TECHNISCHE UNIVERSITÄT MÜNCHEN
Fachgebiet für Bioinformatik

Somatic Mutations in Tumors: Their Detection, Generation Mechanisms and Implications

HONGEN XU

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. Arne Skerra
Prüfer der Dissertation:	
	1. Prof. Dr. Dmitrij Frischmann
	2. Prof. Dr. Ralf Zimmer
	3. Prof. Angelika Schnieke, Ph.D.

Die Dissertation wurde am 22.06.2017 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 06.10.2017 angenommen.

Abstract

Cancer is fundamentally a disease of genome characterized by somatically acquired mutations. Recent advances in high-throughput genomic technologies such as single nucleotide polymorphism (SNP) microarrays and DNA next-generation sequencing have enabled us to explore the mutational landscape of cancer genomes at an unprecedented resolution. Somatic mutations include single nucleotide variants (SNVs), small insertions/deletions (indels), somatic copy number alterations (SCNAs), structural variations (SVs), and epigenetic changes altering gene expression and chromatin structure. On one hand, the characterization of somatic mutations allows the identification of driver mutations and driver genes, providing new insights into the underlying mechanism of tumorigenesis and possibly revealing new therapeutic targets for cancer treatment. On the other hand, the exploration of somatic alterations makes it possible to investigate generation mechanism of somatic mutations, contributing to the understanding of DNA damage and repair processes that have been operative throughout the development of cancer.

In this thesis, we investigated generation mechanisms of somatic mutations, especially SNVs (Chapter 2) and SCNAs (Chapter 3) in diverse tumor types. Taking advantage of available genetic and epigenetic features, we showed that SNV rate in cancer genome is strikingly related to chromatin organization. We also revealed that the strong association between SNV mutation rate and chromatin organization is independent of tissue and mutation types. For SCNAs, we conducted multiple linear regression (MLR) analyses of the pooled SCNA data from The Cancer Genome Atlas Pan-Cancer project. Our MLR model explains $>$30% of the pooled SCNA breakpoint variation, with the explanatory power ranging from 13 to 32% for 11 different cancer types and SCNA types—amplifications and deletions, telomere-bound and interstitial SCNAs and local SCNAs. In addition to confirming previously identified features, 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 database. Using a rare event logistic regression model and an extremely randomized tree classifier, we revealed that genomic features are informative for telling apart common SCNA breakpoint hotspots and non-hotspots.

We also characterized SCNAs and chromosomal breaks in human osteosarcoma (OS, Chapter 4) as well as SNVs, indels, SCNAs and SVs in chicken Marek's Disease (MD)

lymphomas (Chapter 5). OS is the most common primary malignant bone tumor in children and adolescents. We performed a comprehensive assessment of SCNAs in 160 OS samples using whole-genome CytoScan High Density arrays (Affymetrix, Santa Clara, CA). Breakage analysis revealed OS specific unstable 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 particularly chromothripsis were strongly correlated with OS patient clinical outcome. MD is a lymphoproliferative disease in chickens caused by MD Virus, a highly oncogenic α-herpesvirus. We explored the somatic mutational landscape of MD with multiple approaches including whole genome sequencing, whole transcriptome sequencing and SNP microarrays. We identified 54 high-confidence driver genes, of which *IKZF1* encodes a transcription factor associated with chromatin remodeling and is an important player in lymphomagenesis.

Overall, our results contribute to the understanding how somatic mutations drive tumorigenesis and shed light on the molecular mechanisms of somatic mutation generation in cancer.

Keywords: Somatic Mutations; Single Nucleotide Variants (SNVs); Small Insertions and/or Deletions (Indels); Somatic Copy Number Alterations (SCNAs); Structural Variations (SVs); Driver Genes; Generation Mechanism, Osteosarcoma; Marek's Disease.

Zusammenfassung

Krebs ist eine genomische Krankheit, die auf der Entstehung von somatischen Mutationen basiert. Aktuelle Fortschritte bei Hochdurchsatz-Technologien, wie etwa Einzelnukleotid Polymorphismen (SNP) Microarrays und DNA Next-Generation Sequenzierung, ermöglichen uns die Analyse von Mutationen in Krebsgenomen in einer bisher nie dagewesenen Auflösung. Somatische Mutationen sind Einzelnukleotid-Varianten (SNVs), kleine Insertionen/Deletionen (Indels), Änderungen der somatischen Kopienanzahlen (SCNAs), strukturelle Variationen (SVs) und epigenetische Änderungen, die Genexpression und Chromatinstruktur beeinflussen. Auf der einen Seite erlaubt die Charakterisierung somatischer Mutationen die Identifikation von Driver-Mutationen und Driver-Genen, um so neue Erkenntnisse über die zugrundeliegenden Mechanismen der Tumorgenese zu erlangen, die eventuell zu neuen Therapieansätze für die Behandlung von Krebs führen. Auf der anderen Seite ermöglicht die Erforschung von somatischen Veränderungen, die Mechanismen hinter der Entstehung von somatischen Mutationen zu untersuchen, um so die Prozesse von DNA-Schädigung und -Reparatur zu verstehen, die hinter der Entwicklung von Krebs stehen.

Im Rahmen dieser Arbeit haben wir die Entstehungsmechanismen von somatischen Mutationen, im speziellen SNVs (Kapitel 2) und SCNAs (Kapitel 3) in vielfältig Tumoren untersucht. Durch verfügbare genetische und epigenetische Eigenschaften haben wir demonstriert, dass die SNV-Rate im Krebsgenom in auffälliger Weise mit der Chromatinorganisation zusammenhängt. Wir haben außerdem gezeigt, dass der deutliche Zusammenhang zwischen SNV-Mutationsrate und Chromatinorganisation unabhängig von Gewebeart und Mutationstyp ist. Basierend auf den zusammengelegten SCNA-Daten des The Cancer Genome Atlas Pan-Cancer-Projekts haben wir Analysen mittels multipler linearer Regression (MLR) ausgeführt. Unser MLR-Modell erklärt $>30\%$ der SCNA Breakpoint-Variation, wobei die Aussagekraft zwischen 13 und 32% für 11 verschiedene Krebstypen und SCNA-Typen — Vervielfältigungen und Deletionen, Telomer-gebundene und interstitielle SCNAs und lokale SCNAs — liegt. Zusätzlich zum Nachweis bisher identifizierter Eigenschaften haben wir auch weitere neue informative Eigenschaften identifiziert, wie z.B. Distanz zum Telomer, Distanz zum Zentromer und Wiederholungen von geringer Komplexität. Die Ergebnisse der MLR-Analyse wurden außerdem durch einem unabhängigen SCNA-Datensatz aus der Catalogue of Somatic Mutations in Cancer Datenbank verifiziert. Mit einem logistischen Regressionsmodell für seltene Ereignisse und einem extrem randomisierten Entscheidungsbaum-Klassifizierer konnten wir zeigen,

dass mithilfe genomischer Eigenschaften SCNA Breakpoint Hotspots und Nicht-Hotspots auseinandergehalten können werden.

Wir haben außerdem SCNAs und Chromosombrüche in Osteosarkomen beim Menschen (OS, Kapitel 4) und zusätzlich SNVs, Indels, SCNAs und SVs in von der Marek-Krankheit (MD) verursachten Lymphomen bei Hühnern (Kapitel 5) charakterisiert. OS ist der häufigste primäre bösartige Knochentumor bei Kindern und Jugendlichen. Wir haben SCNAs in 160 OS-Proben mittels CytoScan High Density arrays (Affymetrix, Santa Clara, CA) für komplette Genome verglichen. Eine Bruchstellenanalyse hat für OS spezifische instabile Regionen aufgezeigt, in denen sich bekannte OS Tumorsuppressionsgene befinden, unter anderem *TP53*, *RB1*, *WWOX*, *DLG2* und *LSAMP*. Bestimmte Genomeigenschaften, wie etwa Transposons oder nicht-B DNA-bildende Motive, waren deutlich häufiger in der näheren Umgebung von Chromosombruchstellen zu finden. Ein komplexes Bruchmuster — Chromothripsis — scheint ein verbreitetes Phänomen bei OS zu sein. Es konnte gezeigt werden, dass Hyperploidie und speziell Chromothripsis deutlich mit dem klinischen Ergebnis von OS-Patienten zusammenhängen. MD ist eine lymphoproliferative Krankheit bei Hühnern, die vom MD-Virus, bei dem es sich um ein hoch onkogenes α-Herpesvirus handelt, verursacht wird. Wir haben die somatischen Mutationen in MD mithilfe verschiedener Ansätze analysiert, unter anderem Sequenzierung kompletter Genome, Sequenzierung kompletter Transkriptome und SNP Microarrays. Wir konnten 54 Driver-Gene mit großer Gewissheit identifizieren, darunter *IKZF1*, das für einen Transkriptionsfaktor codiert, der mit Chromation-Remodellierung assoziiert wird und eine wichtige Rolle bei der Lymphomagenese spielt.

Zusammenfassend betrachtet tragen unsere Ergebnisse zum Verständnis bei, wie somatische Mutationen Tumorgenese vorantreiben, und beleuchten die molekularen Mechanismen der Entstehung von somatischen Mutationen bei Krebs.

Publications

* Equal contribution

1. Zhang, Y.*, **Xu, H.***, and Frishman, D. (2016) Genomic determinants of somatic copy number alterations across human cancers. *Hum. Mol. Genet.*, 25(5), 1019-1030.

2. 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*, 141(4), 816–828.

3. **Xu, H.**, Pausch, H., Rutkowska, K., Wurmser, C., Reblinger, B., Venhoranta, H., Flisikowska, T., Frishman, D., Zwierzchowski, L., Andersson, M., Fries, R., Kind, A., Schnieke, A., and Flisikowski, K. (2017) Differential transcriptome pattern in maternal and foetal placenta in intrauterine growth restriction. *Biol. Reprod.*, 97(2), 249-257.

4. Flisikowska, T., Stachowiak, M., **Xu, H.**, Wagner, A., Caceres, H. A., Wurmser, C., Wander, C., Pausch, H., Perkowska, A., Fischer, K., Frishman, D., Fries, R., Switonski, M., Kind, A., Saur, D., Schnieke, A. and Flisikowski, K. (2017) Porcine familial adenomatous polyposis model enables systematic analysis of early events in adenoma progression. *Sci. Rep.*, 7, 6613.

5. Steep, A.*, **Xu, H.***, Zhang, Y., Pyrkosz, A., Delany, M.E., Frishman, D., and Cheng, H.H. (2017) Preliminary analysis of somatic mutational landscape of Marek's disease lymphomas in chickens. Manuscript in Preparation.

6. Bobadilla, E.*, Zhang, Y.*, Dehler, S*., Zhao, S., **Xu, H.**, Frishman, D., and Villalba, A.M. (2017) Injury signals uncover a latent regenerative program in mouse neural stem cells. Manuscript in Preparation.

Acknowledgments

It has been almost five years since I started my doctoral study. The present dissertation is a result of several cooperative projects with different researchers. I am very grateful to people who have helped or accompanied me during my doctoral study.

First, I would like to thank Prof. Dmitrij Frishman for giving me the opportunity to work in his group and for introducing me to the fascinating research field of cancer genomics. I appreciate the freedom that I have been given and the stimulating discussions with him. I also appreciate his help in improving my manuscripts and I really learn a lot from it.

I would like to acknowledge the financial support of the China Scholarship Council. I appreciate having such an opportunity to study in Technical University of Munich. I am grateful to Prof. Dmitrij Frishman and Prof. Angelika Schnieke for their financial support for the last year of my doctoral study.

Special thanks go to my main collaborators Yanping Zhang, Dr. Jan Smida, Alexander Steep and Dr. Krzysztof Flisikowski. It was a great pleasure to work with you. I want to thank Alexander Steep for all the lessons he has taught me about cancer genomics and for his insightful comments on this dissertation.

I would like to express my gratitude to my colleagues in Department of Genome-oriented Bioinformatics. Thanks to Léonie Corry, Roswitha Weinbrunn and Claudia Luksch for their administrative help since the moment I came to Germany. Thanks to Drazen Jalsovec for his work in maintaining IT infrastructure. Thanks to Yu Wang for giving me access to a high-performance computer cluster. I would like to thank Anja Mösch for translating the abstract of this dissertation into German. Special thanks go to Yanping Zhang, Fei Qi, Bo Zeng, Jinlong Ru, Hengyuan Liu, Stefanie Kaufmann, Kerstin Haase, Usman Saeed, Peter Hönigschmid, Nermin Pinar Karabulut and Evans Kataka for insightful discussions and interesting conversations.

I appreciate very much the scientific inputs from external researchers, including Travis I. Zack, Yudong Li, Subhajyoti De, Weichen Zhou, Feng Zhang, Norbert Krautenbacher, Haoyang Cai, Peter Van Loo, Ao Li and Yi Qiao.

I am very grateful to Prof. Arne Skerra who kindly agreed to be the Chair of my examining committee. I am also very grateful to Prof. Ralf Zimmer and Prof. Angelika Schnieke who kindly agreed to be the examiner of my dissertation.

I would like to express my gratitude to Yao Lu for her help in preparing figures for this dissertation. I also wish to thank Jinlong Ru and Fei Qi for their bioinformatics assistance.

I would like to thank Chengdong Zheng, Yao Lu, Xingyue Ma, Yuting Xie, Jie Luo, Stefan Steinhauser, Pauline de Jerphanion, Bo Zeng, Yanping Zhang, Shun Li, Kai Li, Fei Qi, Haitao Liu, Saiqi Yang, Fang Yang, Xiangdong Zhao, Tingting Chen, Kun Qian, Yu Zhuang, Guo Chen, Baopeng Ma, Henyuan Liu, Jinlong Ru, Lin Zhao, Hanyin Sun, Jiekui Zou accompanied me during the doctoral study.

Finally, I wish to thank my parents Faqing Xu and Dongna Tian, my wife Yuanyuan Ma and my sisters Fengli Xu, Lihong Xu and Xiaohong Xu. All of your continued support is deeply appreciated.

Contents

List of Tables

List of Figures

Abbreviations

array-CGH	array Comparative Genomic Hybridization
AUC	Area Under the ROC Curve
BAF	B Allele Frequency
BIC	Bayesian Information Criterion
Bisulfit-seq	Bisulfit sequencing
BLCA	Bladder urothelialc Carcinoma
BMR	Background Mutation Rate
BRCA	Breast invasive Carcinoma
CBS	Circular Binary Segmentation
CGC	Cancer Gene Census
ChIP-seq	Chromatin Immunoprecipitation followed by sequencing
CHSs	Common Hotspots
CNP	Copy Number Polymorphism
CNVs	Copy Number Variations
COAD	Colon Adenocarcinoma
COSMIC	Catalog Of Somatic Mutations In Cancer
CTLP	Chromothripsis-Like Pattern
CytoScan-HD	CytoScan High Density
DHS	DNase I Hypersensitive Sites
DPI	Days Post Infection
DSBs	Double Strand Breaks
EMBOSS	European Molecular Biology Open Software Suite
ENCODE	Encyclopedia Of DNA Elements
FDR	False Discovery Rate
FM	Functional Mutation
FoSTeS	Fork Stalling and Template Switching
G4	G-quadruplexes
GATK	Genome Analysis Toolkit
GBM	Glioblastoma Multiforme
GISTIC	Genomic Identification of Significant Targets In Cancer
GO	Gene Ontology
GOSS	Gene Ontology Similarity Score

GSEA	Gene Set Enrichment Analysis
HGMD	Human Gene Mutation Database
HGP	Human Genome Project
HMM	Hidden Markov Model
HNSC	Head and Neck Squamous cell Carcinoma
HPRD	Human Protein Reference Database
HR	Homologous Recombination
ICGC	International Cancer Genome Consortium
Indels	Insertions/Deletions
Kb	Kilo base pair
KEGG	Kyoto Encyclopedia of Genes and Genomes
KIRC	Kidney Renal clear cell Carcinoma
LCRs	Low Copy Repeats
LINEs	Long Interspersed Nuclear Elements
LOH	Loss Of Heiterozygosity
LR	Logistic Regression
LRR	Log R Ratio
LTRs	Long Terminal Repeats
LUAD	Lung Adenocarcinoma
LUSC	Lung Squamous cell Carcinoma
Mb	Mega base pair
MDA	Mean Decrease Accuracy
MD	Marek's Disease
MDI	Mean Decrease Impurity
MDV	Marek's Disease Virus
MLR	Multiple Linear Regression
MMBIR	Micro-homology Mediated Break-Induced Replication
MMEJ	Micro-homology Mediated End Joining
MMR	Mismatch Repair
NAHR	Non-Allelic Homologous Recombination
NCHSs	Non-common hotspots
NER	Nucleotide Excision Repair
NGS	Next Generation Sequencing
NHEJ	Non-Homologous End Joining

NHSs	Non-hotspots
OMIM	Online Mendelian Inheritance in Man
OS	Osteosarcoma
OV	Ovarian serous cystadenocarcinoma
PCR	Polymerase Chain Reaction
PES	Paired End Sequencing
RCVE	Relative Contribution to Variance Explained
READ	Rectum Adenocarcinoma
RELR	Rare Events Logistic Regression
RNA-seq	RNA sequencing
ROC	Receiver Operating Characteristic
RT	Replication Timing
SCs	Self-Chain regions
SCSs	Self-Chain Segments
SCNAs	Somatic Copy Number Alterations
SDs	Segmental Duplications
SES	Single End Sequencing
SINEs	Short Interspersed Nuclear Elements
SMGs	Significantly Mutated Genes
SNPs	Single Nucleotide Polymorphism
SNP-FASST2	SNP-Fast Adaptive States Segmentation Technique 2
SNVs	Single Nucleotide Variants
SRS	Serial Replication Slippage
SSA	Single-Strand Annealing
SVs	Structural Variants
TCGA	The Cancer Genome Atlas
TSG	Tumor Suppressor Genes
UCEC	Uterine Corpus Endometrial Carcinoma
UCSC	University of California, Santa Cruz
UV	Ultra-violet
VIFs	Variance Inflation Factors
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing

1 Literature review

Cancer is a group of more than 200 distinct diseases involving abnormal proliferation of cells with the potential to invade or metastasize to other normal tissues and organs [1]. Since 2010, cancer has been the leading cause of death in China with an estimated 4.29 million new cases and 2.81 million deaths in the year 2015 alone [2]. To effectively diagnose and treat cancer, better understanding of the disease is required. The last century has witnessed a tremendous advance in our knowledge of cancer, and an emerging consensus is that cancer is a disease of the genome.

1.1 Cancer is a disease of the genome

More than a century ago, seminal studies on the development of doubly fertilized sea urchin eggs by Theodor Boveri led to the hypothesis that cancer is caused by chromosomal abnormalities [3], in other words, cancer is "a disease of the genome"[4, 5]. At the beginning of the 20th century, cancer causing chemicals were discovered, however, their cellular targets have not yet been identified [6]. The discovery of DNA as the genetic material of inheritance [7] and the determination of its structure by Watson and Crick [8] indicated that DNA was the cellular target for chemical carcinogens and that these agents generate mutations leading to cancer [6]. The role of genetic mutations in human cancer was confirmed by the discovery of translocation between chromosomes 9 and 22 (known as the "Philadelphia chromosome") in chronic myeloid leukemia [9–11]. The discovery of the Philadelphia chromosome in almost all cases of a specific human cancer strongly supported Boveri's hypothesis that a critical genetic alteration in a single cell could give rise to a tumor [12]. Advances in molecular techniques later allowed the identification of

critical genes involved in the Philadelphia chromosome: v-abl Abelson murine leukemia viral oncogene homolog (*ABL*) on chromosome 9 and breakpoint cluster region (*BCR*) on chromosome 22 [13]. The idea that cancer is a disease of an altered genome attracted wider attention following the discovery that transfer of total genomic DNA from tumor cells into other cells was sufficient to cause transformation [14, 15]. Cloning and characterization of the specific DNA segment responsible for the transformation led to the identification of the first oncogene—*HRAS*, followed by the discovery of the exact point mutation (G >T substitution) in codon 12 resulting in a glycine to valine substitution [16–18]. These landmark findings launched a new era of molecular cancer genetics research that continues to date: identification of mutated genes causally implicated in the development of human cancer (cancer genes) [4, 19].

1.1.1 Cancer genes: oncogenes and tumor suppressor genes

A major aim of cancer studies is to search for genes that are implicated in tumor initiation and development. Based on whether mutations are dominant or recessive at the cellular level, cancer genes can be divided into oncogenes (dominant mutation, a single altered allele is sufficient to initiate cancer) and tumor suppressor genes (TSGs) (recessive mutation, both alleles need to be changed)[19].

The protein products of oncogenes include transcription factors, chromatin remodelers, growth factors, growth factor receptors, signal transducers, and apoptosis regulators [20]. Oncogenes are altered in ways that render them permanently active or active when they are not supposed to [21]. Oncogene activation can be achieved by chromosomal translocations, gene amplifications, intragenic mutations, or by changes in methylation [21]. A common translocation event in Burkitt's lymphoma is a well-characterized example of oncogene activation. Translocations juxtapose *MYC* oncogene to the enhancer elements in the immunoglobulin loci on chromosomes 14q, 22q and 2p, thereby leading to transcriptional deregulation of *MYC* gene [22]. MYC protein, a transcription factor, plays an important role in cell cycle progression and cellular transformation. Amplification of *ERBB2* gene was found in some breast cancers, and is associated with poor clinical outcome [23]. Oncogene gain-of-function mutations often involve critical regulatory regions leading to continuously increased activity of the mutated protein. For example, the most common mutations of *BRAF* gene, amino acid change of a valine to a glutamate at codon 599, results in elevated kinase activity and transformation capability [24].

TSGs normally act to inhibit inappropriate cell growth and division, stimulate apoptosis, and repair DNA [25]. In many tumors, these genes are lost or inactivated by genetic or epigenetic alterations, including non-synonymous mutations, insertion or deletions of variable sizes, and epigenetic silencing [21]. Although for some TSGs haploinsufficiency (loss of only one allele) may contribute to carcinogenesis [26], mutation or loss of both alleles is generally required to facilitate tumor progression [21]. The first tumor suppressor gene *RB1* was identified by studies of the genetic mechanisms underlying retinoblastoma, a rare childhood retinal tumor. Besides the inherited mutation in an allele of *RB1* gene, a retinoblastoma patient normally has an additional mutation event or loss of heterozygosity (LOH) to inactivate the other allele [27]. Among TSGs, DNA repair genes are particularly important in prohibiting tumor development. These genes are responsible for correcting DNA mistakes during normal DNA replication or those induced by mutagens [21]. When these genes are inactivated, mutation rate will be elevated in other genes. Typical examples include *BRCA1* in breast and ovary cancers, and *RECQL4* in bone tumors.

1.1.2 A consistent cancer hallmark—genome instability

Although there are significant differences between cancer types, there are also properties shared by most if not all cancers. These properties, referred to as "cancer hallmarks", include but are not limited to self-sufficiency in growth signals, insensitivity to anti-growth signals, evasion of apoptosis, limitless replicative potential, sustained angiogenesis, and tissue invasion and metastasis [28]. Genome instability is a consistent characteristic crucial to the acquisition of the hallmarks of cancer [29], and plays important roles in tumor initiation and progression. Genome instability is typically subdivided into three categories: nucleotide instability, microsatellite instability and chromosomal instability [30]. Nucleotide instability is characterized by increased frequencies of base-pair mutations and small insertions and deletions. Microsatellite instability, which refers to the expansion and contraction of oligonucleotide repeats in microsatellites, is the consequence of impaired mismatch repair genes. Chromosomal instability, the most prevalent form of genome instability, refers to the changes in the structure and number of chromosomes in cancer cells compared with normal ones. Several mechanisms have been proposed to explain the source of genome instability: defects in DNA repair and mitotic checkpoint genes [30], telomere dysfunction [31], centrosome abnormality and replication stress [32].

1.2 The catalog of somatic mutations in cancer genomes

Somatic cells can accumulate mutations in DNA after conception. These mutations are collectively termed somatic mutations to distinguish them from germline mutations. Somatic mutations will not be transmitted to offspring, while, on the other hand, germline mutations do. Germline mutations account for 5-10% of cancers as high-penetrance variants observed in various hereditary cancer syndromes [33]. For example, germline mutations in *TP53* can cause Li-Fraumeni syndrome, which is characterized by development of a variety of cancer types including sarcomas, breast cancer, brain tumors and leukemia [34]. Inherited alterations in the *BRCA1* and *BRCA2* genes are responsible for the majority of hereditary breast and ovarian cancer syndromes, which are marked by increased risks of breast and ovarian cancer in women [35, 36]. Somatically acquired mutations are the most common cause of sporadic cancers. Somatic mutations include different types of DNA sequence changes: single nucleotide variants (SNVs), small insertions and deletions (indels), somatic copy number alterations (SCNAs), structural variations (SVs), small or large-scale size mutations in mitochondrial genomes, and epigenetic changes altering gene expression and chromatin structure [4]. Recently, sequencing of cancer genomes has led to the discovery of three new classes of complex chromosomal rearrangement: chromothripsis [37], chromoanasynthesis [38], and chromoplexy [39]. Chromothripsis is characterized by tens to hundreds of genomic rearrangements restricted to one or a few chromosomes and an oscillating pattern of DNA copy number states [37]. Based on the similarities shared between chromothripsis and complex genomic rearrangements, a new term of chromoanasynthesis (chromosome reconstitution or chromosome reassortment) was then proposed to describe better the underlying mechanisms [38]. Chromoplexy, on the other hand, is characterized by a closed chain of translocations involving multiple chromosomes, with little or no copy number alterations [39]. Cancer cells may also acquire DNA sequences from various types of viruses, such as human papilloma viruses in cervical cancer and Epstein-Barr viruses in Burkitt's lymphoma [40].

1.3 Technologies for exploring the mutational landscape of the cancer genome

Recent advances in high-throughput genomic technologies such as array comparative genomic hybridization (array-CGH), single nucleotide polymorphism (SNP) genotyping and next-generation sequencing (NGS) have revolutionized the study of cancer genomics by aiding the comprehensive characterization of somatic mutations in tumor cells [41]. Although early cancer genomics projects relied on array-based methods to investigate mRNA expression and DNA copy-number, the most recent large-scale projects such as The Cancer Genome Atlas (TCGA) [42] and the International Cancer Genome Consortium (ICGC) [43] employ a combination of SNP genotyping microarrays and NGS techniques [44].

1.3.1 Single nucleotide polymorphism microarrays

The human genome has been estimated to harbor approximately ten million or more SNPs. Two alleles of a SNP are often arbitrarily labeled as A and B for simplicity. Therefore, for each individual, there are three possible genotypes at each SNP site: AA, BB and AB. SNP microarrays were originally designed to genotype DNA sequences at thousands of SNPs across the human genome. Since their initial development, SNP arrays have been widely used in genome-wide association studies aimed at identifying disease risk loci. Nowadays, the inclusion of copy number polymorphism (CNP) probes in SNP microarrays has made them ideal to identify SCNAs and loss of heterozygosity in cancer [45]. The most commonly used SNP microarrays come from Affymetrix and Illumina. For example, Genome-Wide Human SNP Array 6.0 contains about 1 million SNP probes and 1 million CNP probes. Using these commercial microarrays, the landscape of SCNAs has been characterized across multiple cancer types, generating new insights into how focal SCNAs are frequently altered across several cancer types [46, 47].

1.3.2 Next-generation sequencing techniques

DNA sequencing technology was first developed in 1977 by Frederick Sanger and Walter Gilbert based on different methods: the chain-termination method (known as Sanger sequencing) [48] and the chemical degradation method [49]. A decade later, Applied

Biosystems introduced the first automated sequencing instruments, which were based on capillary electrophoresis and were the main workhouses for the Human Genome Project (HGP) [50]. Using the first generation sequencing technique, the HGP took more than a decade and cost about 3 billion US dollars [51]. The need for faster, more accurate, higher throughput, and cheaper sequencing instruments stimulated the emergence of NGS technologies [52]. NGS technologies are distinct from the first generation sequencing methods in terms of massively parallel analysis, high throughput, and relatively short reads [52, 53]. Three most typical NGS technologies are pyrosequencing method from 454 Life Sciences (purchased by Roche in 2007), sequencing-by-synthesis from Solexa (acquired by Illumina in 2007), and Sequencing by Oligo Ligation Detection from Applied Biosystems (purchased by Life Technologies in 2008 and Life Technologies was then acquired by Thermo Fisher Scientific in 2014) [53]. Different NGS technologies have advantages and drawbacks with regard to read length, throughput, run time, error rate and cost (reviewed in [53]).

As one of the most widely adopted technologies in the NGS industry, Illumina Solexa sequencing provides the highest throughput and the lowest per-base sequencing cost [53]. The Illumina workflows consist of four steps: library preparation, cluster generation, sequencing and data analysis [54]. For library preparation, DNA or cDNA is randomly fragmented into small sizes and each fragment ligated to an adapter at both ends, followed by polymerase chain reaction (PCR) amplification and gel purification. During cluster generation, the library is loaded into a flow cell and the fragments are bound at one end to a solid surface coated with oligonucleotides complementary to the adapters used in the library preparation step. The free end of each fragment hybridizes to a complementary adapter to initiate complementary strand synthesis, which is termed as bridge amplification. Illumina's sequencing-by-synthesis detects single bases as they are introduced into growing DNA strands by using a reversible terminator-based method. There are two commonly used sequencing strategies, single-end sequencing (SES) and paired-end sequencing (PES). SES involves sequencing DNA from only one end, while PES involves sequencing both ends of the DNA fragments and assigning the forward and reverse read pairs [54]. Compared with SES, PES produces twice the number of reads and allows more accurate read alignment. These advantages make PES more suitable for detecting some types of somatic mutations, such as SVs (see below). In the data analysis step, the large NGS data sets demand bioinformatics tools for data analysis and management. For example, the relatively short reads required the development of new alignment

tools [53]. Furthermore, the bioinformatics algorithms used in NGS data analysis should account for biases introduced during the library preparation and sequencing, such as GC content bias [55] and mappability bias [56].

NGS has a series of applications to cancer genomic studies, which include sequencing an entire genome (whole-genome sequencing, WGS), the coding genomic regions (whole-exome sequencing, WES), and the transcriptome (RNA sequencing, RNA-seq) [52, 57]. As coding sequences constitute only 1-2% of the human genome, the cost for WES is lower than WGS. Despite its much higher cost, WGS provides additional information on structural and non-coding variants, which cannot be captured by WES. In addition to quantifying gene expression profiles, RNA-seq can detect alternative splicing and fusion transcripts [58]. NGS can also be applied to cancer epigenomic studies to study epigenetic alterations, DNA methylation changes and histone modifications [52–54]. These technologies include Bisulfite Sequencing (Bisulfite-seq) and Chromatin Immunoprecipitation followed by Sequencing (ChIP-seq). The combination of these NGS technologies will provide us a high-resolution view of the mutational landscape of cancer genome.

1.4 Detection of somatic mutations

Somatic aberrations acquire by tumor cells at different stages of the disease may contain information crucial for understanding the mechanisms of tumor development, progression, metastasis and relapse. To investigate the cancer genome with NGS technologies, it is common practice to simultaneously sequence genomic information from tumor and matched normal (often blood) samples from the same patient. The reads from these two matching samples are aligned to the reference genome using alignment tools (such as Burrows-Wheeler Alignment [59], reviewed in [60]) and differences between the normal genome and the tumor genome characterized [61]. To detect somatic mutations, an intuitive approach would be analyses of tumor and normal independently followed by subtractions of tumor and normal variant calls [62]. Specifically, mutations observed only in the tumor genome but absent in the normal genome are characterized as somatic mutations unique to the tumor sample. It has been found that direct comparison of the aligned reads from the matched two samples yields better results in terms of sensitivity and specificity [61]. However, the detection of somatic alterations from aligned reads is not an easy task. Both sequencing and alignment introduce a number of errors and biases, such as sequencing errors, PCR duplicates, strand bias and ambiguities in short read mapping

[63]. Other confounding factors include tumor sample heterogeneity and tumor impurity contaminated by matched normal samples [57]. In the past decade, many algorithms and softwares have been developed to detect SNVs, small indels, SCNAs, SVs and gene fusions (some popular tools are listed in Table 1.1).

1.4.1 SNV detection

SNVs are the most common alterations in tumor genomes. The last decade has witnessed the development of algorithms to detect SNVs in cancer genomes: SomaticSniper [64], JointSNVMix [65], MuTect [66], Strelka [67], LoFreq [68], VarScan 2 [69] and VarDict [70] (listed in Table 1.1). Most of these methods consider only a subset of errors and biases described above. For example, VarScan2 employs empirically derived filtering parameters, including read position, strandedness, and average mapping quality between reference and variant reads to exclude candidate variants resulting from sequencing or alignment artifacts [69]. MuTect was specifically designed to detect low allele fraction variants due to either tumor heterogeneity or normal cell contamination [66]. It utilizes filters to remove false positives with characteristics corresponding to strand bias or poor mapping quality. Although a number of comparative studies of SNV callers are available [71, 72], there are no concordant recommendations of tools optimally balancing sensitivity and specificity. The varying performances based on different datasets suggest that multi-caller strategies are favorable [57, 63]. Of noteworthy, several machine-learning algorithms, such as MutationSeq [61] and SomaticSeq [73] have been developed. These algorithms trained their classifiers on a series of sequence features from a training dataset, then classifiers were used on a target dataset to distinguish true somatic alterations from false positives. Incorporating the strengths of different somatic mutation detection algorithms, these methods report higher accuracy and robustness [73].

