Dissertation

Epigenetic traces in cell-free DNA

submitted by **DI Peter Ulz, BSc**

for the Academic Degree of
Doctor of Medical Science
(Dr. scient.med)

at the
Medical University of Graz

Diagnostic & Research Institute of Human Genetics

Under the Supervision of
Univ.-Prof. Dr. Michael R. Speicher

2019

Declaration

I hereby declare that this thesis is my own original work and that I have fully acknowledged by name all of those individuals and organizations that have contributed to the research for this thesis. Due acknowledgement has been made in the text to all other material used. Throughout this thesis and in all related publications I followed the "Standards of Good Scientific Practice and Ombuds Committee at the Medical University of Graz".

Date	Signature

Disclosures

Parts of this thesis have been published as a preprint (bioRxiv, DOI: 10.1101/456681) as well as in Ulz P, Perakis S., et al. Inference of transcription factor binding from cell-free DNA enables tumor subtype prediction and early detection. Nature Communications 10, 4666 (2019); DOI:10.1038/s41467-019-12714-4. The following co-authors have directly contributed to the work that resulted in the present thesis and agreed to the use of their work in the presented thesis:

- Samantha Perakis[1]
- Qing Zhou[1]
- Tina Moser[1]
- Jelena Belić[1]
- Isaac Lazzeri[1]
- Albert Wölfler[2]
- Armin Zebisch[2]
- Armin Gerger[3]
- Edgar Petru[4]
- Gunda Pristauz[4]
- Brandon White[3]
- Charles E. S. Roberts[4]
- John St. John[4]
- Michael G. Schimek[6]
- Jochen B. Geigl[1]
- Thomas Bauernhofer[3]
- Heinz Sill[2]
- Christoph Bock[7]
- Ellen Heitzer[1]
- Michael R Speicher[1]

[1]Institute of Human Genetics, Diagnostic and Research Center for Molecular;[2]Department of Hematology, Medical University of Graz;[3]Department of Internal Medicine, Division of Oncology, Medical University of Graz;[4]Department of Obstetrics and Gynecology, Medical University of Graz;[5]Freenome, South San Francisco, USA;[6]Institute of Medical Informatics, Statistics and Documentation, Medical University of Graz;[7]CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences

Acknowledgements

As a doctoral student I received funding from the Medical University of Graz through the Doctoral School of Molecular Medicine and Inflammation (MOLMED). This work was done in projects that received funding from CANCER-ID, a project funded by the Innovative Medicines Joint Undertaking (IMI JU; #115749-1), by the BioTechMed-Graz flagship project EPIAge, and by the Christian Doppler Research Fund for Liquid Biopsies for Early Detection of Cancer. I also would like to express my gratitude to Freenome, which helped out with additional data and suggestions on data analysis in early-stage colorectal cancers.

Personally, I would like to thank Univ.-Prof. Dr. med. univ. Michael Speicher for the continuous support over many years of collaboration even before it became apparent that Bioinformatics is going to be detrimental in our research and for being a fundamental driving force of my professional career. Moreover, I would like to to thank Assoz. Prof.in Priv.-Doz.in Mag.a Dr.in rer.nat. Ellen Heitzer for being a mentor and a constant source of ideas and thoughtful criticism over the years. Sharing crazy ideas on how to extract information from our experiments has undoubtedly enriched both of us. I will be eternally grateful for the help and support from the late Univ.-Ass. Priv.-Doz. Mag. Dr.rer.nat. Dr.scient.med. Thomas Schwarzbraun who unfortunately can't witness my graduation. He believed in me at a time when I had nothing to show for and helped getting me employed at the Institute. Your wit, crazy ideas and your scientific approach to every single problem at hand are missed a lot. Furthermore, I'd like to thank all colleagues at the Institute of Human Genetics for a lively and stimulating working environment that I really enjoyed being part of.

Lastly, I'd like to thank my family, my beloved wife Barbara and my children Benjamin and Magdalena for their love and support. This would have been impossible without you.

Contents

Abbreviations

BAM binary alignment/mapping format

bp base-pair

BWA Burrows-Wheeler Aligner

cfDNA cell-free DNA

cfRNA cell-free RNA

ChIP-seq Chromatin immune precipitation sequencing

CNA copy number alteration

COAD colon adenocarcinoma

CRC colonrectal carcinoma

ctDNA circulating tumor DNA

EBI European Bioinformatics Institute

EGA European Genome-phenome Archive

FPKM fragments per kilobase exon per million reads

GEO gene expression omnibus

MNase-seq micrococcal nuclease sequencing

NDR nucleosome depleted region

NGS next-generation sequencing

PCR polymerase chain reaction

PSA prostate specific antigen

SAM sequence alignment/mapping format

SCNA somatic copy number alteration

SNP single (or simple) nucleotide polymorphism

SNV single (or simple) nucleotide variation

SRA sequence read archive

TSS transcription start site

TF transcription factor

TFBS transcription factor binding site

UCSC University of California in Santa Cruz

List of Figures

List of Tables

Zusammenfassung in Deutsch

Veränderte epigenetische Regulierungen in Genomen von Krebszellen sind ein wichtiger Mechanismus in der Entstehung von Tumorerkrankungen. Allerdings gibt es derzeit kaum Möglichkeiten diese Änderungen mit minimal-invasiven Methoden in soliden Geweben zu untersuchen. In der vorliegenden Arbeit wurde evaluiert, ob sich veränderte epigenetische Regulationsmechanismen über die Analyse zellfreier DNA, insbesondere der Position von Nukleosomen, ableiten lassen. Bestimmte Abschnitte der zellfreien DNA sind durch die Bindung an Histonen vor einem enzymatischen Abbau geschützt und spiegeln deshalb die Lokalisation von Nukleosomen wider. Dadurch sollten Veränderungen der Nukleosomsynchronisierung durch verschiedene biologische Prozesse über nicht-zufällige Verteilung von Fragment durch Sequenzierung der gesamten zellfreien DNA möglich sein. In Proben von metastasierten Patienten konnten sowohl Patienten-, sowie Tumor-spezifische Änderungen detektiert werden. Vor allem Änderungenen der Nukleosomenstruktur durch verändertes Bindungsverhalten von Transkriptionsfaktoren zeigen ein großes Potential veränderte biologische Prozesse des Primärtumors nicht-invasiv zu analysieren. Speziell die Transkriptionsfaktoren AR, HOXB13 und NKX3-1 erlauben die Klassifizierung von Untergruppen bei Prostatakarzinomen. Außerdem konnten neue Transkriptionsfaktoren, die in der Entstehung von Dickdarmkrebs eine Rolle spielen, identifiziert werden. In einer weiteren Kohorte aus Proben von Kolonkarzinompatienten in frühen Stadien konnte ein Klassifikationsalgorithmus basierend auf Transkriptionsfaktormustern erstellt werden, der Karzinompatienten auch schon in frühen Stadien erkennt. Zusätzlich wurde Nukleosomensynchronisierung auch für andere epigenetische Mechanismen untersucht, wobei die einzelnen Signale schwächer sind, was die genaue Erfassung von veränderten Aktivitäten erschwert. Die vorliegenden Analysen um epigenetische Mechanismen in Tumorpatienten nichtinvasiv zu messen, könnte einen großen Teil des Tumorgenoms auch außerhalb codierender

Bereiche für klinische Analysen zugänglich machen.

Abstract in English

Alterations in epigenetic regulatory regions in the genome are important drivers of tumorigenesis, but non-invasive assays for assessing them are lacking. In this work, the feasibility of inferring altered epigenetic regulatory mechanism in solid tumors from their nucleosome footprint in circulating cell-free DNA was evaluated. Since cell-free DNA fragments are associated to histone proteins, changes in nucleosome synchronization should lead to non-random coverage in whole-genome sequencing data. In a set of late-stage cancer samples, patient-specific as well as tumor-specific patterns can be observed using this method. Especially altered nucleosome positioning due to transcription factor binding showed great potential in uncovering biological mechanism of the primary tumor by non-invasive means. Specifically, inferred binding patterns for the transcription factors AR, HOXB13, and NKX3-1 allowed classification of patients by tumor type, including subtypes of prostate cancer, which has clinical implications for the management of patients. Moreover, novel transcription factors that play a role in the carcinogenesis of colorectal cancer were identified. Lastly, transcription factor accessibilities were used to construct a cancer classification algorithm that can reliably detect colon carcinoma samples even at early stages. Moreover, several other epigenetic regulatory mechanism leave a trace in cfDNA fragmentation patterns, however, the respective signal strengths are lower and nucleosome synchronization is more difficult to assess. This approach for mapping tumor-specific alteration of epigenetic regulatory mechanisms in vivo based on blood samples makes a key part of the noncoding genome amenable for clinical analysis.

1. *Introduction*

The analysis of cell-free DNA in the plasma has transformed blood-based diagnostics in several areas of medicine. Especially applications in minimally-invasive prenatal screening and minimally-invasive oncological analytics, cell-free DNA has garnered wide-spread attention. Moreover, cell-free DNA is a promising candidate to replace early-stage cancer screening [1].

1.1. **cfDNA**

Cell-free DNA (cfDNA) was first discovered in 1948 by Mandel and Métais [2]. Leon et al. first described a higher concentration of cfDNA in cancer patients in 1977 [3]. The first evidence of somatic cancer mutations in cfDNA was found in 1999, where mutations in the ras genes were found in hematological diseases [4]. Since then cfDNA has been used as a marker in cancer diagnostics in identifying pathological biology in cancers, treatment recurrence, therapy monitoring and lately also in early-stage cancer detection [5].

1.1.1. *cfDNA in oncology*

Cell-free DNA that is shed directly from the tumor into the peripheral blood is often termed circulating tumor DNA (ctDNA). CtDNA can be used in several stages to aid diagnostics and therapy management throughout a cancer disease. Fragments of the tumor genomes carry many of the characteristics that distinguish them from fragments that derive from healthy (mainly blood-cell) turnover. These include:

- Somatic mutations
- Somatic copy-number alterations (SCNAs)
- Altered methylation

- Altered fragment sizes
- Viral DNA in viral-induced cancer types

All of these properties can be exploited for ctDNA analyses by varying molecular and bioinformatic techniques.

Methods for detecting somatic mutations

Especially the use of somatic mutation based assays is being used widely for cancer diagnostics. While somatic mutations are usually very specific to cancer, large scale cancer genomics projects have found that a very similar mutation spectrum to somatic mutations in cancer can also occur due to clonal expansion of hematological cells of indeterminate potential (CHIP) [6], which eventually can lead to leukemia. This phenomenon and also other findings of clonal expansion of mutated cells in solid healthy tissues are a source of false-positive findings and puts into question any detection method without prior knowledge of somatic mutations in the primary tumor [7].

Somatic mutations in ctDNA can be detected in one of several ways. While Sanger sequencing lacks the sensitivity to detect mutations in lower frequencies, next-generation sequencing (NGS) can be used to track mutations down to a low allele frequency threshold. However, background noise from PCR and the sequencing reaction itself is prohibitive of detecting mutations at allele frequencies lower than approximately 1% [8]. To this end, many commercially available methods for somatic mutation detection rely on molecular barcoding [9]. This technique marks cfDNA molecules before any PCR is performed. Thus, distortion of allelic representation due to preferential amplification can be avoided, but it also allows for polishing sequencing errors by only allowing variants that are found in all fragments that are derived from the same template molecule. Other than sequencing, digital PCR (dPCR) has been applied to detect single mutations [5].

Methods for detecting somatic copy number alterations (SCNAs)

Somatic copy-number alterations are a hallmark of cancer. SCNAs can be detected by random low-coverage whole-genome sequencing and especially late-stage cancer cfDNA samples exhibit a wide variety of detectable SCNAs [10, 11, 12]. However, due to high background noise, many gains and losses can only be detected at tumor fraction >10%

[11]. Focal amplifications (small regions in the genome that are present at high copy-numbers) maybe an especially attractive target for these analyses as the small region may pinpoint important cancer driver genes and the high copy-number might make it amenable for analysis even at lower tumor fractions [12].

Methods for detecting methylation

Methylated CpGs can be detected with two approaches: bisulfite-sequencing and methylation microarrays. Both approaches need to convert unmethylated Cs into Us (which are subsequently read as T) by applying bisulfite-treatment. This procedure can impact the integrity of DNA fragments and thus may lead to loss of already lowly concentrated cfDNA samples. Nevertheless, methylation analyses so far have facilitated analyses of cancer hypomethylation [13] as well as tissue deconvolution [14, 15] Recently, a variation of MeDIP sequencing has been reported for use with cfDNA [16]. In this approach, only methylated fragments are co-immunoprecipitated by the use of methylation sensitive antibodies. Non-methylated fragments are washed away and the remaining fragments can be sequenced.

1.1.2. *ctDNA applications*

The minimally-invasive nature of ctDNA analyses brings many advantages in cancer diagnostics. While molecular analyses of tumor tissue is not going to be replaced any time soon, some applications in cancer diagnostics seem especially suited for using cfDNA/ctDNA analytics:

ctDNA for early cancer detection

A general cancer screening test from peripheral blood that could be applied to a general non-symptomatic population is considered to be the application of ctDNA analytics with the highest impact in oncological medicine. Earlier detection of almost any cancer type increases long-term survival and treatment response. This led to the creation of many different incentives to implement a screening test in both academia and industry. [1]. So far, reports on early cancer detection included the analysis of cfDNA mutations and proteins [17], fragment length distributions [18] and DNA methylation [16]. Apart from

academic institution, several companies are trying to bring early cancer detection to a wider population.

ctDNA for treatment monitoring

One of the advantages of ctDNA analyses over traditional tissue biopsies lies in the fact, that it allows longitudinal sampling due to the minimally invasive procedure. Thus, changes in the tumor genome can be monitored. In fact, the levels of ctDNA (as measured by e.g. mutant allele frequencies of known somatic mutations) has been found to be predictive of survival [19]. Remaining tumor-derived fragments in the peripheral blood are associated with a worse prognosis and indicative of the presence of minimal-residual disease. The absence of ctDNA after treatment is called molecular response, since this might be an independent predictive marker of treatment response [20].

Detection of treatment resistance mechanisms

The longitudinal sampling by venipuncture allows to monitor changes in the tumor genome over a period of time. Certain types of cancer treatment have been shown to induce tumor evolution, a mechanism that allows tumor cells to grow even in the presence of an anti-tutor agent. Research at the Institute of Human Genetics shows evolutionary changes in tumor genome in prostate and colon cancer patients by copy-number analyses of cfDNA samples in late-stage disease before and after therapy [12, 21]. Moreover, it has been shown that ctDNA not only comprises genomic mutations of the primary tumor but could also reflect changes that happened in distant metastases [22].

1.1.3. *cfDNA properties*

The analysis of cfDNA is obstructed by two key properties. Firstly, cfDNA is present in very low concentrations in most patients. While cancer patients show higher concentrations, many laboratory protocols need to be adapted in order to cater for low input amounts of DNA. The same property also puts a lower threshold for the resolution of mutation analyses. Since only a few genomic equivalents are present in the sample, only mutations which are present at a certain allele frequency can be robustly detected. This is further complicated by the fact, that cfDNA fragments coming from the tumor are diluted

by an order of magnitude by cfDNA fragments coming from cells of the hematological system [15].

Secondly, another complication in the analysis of cfDNA is its fragmentation. In healthy patients cfDNA fragment size distributions have a mode at approximately 166bp with a second mode twice this size (see Fig. 1.1) [23, 24]. There are reports that show that cancer-derived fragments may be shorter than that [25], while this might not be an overall property of any kind of cancer [26, 27]. However, a size difference has been established for fetal-derived fragments in the non-invasive prenatal diagnostics setting [28]. Recently, it has been shown that these size differences can also be exploited in order to detect the presence of tumor-derived cfDNA fragments [18].

1.1.4. *Tissues of origin*

Cell-free DNA is a convolution of apoptotic DNA of several cell types. In healthy people mostly hematological cells (including erythrocyte progenitors) contribute DNA, while liver cells, endothelial and neuron cells do so to a lesser degree [15]. In pregnant women, a fraction of the DNA is derived from the placenta, while in cancer patients the tumor itself sheds DNA into the peripheral blood [14]. In cancer patients, methylation has been shown to correctly identify the primary tissue by applying Bayes-based classification on a per-read level [29]. Recent advances in methylation analysis reported advantages of using single purified cell-types as a reference rather than tissues, which themselves are a mixture of individual cell types [15].

Other than cancer, different diseases might also lead to changes in cfDNA composition that might be picked up by tissue-of-origin analyses. In patients after pancreatic islet transplantation, a significant fraction of cfDNA is derived from pancreatic acinar and beta cells, in sepsis patients with liver damage liver-derived fragments are more abundant and in sepsis without liver damage the amount of cfDNA fragments of granulocytes has been reported to be increased [15]. Initial results from analyses of nucleosome positioning also hint at the possibility of tissue-of-origin detection, however, no complete deconvolution was shown and primary tissue of origin also was only shown in 3 out of 5 samples [30].

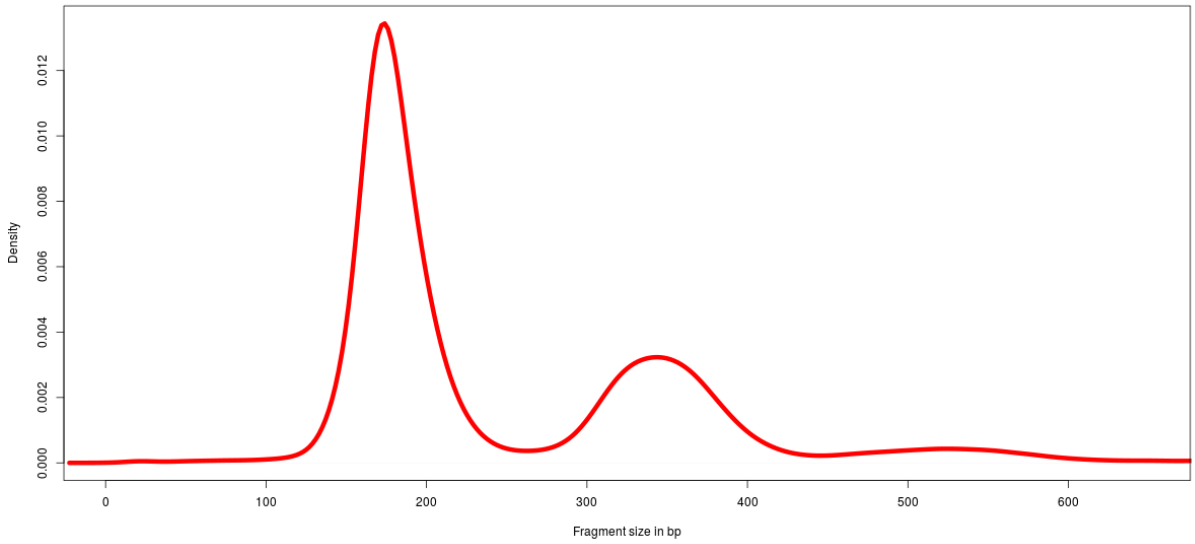

Figure 1.1: Size distribution of ctDNA in a random selection of reads from a healthy control after sequencing. While a large fraction of fragments are presumably derived from mono-nucleosomes (size 167bp), other fragments come from di- and even tri-nucleosomes.

1.1.5. *Nucleosomes*

Within the nucleus of each cell, DNA is packaged in order to reduce the space needed for the large molecule to fit within the nuclear membrane. Packaging occurs hierarchically in several orders. In a very early packaging step, 147bp of DNA is wrapped around a protein complex, which consist of histone octamers [31], thus forming a nucleosome. Between them, a stretch of DNA between 10 to 90bp long connects nucleosomes and is known as linker DNA. Linker DNA is either bound by histone H1 or not bound to any protein [31] (see Fig. 1.2). This first step of packaging DNA reduces space requirements for DNA within the nucleus, giving rise to a total of about 10,000-fold decrease in volume [32]. Recently, nucleosomes also have been described as having a large influence on the mutation rate in nuclear DNA [33].

Traditionally, analyses on nucleosomal derived fragments were performed by using micrococcal nuclease (MNase-seq) to digest bits of DNA that are not protected from cleavage by the histone association. Depending on the time of digestion, fragments are either partially digested or completely digested to only contain the portion of DNA that is directly bound to histones (other than linker histone H1) [34]. However, MNase-seq

has now largely been abandoned for studying chromatin organization in favor of easier protocols and stronger (positive) signals by DNase-Hypersensitivity assays or ATAC-seq.

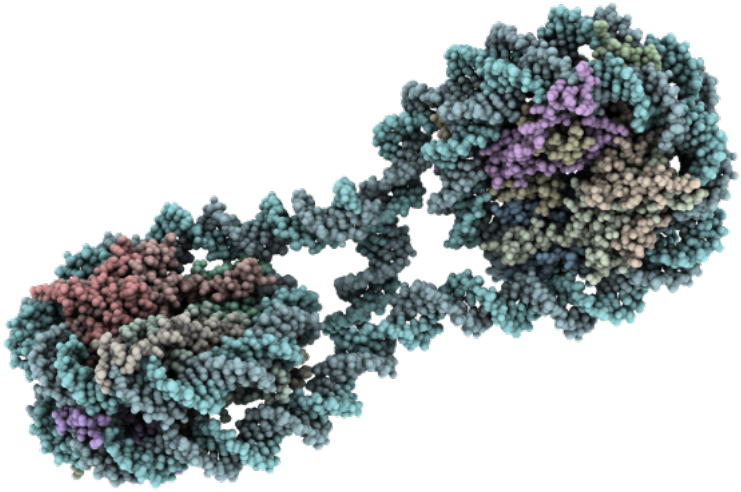

Figure 1.2: DNA (turquoise) is wrapped around two histone octamers, connected by linker sequence. [Image generated by Qutemol (v0.4.1) from PDB accession: 1zbb].

1.1.6. *cfDNA-Nucleosome association*

Generally, cfDNA fragments represent the whole genome, however, when looking in detail, a non-random fragmentation pattern is observable [35, 36]. Several lines of evidence show the association of cfDNA fragments to histone proteins, which preferentially prevent digestion of the DNA fragment:

Fragment lengths

Fragment lengths of cfDNA show a very peculiar distribution. While having an enrichment at 166bp, there is usually also a second peak at twice the size of the first peak (see Fig. 1.1 [5]. This is in line with the DNA lengths that would be expected from degradation in apoptosis and while the first peak most likely represents DNA that is bound to a single octamer of histones, the second peak could represent a fragment where no digestion has taken place between nucleosomes.

In turn, fragmentation patterns of fragments that are derived from mitochondria lack this specific distribution, since they also lack a higher-level DNA compaction mechanism [35, 36]. Apart from mono- and di-nucleosome peaks, cfDNA also exhibits relative en-

richments of fragment lengths at 10.4bp periodic intervals below 166bp. This has been attributed to the helical core of the nucleosome. Special library preparation methods can enrich for these smaller fragments [30, 37].

A/B compartments

Another level of evidence that cfDNA is primarily associated to histones is the finding that predicted peak-to-peak spacing from cfDNA sequencing is associated to hetero- and eu-chromatin regions (sometimes also referred to as A-/B- compartments) [30]. Thus, cfDNA might in fact represent chromatin organization at different levels of detail.

Nucleosomes at ordered nucleosome arrays

A region that shows very stringent positioning of nucleosomes on the peri-centromeric part of chromosome 12 is often used in studies of nucleosome organization to show that methods can enrich for nucleosomally-protected fragments (e.g. from MNase-seq studies). Coverage of whole-genome sequencing of cfDNA has also been shown to display regularly-spaced peaks and a good correlation to signal strength from MNase-seq data [30, 38].

Nucleosomes at transcription start sites

At transcription start sites, DNA needs to be accessible in order for the transcription machinery to facilitate gene expression. This includes transcription factors as well as the polymerase itself [39]. However, in unexpressed genes accessibility is not required and thus accessibility of DNA sequences at TSSs is low. This can be exploited to infer gene expression, simply from coverage data of cfDNA samples. Here, expressed genes that have accessible promotors show lower coverage than genes that are unexpressed. Moreover, strong nucleosome synchronization has been shown for housekeeping genes, especially at the first nucleosome after transcription start [38].

