STRUCTURAL VARIATION DISCOVERY AND GENOTYPING FROM WHOLE GENOME SEQUENCING: METHODOLOGY AND APPLICATIONS

A Dissertation Presented
By

Jiali Zhuang

Submitted to the Faculty of the University of Massachusetts Graduate School of Biomedical Sciences, Worcester in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

(September 15, 2015)

STRUCTURAL VARIATION DISCOVERY AND GENOTYPING FROM WHOLE GENOME SEQUENCING: METHODOLOGY AND APPLICATIONS

A Dissertation Presented
By

Jiali Zhuang

The signatures of the Dissertation Committee signify
completion and approval as to style and content of the Dissertation

Zhiping Weng, Thesis Advisor

William Theurkauf, Member of Committee

Daniel Caffrey, Member of Committee

Manuel Garber, Member of Committee

Ekta Khurana, Member of Committee

The signature of the Chair of the Committee signifies that the written dissertation
meets the requirements of the Dissertation Committee

Jeffrey Bailey, Chair of Committee

The signature of the Dean of the Graduate School of Biomedical Sciences signifies
that the student has met all graduation requirements of the school

Anthony Carruthers, Ph.D.,
Dean of the Graduate School of Biomedical Sciences

Bioinformatics and Computational Biology

September 15, 2015

ACKNOWLEDGEMENT

I would like to thank my thesis adviser Dr. Zhiping Weng for supporting my research by providing funding, supervision, intellectual mentoring and collaborative opportunities. I am particularly grateful for the freedom she allowed me to follow my own interest and explore different paths. The work presented in this dissertation would not have been possible without her support, encouragement and guidance. I would also like to thank Dr. Bill Theurkauf, a major collaborator, for the enlightening discussions on various topics including transposition regulation, genomic stability maintenance and piRNA pathways. As members of my Thesis Research Advisory Committee Dr. Jeffrey Bailey and Dr. Daniel Caffrey spent their precious time on my TRAC meetings and provided critical insights and suggestions that foster the progress of my research. Many people assist in and contribute to the development of the work presented in this dissertation and it is not possible to list all their names here. Particularly I would like to point out that I benefited a lot from the lively and stimulating discussions with all members (including some former members) of the Weng Lab and some members of the Theurkauf Lab.

ABSTRACT

A comprehensive understanding about how genetic variants and mutations contribute to phenotypic variations and alterations entails experimental technologies and analytical methodologies that are able to detect genetic variants/mutations from various biological samples in a timely and accurate manner. High-throughput sequencing technology represents the latest achievement in a series of efforts to facilitate genetic variants discovery and genotyping and promises to transform the way we tackle healthcare and biomedical problems. The tremendous amount of data generated by this new technology, however, needs to be processed and analyzed in an accurate and efficient way in order to fully harness its potential. Structural variation (SV) encompasses a wide range of genetic variations with different sizes and generated by diverse mechanisms. Due to the technical difficulties of reliably detecting SVs, their characterization lags behind that of SNPs and indels. In this dissertation I presented two novel computational methods: one for detecting transposable element (TE) transpositions and the other for detecting SVs in general using a local assembly approach. Both methods are able to pinpoint breakpoint junctions at single-nucleotide resolution and estimate variant allele frequencies in the sample. I also applied those methods to study the impact of TE transpositions on the genomic stability, the inheritance patterns of TE insertions in the population and the molecular mechanisms and potential functional consequences of somatic SVs in cancer genomes.

Table of Contents

PREFACE

The Chapters II and III of this dissertation are adopted from two published works. These works are:

Local sequence assembly reveals a high-resolution profile of somatic structural variations in 97 cancer genomes

Jiali Zhuang and Zhiping Weng*,

Nucl. Acids Res. (2015) doi: 10.1093/nar/gkv831

First published online: August 17, 2015

TEMP: a computational method for analyzing transposable element polymorphism in populations

Jiali Zhuang, Jie Wang, William Theurkauf,* and Zhiping Weng,*

Nucl. Acids Res. (17 June 2014) 42(11): 6826-6838. doi: 10.1093/nar/gku323

CHAPTER I

INTRODUCTION

Genetic variations and mutations are the main driving force behind population phenotypic polymorphism, adaptation, evolution and diseases. Better understanding of the molecular mechanisms underlying those events, their phenotypic consequences, and their inheritance patterns within the population is therefore crucial for unraveling many of the mysteries in biology. For example, associating genetic variants with a disease phenotype may yield clues about which genes are potentially responsible for the disease (known as Genome Wide Association Studies or GWAS); and mutations that alter protein-coding gene sequences often disrupt the normal functioning of the protein product and therefore lead to physiological changes in the cell. Accurate and reliable detection of variants/mutations from various biological samples provides a solid foundation upon which more in-depth characterizations of the system at hand can be carried out. The advent of next-generation sequencing technology makes it possible to survey variants/mutations across the entire genome with reasonable time and cost and therefore completely revolutionizes how genetic variants are studied. The technology itself undergoes a series of paradigmatic shifts and keeps improving in terms of speed, accuracy and cost-efficiency. The cost for sequencing the entire human genome has decreased dramatically since

the publication of the first human genome draft and will soon become affordable to individual patients and consumers. This technological breakthrough opens up new opportunities for studying the genetic variants and mutations at the population level and promises to transform the field of biomedicine. To fully capitalize the enormous power of next-generation sequencing technology, however, it is crucial to implement well-designed, carefully calibrated and rigorously tested computational algorithms.

Diverse types of genetic variants have been identified so far: from single nucleotide polymorphisms (SNPs), to short insertion/deletions (indels) that are shorter than 50 bp, to structural variations that can span several megabases, to aneuploidy where the entire chromosome is duplicated or missing. In this chapter I provide an overview of the evolution of experimental techniques for genetic variant/mutation detection and discuss how they were used to study biological systems historically.

Experimental techniques for genetic variation detection

Traditionally genetic variants/mutations are identified and genotyped by Sanger sequencing technology (or some variations of the technology). This method takes advantage of the labeled di-deoxynucleotides (ddNTPs), which upon incorporation terminate the DNA synthesis reaction. The procedure starts with

mixing fluorescently labeled ddNTPs (ddATP, ddCTP, ddGTP and ddTTP are labeled with different fluorochromes) with normal deoxynucleotides (dNTPs) and initiating the DNA synthesis reaction using the sample DNA as the template. Next the newly synthesized DNAs are separated by size and the fluorescence of their terminal ddNTP read by the fluorescence detector. Finally the sequence of the sample DNA is determined by analyzing the resulting chromatograms. By comparing the sample DNA sequences with reference sequence or the sequence of control DNA, it is relatively straightforward to figure out the variants/mutations if the sequences are of high quality. Very large structural variations that change the karyotypes of the cell can also be detected by cytogenetics. These approaches, however, have very low throughput because they are labor-intensive and time-consuming and therefore only a limited number of loci can be surveyed.

Several array-based techniques were successfully developed that can achieve variants discovery or genotyping in a high-throughput manner. SNP arrays, for example, enable researchers to genotype millions of known SNP sites (Illumina Omni5 chip, for example, contains 5 million markers) across the entire genome in a single experiment. The fundamental principle behind this technique is the hybridization between complementary DNA fragments. DNA fragments representing all the possible genotypes of the SNP sites are prepared and planted on the microchips and then the fluorescence labeled sample DNA

fragments are allowed to hybridize with the fragments on the chip. Perfectly complementary fragments bind to each other with a higher affinity and therefore displays stronger signal, which form the basis for determining the sample genotypes at the measured loci. The drawback of this approach is that it only genotypes pre-determined SNP sites and cannot perform *de novo* single nucleotide variations/mutations discovery. A similar approach called array-based Comparative Genomic Hybridization (aCGH) uses hybridization to detect copy number changes (deletions and duplications) of large genomic regions. When a segment of the genome is deleted the signal from the hybridization will decrease accordingly. And vice versa, when a genomic segment is duplicated the signal will increase. In addition it is able to compare copy number changes of genomic segments between two biological samples by labeling the sample DNAs with two different dyes.

High-throughput sequencing technology emerged in the late 2000s and has since become the major driving force behind the study of genetic variations (X. Zhou et al., 2010). It makes accurate, fast and genome-wide measurements of sample DNA sequences possible (Illumina Genome Analyzer, for instance, can finish a 14Gb run in 7 days with error rate around 1%) and therefore tremendously accelerates the effort to discover and genotype genetic variations in various species. Since it directly sequences the sample DNA molecules it can be used for both the discovery and genotyping without being constrained to known

variation sites. Different commercial companies develop their own patented implementation for high-throughput sequencing (X. Zhou et al., 2010). Illumina for example employs procedures that involve DNA fragment immobilization, solid-phase bridge amplification and sequencing by synthesis through a number of reaction cycles. Different parts of the genome have been the subject of sequencing studies with different emphasis. Exome sequencing aims to identify genetic variations within known exon regions whereas whole genome sequencing has the potential to detect variations across the whole genome. It has the advantage of being cheaper, able to achieve very high coverage and easier to interpret the identified variations (Bamshad et al., 2011; Clark et al., 2011). Targeted sequencing involves capturing DNA sequences from a region of interest (e.g., regions implicated by a GWAS to be associated with a certain disease) and then sequences them. Now with the cost of sequencing experiments rapidly decreasing and our understanding about the non-coding regions of the genome improving, whole genome sequencing has become more and more widely used and promises to reveal previously unknown insights from various biological systems.

Applications of genetic variation discovery and genotyping

Genetic variants discovery and genotyping is not the end itself, instead it serves as the foundation to better understand various biological systems, pathways and

phenotypes. The three most common applications of genetic variants/mutations discovery and genotyping are: Genome Wide Association Studies (GWASs), expression Quantitative Trait Locus (eQTL) and somatic mutation characterization in cancer genomes (Stranger, Stahl, & Raj, 2011; Vogelstein et al., 2013).

GWAS tries to detect genetic variations (typically SNPs or CNVs) that are significantly associated with a trait of interest. It is widely employed to study the genetic underpinnings of complex diseases where a case-control approach is typically used. In this type of studies a large number of individuals from two groups (case: individuals with the disease; control: healthy individuals from similar ethnic background and geological locations) are genotyped for common SNPs genome wide (the number of total SNPs measured varies across studies, ranging from tens of thousands to millions) and rigorous statistical tests are performed to identify significant associations between SNPs and the disease phenotype. Since its inception numerous GWASs have been successfully carried out on various traits and have yielded tens of thousands of significant SNP-trait associations. For instance, multiple loci within the FGF12 gene have recently been found to strongly associate with Kashin-Beck disease (Zhang et al., 2015), suggesting that FGF12 is a novel candidate gene for causing the disease. Another recent study discovered significant associations between polymorphisms in GCKR, SLC17A1 and SLC22A12 genes and the gout phenotype in the Han

Chinese population (Z.-W. Zhou et al., 2015). A GWAS on mucinous ovarian carcinoma susceptibility identifies 3 novel risk associations, two of which have significant eQTL associations with HOXD9 and PAX8 genes, respectively (Ovarian Cancer Association Consortium, Australian Cancer Study, Australian Ovarian Cancer Study Group, 2015). The NHGRI-EBI GWAS catalog is a comprehensive and curated database for SNP-trait associations, which includes around 14,000 statistically significant (P-value < 1e-5) SNP-trait associations as of 2013 (Welter et al., 2014). It is worth noting that SNPs significantly associated with a certain trait are not necessarily the cause of the phenotypic difference. Indeed, vast majority of the significant SNPs in GWASs fall in intergenic or intronic regions and do not change the amino acid sequences of the protein product. It is possible that many of those SNPs are actually not causal themselves but in linkage disequilibrium with some other causal variants such as SVs.

Expression Quantitative Trait Locus (eQTL) works in similar fashion only seeking significant associations between SNPs and gene expressions. Since gene expression level is a quantitative trait, eQTL follows the standard QTL mapping which uses *t*-test or ANOVA to detect significant SNP-trait associations. The identified significant associations may shed important light on the regulatory mechanisms controlling the gene expression profiles. For example, a recent study elucidates the genomic modulators of gene expression in human

neutrophils, reporting 450 novel *cis*-eQTLs in neutrophils (Naranbhai et al., 2015). The Genotype-Tissue Expression (GTEx) consortium recently presented a pilot analysis on data collected from 1641 samples across 43 tissue types from 175 individuals, describing thousands of tissue-specific and shared eQTLs that provide a global view of cellular expression regulatory mechanisms at an unprecedented scale (GTEx Consortium, 2015). eQTL analysis may also facilitate the functional annotation of the variants identified in the genome-wide association studies and the identification of causal or risk genes. For instance, a study illustrates the cell-type-specific regulation of the expression level of asthma-related genes by integrating GWAS-identified SNPs associated with asthma and eQTL information in bronchial epithelial cells and bronchial alveolar lavage (X. Li et al., 2015).

Somatic mutations are genetic alterations that accumulated in the cells during an organism's lifetime. They are the results of DNA lesions induced by mutagens or errors during DNA replication process. The rate of spontaneous DNA mutations is quite low under normal conditions, but when the organism is exposed to high dose of mutagens and/or the cellular DNA proofreading or repair mechanisms are compromised large amount of mutations could accumulate in a short period of time. Many cancer genomes are characterized by a large number of genetic mutations. By comparing the DNA from the cancer samples to the DNA from the healthy tissue of the same individual, researchers are able to identify somatic

mutations by focusing on variants that are only present in the cancer samples. By performing exome sequencing on large numbers of cancer genomes, previous studies have successfully identified genes that are recursively mutated in multiple tumor samples. This approach leads to the discovery of many well-known oncogenes and tumor suppressors such as *TP53*, *PTEN*, *BRCA1* and *SMAD4* (Vogelstein et al., 2013). Catalog of Somatic Mutations In Cancers (COSMIC) is a comprehensive database that hosts compiled information about somatic mutations in human cancers. As of April 2014, COSMIC includes 2,002,811 coding point mutations in over one million tumor samples across all major cancer types (Forbes et al., 2015).

A history of structural variations and their phenotypic implications

All the aforementioned applications can in principle be achieved with any type of genetic variants/mutations but in practice most studies focus on SNPs/indels because reliable and efficient experimental methods and computational algorithms are available to detect and genotype them. In addition to SNPs/indels, another important class of genetic variants/mutations is structural variation that includes deletions, duplications, copy number variations, transposable element movements, inversions and translocations. The study of structural variations started with the observation that segmental duplications and copy number changes are common within the human genome when the first draft of the human

genome became available (Bailey et al., 2002; Bailey, Yavor, Massa, Trask, & Eichler, 2001) and is further motivated by the fact that many regions susceptible to copy number changes contain genes that are known to play crucial roles in various diseases (Ji, Eichler, Schwartz, & Nicholls, 2000).

In the early days of the copy number variation study (before the advent of high throughput sequencing technology), the most widely used experimental techniques are bacterial artificial chromosome (BAC) microarray, array CGH and fluorescent *in situ* hybridization (FISH). For example, Sharp *et al.* examined 130 potential rearrangement hotspots on 47 normal individuals with a targeted BAC microarray and identified 119 copy number polymorphisms (Sharp et al., 2005). In this type of studies regions of interest are cloned into BACs, which are then used in constructing the microarrays. The DNAs extracted from both the sample and the reference are then labeled with Cy3 and Cy5 dyes and the potential copy number changes are inferred from the fluorescence resulting from the hybridization (Sharp et al., 2005). Now with the better annotation of the reference genome and better manufacturing technologies, both standardized and custom aCGH arrays are available that allows the characterization of copy number polymorphisms genome wide. The working principle of aCGH is also based on hybridization and fluorescence signal strength and has been covered in earlier sections. On the computational side, Bailey *et al.* proposed algorithms that detect segmental duplications within the human reference genome by looking for

significant sequence similarities between different genomic regions (Bailey et al., 2001; 2002). FISH is a laborious but reliable experimental technique often used to validate a selected set of predicted SVs.

The continual advancement of sequencing technology enables more efficient and higher-resolution analysis of structural variations. In 2008 Kidd et al. presented arguably the first genome wide SV study with sequencing technology. They cloned the entire genomes of eight individuals from diverse ethnic background into fosmids and sequenced both ends of each clone insert with Sanger sequencing to generate the so-called end-sequence pairs (ESPs). Based on fosmids whose apparent insert size deviate from library mean insert size and several additional filtering and validation steps they reported a total of 1,695 SVs, encompassing more SV types and wider SV size spectrum than possible with array-based approaches (Kidd et al., 2008). The advent of high throughput whole genome sequencing technology gave a further impetus to the characterization of SVs. In 2011, the 1000 Genome Projects consortium presented a comprehensive survey of SVs across 185 human genomes. Combining both whole genome sequencing technology and extensive experimental validation, they reported more than 28,000 SVs (53% of which are mapped with single-nucleotide resolution) and laid the foundation for understanding the population landscape of SVs (Mills et al., 2011).

Transposable elements (TEs) are DNA sequences able to replicate or mobilize themselves within the genome. They were first discovered in plants and later found to be present in most species and even make up of a considerable proportion of multiple genomes (Gogvadze & Buzdin, 2009; McCLINTOCK, 1950). There are two classes of TEs: Class I TEs transpose by the so-called “cut-and-paste” mechanism where the original copy is excised and then integrated at a novel site; Class II TEs transpose through RNA transcript intermediates in the so-called “copy-and-paste” mechanism which results in extra copies of the TE. Most active TEs contain the sequences encoding the transposases needed for their transposition and therefore in a sense they can be viewed as semi-independent parasites within their host genomes. Not surprisingly, most host genomes employ sophisticated defense mechanisms to contain the spread of TEs and preserve their genomic integrity as much as possible (Siomi, Sato, Pezic, & Aravin, 2011). The evolutionary arms race between the TE parasites and host genomes is a fascinating area for future research (Jacobs et al., 2014).

The impacts of diverse types of structural variations on human health have been extensively studied since SVs were first discovered and numerous associations between SVs (copy number variations in particular) and various diseases have been reported. For instance, more than a dozen cancer-related genes including *PIK3CA*, *MYC* and *EGFR* are affected by copy number variations (CNVs) in multiple gastric cancer samples and may contribute to gastric oncogenesis

(Liang, Fang, & Xu, 2015). A large-scale survey of 715 grade II and III gliomas genomes conducted recently revealed that 37 focal regions and 19 chromosomal arms undergo recurrent copy number changes. Among the genes that are affected by the CNVs are well-known oncogenes and tumor suppressors like *MDM4*, *EGFR*, *CCND2* and *RB1* (Suzuki et al., 2015). With increased efficiency and decreased cost of whole genome sequencing experiments, genome-wide associations using CNVs has become popular lately and shed some light on the impact of SVs on diseases. A recent whole wide CNV scan performed on a large cohort of 249 patients and 232 matched controls, for example, identified a 13q12.11 duplication that includes exportin-4 gene to be associated non-alcoholic fatty liver disease (Zain et al., 2015). Genome wide CNV studies using SNP arrays also help identify potential causal genes for colorectal adenomatous polyposis (Horpaopan et al., 2015) and Type II diabetes (Dajani et al., 2015).

Compared with CNVs, the phenotypic implications of other types of SVs are less well understood. That is changing rapidly, however, as in recent years great efforts have been made to catch up with their characterization. Somatic transposable element movements, for instance, have been observed to be prevalent in many cancer genomes and are likely to contribute to the initiation or progress of cancers (Carreira, Richardson, & Faulkner, 2014; Helman et al., 2014). The retro-transpositions of a TE may disrupt protein-coding genes (the a new copy is inserted into one of the exons) or alter regulatory machinery by

either disrupting the regulatory elements or bring with it new regulatory elements to target genes near its new insertion sites. All those events could potentially breaking the proper regulation and homeostasis in the cell and lead to deleterious consequences and have been indeed observed in multiple types of cancers (Helman et al., 2014; Lee et al., 2012; Shukla et al., 2013; Solyom et al., 2012; Tubio et al., 2014). Similar to SNPs/indels, it is challenging to distinguish between SVs that initiate the oncogenesis (driver mutations) and those that are the result of compromised DNA repair machinery (passenger mutations) and may require analysis across large number of cancer samples.

Extensive structural variation polymorphisms have been reported on populations of multiple species (Mills et al., 2011; Mills, Bennett, Iskow, & Devine, 2007; Zichner et al., 2013). A study of SVs within 39 strains derived from a wild Drsophila Melanogaster population in North Carolina (the Drosophila Genetic Reference Panel, DGRP) reported 8,962 deletions and 916 duplications. Futhermore, an eQTL mapping with those SVs revealed functional impact at more than 100 loci (Zichner et al., 2013). Several studies examined the role of TE transpositions in shaping the expression regulatory machinery (González & Petrov, 2009) and suggested that TE movement plays a crucial part in the evolution of new species or phenotypes such as pregnancy in mammals (Lynch, Leclerc, May, & Wagner, 2011).

A brief summary of computational algorithms that detect SVs from paired-end whole genome sequencing datasets

Paired-end high throughput whole genome sequencing has become the most widely employed experimental approach for genome-wide scale discovery and genotyping of structural variations during the last several years. The complexity and scale of the SVs and the relatively short read length permitted by the mainstream high throughput sequencing technology presents a daunting computational challenge for accurate and efficient discovery of SVs.

Many methods and algorithms that attempt to harness the power of pair-end high throughput sequencing for whole genome SV detection have been proposed during the course of past five or six years (K. Chen et al., 2009; Hormozdiari et al., 2010; Layer, Chiang, Quinlan, & Hall, 2014; Rausch et al., 2012; Sindi, Önal, Peng, Wu, & Raphael, 2012; J. Wang et al., 2011; Ye, Schulz, Long, Apweiler, & Ning, 2009). Almost all of them start with mapping the read-pairs to the reference genome and leverage on three types of discordant mapping information that indicate a difference between sample genome and the reference sequence (**Figure 1.1**):

1) Discordant read-pairs: read-pairs whose distance or orientation between the two reads are inconsistent with the reference genome;

2) Split-read alignments: different parts of a read mapped to discontiguous genomic loci;
3) Read-depth: a significant change in the read coverage of a particular genomic region indicating a copy number change of the region.

These three types of information are useful under different circumstances. For example, the discordant read-pair information can potentially detect all types of SVs but nonetheless fails to pinpoint the exact location of the breakpoints; and read-depth information is unable to detect balanced (no change in copy number) SVs such as inversions and translocations. While each of the three types of information is relatively straightforward in itself, it is difficult to integrate these sources under a single computational framework. In addition, the sequencing technology is far from error-free and SV detection algorithms should be able to filter out false positive SV calls resulting from sequencing errors such as substitutions and indels in the sequencing reads, chimeric read-pairs and abnormal library insert sizes.

The most popular computational framework for SV detection by far considers the three types of discordant mapping information in a sequential manner. It first selects all the discordant read-pairs following alignment to the reference genome and then uses various algorithms to cluster read-pairs that appear to support the same SV event. Split-read and read-depth information is then used to further validate or refine the breakpoints of the SVs predicted in the previous step. Only

SV calls that are supported by multiple read-pairs (the cutoff threshold differs from one method to another and is often a parameter tunable by the users) are retained. Breakdancer (K. Chen et al., 2009), VariationHunter (Hormozdiari et al., 2010), Hydra (Quinlan et al., 2011) and DELLY (Rausch et al., 2012) are examples of methods built upon this framework. In other SV discovery tools such as CREST (J. Wang et al., 2011) and Pindel (Ye et al., 2009) the initial SV calling is triggered by soft-clipped reads (split-read mapping) and then further validated or refined by either a second round of alignment (in the case of Pindel) or the assembly of soft-clipped reads (in the case of CREST). TIGRA is an interesting tool that does not predict SVs by itself but try to pinpoint the breakpoints of SVs predicted by other methods through targeted iterative graph assembly (K. Chen et al., 2013). Some more recent algorithms such as GASVPro (Sindi et al., 2012), cnvHiTSeq (Bellos, Johnson, & M Coin, 2012) and LUMPY (Layer et al., 2014) attempt to integrate information from different sources into a statistical model that measures the confidence of each SV prediction and even the possible range of breakpoints (in the case of LUMPY). Due to both sequencing errors and mapping artifacts many SV discovery methods have relatively high false-positive rates, especially for samples with moderate coverage. Handsaker *et al.* proposed an algorithm that leverages on population data to increase both the sensitivity and specificity of genomic deletion detection. The proposed algorithm Genome STRiP, however, is only able to detect deletions and its performance on rare SVs

has not been rigorously evaluated (Handsaker, Korn, Nemesh, & McCarroll, 2011).

