ASSESSMENTS OF INDEL ANNOTATION PROGRAMS AND COMPARATIVE SOMATIC INDEL ANALYSIS IN CANCER GENOMES

by

Jing Chen

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Bioinformatics and Computational Biology

Charlotte

2020

Approved by:

_____

Dr. Jun-tao Guo

_____

Dr. Zhengchang Su

_____

Dr. Xinghua Shi

_____

Dr. Bao-Hua Song

ABSTRACT

JING CHEN. Assessments of indel annotation programs and comparative somatic indel analysis in cancer genomes. (Under the direction of DR. JUN-TAO GUO)

Insertions and deletions (indels) represent the second largest variation type in human genomes and have been implicated in the development of cancer. Accurate indel annotation is of paramount importance in variants analysis in both healthy and disease genomes. Previous studies have shown that existing indel calling methods generally produce high false positives and false negatives, which limits the downstream investigation of the roles of indels in structural and functional effects.

To assess the accuracy of indel calling programs, we carried out a comparative analysis by evaluating 7 general indel calling programs and 4 somatic indel calling programs, using 78 healthy healthy human genomes from the 1000 Genomes Project and 30 cancer samples from The Cancer Genome Atlas (TCGA). We adopted a comprehensive and more stringent indel comparison approach, and an efficient way to use a benchmark for improved performance comparisons for the general indel calling programs. We found that germline indels in healthy genomes derived by combining several indel calling tools could help remove a large number of false positive indels from individual programs without compromising the number of true positives. The performance comparisons of somatic indel calling programs are more complicated due to the lack of a reliable and comprehensive benchmark.

We further performed a comparative analysis of somatic indels in two cancer types, BRCA and LUAD. We compared somatic indels in both coding and non-coding regulatory regions such as transcription factor binding sites (TFBSs). We used an improved algorithm to predict TFBSs in human genomes and analyzed their evolutionary and structural roles. Our comparative results indicated that while there are differences between LUAD and BRCA genomes, both of them show a higher deletion

rate, coding indel rate and frame-shift indel rate. Somatic indels tend to locate in sequences with important functions, including both coding and non-coding regions. This study can serve as the first step in future pan-cancer analysis for identifying key variant markers of cancer genomes.

ACKNOWLEDGEMENTS

I have received a large amount of help and support throughout my Ph.D. study.

Firstly, I would like to thank my advisor, Dr. Jun-tao Guo. He is knowledgeable and supportive of my Ph.D. research. His patience and enthusiasm make him a great advisor. We have many insightful discussions and I learnt a lot from him during the past 5 years.

I would like to acknowledge my committee members: Dr. Zhengchang Su, Dr. Xinghua Shi and Dr. Bao-Hua Song for their support and guidance. They give me valuable suggestions and encouragements.

I would like to thank my lab mates. They are excellent research collaborators and are always willing to answer any research questions. Also, I thank my friends in UNCC for the wonderful time we shared.

In addition, I would like to thank my families: my parents and boyfriend for their support and accompany during my Ph.D. study.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

3'UTR 3'-untranslated regions.

5'UTR 5'-untranslated regions.

bp Base pair.

CDS Coding-sequence.

cEH Codonic enhancer.

ChIP-chip Chromatin immunoprecipitation on microarray.

ChIP-seq Chromatin Immunoprecipitation Sequencing.

COSMIC Catalogue Of Somatic Mutations In Cancer.

CRM Cis-regulatory module.

eExons Exons overlap with enhancers.

FN False negative.

FP False positive.

FS Frame-Shift.

GATK_HC GATK HaplotypeCaller.

GATK_UG GATK Unified Genotyper.

HGSV Human Genome Structural Variation project.

Indel Insertion and deletion.

NCBI National Cancer for Biotechnology Information.

NFS Non-Frame-Shift.

NGS Next-generation sequencing.

NHGRI National Human Genome Research Institute.

NMD Nonsense-mediated mRNA decay.

SNP Single nucleotide polymorphism.

SV Structural variation.

TCGA The Cancer Genome Atlas.

TF Transcription factor.

TFBS Transcription factor binding site.

TP True positive.

CHAPTER 1: Introduction

Human genome consists of coding and non-coding DNA sequences. While the coding sequences generally encode for proteins, the non-coding sequences contain important elements involved in regulation of gene expression, such as transcription factor binding sites (TFBSs) [1, 2, 3]. Comparative genome analysis has revealed a large number of variants, including genetic variants from germline cells and somatic variants [4, 5]. In this dissertation research, we focus on the second largest variation type, insertion and deletion (indel), in human genomes, especially the effects of somatic indels in human cancer genomes. We first carried out a comparative assessment on indel calling programs, including both general indel calling programs and somatic indel calling programs (Chapter 2). We then investigated somatic indels in two cancer genomes, including both the effect of coding indels on protein structures and the overlap between non-coding indels and TFBS (Chapter 3). In Chapter 4, we analyzed the evolution of TFBS in human genomes and their roles in protein structures.

In this chapter, we first review the variants in human genomes and the current indel calling algorithms. We then introduce the state-of-art of genome level prediction of TFBSs and their roles in genetic analysis.

1.1 Variants in human genomes

1.1.1 An overview of variant types and databases

The majority of the DNA sequences are consistent between two individuals and the remaining sequences contribute to the uniqueness of the individual traits [5, 6, 7, 8]. There are three major types of variations in human genomes: single nucleotide polymorphisms (SNPs) [9, 10, 11], insertions and deletions (indels) [7, 8, 12] and

structural variations (SVs) [13, 14, 15]. With the advancement of biotechnology, genome sequencing becomes faster and cheaper, making it easier to quickly sequence more genomes in a very short time and reveal more and novel genome variations. In 2001, the international SNP map working group reported over 1 million SNPs in human genomes [10]. It has been demonstrated that single nucleotide differentiation appears in every 2,000 - 3,000 base pairs (bps) when two human genomes are compared [10]. Mills *et al.* reported 415,436 unique indels from 36 human genomes in 2006 and 2 million in their updated analysis with 79 genomes in 2011 [7, 8]. The Human Genome Structural Variation project (HGSV) identified 1,695 sites of SVs from 8 individuals in 2008 [16]. Besides these studies, the 1000 Genomes Project provided a map of variations in human genomes from 26 populations. The consortium carried out a comparative study among different type of variations in 2,504 human genomes [5, 6, 17, 18]. The results revealed that the ratio of the number of these variations is approximately 750 SNPs : 50 indels : 1 SVs [6]. At the genome level, around 250 - 300 loss of function variants could be found in each person, including frame shift variations, early stops, etc. [6]. At the population level, significant differential variation sites have been detected between populations [6]. For example, a missense variant in gene SLC24A5 distinguished CEU (Utah Residents (CEPH) with Northern and Western European Ancestry) to CHB + JPT (Han Chinese in Beijing, China + Japanese in Tokyo, Japan) [6]. Family trio study indicated that the *de novo* germline mutation rate is about 10^{-8} per bp per generation [6].

dbSNP is the most commonly used database for SNPs and short indels from a large number of species and is organized by the National Human Genome Research Institute (NHGRI) and the National Center for Biotechnology Information (NCBI) [19]. dbVar, which is also established by NCBI, is another variation database that includes insertions, deletions, duplications, inversions, mobile elements, and translocation sites [20]. Currently, dbSNP contains 1,803 million variations and dbVar contains 34.6 mil-

lion variations [19, 20]. There are also clinical related variation databases. ClinVar is designed to explore the relationship between genome variations and human health [21]. The variations in this database are collected from patients and can be used in the clinical studies [21].

1.1.2 Variants in human cancer genomes

Cancer represents a group of diseases characterized by abnormal and uncontrolled cell growth [4]. Under normal conditions, the rates of cell growth and death are delicately controlled to maintain the number of cells. Once the balance is broken and the growth rate is abnormally increased, tumor will be formed. Tumors can be benign tumors or malignant tumors, also called cancer. A benign tumor is not invasive, meaning although there is an abnormal growth of cells, these cells will not spread out to other parts of the body or influence other tissues. Cancer, on the other hand, is much more dangerous. The malignant tumor may destroy normal tissues, migrate to other parts of the body and form new tumors. The benign tumor could evolve to the malignant tumor by several variants on certain sites of genes [4].

The cause for cancer is complicated. A large number of cancer-related studies showed that tumorigenesis may be initiated by variations occurred in several key genes called driver genes, at the genome level [4, 22, 23]. The inherited variations are called germline variations and the newly developed variations are called somatic variations in cancer genomes [4]. While the majority of cancer-related studies revealed a number of somatic variations and their possible roles in cancer development, germline variations have also been implicated in cancer development. Germline variations occurred in gene TP53 may cause Li-Fraumeni syndrome, for instance [24]. The mutations in cancer can also be divided into driver mutations and passenger mutations depending on whether or not the selective growth advantage is affected [4]. Genes with driver mutations are called driver genes [4]. For example, TP53 is a consensus driver gene and the somatic variations in TP53 were found in almost 50% of cancers, including

breast cancer, bladder cancer, ovarian cancer, lung cancer, etc., and one single amino acid change in the p53 protein (like R175H) can lead to an uncontrolled cell growth [24, 25, 26, 27].

The Catalogue Of Somatic Mutations In Cancer (COSMIC) is a commonly used comprehensive database designed to analyze the effect of somatic variations in cancer genomes [28]. Cancer gene census is an ongoing part of the COSMIC project, which collects genes with known relationship with cancers. COSMIC contains more than 8 million variations and covers the majority of variation types [28].

1.2 Indel and indel calling method

1.2.1 Indels in human genomes

Indel is the second largest variant type in human genomes and has been implicated in a number of diseases [12, 29, 30, 31]. Indel refers to insertion or deletion of 1 to 10,000 bp in genomes [7]. An indel with less than 50 bp is called a microindel, and an indel with less than 1,000 bp are usually called a small indel [7, 12, 32, 33]. Some of the large indels may also be considered as SVs since typically the length of SVs is from 1,000 to 3,000,000 bp [15, 34]. Studies have estimated that indels contribute to 16% to 25% of sequence polymorphisms in populations [7, 35, 36, 37]. Like other types of variations, indels can lead to diseases, such as cancer, and alter human traits [12, 38, 39].

The first large-scale indel analysis in the human genome only concentrated to chromosome 22 with ABI resequencing data from 31 samples and reported that 13% of the variants are indels [40]. Later, a genome-wide study identified 2,000 indels, which represents about 20% variants in human genomes [30]. In 2006, Mills *et al.* published a large-scale indel discovery pipeline and reported 415,436 indels from 36 samples [7]. Two years later, Kidd *et al.* reported 796,273 indels in their project, however only around 40,000 indels are consistent with the previous indel set [16]. In 2010, the 1000 Genomes Project published their pilot analysis and reported a total of 1.48 million

indels [6]. In an updated study, Mills *et al.* reported around 2 million indels from 79 samples [8]. These two indel sets have 463,377 common indels [6, 8]. The latest analysis from the 1000 Genomes Project with 2,504 samples revealed 3.6 million of small indels [5].

Indels in different personal genomes vary greatly. In 2007, Levy *et al.* carried out an individual genome analysis and reported 823,396 indels, ranging from 1bp to over 80,000bp in length [41]. Another study reported only 135,262 indels in an Asian individual genome with very short length (from 1bp to 3 bp) [42]. The differences may be due to different sequencing data types and different indel discovery methods/algorithms [12].

A number of studies have been carried out to investigate the structural and functional effects of indels. Chaux *et al.* identified 517,812 indels from the Ensembl database [43]. They found that 724 indels are located in the coding regions and are enriched in loop regions of the proteins [43]. We recently performed a study on coding indels from the 1000 Genomes Project and also found that indels in the coding regions prefer coil secondary structure types [32]. The number of frame-shift (FS) indels are less than expected, which is in agreement with the results by Mills *et al.* [7, 32]. It has been reported that indels have significantly lower density in exons (including UTRs and CDS regions) and CpG islands compared with SNPs [44]. Indel density rate is lower than SNP in TFBSs, suggesting critical roles of indels in non-coding regions [45].

The impact of indels on gene functions in cancer genomes has been explored in several studies. In an analysis of non-frame shift (NFS) indels in cancer genomes, Pagel *et al.* found that pathogenic indels are enriched in helix and strand regions of proteins [46]. In addition, the functional mechanisms of somatic indels may depend on specific cancer types [46]. One study by Yang *et al.* examined the distribution of coding somatic indels in cancer genomes and found that indels tend to locate close

to somatic SNPs within the same patient genome [47]. Non-coding somatic indels in lung adenocarcinomas are reported to target surfactant protein genes [48]. However, a study conducted by Nakagomi *et al.* showed that only 25.7% lung cancer patients have non-coding indels in surfactant-encoding genes [49].

1.2.2 Indel calling programs

One of the key steps in studying indel variants in human genomes is the accurate annotation of indels. Next generation sequencing technology (NGS) has been used to produce large scale sequencing data in recent years and a number of computational programs have been developed to call variants using a variety of algorithms. There are several different types of genome sequencing techniques that may affect large scale identification of indels. The sequencing techniques include traditional ABI reads, 454 reads and Illumina reads. PCR is the most commonly used experimental method for indel validation [12, 41, 42].

In 2010, Mullaney *et al.* reviewed several computational indel calling programs, including SOAP, MAQ, BAW, Pindel, Bowtie and BFAST [12, 50, 51, 52, 53, 54, 55]. Their study showed that these tools have false negative rates ranging from 0.1 to 0.35, indicating a large number of undiscovered indels in human genomes [12]. Hasen *et al.* performed comparative evaluations on 7 indel calling programs (GATK Unified Genotyper, Dindel, Pindel, SAMtools, GATK HaplotypeCaller, VarScan, and Platypus) [53, 56, 57, 58, 59, 60, 61, 62]. They found these programs have issues with high false negative (FN) rates as well as high false positive (FP) rates, and there are only a small proportion of indels that can be called by all tools [56]. In 2018, Li *et al.* evaluated 4 indel calling programs (Platypus, VarScan2, Scalpel, GATK HaplotypeCaller, and GotCloud) with both simulation data and real sequencing data, and they showed that GATK HaplotypeCaller has the best performance [61, 62, 63, 64, 65, 66].

Besides these general indel calling programs, there are also tools designed for so-

matic variants discovery. However, a systematic evaluation on somatic indel specific calling tools is still lacking. Current studies mainly focused on the program performances on somatic SNP callings and very little attention has been paid to the accuracy of somatic indel calling.

1.3 TFBS in human genomes

Transcription is an important biological process, in which DNA is transcribed into RNA molecules. Transcription factors (TFs), a special class of proteins, bind to DNA sequences, called transcription factor binding sites (TFBSs), and regulate gene expression [67, 68]. In addition to non-coding regions, recent studies have shown that TFBSs exist in coding regions, suggesting that these DNA sequences may have dual functions [69, 70]. In this section, we will discuss the TFBSs in human genomes and the corresponding evolution constrains.

1.3.1 TFBS

The expression of genes is controlled by the gene regulatory machinery including RNA polymerases and transcription factors and their corresponding binding sites. A cluster of TFBSs, including enhancers, promoters, silencers, etc., is called a cis-regulatory module (CRM) [71, 72]. Identification of CRMs/TFBSs represents a crucial step in genomic analysis. There are several databases of CRMs/TFBSs derived from different methods [73, 74]. For example, the VISTA Enhancer Browser is a commonly used data source for experimentally verified functional, tissue-specific enhancers in human and mice genomes [75]. Since the enhancers are experimentally verified, this dataset can be used as a benchmark to assess enhancer prediction methods and algorithms [76]. In 2017, Fishilevich *et al.* developed GeneHancer with 285,000 enhancer elements by integrating four enhancer data sources [77]. EnhancerAtlas is another enhancer database and the updated version EnhancerAtlas 2.0 has 13,494,603 enhancers in its current release [78, 79].

A number of experimental approaches have been developed for identifying enhancers in genomes. For example, chromatin immunoprecipitation sequencing (ChIP-seq) and the union of chromatin immunoprecipitation and whole-genome DNA microarrays (ChIP-chip) data can identify different types of TFBSs by chromatin immunoprecipitation [78]. Some factors that are involved in the binding process can also provide information about the locations of binding sites in the genomes, such as enhancer specific factors EP300 and RNA polymerase II [78, 80, 81]. DNase I hypersensitivity sites are the open chromatin regions that may contain TF binding sites [82]. Hi-C technology can help detect enhancer-promoter interactions, using histone modification patterns to identify enhancers, FAIRE, eRNAs data, etc. [83, 84, 85, 86].

1.3.2 TFBS in genome evolutions

More experimental data have revealed the overlap between coding sequences and regulatory regions, suggesting some DNA fragments may have dual functions [69, 70]. They can serve as coding sequences to produce functional proteins and as TFBSs in regulation of gene expressions [69]. However, the percentage of such DNA sequences varies in species and by different CRMs/TFBSs annotations. Birnbaym *et al.* showed that there are 7% and 6% of binding peaks located in protein coding regions in human genomes and mice genomes respectively, based on ChIP-seq data [87]. Using DNase I footprinting method, Stergachis *et al.* found that 14% of coding regions in human genomes can bind TFs [70]. Moreover, coding enhancers not only can regulate itself, they can also regulate nearby genes [69, 88].

A recent study showed that some of the SNPs that are located on eExons (exons overlap with enhancers) may change the enhancer activity, and the corresponding sequences encode important proteins [87]. This study also revealed that both synonymous and non-synonymous mutations can affect enhancer activities [87]. Stergachis *et al.* compared the coding sequences that only coding proteins with those eExons and found that eExons are more conserved [70]. The third position of the codons

within eExons is constrained by the additional function [70]. However, Xing *et al.* reanalyzed the eExons data and showed that the conservation is a result of the base bias [89]. When G/C and A/T are separately considered, there are no significant differences between TF-bound and TF-depleted codons, suggesting eExons are evolutionary neutral [89]. Agoglia *et al.* performed a selection analysis on three exonic enhancers and showed similar results to those from Xing *et al.* Their results also indicated that the enhancer function may not affect protein evolution [90]. Birnbaum *et al.* reported that there are no differences between the effects of synonymous and non-synonymous SNVs, suggesting that enhancer activity does not have evolutionary constrains [87].

CHAPTER 2: Comparative Assessments of Indel Annotations in Healthy and Cancer Genomes with Next-generation Sequencing Data

2.1 Background

Insertion and deletion (indel) is the second-largest genetic variation type in human genomes. On average, one healthy genome differs from the reference genome at about 566,000 sites with indel lengths ranging from 1 to 1,000 base pairs (bps) [5]. Typically, small indels are termed for insertions/deletions of shorter than 50 bps while longer ones are considered as structural variants (SVs) [91, 92]. Besides contributing to genetic variations in healthy population, deleterious indels in both coding and non-coding regions can lead to various types of diseases. For example, coding indels were identified in breast cancer development genes, including AKT1, BRCA1 and CDH1, and the fragile X syndrome is caused by a large insertion in 5'UTR of the FMR1 gene [93, 94]. Several databases with annotated indels have been developed to document these variants, including dbSNP, dbVar, and COSMIC (the Catalogue Of Somatic Mutation In Cancer) [19, 20, 28].

Detection of genomic variations including indels represents one of the most important aspects in human genome analysis. Mills *et al.* reported 2 million unique indels in their updated analysis of 79 genomes in 2011 [7, 8]. The indel set from the 79 genomes is commonly used as a reference for indel analysis in healthy genomes since these indels were annotated with Sanger sequencing data, which reported a 97.2% validation rate [8]. There are also studies focused on somatic indels in cancer genomes. For example, Niu *et al.* analyzed 4,201 non-frame-shift indels and identified more than 6000 mutation clusters on protein 3-dimentional (3D) structures across 19 cancer types [95]. Besides somatic coding indels, non-coding indels also play important

roles in cancer genomes. Imielinski *et al.* found that non-coding somatic indels tend to be enriched in lineage-defining genes in multiple cancer genomes [48].

Next-generation sequencing (NGS) technology has reduced the sequencing cost and produced more genome sequence data. A number of programs have been developed both germline indel and somatic indel identification from NGS data [56, 96, 97]. Current indel calling programs use different algorithms to distinguish sequence errors or alignment errors from real indel variations [13]. General indel calling programs are classified into five major groups: alignment-based methods, split read mapping methods, paired end mapping methods, haplotype based methods, and machine learning-based approaches [13, 56]. A list of indel calling programs with variant types that can be detected and the corresponding algorithms are shown in Table 2.1 [53, 60, 98, 57, 58, 59, 62, 99, 100, 64]. Alignment-based methods, including Dindel, GATK_UG, SAMTools and Varscan, use information from the mapping step and performed statistical approach to find indels [56]. These alignment-based programs differ in the statistical models and processing details [59]. The indel sizes from these alignment-based programs are constrained by the length of sequence reads. Consequently the medium sized indels and large insertions are hard to detect since the workflow relies on the initial alignments [13]. Split read mapping methods, such as Pindel, rely on the discordant reads in the alignment step and can be used to annotate medium sized indels. These methods usually do not use statistical approaches to filter variants [53]. The haplotype-based methods, such as GATK_HC and Platypus, collect candidate haplotypes and identify the variants based on the realignment results on haplotypes [56]. Paired-end read mapping method compares the real and expected distances between paired-end reads to identify potential indels. Indel calling tools used this method can find indel positions. However the exact indel sequences are usually hard to annotate. They are considered more accurate for medium sized indels but not for small indels. Machine learning methods need training data to pre-

Table 2.1: A list of indel calling programs

Programs	General/Somatic	Type of variants	Core algorithms	Notes and references
Dindel	General	Indel	Alignment-based	Bayesian approach [60]
GATK_HC	General	SNP + Indel	Haplotype-based	Collection of candidate haplotypes [61]
GATK_UG	General	SNP + Indel	Alignment-based	Bayesian genotype likelihood model [57]
Pindel	General	Indel	Split read mapping	A pattern growth approach [53]
Platypus	General	SNP + Indel	Haplotype-based	Collection of candidate haplotypes [62]
SAMTools	General	SNP + Indel	Alignment-based	Bayesian model [59]
Varscan	General	SNP + Indel	Alignment-based	Heuristic method [58]
GATK Mutect2	Somatic	SNP + Indel	Allele frequency	Reassembly of haplotypes methods [98, 101]
Strelka	Somatic	SNP + Indel	Allele frequency	Bayesian approach [99]
Strelka2	Somatic	SNP + Indel	Allele frequency	A mixture model [100]
Varscan2	Somatic	SNP + Indel	Heuristic methods	Heuristic and statistical methods [64]

dict true indels [13, 56]. Due to these issues or constraints, paired-end read mapping and machine learning-based methods are not included in this study.

Besides these general indel calling programs, there are tools designed for detecting germline/somatic variants from cancer genomes. Almost all somatic indel calling programs can detect single nucleotide variants, some of them can also detect SVs [102]. Majority of these program use tumor-normal paired data to identify somatic variants, while others can predict with only tumor samples [103]. For programs based on the tumor-normal paired data, the general core algorithms include joint genotype analysis, allele frequency analysis, heuristic threshold, haplotype analysis, and machine learning [103]. In this study, we selected Varscan2, GATK Mutect2, Strelka and Strelka2 for comparative cancer indel analysis based on their good performances reported by several groups [102, 104, 105, 106, 107, 101] (Table 2.1). In general, performance evaluations for somatic indel identification can be done with simulation data and/or real sequence data [104, 106]. While the simulation data can help test different features such as variant allele fractions [106], comparison of indel annotation methods with real NGS data can provide useful guidance for their application in variant analysis in disease genomes. Even though currently there is no gold standard for evaluating somatic indel variants from cancer genomes, several existing databases can provide some useful information [104]. For instance, annotated indels in GATK Resource Bundle and dbSNP can be used to check false positive and indels in COS-

MIC can be used to evaluate positive cases, respectively [19, 28, 104, 108]. However, caution should be taken when using these databases for evaluation purpose as both databases contain only partial data.

Accurate annotation of indels is of paramount importance in studying genetic variations and in identifying disease associated indels [46, 109, 110]. To test the consistency or differences among the general indel calling programs, Hasan *et al.* performed a comparative analysis by using the sequences of chromosome 11 from 78 samples of the 1000 Genomes Project and showed that 78%-89% of the benchmark indels are not identified in a sample by any program and only a very small number of indels are identified by all seven programs [56]. However, the results do not accurately reflect the performance of each program as well as the common indels predicted from different programs. First, they compared the indels from individual genome samples to the pooled indel dataset of 79 genomes. Rare and low frequency variants account for a large proportion of indels and the pooled indel set includes all of them, but an individual sample may contain only a small subset of the pooled indel set [5, 32, 111]. Figure 2.1 shows a schematic example to explain the potential pitfalls of comparing individual samples with a pooled reference set from multiple samples. In this study we applied a pooled-sample based method for more accurate comparative analysis since indels from multiple samples from one program are pooled together to compare with the pooled benchmark indels (Figure 2.1b). In addition, we expanded the comparison with the whole genome sequences instead of only one chromosome.

