TECHNISCHE UNIVERSITÄT MÜNCHEN
Fachgebiet für Bioinformatik

Integrative Analysis of High-throughput Data in Cancer and Neurogenesis

Yanping Zhang

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

genehmigten Dissertation.

Vorsitzender:	Prof. Dr. Martin Klingenspor
Prüfer der Dissertation:	1. Prof. Dr. Dimitrij Frischmann
	2. Prof. Dr. Weihua Chen

Die Dissertation wurde am 19.06.2017 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 25.09.2017 angenommen.

Abstract

In this thesis, we take advantage of current high throughput assays (microarray and next generation sequencing) to examine the pattern of somatic copy number alterations (SCNAs) in cancer genomes, and investigate the role of DNA methylation in neural stem cells (NSCs).

To begin, we analyzed the relationship between genomic architecture and SCNA. This work was done by using the pooled cnv data from The Cancer Genome Atlas Pan-Cancer (TCGA) project. In multiple linear regression (MLR) analyses, previously identified features and several novel features, including distance to telomere, distance to centromere, and low complexity repeats were found to be factors predicting SCNA pattern in cancers. Furthermore, applying a rare event logistic regression model and an random forest classifier, we found that genomic features *e.g.* distance to telomere and direct repeats are effective to predict common SCNA breakpoint hotspots.

We carried out an analysis of SCNAs in 160 osteosarcoma (OS) samples. We found that chromosomal breakages are not randomly distributed in the OS genome and enriched in genomic features with the potential to form DNA secondary structures. We found a number of genes including *TP53*, *ATRX*, *FOXN1*, and *WWOX* located in those broken region tend to become deregulated or deleted. In addition, chromothripsis and aneuploidy are common in OS and predictive of disease outcome.

Finally, we performed an integrative analysis of whole-genome bisulfite sequencing from neural stem cells (NSCs) in injured and non-injured conditions. We found that NSCs are more responsive to brain injury in terms of methylation changes. Furthermore, a substantial number of genomic regions become permissive to transcription after injury in NSCs. We uncovered an injury-induced epigenetic program that encompasses the decommissioning of developmental transcription factors and enhancers selectively in NSCs.

Zusammenfassung

In dieser Arbeit nutzen wir moderne Hochdurchsatz-Assays (Microarray und Next-Generation-Sequenzierung), um Muster in den Somatic Copy Number Alterations (SCNAs) der Krebsgenome zu finden, und untersuchen die Rolle von DNA-Methylation in neuralen Stammzellen (NSCs).

Zu Beginn haben wir den Zusammenhang zwischen Genomarchitektur und SCNA analysiert. Dies geschah unter Verwendung von zusammengelegten CVN-Daten aus dem The-Cancer-Genome-Atlas-Pan-Cancer-Projekt (TCGA). Durch multiple, lineare Regressionsanalysen (MLR) haben wir sowohl bereits bekannte als auch neue Merkmale identifiziert, wie etwa Distanz zum Telomer, Distanz zum Centromer und Low-Complexity-Repeats, die SCNA-Muster in Krebsgewebe vorhersagen. Weiterhin haben wir mittels einer Rare-Event-Logistic-Regression und einem Random-Forest-Klassifizierer herausgefunden, dass genomische Merkmale, zum Beispiel die Distanz zum Telomer und Direct Repeats, effektive Prädiktoren für gemeine SNA-Breakpoint-Hotspots sind.

Wir haben SCNAs in 160 osteosarcoma-Proben (OS) analysiert. Dabei haben wir festgestellt, dass Chromosomal Breakages nicht zufällig verteilt im OS-Genom auftreten und mit genomischen Merkmalen angereichert sind, die potenziell sekundäre DNA-Strukturen bilden können. Wir haben eine Gruppe von Genen bestimmt, darunter TP53, ATRX, FOXN1 und WWOX, die in einer solch zerbrochenen Region dazu neigen, dereguliert oder gelöscht zu werden. Zusätzlich sind chromothripsis und aneuploidy verbreitet in OS und prädiktiv für einen Krankheitsverlauf.

Zum Abschluss haben wir eine integrative Analyse von Whole-Genome Bisulfite Sequencing von neuralen Stammzellen (NSCs) im verletzten und intakten Zustand durchgeführt. Diese hat ergeben, dass NSCs besser auf Hirntraumata ansprechen im Sinne von Methylation Changes. Außerdem wird ein wesentlicher Anteil der Genomregionen in NSCs nach einer Verletzung zur Transkription freigegeben. Wir haben ein traumainduziertes, epigenetisches Programm offen gelegt, das das Decomissioning von Developmental Transcription Factors und Enhancers in NSCs reguliert.

Acknowledgements

First of all, I would like to express my deepest gratitude to my adviser Prof. Dr. Dmitrij Frishman, for his guidance, support, and encouragement. I am very grateful that Prof. Frishman provides me the opportunity to prepare this thesis. I am very thankful to Prof. Dr. WeiHua Chen for investing time in reviewing my thesis and acting as a second supervisor.

I am also very grateful to Prof. Dr. Martin Klingenspor for agreeing to be the chair of my thesis committee.

I would like to express my gratitude to our group members in Weihenstephan. Thanks to Hongen Xu for all the lessons he has taught me about the biology. Thanks to Bo Zeng for many interesting conservations and help. And special thanks to Andery Chursov for sharing his expertise.

I would also like to thank Leonie Corry for her excellent administrative help during my years in the lab.

I want to thank Prof.Dr.Ana Martin-Villalba, Sheng Zhao for their delightful comments on the project that I have been working on.

I want to express special thanks to all my colleagues and friends for proof-reading my thesis and for their good comments that helped me improving it.

Finally, I would like to thank my husband Guang for his continuous support, patient, love and encouragement. My son Xiaoshitou cheered me up after long hours working. I am thankful to my brother Yiping Zhang, my parents Xinguo Zhang and Xiaoxin Xie for all their love and efforts in educating me.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

A great fraction of the genome carries copy number variations (CNVs) [1, 2, 3, 4], which can arise meiotically and also somatically as shown that identical twins differ in CNVs [5] and observed difference in copy number of repeated sequences for different tissues from an individual [6]. Somatic copy-number alteration (SCNA, distinguished from germline copy-number variation) is important in cancer formation and progression by activating oncogenes and inactivating tumor suppressor genes [7, 8, 9, 10].

DNA methylation is arguably the best understood and most widely studied epigenetic modification in mammalian cells; the mechanisms controlling the establishment and maintenance of DNA methylation patterns are well characterized. Functional studies have shown that DNA methylation is an important cell-intrinsic program, which can interact with transcription factors and environmental cues to modulate the normal development and differentiation of neural stem cells (NSCs) [11].

The thesis starts with an introduction to the biological background and mathematical concepts used in the work.

1.1 Copy number variations

CNVs are gains or losses copies of DNA segments, and are a major type of genetic variations that are widely found in human and other mammalian genomes [12]. CNV including genomic deletion, duplication, and complex rearrangement can differ in size ranging from 100 base pairs to several mega base pairs [13]. CNVs are not uniformly distributed across the genome, instead they tend to cluster in discrete regions with a high mutation rate. Selection and mutational biases are found to shape the genomic distribution of CNVs [14].

In the human genome, about half of the CNVs are found to disrupt protein-coding regions [15]. CNV loci encompassing genes may potentially cause gene expression variations [16], alter gene structures, affect epigenetic regulation and contribute to phenotypic variation [17]. A great number of CNVs have been implicated in complex human diseases, such as cancer [18], autism [19], and even susceptibility to HIV [20] due to the effect of CNVs on gene expression and their potentially disruptive effects on gene structure and function. SCNAs often occur during carcinogenesis, leading to the amplification of oncogenes or deletion of tumor suppressor genes [21]. Indeed, quite a few cancer-related genes, such as *KRAS*, *RB1*,*PTEN* [22] have been identified to be affected by SCNAs. Cancer genes are more frequently found in genomic regions with recurrent CNVs, where CNVs are common among tumor samples [23]. Therefore, studies on CNVs can help us to understand the genetic etiology of human diseases.

1.1.1 The mechanism of CNV

CNVs represent a significant of genetic variation. Generally CNVs are formed when DNA double strand breaks (DSBs) are not properly repaired [24]. DSBs occurs in the process of normal cellular metabolic reaction or when cells are exposed to ionizing radiation. The mechanisms leading to change in the copy number include homologous recombination repair and non-homologous repair [25]. Non-homologous repair can further be divided into non-replicative and replicative non-homologous repair.

Homologous recombination repair

Homologous recombination repair including homologous recombination (HR) and single-strand annealing (SSA) pathway requires sequence homology to perform the repair [25]. HR requires longer sequence identity (100 bp to 200bp) than SSA (50bp). Another difference is that SSA always cause small deletions, while mostly HR can repair DNA breaks without generating copy number alterations [24, 25].

Non-allelic homologous recombination (NAHR) between low-copy repeats (LCRs) is the major type of HR. LCRs, also known as segmental duplications, are stretches of DNA with over 90% sequence homology [26]. Non-allelic copies of LCRs other than copies at the usual allelic positions, can sometimes act as the mediators of NAHR. For example, when the two LCR pairs are located on the same chromosome and in the same orientation, NAHR between them will generate duplication and deletion [27]. However, when LCR pairs are on the same chromosome but in different orientation, it will cause inversions. It is also worthy to note that a proportion of NAHR events use repetitive elements such as short interspersed nuclear elements (SINEs), long interspersed element-1 (L1) and long terminal repeat (LTR) retrotransposons, rather than LCRs as homology substrates.

SSA happens when neither of the ends of a two-ended DSB invades homologous sequence. In humans, identical Alu repeats located only a few hundred base pairs from each other have been found to trigger DSB-induced SSA [28]. The longer the sequence between the repeats, the less likely that SSA will repair the DNA break. This length restriction suggests that SSA is only a minor mechanism for the formation of CNVs.

Non-homologous repair

Non-replicative non-homologous repair - Non-homologous end joining (NHEJ) and micro-homology mediated end joining (MMEJ) are two major form of non-replicative non-homologous repair mechanism. NHEJ does not require sequence homology while MMEJ uses microhomology to repair DSBs [29]. NHEJ either rejoins DSB ends accurately or cause small deletions (1-4 bp) and insertions [25]. NHEJ proceeds in four steps: detection of DSB; molecular bridging of both broken DNA ends; modification of the ends to make them compatible and ligatable; and the final ligation step. Although NHEJ is not directly mediated by nor strictly dependent on certain genomic elements in the way that NAHR is dependent on LCRs, it may still be stimulated and regulated by the genomic architecture [30, 31]. MMEJ uses 5-25 bp micro-homologous sequences to anneal at the DSB ends, leading to deletions of sequences flanking the original breaks.

Replicative non-homologous repair - In recent years, replication-based repair mechanisms have been proposed to explain the highly complex CNVs [25, 32, 33] that are difficult to be explained by either the NAHR or NHEJ recombination mechanism. Three mechanisms including fork stalling and template switching (FoSTeS) [34], micro-homology mediated break-induced replication (MMBIR) [35] and serial replication slippage (SRS) are proposed [36]. All of these models require microhomology for re-annealing and assume template DNA can be generated from nearby replication forks [25, 33]. Although these models can also be applied to mediate the formation of simple CNVs, it is hard to distinguish them from NHEJ and MMEJ.

1.1.2 Detection of CNV

Accurate CNVs detection plays an important role in the analysis of cancer genome, which can improve cancer diagnosis and treatment decision. Many research on the techniques of detecting CNVs were performed.

SNP arrays have been applied extensively for detecting copy number variation in tumor cells. SNP arrays use less sample per experiment compared to comparative genomic hybridization (CGH) arrays. Although it is much easier to detect copy number variation due to the next generation sequencing technology, SNP arrays of Illumina and Affymetrix platforms can identify CNV at high resolution without a great reduction in genome-wide coverage. The SNP array-based approaches use computational methods leveraging signals from genotyping and sequencing to infer CNVs. The log R ratio (LRR) represents the logged ratio of observed probe intensity to expected intensity for both alleles, and the B allele frequency (BAF) is the relative proportional of one of the alleles with respect to the total intensity signal. Copy number changes can be detected through LRR and BAF, provided by the SNP array.

For SNP array-based analyses, a number of tools have been developed for identification of regions affected by genomic aberrations. They are based on two commonly used strategy: circular binary segmentation (CBS) method and the hidden Markov model (HMM) method. CBS method is a segmentation of the total probe signals into genomic regions with similar average signal. For the CBS method, a variety of programs have been developed. For example, OncoSNP [37], GenoCNA [38], GPHMM [39] and MixHMM [40] have been developed for copy number analysis of Illumina SNP-array data. PICNIC (Predicting Integral Copy Number in Cancer) , CNNLOH [41], PSCN [42] and TumorBoost [43] are suitable for Affymetrix SNP-array data. For the HMM method, ASCAT (Allele-Specific Copy Number Analysis of Tumors) [44], GAP (Genome Alteration Print) [45] are prevalent programs. ASCAT and GAP allow analysis of both Illumina and Affymetrix SNP-array data of tumor samples. All of these methods detect CNVs using sample-specific breakpoints, not considering different samples simultaneously. Due to the high noise level in the intensity values, the boundaries of the detected CNVs are more likely to vary among individuals.

Common CNV regions (*i.e.* recurrent CNV) tend to occur at the same genomic positions across different individuals. As a result, disease-causing genes are preferably to locate in recurrent CNV regions. Recurrent CNV regions encompassing genes are more probable to harbor driver alterations (functionally significant for disease initiation or progression), while “passenger”alterations (random somatic events irrelevant to pathological events) are more likely to occur in individual-sample specific CNVs. A variety of statistical and computational approaches have been developed for recurrent CNV detection. These methods differ in terms of both input data and the implemented algorithm models. For the input, most of the recurrent CNV detection approaches can be divided into two categories: continuous (log 2 ratio) and discrete (gains/losses). For the algorithms, they can be categorized in different models , such as permutation probabilistic method, null model or none.

1.1.3 Chromothripsis

Recently, the combination of whole-genome sequencing, SNP array and bioinformatics analyses has led to the discovery of a new catastrophic chromosomal rearrangement, termed as chromothripsis. Chromothripsis was first found in a patient with chronic lymphocytic leukemia [46] by an comprehensive analysis of the chromosomal rearrangements. Since the initial discovery, there have been many studies confirming that chromothripsis features were indeed exhibited in many tumor types [47, 48, 49, 50, 51]. Chromothripsis occurs in approximately 2% to 5% of human cancers [46], yet more frequently reaching up to 39% in certain tumor types [49]. Initially, it was thought that chromothripsis was particularly common in bone cancers, but recent studies show that all sarcomas are reported to exhibit increased rates of chromothripsis [47]. The high frequency of chromothripsis in certain tumor types suggests that chromothripsis depends on the genetic and environmental background of cancers. Chromothripsis is a common mechanism that can drive tumorigenesis by initiating the formation of double-minute chromosomes. It can not only lead to the amplification of a single oncogene but also create potent amplicons containing multiple candidate oncogenes [52, 53, 54]. The high number of rearrangements caused by chromothripsis also suggests it might have a higher probability of creating functional oncogenic fusions driving tumorigenesis. This is not a common phenomenon given that the genome only consist of 1% of coding sequences. Besides, chromothripsis can drive cancer by the generation of deletion of one or more tumor suppressor genes at a single catastrophe event. In a insightful analysis of TCGA SNP array data, 72% of chromothripsis events were linked to copy-number variation regions that are recurrently disrupted in cancer [55]. Chromothripsis has also been associated to mutations in

TP53 and an aberrant DNA damage response [46].

Four features distinguish this patterns of rearrangements. First, there are complex adjacencies rather than simple deletions or non-overlapping tandem duplications due to the clustered breakpoints in the chromosome or chromosomal region. Second, despite the large number of rearrangements, the chromosome region oscillates between only two copy number states that is in sharp contrast to conventional clusters of complex rearrangements. Third, the alternation between two copy number states is accompanied by loss and preservation of heterozygosity. Finally, the pattern of end-joining strongly suggests an origin from a DSB.

1.2 Statistical methods for the analysis of CNV data

A variety of different types of regression and classification have been developed in recent years, and the machine learning continues to expand at an impressive rate. Below, we give a brief introduction to the regression and machine learning techniques used in this work.

1.2.1 Multiple linear regression

Multiple linear regression is applied to examine the relationship between one dependent variable Y and multiple independent variables X_i, given the vector of multiple predictors $X^T = X_1 + \ldots + X_n$, the response Y can be predicted via the formula:

$$Y = \alpha + \beta_1 X_1 + \ldots + \beta_n X_n$$

where α, β_p is the intercept and slopes, respectively.

The residual sum of squares (RSS) is used to measure the performance of a regression model, and it is defined as:

$$RSS\left(\hat{Y}, Y\right) = \sum_{i=1}^{N} \left(\hat{Y}_i - Y_i\right)^2$$

where Y_i is the true value for the outcome, and $\hat{Y}_i$ is the expected value for the outcome.

In the least-squares model, the best-fitting line for the observed data is calculated by minimizing the residual sum of the squares.

R-squared

R-squared (R^2), also known as the coefficient of multiple determinations, is a measurement of how close the data are to the fitted regression line. It is used to capture the explanatory power of the regression model. It is defined as:

$$R^2 = \frac{EV}{TV}$$

where EV, TV is the explained variation and total variation, respectively. R^2 ranges from 0 to 100%, where 0 denotes the model explain none of the variability of the response and 100% denotes the response can be fully explained. Generally, the higher the R^2, the better the model fits the data.

Adjusted R-squared

The adjusted R-squared ($\hat{R}^2$) is associated with the number of variables and the number of observations. The performance of R^2 will improve when adding more predictors into the model, but some of that improvement may be due to chance alone. So adjusted R-squared tries to correct for this, and it is defined as:

$$\hat{R}^2 = 1 - \frac{N-1}{N-k-1}(1-R^2)$$

where N is the number of observations and k is the number of predictors.

The variance inflation factor

The variance inflation factor (VIF) for each variable measures the increase of the variance compared to an orthogonal basis. As a rule of thumb, the regression coefficients are poorly estimated due to multicollinearity if any of the VIFs exceeds 10.

1.2.2 Logistic regression for classification

Logistic regression is a statistical method similar to linear regression except that the outcome is measured with a dichotomous variable(true/false, success/failure, yes/no etc.). Simple logistic regression is the regression with one dichotomous characteristic of interest and one independent variable; multiple logistic regression refers to the regression that there is a single dichotomous outcome and a set of independent variables.

The dependent variable is assumed to be a stochastic event in logistic regression. For instance the outcome event is either killed or alive when we analyze a pesticides kill locusts. Logistic regression calculates the probability for bug of getting killed. If the probability of bug getting killed is greater than 0.5 it is denoted dead, if it is less than 0.5 it is denoted alive.

The outcome variable is often coded as 0 or 1, where 1 indicates that the presence of outcome and 0 indicates that the absence of outcome. If we define p as the probability that the outcome is 1 logistic regression is defined as:

$$\hat{p} = \frac{\exp(b_0 + b_1X_1 + \cdots + b_pX_p)}{1 + \exp(b_0 + b_1X_1 + \cdots + b_pX_p)}$$

where $\hat{p}$ is the expected probability of the presence of outcome; X_i is independent variable and b_i is the regression coefficient.

It is worthy to note that multiple linear regression model chooses parameters that minimize the RSS while logistic regression model chooses parameters that maximize the likelihood of observing the sample values.

1.2.3 Rare event logistic regression

Logistic regression clearly interprets the relationship between a dichotomous dependent variable y and a set of predictor variables. Although logistic regression is a popular approach, it may generate extremely biased results when the proportion of the response variable data is imbalanced. King and Zeng [56] have shown that rare events are difficult to predict as the standard application of logistic regression techniques can sharply underestimate the probability for rare events. To correct this bias, they proposed rare-event logistic regression. Specifically, an endogenous stratified sampling of the dataset was first performed, then a prior correction of the intercept was done and finally a correction of the probabilities was calculated to include the estimation uncertainty.

In our data, the response variable data is imbalanced (response variable $y = 0 >> y = 1$). So we decided to use rare event logistic regression due to its ability to deal with unbalanced binary event data.

1.2.4 Random forest

Random forest is an ensemble of decision trees [57]. An example of a decision tree is illustrated in Figure 1.1. It has been applied extensively in the computational biology such as gene expression classification, protein-protein interaction or disease associated genes identification from genome wide association studies.

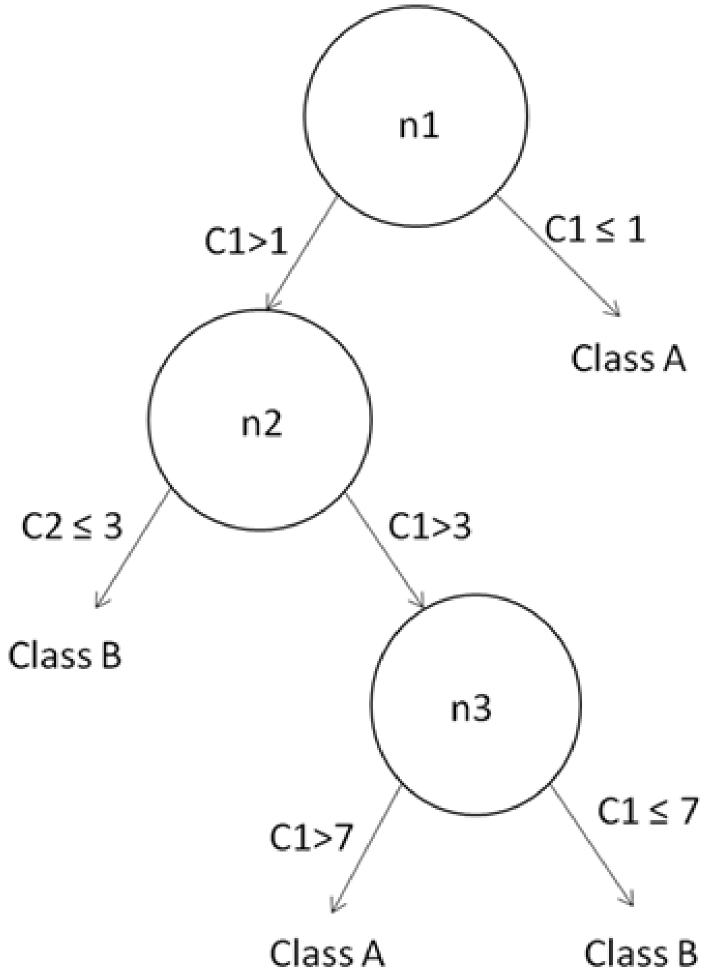


Figure 1.1: An illustration of a decision tree. The decision tree consists of three nodes denoted as n1,n2 and n3. At each node the data is split based on a rule associated to that node and the attribute associated to the vectors denoted as C1,C2 and C3. In the terminal nodes the class is assigned for the vector.

Given a training set $X = X_1, \cdots, X_n$ with response $Y = Y_1, \cdots, Y_n$, the random forest is calculated as follows:

- Sample N cases at random with replacement from X, Y; call these X_b,Y_b. X_b,Y_b should be about 66% of the total training data.

- Train a decision tree f_b on X_b,Y_b. It is important to note that predictor variables (say m) are selected at random out of all the predictor variables and the best split on these m is used to split the node.

- Calculate the misclassification rate - out of bag error rate (OOB) for each tree using the leftover data (33% of the total data). Aggregate error from all trees to determine overall OOB error rate for the classification.

- Repeat step 1 to 3 , b times.
- Each tree gives a classification, and we say the tree “votes”for that class. The forest chooses the classification having the most votes over all the trees in the forest.

After the training, the random forest can be used to classifying new data.

Feature selection

Feature selection consists in identifying a subset of the original input variables that are useful for building a good model. Feature selection can improve the prediction power of the model. For example, it can exclude the predictors that has a negative influence on the model. Besides, feature selection allows for a faster and more cost effective implementations in contexts when there are thousands or even more variables in a dataset. There are many feature selection algorithms and they are all based on the assessment of importance of each feature.

Feature importance

Three evaluation metrics including filter, wrapper and embedded methods assess the importance of features in terms of predictive power of the model. For the filter method, features are removed independently of the model based on criteria of their own properties. Mutual information, pearson correlation coefficient and inter or intra class distance are the common metrics [58]. The wrapper methods treat the variables as inputs and use heuristic search methods for the best subset according to the performance of optimized model. Stepwise regression, the most popular form of feature selection is a wrapper technique. It is a greedy algorithm that adds the best feature (or deletes the worst feature) at each round. The embedded methods typically couple the predictor search algorithm with the estimation of parameters and are usually optimized with a single objective function. It is also worthy to note that feature importance is also applied to establish a ranking of the predictors.

Gini vs Permutation

Several measures are available for feature importance in random forests. Gini importance or mean decrease in impurity (MDI) calculates the importance of each feature as the sum over the number of splits (across all tress) that include the feature, proportionally to the number of samples it splits. Permutation importance or mean decrease in accuracy (MDA) is assessed for each feature by removing the association between that feature and the target. This is achieved by randomly permuting the values of the feature and measuring the resulting increase in error. The influence of the correlated features is also removed.

1.3 Next Generation Sequencing

Next-Generation Sequencing (NGS) technologies have surpassed conventional capillary-based sequencing by the ability of massively parallel sequencing of short DNA fragments [59, 60]. NGS technologies are significantly cheaper, need significantly less DNA and are more accurate and reliable compared with Sanger sequencing. In contrast to hybridization-based technologies, it is not limited to the interrogation of selected probes on an array. Roche/454 [61], Illumina/Solexa [62] and LifeTechnologies/ABI [63] are the first platforms of NGS. Although they differ in specific technical details (Table 1.1), they share general processing steps [60]:

- First, the input DNA is fragmented followed by ligation to platform specific oligonucleotide adapter sequences. This process is called library preparation.
- In a next step, each single library molecule undergoes multiple rounds of amplification in a way that all copies of the same molecule stay clustered in the same position.
- The large number of clusters are then sequenced by alternate cycles of addition of fluorescently marked nucleotides and imaging.

1.3.1 RNA sequencing

NGS methods allow for sequencing the transcribed molecules, a methodology referred to as RNA sequencing (RNA-Seq). RNA-seq is a prevalent NGS methods applied to the investigation of transcriptome. RNA-seq has

Table 1.1: Primary features of the main NGS platforms

Platform	Read Length	Yields per run	Cost per bp	Disadvantages	Advantages
Roche:454	400bp	1Gbp	High	Low throughput High cost per base pair	Long read length De novo genome assembly Transcriptome assembly
Illumina:Solexa	100bp	20Gbp	Low	Short reads Relatively high error rate	High throughput Chip-seq, RNA-seq DNA methylation
Life Technologies: SOLiD	50bp	100Gbp	Low	Very short reads	Low error rate Genome re-sequencing for variant detection

a wider detection range, lower cost and is more sensitive for transcriptome profiling, compared to array-based methods. Besides, it can capture the genome-wide expression profile including lowly expressed genes, which might be missed using the traditional cloning based expression sequence tags (EST) approach. RNA-seq provides a useful tool to identify differentially expressed genes between tumor and normal tissues [64], detect several novel miRNAs in cancerous cells [65]. RNA-seq can also identify allele-specific expression, disease-associated SNPs, novel splice sites and several novel translocations [66, 67, 68, 69].

1.3.2 ChIP sequencing

ChIP sequencing (ChIP-seq) has been widely applied for identification of transcription factor binding sites and a variety of histone modifications. For example, Cheung's group used chip-seq to uncover the AR transcriptional network and found this network plays a critical role in manipulating AR activity for the targeted eradication of prostate cancer cells [70]. They confirmed that ChIP-seq has a important role in the discovery of transcriptional networks. In addition to identification of transcription factor binding sites, ChIP-seq can also been applied to uncover distinct mechanisms associated with differential gene regulation. Taking the important transcription factor nuclear factor κB (*NF-κB*) as an example, Lister.*et.al* recently used ChIP-seq to study the role of lysine methylation of the p65 subunit of the *NF-κB* in differential gene regulation [71]. They demonstrated that mutations in the mutants of lysine (K) 37 and 218/221 of p65 have dramatically different effects due to the fact that methylations of these residues affect different genes by distinct mechanisms. This suggested that cells may use a critical mechanism to differentially regulate NF-κB-dependent genes in different physiological or disease states. The above example reveals that ChIP-seq combined with site mutation of the post-translation modifications of a given transcription factor could help elucidating the fundamental mechanism of transcription factor-governed differential gene regulation.

1.3.3 NGS applied in DNA methylation

NGS is also an important tool to characterize DNA methylome, helping better understanding of specific cell-type expression patterns that is hard to be interpreted at the genetic level. NGS allows analysis of the entire genome so that methylome can be charted at single base-pair resolution.

For example, Lister.*et.al* [72] provides the first whole genome and single-base resolution methylome profiling for both human embryonic stem cells and fetal fibroblasts. They demonstrated that differential methylated regions are close to genes responsible for pluripotency and differentiation.

1.4 DNA methylation

DNA methylation is an epigenetic mechanism. A methyl group is transferred to the C-5 position of the cytosine ring of DNA and the methylation is catalysed by DNA methyltransferases (DNMTs) (Figure 1.2). Different DNMTs function together either as de novo DNMTs, establishing a new methylation pattern to unmodified DNA or as maintenance DNMTs that copy faithfully the DNA methylation pattern from the paternal strand onto the newly synthesized daughter strand. DNA methylation fluctuates during the process of mammalian development. For example, demethylation can occur during the process of cell division or the removal of methylcytosine is caused by an oxidized intermediate [73].

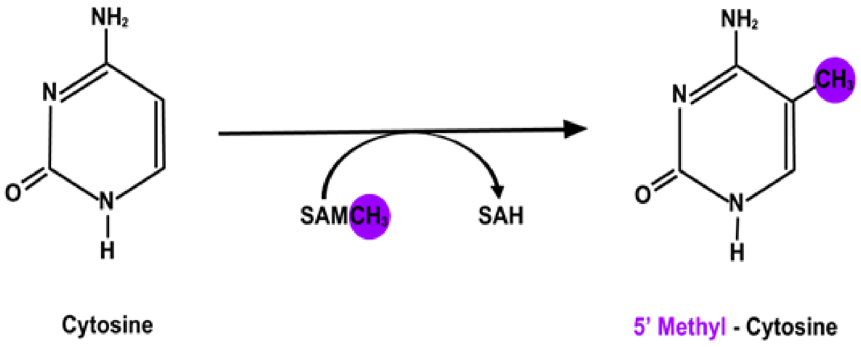


Figure 1.2: Illustration of DNA methylation, which converts cytosine to 5'methyl-cytosine by DNA methyltransferase (DNMT). SAM:S-adenosylmethionine; SAH:S-adenosylhomocysteine.

1.4.1 The genomic context of methylation

In humans, more than 98% of DNA methylation occurs in the context of a cytosine-guanine dinucleotide (CpG) in somatic cells and the result is two methylated cytosines positioned diagonally to each other on opposite strands of DNA. Interestingly, several studies revealed that there is a significant proportion of methylation in non-CpG contexts in pluripotent cell-types and oocytes [74, 75]. In plants, methylated cytosines are located in symmetrical (CG or CHG) or asymmetrical (CHH, where H is A, T, or C) context.

Mammalian genome consists of 1-4% of CpG dinucleotides and around 70% of CpG sites are methylated [76]. Methylated cytosine is inherently mutagenic because of the spontaneous deamination of 5-methylcytosine to thymine. Therefore, CpG motifs are generally depleted within the genome except certain regions with high CpG density, termed as the CpG islands (CGIs). CGIs are generally characterized by a minimum GC content of 50-55% and a defined sequence length often between 200bp-1kbp [77]. A comparison of human and mouse genome shows that the number of CGIs in the mouse genome is more than that in the human genome, but these regions have a lower average CpG density in mouse [78]. CGIs are frequently associated with gene promoter regions and methylation events at these CGIs may shape chromatin and gene transcription states, suggesting a regulatory role for methylation at these promoter-associated regions [79].

The majority of other CpG dinucleotides in the genome are generally methylated in most celluar context, but most of CpG sites comprising CGIs remain unmethylated during development, suggesting that methylation is the default state and CGIs are exceptions to this rule [80, 81]. Although CGIs are known to unmethylated, a small fraction of these loci are fully methylated often in a tissue-specific manner in some specific cell type [82, 83, 78]. Besides, these CGIs are located often in intragenic region and are often discovered in transcribed gene [84]. The propensity for a CGI being methylated is also associated with its CpG density, GC content and enrichment for transcription factor-binding motifs [85, 86, 87].

1.4.2 Role of DNA methylation

DNA methylation is involved in a number of celluar process such as X chromosome dosage compensation, gene imprinting and the maintenance of genome stability [88, 89]. Dysregulation of DNA methylation is implicated in the appearance of several disorders as cancer [90, 91] and a variety of human diseases are caused by defective imprinting [92].

Methylation also contributes to the regulation of gene expression. Studies show that high methylation of gene promoters usually leads to low or no transcription [81] and methylation has therefore been regarded as a repressive epigenetic mark. DNA methylation is frequently associated with active coding regions. For example, methylation is found in the gene body of actively transcribed genes in both plants and mammals [93, 94]. Genome-wide analysis of DNA methylation pattern at the single-base resolution in different physiological and pathological states has revealed that local changes in DNA methylation are associated with cell-type specific variation in gene expression. Moreover, DNA methylation plays an critical role in controlling gene expression during differentiation of stem cells [95, 96]. A genome-wide DNA methylation analysis of human embryonic stem cells (ESCs) and differentiated fetal fibroblasts demonstrated that there is significant differential methylation at genes important for stem cell maintenance and differentiation processes [72]. Furthermore, lineage-specific genes are activated at the appropriate time during development due to the DNA demethylation at enhancers and promoters [97]. In contrast, promoters of stem cell genes become more methylated as cells differentiate [98, 99, 100]. Embryonic stem cells on the other hand feature a significantly higher non-CG context methylation rate as well as methylation pattern modifications during the differentiation. This allows for the cellular differentiation of different cell types with various gene expression patterns.

1.5 Technologies for quantifying DNA methylation

The development of microarray and sequencing technologies provides the genome-wide pattern of DNA methylation, even in cohorts that contain hundreds or thousands of samples [101]. Accurate determination of methylation at CpG dinucleotide positions across the genome is critical for understanding its association with functional regulation. Many technologies for genome-wide DNA methylation analysis have been rapidly developed [101, 81]. These technologies are primarily based on four approaches including microarray, endonuclease digestion, affinity enrichment and bisulfite conversion to discriminate the methylated and unmethylated cytosines.

1.5.1 DNA methylation arrays

Microarrays now have become widely applied to investigate DNA methylation. The Illumina Infinium HumanMethylation450 BeadChip (HumanMethylation 450K) [102] has a predominant role for DNA methylation analysis, being not only the technology adopted by TCGA [103], but also for numerous studies such as the aging process [104] or inter-individual variability [105]. HumanMethylation 450K interrogates over 480,000 CpG sites in human genome. This array covers over 17-fold more CpG sites than 27K DNA Methylation array and therefore allows for a more comprehensive analysis of methylome [106, 107, 102]. Moreover, sample preparation takes only minimal time and each BeadChip contains 12 arrays for DNA hybridization, making this approach suitable for analyzing large cohorts. The HumanMethylation 450K array covers 96% coverage of CpG islands [108], more than 99% promoters of RefSeq genes [109], non-CpG methylated sites and miRNA promoters.

MethylationEPIC BeadChip Infinium microarray interrogates over 850, 000 CpG sites, which includes 413,745 new CpG sites not included in the 450K microarray. It covers the DNA methylation status of other sequences of the genome.

However, methylation data obtained by hybridization microarrays is biased and restricted by the design of the array platform.

1.5.2 Enzymatic DNA methylation analysis

DNA cleaverage with commonly used restriction enzymes like HpaII and MSPI in separate assays and the comparison of the resulting fragment sizes determines the methylated or unmethylated status of genomic DNA. Techniques including HELP (HpaII tiny fragment Enrichment by Ligation-mediated PCR), RLGC (Restriction Landmark Genomic Scanning) and DNA methylation Restriction Enzyme Analysis (MSRE) are based on this approach. One limitation of this method is that they can only detect the methylation at the restriction enzyme recognition sites or adjacent regions.

1.5.3 DNA enrichment methylation methods

Methylated DNA is isolated by antibody immunoprecipitation methods, methyl-CpG binding domains or other protein domains [110, 111, 112]. These methods include Methylated DNA immunoprecipitation(MeDIP), Methyl-CpG-binding domain(MBD) and others. DNA enrichment methylation methods have been used to generate the comprehensive profiling of DNA methylation. They do not damage the DNA like bisulfite treatment. However, the exact methylation state of individual CpG dinucleotides cannot be determined using this approach.

MeDIP is an immunocapturing method that uses a 5-methylcytidine-antibody to specifically recognize 5-methylcytidine(5mC). The methylated fraction of the genome can be analyzed by PCR or by microarray analysis or deep sequencing. However, it is important to note that CpGs are distributed unequally in mammalian genomes and the enrichment of MeDIP fraction depends both on the methylation status of the target sequence and the number of CpGs it contains.

MeDIP antibody, which binds methylated single-stranded molecules containing one or more methylated CpG sites. In contrast, MBD-based strategy uses MBD2b to capture double stranded methylated DNA fragments. DNA fragment size and sequencing read length determine the resolution of MeDIP-seq and MBD-seq. Both of these two methods provide moderate genomic resolution of 100-300bp and they can accurately identify differentially methylated regions (DMRs) between samples, however, the DMRs are only represent the relative methylation differences rather than quantitative differences due to that the detected methylation signals are strongly influenced by sequencing depth. MeDIP is more sensitive to methylation differences in regions that has a low CpG density, while MBD is more sensitive in regions with higher CpG density such as CpG islands [113].

1.5.4 Bisulfite conversion

Bisulfite sequencing (BS-seq) is applied to get the entire methylome of genome. It can determine the exact methylation status of almost every CpG, which secured bisulfite sequencing as the gold-standard method for detection of DNA methylation [114]. BS-seq is arguably the best method to offer unbiased genome-wide DNA methylation profiling [101, 81]. However, BS-seq requires a deep sequencing depth to provide sufficient CpG coverage for the methylation profile, which greatly increases the cost per methylome. MeDIP-seq and MBD-seq only need to produce DNA libraries that covers highly-methylated genomic regions, thus focusing the sequencing reads only on potential regions of interest. BS-seq sequencing covers the whole genome and most of sequencing reads map to unmethylated genomic regions are effectively discarded [115].