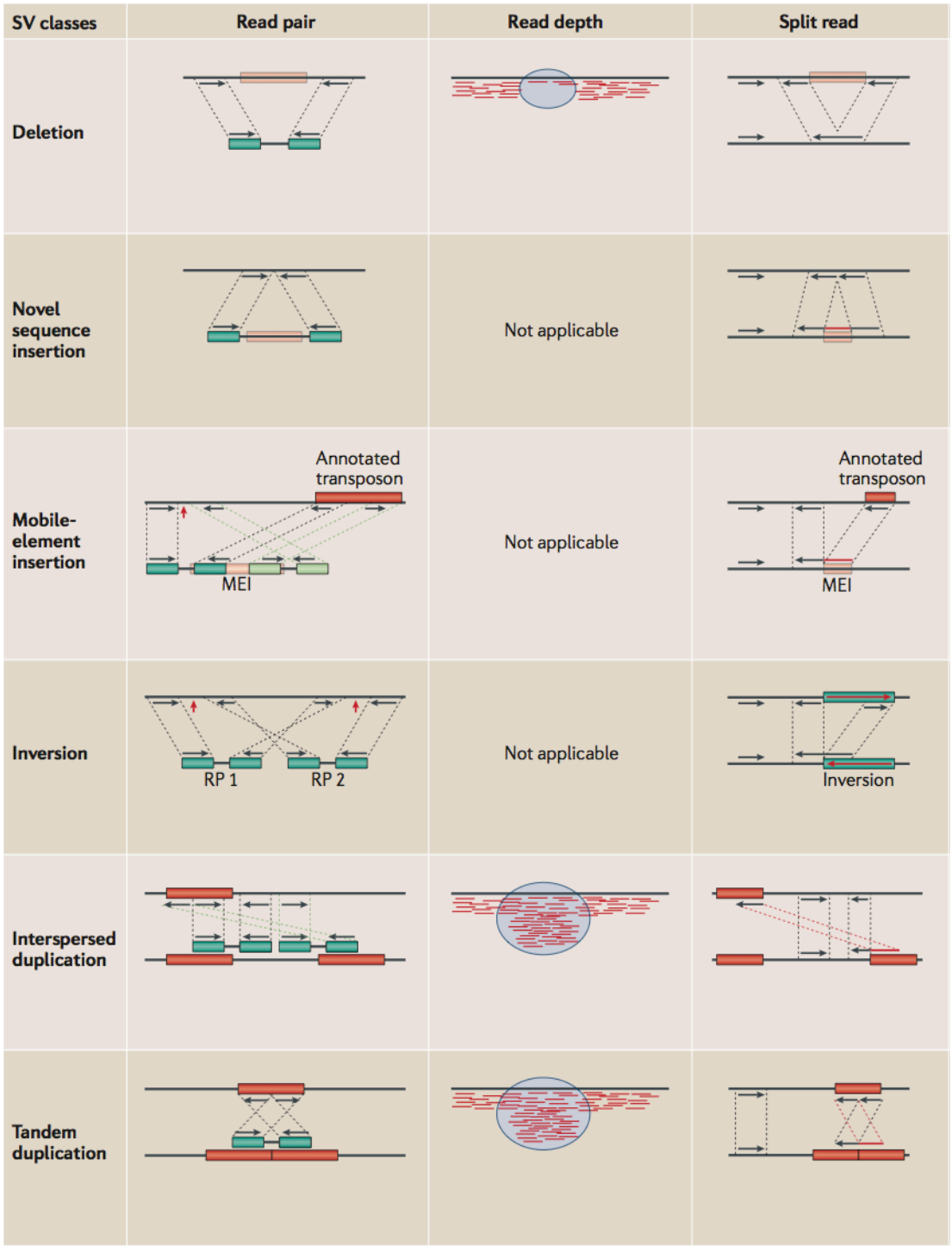
A unique feature that distinguishes the detection of TE insertions that are absent in the reference genome from the detection of other types of SVs is that the inserted TE sequences are highly repetitive. As a result, it is extremely difficult and in some cases even impossible to determine the exact genomic origin of a read deriving from a TE sequence since there are hundreds or even thousands of virtually identical copies of the same TE in the genome. Fortunately under most circumstances it suffices to know what element instead of which copy of that element is inserted and therefore nearly all TE detecting algorithms take advantage of the annotated TE consensus sequences to identify TE insertion events (Keane, Wong, & Adams, 2013; Kofler, Betancourt, & Schlötterer, 2012). When one of the reads in a discordant read-pair or the clipped part of a soft-clipped read can be confidently mapped to a TE consensus sequence, the read-pair or the read implies a TE insertion event. The genome-mapping read or part of the read can be used to estimate the genomic locations of the insertion (Keane et al., 2013; Kofler et al., 2012). The effectiveness of this approach is based on the premise that for a particular TE its consensus sequence is a reliable representation of all the sequences of its active copies within the genome. This assumption typically holds because considerable mutations within a TE copy often abolishes its capacity for transposition (by disrupting the catalytic activity of

the transposase for example) and therefore severely limits the number of mutated copies.

Although the application of associating SNPs/indels with various phenotypes has proven fruitful, concentrating on just one or two types of variants/mutations risks missing the larger picture. Therefore developing novel techniques and algorithms that are able to accurately detect and genotype other types of genetic variants/mutations is of vital importance. As noted before structural variation (SV) discovery from paired-end sequencing datasets presents a significant challenge and no methods developed so far achieve superior performance for all tasks/SV types. In addition, using the number of reads supporting a predicted SV event as a metric for confidence can sometimes be misleading because the coverage is usually not uniform across the entire genome and using a fixed cutoff threshold for the whole genome is often counter-productive. Pinpointing the breakpoints of SVs at single-nucleotide resolution and accurately estimating the SV allele frequency in heterogeneous samples is extremely valuable for a deeper knowledge concerning the formation mechanisms, inheritance patterns and phenotypic significance of those SVs. In this dissertation, I described two novel algorithms I developed during my thesis research for detecting transposable elements (TEs) movement and structural variations from paired-end high-throughput genomic sequencing datasets and the biological insights gleaned from applying them to real biological samples. TEMP (Transposable Element

Movement in Populations) is a method capable of accurately detecting TE transpositions from pooled sequencing datasets and estimating their frequencies within the pool. This method and its applications are discussed at length in Chapter II. laSV is a novel local assembly based algorithm for detecting SVs in general. I applied it to profile somatic SVs in 97 cancer genomes and the findings are described in Chapter III.

Figure 1.1



Schemetic representation of major types of structural variations and how to detect them. Adopted from Alkan *et al.* with permission from Nature Publishing Group (permission ID number: 3718870525416).

CHAPTER II

TEMP: a computational method for analyzing transposable element polymorphism in populations

Summary

Insertions and excisions of transposable elements (TEs) affect both the stability and variability of the genome. Studying the dynamics of transposition at the population level can provide crucial insights into the processes and mechanisms of genome evolution. Pooling genomic materials from multiple individuals followed by high-throughput sequencing is an efficient way of characterizing genomic polymorphisms in a population. Here we describe a novel method named TEMP, specifically designed to detect transposable element movements present with a wide range of frequencies in a population. By combining the information provided by pair-end reads and split reads, TEMP is able to identify both presence and absence of TE insertions in genomic DNA sequences derived from heterogeneous samples; accurately estimate the frequencies of transposition events in the population; and pinpoint junctions of high frequency transposition events at nucleotide resolution. Simulation data indicate that TEMP outperforms other algorithms such as PoPoolationTE, RetroSeq, VariationHunter, and GASVPro. TEMP also performs well on whole-genome human data derived from the 1000 Genomes Project. We applied TEMP to

characterize the TE frequencies in a wild *Drosophila melanogaster* population and study the inheritance patterns of TEs during hybrid dysgenesis. We identified sequence signatures of TE insertion and possible molecular effects of TE movements, such as altered gene expression and piRNA production. TEMP is freely available at github: https://github.com/JialiUMassWengLab/TEMP.git.

Introduction

Transposable element (TE) mobilization is one of the major sources of genomic variation and a potential driving force of evolution (Bennetzen, 2000; Britten, 2010; Hedges & Belancio, 2011). Detecting transposition events within the genome is therefore crucial for understanding the mechanisms by which TEs are regulated and the phenotypic consequences that result from TE movements. The task of detecting TE insertions and excisions falls within the more general category of genomic structural variation detection (Alkan, Coe, & Eichler, 2011). Much progress has been made in discovering structural variations from high-throughput genomic DNA sequencing data (Hormozdiari, Alkan, Eichler, & Sahinalp, 2009; Quinlan et al., 2010; Rausch et al., 2012). So far, most structural variation discovery tools are designed to handle isogenic samples-- i.e., they assume that the sequence reads originate from a single genome or at least the sample is dominated by a single genome (Alkan et al., 2011). However, just as

any other types of genomic variation, it would be extremely useful to estimate the population frequency of polymorphic transposition events. Sequencing a large number of individuals in a population separately is impossible under many circumstances because of the prohibitively high costs and the difficulty in obtaining enough experimental material. Pooled sequencing is a widely employed experimental practice whereby investigators pool tissues from multiple individuals (or organisms) and sequence the DNA (or RNA) without knowing which read originates from which individual (or organism) (Calvo et al., 2010; Futschik & Schlötterer, 2010; S. R. Wang et al., 2013a; Zhu, Bergland, González, & Petrov, 2012). In fact, for many species that cannot be individually cultured in laboratory conditions, pooled sequencing is the only means for obtaining sufficient experimental material as required by state-of-the-art sequencing technologies. When analyzed with an effective computational algorithm, this approach can accurately estimate the population frequency of transposition events.

When applied to pooled sequencing data, methods designed to detect structural variations in largely isogenic samples can only detect variations that are shared by most genomes in the pool. Discovering TE transpositions and estimating their frequencies using a pooled sequencing dataset present some unique computational challenges. Detecting rare TE transposition events with high confidence, identifying reads that are likely to support the same transposition

event, and overcoming biases stemming from the non-uniformity of sequencing depth across the genome are some of the difficulties involved. Kofler *et al.* designed an algorithm named PoPoolationTE to detect novel TE insertions and estimate their population frequency from pooled sequencing data. They applied PoPoolationTE to a natural population of *Drosophila* to study transposon evolution. In this article we present an algorithm named TEMP that uses discordant mapping reads to detect TE polymorphisms relative to a reference genome, pinpoint the position of their junctions within genomic DNA, and estimate their population frequencies from the pooled sequencing data. We demonstrated TEMP's performance by comparing it with PoPoolationTE, RetroSeq (an algorithm designed for detecting TE insertions in individual genomes), and two general-purpose structural variation discovery algorithms VariationHunter, and GASVPro using simulated data. We further used TEMP to analyze several biological datasets in *Drosophila melanogaster* to demonstrate the unique biological insights that can be obtained using our algorithm. TEMP requires a curated library of transposon consensus sequences, and cannot identify transposition events *de novo*. The TEMP software package is freely available at github: https://github.com/JialiUMassWengLab/TEMP.git, or the TEMP webpage: http://zlab.umassmed.edu/TEMP/.

Material and Methods

Sequence mapping and the input files for TEMP

TEMP takes input files in the BAM format obtained by mapping sequencing reads to a reference genome. Throughout this article we used BWA (v0.6.1-r104) (H. Li & Durbin, 2010) as the mapping software and *Drosophila melanogaster* dm3 as the reference genome for mapping. Mapping was done using the BWA aln algorithm with command line options -n 3 -l 100 -R 10000, which allows for 3 mismatches. Other input files required by TEMP are transposon consensus sequences, which can be downloaded from Repbase (Version 17.07, http://www.girinst.org/repbase/), and RepeatMasker files containing the annotated TEs in the reference genome, which can be downloaded from the UCSC Genome Browser (http://genome.ucsc.edu/).

The TEMP method for identifying TE insertions and absence

In order to detect a TE insertion, TEMP first identifies all discordant read pairs (**Figure 2.1**), with one uniquely mapped read (the anchor read, or anchor) and a second read that is unmappable or maps to multiple distant locations. Those non-uniquely mapping reads are then compared to a library of consensus TE sequences. The TE to which the read maps with fewest mismatches determines the type of the TE insertion. For example, if the TE-mapping read maps to the *P-element* sequence then it is likely that there is a *P-element* insertion in the vicinity the anchor. TEMP infers the orientation of the insertion by examining the

genomic strand of the anchor and the transposon strand of the TE-mapping read. (**Figure SII-1**). A single read pair is usually insufficient for inferring the precise junction. Therefore TEMP first attempts to identify a genomic interval that includes the junction, called the interval estimate. This estimate is based on the average insert size of the sequencing library. The junction must be located in the interval beginning at the end of the anchor and extending into the genome by the length of the average insert size. The reads that support the same insertion event (i.e., the same type of TE, in the same genomic strand and with interval estimates that overlap by at least one nucleotide) are clustered and their intersecting region provides a refined interval estimate (**Figure SII-2**).

To detect TE insertions that are present in the reference genome but absent in the sample genomes, TEMP first identifies all read pairs for which the distances between the two genome-mapping reads are significantly longer than average insert size (but less than 10k bps to avoid mapping artifacts) and then examines whether the intervening genomic region spans one or more known TEs as annotated in the reference genome. In order to prevent false positives, we require that both reads are uniquely mapped to the reference genome and that the distance between the two reads (after subtraction of the excised TEs) is consistent with the average insert size of the library. Read pairs that support the same event are clustered (**Figure 2.1b**). These structural alterations could reflect strain-specific excision of DNA elements, which move by a cut and paste

mechanism, or could reflect polymorphic DNA or RNA elements insertions that are specific to the reference genome. TEMP cannot distinguish between these alternatives unambiguously, but if the transposon is an RNA element transposing by the “copy and paste” mechanism and its frequency is close to zero in most of the sample genomes, it is most likely that the element is a polymorphic insertion in the reference genome.

Estimation of new junctions and transposition frequencies in a population

Based on the interval estimates obtained in the previous step, TEMP attempts to determine the new junctions created by transposition events up to base-pair resolution (base estimates) by taking advantage of reads that start in genomic sequence but are interrupted by transposon or non-contiguous genomic sequence (soft-clipped reads; **Figure 2.1 c, d**).

For insertions, TEMP first extends the interval estimates obtained in the detection step by 20 bps in both directions. Soft-clipped reads that map within the extended interval are identified. For each such read, TEMP determines if the clipped portion of the read can be confidently explained by the insertion event (i.e., the clipped sequence corroborates the type and direction of the TE insertion determined by the previous step). We require the clipped portion to be at least 7 nucleotides long and map perfectly to the appropriate TE sequence. When multiple junction estimates are identified, TEMP chooses the one supported by

the most reads. We use a similar approach to estimate the junctions of TEs that are absent in the reference genome. Soft-clipped reads that map near the annotated boundaries of the absent TE are identified and the clipped portion of each read is examined to ensure that it maps to the sequence on the other side of the transposon. We note that base estimates of the junctions are strand-specific as the soft-clipped reads are mapped to only one strand of the genome. When the base estimates are not available, TEMP uses the midpoint of the interval estimates and the annotated TE boundaries as surrogates for insertion and absence, respectively.

For each detected presence or absence of transposon insertion, TEMP first compiles all the read pairs that support the transposition event, which include the discordant read pairs that define the transposition event and the soft-clipped reads that delineate its junctions with genomic sequence. TEMP also keeps track of another set of read pairs that originate from the genomes where the transposition event does not happen; these read pairs span the estimated junctions of the transposition (**Figure 2.1**). TEMP computes the ratio $T/(T+R)$ as an estimate of the population frequency of the transposon, where T stands for the total number of read pairs that support the presence and R stands for the total number of read pairs that are consistent with the absence of the transposon insertion.

The workflow of TEMP is represented in **Figure SII-3 & SII-4**

Simulation analysis

In each experiment, 50 insertions and 50 excisions were randomly placed across chromosome arm 2L of the *Drosophila melanogaster* reference genome. For each simulated insertion, the TE family and insertion site coordinate were selected randomly. The entire sequence of the chosen TE was then inserted at the selected coordinate. For each simulated excision, an annotated transposon (as annotated in the output of the RepeakMasker program) was randomly selected and the entire sequence was deleted. The insertion and deletion operations were carried out using a genomic structural variation simulation package named *RSVSim* (v1.1.1) (Bartenhagen & Dugas, 2013). Simulated read pairs with read length of 90 nucleotides (nt) following a normal distribution of insert sizes (500 ± 50 nt) were then generated from the simulated genome obtained in the previous step with four different sequencing depths (5X, 10X, 20X and 40X) using a profile-based Illumina paired-end reads simulator named *pIRS* (Hu et al., 2012). We used *pIRS* v1.1.0 with options -l 90 -m 500 -v 50 -e 0.0001 -a 0 -g 0, which simulated 90-nt long reads, with mean insert size set at 500 nt, standard deviation of insert sizes at 50 nt, sequencing error rate at 0.0001, no insertions or deletions in the reads, and no GC bias. To simulate various population frequencies of transposition events, we mixed reads generated from the simulated genome with reads generated from the reference genome at

appropriate ratios. Finally we mapped all the reads to dm3 using BWA and fed the mapping results to TEMP and other algorithms to evaluate their performance. The above procedure was repeated 100 times to obtain 5,000 simulated insertions and 5,000 simulated excisions.

To compare TEMP with other algorithms, we generated datasets by combining five independently simulated *Drosophila* chromosome 2L arms. Each simulated chromosome arm was generated as described above, and pair-end reads were simulated at 5X coverage. Each simulated dataset hence contained reads originating from the five simulated chromosome arms with an apparent coverage of 25X, and the process was repeated 20 times. We compared TEMP with PoPoolationTE (Kofler et al., 2012), RetroSeq (Keane et al., 2013), VariationHunter (Hormozdiari et al., 2010), and GASVPro (Sindi et al., 2012) on these datasets. The results are summarized in Table2.1. We evaluated PoPoolationTE (v1.02, https://code.google.com/p/popoolationte/), RetroSeq (https://github.com/tk2/RetroSeq), VariationHunter CommonLaw (v0.04, http://variationhunter.sourceforge.net/Home), and GASVPro-HQ (2013 Oct Release, http://code.google.com/p/gasv/). For PoPoolationTE we followed the typical workflow described at https://code.google.com/p/popoolationte/wiki/Workflow and used the parameters therein. For RetroSeq, we used the BAM format alignment file produced by the BWA aln algorithm as the input and chose the same parameters as described in

the tutorial https://github.com/tk2/RetroSeq/wiki/RetroSeq-Tutorial. For VariationHunter-CL, we first mapped paired-end reads to the reference genome (dm3) using mrfast (v2.6.0.1, http://mrfast.sourceforge.net/) with parameters -min 400 –max 600 -e 3 (which allows for 3 mismatches and defines concordant insert sizes as between 400nt and 600nt), and then ran VH and multiInd_SetCover with default parameters. Structural variations supported by fewer than 8 reads were discarded. For GASVPro-HQ, we used the BAM format alignment file produced by the BWA aln program as the “high quality unique mapping BAM file” and default parameters. GASVPro produced a large number of predictions. We ranked the predictions by log-likelihood ratio and kept the top 250 predictions.

A simulated insertion was correctly recovered if the interval estimate given by TEMP included its true junction with the genomic DNA, and if the transposon family and direction of the insertion were determined correctly. A simulated excision was correctly recovered if TEMP reported the absence of the corresponding transposon. For a simulated transposition event, we considered its junction correctly identified if the base estimate TEMP reported lay within 5 nt of the true junction.

Testing TEMP on pair-end sequencing data from the 1000 genomes project

BAM files containing alignments to the GRCh37 (hg19) human reference genome for four individuals (NA18517, NA19240, NA12156, NA12878) were

downloaded from the data portal of the 1000 genomes project and merged to mimic a pooled sequencing dataset. We then ran TEMP on this dataset and predicted presence and absence of TE insertions with frequency greater than or equal to 20% and covered by more than eight reads. TEMP predictions were compared with previously reported insertions and deletions involving these four individuals as deposited in structural variation database DGV (Database for Genomic Variants, http://dgv.tcag.ca/dgv/app/home). Note that the structural variations in DGV include all types of changes in genomic DNA regardless whether they are caused by TEs.

Hybrid dysgenesis population analysis

The small RNA sequencing and genomic deep sequencing datasets were downloaded from NCBI SRA database (SRP007937) and processed and analyzed as described in Khurana *et al.,* 2011. We define parental transposons as TE insertions with population frequencies greater than 10% in at least one of the parental strains (w^1 or Harwich). The frequency change of a parental transposon is defined by: $FC = F–(H+W)/2$, where F, W, H represent frequency of the transposon in the w^1 x Har; *2-4 day* F1 population, the w^1 population, and the Harwich population, respectively. The junction spanning small RNA reads need to be at least 21 nt long and map perfectly across the genome-transposon junction. We use the piRNA cluster annotation by Brennecke *et al.* (Khurana et al., 2011), which includes 141 clusters in total (excluding the chrX_TAS cluster),

occupying 4,924,944 bp of the dm3 genome. piRNA clusters are the genomic loci from which precursor piRNA transcripts are produced.

The *Drosophila* Genetic Reference Panel (DGRP) datasets

We downloaded genomic deep sequencing data for 53 DGRP inbred lines (Mackay et al., 2012) from NCBI SRA (**Table SII-1**). Except for lines RAL-362, RAL-765 and RAL-517, the other 50 lines each had >20X sequencing coverage. We included those three lines with <20X coverage because they were the only lines with RNA-seq data. We mapped the reads to dm3 with the BWA aln algorithm, allowing for 3 mismatches and then ran TEMP on the BWA output files in the BAM format.

For TE insertion distribution analysis, 11,311 insertions that had frequencies greater than 80% in at least one of the inbred lines were chosen. We profiled the number of insertions in each of the 5 genomic features: promoters (2 kb upstream of an annotated TSS), exons (Flybase annotation), intron/UTR regions (regions within annotated genes but not in exons), intergenic regions (regions more than 2 kb from any annotated genes) and piRNA clusters for each TE family. A binomial test was performed to assess the statistical significance of enrichment or depletion for each TE family in each of the five genomic features and the Benjamini–Hochberg procedure (Benjamini & Hochberg, 1995) was used for multiple testing corrections. Only enrichments and depletions with q-values

lower than 0.15 are shown in **Table SII-2**. The annotation for genes and exons were obtained from FlyBase (Release 5.45) and the annotation for piRNA clusters was from Brennecke *et al*. (Khurana et al., 2011) as described above.

For RNA-seq data, we downloaded seven datasets involving the three lines RAL-362, RAL-765 and RAL-517 and four progeny populations (**Table SII-1**). The samples involving two lines were F1 samples (i.e., the progeny of the two indicated lines separated by "x"). We mapped the reads to the reference genome using Tophat (v2.0.8b with default parameters) and then used Cufflinks (v2.1.1 with default parameters) to compute the expression level for each gene (in FPKM). Thus for the three parental lines (RAL-362, RAL-765 and RAL-517), both genomic sequencing and RNA-seq data were available. To find the TE insertions that could potentially affect gene expression, we looked for genes that had 1) TE insertions with frequencies greater than 20% in their promoters, introns, exons or UTRs in only one of the three parental lines and 2) expression levels in that line were more than two fold higher or lower (with a pseudo count of 0.5 FPKM) than the corresponding expression levels in the other two lines where the insertion was absent. The 48 genes obtained are listed in **Table SII-3** along with their expression levels in all seven lines.

Sequence signature of TE insertions

We ran TEMP on all *Drosophila* genomic sequencing data we had and obtained 14,363 non-redundant insertion events with junctions on both strands detected. By calculating the difference between the coordinates of the junctions on two strands, we were able to estimate the length of target site duplications (TSD) for each such insertion. We investigated the nucleotide composition of the sequence around the junctions by extending 15 bp both upstream and downstream from the midpoint between the junctions of the two strands for each insertion. We then ran MEME on these 30-bp long sequences to report up to five most significant motifs with lengths of 4-15 nt. (MEM was ran with options -dna -mod zoops -nmotifs 5 -minw 4 -maxw 15 -pal, where dna and zoops indicate that there is zero or one motif site per input DNA sequence and pal indicates that we were looking for palindromic motifs.) This procedure was performed for each TE family to identify any sequence motifs that were enriched in the sequences surrounding the junctions. In the mono- and dinucleotide composition analysis, the same length (i.e., 30 bps) of flanking sequences (100 bps upstream and downstream of the junction) was selected as background, and the enrichment for each mono or dinucleotide was measured by the ratio of its frequency in the junction surrounding sequence over its frequency in flanking sequences.

Results

We first describe the general approach that TEMP takes for detecting TE polymorphisms and estimating their population frequencies. We then evaluate TEMP's performance on both simulated datasets and pooled human genome sequencing data. We compare TEMP with four other algorithms (PoPoolationTE, RetroSeq, VariationHunter, and GASVPro) on the simulated data. To showcase how TEMP can be applied to studying biological problems, we use TEMP to investigate the inheritance patterns of polymorphic transpositions in *Drosophila melanogaster* hybrid dysgenic strains. Finally we analyze the genomic sequencing and RNA-seq data of 53 lines in a wild *Drosophila melanogaster* population to learn about the molecular signatures of TE integration sites and potential molecular consequences of TE insertions.

Overview of the TEMP method

TEMP detects the presence or absence of TE insertions in a population of sample genomes using read pairs that are mapped discordantly on a reference genome. Discordant read pairs with one read mapped uniquely to the reference genome and the other read mapped to TE sequence indicate sample-specific TE insertions (**Figure 2.1a**). Sample-specific absence of TEs can be detected by looking for read pairs that are separated by a distance substantially longer than the average insert size of the library and span a TE presents in the reference genome (**Figure 2.1b**). As detailed in Materials and Methods, TEMP can identify

the presence and absence of TE insertions by identifying and sorting through discordant read pairs. TEMP then attempts to estimate the minimal genomic interval that includes the junction, called the interval estimate. For the insertions supported by sufficient numbers of reads, TEMP proceeds to refine the interval estimates to base-pair resolution by taking advantage of soft-clipped reads (**Figure 2.1 c, d**).

In order to estimate the population frequency of transposition events, TEMP assumes that 1) the pool of sample genomes is a faithful representative of the population from which it is drawn and 2) the number of read pairs supporting a transposition event is proportional to the frequency of the event in the pool of sample genomes. For each transposition event, TEMP keeps track of two sets of read pairs (including both discordant and soft clipped read pairs), one set originating from the genomes where the transposition is present (*T* pairs) and the other set from the genomes where the transposition is absent (*R* pairs). TEMP computes the ratio $T/(T+R)$ as an estimate of the population frequency of the transposon (see Materials and Methods for more details).

Assessment of TEMP performance on simulated and biological datasets

As there are no pooled sequencing datasets for which the population frequencies of polymorphic transposition events are known, we first evaluated the performance of TEMP on a simulated dataset. We randomly inserted and deleted TE sequences in chromosome arm 2L of the *Drosophila* reference genome with the *RSVSim* (Bartenhagen & Dugas, 2013) program and generated simulated reads from the simulated genomes using the *pIRS* (Hu et al., 2012) program. The simulated reads were mapped back to chr2L and TEMP was used to detect presence and absence of insertions, resolve the junctions, and estimate the population frequencies of the transposons.

The performance of TEMP depends on the sequencing depth as well as the frequency of the transposition (**Figure 2.2**). TEMP performs better at higher sequencing depth, in terms of higher detection rates (**Figure 2.2 a, b** solid lines), more accurate estimates of population frequency (**Figure 2.2 a, b** dashed lines), higher probability of discovering the junctions (**Figure 2.2 c, d** solid lines) and more correctly recovered junctions (**Figure 2.2 c, d** dashed lines). Frequency of the target transpositions has similar effects. Those instances of presence and absence of insertions with very low frequencies are often undetected and it is difficult to determine their precise junctions because there are only a few reads. False discovery rate (FDR) for TE insertion detection rises slowly with increasing sequencing depth and insertion frequency (**Figure 2.2 e)**. On the other hand, for

TE absence detection the FDR remains low and flat across the range of sequencing depth and transposon frequency (**Figure 2.2f**).

