Abstract

Computational analysis on genomic variation: Detecting and characterizing structural variants in the human genome

Hugo Y.K. Lam

2010

Genomic variation refers to the difference of DNA sequence between two or more individuals. In the past, it was believed that most human sequence variation was attributable to single nucleotide polymorphisms (SNPs), which was estimated to occur every 300–1,000 bases on average when comparing two different chromosomes. Nowadays, with the advance of sequencing technology, we are able to reveal a large number of different variation called structural variation (SV). This kind of variation includes genomic rearrangement such as deletion, insertion and inversion, which are usually defined as >1 kbp in size. These SVs have considerable impact on genomic variation by causing more nucleotide differences between individuals than SNPs and by creating gene duplication or deletion. Even though many recent findings have implicated the importance of SVs such as disease association, the understanding of their formation processes and the ability to identify them are still very limited, which have particularly hampered further studies on a large scale. To this end, this thesis aims to carry out a detailed and large-scale computational analysis on genomic variation. It demonstrates a loss-of-function variation analysis across different eukaryotic genomes by using a database of pseudogene families and an ontology, which reveals the formational bias of pseudogene and its relation with other genomic segments such as segmental duplications (SDs). It goes on to investigate the formation mechanisms of SVs by correlating SDs and copy number variants (CNVs) with genomic repeats such as the Alu elements. Then, it extends the characterization of SVs by using an SV breakpoint library and reveals their formational biases. Finally, it introduces a novel computational approach for reliably and efficiently identifying SVs in a newly-sequenced personal genome.

Computational analysis on genomic variation: Detecting and characterizing structural variants in the human genome

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
in Candidacy for the Degree of
Doctor of Philosophy

by
Hugo Y.K. Lam

Dissertation Director: Mark B. Gerstein

December, 2010

Contents

List of Figures

List of Tables

Acknowledgments

This thesis would not have been possible without the support and encouragement from a number of people.

First and foremost I owe my deepest gratitude to my advisor, Prof. Mark B. Gerstein, who has given me enormous insights and directions on my research and various chances to work on different projects.

My gratitude is also due to the members of the thesis committee, Prof. Hongyu Zhao and Assoc. Prof. Kei-Hoi Cheung, for their advice and support on my work.

Besides my advisor and committee members, I am heartily thankful to Asst. Prof. Philip M. Kim and Dr. Jan O. Korbel for their advice and collaborations.

I would also like to extend my thank to my fellow colleagues, especially Joel Rozowsky, Nitin Bhardwaj, Jasmine Mu, Alexej Abyzov, Jiang Du, Zhengdong Zhang, Ekta Khurana, Zhi Lu, Chong Shou, Tara Gianoulis and everyone in the Gerstein Lab.

Last but not least, I am very grateful to have the love and endless support from my parents and my beloved wife, Carinna, throughout the years.

Chapter 1

Introduction

Genome sequencing, which determines the order of nucleotides that make up the genomes of living organisms, has been a key tool in modern biological research. Due to the enormous cost and time to sequence a whole genome in the past [1], early attempt to discover genomic variants that set the phenotypic difference between individuals was only focused on the differences of single nucleotides, known as single nucleotide polymorphisms (SNPs). Recent advances in sequencing technology now enable scientists to decode the genome more rapidly and efficiently [2], and have paved the way for identifying large-block genomic variants, called structural variants or SVs.

SVs are usually defined as genomic variation ranging from kilobase to megabase, including insertions, deletions and inversions [3] (Figure 1.1). These variants cause more nucleotide differences between individuals than SNPs. Some SVs, in fact, involve thousands of base pairs and can remove whole genes or create additional ones that can have major effects on an organism [4]. For example, SVs are sometimes associated with diseases such as HIV [5] and also with developmental disorders such as Down syndrome [6]. They are a result of sequence rearrangement in the genome caused by various mechanisms, such as non-allelic homologous recombination (NAHR; involving homology-mediated recombination between paralogous sequence blocks), non-homologous end-joining (NHEJ; associated with the repair of DNA double-strand breaks) and transposable element insertion (TEI;

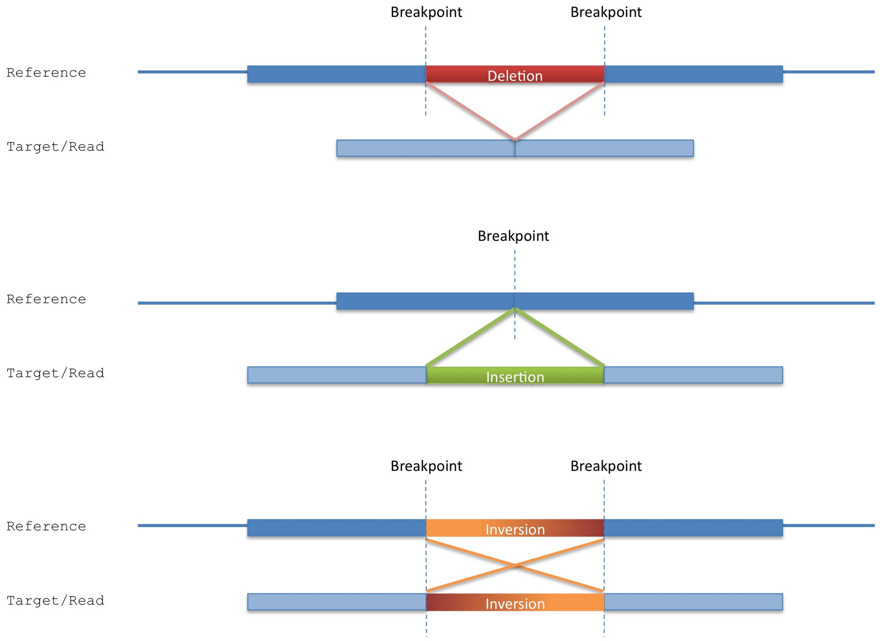


Figure 1.1: Major SV events defined as deletion (upper), insertion (middle) and inversion (lower) relative to the reference genome.

involving mostly long and short interspersed elements). When SVs exhibit difference in copy number between two or more individuals, they are referred as copy number variation (CNV). Due to the biological significance and importance of SV implicated in recent findings, it has become one of the focuses in genetics lately.

To detect CNVs or SVs, comparative genomic hybridization (CGH) and paired-end sequencing are the major techniques being used to date [7]. It has been previously demonstrated that using high-resolution CGH based on tiling array can precisely map copy number in mammalian cells [8]. Various computational methods have also been developed to identify the variants from the array signals. For example, a non-parametric method using a mean-shift-based (MSB) approach has been proposed to detect the changed copy numbers in array-CGH data by determining local modes of the signal [9]. To compensate for the approximate CNV coordinates from CGH experiments, a computational approach called BreakPtr has been proposed to fine-map CNVs. It suggested a predictive resolution ($\sim$ 300 bp) that could enable more precise correlations between CNVs and across individuals

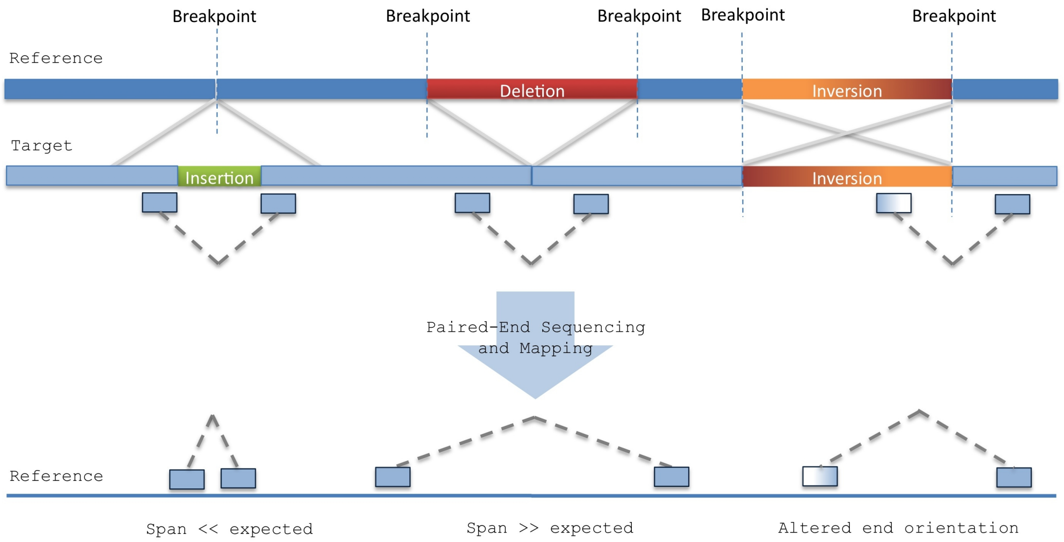


Figure 1.2: SV detection method: paired-end sequencing and mapping.

than previously possible [10].

Different from CGH, which can only detect copy number variants, paired-end sequencing (Figure 1.2) is able to discover copy number balanced events (e.g., inversions). In recent years, a high-throughput and massive sequencing method, called paired-end mapping (PEM), was developed to identify structural variants (down to almost 3kb in size) between genomes using the 454 paired-end sequencing technology [11]. To facilitate the SV detection from massive paired-end sequences, a cross-platform computational framework, PEMer, was then developed for identifying structural variants with sequence reads from different sequencing platforms. The analysis pipeline aims to map SVs at high-resolution by paired-end sequences and has showed improvement on sensitivity and specificity over previous approaches [12].

Since SVs identified by methods such as paired-end mapping may not precisely represent the breakpoints of the events, a split-read analysis approach [13] is commonly used to deduce the exact breakpoint locations of the SVs with reads from second-generation sequencing or from PCR sequences spanning the paired ends. This approach basically

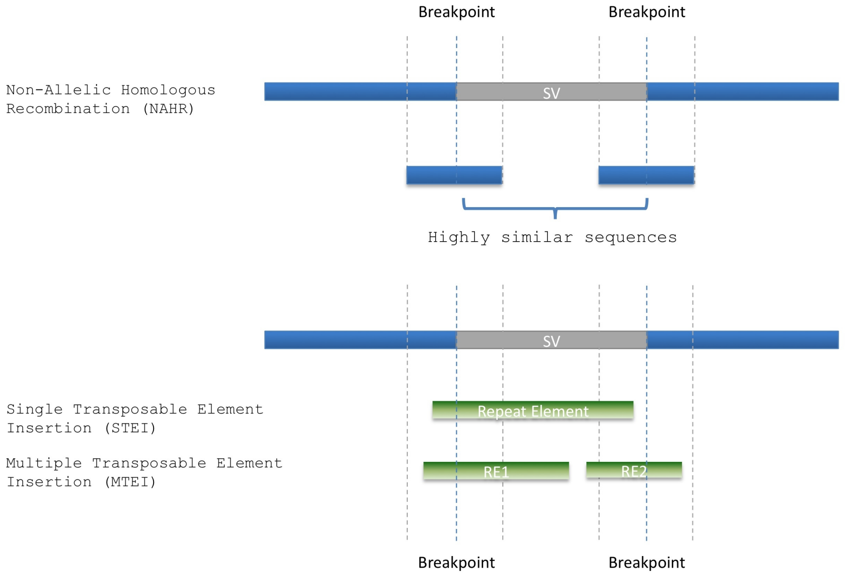

Breakpoint
Breakpoint
Non-Allelic Homologous Recombination (NAHR)
SV
Highly similar sequences
SV
Single Transposable Element Insertion (STEI)
Repeat Element
Multiple Transposable Element Insertion (MTEI)
RE1
RE2
Breakpoint
Breakpoint

Figure 1.3: SV breakpoint junction analysis: examples for classifying SVs as Non-Allelic Homologous Recombination (NAHR) and Transposable Element Insertion (TEI).

aligns a read or sequence to a reference genome and identifies those having one end of the read mapped to one location on a chromosome and the other to a concordant location that represents either a deletion or an insertion event (Figure 1.1). While we have SV calls with breakpoints, we could carry out nucleotide-level analysis on SVs such as deducing their formation mechanisms (Figure 1.3).

This thesis is focused on computational analysis on genomic variation, particularly in humans, including SV, CNV and pseudogene. It involves using high throughput sequencing technologies and computational algorithms to systematically and efficiently identify and characterize SVs. It aims to enhance the understanding of the mechanism and impact of genomic variation and to facilitate such analyses.

1.0.1 Loss-of-function variation

In chapter two [14], we present a database of pseudogene families, Pseudofam, based on the protein families from the Pfam database. It provides resources for analyzing the family structure of pseudogenes including query tools, statistical summaries and sequence alignments. Pseudogenes are loss-of-function gene relics resulted from duplication or retrotranposition events. Like their functional counterparts, they experience variation among different individuals as well as different species. The current version of Pseudofam contains more than 125,000 pseudogenes identified from 10 eukaryotic genomes and aligned within nearly 3000 families (approximately one-third of the total families in PfamA). Pseudofam uses a large-scale parallelized homology search algorithm (implemented as an extension of the PseudoPipe pipeline) to identify pseudogenes. Each identified pseudogene is assigned to its parent protein family and subsequently aligned to each other by transferring the parent domain alignments from the Pfam family. Pseudogenes are also given additional annotation based on an ontology, reflecting their mode of creation and subsequent history. In particular, our annotation highlights the association of pseudogene families with genomic features, such as segmental duplications. In addition, pseudogene families are associated with key statistics, which identify outlier families with an unusual degree of pseudogenization. The statistics also show how the number of genes and pseudogenes in families correlates across different species. Overall, they highlight the fact that housekeeping families tend to be enriched with a large number of pseudogenes.

1.0.2 Segmental duplication and copy number variation

In chapter three [15], we investigate the formation of SDs (operationally defined as >1 kb stretches of duplicated DNA with high sequence identity) and CNVs by examining their large-scale patterns of co-occurrence with different repeats. Alu elements, a major class of genomic repeats, had previously been identified as prime drivers of SD formation. We also observe this association; however, we find that it sharply decreases for younger SDs. Continuing this trend, we find only weak associations of CNVs with Alus. Similarly, we find an association of SDs with processed pseudogenes, which is decreasing for younger SDs and absent entirely for CNVs. Next, we find that SDs are significantly co-localized with each other, resulting in a highly skewed 'power-law' distribution and chromosomal hotspots. We also observe a significant association of CNVs with SDs, but find that an SD-mediated mechanism only accounts for some CNVs ($< 28\%$). Overall, our results imply that a shift in predominant formation mechanism occurred in recent history: ∼40 million years ago, during the 'Alu burst' in retrotransposition activity, non-allelic homologous recombination, first mediated by Alus and then by newly formed CNVs themselves, was the main driver of genome rearrangements; however, its relative importance has decreased markedly since then, with proportionally more events now stemming from other repeats and from non-homologous end-joining. In addition to a coarse-grained analysis, we performed targeted sequencing of 67 CNVs and then analyzed a combined set of 270 CNVs (540 breakpoints) to verify our conclusions.

1.0.3 Structural variation with breakpoints

In chapter four [16], we introduce a new computational approach for identifying SVs and further investigate their formational biases. SVs are a major source of human genomic variation; however, characterizing them at nucleotide resolution remains challenging. Here we assemble a library of breakpoints at nucleotide resolution from collating and standardizing ∼2,000 published SVs. For each breakpoint, we infer its ancestral state (through comparison to primate genomes) and its mechanism of formation (e.g., nonallelic homologous recombination, NAHR). We characterize breakpoint sequences with respect to genomic landmarks, chromosomal location, sequence motifs and physical properties, finding that the occurrence of insertions and deletions is more balanced than previously reported and that NAHR-formed breakpoints are associated with relatively rigid, stable DNA helices. Finally, we demonstrate an approach, BreakSeq, for scanning the reads from short-read sequenced genomes against our breakpoint library to accurately identify previously overlooked SVs, which we then validate by PCR. As new data become available, we expect our BreakSeq approach will become more sensitive and facilitate rapid SV genotyping of personal genomes.

In chapter five, we conclude the thesis with possible future directions, followed by the appendix [17] which introduces a methodolody for predicting protein domain binding sites.

Chapter 2

Analyzing loss-of-function variations in eukaryotic genomes using a family approach

2.1 Introduction

The complexity of the eukaryotic genome is characterized by its large amount of non-protein-coding DNA. This type of DNA typically lies in intergenic regions and was regarded as 'junk' DNA in the past. However, due to the recent advancement of genomic technology, it has been found that intergenic DNA indeed plays an important role in regulatory function and also provides a basis for studying the dynamics and evolution of a genome [18].

Among all the intergenic elements, from transcription factor binding sites to microsatellites, pseudogenes, which are in effect genetic fossils, are the elements most likely to record historical aspects of living genes. Pseudogenes not only capture genes in the past, but also provide precious clues about genome dynamics, such as gene duplication events (for duplicated pseudogenes) and retrotransposition events (for processed pseudogenes). Since proteins in the same family are believed to share a common ancestor giving rise to the shared domain, association of pseudogenes with their parent protein families could reveal the correlation between the generation of pseudogenes and the functions of their parents. This correlation otherwise might not be observable from the study of individual pseudogenes.

A number of experimental and computational approaches have been developed to identify and annotate pseudogenes in eukaryotic genomes [19, 20, 21]. Also, there are a few prior studies that have attempted to analyze pseudogenes using protein families [22, 23]. However, no study thus far has systematically formalized the pseudogene relationships and presented an integrated analysis of several eukaryotes using a family approach. To this end, we aim to develop a large-scale database of pseudogene families of eukaryotes, Pseudofam, that could enable researchers to analyze pseudogenes and relate them to existing genomic information in an integrated fashion.

2.2 A database of pseudogene families

Pseudofam is implemented as an online database, which is available at `http://pseudofam.pseudogene.org`. The web site itself is a thin-client application implemented using Java on the server side and requires only a web browser on the client side. It provides tools for researchers to browse and query the pseudogene families. Moreover, it provides certain useful statistics (described in detail below), such as the enrichment of parent proteins for each family and the correlation of different family parameters between species. The database is also capable of interfacing with other related systems, such as the Ensembl server and the Pfam database. Furthermore, researchers can download the family data sets, including the alignment of the sequences, in flat file formats.

2.3 Assigning pseudogenes to families

Figure 2.1 depicts an overview of the generation of Pseudofam data from the identification of pseudogenes to the formation of the families. DNA sequences of 10 eukaryotic genomes: human, chimpanzee, dog, mouse, rat, fruit fly, mosquito, chicken, zebra fish and worm,

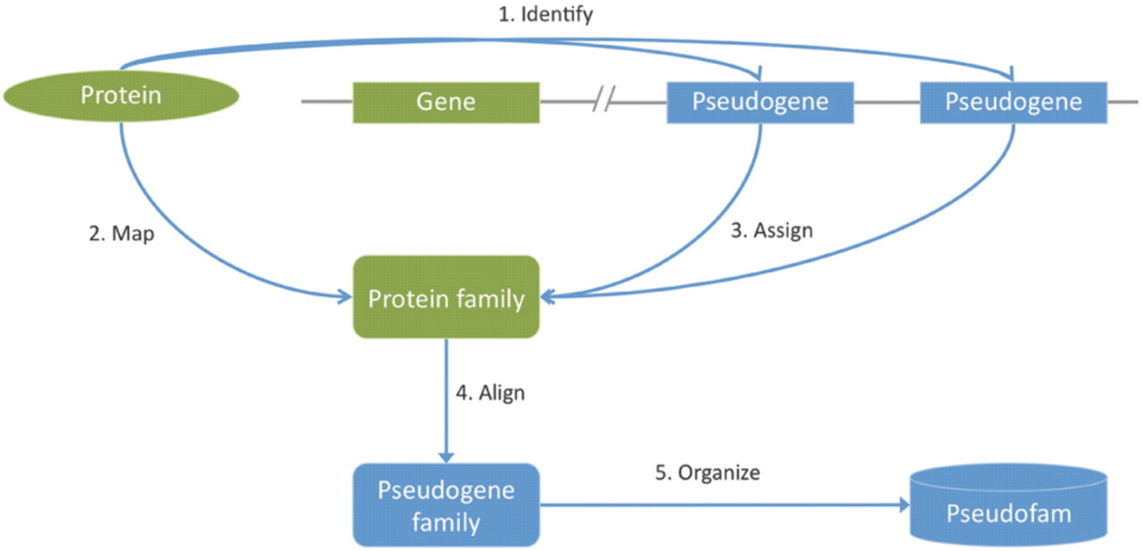


Figure 2.1: The generation of Pseudofam. (1) Identify pseudogenes by existing proteins of the genome. (2) Map all the parent proteins to their protein families. (3) Assign the identified pseudogenes to their parent protein families. (4) Align the pseudogenes in each family to build the pseudogene families. (5) Calculate the key statistics for the families and organize the data into the Pseudofam database.

together with their over 291 000 protein sequences were retrieved from Ensembl (`http://www.ensembl.org`; release 48—December 2007) [24, 25]. Each genome and its associated protein sequences are run through BLAST [26, 27] to identify all genomic regions that share sequence similarity with the given protein sequences. The proteins are divided into groups of queries, which are processed concurrently to reduce overall runtime, while the genomes are used as the BLAST database. The results are then processed using PseudoPipe [21] to identify potential pseudogenes. This analysis pipeline uses tFasty [28] to refine the BLAST alignments and determine frame shifts and other disablements. It takes about 3 days of computational time to complete the identification of pseudogenes in the human genome with our current configuration.

Our current release of Pseudofam contains 3,821 protein families covering all the protein sequences used as input for identifying the 125,272 pseudogenes. The parent proteins of the identified pseudogenes belong to 2,986 pseudogene families. Thus, there are 835 protein families not found to have any pseudogenes. Families for the protein sequences are constructed by mapping the Ensembl peptide IDs to the Pfam ID via mappings available at the BioMart server (`http://www.biomart.org/`; Ensembl Release 48) [29]. Pseudogenes are assigned to the protein families based on the assignments of their parent proteins and then aligned to identify any pseudogene domains by the mechanics described below.

Figure 2.2 shows a schematic representation of our approach in aligning pseudogene domains by transferring their parent domain alignments from the Pfam multiple alignments. Within each family, a pseudogene is first aligned to its parent protein and then to its corresponding protein domain retrieved from the Pfam database (`http://pfam.sanger.ac.uk`; version 22) [30, 31]. After the individual alignments, all the pseudogene domains from distinct species are then aligned together with their parent protein domains. This approach of alignment enables us to accurately align pseudogenes with low levels of similarity and consequently to identify pseudogene domains that might exhibit low similarity to their parent protein domains. The resulting pseudogene domain alignment data provide researchers a means to estimate the mutation rate of genomic elements that evolve under no or less selection pressure [32]. This alignment data is available for download.

2.4 Describing pseudogene families using ontologies

With the family data available at Pseudofam, we can extend our family approach to other potentially related analyses. Since pseudogenes are nonprotein coding and have no direct functions, their relationships with other parts of the genome are often neglected and poorly understood. However, more and more findings have demonstrated pseudogenes, as a gene relic, not only facilitate evolutionary study, but also exhibit substantial interactions in the genome. They have been shown to play different roles in the genome remodeling process, including retrotransposition, duplication and mutation. Recent studies, for example, have shown that some of the pseudogenes may have mediated the formation of segmental duplications (SDs) [15], regulating their parent genes through RNA interference [33], or even been reactivated [34, 35]. As a result, we have developed an ontology (a formal specifica-

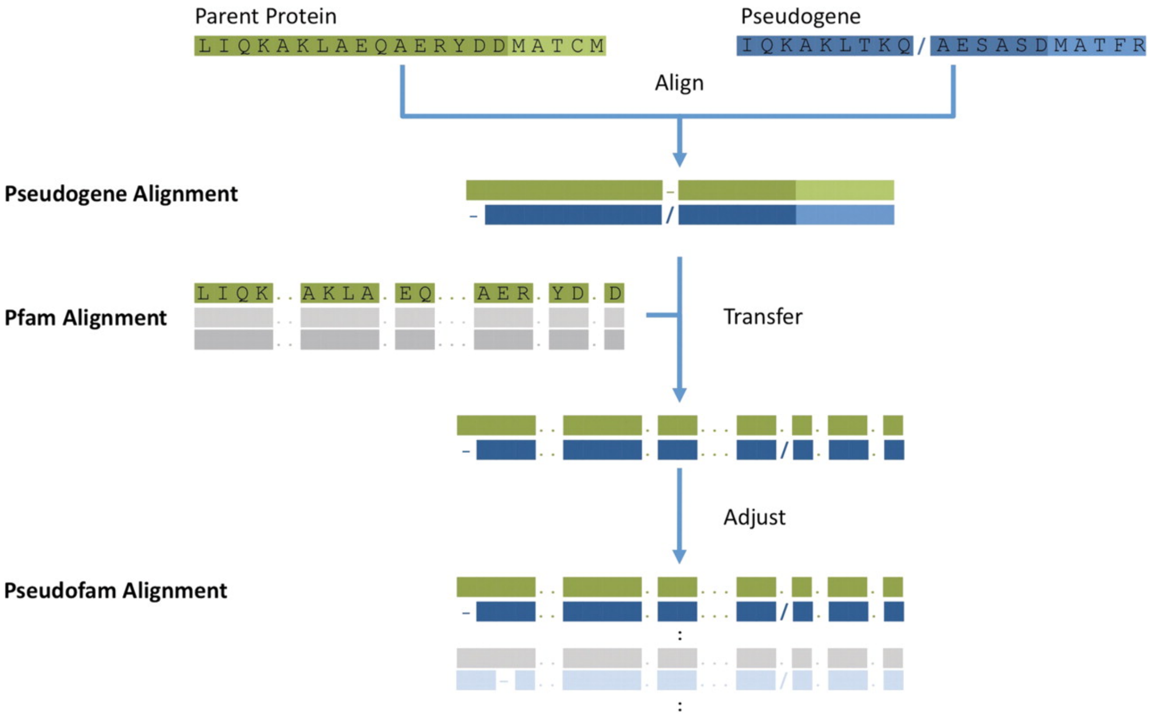


Figure 2.2: The alignment of pseudogene family. Each pseudogene in a family is first aligned to its parent protein. Then, the pseudogene alignment is aligned with the parent protein domain by transferring the corresponding alignment from the Pfam multiple alignments. At last, all the aligned pseudogene domains, including their aligned parent protein domains, will be adjusted together to generate the final alignment.

tion of conceptualization [36]) to illustrate pseudogene family relationships. To facilitate further analysis, we have also formatted our ontology into the Open Biomedical Ontology (OBO) format and annotated our data accordingly.

2.4.1 An upper ontology

The ontology in Figure 2.3 shows an upper ontology depicting the pseudogene family and its relationships. It spans across several domains and involves different domain-specific ontologies, such as the Gene Ontology (GO), Protein Ontology (PO/PrO), Sequence Ontology (SO) and Pseudogene Ontology (Figure 2.4). It basically consists of three parts. The first (in blue) is the core part and family concept that Pseudofam is built upon. The second (in dark gray) is a part that describes certain primary aspects of pseudogenes that are fairly well established, such as their genomic processes of creation (e.g. retrotransposition and duplication). The third (in light gray) is a part that describes the secondary aspects of a pseudogene family (e.g. its association with SDs), as well as terms that are currently in a draft state. These draft terms include unitary (describing pseudogenes mutated directly from a parent gene), orphaned (for pseudogenes whose parent genes were lost after speciation) and transcribed (for apparently active pseudogenes). While the upper ontology is essentially finished, the full Pseudogene Ontology is still being developed in collaboration with a number of other individuals.