Unlike SNPs, indels are more complicated in that there are two different indel types, insertion and deletion. Moreover, for a coding indel, it can be a frame-shift (FS) or non-frame-shift (NFS) indel. Consequently, the way to compare the indels can affect the number of true positives and false positives. Previous studies used a position range of $i \pm 5$ (where i is the indel position) to determine if an indel is the same one as that in the reference set [112]. However, this approach has several disadvantages. First,

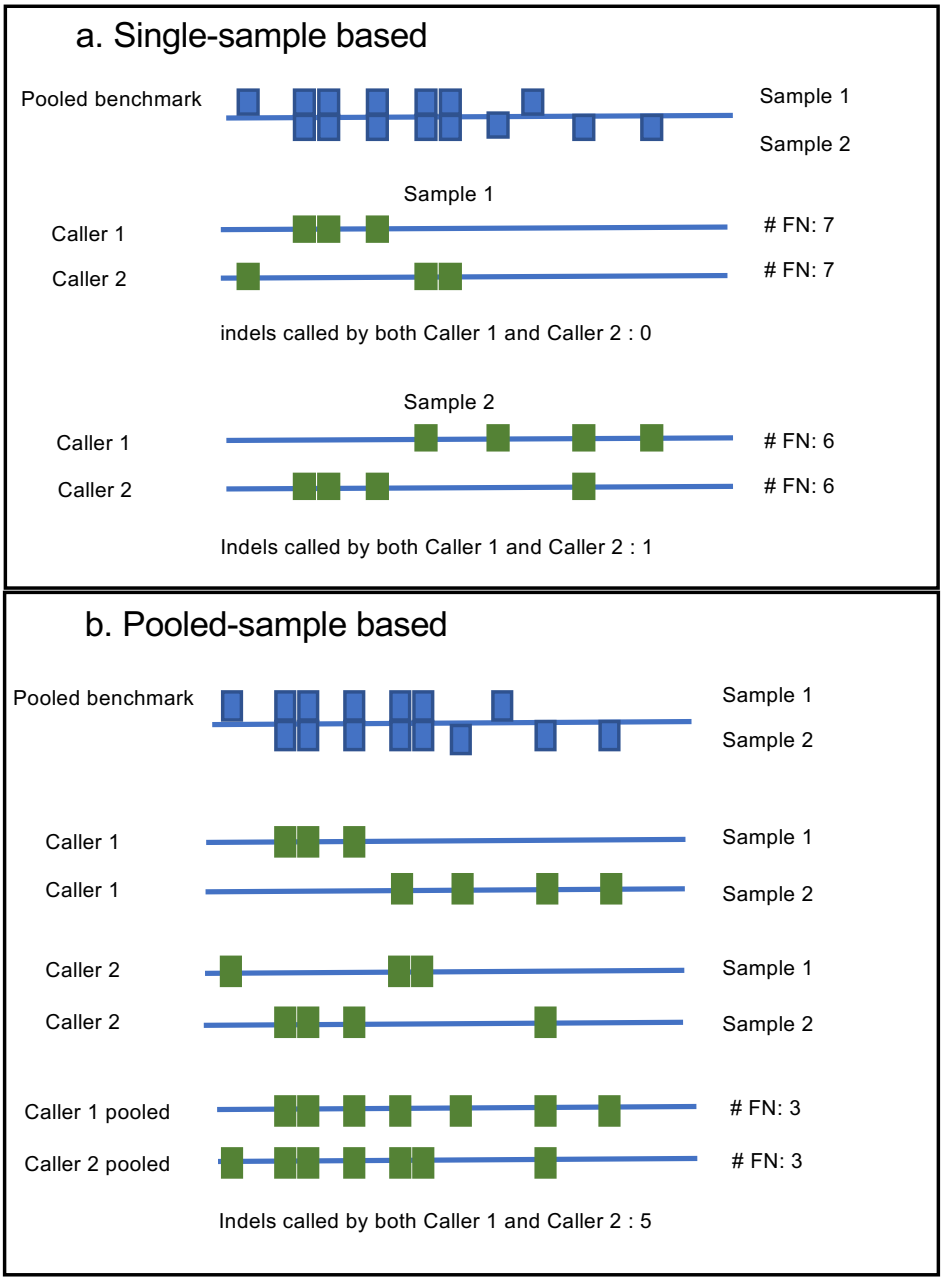

Figure 2.1: Comparison of different methods regarding false negative indels.

A schematic comparison between single-sample based method (a) and pooled-sample based method (b) with a pooled reference benchmark. FN: false negative

the indel types, insertion or deletion, are not considered separately. An insertion and a deletion at the same genome position are two different indels, not the same indel. Secondly, for coding indels, 1 bp difference in a position may result in a totally

different protein sequence due to an open reading frame shift. In light of these issues, we adopted a modified approach by considering indel types (insertion or deletion) as well as positions, which is especially important in germline indel analysis.

2.2 Methods

2.2.1 Datasets

We used the same dataset as Hasan *et al.*, which consists of 78 samples from the 1000 Genomes Project (http://www.internationalgenome.org/) covering five super populations (EUR, EAS, SAS, AMR, and ARF) and 26 sub-populations (three from each sub-populations) to evaluate general indel calling programs [56]. The benchmark is a set of about two million small indels identified by Mills *et al.* [8]. For somatic indel program evaluation, we used a total of 30 tumor-normal paired data, including 10 colon cancer, 10 breast cancer, and 10 bladder cancer samples. The cancer genome sequencing data were downloaded from TCGA with dbGap ID phs000178.v11.p8. A total of 4,970 indels from the latest version of COSMIC (v90) were downloaded for somatic indel evaluations [28].

2.2.2 Evaluation methods

For germline indels from healthy genomes, they are mainly genetic variants with the type and position of the indels presumably conserved in sub-populations or super populations. In other words, they are less random compared with somatic variants and usually do not lead to diseases. Therefore, when evaluating germline indels from healthy genomes, we only count the indels that are located at the same positions with the same insertion or deletion sequences between the samples and the reference as a positive identification. Since somatic indels from cancer genomes are less conserved than the germline indels, we use the typical range of $i \pm 5$ in positions along with the indel types, either insertion or deletion, for comparative evaluation.

Recall, precision and F measure are calculated for performance evaluations (Equa-

tions 2.1 2.2 2.3):

$$Recall = \frac{TP}{TP + FN} \tag{2.1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2.2}$$

$$F = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{2.3}$$

Where TP represents true positive, FP represents false positive, and FN represents false negative. As mentioned in the Background section, for germline indels, the TP, FP and FN are identified by a pooled sample-based method (Figure 2.1b). For somatic indel evaluation, the predicted indels are compared with the annotated indels in the COSMIC database as potential somatic indels (the indel types are classified using the indel labels downloaded from COSMIC). To identify potential false somatic indels, we compared the predictions with the indel set from the GATK Resource Bundle, which is considered as a standard germline indel set for human reference GRCh38 [28, 108].

2.3 Results

2.3.1 General indel calling programs

2.3.1.1 Overall analysis of the predicted indels

The number of true positive and false positive indels from healthy genomes by different programs is listed in Table 2.2. SAMTools calls the largest number of indels, with Platypus ranks the second. The number of the TP indels varies by programs. Dindel has the highest recall (0.78) but with a low precision (0.24). Varscan, which calls the least number of indels, has the highest precision (0.56) as well as the best F value (0.48). GATK_UG and GATK_HC have the second-best F value with relatively good recall and precision.

Table 2.2: Performance of different general indel annotation programs

Tool	TP indels	FP indels	Recall	Precision	F
Varscan	533,101	**424,740**	0.42	**0.56**	**0.48**
GATK_UG	884,763	1,802,477	0.69	0.33	0.45
GATK_HC	948,738	2,026,903	0.74	0.32	0.45
Pindel	446,622	619,846	0.35	0.42	0.38
Dindel	**994,947**	3,097,117	**0.78**	0.24	0.37
Platypus	941,046	3,403,565	0.74	0.22	0.33
SAMTools	930,860	15,083,658	0.73	0.06	0.11

Among all the programs, GATK_HC calls the longest indel with 616 bps. The length distribution is shown in Figure 2.2a (percentages) and Table 2.3 (counts) with the benchmark as a reference. SAMTools has the largest number of short indels for length between 1 bp and 20 bps, which is not surprising since it calls much more indels than any other programs (Table 2.2 and Table 2.3). Pindel has the largest number of indels longer than 50 bps (Supplementary Table 2.3 and Figure 2.2), largely because Pindel uses an algorithm that tends to call longer indels. In terms of mid-length indels between 20 and 50 bps, GATK_HC has the largest number in each category. Percentage-wise, Platypus, Varscan, GATK_UG, SAMTools predict relatively more short indels compared to other three programs. We also compared the programs in terms of indel types, insertion and deletion (Figure 2.2b and Table 2.4). SAMTools has a higher percentage of deletion types while GATK_UG has more insertion types in term of the ratio when compared with the benchmark. Dindel has the most similar insertion/deletion ratio (56.2%/43.8%) to the benchmark (57.6%/42.4%) and it has the highest TP rate for both insertion and deletion types (Table 2.4).

In coding regions, indels can be grouped into FS and NFS types. An NFS indel consists of a multiple of three base pairs, introducing an insertion/deletion of one or more amino acids while keeping the other part of the protein sequence unchanged. In contrast, an FS indel changes the open reading frame (ORF) starting from the site of insertion/deletion, which can produce different protein sequences from the indel

Table 2.3: Indel length (bps) distribution from different programs

Programs	**1-10**	**11-20**	**21-31**	**31-40**	**41-50**	**>50**
Varscan	973,285 (98.38%)	14,468 (1.48%)	1,338 (0.14%)	168 (0.02%)	52 (0.01%)	55 (0.01%)
GATK_UG	3,258,954 (97.01%)	85,365 (2.54%)	11,972 (0.36%)	2,145 (0.06%)	533 (0.02%)	604 (0.02%)
GATK_HC	3,577,313 (90.73%)	251,263 (6.37%)	**63,270 (1.61%)**	**23,372 (0.59%)**	**10,451 (0.27%)**	17,157 (0.44%)
Pindel	1,015,404 (89.54%)	56,060 (4.94%)	19,085 (1.68%)	9,412 (0.83%)	7,268 (0.64%)	**26,746 (2.36%)**
Dindel	4,975,421 (94.44%)	244,824 (4.65%)	36,877 (0.70%)	6,746 (0.13%)	1,839 (0.04%)	2,442 (0.05%)
Platypus	5,644,495 (96.62%)	168,210 (2.88%)	23,381 (0.40%)	4,153 (0.07%)	987 (0.02%)	886 (0.02%)
SAMTools	**19,903,777 (98.11%)**	**326,069 (1.61%)**	45,619 (0.23%)	7,891 (0.04%)	1,940 (0.01%)	1,389 (0.01%)
Benchmark	1,218,248 (95.94%)	33,112 (2.61%)	8,550 (0.67%)	3,817 (0.30%)	2,025 (0.16%)	4,002 (0.32%)

Table 2.4: Indel types (deletion and insertion)

Tool	**Deletion***	**TP Deletion#**	**Insertion***	**TP Insertion#**
Varscan	532,457 (55.59%)	302,541 (41.23%)	425,384 (44.41%)	230,560 (42.63%)
GATK_UG	1,330,229 (49.50%)	491,384 (66.97%)	1,357,011 (50.50%)	393,379 (72.74%)
GATK_HC	1,512,471 (50.83%)	544,906 (74.26%)	1,463,170 (49.17%)	403,832 (74.67%)
Pindel	594,809 (55.77%)	246,059 (33.53%)	471,659 (44.23%)	200,563 (37.08%)
Dindel	2,299,063 (56.18%)	**572,579 (78.03%)**	1,793,001 (43.82%)	**422,368 (78.10%)**
Platypus	2,351,612 (54.13%)	548,043 (74.69%)	1,992,999 (45.87%)	393,003 (72.67%)
SAMTools	**10,467,240 (65.36%)**	550,396 (75.01%)	5,547,278 (34.64%)	380,464 (70.35%)
Benchmark	733,758 (57.57%)	-	540,822 (42.43%)	-

*:Percentage= percentage of deletions/insertions of indels by each program
#:Percentage= percentage of deletions/insertions of benchmark by each program

Table 2.5: Coding indel types (NFS and FS)

Tools	FS*	TP FS#	NFS*	TP NFS#
Varscan	645 (57.95%)	401 (37.79%)	468 (42.05%)	263 (33.04%)
GATK_UG	1,460 (57.46%)	620 (58.44%)	1,081 (42.54%)	491 (61.68%)
GATK_HC	2,826 (60.57%)	703 (66.26%)	1,840 (39.43%)	550 (69.10%)
Pindel	1,099 (57.09%)	363 (34.21%)	826 (42.91%)	271 (34.05%)
Dindel	10,745 (83.11%)	743 (70.03%)	2,183 (16.89%)	579 (72.74%)
Platypus	4,550 (69.96%)	722 (68.05%)	1,954 (30.04%)	596 (74.87%)
SAMTools	**113,244 (94.88%)**	**790 (74.46%)**	**6,111 (5.12%)**	**610 (76.63%)**
Benchmark	1,061 (57.14%)	-	796 (42.86%)	-

*:Percentage = percentage of FS or NFS of indels by each program

#:Percentage = percentage of FS or NFS in benchmark by each program

position. FS indels can also lead to premature termination and the mRNA molecules can be subjected to a surveillance pathway called non-sense-mediated mRNA decay (NMD) [113]. The proportion of NFS and FS coding indels from each program is shown in Figure 2.2c and Table 2.5. GATK_UG, Pindel and Varscan show similar FS/NFS ratios to that of the benchmark while Pindel, SAMTools, and Platypus have a much higher percentage of FS coding indels.

2.3.1.2 Pare-wise comparisons

To check the similarity or difference of indels predicted by two different programs, the overlapped indels from two programs are compared with the benchmark indels. The recall and precision values are presented in Table 2.6, showing a trade-off between recall and precision. When a program is paired with Varscan or Pindel, it usually achieves high precision with smaller number of FPs while having low recall at the same time since these are the two programs that call the lowest number of total indels. The

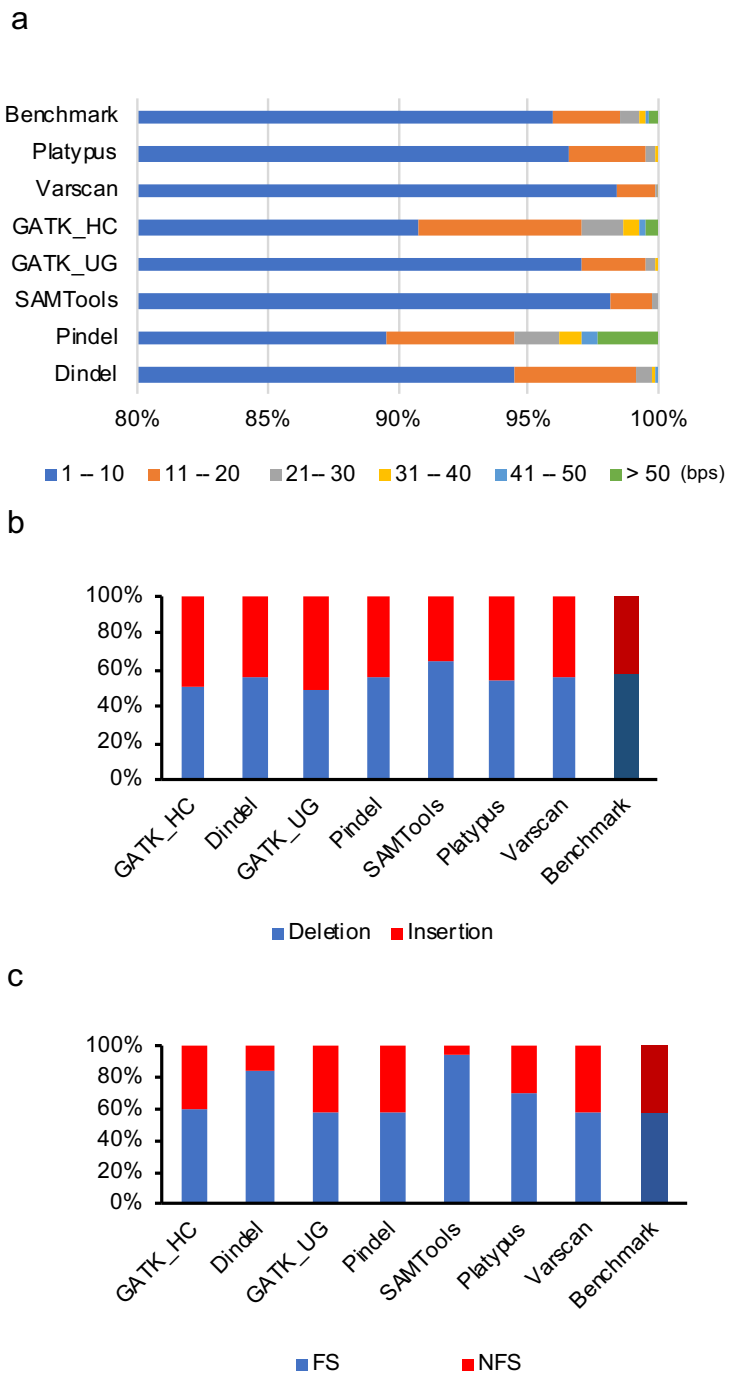

Figure 2.2: Comparisons of indels called by seven general indel calling tools.

(a) Indel size distribution. (b) Indel types distribution. (c) Coding indel types distribution. FS: frame shift, NFS: non-frame shift

indels from Varscan, GATK_UG and Dindel are highly similar. About 94% of indels from Varscan are also annotated by GATK_UG (898,482 out of 957,841) or Dindel (903,756 out of 957,841).

2.3.1.3 Combination of indels from different programs

The results from individual programs have shown that there are a large number of false positive indel predictions from the NGS data (Table 2.2). While false negatives may represent missed opportunities, false positives can result in wrong conclusions and are costly in real applications. We hypothesize that by selecting the consistent indel annotations from different programs, we may be able to remove majority of the false positives while retaining most of the true positives. The underlying idea is that in general, unlike false positives, true indels can be identified by different prediction algorithms. The ones that are program specific have a higher probability to be false positives. In a previous study, Hasan *et al.* showed that only a very small number of indels were called by all seven programs [56]. But as discussed in Background, that conclusion is a result from their approach by comparing the indels from individual samples to the pooled benchmark dataset, which may produces a large number of false negatives. We adopted a pooled sample method for a more meaningful comparison in this study. Figure 2.3 shows a schematic example to explain the differences by counting the overlaps or consistent indels between the two approaches. Among the seven indels called by both caller 1 and caller 2 with the pooled sample method, five of them are true positive indels. However, the single sample approach only identifies two true positives, resulting a very low TP rate from the overlapped indels (Figure 2.3).

Table 2.7 shows the averages of TP indels, FP indels, recall, precision and F values for all possible combinations including individual programs. The results from Hasan *et al.* show that only a small proportion of TP indels (1.51%) are called by all seven programs [56]. With our pooled sample approach, we found that 476,253 indels called

Table 2.6: Pair-wise comparison between general indel calling programs

recall \ precision	**Varscan**	**GATK_UG**	**GATK_HC**	**Pindel**	**Dindel**	**Platypus**	**SAMTools**
Varscan	-	0.40	0.41	0.30	0.41	0.39	0.36
GATK_UG	0.57	-	0.65	0.33	0.66	0.64	0.60
GATK_HC	0.59	0.43	-	0.34	0.72	0.68	0.65
Pindel	0.62	0.59	0.57	-	0.34	0.33	0.31
Dindel	0.58	0.41	0.38	0.56	-	0.70	0.68
Platypus	0.57	0.34	0.38	0.59	0.35	-	0.66
SAMTools	0.55	0.40	0.40	0.60	0.27	0.29	-

Table 2.7: Performance comparison of different program combinations (showing average values)

# of Tools	**TP indels**	**FP indels**	**Recall**	**Precision**	**F**
1	811,440	3,779,758	0.64	0.31	0.37
2	639,772	899,660	0.51	0.48	0.45
3	528,467	496,588	0.41	0.56	0.45
4	450,280	322,289	0.37	0.60	0.44
5	394,064	230,561	0.31	0.64	0.41
6	354,111	179,699	0.28	0.67	0.38
7	326,184	150,069	0.26	0.68	0.37

by all seven programs. Among these indels, 326,184 are TP indels that can be found in the reference set (25.6%).

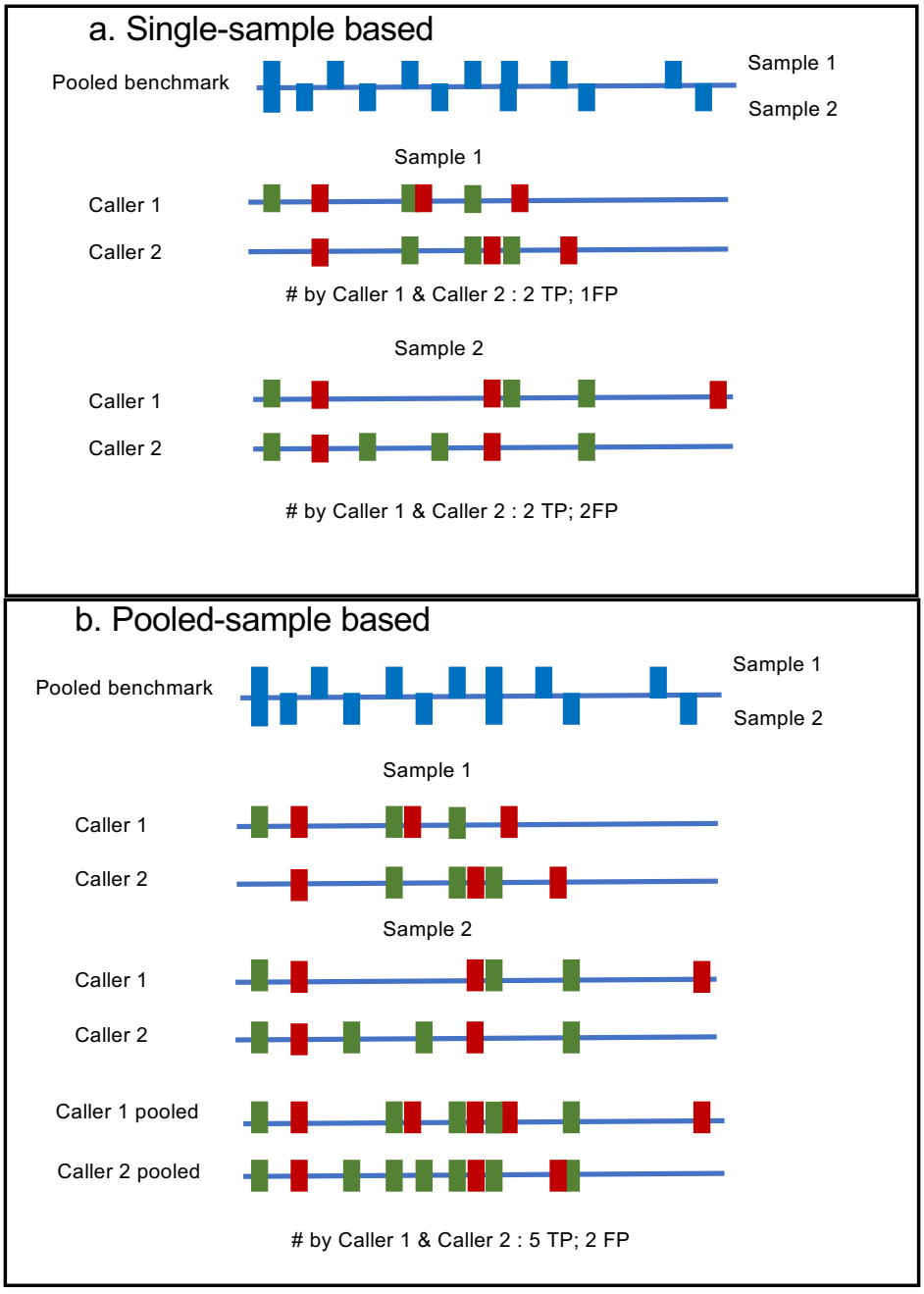

Figure 2.3: Comparison of different methods regarding true indels.

A schematic comparison between single-sample based method (a) and pooled-sample based method (b) with a pooled reference benchmark. Green represents true positives. Red represents false positive predictions. Blue blocks are the benchmark indels.