1.2. Epigenetic regulators

Many regions in the human genome exert some kind of function in order to maintain the cell. These regulatory regions fall in different classes e.g.:

- Transcription factor binding sites
- Enhancers
- Polyadenylation recognition sites
- G-quadruplexes
- Histone modifications
- Transcription start sites

While cfDNA can certainly not recapitulate all these various regulatory mechanisms, it can be used in epigenetic mechanisms that influence nucleosome ordering.

1.2.1. *Transcription factors*

Transcription factors are a special class of proteins, that can influence transcription of certain genes either as monomers or, more commonly, by building multimers. Most of these proteins bind DNA based on a certain sequence (i.e. the sequence motif) and thus recognize their targets [40].

Since transcription factors need to bind DNA, they either need to compete for space on the DNA strand with nucleosomes or interact with nucleosomes. This can also be achieved by cooperation of various transcription factors [41]. Moreover, some transcription factors can directly influence and displace nucleosomes (i.e. pioneer factors, e.g. FOXA1) [42].

Other transcription factors rely on chromatin remodeling systems that displace nucleosomes in order to facilitate their binding [43].

Although transcription initiation and transcription factor binding is a very complicated process, preferential non-random nucleosome positioning has been shown for MNase-seq data, a process where DNA that is not protected from histones is degraded [44]. Thus, similar nucleosome profiles should be detectable in cfDNA as cfDNA recapitulates the DNA digestion in-vivo.

1.2.2. *Enhancers*

Enhancers are genetic elements in the genome that facilitate mRNA transcription even at a large distance to the actual gene locus itself [45]. This long-distance interaction between enhancer elements and the respective gene promoters is mostly facilitated by

three dimensional proximity of those loci with the help of cohesin and CTCF, which construct topologically associated domains [46].

1.2.3. *G-quadruplexes*

G-quadruplexes are special structures in DNA mostly found at nucleosome-depleted regions and telomeres. Especially single-stranded DNA can easily fold into this structure, but also double-stranded DNA can form quadruplex structures, when primed by small molecule ligands [47]. Special sequence motifs have also been shown to prime the formation of quadruplex structures. [47]. Moreover, enhanced formation of quadruplex structures and de-novo formation have beed described to induce and influence transcription activity [48].

1.2.4. *Histone modifications*

Histone proteins act as the first mechanism to package DNA into the small cellular nucleus. DNA that is bound by histones is inaccessible to protein binding, thus many mechanisms that control the genome architecture also need to interact with nucleosomes in one of several ways [39]. Histone proteins can be modified for more granular control of DNA accessibility. Several modifications have been discovered to date on specific rescues of histone proteins, mainly on Lysine residues of histone proteins H3 and H4 [49]:

- Acetylation
- Methylation
- Phosphorylation
- Ubiquitylation

Recently, histone modification have also directly been assessed in cfDNA by direct ChIP-seq on cfDNA fragments by using antibodies against several histone modifications [50].

1.3. **Aim of the thesis**

The aim of this thesis is to explore whether epigenetic regulatory regions in the genome can be analyzed by the analysis of cell-free DNA. In a first step, patterns are being detected

in a set of healthy control data (n=24) and abnormal variations of these patterns are searched for in a set of cfDNA data from late-stage tumor patients.

To this end the following aspects will be investigated in particular:

- Explore patterns in coverage data at sites of known regulatory regions
- Explore variation in these patterns in a set of 24 healthy control samples
- Detect patterns in selected tumor samples and compare to control samples
- Validation of differences in specific regulatory regions by external data

2. *Methods*

2.1. **Probands**

The study was approved by the Ethics Committee of the Medical University of Graz (approval numbers 21-227 ex 09/10 [breast cancer], 21-228 ex 09/10 [prostate cancer], 21-229 ex 09/10 [colorectal cancer], and 29-272 ex 16/17 [High resolution analysis of plasma DNA]), conducted according to the Declaration of Helsinki and written informed consent was obtained from all patients and healthy probands, respectively. Retrospective human plasma samples from patients with colon cancer (Freenome cohort) were acquired from five biobanks for patients diagnosed with COAD and healthy controls. All plasma samples used in this analysis were de-identified prior to receipt, with no key available to re-identify [51].

2.2. **Blood sampling**

Peripheral blood was collected from patients with metastatic prostate, breast and colon cancer at the Department of Oncology and from anonymous healthy donors without known chronic or malignant disease at the Department of Hematology at the Medical University of Graz. CfDNA was isolated from plasma using the QIAamp Circulating Nucleic Acids kit (QIAGEN, Hilden, Germany) in accordance with the manufacturer's protocol. Shotgun libraries were prepared using the TruSeq DNA LT Sample preparation Kit (Illumina, San Diego, CA, USA) following the manufacturer's instructions with three exceptions. First, due to limited amounts of plasma DNA samples we used 5-10 ng of input DNA. Second, we omitted the fragmentation step since the size distribution of the plasma DNA samples was analyzed on a Bioanalyzer High Sensitivity Chip (Agilent Technologies, Santa Clara, CA, USA) and all samples showed an enrichment of fragments in the range of 160 to 340

bp. Third, for selective amplification of the library fragments that have adapter molecules on both ends we used 20-25 PCR cycles. Library preparation was performed by employees by the Institute of Human Genetics Graz [51].

2.3. Sequencing

Control and high-coverage tumor samples were sequenced on the Illumina NovaSeq S4 flowcell at 2x150bp by the Biomedical Sequencing Facility at CeMM, Vienna, Austria. For the control samples, an average of 435,135,450 (range: 352,904,231-556,303,420) paired-end reads were obtained. For the tumor samples (P40_1, P40_2, P147_1, P147_3, P148_1, P148_3, C2_6, C2_7), an average of 688,482,253 reads (range: 541,216,395-870,285,698) were sequenced. Additional samples were sequenced using the Illumina NextSeq platform (B7_1, B13_1, P190_3, P170_2, P179_4, P198_5, P240_1; average sequencing yield: 195,425,394 reads; range: 115,802,787-379,733,061). Low-coverage tumor samples which were used to create single-entity pools were sequenced on either the Illumina Next-Seq or MiSeq platform. This resulted in 382,306,130 reads from 69 prostate cancer samples, 254,490,128 reads from 60 breast cancer samples and 604,080,473 reads from 100 colon cancer samples. Samples from the Freenome cohort were paired-end sequenced on the Illumina NovaSeq platform [51].

2.4. Transcription factor analyses

2.4.1. *Transcription factor binding site definitions*

The GTRD database combines publicly available data of ChIP-seq experiments for transcription factors and combines ChIP-seq peaks to clusters. For every cluster, the samples that support a peak and the position with the highest combined signal is recorded [52]. Here, data from the GTRD database were downloaded and individual BED files per transcription factor were extracted based. The position was recalculated by focusing on the reported point where the meta-cluster has the highest ChIP-seq signal. An additional BED file was created which only includes peaks that are supported by >50% of the maximum number of samples analyzed for this specific transcription factor. All BED files were then converted to hg19 (from original hg38) using the liftOver tool provided by UCSC

[53].

2.4.2. *Transcription factor binding site overlaps*

In order to check whether binding sites of transcription factors overlap, regions of the binding sites from GTRD are increased by 25,50 and 100bp, respectively, on either side using bedtools slop. Subsequently, the number of overlaps is calculated by using bedtools intersect via pybedtools [54] for every transcription factor with every other transcription factor [51].

2.4.3. *Single-end sequencing data preparation*

In order to enhance the nucleosome signal, reads were trimmed to remove parts of the sequencing read that is associated to the linker region. Hence, forward reads were trimmed to only contain base 53-113 (this would be the central 60bp of a 166bp fragment). Reads were then aligned to the human hg19 genome using bwa and PCR-duplicates were removed using samtools rmdup [55]. Average coverage is calculated by bedtools genomecov [54, 51].

2.4.4. *Paired-end sequencing data preparation*

Paired-end reads were aligned to the human hg19 genome using bwa mem [56] and PCR duplicates were marked with picard MarkDuplicates [57].

2.4.5. *MNAse-seq data preparation*

BAM files of MNAse-seq experiments of GM12878 were downloaded from the ENCODE portal. Reads in BAM file were trimmed directly from the BAM file using pysam [54]. In brief, left-most alignment positions in the BAM file were shifted 53bp in the respective direction and the sequence length was adjusted to 60bp. The coverage patterns were then calculated in the same way as the trimmed cell-free DNA sequencing data [51].

2.4.6. *Copy-number variation*

For control data paired-end alignments were subsampled using samtools view to only include 2% of the initial alignments and converted to fastq using samtools fastq [55].

Plasma-Seq [11] was applied to the subsampled FastQ files. In brief, reads are aligned to the human hg19 genome and reads are counted within pre-specified bins. The bin size is determined by the amount of theoretically mappable positions to account for differences in mappability throughout the genome. Read counts are normalized for total amount of reads and GC-content of bins are corrected for by LOESS smoothing over the GC-spectrum. Moreover, corrected read counts are normalized by the mean read counts of non-cancer controls per bin to control for additional positional variation [51].

2.4.7. *Coverage patterns at transcription factor binding sites*

For every transcription factor in the GTRD, coverage patterns were calculated. To this end, coverage data was extracted for every region using pysam count_coverage in a region ± 1000 bp around the defined binding sites [55]. Coverage data at every site are normalized by regional copy-number variation and by mean coverage. For every position around the TFBS, coverage is averaged and 95% confidence intervals are calculated. If >100,000 positions are defined for a transcription factor, 100,000 sites are randomly chosen to be analyzed. For the Freenome cohort, reads were aligned to the human hg38 genome using BWA-MEM 0.7.15 [56]. Midpoints of paired-end reads were extracted and average midpoint counts at transcription factor binding sites (+/- 1250bp) were calculated per transcription factor.In order to better compare the data to the aforementioned samples, a running median (window-size = 30) was applied to the data using the numpy.convolve function in a single dimension [51].

2.4.8. *ATAC-seq data analyses hematological samples*

Raw ATAC-seq data were downloaded from hematological cells from the Gene Expression Omnibus (Accession: GSE74912). The full count matrix was divided by sample. Next, all reads of bins that overlap defined binding sites of a respective transcription factor were summed and divided by the number of total reads within this sample [51].

2.4.9. *ATAC-seq data analyses cancer samples*

Raw ATAC-seq data matrices of TCGA samples were downloaded for colon adenocarcinoma, prostate carcinoma and breast carcinoma samples [58]. Again, data was divided by

sample and bins that overlap a specific transcription factor are summed up and divided by the total amount of reads for a specific sample [51].

2.4.10. *Insert sizes around transcription factor binding sites*

To see whether fragment sizes around transcription factor binding sites were biased, insert size data from paired-end analyses were used. Every position from -1000 until 1000bp from the binding site was traversed and (single-end) reads where the central 3bp around the midpoint are located at this position were fetched using pysam [55]. Also, paired-end alignments from the same sample were fetched and the insert size information was designated to the respective reads. All insert sizes at specific positions relative to the TFBS were then summarized and 1000 data points were sampled and plotted for each position in the range of -1000 to 1000bp from the TFBS.

2.4.11. *Measuring transcription factor binding site size*

In order to measure the size of the transcription factor binding site, the respective coverage pattern was smoothed using a third order Savitzky-Golay filter (window-size: 31). Peaks were identified by searching for data points that were larger than the neighboring 20 data points on either side. Peaks were removed if they resided within 50bp of the center of the supposed binding site. The distance between the closest peaks next to the binding site peak was specified as the transcription factor binding site size. Since binding site estimates are only reasonable if nucleosome synchronization is detectable, we filtered the signals by various criteria:

- High-frequency signal amplitude >0.1
- Mean normalized coverage of the central 100bp <1
- Amount of peaks is less than 15
- Median distance between peaks is >150bp
- The binding site sets comprises over 500 sites

228 binding site sets passed these filters and were used for binding site estimation [51].

2.4.12. *Measures of transcription factor activity*

As we hypothesize that two distinct signals make up the final coverage pattern at varying levels, we sought to deconvolve the pattern into both primary sources. The lower range frequency data is extracted by a Savitzky-Golay filter (3rd order polynomial and window size of 1001). A high-frequency signal is extracted by a different Savitzky-Golay filter (3rd order polynomial and window size of 51). The high frequency signal then is normalized by dividing by the results of the low-frequency signal. The data range of the high frequency signal then is recorded. Since, coverage profiles from transcription factors with few described binding sites are inherently noisier, a LOESS smoothing was performed over the signal range and the amount of described binding sites. The range values were corrected by the smoothed LOESS and ranks of the adjusted range were calculated.

2.4.13. *Signal deconvolution*

In an extension of the frequency-based signal filtering a more complex deconvolution was performed. The signal is assumed to be made up of several underlying basis functions (see Fig. 2.1):

- A sinusoid function with a specific frequency
- A distance element between the nucleosome peaks closest to the binding site
- An amplitude modulating envelope function
- A trending baseline

Firstly, the frequency of the signal with the highest power is determined by wavelet analysis. Then a local trend is calculated using the Savitzky-Golay filters mentioned above (3rd order polynomial and window size of 1001). Moreover, the envelope function is approximated by identifying the peaks and troughs of the coverage signal. Peak identification is done by searching for values that are higher than the neighboring 20 values from the de-trended smoothed coverage signal. Trough identification is done by inverting the sign of the signal and performing the same procedure. A smooth spline is fitted on both the peaks and the troughs of the coverage data. Lastly, the distance element between the first two nucleosome peaks is calculated by finding peaks of the coverage signal and searching for the nearest peaks to the binding site. This is compared to a hypothetical signal that

would com from the cosine function of the frequency that is specified by wavelet analysis and finding and subtracting the difference thereof. Thus, one can assemble all the individual functions in order to reconstruct the individual signal. It starts with the cosine function with the frequency that is calculated in the wavelet analysis. Next, the distance element is incorporated by increasing the start from point zero by the size differences in the nucleosome peaks and replacing the part between zero and the differences with values of -1. Then, the amplitude of the signal is modulated by the envelope function that was approximated using splines from the peaks and the troughs, respectively. Lastly, the trending baseline is added.

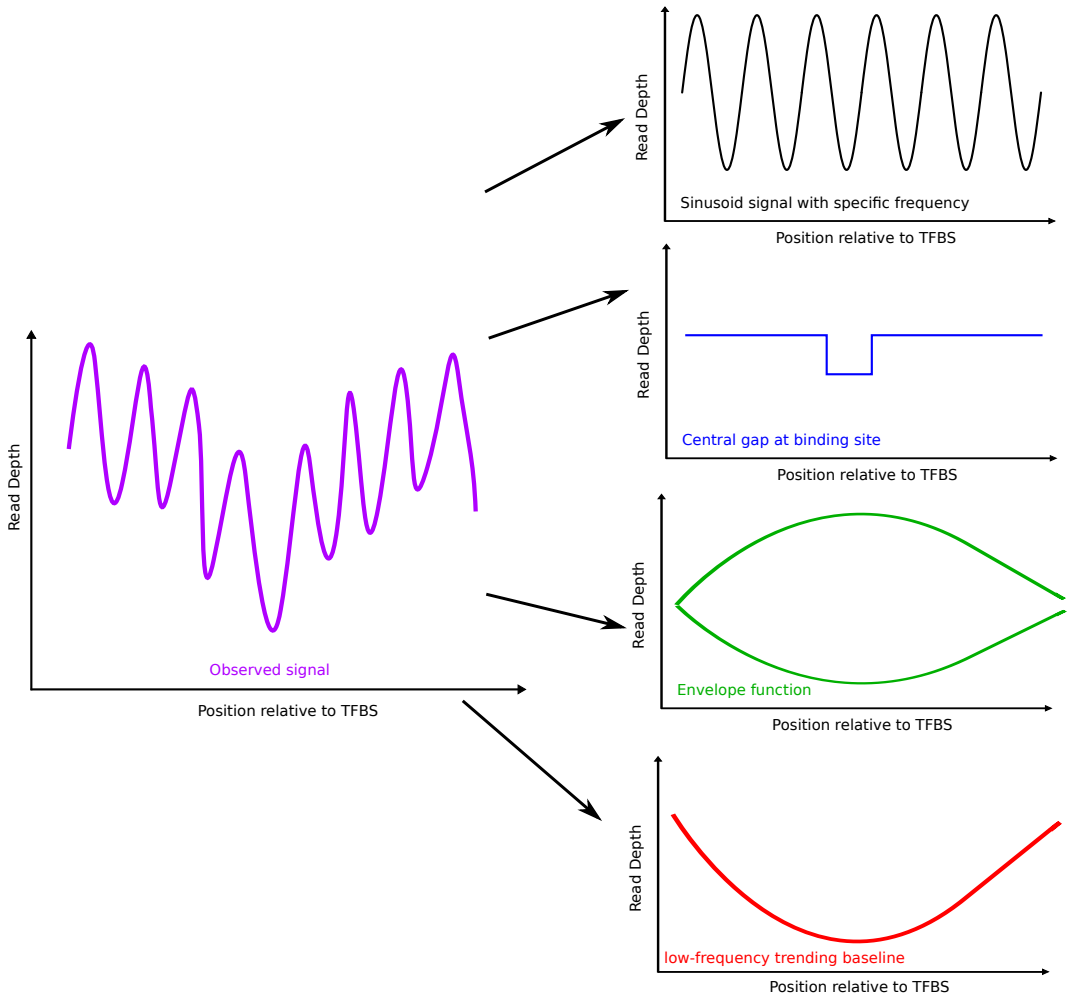

Figure 2.1: The primary signal is deconvolved into four primary signals. A cosine function with a specific frequency, a distal element that separates the most proximal nucleosome peaks to the binding site, an amplitude-modulating envelope function and a trending baseline.

2.4.14. *Comparing tumor and control samples*

In order to compare tumor and control samples, the ranks of the respective transcription factors in the adjusted range values were compared. Rank differences were calculated between a tumor sample and every control sample and mean rank differences were recorded. Moreover, z-scores were calculated for every transcription factor from the accessibility ranks by taking the respective rank and subtracting the mean rank of the control samples and dividing by the standard deviation of the transcription factor ranks of the control samples (RankDiff z-scores). In another round, z-scores were calculated for the overall deviation of transcription factor accessibility in two samples. To this end,accessibility values of each of the 24 healthy samples were compared to the remaining 23 and the rank difference was recorded. The rank differences were used to estimate a normal distribution and z-scores for the rank differences over all transcription factors were calculated (Overall z-scores) [51].

2.4.15. *DNAse hypersensitivity data analysis*

BAM-files from DNAse hypersensitivity experiments were downloaded from the ENCODE database for GM12878 and LNCaP cell lines. Binding site regions of a transcription factor were increased by 25bp on either side using bedtools slop [54]. Coverage at the respective binding sites was extracted using mosdepth [59] and normalized by million mapped reads per sample [51].

2.4.16. *Logistic Regression*

In order to classify samples, logistic regression was applied to the accessibility values of all 504 transcription factors in 769 samples (177 control samples and 592 samples of cancer patients). To this end, the LogisticRegressionCV from the scikit-learn package [60] was applied using 5-fold cross-validation and balanced class weights to correct for the slightly unbalanced sample set. In 100 permutations, samples were split into training and test sets. Training samples were used to establish the model whereas (held-out) test sets were used to estimate the actual performance of the model. Mean performance metrics,as well as ROC curves (based on the prediction probabilities of the LogisticRegression) of the

models were calculated from the 100 permutations [51].

2.5. Enhancer analyses

2.5.1. *Enhancer region definitions*

Enhancer definitions were retrieved from the FANTOM5 database [61]. Predefined enhancer tracks for cell-type and tissue-specific enhancers are readily available at [62].

2.5.2. *Reference region*

In order to compare the average coverage to random regions, a reference region set was created for every enhancer region set. Briefly, for every region 100 regions with the same length are randomly selected from the human genome using the random function in pybedtools [54]. GC-content is calculated for the input region as well as the 100 random regions. The region with the most similar GC-content is saved. Thus, the reference region set contains the same amount of regions, with the same length distribution and very similar GC-content and should be an ideal model for similar genetic regions for comparison of coverage data. Finally, region sets are sorted and gonosomal regions are removed.

2.5.3. *Coverage analysis*

Average coverage per region is calculated using mosdepth.

2.5.4. *Calculate coverage differences in region sets*

For every region set and the corresponding reference set, Cohen's D effect size was calculated:

$d = \frac{\mu_1+\mu_1}{\sqrt{(sd_1^2+sd_1^2)/2}}$

Moreover, the ratio between the mean of the coverage of enhancer regions and the mean of the coverage of reference regions was calculated. Confidence intervals of both measures were calculated by bootstrapping using R's boot package (1000 iterations) [63]. Lastly, t-tests between the distributions were calculated.

2.5.5. *Nucleosome positioning*

Coverage data was extracted as described above for the transcription factors. For every enhancer in a given track, coverage patterns were calculated. To this end, coverage data was extracted for every region using pysam [55] count_coverage in a region ± 1000 bp around the defined startpoint, midpoint and endpoint, respectively. Coverage data at every site are normalized by regional copy-number variation and by mean coverage. For every position around the defined anchorpoint, coverage is averaged and 95% confidence intervals are calculated. If >100,000 positions are defined for an enhancer, 100,000 sites are randomly chosen to be analyzed.

2.6. **Analyses of G-quadruplexes**

2.6.1. *G-quadruplex region definition*

G-quadruplexes were predicted using pqsfinder on the human hg19 genome in R [64]. Regions on chromosome X and Y were removed and predicted quadruplex forming regions were split according to the amount of predicted GC-tetrads (3,4 and 5, respectively). Moreover, only regions with perfect tetrads were considered for the following analyses.

2.6.2. *Coverage analyses*

Coverage analyses were performed similarly to transcription factors by averaging coverage data over all defined G-quadruplex predictions.

2.7. **LOLA definitions**

LOLA provides a database of epigenetic annotations [65]. For this analysis 2323 annotations were downloaded from the LOLA webportal [66] and coverage profiles were generated anchoring at the midpoint of the defined annotations. Accessibility values were calculated similarly to transcription factors, using Savitzky-Golay filters removing low-frequency signals.

2.8. chromHMM

2.8.1. *Definitions*

chromHMM state predictions for a neutrophil was downloaded from the ENCODE roadmap homepage [67]. This contains predictions for several primary cell types. Since a large portion of cfDNA fragments is suspected to come from neutrophil granulocytes, predictions of this cell type were used for further analyses [68].

2.8.2. *Coverage analyses*

Coverage distributions were calculated from the downloaded BED file with mosdepth [59]. This calculates the average coverage at each specified genomic region. Coverage distributions were plotted divided per state using ggplot2 [69] in R [63].

2.8.3. *Fragment length distribution*

In order to check whether fragment size distributions vary between predicted chromatin states, fragment sizes of paired-end sequencing data were extracted. In brief, the chromHMM prediction BED file [67] was split to only contain a single state. Then, for every BED file fragment sizes of paired-end reads were extracted from 1,000 random regions per state. Only reads that overlap the center of that region ($\pm$100bp) were considered. Fragment size distributions were extracted using pysam [55] and plotted in R. In order to normalize for varying number of available fragment sizes, a random subset of 10,000 fragment sizes were used to plot.

2.9. Histone modification

2.9.1. *Annotations*

Annotations of known histone modification were downloaded from both Blueprint and ENCODE as BED files. For all of the mentioned analyses ChIP-seq data from neutrophil cells that each target a single modification were used. Since these annotations only capture broad regions of histone modifications, additionally raw signal in BigWig format were downloaded from the ENCODE database. The highest signal in each region from the

BED file was searched by extracting base-wise fold-enrichment values from the BigWig files using pyBigWig. The position of the highest signal was denoted and output as another (single-position) BED file.

2.9.2. *Fragment lengths at various histone modification*

To see whether fragments display varying fragment sizes, reads were extracted at the regions of the highest ChIP-seq signal (±10bp) and their fragment size (as measured by distance between two paired-end reads) were denoted. Only reads were used where the fragment midpoint (±10bp) overlaps the selected region. In order to assess the variability of the fragment size distributions 100 samples of 10,000 fragment sizes each were compared.