Table 1.1: Computational tools for detecting somatic mutations

Tools	Description	Mutation type	Reference
SomaticSniper	Bayesian probability with posterior filtering	SNVs	[64]
JointSNVMix	Probabilistic graphical model with pre-filtering	SNVs	[65]
MuTect	Bayesian classifier with pre- and post-filtering	SNVs	[66]
MuSE	Markov substitution model for molecular allelic evolution	SNVs	[74]
Pindel	Pattern growth learning approach	Indels	[75]
Dindel	Bayesian model accounting for sequencing, base-calling and mapping errors	Indels	[76]
Indelocator	Information not available	Indels	[77]
Strelka	Bayesian approach with posterior filtering	SNVs, Indels	[67]
LoFreq	Statistical model for sequencing error biases	SNVs, Indels	[68]
SomaticSeq	Ensemble approach with machine learning	SNVs, Indels	[73]
VarScan 2	Fisher exact test, filtering and FDR correction	SNVs, Indels, SCNAs	[69]
VarDict	Fisher exact test with post-filtering	SNVs, Indels, SVs	[70]
GAP[1]	Pattern recognition of segmented and smoothed bi-dimensional profile	SCNAs	[78]
GenoCNA[1]	Continuous time HMM with discrete states	SCNAs	[79]
PICNIC[1]	HMM algorithm with preprocessing transformation	SCNAs	[80]
ASCAT[1]	Goodness-of-fit score of candidate solutions of tumor ploidy and tumor purity	SCNAs	[81]
OncoSNP[1]	Single unified Bayesian framework.	SCNAs	[82]

Continued on next page

Table 1.1 – *Continued from previous page*

Tools	Description	Mutation type	Reference
GPHMM[1]	Global parameter HMM	SCNAs	[83]
ABSOLUTE[1]	Optimization of logarithmic scores	SCNAs	[84]
SegSeq[2]	Local change-point analysis with a subsequent merging procedure	SCNAs	[85]
CNAseg[2]	HMM segmentation with read depth variability correction	SCNAs	[86]
readDepth[2]	CBS algorithm with GC-content and mappability correction	SCNAs	[87]
BIC-seq[2]	Minimizing BIC approach with no read distribution assumption	SCNAs	[88]
Control-FREEC[2]	Sliding window approach with corrections of GC-content and mappability biases	SCNAs	[89]
ExomeCNV[2]	CBS algorithm with an assumption of read Gaussian distribution	SCNAs	[90]
CNAnorm[2]	CBS algorithm with correction of normal cell contamination and tumor aneuploidy	SCNAs	[91]
Patchwork[2]	CBS algorithm with tumor purity and ploidy estimation	SCNA	[92]
HMMcopy[2]	HMM segmentation with GC-content and mappability correction	SCNAs	[93]
OncoSNP-SEQ[2]	HMM segmentation accounting for tumor purity, ploidy and heterogeneity	SCNAs	[94]
CLImAT[2]	Integrated HMM algorithm accounting for tumor purity and ploidy	SCNAs	[95]
PEMer	Read pair based approach with simulation based error models	SVs	[96]
BreakDancer	Read pair based approach	Indels, SVs	[97]
VariationHunter	Read pair based approach	SVs	[98]
SVDetect	Integrated method of read pair and read depth	SVs	[99]
DELLY	Integrated method of read pair and split reads	SVs	[100]

Continued on next page

Table 1.1 – *Continued from previous page*

Tools	Description	Mutation type	Reference
PRISM	Integrated method of read pair and split reads	SVs	[101]
HYDRA	Integrated method of read pair and local assembly	SVs	[102]
CREST	Integrated method of split reads and local assembly	SVs	[103]
cortex_var	*De novo* assembly method using colored de Bruijn graphs	SVs	[104]
Meerkat	Integrated method of read pair, split reads, and assembly	SVs	[105]
LUMPY	Integrated method of read pair, split read and read depth, as well as prior knowledge	SVs	[106]
MapSplice	Gene fusion detection from paired-end or single-end RNA-seq data	Gene fusions	[107]
FusionSeq	Gene fusion detection from paired-end RNA-seq data	Gene fusions	[108]
TopHat-Fusion	Gene fusion detection from paired-end or single-end RNA-seq data	Gene fusions	[109]
SnowShoes-FTD	Gene fusion detection from paired-end RNA-seq data	Gene fusions	[110]
ShortFuse	Gene fusion detection from paired-end RNA-seq data	Gene fusions	[111]
FusionMap	Gene fusion detection from WGS or RNA-seq data (both paired and single end)	Gene fusions	[112]
FusionHunter	Gene fusion detection from paired-end RNA-seq data	Gene fusions	[113]
deFuse	Gene fusion detection from paired-end RNA-seq data	Gene fusions	[114]
Comrad	Integrated gene fusion detection from paired-end RNA-seq and WGS data	Gene fusions	[115]
ChimeraScan	Gene fusion detection from paired-end RNA-seq data	Gene fusions	[116]
nFuse	Integrated gene fusion detection from paired-end RNA-seq and WGS data	Gene fusions	[117]
SOAPfuse	Gene fusion detection from paired-end RNA-seq data	Gene fusions	[118]

Continued on next page

Table 1.1 – *Continued from previous page*

Tools	**Description**	**Mutation type**	**Reference**
INTEGRATE	Integrated gene fusion detection from paired-end RNA-seq and WGS data	Gene fusions	[119]

[1]for SNP array data; [2]for NGS data.

HMM: Hidden Markov Model; CBS: Circular Binary Segmentation; BIC: Bayesian Information Criterion.

1.4.2 Indel detection

Indel detection lags behind the calling of SNVs in terms of sensitivity and specificity [76]. The challenge lies in the lower frequencies of indels than those of SNVs [57, 76] and mapping difficulties of reads overlapping the indel sequence, especially when indels are located in short tandem repeats [76, 120]. Despite these challenges, there are several tools (listed in Table 1.1) available to identify indels from cancer genome sequencing data. These are generally based on approaches that include split reads, gapped alignment and *de novo* assembly [57]. Split read tools (e.g., Pindel [75]) realigned soft-clipped reads to infer indels, however, it is still difficult for these methods to distinguish low frequency true indel calls from false positives derived from alignment errors. Gapped alignment-based tools, such as Dindel [76], Strelka [67] and LoFreq [68], performed local realignments to detect indels. A major drawback of these methods is the reduced sensitivity to detect longer (>30 bp) indels [121]. *De novo* assembly approaches have been developed for indel discovery, including Scalpel [121]. None of the tools described above are able to predict indels of full size spectrum. Therefore, a hybrid algorithm integrating split reads, gapped alignment and *de novo* assembly approaches has recently been designed to detect indels with increased sensitivity [122].

1.4.3 SCNA detection

SCNAs affect a much larger part of the cancer genome than SNVs and indels. Array-CGH [123, 124], SNP genotyping and NGS have been used for detecting SCNAs in cancer. Since SNP arrays allow for the estimation of absolute copy number and allelic content, they have replaced array-CGH and have been widely used in TCGA and ICGC projects. NGS of tumor and matched normal samples enables the efficient detection of SCNAs at base pair resolution. Some widely-used SCNA detecting tools based on SNP arrays and NGS are listed in Table 1.1.

SNP arrays of Illumina and Affymetrix simultaneously measure copy number and allelic ratios at many SNP loci in the genome. For each SNP probe, the log R ratio (LRR) reflects the total signal intensity for both alleles, and the B allele frequency (BAF) is an estimate of the relative proportion of one of the alleles with respect to the total signal intensity. Based on these two complementary information, several computational algorithms have been proposed to detect SCNAs in cancer (listed in Table 1.1). Compared

with CNV detection in germline samples (e.g., QuantiSNP [125] and PennCNV [126], reviewed in [127]), SCNA detection in cancer is much more difficult for several reasons. First, widespread aneuploidy observed in cancer [128] violates the assumption of a baseline copy number of two in germline samples, and the resulting LRR baseline shift affects copy number assignment [129]. Second, contamination from adjacent normal cells causes the LRR and BAF values to converge towards a diploid state proportionally to the degree of contamination [129]. Third, intra-tumor heterogeneity [130] further complicates LRR and BAF signals. Some of the tools listed in Table 1.1 (such as GAP [78], OncoSNP [82] and ABSOLUTE [84]) take into consideration tumor aneuploidy, normal cell contamination and intra-tumor heterogeneity, while others (such as GenoCNA [79], PICNIC [80], ASCAT [81] and GPHMM [83]) account for only one or two factors of them. Although there is disagreement on the performance of GPHMM, a comparative study [129] showed that GAP generally performed better in both simulated and real genotyping data.

NGS provides a feasible alternative to SNP microarrays for detecting SCNAs. Since most studies classify SCNAs as one type of SVs (e.g., [131]), we consider only tools specifically for SCNA detection in this section, and summarize algorithms for SVs in the next section. Read depth information of NGS can be used to estimate copy number, with the underlying hypothesis being that the read depth of a genomic region is positively correlated with the copy number of the region [132]. Compared with germline CNV detection tools (e.g., CNV-seq [133] and CNVnator [134]), SCNA calling algorithms need to account for the special characteristics of SCNAs as well as tumor impurity, aneuploidy and heterogeneity [135]. Table 1.1 lists a number of widely used tools among the research community, of which some account for inherent bias from NGS short reads (e.g., mapping bias and GC-content bias), and others further take into consideration tumor impurity contaminated by normal cells, tumor aneuploidy and tumor heterogeneity. Though comparative studies [136–138] provide guidance for tool selection, lack of a gold standard makes comprehensive benchmarking less reproducible and concordant. Therefore, better benchmark datasets are urgently needed to evaluate different algorithms and further advance the development of new tools [135].

1.4.4 SV detection

SVs account for more polymorphism than SNVs as measured by total number of base pair changes. A number of tools have been developed to detect SVs from NGS data. These

detection methods can be divided into five different strategies: (1) read pair, (2) split-read, (3) read depth, (4) assembly, and (5) combinatorial methods of the above approaches [131, 132, 139]. Read depth based methods have already been described above (SCNA detection section), and the other approaches are discussed in this section. Several popular tools are summarized in Table 1.1, and please refer to comprehensive reviews [132, 139] for an exhaustive list. The read-pair methods are only applicable to paired-end reads but not single-end reads. In paired-end sequencing, the DNA fragments from the same library preparation protocol exhibit a specific insert size distribution. Read-pair methods utilize discordantly mapped paired-reads, in which the mapping span and/or orientation are inconsistent with the reference genome, to identify SVs [131, 132]. The read pair method, the most widely used approach, was applied in PEMer [96], BreakDancer [97], VariationHunter [98], and many other softwares. It can efficiently identify many types of SVs, including insertions, deletions, tandem duplications, inversions, and translocations, but only report approximate breakpoint locations [132, 139]. The split-read methods localize the breakpoints of a SV on the basis of a "split" signal, in which one read from a read pair is mapped to the reference genome while the other fails to map or only partially maps to the genome [131, 132]. The split-read methods can provide base resolutions of SV breakpoints, but are not sensitive to certain types of SVs, i.e., inversions and translocations [139]. As described above in the SCNA detection sections, the read depth methods can only detect duplications and deletions. The assembly methods first reconstruct contigs from short reads and then identify all forms of SV by comparing the assembly contigs with the reference genome [131, 139]. Although in their infancy, the assembly methods provide an unbiased approach to discover SVs and other alterations, as illustrated in cortex_var [104]. As discussed above, each approach has both advantages and drawbacks. Consequently, to overcome the inherent limitations of each approach, one possible solution would be incorporating multiple methods to improve sensitivity and specificity [131, 132, 139]. These combinatorial methods integrated two to four approaches, such as SVDetect [99], DELLY [100], PRISM [101], HYDRA [102], CREST [103], Meerkat [105], and LUMPY [106] (Table 1.1).

1.4.5 Gene fusion detection

Gene fusions may result from SVs, including insertions, deletions, inversions and translocations. Widespread across many cancer types, gene fusions provided fundamental insights into tumorigenesis, and have been successfully used for cancer diagnosis and treat-

ment [140]. Traditionally detected by fluorescence in situ hybridization or DNA microarrays, the advancement of NGS provides an unbiased approach to identify gene fusions either at DNA or RNA level. Leveraging the strengths of high-throughput NGS, a number of tools have been developed to detect gene fusions in the past several years. Table 1.1 lists some popular tools, and a complete list of detection methods can be found in a recent comprehensive review [141]. As demonstrated in Table 1.1, WGS and RNA-seq are two major NGS data used for gene fusion characterization. Although WGS can provide a comprehensive and unbiased view of gene fusions, its higher cost and more intensive computational analysis hinders its application in cancer genomic studies [141, 142]. On the other hand, RNA-seq only sequences about 2% of the whole genome that is transcribed and spliced into mature mRNA. The relatively lower cost and shorter data processing time make RNA-seq popular for gene fusion detection [141, 142]. Recently, comparative studies of detection tools revealed that small overlaps of the fused genes were detected by different tools [143–145], which could be due to the high number of false positives reported by most tools [143, 144]. To reduce false positives, one possible solution would be integrating RNA-seq and WGS data as applied in Comrad [115], nFuse [117] and INTEGRATE [119] to increase the specificity. An alternative solution is to design a meta-caller to combine tools of top performance so as to reprioritize candidate fusion genes [145].

1.5 Identification of driver mutations, genes and pathways

Cancer genome sequencing projects have revealed thousands of somatic mutations in coding and non-coding genomic regions. However, not all somatic alterations in a cancer genome are involved in cancer development. Indeed, only a subset of these mutations drive tumorigenesis and progression (driver mutations), whereas the remainder are non-functional random events caused by the general genomic instability in cancer cells (passenger mutations) [4]. Driver mutations have dramatic impacts on the molecular functions (gain- or loss-of-function) of gene products important for tumor initiation and progression, and provide growth advantages to cancer cells [4, 146]. Undoubtedly, the identification of driver mutations and driver genes would provide new insights into the underlying mechanism of tumorigenesis and the development of new therapeutic targets for cancer treatment. A challenge is to distinguish the relatively small number of driver mutations

from the large number of passenger mutations. There are many computational and statistical algorithms presently available to identify likely driver mutations, genes, and pathways from somatic variants across a cohort of cancer samples. According to their function, these tools can be divided into four general types: variant mapping and annotation, variant effect prediction, driver gene detection, and driver pathway identification [57, 63, 146–148]. Some popular tools for each category are listed in Table 1.2.

1.5.1 Variant mapping and annotation

After the detection of somatic variants, our primary goal is to map them onto annotated functional genomic features and determine their impacts on protein-coding and non-coding transcripts, transcription factor binding sites, and other potential regulatory elements [146]. We defined functional elements characterized by the Encyclopedia of DNA Elements (ENCODE) Consortium as regulatory features, including transcription binding sites, regions of open chromatin, DNase I hypersensitive sites (DHSs), histone modification and chromatin interactions [149–151]. This step also involved a comparison of these variants with databases of known variants, such as dbSNP [152], 1000 genomes [153], Catalog Of Somatic Mutations In Cancer (COSMIC) [154], the Human Gene Mutation Database (HGMD)[155] and the Database of Genomic Variants [156]. There are a variety of tools available to map and annotate variants to genomic features (listed in Table 1.2). Among them, VAT [157] and Oncotator [158] provide annotations of variants at transcript and protein levels, while ANNOVAR [159] and SnpEff [160] have additional support to include annotation of regulatory features. The Ensembl Variant Effect Predictor (VEP) [161] and AnnTools [162] can map and annotate all kinds of somatic variants (SNVs, indels, SCNAs, and SVs), while VARIANT [163] and CRAVAT [164] only consider SNVs.

Table 1.2: Computational tools for detecting driver mutations, genes and pathways

Tools	Function	Description	Mutation type	Reference
ANNOVAR	Variant annotation	Transcripts, protein, and regulatory feature annotation	SNVs, Indels, SCNAs, SVs	[159]
VEP	Variant annotation	Transcripts, protein, and regulatory feature annotation	SNVs, Indels, SCNAs, SVs	[161]
AnnTools	Variant annotation	Transcripts, protein, and regulatory feature annotation	SNVs, Indels, SCNAs, SVs	[162]
SnpEff	Variant annotation	Transcripts, protein, and regulatory feature annotation	SNVs, Indels	[160]
VARIANT	Variant annotation	Transcripts, protein, and regulatory feature annotation	SNVs	[163]
VAT	Variant annotation	Transcripts and protein annotation	SNVs, Indels, SCNAs, SVs	[157]
Oncotator	Variant annotation	Transcripts and protein annotation	SNVs, Indels	[158]
CRAVAT	Variant annotation	Transcripts and protein annotation	SNVs	[164]
SIFT	Functional prediction	Conservation-based prediction	nsSNVs	[165, 166]
MutationAssessor	Functional prediction	Conservation-based prediction	nsSNVs	[167]
PROVEAN	Functional prediction	Alignment-based score	nsSNVs, ifIndels	[168]
MAPP	Functional prediction	Physicochemical-property-based prediction	nsSNVs	[169]
LS-SNP/PDB	Functional prediction	Protein-structure-based prediction	nsSNVs	[170]
transFIC	Functional prediction	Transformed FI score for cancer	nsSNVs	[171]
Condel	Functional prediction	Consensus deleteriousness score of FI scores	nsSNVs	[172]
CanPredict	Functional prediction	Combined prediction based on SIFT, Pfam and GOSS	nsSNVs	[173]
PolyPhen-2	Functional prediction	Naïve Bayes classifier based on structure and alignment	nsSNVs	[174]

Continued on next page

Table 1.2 – *Continued from previous page*

Tools	Function	Description	Mutation type	Reference
CHASM	Functional prediction	Random forest classifier based on diverse features	nsSNVs	[175]
VEST	Functional prediction	Machine learning-based classifier	nsSNVs	[176]
VEST-Indel	Functional prediction	Machine learning-based classifier	if/fsIndels	[177]
SIFT Indel	Functional prediction	Decision tree-based algorithm	if/fsIndels	[178]
FATHMM	Functional prediction	Hidden Markov Models algorithm	nsSNVs, ncSNVs	[179]
MutationTaster	Functional prediction	Naïve Bayes classifier	cSNVs, inSNVs, Indels	[180, 181]
CADD	Functional prediction	Combined Annotation Dependent Depletion	SNVs, Indels	[182]
MuSiC	Driver gene detection	Recurrence-based prediction	SNVs, Indels	[183]
MutSigCV	Driver gene detection	Recurrence-based prediction with variable BMR	SNVs, Indels	[184]
InVex	Driver gene detection	Recurrence-based prediction	SNVs, Indels	[185]
Simon	Driver gene detection	BMR, FI and genetic code redundancy	SNVs, Indels	[186]
OncodriveFM	Driver gene detection	Functional-mutation-based prediction	nsSNVs	[187]
OncodriveCLUST	Driver gene detection	CLUST-based prediction	nsSNVs	[188]
ActiveDriver	Driver gene detection	ACTIVE-based prediction	nsSNVs	[189]
OncodriveFML	Driver gene detection	FI bias in coding and non-coding regions	SNVs	[190]
GSEA	Pathway analysis	Gene Set Enrichment Analysis	SNVs, Indels, SCNAs	[191]
CaMP-GSEA	Pathway analysis	GSEA with Cancer Mutation Prevalence scores	SNVs, Indels, SCNAs	[192]
PathScan	Pathway analysis	Probability model for mutation-enriched pathways	SNVs, Indels, SCNAs	[193]

Continued on next page

Table 1.2 – *Continued from previous page*

Tools	**Function**	**Description**	**Mutation type**	**Reference**
HotNet	Pathway analysis	Heat-diffusion model with known interaction network	SNVs, Indels, SCNAs	[194]
HotNet2	Pathway analysis	Heat-diffusion model with known interaction network	SNVs, Indels, SCNAs	[195]
NetBox	Pathway analysis	Finding significantly mutated network modules	SNVs, Indels, SCNAs	[196]
PSMP	Pathway analysis	Exclusivity based pairwise search for mutational pattern	SNVs, Indels, SCNAs	[197]]
MEMo	Pathway analysis	Driver network identification based on exclusivity	SNVs, Indels, SCNAs	[198]
Dendrix	Pathway analysis	*De novo* driver pathway identification	SNVs, Indels, SCNAs	[199]
Multi-Dendrix	Pathway analysis	*De novo* driver pathway identification	SNVs, Indels, SCNAs	[200]
MDPFinder	Pathway analysis	*De novo* driver pathway identification	SNVs, Indels, SCNAs	[201]
RME	Pathway analysis	*De novo* driver pathway identification	SNVs, Indels, SCNAs	[202]

VEP, Variant Effect Predictor; nsSNVs, non-synonymous SNVs; ifIndels, in-frame Indels; FI, Functional Impact; GOSS, Gene Ontology Similarity Score; if/fsIndels, in-frame and frame-shift Indels; ncSNVs, non-coding SNVs; cSNVs, coding SNVs; inSNVs, intronic SNVs; CADD, Combined Annotation Dependent Depletion; BMR, Background Mutation Rate.

1.5.2 Functional prediction of somatic variants

The exact determination of variant functional effects relies on labor-intensive *in vivo* biological and clinicopathological experiments [203]. Alternatively, *in silico* methods can attempt to predict the effects of variants on the functions of proteins or regulatory elements. Because non-synonymous variants (changes amino acid of protein-coding genes) account for approximately half of the disease-causing mutations deposited in Online Mendelian Inheritance in Man (OMIM) [204] and HGMD [155], they are particularly the subject of recently developed computational methods [147]. These computational approaches typically use the Physicochemical properties of amino acids, evolutionary conservation information (multiple sequence alignments), as well as information about the role of amino acid side chains in three-dimensional protein structure [146]. Based on the underlying methodology, these methods can be classified as "direct methods" or "machine learning methods" [146, 167, 205] (Table 1.2). The direct methods assess the effect of a mutation by a phenomenological score computed based on a particular theoretical model [146, 167]. The machine learning methods use relevant properties (e.g., size and polarity) of both the original and mutant residues, structural information (e.g., surface accessibility and hydrogen bonding), evolutionary conservation and other features, and train these features to distinguish functionally deleterious variants from nonfunctional neutral ones [146, 167]. As listed in Table 1.2, most of these tools can only assess the functional effects of non-synonymous SNVs , for instance, SIFT [165, 166], MutationAssessor [167], PolyPhen-2 [174] and some other extend underlying algorithm to include in-frame and/or frame-shift indels, such as PROVEAN [168], VEST-Indel [177] and SIFT Indel [178]. Using functionally validated missense mutation data collected from literature and database, Martelotto *et al.* [206] benchmarked the performance of 15 algorithms including SIFT [165, 166], MutationAssessor [167], PROVEAN [168], Condel [172], PolyPhen-2 [174], CHASM [175], VEST [176], FATHMM [179] and MutationTaster [180] (Table 1.2). The results showed that the prediction accuracy varies among different tools and the combination of different algorithms can significantly improve the overall accuracy [206].

Most of the tools described above focus exclusively on non-synonymous mutations, with the underlying assumption being that coding mutations do not change amino acid sequence (synonymous mutations) and non-coding mutations are passenger mutations. However, several pilot studies have revealed the important roles of synonymous and non-

coding variants in tumorigenesis [207–211]. Supek *et al.* showed that synonymous mutations frequently change exonic motifs that regulate RNA splicing resulting in abnormal oncogene splicing in tumors [208]. The validation of splice-site mutation consequences requires additional information from transcriptome, as implemented in PVAAS [212] and Veridical [213]. In two side-by-side papers published in the February 2013 issue of Science, Huang *et al.* [209] and Horn *et al.* [210] reported somatic and germline mutations in the core promoter of telomerase reverse transcriptase (*TERT*) gene, which generate novel binding motifs for E-twenty-six (ETS) transcription factors resulting in up to twofold increase in transcription. Therefore, non-coding somatic mutations may represent an alternative oncogenesis mechanism through regulatory potential. With the decreasing costs of whole-genome sequencing, cancer genome projects provide us a wealth of data to examine the consequences and clinical significance of non-coding mutations in cancer. Currently, only a few tools are capable of predicting the functional consequences of non-coding SNVs (e.g., FATHMM [179], MutationTaster [180, 181] and CADD [182]), however, in the coming years we foresee the development of more algorithms to illuminate the crucial role of non-coding somatic mutations in cancer [214].

1.5.3 Detection of driver genes

A driver gene harbors driver mutations, but could also contain passenger mutations [215]. A driver mutation confers selective growth advantages to cancer cells and is positively selected during the evolution of the cancer. An important goal of cancer genomics analyses is the characterization of cancer driver genes [4]. The main strategy generally used for this task is to search for signals of positive selection across a cohort of tumor samples. The most common methods for identifying driver genes is evaluating whether a gene is mutated more frequently than expected from the background mutation rate (BMR) (recurrence). Algorithms utilizing this approach include MuSiC [183], MutSig [184], and InVex [185]. However, a major challenge is to correctly estimate the BMR so as to reduce the number of false positives [146]. The estimation of the BMR takes into account factors such as gene length, mutation type (transitions and transversions), the nucleotide context, DNA region replication timing [216], and gene expression level. Although this approach is successful in detecting frequently mutated driver genes, it rarely detects driver genes mutated at very low frequencies. Recurrence-based methods have also been designed to identify genes that are frequently targeted by SCNAs, for example, GISTIC [217, 218] and RAE [219]. Applying GISTIC to copy-number profiles from a large collection of

tumor samples, it was revealed that some of the recurrently altered regions contain oncogenes or tumor suppressor genes, while most of them have no known cancer genes [46, 47]. The second approach relies on other signals of positive selection, such as a bias accumulation of functional mutations (FM bias) or a clustering of mutations in certain regions or functional sites (phosphorylation sites) of the protein sequences (CLUST bias and ACTIVE bias). This approach has been implicated in Simon [186], OncoDriveFM [187], OncoDriveFML [190], OncoDriveCLUST [188], and ActiveDriver [189]. Advantages of this approach include its independence of the BMR estimation and its ability to detect driver genes with low mutation frequency. However, the performance of some methods (e.g., OncoDriveFM and OncoDriveFML) may be affected by the bias induced by the metrics used to predict the putative impact of somatic variants on protein function. Of noteworthy, OncoDriveFML is able to identify putative drivers in both coding and non-coding genomic regions through the computation of a local FM bias [190]. As discussed above, all the methods have particular biases and shortcomings [146]. Therefore, the combination of several complementary methods allows the balancing of their pros and cons in order to identify a comprehensive and reliable list of driver genes [146, 147, 220].

1.5.4 Identification of driver pathways

Detection of recurrent or driver mutations from a large number of genomic variants greatly reduces the data complexity and makes it possible to identify inherent signaling pathways and biological processes [44]. Pathway and network approaches can provide enhanced understanding of tumorigenesis based on the observation that driver mutations tend to affect genes in signaling, regulatory and metabolic pathways [21, 221]. The pathway view of cancer mutation may explain the phenomenon of mutational heterogeneity that no two tumors, even from the same tumor type, harbor exactly the same set of somatic mutations [221, 222]. In individual tumors, a particular pathway can be perturbed by mutations in multiple genes of the pathway [29, 215]. Collectively, across a cohort of tumors of a cancer type, only a few genes in the pathway are frequently altered and many more rarely mutated. Because the recurrence-based detection algorithms tend to miss these rarely mutated driver genes, pathway and network approaches are generally preferred [63, 148]. Tools for the identification of driver pathways and networks can be classified into three types: gene set analysis methods, interaction network methods, and *de novo* identification methods [57, 63, 148](Table 1.2).

Gene set analysis methods examine the overlap between lists of mutated genes and predefined gene sets from databases of known pathways or other functional groupings, such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) [223] and Gene Ontology (GO) [224]. This approach has been successfully used in gene expression analysis to identify differentially expressed pathways. For example, Gene Set Enrichment Analysis (GSEA) assess whether a group of genes has more high-ranking genes than would be expected by chance [191]. GSEA has been used together with different rank scores or variables to determine the enrichment of mutations in particular pathways or functional groups. CaMP-GESA uses Cancer Mutation Prevalence (CaMP) scores to rank genes [192]. PathScan scores each gene at the patient level and also accounts for the mutation probability variations in gene length [193]. Compared with single gene-oriented methods, gene set analysis ones are more interpretable and statistically powerful [148]. However, one major drawback of gene set analysis methods is that they consider all genes in a single pathway as equally important but ignore the topology of gene interactions [63].

To overcome this limitation, interaction network methods are to examine mutations on large-scale protein-protein interaction networks deposited in databases such as the Human Protein Reference Database (HPRD) [225], Reactome [226] and STRING [227]. The primary goal of this type of approach is to identify significantly mutated subnetworks in the context of a large interaction network. HotNet [194] employs a heat diffusion model to build an "influence graph" including neighborhood information for mutated genes and then identifies recurrently altered subnetworks in more sample than would be expected by chance. Applied to several TCGA cancer types, HotNet identified the significantly mutated Notch pathway in high-grade ovarian serous adenocarcinomas [228] and the SWI/SNF chromatin remodeling complex in clear cell renal cell carcinomas [229]. HotNet has been recently updated to HotNet2 [195] to identify perturbed pathways and protein complexes across pan-cancer types. An alternative method, NetBox [196] is based on the hypothesis that inactivation of multiple functional modules in interaction network leads to cancer phenotype. Notably, MEMo [198] is another approach to find mutated subnetworks based on the observation of mutational mutual exclusivity in various cancer types. The reasoning behind mutual exclusivity is that across a cohort of patients, co-occurring altered genes tend to be in different pathways and mutually exclusive ones in the same pathway [197]. The methods described above require prior knowledge of protein interaction networks, which is far from complete now, and has limited capability to discover novel combinations of mutated genes [63, 148].

To identify novel combinations of mutated genes, an ideal solution is to test the significance of recurrent mutations of all possible combinations of genes. This *de novo* approach is computationally impossible because there would be a huge number of possible combinations of gene sets to evaluate. The discovery of mutual exclusivity pattern [21, 197] in cancer largely reduced this number to one computationally plausible. A few tools have been developed to identify putative driver pathways, such as Dendrix [199], Multi-Dendrix [200], MDPFinder [201] and RME [202]. Although *de novo* approaches avoid bias introduced by prior information in gene set and network approaches, one of their disadvantage is that they focus on a subset of functional combinations of genes and cannot characterize all of such combinations [63].

1.6 Generation mechanism of somatic mutations in cancers

The collective somatic mutations observed in a cancer are the products of DNA damage and DNA repair processes that have been operative throughout the development of cancer. The recent deluge of cancer genomics data provides an unprecedented opportunity for the discovery of the generation mechanisms for somatic alterations in cancer. Below we summarize the new insights into the generation mechanisms for different types of somatic mutations.

1.6.1 SNVs

Studies of mechanistic bases for germline point mutations can provide clues to the underlying mechanisms for SNVs, as the patterns of SNVs in cancer genomes have similarities (and differences) to those of germline SNPs [230]. These studies typically analyzed human nucleotide diversity and DNA sequence divergence between human and other mammals (e.g., chimpanzee) [231]. Pilot researches of germline mutations together with the availability of cancer sequencing data have stimulated studies investigating the generation mechanisms of SNVs in cancer. These studies offered new insights by analyzing associations between somatic mutation rates and genomic features, and by investigating the patterns of somatic mutations (mutation signature) caused by different mutational processes.

Germline point mutation rate is not constant across the genome [232]. Such variation occurs on different scales, including sequence context effects (a best-known example is the hypermutability of a methylated cytosine in a CpG dinucleotide), variation within chromosomes and variation between chromosomes (such as between sex chromosomes and autosomes) [230, 233]. Although the reasons for the mutation rate variation are poorly understood, a number of genomic factors (recombination, replication timing, chromatin structures and nucleosome occupancy) have been found to affect the germline mutation rate. A positive correlation between nucleotide diversity and recombination rate have been observed, suggesting a mutagenic role of recombination through incorrect repair of double stand breaks [234]. Hellmann *et al.* [235] showed that the correlation between human diversity and recombination remain after controlling some confounding factors (e.g., GC and CpG content, simple repeats, and distance to telomeres and centromeres). Stamatoyannopoulos *et al.* observed that mutation rate, as measured in evolution divergence and human SNP diversity, is associated with DNA replication timing [216]. It was further shown that mutation rate is associated with chromatin structure, and that regions of open chromatin have the lowest non-CpG mutation rate, while regions with closed chromatin have the highest rate [236]. However, a following study suggested that this association was probably due to the correlation of chromatin compaction with replication timing [237]. Nucleosome occupancy showed a complex association pattern with mutation rate, in which SNPs are enriched around general nucleosome occupancy but depleted around the positions preferentially occupied by epigenetically modified nucleosomes [238].

A number of genetic and epigenetic features have been proposed to influence the rate of SNVs in cancer, including GC content [239, 240], gene density [239, 240], open and closed chromatin structures [240], nucleosome occupancy [239, 240], DNA replication timing [239–242], three-dimensional chromatin organization [240, 242], and DNase I hypersensitivity (a measure of chromatin accessibility) [243]. Taking advantage of a large number of epigenetic features from more than one hundred cell types, a comprehensive study revealed that chromatin accessibility, histone modifications and replication timing can explain 74-86% of mutation rate variance in cancer genomes [244]. Several mechanisms have been proposed to explain the observed associations. The elevated mutation rate in regions of high GC content is attributed to high frequency of CpG dinucleotides, in which methylated cytosine is vulnerable to deamination to thymidine. The negative association between somatic mutation rate and gene density is probably due to an additional DNA damage repair mechanism—transcription coupled repair (reviewed in [245]).

Somatic mutation rate is elevated in closed heterochromatin and is repressed in open chromatin. This could reflect the ready accessibility to DNA repair complexes in open chromatin or increased exposure to mutagens in closed chromatin, which is located at the nuclear periphery in three-dimensional chromosomal folding [240]. Recently, Supek and Lehner observed that somatic mutations are no longer enriched in closed heterochromatin compared with open chromatin after the inactivation of DNA mismatch repair genes [246]. They further proposed that differential DNA repair, rather than differential mutation supply, is the actual cause for regional mutation rate variations in cancer cells. The lower mutation rates in regions of higher nucleosome occupancy could be explained by the fact that DNA in nucleosome undergoes less spontaneous local conformational fluctuations within double-stranded DNA (DNA breathing) and is thus less accessible [247]. One possibility for the accumulation of SNVs in later replicating regions is that the slowing or stalling of replication fork leads to the formation of hypermutable single-strand DNA [216]. It was further observed that mutation rate is reduced in active regulatory regions (defined by DNase I hypersensitive sites), probably suggesting that active regions are more accessible to DNA repair complex [243]. However, two independent studies showed that mutation rate increased in the center of active promoters [248, 249]. The authors of these two papers associated the elevated mutation rate with reduced level of nucleotide excision repair (NER) which is caused by the binding of transcription-initiation machinery [248, 249]. This discrepancy can be explained by the fact that, although regulatory regions as a whole are more accessible to NER, the accessibility for NER in the core sites is limited because of bound transcription-initiation proteins [250].

Statistical associations between somatic mutation rates and genomic properties do not always imply causal effects of individual features. Analyses of mutation signatures in cancer provide an alternative way to uncover the underlying DNA damage and repair processes or replicative mechanisms to which cancer cells have been exposed [251]. The simple analyses of mutational spectra (C·G $\rightarrow$ A·T, C·G $\rightarrow$ G·C, C·G $\rightarrow$ T·A, T·A $\rightarrow$ A·T, T·A $\rightarrow$ C·G, T·A $\rightarrow$ G·C) showed that some mutational spectra are specific to some tumor types and related exogenous mutagens. For example, increased C·G $\rightarrow$ A·T transversion rate in lung cancer is associated with tobacco carcinogen, while C·G $\rightarrow$ T·A transitions are predominantly in ultraviolet (UV) radiation exposure related melanoma [251, 252]. However, these analyses failed to consider the sequence context (the immediately flanking 5' and 3' bases) of a mutation, which affect the mutation rate of the mutated base [233]. In total, there are 96 possible mutated trinucleotides (six types of substitutions,

and four possible bases at 5' base and four possible bases at 3' base). The large-scale cancer sequencing projects provide us an unprecedented opportunity to detect a complete set of mutation signatures in cancer. Mathematical algorithms [253–256] can also be used to extract mutation signatures and to quantify the contribution of each signature. A comprehensive mathematical analysis identified 21 different mutational signatures (for the characteristics of each signature, refer to [257]) from the somatic mutations of more than 7,000 human cancers of 30 different cancer types [257].