At 20-fold genome coverage, which is easily achievable with current technology even for large mammalian genomes, TEMP is able to detect more than 95% of the presence and absence of insertions with population frequencies exceeding 20%. The average error of frequency estimation is <10% for presence and <9% for absence across the entire frequency range. Among the base estimates of the junctions reported by TEMP, more than 95% of them are correct. These results demonstrate that TEMP is effective in detecting sample specific TE insertion and absence, estimating their population frequency with high accuracy, and pinpointing the precise junctions for some of the transposition instances across a wide range of sequencing depths and transposition frequencies.

We compared TEMP with four other algorithms on a simulated dataset that mimics a pooled sequencing library: PoPoolationTE (Kofler et al., 2012), an algorithm designed for detecting transposon insertions in pooled sequences; RetroSeq (Keane et al., 2013), designed for detecting transposon insertions in individual genomes; and two commonly used general-purpose structural variation discovery tools VariationHunter (Hormozdiari et al., 2010; Hormozdiari, Hajirasouliha, McPherson, Eichler, & Sahinalp, 2011), and GASVPro (Sindi et al., 2012). The results are presented in **Table 2.1**. TEMP achieved better

performance in detecting both presence and absence of TE insertion than the other four methods.

As a method designed for pooled sequencing data, PoPoolationTE performed worse than TEMP in terms of sensitivity (88.50% vs. 98.80%), precision (92.45% vs. 99.50%) and average error of estimating insertion frequency (8.77% vs. 7.27%). RetroSeq is specifically designed for detecting TE insertions, but since it is not intended for handling pooled sequencing data, it does not estimate insertion frequency. RetroSeq achieved a low sensitivity (71.82%) and a high precision (93.95%). Neither PoPoolationTE nor RetroSeq is designed for detecting sample-specific absence of TEs in the reference genome. In comparison, TEMP can detect transposon absence with high sensitivity (93.09%), precision (98.64%) and low error in frequency estimate (7.25%). The two general-purpose structural variation detection algorithms VariationHunter and GASVPro could detect transposon absence, although with lower sensitivity and precision than TEMP (**Table 2.1**). Neither algorithm could detect TE insertion, nor are they designed to estimate transposon frequency. GASVPro produced many false positives in detecting TE absence.

We also assessed TEMP's ability in detecting polymorphic TE transpositions in human genomes using whole-genome datasets generated by the 1000 Genomes Project (Consortium et al., 2012). We pooled the reads from four individuals and

ran TEMP to detect TE insertion and absence relative to the reference genome (GRCh37, see **Table SII-4** for details of TEMP predictions). Since there is little information on experimentally validated TE presence and absence genome-wide, we used structural variations deposited in the Database of Genomic Variants (DGV) for evaluating TEMP (J. R. Macdonald, Ziman, Yuen, Feuk, & Scherer, 2013). Overall, 363 out of the 536 (67.7%) of insertions predicted by TEMP overlapped with insertions for these individuals in DGV and 423 out of the 1593 instances (26.5%) of absence predicted by TEMP overlapped with the deletions in DGV. The percentage of predictions that overlapped insertions and deletions in DGV went up to 81.5% and 95.5% respectively if we considered all individuals deposited in the DGV. **Table SII-4** also lists which DGV insertions or deletions that TEMP predictions matched. Thus TEMP works effectively in detecting transposition events in human genomes.

We evaluated the time complexity of TEMP on the same human whole-genome sequencing dataset. The combined dataset is equivalent to ~12X coverage of the human genome and the insertion analysis took 1,382 minutes on a Dell M605 node with 2 quad core AMD Opterons. The absence analysis took 721 minutes on the same machine.

Identifying potentially selected TE insertions from pooled sequencing of hybrid-dysgenic population

We used TEMP to analyze the pooled genomic sequencing data of a wild type strain of *Drosophila melanogaster* (Harwich or Har in short), a lab strain (w^1), and the offspring populations from crossing Har males with w^1 females. When Har females are mated with *w1* males, the first-generation offspring (F1) are normal; however, when Har males are mated with *w1* females, the offspring suffer from widespread TE transpositions, genomic instability and are initially sterile, a phenomenon known as hybrid dysgenesis (Bucheton, 1973; 1979; Hiraizumi, 1971; Kidwell, 1985) (**Figure 2.6**). As the surviving female dysgenic flies age, they partially recover from the dysgenic phenotypes and begin to produce viable offspring, a change thought to be the result of *de novo* piRNA production in the ovaries (Khurana et al., 2011).

We used TEMP to detect TE insertions relative to the reference genome and estimate their frequencies in each of the parental and progeny populations (**Table SII-5**). This enables us to find insertions that show inheritance patterns potentially under adaptive selection. For a neutral insertion polymorphism, its population frequency in the progeny population should be close to the arithmetic mean of the frequencies in the two parental populations if the inheritance obeys Mendelian segregation. We therefore defined *frequency change* using a simple formula $FC = F–(H+W)/2$, where F, H, W denote the population frequency of a TE insertion in the dysgenic F1 population, the Harwich population, and the w^1

population, respectively. A large positive value of frequency change suggests positive (adaptive) selection whereas a large negative value suggests negative (purifying) selection.

We computed the frequency change for each parental TE insertion (defined as the insertions whose frequencies exceed 10% in at least one of the parental populations) (**Figure 2.3a**). As expected, the vast majority of parental insertions have negative but close to zero frequency changes in the F1 population (**Figure 2.3a**), suggesting that they were under weak purifying selection. The most critical challenge facing the hybrid dysgenic flies was coping with hyperactive transpositions and any trait that helped suppress TE mobilization could be potentially selected for. Insertion of a TE into piRNA clusters can lead to production of piRNAs whose sequences are complementary to the TE, and these piRNAs can in turn silence the corresponding transposon genome-wide (Aravin, Hannon, & Brennecke, 2007; Ghildiyal & Zamore, 2009; Khurana & Theurkauf, 2010). Indeed, among insertions whose frequencies increased by 30% or more in the F1 population ($FC \geq 0.3$), there were more insertions residing within piRNA clusters than expected (p-value = 5.38E-4, hypergeometric test). In contrast, among insertions with $FC \leq -0.3$, there were fewer of them than expected in piRNA clusters (p-value = 8.02E-5, hypergeometric test) (**Figure 2.3a**). We also analyzed the germline DNA isolated from the ovaries of the second-generation progenies (F2) produced by backcrossing F1 dysgenic females to w^1 males.

Again, we computed the *FC* for each parental TE insertion, i.e., those insertions whose frequency exceeded 10% in either F1 or w^1. These F2 females did not suffer from hyperactive TE movement; accordingly, our data support that there were fewer insertions under negative selection than their parents (14.05% with *FC* ≤ -0.3 in F2 vs. 19.09% in F1; p-value = 4.90E-5, X^2-test). Moreover, there was no enrichment for TE insertions in piRNA clusters (p-value = 0.53, hypergeometric test), consistent with the notion that such insertions would not confer significant selective advantages in non-dysgenic individuals (**Figure 2.3b**). We note that according to the Wright-Fisher model with a population size of 100 (200 chromosomes), the probability of *FC* ≤ -0.3 or ≥ 0.3 or more extreme is smaller than 1E-15 (**Table SII-6**). Therefore the sites with *FC* ≤ -0.3 or ≥ 0.3 are likely under selective pressure. Moreover, at 20X sequencing depth TEMP's false discovery rate is 1.17% for sites with frequency at 0.3 (**Figure 2.2e**). Therefore most of the sites with *FC* ≤ -0.3 or ≥ 0.3 represent actual change, not detection error.

We were able to resolve the junctions for one of the insertions that were both strongly selected for and lie within a piRNA cluster. A *pogo* insertion at position 2,378,892-2,378,894 of chromosome arm 2R (within the piRNA cluster 42AB) had a frequency of 96.77% in the Har population and was absent in the w^1 population. In the F1 hybrid dysgenic population the frequency of the same insertion is 88.24%, which far exceeded what would be expected from Mendelian

inheritance-- suggesting that it was under strong positive selection. Evidently, in addition to the F1 embryos that lacked a *pogo* insertion in both alleles, some of the F1 embryos that were heterozygous for this *pogo* insertion did not mature to adulthood. As shown by the large number of piRNA reads that mapped across the unique junctions produced by the insertion, this insertion led to *de novo* production of piRNAs and probably helped repress transposition of the *pogo* element, giving the individuals a selective advantage (**Figure 2.3c**). Interestingly, the same insertion also exhibited higher than expected frequency in the F2 backcross progeny, suggesting a persisted adaptive selection at this locus even though the backcross did not induce hybrid dysgenesis (**Figure 2.3c, bar plots**).

Sequence signatures and potential effects on gene expression of TE insertions

The *Drosophila melanogaster* Genetic Reference Panel (DGRP) is a community resource of inbred lines of fruit flies derived from a wild population (Mackay et al., 2012). In freeze 1.0, the genomes of 168 inbred lines have been sequenced and the sequencing data are publicly available. Moreover the RNA-seq data for three of these lines are also available, as well as the RNA-seq data on four progeny populations of these three lines. We selected 53 lines with the highest genome sequencing coverage and applied TEMP to detect presence and absence of TE insertions. TEMP detected in total 11,316 instances of presence and 1,378

instances of absence of transposons that had frequency greater than 80% in at least one line (**Tables SII-7**). The distribution of TE insertions across the genome showed that most TEs insertions are enriched in intronic and intergenic regions and depleted in exonic regions (**Table SII-2**). This is consistent with a recent report on a related dataset (Cridland, Macdonald, Long, & Thornton, 2013).

We also used TEMP to pinpoint the junctions in both the DGRP datasets and the hybrid dysgenic datasets. TEMP reported the positions for 14,363 non-redundant junctions at base pair resolution, which enabled us to investigate the sequence signatures near the TE insertion sites including the length of target site duplications (TSDs), the dinucleotide composition of target site sequences, and potential sequence motifs at target sites that may reveal the sequence preferences of the integrases (**Figure SII-5**).

There were 44 TE families for which we detected more than 50 non-redundant target sites. Most of these TE families exhibited narrow TSD length distributions (**Figure SII-6**). Strikingly, TEs in the same super-family showed very similar TSD length distributions (**Figure 2.4a**) except for DNA elements. TEs in most LTR/Gypsy super-families showed 4-nt-long TSDs and nearly all TEs in the LTR/Copia and LTR/Pao super-family produced 5-nt-long TSDs. This interesting pattern probably reflects the evolutionary relationship among TEs, as integrases encoded by the TEs within the same super-family are more likely to share similar

sequences and functional features (Linheiro & Bergman, 2012; Nefedova, Mannanova, & Kim, 2011). LINE elements had much longer TSDs and much broader TSD distributions compared with other retro-transposon super-families (**Figure 2.4a**), which is consistent with previous findings about the L1 element in the human genome (Lee et al., 2012; Szak et al., 2002).

We also discovered that the genomic sequences around predicted insertion junctions (±15 nt) of many TEs are AT rich, with the AT and TA dinucleotides being most prevalent (**Figure 2.4b**). The enriched sequence motif around the junctions is a simple dinucleotide repeat for many TEs (**Table SII-8**), which is consistent with the mono- and dinucleotide composition analysis. As exceptions, we detected high-information-content motifs around the insertion sites of the *hopper*, *1360*, *Tirant* and *Transpac* elements, suggesting that their integrases or transposases may possess sequence specificity (**Figure 2.4c**).

TE transposition is one of the main sources of genomic variability. Relating transposition polymorphisms to variations at phenotypic and molecular levels is crucial for understanding how transposition shapes the genomic landscape and contributes to evolution (Daborn et al., 2002; Tsuchiya & Eulgem, 2013; X. Wang, Weigel, & Smith, 2013b). By integrating RNA-seq data for 3 DGRP lines with their respective genomic sequencing data, we searched for TE insertions that are likely to affect gene expression. More specifically, we looked for

insertions that were in the promoter region or the gene body for which expression level of the affected gene changed by more than two fold. We identified 48 insertions that were associated with changes in gene expression. For example, *nrm* encodes a protein important for synaptic target recognition, and a *P-element* insertion at the promoter of *nrm* that is unique to strain RAL-517 is associated with a more than 3-fold increase in *nrm* expression (**Figure 2.5 a, b**). Crosses with a strain showing lower expression produced progeny with intermediate expression levels, strongly suggesting that increased expression is inherited in the F1 generation (**Figure 2.5b**). The correlation between TE insertion and altered gene expression suggests a causal relationship, although other background variants can also contribute to the change in expression. Using TEMP to detect TE insertions and estimate their frequencies genome-wide, users will be able to correlate transposition polymorphisms with phenotypes and biological processes and identify candidate sites for experimental validation.

Discussion

Transposition of TEs is a widespread phenomenon that destabilizes the genome, but may also produce beneficial genetic diversity. The rapid development of high-throughput sequencing techniques offers unprecedented opportunities for detecting TE transpositions in a variety of samples. We described TEMP, an

algorithm that can detect TE insertion and absence, pinpoint their junctions with genomic DNA at base pair resolution, and estimate their frequencies in the population. Our analysis on both simulated and biological datasets demonstrates that TEMP is a reliable and useful tool for studying TE transpositions at both population and molecular levels and can be applied to a wide variety of datasets to accomplish quantitative analysis and generate testable hypotheses.

One limitation of TEMP is that it requires a curated library of transposon consensus sequences, namely the RepBase (Jurka et al., 2005), and cannot identify transposition events *de novo*. Thus for a newly sequenced genome, one first needs to apply *de novo* repeat identification algorithms such as RECON (Bao & Eddy, 2002) or RepeatScout (Price, Jones, & Pevzner, 2005) to build a library of transposon consensus sequences, before one can use TEMP to identify the presence and absence of these transposons in populations of the same species. One idea that may aid such an analysis is to perform *de novo* assembly of the reads of the test population that do not map to the reference genome. This may yield longer sequences which, once aligned back to the reference genome, can reveal insertion or deletion junctions. Instead of transposon consensus sequences, the PoPoolationTE algorithm uses a database of many diverged sequences for each TE family. PoPoolationTE therefore may have a higher sensitivity than TEMP with detecting highly diverged TE copies.

The ability of TEMP to detect insertions genome wide and identify junctions at base pair resolution for thousands of sites enabled us to better understand the molecular mechanisms of TE integration. Linheiro and Bergman examined TSD lengths and target site motifs on 166 DGRP datasets (Linheiro & Bergman, 2012). Among the 25 TEs whose TSD lengths were reported in both studies, 19 of them had exactly the same TSD length (**Table SII-9**). Linheiro and Bergman treated paired-end sequencing data as independent reads and used only junction-spanning reads to identify TE insertion target sites, possibly restricting the sensitivity of their method. Indeed, for each of the 6 TEs that the two studies disagreed, Linheiro and Bergman identified fewer than 20 non-redundant insertion sites, while we identified 50 or more sites. We also compared the target site motifs identified in the two studies and they mostly agree.

Transposition is proposed to produce both beneficial and deleterious changes in genome organization. To determine the utility of TEMP in defining the molecular consequences of transposon insertion, we applied TEMP to analyze dysgenic hybrids as well as 53 strains derived from independent wild populations of *Drosophila* and identified over 14,000 high frequency insertions at base pair resolution. Analysis of RNA-seq data from three of these strains, and the F1 progeny of inter-strain crosses, showed that many of these insertions were linked to heritable changes in gene expression (**Figure 2.5b**). These findings raise the possibility that strain specific transposon insertions that modify gene expression

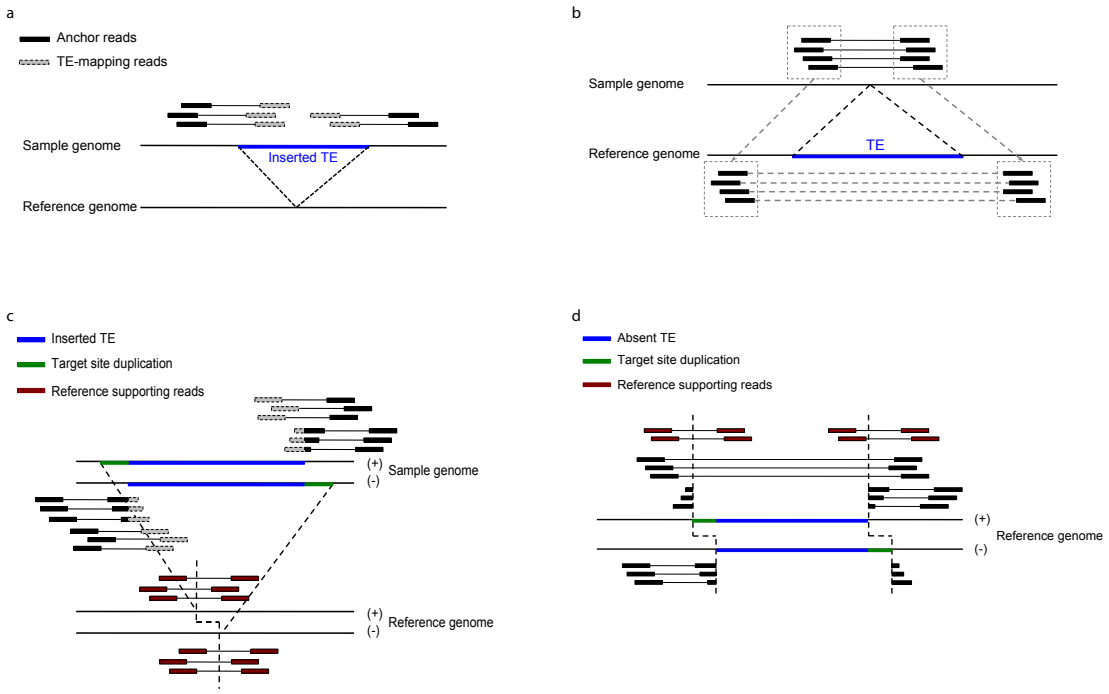
can sweep though populations, perhaps because they provide a reproductive benefit. This can be directly tested by crossing strains and following the inheritance patterns of specific insertions using TEMP and measuring gene expression using RNA sequencing.

Our analysis of hybrid dysgenesis shows that transposition can also alter expression of small non-coding RNAs. Transposons are silenced by piRNAs that are deposited in the oocyte. In the early embryo, the piRNA pool is therefore derived exclusively from the maternal genome. Hybrid dysgenesis is triggered during crosses in which the sperm carry a transposon that is not represented in the maternal genome. Transposon activation in the hybrid germline leads to adult female sterility. We previously showed that paternal introduction of *P-element* transposons activated both the invading *P-element* and resident transposons that were shared by the maternal and paternal genomes. Remarkably, the dysgenic F1 females regained fertility with age, as they silenced *P-element* and resident elements. Furthermore we demonstrated that this was linked to accumulation of *new* transposon insertions (i.e., not in either of the parental genomes) in the heterochromatic clusters that produced piRNAs, and that these insertions were the source of novel piRNAs that appeared to enhance silencing. Here, we used TEMP to estimate the frequencies of *existing* TE insertions in parental and progeny populations, and found that TE insertions within piRNA clusters were under positive selection in F1 dysgenic females. This finding, along with our

earlier study, indicates that both *de novo* and inherited transposon insertions into piRNA clusters are under positive selection in dysgenic hybrids, where they appear to enhance silencing by promoting piRNA production.

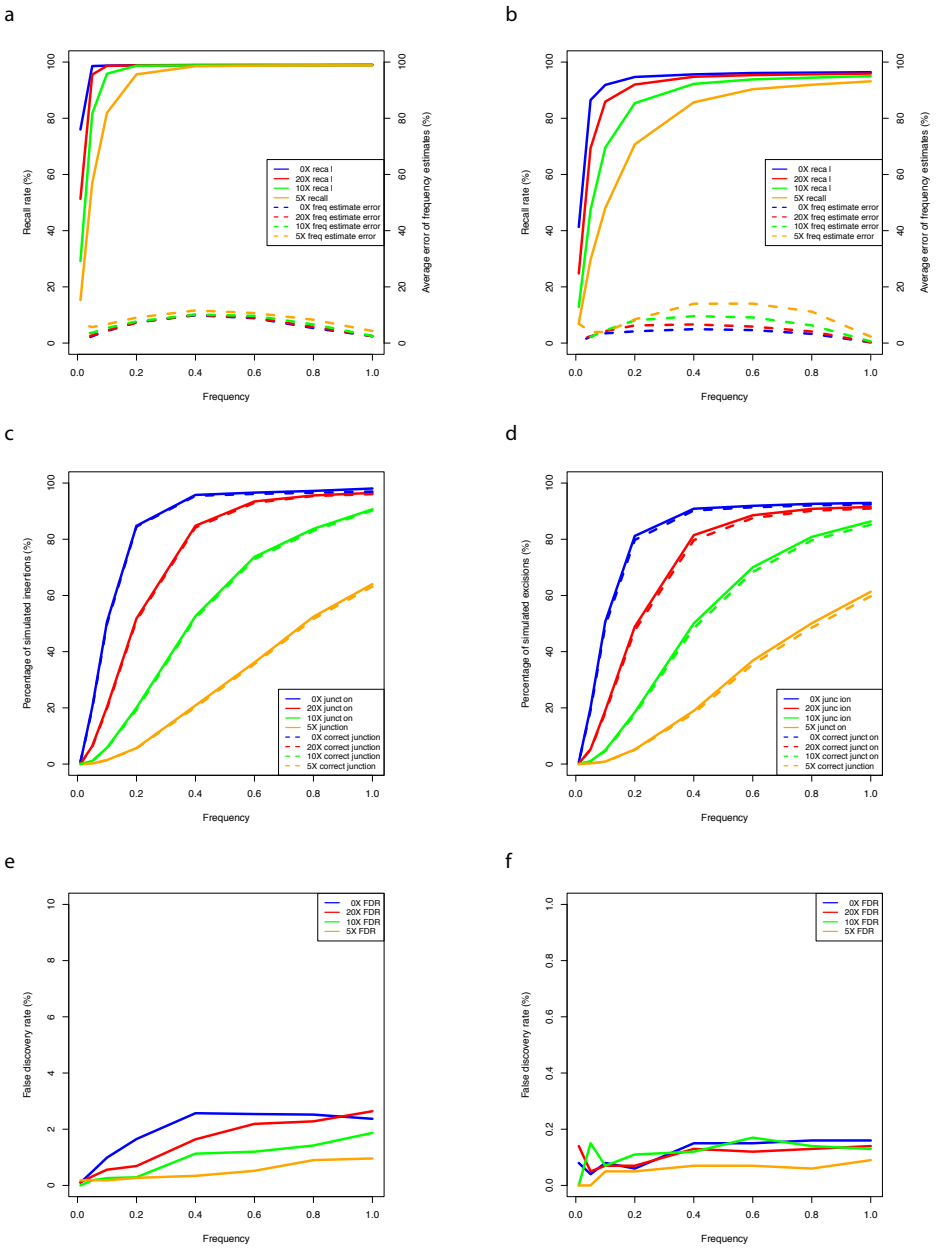
Our studies thus show that transposition can alter both coding and non-coding RNA expression, and suggest that these modifications can generate beneficial genetic variation. The paradigm of sequencing parental and progeny populations, estimating the population frequencies of the transposition polymorphisms with TEMP, and then identifying potentially selected polymorphisms can be applied to a wide range of systems to study the inheritance of transposition polymorphisms and their biological consequences.

Figure 2.1



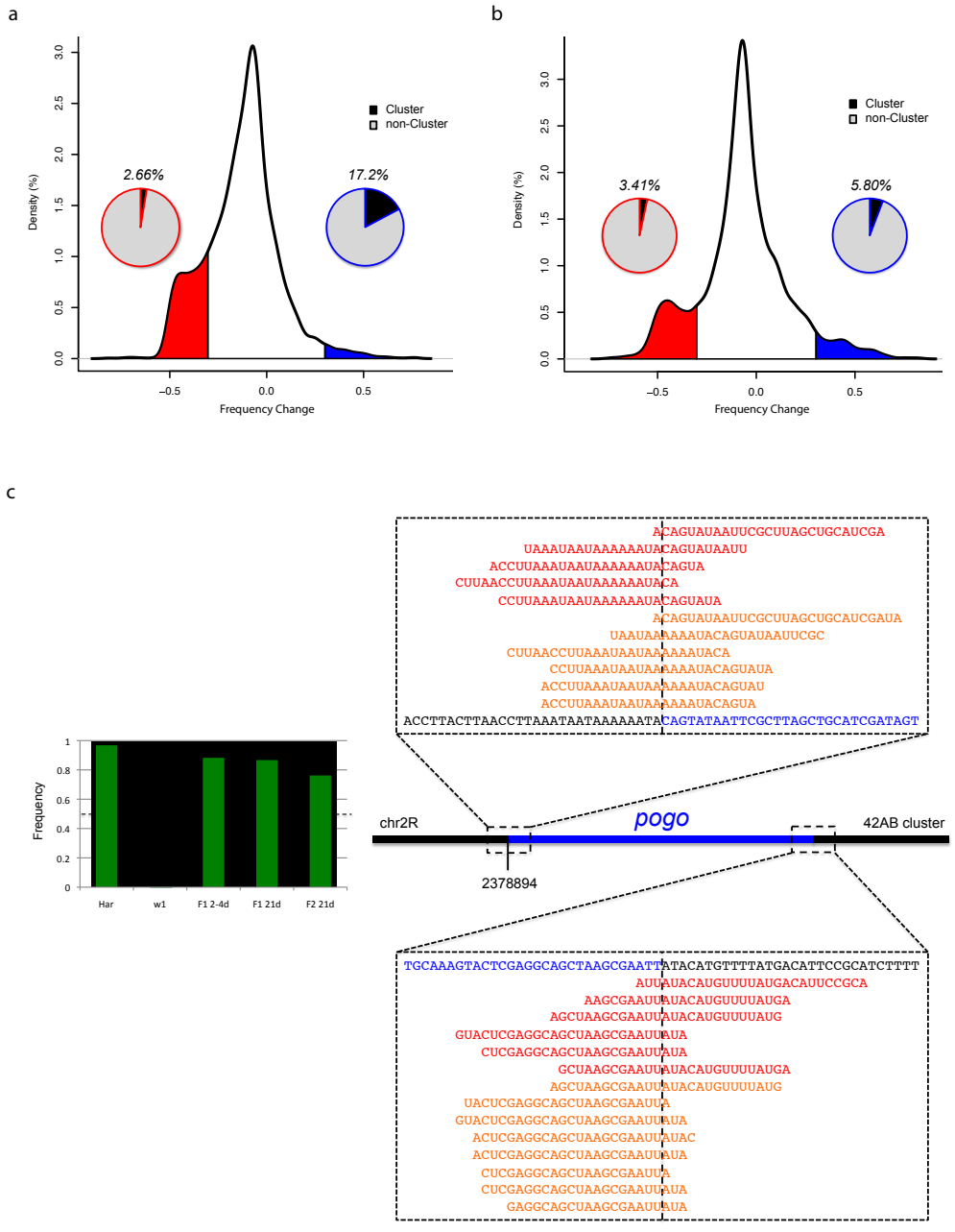
Diagrams depicting how TEMP detects presence (a) and absence (b) of insertion events and estimates junctions at base-pair resolution for presence (c) and absence (d) of insertion events.