2.4.2 Family relationships

Based on the fundamental relationship between protein family and pseudogene, our ontology also depicts the structural and functional relationships tying to a pseudogene family. These relationships could aid in further understanding of various genomic processes. For example, the co-localization of pseudogenes in a shared synteny could indicate their formation before speciation [35, 37], and the presence of pseudogenes in SDs could provide clues about SD formation since both pseudogenes and SDs represent duplicated regions of the genome [38]. Thus, Pseudofam currently provides the human pseudogene dataset

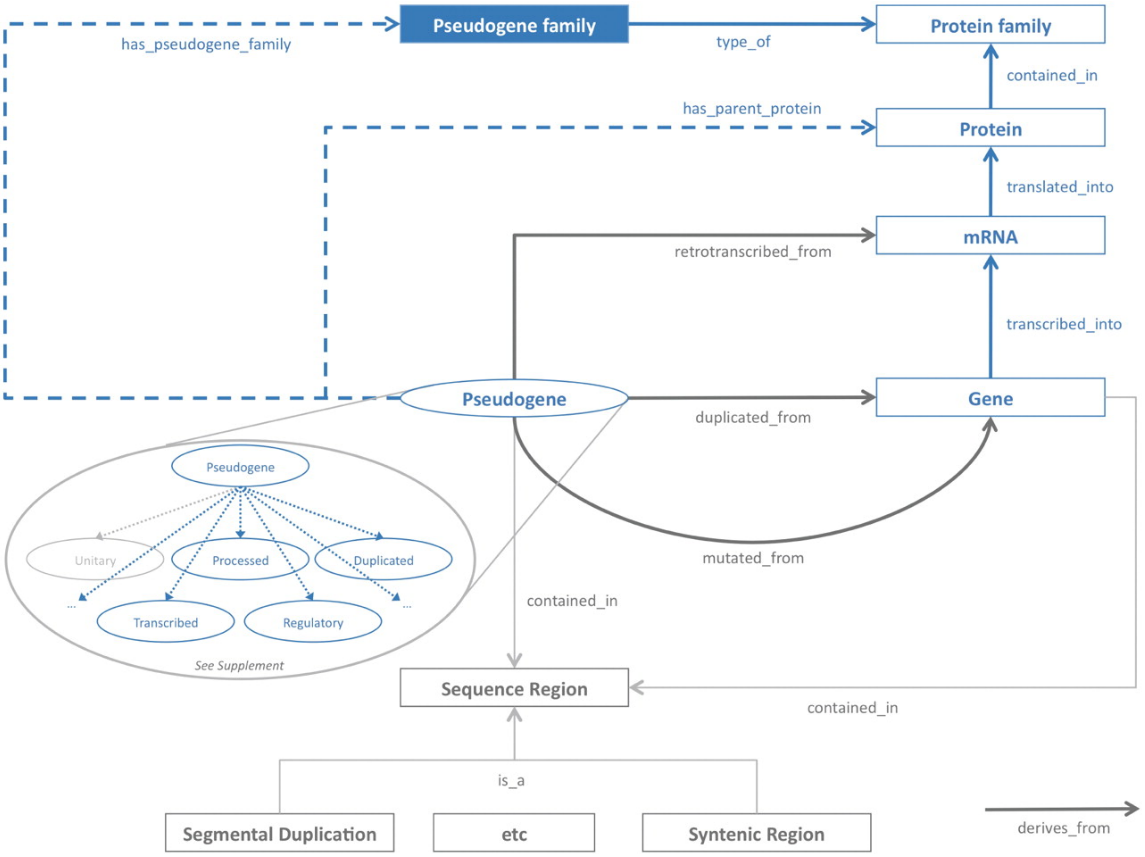


Figure 2.3: The pseudogene family ontology. An upper ontology that describes the various relationships between a pseudogene family and other genomic elements. The solid lines represent direct relationships and the dashed lines represent inferred or indirect relationships. The core part is represented in blue, while the well-established relationships are in dark gray and the secondary aspects of a pseudogene family are in light gray.

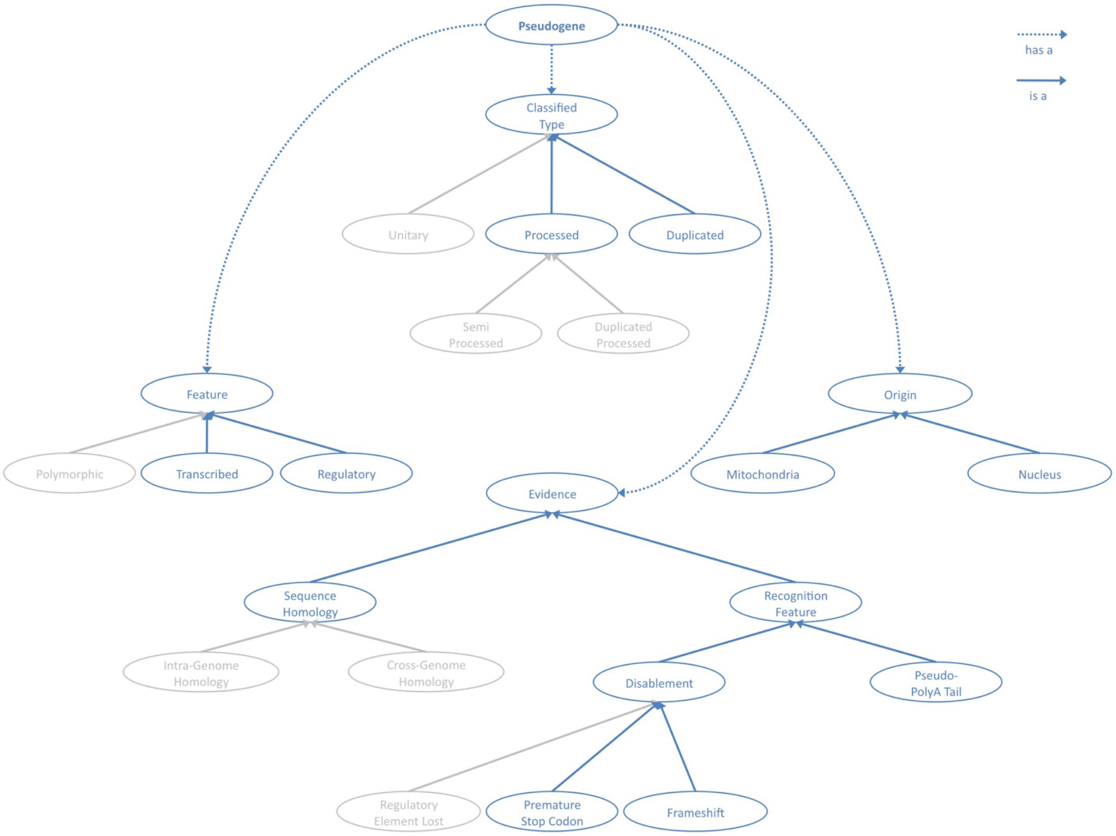


Figure 2.4: A pseudogene ontology in draft. Solid lines represent an 'is-a' relationship and dashed lines represent a 'has-a' relationship. Blue presents solid concepts and light grey represents concepts in draft.

annotated with SD information obtained from the Human Segmental Duplication database at `http://eichlerlab.gs.washington.edu/database.html`. While the SD relationship derives directly from the pseudogenes themselves, the family relationship of a pseudogene is inferred by the protein family relationship of its parent protein and hence is more indirect. Here, we formalize this inferred relationship in a first-order logic on which Pseudofam is built:

$$\forall_p (Pseudogene(p) \wedge$$

$$\exists_r (Protein(r) \wedge hasParentProtein(p, r) \wedge \exists_f (ProteinFamily(f) \wedge contains(f, r)$$

$$\rightarrow hasPseudogeneFamily(p, f))))$$

In words, for all pseudogene p, if there exists a protein r, which is a parent protein of p, and there also exists a protein family f, which contains r, then p has a pseudogene family f. Even though a pseudogene is nonprotein coding, this protein family approach of classification gives us a way to associate domain and function with it. Proteins in the same family are believed to share a common structural domain and function that evolved from a common ancestor. As a result, a family approach allows us to analyze pseudogenes by their functional groups and have a better understanding of their roles in genome rearrangement by relating them to other genomic features.

2.5 Statistical analysis on pseudogene families

To further facilitate analysis with our family data, Pseudofam provides key statistics, such as the degree of pseudogenization and pseudogene-to-gene ratio, for each family both online and in the datasets for download. It also provides a tool to correlate different family parameters between species. To identify outlier families that have an unusual degree of pseudogenization, Pseudofam calculates the enrichment of parent proteins in each family and uses the hypergeometric distribution to calculate P-value, viz:

$$Pr(K = k) = f(k; N, m, n) = \frac{\begin{pmatrix} m \\ k \end{pmatrix} \begin{pmatrix} N - m \\ n - k \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}}$$

This formula calculates the probability $Pr(K)$ of having the observed number of parent proteins k for a given family with n proteins under the hypergeometric distribution. Required for the computation is the total number of proteins N used for identifying the pseudogenes and the corresponding number of parent proteins m. The P-value for a positive enrichment is the $Pr(K \geq k)$ and for a negative enrichment is the $Pr(K \leq k)$. This parent protein approach is preferred over using a random sampling method to calculate the enrichment of pseudogenes because it is more computationally efficient and less susceptible to the changes of the pseudogenes identification algorithm or parameters that may cause the number of pseudogenes identified to fluctuate. The following sections show a brief analysis based on the key statistics provided by Pseudofam.

2.5.1 Degree of pseudogenization

Table 2.1 shows the numbers of protein and pseudogene families in different species and their degree of pseudogenization. It indicates that among the species in our study mammals have a higher percentage (an average of 50%) of families containing pseudogenes than nonmammals (an average of 22%). For instance, human has 3,486 protein families of which 1,790 (51%) are found to have pseudogenes. On the other hand, Drosophila has 2,620 protein families but only 201 (8%) are found to have pseudogenes. Looking at the families individually shows that certain families have a high degree of pseudogenization, while some have no pseudogenes at all. For example, the reverse transcriptase (RNA-dependent DNA polymerase) family has 18 out of 22 (82%) proteins found to have associated pseudogenes. In contrast, the bestrophin protein family, which has 71 proteins, has not been found to have any pseudogenes.

	Protein Family	Pseudogene Family	Pseudogenized
Homo sapiens (HS)	3,486	1,790	51.35%
Pan troglodytes (PT)	3,443	1,906	55.36%
Canis familiaris (CF)	3,151	1,529	48.52%
Mus musculus (MM)	3,461	1,654	47.79%
Rattus norvegicus (RN)	3,138	1,489	47.45%
Anopheles gambiae (AG)	2,715	570	20.99%
Gallus gallus (GG)	2,911	860	29.54%
Drosophila melanogaster (DM)	2,620	201	7.67%
Danio rerio (DR)	3,145	1,125	35.77%
Caenorhabditis elegans (CE)	2,633	360	13.67%
Total	3,821	2,986	78.15%

Table 2.1: Numbers of protein and pseudogene families in different species out of 9,318 PfamA families. The number of protein families represents the total number of families that each has at least one protein in the species. The number of pseudogene families is a subset of the previous number representing the total number of protein families with at least one pseudogene.

2.5.2 Correlation of family sizes across species

Since the mammalian genomes have a substantial number of pseudogene families, they enable us to carry out a more accurate statistical analysis of the correlation of genes and pseudogenes. Table 2.2 shows the Spearman correlation of the family size between the five mammalian genomes in our study. It shows that protein family size has an obviously stronger correlation (0.81) among species than pseudogene family size (0.63). It also shows that the correlation of pseudogene family size decreases when the evolutionary distance increases between the species. For example, human has a correlation of 0.89 with chimpanzee, but only around 0.58 with dog, mouse and rat. Similarly, mouse has a correlation of 0.67 with rat, but only around 0.58 with human, chimpanzee and dog. It supports the theory that pseudogenes in general are evolving under no or less selection pressure relative to functional genes.

2.5.3 Extreme families

The enrichment results (Table 2.3) show that families with housekeeping proteins, such as the GAPDH protein (a NAD-binding enzyme involved in glycolysis and glyconeogenesis),

	HS	PT	CF	MM	RN
HS	1.00	0.92	0.77	0.84	0.75
PT	0.89	1.00	0.79	0.84	0.77
CF	0.60	0.62	1.00	0.78	0.85
MM	0.58	0.60	0.57	1.00	0.80
RN	0.57	0.59	0.59	0.67	1.00

Table 2.2: Spearman's rank correlation of protein family sizes (the upper right) and pseudogene family sizes (the lower left) between different species.

and the ribosomal protein RPL7A (responsible in mRNA-directed protein synthesis in all organisms) [31] have significantly more parent proteins than others. In order to investigate whether proteins having housekeeping functions tend to have more pseudogenes than those with nonhousekeeping functions, we downloaded a total of 575 human housekeeping genes derived from gene expression profiling [39, 40]. We selected all the 197 pseudogene families that contain both the housekeeping and nonhousekeeping genes, and tested the pseudogene-to-gene ratio between these two types of genes using a Wilcoxon signed rank test. We found that the pseudogene-to-gene ratio for housekeeping genes is significantly higher (P-value $<$ 0.04) than for nonhousekeeping genes in such pseudogene families, especially in processed pseudogenes (P-value $<$ 0.01). It has also been reported previously by Gonclaves et al. [41] that housekeeping genes generally have more processed pseudogenes. This could be explained by the relatively constant expression level of housekeeping genes, which boosts their chances of being retrotranscribed.

2.5.4 Correlation with segmental duplications

With the tools, statistics and ontology provided by Pseudofam, we can analyze pseudogenes from a different perspective and integrate pseudogene families with other related datasets to better understand the genome remodeling processes. For example, both pseudogenes and SDs represent duplicated regions of the genome; hence, by analyzing the presence of pseudogenes located in SDs, some precious clues about the generation processes of pseudogene and SD formation can be obtained [42]. It was reported recently by Zheng [38] that

Pfam Acc	Pfam ID	Proteins	Parents	Enrichment	P-value
PF00001	7tm_1	8,548	1,446	1.39	1.21E-39
PF01157	Ribosomal_L21e	113	72	5.25	6.50E-38
PF00044	Gp_dh_N	229	105	3.78	1.69E-36
PF03402	V1R	214	109	3.44	2.95E-35
PF02800	Gp_dh_C	228	103	3.72	3.88E-35
PF07686	V-set	5,355	943	1.45	4.87E-32
PF01248	Ribosomal_L7Ae	237	101	3.51	7.15E-32
PF01352	KRAB	2,127	500	1.67	8.04E-32
PF03953	Tubulin_C	248	95	3.15	7.64E-26
PF07735	FBA_2	159	50	5.98	8.75E-26
PF00091	Tubulin	271	98	2.98	2.18E-24
PF03939	Ribosomal_L23eN	71	44	5.10	4.90E-23
PF00276	Ribosomal_L23	102	52	4.20	1.65E-21
PF00333	Ribosomal_S5	82	46	4.62	1.82E-21
PF00046	Homeobox	2,551	168	0.54	5.91E-21
PF00018	SH3_1	2,960	207	0.58	1.16E-20
PF00237	Ribosomal_L22	59	36	5.02	8.63E-19
PF00076	RRM_1	3,196	556	1.43	2.30E-18
PF03719	Ribosomal_S5_C	91	45	4.07	3.79E-18
PF01391	Collagen	1,289	66	0.42	5.44E-18

Table 2.3: The top 20 protein families (sorted by p-value in ascending order), which have an unusual degree of pseudogenization (p-value < 0.05).

in humans, SDs are more enriched with pseudogenes than genes, with 36.8% pseudogenes located in SDs and 17.8% genes located in SDs. Since genomic duplications have a destabilizing effect [42], it makes sense that the SDs are more enriched with pseudogenes than with genes, because structural variations in pseudogenes have less impact than in genes. This trend also reflects in the correlations of pseudogenes and parent genes of pseudogene families within SDs for the human genome (Figure 2.5), where there is a stronger positive Spearman correlation (0.69) between the numbers of duplicated pseudogenes in pseudogene families and those located in SDs, than that of parent genes (0.41).

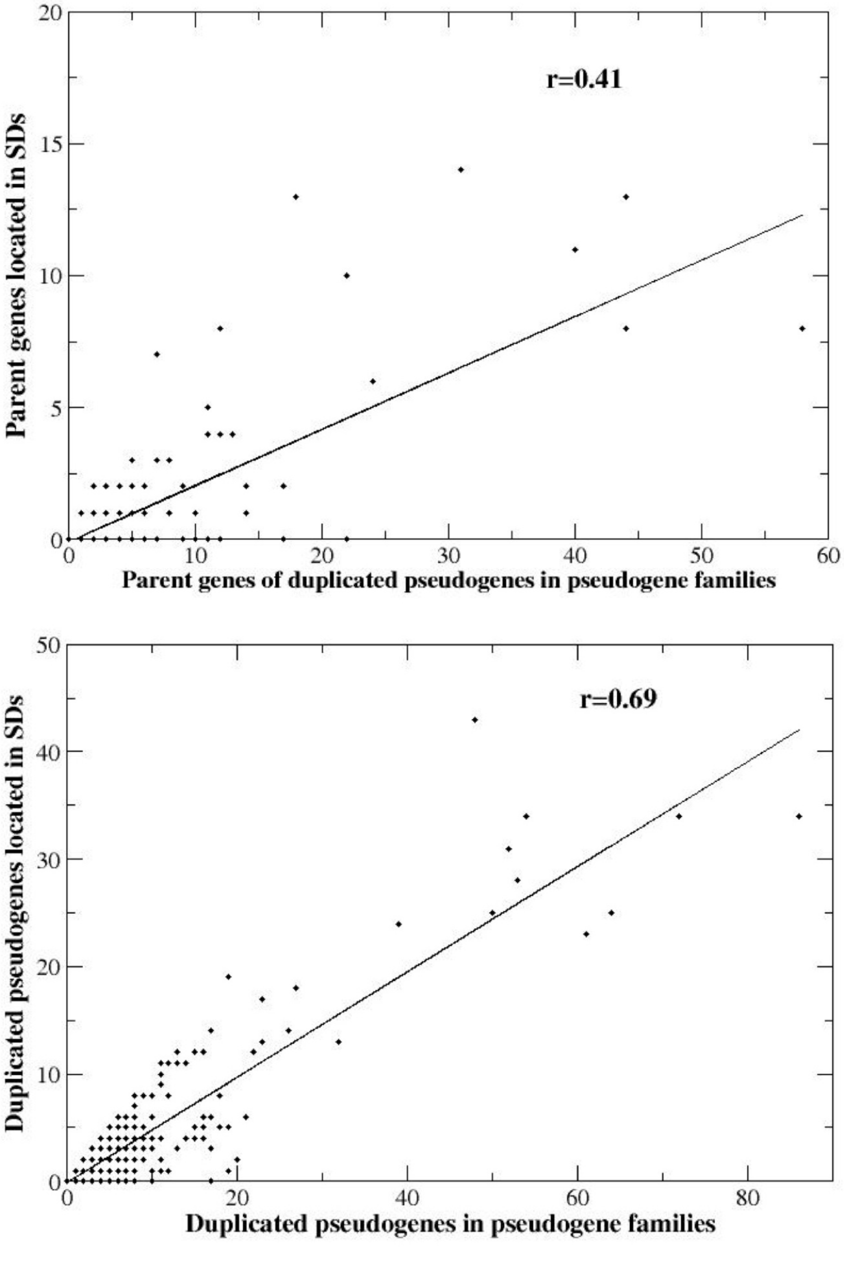


Figure 2.5: Parent genes (upper) and duplicated pseudogenes (lower) within each pseudogene family vs. those located in SDs. r stands for Spearman's rank correlation. p-values for all the correlations are below 2.2 e-16.

Chapter 3

Analysis of copy number variants and segmental duplications in the human genome

3.1 Introduction

With the rapid advances in high-throughput technology, the study of human genome variation is emerging as a major research area. A large fraction of variation in terms of single nucleotide polymorphisms (SNPs) ('point variation') has been mapped and genotyped (The International HapMap Consortium 2005). However, it has recently been recognized that a major fraction of mammalian genetic variation is manifested in an entirely different phenomenon known as 'copy number variation'. In contrast to SNPs, these variations correspond to relatively large (>1 kb according to a widely accepted operational definition) regions in the genome that are either deleted or amplified on certain chromosomes ('block variation') [43, 44, 45, 46, 47, 11]. They are known as 'copy number variants' (CNVs) and are estimated to cover $\sim$ 12% of the human genome, thereby accounting for a major portion of human genetic variation [47, 48]. Some CNVs reach fixation in the population and (if they correspond to duplications) are then visible in the genome as fixed Segmental

Duplications (SDs) [42]. A sizeable fraction (estimated to be 5.2%) of the human genome is covered in either fixed or polymorphic SDs [49, 42]. These are defined as duplicated genomic regions of >1 kb with 90% or greater sequence identity among the duplicates. They are especially widespread in the primate lineage [50]. SDs enclosing entire genes contribute to the expansion of protein families [51]. Some of these duplicated genes may fall out of use, thereby giving rise to pseudogenes. Some duplications that are annotated as SDs may not be fixed in the population, but rather correspond to common CNVs, in particular, common ones that are present in the human reference genome. Current efforts to sequence individual human genomes, such as the 1000 Genomes Project (1000genomes.org), will bring greater certainty about which SDs are fixed and which are polymorphic, more precisely viewed as CNVs.

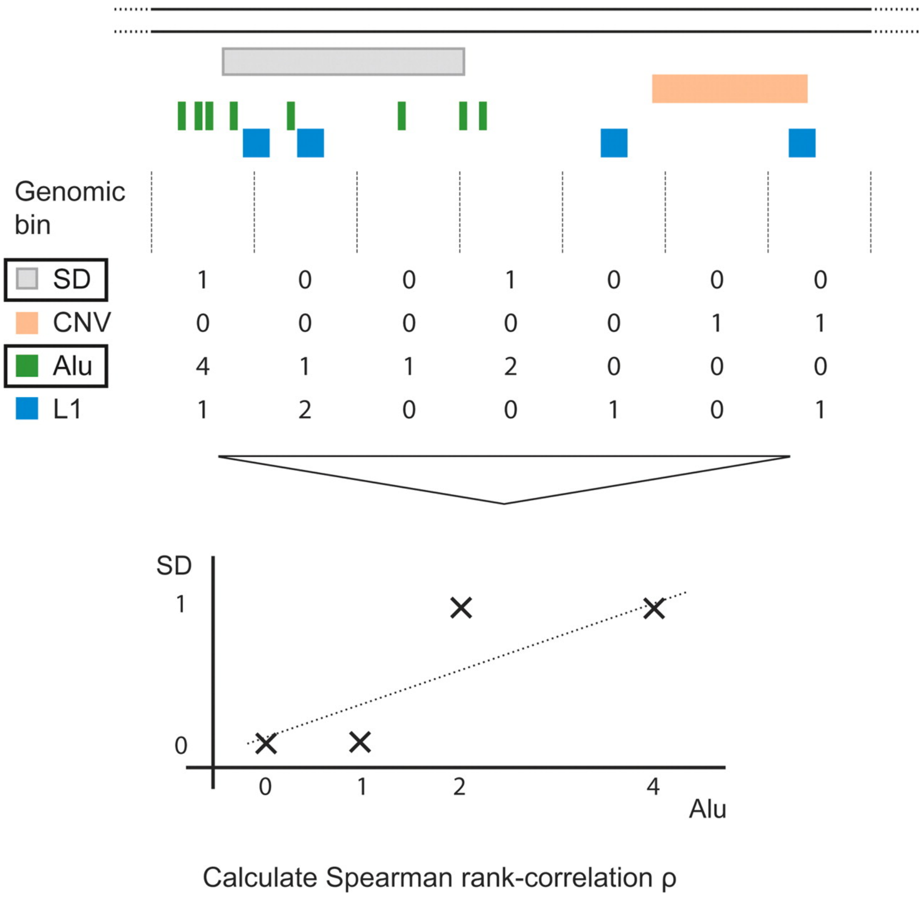
Hitherto, not much was known about mechanisms of CNV formation, but it has been suggested that non-allelic homologous recombination (NAHR) during meiosis can lead to the formation of larger deletions and duplications (or to structural variants such as inversions). In general, recombination mechanisms such as NAHR are mediated by preexisting repeats. Alu elements have been previously implicated in the formation of SDs [52, 53], which is consistent with NAHR-based formation. Likewise, SDs have been suggested as mediating CNV formation [46, 54, 55]. However, not all duplications are thought to arise because of NAHR-based mechanisms: In subtelomeres, a separate mechanism, non-homologous end-joining (NHEJ), has been suggested for SD formation [56, 57]. Furthermore, recent studies have uncovered a mechanism that combines both homologous and non-homologous recombination [58, 59]. Finally, a novel mechanism that involves fork stalling and template switching during replication has been proposed [60].

In this study, we examine formation signatures of both SDs and CNVs in an integrated fashion. Specifically, we first survey genomic features in the human and their occurrence. Among the features that we survey are SD and CNV boundaries as well as common repeat elements, such as Alu and LINE retrotransposons and microsatellites. To assess colocalization of the different features, we follow a two-pronged approach: First, we bin all the features into small sequence bins of 100 kb and examine the associations by computing Spearman (rank) correlation coefficients between two features (e.g., Alu elements and CNV breakpoints) as sketched out in Figure 3.1. This coarse-grained approach is necessary to avoid problems with the comparatively low resolution of current large-scale CNV data (at best 50 kb) [61]. We use the Spearman correlation as a more robust measure to detect nonlinear relationships. A high (statistically significant) correlation implies strong colocalization. We interpret statistical enrichment of colocalized elements as an indicator that these elements might be involved in the formation of SDs or CNVs, respectively. Second, to provide further evidence that the colocalization trends found above are due to actual differences in formation mechanisms, we examined actual breakpoints. Thus far, not many sequences of CNV breakpoints are available. Hence, we performed targeted sequencing of breakpoints, and we analyzed them in combination with a large number of previously sequenced ones. To calculate enrichment of specific features around the breakpoints, we compare the number of intersecting features to randomized global and local regions of the genome. Our results show different signatures of formation for SDs and CNVs. While for SDs (especially older ones), we find a striking enrichment of Alu elements and other repeats in the breakpoint regions, suggesting Alu-mediated formation, we find little evidence for such a mechanism in CNVs. Here, we present evidence for several alternative features that may contribute to the formation of both SDs and CNVs.