Some mutational signatures have been related to endogenous or exogenous DNA damages, DNA repair processes or DNA replication errors [251]. Signature 1A and 1B are characterized by C·G $\rightarrow$ T·A mutations at Np$\underline{\text{C}}$pG trinucleotides ("_" denotes the mutated base) and are observed in many different cancer types [257]. This signature has been linked to mutagenic processes attributed to spontaneous deamination of 5-methylcytosine to thymine. Signatures 2 and 13 are characterized by C·G $\rightarrow$ T·A and C·G $\rightarrow$ G·C mutations at Tp$\underline{\text{C}}$pN trinucleotides and have been found in many cancer types, including breast cancer and bladder cancer [258]. These two signatures result from some highly expressed members of APOBEC (apolipoprotein B mRNA editing enzyme, catalytic polypeptide) enzymes based on similarities in the type and sequence context between mutations caused by APOBEC enzymes *in vitro* and those mutations in cancer [251, 257, 259]. Signature 7, mainly found in malignant melanoma associated with UV radiation [257], has a higher prevalence of C·G $\rightarrow$ T·A sites and CC·GG $\rightarrow$ TT·AA at pyrimidine dimers and is a characteristic feature of transcriptional strand bias [251]. The mutation characteristics and strand bias of signature 7 suggest its formation mechanism through which UV exposure results in pyrimidine dimers followed by transcription coupled repair [251, 257]. Signature 5 is characterized by a broad spectrum of base changes with slightly more C·G $\rightarrow$ T·A and T·A $\rightarrow$ C·G mutations [257]. Recently, signature 5 has been linked to the inactivation of nucleotide excision repair gene *ERCC2* in urothelial cancer [260]. Independent of *ERCC2* mutation status, signature 5 is also associated with smoking history, which provides the first evidence of tobacco-related mutagenesis in urothelial cancer [260]. Signature 10 has been found in some tumors of colorectal and uterine cancer and has a specific pattern of C·G $\rightarrow$ A·T and C·G $\rightarrow$ T·A mutations at Tp$\underline{\text{C}}$pG [251, 257]. The altered proof-reading activity of DNA polymerase Pol ϵ has been proposed to be the underlying mutational process [251, 257].

1.6.2 Indels

Comparative genomic studies of small indel distributions have provided clues to indel generation mechanisms. Indels are not randomly distributed along the genome; indel rates can vary by more than two orders of magnitude [261]. Rates of small indels have been found to be associated with a number of genomic factors, including sequence context (e.g., microsatellites) [261, 262], GC content [263], proximity to telomeres [263], male and female recombination rates [263], and DNA replication errors [263]. The two types of indels are likely generated in part by different mechanism: replication-related factors are more pronounced for deletions, while recombination-related features contributed more to insertions [263]. It was further shown that polymerase slippages are responsible for the majority (75%) of all indels [264]. The remaining indels are mostly simple deletions in regions with complex sequences, and insertions are significantly associated with palindromic sequence features [264]. The latter are compatible with the fork stalling and template switching (FoSTeS) mechanism, which are more frequently associated with SVs [264, 265].

Analysis of indel signatures provided some primary insights into the generation mechanism, although the power to detect indel patterns is relatively limited [251]. Small 1-3 bp indels were found to associate with SNV signature 6 that is characterized by C·G $\rightarrow$ T·A mutations at NpCpG trinucleotides [257]. In colorectal, kidney and prostate cancers, an excess of SNVs associated with signature 6 and small indels [251] was observed in some tumors. This pattern is attributed to the loss of mismatch repair (MMR) genes [257]. Defects in MMR often lead to microsatellite instability—a phenomenon that is characterized by variable numbers of repeats of microsatellites (short repetitive sequences $<$5 bp) and is frequently observed in colorectal and endometrial cancers [266]. Large indels (4-50 bp), by contrast, have been correlated with SNV signature 3 that is characterized by a fairly uniform mutational spectrum [257]. Signature 3 has been reported in breast and ovarian cancers, and tumors associated with signature 3 show increased number of larger indels ($>$3 bp) [257]. Although signature 3 has been linked to the inactivation of *BRCA1* and *BRCA2* genes in homologous recombination double strand break (DSB) repair pathway, the exact mechanism for the elevated indel rate remains elusive [251].

1.6.3 SVs (SCNAs)

A variety of molecular mechanisms have been proposed to explain the formation of SVs (and SCNAs). Generally SVs occur when the repair of DSBs is incomplete [267]. DSBs arise as part of the normal metabolism of the cell (e.g., V(D)J recombination) or as a consequence of exposure to exogenous agents (e.g., ionizing radiation) or from the DNA structures capable of inducing DSBs (such as non-B conformation motifs and DNA transposons) [268, 269]. Repair mechanisms can be divided into three types: homologous recombination repair, non-replicative non-homologous repair, and replicative non-homologous repair [269].

Two molecular pathways, homologous recombination (HR) and single-strand annealing (SSA), use homologous recombination to repair DSBs [269]. They are different from each other in the extent of the required homology: HR requires longer sequence identity than SSA (100 to 200 bp versus 50 bp) [267–269]. Another difference is that SSA always results in small deletions, while HR mostly can repair DSBs without generating CNVs or SVs [267, 269]. A well studied example of HR is non-allelic homologous recombination (NAHR) between low copy repeats (LCRs). LCRs (also known as segmental duplications, SD), are highly homologous sequence elements within the human genome typically 10-300 kb in size, and bear >95% sequence identity [270]. Due to their high degree of sequence identity, non-allelic copies of LCRs, instead of the copies at the usual allelic positions, can sometimes act as the substrates of NAHR. This is the major mechanism responsible for recurrent CNV formation [269]. The relative positioning of LCR pairs can result in different types of genomic rearrangements [271]. Located on the same chromosome in direct orientation or in opposite orientation, or on different chromosomes, NAHR between two LCRs leads to duplication and/or deletion, inversion, and translocation, respectively. SSA can act directly at repeated sequences [269]. In this process, neither of the two DSB ends invades homologous sequences and internal sequence between the two repeats as well as one of the repeats will be deleted. In humans, DSB induced SSA has been observed between identical *Alu* elements [272].

Non-replicative non-homologous repair pathways either do not require homology or need limited micro-homology for repairing DSBs. Non-homologous end joining (NHEJ) does not require a homologous template to guide repair. It either rejoins DSB ends accurately or leads to small deletions (1-4 bp), and sometimes to insertion of free DNA from other genomic regions [269]. The alternative end joining (alt-EJ) mechanism, also called micro-

homology mediated end joining (MMEJ) is an error-prone pathway. In MMEJ, 5-25 bp micro-homologous sequences were used to align DSB ends before joining, thereby resulting in deletions of sequences flanking the original breaks. MMEJ is frequently associated with chromosomal structural changes such as deletions, translocations, and other complex rearrangements [273].

Replicative non-homologous mechanisms have been proposed based on the observation that some of human CNVs are highly complex and hard to be explained by the canonical HR or by end joining pathways [274, 275]. These replication based mechanisms include FoSTeS [276], micro-homology mediated break-induced replication (MMBIR) [277] and serial replication slippage (SRS) [278]. The FoSTeS mechanism has been further generalized to MMBIR [265]. Although these models are different in some aspects, they all assume that the replication fork can stall and template DNA can be introduced via micro-homology from replication forks nearby or over long distances [267, 269, 274]. These mechanisms can result in inversion, tandem duplication, translocation, or more complex rearrangements [265, 277].

In germline cells, insights into the mechanisms underlying SV formation was gained from the studies of genomic disorders [279]. Genomic disorders are a group of diseases caused by the loss or gain of DNA resulting from the inherent human genomic instability at some loci [271]. For example, NAHR is responsible for most of recurrent germline SVs which show breakpoints clustering inside LCRs and recur in multiple patients [268]. In contrast, NHEJ (or MMEJ) accounts for most of non-recurrent SVs that are of different sizes among patients but may share a small regions of overlap within patients [268, 273]. FoSTeS/MMBIR was proposed to explain the observed complex SVs associated with genomic disorders [276, 277].

The landscapes of somatic SVs in cancer are extremely diverse and ranges from very few to hundreds of SVs per patient [280]. Nevertheless, the junctional sequences flanking SV breakpoints at nucleotide resolution enabled the revelation of mechanisms involved in the generation of somatic SVs [251]. Although NAHR was implicated in somatic SVs in cancer, the exact contribution of NAHR remains unknown [268, 281]. The very few overlapping sequences at breakpoints imply that NHEJ and MMEJ are involved in the formation of somatic SVs in cancer [282]. It was suggested that MMEJ occurrence is rarer than NHEJ based on the observation that only 2.5 % of SVs had $>$5 bp micro-homology [282]. A comprehensive study showed that micro-homology based mechanisms (MMEJ

and FoSTeS/MMBIR) contribute more to germline SVs (e.g., deletions, tandem duplications and complex SVs) than to somatic SVs [105]. This phenomenon may suggest that these mechanisms are suppressed in cancer cells or that DNA breakage and replication fork stalling frequently occurred in cancer cells, and non-homology based mechanisms are the easiest way for DNA repair [105].

2

Chromatin organization is a major influence on regional mutation rates in human cancer cells

This chapter reproduce a study [240] published in Nature (Schuster-Böckler B. and Lehner B. Nature, 2012, 488(7412):504-507). Cancer genome sequencing provides an unprecedented opportunity to investigate how mutation rates vary across the genomes of somatic cells. Taking advantage of available genetic and epigenetic features, Schuster-Böckler and Lehner showed that mutation rates in cancer genomes are strikingly related to chromatin organization [240]. They revealed that at the mega base (Mb) scale, a heterochromatin-associated histone modification marker (H3K9me3) explains >40% of mutation-rate variance, and all investigated features account for >55% variance. They also showed that the strong association between somatic mutation rates and chromatin organization is independent of tissue and mutation types. In this work, we reproduced this study using the same data sets in order to offer new insights, if any, into the mutation-rate variance in human somatic cells. Our results are largely consistent with the original study, with the exception being that in our study replication timing is the most prominent predictor for mutation rate in cancer cells.

2.1 Introduction

It has been revealed that germline mutation rates are not constant across the genome [232]. Although the reasons behind mutation rate variance are largely unknown, a number of genetic and epigenetic properties have been found to affect mutation rates including local base composition [232], DNA replication timing [216] and chromatin structure [236]. In a previous study, Hodgkinson *et al.* showed that, at the mega base scale, somatic mutation

rate varies substantially within the human genome, and the associated individual genomic property can only explain very little of the regional variation across the genome [239].

To identify potential causes of mutation rate variance across the genome, the authors of the study [240] compiled a set of genetic and epigenetic features and gathered SNVs from genome sequencing projects of leukemia, melanoma, small cell lung cancer and prostate cancer. The examined features included base composition, CpG content, gene density, DNA replication timing, nucleosome occupancy, long-range chromatin interactions (Hi-C), recombination rate, the density of unique sequences (mappability of 24-base polymers), levels of 18 histone acetylations, levels of 17 histone methylations, and occupancy of RNA polymerase II, the CTCF insulator protein and the histone variant H2AZ.

2.2 Materials and Methods

2.2.1 Data of cancer SNV, germline SNP and human–chimp sequence divergence

Somatic autosomal SNVs were obtained from the supplementary tables of the respective publications: 32 075 SNVs from melanoma [62], 27 354 from prostate cancer [283], 21 708 from lung cancer [284], and 3 874 from leukemia [285]. The leukemia SNVs are not included in the calculation of transition/transversion correlations because the exact alternated nucleotides were not available. All genomic coordinates for somatic SNVs correspond to the human genome assembly hg18. When necessary, the University of California, Santa Cruz (UCSC) *liftOver* tool [286] was used to convert the hg19 coordinates to hg18. The same strategy was applied to germline SNPs, human-chimp divergence data, and genome-wide feature data sets. NCBI dbSNP build 130 comprising of 8 344 654 SNPs was downloaded from the UCSC goldenPath database (`http://hgdownload.cse.ucsc.edu/goldenPath/hg18/database`). Sequence divergence data between *Homo sapiens* and *Pan troglodytes* were extracted from EPO (Enredo–Pecan–Ortheus) whole-genome alignments available from Ensembl release 54. The study [240] showed that the human–chimp alignment covers $>88\%$ of the human genome, yielding approximately 10^8 substitutions across all autosomes. However, we only found about 3.44×10^7 autosomal substitutions.

2.2.2 Genome-wide feature sets

The human genome was split into evenly-sized (1 Mb) windows, and each genome feature measured at 1 Mb scale. GC density is defined as the fraction of all G or C residues per 1 Mb window, and was calculated using UCSC *hgGcPercent* utility. CpG density refers to the fraction of residues in CpG dinucleotides per window. Gene density is the fraction of nucleotides covered by a gene (including exons and introns) per window. Repeat annotation was downloaded from UCSC Genome Browser, and repeat coverage per window was computed using an in-house Perl script. As suggested by the authors of the study [287], replication timing (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. Data for highly positioned nucleosomes were downloaded from `http://liulab.dfci.harvard.edu/NPS/Result/`. These data sets were predicted with the NPS algorithm [288] using micrococcal nuclease digested chromatin data extracted from resting CD4 T cells as reported by *et al.* [289]. Hi-C data for the lymphoblastoid cell line GM06990 were downloaded from the Gene Expression Omnibus database with accession number GSE18199, and eigenvector 2 was used for chromosomes 4 & 5 and eigenvector 1 for other chromosomes [290]. Recombination rates were downloaded from recombRate table of UCSC Genome Browser, and decodeAvg value from the deCODE genetic map [291] was used. The genome-wide uniqueness of 24-base polymers was downloaded from UCSC Genome Browser (wgEncodeDukeUniqueness24bp table). Genomic coordinates for all uniquely mapped reads for 18 histone acetylation markers were downloaded from `http://dir.nhlbi.nih.gov/papers/lmi/epigenomes/hgtcellacetylation.aspx` [292]. Genomic coordinates for all uniquely mapped reads for 17 histone methylation markers, H2AZ, CTCF and RNA PolII binding were downloaded from `http://dir.nhlbi.nih.gov/papers/lmi/epigenomes/hgtcell.aspx` [293]. The genomic coordinates for evolutionarily conserved DNA elements were downloaded from `https://www.broadinstitute.org/mammals-models/29-mammals-project-supplementary-info`, and log-odds Siphy-pi scores were used with a cut-off threshold of 3 [294].

2.2.3 Measurement of cancer SNV, germline SNP, human-chimp sequence divergence and feature sets at 1 Mb resolution

The feature of mappability described above assigns 24-base polymers a value of 1 if they occur uniquely in the genome, 0.5 if they occur twice, 0.33 if they occur three times, and 0 otherwise. The human genome was partitioned into non-overlapping 1 Mb windows, and windows with average mappability less than 0.8 were removed to exclude windows with highly repetitive DNA elements. Genetic and epigenetic features as well as cancer SNVs, germline SNPs and human–chimp sequence divergence were measured as coverage (fraction of a window occupied by the feature) or sum (specifically for replication timing).

2.2.4 Statistical analysis

Pearson correlation analysis and principal component analysis (PCA) were performed in R using the functions *cor.test* and *princomp*, respectively. For PCA, vectors of all genomic features as well as cancer SNVs, germline SNPs and human-chimp divergence at 1 Mb resolution were scaled to mean 0 and standard deviation 1. To identify the minimal informative set of predictive features for cancer SNVs, germline SNPs and human–chimp divergence, linear models were fitted by generalized least-squares estimation. Different models were compared by their Akaike information criterion (AIC), and models with minimal AIC chosen. This procedure was repeated $n - 1$ (*n* equals to the number of features) times, adding one feature to the model at each iteration. The set of features with minimal AIC was chosen as the minimal informative set of predictive features. Percentage explained variance was calculated as the R^2 of a linear regression model using the sets of selected predictive features. Calculations were performed in R using the *AIC*, *gls* and *lm* functions.

2.3 Results

2.3.1 Cancer SNV density is correlated with regional variation in chromatin organization

The human genome was split into 1 Mb windows, and genomic features as well as cancer SNVs, germline SNPs and human–chimp divergence were measured at 1 Mb scale. The correlation coefficient was calculated for all pairwise combinations of genomic features and target features (i.e., cancer SNVs, germline SNPs and human-chimp divergence).

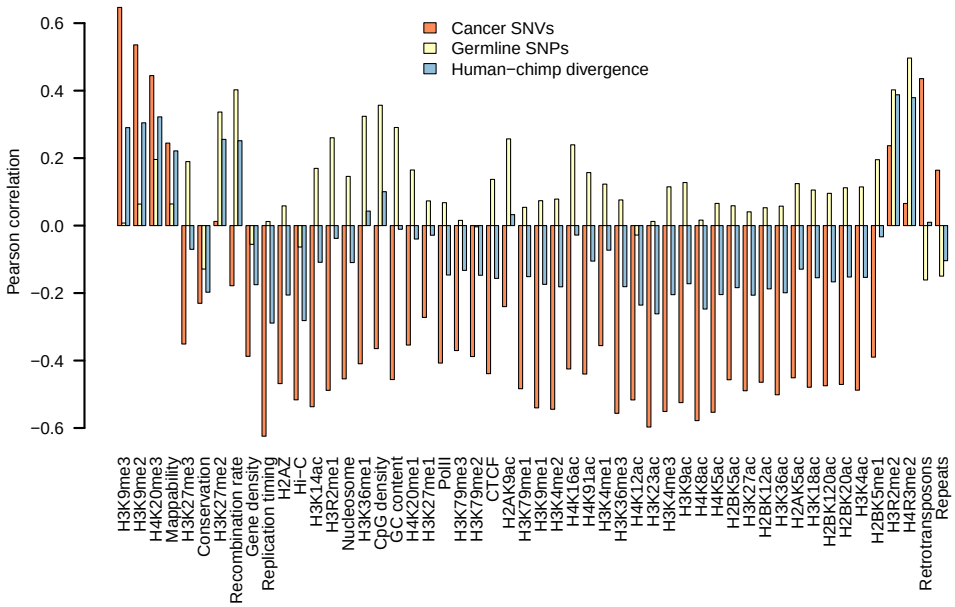


Figure 2.1: Pearson correlation coefficients of cancer SNVs, germline SNPs and human-chimp divergence with genomic features in non-overlapping 1 Mb windows.

Indeed, cancer SNV density is strikingly correlated with many features of closed chromatin organization at 1 Mb scale (Figure 2.1). The repressive histone modification H3K9me3, correlated strongly with cancer SNV density ($r = 0.64, P < 2.2 \times 10^{-16}$). Other repressive histone modification markers also show positive correlations, for instance, H3K9me2 ($r = 0.53, p < 2.2 \times 10^{-16}$) and H4K20me3 ($r = 0.43, p < 2.2 \times 10^{-16}$). In contrast, cancer SNV density negatively correlated with levels of open chromatin associated histone markers, such as H3K4me3 ($r = -0.60, p < 2.2 \times 10^{-16}$) and H3K9ac ($r = -0.59, p < 2.2 \times 10^{-16}$). Anti-correlations are also observed with other ge-

nomic features such as replication timing ($r = -0.66, p < 2.2 \times 10^{-16}$), GC content ($r = -0.46, p < 2.2 \times 10^{-16}$), gene density ($r = -0.40, p < 2.2 \times 10^{-16}$) and the density of highly positioned nucleosomes ($r = -0.48, p < 2.2 \times 10^{-16}$). The authors also showed that these conclusions are upheld when splitting the genome into alternative sizes (i.e., 10 kb, 100 kb and 10 Mb) [240]. Taken together, the authors concluded that regional mutation rate variance is strongly associated with regional variation in chromatin organization [240].

To investigate the inter-dependencies among different genomic features, the correlation coefficient was calculated for all pairwise combinations of genomic features. The results showed that genomic features were clustered into two distinct groups, one consisting of retrotransposons, repeats, H3K9me3, H3K9me2 and H4K20me3 while the other group including other genomic features (Figure 2.2).

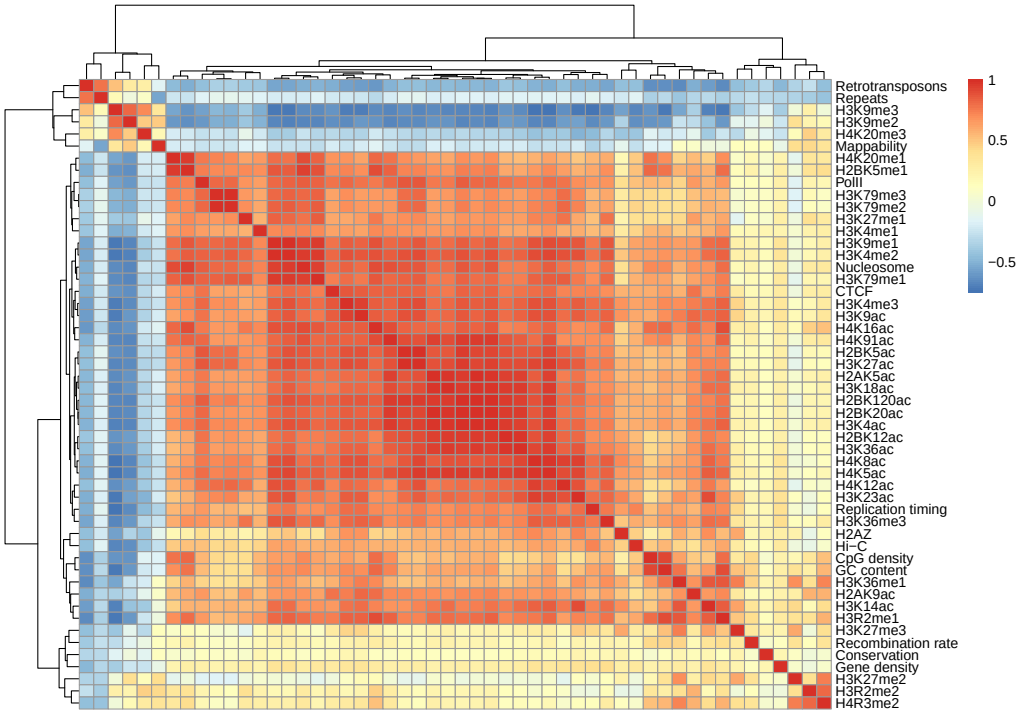


Figure 2.2: The correlation matrix of genomic features at 1 Mb resolution. Red denotes positive and blue negative correlation.

Principal component analysis was further used to investigate the inter-dependencies of features. The results showed that at 1 Mb resolution about 60% of the variance in these genomic features could be explained by a first principal component (Figure 2.3). Many histone modifications and other features associated with either accessible euchromatin or inaccessible heterochromatin have a strong loading on the first principal component (Figure 2.4). For example, the histone modifications H3K9me3, H3K9me2 and H4K20me3

have strong negative loadings on the first component. GC content, gene density, replication timing and many histone modifications associated with accessible euchromatin show strong positive loadings on the first component. Cancer SNV density also shows a strong negative loading on the first component. The authors believe that the result is consistent with the idea that somatic mutation rate is higher in inaccessible heterochromatin domains and lower in accessible euchromatin ones [240]. In contrast, germline SNP density and human-chimp divergence have stronger loadings on the second principal component (Figure 2.4).

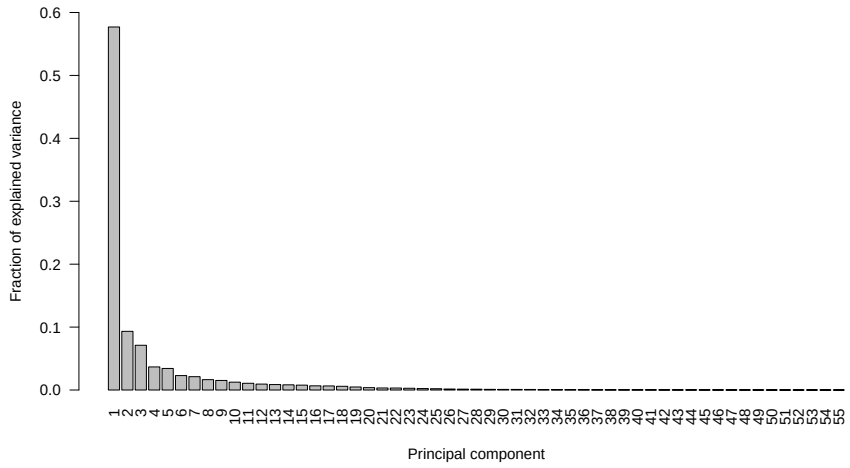


Figure 2.3: Percentage of total variance explained by each principal component.

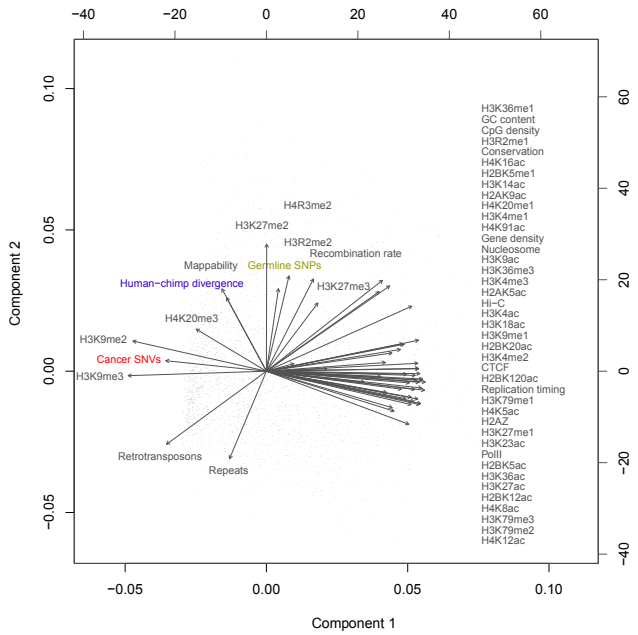


Figure 2.4: Bi-plot of first two principal components. Black dots denote transformed values of individual 1 Mb windows.

2.3.2 The correlation between chromatin organization and mutation rate variance is independent of cancer type, mutation type and genomic context

To investigate whether the correlation between chromatin organization and cancer SNV density are independent of tissue type, the authors also analyzed the mutation data from each tumor type separately. It has been shown that somatic mutations show signatures associated with mutagen exposure such as ultraviolet light exposure in the melanoma [62] and tobacco smoking in the lung cancer [284]. It can be seen from Figure 2.5 that in each cancer type SNV density is positively correlated with repressive histone markers (e.g., H3K9me3, H3K9me2 and H4K20me3), and negatively correlated with genomic features associated with accessible chromatin. These results indicate that the correlation between chromatin organization and mutation rate variance is independent of tumor type.

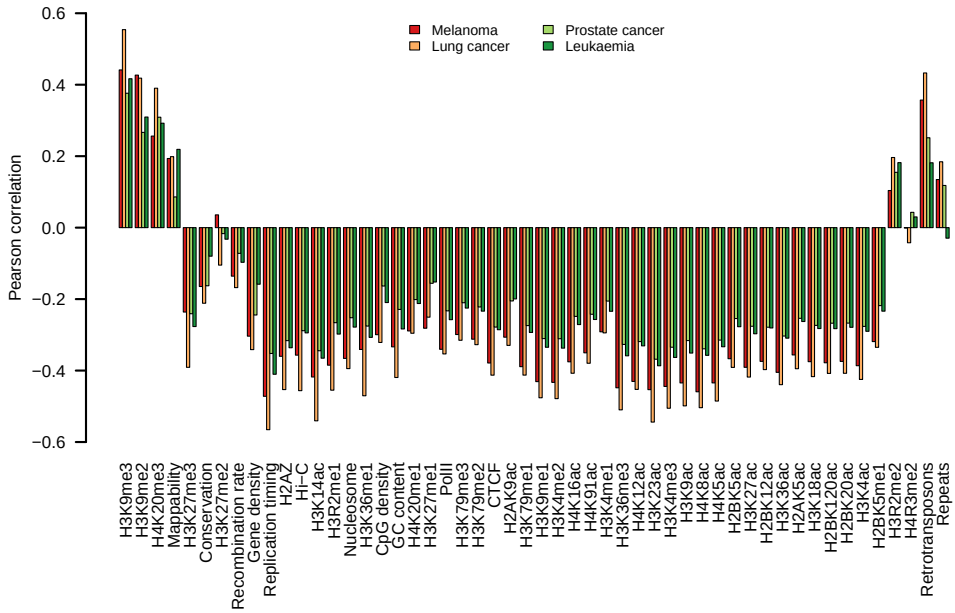


Figure 2.5: Correlation coefficients of SNV density from individual cancer genomes with diverse genetic and epigenetic features at 1 Mb resolution.

To determine whether the correlation between chromatin organization and mutation rate is mutation-type and genomic-context specific, mutations were divided into different categories: transitions or transversions, CpG mutations or non-CpG mutations, mutations in genic or non-genic regions. The results showed that mutation rates are strongly associated with H3K9me3 for mutations of different types and genomic context (Figure 2.6). The correlation between mutation rate and H3K9me3 is strong when only considering SNVs

surrounded by 20 bp of unique sequence, or when excluding evolutionarily conserved bases, or when filtering out regions with extreme GC content (<35% or >75%) (Figure 2.6).

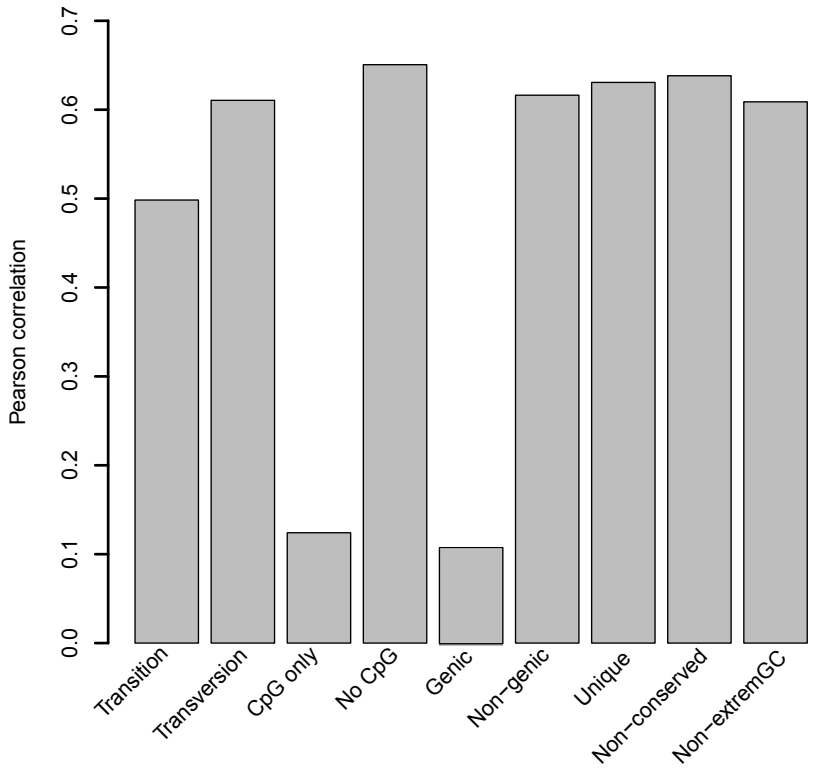


Figure 2.6: Correlation coefficients of cancer SNV density with H3K9me3 for diverse mutation types and genomic context.

Taken together, the association between chromatin organization and mutation rate variance in cancer cells is upheld in diverse tumor types, mutation types and genomic regions.

2.3.3 Improved prediction power for cancer SNV density variation by integrated models

A previous study showed that all genomic features together explain less than 40% of the mutation-rate variance and individual features explain very little [239]. Next, the authors examined whether predictions of variance in mutation rate could be improved by using linear regression models combining the information from multiple genomic features [240]. Using the same procedure, our results showed that all genomic features can explain about 55% of the variance in cancer SNV density along the genome, and that a single

feature —- replication timing (instead of H3K9me3 in the original study [290]) alone can account for more than 42% of the variance (Figure 2.7).

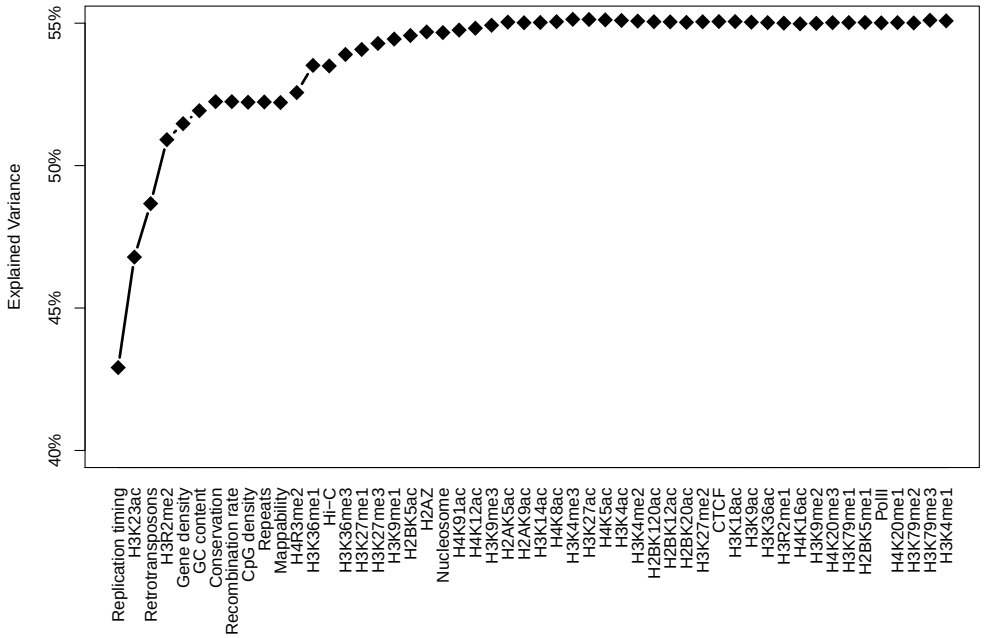


Figure 2.7: Prediction of cancer SNV density variation using integrated models. Cumulative R^2 of linear models, adding the feature on the x axis as a predictor at each step.

2.4 Discussion

Somatic SNVs are not uniformly distributed along the human genome. The authors of the study [240] showed that somatic mutation rate is associated with chromatin organization, irrespective of tumor type, mutation type, or genomic context. They also showed that at the Mb scale, a repressive histone modification marker — H3K9me3 — explains >40% of mutation-rate variance. Using the same data sets and same procedure, we got the results which are largely consistent with those presented in [240]. The only exception is that replication timing is the most prominent predictor for somatic mutation rate in cancer cells. Our results comply with two subsequent studies [241, 242], in which replication timing was found to have a prominent role in shaping SNV landscape in cancer cells.

Both somatic mutagenesis and epigenetic features are highly cell-type specific. Since the data for histone modification markers analyzed in the study [240] were not from the same cell types as the somatic mutations, the authors argued that the actual influence of chromatin organization on regional mutation rates could have been underestimated. Re-

cently, a comprehensive study compared somatic mutations from diverse cancer types to cell-type-specific epigenomic features [244]. The results showed that chromatin accessibility and modification, as well as replication timing, can explain 74-86% of mutation rate variance along cancer genomes.