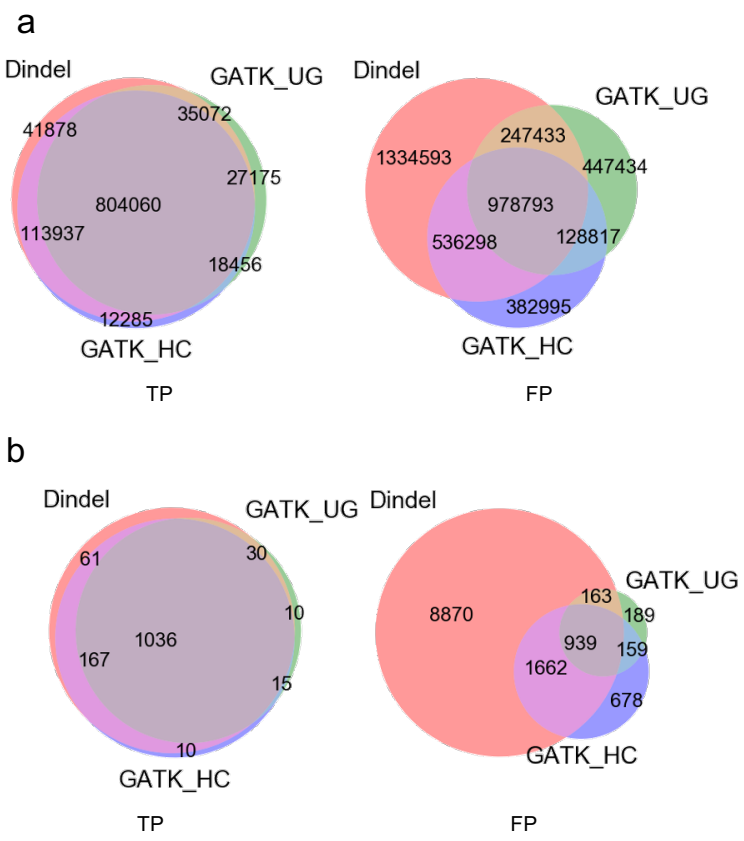

Figure 2.4: Overlapped indels called by GATK_UG, GATK_HC and Dindel. (a) All indels; (b) Coding indels only.

Among all the possible combinations, including the individual programs, a five tool combination of GATK_UG, GATK_HC, Pindel, SAMTools and Dindel has the highest F value (0.53). Dindel has the highest recall (0.78, Table 2.2) and a combination of three tools (GATK_UG, Pindel and SAMTools) has the highest precision (0.69). On average, a combination of 2 or 3 programs has the highest average F values (Table 2.7). Table 2.8 lists top three combinations of two and three programs ranked by F values. As shown in Tables 2.7 and 2.8, adding more programs can remove more false positives than true positives and a combination of three programs seems to have a good balance of recall and precision. Figure 2.4a shows an example of indels called by 3 programs: GATK_UG, GATK_HC and Dindel. There are large

Table 2.8: Top 3 indel annotation program combinations (2 programs and 3 programs)

F rank	Combination of 2 tools	TP	FP	Recall	Precision	F
1	GATK_UG + GATK_HC	822,516	1,107,610	0.65	0.43	0.51
2	GATK_UG + Dindel	839,132	1,226,226	0.66	0.41	0.50
3	GATK_HC + Platypus	871,596	1,403,334	0.68	0.38	0.49
F rank	**Combination of 3 tools**	**TP**	**FP**	**Recall**	**Precision**	**F**
1	GATK_UG + GATK_HC + Dindel	804,060	978,793	0.63	0.45	0.53
2	GATK_UG + GATK_HC + SAMTools	725,419	768,246	0.57	0.49	0.52
3	GATK_UG + GATK_HC + Platypus	778,439	991,540	0.61	0.44	0.51

overlaps among the TP indels either for all indels (Figure 2.4a) or for coding indels only (Figure 2.4b), while the disagreement among the FP indels are much bigger. Therefore, if a low number of false positives is preferred in an application, results from more programs can be used and combined.

2.3.2 Somatic indel calling programs

Unlike general indel calling approaches, the majority of somatic indel annotations need both normal and diseases genome samples and thus are more complicated. Different methods or algorithms have been developed for somatic indel identifications (Table 2.1). In this study, we applied four somatic indel calling programs to three types of cancers. As discussed in Background, there are no benchmark sets available to assess the true positive or false positives for cancer genome indels. But for comparison purposes between programs and cancer types, we can use the COSMIC database with annotated somatic cancer indels and GATK Resource Bundle as potential false positives (or germline indels) to see how much they agree or differ with each other. Since the COSMIC indel set represents only a small portion of real cancer population indels, a small number of indels in COSMIC does not necessarily indicate a large number of false positives from a program. Similarly, an indel found in the germline indel set does not necessarily mean it is a true false positive since there is a single

Table 2.9: Performance comparison of different somatic indel annotation programs

Tools	Total indels	Cancer type	COSMIC indels	Potential germline indels and rate
Strelka	2,186	Bladder	5	884 (0.40)
	5,521	Breast	5	2,536 (0.46)
	14,174	Colon	11	5,227 (0.37)
Strelka2	867	Bladder	0	225 (0.26)
	2,162	Breast	0	768 (0.36)
	9,920	Colon	2	3,583 (0.36)
Varscan2	1,804	Bladder	2	438 (0.24)
	3,796	Breast	4	879 (0.23)
	6,286	Colon	8	831 (0.13)
Mutect2	19,124	Bladder	10	761 (0.04)
	44,373	Breast	16	1,708 (0.04)
	30,503	Colon	31	4,971 (0.16)

cancer sample vs. pooled germline samples problem. Nevertheless, the comparative analysis can provide some insights about these somatic indel calling programs and the similarity or differences among different cancer types.

The number of potential true positive and false positive indels called by four programs are shown in Table 2.9. GATK Mutect2 calls the largest number of indels independent of cancer types and it has the largest overlap with the COSMIC indels and relatively low number of potential germline indels among the four programs. Strelka2 has the smallest numbers of indels for bladder and breast cancer types while Varscan2 calls the lowest number of indels in colon cancer. In terms of cancer types, colon cancer has more indels than the other two cancer types. The number of indels in bladder cancer is much less than the other two types. Taken together, GATK Mutect2 has a better coverage of somatic indels in all three cancer types with relatively low number of germline indels, or potential false positives. Strelka has the second largest number of total indels and COSMIC indels, however, the number of potential germline indels is also high.

As for the length distribution of the somatic indels, GATK Mutect2 calls the longest somatic indel (245 bps) in a cancer genome and identifies more longer indels (Figure

2.5a and Table 2.10). It has 202 indels longer than 50 bps. However, no other programs identify any indels of length 50 or more. The length distributions in terms of cancer types also vary. Even though colon cancer has the largest number of indels, breast cancer has more longer indels (Figure 2.5b and Table 2.10).

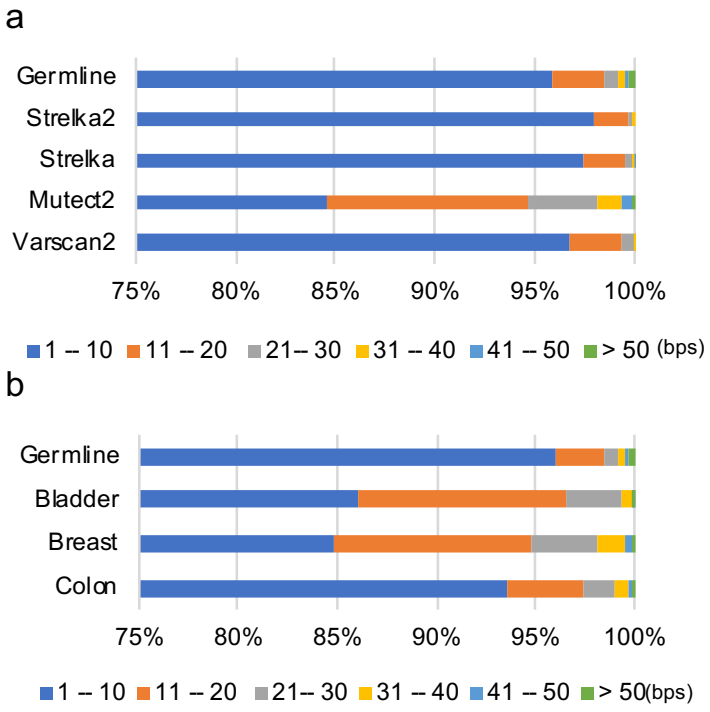

Figure 2.5: Somatic indel size distribution.
(a) Program based; and (b) Cancer type based.

In healthy genomes, there are more germline deletions (57.75%) than germline insertions (42.25%) (Table 2.11) while in cancer indel database COSMIC, the ratios are slightly different with 34.39% of insertions and 52.00% of deletions, while others are assigned as complex indels (13.61%) (Table 2.11) [28]. Except for GATK Mutect2 in bladder and breast cancer genomes, all other programs detect relatively low number of insertions. It is not clear if cancer genomes have relatively fewer insertions or the programs have difficulty in identifying somatic insertions. As for coding indels,

Table 2.10: Somatic indel length distribution

Tools / Cancer types	**1-10**	**11-20**	**21-31**	**31-40**	**41-50**	**>50**
Strelka2	12,699 (97.93%)	219 (1.69%)	43 (0.33%)	6 (0.05%)	0 (0.00%)	0 (0.00%)
Strelka	21,370 (97.41%)	471 (2.15%)	84 (0.38%)	12 (0.05%)	2 (0.01%)	0 (0.00%)
Mutect2	80,177 (84.54%)	9,514 (10.03%)	3,295 (3.47%)	1,280 (1.35%)	368 (0.39%)	202 (0.21%)
Varscan2	11,505 (96.79%)	313 (2.63%)	63 (0.53%)	5 (0.04%)	0 (0.00%)	0 (0.00%)
Bladder	20,629 (86.00%)	2,527 (10.53%)	695 (2.90%)	91 (0.38%)	15 (0.06%)	31 (0.13%)
Breast	48,136 (84.83%)	5,692 (10.03%)	1,823 (3.21%)	801 (1.41%)	206 (0.36%)	89 (0.16%)
Colon	56,986 (93.58%)	2,298 (3.77%)	967 (1.59%)	411 (0.67%)	149 (0.24%)	82 (0.13%)
Germline	1,215,718 (95.95%)	33,004 (2.60%)	8,505 (0.67%)	3,795 (0.30%)	2,012 (0.16%)	3,974 (0.31%)

germline coding indels has slightly more NFS indels (51.63%) than the FS indels (48.37%) (Table 2.12). It is not surprising that the number of FS coding indels is smaller than expected (2 to 1 ratio if there is no selection) in healthy genomes, as FS indels are more deleterious than NFS indels, which are more likely to be removed from the population during evolution. FS indels found in healthy individuals generally are less deleterious and their major role is to contribute to population diversity [32]. In COSMIC cancer indel database, FS indel is the dominant coding indel type (81.05%). Except for Strelka2 in balder cancer, all other programs predict more FS indels than NFS indels in all three cancer types. It should be pointed out that the total numbers of coding indels predicted by Varscan2 and Strelka2 are rather small (Table 2.12).

When the somatic indels called from different programs are compared, the number of similar indels from different programs or the overlapped indels are much smaller especially when more programs are considered (Figure 2.6 and Table 2.13). This is quite different from the germline indels by the general indel annotation programs

Table 2.11: Somatic indel types (deletion and insertion)

Cancer Types	Tools	# of Deletions	Deletion Percentage	# of Insertions	Insertion Percentage
Bladder	Strelka	1,417	64.73%	772	35.27%
	Strelka2	614	70.49%	257	29.51%
	Varscan2	1,128	62.53%	676	37.47%
	Mutect2	7,784	40.03%	11,662	59.97%
Breast	Strelka	3,186	57.57%	2,348	42.43%
	Strelka2	1,413	65.24%	753	34.76%
	Varscan2	2,227	58.67%	1,569	41.33%
	Mutect2	16,273	35.40%	29,701	64.60%
Colon	Strelka	9,280	65.28%	4,936	34.72%
	Strelka2	6,973	70.22%	2,957	29.78%
	Varscan2	5,058	80.46%	1,228	19.54%
	Mutect2	20,022	63.74%	11,390	36.26%
Germline indels		731,665	57.75%	535,343	42.25%
COSMIC indels		3,104	52.00%	1,709	34.39%

Table 2.12: Somatic coding indel types (FS and NFS)

Cancer Types	Tools	# of FS indels	FS Percentage	# of NFS indels	NFS Percentage
Bladder	Strelka	135	53.78%	116	46.22%
	Strelka2	6	20.69%	23	79.31%
	Varscan2	66	63.46%	38	36.54%
	Mutect2	4,292	63.25%	2,494	36.75%
Breast	Strelka	142	62.28%	86	37.72%
	Strelka2	4	80.00%	1	20.00%
	Varscan2	60	71.11%	30	28.89%
	Mutect2	7,651	71.11%	3,109	28.89%
Colon	Strelka	303	78.29%	84	21.71%
	Strelka2	37	88.10%	5	11.90%
	Varscan2	202	82.79%	42	17.21%
	Mutect2	3,218	65.17%	1,720	34.83%
Germline indels		697	48.37%	744	51.63%
COSMIC indels		2,515	81.05%	588	18.95%

Table 2.13: Performance on different number of somatic program combinations (The data shown are average values)

Cancer types	# of Tools	Total indels	COSMIC indels	Potential germline indels and rate
Bladder	1	5,995	4	577 (0.24)
	2	285	1	64 (0.22)
	3	92	1	20 (0.24)
	4	22	0	6 (0.27)
Breast	1	13,963	6	1,463 (0.27)
	2	616	1	185 (0.25)
	3	181	1	47 (0.19)
	4	36	0	5 (0.14)
Colon	1	15,221	13	6,666 (0.26)
	2	3,142	3	948 (0.23)
	3	1,051	2	300 (0.18)
	4	161	1	14 (0.09)

especially the comparison criteria are not as stringent as those used for germline indel comparisons (Table 2.7), in which there are a large number of indels called by all the programs, especially for the true positive indels. Table 2.13 and Figure 2.6 show that when all four programs are used, there are only 22, 36, 161 indels in the bladder, breast, and colon cancel samples respectively. These results suggest that the agreement among different programs is low and it might not be practical to use multiple programs in order to remove false positives in cancer samples as we showed in the germline indel cases since it also dramatically decrease the total number of indels as well as true positives.

2.4 Discussions

Accurate annotation of indels in both healthy and cancer genomes is important for downstream analysis in biological and medical applications. A number of programs have been developed for identifying indels from both healthy genomes for germline indels as well as cancer genomes for somatic indels with NGS data. Comparative analysis and evaluation can provide useful information about each program's performance. The best available benchmark for large-scale germline indels so far is the

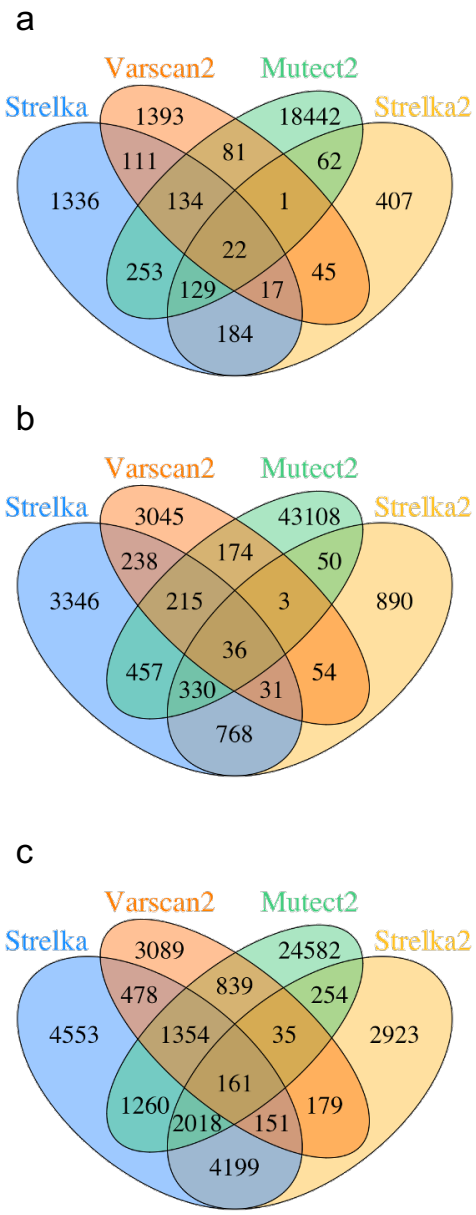

Figure 2.6: Overlapped indel annotations of different cancer types
(a) Bladder cancer; (b) Breast cancer; and (c) Colon cancer

pooled sample indels [8]. One previous comparative study applied this pooled benchmark set and evaluated seven general indel calling programs using chromosome 11 of 78 samples. However, the comparison was carried out between a single sample and the pooled benchmark, which is problematic as shown in Figures 2.1 and 2.3. It may also explain why the study finds little overlap of indels when the results from all seven

tools are combined [56]. In this study we carried out an improved approach to assess the general indel calling programs using the whole genome NGS sequences instead of using one chromosome sequences. More importantly, we adopted a pooled sample vs. a pooled benchmark comparison, which provides more accurate assessment of programs' performances. The new method greatly reduced the number of false negative cases by correctly recognizing the true positives (Figures 2.1 and 2.3). Last but not the least, we adopted a stringent indel comparison approach by considering the exact indel position as well as the indel types, which was not considered in previous studies. It should be noted that even though we applied a pooled sample approach, the comparison is not error free since the samples and the genomes in the benchmark set are different. There are some sample specific indels in both the test set and the benchmark set. Nevertheless, our approach makes the best use of the reference set and provides a more accurate performance evaluations.

These general indel calling programs employ different prediction algorithms and predict different number of indels with different length and type distributions (Tables 2.3, 2.4). There is a tradeoff between the number of true positives and false positives. Some of them recognize a large number of true positive indels but at the same time output more false positive indels. We found that combing indels predicted from several different programs can achieve a good balance of TPs and FPs by removing a large number of false positives while keeping most of the true positives. The idea behind this is that if an indel is a true one, most programs are expected to find it no matter what algorithm is used. On the other hand, if an indel is a false one, it probably will only be predicted by one or a small number of programs. Our results show it is indeed the case and the best TP/FP balance is achieved with two or three different programs (Tables 2.7 and 2.8).

In addition to germline indels, we also carried out program comparisons of somatic indel predictions using 30 cancer samples of three different types. Evaluating

somatic indels is even more challenging because there is no benchmark that can be used for a systematic comparison and cancer indels are more random in terms of indel positions. Nevertheless, by using a common sample sets, we can evaluate the similarity/differences of indels from different somatic indel calling programs and among different cancer types. To get a sense of the potential number of true positive or false positive somatic indels, we compared the predicted indels with the cancer indels in COSMIC database (as potential true positives) and the germline indel set (as potential false positive somatic indels). While each program produced different number of indels with various ratios of indel types (Table 2.9, 2.10 and 2.11), there is a clear trend among different cancer types in general. Bladder cancer has the lowest number of predicted somatic indels and colon cancer has the largest number of predicted somatic indels (Tables 2.9 and 2.11). Secondly, unlike the germline indels, the number of indels predicted by all programs is very small (Table 2.13), suggesting a low agreement among the programs even though the input sequences are the same. Thirdly, the programs identify a small number of insertions. This trend has also been reported by other case studies. For example, 2,233 deletions and 544 insertions were identified from 21 breast cancer genomes by a modified Pindel program, and 680 deletions and 303 insertions were found from a skin cancer genome by Pindel, BWA and GROUPER [114, 115]. In COSMIC database, there are also less insertions compared with deletions (Table 2.11). On the other hand, Sathya *et al.* identified SNP and indel patterns from lung cancer genomes and found more insertions than deletions in both healthy genomes and lung cancer genomes using GATK_UG [116]. Whether the difference in the ratio of insertion and deletion in the cancer genome is caused by the characteristics of the cancer genome or by the algorithms used by the somatic variants calling programs remains to be further studied.

CHAPTER 3: Comparative Somatic Indel Analysis in Cancer Genomes

3.1 Background

Several studies have been carried out to investigate indels in different cancer types. A recent pan-cancer analysis indicated that renal cell carcinoma has the highest number of indels and the highest proportion of indels among 19 cancer types [117]. The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium reported that colorectal adenocarcinoma has a larger number of somatic indels among diverse cancer types [118]. Similar to indel annotation in heathy genomes, different methods and algorithms may lead to different somatic indel annotations. In Chapter 2, we performed a comparative evaluation analysis on four somatic indel calling programs (Strelka, Strelka2, Varscan2 and GATK Mutect2) and three cancer types (colon cancer, breast cancer, and bladder cancer) [64, 98, 99, 100]. We found that the performances from these calling programs vary greatly.

Investigations of somatic coding indels in cancers have been carried out at both domain and protein level. Pagel *et al.* mapped somatic non-frame shift (NFS) indels from COSMIC onto protein structures and found pathogenic variants tend to be enriched in helix and strand regions [46]. Niu *et al.* developed a tool to identify 3D variants clusters on protein structures that can be used in variant-drug interaction analysis in cancer genomes [95]. They identified mutation-mutation and mutation-drug clusters from more than 4,400 samples across 19 cancer types. More than 6,000 clusters were identified at 3D structure level, including both intra-molecular and inter-molecular clusters. Among the 553,496 somatic variants, only 0.76% of them are indels [95]. Yang *et al.* identified about 100 significantly mutated protein domains from 7,260 samples across 21 cancer types. However, all of the 237,716 somatic variants

are SNPs [119]. Similar domain-centric study performed by Peterson *et al.* also only focused on SNPs [120].

Variants in non-coding regions can also cause diseases [121, 122, 123, 124, 125, 126]. Sakthikumar *et al.* investigated non-coding variants in Glioblastoma (GBM) genomes and found that the GBM somatic variants are enriched in non-coding regions of 78 GBM key genes, suggesting the essential role of non-coding somatic variants in cancer [127]. Imielinski *et al.* studied somatic non-coding indels in 79 lung cancer genomes and found that these indels are enriched in surfactant protein genes [48]. However, Nakagomi *et al.* analyzed 113 lung cancer genomes and reported that only 29 of them have non-coding indels in surfactant protein genes [49].

Currently, majority of the small variants analyses in cancer genomes focused on SNPs. Although some of them also include indels, the number of the indels is rather small [95, 119, 120]. Here we focused on both coding and non-coding somatic indels in cancer genomes. The goal of this study is to compare features of indels in different cancer types as well as indels called by different programs. As discussed earlier, indels predicted from different programs vary due to different prediction algorithms. In this study, we applied Strelka and Varscan2 to call somatic indels as previous studies demonstrated that these two programs perform well for somatic variants calling [99, 64, 103, 104]. We selected two cancer types, invasive breast carcinoma (BRCA) and lung adenocarcinoma (LUAD). Several cancer-related studies showed that BRCA has the largest sample size and is one of the most analyzed cancer types, and the proportion of indels in breast cancer ranked second among 19 cancer types [117, 120]. A recent report showed that LUAD has high numbers of exonic somatic variants in several studies [119, 120].

We identified a total of 184,106 somatic indels. Only 9% of these indels are in both BRCA and LUAD cancer types. The somatic indels in BRCA and LUAD have different proportion of indel types, including deletion/insertion, coding/non-

coding and FS/NFS. The somatic coding indels are more likely to be enriched in the important regions of the proteins than those in the healthy genomes. About 30% of the somatic indels in SMGs' non-coding regions overlap with annotated TFBSs.

3.2 Methods

3.2.1 NGS data source and indel calling

We downloaded 436 BRCA and 564 LUAD whole genome sequencing data of TCGA-BRCA and TCGA-LUAD projects from TCGA data portal with dbGaP Study Accession: phs000178.v11.p8 [128, 129]. Both tumor and normal blood/tissue sequencing data were used to call somatic indels. Strelka was used to call somatic indels from all 1,000 samples. For program comparison, Varscan2 was employed to call somatic indels from 564 LUAD genomes [99, 64]. The human genome reference GRCh38.p13 was used in variants calling. The indel set from the GATK Resource bundle with 1,267,008 germline indels was used as the reference of germline indel annotations in healthy human genomes [108]. In previous studies, a position $i \pm 5$ has been used to determine whether two indels are the same, without concerning the indel types (insertion or deletion) [56]. In this study, to distinguish germline indels and somatic indels, we used a more stringent approach by considering the indel types and insertion/deletion sequences in addition to the indel positions. Two indels are considered the same only if both have the same position, indel type and sequences. For cancer type wise and program wise comparison, we used the position $i \pm 5$ and also the indel type to define same indels.

A total of 127 significantly mutated genes (SMGs) across 12 major cancer types identified by Kandoth *et al.* were used to map both coding and non-coding indels [130]. Since these 127 SMGs contain 125 protein coding genes, 1 lncRNA gene and 1 miRNA gene, only 125 protein coding genes are used in this study.

3.2.2 Protein structural analysis

To locate the positions of coding somatic indels, we downloaded all protein coding gene annotations from Ensembl [131]. If a gene has multiple transcripts, the longest transcript is used for mapping. Each transcript was searched against PDB using Blast for known protein structures [132, 133]. If a protein has known structure(s) in PDB, we assigned secondary structure types of the proteins using DSSP, as described in our previous studies. H (α-helix), G (3_{10}-helix) and I (π-helix) states are grouped as helix conformation; E (extended strand) and B (residue in isolated β-bridge) states are grouped as strand conformation and all the remaining states are loop conformation [134]. For proteins without known structures, we searched for highly homologous proteins in PDB with at least 50% coverage and 80% sequence identity. Then the secondary structure types of the template protein were copied to the target protein. If no homologous structures were found in PDB, RaptorX, a highly accurate secondary structure prediction program, was applied to predict secondary structure types with default settings. RaptorX uses conditional neural fields method to predict secondary structure types and achieves 84% accuracy [135]. The structural analysis of the indels from healthy genomes were performed based on 769,743 short coding indels annotated by GATK Resource bundle [108].