Whole-genome bisulte sequencing (WGBS) [72] and tagmentation-based whole-genome bisulte sequencing (T-WGBS) [116] are applied to detect the methylation level of individual CpG sites in the genome. Genomic DNA is first treated with sodium bisulfite to convert cytosine residues to uracil, but leaves 5-methylcytosine residues unaffected. Then, treated genomic DNA is used to perform high-throughput sequencing.

T-WGBS is able to determine methylation level of all CpG dinucleotides in a genome with very limited amounts of input DNA, as low as 10-30 ng compared to 5g for WGBS. Figure 1.3 [116] outlines the T-WGBS method.

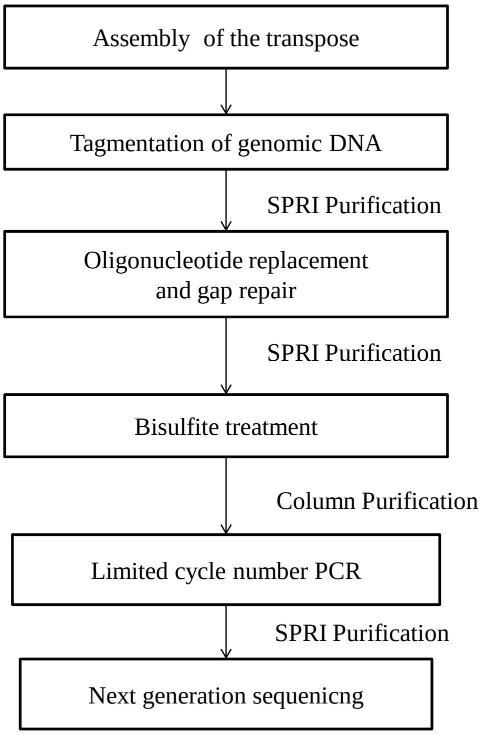

Assembly of the transpose
Tagmentation of genomic DNA
SPRI Purification
Oligonucleotide replacement and gap repair
SPRI Purification
Bisulfite treatment
Column Purification
Limited cycle number PCR
SPRI Purification
Next generation sequenicng

Figure 1.3: Illustration of T-WBS method. SPRI: Solid phase reversible immobilisation.

Reduced representation bisulfite sequencing (RRBS) employs a similar experimental approach to quantify the methylation of less than 10% of all CpG sites in the genome [117], which brings down cost. The genomic DNA is first digested with a methylation insensitive enzyme and cut at CCGG sites, which enriches for CpG rich regions. Next, the restriction fragments are size selected, equipped with adapters, bisulfite converted, PCR amplified, cloned and sequenced.

Analyses of BS-seq data

The best practices apply to BS-seq data contain three steps.

Quality control - Next generation sequencers assign a Phred quality score, which represents the probability of a base calling being wrong, to the called bases. A quality score less than 30 is commonly regarded as a poor quality. PCR artifacts, contamination, untrimmed adapter sequences and problems from sequencing itself can result in low quality read data. Low base call qualities, which often appear towards the end of next generation sequencing reads have to be eliminated because it can lead to inaccurate downstream analysis and data interpretation. Moreover, DNA fragments that are shorter than the read length will cause reads extent into adapter sequences. If a read extends into adapter by only a few bases it may align with mismatches and indels in the adapter region, leading to incorrect mapping. As fragments get shorter and the fraction of adapter sequences increases, the read will not align to the genome. Therefore, checking the quality of raw sequencing reads is the first step. Several tools such as FastQC [118] and PRINSEQ [119] are available to produce general quality assessment. Once the data are checked for quality, they should be processed to remove reads with low-quality bases, adapter sequences, and other contaminating sequences. Tools such as Cutadapt [120], Trimmomatic [121], TrimGalore [122], FASTX-Toolkit [123], which trim adapter or other contaminants based upon user-provided parameters, can be used for performing these operations.

Read mapping - Bisulfite sequencing short-read mapping relies on a reference genome, from which in silico bisulfite-converted genomes are generated for use in read alignment [124, 72]. Several approaches have been developed for the mapping of BS-seq reads, such as BSMAP [125], Bismark [126] , MethylCoder [127], BS Seeker [128], Last [129] and BRAT-BW [130]. These not only differ considerably in terms of alignment speed and flexibility but also in their output information. Note that some of these tools are not just aligners but can additionally extract methylation levels from the alignments such as Bismark and MethylCoder, which enable the end user to explore the biological effects of methylation more quickly. Bismark and BS-Seeker support the directional and the nondirectional BS-seq protocol and use Bowtie2 [131] as an internal read mapper. The tool MethylCoder is more flexible and able to use either GSNAP or Bowtie2. Besides it uses a similar strategy as Bismark and BS Seeker.

Methylation level calling - After alignment, the methylation states for genomic C positions can be estimated: C/T ratio of the mapped reads indicates unmethylated cytosines while C/C matches reveal methylcytosine. However, this is a rather inaccurate method. The main challenge is that sequencing errors, sequence variations, mis-mapping and bisulfite failures can lead to wrong inference of methylation levels. For example, sequence variation is traditionally disregarded in the analysis of WGBS data and a C/T single nucleotide variant would still align to a bisulfite-converted reference, but be regarded as an entirely unmethylated CpG site, even though the CpG site no longer exists. Given that over two thirds of all SNPs occur in a CpG context, having two alleles: C/T or G/A [132], it is important to take sequence variation in consideration to avoid wrong inference of methylation states.

To our knowledge, NGSmethPipe is the first program that conducted a threshold-based detection of sequence variation in bisulfite sequencing experiments [133]. This program reports the genomic postions of detected sequence variation in the output. Recent tools for calling SNP genotypes directly from bisulfite sequencing reads, including Bis-SNP [134] and BS-SNPer [135] have been developed. Bis-SNP, which is based on the Genome Analysis Toolkit (GATK) can identifies SNPs at high precision and estimates methylation levels. It is based on a Bayesian method and takes advantage of both top and bottom DNA strand information to discriminate SNPs from bisulfite conversions. In this way, C>T SNPs are no longer interpreted as unmethylated Cs. However, Bis-SNP supports only the directional BS-Seq protocol since it is not always known which strand non-directional reads originate. The disadvantages of Bis-SNP is that, despite its enhanced model for genotype calling, the methylation levels are simply estimated using the C-T ratio. BS-SNPer is a program for the detection of variation from BS-Seq alignments in standard BAM/SAM format [136]. It implemented a dynamic matrix algorithm and approximate Bayesian modeling and is much faster than Bis-SNP.

Chapter 2

Genomic determinants of somatic copy number alterations across human cancers

This chapter has been published in Zhang, Y., Xu, H., and Frishman, D.(2016) Genomic determinats of somatic copy number alterations across human cancers. *Hum. Mol. Genet.*, 25(5), 1019-1030. I and HongenXu contributed equally to this work. This study was designed by Dmitrij Frishman, Hongen Xu and me. I collected data and did mutliple linear regression, and Hongen Xu did logistic regression and extremely randomized tree classifier. The manuscript was written by myself and Hongen Xu, and corrected by Dmitrij Frishman.

2.1 Abstract

Somatic copy number alterations (SCNAs) play an important role in carcinogenesis. However, the impact of genomic architecture on the global patterns of SCNAs in cancer genomes remains elusive. In this work we conducted multiple linear regression (MLR) analyses of the pooled SCNA data from The Cancer Genome Atlas Pan-Cancer project. We performed MLR analyses for 11 individual cancer types and three different kinds of SCNAs-amplifications and deletions, telomere-bound and interstitial SCNAs and local SCNAs. Our MLR model explains more than 30% of the pooled SCNA breakpoint variation, with the explanatory power ranging from 13 to 32% for different cancer types and SCNA types. In addition to confirming previously identified features [*e.g.* long interspersed element-1 (L1) and short interspersed nuclear elements (SINEs)], we also identified

several novel informative features, including distance to telomere, distance to centromere and low complexity repeats. The results of the MLR analyses were additionally confirmed on an independent SCNA data set obtained from the Catalogue Of Somatic Mutations In Cancer (COSMIC) database. Using a rare event logistic regression model and an extremely randomized tree classifier, we revealed that genomic features are informative for defining common SCNA breakpoint hotspots. Our findings shed light on the molecular mechanisms of SCNA generation in cancer.

2.2 Introduction

Cancer is fundamentally a disease characterized by a diversity of somatic alterations [137]. Recently developed technologies, such as single nucleotide polymorphism (SNP) arrays and next-generation DNA sequencing have created unprecedented opportunities for studying different classes of mutations, including single base substitutions, small indels, genomic rearrangements, and somatic copy number alterations (SCNAs) [137, 138, 139] . The landscape of SCNAs has been charted across different types of cancer, with recurrent SCNAs often pointing at novel oncogenes and tumor suppressor genes [138, 140, 55]. Although SCNAs affect a sizeable fraction of the genome and are functionally important in carcinogenesis, their generation mechanisms are not yet fully understood.

Previous analyses of SCNA data have provided insights into the mechanisms shaping SCNA occurrence [138, 55, 141, 142]. SCNA breakpoints are not uniformly distributed in the genome, but rather tend to be spatially clustered in breakpoint hotspots [141]. For instance, G-quadruplex sequences (G4s) are enriched in the vicinity of SCNA breakpoints, suggesting the contribution of genomic properties to SCNA formation [141]. A recent comparative analysis has identified two types of SCNA breakpoint hotspots-cancer-type-specific SCNA breakpoint hotspots, which are enriched in known cancer genes, and common hotspots (CHSs). The latter can be relatively well predicted from genomic context by a multiple linear regression (MLR) model [143]. However, the model presented in [143] explains only a small part of the SCNA breakpoint variance [with the top four features-indel rate, exon density, substitution rate, and SINE coverage-being collectively responsible for 14% of the variation]. A model considering a much wider spectrum of genomic properties would be expected to better illuminate how different genomic features contribute to the global patterns of SCNAs in cancer genomes.

Many endogenous factors (such as non-B DNA conformations and repetitive sequences) can cause double-strand breaks (DSBs). Subsequent erroneous DNA repairs will result in copy number alterations [141, 144, 25]. Indeed, genome-wide mapping of DSBs has shown that DSB regions are enriched in genomic regions frequently rearranged in cancers [145]. Under certain circumstances, DNA can assemble into non-B conformations at specific sequence motifs including A-phased repeats, G-quadruplex, Z-DNA, inverted repeats, mirror repeats, and direct repeats [146]. The resulting DNA secondary structures have been implicated in the formation of structural alterations including CNVs, inversions and translocations, such as G-quadruplexes [141], Z-DNA [147], cruciforms formed by inverted repeats [148] and triplexes (also known as H-DNA) formed by mirror repeats [149]. Transposable elements are dispersed at high copy numbers throughout the human genome, and non-allelic homologous recombination between different copies of transposable elements can result in CNVs. For example, homologous recombination of non-allelic copies of L1 and human endogenous retroviral elements leads to the formation of CNVs [150, 151]. Moreover, a 13-mer CCNCCNTNNCCNC motif was found to associate with recombination hotspots in humans and was clustered in common mitochondrial deletion hotspots [152]. Recently, Zhou *et al.* [153] have revealed a significant enrichment of human germline and somatic structural variant breakpoints in self-chain (SC) regions, a group of low-copy repeats shorter than 1 kb. Besides the effects of local genomic context on CNV formation, TCGA Pan-Cancer analysis has suggested different mechanisms for telomere-bound SC-NAs and those SCNAs that are interstitial to chromosomes, highlighting the importance of chromosome structure (*e.g.* telomeres and centromeres) [55].

In this study, we selected genomic features, which have been proposed to affect SCNAs across the human genome, of which DSBs, SCs, recombination motifs, and distance to telomeres and centromeres have not been investigated in previous studies. We also include the histone marker H3K9me3, which accounts for more than 40% of mutation rate variation in cancer cells [154]. We built MLR and logistic regression (LR) models to explore the intrinsic basis of observed SCNA patterns. These statistical methods have been successful in contrasting common fragile sites and non-fragile sites [155] and investigating the effects of diverse sequence features on integration sites of DNA transposons [156].

The overview of our study is presented in Figure 2.1. Taking advantage of SCNAs data from the TCGA Pan-Cancer project and collected genomic features, we firstly selected predictors (genomic features) to reduce multicollinearity and identified common SCNA breakpoint hotspots and non-hotspots (NHSs) across Pan-Cancer types. We then built MLR models to investigate whether and how different genomic features contribute to the genome-wide patterns of SCNA breakpoints. We also applied LR and extremely randomized tree classifier to contrast between common SCNA breakpoint hotspots and NHSs. Our MLR models can explain more than 30% of SCNA breakpoint variation. The power of the models remain stable when one considers separately different SCNA types (amplifications and deletions), SCNA types of possible different generation mechanisms (telomere-bound SCNAs and interstitial SCNAs), and SCNAs from different cancer types. We also demonstrate that these genomic features are informative for telling apart common SCNA breakpoint hotspots and NHSs by logistic models and extremely randomized tree classifiers. This suggests that common breakpoint hotspots strongly depend on the local genomic context.

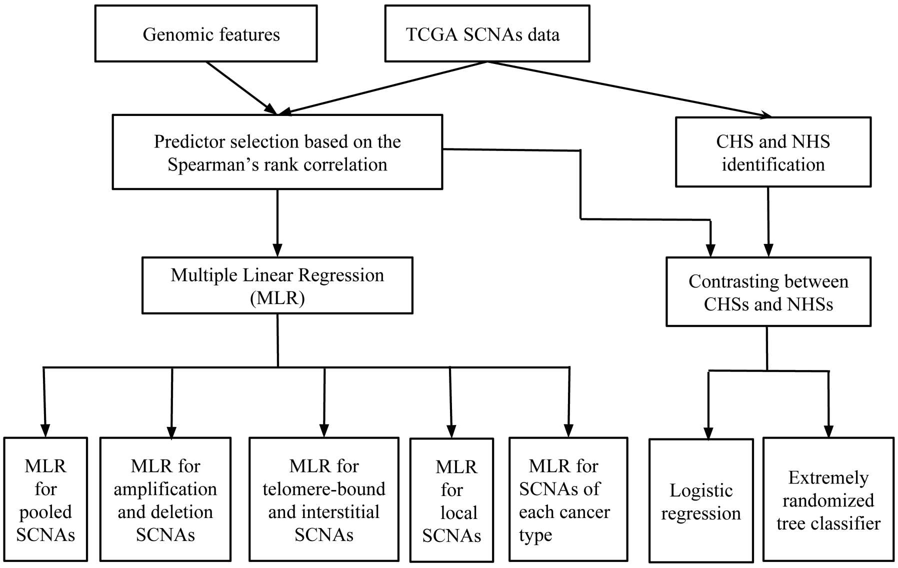

Genomic features
TCGA SCNAs data
Predictor selection based on the Spearman's rank correlation
CHS and NHS identification
Multiple Linear Regression (MLR)
Contrasting between CHSs and NHSs
MLR for pooled SCNAs
MLR for amplification and deletion SCNAs
MLR for telomere-bound and interstitial SCNAs
MLR for local SCNAs
MLR for SCNAs of each cancer type
Logistic regression
Extremely randomized tree classifier

Figure 2.1: An overview of the study design. TCGA: the cancer genome atlas; SCNA: somatic copy number alterations; CHS: common hotspots; NHS: non-hotspots.

2.3 Materials and Methods

2.3.1 SCNA data

The first SCNA data published in [55] were kindly provided by Travis I Zack and Rameen Beroukhim (Dana-Farber Cancer Institute, USA). SCNAs were obtained by mapping the signal intensities from the Affymetrix Genome-Wide Human SNP Array 6.0 in each cancer sample upon removing the probes in regions of recurrent germline CNVs identified from normal tissue samples. The data were provided as files with 105,890 and 96,354 individual SCNAs corresponding to amplifications and deletions. For each individual SCNA the files contain its chromosomal coordinates (chromosome number as well as start and end positions), TCGA barcode (sample identity), amplitude of copy number change and other information. We grouped SNCAs from the same cancer type based on the Pan-Cancer project sample information from http://www.synapse.org (syn1710466). Both boundaries of each SCNA were defined as breakpoints with a precision of about 1 kb (the median inter-marker distance for Affymetrix Genome-Wide Human SNP Array 6.0 is less than 700 bases). In total, we obtained 404,488 SCNA breakpoints from 4,943 samples across 11 cancer types, of which 211,780 and 192,708 breakpoints correspond to amplifications and deletions, respectively (Table 2.1). We also subdivided all SCNAs into two categories: telomere-bound SCNAs, with at least one boundary situated on a telomere, and interstitial SCNAs, with both boundaries interstitial to the chromosome. Specifically, for each chromosome we defined those SCNAs started at the left-most position or ended at the right-most position of the chromosome as telomere-bound SCNAs (see Figure 2.2). All the remaining SCNAs were considered to be interstitial. We further subdivided SCNAs into local and chromosome-level ones. Chromosome-level SCNAs were defined as those having the left boundary at the left-most position and the right boundary at the right-most position in the given chromosome, while all other SCNAs were considered local (Figure 2.2). By definition, all chromosome-level SCNAs are also telomere-bound, and all interstitial SCNAs are also local SCNAs. The second dataset was from the COSMIC database (version 73) [157], and we retrieved 699 492 SCNAs generated by studies other than TCGA (COSMIC study identifiers: 328, 382, 538, 585, 586, 589, and 650).

Table 2.1: Summary of somatic copy number alteration (SCNA) data from The Cancer Genome Atlas (TCGA) Pan-Cancer project

Cancer type	Abbr.	Sample size	SCNA breakpoints	Breakpoints					
				Amplification			Deletion		
				Interstitial	Telomere-bound		Interstitial	Telomere-bound	
				Local	Chr. level	Local	Local	Chr. level	Local
Bladder urothelialc carcinoma	BLCA	90	13344	4562	802	1172	3900	1326	1582
Breast invasive carcinoma	BRCA	745	99574	42268	2624	8792	25 414	8610	11866
Colon adenocarcinoma	COAD	349	21650	4222	2318	2004	6672	3966	2468
Glioblastoma multiforme	GBM	485	28462	10162	2078	1074	10234	2556	2358
Head and neck squamous cell carcinoma	HNSC	270	24272	6990	1130	3068	5586	3320	4178
Kidney renal clear cell carcinoma	KIRC	373	9040	1818	1024	860	1756	2230	1352
Lung adenocarcinoma	LUAD	292	34952	12080	1890	3430	8006	4882	4664
Lung squamous cell carcinoma	LUSC	261	34400	10828	1106	3998	7992	4628	5848
Ovarian serous cystadenocarcinoma	OV	457	92216	41238	2762	10720	19200	7176	11120
Rectum adenocarcinoma	READ	147	12358	2620	1114	1090	3694	2328	1512
Uterine corpus endometrial carcinoma	UCEC	376	34220	18014	1196	2726	6570	2132	3582
Total		3845	404488	154802	18044	38934	99024	43154	50530

Abbr., Abbreviation; Chr., Chromosome

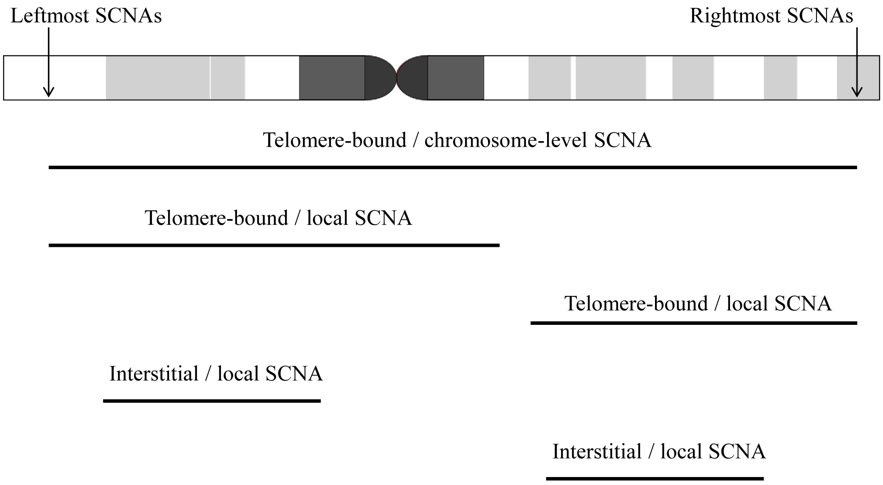

Leftmost SCNAs
Rightmost SCNAs
Telomere-bound / chromosome-level SCNA
Telomere-bound / local SCNA
Telomere-bound / local SCNA
Interstitial / local SCNA
Interstitial / local SCNA

Figure 2.2: Schematic illustration of SCNA categories considered in this work.

2.3.2 Data collection on genomic features

A total of 29 genomic features were considered as potential predictors of the SCNA patterns (Table 2.2). Their genomic coordinates were either obtained from public databases and published studies or identified in this study. All coordinates correspond to the human genome assembly hg19 and, where necessary, the University of California, Santa Cruz (UCSC) liftOver tool was used to convert the hg18 coordinates to hg19 [108].

Chromosomal coordinates of the following genomic features were downloaded from the UCSC Genome Browser [108]: probes of the Affymetrix Genome-Wide Human SNP Array 6.0 (retrieved from the SNP/CNV Arrays track); long terminal repeat (LTR) retrotransposons, L1, L2, SINE, DNA transposons and low-complexity repeats (retrieved from the RepeatMasker track); telomeres, centromeres, and genome assembly gaps (retrieved from the Gap track); microsatellites; simple repeats; CpG islands; exons and SCs. The latter elements are essentially pairs of short (up to 1 kb) low-copy repeats either in direct (+) or inverted (-) orientation [153]. Following [153] we only considered self-chain segments (SCS) consisting of paired SCs located on the same chromosome as well as their spacing gaps with the total lengths of up to 30 kb. Furthermore, we removed any SCSs overlapping with gaps in the human genome assembly (including centromeres, telomeres, heterochromatin regions, etc.) and segmental

Table 2.2: Genomic features used in the regression analyses

Category	Predictor	Measure	Source
DNA conformation	A-phased repeats	Coverage	Non-B DB version 2
	Mirror repeats	Count	Non-B DB version 2
	Direct repeats	Coverage	Non-B DB version 2
	Inverted repeats	Coverage	Non-B DB version 2
	Z-DNA	Coverage	Non-B DB version 2
	G4	log_{10}(count)	Non-B DB version 2
DNA sequence	Microsatellites	Coverage	UCSC Genome Browser
	SINEs	log_{10}(count)	UCSC Genome Browser
	L1	Coverage	UCSC Genome Browser
	L2	Coverage	UCSC Genome Browser
	LTR retrotransposons	Coverage	UCSC Genome Browser
	DNA transposons	Coverage	UCSC Genome Browser
	Low-complexity repeats	Coverage	UCSC Genome Browser
	Double-strand breaks	Coverage	Tchurikov *et al.* (2013)
	Self-chain segments	Coverage	This work
	GC content	Coverage	This work
	Simple repeats	Coverage	UCSC Genome Browser
Gene regulation	H3K9me3	Count	Barski *et al.* (2007)
	CpG islands	Coverage	UCSC Genome Browser
Chromosome structure	Distance to centromere	log_{10}(distance in bp)	This work
	Distance to telomere	log_{10}(distance in bp)	This work
Evolutionary features	Recombination motif	Coverage	This work
	Conserved DNA elements	Count	Siepel *et al.*(2005)
	Indel rate	Coverage	Human-Chimp alignment
	Substitution rate	Coverage	Human-Chimp alignment
Functional features	Replication timing	Sum	Hansen *et al.* (2010)
	Exon	Coverage	UCSC Genome Browser
	miRNA genes	Coverage	miRbase database
	Fragile sites	Yes/no	Fungtammasan *et al.* (2012)

duplications.

Non-B DNA motifs (A-phased repeats, direct repeats, inverted repeats, mirror repeats, G-quardruplexes (G4) and Z-DNA) were downloaded from the non-B DB version 2 [146]. We used the dataset of conserved DNA elements in vertebrates published by Siepel *et al.* [158]. Regions containing DSBs were downloaded from Tchurikov *et al.* [159]. Genomic coordinates for each histone modification marker H3K9me3 in CD4$^+$ T cells were obtained from the study of Barski *et al.* [160]. Replication timing (RT) data for the lymphoblastoid cell line GM06990 were obtained from Hansen *et al.* [161]. For each 1kb window of the genome sequence we obtained percent-normalized tag density values for the six phases of the cell cycle (denoted G1b, S1, S2, S3, S4 and G2). As suggested by the authors, a weighted average of the data based on the progression of each cell cycle was utilized, and RT was defined by the following formula:

$$RT = (0.917 \times G1b) + (0.75 \times S1) + (0.583 \times S2) + (0.417 \times S3) + (0.25 \times S4) + (0 \times G2).$$

Higher RT values correspond to earlier replication events. The percentage of G/C nucleotides (GC coverage) for specific genomic regions was calculated using the *nuc* utitlity, which is part of BEDTools [162]. The genome-wide distribution of the 13-mer CCNCCNTNNCCNC motifs related to recombination hotspots was obtained by *FUZZNUC* searches (as implemented in the European Molecular Biology Open Software Suite package [163]). We obtained the coordinates for fragile sites and miRNA genes from a previous study [155] and miRbase [164], respectively. The rates of nucleotide substitutions and indels were calculated based on human-chimpanzee alignments as described in [143].

2.3.3 Data transformation and prescreening of SCNA predictors

Genomic features described above were considered as potentially affecting the patterns of SCNA occurrence across the genome. We partitioned the human genome into non-overlapping 1 Mb windows, after excluding gaps in the genome assembly. The features were measured as counts (number of copies in a window), coverage (fraction of a window occupied by the feature), distance in base pairs to a telomere or a centromere, or sum (specifically, the sum of the RT values of 1kb fragments in a 1 Mb window) (Table 2.2). All features were evaluated for normality, and if necessary transformed by the logarithm function to approximate it (Table 2.2). In order to improve the efficiency of model selection for the subsequent regression analyses (see below) and reduce the influence of multicollinearity, we performed the same filtering process for the genomic features as in [155, 156]. We used hierarchical clustering to identify clusters of features based on Spearman's rank correlation coefficient using a threshold of 0.8. From each such cluster, we selected one representative feature, thus ensuring relatively low linear dependencies.

2.3.4 Identification of common hotspots and non-hotspots for breakpoints across cancer types

Breakpoint hotspots, i.e. genomic regions in which breakpoints are significantly enriched, were identified according to the method described in [141, 143, 165]. We split the human genome into non-overlapping 1 Mb windows and excluded from consideration windows with extremely low Affymetrix Genome-Wide Human SNP Array 6.0 probe density (below three standard deviations from the mean). The number of breakpoints for each cancer type was counted in each 1 Mb window. The same procedure was applied to SCNA breakpoint positions randomized 1000 times in order to generate the null distribution expected by chance. Randomization and counting of breakpoints were performed using BEDTools [162]. We assumed a normal distribution for the randomly generated samples and computed P-values from the parameterized normal cumulative density function. The windows with false discovery rate (FDR) corrected P <0.05 were defined as breakpoint hotspots. We defined the 1 Mb breakpoint hotspots shared in all 11 cancer types as CHSs and the 1 Mb windows which are not identified as breakpoint hotspot in any cancer type as NHSs. The remaining 1 Mb breakpoint hotspots were defined as non-common hotspots (NCHSs), including hotspots found in only one cancer type and hotspots identified in some, but not all cancer types.

2.3.5 Multiple linear regression analysis

MLR models an approximately continuous response on the predictors. MLR builds the linear relationship between the predictors and the response. All surveyed genomic features measured in 1 Mb segments were used as potential predictors of SCNA occurrence across the human genome. The density of SCNA breakpoints in every 1 Mb window was determined both for all cancer types pooled together and for each cancer type individually. In addition, in each window we also calculated the breakpoint density of copy number amplifications and deletions, as well as telomere-bound and interstitial SCNAs. Further, for each window we also computed the SCNA breakpoint densities after excluding chromosome-level SCNAs with both boundaries located approximately at telomeres. These densities were used as response variables for MLR.

To diagnose multicollinearity of each predictor, variance inflation factors (VIFs) were calculated to avoid problems caused by the instability of the coefficients. R^2 was used to capture the explanatory power of the MLR model. For the MLR model, the RCVE of each predictor was defined as:

$$RCVE = 1 - R^2_{reduced}/R^2_{full}$$

where R^2_{full} and $R^2_{reduced}$ denote the residual sum of squares of the full model (including all of the tested predictors) and the reduced model without the predictor of interest, respectively. Moreover, we tested the robustness of the MLR model by substituting some of the predictors with other highly correlated features. We performed k-fold cross validation [166] of the MLR model by randomly dividing the data into k-folds of the same size, using k-1 folds of the data as a training dataset, and testing the model on the remaining fold. The results from each fold test are combined to produce a single estimate, which we call k-fold MLR. The mean of the k-fold adjusted R^2 for the model and k-fold RCVE for each predictor are denoted as k-fold adjusted R^2 and k-fold RCVE, respectively.

All statistical analyses were performed in the R environment [167]. The MASS [168] and Car [169] packages were used to generate the common diagonostic plots (e.g., residual plots, Q-Q plots) and the QuantPsyc [170] package was used to calculate the standardized coefficient of predictors (with the signs of plus or minus denoting the positive or negative effect that predictors have on the response). The DAAG [171] package was used to perform k-fold cross validation. RCVEs were represented graphically in heatmaps. Predictors with FDR-corrected $P < 0.05$ are considered to be significant.

2.3.6 Distinguishing between common hotspots and non-hotspots by logistic regression

LR was used to distinguish between CHSs (binary response 1) and NHSs (binary response 0) using the same predictors as in the MLR model. To eliminate the possible small-sample size bias we increased the number of CHSs by applying a sliding procedure. Specifically, we divided the human genome into sliding windows of 1 Mb in length with a step size of 100 Kb. We also applied rare events logistic regression (RELR) [56] to reduce the sample imbalance bias. The RELR analysis was performed with the help of the statistical software Zelig (http://gking.harvard.edu/zelig) [172] using the same predictors as in the LR model. We used pseudo R^2 to capture the explanatory power of the LR and RELR models. The relative contribution of each predictor for both models (relative contribution to variance explained, RCVE) was calculated by the formula:

$$RCVE = [(D_0 - D) - (D_0 - D_{(-p)})]/(D_0 - D)$$

where D_0 and D are the null deviance and residual deviance of the model, respectively, and $D_{(-p)}$ is the deviance of the resulting model after removing the predictor of interest.

2.3.7 Distinguishing between common hotspots and non-hotspots by an extremely randomized tree classifier

A classification decision tree [173] is an input-output model represented by a tree structure. As a single decision tree usually suffers from high variance, ensembles of decision trees have been proposed to circumvent this problem. In this work, we applied the extremely randomized tree classifier to distinguish between CHSs and NHSs using the same features as in the MLR and LR models. The extremely randomized tree classifier is implemented in Scikit-Learn, a collection of Python modules of common machine learning algorithms (http://scikit-learn.org) [174]. We chose to build 500 trees to obtain robust results, growing each tree to its full depth. To balance the input data classes, sample weights were passed to the classifier. The predictive performance of the classifier was assessed by AUC obtained on the dataset by 5-fold cross-validation: in each validation round 80% of the data were used as the training data and the remaining 20% were used as the test data. The final AUC values were computed by averaging AUCs over the 5-folds. Feature importance in extremely randomized tree classifiers was assessed based on the mean decrease impurity importance, which gets computed and normalized in Scikit-Learn by default.

2.4 Results

2.4.1 Identification of SCNA breakpoint hotspots

In this work we analyzed data on 404488 SCNA breakpoints [55] in 11 cancer types (Table 2.1). To characterize the genome-wide patterns of SCNA occurrence, we divided the human genome into 1 Mb non-overlapping windows, after removing gaps, and calculated the density of SCNA breakpoints within each window. Based on the randomization procedure described in the Materials and Methods section, we identified 81-331 breakpoint hotspots in individual cancers (FDR-corrected P <0.05). As seen in Figure 2.3 different types of cancer often share breakpoint hotspots, but also have their specific hotspots. Based on the definitions in the Materials and Methods section, we identified 29 CHSs, 1824 NHSs and 685 NCHSs.

2.4.2 Human genomic features

To identify potential correlates of SCNA breakpoint patterns, we compiled a set of diverse genomic features, of which some, including non-B DNA sequences, and transposable elements, were previously investigated for their effects on SCNA breakpoints [143], while several other features, such as distance to centromere and DSBs, are used for this purpose in this work for the first time. In total, we examined 29 features that can be generally

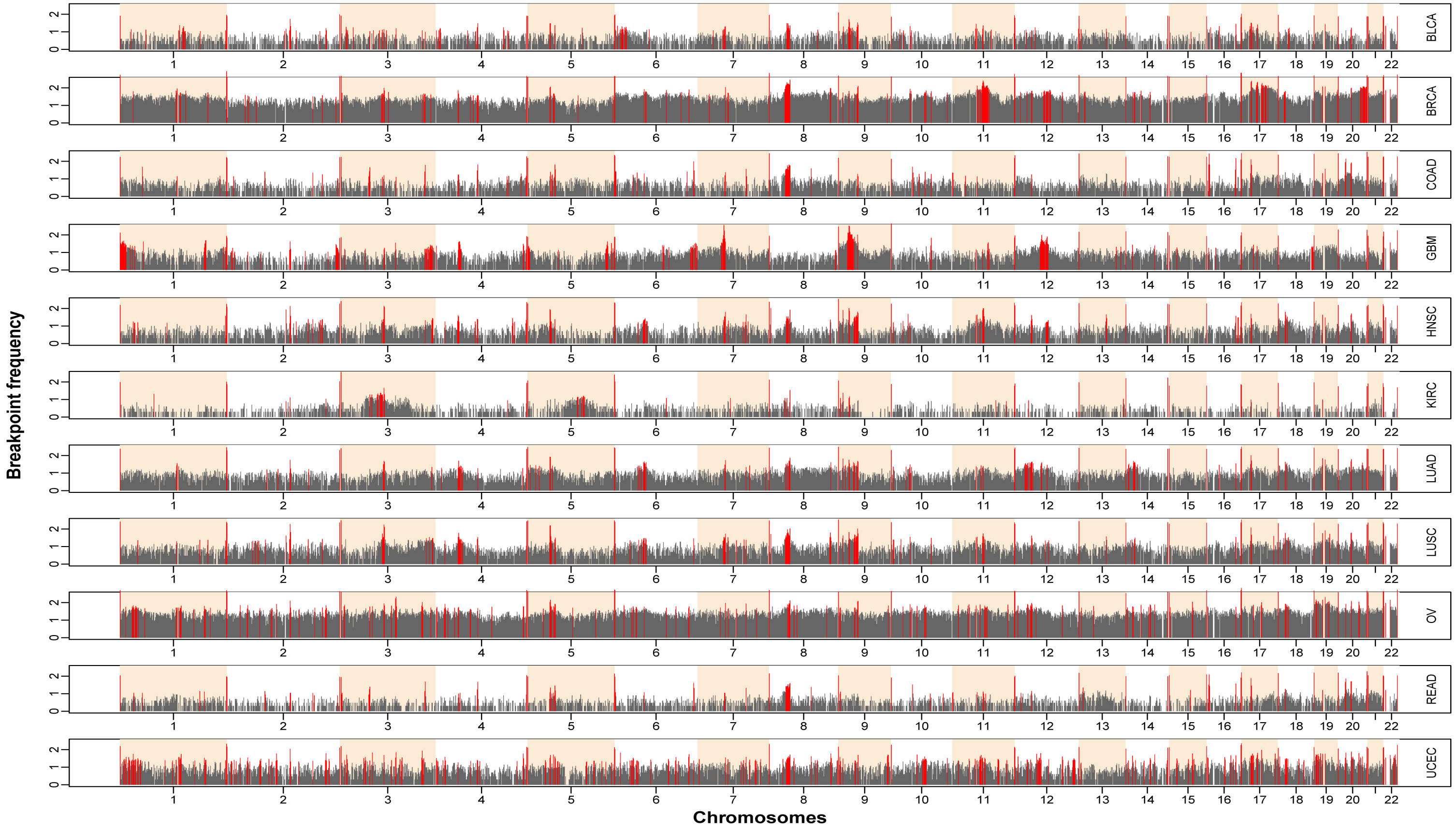


Figure 2.3: The distribution of SCNA breakpoint frequencies in 11 cancer types - BLCA, BRCA, COAD, GBM, HNSC, KIRC, LUAD, LUSC, OV, READ and UCEC (see Table 2.1 for full names), calculated as $= log_{10}$(the number of SCNA breakpoints in each block +1). Breakpoint hotspots in each cancer type are colored in red.

categorized into six groups: non-B DNA conformations; DNA sequence; gene regulation and expression; evolutionary features; chromosome structures; and functional features (Table 2.2). Following Fungtammasan *et al.* [155] and Campos-Sanchez *et al.* [156], we used hierarchical clustering with Spearman‘s rank correlation to remove some strongly correlated features (Figure 2.4). Finally, 25 features were used for subsequent regression analyses.

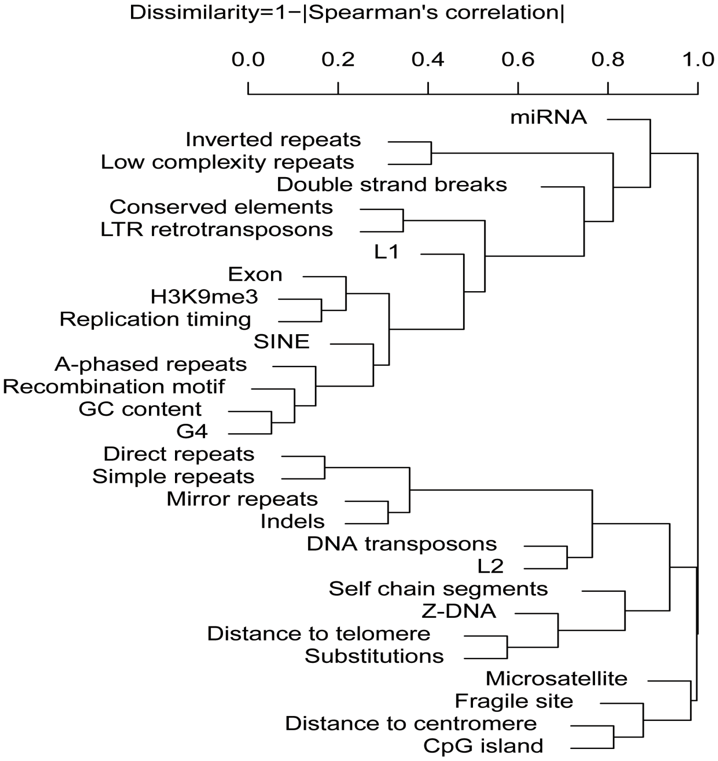


Figure 2.4: Hierarchical clustering of predictors based on their Spearman’s correlation coefficients.