Somatic mutation rates are elevated in heterochromatin-like domains and repressed in open chromatin domains. The authors suggested that this pattern could reflect the variation in DNA accessibility by DNA repair machines between open and closed chromatin domains [240]. A following study by the same group showed that after the inactivation of DNA mismatch repair genes, somatic mutation rates are no longer elevated in closed heterochromatin regions [246]. They further proposed that differential DNA repair, rather than differential mutation supply, is the primary cause of large-scale regional mutation rate variance in cancer cells.

3

Genomic determinants of somatic copy number alterations across human cancers

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 >30% of the pooled SCNA breakpoint variation, with the explanatory power ranging from 13-32% 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.

This chapter has been published in Zhang, Y., Xu, H., and Frishman, D. (2016) Genomic determinants of somatic copy number alterations across human cancers. *Hum. Mol. Genet.*, 25(5), 1019–1030. Yanping Zhang and I contributed equally to this work. This study was designed by Dmitrij Frishman, Yanping Zhang and me. Yanping Zhang collected data and did multiple linear regression, and I did logistic regression and ex-

tremely randomized tree classifier. The manuscript was written by Yanping Zhang and me, and corrected by Dmitrij Frishman.

3.1 Introduction

Cancer is fundamentally a disease characterized by a diversity of somatic alterations [41]. 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) [4, 41, 46]. The landscape of SCNAs has been charted across different types of cancer, with recurrent SCNAs often pointing at novel oncogenes and tumor suppressor genes [42, 46, 47]. Although SCNAs affect a sizable 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 [46, 47, 295, 296]. SCNA breakpoints are not uniformly distributed in the genome, but rather tend to be spatially clustered in breakpoint hotspots [295]. For instance, G-quadruplex sequences (G4s) are enriched in the vicinity of SCNA breakpoints, suggesting the contribution of genomic properties to SCNA formation [295]. 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 [297]. However, the model presented in [297] 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 [268, 269, 295]. Indeed, genome-wide mapping of DSBs has shown that DSB regions are enriched in genomic regions frequently rearranged in cancers [298]. 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 [299]. The resulting DNA secondary structures have been implicated in the formation of structural alterations including copy number variations (CNVs), inversions and translocations, such as G-quadruplexes [295], Z-DNA [300], cruciforms formed by inverted repeats [301] and triplexes (also known as H-DNA) formed by mirror repeats [302]. 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 [303, 304]. Moreover, a 13-mer CCNCCNTNNCCNC motif was found to associate with recombination hotspots in humans and was clustered in common mitochondrial deletion hotspots [305]. Recently, Zhou *et al.* [306] have revealed a significant enrichment of human germline and somatic structural variant breakpoints in self-chain (SC) regions, a group of low-copy repeats (LCRs) 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 SCNAs and those SCNAs that are interstitial to chromosomes, highlighting the importance of chromosome structure (e.g., telomeres and centromeres) [47].

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 >40% of mutation rate variation in cancer cells [240]. 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 [307] and investigating the effects of diverse sequence features on integration sites of DNA transposons [308].

The overview of our study is presented in Figure 3.1. Taking advantage of SCNA 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 >30% of SCNA breakpoint variation. The power of the

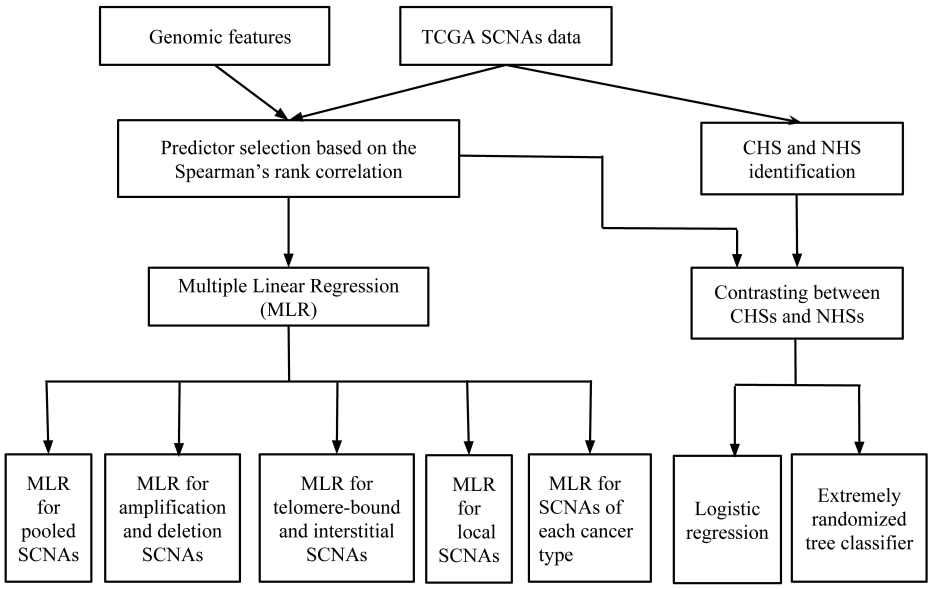


Figure 3.1: An overview of the study design.

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.

3.2 Materials and Methods

3.2.1 SCNA data

The first SCNA data published in [47] 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 SCNAs 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 3.1). 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 3.2). 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 3.2). 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) [154], and we retrieved 699 492 SCNAs generated by studies other than TCGA (COSMIC study identifiers: 328, 382, 538, 585, 586, 589, and 650).

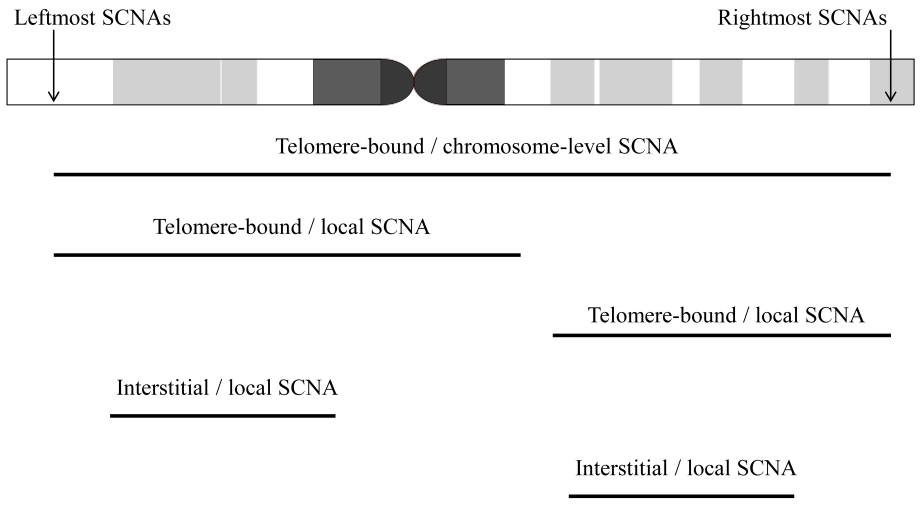


Figure 3.2: Schematic illustration of SCNA categories considered in this work.

Table 3.1: Summary of somatic copy number alteration (SCNA) data from The Cancer Genome Atlas 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	13 344	4 562	802	1 172	3 900	1 326	1 582
Breast invasive carcinoma	BRCA	745	99 574	42 268	2 624	8 792	25 414	8 610	11 866
Colon adenocarcinoma	COAD	349	21 650	4 222	2 318	2 004	6 672	3 966	2 468
Glioblastoma multiforme	GBM	485	28 462	10 162	2 078	1 074	10 234	2 556	2 358
Head and neck squamous cell carcinoma	HNSC	270	24 272	6 990	1 130	3 068	5 586	3 320	4 178
Kidney renal clear cell carcinoma	KIRC	373	9 040	1 818	1 024	860	1 756	2 230	1 352
Lung adenocarcinoma	LUAD	292	34 952	12 080	1 890	3 430	8 006	4 882	4 664
Lung squamous cell carcinoma	LUSC	261	34 400	10 828	1 106	3 998	7 992	4 628	5 848
Ovarian serous cystadenocarcinoma	OV	457	92 216	41 238	2 762	10 720	19 200	7 176	11 120
Rectum adenocarcinoma	READ	147	12 358	2 620	1 114	1 090	3 694	2 328	1 512
Uterine corpus endometrial carcinoma	UCEC	376	34 220	18 014	1 196	2 726	6 570	2 132	3 582
Total		3845	404 488	154 802	18 044	38 934	99 024	43 154	50 530

Abbr., Abbreviation; Chr., Chromosome.

3.2.2 Data collection on genomic features

A total of 29 genomic features were considered as potential predictors of the SCNA patterns (Table 3.2). 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 [286].

Table 3.2: 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)

Chromosomal coordinates of the following genomic features were downloaded from the UCSC Genome Browser [286]: 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 self-chain regions (SCs). The latter elements are essentially pairs of short (up to 1 kb) low-copy

repeats either in direct (+) or inverted (-) orientation [306]. Following [306] we only considered self-chain segments (SCSs) 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 duplications.

Non-B DNA motifs (A-phased repeats, direct repeats, inverted repeats, mirror repeats, G4s and Z-DNA) were downloaded from the non-B DB version 2 [299]. We used the dataset of conserved DNA elements in vertebrates published by Siepel *et al.* [309]. Regions containing DSBs were downloaded from Tchurikov *et al.* [310]. Genomic coordinates for each histone modification marker H3K9me3 in $CD4^+$ T cells were obtained from the study of Barski *et al.* [293]. Replication timing (RT) data for the lymphoblastoid cell line GM06990 were obtained from Hansen *et al.* [287]. 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 [311]. 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 [312]). We obtained the coordinates for fragile sites and miRNA genes from a previous study [307] and miRbase [313], respectively. The rates of nucleotide substitutions and indels were calculated based on human-chimpanzee alignments as described in [297].

3.2.3 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 3.2). All features were evaluated for normality, and if necessary transformed by the logarithm function to approximate it (Table 3.2). 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 [307, 308]. 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.

3.2.4 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 [295, 297, 314]. 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 [311]. 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.

3.2.5 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 relative contribution to variance explained (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 [315] 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 [316]. The *MASS* [317] and *Car* [318] packages were used to generate the common diagonostic plots (e.g., residual plots, Q-Q plots) and the *QuantPsyc* [319] 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* [320] 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.

3.2.6 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) [321] to reduce the sample imbalance bias. The RELR analysis was performed with the help of the statistical software Zelig (`http://gking.harvard.edu/zelig`) [322] 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.

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

A classification decision tree [323] 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`) [324]. 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.

3.3 Results

3.3.1 Identification of SCNA breakpoint hotspots

In this work we analyzed data on 404 488 SCNA breakpoints [47] in 11 cancer types (Table 3.1). 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 3.3 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.

3.3.2 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 [297], 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 categorized into six groups: non-B DNA conformations; DNA sequence; gene regulation and expression; evolutionary features; chromosome structures; and functional features (Table 3.2). Following Fungtammasan *et al.* [307] and Campos-Sánchez *et al.* [308], we used hierarchical clustering with Spearman's rank correlation to remove some strongly correlated features (Figure B.1). Finally, 25 features were used for subsequent regression analyses.

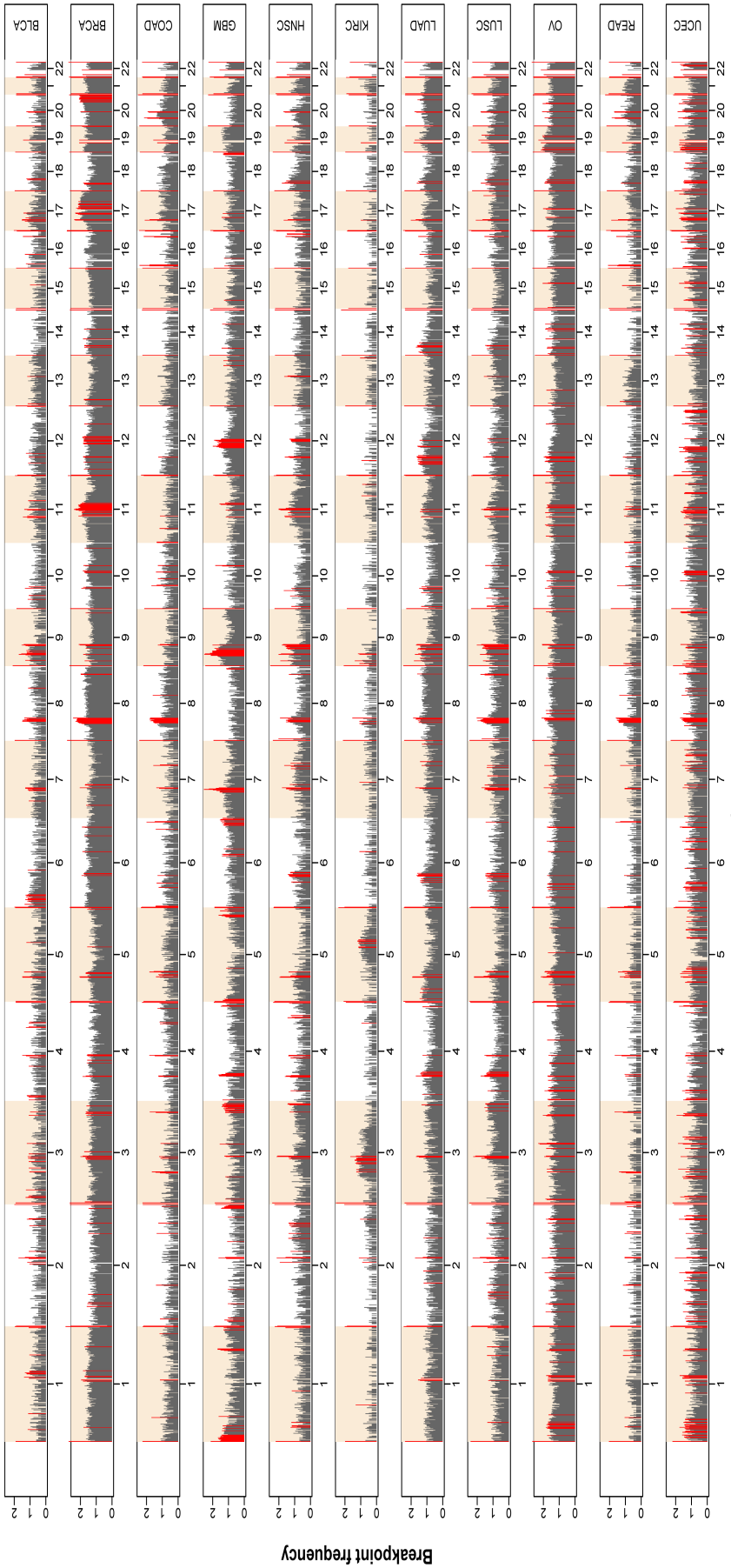


Figure 3.3: The distribution of SCNA breakpoint frequencies in 11 cancer types — BLCA, BRCA, COAD, GBM, HNSC, KIRC, LUAD, LUSC, OV, READ and UCEC (see Table 3.1 for full names), calculated as $= log_{10}$(the number of SCNA breakpoints in each block plus 1). Breakpoint hotspots in each cancer type are colored in red.

3.3.3 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 [307, 308], 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 3.3). 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 RCVE ranging from 0.89 to 2.06% (Table 3.3; Figure 3.4). 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 3.3: The multiple linear regression (MLR) model for pooled SCNA breakpoints

Predictor	SCE	VIF	P-value	RC,%	Five-fold RC,%
Distance to centromere	-0.243	1.265	4.24×10^{-38}	14.55	19.76
Conserved element count	0.113	3.382	1.88×10^{-04}	1.18	1.07
CpG island coverage	0.072	1.133	3.88×10^{-05}	1.43	1.11
Direct repeat coverage	0.425	5.433	7.69×10^{-28}	10.35	11.97
Inverted repeat coverage	0.098	3.330	1.17×10^{-03}	0.89	0.51
L1 coverage	0.136	3.677	1.66×10^{-05}	1.57	1.67
Low complexity repeat coverage	0.142	3.069	8.34×10^{-07}	2.06	2.78
Mirror repeat count	-0.303	4.284	1.12×10^{-18}	6.68	7.70
SINE count	0.223	3.762	4.84×10^{-06}	1.77	1.87
Distance to telomere	-0.419	1.883	2.81×10^{-72}	29.15	32.21
Z-DNA coverage	-0.108	3.146	2.46×10^{-04}	1.14	Not significant
Simple repeat coverage	-0.087	2.434	6.67×10^{-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.

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 A.1, A.2, A.3 and A.4).

	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 3.4: The effect of genomic features in multiple linear regression models. The intensity of color is proportional to the RCVE in each model. Predictors in white color are not significant. See Table 3.1 for full names of cancer types.

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 3.4, 3.5), 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 3.4: The MLR model for SCNA amplification breakpoints

Predictor	SCE	VIF	P-value	RC,%	Five-fold RC,%
Distance to centromere	-0.293	1.265	1.88×10^{-52}	22.39	31.04
Conserved element count	0.118	3.382	1.17×10^{-04}	1.37	1.38
CpG island coverage	0.056	1.133	1.52×10^{-03}	0.93	0.73
Direct repeat coverage	0.347	5.433	7.82×10^{-19}	7.34	5.73
Inverted repeat coverage	0.123	3.330	5.50×10^{-05}	1.50	1.83
L1 coverage	0.121	3.677	1.51×10^{-04}	1.32	0.60
Low-complexity repeat coverage	0.106	3.069	2.73×10^{-04}	1.22	0.07
Mirror repeat count	-0.247	4.284	1.17×10^{-12}	4.70	5.61
SCS coverage	0.065	1.375	9.83×10^{-04}	1.00	Not Significant
SINE count	0.218	8.762	1.06×10^{-05}	1.79	1.34
Distance to telomere	-0.411	1.884	4.54×10^{-68}	29.73	31.79
Simple repeat coverage	-0.120	2.434	4.12×10^{-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.

Table 3.5: The MLR model for SCNA deletion breakpoints

Predictor	SCE	VIF	P-value	RC,%	Five-fold RC,%
Distance to centromere	-0.192	1.265	1.02×10^{-23}	10.23	13.68
Conserved element count	0.099	3.382	1.36×10^{-03}	1.02	0.34
CpG island coverage	0.074	1.133	4.01×10^{-05}	1.68	Not Significant
Direct repeat coverage	0.426	5.433	9.81×10^{-27}	11.66	12.54
L1 coverage	0.131	3.677	5.21×10^{-05}	1.63	1.63
Low-complexity repeat coverage	0.148	3.069	5.67×10^{-07}	2.50	2.09
Mirror repeat count	-0.304	4.284	5.17×10^{-18}	7.56	8.55
SINE count	0.205	8.762	4.32×10^{-05}	1.67	1.19
Distance to telomere	-0.383	1.884	1.42×10^{-58}	27.30	33.00
Z-DNA coverage	-0.119	3.214	8.70×10^{-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.

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 3.3, 3.4 and 3.5). 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 [47]. 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 3.6 and 3.7. 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.

By definition, the breakpoints of chromosome-level SCNAs are fixed at 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 A.5). Notably, the predictors and their effect are similar to those of pooled SCNAs.

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-

Table 3.6: The MLR model for telomere-bounded SCNA breakpoints

Predictor	SCE	VIF	P-value	RC,%	Five-fold RC,%
Distance to centromere	-0.163	1.265	1.35×10^{-18}	6.49	7.48
Conserved element count	0.109	3.382	3.24×10^{-04}	1.07	1.03
CpG island coverage	0.070	1.133	6.38×10^{-05}	1.32	0.22
Direct repeat coverage	0.439	5.433	7.06×10^{-30}	10.91	10.07
L1 coverage	0.160	3.677	3.52×10^{-07}	2.15	2.18
Low-complexity repeat coverage	0.154	3.069	9.67×10^{-08}	2.36	2.20
Mirror repeat count	-0.329	4.284	6.39×10^{-22}	7.78	8.32
SINE count	0.184	8.762	1.57×10^{-04}	1.18	1.10
Distance to telomere	-0.429	1.884	8.74×10^{-76}	29.97	31.98
Z-DNA coverage	-0.115	3.214	9.05×10^{-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.

Table 3.7: The MLR model for interstitial SCNA breakpoints

Predictor	SCE	VIF	P-value	RC,%	Five-fold RC,%
Distance to centromere	-0.349	1.265	6.63×10^{-65}	45.91	53.44
H3K9me3 count	0.143	2.272	9.89×10^{-08}	4.27	2.80
LTR coverage	-0.090	2.206	6.65×10^{-04}	1.74	1.95
SINE count	0.178	8.762	7.12×10^{-04}	1.72	1.53
Simple repeat coverage	-0.122	2.434	1.07×10^{-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.

complexity repeat coverage and SINE count as significant positive predictors for almost all cancer types (Figure 3.4). The distance to telomere, distance to centromere and mirror repeat count remained significant negative predictors for each cancer type (Figure 3.4).

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 3.3). Moreover, the significant predictors and their effects identified in 5-fold MLR are similar to those of MLR (Table 3.3). The 5-fold MLR results for the other MLR models are provided in Tables 3.4-3.7, Tables A.1-A.5 and Figure 3.5. 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 data sets.

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 3.8). The most significant predictors, e.g., distance to telomere, mirror repeats and distance to centromere identified in the MLR model for

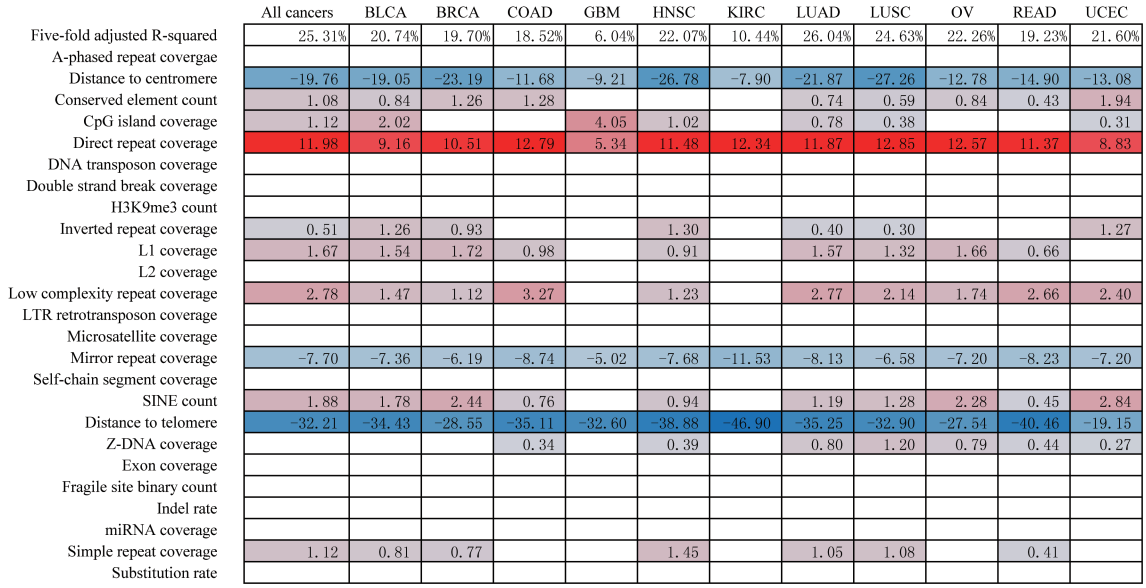


Figure 3.5: The effect of genomic features in 5-fold MLR models. The intensity of color is proportional to the RCVE of each model. Predictors in white color are not significant. See Table 3.1 for full names of cancer types.

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 [297].

Table 3.8: The MLR model 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	2.15×10^{-04}	0.79	0.78
Distance to centromere	-0.086	1.299	1.24×10^{-06}	1.36	1.29
CpG island coverage	0.059	1.198	4.66×10^{-04}	0.71	0.67
H3K9me3 count	-0.153	3.072	2.08×10^{-08}	1.82	1.87
LTR retrotransposon coverage	-0.099	2.230	1.89×10^{-05}	1.06	0.94
Mirror repeat count	-0.128	4.447	9.17×10^{-05}	0.88	0.67
Distance to telomere	-0.212	1.634	5.48×10^{-26}	6.56	7.12
Exon coverage	0.202	3.551	6.70×10^{-12}	2.74	2.87
Indel rate	0.121	5.124	5.85×10^{-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.

3.3.4 Contrasting between common hotspots and non-hotspots 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 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 versus 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 3.9; Figure 3.6). 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 3.9: Rare events logistic regression for contrasting common hotspots with non-hotspots

Predictor	Standardized coefficient	P-value	Relative contribution,%
Conserved elements count	5.029	5.18×10^{-04}	1.01
CpG island coverage	1.825	1.04×10^{-06}	1.14
Direct repeats coverage	11.257	2.16×10^{-11}	5.16
DNA coverage	-5.251	3.82×10^{-05}	2.02
L1 coverage	8.253	1.87×10^{-09}	2.95
L2 coverage	-4.857	2.02×10^{-05}	1.61
Low-complexity repeats coverage	3.746	1.56×10^{-04}	1.08
Mirror repeat count	-2.741	5.41×10^{-03}	0.67
SINE count	10.513	6.26×10^{-08}	2.50
Distance to telomere	-44.259	4.50×10^{-27}	20.70
Z-DNA coverage	-4.025	1.16×10^{-05}	1.61
Simple repeat coverage	-6.701	9.29×10^{-04}	1.02
Explained Deviance			51.83

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.

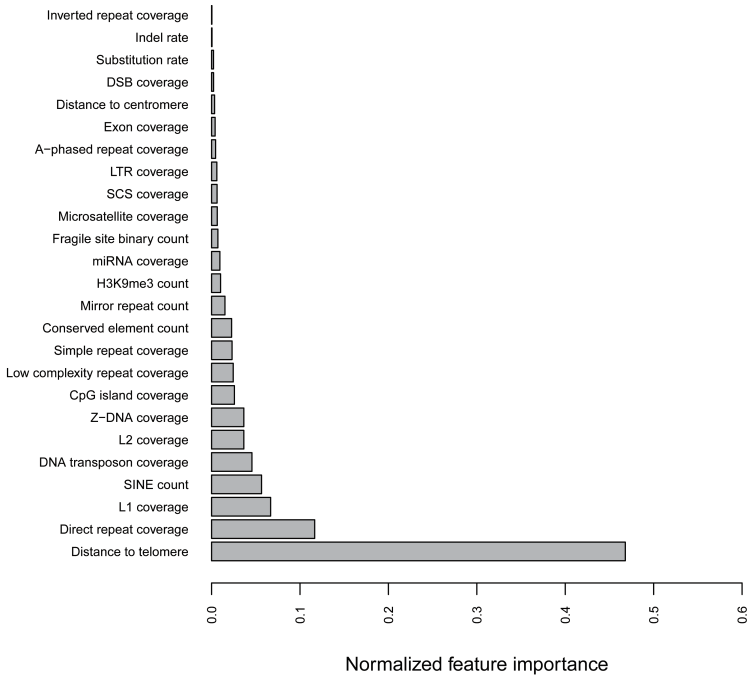


Figure 3.6: The normalized relative contribution of predictors in terms of distinguishing common hotspots and non-hotspots for the rare events logistic regression model.

3.3.5 Extremely randomized tree classifier for telling apart common hotspots and non-hotspots

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 3.7a). The important features determined by the classifier for CHSs are distance to telomere, indel rate, and direct repeats (Figure 3.7b), 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.

3.4 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 [297]. Taking advantage of the TCGA Pan-Cancer SCNA data, we considered a wider range of genomic features than in [297] and performed prescreening of features to reduce the effect of multicollinearity. Our MLR model is more than two times more powerful than that in [297] (32% of break-

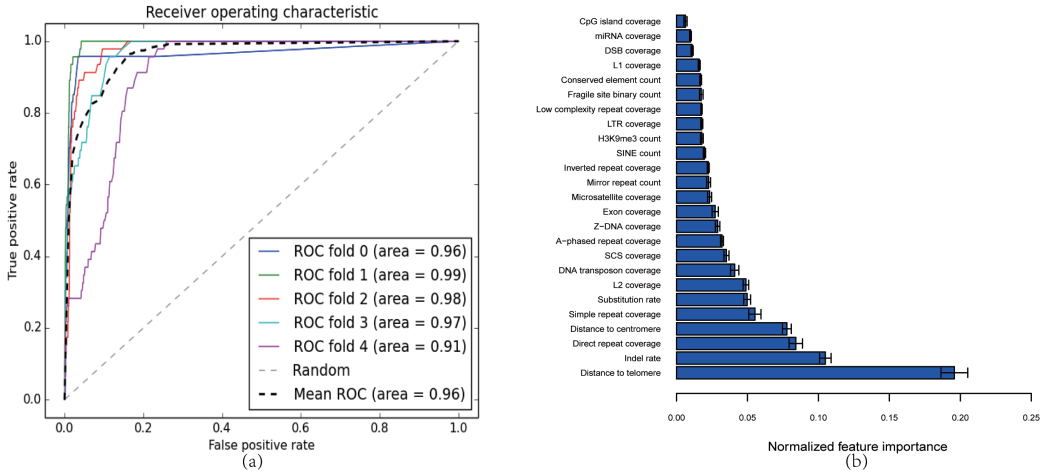


Figure 3.7: Distinguishing common hotspots from non-hotspots from genomic features. (a) ROC-AUC curves of the extremely randomized forests; (b) The normalized relative contribution of predictors in terms of distinguishing CHSs and NHSs.

point 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 [297]. Notably, out of the top four features reported in [297] 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 A.6). 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 3.8). 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 [47]. The frequency of interstitial SCNAs is inversely correlated with their lengths [46, 47], while the telomere-bound ones tend to follow a uniform length distribution [47], 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 [47, 325, 326]. Frequent breakages near centromeres may lead to their dysfunction and further cause chromosomal instability [327], which is a hallmark of diverse cancers [30]. The prevalence of telomere-bound SCNAs in cancers may relate to telomere dysfunction [328], and those breakpoints of telomere-bound SCNAs that are not located in telomeres were speculated to occur at regions with DSBs [47]. 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 [329]. 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 [303, 330]. 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 [331] and thus induce breaks. Moreover, AT-rich repeats constitute unstable regions of the genome, conferring susceptibility to rearrangements [332]. These results suggest a general mechanism of genome instability induced by genomic context.

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 [297] (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 [46, 297], Li *et al.* hypothesized that the high frequency of SCNAs in these CHSs across cancer types is largely due to regionally higher mutation rate [297]. 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.

4

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

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 provides novel clues for understanding the biology of OS.

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, I and Yanping Zhang 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, Yanping Zhang and I conceived the bioinformatics part of the project. I did somatic copy number alteration calling, driver gene identification, tumor subclone decomposition, tumor purity and ploidy estimation. Yanping Zhang performed chromothripsis detection, chromosomal breakage characterization as well as its association with genomic features. The manuscript was written by Jan Smida, Yanping Zhang and me, and edited by Valerie B. O'Leary, Dmitrij Frishman and Michaela Nathrath.

4.1 Introduction

Osteosarcoma (OS) is the most common primary malignant bone tumor in adolescents and young adults [333, 334]. 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 [335–337]. 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% [338], 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 [339] and chromosomal alteration staging systems [336] 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 [30, 340], is either driven by positive selection or originates from sequence-specific unstable regions [30]. Chromosomal fragile sites are specific genomic locations that appear as gaps or breaks on metaphase chromosomes under replication stress [341]. This can be induced by endogenous or exogenous sources, and result in the generation of DNA double strand breaks (DSBs) and genomic instability [342]. A variety of molecular pathways are involved in DSB repair, and, in the case of deficient repair, copy number alterations result.

To identify SCNAs, 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 [343]. However, the analysis of whole genome array data for tumor samples can be difficult 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 [81]. 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 [46]. Analyses of SCNA data generated in recent years have provided insights into driver genes for many tumor types [46, 47]. However, the enormous complexity of genomic aberrations in OS has made it challenging to identify recurrent alterations and genes driving tumorigenesis [335, 337]. Furthermore, in OS the identification of driver genes has been hindered by intra- and inter-tumor heterogeneity and limited sample availability [337, 344–346]. Despite such complications, we and others have revealed recurrent genomic loss in regions containing tumor suppressor genes such as *LSAMP*, *CDKN2A*, *RB1* and *TP53* and most frequent gains at sites including the oncogene *MYC* and the gene *RUNX2* — an important player in osteogenic differentiation [337, 344–347].

Apart from their genomic instability, OSs 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 [37]. There is a strong evidence for an association between chromothripsis and poor outcome in different cancer types, including multiple myeloma [348], neuroblastoma [349] and Sonic-Hedgehog medulloblastoma [350]. Although the mechanisms governing chromothripsis are largely unknown, it has important implications for our understanding of cancer and disease [351], as such detailed analyses of CTLPs may shed light on OS development and progression.

Herein, copy number profiles derived from 160 pre-therapeutic OS biopsies have been analyzed using whole-genome CytoScan High Density (CytoScan HD) arrays (Affymetrix, Santa Clara, CA). Integration of SCNAs for each sample was performed in order to iden-

tify potential genes driving 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 influenced by local genomic context. Most noteworthy, both aneuploidy and CTLP occurrence were found to be correlated with clinical outcome of OS patients.

4.2 Materials and Methods

4.2.1 Tissue samples and patient characteristics

For CytoScan HD array analysis, a set of 160 fresh-frozen tissue samples derived from pretherapeutic biopsies was used. All biopsies were evaluated by an experienced bone pathologist who confirmed the tumor content to be >70% per sample. 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, `http://www.ekbb.ch`, Reference 274/12). The descriptive characteristics of this collection are summarized in Table 4.1 (three samples were excluded due to insufficient copy number profiling quality). The vast majority of the investigated samples (n=141) are classified as high-grade OS. The patients were treated between 1990 and 2012 according to the protocols of the Cooperative German-Austria-Swiss OS Study Group [352](reviewed and approved by the appropriate ethics committees) after informed consent was obtained.

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

DNA from frozen OS tissue was analyzed using the Affymetrix CytoScan HD platform. The raw data are available in the ArrayExpress database [353] 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

Table 4.1: 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	64.5	
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%

(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 informative probes were excluded from further consideration. GISTIC 2.0 (Genomic Identification of Significant Targets In Cancer) integrated in the Nexus copy number software was utilized to identify potential driver SCNAs and genes by evaluating the frequency and amplitude of observed events [217].

Subclone structures were reconstructed for each tumor sample based on the SCNA calling data from the Nexus copy number software. The SubcloneSeeker software [354] was used 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 [354] 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".

4.2.3 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 [286]. 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 [286]. Non-B DNA motifs were obtained from non-B DB v2.0 [299]. Common fragile sites were found to be tissue- and cell-type specific [355]. As tissue-specific data was not available, we obtained genomic coordinates for common fragile sites and non-fragile regions from a previous study [307]. 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 (i.e., item) 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 [311] and in-house Perl scripts.

4.2.4 Detection of chromothripsis-like patterns in osteosarcoma

To detect CTLPs the algorithm described in [356] 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.

4.2.5 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 probeset, the log R ratio (LRR) reflects the ratio of total signal intensity for both alleles against expected signals, and the B allele frequency (BAF) is an estimate of the relative proportion of one of the alleles with respect to the total signal intensity. LRR and BAF values were derived using the *affy2sv* R package [357] together with the Affymetrix Power Tools. A total of 873 normal samples downloaded from the study [358] (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) [83] 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 [126], 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 [126]. Tumor ploidy was further determined following the protocol described in [359]. 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 [360].