Figure 2.2



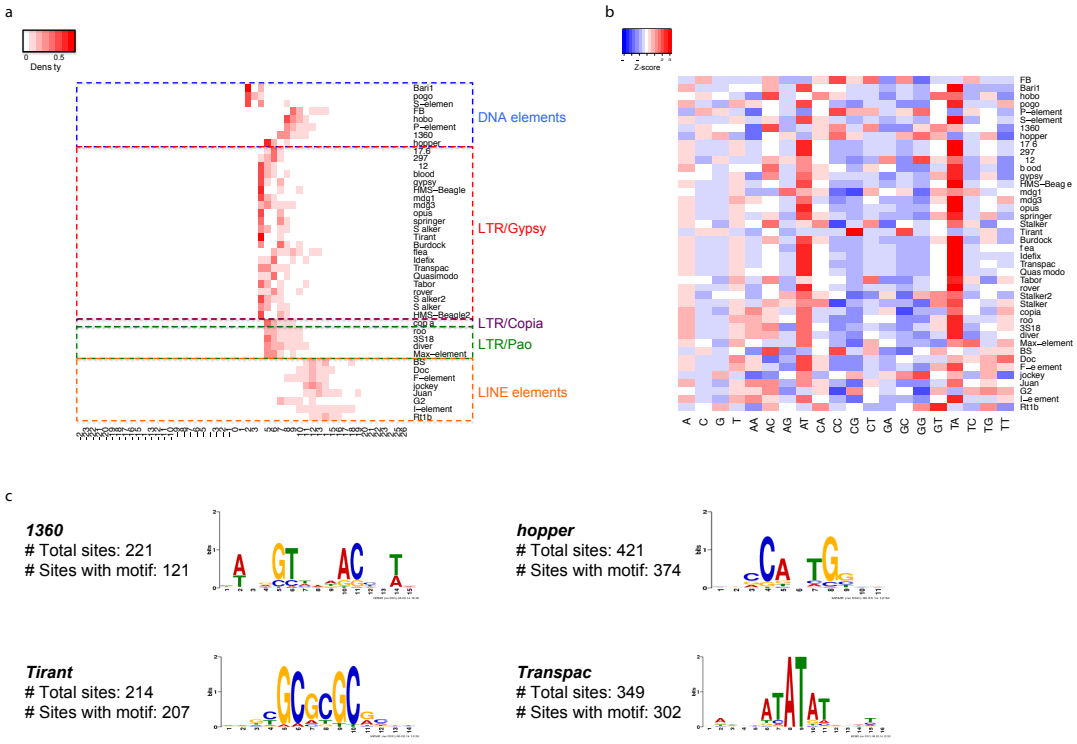
Evaluation of TEMP performance on a simulated dataset and the effects of sequencing depth and the population frequencies of the transposition events on the performance. The sequencing depth is color coded, with blue, red, green and orange denoting coverage 40X, 20X, 10X and 5X, respectively. Detection recall rates (solid lines) and average errors of frequency estimates (dashed lines) are plotted against population frequencies for presence (**a**) and absence (**b**) of insertion events. Percentage of transposition events for which TEMP identified junctions (solid lines) and for which TEMP *correctly* identified junctions (dashed lines) are plotted against population frequencies for presence (**c**) and absence (**d**) of insertion events. FDRs for detecting presence (**e**) and absence (**f**) of insertion events are plotted against population frequencies.

Figure 2.3



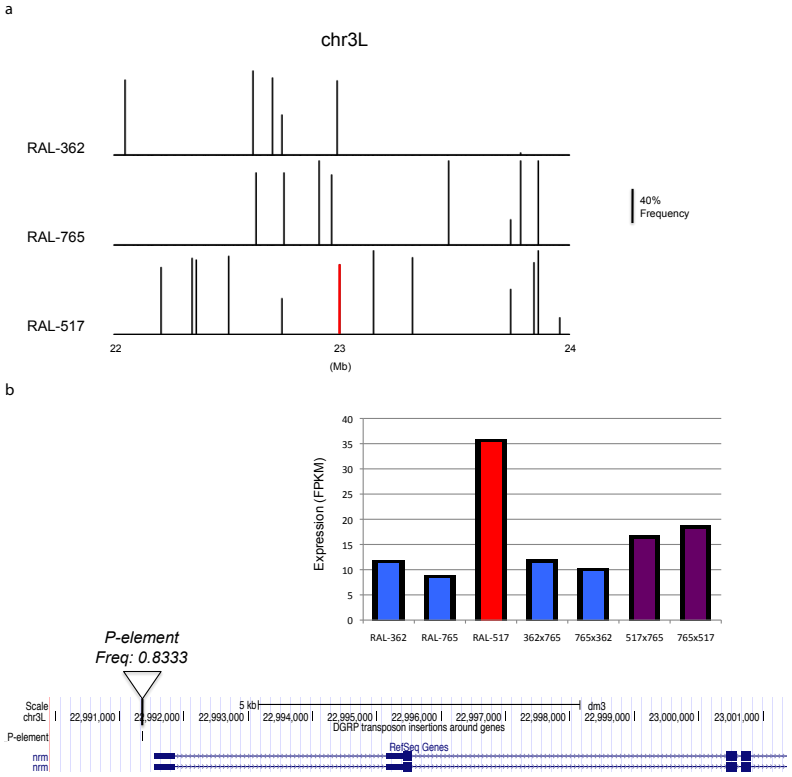
(**a**) Distribution of selection strength acted on parental TEs. Positively selected TEs (blue shaded region) shows enrichment for piRNA cluster residing TEs whereas negatively selected TEs (red shaded region) shows depletion for piRNA cluster residing TEs. The pie charts represent the percentages of piRNA cluster insertions (labeled) among the positively (or negatively) selected TEs. (**b**) Same as (a) except for F2 backcross progeny. (**c**) A *pogo* insertion within the 42AB piRNA cluster is under strong positive selection. It led to the *de novo* production of piRNAs as demonstrated by piRNA reads that span the insertion junctions in two F1 populations, *w1 X Har* 2–4 days (red) and *w1 X Har* 21 days (orange). The bar plots on the left show the frequency of the *pogo* insertion in the parental, F1 and F2 populations.

Figure 2.4



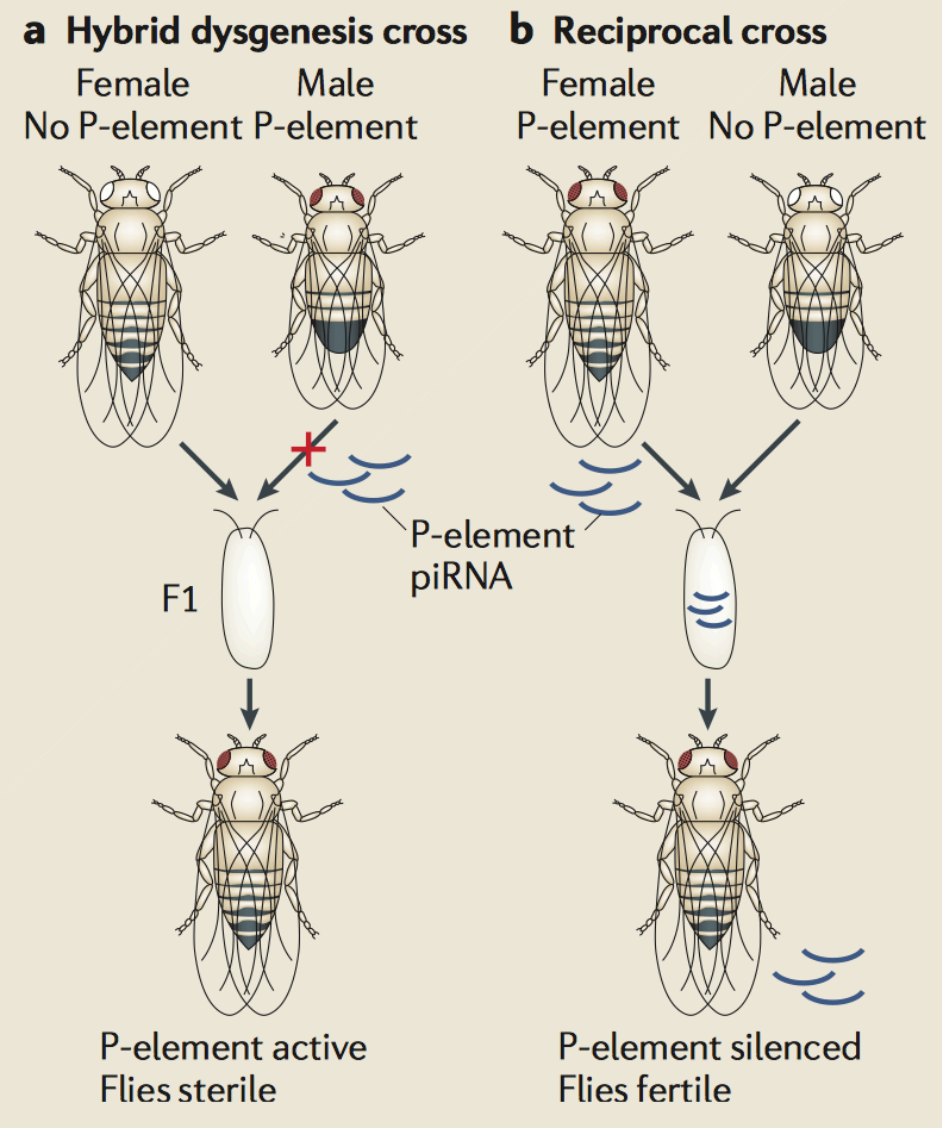
(**a**) Length distribution of TSDs (depletions) for TEs. The TEs are grouped according to families. Negative values on the x-axis denote the length of target site depletions. (**b**) Dinucleotide composition around target sites. Each row is normalized and the Z-score for each entry is color coded with red represents enrichment and blue represents depletion. (**c**) Sequence motifs for TE elements *1360*, *Tirant*, *Transpac* and *hopper*.

Figure 2.5



(**a**) Distribution of unique TE insertions of three DGRP strains in a region of chromosome arm 3L. The heights of the bars are proportional to the estimated population frequencies. The red bar in strain RAL-517 near 23Mb is a *P-element* insertion at the promoter of the *nrm* gene and its detailed view is presented in (b). (**b**) A *P-element* insertion at the promoter regions of the *nrm* gene in strain RAL-517 is correlated with a 3.65-fold increase in its expression level. The bar plot shows the expression level of *nrm* in the three lines as well as the four F1 progeny samples. The expression of *nrm* is higher for the progeny populations of RAL-517 (purple bars) than the progeny produced by crossing the other two strains.

Figure 2.6



A schematic representation of hydrid dysgenic cross. Reproduced from Siomi *et al.* with permission from Nature Publishing Group (Permission ID number: 3720901452157).

Table 2.1

Table 1. Performance comparison between TEMP and transposon or structural variation discovery methods

Measures \ Methods	TEMP	PoPoolationTE	RetroSeq	VariationHunter-CL	GASVPro
TE insertion Sensitivity	***98.80%***	*88.50%*	*71.82%*	*0.00%*	*NA*
TE insertion Precision	***99.50%***	*92.45%*	*93.95%*	*0.00%*	*NA*
Average error of TE insertion frequency estimate	***7.27%***	*8.77%*	*NA*	*NA*	*NA*
TE absence Sensitivity	***93.09%***	*NA*	*NA*	*86.91%*	*72.37%*
TE absence Precision	***98.64%***	*NA*	*NA*	*79.74%*	*61.26%*
Average error of TE absence frequency estimate	***7.25%***	*NA*	*NA*	*NA*	*NA*

Table SII-2

TE	Genomic feature	#insertions in the feature	#insertions genome wide	p-value	Benjamini–Hochberg q-value	log p-value
FBgn0003055_P-element	Promoter	437	652	1.96E-177	5.07E-175	176.71
FBgn0000638_FB	Intron/UTR	284	327	2.04E-94	2.65E-92	93.69
FBgn0000155_roo	Intron/UTR	1011	2016	4.23E-64	3.66E-62	63.37
FBgn0001210_hobo	Intron/UTR	288	531	3.25E-26	2.10E-24	25.49
FBgn0014967_hopper	Intron/UTR	175	321	6.95E-17	3.60E-15	16.16
FBgn0002698_mdg3	Intron/UTR	105	191	4.71E-11	2.03E-09	10.33
FBgn0003519_Stalker	Intergenic	72	101	2.10E-10	7.75E-09	9.68
FBgn0003122_pogo	Intron/UTR	139	278	3.54E-10	1.15E-08	9.45
FBgn0010302_Burdock	Intron/UTR	121	241	3.38E-09	9.73E-08	8.47
FBgn0063897_Stalker4	Intergenic	76	116	2.75E-08	7.12E-07	7.56
FBgn0000199_blood	Intron/UTR	155	340	1.13E-07	2.66E-06	6.95
FBgn0005673_1360	Promoter	60	189	4.19E-07	9.04E-06	6.38
FBgn0004082_Tirant	Intron/UTR	99	203	4.92E-07	9.80E-06	6.31
FBgn0003007_opus	Intron/UTR	174	403	1.62E-06	3.00E-05	5.79
FBgn0001283_jockey	Intron/UTR	165	382	2.86E-06	4.94E-05	5.54
FBgn0069343_TAHRE	piRNA Cluster	7	27	9.60E-06	1.55E-04	5.02
FBgn0040267_Transpac	Intron/UTR	122	275	1.17E-05	1.78E-04	4.93
FBgn0003055_P-element	Intron/UTR	260	652	1.29E-05	1.86E-04	4.89
FBgn0000349_copia	Intron/UTR	166	394	1.49E-05	2.03E-04	4.83
FBgn0000005_297	Intron/UTR	137	317	1.80E-05	2.33E-04	4.75
FBgn0063429_invader2	Intergenic	23	29	1.91E-05	2.36E-04	4.72
FBgn0001283_jockey	Promoter	96	382	2.89E-05	3.40E-04	4.54
FBgn0000006_412	Intron/UTR	188	461	4.31E-05	4.86E-04	4.37
FBgn0014947_flea	Intron/UTR	98	221	8.02E-05	8.66E-04	4.10
FBgn0003122_pogo	Promoter	72	278	9.82E-05	1.02E-03	4.01
FBgn0000652_F-element	Intron/UTR	189	470	1.03E-04	1.03E-03	3.99
FBgn0069343_TAHRE	Intergenic	20	27	3.55E-04	3.41E-03	3.45
FBgn0005673_1360	Intron/UTR	83	189	3.89E-04	3.60E-03	3.41
FBgn0000481_Doc	Intron/UTR	137	336	4.25E-04	3.80E-03	3.37
FBgn0042682_Rt1b	Intron/UTR	42	84	4.52E-04	3.90E-03	3.34
FBgn0043969_diver	Promoter	25	79	9.49E-04	7.92E-03	3.02
FBgn0000481_Doc	piRNA Cluster	21	336	1.05E-03	8.50E-03	2.98
FBgn0005384_3S18	Intron/UTR	54	118	1.21E-03	9.50E-03	2.92
FBgn0004904_TART-A	Intergenic	16	22	1.96E-03	1.49E-02	2.71
FBgn0046110_Juan	Intron/UTR	35	73	3.22E-03	2.39E-02	2.49
FBgn0000349_copia	piRNA Cluster	22	394	0.00328686	2.36E-02	2.48
FBgn0003908_R1A1-element	Intergenic	28	46	0.00348536	2.44E-02	2.46
FBgn0004904_TART-A	piRNA Cluster	4	22	0.00348814	2.38E-02	2.46
FBgn0040267_Transpac	Promoter	64	275	0.00407263	2.70E-02	2.39
FBgn0002698_mdg3	Promoter	46	191	0.0069734	4.52E-02	2.16
FBgn0000006_412	Promoter	98	461	0.00864213	5.46E-02	2.06
FBgn0063429_invader2	piRNA Cluster	4	29	0.00963394	5.94E-02	2.02
FBgn0063455_Stalker2	Intron/UTR	46	107	0.01090051	6.57E-02	1.96
FBgn0001167_gypsy	piRNA Cluster	4	31	0.01218934	7.18E-02	1.91
FBgn0003490_springer	Promoter	12	38	0.01926688	1.11E-01	1.72
FBgn0041728_Rt1a	Intron/UTR	13	24	0.0200765	1.13E-01	1.70
FBgn0043055_Ivk	Promoter	11	34	0.02045411	1.13E-01	1.69
FBgn0004082_Tirant	Promoter	46	203	0.02066483	1.12E-01	1.68
FBgn0063507_G2	Intron/UTR	38	90	0.02627614	1.39E-01	1.58
FBgn0044355_Quasimodo	Intron/UTR	72	185	0.02730409	1.41E-01	1.56

Supplementary Material II

Supplementary material for Chapter II is too large to be included fully here. The whole file can be downloaded from link:

http://nar.oxfordjournals.org/content/42/11/6826/suppl/DC1

CHAPTER III

Local sequence assembly reveals a high-resolution profile of somatic structural variations in 97 cancer genomes

Summary

Genomic structural variations (SVs) are pervasive in many types of cancers. Characterizing their underlying mechanisms and potential molecular consequences is crucial for understanding the basic biology of tumorigenesis. Here, we engineered a local assembly-based algorithm (laSV) to detect SVs with high accuracy from paired-end high-throughput genomic sequencing data and to pinpoint their breakpoints at single base-pair resolution. By applying laSV to 97 tumor-normal paired genomic sequencing datasets across six cancer types produced by The Cancer Genome Atlas Research Network, we discovered that non-allelic homologous recombination is the primary mechanism for generating somatic SVs in acute myeloid leukemia. This finding contrasts with results for the other five types of solid tumors, in which non-homologous end joining and microhomology end joining are the predominant mechanisms. We also found that genes recursively mutated by single nucleotide alterations differed from genes recursively mutated by SVs, suggesting that these two types of genetic alterations play different roles during cancer progression. We further

characterized how the gene structures of the oncogene *JAK1* and the tumor suppressors *KDM6A* and *RB1* are affected by somatic SVs and discussed the potential functional implications of intergenic SVs.

Introduction

Genomic structural variations (SVs) such as deletions, insertions, inversions, translocations and tandem duplications are an important class of genetic variations that underlies genomic diversity in a population (Alkan et al., 2011). A deep and comprehensive understanding of the formation mechanism, genomic distribution and functional impacts of SVs is crucial for studying complex diseases such as cancer.

Performing a comprehensive survey of different SV formation mechanisms and their relative contributions across different cancer types is challenging because it entails precise characterization of the sequence across the breakpoints. Despite extensive efforts, accurately detecting SVs with a high resolution remains a challenge. Most existing SV discovery methods take advantage of three types of signals that are indicative of SVs between the reference genome and the sample genome: changes in the coverage of read pile-up, suggesting copy number alterations (read depth); discordant read pairs with a distance or orientation

between the two reads that is inconsistent with the reference genome (read pair); and reads that can be split into parts that align to discontiguous loci in the reference genome (split reads) (K. Chen et al., 2009; Escaramís et al., 2013; Layer et al., 2014; Quinlan et al., 2010; Rausch et al., 2012; J. Wang et al., 2011; Ye et al., 2009). It is algorithmically challenging to integrate information from these sources; furthermore, reads (or parts of a read) that can be aligned to multiple loci in the reference genome may result in spurious SV calls. Some methods such as TIGRA (K. Chen et al., 2013) try to pinpoint the breakpoints of predicted SVs by assembling reads mapped to the locus. This approach does not avoid the mapping ambiguities since both the SV predictions and the read selection for assembly are based on aligning short reads to the reference genome. A potential alternative is to perform a reference-free *de novo* assembly of the sequencing reads first and then compare the contigs with the reference genome. However, conventional *de novo* assembly methods are not designed for the purpose of SV discovery, especially for samples with a high degree of heterogeneity such as tumor samples. These tools assume that all the reads originate from a single underlying genome and therefore only detect homozygous SVs (Y. Li et al., 2011). In this report, we described a *de novo* local assembly-based SV discovery algorithm, designated laSV, which is able to pinpoint SV breakpoints at a single-nucleotide resolution and estimate the allele frequencies of the detected SVs in the sample.

Double-stranded breaks (DSB) in genomic DNA are detrimental to the cell, and several DSB repair pathways have therefore evolved to protect the cell from such catastrophic events. These pathways do not repair DSBs perfectly, and erroneous repairs are believed to be an important source of SVs (Hastings, Lupski, Rosenberg, & Ira, 2009b), especially in cancers. Homologous recombination (HR) is the method that is most widely used by cells to repair DSBs, and it requires long stretches of homologous sequences at the breakpoints. When HR occurs between non-allelic regions with high sequence similarity, termed non-allelic homologous recombination (NAHR), structural alterations may ensue (Jasin & Rothstein, 2013; Krejci, Altmannova, Spirek, & Zhao, 2012; Mehta & Haber, 2014). Mutations in genes that are key components of the HR pathway, such as *BRCA1* and *BRCA2*, are observed in many types of cancers and deemed the major driving force of genomic instability in these cancers. Nonhomologous end joining (NHEJ), however, does not require sequence homology and often generates very short deletions or insertions at the breakpoint. Key players in this pathway include XRCC5/6 and TP53 (Aparicio, Baer, & Gautier, 2014; Weterings & Chen, 2008). An alternative pathway, known as microhomology-mediated end joining (MMEJ), plays an active role in some cancers (Decottignies, 2013; Ottaviani, Lecain, & Sheer, 2014). MMEJ relies on relatively short stretches of homologous sequence (≤ 25 bp) at the breakpoint (Truong et al., 2013). The molecular details of this pathway are much less well understood compared with NAHR and NHEJ, although it is known to share the

initial end resection step with NAHR (Truong et al., 2013). In another DNA replication-associated repair mechanism, known as fork stalling and template switching (FoSTeS), it has been proposed that when a replication fork is stalled during replication, the polymerase is able to switch to a nearby locus and use it as the template to continue replicating, which often results in complex rearrangements (Hastings, Ira, & Lupski, 2009a; Ottaviani et al., 2014). Mobile element movements (MEs) are deletions that overlap with annotated transposable elements or novel insertions of TEs. Non-template insertions (NIs) are insertions whose sequences do not match any sequence in the reference genome.

Applying laSV to six cancer types, we discovered that in acute myeloid leukemia, NAHR is the major mechanism for generating somatic SVs, while in the other five types of solid tumor, NHEJ and MMEJ are the predominant forces underlying somatic SVs. We further observed that such a preference for DSB repair pathway utilization could be ascribed to the differential expression of several key genes in the HR pathway among the evaluated cancer types. Moreover, we analyzed genes that were affected by somatic SVs and to our surprise we found that genes that were frequently mutated by SVs tended to differ from the genes that were frequently mutated by single nucleotide alterations, which suggests different roles of the two types of mutations during cancer development. We also described in detail examples of complex genomic rearrangements and intragenic

SVs disrupting known oncogenes and tumor suppressors. Finally we characterized the somatic SVs in the intergenic regions and discussed the potential functional implications of SVs the overlap with genomic regulatory elements. The laSV package is freely available at https://github.com/JialiUMassWengLab/laSV/tree/master.

Material and Methods

Detection of putative SVs via *de novo* local assembly

laSV uses the de Bruijn graph as the backbone data structure for assembly. Briefly, a de Bruijn graph contains nodes that represent k-mer sequences and edges that represent the overlap between nodes (k-mers) and encode contiguous sequences as paths within the graph. The construction and storage of a de Bruijn graph is adopted from the CORTEX algorithm (Iqbal, Caccamo, Turner, Flicek, & McVean, 2012). After the construction of the de Bruijn graph from raw reads, laSV maps the reads to the branch sequences using the BWA MEM algorithm and identifies “connected” branches as those covered by the same read or the same read pair (**Figure SIII-1**). The connections of branches are stored as a hash table in the RAM and used for extending contigs during traversing. Next, the de Bruijn graph is traversed in a breadth-first manner to

output the “maximal unambiguous contigs” (MUCs). MUCs are defined as the longest contigs that contain only the connected branches (**Figure SIII-2**). These MUCs are then mapped to the reference genome using BWA MEM, which performs a local alignment. Contig segments that can be mapped to multiple loci in the reference genome are discarded because laSV cannot determine their origin unequivocally. Contigs that can be split-mapped to discontiguous loci of the reference genome are classified as discordant. Discordant contigs are indicative of putative SVs and are retained for further analysis.