3.2.3 Non-coding indel analysis

To study the overlap between indels and TFBS, we used the TFBS set predicted by dePCRM2, a recently developed program for genome scale TFBS prediction with a sensitivity of more than 97% [136]. Using dePCRM2, a total of 25,297,119 non-overlapping TFBSs were predicted with a p-value cutoff at $5\text{x}10^{-6}$. The non-coding regions of the SMGs are defined as UTR regions, introns and $\pm$100 kb intergenic flanking regions, and the coordinates are downloaded from Ensembl [127, 131].

3.3 Results

3.3.1 Comparison of somatic indels between cancer types and programs

Using Strelka, there are 109,856 and 91,159 somatic indels identified from 436 BRCA samples and 564 LUAD samples, respectively. The number of common indels between BRCA and LUAD is 16,909 (Table 3.1). Varscan2 identified 55,345 somatic indels from LUAD samples, much less than the number from Strelka, and 25,278 indels are called by both programs (Table 3.2). As shown in Chapter 2, due to the prediction algorithms, the calling results from these programs may contain germline indels. To filter out these indels from downstream analyses, we compared these predicted somatic indels with the germline indel set from GATK Resource bundle. We found 16.74% and 19.64% of indels in BRCA and LUAD respectively are the same as the germline indels from healthy genomes (Table 3.1). Only 8.14% of the indels from Varscan2 are germline indels (Table 3.2). All the downstream somatic analyses are based on the indels after removing the germline ones from each case.

Similar to the coding germline indels in healthy genome, there are more deletions than insertions in both BRCA and LUAD. The percentages of deletions in both cancer types are slightly higher than those of GATK germline indel set. The common indels in BRCA and LUAD have a different trend with a slightly higher insertion rate (Table 3.1). As for different programs, Varscan2 called more deletions than Strelka (64.19% vs. 61.80%), and the indels identified by both programs also have a higher deletion rate (66.00%) than that for the germline indels (57.75%).

3.3.2 Somatic coding indels in BRCA and LUAD genomes

Compared with the germline indels from healthy genomes (0.11%), the proportion of coding somatic indels is very high in BRCA (5.94%) and much higher in LUAD (10.79%) with Strelka (Table 3.3). Varscan2 also showed 8.96% of the indels in LUAD are coding indels and the common indels between the two programs have even higher

Table 3.1: Somatic indels in BRCA and LUAD with Strelka

Cancer types	Total indels	Overlap with germline indels	Deletions	Insertions
BRCA	109,856	18,392 (16.74%)	54,034 (59.08%)	37,430 (40.92%)
LUAD	91,159	17,900 (19.64%)	45,273 (61.80%)	27,986 (38.20%)
BRCA ∩ LUAD	16,909	4,085 (24.16%)	6,972 (54.37%)	5,852 (45.63%)
Germline	1,267,008	-	731,665 (57.75%)	535,343 (42.25%)

Table 3.2: Somatic indels in LUAD using different programs

Programs	Total indels	Overlap with germline indels	Deletions	Insertions
Strelka	91,159	17,900 (19.64%)	45,273 (61.80%)	27,986 (38.20%)
Varscan2	55,245	4,499 (8.14%)	32,574 (64.19%)	18,172 (35.81%)
Strelka ∩ Varscan2	25,278	1,890 (7.48%)	15,435 (66.00%)	7,953 (34.00%)
Germline	1,267,008	-	731,665 (57.75%)	535,343 (42.25%)

coding rate (14.73%) (Table 3.4). In terms of the deletion/insertion ratio in coding regions, LUAD has more deletion types (around 70%) no matter which program is used for indel identification when compared with that in the germline indels from healthy genomes (Tables 3.3 and 3.4). Around 57.95% of somatic coding indels from BRCA samples are deletions. The common coding indels between BRCA and LUAD have more insertions than deletions (54.08% vs 45.92%) (Table 3.3).

Coding indels can also be divided into FS and NFS indels according to the indel length. If the length of an indel is a multiple of three (NFS), then it only changes the amino acid(s) of the mutated positions, while keeping other parts of the sequence unchanged. Indels with other lengths change the open reading frame and cause a frame shift at the indel site (FS), which are prone to be more deleterious [43, 32, 137]. In theory, the number of FS indels is expected to be twice of NFS indels. Our analysis of the healthy genomes from the 1000 Genomes Project showed that the number of germline FS indels (3,775) is similar to NFS indels (3,662) [32]. The indels from GATK Resource bundle are also similar (697 FS vs. 744 NFS) (Tables 3.3 and 3.4). These results indicate that healthy genomes tend to have less deleterious coding indels. However, for somatic indels in cancer genomes, the number of FS indels is three to four times more than that of NFS indels, especially for the LUAD cancer type, over 80% of the indels are FS indels (Tables 3.3 and 3.4). The coding somatic indels that exist in both BRCA and LUAD genomes have relatively lower ratio of FS indels (60.67%).

We found that 7,756 genes have somatic indels in the CDS regions. In BRCA genomes, 3,979 genes have somatic coding indels predicted by Strelka. For LUAD genomes, Strelka and Varscan2 identified 5,513 and 3,438 genes with coding somatic indels respectively. There are 5,951 genes with somatic coding indels in LUAD after pooling the indels together from the two programs. Between BRCA and LUAD, 2,174 common genes have somatic indels in the CDS regions. Among all these genes,

Table 3.3: Somatic coding indel types in BRCA and LUAD

Cancer types	Total indels	Deletions	Insertions	FS
BRCA	5,432 (5.94%)	3,148 (57.95%)	2,284 (42.05%)	4,049 (74.54%)
LUAD	7,904 (10.79%)	5,584 (70.65%)	2,320 (29.35%)	6,467 (81.82%)
BRCA ∩ LUAD	956 (7.45%)	439 (45.92%)	517 (54.08%)	580 (60.67%)
Germline	1,441 (0.11%)	896 (62.18%)	545 (37.82%)	697 (48.37%)

Table 3.4: Somatic coding indel types in LUAD with different programs

Programs	Total indels	Deletions	Insertions	FS
Strelka	7,904 (10.79%)	5,584 (70.65%)	2,320 (29.35%)	6,467 (81.82%)
Varscan2	4,548 (8.96%)	3,210 (70.58%)	1,338 (29.42%)	3,749 (82.43%)
Strelka ∩ Varscan2	3,445 (14.73%)	2,551 (74.05%)	894 (25.95%)	2,964 (86.04%)
Germline	1,441 (0.11%)	896 (62.18%)	545 (37.82%)	697 (48.37%)

MAP3K1 has the most somatic coding indels in BRCA (45 by Strelka), and TP53 has the most somatic coding indels in LUAD (37 by Strelka and 31 by Varscan2). Both genes are in the list of 125 protein coding SMGs [130].

Previously, we found that both FS and NFS germline coding indels in healthy human genomes tend to locate in terminal regions of the transcripts [32]. There are more NFS somatic coding indels at the terminal regions in both BRCA and LUAD, while FS somatic coding indels are nearly evenly distributed, except for a small peak at around 70% transcript position in BRCA (Figure 3.1).

Since NFS somatic coding indels only affect a part of the protein while keeping the remaining sequence unchanged, we compared the distribution of secondary structure types of these indels with those in the healthy genomes. A total of 181 proteins with NFS somatic mutations in BRCA were found to have known or homologous structures in PDB. For LUAD, the number of such protein is 106 (Strelka) and 51

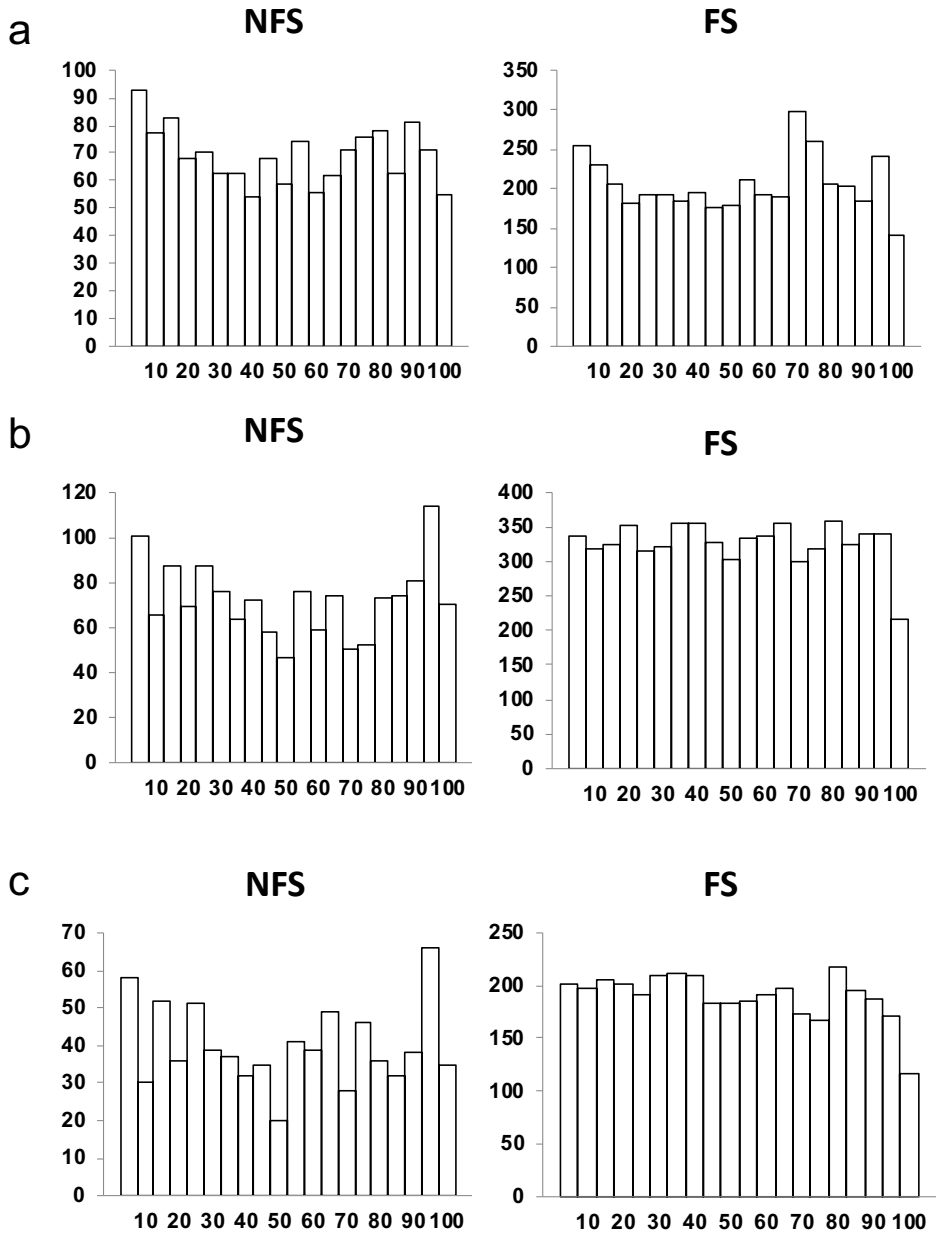

Figure 3.1: Positions of coding indels on proteins.

(a) Somatic coding indels from BRCA by Strelka. (b) Somatic coding indels from LUAD by Strelka. (c) Somatic coding indels from LUAD by Varscan2.

(Varscan2). For the proteins without known structures, we used RaptorX to predict the secondary structure types for each amino aide, as described in Methods. The distributions between two cancer types are different with a p-value of 0.003.

When compared with the general protein secondary structure type distribution (background), we found that somatic coding NFS indels in cancer genomes have more loop structures (57.38%, 59.24% and 59.28% for BRCA/Strelka, LUAD/Strelka and LUAD/Varscan2, respectively) with low strand types (10.59%, 9.96% and 10.38% for BRCA/Strelka, LUAD/Strelka and LUAD/Varscan2, respectively) [138]. When compared with the distribution of secondary structure types of NSF germline indels from healthy genomes, these somatic indels have more in helix and strand, fewer in the loop types, as shown in Figure 3.2. These distributions are significantly different with p-values of chi-square tests less than $2.2\text{x}10^{-16}$. The result indicated that the NFS somatic indels in BRCA and LUAD genomes tend to be in core secondary structure types, helix and strand, when compared with those of germline NFS indels.

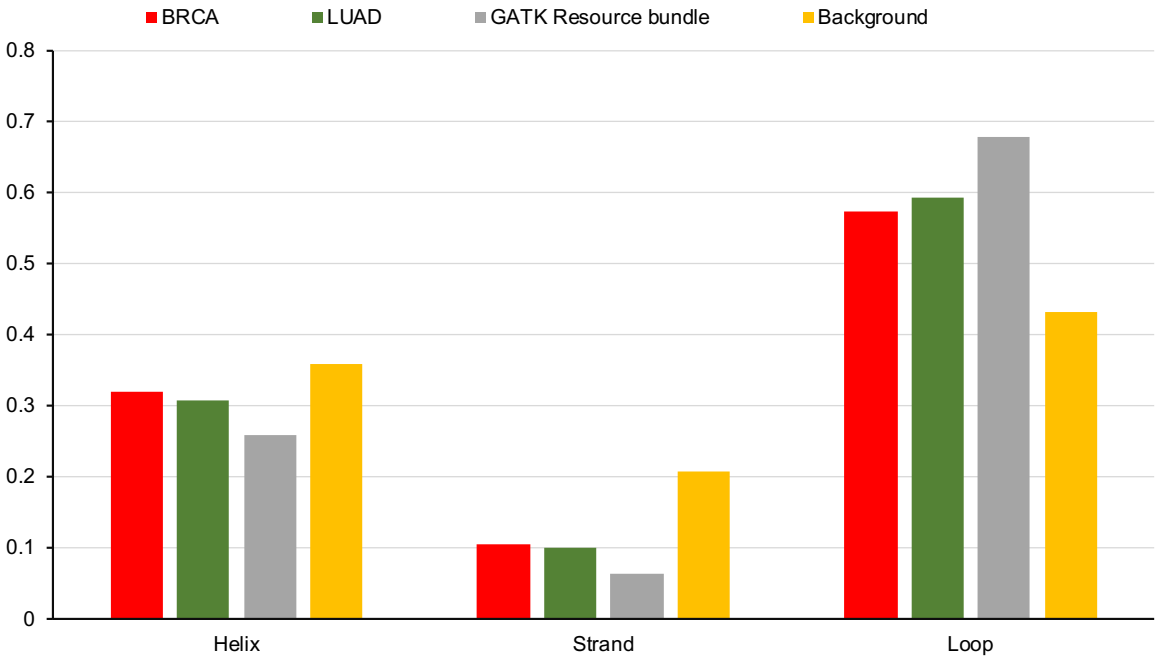

Figure 3.2: Distribution of secondary structure types of somatic NFS indels and germline NFS indels.

Table 3.5: Somatic non-coding indels in BRCA and LUAD

Cancer types	Total indels	Indels in 5'UTR	Indels in 3'UTR	Indels overlapping with TFBS
BRCA	91,464	320 (0.35%)	1,574 (1.72%)	16,689 (18.25%)
LUAD	73,259	335 (0.46%)	974 (1.33%)	20,730 (28.30%)
BRCA ∩ LUAD	12,824	103 (0.80%)	442 (3.45%)	4,193 (32.70%)
Germline	1,267,008	882 (0.07%)	12,680 (1.00%)	178,218 (14.07%)

3.3.3 Non-coding exon somatic indels in BRCA and LUAD genomes

For non-coding somatic indels in BRCA, we found that there are 320 and 1,574 indels in the 5'UTR and 3'UTR regions respectively. For somatic indels in LUAD by Strelka, there are 335 and 974 indels in the 5'UTR and 3'UTR regions respectively, while the numbers are 356 in 5'UTR and 1,257 in 3'UTR by Varscan2 (Tables 3.5 and 3.6). These results show that there are more somatic indels in 3'UTR regions than in 5'UTR regions and Varscan2 predicts more non-coding indels especially in the 3'UTR regions even though it calls fewer total somatic indels (Table 3.2). For germline indels in healthy genomes, 0.07% and 1% of them are in the 5'UTR and 3'UTR respectively. In BRCA and LUAD genomes, the indels are enriched in both 5'UTR (0.35%, 0.46% and 0.70% for BRCA/Strelka, LUAD/Strelka and LUAD/Varscan2, respectively) and 3'UTR (1.72%, 1.33% and 2.48% for BRAC/Strelka, LUAD/Strelka and LUAD/Varscan2, respectively) (Tables 3.5 and 3.6). Compared with the germline indels, somatic non-coding exon indels in cancer genomes are also enriched in the predicted TFBS sequences (18.25%, 28.30% and 23.78% for BRCA/Strelka, LUAD/Strelka and LUAD/Varscan2, respectively) (Tables 3.5 and 3.6).

Table 3.6: Somatic non-coding indels in LUAD by Strelka and Varscan2

Programs	Total indels	Indels in 5'UTR	Indels in 3'UTR	Indels overlapping with TFBS
Strelka	73,259	335 (0.46%)	974 (1.33%)	20,730 (28.30%)
Varscan2	50,746	356 (0.70%)	1,257 (2.48%)	12,066 (23.78%)
Strelka ∩ Varscan2	23,388	202 (0.86%)	613 (2.62%)	6,717 (28.72%)
Germline	1,267,008	882 (0.07%)	12,680 (1.00%)	178,218 (14.07%)

Table 3.7: Somatic indels in SMGs of BRCA and LUAD

Cancer types	Total indels	Indels in SMGs	Indels in SMGs' CDS	Coding indels overlap with TFBS	Indels in SMGs' non-coding region	Non-coding indels overlap with TFBS
BRCA	91,464	1,141 (1.25%)	349 (30.59%)	172 (49.28%)	2,005 (2.19%)	593 (29.58%)
LUAD	73,259	757 (1.03%)	267 (35.27%)	132 (49.44%)	1,390 (1.90%)	423 (30.43%)
BRCA ∩ LUAD	12,824	197 (1.54%)	28 (14.21%)	17 (60.71%)	431 (3.36%)	146 (33.87%)
Germline	1,267,008	4,820 (0.38%)	11 (0.23%)	5 (45.45%)	15,287 (1.21%)	2,518 (16.47%)

3.3.4 Somatic indels on SMGs

After mapping all the somatic indels to the annotated 125 SMGs, we found that somatic indels in cancer genomes are enriched in SMGs, when compared with the germline indels from healthy genomes (Tables 3.7 and 3.8). In addition, somatic indels are more enriched in SMG's coding regions in both cancer genomes. In healthy genomes, there are only 0.23% SMG indels in coding regions, but in cancer genomes, 30.59% and 35.27% of SMG somatic indels are in the coding regions in BRCA and LUSAD respectively (Table 3.7). The common indels identified by both Varscan2 and Strelka showed that 44.6% of the SMG indels are in the coding regions (Table 3.8).

Among the 125 SMGs, 70 of them have BRCA somatic indels in CDS regions while

Table 3.8: Somatic indels in LUAD SMGs with different programs

Programs	Total indels	Indels in SMGs	Indels in SMGs' CDS	Coding indels overlap with TFBS	Indels in SMGs' non-coding region	Non-coding indels overlap with TFBS
Strelka	73,259	757 (1.03%)	267 (35.27%)	132 (49.44%)	1,390 (1.90%)	423 (30.43%)
Varscan2	50,746	660 (1.30%)	180 (27.27%)	90 (50.00%)	1,128 (2.22%)	299 (26.51%)
Strelka ∩ Varscan2	23,388	361 (1.54%)	161 (44.60%)	80 (49.69%)	502 (2.15%)	167 (33.27%)
Germline	1,267,008	4,820 (0.38%)	11 (0.23%)	5 (45.45%)	15,287 (1.21%)	2,518 (16.47%)

Table 3.9: The number of SMGs with somatic indels in BRCA and LUAD

Cancer types	# of SMGs with indels	# of SMGs with indels in CDS regions	# of SMGs with indels in non-coding regions	Both CDS and non-coding regions
BRCA	118	70	125	70
LUAD	114	71	125	71
BRCA ∩ LUAD	74	14	106	12
Germline	119	9	125	9

in LUAD, there are 71 SMGs and 53 SMGs (50 overlapped SMGs) have somatic coding indels by Strelka and Varscan2 respectively (Tables 3.9 and 3.10). When the SMGs with somatic coding indels were compared between BRAC and LUAD, only 14 SMGs appear in both cancer types, suggesting different mutation/variant patterns in different cancer types while there are some commonality between cancer types. These somatic indels have higher FS rate than all coding somatic indels and much higher than the rates in germline indels (Tables 3.3, 3.4, 3.11, and 3.12), suggesting these SMGs contain more deleterious variations in cancer genomes.

For the 125 annotated SMGs, we collect the UTR regions, introns and 100-kbp flanking intergenic regions as the non-coding regions of SMGs (see Methods). All SMGs have somatic indels in non-coding regions by Strelka and three SMGs do not have somatic indels by Varscan2 (Tables 3.9 and 3.10). Interestingly, all SMGs also have germline indels in non-coding regions, but only 9 SMGs have germline indels

Table 3.10: The number of SMGs with somatic indels in LUAD with different programs

Programs	# of SMGs with indels	# of SMGs with indels in CDS regions	# of SMGs with indels in non-coding regions	Both CDS and non-coding regions
Strelka	114	71	125	71
Varscan2	110	53	122	53
Strelka ∩ Varscan2	92	50	111	46
Germline	119	9	125	9

Table 3.11: Distribution of somatic indels from BRCA and LUAD in SMGs

Cancer types	Indels in SMGs	Indels in SMGs' CDS	Coding indels overlap with TFBS	Indels in SMGs' non-coding region	Non-coding indels overlap with TFBS
BRCA	1,141 (Del: 59.86%; Ins: 40.14%)	349 (Del: 63.61%; Ins: 36.39%) (FS: 82.81%; NFS: 17.19%)	172 (Del: 68.60%; Ins: 31.40%)	2,005 (Del: 56.06%; Ins: 43.94%)	593 (Del: 53.87%; Ins: 46.13%)
LUAD	757 (Del: 62.62%; Ins: 37.38%)	267 (Del: 73.41%; Ins: 26.59%) (FS: 82.77%; NFS: 17.23%)	132 (Del: 80.30%; Ins: 19.70%)	1,390 (Del: 58.35%; Ins: 41.65%)	423 (Del: 61.36%; Ins: 38.64%)
BRCA ∩ LUAD	197 (Del: 57.36%; Ins: 42.41%)	28 (Del: 53.57%; Ins: 46.43%) (FS: 82.14%; NFS: 17.86%)	17 (Del: 64.71%; Ins: 36.29%)	431 (Del: 54.06%; Ins: 45.94%)	146 (Del: 53.28%; Ins: 46.72%)
Germline	4,820 (Del: 56.24%; Ins: 43.76%)	11 (Del: 54.55%; Ins: 45.45%) (FS: 27.27%; NFS: 72.73%)	5 (Del: 20.00%; Ins: 80.00%)	15,287 (Del: 56.34%; Ins: 43.66%)	2,518 (Del: 55.95%; Ins: 43.05%)

Table 3.12: Distribution of somatic indels in SMGs from LUAD with different programs

Programs	**Indels in SMGs**	**Indels in SMGs' CDS**	**Coding indels overlap with TFBS**	**Indels in SMGs' non-coding region**	**Non-coding indels overlap with TFBS**
Strelka	757 (Del: 62.62%; Ins: 37.38%)	267 (Del: 73.41%; Ins: 26.59%) (FS: 82.77%; NFS: 17.23%)	132 (Del: 80.30%; Ins: 19.70%)	1,390 (Del: 58.35%; Ins: 41.65%)	423 (Del: 61.36%; Ins: 38.64%)
Varscan2	660 (Del: 66.67%; Ins: 33.33%)	180 (Del: 75.56%; Ins: 24.44%) (FS: 87.78%; NFS: 12.22%)	90 (Del: 78.89%; Ins: 21.11%)	1,128 (Del: 62.94%; Ins: 37.06%)	299 (Del: 61.38%; Ins: 38.62%)
Strelka ∩ Varscan2	361 (Del: 67.59%; Ins: 32.41%)	161 (Del: 75.78%; Ins: 24.22%) (FS: 86.34%; NFS: 13.66%)	80 (Del: 78.75%; Ins: 21.25%)	502 (Del: 63.15%; Ins: 36.85%)	167 (Del: 63.75%; Ins: 36.25%)
Germline	4,820 (Del: 56.24%; Ins: 43.76%)	11 (Del: 54.55%; Ins: 45.45%) (FS: 27.27%; NFS: 72.73%)	5 (Del: 20.00%; Ins: 80.00%)	15,287 (Del: 56.34%; Ins: 43.66%)	2,518 (Del: 55.95%; Ins: 43.05%)

in coding regions (Tables 3.9 and 3.10). Among all these SMGs, ATM has the most number of BRCA non-coding indels (45 by Strelka) and EPHB6 has the most number of LUAD non-coding indels (36 by Strelka and 40 by Varscan2). When both cancer types are considered, AJUBA has the most somatic indels in its non-coding regions (15 by Strelka). We also checked the overlap between these non-coding somatic indels and the predicted TFBS sequences. There are 29.58%, 30.43% and 26.51% somatic non-coding indels on SMGs from BRCA/Strelka, LUAD/Strelka and LUAD/Varscan2 overlap with TFBS, higher than that in the germline indels in healthy genomes (16.47%) (Tables 3.7 and 3.8). Those non-coding indels that appear in both BRCA and LUAD have an even higher percentage to overlap with TFBS sequences (33.87%). These results indicate that somatic indels in BRCA and LUAD genomes are enriched in TFBSs. In Chapter 4, we found that there are 2.8% predicted TFBS in coding regions. So we checked the overlap between coding indels on SMGs and TFBS sequences. There are 49.28%, 49.44% and 50.00% somatic coding indels on SMGs from BRCA/Strelka, LUAD/Strelka and LUAD/Varscan2 overlap with TFBS, slightly higher than that in the germline indels in healthy genomes (45.45%) (Tables 3.7 and 3.8). The importance of TFBSs in the regulation of gene expression suggests that these indels may disrupt the normal functions of transcriptional machinery in the cancer development or as results of the disease.