4.3 Results

4.3.1 Overview of somatic copy number alterations in osteosarcoma

The SCNA landscape of pre-treatment tissue samples (n = 160) from OS patients (characteristics of whom are provided in Table 4.1) 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 4.1. In our collections, 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.

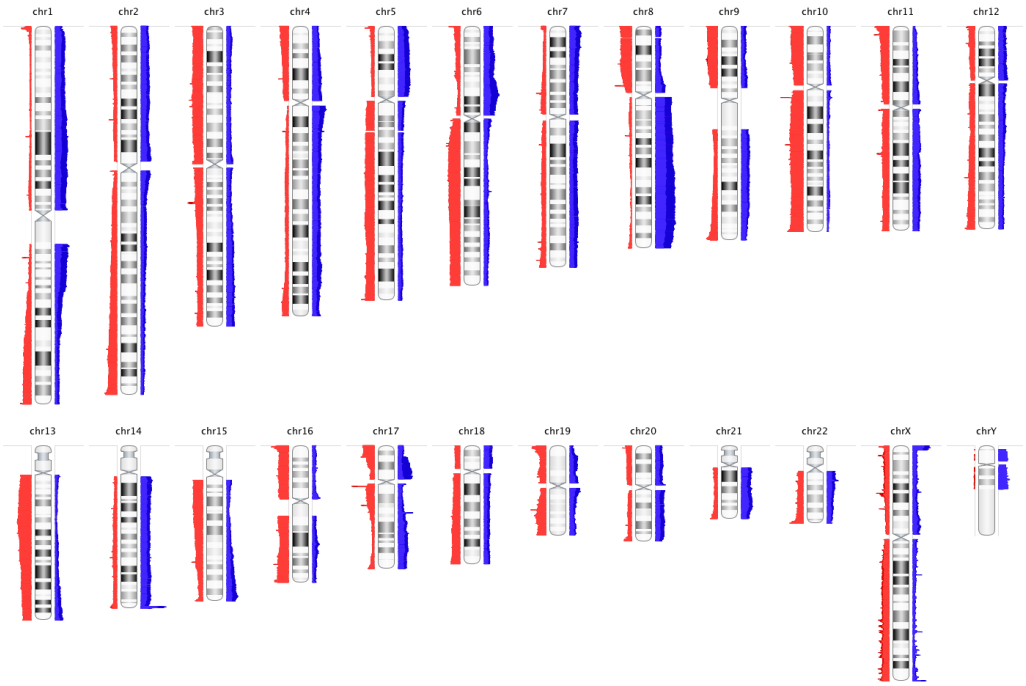


Figure 4.1: Genome-wide frequency plot of SCNAs in 157 OS samples. Copy number losses and gains are in red and blue, respectively.

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

GISTIC 2.0 [217] 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 4.2; genomic locations of these regions have been listed in Supplementary Table A.7). The annotation of GISTIC regions revealed 101 targeted genes (listed in Supplementary Table A.8), 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) [361] — 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 [344, 362, 363]. Besides these well-known OS driver genes, GISTIC regions contained several other OS-related genes, such as *RUNX2* and *DLG2* [344, 364].

Analysis also revealed novel or recently described genes — *FOXN1* and *WWOX*. *FOXN1* (17q11.2) is the main transcriptional regulator of thymic epithelial cell development, differentiation, and function [365]. 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 [366]. 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 [367–369]. 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 [370].

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 occur is a powerful means for identifying genes with fundamental roles in oncogenesis. SubcloneSeeker [354] succeeded in inferring subclone structures for 99.4% of tumors (156 out of 157). The mean number of predicted subclone

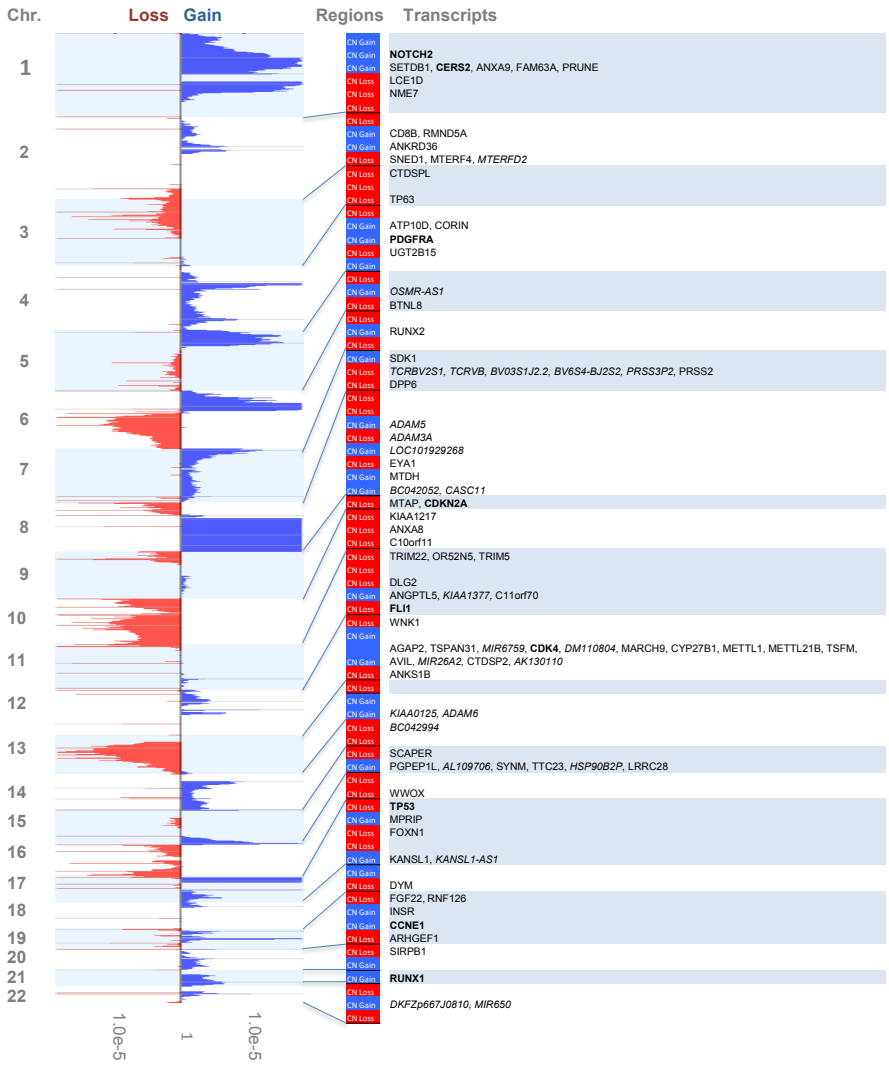


Figure 4.2: Significantly altered regions and genes contained therein with copy number alterations in osteosarcoma as identified by GISTIC analysis

structures for each tumor was 8.5 (ranging from 1 to 45). Thirty-six tumors had >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 ($\approx 90\%$) of SCNAs were subclonal events [363]. Thirty-four tumors had clonal SCNAs overlapping 1-10 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.

4.3.3 Breakage analyses reveal osteosarcoma-specific unstable 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 $\approx$ 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 the $\log_2$ signal value alteration between two adjacent genomic segments $>$0.3 (Figure 4.3), which is more stringent than the cutoff of 0.23 previously used [371]. 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 mega base 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 the density of chromosomal breaks per block calculated. 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", 11% are located within common fragile sites, while 46% overlapped with non-fragile sites [307], indicating apparent OS-specific instability characteristics.

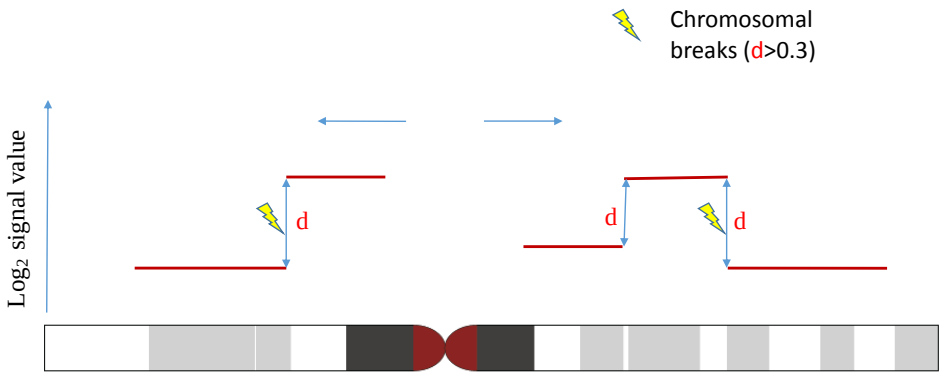


Figure 4.3: Schematic illustration of chromosomal breaks. "d" means log_2 value changes between two adjacent genomic segments at a specific genomic position.

Some of the OS-associated tumor suppressor genes [347], including *TP53*, *RB1*, *WWOX*,

DLG2 and *LSAMP*, but no known OS oncogenes, were located in these broken regions (Figure 4.4). 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 (10 662 versus 4 579, *P*-value=0), implicating chromosomal breakage as a clonal event of early occurrence in tumorigenesis.

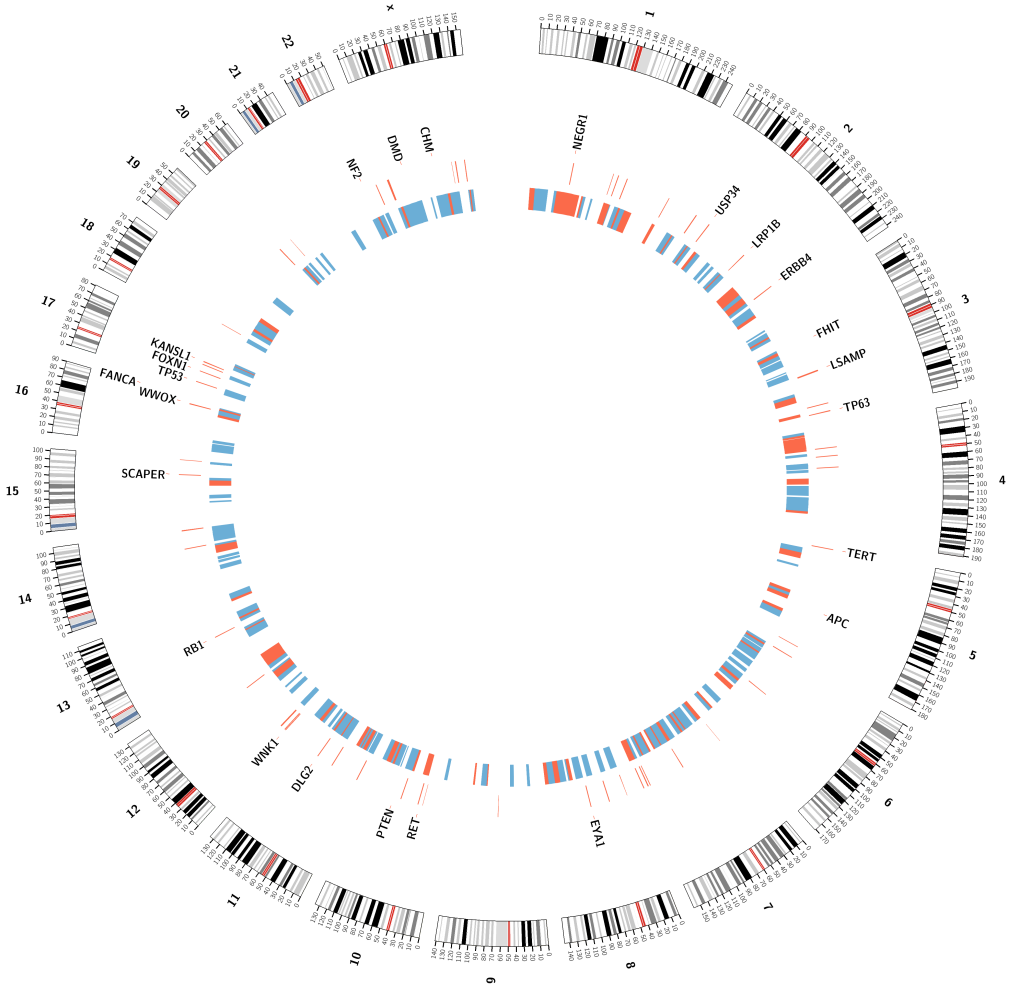


Figure 4.4: The genomic landscape of chromosomal breaks and associated genes in OS. The outermost circle represents chromosomes and cytogenetic bands. The next circle represents known OS driver genes and other genes as listed in Table 4.2. The third circle represents "broken regions". The innermost circle shows common fragile sites and non-fragile regions in red and blue, respectively.

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 limited by a degree of uncertainty in calling the location of chromosomal breaks (due to the inter-probe distance characteristic

for CytoScan HD arrays), can nevertheless 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 4.2) have been previously shown to be associated with OS (*DLG2*, *WWOX*, *TP53*, *RB1*, *LSAMP*, *PTEN*, and *APC* [347]) 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 4.5 the largest number of chromosomal breaks was located in the first intron of this gene [344, 362].

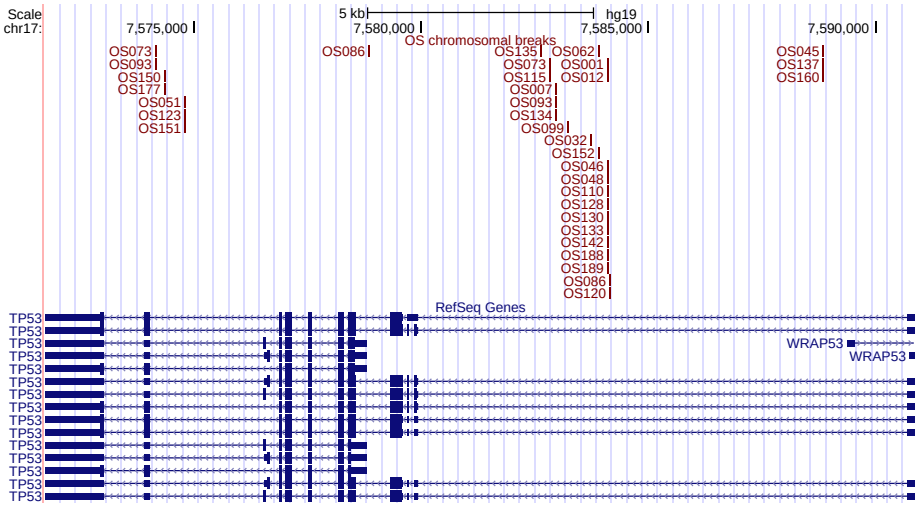


Figure 4.5: Plot of chromosomal breaks around the *TP53* gene.

4.3.4 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 1 Mb 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 [268, 295, 306]. 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 [295, 297, 372]. As expected, SCNA breakpoints and chromosomal breakage are highly correlated (*P*-value $<2.20 \times 10^{-16}$, Spearman rho = 0.76). In

Table 4.2: Genes frequently targeted by chromosomal breaks in osteosarcoma that were previously shown to associate with osteosarcoma 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 [347] are in bold;
gene names identified by GISTIC analysis in this study are in italics.

addition, it was also noted that SCNA breakpoints and chromosomal breaks were significantly correlated with diverse genomic properties, including Alu, L1, Cruciform, G4, Slip, Triplex, Z-DNA, exon density, and indel rate (Bonferroni corrected *P*-values <0.01; Table 4.3).

We further examined the association of genomic properties to chromosomal breaks at a higher resolution. Specifically, windows of 10, 20, 50, and 100 kb centered around each chromosomal break were analyzed with subsequently merging of overlapped windows. 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 4.4; Bonferroni corrected *P*-values <0.01, Mann-Whitney test). These genomic features have been associated with

Table 4.3: 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.

SCNA breakpoints in different cancer types [297], suggesting that OS 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 4.4), indicating that OS has perhaps very specific breakage characteristics that include already known common fragile sites as well as unique sites of instability.

4.3.5 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 within this dataset, implying that chromothripsis is a widespread phenomenon in OS. This incidence rate was largely consistent with a previous study of

Table 4.4: 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).

a small sample size of bone cancers [37]. CTLPs had a tendency to occur frequently on chromosomes 8 (11.5%) and 17 (9.2%). The OncoPrint shown in Figure 4.6 provides an overview of SCNAs in specific genes and CTLP affecting individual samples. Chromosomal aberrations in *TP53* occurred 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 analyzed three genes — *RB1*, *WWOX* and *DLG2* — that frequently harbor structural variation in OS [344]. 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 harbored aberrations in *DLG2*, compared with 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 CTLPs in OS.

Furthermore, an investigation of the association between CTLPs and clinical data was performed [373]. 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 4.7a, 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=7.06×10^{-4}).

A successful estimation was made of tumor ploidy and tumor content for 90.4% (142/157)

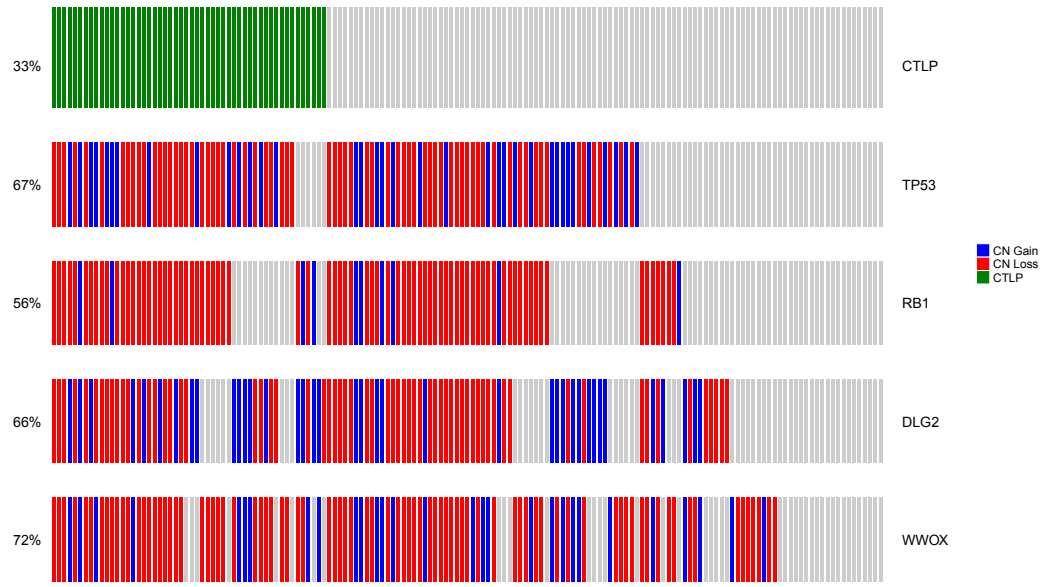


Figure 4.6: OncoPrint showing the distribution of SCNAs (CN gain and CN loss) for genes *TP53*, *RB1*, *DLG2* and *WWOX* and CTLP in OS patients (column). Each bar represents a sample. Green bars indicate samples with CTLP. Red and blue bars indicate samples with CN loss and CN gain for a specific gene, 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.

of samples using the GPHMM algorithm. These OS biopsies were estimated to have on average 37.5% normal tissue contamination with a median ploidy of 2.7n. Following the procedures for chromosome number estimation (as described in the "Materials and Methods"), the distribution of chromosome numbers was plotted in 142 samples to clearly demonstrate a two ploidy status of the tumor genome (Figure 4.7b). Near-diploid was defined for tumors with chromosome number $<$60 and DNA index $<$1.3 (see "Materials and Methods" for details), without consideration for SCNAs presence or absence in tumors. Near-tetraploid tumors had greater chromothripsis events than those classified as near-diploid (Figure 4.7c, *P*-value=4.60×10^{-3}, Fisher's exact test). This was compatible with results from a recent study linking chromothripsis with hyperploidy [374]. Patients with tumors exhibiting near-tetraploid genomes had poorer survival compared with patients having tumors with estimated ploidy of ≈ 2 (Figure 4.7d).

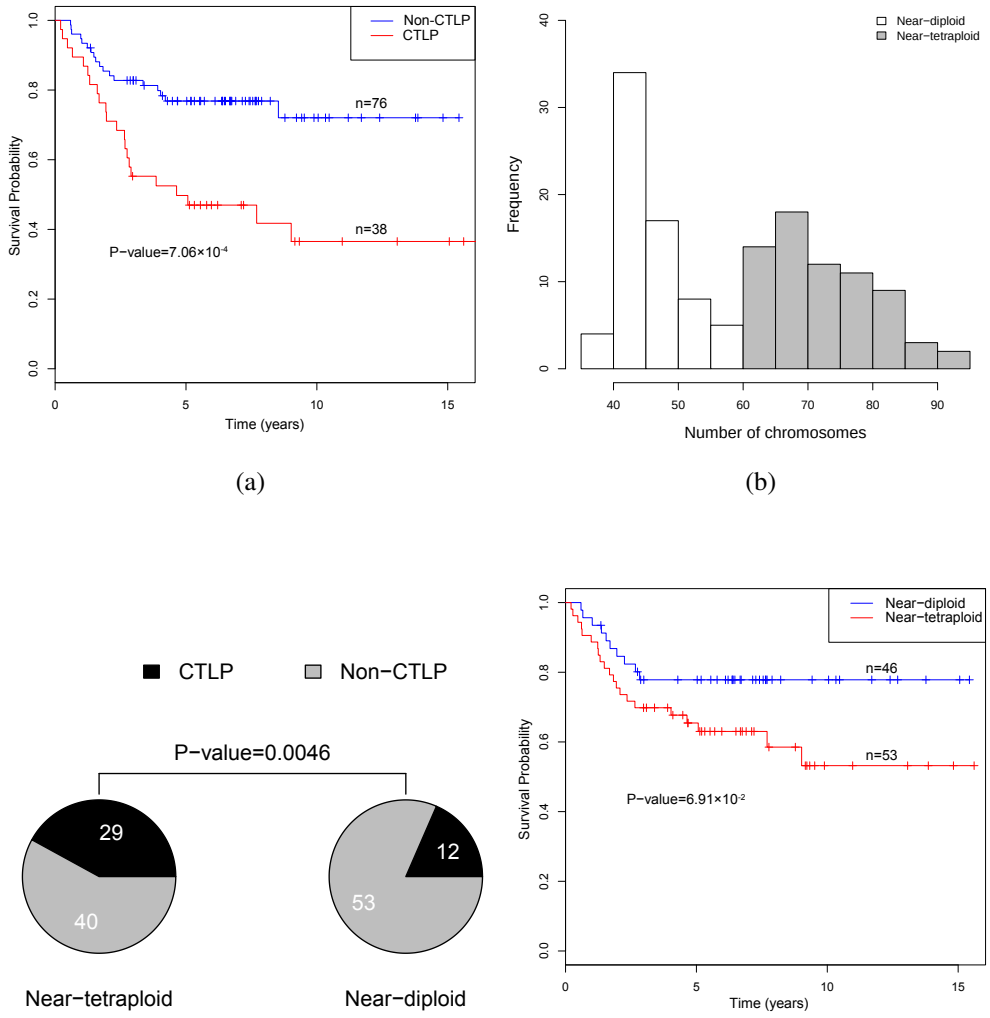


Figure 4.7: Clinical implications of chromothripsis and ploidy. (a) Kaplan-Meier survival curves for CTLPs versus non-CTLP cases. The *P*-value is based on the log-rank test; (b) Distribution of chromosome numbers in 142 OS samples, displaying the two ploidy status of tumor genomes; (c) Association of the ploidy status with chromothripsis; (d) Kaplan-Meier survival curves for near-tetraploid samples versus near-diploid samples. The *P*-value is based on the log-rank test.

4.4 Discussion

Rarity and genomic complexity, as well as marked intra- and intertumoral heterogeneity, have challenged the molecular characterization of OS etiology [347]. 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 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 [363]. 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 $>$80% of OS [363].

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 [375, 376]. In previous reports by others, the function loss of WWOX has been linked to chromosomal deletions and translocations as well as loss of expression [367, 369]. 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 more likely to be of early occurrence. The results are consistent with the hypothesis that loss of WWOX expression is an early event in OS pathogenesis [369].

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. Our observations comply with the findings that unstable sites are tissue specific [355]. It is noteworthy that OS-associated tumor suppressor genes including *TP53*, *RB1*, *WWOX*, *DLG2*, and *LSAMP* [347] 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 [330]. 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 [329]. Therefore, it could be speculated that breakages probably occur at OS-specific fragile sites with the potential to form stable secondary structures (e.g., non-B DNA structures) and to consequently stall the replication fork.

Based on 20 patients including 9 OSs and 11 chordomas, Stephens *et al.* [37] estimated that 25% of bone cancers were associated with chromothripsis. In our dataset, CTLPs occurred in about one third of patients, suggesting that chromothripsis is a widespread phenomenon in OS. Massive genomic rearrangement raised by chromothripsis apparently represents an important mechanism of carcinogenesis, as distinct from progressive accumulation.

Although the underlying cause of chromothripsis is not fully understood, several hypotheses have been recently proposed [37, 377, 378]. Firstly, chromothripsis might occur by ionizing radiation induced DNA damage at a short or long stretch of the chromosome [37, 377]. Secondly, telomere attrition may cause dicentric chromosomes which persist through mitosis developing into chromatin bridges that further generate single-stranded DNA and trigger DNA repair [37, 379]. Thirdly, abortive apoptosis has also been considered as a possible mechanism [377], but it does not provide a reasonable explanation for the localization of DNA shattering [378]. Fourthly, premature chromosome compaction, in which chromosomes are induced to undergo chromosome condensation before completing DNA replication, results in shattering of the incompletely replicated chromosome [380]. An appealing explanation for chromothripsis is that the localized damage could occur in one or two chromosomes (or chromosome part) physically isolated from other chromosomes [381]. The so-called nuclear structure “micronuclei” are widely observed in cancer cell lines. Taking advantage of live cell imaging and single-cell genome sequencing, Zhang *et al.* demonstrated that chromatid fragmentation and subsequent reassembly occur in the micronucleus and can generate localized genomic rearrangements, some of which recapitulate all features of chromothripsis [382]. Investigation of the association between lesions in specific genes and chromothripsis will offer some insights into the impact of chromothripsis in cancerogenesis. Our analysis indicates that SCNAs in the

TP53, *RB1*,*WWOX* and *DLG2* genes are strongly associated with CTLPs in OS. Among them, *DLG2* frequently shows breakages in OS and may be a preferential target for chromothripsis and breakage [344]. *RB1* is significantly copy-number altered in OS, while the other candidate, *TP53*, has already been linked to chromothripsis in medulloblastoma [350]. Utilizing an *in vitro* cell-based system, chromothripsis has been recently linked to hyperploidy [374]. Indeed, we have shown that compared with diploid tumors, those which are hyperploid had a greater chance to harbor chromothripsis events and less favorable outcomes.

5

Preliminary analysis of somatic mutational landscape of Marek's disease lymphomas in chickens

This chapter present the preliminary results of an ongoing project collaborated with Alexander Steep from the Avian Disease and Oncology Laboratory in Michigan State University. This chapter was written with the help from him. He did wet lab experiments and I did the most of bioinformatics analyses. Marek's Disease Virus (MDV) induce Marek's Disease (MD) in chickens, which is characterized by T-cell lymphomas. It was estimated that MD costs the world-wide poultry industry 1-2 billion US dollars per year. Although vaccination against MDV-induced transformation has been successful in stopping the formation of neoplasms in infected chickens, high-density poultry rearing practices have induced MDV evolution and increased MDV virulence. To develop more sustainable MD control measures, a fundamental understanding of the molecular etiology of tumorigenesis is needed. Here we used multiple approaches including whole genome sequencing, whole transcriptome sequencing, and DNA microarrays to explore the somatic mutational landscape of MD. We identified 54 high confidence driver genes, of which *IKZF1* encodes a transcription factor associated with hematopoietic cell differentiation and has been linked to the development of lymphoid leukemia. Our results contribute to the understanding how somatic mutations drive transformation and lymphomagenesis, and will likely be used to guide future disease control.

5.1 Introduction

Marek's Disease (MD) is a lymphoproliferative disease in chickens caused by Marek's Disease Virus (MDV), a highly oncogenic α-herpesvirus [383]. Originally described as a sporadic chronic disease, MD has increased severity drastically since its discovery in 1907 [383], and today is manifested by paralysis, chronic wasting, and most notably the development of multiple lymphomas in the viscera and musculature [384]. It was estimated that MD costs the world-wide poultry industry 1-2 billion US dollars per year, which is likely an underestimation due to under-reporting [385]. The main strategies for disease control focus on good housekeeping, optimization of genetic resistance to MD, and vaccination against MDV-induced transformation [386]. As the first-developed cancer vaccine, the vaccine against MDV successfully stops the formation of neoplasms in infected chickens [387]. Despite the success of vaccination, multiple vaccine breaks occurred throughout the second half of the 20$^{\text{th}}$ century [383, 388]. One explanation for vaccine breaks is that vaccination control enhances the fitness and transmission of highly virulent strains when vaccines are "leaky" and let the host survive but do not stop viral proliferation and transmission [389]. To develop more sustainable MD control measures, a fundamental understanding of the molecular etiology of tumorigenesis (e.g., somatic mutations driving transformation and tumorigenesis) is needed [388].

Virulent MDV goes through four overlapping stages of infection: early cytolytic, latent, late cytolytic, and transformation [383, 390]. MDV infection begins when chickens inhale dander shed from infected chickens. Infectious dander is taken up by phagocytic cells in the upper respiratory tract [391] and the virus is transported from the lungs to the lymphoid tissues: the spleen, thymus and bursa of Fabricius [392]. In these organs the early cytolytic phase, characterized by the infection of B lymphocytes, is evident between 3 and 6 days post infection (dpi) [393]. The infected B cells induce the activation of CD4$^+$ T cells, which in turn become infected and serve as the primary vehicle for MDV multiplication and dissemination. From 7 dpi MDV enters a latent stage defined as the presence and maintenance of viral genome without virus production [383]. Infected T cells carry the virus to the feather follicle epithelium, where infectious virus is assembled and shed into the environment [394]. In susceptible and/or unvaccinated birds, a second cytolytic phase occurs between 14 and 21 dpi and infected CD4$^+$ T cells become transformed and develop into fatal lymphomas [384]. The cell subpopulation transformed by MDV are identical to those in which latent infection is established, suggesting that latent stage is

necessary for transformation [383].

The rapid onset of MD lymphomas suggests that genes carried on the MDV genome and somatic alterations are directly involved in oncogenic transformation and lymphomagenesis. A major MDV oncogene is *Meq* (MDV EcoRI Q), which encodes a basic leucine zipper (bZIP) transcription factor [395]. Meq is expressed in lytically infected cells and T-cell tumors with a variety of functions including transactivation, DNA binding, chromatin remodeling and transcriptional regulation [383, 396]. Previous studies indicate that Meq is necessary but not sufficient for MDV transformation, as deletion of *Meq* from a very virulent MDV strain results in no lymphomas [397] and Meq is encoded and expressed in non-oncogenic MDV strains [398]. It has been revealed that MDV can integrate into chicken chromosome and the integration seems to be random and common [399]. MDV integration is necessary but does not guarantee transformation [400], suggesting that additional somatic alterations are needed. Given that MDV oncogenes (e.g. *Meq*) regulate many host genes and pathways, it is believed that the interplay among somatic alterations, MDV-induced gene expression regulation, and MDV integration plays a fundamental role in MDV-induced transformation.

To test the hypothesis whether or not somatic mutations contribute to MDV-induced transformation, we use multiple approaches including whole genome sequencing (WGS), whole transcriptome sequencing (i.e., RNA sequencing), and DNA microarrays to explore the somatic mutational landscape of MD. The combination of approaches has not yet applied to MD and is expected to reveal novel insight into MDV-induced tumorigenesis. We identified 54 high confidence driver genes, whose functions include cell adhesion, cell signaling, cellular proliferation, cell differentiation and immune response. Notably, we found that disruptive mutations together with low gene expression of *IKZF1* occurred in 12 of 26 (46%) MD tumors. IKZF1 has been found to play an important role in hematopoietic cell differentiation and its loss of function has been linked to the development of lymphoid leukemia. Our results contribute to the understanding of how somatic mutations drive transformation and lymphomagenesis, and will likely be used to guide future disease control.

5.2 Materials and Methods

5.2.1 Experimental birds, materials, and tissue sampling

White Leghorn chicken lines 6 and 7 differ greatly in susceptibility to MD (line 6 resistant and line 7 susceptible), and have been extensively studied to characterize the genetic basis for MD resistance. In this experiment, we used F_1 progenies of lines 7×6 and parental lines maintained in Avian Disease and Oncology Laboratory, United States Department of Agriculture. A total of 200 F_1 chicks were inoculated at hatch with 1 000 plaque forming unit (pfu) of the MDV JM/102W strain, which is classified as virulent type in the virulent spectrum of MDV ranging from mildly virulent (m) to virulent (v), very virulent (vv) and very virulent plus (vv+) [401]. After 8 weeks post-inoculation or until moribund (4-8 weeks), chickens were necropsied for large late-stage tumors for the sake of easy discrimination between tumor and surrounding tissue. In total, we collected 162 tumors from 54 birds, and 64 of these tumors from 36 birds were used for different purposes: SNP array genotyping, WGS and RNA sequencing.

Blood sampled from a pool of 6 F_1 birds not challenged with MDV was used as controls for SNP array genotyping. For controls used in WGS, liver or the other gonad was sampled from the infected bird. For controls used in RNA sequencing, RNAs were isolated from purified $CD4^+$ T cells, the cell type most likely to be transformed by MDV, from unchallenged chickens of 2 and 6 weeks of age.

5.2.2 Whole genome sequencing, whole transcriptome sequencing and SNP array genotyping

DNA from 22 unchallenged F_1 birds and 26 MD gonadal tumors specifically selected from the 162 MD tumors mentioned above were subject to WGS. Four tumors (26-22=4) were sequenced twice with a biological replicate using DNA isolated from adjacent cells. Parental lines 6 and 7 were also sequenced to provide reference information. For RNA sequencing, RNAs from 8 unchallenged F_1 chickens and the same 26 MD tumors were subjected to sequencing. DNA from 6 F_1 birds and 72 MD tumors were subject to genotyping on a custom 15K Affymetrix SNP array. A pool of 6 Line 6 birds and a pool of 6 Line 7 birds were used as reference samples.

Libraries for WGS and RNA sequencing were constructed using Illumina TruSeq Nano DNA Library Preparation Kit and Illumina TruSeq Stranded mRNA LT kit, respectively. Libraries were sequenced with an Illumina HiSeq 2000 to obtain 125 bp paired-end reads. Base calling was performed with the Illumina Real Time Analysis v1.18.64 and the output was demultiplexed and converted to FastQ format with the Illumina Bcl2fastq v1.8.4.