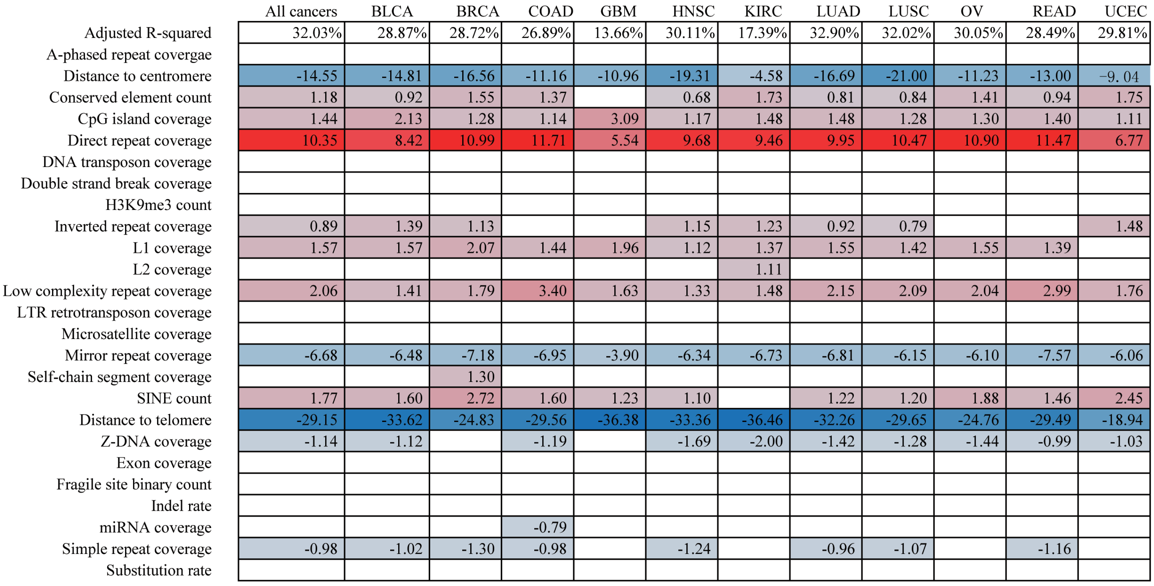
2.4.3 Impact of genomic features on the frequencies of SCNA breakpoints

We examined to what extent the observed genome-wide patterns of breakpoints could be explained by genomic features. Following an approach similar to the one described in [155, 156], the density of SCNA breakpoints (response) calculated in each 1 Mb window was represented as a function of the 25 genomic features (predictors) measured in the same 1 Mb window. The resulting MLR model accounted for 31.36% of the variation in the breakpoint density and contained 11 significant predictors (Table 2.3). The predictor with the strongest positive effect in the model is direct repeat coverage (10.35%). Other predictors with a significant positive effect are L1 coverage, low-complexity repeat coverage, SINE count, conserved DNA element count, CpG island coverage, and inverted repeat coverage with the relative contribution to variance explained (RCVE) ranging from 0.89% to 2.06% (Table 2.3; Figure 2.5). The predictors with the strongest negative effect are distance to telomere (29.15%) and distance to centromere (14.55%). Less significant predictors with a negative effect are mirror repeat count (6.68%), Z-DNA coverage (1.14%) and simple repeat coverage (0.98%).

Table 2.3: The MLR model for pooled SCNA breakpoints

Predictor	SCE	VIF	P-value	RC,%	Five-fold RC,%
Distance to centromere	-0.243	1.265	4.24E-38	14.55	19.76
Conserved element count	0.113	3.382	1.88E-04	1.18	1.07
CpG island coverage	0.072	1.133	3.88E-05	1.43	1.11
Direct repeat coverage	0.425	5.433	7.69E-28	10.35	11.97
Inverted repeat coverage	0.098	3.330	1.17E-03	0.89	0.51
L1 coverage	0.136	3.677	1.66E-05	1.57	1.67
Low complexity repeat coverage	0.142	3.069	8.34E-07	2.06	2.78
Mirror repeat count	-0.303	4.284	1.12E-18	6.68	7.70
SINE count	0.223	3.762	4.84E-06	1.77	1.87
Distance to telomere	-0.419	1.883	2.81E-72	29.15	32.21
Z-DNA coverage	-0.108	3.146	2.46E-04	1.14	Not significant
Simple repeat coverage	-0.087	2.434	6.67E-04	0.98	1.12
Adjusted R^2					31.36
Five-fold adjusted R^2					25.31

SCE, standardized coefficient; VIF, variance inflation factor; RC, relative contribution



	All cancers	BLCA	BRCA	COAD	GBM	HNSC	KIRC	LUAD	LUSC	OV	READ	UCEC
Adjusted R-squared	32.03%	28.87%	28.72%	26.89%	13.66%	30.11%	17.39%	32.90%	32.02%	30.05%	28.49%	29.81%
A-phased repeat covergae												
Distance to centromere	-14.55	-14.81	-16.56	-11.16	-10.96	-19.31	-4.58	-16.69	-21.00	-11.23	-13.00	−9. 04
Conserved element count	1.18	0.92	1.55	1.37		0.68	1.73	0.81	0.84	1.41	0.94	1.75
CpG island coverage	1.44	2.13	1.28	1.14	3.09	1.17	1.48	1.48	1.28	1.30	1.40	1.11
Direct repeat coverage	10.35	8.42	10.99	11.71	5.54	9.68	9.46	9.95	10.47	10.90	11.47	6.77
DNA transposon coverage												
Double strand break coverage												
H3K9me3 count												
Inverted repeat coverage	0.89	1.39	1.13			1.15	1.23	0.92	0.79			1.48
L1 coverage	1.57	1.57	2.07	1.44	1.96	1.12	1.37	1.55	1.42	1.55	1.39	
L2 coverage							1.11					
Low complexity repeat coverage	2.06	1.41	1.79	3.40	1.63	1.33	1.48	2.15	2.09	2.04	2.99	1.76
LTR retrotransposon coverage												
Microsatellite coverage												
Mirror repeat coverage	-6.68	-6.48	-7.18	-6.95	-3.90	-6.34	-6.73	-6.81	-6.15	-6.10	-7.57	-6.06
Self-chain segment coverage			1.30									
SINE count	1.77	1.60	2.72	1.60	1.23	1.10		1.22	1.20	1.88	1.46	2.45
Distance to telomere	-29.15	-33.62	-24.83	-29.56	-36.38	-33.36	-36.46	-32.26	-29.65	-24.76	-29.49	-18.94
Z-DNA coverage	-1.14	-1.12		-1.19		-1.69	-2.00	-1.42	-1.28	-1.44	-0.99	-1.03
Exon coverage												
Fragile site binary count												
Indel rate												
miRNA coverage				-0.79								
Simple repeat coverage	-0.98	-1.02	-1.30	-0.98		-1.24		-0.96	-1.07		-1.16	
Substitution rate												

Figure 2.5: The effect of genomic features in MLR models. The intensity of color is proportional to the RCVE in each model. Predictors in white color are not significant. See Table 2.1 for full names of cancer types.

We repeated the same analysis replacing some of the predictors with highly correlated predictors. For example, A-phased repeat coverage was replaced with GC content, recombination motif coverage or G4 count and we observed slight changes in both the RCVE of predictors and R^2 of models. Most of genomic features remained significant in these alternative models (Tables 2.4, 2.5, 2.6, 2.7).

Table 2.4: Alternative MLR replacing A-phased repeat with GC content

Predictor	SCE	VIF	P-value	RC,%	Five-fold RC,%
Distance to centromere	-0.244	1.261	1.47E-38	14.71	19.93
Conserved element count	0.117	3.418	1.18E-04	1.25	1.19
CpG island coverage	0.074	1.135	2.39E-05	1.51	1.29
Direct repeat coverage	0.436	5.332	9.84E-30	11.09	13.32
L1 coverage	0.134	3.659	2.07E-05	1.53	1.79
Low-complexity repeat coverage	0.140	3.084	1.38E-06	1.97	2.71
Mirror repeat count	-0.309	4.324	2.93E-19	6.90	8.08
SINE count	0.246	9.761	1.75E-06	1.94	1.95
Distance to telomere	-0.418	1.864	1.90E-72	29.16	32.51
Simple repeat coverage	-0.085	2.383	8.22E-04	0.95	1.04
Adjusted R^2					31.41
Five-fold adjusted R^2					24.40

SCE, standardized coefficient; VIF, variance inflation factor; RC, relative contribution

Table 2.5: Alternative MLR replacing A-phased repeat with recombination motif

Predictor	SCE	VIF	P-value	RC,%	Five-fold RC,%
Distance to centromere	-0.243	1.260	2.46E-38	14.61	19.80
Conserved element count	0.116	3.393	1.38E-04	1.23	1.16
CpG island coverage	0.073	1.132	2.77E-05	1.49	1.15
Direct repeat coverage	0.429	5.244	2.45E-29	10.93	13.26
Inverted repeat coverage	0.096	3.330	1.46E-03	0.86	0.44
L1 coverage	0.139	3.664	1.05E-05	1.64	1.88
Low-complexity repeat coverage	0.144	3.082	6.25E-07	2.10	2.85
Mirror repeat count	-0.300	4.294	2.53E-18	6.52	7.79
SINE count	0.252	10.209	1.66E-06	1.94	2.06
Distance to telomere	-0.416	1.869	6.81E-72	28.91	31.88
Z-DNA coverage	-0.096	3.334	1.46E-03	0.86	-0.22
Simple repeat coverage	-0.086	2.364	7.03E-04	0.97	1.08
Adjusted R^2					31.42
Five-fold adjusted R^2					24.43

SCE, standardized coefficient; VIF, variance inflation factor; RC, relative contribution

Table 2.6: Alternative MLR replacing A-phased repeat with G4

Predictor	SCE	VIF	P-value	RC,%	Five-fold RC,%
Distance to centromere	-0.243	1.260	3.28E-38	14.60	19.81
Conserved element count	0.108	3.510	4.85E-04	1.03	0.88
CpG island coverage	0.072	1.133	4.22E-05	1.42	1.19
Direct repeat coverage	0.425	5.336	2.47E-28	10.56	12.55
Inverted repeat coverage	0.100	3.319	8.91E-04	0.94	0.57
L1 coverage	0.133	3.753	3.07E-05	1.47	1.58
Low-complexity repeat coverage	0.139	3.199	2.51E-06	1.88	2.48
Mirror repeat count	-0.301	4.332	2.56E-18	6.54	7.73
SINE count	0.205	8.261	1.50E-05	1.59	1.66
Distance to telomere	-0.419	1.869	1.12E-72	29.35	32.54
Z-DNA coverage	-0.125	3.837	1.06E-04	1.27	0.53
Simple repeat coverage	-0.094	2.342	2.07E-04	1.17	1.30
Adjusted R^2					31.35
Five-fold adjusted R^2					24.21

SCE, standardized coefficient; VIF, variance inflation factor; RC, relative contribution

Table 2.7: Alternative MLR replacing H3K9me3 with replication timing

Predictor	SCE	VIF	P-value	RC,%	Five-fold RC,%
Distance to centromere	-0.244	1.258	1.01E-38	14.77	19.74
Conserved element count	0.115	3.387	1.41E-04	1.23	1.16
CpG island coverage	0.071	1.133	5.01E-05	1.39	1.03
Direct repeat coverage	0.417	5.420	4.77E-27	10.01	11.51
Inverted repeat coverage	0.103	3.322	5.75E-04	1.00	0.70
L1 coverage	0.140	3.667	9.81E-06	1.65	1.86
Low-complexity repeat coverage	0.145	3.073	5.12E-07	2.14	2.87
Mirror repeat count	-0.298	4.302	3.96E-18	6.45	7.40
SINE count	0.198	7.809	1.65E-05	1.57	1.49
Distance to telomere	-0.422	1.879	3.42E-73	29.49	32.27
Z-DNA coverage	-0.118	2.837	2.25E-05	1.52	0.16
Simple repeat coverage	-0.088	2.335	4.43E-04	1.04	1.14
Adjusted R^2					31.43
Five-fold adjusted R^2					24.67

SCE, standardized coefficient; VIF, variance inflation factor; RC, relative contribution

We next applied MLR for breakpoints of two SCNA types-amplifications and deletions-separately. The MLR model explained 29.52% (amplifications) and 27.88% (deletions) of response variance. Notably, the predictors and the sign of their effect revealed by these two MLR models are similar to those of pooled SCNA breakpoints (Tables 2.8, 2.9), although some differences were apparent. For instance, Z-DNA repeat coverage, which had negative effect when both types of breakpoints were considered, disappeared in the MLR model for amplification breakpoints. Likewise, inverted repeat coverage lost its positive effect in the MLR model for deletion breakpoints.

Table 2.8: MLR for SCNA amplification breakpoints

Predictor	SCE	VIF	P-value	RC,%	Five-fold RC,%
Distance to centromere	-0.293	1.265	1.88E-52	22.39	31.04
Conserved element count	0.118	3.382	1.17E-04	1.37	1.38
CpG island coverage	0.056	1.133	1.52E-03	0.93	0.73
Direct repeat coverage	0.347	5.433	7.82E-19	7.34	5.73
Inverted repeat coverage	0.123	3.330	5.50E-05	1.50	1.83
L1 coverage	0.121	3.677	1.51E-04	1.32	0.60
Low-complexity repeat coverage	0.106	3.069	2.73E-04	1.22	0.07
Mirror repeat count	-0.247	4.284	1.17E-12	4.70	5.61
SCS coverage	0.065	1.375	9.83E-04	1.00	Not Significant
SINE count	0.218	8.762	1.06E-05	1.79	1.34
Distance to telomere	-0.411	1.884	4.54E-68	29.73	31.79
Simple repeat coverage	-0.120	2.434	4.12E-06	1.96	Not Significant
Adjusted R^2					29.52
Five-fold adjusted R^2					21.46

SCE, standardized coefficient; VIF, variance inflation factor; RC, relative contribution

Distance to telomere is a predictor with the strongest negative effect for both pooled SCNA breakpoints and the breakpoints corresponding to the two individual SCNA types-amplifications and deletions (Tables 2.3,

Table 2.9: MLR for SCNA deletion breakpoints

Predictor	SCE	VIF	P-value	RC,%	Five-fold RC,%
Distance to centromere	-0.192	1.265	1.02E-23	10.23	13.68
Conserved element count	0.099	3.382	1.36E-03	1.02	0.34
CpG island coverage	0.074	1.133	4.01E-05	1.68	Not Significant
Direct repeat coverage	0.426	5.433	9.81E-27	11.66	12.54
L1 coverage	0.131	3.677	5.21E-05	1.63	1.63
Low-complexity repeat coverage	0.148	3.069	5.67E-07	2.50	2.09
Mirror repeat count	-0.304	4.284	5.17E-18	7.56	8.55
SINE count	0.205	8.762	4.32E-05	1.67	1.19
Distance to telomere	-0.383	1.884	1.42E-58	27.30	33.00
Z-DNA coverage	-0.119	3.214	8.70E-05	1.54	Not Significant
Adjusted R^2					27.88
Five-fold adjusted R^2					19.48

SCE, standardized coefficient; VIF, variance inflation factor; RC, relative contribution

2.8, 2.9). In order to remove the confounding effect of this parameter, we next divided SCNAs into two categories: telomere-bound SCNAs, with one boundary located in the telomere and interstitial SCNAs, with both boundaries interstitial to the chromosome [55]. MLR models accounted for 31.90 and 20.24% of the variation for telomere-bound SCNAs and interstitial SCNAs, respectively. Significant predictors of telomere-bound and interstitial SCNAs are listed in Tables 2.10 and 2.11. Distance to telomere is a dominant predictor for telomere-bound SCNAs (relative contribution of 29.97%), while for interstitial SCNAs the most significant predictor is distance to centromere (relative contribution of 45.91%). Distance to centromere and SINEs are also significant for both SCNA types. However, the relative contribution of distance to centromere is substantially reduced for the telomere-bound SCNAs compared with interstitial SCNAs. Moreover, the other significant predictors for telomere-bound SCNAs are quite different from the significant predictors for the interstitial SCNAs.

Table 2.10: MLR for telomere-bounded SCNA breakpoints

Predictor	SCE	VIF	P-value	RC,%	Five-fold RC,%
Distance to centromere	-0.163	1.265	1.35E-18	6.49	7.48
Conserved element count	0.109	3.382	3.24E-04	1.07	1.03
CpG island coverage	0.070	1.133	6.38E-05	1.32	0.22
Direct repeat coverage	0.439	5.433	7.06E-30	10.91	10.07
L1 coverage	0.160	3.677	3.52E-07	2.15	2.18
Low-complexity repeat coverage	0.154	3.069	9.67E-08	2.36	2.20
Mirror repeat count	-0.329	4.284	6.39E-22	7.78	8.32
SINE count	0.184	8.762	1.57E-04	1.18	1.10
Distance to telomere	-0.429	1.884	8.74E-76	29.97	31.98
Z-DNA coverage	-0.115	3.214	9.05E-05	1.27	0.60
Adjusted R^2					31.90
Five-fold adjusted R^2					24.40

SCE, standardized coefficient; VIF, variance inflation factor; RC, relative contribution

By definition, the breakpoints of chromosome-level SCNAs are fixed at

Table 2.11: MLR for intersttial SCNA breakpoints

Predictor	SCE	VIF	P-value	RC,%	Five-fold RC,%
Distance to centromere	-0.349	1.265	6.63E-65	45.91	53.44
H3K9me3 count	0.143	2.272	9.89E-08	4.27	2.80
LTR coverage	-0.090	2.206	6.65E-04	1.74	1.95
SINE count	0.178	8.762	7.12E-04	1.72	1.53
Simple repeat coverage	-0.122	2.434	1.07E-05	2.91	2.58
Adjusted R^2					20.24
Five-fold adjusted R^2					14.95

SCE, standardized coefficient; VIF, variance inflation factor; RC, relative contribution

telomeres. We therefore excluded chromosome-level SCNAs from all the pooled SCNAs before conducting MLR analyses. We found that the model could explain 30.36% of the variation and included 10 significant predictors (Table 2.12). Notably, the predictors and their effect are similar to those of pooled SCNAs.

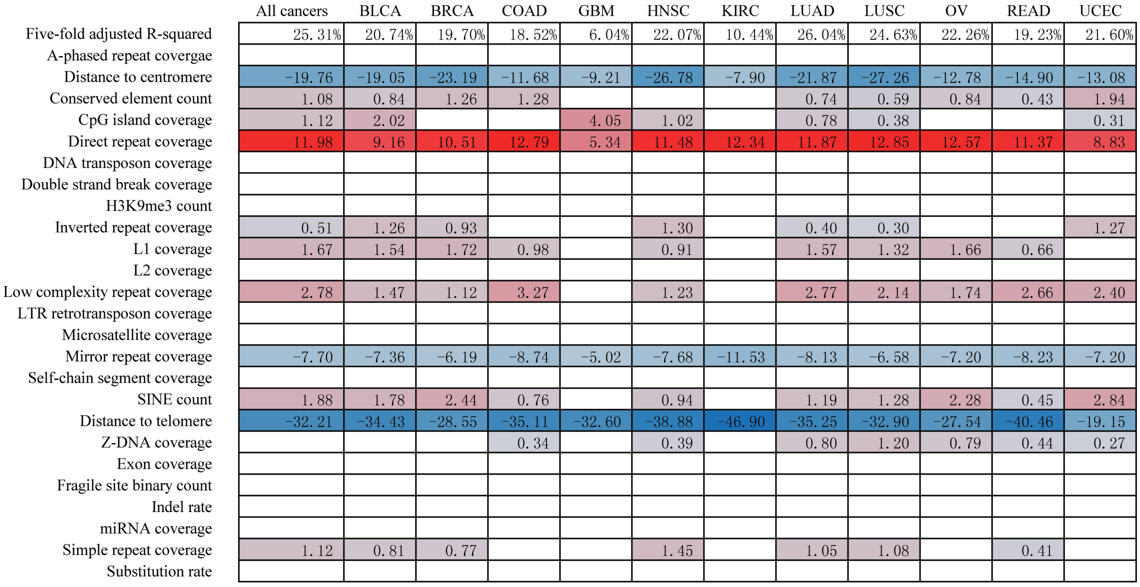
Table 2.12: MLR for SCNA breakpoints after excluding chromosome-level SCNAs

Predictor	SCE	VIF	P-value	RC,%	Five-fold RC,%
Distance to centromere	339	1.265	1.24E-69	29.30	41.94
Conserved element count	0.097	3.382	1.49E-03	0.89	0.67
CpG island coverage	0.086	1.133	1.01E-06	2.13	0.01
Direct repeat coverage	0.370	5.433	2.38E-21	8.11	10.09
Inverted repeat coverage	0.114	3.330	1.60E-04	1.26	1.39
Low-complexity repeat coverage	0.092	3.069	1.52E-03	0.89	0.52
Mirror repeat count	-0.229	4.284	3.00E-11	3.94	3.53
SINE count	0.222	8.762	6.40E-06	1.81	1.73
Distance to telomere	-0.391	1.884	1.38E-62	26.08	30.43
Simple repeat coverage	-0.115	2.434	8.58E-06	1.76	1.78
Adjusted R^2					30.36
Five-fold adjusted R^2					22.48

SCE, standardized coefficient; VIF, variance inflation factor; RC, relative contribution

We also performed similar analyses for each cancer type and found the adjusted R^2 of models to be greater than 26% for all cancer types except for glioblastoma multiforme (13.66%) and kidney renal clear cell carcinoma (17.39%). Similar to the MLR model of the pooled SCNA breakpoints, we identified direct repeat coverage, L1 coverage, low-complexity repeat coverage and SINE count as significant positive predictors for almost all cancer types (Figure 2.5). The distance to telomere, distance to centromere and mirror repeat count remained significant negative predictors for each cancer type (Figure 2.5).

We also conducted 5-fold cross validation for all the MLR models. While the MLR model trained over the pooled breakpoint dataset yielded an adjusted R^2 of 31.36%, the R^2 of the 5-fold MLR built from the pooled breakpoint dataset was 25.31% (Table 2.3). Moreover, the significant predictors and their effects identified in 5-fold MLR are similar to those of MLR (Table 2.3). The 5-fold MLR results for the other MLR models are provided in Tables 2.4-2.12 and Figure 2.6. The consistency between the MLR model and 5-fold MLR model indicates that the MLR model demonstrates good predictive ability and generalizes well on validation datasets.



	All cancers	BLCA	BRCA	COAD	GBM	HNSC	KIRC	LUAD	LUSC	OV	READ	UCEC
Five-fold adjusted R-squared	25.31%	20.74%	19.70%	18.52%	6.04%	22.07%	10.44%	26.04%	24.63%	22.26%	19.23%	21.60%
A-phased repeat covergae												
Distance to centromere	−19.76	−19.05	−23.19	−11.68	−9.21	−26.78	−7.90	−21.87	−27.26	−12.78	−14.90	−13.08
Conserved element count	1.08	0.84	1.26	1.28				0.74	0.59	0.84	0.43	1.94
CpG island coverage	1.12	2.02			4.05	1.02		0.78	0.38			0.31
Direct repeat coverage	11.98	9.16	10.51	12.79	5.34	11.48	12.34	11.87	12.85	12.57	11.37	8.83
DNA transposon coverage												
Double strand break coverage												
H3K9me3 count												
Inverted repeat coverage	0.51	1.26	0.93			1.30		0.40	0.30			1.27
L1 coverage	1.67	1.54	1.72	0.98		0.91		1.57	1.32	1.66	0.66	
L2 coverage												
Low complexity repeat coverage	2.78	1.47	1.12	3.27		1.23		2.77	2.14	1.74	2.66	2.40
LTR retrotransposon coverage												
Microsatellite coverage												
Mirror repeat coverage	−7.70	−7.36	−6.19	−8.74	−5.02	−7.68	−11.53	−8.13	−6.58	−7.20	−8.23	−7.20
Self-chain segment coverage												
SINE count	1.88	1.78	2.44	0.76		0.94		1.19	1.28	2.28	0.45	2.84
Distance to telomere	−32.21	−34.43	−28.55	−35.11	−32.60	−38.88	−46.90	−35.25	−32.90	−27.54	−40.46	−19.15
Z-DNA coverage				0.34		0.39		0.80	1.20	0.79	0.44	0.27
Exon coverage												
Fragile site binary count												
Indel rate												
miRNA coverage												
Simple repeat coverage	1.12	0.81	0.77			1.45		1.05	1.08		0.41	
Substitution rate												

Figure 2.6: The effect of genomic features in 5-fold MLR models. The intensity of color is proportional to the RCVE of each model. Predictors in white color are not significant. See Table 2.1 for full names of cancer types.

We also assessed the generalization ability of our MLR model on an independent dataset obtained from the COSMIC database (see Materials and Methods section). On this dataset the MLR model and the 5-fold MLR model accounted for 41.16% and 36.99% of breakpoint variation, respectively (Table 2.13). The most significant predictors, e.g., distance to telomere, mirror repeats and distance to centromere identified in the MLR model for pooled breakpoints from TCGA are also found to be significant in the MLR model on the independent dataset. However, predictors, including exon coverage, H3K9me3 count, LTR retrotransposon coverage, and indel rate, gained significance in this data set. Exon coverage and indel rate are among the top four features in the model presented in [143].

2.4.4 Contrasting between CHSs and NHSs by logistic regression

We investigated how genomic context affects the distribution of common breakpoint hotspots in cancer genomes. To this end we built a standard LR model using 25 features. The final standard LR model had a pseudo

Table 2.13: MLR for SCNA breakpoints from an independent data set

Predictor	SCE	VIF	P-value	RC,%	Five-fold RC,%
A-phased repeats coverage	-0.133	5.312	2.15E-04	0.79	0.78
Distance to centromere	-0.086	1.299	1.24E-06	1.36	1.29
CpG island coverage	0.059	1.198	4.66E-04	0.71	0.67
H3K9me3 count	-0.153	3.072	2.08E-08	1.82	1.87
LTR retrotransposon coverage	-0.099	2.230	1.89E-05	1.06	0.94
Mirror repeat count	-0.128	4.447	9.17E-05	0.88	0.67
Distance to telomere	-0.212	1.634	5.48E-26	6.56	7.12
Exon coverage	0.202	3.551	6.70E-12	2.74	2.87
Indel rate	0.121	5.124	5.85E-04	0.68	0.69
Adjusted R^2					41.16
Five-fold adjusted R^2					36.99

SCE, standardized coefficient; VIF, variance inflation factor; RC, relative contribution

R^2 51.83% and comprised two highly significant genomic features: distance to telomere (individual contribution 20.70%) and direct repeat coverage (individual contribution 5.16%).

However, the standard LR model may suffer from small-sample bias and class imbalance. In this work, the sample size of CHSs is small (sample size: 29) and sample sizes for NHSs and CHSs are imbalanced (1824 vs 29). For this reason, besides standard LR, we performed the rare events logistic regression (RELR). The estimates of a RELR model are corrected for class imbalance. Moreover, to eliminate the possible small-sample bias, we increased the number of common cancer hotspots by a sliding process, in which we divided the human genome into 1 Mb overlapping widows with a step size of 100 kb. Following the hotspot identification procedure described in Materials and Methods section, we identified 231 CHSs. The RELR model has a pseudo R^2 51.83% and contains 12 significant predictors (Table 2.14; Figure 2.7). The strongest feature discriminating CHSs and NHSs was distance to telomere (individual contribution 20.70%). This was a negative predictor, indicating that CHSs tend to be positioned closely to telomere. Direct repeat coverage is the strongest significant positive predictor (with the individual contribution of 5.16%), which implies that CHSs are located preferably in a genomic context that is enriched in direct repeats. We also performed RELR to contrast between non-common hotspots (NCHSs) and NHSs as well as between NCHSs and CHSs. We found that genomic features cannot discriminate between them (data not shown).

Table 2.14: RELR for contrasting CHSs with NHSs

Predictor	Standardized coefficient	P-value	Relative contribution,%
Conserved elements count	5.029	5.18E-04	1.01
CpG island coverage	1.825	1.04E-06	1.14
Direct repeats coverage	11.257	2.16E-11	5.16
DNA coverage	-5.251	3.82E-05	2.02
L1 coverage	8.253	1.87E-09	2.95
L2 coverage	-4.857	2.02E-05	1.61
Low-complexity repeats coverage	3.746	1.56E-04	1.08
Mirror repeat count	-2.741	5.41E-03	0.67
SINE count	10.513	6.26E-08	2.50
Distance to telomere	-44.259	4.50E-27	20.70
Z-DNA coverage	-4.025	1.16E-05	1.61
Simple repeat coverage	-6.701	9.29E-04	1.02
Explained Deviance			51.83

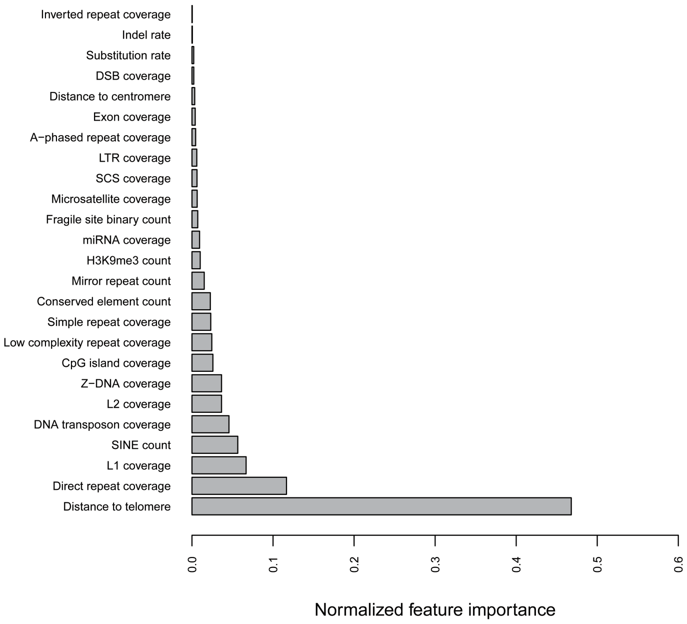


Figure 2.7: The normalized relative contribution of predictors in terms of distinguishing CHSs and NHSs for the RELR model.

Interestingly, the important features determined by the model, such as distance to telomere, direct repeat coverage, distance to centromere and L1 coverage, were also identified to have significant effects on SCNA breakpoint in the MLR models.

2.4.5 Extremely randomized tree classifier for telling apart CHSs and NHSs

We applied the extremely randomized tree classifier to distinguish CHSs and NHSs using the same 25 features. For the CHSs, this classifier reaches the area under the receiver operating characteristic (ROC) curve (AUC) of 0.96 (Figure 2.8a). The important features determined by the classifier for CHSs are distance to telomere, indel rate, and direct repeats (Figure 2.8b), which is generally consistent with the predictors identified in the RELR model. These results suggest that the positions of common breakpoint hotspots can be reasonable well predicted from local genomic properties.

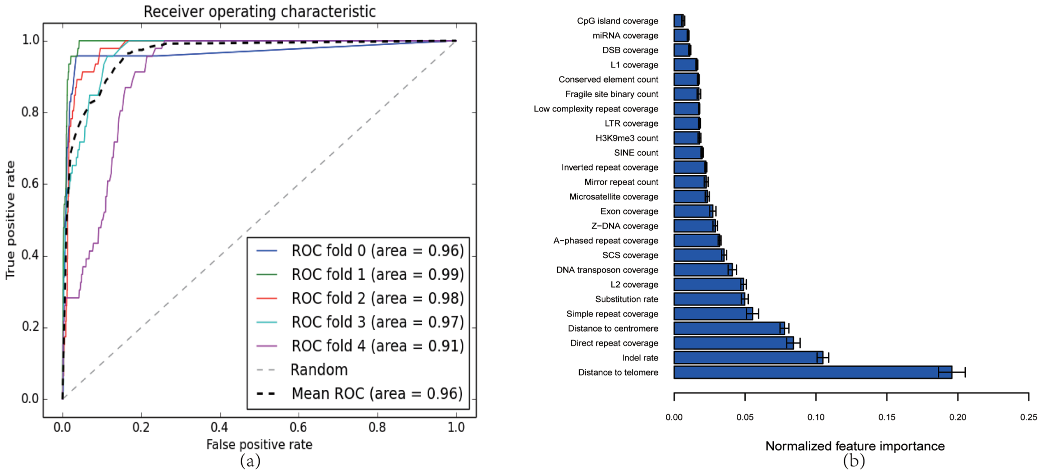


Figure 2.8: Distinguishing CHSs from NHSs from genomic features. (a) ROC-AUC curves of the extremely randomized forests. (b) The normalized relative contribution of predictors in terms of distinguishing CHSs and NHSs.

2.5 Discussion

Using a MLR model trained on 19 genomic properties, a previous study revealed top four genomic features, including indel rate, exon density, substitution rate and SINE coverage, contributing to SCNA breakpoint formation [143]. Taking advantage of the TCGA Pan-Cancer SCNA data, we considered a wider range of genomic features than in [143] and performed prescreening of features to reduce the effect of multicollinearity. Our MLR model is more than two times more powerful than that in [143] (32% of breakpoint variance explained versus 14%) and maintains its strong performance upon 5-fold cross validation. By including six novel genomic features, our models revealed two novel predictors-distance to telomere and distance to centromere, which made the strongest contribution to our model (relative contribution of 29.15 and 10.35% to MLR model for pooled SCNA breakpoints). The inclusion of these two features may explain the superiority of our model compared with that described in [143]. Notably, out of the top four features reported in [143] SINE coverage ranked sixth in predictive importance in our model, while the other three features-indel rate, exon density and substitution rate-were not among the significant predictors in our model (rank below 13th, see Table 2.15). When applying the same model to an independent data set, exon density and indel rate have some predictive power and rank second and last, respectively (Table 2.13). We, thus, encountered some discrepancies between the results obtained on the TCGA data and the independent COSMIC dataset. However, we found that distance to telomere, distance to centromere, CpG island coverage and mirror repeat count affect SCNA formation in both data sets, and the general consistency of the results obtained on these two datasets emphasizes the reliability of our findings. The power of the models was upheld for different SCNA types (amplifications and deletions), for SCNAs generated by distinct mechanisms (telomere-bound SCNAs and interstitial SCNAs) and for SCNAs from different cancer types. The TCGA Pan-Cancer analysis has revealed two types of SCNAs: interstitial SCNAs and telomere-bound ones [55]. The frequency of interstitial SCNAs is inversely correlated with their lengths [138, 55], while the telomere-bound ones tend to follow a uniform length distribution [55], which reflects distinct mechanisms underlying their formation. Indeed, in our study distance to centromere contributes strongly to the MLR model for interstitial SCNAs, while distance to centromere has a much smaller role than distance to telomere and direct repeat coverage in the MLR model for telomere-bound SCNAs. According to the MLR model the breakpoints of interstitial SCNAs are overrepresented close to centromeres, which is consistent with the previous observations [55, 175, 176]. Frequent breakages near centromeres may lead to their dysfunction and further cause chromosomal instability [177], which is a hallmark of diverse cancers [178]. The prevalence of telomere-bound SCNAs in cancers may relate to telomere dysfunction [179], and those breakpoints of telomere-bound SCNAs that are not located in telomeres were speculated to occur at regions with DSBs [55]. Our MLR models for telomere-bound SCNAs favor this hypothesis and demonstrate frequent occurence of DSBs in regions enriched in direct repeats. Direct repeats have been documented previously to cause hairpins and to overlap with chromosome regions undergoing somatic rearrangements [180]. The high prediction power of direct repeats in every cancer type suggests their significant common role in shaping the distribution of SCNA breakpoints. We also demonstrate that mirror repeat count, L1 coverage, SINE count, low-complexity repeat coverage and several other features have important albeit smaller roles in our MLR models. SINEs and L1 have been extensively studied for their roles in non-allelic homologous recombination, which leads to deletions, duplications and inversions [150, 181]. The significant positive effect of low-complexity repeats for all cancer types is in line with the fact that they are usually AT-rich and prone to causing the replication fork to pause or stall [182] and thus induce breaks. Moreover, AT-rich repeats constitute unstable regions of the genome, conferring susceptibility to rearrangements [183]. These results suggest a general mechanism of genome instability induced by genomic context.

Table 2.15: List of all features ranked by relative contribution to SCNA breakpoints formation in MLR model

Predictor	Relative contribution,%	Rank
Distance to telomere	29.15	1
Distance to centromere	14.55	2
Direct repeat coverage	10.35	3
Mirror repeat count	6.68	4
Low-complexity repeat coverage	2.06	5
SINE count	1.77	6
L1 coverage	1.57	7
CpG island coverage	1.44	8
Z-DNA coverage	1.14	9
Conserved element count	1.18	10
Simple repeat coverage	0.98	11
Inverted repeat coverage	0.89	12
H3K9me3 count	0.48	13
Indel rate	0.35	14
Exon coverage	0.20	15
DNA transposon coverage	0.13	16
Microsatellite coverage	0.12	17
Double strand break coverage	0.10	18
L2 coverage	0.07	19
A-phased repeat coverage	0.05	20
Self-chain segment coverage	0.04	21
Substitution rate	0.04	22
miRNA coverage	0.03	23
LTR retrotransposon coverage	0.01	24
Fragile site count	0.00	25

Using the same 25 genomic features to contrast CHSs and NHSs of SCNA breakpoints, we applied extremely tree classifiers to train the model and obtained a more powerful model compared with that in [143] (AUC: 0.96 versus 0.75). RELR and extremely tree classifiers both revealed distance to telomere and direct repeat coverage as being particularly potent in distinguishing CHSs and NHSs of SCNA breakpoints. The consistency of the results obtained by rare-event logistic models and extremely tree classifiers corroborates the robustness of our conclusions. It is noteworthy that indel rate is an important predictor in extremely tree classifiers, but not in rare event logistic models. The strong contrast between CHSs and NHSs for SCNA breakpoints in terms of the distance to telomere and direct repeat coverage indicates that CHSs strongly depend on the local genomic context. Given that only few known cancer genes are located in common breakpoint hotspot regions [138, 143], Li *et al.* hypothesized that the high frequency of SCNAs in these CHSs across cancer types is largely due to regionally higher mutation rate [143]. The regions with intrinsically higher mutation rate are independent of tumor type (or tissue origin) and are usually shared across different caner types. Since the regions enriched in direct repeats and/or those close to telomeres are susceptible to mutations, our models comply with this hypothesis.

