi, Lei Shi, Stanley B Pounds, Catherine L Carmichael, Katherine E Masih, Sarah M Morris, R Coleman Lindsley, et al. Genomic subtyping and therapeutic targeting of acute erythroleukemia. *Nature Genetics*, 51(4):694–704, 2019.

[147] Alexandre Fagnan, Frederik Otzen Bagger, Maria-Riera Piqué-Borràs, Cathy Ignacimouttou, Alexis Caulier, Cécile K Lopez, Elie Robert, Benjamin Uzan, Véronique Gelsi-Boyer, Zakia Aid, et al. Human erythroleukemia genetics and transcriptomes identify master transcription factors as functional disease drivers. *Blood, The Journal of the American Society of Hematology*, 136(6):698–714, 2020.

[148] June Takeda, Kenichi Yoshida, Masahiro M Nakagawa, Yasuhito Nannya, Akinori Yoda, Ryunosuke Saiki, Yotaro Ochi, Lanying Zhao, Rurika Okuda, Xingxing Qi, et al. Amplified epor/jak2 genes define a unique subtype of acute erythroid leukemia. *Blood cancer discovery*, 3(5):410–427, 2022.

[149] Anthony Wl Lo, Laure Sabatier, Bijan Fouladi, Géraldine Pottier, Michelle Ricoul, and John P Murnane. Dna amplification by breakage/fusion/bridge cycles initiated by spontaneous telomere loss in a human cancer cell line. *Neoplasia*, 4(6):531–538, 2002.

[150] Arash Ronaghy, Shimin Hu, Zhenya Tang, Wei Wang, Guilin Tang, Sanam Loghavi, Shaoying Li, Beenu Thakral, L Jeffrey Medeiros, and Tariq Muzzafar. Myeloid neoplasms associated with t (3; 12)(q26. 2; p13) are clinically aggressive, show myelodysplasia, and frequently harbor chromosome 7 abnormalities. *Modern Pathology*, 34(2):300–313, 2021.

[151] Tina Nilsson, Ahmed Waraky, Anders Östlund, Susann Li, Anna Staffas, Julia Asp, Linda Fogelstrand, Jonas Abrahamsson, and Lars Palmqvist. An induced pluripotent stem cell t (7; 12)(q36; p13) acute myeloid leukemia model shows high expression of mnx1 and a block in differentiation of the erythroid and megakaryocytic lineages. *International Journal of Cancer*, 151(5):770–782, 2022.

[152] Manja Meggendorfer, Marietta Truger, Wencke Walter, Constance Baer, Stephan Hutter, Niroshan Nadarajah, Wolfgang Kern, Torsten Haferlach, and Claudia Haferlach. Detecting the unusual without compromising diagnostic accuracy-a prospective wgs/wts pilot study in acute leukemias provided additional information for diagnosis, prognosis and treatment. *Blood*, 140(Supplement 1):4959–4960, 2022.

[153] Elena Poli, Vanessa Barbon, Silvia Lucchetta, Manuela Cattelan, Luisa Santoro, Angelica Zin, Giuseppe Maria Milano, Ilaria Zanetti, Gianni Bisogno, and Paolo Bonvini. Immunoreactivity against fibroblast growth factor 8 in alveolar rhabdomyosarcoma patients and its involvement in tumor aggressiveness. *Oncoimmunology*, 11(1):2096349, 2022.

[154] Saskia Sydow, Paul Piccinelli, Shamik Mitra, Panagiotis Tsagkozis, Asle Hesla, Camila BR De Mattos, Jan Köster, Linda Magnusson, Jenny Nilsson, Adam Ameur, et al. Mdm2 amplification in rod-shaped chromosomes provides clues to early stages of circularized gene amplification in liposarcoma. *Communications biology*, 7(1):606, 2024.

[155] Raf Sciot. Mdm2 amplified sarcomas: a literature review. *Diagnostics*, 11(3):496, 2021.

[156] Stacey J Baker, Poulikos I Poulikakos, Hanna Y Irie, Samir Parekh, and E Premkumar Reddy. Cdk4: a master regulator of the cell cycle and its role in cancer. *Genes & cancer*, 13:21, 2022.