2.10. **Differential nucleosome protection between paired cancer samples**

Using bedtools makewindows function [54], 100bp windows of the human hg19 genome were generated. Average coverage in these 100bp windows was calculated in samples P148_1 and P148_3 using mosdepth [59]. Since both samples harbor copy-number alterations to quite some extent, copy-numbers of the respective 100bp window (measured by Plasma-seq [11]) were noted for both samples.

In a next step a null model was created in order to better estimate the variation in coverage differences between 100bp windows. To this end, for both samples, 100,000 random coverage values were extracted, subtracted from each other and the mean and standard deviation of the differences was calculated. Windows with coverage of 0 in either sample were excluded. Finally, Z-scores are calculated for every window by calculating coverage differences, subtracting the mean of the null model and finally divide the result by the standard deviation. Hence, the resulting Z-scores specify how many standard deviations the coverage difference in that particular window is away from the mean difference.

LOLA enrichment is being performed in order to investigate overlaps with predefined epigenetic features [65]. Separate analyses for windows with higher coverage in sample 1 and sample 3 are performed.

3. *Results*

3.1. **Transcription factors**

3.1.1. *Definition of transcription factor binding sites*

Transcription factor binding sites are typically analyzed by ChIP-seq, in which DNA bound to a transcription factor and the respective factor are co-precipitated. A sequencing library is generated by the precipitated DNA after washing away all proteins. Single experiments are subject to noise coming from background signals due to non-specific immunoprecipitation. The GTRD database combines publicly available data of ChIP-seq experiments for transcription factors and combines ChIP-seq peaks to clusters. For every cluster, the samples that support a peak and the position with the highest combined signal is recorded [52]. The highest peak of the combined data gives a good starting point for anchoring multiple coverage profiles. In total, transcription factor binding sites of 676 transcription factors were used for further analyses. Because of the potentially high number of TFBS to which TFs bind with variable frequencies, we generated average TFBS occupancy patterns based on three different stringency criteria: first, for all TFBS for all tissue samples in the GTRD; second, for those peaks supported by >50% of the maximum number of samples (subsequently referred to as ">50%-TFBSs"; in these two analyses all 676 TFs from the GTRD were included), and third, we took the 1,000 TFBSs, which were supported by the majority of samples from each TF ("1,000-msTFBSs"; 504 TFs fulfilled this criterion).

3.1.2. *Coverage patterns at binding sites*

In order to check for signs of nucleosome ordering around transcription factor binding sites, 24 samples of healthy probands were sequenced (12 male, 12 female). Average coverage relative to transcription factor binding sites were recorded for every transcription factor in every sample and indeed preferred nucleosome occupancy was found for many transcription factors (see Fig. 3.1). Especially in transcription factors that are important in hematopoiesis, such as PU.1, LYL1, SPIB, a non-random coverage pattern of synchronized nucleosome positioning was detected.

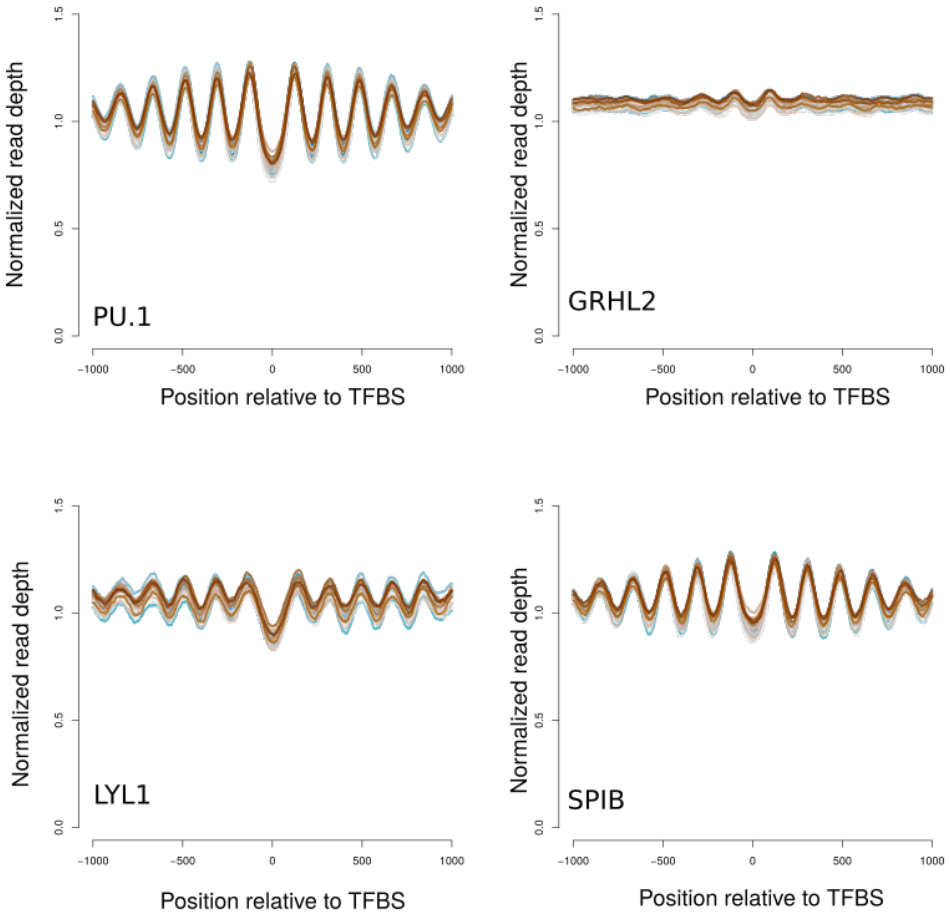

Figure 3.1: **Coverage analysis for hematological TFs and GRHL2.** Coverage analysis for transcription hematological factors PU.1, LYL1, SPIB and epithelial transcription factor GRHL2, shows high-amplitude non-random coverage around the binding sites in 24 healthy healthy samples

3.1.3. *High molecular weight DNA*

Coverage patterns around transcription factor binding sites were analyzed in whole-genome sequencing data of conventional high-molecular weight DNA. This was done in order to exclude possible biases due to biased sequence and GC-content, respectively. While some variation in the patterns are visible, they differ to a great extent from the coverage patterns seen in cfDNA patterns (see Fig. 3.2).

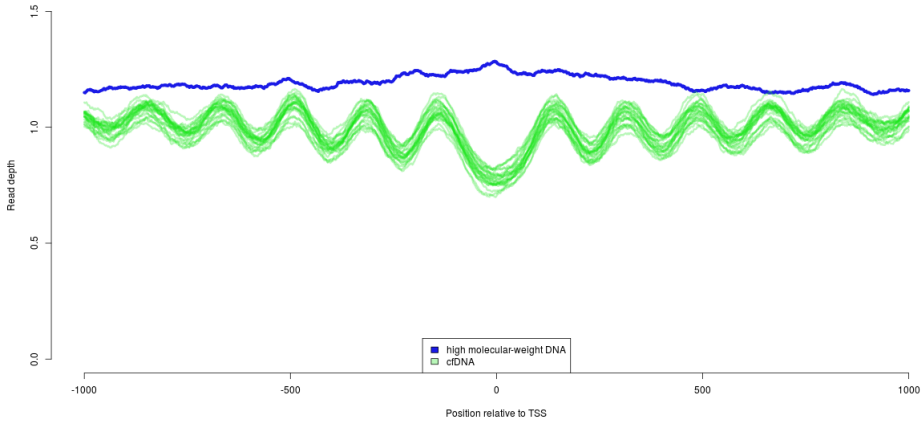

Figure 3.2: **Coverage patterns of high-molecular-weight DNA.** High-molecular weight DNA (blue) shows no non-random sinusoid signal of preferential nucleosome positioning in comparison to the healthy samples for transcription factor ELF-5

3.1.4. *Comparison to MNase-seq data*

The nucleosome association of DNA can also be analyzed by in-vitro digestion of DNA not bound to nucleosomes. This is done using micrococcal nuclease and is called MNase-seq. Since this should recapture very similar signals then are present in cfDNA, coverage patterns are compared for every transcription factor. While most of the patterns are indeed very similar, some notable differences were observed in transcription factors that play a role in B-cell development.

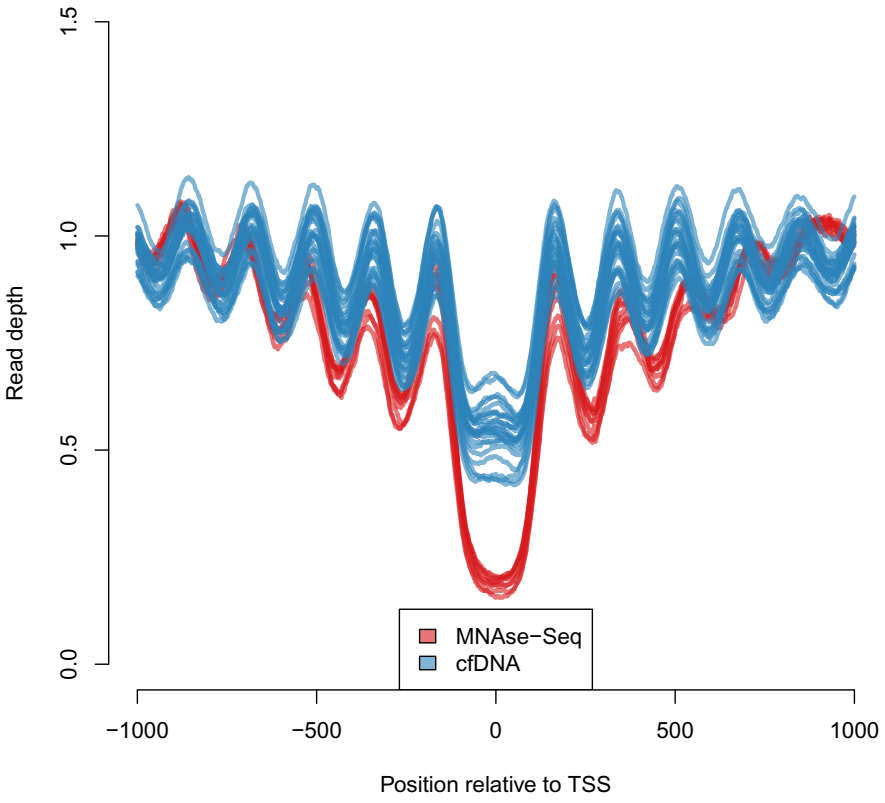

Figure 3.3: **Coverage analysis in MNase-seq data.** Very similar non-random nucleosome positioning is observed in both MNase-seq data from GM12878 (red) and cfDNA data (blue) for transcription factor CREB

3.1.5. *Overlap between transcription factor binding sites*

Many transcription factors are binding in large complexes in order to facilitate transcription. Thus, overlapping binding sites in transcription factors that regulate the same genes are expected to be found in the GTRD database. Here, transcription factor binding sites (at the basepair level) were enlarged by 25b, 50bp and 100bp on either side and overlaps were calculated. Overall, most transcription factors show no overlap as depicted in figure 3.4. The transcription factors with the highest overlaps are mostly from the same protein family or otherwise related.

The transcription factors with the most overlap (restricting to transcription factors that have over 1,000 binding sites defined in GTRD) are shown in table 3.1.

Table 3.1: Transcription factors with the highest overlap of binding sites

TF1	**TF2**	**Overlap Fraction**
USF-1	USF-2	0.763
PBX-2	PBX-3	0.701
CREB	ATF-1	0.669
REST	MIER2	0.592
ZNF143	SIX5	0.585
ZNF75A	GABPA	0.582
JunD	Fra-2	0.574
Dp-1	E2F-4	0.571
CREB	CREM	0.567
c-Jun	JunB	0.558

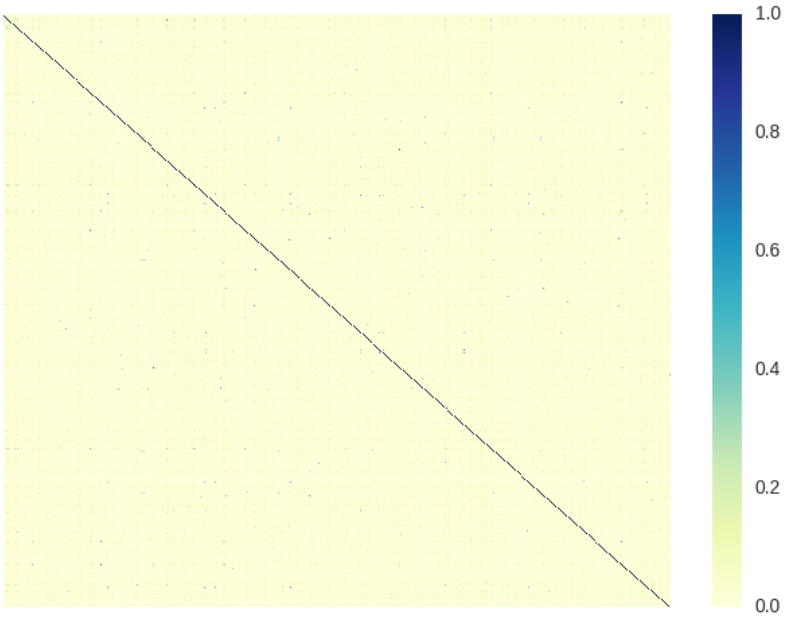

Figure 3.4: **Overlaps of binding sites between transcription factors.** Transcription factors are plotted on both axes. Overlaps are depicted by colors (colorscale on the right. Most transcription factors display very few overlapping sites.

3.1.6. *Fragment sizes*

Since nucleosomes are phased in certain transcription factors in the blood, variability of fragment sizes around phased nucleosomes were analyzed. To this end, fragment sizes were recorded around CTCF transcription factor binding sites. Indeed, fragment sizes vary a lot along the surrounding regions of CTCF binding sites (see Fig. 3.5). The shape of this pattern has been noted in analyses of MNase-seq data [32].

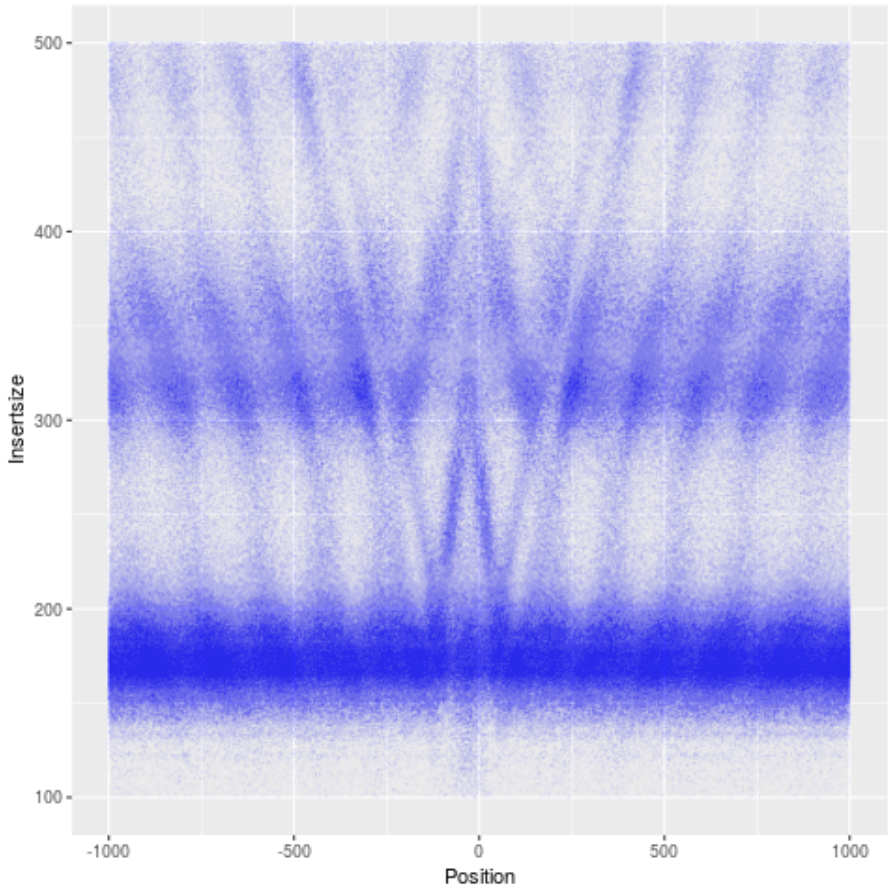

Figure 3.5: **Insertsizes around CTCF.**Fragment sizes of cfDNA fragments vary a lot along CTCF binding sites.

3.1.7. *Binding site sizes*

Due to transcription factor binding synchronized nucleosome positioning was observed in many transcription factors and while the spacing between nucleosomes are constant throughout the whole ±1,000 bp range, the nucleosome peaks immediately adjacent to the defined binding site display variable distances.

While many transcription factors lead to regular spacing of nucleosomes (e.g. AP4), some coverage patterns display very wide troughs with low representation of fragments (e.g. CREM) (see Fig. 3.6).

In total, 55 coverage profiles showed a central gap that exceeded 300 bp, from these 26 had binding sites close to di-nucleosomal sizes (312-352 bps). Since the wide spacing around transcription factor binding site might be an effect of CpG island promotors adjacent to the binding sites, overlap of transcription factors,that display wide spacing

of central nucleosomes with pre-defined CpG islands was analyzed. Indeed, transcription factors with binding sites >300 bp show both an enrichment in CpG islands ($p=4.2x10^{-11}$; two-tailed Mann-Whitney U test) as well as transcription start sites ($p=8.5x10^{-12}$; two-tailed Mann-Whitney U test). (Fig. 3.7)

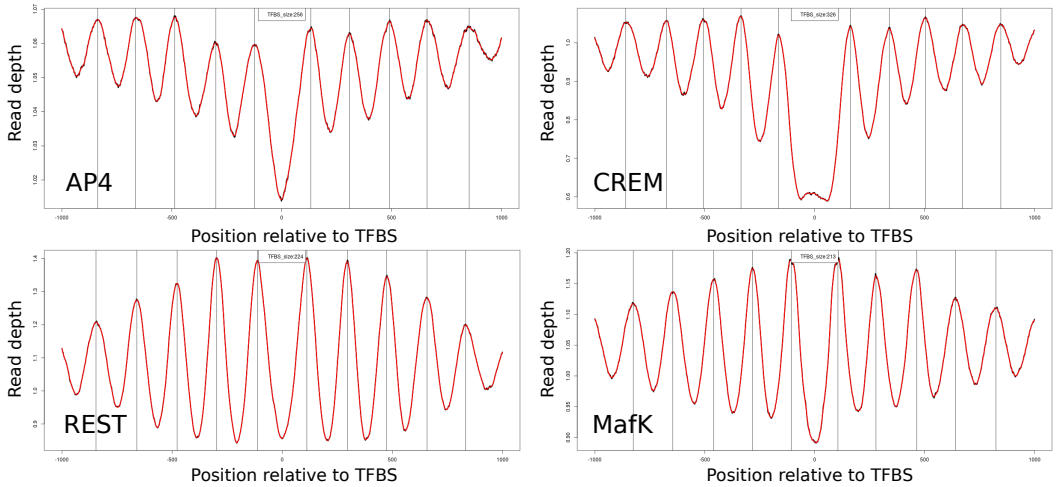

Figure 3.6: **Binding site size.**The size of the transcription factor binding site varies between transcription factors. Especially, CREM displays a large gap between adjacent nucleosome peaks.

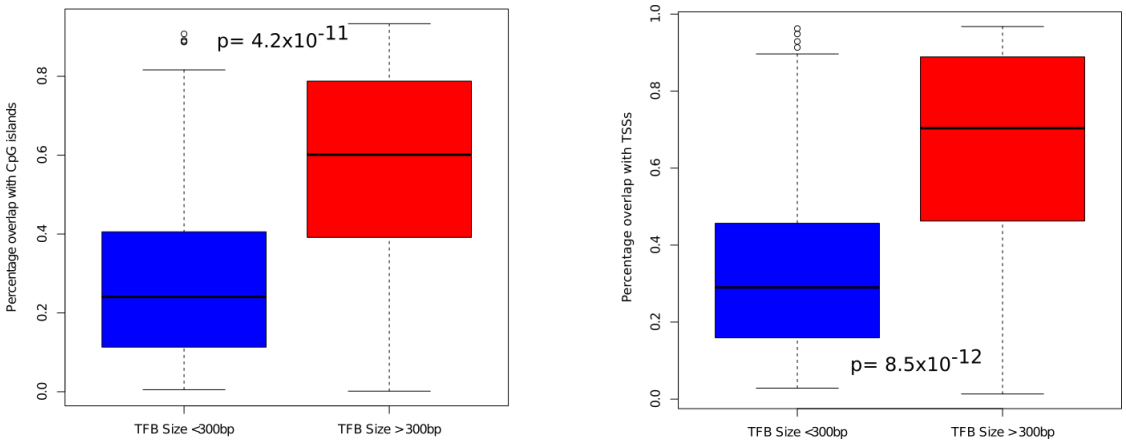

Figure 3.7: **Binding site sizes overlap with CpG sites.** Transcription factors with a large binding site (>300bp) show a higher overlap with pre-defined CpG islands than transcription factors with smaller binding site sizes.

3.1.8. *CTCF*

CTCF displays the largest signal throughout all the analyses performed. This might be due to the fact, that CTCF, apart from its role as transcription factor, also serves

the special purpose of regulating topological genome architecture in conjunction with the protein Cohesin [46]. These two proteins create topologically associated domains (TADs) that are conserved very well throughout evolution, since this association has a strong influence on encapsulation and gene expression regulation. CTCF may bind to a very diverse set of genomic sequences due to its combinatorial use of 11 different zinc finger motifs [70]. Here, CTCF sites in the GTRD were further split into sites that overlap transcription start sites (TSSs), are distant to TSS (>2kbp), as well as sites that overlap previously reported TAD boundary regions [71] and sites that do not overlap those boundary regions. These analyses were done for both, all sites available in the GTRD, and sites that are supported by more than 50% of the samples, respectively. In addition, data from an analysis that elucidated ultra-conserved CTCF elements in various mammals were plotted for comparison in both plots [72] (see Fig. 3.8).

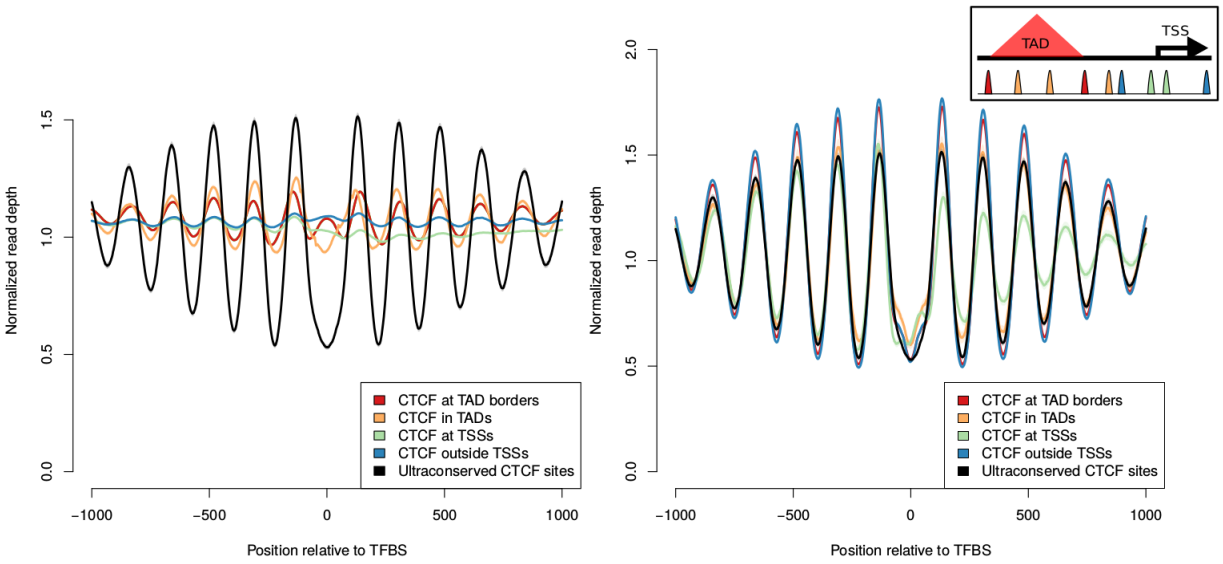

Figure 3.8: **Coverage patterns at CTCF sites.** Coverage analyses for all CTCF sites in the GTRD (left plot) divided by proximity to TSSs and overlap with previously described TAD boundary regions [71], and the same analyses for the sites that are supported by more than 50% of the samples analyzed (">50%-TFBSs"). For comparison an additional set of 5,000 ultra-conserved CTCF sites is plotted [72]. The box on the upper right explains the color code. No graph was used from the cited publications.

