Dissertation

submitted to the

Combined Faculties for the Natural Sciences and for Mathematics

of the Ruperto-Carola University of Heidelberg, Germany

for the degree of

Doctor of Natural Sciences

presented by

Jelena Tica

M.Sc. Molecular Biology

born in: Banja Luka, Bosnia and Herzegovina

Date of oral examination: 18^{th} May 2015

INVESTIGATING ORIGIN AND FUNCTIONAL IMPACT OF GENOMIC STRUCTURAL VARIANTS WITH NEXT-GENERATION SEQUENCING

Referees:

Dr. Lars Steinmetz

Prof. Dr. Rainer König

Abstract

Genomic variants play an important role in phenotypic variation and have significant impact on a disease development. Due to the technology limitations, inference of genomic variants and their potential consequence on phenotype was until recently restricted. Only with the advent of next-generation sequencing (NGS) approaches, could a vast majority of genomic variants be successfully identified for the first time. In my PhD Thesis I will present my work on structural variants (SVs), their formation mechanism and their functional impact.

The first part of my Thesis focuses on structural variants in non-human primates, studies of which using NGS have not been pursued prior to the research studies we carried out. In order to inspect the origin and functional impact of SV formation mechanisms, we constructed a comprehensive SV map based on the fibroblast-derived DNA from three different species: chimpanzee, orangutan and rhesus macaque. We noted striking differences in the activity of homology-related SV formation mechanisms between the great apes and rhesus macaques, with a third of the chimpanzee and orangutan SVs inferred to be formed by non-allelic homologous recombination compared with only 2% of the macaque SVs. One additional key finding was the presence of a markedly higher mobile element activity in macaques compared to the other non-human primates studies. Additionally, we could show that long L1 elements surpassed *Alu* activity in chimpanzee and orangutan as opposed to macaque where *AluMacYa3* dominates the genomic landscape causing a burst of relatively short SVs. In addition to inserting into genome, active L1 elements possess the ability to mobilize 3' flanking DNA to different genomic loci as transductions. By combining translocation and L1 discovery pipelines we further developed a novel computational methodology, termed TIGER, for the discovery of polymorphic L1-mediated 3' transductions. We employed TIGER to a deeply sequenced human genome and to aforementioned non-human primates species to characterize transductions. TIGER enables studying germline L1-mediated 3' transductions, making a relevant structural variation class amenable for population and disease studies for the first time.

In the second part of my Thesis, I discuss the differences in the formation mechanisms of both germline and somatic SVs in the human genome. Our *de novo* mechanism classification analyses performed on four previously published SV datasets revealed that almost half of germline human SVs are due to mechanisms independent of homology, followed by homology-related DNA repair, mobile elements and variable number of tandem repeats. We also investigated the formation of somatic SVs in four medulloblastoma brain tumor patients with a germline *TP53* mutation (Li-Fraumeni syndrome). In contrast to the germline SVs, our analyses of rearrangement breakpoints in medulloblastoma in the context of mutated *TP53*, rather support a model of massive DNA double strand breaks known as chromothripsis, followed by exclusive homology-independent repair.

Zusammenfassung

Genomische Varianten sind von großer Bedeutung für phänotypische Unterschiede; somit auch bei der Entwicklung von Krankheiten. Bis jetzt war die Untersuchung von Genomvarianten und ihren potentiellen Auswirkungen auf den Phänotypen von technischen Möglichkeiten beschränkt. Durch das Aufkommen der Hochdurchsatz-Sequenzierungsmethoden (NGS) kann nun erstmalig erfolgreich eine große Anzahl von Genomvarianten identifiziert werden. In meiner Doktorarbeit erläutere ich das Auftreten großer struktureller Variationen (SVs), ihrer Bildungsmechanismen und funktionelle Bedeutung.

Der erste Teil meiner Dissertation bezieht sich auf SVs in Primaten. Vor der vorliegenden Arbeit wurden derartige Analysen, noch nicht verfolgt unter Verwendung von NGS. Um das Auftreten und die funktionelle Auswirkung von SVs erforschen zu können, konstruierten wir umfassende SV-Listen basierend auf der Fibroblasten-DNS dreier Spezies: Schimpanse, Orang-Utan und Rhesusaffe. Wir haben markante Unterschiede in der Aktivität von Homologie-abhängiger SV-Bildung zwischen Menschffen und Rhesusaffen festgestellt, wodurch ein Drittel der SVs bei Schimpanse und Orang-Utan durch nichthomologe Rekombination entstehen, im Gegensatz von nur 2% der SVs beim Rhesusaffen. Ein weiteres Schlüsselergebnis war die eindeutig höhere Aktivität von mobilen Elementen im Rhesusaffen verglichen mit den anderen untersuchten Primaten. Es konnte auch gezeigt werden, dass die Aktivität langer L1 Elemente die *Alu*-Aktivität im Schimpansen und Orang-Utan übertrifft, verglichen mit dem Rhesusaffen, bei welchem kurze *AluMacYa3* das Genom dominieren. Zusätzlich können aktive L1 Elemente auch 3' benachbarte DNS Sequenzen mobilizieren in andere Genomregionen durch den Prozess der Transduktion einbauen. Wir haben einen Algorithmus namens TIGER entwickelt, der die Methoden zur Detektion von sowohl L1- als auch Translokationen verbindet und mit dessen Hilfe polymorphe, L1-vermittelte 3' Transduktionen aufdeckt. Wir haben TIGER zur Charakterisierung von Transduktionen bei Primaten und für das humane Genom verwendet. TIGER ermöglicht somit erstmals die Untersuchung L1-vermittelter 3' Transduktionen in der Keimbahn, und damit die Erforschung dieser bedeutenden SVs innerhalb von Populationen und ihr Auftreten bei Erkrankungen.

Im zweiten Teil bespreche ich die Entstehungsmechanismen von sowohl Keimbahn- als auch somatischen SV Datensätzen im humanen Genom. Unsere *de novo* Klassifikationsanalyse, basierend auf bereits publizierten SVs zeigt, dass fast die Hälfte der humanen Keimbahn SVs durch nichthomologe Mechanismen hervorgerufen werden. Weiterhin haben wir das Auftreten somatischer SVs in vier Medulloblastom Gehirntumorpatienten mit Keimbahn-*TP53*-Mutation (Li-Fraumeni Syndrom) untersucht. Im Gegensatz zur Keimbahnanalyze, unterstützen unsere Analysen von SVs beim Medulloblastom eher das Modell massiver Doppelstrangbrüche, durch einen Chromothripsis genannten Mechanismus, welche alleinig von nicht-homologie-abhängigerer DNS-Reparatur korrigiert werden.

Acknowledgements

My PhD would not have been possible without the guidance and the help of people who in one way or another contributed to this work, for which I am truly grateful.

First and foremost, my utmost gratitude to my supervisor Dr. Jan Korbel whose encouragement I will never forget. I would like to thank him for giving me the opportunity to work with him and his group and for helping me with all the obstacles I encountered during my research.

I further wish to thank my TAC members, Dr. Lars Steinmetz, Dr. Rainer König and Dr. Paul Flicek, whose feedback and suggestions enabled me to develop scientifically and progress in my work. Thanks and appreciation to the collaborators in Heidelberg and Boston: without permission to use their experimental data and sharing all the valuable information, this research would not see the light of the day. Special thanks to Eunjung Lee, Rebecca Iskow and Omer Gokcumen for their support and collaboration in the primate sequencing project.

My thanks to the entire Korbel group for their friendship, guidance and for providing a pleasant working atmosphere. I wish to acknowledge Adrian Stütz for his support in the primate transduction project. I would also like to thank Giorgia Guglielmi, Christopher Buccitelli, Nina Habermann and Adrian Stütz for their feedback regarding the work presented in this Thesis. Special thanks to Maia Segura Wang, Stephanie Sungalee, Thomas Zichner and Verena Tischler, for being there when I needed them, as well as for their patience and advices. They were my small lab family here in Heidelberg and we went together though 'thick and thin'.

I also thank my friends for being there during the past four years: Lidija Stanišić and Marija Balint for all the great times we have spent during my short home visits; Matilda Maleš and Giorgia Guglielmi for all coffees, cakes and short and long talks during late afternoons at EMBL... Their support and encouragements made my PhD experience enjoyable. Many thanks to Igor Jukić, for always being there and providing a boost of motivation when I needed it the most.

Finally, my endless gratitude to my family: my mom, my dad, my grandmother and my grandfather - without them and their selfless support, understanding, patience and love during my education path, none of this would have been possible.

Za mamu i tatu...

Contents

Abbreviations

1000GP	1000 Genomes Project
aCGH	Array-Comparative Genomic Hybridization
BAF	B Allele Frequency
bp	base pair
cDNA	Complementary Deoxyribonucleic Acid
CGH	Comparative Genomic Hybridization
ChIP	Chromatin Immunoprecipitation
chr	chromosome
CNV	Copy Number Variant
DNA	Deoxyribonucleic Acid
DSB	Double Strand Break
EMBL	European Molecular Biology Laboratory
FDR	False Discovery Rate
FISH	Flourescent *In Situ* Hybridization
FoSTeS	Fork Stalling and Template Switching
GRIP	Gene Retrocopy Insertion Polymorphism
HERV	Human Endogenous Retrovirus
HR	Homologous Recombination
ICGC	International Cancer Genome Consortium
Indel	Insertion/deletion
Kb	Kilobase(s)
LINE	Long Interspersed Nuclear Element
LFS	Li-Fraumeni Syndrome
LOH	Loss of heterozygosity
LTR	Long Terminal Repeat
MB	Medulloblastoma
Mb	Megabase(s)
ME	Mobile Element
MEI	Mobile Element Insertion

mRNA	Messenger Ribonucleic Acid
MMBIR	Microhomology-Mediated Break-Induced Replication
Mya	Million years ago
Myr	Million years
NAHR	Non-Allelic Homologous Recombination
NHEJ	Non-Homologous End-Joining
NHR	Non-Homologous Rearrangement
NGS	Next Generation Sequencing
Nt	Nucleotide(s)
pA/polyA	Polyadenylation
PCR	Polymerase Chain Reaction
RMD	Retrotransposon-Mediated Deletion
RNA	Ribonucleic Acid
SA	Single-Anchored
SD	Segmental Duplication
SHH	Sonic-Hedgehog
SINE	Short Interspersed Nuclear Element
SMRT	Single-Molecule Real Time sequencing
SNP	Single Nucleotide Polymophism
SNV	Single Nucleotide Variant
SV	Structural Variant
SVA	SINE-VNTR-*Alu*
TL	Translocation
TPRT	Target-Primed Reverse Trancription
TS	Transduced Sequence
TSD	Target-Site Duplication
VNTR	Variable Number of Tandem Repeats
UCSC	University of California, Santa Cruz
WES	Whole-Exome Sequencing
WGS	Whole-Genome Sequencing
WHO	World Health Organization

Chapter 1

Introduction

Genetics is an important field in the life sciences with its main purposes being to uncover the function of genes and the nature of heredity and genetic variation in living organisms. The word genetics derives from the Ancient Greek word γένεσις - genesis which means 'origin' in common English. Some of the earliest hereditary theories were established by Hippocrates and Aristotle in Ancient Greece. Their theories about the inheritance of acquired characters remained as the accepted standard until the 19^{th} century. Understanding genetics as a process began with the work of Augustinian friar Gregor Johann Mendel in the mid-19^{th} century. Mendel, the founder of modern genetics, is widely known for his inheritance experiments on pea plants (*Pisum sativum*). He bred over 25,000 pea plants and observed that certain inheritable traits follow simple statistical rules, today known as Mendel's Laws of Inheritance [Mendel, 1866].

Today, with advances in technology, we know that all genetic information is stored in a molecule named deoxyribonucleic acid (DNA). DNA encodes for a whole set of genetic instruction needed for successful development and function of an organism. As its name suggests, DNA is a nucleic acid, which together with proteins and carbohydrates make up the major macromolecules essential for all life. In 1953, James D. Watson and Francis Crick discovered that DNA is composed of two separate polynucleotide strands, coiled together into a double-stranded helix [Watson and Crick, 1953]. Each strand is composed of nucleotides or nitrogen bases, a monosaccharide sugar called deoxyribose and a phosphate group. Nitrogen bases are commonly referred to by letters A, C, G and T which stand for adenine, guanine, thymine and cytosine, respectively. In eukaryotes, DNA is packed into higher structures called chromosomes. The human genome, for instance, is a diploid genome consisting of 23 pairs of chromosomes, from which 22 are autosomes (1-22) and 1 is an allosome pair (sex chromosomes, X and Y).

Comparing DNA across and within species reveals a wealth of differences - across different species, between two individual genomes from the same species and even within the same individual

between tissues. For example, the genomes of two healthy humans can differ by as much as 1% [Pang et al., 2010]. Some of the variants are present in the germline, meaning that they are passed on through generations, whereas some are somatic - acquired postnatally and thus not inheritable. Due to recent technological advances, today it is possible to obtain a whole-genome sequence in less than a day and study genome-wide repertoires of variants at once. One such revolutionizing technology, DNA sequencing, has generated large amounts of sequenced genomes. Despite having multiple advantages, this technology comes with certain challenges, all of which will be discussed throughout this chapter.

During my PhD, I analyzed whole-genome sequencing data from human and non-human primates, in a context of disease and evolution, respectively. In both projects I worked on structural variant mechanism formation, with specific interest in retrotransposons - variants able to move within a genome.

1.1 Genomic variations

Genomic variation encompasses the set of differences observed between DNA sequences. Although two individual genomes of the same species are usually very similar, every existing genome is unique. This naturally occurring variation permits flexibility and survival of the population. Many variants are neutral; neither beneficial nor detrimental. They do not affect an organism's ability to survive and reproduce and are subsequently inheritable. Other variant effects can be either positive or negative, resulting in various phenotypic responses, from differences in physical appearance to predisposition to different diseases.

The functional impact of genomic variation is highly dependent on location in the genome and the size of affected area. Based on their size, genomic variants are usually split into two categories: (1) small-scale variants ranging from 1 to few base pairs, and (2) large-scale variants, which are typically defined as larger than 50 base pairs [Mills et al., 2011]. Larger variants have higher chances of affecting genes and gene regulatory regions, potentially causing changes in gene expression and regulation. However, small variants involving only few base pairs can ultimately have similar effects.

In the past decade, there have been various research efforts to establish detailed catalogs of genomic variants with the aim to understand the role of variants in their individual, population-scale and disease context.

1.1.1 Small-scale variants

Small-scale variants are typically defined as genomic variants up to 50 basepairs (bp) in size. Variants affecting only 1 bp are known as single nucleotide variants (SNVs), whereas those ranging from 2 bp to roughly 50 bp are termed indels (short insertions/deletions).

SNVs with a frequency that is higher than 1% in a given population, are usually considered polymorphisms and therefore called single nucleotide polymorphisms (SNPs). In general, SNVs are the most abundant form of genomic variations with ∼3 million SNVs commonly occurring across human genomes [The 1000 Genomes Project Consortium, 2012, The International HapMap, 2003]. In terms of SNV categorization, they can be split into transitions and transversions [Freese, 1959]. A transition involves the replacement of a purine base (adenine (A), guanine (G)) with another purine (A↔G) or a pyrimidine base (cytosine (C), thymine (T)) with another pyrimidine (C↔T), whereas whereas transversions involve the substitution of a purine with a pyrimidine or vice versa (A↔, A↔T, G↔C, G↔T). In humans, transitions are twice as common as transversions [Zhang and Gerstein, 2003] and the most frequent transition accounting for almost half of human SNVs is C→T, caused by spontaneous deamination of 5-methylcytosine [Shen et al., 1994].

SNVs may occur anywhere in the genome and although they are scarcer in protein-coding regions [Barreiro et al., 2008, The International SNP Map Working Group, 2001], non-coding SNVs may still alter gene splicing or transcription factor binding. Many coding SNVs are silent with no effect on phenotype. However, others (termed nonsynonymous SNVs) can directly alter the amino acid sequence, causing the emergence of a premature stop codons or even a shift in the open reading frame. Both states usually result in a truncated or damaged protein that is typically functionally different or even obsolete [Ng and Henikoff, 2006].

Indels are a less abundant form of genomic variations. Similar to SNVs, they are usually depleted from protein-coding regions and rather tend to cluster within repetitive sequences. During replication, indels can occur due to DNA polymerase slippage, resulting in an expansion or a shortening of tandem repeats [Montgomery et al., 2013, Mullaney et al., 2010]. In gene-coding regions, if the number of added/removed nucleotides is not corresponding to a complete codon (i.e. to three consecutive nucleotides), indels produce frameshift mutations, leading to nonfunctional proteins in most cases [Hu and Ng, 2012, Nagy and Maquat, 1998]. Therefore, they are less likely to be observed in comparison to non-frameshift indels, which result in an entire amino acid being inserted or deleted and are thus less damaging.

As previously mentioned, both SNVs and indels can alter the protein sequence in similar ways with analogous phenotypic consequences. Of note, both variant types can be associated with increased risks for several diseases, including cancer [Frazer et al., 2009, Yang et al., 2010].

1.1.2 Large-scale variants

Genomic variants larger than ∼50 bp are usually defined as large-scale structural variants (SVs). SVs vary in size and therefore can involve both microscopic and submicroscopic events, ranging from several kilobases up to a few megabases [Baker, 2012, Feuk et al., 2006]. Known SV types involve copy-number variants (CNVs), such as deletions and duplications, as well as balanced SVs (inversions and translocations) and insertions.

Mobile elements insertions (MEIs) represent a very interesting SV class, since they have the ability to amplify themselves and subsequently insert into various genomic locations. Although historically labeled as 'junk' DNA, it is worth mentioning that in mammals nearly 50% of the genome is composed of various repetitive sequences [Cordaux and Batzer, 2009]. Recent research suggests that MEIs are one of the key players in genomic structural variation formation [Gokcumen et al., 2013, Helman et al., 2014, Kidd et al., 2010, Lee et al., 2012, Mills et al., 2011, Tubio et al., 2014]. They can generate local genomic instability and disrupt gene activity directly by inserting into a gene or a regulatory element. Indirectly, MEIs can also produce additional genomic rearrangements in the form of deletions, duplications and inversions [Cordaux and Batzer, 2009, Gilbert et al., 2002].

The fraction of the genome affected and consequent phenotypic impact of SVs is larger than that of SNVs. Unsurprisingly, SV associations have been made with both diverse diseases, as well as with normal traits (reviewed in Onishi-Seebacher and Korbel [2011], Weischenfeldt et al. [2013]). For example, the salivary amylase gene (*AMY1*) copy-number is positively correlated with salivary amylase protein levels and the ability to digest starch. Populations with no salivary amylase, such as chimpanzees, tend to consume little or no starch, whereas human populations with greater *AMY1* copy-number have traditionally starch-rich diet [Perry et al., 2007].

SVs are a less studied class of genetic variants than SNVs due to the technical limitations of their detection. During the last 50 years, there have been numerous examples linking SVs with various disease phenotypes. Trisomy of chromosome 21 is a well characterized structural variant causing Down syndrome Korbel et al. [2009b]. Another example involves a recurrent 400 kb inversion in the factor VIII gene causing hemophilia A [Lakich et al., 1993, Naylor et al., 1993]. In cancer, one of the first rearrangements discovered was a translocation of the *Abl1* gene on chromosome 9 to a part of the *BCR* gene on chromosome 22, resulting in a fusion gene. This fusion between chromosome 9 and 22 is known as the Philadelphia chromosome, and the major driving event behind chronic myelogenous leukemia (CML) [Nowell C. and Hungerford A., 1960]. In general, SVs can have various functional consequences on the genome and phenotype [Weischenfeldt et al., 2013]. For example, they can alter gene coding regions by removing part of a gene or by fusing genes together. Deletions and duplications can give rise to different gene copy-numbers, thereby causing gene dosage changes. Apart from affecting genes directly, structural variants can have a

positional effect where a resulting change in position of either a gene or a regulatory element may result in altered gene expression (Figure 1.1). To better characterize individual genomes, large genome consortia are investing massive efforts in variant detection strategies and technologies with the aim of facilitating the identification of structural variants at base-pair resolution.

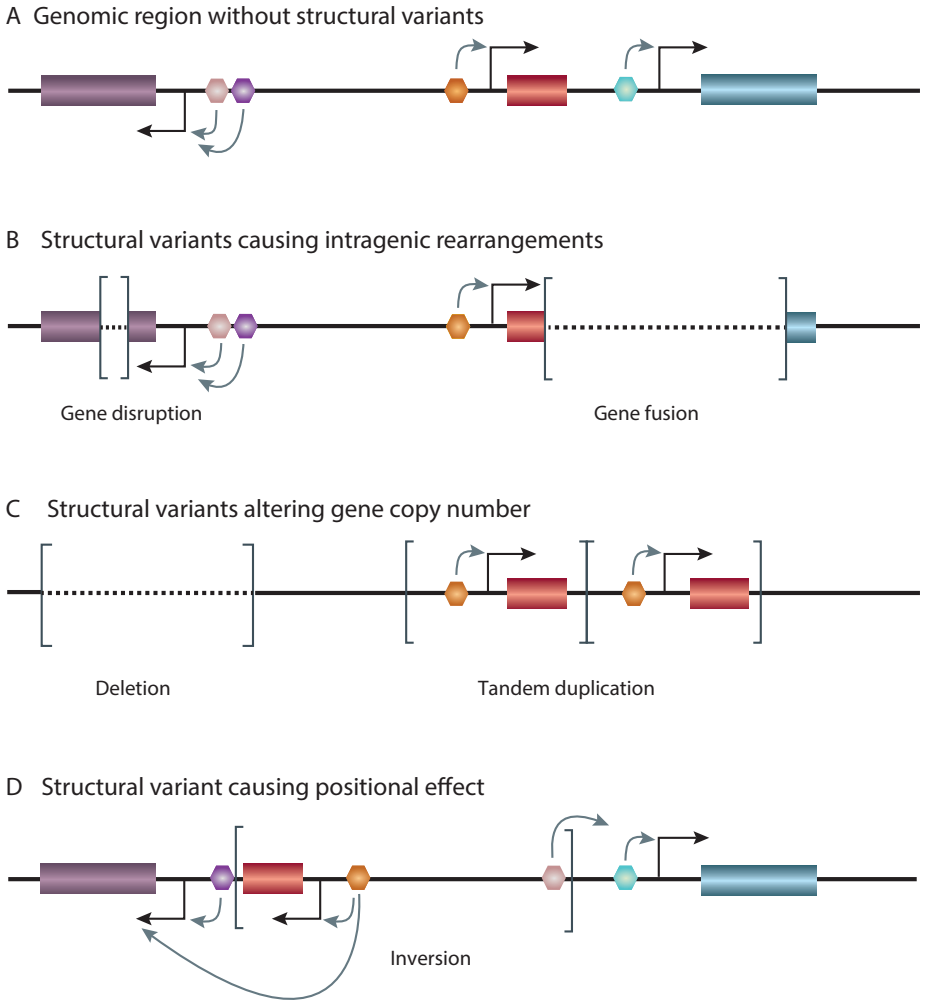

FIGURE 1.1: Functional consequences of structural variants. (A) Genes (boxes) are regulated by the collective and combinatorial input of regulatory elements (hexagons, different colors indicate tissue-specificity); (B)–(D) Structural variants (square brackets) can lead to various phenotypic consequences: SVs can alter gene coding regions by removing part of a gene or fusing different coding regions together (B). Deletions or duplications can result in different gene copy-number and cause gene dosage changes (C). Structural variants can have positional effect manifesting in altered gene expression (D). Figure adapted from [Weischenfeldt et al., 2013].

1.2 Detection of genomic variants

To understand genomic variants, it is of utmost importance to be able to reliably detect and characterize them. Within the past ten years, many experimental and computational approaches have been developed to identify variants of all sizes and complexities. Hybridization-based microarray approaches, single molecule analyses and next-generation sequencing methods will be described in more detail below, with a focus on next-generation sequencing (NGS) as the majority of the data I analyzed during my PhD was obtained with NGS.

Understanding the structure and location of structural variants traditionally required a visualization step at the single-molecule level. Fluorescent in situ hybridization (FISH) and spectral karyotyping allowed visualization of structural variants by microscopy [Feuk et al., 2006]. Although these methods allow inspection of microscopic SVs, their application is limited to particularly large structural differences (∼500 kb to 5 Mb) and are not suitable for high throughput population-scale analyses [Alkan et al., 2011].

1.2.1 Hybridization-based microarray approaches

For a long time microarray technologies have been used as a standard for CNV discovery and genotyping, primarily through array comparative genomic hybridization (array CGH) and SNP microarray approaches [Iafrate et al., 2004, Pinkel et al., 1998, Snijders et al., 2001]. Though the idea behind both of these approaches is similar and the way each molecular assay is performed differs.

Array CGH. Array CGH platforms are based on hybridization of typically two labeled samples (test sample and reference sample) onto set of long oligonucleotides. The ratio observed between the sample and the reference is taken as an inference of copy-number state [Picard et al., 2005]. In general, signal from at least three to roughly ten consecutive probes is used to detect CNVs. Currently, array CGH platforms are capable of detecting CNVs as small as 500 bp with relatively precise breakpoints allowing to identify variant specific sequence motifs [Alkan et al., 2011].

SNP microarray. In comparison to array CGH, SNP microarray platforms use only one sample per microarray, requiring subsequent signal intensity clustering of each probe in many samples [Cooper et al., 2008]. Another difference lies in the probe design, as every SNP probe takes advantage of known single nucleotide polymorphisms between two DNA sequences. Although SNP arrays have lower signal-to-noise ratio compared to array CGH, they are capable of differentiating alleles through B allele frequency (BAF) measure calculation [LaFramboise, 2009].

Advantages and limitations. Advantages of hybridization-based microarray approaches include their low cost and high throughput. For instance, array CGH have custom, high-probe-density arrays readily available, whereas SNP arrays make use of public SNP data, both providing an opportunity to detect CNVs in large data sets [Alkan et al., 2011]. Despite their widespread application, hybridization based approaches suffer from certain limitations, including possible cross-hybridizations of probes. Perhaps the most obvious limitation is their inability to identify balanced SVs, such as inversions and translocations. Additionally, even in the case of unbalanced SVs, such as duplications, the location and structure of the duplicated sequence cannot be determined [Weiss et al., 1999]. Both hybridization methods lack sensitivity and are in general limited in resolution [Forozan et al., 1997]. Lastly, arrays perform poorly in repeat-rich and duplicated regions due to the assumption that each location in the reference genome is diploid (which is not true in duplicated sequences) [Oostlander et al., 2004].

1.2.2 Next-generation sequencing

The first ever DNA sequencing method was developed by Frederick Sanger and his colleagues in 1977 [Sanger and Coulson, 1975, Sanger et al., 1977]. Sanger sequencing is based on the selective incorporation of chain-terminating dideoxynucleotides by DNA polymerase during in vitro DNA replication. Although it has been replaced by NGS in many large-scale projects, the Sanger method remains in wide use, primarily for smaller-scale projects and for obtaining long contiguous DNA sequence reads.

Today, NGS is the most widely used approach to identify structural variations. It was developed in order to complement hybridization-based approaches and ultimately even fully replace them as a platform for SV discovery and genotyping. During the past years, high demand for low cost and high-throughput sequencing have driven the development of various sequencing platforms now able to produce millions of sequences simultaneously in parallel. Throughout my PhD, I have worked with data generated solely by Illumina machines, as they are presently the most widely used platform. Therefore, when referring to methods for variant discovery, I will focus on Illumina technology. Still, it is important to remark that other platforms exist, including Pacific Biosciences, Ion-Torrent, Oxford Nanopore and others [Mardis, 2013, Shendure and Ji, 2008, Zhao et al., 2013].

The general Illumina NGS approach combines the rescue and capture of paired ends, massive sequencing, and a computational approach to map DNA reads onto a reference genome. Once double-stranded genomic DNA is isolated from cells or tissues and purified, it is sheared into ∼200 to ∼500 bp fragments [Campbell et al., 2008], followed by addition of the platform-supplied adapters (Figure 1.2). Single-stranded adapter-bound fragments are subsequently attached to the complementary adapters on the platform flowcell, and the DNA polymerase with unlabeled

nucleotides is added to initiate bridge amplification [Mardis, 2013]. This step is required to create local clusters of each DNA fragment. Importantly, the density of initially bound fragments has to be low enough to allow sparse cluster formation and subsequent signal recognition. At this point, the actual sequencing cycles begin in sequencing-by-synthesis manner. The primer binds to the single-stranded DNA fragment found in previously generated clusters and the DNA polymerase incorporates one of four nucleotides (A, C, G or T), complementary to the nucleotide in DNA fragment. Each base is labeled with a different fluorescent dye (fluorophore) and emits a unique signal. After laser excitation, the emitted fluorescence is captured, the incorporated base identified and the fluorophore cleaved. The last step allows the incorporated base to become unblocked and the synthesis (incorporation of another nucleotide) to proceed. With the current Illumina chemistry, the sequencing cycles are usually repeated 100 times, providing the sequence length of 100 bp. Each fragment can be sequenced from one side or from both resulting in either 100 bp single-end read, or 100 bp paired-end reads, respectively. For paired-end libraries, both paired-ends belong to 200-500 bp DNA fragments, resulting in an measure of distance between the two reads. The size of the whole fragment (sequenced 100 bp reads + distance between reads) is referred to as 'insert size' (Figure 1.2).

In order to detect variants using NGS data, various computational and bioinformatics approaches have been developed. The choice of methods depends largely on the variant type of interest and their size. The general idea of each strategy consists of mapping sample reads to the reference genome and identifying 'discordant', i.e. abnormal signatures or patterns suggestive of an SV.

Read-depth. Similar to arrays, read-depth approaches successfully detect unbalanced SVs: deletions and duplications. The general workflow consists of mapping the reads against the reference genome, followed by dividing the genome into windows and calculating the ratio between read-depth in each window and the average read-depth of the whole sample. The method assumes a random distribution in mapping depth of the sequenced sample and therefore results in significantly higher read-depth within duplicated regions [Bailey et al., 2002] or reduced read-depth in deleted regions. As mentioned before, read-depth approaches are not able to detect balanced SVs since translocations and inversions do not result in read-depth changes in comparison to neighboring regions. The resolution of this method relies on the window size chosen, but will nonetheless never reach nucleotide breakpoint precision.

Paired-reads. Paired-reads or read-pairs take advantage of the orientation and the span of sequenced reads and use this information to identify potential SVs [Korbel et al., 2007]. As described before, genomic DNA is sheared into fragments of a certain size (e.g. 500 bp) and the ends (100 bp at each end) are sequenced. After sequencing, the initial step consists of mapping the sample paired-reads to the reference genome and subsequently reporting any discordancy in the mapping which is inconsistent with the reference [Korbel et al., 2007, 2009a]. For instance, paired-reads that map further away in the reference genome indicate deletion in the sample

(the insert-size is larger than expected). Similarly, the insert-size that is smaller than expected would be indicative of insertion. Translocations are usually detected if one read maps to one chromosome and the other read to another chromosome. Any inconsistency in orientation upon mapping the reads to the reference can be used to predict inversions or tandem duplications. In case of a MEI, usually the cluster of so-called single-anchored reads can be found, where one read maps to the reference and the other is unmapped indicating novel insertion.

The advantage of using paired-reads in comparison to read-depth is the ability to detect all variant types. Although this approach provides more accuracy than read-depth method, the identified SVs usually also lack nucleotide breakpoint precision and the resolution of the SVs detected often depends on the expected insert-size.

Splitreads. The aim of the splitread approach is to define a breakpoint to the single basepair resolution. Upon mapping of the reads onto the reference genome, some of the reads will remain unmapped or single-anchored. Those reads can be 'split' in order to locally map parts of the read separately where one part of the read maps in a certain distance from the other part. For instance, if deletion in the sample occurred, there can be reads spanning the breakpoint in the sample. Once mapped in the splitread fashion onto the reference, those reads will be broken and parts will map further away defining the exact nucleotide breakpoints in the reference.

The splitread approach is essential for determining *de novo* SV formation mechanisms [Lam et al., 2010]. However, inferring the exact breakpoint can be computationally challenging due to the complexity of local realignments and also due to the frequent association of SVs with repeat sequence. Many algorithms providing splitread analysis first infer SVs using paired-read and subsequently add the splitread information [Rausch et al., 2012b, Ye et al., 2009].

Sequence assembly. To detect any genomic variant, the easiest way would be to assemble the sequenced genome i.e. to put the sequenced fragments together. In theory, given reads that are long and accurate enough, *de novo* sequence assembly would allow the reliable and precise definition of SVs. However, assembly approaches are usually limited to combining *de novo* and local assembly to generate longer contigs based on sequenced reads [Alkan et al., 2011]. These contigs can be compared and mapped to the reference genome to identify possible variants. All variant types can be inferred using assembly with the nucleotide breakpoint resolution, although the repetitive regions are extremely difficult to assemble. Due to the current computational challenges and the high cost associated with depth of sequencing needed for accurate assembly, this approach is still not widely used for SV detection, but it is important to note that with time it could become the most powerful method to detect genomic variants.

Advantages and limitations. NGS techonologies allow detection and characterization of different SV types. In general, sequencing based approaches are largely unbiased and using the combination of different methods helps to obtain the most comprehensive SV dataset. Each

separate approach has its own limitation and disadvantages depending on the variant type and size. For instance, read-depth is limited to unbalanced CNVs but at the same time this is the only method able to predict accurate copy-number. Furthermore, read-depth performs poorly when it comes to nucleotide breakpoint resolution identification and resolving ambiguous read mapping in the repetitive regions. Some examples of the tools developed for the read-depth analysis are: CNVnator [Abyzov et al., 2011], CopySeq [Waszak et al., 2010] and BICseq [Xi et al., 2010].

Paired-read mapping is currently widely used for detection of SVs as it is capable of identifying all SV types. There are many tools currently available that employ paired-read mapping information, such as DELLY [Rausch et al., 2012b], VariationHunter [Hormozdiari et al., 2010], BreakDancer [Chen et al., 2009] and GenomeSTRiP [Handsaker et al., 2011]. Despite its many advantages compared to read-depth, using paired-end read approaches still has problems with reliable breakpoint identification. Splitread approaches overcome this problem and successfully accurately describe breakpoints. However, it requires substantial computational power and it is only reliable in the unique regions of the genome. Pindel [Ye et al., 2009] is one of the tools providing such splitread identification.

Overall, the most promising method is certainly whole-genome sequencing (WGS) assembly as it should allow unbiased comparison between two independently assembled genomes. Still, it is inaccurate in repetitive and duplicated regions due to the presence of multiple identical sequences in said regions of the genome. Assembly is also extremely computationally expensive, sometimes even to the point that the whole process collapses [Alkan et al., 2011]. Well known *de novo* algorithms for whole-genome or local assembly include ABySS [Simpson et al., 2009], SOAPdenovo [Li et al., 2010], HYDRA [Quinlan et al., 2010] and TIGRA [Chen et al., 2014].

Many of the above mentioned tools and approaches are either completely incapable of detecting MEIs or have substantial issues in repetitive regions. Undeniably, identification of MEIs has always been hindered in NGS approaches. Due to being present at many different locations in the genome, MEIs can create ambiguities upon alignment and assembly, resulting in detection biases and errors [Treangen and Salzberg, 2011]. Many of the SV detection tools developed in the past completely ignored MEIs. Nevertheless, recent efforts have improved detection accuracy and new algorithms have been designed exclusively for MEI exploration. Some examples of such tools are TEA [Lee et al., 2012], Tangram [Wu et al., 2014], Mobster [Thung et al., 2014] and Retroseq [Keane et al., 2013].

With recent advancement in sequencing technology and the decline in sequencing costs, NGS technologies have become the standard choice among methods to detect variants in many laboratories. Although NGS based approaches have similarities with arrays, NGS provides better general accuracy. Both approaches require the presence of another genome, in order to be able

to compare the investigated sample to a standard. In the case of NGS this is usually a publicly available reference genome found in public databases (e.g. University of California Santa Cruz - UCSC contains reference sequences and working draft assemblies for a large collection of genomes; http://genome.ucsc.edu/), whereas for arrays it is typically an arbitrarily chosen sample. Additionally, hybridization-based approaches depend on probes that are designed based on sequences present in the reference assembly [Alkan et al., 2011]. In order to minimize all above listed limitations and to improve sensitivity and specificity, there are currently many tools that incorporate multiple methodologies being developed (e.g paired-reads with read-depth).

Next-generation sequencing applications. Besides whole-genome *de novo* sequencing, NGS technologies have found application in many other areas. Whole-exome sequencing (WES) is one of NGS adaptations used to detect variants exclusively in protein-coding regions of the genome [Majewski et al., 2011, Singleton, 2011]. In comparison to WGS, WES is applied on 1% of the human genome occupied by exons, resulting in significantly cheaper and faster throughput. However, WES can only identify smaller variants found in the coding region of genes which affect protein function, omitting larger SVs in non-coding regions. Those variants can be also associated with diseases and found using other methods such as WGS.

Another application of NGS is RNA sequencing (RNA-seq) or whole-transcriptome sequencing [Chu and Corey, 2012]. This approach is quite similar to WGS approaches. In general, a population of RNAs is isolated and converted into a complementary DNA (cDNA) library, which is then sequenced in the same fashion as DNA. RNA-Seq provides a far more precise measurement of levels of transcripts and their isoforms, compared to similar methods [Wang et al., 2009].

NGS can be adapted and used for other analyses such as ChIP-sequencing (ChIP-seq) [Deliard et al., 2013], various chromosome conformation capture assays such as 3C, 4C, 5C and Hi-C [de Wit and de Laat, 2012] and whole-genome bisulfite sequencing (methyl-seq) [Lou et al., 2014]. ChIPseq combines chromatin immunoprecipitation (ChIP) with massively parallel DNA sequencing to identify the binding sites of DNA-associated proteins while methyl-seq focuses on determining the methylation status of a DNA segment.

In the last five years there have been many advances in NGS technology and application. Although NGS design with complementary computational approaches as described above is currently widely used in the scientific community, sequence reads produced this way are sometimes too short to overcome genomic complexity and can create biases upon alignment to the reference. Quite recently, there have been long-read sequencing technologies developed that allow generation of reads longer than 5 kb. Pacific Biosciences's (PacBio) single-molecule real time sequencing (SMRT) technology [Eid et al., 2009] and Oxford Nanopore's MinION (general methodology described in Clarke et al. [2009]) each sequence single DNA (or RNA) molecules by synthesis and are often therefore called third generation sequencing approaches. Currently, these technologies

can only produce somewhat inaccurate sequences and require special algorithms for analysis. However, long-reads allow easier *de novo* assembly compared to short-read sequencing and help fill the gaps between the different technologies allowing the successful identification of SVs in repetitive regions. Taken all this into account, long-read technology has the potential to grow fast and become widely used in the future.

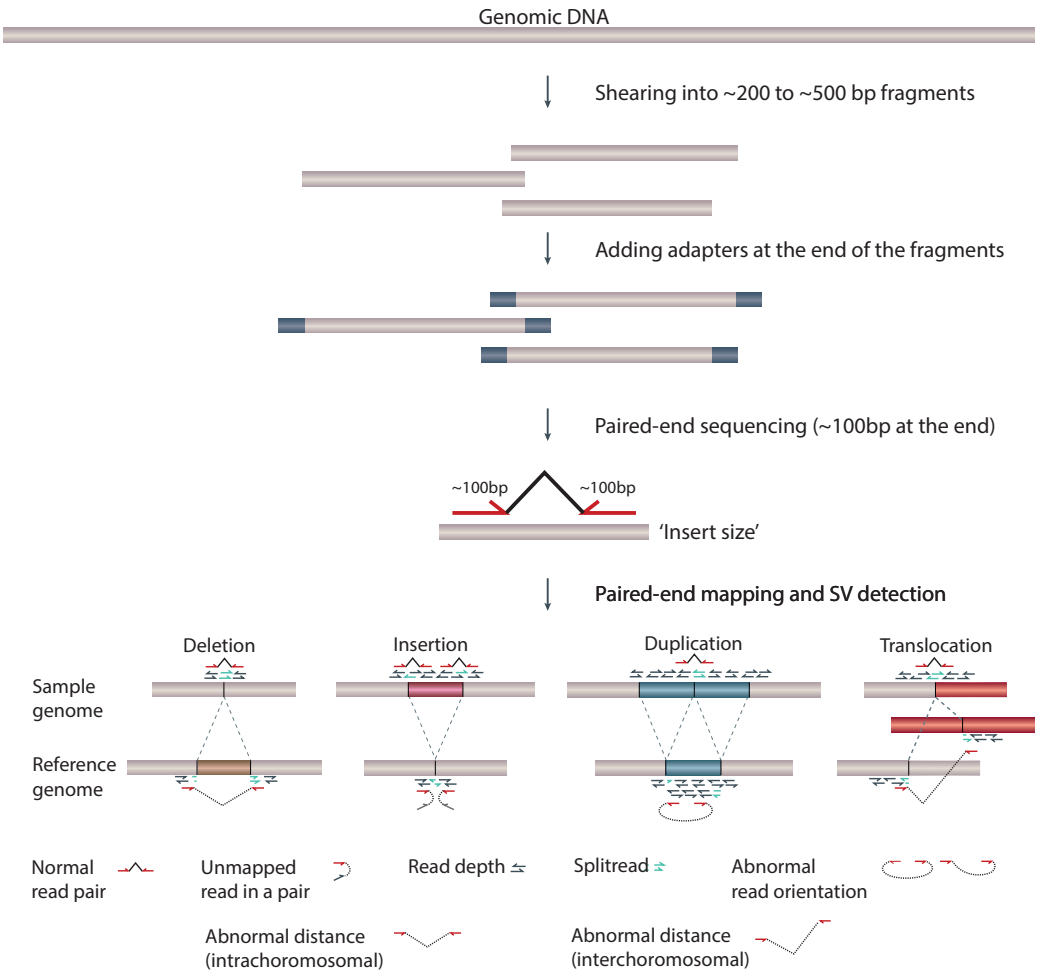

FIGURE 1.2: Structural variant detection and classes. The upper panel shows the preparation of genomic DNA for paired-end sequencing and subsequent SV detection consisting of 1. Shearing DNA into fragments of roughly equal size, 2. Adding adapters at the end of each fragment, 3. Sequencing ends of each fragment (here 100 bp) and 4. Paired-end mapping and SV detection. The lower panel represents different types of SVs and various ways to detect them. Structural variants comprise unbalanced copy-number variations $\geq$50 bp, including deletions and duplications, and variations such as translocations and insertions. Widely applied DNA-sequencing-based approaches for structural variant detection using the relative orientation, position and read depth of paired-end DNA sequencing reads are indicated. Figure adapted from [Weischenfeldt et al., 2013]

1.3 *De novo* structural variant formation mechanisms

Although most of the SVs are common in the population, *de novo* SV formation is believed to occur constantly in the germline. Recent studies have shown that SV formation mechanisms can be classified into four major groups: non-homologous rearrangements (NHR) associated with non-homologous end-joining (NHEJ) and replication-based mechanisms (microhomology-mediated break-induced replication (MMBIR) and fork stalling and template switching (FoSTeS)), mobile element insertions (MEI), variable number of tandem repeats (VNTR) and non-allelic homologous recombination (NAHR) events [Hastings et al., 2009b, Lam et al., 2010, Onishi-Seebacher and Korbel, 2011]. Advances in NGS analyses opened up the possibility to reliably predict precise SV breakpoints. In particular, splitread information enabled exploration of SV junction sequences and thus the inference of the mechanisms underlying SV formation [Lam et al., 2010]. Due to the focus of this Thesis, *de novo* SV formation mechanisms will be discussed below in context of ME-related mechanisms and mechanisms independent of ME.

1.3.1 SV formation mechanisms independent from mobile elements

As described above, there are several MEI-independent mechanisms that can lead to the emergence of SVs. In humans, roughly 28% of all SVs detectable by NGS approaches arise through homologous recombination, whereas non-homology-based mechanisms are responsible for almost half of all human SVs (∼45%). The remaining 27% occurs due to VNTRs (∼5%) and MEs [Kidd et al., 2010, Lam et al., 2010, Mills et al., 2011].

