Peter Ulz, BSc

Nucleosome protection of circulating tumor DNA

MASTER THESIS

to achieve the university degree of

Diplom-Ingenieur

Individual master's degree programme: Computational Biology

submitted to

Graz, University of Technology

Supervisors

Univ.-Prof. Dr. Christoph Wilhelm Sensen

Institute of Molecular Biotechnology

Univ.-Prof. Dr. Michael Speicher

Institute of Human Genetics

Medical University Graz

Graz, October 2016

Affidavit

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present master‘s thesis.

Date   Signature

Abstract

Circulating tumor DNA (ctDNA) is mainly shed into the blood circulation by cells going through apoptosis, a process during which DNA is fragmented by DNase digestion before it is released. This fragmentation step is not random, since DNA bound to a histone complex (*i.e.* the nucleosome) is preferentially protected from degradation. It is also known that in a certain region before the transcription starts in a gene, nucleosomes are actively removed from the DNA in order to facilitate transcription.

Here, this fact is used to deduce the expression of genes based on the relative underrepresentation of fragments at the transcription start site in genes which are transcribed to mRNA.

Whole-genome sequencing data of plasma of healthy controls is used to establish an algorithm which predicts the expression status of genes based on Support Vector Machines. These results are compared to previously published data of gene expression analyses of circulating RNA.

In a next step, whole-genome sequencing of cell-free DNA from two patients with metastasized breast cancer was performed to establish the applicability of this approach to cancer samples.

Expressed genes differ greatly in their coverage profile around transcription start sites in healthy controls. This signal can be exploited to infer the expression status of genes. Moreover, this approach also works in plasma DNA from tumor patient, provided that the fraction of tumor-derived DNA is high enough.

The analysis of gene expression from ctDNA could enhance the application of liquid biopsies and may be used in various fields of medicine, apart from cancer.

Contents

Abbreviations

2K-TSS 2 kilobases around transcription start site

BAM binary alignment/mapping format

BWA Burrows-Wheeler Aligner

cfDNA cell-free DNA

cfRNA cell-free RNA

CNA copy number alteration

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

PCR polymerase chain reaction

RMA robust multi-array average

SAM sequence alignment/mapping format

SNP single (or simple) nucleotide polymorphism

SNV single (or simple) nucleotide variation

SRA sequence read archive

TSS transcription start site

UCSC University of California in Santa Cruz

1. *Introduction*

Liquid biopsies, *i.e.* the analysis of circulating tumor cells (CTCs) and circulating tumor DNA (ctDNA), are evolving to become promising tools to study characteristics of the tumor genome non-invasively in patients with cancer from the peripheral blood. In particular, ctDNA is intensively investigated as a biomarker in translational and clinical research as it may reflect the tumor burden and may provide information on therapy response and the development of therapy resistance.

1.1. **Tumorgenetics**

Cancer is a heterogeneous disease which can affect almost every organ in the human body. It is now widely accepted that the transformation of a healthy cell to a tumor cell is due to mutations in the cancer cell's DNA which gives rise to certain characteristics that allow the cancer cell to grow uncontrollably [1]. Many different types of mutations have been described from single nucleotide substitutions to mutations affecting whole chromosomes or even whole genome duplications.

Mutations in the cancer genome can be divided into driver and passenger mutations. Driver mutations are directly affecting the cancer cell's ability to grow whereas passenger mutations have no positive effect on the tumorigenic capabilities [2].

To date, many genes have been found which recurrently harbor driver events and thus may promote tumor growth. Usually, these genes fall in either of two classes [3]:

- Tumor suppressor genes

  Tumor suppressor genes control the cell cycle in healthy cells. Only when they are inactivated by mutation, tumor growth can circumvent this control mechanism. Usually mutations in these genes destroy the protein function and are distributed throughout the whole gene. *RB1*, for example, was the first identified tumor suppressor gene [1, 3, 4].

- Oncogenes

Oncogenes usually do not have a known function in the cell cycle as long as they are in their normal state. However, certain mutations can activate or amplify a function of these genes which cause the cell to grow indefinitely [3, 5].
KRAS is an example of a well-studied oncogene [3].

The origin of mutations can be manifold:

- Exposure to mutagenic substances
- Inheritance of cancer-predisposing mutations
- Random errors due to non-perfect DNA replication

Understanding the role of cancer genetics becomes increasingly important as more and more cancer therapies targeting molecular characteristics become available [6].

1.2. Tumor evolution

The genetic landscape of cancer is very complex. Several mutations are necessary to transform a healthy cell to a tumor cell, however, when the tumor reaches a malignant stage, it comprises a multitude of different cells comprising subclones of the overall tumor. These are cells which originate from a single cell but differentiate (both genetically and phenotypically) during the course of disease [3, 7]. Subclones may have a different set of somatic mutations and are formed by a process termed ”tumor evolution”. This impedes analysis of somatic mutations, but may also ultimately lead to a failure of a therapy targeting a molecular feature which might not be present in every subclone[8]. To study heterogeneity and clonal evolution of a tumor throughout therapy, the analysis of ctDNA promises a significant advancement, since repeated sampling is a lot easier compared to tissue biopsies [9].

1.3. Circulating tumor DNA

Cell-free DNA (cfDNA) is DNA circulating in the liquid parts of the peripheral blood bound to histones in form of nucleosomes, which can be isolated by centrifugation of all cellular parts and extraction of DNA from the plasma or serum [10, 11, 12]. Although most extracellular DNA in the plasma is present in the form of nucleosomes, at least some of it may be present in membranous vesicles [13].

In cancer patients, a fraction of the cell-free DNA originates from tumor cells and may be used to characterize the tumor genome non-invasively [11]. The presence of cell-free DNA has been shown for the first time in 1948 by Mandel and Métais [14], while Leon *et al.* were the first to identify changes in plasma of cancer patients [15] around 30 years later.

The decreasing cost of sequencing in recent years has now opened up many new possibilities, including whole genome sequencing of ctDNA [16], high-resolution detection of mutations in ctDNA [17] and methylation analyses [18].

1.3.1. *cfDNA release*

Although the analysis of cfDNA has been in wide use over the last years, very little is known about the mechanisms of DNA release and the involved kinetics. Apoptosis and necrosis have been identified as important contributors of cfDNA [10], however, also active release of DNA by the cells into the circulation was postulated [19].

However, cell-free DNA is only released in very small quantities, which complicates further downstream analyses [9].

1.3.2. *Size*

Apart from very low concentrations, another complication in the analysis of cfDNA is its fragmentation. While there is agreement on the size of cfDNA in healthy patients [20], many conflicting results were shown for ctDNA in patients with cancer [21]. While one group showed that ctDNA is shorter than cfDNA and yields 10bp-periodic peaks below the cfDNA modal peak of 166bp [22], results from the Institute of Human Genetics in Graz show size distributions of multiples of 166bp in metastasized cancer patients (see figure 1.1) [23].

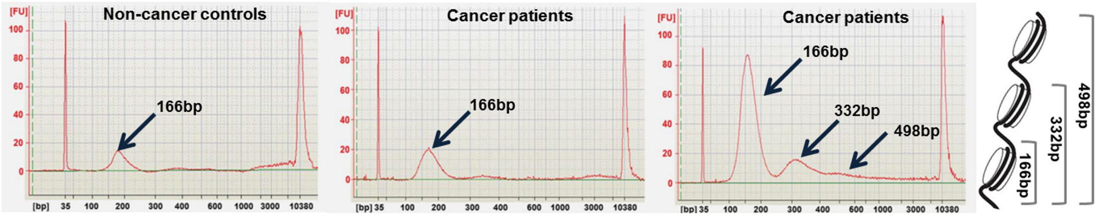

Figure 1.1: Size distribution of ctDNA in cancer patients occur in multiples of 166bp. [Image taken from Heitzer *et al.* [9]].

It has been hypothesized that the particular size distribution comes from preferential protection of DNA from nucleases via binding of nucleosomes [22]. This is supported by an analysis of cell-free DNA which found cfDNA occurring in clusters along the genome which correlate to nucleosomal arrays [24] (figure 1.2).

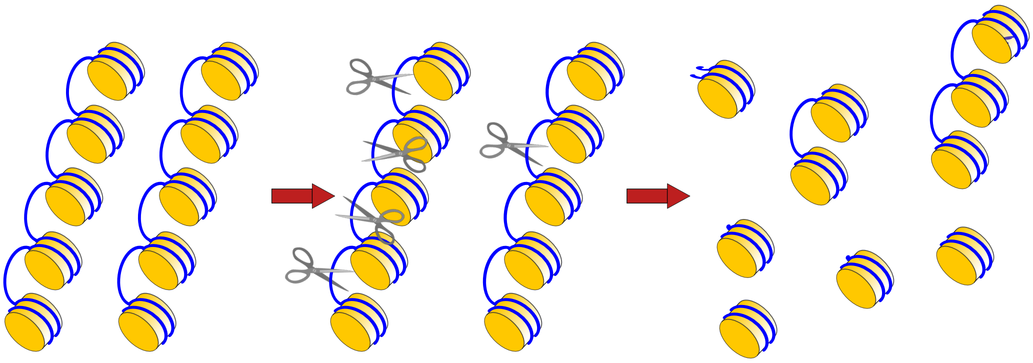

Figure 1.2: Circulating cell-free DNA is released by various cell types and is digested by nucleases during apoptosis. DNA that is bound to nucleosomes is protected from digestion and are thus more likely available for down-stream analyses.

1.4. Nucleosomes

1.4.1. *Organization*

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 [25], 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 [25] (see figure 1.3). This first step of packaging DNA reduces space requirements for DNA about 10,000-fold [26].

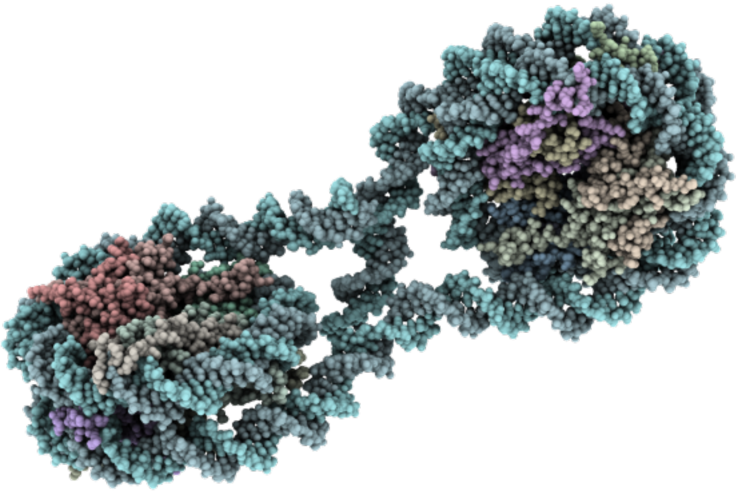

Figure 1.3: DNA (turquoise) is wrapped around two histone octamers, connected by linker sequence. [Image generated by Qutemol (v0.4.1) from PDB accession: 1zbb].

1.4.2. *Sequence specificity*

The positioning of nucleosomes along the genome is affected by the DNA sequence and its ability to rotate around the histone octamer. Thus, sequences which do not favor such rotational positions tend to inhibit binding of the nucleosome. Homopolymeric sequences such as poly(dA:dT) sites tend to have inhibitory effect on nucleosome binding, whereas 10bp-periodic alternating dinucleotides of GC and (AA/TT/TA) strongly favor binding of nucleosomes (see figure 1.4)[27].

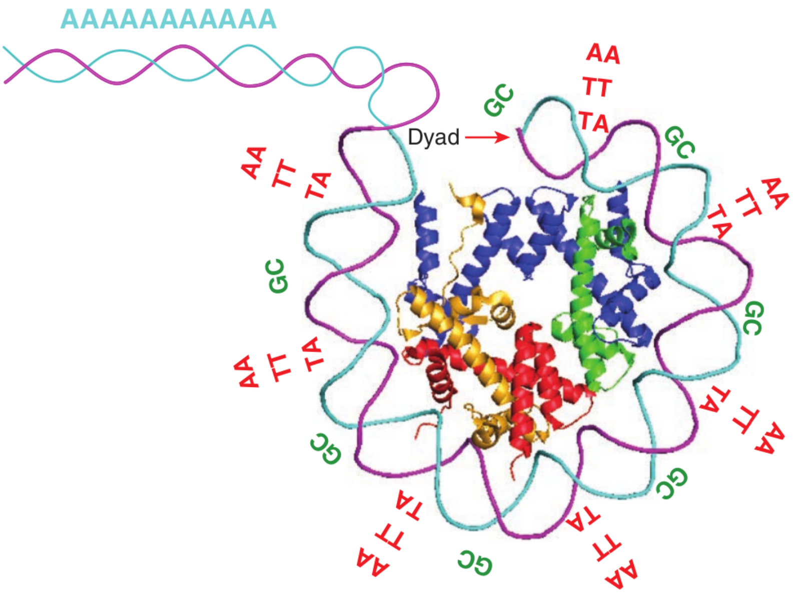

Figure 1.4: Periodic dinucleotides of GC as well as AA/TT/TA favor nucleosome binding as these structures may provide enough bending to wrap around nucleosomes. [Image taken from Struhl *et al.* [27]].

1.4.3. *MNase-seq*

In order to study nucleosome positioning along the genome, a common approach is to digest chromatin using a nuclease derived from Micrococci termed micrococcal nuclease (MNase). Here, DNA is digested unspecifically, however, DNA bound to nucleosomes is protected from this process which allows preparation of every sequence initially bound to nucleosomes. After treatment, DNA can be analyzed either via microarray analysis or via next generation sequencing [26].

1.4.4. *Nucleosome depleted region at transcription start sites*

Transcription start sites (TSS) tend not to harbor a lot of nucleosomes, since AT-rich promotor regions do not bend enough to wrap around the histone octamer [28]. Moreover, chromatin remodeling complexes maintain the nucleosome depleted region (NDR) by sliding nucleosomes away in order to ensure accessibility of transcription factors and Polymerase II [29, 30]. In addition, the first nucleosome after the transcription start (+1 nucleosome) is strongly positioned *in vivo* but not *in vitro* [27]. The strong positioning then decreases for the following nucleosomes and may be attributable to the strong positioning of the first nucleosome and the restriction on linker length [27]. Adding ATP-dependent nucleosome remodelers to cell-free extract of yeast enhances the nucleosome depletion at promotors also *in vitro* to nearly the same way as it is observed in vivo [27].

1.4.5. *Nucleosomes at transcription factor binding sites*

Furthermore, transcription factor binding sites seem to be depleted of nucleosomes *in vivo*, although a high occupancy of these sites was predicted *in vitro* [31].

1.5. **Aim of the thesis**

The aim of this thesis is to explore whether it is possible to predict the gene expression of a tumor, based on coverage differences in sequencing data due to the pattern of nucleosomal binding on the plasma DNA.

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

- Confirm the nucleosome association of cfDNA in healthy samples
- Analyze coverage differences around transcription start sites between expressed and unexpressed genes
- Select features to predict expression of every gene
- Validation of gene expression prediction
- Explore whether gene expression prediction can be performed in tumor samples

2. *Methods*

2.1. **Data sets**

Several data sets were used for the subsequent analyses

2.1.1. *Pool of cfDNA samples (n=179)*

A set of 179 cell-free DNA samples (cancer patients and non-cancer controls) which have been analyzed using paired-end low-coverage whole-genome sequencing were used to explore differences in fragment size between nuclear DNA and mitochondrial DNA.

2.1.2. *cfDNA samples of non-cancer controls (n=104)*

Cell-free DNA samples of 104 non-cancer controls which were subjected to single-end low-coverage whole-genome sequencing were used to deduce coverage differences in the transcription start region and to setup and validate the gene expression prediction.

2.1.3. *cfRNA data of non-pregnant women (n=4)*

Raw cfRNA sequencing data of 4 non-pregnant women were downloaded from the Sequence Read Archive (SRA) [32] (Accessions: SRR1296080, SRR1296081, SRR1296082 nad SRR1296083). This data were generated by Koh *et al.* [33].

In addition microarray (Affymetrix Human Gene 1.0 ST Array) data of the same samples were downloaded from the Gene Expression Omnibus (GEO) [34]. This datasets contained precomputed Robust multi-array average (RMA) values for 48 samples, 4 of which were from non-pregnant women (GSM1370906, GSM1370907, GSM1370908 and GSM1370909). Data of non Ref-Seq transcripts was discarded and RMA values for all four samples were averaged.

2.1.4. *MNase-seq data of GM12878*

Results of MNase-seq analyses from the cell-line GM12878 were downloaded as BigWig files from the UCSC Genome Browser [35].

2.1.5. *ctDNA of breast cancer samples*

Plasma DNA of two patients with metastasized breast cancer were used for testing the applicability on tumor-derived DNA. Samples have been isolated and prepared as described below by members of the Institute of Human Genetics.

2.1.6. *RNA-seq of primary tumors*

In order to compare gene expression predictions from plasma DNA of breast cancer patients, RNA-seq was performed on the primary tumor tissues of both breast cancer patients. This was done at the Institute of Pathology.

2.1.7. *Functional enrichment of Top100 genes*

Functional enrichment analysis on the Top100 genes from chromosome 1q (for B7) and gained regions of chromosome 8 (for B13) was performed using DAVID [36]. Gene ontology terms [37] and pathways from the KEGG [38], Biocarta [39] and Reactome [40] database were annotated.

2.2. **Plasma DNA preparation**

Plasma DNA was prepared using the QIAamp DNA Blood Mini Kit (Qiagen, Hilden, Germany) as previously described [23]. Samples selected for sequencing library construction were analyzed on the Bioanalyzer instrument (Agilent Technologies, Santa Clara, CA, USA) to observe the plasma DNA size distribution. DNA preparation was done by members of the Institute of Human Genetics.

2.3. **Plasma DNA sequencing**

Shotgun libraries of plasma DNA and tumor DNA were prepared using the TruSeq DNA Nano library preparation kit by Illumina (Illumina, San Diego, CA, USA) with a starting amount of 5- 10ng according to the protocol [41]. However, due to the low DNA input, we increased the amount of PCR cycles to 25. Furthermore, the fragmentation step was omitted due to the degradation of plasma DNA. Libraries were sequenced on the Illumina MiSeq and NextSeq sequencers by members of the Institute of Human Genetics.

2.4. Analysis environment

The following analysis tools are used:

- GNU make (4.1)
- R (2.14.1)
  - library e1071 (1.6-7)
- R (3.2.3)
  - library MASS (7.3-45)
- Python (2.7.3)
  - library sys
  - library argparse
  - library subprocess
  - library numpy
  - library scipy
  - library os.path
  - library multiprocessing
- Java (2.7.3)
- zcat
- samtools (v.0.1.18)
- picard (1.128)
- bwa (0.7.4-r385)
- tophat (2.3.7)
- cufflinks (2)
- VarScan2 (2.2)

Details of the code used to produce the results can be found at GitHub [42].