3.1.9. *Measuring accessibility*

To decipher the underlying structure of an average TFBS profile we conducted calculations separately for TFBS within and outside of TSSs (see Fig. 3.9).

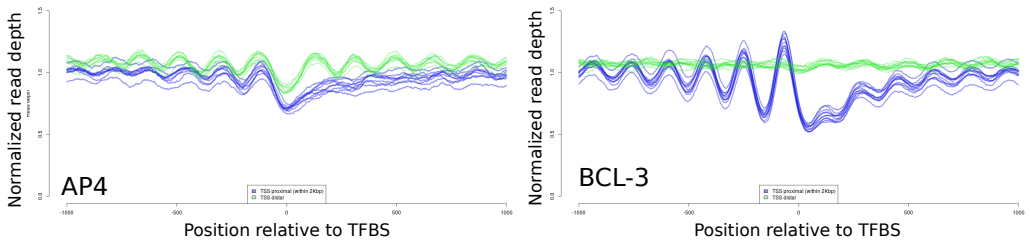

Figure 3.9: **Coverage patterns at TSSs.** Regions that are close to transcription start sites (TSSs) (depicted in blue) differ in their coverage profile from regions that are distal to TSSs (>2kbp, depicted in green).

Since local loss of nucleosomes before transcription start sites have a direct influence on the amplitude of the nucleosome peaks, two Savitzky-Golay filters using different window sizes were used to extract the two separate signals. The effect of the low-frequency signal was removed from the high-frequency signal and the amplitude of this "corrected" high-frequency signal served as a proxy for transcription factor accessibility (see Fig. 3.10a).

Subsequently, we recorded the data range (maximum minus the minimum of the data values, corresponds to the amplitude) of the high-frequency signal, corrected them by LOESS smoothing as they depend on the number of TFBSs (see Fig. 3.10b), and then used calculated ranks (see Methods) as measure for the accessibility of each TFBS. As TF binding only opens or "primes" its target enhancers, without necessarily activating them per se [73], we refer to the rank values as "accessibility score". With the exception of the 1,000-msTFBSs evaluations we normalized for the number of defined binding sites. As a further mean to explore TF accessibility we reconstructed an unbiased, de-trended signal at a period between 135 and 235 bp by wavelet analysis and summed up the powers of the signal across the 2,000bp flanking TFBSs (Fig. 3.10c-d). To further explore Savitzky-Golay filtering and wavelet analysis we used cfRNA data [74] and observed significantly reduced accessibility for unexpressed TFs (i.e. <0.01 FPKM [Fragments Per Kilobase exon per Million reads]) as compared to the accessibility of expressed (i.e. >10 FPKM) TFs (>50%-TFBSs; Savitzky-Golay filtering: $p=1.75x10^{-13}$; the sum of powers (wavelet analysis): p=0.0004049; 1,000-msTFBSs; Savitzky-Golay filtering: p=1.254x10-11; Mann-Whitney-U test each) (see Fig. 3.11). These differences were also significant when we compared the adjusted ranges to mean DNase coverage (>50%-TFBSs; Savitzky-Golay filtering: $p<2.2x10^{-16}$; the sum of powers (wavelet analysis): p<2.2x10-

16; 1,000-msTFBSs; Savitzky-Golay filtering: $p<2.2x10^{-16}$; Mann-Whitney-U test each) (see Fig. 3.11). We then defined detection thresholds for TFBS accessibilities deviating from the normal samples as ± 3 mean of the standard deviation (as a z-score of 3). For assessments based on all or >50%-TFBSs the detection thresholds for our normalized accessibility score were ± 253 and for the 1,000-msTFBSs, which have a lesser number of analyzable TFs, ± 88.

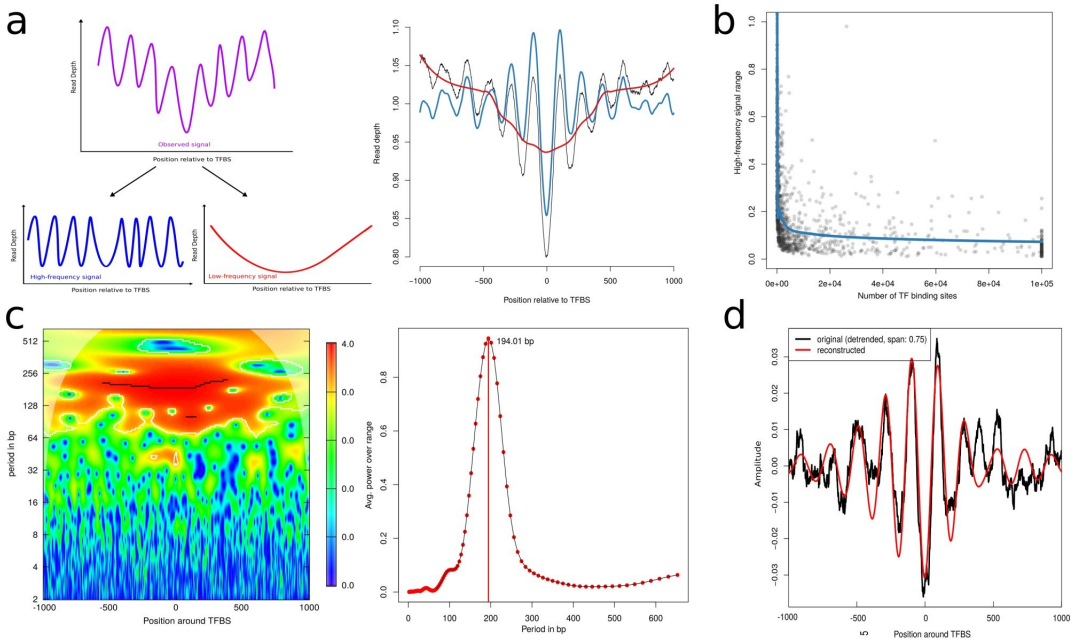

Figure 3.10: **Analysis of transcription factor accessibility.**
a) Coverage profiles are first filtered in a high-frequency and a low-frequency signal. The amplitude of the high-frequency signal is used to infer accessibility of the transcription factor.
b) Since the amount of binding sites defined per transcription factor varies, a LOESS model was calculated to remove these effects. c) Left: Heatmap of periods around the transcription factor binding site by wavelet analysis. The quantiles of the signal power distribution are displayed by colors. Right: The average power of periods for the transcription factor GRHL2 shows a peak at 194.01 bp.
d) The de-trended reconstructed signal after wavelet analysis compared to the original (de-trended) coverage profile of GRHL2. Parts of this figure were published in [51].

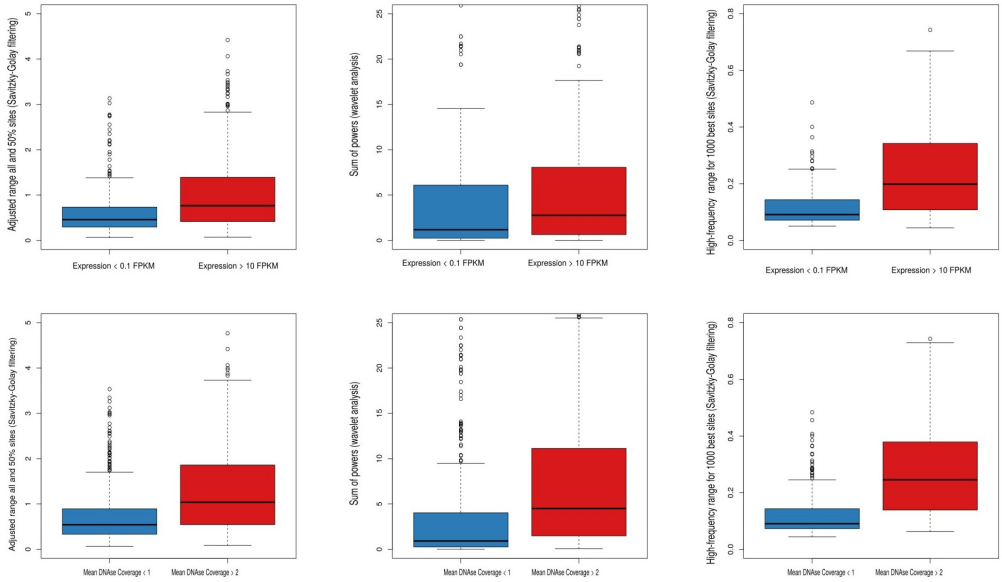

Figure 3.11: **Validation of transcription factor accessibility.**
Upper Panel) Transcription factors that are expressed in blood (as measured by Koh et al [74] with >10 FPKM consistently show higher accessibility when compared to transcription factors that are expressed <0.1 FPKM.
Lower Panel) Transcription factors that display a high average coverage at DNAse-seq experiments in GM12878 also show higher accessibilities when compared to transcription factors with low average coverage at their binding sites.

3.1.10. *Signal deconvolution*

As a further means to analyze the coverage profile, a more general approach of deconvolving the signal was taken. Basically, this is an extension of the frequency filtering with more parameters. The basis for the preferential nucleosome signal is a sinusoid function having a specific frequency with several modifications(see Fig. 2.1).

- A gap between nucleosome peaks next to the binding site
- An envelope function that alters the amplitude
- A shifting baseline with a trough exactly at the binding site

Deconvolution and reconstruction of the coverage signal yields a comparable signal for some transcription factors, but not for coverage profiles that are subject to a lot of background noise. Also, the bias that results from transcription factor binding can be estimated quite well (see Fig 3.12).

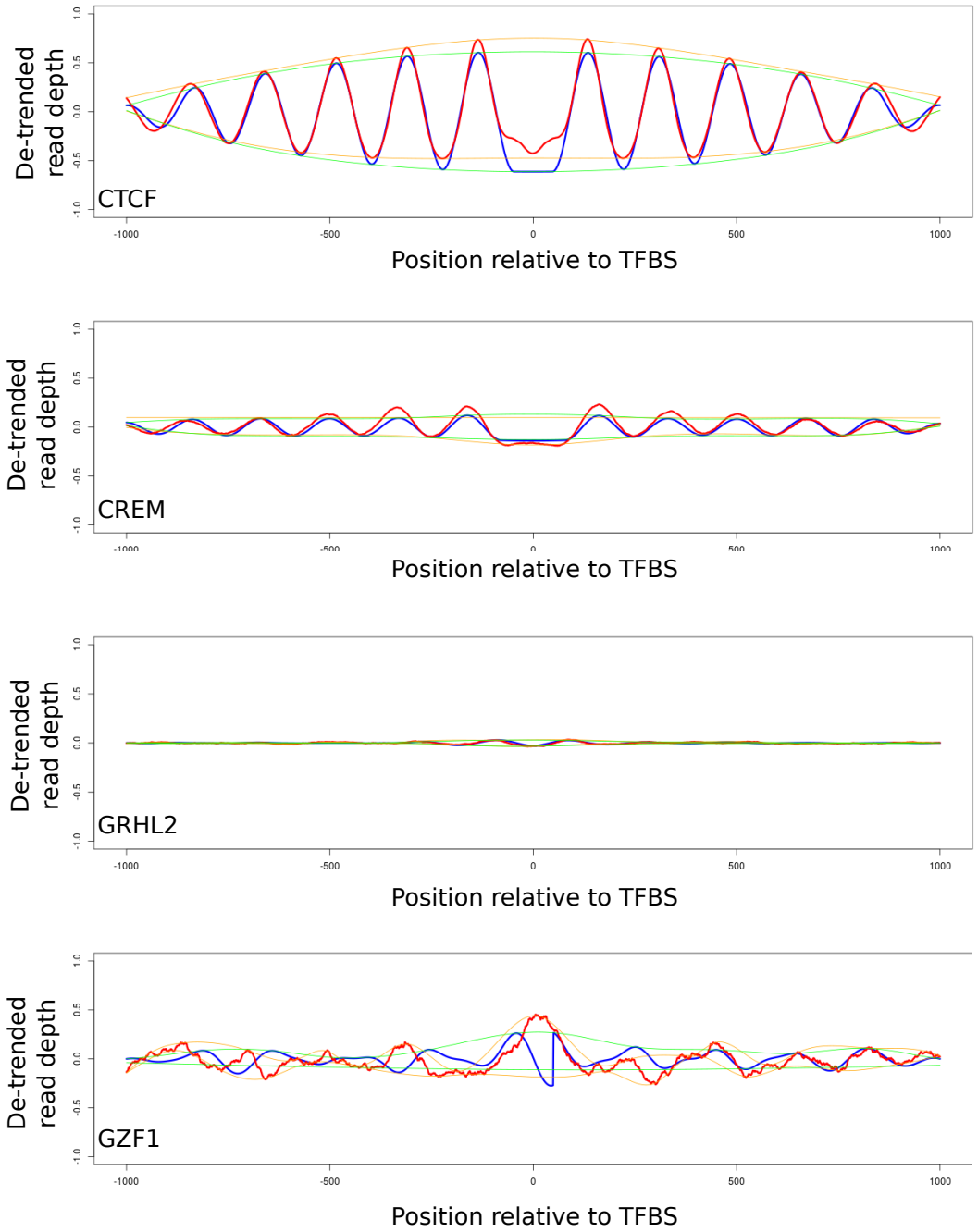

Figure 3.12: **Signal mixture deconvolution results.** The primary signal is deconvoluted into four primary signals and then reconstructed. The original (de-trended) signal is displayed in red, the reconstructed signal in blue. Orange and green depict the envelope function of the original signal and the reconstructed signal, respectively. This reconstruction works well in transcription factors with a high amplitude (CTCF) and also can give quite accurate estimations about a bias around the binding site (CREM). Also, low-signal are generally no problem (GRHL2), however, high background noise or a bad estimation of the frequency result in reconstructions that do not reflect the original signal (GZF1).

3.1.11. *Tissue-specificity of TFs*

In order to identify lineage-specific TFs, publicly available ATAC-seq data [75] was used and coverage data on the predefined TF binding sites was recorded. Different tissue-specific TFs were observed, some of which are elevated in all cancer tissues, when compared to ATAC-seq data of hematological cells (GRH-L2). As a confirmation for the robustness and reproducibility of lineage-specific TFs in cfDNA, pools of multiple samples generated by shallow-coverage whole-genome sequencing of cfDNA (<0.2x) [11] were analyzed. TFs with increased accessibility in the majority or all samples, i.e. lineage-specific TFs, should appear with an increased accessibility score whereas others will be averaged out. To this end, cfDNA samples from prostate (n=69), colon (n=100) and breast (n=60) cancer cases were pooled and TF accessibility calculated. The epithelial TF GRHL2 and the hematopoietic TFs reiterated their increased and decreased, respectively, accessibility patterns (3.13). Within the prostate cancer cfDNA pool the lineage-specific TFs AR, HOXB13, and NKX3-1 showed the expected increased accessibilities (see Fig. 3.13), suggesting that these features are universally present in prostate cancer and may be suitable for the identification of tumor-of-origin.

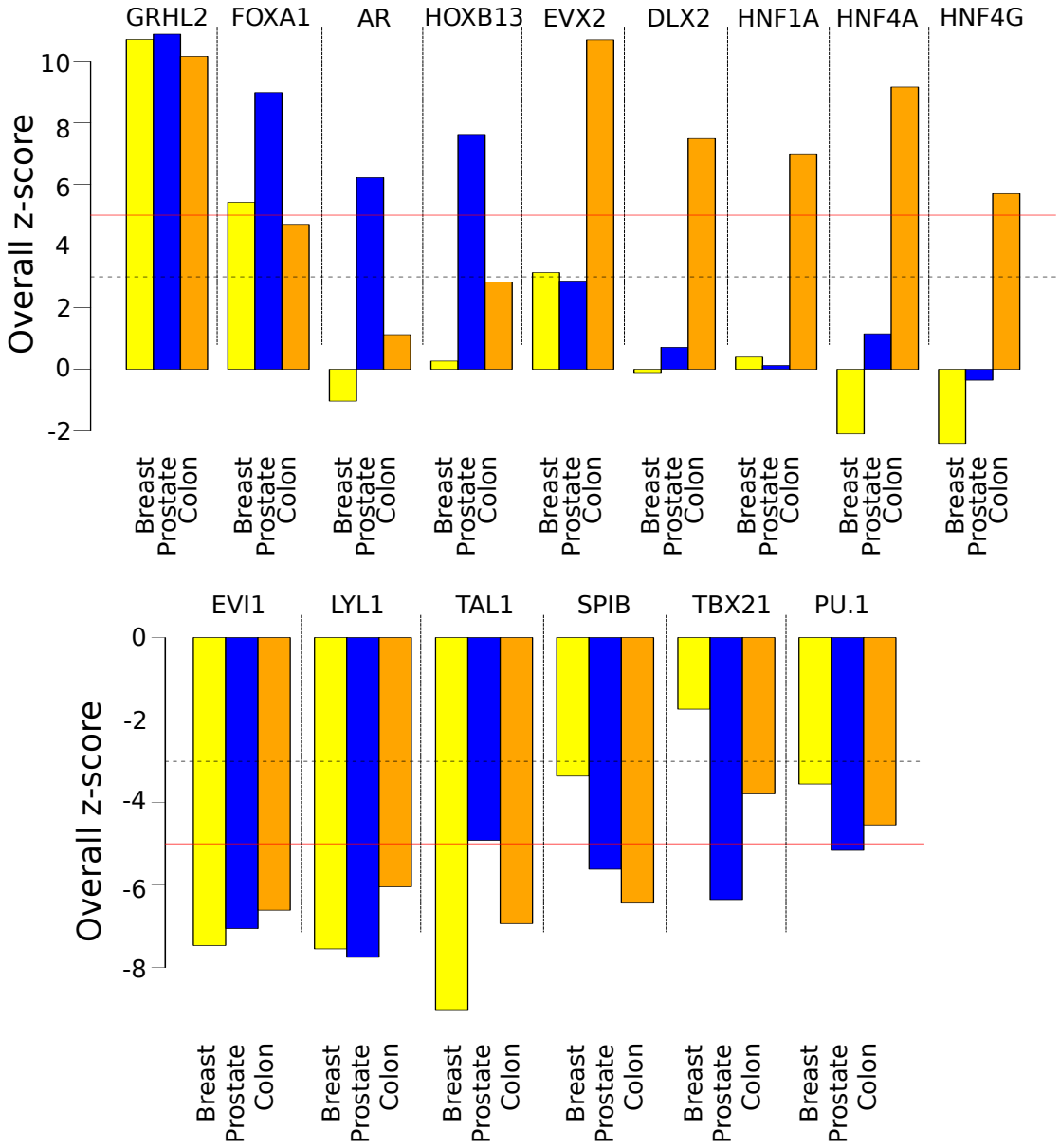

Figure 3.13: **TF accessibility in pooled samples.** Accessibility of pooled cancer samples when compared to healthy controls. Top panel: TFs with increased accessibility, bottom panel: TFs with decreased accessibility. While some TFs show higher accessibility in all cancer types (e.g. GRHL2), some are specific to prostate cancer (AR, HOXB13, and some are specific to colorectal cancer (EVX2, DLX2, HNF1A, HNF4A, HNF4G). This figure was published in [51].

3.1.12. *Single-sample WGS analyses*

As a next step, high-coverage data of late-stage breast, prostate and colon cancer samples were generated and used to identify aberrations of TF accessibilities. For baseline estimation of TF accessibilities, aforementioned data of 24 healthy control samples was used.

The amplitude of CTCF remained similar throughout all samples regardless whether the cfDNA was derived from healthy controls or from patients with cancer (Fig. 3.14a). This was consistent with DNase hypersensitivity assays from the ENCODE database for cell lines GM12878, LNCaP (androgen-sensitive human prostate adenocarcinoma cell line) and HCT116 (human colon cancer cell line) (Fig. 3.14a). However, patients with cancer have an increased fraction of ctDNA, which alters the balance between DNA from hematopoietic versus epithelial cells within cfDNA. Accordingly, the amplitudes for the hematopoietic TFs PU.1, LYL1, and SPIB decreased whereas the amplitude for the epithelial TF GRHL2 increased (Fig. 3.14b). These observations were again consistent with DNase hypersensitivity assays (Fig. 3.14b). As another example for a well-established TF we analyzed FOXA1, a TF widely expressed in different tissues where it controls cellular differentiation and organ function [76]. Furthermore, FOXA1 cooperates with nuclear hormone receptors in endocrine-driven tumors of the breast and prostate [77] and in prostate its expression has been associated with castration-resistant prostate cancer (CRPC) [78]. Indeed, consistent with DNase hypersensitivity assays, we observed preferentially increased accessibility of FOXA1 in the plasma samples of prostate and breast cancer patients (Fig. 3.14c). Overall transcription factor accessibility deviation from healthy control samples as measured in Z-scores can be seen in Fig. 3.15.

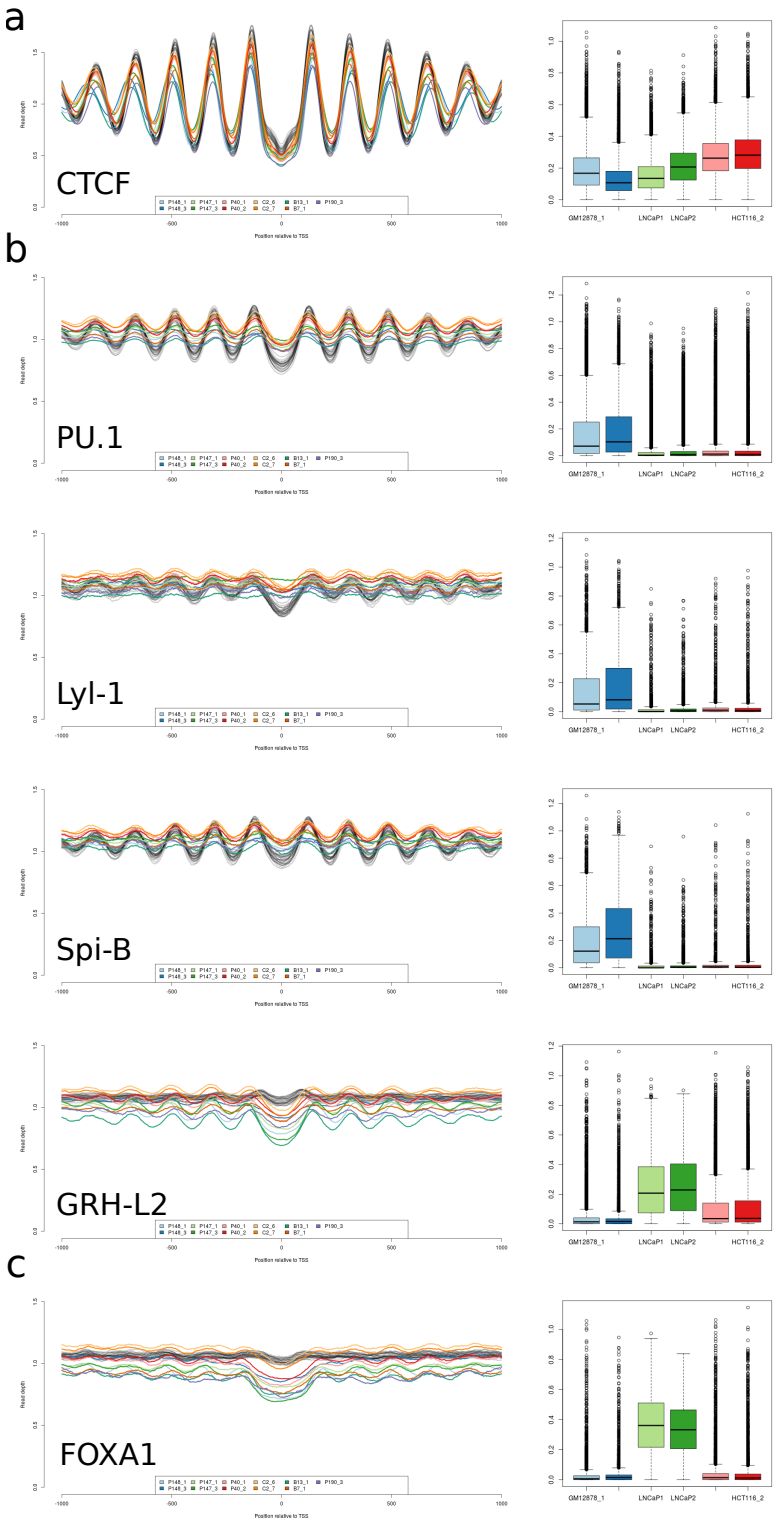

Figure 3.14: **Coverage profile in cancer samples.** Coverage profiles in cancer samples show comparable profiles for CTCF (a), decreased accessibility for hematopoietic transcription factors (b) and increased accessibility for FOXA1 in plasma samples of prostate and breast cancer (c). Changes in accessibility were concordant with analyses from DNAse Hypersensitivity experiments in cell lines (right panels). Parts of this figure were published in [51].