3.2 Results

3.2.1 Segmental duplications follow a power-law pattern in the human genome

We believe that SDs should be the result of CNVs reaching fixation. Also, it has been suggested that CNV formation is partly mediated by SDs [46, 62, 54]. Taken together, this would imply that SD formation would preferentially occur in regions with many previously existing SDs. That is, an SD-rich region would generate more CNVs than other regions,



Genomic bin							
▢ SD	1	0	0	1	0	0	0
🟧 CNV	0	0	0	0	0	1	1
🟩 Alu	4	1	1	2	0	0	0
🟦 L1	1	2	0	0	1	0	1

Figure 3.1: Schematic representation of the overall analysis methodology. For the coarse-grained analysis, genomic features are surveyed. First, the number of features in each genomic bin is counted. Then the overall pairwise correlation is measured (using Spearman rank correlation or Wilcoxon rank-sum tests).

some of which, in turn, become fixed as SDs. This phenomenon represents one form of a preferential attachment mechanism ('the rich get richer'). This mechanism has been well studied in the physics literature, and it is known that it generally leads to a power-law distribution in terms of the regions [63]. Note, however, that while a preferential attachment mechanism does generally lead to a power-law distribution, the inverse is not necessarily the case. A power-law or scale-free distribution corresponds to a distribution with a very long tail [64]. For our case, this would mean that there should be an extreme imbalance in the distribution of SDs, that is, a few regions in the genome would be very rich in SDs, while most would contain no or very few SDs. Intuitively, the phenomenon of preferential attachment led to an enrichment of SDs in regions already rich in SDs and resulting in a highly skewed distribution. Hence, if SD-mediated NAHR is a major factor contributing to new SDs, we would expect the density of SDs to be distributed according to a power law throughout the human genome. Indeed, when analyzing different regions in the human genome for ends of SDs harbored, we observe a distinct power law (Figure 3.2). This power-law behavior is consistent with the existence of rearrangement 'hot spots' [65]. This result, taken together with the aforementioned theoretical notions, supports the hypothesis that SD formation is mediated by pre-existing SDs. The power-law distribution is independent of SD size, age, or the binning procedure (Methods).

3.2.2 Segmental duplications co-occur best with other segmental duplications of similar age

Furthermore, an SD-mediated NAHR mechanism would imply that recent SDs should co-occur with older segmental duplications. Hence, if we bin SDs according to sequence similarity between the duplicates (viewing sequence similarity between the duplicates as approximate age since they diverge after duplication), we should see a significant co-occurrence between different bins. Indeed, we observe a significant correlation between SDs in different age groups (sequence identity) (Figure 3.3). Strikingly, we observe that the best co-occurrence for the SDs of any given age bin is with the SDs in the 'neighboring'

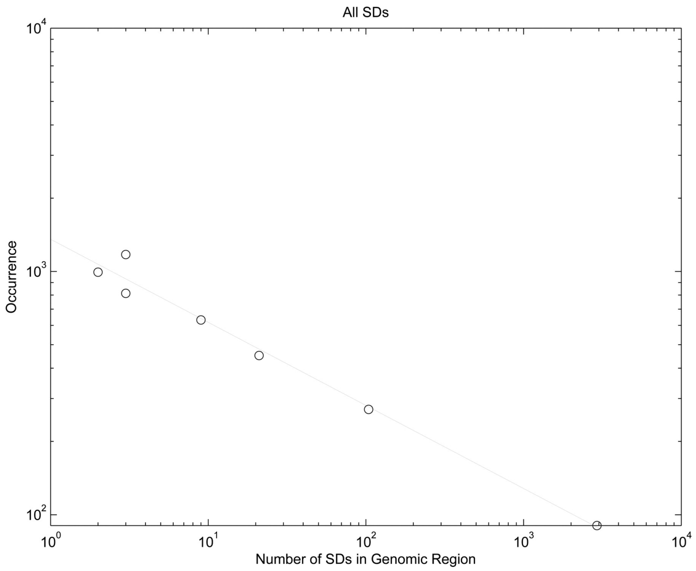


Figure 3.2: Segmental duplications are distributed according to a power law in the human genome. As can be seen, segmental duplications follow a power-law distribution, that is, while most regions in the genome are relatively poor in SDs, there are a small number of regions with much higher SD occurrence ($p(x) \sim x^{-0.31}$). This is indicative of a preferential attachment ('rich get richer') mechanism.

bin (i.e., the bin slightly older), consistent with an SD-mediated NAHR. Note that this result would also be consistent with different regions being susceptible to chromosomal rearrangements at different times. However, without a preferential attachment mechanism, we are very unlikely to observe a power-law distribution as in Figure 3.2. Finally, we observe that this correlation is best for old SDs and gets successively worse as we move toward more and more recent SDs. This may be indicative of a trend of changing SD formation behavior, as we discuss below.

3.2.3 Alu-mediated NAHR is an additional mechanism to preferential attachment

As another mechanism for SD formation, NAHR mediated by Alu retrotransposons has been proposed [52]. Note that Alu repeats are the most common repeat element in the human genome with about a million copies. We set out to examine this mechanism and find that SDs show highly significant colocalization with Alu elements (Figure 3.4B), consistent

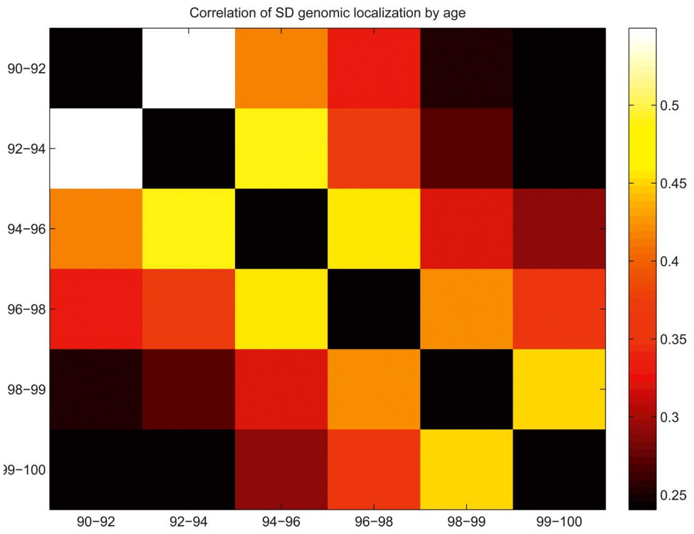


Figure 3.3: Heatmap of associations of SDs in different sequence identity bins. SDs co-occur best with pre-existing SDs of similar age, and this trend appears to be stronger for older SDs. Associations are given as Spearman rank correlations of the number of occurrences in genomic bins. All correlations are highly significant (P-value << 0.00001).

with earlier reports [53, 42]. This trend is decreasing rapidly for younger SDs (Figure 3.4B), while the oldest (most divergent) SDs associate most strongly with Alus. In line with this result, we find that most SDs have a sequence identity similar to Alu elements (90%) (Figure 3.5). The abundance of both retrotransposed elements and SDs then decreases with rising sequence identity, in sync. SDs also appear to colocalize with LINE/L1 repeats, but this association is much weaker and might be reflective of colocalization of Alus and L1 repeats [66]. We also find evidence that Alu-mediated mechanisms and preferential attachment mechanisms may be complementary. That is, SDs that colocalize strongly with Alus show weaker correlation with pre-existing SDs (Figure 3.4A) than those that appear in Alu-poor regions. This result holds true for SDs of any sequence identity bin. It suggests that a certain group of SDs is likely to have been formed by an Alu-mediated mechanism, and another disjoint group is a more likely candidate for a mechanism involving pre-existing SDs.

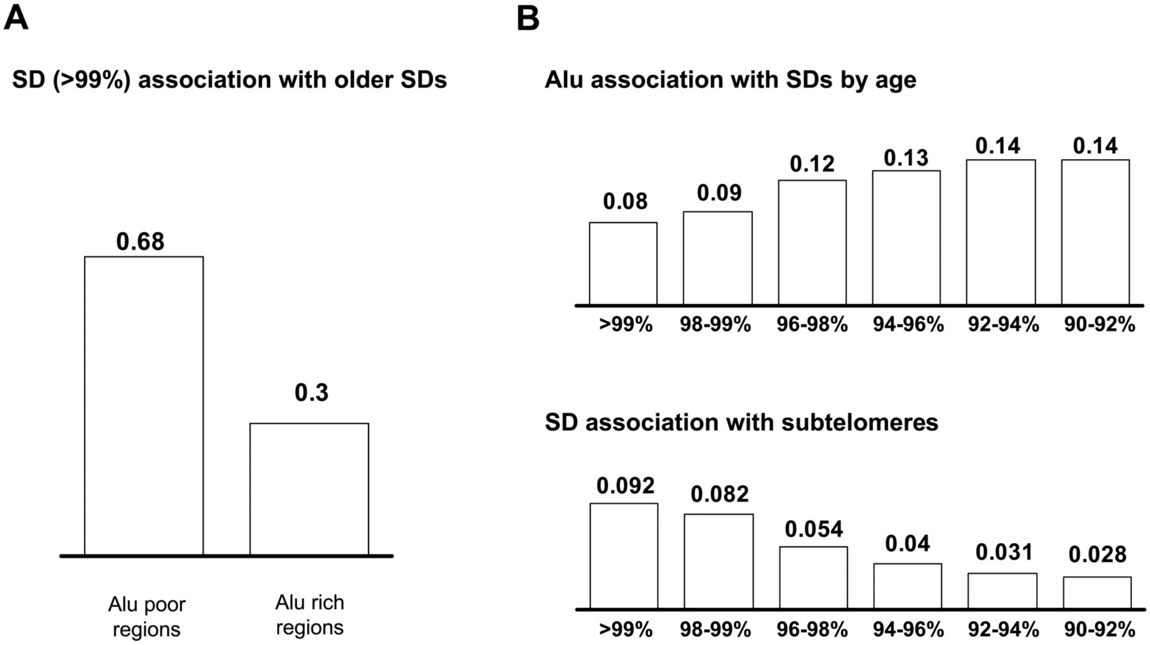


Figure 3.4: (A) Alu-mediated NAHR and preferential attachment are two complementary mechanisms for SD formation. In Alu-rich regions (>10 Alu elements per 10 kb), the association of SDs and pre-existing SDs is much lower than in Alu-poor regions (no Alu elements per 100 kb). Associations are given as Spearman rank correlations of the number of occurrences in genomic bins. All correlations are highly significant (P-value << 0.00001). (B) Association of Alu elements and SDs is highest for the oldest (∼40 Mya) SDs and drops significantly for recent SDs. At the same time, preference for subtelomeric regions and a presumed NHEJ mechanism rises. Associations are given as Spearman rank correlations of the number of occurrences in genomic bins. All correlations are highly significant (P-value << 0.00001).

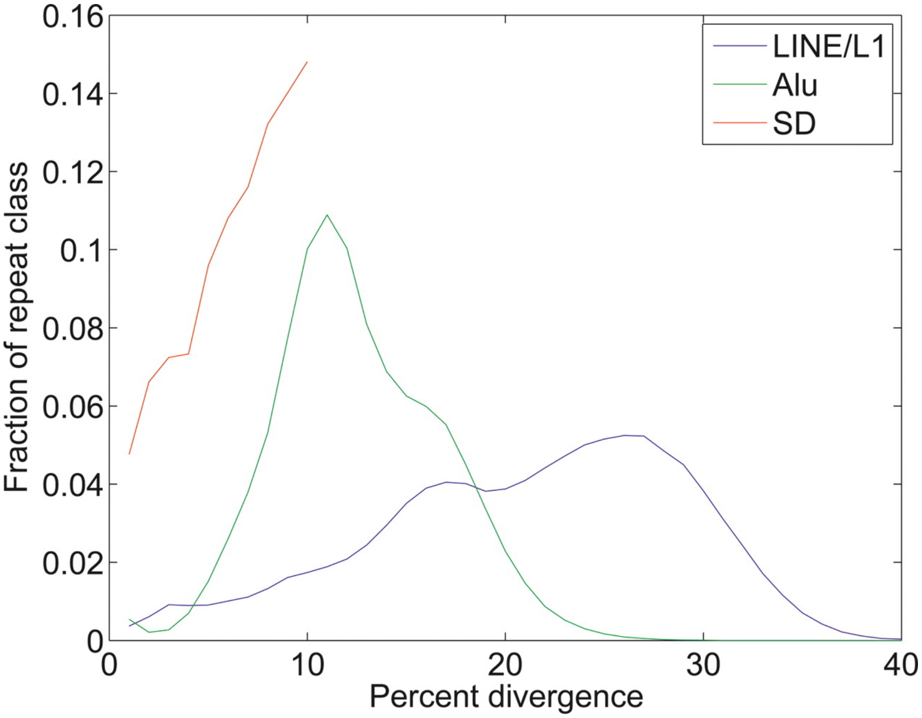


Figure 3.5: Sequence divergence of repeat elements in the human genome. As approximate age, the sequence divergence shows a burst of Alu activity ∼40 Mya and a marked decrease afterward. The distribution of (active) LINE elements is somewhat more even. The relative number of SDs decreases in a fashion similar to the Alu elements.

3.2.4 Processed pseudogenes show significant association with SDs

Processed pseudogenes were formed in a way similar to Alu retro transposons, that is, they parasitize the same LINE retrotransposition machinery and are also thought to have been mostly formed during the Alu burst ∼40 million years ago (Mya) [67]. The obvious difference is that there are a much greater variety of pseudogenes than Alu elements. Therefore, it is less likely for any given processed pseudogene to find a nearby matching partner to recombine with, which is a prerequisite for genome rearrangement via homologous recombination. Despite this, we find a strong enrichment of processed pseudogenes with SDs (Figure 3.6). To evaluate whether these pseudogenes actually contributed to the formation of SDs, we performed a detailed breakpoint analysis of SDs. For a number of cases (144), we find matching processed pseudogenes at the matching SD junction regions of duplicated regions. In an additional 78 cases, we find processed pseudogenes at both SD junctions that have different parent genes, but are highly similar ($>$ 95% sequence identity) over stretches of at least 200 bp. Note that many pseudogenes have different parents but still show high sequence identity. While these numbers are highly significant (P-values $\leq$ 0.001), they are relatively small compared to the total number of processed pseudogenes in the human genome (9747; www.pseudogene.org). One reason may be that the recombination process requires the pairing of two separate and matching pseudogenes. Since there are far fewer matching pseudogenes than Alu elements, this may have led to the formation of much fewer SDs. These results suggest that pseudogenes did contribute to SD formation, albeit only in a small number of cases.

3.2.5 Copy number variants co-occur with segmental duplications

It has been noted previously that CNVs co-occur with SDs, and SD-mediated NAHR has been suggested as a possible mechanism of CNV formation [46, 68, 69, 54]. In line with this, CNVs have been viewed as the drifting, polymorphic form of SDs. This view implies that CNVs should follow a similar pattern of distribution as very young SDs (i.e., SDs of

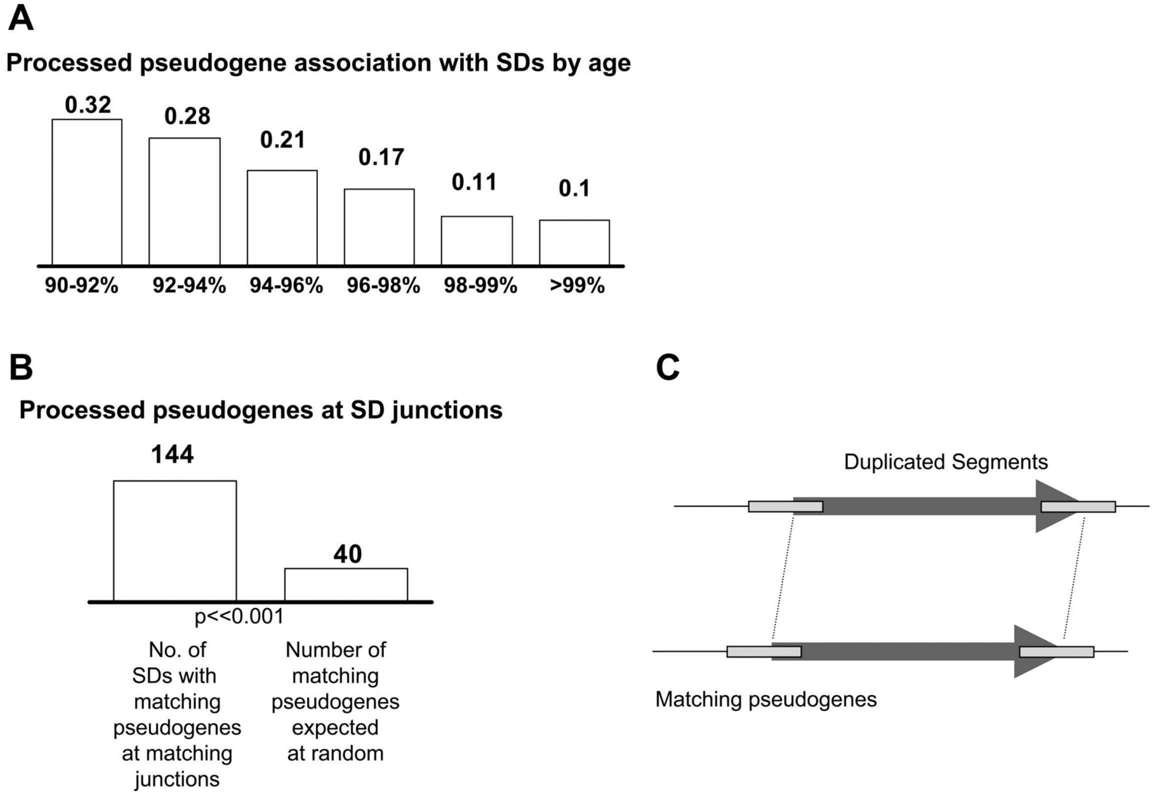


Figure 3.6: (A) Pseudogene association with SDs. Just like Alu elements, pseudogenes colocalize very strongly with old SDs and less so with younger SDs. All correlations are highly significant (P-value >> 0.00001). (B) Detailed SD junction analysis. A total of 144 SDs showed matching processed pseudogenes at both junctions, that is, both pseudogenes have the same parent gene and show high homology. When picking random genomic regions of the same size and number as SDs, no matching pseudogenes were ever found to overlap both SD junctions. When using a randomized offset of ± 5 kb to account for potential sequence biases, an average of 40 matching pseudogenes were found, but in 1000 trials, never more than 43. (C) Schematic of matching processed pseudogenes at SD junctions. The processed pseudogenes overlap matching SD junctions at both duplicated segments, making them likely candidates for having mediated NAHR.

very high sequence similarity), since they would have been created by similar mechanisms. When analyzing SD and CNV distributions in the genome, we indeed find that there is a significant overlap (Figure 3.7A). However, the correlation between SD and CNV occurrence is smaller than may be expected. We find that maximally 28% of CNVs were formed by an SD-mediated mechanism, that is, lie in a region with a nearby SD. This is an upper bound estimate, since proximity does not imply causality. From another perspective, one may (perhaps naively) expect that the similarity in distribution of CNVs and SDs of $> 99\%$ sequence identity should be comparable to the similarity between the distributions of SDs of $> 99\%$ sequence identity and SDs of $98\% - 99\%$ identity. However, we find that the correlation for CNVs and young SDs (rank correlation of 0.14) is lower than the one for 'very old' ($90\% - 92\%$ sequence identity) and 'very young' ($> 99\%$ sequence identity) SDs (rank correlation of 0.24). In other words, $\sim 60\%$ of 'very young' SDs could be the result of NAHR mediated by older SDs. Conversely, the same can be said of only 28% of CNVs. This may be consistent with the fact that CNVs are polymorphic SDs.

3.2.6 Copy number variants do not show any significant association with Alu elements

If CNVs and SDs are formed by similar processes, one might assume that CNVs would also show association with Alu elements. However, we find that CNVs show no significant association with Alu elements (Figure 3.7B). Previous studies found weak associations of CNVs with Alu elements [55], but they are much weaker than the ones found for SDs (of any sequence identity bracket). Indeed, when controlling for SD content, the association becomes even weaker.

This result implies that an Alu-mediated mechanism is an unlikely candidate for CNV formation. It is consistent with reports that Alu-mediated NAHR was most common during or shortly after the burst of Alu activity ~40 Mya and has since declined [70, 71]. Hence, the formation of CNVs and some SDs is probably mediated by different phenomena. One might argue that some of this difference is due to the different methods of experimental

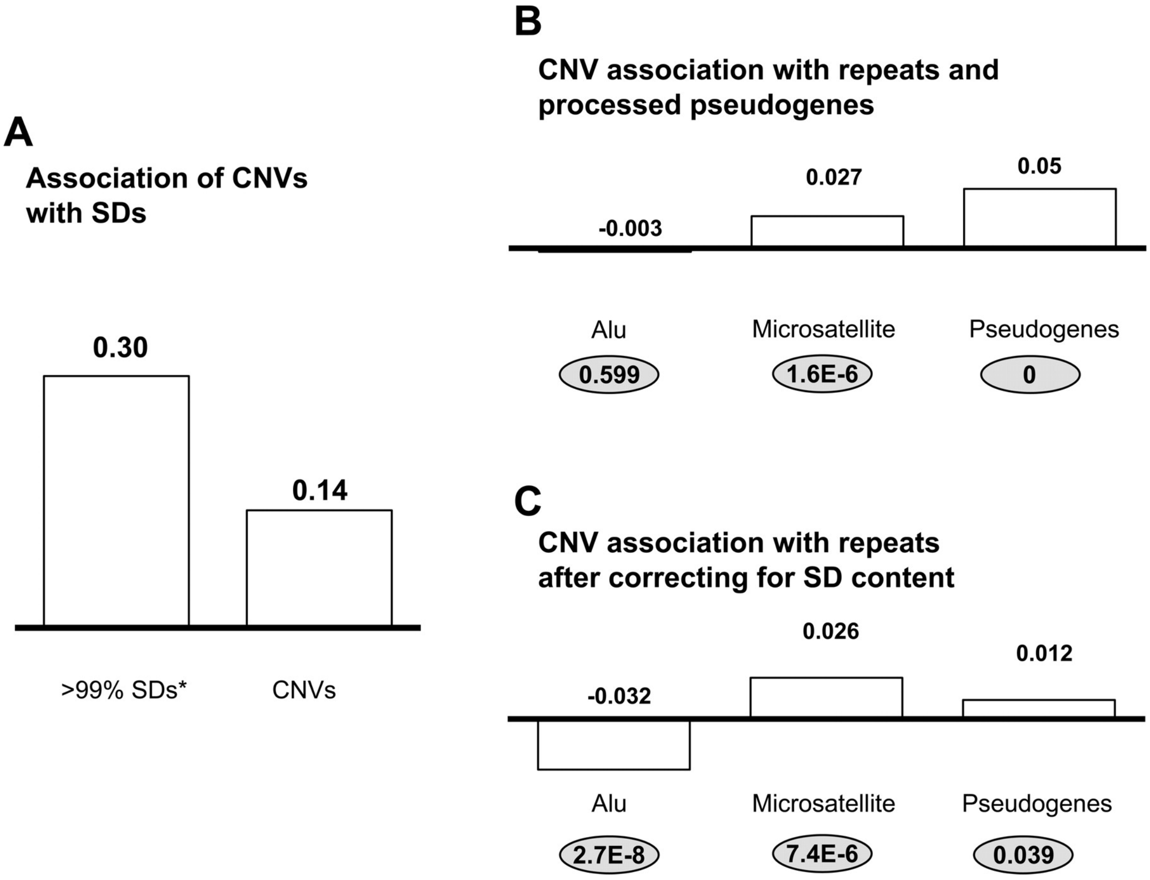


Figure 3.7: (A) Association of SDs and CNVs. Shown is the association of SDs ($90\% - 99\%$ sequence identity) with (left bar) ‘young’ SDs ($> 99\%$ sequence identity) and (right bar) CNVs. CNVs colocalize with SDs, but much more weakly than with very young SDs. Associations are given as Spearman rank correlations of the number of occurrences in genomic bins. All correlations are highly significant (P-value $<< 0.00001$). (B) CNV association with different human repeat elements. CNVs associate weakly with L1 elements and microsatellites, but show no association with Alu elements. (C) CNV association with human repeat elements after correcting for SD content. There is almost no significant association; the observed depletion in Alu elements may be due to a preference of CNVs for subtelomeric regions. Associations are given as Spearman rank correlations of the number of occurrences in genomic bins. P-values of the correlations are given in the bubbles.

Repeat Type	Frequency	Global enrichment	P-value	Local enrichment	P-value
Alu	0.09	0.94	3.24E-01	1.13	1.74E-01
SD	0.41	2.57	2.14E-07	1.17	2.64E-01
L1	0.24	1.48	1.03E-07	1.12	7.16E-02
L2	0.01	0.47	1.72E-02	0.52	2.31E-02
Microsatellite	0.03	3.91	6.74E-11	3.11	2.99E-07
LTR	0.09	1.14	1.71E-01	0.89	1.97E-01
P.Pseudogene	0.01	2.08	9.55E-02	1.66	1.98E-01
GC	0.39	0.96	7.24E-03	0.97	3.00E-02

Table 3.1: Association of SV breakpoints with several classes of repetitive elements. The relative enrichment (global) gives the enrichment relative to the global genomic background. The local relative enrichment gives the enrichment relative to a 50kb window around the breakpoint.

determination—SDs are read directly from the genome, and CNVs used in this study are determined using microarrays. Therefore, we computed associations between Alus and CNVs that were determined using very different methodologies, including different kinds of microarrays and paired-end sequencing. We conclude that Alu elements, while active in genome rearrangements in the past, do not currently play a major role in the formation of CNVs. It should be pointed out that this result does not contradict the notion of CNVs as drifting SDs—it simply suggests that the mechanism of CNV/SD formation may have undergone significant change in the past 40 million years.

The absence of association with Alu elements and the weakness of colocalization with SDs leads to the question of which genomic features are relevant for CNV formation. It has been suggested that microsatellite repeats have a role in mediation of chromosome rearrangements [72]. An association of SD junctions with microsatellites has previously been pointed out [52]. Hence, we examined whether they would associate with known CNVs. We indeed find that microsatellite repeats show a highly significant colocalization with CNVs (Figure 3.7B,C and Table 3.1), even after correcting for SD abundance.

3.2.7 Analysis of sequenced breakpoints

A difference between SDs and most of the current CNV data is that SD breakpoints are known exactly, whereas for CNVs only their approximate locations are known (based on CGH experiments). As mentioned above, most of the current data has a resolution of at best 50 kb [61]. To make authoritative statements about formation signatures, one has to analyze the exact sequences surrounding the breakpoints. Therefore, we performed targeted sequencing of a number of representative CNV breakpoints and identified a total of 134 breakpoints (Table 3.2). We combined this with previously sequenced breakpoints [11] to analyze a total set of 540 breakpoints, representative of all CNV events. To verify the trends we identified using the large-scale data, we analyzed the enrichment of different repeat elements in the immediate vicinity of the breakpoints and the existence of matching repeats flanking both sides of the breakpoints. To control for local sequence biases, we calculated the enrichment both with respect to the entire genome (global enrichment) and a 50-kb region around the breakpoint (local enrichment) (Table 3.1). We find only an extremely weak association with Alu elements, confirming the above trend. In total, we find 29% of the breakpoints to be associated with LINE repeats and another 2% to be associated with SDs. Nine percent were flanked by other repeat elements (e.g., LTR and others). The remainder (60%) of breakpoints did not show any homology signature. We should note here that the paired-end matching (using short sequence reads) approach is likely to bias somewhat against repeat-rich regions, and hence the fraction of NAHR-mediated CNVs may be higher in reality. This may also explain the discrepancy between the above found fraction of SD-mediated CNVs (maximally 28%) with the one found here ($\sim 2\%$). However, many exhibit signatures that may be indicative of non-homologous end-joining (NHEJ). Specifically, 40% of the breakpoints show the so-called microhomologies that can be a sign of NHEJ [73]. Another 14% exhibit microinsertions, which have also been implicated in NHEJ. We hence estimate that the latter CNVs may have been formed by double-strand breakage and NHEJ. Aside from these sequence signatures, there is also biophysical evidence: Breakpoints are enriched in regions that are known to be genomically unstable. We find that breakpoint regions tend to lie in GC-poor regions (Table 3.1), which are known to be thermodynamically less stable. Moreover, NHEJ breakpoints tend to lie in significantly less stable regions than NAHR breakpoints (P-value < 0.01). Also, we find that a few NHEJ breakpoints lie in the unstable subtelomeric regions, while no NAHR breakpoints do. We hence hypothesize that random breakage followed by NHEJ is one major mechanism for CNV formation.