3.4 Discussion

With the development of biotechnology, especially the invention of the NGS technology, a large number of genomes have been sequenced with a variety of cancer types. Somatic variations in cancer genomes have been one of the main focuses in cancer studies, including variants in both coding and non-coding regions [4]. However, most of the cancer genome studies focused on SNPs [95, 120, 127]. In this study, we analyzed the somatic indels in both BRCA and LUAD genomes.

The somatic indels from different cancer types vary greatly. There are only 16,909

indels in both cancer types, which account for 15.39% of BRCA and 18.55% of LUAD somatic indels respectively. Different somatic indel annotation programs also produce different results for the same genome. A total of 25,278 somatic indels are predicted by both Strelka and Varscan2 in LUAD genomes, representing 27.73% and 45.67% of indels from Strelka and Varscan2 respectively. Among these indels, many of them are germline indels and are removed for feature analyses.

The percentage of deletions in LUAD genomes is higher than that in BRCA genomes (Tables 3.1 and 3.2). Somatic indels that exist in both BRCA and LUAD genomes may represent common cancer disease features. The indels called by both prediction programs may have higher confidence to be true indels as discussed in Chapter 2. Compared with germline indels in healthy genomes, somatic indels in cancer genomes have higher proportion in CDS regions. Coding somatic indels also have higher rate of FS types, especially in SMGs (Tables, 3.3, 3.4, 3.11 and 3.12). This phenomenon is not surprising since FS indels are prone to be deleterious [43, 32, 137]. The results of protein secondary structure analysis show that somatic indels are enriched in helix and strand but depleted in loop, when compared with germline indels. Somatic indels in cancer genomes may affect the core structural elements, which in turn change protein structures and functions, leading to disease development.

As reported by Kandoth *et al.*, the top 3 most frequently mutated genes in BRCA are TP53, GATA3 and MLL3, and the top 3 most frequently mutated genes in LUAD are TP53, MLL3 and TSHZ3 [130]. In our study, we focus on the somatic indels on these SMGs and find that TP53, ATM203 and TSHZ2 are the top 3 genes with most indel variations in BRCA. For LUAD, the top 3 genes are TP53, STK11 and NAV3 revealed by Strelka and TP53, EPHB6 and ATM203 predicted by Varscan2. Additional work can help explain if these differences are caused by variation types or algorithms for somatic indel prediction.

Last but not the least, non-coding somatic indels may also play important roles in

cancer development. Somatic indels in BRCA and LUAD are enriched in both 5'UTR and 3'UTR, as well TFBS sequences, suggesting changes in regulatory regions are a big part of the cancer disease mechanisms.

CHAPTER 4: Evolution of Exonic Enhancers in the Human Genome

4.1 Background

Eukaryotic genomes contain largely two types of functional sequences: coding-sequences (CDSs) that encode proteins or RNAs, and cis-regulatory sequences or modules (CRMs) such as promoters and enhancers that control the expression of target CDSs [139, 140, 141]. Usually CRMs are located in the intergenic regions or in the introns of genes [142, 143, 144], therefore they do not overlap with CDSs as well as 5'-untranslated regions (5'-UTRs) and 3'-UTRs at the start and end of the first exons and last exons, respectively. However, it has long been known that in some cases CDSs, 5'-UTRs and 3'-UTRs can also function as CRMs, particularly enhancers for different genes [69, 88, 145, 146, 147, 148, 149]. Since mutations of such codonic enhancers (cEHs) and UTR enhancers (uEHs) can result in diseases by altering their enhancer activities [69, 87, 150, 151, 152], it is important to understand the prevalence and evolution of exonic enhancers (eEHs) in the human genome [69, 153]. More recently, it was reported that at least 15% of codons in the human genome were hypersensitive to DNase I treatment in 81 human cell types and thus were likely in dual-use for defining both specific amino acid sequences for protein functions and specific transcription factors (TFs) binding sites (TFBSs) for gene transcriptional regulation [70]. These so-called duons were found to be more conserved than non-duon codons at fourfold degenerate sites, and thus thought to be under additional selection constraints owing to the dual-use [70]. It was also reported that mutations in these duons could lead to diseases by altering the activities of relevant cEHs [154]. However, a recent reanalysis of the duons showed that the so-called duons were in fact not more conserved than non-duon codons when the biased usage of A/T and

C/G at the third position of the 21 synonymous codon sets was taken into account [89]. As most of synonymous sites in duons were evolutionarily neutral, these authors cast doubt on any role of the duons in transcriptional regulation [89]. A similar conclusion was drawn based on an analysis of three known cEHs [90]. Therefore, although there is no doubt about the existence of cEHs and uEHs, their prevalence and how they evolve are under hot debate [89, 90]. On the other hand, it has been shown that DNase I hypersensitive sites (DHSs) are not a reliable predictor of TFBSs due to their low resolution and high false positives [155, 156, 157, 158]. It is highly likely that a considerable proportion of duons predicted solely based on DNase I hypersensitivity signals [70] may not contain any TFBSs, or only some part of it harbors TFBSs. Therefore, the evolutionary neutrality of the duons observed by Xing and He [89] might be due to the inaccuracy of the predicted duons used in their analysis. To clarify these contradictory results, sufficient experimentally identified eEHs are needed, however, such a dataset is still lacking. Thus, a set of predicted CRMs with high accuracy might be the solution.

A recently developed dePCRM2 was used to predict de novo CRMs and constituent TFBSs, using a large amount of integrated TF ChIP-seq datasets. dePCRM2 divides the TF binding peaks sequences into two sets: CRM candidates (CRMCs) and non-CRMs. Each CRMC is evaluates by its TFBSs element(s) [136]. We applied dePCRM2 to predict CRMC with 6,092 TF ChIP-seq datasets consisting 779 types of TFs from 2,631 divers cell/tissue types. we found 1,404,973 CRMCs with a total of 1,359,824,275bp (56.84%), while the non-CRMs have 1,032,664,424bp (43.16%). The CRMCs and non-CRMs represent 44.03 and 33.44% of the human genome sequence respectively. Validation on experimentally determined CRM function-related sequence elements such as VISTA enhancers and ClinVar variants indicates that our predicted CRMCs are highly accurate, with an estimated false positive rate (FPR) of 0.045% and false negative rate (FNR) of 1.27%. FPR and FNR for predicted

CRMs decrease rapidly with the decrease of p-values. For instance, at a highly stringent p-value cutoff of $5x10^{-6}$, we predicted 428,628 CRMs and 38,507,543 constituent TFBSs, covering 31.81% and 12.88% of the genome, respectively. As expected, the majority of these predicted CRMs (94.42%) and constituent TFBSs (93.65%) sites are located in non-exonic sequences (NESs). Surprisingly, the remaining 5.58% of the CRMs and 6.35% of TFBSs sites overlap exons (CDSs or UTRs). The predicted CRMs in NESs tend to be under either stronger purifying selection or stronger positive selection than the non-CRMs. Thus, it is interesting to see how the predicted TFBSs in exons evolve compared to their counterparts in non-CRMs. These putative exonic TFBSs (eTFBSs) also provide us an opportunity to re-examine the duon hypothesis and associated fundamental questions for possible dual-use of codons and UTRs.

4.2 Methods

4.2.1 Datasets

The non-CRMs as well as CRMs and constituent TFBSs predicted at p-value $\leq$ $5x10^{-6}$ in the human genome were downloaded from the original paper (in submission). The human genome assembly version GRCh38.p13 was used as the reference genome. Annotations (CDSs, 5'-UTRs and 3'-UTRs) of genes and verified transcripts were downloaded from the Ensembl Release 100 (https://useast.ensembl.org/index.html). For each gene, the longest annotated transcript (agreed by different databases and supported by mRNA data) was selected for coding region assignments. Coordinates of proteins with known structures were download from the Protein Data Bank (PDB) [133]. The ChIA-PET dataset in the K562 cell line was downloaded from the GEO database with the access number GSE33664. TF ChIP-seq datasets were downloaded from the Cistrome database [159].

4.2.2 Assignment of secondary structure types

For proteins with known structures in PDB, we assigned secondary structure types of amino acid sequences using the DSSP program [134] as described previously in Chapter 3. For proteins without known structures, we assigned secondary structure types of residues as follows. We first generated a non-redundant protein sequence set (less than 30% sequence identity) for all annotated proteins in the human genome using CD-HIT [160]. If a protein in the dataset has a highly homologous protein with known structure in PDB with at least 50% coverage and 80% sequence identity, the secondary structure types of the template protein were copied to the target protein. If no homologous structures were found in PDB, RaptorX, a highly accurate secondary structure prediction program, was applied to predict secondary structure types with default settings [161].

4.2.3 Assignment of conservation scores

The GERP(Genomic Evolutionary Rate Profiling)[162] and phyloP [163] scores of each nucleotide site in the human genome were downloaded from the UCSC Genome Browser database [164].

4.2.4 Analysis of chromatin interactions

We identified eTFBSs that overlap with any TF ChIP-seq binding peak within 500 bp region centering on its summit, collected in the K562 cells that were used in our CRM and TFBS prediction (in submission), with an assumption that they are active in the cells, while the remaining eTFBSs that do not overlap any TF ChIP-seq binding beaks in the cells are inactive. We calculated significant interactions between two loci using ChiaSig (Paulden, *et al.* 2014) with the ChIA-PET reads from the K562 cells. We then counted the number of the putative active eTFBSs in close proximity to promoters that are at least 5,000 bp away from the eTFBSs to exclude the promoter of its own gene. As a control, we performed the same analysis on the

putative inactive eTFBSs.

4.2.5 Statistical analysis

We used Kolmogorov-Smirnov test, Mann-Whitney U test and χ^2 test to evaluate statistical significant levels of hypothesis tests as indicated in the text and figure legends.

4.3 Results

4.3.1 eTFBSs distribution in CDSs and UTRs

To analyze the evolution of eEHs, we used the 428,628 CRMs containing 38,507,543 putative TFBSs predicted at a stringent p-value cutoff $5x10^{-6}$ to minimize FPR, and focused on the TFBSs in the CRMs. After merging overlapping TFBSs, we ended up with 25,297,119 non-overlapping TFBSs with a total length of 397,703,041bp, covering 12.88 % of the genome (3,088,269,832bp). Of these TFBS sites, 372,677,171bp (93.71%), 11,121,887bp (2.80%), 3,618,277bp (0.91%) and 10,244,058bp (2.58%) are located in annotated NESs (nTFBSs), CDSs (cTFBSs), 5'-UTRs (5'-uTFBSs) and 3'-UTRs (3'-uTFBSs) (Figure 4.1a), respectively. If a site is annotated in both CDS and UTR of different transcripts, we consider it as a CDS site. If a site is annotated in both a 5'-UTR and a 3'-UTR of different transcripts, we group them as others with a total number of 41,648bp (0.01%) (Figure 4.1a). To make sure these 24,984,222bp predicted eTFBS sites are transcribed as exons in protein-coding genes, we mapped them to the 19,694 experimentally verified mRNA transcripts in human tissues, and found that 20,291,510bp (81.22%) could be mapped to one of the 19,060 mRNA transcripts. Specifically, 11,102,033pb (54.71%), 1,514,454bp (7.46%) and 7,675,023bp(37.82%) were mapped to verified CDSs, 5'-UTRs and 3'-UTRs of 18,971, 12,903 and 13,237 genes, respectively (Figure 4.1b). Our subsequent analyses focus on these eTFBSs in verified exons in the mRNAs, which comprise 33.85%, 48.78% and 32.48% of the total length of annotated CDSs (32,797,669bp, average length

1,729bp), 5'-UTRs (3,104,596bp, average length 241bp) and 3'-UTRs (23,630,714bp, average length 1,785bp), respectively (Figure 4.1c). Thus, CDSs and 3'-UTRs have similar eTFBS densities that are lower than that in 5'-UTRs. The 19,060 transcribed genes containing at least one eTFBS harbor an average of 56 potentially overlapping eTFBSs, forming eEHs (Figure 4.1d). As shown in Figure 4.1e, cTFBSs and 3'-uTFBSs tend to be clustered at the 5'-end and 3'-end of CDSs and 3'-UTRs, while 5'-uTFBSs are evenly located along most part of 5'-UTRs but tend to avoid the 5'-end and 3'-end.

4.3.2 C/G contents at degenerate sites in cTFBSs, 5'-uTFBSs and nTFBSs

The key evidence supporting the duon hypothesis was the observation that fourfold degenerate sites in duons were more conserved than those in non-duon codons [70]. However, a reanalysis of the original DHSs data found that the duons actually evolved similarly to non-duon codons when the conservation levels or substitution rates of A/T and C/G were compared separately, using either phyloP [163] conservation scores or substitute rates derived from the most recent common ancestor of humans and gorilla compared with chimpanzee [89]. Thus, the authors augured that the earlier conclusions [70] were incorrectly drawn due to the higher C/G frequencies at the fourfold degenerate sites in duons than those in non-duon CDSs, and that C/G at the fourfold degenerate sites tend to be more conserved than A/T [89]. This result casts doubts on the validity of the duon hypothesis [89]. Therefore, it is interesting to see how our predicted cTFBSs and uTFBSs evolve compared with the counterparts in non-CRMs.

To this end, we first compared the usage of the degenerate third codon positions of 21 synonymous codon sets in the cTFBSs with those in the non-CRMs (see Materials and Methods). As shown in Figure 4.2a, C/G are more preferred at the degenerate third positions in the cTFBSs than those in non-CRM CDSs ($P < 2.2\text{x}10^{-16}$, χ^2 test) for all the synonymous codon sets, which is in agreement with the earlier finding that

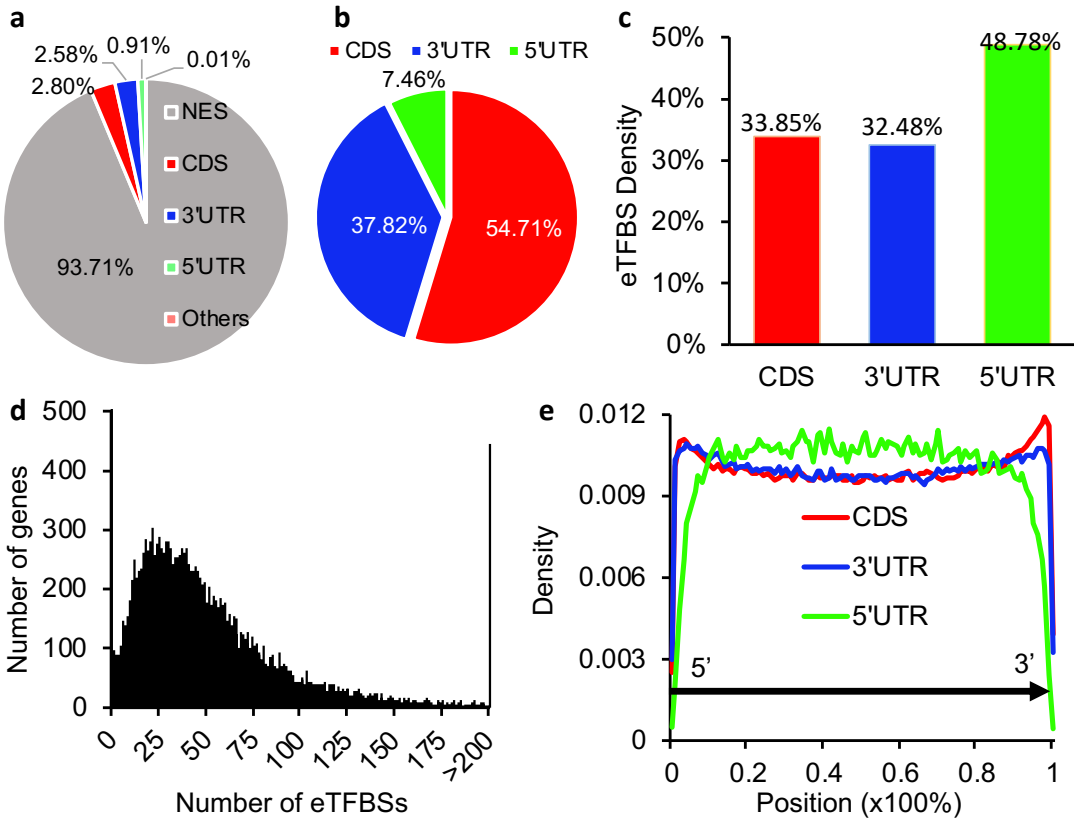

Figure 4.1: Properties of the predicted TFBSs located in annotated exons in the human genome.

a. Distribution of predicted TFBSs in non-exonic sequences (NESs), coding sequences (CDSs), 5'-UTRs, 3'-UTRs and in sequences in both 5'-UTR and 3'-UTR annotations (others). b. Distribution of TFBSs in CDSs, 5'-UTRs and 3'-UTRs in experimentally verified exons. c. Percentage of the length of experimentally verified CDSs, 5'-UTRs and 3'-UTRs that are predicted as exonic TFBSs (eTFBSs). d. Number of transcribed genes containing different numbers of predicted eTFBSs. e. Distribution of the predicted eTFBSs along the CDSs, 5'-UTRs and 3'-UTRs of genes from the 5'-end to the 3'-end as indicated by the horizontal arrow.

C/G was more preferred at the degenerate third positions of the synonymous codon sets in the duons than those in non-duon CDSs [89]. Furthermore, although the human genome is A/T-rich (59% A/T vs. 41% C/G) [165], there are more C/G than A/T in the cTFBSs for all the synonymous sets, except for Asn (AA*), Ile(AT*), Ser(TC*) and Thr(AC*), while this is true only for Gln (CA*), Leu(CT* and TT*) and Val (GT*) in non-CRM CDSs (Figure 4.2b). Overall, degenerate third positions in cTFBSs have higher C/G contents than those in non-CRM CDSs (Figure 4.2c).

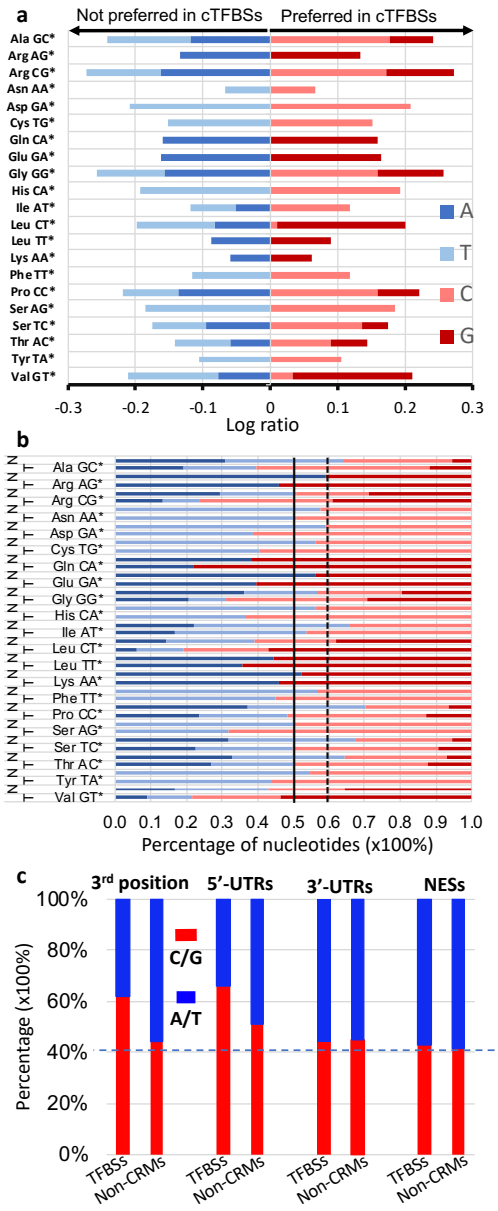

Figure 4.2: Biased distribution of A/T and C/G in predicted TFBSs.

a. Preferences of A/T and C/G in the third positions of the 21 synonymous codon sets in the predicted codonic TFBSs (cTFBSs) relative to the those in non-CRM CDSs. b. Elevation of C/G contents in the third positions of the codons in cTFBSs and non-CRM CDSs. The dotted vertical line indicates the neutral expectation (59%) of A/T contents in the genome. c. A/T and C/G contents in the predicted TFBSs in CDS (the third codon position), 5'-UTRs, 3'-UTRs, and NESs in comparison with the those in their counterparts in non-CRMs. The dotted horizontal line indicates the neutral expectation (41%) of C/G contents in the genome. P-values for the difference of frequencies of A/T and C/G between TFBSs and those in their counterparts in non-CRM CDSs were computed using χ^2 test, and $P < 2.2\text{x}10^{-16}$ for all the comparisons.

Interestingly, 5'-uTFBSs and nTFBSs also have higher C/G contents than respective counterparts in non-CRM 5'-UTRs and non-CRM NESs, but the opposite is true for 3'-uTFBSs (Figure 4.2c). Furthermore, C/G contents in the degenerate third codon positions, 5'-UTRs, 3'-UTRs are higher than the neutral expectation (41%), regardless of their location in TFBSs or in non-CRMs (Figure 4.2c). It has been observed that C/G contents at fourfold degenerate sites are consistently elevated above the neutral expectation [166], this is particularly true for the sites in the cTFBSs. Interestingly, C/G contents in nTFBSs are also higher than the neutral expectation (41%) and those in non-CRM NESs, which are close the neutral expectation (Figure 4.2c).

4.3.3 Evolutionary constraints of cTFBSs

Since GERP [167] and phyloP [163] scores of nucleotide sites in the human genome are widely used to quantify their conservation levels, we compared the GERP and phyloP scores of A/T and C/G at the non-degenerate sites at the first and second codon positions and degenerate sites at the third positions of the 21 synonymous codon sets in the cTFBSs with those in the non-CRM CDSs. A large positive conservation score (GERP or phyloP) suggests a higher possibility of purifying selection; on the other hand, a small negative conservation score of a site may be a result of positive selection. A conservation score close to zero indicates that the site is likely under neutral selection. As previously described, a site with a conservation score of more than 1 is considered under purifying selection, while a score of less than 1 is considered under positive selection. The remaining sites are under neutral selection. The proportion of a set of nucleotide sites is considered to be under positive selection, selectively neutral or under purifying selection, as the size of the area of the conservation scores under the curve of the density distribution for three blocks (minimum, -1), [-1, 1] and (1, maximum), respectively. Evolutionary constraints on non-degenerate sites at the first and second codon positions reflect the selection pressure from the requirement of amino acids encoded by the sites for protein structure and functions

and other functions of CDSs, if any; while the evolutionary constraints on degenerate sites at the third codon positions indicate the selection pressure on the sites required for their functions other than protein-coding.