Chapter 3

Genome-wide analysis of somatic copy number alterations and chromosomal breakages in osteosarcoma

This chapter has been published in Smida, J., Xu, H., Zhang, Y., Baumhoer, D., Ribi, S., Kovac, M., von Luettichau, I., Bielack, S., O'Leary, V., Leib-Mösch, C., Frishman, D., and Nathrath, M. (2017) Genome-wide analysis of somatic copy number alterations and chromosomal breakages in osteosarcoma. *Int. J. Cancer*, DOI:10.1002/ijc.30778. Jan Smida, Hongen Xu and I contributed equally to this work. This study was designed and initiated by Jan Smida, Christine Leib-Mösch, Dmitrij Frishman and Michaela Nathrath. Daniel Baumhoer, Irene von Luettichau, Stefan Bielack and Michaela Nathrath collected osteosarcoma samples and the corresponding clinical data. Jan Smida, Sebastian Ribi, and Michal Kovac performed SNP array genotyping. Dmitrij Frishman, Hongen Xu and I conceived the bioinformatics part of the project. Hongen Xu did somatic copy number alteration calling, driver gene identification, tumor subclone decomposition, tumor purity and ploidy estimation. I performed chromothripsis detection, chromosomal breakage characterization as well as its association with genomic features. The manuscript was written by Jan Smida, Hongen Xu and me, and edited by Valerie B. O'Leary, Dmitrij Frishman and Michaela Nathrath.

3.1 Abstract

Osteosarcoma (OS) is the most common primary malignant bone tumor in children and adolescents. It is characterized by highly complex karyotypes with structural and numerical chromosomal alterations. The observed OS-specific characteristics in localization and frequencies of chromosomal breakages strongly implicate a specific set of responsible driver genes or a specific mechanism of fragility induction. In this study, a comprehensive assessment of somatic copy number alterations (SCNAs) was performed in 160 OS samples using whole-genome CytoScan High Density arrays (Affymetrix, Santa Clara, CA). Genes or regions frequently targeted by SCNAs were identified. Breakage analysis revealed OS specific fragile regions in which well-known OS tumor suppressor genes, including *TP53*, *RB1*, *WWOX*, *DLG2*, and *LSAMP* are located. Certain genomic features, such as transposable elements and non-B DNA-forming motifs were found to be significantly enriched in the vicinity of chromosomal breakage sites. A complex breakage pattern - chromothripsis - has been suggested as a widespread phenomenon in OS. It was further demonstrated that hyperploidy and in particular chromothripsis were strongly correlated with OS patient clinical outcome. The revealed OS-specific fragility pattern may provide a basis for patient prognosis and offer a vital platform for therapeutic intervention in the future.

3.2 Introduction

Osteosarcoma (OS) is the most common primary malignant bone tumor in adolescents and young adults [184, 185]. It is characterized by a complex karyotype with a high degree of aneuploidy and numerous structural aberrations such as somatic copy number alterations (SCNAs) and genomic rearrangements [186, 187, 188]. Curative treatment of OS is based on multi-agent chemotherapy in addition to complete surgery. For patients with localized extremity disease 10-year event-free survival rates reach approximately 60% [189], but have plateaued during the past decades. Further improvement in cure rates will most likely depend on an increased knowledge about the underlying molecular mechanisms of this disease.

Although several predictors, such as gene expression profiles [190] and chromosomal alteration staging systems [188] have been proposed to anticipate tumor response to chemotherapy, common markers of prognostic and therapeutic value remain to be identified. Genomic instability is a hallmark of most cancers, including OS [191, 178]. Recurrent genomic instability in cancer is either driven by positive selection or originates from sequence-specific unstable regions [178]. Chromosomal fragile sites are specific genomic locations that appear as gaps or breaks on metaphase chromosomes under replication stress [192]. Replication stress can be induced by endogenous or exogenous sources, and result in the generation of DNA double strand breaks (DSBs) and genomic instability [193]. A variety of molecular pathways are involved in DSB repair, and, in the case of deficient repair, copy number alterations result.

To identify SCNAs, an array-based copy number profiling has been utilized as an alternative to next generation sequencing due to its lower consumption of precious biopsy material. DNA copy number profiling was generally opted for over gene expression, as it provided relatively stable profiles enabling differentiation of clinically relevant genetic subgroups [194]. However, the analysis of whole genome array data for tumor samples can be challenging due to the fact that the total DNA amount in a cancer cell can differ significantly from a diploid state, and tumor tissues often contain some proportion of normal cells [44]. SCNAs have the potential to inactivate tumor suppressor genes or activate oncogenes, and consequently play fundamental roles in gene regulation and pathobiological processes in cancer [138]. Analyses of SCNA data generated in recent years have provided insights into driver genes for many tumor types [138, 55]. However, the enormous complexity of genomic aberrations in OS has made it challenging to identify recurrent alterations and genes driving tumorigenesis [186, 187]. Furthermore, in OS the identification of driver genes has been hindered by intra- and inter-tumor heterogeneity and limited sample availability [195, 187, 196, 197]. Despite such difficulties, we and others have revealed recurrent genomic loss regions containing tumor suppressor genes such as *LSAMP*, *CDK2NA*, *RB1*, and *TP53* and most frequent gains including the oncogene *MYC* and the gene *RUNX2* - an important player in osteogenic differentiation [195, 198, 187, 196, 197].

Apart from their genomic instability, osteosarcomas show a disease specific SCNA pattern. The phenomenon of chromothripsis represents an important mechanism of carcinogenesis that differs from progressive accumulation of genomic rearrangements. The simultaneous fragmentation of distinct chromosomal regions (breakpoints showing a specific, non-random distribution) and subsequent imperfect reassembly of those fragments leads to a specific SCNA pattern (chromothripsis like pattern, CTLP). The initial discovery indicated that chromothripsis is a widespread phenomenon, which can be seen in 2% - 3% of all cancers, most notably in 25% of bone cancers [46]. There is a strong evidence for an association between chromothripsis and poor outcome in different cancer types, including multiple myeloma [199], neuroblastoma [200] and Sonic-Hedgehog medulloblastoma [53]. Although the mechanisms governing chromothripsis are largely unknown, it has important implications for our understanding of cancer and disease [201], as such detailed analyses of chromothripsis-like patterns may shed light on OS development and progression.

Herein, copy number profiles derived from 160 pre-therapeutic osteosarcoma biopsies have been analysed using whole-genome CytoScan High Density (CytoScan HD) arrays (Affymetrix, Santa Clara, CA). SCNAs for each sample were integrated to identify potential genes that may drive OS oncogenesis. Previously found OS driver genes were identified as well as other OS-related genes. Chromosomal breakages were found to be spatially clustered in certain locations, termed “broken regions”, harboring the regarded OS tumor suppressor genes *TP53*, *RB1*, *WWOX*, *DLG2*, and *LSAMP*. Furthermore, chromosomal breakages in these regions occurred early and were determined by local genomic context. Most noteworthy, both aneuploidy and CTLP occurrence were found to be correlated with clinical outcome of OS patients.

3.3 Methods

3.3.1 Tissue samples and patient characteristics

For CytoScan HD array analysis, a set of 160 fresh-frozen tissue samples derived from pretherapeutic biopsies was used. The patient cohort samples were obtained according to the guidelines and approval of the Research Ethics Board at the Faculty of Medicine of the Technical University of Munich (Technische Universität München, Reference 1867/07) and local ethical committee of Basel, Switzerland (Ethikkommission beider Basel EKBB, www.ekbb.ch, Reference 274/12). The descriptive characteristics of this collective are summarized in Table 3.1. The vast majority of the investigated samples (n=141) are classified as high-grade osteosarcoma. The patients were treated between 1990 and 2012 according to the protocols of the Cooperative German-Austria-Swiss Osteosarcoma Study Group (reviewed and approved by the appropriate ethics committees) after informed consent was obtained.

Table 3.1: Clinical characteristics of 157 osteosarcoma patients

Descriptive statistics		
Sex	**n=157**	
Male	83	
Female	74	
Age at diagnosis(years)	**n=157**	
Average	20.08	
Median	15	
Range	3-85	
Metastases	**n=143**	
Yes	61	
No	82	
Observation period (months)	**n=147**	
Average	64	
Median	56.2	
Range	0.24-204.5	
Response to neoadjuvant treatment	**n=128**	
Good	64	
Poor	64	
Survival	**n=130**	
Alive	90	
Deceased	40	
Event (relapse or death)	**n=143**	
Yes	60	
No	83	
Overall survival	5-year: 74.8%	10-year: 62.9%
Grouped by event status	**5-year**	**10-year**
Event	25.5%	27.3%
Grouped by response to chemotherapy	**5-year**	**10-year**
Good response	90.2%	83.6%
Poor response	66.7%	61.1%

3.3.2 SCNA calling, driver gene identification, and tumor subclone decomposition

DNA from frozen osteosarcoma tissue was analysed using the Affymetrix CytoScan HD platform. The raw data are available in the ArrayExpress database [202] under accession number E-MTAB-4815. Nexus copy number software version 7.5 (obtained from BioDiscovery, Inc.) was used to process CEL files. Copy number alterations were called using the Single Nucleotide Polymorphism Fast Adaptive States Segmentation Technique 2 (SNP-FASST2) segmentation algorithm together with quadratic correction implemented in Nexus. Sample- and chromosome-specific thresholds defining copy number gain, copy number loss, high copy gain, and homozygous copy loss were based on true diploid regions in individual tumor sample (performed using Nexus with subsequent manual curation by experts from BioDiscovery, Inc.). SCNAs with fewer than 20 probes were excluded from further consideration. GISTIC 2.0 (Genomic Identification of Significant Targets In Cancer) integrated in the Nexus copy number software was utilised to identify potential driver SCNAs and genes by evaluating the frequency and amplitude of observed events [203].

Subclone structures were reconstructed for each tumor sample based on the SCNA calling data from the Nexus copy number software. The SubcloneSeeker software [204] was utilized to decompose tumor subclone structures. In this study, a subclone was defined as a collection of cells in the tumor sample that contained the same set of SCNAs. The segmental mean values of each segment generated by SNP-FASST2 was used as input for the SubcloneSeeker software [204] to reconstruct the clonal structures for each patient. The *segtxt2db* and *ssmain* applications were employed to cluster the segments based on their cell prevalence values and to enumerate the clonal structures. The results were exported using the *treeprint* utility. We refer to the SCNAs that occurred at the root node of the subclone tree as clonal SCNAs and to all others as subclonal ones.

3.3.3 Definitions of chromosomal breakages and their association with genomic features

We defined genomic starts and ends of SCNAs as SCNA breakpoints although their exact chromosomal positions could not be determined. Breakpoints situated upstream of the first or downstream of the last CytoScan HD probe on the same chromosome as well as those located in telomeres or centromeres were ignored. We defined a genomic position to be a chromosomal break when the $\log_2$ signal value alteration between two adjacent genomic segments (from centromere to telomere) was >0.3.

An association was determined between chromosomal breakages and multiple genomic features as obtained from public databases and published studies or as identified in the current study. All genomic coordinates of the features correspond to the human genome assembly hg19 and, when necessary, the University of California, Santa Cruz (UCSC) *liftOver* tool was used to convert the hg18 coordinates to hg19 [108]. Specifically, chromosomal coordinates for Alu repeats, DNA transposons, L1 and long terminal repeat (LTR) retrotransposons, exons, and conserved elements (the PhyloP46wayPrimates table) were downloaded from UCSC Genome Browser [108]. Non-B DNA motifs were obtained from non-B DB v2.0 [146]. Genomic coordinates for common fragile sites and non-fragile regions were obtained from a previous study [155]. We defined nucleotide substitution (or insertions/deletions, indels) rate as the ratio of the total number of substitutions (or indels) to the total number of nucleotides in the human-chimpanzee alignments (from UCSC Genome Browser).

The density of SCNA breakpoints, chromosomal breaks or genomic features were defined as the ratio of total base pairs belonging to the item to the total length of the genomic region. The subdivision of the genome, shuffling, and feature density calculation were performed using BEDTools [162] and in-house Perl scripts.

3.3.4 Detection of chromothripsis-like patterns in osteosarcoma

To detect chromothripsis-like patterns (CTLPs) the algorithm described in [47] was applied to identify clustering of copy number changes in the genome. Default settings were used except for the parameter of $\log_2$ signal value difference between two adjacent segments (set to 0.2). CTLP samples were determined by the evidence of the copy number switching its status at least 12 times ($SwitchNo \geq 12$) and $\log_{10}$ of likelihood ratio greater than 8 ($\log_{10} LR \geq 8$) within a single chromosome.

3.3.5 Estimation of tumor purity and ploidy

SNP-based DNA microarrays allow simultaneous measurement of the allele-specific copy number at many different SNP loci in the genome. For each probe set, the log R ratio (LRR) reflects the ratio of total intensity signals for both alleles to expected signal, and the B allele frequency (BAF) is an estimate of the relative proportion of one of the alleles with respect to the total intensity signal. LRR and BAF values were derived using the affy2sv R package [205] together with the Affymetrix Power Tools. A total of 873 normal samples downloaded from the study [206] (Gene Expression Omnibus accession number: GSE59150) were also processed using *affy2sv*. The resulting LRR and BAF were used as input for the GPHMM algorithm (version 1.4) [39] to obtain an estimation of normal cell contamination and absolute copy number of genomic segments for each sample. Population frequency of the B allele file required for running GPHMM was created using the Perl script *compile_pfb.pl* in PennCNV [207], with BAF values from the 873 normal samples as input. Another required file - GC model file (GC content flanking SNP markers) - was generated using the Perl script *cal_gc_snp.pl* in PennCNV [207]. Tumor ploidy was further determined following the protocol described in [208]. Specifically, the chromosome arm count in a tumor genome was estimated based on the absolute copy number of genomic segments in the pericentric region. The copy number of the corresponding arm was set to the absolute copy number of the segments in the pericentric region if its size was $\geq$1.5 Mb. Otherwise, if the size of the pericentric segments was <1.5 Mb, the copy number of the chromosome arm was approximated by the average copy number of all segments on that chromosome arm. Tumor ploidy was assigned for each tumor sample based on chromosome counts and the DNA index, defined as the average copy number of the tumor genome divided by 2. Tumor ploidy was set at 2 (near-diploid genome) for chromosome counts <60 and DNA index <1.3, and set at 4 (near-tetraploid genome) for chromosome counts $\geq$ 60 and DNA index $\geq$1.3 [209].

3.4 Results

3.4.1 Overview of SCNAs in osteosarcoma

The SCNA landscape of pre-treatment tissue samples (n = 160) from osteosarcoma patients (characteristics of whom are provided in Table 3.1) was profiled using Affymetrix CytoScan HD arrays. Three samples were excluded from copy number analysis due to insufficient data quality. A genome-wide frequency plot of SCNAs is shown in Figure 3.1. In our collective the median size of the SCNAs was 1.2 Mb with the OS genome having on average 209 SCNA events. Regional gains and losses of various sizes were observed, ranging from entire chromosomes to minor genomic segments. Many oncogenes and tumor suppressor genes were located within these sites. No significant correlation was noted between the total SCNA number, size, or median in relation to age or gender. An apparent correlation trend was evident for total SCNA size and survival, although perhaps due to insufficient power this did not reach significance.

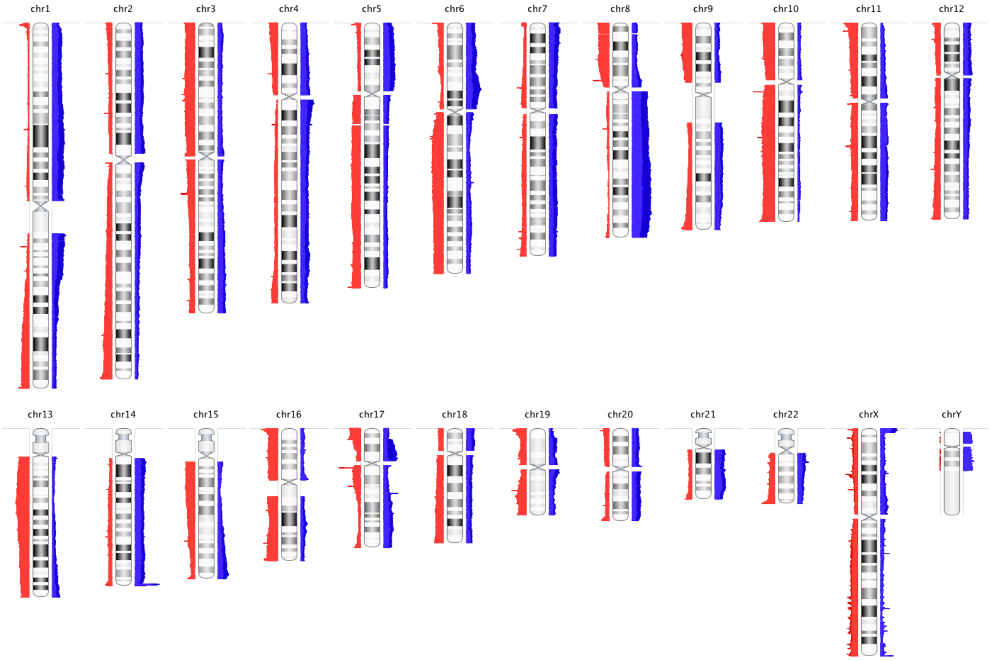


Figure 3.1: Genome-wide frequency plot of somatic copy number alterations in 157 osteosarcoma samples. Copy number losses and gains are in red and blue, respectively.

3.4.2 GISTIC analysis and tumor subclone decomposition uncover key driver genes affected by SCNAs in osteosarcoma

GISTIC 2.0 [203] is a tool to identify genes targeted by SCNAs that may drive cancer development. The X and Y chromosomes were excluded from the analysis and were analyzed separately in gender specific subsets of OS patients. GISTIC identified 88 regions significantly altered in 157 OS samples (Figure 3.2; genomic locations of these regions have been listed in Supplementary Table 6.1). The annotation of GISTIC regions revealed 101 targeted genes (listed in Supplementary Table 6.2), of which the vast majority (74 transcripts) were protein-coding genes. Nine genes listed in the Catalogue of Somatic Mutations in Cancer (COSMIC) Cancer Gene Census (CGC) [157] - namely *NOTCH2*, *PDGFRA*, *CDK4*, *CCNE1*, and *RUNX1* were located in copy-number gain regions, while *CDKN2A*, *FLI1*, *TP53*, and *ATRX* were identified in copy-number loss regions. *TP53* and *ATRX*, often targeted by SCNAs, have been reported by us and others as important driver genes in OS [195, 210, 211]. Besides these well-known OS driver genes, GISTIC regions contained several other OS-related genes, such as *RUNX2* and *DLG2* [195, 212].

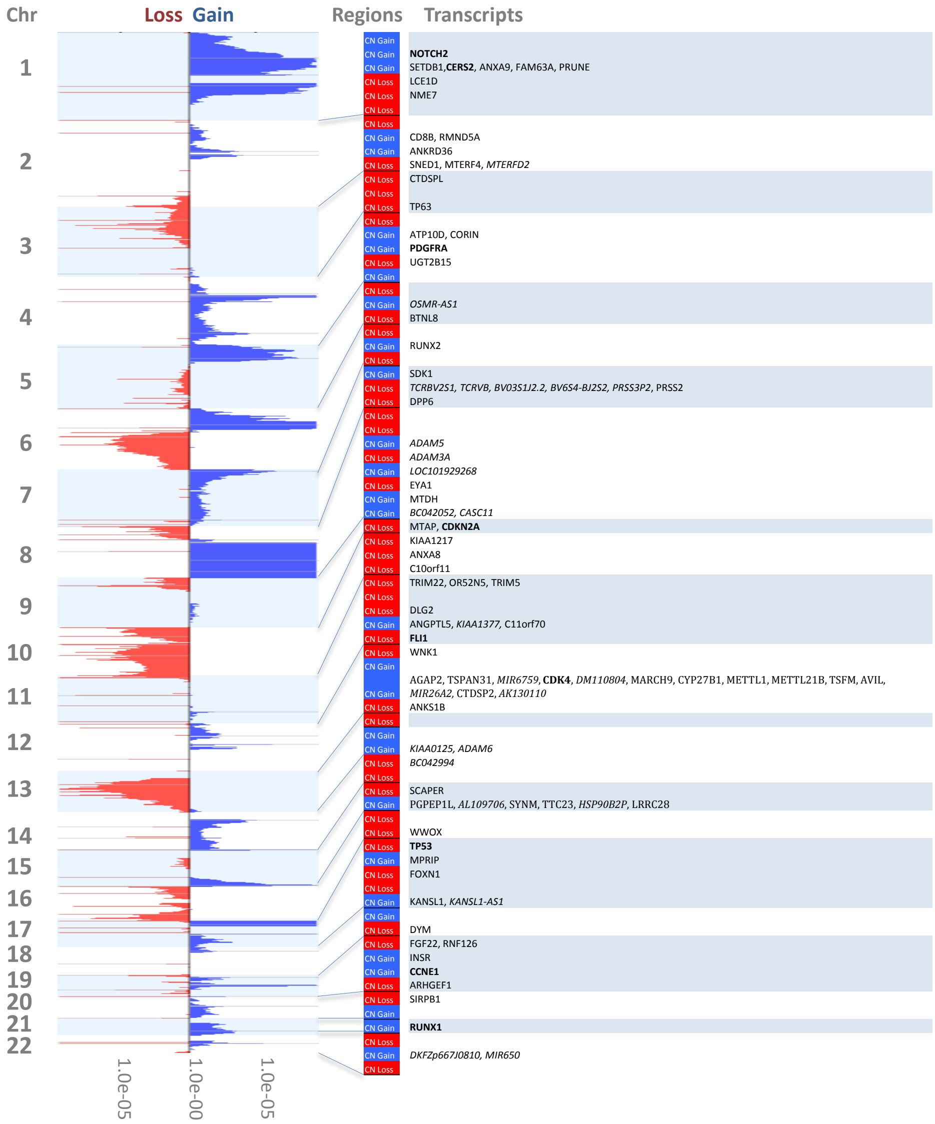


Figure 3.2: Significantly altered regions and genes contained therein with copy number alterations in osteosarcoma as identified by GISTIC analysis.

Analysis also revealed novel or recently described genes - *FOXN1* and *WWOX*. *FOXN1* (17q11.2) is the main transcriptional regulator of the development, differentiation, and function of thymic epithelial cells [213]. Although it directly or indirectly regulates expression of a broad variety of genes, it has not been found to date to be associated with cancer and, in particular OS. The *WWOX* gene (16q23.1) spans a common fragile site FRA16D, associated with DNA instability in cancer [214]. Recently, a series of reports demonstrated the relevance of reduced or absent *WWOX* expression in various cancer types, including OS, presumably due to chromosomal deletions and translocations within the *WWOX* gene highlighting an essential role for *WWOX* in tumor suppression and genomic stability [215, 216, 217]. Besides the tumor suppressor and pro-apoptotic activity of *WWOX* in OS, its role in osteogenic differentiation and interaction with *RUNX2* has recently been elucidated [218].

A malignant tumor often consists of genetically distinct cell populations, referred to as tumor subclones, with each possessing a specific mutation subset. Determination of the order in which SCNA mutations occurred is a powerful means for identifying genes with fundamental roles in oncogenesis. SubcloneSeeker [204] succeeded in inferring subclone structures for 99.4% of tumors (156 out of 157). The mean number of predicted subclone structures for each tumor was 8.5 (ranging from 1 to 45). Thirty-six tumors had greater than 10 possible subclone structures, which may be due to the complex nature of such tumor samples. Next, an investigation was undertaken as to whether or not SCNAs overlapping with putative genes (identified by GISTIC) were clonal events. Previously reported findings as revealed by alternative approaches were confirmed, to show that even for the well-known OS driver genes such as *TP53* and *RB1*, the majority (∼90%) of SCNAs were subclonal events [210]. Thirty-four tumors had clonal SCNAs overlapping one to ten driver genes, such as *TP53*, *RB1*, *DLG2*, *WWOX*, *TERT*, *FOXN1*, *APC*, *PTEN*, *LSAMP*, *ATRX*, and *CDKN2A*. No single gene had clonal SCNAs in the majority of tumors.

3.4.3 Breakage analyses reveal osteosarcoma-specific fragile regions

DNA breakage is a prerequisite for cancer-associated genomic aberrations, including amplifications, deletions, inversions, and translocations. The genomic start and end of SCNAs were defined as breakpoints with a precision of ∼1 kb (average inter-probe distance for CytoScan HD Array is <1 kb). Since whole genome arrays have reduced ability for inversion and/or translocation detection, the chromosomal breakage landscape was investigated, which strongly indicated the prevalence of genomic rearrangements. The criterion for considering a SCNA breakpoint as a chromosomal break was based on thelog$_2$ signal value alteration between two adjacent genomic segments >0.3 (Figure 3.3), which is more stringent than the cutoff of 0.23 used in [219]. In total, 62,172 SCNA breakpoints and 19,810 chromosomal breaks were identified in 157 OS samples. The number of chromosomal breaks per sample ranged from 17 to 425, with a median value of 114. The number of breaks per megabyte ranged from 4 (chromosome 2) to 14 (chromosome 17). In order to further examine the landscape of chromosomal breaks across different chromosomes, each chromosome was divided into non-overlapping 1 Mb regions following gap exclusion in the genome assembly and calculated the density of chromosomal breaks per block. Results showed that 2% of genomic regions (61/3060) were significantly enriched for chromosomal breaks (Bonferroni corrected P-values <0.1). Out of these "broken regions", 13% are located within common fragile sites, while 49% overlapped with non-fragile sites [155], indicating apparent OS-specific fragility characteristics.

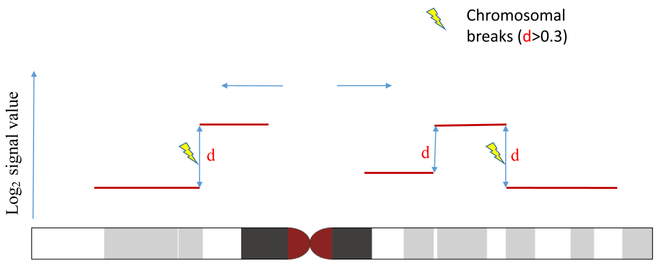


Figure 3.3: Schematic illustration of chromosomal breaks. "d"means log$_2$ value changes between two adjacent genomic segments at a specific genomic position.

Some of the OS-associated tumor suppressor genes [198], including *TP53*, *RB1*, *WWOX*, *DLG2*, and *LSAMP*, but no known OS oncogenes, were located in these broken regions (Figure 3.4). To determine the evolutionary order in which SCNAs occurred in these areas, a comparison was made with clonal SCNAs obtained by the SubcloneSeeker analysis. An enrichment of clonal SCNAs was found in these broken regions compared to randomly generated ones (10662 vs 4579, P-value=0), implicating chromosomal breakage is a clonal event of early occurrence in tumorigenesis.

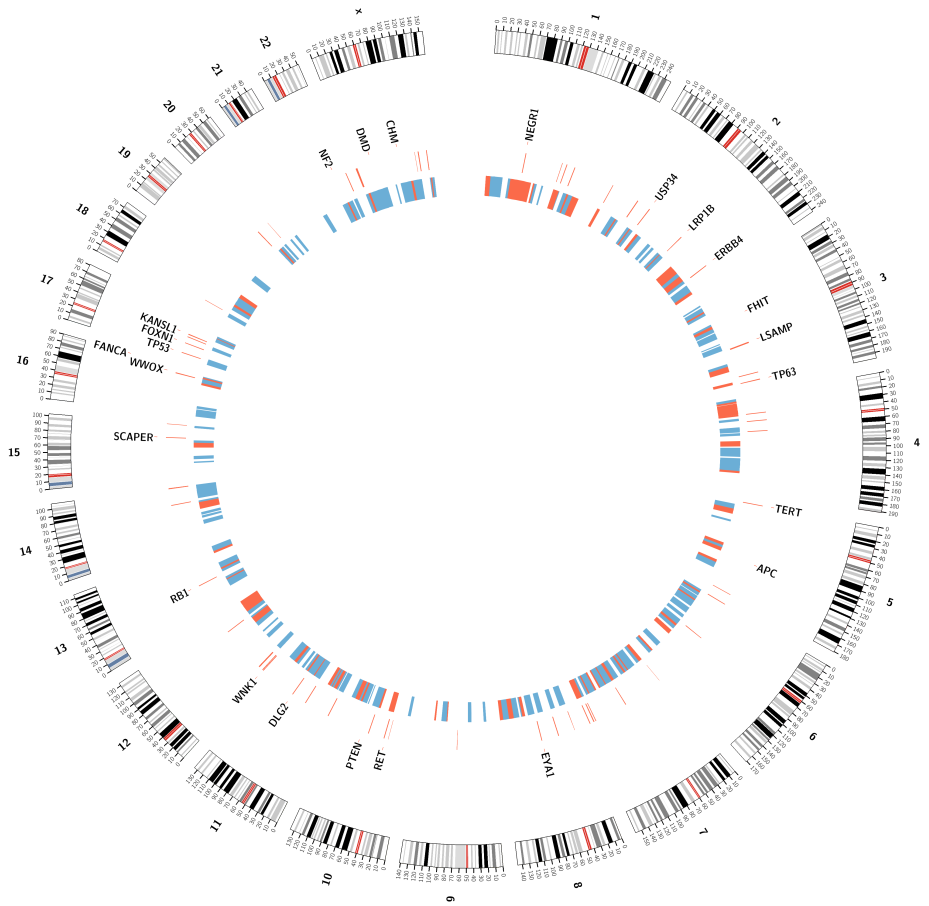


Figure 3.4: The genomic landscape of chromosomal breaks and associated genes in osteosarcoma. The outermost circle represents chromosomes and cytogenetic bands. The next circle represents known OS driver genes and other genes as listed in Table 3.2. The third circle represents "broken regions ". The innermost circle shows common fragile sites and non-fragile regions in red and blue respectively.

In order to identify genes prone to breakage in OS, we compared the distribution of actual chromosomal breaks to a background distribution obtained by shuffling the position of chromosomal breaks 1,000 times. This approach, while admittedly suffering from some uncertainty in calling the location of chromosomal breaks due to the inter-probe distance characteristic for CytoScan HD arrays, can provide clues as to which genes are prone to breakage in OS. A total of 343 genes were found to harbor chromosomal breaks significantly more frequently than would be expected by chance (Bonferroni corrected P-values < 0.01). Of these, 24 genes (listed in Table 3.2) have been previously shown to be associated with OS (*DLG2*, *WWOX*, *TP53*, *RB1*, *LSAMP*, *PTEN*, and *APC* [198]) and other tumors (*DMD*, *EYA1*, *SCAPER*, *WNK1*, *KANSL1*, *TP63*, *FOXN1*, and *CHM*) and found by GISTIC analysis. *TP53* was selected to demonstrate the distribution of chromosomal breaks along the gene. As seen in Figure 3.5 the largest number of chromosomal breaks was located in the first intron of this gene [195, 211].

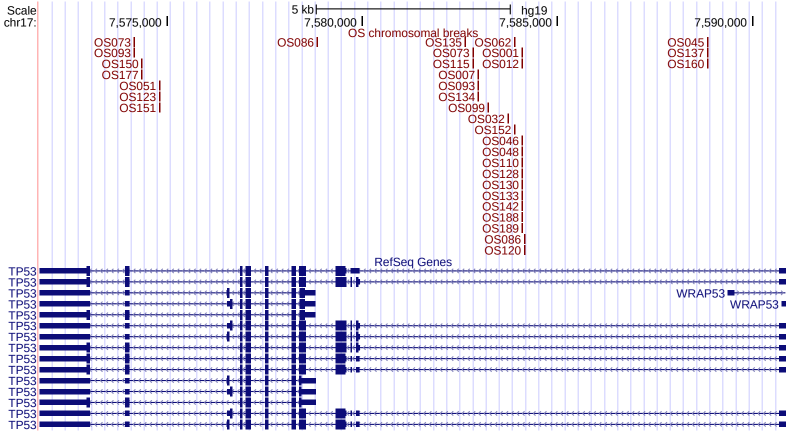


Figure 3.5: Plot of chromosomal breaks around the *TP53* gene.

3.4.4 Chromosomal breakage in osteosarcoma is dependent on local genomic context

To examine whether chromosomal breakages in OS were associated with the local genomic context, we investigated the joint distributions of chromosomal breaks, SCNA breakpoints and multiple genomic features within a 1Mb genomic window. Previous studies have shown that DNA breakage can be induced by DNA structures such as non-B DNA conformations, including Cruciform, G-quadruplexes (G4), Slip, Triplex, and Z-DNA, and by highly homologous genomic repeats, such as L1 and Alu [141, 144, 153]. Further features considered in this analysis were common fragile sites, evolutionarily conserved elements, substitution rate, indel rate and exon density which have been associated with SCNA breakpoints [141, 143, 220]. As expected, SCNA breakpoints and chromosomal breakage are highly correlated (P-value $< 2.20 \times 10^{-16}$, Spearman rho $= 0.76$). In addition, it was also

Table 3.2: Genes frequently targeted by chromosomal breaks in OS that were previously shown to associate with OS or other tumors

Gene	Chromosome	Start	End	OMIM	Count	% OS
DLG2	11	83 166 055	85 338 314	603583	113	27.39
WWOX	16	78 133 309	79 246 564	605131	102	31.85
DMD	X	31 137 344	33 357 726	300377	71	17.83
EYA1	8	72 109 667	72 274 467	601653	62	20.38
SCAPER	15	76 640 526	77 176 217	611611	61	19.75
ERBB4	2	212 240 441	213 403 352	600543	43	12.74
FHIT	3	59 735 035	61 237 133	601153	42	8.28
WNK1	12	862 088	1 020 618	605232	40	14.01
KANSL1	17	44 107 281	44 302 740	612452	40	21.66
LRP1B	2	140 988 995	142 889 270	608766	39	12.74
TP53	17	7 571 719	7 590 868	191170	34	19.75
TP63	3	189 349 215	189 615 068	603273	34	10.83
USP34	2	61 414 589	61 697 849	615295	29	11.46
TERT	5	1 253 286	1 295 162	187270	28	10.19
FOXN1	17	26 850 958	26 865 175	600838	25	15.92
NF2	22	29 999 544	30 094 589	607379	25	6.37
RB1	13	48 877 882	49 056 026	614041	24	8.28
NEGR1	1	71 868 624	72 748 277	613173	21	7.01
CHM	X	85 116 184	85 302 566	300390	21	7.01
LSAMP	3	115 521 209	116 164 385	603241	19	8.92
PTEN	10	89 623 194	89 728 532	601728	11	3.82
APC	5	112 043 201	112 181 936	611731	10	3.18
RET	10	43 572 516	43 625 797	164761	8	4.46
FANCA	16	89 803 958	89 883 065	607139	6	2.55

All genomic coordinates are based on human genome assembly hg19;
Count: the total number of chromosomal breaks found in gene regions;
% OS: percent of OS samples affected by chromosomal breaks;
gene names previously associated with OS are in bold;
gene names identified by GISTIC analysis in this study are in italics.

noted that SCNA breakpoints and chromosomal breaks were significantly correlated with diverse genomic properties, including Alu, L1, Cruciform, G4, Slip, Triplex, Z-DNA, conserved elements, exon density, and indel rate (Bonferroni corrected P-values <0.01; Table 3.3).

We further examined the association of genomic properties to chromosomal breaks at a higher resolution. Specifically, windows of 10 kb, 20 kb, 50 kb, and 100 kb centred around each chromosomal break were analysed with subsequently overlapped windows merged. For each window, the density of each feature was computed and determined as to whether the feature was enriched compared to the remaining regions. Compared with random expectation, the vicinity of chromosomal breaks was significantly enriched for several genomic features, including genomic repeats, non-B DNA conformation forming motifs, conserved elements, exon density, substitution rate and indel rate (Table 3.4; Bonferroni corrected P-values < 0.01, Mann-Whitney test). These genomic features have been associated with SCNA breakpoints in different cancer types [143], suggesting that OS

Table 3.3: Correlations among SCNA breakpoints, chromosomal breaks and genomic features

Chromosomal Breakage	Genomic Features	P-values	Spearman Rho
Chromosomal breaks	**Alu**	6.01×10^{-29}	0.20
	DNA transposons	1.11×10^{-2}	0.05
	L1	1.36×10^{-12}	0.13
	LTR retrotransposons	3.31×10^{-6}	0.08
	Cruciform	1.67×10^{-17}	0.15
	G4	7.75×10^{-21}	0.17
	Slip	3.00×10^{-38}	0.23
	Triplex	4.47×10^{-13}	0.13
	Z-DNA	1.63×10^{-31}	0.21
	Conserved elements	2.92×10^{-5}	0.08
	Exon density	1.67×10^{-15}	0.14
	Common fragile sites	1.75×10^{-2}	-0.04
	Substitution rate	1.69×10^{-14}	0.14
	Indel rate	6.88×10^{-20}	0.16
SCNA breakpoints	**Alu**	1.50×10^{-52}	0.27
	DNA transposons	1.85×10^{-5}	0.08
	L1	4.52×10^{-25}	0.19
	LTR retrotransposons	5.63×10^{-3}	0.05
	Cruciform	1.16×10^{-11}	0.12
	G4	2.69×10^{-49}	0.26
	Slip	8.66×10^{-48}	0.26
	Triplex	3.48×10^{-21}	0.17
	Z-DNA	8.73×10^{-27}	0.19
	Conserved elements	5.36×10^{-1}	0.01
	Exon density	2.27×10^{-42}	0.24
	Common fragile sites	1.25×10^{-2}	-0.05
	Substitution rate	5.26×10^{-2}	0.01
	Indel rate	5.00×10^{-8}	0.10

Genomic features with Bonferroni corrected P-values less than 0.01 are in bold.

is similar to other cancers in regards to chromosomal breakage occurrence. Of note, common fragile sites were not preferentially associated with chromosomal breaks at any genomic resolution investigated in this study (Table 3.4), indicating that OS has perhaps very specific breakage characteristics that include already known common fragile sites as well as unique sites of instability.