3.3 Discussion

We have presented results that suggest changes in the formation of large genome rearrangements over the past 40 Mya. Our results suggest that shortly after the burst in Alu activity, Alu- or pseudogene-mediated mechanisms were predominant in the formation of SDs. The formed SDs then presented highly homologous regions themselves and were active shortly after formation in generating new SDs. However, it is striking to see that the association of SDs with Alu elements is decreasing with decreasing age of the SD (increasing sequence similarity between the duplicates) (Figure 3.4B). Likewise, the colocalization of SDs with their younger counterparts is decreasing. These trends are indicative of a lesser contribution of homology-mediated mechanisms for SD formation. At almost the same rate, preference of SDs for subtelomeric regions in the genome is increasing (Figure 3.4B). Genesis of SDs in subtelomeric regions is largely due to a mechanism based on NHEJ mediated by microhomologies (<25 bp homology), rather than a NAHR mechanism mediated by larger matching repeats [56]. Note that an alternative hypothesis for the enrichment of SD breakpoints in Alu-rich regions is the clustering of Alu elements [70, 71].

The lack of association of CNVs with Alu elements is quite surprising, as concurrent Alu-Alu recombination has been reported in the literature [74, 75]. However, our results indicate that while Alu-Alu recombination used to be a major force in shaping genome rearrangements, in the very recent genome evolution it did not leave a significant signature.

Chromosome	Start	End	Mechanism	Repeat
1	147600602	147986401	NAHR 272bp homology	SD
1	154793347	154795560	NAHR 19bp homology	None
1	157227979	157232826	NHEJ 4bp microhomology	None
1	208144678	208152601	NHEJ 6bp microinsertion	None
1	246118115	246124262	NAHR 14bp homology	None
2	126159721	126168302	NHEJ 4bp microhomology	None
2	146579091	146593333	NHEJ 2bp microhomology	None
2	54418997	54420978	NHEJ 3bp microinsertion	None
2	90959251	90972058	NAHR 205bp homology	Satellite
3	10201175	10203945	NHEJ 4bp microhomology	None
3	121644332	121647642	NHEJ 10bp microinsertion	None
3	188063727	188068042	NHEJ 45bp microinsertion	None
3	47465673	47468445	NHEJ 2bp microhomology	None
3	62639438	62670706	NHEJ 3bp microhomology	None
4	106926782	106936575	NAHR (repeat)	LINE/L1
4	108347263	108351179	NHEJ 11bp microinsertion	None
4	142450233	142452513	NHEJ 5bp microhomology	None
4	165024355	165039560		None
4	42457435	42464300	NAHR (repeat)	LINE/L1
4	58180961	58185488	NAHR (repeat)	LINE/L1
4	79488158	79494220	NAHR 14bp homology	None
5	10579961	10585291	NAHR 105bp homology	SINE/Alu
5	177754281	177756656	NHEJ 8bp microhomology	None
5	49471345	49476325	NAHR 303bp homology	Satellite/centr
5	57715747	57721855	NHEJ 4bp microhomology	None
5	71386	76029	NHEJ 3bp microhomology	SD
6	165644659	165652123	NHEJ 3bp microhomology	None
6	34045807	34050676	NHEJ 8bp microinsertion	None
7	113203412	113209444	NAHR 15bp homology	None
8	2116965	2122377	NHEJ 1bp microhomology	None
8	25122602	25126570	NHEJ 7bp microhomology	None
8	584397	589415	NHEJ 3bp microinsertion	None
8	73950329	73956378	NAHR 10bp homology	None
9	112516996	112519927	NHEJ 4bp microhomology	None
9	70927942	70933175	NHEJ 2bp microhomology	None
9	73446481	73449953	NHEJ 3bp microhomology	None
9	84854269	84860328	NAHR 15bp homology	None

10	114102173	114106649	NHEJ 2bp microhomology	None
10	128578838	128582206	NHEJ 10bp microinsertion	None
10	4427701	4431391	NHEJ 1bp microhomology	None
10	5627110	5677111	NHEJ 6bp microhomology	None
10	84117799	84120345	NHEJ 5bp microhomology	None
12	11075858	11142017	NAHR 170bp homology	SD
12	128624266	128628228		None
12	15909933	15912931	NHEJ 1bp microinsertion	None
12	38587965	38602082	NHEJ 13bp microinsertion	None
12	55618220	55663208	NAHR (repeat)	SD
12	94757723	94760459	NAHR 11bp homology	None
13	33033730	33042822		None
13	56650541	56686865	NHEJ 3bp microhomology	None
13	71705623	71710360	NHEJ 5bp microinsertion	None
14	105282154	105397044	NHEJ 3bp microhomology	None
14	34184839	34192011	NHEJ 2bp microhomology	None
14	73076457	73108631	NAHR 256bp homology	LINE/L1
14	81568863	81573084	NHEJ 10bp microinsertion	None
15	22009161	22111478	NAHR (repeat)	LTR/ERVL
15	68808907	68814563	NAHR 14bp homology	LINE/L1
16	29167046	86811700	NAHR 264bp homology	SD
16	76929139	76942400		None
18	14542177	14558726	NHEJ 8bp microhomology	SD
18	45948971	45952385	NHEJ 4bp microinsertion	None
20	28122727	28149711	NAHR (repeat)	SD
20	42760727	42762938	NHEJ 1bp microhomology	None
20	7044793	7050847	NAHR 12bp homology	None
21	19758801	19765198		None
22	27963089	27965391	NHEJ 3bp microhomology	None

Table 3.2: Newly sequenced CNV breakpoints. Most sequenced breakpoints show small homologies indicative of NHEJ. Furthermore, some breakpoints have microinsertions, which also indicate an NHEJ mechanism. Finally, some breakpoints show larger homologies, which suggest NAHR.

Furthermore, our sequenced breakpoints confirm that there is no significant enrichment of Alu elements near the breakpoints. Note, however, that there may be some bias of the sequencing method against Alu repeats. Moreover, it is in line with the emerging trend of decreasing Alu association of SDs. It is likely the result of the decrease in Alu activity since the Alu burst, which led to continuing Alu divergence and hence, diminishing probability of Alu-mediated NAHR. This finding is further bolstered by the fact that most SDs have a similar sequence divergence (age) as most Alus, that is, they were likely created around the Alu burst. While association does not imply causality, the lack of association (such as here, with Alu elements and CNVs) certainly implies lack of causality. In other words, it would be hard to argue that Alu elements are the predominant mediator of CNV formation solely based on the observation of colocalization. Thus, our observations provide strong evidence against the involvement of Alu elements in CNV formation.

On the other hand, it has previously been suggested that CNVs associate with SDs, and we find this trend persisting. However, SDs-mediated CNV formation can only account for a minority of the CNVs found ($< 10\%$ based on our sequenced breakpoints). Therefore, other mechanisms have to be at work as well. We suggest the following two possibilities for alternative mechanisms: First, we find associations of CNVs with other repeats, namely, microsatellites and LINE repeats. Large-scale associations only give weak evidence for this connection, but the presence of matching repeats in the immediate vicinity of the sequenced breakpoints makes a stronger case for microsatellites and LINE involvement in CNV formation. Since microsatellites have been implicated in genome rearrangements, an involvement in CNV formation would certainly be sensible [72]. Second, our findings are also suggestive of an increased role of NHEJ-based mechanisms for the generation of CNVs, which accounts for many of the breakpoints that were not associated with any known repeat. Indeed, we find an association of CNVs toward subtelomeric regions (P-value $<$ 0.001), where double-strand breakage and NHEJ are known to be prevalent. Moreover, in the sequenced breakpoint data, we find some indication that NHEJ is an alternative mechanism for CNV formation, such as the microhomologies present in many breakpoint sequences.

In summary, we find evidence for formation of duplications via NAHR that was mediated by repeat elements. While the colocalization does not imply causality, this mechanism has been proposed before and is supported by several pieces of data for SDs. It also explains nicely the decrease of colocalization of SDs with Alus and with each other. This leads to a coherent picture: $\sim$40 Mya, there was a peak in Alu activity, known as the Alu burst (Figure 3.8). The burst created a high number of repeat elements that served as templates for NAHR. Hence, ectopic recombination took place at a high rate and set off extensive genome rearrangement, thereby creating many SDs. The SDs themselves then could also serve as NAHR templates, 'feeding the fire' of recombination. This also nicely explains the existence of the rearrangement hot spots in the current human genome. Therefore, the majority of SDs that we find have low sequence identity ($\sim$ 90%), similar to Alu elements stemming from the burst, suggesting that they were formed during a similar time. Moreover, the number of SDs decreases with rising sequence identity, in sync with the decrease of Alu repeats (correlation $r = 0.92$, $P < 0.001$) (Figure 3.5). This is consistent with our hypothesis that the decline in retrotransposition activity then led to an overall decline in genome rearrangements. Moreover, the relative importance of other repeat elements, such as LINE elements or microsatellites, in terms of mediating NAHR increased; while they were created in the genome at a basal level, the strong effect of the Alu burst had previously masked their influence. This is why we find a stronger signature of enrichment of these elements with CNV breakpoint regions. Finally, other mechanisms play a much bigger role in reshaping the genome today, again consistent with the fact that a majority of current CNV breakpoints exhibit signatures suggesting a formation through NHEJ.

Aside from the factors discussed above, selection could have influenced the sequence signatures found around SDs or CNVs. Many SDs may have undergone some kind of selection during their way to fixation. In contrast, most CNVs are likely to be neutral, even though, analogous to SNPs, some may have been selected for or against [55, 11, 4]. As a result, one may assume that the differences between CNVs and SDs pointed out

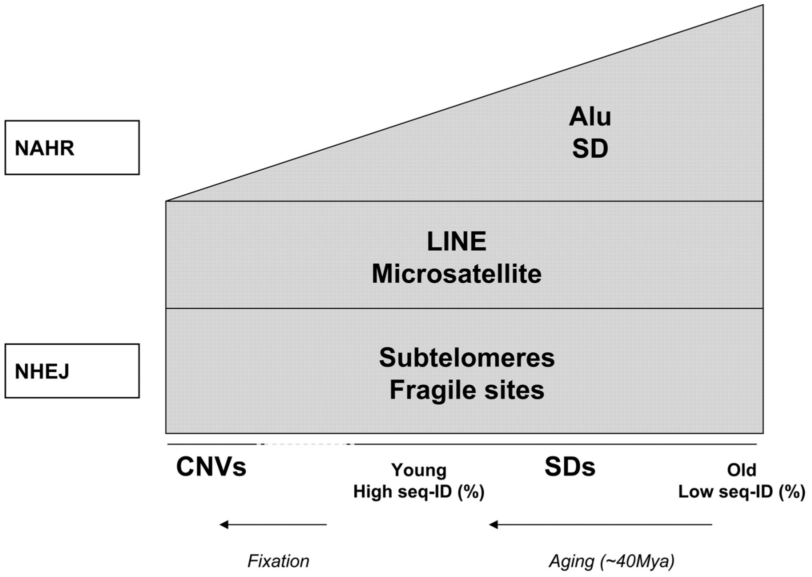


Figure 3.8: A schematic of the change of formation mechanism over the last 40 million years in the mammalian lineage.

above could be due to selection. The most striking difference is certainly the difference in association with Alu elements; if selection were responsible for this difference, two scenarios are possible: First, Alu elements in the vicinity of SDs could lead to preferential fixation of these SDs. It is hard to imagine how Alu elements in the genomic neighborhood should influence the fixation of SDs; therefore, we deem this scenario very unlikely. Second, Alu elements in the vicinity of CNV were removed by negative selection. This possibility is equally unlikely, and we believe that the far more parsimonious explanation is that Alu elements had a predominant role in past SD, but not in present CNV formation.

3.4 Methods

3.4.1 Sequence data preparation

We used the segmental duplications database from the University of Washington (`http://eichlerlab.gs.washington.edu/database.html`) based on the build 36 genome [49]. We binned all existing SDs into sequence identity categories and different size categories. To enable comparison with low-resolution copy number variation data, we finally binned all segmental duplications according to genomic coordinate. We varied the bin size from 10 kb to 1 Mb. Because the copy number variant mapping resolution is at most 50 kb for the techniques employed in the used data sets [61], we report the results for calculations with a bin size of 100 kb. For copy number variants, we used three separate data sets, based on three different assay methodologies. The three-way comparison should avoid biases that may have been introduced by a single method. First, we used the recent set from the Human Copy Variation Consortium, which was based on microarray methods [47]. Secondly, the structural variation data based on Fosmid-paired-end sequencing was used [45]. Finally, a comparison of two different genome assemblies has revealed putative copy number variations [76].

3.4.2 Breakpoint sequencing

A total of 67 CNVs identified by the paired-end matching (PEM) were sequenced using the following approach. PCR fragments were extracted either by gel purification or gel extraction with Millipore Ultrafree-DA centrifugal filter devices (Millipore Corp.) or by bead purification from the reaction mixture with Agencourt AMPure (Agencourt Biocience Corporation). Amplified fragment pools (50–150 fragments each) were randomly sheared by nebulization, converted to blunt ends, and adaptors ligated with the GS DNA Library Preparation kit according to the manufacturer's protocols (454 Life Sciences; Roche Diagnostics). The resulting single-stranded DNA shotgun libraries were then sequenced with 454 Sequencing. Both the resulting reads (median length = 250 bp) and contigs generated by 454's de novo assembler Newbler (see software user manual; 454 Life Sciences and Roche Diagnostics) were scanned for the respective SV breakpoints with BLAST alignment against the human reference genome; we required best hits to the genome for both portions of a read/contig matching on either side of a candidate breakpoint junction.

3.4.3 Repeat analysis

Different kinds of repeats were identified using the genome annotation on the UCSC Genome Browser, based on the output of RepeatMasker. As above, distributions of Alu elements, LINE elements, and microsatellites were binned according to their genomic coordinates. Recombination hot spot data were taken from the HapMap recombination data [77]. Data for the processed pseudogenes were obtained from Pseudogene.org [78].

3.4.4 Computation of associations

Coarse-grained colocalization was assessed by computing the Spearman rank correlations between the binned distributions of each feature (SD occurrence, CNV occurrence or repeat occurrence) per bin. This measure is an accurate and robust measure of association and is independent of any assumptions of the distribution of the respective features. We used a bin size of 100 kb for the analysis, but changes in the binning procedure did not have an effect on our results. This coarse-grained approach can identify larger-scale trends. It is especially suitable for the analysis of CNV associations because of the current low-resolution mapping of their breakpoints. However, it may not be able to pinpoint exact breakpoint characteristics.

For sequenced breakpoints, we calculated enrichments both in a global and a local context. In a global context, we compared the average number of a random nucleotide in the genome intersecting with a given genomic element to that of a breakpoint. Since this may be biased by local genomic context, we also calculated the average number of a random nucleotide intersecting with a given genomic element in a 50-kb window around the breakpoint.

3.4.5 Detailed SD breakpoint analysis for processed pseudogenes

For a detailed analysis of processed pseudogene enrichment at SD breakpoints, we analyzed all SD junctions for overlap with pseudogenes. Because of potential sequencing and alignment errors, we defined the SD junction as ± 5 bp around the annotated breakpoint. We then looked for SDs where pseudogenes overlapped the SD start and end junctions in both duplicated segments. For each of these, we then compared the parent genes of the two pseudogenes that overlapped the SD junctions. For pseudogenes with different parent genes, we compared their sequence similarity using FASTA.

To assess the significance of the overlap between the processed pseudogenes and SD junctions, we first picked genomic regions of the same size and number as SDs at random and compared the overlap with processed pseudogenes. No matching junctions that had matching pseudogenes were found. As a second procedure that captures potential sequence biases, we randomized the SD junctions in a 50-kb window around the actual junctions and calculated their overlap with matching pseudogenes.

3.4.6 CNV breakpoint analysis

To complement the coarse-grained approach, we analyzed a set of 540 sequenced breakpoints, a combination of the breakpoints from [11] and the newly sequenced breakpoints above. We analyzed the occurrence of breakpoints in known repeat sequences from RepeatMasker. Furthermore, we analyzed each breakpoint for the occurrence of microhomologies and microinsertions. All calculations were carried out using custom code in Matlab, R, and Perl.

Chapter 4

Nucleotide-resolution analysis of structural variants using a breakpoint library

4.1 Introduction

Structural variation of large segments (>1 kb), including copy-number variation and unbalanced inversion events, is widespread in human genomes [44, 79, 45, 47, 11, 80] with ~20,000 SVs presently reported in the Database of Genomic Variants (DGV) [79]. These SVs have considerabe impact on genomic variation by causing more nucleotide differences between individuals than single-nucleotide polymorphisms [47, 11, 80] (SNPs). In several genomic loci, rates of SV formation could even be orders of magnitude higher than rates of single nucleotide substitution [81, 82]. To measure the influence on human phenotypes of common SVs (that is, those present at substantial allele frequencies in populations) and de novo formed SVs, several studies have mapped SVs across individuals. They reported associations of SVs with normal traits and with a range of diseases, including cancer, HIV, developmental disorders and autoimmune diseases [6, 83, 84, 85, 5, 86]. Although most SVs listed in DGV are presumably common, de novo SV formation is believed to occur constantly in the germline and several mutational mechanisms have been proposed [87].

Nevertheless, so far our understanding of SVs and the way we analyze SV maps is limited by the limited resolution of most recent surveys, such as those solely based on microarrays, which have not revealed the precise start and end coordinates (that is, breakpoints) of the SVs. This has hampered our understanding of the extent and effects of SVs in humans, as mapping at breakpoint resolution can reveal SVs that intersect with exons of genes or that lead to gene fusion events [11, 15].

The lack of nucleotide-resolution maps has further prevented systematic deduction of the processes involved in SV formation, such as whether common SVs emerged initially as insertions or deletions at ancestral genomic loci. Instead, operational definitions have been applied for classifying common SVs into gains, losses, insertions and deletions based on either allele frequency measurements, or the ‘human reference genome’ (hereafter also referred as the reference genome) that was originally derived from a mixed pool of individuals [88]. Thus, inference of the ancestral state of an SV locus is crucial for relating SV surveys to primate genome evolution and population genetics.

The lack of data at nucleotide resolution has also limited the number of SVs for which the likely mutational mechanisms of origin have been inferred. These mechanisms are thought to include (i) NAHR involving homology-mediated recombination between paralogous sequence blocks; (ii) nonhomologous recombination (NHR) associated with the repair of DNA double-strand breaks (that is, nonhomologous end-joining) or with the rescue of DNA replication-fork stalling events (that is, fork stalling and template switching [60]); (iii) variable number of tandem repeats (VNTRs) resulting from expansion or contraction of simple tandem repeat units; and (iv) transposable element insertions (TEIs) involving mostly long and short interspersed elements (LINEs and SINEs) and combinations thereof, along with other types of TEI-associated events (e.g., processed pseudogenes).

Finally, owing to the lack of resolution of most SV maps, junction sequences (the flanking sequences of breakpoints) have thus far not been exploited for testing the presence of SVs in an individual in a similar fashion to the way SNPs can be directly detected by

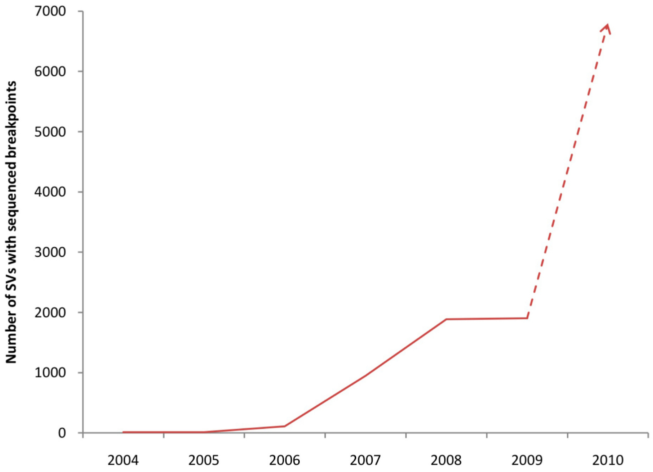


Figure 4.1: Number of SVs with sequenced breakpoints. The number of SVs with sequenced breakpoints has been increasing rapidly since 2006, due to the recent advance of sequencing technologies, and will grow substantially in the near future (indicated by the dashed line with arrow) owing to collaborative efforts such as the 1000 Genomes Project.

oligonucleotide chips with probes designed for each polymorphism.

Recent advances in microarray technology and large-scale DNA sequencing have paved the way for high-resolution SV maps. To date, nearly 2,000 SVs have been fine-mapped at nucleotide level and efforts such as the 1000 Genomes Project (`http://1000genomes.org`), which will soon sequence >1,000 human genomes, might in the near future report many more SVs at such resolution (Figure 4.1). Thus far, however, no study has leveraged the potential of collectively analyzing breakpoint-level SV data.

Here we present a comprehensive analysis of a library of nearly 2,000 SVs assembled from eight recent surveys that involve individuals from three distinct populations. We demonstrate four uses of the breakpoint library—mapping structural variation at high resolution, revealing ancestral states of variants, inferring mechanisms of variant formation and correlating the inferred mechanisms with DNA sequence features. We found several lines of evidence consistent with a nonuniform distribution of SV formation mechanisms and with locus-specific sequence properties, such as DNA helix stability, chromatin accessibility and the propensity for a DNA sequence to recombine, which may predispose genomic regions to SV-mutational processes.

4.2 Results

4.2.1 Generation of a standardized SV breakpoint library

We compiled a set of breakpoints from eight published sources (Figure 4.2). In accordance with a previously proposed operational definition [3], we defined SVs to be deletions, insertions and inversions reported relative to the reference genome with a size of 1 kb or larger. As our initial library encompassed SVs mapped using different types of evidence, sequencing technologies and genome assembly versions, an essential first step was library standardization. We therefore implemented a computational pipeline for generating a unified, nonredundant breakpoint library (Methods).

The pipeline yielded a nonredundant set of 1,889 SVs that were initially annotated as deletions (1,409), insertions (419) or inversions (61) relative to the reference genome. This set, which represents the most exhaustive compilation to date of SV breakpoints in phenotypically normal individuals, is available at `http://sv.gersteinlab.org/breakseq`. It also has been deposited into the BreakDB database [12] (`http://sv.gersteinlab.org/breakdb`).

4.2.2 High-resolution mapping of SVs from short-read sequencing data

Personal genomics endeavors based on next-generation sequencing technology [89, 92, 93] typically detect genomic variation by mapping relatively short sequencing reads directly onto the reference genome. Although many short indels (<1 kb) can be accurately identified with such an approach, SVs 1 kb are commonly missed, or not identified at nucleotide (that is, breakpoint-level) resolution. This is probably because of the difficulty in constructing accurate sequence alignments from short reads (e.g., 36 mers), especially if they involve long sequence gaps or span breakpoints.

We thus devised an approach, BreakSeq, for detecting SVs by aligning raw reads directly onto SV breakpoint junctions of the alternative, nonreference, alleles contained in our library (Figure 4.3a, Methods). Briefly, the genomic coordinates of each breakpoint in

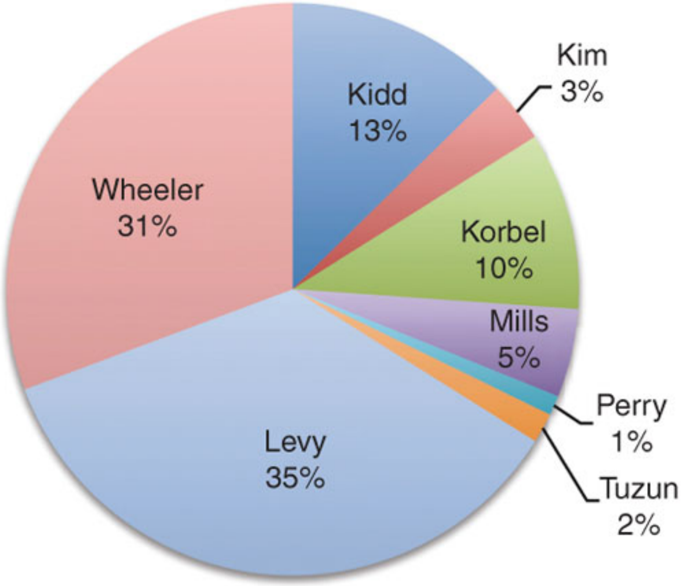
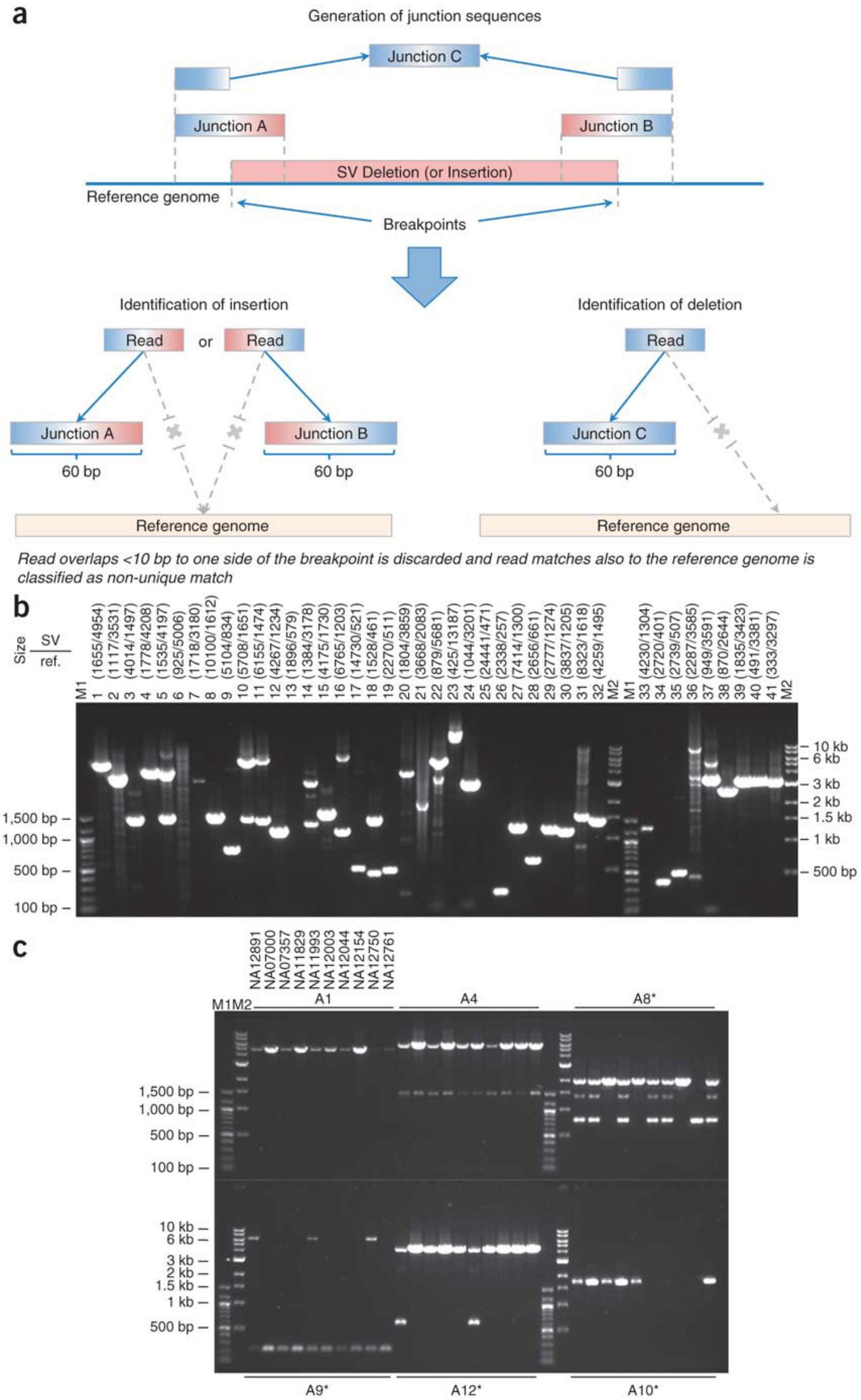


Figure 4.2: Composition of the SV breakpoint library. SVs in the library were based on different SV-mapping and breakpoint-sequencing strategies. A large fraction (44%) of the breakpoints were based on data generated using 454/Roche sequencing, including resequencing of an individual human genome (Wheeler [89], 602 SVs) and sequencing of breakpoints in two individuals after high-resolution and massive paired-end mapping (Korbel [11] and Kim [15], 264 SVs). The remaining 56% of the breakpoints were identified using other approaches, including Sanger capillary sequencing of breakpoints identified by whole-genome shotgun sequencing and assembly of an individual human genome (Levy [48], 694 SVs), fosmid-paired-end sequencing carried out in multiple individuals (Tuzun [45] and Kidd [80], 281 SVs), breakpoints mined from SNP discovery DNA resequencing traces (Mills [90], 98 SVs), and tiling-array-based comparative genomic hybridization followed by breakpoint sequencing (Perry [91], 22 SVs). Fewer than five breakpoints were reported in two genomes sequenced using short 36-bp reads (Illumina/Solexa) [92, 93], presumably owing to the complex DNA sequence patterns frequently associated with breakpoints [11, 80, 91].

the standardized library are used to extract 30 bp of flanking sequence from the reference genome. These 30-bp flanking sequences are concatenated into 60-bp junction sequences. Thus, a deletion event is represented with a single junction sequence in the library (containing the sequence flanking its single breakpoint), whereas an insertion has both left and right junction sequences (containing the sequence flanking each of its two breakpoints). DNA reads from personal genomes are aligned against the junction sequences. Successful alignment requires a read to overlap a junction sequence by at least 10 bp on each side of the breakpoint. This approach is conceptually similar to using a library of exon splice junctions in transcriptome analyses, which leads to considerably better coverage of alternatively spliced transcripts than restricting the analysis to reference genome sequences lacking splice junctions [94].