2.5. Nucleosomal association

2.5.1. *Fragment size at mitochondrial genome*

Insert size of 179 low-coverage whole-genome plasma sequencing datasets were aligned to the human hg19 genome using bwa backtracking (version 0.7.4 [43]), since second reads are about 35bp. A SAM file was generated using the sampe command provided by bwa and converted to BAM files [44]. Individual BAM files were merged subsequently. Mitochondrial reads were extracted by samtools view with the region filtering option [45] and subsequently, insert sizes were calculated using Picard's CollectInsertSizeMetrics function [46].

In order to get reads from the nuclear genome, the merged BAM file was downsampled using Picard's DownsampleSam function [46] and only a fraction of 0.0002 of all the reads were kept to produce a comparable amount of samples for insert-size calculation. Hits to the mitochondrial genome were discarded and insert sizes were again calculated using Picard [46].

Insert size distributions were then plotted in R.

2.5.2. *Read trimming*

In order to get a cleaner signal of nucleosome-associated region with the cell-free DNA, reads were trimmed to 60bp, starting from base 53 to base 113. This was done, since cell-free DNA has a size modus at 166bp, thus base 53-113 should represent the center of the molecule and the region most likely be bound to nucleosomes (see figure 2.1). Read trimming was done using fastx_trimmer tool provided by the FASTX toolkit [47].

2.5.3. *Alignment*

Trimmed reads of the control samples were aligned to the human hg19 genome using bwa [43]. PCR duplicates were marked using samtool's rmdup function [45] and merged to produce a single BAM file containing all individual control BAM files.

2.5.4. *Coverage at nucleosome array*

From the merged control BAM files, WIG files containing coverage information in the region with ordered nucleosome arrays [48] were generated by a wrapper script which

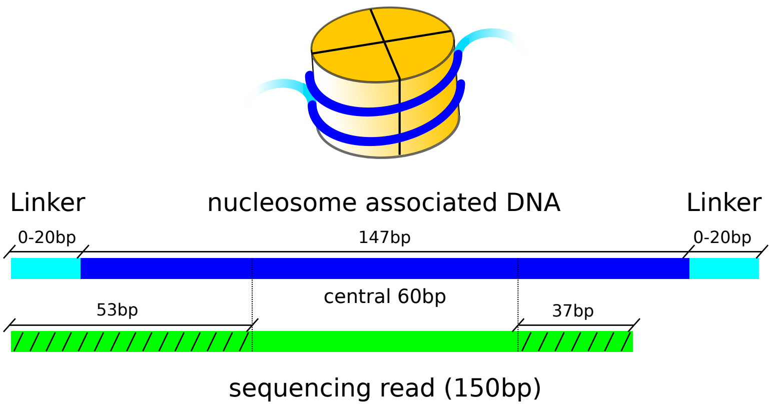

Figure 2.1: Reads were trimmed to 60bp (from base 53 to base 113) in order to only include the most central portion of a (hypothetical) 166bp cfDNA fragment, which should be the portion with the highest association to nucleosomes.

uses the bam_to_wiggle.py script provided by Brad Chapman via GitHub [49].

Wiggle files were loaded into the UCSC Genome Browser and MNase-Seq results of GM12878 from the ENCODE project [50] were visualized together with coverage data from merged control cfDNA.

2.5.5. *Correlation at nucleosome array*

Based on the wiggle files, Pearson's and Spearman's correlation coefficients were calculated between the merged controls and the MNase-Seq results from GM12878 from the ENCODE project in R [51].

2.6. Coverage profile at transcription start sites

The coverage of transcription start sites (TSS) was extracted from the trimmed and aligned single-end reads with the samtools depth [45] command and was normalized by the mean coverage of the regions from -3000 to -1000 bp from TSS and 1000 to 3000 bp, respectively. This is done to compare for copy number differences and to normalize for the variable total read input. Normalized coverage values were then averaged and confidence regions were calculated to check for the variability of the signal. Averaging was done on the following gene sets:

- Housekeeping genes
- Unexpressed genes in FANTOM5
- 1000 highest and lowest expressed genes from cfRNA studies

2.6.1. *Housekeeping genes*

Housekeeping genes as defined by Eisenberg and Levanon [52] were used to calculate a normalized TSS profile.

2.6.2. *FANTOM5*

The FANTOM5 project, which catalogs gene expression for a diverse set of tissues, was used to derive genes unexpressed in all tissues [53]. To this end, raw data was downloaded from the EBI expression atlas [54] and genes were searched which were expressed <0.1 FPKM in all of 56 tissues.

2.6.3. *cfRNA gene expression microarrays*

The 1000 highest and lowest expressed genes were extracted from preanalyzed gene-expression microarray data (Affymetrix Human Gene 1.o ST Array) of circulating cell free RNA [33] of four non-pregnant women. RMA values were averaged and ranked and only genes which code for mRNAs in RefSeq were used for subsequent analyses.

2.6.4. *cfRNA RNA-seq analysis*

FastQ files of the four non-pregnant women were downloaded from the SRA, aligned to the human genome using Tophat (v2.3.7) [55] and FPKM were calculated using cufflinks2 [56].

2.7. **Gene expression prediction**

2.7.1. *Feature extraction*

Two parameters were used for the identification/prediction of genes into an expressed and unexpressed subset.

- The coverage between TSS-1000bp and TSS+1000bp (2K-TSS coverage)
- The coverage between TSS-150bp and TSS+50bp (NDR coverage)

For every TSS in RefSeq, parameters were extracted and divided by the relative copy number of that region identified in the copy number alteration (CNA) analysis step.

2.7.2. *Kernel density estimation*

In order to see whether both coverage signals combined yield a signal which can distinguish expressed from unexpressed genes, multivariate Kernel Density Estimation (KDE) [57] was performed on the coverage signals of the Top 1000 and the Bottom1000 genes respectively. To this end, the kde2d function provided by the MASS package was used in R.

2.7.3. *Machine learning*

In order to predict the expression status of individual genes, we used Support Vector Machines (SVM). To this end, an implementation of SVM provided by the e1071 package within R was used. As a training set for expressed genes, we used a random subset of 300 housekeeping genes out of 3,804 housekeeping genes which are expressed uniformly in multiple tissues and for unexpressed genes a random subset of 300 genes out of 670 reported to be unexpressed in most tissues by the FANTOM5 project [53]. Every gene not used as training data was predicted. Random subset selection and prediction was repeated a 1,000 times and prediction status for each TSS was recorded. We considered a gene to be expressed when the prediction consent of all the iterations was higher than 75%.

2.7.4. *Classification validation*

We conducted detailed analyses of sensitivity, specificity, accuracy, precision, and F1-score for different groups of genes, such as Top 100, Top 1000 and Top 5000 genes (*i.e.* the 100, 1,000, and 5,000, respectively, most highly expressed genes).

$$sensitivity = \frac{true\,positive}{true\,positive + false\,negative} \tag{2.1}$$

$$specificity = \frac{true\,negative}{true\,negative + false\,positive} \tag{2.2}$$

$$accuracy = \frac{true\,positive + true\,negative}{total\,number\,of\,genes} \tag{2.3}$$

$$F1 - score = \frac{2 * true\,positive}{2 * true\,positive + false\,positive + false\,negative} \tag{2.4}$$

In order to perform these analyses also for the full set of expressed genes, different FPKMs were considered as threshold to distinguish between transcribed and unexpressed genes because low- abundance transcripts might not represent active transcripts but rather technical or biological noise. First, a FPKM threshold of 1 was used, as several previous studies had used such a value as a fixed threshold. Second, a FPKM value of 0.44 was used as a reliable and robust threshold between active and background gene expression, which was established in a recent study based on large-scale studies such as the ENCODE project [58] (table 3.1).

2.7.5. *In-silico dilution simulation*

Dilution simulations to test the reliability of the prediction at varying tumor fractions were performed. To this end, the distribution of the 2K-TSS and the NDR coverage parameters of the 1000 least expressed were modeled as a normal distribution with mean and standard deviation calculated from the two parameters from the aforementioned genes. Subsequently, random numbers from these distributions were added to the parameters of the Top 1000 expressed genes at varying proportions (denoted below as λ) to simulate a dilution of the signal (Top 1000 genes) with background noise (Bottom 1000 genes) having mean μ and standard deviation σ. Prediction accuracy was measured for every dilution.

$$2K - TSS(\lambda) = 2K - TSS * \lambda + (\lambda - 1) * N_{2K-TSS}(\mu_{Bottom1000}, \sigma_{Bottom1000}) \tag{2.5}$$

$$NDR(\lambda) = NDR * \lambda + (\lambda - 1) * N_{NDR}(\mu_{Bottom1000}, \sigma_{Bottom1000}) \tag{2.6}$$

2.7.6. *Quantitative analysis*

In order to test whether the 2K-TSS and the NDR coverage contain quantitative information about gene expression, we annotated every TSS from the merged controls with the FPKM values of the respective genes from the aforementioned plasma RNA-seq experiments [33]. FPKM values were ranked and percentiles of the ranks were calculated. Subsequently, data from the 2K-TSS and NDR coverage parameters were binned and average percentile of every (integer) bin was calculated. Bins containing 10 TSSs or less were discarded.

2.7.7. *Linear Model*

By employing multiple regression analysis, a more detailed quantitative prediction was investigated. To this end, a linear model was fitted by using the glm function in R and specifying 2K-TSS and NDR coverage as predictors and the FPKM percentile as response variable.

2.8. **Tumor samples**

2.8.1. *Single nucleotide variant (SNV) identification*

Paired-end reads of both plasma samples from two breast cancer patients were aligned to the human (hg19) genome using bwa [43]. PCR duplicates were removed using samtools rmdup function and a pileup file of every position in the genome was created using mpileup function also provided by samtools [45]. Variant calls were generated using VarScan 2 (version 2.2) using standard parameters (minimum coverage: 8, minimum reads supporting variant: 2, minimum variation frequency, 1%, minimum average quality 15 and p-value threshold 0.99) [59].

2.8.2. *SNV filtering*

SNVs were filtered to identify possible driver mutations of the tumor. Moreover, germline mutations are also in the initial call set, since no germline DNA was available for comparison. In a first step SNVs are annotated using annovar [60], which adds information from several databases to the identified SNP calls. The first filtering step removes every mutation outside of protein coding sequences, since these mutations are hard to interpret. Subsequently, synonymous SNVs are removed as well as SNVs with a high allele frequency (>1%) in a healthy population (as determined by the 1000genomes project [61], ExAc database 2 [62] and the Exome sequencing project [63]). In a last step, SNVs are removed which lie within segmental duplications, since these are enriched for artifacts.

2.8.3. *Copy number alteration analysis*

Raw (single-end) reads of the two breast cancer samples and the merged controls were aligned to the human hg19 genome using Burrows-Wheeler aligner (bwa) [43] where the pseudo-autosomal region of the Y-chromosome was masked. PCR duplicates were removed and reads were counted in 50,000 genome bins, each containing the same amount of mappable positions (approximately 56kbp). Raw read counts were normalized by the median bin count and GC correction was done using LOWESS smoothing. Furthermore, corrected read counts were normalized by mean bin counts of cfDNA samples of 10 non-cancer controls and segmented using both CBS and GLAD (which partition the raw copy number data into segments of similar copy number) provided by the CGHweb framework [64].

2.8.4. *Insert size estimation*

Fragment lengths were calculated by making use of the paired-end sequencing approach. Both ends of a fragment are aligned to the genome and the distance between the alignments are calculated. Paired-end reads were aligned (simultaneously) using bwa mem [43]. After PCR duplicate removal with samtools [45], insert sizes were calculated using Picard (version 1.128) [46]. For region-specific insert size calculations, parts of the aligned BAM files were extracted using samtools view and insert size calculation again performed using Picard.

Insert size distribution was estimated using Kernel Density Estimation (KDE), calculated by using R's density function on a random sample of 100,000 insert sizes from the respective chromosomal region. Confidence intervals were obtained by repeated sampling (n=1000) of the insert sizes.

2.8.5. *Focal amplifications*

From the segmented copy number data, focal amplifications were identified based on certain evaluation criteria [65]:

- Segment should be <20 Mbp
- Log_2-ratio of the segment must be >0.2
- Segment should contain at least one but not more than 100 genes

- Log_2-ratio of the segment must >0.2 higher than average $\log_2$-ratio of neighboring 20 Mb if it contains a known tumor gene
- Log_2-ratio of the segment must >0.58 (corresponds to 3 copies) higher than average $\log_2$-ratio of neighboring 20 Mb if it contains a known tumor gene
- Segment should not contain segmental duplications in >50% of its size
- Segment should not overlap with known entries in DGVar

2.8.6. *Tumor fraction estimation*

The tumor fraction of the two breast cancer samples was estimated by applying ABSOLUTE [66] to the segmented $\log_2$-ratios obtained by the copy number alteration (CNA) analysis. ABSOLUTE takes the copy number information obtained in the prior step and compares this to a database of known cancer karyotypes and tries to find the best fit using maximum likelihood. Plausible karyotypes can be selected and ABSOLUTE can calculate the ploidy, purity and subclonality of a tumor sample. The detailed workflow can be seen in figure 2.2.

2.8.7. *Isoform discrimination*

Expressed isoforms were determined by calculating the distance of the two parameters between the TSS in the merged control data and the tumor patient after normalizing both parameters in both data sets. TSSs which lead to higher expression in the tumor should decrease in both the 2K-TSS and the NDR coverage when compared to the same TSS in control data.

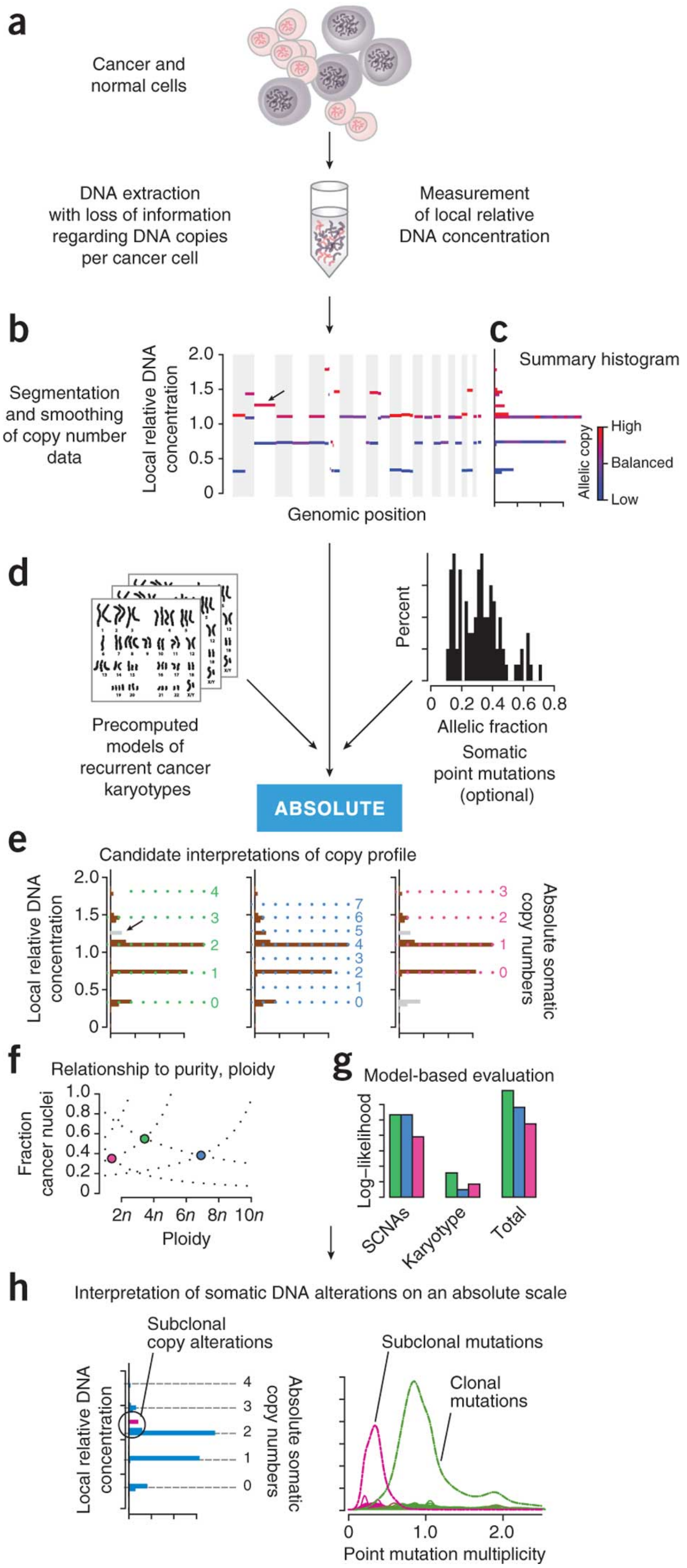
a
Cancer and normal cells
DNA extraction with loss of information regarding DNA copies per cancer cell
Measurement of local relative DNA concentration
b
Segmentation and smoothing of copy number data
Local relative DNA concentration
2.0
1.5
1.0
0.5
0
Genomic position
c
Summary histogram
Allelic copy
High
Balanced
Low
d
Precomputed models of recurrent cancer karyotypes
Percent
0 0.2 0.4 0.6 0.8
Allelic fraction
Somatic point mutations (optional)
ABSOLUTE
e
Candidate interpretations of copy profile
Local relative DNA concentration
Absolute somatic copy numbers
f
Relationship to purity, ploidy
Fraction cancer nuclei
1.0 0.8 0.6 0.4 0.2 0
2n 4n 6n 8n 10n
Ploidy
g
Model-based evaluation
Log–likelihood
SCNAs
Karyotype
Total
h
Interpretation of somatic DNA alterations on an absolute scale
Subclonal copy alterations
Local relative DNA concentration
Absolute somatic copy numbers
Subclonal mutations
Clonal mutations
0 1.0 2.0
Point mutation multiplicity

Figure 2.2: ABSOLUTE estimates the tumor fraction by comparing known tumor karyotypes to the distribution of copy numbers within a sample. In addition ABSOLUTE may use mutated allele fractions of somatic SNPs to further enhance the estimation. [Image taken from [66]].

3. *Results*

3.1. **Nucleosome association of cfDNA**

In order to get further evidence that cell-free DNA is associated with nucleosomes, two analyses were done:

- Comparing fragment sizes of mitochondrial DNA and nuclear cell-free DNA
- Compare coverage signal of non-cancer cfDNA controls to MNase-seq data

3.1.1. *Fragment size*

Since mitochondrial DNA is not packaged in the same way as nuclear DNA, insert sizes of paired-end reads (as a proxy to cfDNA fragment size) mapping to the mitochondrial genome should be different than that of reads mapping to the nuclear genome. Hence, we used paired-end sequencing data of 179 plasma cfDNA samples (of both healthy individuals and patients with cancer) and measured the insert sizes (see figure 3.1).