3.4.5 Clinical implications of chromothripsis-like patterns and hyperploidy

Applying the CTLP detecting algorithm to the OS SCNA dataset a total of 87 chromosomes from 52 patients passed the threshold and were termed CTLP cases. CTLP occurred in 33.1% of patients in this dataset, implying that chromothripsis is a widespread phenomenon in OS. This incidence rate was largely consistent with a previous study of a small sample size of bone cancers [46]. CTLPs had a tendency to occur frequently on chromosomes 8 (11.5%) and 17 (9.2%). The OncoPrint shown in Figure 3.6 provides an overview of SCNAs in specific genes and CTLP affecting

Table 3.4: Correlation between chromosomal breaks and genomic features

Genomic features	**Enrichment in genomic regions centered at chromosomal breaks**			
	10 kb	20 kb	50 kb	100 kb
Alu	+	+	+	+
DNA transposons	+	+	+	+
L1	+	+	+	+
LTR retrotransposons	+	+	+	+
Cruciform		+	+	+
G4	+	+	+	+
Slip	+	+	+	+
Triplex			+	+
Z-DNA		+	+	+
Conserved elements	+	+	+	
Exon density	+	+		
Common fragile sites				
Substitution rate	+	+	+	+
Indel rate	+	+	+	+

+ denotes enrichment of genomic features in genomic windows centered at chromosomal breaks (Bonferroni corrected P-values <0.01).

individual samples. Chromosomal aberrations in *TP53* occured in 88% (46/52) of CTLP patients, compared to 56% (59/105) of non-CTLP cases (P-value = 1.0×10^{-4}, two-tailed Fisher‘s exact test). We analysed three genes - *RB1*, *WWOX* and *DLG2* - that frequently harbor structural variation in OS [195]. Chromosomal alterations in *RB1* occur in 73% (38/52) of CTLP cases, but only in 48% (50/105) of non-CTLP samples (P-value = 3.5×10^{-3}, two-tailed fisher‘s exact test). Chromosomal aberrations in *WWOX* occur in 85% (44/52) and 66% (69/105) CTLP and non-CTLP samples, respectively (P-value= 1.4×10^{-2}, two-tailed fisher‘s exact test). Finally, 83% (43/52) of CTLP cases harboured aberrations in *DLG2*, compared to 57% (60/105) of non-CTLP cases (P-value = 1.3×10^{-3}, two-tailed fisher‘s exact test). These observations indicate that chromosomal aberrations in *TP53*, *RB1*, *WWOX* and *DLG2* genes are strongly associated with chromothripsis-like patterns in OS.

Furthermore, an investigation of the association between chromothripsis-like patterns and clinical data was performed [221]. As follow-up clinical data was available for 114 patients, CTLP was detected in 33% (38/114) of this cohort. Notably, as shown in Figure 3.7, Kaplan-Meier analysis revealed that patients with CTLP patterns in their tumors showed significantly curtained survival expectancies compared to those without CTLP (log-rank test, P-value = 7.1×10^{-4}).

A successful estimation was made of tumor ploidy and tumor content for 90.4% (142 /157) of samples using the GPHMM algorithm. These osteosarcoma biopsies were estimated to have on average 37.5% normal tissue

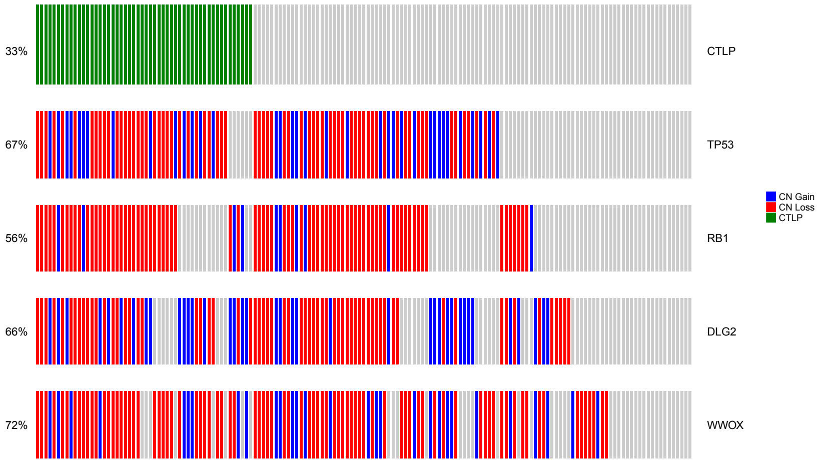


Figure 3.6: OncoPrint showing the distribution of SCNAs (CN gain and CN loss) for genes *TP53*, *RB1*, *DLG2* and *WWOX* and chromothripsis-like pattern (CTLP) in osteosarcoma patients (column). Each bar represents a sample. Green bars indicate samples with CTLP. Red and blue bars indicate samples with CN loss and CN gain for a specific gene, respectively. Gray bars represent samples without CTLP or without CN changes for a specific gene. The numbers on the left show what percentage of samples is affected by CTLP or CN changes for a specific gene.

contamination with a median ploidy of 2.7n. Following the procedures for chromosome number estimation (as described in the Methods), the distribution of chromosome numbers was plotted in 142 samples to clearly demonstrate a two ploidy status of the tumor genome (Figure 3.8a). Near-tetraploid tumors had greater chromothripsis events than diploid ones (Figure 3.8b, P-value = 0.0046, Fisher‘s exact test). This was compatible with results from a recent study linking chromothripsis with hyperploidy [222]. Patients with tumors exhibiting near-tetraploid genomes had poorer survival compared to patients having tumors with estimated ploidy of around 2 (Figure 3.8c).

3.5 Discussion

Rarity and genomic complexity, as well as marked intra- and intertumoral heterogeneity, have challenged the molecular characterization of osteosarcoma etiology [198]. Given the difficulty in acquiring a large cohort of samples in this rare tumor, we integrated DNA copy number profiles of 160 pretherapeutic biopsies to identify recurrent genomic changes and driver genes. Genome-wide profiles were performed on Affymetrix CytoScan HD platform, which has the highest resolution of SNP and non-polymorphic

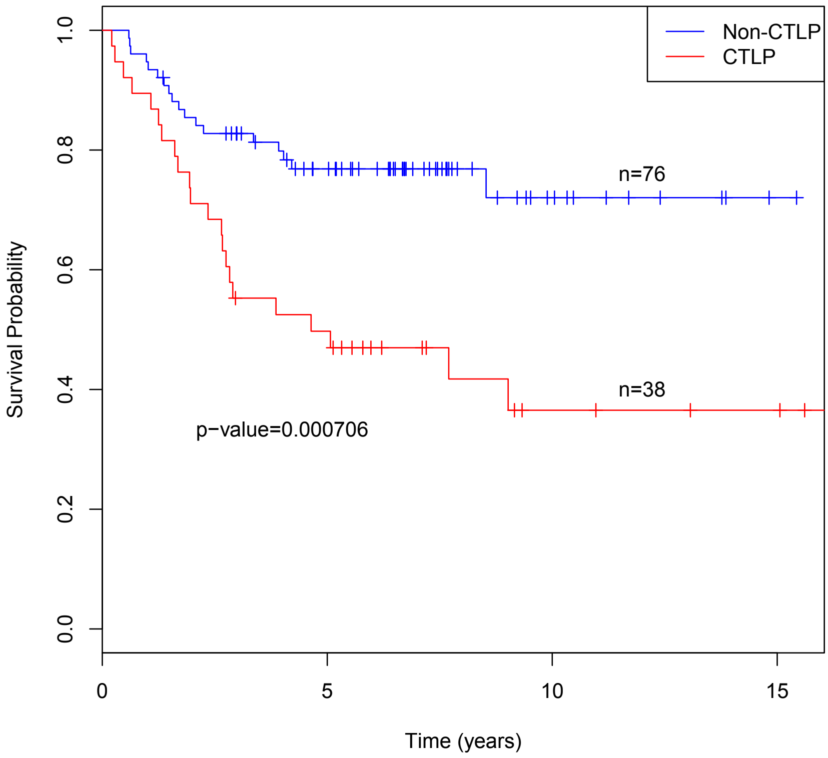
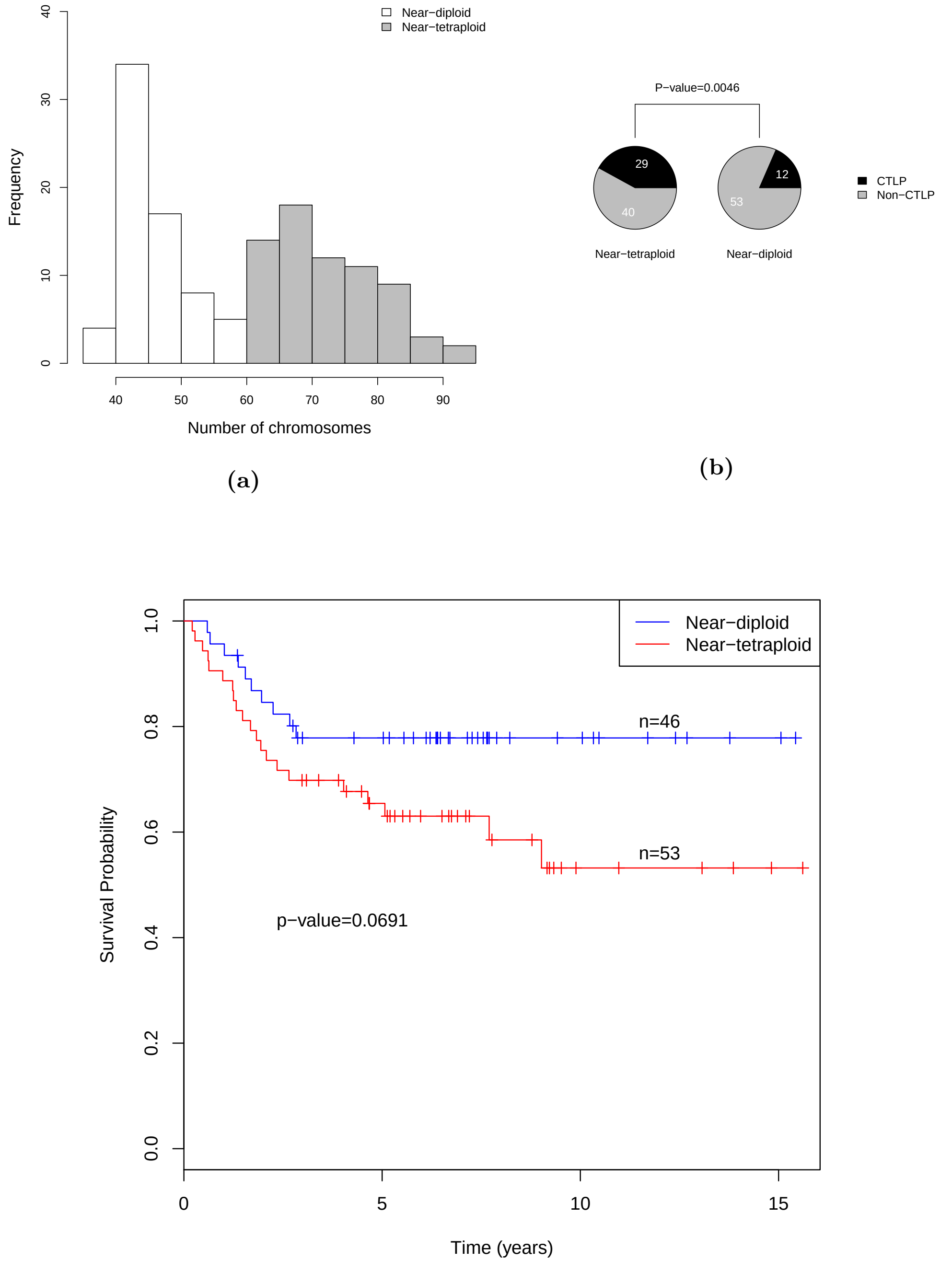


Figure 3.7: Kaplan-Meier survival curves for chromothripsis-like patterns (CTLPs) versus non-CTLP cases. The P-value is based on the log-rank test.



(a)

(b)

(c)

Figure 3.8: Ploidy estimation and its clinical implications. (a) Distribution of chromosome numbers in 142 osteosarcoma samples, displaying the 2 ploidy status of tumor genomes. (b) Association of the ploidy status with chromothripsis. (c) Kaplan-Meier survival curves for near-tetraploid samples versus near-diploid samples. The P-value is based on the log-rank test.

probes for detecting human chromosomal alterations. Copy number analyses confirmed high genomic instability in the OS biopsies, with the vast majority of samples (82%) exhibiting highly complex altered genomes. The unstable genome in the majority of OS is probably due to the deficiency in homologous recombination repair [210]. The *BRCA1/2*(important players in homologous recombination pathway) deficiency associated characteristics in single base substitutions, and large-scale genome instability signatures are evident in more than 80% of OS [210].

Using GISTIC, we identified a number of genes which are frequently targeted in OS, including already known driver genes (*e.g.* *TP53* and *ATRX*) as well as other OS-related genes, such as *WWOX*. *WWOX* is a putative tumor suppressor gene encompassing a common fragile site FRA16D, which is a frequent target of chromosomal rearrangement in multiple cancers. The absence or reduced expression of *WWOX* have been linked to poor prognosis in a wide variety of cancers, particularly in ovarian cancer and OS [223, 224]. In previous reports by others, the function loss of *WWOX* has been linked to chromosomal deletions and translocations as well as loss of expression [215, 217]. In this study, we showed that 32% of OS samples have at least one chromosomal break within the *WWOX* gene, supporting the *WWOX* inactivation by chromosomal rearrangements. We further showed that *WWOX* gene was located in“broken regions”(discussed below) and SCNAs and chromosomal breaks in those regions were more likely to occur early. The results are consistent with the hypothesis that loss of *WWOX* expression is an early event in the pathogenesis of OS [217].

Genome-wide analysis revealed that chromosomal breaks are not randomly distributed and clustered in “broken regions”. About half of these regions overlapped with non-fragile sites, strongly suggestive of OS-specific fragility. It is noteworthy that OS-associated tumor suppressor genes including *TP53*, *RB1*, *WWOX*, *DLG2*, and *LSAMP*[198] are situated in the “broken regions”. SCNAs in those broken regions were more likely to be clonal events as opposed to those expected by chance. The early occurrence of breakages and the presence of multiple tumor suppressor genes in such regions may explain the complex and aggressive nature of OS.

We further revealed that SCNA breakpoints and chromosomal breaks were significantly correlated with diverse genomic properties, including Alu, L1, cruciform, G4, slip, triplex, Z-DNA, conserved elements, exon density, and indel rate. Genomic repeats such as L1 and Alu are interspersed throughout the human genome at high copy numbers, and non-allelic homologous recombination events between different copies lead to duplications, deletions, and inversions [181]. Repetitive DNA motifs may fold into non-B DNA conformation, thereby serving as chromosomal targets for DNA repair and recombination leading to the formation of structural variations including CNVs, inversions and translocations [180]. Therefore, it could be speculated that breakages probably occur at OS-specific fragile sites with the potential to form stable secondary structures (*i.e.* non-B DNA structures) and to consequently stall the replication fork.

Based on 20 patients including 9 osteosarcomas and 11 chordomas, Stephens *et.al.* [46] estimated that 25% of bone cancers were associated with chromothripsis. In our dataset, chromothripsis-like patterns occurred in about one third of patients, suggesting that chromothripsis is a widespread phenomenon in OS. Massive genomic rearrangement raised by the phenomenon of chromothripsis apparently represents an important mechanism of carcinogenesis, as distinct from progressive accumulation. As observed by Stephens [46], a single catastrophic event can occur while the chromosomes are being condensed for mitosis. However, the underlying cause of chromosomal damage is still unknown. Our analysis indicates that SCNAs in the *TP53*, *RB1* and *DLG2* genes are strongly associated with chromothripsis-like patterns in OS. Among them, *DLG2* frequently shows breakages in OS and may be a preferential target for chromothripsis and breakage [195]. *RB1* is significantly copy-number altered in OS, while the other candidate, *TP53* has already been linked to chromothripsis in medulloblastoma [53]. Utilizing an in vitro cell-based system, chromothripsis has been recently linked to hyperploidy [222]. Indeed, we have shown that compared with the diploid tumors, the hyperploid ones had more chance to get chromothripsis events and less favourable outcomes.

3.6 Conclusions

A comprehensive characterization of somatic copy number alterations (SCNAs) in a large cohort (n = 160) of osteosarcoma samples was undertaken in this study. A high percentage (98%) of the analysed OS samples were of sufficient quality for data analysis. The high degree of aneuploidy and large-size copy number alterations in OS was confirmed. Using GISTIC, a number of genes that are frequently targeted in OS were identified, of which *TP53*, *ATRX*, *FOXN1*, and *WWOX* are already known tumor suppressors associated with OS and other tumor types. Genome-wide analysis of chromosomal breaks revealed a tendency for confinement to genomic regions harbouring OS-associated tumor suppressor genes including *TP53*, *RB1*, *WWOX*, *DLG2*, and *LSAMP*. Breakage susceptibility in OS was found to be largely dependent on local genomic context. A complex breakage pattern - chromothripsis - has been suggested as a widespread phenomenon in OS correlated with OS patient survival. Through unlocking an OS-specific fragility pattern, a specific code has been revealed that may provide a basis for patient prognosis and offer a vital platform for therapeutic intervention in the future.

Chapter 4

Injury signals uncovered a regenerative program in mouse neural stem cells

This chapter is in Bobadilla,E. , Zhang,Y., Dehler,S., Xu, H. Frishman,D., Villalba,A.M. (2017) Injury signals uncover a regenerative program in mouse neural stem cells (manuscript in preparation). Enric Llorens-Bobadilla, I and Sascha Dehler contributed equally to this work. This study was designed and initiated by Enric Llorens-Bobadilla, Dmitrij Frishman and Ana Martin-Villalba. Enric Llorens-Bobadilla and Sascha Dehler did the biological experiments. Dmitrij Frishman and I conceived the bioinformatics part of this project. I did the methylation level calling, differentially methylated regions identification, annotation of the differentially methylated regions. The manuscript are now written by Enric Llorens-Bobadilla, Sascha Dehler and me and will be edited by Dmitrij Frishman and Ana Martin-Villalba.

4.1 Introduction

Stem cells display two unique characteristics: to self-renew and to differentiate into multiple cell types. The adult central nervous system (CNS) contains neural stem cells (NSCs) that are crucial for both brain development and adult neurogenesis. NSCs go through either symmetric division that generates two daughter NSCs, which have identical stem cell properties as the parental cell, or asymmetric division, which yieds one identical daughter NSC and more mature progenitors of all neural lineages by asymmetric division.

NSCs are present in few specialized areas of adult mammalian brain, such as the subventricular zone (SVZ) of the lateral ventricles and the subgranular layer of the dentate of the hippocampus [225, 226]. NSCs in the SVZ give rise to neuroblasts, which migrate to the olfactory bulb (OB) along the rostral migratory stream (RMS) and generate new neurons [227, 228].

It is believed that extracellular factors and intracellular process control cell fate specification and differentiation of NSCs [229]. Epigenetic mechanisms, such as DNA methylation are critical in cell type specification and tissue development. Previous studies have revealed that epigenetics play an essential role in the development of NSC, and that NSCs are prevented from differentiation when inhibiting DNA methylation [230, 231]. However, little is known about the role of epigenetic mechanism in regulating injury-induced neurogenesis.

Here we studied the molecular response of endogenous NSCs to ischemic injury. We examined how injury signals affect the DNA methylome of NSCs by whole genome bisulfite sequencing. We uncovered an injury-induced epigenetic program that encompasses the decommissioning of developmental transcription factors and enhancers selectively in NSCs.

4.2 Materials and Methods

4.2.1 Mouse sample

Young adult mice are exposed to ischemic injury and two days later NSCs (GLASTs$^+$Prom1$^+$) from their in vivo niche are isolated as previously described [232]. For comparative analyses, PSA-NCAM$^+$ neuroblasts (NBs, the NSC progeny) and Prom1$^+$ niche oligodendrocytes (OLs, a postmitotic glial cell type) are simultaneously profiled.

4.2.2 Tagmentation-based whole genome bisulfite sequencing

20-30ng of genomic DNA were isolated from freshly sorted cells for tagmentation-based whole genome bisulfite sequencing (T-WGBS) libraries. It is processed as previously described in collaboration with Dr. Dieter Weichenhan [116, 233]. All sequencing runs were conducted at the DKFZ Genomics and Proteomics core facility. T-WGBS libraries were sequenced in an Illumina HiSeq 2000 for 101bp paired-end sequencing.

4.2.3 Preprocessing and read mapping

Raw T-WGBS reads were analyzed with the FastQC quality control tool (v0.10.1). Trim_galore_0.4 [122] was used to trim off adapter sequences and remove bases with a Phred base quality score below 30. Phred describes the error probability (P) of a single base call in a way that

$$phred - score = -10 \log_{10}(P)$$

Thus, a phred-score of 30, for example, gives an error probability of 0.1% and hence an accuracy of 99.9%.

Trimmed reads were aligned to mouse reference genome (GRCm38/Ensembl) using BWA-Meth_0.1 [234] with the default settings. Duplicate reads were removed after alignment using the function *MarkDuplicates* from the Picard tool [235].

4.2.4 Detection of CpG methylation level

BS-SNPer is a program for variation detection of Bisulfite sequencing using approximate Bayesian modeling. CpG methylation calls were made using BS-SNPer [135], with the following parameters: -minquali 20, -mincov 10, -mapvalue 30. The output CpG methylation are strand-independent where counts from the two Cs in a CpG and its reverse complement (position i on the plus strand and position i+1 on the minus strand) were combined and assigned to the position of the C in the plus strand. Moreover, we used the script "*filterCG_SNP.pl*"of BS-SNPer to discard C nucleotides that have been confirmed to be C>T SNPs.

4.2.5 Identification of differentially methylated regions

The methylation value of strand-merged CpG sites from the two injured and two non-injuried replicates were used to identify injury-induced difference in methylation, using the R package bsseq [124]. The smoothing was carried out with the parameters ns=20, h=250, maxGap=100,000,000. Loci with coverage ≥ 4 in all the replicates of non-inury and injuried samples were retained. Injury-induced differentially methylated regions (DMRs) were identified using t-statistic quantile cut-offs of 0.01 and 0.99, requiring with at least 5 CpGs per DMR and exhibiting a mean methylation change at least 20% between injuried and non-injuried samples.

Hypo DMRs (hypoDMRs) are statistically significant regions with mean methylation value of injuried samples less than that of non-injuried samples.

Hyper DMRs (hyperDMRs) are statistically significant regions with mean methylation value of injuried samples greater than that of non-injuried samples.

4.2.6 Genomic and functional annotation of CpG sites

Gene locations were defined based on the Ensembl/GRCm38 assembly. The 5'-most transcript start site (TSS) on the plus strand were selected as the single TSS of genes with multiple transcripts. The reverse (3'-most TSS) was done for genes on the minus strand. We limited our analysis to protein-coding genes, resulting in 22,082 TSSs in total. We extracted the exon coordinates from the transcript annotation file and removed overlapping exons. The exonic region was subtracted from the genic region to get the intronic regions and the gene region was subtracted from the whole genome to get the intergenic regions. CpG islands were downloaded from the UCSC [108]. For location of a site relative to a gene, we used these categories: $TSS \pm 500bp$ (from 1 to 500bp downstream or upstream of the TSS), $TSS \pm 2kb$ (from 1 to 2000bp downstream or upstream of the TSS). Promoter is defined from 1 to 1000bp upstream and 1 to 200bp downstream of TSS.

PhastCons [158] conservation scores from alignment of 59 vertebrate genomes with mouse genome were obtained from the UCSC genome Browse [108]. We examined the conservation status of hypoDMRs. PhastCons conservation score was calculated for 50-bp windows of 1kb up- and downstream around the center of all hypoDMRs. The average PhastCons conservation for each window was plotted.

4.2.7 Motif analysis

We used the hypergeometric optimization of motif enrichment (HOMER) [236] tool to searched for enrichment of known motifs within hypoDMRs with the parameters -size given and -cpg , and otherwise default parameters. The known motifs used in our analysis were derived from the HOMER tool [236].

4.2.8 Gene ontology and pathway analysis

Genomic regions of hypoDMRs were used as input in GREAT package [237]. The genome mm10 was used as reference and the nearest gene within a 1000kb distance was considered to be associated with a particular region.

For hierarchical representation, the nearest gene for the hypoDMRs were used as input in ClueGo [238], which was run as a plugin of Cytoscape v 3.4.0 [239].

4.2.9 Histone marks of development enhancers analysis

Previous published ChIP-seq peak data for histone modifications of H3k04me1 and H3k27ac of E14.5 developing brain (Dev.Brain) and Olfactory bulb (OB) (Table 4.1) were downloaded from the mouse Encode/LICR [240]. The peak regions were converted to mm10 using UCSC *liftOver* [108].

Table 4.1: Summary of ChIP-seq peak data for histone modifications

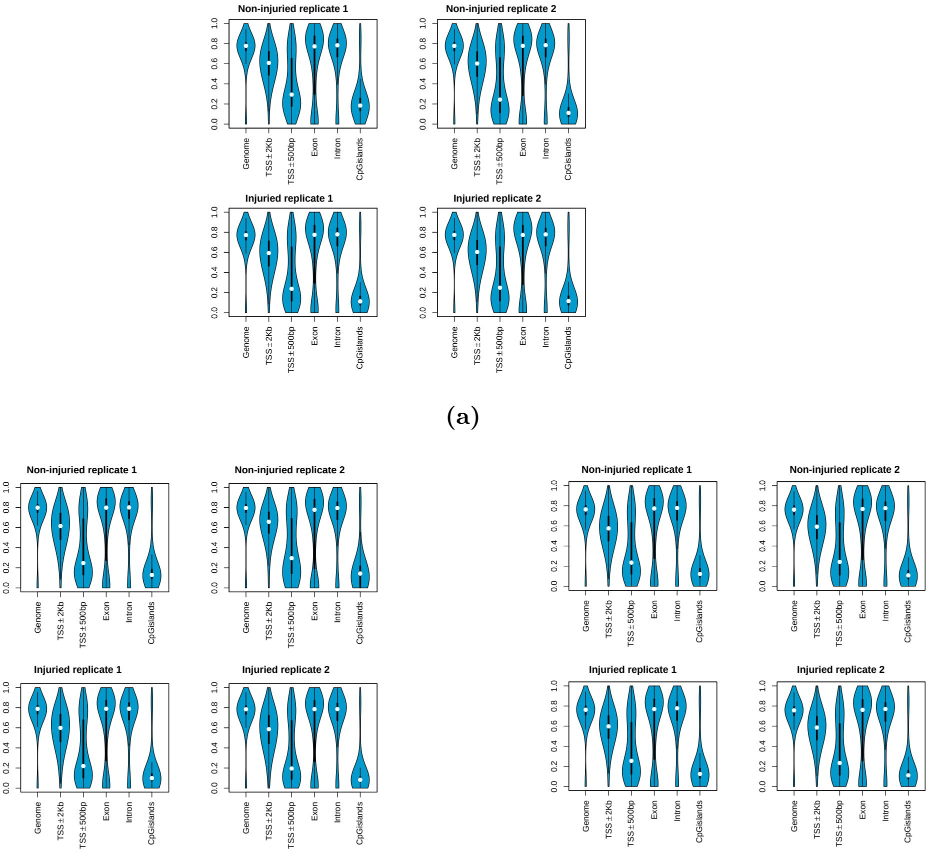
Cell Type	H3k04me1 Peak Regions	H3k27ac Peak Regions
E14.5 developing brain	131394	36495
Olfactory bulb	94715	36596

We investigated whether hypoDMRs offer a rich source of enhancers compared with random genomic sequences. We compared the number of hypoDMRs that overlap with the enhancer peak regions against the random background expectation. To assemble background datasets, we generated for each enhancer peak data set 200 randomized data sets, matched for enhancer region size, chromosome. We also exclude DAC blacklist genomic regions obtained from UCSC table browsers [108] because these regions contain signal artifacts in sequencing experiments. The numbers of overlapping hypoDMRs were averaged to get the random background, and the enrichment of the observed overlap against the random background was assessed using a one-sided binomial test.

4.3 Results

4.3.1 Whole-genome DNA methylation analysis in the injuried SVZ

In order to study the DNA methylation profiles of small amounts of sorted cell populations at genome-wide resolution we applied the recently developed tagmentation-based whole genome bisulfite sequencing (T-WGBS). In total, two replicates for each cell type and condition were sequenced to obtain an average coverage of 15X per replicate (30X combined). On average, we measured the methylation level of 19 million CpG sites (coverage$\geq$4) per sample (range of 17 million to 19 million). As expected, most CpG



(a)

(b)

(c)

Figure 4.1: DNA methylation levels of different genomic features. (a) DNA methylation levels of genomic features in injuried and non-injuried NSCs. The thickness of the bars indicates densities of CpGs at the y axis ratio, and the white circle indicates the median. (b) Equivalent plot for DNA methylation in NBs. (c) Equivalent plot for DNA methylation in OLs.

sites were highly methylated throughout the genome except very close to transcription start sites (TSS) and CpG islands (Figure 4.1a-4.1c).

To examine the global distribution of DNA methylation, we divided the genome into 10-kb bins and calculated CpG methylation in the injured and non-injuried samples (Figure 4.2a). Both injured and non-injuried methylomes were highly methylated, although non-injuried methylomes containing a little higher abundance of mCG (Figure 4.2).

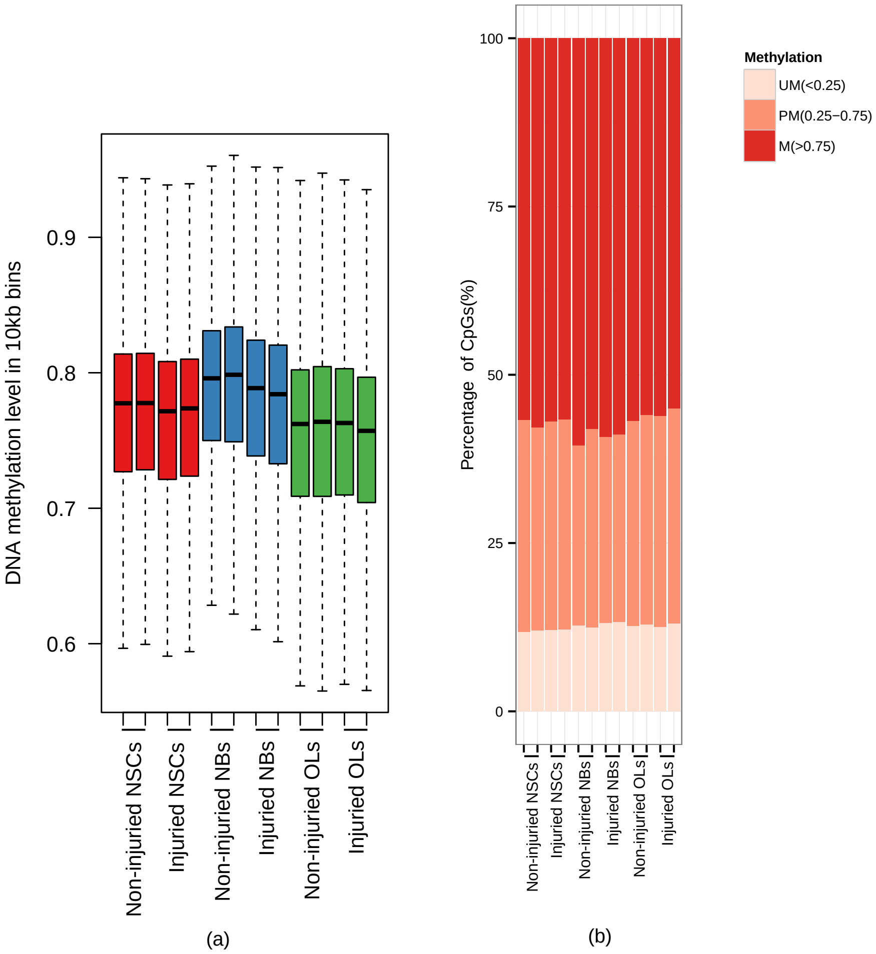


Figure 4.2: Genome-wide methylation level in NSCs, NBs and OLs. (a) Box plots of the percent mCG distribution for each sample, calculated from non-overlapping 10-kb bins spanning the mouse genome. (b) Percentage of highly methylated CpGs, partially methylated CpGs and unmethylated CpGs for each sample.

4.3.2 DNA methylation changes caused by injury

We hypothesized that epigenetic changes caused by brain injury may contribute to increased fate plasticity. Thus, we conducted analysis on the methylation changes after ischemic injury. To explore how the methylome changes to the response of NSCs, NBs and OLs to injury signals, we used bsseq [124] to analyze differential CpG methylation of non-injuried and injuried samples of each cell type separately. Only loci with coverage ≥4 in all samples were retained, which left 17625739, 16926719 and 18161496 CpG sites for NSCs, NBs and OLs, respectively. We defined injury-induced differential methylated regions (DMRs) as a region of 5 or more CpG sites exhibiting a significant difference in methylation between the two groups and an absolute mean methylation difference above 0.2. Using these criteria, we identified 2735, 1125 and 1087 DMRs for NSCs, NBs and OLs, respectively (Figure 4.3).

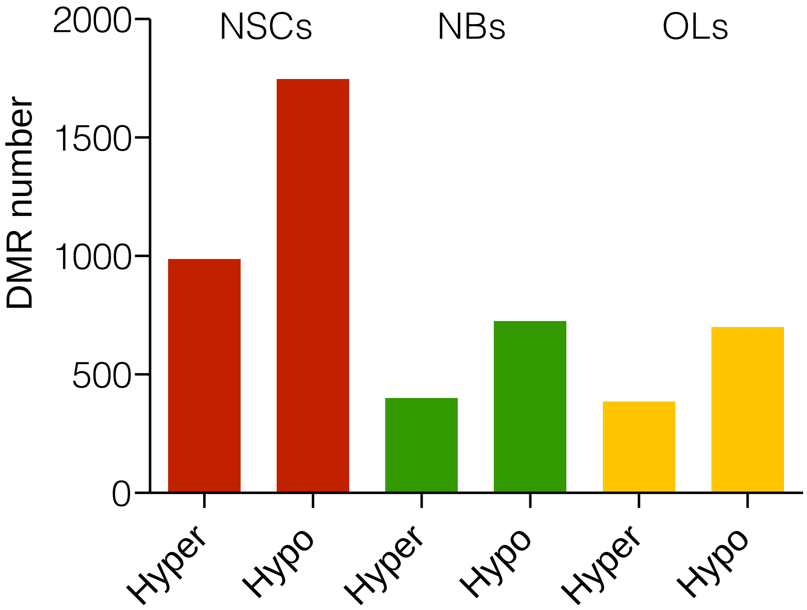


Figure 4.3: The number of injury-induced differential methylated regions(DMRs) in NSCs, NBs and OLs. Hypo: hypo methylated. Hyper: hyper methylated.

We found that only a small proportion of (5.9%) hypoDMRs overlapped with a promoter (Figure 4.4a). The vast majority of hypoDMRs were located distal to TSSs (Figure 4.4b). Moreover, hypoDMRs occurred in genomic regions that show increased level of evolutionary conservation (Figure 4.4c).

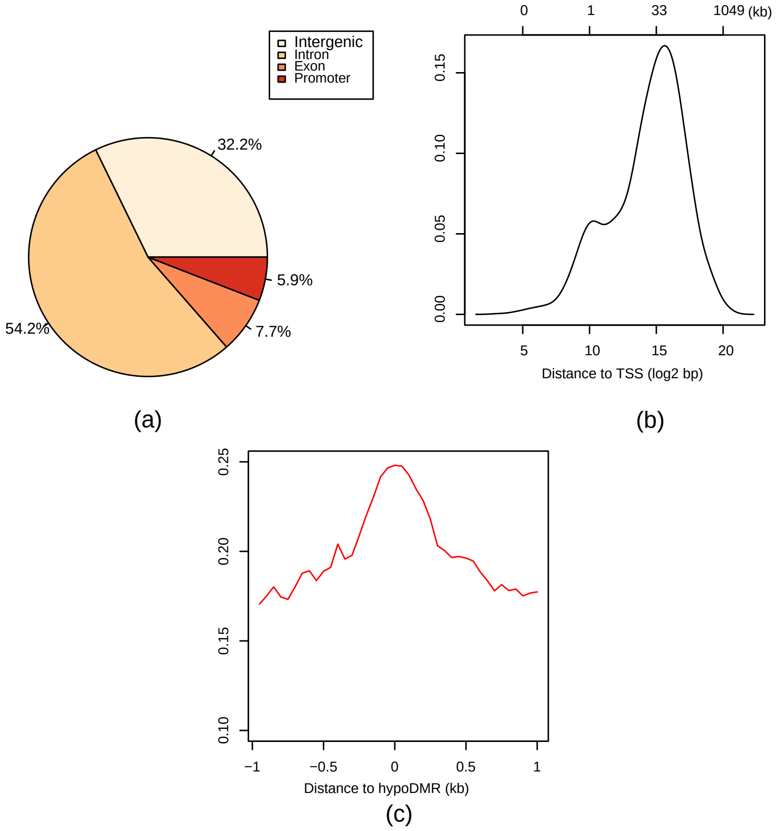


Figure 4.4: Injury-induced hypoDMRs in NSCs. (a) Pie charts showing the distribution of hypoDMRs in different genomic regions. (b) Distribution of distances of hypoDMR to the nearest TSS. (c) Average phastCons conservation score within 50-bp windows, around the center of hypoDMRs.

4.3.3 Transcription factor binding sites at injury-induced DMRs in NSCs

We next examined whether hypoDMRs are enriched with transcription factor binding sites (TFBS) that may potentially expose the epigenetic changes to an injury-induced upstream regulator. We applied the HOMER tool [236] to do the TFBS enrichment analysis across hypoDMRs. We found that hypoDMRs are enriched with TFBS that are associated with the injury-induced TFs such as C-jun or Stat1. Interestingly, Isl1 appeared among the top-enriched TFBS exclusively in NSCs (1e-11) (Figure 4.5).