As expected, most (>60%) of non-degenerate sites at both the first and second codon positions are under purifying selection (Figures 4.3a - 4.3d). However, either A/T or C/G at both positions in the cTFBSs tend to have higher proportions of sites under purifying selection (68.24% vs. 61.05% for A/T and 71.00% vs. 65.97% for C/G at the first codon position sites, and 79.05% vs. 72.46% for A/T and 72.45% vs. 65.32% for C/G at the second codon position sites, measured by the GERP scores) than those in the non-CRM CDSs ($p < 2.2x10^{-16}$, KS-test). In general, a small portion (<15%) of these non-degenerate sites are under positive selection; but either A/T or C/G at both codon positions in the cTFBSs tend to have lower proportions of sites under positive selection (14.22% vs. 15.44% for A/T and 9.39% vs. 9.68% for C/G at the first codon position sites, and 8.47% vs. 9.54% for A/T and 9.67% vs. 10.78% for C/G at the second codon position sites) than the non-CRM CDSs ($p < 2.2x10^{-16}$, KS-test). Interestingly, although in general an intermediate portion (around 20%) of these non-degenerate sites are selectively neutral or nearly so, either A/T or C/G at both codon positions in the non-CRM CDSs tend to have higher proportions of sites to be selectively neutral or nearly so (17.54% vs. 23.51% for A/T and 19.61% vs. 24.35% for C/G at the first codon position, and 12.49% vs. 17.99% for A/T and 17.88% vs. 23.90% for C/G at the second codon position) than those in the cTFBSs ($p < 2.2x10^{-16}$, KS-test), as indicated by the much larger peaks around score 0 in their density graphs than those of both A/T and C/G in the both positions in the cTFBSs (Figures 4.3a and 4.3c). These results indicate that the non-degenerate sites at the first and second codon positions in cTFBSs are more likely to be conserved than those in the non-CRM CDSs that contain a higher proportion of selectively neutral sites, suggesting that at least a small portion (5-7%) of non-degenerate sites

are in dual-use. By contrast, only about a third of the degenerate sites at the third codon positions are under purifying selection, but A/T in the cTFBSs have a lower proportion (19.01% vs. 21.91%), while C/G in the cTFBSs have a higher proportion (33.84% vs. 32.24%), of sites under purifying selection, than those in the non-CRM CDSs (Figures 4.3e - 4.3f) ($p < 2.2x10^{-16}$, KS-test). Compared to non-degenerate sites at the first and second codon positions (<15%, Figures 4.3a - 4.3d), a much higher proportion (>25%) of degenerate sites at the third codon positions are under positive selection (Figures 4.3e - 4.3f). However, both A/T and C/G at the third codon positions in the cTFBSs have higher proportions (41.12% vs. 32.47% for A/T and 29.45% vs. 27.24% for C/G) of sites under positive selection than those in the non-CRM CDSs (Figures 4.3e - 4.3f) ($p < 2.2x10^{-16}$, KS-test). Furthermore, more than a third of the degenerate sites at the third positions are selectively neutral or nearly so (Figures 4.3e - 4.3f), however, both A/T and C/G in the cTFBSs have a lower proportion of neutrality (39.88% vs. 45.62% for A/T and 36.7% vs. 40.52% for C/G) than the non-CRM CDSs($p < 2.2x10^{-16}$, KS-test). These results suggest that the A/T at the degenerate sites in the cTFBSs evolve faster than those in non-CRM CDSs, while C/C at the degenerate sites in the cTFBSs evolve either faster or slower than those in non-TFBS CDSs. This conclusion therefore is in sharp contrast to either of the earlier two observations that the degenerate sites in duons are more conserved than those in non-duons (19), or that both A/T and C/G at the degenerate sites in the duons are similarly or slightly more conserved compared with those in non-duon CDSs(21), indicating that our predicted cTFBSs are distinct from the earlier predicted duons based on DHSs. Taken together, our results indicate that non-degenerate sites at the first and second positions of codons in the cTFBSs are more likely to be conserved than those in non-CRM CDSs, while the degenerate sites at the third codon positions in cTFBSs are more likely to be under either positive selection or purifying selection than those in non-CRM CDSs.

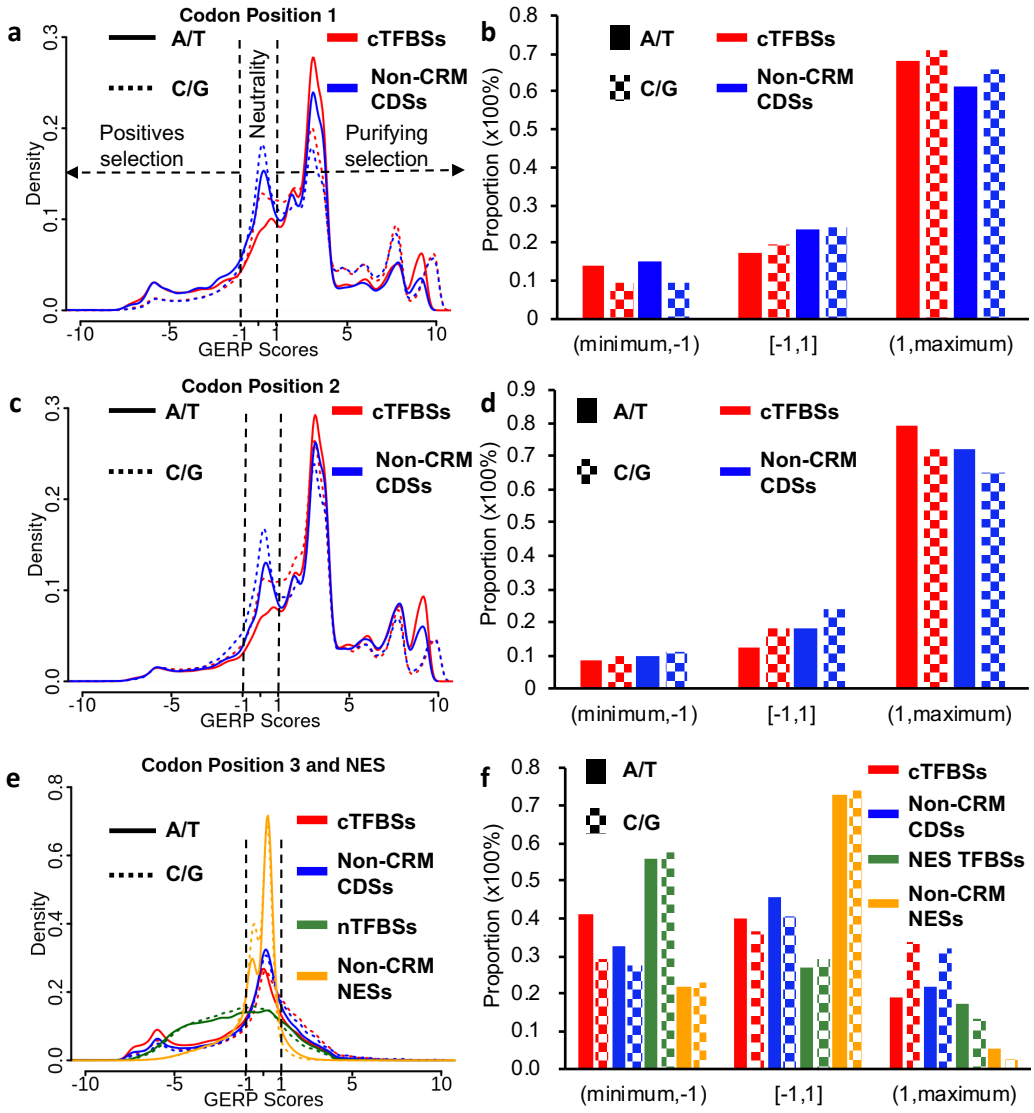

Figure 4.3: Comparison of GERP scores in the predicted cTFBSs and those in non-CRM CDSs.

Distributions of GERP scores of A/T and C/G at sites in the first (a), second (c) and third (e) positions of the synonymous codons in the predicted cTFBSs and non-CRM CDSs. Proportions of A/T and C/G at sites that are under positive selection (GERP scores in (minimum, -1)), purifying selection (GERP scores in (1, maximum)), or evolutionarily neutral (GERP scores in [-1, 1]), at the first (b), second (d) and third (f) positions in the synonymous codons in the predicted cTFBSs and non-CRM CDSs. GERP scores of A/T and C/G at sites in the third positions (e and f) are also compared with those in NESs. P-values were computed using Kolmogorov-Smirnov test for the density.

In addition, we reason that if the degenerate sites in non-CRM CDSs are selectively neutral, then they should evolve similarly to sites in non-CRM NESs; and if the only

function of degenerate sites in the cTFBSs is for TF binding, then they should evolve similarly to nTFBSs. As shown in Figures 4.3e and 4.3f, both A/T and C/G at the degenerate sites in non-CRM CDSs are more likely to be under either positive selection (32.47% vs. 22.08% for A/T and 27.24% vs. 23.14% for C/G) or purifying selection (21.91% vs. 5.24% for A/T and 32.24% vs. 2.74% for C/G), and less likely to be evolutionally neutral, than those in non-CRM NESs, suggesting that many degenerate sites in non-CRM CDSs may have some biological functions other than TF binding. On the other hand, both A/T and C/G at the degenerate sites in cTFBSs are more likely to be under purifying selection (19.01% vs. 17.28% for A/T and 33.84% vs. 13.36% for C/G), a small portion of A/T are more likely under strongly positive selection (GERP score $<$ -6.5), than those in the nTFBSs (Figure 4.3e and 4.3f), suggesting that a small portion of A/T (around 2%) and up to 20% C/G at these degenerate sites might be in dual-use. The same results are obtained when the phyloP scores were used for these analyses (Figure 4.4a - 4.4f).

4.3.4 Evolutionary constraints of uTFBSs

As we indicated earlier, 1,514,454bp (7.46%) and 7,675,023bp(37.82%) of our predicted eTFBS sites were mapped to verified 5'-UTRs and 3'-UTRs, respectively (Figure 4.1b). To see how these putative 5'-uTFBS and 3'-uTFBS sites evolve, we compared the GERP scores of their A/T and G/C with those in non-CRM UTRs. As shown in Figures 4.5a - 4.5d, only about 25% of both A/T and C/G in both 5'-uTFBSs and 3'-uTFBSs are selectively neutral, and they are much more likely to be under either purifying selection (22.69% vs. 6.38% for A/T and 29.49% vs. 5.68% for C/G in 5'-UTRs; and 25.96% vs. 7.22% for A/T and 21.70% vs. 3.70% for C/G in 3'-UTRs) or positive selection (56.78% vs. 29.67% for A/T and 42.43% vs. 26.89% for C/G in 5'-UTRs; and 50.24% vs. 28.07% for A/T and 52.11% vs. 27.45% for C/G in 3'-UTRs) than those in respective non-CRM UTRs. Therefore, both A/T and C/G in uTFBSs evolve similar to those in the nTFBSs. However, Both A/T

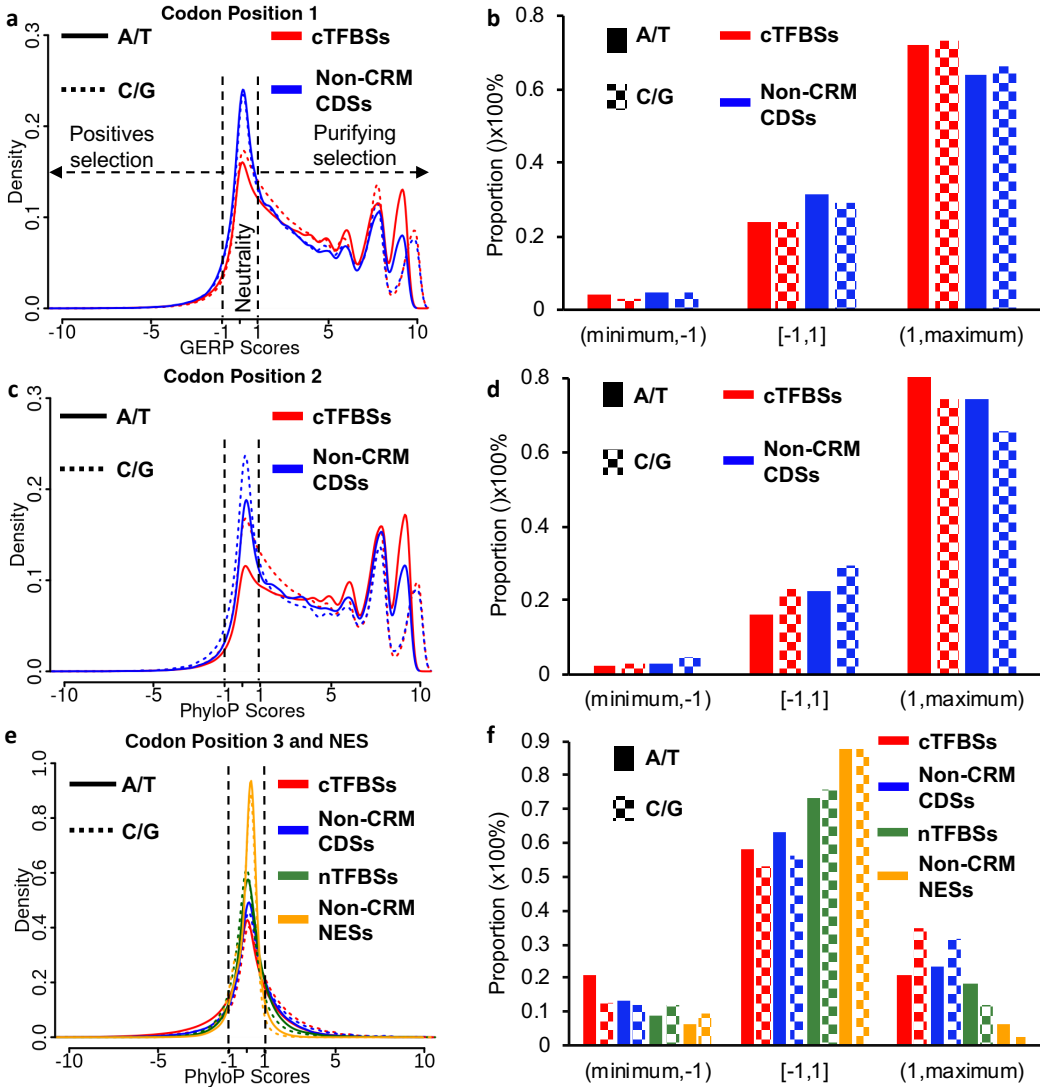

Figure 4.4: Comparison of phyloP scores in the predicted cTFBSs and those in non-CRM CDSs.

Distributions of phyloP scores of A/T and C/G at sites in the first (a), second (c) and third (e) positions of the synonymous codons in the predicted cTFBSs and non-CRM CDSs. Proportions of A/T and C/G at sites that are under positive selection (phyloP scores in (minimum, -1)), purifying selection (phyloP scores in (1, maximum)), or evolutionarily neutral (phyloP scores in [-1, 1]), at the first (b), second (d) and third (f) positions in the synonymous codons in the predicted cTFBSs and non-CRM CDSs. phyloP scores of A/T and C/G at sites in the third positions (e and f) are also compared with those in NESs. P-values were computed using Kolmogrov-Smirnov test for the density.

and C/G in uTFBSs tend to be more likely to be under purifying selection (22.69% vs. 17.28% for A/T and 29.49% vs. 13.36% for C/G in 5'-UTRs; and 25.96% vs.

17.28% for A/T and 21.70% vs. 13.36% for C/G in 3'-UTRs) than those in nTFBSs (Figure 4.5a - 4.5d), suggesting that at least a portion of uTFBSs are in dual-use. Interestingly, both A/T and C/G in both non-CRM 5'-UTR and non-CRM 3'-UTR sites are more likely to be under either purifying selection (6.38% vs. 5.24% for A/T and 5.68% vs. 2.74% for C/G in 5'-UTRs; and 7.22% vs. 5.24% for A/T and 3.70% vs. 2.74% for C/G in 3'-UTRs) or positive selection (29.67% vs. 22.08% for A/T and 26.89% vs. 23.14% for C/G in 5'-UTRs; and 28.07% vs. 22.08% for A/T and 27.45% vs. 23.14% for C/G in 3'-UTRs), than those in non-CRM NESs (Figures 4.5a - 4.5d), suggesting that non-CRM UTR sites might have biological functions other than TF binding, such as coding for ribosome entry sites in 5'-UTRs [168] and biding sites for RNA-binding protein (RBPs) in 3'-UTRs [169]. The similar evolutionary behaviors of uTFBSs to that of nTFBSs strongly suggest that uTFBSs might function as TFBSs. Similar results were seen using the phyloP scores (Figures 4.6a - 4.6d).

4.3.5 Location of cTFBSs on protein structures

One of the puzzles for the dual-use of CDSs is how the two irrelevant functions of a DNA sequence can be possibly co-evolved. To address this, we mapped the amino acids encoded by the cTFBSs to known 3D structures of proteins in PDB (Materials and Methods). To reduce the biased distribution of structures to some protein families, we generated a 30% identity non-redundant protein set (MATERIALS AND METHODS) whose CDSs contain a total of 7,266,384 bp cTFBS sites. Amino acids encoded by 544,490bp (7.49%) out of the 7,266,384 bp cTFBS sites could be mapped to 1,761 known protein structures. These mapped amino acids are enriched in loops (48.73%) compared with the proportion of length of loops in host proteins (46.65%, $p < 2.2x10^{-16}$) as well as in all proteins with known structures(40) (47.10%, $p < 2.2x10^{-16}$) (Figure 4.7a). On the other hand, these mapped amino acids are underpresented in helices (33.53%) and strands (17.74%) compared with the proportions of the lengths of helices and strands in the host proteins (34.31% and 19.04%, respec-

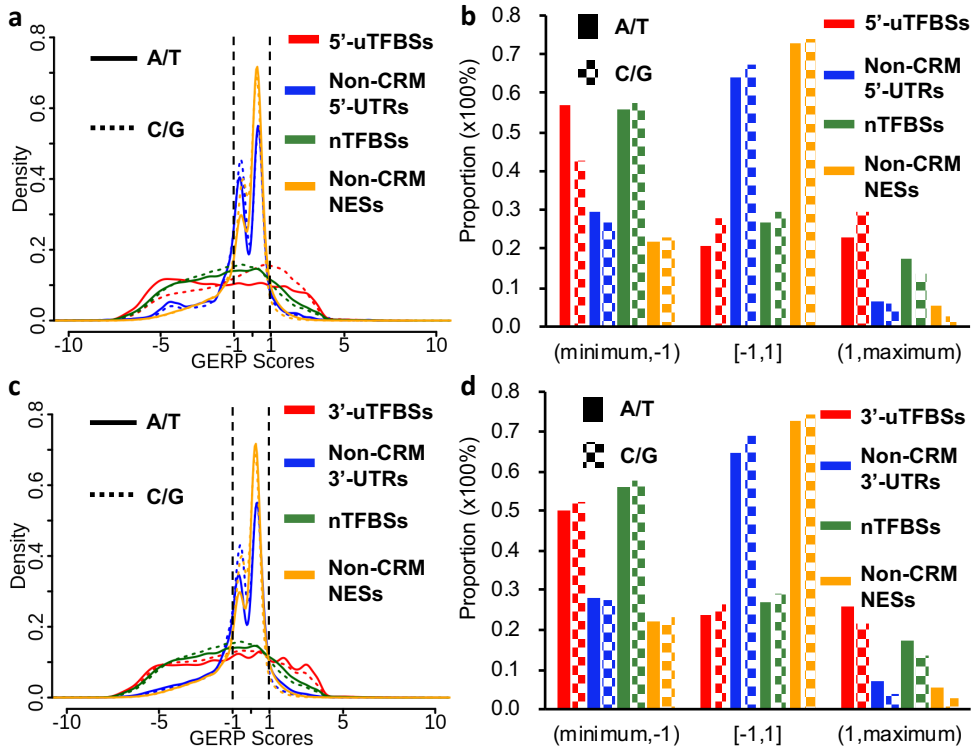

Figure 4.5: Comparison of GERP scores of the predicted uTFBSs, non-CRM UTRs, nTFBSs and non-CRM NESs.

Distributions of GERP scores of A/T and C/G sites in 5'-uTFBSs (a), 3'-uTFBSs (b) and their counterparts in non-CRMs, in comparison with those in nTFBSs and non-CRM NESs. Proportions of the sites that are under positive selection (GERP scores in (minimum, -1)), purifying selection (GERP scores in (1, maximum)), or selectively neutral (GERP scores in [-1, 1]), in 5'-UTRs (c) and 3'-UTRs (d) and their counterparts in non-CRMs, in comparison with those in nTFBSs and non-CRM NESs. P-values were computed using Kolmogorov-Smirnov test for the density

tively, $p < 2.2x10^{-16}$) as well as in all proteins with known structures [138] (34.21% and 18.69%, respectively, $p < 2.2x10^{-16}$) (Figure 4.7a). For the amino acids encoded by the remaining 6,721,894bp (92.51%) cTFBSs sites, which could not be mapped to any known protein structures, we predicted their secondary structure types using RaptorX [161]. A similar pattern was found, in which the amino acids encoded by these cTFBSs are enriched in loops but depleted in helices and strands (Figure 4.7b). Specifically, 60.79%, 30.01%, and 9.20% of the peptides encoded by cTFBSs were predicted to adopt loops, helices and strands, respectively, while these proportions

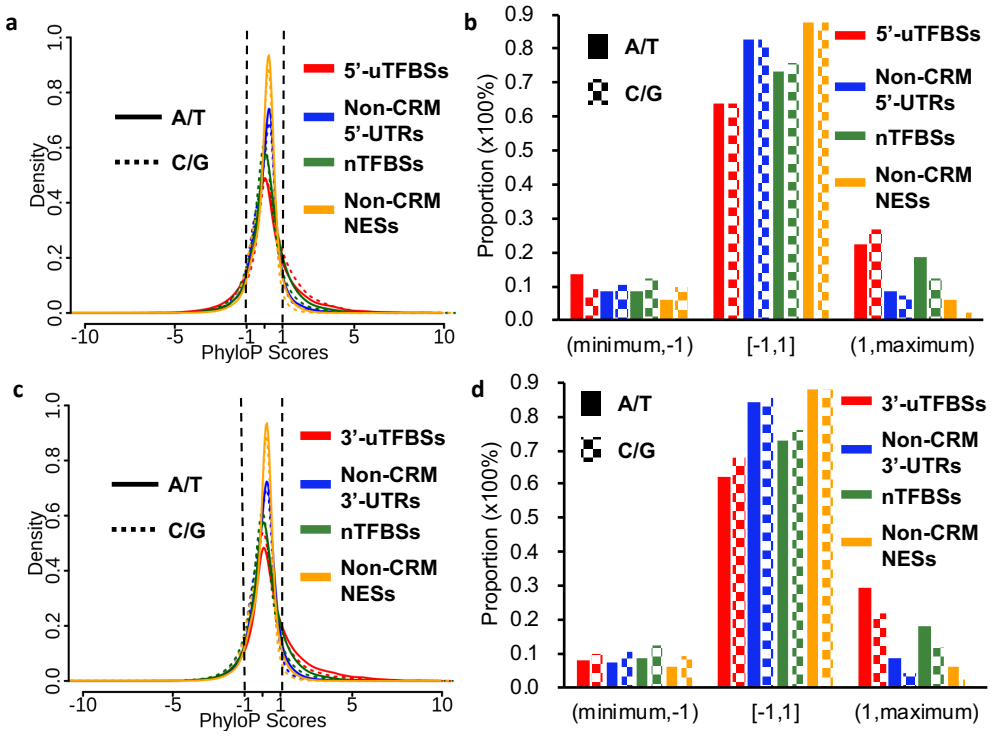

Figure 4.6: Comparison of phyloP scores of the predicted uTFBSs, non-CRM UTRs, nTFBSs and non-CRM NESs.

Distributions of phyloP scores of A/T and C/G sites in 5'-uTFBSs (a), 3'-uTFBSs (b) and their counterparts in non-CRMs, in comparison with those in nTFBSs and non-CRM NESs. Proportions of the sites that are under positive selection (phyloP scores in (minimum, -1)), purifying selection (phyloP scores in (1, maximum)), or selectively neutral (phyloP scores in [-1, 1]), in 5'-UTRs (c) and 3'-UTRs (d) and their counterparts in non-CRMs, in comparison with those in nTFBSs and non-CRM NESs. P-values were computed using Kolmogrov-Smirnov test for the density

are 57.30%, 31.92%, and 10.78% ($p < 2.2x10^{-16}$), respectively, for the host proteins; and 58.20%, 31.44%, and 10.36% ($p < 2.2x10^{-16}$), respectively, for all the proteins with predicted secondary structures (Figure 4.7b). It has been shown that loops are generally less conserved than helices and strands[138], and we see the same results for both known (Figures 4.8a and 4.8b) or predicted (Figures 4.8c and 4.8d) secondary structure types. Since the folding of a core protein is mainly determined by its helix and strand structures, changes in amino acids in the loops are less likely to alter the overall structure and function of the protein. In this regard, it is not surprising that the predicted cTFBSs tend to encode amino acids in loops where purifying selection

are weaker (Figures 4.3a - 4.3d), and therefore they can adopt for specific TF binding without compromising the overall structures of proteins.

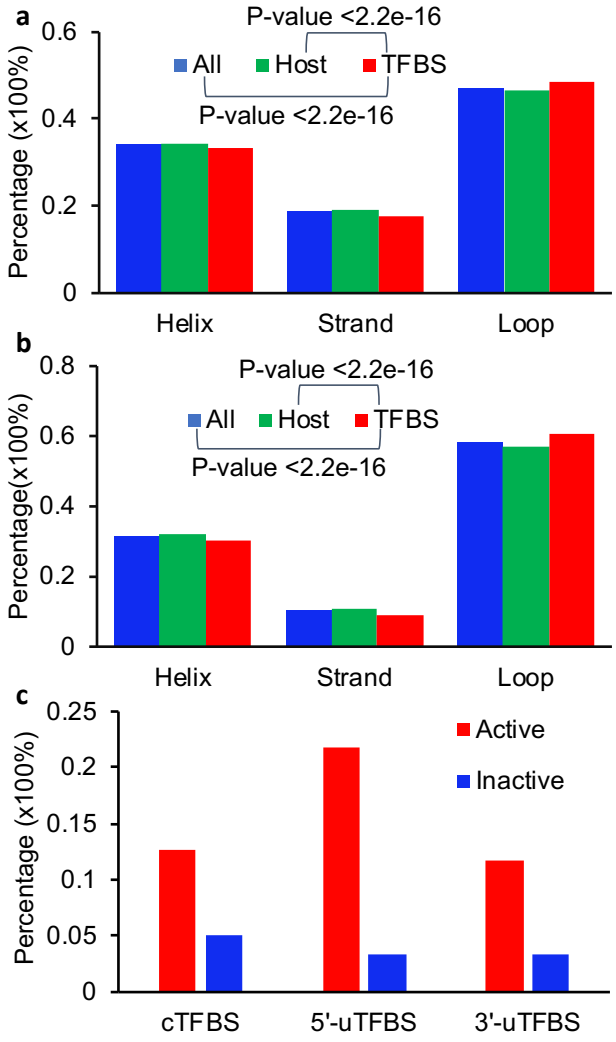

Figure 4.7: Preference of secondary structure types of amino acids encoded by the cTFBSs.

a. Proportions of cTFBS-encoded amino acids mapped to known loops, helices, and strands in comparison with those in the host proteins and in all known protein structures. b. Proportions of cTFBS-encoded amino acids mapped to predicted loops, helices, and strands in comparison with those in the host proteins and in all predicted secondary structures. c. Enrichment of putative active eTFBSs for close physical proximity with distal promoters in the K562 cells comparison with putative inactive eTFBSs. P-values were computed using χ^2 test for the proportion plots.