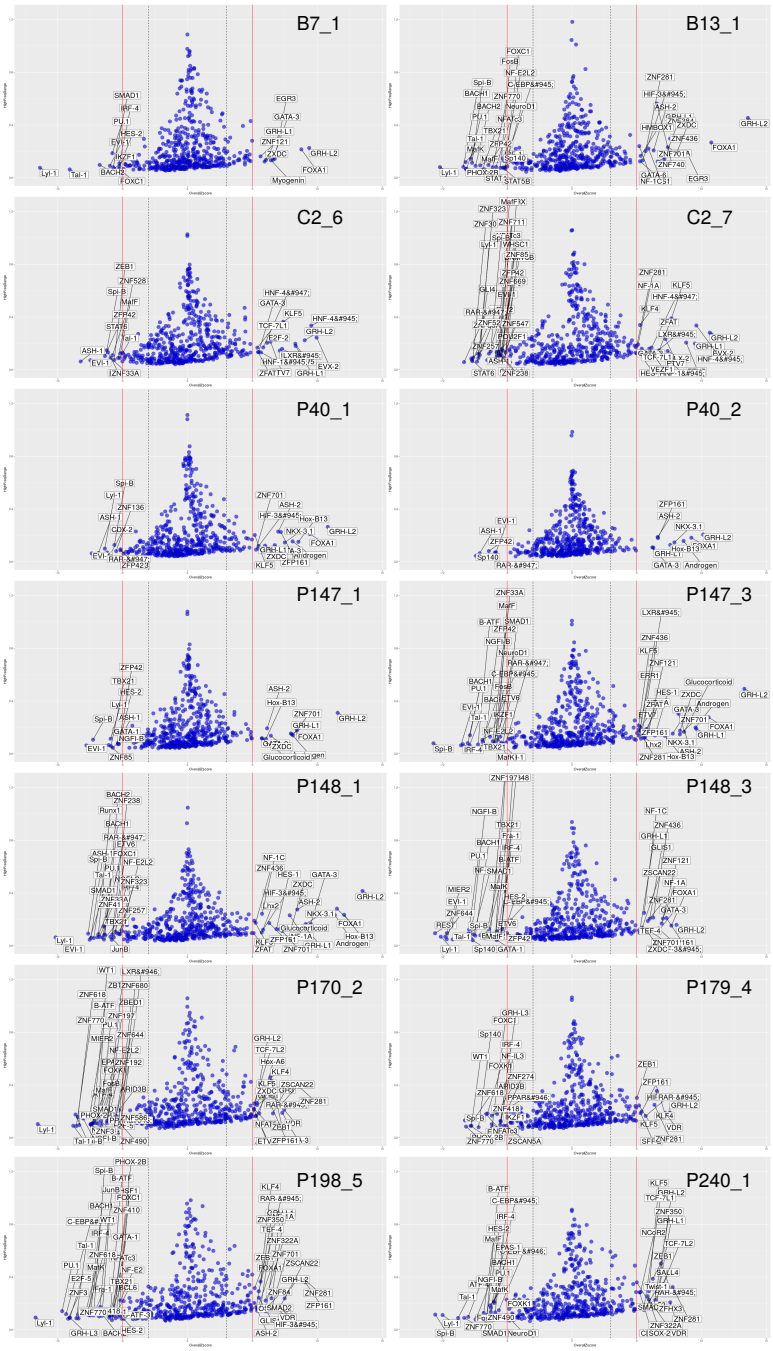

Figure 3.15: **TF accessibility in single samples.** Accessibility of single cancer samples relative to control samples, measured in Z-scores (X-acis). Y-axis denotes accessibility values. TFs with a deviation of Z-score >5 are highlighted.

3.1.13. *Differences in accessibility in paired-cancer samples*

To address the question whether TF accessibility remains stable over time two samples each from 4 patients were analyzed (P40, P147, P148, C2). No significant differences for three of the four plasma sample pairs was found (Controls: Median: 0.8404 ± 0.0196

(IQR); P40: 0.8620; P147: 0.8370; C2: 0.8719; each Kendall's Tau).

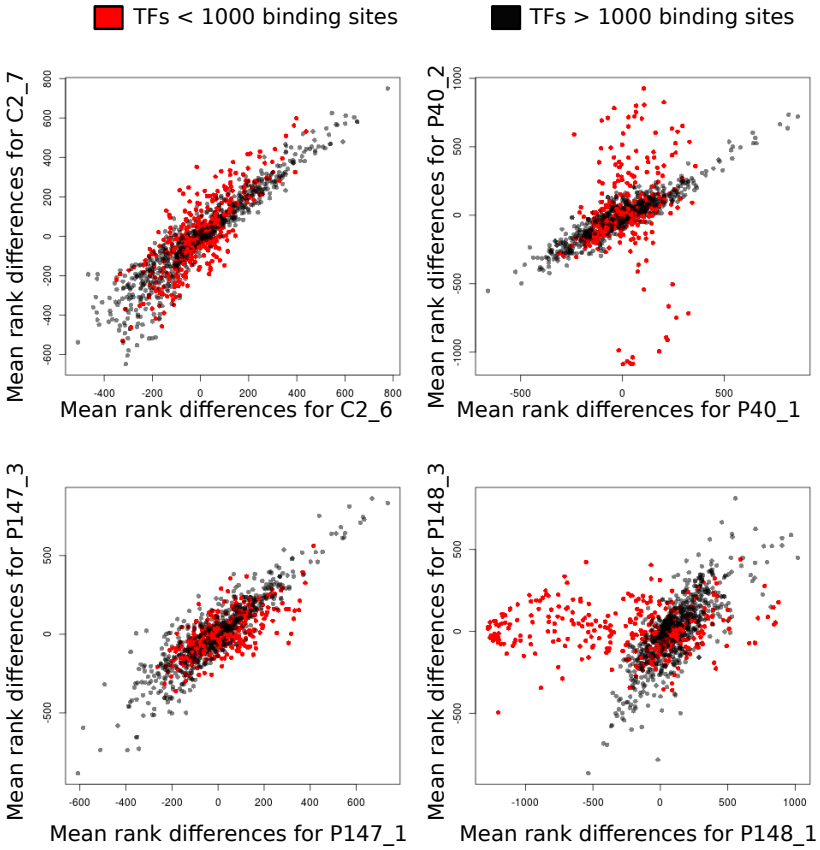

Figure 3.16: **Correlation of paired samples.** Correlation analyses of accessibilities in paired cfDNA samples of cancer patients. Only P148 shows a reduced correlation between samples of two different time points.

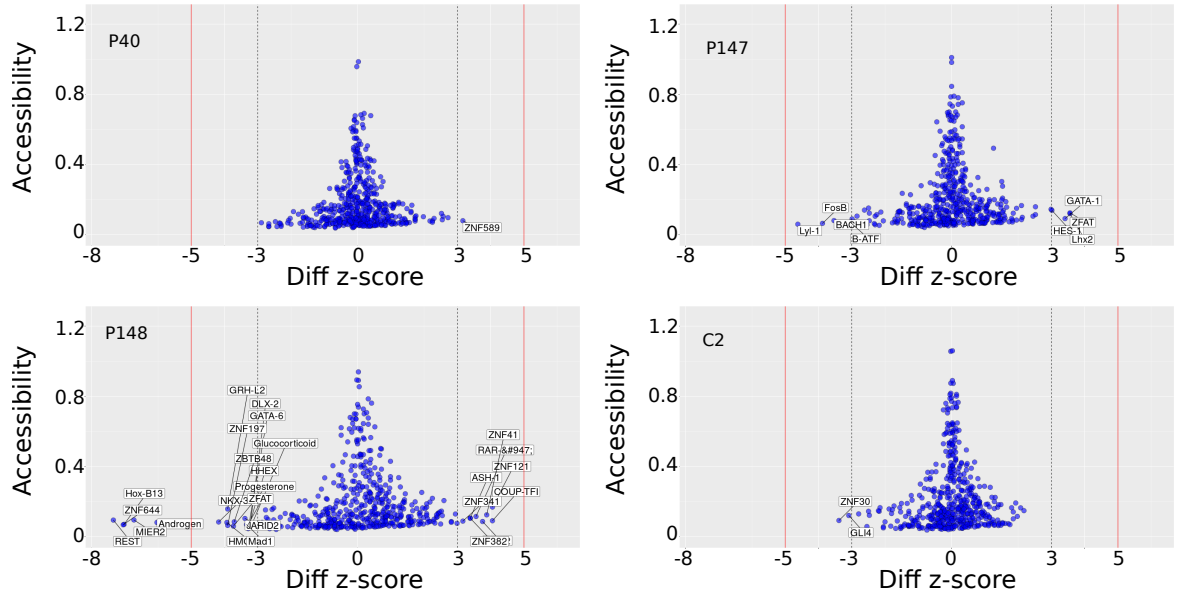

Figure 3.17: **Pairwise analysis of cancer samples.** Pairwise analyses of cancer samples show that only in P148 extensive differences between subsequent samples are found.

However, case P148 showed significant TFBS accessibility changes (Kendall's Tau: 0.7573) (Fig. 3.16). Within 12 months, the time interval between plasma samples P148_1 and P148_3, the prostate adenocarcinoma transdifferentiated to a neuroendocrine tumor, which was accompanied by a decrease of PSA (prostate-specific antigen) and an increase of NSE (neuron-specific enolase) [12]. The involvement of several TFs in such a transdifferentiation process has been extensively studied ([79, 80, 81]) and we used the z-score statistic to explore these TFs in cfDNA (Fig 3.17). Neuroendocrine tumors are no longer an androgen-dependent stage of prostate cancer [81] and consequently accessibility of AR binding sites is no longer needed, which was actually reflected in cfDNA (Fig. 3.20a). In addition, due to its close cooperation with nuclear hormone receptors in prostate cancer [77], accessibility to FOXA1 was correspondingly reduced (Fig. 3.20a). Furthermore, the change in the cell type identity became apparent as reduced accessibility to the binding sites of the prostate specific lineage TFs HOXB13 and NKX3-1 (Fig. 3.20c).

Similarly, the accessibility of the epithelial TF GRHL2 decreased substantially (Fig. 3.18). Importantly, we also observed changes in TFs associated with neuronal development. Hypoxia occurs frequently in advanced solid tumors and may facilitate the development of prostate adenocarcinoma to an androgen-independent state [82]. Indeed, the accessibility of GLI-similar 1 (GLIS1), a TF whose expression is dramatically increased

under hypoxic conditions [83], was significantly augmented (Fig. 3.18).

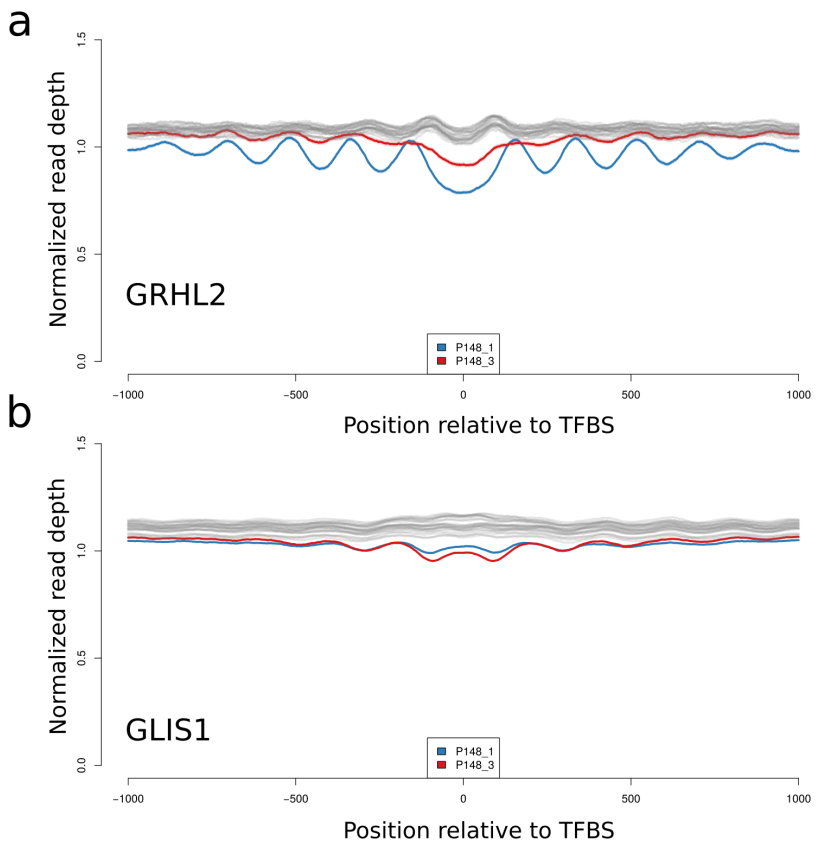

Figure 3.18: **GRHL2 and GLIS1 in P148.** Accessibility of GRHL2 decreases after the trans-differentiation to a neuroendocrine carcinoma in P148. On the other hand, accessibility of GLIS1 increases.

Furthermore, an additional transcription factor deviated a lot between the two samples of the same patient: repressor element-1 (RE-1) silencing transcription factor (REST). This transcription factor has already been shown to be important in neuroendocrine differentiation in prostate cancer [84] (Fig. 3.20c). This case suggested that a panel of TFs may be suitable to distinguish between different prostate cancer subtypes. In order to test whether tumor subtype classification based on TFBSs from cfDNA is possible whole-genome sequencing data of cfDNA from 4 further prostate cases (P170_2, P179_4, P198_5, and P240_1) were analyzed in addition as a proof-of-principle, all characterized by low PSA and high NSE values. Since only 50 million reads were available for these samples, TF accessibility values from down-sampled cfDNA data for P148_1 (819,607,690 reads) and P148_3 (768,763,081 reads) to 50 million reads were compared to TF accessibilities using all reads (see Fig. 3.19). The reduction of reads resulted in an increase

of noise levels, which was dependent on the number of TFBSs and neglectable for TFs with more than 1,000 TFBSs. Accordingly, accessibility analyses for the aforementioned highly relevant TFs involved in transdifferentiation to neuroendocrine carcinoma were not affected. Analysis of the 4 putative neuroendocrine prostate cancer samples again showed decreased accessibilities for TFs AR, FOXA1, HOX-B13, and NKX3-1, or the increased accessibility of N-MYC (Fig. 3.20). Interestingly, decreased accessibility of REST was detected only in two of these four cases (P170_2 and P198_5; Fig. 3.20), which is consistent with reports that REST down-regulation is usually observed in 50% of neuroendocrine prostate cancer cases [81]. Only in these two cases, GLIS1 showed increased accessibility (z-scores: P170_2: 4.3; P198_5: 4.4), suggesting that this hypoxia-associated TF may be linked to REST downregulation.

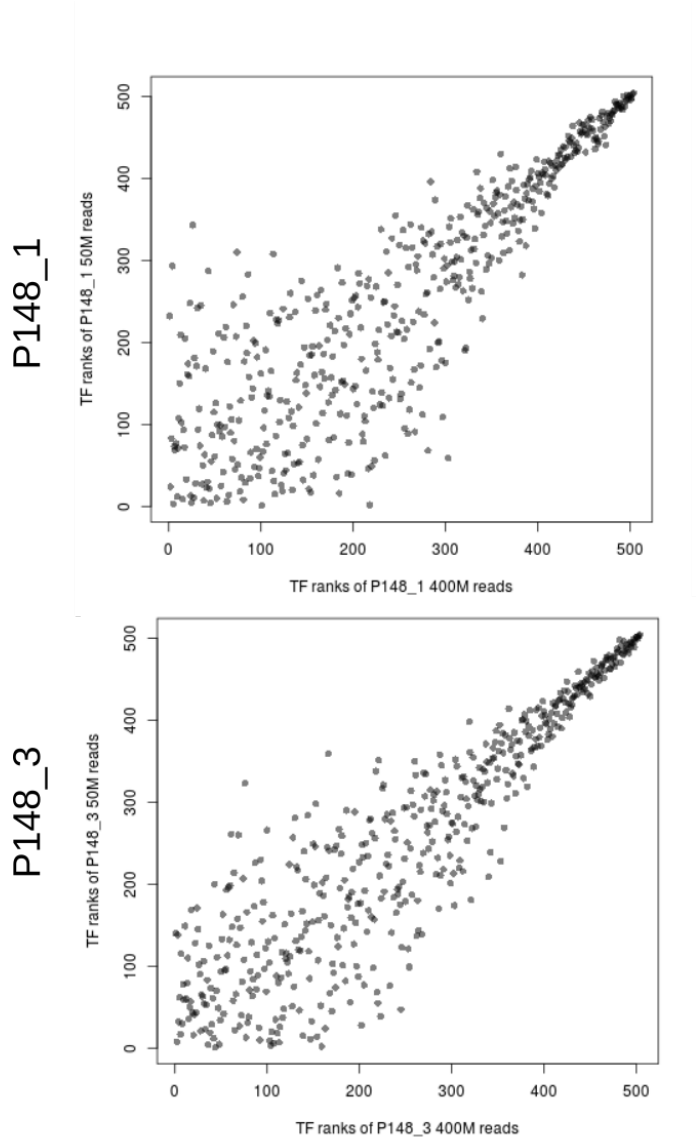

Figure 3.19: **Downsampling of P148_1 and P148_3.** Transcription factor accessibilities of downsampled data were compared to values generated from all available reads were compared for samples P148_1 and P148_3, respectively. TFs with high accessibility showed very little noise, while TFs with low accessibility were affected more strongly.

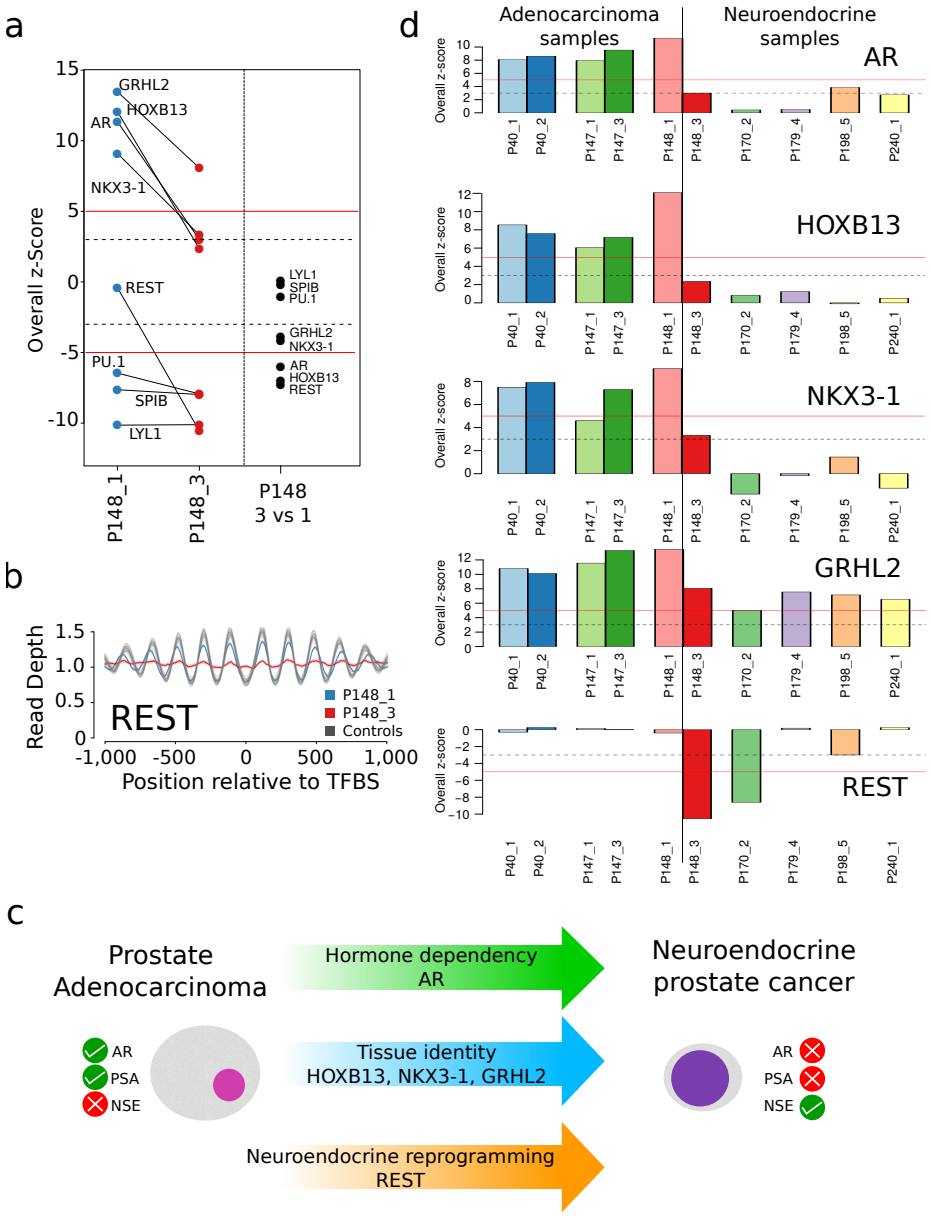

Figure 3.20: **TF changes in neuroendocrine cancers.** Transcription factor profiles change between samples of different time points in P148. In a set of prostate cancer samples with high NSE, many of the transcription factor profiles resemble the latter sample of P148. However, REST seems to be deactivated only in 2 of four samples. This figure was published in [51].

3.1.14. *Allele frequency thresholds*

In order to see, whether TF accessibility may be used in early cancer stages, a large database of early colon cancer samples (n=592) and controls provided by Freenome were used (n=177). TF accessibilities for 6 TFs that were identified as important in colorectal

cancer were analyzed and compared to estimated tumor fractions (based on ichorCNA [10]) (see Fig. 3.21a). All TFs show significantly different accessibilities in samples with a tumor fraction over 1%, while some TFs show significantly different accessibilities for all samples.

3.1.15. *Early cancer detection*

Subsequently, in a different set of colorectal cancer patients that is enriched for early stage, a logistic regression model is trained to see whether early cancer detection based on transcription factor accessibilities might be feasible. TF accessibilities are calculated for all the samples and a logistic regression model is trained. In 100 iterations data are separated into training (90%) and test sets (10%) and a class-balanced logistic regression model provided by scikit-learn is trained, while performance metrics are evaluated on the test data. Stage I cancer samples are detected with an accuracy of 71.8%, while stage II samples are detected with an accuracy of 75.5%. ROC curves for every of 100 iterations of train-test splits can be seen in Fig. 3.21b. Reported values are average values of the 100 random train-test splits.