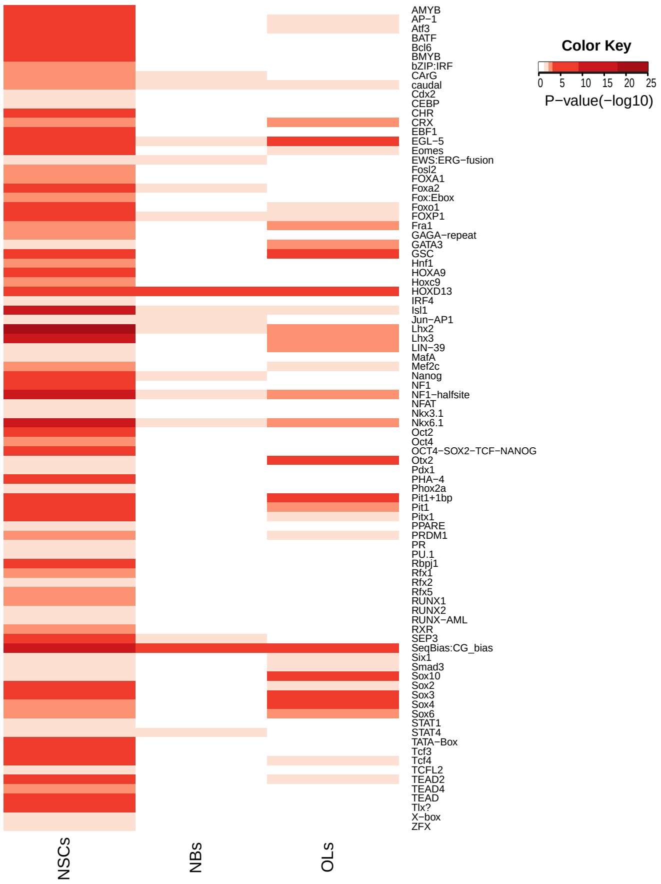


Figure 4.5: Heatmap representing the enrichment of transcription factor binding motifs for injury-induced hypoDMRs in each cell type. Each row represents a motif.

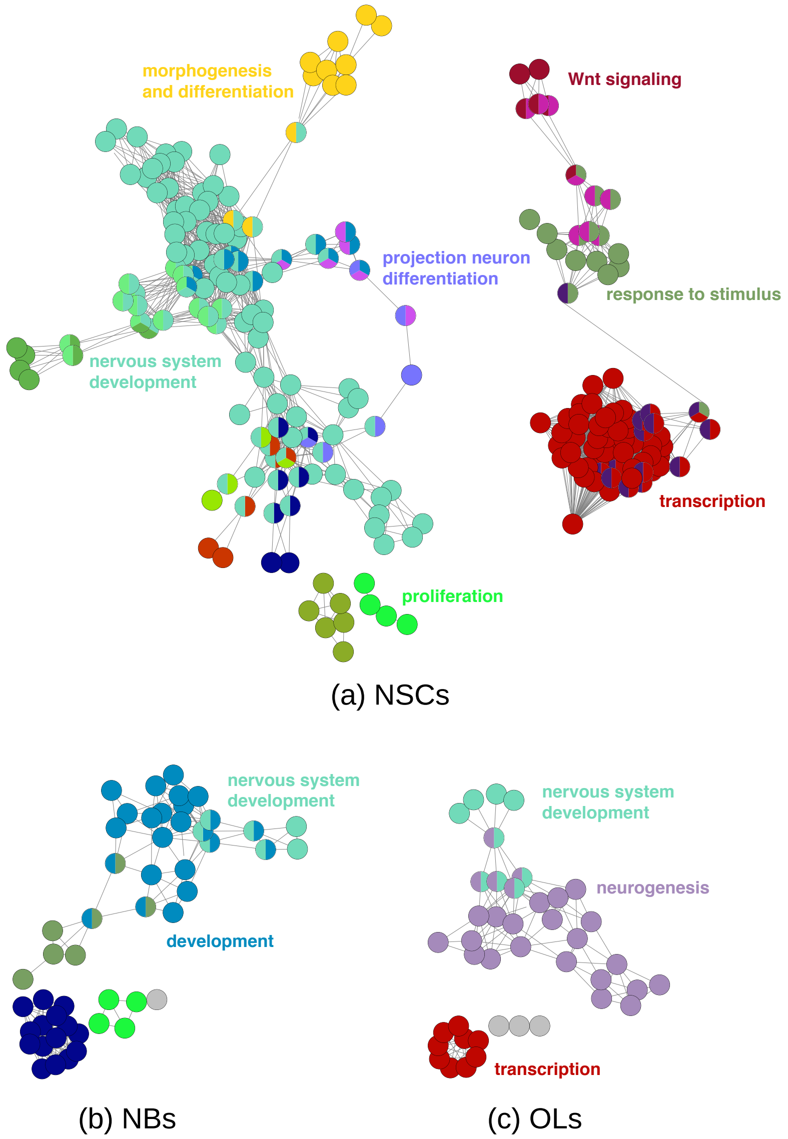
4.3.4 Specific injury-induced demethylation of developmental transcription factor enhancers in NSCs

Enhancers work as cis-regulatory elements that activate gene expression and can operate tens of thousands of base pairs away from the target gene [241]. To study the function of the genes associated to our hypoDMRs we used the genomic regions enrichment of annotations tool (GREAT)

[237] to map the nearest gene within a 1000kb distance. We revealed a high enrichment for transcription factors (TFs) among the genes associated to hypo DMRs (102 regions mapped to 72 known TFs, FDR<2.3e-14). These TFs included several known regulators neuronal subtype specification, such as *Bcl11b*, *Pax6* or *Neurod2*. To globally visualize the gene ontology term enrichment we created Cytoscape maps. This representation uncovered a very elaborate network of highly enriched gene ontology terms in NSCs (Figure 4.6a). Specifically, hypoDMRs in NSCs are significantly near genes that are responsible for developmental process, neurogenesis and differentiation-related categories. We also did the same analyses on injury-induced hypoDMRs in NBs and OLs we observed a much more modest functional enrichment (Figure 4.6b-4.6c). Together, the genes under putative control by the enhancers that become hypomethylated after injury are enriched in functions relevant to tissue repair, including the re-activation of a developmental program.

4.3.5 Injury induces demethylation at developmental enhancers in NSCs

To further characterize these hypoDMRs we compared them to ENCODE two brain ChIP-seq datasets of histone modification [240]. Specifically, one is adult OB, the ‘default‘ program for NSCs; and the other is E14.5 Dev.Brain, because it is a developmental period where multiple neuronal subtypes instead of that ‘default‘ program are being generated. Following the enrichment method described in section 4.2.9, we revealed that hypoDMRs are more closely correlated with the enhancer landscape of the developing brain than with the adult olfactory bulb (Figure 4.7). Specifically, 62% of the regions (1091 out of 1747) overlapped with E14.5 Dev.Brain H3k04me1-marked enhancers, compared to 24% (428 out of 1747) overlapped with adult OB enhancers (Figure 4.7). Similarly, the active enhancer mark H3k27ac, also significantly associated correlated with hypoDMRs, and 11% and 12% overlapped with active enhancers in E14.5 Dev.Brain and adult OB respectively. In summary, a subset of developmental enhancers becomes permissive, through loss of DNA methylation, in NSCs after ischemic injury.



(a) NSCs

(b) NBs

(c) OLs

Figure 4.6: Injury-induced demethylation at enhancers. (a) Enriched GO network groups using ClueGO. P<0.00001 are shown. Each node represents a biological process. Edges represent connections between the nodes and the length of each edge reflects the relatedness of two processes. Node color, represents the class that they belong. Mixed coloring means that the specific node belongs to multiple classes. Ungrouped terms are not shown. (b) Equivalent plots for hypoDMRs in NBs. (c) Equivalent plots for hypoDMRs in OLs.

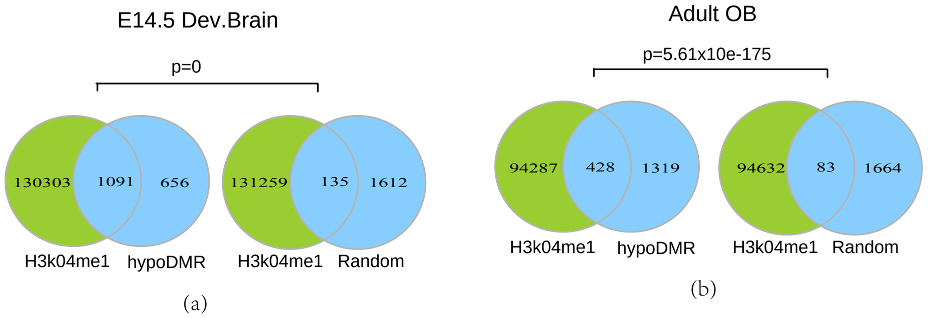


Figure 4.7: Venn diagram showing the enrichment of hypoDMRs in NSCs on enhancers marked by H3k04me1 and H3k27ac. P values are calculated by the bionomial test. (a) Overlap between hypoDMRs with H3k04me1 of E14.5 Dev.Brain tissue compared with that of hypoDMRs with random regions. (b) Overlap between hypoDMRs with H3k04me1 of adult OB tissue compared with that of hypoDMRs with random regions.

4.4 Conclusions

In this part, we have investigated how the DNA methylome integrates injury signals to mediate cell plasticity. To achieve this, we analyzed the whole genome bisulfite sequencing data of non-injuried and injuried samples in NSCs,NBs and OLs.

Once mapped to the genome, we confirmed the previous reported DNA methylation pattern in somatic tissues-CpG sites were ubiquitously methylated throughout the genome except near transcription start sites (TSSs) and in CpG islands. Interestingly, we found that less than 6% of hypoDMRs in NSCs overlapped with a putative promoter and that the vast majority of these sites were located distal to TSSs. We were able to show that injury-induced demthylation regions in NSCs show increased level of evolutionary conservation, a finding supporting that they are functional important.

Previous work reported that histone modifications such as H3k27ac and H3k04me1 correlate with DNA hypomethylation at active enhancers [242]. We confirmed this correlation in the injury-induced DMRs, being stronger at enhancers known to be active in the developing brain, suggesting epigenetic priming for future differentiation.

We have further conducted an analysis of transcription factor binding site to decipher the possible pathway responsible for injury-induced hypomethylation at enhancers. We revealed that several motifs for injury-induced TFs are enriched within the hypoDMRs. Interestingly, we found that Isl1 is among the top-enriched TFBS exclusively in NSCs (1e-11). Given that Isl1 is required during development for the generation of striatal medium spiny neurons, these results suggest that the unmasking of developmental enhancers, might facilitate the generation of striatal neurons after injury.

Our functional enrichment analysis revealed a high enrichment for transcription factors (TFs) among the genes associated to hypoDMRs (102 regions mapped to 72 known TFs, FDR<2.3e-14). These TFs included several known regulators neuronal subtype specification, such as *Bcl11b*, *Pax6* or *Neurod2*. The gene ontology analysis uncovered that injury-induced hypoDMRs are significantly near genes that are responsible for developmental process, neurogenesis and differentiation-related categories.

Chapter 5

Summary

This work aims to increase the understanding of mechanisms of copy number variation (CNV) in cancer genomes and the role of DNA methylation during differentiation of neural stem cells (NSCs).

In the first part of this work we investigated the somatic copy number alterations (SCNAs) from different cancers. We collected different genomic features including DNA conformation, sequence-based such as repeats (LINE, LTR, SINE), structural, gene regulation, evolutionary and functional features. Based on the SCNA data from 11 individual cancer types from TCGA we applied multiple linear regression to identify the potential predictors for SCNA patterns. Our findings showed that distance to telomere, distance to centromere and direct repeats coverage are the strong correlates for SCNA generation in cancers.

As breakpoints of SCNAs are not randomly distributed across the genome, they tend to cluster in regions and some of these regions are statistical significant, termed as breakpoint hotspots. We investigated how genomic context contribute to the pattern of common breakpoint hotspots in cancer genomes. Based on the statistical methods including rare event logistic regression and random forest we revealed that distance to telomere and direct repeats coverage are able to distinguish common hotspots and non-hotspots of SCNA breakpoints.

The second part of this work is focused on characterizing SCNAs in osteosarcoma (OS). The complexity of OS genome drives us to characterize the specific set of driver genes. Based on the SCNA breakpoint data, we detected a number of genes more likely to be targeted by breakpoints, including well-known driver genes (*e.g. TP53* and *ATRX*) and other OS-related genes, such as *WWOX*. Our findings were also confirmed by the gene set identified by a permutation statistical method for breakage analysis.

Previous studies have shown that DNA breakage can occur invariably at non-B DNA structure-forming sequences or highly homologous genomic repeats. Thus, we investigated the association between chromosomal breaks, SCNA breakpoints and multiple genomic features. Confirming previous research we showed that SCNA breakpoints and chromosome breaks were significantly enriched in diverse genomic features such as Alu, L1, cruciform and indel rate. Moreover, we found that half of breaks hotspots overlapped with non-fragile sites, suggesting the specific fragility of OS genome.

Many studies confirmed that chromothripsis were occurred in many tumor types, especially in bone cancers. We suggest that chromothripsis is a prevalent phenomenon in OS. Applying a chromothripsis detection method, we found that chromothripsis-like patterns (CTLP) occurred in about one third of patients. Although the cause of this catastrophe event is still unknown, our findings showed that chromothripsis-like patterns are strongly correlated with SCNAs in the *TP53*, *RB1* and *RUNX2* genes. We further investigated the relationship between chromothripsis-like pattern and clinical data. Our analysis revealed that the survival time of patients with CTLP patterns in their tumors is significantly shorter than those without CTLP.

In the last part of this work we analyzed the whole-genome bisulfite sequencing data for NSCs in injuried and non-injuried conditions. We examined the global distribution of DNA methylation by dividing the genome into 10-kb bins and calculated CpG methylation in the injured and non-injuried samples. As expected for mammalian cells, most CpG dinucleotides were highly methylated throughout the genome except near transcription start sites (TSSs) and CpG islands.

Given that over two thirds of all Single Nucleotide Polymorphisms (SNPs) occur in a CpG context, BS-SNPer considers sequence variation to avoid wrong inference of methylation state. Next, we applied BS-SNPer program to identify the differentially methylated regions (DMRs) for each cell type. Based on the PhastCons conservation scores from alignment of 59 vertebrate genomes with mouse genome, we calculated the average conservation scores for the DMRs, 1kb upstream and 1kb downstream regions, respectively. Our results demonstrated that hypoDMRs occurred in genomic regions with elevated level of evolutionary conservation, supporting their functional importance.

We also found a high enrichment for transcription factors (TFs) among the genes related to hypoDMRs (102 regions mapped to 72 known TFs, FDR<2.3e-14). These TFS included several interesting regulators for neuronal subtype specification, such as *Bcl11b*, *Pax6* or *Neurod2*. By integrating our methylation maps with ChIP-seq data on two histone marks, we showed that active demethylation occurs almost at distal regulatory elements, particularly enhancers. We further conducted an analysis of transcription factor binding sites for injury-induced hypoDMRs. We revealed that a very interesting motif Isl1 is significantly enriched in hypoDMRs in NSCs. Isl1 is necessary for the generation of striatal medium spiny neurons, suggesting that injury may activate the devleopmental enhancers to facilitate the generation of striatal neurons.

Chapter 6

Appendix

6.1 Supplementary Tables

Table 6.1: Genomic regions significantly altered identified by GISTIC in 157 osteosarcoma samples

Chr.[1]	Region	Extended Region	Type	Genes
chr1	chr1:72768081-72771450	chr1:72768081-72771450	CN Gain	
chr1	chr1:120532528-120540803	chr1:120532228-121119145	CN Gain	NOTCH2
chr1	chr1:150915428-150986518	chr1:150106621-151292631	CN Gain	SETDB1; CERS2; ANXA9; FAM63A; PRUNE
chr1	chr1:152762026-152771308	chr1:152761930-152771308	CN Loss	LCE1D
chr1	chr1:169225449-169242083	chr1:169225449-169242083	CN Loss	NME7
chr1	chr1:248758246-248787569	chr1:248753426-248794436	CN Loss	
chr2	chr2:34696356-34729740	chr2:34696356-34729740	CN Loss	
chr2	chr2:87021286-87054784	chr2:86863077-88263441	CN Gain	CD8B; RMND5A
chr2	chr2:97765044-97889750	chr2:97449536-98128314	CN Gain	ANKRD36
chr2	chr2:242013345-242045252	chr2:241988330-242195981	CN Loss	SNED1; MTERF4; MTERFD2
chr3	chr3:37983108-37986935	chr3:37983108-37986935	CN Loss	CTDSPL
chr3	chr3:116548005-116553148	chr3:116530653-116677267	CN Loss	
chr3	chr3:189362262-189363677	chr3:189362262-189371001	CN Loss	TP63
chr4	chr4:34783101-34824462	chr4:34783101-34828255	CN Loss	
chr4	chr4:47585962-47633769	chr4:47274810-47643922	CN Gain	ATP10D; CORIN
chr4	chr4:55144803-55146541	chr4:54583847-55227042	CN Gain	PDGFRA
chr4	chr4:69495772-69521133	chr4:69495772-69521133	CN Loss	UGT2B15
chr4	chr4:161950067-162007018	chr4:160234964-162282493	CN Gain	
chr5	chr5:6522965-6525445	chr5:6522965-6525445	CN Loss	
chr5	chr5:38738377-38760633	chr5:38585742-38917416	CN Gain	OSMR-AS1
chr5	chr5:180377034-180410761	chr5:180375094-180424577	CN Loss	BTNL8
chr6	chr6:255666-257069	chr6:255666-257417	CN Loss	
chr6	chr6:45448960-45459235	chr6:45269549-45709252	CN Gain	RUNX2
chr6	chr6:77438359-77455244	chr6:77438359-77455244	CN Loss	
chr7	chr7:3971188-4071542	chr7:3770143-5137384	CN Gain	SDK1
chr7	chr7:142476621-142481638	chr7:142476621-142486098	CN Loss	TCRBV2S1; TCRVB; PRSS3P2; PRSS2
chr7	chr7:154391477-154399616	chr7:154391477-154400278	CN Loss	DPP6
chr8	chr8:1659358-1676610	chr8:492396-1676610	CN Loss	
chr8	chr8:24974355-24989291	chr8:24974355-24989291	CN Loss	
chr8	chr8:39208722-39226339	chr8:39026273-39226339	CN Gain	ADAM5
chr8	chr8:39248531-39352993	chr8:39238548-39386079	CN Loss	ADAM3A
chr8	chr8:49554073-49572201	chr8:48810937-50417372	CN Gain	LOC101929268
chr8	chr8:72215337-72216222	chr8:72215310-72216684	CN Loss	EYA1
chr8	chr8:98718483-98733201	chr8:98240419-98790083	CN Gain	MTDH
chr8	chr8:128735487-128738992	chr8:128305898-129002357	CN Gain	BC042052; CASC11
chr9	chr9:21968624-21976768	chr9:21850263-22028704	CN Loss	MTAP; CDKN2A
chr10	chr10:24376468-24378414	chr10:24376468-24379860	CN Loss	KIAA1217
chr10	chr10:47058829-47061065	chr10:47057570-47061065	CN Loss	ANXA8
chr10	chr10:78257335-78261389	chr10:78257335-78261389	CN Loss	C10orf11

Continued on next page

Table 6.1 – *Continued from previous page*

Chr.	Region	Extended Region	Type	Genes
chr11	chr11:5797748-5808726	chr11:5784971-5809277	CN Loss	TRIM22; OR52N5; TRIM5
chr11	chr11:55374167-55403443	chr11:55374167-55433103	CN Loss	
chr11	chr11:84184013-84184955	chr11:84159254-84222629	CN Loss	DLG2
chr11	chr11:101517518-101927296	chr11:101316304-102237928	CN Gain	ANGPTL5; KIAA1377; C11orf70
chr11	chr11:128681554-128683826	chr11:128679603-128683826	CN Loss	FLI1
chr12	chr12:869296-873583	chr12:867422-874562	CN Loss	WNK1
chr12	chr12:34383785-34485085	chr12:34261964-35800000	CN Gain	
chr12	chr12:58135816-58305277	chr12:58124923-58322883	CN Gain	AGAP2; TSPAN31; MIR6759; CDK4; DM110804; MARCH9; CYP27B1; METTL1; METTL21B; TSFM; AVIL; MIR26A2; CTDSP2; AK130110
chr12	chr12:99795602-99798726	chr12:99795602-99800925	CN Loss	ANKS1B
chr13	chr13:38071673-38086565	chr13:38071673-38086565	CN Loss	
chr14	chr14:23100225-23120359	chr14:22844274-23307453	CN Gain	
chr14	chr14:106335832-106489591	chr14:106335832-106527892	CN Gain	KIAA0125; ADAM6
chr14	chr14:106557833-106603522	chr14:106536937-106603522	CN Loss	BC042994
chr14	chr14:106885733-106920359	chr14:106885733-106920359	CN Loss	
chr15	chr15:76879983-76895555	chr15:76879983-76895555	CN Loss	SCAPER
chr15	chr15:99530128-99880948	chr15:99300869-99959809	CN Gain	PGPEP1L; AL109706; SYNM; TTC23; HSP90B2P; LRRC28
chr16	chr16:19944410-19968380	chr16:19944410-19968380	CN Loss	
chr16	chr16:78372017-78382206	chr16:78372017-78384869	CN Loss	WWOX
chr17	chr17:7582979-7583221	chr17:7578835-7583723	CN Loss	TP53
chr17	chr17:17037165-17065229	chr17:16991233-17074052	CN Gain	MPRIP
chr17	chr17:26843566-26848243	chr17:26843402-26848243	CN Loss	FOXN1
chr17	chr17:39423181-39430490	chr17:39423181-39430490	CN Loss	
chr17	chr17:44223496-44279974	chr17:44213141-44279974	CN Gain	KANSL1
chr18	chr18:11252274-11464401	chr18:10812801-11589974	CN Gain	
chr18	chr18:46944321-46952804	chr18:46944321-46953209	CN Loss	DYM
chr19	chr19:638104-658093	chr19:638104-1291591	CN Loss	FGF22; RNF126
chr19	chr19:7151245-7195285	chr19:7146765-7302221	CN Gain	INSR
chr19	chr19:30299491-30321146	chr19:30284135-30344003	CN Gain	CCNE1
chr19	chr19:42422360-42428514	chr19:42422120-42428735	CN Loss	ARHGEF1
chr20	chr20:1560269-1560674	chr20:1557189-1560674	CN Loss	SIRPB1
chr20	chr20:29917644-29956205	chr20:29433517-30040495	CN Gain	
chr21	chr21:37237166-37248079	chr21:37064469-37368136	CN Gain	RUNX1
chr22	chr22:19570331-19572970	chr22:19570331-19572970	CN Loss	
chr22	chr22:23146865-23207698	chr22:23146262-23240129	CN Gain	DKFZp667J0810; MIR650
chr22	chr22:51105118-51106136	chr22:51104136-51106136	CN Loss	
chrX	chrX:825934-826729	chrX:821776-826729	CN Loss	
chrX	chrX:2302238-2302530	chrX:2302238-2302530	CN Gain	
chrX	chrX:6659340-6659459	chrX:6659303-6661807	CN Loss	
chrX	chrX:31458638-31458832	chrX:31457616-31459915	CN Loss	
chrX	chrX:76948103-76949541	chrX:76896688-77032001	CN Loss	
chrX	chrX:85291897-85293444	chrX:85291897-85295272	CN Gain	
chrX	chrX:115135704-115138008	chrX:115135704-115153407	CN Loss	
chrX	chrX:122900376-122900406	chrX:122900268-122900751	CN Loss	
chrX	chrX:136493788-136495362	chrX:136493788-136495561	CN Loss	
chrX	chrX:147320320-147320888	chrX:147318675-147326708	CN Loss	
chrX	chrX:153963340-153963495	chrX:153960395-153963495	CN Loss	
chrX	chrX:155086346-155086387	chrX:155086346-155086387	CN Gain	
chrY	chrY:20836985-21024837	chrY:17235271-22252906	CN Loss	
chrY	chrY:22275025-22410762	chrY:22264667-22465913	CN Gain	

[1] Chromosome

Table 6.2: Genes contained in the regions of frequent copy number alterations as identified by GISTIC analysis.

Gene Symbol	Chromosome	Start	End	Length
ADAM3A	chr8	39308563	39380508	71946
ADAM5	chr8	39172181	39274897	102717
ADAM6	chr14	106435817	106438358	2542
AGAP2	chr12	58118075	58135944	17870
AK130110	chr12	58230875	58236325	5451
AL109706	chr15	99571772	99574275	2504
ANGPTL5	chr11	101761404	101787253	25850
ANKRD36	chr2	97779232	97930257	151026
ANKS1B	chr12	99128568	100378432	1249865
ANXA8	chr10	47011755	47174143	162389
ANXA9	chr1	150954498	150968114	13617
ARHGEF1	chr19	42387266	42434296	47031
ATP10D	chr4	47487409	47595503	108095
ATRX[1]	chrX	76760355	77041755	281401
AVIL	chr12	58191159	58209852	18694
BC042052	chr8	128698587	128746211	47625
BC042994	chr14	106576813	106598011	21199
BC062752	chrY	20934593	20981392	46800
BTNL8	chr5	180326076	180377906	51831
BV03S1J2.2	chr7	142428689	142499111	70423
BV6S4-BJ2S2	chr7	142462183	142494293	32111
C10orf11	chr10	77542518	78317126	774609
C11orf70	chr11	101918168	101955291	37124
CASC11	chr8	128712852	128746213	33362
CCNE1[1]	chr19	30302900	30315215	12316
CD8A	chr2	87011727	87035519	23793
CD8B	chr2	87042459	87089047	46589
CDK4[1]	chr12	58141509	58146230	4722
CDKN2A[1]	chr9	21967750	21994490	26741
CERS2	chr1	150937648	150947479	9832
CHM	chrX	85116184	85302566	186383
CORIN	chr4	47596014	47840123	244110
CTDSP2	chr12	58213709	58240747	27039
CTDSPL	chr3	37903668	38025960	122293
CYP27B1	chr12	58156116	58160976	4861
DHRSX	chrX	2137554	2419015	281462
DKFZp667J0810	chr22	22786692	23248968	462277
DLG2	chr11	83166055	85338314	2172260
DM110804	chr12	58145424	58145484	61
DMD	chrX	31137344	33357726	2220383
DPP6	chr7	153584181	154686000	1101820
DYM	chr18	46570171	46987079	416909
EYA1	chr8	72109667	72274467	164801
FAM63A	chr1	150969300	150980854	11555
FGF22	chr19	639925	643703	3779
FLI1[1]	chr11	128556429	128683162	126734
FOXN1	chr17	26833277	26865175	31899
GAB3	chrX	153903526	153979858	76333
HSP90B2P	chr15	99797729	99800481	2753
INSR	chr19	7112265	7294011	181747
KANSL1	chr17	44107281	44302740	195460
KANSL1-AS1	chr17	44270938	44274089	3152
KIAA0125	chr14	106355979	106398502	42524
KIAA1217	chr10	23983674	24836777	853104
KIAA1377	chr11	101785745	101871796	86052
LCE1D	chr1	152769226	152770657	1432
LOC101929268	chr8	49464126	49611069	146944
LRRC28	chr15	99791566	99927280	135715
MARCH9	chr12	58148880	58154193	5314
METTL1	chr12	58162350	58165914	3565
METTL21B	chr12	58166382	58176324	9943
MIR26A2	chr12	58218391	58218475	85
MIR650	chr22	23165269	23165365	97
MIR6759	chr12	58142400	58142465	66
MPRIP	chr17	16946073	17095962	149890
MTAP	chr9	21802634	22029593	226960
MTDH	chr8	98656406	98742488	86083
MTERF4	chr2	242026508	242041747	15240

Continued on next page

Table 6.2 – *Continued from previous page*

Gene Symbol	Chromosome	Start	End	Length
MTERFD2	chr2	242034544	242041747	7204
NME7	chr1	169101767	169337201	235435
NOTCH2[1]	chr1	120454175	120612317	158143
OR52N5	chr11	5798863	5799897	1035
OSMR-AS1	chr5	38693314	38845931	152618
PDGFRA[1]	chr4	54243819	55164412	920594
PGPEP1L	chr15	99511458	99551024	39567
PRSS2	chr7	142479907	142481378	1472
PRSS3P2	chr7	142478756	142482399	3644
PRUNE	chr1	150980972	151008189	27218
RMND5A	chr2	86947413	88038768	1091356
RNF126	chr19	647525	663233	15709
RUNX1[1]	chr21	36160097	37357047	1196951
RUNX2	chr6	45296053	45518819	222767
SCAPER	chr15	76640526	77197744	557219
SDK1	chr7	3341079	4308631	967553
SETDB1	chr1	150898814	150937220	38407
SIRPB1	chr20	1545028	1600689	55662
SNED1	chr2	241938254	242033643	95390
SYNM	chr15	99645285	99675800	30516
TCRBV2S1	chr7	142334185	142494579	160395
TCRVB	chr7	142353890	142500213	146324
TP53[1]	chr17	7565096	7590868	25773
TP63	chr3	189349215	189615068	265854
TRIM22	chr11	5710816	5821759	110944
TRIM5	chr11	5684424	5959849	275426
TSFM	chr12	58176527	58196639	20113
TSPAN31	chr12	58138783	58142026	3244
TTC23	chr15	99676527	99791431	114905
TTTY9A	chrY	20891767	20901083	9317
UGT2B15	chr4	69512314	69536494	24181
WNK1	chr12	862088	1020618	158531
WWOX	chr16	78133309	79246564	1113256

[1] Genes with gene symbols in bold are listed in Cancer Gene Census of COSMIC.

References

[1] A. J. Iafrate, L. Feuk, M. N. Rivera, M. L. Listewnik, P. K. Donahoe, Y. Qi, S. W. Scherer, and C. Lee. **Detection of large-scale variation in the human genome**. *Nat Genet*, **36**(9):949–51, 2004. 1

[2] R. Khaja, J. Zhang, J. R. MacDonald, Y. He, A. M. Joseph-George, J. Wei, M. A. Rafiq, C. Qian, M. Shago, L. Pantano, H. Aburatani, K. Jones, R. Redon, M. Hurles, L. Armengol, X. Estivill, R. J. Mural, C. Lee, S. W. Scherer, and L. Feuk. **Genome assembly comparison identifies structural variants in the human genome**. *Nat Genet*, **38**(12):1413–8, 2006. 1

[3] R. Redon, S. Ishikawa, K. R. Fitch, L. Feuk, G. H. Perry, T. D. Andrews, H. Fiegler, M. H. Shapero, A. R. Carson, W. Chen, E. K. Cho, S. Dallaire, J. L. Freeman, J. R. Gonzalez, M. Gratacos, J. Huang, D. Kalaitzopoulos, D. Komura, J. R. MacDonald, C. R. Marshall, R. Mei, L. Montgomery, K. Nishimura, K. Okamura, F. Shen, M. J. Somerville, J. Tchinda, A. Valsesia, C. Woodwark, F. Yang, J. Zhang, T. Zerjal, J. Zhang, L. Armengol, D. F. Conrad, X. Estivill, C. Tyler-Smith, N. P. Carter, H. Aburatani, C. Lee, K. W. Jones, S. W. Scherer, and M. E. Hurles. **Global variation in copy number in the human genome**. *Nature*, **444**(7118):444–54, 2006. 1

[4] K. K. Wong, R. J. deLeeuw, N. S. Dosanjh, L. R. Kimm, Z. Cheng, D. E. Horsman, C. MacAulay, R. T. Ng, C. J. Brown, E. E. Eichler, and W. L. Lam. **A comprehensive analysis of common copy-number variations in the human genome**. *Am J Hum Genet*, **80**(1):91–104, 2007. 1

[5] C. E. Bruder, A. Piotrowski, A. A. Gijsbers, R. Andersson, S. Erickson, T. Diaz de Stahl, U. Menzel, J. Sandgren, D. von Tell, A. Poplawski, M. Crowley, C. Crasto, E. C. Partridge, H. Tiwari, D. B. Allison, J. Komorowski, G. J. van Ommen, D. I. Boomsma, N. L. Pedersen, J. T. den Dunnen, K. Wirdefeldt, and J. P. Dumanski. **Phenotypically concordant and discordant monozygotic twins display different DNA copy-number-variation profiles**. *Am J Hum Genet*, **82**(3):763–71, 2008. 1

[6] A. Piotrowski, C. E. Bruder, R. Andersson, T. Diaz de Stahl, U. Menzel, J. Sandgren, A. Poplawski, D. von Tell, C. Crasto, A. Bogdan, R. Bartoszewski, Z. Bebok, M. Krzyzanowski, Z. Jankowski, E. C. Partridge, J. Komorowski, and J. P. Dumanski. **Somatic mosaicism for copy number variation in differentiated human tissues**. *Hum Mutat*, **29**(9):1118–24, 2008. 1

[7] Network Cancer Genome Atlas. **Comprehensive molecular portraits of human breast tumours**. *Nature*, **490**(7418):61–70, 2012. 1

[8] Network Cancer Genome Atlas Research. **Comprehensive genomic characterization of squamous cell lung cancers**. *Nature*, **489**(7417):519–25, 2012. 1

[9] P. J. Stephens, D. J. McBride, M. L. Lin, I. Varela, E. D. Pleasance, J. T. Simpson, L. A. Stebbings, C. Leroy, S. Edkins, L. J. Mudie, C. D. Greenman, M. Jia, C. Latimer, J. W. Teague, K. W. Lau, J. Burton, M. A. Quail, H. Swerdlow, C. Churcher, R. Natrajan, A. M. Sieuwerts, J. W. Martens, D. P. Silver, A. Langerod, H. E. Russnes, J. A. Foekens, J. S. Reis-Filho, L. van 't Veer, A. L. Richardson, A. L. Borresen-Dale, P. J. Campbell, P. A. Futreal, and M. R. Stratton. **Complex landscapes of somatic rearrangement in human breast cancer genomes**. *Nature*, **462**(7276):1005–10, 2009. 1

[10] B. A. Weir, M. S. Woo, G. Getz, S. Perner, L. Ding, R. Beroukhim, W. M. Lin, M. A. Province, A. Kraja, L. A. Johnson, K. Shah, M. Sato, R. K. Thomas, J. A. Barletta, I. B. Borecki, S. Broderick, A. C. Chang, D. Y.

Chiang, L. R. Chirieac, J. Cho, Y. Fujii, A. F. Gazdar, T. Giordano, H. Greulich, M. Hanna, B. E. Johnson, M. G. Kris, A. Lash, L. Lin, N. Lindeman, E. R. Mardis, J. D. McPherson, J. D. Minna, M. B. Morgan, M. Nadel, M. B. Orringer, J. R. Osborne, B. Ozenberger, A. H. Ramos, J. Robinson, J. A. Roth, V. Rusch, H. Sasaki, F. Shepherd, C. Sougnez, M. R. Spitz, M. S. Tsao, D. Twomey, R. G. Verhaak, G. M. Weinstock, D. A. Wheeler, W. Winckler, A. Yoshizawa, S. Yu, M. F. Zakowski, Q. Zhang, D. G. Beer, II Wistuba, M. A. Watson, L. A. Garraway, M. Ladanyi, W. D. Travis, W. Pao, M. A. Rubin, S. B. Gabriel, R. A. Gibbs, H. E. Varmus, R. K. Wilson, E. S. Lander, and M. Meyerson. **Characterizing the cancer genome in lung adenocarcinoma**. *Nature*, **450**(7171):893–8, 2007. 1

[11] B. Juliandi, M. Abematsu, and K. Nakashima. **Epigenetic regulation in neural stem cell differentiation**. *Dev Growth Differ*, **52**(6):493–504, 2010. 1

[12] D. F. Conrad, D. Pinto, R. Redon, L. Feuk, O. Gokcumen, Y. Zhang, J. Aerts, T. D. Andrews, C. Barnes, P. Campbell, T. Fitzgerald, M. Hu, C. H. Ihm, K. Kristiansson, D. G. Macarthur, J. R. Macdonald, I. Onyiah, A. W. Pang, S. Robson, K. Stirrups, A. Valsesia, K. Walter, J. Wei, Consortium Wellcome Trust Case Control, C. Tyler-Smith, N. P. Carter, C. Lee, S. W. Scherer, and M. E. Hurles. **Origins and functional impact of copy number variation in the human genome**. *Nature*, **464**(7289):704–12, 2010. 1

[13] F. Zhang, W. Gu, M. E. Hurles, and J. R. Lupski. **Copy number variation in human health, disease, and evolution**. *Annu Rev Genomics Hum Genet*, **10**:451–81, 2009. 1

[14] G. H. Perry, J. Tchinda, S. D. McGrath, J. Zhang, S. R. Picker, A. M. Caceres, A. J. Iafrate, C. Tyler-Smith, S. W. Scherer, E. E. Eichler, A. C. Stone, and C. Lee. **Hotspots for copy number variation in chimpanzees and humans**. *Proc Natl Acad Sci U S A*, **103**(21):8006–11, 2006. 1

[15] J. Sebat, B. Lakshmi, J. Troge, J. Alexander, J. Young, P. Lundin, S. Maner, H. Massa, M. Walker, M. Chi, N. Navin, R. Lucito, J. Healy, J. Hicks, K. Ye, A. Reiner, T. C. Gilliam, B. Trask, N. Patterson, A. Zetterberg, and M. Wigler. **Large-scale copy number polymorphism in the human genome**. *Science*, **305**(5683):525–8, 2004. 2

[16] B. E. Stranger, M. S. Forrest, M. Dunning, C. E. Ingle, C. Beazley, N. Thorne, R. Redon, C. P. Bird, A. de Grassi, C. Lee, C. Tyler-Smith, N. Carter, S. W. Scherer, S. Tavare, P. Deloukas, M. E. Hurles, and E. T. Dermitzakis. **Relative impact of nucleotide and copy number variation on gene expression phenotypes**. *Science*, **315**(5813):848–53, 2007. 2

[17] G. H. Perry, N. J. Dominy, K. G. Claw, A. S. Lee, H. Fiegler, R. Redon, J. Werner, F. A. Villanea, J. L. Mountain, R. Misra, N. P. Carter, C. Lee, and A. C. Stone. **Diet and the evolution of human amylase gene copy number variation**. *Nat Genet*, **39**(10):1256–60, 2007. 2

[18] S. J. Diskin, C. Hou, J. T. Glessner, E. F. Attiyeh, M. Laudenslager, K. Bosse, K. Cole, Y. P. Mosse, A. Wood, J. E. Lynch, K. Pecor, M. Diamond, C. Winter, K. Wang, C. Kim, E. A. Geiger, P. W. McGrady, A. I. Blakemore, W. B. London, T. H. Shaikh, J. Bradfield, S. F. Grant, H. Li, M. Devoto, E. R. Rappaport, H. Hakonarson, and J. M. Maris. **Copy number variation at 1q21.1 associated with neuroblastoma**. *Nature*, **459**(7249):987–91, 2009. 2

[19] J. Sebat, B. Lakshmi, D. Malhotra, J. Troge, C. Lese-Martin, T. Walsh, B. Yamrom, S. Yoon, A. Krasnitz, J. Kendall, A. Leotta, D. Pai, R. Zhang, Y. H. Lee, J. Hicks, S. J. Spence, A. T. Lee, K. Puura, T. Lehtimaki, D. Ledbetter, P. K. Gregersen, J. Bregman, J. S. Sutcliffe, V. Jobanputra, W. Chung, D. Warburton, M. C. King, D. Skuse, D. H. Geschwind, T. C. Gilliam, K. Ye, and M. Wigler. **Strong association of de novo copy number mutations with autism**. *Science*, **316**(5823):445–9, 2007. 2

[20] E. Gonzalez, H. Kulkarni, H. Bolivar, A. Mangano, R. Sanchez, G. Catano, R. J. Nibbs, B. I. Freedman, M. P.