[157] Zeng Yang, Bo Wei, Anbang Qiao, Popo Yang, Wenhui Chen, Dezhi Zhen, and Xiaojian Qiu. A novel ezh2/nxph4/cdkn2a axis is involved in regulating the proliferation and migration of non-small cell lung cancer cells. *Bioscience, Biotechnology, and Biochemistry*, 86(3):340–350, 2022.

[158] Stephen L Chan, Yan Cui, Andrew Van Hasselt, Hongyu Li, Gopesh Srivastava, Hongchuan Jin, Ka M Ng, Yajun Wang, Kwan Y Lee, George SW Tsao, et al. The tumor suppressor wnt inhibitory factor 1 is frequently methylated in nasopharyngeal and esophageal carcinomas. *Laboratory investigation*, 87(7):644–650, 2007.

[159] Eloise Mastrangelo and Mario Milani. Role and inhibition of gli1 protein in cancer. *Lung Cancer: Targets and Therapy*, pages 35–43, 2018.

[160] Delila Gasi Tandefelt, Joost Boormans, Karin Hermans, and Jan Trapman. Ets fusion genes in prostate cancer. *Endocrine-related cancer*, 21(3):R143–R152, 2014.

[161] Navatha Shree Polavaram, Samikshan Dutta, Ridwan Islam, Arup K Bag, Sohini Roy, David Poitz, Jeffrey Karnes, Lorenz C Hofbauer, Manish Kohli, Brian A Costello, et al. Tumor-and osteoclast-derived nrp2 in prostate cancer bone metastases. *Bone research*, 9(1):24, 2021.

[162] Ridwan Islam, Juhi Mishra, Navatha Shree Polavaram, Sreyashi Bhattacharya, Zhengdong Hong, Sanika Bodas, Sunandini Sharma, Alyssa Bouska, Tyler Gilbreath, Ahmed M Said, et al. Neuropilin-2 axis in regulating secretory phenotype of neuroendocrine-like prostate cancer cells and its implication in therapy resistance. *Cell reports*, 40(3), 2022.

[163] Michael Scherer, Almut Nebel, Andre Franke, Jörn Walter, Thomas Lengauer, Christoph Bock, Fabian Müller, and Markus List. Quantitative comparison of within-sample heterogeneity scores for dna methylation data. *Nucleic acids research*, 48(8):e46–e46, 2020.

[164] Matthias Farlik, Florian Halbritter, Fabian Müller, Fizzah A Choudry, Peter Ebert, Johanna Klughammer, Samantha Farrow, Antonella Santoro, Valerio Ciaurro, Anthony Mathur, et al. Dna methylation dynamics of human hematopoietic stem cell differentiation. *Cell stem cell*, 19(6):808–822, 2016.

[165] Natalia Carreras-Gallo, Varun B Dwaraka, Dereje D Jima, David A Skaar, Tavis L Mendez, Antonio Planchart, Wanding Zhou, Randy L Jirtle, Ryan Smith, and Cathrine Hoyo. Creation and validation of the first infinium dna methylation array for the human imprintome. *Epigenetics Communications*, 4(1):5, 2024.

[166] Frank Court, Chiharu Tayama, Valeria Romanelli, Alex Martin-Trujillo, Isabel Iglesias-Platas, Kohji Okamura, Naoko Sugahara, Carlos Simón, Harry Moore, Julie V Harness, Hans Keirstead, et al. Genome-wide parent-of-origin dna methylation analysis reveals the intricacies of human imprinting and suggests a germline methylation-independent mechanism of establishment. *Genome research*, 24(4):554–569, 2014.

[167] Angela Sparago, Flavia Cerrato, Maria Vernucci, Giovanni Battista Ferrero, Margherita Cirillo Silengo, and Andrea Riccio. Microdeletions in the human h19 dmr result in loss of igf2 imprinting and beckwith-wiedemann syndrome. *Nature genetics*, 36(9):958–960, 2004.