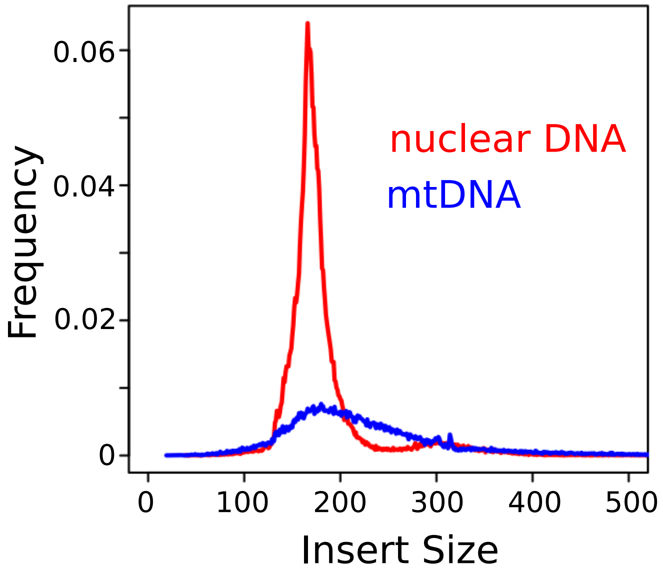

Figure 3.1: Comparing insert sizes of nuclear and mitochondrial DNA from paired-end sequences from 179 individuals. Insert sizes of fragments mapping to the nuclear genome have a distinct pattern with a mode around 166bp, while the length distribution of mitochondrial fragments appears to be wider distributed.

3.1.2. *MNase-seq comparison*

Chromosome 12 harbors a region close to the centromere, where 400 nucleosomes are found in an ordered array [48]. The coverage signal from cfDNA of 104 controls resembles the coverage enrichment signal from MNase-seq results produced in the ENCODE project (see figure 3.2).

Signals for the whole region (hg19: chr21:34,484,733-34,560,733) show a high correlation (Pearson correlation coefficient: 0.709, Spearman correlation coefficient: 0.708) (see figure 3.3).

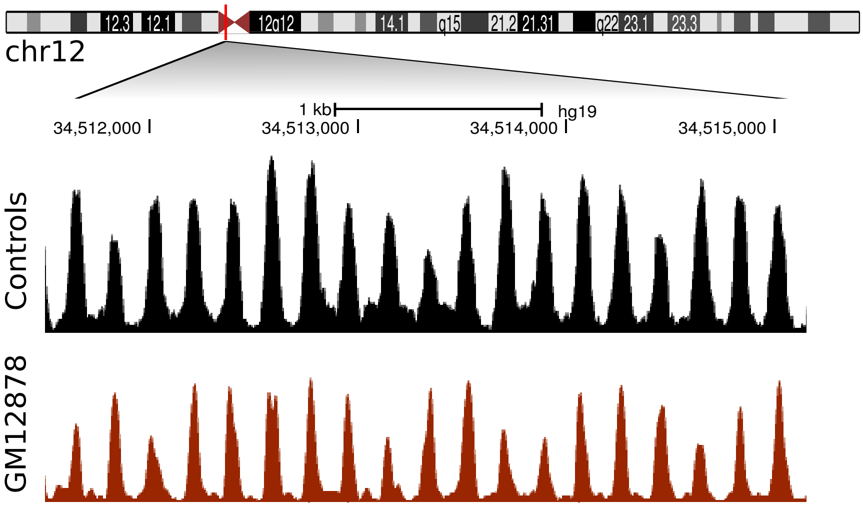

Figure 3.2: The coverage signal of cfDNA fragments resembles the signal enrichment from MNase-seq studies on the cell line GM12878 in a region of ordered nucleosomes on chromosome 12.

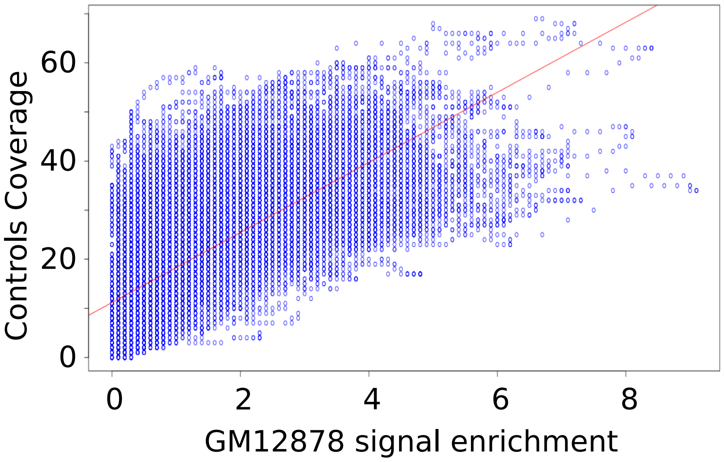

Figure 3.3: The coverage signal of cfDNA fragments correlates well to the signal enrichment from MNase-seq studies on the cell line GM12878.

3.2. Coverage at transcription start sites

Since transcription start sites of expressed genes should not be bound to nucleosomes to facilitate access of the transcription machinery, the coverage in that region should be lower in cfDNA. This has already been demonstrated in MNase-seq data [30]. Here, we used cfDNA sequencing data from 104 non-cancer controls to verify this observation.

3.2.1. *Housekeeping genes*

Since housekeeping genes should be expressed in every tissue, the coverage profile around the transcription start should give clues whether a nucleosome depleted region can be identified in cfDNA the same way as in MNase-seq experiments [30].

Indeed, the coverage signals drops around the TSS and on both sides and a repetitive pattern on both the 5' and the 3' end are visible (see figure 3.4).

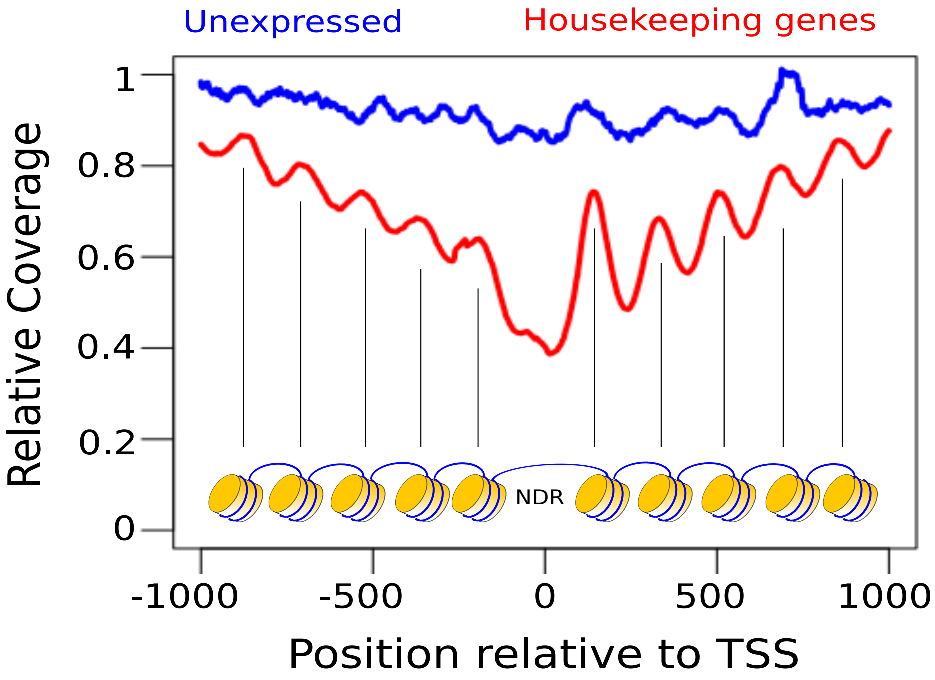

Figure 3.4: The coverage profile around transcription starts of housekeeping genes [52] in comparison to genes which are unexpressed in all tissues (as determined in the FANTOM5 project [53]) shows a distinct pattern of well-positioned nucleosomes in the former but not the latter. Vertical black lines denote recurrent nucleosome dyads, deduced from the peak in the coverage signal.

3.2.2. *Cell-free RNA*

To further confirm this effect, cell-free RNA (cfRNA) microarray analyses were obtained from four healthy (non-pregnant) women [33] and a coverage profile around the TSS was calculated for the 1000 highest (Top1000) and lowest expressed genes (Bottom1000) (see figure 3.5. This was also done for MNase-seq results from the ENCODE project to verify the nucleosome association (see figure 3.6).

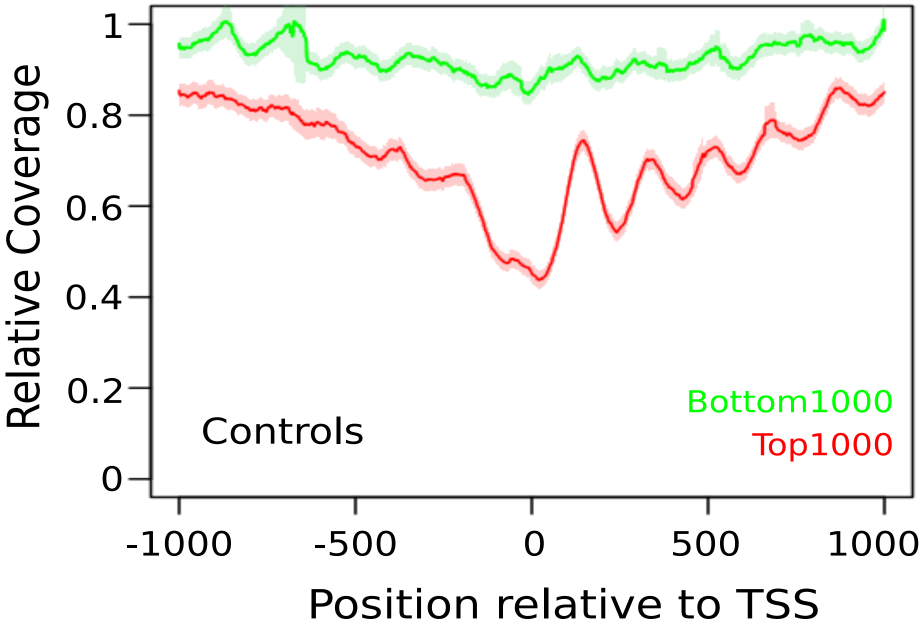

Figure 3.5: The coverage profile around transcription starts of the 1000 highest and lowest expressed genes [33] in 104 healthy individuals. Shaded areas represent 95% confidence intervals.

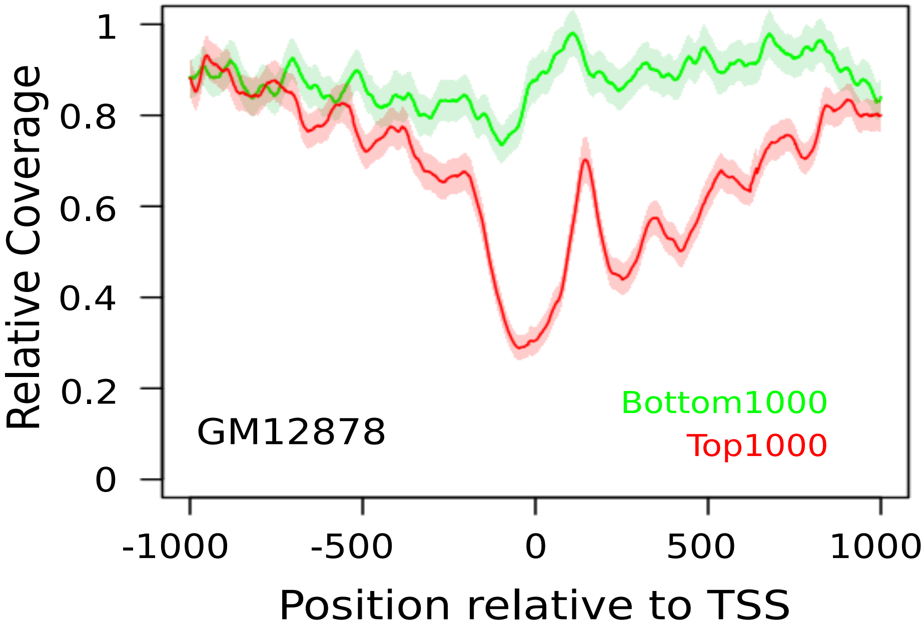

Figure 3.6: The coverage profile around transcription starts of the 1000 highest and lowest expressed genes [33] in the GM12878 MNase-seq dataset. Shaded areas represent 95% confidence intervals.

RNA-seq data were also available from the same four individuals. In a subsequent analysis, the TSS profile for gene subsets at varying FPKMs was analyzed and the strongest signal was seen in genes with the highest gene expression (figure 3.7).

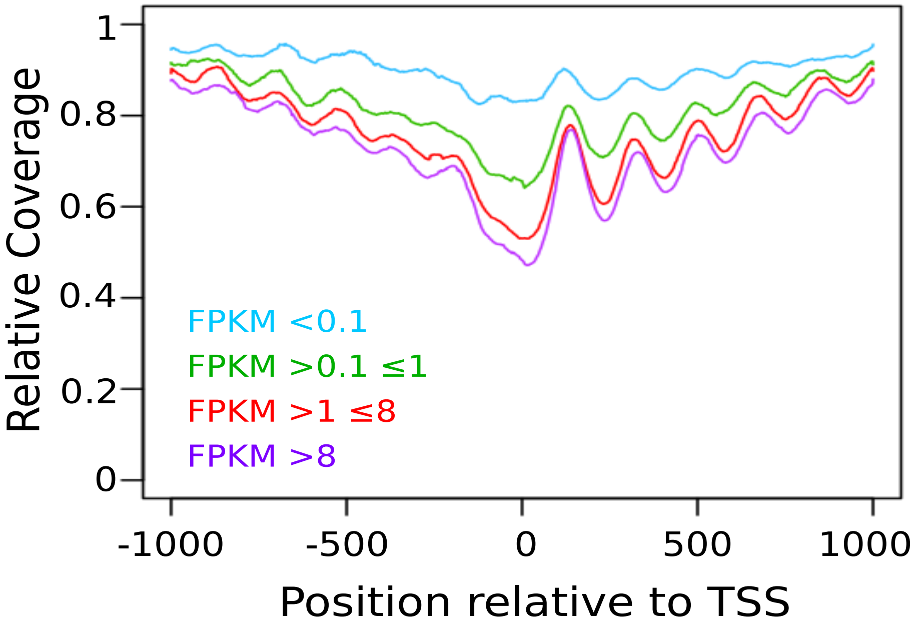

Figure 3.7: The coverage profile around transcription starts depends on the expression as measured by RNA-seq [33].

3.3. Prediction of expression status

As a next step, features were identified in order to distinguish between expressed and unexpressed genes for every single gene. It seems that not only the coverage directly around the transcription start seems to be lower for expressed genes, but also the coverage in a 2,000bp window around the TSS seems to be smaller. Thus, both, the 2,000bp window around the TSS (subsequently called "2K-TSS coverage") and a smaller region around the TSS (-150bp to 50bp from TSS; subsequently called "NDR coverage") were chosen as features. This leads to a reduction in complexity for subsequent analyses, since only 2 parameters (instead of 2,000) need to be analyzed

3.3.1. *Feature testing*

The distribution of the features for the 1000 highest and lowest expressed genes seem to discriminate well expressed from unexpressed genes. However, the distinction is more pronounced in the 2K-TSS Coverage than in the NDR coverage (figure 3.8).

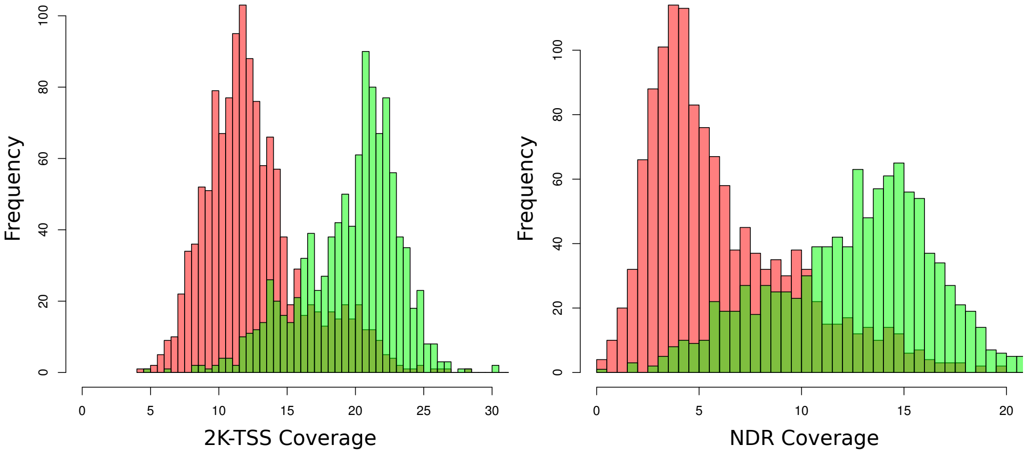

Figure 3.8: Histograms of distribution of the two features (2K-TSS coverage and NDR coverage) separates the highest 1000 expressed genes (displayed in red) from the lowest expressed genes (displayed in green) in cfRNA [33].

Also, multivariate distribution analysis by kernel density estimation results in two distinct peaks with high density (figure 3.9) in the merged control samples. Peaks most likely represent the Top 1000 genes (bottom left) and Bottom 1000 genes (top right). The distribution of the two features in two dimensions of every gene (not just the Top and Bottom 1000) yields a dense point cloud 3.10.

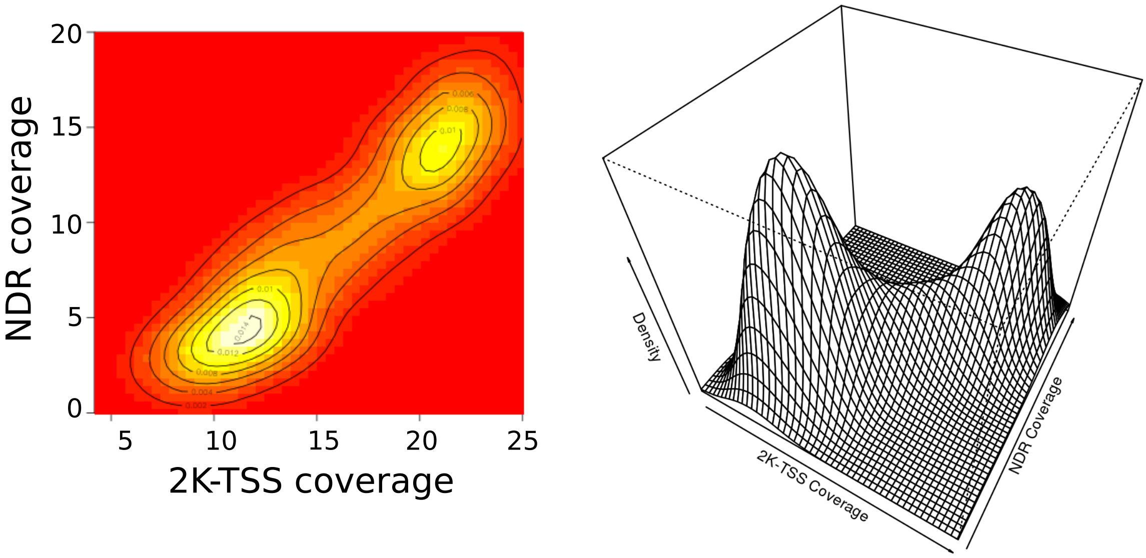

Figure 3.9: Kernel density estimation of the two features used to predict expressed genes from the 1000 highest and lowest expressed genes [33].