To demonstrate the utility of our approach for mapping personal SVs at high resolution, we mapped short reads from three personal genomes sequenced with Illumina/Solexa technology. These included two previously published genomes [92, 93] from individuals of Nigerian (Yoruba from Ibadan, YRI) and Han Chinese (HCH) origins. The third genome was from a HapMap individual of European ancestry (CEPH) that was sequenced recently

Figure 4.3 *(following page)*: Mapping breakpoints using the library. (a) Overview of the BreakSeq approach. Breakpoints are used to generate junction sequences spanning breakpoints (upper)Xthe 30 bp of sequence flanking each side of the breakpoint (60 bp total). Then, DNA reads are aligned to the junction sequences (lower). Alignment results are interpreted as follows. In the case of insertions relative to the reference genome (left), sequences A and B represent the left and right breakpoint junction sequences of the nonreference SV allele, respectively. In the case of deletions (right), sequence C represents the junction sequence of the nonreference SV allele. Solid lines with arrows, successful alignments. Dashed lines with crosses, no proper alignment. (b) Representative PCR validation of detected SVs in NA12891. Primers flanking each SV were used to amplify 41 different genomic regions. Expected band sizes for the reference and nonreference SV alleles are given at the top of each lane. The difference in size of the products for the reference and nonreference alleles confirmed the presence of the SVs for all loci except 6, 13 (confirmed by LongAmp Taq in a separate experiment), 21, 25 and 36. M1 is a 100-bp marker and M2 is a 1-kb marker. (c) A subset of SVs, which were confirmed by sequencing, was analyzed in nine additional genomic DNA samples (HapMap individuals with ancestry in Europe) to test for SV frequency within the CEPH population. An asterisk indicates that the SV is present polymorphically.


a
Generation of junction sequences
Junction C
Junction A
Junction B
SV Deletion (or Insertion)
Reference genome
Breakpoints
Identification of insertion
Read
or
Read
Junction A
Junction B
60 bp
60 bp
Reference genome
Identification of deletion
Read
Junction C
60 bp
Reference genome
Read overlaps <10 bp to one side of the breakpoint is discarded and read matches also to the reference genome is classified as non-unique match
b
Size SV/ref.
M1
1 (1655/4954)
2 (1117/3531)
3 (4014/14497)
4 (1778/4208)
5 (1535/4197)
6 (925/5006)
7 (1718/3180)
8 (10100/1612)
9 (5104/834)
10 (5708/1651)
11 (6155/1474)
12 (4267/1234)
13 (1896/579)
14 (1384/3178)
15 (4175/17330)
16 (6765/1203)
17 (14730/521)
18 (1528/461)
19 (2270/511)
20 (1804/3859)
21 (3668/2083)
22 (879/5681)
23 (425/13187)
24 (10444/3201)
25 (24441/471)
26 (2338/257)
27 (7414/1300)
28 (2656/661)
29 (2777/1274)
30 (3837/1205)
31 (8323/1618)
32 (4259/1495)
M2
M1
33 (4230/1304)
34 (2720/401)
35 (2739/507)
36 (2287/3585)
37 (949/3591)
38 (870/2644)
39 (1835/3423)
40 (491/3381)
41 (333/3297)
M2
1,500 bp –
1,000 bp –
500 bp –
100 bp –
– 10 kb
– 6 kb
– 3 kb
– 2 kb
– 1.5 kb
– 1 kb
– 500 bp
c
NA12891
NA07000
NA07357
NA11829
NA11993
NA12003
NA12044
NA12154
NA12750
NA12761
M1M2
A1
A4
A8*
1,500 bp –
1,000 bp –
500 bp –
100 bp –
10 kb –
6 kb –
3 kb –
2 kb –
1.5 kb –
1 kb –
500 bp –
A9*
A12*
A10*

in the pilot phase of the 1000 Genomes Project. To prioritize the SV calls generated by BreakSeq, we developed a scoring system based on supportive read-matches (the number of reads that map to a breakpoint; Methods) and distinguished low-support SV calls (with 1 to 4 supportive read-matches) from high-support SV calls. For the HCH, CEPH (NA12891) and YRI (NA18507) genomes, we identified 158, 219 and 179 SVs, respectively. Several SVs were shared among the three, suggesting that they may represent common alleles. For example, among the high-support calls, we found that 57 SVs were shared between the YRI and HCH genomes, 62 between the YRI and NA12891 genomes, 52 between the HCH and NA12891 genomes, and 42 were common to all three genomes.

To validate these results, we used PCR to test 24 insertion and 33 deletion calls predicted in NA12891 relative to the reference genome. Specifically, PCR amplification of predicted nonreference SV alleles [11] was used as a means for validation. In 48 cases the predicted SVs were validated, and in one case the reaction was inconclusive (Figure 4.3b and 4.8). Furthermore, seven reactions neither revealed the reference allele nor the predicted SV allele. (This primer failure rate can be explained by repetitive and GC-rich sequences that occur in association with SVs.) Finally, in a single case only the reference allele was found, suggesting either a false-positive prediction or the inability to amplify the event band of a predicted size of 7.5 kb.

We then sequenced 12 of the PCR-validated amplicons with Sanger capillary sequencing and confirmed the predicted breakpoint in all—that is, the Sanger-sequenced junction was identical to that in the library, with few single base-pair differences (presumable SNPs). We also analyzed a panel including nine unrelated CEPH individuals for the presence of six of the sequenced SVs and found that most SVs (four) were present polymorphically, whereas the remaining SVs likely represent rare alleles (Figure 4.3c). All together, 48 out of 57 predicted SVs (84.2%) were confirmed successfully, and the validation rate was estimated at 98% (48 out of 49) based on the PCR reactions that could be scored, demonstrating high specificity. Notably, as about half of our validated SVs were low-support SV calls, our validations demonstrate that accurate calls are generated both at high- and low-support levels. This suggests that BreakSeq may perform reasonably even in conjunction with low-coverage sequencing projects.

4.2.3 Inferring ancestral states of SV loci by comparing breakpoint junctions to primate genomes

Global SV surveys have so far reported SV events such as insertions and deletions using operational definitions—that is, comparisons with the human reference genome or allele frequency measurements. However, we reasoned that a systematic assessment of SV formation requires an unambiguous discrimination of SV event types—that is, one minimally affected by ascertainment biases. As the human reference genome presumably contains a mixture of common and rare SV alleles, it can serve only as a provisional reference for classifying SVs as insertions or deletions. Likewise, allele frequency measurements are of limited use in the context of classifying SVs into 'gains' and 'losses', as they may be affected by population-specific allele frequencies. In fact, ancestral state assignments facilitate systematic surveys of SVs in the context of studies focusing on human genome evolution, SV formational processes as well as minor and/or major allele assignment (as the ancestral allele often corresponds to the major one).

We therefore devised a framework that automatically assigns ancestral states of SV genomic loci based on a comparison of SV breakpoint junction sequence with the corresponding syntenic segments from the chimpanzee, orangutan and macaque genomes. Our approach (Figure 4.4a and Methods) involves extracting ± 500-bp flanking sequences around each breakpoint junction, combining them into putative ancestral regions (stretches resembling the allele present in the reference genome and stretches resembling the alternative allele), and then comparing the regions with syntenic primate genome sequences to deduce the most likely ancestral state. We defined SV loci as 'rectifiable' if unambiguous high-quality alignments to putative ancestral regions could be constructed for the loci in any primate genomes.

Overall, ancestral states of 1,281 (70%) out of 1,828 SV indel events could be assigned.

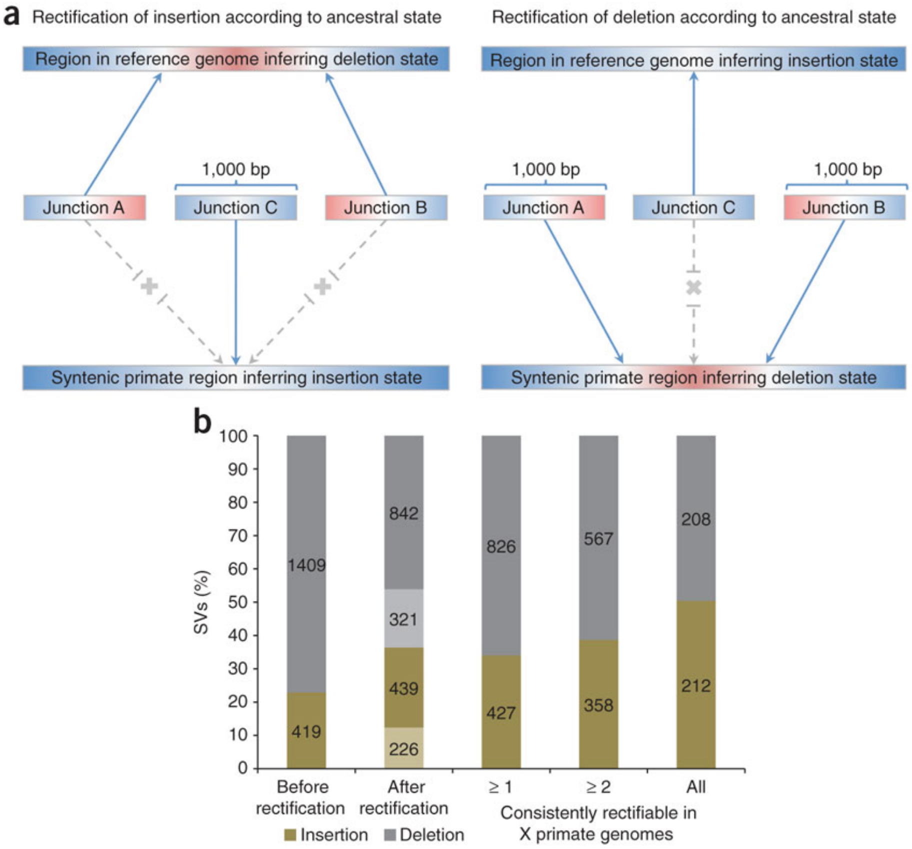


Figure 4.4: Ancestral state classification. (a) Junction sequences are aligned onto syntenic regions of a nonhuman primate genome to infer SV ancestral states. For rectifying an SV insertion event (from deletion) according to ancestral state (left), sequences A and B represent the junction sequences of the reference SV allele, whereas sequence C represents the junction sequence of the nonreference SV allele. For rectifying an SV deletion event (from insertion) according to ancestral state (right), sequence C represents the junction sequence of the reference SV allele and sequences A and B represent the junction sequences of the nonreference SV allele. Solid lines with arrows indicate successful alignments and dashed lines with crosses indicate no proper alignment. (b) Results of classifying SVs as insertions or deletions according to ancestral state. An SV event is defined as 'rectifiable' (indicated by darker color) if unambiguous high-quality alignments to putative ancestral regions could be constructed for the loci in any primate genomes (regardless of whether the classification is changed), and as 'unrectifiable' (represented by lighter color) if not.

For the vast majority of these (1,142), the chimpanzee genome contributed to the ancestral state assignment. For an additional 139 cases located in hard-to-align regions in the chimpanzee genome (e.g., sequence assembly gaps), the ancestral state was inferred based on aligning junctions to the orangutan and macaque genomes. After ancestral state assignment, 665 SVs (36%) were classified as insertions and 1,163 (64%) as deletions. Furthermore 925 out of the 1,281 events were consistently rectifiable in at least two genomes. Of those, 420 were consistently rectifiable in all three genomes, with an approximate balance between insertions (212) and deletions (208) (4.4b). We note that this balance differs substantially from earlier provisional SV classifications, which were strongly biased toward deletions, probably owing to the difficulty of many SV detection approaches in identifying insertions relative to the reference genome.

4.2.4 Inferring mechanisms of SV formation

Breakpoint junction sequences can also be used to deduce the molecular mechanisms of origin for SVs. To systematically classify SVs in our library, we evaluated previously reported signatures of particular formation mechanisms (such as VNTR, TEI, NAHR and NHR) with a computational pipeline (Figure 4.6a and Methods). TEIs can be identified by the underlying genomic signatures of transposable elements; VNTRs, by underlying tandem repeats and low-complexity DNA; NAHRs, by the extended stretches of high sequence identity at the breakpoint junctions; and NHRs, by events lacking the former patterns. Parameters of the pipeline were chosen so as to yield results comparable to those achieved manually; in this regard, we confirmed the applicability of the chosen parameters by performing a sensitivity analysis (Methods).

We found, consistent with earlier findings based on considerably smaller data sets [11, 91], that NHR events constitute the most abundant mechanism of SV formation in the genome (Figure 4.6b). Our analyses inferred NHR as the formation mechanism for nearly half of all SVs in our set (45%), whereas 28% involved NAHRs, 21% involved TEIs, 5% involved VNTRs and 2% were ambiguous. Although VNTRs have the ability to contract and expand more than a kilobase, most of the 92 VNTRs identified in this study involved simple repeat units <1 kb in size. We thus reasoned that they do not fall strictly into the stringent SV definition given above and excluded VNTRs from most of the remaining analyses below. Additionally, for NAHR and TEI mechanisms, we focused on the high-confidence sets in the analyses unless indicated otherwise (Methods).

We then analyzed SV formation mechanisms of 1,281 rectifiable SV-indel events. As discussed above, SVs were provisionally mostly reported as deletions owing to ascertainment biases [11, 15, 89], regardless of the respective formation mechanisms. For example, despite the fact that retrotransposons are thought to move within the genome by a 'copy-and-paste process' involving reverse transcription of RNA intermediates and insertion of full-length or fragmented mobile elements [95], most TEIs were previously annotated as deletions. Nevertheless, our ancestral state analysis revised the actual locus origin for a considerable number of SVs, and helped to resolve this apparent contradiction. Our results show that nearly all SVs associated with transposable elements for which ancestral states could be assigned were categorized as insertions (98%).

Through manual inspection, we found that the remaining transposable elementVassociated deletions can be reasonably explained as NHR-mediated SV deletions in regions of concentrated transposon annotations, which are difficult to distinguish from retrotranspositions. This shows that using the class name TEI was justified in retrospect, and that our ancestral analysis pipeline is able to produce results consistent with prior knowledge on the formation mechanism of TEI. On the other hand, even after classification by ancestral states, NAHR and NHR events were mostly annotated as deletions (Figure 4.6c), which may be due to biases of these formation mechanisms toward deletions (as previously reported for NAHR [81]) or due to biases in SV detection methods toward ascertaining deletions in ancestral loci.

Further analysis of TEI events showed that they involved LINEs, SINEs, LTR-elements, composite retrotransposons and processed pseudogenes. Our results show that LINE-1s (L1s) represent the most abundant class at the given size range (>1 kb) as expected

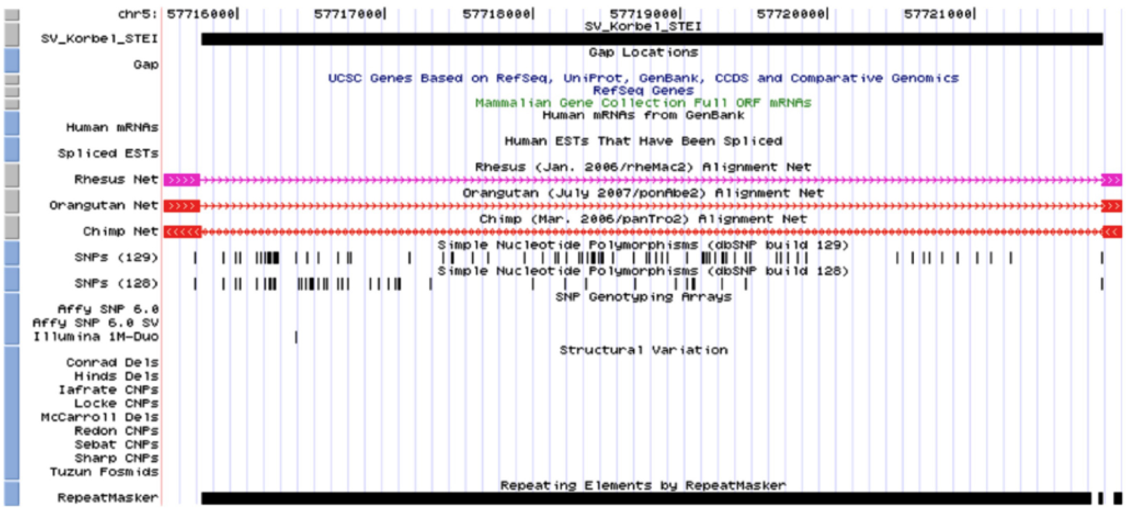

chr5: 57716000| 57717000| 57718000| 57719000| 57720000| 57721000|
SV_Korbel_STEI
Gap
Gap Locations
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
RefSeq Genes
Mammalian Gene Collection Full ORF mRNAs
Human mRNAs
Human mRNAs from GenBank
Spliced ESTs
Human ESTs That Have Been Spliced
Rhesus Net
Rhesus (Jan. 2006/rheMac2) Alignment Net
Orangutan Net
Orangutan (July 2007/ponAbe2) Alignment Net
Chimp Net
Chimp (Mar. 2006/panTro2) Alignment Net
SNPs (129)
Simple Nucleotide Polymorphisms (dbSNP build 129)
SNPs (128)
Simple Nucleotide Polymorphisms (dbSNP build 128)
Affy SNP 6.0
Affy SNP 6.0 SV
Illumina 1M-Duo
SNP Genotyping Arrays
Structural Variation
Conrad Dels
Hinds Dels
Iafrate CNPs
Locke CNPs
McCarroll Dels
Redon CNPs
Sebat CNPs
Sharp CNPs
Tuzun Fosmids
RepeatMasker
Repeating Elements by RepeatMasker

Figure 4.5: UCSC browser view of an example of active L1 insertion in the human genome. The top bar represents an insertion event of a SV (STEI) after rectification. The colored arrows represent the corresponding syntenic regions in the other primate genomes. The bottom bar represents the RepeatMasker annotated L1, which is also annotated as active by Mills et al., 2006.

[96], with 71% of the TEIs mediated by LINE/L1 transposable elements. Although many transposable elements in the human genome have lost their ability to retrotranspose autonomously, several full-length elements, including 147 L1s, are still implicated in recent or ongoing retrotransposition activity [95]. Interestingly, our results suggest the possible recent activity in the human population of at least 84 L1 elements, which were reported by our pipeline as 'full-length' with poly-A tracts and target-site duplications. To the best of our knowledge, 38 of these putative active mobile elements have not yet been implicated with recent L1 activity (Figure 4.6b, Table 4.1). The remaining TEIs include three potential processed pseudogenes that were identified on the basis of their spliced primary transcripts, poly-A tracts and target site duplications (Figure 4.6b, Table 4.1).

We then focused on SVs associated with NAHR and NHR. Because these SVs mostly involve deletions relative to ancestral sequence, we reasoned that they might represent a particularly interesting class of SVs with potential impact on conserved DNA sequence. In fact, we found that 41% and 33% of the NAHR and NHR-based deletions, respectively, intersect with annotated exons from RefSeq genes (Methods) and thus may have a functional

Chr	Start	End	SV Size	STEI
chr1	216249121	216255226	6106	LINE/L1
chr2	4759167	4765234	6068	LINE/L1
chr3	89592655	89598696	6042	LINE/L1
chr4	18688549	18694648	6100	LINE/L1
chr4	81107079	81113123	6045	LINE/L1
chr4	167913617	167919651	6035	LINE/L1
chr5	151436616	151442649	6034	LINE/L1
chr6	85374869	85380945	6077	LINE/L1
chr7	96313827	96319934	6108	LINE/L1
chr10	6451585	6457650	6066	LINE/L1
chr20	53868019	53874024	6006	LINE/L1
chr10	6451587	6457652	6066	LINE/L1
chr2	4759186	4765250	6065	LINE/L1
chr20	53868031	53874036	6006	LINE/L1
chr4	18688564	18694663	6100	LINE/L1
chr4	81107087	81113131	6045	LINE/L1
chr5	151436624	151442657	6034	LINE/L1
chr6	85374871	85380947	6077	LINE/L1
chr6	86765474	86771525	6052	LINE/L1
chr7	96313838	96319945	6108	LINE/L1
chrX	80983310	80989360	6051	LINE/L1
chr4	147444746	147444747	6049	LINE/L1
chr5	89486537	89486538	6089	LINE/L1
chr3	75963555	75969765	6211	LINE/L1
chr20	53868018	53874023	6006	LINE/L1
chr5	151436607	151442640	6034	LINE/L1
chr10	91707613	91707614	6047	LINE/L1
chr15	52916843	52916844	6075	LINE/L1
chr15	81348674	81348675	6055	LINE/L1
chr18	49679741	49679742	6047	LINE/L1
chr2	169813375	169819421	6047	LINE/L1
chr2	41904907	41904908	6049	LINE/L1
chr3	20723908	20723909	6101	LINE/L1
chr4	82425590	82425591	6078	LINE/L1
chr4	88487270	88493324	6055	LINE/L1
chr6	72856220	72856221	6103	LINE/L1
chr6	86765466	86771517	6052	LINE/L1
chr8	123713059	123713060	6089	LINE/L1
chr1	166291199	166292371	1173	Pssd. Pseudogene
chr19	23825024	23825025	1147	Pssd. Pseudogene
chr19	14593346	14595142	1797	Pssd. Pseudogene

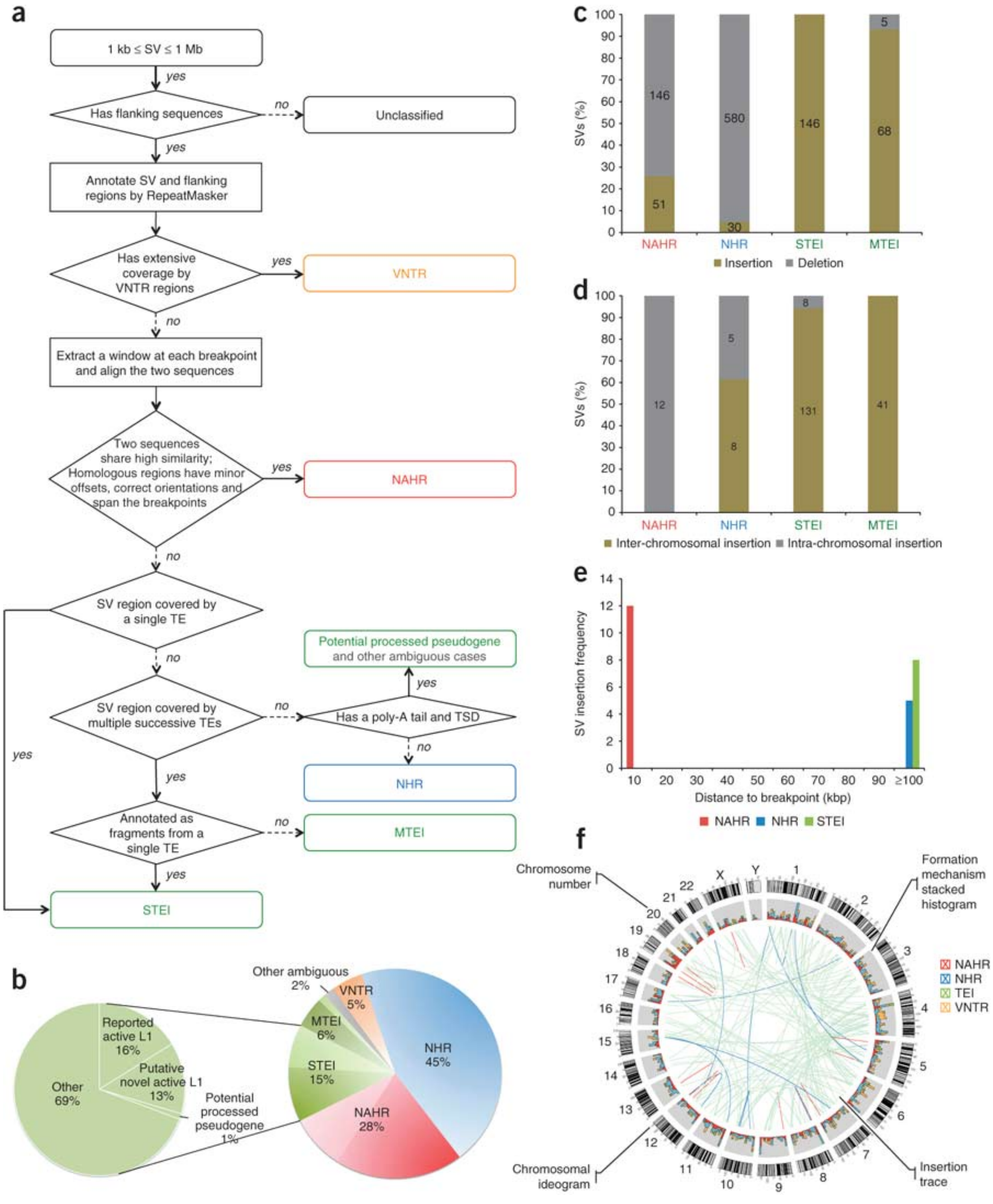
Table 4.1: Potential active L1 elements (38) and processed pseudogenes (3).

impact. On the other hand, insertions generated by NAHR or NHR have thus far received little attention, presumably due to difficulties in tracing these. Therefore, we extended our analysis to infer the most likely loci of origin of the inserted DNA sequences for 427 consistently rectifiable insertions (Methods). We found that NAHR insertions usually involve nearby sequence stretches stemming from the same chromosome as would be expected from the NAHR duplication mechanism. On the contrary, TEIs were found to originate randomly from inter-chromosomal locations in the genome, probably owing to the nature of retrotransposition of RNA intermediates. Furthermore, NHR-based insertions commonly involve both intra- and inter-chromosomal rearrangements (Figure 4.6d–f).