Non-allelic homologous recombination. NAHR involves homology-based recombination between two highly similar or identical sequences. When this occurs, sequences that lie between the repeats that recombine can be either duplicated or deleted, thus resulting in copy-number change (Figure 1.3). Given that the orientation of the recombining sequences is different, the resulting SV will be an inversion. Finally, translocation can arise if the segments that recombine come from two different chromosomes. The homologous sequences might be highly repetitive in the genome or occur only twice or a few times (i.e. low-copy repeats, LCRs, or segmental duplications, SDs) [Shaw and Lupski, 2005].

Replication-based mechanisms. During DNA replication, the replication fork can stall and switch templates using microhomology from a complementary template to continue the process [Zhang et al., 2009], resulting in a FoSTeS repair mechanism (Figure 1.3). As the whole model is replication-based, it is thought to occur during mitosis [Lee et al., 2007]. The involved forks can be separated by a range of linear distances, but in three-dimensional space they may be in a close physical proximity, resulting in all SVs types: deletions, duplications, inversions, translocations and even complex rearrangements (Figure 1.3). MMBIR is essentially a generalized form of

FoSTeS mechanisms following a replication fork collapse in cells under stress [Hastings et al., 2009a,b].

Non-homologous end joining. Aside from homology-based mechanisms, NHEJ is one of the major mechanisms used to repair double-strand DNA breaks (DSBs) in different organisms from bacteria to mammals [Gu et al., 2008, Lieber et al., 2003]. DSBs can be caused by molecular processes such as V(D)J recombination or by ionizing radiation and reactive oxygen species (ROS). In these cases, NHEJ mechanism machinery detects both broken DNA ends, modifies the ends to make them compatible and finally ligates them together (Figure 1.3). As a result of end processing, NHEJ can leave a 'repair scar' at the site of ligation [Lieber, 2008]. If the NHEJ repair is erroneous, deletions, duplications and tranlocations can arise.

Variable number of tandem repeats. VNTRs can be found in a genome where a short nucleotide sequence is organized as a tandem repeat [Brookes, 2013]. Deletions and duplications arise in VNTR-rich regions, due to expansion or contraction of simple tandem repeat units during recombination or replication (Figure 1.3) [Onishi-Seebacher and Korbel, 2011]. VNTR regions often show variations in number of repeats between individuals. This variation is inherited and therefore can serve for personal or parental identification by genetics and forensics.

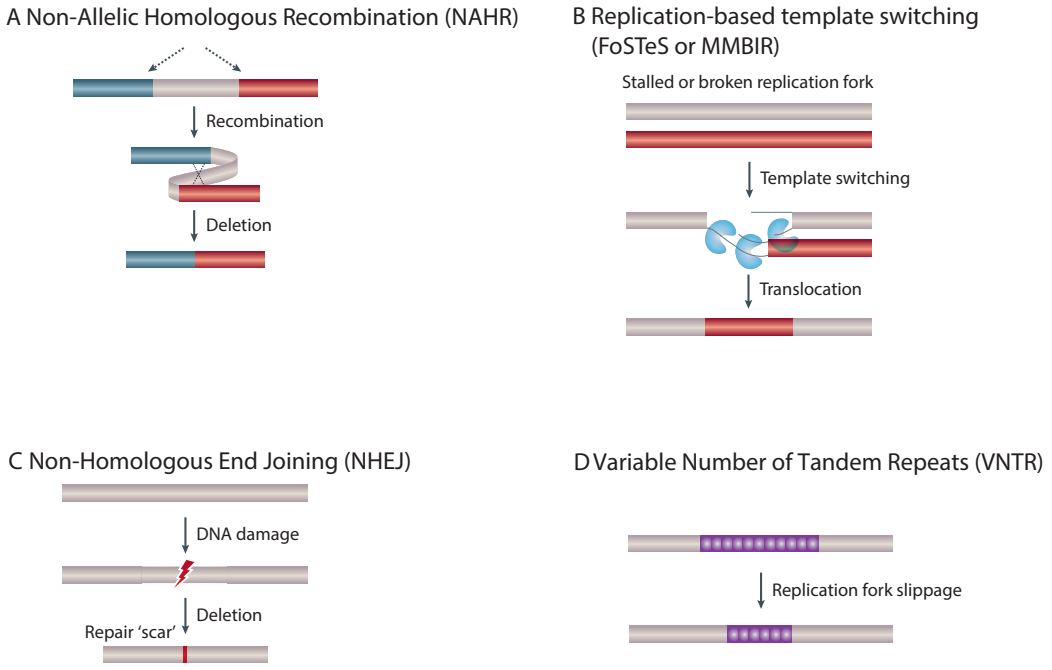

FIGURE 1.3: *De novo* SV formation mechanisms independent from mobile elements: Non-Allelic Homologous Recombination (NAHR), Replication-based mechanisms (FoSTeS/MMBIR), Non-Homologous End Joining (NHEJ) and Variable Number Tandem Repeats (VNTR). Figure adapted from [Weischenfeldt et al., 2013].

1.3.2 SV formation mechanisms induced by mobile elements

Mobile or transposable elements were first discovered in maize plants by Barbara McClintock in the 1940s [McClintock, 1956]. The identification of transposition demonstrated that genes can change their location within the genome and by doing so alter the gene's expression. Although the importance of transposons was recognized, it took roughly half a century for science to truly begin to understand how transposons behave [Cordaux and Batzer, 2009].

The completion of the first human genome sequence revealed that almost half of the human genome is composed of various transposable sequences [Lander et al., 2001], many of which are inactive ('fossilized') elements. Transposons have the ability to move within genome and can directly and indirectly cause the formation of SVs. Usually, transposons are separated into two major classes: RNA transposons or retrotransposons (class I) and DNA transposons (class II). DNA transposons constitute ∼3% of the human genome and move through a cut-and-paste mechanism by excising themselves and inserting elsewhere. They are not active in the human genome anymore, but were during early primate evolution roughly ∼37 million years (Myr) ago [Pace and Feschotte, 2007].

Retrotransposons (also referred to as MEIs) belong to an active class of transposons and move through an RNA intermediate via a copy-and-paste mechanism. The mechanism retrotransposons use to move within genome is referred to as target-primed reverse transcription mechanism (TPRT) and involves reverse-transcription of the retrotransposon messenger RNA (mRNA) into cDNA and final insertion of cDNA into a target chromosome [Malik et al., 1999]. During the process of ME insertion, TPRT produces short (∼15 bp) target site duplications (TSDs) at the flanks of the newly integrated retrotransposon (Figure 1.4 [Kojima, 2011]. Depending on the presence or absence of long-terminal repeats (LTRs), retrotransposons can be further split into two categories: LTR and non-LTR retrotransposons. Human LTR elements are known as endogenous retroviruses (ERV) and they occupy roughly ∼8% of the human genome. To date, no human ERV (HERV) has been identified as a cause for disease and therefore it is believed that ERVs have limited activity, if any [Mills et al., 2006]. The major contributor to the ME insertions and therefore overall transposon activity is the non-LTR retrotransposon class composed of long and short interspersed elements: *Alu*, SVA and L1 (long interspersed nuclear element, LINE1) [Stewart et al., 2011]. Although all three elements are active today in mammalian genomes, they differ significantly in their size and structure. Short interspersed nuclear elements (SINE) or *Alu*s are usually ∼300 bp in size, whereas SVAs represent SINE-VNTR-*Alu* composite elements and are ∼2 kb long [Ostertag et al., 2003]. Both *Alu* and SVA elements are non-autonomous elements and depend on the L1's molecular machinery to successfully mobilize throughout the genome. L1 elements are the longest of all non-LTR retrotransposons with ∼6 kb in full length. They contain two open-reading frames (ORFs), which encode proteins needed for retrotransposition: a RNA

binding protein, an endonuclease and a reverse-transcriptase [Cordaux and Batzer, 2009]. There are many non-LTR elements present in mammalian genomes. For instance, there are more than 500,000 L1 elements in the human genome, but less than 100 of them still remain active [Brouha et al., 2003]. *Alu* elements are the most successful in number of insertions with >1,000,000 copies [Lander et al., 2001], whereas SVAs are the youngest elements and have roughly 3,000 elements inserted into the human genome [Ostertag et al., 2003, Wang et al., 2005].

Impact of mobile elements on evolution. An important impact of retrotransposons on evolutionary dynamics is represented by the emergence of various subfamilies belonging to each class. For instance, *Alu* elements expansion started ∼65 million years ago (Mya) and during their continuous mobilization, 200 different subfamilies emerged [Price et al., 2004]. SVAs, being the youngest element, evolved during the ∼25 Mya of hominoid evolution and have only six existing subfamilies in mammalian genomes [Wang et al., 2005]. The diagnostic nucleotide substitutions and deletions or insertions defining a subfamily tend to accumulate hierarchically, indicating the existence of few 'master' elements involved in the retrotransposition [Batzer and Deininger, 2002]. It is estimated that the average human genome carries 80-100 active L1, six of which are known as 'hot L1' elements probably driving the whole retrotransposition process [Brouha et al., 2003, Seleme et al., 2006].

Due to their activity and accumulation over time, all retrotransposon classes have had major effects on primate genome evolution. Looking at the impact of MEs on human genome size, L1 and *Alu* elements have so far contributed ∼750 Mb to the whole genome [Lander et al., 2001]. Still, since retrotranspositon is an ongoing process occurring through a copy-and-paste mechanism, it continuously creates more genomic sequence. Although *Alu*, L1 and SVA have effect on evolution due to heritable retrotransposition in the germline, it is worth to note that this process takes place in somatic tissues as well. Retrotransposition-mediated somatic variations have been implicated in brain development [Baillie et al., 2011, Muotri et al., 2005], early embryonic development [Kano et al., 2009], cancer [Helman et al., 2014, Lee et al., 2012, Tubio et al., 2014] and other diseases [Deininger and Batzer, 1999], opening a variety of questions on ME behavior in context of somatic diseases.

Local genomic instability caused by mobile elements. MEs can facilitate genomic instability in many ways. By inserting into genes or gene regulatory regions, retrotransposons can not only alter protein function, but also influence genome evolution on various scales. There are numbers of heritable genetic disorders caused by direct ME insertions, some examples being hemophilia, cystic fibrosis, Apert syndrome, neurofibromatosis, Duchenne muscular dystrophy, β-thalassemia, hypercholesterolemia, and breast and colon cancers [Chen et al., 2005, Deininger and Batzer, 1999, Kazazian et al., 1988]. Other ways MEs can contribute to the genomic instability is by creating and repairing DNA DSB [Gasior et al., 2006, Morrish et al., 2002], promoting

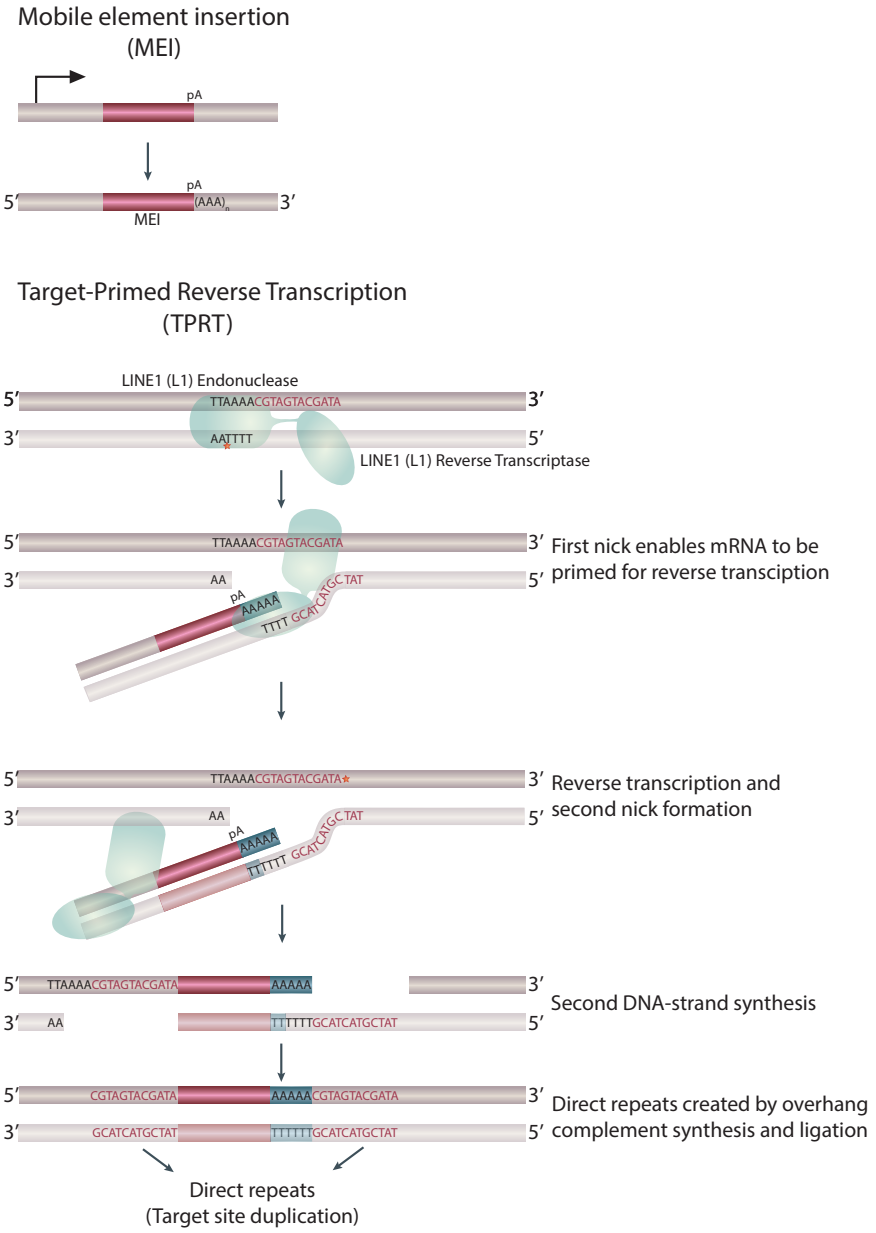

FIGURE 1.4: Target-primed reverse transcription mechanism (TPRT). Upper part shows mobile element insertion through copy-and-paste mechanism. Lower panel shows detailed molecular mechanism of target-primed reverse transcription. Mobile element insertion is mediated by the L1 endonuclease domain (green rectangle), which creates a first nick (red star) at the genomic site of insertion at the TTAAAA target sequence. L1 reverse transcriptase (green oval) uses this nick to prime the mRNA for reverse transcription (the parental mRNA serves as template), followed by second-strand nick generation and subsequent second DNA-strand synthesis. As a result of this process, duplication of the flanking sequence at the target occurs (target site duplication, TSD), which is one of the molecular signatures of retrotransposition. Another signature typical to ME insertion is polyadenylation tail emergence between TSD and ME at the 3' end of inserted element. Figure adapted from [Cordaux and Batzer, 2009, Kaessmann et al., 2009].

homology-mediated deletion (e.g. between two homologous ME sequences) or serving as a template for DNA repair, respectively. They can also serve as a source of microsatellites [Arcot et al., 1995] and undergo gene conversion by replacing older homologous elements with younger elements [Kass et al., 1995].

Genomic rearrangements as a result of mobile element activity. Upon insertion of L1 and *Alu* elements at new target loci, adjacent genomic sequences can sometimes be deleted. It has been shown that this process occurs naturally in the human and chimpanzee genomes and human-chimpanzee genome comparisons have detected one insertion-mediated deletion event which happened in the past ∼6 million years (Myr) and caused loss of a functional gene [Callinan et al., 2005, Han et al., 2005]. Another way MEs create genomic rearrangements is through recombination between non-allelic homologous elements at post-insertion. In pathological contexts, >70 *Alu* retrotransposon-mediated deletions (RMDs) have been reported, whereas only three disease-causing L1s are responsible for various cancers and genetic disorders [Cordaux and Batzer, 2009]. Despite having identified only 492 *Alu* RMD events and 73 L1 RMD events in the human genome that happened after human–chimpanzee divergence, these events have collectively removed nearly 1 Mb of genomic sequence, indicating their evolutionary importance [Cordaux, 2008, Han et al., 2008, Sen et al., 2006]. One interesting aspect of recombination between *Alu* elements is the origin and expansion of human SDs. SDs are large (>10 kb), nearly identical duplicated genomic regions. Since their boundaries are enriched in *Alu* elements, it is believed that they emerged through *Alu* recombination-mediated duplication [Bailey et al., 2003].

Additional to duplicating themselves, MEs are also capable of carrying neighboring genomic sequence and inserting it elsewhere in the genome. This process is known as retrotransposon-mediated transduction. During transcription, the RNA machinery sometimes skips a weak transcription termination signal - polyadenylation (polyA) signal, 5'-AATAAA-3' for L1 and SVA [Kaer and Speek, 2013] - located at the 3' end of a mobile element, thereby terminating the RNA synthesis at any downstream polyA signal. As a consequence, downstream flanking sequence is then mobilized together with the retrotransposon. Similarly, 5' transduction can occur if the retrotransposon is using an upstream 5' promoter, with the 5' sequence between the promoter and the retrotransposon getting transcribed and inserted elsewhere with the ME [Cordaux and Batzer, 2009]. This process can have an impact on various disorders and cancers [Solyom et al., 2012a, Tubio et al., 2014], as well as gene evolution, if the transduced sequence contains coding genes. Such example include multiple SVA-mediated acyl-malonyl condensing enzyme 1 (*AMAC1*) transductions, that have led to the formation of a new gene family during recent human evolution [Xing et al., 2006]. This whole gene transduction event happened after the divergence of African apes from orangutans but before the divergence of humans, chimpanzees, and gorillas, approximately ∼7 to 14 Mya. Recent findings show that transduced sequence can not only shuffle exons and genes, but also insert into a gene. For instance, a transduced sequence

insertion into the *dystrophin* gene can likely have a major role in Duchenne muscular dystrophy development [Solyom et al., 2012a]. Up to now, studies have shown that 3' transductions are a relatively frequent event in the human reference genome with ∼10% of L1 and SVA insertions being associated with 3' transduction events [Damert et al., 2009, Goodier et al., 2000, Hancks and Kazazian, 2012, Moran et al., 1999, Pickeral, 2000, Xing et al., 2006]. In humans, a few studies have looked into non-reference 3' L1-transductions in germline using capillary sequencing data [Kidd et al., 2010] and in disease contexts using NGS [Tubio et al., 2014]. The latter study explored possible exon-shuffling induced by cancer-specific transductions, which revealed the relevance of this form of variation, at least when occurring somatically in human disease.

In contrast to transduction, where MEs carry additional sequence to another location, the process known as gene retrotransposition uses retrotransposition machinery to duplicate gene sequences. As L1 elements encode proteins needed for retrotransposition, sometimes their machinery gets hijacked by host mRNA transcripts including *Alu* and SVA elements [Esnault et al., 2000]. Gene mRNA can then subsequently get inserted into another location as an intronless gene. To become functional, duplicated genes must acquire new regulatory regions in the target locus. Currently, it is widely accepted to call novel non-functional gene retrotransposed copies 'pseudogenes', and their functional counterparts 'retrogenes' or 'gene retrocopy insertion polymorphisms' (GRIPs). Similar to transductions, retrogenes have been important in the formation of new primate genes [Babushok et al., 2007, Kaessmann et al., 2009, Oliver and Greene, 2011] and it has been estimated that at least one new retrogene has emerged every million years in the human lineage over the past ∼65Myr [Marques et al., 2005]. An interesting example in evolution is the origin of the gene *TRIMCyp*, which arose when a L1 retrotransposon catalyzed the insertion of a cyclophilin A (*CypA*) cDNA into the *TRIM5* locus. In Old World Monkeys *TRIM5* blocks human immunodeficiency virus type 1 (HIV-1) infection, whereas in humans HIV-1 is protected by *CypA* binding to viral capsid. After the divergence of New and Old World monkeys, the retrotransposition of *CypA* and subsequent insertion into *TRIM5* occurred in owl monkeys, resulting in chimeric gene *TRIMCyp* which enables post-entry restriction of HIV-1 [Sayah et al., 2004]. Recently, due to advances in NGS, many tools have been developed to identify novel retrogenes in human, although some of the studies also looked into chimpanzee and mouse genome [Ewing et al., 2013, Schrider et al., 2013].

1.4 Primate evolution and genome differences

Non-human primates are important organisms for evolutionary studies due to their genetic and phenotypic similarities to humans. The first primates appeared around 65 Mya and humans diverged from this lineage 6 Mya, resulting in over 50 Myr of shared ancestry. Additionally, primates are extensively studied because of their diverse physiological and behavioral differences

as well as varied habitats. Most of the non-human primate species are endangered, presenting substantial challenges to primate research. Nevertheless, important results have emerged from studying primates, such as the development of yellow fever vaccine, the culturing of poliovirus resulting in a polio vaccine, and the significant discoveries regarding visual processing in the brain [Leader and Stark, 1987].

1.4.1 Evolutionary aspect of primate lineages

In evolution, primates are divided into two distinct monophyletic suborders: *Strepsirrhini*, or wet-nosed primates and *Haplorhini*, or dry-nosed primates consisting of tarsiers and simians. Simians are further split into geographically divided Old World and New World monkeys. New World monkeys (NWM) are found in Central and South America and portions of Mexico, whereas Old World monkeys (OWM) are native to Africa and Asia. apart from geographical separation, OWM and NWM differ in their physiological appearances and lifestyles. Due to the content of this Thesis, OWM will be discussed below in more details with specific focus on genomes of rhesus macaque and some of the great apes (orangutan and chimpanzee). Figure 1.5 represents evolutionary relationship and divergence of rhesus macaque, orangutan, chimpanzee and human lineages.

1.4.2 Properties of non-human primate genomes

Due to their similarity with humans, primates are especially valuable in biomedical studies such as neuroscience and various studies of infectious diseases and drug design [Conlee et al., 2004]. Non-human primates, such as marmosets, macaques, baboons and chimpanzees, whether wildtype or bred in captivity, are commonly used in such research. Sequencing of non-human primates opened another chapter in evolutionary studies, allowing scientists to look at differences between genomes on a basepair level. Following the initial draft sequence of the human genome in 2001 [Lander et al., 2001], WGS of the non-human primates genomes instantly became one of the highest priorities.

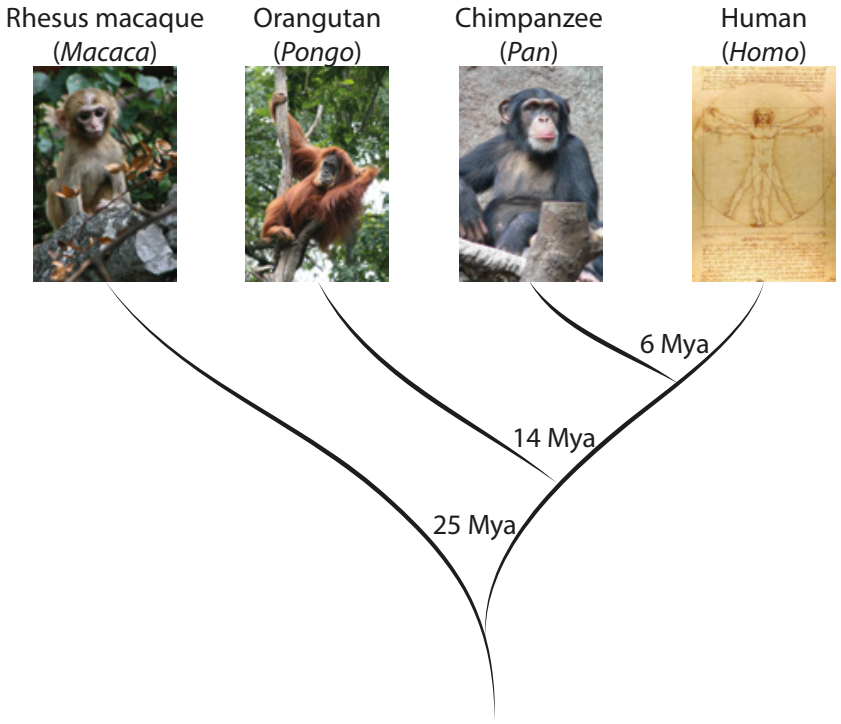

FIGURE 1.5: Evolutionary relationships of rhesus macaque, orangutan, chimpanzee and human represented as a reduced primate phylogenetic tree. Rhesus macaque diverged from the great apes/human common linage ∼25 Mya. Orangutan lineage split from human/chimpanzee 14 Mya, and finally human separation from the rest occurred 6 Mya. Photographs obtained from Wikipedia (http://www.wikipedia.org/) under the Creative Commons License.

Today, among other primates, the chimpanzee [The Chimpanzee Sequencing and Analysis Consortium, 2005], orangutan [Locke et al., 2011], Indian [Rhesus Macaque Genome Sequencing and Analysis Consortium, 2007] and Chinese [Yan et al., 2011] rhesus macaque genomes have been successfully sequenced. The comparisons between genomes revealed that many more basepair differences occurred due to indels and larger SVs (duplications, deletions, insertions, and bursts of retrotransposition events), rather than to SNVs [Marques-Bonet et al., 2009]. Specifically, sequencing of the chimpanzee genome (*Pan troglodytes*) highlighted that gene duplications are responsible for most differences between humans and chimpanzees observed on a genome level. Also, analysis of orangutan (*Pongo abelii*) and chimpanzee genomes showed that their smaller chromosomes 2A and 2B fused together in the human lineage to form human chromosome 2.

Repetitive elements in non-human primates are present in a similar percentage to the human genome and overall they might play an important role in the formation of large SVs formation [Marques-Bonet et al., 2009]. One such type would be SDs identified as duplications of homologous sequences ($\geq$90% identity) that can subsequently recombine and result in copy-number changes. Indeed, it has been shown that SDs in human and great apes tend to be larger, more

complex, and more interspersed [Bailey and Eichler, 2006], compared to other species. An interesting class of repetitive elements are retrotransposons, currently active in all primate genomes. Detailed inspection of the non-human primates reference genomes revealed relatively stable reference *Alu*, L1 and SVA numbers between species, with only a minority of MEs considered to be polymorphic. For instance, the orangutan genome has a dramatically lower number of lineage-specific *Alu* repeats compared to the chimpanzee genome, indicating different *Alu* insertion rates [Locke et al., 2011]. Gokcumen et al. [2013] have recently performed analyses on polymorphic MEs in non-human primates genomes and demonstrated previously reported *Alu* quiescence in orangutan genome, and additionally reported strikingly high *Alu* numbers in rhesus macaque genome compared to the great apes.

1.5 Human tumors and cancer

Tumor or neoplasm (Greek; *neo* new; *plasma* formation, creation) is an abnormal growth of tissue. [Cooper, 1992]. According to the International Statistical Classification of Diseases and Related Health Problems (ICD) of The World Health Organization (WHO), tumors can be classified as benign, malignant, *in situ* neoplasms, and tumors of uncertain or unknown behavior (who.int/classifications/icd10/ browse/2015/en#/II). Malignant tumors are typically referred to as cancer, a term that describes a large group of different diseases represented by uncontrolled growth of abnormal cells in the body. In order to become malignant, tumor cells adopt various properties through a multistep process involving somatic genetic variants. These 'hallmarks' of cancer include: evading cell death (apoptosis) and growth suppressors in general, inducing growth of new blood vessels (angiogenesis), initiating and allowing replicative immortality, activating metastasis and invasion and finally preserving proliferative signaling [Hanahan and Weinberg, 2000, 2011]. In contrast, benign tumors have a slower growth rate and do not metastasize or spread to other parts of the body [Cooper, 1992].

On a genomic level, cancer is essentially an alteration of growth and differentiation pathways transforming a normal cell into a malignant one. Two types of genes are usually affected: proto-oncogenes and tumor suppressors [Croce, 2008, Knudson, 2001]. Proto-oncogenes are growth- and survival-promoting genes, usually highly expressed or mutated in cancer. In contrast, tumor suppressor genes inhibit cell division and help with DNA repair, preventing cells from becoming malignant. If a tumor suppressor gene is lost or affected by one of the many mutation types, the resulting protein might exhibit a loss of function or even be completely absent from the cell. Importantly, while proto-oncogenes cause cancer when activated, tumor suppressors do so when inactivated. Unlike oncogenes, which require single allele mutations to become active, tumor suppressors usually undergo 'two-hit' mutations affecting both alleles, in order for full inactivation to take effect [Knudson, 1971].

Cancers arise as a consequence of somatic variant acquisition in the genome. Accumulation of different rearrangements, SVs and SNVs, in a formerly healthy genome cause normal cellular functions to be damaged resulting in the formation of malignant cells. One well-known example is the previously described (see 1.1.2 Large-scale variants section) fusion gene *BCR-Abl* that gains oncogenic property through translocation of the tyrosine kinase gene *Abl* from chromosome 9 to the break point cluster (*BCR*) gene on chromosome 22 [Nowell C. and Hungerford A., 1960]. Similar to natural variations in the germline, somatic alterations can encompass different variant types: copy-number changes such as gene deletions, duplications, amplifications, translocations and complex rearrangements, occurrence of small variants including nucleotide base substitutions in genes and regulatory regions. It has been reported that cancer cells can also acquire exogenous viral DNA which then leads to cancer development (e.g. human papilloma virus, Epstein Barr virus, hepatitis B virus) [Stratton et al., 2009, Talbot and Crawford, 2004].

Despite the technological advances in NGS and subsequent computational analyses, a need to develop new approaches still exists, especially when analyzing tumor data. Therefore, many large consortia dealing with these types of challenges have been formed in order to maximize the resources used and standardize the results. Following initial germline consortia such as the Human Genome Project (HGP) and the HapMap consortium, the International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA) have tried to characterize somatic variations occurring in various cancers. For a minor period of my PhD, I have contributed to the ICGC project involving by describing the occurrence of a novel SV formation mechanism (chromothripsis) in childhood brain tumors (medulloblastoma), which will be further discussed in the following sections.

1.5.1 Complex chromosomal alterations in cancer

In a genomic context, cancer is thought to be driven by somatically acquired point mutations and genomic rearrangements occurring in a progressive manner [Knudson, 1971, Nowell, 1976, Stratton et al., 2009]. This model suggests that tumorigenesis involves the progressive development of a tumor through multiple cycles of mutation and clonal expansion of the fittest cells ultimately leading to malignancy. However, there are examples of cancer development that are better described by a 'punctuated equilibrium' model rather than a 'progressive' one. This would involve bursts of somatic mutation in a short period of time. Recent studies have shown that some of the tumor genomes actually show a 'non-progressive' pattern whereby a chromosome seems to have been shattered and then reshuffled. The phenomenon, known as chromothripsis (Greek; *chromo* from chromosome; *thripsis*, for shattering) and it is thought to involve a single catastrophic event, rather than the progressive mutation acquisition of rearrangements (Figure

1.6). As a relatively novel mechanism, it is important to highlight the following about chromothripsis: (a) the mechanistic basis of this phenomenon is not fully understood, (b) there is a strong association of chromothripsis with poor prognosis (recently reported in several different malignancies:[Hirsch et al., 2012, Magrangeas et al., 2011, Molenaar et al., 2012, Rausch et al., 2012a]) and (c) chromothripsis occurs in many different cancer types, where it is thought to be crucial for cancer development (2-3% of all cancers [Stephens et al., 2011]). Since cancer genomes can acquire a large number of somatic DNA alterations, with dozens affecting a single chromosome in some cases, it is generally difficult to distinguish chromothripsis events from DNA alterations that occurred through a stepwise process.

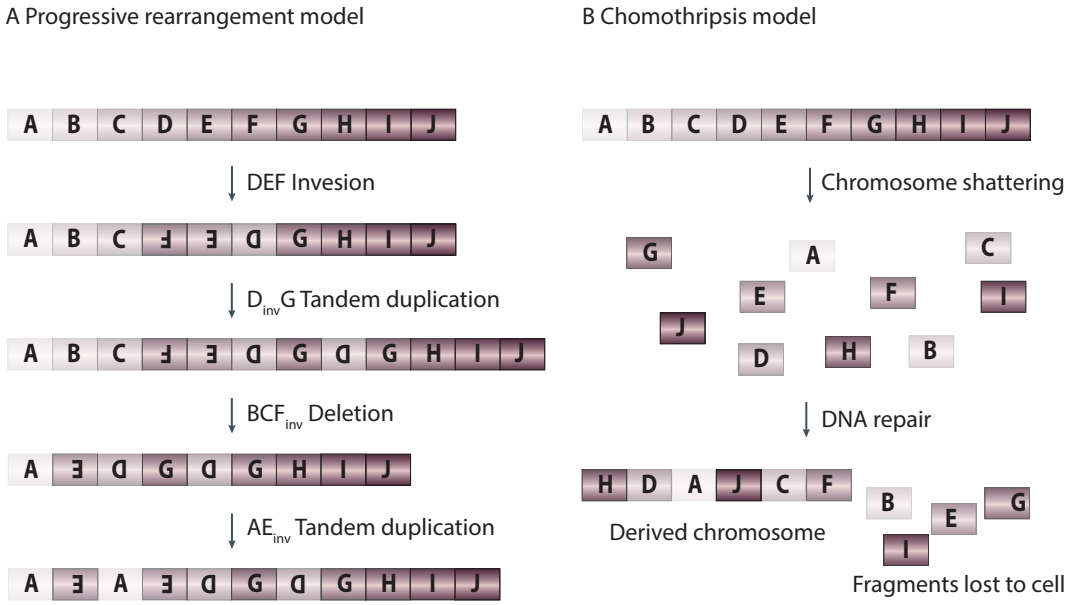

FIGURE 1.6: Difference between progressive cancer model and chromothripsis. Differently colored rectangles mark chromosomal segments that can be affected by structural variant (*inv* stands for inverted segment)(A) In progressive rearrangement model, mutations of different types are occurring in a stepwise fashion. (B) Chromothripsis induces shattering of usually one chromosome by DSB. DNA repair stitches some of the pieces together randomly, resulting in a derived chromosome. Other fragments are not included and subsequently they are lost to the cell. Figure adapted from [Stephens et al., 2011].

Korbel and Campbell [2013] described criteria to differentiate chromothripsis from a progressive model in an unbiased, statistically significant way. One criterion involves testing for the characteristic localized clustering of DNA breaks involved in chromothripsis. In contrast, stepwise alterations do not show a similar level of breakpoint clustering as chromothripsis. Another important difference when compared to the progressive model in chromothripsis genomes is the regularity of oscillating copy-number (CN) states, where alternations between only 2 or 3 CN states can be observed. Importantly, chromothripsis associated events usually affect a single parental chromosome, whereas stepwise alterations do not normally show such preference.

As a result of this single catastrophic event, chromosomal fragments are either lost or retained in a derivative chromosome created by the stochastic ligation of the remaining fragments. By comparison, progressive events are usually biased toward certain rearrangement forms, and thus will not show such random patterns of DNA segment order and fragment joining. Circular derivative chromosomes, known as 'double-minute chromosomes', can also be formed through this process and facilitate the amplification of oncogenes [Rausch et al., 2012a, Stephens et al., 2011]. Due to the variability of observed chromothripsis events in different cancer samples and tumor heterogeneity, implementing the aforementioned criteria for statistical assessment of chromothripsis does comes with certain challenges [Korbel and Campbell, 2013]. Though many formation mechanisms have been speculated, the clear mechanism and cause of chromothripsis has yet to be discovered.

1.5.2 Medulloblastoma susceptibility to chromothripsis

One of the cancer types that commonly harbors chromothripsis is medulloblastoma – a highly malignant pediatric brain tumor. It originates from the external granular layer cells of the cerebellum, and although it affects both children and adults, it is most common tumor in children. The survival prognosis is typically better for younger population, with 60%, 52%, and 47% survival rate at 5 years, 10 years, and 20 years, respectively [Smoll, 2012].

According to National Cancer Institute (NCI), several studies have split medulloblastoma into four molecular subtypes: (1) subtype 1 medulloblastoma with aberrations in the WNT signaling pathway, (2) subtype 2 medulloblastoma with aberrations in the Sonic-Hedgehog (SHH) pathway, (3) group 3 with presence of isochromosome 17q (abnormally long chr17, due to loss of short arm and duplication of long arm (i17q)) and *MYC* gene amplification, and (4) group 4 with *CDK6* and *MYCN* amplification [Kool et al., 2012, Northcott et al., 2012a, Taylor et al., 2012]. Medulloblastoma is a recognized Li-Fraumeni syndrome (LFS) tumor [Li and Fraumeni JR, 1969], which is linked to germline mutations of the *TP53* tumor-suppressor gene [Varley, 2003]. Aside from medulloblastoma, LFS malignancies include breast cancer, sarcoma and adrenal gland carcinomas.

Rausch et al. [2012a] have shown that chromothripsis is abundant in SHH subtype medulloblastomas with *TP53* mutation. This suggests a possible priming effect of certain genetic factors on chromothripsis that may shed additional light on the mechanistic basis of this unusual phenomenon, which appears to be crucial for the development of some aggressive cancers. Other than *TP53* [Malkin et al., 1990] which is related to chromothripsis, other genes whose functions are involved in SHH-medulloblastoma include *SUFU* [Taylor et al., 2002], *HIC1* and *PTCH1* [Briggs et al., 2008].

1.6 Motivation and background

Genomic SVs are defined as genetic polymorphisms leading to variation in structure of the genomic material. SVs are approximately larger than 50 bp, but they do vary in size and therefore can be microscopic and submicroscopic events, ranging from several kilobases up to few megabases. Known SV types involve CNVs, such as deletions and duplications, as well as inversions, insertions and translocations. In comparison to SNPs, they are less studied classes of genetic variation, even though the fraction of the genome affected by SVs is larger than that accounted by SNPs. As previously described, SVs have significant impact on phenotypic variation.[Mills et al., 2011].

Understanding structural variants was the main focus of my PhD research. Given the abundance of SVs in the genome, and given that widespread phenotypic effects have already been linked with SVs, it was my specific goal to understand mechanisms of SV formation in germline and in somatic tissues. My PhD work involved inferring the SV formation mechanisms in non-human primates and germline of the human genome. As a part of a collaboration with Charles Lee's group at Harvard Medical School, I have taken advantage of non-human primate DNA sequencing data that has been generated in the Korbel group to investigate SV formation mechanisms in these non-human species. One goal of this study was to investigate how formation mechanisms differ both in intra- and inter-species relations. Towards the end of my PhD, my focus shifted to the specific class of SVs - MEIs in both human and non-human primate species with specific interest in their ability to mobilize additional genomic sequences. Another part of my PhD research was to infer the formation of somatic SVs in cancer. When studying cancer on a genetic level, faults in two types of genes are especially important: oncogenes, which can drive the growth of cancer cells, and tumor-suppressor genes, which can prevent cancer from developing. Somatic structural variations can give rise to a cancer by affecting these genes. Therefore by studying differences of SVs between healthy and cancerous tissue one would be able to better understand tumorigenesis and the disease itself.

Chapter 2 describes MEI distributions in non-human primates, specifically in *Pan troglodytes* (chimpanzee), *Pongo abelii* (orangutan), and *Macaca mulatta* (rhesus macaque) [Gokcumen et al., 2013] with more focus on previously uncharacterized MEIs and their differences observed in non-human primates. In addition to the SV maps generated in each species (Chapter 4 [Gokcumen et al., 2013]), we generated a comprehensive MEI datasets in chimpanzee, orangutan and rhesus macaque consisting of: (1) reference-derived polymorphic MEIs (using BreakSeq [Lam et al., 2010]), (2) reference-derived species-specific fixed MEIs (using similar approach as Mills et al. [2007]), (3) novel, non-reference MEIs (using TEA [Lee et al., 2012]). Compared to the great apes, we discovered a notable excess of *Alu* activity in rhesus macaque, with *AluMacYa3* being the most dominating MEI subfamily. In the great apes we studied, the L1Pt family in

chimpanzees and the L1PA3 family in orangutans surpassed *Alu* elements, showing that the polymorphic L1 elements dominate the respective MEI landscapes.

Chapter 3 describes further rearrangements caused by mobile elements. The main focus of the study presented here will be on active L1 elements with the ability to mobilize 3' flanking DNA to different genomic loci. I combined two independent translocation [Rausch et al., 2012b] and L1 discovery pipelines [Lee et al., 2012] to create a novel computational methodology, termed TIGER (Transductions In GERmline) for the discovery of polymorphic L1-mediated 3' transductions. Several studies focused on fixed 3' transduced sequences in the human reference genome, reported that 3' transduction is relatively frequent [Damert et al., 2009, Goodier et al., 2000, Hancks and Kazazian, 2012, Moran et al., 1999, Pickeral, 2000, Xing et al., 2006]. In contrast, our results generated by TIGER identify significant differences in L1-mediated 3' transduction rates across non-human primate species and indicate species-specific L1 subtypes involved in this process.

Chapter 4 focuses on SV formation in non-human primates described in Chapter 2. Our main aim in this study was to build comprehensive SV maps in aforementioned species, in order to explore different SV landscapes and obtain a deeper evolutionary insight. We performed massively parallel sequencing of fibroblast-derived genomic DNA from five unrelated chimpanzee, orangutan, and rhesus macaque individuals to generate deletion and duplication datasets. Using BreakSeq [Lam et al., 2010], I performed *de novo* formation mechanism analysis on each SV map to describe differences in SVs observed between each species. Our results indicated a marked increase of NAHR-mediated SVs in orangutans and chimpanzees, whereas in rhesus macaque we observed dominance of MEI-related mechanism (described in Chapter 2).

Chapter 5 describes SV formation mechanisms in the human germline and in somatic tissues. I adapted and used the BreakSeq software [Lam et al., 2010] in order to infer *de novo* formation mechanisms in previously published structural variation datasets: Mills et al. [2011], Conrad et al. [2010], Kidd et al. [2010] and Lam et al. [2010], as well as to put them in a relation to SV formation mechanisms we observed in childhood brain tumor, medulloblastoma [Rausch et al., 2012a]. In brief, our analyses of SVs detected in germline are consistent with previous findings, which indicate that almost half of human deletions form through NHR mechanism involving NHEJ or MMBIR repair, with the rest forming through VNTR, MEI and NAHR related mechanisms. In contrast to the germline study, rearrangement breakpoints we observed in medulloblastoma support a model of massive DNA double strand breaks [Stephens et al., 2011], followed by NHEJ-mediated repair.

Chapter 6 summarizes studies presented in this Thesis and provides conclusions and future outlook of genomics involving variant discovery. In the remaining part, all supplementary data and detailed methods of the corresponding chapters are presented in a form of appendices:

Appendix A focuses on supplementary figures and tables and Appendix B on detailed methods for each chapter. Appendix C outlines scientific publications including me as one of the authors.

The work presented throughout this Thesis is mostly a collaborative effort involving many sides that contributed with analyses, feedback, ideas and support. Therefore, before each chapter, I indicated the input of collaborators involved in presented studies, as well as my personal contribution.

Chapter 2

Mobile element insertion landscape in non-human primates

Retrotransposon datasets in non-human primates will be a focus of this chapter. For all reference-derived retrotransposons, the size and subfamily of each mobile element was determined, whereas for the non-reference (novel) mobile elements insertions, such analysis was not possible at the time, due to algorithm limitations. The results presented throughout this chapter are partially unpublished, whereas the rest were reported in the following publication:

> Gokcumen O.*, Tischler V.*, Tica J., Zhu Q., Iskow R. C., Lee E., Fritz M. H.-Y., Langdon A., Stütz A. M., Pavlidis P. et al. Primate genome architecture influences structural variation mechanisms and functional consequences. *Proceedings of the National Academy of Sciences of the United States of America*, 110(39):15764-9, September 2013.

Contribution

I performed the generation and subsequent analyses of all reference-derived mobile element maps, as well as analyses on novel, non-reference mobile element lists provided by Eunjung Lee and Peter Park. Validation of species-specific retrotransposons was designed and performed by Rebecca Iskow. The published part of this study was a collaboration between our group and Charles Lee's group at Harvard Medical School in Boston and both Charles Lee and Jan Korbel provided a significant feedback and supervised the analyses. The unpublished part was supervised by Jan Korbel, who together with Rebecca Iskow, Omer Gokcumen and Verena Tischler contributed with numerous discussions and general feedback.