[168] Dieter Weichenhan, Anna Riedel, Charlotte Meinen, Alisa Basic, Reka Toth, Marion Bähr, Pavlo Lutsik, Joschka Hey, Etienne Sollier, Umut H Toprak, et al. Translocation t (6; 7) in aml-m4 cell line gdm-1 results in mnx1 activation through enhancer-hijacking. *Leukemia*, 37(5):1147–1150, 2023.

[169] Simone S Riedel, Congcong Lu, Hongbo M Xie, Kevin Nestler, Marit W Vermunt, Alexandra Lenard, Laura Bennett, Nancy A Speck, Ichiro Hanamura, Julie A Lessard, et al. Intrinsically disordered meningioma-1 stabilizes the baf complex to cause aml. *Molecular cell*, 81(11):2332–2348, 2021.

[170] Frank G Rücker, Lars Bullinger, Carsten Schwaenen, Daniel B Lipka, Swen Wessendorf, Stefan Fröhling, Martin Bentz, Simone Miller, Claudia Scholl, Richard F Schlenk, et al. Disclosure of candidate genes in acute myeloid leukemia with complex karyotypes using microarray-based molecular characterization. *Journal of clinical oncology*, 24(24):3887–3894, 2006.

[171] Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R Ledsam, Agnieszka Grabska-Barwinska, Kyle R Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature methods*, 18(10):1196–1203, 2021.

[172] Alexander Karollus, Thomas Mauermeier, and Julien Gagneur. Current sequence-based models capture gene expression determinants in promoters but mostly ignore distal enhancers. *Genome biology*, 24(1):56, 2023.

[173] Calum Gabbutt, Ryan O Schenck, Daniel J Weisenberger, Christopher Kimberley, Alison Berner, Jacob Househam, Eszter Lakatos, Mark Robertson-Tessi, Isabel Martin, Roshani Patel, et al. Fluctuating methylation clocks for cell lineage tracing at high temporal resolution in human tissues. *Nature biotechnology*, 40(5):720–730, 2022.

[174] Christine Birdwell, Warren Fiskus, Tapan M Kadia, Courtney D DiNardo, Christopher P Mill, and Kapil N Bhalla. Evi1 dysregulation: impact on biology and therapy of myeloid malignancies. *Blood cancer journal*, 11(3):64, 2021.

[175] Swapan Mallick, Heng Li, Mark Lipson, Iain Mathieson, Melissa Gymrek, Fernando Racimo, Mengyao Zhao, Niru Chennagiri, Susanne Nordenfelt, Arti Tandon, et al. The simons genome diversity project: 300 genomes from 142 diverse populations. *Nature*, 538(7624):201–206, 2016.

[176] David Benjamin, Takuto Sato, Kristian Cibulskis, Gad Getz, Chip Stewart, and Lee Lichtenstein. Calling somatic snvs and indels with mutect2. *BioRxiv*, page 861054, 2019.

[177] Konrad J Karczewski, Laurent C Francioli, Grace Tiao, Beryl B Cummings, Jessica Alföldi, Qingbo Wang, Ryan L Collins, Kristen M Laricchia, Andrea Ganna, Daniel P Birnbaum, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581(7809):434–443, 2020.

[178] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, et al. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome research*, 20(9):1297–1303, 2010.

[179] Boris Muzellec, Maria Teleńczuk, Vincent Cabeli, and Mathieu Andreux. Pydeseq2: a python package for bulk rna-seq differential expression analysis. *Bioinformatics*, 39(9):btad547, 2023.

[180] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15:1–21, 2014.

[181] Zhuoqing Fang, Xinyuan Liu, and Gary Peltz. Gseapy: a comprehensive package for performing gene set enrichment analysis in python. *Bioinformatics*, 39(1):btac757, 2023.

[182] Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P Mesirov, and Pablo Tamayo. The molecular signatures database hallmark gene set collection. *Cell systems*, 1(6):417–425, 2015.

[183] Stuart B Hay, Kyle Ferchen, Kashish Chetal, H Leighton Grimes, and Nathan Salomonis. The human cell atlas bone marrow single-cell interactive web portal. *Experimental hematology*, 68:51–61, 2018.

[184] ENCODE Project Consortium et al. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57, 2012.