Quinones, M. J. Bamshad, K. K. Murthy, B. H. Rovin, W. Bradley, R. A. Clark, S. A. Anderson, J. O'Connell R, B. K. Agan, S. S. Ahuja, R. Bologna, L. Sen, M. J. Dolan, and S. K. Ahuja. **The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility**. *Science*, **307**(5714):1434–40, 2005. 2

[21] C. Curtis, S. P. Shah, S. F. Chin, G. Turashvili, O. M. Rueda, M. J. Dunning, D. Speed, A. G. Lynch, S. Samarajiwa, Y. Yuan, S. Graf, G. Ha, G. Haffari, A. Bashashati, R. Russell, S. McKinney, Metabric Group, A. Langerod, A. Green, E. Provenzano, G. Wishart, S. Pinder, P. Watson, F. Markowetz, L. Murphy, I. Ellis, A. Purushotham, A. L. Borresen-Dale, J. D. Brenton, S. Tavare, C. Caldas, and S. Aparicio. **The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups**. *Nature*, **486**(7403):346–52, 2012. 2

[22] T. Santarius, J. Shipley, D. Brewer, M. R. Stratton, and C. S. Cooper. **A census of amplified and overexpressed human cancer genes**. *Nat Rev Cancer*, **10**(1):59–64, 2010. 2

[23] R. Beroukhim, G. Getz, L. Nghiemphu, J. Barretina, T. Hsueh, D. Linhart, I. Vivanco, J. C. Lee, J. H. Huang, S. Alexander, J. Du, T. Kau, R. K. Thomas, K. Shah, H. Soto, S. Perner, J. Prensner, R. M. Debiasi, F. Demichelis, C. Hatton, M. A. Rubin, L. A. Garraway, S. F. Nelson, L. Liau, P. S. Mischel, T. F. Cloughesy, M. Meyerson, T. A. Golub, E. S. Lander, I. K. Mellinghoff, and W. R. Sellers. **Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma**. *Proc Natl Acad Sci U S A*, **104**(50):20007–12, 2007. 2

[24] M. Cardoso-Moreira, J. R. Arguello, and A. G. Clark. **Mutation spectrum of Drosophila CNVs revealed by breakpoint sequencing**. *Genome Biol*, **13**(12):R119, 2012. 2

[25] P. J. Hastings, J. R. Lupski, S. M. Rosenberg, and G. Ira. **Mechanisms of change in gene copy number**. *Nat Rev Genet*, **10**(8):551–64, 2009. 2, 3, 25

[26] P. Stankiewicz and J. R. Lupski. **Genome architecture, rearrangements and genomic disorders**. *Trends Genet*, **18**(2):74–82, 2002. 2

[27] J R Lupski. **Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits**. *Trends Genet*, **14**(10):417–22, 1998. 3

[28] Beth Elliott, Christine Richardson, and Maria Jasin. **Chromosomal translocation mechanisms at intronic alu elements in mammalian cells**. *Mol Cell*, **17**(6):885–94, 2005. 3

[29] M. McVey and S. E. Lee. **MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings**. *Trends Genet*, **24**(11):529–38, 2008. 3

[30] C. J. Shaw and J. R. Lupski. **Implications of human genome architecture for rearrangement-based disorders: the genomic basis of disease**. *Hum Mol Genet*, **13 Spec No 1**:R57–64, 2004. 3

[31] P. Stankiewicz, C. J. Shaw, J. D. Dapper, K. Wakui, L. G. Shaffer, M. Withers, L. Elizondo, S. S. Park, and J. R. Lupski. **Genome architecture catalyzes nonrecurrent chromosomal rearrangements**. *Am J Hum Genet*, **72**(5):1101–16, 2003. 3

[32] A. R. Quinlan and I. M. Hall. **Characterizing complex structural variation in germline and somatic genomes**. *Trends Genet*, **28**(1):43–53, 2012. 3

[33] F. Zhang, C. M. Carvalho, and J. R. Lupski. **Complex human chromosomal and genomic rearrangements**. *Trends Genet*, **25**(7):298–307, 2009. 3

[34] J. A. Lee, C. M. Carvalho, and J. R. Lupski. **A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders**. *Cell*, **131**(7):1235–47, 2007. 3

[35] P. J. Hastings, G. Ira, and J. R. Lupski. **A microhomology-mediated break-induced replication model for the origin of human copy number variation**. *PLoS Genet*, **5**(1):e1000327, 2009. 3

[36] J. M. Chen, N. Chuzhanova, P. D. Stenson, C. Ferec, and D. N. Cooper. **Complex gene rearrangements caused by serial replication slippage**. *Hum Mutat*, **26**(2):125–34, 2005. 3

[37] Christopher Yau, Dmitri Mouradov, Robert Jorissen, Stefano Colella, Ghazala Mirza, Graham Steers, Adrian Harris, Jiannis Ragoussis, Oliver Sieber, and Christopher Holmes. **A statistical approach for detecting genomic aberrations in heterogeneous tumor samples from single nucleotide polymorphism genotyping data**. *Genome Biology*, **11**(9):R92, 2010. 4

[38] Wei Sun, Fred Wright, Zhengzheng Tang, Silje Nordgard, Peter Van Loo, Tianwei Yu, Vessela Kristensen, and Charles Perou. **Integrated study of copy number states and genotype calls using high-density SNP arrays**. *Nucleic acids research*, **37**(16):5365–5377, 2009. 4

[39] Ao Li, Zongzhi Liu, Kimberly Lezon-Geyda, Sudipa Sarkar, Donald Lannin, Vincent Schulz, Ian Krop, Eric Winer, Lyndsay Harris, and David Tuck. **GPHMM: an integrated hidden Markov model for identification of copy number alteration and loss of heterozygosity in complex tumor samples using whole genome SNP arrays**. *Nucleic Acids Res*, **39**(12):4928–41, 2011. 4, 57

[40] Zongzhi Liu, Ao Li, Vincent Schulz, Min Chen, and David Tuck. **MixHMM: Inferring Copy Number Variation and Allelic Imbalance Using SNP Arrays and Tumor Samples Mixed with Stromal Cells**. *PLoS ONE*, **5**(6):e10909, 2010. 4

[41] H. Goransson, K. Edlund, M. Rydaker, M. Rasmussen, J. Winquist, S. Ekman, M. Bergqvist, A. Thomas, M. Lambe, R. Rosenquist, L. Holmberg, P. Micke, J. Botling, and A. Isaksson. **Quantification of normal cell fraction and copy number neutral LOH in clinical lung cancer samples using SNP array data**. *PLoS One*, **4**(6):e6057, 2009. 4

[42] H. Chen, H. Xing, and N. R. Zhang. **Estimation of parent specific DNA copy number in tumors using high-density genotyping arrays**. *PLoS Comput Biol*, **7**(1):e1001060, 2011. 4

[43] H. Bengtsson, P. Neuvial, and T. P. Speed. **TumorBoost: normalization of allele-specific tumor copy numbers from a single pair of tumor-normal genotyping microarrays**. *BMC Bioinformatics*, **11**:245, 2010. 4

[44] P. Van Loo, S. H. Nordgard, O. C. Lingjaerde, H. G. Russnes, I. H. Rye, W. Sun, V. J. Weigman, P. Marynen, A. Zetterberg, B. Naume, C. M. Perou, A. L. Borresen-Dale, and V. N. Kristensen. **Allele-specific copy number analysis of tumors**. *Proc Natl Acad Sci U S A*, **107**(39):16910–5, 2010. 4, 53

[45] Tatiana Popova, Elodie Manié, Dominique Stoppa-Lyonnet, Guillem Rigaill, Emmanuel Barillot, and Marc Henri Stern. **Genome Alteration Print (GAP): a tool to visualize and mine complex cancer genomic profiles obtained by SNP arrays**. *Genome biology*, **10**(11):R128, 2009. 4

[46] P. J. Stephens, C. D. Greenman, B. Fu, F. Yang, G. R. Bignell, L. J. Mudie, E. D. Pleasance, K. W. Lau, D. Beare, L. A. Stebbings, S. McLaren, M. L. Lin, D. J. McBride, I. Varela, S. Nik-Zainal, C. Leroy, M. Jia, A. Menzies, A. P. Butler, J. W. Teague, M. A. Quail, J. Burton, H. Swerdlow, N. P. Carter, L. A. Morsberger, C. Iacobuzio-Donahue, G. A. Follows, A. R. Green, A. M. Flanagan, M. R. Stratton, P. A. Futreal, and P. J. Campbell. **Massive genomic rearrangement acquired in a single catastrophic event during cancer development**. *Cell*, **144**(1):27–40, 2011. 5, 6, 54, 66, 72

[47] H. Cai, N. Kumar, H. C. Bagheri, C. von Mering, M. D. Robinson, and M. Baudis. **Chromothripsis-like patterns are recurring but heterogeneously distributed features in a survey of 22,347 cancer genome screens**. *BMC Genomics*, **15**:82, 2014. 5, 57

[48] Tae-Min Kim, Ruibin Xi, Lovelace J Luquette, Richard W Park, Mark D Johnson, and Peter J Park. **Functional genomic analysis of chromosomal aberrations in a compendium of 8000 cancer genomes**. *Genome Res*, **23**(2):217–27, 2013. 5

[49] A. Malhotra, M. Lindberg, G. G. Faust, M. L. Leibowitz, R. A. Clark, R. M. Layer, A. R. Quinlan, and I. M. Hall. **Breakpoint profiling of 64 cancer genomes reveals numerous complex rearrangements spawned by homology-independent mechanisms**. *Genome Res*, **23**(5):762–76, 2013. 5

[50] A. M. Patch, E. L. Christie, D. Etemadmoghadam, D. W. Garsed, J. George, S. Fereday, K. Nones, P. Cowin, K. Alsop, P. J. Bailey, K. S. Kassahn, F. Newell, M. C. Quinn, S. Kazakoff, K. Quek, C. Wilhelm-Benartzi, E. Curry, H. S. Leong, Group Australian Ovarian Cancer Study, A. Hamilton, L. Mileshkin, G. Au-Yeung, C. Kennedy, J. Hung, Y. E. Chiew, P. Harnett, M. Friedlander, M. Quinn, J. Pyman, S. Cordner, P. O'Brien, J. Leditschke, G. Young, K. Strachan, P. Waring, W. Azar, C. Mitchell, N. Traficante, J. Hendley, H. Thorne, M. Shackleton, D. K. Miller, G. M. Arnau, R. W. Tothill, T. P. Holloway, T. Semple, I. Harliwong, C. Nourse, E. Nourbakhsh, S. Manning, S. Idrisoglu, T. J. Bruxner, A. N. Christ, B. Poudel, O. Holmes, M. Anderson, C. Leonard, A. Lonie, N. Hall, S. Wood, D. F. Taylor, Q. Xu, J. L. Fink, N. Waddell, R. Drapkin, E. Stronach, H. Gabra, R. Brown, A. Jewell, S. H. Nagaraj, E. Markham, P. J. Wilson, J. Ellul, O. McNally, M. A. Doyle, R. Vedururu, C. Stewart, E. Lengyel, J. V. Pearson, N. Waddell, A. deFazio, S. M. Grimmond, and D. D. Bowtell. **Whole-genome characterization of chemoresistant ovarian cancer**. *Nature*, **521**(7553):489–94, 2015. 5

[51] N. Waddell, M. Pajic, A. M. Patch, D. K. Chang, K. S. Kassahn, P. Bailey, A. L. Johns, D. Miller, K. Nones, K. Quek, M. C. Quinn, A. J. Robertson, M. Z. Fadlullah, T. J. Bruxner, A. N. Christ, I. Harliwong, S. Idrisoglu, S. Manning, C. Nourse, E. Nourbakhsh, S. Wani, P. J. Wilson, E. Markham, N. Cloonan, M. J. Anderson, J. L. Fink, O. Holmes, S. H. Kazakoff, C. Leonard, F. Newell, B. Poudel, S. Song, D. Taylor, N. Waddell, S. Wood, Q. Xu, J. Wu, M. Pinese, M. J. Cowley, H. C. Lee, M. D. Jones, A. M. Nagrial, J. Humphris, L. A. Chantrill, V. Chin, A. M. Steinmann, A. Mawson, E. S. Humphrey,

E. K. Colvin, A. Chou, C. J. Scarlett, A. V. Pinho, M. Giry-Laterriere, I. Rooman, J. S. Samra, J. G. Kench, J. A. Pettitt, N. D. Merrett, C. Toon, K. Epari, N. Q. Nguyen, A. Barbour, N. Zeps, N. B. Jamieson, J. S. Graham, S. P. Niclou, R. Bjerkvig, R. Grutzmann, D. Aust, R. H. Hruban, A. Maitra, C. A. Iacobuzio-Donahue, C. L. Wolfgang, R. A. Morgan, R. T. Lawlor, V. Corbo, C. Bassi, M. Falconi, G. Zamboni, G. Tortora, M. A. Tempero, Initiative Australian Pancreatic Cancer Genome, A. J. Gill, J. R. Eshleman, C. Pilarsky, A. Scarpa, E. A. Musgrove, J. V. Pearson, A. V. Biankin, and S. M. Grimmond. **Whole genomes redefine the mutational landscape of pancreatic cancer**. *Nature*, **518**(7540):495–501, 2015. 5

[52] D. W. Garsed, O. J. Marshall, V. D. Corbin, A. Hsu, L. Di Stefano, J. Schroder, J. Li, Z. P. Feng, B. W. Kim, M. Kowarsky, B. Lansdell, R. Brookwell, O. Myklebost, L. Meza-Zepeda, A. J. Holloway, F. Pedeutour, K. H. Choo, M. A. Damore, A. J. Deans, A. T. Papenfuss, and D. M. Thomas. **The architecture and evolution of cancer neochromosomes**. *Cancer Cell*, **26**(5):653–67, 2014. 5

[53] T. Rausch, D. T. Jones, M. Zapatka, A. M. Stutz, T. Zichner, J. Weischenfeldt, N. Jager, M. Remke, D. Shih, P. A. Northcott, E. Pfaff, J. Tica, Q. Wang, L. Massimi, H. Witt, S. Bender, S. Pleier, H. Cin, C. Hawkins, C. Beck, A. von Deimling, V. Hans, B. Brors, R. Eils, W. Scheurlen, J. Blake, V. Benes, A. E. Kulozik, O. Witt, D. Martin, C. Zhang, R. Porat, D. M. Merino, J. Wasserman, N. Jabado, A. Fontebasso, L. Bullinger, F. G. Rucker, K. Dohner, H. Dohner, J. Koster, J. J. Molenaar, R. Versteeg, M. Kool, U. Tabori, D. Malkin, A. Korshunov, M. D. Taylor, P. Lichter, S. M. Pfister, and J. O. Korbel. **Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations**. *Cell*, **148**(1-2):59–71, 2012. 5, 54, 72

[54] J. Z. Sanborn, S. R. Salama, M. Grifford, C. W. Brennan, T. Mikkelsen, S. Jhanwar, S. Katzman, L. Chin, and D. Haussler. **Double minute chromosomes in glioblastoma**

multiforme are revealed by precise reconstruction of oncogenic amplicons. *Cancer Res*, **73**(19):6036–45, 2013. 5

[55] T. I. Zack, S. E. Schumacher, S. L. Carter, A. D. Cherniack, G. Saksena, B. Tabak, M. S. Lawrence, C. Z. Zhsng, J. Wala, C. H. Mermel, C. Sougnez, S. B. Gabriel, B. Hernandez, H. Shen, P. W. Laird, G. Getz, M. Meyerson, and R. Beroukhim. **Pan-cancer patterns of somatic copy number alteration**. *Nat Genet*, **45**(10):1134–40, 2013. 5, 24, 25, 27, 34, 41, 47, 53

[56] G. King and L. Zeng. **Logistic Regression in Rare Events Data**. *Polit. Anal.*, **9**:137–163, 2001. 8, 33

[57] L. Breiman. **Random forests**. *Machine Learning*, **45**(1):5–32, 2001. 8

[58] Forman George. **An extensive empirical study of feature selection metrics for text classification**. *Journal of Machine Learning Research*, **3**:1289–1305, 2003. 10

[59] M. L. Metzker. **APPLICATIONS OF NEXT-GENERATION SEQUENCING Sequencing technologies - the next generation**. *Nature Reviews Genetics*, **11**(1):31–46, 2010. 11

[60] J. Shendure and H. L. Ji. **Next-generation DNA sequencing**. *Nature Biotechnology*, **26**(10):1135–1145, 2008. 11

[61] D. A. Wheeler, M. Srinivasan, M. Egholm, Y. Shen, L. Chen, A. McGuire, W. He, Y. J. Chen, V. Makhijani, G. T. Roth, X. Gomes, K. Tartaro, F. Niazi, C. L. Turcotte, G. P. Irzyk, J. R. Lupski, C. Chinault, X. Z. Song, Y. Liu, Y. Yuan, L. Nazareth, X. Qin, D. M. Muzny, M. Margulies, G. M. Weinstock, R. A. Gibbs, and J. M. Rothberg. **The complete genome of an individual by massively parallel DNA sequencing**. *Nature*, **452**(7189):872–U5, 2008. 11

[62] D. R. Bentley, S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, C. G. Brown, K. P. Hall, D. J. Evers, C. L. Barnes, H. R. Bignell, J. M. Boutell, J. Bryant, R. J. Carter, R. K. Cheetham, A. J. Cox, D. J. Ellis, M. R.

Flatbush, N. A. Gormley, S. J. Humphray, L. J. Irving, M. S. Karbelashvili, S. M. Kirk, H. Li, X. H. Liu, K. S. Maisinger, L. J. Murray, B. Obradovic, T. Ost, M. L. Parkinson, M. R. Pratt, I. M. J. Rasolonjatovo, M. T. Reed, R. Rigatti, C. Rodighiero, M. T. Ross, A. Sabot, S. V. Sankar, A. Scally, G. P. Schroth, M. E. Smith, V. P. Smith, A. Spiridou, P. E. Torrance, S. S. Tzonev, E. H. Vermaas, K. Walter, X. L. Wu, L. Zhang, M. D. Alam, C. Anastasi, I. C. Aniebo, D. M. D. Bailey, I. R. Bancarz, S. Banerjee, S. G. Barbour, P. A. Baybayan, V. A. Benoit, K. F. Benson, C. Bevis, P. J. Black, A. Boodhun, J. S. Brennan, J. A. Bridgham, R. C. Brown, A. A. Brown, D. H. Buermann, A. A. Bundu, J. C. Burrows, N. P. Carter, N. Castillo, M. C. E. Catenazzi, S. Chang, R. N. Cooley, N. R. Crake, O. O. Dada, K. D. Diakoumakos, B. Dominguez-Fernandez, D. J. Earnshaw, U. C. Egbujor, D. W. Elmore, S. S. Etchin, M. R. Ewan, M. Fedurco, L. J. Fraser, K. V. F. Fajardo, W. S. Furey, D. George, K. J. Gietzen, C. P. Goddard, G. S. Golda, P. A. Granieri, D. E. Green, D. L. Gustafson, N. F. Hansen, K. Harnish, C. D. Haudenschild, N. I. Heyer, M. M. Hims, J. T. Ho, A. M. Horgan, et al. **Accurate whole human genome sequencing using reversible terminator chemistry**. *Nature*, **456**(7218):53–59, 2008. 11

[63] K. J. McKernan, H. E. Peckham, G. L. Costa, S. F. McLaughlin, Y. T. Fu, E. F. Tsung, C. R. Clouser, C. Duncan, J. K. Ichikawa, C. C. Lee, Z. Zhang, S. S. Ranade, E. T. Dimalanta, F. C. Hyland, T. D. Sokolsky, L. Zhang, A. Sheridan, H. N. Fu, C. L. Hendrickson, B. Li, L. Kotler, J. R. Stuart, J. A. Malek, J. M. Manning, A. A. Antipova, D. S. Perez, M. P. Moore, K. C. Hayashibara, M. R. Lyons, R. E. Beaudoin, B. E. Coleman, M. W. Laptewicz, A. E. Sannicandro, M. D. Rhodes, R. K. Gottimukkala, S. Yang, V. Bafna, A. Bashir, A. MacBride, C. Alkan, J. M. Kidd, E. E. Eichler, M. G. Reese, F. M. De la Vega, and A. P. Blanchard. **Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding**. *Genome Research*,

19(9):1527–1541, 2009. 11

[64] B. B. Tuch, R. R. Laborde, X. Xu, J. Gu, C. B. Chung, C. K. Monighetti, S. J. Stanley, K. D. Olsen, J. L. Kasperbauer, E. J. Moore, A. J. Broomer, R. Tan, P. M. Brzoska, M. W. Muller, A. S. Siddiqui, Y. W. Asmann, Y. Sun, S. Kuersten, M. A. Barker, F. M. De La Vega, and D. I. Smith. **Tumor transcriptome sequencing reveals allelic expression imbalances associated with copy number alterations**. *PLoS One*, **5**(2):e9317, 2010. 13

[65] D. D. Jima, J. Zhang, C. Jacobs, K. L. Richards, C. H. Dunphy, W. W. Choi, W. Y. Au, G. Srivastava, M. B. Czader, D. A. Rizzieri, A. S. Lagoo, P. L. Lugar, K. P. Mann, C. R. Flowers, L. Bernal-Mizrachi, K. N. Naresh, A. M. Evens, L. I. Gordon, M. Luftig, D. R. Friedman, J. B. Weinberg, M. A. Thompson, J. I. Gill, Q. Liu, T. How, V. Grubor, Y. Gao, A. Patel, H. Wu, J. Zhu, G. C. Blobe, P. E. Lipsky, A. Chadburn, S. S. Dave, and Consortium Hematologic Malignancies Research. **Deep sequencing of the small RNA transcriptome of normal and malignant human B cells identifies hundreds of novel microRNAs**. *Blood*, **116**(23):e118–27, 2010. 13

[66] M. F. Berger, J. Z. Levin, K. Vijayendran, A. Sivachenko, X. Adiconis, J. Maguire, L. A. Johnson, J. Robinson, R. G. Verhaak, C. Sougnez, R. C. Onofrio, L. Ziaugra, K. Cibulskis, E. Laine, J. Barretina, W. Winckler, D. E. Fisher, G. Getz, M. Meyerson, D. B. Jaffe, S. B. Gabriel, E. S. Lander, R. Dummer, A. Gnirke, C. Nusbaum, and L. A. Garraway. **Integrative analysis of the melanoma transcriptome**. *Genome Res*, **20**(4):413–27, 2010. 13

[67] L. Conde, P. M. Bracci, R. Richardson, S. B. Montgomery, and C. F. Skibola. **Integrating GWAS and expression data for functional characterization of disease-associated SNPs: an application to follicular lymphoma**. *Am J Hum Genet*, **92**(1):126–30, 2013. 13

[68] C. A. Maher, C. Kumar-Sinha, X. Cao, S. Kalyana-Sundaram, B. Han, X. Jing, L. Sam, T. Barrette, N. Palanisamy, and A. M. Chinnaiyan. **Transcriptome**

sequencing to detect gene fusions in cancer. *Nature*, **458**(7234):97–101, 2009. 13

[69] J. Supper, C. Gugenmus, J. Wollnik, T. Drueke, M. Scherf, A. Hahn, K. Grote, N. Bretschneider, B. Klocke, C. Zinser, K. Cartharius, and M. Seifert. **Detecting and visualizing gene fusions**. *Methods*, **59**(1):S24–8, 2013. 13

[70] K. R. Chng, C. W. Chang, S. K. Tan, C. Yang, S. Z. Hong, N. Y. Sng, and E. Cheung. **A transcriptional repressor co-regulatory network governing androgen response in prostate cancers**. *EMBO J*, **31**(12):2810–23, 2012. 13

[71] T. Lu, M. Yang, D. B. Huang, H. Wei, G. H. Ozer, G. Ghosh, and G. R. Stark. **Role of lysine methylation of NF-kappaB in differential gene regulation**. *Proc Natl Acad Sci U S A*, **110**(33):13510–5, 2013. 13

[72] R. Lister, M. Pelizzola, R. H. Dowen, R. D. Hawkins, G. Hon, J. Tonti-Filippini, J. R. Nery, L. Lee, Z. Ye, Q. M. Ngo, L. Edsall, J. Antosiewicz-Bourget, R. Stewart, V. Ruotti, A. H. Millar, J. A. Thomson, B. Ren, and J. R. Ecker. **Human DNA methylomes at base resolution show widespread epigenomic differences**. *Nature*, **462**(7271):315–22, 2009. 14, 16, 18, 20

[73] W. A. Pastor, L. Aravind, and A. Rao. **TETonic shift: biological roles of TET proteins in DNA demethylation and transcription**. *Nat Rev Mol Cell Biol*, **14**(6):341–56, 2013. 14

[74] S. Tomizawa, H. Kobayashi, T. Watanabe, S. Andrews, K. Hata, G. Kelsey, and H. Sasaki. **Dynamic stage-specific changes in imprinted differentially methylated regions during early mammalian development and prevalence of non-CpG methylation in oocytes**. *Development*, **138**(5):811–20, 2011. 15

[75] M. J. Ziller, F. Muller, J. Liao, Y. Zhang, H. Gu, C. Bock, P. Boyle, C. B. Epstein, B. E. Bernstein, T. Lengauer, A. Gnirke, and A. Meissner. **Genomic distribution and inter-sample variation of non-CpG methylation across human cell types**. *PLoS Genet*, **7**(12):e1002389, 2011. 15

[76] A. P. Bird. **DNA methylation and the frequency of CpG in animal DNA**. *Nucleic Acids Res*, **8**(7):1499–504, 1980. 15

[77] R. S. Illingworth and A. P. Bird. **CpG islands–'a rough guide'**. *FEBS Lett*, **583**(11):1713–20, 2009. 15

[78] R. S. Illingworth, U. Gruenewald-Schneider, S. Webb, A. R. Kerr, K. D. James, D. J. Turner, C. Smith, D. J. Harrison, R. Andrews, and A. P. Bird. **Orphan CpG islands identify numerous conserved promoters in the mammalian genome**. *PLoS Genet*, **6**(9):e1001134, 2010. 15

[79] A. M. Deaton and A. Bird. **CpG islands and the regulation of transcription**. *Genes Dev*, **25**(10):1010–22, 2011. 15

[80] A. Meissner, T. S. Mikkelsen, H. Gu, M. Wernig, J. Hanna, A. Sivachenko, X. Zhang, B. E. Bernstein, C. Nusbaum, D. B. Jaffe, A. Gnirke, R. Jaenisch, and E. S. Lander. **Genome-scale DNA methylation maps of pluripotent and differentiated cells**. *Nature*, **454**(7205):766–70, 2008. 15

[81] M. M. Suzuki and A. Bird. **DNA methylation landscapes: provocative insights from epigenomics**. *Nat Rev Genet*, **9**(6):465–76, 2008. 15, 16, 18

[82] G. Auclair, S. Guibert, A. Bender, and M. Weber. **Ontogeny of CpG island methylation and specificity of DNMT3 methyltransferases during embryonic development in the mouse**. *Genome Biol*, **15**(12):545, 2014. 15

[83] R. Illingworth, A. Kerr, D. Desousa, H. Jorgensen, P. Ellis, J. Stalker, D. Jackson, C. Clee, R. Plumb, J. Rogers, S. Humphray, T. Cox, C. Langford, and A. Bird. **A novel CpG island set identifies tissue-specific methylation at developmental gene loci**. *PLoS Biol*, **6**(1):e22, 2008. 15

[84] A. K. Maunakea, R. P. Nagarajan, M. Bilenky, T. J. Ballinger, C. D'Souza, S. D. Fouse, B. E. Johnson, C. Hong, C. Nielsen, Y. Zhao, G. Turecki, A. Delaney, R. Varhol, N. Thiessen, K. Shchors, V. M. Heine, D. H. Rowitch, X. Xing, C. Fiore, M. Schillebeeckx, S. J. Jones, D. Haussler, M. A. Marra, M. Hirst, T. Wang, and J. F. Costello. **Conserved role of intragenic DNA methylation**

in regulating alternative promoters. *Nature*, **466**(7303):253–7, 2010. 15

[85] A. R. Krebs, S. Dessus-Babus, L. Burger, and D. Schubeler. **High-throughput engineering of a mammalian genome reveals building principles of methylation states at CG rich regions**. *Elife*, **3**:e04094, 2014. 15

[86] C. Marchal and B. Miotto. **Emerging concept in DNA methylation: role of transcription factors in shaping DNA methylation patterns**. *J Cell Physiol*, **230**(4):743–51, 2015. 15

[87] E. Wachter, T. Quante, C. Merusi, A. Arczewska, F. Stewart, S. Webb, and A. Bird. **Synthetic CpG islands reveal DNA sequence determinants of chromatin structure**. *Elife*, **3**:e03397, 2014. 15

[88] S. Guibert and M. Weber. **Functions of DNA methylation and hydroxymethylation in mammalian development**. *Curr Top Dev Biol*, **104**:47–83, 2013. 15

[89] R. J. Klose and A. P. Bird. **Genomic DNA methylation: the mark and its mediators**. *Trends Biochem Sci*, **31**(2):89–97, 2006. 15

[90] A. Eden, F. Gaudet, A. Waghmare, and R. Jaenisch. **Chromosomal instability and tumors promoted by DNA hypomethylation**. *Science*, **300**(5618):455, 2003. 15

[91] J. S. You and P. A. Jones. **Cancer genetics and epigenetics: two sides of the same coin?** *Cancer Cell*, **22**(1):9–20, 2012. 15

[92] J. Peters. **The role of genomic imprinting in biology and disease: an expanding view**. *Nat Rev Genet*, **15**(8):517–30, 2014. 15

[93] M. P. Ball, J. B. Li, Y. Gao, J. H. Lee, E. M. LeProust, I. H. Park, B. Xie, G. Q. Daley, and G. M. Church. **Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells**. *Nat Biotechnol*, **27**(4):361–8, 2009. 16

[94] X. Zhang, J. Yazaki, A. Sundaresan, S. Cokus, S. W. Chan, H. Chen, I. R. Henderson, P. Shinn, M. Pellegrini, S. E.

Jacobsen, and J. R. Ecker. **Genome-wide high-resolution mapping and functional analysis of DNA methylation in arabidopsis**. *Cell*, **126**(6):1189–201, 2006. 16

[95] G. L. Sen, J. A. Reuter, D. E. Webster, L. Zhu, and P. A. Khavari. **DNMT1 maintains progenitor function in self-renewing somatic tissue**. *Nature*, **463**(7280):563–7, 2010. 16

[96] Z. D. Smith and A. Meissner. **DNA methylation: roles in mammalian development**. *Nat Rev Genet*, **14**(3):204–20, 2013. 16

[97] D. Palacios, D. Summerbell, P. W. Rigby, and J. Boyes. **Interplay between DNA methylation and transcription factor availability: implications for developmental activation of the mouse Myogenin gene**. *Mol Cell Biol*, **30**(15):3805–15, 2010. 16

[98] K. Kim, A. Doi, B. Wen, K. Ng, R. Zhao, P. Cahan, J. Kim, M. J. Aryee, H. Ji, L. I. Ehrlich, A. Yabuuchi, A. Takeuchi, K. C. Cunniff, H. Hongguang, S. McKinney-Freeman, O. Naveiras, T. J. Yoon, R. A. Irizarry, N. Jung, J. Seita, J. Hanna, P. Murakami, R. Jaenisch, R. Weissleder, S. H. Orkin, I. L. Weissman, A. P. Feinberg, and G. Q. Daley. **Epigenetic memory in induced pluripotent stem cells**. *Nature*, **467**(7313):285–90, 2010. 16

[99] J. M. Polo, S. Liu, M. E. Figueroa, W. Kulalert, S. Eminli, K. Y. Tan, E. Apostolou, M. Stadtfeld, Y. Li, T. Shioda, S. Natesan, A. J. Wagers, A. Melnick, T. Evans, and K. Hochedlinger. **Cell type of origin influences the molecular and functional properties of mouse induced pluripotent stem cells**. *Nat Biotechnol*, **28**(8):848–55, 2010. 16

[100] C. S. Schmidt, S. Bultmann, D. Meilinger, B. Zacher, A. Tresch, K. C. Maier, C. Peter, D. E. Martin, H. Leonhardt, and F. Spada. **Global DNA hypomethylation prevents consolidation of differentiation programs and allows reversion to the embryonic stem cell state**. *PLoS One*, **7**(12):e52629, 2012. 16

[101] P. W. Laird. **Principles and challenges of genomewide DNA methylation analysis**. *Nat Rev Genet*, **11**(3):191–203, 2010. 16, 18

[102] J. Sandoval, H. Heyn, S. Moran, J. Serra-Musach, M. A. Pujana, M. Bibikova, and M. Esteller. **Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome**. *Epigenetics*, **6**(6):692–702, 2011. 16, 17

[103] D. J. Weisenberger. **Characterizing DNA methylation alterations from The Cancer Genome Atlas**. *J Clin Invest*, **124**(1):17–23, 2014. 16

[104] H. Heyn, N. Li, H. J. Ferreira, S. Moran, D. G. Pisano, A. Gomez, J. Diez, J. V. Sanchez-Mut, F. Setien, F. J. Carmona, A. A. Puca, S. Sayols, M. A. Pujana, J. Serra-Musach, I. Iglesias-Platas, F. Formiga, A. F. Fernandez, M. F. Fraga, S. C. Heath, A. Valencia, I. G. Gut, J. Wang, and M. Esteller. **Distinct DNA methylomes of newborns and centenarians**. *Proc Natl Acad Sci U S A*, **109**(26):10522–7, 2012. 17

[105] H. Heyn, S. Moran, I. Hernando-Herraez, S. Sayols, A. Gomez, J. Sandoval, D. Monk, K. Hata, T. Marques-Bonet, L. Wang, and M. Esteller. **DNA methylation contributes to natural human variation**. *Genome Res*, **23**(9):1363–72, 2013. 17

[106] M. Bibikova, B. Barnes, C. Tsan, V. Ho, B. Klotzle, J. M. Le, D. Delano, L. Zhang, G. P. Schroth, K. L. Gunderson, J. B. Fan, and R. Shen. **High density DNA methylation array with single CpG site resolution**. *Genomics*, **98**(4):288–95, 2011. 17

[107] S. Dedeurwaerder, M. Defrance, E. Calonne, H. Denis, C. Sotiriou, and F. Fuks. **Evaluation of the Infinium Methylation 450K technology**. *Epigenomics*, **3**(6):771–84, 2011. 17

[108] L. R. Meyer, A. S. Zweig, A. S. Hinrichs, D. Karolchik, R. M. Kuhn, M. Wong, C. A. Sloan, K. R. Rosenbloom, G. Roe, B. Rhead, B. J. Raney, A. Pohl, V. S. Malladi, C. H. Li, B. T. Lee, K. Learned, V. Kirkup, F. Hsu, S. Heitner, R. A. Harte, M. Haeussler, L. Guruvadoo, M. Goldman, B. M. Giardine, P. A. Fujita, T. R. Dreszer, M. Diekhans, M. S. Cline, H. Clawson, G. P. Barber, D. Haussler, and W. J. Kent. **The UCSC Genome Browser**

database: extensions and updates 2013. *Nucleic Acids Res*, **41**(Database issue):D64–D69, 2013. 17, 29, 56, 78, 79

[109] K. D. Pruitt, T. Tatusova, G. R. Brown, and D. R. Maglott. **NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy**. *Nucleic Acids Res*, **40**(Database issue):D130–5, 2012. 17

[110] T. A. Down, V. K. Rakyan, D. J. Turner, P. Flicek, H. Li, E. Kulesha, S. Graf, N. Johnson, J. Herrero, E. M. Tomazou, N. P. Thorne, L. Backdahl, M. Herberth, K. L. Howe, D. K. Jackson, M. M. Miretti, J. C. Marioni, E. Birney, T. J. Hubbard, R. Durbin, S. Tavare, and S. Beck. **A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis**. *Nat Biotechnol*, **26**(7):779–85, 2008. 17

[111] F. V. Jacinto, E. Ballestar, and M. Esteller. **Methyl-DNA immunoprecipitation (MeDIP): hunting down the DNA methylome**. *Biotechniques*, **44**(1):35, 37, 39 passim, 2008. 17

[112] D. Serre, B. H. Lee, and A. H. Ting. **MBD-isolated Genome Sequencing provides a high-throughput and comprehensive survey of DNA methylation in the human genome**. *Nucleic Acids Res*, **38**(2):391–9, 2010. 17

[113] S. S. Nair, M. W. Coolen, C. Stirzaker, J. Z. Song, A. L. Statham, D. Strbenac, M. D. Robinson, and S. J. Clark. **Comparison of methyl-DNA immunoprecipitation (MeDIP) and methyl-CpG binding domain (MBD) protein capture for genome-wide DNA methylation analysis reveal CpG sequence coverage bias**. *Epigenetics*, **6**(1):34–44, 2011. 18