2.1 Motivation and background

Due to their markedly high abundance in mammalian genomes, retrotransposons or mobile element insertions (MEIs) are an especially interesting group of large variants. Nearly half of the human genome is derived from transposable elements [Lander et al., 2001], but the vast majority of these elements are fixed in the population (i.e., present in all individuals of a species and not polymorphic) and inactive (i.e., incapable of creating new insertions) [Mills et al., 2007]. *Alu*, L1, and SVA families, representing a subset of retrotransposons capable of spawning new insertions, tend to be polymorphic in the population [Iskow et al., 2010]. MEIs can affect genes and their function directly by disrupting an exon and hence changing the protein sequence. They can also disable the gene indirectly by altering its expression levels through regulatory element disruption. Ultimately, both scenarios can result in genetic disorders and therefore retrotransposons can be considered as endogenous insertional mutagens.

Although many studies looked into retrotransposons in the human genome, up to date far less is known about MEIs within non-human primate genomes. Recent studies have shown that there is a reduction of *Alu* retrotransposition in orangutans, which implies a limited MEI threat to the genome [Locke et al., 2011, Walker et al., 2012]. However, the overall extent of different retrotransposon classes on non-human primate genomes was never investigated in depth, due to the lack of corresponding reference genomes or adequate tools to analyze such data.

In this chapter I will present three comprehensive MEI datasets in chimpanzee, orangutan and rhesus macaque: (1) reference-derived polymorphic MEIs, (2) reference-derived species-specific fixed MEIs, (3) novel (non-reference) MEIs (Appendix A, Figure A.2). Previously undetected retrotransposon polymorphisms and their genomic features will be a main focus of the chapter, together with different methods used to identify MEIs in non-human primate species.

2.2 Polymorphic MEI distribution in non-human primates and human

In order to construct SV maps in non-human primate genomes, we performed massively parallel sequencing of fibroblast-derived genomic DNA from five unrelated chimpanzee (*Pan troglodytes*), orangutan (*Pongo abelii*), and rhesus macaque (*Macaca mulatta*) individuals (described in details in Chapter 4). Using two recently developed computational methods, we identified polymorphic retrotransposon insertions that are (1) present in the sample genome, but absent from its respective reference genome (non-reference MEIs) or, (2) present in the reference genome, but absent in one or more of the samples for that species (reference MEIs). For the former approach, we

used the TEA algorithm [Lee et al., 2012] (for more details see Chapter 4 Methods) and successfully mapped 764, 2,548, and 15,566 non-reference MEIs in chimpanzee, orangutan, and rhesus macaque, respectively. To analyze non-insertion polymorphisms we used the BreakSeq algorithm which internally overlaps deletion and duplication predictions with known retrotransposons [Lam et al., 2010] and identified 90, 315 and 1,124 reference-derived MEIs in chimpanzee, orangutan, and rhesus macaque, respectively (Figure 2.1, for more details see Chapter 4 Methods) .

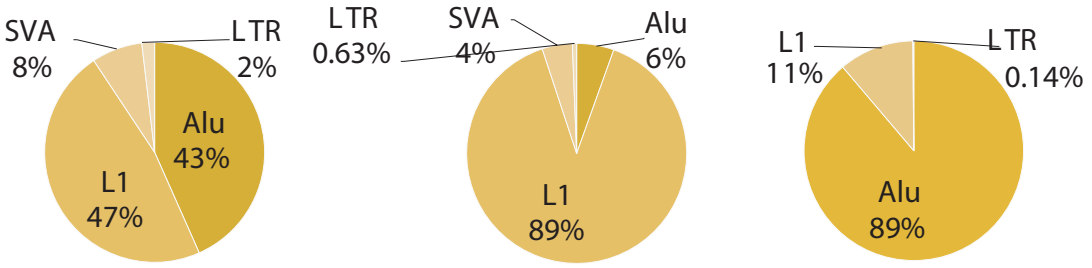

FIGURE 2.1: Breakdown of MEIs identified as reference or non-reference transposable element insertions. LTR, endogenous retrovirus-associated long terminal repeats; SVA, SINE–VNTR–*Alu* composite mobile elements.

In the great apes we studied, the relative abundance of polymorphic L1 elements surpassed *Alu* elements, with the L1Pt family in chimpanzees and the L1PA3 family in orangutans dominating the respective MEI landscapes, whereas the *AluMacYa3* was shown to be the most dominating MEI subfamily in macaques (subfamily assignments based on reference MEIs; Figure 2.1 and 2.2). Indeed, analysis of both reference-derived and novel MEIs showed a markedly higher *Alu* activity in macaques as opposed to great apes (P<2.2x10^{-16}; two-sided Fisher's exact test). This ultimately led to a pronounced increase of small SVs in macaques corresponding to the size of *Alu* elements - ∼300 bp (See Appendix A, Figure A.3, Figure A.16). Polymorphic *Alu* insertions were found at a proportionally lower rate in orangutans compared with chimpanzees (from 43% of all MEIs in chimpanzees to 6% in orangutans; P<2.6 x 10^{-100}, two-sided Fisher's exact test; Figure 2.1).

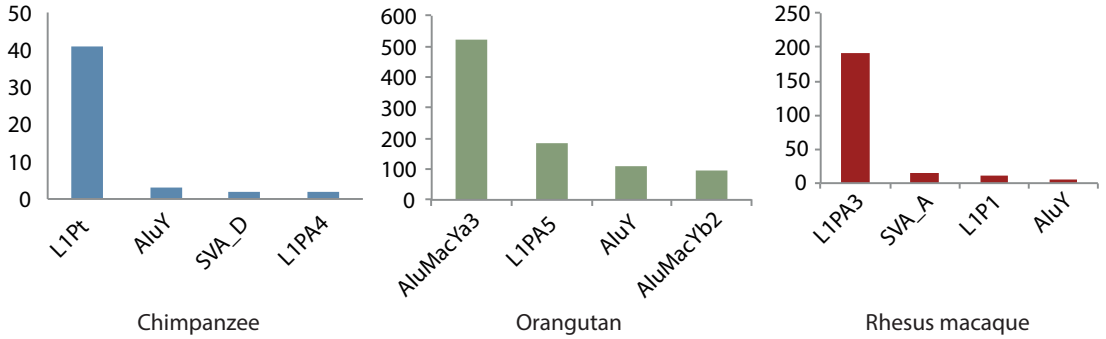

FIGURE 2.2: Four most abundant MEI subfamilies in each species detected as polymorphic reference-derived MEIs.

Under the neutral theory of molecular evolution, the rate of evolutionary change in genomes is largely determined by the mutation rate [Khaitovich et al., 2006]. Many of these changes are neutral and accumulate over time with constant rates. As they are not under any type of selection, most of them are not responsible for effect on phenotype.

Since we observed strong differences in numbers of SVs mediated by MEI process between rhesus macaque and great apes, we wanted to address if they form under the constant rate in each species. Under the assumption that SNP and SV mutation rates are approximately similar across primate species, the number of observed SNPs and SVs should correlate. Indeed, a strong correlation between the number of SNPs and the number of L1 events was observed (r^2 value = 0.76; Figure 2.3), whereas weaker or no correlation was detected between SNPs and *Alu* element insertions ($r^2 = 0.45$). This finding further supports the notion that *Alu* and non-allelic homologous recombination (described in Chapter 4) formation rates have changed considerably in recent primate evolution (Figure 2.3).

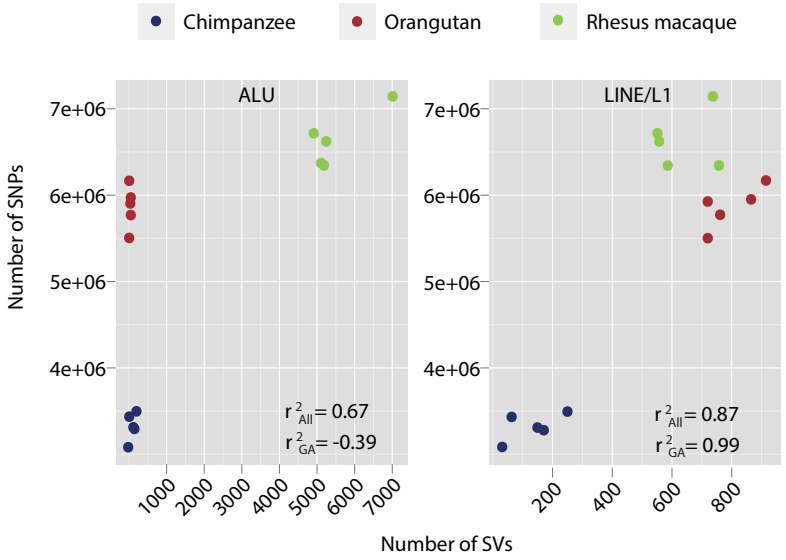

FIGURE 2.3: Correlation in the abundance of SNPs and MEIs. Dots represent different samples. r^2_{All} = Pearson correlation coefficient for all three studied primate species; r^2_{GA} = Pearson correlation coefficient for studied great ape species.

2.3 Species-specific fixed MEIs

Apart from investigating polymorphic MEIs, we also inspected fixed reference mobile elements in human and non-human primate genomes. When looking at the non-human primate reference genomes, numbers of MEIs present in the reference genomes are relatively stable throughout the primate tree with 1,000,000 *Alu* elements, 900,000 L1 copies and 4,000 SVA elements per genome (Appendix A, Figure A.1). However, many of those elements were present in the ancestral genome and are therefore shared between human, chimpanzee, orangutan and rhesus macaque. In order to investigate reference elements exclusively present in a single species, we decided to adopt an approach to detect species-specific MEIs, described in Mills et al. [2007]. In brief, pairwise whole-genome alignments between human, chimpanzee, gorilla, orangutan and rhesus macaque (Figure 2.4 A) were used in order to obtain species-specific MEIs (see Methods for details). The gorilla reference genome was added to the analysis to improve specificity of each identified MEI. To recover a MEI differentially present in the two genomes, we looked for alignment gaps present in one genome and a MEI in another genome (Figure 2.4 B). By combining all alignments of one species versus all others (e.g. for human as a query species: human-chimpanzee, human-gorilla, human-orangutan and human-rhesus macaque) and subsequently taking all the query-specific

MEIs where other species exhibited gaps in the alignment, we successfully derived a species-specific MEI lists.

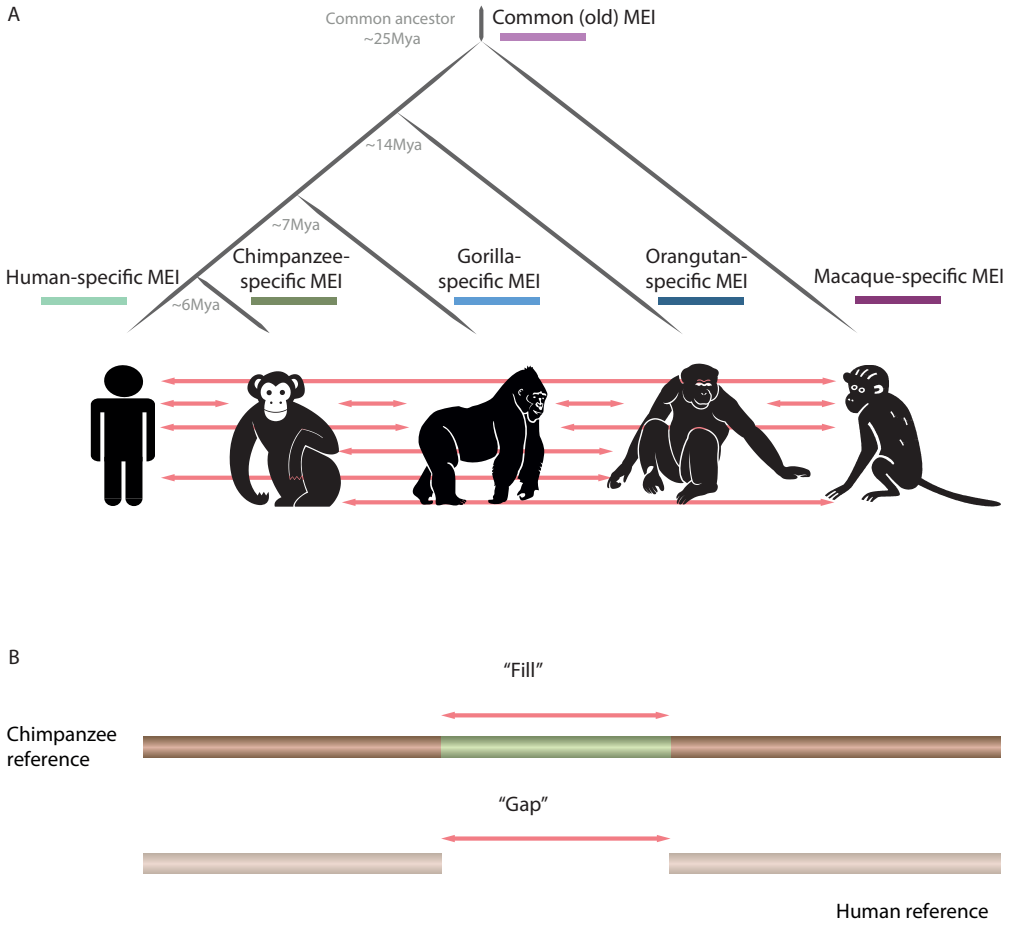

FIGURE 2.4: Overview of fixed species-specific MEI discovery pipeline. (A) To delineate species-specific MEIs, pairwise whole-genome reference alignments were performed with all possible combinations (e.g. for human as a query sequence: human-chimpanzee, human-gorilla, human-orangutan and human-rhesus macaque pairwise alignments were taken into consideration). (B) In loci where one species exhibits alignment gap (e.g. chimpanzee) and the other has 'fill' sequence (e.g. human) this is considered to be species-specific sequence, subsequently checked if it overlaps a MEI, indicating a species-specific MEI (in this case, human-specific MEI).

In human we identified 5,903 human-specific *Alu* elements, 1,641 L1 elements and 583 SVA elements. Great apes have similar numbers of species-specific *Alu* elements with 2,245, 2,212 and 1,886 *Alu* events in chimpanzee, gorilla and orangutan, respectively. The main difference in great apes comes from L1 elements, with 6,347 L1 elements in orangutan, compared to 1,598 and 1,399 L1 events in gorilla and chimpanzee. Orangutans have the highest count of species-specific SVA elements (354), whereas gorilla and chimpanzee have 298 and 216 species-specific SVA elements. In rhesus macaque, we again observed dominance of *Alu* elements with 55,941 macaque-specific *Alu* events compared to 11,010 L1 events.

To test our approach, we used chimpanzee-specific MEIs, detected exclusively from chimpanzee-human pairwise alignment and compared it to the Mills et al. [2007] dataset, which was generated using a similar approach (Appendix A, Figure A.5 B). We successfully recovered 78% of the Mills et al. [2007] chimpanzee-specific MEIs, whereas the remaining 22% were probably undetected due to discrepancies between reference builds used and the inability to recover all coordinates (Mills et al. [2007] used *panTro1-hg17* for the pairwise alignment, whereas we used *panTro3-hg19*). To further curate each mobile element, we assessed a range of MEI diagnostic features, including delineating the MEI target site duplication (TSD), poly-A tail, the MEI length, 5' truncations, if present, and 3' transduction (see Methods and Appendix A, Table A.1 and Figure A.4). TSD values we observed were consistent with previous reports [Dewannieux et al., 2003, Lee et al., 2012]. We also performed experimental validations using primers specific to the pairwise aligned sequence (left and right of the predicted species-specific MEI), to confirm presence of species-specific MEIs in one species and absence in another (Appendix A, Figure A.5 A). Out of five tested loci, all five were successfully validated (FDR=0%).

As indicated before, differentially present retrotransposons exist in one species' reference genome, but are absent from another closely related species. Some of the differentially present elements may be polymorphic, but it is likely that most have reached fixation. Together with polymorphic retrotransposons, the combined dataset is referred to as 'recent' retrotransposition events. The breakdown of species-specific fixed elements follows the same distribution of *Alu*, L1 and SVA elements observed in polymorphic MEIs, with quiescence of *Alu* elements in the great apes and dominant *Alu* activity in rhesus macaque. In contrast, we show that lineage-specific MEI calls, present in the species of interest (query species), but also in at least one other species (excluding species-specific MEIs present in query species only), follow a completely different distribution (pairwise comparisons between lineage-specific and species-specific sets per species: $P<0.05$, Pearson's Chi-squared test; Figure 2.5). As one element is shared between two genomes, it is very likely that it is also present in the ancestral genome, although it can be subsequently deleted in some populations. The lineage-specific MEI distributions, therefore, differ across species due to their fixation in the ancestral genome and subsequent lineage sharing of these MEIs after divergence.

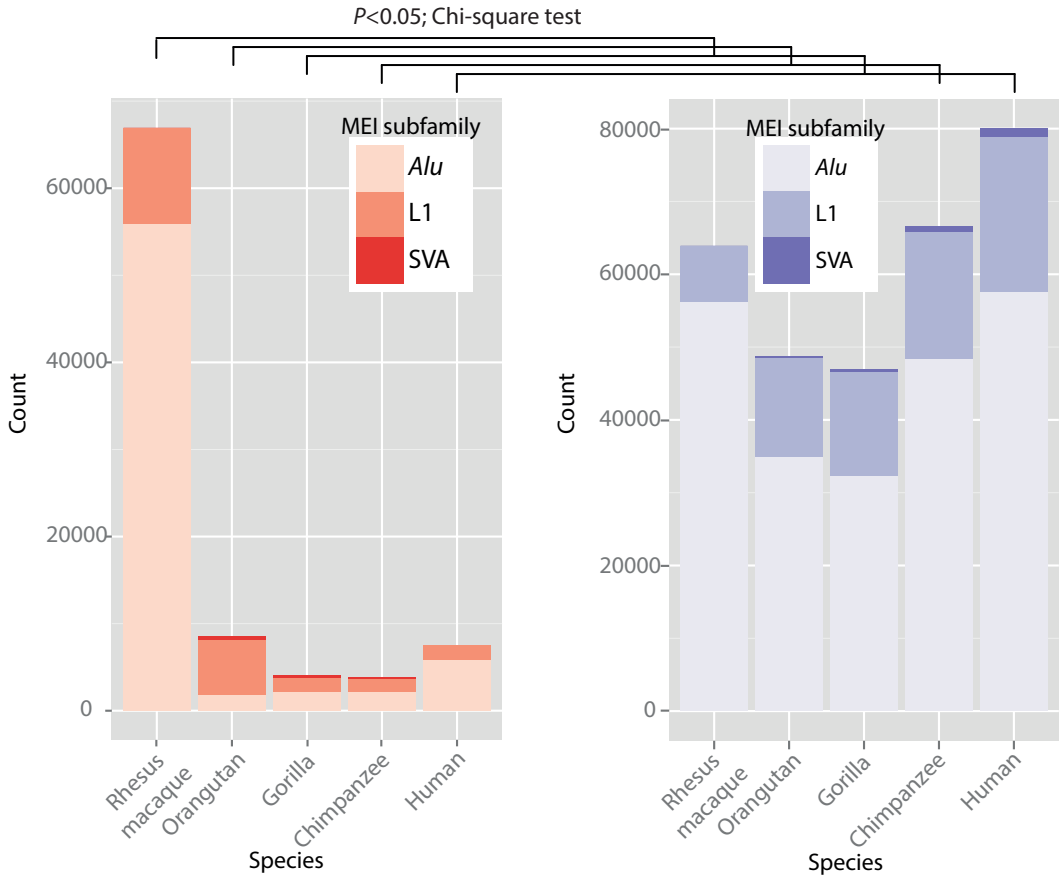

FIGURE 2.5: Species-specific (recent) fixed MEI and lineage-specific (ancestral) MEI distributions in human, chimpanzee, gorilla, orangutan and human lineages. $P<0.05$, Pearson's Chi-squared test for each pairwise comparison between species-specific and lineage specific elements.

2.4 Discussion

Reference genome assemblies of the chimpanzee [The Chimpanzee Sequencing and Analysis Consortium, 2005], orangutan [Locke et al., 2011] and rhesus macaque [Rhesus Macaque Genome Sequencing and Analysis Consortium, 2007], together with human reference [Lander et al., 2001] have provided valuable resources for variant discovery and annotation. Since all mammalian genome have a high content of mobile transposable elements, in-depth inspection of MEIs in primates was needed to comprehensively understand their impact on genomic landscapes. In this study we successfully identified fixed and polymophic MEIs in several non-human primates.

Our comprehensive dataset consists of: (1) polymorphic reference, (2) polymorphic non-reference and (3) fixed species-specific MEIs detected in chimpanzee, orangutan and rhesus macaque. By

identifying MEIs in non-human primates, we observed a notable excess of *Alu* activity in rhesus macaque compared with the great apes. Similar to the great apes, about 15% of all SVs in human are formed by MEI-related mechanism [Mills et al., 2011], indicating a similar rate of MEI insertions in great apes and human lineages. Since *Alu* retrotransposition represents the most active of human mobile elements [Mills et al., 2011, Stewart et al., 2011], our findings suggest a rapid turnover of active transposable DNA sequences, leading to a divergent set of species-specific MEIs. Together with SVs formed by non-allelic homologous recombination (discussed in Chapter 4), we speculate that fixed species-specific MEIs will likely further accumulate differentially in great ape and Old World monkey lineages, thereby promoting additional diversification in those lineages.

Chapter 3

Novel L1-mediated 3' sequence transductions

In this chapter, further rearrangements caused by mobile elements will be discussed. Particular focus will be on active L1 elements with the ability to mobilize 3' flanking DNA to different genomic loci. By combining translocation and L1 discovery pipelines a novel computational methodology, termed TIGER, was developed for the discovery of polymorphic L1-mediated 3' transductions. The methodology and results presented throughout this chapter are a part of the following manuscript in preparation:

> Tica J., Lee E., Untergasser A., Gokcumen O., Park P. J., Stütz A. M.*, Korbel J. O.* TIGER: Detection of L1-mediated 3' Transductions In GERmline using next-generation sequencing data (*Manuscript in preparation*)

Contribution

I performed generation of translocation calls, and all analyses on non-reference mobile elements lists generated by Eunjung Lee. I also developed the TIGER tool for reliable discovery of L1-mediated 3' transductions and designed primers needed for experimental validations performed by Adrian Stütz. Bernd Klaus provided input for statistical analyses and performed goodness-of fit test for the species-specific transduction rates. Benjamin Raeder and Adrian Stütz prepared non-human primate samples for single-molecule sensing experiment (MinION). Subsequent computational analyses of MinION reads were generated by Andreas Untergasser. I analyzed Pacific Biosciences (PacBio) reads for NA12878 human sample. The analysis of retrogenes presented in this chapter was performed by Verena Tischler with my support. This project was supervised by Jan Korbel and Adrian Stütz who provided significant support and feedback on analyses and results presented in this chapter.

3.1 Motivation and background

Due to their ability to move within the genome, MEIs are an important source of genomic structural variants forming *de novo*. Active L1 elements are ∼6 kb in length and contain two open-reading frames (ORFs), which encode proteins required for retrotransposition.

As indicated in Chapter 1, MEIs belong to an active class of transposons, moving within genome through copy-and-paste mechanism known as TPRT. Upon transcription, the RNA polymerase sometimes skips weak transcription termination signals (polyadenylation (polyA) signal, 5'-AATAAA-3' for L1 and SVA), and subsequently terminates RNA synthesis at downstream, 3' polyA signal. The consequence of 3' transduction process is the mobilization of downstream flanking sequence together with the retrotransposon. If the transduced sequence contains genes or other functional elements, transduction of such sequence can be a source of new structural variants contributing to diseases [Solyom et al., 2012a, Tubio et al., 2014] and gene evolution [Xing et al., 2006]. MEIs are also capable of mobilizing the upstream, 5' sequence. If an 5' promoter upstream of L1 or SVA reads through the downstream sequence including the element, it will create a new 5' start of the transcript [Damert et al., 2009] and subsequently carry the additional 5' transduced sequence to a new genomic locus.

Apart from L1 *cis* preference for their encoding RNA, L1 can additionally act in *trans* to promote retrotransposition of mutant L1s and other cellular mRNAs [Wei et al., 2001]. Insertion of such mRNA results in an intronless gene duplication known as retrogene insertion (Figure 3.1). In a recent study, Ewing et al. [2013] have shown that retrogenes are a widespread phenomenon, present in different species, as well as in cancerous somatic tissues.

As indicated, retrotransposon activity usually results in an novel insertion, but the outcome of such process can be different in each iteration. Therefore, the inserted element can be a standard full-length sequence, can be accompanied by the transduced sequence, but can also be severely truncated (Figure 3.1). Essentially all combinations are possible, including truncations and transductions in the same inserted element. As described in Chapter 1, Solyom et al. [2012a] have shown that L1-mediated 3' transduction with severe 5' truncation can insert into the *dystrophin* gene and ultimately lead to genetic disease known as Duchenne muscular dystrophy.

Few recent studies have looked at non-reference L1 transduction events (i.e. transductions leading to insertions of sequence not present in the reference genome), with L1 transduction events representing the most common form of mobile element mediated transduction in humans and non-human primates. Kidd et al. [2010] characterized polymorphic non-reference 3' L1-transductions using capillary sequenced fosmid libraries, whereas another more recent study investigated the importance of somatic L1 transductions in cancer genomes [Tubio et al., 2014]. Due to the lack of tools for the discovery of polymorphic L1 transduction events in the germline, however,

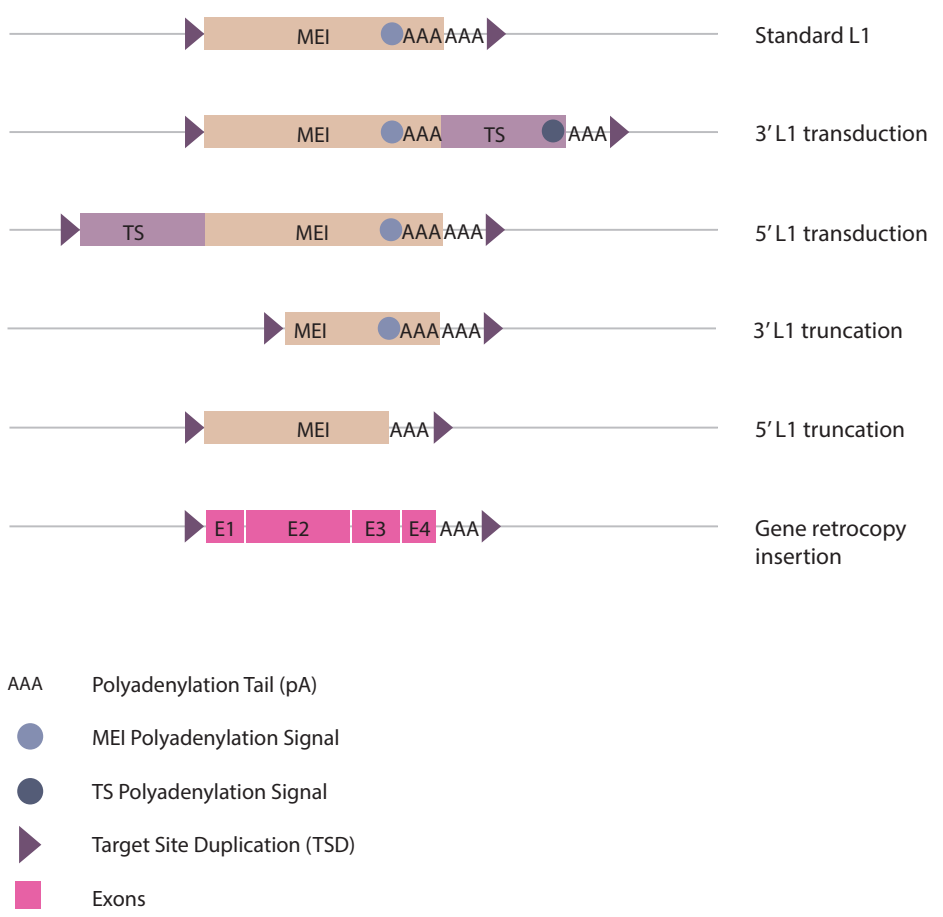

FIGURE 3.1: Structural variants of ME (L1) elements. Standard L1 is full-length (∼6 kb) and obtains polyadenylation tail (polyA) and target site duplication (TSD) upon insertion. Structures of 3'-transduction (3' additional sequence), 5'-transducing (5' additional sequence), 3'-truncated (3' sequence missing), 5'-truncated (5' sequence missing) and gene retrocopy insertions (intronless gene copy insertion) are depicted below.

systematic assessments of transduction extent and activity based on analyzing non-reference transduction events have not yet been performed.

In this chapter I will present a novel tool for the detection of 3' L1-mediated TRanductions In GERmline (TIGER) genomes using next generation sequencing data. For this purpose, we used a combination of novel L1 insertion [Lee et al., 2012] as well as translocations predictions [Rausch et al., 2012b] to identify an insertion of L1 and 3' transduced sequence. We applied TIGER to 5 individuals each of three different non-human primate species (chimpanzee, orangutan and rhesus macaque) presented in Chapter 2 as well as a well-characterized human genome, to test its ability to identify transductions and to characterize L1 transduction activities in different primate species. Our results identify significant differences in L1-mediated 3' transduction rates across primate species and indicate species-specific L1 subtypes in this process. Additionally, we

applied previously published tool GRIPper [Ewing et al., 2013] to infer gene retrocopies in same non-human primates and have shown that the rate of retrogene insertion varies across species.

3.2 Identification of L1-mediated 3' transductions

Improvement of NGS sequencing techniques as well as the computational approaches dealing with variant discovery enabled further exploration of genomic SVs. MEIs present one of the most difficult variant type to study, as they are highly repetitive and can create ambiguities upon alignment and assembly, resulting in detection biases and errors. Recently, many algorithms have been developed in order to improve the detection accuracy of MEIs [Keane et al., 2013, Lee et al., 2012, Thung et al., 2014, Wu et al., 2014].

In this study, we decided to build on current knowledge and use already published tools to develop a novel method (TIGER) that accurately detects L1-mediated transduction events in germline. As a proof-of-principle, we applied TIGER to chimpanzee, orangutan and rhesus macaque whole-genome sequencing (WGS) data from five individuals per species, sequenced between 14.4-28.8x [Gokcumen et al., 2013] and additionally a human sample NA12878 (HapMap/1000GP CEU daughter [Abecasis et al., 2010, The 1000 Genomes Project Consortium, 2012], downsampled to ∼21x using 3 independent technical replicates). TIGER uses a combination of (1) non-reference L1 insertions (in this study discovered by TEA [Lee et al., 2012], (2) translocation (TL) calls (here identified by DELLY [Rausch et al., 2012b]) as well as (3) single-anchored (SA) reads obtained directly from a BAM (Binary Alignment/Map) file [Li et al., 2009]. SA and TL reads are found as discordantly mapped read pairs, either having one read mapped and the mate unmapped (SA) or both read and mate mapped onto two different chromosome (TL) [Korbel et al., 2007].

In brief, we looked for the overlap between non-reference L1 insertion and at least one TL read, which implies the presence of L1-mediated transduction, manifesting as insertion of L1 element accompanied by additional unique sequence originating from another chromosome. For every identified L1-TS candidate region, all discordant (TL or SA) reads mapping within this region were subsequently obtained and their respective mates realigned onto the corresponding reference genome to detect the possible source element. To identify the most probable source region per insertion locus, we required a set of uniquely mapping mates to cluster on one chromosomal region in an overlapping fashion. In addition to the unique transduced sequence, we searched for a cluster of repetitive reads mapping randomly multiple times in the genome indicating a L1 element presence. To prevent any reference biases, all predicted L1-TS insertion regions were filtered for overlap with corresponding SD dataset (using the combined dataset described in

Chapter 4) as well as the presence of a reference L1 at the insertion site (Figure 3.2, for further details see Methods and Appendix A, Figure A.6).

As described in Chapter 1, typical MEI sequence usually contains a polyadenylation tail and is flanked by a target site duplication (TSD) upon insertion. The same is true for L1 carrying a transduced sequence. In order to further annotate TIGER transductions, we associated each predicted L1-TS with the corresponding TSD values from original L1 insertion file generated by TEA [Lee et al., 2012], whereas a putative presence of a polyA tail was evaluated by searching for six consecutive non-reference A's or T's (AAAAAA/TTTTTT) in each read uniquely mapping and clustering in the source loci.

We hypothesized that transductions, same as solo-L1 insertions are driven by a species-specific active L1 elements. In order to assess which subfamily class promotes retrotransposition in each species, we derived subfamily-specific consensus sequence from all full-length (>6 kb) primate active L1 elements and remapped repetitive reads indicating L1 presence. Best mapping suggests most probable L1 subfamily causing the transduction in each locus.

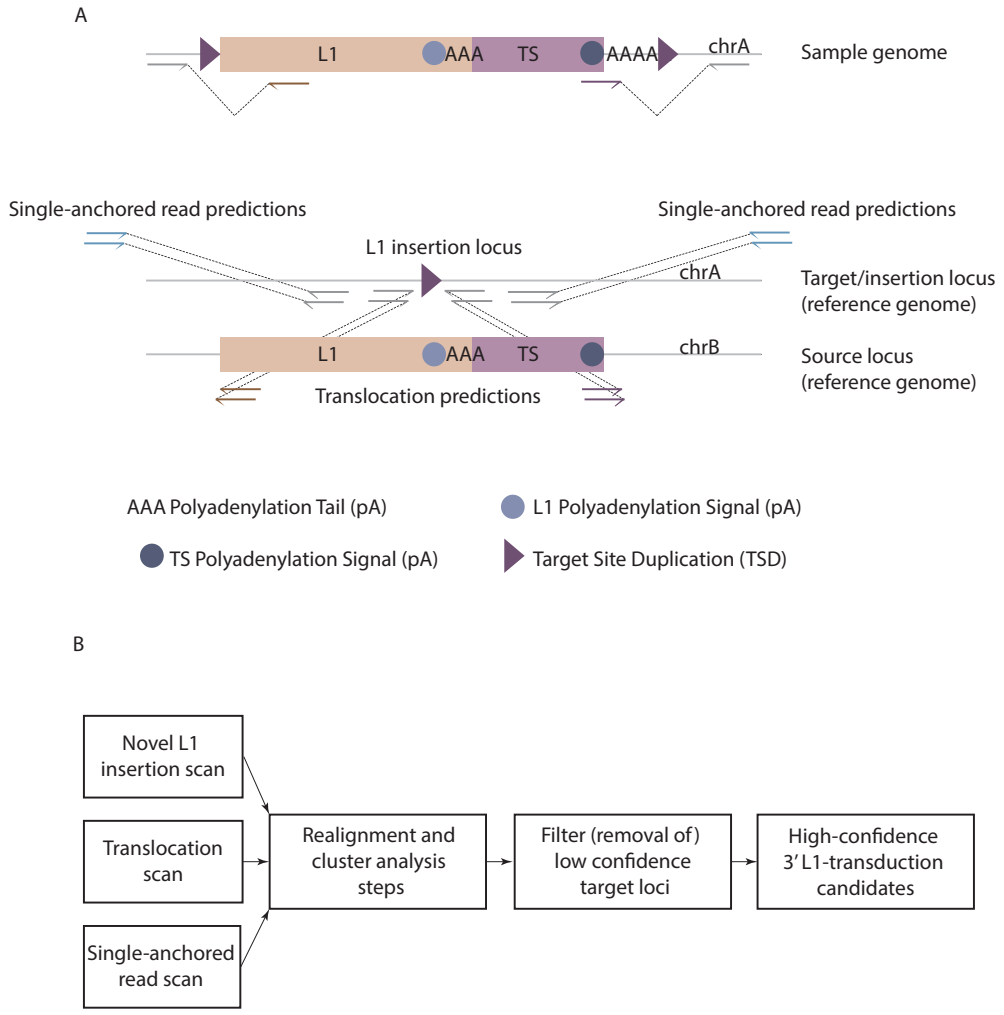

FIGURE 3.2: TIGER approach. (A) L1-TS insertions are typically composed of flanking target site duplications (TSDs), L1 sequence and unique TS sequence followed by a non-reference polyA tail. To detect such events, candidate regions are chosen based on an overlap between L1 insertion loci, the paired-ends indicative for the translocation (TL) of unique DNA stretches between chromosomes, and remapped single anchored (SA) reads in the reference genome. (B) A combination of L1 insertion, translocation-indicating and single-anchored reads are used to detect L1-TS insertion candidates, whereby TL and SA mate reads are realigned to correctly place them on the genome. Candidate regions are subjected to filtering in order to remove low confidence loci, resulting in a high confidence L1-TS insertion calls.

3.3 L1-mediated 3' transductions in non-human primates

We subjected the entire set of 15 non-human primate individuals (comprising of 5 chimpanzee, 5 orangutan, and 5 macaque individuals [Gokcumen et al., 2013]) to TIGER. In total, 275 non-redundant L1-mediated 3' transductions were detected: 71 in rhesus macaque, 191 in orangutan

and 12 in chimpanzee (Appendix A, Figure A.7, Figure A.8), with the average number of L1-TS per individual amounting to 27.8 in macaque, 62.4 in orangutan and 6 in chimpanzee.

To assess the ability of TIGER to identify L1 transduction events, we analyzed a set of recently published non-human primate genomes presented in Chapter 2 [Gokcumen et al., 2013] with our tool. An example of the computational evidence of a TIGER transduction is shown in Figure 3.3, where a unique sequence on chimpanzee chr7:6620368-6620628 was predicted to insert into chr10:54643580-54643593 region (TSD=13 bp). Out of all discordantly mapping reads in the target locus, we found a cluster of unique reads mapping to the source chromosome 7, as well as a cluster of repetitive reads indicating presence of L1. Some of the uniquely mapping reads carry a non-reference polyA tail indicating important evidence for a transduction in contrast to a regular translocation. Also, a few bp upstream of the polyA tail, the new polyA signal which caused the transduced sequence transcription to terminate can be seen.

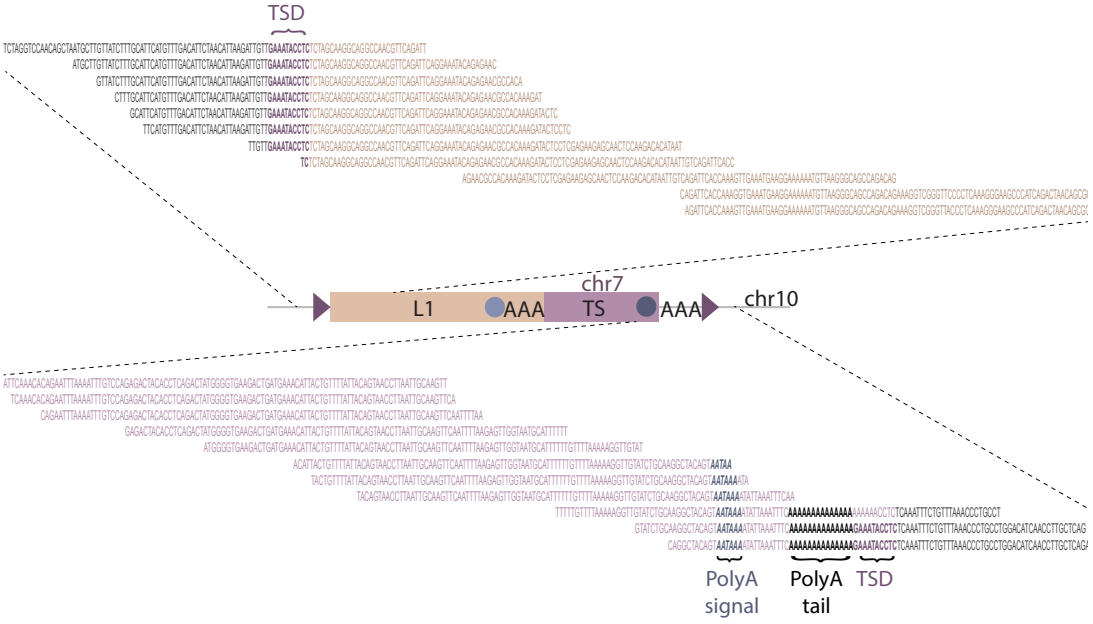

FIGURE 3.3: Computational analysis of insertion chr7:6620368-6620628 into chr10:54643580-54643593 region in the chimpanzee sample PR01171: (A) There were 29 predicted unique reads clustering to source chr7 with an average 'uniqueness' of 1, indicating that each read maps only once in the reference genome and fulfilling the criteria of being smaller than 3 (arbitrary cutoff for 'uniqueness'). Out of 29 reads, 7 carry part of a non-reference polyA tail indicating important evidence for a transduction in contrast to a translocation (only subset of reads shown).

To make sure that the reads are correctly mapped (as shown in Figure 3.3), all TL and SA unmapped mates were realigned to the corresponding reference genome using the BLAT software

[Kent, 2002]. The realignment is a crucial step for high quality predictions, which subsequently facilitates correct read clustering and therefore adds substantial detection power to the TIGER tool. After realignment, many of previously unmapped reads were placed correctly onto the reference genome, identifying the true L1-TS event instead of a regular translocation (visualized with the IGV software [Robinson et al., 2011, Thorvaldsdóttir et al., 2013], Figure 3.4).

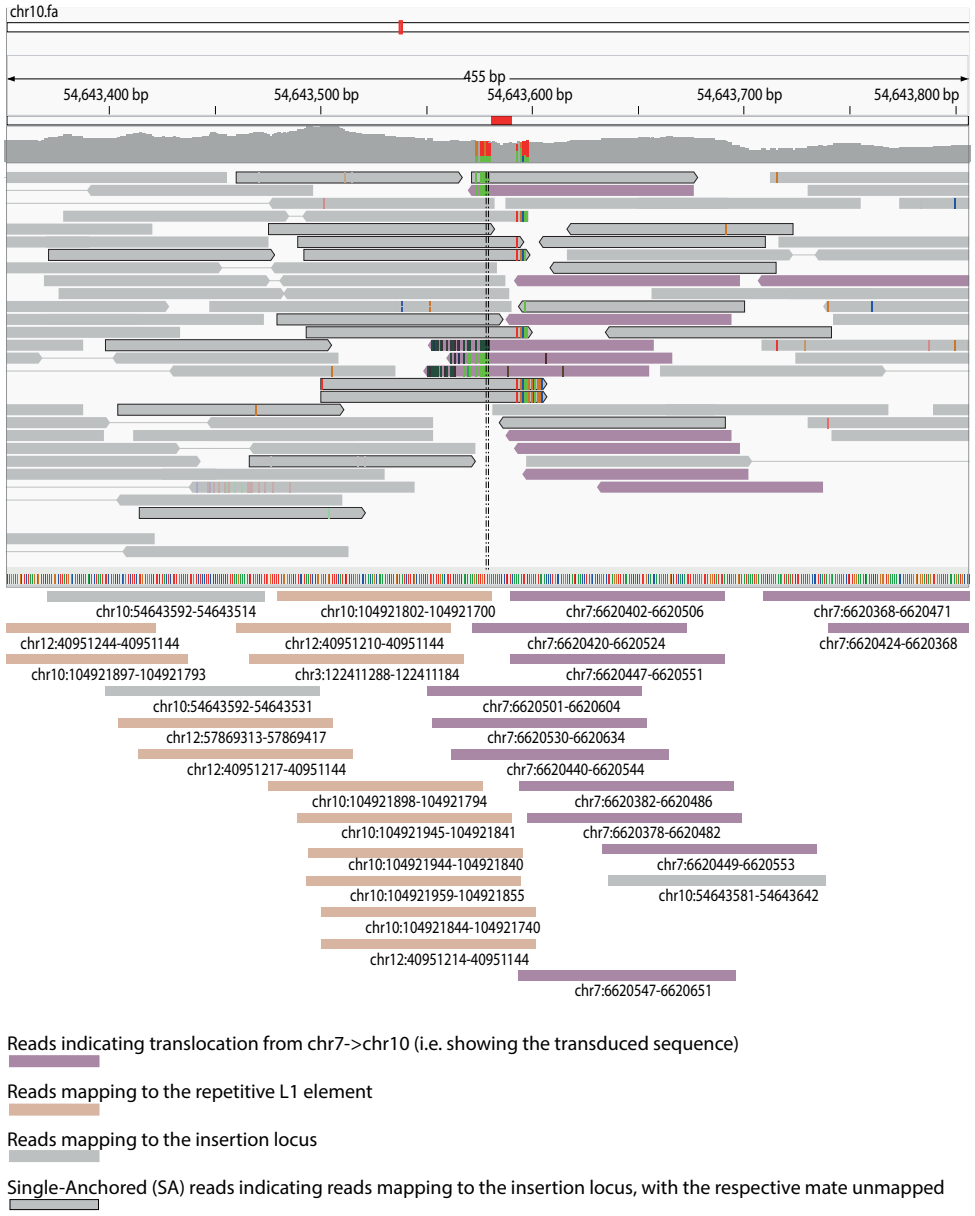

FIGURE 3.4: Computational analysis of insertion chr7:6620368-6620628 into chr10:54643580-54643593 region in the chimpanzee sample PR01171: chr10:54643580-54643593 region depicted using the Integrative Genomics Viewer (IGV) initially showed only TL reads with the polyA stretch indicating a potential transduction (cluster of reads on the left side of the breakpoint). After realignment, many of these reads were placed correctly onto the reference genome, which allowed us to add another track with reads clustering on both sides of the breakpoint identifying the L1-TS correctly.