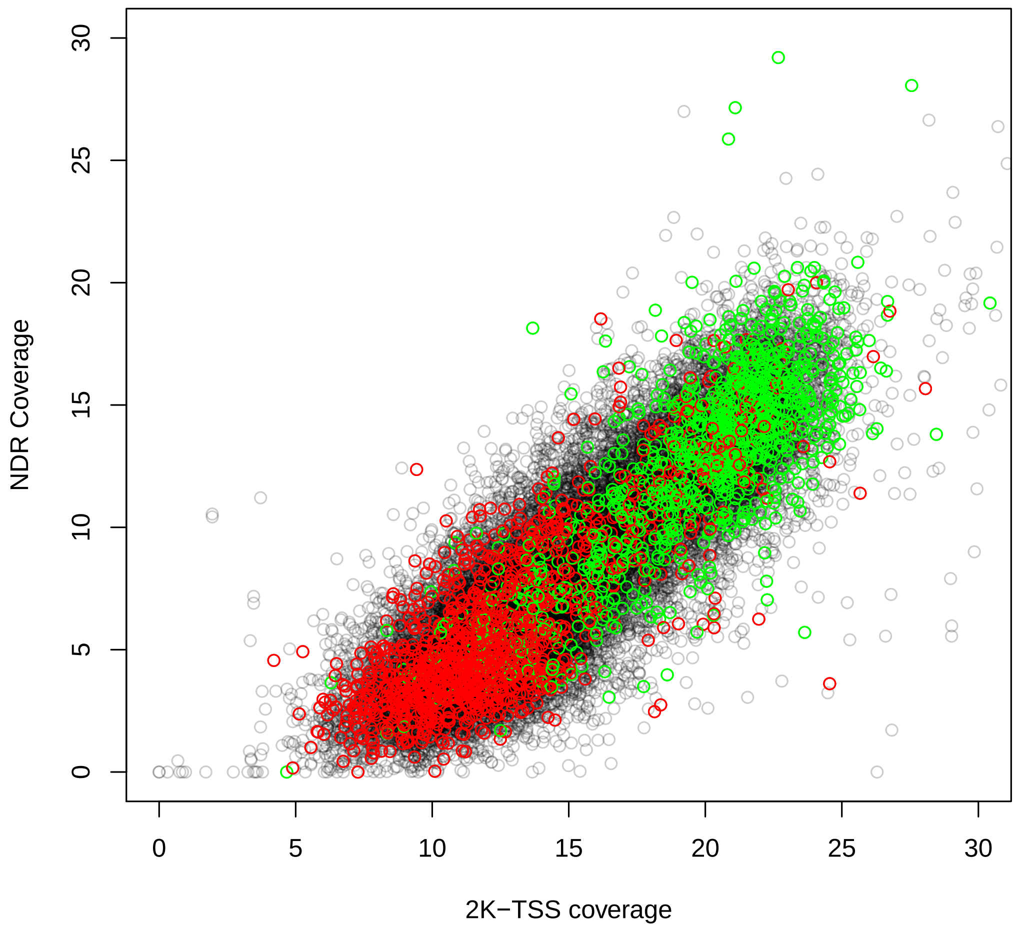

Figure 3.10: Scatter plot of the two features for every transcription start site. Top1000 and Bottom1000 genes are marked in red and green, respectively. All other genes are plotted in gray.

3.3.2. *Machine learning*

To predict the expression status of single TSSs, we applied machine learning on the two features. A random subset of 300 housekeeping genes [52] and 300 genes unexpressed in FANTOM5 dataset [53] was used to train support vector machines for 1000 iterations (each with a different random gene subset). Expression status was predicted when prediction consent was >75% in all 1000 iterations.

3.3.3. *Accuracy*

We again used the cfRNA dataset by [33] to test the accuracy of the expression status prediction. Of the Top 100 and Bottom 100 expressed genes, 91% were predicted correctly, while the accuracy dropped a little when using the Top 1000 and Bottom 1000 expressed genes (83%) (figure 3.11). Detailed performance characteristics are displayed in table 3.1.

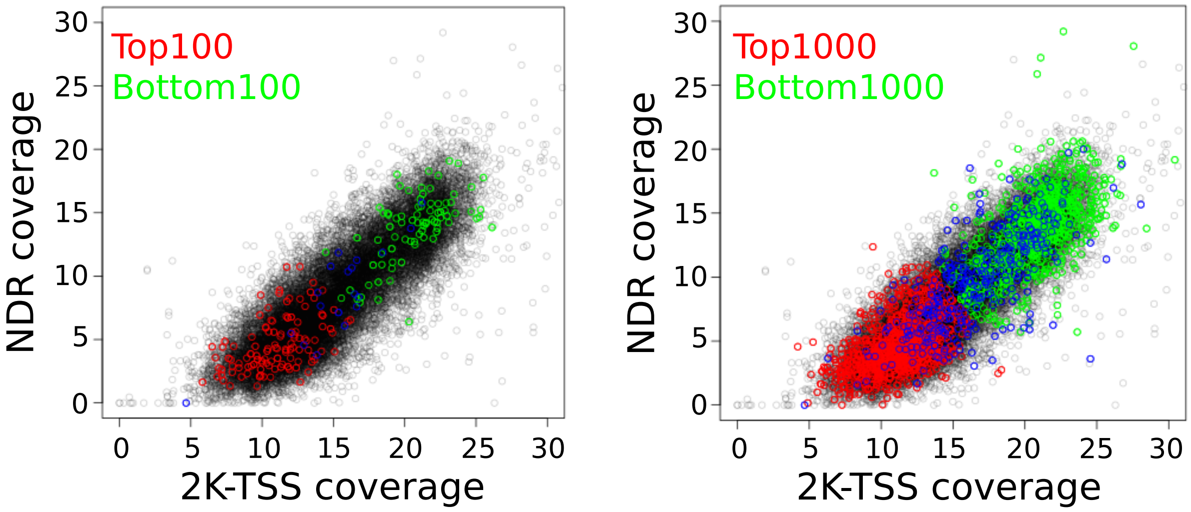

Figure 3.11: Prediction of Top100/Bottom100 and Top1000/Bottom1000 genes showed accuracies of 0.91 and 0.83, respectively.

Table 3.1: Performance characteristics of the expression prediction algorithm using various test sets.

Test set	Sensitivity	Specificity	Accuracy	Precision	F1-score
Top100	0.91	0.91	0.91	0.94	0.92
Top1000	0.81	0.86	0.83	0.88	0.84
Top5000	0.78	0.73	0.76	0.77	0.77
All (FPKM:1)	0.72	0.68	0.70	0.72	0.72
All (FPKM: 0.44)	0.69	0.72	0.71	0.79	0.74

3.3.4. *In-silico dilution series*

In cancer patients, cell-free DNA is a mixture of cfDNA from hematopoietic cells and from tumor tissue. In-silico dilution series were performed to identify a minimum signal fraction, which can still be used to infer the expression status. To this end, the distribution of both features for the Bottom 1000 genes was modeled as Gaussian distributions and random numbers of these distributions were added to the feature data of the Top 1000 genes at varying degrees (figure A.1).

Subsequently, for every dilution between one and 100%, accuracy of the expression prediction was measured. At a dilution of 75% the accuracy was still 70% (figure 3.12).

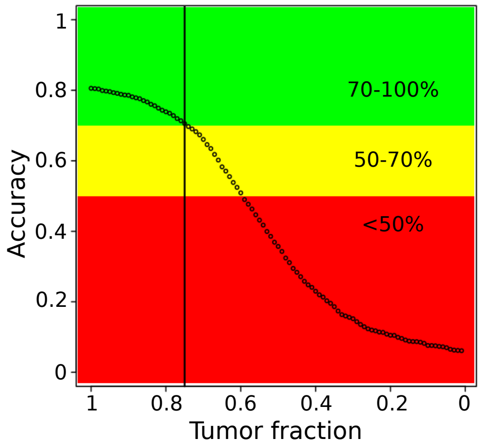

Figure 3.12: Prediction of diluted feature sets showed an accuracy of 70% at a dilution of 75%, however, accuracy declined rapidly in higher dilutions.

3.4. Quantitative relationships

3.4.1. *Correlation analysis*

The TSS profiles of genes with varying FPKM content (as measured by data from Koh *et al.*[33]) suggest a relationship between gene expression and both coverage values (*i.e.* NDR coverage and 2K-TSS coverage). To further elucidate this relationship, correlation analyses were performed independently on both coverage parameters. While no significant correlation was found between the parameters and the direct FPKM values, a significant correlation was found with the FPKM percentiles (Pearson correlation coefficients: 2K-TSS coverage: -0.356, p<2.2x10^{-16}, NDR coverage: -0.327, p<2.2x10^{-16}, see figure 3.13).

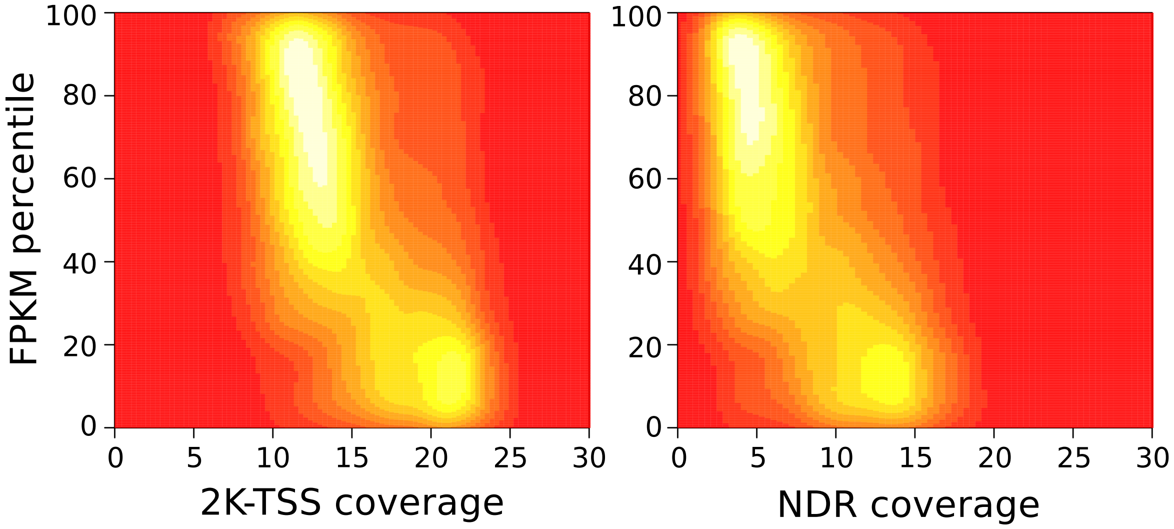

Figure 3.13: Correlation analyses of FPKM percentiles as measured by data from Koh *et al.*[33] and 2K-TSS coverage and NDR coverage, respectively, show a statistically significant) negative correlation

3.4.2. *Semi-quantitative analysis*

For semiquantitative analyses, genes were ranked by gene expression and split into deciles. Next, means of both, 2K-TSS and NDR coverage were plotted for every decile next to the raw data. Although variation is very high, on average, the relationship between gene expression and coverage is obvious (see figure 3.14).

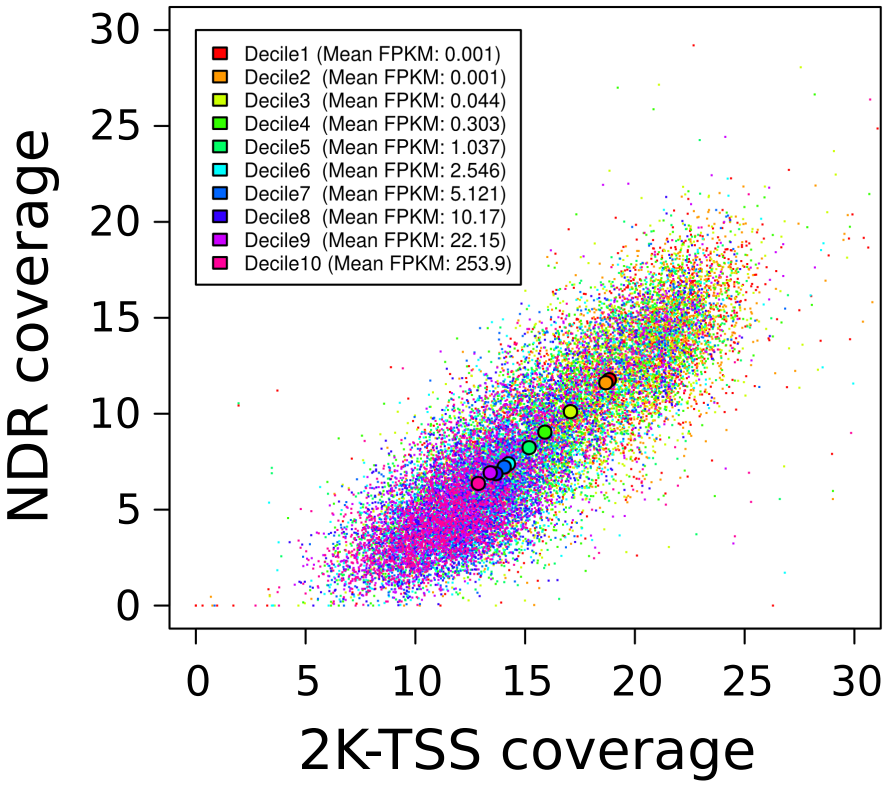

Figure 3.14: Means of coverage parameters from genes grouped into deciles based on their expression show a quantitative relationship between gene expression and promotor coverage.

3.4.3. *Binned data*

To further confirm this relationship we group every TSS into bins according to their location on the scatter plot. The scatter plot was divided in 30x30 fields and FPKM percentiles for every gene in the respective field was averaged. While this initially looked noisy (figure 3.15, left panel), after removing fields where 10 or less data points were available, the quantitative relationship is very clear (figure 3.15, right panel).

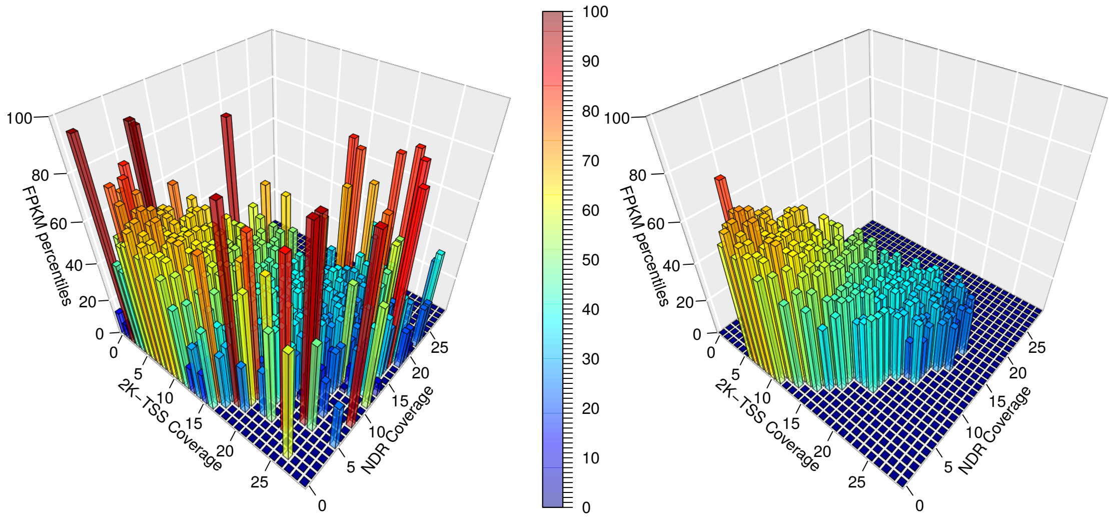

Figure 3.15: When genes are grouped into bins on the scatter plot and their FPKM percentiles averaged the quantitative relationship becomes clear. After removing bins with 10 or less data points this relationship is even more pronounced.

3.4.4. *Multiple regression*

Ideally, this relationship could be directly exploited to predict the percentile of any TSS analyzed. To see whether this would be possible, a two-dimensional linear model was fitted to the data (figure 3.16). While F-statistics for the model were statistically significant (p<2.2 x 10^{-16}), the model had a standard deviation of 26.88 percentiles and only 13.3% of the variance of the data could be modelled via the regression model (multiple R^2).

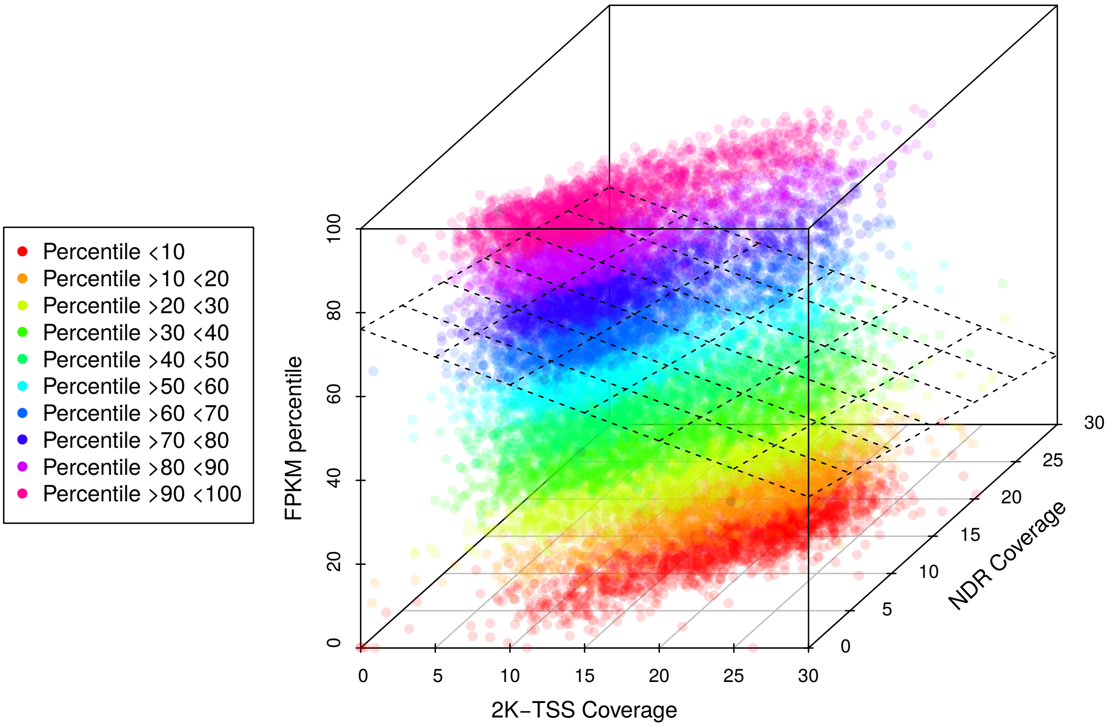

Figure 3.16: A linear model was fitted to the data in order to predict the FPKM percentile from the coverage data. The plane depicts prediction based on the multiple regression model.