[185] Viral Shah, George Giotopoulos, Hikari Osaki, Markus Meyerhöfer, Eshwar Meduri, Benedict Schubert, Haiyang Yun, Sarah J Horton, Shuchi Agrawal-Singh, Patricia S Haehnel, et al. Acute resistance to bet inhibitors remodels compensatory transcriptional programs via p300 coactivation. *bioRxiv*, pages 2022–09, 2022.

[186] Zhenxian Zheng, Shumin Li, Junhao Su, Amy Wing-Sze Leung, Tak-Wah Lam, and Ruibang Luo. Symphonizing pileup and full-alignment for deep learning-based long-read variant calling. *Nature Computational Science*, 2(12):797–803, 2022.

[187] Marcel Martin, Murray Patterson, Shilpa Garg, Sarah O Fischer, Nadia Pisanti, Gunnar W Klau, Alexander Schöenhuth, and Tobias Marschall. Whatshap: fast and accurate read-based phasing. *BioRxiv*, page 085050, 2016.

[188] Etienne Sollier, Jessica Heilmann, Clarissa Gerhauser, Michael Scherer, Christoph Plass, and Pavlo Lutsik. Figeno: multi-region genomic figures with long-read support. *Bioinformatics*, 40(6), 2024.

MANUSCRIPTS, CONFERENCE TALKS, AND POSTER PRESENTATIONS

First-author manuscripts

- **Etienne Sollier**, Jack Kuipers, Koichi Takahashi, Niko Beerenwinkel and Katharina Jahn. **COMPASS: Joint copy number and mutation phylogeny reconstruction from single-cell amplicon sequencing data**. Nature Communications, August 2023.

- **Etienne Sollier**, Jessica Heilmann, Clarissa Gerhäuser, Michael Scherer, Christoph Plass and Pavlo Lutsik. **Figeno: multi-region genomic figures with long read support**. Bioinformatics, June 2024.

- Dieter Weichenhan*, Anna Riedel*, **Etienne Sollier***, Umut H. Toprak*, Joschka Hey, Kersten Breuer, Justyna A. Wierzbinska, Aurore Touzart, Pavlo Lutsik, Marion Bähr, Anders Östlund, Tina Nilsson, Susanna Jacobsson, Marcel Edler, Ahmed Waraky, Yvonne Lisa Behrens, Gudrun Göhring, Brigitte Schlegelberger, Clemens Steinek, Hartmann Harz, Heinrich Leonhardt, Anna Dolnik, Dirk Reinhard, Lars Bullinger, Lars Palmqvist, Daniel B. Lipka, Christoph Plass. **Altered enhancer-promoter interaction leads to *MNX1* expression in pediatric acute myeloid leukemia with t(7;12)(q36;p13)**. Blood Advances, August 2024. *co-first authors

- **Etienne Sollier***, Anna Riedel*, Umut H. Toprak*, Justyna A. Wierbinska, Dieter Weichenhan,Jan Philipp Schmid, Mariam Hakobyan, Aurore Touzart, Ekaterina Jahn, Binje Vick, Fiona Brown-Burke, Katherine Kelly, Simge Kelekci, Anastasija Pejkovska, Ashish Goyal, Marion Bähr, Kersten Breuer, Mei-Ju May Chen, Maria Llamazares-Prada, Mark Hartmann, Maximilian Schönung, Nádia Correia, Andreas Trumpp, Yomn Abdullah, Ursula Klingmüller, Sadaf S. Mughal, Benedikt Brors, Frank Westermann, Matthias Schlesner, Sebastian Vosberg, Tobias Herold, Philipp A. Greif, Dietmar Pfeifer, Michael Lübbert, Thomas Fischer, Florian Heidel, Claudia Gebhard, Wencke Walter, Torsten Haferlach, Ann-Kathrin Eisfeld, Krzysztof Mrózek, Deedra Nicolet, Lars Bullinger, Leonie Smeenk, Claudia Erpelinck, Roger Mulet-Lazaro, Ruud Delwel, Aurélie Ernst, Michael Scherer, Pavlo

Lutsik, Irmela Jeremias, Konstanze Döhner, Hartmut Döhner, Daniel B. Lipka, Christoph Plass. **Pyjacker identifies enhancer hijacking in rearranged AML genomes including del(7q) AML**. In preparation. *co-first authors.