Additionally, clustering of at least four DNA sequence reads with a mean size of 101 bp on the same source chromosome offered us the possibility to construct extended sequence stretches that reflect the portion of unique DNA sequence transduced. Sizes of predicted TSs calculated based on the paired-end read clustering varied between 64 bp-361 bp in macaque, 90 bp-260 bp in chimpanzee and 74 bp-437 bp in orangutan. This indicates that the minimal sequence requirement to predict transductions with TIGER using NGS data is approximately 50 bp, whereas the upper value represents the maximum length we were able to computationally assemble.

We further characterized the predicted TS sources to check if they mobilize any exons in the genome, and indeed found one candidate in orangutan and one in macaque. As the evidence for L1 predicted to insert was minimal (only three reads supporting L1 insertion), we initially thought they might be orphan transductions lacking the L1. Surprisingly, after closer inspection and experimental validations, these insertions turned out to indicate retrogenes insertions (gene retrocopy insertion polymorphisms, GRIPs), sharing the diagnostic features such as TSD and polyA with L1-TS and being mobilized by the L1 machinery. Although the TIGER tool is not specifically designed to detect GRIPs, particularly due to the absence of MEIs in GRIPs, it can be used for that purpose if the input set is changed accordingly. We inspected all other transduction candidates generated by TIGER and confirmed that no additional events corresponded to GRIPs. Of all transductions inferred by TIGER in rhesus macaque, 40 source regions are originating from intron sequence and 25 are inserting into an intron, out of 71 source-target predictions in total; in orangutan out of 191 transduction calls, 57 target regions are identified to be introns and 59 are predicted as sources-introns; and in chimpanzee 5 sources are intron sequences and 4 are inserting into an intron.

3.4 Experimental validation of primate-specific L1-mediated transductions

Experimental validations were performed on 52 randomly chosen calls (∼20% of all predicted calls) by PCR with a combination of an 'outer' and 'inner' primer pair and capillary (Sanger) sequencing (Figure 3.5, for details see Methods). Primers were designed to bind to unique regions at least 100 bp away of the target intergration site using in house primer design tool. Due to severe 5' truncations of predicted insertions, expected sizes of events (∼6 kb for solo-L1 and >6 kb for L1-TS) could not be used to assess the accuracy of predicted L1-transduction compared to solo-L1. Additionally, capillary sequencing with 'outer' primers alone was not able to confirm the TS because of inability to read through polyA tail in the MEI and a polyA tail in the TS and yield TS sequence located between two polyA tails. Therefore, we subsequently designed the 'inner' set of primers (within predicted source) using Primer3Plus [Untergasser et al., 2007] followed by

PCR amplification and another round of capillary sequencing. For a true L1-TS event, sequences obtained from the 'inner' primer pair binding to the TS, should match uniquely to the source chromosome, flanked by a non-reference polyA stretch on one side and a polyT stretch on the other side, indicating the end of the transduced sequence and the end of the L1 responsible for the transduction, respectively. To identify the negative result, we observed two possible scenarios: (1) the reference sequence with no insertion and (2) an insertion of L1 element with its non-reference polyA tail. In both cases no additional transduced sequence was seen and TIGER prediction was deemed to be wrong. In case of PCR failure (tested with two independent 'outer' primer pairs), the validation result was marked as unclear. The same indication was applied to sequencing failure results because the sequence identity was not confirmed, despite observing a band larger than the expected reference size indicating an insertion.

We have successfully validated 7 L1-TS calls in chimpanzee, 28 in orangutan and 17 in rhesus macaque (in total, 43 L1 transduction; Figure 3.5 C and Appendix A, Figure A.9). We assessed the False Discovery Rate (FDR) of 17%, with similar validation successes in different non-human primate species (Table 3.1, Figure 3.5). Upon closer inspection, we found that 7 out of 9 negative loci were insertion-negative indicating that only the reference genome sequence was observed (false MEI insertion prediction), whereas only 2 TIGER predictions were transduction negative, confirming an solo-L1 insertion without a transduced sequence. In general, the FDR therefore is highly depending on the quality of the MEI input predictions, in addition to the evidence from paired-ends indicative for translocated sequence. Our overall FDR of 17% is in agreement with a recent assessment of FDR for the TEA MEI caller [Lee et al., 2012] used in this study for germline L1 insertions (FDR: 16-24%, [Keane et al., 2013]). Furthermore, in 43 out of 45 examples where an MEI turned out to be correctly inferred, our PCR validations verified that unique sequence stretches were transduced, indicating high accuracy of the transduction calls made by TIGER.

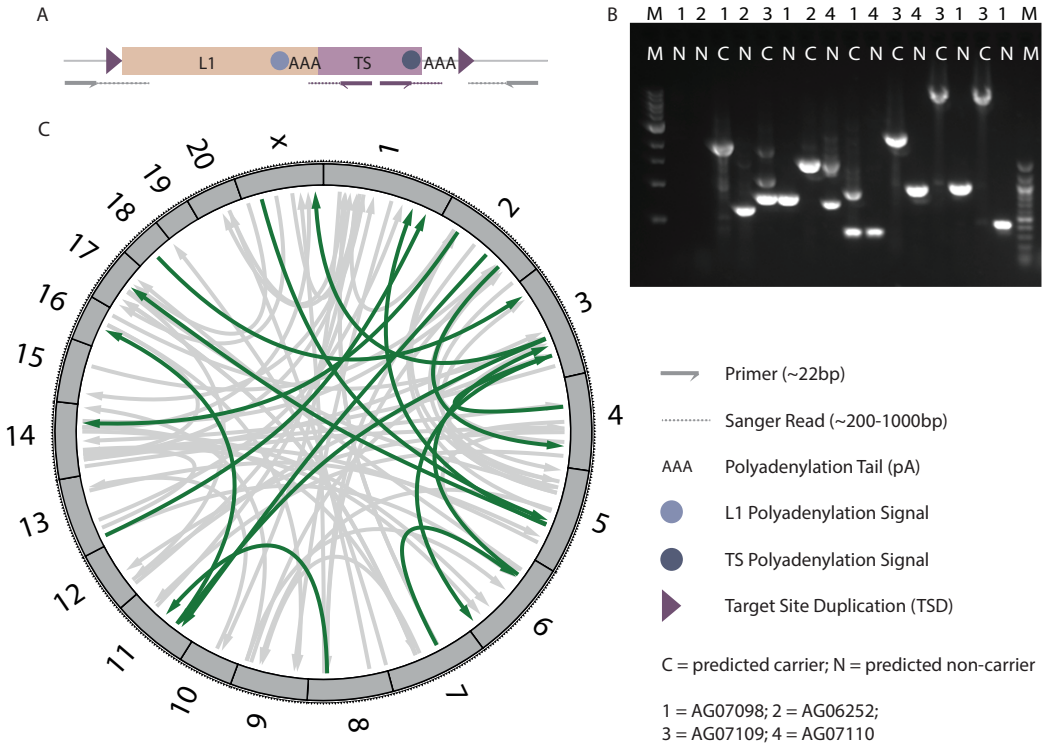

FIGURE 3.5: Experimental validations of computationally predicted L1-mediated 3' transductions. (A) General primer design: Outer (gray arrows) primers were placed outside of the event into the target locus in order to amplify the L1-TS insertion allele and/or the reference genome allele. On the left side of the locus, the corresponding Sanger sequence (dotted line) starts with a unique match to the target site, then splits and matches to multiple positions in the genome indicating the L1 element. On the right side, another corresponding Sanger sequence will also match uniquely to the target site and end with a polyA/T stretch not seen in the reference genome. In order to confirm the presence and the origin of the transduced sequence (source locus), a 2^{nd} set of primers (purple arrows), inside the predicted unique TS, is necessary together with further Sanger sequencing. (B) A subset of rhesus macaque L1-TS PCRs using the outer primers are shown: for predicted carrier (C) and non-carrier (NC) samples. In case of an L1-TS insertion, a larger band than the reference band in NC is seen; heterozygote samples show both bands whereas homozygous L1-TS insertions show only the higher band. (C) A circos plot (http://circos.ca/) shows the distribution for all rhesus macaque L1-TS predictions, the 14 experimentally validated insertions are depicted in green arrows. Arrows indicate the direction of the source inserting into the target locus.

The longest transduction we detected after experimental validation and sequencing was 6000 bp, whereas the smallest was 300 bp. This range indicates that we are not able to always recover the full TS sequence size with the TIGER tool, as the longest TS stretch computationally predicted was 437 bp. Experimental results indicate prevalence of 5' truncated L1 elements accompanying TS. In rhesus macaque, 2 out of 14 validated L1-TS sequences are approximately ~6 kb long indicating a full-length insertion. In chimpanzee only one insertion is ~6 kb long, whereas in orangutan none of the inserted L1-TS is full-length (Figure 3.6). The only full-length chimpanzee insertion was presented earlier (Figure 3.3, Figure 3.4): chr7 TS insertion into chr10 target loci.

Due to the limitations of the paired-end mapping, we have been able to computationally recover 260 bp of the TS, instead of 6 kb. Interestingly, L1 elements responsible for transductions were frequently not found in the reference genome upstream/downstream of the inferred source indicating formation through active polymorphic L1 elements.

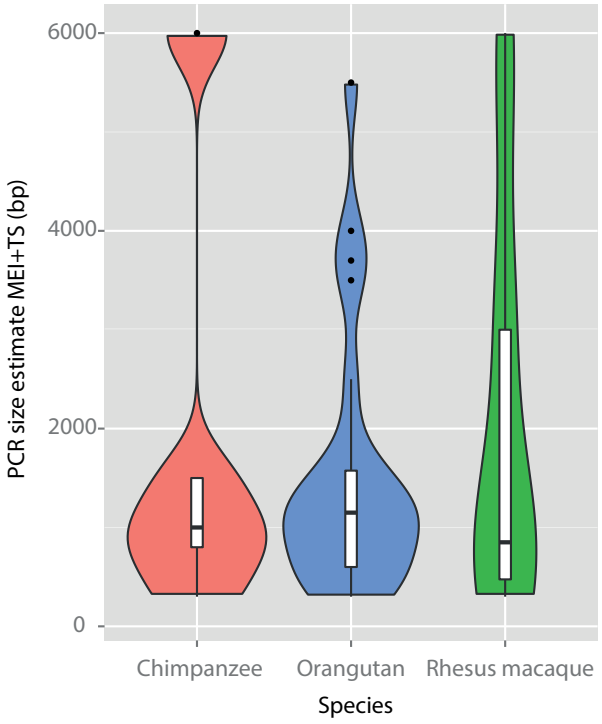

FIGURE 3.6: L1-TS insertion sizes based on experimental results. The size range detected after experimental validations and sequencing was between 300 bp and 6000 bp, indicating that some of the validated predictions contain severely truncated L1, whereas fewer are full-length L1 elements accompanying 3' transduction sequence. In comparison to chimpanzee and orangutan, rhesus macaque has a slight shift of size distribution towards larger insertions (more uniformly distributed compared to the great apes).

We also used the long read technology to get a deeper insight into L1-TS insertion as they allowed us to recover the entire inserted sequence. For instance, MinION long reads spanning the rhesus macaque L1-TS insertion locus on chromosome 5 (chr5:113783999-113784017) indicated ∼1500 bp long insertion. Inspection of inserted sequence revealed 742 bp long L1 element and 644 bp long TS including polyA tail. MinION reads also confirmed that L1 sequence inserted together with TS was severely 5' truncated, formerly apparent from the experimental validations (size of the band ∼1000 bp long) (Figure 3.7).

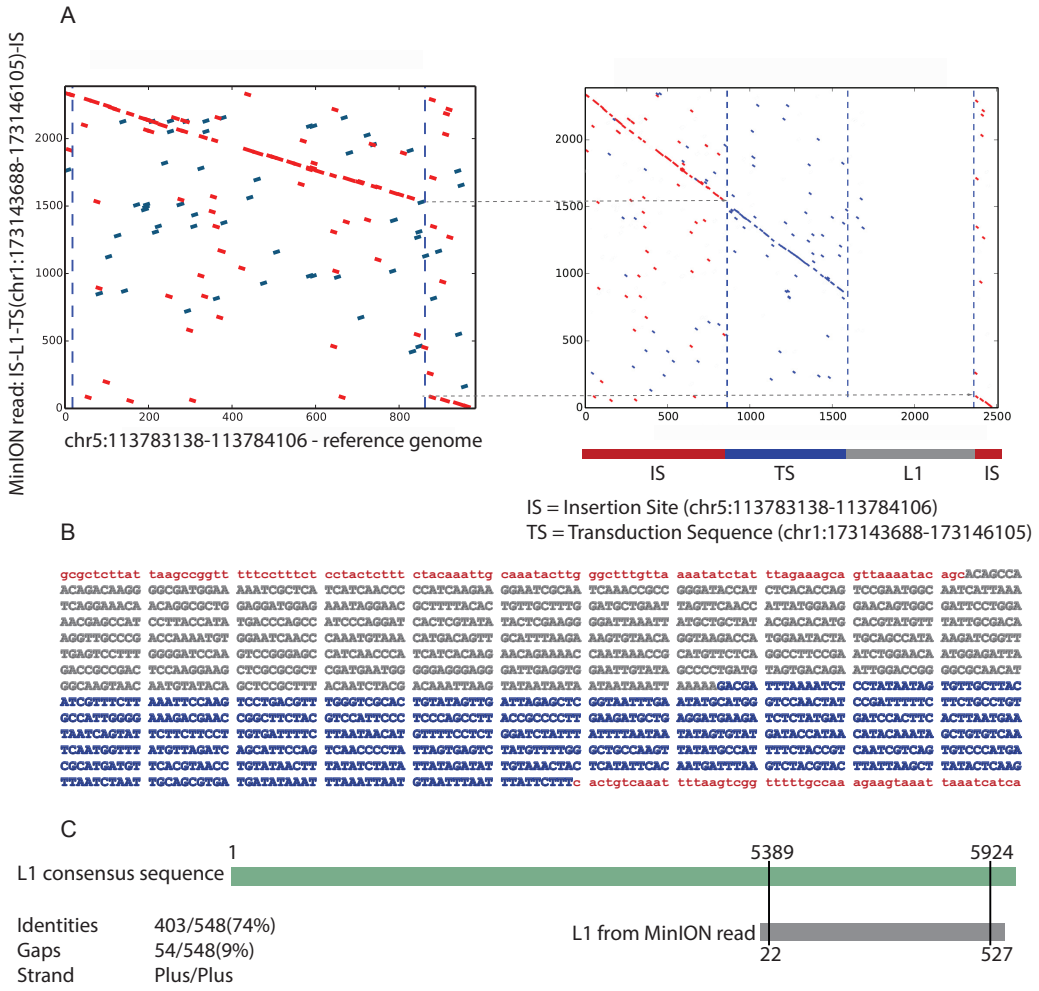

FIGURE 3.7: MinION long read L1-TS confirmation in rhesus macaque: (A) Dotplot with reference genome sequence shown on x axis and MinION long read on y axis: ∼1500 bp shift indicates an insertion. (B) Inspection of inserted sequence revealed 742 bp long L1 element and 644 bp long TS including polyA tail (based on one MinION read). (C) L1 inserted together with TS was severely 5' truncated, shown in relation with ∼6 kb long L1 consensus sequence (pairwise-alignment performed with BLAST [Altschul et al., 1990]).

To investigate properties of transduced sequences as well as target loci, we looked at the replication timing of source and target regions. Due to the lack of replication timing data in non-human primates, we have converted the primate transduction chromosomal coordinates to human using the liftOver tool [Hinrichs et al., 2006] and searched for overlap with replication time values from human fibroblast cell line (downloaded from http://www.replicationdomain.com/, for details see Methods). As previously reported, early replicating regions are significantly depleted of L1 insertions, which rather tend to occur in late replicating parts of the genome [Hansen et al., 2010]. We have observed the same phenomena, with no significant difference in replication time between target and source sequences. However, in target regions we noticed a slight shift of replication time distributions towards more negative values, indicating tendency of *de novo* insertions to occur in 'even later' replicating regions compared to source elements. This is further supported by distribution of reference and polymorphic L1 replication timing values, which can be treated as global 'sources' and 'targets' for L1 insertions, respectively (Appendix A, Figure A.11). Prior studies also reported a strong positive correlation between GC content and early replication [Costantini and Bernardi, 2008, Watanabe et al., 2002], indicating that late replicating regions are AT rich. As described in Chapter 1, upon insertion, L1 endonuclease generates a single-stranded 'nick' in the genomic DNA at the 5'-TTAAAA-3', further supporting insertion mechanism in late replicating regions. Indeed, when looking at the target sequence motives (TSD sequences longer than 8 bp that get duplicated after insertion), they are almost always AT rich.

3.5 L1-mediated 3' transductions in human

While the focus of TS analysis was on primate genomes, we also investigated the ability of TIGER to identify non-reference L1-mediated transductions in humans by analyzing the well characterized CEU sample NA12878 [Abecasis et al., 2010, The 1000 Genomes Project Consortium, 2012]. We downsampled NA12878 reads generated by the 1000 Genomes Project to yield three 'technical replicates' with similar coverage as the primate samples (∼20X) (NA12878_1, NA12878_2 and NA12878_3). TIGER predicted 6 L1-TS calls in NA12878_1, 4 in NA12878_2, and 1 in NA12878_3, respectively. After merging, the total number of non-overlapping predicted transductions for NA12878 was 6.

Evaluation of these human L1 transduction calls was done using a single-molecule, long read Pacific Biosciences (PacBio) dataset [Eid et al., 2009] of the same NA12878 sample (Figure 3.8). This technology allows resolving complex structural variations as well as low complexity regions such as MEI due to the read length of up to ∼8 kb. As shown in Figure 3.8, PacBio long reads revealed insertion of 908 bp long L1 element accompanied by 126 bp long TS including polyA tail into chromosome 4 (chr4:104210671-104214687 region). Again, as apparent from the size,

L1 inserted was severely 5' truncated compared with ∼6 kb full-length L1 element. All together, 4 out of 6 TIGER candidates in NA12878 were confirmed to contain L1-TS (2 by PacBio reads and 2 found in Kidd et al. [2010]), whereas 1 showed only an L1 insertion without transduced sequence and the last locus remained inconclusive due to no spanning long reads. In human data we observed 3 L1-TS insertions in introns and 3 sources predicted to fall in intron regions.

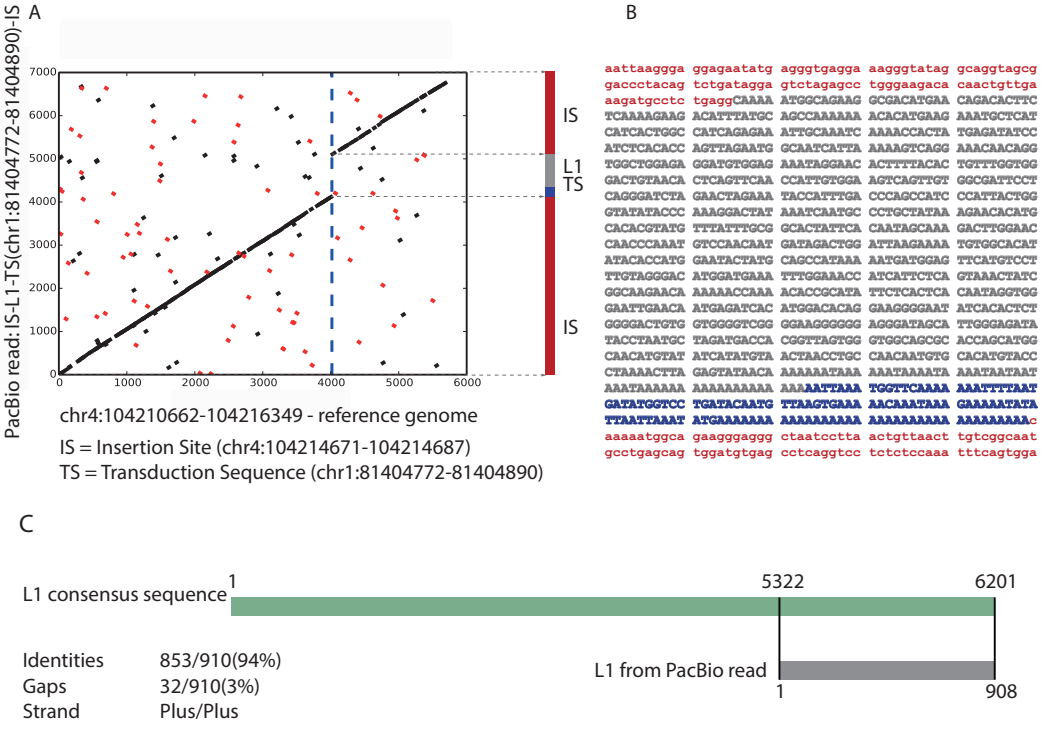

FIGURE 3.8: Pacific Biosciences long reads confirm L1-TS insertion into the human chr4:104210671-104214687 region. (A) Dotplot with reference genome sequence shown on the x axis and PacBio long read on y axis: ∼1000 bp shift indicates an insertion. (B) Inspection of inserted sequence revealed 908 bp long L1 element and 126 bp long TS including polyA tail (consensus sequence created from all PacBio reads by multiple sequence alignment (MSA) [Lassmann and Sonnhammer, 2005]). (C) L1 inserted together with TS was severely 5' truncated, shown in relation with ∼6 kb long L1 consensus sequence (pairwise-alignment performed with BLAST [Altschul et al., 1990]).

3.6 Species-specific L1-mediated 3' transduction rates

In order to calculate the rate of transductions per species, the number of high confidence TIGER calls was divided by the number of high-confidence non-reference L1 insertions identified by TEA [Lee et al., 2012]. These rate estimates showed significant differences between species with 2.5%±1.1 CI (t-test, 95% confidence intervals) transduction rate in chimpanzee, 8.8%±1.4 in orangutan and 5.5%±1.2 in macaque (Tweedie goodness-of-fit, chimpanzee-orangutan (P=0.000037), chimpanzee-macaque (P=0.000073) and orangutan-macaque (P=0.0003); Table 3.1). With the

exception of orangutans, these rate estimates for non-human primates are lower than the previously reported human L1 transduction rates of 10-25% (Table 3.2) which were pursued either on reference transductions many of which are likely not polymorphic [Pickeral, 2000, Szak et al., 2003, Xing et al., 2006] or on somatic L1 transductions [Helman et al., 2014, Solyom et al., 2012b, Tubio et al., 2014].

TABLE 3.1: Summary of TIGER results. *Cov* = physical coverage (sequencing coverage is presented in Appendix A, Table A.8), *L1-TS* = L1-mediated transduction, *TSs* = transduced sequences. PR00738 and PR00818 chimpanzee samples have higher coverage in comparison to Gokcumen et al. [2013] due to the BAM file [Li et al., 2009] merging (from different libraries).

Species	**Sample**	**Cov**	**L1**	**L1-TS**	**L1-TS rate***	**Validated TSs**
Rhesus macaque	AG06249	26.0	449	29	5.5±1.2**	14/17
	AG06252	29.2	620	28		
	AG07098	21.7	424	26		
	AG07109	23.7	473	28		
	AG07110	18.6	635	28		
Orangutan	AG06105	19.2	663	52	8.8±1.4**	24/28
	AG06209	24.2	803	81		
	GM04272	24.0	649	62		
	PR00054	23.3	775	70		
	PR01110	17.2	633	47		
Chimpanzee	PR00226	32.2	214	4	2.5±1.1**	5/7
	PR00738	32.9	246	7		
	PR00818	28.2	223	4		
	PR01106	19.8	148	3		
	PR01171	18.8	132	5		

*Determined based on ratio between TIGER transductions and L1 insertions. 95% confidence intervals were calculated using one sample t-test.

**Significantly different based on goodness-of-fit test (Tweedie model): chimpanzee-macaque: P=0.000073; chimpanzee-orangutan: P=0.000037; macaque-orangutan: P=0.0003.

We estimated the total amount of high-confidence L1 calls based on the three downsampled human genomes to be 90 and therefore the transduction rate based on TIGER predictions to be 6.7%. This is slightly lower than presented in other human genome studies so far [Helman et al., 2014, Kidd et al., 2010, Pickeral, 2000, Solyom et al., 2012b, Szak et al., 2003, Tubio et al., 2014, Xing et al., 2006] which reported roughly ∼10% transduction rate in human genome. The study from 2010 [Kidd et al., 2010] reported that 20% of all L1 predicted in nine human genomes carry additional sequence as transductions. After reevaluating these results by applying the following

filters as used in our study: (1) focusing on novel L1 and L1-TS insertions, (2) ignoring reference MEIs and *Alu*/SVA elements, (3) requiring a minimum length of TS to be 50 bp and, (4) absence of reference MEI as well as SDs in the insertion loci, the rate would translate into 7.5% (5 L1-TS out of 66 high-confidence L1 insertions in total). Reassuringly, out of 6 transductions in Kidd et al. [2010] study they observe in nine human genomes, 2 match to transductions we observe in our NA12878 sample.

Table 3.2: Comparison of 3' transduction rates (%)

Human somatic TS**	**Human somatic partnered TS****	**Human somatic orphan TS****	**Human germline TS*****
22.4	10.2	12.1	7.5

*Determined using TIGER approach

**Adapted from Tubio et al. [2014]

***Adapted from Kidd et al. [2010] with identical parameters used in TIGER

Recently, somatic transductions in human cancers were studied using next generation sequencing data [Tubio et al., 2014]. In our study investigating the germline, we find similarities such as similar rate of L1 partnered-transductions in some of the tumor samples (Appendix A, Table A.2) but also differences to their somatic results, indicating that these might have different properties. Tubio et al. [2014] observed the existence of L1-master elements driving somatic transduction insertions into multiple regions in the human cancer genomes. This is in contrast to our observations of germline transductions, which present a one-to-one pattern (one source inserts into one target). The transduction rate of cancer-specific somatic L1-TS was shown to be as high as 22.4%, but that rate is highly dependent on a tumor type. For instance, tumors that show high variability in numbers of predicted L1-TS events are colon, lung and prostate cancer, where the rate vary from 0-100% (mean values are 26.7%, 34.3% and 10.3% for colon, lung and prostate cancer, respectively). Additionally, the L1-TS elements predicted in cancer can be split up into two classes: partnered- and orphan-TS, contributing with 10.2% and 12.2% to the transduction rate (Table 3.2). TIGER is developed for exploration of partnered-transductions exclusively, requiring both L1 and TS to be inserted. Taking that into consideration, our transduction rate predicted in human genome can be treated as similar to the one predicted in cancer. Moreover, rates of partnered transductions exclusively in previously mentioned tumor types are lower by more than half of total transduction rate with 5.4%, 12.5% and 1.59% mean rate in colon, lung and prostate cancer, indicating that orphan transductions dominate cancer L1 landscape.

3.7 Species-specific subfamilies driving transductions

Based on the reference polymorphic MEI elements analysis in each species presented in Chapter 2, L1 subfamilies were shown to differ in non-human primate genomes as well as in humans [Gokcumen et al., 2013]. When looking at both full-length and all L1 elements in the corresponding reference genomes, L1 subfamily divergence is apparent (Appendix A, Table A.3). Reference L1PA6-L1PA8 are pretty similar between the three non-human primate species; L1PA5 is specific to rhesus macaque, and L1PA2 is specific to chimpanzee and human (L1HS and L1PA2 diverged after chimpanzee-orangutan divergence, and L1HS (L1PA1) is mostly human-specific). To test if transduction events are driven by different subfamilies in each species, therefore resulting in different transduction rates, we performed an analysis to assess which subfamily dominates the non-reference L1 insertion landscape. In brief, repetitive reads were remapped to the consensus L1 subfamilies and best mapping with smallest mismatch was reported (for details see Methods).

Our results indicated that most of the L1 insertions in rhesus macaque belong to the L1CER subfamily evolved from macaque-specific L1PA5 [Han et al., 2007]. The same is also true for L1-TS calls, as most of the L1 accompanying TS are L1CER (Figure 3.9; Appendix A, Figure A.13). In orangutan, dominating subfamilies are found to be L1PA3, whereas in chimpanzee L1Pt drive most of the L1 insertion, as well as L1-TS insertions. Differences observed in subfamily distribution can explain numbers of L1 insertions, directly contributing different transduction rates across species (Figure 3.9; Appendix A, A.13 for L1 insertions).

3.8 Retrogene insertions in non-human primates

As indicted above, although TIGER is not specifically designed to detect GRIPs, we successfully identified two genes that inserted as intronless copies into target regions. In orangutan we detected unannotated *TMEM126B* retrogene insertion into chromosome 6 (chr6:80007306-80007324) and in rhesus macaque *PABPC4* retrogene insertion in chromosome 7 (chr7:65507000-65507014). Both genes are protein coding, with *TMEM126B* producing mitochondrial transmembrane protein, and *PABPC4* cytoplasmic poly(A) binding protein.

In addition, we also used the GRIPper tool [Ewing et al., 2013] on our 15 non-human primates to discover retrogenes specific to each species. GRIPper essentially searches for discordant mapping of paired reads, where one read maps to an exon of a gene and the other to another genomic location, indicating possible gene duplication. By using GRIPper, we identified a total of 35, 11 and 62 GRIPs in chimpanzee, orangutan and rhesus macaque, respectively (7, 2 and 12 on average per individual in chimpanzee, orangutan and rhesus macaque, respectively). As most of the GRIPs were predicted in multiple samples, we merged them and the final non-redundant numbers were 32 in rhesus macaque, 10 in orangutan and 24 in chimpanzee (Table 3.3; Appendix A, Table A.4, Table A.5, Table A.6). In rhesus macaque, the *PABPC4* retrogene

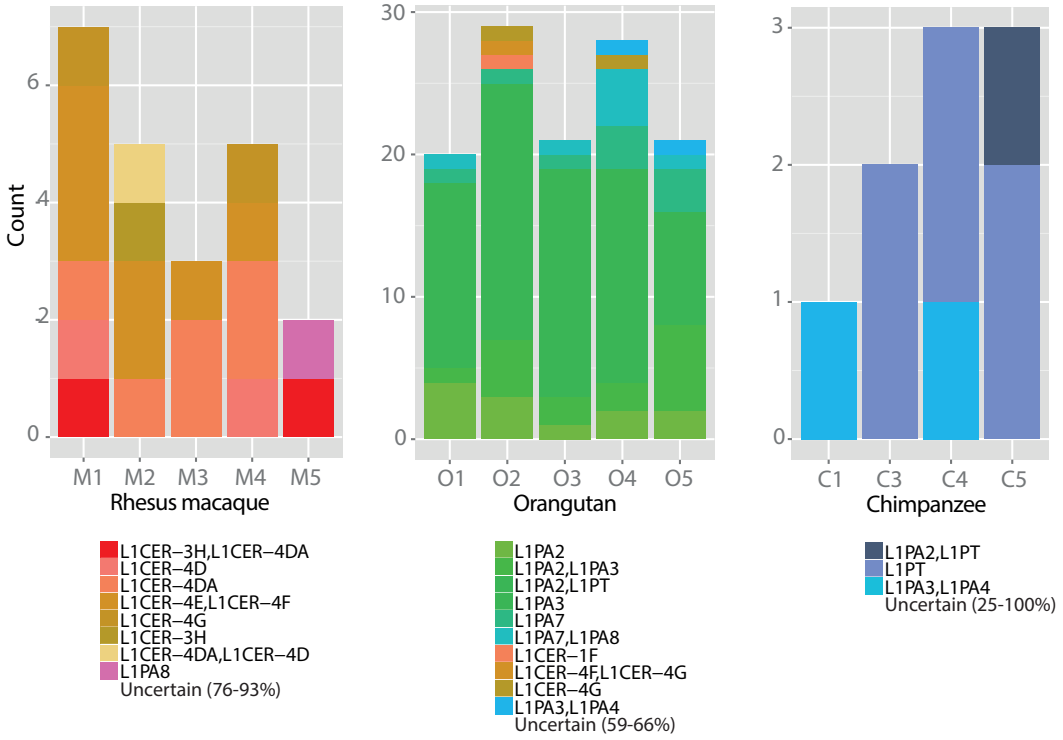

FIGURE 3.9: Differences in L1 subfamily driving the L1-TS insertions: In rhesus macaque, rhesus-specific L1CER subfamilies contribute to the most L1-TS insertions, whereas in orangutan dominating L1 subfamily are L1PA2 and L1PA3. Most of the L1 insertions in chimpanzee are driven by L1PA2/L1PA3 and chimpanzee-specific L1PT subfamily. 'Uncertain' subtype indicates that L1-TS had more than two predicted subfamilies, subsequently merged into 'uncertain' class. Values in parenthesis indicate how many of the predicted L1-TS are 'uncertain' in five individuals.

insertion, previously detected by TIGER, was identified by GRIPper as well. Interestingly, observed GRIP numbers show similar distribution as our *Alu* predictions in same species (see Chapter 2), whereas no correlation was observed between GRIPs and L1 elements, despite L1-encoded proteins driving the retrotransposition (and therefore retrogene insertion) in general.

TABLE 3.3: Gene retrocopy polymorphic insertions (GRIPs) in non-human primates.

Species	**Total GRIPs**	**Non-redundant GRIPs**	**Average number of GRIPs per sample**
Chimpanzee	35	24	7
Orangutan	11	10	2
Rhesus macaque	62	32	12

To test our GRIP predictions, we performed experimental validations and successfully validated 6/8 GRIPs (*EIF3*, *SDHB* and *NDUFB8* and *USP8* in chimpanzee, *UQCRB* in orangutan and *ACTG1*in rhesus macaque). We further subjected the remaining 6 positive GRIPs to Sanger sequencing and confirmed the presence of exon-exon junctions indicative of an intronless gene duplication. As a proof-of-principle, we wanted to check if GRIPs detected in our samples overlap with Ewing et al. [2013] GRIP dataset discovered in ten chimpanzee samples. Out of 24 predicted GRIPs in our samples and 19 predictions from Ewing et al. [2013] GRIP dataset, 8 inferred GRIPs overlapped (*SDHB*, *NDUFB8*, *EIF3*, *TRA2A*, *PHF23*, *NCBP2*, *LOC458071* and *CCT8*), 3 of which we have experimentally validated (*EIF3*, *SDHB* and *NDUFB8*). Since two datasets are independent, meaning no sample is shared between them and none of the samples are related, we assume that 8 shared GRIPs are either common GRIPs shared in population or represent GRIPs absent from the reference genome (i.e. private deletions in the reference sample).

3.9 Discussion

In order to inspect the impact of L1 elements successfully mobilizing unique sequence in primates, we have developed TIGER, a novel approach to detect L1-mediated 3' transduction events in germline. As indicated, transductions are an important class of structural variations previously poorly explored due to limitations in NGS approaches. TIGER successfully overcomes those challenges by utilizing short read data to detect L1-TS events in rhesus macaque, orangutan, chimpanzee and human. The high experimental validation rate confirms the reliability and quality of computationally predicted L1-TS calls. Transduction rates for non-human primates are highly variable dependent on the species, with orangutan having the closest rate to human. Rhesus macaque and chimpanzee have significantly lower TS rates than orangutan, whereas observed TSs in humans are concordant with previous studies [Helman et al., 2014, Kidd et al., 2010, Pickeral, 2000, Solyom et al., 2012b, Szak et al., 2003, Tubio et al., 2014, Xing et al., 2006]. We estimate, there are several reasons why we observe slightly lower transduction rate in comparison to studies so far: (1) we looked at novel germline L1-TS insertions, not fixed (reference) transductions or somatic events, (2) the focus of our study were transductions translocating from one chromosome to another, (3) we performed an additional filtering steps based on overlap with low confidence regions and 4) TIGER is limited to detect transductions $\pm$50 bp and requires at least part of unique sequence. Using TIGER we are investigating only L1-TS calls that would result in translocations, originating from one chromosome and inserting into a different one. This presents a potential limitation of TIGER as we might be missing transductions occurring on the same chromosome. We have additionally investigated our samples for such events using discordant paired-end reads (deletion and duplication calls) and have not found a pattern

suggesting transductions where the same chromosome presents both a source and a target. However, together with requiring the absence of L1 at the L1-TS insertion locus and limitation to detect orphan transductions, this might suggest why we observe smaller transduction rates than previous studies [Helman et al., 2014, Kidd et al., 2010, Pickeral, 2000, Solyom et al., 2012b, Szak et al., 2003, Tubio et al., 2014, Xing et al., 2006].

Although our detection approach did not necessarily differentiate between 3' and 5' transductions, we have not observed any 5' transduction events driven by an upstream promoter. Due to the fact that L1 elements belong to the autonomous retrotransposition-competent MEI class known to mobilize non-repetitive sequences, we exclusively focused on L1-mediated transduction detection. As previously shown, SVA elements are also capable of transducing unique DNA sequences. For example, reference SVA element was shown to be responsible for *AMAC* gene duplication before human-great apes divergence [Xing et al., 2006]. However, since SVA elements are completely absent from rhesus macaque genome and previously observed non-reference SVA elements in other species were relatively low in numbers [Gokcumen et al., 2013], SVA elements presented a limited dataset for our analysis. In contrast to both L1 and SVA, *Alu*-mediated transductions are so far not known to occur. *Alu* elements are often very small in length, they possess high sequence similarity and they are present in large numbers per genome which can present additional challenges and lead to problems when trying to identify novel *Alu* insertions carrying TS.

Interestingly, our analyses show that most of the L1 accompanying TS are polymorphic and subsequently deleted from the source location, likely due to population bottlenecks resulting in lost L1 source alleles. Similar approaches in cancer context show that one source L1-master element causes several transductions [Tubio et al., 2014]. This difference likely occurs due to different suppression of active L1 elements, where in healthy germline tissue such activity would be preferentially silenced and therefore L1-TS source allele would be lost. Consistently with previous studies [Tubio et al., 2014], L1 elements belonging to the L1-TS sequence insertion are severely 5' truncated, resulting in insertions smaller than expected given that TS is accompanied by full-length L1. The truncation of L1 likely happens in order for cell to prevent further retrotransposition of inserted L1 elements.

In summary, the development of the TIGER tool able to detect ME mediated transductions in germline is important for several reasons: transductions are a largely unexplored form of ME driven mechanism and the portion of such events is relatively high among all novel MEIs. In non-human primate genomes, they are an abundant form of L1 events, contributing to L1 diversity of the corresponding genomes. Also, by mobilizing unique sequences including GRIPs, L1 elements are in general responsible for duplication and shuffling of different genomic segments adding to the overall genomic diversity.

Chapter 4

SV formation differences in non-human primate species

Throughout this chapter, the structural variants detected in three non-human primate species (chimpanzee, orangutan and rhesus macaque) will be covered. For all predicted SVs with nucleotide breakpoint resolution, *de novo* formation mechanisms were inferred. Additionally, the map of novel mobile elements in every species has been identified and, together with the other SV mechanisms, the rate of formation was calculated and compared across species. The analyses and results presented in this chapter are based on the following publication:

> Gokcumen O.*, Tischler V.*, Tica J., Zhu Q., Iskow R. C., Lee E., Fritz M. H.-Y., Langdon A., Stütz A. M., Pavlidis P. et al. Primate genome architecture influences structural variation mechanisms and functional consequences. *Proceedings of the National Academy of Sciences of the United States of America*, 110(39):15764-9, September 2013.

Contribution

In this study, I performed the mechanism classification analyses, as well as the ancestral state inference on SV dataset generated by Verena Tischler. Additionally, in order to perform this analysis on non-human primates, I modified and applied previously published software [Lam et al., 2010]. I also performed analyses on novel non-reference mobile element lists provided by Eunjung Lee and Peter Park. This study was a collaboration between our laboratory and Charles Lee's group at Harvard Medical School in Boston. Omer Gokcumen managed the sample acquisition and coordinated all analyses performed in this study. Verena Tischler identified and characterized SVs in non-human primate genomes and performed several analyses based on the mechanism maps I generated (e.g. Monte Carlo simulations). SNV maps were identified by Qihui Zhu. Amy Langdon and Rebecca Iskow designed and performed SV PCRs and FDR assessments. Duplicative insertion sources were detected by Markus Hsi-Yang Fritz, based on

results of ancestral state analysis I implemented. Sequencing libraries were prepared by Adrian Stütz. Charles Lee and Jan Korbel supervised this study and, together with Omer Gokcumen and Verena Tischler, provided significant feedback on results presented in this chapter.

4.1 Motivation and background

As stated earlier in this Thesis, recent advances in the application of massively parallel sequencing (MPS) have enabled the discovery of large-scale variants ($\geq$50 bp). While SVs are presumed to have a major role in primate evolution and phenotypic variation [Varki et al., 2008], analyses of SV formation mechanisms have not been actively pursued in non-human primate species, due to the lack of inter- and intra-species nucleotide-resolution maps. Distinct activities of SV formation mechanisms may explain the differential genomic impact of SVs, making it necessary to understand how SVs actually form and emerge through evolution. Reference genome assemblies (Appendix A, Table A.7 shows primate genome statistics in a relation to the human genome) of the chimpanzee [The Chimpanzee Sequencing and Analysis Consortium, 2005], orangutan [Locke et al., 2011] and rhesus macaque [Rhesus Macaque Genome Sequencing and Analysis Consortium, 2007] provided general insight into variants present in primates. In addition, array-based approaches have supported those findings with identification of primate copy-number variants (CNVs) [Gazave et al., 2011, Gokcumen et al., 2011, Lee et al., 2008, Perry et al., 2007]. However, despite progress in assessing SNPs in primates [Auton et al., 2012, Locke et al., 2011, Prüfer et al., 2012, Yan et al., 2011], at the time we started this project, no other study had focused on inter- and intraspecies SVs.

In this chapter I will present our efforts to build comprehensive SV maps in *Pan troglodytes* (chimpanzee), *Pongo abelii* (orangutan), and *Macaca mulatta* (rhesus macaque), followed by SV formation mechanism assessment in each species in order to obtain a deeper evolutionary insight on different SV landscapes.

4.2 Structural variant differences in chimpanzee, orangutan and rhesus macaque

In order to perform polymorphic MEI discovery in non-human primate genomes, we used the chimpanzee, orangutan and rhesus macaque samples. We used 101 bp Illumina paired-end DNA reads with the average sequencing coverage ranging from 15x to 28x (Appendix A, Table A.8). Such coverage is estimated to result in the identification of 70–80% of deletion polymorphisms with >90% accuracy [Mills et al., 2011, Sudmant et al., 2010]. When combining our deletion, duplication, and MEI sets, we inferred a total of 6,947, 9,481, and 22,027 SVs in these species (Appendix A, Figure A.14, upper panel).