4.2.5 Insights into SV formational biases

Finally, we analyzed the relationship between mechanisms of SV formation and sequence features located near to the breakpoints (including chromosomal landmarks, recombination hotspots, repeat sequences, GC content, short DNA motifs and microhomology regions). Briefly, we first extracted the DNA sequences flanking both sides of each breakpoint junction. In the case of insertions, junction sequences included flanking DNA reconstructed

Figure 4.6 *(following page)*: Inferring mechanisms of SV formation. (a) Pipeline for classifying SV-formation mechanisms. TE, transposable element. TSD, target site duplication. (b) Mechanisms of formation inferred for SVs in the library (larger circle on right). In NAHR (red) and MTEI/STEI (green), darker wedges represent high-confidence classification subsets, and lighter wedges are extended subsets. STEI is further subdivided in the left circle according to the fraction of previously reported L1 insertions26, novel L1 insertions and processed pseudogene insertions in our data set. STEI, single transposable element insertion; MTEI, multiple transposable element insertion. (c) SV-indel distribution for all rectifiable events, broken down by formation mechanism. (d) Distribution of inter- versus intra-chromosomal events for all consistently rectifiable insertions, broken down by formation mechanism. (e) Distances of putative ancestral loci to insertion sites for all consistently rectifiable intra-chromosomal insertions, showing that intra-chromosomal NAHR insertions usually involve nearby sequences, whereas TEIs and NHR-associated insertions usually involve distant sequences. (f) Genome-wide view of insertion trace. The outermost circle represents chromosomal ideograms; the second circle represents SV formational mechanisms of 1,554 events in a stacked histogram. The lines in the innermost circle indicate the origin of the insertion sequences in the human genome for all 321 consistently rectifiable insertions.


a
1 kb ≤ SV ≤ 1 Mb
yes
Has flanking sequences
no
Unclassified
yes
Annotate SV and flanking regions by RepeatMasker
Has extensive coverage by VNTR regions
yes
VNTR
no
Extract a window at each breakpoint and align the two sequences
Two sequences share high similarity; Homologous regions have minor offsets, correct orientations and span the breakpoints
yes
NAHR
no
SV region covered by a single TE
no
Potential processed pseudogene and other ambiguous cases
yes
SV region covered by multiple successive TEs
no
Has a poly-A tail and TSD
no
NHR
yes
yes
Annotated as fragments from a single TE
no
MTEI
yes
STEI
b
Other 69%
Reported active L1 16%
Putative novel active L1 13%
Potential processed pseudogene 1%
Other ambiguous 2%
VNTR 5%
MTEI 6%
STEI 15%
NAHR 28%
NHR 45%
c
SVs (%)
100
90
80
70
60
50
40
30
20
10
0
146
51
580
30
146
5
68
NAHR
NHR
STEI
MTEI
Insertion
Deletion
d
SVs (%)
100
90
80
70
60
50
40
30
20
10
0
12
5
8
8
131
41
NAHR
NHR
STEI
MTEI
Inter-chromosomal insertion
Intra-chromosomal insertion
e
SV insertion frequency
14
12
10
8
6
4
2
0
10
20
30
40
50
60
70
80
90
≥100
Distance to breakpoint (kbp)
NAHR
NHR
STEI
f
Chromosome number
Formation mechanism stacked histogram
NAHR
NHR
TEI
VNTR
Chromosomal ideogram
Insertion trace
X
Y
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22

from the inserted sequence. We also generated two random background sets, one by randomly picking sequences from the reference genome (global background), and the other by randomly picking DNA sequences from the local sequence context specific to each mechanistic class (local background). We then identified sequence features in the flanking regions of each breakpoint and calculated their enrichment with P-values based on randomization tests (Methods).

We correlated SVs with chromosomal landmarks and found that NAHR events are significantly (P $\leq$ 1E-05) more proximal to telomeres and human-chimp synteny block boundaries than the other mechanistic classes. Moreover, we observed that VNTRs are significantly (P $\leq$ 1E-10) enriched in centromeric and pericentromeric regions, as expected (Figure 4.7a). These results demonstrate a nonuniform distribution of SV formation mechanisms in the human genome (Figure 4.6f).

We correlated SVs with recombination hotspots [97] and observed that they are significantly enriched for NAHR events (1.5-fold enrichment; P = 2.96E-03). Recombination hotspots are typically enriched for segmental duplications [62], which may act as mediators for NAHR during meiotic recombination. We further observed biases toward recombination hotspots for TEIs (Table 4.2), but not for NHR-mediated events. Whereas the accumulation of TEIs might in part be due to the formation of such elements by NAHR-mediated recombination involving interspersed repeat sequence, the lack of an enrichment for NHR indicates that DNA double-strand breaks occurring during recombination might be insufficient for initiating double-strand repair mediated by nonhomologous end-joining.

We assessed associations between SV formation mechanisms and common repeat elements in the genome. For example, NAHR events have previously been reported to be associated with various types of genomic DNA repeats, in particular segmental duplications [11, 80, 15]. After classification of NAHR events by our pipeline, we confirmed that significant (P $\sim$ 0) associations with segmental duplications are present both for NAHR-insertions (3.9-fold) and NAHR-deletions (7.4-fold). Furthermore, we found NAHR significantly (P $\sim$ 0) associated with the SINE/Alu class of mobile elements. On the other hand, LINE

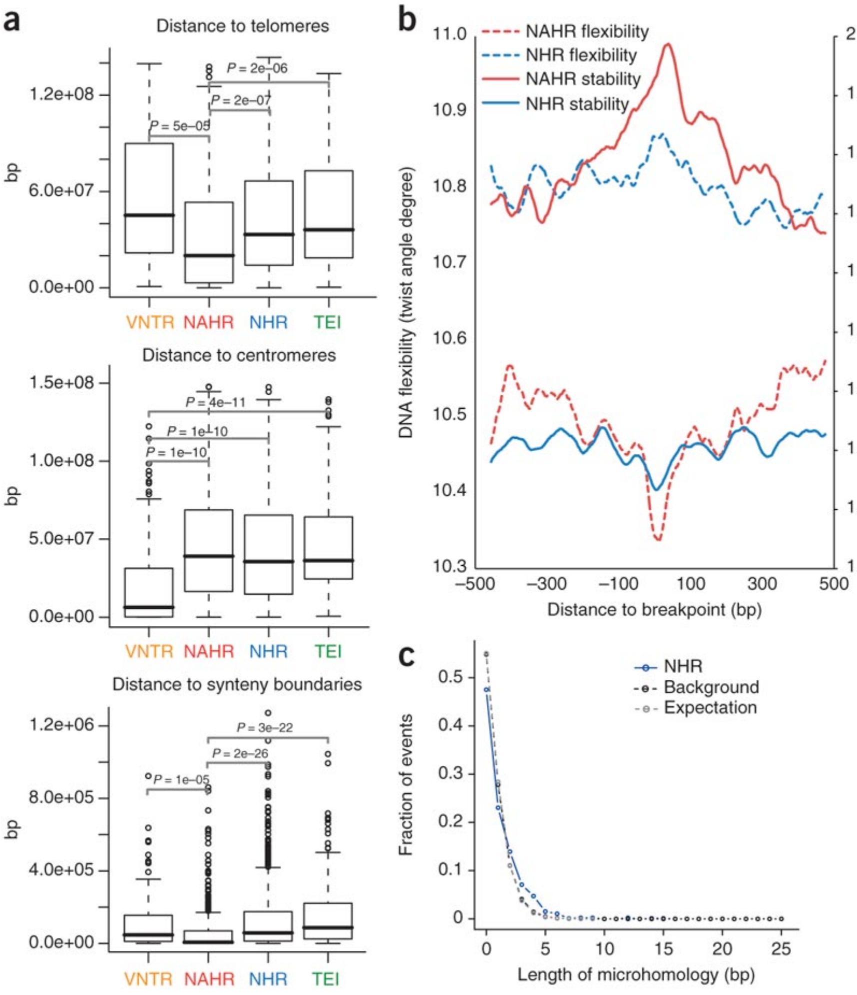


Figure 4.7: Analysis of breakpoint features. (a) Distance to chromosomal landmarks. Brackets indicate significantly different classes (P < 0.05 in Wilcoxon rank sum test after multiple hypothesis test correction by the Holm method). NAHR events are found to be significantly closer to telomeres and human-chimpanzee synteny block boundaries than the other mechanistic classes; VNTRs are significantly enriched in centromeric and pericentromeric regions. (b) DNA flexibility (dashed lines and left y-axis) and helix stability (solid lines and right y-axis) around NAHR and NHR breakpoints. (c) Distribution of NHR events with different lengths of microhomologies at the breakpoints. Microhomologies are significantly enriched in NHR breakpoints compared to a random background (KS test, P = 2.43E-11).

NAHR	Observed	Global Enrich.	P-value	Local Enrich.	P-value
Recomb. Hotspots	0.13	1.51	1.75E-03	1.04	4.03E-01
SINE/Alu	0.29	2.72	0.00E+00	2.17	0.00E+00
SINE/MIR	0	0.14	1.85E-03	0.23	2.30E-02
LINE/L1	0.1	0.59	1.54E-04	0.67	3.50E-03
LINE/L2	0	0.13	1.18E-03	0.22	1.79E-02
Dupl. Pseudogene	0.03	4.51	8.17E-08	0.96	4.52E-01
Pssd. Pseudogene	0	1.59	2.86E-01	0.66	3.13E-01
SD	0.96	5.95	0.00E+00	1.16	1.34E-01
GC	0.49	1.21	0.00E+00	1.06	1.61E-05
Flexibility	10.33	0.96	0.00E+00	0.98	3.90E-04
Helix Stability	1.99	1.07	0.00E+00	1.03	4.88E-07
NHR	**Observed**	**Global Enrich**	**P-value**	**Local Enrich.**	**P-value**
Recomb. Hotspots	0.09	1.02	4.26E-01	1.03	3.90E-01
SINE/Alu	0.13	1.16	5.45E-02	1.13	9.43E-02
SINE/MIR	0.02	0.74	1.00E-01	0.9	3.21E-01
LINE/L1	0.18	1.04	3.04E-01	0.95	2.62E-01
LINE/L2	0.02	0.62	2.27E-02	0.66	3.68E-02
Dupl. Pseudogene	0.01	1.29	2.64E-01	0.98	4.77E-01
Pssd. Pseudogene	0	0.91	4.50E-01	0.68	2.95E-01
SD	0.33	2.06	1.86E-06	0.9	2.55E-01
GC	0.41	1.01	3.19E-01	1	4.51E-01
Flexibility	10.86	1.01	1.87E-02	1.01	1.85E-02
Helix Stability	1.85	0.99	1.93E-02	0.99	2.25E-02
TEI	**Observed**	**Global Enrich**	**P-value**	**Local Enrich.**	**P-value**
Recomb. Hotspots	0.12	1.39	3.11E-02	0.9	2.77E-01
SINE/Alu	0.05	0.47	1.45E-03	0.7	9.48E-02
SINE/MIR	0	0.14	1.01E-02	0.24	5.20E-02
LINE/L1	0.29	1.66	7.88E-07	0.76	1.43E-03
LINE/L2	0.01	0.25	1.49E-02	0.42	1.00E-01
Dupl. Pseudogene	0	0	1.21E-01	0	2.01E-01
Pssd. Pseudogene	0	0	2.17E-01	0	2.09E-01
SD	0.09	0.57	1.55E-01	1.04	4.70E-01
GC	0.31	0.76	0.00E+00	0.76	0.00E+00
Flexibility	10.38	0.96	0.00E+00	0.96	0.00E+00
Helix Stability	1.84	0.99	3.71E-02	0.99	4.00E-02

Table 4.2: Enrichment analysis of features at breakpoint junctions generated by different mechanisms.

elements (both the L1 and L2 classes) were significantly ($P \leq$ 1E-03) depleted among the NAHR events in our set whereas NHR events did not show significant enrichment (or depletion, except marginally for L2) with genomic repeat-structure (Table 4.2).

We analyzed various features related to the physical properties of DNA at SV breakpoint junctions. In contrast to NHR, NAHR events were found to be biased toward GC-rich regions (Table 4.2). A possible explanation for this bias is the known GC-richness of recombination hotspots [98], which we found to be significantly ($P =$ 2.96E-03) enriched for NAHR events. Further, our results may indicate SV formation biases owing to DNA duplex stability. We thus extended our analyses by two additional features: DNA helix stability predicted by calculating the average of the dissociation free energy of each overlapping dinucleotide [99], and DNA flexibility based on the calculation of the average of the twist angle among each overlapping dinucleotide [100]. Our results indicate that in contrast to NAHR, NHR events are associated with high DNA flexibility and low helix stability, both of which are believed to be markers of fragility [101]. This is possibly due to sequence-specific biases for SV formation (Table 4.2). We went on to characterize the change of these fragility marker signatures in a region of ± 500 bp around the breakpoint by smoothing the signal with a 50-bp sliding window. Interestingly, we observed that the strength of the marker signatures was most extreme at or very close to the SV breakpoints (Figure 4.7b).

We reasoned that our comprehensive breakpoint junction library may enable us to identify simple DNA sequence motifs associated with SV breakpoints. Thus, we used the MEME tool [102] to carry out a comprehensive search for DNA motifs (6–12 nt, Methods) and found a significant enrichment (2.1-fold; $P \sim 0$) of the dinucleotide repeat $(TG)_6$ near breakpoints of NHR events, a sequence motif that fits with their relatively neutral GC content as shown above. We further analyzed all the NHR breakpoint sequences and found that the maximum consecutive occurrence of the TG-dinucleotide was 26. The MEME search did not reveal significantly enriched sequence motifs near NAHR or TEI events. Nevertheless, we used the MAST tool [102] to search for the DNA sequence motif ‘CCNCCNTNNCCNC’ that recently was reported to be associated with chromosomal recombination hotspots [103], and found a significant enrichment (1.5-fold; P $\sim$ 0) of the motif near NAHR-associated SVs, but not near NHR- or TEI-associated SVs.

Previous studies have observed the occurrence of stretches of short repeating sequences of 2 to ∼10 bp (that is, microhomologies) at the breakpoints of NHR events [60, 56]. We used our breakpoint junction library to scan NHR breakpoints for microhomology stretches of different lengths, and observed statistical enrichment relative to a random background (1.4-fold on average; KS test, P = 2.43E-11; Figure 4.7c) as expected. This suggests a strong association of microhomology stretches with SV formation by nonhomologous end-joining [56] or fork stalling and template switching [60].

4.3 Discussion

In this study we presented a comprehensive library of 1,889 nonredundant SVs identified by breakpoint-resolution mapping in eight studies. Our approach, BreakSeq, leverages a breakpoint junction library for SV detection. Whereas other computational approaches for SV detection (such as paired-end mapping [11, 104], DNA read-depth analysis [105, 106, 9] and split-read alignment analysis [13]) remain essential for identifying previously unknown SVs (a process that typically involves targeted PCR and sequencing), our approach serves as a tool for rapidly identifying specific SV alleles in personal genomics data. Specifically, by mining personal genomes for sequences present in the breakpoint junction library, BreakSeq leverages alternative, nonreference genomic sequence data to rapidly detect previously described SVs that short-read based personal genomics surveys commonly fail to ascertain. As such, BreakSeq enables a step towards overcoming reference bias, which is the favoring in ascertainment of SV alleles present in the human reference genome sequence.

We foresee that the utility of BreakSeq will increase as data sets grow (e.g., when SV calls from the 1000 Genomes Project are published). As our approach has a linear time complexity (Methods), it is easily extendable to larger data sets. In this regard, the size of our junction library currently comprises 0.004% of the reference genome in terms of nucleotide bases, and even a 100-fold increase of its size (>0.2 million SVs; ~10 times of DGV) will result in a data set considerably smaller than the reference genome. Thus, applying BreakSeq in personal genomics studies adds negligible computing efforts (compared to SNP genotyping) and at the same time dramatically improves SV calling. The library will be updated regularly to serve the personal genomics community in enabling precise SV detection with various next-generation sequencing platforms.

In addition to enabling accurate SV mapping, our junction library allows characterizing SV ancestral states. Whereas the ancestral states of SNPs and small indels have been inferred according to ancestral alignments in earlier studies [107, 108], we here report systematic ancestral state inference for SVs. When applying our new classification approach to 1,281 SVs, we found that overall there is a balance of insertions and deletions, unlike most currently published SV sets that display a considerable bias toward deletions. It should be noted that the nonhuman primate genomes used in our ancestral state inference correspond to single animals, which certainly do not represent idealized ancestral genomes. Nonetheless, we reasonably assume that SV loci can be classified at high confidence when ancestral states can be consistently inferred across three distinct primates.

Furthermore, we have developed a computational pipeline for classifying SVs according to their formation mechanisms and for analyzing various DNA sequence characteristics of the affected genomic loci. Together with the ancestral state analysis, this allowed us to analyze SV formation processes with respect to likely ancestral loci, an analysis that revealed some insights into SV formation. For example, our analyses suggest that the physical properties of the underlying DNA sequence influence locus-specific propensities for different SV formation mechanisms. We observed that NAHR-based SVs are associated with a relatively high GC content and with recombination hotspots, indicating that double-strand breaks occurring specifically during meiotic recombination contribute to NAHR-associated SV formation. On the other hand, NHR breakpoint regions appear to have lower DNA stability and higher flexibility, features that may increase the chance of double-strand breaks in general. Overall, our analysis reveals formational biases underlying SV formation and conforms to the fact that NAHR is driven by recombination between repeat sequences, whereas NHR is likely driven by DNA repair and replication errors.

By applying BreakSeq on a large scale, we envisage that it could be used for genotyping and determining SV allele frequencies. In fact, it should be possible to put each of the breakpoint sequences in our library directly onto a commercially available SNP chip, which could be used to precisely assess SV genotypes simultaneously with all of the SNPs in an individual. (This should add only a small number of probes to the ∼1 M probes already on commercial chips.)

Lastly, we note that as our approach depends on current SV lists, it is inevitably affected by their existing biases owing to presently applied technologies. Likely biases include the difficulty in mapping insertions relative to the reference genome and in ascertaining SVs in repetitive regions, for example, segmentally duplicated sequences. We anticipate that in the near future, as technologies advance in terms of read-lengths, inherent biases against repeat-rich sequences will be further reduced and the mapping of SVs onto our junction library will further improve, making it essentially comparable to SNP genotyping. In this regard, as thousands of human genomes will be sequenced in the coming years, there will be a huge demand for reliable and accurate SV mapping and SV genotyping.

4.4 Methods

4.4.1 Data preparation

Our initial breakpoint library altogether represented 1,961 SVs identified at high precision based on the Nationl Center for Biotechnology Information (NCBI) build 36 of the human genome. It was compiled from eight different published sources based on paired-end mapping [11, 15], fosmid-paired-end sequencing [45, 80], Sanger capillary sequencing [48], resequencing of an individual human genome using second-generation sequencing [89], DNA resequencing traces for SNP discovery projects (support by at least two reads was required for an SV to be included in our data set) [90], and high-resolution array-based comparative genomic hybridization [91]. For the 253 SVs identified through fosmid-paired-end sequencing [45, 80], 387 published sequenced clones originally used to identify SVs in NCBI build 35 were realigned to the NCBI build 36 human genome before inclusion in the library. A split-read analysis was then carried out using BLAT to infer the breakpoints of the events. For the 98 SVs from resequencing traces [90], the liftover tool available at the UCSC genome browser (`http://genome.ucsc.edu/`) was used to convert the breakpoint coordinates from human NCBI build 35 to build 36. All SVs in our analysis were between 1 kb and 1 Mb in length (that is, we removed events >1 Mb, reasoning that they may be lower in confidence). After accounting for redundancy, our standardized breakpoint library consisted of 1,889 SVs that were used in all subsequent calculations and analyses.

4.4.2 SV mechanism classification pipeline

Four major steps were involved in our procedure to classify SV formation mechanisms. First, SVs were examined for extensive coverage by tandem repeats and regions of low complexity (here, low-complexity DNA refers to micro-satellite DNA, polypurine/polypyrimidine stretches, and regions of extremely high AT or GC content, as defined by the RepeatMasker program; `http://www.repeatmasker.org/`) to identify instances of expansion or contraction of VNTRs. Second, ± 100-bp flanking sequences derived from both breakpoint junctions were aligned against each other to scan for blocks of extensive homology. SVs were classified as 'high-confidence NAHR' if the homologous blocks had a minimum sequence identity of 85%, a minimum length of 50 bp for the identical sequences, a maximum offset of 20 bp between the homologous blocks, correct orientations and covered the breakpoints. SVs displaying at least three but not all of the above criteria were classified as 'extended NAHR'. Third, SVs aligning to known interspersed mobile elements carrying the common diagnostic features of corresponding transposable elements, that is, target site duplications and poly-A tracts [95], were classified as 'high-confidence TEIs'. Events missing one or more of the diagnostic features were classified as 'extended TEIs'. TEIs were furthered categorized as single transposable element insertions (STEIs) if a single element was involved and multiple transposable element insertions (MTEIs) if multiple elements appeared to be involved. Furthermore, full-length TEIs were discriminated from transposable element fragments and transposable element subfamilies were also recorded. Through identification of spliced protein-coding gene sequences and TEI-diagnostic features, processed pseudogenes likely inserted via a TEI-associated mechanism were also identified. Finally, SVs lacking signatures of any of the above diagnostic sequence features were classified as NHR events.

4.4.3 Sensitivity analysis for the SV mechanism classification

Sensitivity analysis was performed on five key parameters used in the mechanism classification pipeline. Classification results were examined as each parameter was varied over a large range while fixing the other parameters at default values. First, the cutoff for the length of homologous blocks in the flanking sequences alignment for classifying NAHR events (NAHRhomolen) was varied from 10 to 150 bp with a step size of 10 bp. Second, the cutoff for the percentage identity of homologous blocks in the flanking sequences alignment for classifying NAHR events (NAHRpct) was varied from 70 to 100% with a step size of 1%. Third, the cutoff for the coverage of VNTR regions in the SV was varied from 0 to 100% with a step size of 5%. Fourth, the window size used to examine the consistency of the transposable element boundary with a breakpoint for classifying STEI and MTEI events (TEIwin) was varied from 10 to 400 bp with a step size of 10 bp. Finally, the gap size used to examine whether adjacent transposable elements can be joined for classifying MTEI events (TEIgap) was varied from 0 to 300 bp with a step size of 10 bp. Default values for NAHRhomolen, NAHRpct, VNTRcutoff, TEIwin and TEIgap used in the pipeline were 50, 85, 50, 200 and 150, respectively.

4.4.4 Analysis of ancestral state

For a 'deletion' relative to the reference genome, a ± 500-bp flanking sequence at each breakpoint was extracted to obtain two sequences of 1,000 bp representing both the left (A) and right (B) breakpoint junction sequences. Then a 1,000-bp junction sequence at the breakpoint of the alternative allele, representing 500 bp upstream and downstream of the left and right breakpoints, respectively (C), was also extracted. If C aligned onto a nonhuman primate genome (that is, a potential ancestral genomic locus) at high-quality and with better length and sequence identity (represented by the BLAT score) than A and B, then the event was rectified as an insertion relative to the ancestral genome. Conversely, for an 'insertion' relative to the reference genome, the A, B (alternative allele) and C (reference allele) junction sequences of the event were extracted. If A and B both displayed an alignment better than C onto a nonhuman primate genome, the event was rectified as a deletion relative to the ancestral genome.

All the alignments were performed using BLAT on the chimpanzee (panTro2), macaque (rheMac2), and orangutan (ponAbe2) genomes, the sequences of which were downloaded from the UCSC genome browser (`http://genome.ucsc.edu/`). The Net alignments [109, 110] from UCSC were also downloaded and the top level was chosen to verify that the alignment of the junction sequences were in the syntenic regions of the corresponding SVs. Because all the primate ancestral genomes are highly similar, the alignment identity and coverage were required to be >90%. Furthermore, the length ratio of target versus query was required not to exceed a deviation of 10%.

SVs were classified as 'rectifiable' if unambiguous high-quality alignments to putative ancestral regions could be constructed in any nonhuman primate genome. Particularly, an SV was classified as 'rectified' if its state was changed from its original state to another after the analysis (from deletion to insertion, or vice versa). The state of each SV was then assigned based on the closest nonhuman primate genome (e.g., from chimpanzee to orangutan and to macaque) in which a corresponding syntenic region existed. SVs were considered as 'consistently rectifiable' if they were rectified to the same state with no inconsistent ancestral assignment inferred.

4.4.5 Insertion trace

After rectification based on the ancestral state analysis, all insertions that were consistently rectifiable were aligned onto the human reference genome to scan for the presumable origin of the inserted sequences. Because the inserted sequence of an event rectified from a deletion is already present in the reference genome, any alignments overlapping with >50% of the SV region were discarded and the next best match was chosen. BLAT alignments tracing inserted sequences were required to have a sequence identity >90%.

4.4.6 Enrichment calculation

To calculate the enrichment and P-value for each feature and repeat association with breakpoints, a nonparametric randomization test based on sampling was employed. For the observed samples, the exact coordinates of the breakpoints were taken for location-dependent computation and sequences flanking the breakpoints were extracted for sequence-dependent computation. A random global background was generated by randomly sampling a set of coordinates, or sequences with the same length, of the same amount from the reference genome (build 36). Similarly, a local background was generated by randomly sampling in a 10-kb window at the breakpoints. The sampling was repeated 1,000 times with replacement and the observed statistic of the breakpoints was tested against the sampling distribution based on the whole genome. The enrichment value was calculated by comparing the observed statistic over the mean of the statistics of the samplings. Then, the P-value of the enrichment was calculated by counting the number of samplings that yielded a statistic as extreme as, or more extreme than, the observed one. The enrichment was reported as significant for any $P < 0.05$.

4.4.7 Correlation of chromosomal landmarks

Distance to telomeres was calculated from the midpoint of an SV to the end of the chromosome in the same arm. Distances to centromeres and pericentromeric gaps were calculated from the midpoint of an SV to the closest centromeric or pericentromeric gap boundary on the same chromosome. Distance to the closest synteny block boundary was calculated by computing the distance from each breakpoint to the closest synteny block boundary and then taking the average for the two breakpoints. Synteny block boundaries were taken from the human-chimpanzee Net alignment file [109, 110] available at the UCSC genome browser and the 'gap' type was excluded from the analysis. A Wilcoxon rank sum test was then done to compare the distance measurements of different formation mechanisms in a pair-wise fashion, followed by a correction for multiple hypothesis testing using the Holm method.