Genotyping and estimation of SV allele frequencies

laSV further validates the putative SVs by mapping the raw reads to sequences that represent both the putative SV alleles derived from the assembled contigs and the corresponding alleles in the reference genome using the BWA aln algorithm. SV and reference alleles are prepared by extending 500 bp from the breakpoints in both directions. SV calls with fewer than four read pairs mapping to the variant allele are most likely false positives and are discarded. Based on the number of reads mapped to the variant allele and the corresponding reference allele, laSV estimates the frequency of the variant allele using the formula $F = \frac{C_V}{C_V + C_R}$, with effective coverages $C_V = \frac{V}{l_V}$ and $C_R = \frac{R_1 + R_2}{2l_R}$, where V, R_1 and R_2 represent the number of SV-supporting reads, the number of reads supporting reference locus 1 and the number of reads supporting reference locus

2, respectively (**Figure SIII-3**). Effective lengths l_V and l_R are given by $l_V = \sum_{i=h}^{1000} \lambda(i)(i-h)$ and $l_R = \sum_{i=0}^{1000} \lambda(i)i$, where h is the homologous sequence length and $\lambda(i)$ is the proportion of reads with fragment size i in the library (Trapnell et al., 2012).

After performing *de novo* SV discovery in the cancer genomes, we genotyped all of the putative SVs in the matched normal genomes. The SVs present in the cancer genomes with a ≥ 10% allele frequency that were supported by ≥ 4 read pairs and were absent from the matched normal genomes were considered somatic SVs.

Validation of NA12878 SVs using long-read sequencing datasets

We validated the SV calls in an individual with European ancestry using the long-read sequencing datasets for the same individual provided by Moleculo and PacBio. The datasets were downloaded from the 1000 Genomes Project FTP site:

ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/phase3/integrated_sv_map/supporting/NA12878/moleculo/, and

ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/phase3/integrated_sv_map/supporting/NA12878/pacbio/.

An SV was considered validated if there are 2 PacBio reads or 1 Moleculo read that supported the same type of SV with a breakpoint within 6 nt of that identified by laSV.

Simulated datasets for comparing SV detection algorithms

To compare the performance of laSV with several other methods, we generated simulated datasets each with 100 deletions, inversions and tandem duplications randomly inserted across human chromosome 9 using the SV simulation tool RSVSim (Bartenhagen & Dugas, 2013). Paired-end Illumina sequencing reads at 30X coverage with the mean and variance of the fragment size 400 bp and 50 bp respectively were then simulated using the pIRS software (Hu et al., 2012). The process was repeated for 100 times to produce 10,000 deletions, inversions and tandem duplications in total.

Classification of SV mechanisms

Our inference regarding the SV formation mechanism is based on the homology length, defined as the length of the homologous sequence between the two

breakpoint loci (**Figure SIII-4**). We define breakpoints with a homology length ≤ 2 and ≥ –10 (a negative homology length indicates insertion at the breakpoint) as being generated by NHEJ, those with a homology length > 2 and ≤ 25 as being generated by MMEJ, and those with a homology length > 25 as being generated by NAHR. Breakpoints with a more than 10 nt insertion at the breakpoint are classified as non-template insertions.

Detection of complex rearrangements

We used breakpoint graphs, as described by Pevzner (Pevzner, 2000), for the detection of complex rearrangement. Briefly, each node in the graph represents a genomic position, and two nodes are connected by a “breakpoint edge” if there is an SV bringing the two genomic positions together. Two nodes are connected by an “adjacency edge” if the distance between the two genomic positions is shorter than 100 Kb, and the weight of the edge is defined as the genomic distance between the two positions. An alternating path in the graph is defined as a path consisting of adjacency edges and breakpoint edges in an alternating fashion. A shortest alternating path in the graph that contains at least two breakpoint edges represents a potential complex rearrangement. The shortest alternating path between all pairs of nodes can be computed using a variant of the Dijkstra algorithm, as described by Brown (Brown, 1974).

Whole-genome sequencing and RNA-seq datasets

All of the whole-genome sequencing and RNA-seq datasets used in this study were produced by The Cancer Genome Atlas (TCGA) Research Network. The full list of samples used is listed in Supplementary Table SIII-1. The FASTQ raw sequence reads of genomic DNA were downloaded from CGHub (https://cghub.ucsc.edu/), and transcriptome RNA-seq data were obtained from the Data Portal of TCGA (https://tcga-data.nci.nih.gov/tcga/).

Analysis of intergenic SVs

Intergenic SVs (SVs that do not overlap with any genes) in BRCA, CESC, GBM, AML and UCEC were overlapped with the ENCODE DNaseI Hypersensitivity Uniform Peaks from the cell lines MCF-7, Helas3, Gliobla, K562 and Ishikawa, respectively. For enrichment simulation analysis, 20,000 random SV sets were generated for each of the five types of cancer, with each random set exhibiting exactly the same number of SVs and same SV length distribution as the real set. Each random SV set was overlapped with the DNase HS peaks in the corresponding cell type, and empirical p-values were computed as the fraction of random sets showing more overlap than the observed SV set.

Results

Detection of SVs

The overall workflow of laSV is depicted in Figure 3.1. It first uses raw sequence reads in FASTQ format as input and performs reference-free local assembly using de Bruijn graphs to generate contigs. Next, it aligns those contigs to the reference genome and detects all discordant alignments, i.e., different parts of the same contig mapped to discontiguous loci of the reference genome, which are indicative of putative SVs. Finally it maps the raw sequence reads to both the variant allele sequence (obtained from the assembled contigs) and the corresponding reference allele and estimates the variant allele frequencies of the putative SVs based on the ratio of variant-supporting reads over reference-supporting reads. This approach naturally integrates read-pair and split-read information, and by producing contigs that are much longer than raw sequence reads, it avoids some mapping ambiguities and, hence, achieves higher accuracy. We use a local assembly approach to avoid aggressively pruning the de Bruijn graphs, preserving true SVs present at low allele frequencies in the sample. Moreover, the reference-free assembly makes it possible to capture novel sequences that are not present in the reference genome.

To evaluate the accuracy of our method, we ran laSV on a high-coverage whole-genome DNA sequencing dataset from an individual of European descent (NA12878) produced by the 1000 Genomes Project and validated its results by comparing the calls with Moleculo and PacBio long-read sequencing datasets for the same individual. Among the SVs predicted by laSV with allele frequencies above 10%, 91.54% (1,687 out of 1,843) of the deletions and 94.93% (262 out of 276) of the non-template insertions were supported by the long-read sequencing datasets, suggesting that most of the laSV predictions were correct.

We also compared laSV and other SV detection methods CREST (J. Wang et al., 2011), pindel (Ye et al., 2009), delly (Rausch et al., 2012) and lumpy (Layer et al., 2014) on both simulated datasets (see Methods) and the NA12878 sequencing dataset. On simulated datasets, laSV achieved 99.20%, 99.46%, 99.51% precision rates and 83.18%, 85.98%, 81.34% recall rates for deletions, inversions and tandem duplications, respectively. Compared with the other methods, laSV has high specificity while maintaining good sensitivity (Figure SIII-5). On the NA12878 dataset, laSV outperforms the other methods in specificity (Figure SIII-6). These results show that laSV is able to make reliable predictions for various SV types.

We applied laSV to 97 cancer-normal paired high-coverage whole-genome sequencing datasets across six cancer types: uterine corpus endometrial

carcinoma (UCEC), glioblastoma multiforme (GBM) (Brennan et al., 2013; Cancer Genome Atlas Research Network, 2008), sarcoma (SARC), cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), breast invasive carcinoma (BRCA) (Cancer Genome Atlas Network, 2012) and acute myeloid leukemia (AML) (The Cancer Genome Atlas Research Network, 2013), produced by The Cancer Genome Atlas (TCGA) Research Network (Supplementary Table SIII-1). We identified somatic SVs as those that were present in the cancer sample but absent in the normal tissue of the corresponding individual. A total of 35,396 somatic SVs were detected, and we observed a high degree of heterogeneity in terms of the total number of somatic SVs, the composition of different SV types and the possible contributions of different SV mechanisms across samples, even within the same cancer type (**Figure 3.2**). An analysis of the SV length distribution reveals that most of the somatic SVs are very short (a few hundred base pairs) deletions and inversions (**Figure SIII-7**), which are possibly the product of error-prone DNA repairs and may have limited phenotypic impact.

We asked whether the SVs in CESC were due to human papillomavirus (HPV), which is a major cause for CESC. We aligned all the contigs assembled from CESC samples to the genomes of all 175 HPV strains downloaded from PaVE (http://pave.niaid.nih.gov/) using the BLAT algorithm (Kent, 2002). None of the

contigs indicative of SVs could be aligned to the HPV genomes, suggesting that the SVs we identified were not caused by HPV.

A survey of molecular mechanisms underlying somatic SVs

Because laSV has the capability to pinpoint breakpoints with a single-nucleotide resolution, we were able to infer the molecular mechanisms underlying the somatic SVs that we detected based on sequence homology at breakpoints (for more details about sequence homology please see the methods and materials section and **Figure SIII-4**) (**Figure 3.2b**). In all five of the solid tumor types we analyzed, NHEJ and MMEJ appear to be the predominant forces driving somatic SVs, which is consistent with previous reports (Malhotra et al., 2013; Yang et al., 2013). Surprisingly, in acute myeloid leukemia (AML), most somatic SVs show long stretches of homologous sequences across breakpoints and are probably the result of NAHR. To ensure that this difference is not an artifact due to the choice of homology length cutoffs for classifying the three mechanisms, we plotted the distributions of sequence homology lengths at SV breakpoints across all of the samples we analyzed (**Figure 3.2c**). Despite substantial heterogeneity among samples within the same cancer type, AML samples generally exhibit longer sequence homology at breakpoints than the other cancer types (p-values

are 1.46e-4, 1.20e-3, 8.69e-3, 2.36e-5 and 0.0153 versus BRCA, CESC, GBM, SARC and UCEC, respectively; Wilcoxon rank sum test).

What might be the reasons for such a differential preference for DSB repair pathways among cancer types? We found that for a third of the known genes in the HR pathway (6/18), the expression level is significantly higher in AML than in all the other cancer types (q-value < 0.01; **Figure SIII-9**). The genes that are more abundantly expressed in AML include *BRCA2*, *FAM175A* and *BRIP1*, which encode proteins known as RAD51 mediators, known to be crucial for recruiting RAD51 to damaged sites and initiating the HR pathway upon DNA damage (Krejci et al., 2012). Perhaps these more highly expressed HR genes increase the activity of the HR pathway in AML and lead to a higher proportion of SVs produced by NAHR than in the other cancers. An alternative explanation is that some unknown mechanism generates extensive DNA damages that require HR pathway for repairing and as an emergency response the expression levels of those genes are up-regulated.

Identification of complex genomic rearrangements

Complex genomic rearrangements are defined as SVs that are formed in a single event and involve multiple breakpoints. One class of replication-based mechanisms capable of generating such complex rearrangements is replication

fork stalling and template switching (FoSTes) and more generally microhomology-mediated break-induced replication (MMBIR) (Hastings et al., 2009a). Another mechanism is chromothripsis, massive chromosomal rearrangements that occur during a single catastrophic event within a localized genomic region (Stephens et al., 2011). To identify potential complex rearrangements, we developed a graph-based algorithm to connect breakpoints that are proximal to each other (see methods and materials section for details).

Figure 3.3a shows an example of complex rearrangement likely due to MMBIR. In the gene body of *MEGF8*, there is a 749 bp deletion and in its place is a segment of 5,584 bp that includes the 3´ portion of *PPR19* and the 5´ portion of *TMEM145*, two genes upstream of *MEGF8*. The two breakpoints exhibit 3-bp and 1-bp homology, respectively. This rearrangement effectively creates two fused genes, *MEGF8–PPR19* with exons 1-19 of *MEGF8* and *TMEM145–MEGF8* with exons 20-42 of *MEGF8*. RNA-seq data from the same individual indicates that the expression level of exons 1-19 of *MEGF8* is 1.24-fold higher than exons 20-42 of *MEGF8* (**Figure 3.3a**). The *TMEM145–MEGF8* chimeric transcript likely undergoes nonsense-mediated decay due to a premature stop codon caused by the fusion and the reads mapping to exons 20-42 of *MEGF8* are from the wild type copy of the *MEGF8* gene in the sister chromosome.

In addition, we noticed that in some of the samples, there are a large number of breakpoints concentrated within localized genomic regions. For instance, in one SARC sample, the vast majority of breakpoints fall within four narrow genomic regions in chromosomes 1, 5, 12 and 14 (**Figure 3.3b**; left). There are also many novel adjacencies connecting fragments of these four regions, suggesting extensive rearrangements possibly as a result of faulty DNA repairs in response to chromothripsis. Another SARC sample shows similar patterns with different genomic loci being affected (**Figure 3.3b**; right).

Somatic SVs that overlap protein-coding genes

In the samples we studied, there were a total of 17,184 protein-coding genes overlapping at least one SV in at least one sample (Supplementary Table SIII-2). Most of those SVs probably do not confer a growth advantage to the tumor cells carrying them and, hence, are so-called “passenger mutations”. However, there was a significant enrichment of known oncogenes and tumor suppressors (Vogelstein et al., 2013) (p-value=0.013; hypergeometric test) among the genes affected by somatic SVs. Furthermore, when we restricted our analysis to somatic SVs present with a 20% or higher allele frequencies, the enrichment was more significant (p-value=6.5e-3; hypergeometric test), suggesting that SVs having phenotypic consequences are more likely to cause a cancer subclone to be selected and increase in frequency.

When we performed gene ontology analysis on genes that are affected by SVs in multiple samples for each of the six cancer types (**Figure SIII-10**), we observed enrichment for processes such as immune responses, keratinization, metallopeptidase-related processes and cell-cell adhesion. Mutations in genes belonging to these biological processes and pathways are unlikely to cause tumorigenesis. Instead, they might confer a growth advantage on tumor cells and allow them to evade the immune system and metastasize.

Three of the cancer types we analyzed (BRCA, GBM and AML) have been extensively studied before by the TCGA consortium (Brennan et al., 2013; Cancer Genome Atlas Network, 2012; The Cancer Genome Atlas Research Network, 2013). Whole-exome sequencing was performed on hundreds of samples for each of these three cancer types to identify recurrently mutated genes. We compared the lists of genes showing recurrent single-nucleotide alterations (SNAs) and indels with the genes that we found to be affected by SVs and asked whether the same genes tended to harbor both SNAs/indels and SVs. In BRCA, the overlap between genes showing recurrent point mutations and genes affected by SVs was statistically significant (p-value=0.0184, hypergeometric test). In GBM and AML, however, there was no significant overlap between recurrently point-mutated and SV-mutated genes (p-

value=0.378 and 0.093 for GBM and AML, respectively), suggesting that SNAs/indels and SVs may play different roles during cancer development.

For all of the genes harboring SVs or SNAs in a given cancer type (point-mutation data are not available for SARC), we correlated the number of samples where SV-induced mutations occurred with the number of samples where point mutations occurred. We observed negative correlations for all five cancer types, with Pearson correlation coefficients r=–0.537, –0.785, –0.713, –0.293 and –0.697 (all p-values < 1e-100) for UCEC, GBM, CESC, BRCA and AML, respectively, further indicating that genes show recurrent point mutations are less likely to harbor SV mutations (**Figure 3.4a**).

To test whether this negative correlation was due to decreased power of detecting SNAs in deleted regions, we assessed the relative impact of deletions in each cancer type. Since all SNAs are in coding regions (CDS), we compared the total deleted length of CDS with the duplicated length of CDS in each cancer type (Supplementary Table SIII-3). Our results revealed no strong bias towards deletion over duplication and therefore the aforementioned negative correlations are unlikely to be the result of compromised SNA detection power due to deletions. Furthermore, since the breakpoints of the SVs fall predominantly in intergenic and intronic regions far away from CDS, it is also unlikely that the negative correlations are caused by the effect of SNAs on SV detection power.

The tumor suppressor *KDM6A* encodes a lysine-specific demethylase that catalyzes the demethylation of tri- and di-methylated H3K27. Missense and nonsense mutations in this gene have been reported in multiple cancer types (Sengoku & Yokoyama, 2011). In one of the CESC samples, laSV detected a 148,495 bp deletion that eliminates exons 3-28 of *KDM6A* (**Figure 3.4b**). This deletion leads to a much shortened transcript, which if translated, encodes a nonfunctional protein because the JmjC catalytic domain is deleted. Based on RNA-seq data from the same individual, we observed 48 reads that map across the exon 2–exon 29 junction, indicating that the mutated version of the *KDM6A* gene was indeed transcribed.

The oncogene *JAK1* encodes a non-receptor tyrosine kinase whose hyperactivity has been implicated in multiple cancer types, including breast cancer, colorectal cancer and lung cancer (Ren et al., 2013; Song, Rawal, Nemeth, & Haura, 2011). We observed a 22,471 bp tandem duplication that includes exons 6-12 in one of the BRCA samples (**Figure 3.4c**). At the protein level, this duplication leads to an extra copy of a portion of the FERM domain, the entire SH2 domain, and the SH2-pseudokinase linker. Recent biochemical studies have shown that the FERM domain and the SH2 domain of JAK family proteins are crucial for binding to the cytoplasmic region of the cytokine receptors (Babon, Lucet, Murphy, Nicola, & Varghese, 2014; Wallweber, Tam, Franke, Starovasnik, & Lupardus,

2014). Perhaps the duplication increases the affinity with which JAK1 binds to the cytokine receptor or shifts the relative position of the kinase domain with respect to the cytokine receptor, disrupting proper regulation.

In both of the above examples, the mutated genes are still translated in-frame. In other cases, structural variations may also cause a frameshift and, thus, grossly alter the amino acid sequences of the protein product. *RB1* is a negative regulator of the cell cycle and was the first discovered tumor suppressor (Benavente & Dyer, 2015). In one of the BRCA samples, we observed a tandem duplication that included exons 7-12 of the *RB1* gene. This results in a frameshift that leads to a premature stop codon (**Figure 3.4d**). In the same individual, we observed a 6-fold decrease in *RB1* expression in the tumor tissue compared with the nearby normal tissue. The premature stop codon is located 2,077 nt upstream of the last exon-exon junction and may have triggered nonsense-mediated decay, leading to the decreased *RB1* level.

Some intergenic SVs may impact genomic regulation

SVs that do not overlap any gene are usually ignored due to the difficulty of evaluating their possible effects. For five of the six cancer types we analyzed (except SARC) we were able to find DNaseI sequencing data produced by ENCODE on cell types corresponding to the same tissue. We then intersected

the intergenic SVs with DNase hypersensitive sites (DHSs) in the corresponding cell type. Overall, a background level of 1.10% (356/32455) intergenic SVs overlapped with DNase hypersensitive regions. Nevertheless, DHS-overlapping SVs have higher allele frequencies than the non-overlapping SVs (p-value=3.73e-4, Wilcoxon Rank Sum test, **Figure SIII-11**), indicating that DHS-overlapping SVs are more likely to confer a growth advantage.

BCL9 is an oncogene that encodes an important component of the Wnt pathway. BCL9 interacts with β-catenin to enhance its transcriptional activity and is implicated in several types of cancer (la Roche, Worm, & Bienz, 2008; Sampietro et al., 2006). In one of the BRCA samples, we observed a 22,847-bp duplication upstream of the *BCL9* gene that overlaps with two DHSs in MCF-7 cells (**Figure 3.5**). One of the DHSs is bound by the transcription factors E2F1, CTCF, RAD21 and MAX. Moreover, the Pol II ChIA-PET data indicate that there is a chromatin interaction between the DHS and the promoter of the *BCL9* gene, which suggests that the DHS may regulate *BCL9* transcription. Indeed, we observed a 63.52% increase in *BCL9* expression in the tumor sample compared with the matched normal sample. It is likely that the duplication of the regulatory DHS leads to an elevated expression level of *BCL9*.

Discussion

Cancer is a group of complex diseases driven by various genetic and epigenetic alterations. Previous surveys on genetic alterations in cancer have mostly focused on single-nucleotide mutations in protein-coding sequences, fusion transcripts and copy number alterations of large genomic segments (Vogelstein et al., 2013). In this article, we reported a novel algorithm, laSV, that is capable of detecting genomic SVs across a wide spectrum of sizes from highly heterogeneous tumor samples and pinpointing their breakpoints at a single-nucleotide resolution. Applying this algorithm to 97 high-coverage whole-genome sequencing datasets across six cancer types, we observed several interesting phenomena.

Because laSV supports nucleotide-resolution delineation of SV breakpoints, we examined the prevalence of different breakpoint formation mechanisms across all of the samples that we analyzed. To our knowledge, there have been two studies conducted thus far that have comprehensively surveyed breakpoint formation mechanisms across multiple cancer types (Malhotra et al., 2013; Yang et al., 2013). Both studies included only solid tumors and concluded that most breakpoints exhibit little or no homology and were therefore probably formed via NHEJ or MMEJ. We observed similar patterns in the five types of solid tumors that we analyzed. In AML, however, most of the breakpoints showed homologous sequences much longer than those needed for NHEJ and MMEJ, suggesting that

NAHR is the predominant mechanism of breakpoint formation. Such a preference for different breakpoint formation mechanisms might provide important insight into the course of evolution taken by different cancer types and have implications for the development of cancer type-specific treatments.

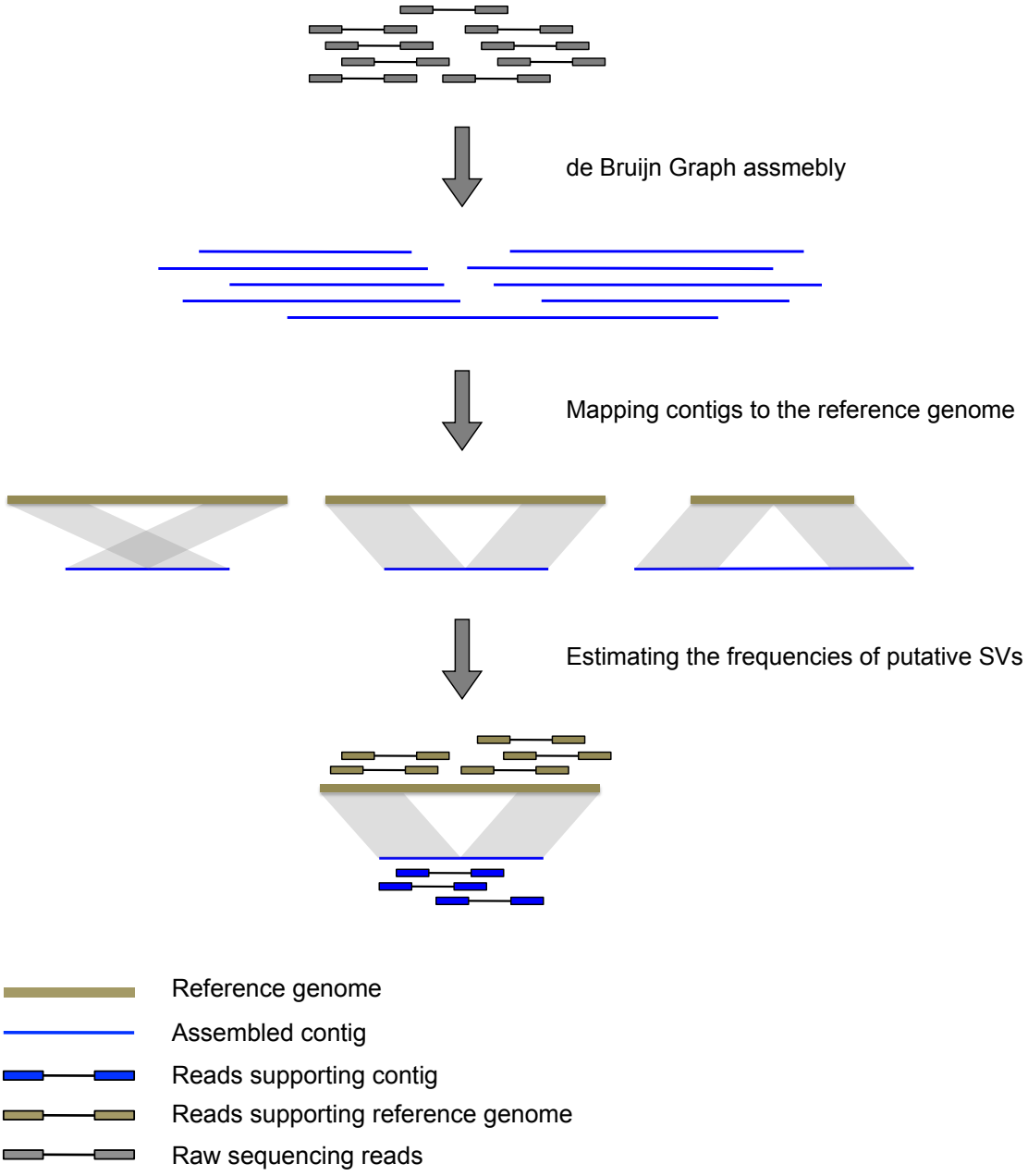
At present laSV employs BWA to align assembled contigs to the reference and predict SV breakpoints. There are other methods, such as YAHA (Faust & Hall, 2012) and AGE (Abyzov & Gerstein, 2011), that specialize in aligning long sequences and detecting potential breakpoints. In the future it would be interesting to explore how laSV performs using these software for contigs alignment. In addition to reflecting the confidence level of the SV calls, the SV allele frequency computed by laSV could also be useful in some other applications, such as distinguishing driver mutations from passenger mutations since driver mutations tend to occur early on during the tumor development and therefore be present in most of the cells in the tissue. Furthermore, when samples from different stages of the tumor development or from different metastasized sites are available, it would be informative to compare the SV allele frequencies across those samples as they may reveal how the cancer progressed and adapted to new metastasized locations.

The fact that genes that show recurrent SNAs do not appear to be preferentially mutated by SVs is noteworthy. Considering that the spontaneous mutation rate

for point mutations is much higher than for SVs (Itsara et al., 2010), we hypothesize that tumorigenesis is often initiated by point mutations and that most SVs occur later during cancer development, when DNA repair mechanisms are compromised. Because additional SV mutations in a gene already disrupted by cancer-causing point mutations rarely enhance the cancer phenotype, they are unlikely to be selected for in tumor tissues. The observation that genes that are recurrently affected by SVs are enriched for pathways such as cytoskeleton metabolism, immune response and cell-cell adhesion, which are unlikely to cause uncontrolled cell proliferation but may contribute to the migration, immune defense evasion and metastasis of cancer cells, further supports our hypothesis. The characterization of SVs in cancer lags behind that of SNAs/indels because whole-genome sequencing is more costly than exome sequencing. More cancer genomes need to be sequenced to more accurately identify genes recurrently affected by SV mutations for various cancer types. Our results indicated that in addition to point mutations, gains/losses of large genomic segments and transcript fusions, intragenic SVs can also have a significant impact on the expression levels and products of protein-coding genes, as demonstrated by the examples of *KDM6A*, *JAK1* and *RB1*. Therefore, laSV, with its ability to accurately detect more subtle SVs, will be a valuable tool for future surveys of genetic alterations in cancers.