Similarly to the 1000GP [Mills et al., 2011], which used integrative approach to detect SVs in human, we applied different algorithms to construct SV maps in non-human primate species. The approaches we used looked for different signatures, such as (1) discordant mapping of paired-reads, (2) splitread support for breakpoint identification and (3) read-depth indication of copy number change, in order to detect CNVs. A combination of three independent available computational tools: DELLY [Rausch et al., 2012b], GenomeSTRiP [Handsaker et al., 2011] and CNVnator [Abyzov et al., 2011] were used to construct comprehensive datasets (see Methods). Using this strategy, we have successfully identified 2,680, 4,983, and 3,905 polymorphic deletions and inferred 1,499, 1,095, and 807 polymorphic duplications (Table 4.1) in chimpanzees, orangutans, and macaques, respectively. Fixed duplications present in all five individuals per species and absent from the corresponding reference genome were also identified, with 1,910 duplications in chimpanzees, 540 in orangutans, and 625 in rhesus macaques. Of all the predicted SVs, we were able to map ∼51% of all deletions and ∼18% of all duplications at breakpoint resolution.

In addition to identifying deletions and duplications, we also investigated polymorphic MEIs in these species (presented in Chapter 2). Non-reference MEIs were detected based on deletion and duplication datasets and subsequently analyzed separately from other CNVs. Although excisions of a mobile element is essentially non-existent, many of the deletions detected in non-human species emerged mechanistically through an MEI-mediated process. Since every deletion is detected compared to the corresponding reference genome, a MEI deletion is actually detected as an in the reference and subsequently annotated as a 'reference MEI'.

We also looked at novel MEIs not present in the reference genomes, but rather exclusive to our sample. Using the TEA tool [Lee et al., 2012], we mapped 764, 2,548, and 15,566 non-reference MEIs in chimpanzee, orangutan, and rhesus macaque, respectively (see Methods). The reference and novel (non-reference) transposable elements, which we inferred to be polymorphically absent/present in some individuals, consist of 858, 2,863, and 16,690 mobile element insertions in chimpanzees, orangutans, and macaques, respectively (Table 4.1).

TABLE 4.1: Non-human primates genome sequencing and SV detection information. Chimpanzee, orangutan and rhesus macaque sequencing and mapping details are listed: raw bases sequenced, successfully mapped bases, mean and total coverage; as well as number of polymorphic deletions, polymorphic and fixed duplication, novel and reference MEIs detected.

	Chimpanzee	Orangutan	Rhesus macaque
Total raw bases (Gb)	358.94	332.43	299.37
Total mapped bases (%)	82.59	79.44	80.34
Mean coverage per species	19X	17X	17X
Total coverage per species	96X	86X	85X
Polymorphic deletions*	2680	4983	3905
Polymorphic duplications*	1499	1095	807
Fixed unannotated duplications*	1910	540	625
Novel polymorphic MEI insertions ('non-reference MEI')	764	2548	15566
Polymorphic MEIs ('reference MEI')	94	315	1124

*dataset excluding reference MEIs

To assess the quality of our deletion callset, we verified 42 of 50 randomly sampled variant sites using PCR (Appendix A, Figure A.15 A). As we also investigated polymorphic mobile element insertions [Lee et al., 2012] in these species, we validated 42 of 49 (86%) randomly selected unique MEIs by PCR (Appendix A, Figure A.15 B). The validations were performed by using forward and reverse primers outside of the putative deletion, whereas MEI presence/absence was confirmed via the PCR band size.

4.3 *De novo* SV formation mechanisms in non-human primates

Of all predicted CNVs, we were able to map on average 51% of all deletions and 18% of all duplications at breakpoint resolution. This dataset was used to predict SV formation mechanisms in chimpanzees, orangutan and rhesus macaque and to distinguish MEIs, nonallelic homologous recombination (NAHR), variable number of tandem repeat (VNTR) expansion or contraction, and nonhomology-associated rearrangements (such as nonhomologous end joining (NHEJ) or microhomology-mediated break-induced replication(MMBIR)).

Our analysis of *de novo* SV mechanism formation revealed markedly higher MEI activity in rhesus macaque compared to the great apes (Figure 4.1, details described in Chapter 2). Furthermore, we noted striking differences in the activity of NAHR events between the great apes and macaques

(Figure 4.1). In rhesus macaque, only 2% of all SVs were inferred to be formed by NAHR compared with 28% of the chimpanzee and orangutan SVs ($P<2.2\text{x}10^{-16}$; two-sided Fisher's exact test). Based on previous SV studies performed on the human genome [Kidd et al., 2010, Lam et al., 2010, Mills et al., 2011], we conclude that there is a similar rate of NAHR-based SV formation throughout great ape lineage, including humans (22%-28% of human SVs emerge due to NAHR mechanism).

Each SV formation mechanism tends to be associated with specific size spectra [Kidd et al., 2010, Lam et al., 2010, Mills et al., 2011], indicating that observed differences in formation mechanisms reflect SV size variation. On average, in non-human primates, NAHR-mediated SVs tend to be larger than the size of NHR-mediated SVs (Appendix A, Figure A.16). The mean sizes of NAHR events were estimated to be 16.5 kb in chimpanzee, 7.4 kb in orangutan, and 11.3 kb in rhesus macaque, whereas NHR-associated events were predicted to be shorter with 7.8 kb, 5.7 kb and 3.1 kb in chimpanzee, orangutan and rhesus macaque, respectively. To investigate whether NAHR events are significantly larger than expected based on random mechanism assignments, we performed a Monte Carlo simulation-based approach. The total amount of genomic sequence occupied by NAHR was 11.6 Mb in chimpanzee, 12.7 Mb in orangutan and 4.4 Mb in rhesus macaque. In comparison, NHR events occupied 5.8 Mb, 8.9 Mb and 6.6 Mb in chimpanzee, orangutan and rhesus macaque, respectively. These results indicate the marked excess of NAHR events in the great apes compared to rhesus macaque. We performed 1000 permutations in each primate species, by keeping SV size assignments constant and permuting the mechanism assignments. Subsequently, we calculated the total amount of genomic sequence occupied by randomly-assigned NAHR-labeled and NHR-labeled SVs in each iteration. Our observation confirmed that NAHR-mediated SVs were larger than SVs formed by other mechanisms in all three primate species ($P<0.001$, $P=0.037$ and $P<0.001$ in chimpanzee, orangutan, and rhesus macaque, respectively; empirically calculated P-values based on permutation). NHR-mediated SVs did not display a trend towards larger SVs ($P=0.41$, $P=0.64$ and $P=0.99$ in chimpanzee, orangutan, and rhesus macaque, respectively; empirically calculated pvalues). Additionally, we randomly picked 20% out of all NAHR-mediated SVs (to account for the 5-fold difference between great apes and rhesus macaque) and calculated the total size of sequence occupied. Both chimpanzee and orangutan displayed a smaller genomic impact of NAHR-mediated SVs than rhesus macaque ($P<0.005$; permutation-based empirical P-value), with an average of 2.3 Mb in chimpanzee and 2.5 Mb of sequence occupied in orangutan, compared to the 4.4 Mb that are occupied by NAHR-mediated SVs in the macaque.

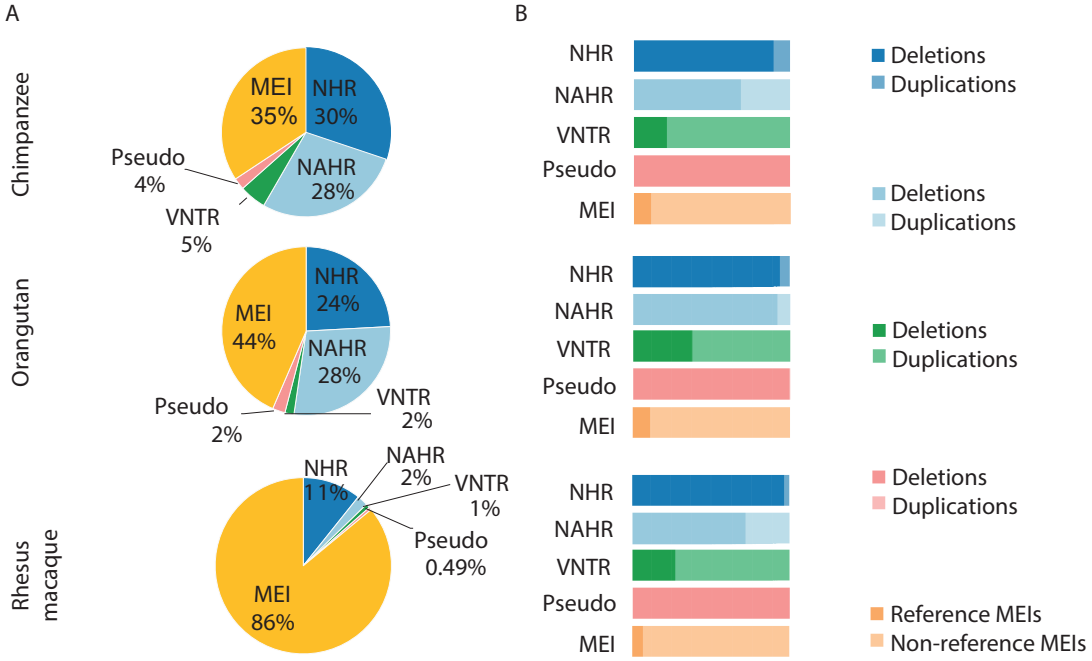

FIGURE 4.1: *De novo* SV formation mechanism distribution in non-human primates: chimpanzee, orangutan and rhesus macaque (top to bottom). (A) Proportion of SV formation mechanisms in each species. (B) Breakdown of SV type contribution to each mechanism: deletions, duplications, reference and non-reference mobile element insertions (MEIs). NHR = non-homologous rearrangement; NAHR = non-allelic homologous recombination; VNTR = variable number of tandem repeats; Pseudo = pseudogene; MEI = mobile element insertion.

4.4 Comparison of SV formation mechanisms rates between species

Under the assumption that the numbers of observed SNPs and SVs should correlate, as described in Chapter 2, we inspected whether the number of detected SNPs and NAHR in non-human primates correlate or not. Indeed, when looking at the number of NHR-mediated mechanism in all individuals and the number of predicted SNPs, we observed a strong correlation between the number of SNPs and the number of non-homology-associated rearrangement (r^2 value = 0.98; Figure 4.2). A weaker correlation or no correlation was observed between SNPs and NAHR events (r^2 = ~0), further supporting the notion that NAHR formation rates have changed considerably in recent primate evolution. The same correlation analysis was performed for SNPs and MEIs detected in non-human primate species and was described in Chapter 2.

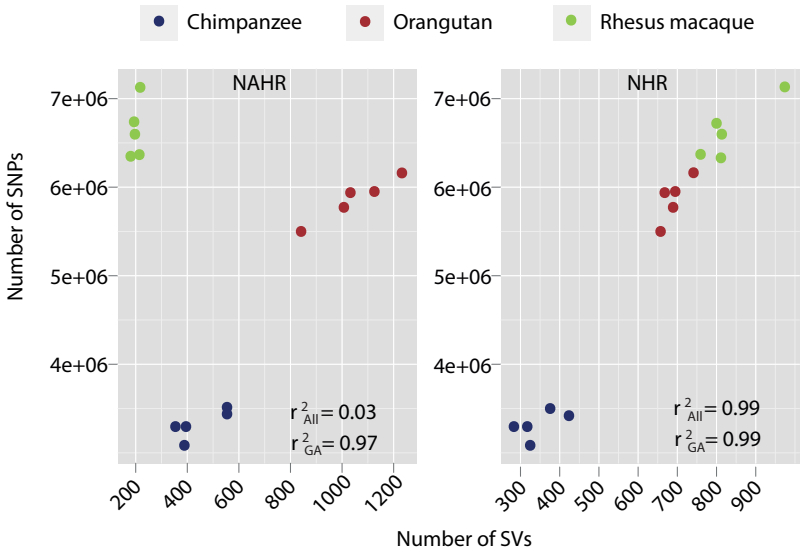

FIGURE 4.2: Correlation in the abundance of SNPs and SVs formed by different mechanisms. Dots represent different samples. r^2_{All} = Pearson correlation coefficient for all three studied primate species; r^2_{GA} = Pearson correlation coefficient for studied great ape species.

Additionally, we also determined the ancestral state of all deletions and duplications relative to the reference genome and mapped them at the nucleotide resolution [Lam et al., 2010, Mills et al., 2011]. Essentially, for every SV two alleles were designed: reference (no deletion/duplication) and alternative (deleted/duplicated allele) and both of them were aligned onto syntenic net alignments of other species (for example human reference and alternative alleles were aligned onto chimpanzee, orangutan, macaque and marmoset syntenic regions downloaded from the UCSC Genome Browser). In the case of a deletion, if the alternative allele maps with better sequence identity and length onto one of four different genomes, the event was rectified as an 'insertion' in the reference genome, rather than a deletion in a sample (see Methods for details). These insertions were subjected to BLAT alignments [Kent, 2002] in order to find a donor locus. We refer to all the ancestral insertions, for which we could delineate the source locus, as duplicative insertions. The analyses we performed showed an excess of intrachromosomal over interchromosomal duplicative insertions (i.e., SVs arising from the insertion of duplicated sequence) in great apes and a marked depletion of intrachromosomal duplicative insertions in macaques (Figure 4.3 left panel; P<0.01, two-sided Fisher's exact test). When looking at the NAHR compared to the other mechanisms, we observe a similar effect: NAHR seems to be the dominating mechanism for duplicative insertions in the great apes compared to macaque, where other mechanisms surpass NAHR-mediated formation (Figure 4.3, right panel).

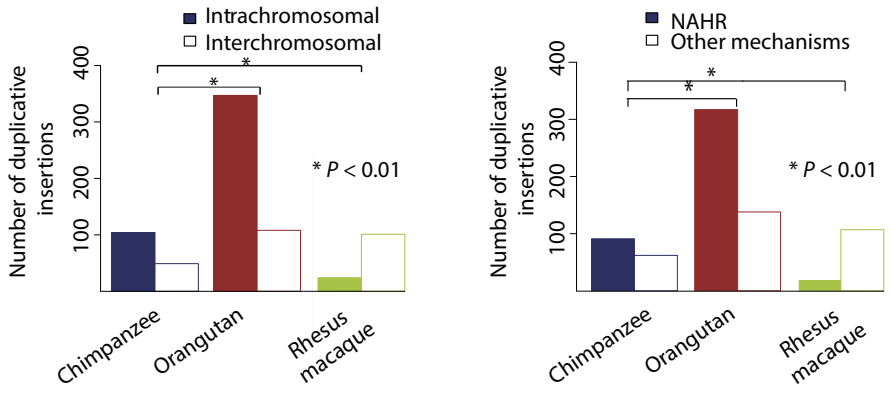

FIGURE 4.3: Breakdown of intrachromosomal and interchromosomal duplicative insertions (P value computed using a two-sided Fisher's exact test) and breakdown of duplicative insertions mediated by NAHR and other mechanisms.

A high rate of NAHR events can be explained by recent burst of SDs in the great apes [Marques-Bonet et al., 2009], where segmental duplications (SDs) can act as mediators of NAHR [Hastings et al., 2009b]. We assessed comparable SD dataset (see Methods) in the non-human primate species, and showed that SDs comprise 4.7-5.4% of the genomes of great apes compared with only 1.6% of the macaque genome (i.e., 2.6- to 3.4-fold relative increase; P<0.0008, two-sided Fisher's exact test).

4.5 Discussion

Massively parallel sequencing enabled the creation of SV maps not just in human, but also in other species. In this study we have provided comprehensive SV maps for chimpanzee, rhesus macaque and orangutan genomes, and have shown differences in their formation mechanisms. In the recent primate history, specifically during the last 25 Myr, MEI and NAHR activity in particular seems to have gone through rapid evolutionary change.

Our analyses show a marked increase of NAHR-mediated SVs in orangutans and chimpanzees. In all species analyzed in our study, NAHR-mediated SVs were, on average, larger than other SV classes (NHR, MEI and VNTR). The observed increase in the number of NAHR-associated SVs in great apes, compared to rhesus macaque, demonstrates a high nucleotide-level impact of this SV type in these species. Apart from being larger, NAHR events often intersect genes and have been associated with various genomic disorders [Stankiewicz and Lupski, 2002, Weischenfeldt et al., 2013]. We propose that fixed NAHR and MEI events will likely further accumulate differentially between great apes and OWM lineage, and thus will continue to contribute to their diversification. Based on our results, it is more likely that the emerging variants in either

chimpanzee or orangutan will continue to form through NAHR-mediated mechanism compared to rhesus macaque, where retrotransposons are the most dominating formation mechanism.

Among the great apes, the burst of SDs [Marques-Bonet et al., 2009] linked to the NAHR mechanism, as well as abundance of MEIs in the OWM lineage compared with the great ape lineage [Locke et al., 2011] have been previously reported. Our results confirm these observations by providing strong evidence for lineage-specific activities of NAHR and retrotransposition influencing species variant landscapes at the genome-wide scale.

Chapter 5

Comparison of germline and somatic SVs in human

This chapter will focus on *de novo* structural variant formation mechanisms in healthy human individuals as well as in the context of disease. Germline SV formation mechanisms were assessed based on previously published data for the purpose of the following review written by a former postdoc in Korbel group:

> Onishi-Seebacher M. and Korbel J. O. Challenges in studying genomic structural variant formation mechanisms: the short-read dilemma and beyond. *BioEssays: news and reviews in molecular, cellular and developmental biology*, 33(11):840–50, November 2011 (Acknowledgments: "We thank ... and Jelena Tica for assistance with formation mechanism analysis...")

Furthermore, germline SVs will be compared with complex somatic alterations and the formation mechanisms found in SHH medulloblastoma patients with Li-Fraumeni syndrome (LFS). This study was a part of International Cancer Genome Consortium (ICGC) Pediatric Brain Tumor Research Project and a collaboration with Peter Lichter's and Stefan M. Pfister's groups at DKFZ, Germany. Most of the results were published in the following research article:

> Rausch T.*, Jones D.*, Zapatka M.*, Stütz A.*, Zichner T., Weischenfeldt J., Jäger N., Remke M., Shih D., Northcott P., Pfaff E., Tica J. et al. Genome Sequencing of Pediatric Medulloblastoma Links Catastrophic DNA Rearrangements with TP53 Mutations. *Cell*, 148(1-2):59–71, January 2012

Contribution

I performed *de novo* structural variant mechanism classification analyses presented in this chapter. Based on my results, Megumi Onishi-Seebacher performed the formation mechanism analysis and generated the original figure showing mechanism classification for previously published human structural variants. For the medulloblastoma project, Tobias Rausch implemented the structural variant (SV) prediction tool DELLY to generate SV lists. Thomas Zichner, Tobias Rausch and Jan Korbel provided significant feedback on results presented in this chapter.

5.1 Motivation and background

Advancements in sequencing technology have enabled faster, more reliable and precise identification of SVs. Following the development of NGS approaches, the number of computational tools analyzing NGS data has constantly been increasing. This has led to large amounts of data in a need of correct characterization. The nature of the mechanisms involved in *de novo* SV formation are important considering that this process occurs continuously during a life of an organism. Inference of variant formation mechanisms was hindered in the past due to technological limitations. Today, the advent of algorithms able to predict a wide variety of SVs across a broad range of sizes at single-nucleotide resolution has afforded better variant characterization and even mechanism classification.

The mechanism through which a SV is formed is an important characteristic that can help decipher the true functional impact of a single or multiple complex rearrangements. The underlying mechanism of genome breakage can explain whether a SV is a result of homology based recombination, mobile element insertion, replication errors or double-stranded DNA breaks. In a disease context, the benefit of understanding variant formation lies in the possibility of deciphering the origin of potentially harmful, disease causing SVs and differentiating them from other, less damaging variants. Even studying how neutral variants in a healthy individual form can help to disentangle variant evolution and specific differences across species and individuals.

Throughout this chapter I will describe *de novo* SV formation mechanisms based on previously published structural variation datasets: Mills et al. [2011], Conrad et al. [2010], Kidd et al. [2010] and Lam et al. [2010], and put them in a relation to SV formation mechanisms we observed in childhood brain tumor, medulloblastoma [Rausch et al., 2012a].

5.2 SV formation mechanisms in human germline

As previously mentioned, identifying precise variant breakpoints is crucial for the reliable characterization of each event. In the last five years, many studies have used different approaches to determine sequence breakpoints. For instance, Conrad et al. [2010] used hybridization-based DNA capture and 454 sequencing to sequence copy-number variants (CNVs), focusing on deletions. For 315 deletions discovered at that time, the reconstruction of the molecular events was not possible, although certain microhomology and insertion signatures were identified. Kidd et al. [2010] looked at 1,054 structural variants with the breakpoint resolution based on capillary end sequencing of 13.8 million fosmid clones from 17 human genomes. Predominant mechanisms of origin were shown to be microhomology-mediated processes involving short (2–20 bp) stretches of homologous sequence (28%), nonallelic homologous recombination (22%), and L1 retrotransposition (19%). In the same year, Lam et al. [2010] developed a tool named BreakSeq which finally allowed to classify SVs based on their formation mechanism: MEIs, nonallelic homologous recombination (NAHR), variable number of tandem repeat (VNTR) expansion or contraction, and nonhomology-associated rearrangements (such as nonhomologous end joining (NHEJ) or microhomology-mediated break-induced replication(MMBIR)). The algorithm was developed and tested on a non-redundant set of 1,889 previously published SVs. Almost half (45%) of the SVs were shown to originate through NHR processes, whereas 28% involved homology (NAHR), 21% were MEIs, 5% involved VNTRs and 2% were ambiguous. A year later, the BreakSeq analysis was expanded to a set of 185 human genomes with 22,025 deletions and 6,000 additional SVs, including insertions and tandem duplications [Mills et al., 2011]. The results of this study were consistent with previous findings, confirming the NHR as the dominating deletion mechanism, and MEI as the dominating mechanism of insertion.

To confirm the aforementioned findings and show the dominance of NHR mechanism in the human germline, we collected all the predicted deletions with breakpoint resolution from these four studies and performed mechanism classification using BreakSeq. Our findings revealed that NHR is indeed the most dominating mechanism of SV formation in the human genome, independent of the discovery method (Figure 5.1). Although it seems that most SVs do not require homology to form, this observation might change in the future, due to technological improvements. For example, repetitive and mobile elements are currently challenging to study, but this obstacle might be solved with the use of long reads able to span the whole insertion or repetitive sequence. Additionally, many of the SVs studied in the human genome are currently relatively simple, with more complex rearrangements being ignored. In fact, Chiang et al. [2012] showed high incidence of complex rearrangements (19.2%) in germline, indicating that this kind of mechanism is not exclusive to cancer cells.

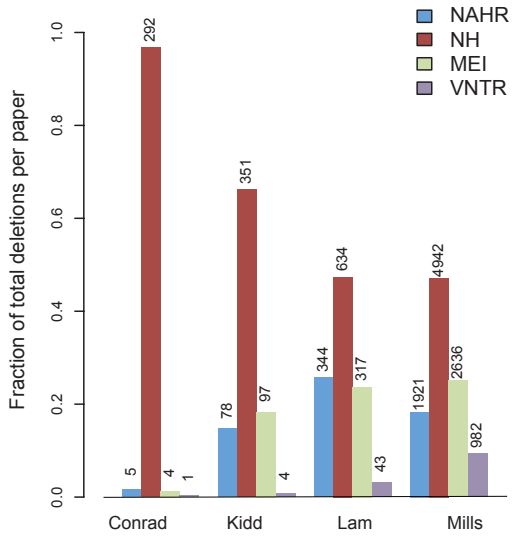

FIGURE 5.1: Comparison of different mechanisms inferred in previously published human deletions [Conrad et al., 2010, Kidd et al., 2010, Lam et al., 2010, Mills et al., 2011]. The total number of deletions is indicated above each bar. Distributions of *de novo* variant formation is relatively similar in all four datasets. Conrad dataset has a slightly higher degree of NHR events due to the array-based approach used to detect variants. All mechanisms were detected using BreakSeq tool [Lam et al., 2010], and subset of NHR undetectable by BreakSeq was assessed by identification of 50 bp flanking homologous sequences with 85% identity.

5.3 Formation of complex rearrangements in medulloblastoma

In contrast to healthy human individuals, genomes of common and rare diseases often harbor very specific and more complex variants. One example of such a disease is a childhood brain tumor - medulloblastoma, which causes the highest cancer-related mortality in children. As described in Chapter 1, medulloblastoma is one of the recognized Li-Fraumeni syndrome (LFS) tumors [Li and Fraumeni JR, 1969]. The LFS patients often carry heterozygous *TP53* germline mutation, which affects the p53 tumor suppressor [Malkin et al., 1990].

We have investigated four Sonic-Hedgehog medulloblastoma (SHH-MB) patients, which have tumors arising from the part of a brain called cerebellum [Bühren et al., 2000]. All the analyses were performed on tumor and paired normal tissues from the same patient, including whole-genome paired-end sequencing and subsequent variant discovery (patient information and sequencing details can be found in Appendix A, Table A.9). One of the major findings revealed a frequent incidence of massive genomic rearrangements localized on individual chromosomes. These findings are consistent with previously proposed model for tumorigenesis, termed chromothripsis [Stephens et al., 2011]. Moreover, this single catastrophic event usually involves shattering of

one or a few chromosomes followed by random assembly of the fragmented pieces, a mechanism fundamentally different from the progressive acquisition of mutations [Knudson, 1971, Nowell, 1976, Stratton et al., 2009]. In this study we have discovered a novel link between chromothripsis and *TP53* mutations, providing a possible explanation of how p53 status can influence massive rearrangements.

Following whole-genome paired-end sequencing, we discovered large-scale structural variants using DELLY software [Rausch et al., 2012b]. The variation landscape consisted of deletions, tandem duplication, inversions and interchromosomal rearrangements consistent with translocations paired-end signatures. All variants were subjected to filtering based on quality (see Methods) and overlap with previously discovered 1000 Genomes Project variants as well as variants present in the paired control tissue, to account for germline-specific SVs. Tumor-specific variants were also differentiated from those involved in the chromothripsis catastrophic event present in only a few chromosomes per patient. Every variant identified by paired-reads was additionally fine-mapped using a splitread approach, which also provided the possibility to investigate SV formation mechanisms (Table 5.1).

Table 5.1: Structural variants with breakpoint resolution discovered in SHH medulloblastoma patients used for mechanism classification.

Sample	**SV type**	**Chromothripsis**	**Tumor-specific**	**Germline**
LFS-MB1	Deletion	7	14	1343
	Tandem duplication	0	8	342
	Inversion	4	3	103
	Interchromosomal	6	0	0
LFS-MB2	Deletion	5	4	1468
	Tandem duplication	2	5	287
	Inversion	9	10	118
	Interchromosomal	8	0	0
LFS-MB4	Deletion	7	11	1372
	Tandem duplication	0	0	261
	Inversion	15	17	107
	Interchromosomal	0	0	0

Analysis of breakpoint sequence signatures of the three datasets: (1) germline-specific, (2) tumor-specific not related to chromothripsis and (3) chromothripsis-related variants was performed using the BreakSeq tool [Lam et al., 2010]). Germline specific-variants showed formation mechanism profiles consistent with the analysis we performed on previously published data, with NHR being the most dominant mechanism. We observed higher numbers of MEI-related formation mechanisms than previously reported, occurring probably due to better annotation of

updated reference genome build used for this analyses (hg19). Chromothripsis-related variants revealed short microhomology tracts (2-4 bp), compatible with nonhomologous end-joining (NHEJ)-mediated double-strand repair, or microhomology-mediated break-induced replication (MMBIR) [Hastings et al., 2009a, Lee et al., 2007] (Figure 5.2; Appendix A, Table A.10, Table A.11). In a few cases, we detected short non-template insertions at the breakpoint junctions (Appendix A, Table A.12). NHEJ-mediated repair during chromothripsis seems to be the most plausible explanation for repair of shattered DNA fragments, as we have not observed templated insertions, commonly related to the replication-based mechanisms (MMBIR). Moreover, in some cases, during the repair of shattered pieces, circular, so called 'double-minute chromosomes' can be formed (detected in LFS-MB1, LFS-MB2 and LFS-MB3) and typically carry oncogenes (such as *MYCN* and *GLI2* in LFS-MB4) (for details, see Rausch et al. [2012a]). The complexity of massive-rearrangements observed in these patients, together with formation of double-minute chromosomes, makes the MMBIR repair mechanism improbable in a chromothripsis model.

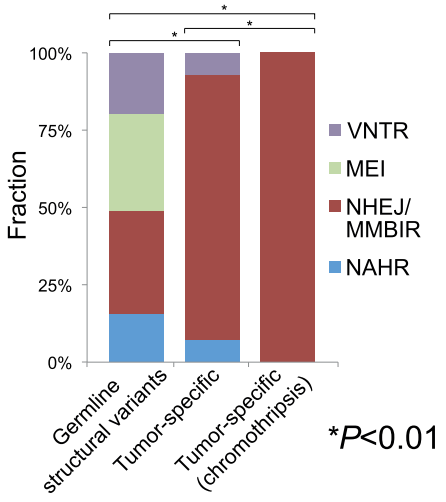

FIGURE 5.2: Rearrangement formation mechanisms analysis. Polymorphic genomic structural variants detected in the germline are shown for comparison. *P* values, indicating significant differences between the distributions of inferred formation mechanisms, are based on Chi-square tests.

5.4 Discussion

Advancements in DNA sequencing technology have had a major impact on the genomics research community. Massively parallel next-generation whole-genome sequencing has allowed faster and reliable sequencing of multiple genomes at the same time. This has further enabled the characterization of variants present in an individual compared to the reference genome, in one population

relative to another or in a pathological versus physiological context. Until recently, the origin of variants was not simple to ascertain due to the technological limitations of identifying precise breakpoints. However, splitread approaches have provided facilitated the determination of SV breakpoints accurately at single-nucleotide resolution, an essential step prior to classifying each variant based on their formation mechanism.

Our analyses of SVs detected in germline [Conrad et al., 2010, Kidd et al., 2010, Lam et al., 2010, Mills et al., 2011] support previous reports of variant origin: almost half of every deletion dataset is predicted to form through NHR mechanism involving either NHEJ or MMBIR repair. With computational tools able to accurately resolve SVs forming in repetitive areas and better annotated reference builds, the number of predicted MEI-related mechanisms and VNTRs increased in Kidd et al. [2010], Lam et al. [2010], Mills et al. [2011] compared to Conrad et al. [2010]. Some of the NHR bias in Conrad et al. [2010] study probably occurred due to the partial array-based approach they used for variant discovery, resulting in variants with no defined breakpoints. As sequencing and algorithms used to analyze the data improve, the observed distribution of mechanisms present in the human germline might change. Long read technology, for instance, can span some of the variants smaller than 8 kb and can help identify the exact process of the variant origin.

In contrast to germline studies, our analyses of rearrangement breakpoints in SHH-MB support a model of massive DNA double strand breaks [Stephens et al., 2011], followed by NHEJ-mediated repair. Similar to NHEJ, replication-based MMBIR repair mechanism can also result in complex alterations with multiple breakpoints [Hastings et al., 2009a]. MMBIR mechanism is often associated with the presence of templated insertions at breakpoint junctions, as well as longer tracts of microhomology compared to NHEJ [Ottaviani et al., 2014]. The microhomology in chromothripsis-related rearrangements, if present, is 2-4 bp long, which is consistent with canonical NHEJ repair [Lieber, 2010]. The lack of templated insertions, as well as short microhomologies observed at the breakpoint junctions, led us to believe that complex alterations in chromothripsis get repaired by NHEJ. Although the reason why chromothripsis occurs as well as the responsible underlying mechanisms involved are not yet fully understood, there are several theories that might explain this phenomenon. One of the most important characteristics of chromothripsis is the occurrence of many complex localized chromosomal rearrangements, indicating that DNA needs to be as condensed as possible for shattering to occur in a single chromosome (i.e. in mitosis) [Forment et al., 2012, Maher and Wilson, 2012]. Micronuclei formation [Crasta et al., 2012, Forment et al., 2012] is accepted as the most probable model of chromothripsis. During cell proliferation, mitotic errors and defective chromosomal segregation arise, causing a single or very few chromosomes to be enclosed and isolated in micronuclei. These chromosomes are prone to slower and defective DNA replication, resulting in broken chromosomes. Other theories concerning the emergence of chromothripsis include ionizing radiation during mitosis

[Maher and Wilson, 2012] and telomere attrition (shortening) [Tubio and Estivill, 2011]. Regarding the possible repair mechanisms of broken chromosome fragments, breakpoint analyses indicate that repair occurs by either NHEJ [Kloosterman et al., 2011, 2012] or replication-based mechanisms (FoSTeS/MMBIR) [Liu et al., 2011]. Based on our analyses, these breaks are likely to be repaired by low-fidelity, error prone NHEJ, which plays a greater role when levels of p53 are reduced [Dahm-Daphi et al., 2005]. The ongoing technological improvements will undoubtedly allow better and more correct characterization of variants and complex rearrangements, and it will be interesting to see how much our findings will deviate from future predictions and mechanism classification analyses.

Chapter 6

Summary, conclusions and future directions

The work presented in this Thesis addresses several aspects of genomic variations, their origin, mechanism of formation and potential impact on the genomic landscape. In Chapter 2, a comprehensive map of mobile elements in non-human primates (chimpanzee, orangutan and rhesus macaque) was presented, followed by Chapter 3 where the analyses were expanded on specific L1 elements mobilizing additional unique sequence (L1-mediated 3' transductions) and intronless gene duplications (gene retrocopies), both mediated by the retrotransposition. Chapter 4 focuses on all other *de novo* SV formation mechanism in non-human primates, excluding MEIs. Finally, Chapter 5 addresses SV mechanism formation in healthy human individuals as well as formation of massive rearrangements in a specific context of a pediatric brain tumor.

In general, advancements in DNA sequencing approaches had a major impact on work presented throughout this Thesis. Without defining precise nucleotide breakpoints of each variant, analyses such as SV mechanism classification would not have been possible. Additionally, until recently, MEIs presented a challenge for reliable identification, due to their repetitiveness, high numbers and sequence similarity. Improvements of SV and MEI detection algorithms and development of new, more reliable approaches allows accurate characterization of all genomic variants, ultimately leading to better understanding of biological processes and phenotypic variation. For instance, long read technology, although relatively new, already enabled discovery and characterization of complex and repetitive variants. In this Thesis it was presented in the context of L1-mediated transductions, a variant class hard to completely resolve using traditional short-read sequences.

Mobile element have a markedly higher activity in rhesus macaque compared to the great apes.

As indicated in Chapter 1, the major contributors of transposon activity in mammals are long and short interspersed elements: *Alu*, SVA and L1 [Stewart et al., 2011]. Throughout evolution they have been remarkably successful and therefore they currently occupy almost 50% of the mammalian genomes [Lander et al., 2001]. Although many of them are in an inactive form, due to the sequence deterioration and accumulation of deleterious changes, some of the elements still remain active. In fact, Hancks and Kazazian [2012] reviewed 96 (25 L1, 60 *Alu*, 7 SVA, or 4 polyA) retrotransposition events in the literature resulting in single-gene diseases.

Reference genome assemblies of the chimpanzee [The Chimpanzee Sequencing and Analysis Consortium, 2005], orangutan [Locke et al., 2011] and rhesus macaque [Rhesus Macaque Genome Sequencing and Analysis Consortium, 2007], together with the existence of the human reference genome [Lander et al., 2001] served as a valuable resource for variant discovery and annotation in general. Since all mammalian genome have a high content of mobile transposable elements, in-depth inspection of MEIs in non-human primates and human was needed to comprehensively understand their impact on genomic landscapes.

Our study presented in Chapter 2 provided a comprehensive MEI dataset consisting of polymorphic and fixed, recent species-specific MEIs detected in chimpanzee, orangutan and rhesus macaque. In rhesus macaque, we observed a notable excess of *Alu* activity compared with the great apes, exclusively due to the *AluMacYa3* rhesus-specific subfamily [Liu et al., 2009], which evolved after the divergence of great apes-human lineage from rhesus macaque branch ∼25 Mya. Similar to human, we have observed that about 15% of all SVs in the great apes forming by MEI-related mechanism [Mills et al., 2011], indicating similar rate of MEI insertions in the great apes-human lineage. However, in orangutan, *Alu* elements are found to be quiescent [Gokcumen et al., 2013, Locke et al., 2011] which is in contrast with human, where *Alu* represents the most active human mobile element [Mills et al., 2011, Stewart et al., 2011]. Therefore, our findings suggest a rapid turnover of active transposable DNA sequences, leading to a divergent set of species-specific MEIs. We believe that those species-specific MEIs will likely further accumulate differentially primate genomes and promote additional diversification in great ape and macaque lineages.

In the great apes specifically and in the human genome, hominid-specific composite SVA elements also continue to evolve. They are ∼2 kb long, non-coding RNAs mobilized by L1 *in trans* [Wang et al., 2005] with different subfamilies active in each species. Recently, LAVA elements, closely related to the hominoid-specific SVA element, were discovered and characterized in gibbon [Carbone et al., 2012]. Both LAVA and SVA share the 'VA' part (VNTR and *Alu*-like sequence). However, instead of the SVA-specific SINE-R region, LAVA elements contain unique sequence

sections as well as ancient *Alu* and L1 sequence. Existence of such element further supports unique, species-specific independent genome evolution.

Apart from MEIs in the context of evolution, they are particularly interesting in the context of disease biology. As such, it has been shown that somatic MEIs can insert into genes, promoting cancer [Miki et al., 1992]. However, whether somatic ME insertions are cause or consequence of the disease is yet to be uncovered. Finally, due to their complexity, it is worth to note that MEI regulation is a growing field of research, including studies on different regulation stages, such as transcription, post-transcription and post-translation [Hancks and Kazazian, 2012].

L1-mediated 3' transduction rates differ between species.

Until recently, little was known about MEI evolutionary influence within non-human primate genomes in comparison to human. Although differences in MEI numbers and their activity have been observed in primates [Gokcumen et al., 2013, Locke et al., 2011], the extent and further characterization of such elements has been lacking.

In order to inspect MEI-related SVs further in depth, we have developed TIGER, a novel approach to detect L1-mediated transduction events in germline. Our aim was to identify and characterize all L1-mediated transductions in chimpanzee, rhesus macaque and orangutan. Although analyses we performed were not discriminating against 5' transductions, our results indicated no such event. The highest number of L1-mediated transductions was observed in orangutan, and the smallest in chimpanzee, indicating different evolutionary dynamics of L1 elements in primates. As sequencing coverage in each species was comparable, we concluded that sequencing itself could not affect subsequent variant identification and observed difference in numbers of identified events. Further support came from the L1-mediated 3' transduction rates we calculated based on total number of L1 elements, coverage and portion of L1 elements detected as transduction events. Due to the differences in predicted species-specific rates, we hypothesized that different subfamilies might drive overall retrotransposition in primates. Indeed, our subfamily analyses revealed that L1-TS sequence insertions in rhesus macaque occurred due to the activity of macaque-specific L1CRE element, whereas in orangutan primate-specific L1PA3 is mostly responsible for transduction and overall retrotransposition as well. In chimpanzee, chimpanzee-specific L1Pt seems to be dominant subfamily driving L1 retrotransposition. Based on our analysis, rhesus macaque and chimpanzee have significantly lower TS rates than orangutan, whereas observed TSs in humans are concordant with previous studies, when applying consistent parameters. However, our predicted L1-TS rate in human is still slightly lower than previously published rates [Helman et al., 2014, Kidd et al., 2010, Pickeral, 2000, Solyom et al., 2012b, Szak et al., 2003, Tubio et al., 2014, Xing et al., 2006]. There are several reasons why we observe this discrepancy. In brief, we looked at novel (non-reference) L1-TS germline insertions, whereas other studies mostly looked at reference or somatic L1-TSs. In addition, we used a more stringent approach requiring absence of low-complexity sequence stretches in

insertion loci. Another discrepancy might be the sole focus of our study, which are transductions translocating from one chromosome to another. Although we looked for presence of same chromosome transductions (both sources and targets located at the same chromosome), we have not observed such events in our dataset.

In an independent analysis, we looked at retrogene insertions (GRIPs) in non-human primates as another retrotransposition-mediated SV class. The highest number of GRIPs was observed in rhesus macaque and the smallest in orangutan, showing similar distribution as our *Alu* predictions in same species. As we had not observed such correlation with L1 elements, we concluded that L1 retrotransposition machinery is equally hijacked by both *Alu* elements and GRIPs.

In summary, we developed and will provide our tool TIGER as the first existing approach to detect L1-mediated 3' insertions in the germline to open up this important class of germline structural variation for population and disease studies. We hope this will enable further studies of polymorphic 3' transduction events and better characterization of such events. We foresee possible TIGER application to new evolutionary studies between species, as well as in cancer genomes to study germline transductions as a potential hereditary predisposing factor.

NAHR-mediated SVs are markedly increased in the great apes compared to rhesus macaque.

Recent advances in the application of massively parallel sequencing have not only enabled discovery of MEIs, but also other large-scale variants $\geq$50bp. Due to the lack of inter- and intra-species nucleotide-resolution maps until the time when I started my PhD, analyses of SV formation mechanisms have not been actively pursued in non-human primate species.

In our study we have provided comprehensive SV maps in previously mentioned chimpanzee, rhesus macaque and orangutan individuals, and showed differences in their formation mechanisms [Gokcumen et al., 2013]. In order to detect *de novo* SV formation mechanism, we needed to precisely define a single-nucleotide breakpoint. Such stringency is required not to cause any bias in differentiating four main SV mechanisms: NHR, NAHR, MEI and VNTR. The portion of SVs forming by NAHR-mediated mechanism in the great apes was discovered to be increased in comparison to NAHR-forming SVs in rhesus macaque. However, all NAHR-mediated SVs we analyzed were, on average, larger than other SV classes. In humans, the portion of SVs forming by NAHR is similar to the great apes observed in this study. In addition, such larger event were shown to often intersect genes and have been associated with various genomic disorders [Stankiewicz and Lupski, 2002, Weischenfeldt et al., 2013]. The high rate of NAHR in great apes may occur due to the burst of recent segmental duplications [Marques-Bonet et al., 2009], capable of mediating NAHR processes [Hastings et al., 2009b]. This shows that SV formation mechanism is closely related to the genome architecture of each individual or species in general.

We propose that fixed NAHR and MEI will likely further accumulate differentially through the activity of their polymorphic species-specific counterparts in each species. Accumulating differences will continue to contribute to the diversification of chimpanzee, orangutan and rhesus macaque. In the case of great apes this likely means that NAHR-derived SVs will continue to propagate faster, in comparison to rhesus macaque. As previously indicated, burst of species-specific *Alu* elements will promote further accumulation of short SVs in rhesus macaque in contrast to orangutan, where we observed quiescence of such short repeats.

Our study provided comparison of species ranging from the OWMs to the great apes. Due to limitations in technology, such large scale analyses including identification of different SV types and variant formation assessment, was not possible. Approaches we developed could be expanded to even more species, including NWM, e.g. marmosets, to get an even deeper insight into SV evolution and ancestral state of each variant, ultimately allowing us to fully reconstruct genome evolution and possible selective pressures acting upon it.

Difference in formation of germline and medulloblastoma-associated somatic SVs in the human genome.

In order to discover formation mechanism of human specific SVs, we employed similar approaches as presented previously in the non-human primate studies. Before us, many studies have already identified mechanisms through SVs form [Conrad et al., 2010, Kidd et al., 2010, Lam et al., 2010, Mills et al., 2011]. However, two of the studies Conrad et al. [2010], Kidd et al. [2010] were limited in such assessment due to the absence of algorithm providing reliable mechanism classification analysis. Lam et al. [2010] have provided such tool and both Lam et al. [2010] and Mills et al. [2011] have used it to infer mechanisms in their deletion dataset. In order to compare all four datasets, we repeated the whole analysis on deletion breakpoints predicted in each study. As previously reported in [Lam et al., 2010, Mills et al., 2011], almost half of every deletion dataset forms through NHR mechanism, followed by NAHR, MEI and VNTR. We observed some differences between each dataset, probably due to the approaches these studies used to generate breakpoint datasets [Onishi-Seebacher and Korbel, 2011]. Compared to the earliest study [Conrad et al., 2010], numbers of MEI- and VNTR-derived SVs increased in later studies [Kidd et al., 2010, Lam et al., 2010, Mills et al., 2011], indicating that improvement in experimental technologies as well as computational approaches used to analyze such data may, in future, alter results currently observed. With further development of approaches, complex rearrangements and repetitive sequences may become easier to resolve, ultimately leading to different distributions of SV formation mechanisms we observed in this and other studies.