4.4.8 Feature computation

We considered the following features at SV breakpoints in our analysis: GC content, helix stability and DNA flexibility. All features were computed for sequences within 50 bp of the breakpoints or randomly extracted from the genome. GC content was calculated by computing the percentage of guanine and cytosine nucleotides over the given length of the sequence. Helix stability of the DNA duplex was predicted by calculating the average of the dissociation free energy of each overlapping dinucleotide [99]. Similarly, DNA flexibility was estimated by calculating the average of the twist angle among all overlapping dinucleotides [100]. To observe the change of the DNA flexibility and helix stability around a breakpoint, values at each nucleotide were smoothed using a sliding window of 50 bp, which was slid across an interval of 1 kb centered on the breakpoint.

4.4.9 Repeat association

The association of repeat elements and pseudogenes was calculated by intersecting the relevant data sets. Each element was overlapped with a breakpoint and the average number of overlapping elements for all the input breakpoints was calculated. Repeat elements in the human genome build 36 were downloaded from the RepeatMasker track of the UCSC genome browser (March 2006 assembly). Only the elements annotated with repeat classes SINE and LINE were included in this analysis. In total, there were 1,783,897 SINE elements and 1,407,547 LINE elements of which 1,193,509 were Alu elements and 927,909 were L1 elements, respectively. For the pseudogene analysis, we used PseudoPipe [111] to identify pseudogenes in the genome based on the protein annotations in the Ensembl database (release 48). This analysis involved 2,454 duplicated pseudogenes and 10,999 processed pseudogenes.

4.4.10 Motif discovery

MEME was used to discover sequence motifs near SV breakpoints and to generate position weight matrices (PWMs) for significantly enriched motifs. The input data to MEME were sequences of 200 bp centered on the breakpoints. Motif width was allowed to range from 6 bp to 12 bp. For SVs classified as NAHR-mediated we also looked for an overrepresentation of a previously described sequence motif specific to recombination hotspots [103]. The recombination-hotspot motif was converted into a PWM by considering the average genomic frequencies of the four bases ACGT (0.295, 0.205, 0.205, 0.295) and by adding pseudocounts of 1. After identifying the motifs, MAST was applied to search for a motif match in the original set and the global background set. The P-value cutoff for each motif match was $P < 0.0001$ and a randomization test was performed as described above to calculate the enrichment P-values for each motif.

4.4.11 Microhomology enrichment analysis

The lengths of the microhomology sequences at the breakpoints of NHR-mediated events were compared with the local background and a theoretical distribution. The theoretical expectation was calculated by assuming independence between genomic positions and a uniform distribution of the four nucleotides (ATCG) in the genome. The formula $P \times (1-P)^2 \times (i+1)$ was used to calculate the probability of observing homology of a specific length, where i is the length of homology and P is the probability of observing the same pair of nucleotides at the given genomic positions (that is, $P = p(A)^2 + p(T)^2 + p(C)^2 + p(G)^2$ and $p(A, C, G, T) = (0.295, 0.205, 0.205, 0.295)$ were estimated from the local background). A one-sided Kolmogorov-Smirnov test (KS-test) was performed to test the enrichment of microhomologies in NHR compared to the local background. The size of the effect was calculated as the fold enrichment of microhomology stretches between NHR and the background.

4.4.12 Mapping SVs with a junction library

The breakpoint junction mapping approach that we developed works as follows. The junction library for SV mapping is created by joining 30 bp flanking sequences on each side of a breakpoint. A deletion event is represented with a single junction sequence in the library, while an insertion has both a left and right junction sequence corresponding to each of its breakpoints. DNA reads from personal genomes are aligned against the junction library. Reads are required to overlap a breakpoint by at least 10 bp on each side. All successfully mapped reads are then aligned against the reference genome. Only those reads that do not map onto the reference genome are labeled as ‘unique’ in the personal genome; the other reads are labeled as ‘nonunique’. A short-read aligner, Bowtie [112], is used to perform all the alignments (allowing for two mismatches). To score the SV candidates on the basis of supportive hits, the following formula is used:

$$Si = \max(0, \log_2 Ti - \log_2 Ri)$$

where Si is the score representing the effective number of hits (supportive hits) in $\log_2$ scale for SV i, with unique and nonunique hits denoted as Ti and Ri respectively. If Ti or Ri is 0, the log term is replaced by 0. A score of 1 thus indicates 2 supportive hits, whereas scores >2 (high-support) indicate the presence of >4 supportive hits.

The mapping process showed a linear time complexity in practice. On average, it required 8 h to run our junction-mapping program (open-sourced and available for download at `http://sv.gersteinlab.org/breakseq)` against a sequenced genome at 40× physical coverage on a 3GHz quad-core computer node with 16GB physical memory.

4.4.13 Intersection of the breakpoint junction library with RefSeq genes

RefSeq gene annotations were downloaded from the UCSC Genome Browser. Intersection of the SVs in our breakpoint junction library and RefSeq genes were found by comparing the start- and end- coordinates of the two datasets. For insertion events whose inserted sequences could be traced, the positions from which the insertions were derived were compared to the RefSeq gene annotations. In particular, 60 out of 146 NAHR deletions and 193 out of 580 NHR deletions intersected with annotated exons from RefSeq genes. Insertions were also found to have an impact on coding regions, with 19 out of 51 NAHR insertions and 11 out of 30 NHR insertions intersecting with the exons. These included cases where exons at the insertion site were altered by the insertion event (19 NAHRs and 7 NHRs) and where the inserted sequence was itself derived from exonic DNA (3 NAHRs and 6 NHRs).

4.4.14 PCR validation

We tested by PCR validation 24 insertion and 33 deletion calls predicted in NA12891 relative to the reference genome. Specifically, we designed PCR primers as previously described [11] and amplified the predicted nonreference SV alleles. For the PCR, 10ng of

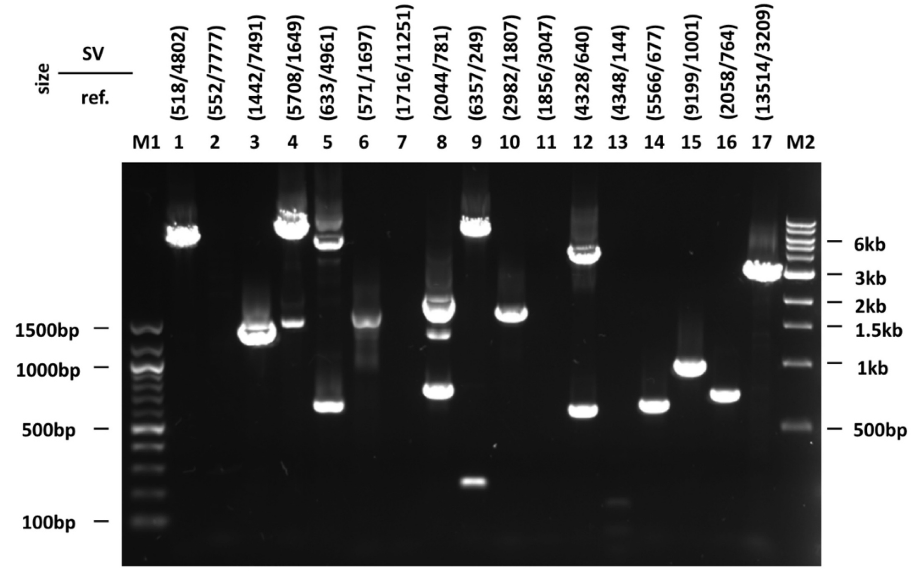


Figure 4.8: Additional PCR validation for predicted SVs. The figure displays 17 additional genomic regions on which PCR validations were carried out (expected band sizes for the reference and non-reference SV alleles are shown at the top). SVs mapped in NA12891 were analyzed by PCR using SV flanking primers. The difference in size of the products for the reference and non-reference alleles confirmed the presence of the SVs for all loci except 2, 3, 5, 7 and 11. M1 is a 100bp marker and M2 is a 1kb marker.

genomic DNA (Coriell Institute were used with the SequalPrep Long PCR Kit (Invitrogen) in 20 gl volumes using the following PCR conditions in a C1000 thermocycler (BioRad): 94 °C for 3 min, followed by 10 cycles of 94 °C for 10 s, 60 °C for 30 s and 68 °C for 10 min and 25 cycles of 94 °C for 10 s, 56 °C for 30 s and 68 °C for 10 min (+10 s/cycle), followed by a final cycle of 72 °C for 10 min. Some of the reactions that failed with the SequalPrep enzyme were amplified with the LongAmp Taq DNA Polymerase (NEB) or the iProof High Fidelity DNA Polymerase (Biorad). PCR products were analyzed on a 1% agarose gel stained with Sybr Safe Dye (Invitrogen). Marker M1 was a 100-bp ladder whereas M2 corresponded to a 1-kb ladder (500, 1,000, 1,500, 2,000, 3,000, etc) (NEB).

Chapter 5

Conclusion

In this thesis, we demonstrate that with the tools, statistics and ontology provided by Pseudofam, we can analyze pseudogenes from a different perspective and integrate pseudogene families with other related datasets to better understand the genome remodeling processes. For example, both pseudogenes and SDs represent duplicated regions of the genome; hence, by analyzing the presence of pseudogenes located in SDs, some precious clues about the generation processes of pseudogene and SD formation can be obtained. In particular, comparing the substitution rates of a pseudogene and its parent gene with their enclosing SD segments shall reveal details about their origin and time of formation.

We then present evidence for different formation mechanisms of SVs in the human genome. Our result suggests that currently occurring copy number variants appear to follow a pattern somewhat similar to young segmental duplications and decidedly different from older segmental duplications. We show a shift from a prevalence of Alu-mediated generation of old SDs toward other mechanisms for more recent SDs. The weakness of association of CNVs with Alu elements can be viewed as the natural extension of this trend, as CNVs are usually ‘very young’ SDs. This trend is consistent with the current models that propose a decrease of Alu activity after the ‘Alu burst’ ∼40 Mya. Finally, we present results suggesting that while some CNVs are formed through NAHR, a large fraction of them are formed through NHEJ. These trends are present in the large amounts of low-resolution data as well as found confirmed in the substantial number of sequenced breakpoints.

To pinpoint the effects of SV and to characterize them, we thus extend our analysis to a large-scale study of nucleotide-resolution SVs. Our BreakSeq approach uses a library of previously discovered SVs that have breakpoint information to help researchers rapidly scan for and characterize SVs in a newly sequenced personal genome. Furthermore, it has been implemented as a computational pipeline that not only identifies SVs in a personal genome, but also deduces the formation mechanism and ancestral state of an SV. Overall, our analysis reveals the formational biases underlying SV formation and conforms to the fact that NAHR is driven by recombination between repeat sequences, whereas NHR is likely driven by DNA repair and replication errors. By applying BreakSeq on a large scale, we envisage that it could be used for genotyping and determining SV allele frequencies. In fact, it should be possible to put each of the breakpoint sequences in our library directly onto a commercially available SNP chip, which could be used to precisely assess SV genotypes simultaneously with all of the SNPs in an individual.

In the future, we will look into transposons that contribute a substantial part of the SVs in the human genome. Transposons are DNA sequences that move around to different positions within the genome, which consist of DNA transposons and retrotransposons. Though $\sim 45\%$ of the human genome are occupied by transposons or alike, there are less than 0.05% still active. It is estimated that about 35–50 subfamilies of Alu, L1, and SVA elements remain actively mobile, which not only produce genetic diversity but also cause diseases by gene disruption. However, there is still no systematic and efficient way to identify active mobile elements and their variation among individuals. To this end, we aim to identify active mobile elements in a genome-wide fashion. For example, a microarray with probes that represent both ends of known L1 elements can be used to capture candidates for paired-end sequencing and subsequent alignment analyses to deduce which are still active.

Chapter 6

Appendix: Automated Motif Analysis for Predicting Targets of Modular Protein Domains

6.1 Abstract

Many protein interactions, especially those involved in signaling, involve short linear motifs consisting of 5–10 amino acid residues that interact with modular protein domains such as the SH3 binding domains and the kinase catalytic domains. One straightforward way of identifying these interactions is by scanning for matches to the motif against all the sequences in a target proteome. However, predicting domain targets by motif sequence alone without considering other genomic and structural information has been shown to be lacking in accuracy. Thus, we developed an efficient search algorithm to scan the target proteome for potential domain targets and to increase the accuracy of each hit by integrating a variety of pre-computed features, such as conservation, surface propensity, and disorder. The integration is performed using naive Bayes and a training set of validated experiments. By integrating a variety of biologically relevant features to predict domain targets, we demonstrated a notably improved prediction of modular protein domain targets.

Combined with emerging high-resolution data of domain specificities, we believe that our approach can assist in the reconstruction of many signaling pathways.

6.2 Background

Important protein-protein interactions (e.g., those involved in signal transduction) are often mediated by modular protein domains [113]. These domains often work in a mix-and-match fashion, thereby acting as the building blocks of signaling pathways [114]. Examples include the SH3 and WW domains that bind proline-rich motifs [115], and the serine/threonine kinase domain that specifically phosphorylates the hydroxyl group of serine and threonine [116]. Throughout we will refer to these collectively as ‘domains’. Since these kinds of domains play an important role in the assembly, regulatory and signaling activities of the cell [115, 117, 118], accurate prediction of their targets is crucial to understanding many biological pathways [119, 120]. As a result, various techniques have been developed to predict domain targets and to enhance the prediction. Earlier studies have tried to use consensus sequences from phage display experiments to predict the targets of peptide-binding domains [121]. Also, a modern peptide library screening approach, which is commonly used to determine phosphorylation motifs for kinases, has shown to have high accuracy in determining domain specificity [122]. Both approaches have in common that they identify the specificity of each domain in a position-specific manner, yielding a Position Specific Scoring Matrix (PSSM; also known as Position Weight Matrix, PWM). Furthermore, many studies have demonstrated various ways to improve prediction performance using genomic information. For instance, comparative genomics and secondary structure information have been used to increase the performance of SH3 target prediction [123, 124]. Nevertheless, to date the prediction of biologically relevant targets of these domains has yet to be addressed in an automated and integrated fashion. To this end, we present an automated process, which integrates comparative genomic (i.e., sequence conservation) and structural genomic (i.e., surface propensity and peptide disorder) data with traditional profile scanning method to predict domain targets based on experimental screening result (e.g. peptide library screening) or their derived PSSMs. The process is fully automated and implemented as an online server. The implementation is open-source and also available for download at `http://motips.gersteinlab.org`.

6.3 Results and Discussion

6.3.1 An Automated Pipeline Process

Our approach first converts the input data into a PSSM and then normalizes it. Secondly, it scans the target proteome by using the normalized PSSM and generates a hit list of potential domain targets. Following the motif scanning, it computes the conservation score, solvent accessibility score, and disorder score for each motif hit based on the pre-computed scores for each protein residue. It then integrates these genomic features with the motif matching scores and the number of hits per protein by naive Bayes to predict the optimal targets based upon a validated training set. Lastly, it sorts the motif hits by their likelihood of having interaction with the domain and consolidates them into unique protein hits.

6.3.2 Data Conversion and Normalization

A number of experimental approaches, such as phage display and peptide library screening (Figure 6.1), have been developed to identify domain binding and phosphorylation targets. However, data from different experiments result in different formats that always complicate the data analysis process. To keep the process consistent and standardized, these data are converted into PSSM followed by normalization (for supported input formats, see System Implementation and Availability).

Our approach employs two different ways to normalize the input data. The first approach is designed for signal data from experiments such as from peptide library screening. It normalizes the signal score for each amino acid at each position by the following equation

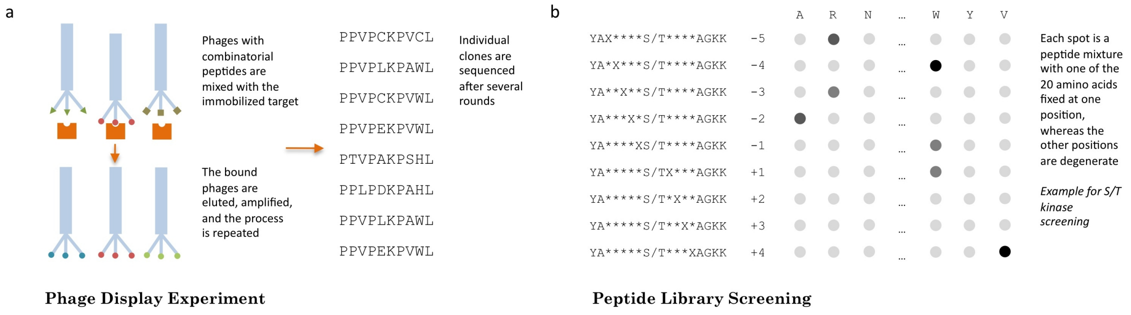

a
Phages with combinatorial peptides are mixed with the immobilized target
The bound phages are eluted, amplified, and the process is repeated
PPVPCKPVCL
PPVPLKPAWL
PPVPCKPVWL
PPVPEKPVWL
PTVPAKPSHL
PPLPDKPAHL
PPVPLKPAWL
PPVPEKPVWL
Individual clones are sequenced after several rounds
Phage Display Experiment
b
A
R
N
...
W
Y
V
YAX****S/T****AGKK -5
YA*X***S/T****AGKK -4
YA**X**S/T****AGKK -3
YA***X*S/T****AGKK -2
YA****XS/T****AGKK -1
YA*****S/TX***AGKK +1
YA*****S/T*X**AGKK +2
YA*****S/T**X*AGKK +3
YA*****S/T***XAGKK +4
Each spot is a peptide mixture with one of the 20 amino acids fixed at one position, whereas the other positions are degenerate
Example for S/T kinase screening
Peptide Library Screening

Figure 6.1: Experiments for motif identification. a) The phage display experiment identifies potential target peptides of short sequences, and b) the peptide library screening measures the binding specificity at position level. The resulting experimental data of such experiments can be converted into a Position Specific Scoring Matrix (PSSM).

$$Z_{ca} = \frac{S_{ca}}{\sum_{i}^{m} S_{ci}} \times m \tag{6.1}$$

where Z_{ca} is the normalized score for amino acid a at position c, which has a signal score S_{ca}, and m is the total number of amino acids. Equation (6.1) thus computes the weight for each amino acid at each position and scales it up by the total number of amino acids. However, to consider the known specificity for domains such as the serine/threonine kinase domain, which have fixed amino acid targets (e.g., serine and threonine) at a certain position in the binding motif, a score of 0 is automatically assigned to every other amino acid that is not expected at that position. To indicate the slight probability of observing the fixed amino acids at other positions, a pseudo-count of 1 is assigned to each of them at these non-specific positions.

The second way of normalization is designed for peptide data from experiments such as from phage display experiment. Our approach employs the pseudo-count method based on substitution probabilities to complement the incomplete or imperfect representation of a position in the original peptide data [125]. Pseudo-counts are needed since this kind of experiments significantly undersample sequence space, thereby severely penalizing rare residues. It calculates the probability pca of amino acid a at position c by equation (6.2) as follows

$$P_{ca} = \frac{n_{ca} + b_{ca}}{N_c + B_c} \tag{6.2}$$

$$B_c = \psi \times R_c \tag{6.3}$$

where n_{ca} and b_{ca} are the count and pseudo-count for amino acid a at position c, while N_c and B_c are the total count and pseudo-count for all amino acids. The total pseudo-count B_c is calculated from equation (6.3) with ψ as an empirically chosen positive number (default to 5) and R_c as the unique count for all amino acids at position c. Taking different substitution probabilities of different amino acids into consideration, substitution matrixes such as the BLOSUM 62 [126, 127] and McLachlan [128] matrixes are used to calculate pseudo-count bca by equation (6.4) shown as the following

$$b_{ca} = B_c \times \sum_{i}^{m} \frac{n_{ci}}{N_c} \times \frac{q_{ia}}{Q_i}; Q_i = \sum_{i}^{m} q_{ia} \tag{6.4}$$

where q_{ia} is the substitution probability for amino acid a replaced by i, and Q_i is the substitution probability for a replaced by any amino acid. In addition to the pseudo-count method based on substitution probabilities, we also provide alternative pseudo-count methods based on flat counting (adding 1 to all values) and entropy (adding a pseudo-count proportional to the entropy of each position to its corresponding values).

6.3.3 Motif Scanning and Scoring

To scan the target proteome for potential domain targets and to score them, our approach uses a window-sliding method based on a normalized PSSM similar to the method used in Scansite [129, 130]. For each protein in the target proteome, it slides a window of size equivalent to the length of the motif on the peptide sequence by every single amino acid (Figure 6.2). Based on the scoring matrix, the score for each window sequence is calculated by equation (6.5)

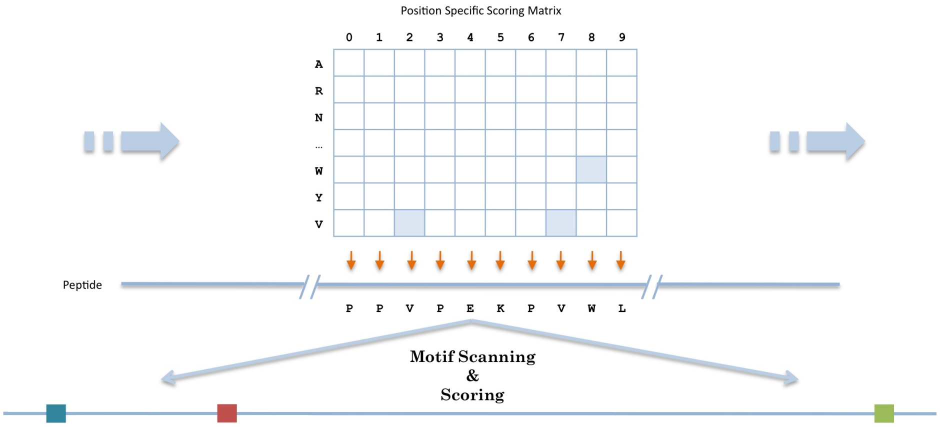


Figure 6.2: Motif scanning and scoring. Identify potential target sites of the domain by sliding a Position Specific Scoring Matrix (PSSM) across the peptides in the proteome and comparing the motif matching scores for each window.

$$E' = \sum_{c}^{l} - \log_2 \left(\frac{S_{ca}}{\sum_{i}^{m} S_{ci}} \right) \tag{6.5}$$

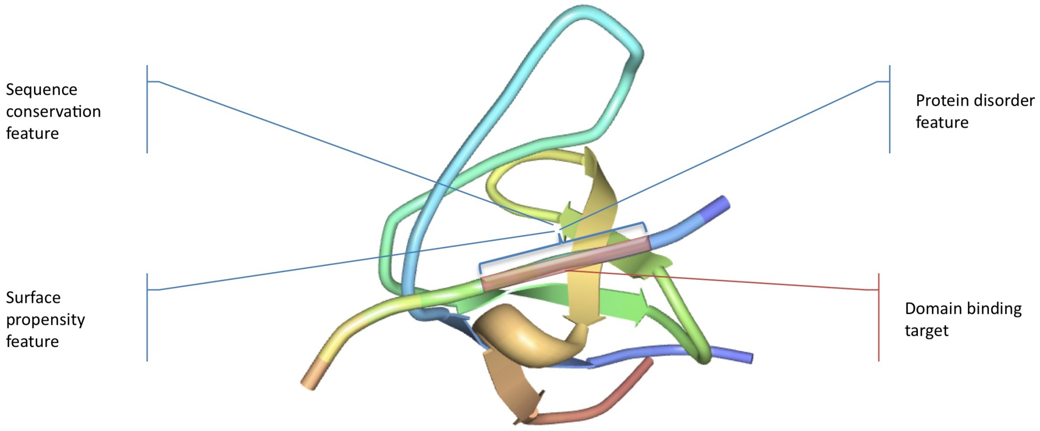
where l is the length of the motif and S_{ca} is the score for the amino acid a at position c in the window sequence. This equation is also used to calculate an optimal score of the motif where S_{ca} is the maximum score at position c in the scoring matrix. Then the final normalized score E for the window sequence is calculated by equation (6.6)

$$E = \frac{E'_{sequence} - E'_{optimal}}{E'_{optimal}} \tag{6.6}$$

To improve the efficiency of the scanning algorithm, each motif hit is compared immediately to a sorted hit list of fixed size (currently 2,000 hits) and will only be retained if it has a more significant score than the least significant one in the list.

6.3.4 Structural Features and Scoring

Although a profile-matching scan could identify possible domain targets, it does not take into account the structural information of the target sequences that are also related to



Yeast Sho1 SH3 domain with Pbs2 peptide (Image source: PDB)

Figure 6.3: A peptide-binding domain example. A peptide-binding domain, such as the SH3 domain, recognizes the binding site on a peptide which exhibits certain structural and conservation features including surface propensity, protein disorder, and sequence conservation.

protein-protein interactions. For instances, sequences exposed on the surface should be more accessible than those that are buried; sequences that are unfolded should be more easily bound than those that are folded; and structures that are highly conserved among close species could have more biological significance. Taking these factors into account, our approach includes three major structural and conservation features in the prediction, which are surface propensity, protein disorder, and sequence conservation, to complement the motif scanning score (Figure 6.3).

The degree of surface propensity of a given sequence is measured by its relative solvent accessibility, which represents the extent of residue solvent exposure. It is predicted by a protein structure prediction program, SABLE, which uses a neural network-based regression algorithm [131]. To measure the disorder of the sequence, DISOPRED, a neural networks and PSI-BLAST-based approach is used to estimate the probability of the region being disordered [132, 133]. For measuring the conservation of the sequence structure, orthologs of the sequence are identified using INPARANOID [134]. Following the ortholog identification, the sequences in the orthologous groups are aligned with MUSCLE [135] and a conservation score for each position in the sequence is estimated by its entropy using AL2CO [136].

For each protein in each proteome being studied, the solvent accessibility, disorder and conservation scores are pre-computed for each residue. As a result, the scores for the motif hits could be calculated in a timely manner.

6.3.5 Feature Integration and Target Prediction

In addition to calculating the structural and conservation scores for each motif hit, the number of hits per protein is also calculated as a feature for the hit. Our approach then applies a Bayesian learning algorithm to integrate all the aforementioned features, including the motif scanning score, solvent accessibility score, disorder score, conservation score, and number of hits per protein, to predict potential domain targets. Because of the simplicity and efficiency of the naive Bayes model, it is employed to build a classifier based on a validated training set under the assumption of independence of the features. In particular, the default models (i.e., the SH3 model based on Sho1 and the S/T kinase model based on Prk1) used a number of experimentally determined interaction pairs [137, 138] as the gold-standard positives to train the algorithm. Moreover, a set of paired proteins in which each pair was annotated to always localize to two different compartments (for example, nucleus only and cytoplasm only in the Gene Ontology) in the cell was selected as the gold-standard negatives. The conditional probability can then be calculated from the given features based on equation (6.7)

$$p(I|F_1, \ldots, F_n) \propto p(I) \prod_{i=1}^{n} p(F_i|I) \tag{6.7}$$

where I is the class variable (i.e., interaction or non-interaction), F is the feature such as the motif scanning score, and n is the total number of features. To assess the independence of the features, pair-wise correlation coefficients were calculated. The results showed the pair-wise correlation coefficients have an average of 0.23 for the SH3 model and 0.18 for the

S/T kinase model, indicating the features are to a large extent independent. Furthermore, since the independency assumption is not harmful for data pre-processed with Principal Component Analysis (PCA) [139], we performed PCA to transform the possibly correlated features into uncorrelated features. The first three principal components were chosen to build a naive Bayes model followed by a stratified 10-fold cross-validation. The Area Under Curve (AUC; 89.1 for the SH3 model and 75.9% for the S/T kinase model) of the Receiver Operating Curve (ROC) resulting from the PCA transformation was then compared to the AUC (91.8% for the SH3 model and 78.6% for the S/T kinase model) without the PCA. No significant deviation of performance was observed between the predictions without PCA and those with PCA, indicating no strong dependency among the original features.