[114] M. Frommer, L. E. McDonald, D. S. Millar, C. M. Collis, F. Watt, G. W. Grigg, P. L. Molloy, and C. L. Paul. **A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands**. *Proc Natl Acad Sci U S A*, **89**(5):1827–31, 1992. 18

[115] M. D. Robinson, A. L. Statham, T. P. Speed, and S. J. Clark. **Protocol matters: which methylome are you actually studying?** *Epigenomics*, **2**(4):587–98, 2010. 18

[116] A. Adey and J. Shendure. **Ultra-low-input, tagmentation-based whole-genome bisulfite sequencing**. *Genome Res*, **22**(6):1139–43, 2012. 18, 19, 76

[117] A. Meissner, A. Gnirke, G. W. Bell, B. Ramsahoye, E. S. Lander, and R. Jaenisch. **Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis**. *Nucleic Acids Res*, **33**(18):5868–77, 2005. 19

[118] http://www.bioinformatics.babraham.ac.uk/projects/fastqc/, 2016. Accessed on 2016-01-07. 20

[119] R. Schmieder and R. Edwards. **Quality control and preprocessing of metagenomic datasets**. *Bioinformatics*, **27**(6):863–4, 2011. 20

[120] Marcel Martin. **Cutadapt removes adapter sequences from high-throughput sequencing reads**. *2011*, **17**(1), 2011. 20

[121] A. M. Bolger, M. Lohse, and B. Usadel. **Trimmomatic: a flexible trimmer for Illumina sequence data**. *Bioinformatics*, **30**(15):2114–20, 2014. 20

[122] https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/, 2016. Accessed on 2016-01-07. 20, 77

[123] http://hannonlab.cshl.edu/fastx_toolkit/, 2016. Accessed on 2016-01-07. 20

[124] K. D. Hansen, B. Langmead, and R. A. Irizarry. **BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions**. *Genome Biol*, **13**(10):R83, 2012. 20, 77, 82

[125] R. Lister and J. R. Ecker. **Finding the fifth base: genome-wide sequencing of cytosine methylation**. *Genome Res*, **19**(6):959–66, 2009. 20

[126] F. Krueger and S. R. Andrews. **Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications**. *Bioinformatics*, **27**(11):1571–2, 2011. 20

[127] B. Pedersen, T. F. Hsieh, C. Ibarra, and R. L. Fischer. **MethylCoder: software pipeline for bisulfite-treated sequences**. *Bioinformatics*, **27**(17):2435–6, 2011. 20

[128] P. Y. Chen, S. J. Cokus, and M. Pellegrini. **BS Seeker: precise mapping for bisulfite sequencing**. *BMC Bioinformatics*, **11**:203, 2010. 20

[129] M. C. Frith, R. Mori, and K. Asai. **A mostly traditional approach improves alignment of bisulfite-converted DNA**. *Nucleic Acids Res*, **40**(13):e100, 2012. 20

[130] E. Y. Harris, N. Ponts, K. G. Le Roch, and S. Lonardi. **BRAT-BW: efficient and accurate mapping of bisulfite-treated reads**. *Bioinformatics*, **28**(13):1795–6, 2012. 20

[131] B. Langmead and S. L. Salzberg. **Fast gapped-read alignment with Bowtie 2**. *Nat Methods*, **9**(4):357–9, 2012. 20

[132] D. J. Tomso and D. A. Bell. **Sequence context at human single nucleotide polymorphisms: overrepresentation of CpG dinucleotide at polymorphic sites and suppression of variation in CpG islands**. *J Mol Biol*, **327**(2):303–8, 2003. 21

[133] http://bioinfo2.ugr.es/NGSmethPipe/, 2016. Accessed on 2016-01-07. 21

[134] Y. Liu, K. D. Siegmund, P. W. Laird, and B. P. Berman. **Bis-SNP: combined DNA methylation and SNP calling for Bisulfite-seq data**. *Genome Biol*, **13**(7):R61, 2012. 21

[135] S. Gao, D. Zou, L. Mao, H. Liu, P. Song, Y. Chen, S. Zhao, C. Gao, X. Li, Z. Gao, X. Fang, H. Yang, T. F. Orntoft, K. D. Sorensen, and L. Bolund. **BS-SNPer: SNP calling in bisulfite-seq data**. *Bioinformatics*, **31**(24):4006–8, 2015. 21, 77

[136] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and Subgroup Genome Project Data Processing. **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics*, **25**(16):2078–9, 2009. 21

[137] M. R. Stratton. **Exploring the genomes of cancer cells: progress and promise**. *Science*, **331**(6024):1553–8, 2011. 24

[138] Rameen Beroukhim, Craig H. Mermel, Dale Porter, Guo Wei, Soumya Raychaudhuri, Jerry Donovan, Jordi Barretina, Jesse Boehm, Jennifer Dobson, Mitsuyoshi

Urashima, Kevin Mc Henry, Reid Pinchback, Azra Ligon, Yoon-Jae Cho, Leila Haery, Heidi Greulich, Michael Reich, Wendy Winckler, Michael Lawrence, Barbara Weir, Kumiko Tanaka, Derek Chiang, Adam Bass, Alice Loo, Carter Hoffman, John Prensner, Ted Liefeld, Qing Gao, Derek Yecies, Sabina Signoretti, Elizabeth Maher, Frederic Kaye, Hidefumi Sasaki, Joel Tepper, Jonathan Fletcher, Josep Tabernero, JosÃ© Baselga, Ming-Sound Tsao, Francesca Demichelis, Mark Rubin, Pasi Janne, Mark Daly, Carmelo Nucera, Ross Levine, Benjamin Ebert, Stacey Gabriel, Anil Rustgi, Cristina Antonescu, Marc Ladanyi, Anthony Letai, Levi Garraway, Massimo Loda, David Beer, Lawrence True, Aikou Okamoto, Scott Pomeroy, Samuel Singer, Todd Golub, Eric Lander, Gad Getz, William Sellers, and Matthew Meyerson. **The landscape of somatic copy-number alteration across human cancers**. *Nature*, **463**(7283):899–905, 2010. 24, 47, 49, 53

[139] M. R. Stratton, P. J. Campbell, and P. A. Futreal. **The cancer genome**. *Nature*, **458**(7239):719–724, 2009. 24

[140] Cancer Genome Atlas Research Network. **The Cancer Genome Atlas Pan-Cancer analysis project**. *Nat Genet*, **45**(10):1113–1120, 2013. 24

[141] S. De and F. Michor. **DNA secondary structures and epigenetic determinants of cancer genome evolution**. *Nat Struct Mol Biol*, **18**(8):950–5, 2011. 24, 25, 31, 64

[142] G. Fudenberg, G. Getz, M. Meyerson, and L. A. Mirny. **High order chromatin architecture shapes the landscape of chromosomal alterations in cancer**. *Nat Biotechnol*, **29**(12):1109–13, 2011. 24

[143] Y. Li, L. Zhang, R. L. Ball, X. Liang, J. Li, Z. Lin, and H. Liang. **Comparative analysis of somatic copy-number alterations across different human cancer types reveals two distinct classes of breakpoint hotspots**. *Hum Mol Genet*, **21**(22):4957–65, 2012. 24, 31, 34, 43, 46, 47, 48, 49, 64, 65

[144] W. Gu, F. Zhang, and J. R. Lupski. **Mechanisms for human genomic rearrangements**. *Pathogenetics*, **1**(1):4, 2008. 25, 64

[145] N. Crosetto, A. Mitra, M. J. Silva, M. Bienko, N. Dojer, Q. Wang, E. Karaca, R. Chiarle, M. Skrzypczak, K. Ginalski, P. Pasero, M. Rowicka, and I. Dikic. **Nucleotide-resolution DNA double-strand break mapping by next-generation sequencing**. *Nat Methods*, **10**(4):361–5, 2013. 25

[146] R. Z. Cer, D. E. Donohue, U. S. Mudunuri, N. A. Temiz, M. A. Loss, N. J. Starner, G. N. Halusa, N. Volfovsky, M. Yi, B. T. Luke, A. Bacolla, J. R. Collins, and R. M. Stephens. **Non-B DB v2.0: a database of predicted non-B DNA-forming motifs and its associated tools**. *Nucleic Acids Res*, **41**(Database issue):D94–D100, 2013. 25, 30, 56

[147] G. Wang, L. A. Christensen, and K. M. Vasquez. **Z-DNA-forming sequences generate large-scale deletions in mammalian cells**. *Proc Natl Acad Sci USA*, **103**(8):2677–82, 2006. 25

[148] H. Inagaki, T. Ohye, H. Kogo, T. Kato, H. Bolor, M. Taniguchi, T. H. Shaikh, B. S. Emanuel, and H. Kurahashi. **Chromosomal instability mediated by non-B DNA: cruciform conformation and not DNA sequence is responsible for recurrent translocation in humans**. *Genome Res*, **19**(2):191–8, 2009. 25

[149] G. Wang and K. M. Vasquez. **Naturally occurring H-DNA-forming sequences are mutagenic in mammalian cells**. *Proc Natl Acad Sci USA*, **101**(37):13448–53, 2004. 25

[150] K. Han, J. Lee, T. J. Meyer, P. Remedios, L. Goodwin, and M. A. Batzer. **L1 recombination-associated deletions generate human genomic variation**. *Proc Natl Acad Sci USA*, **105**(49):19366–71, 2008. 25, 48

[151] I. M. Campbell, T. Gambin, P. Dittwald, C. R. Beck, A. Shuvarikov, P. Hixson, A. Patel, A. Gambin, C. A. Shaw, J. A. Rosenfeld, and P. Stankiewicz. **Human endogenous retroviral elements promote genome instability via nonallelic homologous recombination**. *BMC Biol*, **12**(1):74, 2014. 25

[152] S. Myers, C. Freeman, A. Auton, P. Donnelly, and G. McVean. **A common sequence motif associated with recombination hot spots and genome instability in humans**. *Nat Genet*, **40**(9):1124–9, 2008. 25

[153] W. Zhou, F. Zhang, X. Chen, Y. Shen, J. R. Lupski, and L. Jin. **Increased genome instability in human DNA segments with self-chains: homology-induced structural variations via replicative mechanisms**. *Hum Mol Genet*, **22**(13):2642–51, 2013. 25, 29, 64

[154] B. Schuster-Bockler and B. Lehner. **Chromatin organization is a major influence on regional mutation rates in human cancer cells**. *Nature*, **488**(7412):504–507, 2012. 25

[155] Arkarachai Fungtammasan, Erin Walsh, Francesca Chiaromonte, Kristin A. Eckert, and Kateryna D. Makova. **A genome-wide analysis of common fragile sites: What features determine chromosomal instability in the human genome?** *Genome Res*, **22**(6):993–1005, 2012. 25, 31, 36, 57, 62

[156] Rebeca Campos-SÃ¡nchez, AurÃ©lie Kapusta, CÃ©dric Feschotte, Francesca Chiaromonte, and Kateryna D. Makova. **Genomic Landscape of Human, Bat and Ex Vivo DNA Transposon Integrations**. *Mol Biol Evol*, **31**(7):1816–1832, 2014. 25, 31, 36

[157] S. A. Forbes, D. Beare, P. Gunasekaran, K. Leung, N. Bindal, H. Boutselakis, M. Ding, S. Bamford, C. Cole, S. Ward, C. Y. Kok, M. Jia, T. De, J. W. Teague, M. R. Stratton, U. McDermott, and P. J. Campbell. **COSMIC: exploring the world's knowledge of somatic mutations in human cancer**. *Nucleic Acids Res*, **43**(Database issue):D805–11, 2015. 27, 59

[158] A. Siepel, G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. W. Hillier, S. Richards, G. M. Weinstock, R. K. Wilson, R. A. Gibbs, W. J. Kent, W. Miller, and D. Haussler. **Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes**. *Genome Res*, **15**(8):1034–50, 2005. 30, 78

[159] N. A. Tchurikov, O. V. Kretova, D. M. Fedoseeva, D. V. Sosin, S. A. Grachev, M. V. Serebraykova, S. A. Romanenko, N. V. Vorobieva, and Y. V. Kravatsky. **DNA double-strand breaks coupled with PARP1 and HNRNPA2B1 binding sites flank coordinately expressed domains in human chromosomes**. *PLoS Genet*, **9**(4):e1003429, 2013. 30

[160] A. Barski, S. Cuddapah, K. Cui, T. Y. Roh, D. E. Schones, Z. Wang, G. Wei, I. Chepelev, and K. Zhao. **High-resolution profiling of histone methylations in the human genome**. *Cell*, **129**(4):823–37, 2007. 30

[161] R. S. Hansen, S. Thomas, R. Sandstrom, T. K. Canfield, R. E. Thurman, M. Weaver, M. O. Dorschner, S. M. Gartler, and J. A. Stamatoyannopoulos. **Sequencing newly replicated DNA reveals widespread plasticity in human replication timing**. *Proc Natl Acad Sci USA*, **107**(1):139–144, 2010. 30

[162] A. R. Quinlan and I. M. Hall. **BEDTools: a flexible suite of utilities for comparing genomic features**. *Bioinformatics*, **26**(6):841–2, 2010. 31, 32, 57

[163] P. Rice, I. Longden, and A. Bleasby. **EMBOSS: the European Molecular Biology Open Software Suite**. *Trends Genet*, **16**(6):276–7, 2000. 31

[164] Ana Kozomara and Sam Griffiths-Jones. **miRBase: integrating microRNA annotation and deep-sequencing data**. *Nucleic Acids Res*, **39**(Database issue):D152–7, 2011. 31

[165] S. De, B. S. Pedersen, and K. Kechris. **The dilemma of choosing the ideal permutation strategy while estimating statistical significance of genome-wide enrichment**. *Brief Bioinform*, **15**(6):919–928, 2014. 31

[166] David L. Olson and Dursun Delen. *Advanced Data Mining Techniques.* Springer-Verlag, Berlin, Germany, 2008. 32

[167] R Development Core Team. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria, 2011. 33

[168] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S.* Springer, New York, 2002. 33

[169] J. Fox and S. Weisberg. *An R Companion to Applied Regression.* Sage, Thousand Oaks, CA, 2011. 33

[170] T.D. Fletcher. *QuantPsyc: Quantitative Psychology Tools*, 2012. 33

[171] J.H. Maindonald and W.J Braun. *Data Analysis and Graphics Using R.* Cambridge University Press, Cambridge, UK, 2010. 33

[172] K. Imai, G. King, and O. Lau. **Toward A Common Framework for Statistical Analysis and Development.** *J. Comput. Graph. Stat.*, **17**(4):1–22, 2008. 33

[173] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. *Classification and Regression Trees.* Wadsworth, Belmont, CA, 1984. 33

[174] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. **Scikit-learn: Machine Learning in Python**. *J. Mach. Learn. Res.*, **12**:2825–2830, 2011. 34

[175] A. E. Alsop, A. E. Teschendorff, and P. A. Edwards. **Distribution of breakpoints on chromosome 18 in breast, colorectal, and pancreatic carcinoma cell lines**. *Cancer Genet Cytogenet*, **164**(2):97–109, 2006. 47

[176] D. Q. Nguyen, C. Webber, and C. P. Ponting. **Bias of selection on human copy-number variants**. *PLoS Genet*, **2**(2):e20, 2006. 47

[177] A. L. Manning, M. S. Longworth, and N. J. Dyson. **Loss of pRB causes centromere dysfunction and chromosomal instability**. *Genes Dev*, **24**(13):1364–76, 2010. 47

[178] S. Negrini, V. G. Gorgoulis, and T. D. Halazonetis. **Genomic instability–an evolving hallmark of cancer**. *Nat Rev Mol Cell Biol*, **11**(3):220–8, 2010. 47, 53

[179] S E Artandi, S Chang, S L Lee, S Alson, G J Gottlieb, L Chin, and R A DePinho. **Telomere dysfunction promotes non-reciprocal translocations and epithelial cancers in mice**. *Nature*, **406**(6796):641–5, 2000. 47

[180] J. Zhao, A. Bacolla, G. Wang, and K. M. Vasquez. **Non-B DNA structure-induced genetic instability and evolution**. *Cell Mol Life Sci*, **67**(1):43–62, 2010. 47, 72

[181] M. K. Konkel and M. A. Batzer. **A mobile threat to genome stability: The impact of non-LTR retrotransposons upon the human genome**. *Semin Cancer Biol*, **20**(4):211–21, 2010. 48, 72

[182] C. H. Freudenreich. **Chromosome fragility: molecular mechanisms and cellular consequences**. *Front Biosci*, **12**:4911–24, 2007. 48

[183] L. Edelmann, E. Spiteri, N. McCain, R. Goldberg, R. K. Pandita, S. Duong, J. Fox, D. Blumenthal, S. R. Lalani, L. G. Shaffer, and B. E. Morrow. **A common breakpoint on 11q23 in carriers of the constitutional t(11;22) translocation**. *Am J Hum Genet*, **65**(6):1608–16, 1999. 48

[184] S. S. Bielack, B. Kempf-Bielack, D. Branscheid, D. Carrle, G. Friedel, K. Helmke, M. Kevric, G. Jundt, T. Kuhne, R. Maas, R. Schwarz, A. Zoubek, and H. Jurgens. **Second and subsequent recurrences of osteosarcoma: presentation, treatment, and outcomes of 249 consecutive cooperative osteosarcoma study group patients**. *J Clin Oncol*, **27**(4):557–65, 2009. 52

[185] Lisa Mirabello, Rebecca J Troisi, and Sharon A Savage. **Osteosarcoma incidence and survival rates from 1973 to 2004: data from the Surveillance, Epidemiology, and End Results Program**. *Cancer*, **115**(7):1531–43, 2009. 52

[186] Jane Bayani, Maria Zielenska, Ajay Pandita, Khaldoun Al-Romaih, Jana Karaskova, Karen Harrison, Julia A Bridge, Poul Sorensen, Paul Thorner, and Jeremy A Squire. **Spectral karyotyping identifies recurrent complex rearrangements of chromosomes 8, 17, and 20 in osteosarcomas**. *Genes Chromosomes Cancer*, **36**(1):7–16, 2003. 52, 53

[187] M. L. Kuijjer, H. Rydbeck, S. H. Kresse, E. P. Buddingh, A. B. Lid, H. Roelofs, H. Burger, O. Myklebost, P. C. Hogendoorn, L. A. Meza-Zepeda, and A. M. Cleton-Jansen. **Identification of osteosarcoma driver genes by integrative analysis of copy number and gene expression data**. *Genes Chromosomes Cancer*, **51**(7):696–706, 2012. 52, 53

[188] J. Smida, D. Baumhoer, M. Rosemann, A. Walch, S. Bielack, C. Poremba, K. Remberger, E. Korsching, W. Scheurlen, C. Dierkes, S. Burdach, G. Jundt, M. J. Atkinson, and M. Nathrath. **Genomic alterations and allelic imbalances are strong prognostic predictors in osteosarcoma**. *Clin Cancer Res*, **16**(16):4256–67, 2010. 52

[189] Anja Luetke, Paul A Meyers, Ian Lewis, and Heribert Juergens. **Osteosarcoma treatment - where do we stand? A state of the art review**. *Cancer Treat Rev*, **40**(4):523–32, 2014. 52

[190] Tsz-Kwong Man, Murali Chintagumpala, Jaya Visvanathan, Jianhe Shen, Laszlo Perlaky, John Hicks, Mark Johnson, Nelson Davino, Jeffrey Murray, Lee Helman, William Meyer, Timothy Triche, Kwong-Kwok Wong, and Ching C Lau. **Expression profiles of osteosarcoma that can predict response to chemotherapy**. *Cancer Res*, **65**(18):8142–50, 2005. 52

[191] J. W. Martin, J. A. Squire, and M. Zielenska. **The genetics of osteosarcoma**. *Sarcoma*, **2012**:627254, 2012. 53

[192] S. G. Durkin and T. W. Glover. **Chromosome fragile sites**. *Annu Rev Genet*, **41**:169–92, 2007. 53

[193] Michelle K. Zeman and Karlene A. Cimprich. **Causes and consequences of replication stress**. *Nat Cell Biol*, **16**(1):2–9, 2014. 53

[194] Serge J Smeets, Ulrike Harjes, Wessel N van Wieringen, Daoud Sie, Ruud H Brakenhoff, Gerrit A Meijer, and Bauke Ylstra. **To DNA or not to DNA? That is the question, when it comes to molecular subtyping for the clinic!** *Clin Cancer Res*, **17**(15):4959–64, 2011. 53

[195] X. Chen, A. Bahrami, A. Pappo, J. Easton, J. Dalton, E. Hedlund, D. Ellison, S. Shurtleff, G. Wu, L. Wei, M. Parker, M. Rusch, P. Nagahawatte, J. Wu, S. Mao, K. Boggs, H. Mulder, D. Yergeau, C. Lu, L. Ding, M. Edmonson, C. Qu, J. Wang, Y. Li, F. Navid, N. C. Daw, E. R. Mardis, R. K. Wilson, J. R. Downing, J. Zhang, M. A. Dyer, and Project St. Jude Children's Research Hospital-Washington University Pediatric Cancer Genome. **Recurrent somatic structural variations contribute to tumorigenesis in pediatric osteosarcoma**. *Cell Rep*, **7**(1):104–12, 2014. 53, 59, 64, 67, 72

[196] Jennifer A Perry, Adam Kiezun, Peter Tonzi, Eliezer M Van Allen, Scott L Carter, Sylvan C Baca, Glenn S Cowley, Ami S Bhatt, Esther Rheinbay, Chandra Sekhar Pedamallu, Elena Helman, Amaro Taylor-Weiner, Aaron McKenna, David S DeLuca, Michael S Lawrence, Lauren Ambrogio, Carrie Sougnez, Andrey Sivachenko, Loren D Walensky, Nikhil Wagle, Jaume Mora, Carmen de Torres, Cinzia Lavarino, Simone Dos Santos Aguiar, Jose Andres Yunes, Silvia Regina Brandalise, Gabriela Elisa Mercado-Celis, Jorge Melendez-Zajgla, Rocío Cárdenas-Cardós, Liliana Velasco-Hidalgo, Charles W M Roberts, Levi A Garraway, Carlos Rodriguez-Galindo, Stacey B Gabriel, Eric S Lander, Todd R Golub, Stuart H Orkin, Gad Getz, and Katherine A Janeway. **Complementary genomic approaches highlight the PI3K/mTOR pathway as a common vulnerability in osteosarcoma**. *Proc Natl Acad Sci U S A*, **111**(51), 2014. 53

[197] Kathrin Poos, Jan Smida, Doris Maugg, Gertrud Eckstein, Daniel Baumhoer, Michaela Nathrath, and Eberhard Korsching. **Genomic heterogeneity of osteosarcoma - shift from single candidates to functional modules**. *PLoS One*, **10**(4), 2015. 53

[198] M. Kansara, M. W. Teng, M. J. Smyth, and D. M. Thomas. **Translational biology of osteosarcoma**. *Nat Rev Cancer*, **14**(11):722–35, 2014. 53, 62, 64, 68, 71

[199] Florence Magrangeas, Hervé Avet-Loiseau, Nikhil C

Munshi, and Stéphane Minvielle. **Chromothripsis identifies a rare and aggressive entity among newly diagnosed multiple myeloma patients**. *Blood*, **118**(3):675–8, 2011. 54

[200] J. J. Molenaar, J. Koster, D. A. Zwijnenburg, P. van Sluis, L. J. Valentijn, I. van der Ploeg, M. Hamdi, J. van Nes, B. A. Westerman, J. van Arkel, M. E. Ebus, F. Haneveld, A. Lakeman, L. Schild, P. Molenaar, P. Stroeken, M. M. van Noesel, I. Ora, E. E. Santo, H. N. Caron, E. M. Westerhout, and R. Versteeg. **Sequencing of neuroblastoma identifies chromothripsis and defects in neuritogenesis genes**. *Nature*, **483**(7391):589–93, 2012. 54

[201] C. A. Maher and R. K. Wilson. **Chromothripsis and human disease: piecing together the shattering process**. *Cell*, **148**(1-2):29–32, 2012. 54

[202] N. Kolesnikov, E. Hastings, M. Keays, O. Melnichuk, Y. A. Tang, E. Williams, M. Dylag, N. Kurbatova, M. Brandizi, T. Burdett, K. Megy, E. Pilicheva, G. Rustici, A. Tikhonov, H. Parkinson, R. Petryszak, U. Sarkans, and A. Brazma. **ArrayExpress update-simplifying data submissions**. *Nucleic Acids Research*, **43**(D1):D1113–D1116, 2015. 55

[203] C. H. Mermel, S. E. Schumacher, B. Hill, M. L. Meyerson, R. Beroukhim, and G. Getz. **GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers**. *Genome Biol*, **12**(4), 2011. 56, 59

[204] Yi Qiao, Aaron R Quinlan, Amir A Jazaeri, Roeland Gw Verhaak, David A Wheeler, and Gabor T Marth. **SubcloneSeeker: a computational framework for reconstructing tumor clone structure for cancer variant interpretation and prioritization**. *Genome Biol*, **15**(8):443, 2014. 56, 61

[205] C. Hernandez-Ferrer, I. Quintela Garcia, K. Danielski, A. Carracedo, L. A. Perez-Jurado, and J. R. Gonzalez. **affy2sv: an R package to pre-process Affymetrix CytoScan HD and 750K arrays for SNP, CNV, inversion and mosaicism calling**. *BMC Bioinformatics*, **16**:167, 2015. 57

[206] Mohammed Uddin, Bhooma Thiruvahindrapuram, Susan Walker, Zhuozhi Wang, Pingzhao Hu, Sylvia Lamoureux, John Wei, Jeffrey R MacDonald, Giovanna Pellecchia, Chao Lu, Anath C Lionel, Matthew J Gazzellone, John R McLaughlin, Catherine Brown, Irene L Andrulis, Julia A Knight, Jo-Anne Herbrick, Richard F Wintle, Peter Ray, Dimitri J Stavropoulos, Christian R Marshall, and Stephen W Scherer. **A high-resolution copy-number variation resource for clinical and population genetics**. *Genet Med*, **17**(9):747–52, 2014. 57

[207] K. Wang, M. Li, D. Hadley, R. Liu, J. Glessner, S. F. Grant, H. Hakonarson, and M. Bucan. **PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data**. *Genome Res*, **17**(11):1665–74, 2007. 57

[208] T. Popova, E. Manié, and M. Stern. **Genomic Signature of Homologous Recombination Deficiency in Breast and Ovarian Cancers.** *Bio-protocol*, **3**(13), 2013. 58

[209] Tatiana Popova, Elodie Manié, Guillaume Rieunier, Virginie Caux-Moncoutier, Carole Tirapo, Thierry Dubois, Olivier Delattre, Brigitte Sigal-Zafrani, Marc Bollet, Michel Longy, Claude Houdayer, Xavier Sastre-Garau, Anne Vincent-Salomon, Dominique Stoppa-Lyonnet, and Marc-Henri Stern. **Ploidy and large-scale genomic instability consistently identify basal-like breast carcinomas with BRCA1/2 inactivation**. *Cancer Res*, **72**(21):5454–62, 2012. 58

[210] Michal Kovac, Claudia Blattmann, Sebastian Ribi, Jan Smida, Nikola S. Mueller, Florian Engert, Francesc Castro-Giner, Joachim Weischenfeldt, Monika Kovacova, Andreas Krieg, Dimosthenis Andreou, Per-Ulf Tunn, Hans Roland Durr, Hans Rechl, Klaus-Dieter Schaser, Ingo Melcher, Stefan Burdach, Andreas Kulozik, Katja Specht, Karl Heinimann, Simone Fulda, Stefan Bielack, Gernot Jundt, Ian Tomlinson, Jan O. Korbel, Michaela Nathrath, and Daniel Baumhoer. **Exome sequencing of osteosarcoma reveals mutation signatures reminiscent of BRCA deficiency**. *Nat Commun*, **6**:8940, 2015. 59, 61, 71

[211] Sebastian Ribi, Daniel Baumhoer, Kristy Lee, Audrey S M Teo, Babita Madan, Kang Zhang, Wendy K Kohlmann, Fei Yao, Wah Heng Lee, Qiangze Hoi, Shaojiang Cai, Xing Yi Woo, Patrick Tan, Gernot Jundt, Jan Smida, Michaela Nathrath, Wing-Kin Sung, Joshua D Schiffman, David M Virshup, and Axel M Hillmer. **TP53 intron 1 hotspot rearrangements are specific to sporadic osteosarcoma and can cause Li-Fraumeni syndrome**. *Oncotarget*, **6**(10):7727–40, 2015. 59, 64

[212] J. W. Martin, M. Zielenska, G. S. Stein, A. J. van Wijnen, and J. A. Squire. **The Role of RUNX2 in Osteosarcoma Oncogenesis**. *Sarcoma*, **2011**:282745, 2011. 59

[213] L. Weiner, R. Han, B. M. Scicchitano, J. Li, K. Hasegawa, M. Grossi, D. Lee, and J. L. Brissette. **Dedicated epithelial recipient cells determine pigmentation patterns**. *Cell*, **130**(5):932–42, 2007. 61

[214] M. Mangelsdorf, K. Ried, E. Woollatt, S. Dayan, H. Eyre, M. Finnis, L. Hobson, J. Nancarrow, D. Venter, E. Baker, and R. I. Richards. **Chromosomal fragile site FRA16D and DNA instability in cancer**. *Cancer Res*, **60**(6):1683–9, 2000. 61

[215] R. I. Aqeilan, M. Abu-Remaileh, and M. Abu-Odeh. **The common fragile site FRA16D gene product WWOX: roles in tumor suppression and genomic stability**. *Cell Mol Life Sci*, **71**(23):4589–99, 2014. 61, 71

[216] M. S. Schrock and K. Huebner. **WWOX: a fragile tumor suppressor**. *Exp Biol Med (Maywood)*, **240**(3):296–304, 2015. 61

[217] J. Yang, D. Cogdell, D. Yang, L. Hu, H. Li, H. Zheng, X. Du, Y. Pang, J. Trent, K. Chen, and W. Zhang. **Deletion of the WWOX gene and frequent loss of its protein expression in human osteosarcoma**. *Cancer Lett*, **291**(1):31–8, 2010. 61, 71

[218] S. Del Mare and R. I. Aqeilan. **Tumor Suppressor WWOX inhibits osteosarcoma metastasis by modulating RUNX2 function**. *Sci Rep*, **5**:12959, 2015. 61

[219] S. Zheng, J. Fu, R. Vegesna, Y. Mao, L. E. Heathcock, W. Torres-Garcia, R. Ezhilarasan, S. Wang, A. McKenna, L. Chin, C. W. Brennan, W. K. Yung, J. N. Weinstein, K. D. Aldape, E. P. Sulman, K. Chen, D. Koul, and R. G. Verhaak. **A survey of intragenic breakpoints in glioblastoma identifies a distinct subset associated with poor survival**. *Genes Dev*, **27**(13):1462–72, 2013. 62

[220] Y. Zhang, H. Xu, and D. Frishman. **Genomic determinants of somatic copy number alterations across human cancers**. *Hum Mol Genet*, **25**(5):1019–30, 2016. 64

[221] J. V. Forment, A. Kaidi, and S. P. Jackson. **Chromothripsis and cancer: causes and consequences of chromosome shattering**. *Nat Rev Cancer*, **12**(10):663–70, 2012. 67

[222] B. R. Mardin, A. P. Drainas, S. M. Waszak, J. Weischenfeldt, M. Isokane, A. M. Stutz, B. Raeder, T. Efthymiopoulos, C. Buccitelli, M. Segura-Wang, P. Northcott, S. M. Pfister, P. Lichter, J. Ellenberg, and J. O. Korbel. **A cell-based model system links chromothripsis with hyperploidy**. *Mol Syst Biol*, **11**(9):828, 2015. 68, 72

[223] K. C. Kurek, S. Del Mare, Z. Salah, S. Abdeen, H. Sadiq, S. H. Lee, E. Gaudio, N. Zanesi, K. B. Jones, B. DeYoung, G. Amir, M. Gebhardt, M. Warman, G. S. Stein, J. L. Stein, J. B. Lian, and R. I. Aqeilan. **Frequent Attenuation of the WWOX Tumor Suppressor in Osteosarcoma Is Associated with Increased Tumorigenicity and Aberrant RUNX2 Expression**. *Cancer Research*, **70**(13):5577–5586, 2010. 71

[224] M. I. Nunez, D. G. Rosen, J. H. Ludes-Meyers, M. C. Abba, H. Kil, R. Page, A. J. Klein-Szanto, A. K. Godwin, J. Liu, G. B. Mills, and C. M. Aldaz. **WWOX protein expression varies among ovarian carcinoma histotypes and correlates with less favorable outcome**. *BMC Cancer*, **5**:64, 2005. 71

[225] X. Duan, E. Kang, C. Y. Liu, G. L. Ming, and H. Song. **Development of neural stem cell in the adult brain**. *Curr Opin Neurobiol*, **18**(1):108–15, 2008. 76

[226] F. H. Gage. **Mammalian neural stem cells**. *Science*, **287**(5457):1433–8, 2000. 76

[227] R. A. Ihrie and A. Alvarez-Buylla. **Lake-front property: a unique germinal niche by the lateral ventricles of the adult brain**. *Neuron*, **70**(4):674–86, 2011. 76

[228] G. L. Ming and H. Song. **Adult neurogenesis in the mammalian brain: significant answers and significant questions**. *Neuron*, **70**(4):687–702, 2011. 76

[229] B. Juliandi, M. Abematsu, and K. Nakashima. **Epigenetic regulation in neural stem cell differentiation**. *Dev Growth Differ*, **52**(6):493–504, 2010. 76

[230] M. Namihira, J. Kohyama, M. Abematsu, and K. Nakashima. **Epigenetic mechanisms regulating fate specification of neural stem cells**. *Philos Trans R Soc Lond B Biol Sci*, **363**(1500):2099–109, 2008. 76

[231] R. P. Singh, K. Shiue, D. Schomberg, and F. C. Zhou. **Cellular epigenetic modifications of neural stem cell differentiation**. *Cell Transplant*, **18**(10):1197–211, 2009. 76

[232] E. Llorens-Bobadilla, S. Zhao, A. Baser, G. Saiz-Castro, K. Zwadlo, and A. Martin-Villalba. **Single-Cell Transcriptomics Reveals a Population of Dormant Neural Stem Cells that Become Activated upon Brain Injury**. *Cell Stem Cell*, **17**(3):329–40, 2015. 76

[233] Q. Wang, L. Gu, A. Adey, B. Radlwimmer, W. Wang, V. Hovestadt, M. Bahr, S. Wolf, J. Shendure, R. Eils, C. Plass, and D. Weichenhan. **Tagmentation-based whole-genome bisulfite sequencing**. *Nat Protoc*, **8**(10):2022–32, 2013. 76

[234] B.S. Pedersen, K. Eyring, S. De, I.V. Yang, and D.A. Schwartz. **Fast and Accurate Alignment of Long Bisulfite-Seq Reads.** http://arxiv.org/. Accessed on 2016-05-07. 77

[235] Picard. **Picard: a set of tools (in java) for working with next generation sequencing data in the bam format.** http://sourceforge.net/projects/picard. Accessed on 2016-05-07. 77

[236] S. Heinz, C. Benner, N. Spann, E. Bertolino, Y. C. Lin, P. Laslo, J. X. Cheng, C. Murre, H. Singh, and C. K. Glass. **Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities**. *Mol Cell*, **38**(4):576–89, 2010. 78, 83

[237] C. Y. McLean, D. Bristor, M. Hiller, S. L. Clarke, B. T. Schaar, C. B. Lowe, A. M. Wenger, and G. Bejerano. **GREAT improves functional interpretation of cis-regulatory regions**. *Nat Biotechnol*, **28**(5):495–501, 2010. 78, 85

[238] G. Bindea, B. Mlecnik, H. Hackl, P. Charoentong, M. Tosolini, A. Kirilovsky, W. H. Fridman, F. Pages, Z. Trajanoski, and J. Galon. **ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks**. *Bioinformatics*, **25**(8):1091–3, 2009. 79

[239] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. **Cytoscape: a software environment for integrated models of biomolecular interaction networks**. *Genome Res*, **13**(11):2498–504, 2003. 79

[240] Encode Consortium Mouse, J. A. Stamatoyannopoulos, M. Snyder, R. Hardison, B. Ren, T. Gingeras, D. M. Gilbert, M. Groudine, M. Bender, R. Kaul, T. Canfield, E. Giste, A. Johnson, M. Zhang, G. Balasundaram, R. Byron, V. Roach, P. J. Sabo, R. Sandstrom, A. S. Stehling, R. E. Thurman, S. M. Weissman, P. Cayting, M. Hariharan, J. Lian, Y. Cheng, S. G. Landt, Z. Ma, B. J. Wold, J. Dekker, G. E. Crawford, C. A. Keller, W. Wu, C. Morrissey, S. A. Kumar, T. Mishra, D. Jain, M. Byrska-Bishop, D. Blankenberg, B. R. Lajoie, G. Jain, A. Sanyal, K. B. Chen, O. Denas, J. Taylor, G. A. Blobel, M. J. Weiss, M. Pimkin, W. Deng, G. K. Marinov, B. A. Williams, K. I. Fisher-Aylor, G. Desalvo, A. Kiralusha, D. Trout, H. Amrhein, A. Mortazavi, L. Edsall, D. McCleary, S. Kuan, Y. Shen, F. Yue, Z. Ye, C. A. Davis, C. Zaleski, S. Jha, C. Xue, A. Dobin, W. Lin, M. Fastuca, H. Wang, R. Guigo, S. Djebali, J. Lagarde, T. Ryba, T. Sasaki, V. S. Malladi,

M. S. Cline, V. M. Kirkup, K. Learned, K. R. Rosenbloom, W. J. Kent, E. A. Feingold, P. J. Good, M. Pazin, R. F. Lowdon, and L. B. Adams. **An encyclopedia of mouse DNA elements (Mouse ENCODE)**. *Genome Biol*, **13**(8):418, 2012. 79, 85

[241] A. Visel, E. M. Rubin, and L. A. Pennacchio. **Genomic views of distant-acting enhancers**. *Nature*, **461**(7261):199–205, 2009. 84

[242] C. Buecker and J. Wysocka. **Enhancers as information integration hubs in development: lessons from genomics**. *Trends Genet*, **28**(6):276–84, 2012. 87