In an independent study, we looked at formation of rearrangements in medulloblastoma brain tumor patients with a germline *TP53* mutation (Li-Fraumeni syndrome) [Rausch et al., 2012a]. Genomes we used harbor heterozygous *TP53* germline mutation, affecting p53 tumor suppressor [Malkin et al., 1990]. Our analyses of rearrangement breakpoints in a particular type of

medulloblastoma (Sonic-Hedgehog medulloblastoma, SHH-MB) revealed massive DNA double strand breaks, consistent with Stephens et al. [2011] findings of one-step catastrophic event termed chromothripsis. Such rearrangements were observed to form exclusively through non-homology mediated mechanism. As chromothripsis-related variants usually possessed short microhomology tracts in our dataset, we presumed that they formed through nonhomologous end-joining (NHEJ)-mediated double-strand repair, rather than microhomology-mediated break-induced replication (MMBIR). Although both mechanisms can result in complex alterations with multiple breakpoints [Hastings et al., 2009a], NHEJ is often associated with short microhomology tracts [Lieber, 2010]. In contrast, templated insertions are indicative of MMBIR mechanism, which we have not observed in our dataset.

As previously discussed, the real underlying mechanism of chromothripsis-related SVs is not yet fully understood, although there are several theories that might explain reason why such shattering occurs. Moreover, recent studies have found evidence for chromothripsis in different cancers [Magrangeas et al., 2011, Molenaar et al., 2012, Northcott et al., 2012b, Rausch et al., 2012a] as well as in germline [Chiang et al., 2012, Kloosterman et al., 2011].This further implies the need for correct identification and characterization of such unique, but complex event.

Ongoing development of both existing and emerging approaches will undoubtedly allow more reliable identification of simple and complex variants. Experimental, as well as computational improvements will provide more accurate datasets, easy to assemble and assess. Together with data standardization and integration, predictions emerging from such approaches will be of high quality, ultimately changing how we currently observe variants and their complexity.

Appendix A

Supplementary figures and tables

A.1 Supplementary information for Chapter 2

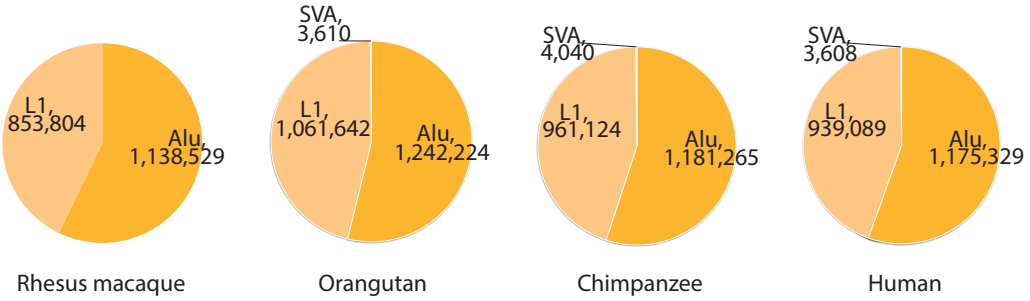

FIGURE A.1: Distributions of all *Alu*, L1 and SVA elements in human, chimpanzee, orangutan and rhesus macaque. Note that rhesus macaque has no SVA elements, as they emerged after the divergence of the great apes exclusively in great apes lineage.

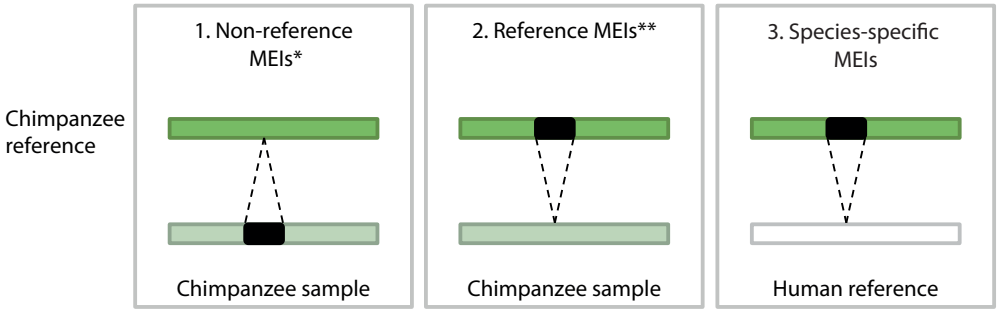

FIGURE A.2: Approaches to study MEIs in non-human primate genomes. (1) Non-reference polymorphic MEIs are discovered as insertions in the sample genome compared to the reference (*TEA tool [Lee et al., 2012]). (2) Reference polymorphic MEIs are identified as deletions in the sample, in the loci where reference possesses *Alu*, L1 or SVA element (**BreakSeq tool [Lam et al., 2010]). (3) Species-specific MEIs are fixed elements detected from whole-genome pairwise alignments and eliminating all shared MEIs (based on approach presented in Mills et al. [2007]).

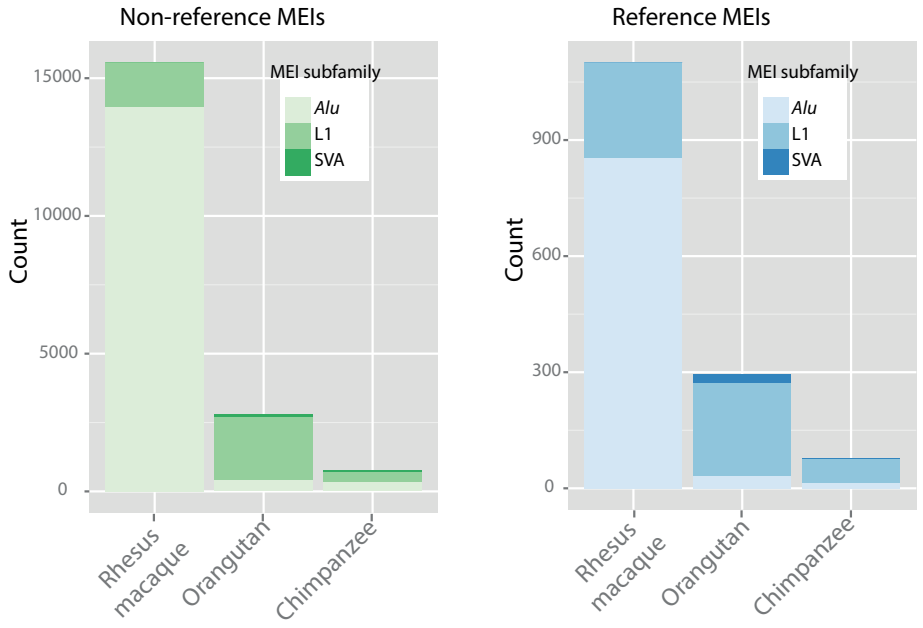

FIGURE A.3: Breakdown of separated datasets - non-reference and reference polymorphic mobile elements in rhesus macaque, orangutan and chimpanzee.

TABLE A.1: Target-site duplications (TSD) detected in non-human primates in three MEI datasets. Mean and median values are indicated (median in parentheses). 'Reference-fixed dataset' presents 'species-specific dataset' with TSD values calculated by the TSDfinder tool [Szak et al., 2002]. 'Reference-polymophic' TSD values are derived from the BreakSeq output [Lam et al., 2010]. Note that BreakSeq looks at microhomologies at the breakpoints of deletions and duplications for clues indicating their potential mechanism, which might not reflect MEI-specific TSD values. 'Novel-polymophic (non-reference) dataset' was generated with TEA [Lee et al., 2012] and the TSD values were extracted from the TEA output.

	Reference-fixed	**Reference-poymorphic**	**Novel-polymorphic**
Chimpanzee	14.08 (14)	8.14 (7)	12.56 (14)
Orangutan	13.76 (14)	9.48 (12)	11.28 (13)
Rhesus macaque	13.96 (14)	10.69 (13)	12.98 (14)

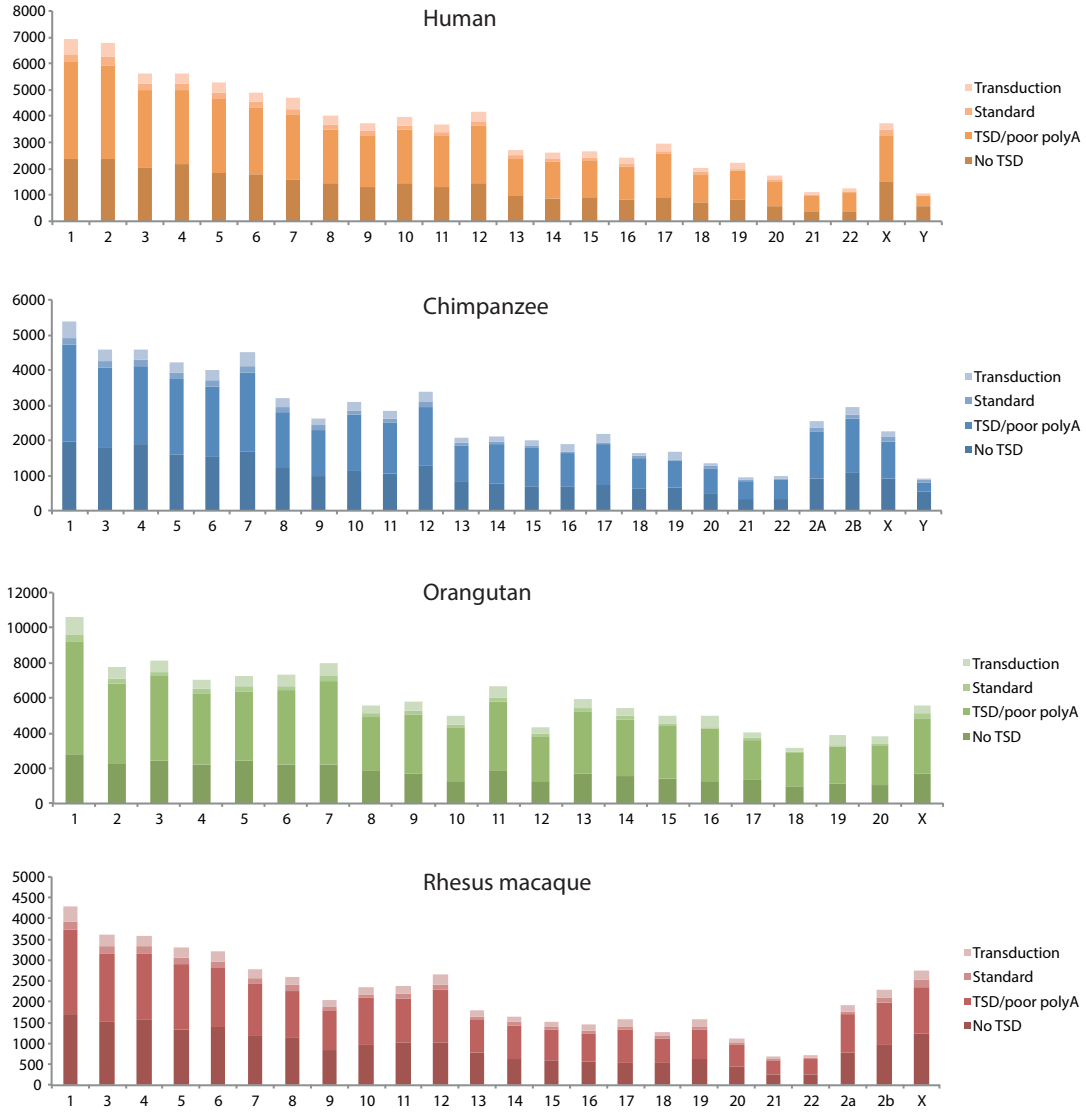

FIGURE A.4: Combined lineage-specific and species-specific fixed MEIs containing recent species-specific MEIs and ancient shared retrotransposon (breakdown per chromosome in human, chimpanzee, orangutan and rhesus macaque). Annotations were done using TSDfinder [Szak et al., 2002]. *No TSD* = MEI detected has no target-site duplication observed in the MEI flanking region. *TSD/poor polyA* = MEI detected has TSD, but has poor/weak polyadenylation (polyA) tail. *Standard* = Canonical MEI insertion with TSD and strong polyA tail. *Transduction* = MEI potentially carries additional unique sequence.

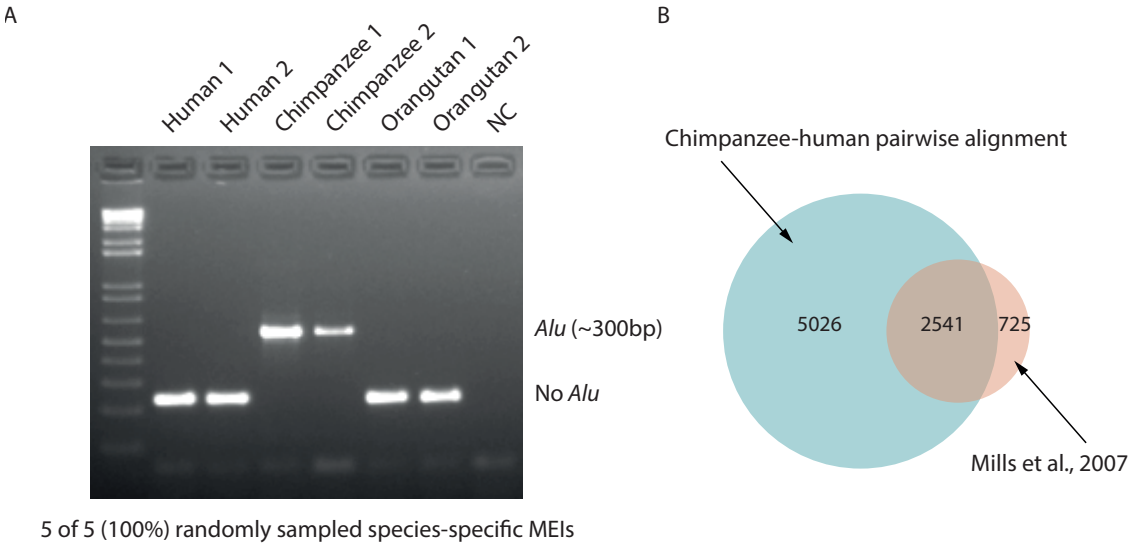

FIGURE A.5: Validations of species-specific reference-derived fixed MEIs. (A) Experimental validation of chimpanzee-specific MEIs: two chimpanzee-specific MEIs are absent from two human and two orangutan genomes. *NC* = negative control. (B) Overlap between chimpanzee-specific elements detected in our study and chimpanzee-specific elements from Mills et al. [2007] dataset. 78% of all elements are shared.

A.2 Supplementary information for Chapter 3

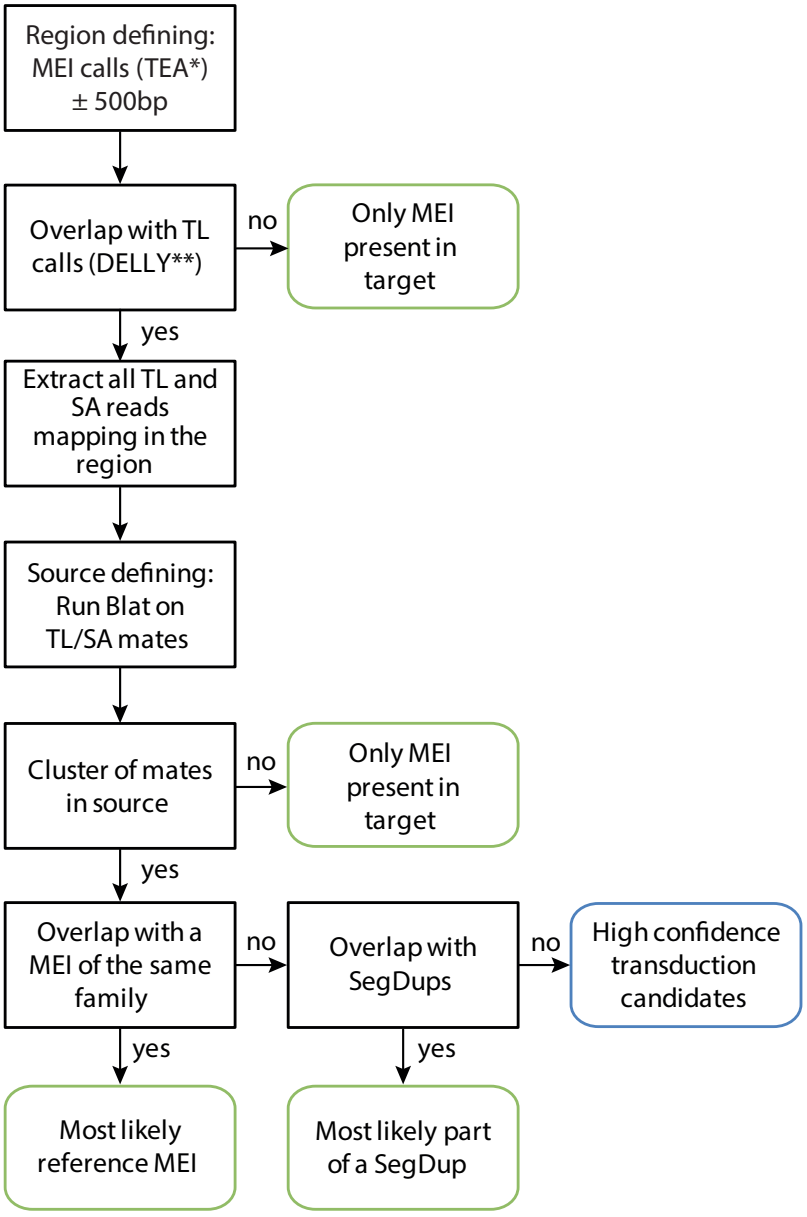

FIGURE A.6: TIGER approach: Each L1 coordinate is extended for additional 500 bp (L1 insertion±500bp). If the overlap between this region and at least one translocation exist, the repetitive L1 element and an additional unique sequence originating from another chromosome are thought to insert together. Once this signature is found and candidate loci are identified, all TL and SA reads mapping with one read to the predicted ±500bp surrounding insertion region are obtained from the BAM file and mates are realigned to the corresponding reference genome using UCSC standalone Blat software. All predicted insertion regions are filtered for overlap with corresponding segmental duplication (using the combined dataset presented in Chapter 4) as well as the presence of a reference L1 at the insertion.

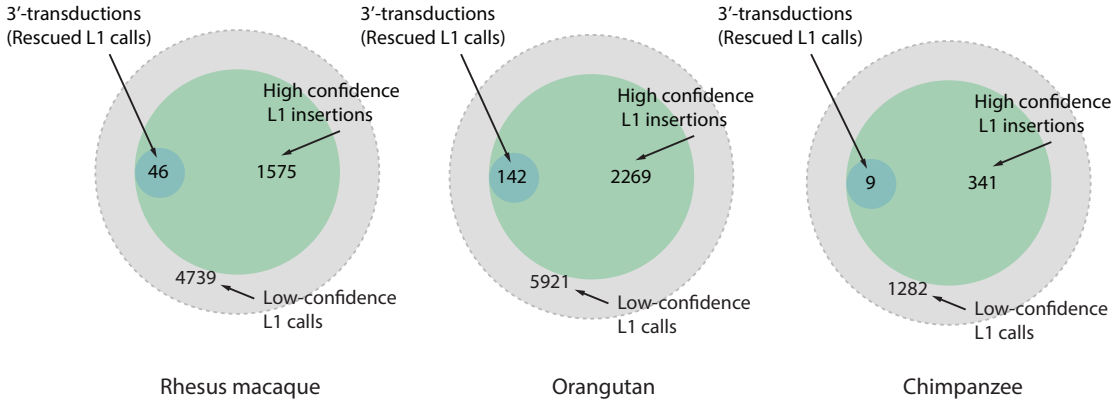

FIGURE A.7: Novel L1-TS calls are contributing to L1 diversity. Previously undetected L1-TS calls (blue circle) can be rescued by TIGER, which takes low confidence L1 callsets (gray circle) and looks for overlap with translocation calls, resulting in extra 46, 142 and 9 L1-TS calls in macaque, orangutan and chimpanzee, respectively. These calls would be lost using standard MEI callers, due to the stringency filtering requiring support for L1 call on both sides. Naturally, L1-TS calls usually have support for L1 insertion only on one side, whereas on the other unique TS is supported.

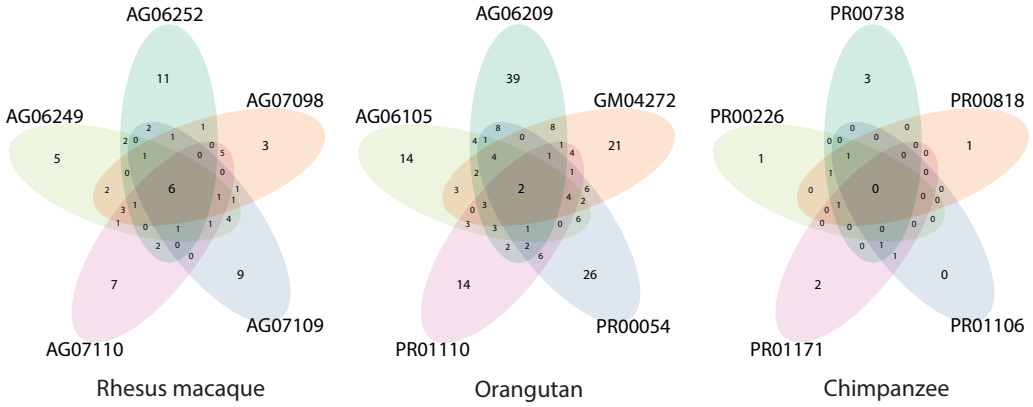

FIGURE A.8: Predicted L1-TS insertions in five individuals per species. In rhesus macaque 6/71 are shared between all five individuals, whereas in orangutan only two and in chimpanzee none are shared between all five individuals.

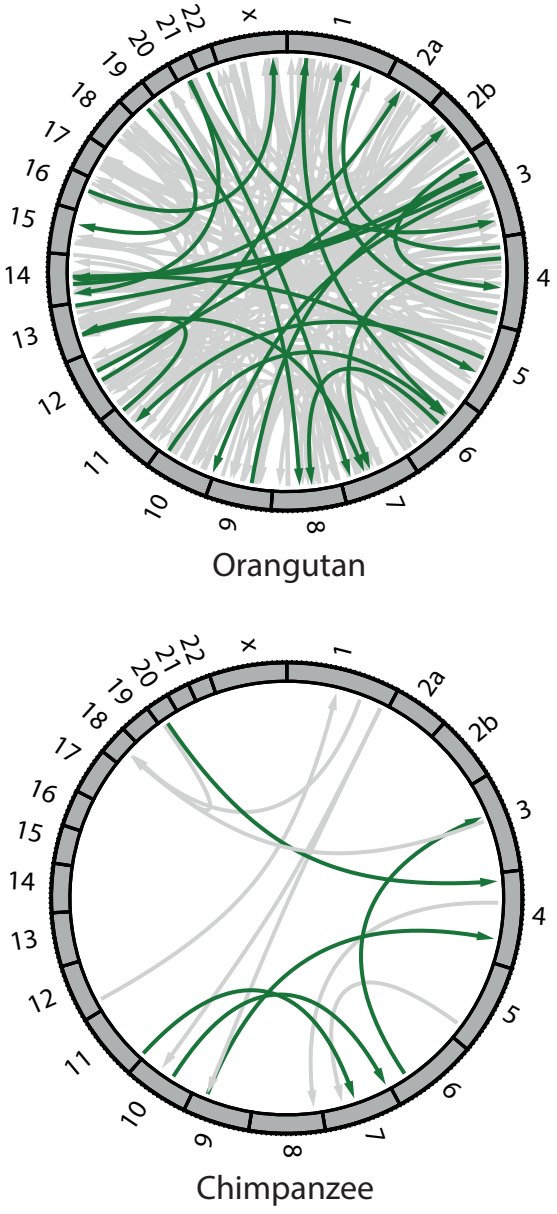

FIGURE A.9: Circos plot (http://circos.ca/) showing the distribution for all orangutan (upper image) and chimpanzee (lower image) L1-TS predictions, the 24 and five experimentally validated insertions in orangutan and chimpanzee, respectively, are depicted in green arrows. Arrows indicate the direction of the source inserting into the target locus.

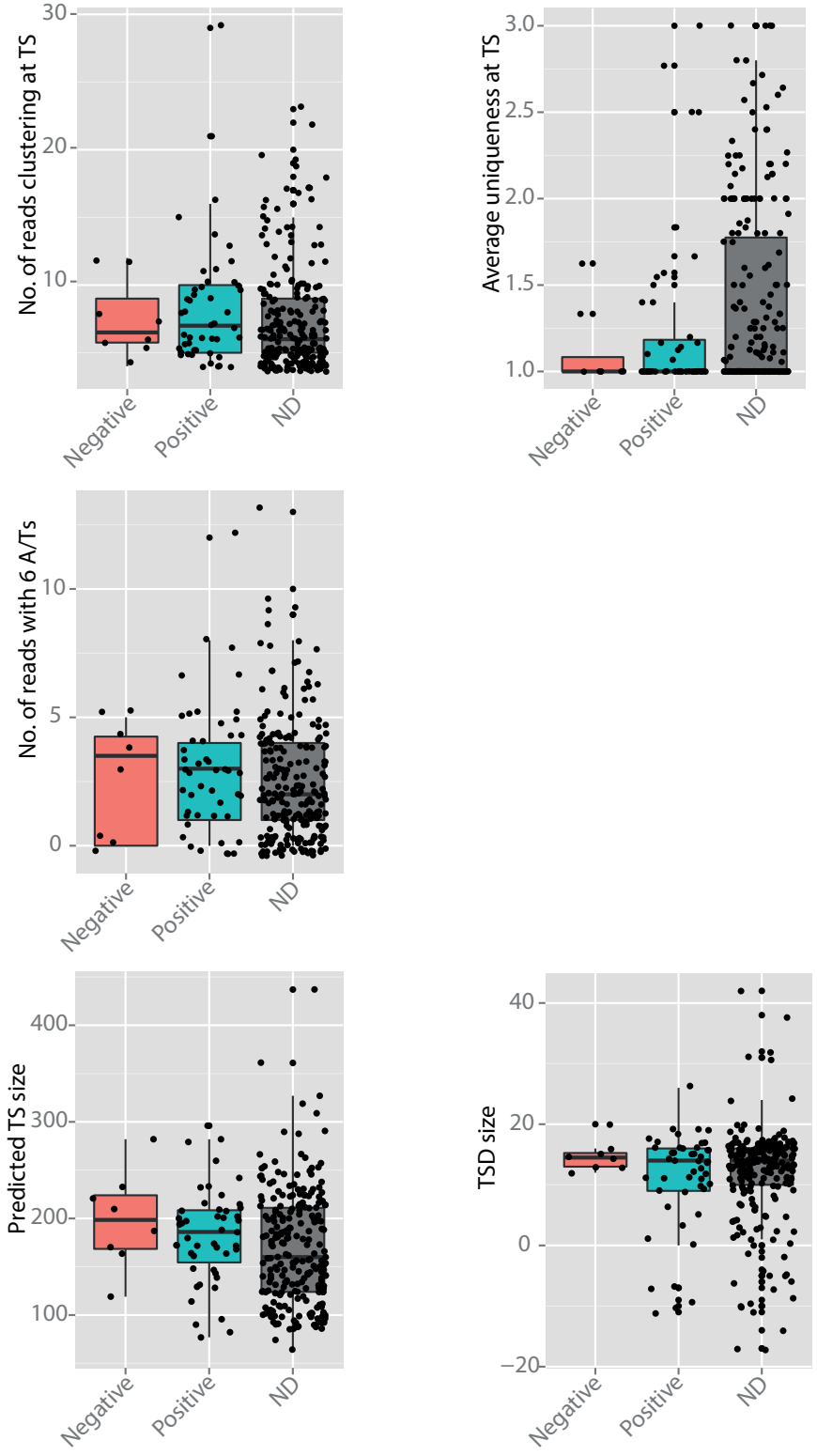

FIGURE A.10: Comparing number of clustered reads at TS site (first barplot), average uniqueness value of clustered TS reads (second barplot), number of reads with six consecutive A's/T's (third barplot), predicted TS size (fourth barplot) and TSD size with experimental validations (positive, negative and ND-non-determined calls). Parameters chosen did not exhibit significant differences between positive and negative calls (Welch t-test, $P>0.05$ for all here presented values, except TSD size ($P=0.0136$).

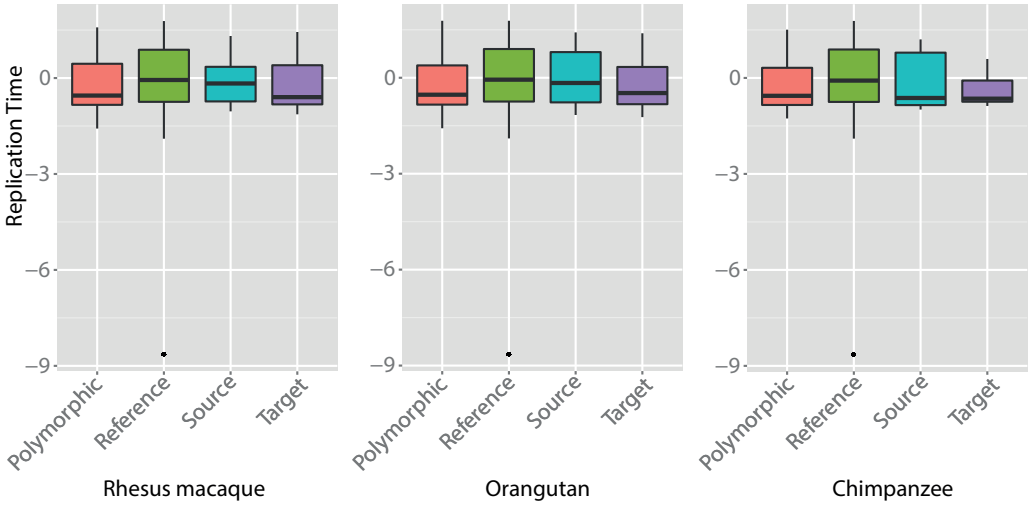

FIGURE A.11: Replication times in relation to predicted sources/targets: Both regions predicted as sources and targets fall more frequently in late replicating regions (negative values indicate late replicating regions, Appendix A, Figure A.11). Based on t-test statistics, observed difference between sources and targets replication time values in orangutan (P=0.025) is significant, whereas the same difference in macaque and chimpanzee is not (P=0.103 and and chimpanzee (P=0.343), respectively). Similar trend is observed in polymorphic L1 insertion (L1 targets) when compared to reference L1 events (potential L1 sources). Difference between those two categories is significant (Welch t-test P=2.2x10^{-16}) in all three species, probably affected by difference in number of datapoints (reference L1s are far greater in numbers than polymorphic L1s).

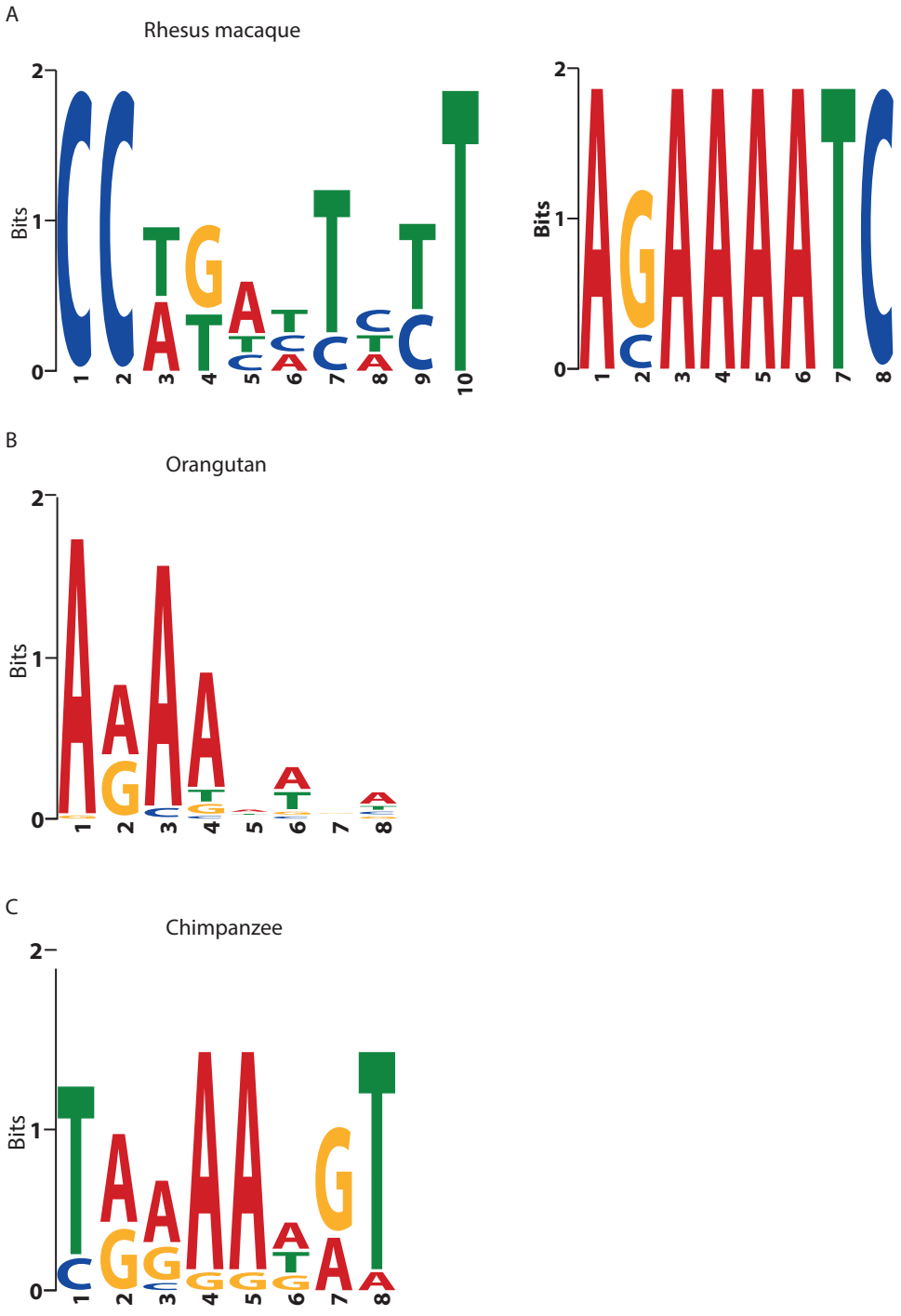

FIGURE A.12: Motives predicted (using MEME tool suite [Bailey and Elkan, 1994], http://meme.nbcr.net/meme/cgi-bin/meme.cgi) in the target site of L1-TS insertion in rhesus macaque, orangutan and chimpanzee based on TSDs with more than 8 bp length. (A) In rhesus macaque for only 15 sites out of 65, motif could be constructed (motif on the left derived from 10 sites and motif on the right derived from 10 sites). (B-C) Orangutan and chimpanzee L1-TS predictions with TSDs larger than 8 bp have one prominent motif depicted in the Figure. Note that all TSD sequences are in forward orientation.

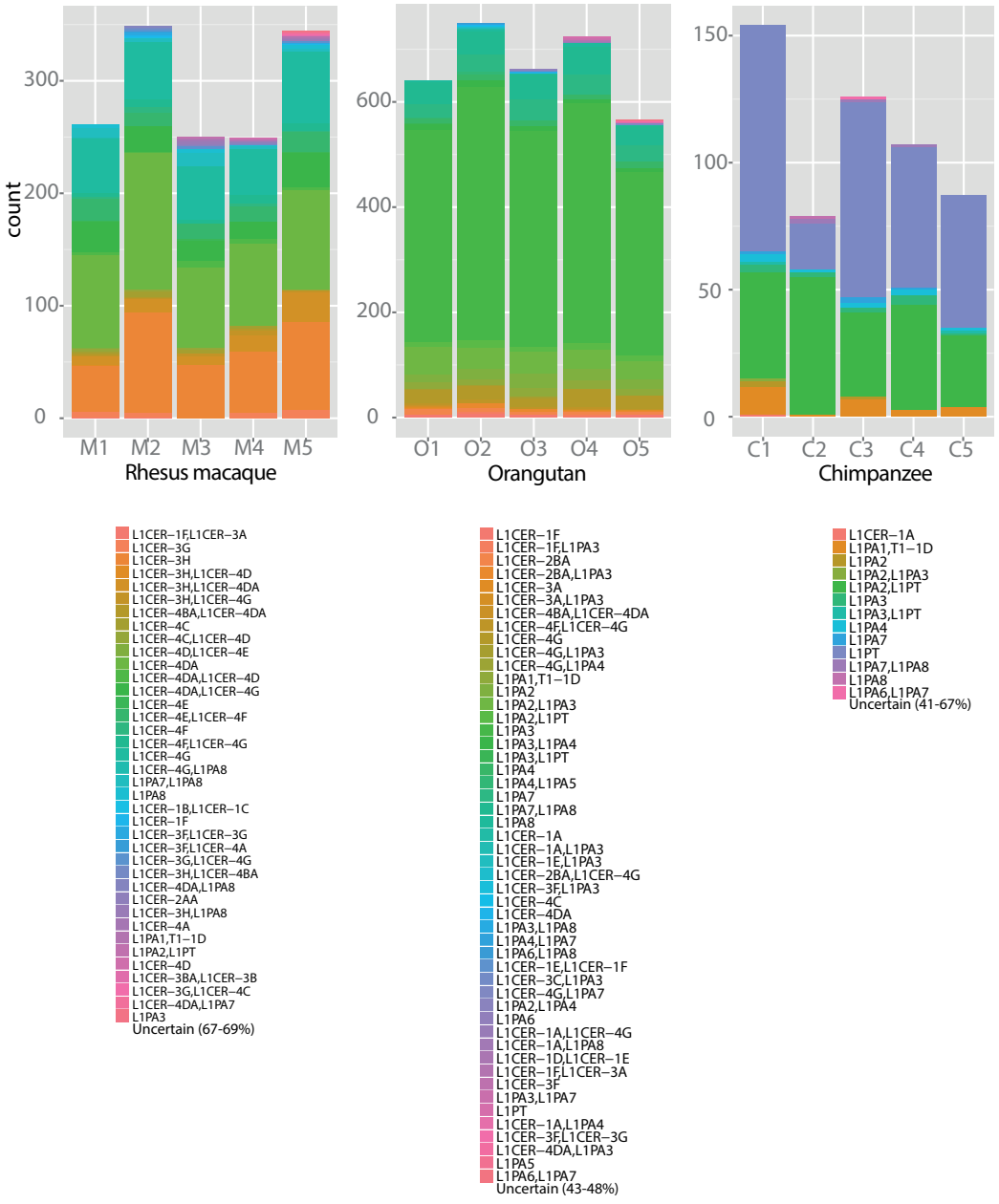

FIGURE A.13: Differences in L1 subfamily driving the solo-L1 insertions: In rhesus macaque, rhesus-specific L1CER subfamilies contribute to the most solo-L1 insertions, whereas in orangutan dominating L1 subfamily are L1PA2 and L1PA3. Most of the L1 insertions in chimpanzee are driven by L1PA2/L1PA3 and chimpanzee-specific L1PT subfamily. Note that colors cannot be compared across species (i.e. green fraction in rhesus macaque individuals mostly indicates L1CER subfamily, whereas in orangutan and chimpnzee individuals it denotes L1PA).'Uncertain' subtype indicates that L1 had more than two predicted subfamilies, subsequently merged into 'uncertain' class. Values in parenthesis indicate how many of the predicted L1 are 'uncertain' in five individuals.

Table A.2: Transduction rates calculated based on Tubio et al dataset: Tubio et al. dataset show high variability in transduction rates depending on a tumor type. In several colon, lung and prostate cancer samples, predicted somatic L1-TS predictions have comparable transduction rates to L1-TS calls in human germline predicted by TIGER.

Tumor type	**Total transductions rate**	**Partnered transduction rate**
Colon LS-1034	15.9	9.1
Lung PD7355	9.8	4.3
Lung NCI H2087	22.4	9.2
Lung TCGA-60-2695	25.0	9.6
Lung TCGA-60-2711	17.3	7.7
Lung TCGA-60-2722	33.3	8.3
Prostate PD11334a-e	15.2	9.0

Table A.3: Reference L1 elements in non-human primate species. Reference L1PA6-L1PA8 are pretty similar between the three non-human primate species; L1PA5 is specific to rhesus macaque, and L1PA2 is specific to chimpanzee and human (L1HS and L1PA2 diverged after chimpanzee-orangutan divergence, and L1HS (L1PA1) is mostly human-specific).

	Chimpanzee FL	**Orangutan FL**	**Rhesus macaque FL**	**Chimpanzee All**	**Orangutan All**	**Rhesus macaque All**
L1HS	8	11	6	100	133	96
L1PA2	255	1	0	3603	1278	24
L1PA3	662	786	12	10445	26714	379
L1PA4	849	661	1	11540	11700	454
L1PA5	668	580	2258	11062	11212	34204
L1PA6	572	504	487	5857	6072	6014
L1PA7	656	551	554	12686	12591	11913
L1PA8	145	131	118	7570	7562	7239

TABLE A.4: Overview of inferred retrogene presence in chimpanzee based on GRIPper [Ewing et al., 2013].

Species	**Ensembl ID**	**Gene Name**
Chimpanzee	ENSPTRG00000000242	SDHB
Chimpanzee	ENSPTRG00000002851	NDUB8
Chimpanzee	ENSPTRG00000003326	EIF3F
Chimpanzee	ENSPTRG00000006207	GMPR2
Chimpanzee	ENSPTRG00000007065	USP8
Chimpanzee	ENSPTRG00000008666	PHF23
Chimpanzee	ENSPTRG00000008742	RPL26L1
Chimpanzee	ENSPTRG00000010019	MYO5B
Chimpanzee	ENSPTRG00000010480	ILF3
Chimpanzee	ENSPTRG00000011268	H2RFV3
Chimpanzee	ENSPTRG00000013822	CCT8
Chimpanzee	ENSPTRG00000014888	SHISA5
Chimpanzee	ENSPTRG00000015789	NCBP2
Chimpanzee	ENSPTRG00000016217	HRNPDL
Chimpanzee	ENSPTRG00000017052	RPS23
Chimpanzee	ENSPTRG00000017737	NOL7
Chimpanzee	ENSPTRG00000018547	NUS1
Chimpanzee	ENSPTRG00000018988	TRA2A
Chimpanzee	ENSPTRG00000020252	LYPLA1
Chimpanzee	ENSPTRG00000021132	NUTM2F
Chimpanzee	ENSPTRG00000022902	SMARCE1
Chimpanzee	ENSPTRG00000023658	GMFB
Chimpanzee	ENSPTRG00000029858	NovelPseudogene
Chimpanzee	ENSPTRG00000030440	TMSB10
Chimpanzee	ENSPTRG00000030975	REXO1L1

TABLE A.5: Overview of inferred retrogene presence in orangutan based on GRIPper [Ewing et al., 2013].

Species	**Ensembl ID**	**Gene Name**
Orangutan	ENSPPYG00000002738	NovelPseudogene
Orangutan	ENSPPYG00000003396	ARL14EP
Orangutan	ENSPPYG00000006831	H2NPC8
Orangutan	ENSPPYG00000013974	HIGD1A
Orangutan	ENSPPYG00000014074	RAB5A
Orangutan	ENSPPYG00000014346	AP2M1
Orangutan	ENSPPYG00000015527	H2PFR7
Orangutan	ENSPPYG00000018763	UQCRB
Orangutan	ENSPPYG00000019647	GOLGA2
Orangutan	ENSPPYG00000020709	UTP14A

TABLE A.6: Overview of inferred retrogene presence in rhesus macaque based on GRIPper [Ewing et al., 2013].