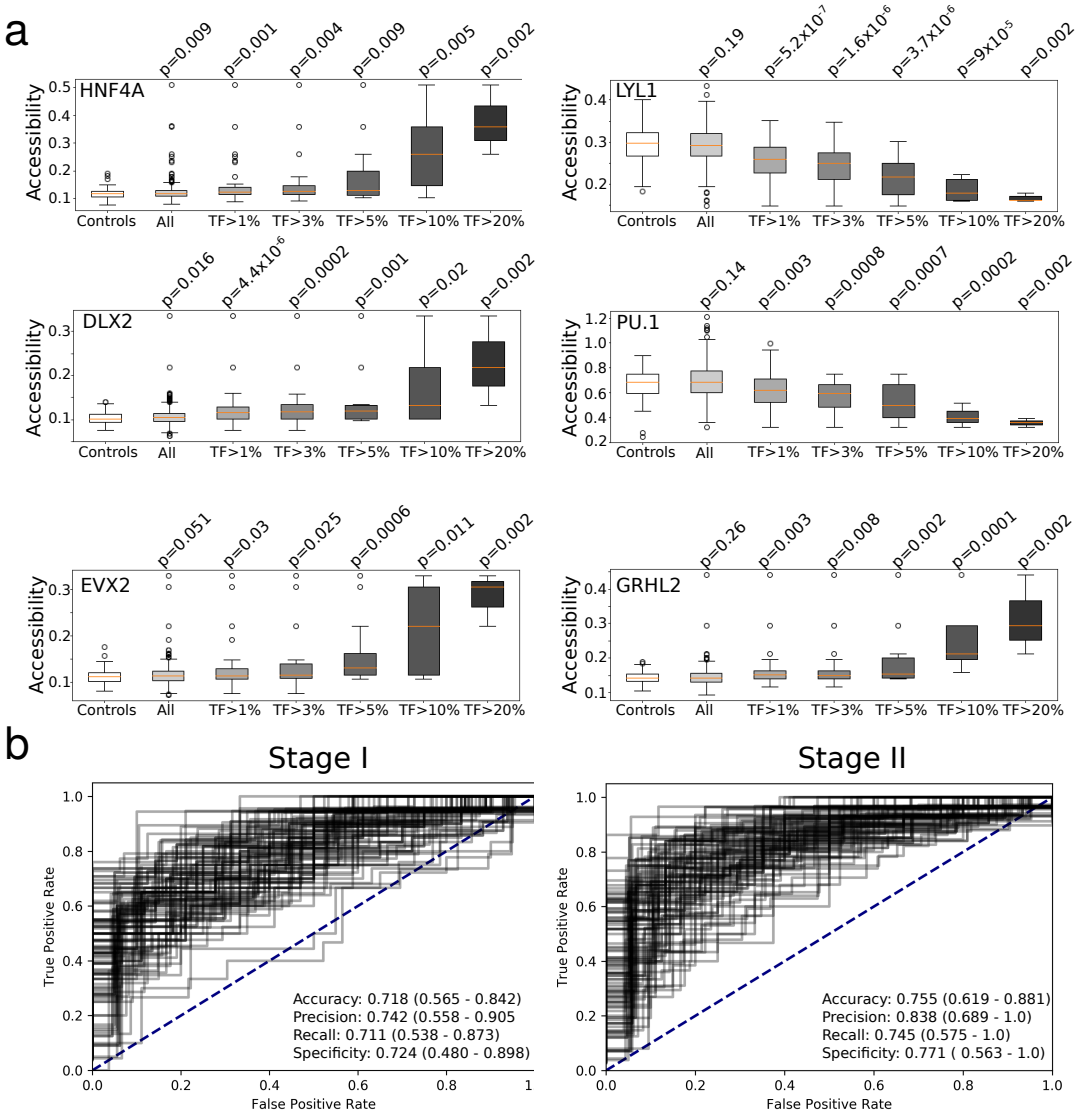

Figure 3.21: **Early detection of colorectal cancer.** a) Transcription factor accessibility of 6 TFs that were identified in colorectal cancer samples. b) Logistic regression analyses on stage I and stage II colorectal cancer samples show ability of TF accessibilities to detect cancer, even in early stages. Each ROC curve represents a single iteration in which data is randomly split into training (90%) and test sets (10%). This figure was published in [51].

3.2. Enhancers

Enhancers play an important part in gene expression through wide-range DNA-DNA contacts and transcription signal amplification. However, enhancers themselves are transcribed by Polymerase II [45] and thus chromatin needs to be regulated [85].

3.2.1. *Enhancer definitions*

Enhancer definitions were obtained from the Slidebase database which provides pre-specified enhancer tracks for cell-type specific enhancers, tissue-specific enhancers as well as other groupings of enhancers found by Andersson et al. [86].

3.2.2. *Region coverages*

In a first analysis, distribution of the average coverage at pre-specified enhancer regions was calculated and compared to a reference set with the same amount of regions and similar GC-content. Overall, coverage differences had small effect sizes and a low signal-to-noise ratio. Only a few enhancer sets show marked differences in average coverage across the regions. The largest signal is found in ubiquitous enhancers, i.e. enhancers that are expressed consistently between cell-types and tissues, respectively (see Fig. 3.22).

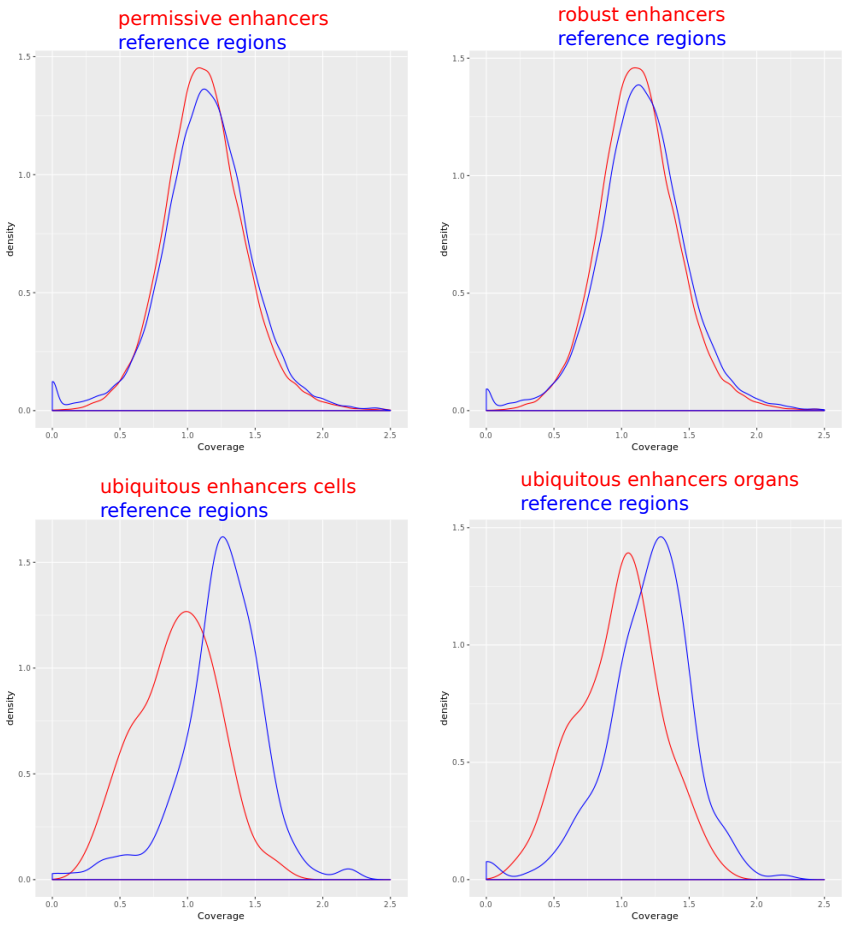

Figure 3.22: **Enhancer coverage in extensive sets.** Coverage analysis between pre-defined enhancer tracks and reference regions.

In tissue-specific enhancer regions, blood-specific enhancers show a large signal, while most other tissue-specific show no detectable difference in coverages. Unexpectedly, enhancer regions that are specifically detected in the umbilical cord show a large difference, however, for this tissue only 10 enhancers were defined (see Fig. 3.23).

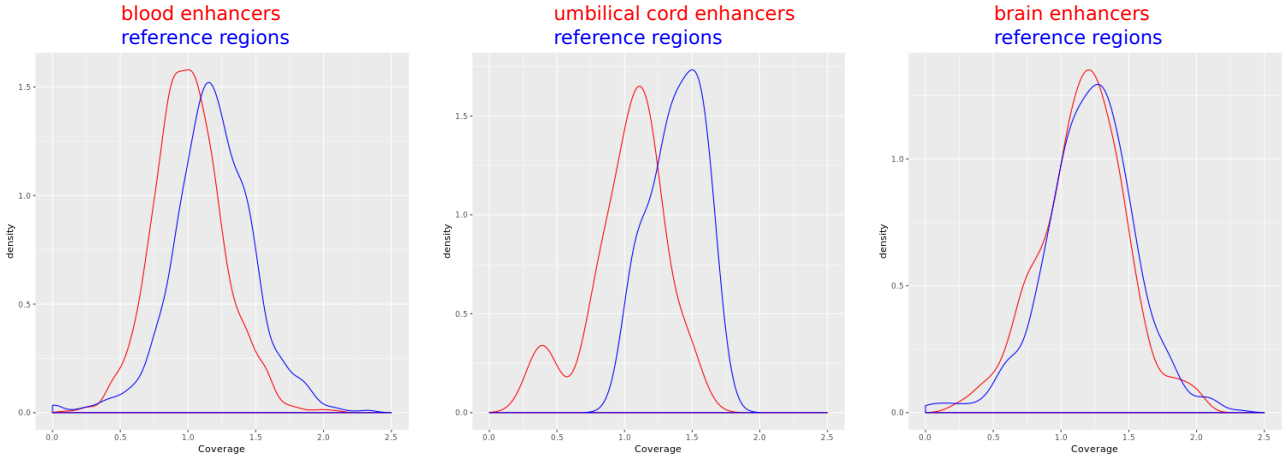

Figure 3.23: **Enhancer coverage in tissue-specific enhancer sets.** Coverage analysis between pre-defined tracks for tissue-specific enhancers (red) and reference regions (blue).

The analysis of cell-type specific enhancers shows a large signals for neutrophil granulocytes and smaller deviations in the distributions for hematological cell types. In comparison, brain cell specific enhancers show no detectable signal (see Fig. 3.24).

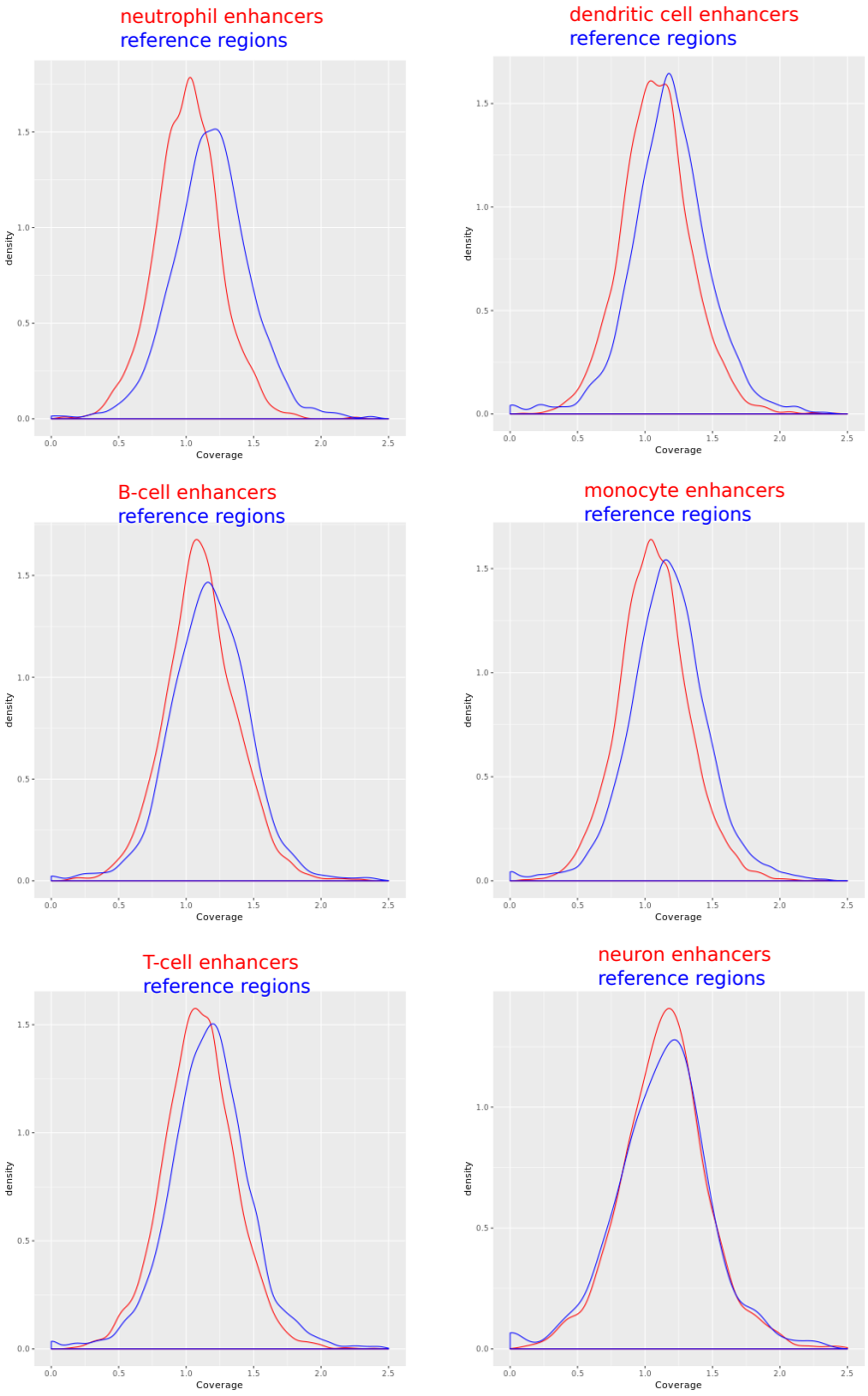

Figure 3.24: **Enhancer coverage in cell-type specific enhancer sets.** Coverage analysis between pre-defined tracks for cell-type specific enhancers and reference regions.

In order to measure the decrease in coverage between enhancer regions and reference regions, several measures were computed:

- Cohen's D effect size
- Ratio between means

- T-test

These measures allow ranking of the tissues. While blood is within the highest scoring tissues, several other non-obvious tissues reach high levels of Cohen's D (see Table 3.2). This might be due to the low number of defined enhancers for these tissues and a high background noise in coverage measurement.

Table 3.2: Tissues with highest Cohen's d for tissue-specific enhancers and bootstrapped 95% confidence intervals. While blood is among the top-scoring tissues many other tissues might be in due to high-noise levels caused by a low number of defined enhancers

Tissue	**Cohen's D**	**Lower Bound**	**Upper Bound**	**n**
umbilical cord	1.45	0.25	2.17	10
penis	0.52	-0.05	1.06	20
skin of body	0.51	-0.18	1.13	20
spinal cord	0.48	0.06	0.80	41
lymph node	0.47	-0.01	0.98	29
vagina	0.45	0.12	0.77	62
blood	0.42	0.23	0.56	1303
meninx	0.39	0.16	0.61	96
large intestine	0.31	0.09	0.50	205
skeletal muscle tissue	0.30	0.00	0.58	92

For cell types ranking yields neutrophil granulocytes first, a cell type that is supposed to be a very high contributor of cell-free DNA [14]. However, high rankings of other cell-types are more difficult to justify (see Table 3.3).

Table 3.3: Cell types with highest Cohen's d for cell-type-specific enhancers and bootstrapped 95% confidence intervals. Neutrophils show the largest signal, however, non-blood cell types also reach a high effect size.

Cell Type	**Cohen's D**	**Lower Bound**	**Upper Bound**	**n**
neutrophil	0.62	0.56	0.69	1954
smooth muscle cell of trachea	0.45	0.17	0.71	93
mesothelial cell	0.42	0.27	0.57	343
fibroblast of choroid plexus	0.42	0.20	0.62	198
fibroblast of lymphatic vessel	0.41	0.22	0.60	231
macrophage	0.41	0.33	0.48	1332
cardiac myocyte	0.37	0.13	0.61	154
chondrocyte	0.37	0.21	0.52	323
melanocyte	0.36	0.25	0.47	468
urothelial cell	0.34	0.19	0.49	377

As noise might be driving many of the high rankings, one could alternatively use the lower bound of the confidence interval which was calculated by bootstrapping for evaluations. High noise would also result in high confidence intervals. Then, in the

tissues umbilical cord enhancers would still be the highest ranking tissue, followed by blood. In the cell-specific enhancer, neutrophils would still be the highest ranking followed by macrophages and mesothelial cells.

3.2.3. *Nucleosome positioning*

As enhancer regions have been found to be transcribed bi-directionally [86] and controlled nucleosome positioning has been described, more detailed analyses in preferred nucleosome positioning was performed. To this end, coverage profiles were generated anchored to the start of the enhancer, the midpoint of the enhancer and the defined endpoint of the enhancers. Indeed, loss of coverage at the midpoints and preferential nucleosome positioning just outside the regions of transcription can be detected, however, the synchronization signal is only marginally greater than noise for neutrophil-specific enhancers (see Fig. A.1).

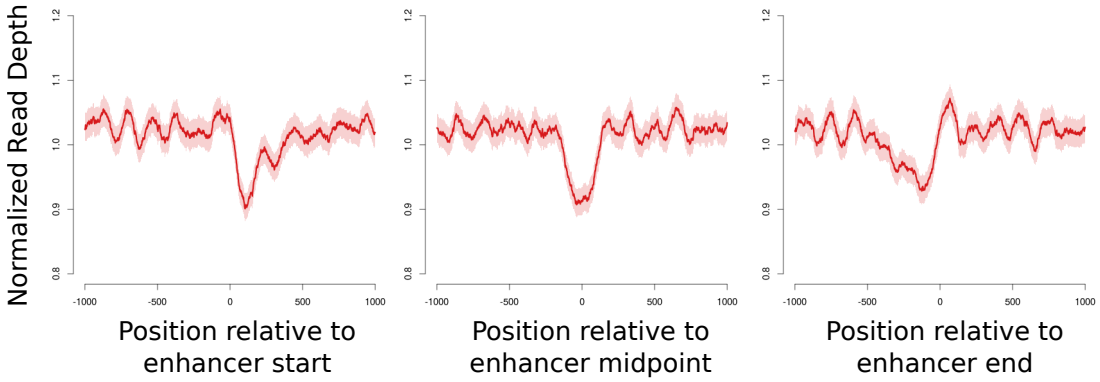

Figure 3.25: **Neutrophil enhancers coverage profile.** Coverage profile at start-points, midpoints and endpoints of neutrophil specific enhancer definitions. The y-axis has been zoomed in, in order to increase signal dynamics.

3.3. Transcription Start sites (TSS)

As clear preferential nucleosome positioning around transcription start sites has been shown already [38], coverage profiles of TSSs of housekeeping genes [87] were plotted (see Fig. 3.26.).

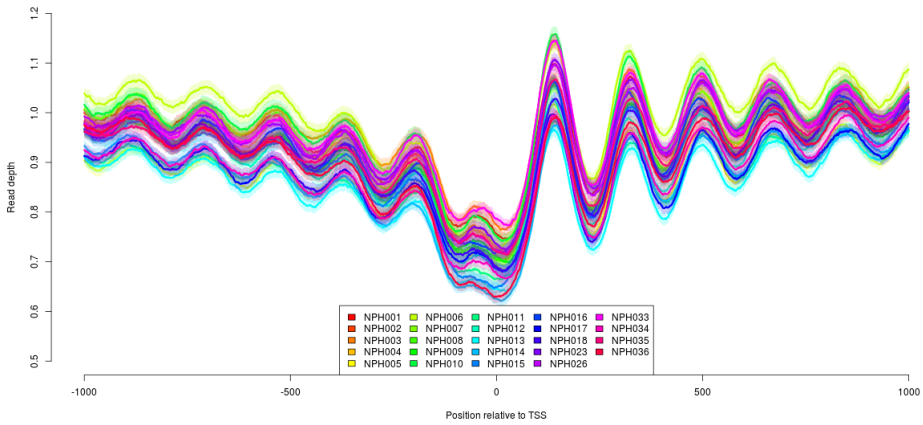

Figure 3.26: **Coverage profile at TSSs.** Coverage profile of 24 control samples at transcription start sites (TSSs) of housekeeping genes [87] recapitulates earlier findings [38]. No graph was used from the cited publications.

3.3.1. *APPRIS principal isoforms*

However, although the profile strongly resembles already published profiles, there might still be room for improving the approach. One drawback of the approach is that it relies on the RefGene database of the UCSC Genome Browser [53], which is based on data from the RefSeq project [88]. While this database gives a comprehensive view of isoforms and genes that have undergone a stringent set of validation, including manual curation, some isoforms of genes are likely more important than others. In earlier approaches every isoform of a gene was incorporated in the averaged coverage analysis.

In an attempt to catalogue the most important isoforms, Rodriguez et al. used conservation data, structural and functional information and annotated principal isoforms for 85% of protein-coding genes [89]. When constricting TSS profile coverage analyses only on principal isoforms of housekeeping genes for the 24 control samples, indeed a greater signal dynamics can be seen (see Fig. 3.27).

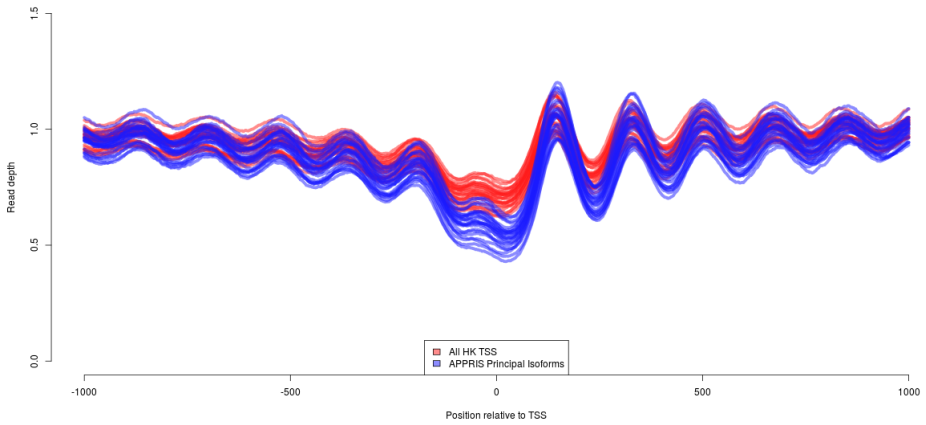

Figure 3.27: **TSS profiles for APPRIS isoforms.** Coverage profile of 24 control samples at transcription start sites (TSSs) for principal isoforms according to the APPRIS database [89] (blue) compared to the traditional approach including every isoform in RefSeq [88] (red). By removing alternative splice variants from the analyses, the nucleosome signal shows higher dynamics. No graph was used from the cited publications.

3.4. Chromatin State Prediction

3.4.1. *chromHMM states*

In an effort to predict Chromatin states from Histone modification and DNAse accessibility data, Ernst and Kellis developed chromHMM. This software uses a Hidden-Markov-Model to predict one of initially 15 states (25 states in newer versions) of chromatin using ChIP-seq data and DNAse Accessibility assays [90].

The predicted states contain:

- TssA Active TSS
- PromU Promoter Upstream TSS
- PromD1 Promoter Downstream TSS 1
- PromD2 Promoter Downstream TSS 2
- Tx5 Transcribed - 5' preferential
- Tx Strong transcription
- Tx3 Transcribed - 3' preferential
- TxWk Weak transcription
- TxReg Transcribed & regulatory (Prom/Enh)

- TxEnh5 Transcribed 5' preferential and Enh
- TxEnh3 Transcribed 3' preferential and Enh
- TxEnhW Transcribed and Weak Enhancer
- EnhA1 Active Enhancer 1
- EnhA2 Active Enhancer 2
- EnhAF Active Enhancer Flank
- EnhW1 Weak Enhancer 1
- EnhW2 Weak Enhancer 2
- EnhAc Primary H3K27ac possible Enhancer
- DNase Primary DNase
- ZNF/Rpts ZNF genes & repeats
- Het Heterochromatin
- PromP Poised Promoter Pink
- PromBiv Bivalent Promoter
- ReprPC Repressed Polycomb
- Quies Quiescent/Low

In a first attempt, coverage distributions along the predicted chromatin states of a neutrophil cell in sample NPH001 were recorded 3.28.