3.5. Pattern correlation

Since expressed genes leave a very distinct coverage pattern around transcription start sites (figure 3.4), a possible alternative for expression prediction could be the correlation of the coverage pattern of a single gene to the pattern obtained from housekeeping genes. As a model for expressed genes, the mean of the coverage pattern of housekeeping genes [52] were used. Pearson correlation values for the Top1000 genes were slightly above 0 (Mean Pearson correlation coefficient Top1000 genes: 0.0055, Bottom1000 genes: -0.0032) but the distributions of Top1000 and Bottom1000 genes overlap greatly when looking at the whole 1000bp around the TSS (figure 3.17).

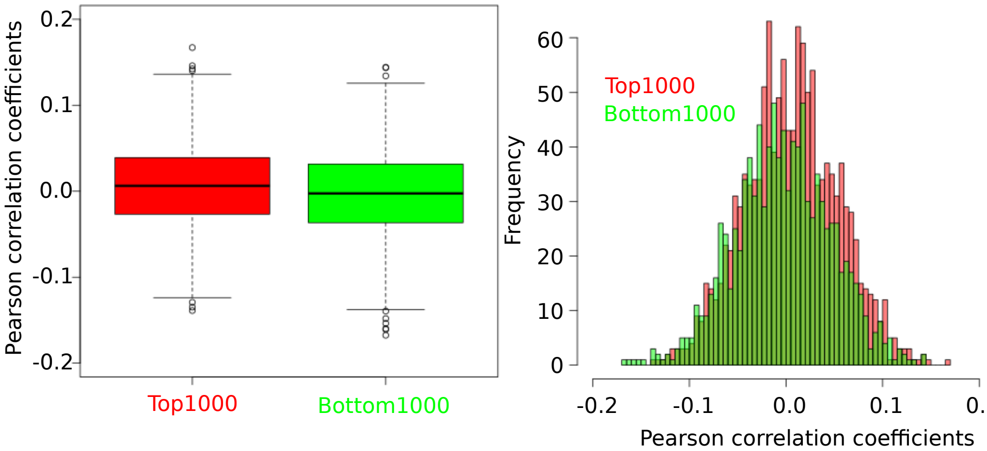

Figure 3.17: Using Pearson correlation to mean coverage values of housekeeping genes no difference was found between Top1000 (red) and Bottom1000 genes (green) when comparing the whole 1,000bp around the transcription start. Overlaps in histogram are shown in dark green.

Correlation analysis is improved a lot when the analysis is focused at the central 200bp around the transcription start. While correlation coefficients vary a lot more, the difference between Top1000 and Bottom1000 genes is markedly more pronounced (figure 3.18, Mean Pearson correlation coefficient Top1000 genes: 0.285, Bottom1000 genes: 0.040).

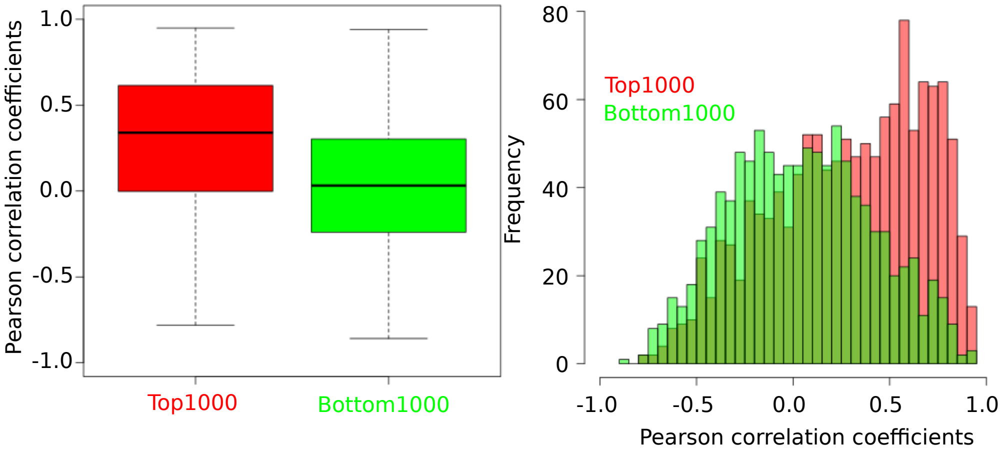

Figure 3.18: Using Pearson correlation the difference in mean correlation coefficients is more pronounced when focusing on the central 200bp around the transcription start. Overlaps in histogram are shown in dark green.

3.6. Tumor samples

In order to investigate whether transcription status can be derived from plasma samples of patients with cancer, we sequenced two samples of patients with metastatic breast cancer

(B7 and B13) at high coverage (411 and 455 mio. reads, respectively). Additionally, the transcriptome of the primary tumors was sequenced using the Ion Proton, in order to compare expression predictions to actual data.

3.6.1. *Single-nucleotide variants (SNVs)*

Since we generated a total coverage of approximately 20-23, single-nucleotide variants of both plasma samples were analyzed. Since the coverage is rather low, only somatic mutations with high frequency should be detected alongside germline sequence variations. In sample B7_1 approximately 5 mio. SNVs were identified whereas in sample B13_1, roughly 4 mio. variants were found. By filtering putative unimportant or germline variations and focusing on variants already seen in the cosmic database, 118 and 85 variants remain, respectively (see table 3.2).

Table 3.2: Results of single variant analyses from whole-genome sequencing. Initial variant calls are reduced in various steps to identify possible artifacts or germline variants. AF denotes allele frequency of healthy individuals.

Sample	**Total SNVs**	**Exonic SNVs**	**No synonymous**	**AF <1%**	**No SegDup**	**Cosmic**
B7_1	4,972,013	43,088	25,771	16,072	3,562	118
B13_1	4,026,013	28,934	17,198	9,926	2,547	85

3.6.2. *Copy number variants*

Copy number alterations were analyzed using read-depth analyses of 50,000 genomic bins. Copy number alterations of varying size were identified. The most prominent copy number alterations are a high-level gain on chromosome arm 1q and a focal amplification of a region containing the *CCND1* gene in B7 and high-level focal amplifications of regions containing *FGFR1* and *ERBB2* in B13 (figure 3.19). CNA profiles were also generated for DNA obtained from tissue biopsies of the respective primary tumors and correlation analysis showed high concordance in CNA levels for the respective samples. (see figure A.2).

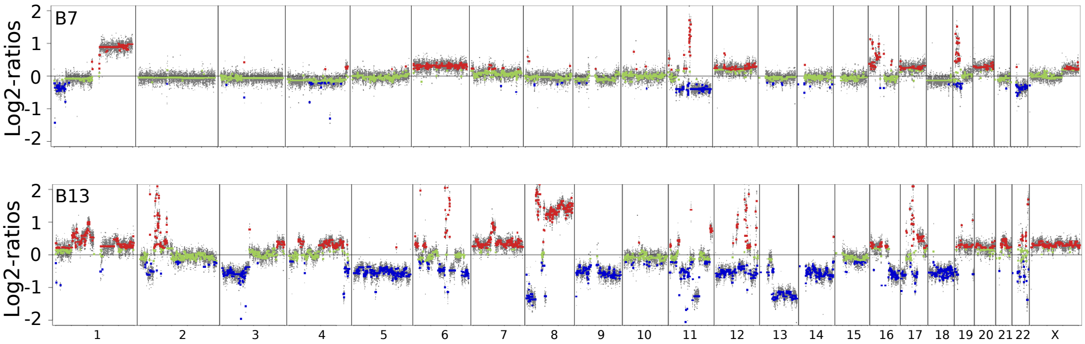

Figure 3.19: Copy number analysis reveals substantial copy number variation along the genome for breast cancer samples B7 and B13.

3.6.3. *Focal amplifications*

A recurrent type of somatic copy number alterations are amplifications of small regions (focal) which can reach high copy numbers. Often, tumor driver genes reside within these focal events and are amplified to very high copy numbers [67]. Here, focal events of plasma samples of both B7 and B13 were analyzed in order to check for possible amplified tumor driver genes.

3.6.4. *Insert size*

Via sequencing of both ends from a fragment, the initial fragment lengths can be reconstructed. By analyzing insert sizes on gained regions (chr1q in B7 and chr8q in B13), lost regions (chr11q in B7 and chr13q in B13) and copy number neutral regions (chr15 for both samples), size distributions of DNA from hematopoietic origin can be compared to DNA from the tumor. Fragments from gained regions should be enriched for tumor DNA while fragments of lost regions should be relatively enriched for normal (non-tumor derived) DNA. Fragment lengths of the different regions look distinct for the different regions in B7 (figure 3.20), but not for B13 (figure 3.21). In sample B7, fragment length analysis suggests that tumor-derived DNA has a larger portion of dinucleosomal-peaks, since dinucleosomal peak is higher for fragments from chromosome 1q, which should be enriched for tumor DNA.

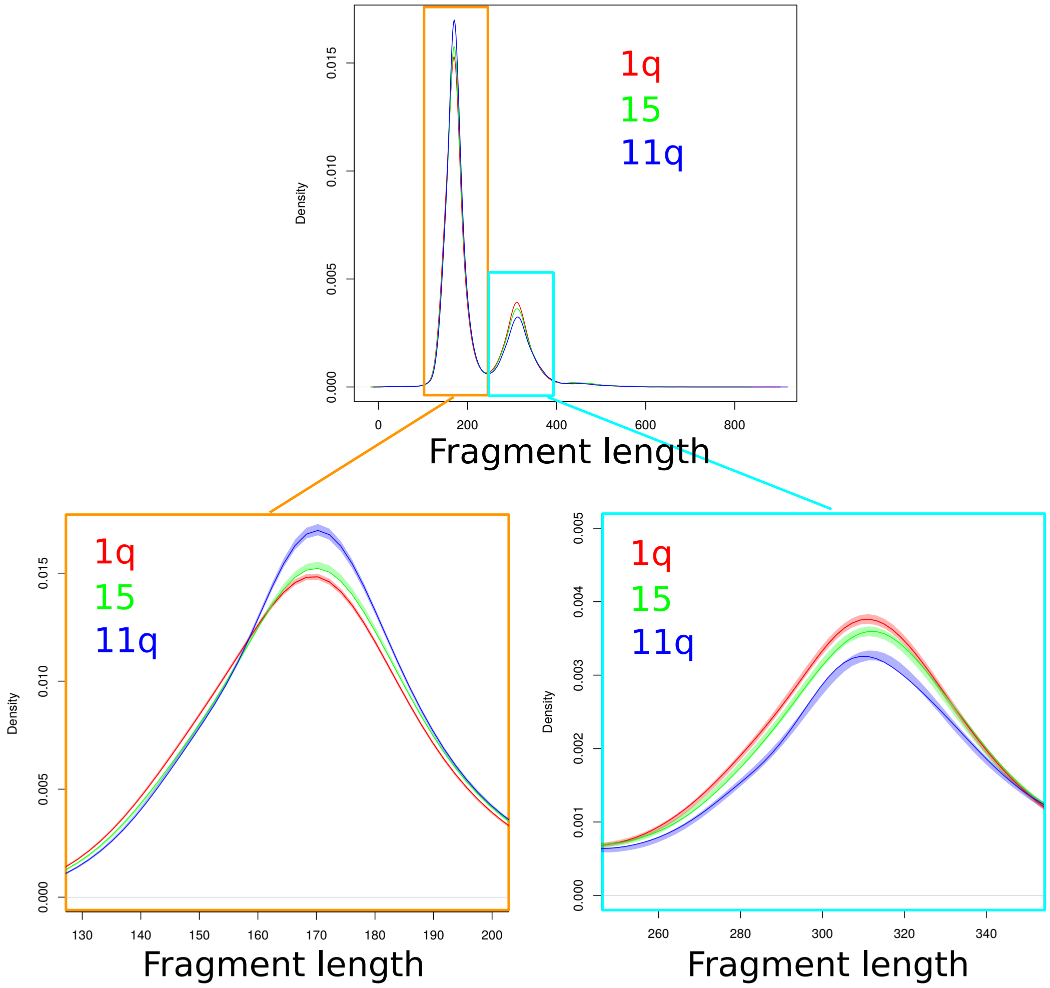

Figure 3.20: Fragment lengths of cfDNA fragments of sample B7. Fragment lengths from gained regions (1q, red) look different than fragment lengths from lost regions (11q, blue). Fragment lengths from copy number neutral regions (15, green) are between both distributions. Shaded areas represent 95% confidence intervals.

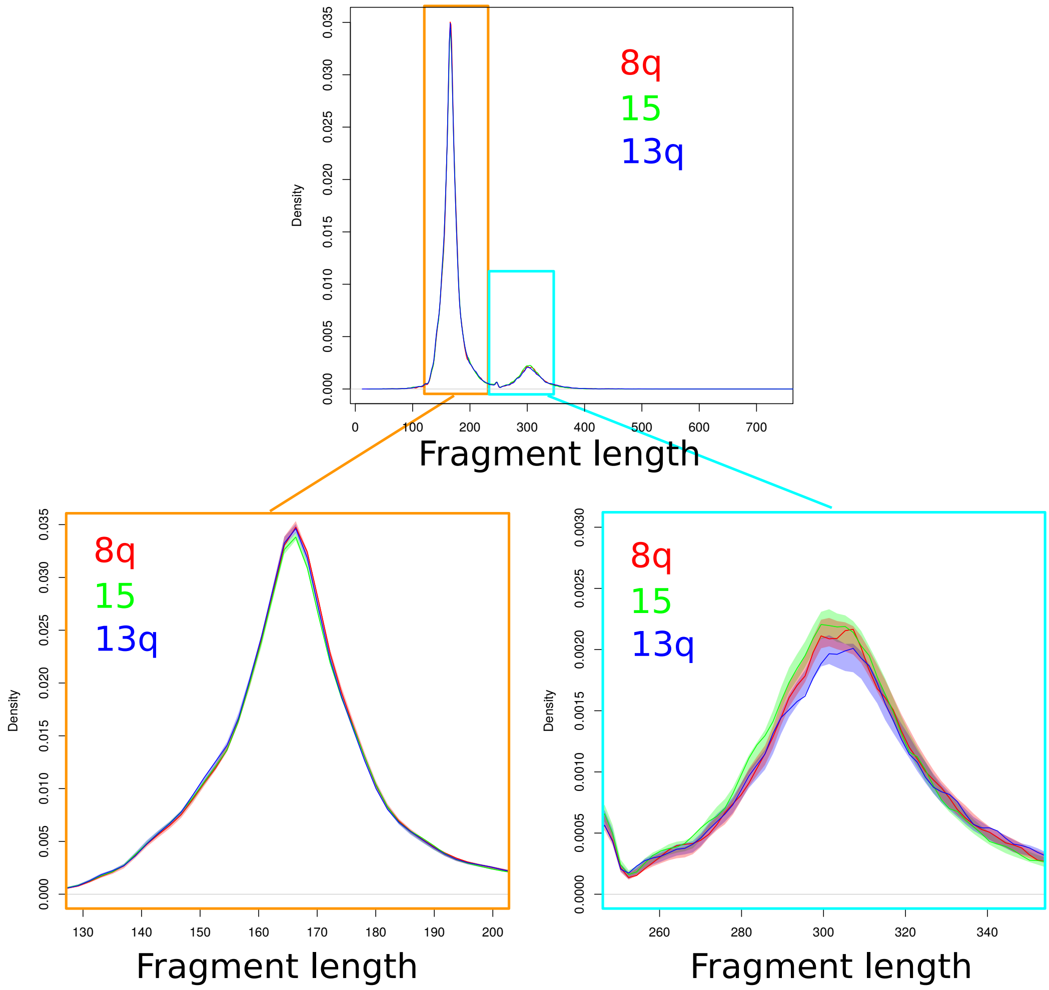

Figure 3.21: Fragment lengths of cfDNA fragments of sample B13. Fragment lengths from gained regions (8q, red) look very similar to fragment lengths from lost regions (13q, blue) and copy number neutral regions (15, green). Shaded areas represent 95% confidence intervals.

3.6.5. *TSS profile*

Since housekeeping genes should be expressed in every tissue, a coverage profile of the housekeeping genes versus unexpressed genes from the FANTOM5 project were established for both, B7 and B13 (see figure 3.22). The signal of housekeeping genes and unexpressed genes resembles the patterns derived from non-cancer controls and thus again indicate the nucleosome occupancy difference between expressed and unexpressed genes.

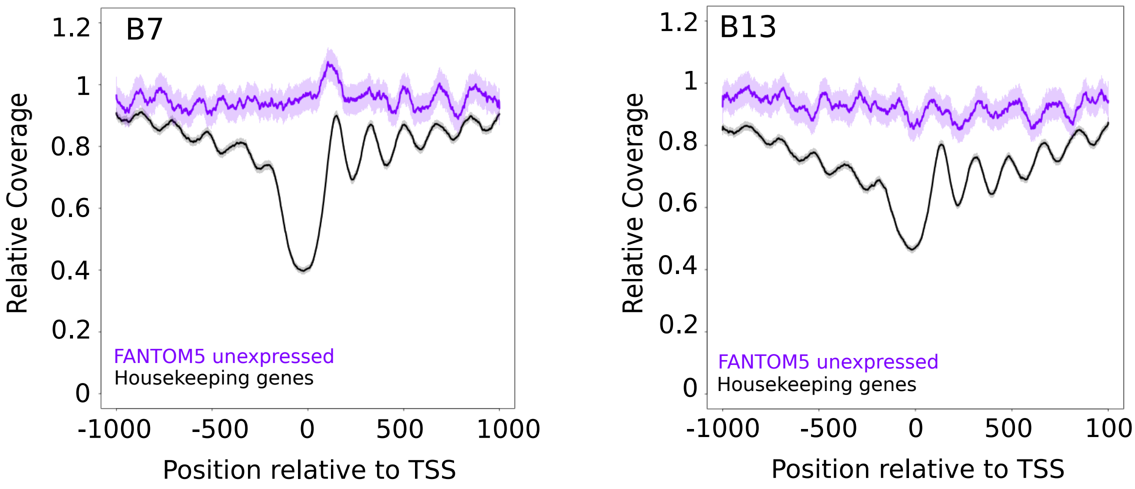

Figure 3.22: The coverage profile around transcription starts of housekeeping genes [52] in comparison to genes which are unexpressed in all tissues (as determined in the FANTOM5 project [53]) shows a distinct pattern in both breast cancer patients.