Species	**Ensembl ID**	**Gene Name**
Rhesus macaque	ENSMMUG00000000486	FABP5
Rhesus macaque	ENSMMUG00000001682	PDIA3
Rhesus macaque	ENSMMUG00000002722	SH3TC1
Rhesus macaque	ENSMMUG00000003239	CYP11A1
Rhesus macaque	ENSMMUG00000003444	TMSB10
Rhesus macaque	ENSMMUG00000004441	S100A11
Rhesus macaque	ENSMMUG00000005098	PEBP1
Rhesus macaque	ENSMMUG00000006064	PPDPF
Rhesus macaque	ENSMMUG00000007341	IGLL1
Rhesus macaque	ENSMMUG00000007497	PABPC4
Rhesus macaque	ENSMMUG00000008177	PLA2G12A
Rhesus macaque	ENSMMUG00000008277	MRPS33
Rhesus macaque	ENSMMUG00000009499	F7HBC2
Rhesus macaque	ENSMMUG00000012054	ACTG1
Rhesus macaque	ENSMMUG00000012325	CNIH
Rhesus macaque	ENSMMUG00000012463	DDX46
Rhesus macaque	ENSMMUG00000012637	CCDC56
Rhesus macaque	ENSMMUG00000013182	GDI2
Rhesus macaque	ENSMMUG00000013618	OR51F1
Rhesus macaque	ENSMMUG00000013916	ERP29
Rhesus macaque	ENSMMUG00000014256	TMSB4X
Rhesus macaque	ENSMMUG00000016898	DUT
Rhesus macaque	ENSMMUG00000018843	NDUFAF4
Rhesus macaque	ENSMMUG00000020028	ExoSC1
Rhesus macaque	ENSMMUG00000020594	GINS2
Rhesus macaque	ENSMMUG00000021820	PARP1
Rhesus macaque	ENSMMUG00000022445	BNIP3
Rhesus macaque	ENSMMUG00000022639	KCTD3
Rhesus macaque	ENSMMUG00000022819	PNKD
Rhesus macaque	ENSMMUG00000022900	PDGFC
Rhesus macaque	ENSMMUG00000030260	TBC1D3F
Rhesus macaque	ENSMMUG00000032158	NANOGP1

A.3 Supplementary information for Chapter 4

TABLE A.7: hg19 (human), panTro3 (chimpanzee), ponAbe2 (orangutan) and rheMac2 (rhesus macaque) reference genome size statistics (obtained from http://genomewiki.ucsc.edu/.) *Chr* = chromosome, *Cov* = coverage.

	Chr count	**Total size**	**Non-N bases**	**N base count**	**% masked**	**Cov**
hg19	93	3,137,161,264	2,897,310,462	239,850,802	50.63	20X
panTro3	24,132	3,307,960,432	2,900,529,764	407,430,668	50.64	6X
ponAbe2	55	3,446,771,396	3,093,543,172	353,228,224	50.89	6X
rheMac2	22	2,864,106,071	2,646,668,809	217,437,262	48.28	5.1X

Table A.8: Overview of sample properties and sequencing statistics. Raw base and mapped bases are shown in gigabases(Gb).

Id	**Species**	**Sex**	**Age**	**IS**	**Raw bases**	**Reference**	**Mapped bases**	**% aligned**	**Cov (X)**
PR00226	Chimpanzee	M	12 years	322	106.6	panTro3	88.22	82.8	28
PR00738	Chimpanzee	M	8 years	301	45.38	panTro3	36.99	81.5	12.7
PR00818	Chimpanzee	F	29 years	288	80.37	panTro3	66.26	82.4	21.1
PR01106	Chimpanzee	M	14 years	268	63.78	panTro3	52.81	82.6	16.8
PR01171	Chimpanzee	M	16 years	256	62.81	panTro3	52.17	83.1	16.7
AG06105	Orangutan (Sumatra)	F	26 years	283	57.17	ponAbe2	45.83	80.2	15
AG06209	Orangutan (Sumatra)	F	28 years	256	71.02	ponAbe2	57.06	80.3	18.7
GM04272	Orangutan (Sumatra)	M	10 years	288	74.72	ponAbe2	58.56	78.4	18.9
PR00054	Orangutan (Sumatra)	M	4 years	272	71.72	ponAbe2	56.88	79.3	18.6
PR01110	Orangutan (Sumatra)	F	17 years	262	57.8	ponAbe2	45.78	79.2	14.9
AG06249	Rhesus macaque (India)	F	9 months	335	62.28	rheMac2	50.32	80.8	17.5
AG06252	Rhesus macaque (India)	F	5 years	327	74.91	rheMac2	58.61	78.2	20.4
AG07098	Rhesus macaque (India)	F	8 days	323	54.34	rheMac2	44.18	81.3	15.4
AG07109	Rhesus macaque (India)	F	27 years	348	55.21	rheMac2	44.59	80.8	15.5
AG07110	Rhesus macaque (India)	F	27 years	326	52.63	rheMac2	42.8	81.3	15

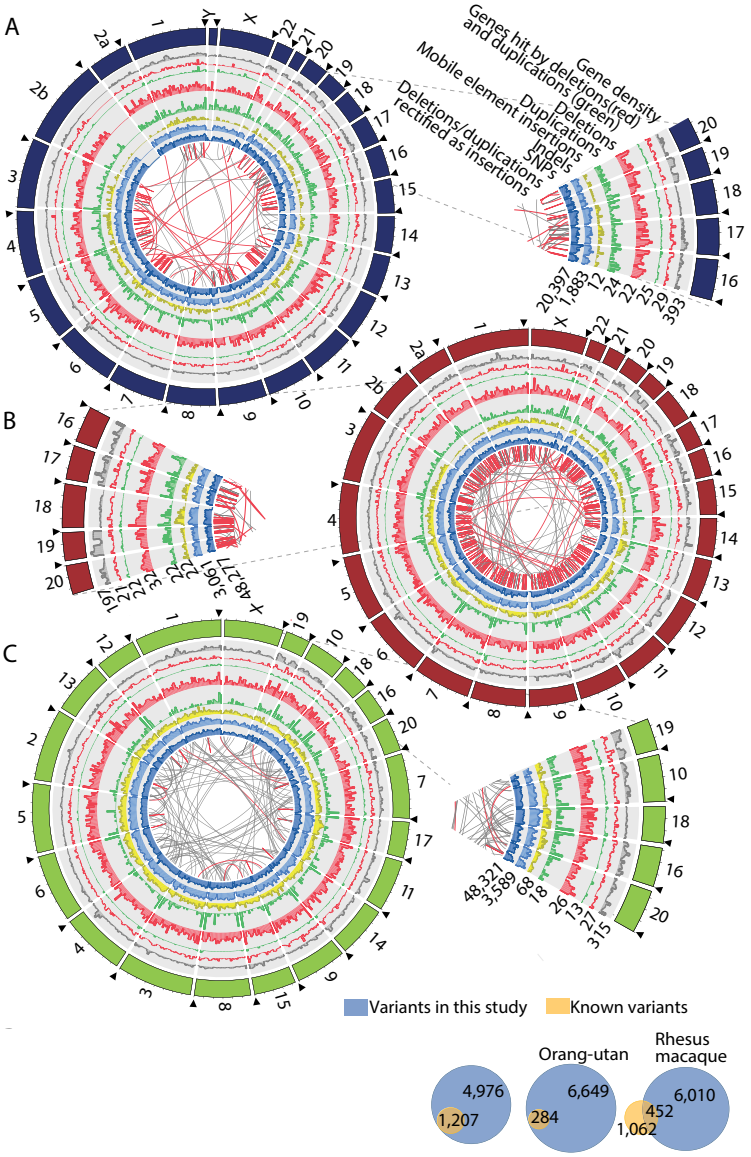

FIGURE A.14: Overview of genomic sequence variants in (A) chimpanzee, (B) orangutan, and (C) rhesus macaque. Black arrowheads mark the start of the chromosomes. Macaque chromosomes are sorted according to orthology with respect to human. The missing part of chromosome 2b in chimpanzee is caused by a large telomeric reference genome gap. Connecting lines at the inside of each plot depict the movement of duplicative insertions (i.e., deletions and duplications rectified as insertions) [Lam et al., 2010]. Red connecting lines indicate NAHR events, and gray connecting lines indicate non-NAHR events (MEIs excluded). Pie slices zoom into the respective circos plots. Heights for different variant types in the circos plots are relative to the abundance of the respective variant type (numbers at the lower edge of the pie slices indicate the maximum value in a 5Mb bin for each variant type in the whole subcircle). Venn diagrams (lower panel) depict the proportion of previously reported structural variants based on published aCGH-based surveys (excluding non-reference MEIs).

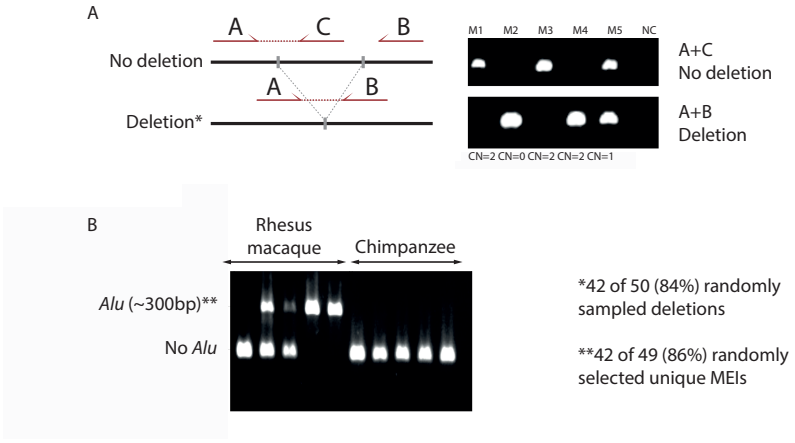

FIGURE A.15: Validations of computationally predicted SVs. (A) Polymerase chain reaction (PCR) based verification of the rhesus macaque deletion. NC = negative control; M1-M5 = macaque samples; CN = copy number; A, B and C = primers designed for deletion validation (A+C=no deletion; A+B=deletion). (B) Quality assessment of non-reference MEIs in rhesus macaque compared to chimpanzee (polymophic MEIs within one species, absent from another).

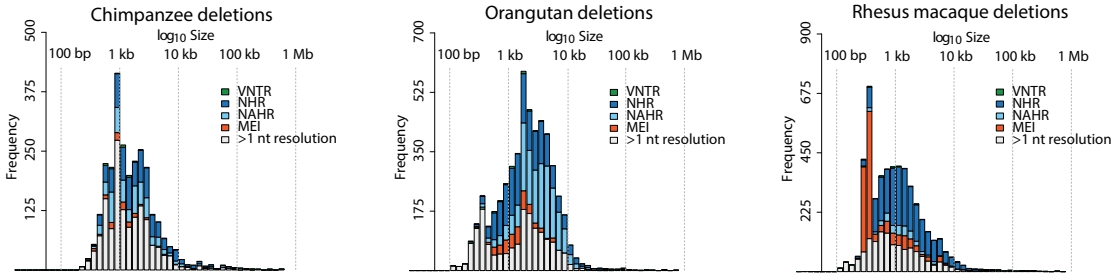

FIGURE A.16: Size distributions of deletions in chimpanzee, orangutan and rhesus macaque.

A.4 Supplementary information for Chapter 5

TABLE A.9: Patient and genome sequencing information. Patient information was obtained from Jones et al. [2012]. M stage belongs to the TNM classification system: T = tumor, N = node, M = metastasis. The number following the letter marks that distant metastases were found.

	LFS-MB1	**LFS-MB2**	**LFS-MB3**	**LFS-MB4**
Age (years)	11	14	12	12
Gender	F	M	M	M
MB type	SHH	SHH	SHH	SHH
M stage	M0	M0	M3	NA
Tumor bases sequenced	109 x 10^9	120 x 10^9	37 x 10^9	143 x 10^9
Paired normal tissue bases sequenced	116 x 10^9	125 x 10^9	17 x 10^9	114 x 10^9
Tumor physical coverage (span coverage)	43.5x	45.8x	77.8x	112.2x
Paired normal physical coverage	41.6x	51.6x	3.3x	49.5x
Tumor sequencing coverage	30.8x	34.6x	8.9x	38.5x
Paired normal sequencing coverage	31.4x	36.7x	4.6x	34.4x

TABLE A.10: Distribution of different *de novo* formation mechanisms observed in germline-specific, tumor-specific (no chromothripsis) and chromothripsis-related structural variants.

	MEI	**NAHR**	**NHEJ/MMBIR**	**VNTR**
Germline	1683	838	1798	1082
Tumor-specific	0	5	62	5
Chromothripsis	0	0	63	0

TABLE A.11: Summary of microhomology (MH) lengths observed in in germline-specific, tumor-specific (no chromothripsis) and chromothripsis-related structural variants. *Min* = minimum, *Max* = maximum. N=5401,72,63 for germline-specific, tumor-specific (no chromothripsis) and chromothripsis-related structural variants, respectively.

	Min MH	**Max MH**	**Mean MH**	**Median MH**
Germline	0	37	2.42	2
Tumor-specific	0	10	1.40	1
Chromothripsis	0	4	0.86	0

TABLE A.12: Summary of non-template microinsertions (MI) lengths observed in in germline-specific, tumor-specific (no chromothripsis) and chromothripsis-related structural variants. *Min* = minimum, *Max* = maximum. N=5401,72,63 for germline-specific, tumor-specific (no chromothripsis) and chromothripsis-related structural variants, respectively.

	Min MI	**Max MI**	**Mean MI**	**Median MI**
Germline	0	50	1.42	0
Tumor-specific	0	18	0.65	0
Chromothripsis	0	10	0.68	0

Appendix B

Methods

B.1 Methods for Chapter 2

This Thesis is mostly a result of collaborative effort and many of the methods we have already described in the corresponding research article indicated before each Chapter. Therefore, previously published methods presented here are based on methods described in each publication.

Data access

The sequencing data have been deposited in the European Nucleotide Archive, www.ebi.ac.uk/ena/ (accession no. ERP002376). In addition, all the callsets are available at http://www.korbel.embl.de/primate_sv/.

Samples

Fibroblast-derived cell lines from unrelated chimpanzee, orangutan and macaque individuals (five samples each) were obtained from the Coriell Cell repository, following the acquisition of federal (Federal Fish and Wildlife Permit, USA – Permit: MA232608-0) and institutional permissions.

Sequencing library preparation

5μg of high molecular weight genomic DNA were fragmented to 250-350 bp insert size with a Covaris S2 device (Covaris, Inc.), followed by sequencing on an Illumina Hiseq2000 instrument. Sequenced reads were aligned to the respective reference genomes of each species in paired-end mode using the alignment software ELAND, version 2 (Illumina). The alignment files were converted to SAM/BAM format using SAMtools [Li et al., 2009] and subjected to various variant discovery pipelines.

Species-specific MEI dataset generation

Each pairwise alignment was downloaded from the UCSC Genome Browser

(http://hgdownload.cse.ucsc.edu) in the form of net/chain files. Whole-genome assemblies alignments were performed by the BLASTZ/LASTZ alignment program [Harris and Chiaromonte, 2007], available from Webb Miller's lab at Penn State University (http://www.bx.psu.edu/miller_lab/). For each pairwise alignments, loci where query sequence has 'fill' sequence and the control sequence has alignment 'gap' are chosen for subsequent analyses. In case of chimpanzee, for example, chimpanzee (*pantro3*) would be a query sequence and human (*hg19*), gorilla (*gorGor3*), orangutan (*ponAbe2*) and rhesus macaque (*rheMac2*) control sequences. Compared to all of them, chimpanzee genome must contain specific 'fill' sequences whereas in the alignment each control sequence would have a gap. Those 'fill' sequences are subjected to overlap with the corresponding RepeatMasker file (downloaded from http://hgdownload.soe.ucsc.edu/) containing *Alu*, L1 and SVA repeats. 80% reciprocal overlap between 'fill' sequences and any MEI was required in order to be identified as species-specific MEI. In addition, all shared MEIs between species eliminated to generate species-specific MEI dataset, were subsequently collected in a list of lineage-specific calls, to allow us to compare recent and ancient fixed MEIs.

Species-specific MEI data annotation

To annotate every species-specific MEI, the TSDfinder tool was used [Szak et al., 2002]. Species-specific dataset was split into separate files containing MEIs on each chromosome and subjected to target-site duplication (TSD), polyadenylation tail, potential truncation and transduction evaluation using default setting of the TSDfinder.

Validation of species-specific MEIs

To verify the species-specific generation approach, chimpanzee-specific dataset we generated from chimpanzee-human pairwise alignment was compared to chimpanzee-specific dataset from Mills et al. [2007]. 50% reciprocal overlap between two instances in two datasets was required to be identified as same element. 78% Mills et al. [2007] MEI were successfully recovered.

Additionally, experimental validations for five random selected loci were performed by designing primers outside of the predicted species-specific event. Due to species pairwise alignments, each MEI flanking sequence is always identical between species, allowing to differentiate presence/absence of the same element in different species.

B.2 Methods for Chapter 3

Computational prediction of L1-TS candidates by TIGER

To identify possible L1-transduction insertion events, intersectBed from BEDTools [Quinlan and Hall, 2010] was used to obtain an intersection between non-reference L1 insertion and at least one translocation (TL) read. As non-reference L1 insertion coordinates usually indicate short

TSD sequence, each coordinate is extended for additional 500 bp (L1 insertion±500 bp). The existing overlap could indicate the presence of both a repetitive L1 element and an additional unique sequence that is originating from another chromosome, respectively. Once this signature is found and candidate loci identified, all TL and single-anchored (SA) reads mapping with one read to the predicted ±500 bp surrounding insertion region are, in addition, obtained directly from the BAM file [Li et al., 2009]. Their respective mates were realigned to the corresponding reference genome using UCSC standalone BLAT software (version 34) [Kent, 2002] to either confirm (with TL reads) or discover (with SA reads) the source chromosome. Although TL reads have both mates mapped, sometimes mapping creates artifacts, especially if the read sequence aligns to the repetitive portion of the genome. Essentially, if the read is repetitive, it can be aligned onto several places in genome and depending on the mapper and parameters used, the read itself can be deemed unmapped (creating single-anchored and fully unmapped paired-end reads) or can be randomly placed to any of these positions (creating TL artifacts).

BLAT [Kent, 2002] mapping positions were further subjected to filtering based on the highest bit-score to find the highest confidence reference match of all possible matches. Additionally, the total number of possible matches (TM) was recorded, allowing to distinguish repetitive from unique regions in the genome (i.e. repetitive regions will have relatively high TM due to their mapping to multiple locations in the genome, in contrast to unique regions with relatively low TM). In addition, only reads mapping with at least 50 bp to the reference genome were taken into consideration (alignment length from BLAT output [Kent, 2002], AL=50 bp).

Finally, cluster of reads mapping uniquely to the region in the reference genome (at least 4 reads clustering together in a same region) as well as cluster of reads mapping multiple times in the genome were determined per insertion locus. This indicates an insertion of a unique transduced sequence and a repetitive L1 sequence. As our samples were sequenced up to 25x coverage (typical coverage used for whole genome sequencing, WGS), upper limit of clustered reads at one source locus was determined to be 30. To confirm that the transduction sequence is indeed unique, the mean of all unique read-specific TM values per locus was calculated and set to be $\leq$3. If the transduced sequence does not satisfy the latter condition, it would indicate either high repetitiveness of transduction or even present a case with no transduction at all (i.e. only L1 insertion with 4 reads clustering at any reference L1 loci). Importantly, in order to get the longest stretch of the source loci possible, reads were clustered in an overlapping fashion (i.e. gap between reads was not allowed). All predicted insertion regions were filtered for overlap with corresponding segmental duplication (using the combined dataset presented in Chapter 4) as well as the presence of a reference L1 at the insertion in order to prevent possible false calling (Appendix A, Figure A.6) .

To annotate and further characterize predicted L1-TS sequences, TSD values were directly extracted from the TEA output [Lee et al., 2012], whereas a putative presence of a polyadenylation

tail (polyA tail) was additionally evaluated by searching for six consecutive non-reference A's or T's (AAAAAA/TTTTTT) in each read.

Parameters chosen with TIGER were shown to be optimal for L1-TS detection with ∼25x coverage data. In order to test each possible parameter, experimentally validated predictions were compared with negative predictions. TSD size, presence of polyA tail, size of predicted TS, number of clustered reads at the predicted source loci and average *TM* values per locus were compared and there were no significant differences in distribution observed, indicating that further parameter adjustment will not result in higher-quality predictions (Appendix A, Figure A.10).

Data used for L1-TS discovery

The TIGER tool was applied to ∼25x WGS data: three different non-human primate species – chimpanzee, orangutan and macaque (five individuals per species, sequenced between 14.4-28.8x, [Gokcumen et al., 2013]) and a human sample NA12878 (HapMap/1000GP CEU daughter from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/ [Abecasis et al., 2010, The 1000 Genomes Project Consortium, 2012]), downsampled to ∼21x using three independent technical replicates.

To facilitate a comparison of non-human primates to human, a high-coverage human sample was downsampled to ∼21x using Downsample.jar from PicardTools version 1.52 (http://picard.sourceforge.net.) using predefined random-seed value and default random seed value (two technical replicates). Additionally, one technical replicate was downsampled using Samtools 0.1.19 (samtools view –s option) [Li et al., 2009], to exclude any data generation bias. Both non-human primates and human datasets were sequenced using the Illumina sequencing platform, have 101 bp paired-end reads, and have comparable sequencing coverage of ∼20x.

The WGS data was aligned onto the corresponding reference genome builds: human *hg19*, chimpanzee *panTro3*, orangutan *ponAbe2* and rhesus macaque *rheMac2* using commercial software Eland v2 from Illumina for the non-human primate species data and with BWA [Li and Durbin, 2009] for the human data. For each species, translocation calls were inferred by running Delly v0.0.11 (jumpy_v0.0.11) [Rausch et al., 2012b]. Non-human primate ME calls were determined by an improved version of TEA (see below, [Lee et al., 2012]) and all L1 calls were considered as a source for possible transductions (i.e. low confidence L1 calls often lack support from both the 5' and 3' end of the insertion point because they can actually be L1-TS events).

Non-reference L1 insertion discovery

The TEA pipeline [Lee et al., 2012] was used to perform non-reference MEI discovery. TEA detects an MEI by identifying 1) clusters of 'repeat-anchored mate' (RAM) reads, which are uniquely mapped to the reference genome and have paired mates that map to ME sequence library, and 2) partially-aligned reads spanning the insertion breakpoints ('clipped reads'), whose

unaligned tail sequences match the inserted ME. The ME sequence library was built by concatenating multiple consensus sequences of ME subfamilies separated by 200 'N' nucleotide spacers. For L1, consensus sequences for L1HS, L1PA3, L1PA5, L1Pt were used. To better detect L1 events with transduction, RAM clusters having RAMs appear only one side of the insertion in L1 candidate sets were included (low confidence L1 calls). At least three RAMs and at least one clipped read on either or both sides of an insertion was required.

Estimating the size and subfamily of L1 insertions

To assess subfamily driving L1 insertions as well as L1-mediated transductions, sequences of clipped reads and RAM mates were assembled into longer contig sequences using CAP3 [Huang and Madan, 1999] to estimate the size and subfamily of L1 elements. Consensus sequences of 42 L1 subfamilies (T1-1D, L1PA1-L1PA8, chimpanzee-specific L1Pt, rhesus macaque-specific 32 L1CER elements [Han et al., 2007]) were compiled, and 921 bp 3' end sequence of each subfamily after multiple sequence alignment using ClustalW2 (http://www.ebi.ac.uk/Tools/msa/clustalw2/) was used for estimating L1 size and subfamily. Subfamilies whose consensus sequences have the smallest mismatch with the RAM and clipped read contigs were reported. Inversion within an L1 was detected from inconsistent mapping orientations of contigs reconstructed from both ends of the insertion. The insertion size was not estimated for insertions with inversions. All L1 annotated with three or more subfamilies were identified as 'Uncertain' group, whereas L1 annotated with one or two subfamilies were reported.

Design of experimental validations

A combination of PCR with several primer combinations and capillary sequencing was necessary to validate transduction prediction and calculate a TIGER FDR rate. Outer primers were designed to bind to unique regions at least 100 bp away of the target integration site using in house primer design tool and inner set of primers (inside TS) were designed using Primer3Plus (http://www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi/) [Untergasser et al., 2007]. PCR primers were obtained from Sigma. PCRs were preformed using 10ng of genomic DNA (Coriell) in 25μl volumes using the Sequalprep Long PCR reagents (Life technologies) in a 96 well plate using the DNA Engine Tetrade 2 thermocycler (BioRad). PCR conditions were: 94 °C for 3min, followed by 10 cycles of 94 °C for 10s, 62 °C for 30s and 68 °C for 6min and 25 cycles of 64 °C for 10s, 60 °C for 30s and 68 °C for 8min, followed by a final cycle of 72 °C for 10min. PCR products were analyzed on a 0.8% agarose gel stained with Sybr Safe Dye (Life Technologies) and a 100 bp ladder and 1 kb ladder (NEB). If necessary, gel bands were cut with a scalpel, gel extracted with the Nucleospin Gel and PCR Cleanup kit (Macherey-Nagel) and send for capillary sequencing (GATC Biotech AG). Sequence chromatograms were manually inspected and sequences were analyzed by BLAT [Kent, 2002].

Oxford Nanopore MinION library sequencing

The purified amplicon PCR DNA pool was used with the Genomic DNA Sequencing kit (version SQK-MAP002) for MinION library prep as part of the MinION Access Programme (Oxford Nanopore Technologies). Briefly, 1.5-2μg of amplicon pool DNA and 5μl of DNA-CS were end repaired using the End repair module reagents (NEB) for 30min at 20 °C, purified with 0.5 volumes of AMPure XP beads and eluted in 25.2μl nuclease free water. A-tailing (NEB) was performed in 30μl for 30min at 37 °C and followed by adapter ligation (Oxford Nanopore Technologies) by adding 10μl adapter mix, 10μl of HP adapter and 50μl of Blunt T/A ligase mix (NEB) and incubation for 10min at 20 °C. A special AMPure XP cleanup step (0.4x volume) was performed, using 150μl of provided wash buffer instead of 70% EtOH once and elution into 25μl provided elution buffer without a drying step. Next, Tether annealing was performed by adding 10μl Tether mix and incubation for 10min at 20 °C and followed by the Library conditioning step by addition of 15μl HP motor mix and incubation o.n. at 20 °C at 750rpm. Briefly before the MinION sequencing run, 6μl of prepared library was mixed with 140μl EP buffer and 4μl of Fuel mix, gently mixed to produce the final library and loaded on a primed MinION flowcell (version FLO-MAP001 and FLO-MAP002). MinION flowcells were used with the software client Metrichor v 0.17.39962, the sequencing software MinKNOW v 0.46.1.9 and the 2D Workflow v1.7. Flowcells with more than 200 active pores in the MAP_Platform_QC run were used for a sequencing run. First, the flowcell was primed with 150μl EP buffer and 10min waiting twice before 150μl of final amplicon library were loaded and started sequencing with the MAP_Amplicon48hSequencing_run script producing fast5 files for analysis.

Goodness-of-fit statistical test of predicted-transduction rates

To test whether the differences between transduction rates were indeed significant, the L1-TS dataset was fit to a Poisson linear model (Tweedie model). The coefficients taken into account to fit predicted transduction numbers were: species, number of all high-confidence solo-L1 insertions and physical coverage, to ensure none of the mentioned coefficients would create bias. The goodness-of-fit was assessed by taking the residual values against the fitted values and subsequently calculated P-value for every pairwise comparison: chimpanzee-orangutan (P=0.000037), chimpanzee-macaque (P=0.000073) and orangutan-macaque (P=0.0003).

Replication time analysis

Values for the replication time analysis were extracted from the Replication Domain database (http://www.replicationdomain.com/ [Weddington et al., 2008]). As replication timing maps do not exist for the non-human primate species, replication time values were obtained for human fibroblast cell line data, as a comparable dataset (hFib cell line - Homo sapiens (build *hg19*) public dataset, ChipID: 552613A05_2012-12-22_hFib-2, array design name: 100710_HG19_WG_CGH_PERF_UX6, PMID: 24685138, Nimblegen platform).

To convert the values from rhesus macaque, orangutan and chimpanzee genome assembly (*rheMac2*, *ponAbe2* and *panTro3* assemblies) to *hg19* human coordinates, the liftOver tool from the UCSC Genome Browser (http://genome.ucsc.edu/,[Fujita et al., 2011]). To overlap transduction predictions (with converted coordinates) with the replication time values intesectBed option from Bedtools was used with default values [Quinlan and Hall, 2010].

Identification of retrogene insertions

A published method GRIPper [Ewing et al., 2013] was used to infer retrogene insertions based on non-human primates sequencing data [Gokcumen et al., 2013]. All parameters were kept at default values, apart from minPeakSize, which was set to 4 reads, and the insert size which was adjusted to the insert size of our sequencing libraries (200-350 bp).

As GRIPper requires gene annotation files to infer retrogenes, gene annotation files were generated for chimpanzee, orangutan and rhesus macaque. Exon annotations were downloaded for each species from ENSEMBL (http://www.ensembl.org/info/data/ftp/index.html) and reformatted to the required input format. Gene information was generated from the 'Genes and Predictions Tracks' available at the UCSC Table Browser (http://genome.ucsc.edu/cgi-bin/hgTables). Repeat annotation files were formatted to the required input format based on the RepeatMasker annotation available via the UCSC genome browser (http://genome.ucsc.edu/cgi-bin/hgTables). Known pseudogenes were downloaded from Ensembl BioMart (http://www.ensembl.org/biomart/martview). All annotations were kept consistent with reference genome builds used for alignment of sequenced short reads.

PCR validation of retrogene insertions

Primers for PCR validation were designed using Primer3Plus (http://www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi/ [Untergasser et al., 2007]) with default parameters to test if the insertion of the predicted source gene in the predicted insertion locus is true. The forward primer was placed maximum 500 bp upstream of the inferred insertion point and the reverse primer was placed in an exon of the source gene. In addition, the reverse primer was designed to account for the predicted orientation of the new GRIP. The presence of exon-exon junctions indicative of a GRIP was confirmed by Sanger sequencing the PCR products and subsequent alignment of sequenced products to the respective reference genomes using BLAT [Kent, 2002].

B.3 Methods for Chapter 4

Discovery of copy-number variants

Deletions and duplications were discovered by combining three different copy-number variant signatures: (1) discordantly mapped paired-end reads, (2) split reads, and (3) abnormal read depth signatures. With the aim to define the most comprehensive dataset, DELLY version 0.0.4 [Rausch et al., 2012b], GenomeSTRiP version 1.03 [Handsaker et al., 2011] and CNVnator version 0.2.2 [Abyzov et al., 2011] were used. DELLY utilizes paired-end mapping and split-reads to define breakpoint-resolution SV calls, whereas CNVnator performs read-depth approach analysis to identify CNVs. GenomeSTRiP essentially integrates read-depth and paired-read based discovery approaches and performs population-based deletion calling. To detect tandem duplications and deletions, DELLY was used with default parameters. CNVnator was used for tandem and dispersed duplication discovery, as well as for deletion discovery by applying window sizes between 100 bp and 300 bp depending on the genomic read coverage of a samples. At least 2 supporting read pairs were required to trigger a splitread analysis in search for deletion and duplication breakpoints.

Calls generated by each of the three methods were filtered and merged based on certain overlap between coordinates of different callers. For instance, 2 calls generated by splitread approach were merged together if they had absolutely the same breakpoint predicted. Paired-end mode detected calls were merged together if they displayed intersecting confidence intervals, assuming intervals of $\pm$100 bp at the breakpoints. Since read-depth approach has even lower resolution, CNVnator calls were merged assuming a confidence interval of $\pm$300 bp at the breakpoints. GenomeSTRiP deletion calls displaying >50% reciprocal overlap were merged with a combined DELLY/CNVnator deletion dataset. The coordinated were taken from DELLY if they were identified at the breakpoint resolution and otherwise they were based on GenomeSTRiP output. Deletion calls observed in all 5 samples of a species showing a >50% reciprocal overlap with reference assembly gaps were removed to ensure high quality of the deletion set. DELLY/CNVnator duplication dataset was independently verified using the read-depth based copy-number genotype assessment algorithm CopySeq version 1.7.1 [Waszak et al., 2010], using default parameters.

Our final dataset was categorized into the '*discovery dataset*' and the '*breakpoint dataset*' (i.e., SV calls with DELLY-based split-read support. Reference MEIs that were detected as deletions relative to the reference genome were separated from our deletion set and analyzed along with the non-reference MEIs in our MEI set. The remaining SVs with breakpoint resolution were used for assessment of SV formation mechanisms and for ancestral state determination.

SV formation mechanism assignment

BreakSeq (version 1.3 with default parameters) [Lam et al., 2010] was used to infer formation mechanisms for deletions and duplications mapped with nucleotide resolution breakpoints.

Roughly ∼51% of all deletions and ∼18% were successfully mapped at the breakpoint resolution. In order to perform mechanism classification, the coordinates of SVs predicted by DELLY [Rausch et al., 2012b] had to be adapted as follows: *BreakSeqStart=DellyStart+1* and *BreakSeqEnd=DellyEnd-1*, due to the discrepancy between DELLY and BreakSeq interpretation of 1-based coordinate system. For every species, species-specific RepeatMasker and the corresponding reference genome downloaded from UCSC (http://genome.ucsc.edu/cgi-bin/hgGateway) was used in order to obtain mechanisms specific for each species.

Ancestral state inference

The ancestral state analysis was performed using the BreakSeq package [Lam et al., 2010]. For deletions or duplications as identified relative to the respective reference genome, two different alleles were taken into account for ancestral state determination: (1) the reference allele, for which ±500 bp flanking sequences were extracted at each breakpoint representing both left and right reference junction sequences; (2) the alternative (deleted/duplicated) allele, for which also ±500 bp breakpoint flanking sequences were extracted. The respective junction sequences were extracted from each species and were aligned to the genomes of the other species (e.g., rhesus macaque junction sequences (query species) were aligned on the marmoset (*calJac3*), orangutan (*ponAbe2*), chimpanzee (*panTro3*) and human (*hg19*) reference genomes, and so forth).

The alignment was performed using BLAT [Kent, 2002] on the syntenic regions of the corresponding SV (top levels of the Net alignments downloaded from UCSC (http://genome.ucsc.edu/cgi-bin/hgGateway) for each species). For example, when assessing SVs identified as deletion by SV discovery pipeline, if the alternative junction sequence from one species mapped with better sequence identity and length (compared with the reference junction sequence from the inspected species) onto one of the four corresponding syntenic regions, the event was rectified as 'insertion'; if the reference junction sequences from the inspected species mapped better than alternative junction sequences, the event was rectified as 'deletion' (see Lam et al. [2010] for details). Events were 'unrectifiable', if we failed to identify an alignment between the junction sequences obtained from the query species and the corresponding syntenic regions from the other species. Deletions and duplications rectified as insertions indicate that an insertion into ancestral genomic sequence, rather than a sequence deletion, has occurred. The respective sequences were subjected to BLAT analysis to determine the donor locus.

Non-reference mobile element insertion discovery

The TEA pipeline [Lee et al., 2012] was used to perform non-reference MEI discovery. The repeat sequence library required by TEA was constructed by concatenating multiple consensus subfamily sequences separated by multiple 'N' nucleotide spacers. To represent L1/LINE elements, consensus subfamily sequences for L1HS, L1PA3, L1PA5, L1Pt were used; for *Alu* elements, consensus subfamily sequences for *AluJb*, *AluSx*, *AluY*, *AluMacYa3*, *AluYe5a2_Pongo*,

AluYc1a5_Pongo, and *AluYe5b5_Pongo* [Walker et al., 2012] were used; for SVA elements, the sequences of six SVA subfamilies (SVA_A/B/C/D/E/F) and of the general SVA consensus sequence were used.

Candidate insertion sites were considered as high-confidence if they satisfied the following criteria: (1) more than three supporting repeat-anchored mate (RAM) reads were observed, and at least one RAM on each side of the insertion was observed; (2) at least one positive and negative strand soft-clipped read was observed within the RAM cluster boundary; (3) the gap between two insertion breakpoints defined by negative and positive strand clipped reads was within [-20, 50]; (4) the ratio of well-aligned clipped reads over all clipped reads was at least 0.5. Insertion loci within 500 bp margin from the instances of the same mobile element family annotated in the reference genome were removed. Mobile element insertions located in gapped regions of the reference genome were annotated as such and removed from the final data set. Following their discovery in individual samples, non-reference *Alu*, L1 and SVA insertions were merged across samples. The list of non-reference MEIs was merged with the list of reference MEIs (mobile element insertions identified as a deletion relative to the respective reference genome) for pursuing SV formation mechanism analyses.

SNP discovery

SNPs were identified using the Genome Analysis Toolkit (GATK) McKenna et al. [2010] and Samtools [Li et al., 2009]. GATK base quality score recalibration and realignment was subsequently applied, and SNP discovery and genotyping across all samples simultaneously was performed using standard hard filtering parameters. The consensus of multiple primary callsets from GATK and Samtools was used for further analysis. For each sample, a series of filters were applied to remove potential false positives. Candidate SNPs mapping to gaps in the reference were removed or segmental duplications, as well as SNPs with a Phred quality score $\leq$10 were excluded. Also SNPs within 10 bp of each other were discarded, in order to minimize the rate of false positives caused by recent segmental duplications. For orangutan and rhesus macaque, those SNPs falling into regions in the reference genome with low consensus quality score $<$90 (on a scale of 1-97, based on the Phred scores of underlying whole-genome shotgun reads) and $<$60 (on a scale of 1-60), respectively. By Sanger sequencing we validated 238 out of 241 SNPs, with false discovery rate (FDRs) of 1.2%.

Segmental duplication maps

High-copy repeats annotated in the UCSC RepeatMasker table (http://hgdownload.soe.ucsc.edu/) were initially removed from each reference assembly and segmental duplications (SDs) were identified by aligning each chromosome with itself (intrachromosomal SDs) and to all other chromosomes (interchromosomal SDs). Maximal exact matches (MEMs) of a minimal length = 17 bp were computed using the vmatch software

(http://www.vmatch.de). Using CHAINER [Abouelhoda and Ohlebusch, 2004] MEMs were then connected with the following parameters: -length 34 -gc 100 -lw 8 and the created chains were extended using -length 100 -gc 1000 -lw 14 (this recursive chaining strategy is described in [Abouelhoda et al., 2008]). High-copy repeat sequences were re-inserted into the resulting chains and chains smaller than 1000 bp were discarded. The remaining chains were globally aligned using EMBOSS 'stretcher' when the sequence length was greater than 100 Mb, or EMBOSS 'needle' otherwise. Alignments showing less than 90% sequence identity, or a gap percentage larger than 30% were discarded.

Novelty of variant calls

In order to inspect novelty of variant calls, merged SV calls (high confidence discovery dataset, deletions and duplications) were compared to published SV calls from array CGH experiments. Calls were compiled from dbVar [Auton et al., 2012, Gokcumen et al., 2011, Yan et al., 2011] and signature papers [Lee et al., 2012, Prüfer et al., 2012, The 1000 Genomes Project Consortium, 2012] and converted to the respective reference genomes *panTro3*, *ponAbe2* and *rheMac2* using the liftOver tool (http://genome.ucsc.edu/cgi-bin/hgLiftOver). The overlap cutoff was set to a minimum of 1 bp between known SV and novel SV from our datasets (Figure A.14, lower panel).

B.4 Methods for Chapter 5

LFS-MB data access

Sequence data analyzed in this study can be accessed from European Genome-phenome Archive (EGA) through the following accession number: EGAS00001000085.

Patients

Patients Informed consent and an ethical vote (Institutional Review Board) were obtained according to ICGC guidelines. No patient underwent chemotherapy or radiotherapy prior to the surgical removal of the primary tumor.

Sequencing of paired-reads and paired-end mapping of LFS-MB data

The sequencing of tumor and corresponding germline sample pairs (5μg DNA each) was performed on Illumina HiSeq 2000 and Genome Analyzer II instruments using paired-end libraries. The raw length of the read was 101 bp, and the median insert size was 285-325 bp ('Illumina paired-end [PE] protocol'). Sequenced reads were aligned onto the human reference (*hg19* assembly downloaded from UCSC Genome Browser, http://genome.ucsc.edu/) using Illumina's ELAND version 2, followed by conversion of raw files into SAM/BAM format [Li et al., 2009] and subsequent variant discovery by DELLY [Rausch et al., 2012b].

Structural variant discovery in LFS-MB data

All discordantly (abnormally) mapped reads-pairs were used as a signature to detect structural variations: deletions, tandem duplications, inversions and interchromosomal rearrangements consistent with translocation signatures. Tumor-specific calls were determined based on absence of 80% reciprocal overlap with variants in matched normal sample or 50% reciprocal overlap with known 1000 Genome Project (1000GP) variants. High-quality deletions were determined based on at least four supporting pairs, or minimum of two supporting pairs and a supporting breakpoint-spanning splitread. For tandem duplication, we required a minimum of two supporting pairs and one split read. High quality inversions and interchromosomal rearrangement were identified based on at least two supporting pairs (both breakpoints) and one splitread (one breakpoint). For all variants larger than 100 kb we required additional read-depth-based support.

SV formation mechanism analysis

SV formation mechanism inference was performed using the BreakSeq tool [Lam et al., 2010], version 1.3. Short template or non-template insertions (microinsertions) were inferred using DELLY tool [Rausch et al., 2012b]. A random subset of microhomologies and microinsertions automatically detected with BreakSeq and DELLY, respectively, were validated using BLAT at the UCSC Genome Browser (https://genome.ucsc.edu/cgi-bin/hgBlat?command=start).

For Conrad et al. [2010], Lam et al. [2010] and Mills et al. [2011] datasets, BreakSeq was used with hg18 version of human reference genome and the corresponding RepeatMasker file, whereas for Kidd et al. [2010] dataset hg17 version was used (all downloaded from UCSC - http://genome.ucsc.edu/cgi-bin/hgGateway). In the LFS-MB study, *hg19* version of human reference genome and the corresponding RepeatMasker was used in conjunction with the BreakSeq tool. In order to use a given SV for the mechanism classification analysis, a minimum of four supporting pairs was required. Additionally calls found in highly-amplified regions were not taken into consideration.

Appendix C

List of publications

Rausch T.*, Jones D.*, Zapatka M.*, Stütz A.*, Zichner T., Weischenfeldt J., Jäger N., Remke M., Shih D., Northcott P., Pfaff E., Tica J. et al. Genome Sequencing of Pediatric Medulloblastoma Links Catastrophic DNA Rearrangements with TP53 Mutations. *Cell*, 148(1-2):59–71, January 2012

Gokcumen O.*, Tischler V.*, Tica J., Zhu Q., Iskow R. C., Lee E., Fritz M. H.-Y., Langdon A., Stütz A. M., Pavlidis P. et al. Primate genome architecture influences structural variation mechanisms and functional consequences. *Proceedings of the National Academy of Sciences of the United States of America*, 110(39):15764–9, September 2013.

Tica J., Lee E., Klaus B., Gokcumen O., Park P. J., Stütz A. M.*, Korbel J. O.* TIGER: Detection of L1-mediated 3' Transductions In GERmline using next-generation sequencing data (*Manuscript in preparation*)

*These authors contributed equally

Bibliography

Abecasis G. R., Altshuler D., Auton A., Brooks L. D., Durbin R. M., Gibbs R. A., Hurles M. E., and McVean G. A. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–73, October 2010.

Abouelhoda M. I. and Ohlebusch E. CHAINER: Software for comparing genomes. *Short paper at ISMB/ECCB 2004. (12th International Conference on Intelligent Systems for Molecular Biology/3rd European Conference on Computational Biology)*, 2004.

Abouelhoda M. I., Kurtz S., and Ohlebusch E. CoCoNUT: an efficient system for the comparison and analysis of genomes. *BMC bioinformatics*, 9:476, January 2008.

Abyzov A., Urban A. E., Snyder M., and Gerstein M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome research*, 21(6):974–84, June 2011.

Alkan C., Coe B. P., and Eichler E. E. Genome structural variation discovery and genotyping. *Nature reviews. Genetics*, 12(5):363–76, May 2011.

Altschul S. F., Gish W., Miller W., Myers E. W., and Lipman D. J. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, October 1990.

Arcot S. S., Wang Z., Weber J. L., Deininger P. L., and Batzer M. A. Alu Repeats: A Source for the Genesis of Primate Microsatellites. *Genomics*, 29(1):136–144, 1995.