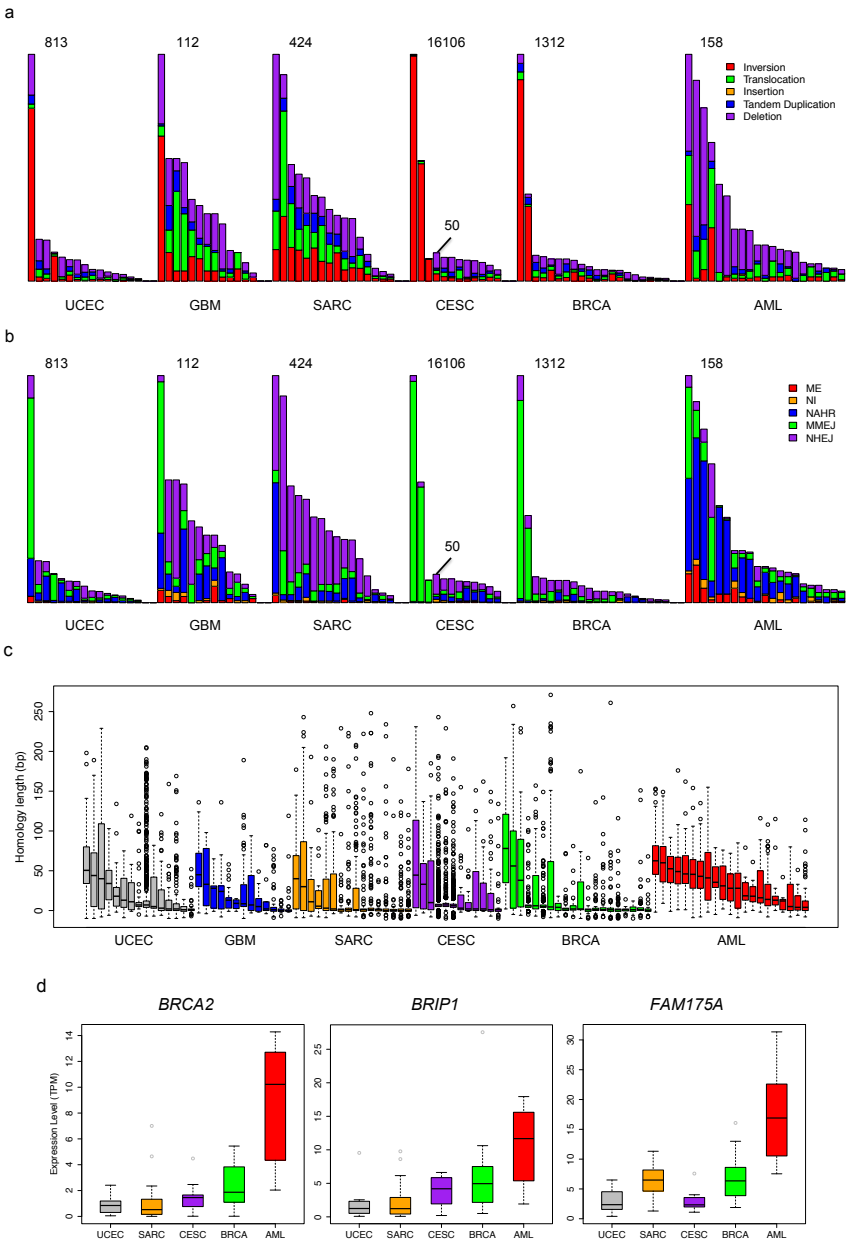
Previous reports on cancer research have mostly focused on genetic alterations within or including coding regions. A recent study suggests that a large fraction of the non-coding portion of the human genome may contain regulatory elements (ENCODE Project Consortium, 2012). We found that some of the SVs in non-coding regions overlap with DNase hypersensitive sites and might have some regulatory impacts. The example of *BCL9* that we described demonstrates how SV discovery in the non-coding region can, when considered in conjunction with the rich information accumulated by the ENCODE consortium, shed new light on regulatory alterations in cancer. With our rapidly expanding knowledge regarding the various regulatory elements in the human genome, further studies will be carried out to interrogate the roles of non-coding regulatory regions in cancer. The accurate identification of more subtle structural variations and the precise determination of their breakpoints will be crucial for the success of such investigations.

Figure 3.1



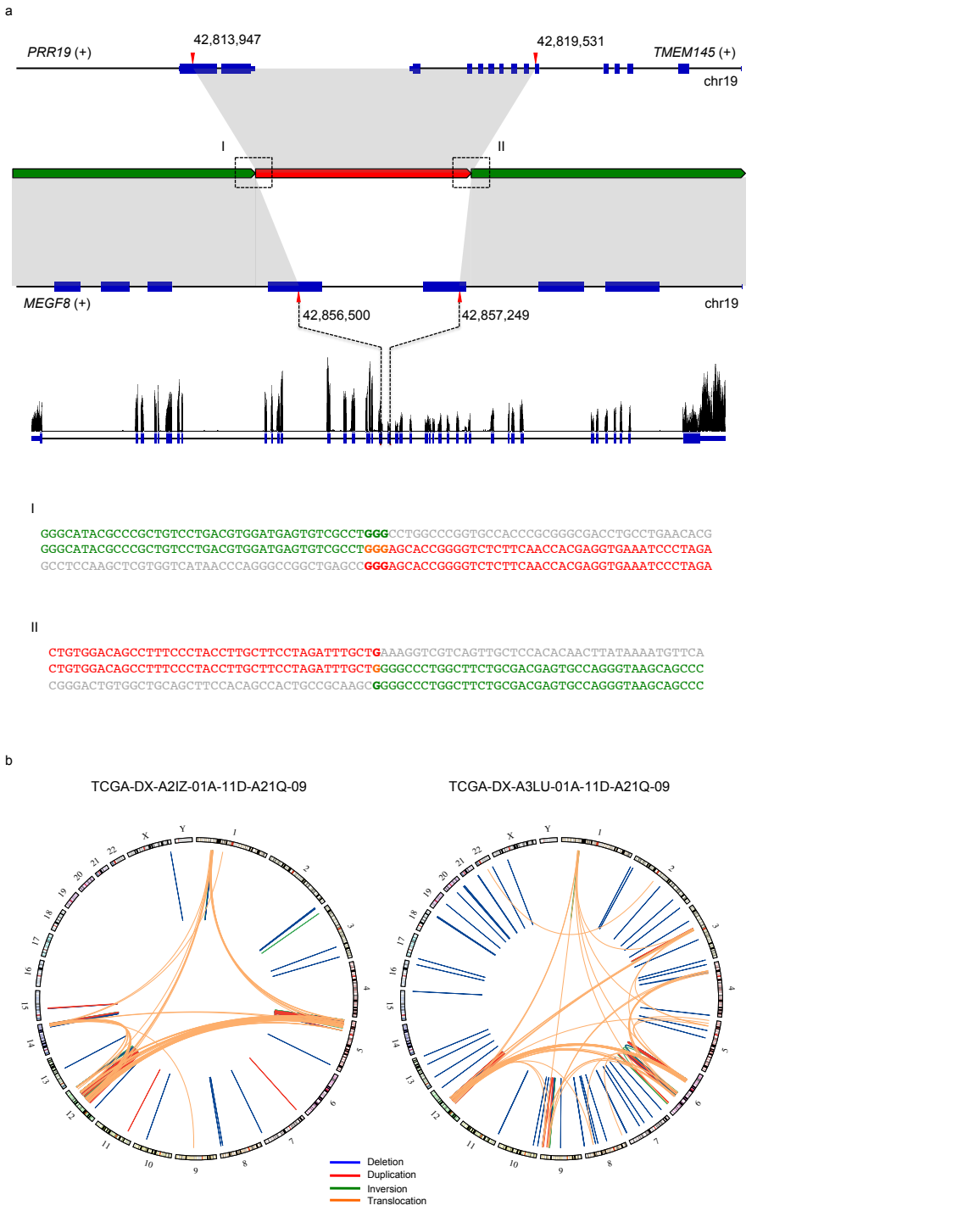
A schematic representation of the workflow of laSV.

Figure 3.2



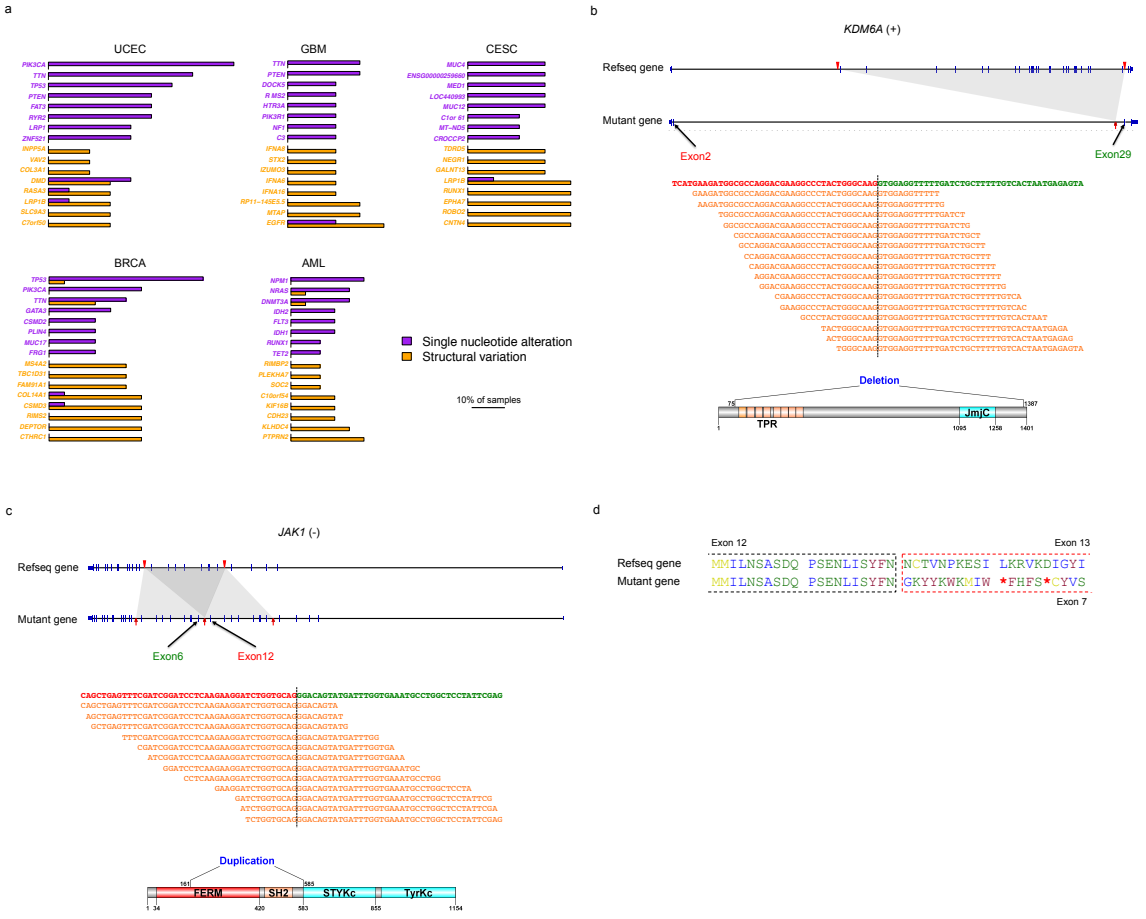
An overview of the SVs across all of the samples we analyzed. The distribution of (**A**) different types of SVs, (**B**) different breakpoint mechanisms and (**C**) breakpoint homology sequence lengths across all 97 samples. The evaluated cancer types are indicated at the bottom of each panel. For CESC, two different scales are used because three of the samples contain many more SVs than the remaining samples. (**D**) Expression levels of three key genes of the HR pathway across different cancer types. Samples within the same cancer type are ranked by the total number of somatic SVs in descending order in (A) and (B) and are ranked by the median homology sequence length in descending order in (C). TPM is transcripts per million, a means of gene expression quantification used by the RSEM algorithm, in which the total number of transcripts in a cell is normalized to one million. RNA-seq data were not available for the GBM samples. ME: mobile element; NI: non-template insertion; NAHR: non-allelic homologous recombination; MMEJ: micro-homology mediated end-joining; NHEJ: non-homologous end joining.

Figure 3.3



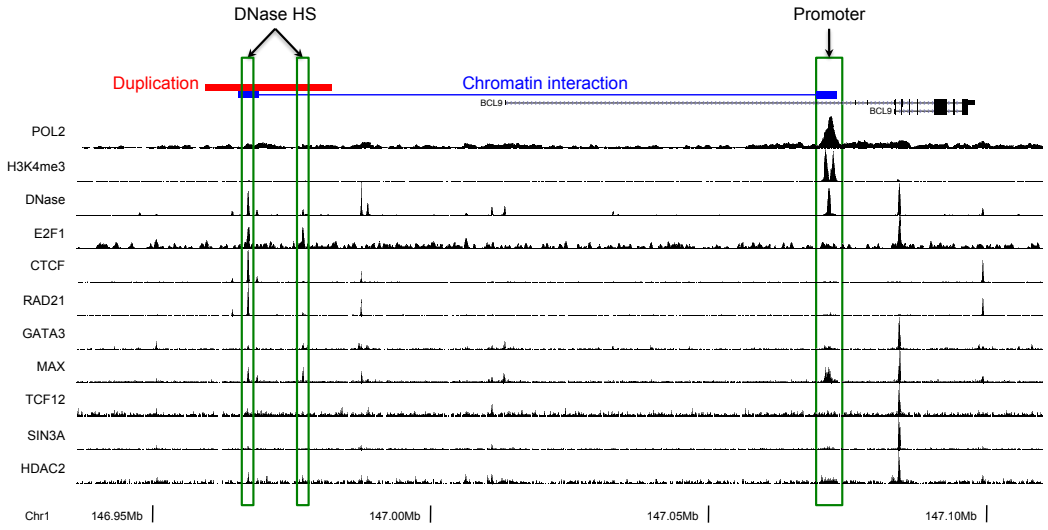
Examples of complex SVs. (**A**) An example of an MMBIR. Boxes (I) and (II) show the two breakpoint sequences. The characters in bold indicate homologous sequences. (**B**) Two examples of chromothripsis.

Figure 3.4



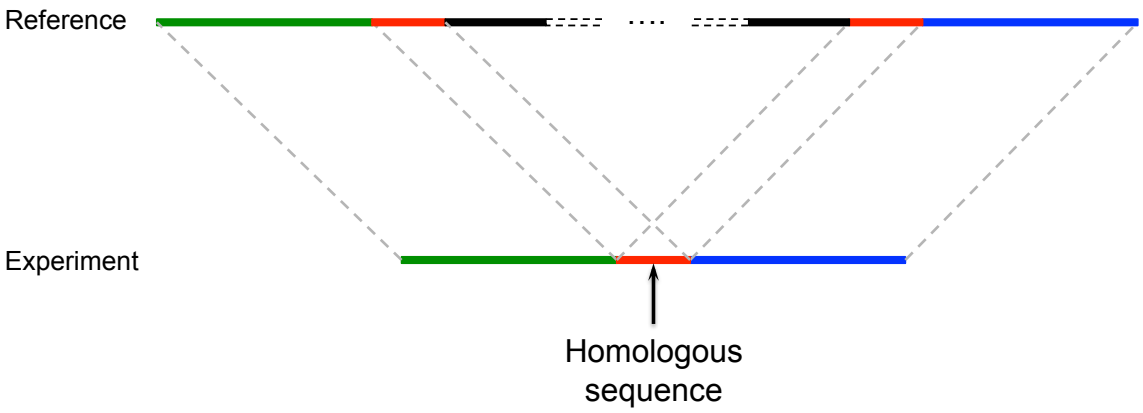
The impact of somatic SVs on protein-coding genes. (**A**) Comparison of genes frequently affected by point mutations with those frequently affected by SVs. Purple bars indicate the percentage of samples in which the gene carries SNAs while orange bars indicate the percentage of samples in which the gene carries SVs. The purple genes are the ones most frequently affected by SNAs within each cancer type; orange genes are the ones mostly frequently affected by SVs. (**B**) A deletion within the tumor suppressor *KSDM6A*. The black dashed line indicates the exon–exon junction. The orange sequences are representative RNA-seq reads that map across the junction. (**C**) A tandem duplication within the oncogene *JAK1*. (**D**) Amino acid sequences of the wild-type and duplicated versions of *RB1*. The red asterisks indicate stop codons.

Figure 3.5



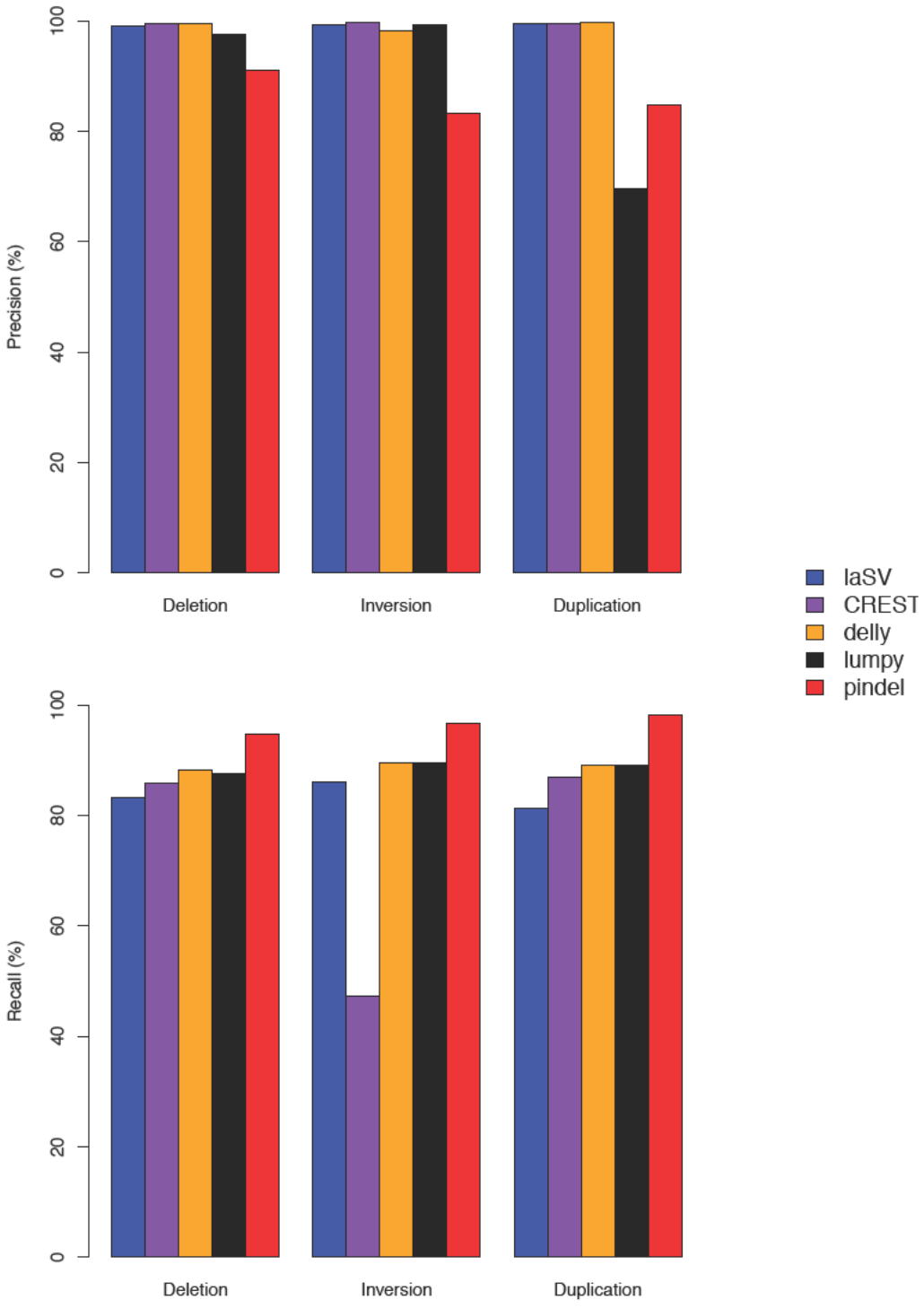
An example of somatic SV affecting intergenic gene regulatory elements. In one of the BRCA samples a tandem duplication spans a DNase hypersensitive site containing regulatory elements of oncogene *BCL9*.

Figure SIII-4



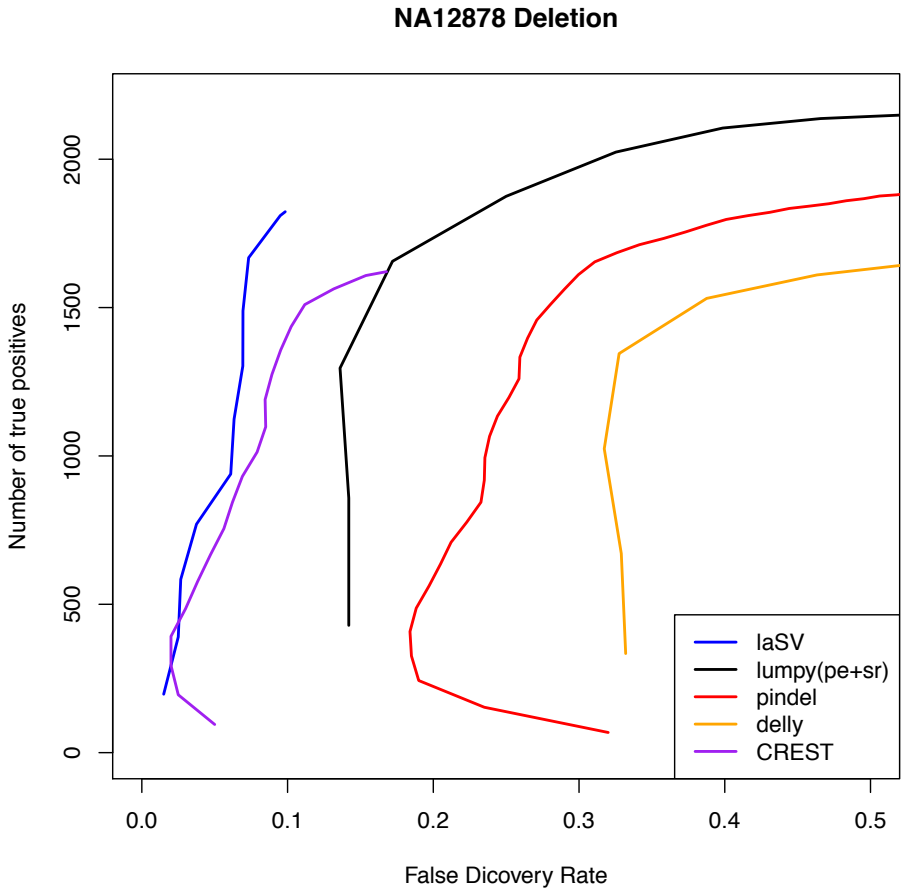
A schematic representation of how homologous sequence is defined.

Figure SIII-5



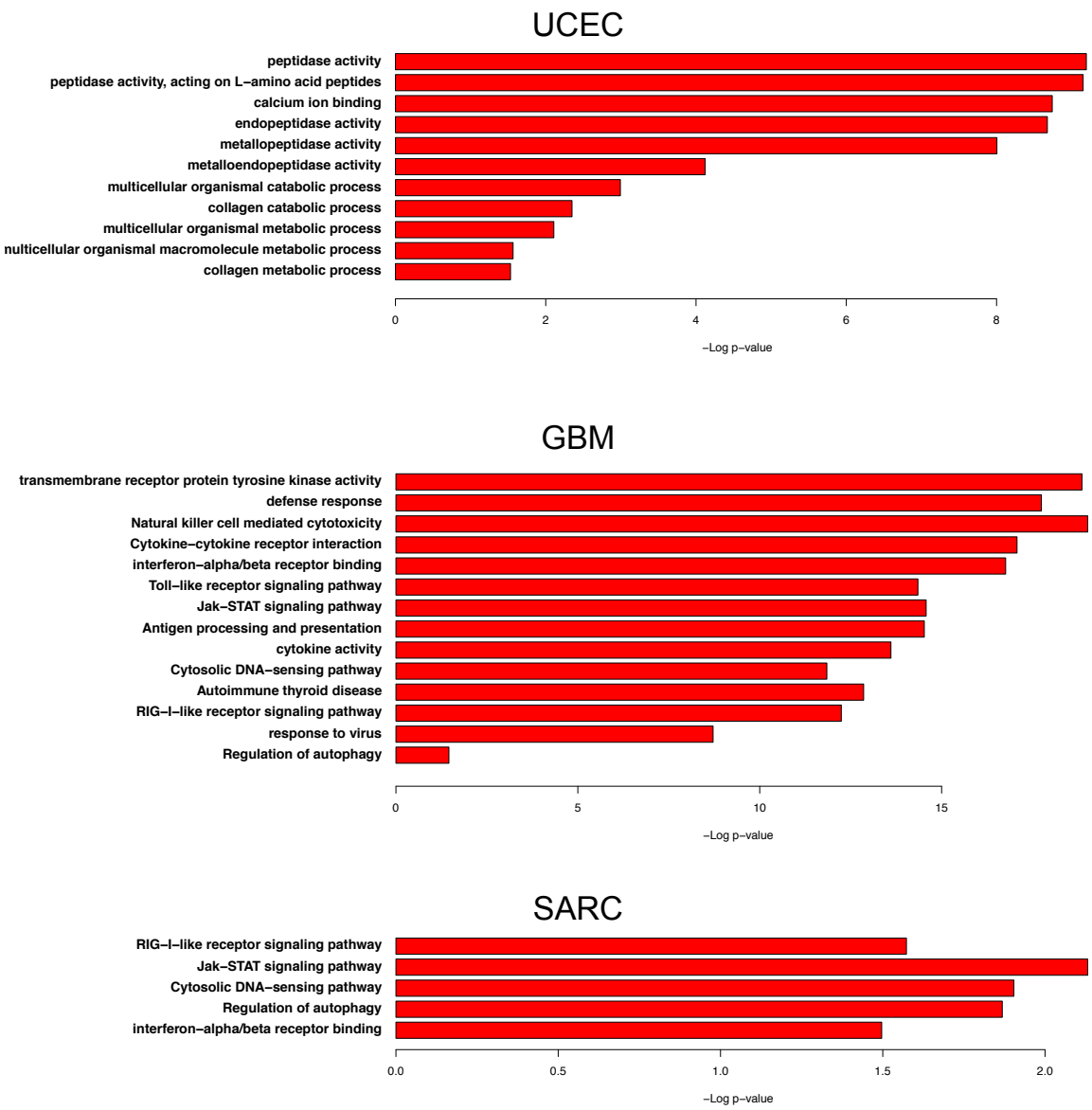
Comparison of the performance of SV detection methods on simulated datasets.