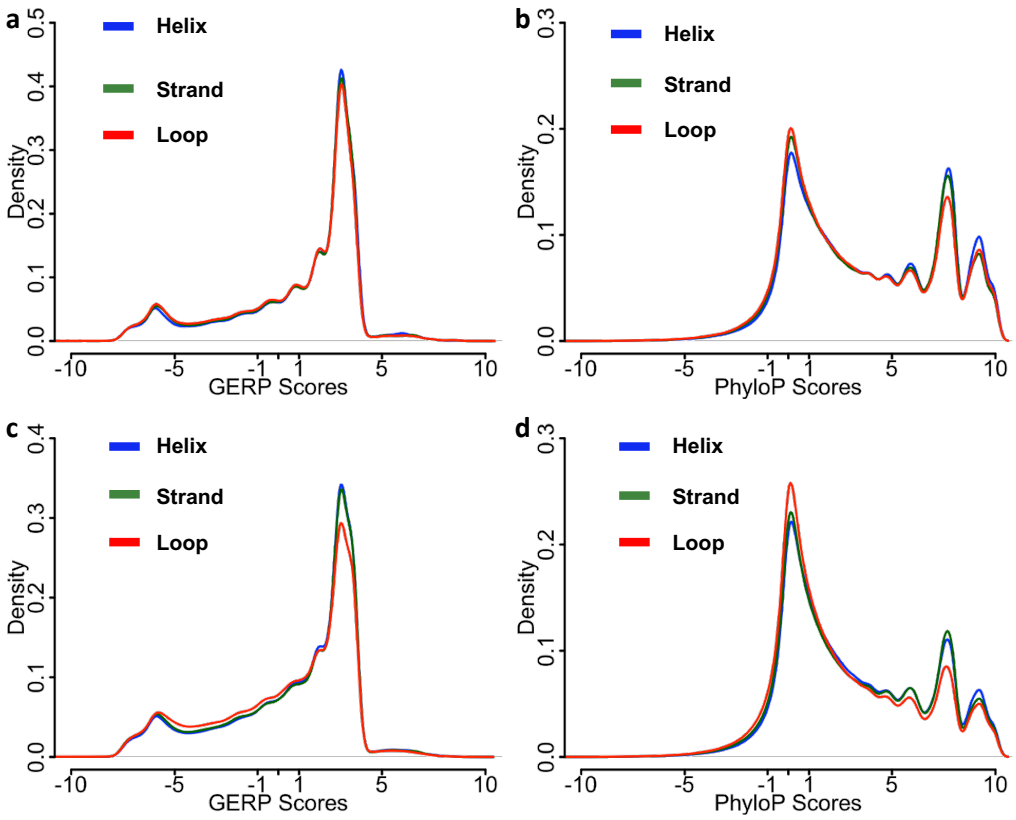

Figure 4.8: Comparison of conservation scores of secondary structure types of amino acids encoded by the cTFBSs.

Distributions of GERP (a) and phyloP (b) scores of cTFBSs encoding amino acids mapped to known loops, helices, and strands in comparison with those in the host proteins and in all known protein structures. Distributions of GERP (c) and phyloP (d) scores of cTFBSs encoding amino acids mapped to predicted loops, helices, and strands in comparison with those in the host proteins and in all predicted secondary structures. P-values were computed using Kolmogorov-Smirnov test

4.3.6 eTFBSs on chromatin structures

Using chromatin conformation capture techniques[170] such as HiC[171], it is now well established that the linear genomic DNA is folded into highly conserved topologically associating domains (TADs) in the nucleus, where enhancers interact with promoters via looping over long distances for transcriptional regulation [172, 173]. Therefore, we hypothesize that eTFBSs must be in close physical proximity to the promoters of target genes in TADs to carry out their transcriptional regulatory functions, although they may be linearly far away. To test this hypothesis, we used a

paired-end tag sequencing (ChIA-PED) dataset generated in K562 cells using an antibody against hypomethylated Pol II at Ser2 [174]. ChIA-PED is a variant of HiC, which combines ChIP-seq with HiC to identify DNA loci that are physically close to sequences bound by the antibody-targeted protein [174]. Pol II hypomethylated at Ser2 stalls at promoters in the preinitiation complex [175], thus the resulting ChIA-PED data are enriched for reads from loci in close physical proximity to promoters. As an eTFBS tends to be close to the promoter of its own gene, we only consider a promoter that is at least 5,000pb away from an eTFBS. Of the predicted non-redundant 1,047,183 cTFBSs, 129,518 5'-uTFBSs and 733,089 3'-uTFBSs, 606,067, 112,981 and 389,627, respectively, overlap with at least one of the TF ChIP-seq binding peak in K562 cells (Materials and Methods), which are presumably active in the cells; while the remaining cTFBSs, 5'-uTFBSs and 3'-uTFBSs are presumably inactive in the cells. As shown in Figure 4.7c, these putative active cTFBSs (77,190 out of 606,067, 12.74%), 5'-uTFBSs (24,601 out of 112,981, 21.77%) and 3'-uTFBSs (45,513 out of 389,627, 11.68%) are all highly enriched for close physical proximity to distal promoters compared with putative inactive cTFBSs (22,572 out of 441,116, 5.12%, $p < 2.2\text{x}10^{-16}$), 5'-uTFBSs (556 out of 16,537, 3.36%, $p < 2.2\text{x}10^{-16}$) and 3'-uTFBSs (11,247 out of 343,471, 3.27%, $p < 2.2\text{x}10^{-16}$). These results unequivocally demonstrate that the predicted active eTFBSs indeed tend to be in close physical proximity to distal promoters, consistent with their possible roles as TFBSs in CRMs.

4.4 DISCUSSION

It has long been known that in addition to encoding amino acids, CDSs can also code for other information, including splicing enhancers [176], overlapping ORFs [177], lincRNA [178] and transcriptional enhancers [69, 88, 145, 146, 147, 148, 149]. It has been estimated that up to 25% of codons in the human genome may code for such overlapping functions based on selection constraints on the degenerate sites in synonymous codons [179]. Although there are numerous experimentally verified cases

for each of these dual-uses of codons, it is under hot debate what selection constraints these codons have undertaken in the course of evolution [70, 176, 180, 181], how prevalent they are in the human genome [70, 182], and how it is possible for a sequence to evolve for two unrelated functions. For instance, Stergachis and colleagues showed that duons predicted using DHSs [70] are under strong purifying selection and up to 15% codons are in dual-use, while others [89, 90] conclude that they are largely selectively neutral. The discrepancies may result from different methods employed [70, 89, 176] and/or different datasets used [70, 90, 176, 182]. The lack of a large high quality positive and negative datasets for cEHs has hampered clarifying these contradictions and addressing related issues of the dual-use of exons. Our recent prediction of large sets of CRMs and non-CRMs from 77.5% of the human genome provides us an opportunity to address these issues.

Using the TFBSs in the CRMs that are located in experimentally verified exons, we found that non-degenerate sites at both the first and second codon positions in the cTFBSs are more likely to be under purifying selection and less likely to be evolutionarily neutral (Figures 4.3a - 4.3d, 4.4 a - 4.4d) than those in non-CRM CDSs. Moreover, A/T at degenerate sites at the third codon positions are more likely be to under strong positive selection, while C/G are more likely to be under either purifying selection or positive selection than those in non-CRM CDSs (Figures 4.3e and 4.3f, 4.4e and 4.4f). Thus, cTFBSs in general are under more evolutionary constraints (either positive selection or purifying selection) than non-CRM CDSs. In this sense, cTFBSs evolve similarly to nTFBSs, although the former tend to be under stronger purifying selection than the latter (Figures 4.3 and 4.4), suggesting that at least some non-degenerate sites as well degenerate sites are in dual-use. For the degenerate sites, the dual-use might be for TFBSs and other non-amino acid-coding functions such as splicing enhancers [181]. This conclusion therefore is different from two earlier contradictory ones by Stergachis *et al.* [70] and Xing and He [89] using

the same set of duons derived from DHSs. We (in submission) and others have shown that enhancers predicted based solely on DHSs have high FPRs [155, 156, 157, 158]. Therefore, it is highly likely that a high FPR of the predicted duons might account for the discrepancy [70, 89]. In addition, both 5'-uTFBSs and 3'-uTFBSs also evolve very similarly to nTFBSs in that they all are more likely to be under either positive selection or purifying selection than their counterparts in non-CRMs (Figures 4.5a - 4.5d, 4.6a - 4.6d), although uTFBSs are more likely to be under strong purifying selection than nTFBSs (Figures 4.5a and 4.5b, 4.6a and 4.6b), suggesting that at least some uTFBSs might be in dual-use.

The observation that C/G contents in degenerate third codon positions, 5'-UTRs, 3'-UTRs and nTFBSs are higher than neutral expectation (41%) suggests that G/C-biased gene conversion (GCBC) [183, 184] rate at these sites might be higher than expected. This might be related to the functions of these sites that are more often to be nucleosome-free than expected. It is well documented that nucleosome-free state is related to a high mutation rate and more GCBC events, and thus higher C/G contents [183, 184]. Interestingly, C/G contents in degenerate sites in cTFBSs, 5'-uTFBSs and nTFBSs are even higher than those in their counterparts in non-CRMs, which is consistent with their roles as TFBSs that might be even more often nucleosome-free than their counterparts in non-CRMs. In contrast, 3'-uTFBS sites have lower C/G contents than non-CRM 3'-UTR sites, suggesting that these non-CRM 3'-UTR sites that tend to be located in the middle of 3'-UTRs (Figure 4.1e) might be more likely constrained for encoding higher C/G-rich binding sites of RBPs [185, 186, 187] and miRNA response elements [188], or they are more likely to be nucleosome-free for unknown reasons.

The paradox concerning the dual-use of exons is how a DNA sequence could evolve two unrelated functions such as TF binding and encoding amino acids as well as UTR functions, because, for instance, the amino acid coding requires the DNA sequence to

specify amino acids that plays a role in the protein's function and at the same time, the TF binding demands it to adopt an interface that is specifically bound by a TF. Our findings might provide an answer to the puzzle. For cTFBSs, as a considerable proportion of even non-degenerate sites at the first and second codon positions in non-CRM CDSs are selectively neutral or nearly so (Figures 4.3a - 4.3d, 4.4 a - 4.4d), hence, they are potentially allowed to evolve into cTFBSs without detrimental effects. Interestingly, the proportion of selectively neutral non-degenerate sites is smaller in cTFBSs than that in non-CRM CDSs (Figures 4.3a - 4.3d, 4.4 a - 4.4d), suggesting that the number of such neutral sites is limited in a protein as expected, and once they become cTFBSs, they are under purifying selection. This might explain why non-degenerate sites in cTFBSs tend to be more conserved than those in non-CRM CDSs (Figures 4.3a - 4.3d, 4.4 a - 4.4d).

Moreover, amino acids encoded by cTFBSs tend to be located in structurally and functional less critical loops, and avoid structurally and functional more important helices and strands (Figures 4.7a and 4.7b), reducing detrimental effects of evolving codons into cTFBS sites. The preferential location of cTFBSs at 5'-ends and 3'-ends of CDSs (Figure 4.1e) suggests that cTFBSs tend to encode at N-termini and C-termini of proteins. These termini probably have less crucial functions. Therefore, it seems that while dual-use of some cTFBSs is possible, nature attempts to avoid it. The strongly purifying selection on the non-degenerate sites and either strong purifying selection or strong positive selection on the degenerate sites in cTFBSs might suggest a scenario of how a CDS evolves into a cTFBS: nature chooses a codon whose non-degenerate sites match critical positions in a desired TFBS, and the non-degenerate site either mutates to the desired nucleotide or remain the same if it matches the desired site. For 5'-uTFBSs, they tend to be located at the middle and avoid the two ends of 5'-UTRs (Figure 4.1e), where 5'-UTR function-related sequences are encoded, such as transcription start sites (TSSs), ribosome entry sites and upstream

open-reading frames [168, 189]. For 3'-uTFBSs, they tend to be located at the two ends and avoid the middle of 3'-UTRs (Figure 4.1e), where 3'-UTR function-related sequences might be coded, such as polyadenylation sites, miRNA response elements and RBP binding sites [169]. Therefore, as in the case of cTFBSs, it seems that while dual-use of some UTRs is possible, nature attempts to avoid it.

We found that cTFBSs, 5'-uTFBSs and 3'-uTFBSs comprise 33.85%, 48.78% and 32.48% of the total length of annotated CDSs, 5'-UTRs, and 3'-UTRs, respectively (Figure 4.1c). And when the entire CRMs overlapping exons are considered, about 80% of the total length of exons are covered by predicted CRMs. Therefore, eTFBSs and eEHs might be more prevalent than originally thought. Then the question is, when only 4.1% of the human genome code for exons [190] and the remaining 95.9% are NESs that can be potentially used to encode enhancers, why are exons exploited to encode TFBSs in enhancers? Our finding that all active cTFBSs, 5'-uTFBSs and 3'-uTFBSs tend to be in close physical proximity to distal promoters in TADs may provide an explanation. When chromatins are folded into conserved 3D structures [191], so that transcriptionally related genes, promoters, and enhancers are brought in close physical proximity in compartments such as TADs [172, 173, 192], it is highly likely that there is no NESs in close proximity to a promoter to function as its enhancer duo to space constraints [193]. In such a scenario, a few nucleotides in a CDS or UTR in proximity to a promoter, which codes for less important amino acids such as those in some loops, or a less critical part of the UTR, may well likely be opted for cTFBSs. In this regard, it seems that dual-use of some exons is unavoidable, and nature chooses less critical exons for eTFBSs, thereby avoiding the dilemma of evolving a sequence for two unrelated functions.

REFERENCES

[1] D. M. Church, V. A. Schneider, T. Graves, K. Auger, F. Cunningham, N. Bouk, H.-C. Chen, R. Agarwala, W. M. McLaren, G. R. Ritchie, *et al.*, "Modernizing reference genome assemblies," *PLoS Biol*, vol. 9, no. 7, p. e1001091, 2011.

[2] I. H. G. S. Consortium *et al.*, "Finishing the euchromatic sequence of the human genome," *Nature*, vol. 431, no. 7011, p. 931, 2004.

[3] V. A. Schneider, T. Graves-Lindsay, K. Howe, N. Bouk, H.-C. Chen, P. A. Kitts, T. D. Murphy, K. D. Pruitt, F. Thibaud-Nissen, D. Albracht, *et al.*, "Evaluation of grch38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly," *Genome research*, vol. 27, no. 5, pp. 849–864, 2017.

[4] B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz, and K. W. Kinzler, "Cancer genome landscapes," *science*, vol. 339, no. 6127, pp. 1546–1558, 2013.

[5] . G. P. Consortium *et al.*, "A global reference for human genetic variation," *Nature*, vol. 526, no. 7571, pp. 68–74, 2015.

[6] . G. P. Consortium *et al.*, "A map of human genome variation from population-scale sequencing," *Nature*, vol. 467, no. 7319, p. 1061, 2010.

[7] R. E. Mills, C. T. Luttig, C. E. Larkins, A. Beauchamp, C. Tsui, W. S. Pittard, and S. E. Devine, "An initial map of insertion and deletion (indel) variation in the human genome," *Genome research*, vol. 16, no. 9, pp. 1182–1190, 2006.

[8] R. E. Mills, W. S. Pittard, J. M. Mullaney, U. Farooq, T. H. Creasy, A. A. Mahurkar, D. M. Kemeza, D. S. Strassler, C. P. Ponting, C. Webber, *et al.*, "Natural genetic variation caused by small insertions and deletions in the human genome," *Genome research*, vol. 21, no. 6, pp. 830–839, 2011.

[9] R. A. Gibbs, J. W. Belmont, P. Hardenbol, T. D. Willis, F. Yu, H. Yang, L.-Y. Ch'ang, W. Huang, B. Liu, Y. Shen, *et al.*, "The international hapmap project," 2003.

[10] R. Sachidanandam, D. Weissman, S. C. Schmidt, J. M. Kakol, L. D. Stein, G. Marth, S. Sherry, J. C. Mullikin, B. J. Mortimore, D. L. Willey, *et al.*, "A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms," *Nature*, vol. 409, no. 6822, pp. 928–934, 2001.

[11] I. H. Consortium *et al.*, "A haplotype map of the human genome," *Nature*, vol. 437, no. 7063, p. 1299, 2005.

[12] J. M. Mullaney, R. E. Mills, W. S. Pittard, and S. E. Devine, "Small insertions and deletions (indels) in human genomes," *Human molecular genetics*, vol. 19, no. R2, pp. R131–R136, 2010.

[13] H. J. Abel and E. J. Duncavage, "Detection of structural dna variation from next generation sequencing data: a review of informatic approaches," *Cancer genetics*, vol. 206, no. 12, pp. 432–440, 2013.

[14] A. J. Iafrate, L. Feuk, M. N. Rivera, M. L. Listewnik, P. K. Donahoe, Y. Qi, S. W. Scherer, and C. Lee, "Detection of large-scale variation in the human genome," *Nature genetics*, vol. 36, no. 9, pp. 949–951, 2004.

[15] J. Sebat, B. Lakshmi, J. Troge, J. Alexander, J. Young, P. Lundin, S. Månér, H. Massa, M. Walker, M. Chi, *et al.*, "Large-scale copy number polymorphism in the human genome," *Science*, vol. 305, no. 5683, pp. 525–528, 2004.

[16] J. M. Kidd, G. M. Cooper, W. F. Donahue, H. S. Hayden, N. Sampas, T. Graves, N. Hansen, B. Teague, C. Alkan, F. Antonacci, *et al.*, "Mapping and sequencing of structural variation from eight human genomes," *Nature*, vol. 453, no. 7191, pp. 56–64, 2008.

[17] P. H. Sudmant, T. Rausch, E. J. Gardner, R. E. Handsaker, A. Abyzov, J. Huddleston, Y. Zhang, K. Ye, G. Jun, M. H.-Y. Fritz, *et al.*, "An integrated map of structural variation in 2,504 human genomes," *Nature*, vol. 526, no. 7571, pp. 75–81, 2015.

[18] . G. P. Consortium *et al.*, "An integrated map of genetic variation from 1,092 human genomes," *Nature*, vol. 491, no. 7422, pp. 56–65, 2012.

[19] S. T. Sherry, M.-H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin, "dbsnp: the ncbi database of genetic variation," *Nucleic acids research*, vol. 29, no. 1, pp. 308–311, 2001.

[20] I. Lappalainen, J. Lopez, L. Skipper, T. Hefferon, J. D. Spalding, J. Garner, C. Chen, M. Maguire, M. Corbett, G. Zhou, *et al.*, "Dbvar and dgva: public archives for genomic structural variation," *Nucleic acids research*, vol. 41, no. D1, pp. D936–D941, 2012.

[21] M. J. Landrum, J. M. Lee, M. Benson, G. Brown, C. Chao, S. Chitipiralla, B. Gu, J. Hart, D. Hoffman, J. Hoover, *et al.*, "Clinvar: public archive of interpretations of clinically relevant variants," *Nucleic acids research*, vol. 44, no. D1, pp. D862–D868, 2016.

[22] P. C. Nowell, "The clonal evolution of tumor cell populations," *Science*, vol. 194, no. 4260, pp. 23–28, 1976.

[23] E. R. Fearon and B. Vogelstein, "A genetic model for colorectal tumorigenesis," *cell*, vol. 61, no. 5, pp. 759–767, 1990.

[24] S. Masciari, D. A. Dillon, M. Rath, M. Robson, J. N. Weitzel, J. Balmana, S. B. Gruber, J. M. Ford, D. Euhus, A. Lebensohn, *et al.*, "Breast cancer phenotype in women with tp53 germline mutations: a li-fraumeni syndrome consortium

effort," *Breast cancer research and treatment*, vol. 133, no. 3, pp. 1125–1130, 2012.

[25] S. Damineni, V. R. Rao, S. Kumar, R. R. Ravuri, S. Kagitha, N. R. Dunna, R. Digumarthi, and V. Satti, "Germline mutations of tp53 gene in breast cancer," *Tumor Biology*, vol. 35, no. 9, pp. 9219–9227, 2014.

[26] C. G. A. R. Network *et al.*, "Comprehensive molecular profiling of lung adenocarcinoma," *Nature*, vol. 511, no. 7511, pp. 543–550, 2014.

[27] L. Chen, F. Rashid, A. Shah, H. M. Awan, M. Wu, A. Liu, J. Wang, T. Zhu, Z. Luo, and G. Shan, "The isolation of an rna aptamer targeting to p53 protein with single amino acid mutation," *Proceedings of the National Academy of Sciences*, vol. 112, no. 32, pp. 10002–10007, 2015.

[28] S. A. Forbes, D. Beare, H. Boutselakis, S. Bamford, N. Bindal, J. Tate, C. G. Cole, S. Ward, E. Dawson, L. Ponting, *et al.*, "Cosmic: somatic cancer genetics at high-resolution," *Nucleic acids research*, vol. 45, no. D1, pp. D777–D783, 2017.

[29] F. S. Collins, M. L. Drumm, J. L. Cole, W. K. Lockwood, G. V. Woude, and M. C. Iannuzzi, "Construction of a general human chromosome jumping library, with application to cystic fibrosis," *Science*, vol. 235, no. 4792, pp. 1046–1049, 1987.

[30] J. L. Weber, D. David, J. Heil, Y. Fan, C. Zhao, and G. Marth, "Human diallelic insertion/deletion polymorphisms," *The American Journal of Human Genetics*, vol. 71, no. 4, pp. 854–862, 2002.

[31] T. R. Bhangale, M. J. Rieder, R. J. Livingston, and D. A. Nickerson, "Comprehensive identification and characterization of diallelic insertion–deletion polymorphisms in 330 human candidate genes," *Human molecular genetics*, vol. 14, no. 1, pp. 59–69, 2005.

[32] M. Lin, S. Whitmire, J. Chen, A. Farrel, X. Shi, and J.-t. Guo, "Effects of short indels on protein structure and function in human genomes," *Scientific reports*, vol. 7, no. 1, pp. 1–9, 2017.

[33] K. D. Gonzalez, K. A. Hill, K. Li, W. Li, W. A. Scaringe, J.-C. Wang, D. Gu, and S. S. Sommer, "Somatic microindels: analysis in mouse soma and comparison with the human germline," *Human mutation*, vol. 28, no. 1, pp. 69–80, 2007.

[34] L. Feuk, A. R. Carson, and S. W. Scherer, "Structural variation in the human genome," *Nature Reviews Genetics*, vol. 7, no. 2, pp. 85–97, 2006.

[35] J. Berger, T. Suzuki, K.-A. Senti, J. Stubbs, G. Schaffner, and B. J. Dickson, "Genetic mapping with snp markers in drosophila," *Nature genetics*, vol. 29, no. 4, pp. 475–481, 2001.

[36] S. R. Wicks, R. T. Yeh, W. R. Gish, R. H. Waterston, and R. H. Plasterk, "Rapid gene mapping in caenorhabditis elegans using a high density polymorphism map," *Nature genetics*, vol. 28, no. 2, pp. 160–164, 2001.

[37] E. Dawson, Y. Chen, S. Hunt, L. J. Smink, A. Hunt, K. Rice, S. Livingston, S. Bumpstead, R. Bruskiewich, P. Sham, *et al.*, "A snp resource for human chromosome 22: extracting dense clusters of snps from the genomic sequence," *Genome research*, vol. 11, no. 1, pp. 170–178, 2001.

[38] S. B. Montgomery, D. L. Goode, E. Kvikstad, C. A. Albers, Z. D. Zhang, X. J. Mu, G. Ananda, B. Howie, K. J. Karczewski, K. S. Smith, *et al.*, "The origin, evolution, and functional impact of short insertion–deletion variants identified in 179 human genomes," *Genome research*, vol. 23, no. 5, pp. 749–761, 2013.

[39] N. A. Chuzhanova, E. J. Anassis, E. V. Ball, M. Krawczak, and D. N. Cooper, "Meta-analysis of indels causing human genetic disease: mechanisms of mutagenesis and the role of local dna sequence complexity," *Human mutation*, vol. 21, no. 1, pp. 28–44, 2003.

[40] J. Mullikin, S. Hunt, C. Cole, B. Mortimore, C. Rice, J. Burton, L. Matthews, R. Pavitt, R. Plumb, S. Sims, *et al.*, "An snp map of human chromosome 22," *Nature*, vol. 407, no. 6803, pp. 516–520, 2000.