Other manuscripts

- Charlotte Smith, Ashish Goyal, Dieter Weichenhan, Eric Allemand, Anand Mayakonda, Umut Toprak, Anna Riedel, Estelle Balducci, Manisha Manojkumar, Anastasija Pejkovska, Oliver Mücke, **Etienne Sollier**, Ali Bakr, Kersten Breuer, Pavlo Lutsik, Olivier Hermine, Salvatore Spicuglia, Vahid Asnafi, Christoph Plass and Aurore Touzart. **TAL1 activation in T-cell acute lymphoblastic leukemia: a novel oncogenic 3' neo-enhancer**. Haematologica, January 2023.
- Dieter Weichenhan, Anna Riedel, Charlotte Meinen, Alisa Basic, Reka Toth, Marion Bähr, Pavlo Lutsik, Joschka Hey, **Etienne Sollier**, Umut H. Toprak, Simge Kelekçi, Yu-Yu Lin, Mariam Hakobyan, Aurore Touzart, Ashish Goyal, Justyna A. Wierzbinska, Matthias Schlesner, Frank Westermann, Daniel B. Lipka and Christoph Plass. **Translocation t(6;7) in AML-M4 cell line GDM-1 results in *MNX1* activation through enhancer-hijacking**. Leukemia, May 2023.
- Ali Bakr, Giuditta Della Corte, Olivera Veselinov, Simge Kelekçi, Mei-Ju May Chen, Yu-Yu Lin, Gianluca Sigismondo, Marika Iacovone, Alice Cross, Rabail Syed, Yunhee Jeong, **Etienne Sollier**, Chun-Shan Liu, Pavlo Lutsik, Jeroen Krijgsveld, Dieter Weichenhan, Christoph Plass, Odilia Popanda and Peter Schmezer. **ARID1A regulates DNA repair through chromatin organization and its deficiency triggers DNA damage-mediated anti-tumor immune response**. Nucleic Acids Research, April 2024.

Conference talks

- **COMPASS: Joint copy number and mutation phylogeny from amplicon single-cell sequencing data**. *ISMB/ECCB2023*. Lyon, 27.07.2023
- **Nanopore sequencing enables detection of enhancer hijacking with allele-specific methylation**. *Berlin Nanopore Day*. Berlin, 16.11.2023
- **Nanopore sequencing enables detection of enhancer hijacking with allele-specific methylation**. *Nanopore WYMM Tour Münich*. Münich, 27.02.2024

Poster presentations

- **Enhancer hijacking in ckAML**. *EHA2023*. Frankfurt, 09.06.2023
- (upcoming) **Figeno: (epi)genomics visualizations, and application to enhancer hijacking**. *ECCB2024*. Turku, 18-19.09.2024

ACKNOWLEDGEMENTS

I would like to thank the following people who helped me throughout my PhD:

- Christoph Plass, for offering me the opportunity to work in his lab and pursue an exciting project, and for being frequently available for discussions.
- Pavlo Lutsik, for co-supervising me and providing guidance.
- Benedikt Brors, for being my first PhD examiner and a member of my thesis advisory committee, as well as for the insightful discussions during the TAC meetings.
- Lennart Hilbert, for being part of my thesis advisory committee and his advice.
- Aurélie Ernst, for the discussions on chromothripsis.
- All the members of the Cancer Epigenomics division, for being such a nice division to work in. In particular, I would like to thank
  - the bioinformatics office (Katherine, Yunhee, Nan, Anand, Joschka), for the nice working environments.
  - the AML team (Anna, Simge, Katherine, Elena, Fiona) for providing the best meeting of the week.
  - all the people who helped me by performing experiments in the lab, especially Jessica and Fiona.
- The TSG78 Heidelberg, for being such a nice and welcoming running club.
- My parents, for always being very supportive throughout my studies.
- Caroline, for filling my heart with joy and motivating me through the difficult times.