3.6.6. *Tumor fraction*

Cell-free DNA in a tumor patient is always a mixture of DNA from the tumor (ctDNA) and DNA from hematopoietic cells. In early disease stages, the fraction of tumor-derived DNA can be very low while it usually is higher in advanced/metastatic disease stages. ABSOLUTE [66] was used for tumor fraction estimation from copy number segments. This analysis compares copy number ratios to model karyotypes by maximum likelihood and can then estimate the contaminating normal fraction. Here, B7 was estimated to have a tumor fraction of 45%, while B13 was estimated to have a higher tumor fraction of 72% (see figure 3.23).

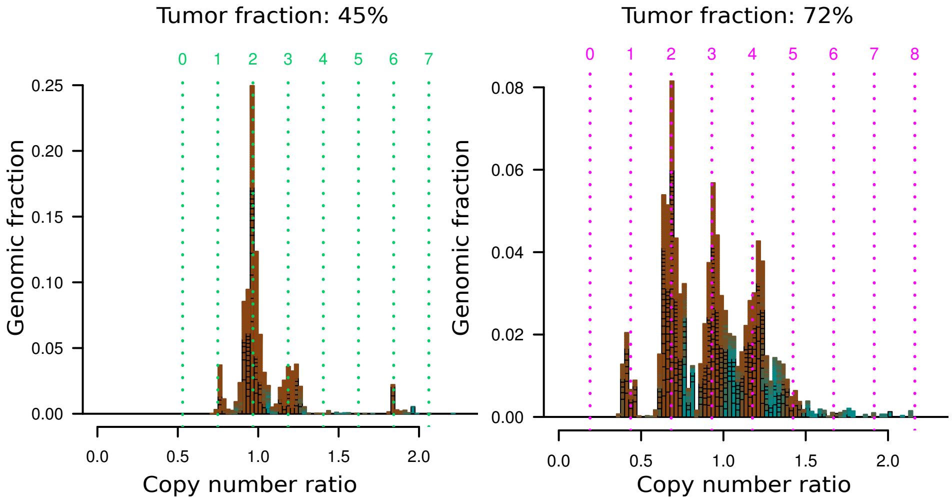

Figure 3.23: Tumor fraction estimation by ABSOLUTE from sample B7 (left) and B13 (right). Numbers in green and purple, respectively, denote absolute copy numbers estimated to derive the tumor fraction.

3.6.7. *Relative tumor fraction*

By in-silico dilution a minimum tumor fraction of 75% was established in order to still classify genes as expressed or unexpressed at 70% accuracy. Since both breast cancer samples do not reach this level throughout the genome, we focused on regions with copy number gains. These regions should represent higher tumor fractions, due to the contribution of more copies than the healthy cells.

The relative tumor fraction thus depends on the copy number ratio and the overall tumor fraction.

The true copy number of a region ($\mathrm{cp_i}$) can be calculated from the $\log_2$-ratio of this region $\mathrm{lr_i}$ and the overall tumor fraction (tf).

$$cp_i = \frac{2 * 2^{lr_i} - 2(1 - tf)}{tf} \tag{3.1}$$

Subsequently, the relative tumor fraction $\mathrm{rtf_i}$ can be calculated:

$$rtf_i = \frac{tf * cp_i}{tf * cp_i + (1 - tf) * 2} \tag{3.2}$$

Using these equations, the relative tumor fraction of every (genome-wide) tumor fraction and a respective copy number ratio can be calculated (see figure 3.24).

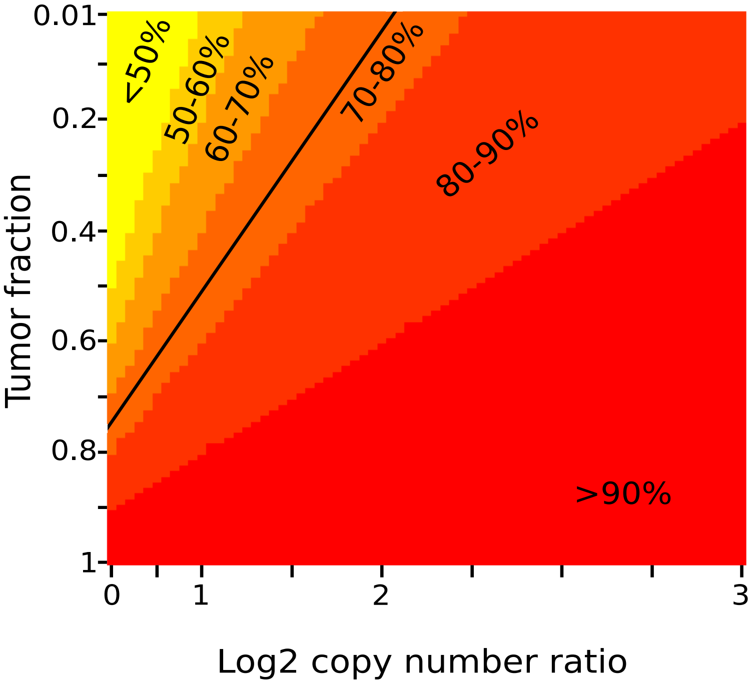

Figure 3.24: For every possible (genome-wide) tumor fraction and a wide range of copy number $\log_2$-ratios a relative tumor fraction is depicted. The black line indicates a relative tumor fraction of 75%.

3.6.8. *Tumor RNA-Seq*

In order to evaluate gene expression predictions, RNA-Seq from the primary tumor material was done for both patients. Since in-silico dilution showed that a relative tumor fraction of >75% was needed for the gene expression prediction to work, the Top100 expressed genes in regions with high relative tumor fraction were extracted (chromosome 1q for patient B7 and gained regions on chromosome 8 for patient B13, respectively).

In B7 these genes showed an enrichment in gene ontology hits, which were associated with nucleosome formation, while no significant enrichment for any gene ontology term was found for the Top100 genes of patient B13.

For pathway enrichment, the Top100 genes of patient B7 showed a significant enrichment of the telomere maintenance pathway, while the Top100 genes of B13 showed significant enrichments in pathways connected to protein biosynthesis.

3.6.9. *Expression prediction in tumor samples*

Focal amplifications:

Since focal amplifications can reach high copy numbers, expression prediction should be feasible in these regions even if the (genome-wide) tumor fraction of a sample is low. In B7, a region on chromosome 11, harboring *CCND1* was focally amplified which is a common alteration in breast cancer [67]. B13 harbors several focal amplifications including an alteration of *ERBB2* on chromosome 17 as well as an amplification on chromosome 8 which contains the *FGFR1* gene. Gene expression was predicted and the FPKMs of the genes (obtained by RNA-seq of the primary tumor) were analyzed and compared. Genes which were predicted to be expressed had statistically significant higher expression values than genes predicted to be unexpressed 3.25.

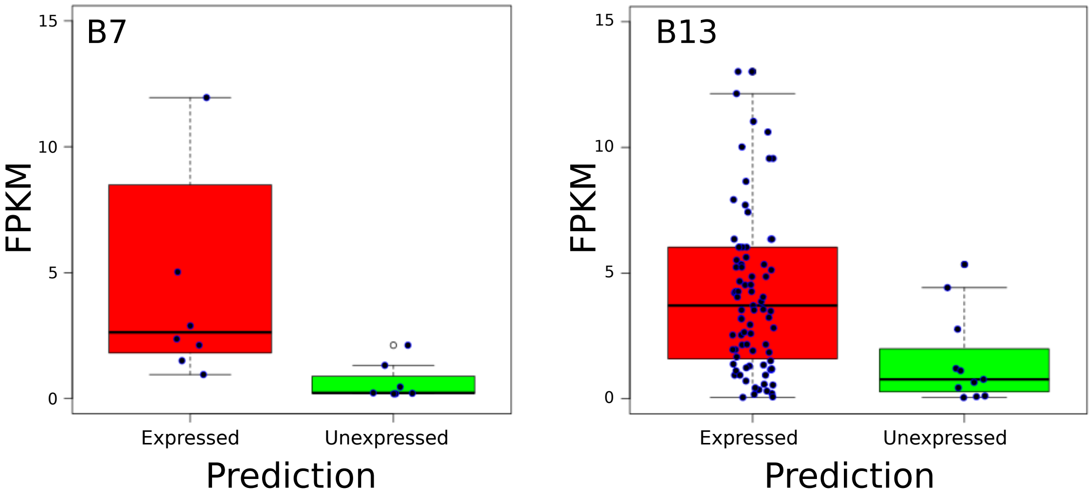

Figure 3.25: FPKMs for genes predicted to be expressed or not expressed in focal amplifications of 11q13.3 (15 TSS in 15 genes including *CCND1*; n expressed = 8; n unexpressed = 7) in B7 (left) and in both 8p11 (39 TSSs in 31 genes including *FGFR1*) and 17q12 (59 TSSs in 46 genes including *ERBB2*) (right) in B13 (n expressed = 87; n unexpressed =11). Blue dots represent genes located in the amplicons. Outliers including *CCND1* (FPKM of 50 in B7) and *ERBB2* (FPKM of 15 in B13) are not shown due to scaling. The differences were statistically highly significant (One-sided Mann Whitney U tests; B7: mean expressed : 9.7 FPKM, sd expressed : 17.0, mean unexpressed : 0.7 FPKM, sd unexpressed : 0.8, *p=0.003*; B13: mean expressed : 5.7 FPKM, sd expressed : 9.7, mean unexpressed : 1.5 FPKM; sd unexpressed : 1.8, *p=0.001*).

Top100 genes in amplifications:

Subsequently, the Top100 genes of amplified regions were analyzed. These regions included the amplification of chromosome 1q in sample B7 and chromosome 8q plus an additional region on the short arm of chromosome 8 which includes the *FGFR1* gene. Expression data of genes within these regions were ranked and the expression prediction status recorded, The correct prediction was achieved in 86.1% (B7) and 88.1%, respectively.

When looking at the Top100 expressed genes in all gained regions of B13 (*i.e.* $\log_2$-ratio >0.2), still 78% were correctly classified as expressed.

Isoform specific prediction:

Since every transcription start site is predicted independently, isoforms from the same gene coming from distinct TSSs might differ in their coverage profile. As a starting point *ERBB2* was analyzed in B13, which harbors an amplification containing this gene. Due to the amplification, most of the DNA in this region should originate from tumor cells. There are obvious differences when comparing the averaged profile of both isoforms of this gene to healthy controls (figure 3.26).

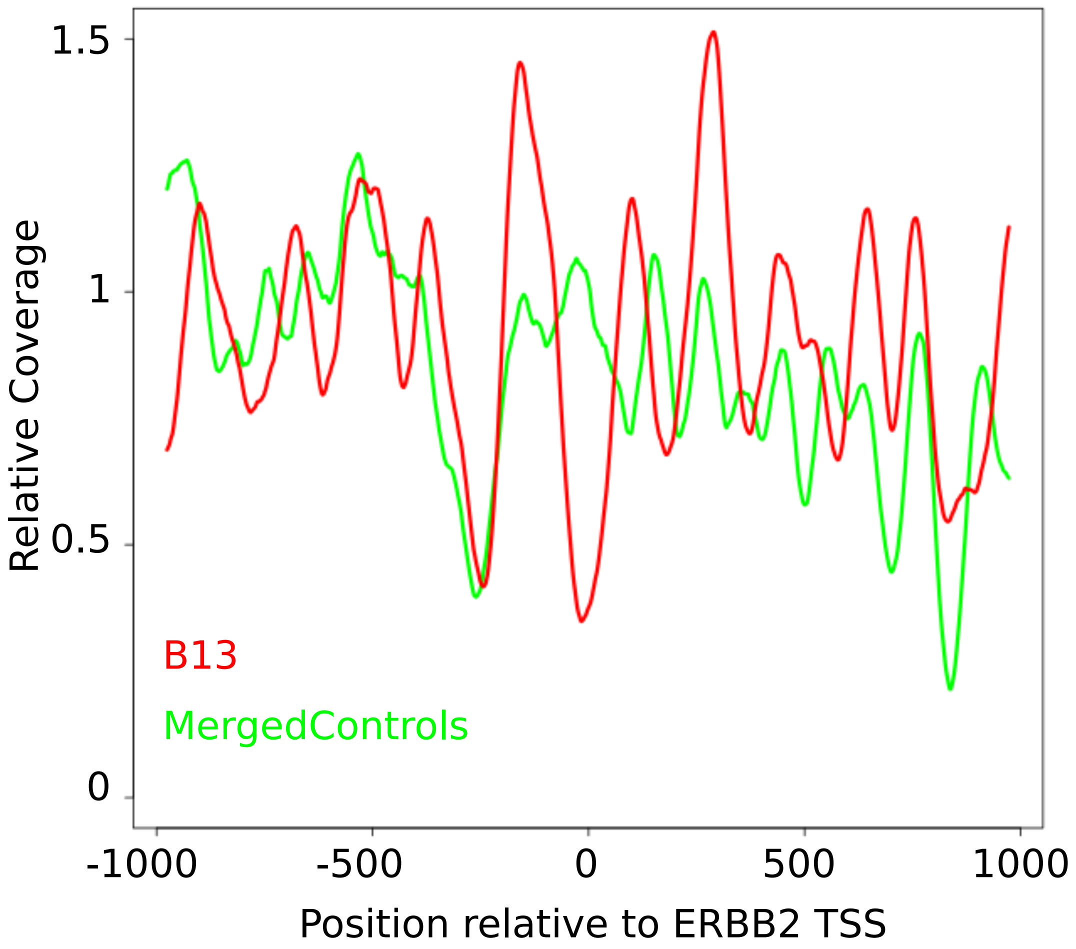

Figure 3.26: The coverage profile of *ERBB2* looks differently for B13 and healthy controls. Especially at the actual transcription start, coverage of B13 drops dramatically, thus likely representing the nucleosome depleted region. Both signals show a reduced coverage 250bp before the TSS. Since there is no gene within that range of either of the two TSS, this might constitute a site with very low nucleosome affinity.

For the detection of isoform specific expression and deduction, the difference between merged healthy controls and the tumor samples were calculated for every TSS and for both parameters. Overexpressed genes or isoforms should deviate more in the negative direction and no change is expected for genes with normal expression. Differences of both parameters were subsequently combined to a euclidean distance. These distances were calculated for both isoforms of *ERBB2* and interestingly one isoform (NM_004448) deviated a lot more from the healthy controls than the other (NM_001005862) (figure 3.27). This was confirmed by RNA-seq, where isoform NM_004448 had 11.4 FPKM, while isoform NM_001005862 was expressed at 4.4 FPKM. The same approach was used to analyze 9 isoforms of *FGFR1* coming from two distinct TSSs. Distance analysis also correctly identified the TSS giving rise to the isoforms with higher expression (TSS1: 3 isoforms, FPKM sum: 6.5; TSS2: 6 isoforms: FPKM sum 3.0).

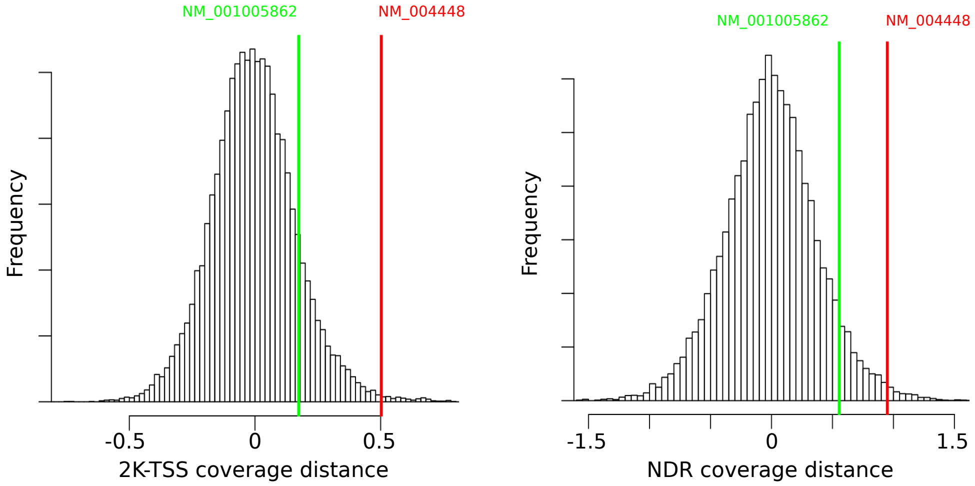

Figure 3.27: Calculating euclidean distances from the normalized data of healthy controls to the respective TSS in tumor samples suggests that isoform NM_004448 is higher expressed than isoform NM_001005862. This was confirmed by RNA-seq

The analysis was then extended to every gene in the focal amplifications of samples B7 and B13. Of 93 genes in those amplification, 8 genes had more than 1 TSS which gave rise to isoforms with at least 2 FPKM difference (including *ERBB2* and *FGFR1*). The TSS leading to the higher expression of isoforms was correctly predicted in 7 of these genes (figure 3.3).

Table 3.3: Distances between healthy controls and tumor samples B7 and B13 were calculated for 8 genes which have several TSSs which give rise to isoforms having a FPKM difference of at least 2. The TSS leading to higher expression had larger distances in 7 of the 8 genes.

Sample	**Gene**	**TSS**	**Isoforms**	**FPKM**	**Distance**
B13	*DDHD2*	chr8:38089008	2	4.85	0.21
		chr8:38089470	1	0.01	0.09
	GRB7	chr17:37894161	1	2.95	1.26
		chr17:37894575	1	1.59	0.91
		chr17:37895023	1	1.33	0.84
		chr17:37896219	1	0.14	0.63
	PPP1R1B	chr17:37784750	2	9.24	0.69
		chr17:37783176	1	3.77	0.89
	GSDMB	chr17:38074903	3	4.27	0.58
		chr17:38073793	1	<0.01	0.30
	ANK1	chr8:41522804	3	4.93	0.44
		chr8:41655140	4	0.35	0.11
		chr8:41754280	1	0.01	0.17
	ERBB2	chr17:37856230	1	11.37	1.07
		chr17:37844336	1	4.43	0.58
	FGFR1	chr8:38325363	3	6.53	0.62
		chr8:38326352	6	3.02	0.36
B7	*ANO1*	chr11:69924407	1	2.06	0.24
		chr11:69931515	1	0.06	0.06

4. *Discussion*

The aim of this thesis was to deduce gene expression from plasma DNA by applying coverage analysis to a set of non-cancer controls at the transcription start site of genes, known to be universally expressed and comparing them to genes which are universally unexpressed. In a later step, the approach was applied to plasma DNA samples from tumor patients to see whether this can be used to infer functional information about the primary tumor.