5.2.3 Analyses of whole-genome sequencing data

5.2.3.1 Read trimming, quality control, mapping, and post-processing

Adapters were trimmed from raw reads with Trimmomatic [402]. Read trimming via base quality was performed using Sickle [403]. Trimmed reads were quality checked by using FastQC [404]. The mapping and post-processing pipeline was designed following the Genome Analysis Toolkit (GATK) best practices [405]. Trimmed reads of ample quality were mapped to the chicken genome assembly Gallus_gallus-5.0 using BWA-MEM in the BWA package [406]. Information for read groups was added using Picard (available at `https://broadinstitute.github.io/picard/`). At both "sequencing lane" level and "merged lane" level, duplicate read marking and local indel realignment were performed using Picard and GATK (available at `https://software.broadinstitute.org/gatk/`), respectively. Versions and parameters for these tools can be found at `https://github.com/hongenxu/MDV_proj`.

5.2.3.2 Detection of SNVs, indels, SCNAs, LOH and SVs

Somatic mutations include single nucleotide variants (SNVs), small insertions and/or deletions (indels), somatic copy number alterations (SCNAs), structural variants (SV) and loss of heterozygosity (LOH). The detection of somatic mutations was performed using tumor and matched normal whole genome BAM files generated in the steps described above. Different callers use different calling strategies based on their underlying statistical or hierarchical assumptions. It is generally believed that candidate mutations detected by several independent algorithms is less likely to include false positives than those by a single algorithm alone [57]. We employed multiple software tools to each type of somatic mutations including SNVs, indels, SCNAs, SVs and LOH. Versions and parameters for these tools can be found at `https://github.com/hongenxu/MDV_proj`.

We used a series of software packages including MuSE [74], MuTect [66], JointSNVMix2

[65], SomaticSniper [64], VarDict [70], and VarScan 2 [69] to detect somatic SNVs. For indel detection, we used Indelocator [77], VarDict, VarScan 2 and LoFreq [68]. The details for post-filtering of SNV and/or indel calls by each tool are available at `https://github.com/hongenxu/MDV_proj`. All filtered SNV and indel calls by different callers were combined using a post somatic mutation calling workflow—SomaticSeq [73]. We selected SNV candidates called by at least two callers for the following analyses.

To characterize SCNAs in MD lymphomas, Control-FREEC [89, 407] and copyCat (available at `https://github.com/chrisamiller/copyCat`) were used. For control-FREEC, we further filtered out somatic SCNAs and LOH failed in both Wilcoxon test and Kolmogorov–Smirnov test ($P > 0.05$). Details for generating chicken GC content and mappability data for running each tool can be found at `https://github.com/hongenxu/MDV_proj`.

WGS allows us to characterize somatic SVs and their breakpoints in base-pair resolution. In order to reduce the number of false positives, we used an integrated approach combining three callers: BreakDancer [97], Delly [100] and novoBreak [408]. Versions, parameters and post filtering strategies for these tools can be found at `https://github.com/hongenxu/MDV_proj`.

5.2.3.3 Inference of somatic mutational signatures from SNVs

Somatic mutational signatures are patterns in the occurrence of somatic SNVs that linked to potential mutagenic processes. In order to infer the mutational signatures of MD, we used the R package *SomaticSignatures* [255], in which a mutation spectrum was decomposed with non-negative matrix factorization algorithm or principal component analysis. The decomposition was performed on a known number of signatures ranging from 2 to 8 using non-negative matrix factorization based. The optimal number of signatures ($r = 5$) was manually chosen based on the residuals sum of squares and the explained variance between the observed and fitted mutational spectrum. The signature analysis was repeated 5 times with the same results obtained after each run.

5.2.3.4 Determination of driver mutations, genes and pathways

To determine which of the detected somatic mutations are likely driver mutations, we used PROVEAN (Protein Variation Effect Analyzer) to predict whether or not a non-synonymous SNV has an impact on the function of a protein [168]. With somatic SNV

and indel calls, significantly mutated genes (SMGs) were identified using three tools of different strategies: mutation recurrence (MuSiC [183]), CLUST bias of mutations (OncodriveCLUST [188]), and network analysis (MUFFINN [409]). For MuSiC analysis, a gene was considered to be a SMG if its false discovery rate (FDR) ≤ 0.2 for at least two of three tests: Fisher's Combined P-value test, Likelihood Ratio test, and the Convolution test. For OncodriveCLUST analysis, the minimum number of mutations to include a gene was set to 1. Genes with q-value < 0.05 were considered as SMGs. Since MUFFINN is designed for human cancers, human orthologs of chicken genes downloaded from InParanoid version 8.0 [410] were used for MUFFINN analysis. Somatic mutation data for chicken genes were subject to analysis as a list of human orthologs and mutation count pairs. We considered 9 mutations types, including missense mutations, nonstop mutations, nonsense mutations, mutations at translations start sites, mutations at splice sites, in frame insertions, in frame deletions, frame shift insertions, and frame shift deletions. A gene was considered to be a SMG if it was ranked in top 100 by at least two of four different combinations between network algorithms (DNmax and DNsum) and functional networks (HumanNet and STRING v10) for MUFFINN. Versions and parameters for running these tools can be found at `https://github.com/hongenxu/MDV_proj`. The mutation waterfall plot and mutation hotspot plot were created using the R package *GenVisR* [411]. The recurrent LOH regions were detected using the HD-CNV tool [412].

5.2.4 Analyses of whole transcriptome sequencing data

5.2.4.1 Read trimming, quality control, mapping, and post-processing

Sequencing adapters were trimmed from raw reads with Trimmomatic [402]. The FastQC tool [404] was used to check the quality of trimmed reads. Trimmed reads were mapped to chicken genome assembly Gallus_gallus-5.0 using the 2-pass mapping of the STAR aligner with default parameters [413]. Duplicate reads were marked with the MarkDuplicates command of Picard (available at `http://broadinstitute.github.io/picard`). Versions and parameters for these tools can be found at `https://github.com/hongenxu/MDV_proj`.

5.2.4.2 Differential expression analysis

The number of aligned reads (skipping duplicate reads) within each gene were counted by using the featureCounts tool implemented in the subread package [414]. Gene annotations for the chicken genome was downloaded from Ensembl (Ensembl release 86). Normalization of read counts and estimation of fold change was carried out using R pacakage *DESeq2* [415]. Differentially expressed genes were selected using the function *results* with $alpha = 1 \times 10_{-5}$ and $lfcThreshold = 1$.

5.2.4.3 Gene ontology enrichment analysis of differentially expressed genes

Gene ontology enrichment analysis was performed at Gene Ontology Consortium website (available at `http://geneontology.org/page/go-enrichment-analysis`), which is based on PANTHER [416]. PANTHER Overrepresentation Test (release 20170413) and Gene Ontology Database (release 2017-05-25) were used. The analyses were performed for up-regulated and down-regulated genes in MD lymphomas, separately.

5.2.5 Analyses of DNA microarray data

A custom 15K Affymetrix SNP array was used to genotype 72 MD tumors and a pool of six F_1 uninfected birds. SNP arrays allow simultaneous measurement of the allele-specific copy number at multiple loci in the genome. For each probeset, the log R ratio (LRR) reflects the ratio of total signal intensity for both alleles to expected signals, and the B allele frequency (BAF) is an estimate of the relative proportion of one of the alleles with respect to the total signal intensity. LRR and BAF values were calculated using the tool PennCNV-Affy [126] following the guidance at `http://penncnv.openbioinformatics.org/en/latest/user-guide/affy/`. The resulting LRR and BAF were used as input for ASCAT [81] and genoCNA [79] to identify allele-specific copy number. Population frequency of the B allele (PFB) required for running genoCNA was created using the Perl script *compile_pfb.pl* in PennCNV [126] with BAF values from the six normal samples as input. Versions, parameters and detailed usage for these tools can be found at `https://github.com/hongenxu/MDV_proj`.

5.3 Results and Discussion

5.3.1 The overview of the study design

The overview of our study is presented in Figure 5.1. The high inbreeding of both lines 6 and 7 as well as the heterozygous nature of F_1 progeny enable us to characterize tumor specific somatic alterations in reference to both parental lines. A total of 200 F_1 chicks were challenged at hatch with JM/102W MDV strain. The main reason this strain was chosen is that it preferentially induces large, fairly homogeneous gonadal tumors instead of diffuse spleen tumors. From these birds, we collected 162 tumors and used them on different technology platforms. Of these, 72 together with 6 F_1 uninfected birds were subject to a custom Affymetrix SNP array. A total of 26 tumors (with 4 replicates) and matched normal tissues were characterized by WGS. The transcriptomes of 26 tumors and 8 normal controls were profiled by RNA-sequencing. The control samples for SNP array genotyping and RNA sequencing were not from infected chickens. We argue that the high level of inbreeding and genetic consistency of F_1 progeny render it unnecessary to have controls for each individual.

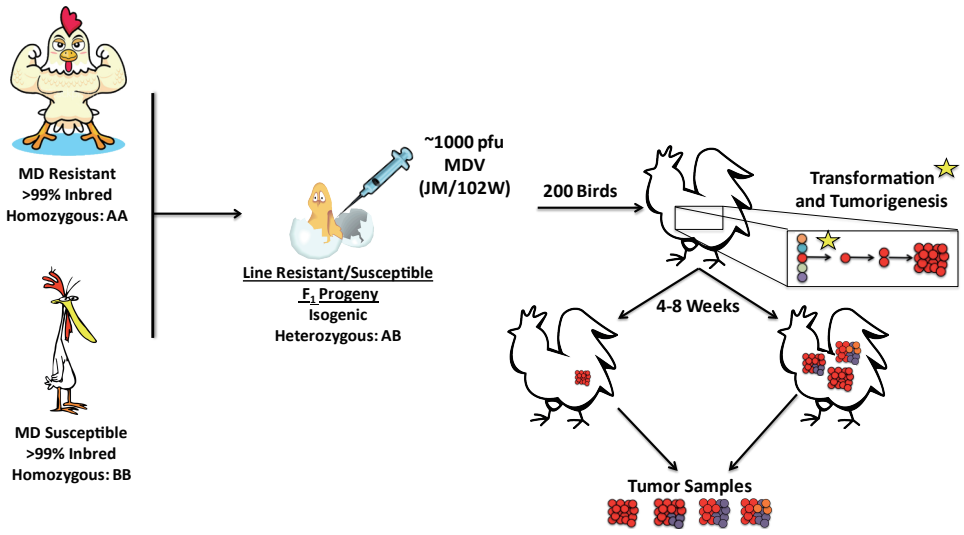


Figure 5.1: An overview of the study design.

The SNP array has the capacity to survey $\approx$ 9K SNPs, the majority of which are completely fixed and differ between parental lines and are therefore heterozygous in F_1 chickens. WGS yielded a median of 16$\times$ (ranging from 12$\times$ to 22$\times$) and 12$\times$ (ranging from

9× to 15×) depth of coverage for normal and tumor tissues, respectively. For RNA sequencing, two samples were removed either due to low number of reads or low mapping rate.

5.3.2 Somatic SNVs and indels in MD lymphomas

WGS of 26 paired tumor and normal samples (including 4 biological replicates) allowed us to identify somatic SNVs and indels. We employed 6 callers for characterizing SNVs and 4 callers for indels. We used filtering strategies designed in each caller or recommended by post-processing workflow—SomaticSeq [73] to reduce false positives. For SNVs, we only consider those identified by at least two tools. We characterized a median of 5114 (ranges from 3062 to 7150) somatic SNVs/indels per tumor. In the coding regions of the chicken genome, we detected a median of 56.5 (ranges 25 to 76) non-synonymous SNVs. This number is larger than pediatric and liquid cancers such as glioblastoma, medulloblastoma and leukemia, but less than most adult solid tumors, such as head and neck squamous cell carcinoma and colorectal cancer [215].

To understand the DNA damage and repair processes those have been operative in MD, we inferred the mutational signatures from mutation spectra using the SomaticSignatures package with non-negative matrix factorization method [255]. For MD, the most common substitution was the C·G→T·A changes, followed by the T·A→C·G changes (Figure 5.2a). We identified five independent mutational signatures in the MD cohort (Figure 5.2b), and these signatures matched the previously identified signatures described in the Catalogue of Somatic Mutations in Cancer (COSMIC) database (`http://cancer.sanger.ac.uk/cosmic/signatures`). Signature S1 and S2 are mainly characterized by C·G→T·A substitutions, and corresponds to COSMIC signature 7 and 11, respectively. These two signatures are strongly represented in glioblastoma and skin cutaneous melanoma, respectively (Figure 5.2c). Signatures S3 and S4 show a broad distribution across the motifs and may match COSMIC signature 4. Signature S3 occurred in head and neck squamous cell carcinoma and thyroid carcinoma, whereas signature S4 was mainly found in lung adenocinoma and lung squamous cell carcinoma (Figure 5.2c). Signature S5 closely resembles COSMIC signature 5, which is characterized by a broad spectrum of base changes and more C·G→T·A and T·A→C·G changes. Notably, signature S5 contributed >75% in MD and >50% in kidney renal clear cell carcinoma. A recent study has linked COSMIC signature 5 in urothelial tumors to a nucleotide excision repair gene—

ERCC2 [260]. An etiology of signature 5 in MD remains to be elucidated.

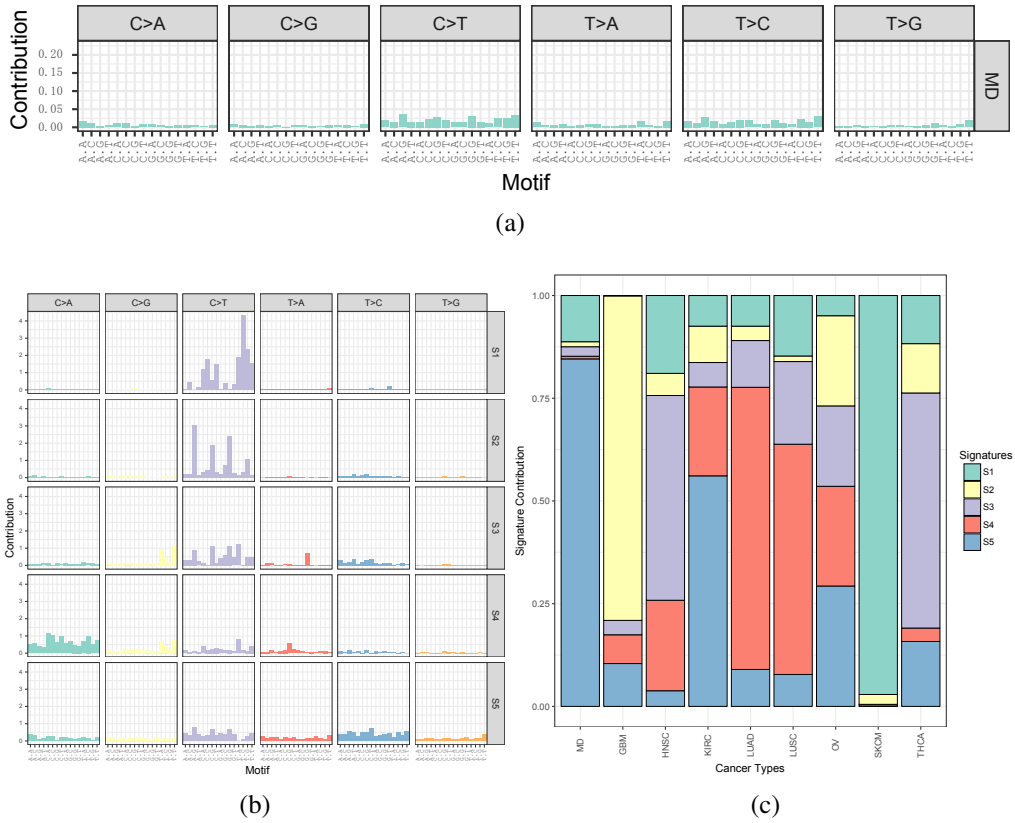


Figure 5.2: Mutational signatures of MD lymphomas. (a) Relative frequency of 96 motifs defined by the base substitution class (top) and the 5' and 3' adjacent bases (bottom) in MD lymphomas. (b) Composition of five somatic signatures (named S1 to S5) estimated from the matrix of mutation counts across MD lymphomas. (c) The contribution of five somatic signatures to MD lymphomas and eight cancer types of The Cancer Genome Atlas (TCGA) project. MD: Marek's Disease; GBM: Glioblastoma Multiforme; HNSC: Head and Neck Squamous Cell Carcinoma; KIRC: Kidney Renal Clear Cell Carcinoma; LUAD: Lung Adenocarcinoma; LUSC: Lung Squamous Cell Carcinoma; OV: Ovarian Serous Cystadenocarcinoma, SKCM: Skin Cutaneous Melanoma, THCA: Thyroid Carcinoma.

To determine the accuracy of SNV calling and custom filtering, we used the genotyping information from custom 15K Affymetrix SNP array performed on the same samples. The SNP array has 13 665 SNPs, of which, only 49 were also present in the WGS. Thus, genotyping data are not sufficient to validate SNV calls from WGS. Targeted sequencing will be used to validate SNVs and evaluate the accuracy of our SNV calling and custom filtering.

5.3.3 Somatic SCNAs, LOH and SVs in MD lymphomas

The package copyCat can detect SCNAs by measuring the coverage depth of massive sequencing of the genome. Compared with copyCat, Control-FREEC [89, 407] is additionally able to detect LOH. We identified 542 and 253 SCNAs (including gains and losses) using copyCat and Control-FREEC, respectively. Only 9 SCNAs were detected by both tools, suggesting that there were many false positives. Notably, Control-FREEC detected a total of 2779 LOH events. We used HD-CNV [412] to identify recurrent LOH regions and got 381 genomic regions with a median size of 163 Kb (ranging from 10 kb to 11.4 Mb).

SNP microarrays have allowed simultaneous detection of copy number and copy-neutral changes on the same array. We used two methods— genoCNA [79] and ASCAT [81] to identify copy number states and genotype calls. The tool genoCNA identified 138 copy number gains, 146 copy number losses and 458 LOH events. Next we used genoCNA results from the same samples to validate SCNAs and LOH events characterized by Control-FREEC. The results showed that for copy number gains and losses, there is no overlap of SCNA calls between WGS and SNP microarray genotyping. But for LOH, 99 LOH events were called by two platforms. These LOH events were additionally confirmed by ASCAT raw outputs, although ASCAT needs additional parameters to infer the final genotype calls for our custom SNP array. The SCNA and LOH calling from both WGS and SNP array indicated that LOH events are frequent in MD lymphomas. We will use target sequencing to validate the common LOH events called by these two platforms and annotate these events to reveal their potential roles in MD lymphoma.

To identify somatic SVs in cancer genomes, computational methods are required to identify SV events and determine their breakpoints in base-pair resolution from the massive amounts of reads generated by a NGS experiment. Computational tools normally use three types of approaches: assembly, read pairs, and split reads [139]. In order to create a reliable list of SVs, we applied three methods based on complementary approaches: BreakDancer [97] (read pairs), Delly [100] (read pairs and split reads), and novoBreak [408] (assembly). BreakDancer, Delly and novoBreak identified 917, 1451 and 163 SVs, respectively. In order to remove false positives, we only considered SV events called by at least two callers. The analysis resulted in 39 high-confidence SV calls, including 28 deletions and 11 inversions. The recurrent SVs (in more than 1 sample) overlapped with genes such as *DCLK1* and *CTCF*. The encoded protein of *DCLK1* is involved in several

different cellular processes, including neuronal migration, neuronal apoptosis and neurogenesis. The gene *CTCF* encodes a transcription factor with 11 highly conserved zinc finger domains, which is involved in many cellular processes including transcriptional regulation and chromatin structure regulation [417]. CTCF has been suggested as a potential tumor suppressor factor as it modulates the expression of several key-regulators of differentiation, cellular senescence, cell cycle control and progression [418].

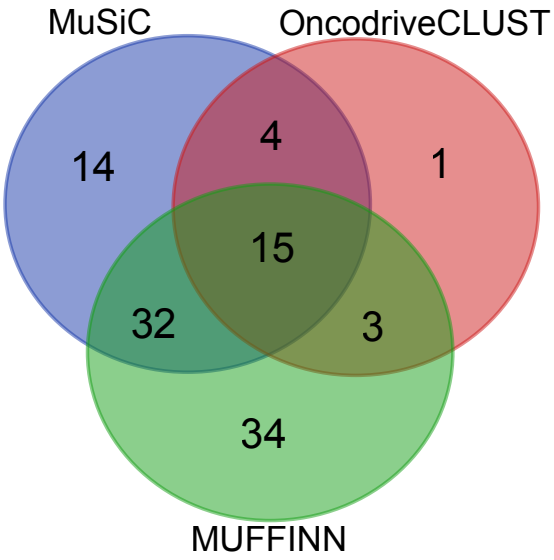
5.3.4 Driver genes and mutations

To identify a comprehensive and reliable list of driver genes, we applied three methods (MuSiC [183], OncodriveCLUST [188] and MUFFINN [409]) based on complementary approaches [220]. MuSiC identifies driver gens based on the frequency of mutations observed in genes across a cohort of tumors, whereas OncodriveCLUST relies on detecting a biased accumulation of mutations in certain regions of protein sequences. MUFFINN is a pathway-centric method, which accounts not only for mutations in individual genes but also in their neighbors in functional networks.

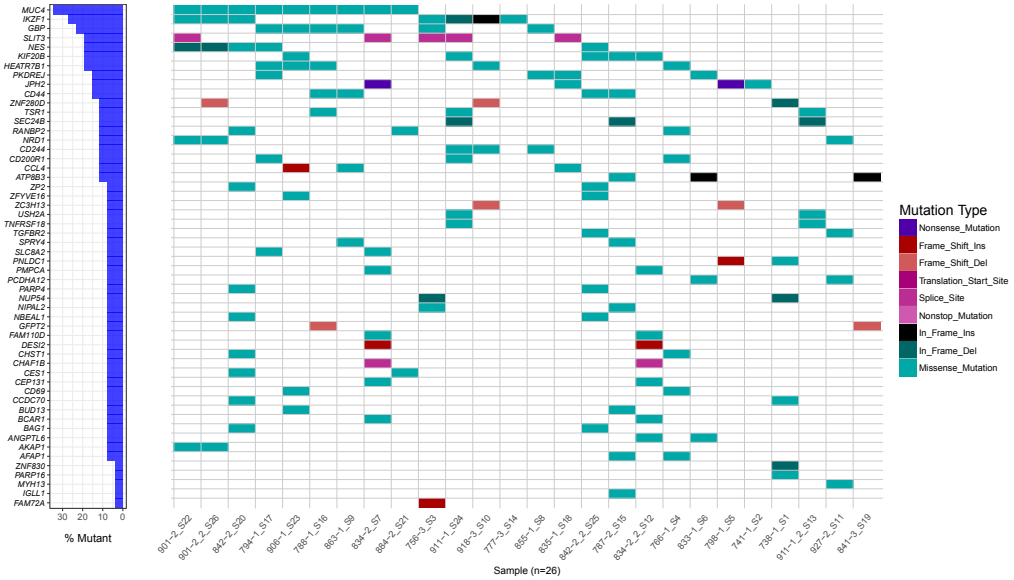
We identified different numbers of driver genes using the three methods (Figure 5.3a). We considered genes identified as drivers by at least two methods high-confidence drivers [220]. In total, we identified 54 high-confidence driver genes, and their mutation frequency and types are shown in Figure 5.3b. Genes are ordered by the presence of mutations in more to less samples. We only considered nonsense mutations, frame shift insertions, frame shift deletions, mutations at translation start sites, mutations at splice sites, nonstop mutations, in frame insertions, in frame deletions and missense mutations. The high-confidence driver genes with higher mutation frequency include *MUC4* and *IKZF1*. *MUC4* encodes a mucin glycoproteins MUC4, which may play important roles in epithelial renewal and differentiation. MUC4 can serve as a ligand for the receptor tyrosine kinase ERBB2, regulating its phosphorylation [419]. The roles of MUC4 in MD lymphomas remain to be elucidated. *IKZF1* encodes a transcription factor associated with chromatin remodeling, and its roles in lymphomas will be discussed in the following sections.

5.3.5 Differentially expressed genes in MD lymphomas

We identified a total of 1 755 genes showing significantly different expression levels between MD lymphomas and normal controls, of which 1 394 and 361 genes were expressed



(a)



(b)

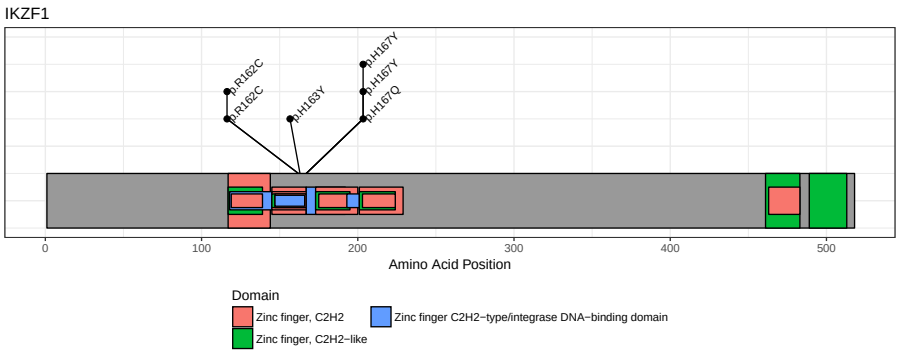
Figure 5.3: Significantly mutated genes in MD lymphomas. (a) Venn diagram showing the number of driver genes identified by each method. The names of the genes detected by 2 or more methods are shown in subfigure B. (b) The mutation waterfall plot for 54 high-confidence driver genes in MD. Driver genes and samples are displayed in rows and columns, respectively. The frequency of mutations for the genes in WGS cohort are shown in the left bar plot.

higher and lower in tumors, respectively. Unexpectedly, none of the differentially expressed genes in MD lymphomas were in the list of high confidence driver genes identified above. Gene ontology enrichment analysis of the up-regulated genes in MD lymphomas revealed an enrichment in genes associated with multicellular organismal signaling, bicellular tight junction and extracellular matrix component. For down-regulated genes in

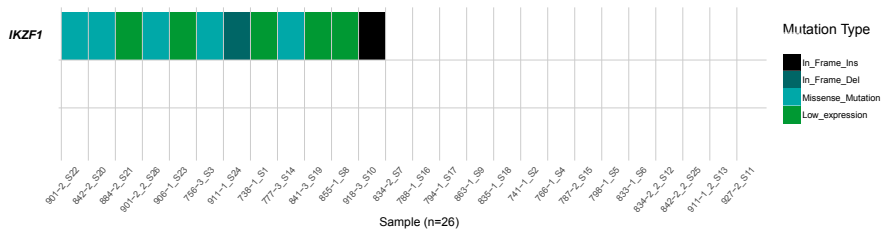
tumors, gene ontology enrichment analysis showed that they are enriched in genes related to leukocyte proliferation, leukocyte differentiation, leukocyte activation, hemopoiesis, cell activation, and positive regulation of immune system process.

5.3.6 Ikaros's Role in MD Lymphomas

The Ikaros gene family, includes IKAROS family zinc finger 1 (*IKZF1*), *IKZF2* and *IKZF3*, encodes transcription factors that belong to the family of zinc-finger DNA-binding proteins associated with chromatin remodeling. The corresponding proteins, also known as Ikaros, Aiolos and Helios, are involved in regulation of lymphoid development and differentiation [420]. The Ikaros protein (IKZF1) contains two separate regions of zinc-finger domains: 4 DNA-binding zinc fingers near the N-terminus and 2 zinc fingers for protein-protein interactions near the C-terminus (Figure 4A) [421]. Several alternatively spliced isoforms have been described for the *IKZF1* gene, and these isoforms differ in the number of N-terminal DNA-binding zinc finger motifs, resulting in proteins with and without DNA-binding properties. Mutations of *IKZF1* result in the loss of Ikaros function, and have been identified as an important event in the development of acute lymphoblastic leukemia with Philadelphia chromosome [422].



(a)



(b)

Figure 5.4: *IKZF1* gene in MD lymphomas. (a) Mutation hotspot graphic for *IKZF1* gene. (b) The mutation and gene expression waterfall plot for *IKZF1* gene.

In MD lymphomas, we showed that *IKZF1* gene harbors SNVs or indels in coding regions in 7 of 26 samples, with missense SNVs in 5 samples, in frame insertion in 1 sample and in frame deletion in 1 sample (Figure 3B). The mutation hotspot plot for *IKZF1* shows that mutations (6 missense mutations in 5 samples) are clustered in the second DNA-binding zinc finger near the N-terminus (Figure 5.4a). This gene got its name because its protein's function is very susceptible to changes in gene expression. Next we investigated the expression level of *IKZF1* in MD lymphomas and controls using RNA sequencing data. If we consider the average expression level of controls as baseline, we found 5 MD samples with much lower gene expression ($<$ one fourth of baseline). Notably, these 5 samples and 7 samples with SNVs or indels are mutually exclusive (Figure 5.4b), which suggests that in addition to mutations, low gene expression may represent another way to cause Ikaros to lose efficacy.

6 Summary

Cancer is a disease of the genome triggered by somatic mutations. Characterizing the nature and importance of these somatic alterations has been the goal of tumor biologists for several decades. On one hand, the characterization of somatic mutations allows the identification of driver mutations and driver genes, providing new insights into the underlying mechanism of tumorigenesis and possibly revealing new therapeutic targets for cancer treatment. On the other hand, the exploration of somatic alterations makes it possible to investigate generation mechanism of somatic alterations, contributing to the understanding of DNA damage and repair processes that have been operative throughout the development of cancer. This dissertation detected somatic copy number alterations (SCNAs) and chromosomal breaks in human osteosarcoma as well as single nucleotide variants (SNVs), small insertions/deletions (indels), SCNAs, structural variants (SVs) in chicken Marek's disease lymphomas. It also investigated generation mechanisms of somatic mutations, especially SNVs and SCNAs in multiple tumor types.

In the Chapter 1 of this dissertation, we reviewed related literatures in cancer genomics. We first introduced the concept of "cancer is a disease of genome", then the catalog of somatic mutations in cancer, followed by high-throughput genomic technologies (next-generation sequencing and whole-genome genotyping microarrays) used for exploring somatic mutations in cancers. We then focus on summarizing computational tools used for detecting somatic mutations including SNVs, indels, SCNAs, SVs and gene fusions, for mapping, annotating and functional prediction of somatic mutations, and for detecting driver genes and pathways from somatic alterations. Finally, we highlighted recent studies providing new insights into the generation mechanisms of SNVs, indels and SCNAs (and SVs) in cancer genome.

In Chapter 2, we aimed to reproduce a study published in Nature (Schuster-Böckler B. and Lehner B. Nature, 2012, 488(7412):504-507) to offer new insights, if any, into the mutation-rate (especially SNV rate) variance in human cancer cells. Cancer genome sequencing provides an unprecedented opportunity to investigate how mutation rates vary across the genomes of somatic cells. Taking advantage of available genetic and epigenetic features, Schuster-Böckler and Lehner have shown that mutation rates in cancer genomes are strikingly related to chromatin organization. They showed that at the mega base scale, a heterochromatin-associated histone modification marker — H3K9me3 — explains >40% of mutation-rate variance, and all investigated features account for >55% variance. They also showed that the strong association between somatic mutation rates and chromatin organization is independent of tissue and mutation types. Using the same data sets and same procedure, our results are largely consistent with the original study, with the exception being that replication timing is the most prominent predictor for mutation rate in cancer cells. Our results comply with two subsequent studies [241, 242], in which replication timing was found to play an important role in shaping SNV landscape in cancer cells.

In Chapter 3, we investigated the generation mechanisms of SCNAs in cancer. SCNAs play an important role in carcinogenesis. However, the impact of genomic architecture on the global patterns of SCNAs in cancer genomes remains elusive. We conducted multiple linear regression (MLR) analyses of the pooled SCNA data from The Cancer Genome Atlas Pan-Cancer project. Our MLR model explains >30% of the pooled 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. 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. Our MLR model is more than two times more powerful than that in [297] (32% of breakpoint variance explained versus 14%) and maintains its strong performance upon 5-fold cross validation. The inclusion of 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), may explain the superiority of our model compared with that described in [297]. Using a rare event logistic regression model and an extremely randomized tree classifier, we revealed that genomic features are informative for telling apart common SCNA breakpoints breakpoint hotspot and non-hotspots. This suggests that common breakpoint hotspots strongly depend on the local genomic context. Our findings shed light on the molecular mechanisms of SCNA generation in cancer.

In Chapter 4, we performed a genome-wide analyses of SCNAs and chromosomal breaks in osteosarcoma (OS). OS is the most common primary malignant bone tumor in children and adolescents and is characterized by highly complex karyotypes with structural and numerical chromosomal alterations. The identification of driver genes for OS has been hindered by intra- and intertumor heterogeneity and limited sample availability. A comprehensive assessment of SCNAs was performed in 160 OS samples using whole-genome CytoScan High Density arrays, of which 98% of the analyzed samples were of sufficient quality for data analysis. A high degree of aneuploidy and large-scale copy number alterations in OS were 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 (i.e., broken regions) harboring OS-associated tumor suppressor genes including *TP53*, *RB1*, *WWOX*, *DLG2*, and *LSAMP*. We showed that 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. 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, suggesting the independence of breakage susceptibility on local genomic context. We speculated that breakages probably occur at OS specific fragile sites with the potential to form stable secondary structures (e.g., non-B DNA structures) and to consequently stall the replication fork. A complex breakage pattern — chromothripsis — has been suggested as a widespread phenomenon in OS. It was further demonstrated that hyperploidy and particularly chromothripsis were strongly correlated with OS patient clinical outcome. The revealed OS-specific fragility pattern provides novel clues for understanding the biology of OS and may provide a basis for patient prognosis in the future.

In Chapter 5, we explored the somatic mutational landscape of Marek's Disease (MD) in

chickens. MD, which is caused by Marek's Disease Virus (MDV), is a serious chronic disease most obviously manifested by malignant T-cell lymphomas. Annual world-wide losses due to MD were estimated to be roughly 1-2 billion US dollars. Although vaccination against MDV has been successful in stopping the formation of neoplasms in infected chickens, high-density poultry rearing practices and vaccination control have induced MDV evolution and increased MDV virulence as shown by multiple vaccine breaks throughout the second half of the 20th century. To address whether somatic alterations are necessary for MDV-induced transformation, we used multiple approaches (whole genome sequencing, whole transcriptome sequencing and SNP genotyping arrays) to chart the somatic mutational landscape of MD. We identified 54 high-confidence driver genes, some of which function in cell adhesion, cell signaling, cellular proliferation, cell differentiation and immune response. Notably, we found that disruptive mutations together with low gene expression of *IKZF1* occurred in 12 of 26 (46%) MD tumors. IKZF1 has been found to have crucial function in hematopoietic cell differentiation have been identified as an important player in the development of acute lymphoblastic leukemia with Philadelphia chromosome [422]. Our results will contribute to the understanding how somatic mutations drive transformation and lymphomagenesis in MD.