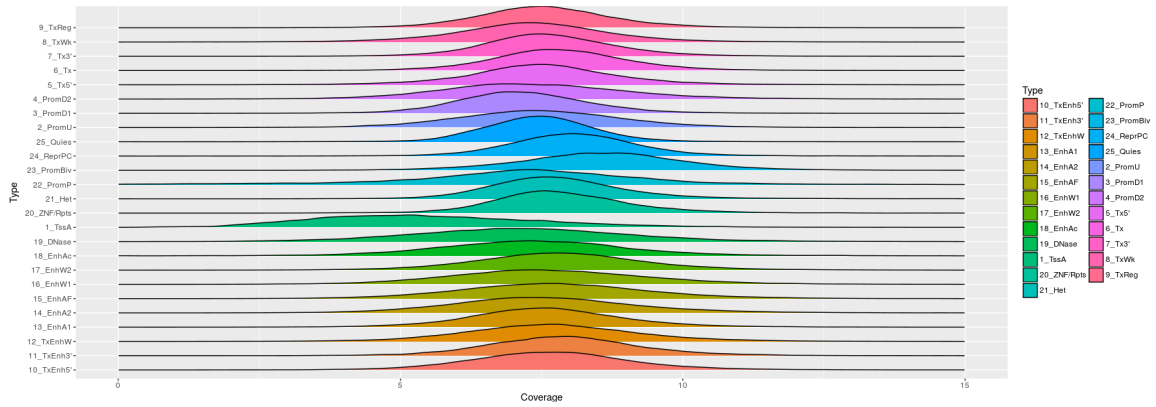

Figure 3.28: **Coverage at predicted chromHMM states.** Coverage distribution at predicted chromatin states from chromHMM. This uses the predicted states in a neutrophil cell.

3.4.2. *Fragment length distribution*

Chromatin accessibility and histone modifications are a primary signal that differentiates chromatin states. Hence, differences in fragment sizes might be seen in varying predicted chromatin states as repositioning of nucleosomes around those states have a direct impact on the available DNA fragments found in plasma.

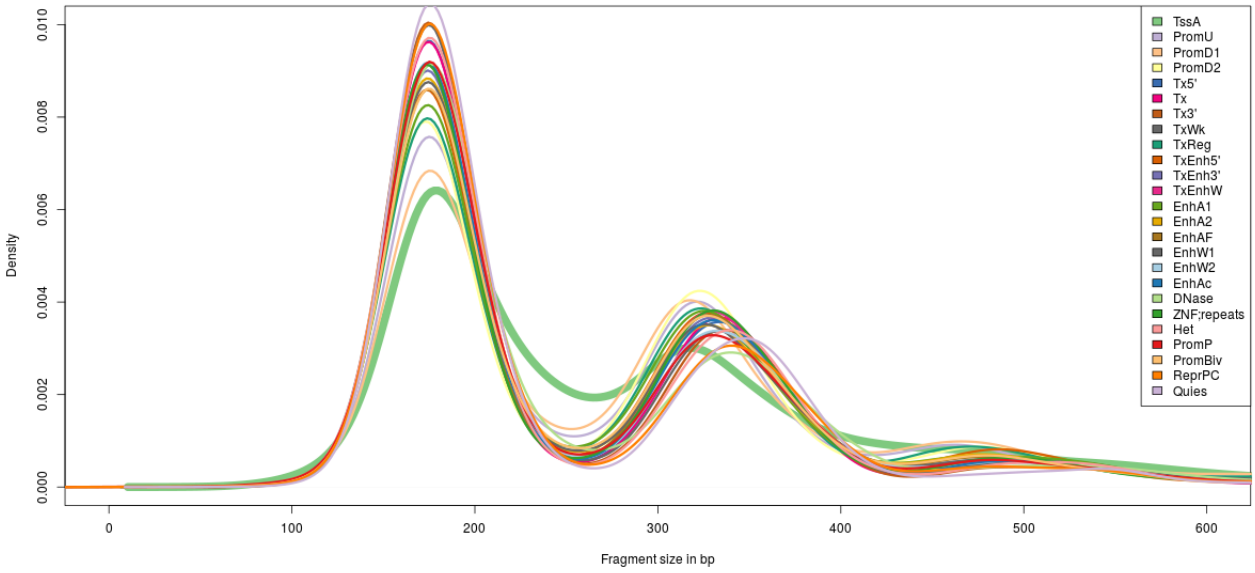

Figure 3.29: **Fragment sizes at predicted chromHMM states.** Fragment size distributions in various predicted chromHMM states show low variability with the exveption of active TSS (green and bigger line width). Both mono-nucleosomal and di-nucleosomal peaks are lower than for other predicted states.

3.5. **Histone modifications**

Since cell-free DNA is mainly associated to histones, I explored whether histone modification have an impact on fragment lengths. To this end, signal peaks of ChIP-seq targetting histone modifications in neutrophils were extracted (H2AFZ data was only available for GM12878) and the fragment length of reads covering these positions were compared. Fragment length distributions around potential histone modification sites show little aberration with the exception of H3K27 methylated histones, where mononucleosomal fragments seem to be slightly enriched (see Fig. 3.30).

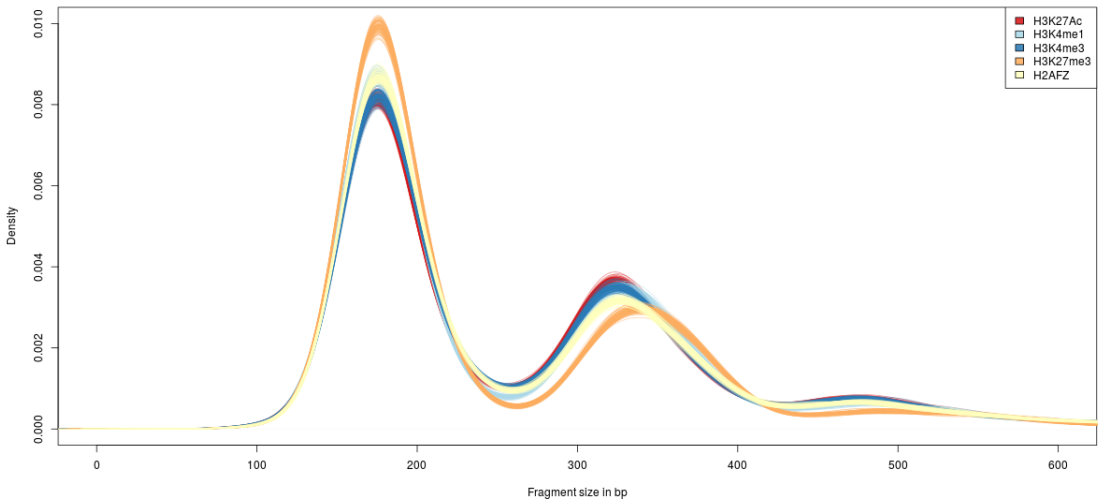

Figure 3.30: **Fragment lengths at histone modifications.** Fragment lengths distribution at potential histone modification sites derived from ChIP-seq data of neutrophil granulocytes.

3.6. G-quadruplex regions

Nucleosome occupancy in G-quadruplex predictions were performed as mentioned above. For a control sample G-quadruplex predictions with three (n=522,970), four (n=41,847) and five tetrads (n=2,204) were used. The sites show a drop in coverage directly at the predicted site and preferred nucleosome positioning in both directions. High-molecular weight DNA, however, also shows the same central drop in coverage at the site. Thus, this drop is likely not from preferred nucleosome positioning in cfDNA (see Fig. 3.31).

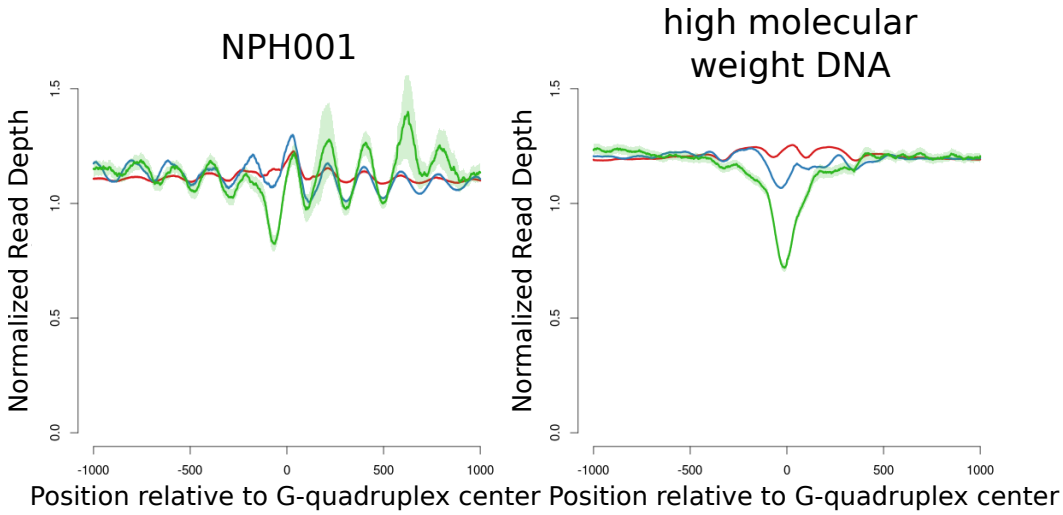

Figure 3.31: **Coverage profiles of G-Quadruplexes.** Coverage analysis around predicted G-quadruplex formation sites. Left panel: A control sample (NPH001) shows a drop in coverage at the center around predicted sites and preferred nucleosome formation extending in both directions. This is true for sites with 3 tetrads (red), 4 tetrads (blue) and 5 tetrads (green). However, at least for the central deep, high-coverage whole-genome sequencing data of high-molecular weight DNA shows the same drop at the center and thus might not constitute a functional property of cfDNA.

3.7. LOLA definitions

A wealth of large-scale experiments were conducted within the last years in order to map a plethora of epigenetic features in the human genome. Sheffield et al. [65] compiled a large database of already annotated epigenetic features in order to provide locus-overlap analyses. While many of them might not directly have an influence on nucleosome positioning, all of the available annotations were screened for non-random coverages. Altogether, 2323 annotations were provided and nucleosome positioning around the midpoints of the annotations were analyzed. Similar to the analysis of transcription factors, accessibility values were calculated and hits with the highest accessibility were examined. The top 10 annotations (see Table 3.4) are all derived from DNAse clusters that were identified using self-organizing maps from publicly-available DNAse Hypersensitivity sequencing [91]. Expectedly, the top accessible DNAse clusters all correspond to sites that contain CTCF binding motifs (dnase.genome.duke.edu/clusterDetail.php?clusterID=491).

Table 3.4: Top 10 accessibility values of annotations from the LOLA database. All of the high ranking annotations correspond to DNAse clusters [91].

Annotation	**Accessibility**
491	1.4004425505
392	1.3733626494
395	1.3643121365
247	1.3017494603
536	1.2594648993
346	1.2521795528
289	1.2388932084
295	1.2034430941
492	1.1834441422
34	1.1785349851

3.8. Differential nucleosome protection between paired cancer samples

Nucleosome synchronization has been found for many different epigenetic regulatory mechanisms, however, their analyses heavily depends on the quality of the regulatory feature definition. One possible alternative to relying on regulatory feature definition is the general analysis of differential coverage between pairs of samples. To this end the genome is divided into windows of 100bp and the mean coverage is calculated using mosdepth. For sample P148, coverage values of both timepoints were subtracted to search for unequally represented windows that may exhibit differential nucleosome protection between the samples. In total, 2098 windows display higher coverage in sample 1 (Z-score >3), whereas 16,587 windows display higher coverage in sample 3 (Z-score <-3) (see Fig. 3.32).

Figure 3.32: **Aberrant Differential windows.** Differential coverage analysis between two samples of the same patient that transdifferentiated from adenocarcinoma to a neuroendocrine carcinoma results in 2,098 windows with higher coverage in sample 1 (upper plot, red) and 16,587 windows with higher coverage in sample 3 (lower plot, blue), respectively. Here, the distribution of differentially covered windows across the genome is depicted.

Furthermore, LOLA was performed to screen windows with differential coverage for overlaps with already known epigenetic features. To this end, LOLA uses a collection of various databases that contain defined epigenetic features and runs Fisher's exact tests on

those. Only a single overlap gives a statistically significant overlap after multiple-testing correction: CpG-islands (as provided by the UCSC genome browser, $p=10^{-8.3}$, $q=10^{-2.53}$, see Tables 3.5 and Table 3.6).

Table 3.5: Results of LOLA analysis of windows that display higher coverage in sample 1 as measured by Z-score analysis.

Filename	**logOddsRatio**	**pValueLog**	**qValue**
cpgIslandExt.bed	2.7182272373 8.2956428562	2.53120400023894E-08	
vistaEnhancers.bed	0.6257402088	0.0981304754	1
vistaEnhancers_colNames.bed	0.6257402088	0.0981304754	1
laminB1Lads.bed	0.1925050437	0	1
numtSAssembled.bed	0	0	1

Table 3.6: Results of LOLA analysis of windows that display higher coverage in sample 3 as measured by Z-score analysis.

Filename	**logOddsRatio**	**pValueLog**	**qValue**
numtSAssembled.bed	0.4210156038	0.042396413	1
cpgIslandExt.bed	0.7131289314	8.4363905517206E-05	1
vistaEnhancers.bed	0.0790778766	0.000001409	1
vistaEnhancers_colNames.bed	0.0790778766	0.000001409	1
laminB1Lads.bed	0.3214436303	0	1

4. *Discussion*

In this study, the effect of various epigenetic mechanisms on fragmentation patterns and representation of fragments in cfDNA was explored.

4.1. **Transcription factors**

In general, the analysis of transcription factor binding sites give the cleanest signal of preferential nucleosome positioning throughout the presented work.

4.1.1. *Evaluation of binding sites*

Coverage patterns are robust throughout all the control samples analyzed. Moreover, the pattern is unique to cfDNA and comparable to MNase-seq data. However, some profiles showed deviations in the amplitude of nucleosome synchronization. Since MNase-seq data comes from isolated and immortalized B-cells, some deviations to cfDNA, which is a mixture of several cell types [15], is expected. Moreover, transcription factor coverage profiles are comparable to already published transcription factor profiles from MNase-seq [92] and cfDNA [30]. Overlaps between transcription factors are rare, however, transcription factors from the same protein family show the largest overlap, which is in line with previous reports [92].

4.1.2. *Fragment sizes*

Fragment size distributions around CTCF binding sites confirm strict nucleosome positioning. Nucleosome positions show an enrichment in mono and di-nucleosomal fragments whereas positions between nucleosomes show intermediate fragment sizes. Especially position close to the center of CTCF binding show a special enrichment of intermediate

fragment sizes which might be due to the protection of both a nucleosome and CTCF or simply a partially digested longer fragment with only one intact nucleosome since the second nucleosome was removed from CTCF. This distribution of fragments is in line with earlier reports of preferential nucleosome positioning around transcription factor binding sites derived from MNase-seq [32].

4.1.3. *Binding site sizes*

Analysis of binding site sizes shows a large variability in the central binding site size. There was no obvious correlation between DNA binding domain and the size. However, only 55 transcription factors could be assessed, since reasonable estimates for a central lack of nucleosomes can only be given, when a certain degree of nucleosome synchronization is present. Thus a multitude of transcription factors that may elucidate biological relationships between size and any other biological property can't be used. Still, a significant difference between TFs that have a large binding site and TFs that have a smaller binding site has been found, where the former shows a higher proportion of CpG-island associated binding sites.

4.1.4. *CTCF*

CTCF shows the largest signal of nucleosome synchronization in the genome, compared to every other signal that has been screened in this thesis. It binds abundantly in the genome and apart from being a transcription factor also regulates the 3-dimensional structure of genomic regions [46]. Moreover, CTCF has a long DNA binding residence time, which may also increase a signal captured in cfDNA [93]. Several TFs show overlaps with CTCF (e.g. CTCFL, ZBTB2). These also show high nucleosome synchronization, however, it is hard to distinguish the effects of those TFs when a large portion of the signal might be driven by CTCF, since the coverage profile is generated by all TFs acting in cfDNA simultaneously. CTCFL share nearly identical binding domains, however it has been found to only be active in germ cells [70]. However, no report was found that would show a connection between ZBTB2 and CTCF. These are examples of transcription factors that share the same binding sequence, that the presented method cannot distinguish very well. Thus, special care must be taken in interpreting single TF signatures.

4.1.5. *Measuring accessibility*

Several measures of accessibility are introduced. While it is reasonable to use all defined transcription factor binding sites, as this would decrease noise of random sampling, the correction for the number of sites in the LOESS model inherently reduces dynamic variation. Thus, using an equal number of binding sites per TF might be beneficial since the background noise should be approximately similar. Indeed, the performance of the 1,000-msTFBSs sites is better than the other models. Wavelet analysis performs worst, which might be due to the assumption that only a single signal with a given period contributes to the overall profile. This might be oversimplifying the real situation.

One drawback of limiting the analysis to a 1,000 binding sites per TF is that a much smoother coverage signal could be obtained when more binding sites can be used. Moreover, using exactly 1,000 sites might lead to inclusion of binding sites with less experimental support in TFs that bind less often within the genome. While the analysis of single binding sites would be beneficial and would probably give a lot more granular information, obtaining single values per TF are easier to interpret.

One possible improvement of the methodology could lie in using more uniform binding sites. For example, binding sites in the vicinity of transcription start sites will likely lead to a very different nucleosome profile as many other transcription factors and proteins will also bind to the genome exert their influence on nucleosomal positioning.

In principle the method of detecting TF accessibility is not dissimilar to methods, developed for TF footprinting in DNase-seq and ATAC-seq data, however, since no direct effect of TF binding can be observed in cfDNA, but rather less direct nucleosome positioning, these methods are not directly applicable [94].

4.1.6. *Signal deconvolution*

A more general model of analyzing several aspects of the coverage patterns were performed in the signal deconvolution step. While the reconstruction of the individual parts of the signal work surprisingly well (see Fig. 3.12), it fails to capture the true signal in some transcription factors. Especially the measurement of the "central gap" in TF binding relies on a pronounced nucleosome synchronization and thus might be the most sensible

part in the deconvolution process.

4.1.7. *Tissue-specificity of TFs*

By analyzing ATAC-seq data, pooled cancer cfDNA samples and individual cancer cfDNA samples, the discovery of tissue-specific TFs is supported by several orthogonal measurements. While the TFs detected in prostate cancer are well known in the literature and thus are supported even more [95, 78], also the TFs that were detected in colorectal cancer but had no known function in that tissue types have recently gained support from the literature [96, 97]. In breast, cancer no tissue-specific TF could be identified. Only TFs that show elevated accessibilities in all (epithelial) cancer types (GRH-L2, FOXA1) showed also higher accessibilities. This might be due to the mixture of different sub-types which might dilute subtype-specific TF signals. Another possibility could also be that no single TF is active in breast tissue or no TF that would be tissue specific leads to changes in nucleosome positioning.

4.1.8. *Single-sample WGS analyses*

In single samples, deviations in accessibility from healthy control sample was investigated by using a Z-score approach. While this is a simple approach it might not necessarily be the most sensitive approach to detecting altered TF accessibility. However, it's simplicity makes it very intuitive and the resulting TFs make sense. In prostate cancer patients the predominant deviations from healthy controls are TFs that belong to the Androgen receptor signaling cascade. These include AR, HOXB13 and NKX3-1. In the samples of the colorectal cancer patient TFs that were previously identified in the pooled cancer samples and the ATAC-seq data sets, are again found to be different from controls. Again, in single breast cancer samples no tissue-specific TF was identified

4.1.9. *Differences in accessibility in paired-cancer samples*

In order to see whether tumor evolution can be traced in subsequent samples from the same patient, differential analyses of TF accessibility was performed in a set of 4 patients (3 prostate cancer and 1 colon cancer patients). While all of those patients displayed some differences in somatic copy number changes from shallow whole-genome sequencing (see

Appendix), only patient P148 showed a very different pattern. Moreover, clinically this patient progressed from an adenocarcinoma to a neuroendocrine tumor of the prostate, marked by high NSE levels and reduced PSA levels. Also, the amplification of the AR receptor on chromosome X was lost, which could mean that proliferating cells of this tumor no longer needed the Androgen receptor based growth signal. Thus, we only find large differences in transcription factor accessibility in this patient. All of the changes are in line with the assumption that the primary tumor became independent of the androgen receptor signaling cascade: Accessibility of the Androgen receptor itself and the TFs that are associated to this (NKX3-1, HOXB13) showed markedly reduced accessibility. Moreover, accessibility of REST was reduced by a large margin. This TF represses genes that are important for neurologic development and thus also falls in line with the other observations [84]. The other three patients showed no detectable difference in TF accessibility. Patient P40 developed a new amplification of the Androgen receptor under anti-hormonal therapy, Patient P147 developed a new focal amplification of the RET oncogene and patient C2 lost an amplification of the KRAS oncogene. Although all of these SCNA changes under therapy can be considered small in terms of genomic size, we were initially suspecting to see changes in the underlying biological mechanisms. Since, none of these changes could be detected, the events probably represent mechanisms of the tumor cells to maintain homeostasis under the pressure of tumor treatment, however this remains to be proven.

4.1.10. *Allele frequency thresholds*

As ctDNA in earlier cancer stages show a lot lower levels than samples of patients with metastasized disease, allele frequency thresholds of detection were estimated in samples of a separate cohort of patients with colorectal cancer. This cohort contains a lot of samples in earlier stages. IchorCNA was used to estimate tumor fractions in this samples in order to see the critical amount of tumor fraction needed to detect alterations in TF binding patterns [10]. Even at 1% tumor fraction, TF accessibility differences from cancer samples to control samples are statistically significant, although effect sizes are low. This holds true for both, TFs that show higher accessibility in CRC (HNF4A, DLX2, EVX2, GRHL2), as well as TFs that show lower accessibility in CRC (Lyl1, PU.1).

4.1.11. *Early cancer detection*

In a separate of samples that is enriched for early colorectal cancer stages, logistic regression enables early cancer detection at a reasonable performance. While in itself performance is lower than has been reported for other analyses [18, 17], transcription factor analyses might be complementary.

4.2. **Enhancers**

4.2.1. *Enhancer definitions*

Enhancers are defined by CAGE sequencing data of the FANTOM5 project, where bidirectional transcripts define an enhancer [62]. The advantage of using this approach rather than chromatin accessibility directly (via DNase-seq or ATAC-seq), is that the ends are better characterized. CAGE sequencing itself should be able to map 5' ends of any given RNA fragment very precisely [86], but whether nucleosome synchronization relative to transcription start sites of enhancer RNAs are very specific is not known.

4.2.2. *Region coverages*

Analyses of coverage distributions show a signal for cells that typically contribute DNA into the bloodstream. However, the effect size of these differences is low and variation of per-region coverage values is high. Thus, enhancer analyses based on region coverages might work well in cases of high-tutor fraction samples, where signals are expected to be high. However, analysis of activation patterns in single enhancers would be more favorable, but the low signal strength is prohibitive in these analyses.

4.2.3. *Nucleosome positioning*

CAGE-sequencing data seems to correctly identify the Enhancer per se as enhancers of cell-types that are expected to contribute a lot of DNA to the whole cfDNA fraction exhibit strong nucleosome synchronization. However the start point might not be ideally pinpointed, which makes the analysis of collapsed regions to identify synchronized nucleosomes hard. Moreover, while there are reports that enhancers show nucleosome positioning [98], the data presented here only show marginal synchronized nucleosomes when

collapsed along enhancer starts, ends, or midpoints. Even at cell types that were shown to represent a large fraction within cfDNA (i.e. neutrophil granulocytes) [15], nucleosome positioning is only marginally reduced.

4.3. Transcription start sites (TSS)

While analyses of cfDNA coverage around transcription start sites in order to infer gene expression already show promising results [38], there undoubtedly is room for improvement. Especially the linear relationship between TSS coverage patterns and relative gene expression have not been delineated satisfactorily. Moreover, gene expression prediction only was successful at substantial tumor fractions. Here, the selective use of isoforms that have sufficient experimental is analyzed as one way to improve upon the signal strength of nucleosome positioning.

4.3.1. *APPRIS principal isoforms*

One of the possible ways of increasing performance of a gene expression classifier is in selection of "important" isoforms of genes [89]. RefSeq was designed in order to comprehensively map many isoforms of genes. However, some isoforms are preferred and transcribed more readily in individual tissues. Using RNA-seq data the authors of the APPRIS project aim to identify "principle" isoforms of genes. The data shown here shows that subsetting the analyses to principal isoforms leads to an increase in signal in housekeeping genes. Furthermore, the analysis of single transcription start sites might yield more granular information than simply averaging over all available, or even principal isoforms for any given gene.