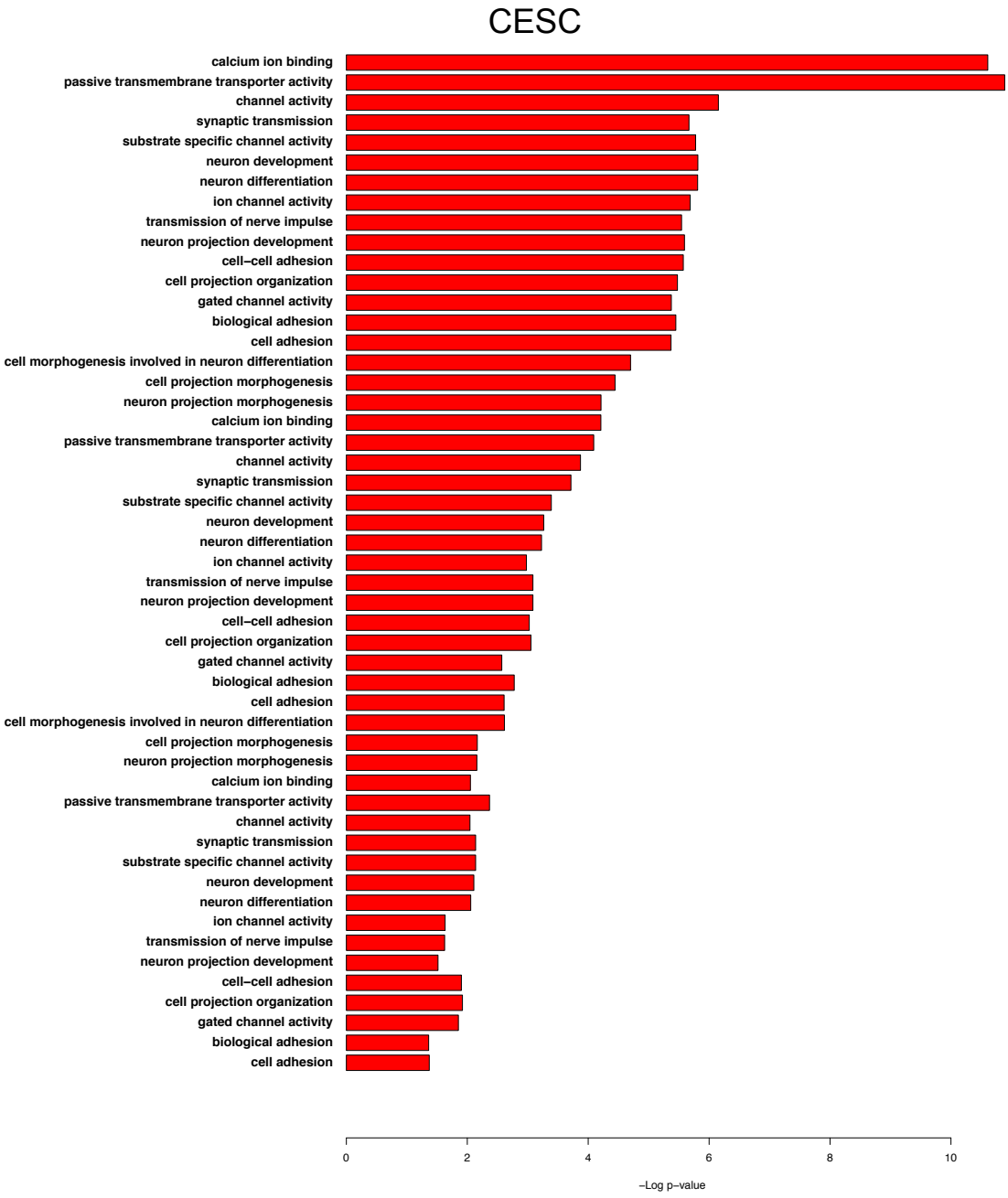
Figure SIII-6

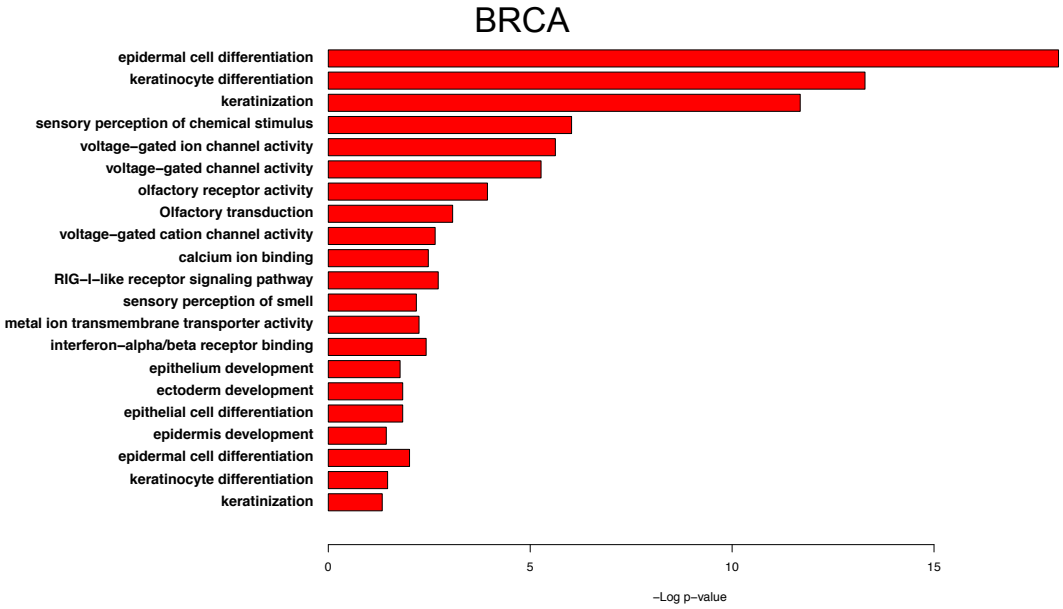


Comparison of the performance of SV detection methods on a high-coverage whole-genome sequencing dataset on NA12878 cells.

Figure SIII-10




CESC
calcium ion binding
passive transmembrane transporter activity
channel activity
synaptic transmission
substrate specific channel activity
neuron development
neuron differentiation
ion channel activity
transmission of nerve impulse
neuron projection development
cell–cell adhesion
cell projection organization
gated channel activity
biological adhesion
cell adhesion
cell morphogenesis involved in neuron differentiation
cell projection morphogenesis
neuron projection morphogenesis
calcium ion binding
passive transmembrane transporter activity
channel activity
synaptic transmission
substrate specific channel activity
neuron development
neuron differentiation
ion channel activity
transmission of nerve impulse
neuron projection development
cell–cell adhesion
cell projection organization
gated channel activity
biological adhesion
cell adhesion
cell morphogenesis involved in neuron differentiation
cell projection morphogenesis
neuron projection morphogenesis
calcium ion binding
passive transmembrane transporter activity
channel activity
synaptic transmission
substrate specific channel activity
neuron development
neuron differentiation
ion channel activity
transmission of nerve impulse
neuron projection development
cell–cell adhesion
cell projection organization
gated channel activity
biological adhesion
cell adhesion
0
2
4
6
8
10
−Log p−value



Gene Ontology enrichment of genes mutated by SVs in multiple samples. The enrichment analysis was carried out with DAVID Functional Annotation Tools (http://david.abcc.ncifcrf.gov) including terms in GO biological process, GO molecular function and KEGG pathway categories. The p-values were corrected for multiple testing using the Benjamini-Hochberg method. AML is not included here because there is no significant enrichment for any term.

Supplementary Material III

Supplementary material for Chapter III is too large to be included here. The whole file can be downloaded from link:

http://nar.oxfordjournals.org/content/early/2015/08/16/nar.gkv831/suppl/DC1

CHAPTER IV

DISCUSSIONS

The advent and advancement of high-throughput sequencing technology is a tremendous boost to the quest for better understanding the relationship between genotypes and phenotypes. The availability of high quality reference genomes for multiple species, combined with genome-wide high-throughput technologies for variants discovery and genotyping enables researchers to survey a large number of genomes in a time and cost efficient manner. The two methods described in this dissertation represent the latest efforts of harnessing the power of high-throughput sequencing for genetic variants/mutations discovery and genotyping. By providing a computational framework that selects and integrates reads incompatible with the reference and reconstructs the sample genome at the loci where it differs from the reference genome, TEMP and laSV are able to detect TE transpositions and SVs accurately, define the breakpoints at single nucleotide resolution and estimate the allele frequencies of the variants/mutations within the samples. The allele frequencies can also serve as confidence indicators for the predicted SVs. Since they are normalized for the local sequencing depth the allele frequencies are a more robust benchmark for measuring confidence level than simply the number of supporting reads. Validation on both simulated datasets and real biological datasets showed that the predictions made by the

two methods are reliable and they outperform other state-of-the-art tools in some important aspects.

Although there could be potential overlap between the applications between TEMP and laSV, these two methods are intended for different purposes. laSV employs a *de novo* assembly approach and compares the assembled contigs with the reference genome to predict putative SVs. This makes it quite versatile and able to detect all types of SVs including TE insertions at least in theory. In practice, however, two factors limit laSV's power for detecting novel TE insertion events. First of all, the inserted TE sequences are highly repetitive (there are usually hundreds or even thousands of almost identical copies in the genome) and the effort for *de novo* sequence assembly is usually severely hindered by the presence of repetitive sequences. Secondly, the success of the assembly approach depends heavily on sufficient sequencing depth because it requires enough overlap between neighboring reads in the region. For those TE insertion events present in the sample with relatively low allele frequencies, the small number of reads derived from the insertion allele makes it impossible to assemble across the insertion breakpoints and therefore severely limits the sensitivity of laSV in terms of detecting low-frequency TE insertion events.

TEMP, on the other hand, is specifically designed for detecting TE insertions and absences present with both high and low allele frequencies within the samples. Since it detects transposition events based on discordantly paired reads, is has no problem picking up TE insertions that are supported by only a few reads and therefore achieves higher sensitivity. TEMP takes advantage of the known consensus sequences of the active transposable elements to accomplish both high sensitivity and high specificity. But this approach also limits its utility, making it far less versatile than laSV and can only be used for detecting TE transposition events.

As I mentioned in the previous chapters, genomic structural variations is a class of highly complex and heterogeneous genetic variants and hence it is almost impossible to have a method or algorithm that can have superb performance on all types of SVs. That is our main motivation for developing these two different methods presented in this dissertation with laSV aiming at discovering SVs in general with relatively high allele frequency; and TEMP focusing specifically on TE transposition events across the entire allele frequency spectrum. Deploying a combination of different complementing methods is essential for obtaining a comprehensive picture about the global genetic variants/mutations landscape and crucial for the success of many projects.

In the following sections I will summarize the new innovation trends in the high throughput sequencing technology and discuss their potential repercussions on the evolution of the SV discovery algorithms.

Limitations of current high-throughput sequencing technology

Short read length and fragment size

Despite its enormous power and tremendous success in recent years, current popular high-throughput sequencing technology has its own limitations, most notably the relatively short read length and fragment size. As a consequence of this technical limitation the analysis that involves highly repetitive sequences longer than the fragment size (e.g., active transposons, segmental duplicated regions) is very challenging and some are even impossible. Here I discuss a few scenarios where interesting and important analysis are hampered due to the short read length and fragment size of current high-throughput sequencing technology.

1. In the genomes of many species multiple highly similar copies of the same transposon element are present. It would be informative to know which of the copies are active and responsible for the newly discovered

transposition events in the sample genomes because it will enable researchers to study the potential factors controlling the activeness of transposons such as genomic context (close to or far away from genes), DNA methylation at the promoter, chromatin state (open versus closed), and so on. Such information is almost unavailable, however, when both the read length and the fragment size of the sequencing library is substantially shorter than the transposon length. This is because in such situation we can only resolve the two ends of the inserted transposon sequence, which in most cases is insufficient to uniquely identify the original copy.

2. When the breakpoints of an SV fall within highly repetitive sequences longer than the fragment size of the sequencing library the SV will most likely eludes detection since the reads are too short to uniquely resolve either of the breakpoints and the read pairs cannot span the repetitive sequence due to the short fragment size. Because it is possible that a non-allelic homologous recombination happens between two highly similar sequences and thus causes an SV, this limitation might sometimes lead to false negatives in SV detection from high-throughput sequencing datasets.

3. The short read length and fragment size also put constraints on the effort for haplotype phasing in diploid genomes. For multiple variants loci, it is very useful to know their exact genotypes on each of the parental

chromosomes (i.e., resolving haplotypes). Such information is necessary for linkage disequilibria detection among multiple loci and greatly facilitates the imputation of low-frequency variants. When the distance between two variant loci is greater than the fragment size, they will never appear on the same fragment and it is therefore impossible to accomplish phasing based on high-throughput sequencing alone. There are computational software developed that use population/pedigree data and statistical models to accomplish haploid phasing from high-throughput sequencing datasets. But in cases where such population information is not available (which is true for nearly all non-model organisms) this approach will not work.

Requirement for relatively large quantity of samples

A typical Illumina paired-end sequencing library requires 2-5μg double-stranded DNA, which have to be harvested from a large number of cells. This requirement precludes the possibility of surveying individual genomes separately in heterogeneous samples such as tumor tissues. In the context of genetic variants analysis in samples containing multiple heterogeneous genomes, it is informative to know if two variants/mutations happen in the same genome and how often do they reside in the same genome. Such information allows us to investigate the relationship between pairs of genetic variants/mutations, for example, does one

mutation precedes another; is any mutation causing a higher mutation rate genome wide; is there any epistatic interactions among variants/mutations, etc. With sample DNA coming from a mixed pool of different genomes and relatively short fragment size, it is extremely difficult and in most cases impossible to figure out such co-occurrence information between any pair of variants/mutations on mainstream short-read sequencing platforms.

New development in high-throughput sequencing technology

Several commercial companies are developing novel technologies that can overcome the aforementioned limitations of current popular high-throughput sequencing technology. Pacific Biosciences RS II platform for example is able to sequence long stretches of DNA (10-15kb) in a single-molecule manner (Eisenstein, 2015). The downside of this platform at the current stage is that it is relatively expensive and the sequencing error rate is quite high. An alternative strategy that has shown some promise is to use biochemical tricks to assign short reads to genomic addresses, instead of producing ‘true’ long reads. The GemCode platform developed by 10X Genomics, for instance, partitions long DNA fragments (on average 50 kb) in to oil-encased droplets with each uniquely labeled by a 14-base barcode sequence and then the barcoded products are directly transferred to standard Illumina sequencing. Once the sequencing is finished, a software reconstruct the original long DNA fragments based on the

barcode information. This approach is compatible with widely used Illumina platforms and therefore requires minimum extra expenses on instruments. But as a consequence it is also subject to the biases and limitations of the Illumina platforms (Eisenstein, 2015).

There is also progress in the area of single-cell sequencing, which will allow researchers to study the genomes of single cells (Baslan & Hicks, 2014; Y. Wang & Navin, 2015). The major technical obstacle in sequencing extremely low amount of DNA lies in pre-sequencing DNA amplification. Since the standard sequencing technology requires much larger amount of sample DNAs than a single cell can offer, whole genome amplification (WGA) is an essential step for any attempt to sequence the genomes from a single cell. At present, the most widely used WGA methods are degenerative-oligonucleotide-PCR (DOP-PCR) and multiple-displacement-amplification (MDA), both of which have its own defects. DOP-PCR generally faithfully retains copy number levels during WGA but its low physical coverage makes it ill suited for detecting SNPs/indels or mutations in the sample genome. On the other hand although MDA is able to achieve high physical coverage, it often causes non-uniformity coverage and distorts the copy number of genomic segments (Y. Wang & Navin, 2015). Another method called multiple annealing- and looping-based amplification (MALBAC) is able to obtain both copy number information and single nucleotide variations but tends to generate high false positive rates (Zong, Lu, Chapman, &

Xie, 2012). All current approaches tend to introduce extensive technical errors and as a result any potential biological signals detected from such procedures need to be carefully validated by more reliable molecular biology techniques such as single-cell qPCR (Y. Wang & Navin, 2015).

The potential impacts of these new developments on SV discovery and genotyping efforts

As mentioned in the previous section, the efforts in sequencing technology innovation is directed largely at preserving the information contained within the sample molecules as much as possible. In the future, the sequencing depth of a library will be derived more from the length of the reads instead of the number of reads while extensive PCR amplifications will be avoided as much as possible. This means that the next generation of SV discovery algorithms will likely be handling libraries with fewer sequencing reads and longer read lengths.

In principle at least, longer reads will make it easier to detect SVs algorithmically because the problem essentially regresses back to the canonical local sequence alignment problem, which has been extensively studied and highly reliable and efficient algorithms such as Smith-Waterman algorithm and its variations have been proposed and implemented. The need for complicated heuristics that seek to cluster sequencing read-pairs and predict the potential SV events that

generate them based on convoluted rules will gradually disappear. The detection of complex genomic rearrangements will also likely to benefit enormously from the movement towards longer sequencing reads. This is because complex genomic rearrangements often lead to read-pairs that cannot be explained by simple SV events and even reads that cannot be mapped to the reference genome and hence eludes conventional SV discovery algorithms. Long sequencing reads may potentially cover the entire region and reveal all the rearrangements at once. The decreasing number of reads, however, may present a challenge for putative SV validation and SV allele frequency estimation because a smaller sample size (number of reads) compromises the powers of statistical models. As a consequence, the focus of future SV detection algorithms may shift away from inferring putative SVs towards validating putative SVs presented in the long reads with sound and rigorous statistical models to avoid false positives introduced by sequencing errors while preserving valid predictions. A potential alternative is to perform short-read targeted sequencing in regions around the SVs predicted by long-reads sequencing experiment. This approach allows one to detect potential SV alleles with long-read sequencing and then validate the predictions (and/or estimate their allele frequencies) with a large number of short reads. With the long-read library providing sequence information at and around the potential SV alleles and the short-read library offering sufficient statistical power for validation and frequency estimation, this combination might yield the best performance.

Concluding remarks

Despite continual progress and improvements in the long-read sequencing and single-cell sequencing technologies, substantial time and efforts are still needed before either technology can mature into accurate, fast and affordable sequencing platforms. At the same time, new datasets are being produced from short-read sequencing platforms at an unprecedented pace thanks to the decreasing cost and improved accessibility for such experiments. For instance, The Cancer Genome Atlas (TCGA) and International Cancer Genome Consortium (ICGC) alone hosted thousands of whole genome sequencing datasets that have yet to be adequately analyzed. Despite all the shortcomings of current short-read sequencing technology, these data if analyzed properly still hold great potential for shedding new light onto various biological mysteries. This is the reason why it is still necessary to focus on analysis that are possible under short-read sequencing technology and continue to design and improve algorithms based on this technology.

Another area that requires extensive efforts in the near future is the functional elucidation and annotation of genetic variants and mutations especially those only affecting intergenic regions. It is well known that the majority of genetic variants/mutations are in intergenic regions but ascertaining whether or how they

alter any physiological processes still remain elusive due to our limited knowledge about the non-coding regions of the genome. That limitation, however, is being shaken by the multiple consortium efforts to interrogate genomic regulatory elements such as ENCODE and Epigenetic Roadmap. The epistatic interaction among variants/mutations is also attracting attentions lately. Phenotypic traits are usually determined by complicated biological pathways and different variants/mutations affecting different components of the pathway may exert synergistic effects on the manifested phenotypic traits. The elucidation of such complex phenomena entails more sophisticated statistical models such as Bayesian Network and much larger sample sizes to ensure sufficient statistical power.

The possibility of obtaining high-quality sequencing result from small amount of DNA samples may eventually allow the detection of mutations from cell-free tumor DNAs (ctDNAs). ctDNAs are DNA fragments released by tumor cells into the blood circulation that may carry cancer specific mutations. The ability to reliably detecting or genotyping ctDNAs holds great promise for developing biopsy-free biomarkers for early cancer diagnosis.

The field of biomedical research is experiencing an exciting transition. With the rapid development of high-throughput technologies researchers are now examining biological systems and disease associated problems in an

increasingly holistic and population-centered manner. This trend promises to fundamentally change the way we understand biological systems and diseases and dramatically improve the efficacy of medicine. Accurate, efficient and well-designed computational algorithms will prove instrumental in making this transition a reality.

REFERENCE:

Abyzov, A., & Gerstein, M. (2011). AGE: defining breakpoints of genomic structural variants at single-nucleotide resolution, through optimal alignments with gap excision. *Bioinformatics (Oxford, England)*, *27*(5), 595–603. http://doi.org/10.1093/bioinformatics/btq713

Alkan, C., Coe, B. P., & Eichler, E. E. (2011). Genome structural variation discovery and genotyping. *Nature Reviews. Genetics*, *12*(5), 363–375. http://doi.org/10.1038/nrg2958

Aparicio, T., Baer, R., & Gautier, J. (2014). DNA double-strand break repair pathway choice and cancer. *DNA Repair*, *19*, 169–175. http://doi.org/10.1016/j.dnarep.2014.03.014

Aravin, A. A., Hannon, G. J., & Brennecke, J. (2007). The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. *Science (New York, N.Y.)*, *318*(5851), 761–764. http://doi.org/10.1126/science.1146484

Babon, J. J., Lucet, I. S., Murphy, J. M., Nicola, N. A., & Varghese, L. N. (2014). The molecular regulation of Janus kinase (JAK) activation. *Biochemical Journal*, *462*(1), 1–13. http://doi.org/10.1007/s11248-014-9795-y

Bailey, J. A., Gu, Z., Clark, R. A., Reinert, K., Samonte, R. V., Schwartz, S., et al. (2002). Recent segmental duplications in the human genome. *Science (New York, N.Y.)*, *297*(5583), 1003–1007. http://doi.org/10.1126/science.1072047

Bailey, J. A., Yavor, A. M., Massa, H. F., Trask, B. J., & Eichler, E. E. (2001). Segmental duplications: organization and impact within the current human genome project assembly. *Genome Research*, *11*(6), 1005–1017. http://doi.org/10.1101/gr.187101

Bamshad, M. J., Ng, S. B., Bigham, A. W., Tabor, H. K., Emond, M. J., Nickerson, D. A., & Shendure, J. (2011). Exome sequencing as a tool for Mendelian disease gene discovery. *Nature Reviews. Genetics*, *12*(11), 745–755. http://doi.org/10.1038/nrg3031

Bao, Z., & Eddy, S. R. (2002). Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Research*, *12*(8), 1269–1276. http://doi.org/10.1101/gr.88502

Bartenhagen, C., & Dugas, M. (2013). RSVSim: an R/Bioconductor package for the simulation of structural variations. *Bioinformatics (Oxford, England)*. http://doi.org/10.1093/bioinformatics/btt198

Baslan, T., & Hicks, J. (2014). Single cell sequencing approaches for complex biological systems. *Current Opinion in Genetics & Development*, *26*, 59–65. http://doi.org/10.1016/j.gde.2014.06.004

Bellos, E., Johnson, M. R., & M Coin, L. J. (2012). cnvHiTSeq: integrative models for high-resolution copy number variation detection and genotyping using population sequencing data. *Genome Biology*, *13*(12), R120.

http://doi.org/10.1186/gb-2012-13-12-r120
Benavente, C. A., & Dyer, M. A. (2015). Genetics and epigenetics of human retinoblastoma. *Annual Review of Pathology*, *10*, 547–562. http://doi.org/10.1146/annurev-pathol-012414-040259
Benjamini, Y., & Hochberg, Y. (1995). JSTOR: Journal of the Royal Statistical Society. Series B (Methodological), Vol. 57, No. 1 (1995), pp. 289-300. *Journal of the Royal Statistical Society Series B …*.
Bennetzen, J. L. (2000). Transposable element contributions to plant gene and genome evolution. *Plant Molecular Biology*, *42*(1), 251–269.
Brennan, C. W., Verhaak, R. G. W., McKenna, A., Campos, B., Noushmehr, H., Salama, S. R., et al. (2013). The Somatic Genomic Landscape of Glioblastoma. *Cell*, *155*(2), 462–477. http://doi.org/10.1016/j.cell.2013.09.034
Britten, R. J. (2010). Transposable element insertions have strongly affected human evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(46), 19945–19948. http://doi.org/10.1073/pnas.1014330107
Brown, J. R. (1974). Shortest alternating path algorithms. *Networks*, *4*(4), 311–334. http://doi.org/10.1002/net.3230040404
Bucheton, A. (1973). [Study of non Mendelian female sterility in Drosophila melanogaster. Hereditary transmission of the degree of efficacy of the reactor factor]. *Comptes Rendus Hebdomadaires Des Séances De l'Académie Des Sciences. Série D: Sciences Naturelles*, *276*(4), 641–644.
Bucheton, A. (1979). Non-Mendelian female sterility in Drosophila melanogaster: influence of aging and thermic treatments. III. Cumulative effects induced by these factors. *Genetics*, *93*(1), 131–142.
Calvo, S. E., Tucker, E. J., Compton, A. G., Kirby, D. M., Crawford, G., Burtt, N. P., et al. (2010). High-throughput, pooled sequencing identifies mutations in NUBPL and FOXRED1 in human complex I deficiency. *Nature Genetics*, *42*(10), 851–858. http://doi.org/10.1038/ng.659
Cancer Genome Atlas Network. (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, *490*(7418), 61–70. http://doi.org/10.1038/nature11412
Cancer Genome Atlas Research Network. (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, *455*(7216), 1061–1068. http://doi.org/10.1038/nature07385
Carreira, P. E., Richardson, S. R., & Faulkner, G. J. (2014). L1 retrotransposons, cancer stem cells and oncogenesis. *The FEBS Journal*, *281*(1), 63–73. http://doi.org/10.1111/febs.12601
Chen, K., Chen, L., Fan, X., Wallis, J., Ding, L., & Weinstock, G. (2013). TIGRA: A Targeted Iterative Graph Routing Assembler for breakpoint assembly. *Genome Research*. http://doi.org/10.1101/gr.162883.113
Chen, K., Wallis, J. W., McLellan, M. D., Larson, D. E., Kalicki, J. M., Pohl, C. S., et al. (2009). BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature Methods*, *6*(9), 677–681.