Auton A., Fledel-Alon A., Pfeifer S., Venn O., Ségurel L., Street T., Leffler E. M., Bowden R., Aneas I., Broxholme J. et al. A Fine-Scale Chimpanzee Genetic Map from Population Sequencing. *Science*, 336(6078):193–198, 2012.

Babushok D. V., Ohshima K., Ostertag E. M., Chen X., Wang Y., Mandal P. K., Okada N., Abrams C. S., and Kazazian H. H. A novel testis ubiquitin-binding protein gene arose by exon shuffling in hominoids, August 2007.

Bailey J. A. and Eichler E. E. Primate segmental duplications: crucibles of evolution, diversity and disease. *Nat Rev Genet*, 7(7):552–564, July 2006.

Bailey J. A., Gu Z., Clark R. A., Reinert K., Samonte R. V., Schwartz S., Adams M. D., Myers E. W., Li P. W., and Eichler E. E. Recent Segmental Duplications in the Human Genome. *Science*, 297(5583):1003–1007, 2002.

Bailey J. A., Liu G., and Eichler E. E. An Alu Transposition Model for the Origin and Expansion of Human Segmental Duplications, October 2003.

Bailey T. L. and Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology*, 2:28–36, 1994.

Baillie J. K., Barnett M. W., Upton K. R., Gerhardt D. J., Richmond T. a., De Sapio F., Brennan P., Rizzu P., Smith S., Fell M. et al. Somatic retrotransposition alters the genetic landscape of the human brain. *Nature*, pages 2–5, October 2011.

Baker M. Structural variation: the genome's hidden architecture. *Nat Meth*, 9(2):133–137, February 2012.

Barreiro L. B., Laval G., Quach H., Patin E., and Quintana-Murci L. Natural selection has driven population differentiation in modern humans. *Nat Genet*, 40(3):340–345, March 2008.

Batzer M. a. and Deininger P. L. Alu repeats and human genomic diversity. *Nature reviews. Genetics*, 3(5):370–9, May 2002.

Briggs K. J., Corcoran-Schwartz I. M., Zhang W., Harcke T., Devereux W. L., Baylin S. B., Eberhart C. G., and Watkins D. N. Cooperation between the Hic1 and Ptch1 tumor suppressors in medulloblastoma, March 2008.

Brookes K. J. The VNTR in complex disorders: the forgotten polymorphisms? A functional way forward? *Genomics*, 101(5):273–81, May 2013.

Brouha B., Schustak J., Badge R. M., Lutz-Prigge S., Farley A. H., Moran J. V., and Kazazian H. H. Hot L1s account for the bulk of retrotransposition in the human population. *Proceedings of the National Academy of Sciences of the United States of America*, 100(9):5280–5, April 2003.

Bühren J., Christoph A. H., Buslei R., Albrecht S., Wiestler O. D., and Pietsch T. Expression of the neurotrophin receptor p75NTR in medulloblastomas is correlated with distinct histological and clinical features: evidence for a medulloblastoma subtype derived from the external granule cell layer. *Journal of neuropathology and experimental neurology*, 59(3):229–240, March 2000.

Callinan P. A., Wang J., Herke S. W., Garber R. K., Liang P., and Batzer M. A. Alu Retrotransposition-mediated Deletion. *Journal of Molecular Biology*, 348(4):791–800, 2005.

Campbell P. J., Stephens P. J., Pleasance E. D., O'Meara S., Li H., Santarius T., Stebbings L. A., Leroy C., Edkins S., Hardy C. et al. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat Genet*, 40(6):722–729, June 2008.

Carbone L., Harris R. A., Mootnick A. R., Milosavljevic A., Martin D. I. K., Rocchi M., Capozzi O., Archidiacono N., Konkel M. K., Walker J. A. et al. Centromere Remodeling in Hoolock leuconedys (Hylobatidae) by a New Transposable Element Unique to the Gibbons. *Genome Biology and Evolution*, 4(7):760–770, 2012.

Chen J.-M., Stenson P. D., Cooper D. N., and Férec C. A systematic analysis of LINE-1 endonuclease-dependent retrotranspositional events causing human genetic disease. *Human genetics*, 117(5):411–27, September 2005.

Chen K., Wallis J. W., McLellan M. D., Larson D. E., Kalicki J. M., Pohl C. S., McGrath S. D., Wendl M. C., Zhang Q., Locke D. P. et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature methods*, 6(9):677–81, September 2009.

Chen K., Chen L., Fan X., Wallis J., Ding L., and Weinstock G. TIGRA: a targeted iterative graph routing assembler for breakpoint assembly. *Genome research*, 24(2):310–7, February 2014.

Chiang C., Jacobsen J. C., Ernst C., Hanscom C., Heilbut A., Blumenthal I., Mills R. E., Kirby A., Lindgren A. M., Rudiger S. R. et al. Complex reorganization and predominant non-homologous repair following chromosomal breakage in karyotypically balanced germline rearrangements and transgenic integration. *Nature genetics*, 44(4):390–7, S1, April 2012.

Chu Y. and Corey D. R. RNA Sequencing: Platform Selection, Experimental Design, and Data Interpretation, August 2012.

Clarke J., Wu H.-C., Jayasinghe L., Patel A., Reid S., and Bayley H. Continuous base identification for single-molecule nanopore DNA sequencing. *Nat Nano*, 4(4):265–270, April 2009.

Conlee K. M., Hoffeld E. H., and Stephens M. L. A demographic analysis of primate research in the United States. *Alternatives to laboratory animals : ATLA*, 32 Suppl 1:315–322, June 2004.

Conrad D. F., Bird C., Blackburne B., Lindsay S., Mamanova L., Lee C., Turner D. J., and Hurles M. E. Mutation spectrum revealed by breakpoint sequencing of human germline CNVs. *Nat Genet*, 42(5):385–391, May 2010.

Cooper G. M. *Elements of Human Cancer*. Biology Series. Jones and Bartlett Publishers, 1992. ISBN 9780867201918.

Cooper G. M., Zerr T., Kidd J. M., Eichler E. E., and Nickerson D. A. Systematic assessment of copy-number variant detection via genome-wide single nucleotide polymorphism genotyping, October 2008.

Cordaux R. The human genome in the LINE of fire. *Proceedings of the National Academy of Sciences of the United States of America*, 105(49):19033–4, December 2008.

Cordaux R. and Batzer M. a. The impact of retrotransposons on human genome evolution. *Nature reviews. Genetics*, 10(10):691–703, October 2009.

Costantini M. and Bernardi G. Replication timing, chromosomal bands, and isochores. *Proceedings of the National Academy of Sciences of the United States of America*, 105(9):3433–3437, March 2008.

Crasta K., Ganem N. J., Dagher R., Lantermann A. B., Ivanova E. V., Pan Y., Nezi L., Protopopov A., Chowdhury D., and Pellman D. DNA breaks and chromosome pulverization from errors in mitosis. *Nature*, January 2012.

Croce C. M. Oncogenes and Cancer. *New England Journal of Medicine*, 358(5):502–511, 2008.

Dahm-Daphi J., Hubbe P., Horvath F., El-Awady R. A., Bouffard K. E., Powell S. N., and Willers H. Nonhomologous end-joining of site-specific but not of radiation-induced DNA double-strand breaks is reduced in the presence of wild-type p53. *Oncogene*, 24(10):1663–1672, March 2005.

Damert A., Raiz J., Horn A. V., Löwer J., Wang H., Xing J., Batzer M. a., Löwer R., and Schumann G. G. 5'-Transducing SVA retrotransposon groups spread efficiently throughout the human genome. *Genome research*, 19(11):1992–2008, November 2009.

de Wit E. and de Laat W. A decade of 3C technologies: insights into nuclear organization, January 2012.

Deininger P. L. and Batzer M. a. Alu repeats and human disease. *Molecular genetics and metabolism*, 67(3):183–93, July 1999.

Deliard S., Zhao J., Xia Q., and Grant S. F. A. Generation of High Quality Chromatin Immunoprecipitation DNA Template for High-throughput Sequencing (ChIP-seq). 19(74):e50286, 2013.

Dewannieux M., Esnault C., and Heidmann T. LINE-mediated retrotransposition of marked Alu sequences. *Nature genetics*, 35(1):41–48, September 2003.

Eid J., Fehr A., Gray J., Luong K., Lyle J., Otto G., Peluso P., Rank D., Baybayan P., Bettman B. et al. Real-Time DNA Sequencing from Single Polymerase Molecules. *Science*, 323(5910): 133–138, 2009.

Esnault C., Maestre J., and Heidmann T. Human LINE retrotransposons generate processed pseudogenes. *Nat Genet*, 24(4):363–367, April 2000.

Ewing A. D., Ballinger T. J., Earl D., Harris C. C., Ding L., Wilson R. K., and Haussler D. Retrotransposition of gene transcripts leads to structural variation in mammalian genomes. *Genome biology*, 14(3):R22, January 2013.

Feuk L., Carson A. R., and Scherer S. W. Structural variation in the human genome. *Nat Rev Genet*, 7(2):85–97, February 2006.

Forment J. V., Kaidi A., and Jackson S. P. Chromothripsis and cancer: causes and consequences of chromosome shattering. *Nat Rev Cancer*, 12(10):663–670, October 2012.

Forozan F., Karhu R., Kononen J., Kallioniemi A., and Kallioniemi O.-P. Genome screening by comparative genomic hybridization. *Trends in Genetics*, 13(10):405–409, 1997.

Frazer K. A., Murray S. S., Schork N. J., and Topol E. J. Human genetic variation and its contribution to complex traits. *Nat Rev Genet*, 10(4):241–251, April 2009.

Freese E. The difference between spontaneous and base-analogue induced mutations of phage T4, April 1959.

Fujita P. A., Rhead B., Zweig A. S., Hinrichs A. S., Karolchik D., Cline M. S., Goldman M., Barber G. P., Clawson H., Coelho A. et al. The UCSC Genome Browser database: update 2011. *Nucleic acids research*, 39(Database issue):D876–82, January 2011.

Gasior S. L., Wakeman T. P., Xu B., and Deininger P. L. The Human LINE-1 Retrotransposon Creates {DNA} Double-strand Breaks. *Journal of Molecular Biology*, 357(5):1383–1393, 2006.

Gazave E., Darre F., Morcillo-Suarez C., Petit-Marty N., Carreno A., Marigorta U. M., Ryder O. A., Blancher A., Rocchi M., Bosch E. et al. Copy number variation analysis in the great apes reveals species-specific patterns of structural variation. *Genome research*, 21(10):1626–1639, October 2011.

Gilbert N., Lutz-Prigge S., and Moran J. V. Genomic Deletions Created upon LINE-1 Retrotransposition. *Cell*, 110(3):315–325, 2002.

Gokcumen O., Babb P., Iskow R., Zhu Q., Shi X., Mills R., Ionita-Laza I., Vallender E., Clark A., Johnson W. et al. Refinement of primate copy number variation hotspots identifies candidate genomic regions evolving under positive selection. *Genome Biology*, 12(5):R52, 2011.

Gokcumen O., Tischler V., Tica J., Zhu Q., Iskow R. C., Lee E., Fritz M. H.-Y., Langdon A., Stütz A. M., Pavlidis P. et al. Primate genome architecture influences structural variation mechanisms and functional consequences. *Proceedings of the National Academy of Sciences of the United States of America*, 110(39):15764–9, September 2013.

Goodier J. L., Ostertag E. M., and Kazazian H. H. Transduction of 3'-flanking sequences is common in L1 retrotransposition. *Human molecular genetics*, 9(4):653–7, March 2000.

Gu W., Zhang F., and Lupski J. R. Mechanisms for human genomic rearrangements. *PathoGenetics*, 1(1):4, January 2008.

Han K., Sen S. K., Wang J., Callinan P. a., Lee J., Cordaux R., Liang P., and Batzer M. a. Genomic rearrangements by LINE-1 insertion-mediated deletion in the human and chimpanzee lineages. *Nucleic acids research*, 33(13):4040–52, January 2005.

Han K., Konkel M. K., Xing J., Wang H., Lee J., Meyer T. J., Huang C. T., Sandifer E., Hebert K., Barnes E. W. et al. Mobile DNA in Old World monkeys: a glimpse through the rhesus macaque genome. *Science (New York, N.Y.)*, 316(5822):238–40, April 2007.

Han K., Lee J., Meyer T. J., Remedios P., Goodwin L., and Batzer M. a. L1 recombination-associated deletions generate human genomic variation. *Proceedings of the National Academy of Sciences of the United States of America*, 105(49):19366–71, December 2008.

Hanahan D. and Weinberg R. A. The Hallmarks of Cancer. *Cell*, 100(1):57–70, January 2000.

Hanahan D. and Weinberg R. a. Hallmarks of cancer: the next generation. *Cell*, 144(5):646–74, March 2011.

Hancks D. C. and Kazazian H. H. Active human retrotransposons: variation and disease. *Cordaux, Richard Batzer, Mark a*, 22(3):191–203, June 2012.

Handsaker R. E., Korn J. M., Nemesh J., and McCarroll S. a. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nature genetics*, 43(3): 269–76, March 2011.

Hansen R. S., Thomas S., Sandstrom R., Canfield T. K., Thurman R. E., Weaver M., Dorschner M. O., Gartler S. M., and Stamatoyannopoulos J. a. Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proceedings of the National Academy of Sciences of the United States of America*, 107(1):139–44, January 2010.

Harris R. S. and Chiaromonte F. Improved pairwise alignment of genomic DNA. Technical report, 2007.

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

Hastings P. J., Lupski J. R., Rosenberg S. M., and Ira G. Mechanisms of change in gene copy number. *Nature reviews. Genetics*, 10(8):551–64, August 2009b.

Helman E., Lawrence M. S., Stewart C., Sougnez C., Getz G., and Meyerson M. Somatic retrotransposition in human cancer revealed by whole-genome and exome sequencing. *Genome research*, 24(7):1053–63, July 2014.

Hinrichs A. S., Karolchik D., Baertsch R., Barber G. P., Bejerano G., Clawson H., Diekhans M., Furey T. S., Harte R. A., Hsu F. et al. The UCSC Genome Browser Database: update 2006. *Nucleic acids research*, 34(Database issue):D590–8, January 2006.

Hirsch D., Kemmerling R., Davis S., Camps J., Meltzer P. S., Ried T., and Gaiser T. Chromothripsis and focal copy number alterations determine poor outcome in malignant melanoma. *Cancer research*, December 2012.

Hormozdiari F., Hajirasouliha I., Dao P., Hach F., Yorukoglu D., Alkan C., Eichler E. E., and Sahinalp S. C. Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. *Bioinformatics (Oxford, England)*, 26(12):i350–7, June 2010.

Hu J. and Ng P. C. Predicting the effects of frameshifting indels. *Genome biology*, 13(2):R9, January 2012.

Huang X. and Madan A. CAP3: A DNA sequence assembly program. *Genome research*, 9(9): 868–877, September 1999.

Iafrate A. J., Feuk L., Rivera M. N., Listewnik M. L., Donahoe P. K., Qi Y., Scherer S. W., and Lee C. Detection of large-scale variation in the human genome. *Nat Genet*, 36(9):949–951, September 2004.

Iskow R. C., McCabe M. T., Mills R. E., Torene S., Pittard W. S., Neuwald A. F., Van Meir E. G., Vertino P. M., and Devine S. E. Natural mutagenesis of human genomes by endogenous retrotransposons. *Cell*, 141(7):1253–1261, June 2010.

Jones D. T. W., Jager N., Kool M., Zichner T., Hutter B., Sultan M., Cho Y.-J., Pugh T. J., Hovestadt V., Stutz A. M. et al. Dissecting the genomic complexity underlying medulloblastoma. *Nature*, 488(7409):100–105, August 2012.

Kaer K. and Speek M. Retroelements in human disease. *Gene*, 518(2):231–241, 2013.

Kaessmann H., Vinckenbosch N., and Long M. RNA-based gene duplication: mechanistic and evolutionary insights. *Nature reviews. Genetics*, 10(1):19–31, January 2009.

Kano H., Godoy I., Courtney C., Vetter M. R., Gerton G. L., Ostertag E. M., and Kazazian H. H. L1 retrotransposition occurs mainly in embryogenesis and creates somatic mosaicism. *Genes & development*, 23(11):1303–12, June 2009.

Kass D. H., Batzer M. A., and Deininger P. L. Gene conversion as a secondary mechanism of short interspersed element (SINE) evolution., January 1995.

Kazazian H. H., Wong C., Youssoufian H., Scott A. F., Phillips D. G., and Antonarakis S. E. Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. *Nature*, 332(6160):164–166, March 1988.

Keane T. M., Wong K., and Adams D. J. RetroSeq: transposable element discovery from next-generation sequencing data. *Bioinformatics (Oxford, England)*, 29(3):389–90, February 2013.

Kent W. J. BLAT—The BLAST-Like Alignment Tool. *Genome Research*, 12(4):656–664, March 2002.

Khaitovich P., Enard W., Lachmann M., and Paabo S. Evolution of primate gene expression. *Nat Rev Genet*, 7(9):693–702, September 2006.

Kidd J. M., Graves T., Newman T. L., Fulton R., Hayden H. S., Malig M., Kallicki J., Kaul R., Wilson R. K., and Eichler E. E. A human genome structural variation sequencing resource reveals insights into mutational mechanisms. *Cell*, 143(5):837–47, November 2010.

Kloosterman W. P., Guryev V., van Roosmalen M., Duran K. J., de Bruijn E., Bakker S. C. M., Letteboer T., van Nesselrooij B., Hochstenbach R., Poot M. et al. Chromothripsis as a mechanism driving complex de novo structural rearrangements in the germline. *Human molecular genetics*, 20(10):1916–24, May 2011.

Kloosterman W. P., Tavakoli-Yaraki M., van Roosmalen M. J., van Binsbergen E., Renkens I., Duran K., Ballarati L., Vergult S., Giardino D., Hansson K. et al. Constitutional chromothripsis rearrangements involve clustered double-stranded DNA breaks and nonhomologous repair mechanisms. *Cell reports*, 1(6):648–55, June 2012.

Knudson A. G. Mutation and Cancer: Statistical Study of Retinoblastoma, April 1971.

Knudson A. G. Two genetic hits (more or less) to cancer. *Nat Rev Cancer*, 1(2):157–162, November 2001.

Kojima K. K. Alu monomer revisited: recent generation of Alu monomers. *Molecular biology and evolution*, 28(1):13–5, January 2011.

Kool M., Korshunov A., Remke M., Jones D., Schlanstein M., Northcott P., Cho Y.-J., Koster J., Schouten-van Meeteren A., van Vuurden D. et al. Molecular subgroups of medulloblastoma: an international meta-analysis of transcriptome, genetic aberrations, and clinical data of WNT, SHH, Group 3, and Group 4 medulloblastomas. *Acta Neuropathologica*, 123(4):473–484, 2012.

Korbel J. O. and Campbell P. J. Criteria for inference of chromothripsis in cancer genomes. *Cell*, 152(6):1226–36, March 2013.

Korbel J. O., Urban A. E., Affourtit J. P., Godwin B., Grubert F., Simons J. F., Kim P. M., Palejev D., Carriero N. J., Du L. et al. Paired-end mapping reveals extensive structural variation in the human genome. *Science (New York, N.Y.)*, 318(5849):420–6, October 2007.

Korbel J. O., Abyzov A., Mu X. J., Carriero N., Cayting P., Zhang Z., Snyder M., and Gerstein M. B. PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome biology*, 10(2): R23, January 2009a.

Korbel J. O., Tirosh-Wagner T., Urban A. E., Chen X.-N., Kasowski M., Dai L., Grubert F., Erdman C., Gao M. C., Lange K. et al. The genetic architecture of Down syndrome phenotypes revealed by high-resolution analysis of human segmental trisomies. *Proceedings of the National Academy of Sciences*, 106(29):12031–12036, 2009b.

LaFramboise T. Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances, July 2009.

Lakich D., Kazazian H. H., Antonarakis S. E., and Gitschier J. Inversions disrupting the factor VIII gene are a common cause of severe haemophilia A. *Nat Genet*, 5(3):236–241, November 1993.

Lam H. Y. K., Mu X. J., Stütz A. M., Tanzer A., Cayting P. D., Snyder M., Kim P. M., Korbel J. O., and Gerstein M. B. Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. *Nature biotechnology*, 28(1):47–55, January 2010.

Lander E. S., Linton L. M., Birren B., Nusbaum C., Zody M. C., Baldwin J., Devon K., Dewar K., Doyle M., FitzHugh W. et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, February 2001.

Lassmann T. and Sonnhammer E. L. L. Kalign–an accurate and fast multiple sequence alignment algorithm. *BMC bioinformatics*, 6:298, January 2005.

Leader R. W. and Stark D. The importance of animals in biomedical research. *Perspectives in biology and medicine*, 30(4):470–485, 1987.

Lee A. S., Gutierrez-Arcelus M., Perry G. H., Vallender E. J., Johnson W. E., Miller G. M., Korbel J. O., and Lee C. Analysis of copy number variation in the rhesus macaque genome identifies candidate loci for evolutionary and human disease studies. *Human molecular genetics*, 17(8):1127–1136, April 2008.

Lee E., Iskow R., Yang L., Gokcumen O., Haseley P., Luquette L. J., Lohr J. G., Harris C. C., Ding L., Wilson R. K. et al. Landscape of somatic retrotransposition in human cancers. *Science (New York, N.Y.)*, 337(6097):967–71, August 2012.

Lee J. a., Carvalho C. M. B., and Lupski J. R. A DNA replication mechanism for generating non-recurrent rearrangements associated with genomic disorders. *Cell*, 131(7):1235–47, December 2007.

Li F. P. and Fraumeni JR J. F. Soft-Tissue Sarcomas, Breast Cancer, and Other NeoplasmsA Familial Syndrome? *Annals of Internal Medicine*, 71(4):747–752, October 1969.

Li H. and Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25:1754–1760, 2009.

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., and Durbin R. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16):2078–2079, August 2009.

Li R., Zhu H., Ruan J., Qian W., Fang X., Shi Z., Li Y., Li S., Shan G., Kristiansen K. et al. De novo assembly of human genomes with massively parallel short read sequencing. *Genome research*, 20(2):265–72, February 2010.

Lieber M. R. The mechanism of human nonhomologous DNA end joining. *The Journal of biological chemistry*, 283(1):1–5, January 2008.

Lieber M. R. The mechanism of double-strand DNA break repair by the nonhomologous DNA end-joining pathway. *Annual review of biochemistry*, 79(D):181–211, January 2010.

Lieber M. R., Ma Y., Pannicke U., and Schwarz K. Mechanism and regulation of human non-homologous DNA end-joining. *Nature reviews. Molecular cell biology*, 4(9):712–20, September 2003.

Liu G. E., Alkan C., Jiang L., Zhao S., and Eichler E. E. Comparative analysis of Alu repeats in primate genomes. *Genome Research*, 19(5):876–885, May 2009.

Liu P., Erez A., Sreenath Nagamani S. C., Dhar S. U., Kołodziejska K. E., Dharmadhikari A. V., Cooper M. L., Wiszniewska J., Zhang F., Withers M. A. et al. Chromosome catastrophes involve replication mechanisms generating complex genomic rearrangements, September 2011.

Locke D. P., Hillier L. W., Warren W. C., Worley K. C., Nazareth L. V., Muzny D. M., Yang S.-P., Wang Z., Chinwalla A. T., Minx P. et al. Comparative and demographic analysis of orang-utan genomes. *Nature*, 469(7331):529–33, January 2011.

Lou S., Lee H.-M., Qin H., Li J.-W., Gao Z., Liu X., Chan L., KL Lam V., So W.-Y., Wang Y. et al. Whole-genome bisulfite sequencing of multiple individuals reveals complementary roles of promoter and gene body methylation in transcriptional regulation. *Genome Biology*, 15(7): 408, 2014.

Magrangeas F., Avet-Loiseau H., Munshi N. C., and Minvielle S. Chromothripsis identifies a rare and aggressive entity among newly diagnosed multiple myeloma patients. *Blood*, 118(3): 675–8, July 2011.

Maher C. A. and Wilson R. K. Chromothripsis and Human Disease: Piecing Together the Shattering Process, January 2012.

Majewski J., Schwartzentruber J., Lalonde E., Montpetit A., and Jabado N. What can exome sequencing do for you? *Journal of Medical Genetics*, 48(9):580–589, 2011.

Malik H. S., Burke W. D., and Eickbush T. H. The age and evolution of non-LTR retrotransposable elements. *Molecular biology and evolution*, 16(6):793–805, June 1999.

Malkin D., Li F. P., Strong L. C., Fraumeni J. F., Nelson C. E., Kim D. H., Kassel J., Gryka M. A., Bischoff F. Z., Tainsky M. A. et al. Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. *Science*, 250(4985):1233–1238, November 1990.

Mardis E. R. Next-generation sequencing platforms. *Annual review of analytical chemistry (Palo Alto, Calif.)*, 6:287–303, January 2013.

Marques A. C., Dupanloup I., Vinckenbosch N., Reymond A., and Kaessmann H. Emergence of young human genes after a burst of retroposition in primates. *PLoS biology*, 3(11):e357, November 2005.

Marques-Bonet T., Ryder O. a., and Eichler E. E. Sequencing primate genomes: what have we learned? *Annual review of genomics and human genetics*, 10:355–86, January 2009.

McClintock B. Controlling elements and the gene. *Cold Spring Harbor Symposia on Quantitative Biology*, 21:197–216, January 1956.

McKenna A., Hanna M., Banks E., Sivachenko A., Cibulskis K., Kernytsky A., Garimella K., Altshuler D., Gabriel S., Daly M. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, 20(9):1297–1303, September 2010.

Mendel G. Versuche über Pflanzen-Hybriden. *Verhandlungen des naturforschenden Vereines in Brünn*, 42:3–47, 1866.

Miki Y., Nishisho I., Horii A., Miyoshi Y., Utsunomiya J., Kinzler K. W., Vogelstein B., and Nakamura Y. Disruption of the APC gene by a retrotransposal insertion of L1 sequence in a colon cancer. *Cancer research*, 52(3):643–645, February 1992.

Mills R. E., Bennett E. A., Iskow R. C., Luttig C. T., Tsui C., Pittard W. S., and Devine S. E. Recently mobilized transposons in the human and chimpanzee genomes. *American journal of human genetics*, 78(4):671–9, April 2006.

Mills R. E., Bennett E. A., Iskow R. C., and Devine S. E. Which transposable elements are active in the human genome? *Trends in genetics : TIG*, 23(4):183–91, April 2007.

Mills R. E., Walter K., Stewart C., Handsaker R. E., Chen K., Alkan C., Abyzov A., Yoon S. C., Ye K., Cheetham R. K. et al. Mapping copy number variation by population-scale genome sequencing. *Nature*, 470(7332):59–65, February 2011.

Molenaar J. J., Koster J., Zwijnenburg D. a., van Sluis P., Valentijn L. J., van der Ploeg I., Hamdi M., van Nes J., Westerman B. a., van Arkel J. et al. Sequencing of neuroblastoma identifies chromothripsis and defects in neuritogenesis genes. *Nature*, 483(7391):589–93, March 2012.

Montgomery S. B., Goode D., Kvikstad E., Albers C. A., Zhang Z., Mu X. J., Ananda G., Howie B., Karczewski K. J., Smith K. S. et al. The origin, evolution and functional impact of short insertion-deletion variants identified in 179 human genomes. *Genome Research*, 23(5): gr.148718.112—-761, March 2013.

Moran J. V., DeBerardinis R. J., and Kazazian H. H. Exon Shuffling by L1 Retrotransposition. *Science*, 283(5407):1530–1534, 1999.

Morrish T. a., Gilbert N., Myers J. S., Vincent B. J., Stamato T. D., Taccioli G. E., Batzer M. a., and Moran J. V. DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. *Nature genetics*, 31(2):159–65, June 2002.

Mullaney J. M., Mills R. E., Pittard W. S., and Devine S. E. Small insertions and deletions (INDELs) in human genomes. *Human molecular genetics*, 19(R2):R131–6, October 2010.

Muotri A. R., Chu V. T., Marchetto M. C. N., Deng W., Moran J. V., and Gage F. H. Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. *Nature*, 435(7044): 903–10, June 2005.

Nagy E. and Maquat L. E. A rule for termination-codon position within intron-containing genes: when nonsense affects {RNA} abundance. *Trends in Biochemical Sciences*, 23(6):198–199, 1998.

Naylor J., Brinke A., Hassock S., Green P. M., and Giannelli F. Characteristic mRNA abnormality found in half the patients with severe haemophilia A is due to large DNA inversions. *Human Molecular Genetics*, 2(11):1773–1778, November 1993.

Ng P. C. and Henikoff S. Predicting the effects of amino acid substitutions on protein function. *Annual review of genomics and human genetics*, 7:61–80, January 2006.

Northcott P. A., Jones D. T. W., Kool M., Robinson G. W., Gilbertson R. J., Cho Y.-J., Pomeroy S. L., Korshunov A., Lichter P., Taylor M. D. et al. Medulloblastomics: the end of the beginning. *Nat Rev Cancer*, 12(12):818–834, December 2012a.

Northcott P. A., Shih D. J. H., Peacock J., Garzia L., Morrissy A. S., Zichner T., Stutz A. M., Korshunov A., Reimand J., Schumacher S. E. et al. Subgroup-specific structural variation across 1,000 medulloblastoma genomes. *Nature*, 488(7409):49–56, August 2012b.

Nowell P. The clonal evolution of tumor cell populations. *Science*, 194(4260):23–28, October 1976.

Nowell C. P. and Hungerford A. D. A Minute Chromosome in Human Chronic Granulocytic Leukemia. *Science*, 1960.

Oliver K. R. and Greene W. K. Mobile DNA and the TE-Thrust hypothesis: supporting evidence from the primates. *Mobile DNA*, 2(1):8, January 2011.

Onishi-Seebacher M. and Korbel J. O. Challenges in studying genomic structural variant formation mechanisms: the short-read dilemma and beyond. *BioEssays : news and reviews in molecular, cellular and developmental biology*, 33(11):840–50, November 2011.

Oostlander A. E., Meijer G. A., and Ylstra B. Microarray-based comparative genomic hybridization and its applications in human genetics. *Clinical Genetics*, 66(6):488–495, 2004.

Ostertag E. M., Goodier J. L., Zhang Y., and Jr H. H. K. SVA Elements Are Nonautonomous Retrotransposons that Cause Disease in Humans. *The American Journal of Human Genetics*, 73(6):1444–1451, 2003.

Ottaviani D., LeCain M., and Sheer D. The role of microhomology in genomic structural variation. *Trends in Genetics*, 30(3):85–94, January 2014.

Pace J. K. and Feschotte C. The evolutionary history of human DNA transposons: Evidence for intense activity in the primate lineage. *Genome Research*, 17(4):422–432, April 2007.

Pang A. W., MacDonald J. R., Pinto D., Wei J., Rafiq M. A., Conrad D. F., Park H., Hurles M. E., Lee C., Venter J. C. et al. Towards a comprehensive structural variation map of an individual human genome. *Genome biology*, 11(5):R52, January 2010.

Perry G. H., Dominy N. J., Claw K. G., Lee A. S., Fiegler H., Redon R., Werner J., Villanea F. A., Mountain J. L., Misra R. et al. Diet and the evolution of human amylase gene copy number variation, October 2007.

Picard F., Robin S., Lavielle M., Vaisse C., and Daudin J.-J. A statistical approach for array CGH data analysis. *BMC bioinformatics*, 6:27, January 2005.

Pickeral O. K. Frequent Human Genomic DNA Transduction Driven by LINE-1 Retrotransposition. *Genome Research*, 10(4):411–415, April 2000.

Pinkel D., Segraves R., Sudar D., Clark S., Poole I., Kowbel D., Collins C., Kuo W.-L., Chen C., Zhai Y. et al. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat Genet*, 20(2):207–211, October 1998.

Price A. L., Eskin E., and Pevzner P. A. Whole-genome analysis of Alu repeat elements reveals complex evolutionary history. *Genome Research*, 14(11):2245–2252, November 2004.

Prüfer K., Munch K., Hellmann I., Akagi K., Miller J. R., Walenz B., Koren S., Sutton G., Kodira C., Winer R. et al. The bonobo genome compared with the chimpanzee and human genomes. *Nature*, 486(7404):527–31, June 2012.

Quinlan A. R. and Hall I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, 26(6):841–2, March 2010.

Quinlan A. R., Clark R. a., Sokolova S., Leibowitz M. L., Zhang Y., Hurles M. E., Mell J. C., and Hall I. M. Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. *Genome research*, 20(5):623–35, May 2010.

Rausch T., Jones D., Zapatka M., Stütz A., Zichner T., Weischenfeldt J., Jäger N., Remke M., Shih D., Northcott P. et al. Genome Sequencing of Pediatric Medulloblastoma Links Catastrophic DNA Rearrangements with TP53 Mutations. *Cell*, 148(1-2):59–71, January 2012a.

Rausch T., Zichner T., Schlattl A., Stütz A. M., Benes V., and Korbel J. O. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics (Oxford, England)*, 28(18):i333–i339, September 2012b.

Rhesus Macaque Genome Sequencing and Analysis Consortium. Evolutionary and Biomedical Insights from the Rhesus Macaque Genome. *Science*, 316(5822):222–234, 2007.

Robinson J. T., Thorvaldsdottir H., Winckler W., Guttman M., Lander E. S., Getz G., and Mesirov J. P. Integrative genomics viewer. *Nat Biotechnol*, 29:24–26, 2011.

Sanger F. and Coulson A. R. A rapid method for determining sequences in {DNA} by primed synthesis with {DNA} polymerase. *Journal of Molecular Biology*, 94(3):441–448, 1975.

Sanger F., Nicklen S., and Coulson A. R. DNA sequencing with chain-terminating inhibitors, December 1977.

Sayah D. M., Sokolskaja E., Berthoux L., and Luban J. Cyclophilin A retrotransposition into TRIM5 explains owl monkey resistance to HIV-1. *Nature*, 430(6999):569–573, July 2004.

Schrider D. R., Navarro F. C. P., Galante P. a. F., Parmigiani R. B., Camargo A. a., Hahn M. W., and de Souza S. J. Gene copy-number polymorphism caused by retrotransposition in humans. *PLoS genetics*, 9(1):e1003242, January 2013.

Seleme M. D. C., Vetter M. R., Cordaux R., Bastone L., Batzer M. a., and Kazazian H. H. Extensive individual variation in L1 retrotransposition capability contributes to human genetic diversity. *Proceedings of the National Academy of Sciences of the United States of America*, 103(17):6611–6, April 2006.

Sen S. K., Han K., Wang J., Lee J., Wang H., Callinan P. A., Dyer M., Cordaux R., Liang P., and Batzer M. A. Human Genomic Deletions Mediated by Recombination between Alu Elements, July 2006.

Shaw C. J. and Lupski J. R. Non-recurrent 17p11.2 deletions are generated by homologous and non-homologous mechanisms. *Human genetics*, 116(1-2):1–7, January 2005.

Shen J. C., Rideout W. M., and Jones P. A. The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA., March 1994.

Shendure J. and Ji H. Next-generation DNA sequencing. *Nat Biotech*, 26(10):1135–1145, October 2008.

Simpson J. T., Wong K., Jackman S. D., Schein J. E., Jones S. J. M., and Birol I. ABySS: a parallel assembler for short read sequence data. *Genome research*, 19(6):1117–23, June 2009.

Singleton A. B. Exome sequencing: making hay while the sun shines, October 2011.

Smoll N. R. Relative survival of childhood and adult medulloblastomas and primitive neuroectodermal tumors (PNETs). *Cancer*, 118(5):1313–1322, 2012.

Snijders A. M., Nowak N., Segraves R., Blackwood S., Brown N., Conroy J., Hamilton G., Hindle A. K., Huey B., Kimura K. et al. Assembly of microarrays for genome-wide measurement of DNA copy number. *Nat Genet*, 29(3):263–264, November 2001.

Solyom S., Ewing A. D., Hancks D. C., Takeshima Y., Awano H., Matsuo M., and Kazazian H. H. Pathogenic orphan transduction created by a non-reference LINE-1 retrotransposon, February 2012a.

Solyom S., Ewing A. D., Rahrmann E. P., Doucet T. T., Nelson H. H., Burns M. B., Harris R. S., Sigmon D. F., Casella A., Erlanger B. et al. Extensive somatic L1 retrotransposition in colorectal tumors. *Genome research*, pages 2328–2338, October 2012b.

Stankiewicz P. and Lupski J. R. Genome architecture, rearrangements and genomic disorders. *Trends in genetics : TIG*, 18(2):74–82, February 2002.

Stephens P. J., Greenman C. D., Fu B., Yang F., Bignell G. R., Mudie L. J., Pleasance E. D., Lau K. W., Beare D., Stebbings L. a. et al. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell*, 144(1):27–40, January 2011.

Stewart C., Kural D., Strömberg M. P., Walker J. a., Konkel M. K., Stütz A. M., Urban A. E., Grubert F., Lam H. Y. K., Lee W.-P. et al. A Comprehensive Map of Mobile Element Insertion Polymorphisms in Humans. *PLoS Genetics*, 7(8):e1002236, August 2011.

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

Sudmant P. H., Kitzman J. O., Antonacci F., Alkan C., Malig M., Tsalenko A., Sampas N., Bruhn L., Shendure J., Project . G. et al. Diversity of Human Copy Number Variation and Multicopy Genes. *Science*, 330(6004):641–646, 2010.

Szak S. T., Pickeral O. K., Makalowski W., Boguski M. S., Landsman D., and Boeke J. D. Molecular archeology of L1 insertions in the human genome. *Genome biology*, 3(10):research0052, September 2002.

Szak S. T., Pickeral O. K., Landsman D., and Boeke J. D. Identifying related L1 retrotransposons by analyzing 3' transduced sequences. *Genome biology*, 4(5):R30, January 2003.

Talbot S. J. and Crawford D. H. Viruses and tumours – an update. *European Journal of Cancer*, 40(13):1998–2005, 2004.

Taylor M. D., Liu L., Raffel C., Hui C.-c., Mainprize T. G., Zhang X., Agatep R., Chiappa S., Gao L., Lowrance A. et al. Mutations in SUFU predispose to medulloblastoma. *Nat Genet*, 31(3):306–310, July 2002.

Taylor M., Northcott P., Korshunov A., Remke M., Cho Y.-J., Clifford S., Eberhart C., Parsons D., Rutkowski S., Gajjar A. et al. Molecular subgroups of medulloblastoma: the current consensus. *Acta Neuropathologica*, 123(4):465–472, 2012.

The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, November 2012.

The Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, 437(7055):69–87, September 2005.

The International HapMap. The International HapMap Project. *Nature*, 426(6968):789–96, December 2003.

The International SNP Map Working Group. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, 409(6822):928–933, February 2001.

Thorvaldsdóttir H., Robinson J. T., and Mesirov J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, 14(2):178–192, 2013.

Thung D. T., de Ligt J., Vissers L. E., Steehouwer M., Kroon M., de Vries P., Slagboom E. P., Ye K., Veltman J. a., and Hehir-Kwa J. Y. Mobster: accurate detection of mobile element insertions in next generation sequencing data. *Genome Biology*, 15(10):488, October 2014.

Treangen T. J. and Salzberg S. L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews. Genetics*, 13(1):36–46, November 2011.

Tubio J. M. C., Li Y., Ju Y. S., Martincorena I., Cooke S. L., Tojo M., Gundem G., Pipinikas C. P., Zamora J., Raine K. et al. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. *Science*, 345(6196):1251343–1251343, July 2014.

Tubio J. M. C. and Estivill X. Cancer: When catastrophe strikes a cell. *Nature*, 470(7335): 476–477, February 2011.

Untergasser A., Nijveen H., Rao X., Bisseling T., Geurts R., and Leunissen J. a. M. Primer3Plus, an enhanced web interface to Primer3. *Nucleic acids research*, 35(Web Server issue):W71–4, July 2007.

Varki A., Geschwind D. H., and Eichler E. E. Explaining human uniqueness: genome interactions with environment, behaviour and culture. *Nature reviews. Genetics*, 9(10):749–63, October 2008.

Varley J. M. Germline TP53 mutations and Li-Fraumeni syndrome. *Human Mutation*, 21(3): 313–320, 2003.

Walker J. a., Konkel M. K., Ullmer B., Monceaux C. P., Ryder O. a., Hubley R., Smit A. F., and Batzer M. a. Orangutan Alu quiescence reveals possible source element: support for ancient backseat drivers. *Mobile DNA*, 3:8, January 2012.

Wang H., Xing J., Grover D., Hedges D. J., Han K., Walker J. a., and Batzer M. a. SVA elements: a hominid-specific retroposon family. *Journal of molecular biology*, 354(4):994–1007, December 2005.

Wang Z., Gerstein M., and Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics*, 10(1):57–63, January 2009.

Waszak S. M., Hasin Y., Zichner T., Olender T., Keydar I., Khen M., Stütz A. M., Schlattl A., Lancet D., and Korbel J. O. Systematic inference of copy-number genotypes from personal genome sequencing data reveals extensive olfactory receptor gene content diversity. *PLoS computational biology*, 6(11):e1000988, January 2010.

Watanabe Y., Fujiyama A., Ichiba Y., Hattori M., Yada T., Sakaki Y., and Ikemura T. Chromosome-wide assessment of replication timing for human chromosomes 11q and 21q: disease-related genes in timing-switch regions. *Human Molecular Genetics*, 11(1):13–21, 2002.

Watson J. D. and Crick F. H. C. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*, 171:737–738, April 1953.

Weddington N., Stuy A., Hiratani I., Ryba T., Yokochi T., and Gilbert D. M. ReplicationDomain: a visualization tool and comparative database for genome-wide replication timing data. *BMC bioinformatics*, 9:530, 2008.

Wei W., Gilbert N., Ooi S. L., Lawler J. F., Ostertag E. M., Kazazian H. H., Boeke J. D., and Moran J. V. Human L1 Retrotransposition: cis Preference versus trans Complementation. *Molecular and Cellular Biology*, 21(4):1429–1439, February 2001.

Weischenfeldt J., Symmons O., Spitz F., and Korbel J. O. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nature Reviews Genetics*, 14(2): 125–138, January 2013.

Weiss M. M., Hermsen M. A., Meijer G. A., van Grieken N. C., Baak J. P., Kuipers E. J., and van Diest P. J. Comparative genomic hybridisation., October 1999.

Wu J., Lee W.-P., Ward A., Walker J. a., Konkel M. K., Batzer M. a., and Marth G. T. Tangram: a comprehensive toolbox for mobile element insertion detection. *BMC genomics*, 15(1):795, January 2014.

Xi R., Luquette J., Hadjipanayis A., Kim T.-M., and Park P. J. BIC-seq: a fast algorithm for detection of copy number alterations based on high-throughput sequencing data. *Genome Biology*, 11(Suppl 1):O10, 2010.

Xing J., Wang H., Belancio V. P., Cordaux R., Deininger P. L., and Batzer M. a. Emergence of primate genes by retrotransposon-mediated sequence transduction. *Proceedings of the National Academy of Sciences of the United States of America*, 103(47):17608–13, November 2006.

Yan G., Zhang G., Fang X., Zhang Y., Li C., Ling F., Cooper D. N., Li Q., Li Y., van Gool A. J. et al. Genome sequencing and comparison of two nonhuman primate animal models, the cynomolgus and Chinese rhesus macaques. *Nat Biotech*, 29(11):1019–1023, November 2011.

Yang H., Zhong Y., Peng C., Chen J.-Q., and Tian D. Important role of indels in somatic mutations of human cancer genes. *BMC medical genetics*, 11:128, January 2010.

Ye K., Schulz M. H., Long Q., Apweiler R., and Ning Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics (Oxford, England)*, 25(21):2865–71, November 2009.

Zhang F., Khajavi M., Connolly A. M., Towne C. F., Batish S. D., and Lupski J. R. The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. *Nature genetics*, 41(7):849–53, July 2009.

Zhang Z. and Gerstein M. Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes. *Nucleic Acids Research*, 31(18):5338–5348, September 2003.

Zhao M., Wang Q., Wang Q., Jia P., and Zhao Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC bioinformatics*, 14 Suppl 1(Suppl 11):S1, January 2013.