Finally, the motif hits from the domain of interest are classified under the selected model and sorted by their likelihood of having an interaction with the domain. Hits for the same protein are consolidated into one single hit represented by the most likely target. Genomic information that is not used in the prediction, such as protein-protein interaction data, localization data and phosphorylome data, could also be integrated easily with the tab-delimited hit list for further analysis while phosphorylation prediction data from mass spectrometry experiments can be used as cross-validation.

6.3.6 Prediction Performance

To assess the prediction performance of our approach, we benchmarked with two existing methods: 1. the Eukaryotic Linear Motif (ELM) database [140], which predicts functional sites in eukaryotic proteins by patterns with context-based rules and logical filters such as the structure filter; and 2. the Scansite method [129], which uses a motif profile-scoring approach to predict sites within proteins that are likely to be phosphorylated or bind to domains. Based on the SH3 interactome data [137], a model for the SH3 domain was trained with the Sho1 interactions. Then, we performed our prediction, requiring a likelihood value above 0.9, on 10 other different SH3 proteins by using the aforementioned model. We compared our results with the predictions from the ELM database (data retrieved from

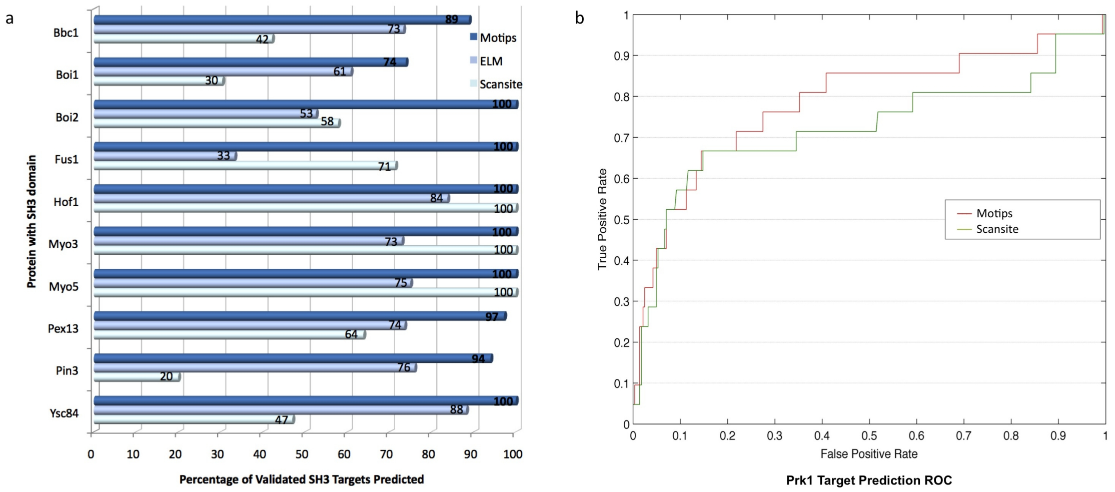


Figure 6.4: Targets prediction performance. a) The benchmark of SH3 target prediction based on the validated targets for 10 different SH3 proteins, and b) the Receiver Operating Curve (ROC) comparing the prediction performance for the Prk1 kinase targets.

the web server using a Python program for 5 different SH3 ligands available on the server) and from the Scansite scanning (which requires a score not more than 3 fold of the optimal score). Our results (Figure 6.4) show that on average our prediction has a 49% increase in accuracy in predicting the validated targets of the SH3 proteins when compared to the ELM prediction. When compared to the profile-scoring method of Scansite, our prediction is almost twice as accurate (90% higher). In addition to predicting SH3 targets, our approach was employed to predict Prk1 phosphorylation sites [138]. A stratified 10-fold cross-validation has shown a performance increase (Figure 6.4; 79% AUC in a ROC curve) when compared to the profile-scoring method (72% AUC).

6.3.7 System Implementation and Availability

The motif analyzing process mentioned above is implemented as an online server, which allows researchers to upload their experimental data representing the motifs of the domains and to predict the targets.

Our pipeline supports various input data formats. For specific analysis software, it currently supports the Gene Pix Result format (`http://www.moleculardevices.com/pages/`

software/gn_genepix_file_formats.html#gpr) that is usually used for peptide library screening data, and the BRAIN project's peptide format (http://www.baderlab.com/Software/BRAIN/PeptideFile) that is usually used for phage display experiments. For general purposes, it supports the FASTA format (i.e., a set of peptides with the same length that represent the possible interacting sites) and the Nx20 format (i.e., a tab-delimited format that represents the positional scores of a motif profile with the first row labeled with the amino acid residues and the subsequent rows as the different positions). The pipeline currently has a compilation of 20 proteomes consisting of 14 yeast proteomes (*S. cerevisiae, C. albicans, D. hansenii, C. glabrata, K. lactis, N. crassa, S. bayanus, S. castelli, S. kluyveri, S. kudriavzevii, S. mikatae, S. paradoxus, S. pombe, Y. lipolytica*), 2 worm proteomes (*C. briggsae, C. elegans*), and 4 mammalian proteomes (*C. familiaris, P. troglodytes, M. musculus, H. sapiens*).

The feature scores were pre-computed and the default prediction models, which could be replaced by a user-defined training set (a tab-delimited file with the gene on the first column and a logical value on the second indicating the interaction), were also built. Moreover, the analyzing process is implemented as an asynchronous multi-threading pipeline process so the prediction results can be delivered to the users via email offline, in addition to being displayed online. Furthermore, the entire system is built using the Java programming language under a Model View Controller architecture in which the analysis process is implemented as a standalone open-sourced program. Therefore, the process could be customized by researchers and executed in command line on multiple platforms. The naive Bayes classification is performed using Weka, the open-source Java data mining software [141].

The standalone pipeline and database are available for download at the MOTIPS server at http://motips.gersteinlab.org.

6.4 Conclusions

By integrating a variety of biologically relevant features and using a Bayesian learning algorithm to predict domain targets, our approach has improved the domain binding and phosphorylation target predictions notably compared to using only profile-matching scan. We believe our approach is versatile enough to predict targets of domains of different kinds, and its implementation as an online public server could facilitate researchers in predicting domain targets more accurately.

Bibliography

[1] Collins, F. S., Morgan, M. & Patrinos, A. The Human Genome Project: lessons from large-scale biology. *Science* **300**, 286–90 (2003).

[2] Schuster, S. C. Next-generation sequencing transforms today's biology. *Nat Methods* **5**, 16–8 (2008).

[3] Feuk, L., Carson, A. & Scherer, S. Structural variation in the human genome. *Nat Rev Genet* **7**, 85–97 (2006). 10.1038/nrg1767.

[4] Hurles, M. E., Dermitzakis, E. T. & Tyler-Smith, C. The functional impact of structural variation in humans. *Trends Genet* **24**, 238–45 (2008).

[5] Gonzalez, E. *et al.* The influence of ccl3l1 gene-containing segmental duplications on hiv-1/aids susceptibility. *Science* **307**, 1434–1440 (2005). 10.1126/science.1101160.

[6] Korbel, J. O. *et al.* The genetic architecture of Down syndrome phenotypes revealed by high-resolution analysis of human segmental trisomies. *Proc Natl Acad Sci U S A* **106**, 12031–6 (2009).

[7] Medvedev, P., Stanciu, M. & Brudno, M. Computational methods for discovering structural variation with next-generation sequencing. *Nat Methods* **6**, S13–20 (2009).

[8] Urban, A. E. *et al.* High-resolution mapping of DNA copy alterations in human chromosome 22 using high-density tiling oligonucleotide arrays. *Proc Natl Acad Sci U S A* **103**, 4534–9 (2006).

[9] Wang, L.-Y., Abyzov, A., Korbel, J. O., Snyder, M. & Gerstein, M. MSB: a mean-shift-based approach for the analysis of structural variation in the genome. *Genome Res* **19**, 106–17 (2009).

[10] Korbel, J. O. *et al.* Systematic prediction and validation of breakpoints associated with copy-number variants in the human genome. *Proc Natl Acad Sci U S A* **104**, 10110–5 (2007).

[11] Korbel, J. O. *et al.* Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**, 420–6 (2007).

[12] Korbel, J. O. *et al.* Pemer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biol* **10**, R23 (2009). 10.1186/gb-2009-10-2-r23.

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

[14] Lam, H. Y. K. *et al.* Pseudofam: the pseudogene families database. *Nucleic Acids Res* **37**, D738–43 (2009).

[15] Kim, P. M. *et al.* Analysis of copy number variants and segmental duplications in the human genome: Evidence for a change in the process of formation in recent evolutionary history. *Genome Res* **18**, 1865–74 (2008).

[16] Lam, H. Y. K. *et al.* Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. *Nat Biotechnol* **28**, 47–55 (2010).

[17] Lam, H. Y. K. *et al.* MOTIPS: Automated Motif Analysis for Predicting Targets of Modular Protein Domains. *BMC Bioinformatics* **11**, 243 (2010).

[18] Gerstein, M. & Zheng, D. The real life of pseudogenes. *Sci Am* **295**, 48–55 (2006).

[19] Ortutay, C. & Vihinen, M. PseudoGeneQuest - service for identification of different pseudogene types in the human genome. *BMC Bioinformatics* **9**, 299 (2008).

[20] Yao, A., Charlab, R. & Li, P. Systematic identification of pseudogenes through whole genome expression evidence profiling. *Nucleic Acids Res* **34**, 4477–85 (2006).

[21] Zhang, Z. *et al.* PseudoPipe: an automated pseudogene identification pipeline. *Bioinformatics* **22**, 1437–9 (2006).

[22] Harrison, P. M. & Gerstein, M. Studying genomes through the aeons: protein families, pseudogenes and proteome evolution. *J Mol Biol* **318**, 1155–74 (2002).

[23] Liu, Y., Harrison, P. M., Kunin, V. & Gerstein, M. Comprehensive analysis of pseudogenes in prokaryotes: widespread gene decay and failure of putative horizontally transferred genes. *Genome Biol* **5**, R64 (2004).

[24] Flicek, P. *et al.* Ensembl 2008. *Nucleic Acids Res* **36**, D707–14 (2008).

[25] Stoesser, G. *et al.* The EMBL Nucleotide Sequence Database. *Nucleic Acids Res* **30**, 21–6 (2002).

[26] Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403–10 (1990).

[27] Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389–402 (1997).

[28] Pearson, W. R., Wood, T., Zhang, Z. & Miller, W. Comparison of DNA sequences with protein sequences. *Genomics* **46**, 24–36 (1997).

[29] Durinck, S. *et al.* BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* **21**, 3439–40 (2005).

[30] Bateman, A. *et al.* The Pfam protein families database. *Nucleic Acids Res* **28**, 263–6 (2000).

[31] Finn, R. D. *et al.* Pfam: clans, web tools and services. *Nucleic Acids Res* **34**, D247–51 (2006).

[32] Zhang, Z. & Gerstein, M. Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes. *Nucleic Acids Res* **31**, 5338–48 (2003).

[33] Tam, O. H. *et al.* Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature* **453**, 534–8 (2008).

[34] Doxiadis, G. G. M. *et al.* Reactivation by exon shuffling of a conserved HLA-DR3-like pseudogene segment in a New World primate species. *Proc Natl Acad Sci U S A* **103**, 5864–8 (2006).

[35] Sassi, S. O., Braun, E. L. & Benner, S. A. The evolution of seminal ribonuclease: pseudogene reactivation or multiple gene inactivation events? *Mol Biol Evol* **24**, 1012–24 (2007).

[36] Gruber, T. R. A translation approach to portable ontology specifications. *Knowl Acquis* **5**, 199–220 (1993).

[37] Svensson, O., Arvestad, L. & Lagergren, J. Genome-wide survey for biologically functional pseudogenes. *PLoS Comput Biol* **2**, e46 (2006).

[38] Zheng, D. Asymmetric histone modifications between the original and derived loci of human segmental duplications. *Genome Biol* **9**, R105 (2008).

[39] Eisenberg, E. & Levanon, E. Y. Human housekeeping genes are compact. *Trends Genet* **19**, 362–5 (2003).

[40] Su, A. I. *et al.* Large-scale analysis of the human and mouse transcriptomes. *Proc Natl Acad Sci U S A* **99**, 4465–70 (2002).

[41] Goncalves, I., Duret, L. & Mouchiroud, D. Nature and structure of human genes that generate retropseudogenes. *Genome Res* **10**, 672–8 (2000).

[42] Bailey, J. A. & Eichler, E. E. Primate segmental duplications: crucibles of evolution, diversity and disease. *Nat Rev Genet* **7**, 552–64 (2006).

[43] Iafrate, A. J. *et al.* Detection of large-scale variation in the human genome. *Nat Genet* **36**, 949–51 (2004).

[44] Sebat, J. *et al.* Large-scale copy number polymorphism in the human genome. *Science* **305**, 525–8 (2004).

[45] Tuzun, E. *et al.* Fine-scale structural variation of the human genome. *Nat Genet* **37**, 727–32 (2005).

[46] Freeman, J. L. *et al.* Copy number variation: new insights in genome diversity. *Genome Res* **16**, 949–61 (2006).

[47] Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444**, 444–54 (2006).

[48] Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol* **5**, e254 (2007).

[49] Bailey, J. A. *et al.* Recent segmental duplications in the human genome. *Science* **297**, 1003–7 (2002).

[50] Cheng, Z. *et al.* A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature* **437**, 88–93 (2005).

[51] Korbel, J. O. *et al.* The current excitement about copy-number variation: how it relates to gene duplications and protein families. *Curr Opin Struct Biol* **18**, 366–74 (2008).

[52] Bailey, J. A., Liu, G. & Eichler, E. E. An Alu transposition model for the origin and expansion of human segmental duplications. *Am J Hum Genet* **73**, 823–34 (2003).

[53] Zhou, Y. & Mishra, B. Quantifying the mechanisms for segmental duplications in mammalian genomes by statistical analysis and modeling. *Proc Natl Acad Sci U S A* **102**, 4051–6 (2005).

[54] Sharp, A. J., Cheng, Z. & Eichler, E. E. Structural variation of the human genome. *Annu Rev Genomics Hum Genet* **7**, 407–42 (2006).

[55] Cooper, G. M., Nickerson, D. A. & Eichler, E. E. Mutational and selective effects on copy-number variants in the human genome. *Nat Genet* **39**, S22–9 (2007).

[56] Linardopoulou, E. V. *et al.* Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication. *Nature* **437**, 94–100 (2005).

[57] Conrad, D. F. & Hurles, M. E. The population genetics of structural variation. *Nat Genet* **39**, S30–6 (2007).

[58] Richardson, C., Moynahan, M. E. & Jasin, M. Double-strand break repair by interchromosomal recombination: suppression of chromosomal translocations. *Genes Dev* **12**, 3831–42 (1998).

[59] Bauters, M. *et al.* Nonrecurrent MECP2 duplications mediated by genomic architecture-driven DNA breaks and break-induced replication repair. *Genome Res* **18**, 847–58 (2008).

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

[61] Coe, B. P. *et al.* Resolving the resolution of array CGH. *Genomics* **89**, 647–53 (2007).

[62] Sharp, A. J. *et al.* Segmental duplications and copy-number variation in the human genome. *Am J Hum Genet* **77**, 78–88 (2005).

[63] Albert, R. & Barabási, A.-L. Statistical mechanics of complex networks. *Rev Mod Phys* **74**, 47–97 (2002).

[64] Barabási, A.-L. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509–12 (1999).

[65] Jiang, Z. *et al.* Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution. *Nat Genet* **39**, 1361–8 (2007).

[66] Kazazian, H. H. J. Mobile elements: drivers of genome evolution. *Science* **303**, 1626–32 (2004).

[67] Zhang, Z., Harrison, P. & Gerstein, M. Identification and analysis of over 2000 ribosomal protein pseudogenes in the human genome. *Genome Res* **12**, 1466–82 (2002).

[68] Goidts, V. *et al.* Complex patterns of copy number variation at sites of segmental duplications: an important category of structural variation in the human genome. *Hum Genet* **120**, 270–84 (2006).

[69] Perry, G. H. *et al.* Hotspots for copy number variation in chimpanzees and humans. *Proc Natl Acad Sci U S A* **103**, 8006–11 (2006).

[70] Jurka, J. Evolutionary impact of human Alu repetitive elements. *Curr Opin Genet Dev* **14**, 603–8 (2004).

[71] Jurka, J., Kohany, O., Pavlicek, A., Kapitonov, V. V. & Jurka, M. V. Duplication, coclustering, and selection of human Alu retrotransposons. *Proc Natl Acad Sci U S A* **101**, 1268–72 (2004).

[72] Ugarkovic, D. & Plohl, M. Variation in satellite DNA profiles–causes and effects. *EMBO J* **21**, 5955–9 (2002).

[73] Lieber, M. R., Ma, Y., Pannicke, U. & Schwarz, K. Mechanism and regulation of human non-homologous DNA end-joining. *Nat Rev Mol Cell Biol* **4**, 712–20 (2003).

[74] Nystrom-Lahti, M. *et al.* Founding mutations and Alu-mediated recombination in hereditary colon cancer. *Nat Med* **1**, 1203–6 (1995).

[75] Deininger, P. L. & Batzer, M. A. Alu repeats and human disease. *Mol Genet Metab* **67**, 183–93 (1999).

[76] Khaja, R. *et al.* Genome assembly comparison identifies structural variants in the human genome. *Nat Genet* **38**, 1413–8 (2006).

[77] Consortium, I. H. A haplotype map of the human genome. *Nature* **437**, 1299–320 (2005).

[78] Karro, J. E. *et al.* Pseudogene.org: a comprehensive database and comparison platform for pseudogene annotation. *Nucleic Acids Res* **35**, D55–60 (2007).

[79] Iafrate, A. J. *et al.* Detection of large-scale variation in the human genome. *Nat Genet* **36**, 949–951 (2004). 10.1038/ng1416.

[80] Kidd, J. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56–64 (2008). 10.1038/nature06862.

[81] Turner, D. J. *et al.* Germline rates of de novo meiotic deletions and duplications causing several genomic disorders. *Nat Genet* **40**, 90–5 (2008).

[82] van Ommen, G.-J. Frequency of new copy number variation in humans. *Nat Genet* **37**, 333–334 (2005). 10.1038/ng0405-333.

[83] Sharp, A. J. *et al.* Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. *Nat Genet* **38**, 1038–42 (2006).

[84] McCarroll, S. A. *et al.* Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. *Nat Genet* **40**, 1107–12 (2008).

[85] de Cid, R. *et al.* Deletion of the late cornified envelope LCE3B and LCE3C genes as a susceptibility factor for psoriasis. *Nat Genet* **41**, 211–5 (2009).

[86] Aitman, T. *et al.* Copy number polymorphism in fcgr3 predisposes to glomerulonephritis in rats and humans. *Nature* **439**, 851–855 (2006). 10.1038/nature04489.

[87] Hastings, P., Lupski, J., Rosenberg, S. & Ira, G. Mechanisms of change in gene copy number. *Nat Rev Genet* **10**, 551–564 (2009). 10.1038/nrg2593.

[88] Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001). 10.1038/35057062.

[89] Wheeler, D. *et al.* The complete genome of an individual by massively parallel dna sequencing. *Nature* **452**, 872–876 (2008). 10.1038/nature06884.

[90] Mills, R. E. *et al.* An initial map of insertion and deletion (indel) variation in the human genome. *Genome Res* **16**, 1182–1190 (2006). 10.1101/gr.4565806.

[91] Perry, G. H. *et al.* The fine-scale and complex architecture of human copy-number variation. *Am J Hum Genet* **82**, 685–695 (2008). 10.1016/j.ajhg.2007.12.010.

[92] Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–9 (2008).

[93] Wang, J. *et al.* The diploid genome sequence of an asian individual. *Nature* **456**, 60–65 (2008). 10.1038/nature07484.

[94] Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**, 57–63 (2009).

[95] Mills, R. E., Bennett, E. A., Iskow, R. C. & Devine, S. E. Which transposable elements are active in the human genome? *Trends Genet* **23**, 183–91 (2007).

[96] Xing, J. *et al.* Mobile elements create structural variation: analysis of a complete human genome. *Genome Res* **19**, 1516 (2009). 10.1101/gr.091827.109.

[97] Myers, S., Bottolo, L., Freeman, C., McVean, G. & Donnelly, P. A fine-scale map of recombination rates and hotspots across the human genome. *Science* **310**, 321–4 (2005).

[98] Meunier, J. & Duret, L. Recombination drives the evolution of GC-content in the human genome. *Mol Biol Evol* **21**, 984–90 (2004).

[99] Breslauer, K. J., Frank, R., Blocker, H. & Marky, L. A. Predicting dna duplex stability from the base sequence. *Proc Natl Acad Sci U S A* **83**, 3746–3750 (1986). 10.1073/pnas.83.11.3746.

[100] Sarai, A., Mazur, J., Nussinov, R. & Jernigan, R. L. Sequence dependence of DNA conformational flexibility. *Biochemistry* **28**, 7842–9 (1989).

[101] Bailey, J. & Eichler, E. Primate segmental duplications: crucibles of evolution, diversity and disease. *Nat Rev Genet* **7**, 552–564 (2006). 10.1038/nrg1895.

[102] Bailey, T. *et al.* Meme suite: tools for motif discovery and searching. *Nucleic Acids Res* **37**, W202 (2009). 10.1093/nar/gkp335.

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

[104] Lee, S., Cheran, E. & Brudno, M. A robust framework for detecting structural variations in a genome. *Bioinformatics* **24**, i59 (2008). 10.1093/bioinformatics/btn176.

[105] Campbell, P. J. *et al.* Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat Genet* **40**, 722–9 (2008).

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

[107] Paten, B. *et al.* Genome-wide nucleotide-level mammalian ancestor reconstruction. *Genome Res* **18**, 1829–1843 (2008). 10.1101/gr.076521.108.

[108] Spencer, C. C. A. *et al.* The influence of recombination on human genetic diversity. *PLoS Genet* **2**, e148 (2006). 10.1371/journal.pgen.0020148.

[109] Kent, W. J., Baertsch, R., Hinrichs, A., Miller, W. & Haussler, D. Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes. *Proc Natl Acad Sci U S A* **100**, 11484–11489 (2003). 10.1073/pnas.1932072100.

[110] Schwartz, S. *et al.* Human-mouse alignments with BLASTZ. *Genome Res* **13**, 103–7 (2003).

[111] Zhang, Z. *et al.* Pseudopipe: an automated pseudogene identification pipeline. *Bioinformatics* **22**, 1437–1439 (2006). 10.1093/bioinformatics/btl116.

[112] Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biol* **10**, R25 (2009). 10.1186/gb-2009-10-3-r25.

[113] Zarrinpar, A., Park, S.-H. & Lim, W. A. Optimization of specificity in a cellular protein interaction network by negative selection. *Nature* **426**, 676–80 (2003).

[114] Pawson, T. & Nash, P. Assembly of cell regulatory systems through protein interaction domains. *Science* **300**, 445–52 (2003).

[115] Zarrinpar, A., Bhattacharyya, R. P. & Lim, W. A. The structure and function of proline recognition domains. *Sci STKE* **2003**, RE8 (2003).

[116] Hanks, S. K., Quinn, A. M. & Hunter, T. The protein kinase family: conserved features and deduced phylogeny of the catalytic domains. *Science* **241**, 42–52 (1988).

[117] Zeng, G. & Cai, M. Regulation of the actin cytoskeleton organization in yeast by a novel serine/threonine kinase Prk1p. *J Cell Biol* **144**, 71–82 (1999).

[118] Pawson, T. Protein modules and signalling networks. *Nature* **373**, 573–80 (1995).

[119] Tong, A. H. Y. *et al.* A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science* **295**, 321–4 (2002).

[120] Landgraf, C. *et al.* Protein interaction networks by proteome peptide scanning. *PLoS Biol* **2**, E14 (2004).

[121] Tonikian, R., Zhang, Y., Boone, C. & Sidhu, S. S. Identifying specificity profiles for peptide recognition modules from phage-displayed peptide libraries. *Nat Protoc* **2**, 1368–86 (2007).

[122] Hutti, J. E. *et al.* A rapid method for determining protein kinase phosphorylation specificity. *Nat Methods* **1**, 27–9 (2004).

[123] Beltrao, P. & Serrano, L. Comparative genomics and disorder prediction identify biologically relevant SH3 protein interactions. *PLoS Comput Biol* **1**, e26 (2005).

[124] Jansen, R. *et al.* A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* **302**, 449–53 (2003).

[125] Henikoff, J. G. & Henikoff, S. Using substitution probabilities to improve position-specific scoring matrices. *Comput Appl Biosci* **12**, 135–43 (1996).

[126] Eddy, S. R. Where did the BLOSUM62 alignment score matrix come from? *Nat Biotechnol* **22**, 1035–6 (2004).

[127] Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* **89**, 10915–9 (1992).

[128] McLachlan, A. D. Repeating sequences and gene duplication in proteins. *J Mol Biol* **64**, 417–37 (1972).

[129] Yaffe, M. B. *et al.* A motif-based profile scanning approach for genome-wide prediction of signaling pathways. *Nat Biotechnol* **19**, 348–53 (2001).

[130] Obenauer, J. C., Cantley, L. C. & Yaffe, M. B. Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res* **31**, 3635–41 (2003).

[131] Adamczak, R., Porollo, A. & Meller, J. Accurate prediction of solvent accessibility using neural networks-based regression. *Proteins* **56**, 753–67 (2004).

[132] Jones, D. T. & Ward, J. J. Prediction of disordered regions in proteins from position specific score matrices. *Proteins* **53 Suppl 6**, 573–8 (2003).

[133] Ward, J. J., McGuffin, L. J., Bryson, K., Buxton, B. F. & Jones, D. T. The DISOPRED server for the prediction of protein disorder. *Bioinformatics* **20**, 2138–9 (2004).

[134] Remm, M., Storm, C. E. & Sonnhammer, E. L. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol* **314**, 1041–52 (2001).

[135] Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792–7 (2004).

[136] Pei, J. & Grishin, N. V. AL2CO: calculation of positional conservation in a protein sequence alignment. *Bioinformatics* **17**, 700–12 (2001).

[137] Tonikian, R. *et al.* Bayesian modeling of the yeast SH3 domain interactome predicts spatiotemporal dynamics of endocytosis proteins. *PLoS Biol* **7**, e1000218 (2009).

[138] Mok, J. *et al.* Deciphering protein kinase specificity through large-scale analysis of yeast phosphorylation site motifs. *Sci Signal* **3**, ra12 (2010).

[139] Turhan, B. & Bener, A. Analysis of naive bayes' assumptions on software fault data: An empirical study. *Data Knowl. Eng.* **68**, 278–290 (2009).

[140] Puntervoll, P. *et al.* ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Res* **31**, 3625–30 (2003).

[141] Frank, E., Hall, M., Trigg, L., Holmes, G. & Witten, I. H. Data mining in bioinformatics using Weka. *Bioinformatics* **20**, 2479–81 (2004).