Appendices

A

Supplementary Tables

Table A.1: Alternative MLR model replacing A-phased repeat with GC content

Predictor	SCE	VIF	P-value	RC,%	Five-fold RC,%
Distance to centromere	-0.244	1.261	1.47×10^{-38}	14.71	19.93
Conserved element count	0.117	3.418	1.18×10^{-04}	1.25	1.19
CpG island coverage	0.074	1.135	2.39×10^{-05}	1.51	1.29
Direct repeat coverage	0.436	5.332	9.84×10^{-30}	11.09	13.32
L1 coverage	0.134	3.659	2.07×10^{-05}	1.53	1.79
Low-complexity repeat coverage	0.140	3.084	1.38×10^{-06}	1.97	2.71
Mirror repeat count	-0.309	4.324	2.93×10^{-19}	6.90	8.08
SINE count	0.246	9.761	1.75×10^{-06}	1.94	1.95
Distance to telomere	-0.418	1.864	1.90×10^{-72}	29.16	32.51
Simple repeat coverage	-0.085	2.383	8.22×10^{-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 A.2: Alternative MLR model replacing A-phased repeat with recombination motif

Predictor	SCE	VIF	P-value	RC,%	Five-fold RC,%
Distance to centromere	-0.243	1.260	2.46×10^{-38}	14.61	19.80
Conserved element count	0.116	3.393	1.38×10^{-04}	1.23	1.16
CpG island coverage	0.073	1.132	2.77×10^{-05}	1.49	1.15
Direct repeat coverage	0.429	5.244	2.45×10^{-29}	10.93	13.26
Inverted repeat coverage	0.096	3.330	1.46×10^{-03}	0.86	0.44
L1 coverage	0.139	3.664	1.05×10^{-05}	1.64	1.88
Low-complexity repeat coverage	0.144	3.082	6.25×10^{-07}	2.10	2.85
Mirror repeat count	-0.300	4.294	2.53×10^{-18}	6.52	7.79
SINE count	0.252	10.209	1.66×10^{-06}	1.94	2.06
Distance to telomere	-0.416	1.869	6.81×10^{-72}	28.91	31.88
Z-DNA coverage	-0.096	3.334	1.46×10^{-03}	0.86	-0.22
Simple repeat coverage	-0.086	2.364	7.03×10^{-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 A.3: Alternative MLR model replacing A-phased repeat with G4

Predictor	SCE	VIF	P-value	RC,%	Five-fold RC,%
Distance to centromere	-0.243	1.260	3.28×10^{-38}	14.60	19.81
Conserved element count	0.108	3.510	4.85×10^{-04}	1.03	0.88
CpG island coverage	0.072	1.133	4.22×10^{-05}	1.42	1.19
Direct repeat coverage	0.425	5.336	2.47×10^{-28}	10.56	12.55
Inverted repeat coverage	0.100	3.319	8.91×10^{-04}	0.94	0.57
L1 coverage	0.133	3.753	3.07×10^{-05}	1.47	1.58
Low-complexity repeat coverage	0.139	3.199	2.51×10^{-06}	1.88	2.48
Mirror repeat count	-0.301	4.332	2.56×10^{-18}	6.54	7.73
SINE count	0.205	8.261	1.50×10^{-05}	1.59	1.66
Distance to telomere	-0.419	1.869	1.12×10^{-72}	29.35	32.54
Z-DNA coverage	-0.125	3.837	1.06×10^{-04}	1.27	0.53
Simple repeat coverage	-0.094	2.342	2.07×10^{-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 A.4: Alternative MLR model replacing H3K9me3 with replication timing

Predictor	SCE	VIF	P-value	RC,%	Five-fold RC,%
Distance to centromere	-0.244	1.258	1.01×10^{-38}	14.77	19.74
Conserved element count	0.115	3.387	1.41×10^{-04}	1.23	1.16
CpG island coverage	0.071	1.133	5.01×10^{-05}	1.39	1.03
Direct repeat coverage	0.417	5.420	4.77×10^{-27}	10.01	11.51
Inverted repeat coverage	0.103	3.322	5.75×10^{-04}	1.00	0.70
L1 coverage	0.140	3.667	9.81×10^{-06}	1.65	1.86
Low-complexity repeat coverage	0.145	3.073	5.12×10^{-07}	2.14	2.87
Mirror repeat count	-0.298	4.302	3.96×10^{-18}	6.45	7.40
SINE count	0.198	7.809	1.65×10^{-05}	1.57	1.49
Distance to telomere	-0.422	1.879	3.42×10^{-73}	29.49	32.27
Z-DNA coverage	-0.118	2.837	2.25×10^{-05}	1.52	0.16
Simple repeat coverage	-0.088	2.335	4.43×10^{-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.

Table A.5: The MLR model for SCNA breakpoints after excluding chromosome-level SCNAs

Predictor	SCE	VIF	P-value	RC,%	Five-fold RC,%
Distance to centromere	-0.339	1.265	1.24×10^{-69}	29.30	41.94
Conserved element count	0.097	3.382	1.49×10^{-03}	0.89	0.67
CpG island coverage	0.086	1.133	1.01×10^{-06}	2.13	0.01
Direct repeat coverage	0.370	5.433	2.38×10^{-21}	8.11	10.09
Inverted repeat coverage	0.114	3.330	1.60×10^{-04}	1.26	1.39
Low-complexity repeat coverage	0.092	3.069	1.52×10^{-03}	0.89	0.52
Mirror repeat count	-0.229	4.284	3.00×10^{-11}	3.94	3.53
SINE count	0.222	8.762	6.40×10^{-06}	1.81	1.73
Distance to telomere	-0.391	1.884	1.38×10^{-62}	26.08	30.43
Simple repeat coverage	-0.115	2.434	8.58×10^{-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.

Table A.6: 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

Table A.7: 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
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	

Continued on next page

Table A.7 – *Continued from previous page*

Chr.	Region	Extended Region	Type	Genes
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 A.8: 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

Continued on next page

Table A.8 – *Continued from previous page*

Gene Symbol	Chromosome	Start	End	Length
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
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

Continued on next page

Table A.8 – *Continued from previous page*

Gene Symbol	Chromosome	Start	End	Length
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.

B

Supplementary Figures

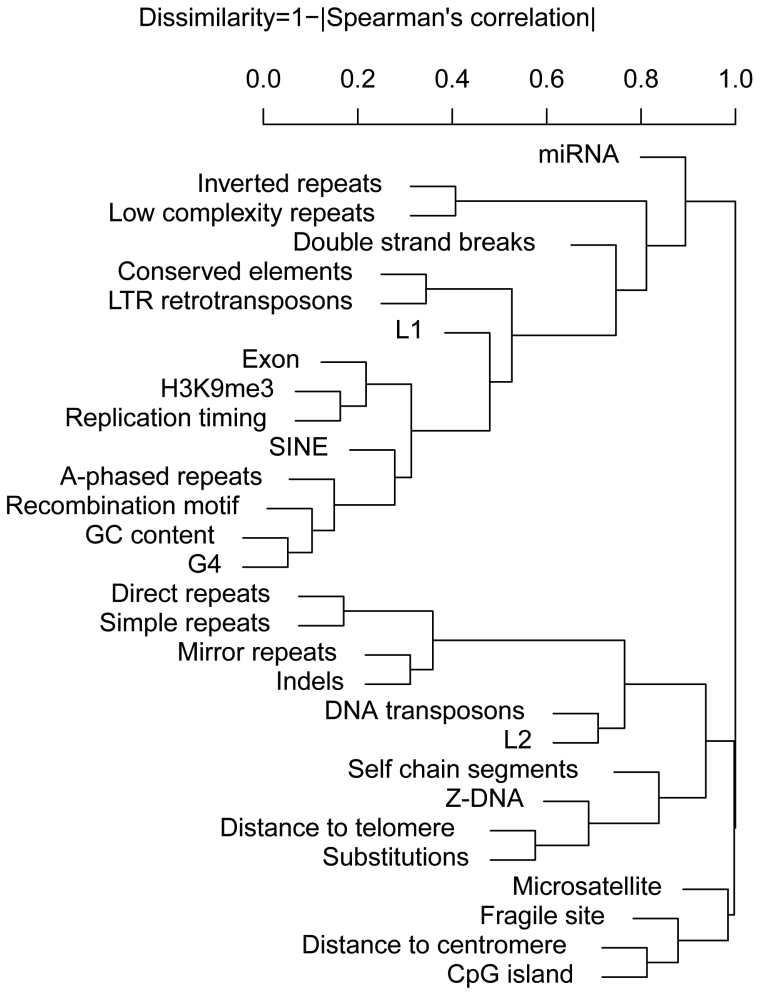


Figure B.1: Hierarchical clustering of predictors based on their Spearman's correlation coefficients.

References

1. National Cancer Institute. *What Is Cancer?* 2016. <http://www.cancer.gov/about-cancer/understanding/what-is-cancer>.
2. Chen, W. *et al.* Cancer statistics in China, 2015. *CA Cancer J Clin* **66,** 115–132 (2016).
3. Boveri, T. Concerning the origin of malignant tumours by Theodor Boveri. Translated and annotated by Henry Harris. *J Cell Sci* **121 Suppl 1,** 1–84 (2008).
4. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458,** 719–724 (2009).
5. Garraway, L. A. & Lander, E. S. Lessons from the cancer genome. *Cell* **153,** 17–37 (2013).
6. Loeb, L. A. & Harris, C. C. Advances in chemical carcinogenesis: a historical review and prospective. *Cancer Res* **68,** 6863–6872 (2008).
7. Avery, O. T., Macleod, C. M. & McCarty, M. Studies on the chemical nature of the substance inducing transformation of pneumococcal types : induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. *J Exp Med* **79,** 137–158 (1944).
8. Watson, J. D. & Crick, F. H. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature* **171,** 737–738 (1953).
9. Nowell, P. C. & D., H. A minute chromosome in human chronic granulocytic leukemia. *Science,* 1497–1501 (1960).
10. Rowley, J. D. A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. *Nature* **243,** 290–293 (1973).

11. Tough, I. M. *et al.* Cytogenetic studies in chronic myeloid leukaemia and acute leukaemia associated with monogolism. *Lancet* **1,** 411–417 (1961).
12. Nowell, P. C. Discovery of the Philadelphia chromosome: a personal perspective. *J Clin Invest* **117,** 2033–2035 (2007).
13. Groffen, J. *et al.* Philadelphia chromosomal breakpoints are clustered within a limited region, bcr, on chromosome-22. *Cell* **36,** 93–99 (1984).
14. Krontiris, T. G. & Cooper, G. M. Transforming activity of human tumor DNAs. *Proc Natl Acad Sci U S A* **78,** 1181–1184 (1981).
15. Shih, C., Padhy, L. C., Murray, M. & Weinberg, R. A. Transforming genes of carcinomas and neuroblastomas introduced into mouse fibroblasts. *Nature* **290,** 261–264 (1981).
16. Tabin, C. J. *et al.* Mechanism of activation of a human oncogene. *Nature* **300,** 143–149 (1982).
17. Reddy, E. P., Reynolds, R. K., Santos, E. & Barbacid, M. A point mutation is responsible for the acquisition of transforming properties by the T24 human bladder-carcinoma oncogene. *Nature* **300,** 149–152 (1982).
18. Goldfarb, M., Shimizu, K., Perucho, M. & Wigler, M. Isolation and preliminary characterization of a human transforming gene from T24 bladder carcinoma cells. *Nature* **296,** 404–409 (1982).
19. Futreal, P. A. *et al.* A census of human cancer genes. *Nat Rev Cancer* **4,** 177–183 (2004).
20. Croce, C. M. Oncogenes and cancer. *N Engl J Med* **358,** 502–511 (2008).
21. Vogelstein, B. & Kinzler, K. W. Cancer genes and the pathways they control. *Nat Med* **10,** 789–799 (2004).
22. Dallafavera, R. *et al.* Human c-myc onc gene is located on the region of chromosome 8 that is translocated in Burkitt lymphoma cells. *Proc Natl Acad Sci U S A* **79,** 7824–7827 (1982).
23. Press, M. F. *et al.* HER-2/neu gene amplification characterized by fluorescence in situ hybridization: poor prognosis in node-negative breast carcinomas. *J Clin Oncol* **15,** 2894–2904 (1997).
24. Davies, H. *et al.* Mutations of the BRAF gene in human cancer. *Nature* **417,** 949–954 (2002).
25. Chial, H. Tumor suppressor (TS) genes and the two-hit hypothesis. *Nature Education* **1,** 177 (2008).

26. Berger, A. H. & Pandolfi, P. P. Haplo-insufficiency: a driving force in cancer. *J Pathol* **223,** 137–146 (2011).
27. Knudson A. G., J. Mutation and cancer: statistical study of retinoblastoma. *Proc Natl Acad Sci U S A* **68,** 820–823 (1971).
28. Hanahan, D. & Weinberg, R. The hallmarks of cancer. *Cell* **100,** 57–70 (2000).
29. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144,** 646–674 (2011).
30. Negrini, S., Gorgoulis, V. G. & Halazonetis, T. D. Genomic instability–an evolving hallmark of cancer. *Nat Rev Mol Cell Biol* **11,** 220–228 (2010).
31. De Lange, T. Telomere-related genome instability in cancer. *Cold Spring Harb Symp Quant Biol* **70,** 197–204 (2005).
32. Magdalou, I., Lopez, B. S., Pasero, P. & Lambert, S. A. The causes of replication stress and their consequences on genome stability and cell fate. *Semin Cell Dev Biol* **30,** 154–164 (2014).
33. Nagy, R., Sweet, K. & Eng, C. Highly penetrant hereditary cancer syndromes. *Oncogene* **23,** 6445–6470 (2004).
34. Malkin, D. *et al.* Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. *Science* **250,** 1233–1238 (1990).
35. Miki, Y. *et al.* A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science* **266,** 66–71 (1994).
36. Wooster, R. *et al.* Identification of the breast cancer susceptibility gene BRCA2. *Nature* **378,** 789–792 (1995).
37. Stephens, P. *et al.* Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144,** 27–40 (2011).
38. Liu, P. *et al.* Chromosome catastrophes involve replication mechanisms generating complex genomic rearrangements. *Cell* **146,** 889–903 (2011).
39. Baca, S. *et al.* Punctuated evolution of prostate cancer genomes. *Cell* **153,** 666–677 (2013).
40. McLaughlin-Drubin, M. E. & Munger, K. Viruses associated with human cancer. *Biochim Biophys Acta* **1782,** 127–150 (2008).
41. Stratton, M. R. Exploring the genomes of cancer cells: progress and promise. *Science* **331,** 1553–1558 (2011).
42. The Cancer Genome Atlas Research Network. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* **45,** 1113–1120 (2013).

43. International Cancer Genome Consortium. International network of cancer genome projects. *Nature* **464,** 993–998 (2010).
44. Gao, J., Ciriello, G., Sander, C. & Schultz, N. Collection, integration and analysis of cancer genomic profiles: from data to insight. *Curr Opin Genet Dev* **24C,** 92–98 (2014).
45. LaFramboise, T. Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic Acids Res* **37,** 4181–4193 (2009).
46. Beroukhim, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463,** 899–905 (2010).
47. Zack, T. I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nat Genet* **45,** 1134–1140 (2013).
48. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74,** 5463–5467 (1977).
49. Maxam, A. M. & Gilbert, W. A new method for sequencing DNA. *Proc Natl Acad Sci U S A* **74,** 560–564 (1977).
50. Collins, F. S., Morgan, M. & Patrinos, A. The Human Genome Project: lessons from large-scale biology. *Science* **300,** 286–290 (2003).
51. Wikipedia. *Human Genome Project* <https://en.wikipedia.org/wiki/Human_Genome_Project>.
52. Liu, L. *et al.* Comparison of next-generation sequencing systems. *J Biomed Biotechnol* **2012,** 251364 (2012).
53. Van Dijk, E. L., Auger, H., Jaszczyszyn, Y. & Thermes, C. Ten years of next-generation sequencing technology. *Trends Genet* **30,** 418–426 (2014).
54. Illumina. *An Introduction to Next-Generation Sequencing Technology* 2016.
55. Benjamini, Y. & Speed, T. P. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res* **40,** e72 (2012).
56. Schwartz, S., Oren, R. & Ast, G. Detection and removal of biases in the analysis of next-generation sequencing reads. *PLoS One* **6,** e16685 (2011).
57. Ding, L., Wendl, M. C., McMichael, J. F. & Raphael, B. J. Expanding the computational toolbox for mining cancer genomes. *Nat Rev Genet* **15,** 556–570 (2014).
58. Ozsolak, F. & Milos, P. M. RNA sequencing: advances, challenges and opportunities. *Nat Rev Genet* **12,** 87–98 (2011).
59. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).

60. Li, H. & Homer, N. A survey of sequence alignment algorithms for next-generation sequencing. *Brief Bioinform* **11,** 473–483 (2010).
61. Ding, J. *et al.* Feature based classifiers for somatic mutation detection in tumour-normal paired sequencing data. *Bioinformatics* **28,** 167–175 (2011).
62. Pleasance, E. D. *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463,** 191–196 (2010).
63. Raphael, B. J., Dobson, J. R., Oesper, L. & Vandin, F. Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. *Genome Med* **6,** 5 (2014).
64. Larson, D. *et al.* SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics* **28,** 311–317 (2012).
65. Roth, A. *et al.* JointSNVMix: a probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data. *Bioinformatics* **28,** 907–913 (2012).
66. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* **31,** 213–219 (2013).
67. Saunders, C. *et al.* Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* **28,** 1811–1817 (2012).
68. Wilm, A. *et al.* LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res* **40,** 11189–11201 (2012).
69. Koboldt, D. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* **22,** 568–576 (2012).
70. Lai, Z. *et al.* VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res* **44,** e108 (2016).
71. Wang, Q. *et al.* Detecting somatic point mutations in cancer genome sequencing data: a comparison of mutation callers. *Genome Med* **5,** 91 (2013).
72. Krøigård, A., Thomassen, M., Lænkholm, A.-V., Kruse, T. & Larsen, M. Evaluation of nine somatic variant callers for detection of somatic mutations in exome and targeted deep sequencing data. *PLoS One* **11,** e0151664 (2016).
73. Fang, L. T. *et al.* An ensemble approach to accurately detect somatic mutations using SomaticSeq. *Genome Biol* **16,** 197 (2015).
74. Fan, Y. *et al.* MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. *Genome Biol* **17,** 178 (2016).

75. Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25,** 2865–2871 (2009).

76. Albers, C. *et al.* Dindel: accurate indel calls from short-read data. *Genome Res* **21,** 961–973 (2011).

77. Banerji, S. *et al.* Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature* **486,** 405–409 (2012).

78. Popova, T. *et al.* Genome Alteration Print (GAP): a tool to visualize and mine complex cancer genomic profiles obtained by SNP arrays. *Genome Biol* **10,** R128 (2009).

79. Sun, W. *et al.* Integrated study of copy number states and genotype calls using high-density SNP arrays. *Nucleic Acids Res* **37,** 5365–5377 (2009).

80. Greenman, C. *et al.* PICNIC: an algorithm to predict absolute allelic copy number variation with microarray cancer data. *Biostatistics* **11,** 164–175 (2010).

81. Van Loo, P. *et al.* Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci U S A* **107,** 16910–16915 (2010).

82. Yau, C. *et al.* A statistical approach for detecting genomic aberrations in heterogeneous tumor samples from single nucleotide polymorphism genotyping data. *Genome Biol* **11,** R92 (2010).

83. Li, A. *et al.* 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,** 4928–4941 (2011).

84. Carter, S. *et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nat Biotechnol* **30,** 413–421 (2012).

85. Chiang, D. Y. *et al.* High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat Methods* **6,** 99–103 (2009).

86. Ivakhno, S. *et al.* CNAseg–a novel framework for identification of copy number changes in cancer from second-generation sequencing data. *Bioinformatics* **26,** 3051–3058 (2010).

87. Miller, C., Hampton, O., Coarfa, C. & Milosavljevic, A. ReadDepth: a parallel R package for detecting copy number alterations from short sequencing reads. *PLoS One* **6,** e16327 (2011).

88. Xi, R. *et al.* Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. *Proc Natl Acad Sci U S A* **108,** E1128–E1136 (2011).

89. Boeva, V. *et al.* Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics* **28,** 423–425 (2012).
90. Sathirapongsasuti, J. F. *et al.* Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. *Bioinformatics* **27,** 2648–2654 (2011).
91. Gusnanto, A., Wood, H., Pawitan, Y., Rabbitts, P. & Berri, S. Correcting for cancer genome size and tumour cell content enables better estimation of copy number alterations from next-generation sequence data. *Bioinformatics* **28,** 40–47 (2012).
92. Mayrhofer, M., DiLorenzo, S. & Isaksson, A. Patchwork: allele-specific copy number analysis of whole-genome sequenced tumor tissue. *Genome Biol* **14,** R24 (2013).
93. Ha, G. *et al.* Integrative analysis of genome-wide loss of heterozygosity and monoallelic expression at nucleotide resolution reveals disrupted pathways in triple-negative breast cancer. *Genome Res* **22,** 1995–2007 (2012).
94. Yau, C. OncoSNP-SEQ: a statistical approach for the identification of somatic copy number alterations from next-generation sequencing of cancer genomes. *Bioinformatics* **29,** 2482–2484 (2013).
95. Yu, Z., Liu, Y., Shen, Y., Wang, M. & Li, A. CLImAT: accurate detection of copy number alteration and loss of heterozygosity in impure and aneuploid tumor samples using whole-genome sequencing data. *Bioinformatics* **30,** 2576–2583 (2014).
96. Korbel, J. O. *et al.* PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biol* **10,** R23 (2009).
97. Chen, K. *et al.* BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods* **6,** 677–681 (2009).
98. Hormozdiari, F., Alkan, C., Eichler, E. E. & Sahinalp, S. C. Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Res* **19,** 1270–1278 (2009).
99. Zeitouni, B. *et al.* SVDetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data. *Bioinformatics* **26,** 1895–1896 (2010).
100. Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28,** i333–i339 (2012).
101. Jiang, Y., Wang, Y. & Brudno, M. PRISM: pair-read informed split-read mapping for base-pair level detection of insertion, deletion and structural variants. *Bioinformatics* **28,** 2576–2583 (2012).
102. Quinlan, A. R. *et al.* Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. *Genome Res* **20,** 623–635 (2010).

103. Wang, J. *et al.* CREST maps somatic structural variation in cancer genomes with base-pair resolution. *Nat Methods* **8,** 652–654 (2011).

104. Iqbal, Z., Caccamo, M., Turner, I., Flicek, P. & McVean, G. De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nat Genet* **44,** 226–232 (2012).

105. Yang, L. *et al.* Diverse mechanisms of somatic structural variations in human cancer genomes. *Cell* **153,** 919–929 (2013).

106. Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol* **15,** R84 (2014).

107. Wang, K. *et al.* MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res* **38,** e178 (2010).

108. Sboner, A. *et al.* FusionSeq: a modular framework for finding gene fusions by analyzing paired-end RNA-sequencing data. *Genome Biol* **11,** R104 (2010).

109. Kim, D. & Salzberg, S. L. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biol* **12,** R72 (2011).

110. Asmann, Y. *et al.* A novel bioinformatics pipeline for identification and characterization of fusion transcripts in breast cancer and normal cell lines. *Nucleic Acids Res* **39,** e100 (2011).

111. Kinsella, M., Harismendy, O., Nakano, M., Frazer, K. & Bafna, V. Sensitive gene fusion detection using ambiguously mapping RNA-Seq read pairs. *Bioinformatics* **27,** 1068–1075 (2011).

112. Ge, H. *et al.* FusionMap: detecting fusion genes from next-generation sequencing data at base-pair resolution. *Bioinformatics* **27,** 1922–1928 (2011).

113. Li, Y., Chien, J., Smith, D. & Ma, J. FusionHunter: identifying fusion transcripts in cancer using paired-end RNA-seq. *Bioinformatics* **27,** 1708–1710. (2011).

114. McPherson, A. *et al.* deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. *PLoS Comput Biol* **7,** e1001138 (2011).

115. McPherson, A. *et al.* Comrad: detection of expressed rearrangements by integrated analysis of RNA-Seq and low coverage genome sequence data. *Bioinformatics* **27,** 1481–1488 (2011).

116. Iyer, M., Chinnaiyan, A. & Maher, C. ChimeraScan: a tool for identifying chimeric transcription in sequencing data. *Bioinformatics* **27,** 2903–2904 (2011).

117. McPherson, A. *et al.* nFuse: discovery of complex genomic rearrangements in cancer using high-throughput sequencing. *Genome Res* **22,** 2250–2261 (2012).

118. Jia, W. *et al.* SOAPfuse: an algorithm for identifying fusion transcripts from paired-end RNA-Seq data. *Genome Biol* **14,** R12 (2013).

119. Zhang, J. *et al.* INTEGRATE: gene fusion discovery using whole genome and transcriptome data. *Genome Res* **26,** 108–118 (2016).

120. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* **18,** 1851–1858 (2008).

121. Narzisi, G. *et al.* Accurate de novo and transmitted indel detection in exome-capture data using microassembly. *Nat Methods* **11,** 1033–1036 (2014).

122. Yang, R., Nelson, A. C., Henzler, C., Thyagarajan, B. & Silverstein, K. A. ScanIndel: a hybrid framework for indel detection via gapped alignment, split reads and de novo assembly. *Genome Med* **7,** 127 (2015).

123. Kallioniemi, A. *et al.* Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science* **258,** 818–821 (1992).

124. Pinkel, D. *et al.* High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat Genet* **20,** 207–211 (1998).

125. Colella, S. *et al.* QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res* **35,** 2013–2025 (2007).

126. Wang, K. *et al.* PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res* **17,** 1665–1674 (2007).

127. Winchester, L., Yau, C. & Ragoussis, J. Comparing CNV detection methods for SNP arrays. *Brief Funct Genomic Proteomic* **8,** 353–366 (2009).

128. Gordon, D. J., Resio, B. & Pellman, D. Causes and consequences of aneuploidy in cancer. *Nat Rev Genet* **13,** 189–203 (2012).

129. Mosen-Ansorena, D., Aransay, A. M. & Rodriguez-Ezpeleta, N. Comparison of methods to detect copy number alterations in cancer using simulated and real genotyping data. *BMC Bioinformatics* **13,** 192 (2012).

130. Burrell, R. A., McGranahan, N., Bartek, J. & Swanton, C. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature* **501,** 338–345 (2013).

131. Alkan, C., Coe, B. P. & Eichler, E. E. Genome structural variation discovery and genotyping. *Nat Rev Genet* **12,** 363–376 (2011).

132. Zhao, M., Wang, Q., Wang, Q., Jia, P. & Zhao, Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics* **14,** S1 (2013).

133. Xie, C. & Tammi, M. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics* **10,** 80 (2009).

134. Abyzov, A., Urban, A., Snyder, M. & Gerstein, M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res* **21,** 974–984 (2011).

135. Liu, B. A. *et al.* Computational methods for detecting copy number variations in cancer genome using next generation sequencing: principles and challenges. *Oncotarget* **4,** 1868–1881 (2013).

136. Duan, J., Zhang, J.-G., Deng, H.-W. & Wang, Y.-P. Comparative studies of copy number variation detection methods for next-generation sequencing technologies. *PLoS One* **8,** e59128 (2013).

137. Alkodsi, A., Louhimo, R. & Hautaniemi, S. Comparative analysis of methods for identifying somatic copy number alterations from deep sequencing data. *Brief Bioinform* **16,** 242–254 (2015).

138. Nam, J. Y. *et al.* Evaluation of somatic copy number estimation tools for whole-exome sequencing data. *Brief Bioinform* **17,** 185–192 (2016).

139. Lin, K., Smit, S., Bonnema, G., Sanchez-Perez, G. & de Ridder, D. Making the difference: integrating structural variation detection tools. *Brief Bioinform* **16,** 852–864 (2015).

140. Mertens, F., Johansson, B., Fioretos, T. & Mitelman, F. The emerging complexity of gene fusions in cancer. *Nat Rev Cancer* **15,** 371–381 (2015).

141. Kumar, S., Razzaq, S. K., Vo, A. D., Gautam, M. & Li, H. Identifying fusion transcripts using next generation sequencing. *Wiley Interdiscip Rev RNA* **7,** 811–823 (2016).

142. Wang, Q., Xia, J., Jia, P., Pao, W. & Zhao, Z. Application of next generation sequencing to human gene fusion detection: computational tools, features and perspectives. *Brief Bioinform* **14,** 506–519 (2013).

143. Carrara, M. *et al.* State-of-the-art fusion-finder algorithms sensitivity and specificity. *Biomed Res Int* **2013,** 340620 (2013).

144. Kumar, S., Vo, A., Qin, F. & Li, H. Comparative assessment of methods for the fusion transcripts detection from RNA-Seq data. *Sci Rep* **6,** 21597 (2016).

145. Liu, S. *et al.* Comprehensive evaluation of fusion transcript detection algorithms and a meta-caller to combine top performing methods in paired-end RNA-seq data. *Nucleic Acids Res* **44,** e47 (2015).

146. International Cancer Genome Consortium. Computational approaches to identify functional genetic variants in cancer genomes. *Nat Methods* **10,** 723–729 (2013).

147. Tamborero, D., Gonzalez-Perez, A. & Lopez-Bigas, N. Identification of oncogenic driver mutations. *Jpn J Exp Med.* **32** (2014).

148. Zhang, J. & Zhang, S. The discovery of mutated driver pathways in cancer: models and algorithms. *arXiv* **1604.01298** (2016).

149. The ENCODE Project Consortium. Identification and analysis of functional elements in 1the human genome by the ENCODE pilot project. *Nature* **447,** 799–816 (2007).

150. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489,** 75–82 (2012).

151. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).

152. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29,** 308–311 (2001).

153. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).

154. Forbes, S. A. *et al.* COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res* **43,** D805–D811 (2015).

155. Stenson, P. D. *et al.* The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum Genet* **133,** 1–9 (2014).

156. MacDonald, J. R., Ziman, R., Yuen, R. K., Feuk, L. & Scherer, S. W. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res* **42,** D986–D992 (2014).

157. Habegger, L. *et al.* VAT: a computational framework to functionally annotate variants in personal genomes within a cloud-computing environment. *Bioinformatics* **28,** 2267–2269 (2012).

158. Ramos, A. H. *et al.* Oncotator: cancer variant annotation tool. *Hum Mutat* **36,** E2423–E2429 (2015).

159. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38,** e164 (2010).

160. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* **6,** 80–92 (2012).

161. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* **17,** 122 (2016).

162. Makarov, V. *et al.* AnnTools: a comprehensive and versatile annotation toolkit for genomic variants. *Bioinformatics* **28,** 724–725 (2012).

163. Medina, I. *et al.* VARIANT: Command Line, Web service and Web interface for fast and accurate functional characterization of variants found by Next-Generation Sequencing. *Nucleic Acids Res* **40,** W54–W58 (2012).

164. Douville, C. *et al.* CRAVAT: cancer-related analysis of variants toolkit. *Bioinformatics* **29,** 647–648 (2013).

165. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* **4,** 1073–1082 (2009).

166. Ng, P. C. & Henikoff, S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res* **31,** 3812–3814 (2003).

167. Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res* **39,** e118 (2011).

168. Choi, Y., Sims, G. E., Murphy, S., Miller, J. R. & Chan, A. P. Predicting the functional effect of amino acid substitutions and indels. *PLoS One* **7,** e46688 (2012).

169. Stone, E. A. & Sidow, A. Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. *Genome Res* **15,** 978–986 (2005).

170. Ryan, M., Diekhans, M., Lien, S., Liu, Y. & Karchin, R. LS-SNP/PDB: annotated non-synonymous SNPs mapped to Protein Data Bank structures. *Bioinformatics* **25,** 1431–1432 (2009).

171. Gonzalez-Perez, A., Deu-Pons, J. & Lopez-Bigas, N. Improving the prediction of the functional impact of cancer mutations by baseline tolerance transformation. *Genome Med* **4,** 89 (2012).

172. Gonzalez-Perez, A. & Lopez-Bigas, N. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *Am J Hum Genet* **88,** 440–449 (2011).

173. Kaminker, J. S., Zhang, Y., Watanabe, C. & Zhang, Z. CanPredict: a computational tool for predicting cancer-associated missense mutations. *Nucleic Acids Res* **35,** W595–W598 (2007).

174. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat Methods* **7,** 248–249 (2010).

175. Carter, H. *et al.* Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer Res* **69,** 6660–6667 (2009).

176. Carter, H., Douville, C., Stenson, P. D., Cooper, D. N. & Karchin, R. Identifying Mendelian disease genes with the variant effect scoring tool. *BMC Genomics* **14 Suppl 3,** S3 (2013).

177. Douville, C. *et al.* Assessing the pathogenicity of insertion and deletion variants with the Variant Effect Scoring Tool (VEST-Indel). *Hum Mutat* **37,** 28–35 (2016).

178. Hu, J. & Ng, P. C. SIFT Indel: predictions for the functional effects of amino acid insertions/deletions in proteins. *PLoS One* **8,** e77940 (2013).

179. Shihab, H. A. *et al.* Ranking non-synonymous single nucleotide polymorphisms based on disease concepts. *Hum Genomics* **8,** 11 (2014).

180. Schwarz, J. M., Rodelsperger, C., Schuelke, M. & Seelow, D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods* **7,** 575–576 (2010).

181. Schwarz, J. M., Cooper, D. N., Schuelke, M. & Seelow, D. MutationTaster2: mutation prediction for the deep-sequencing age. *Nat Methods* **11,** 361–362 (2014).

182. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46,** 310–315 (2014).

183. Dees, N. D. *et al.* MuSiC: identifying mutational significance in cancer genomes. *Genome Res* **22,** 1589–1598 (2012).

184. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499,** 214–218 (2013).

185. Hodis, E. *et al.* A landscape of driver mutations in melanoma. *Cell* **150,** 251–263 (2012).

186. Youn, A. & Simon, R. Identifying cancer driver genes in tumor genome sequencing studies. *Bioinformatics* **27,** 175–181 (2011).

187. Gonzalez-Perez, A. & Lopez-Bigas, N. Functional impact bias reveals cancer drivers. *Nucleic Acids Res* **40,** e169 (2012).

188. Tamborero, D., Gonzalez-Perez, A. & Lopez-Bigas, N. OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics* **29,** 2238–2244 (2013).

189. Reimand, J. & Bader, G. D. Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. *Mol Syst Biol* **9,** 637 (2013).

190. Mularoni, L., Sabarinathan, R., Deu-Pons, J., Gonzalez-Perez, A. & Lopez-Bigas, N. OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. *Genome Biol* **17,** 128 (2016).

191. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102,** 15545–15550 (2005).

192. Lin, J. *et al.* A multidimensional analysis of genes mutated in breast and colorectal cancers. *Genome Res* **17,** 1304–1318 (2007).

193. Wendl, M. C. *et al.* PathScan: a tool for discerning mutational significance in groups of putative cancer genes. *Bioinformatics* **27,** 1595–1602 (2011).

194. Vandin, F., Upfal, E. & Raphael, B. J. Algorithms for detecting significantly mutated pathways in cancer. *J Comput Biol* **18,** 507–522 (2011).