4.4. Chromatin State Prediction

Chromatin state prediction from a plethora of individual chromatin marks has been shown to characterize genomic regions into distinct classes [90]. While the reality of genomic organization likely is more complex than purported by any chromatin segmentation algorithm, the most important information is still kept and the complexity of various differing epigenetic marks across the whole genome are reasonably diminished.

4.4.1. *Coverage at chromHMM states*

On a coverage level, chromHMM predictions from a neutrophil cell (as the cell that should contribute the largest fraction of cfDNA in a healthy sample) show very similar mean coverages. The variance in the distribution is most readily explained by varying number and sizes of the single predictions. Only transcription start sites show a markedly different distribution, having lower coverage than other states.

4.4.2. *Fragment length distribution*

Fragment length distributions at predicted chromHMM states show also very similar fragment lengths again, with transcription start sites showing an enrichment of fragments between mono- and di-nucleosomal reads. The low difference between chromatin states does not warrant further analyses, however, the enrichment if intermediate fragment lengths at TSS might be an option to increase signal in TSS analyses.

4.5. **Histone modifications**

Again, fragment length distribution of fragments that are supposedly derived from histones with modifications look very similar with only slight deviations from one another. The only modification that seems to lead to an enrichment of mono-nucleosomal reads is H3K27me3. H3K27me3 is reported to be associated with the Polycomb repressive complex 2 which leads to repression of down-stream gene transcription [99].

4.6. **G-quadruplex regions**

G-quadruplex regions are abundant in the genome, however, their influence on nucleosome formation is unclear [48]. Moreover, the motif that facilitates G-quadruplex formation is a repetitive sequence that is hard to align with short reads. Since cfDNA molecules are inherently short, this can't be overcome easily using a different technology. The repetitive sequence is hard to map and has a high GC-content, two factors that make coverage analyses impossible. This is exemplified in the analysis of high-molecular weight DNA that shows similar features as cfDNA samples. Hence, G-quadruplexes might be a source of coverage variability in cfDNA data, but using it diagnostically will be very challenging.

4.7. LOLA definitions

For a more general, less targeted overview of possible differences in chromatin accessibility by various epigenetic regulators, the databases that were compiled in the LOLA software distribution are tested for nucleosome synchronization. DNase Hypersensitivity clusters that contain the CTCF binding motif showed the highest accessibility, possibly due to the properties of CTCF as a mediator of topological domains the genome. Other than transcription factor binding sites or DNase Hypersensitivity clusters, no other type of regulatory annotation shows significant nucleosome positioning. Since the region databases were not designed to capture nucleosome synchronization and are too large to capture chromatin synchronization, chromatin synchronization is difficult to assess. Here, the midpoint of any given region was used to anchor the coverage signal and overlap averaged patterns. Even small deviations in the definition of these anchor points will smooth out signals of preferential nucleosome positioning, even when strict regulation would be present. Thus, the use of the GTRD database in the analysis of transcription factor binding sites is a lot better suited for these kind of analysis, as averaging of coverage signals is performed based on ChIP-seq signal peaks at a single base pair.

4.8. Differential nucleosome protection between paired cancer samples

A different way of looking at differential chromatin accessibility in the prostate cancer patient is too first identify regions of differential coverage and try to interpret them later on. Copy-number alterations, GC bias and other external factors can influence these results however, since many processes have influence on the final coverage. 2,098 and 16,587 100bp windows are found with differential coverage and the distribution of the windows seems random enough that copy-numbers cannot reasonably be assumed to be the primary signal source. Interpretation with LOLA [65] results in a single epigenetic regulator that seems to be the main source of differential coverage: CpG islands.

4.9. Conclusion

In this study, the influence of epigenetic regulation on cfDNA is examined by looking at whole-genome sequencing data and correlate various properties of the data to known

regulators of chromatin accessibility. While many potential correlations are found through a large number of regulators, only a few seem to leave a trace in cfDNA that seems to be worth exploiting in liquid biopsy-based diagnostics. The direct influence of transcription factors on nucleosome synchronization and chromatin accessibility seems to be a very worthwhile approach of deciphering tumor biology through non-invasive ways. Especially the possibility of repeated measures makes this approach attractive to get a glimpse on changing tumor biology over the course of a disease.

References

[1] Heitzer E, Haque IS et al. **Current and future perspectives of liquid biopsies in genomics-driven oncology**. *Nature Reviews Genetics* 2018, **20**.

[2] Mandel P and Métais P. **Les acides nucléiques du plasma sanguin chez l'homme**. *Comptes rendus des séances de la Société de biologie et de ses filiales* 1948, **142**: 3–4.

[3] Leon SA, Shapiro B et al. **Free DNA in the serum of cancer patients and the effect of therapy**. *Cancer Research* 1977, **37**: 646–650.

[4] Anker P, Mulcahy H et al. **Detection of circulating tumour DNA in the blood (plasma/serum) of cancer patients**. *Cancer Metastasis Rev.* 1999, **18**.

[5] Heitzer E, Ulz P et al. **Circulating Tumor DNA as a Liquid Biopsy for Cancer**. *Clinical Chemistry* 2015, **61**: 112–123.

[6] Sperling AS,Gibson CJ et al. **The genetics of myelodysplastic syndrome: from clonal hematopoiesis to secondary leukemia**. *Nature Reviews Cancer* 2017, **17**.

[7] Martincorena I,Fowler JC et al. **Somatic mutant clones colonize the human esophagus with age**. *Science* 2018, **362**.

[8] Quail MA, Kozarewa I et al. **A large genome centre's improvements to the Illumina sequencing system**. *Nature Methods* 2008, **5**.

[9] Newman AM, Bratman SV et al. **An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage**. *Nature Medicine* 2014, **20**.

[10] Adalsteinsson V,Ha G et al. **Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors**. *Nature Communications* 2017, **8**.

[11] Heitzer E, Ulz P et al. **Tumor-associated copy number changes in the circulation of patients with prostate cancer identified through whole-genome sequencing**. *Genome Medicine* 2013, **5**.

[12] Ulz P, Belic J et al. **Whole-genome plasma sequencing reveals focal amplifications as a driving force in metastatic prostate cancer**. *Nature Communications* 2016, **7**.

[13] Chan KC, Jiang P et al. **Noninvasive detection of cancer-associated genome-wide hypomethylation and copy number aberrations by plasma DNA bisulfite sequencing**. *Proceedings of the National Academy of Sciences* 2013, **110**: 18761–8.

[14] Sun K, Jiang P, et al. **Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments**. *Proceedings of the National Academy of Sciences* 2015: E5503–E5512.

[15] Moss J, Magenheim J et al. **Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease**. *Nature Communications* 2018, **9**.

[16] Shen SY,Singhania R et al. **Sensitive tumour detection and classification using plasma cell-free DNA methylomes**. *Nature* 2018, **563**.

[17] Cohen JD,Li L et al. **Detection and localization of surgically resectable cancers with a multi-analyte blood test**. *Science* 2018, **359**.

[18] Mouliere F, Chandrananda D et al. **Enhanced detection of circulating tumor DNA by fragment size analysis**. *Science Translational Medicine* 2018, **10**.

[19] Tarazona N, Gimeno-Valiente F et al. **Targeted next-generation sequencing of circulating-tumor DNA for tracking minimal residual disease in localized colon cancer**. *Annals of Oncology* 2019.

[20] Kurtz DM,Scherer F et al. **Circulating Tumor DNA Measurements As Early Outcome Predictors in Diffuse Large B-Cell Lymphoma**. *Journal of Clinical Oncology* 2018, **36**.

[21] Mohan S, Heitzer E et al. **Changes in Colorectal Carcinoma Genomes under Anti-EGFR Therapy Identified by Whole-Genome Plasma DNA Sequencing**. *PLoS Genetics* 2014.

[22] Abbosh C,Birkbak NJ et al. **Phylogenetic ctDNA analysis depicts early stage lung cancer evolution**. *Nature* 2017, **545**.

[23] Heitzer E, Auer M et al. **Establishment of tumor-specific copy number alterations from plasma DNA of patients with cancer**. *International Journal of Cancer* 2013, **133**: 346–357.

[24] Fan HC, Blumenfeld YJ et al. **Analysis of the Size Distributions of Fetal and Maternal Cell-Free DNA by Paired-End Sequencing**. *Clinical Chemistry* 2010, **56**: 1279–1286.

[25] Mouliere F and Rosenfeld N. **Circulating tumor-derived DNA is shorter than somatic DNA in plasma**. *Proceedings of the National Academy of Sciences* 2015, **112**: 3178–3179.

[26] Moser T, Ulz P et al. **Single-Stranded DNA Library Preparation Does Not Preferentially Enrich Circulating Tumor DNA**. *Clinical Chemistry* 2017, **63**.

[27] Jiang P, Chan CW et al. **Lengthening and shortening of plasma DNA in hepatocellular carcinoma patients**. *Proceedings of the National Academy of Sciences* 2015, **112**: 1317–1325.

[28] Chan KC, Zhang J et al. **Size Distributions of Maternal and Fetal DNA in Maternal Plasma**. *Clinical Chemistry* 2004, **50**: 88–92.

[29] Kang S, Li Q et al. **CancerLocator: non-invasive cancer diagnosis and tissue-of-origin prediction using methylation profiles of cell-free DNA**. *Genome Biology* 2017, **18**.

[30] Snyder MW, Kircher M et al. **Cell-free DNA Comprises an In Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin**. *Cell* 2016, **164**: 57–68.

[31] Richmod TJ and Davey CA. **The structure of DNA in the nucleosome core**. *Nature* 2003, **423**: 145–150.

[32] Zentner GE and Henikoff S. **Surveying the epigenomic landscape, one base at a time**. *Genome Biology* 2012, **13**: 250.

[33] Pich O, Muinos F et al. **Somatic and Germline Mutation Periodicity Follow the Orientation of the DNA Minor Groove around Nucleosomes**. *Cell* 2018, **175**: 1074–1087.

[34] Kundaje A, Kyriazopoulou-Panagiotopoulou S et al. **Ubiquitous heterogeneity and asymmetry of the chromatin environment at regulatory elements**. *Genome Research* 2012, **22**: 1735–47.

[35] Chandrananda D, Thorne NP et al. **High-resolution characterization of sequence signatures due to non-random cleavage of cell-free DNA**. *BMC Medical Genomics* 2015, **8**: 29.

[36] Ivanov M, Baranova A et al. **Non-random fragmentation patterns in circulating cell-free DNA reflect epigenetic regulation**. *BMC Genomics* 2015, **16**: S1.

[37] Vishwanath RI. **Nucleosome positioning: bringing order to the eukaryotic genome**. *Trends in Cell Biology* 2012, **22**: 250–256.

[38] Ulz P, Thallinger GG et al. **Inferring expressed genes by whole-genome sequencing of plasma DNA**. *Nature Genetics* 2016, **48**: 1273–1278.

[39] Venkatesh S and Workman JL. **Histone exchange, chromatin structure and the regulation of transcription**. *Nature Reviews Molecular Cell Biology* 2015, **16**: 178–189.

[40] Lambert SA, Jolma A et al. **The Human Transcription factors**. *Cell* 2018, **172**: 650–665.

[41] Adams CC and Workman JL. **Binding of disparate transcriptional activators to nucleosomal DNA is inherently cooperative**. *Molecular Cell Biology* 1995, **15**: 1405–1421.

[42] Iwafuchi-Doi M, Donahue G et al. **The pioneer transcription factor FoxA maintains an accessible nucleosome configuration at enhancers for tissue-specific gene activation**. *Molecular Cell* 2016, **62**: 79–91.

[43] Rippe K, Schrader A et al. **DNA sequence- and conformation-directed positioning of nucleosomes by chromatin-remodeling complexes**. *Proceedings of the National Academy of Sciences* 2007, **104**: 15635–15640.

[44] Wang J, Zhuang J et al. **Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors**. *Genome Research* 2012, **22**: 1798–1812.

[45] Smith I and Shilatifard A. **Enhancer biology and enhanceropathies**. *Nucleic Structural & Molecular Biology* 2014, **21**: 210–219.

[46] Ong CT and Corces VG. **CTCF: an architectural protein bridging genome topology and function**. *Nucleic Structural & Molecular Biology* 2014, **15**: 234–246.

[47] Todd AK, Johnston M et al. **Highly prevalent putative quadruplex sequence motifs in human DNA**. *Nucleic Acids Research* 2005, **33**: 2901–2907.

[48] Hänsel-Hertsch R, Beraldi D et al. **G-quadruplex structures mark human regulatory chromatin**. *Nature Genetics* 2016, **48**: 1267–1271.

[49] Berger SL. **Histone modifications in transcriptional regulation**. *Curries Opin Genet Dev* 2002, **2**: 142–148.

[50] Sadeh R,Falkoff G et al. **ChIP-seq of plasma cell-free nucleosomes identifies cell-of-origin gene expression programs**. *bioRxiv* 2019.

[51] Ulz P, Perakis S et al. **Inference of transcription factor binding from cell-free DNA enables tumor subtype prediction and early detection**. *Nature Communications* 2019, **10**.

[52] Yevshin I, Sharipov R et al. **GTRD: a database of transcription factor binding sites identified by ChIP-seq experiments**. *Nucleic Acids Research* 2017, **45**: D61–D67.

[53] Kent WJ, Sugnet CW et al. **The human genome browser at UCSC**. *Genome Research* 2002, **12**: 996–1006.

[54] Quinlan AR and Hall IM. **BEDTools: a flexible suite of utilities for comparing genomic features**. *Bioinformatics* 2010, **26**: 841–842.

[55] Li H, Handsaker B et al. **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics* 2009, **16**: 2078–2079.

[56] Li H and Durbin R. **Fast and accurate short read alignment with Burrows–Wheeler transform**. *Bioinformatics* 2009, **14**: 1754–1760.

[57] Picard. `http://picard.sourceforge.net`. [Online; accessed 19-July-2015]. 2015.

[58] TCGA ATAC-seq data. `https://api.gdc.cancer.gov/data/f0094e76-4a80-4ee1-9ad0-8ffb94aff5f7`. [Online; accessed 23-August-2018]. 2018.

[59] Pedersen BS and Quinlan AR. **Mosdepth: quick coverage calculation for genomes and exomes**. *Bioinformatics* 2018, **34**: 867–868.

[60] Pedregosa F, Varoquaux G et al. **Scikit-learn: Machine Learning in Python**. *Journal of Machine Learning Research* 2011, **12**: 2825–2830.

[61] Lizio M, Harshbarger J et al. **Gateways to the FANTOM5 promoter level mammalian expression atlas**. *Genome Biology* 2015, **16**: 22.

[62] SlideBase database of Human Enhancers. `http://slidebase.binf.ku.dk/human_enhancers/presets`. [Online; accessed 5-September-2018]. 2018.

[63] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2017. URL: `https://www.R-project.org/`.

[64] Hon J, Martinek, T et al. **pqsfinder: an exhaustive and imperfection-tolerant search tool for potential quadruplex-forming sequences in R**. *Bioinformatics* 2017, **33**: 3373–3379.

[65] Sheffield NC and Bock C. **LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor**. *Bioinformatics* 2016, **22**: 568–576.

[66] LOLA databases. `http://databio.org/lola/`. [Online; accessed 1-December-2018]. 2018.

[67] chromHMM state predicitons. `https://egg2.wustl.edu/roadmap/web_portal/imputed.html`. [Online; accessed 1-December-2018]. 2018.

[68] chromHMM state predicitons for a neutrophil granulocyte. `https://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModels/imputed12marks/jointModel/final/E030_25_imputed12marks_dense.bed.gz`. [Online; accessed 1-December-2018]. 2018.

[69] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN: 978-3-319-24277-4. URL: `https://ggplot2.tidyverse.org`.

[70] Pugacheva EM, Rivero-Hinojosa S et al. **Comparative analyses of CTCF and BORIS occupancies uncover two distinct classes of CTCF binding genomic regions**. *Genome Biology* 2015, **16**: 161.

[71] Xiong J, Dadon DB et al. **3D chromosome regulatory landscape of human pluripotent cells**. *Cell Stem Cell* 2016, **18**: 262–275.

[72] Schmidt D, Schwalie PC et al. **Waves of Retrotransposon Expansion Remodel Genome Organization and CTCF Binding in Multiple Mammalian Lineages**. *Cell* 2012, **148**: 355–348.

[73] Jacobs J, Atkins M et al. **The transcription factor Grainy head primes epithelial enhancers for spatiotemporal activation by displacing nucleosomes**. *Nature Genetics* 2018, **50**: 1011–1020.

[74] Koh W, Pan W et al. **Noninvasive in vivo monitoring of tissue-specific global gene expression in humans**. *Proceedings of the National Academy of Sciences* 2014, **111**: 7361–7366.

[75] Corces MR, Granja JM et al. **The chromatin accessibility landscape of primary human cancers**. *Science* 6413 2018, **362**.

[76] Friedman JR and Kaestner KH. **The Foxa family of transcription factors in development and metabolism**. *Cellular and Molecular Life Sciences* 2006, **63**: 2317–2328.

[77] Augello MA, Hickey TE et al. **FOXA1: master of steroid receptor function in cancer**. *The EMBO journal* 2011, **30**: 3885–3894.

[78] Labbe DP, Brown M et al. **Transcriptional Regulation in Prostate Cancer**. *Cold Spring Harbor perspectives in medicine* 2018.

[79] Beltran H, Prandi D et al. **Divergent clonal evolution of castration-resistant neuroendocrine prostate cancer**. *Nature Medicine* 2016, **22**: 298–305.

[80] Puca L, Bareja R et al. **Patient derived organoids to model rare prostate cancer phenotypes**. *Nature Communications* 2018, **9**.

[81] Puca L, Vlachostergios PJ et al. **Neuroendocrine Differentiation in Prostate Cancer: Emerging Biology, Models, and Therapies**. *Cold Spring Harbor perspectives in medicine* 2018.

[82] Yamasaki M, Nomura T et al. **Chronic hypoxia induces androgen-independent and invasive behavior in LNCaP human prostate cancer cells**. *Urologic oncology* 2013, **31**: 1124–1131.

[83] Khalesi E, Nakamura H et al. **The Krüppel-like zinc finger transcription factor, GLI-similar 1, is regulated by hypoxia-inducible factors via non-canonical mechanisms**. *Biochemical and biophysical research communications* 2013, **441**: 499–506.

[84] Flores-Morales A, Bergmann TB et al. **Proteogenomic characterization of patient-derived xenografts highlights the role of REST in neuroendocrine differentiation of castration-resistant prostate cancer**. *Clinical Cancer Research* 2019, **25**: 595–608.

[85] Heintzman ND, Stuart RK et al. **Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome**. *Nature Genetics* 2007, **39**: 311–318.

[86] Andersson R, Gebhard C et al. **An atlas of active enhancers across human cell types and tissues**. *Nature* 2014, **507**: 455–461.

[87] Eisenberg E and Levanon EY. **Human housekeeping genes, revisited**. *Trends In Genetics* 2013, **29**: 569–574.

[88] O'Leary NA, Wright MW et al. **Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation**. *Nucleic Acids Research* 2016, **44**: D733–745.

[89] Rodriguez JM, Maietta P et al. **APPRIS: annotation of principal and alternative splice isoforms**. *Nucleic Acids Research* 2013, **41**: D110–117.

[90] Ernst J and Kellis M. **ChromHMM: automating chromatin-state discovery and characterization**. *Nature Methods* 2012, **9**: 215–216.

[91] Sheffield NC, Thurman RE et al. **Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions**. *Genome Research* 2013, **23**: 777–788.

[92] Nie Y, Cheng X et al. **Nucleosome organization in the vicinity of transcription factor binding sites in the human genome**. *BMC Genomics* 2014, **15**: 493.

[93] Sung MH, Baek S et al. **Genome-wide footprinting: ready for prime time?** *Nature Methods* 2016, **13**.3: 222–228.

[94] Baek S, Goldstein I et al. **Bivariate Genomic Footprinting Detects Changes in Transcription Factor Activity**. *Cell Reports* 2017, **23**.

[95] Tan PY,Chang CW et al. **Integration of regulatory networks by NKX3-1 promotes androgen-dependent prostate cancer survival**. *Molecular and cellular Biology* 2012, **32**: 399–414.

[96] Chen L, Toke NH et al. **A reinforcing HNF4–SMAD4 feed-forward module stabilizes enterocyte identity**. *Nature Genetics* 2019, **51**: 777–785.

[97] Zhang B, Wang J et al. **Proteogenomic characterization of human colon and rectal cancer**. *Nature* 2014, **513**: 382–387.

[98] Gaffney DJ, McVicker G et al. **Controls of Nucleosome Positioning in the Human Genome**. *PLoS Genetics* 2012, **8**: e1003036.

[99] Ferrari KJ, Scelfo A et al. **Polycomb-dependent H3K27me1 and H3K27me2 regulate active transcription and enhancer fidelity.** *Molecular Cell* 2014, **9**: 49–62.

A. *Appendix*

A.1. **Supplementary figures**

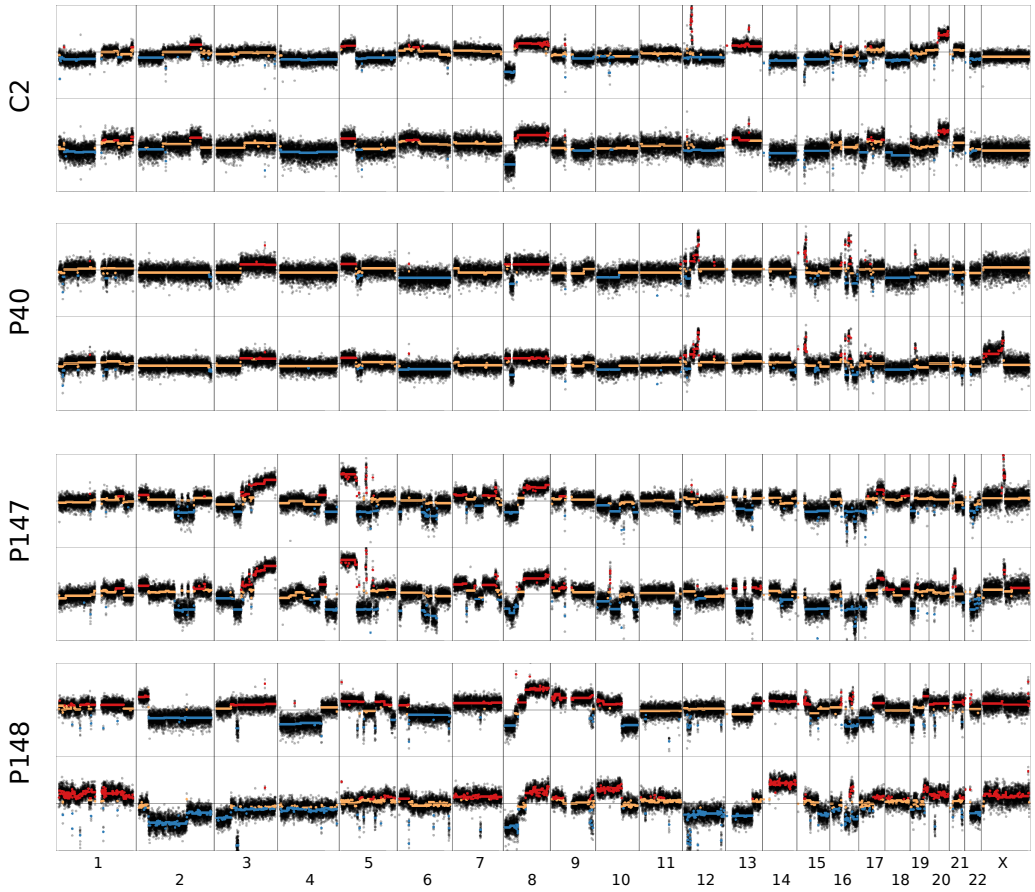

Figure A.1: Copy-number analyses of paired cell-free DNA samples from cancer patients by Plasma-Seq. Copy-number gains are depicted in red, losses in blue and balanced regions are depicted in orange. Raw data of individual 56kbp bins are depicted in black.