4.1. Nucleosome association of plasma DNA

In a first step, the association of ctDNA to histone proteins in the form of nucleosomes was investigated. To this end, fragment lengths of molecules from mitochondrial DNA were compared to the lengths of molecules from nuclear DNA (see figure 3.1). Since mitochondrial DNA packaging works differently than packaging of nuclear DNA [68], fragment lengths after *in vivo* DNA digestion are expected to differ. Indeed, fragment lengths of nuclear DNA show a distinct mode at approximately 166bp which is consistent with the length of DNA wrapped around a histone-octamer (147bp) plus additional bases from the linker region. Conversely, mitochondrial DNA fragments exhibit a vastly different pattern which shows a broader distribution of fragment lengths and no bimodal structure. This finding is consistent with previous reports [24].

Moreover, association of ctDNA with histone proteins was shown by analyzing coverage peaks in a region, which has been described to contain a well-ordered array of preferential nucleosome formation on chromosome 12 [48]. Here, sequencing coverage (as a proxy to the representation of fragments in the circulation) of cell-free DNA in a set of non-cancer controls was compared to the sequencing coverage obtained from MNase-seq experiments from the cell line GM12878 (see figure 3.2). A high correlation of both signals in this region confirms the putative association of cell-free DNA to nucleosomes.

These analyses suggest that cell-free DNA indeed is associated with histone proteins in the form of nucleosomes (*i.e.* DNA wrapped around an octamer of histone proteins), which is in line with prior reports [9, 10, 24].

4.2. Coverage analysis at transcription start

In a next step, sequencing data of 104 non-cancer controls was merged and the mean coverage signal of 3,509 housekeeping genes was compared to the mean coverage signal of 670 genes reported to be unexpressed in all tissues according to the FANTOM5 database [52, 53]. While expressed genes show a lower overall coverage in a region 2,000bp around the transcription start site, they also show a pronounced drop in the coverage signal in a region approximately starting from -150bp before the transcription start to 50bp after the transcription start. This is in line with reports of results from MNase-seq studies, which also showed a lower representation of reads at the transcription start site until approximately 1,000bp into the gene body [30]. MNase-seq should yield comparable results to the sequencing of plasma DNA, since DNA is also preferentially digested in linker regions between nucleosomes than DNA that is bound to histones. However, in MNase-seq, this is done *in vitro*, while DNA digestion happens *in vivo* in plasma DNA.

Moreover, the coverage pattern around the transcription start site shows a wave-like pattern from the transcription start itself, extending in both directions. This might be due to the restricted possibilities of nucleosome formation, since active processes maintain the nucleosome depleted region [29]. As nucleosome formation occurs preferably in regions having a specific sequence composition [25], possible nucleosome forming sites are limited [30].

The analysis of coverage profiles of genes reported to be highly expressed in cfRNA compared to genes which are expressed at low levels showed large differences in the TSS region [33]. Moreover, grouping genes from the cfRNA analysis into several subgroups, the profile seems to depend on the expression level of the respective genes, with highly expressed genes showing a more pronounced drop in average coverage.

Taken together, the data show a distinct pattern in the coverage profile of expressed genes when compared to unexpressed genes and thus coverage information might be sufficient to detect gene expression.

While there is evidence that points to preferential DNA digestion by DNase [30], there might be alternative explanations for the depletion of DNA from the NDR region in the cell-free DNA. Although this DNA might not be present in the liquid compartment of the blood circulation, it might still be present in different compartments, *e.g.* exosomes [69]. However, due to the similarity of the coverage patterns of *in vitro* digested DNA (MNase-seq of GM12878, see figure 3.6) and cfDNA (see figure 3.5), preferential digestion seems to be the more parsimonious explanation.

4.3. Gene expression prediction

Since coverage profiles of expressed and unexpressed genes differed, features were selected in order to predict the expression status of each gene in the set of non-cancer controls. To this end, two features were tested:

- The mean coverage of the region from 1,000bp before the TSS to 1,000bp after the TSS (2K-TSS coverage)
- The mean coverage of the region from 150bp before the TSS to 50bp after the TSS (NDR coverage)

Both features were tested on the 1,000 highest and lowest expressed gene from cfRNA as reported by Koh *et al.* [33]. While the 2K-TSS coverage parameter showed a clearer separation, both features showed a large difference between the two gene sets (see figure 3.8). When analyzing both features together for the 1,000 highest and lowest expressed genes, the distribution of the data points shows a bimodal structure, with the Top 1000 genes having a lower signal in both, 2K-TSS and NDR coverage.

Next, a support vector machines (SVM) based prediction analysis was done, which used housekeeping genes (n=3,509) and genes shown to be unexpressed (n=670) in order to learn the distribution of the two coverage parameters. The prediction step was then evaluated against several gene sets in order to investigate the sensitivity and specificity of the approach. The sensitivity and specificity decreases with more genes being added to the test set, however, even for the 5,000 highest and lowest expressed genes, a sensitivity of 0.78 and a specificity of 0.73 was detected.

4.4. In-silico dilution

Since plasma DNA in a tumor patient is a mixture of healthy DNA and tumor-derived DNA, in-silico dilution was performed to see which tumor DNA fraction could still be used for gene expression prediction. To this end, data of the Top 1000 genes was mixed with random noise at varying degrees and sensitivity for all the dilutions was measured. At a dilution if 75%, the sensitivity was still 70% (see figure 3.12). While a total tumor fraction of 75% might only be available in patients with very late-stage metastasized cancer, regions having copy-number alterations might still be amenable to this approach, since here genomic regions are relatively enriched for tumor DNA.

4.5. Quantitative analysis

As the analysis of the coverage profile of genes split into subgroups suggested that the profile might allow a more nuanced quantitative inference of gene expression, this aspect was investigated further. Correlation analysis of the two features (2K-TSS coverage and NDR-coverage, respectively) to FPKM percentiles showed a good correlation (see figure 3.13, and furthermore grouping genes into deciles (based on the expression values obtained from Koh *et al.* [33]) showed expression dependent mean values for the two features (see figure 3.14).

Moreover, after binning genes into groups, depending on their location on the 2D-scatterplot and calculating mean gene expression values per bin, a quantitative relationship between the two coverage parameters and gene expression was noted.

However, the signal seems to be too noisy for predicting actual gene expression values per gene. This was shown by applying a multiple regression model on the data. While the F-statistics of this model were highly significant only 13% of the variation could be explained by gene expression, while the remaining 87% of the variation is due to noise.

These analyses suggest, that while some quantitative relationships between coverage and gene expression can be found, the coverage signal described above cannot be used to predict actual gene expression for a single gene in it's current state. A more detailed analysis and a more nuanced prediction of gene expression might be available by sequencing data with higher coverage. Also, sequencing a set of cell-free DNA of non-cancer control, each with high coverage, might lead to a more detailed quantitative result of expression prediction, since this would allow to get a glimpse on biological variation of the coverage profile on a per-gene basis.

4.6. Pattern correlation

Since the (averaged) coverage pattern of housekeeping genes around the transcription start is very distinct from the pattern of unexpressed genes (see figure 3.4) an additional method of gene expression prediction was investigated based on pattern correlation. Here, the mean coverage pattern of housekeeping genes was compared to the coverage pattern of the 1,000 highest and 1,000 lowest expressed genes (as determined by Koh *et al.* [33]) using correlation analysis.

Calculating the correlation in the whole 2,000bp window, the 1,000 highest expressed genes show only a marginally higher correlation coefficient. This might be due to the fact that the pattern is more stringent in the small region directly at the transcription start, since nucleosome positioning is more restricted here [30]. Moving further away from the nucleosome depleted region, nucleosome positioning may occur at several different places.

Hence, using only the central 200bp around the TSS for correlation calculation the signal is improved and the 1,000 highest expressed genes show higher correlation coefficients. However, correlation coefficients vary a lot and both gene sets cannot be separated as well as in the two coverage features used for the actual gene prediction. The high variation in correlation coefficients may be explained by the restricted size of the region. In a small genomic region, a single 60bp sequence has more effect on the correlation than a single sequence in a larger genomic region.

4.7. Tumor samples

To investigate, whether this approach can be used to inform about gene expression in ctDNA samples, plasma DNA of two metastasized breast cancer patients were analyzed. These samples exhibited a wide array of copy number alterations throughout the genome (see figure 3.19) and thus represented samples with a high tumor fraction. In fact, the total tumor fraction was estimated to be 45% and 72% for B7 and B13, respectively. While these tumor fractions are relatively high, tumor fractions of up to 90% have been reported already [10].

4.7.1. *Single Nucleotide Variants*

Single nucleotide variants were identified directly from plasma, however, since no corresponding germline DNA was available in order to detect possible somatic mutations, the analyses described here may be of limited use. Furthermore, the relatively low sequencing coverage does not allow for an accurate detection of variants with low mutant allele frequency, which might be expected due to the presence of non-cancer derived DNA in the blood circulation.

4.7.2. *Copy Number Alterations*

Extensive copy number alterations were identified in both plasma DNA samples. The plasma DNA of patient B7 shows a high-level amplification of the region around the gene *CCND1*, which is a common alteration in breast cancer tumors, as well as a approximately 6 copies of the long arm of chromosome 1, according to the ABSOLUTE algorithm [66].

Patient B13 harbors even more alterations, including a potentially druggable amplification of the *ERBB2* gene on chromosome 17 and a very high copy number of the long arm of chromosome 8, which includes the oncogene *MYC*. Furthermore, a focal amplification of the *FGFR1* gene was detected on the short arm of chromosome 8.

Copy number alterations were also analyzed for the respective primary tumors of both patients and generally resembled the plasma DNA copy number profiles. While the amount of tumor DNA in patient B7 seems to be similar to the amount of tumor DNA in the primary tumor samples, patient B13 seems to have a higher tumor fraction in the plasma than in the primary tumor.

4.7.3. *Tumor RNA-seq*

By RNA-seq the gene expression of the respective primary tumors was analyzed in order to compare those with gene expression predictions later on.

Since gene expression prediction is only possible in regions which show copy number gains, the Top100 genes of gained regions in both patients were extracted (chromosome 1q for patient B7 and gained regions of chromosome 8 in patient B13, respectively). Functional enrichment of these highly expressed genes showed enrichments for nucleosome formation and telomere maintenance in patient B7, which might explain the high tumor fraction of 45% in the peripheral blood of this patient.

In patient B13 no significant functional enrichment was found in the gene ontology terms, however, enriched pathways included processes that regulate protein biosynthesis and gene expression and thus might represent the active state and a high level of protein synthesis in these tumor cells.

4.7.4. *Fragment length analysis*

Through paired-end sequencing, it is possible to calculate the fragment length of the initial fragments which were used to create the library. While there is agreement on the modal structure of the fragment length distribution with the main mode around 166bp [20], some conflicting reports were published about the fragment length deviation of tumor-derived DNA in the plasma [22, 23]. Here, insert sizes of plasma DNA from both cancer patients were analyzed, and fragment length distribution of regions which should be enriched for tumor-derived DNA (*i.e.* regions harboring copy number gains) were compared to fragment length distributions of regions which should be enriched for non-tumor derived DNA (*i.e.* regions harboring copy number losses) and regions with no copy number alteration which is. This represents an analogous approach to an analysis which has been previously reported [21].

In patient B13 no difference in the fragment length distributions between these regions were identified. This can be explained in two ways: Either the tumor fraction and the non-tumor derived fraction of the plasma-DNA share the same fragment length characteristics, or there might be virtually no non-tumor derived DNA in the peripheral blood, since these fragments are expected to show a different size distribution [21].

Conversely, fragment length distributions of plasma DNA fragments differ in patient B7 in the different regions. Regions enriched with tumor DNA show a higher proportion of di-nucleosomal fragments than regions enriched for non-tumor derived DNA. This indicates that the kinetics of DNA digestion might differ, depending on the source of DNA.

Interestingly, no fragments smaller than 166bp were identified, although this has been reported [22]. This might be attributable to differences in the sequencing library preparation, which includes size selection steps in order to ensure proper sequencing.

4.7.5. *Coverage profile at TSS*

Since housekeeping genes should be expressed regardless of the tissue, the coverage pattern of housekeeping genes and genes known to be unexpressed were analyzed for both plasma DNA samples from tumor samples (see figure 3.22). This coverage pattern resembles those from non-cancer controls and thus suggests that DNA fragments in these samples are also associated to nucleosomes. Hence these samples should be amenable to gene expression prediction.

4.7.6. *Gene expression prediction in tumor samples*

In-silico dilution of the 1,000 highest and lowest expressed genes suggests that a signal reduction to 75% may still allow prediction of expressed genes. Hence, only regions which exhibited copy-number gains were used for gene prediction. This included genes in focal amplification (as defined in a previous report [65]), as well as broader regions with copy number gains (*i.e.* chromosome 1q for patient B7 and gained regions of chromosome 8 in patient B13).

Gene expression prediction of genes in focally amplified regions were compared to gene expression values obtained from RNA-Seq of the respective primary tumors. Here, FPKMs of genes predicted to be expressed were significantly higher than genes predicted to be unexpressed.

In a further step, the top 100 expressed genes (as measured by RNA-seq of the primary tumors) in regions with copy number gains were extracted and their expression status predicted from the coverage pattern in the plasma DNA. 86.1% of the genes were predicted to be expressed in B7 and 88.1% of these genes were predicted to be expressed in B13.

These analyses suggest that in certain regions of the tumor genome, gene expression can be predicted from plasma DNA alone. However, plasma DNA must exhibit a certain tumor fraction in order to ensure that the coverage signal can be used to predict expression status. If the total tumor fraction is less than 75%, regions with copy number gains can be used, since those may show a larger relative tumor fraction.

4.7.7. *Prediction of isoform specific expression differences*

Many genes in the human genome give rise to several mRNA transcripts which in turn can code for different proteins. While many isoforms of the same gene share the same transcript start, some isoforms use alternative starts.

Here, genes in regions exhibiting copy number gains were identified which have multiple transcription starts. In B13, eight genes were identified and in B7 only one gene have multiple TSS. When calculating distances in both features (*i.e.* 2K-TSS coverage and NDR coverage), transcription starts which give rise to transcripts with higher expression deviated a lot more from non-cancer control data than transcription starts which give rise to less expressed transcripts.

While these analyses are based on very few genes, it indicates that more active isoforms of genes might be predicted from coverage parameters as long as they use a different transcription start.

4.8. **Limitations**

Through detection of coverage differences at transcription starts, the data shown here suggests that inferences about gene expression may be possible directly from cell-free DNA analyses. However, in practice there might be limitations to this approach.

A large tumor fraction is needed, since cfDNA fragments from non-cancer cells create a background noise that clouds the signal. Here, only genomic regions exhibiting copy number gains were used in order to overcome this issue, but cancers without copy number alterations might not be amenable to this approach. Also, cfDNA samples of cancer patients in earlier stages which usually show lower overall tumor fractions are possibly not within reach of this method. Furthermore, some types of cancer consistently spread less DNA into the periphery (*e.g.* brain and renal cancers) [70].

Moreover, some biological processes may also lead to a lower representation of fragments from the NDR, although not leading to an increased expression. For example, RNA polymerases may bind to the promotor region (thus requiring nucleosomes to be depleted in this region), although they do not elongate a mRNA molecule [71, 72].

4.9. **Context**

While statical information (*e.g.* somatic point mutations, copy number alterations) about a tumors genome has been amenable to analysis from ctDNA for some time [9], the inference of expressed genes from read-depth analyses may allow for functional analyses of liquid biopsies.

The coverage signal caused by nucleosome occupancy has already been used before to inform about tissue-of-origin of cancers [73] and while this group also touches upon expression specific differences in their signals, the study presented here greatly expands about the inference of expression and the correlation to nucleosome occupancy.

Another study tried to discern between housekeeping genes and tissue-specific genes [74] based on exome enrichment data of 3 cfDNA samples. While they show some differences between ubiquitously expressed genes and tissue specific genes, the number of genes analyzed was very small. Exome enrichment data does not seem suitable for this analysis since the most important region (the region immediately before transcription start) is not sequenced. Also, their analysis only works for genes with a large first exon (so that it contains at least 3 nucleosomes).

4.10. **Outlook**

As there are great interests in non-invasive tumor monitoring, the approach described here adds an additional layer of information which might be accessed by the analysis of ctDNA.

However, generating 400 million reads per sample is still financially challenging and thus there are still obstacles to overcome to implement a useful application that benefits patients and informs the treating physician using this approach.

One way of improving the signal strength might be to enrich for promotor region of either every gene, or interesting subsets of genes in order to have high sequencing coverage in those regions. This may substantially decrease sequencing costs, however, it also introduces a potential bias, due to preferential hybridization of certain regions and additional PCR cycles needed to create a library that can be sequenced.

In addition, cell-free DNA from other sources might also give additional information about tumors.

Previous studies showed analyses from cell-free DNA from other body fluids including urine, stool, cerebro-spinal fluid, stool and pleural fluid [75]. Provided the DNA derived from these fluids are also associated to nucleosomes, there might be some more possibilities to infer gene expression.

Recently, DNA found in exosomal structures (*i.e.* small extracellular vesicles containing different molecules) attracted attention as possible biomarkers, especially since the membranous particle might be a means of horizontal gene transfer between cells [69]. This might be an especially rewarding topic for future research, since oncogenic transformation of susceptible cells due to transfection-like uptake of cfDNA has been proposed [76].

5. *Conclusion*

This study suggests that nucleosomes have a great influence on the coverage pattern produced when sequencing cell-free DNA in plasma and that this information can be exploited to infer the expression status of gene. This seems to hold true even for data from plasma DNA of cancer patients, where cell-free DNA is a mixture of DNA from the cancer and healthy tissue.

Thus, the study presented here, is, to the best of knowledge, the first general approach to infer expression from nucleosome occupancy and might pave the way for novel biological applications.

Especially metastasized cancer seems to be a suitable target since ctDNA fractions are generally high and should thus be amenable for these analyses. Apart from what has been shown here, several different scenarios may be exploited: Metastasized cancer genomes are variable due to tumor evolution and selection caused by treatment [8]. Here, information about gene expression may elucidate resistance mechanisms or general mechanisms about tumor evolution.

Apart from cancer, other types of disease also lead to increased cfDNA release including myocardial infarction [77], brain injuries and aging [78]. Since the cfDNA from other tissues are practically indistinguishable from blood cell derived cfDNA, only total amount of cfDNA were analyzed in several studies.