195. Leiserson, M. D. M. *et al.* Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat Genet* **47,** 106–114 (2015).

196. Cerami, E., Demir, E., Schultz, N., Taylor, B. S. & Sander, C. Automated network analysis identifies core pathways in glioblastoma. *PLoS One* **5,** e8918 (2010).

197. Yeang, C. H., McCormick, F. & Levine, A. Combinatorial patterns of somatic gene mutations in cancer. *FASEB J* **22,** 2605–2622 (2008).

198. Ciriello, G., Cerami, E., Sander, C. & Schultz, N. Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res* **22,** 398–406 (2012).

199. Vandin, F., Upfal, E. & Raphael, B. J. De novo discovery of mutated driver pathways in cancer. *Genome Res* **22,** 375–385 (2012).

200. Leiserson, M. D., Blokh, D., Sharan, R. & Raphael, B. J. Simultaneous identification of multiple driver pathways in cancer. *PLoS Comput Biol* **9,** e1003054 (2013).

201. Zhao, J., Zhang, S., Wu, L. Y. & Zhang, X. S. Efficient methods for identifying mutated driver pathways in cancer. *Bioinformatics* **28,** 2940–2947 (2012).

202. Miller, C. A., Settle, S. H., Sulman, E. P., Aldape, K. D. & Milosavljevic, A. Discovering functional modules by identifying recurrent and mutually exclusive mutational patterns in tumors. *BMC Med Genomics* **4,** 34 (2011).

203. Chin, L. & Gray, J. W. Translating insights from the cancer genome into clinical practice. *Nature* **452,** 553–563 (2008).

204. Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F. & Hamosh, A. OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. *Nucleic Acids Res* **43,** D789–D798 (2015).

205. Chen, J., Sun, M. & Shen, B. Deciphering oncogenic drivers: from single genes to integrated pathways. *Brief Bioinform* **16,** 413–428 (2015).

206. Martelotto, L. G. *et al.* Benchmarking mutation effect prediction algorithms using functionally validated cancer-related missense mutations. *Genome Biol* **15,** 484 (2014).

207. Gartner, J. *et al.* Whole-genome sequencing identifies a recurrent functional synonymous mutation in melanoma. *Proc Natl Acad Sci U S A* **110,** 13481–13486 (2013).

208. Supek, F., Miñana, B., Valcárcel, J., Gabaldón, T. & Lehner, B. Synonymous mutations frequently act as driver mutations in human cancers. *Cell* **156,** 1324–1335 (2014).

209. Huang, F. W. *et al.* Highly recurrent TERT promoter mutations in human melanoma. *Science* **339,** 957–959 (2013).

210. Horn, S. *et al.* TERT promoter mutations in familial and sporadic melanoma. *Science* **339,** 959–961 (2013).

211. Weinhold, N., Jacobsen, A., Schultz, N., Sander, C. & Lee, W. Genome-wide analysis of noncoding regulatory mutations in cancer. *Nat Genet* **46,** 1160–1165 (2014).

212. Wang, L., Nie, J. & Kocher, J.-P. PVAAS: identify variants associated with aberrant splicing from RNA-seq. *Bioinformatics* **31,** 1668–1670 (2015).

213. Viner, C, Dorman, S., Shirley, B. & Rogan, P. Validation of predicted mRNA splicing mutations using high-throughput transcriptome data. *F1000Research* **3** (2014).

214. Piraino, S. W. & Furney, S. J. Beyond the exome: the role of non-coding somatic mutations in cancer. *Ann Oncol* **27,** 240–248 (2016).

215. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339,** 1546–1558 (2013).

216. Stamatoyannopoulos, J. A. *et al.* Human mutation rate associated with DNA replication timing. *Nat Genet* **41,** 393–395 (2009).

217. Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol* **12,** R41 (2011).

218. Beroukhim, R. *et al.* Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc Natl Acad Sci U S A* **104,** 20007–20012 (2007).

219. Taylor, B. S. *et al.* Functional copy-number alterations in cancer. *PLoS One* **3,** e3179 (2008).

220. Tamborero, D. *et al.* Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Sci Rep* **3,** 2650 (2013).

221. Ding, L. *et al.* Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* **455,** 1069–1075 (2008).

222. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455,** 1061–1068 (2008).

223. Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M. & Hirakawa, M. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res* **38,** D355–D360 (2010).

224. The Gene Ontology Consortium. Gene Ontology Consortium: going forward. *Nucleic Acids Res* **43,** D1049–D1056 (2015).

225. Prasad, T. S. K. *et al.* Human Protein Reference Database-2009 update. *Nucleic Acids Res* **37,** D767–D772 (2009).

226. Croft, D. *et al.* The Reactome pathway knowledgebase. *Nucleic Acids Res* **42,** D472–D477 (2014).

227. Szklarczyk, D. *et al.* STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res* **43,** D447–D452 (2015).

228. The Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature* **474,** 609–615 (2011).

229. The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* **499,** 43–49 (2013).

230. Hodgkinson, A. & Eyre-Walker, A. Variation in the mutation rate across mammalian genomes. *Nat Rev Genet* **12,** 756–766 (2011).

231. Arnheim, N. & Calabrese, P. Understanding what determines the frequency and pattern of human germline mutations. *Nat Rev Genet* **10,** 478–488 (2009).

232. Wolfe, K. H., Sharp, P. M. & Li, W. H. Mutation rates differ among regions of the mammalian genome. *Nature* **337,** 283–285 (1989).

233. Ellegren, H., Smith, N. G. C. & Webster, M. T. Mutation rate variation in the mammalian genome. *Curr Opin Genet Dev* **13,** 562–568 (2003).

234. Lercher, M. J. & Hurst, L. D. Human SNP variability and mutation rate are higher in regions of high recombination. *Trends Genet* **18,** 337–340 (2002).

235. Hellmann, I. *et al.* Why do human diversity levels vary at a megabase scale? *Genome Res* **15,** 1222–1231 (2005).

236. Prendergast, J. G. *et al.* Chromatin structure and evolution in the human genome. *BMC Evol Biol* **7,** 72 (2007).

237. Chen, C. L. *et al.* Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes. *Genome Res* **20,** 447–457 (2010).

238. Tolstorukov, M. Y., Volfovsky, N., Stephens, R. M. & Park, P. J. Impact of chromatin structure on sequence variability in the human genome. *Nat Struct Mol Biol* **18,** 510–515 (2011).

239. Hodgkinson, A., Chen, Y. & Eyre-Walker, A. The large-scale distribution of somatic mutations in cancer genomes. *Hum Mutat* **33,** 136–143 (2012).

240. Schuster-Bockler, B. & Lehner, B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature* **488,** 504–507 (2012).

241. Woo, Y. H. & Li, W. H. DNA replication timing and selection shape the landscape of nucleotide variation in cancer genomes. *Nat Commun* **3,** 1004 (2012).

242. Liu, L., De, S. & Michor, F. DNA replication timing and higher-order nuclear organization determine single-nucleotide substitution patterns in cancer genomes. *Nat Commun* **4,** 1502 (2013).

243. Polak, P. *et al.* Reduced local mutation density in regulatory DNA of cancer genomes is linked to DNA repair. *Nat Biotechnol* **32,** 71–75 (2013).

244. Polak, P. *et al.* Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature* **518,** 360–364 (2015).

245. Hanawalt, P. C. & Spivak, G. Transcription-coupled DNA repair: two decades of progress and surprises. *Nat Rev Mol Cell Biol* **9,** 958–970 (2008).

246. Supek, F. & Lehner, B. Differential DNA mismatch repair underlies mutation rate variation across the human genome. *Nature* **521,** 81–84 (2015).

247. Fei, J. Y. & Ha, T. Watching DNA breath one molecule at a time. *Proc Natl Acad Sci U S A* **110,** 17173–17174 (2013).

248. Perera, D. *et al.* Differential DNA repair underlies mutation hotspots at active promoters in cancer genomes. *Nature* **532,** 259–263 (2016).

249. Sabarinathan, R., Mularoni, L., Deu-Pons, J., Gonzalez-Perez, A. & Lopez-Bigas, N. Nucleotide excision repair is impaired by binding of transcription factors to DNA. *Nature* **532,** 264–267 (2016).

250. Khurana, E. Cancer genomics: Hard-to-reach repairs. *Nature* **532,** 181–182 (2016).

251. Helleday, T., Eshtad, S. & Nik-Zainal, S. Mechanisms underlying mutational signatures in human cancers. *Nat Rev Genet* **15,** 585–598 (2014).

252. Alexandrov, L. B. & Stratton, M. R. Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. *Curr Opin Genet Dev* **24,** 52–60 (2014).

253. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering signatures of mutational processes operative in human cancer. *Cell Reports* **3,** 246–259 (2013).

254. Fischer, A., Illingworth, C. J., Campbell, P. J. & Mustonen, V. EMu: probabilistic inference of mutational processes and their localization in the cancer genome. *Genome Biol* **14,** R39 (2013).

255. Gehring, J. S., Fischer, B., Lawrence, M. & Huber, W. SomaticSignatures: inferring mutational signatures from single-nucleotide variants. *Bioinformatics* **31,** 3673–3675 (2015).

256. Shiraishi, Y., Tremmel, G., Miyano, S. & Stephens, M. A simple model-based approach to inferring and visualizing cancer mutation signatures. *PLos Genet* **11,** e1005657 (2015).

257. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500,** 415–421 (2013).

258. Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149,** 979–993 (2012).

259. Roberts, S. A. *et al.* An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nat Genet* **45,** 970–976 (2013).

260. Kim, J. *et al.* Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nat Genet* **48,** 600–606 (2016).

261. Tanay, A. & Siggia, E. D. Sequence context affects the rate of short insertions and deletions in flies and primates. *Genome Biol* **9,** R37 (2008).

262. Kondrashov, A. S. & Rogozin, I. B. Context of deletions and insertions in human coding sequences. *Hum Mutat* **23,** 177–185 (2004).

263. Kvikstad, E. M., Tyekucheva, S., Chiaromonte, F. & Makova, K. D. A macaque's-eye view of human insertions and deletions: differences in mechanisms. *PLoS Comput Biol* **3,** 1772–1782 (2007).

264. Montgomery, S. B. *et al.* The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes. *Genome Res* **23,** 749–761 (2013).

265. Zhang, F. *et al.* The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. *Nat Genet* **41,** 849–853 (2009).

266. Kim, T.-M., Laird, P. & Park, P. The landscape of microsatellite instability in colorectal and endometrial cancer genomes. *Cell* **155,** 858–868 (2013).

267. Cardoso-Moreira, M., Arguello, J. R. & Clark, A. G. Mutation spectrum of Drosophila CNVs revealed by breakpoint sequencing. *Genome Biol* **13,** R119 (2012).

268. Gu, W., Zhang, F. & Lupski, J. R. Mechanisms for human genomic rearrangements. *Pathogenetics* **1,** 4 (2008).

269. Hastings, P. J., Lupski, J. R., Rosenberg, S. M. & Ira, G. Mechanisms of change in gene copy number. *Nat Rev Genet* **10,** 551–564 (2009).

270. Stankiewicz, P. & Lupski, J. R. Genome architecture, rearrangements and genomic disorders. *Trends Genet* **18,** 74–82 (2002).

271. Lupski, J. R. Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet* **14,** 417–422 (1998).

272. Elliott, B., Richardson, C. & Jasin, M. Chromosomal translocation mechanisms at intronic alu elements in mammalian cells. *Mol Cell* **17,** 885–894 (2005).

273. McVey, M. & Lee, S. E. MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings. *Trends Genet* **24,** 529–538 (2008).

274. Zhang, F., Carvalho, C. M. & Lupski, J. R. Complex human chromosomal and genomic rearrangements. *Trends Genet* **25,** 298–307 (2009).

275. Quinlan, A. R. & Hall, I. M. Characterizing complex structural variation in germline and somatic genomes. *Trends Genet* **28,** 43–53 (2012).

276. Lee, J. A., Carvalho, C. M. & Lupski, J. R. A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell* **131,** 1235–1247 (2007).

277. Hastings, P. J., Ira, G. & Lupski, J. R. A microhomology-mediated break-induced replication model for the origin of human copy number variation. *PLos Genet* **5,** e1000327 (2009).

278. Chen, J. M., Chuzhanova, N., Stenson, P. D., Ferec, C. & Cooper, D. N. Complex gene rearrangements caused by serial replication slippage. *Hum Mutat* **26,** 125–134 (2005).

279. Carvalho, C. M. & Lupski, J. R. Mechanisms underlying structural variant formation in genomic disorders. *Nat Rev Genet* **17,** 224–238 (2016).

280. Stephens, P. *et al.* Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature* **462,** 1005–1010 (2009).

281. Darai-Ramqvist, E. *et al.* Segmental duplications and evolutionary plasticity at tumor chromosome break-prone regions. *Genome Res* **18,** 370–379 (2008).

282. Drier, Y. *et al.* Somatic rearrangements across cancer reveal classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability. *Genome Res* **23,** 228–235 (2013).

283. Berger, M. F. *et al.* The genomic complexity of primary human prostate cancer. *Nature* **470,** 214–220 (2011).

284. Pleasance, E. D. *et al.* A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* **463,** 184–190 (2010).

285. Puente, X. S. *et al.* Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. *Nature* **475,** 101–105 (2011).

286. Meyer, L. R. *et al.* The UCSC Genome Browser database: extensions and updates 2013. *Nucleic Acids Res* **41,** D64–D69 (2013).

287. Hansen, R. S. *et al.* Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proc Natl Acad Sci U S A* **107,** 139–144 (2010).

288. Zhang, Y., Shin, H., Song, J. S., Lei, Y. & Liu, X. S. Identifying positioned nucleosomes with epigenetic marks in human from ChIP-Seq. *BMC Genomics* **9,** 537 (2008).

289. Schones, D. E. *et al.* Dynamic regulation of nucleosome positioning in the human genome. *Cell* **132,** 887–898 (2008).

290. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326,** 289–293 (2009).

291. Kong, A. *et al.* A high-resolution recombination map of the human genome. *Nat Genet* **31,** 241–247 (2002).

292. Wang, Z. *et al.* Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat Genet* **40,** 897–903 (2008).

293. Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129,** 823–837 (2007).

294. Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478,** 476–482 (2011).

295. De, S. & Michor, F. DNA secondary structures and epigenetic determinants of cancer genome evolution. *Nat Struct Mol Biol* **18,** 950–955 (2011).

296. Fudenberg, G., Getz, G., Meyerson, M. & Mirny, L. A. High order chromatin architecture shapes the landscape of chromosomal alterations in cancer. *Nat Biotechnol* **29,** 1109–1113 (2011).

297. Li, Y. *et al.* Comparative analysis of somatic copy-number alterations across different human cancer types reveals two distinct classes of breakpoint hotspots. *Hum Mol Genet* **21,** 4957–4965 (2012).

298. Crosetto, N. *et al.* Nucleotide-resolution DNA double-strand break mapping by next-generation sequencing. *Nat Methods* **10,** 361–365 (2013).

299. Cer, R. Z. *et al.* Non-B DB v2.0: a database of predicted non-B DNA-forming motifs and its associated tools. *Nucleic Acids Res* **41,** D94–D100 (2013).

300. Wang, G., Christensen, L. A. & Vasquez, K. M. Z-DNA-forming sequences generate large-scale deletions in mammalian cells. *Proc Natl Acad Sci U S A* **103,** 2677–2682 (2006).

301. Inagaki, H. *et al.* Chromosomal instability mediated by non-B DNA: cruciform conformation and not DNA sequence is responsible for recurrent translocation in humans. *Genome Res* **19,** 191–198 (2009).

302. Wang, G. & Vasquez, K. M. Naturally occurring H-DNA-forming sequences are mutagenic in mammalian cells. *Proc Natl Acad Sci U S A* **101,** 13448–13453 (2004).

303. Han, K. *et al.* L1 recombination-associated deletions generate human genomic variation. *Proc Natl Acad Sci U S A* **105,** 19366–19371 (2008).

304. Campbell, I. M. *et al.* Human endogenous retroviral elements promote genome instability via nonallelic homologous recombination. *BMC Biol* **12,** 74 (2014).

305. Myers, S., Freeman, C., Auton, A., Donnelly, P. & McVean, G. A common sequence motif associated with recombination hot spots and genome instability in humans. *Nat Genet* **40,** 1124–1129 (2008).

306. Zhou, W. *et al.* Increased genome instability in human DNA segments with self-chains: homology-induced structural variations via replicative mechanisms. *Hum Mol Genet* **22,** 2642–2651 (2013).

307. Fungtammasan, A., Walsh, E., Chiaromonte, F., Eckert, K. A. & Makova, K. D. A genome-wide analysis of common fragile sites: What features determine chromosomal instability in the human genome? *Genome Res* **22,** 993–1005 (2012).

308. Campos-Sánchez, R., Kapusta, A., Feschotte, C., Chiaromonte, F. & Makova, K. D. Genomic landscape of human, bat, and ex vivo DNA transposon integrations. *Mol Biol Evol* **31,** 1816–1832 (2014).

309. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15,** 1034–1050 (2005).

310. Tchurikov, N. A. *et al.* DNA double-strand breaks coupled with PARP1 and HNRNPA2B1 binding sites flank coordinately expressed domains in human chromosomes. *PLoS Genet* **9,** e1003429 (2013).

311. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).

312. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16,** 276–277 (2000).

313. Kozomara, A. & Griffiths-Jones, S. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res* **39,** D152–D157 (2011).

314. De, S., Pedersen, B. S. & Kechris, K. The dilemma of choosing the ideal permutation strategy while estimating statistical significance of genome-wide enrichment. *Brief Bioinform* **15,** 919–928 (2014).

315. Olson, D. L. & Delen, D. *Advanced Data Mining Techniques* (Springer-Verlag, Berlin, Germany, 2008).

316. R Development Core Team. *R: A language and environment for statistical computing.* (R Foundation for Statistical Computing, Vienna, Austria, 2011).

317. Venables, W. N. & Ripley, B. D. *Modern Applied Statistics with S* (Springer, New York, 2002).

318. Fox, J. & Weisberg, S. *An R Companion to Applied Regression* (Sage, Thousand Oaks, CA, 2011).

319. Fletcher, T. *QuantPsyc: Quantitative Psychology Tools* (2012).

320. Maindonald, J. & Braun, W. *Data Analysis and Graphics Using R* (Cambridge University Press, Cambridge, UK, 2010).

321. King, G. & Zeng, L. Logistic regression in rare events data. *Polit Anal* **9,** 137–163 (2001).

322. Imai, K., King, G. & Lau, O. Toward a common framework for statistical analysis and development. *J Comput Graph Stat* **17,** 1–22 (2008).

323. Breiman, L., Friedman, J., Olshen, R. & Stone, C. *Classification and Regression Trees* (Wadsworth, Belmont, CA, 1984).

324. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J Mach Learn Res* **12,** 2825–2830 (2011).

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

326. Nguyen, D. Q., Webber, C. & Ponting, C. P. Bias of selection on human copy-number variants. *PLoS Genet* **2,** e20 (2006).

327. Manning, A. L., Longworth, M. S. & Dyson, N. J. Loss of pRB causes centromere dysfunction and chromosomal instability. *Genes Dev* **24,** 1364–1376 (2010).

328. Artandi, S. E. *et al.* Telomere dysfunction promotes non-reciprocal translocations and epithelial cancers in mice. *Nature* **406,** 641–645 (2000).

329. Zhao, J., Bacolla, A., Wang, G. & Vasquez, K. M. Non-B DNA structure-induced genetic instability and evolution. *Cell Mol Life Sci* **67,** 43–62 (2010).

330. Konkel, M. K. & Batzer, M. A. A mobile threat to genome stability: The impact of non-LTR retrotransposons upon the human genome. *Semin Cancer Biol* **20,** 211–221 (2010).

331. Freudenreich, C. H. Chromosome fragility: molecular mechanisms and cellular consequences. *Front Biosci* **12,** 4911–4924 (2007).

332. Edelmann, L. *et al.* A common breakpoint on 11q23 in carriers of the constitutional t(11;22) translocation. *Am J Hum Genet* **65,** 1608–1616 (1999).

333. Mirabello, L., Troisi, R. J. & Savage, S. A. Osteosarcoma incidence and survival rates from 1973 to 2004: data from the Surveillance, Epidemiology, and End Results Program. *Cancer* **115,** 1531–1543 (2009).

334. Bielack, S. S. *et al.* Second and subsequent recurrences of osteosarcoma: presentation, treatment, and outcomes of 249 consecutive cooperative osteosarcoma study group patients. *J Clin Oncol* **27,** 557–565 (2009).

335. Bayani, J. *et al.* Spectral karyotyping identifies recurrent complex rearrangements of chromosomes 8, 17, and 20 in osteosarcomas. *Genes Chromosomes Cancer* **36,** 7–16 (2003).

336. Smida, J. *et al.* Genomic alterations and allelic imbalances are strong prognostic predictors in osteosarcoma. *Clin Cancer Res* **16,** 4256–4267 (2010).

337. Kuijjer, M. L. *et al.* Identification of osteosarcoma driver genes by integrative analysis of copy number and gene expression data. *Genes Chromosomes Cancer* **51,** 696–706 (2012).

338. Luetke, A., Meyers, P. A., Lewis, I. & Juergens, H. Osteosarcoma treatment - where do we stand? A state of the art review. *Cancer Treat Rev* **40,** 523–532 (2014).

339. Man, T.-K. *et al.* Expression profiles of osteosarcoma that can predict response to chemotherapy. *Cancer Res* **65,** 8142–8150 (2005).

340. Martin, J. W., Squire, J. A. & Zielenska, M. The genetics of osteosarcoma. *Sarcoma* **2012,** 627254 (2012).

341. Durkin, S. G. & Glover, T. W. Chromosome fragile sites. *Annu Rev Genet* **41,** 169–192 (2007).

342. Zeman, M. K. & Cimprich, K. A. Causes and consequences of replication stress. *Nat Cell Biol* **16,** 2–9 (2014).

343. Smeets, S. J. *et al.* To DNA or not to DNA? That is the question, when it comes to molecular subtyping for the clinic! *Clin Cancer Res* **17,** 4959–4964 (2011).

344. Chen, X. *et al.* Recurrent somatic structural variations contribute to tumorigenesis in pediatric osteosarcoma. *Cell Rep* **7,** 104–112 (2014).

345. Perry, J. A. *et al.* Complementary genomic approaches highlight the PI3K/mTOR pathway as a common vulnerability in osteosarcoma. *Proc Natl Acad Sci U S A* **111,** E5564–E5573 (2014).

346. Poos, K. *et al.* Genomic heterogeneity of osteosarcoma - shift from single candidates to functional modules. *PLoS One* **10,** e0123082 (2015).

347. Kansara, M., Teng, M. W., Smyth, M. J. & Thomas, D. M. Translational biology of osteosarcoma. *Nat Rev Cancer* **14,** 722–735 (2014).

348. Magrangeas, F., Avet-Loiseau, H., Munshi, N. C. & Minvielle, S. Chromothripsis identifies a rare and aggressive entity among newly diagnosed multiple myeloma patients. *Blood* **118,** 675–678 (2011).

349. Molenaar, J. J. *et al.* Sequencing of neuroblastoma identifies chromothripsis and defects in neuritogenesis genes. *Nature* **483,** 589–593 (2012).

350. Rausch, T. *et al.* Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations. *Cell* **148,** 59–71 (2012).

351. Maher, C. A. & Wilson, R. K. Chromothripsis and human disease: piecing together the shattering process. *Cell* **148,** 29–32 (2012).

352. Bielack, S. S. *et al.* Prognostic factors in high-grade osteosarcoma of the extremities or trunk: an analysis of 1,702 patients treated on neoadjuvant cooperative osteosarcoma study group protocols. *J Clin Oncol* **20,** 776–790 (2002).

353. Kolesnikov, N. *et al.* ArrayExpress update-simplifying data submissions. *Nucleic Acids Res* **43,** D1113–D1116 (2015).

354. Qiao, Y. *et al.* SubcloneSeeker: a computational framework for reconstructing tumor clone structure for cancer variant interpretation and prioritization. *Genome Biol* **15,** 443 (2014).

355. Sarni, D. & Kerem, B. The complex nature of fragile site plasticity and its importance in cancer. *Curr Opin Cell Biol* **40,** 131–136. ISSN: 1879-0410 (June 2016).

356. Cai, H. *et al.* Chromothripsis-like patterns are recurring but heterogeneously distributed features in a survey of 22,347 cancer genome screens. *BMC Genomics* **15,** 82 (2014).

357. Hernandez-Ferrer, C. *et al.* 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).

358. Uddin, M. *et al.* A high-resolution copy-number variation resource for clinical and population genetics. *Genet Med* **17,** 747–752 (2014).

359. Popova, T., Manié, E. & Stern, M. Genomic signature of homologous recombination deficiency in breast and ovarian cancers. *Bio-protocol* **3,** e814 (2013).

360. Popova, T. *et al.* Ploidy and large-scale genomic instability consistently identify basal-like breast carcinomas with BRCA1/2 inactivation. *Cancer Res* **72,** 5454–5462 (2012).

361. Forbes, S. A. *et al.* COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res* **39,** D945–D950 (2011).

362. Ribi, S. *et al.* TP53 intron 1 hotspot rearrangements are specific to sporadic osteosarcoma and can cause Li-Fraumeni syndrome. *Oncotarget* **6,** 7727–7740 (2015).

363. Kovac, M. *et al.* Exome sequencing of osteosarcoma reveals mutation signatures reminiscent of BRCA deficiency. *Nat Commun* **6,** 8940 (2015).

364. Martin, J. W., Zielenska, M., Stein, G. S., van Wijnen, A. J. & Squire, J. A. The role of RUNX2 in osteosarcoma oncogenesis. *Sarcoma* **2011,** 282745 (2011).

365. Weiner, L. *et al.* Dedicated epithelial recipient cells determine pigmentation patterns. *Cell* **130,** 932–942 (2007).

366. Mangelsdorf, M. *et al.* Chromosomal fragile site FRA16D and DNA instability in cancer. *Cancer Res* **60,** 1683–1689 (2000).

367. Aqeilan, R. I., Abu-Remaileh, M. & Abu-Odeh, M. The common fragile site FRA16D gene product WWOX: roles in tumor suppression and genomic stability. *Cell Mol Life Sci* **71,** 4589–4599 (2014).

368. Schrock, M. S. & Huebner, K. WWOX: a fragile tumor suppressor. *Exp Biol Med (Maywood)* **240,** 296–304 (2015).

369. Yang, J. *et al.* Deletion of the WWOX gene and frequent loss of its protein expression in human osteosarcoma. *Cancer Lett* **291,** 31–38 (2010).

370. Del Mare, S. & Aqeilan, R. I. Tumor Suppressor WWOX inhibits osteosarcoma metastasis by modulating RUNX2 function. *Sci Rep* **5,** 12959 (2015).

371. Zheng, S. *et al.* A survey of intragenic breakpoints in glioblastoma identifies a distinct subset associated with poor survival. *Genes Dev* **27,** 1462–1472 (2013).

372. Zhang, Y., Xu, H. & Frishman, D. Genomic determinants of somatic copy number alterations across human cancers. *Hum Mol Genet* **25,** 1019–1030 (2016).

373. Forment, J. V., Kaidi, A. & Jackson, S. P. Chromothripsis and cancer: causes and consequences of chromosome shattering. *Nat Rev Cancer* **12,** 663–670 (2012).

374. Mardin, B. R. *et al.* A cell-based model system links chromothripsis with hyperploidy. *Mol Syst Biol* **11,** 828 (2015).

375. Nunez, M. I. *et al.* WWOX protein expression varies among ovarian carcinoma histotypes and correlates with less favorable outcome. *BMC Cancer* **5,** 64 (2005).

376. Kurek, K. C. *et al.* Frequent Attenuation of the WWOX Tumor Suppressor in Osteosarcoma Is Associated with Increased Tumorigenicity and Aberrant RUNX2 Expression. *Cancer Res* **70,** 5577–5586 (2010).

377. Tubio, J. M. C. & Estivill, X. Cancer: When catastrophe strikes a cell. *Nature* **470,** 476–477 (2011).

378. Leibowitz, M. L., Zhang, C. Z. & Pellman, D. Chromothripsis: a new mechanism for rapid karyotype evolution. *Annu Rev Genet* **49,** 183–211 (2015).

379. Maciejowski, J., Li, Y., Bosco, N., Campbell, P. J. & de Lange, T. Chromothripsis and kataegis induced by telomere crisis. *Cell* **163,** 1641–1654 (2015).

380. Meyerson, M. & Pellman, D. Cancer genomes evolve by pulverizing single chromosomes. *Cell* **144,** 9–10 (2011).

381. Crasta, K. *et al.* DNA breaks and chromosome pulverization from errors in mitosis. *Nature* **482,** 53–58 (2012).

382. Zhang, C.-Z. *et al.* Chromothripsis from DNA damage in micronuclei. *Nature* **522,** 179–184 (2015).

383. Osterrieder, N., Kamil, J. P., Schumacher, D., Tischer, B. K. & Trapp, S. Marek's disease virus: from miasma to model. *Nat Rev Microbiol* **4,** 283–294 (2006).

384. Jarosinski, K. W., Tischer, B. K., Trapp, S. & Osterrieder, N. Marek's disease virus: lytic replication, oncogenesis and control. *Expert Rev Vaccines* **5,** 761–772 (2006).

385. Davison, F. & Nair, V. *Marek's disease: an evolving problem* (Academic Press, 2004).

386. Steep, A. *Discovery of somatic driver variations of MDV-induced transformations and genetic elements demonstrating resistance to Marek's Disease via cytogenetic analysis, genome-wide selection, and next-generation sequencing* tech. rep. (2014).

387. Churchill, A., Payne, L. & Chubb, R. Immunization against Marek's disease using a live attenuated virus. *Nature* **221,** 744–747 (1969).

388. Nair, V. Evolution of Marek's disease–a paradigm for incessant race between the pathogen and the host. *Vet J* **170,** 175–183 (2005).

389. Read, A. F. *et al.* Imperfect vaccination can enhance the transmission of highly virulent pathogens. *PLoS Biol* **13,** e1002198 (2015).

390. McPherson, M. & Delany, M. Virus and host genomic, molecular, and cellular interactions during Marek's disease pathogenesis and oncogenesis. *Poult Sci* **95,** 412–429 (2016).

391. Beasley, J., Patterson, L., McWade, D., *et al.* Transmission of Marek's disease by poultry bouse dust and chicken dander. *Am J Vet Res* **31,** 339–344 (1970).

392. Barrow, A. D., Burgess, S. C., Baigent, S. J., Howes, K. & Nair, V. K. Infection of macrophages by a lymphotropic herpesvirus: a new tropism for Marek's disease virus. *J Gen Virol* **84,** 2635–2645 (2003).

393. Calnek, B. & Witter, R. L. Marek's disease—a model for herpesvirus oncology. *Crit Rev Microbiol* **12,** 293–320 (1985).

394. Johnson, E., Burke, C., Fredrickson, T. & DiCapua, R. Morphogenesis of Marek's disease virus in feather follicle epithelium. *J Natl Cancer Inst* **55,** 89–99 (1975).

395. Jones, D., Lee, L., Liu, J.-L., Kung, H.-J. & Tillotson, J. K. Marek disease virus encodes a basic-leucine zipper gene resembling the fos/jun oncogenes that is highly expressed in lymphoblastoid tumors. *Proc Natl Acad Sci U S A* **89,** 4042–4046 (1992).

396. Kung, H.-J. *et al.* Meq: an MDV-specific bZIP transactivator with transforming properties. *Curr Top Microbiol Immunol* **255,** 245–260 (2001).

397. Lupiani, B. *et al.* Marek's disease virus-encoded Meq gene is involved in transformation of lymphocytes but is dispensable for replication. *Proc Natl Acad Sci U S A* **101,** 11815–11820 (2004).

398. Ajithdoss, D. K. *et al.* In vitro characterization of the Meq proteins of Marek's disease virus vaccine strain CVI988. *Virus Res* **142,** 57–67 (2009).

399. Delecluse, H.-J. & Hammerschmidt, W. Status of Marek's disease virus in established lymphoma cell lines: herpesvirus integration is common. *J Virol* **67,** 82–92 (1993).

400. Robinson, C. M., Hunt, H. D., Cheng, H. H. & Delany, M. E. Chromosomal integration of an avian oncogenic herpesvirus reveals telomeric preferences and evidence for lymphoma clonality. *Herpesviridae* **1,** 5 (2010).

401. Witter, R. Increased virulence of Marek's disease virus field isolates. *Avian Dis,* 149–163 (1997).

402. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30,** 2114–2120 (2014).

403. Joshi, N. A. & Fass, J. N. *Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files* https://github.com/najoshi/sickle. 2011.

404. Andrews, S. *FastQC: a quality control tool for high throughput sequence data* https://www.bioinformatics.babraham.ac.uk/projects/fastqc/. 2010.

405. Auwera, G. A. *et al.* From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr Protoc Bioinformatics* **43,** 1–33 (2013).

406. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv,* 1303.3997 (2013).

407. Boeva, V. *et al.* Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. *Bioinformatics* **27,** 268–269 (2011).

408. Chong, Z. *et al.* novoBreak: local assembly for breakpoint detection in cancer genomes. *Nat Methods* **14,** 65–67 (2017).

409. Cho, A. *et al.* MUFFINN: cancer gene discovery via network analysis of somatic mutation data. *Genome Biol* **17,** 129 (2016).

410. Sonnhammer, E. L. & Östlund, G. InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Res* **43,** D234–D239 (2015).

411. Skidmore, Z. L. *et al.* GenVisR: Genomic Visualizations in R. *Bioinformatics* **32,** 3012–3014 (2016).

412. Butler, J. L., Osborne Locke, M. E., Hill, K. A. & Daley, M. HD-CNV: hotspot detector for copy number variants. *Bioinformatics* **29,** 262–263 (2013).

413. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).

414. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30,** 923–930 (2014).

415. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15,** 550 (2014).

416. Mi, H., Muruganujan, A., Casagrande, J. T. & Thomas, P. D. Large-scale gene function analysis with the PANTHER classification system. *Nat Protoc* **8,** 1551–1566 (2013).

417. Phillips, J. E. & Corces, V. G. CTCF: master weaver of the genome. *Cell* **137,** 1194–1211 (2009).

418. Fiorentino, F. P. & Giordano, A. The tumor suppressor role of CTCF. *J Cell Physiol* **227,** 479–492 (2012).

419. Carraway, K. L., Theodoropoulos, G., Kozloski, G. A. & Carothers Carraway, C. A. Muc4/MUC4 functions and regulation in cancer. *Future Oncol* **5,** 1631–1640 (2009).

420. Rebollo, A. & Schmitt, C. Ikaros, Aiolos and Helios: transcription regulators and lymphoid malignancies. *Immunol Cell Biol* **81,** 171–175 (2003).

421. Yang, L., Luo, Y. & Wei, J. Integrative genomic analyses on Ikaros and its expression related to solid cancer prognosis. *Oncol Rep* **24,** 571 (2010).

422. Mullighan, C. G. *et al.* BCR-ABL1 lymphoblastic leukaemia is characterized by the deletion of Ikaros. *Nature* **453,** 110–114 (2008).