http://doi.org/10.1038/nmeth.1363
Clark, M. J., Chen, R., Lam, H. Y. K., Karczewski, K. J., Chen, R., Euskirchen, G., et al. (2011). Performance comparison of exome DNA sequencing technologies. *Nature Biotechnology*, *29*(10), 908–914. http://doi.org/10.1038/nbt.1975
Consortium, T. 1. G. P., The 1000 Genomes Consortium Participants are arranged by project role, T. B. I. A. A. F. A. W. I. E. F. P. I. A. P. L. A. I., author, C., committee, S., Medicine, P. G. B. C. O., BGI-Shenzhen, et al. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, *490*(7422), 56–65. http://doi.org/10.1038/nature11632
Cridland, J. M., Macdonald, S. J., Long, A. D., & Thornton, K. R. (2013). Abundance and Distribution of Transposable Elements in Two Drosophila QTL Mapping Resources. *Molecular Biology and Evolution*, *30*(10), 2311–2327. http://doi.org/10.1093/molbev/mst129
Daborn, P. J., Yen, J. L., Bogwitz, M. R., Le Goff, G., Feil, E., Jeffers, S., et al. (2002). A single p450 allele associated with insecticide resistance in Drosophila. *Science (New York, N.Y.)*, *297*(5590), 2253–2256. http://doi.org/10.1126/science.1074170
Dajani, R., Li, J., Wei, Z., Glessner, J. T., Chang, X., Cardinale, C. J., et al. (2015). CNV Analysis Associates AKNAD1 with Type-2 Diabetes in Jordan Subpopulations. *Scientific Reports*, *5*, 13391. http://doi.org/10.1038/srep13391
Decottignies, A. (2013). Alternative end-joining mechanisms: a historical perspective. *Frontiers in Genetics*, *4*, 48. http://doi.org/10.3389/fgene.2013.00048
Eisenstein, M. (2015, May). Startups use short-read data to expand long-read sequencing market. *Nature Biotechnology*, pp. 433–435. http://doi.org/10.1038/nbt0515-433
ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74. http://doi.org/10.1038/nature11247
Escaramís, G., Tornador, C., Bassaganyas, L., Rabionet, R., Tubio, J. M. C., Martínez-Fundichely, A., et al. (2013). PeSV-Fisher: identification of somatic and non-somatic structural variants using next generation sequencing data. *PloS One*, *8*(5), e63377. http://doi.org/10.1371/journal.pone.0063377
Faust, G. G., & Hall, I. M. (2012). YAHA: fast and flexible long-read alignment with optimal breakpoint detection. *Bioinformatics (Oxford, England)*, *28*(19), 2417–2424. http://doi.org/10.1093/bioinformatics/bts456
Forbes, S. A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., et al. (2015). COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Research*, *43*(Database issue), D805–11. http://doi.org/10.1093/nar/gku1075
Futschik, A., & Schlötterer, C. (2010). The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. *Genetics*,

186(1), 207–218. http://doi.org/10.1534/genetics.110.114397
Ghildiyal, M., & Zamore, P. D. (2009). Small silencing RNAs: an expanding universe. *Nature Reviews. Genetics*, *10*(2), 94–108. http://doi.org/10.1038/nrg2504
Gogvadze, E., & Buzdin, A. (2009). Retroelements and their impact on genome evolution and functioning. *Cellular and Molecular Life Sciences : CMLS*, *66*(23), 3727–3742. http://doi.org/10.1007/s00018-009-0107-2
González, J., & Petrov, D. A. (2009). The adaptive role of transposable elements in the Drosophila genome. *Gene*, *448*(2), 124–133. http://doi.org/10.1016/j.gene.2009.06.008
GTEx Consortium. (2015). Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science (New York, N.Y.)*, *348*(6235), 648–660. http://doi.org/10.1126/science.1262110
Handsaker, R. E., Korn, J. M., Nemesh, J., & McCarroll, S. A. (2011). Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nature Genetics*, *43*(3), 269–276. http://doi.org/10.1038/ng.768
Hastings, P. J., Ira, G., & Lupski, J. R. (2009a). A microhomology-mediated break-induced replication model for the origin of human copy number variation. *PLoS Genetics*, *5*(1), e1000327. http://doi.org/10.1371/journal.pgen.1000327
Hastings, P. J., Lupski, J. R., Rosenberg, S. M., & Ira, G. (2009b). Mechanisms of change in gene copy number. *Nature Reviews. Genetics*, *10*(8), 551–564. http://doi.org/10.1038/nrg2593
Hedges, D. J., & Belancio, V. P. (2011). Restless genomes humans as a model organism for understanding host-retrotransposable element dynamics. *Advances in Genetics*, *73*, 219–262. http://doi.org/10.1016/B978-0-12-380860-8.00006-9
Helman, E., Lawrence, M. S., Stewart, C., Sougnez, C., Getz, G., & Meyerson, M. (2014). Somatic retrotransposition in human cancer revealed by whole-genome and exome sequencing. *Genome Research*, *24*(7), 1053–1063. http://doi.org/10.1101/gr.163659.113
Hiraizumi, Y. (1971). Spontaneous recombination in Drosophila melanogaster males. *Proceedings of the National Academy of Sciences of the United States of America*, *68*(2), 268–270.
Hormozdiari, F., Alkan, C., Eichler, E. E., & Sahinalp, S. C. (2009). Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Research*, *19*(7), 1270–1278. http://doi.org/10.1101/gr.088633.108
Hormozdiari, F., Hajirasouliha, I., Dao, P., Hach, F., Yorukoglu, D., Alkan, C., et al. (2010). Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. *Bioinformatics (Oxford, England)*, *26*(12), i350–7. http://doi.org/10.1093/bioinformatics/btq216
Hormozdiari, F., Hajirasouliha, I., McPherson, A., Eichler, E. E., & Sahinalp, S. C.

(2011). Simultaneous structural variation discovery among multiple paired-end sequenced genomes. *Genome Research*, *21*(12), 2203–2212. http://doi.org/10.1101/gr.120501.111

Horpaopan, S., Spier, I., Zink, A. M., Altmüller, J., Holzapfel, S., Laner, A., et al. (2015). Genome-wide CNV analysis in 221 unrelated patients and targeted high-throughput sequencing reveal novel causative candidate genes for colorectal adenomatous polyposis. *International Journal of Cancer. Journal International Du Cancer*, *136*(6), E578–89. http://doi.org/10.1002/ijc.29215

Hu, X., Yuan, J., Shi, Y., Lu, J., Liu, B., Li, Z., et al. (2012). pIRS: Profile-based Illumina pair-end reads simulator. *Bioinformatics (Oxford, England)*, *28*(11), 1533–1535. http://doi.org/10.1093/bioinformatics/bts187

Iqbal, Z., Caccamo, M., Turner, I., Flicek, P., & McVean, G. (2012). De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nature Genetics*, *44*(2), 226–232. http://doi.org/10.1038/ng.1028

Itsara, A., Wu, H., Smith, J. D., Nickerson, D. A., Romieu, I., London, S. J., & Eichler, E. E. (2010). De novo rates and selection of large copy number variation. *Genome Research*, *20*(11), 1469–1481. http://doi.org/10.1101/gr.107680.110

Jacobs, F. M. J., Greenberg, D., Nguyen, N., Haeussler, M., Ewing, A. D., Katzman, S., et al. (2014). An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. *Nature*, *516*(7530), 242–245. http://doi.org/10.1038/nature13760

Jasin, M., & Rothstein, R. (2013). Repair of Strand Breaks by Homologous Recombination. *Cold Spring Harbor Perspectives in Biology*, *5*(11), a012740–a012740. http://doi.org/10.1101/cshperspect.a012740

Ji, Y., Eichler, E. E., Schwartz, S., & Nicholls, R. D. (2000). Structure of chromosomal duplicons and their role in mediating human genomic disorders. *Genome Research*, *10*(5), 597–610.

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., & Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research*, *110*(1-4), 462–467. http://doi.org/10.1159/000084979

Keane, T. M., Wong, K., & Adams, D. J. (2013). RetroSeq: transposable element discovery from next-generation sequencing data. *Bioinformatics (Oxford, England)*, *29*(3), 389–390. http://doi.org/10.1093/bioinformatics/bts697

Kent, W. J. (2002). BLAT--the BLAST-like alignment tool. *Genome Research*, *12*(4), 656–664. http://doi.org/10.1101/gr.229202

Khurana, J. S., & Theurkauf, W. (2010). piRNAs, transposon silencing, and Drosophila germline development. *The Journal of Cell Biology*, *191*(5), 905–913. http://doi.org/10.1083/jcb.201006034

Khurana, J. S., Wang, J., Xu, J., Koppetsch, B. S., Thomson, T. C., Nowosielska, A., et al. (2011). Adaptation to P element transposon invasion in Drosophila melanogaster. *Cell*, *147*(7), 1551–1563. http://doi.org/10.1016/j.cell.2011.11.042

Kidd, J. M., Cooper, G. M., Donahue, W. F., Hayden, H. S., Sampas, N., Graves, T., et al. (2008). Mapping and sequencing of structural variation from eight human genomes. *Nature*, *453*(7191), 56–64. http://doi.org/10.1038/nature06862

Kidwell, M. G. (1985). Hybrid dysgenesis in Drosophila melanogaster: nature and inheritance of P element regulation. *Genetics*, *111*(2), 337–350.

Kofler, R., Betancourt, A. J., & Schlötterer, C. (2012). Sequencing of pooled DNA samples (Pool-Seq) uncovers complex dynamics of transposable element insertions in Drosophila melanogaster. *PLoS Genetics*, *8*(1), e1002487. http://doi.org/10.1371/journal.pgen.1002487

Krejci, L., Altmannova, V., Spirek, M., & Zhao, X. (2012). Homologous recombination and its regulation. *Nucleic Acids Research*, *40*(13), 5795–5818. http://doi.org/10.1093/nar/gks270

la Roche, de, M., Worm, J., & Bienz, M. (2008). The function of BCL9 in Wnt/beta-catenin signaling and colorectal cancer cells. *BMC Cancer*, *8*, 199. http://doi.org/10.1186/1471-2407-8-199

Layer, R. M., Chiang, C., Quinlan, A. R., & Hall, I. M. (2014). LUMPY: A probabilistic framework for structural variant discovery. *Genome Biology*, *15*(6), R84. http://doi.org/10.1186/gb-2014-15-6-r84

Lee, E., Iskow, R., Yang, L., Gokcumen, O., Haseley, P., Luquette, L. J., et al. (2012). Landscape of somatic retrotransposition in human cancers. *Science (New York, N.Y.)*, *337*(6097), 967–971. http://doi.org/10.1126/science.1222077

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, *26*(5), 589–595. http://doi.org/10.1093/bioinformatics/btp698

Li, X., Hastie, A. T., Hawkins, G. A., Moore, W. C., Ampleford, E. J., Milosevic, J., et al. (2015). eQTL of bronchial epithelial cells and bronchial alveolar lavage deciphers GWAS-identified asthma genes. *Allergy*. http://doi.org/10.1111/all.12683

Li, Y., Zheng, H., Luo, R., Wu, H., Zhu, H., Li, R., et al. (2011). structural variation in two human genomes mapped at single-nucleotide resolution by whole genomede novo assembly. *Nature Biotechnology*, *29*(8), 725–732. http://doi.org/10.1038/nbt.1904

Liang, L., Fang, J.-Y., & Xu, J. (2015). Gastric cancer and gene copy number variation: emerging cancer drivers for targeted therapy. *Oncogene*. http://doi.org/10.1038/onc.2015.209

Linheiro, R. S., & Bergman, C. M. (2012). Whole genome resequencing reveals natural target site preferences of transposable elements in Drosophila melanogaster. *PloS One*, *7*(2), e30008. http://doi.org/10.1371/journal.pone.0030008

Lynch, V. J., Leclerc, R. D., May, G., & Wagner, G. P. (2011). Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. *Nature Genetics*, *43*(11), 1154–1159.

http://doi.org/10.1038/ng.917
Macdonald, J. R., Ziman, R., Yuen, R. K. C., Feuk, L., & Scherer, S. W. (2013). The database of genomic variants: a curated collection of structural variation in the human genome. *Nucleic Acids Research*. http://doi.org/10.1093/nar/gkt958
Mackay, T. F. C., Richards, S., Stone, E. A., Barbadilla, A., Ayroles, J. F., Zhu, D., et al. (2012). The Drosophila melanogaster Genetic Reference Panel. *Nature*, *482*(7384), 173–178. http://doi.org/10.1038/nature10811
Malhotra, A., Lindberg, M., Faust, G. G., Leibowitz, M. L., Clark, R. A., Layer, R. M., et al. (2013). Breakpoint profiling of 64 cancer genomes reveals numerous complex rearrangements spawned by homology-independent mechanisms. *Genome Research*, *23*(5), 762–776. http://doi.org/10.1101/gr.143677.112
McCLINTOCK, B. (1950). The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Sciences of the United States of America*, *36*(6), 344–355.
Mehta, A., & Haber, J. E. (2014). Sources of DNA double-strand breaks and models of recombinational DNA repair. *Cold Spring Harbor Perspectives in Biology*, *6*(9), a016428. http://doi.org/10.1101/cshperspect.a016428
Mills, R. E., Bennett, E. A., Iskow, R. C., & Devine, S. E. (2007). Which transposable elements are active in the human genome? *Trends in Genetics*, *23*(4), 183–191. http://doi.org/10.1016/j.tig.2007.02.006
Mills, R. E., Walter, K., Stewart, C., Handsaker, R. E., Chen, K., Alkan, C., et al. (2011). Mapping copy number variation by population-scale genome sequencing. *Nature*, *470*(7332), 59–65. http://doi.org/10.1038/nature09708
Naranbhai, V., Fairfax, B. P., Makino, S., Humburg, P., Wong, D., Ng, E., et al. (2015). Genomic modulators of gene expression in human neutrophils. *Nature Communications*, *6*, 7545. http://doi.org/10.1038/ncomms8545
Nefedova, L. N., Mannanova, M. M., & Kim, A. I. (2011). Integration specificity of LTR-retrotransposons and retroviruses in the Drosophila melanogaster genome. *Virus Genes*, *42*(2), 297–306. http://doi.org/10.1007/s11262-010-0566-4
Ottaviani, D., Lecain, M., & Sheer, D. (2014). The role of microhomology in genomic structural variation. *Trends in Genetics : TIG*, *30*(3), 85–94. http://doi.org/10.1016/j.tig.2014.01.001
Ovarian Cancer Association Consortium, Australian Cancer Study, Australian Ovarian Cancer Study Group. (2015). Genome-wide significant risk associations for mucinous ovarian carcinoma. *Nature Genetics*, *47*(8), 888–897. http://doi.org/10.1038/ng.3336
Pevzner, P. (2000). Computational Molecular Biology. MIT Press.
Price, A. L., Jones, N. C., & Pevzner, P. A. (2005). De novo identification of repeat families in large genomes. *Bioinformatics (Oxford, England)*, *21 Suppl 1*, i351–8. http://doi.org/10.1093/bioinformatics/bti1018
Quinlan, A. R., Boland, M. J., Leibowitz, M. L., Shumilina, S., Pehrson, S. M.,

Baldwin, K. K., & Hall, I. M. (2011). Genome sequencing of mouse induced pluripotent stem cells reveals retroelement stability and infrequent DNA rearrangement during reprogramming. *Cell Stem Cell*, *9*(4), 366–373. http://doi.org/10.1016/j.stem.2011.07.018

Quinlan, A. R., Clark, R. A., Sokolova, S., Leibowitz, M. L., Zhang, Y., Hurles, M. E., et al. (2010). Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. *Genome Research*, *20*(5), 623–635. http://doi.org/10.1101/gr.102970.109

Rausch, T., Zichner, T., Schlattl, A., Stütz, A. M., Benes, V., & Korbel, J. O. (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics (Oxford, England)*, *28*(18), i333–i339. http://doi.org/10.1093/bioinformatics/bts378

Ren, Y., Zhang, Y., Liu, R. Z., Fenstermacher, D. A., Wright, K. L., Teer, J. K., & Wu, J. (2013). JAK1 truncating mutations in gynecologic cancer define new role of cancer-associated protein tyrosine kinase aberrations. *Scientific Reports*, *3*. http://doi.org/10.1038/srep03042

Sampietro, J., Dahlberg, C. L., Cho, U. S., Hinds, T. R., Kimelman, D., & Xu, W. (2006). Crystal structure of a beta-catenin/BCL9/Tcf4 complex. *Molecular Cell*, *24*(2), 293–300. http://doi.org/10.1016/j.molcel.2006.09.001

Sengoku, T., & Yokoyama, S. (2011). Structural basis for histone H3 Lys 27 demethylation by UTX/KDM6A. *Genes & Development*, *25*(21), 2266–2277. http://doi.org/10.1101/gad.172296.111

Sharp, A. J., Locke, D. P., McGrath, S. D., Cheng, Z., Bailey, J. A., Vallente, R. U., et al. (2005). Segmental duplications and copy-number variation in the human genome. *The American Journal of Human Genetics*, *77*(1), 78–88. http://doi.org/10.1086/431652

Shukla, R., Upton, K. R., Muñoz-Lopez, M., Gerhardt, D. J., Fisher, M. E., Nguyen, T., et al. (2013). Endogenous retrotransposition activates oncogenic pathways in hepatocellular carcinoma. *Cell*, *153*(1), 101–111. http://doi.org/10.1016/j.cell.2013.02.032

Sindi, S. S., Önal, S., Peng, L. C., Wu, H.-T., & Raphael, B. J. (2012). An integrative probabilistic model for identification of structural variation in sequencing data. *Genome Biology*, *13*(3), R22. http://doi.org/10.1186/gb-2012-13-3-r22

Siomi, M. C., Sato, K., Pezic, D., & Aravin, A. A. (2011). PIWI-interacting small RNAs: the vanguard of genome defence. *Nature Reviews. Molecular Cell Biology*, *12*(4), 246–258. http://doi.org/10.1038/nrm3089

Solyom, S., Ewing, A. D., Rahrmann, E. P., Doucet, T., Nelson, H. H., Burns, M. B., et al. (2012). Extensive somatic L1 retrotransposition in colorectal tumors. *Genome Research*, *22*(12), 2328–2338. http://doi.org/10.1101/gr.145235.112

Song, L., Rawal, B., Nemeth, J. A., & Haura, E. B. (2011). JAK1 Activates STAT3 Activity in Non-Small-Cell Lung Cancer Cells and IL-6 Neutralizing Antibodies Can Suppress JAK1-STAT3 Signaling. *Molecular Cancer Therapeutics*, *10*(3), 481–494. http://doi.org/10.1158/1535-7163.MCT-10-0502

Stephens, P. J., Greenman, C. D., Fu, B., Yang, F., Bignell, G. R., Mudie, L. J., et al. (2011). Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell*, *144*(1), 27–40. http://doi.org/10.1016/j.cell.2010.11.055

Stranger, B. E., Stahl, E. A., & Raj, T. (2011). Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics*, *187*(2), 367–383. http://doi.org/10.1534/genetics.110.120907

Suzuki, H., Aoki, K., Chiba, K., Sato, Y., Shiozawa, Y., Shiraishi, Y., et al. (2015). Mutational landscape and clonal architecture in grade II and III gliomas. *Nature Genetics*, *47*(5), 458–468. http://doi.org/10.1038/ng.3273

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

The Cancer Genome Atlas Research Network. (2013). Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia. *New England Journal of Medicine*, *368*(22), 2059–2074. http://doi.org/10.1056/NEJMoa1301689

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., et al. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols*, *7*(3), 562–578. http://doi.org/10.1038/nprot.2012.016

Truong, L. N., Li, Y., Shi, L. Z., Hwang, P. Y.-H., He, J., Wang, H., et al. (2013). Microhomology-mediated End Joining and Homologous Recombination share the initial end resection step to repair DNA double-strand breaks in mammalian cells. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(19), 7720–7725. http://doi.org/10.1073/pnas.1213431110

Tsuchiya, T., & Eulgem, T. (2013). An alternative polyadenylation mechanism coopted to the Arabidopsis RPP7 gene through intronic retrotransposon domestication. *Proceedings of the National Academy of Sciences of the United States of America*. http://doi.org/10.1073/pnas.1312545110

Tubio, J. M. C., Li, Y., Ju, Y. S., Martincorena, I., Cooke, S. L., Tojo, M., et al. (2014). Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. *Science (New York, N.Y.)*, *345*(6196), 1251343–1251343. http://doi.org/10.1126/science.1251343

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., & Kinzler, K. W. (2013). Cancer genome landscapes. *Science (New York, N.Y.)*, *339*(6127), 1546–1558. http://doi.org/10.1126/science.1235122

Wallweber, H. J. A., Tam, C., Franke, Y., Starovasnik, M. A., & Lupardus, P. J. (2014). Structural basis of recognition of interferon. *Nature Publishing Group*, *21*(5), 443–448. http://doi.org/10.1038/nsmb.2807

Wang, J., Mullighan, C. G., Easton, J., Roberts, S., Heatley, S. L., Ma, J., et al. (2011). CREST maps somatic structural variation in cancer genomes with base-pair resolution. *Nature Methods*, *8*(8), 652–654.

http://doi.org/10.1038/nmeth.1628
Wang, S. R., Carmichael, H., Andrew, S. F., Miller, T. C., Moon, J. E., Derr, M. A., et al. (2013a). Large-scale pooled next-generation sequencing of 1077 genes to identify genetic causes of short stature. *The Journal of Clinical Endocrinology and Metabolism*, *98*(8), E1428–37. http://doi.org/10.1210/jc.2013-1534
Wang, X., Weigel, D., & Smith, L. M. (2013b). Transposon variants and their effects on gene expression in Arabidopsis. *PLoS Genetics*, *9*(2), e1003255. http://doi.org/10.1371/journal.pgen.1003255
Wang, Y., & Navin, N. E. (2015). Advances and applications of single-cell sequencing technologies. *Molecular Cell*, *58*(4), 598–609. http://doi.org/10.1016/j.molcel.2015.05.005
Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., et al. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, *42*(Database issue), D1001–6. http://doi.org/10.1093/nar/gkt1229
Weterings, E., & Chen, D. J. (2008). The endless tale of non-homologous end-joining. *Cell Research*, *18*(1), 114–124. http://doi.org/10.1038/cr.2008.3
Yang, L., Luquette, L. J., Gehlenborg, N., Xi, R., Haseley, P. S., Hsieh, C.-H., et al. (2013). Diverse Mechanismsof Somatic Structural Variations in Human Cancer Genomes. *Cell*, *153*(4), 919–929. http://doi.org/10.1016/j.cell.2013.04.010
Ye, K., Schulz, M. H., Long, Q., Apweiler, R., & Ning, Z. (2009). Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics (Oxford, England)*, *25*(21), 2865–2871. http://doi.org/10.1093/bioinformatics/btp394
Zain, S. M., Mohamed, Z., Pirmohamed, M., Tan, H. L., Alshawsh, M. A., Mahadeva, S., et al. (2015). Copy number variation in exportin-4 (XPO4) gene and its association with histological severity of non-alcoholic fatty liver disease. *Scientific Reports*, *5*, 13306. http://doi.org/10.1038/srep13306
Zhang, F., Dai, L., Lin, W., Wang, W., Liu, X., Zhang, J., et al. (2015). Exome sequencing identified FGF12 as a novel candidate gene for Kashin-Beck disease. *Functional & Integrative Genomics*. http://doi.org/10.1007/s10142-015-0462-z
Zhou, X., Ren, L., Meng, Q., Li, Y., Yu, Y., & Yu, J. (2010). The next-generation sequencing technology and application. *Protein & Cell*, *1*(6), 520–536. http://doi.org/10.1007/s13238-010-0065-3
Zhou, Z.-W., Cui, L.-L., Han, L., Wang, C., Song, Z.-J., Shen, J.-W., et al. (2015). Polymorphisms in GCKR, SLC17A1 and SLC22A12 were associated with phenotype gout in Han Chinese males: a case-control study. *BMC Medical Genetics*, *16*(1), 66. http://doi.org/10.1186/s12881-015-0208-8
Zhu, Y., Bergland, A. O., González, J., & Petrov, D. A. (2012). Empirical validation of pooled whole genome population re-sequencing in Drosophila melanogaster. *PloS One*, *7*(7), e41901.

http://doi.org/10.1371/journal.pone.0041901
Zichner, T., Garfield, D. A., Rausch, T., Stütz, A. M., Cannavó, E., Braun, M., et al. (2013). Impact of genomic structural variation in Drosophila melanogaster based on population-scale sequencing. *Genome Research*, *23*(3), 568–579. http://doi.org/10.1101/gr.142646.112
Zong, C., Lu, S., Chapman, A. R., & Xie, X. S. (2012). Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science (New York, N.Y.)*, *338*(6114), 1622–1626. http://doi.org/10.1126/science.1229164