In summary, the coverage signal caused by nucleosome occupancy can be exploited to make inferences about the expression of genes and isoforms. This expands the use of circulating cell-free DNA and may elucidate basic mechanisms of gene regulation and expression variation in cancer and other diseases.

References

[1] Hanahan D and Weinberg R. **The Hallmarks of Cancer**. *Cell* 2000, **100**: 57–70.

[2] Stratton M, Campbell PJ et al. **The cancer genome**. *Nature* 2009, **458**: 719–723.

[3] Vogelstein B, Papdopoulos N et al. **Cancer Genome Landscapes**. *Science* 2013, **339**: 1546–1558.

[4] Burkhart DL and Sage J. **Cellular mechanisms of tumour suppression by the retinoblastoma gene**. *Nature reviews Cancer* 2008, **8**: 671–682.

[5] Tabin CJ, Bradley SM , et al. **Mechanism of activation of a human oncogene**. *Nature* 1982, **300**: 143–149.

[6] Schilsky RL. **Implementing personalized cancer care**. *Nature Reviews Clinical Oncology* 2014, **11**: 432–438.

[7] Gerlinger M, Rowan AJ , et al. **Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing**. *New England Journal of Medicine* 2012, **366**: 883–892.

[8] Sidow A and Spies N. **Concepts in solid tumor evolution**. *Trends in Genetics* 2015, **31**: 208–214.

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

[10] Jahr S, Hentze H et al. **DNA Fragments in the Blood Plasma of Cancer Patients: Quantitations and Evidence for Their Origin from Apoptotic and Necrotic Cells**. *Cancer Research* 2001, **61**: 1659–1665.

[11] Crowley E, Di Nicolantonio F et al. **Liquid biopsy: monitoring cancer-genetics in the blood**. *Nature Reviews Clinical Oncology* 2013, **10**: 472–484.

[12] Holdenrieder S, Nagel D et al. **Clinical relevance of circulating nucleosomes in cancer**. *Annals of the New York Academy of Sciences* 2008, **1137**: 180–189.

[13] Peters DL and Pretorius PJ. **Origin, translocation and destination of extracellular occurring DNA — A new paradigm in genetic behaviour**. *Clinica Chimica Acta* 2011, **412**: 806–811.

[14] 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.

[15] 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.

[16] Chan KC, Jiang P et al. **Cancer Genome Scanning in Plasma: Detection of Tumor-Associated Copy number Aberrations, Single-Nucleotide Variants, and Tumoral Heterogeneity by Massively Parallel Sequencing**. *Clinical Chemistry* 2013, **59**: 211–224.

[17] Forshew T, Murtaza M et al. **Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA**. *Sci Transl Med* 2012, **4**: 136ra68.

[18] 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.

[19] Stroun M, Lyautey J et al. **About the possible origin and mechanism of circulating DNA Apoptosis and active DNA release**. *Clinica Chimica Acta* 2001, **313**: 139–142.

[20] 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.

[21] 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.

[22] 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.

[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] 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.

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

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

[27] Struhl K and Segal E. **Determinants of nucleosome positioning**. *Nature Structural & Molecular Biology* 2013, **20**: 267–273.

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

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

[30] Valouev A, Johnson SM, et al. **Determinants of nucleosome organization in primary human cells**. *Nature* 2012, **474**: 516–520.

[31] 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.

[32] Sequence Read Archive. `http://www.ncbi.nlm.nih.gov/sra`. [Online; accessed 12-December-2015]. 2015.

[33] 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.

[34] Gene expression omnibus. `http://www.ncbi.nlm.nih.gov/geo`. [Online; accessed 30-November-2015]. 2015.

[35] GM12878 MNase-seq. `http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeSydhNsome/wgEncodeSydhNsomeGm12878Sig.bigWig`. [Online; accessed 14-September-2015]. 2015.

[36] Huang DW, Sherman BT et al. **Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources**. *Nature Protocols* 2009, **4**: 44–57.

[37] Ashburner M, Ball CA et al. **Gene ontology: tool for the unification of biology**. *Nature Genetics* 2000, **25**: 25–29.

[38] Kanehisa M, Sato Y et al. **KEGG as a reference resource for gene and protein annotation**. *Nucleic Acids Research* 2016, **44**: D457–D462.

[39] Nishimura D. **BioCarta**. *Biotech Software & Internet Report* 2001, **2**: 117–120.

[40] Joshi-Tope G, Gillespie M et al. **Reactome: a knowledgebase of biological pathways**. *Nucleic Acids Research* 2005, **33**: D428.

[41] Illumina DNA nano library preparation protocol. `http://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/samplepreps_truseq/truseqnanodna/truseq-nano-dna-library-prep-guide-15041110-d.pdf`. [Online; accessed 12-September-2016]. 2016.

[42] Github Code. `https://github.com/PeterUlz/Nucleosome_ctDNA`. [Online; accessed 15-September-2016]. 2016.

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

[44] SAM format. `http://samtools.github.io/hts-specs/SAMv1.pdf`. [Online; accessed 19-July-2015]. 2015.

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

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

[47] FASTX Toolkit. `http://hannonlab.cshl.edu/fastx_toolkit/index.html`. [Online; accessed 19-July-2015]. 2015.

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

[49] BAM to wiggle script. `https://github.com/chapmanb/bcbb/blob/master/nextgen/scripts/bam_to_wiggle.py`. [Online; accessed 30-November-2015]. 2015.

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

[51] Mukaka MM. **A guide to appropriate use of Correlation coefficient in medical research**. *Malawi Medical Journal* 2012, **24**: 69–71.

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

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

[54] EBI Expression Atlas. `http://www.ebi.ac.uk/gxa/experiments/E-MTAB-3358`. [Online; accessed 22-December-2015]. 2015.

[55] Kim D, Pertea G et al. **TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions**. *Genome Biology* 2013, **14**: R36.

[56] Trapnell C, Williams BA et al. **Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation**. *Nature Biotechnology* 2010, **28**: 511–5.

[57] Cacoullos T. **Estimation of multivariate density**. *Annals of the Institute of Statistical Mathematics* 1966, **18**: 179–189.

[58] Hart T, Komori HK et al. **Finding the active genes in deep RNA-seq gene expression studies**. *BMC Genomics* 2013, **14**: 778.

[59] Koboldt D, Zhang Q et al. **VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing**. *Genome Research* 2012, **22**: 568–576.

[60] Wang K, Li M et al. **ANNOVAR: Functional annotation of genetic variants from next-generation sequencing data**. *Nucleic Acids Research* 2010, **38**: e164.

[61] The 1000 Genomes Project Consortium. **A global reference for human genetic variation**. *Nature* 2015, **526**: 68–74.

[62] Lek M, Karczewski KJ, et al. **Analysis of protein-coding genetic variation in 60,706 humans**. *Nature* 2016, **536**: 285–291.

[63] Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP). `http://evs.gs.washington.edu/EVS/`. [Online; accessed 12-October-2015]. 2015.

[64] Lai W, Choudhary V et al. **CGHweb: a tool for comparing DNA copy number segmentations from multiple algorithms**. *Bioinformatics* 2008, **24**: 1014–5.

[65] Ulz P, Heitzer E , Speicher MR. **Co-occurrence of MYC amplification and TP53 mutations in human cancer**. *Nature Genetics* 2016, **48**: 104–106.

[66] Carter SL, Cibulskis K et al. **Absolute quantification of somatic DNA alterations in human cancer**. *Nature Biotechnology* 2012, **30**: 413–421.

[67] Beroukhim R, Mermel CH et al. **The landscape of somatic copy-number alteration across human cancers**. *Nature* 2010, **463**: 899–905.

[68] Ngo HB, Lovely GA et al. **Distinct structural features of TFAM drive mitochondrial DNA packaging versus transcriptional activation**. *Nature Communications* 2014, **5**: 3077.

[69] Thakur BK, Zhang H et al. **Double-stranded DNA in exosomes: a novel biomarker in cancer detection**. *Cell Research* 2014, **24**: 766–760.

[70] Bettegowda C, Sausen M et al. **Detection of Circulating Tumor DNA in Early- and Late-Stage Human Malignancies**. *Science Translational Medicine* 2014, **6**: 224ra24.

[71] Schones DE, Cui K, et al. **Dynamic regulation of nucleosome positioning in the human genome**. *Cell* 2008, **132**: 887–898.

[72] Adelman K and Lis JT. **Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans.** *Nature Reviews Genetics* 2012, **13**: 720–731.

[73] 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.

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

[75] Patel KM and Tsui DWY. **The translational potential of circulating tumour DNA in oncology**. *Clinical Biochemistry* 2015, **48**: 957–961.

[76] Garcia-Olmo DC, Dominguez C et al. **Cell-free nucleic acids circulating in the plasma of colorectal cancer patients induce the oncogenic transformation of susceptible cultured cells**. *Cancer Research* 2010, **70**: 560–567.

[77] Chang CP, Chia RH et al. **Elevated cell-free serum DNA detected in patients with myocardial infarction**. *Clinica Chimica Acta* 2003, **327**: 95–101.

[78] Jylhava J, Nevalainen T et al. **Characterization of the role of distinct plasma cell-free DNA species in age-associated inflammation and frailty**. *Aging Cell* 2013, **12**: 338–397.

List of Figures

List of Tables

A. *Appendix*

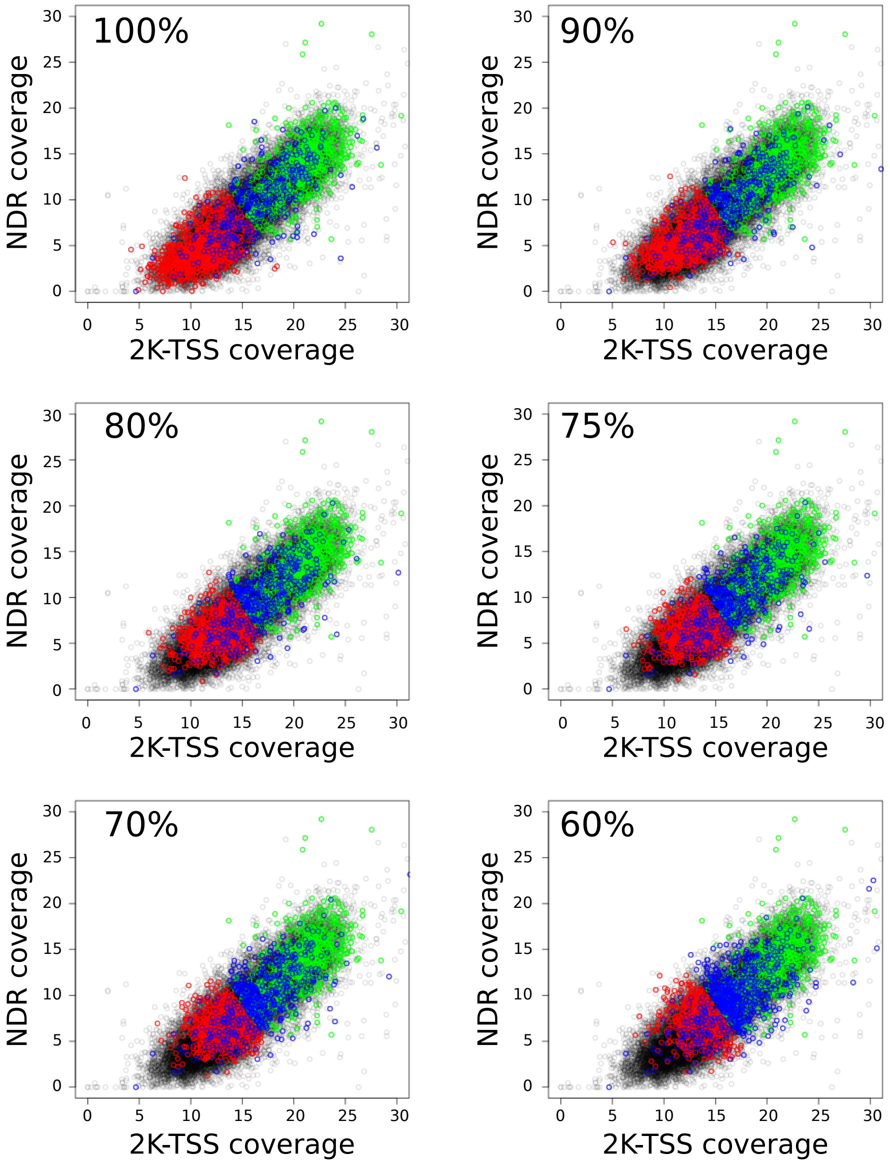

Figure A.1: Scatter plot of the two features for every transcription start site at dilutions of 100%, 90%, 80%, 75%, 70% and 60%.

Table A.1: Significant functional enrichment hits of the Top100 genes on gained genomic regions of both breast cancer patients to gene ontology terms [37]. CC corresponds to Gene Ontology: Cellular Compartment, BP corresponds to Gene Ontology: Biological Process. No significant hit was found for the Top100 genes of patient B13.

Sample	**Category**	**Term**	**P-value**	**corrected P-value**
B7_1	CC	nucleosome	2.0x10 $^{-5}$	3.4x10 $^{-3}$
B7_1	CC	protein DNA complex	9.2x10 $^{-5}$	7.8x10 $^{-3}$
B7_1	BP	nucleosome organization	1.6x10 $^{-4}$	1.9x10 $^{-2}$
B7_1	BP	protein-DNA complex assembly	1.5x10 $^{-4}$	2.2x10 $^{-2}$
B7_1	BP	chromatin assembly	1.2x10 $^{-4}$	2.3x10 $^{-2}$
B7_1	BP	DNA packaging	4.6x10 $^{-5}$	2.7x10 $^{-2}$
B7_1	BP	nucleosome assembly	1.0x10 $^{-5}$	2.9x10 $^{-2}$

Table A.2: Significant functional enrichment hits of the Top100 genes on gained genomic regions of both breast cancer patients to pathways as defined by KEGG [38], Biocarta [39] and Reactome [40] databases.

Sample	**Database**	**Term**	**P-value**	**corrected P-value**
B7_1	KEGG	Systemic lupus erythematosus	6.7x10 $^{-5}$	3.9x10 $^{-3}$
B7_1	Reactome	Telomere Maintenance	4.6x10 $^{-4}$	1.3x10 $^{-2}$
B13_1	Reactome	Gene expression	3.9x10 $^{-4}$	6.7x10 $^{-3}$
B13_1	Reactome	Metabolism of proteins	3.2x10 $^{-4}$	1.1x10 $^{-2}$
B13_1	Reactome	3'-UTR mediated translational regulation	1.9x10 $^{-3}$	2.1x10 $^{-2}$

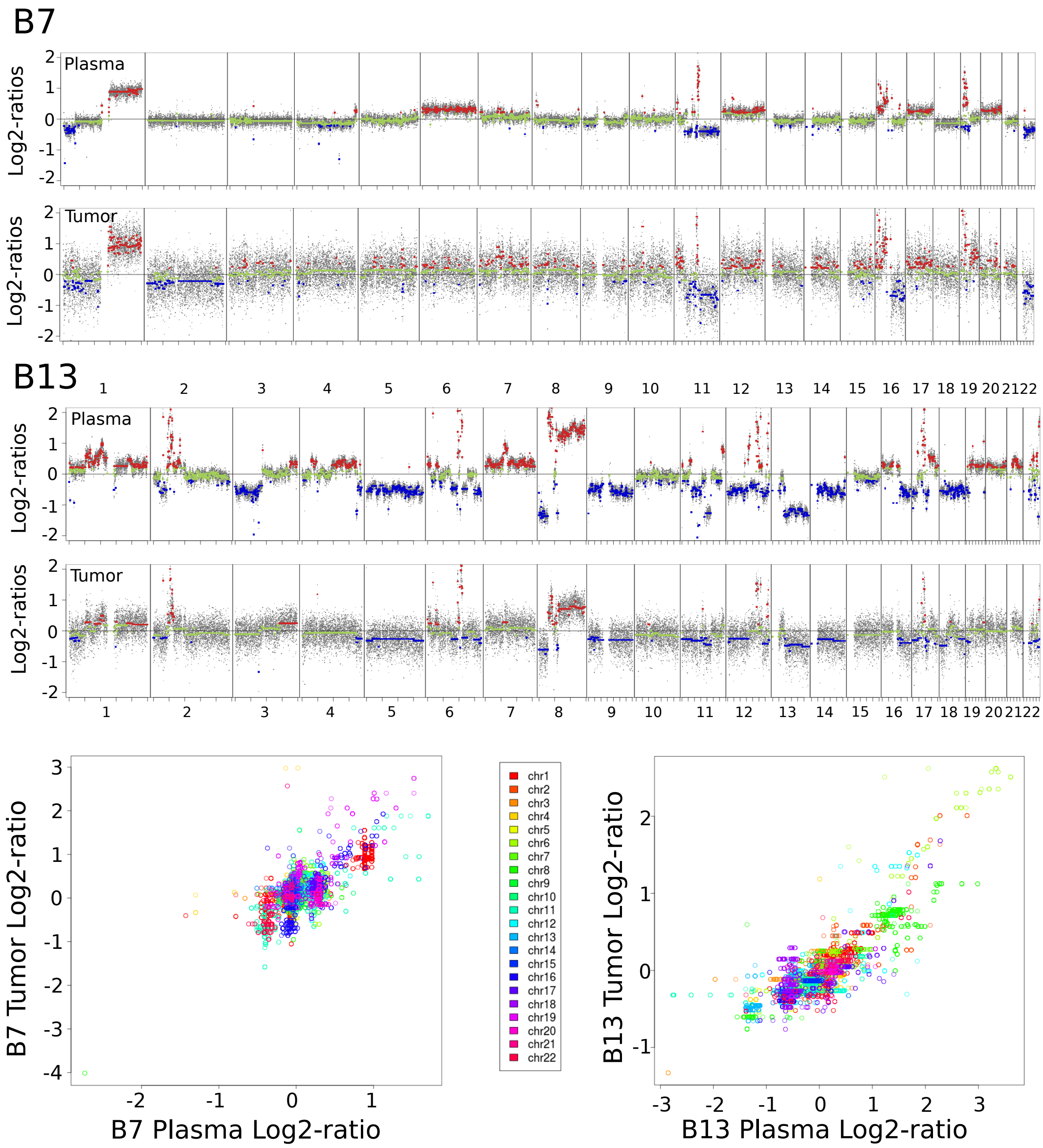

Figure A.2: Read-depth analysis of cfDNA and tissue biopsies of B7 and B13 shows copy number alterations throughout the genome, including focal amplifications of *CCND1* (B7; chromosome 11) and *FGFR1* (chromosome 8p) and *ERBB2* (chromosome 17). CNAs obtained from plasma and tissue biopsies correlate well for both samples.