[41] S. Levy, G. Sutton, P. C. Ng, L. Feuk, A. L. Halpern, B. P. Walenz, N. Axelrod, J. Huang, E. F. Kirkness, G. Denisov, *et al.*, "The diploid genome sequence of an individual human," *PLoS Biol*, vol. 5, no. 10, p. e254, 2007.

[42] J. Wang, W. Wang, R. Li, Y. Li, G. Tian, L. Goodman, W. Fan, J. Zhang, J. Li, J. Zhang, *et al.*, "The diploid genome sequence of an asian individual," *Nature*, vol. 456, no. 7218, pp. 60–65, 2008.

[43] N. de la Chaux, P. W. Messer, and P. F. Arndt, "Dna indels in coding regions reveal selective constraints on protein evolution in the human lineage," *BMC evolutionary biology*, vol. 7, no. 1, p. 191, 2007.

[44] K. Neininger, T. Marschall, and V. Helms, "Snp and indel frequencies at transcription start sites and at canonical and alternative translation initiation sites in the human genome," *PloS one*, vol. 14, no. 4, p. e0214816, 2019.

[45] T. G. Clark, T. Andrew, G. M. Cooper, E. H. Margulies, J. C. Mullikin, and D. J. Balding, "Functional constraint and small insertions and deletions in the encode regions of the human genome," *Genome biology*, vol. 8, no. 9, p. R180, 2007.

[46] K. A. Pagel, D. Antaki, A. Lian, M. Mort, D. N. Cooper, J. Sebat, L. M. Iakoucheva, S. D. Mooney, and P. Radivojac, "Pathogenicity and functional impact of non-frameshifting insertion/deletion variation in the human genome," *PLoS computational biology*, vol. 15, no. 6, p. e1007112, 2019.

[47] H. Yang, Y. Zhong, C. Peng, J.-Q. Chen, and D. Tian, "Important role of indels in somatic mutations of human cancer genes," *BMC medical genetics*, vol. 11, no. 1, p. 128, 2010.

[48] M. Imielinski, G. Guo, and M. Meyerson, "Insertions and deletions target lineage-defining genes in human cancers," *Cell*, vol. 168, no. 3, pp. 460–472, 2017.

[49] T. Nakagomi, Y. Hirotsu, T. Goto, D. Shikata, Y. Yokoyama, R. Higuchi, S. Otake, K. Amemiya, T. Oyama, H. Mochizuki, *et al.*, "Clinical implications of noncoding indels in the surfactant-encoding genes in lung cancer," *Cancers*, vol. 11, no. 4, p. 552, 2019.

[50] R. Li, Y. Li, K. Kristiansen, and J. Wang, "Soap: short oligonucleotide alignment program," *Bioinformatics*, vol. 24, no. 5, pp. 713–714, 2008.

[51] H. Li, J. Ruan, and R. Durbin, "Mapping short dna sequencing reads and calling variants using mapping quality scores," *Genome research*, vol. 18, no. 11, pp. 1851–1858, 2008.

[52] H. Li and R. Durbin, "Fast and accurate short read alignment with burrows–wheeler transform," *bioinformatics*, vol. 25, no. 14, pp. 1754–1760, 2009.

[53] K. Ye, M. H. Schulz, Q. Long, R. Apweiler, and Z. Ning, "Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads," *Bioinformatics*, vol. 25, no. 21, pp. 2865–2871, 2009.

[54] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, "Ultrafast and memory-efficient alignment of short dna sequences to the human genome," *Genome biology*, vol. 10, no. 3, p. R25, 2009.

[55] N. Homer, B. Merriman, and S. F. Nelson, "Bfast: an alignment tool for large scale genome resequencing," *PloS one*, vol. 4, no. 11, p. e7767, 2009.

[56] M. S. Hasan, X. Wu, and L. Zhang, "Performance evaluation of indel calling tools using real short-read data," *Human genomics*, vol. 9, no. 1, p. 20, 2015.

[57] M. A. DePristo, E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. Del Angel, M. A. Rivas, M. Hanna, *et al.*, "A framework for variation discovery and genotyping using next-generation dna sequencing data," *Nature genetics*, vol. 43, no. 5, p. 491, 2011.

[58] D. C. Koboldt, K. Chen, T. Wylie, D. E. Larson, M. D. McLellan, E. R. Mardis, G. M. Weinstock, R. K. Wilson, and L. Ding, "Varscan: variant detection in massively parallel sequencing of individual and pooled samples," *Bioinformatics*, vol. 25, no. 17, pp. 2283–2285, 2009.

[59] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin, "The sequence alignment/map format and samtools," *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, 2009.

[60] C. A. Albers, G. Lunter, D. G. MacArthur, G. McVean, W. H. Ouwehand, and R. Durbin, "Dindel: accurate indel calls from short-read data," *Genome research*, vol. 21, no. 6, pp. 961–973, 2011.

[61] R. Poplin, V. Ruano-Rubio, M. A. DePristo, T. J. Fennell, M. O. Carneiro, G. A. Van der Auwera, D. E. Kling, L. D. Gauthier, A. Levy-Moonshine, D. Roazen, *et al.*, "Scaling accurate genetic variant discovery to tens of thousands of samples," *BioRxiv*, p. 201178, 2017.

[62] A. Rimmer, H. Phan, I. Mathieson, Z. Iqbal, S. R. Twigg, A. O. Wilkie, G. McVean, and G. Lunter, "Integrating mapping-, assembly-and haplotype-based approaches for calling variants in clinical sequencing applications," *Nature genetics*, vol. 46, no. 8, pp. 912–918, 2014.

[63] D. Li, W. Kim, L. Wang, K.-A. Yoon, B. Park, C. Park, S.-Y. Kong, Y. Hwang, D. Baek, E. S. Lee, *et al.*, "Comparison of indel calling tools with simulation data and real short-read data," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 16, no. 5, pp. 1635–1644, 2018.

[64] D. C. Koboldt, Q. Zhang, D. E. Larson, D. Shen, M. D. McLellan, L. Lin, C. A. Miller, E. R. Mardis, L. Ding, and R. K. Wilson, "Varscan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing," *Genome research*, vol. 22, no. 3, pp. 568–576, 2012.

[65] H. Fang, E. A. Bergmann, K. Arora, V. Vacic, M. C. Zody, I. Iossifov, J. A. O'Rawe, Y. Wu, L. T. J. Barron, J. Rosenbaum, *et al.*, "Indel variant analysis of short-read sequencing data with scalpel," *Nature protocols*, vol. 11, no. 12, p. 2529, 2016.

[66] G. Jun, M. K. Wing, G. R. Abecasis, and H. M. Kang, "An efficient and scalable analysis framework for variant extraction and refinement from population-scale dna sequence data," *Genome Research*, vol. 25, no. 6, pp. 918–925, 2015.

[67] I. Yevshin, R. Sharipov, T. Valeev, A. Kel, and F. Kolpakov, "Gtrd: a database of transcription factor binding sites identified by chip-seq experiments," *Nucleic acids research*, p. gkw951, 2016.

[68] B. Lenhard and W. W. Wasserman, "Tfbs: Computational framework for transcription factor binding site analysis," *Bioinformatics*, vol. 18, no. 8, pp. 1135–1136, 2002.

[69] N. Hirsch and R. Y. Birnbaum, "Dual function of dna sequences: protein-coding sequences function as transcriptional enhancers," *Perspectives in biology and medicine*, vol. 58, no. 2, pp. 182–195, 2015.

[70] A. B. Stergachis, E. Haugen, A. Shafer, W. Fu, B. Vernot, A. Reynolds, A. Raubitschek, S. Ziegler, E. M. LeProust, J. M. Akey, *et al.*, "Exonic transcription factor binding directs codon choice and affects protein evolution," *Science*, vol. 342, no. 6164, pp. 1367–1372, 2013.

[71] S. B.-T. de Leon and E. H. Davidson, "Gene regulation: gene control network in development," *Annu. Rev. Biophys. Biomol. Struct.*, vol. 36, pp. 191–212, 2007.

[72] D. M. Jeziorska, K. W. Jordan, and K. W. Vance, "A systems biology approach to understanding cis-regulatory module function," in *Seminars in cell & developmental biology*, vol. 20, pp. 856–862, Elsevier, 2009.

[73] C. A. Davis, B. C. Hitz, C. A. Sloan, E. T. Chan, J. M. Davidson, I. Gabdank, J. A. Hilton, K. Jain, U. K. Baymuradov, A. K. Narayanan, *et al.*, "The encyclopedia of dna elements (encode): data portal update," *Nucleic acids research*, vol. 46, no. D1, pp. D794–D801, 2018.

[74] J. Lonsdale, J. Thomas, M. Salvatore, R. Phillips, E. Lo, S. Shad, R. Hasz, G. Walters, F. Garcia, N. Young, *et al.*, "The genotype-tissue expression (gtex) project," *Nature genetics*, vol. 45, no. 6, pp. 580–585, 2013.

[75] A. Visel, S. Minovitsky, I. Dubchak, and L. A. Pennacchio, "Vista enhancer browserâa database of tissue-specific human enhancers," *Nucleic acids research*, vol. 35, no. suppl_1, pp. D88–D92, 2007.

[76] M. Niu, E. Tabari, P. Ni, and Z. Su, "Towards a map of cis-regulatory sequences in the human genome," *Nucleic acids research*, vol. 46, no. 11, pp. 5395–5409, 2018.

[77] S. Fishilevich, R. Nudel, N. Rappaport, R. Hadar, I. Plaschkes, T. Iny Stein, N. Rosen, A. Kohn, M. Twik, M. Safran, *et al.*, "Genehancer: genome-wide integration of enhancers and target genes in genecards," *Database*, vol. 2017, 2017.

[78] T. Gao, B. He, S. Liu, H. Zhu, K. Tan, and J. Qian, "Enhanceratlas: a resource for enhancer annotation and analysis in 105 human cell/tissue types," *Bioinformatics*, vol. 32, no. 23, pp. 3543–3551, 2016.

[79] T. Gao and J. Qian, "Enhanceratlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species," *Nucleic acids research*, vol. 48, no. D1, pp. D58–D64, 2020.

[80] A. Visel, M. J. Blow, Z. Li, T. Zhang, J. A. Akiyama, A. Holt, I. Plajzer-Frick, M. Shoukry, C. Wright, F. Chen, *et al.*, "Chip-seq accurately predicts tissue-specific activity of enhancers," *Nature*, vol. 457, no. 7231, pp. 854–858, 2009.

[81] T.-K. Kim, M. Hemberg, J. M. Gray, A. M. Costa, D. M. Bear, J. Wu, D. A. Harmin, M. Laptewicz, K. Barbara-Haley, S. Kuersten, *et al.*, "Widespread transcription at neuronal activity-regulated enhancers," *Nature*, vol. 465, no. 7295, pp. 182–187, 2010.

[82] R. E. Thurman, E. Rynes, R. Humbert, J. Vierstra, M. T. Maurano, E. Haugen, N. C. Sheffield, A. B. Stergachis, H. Wang, B. Vernot, *et al.*, "The accessible chromatin landscape of the human genome," *Nature*, vol. 489, no. 7414, pp. 75–82, 2012.

[83] M. J. Fullwood, M. H. Liu, Y. F. Pan, J. Liu, H. Xu, Y. B. Mohamed, Y. L. Orlov, S. Velkov, A. Ho, P. H. Mei, *et al.*, "An oestrogen-receptor-α-bound human chromatin interactome," *Nature*, vol. 462, no. 7269, pp. 58–64, 2009.

[84] K. J. Gaulton, T. Nammo, L. Pasquali, J. M. Simon, P. G. Giresi, M. P. Fogarty, T. M. Panhuis, P. Mieczkowski, A. Secchi, D. Bosco, *et al.*, "A map of open chromatin in human pancreatic islets," *Nature genetics*, vol. 42, no. 3, p. 255, 2010.

[85] J. Cotney, J. Leng, S. Oh, L. E. DeMare, S. K. Reilly, M. B. Gerstein, and J. P. Noonan, "Chromatin state signatures associated with tissue-specific gene expression and enhancer activity in the embryonic limb," *Genome research*, vol. 22, no. 6, pp. 1069–1080, 2012.

[86] R. Andersson, C. Gebhard, I. Miguel-Escalada, I. Hoof, J. Bornholdt, M. Boyd, Y. Chen, X. Zhao, C. Schmidl, T. Suzuki, *et al.*, "An atlas of active enhancers across human cell types and tissues," *Nature*, vol. 507, no. 7493, pp. 455–461, 2014.

[87] R. Y. Birnbaum, R. P. Patwardhan, M. J. Kim, G. M. Findlay, B. Martin, J. Zhao, R. J. Bell, R. P. Smith, A. A. Ku, J. Shendure, *et al.*, "Systematic dissection of coding exons at single nucleotide resolution supports an additional role in cell-specific transcriptional regulation," *PLoS Genet*, vol. 10, no. 10, p. e1004592, 2014.

[88] R. Y. Birnbaum, E. J. Clowney, O. Agamy, M. J. Kim, J. Zhao, T. Yamanaka, Z. Pappalardo, S. L. Clarke, A. M. Wenger, L. Nguyen, *et al.*, "Coding exons function as tissue-specific enhancers of nearby genes," *Genome research*, vol. 22, no. 6, pp. 1059–1068, 2012.

[89] K. Xing and X. He, "Reassessing the âduonâ hypothesis of protein evolution," *Molecular biology and evolution*, vol. 32, no. 4, pp. 1056–1062, 2015.

[90] R. M. Agoglia and H. B. Fraser, "Disentangling sources of selection on exonic transcriptional enhancers," *Molecular biology and evolution*, vol. 33, no. 2, pp. 585–590, 2016.

[91] M. Baker, "Structural variation: the genome's hidden architecture," *Nature methods*, vol. 9, no. 2, pp. 133–137, 2012.

[92] D. L. Cameron, L. Di Stefano, and A. T. Papenfuss, "Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software," *Nature communications*, vol. 10, no. 1, pp. 1–11, 2019.

[93] N. A. O'Leary, M. W. Wright, J. R. Brister, S. Ciufo, D. Haddad, R. McVeigh, B. Rajput, B. Robbertse, B. Smith-White, D. Ako-Adjei, *et al.*, "Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation," *Nucleic acids research*, vol. 44, no. D1, pp. D733–D745, 2016.

[94] P. J. Stephens, P. S. Tarpey, H. Davies, P. Van Loo, C. Greenman, D. C. Wedge, S. Nik-Zainal, S. Martin, I. Varela, G. R. Bignell, *et al.*, "The landscape of cancer genes and mutational processes in breast cancer," *Nature*, vol. 486, no. 7403, pp. 400–404, 2012.

[95] B. Niu, A. D. Scott, S. Sengupta, M. H. Bailey, P. Batra, J. Ning, M. A. Wyczalkowski, W.-W. Liang, Q. Zhang, M. D. McLellan, *et al.*, "Protein-structure-guided discovery of functional mutations across 19 cancer types," *Nature genetics*, vol. 48, no. 8, pp. 827–837, 2016.

[96] S. L. Carter, K. Cibulskis, E. Helman, A. McKenna, H. Shen, T. Zack, P. W. Laird, R. C. Onofrio, W. Winckler, B. A. Weir, *et al.*, "Absolute quantification of somatic dna alterations in human cancer," *Nature biotechnology*, vol. 30, no. 5, pp. 413–421, 2012.

[97] X. Liu, J. Wang, and L. Chen, "Whole-exome sequencing reveals recurrent somatic mutation networks in cancer," *Cancer letters*, vol. 340, no. 2, pp. 270–276, 2013.

[98] K. Cibulskis, M. S. Lawrence, S. L. Carter, A. Sivachenko, D. Jaffe, C. Sougnez, S. Gabriel, M. Meyerson, E. S. Lander, and G. Getz, "Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples," *Nature biotechnology*, vol. 31, no. 3, pp. 213–219, 2013.

[99] C. T. Saunders, W. S. Wong, S. Swamy, J. Becq, L. J. Murray, and R. K. Cheetham, "Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs," *Bioinformatics*, vol. 28, no. 14, pp. 1811–1817, 2012.

[100] S. Kim, K. Scheffler, A. L. Halpern, M. A. Bekritsky, E. Noh, M. Källberg, X. Chen, Y. Kim, D. Beyter, P. Krusche, *et al.*, "Strelka2: fast and accurate calling of germline and somatic variants," *Nature methods*, vol. 15, no. 8, pp. 591–594, 2018.

[101] D. Benjamin, T. Sato, K. Cibulskis, G. Getz, C. Stewart, and L. Lichtenstein, "Calling somatic snvs and indels with mutect2," *BioRxiv*, p. 861054, 2019.

[102] A. B. Krøigård, M. Thomassen, A.-V. Lænkholm, T. A. Kruse, and M. J. Larsen, "Evaluation of nine somatic variant callers for detection of somatic mutations in exome and targeted deep sequencing data," *PloS one*, vol. 11, no. 3, p. e0151664, 2016.

[103] C. Xu, "A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data," *Computational and structural biotechnology journal*, vol. 16, pp. 15–24, 2018.

[104] L. Cai, W. Yuan, Z. Zhang, L. He, and K.-C. Chou, "In-depth comparison of somatic point mutation callers based on different tumor next-generation sequencing depth data," *Scientific reports*, vol. 6, p. 36540, 2016.

[105] N. D. Roberts, R. D. Kortschak, W. T. Parker, A. W. Schreiber, S. Branford, H. S. Scott, G. Glonek, and D. L. Adelson, "A comparative analysis of algorithms for somatic snv detection in cancer," *Bioinformatics*, vol. 29, no. 18, pp. 2223–2230, 2013.

[106] D. H. Spencer, M. Tyagi, F. Vallania, A. J. Bredemeyer, J. D. Pfeifer, R. D. Mitra, and E. J. Duncavage, "Performance of common analysis methods for detecting low-frequency single nucleotide variants in targeted next-generation sequence data," *The Journal of Molecular Diagnostics*, vol. 16, no. 1, pp. 75–88, 2014.

[107] Q. Wang, P. Jia, F. Li, H. Chen, H. Ji, D. Hucks, K. B. Dahlman, W. Pao, and Z. Zhao, "Detecting somatic point mutations in cancer genome sequencing data: a comparison of mutation callers," *Genome medicine*, vol. 5, no. 10, p. 91, 2013.

[108] "Gatk resource bundle." https://storage.cloud.google.com/genomics-public-data/resources/broad/hg38/v0/Mills_and_1000G_gold_standard.indels.hg38.vcf.gz.

[109] H. Zhao, Y. Yang, H. Lin, X. Zhang, M. Mort, D. N. Cooper, Y. Liu, and Y. Zhou, "Ddig-in: discriminating between disease-associated and neutral non-frameshifting micro-indels," *Genome biology*, vol. 14, no. 3, p. R23, 2013.

[110] L. Folkman, Y. Yang, Z. Li, B. Stantic, A. Sattar, M. Mort, D. N. Cooper, Y. Liu, and Y. Zhou, "Ddig-in: detecting disease-causing genetic variations due to frameshifting indels and nonsense mutations employing sequence and structural properties at nucleotide and protein levels," *Bioinformatics*, vol. 31, no. 10, pp. 1599–1606, 2015.

[111] M. Ferlaino, M. F. Rogers, H. A. Shihab, M. Mort, D. N. Cooper, T. R. Gaunt, and C. Campbell, "An integrative approach to predicting the functional effects of small indels in non-coding regions of the human genome," *BMC bioinformatics*, vol. 18, no. 1, pp. 1–8, 2017.

[112] P. Krawitz, C. Rödelsperger, M. Jäger, L. Jostins, S. Bauer, and P. N. Robinson, "Microindel detection in short-read sequence data," *Bioinformatics*, vol. 26, no. 6, pp. 722–729, 2010.

[113] S. Brogna and J. Wen, "Nonsense-mediated mrna decay (nmd) mechanisms," *Nature structural & molecular biology*, vol. 16, no. 2, p. 107, 2009.

[114] E. D. Pleasance, R. K. Cheetham, P. J. Stephens, D. J. McBride, S. J. Humphray, C. D. Greenman, I. Varela, M.-L. Lin, G. R. Ordóñez, G. R. Bignell, *et al.*, "A comprehensive catalogue of somatic mutations from a human cancer genome," *Nature*, vol. 463, no. 7278, pp. 191–196, 2010.

[115] S. Nik-Zainal, L. B. Alexandrov, D. C. Wedge, P. Van Loo, C. D. Greenman, K. Raine, D. Jones, J. Hinton, J. Marshall, L. A. Stebbings, *et al.*, "Mutational processes molding the genomes of 21 breast cancers," *Cell*, vol. 149, no. 5, pp. 979–993, 2012.

[116] B. Sathya, A. P. Dharshini, and G. R. Kumar, "Ngs meta data analysis for identification of snp and indel patterns in human airway transcriptome: A preliminary indicator for lung cancer," *Applied & translational genomics*, vol. 4, pp. 4–9, 2015.

[117] S. Turajlic, K. Litchfield, H. Xu, R. Rosenthal, N. McGranahan, J. L. Reading, Y. N. S. Wong, A. Rowan, N. Kanu, M. Al Bakir, *et al.*, "Insertion-and-deletion-derived tumour-specific neoantigens and the immunogenic phenotype: a pan-cancer analysis," *The lancet oncology*, vol. 18, no. 8, pp. 1009–1021, 2017.

[118] I. The, T. P.-C. A. of Whole, G. Consortium, *et al.*, "Pan-cancer analysis of whole genomes," *Nature*, vol. 578, no. 7793, p. 82, 2020.

[119] F. Yang, E. Petsalaki, T. Rolland, D. E. Hill, M. Vidal, and F. P. Roth, "Protein domain-level landscape of cancer-type-specific somatic mutations," *PLoS Comput Biol*, vol. 11, no. 3, p. e1004147, 2015.

[120] T. A. Peterson, I. I. M. Gauran, J. Park, D. Park, and M. G. Kann, "Oncodomains: A protein domain-centric framework for analyzing rare variants in tumor samples," *PLoS computational biology*, vol. 13, no. 4, p. e1005428, 2017.

[121] F. Zhang and J. R. Lupski, "Non-coding genetic variants in human disease," *Human molecular genetics*, vol. 24, no. R1, pp. R102–R110, 2015.

[122] D. E. Arking, A. Pfeufer, W. Post, W. L. Kao, C. Newton-Cheh, M. Ikeda, K. West, C. Kashuk, M. Akyol, S. Perz, *et al.*, "A common genetic variant in the nos1 regulator nos1ap modulates cardiac repolarization," *Nature genetics*, vol. 38, no. 6, pp. 644–651, 2006.

[123] A. Kapoor, R. B. Sekar, N. F. Hansen, K. Fox-Talbot, M. Morley, V. Pihur, S. Chatterjee, J. Brandimarto, C. S. Moravec, S. L. Pulit, *et al.*, "An enhancer

polymorphism at the cardiomyocyte intercalated disc protein nos1ap locus is a major regulator of the qt interval," *The American Journal of Human Genetics*, vol. 94, no. 6, pp. 854–869, 2014.

[124] D. Spieler, M. Kaffe, F. Knauf, J. Bessa, J. J. Tena, F. Giesert, B. Schormair, E. Tilch, H. Lee, M. Horsch, *et al.*, "Restless legs syndrome-associated intronic common variant in meis1 alters enhancer function in the developing telencephalon," *Genome research*, vol. 24, no. 4, pp. 592–603, 2014.

[125] D. E. Bauer, S. C. Kamran, S. Lessard, J. Xu, Y. Fujiwara, C. Lin, Z. Shao, M. C. Canver, E. C. Smith, L. Pinello, *et al.*, "An erythroid enhancer of bcl11a subject to genetic variation determines fetal hemoglobin level," *Science*, vol. 342, no. 6155, pp. 253–257, 2013.

[126] R. Stadhouders, S. Aktuna, S. Thongjuea, A. Aghajanirefah, F. Pourfarzad, W. van IJcken, B. Lenhard, H. Rooks, S. Best, S. Menzel, *et al.*, "Hbs1l-myb intergenic variants modulate fetal hemoglobin via long-range myb enhancers," *The Journal of clinical investigation*, vol. 124, no. 4, pp. 1699–1710, 2014.

[127] S. Sakthikumar, A. Roy, L. Haseeb, M. E. Pettersson, E. Sundström, V. D. Marinescu, K. Lindblad-Toh, and K. Forsberg-Nilsson, "Whole-genome sequencing of glioblastoma reveals enrichment of non-coding constraint mutations in known and novel genes," *Genome Biology*, vol. 21, no. 1, pp. 1–22, 2020.

[128] "National cancer institute cdc data portal." https://portal.gdc.cancer.gov/.

[129] "National institutes of health the cancer genome atlas (tcga)." ncbi.nlm.nih.gov/gap/.

[130] C. Kandoth, M. D. McLellan, F. Vandin, K. Ye, B. Niu, C. Lu, M. Xie, Q. Zhang, J. F. McMichael, M. A. Wyczalkowski, *et al.*, "Mutational landscape and significance across 12 major cancer types," *Nature*, vol. 502, no. 7471, pp. 333–339, 2013.

[131] A. D. Yates, P. Achuthan, W. Akanni, J. Allen, J. Allen, J. Alvarez-Jarreta, M. R. Amode, I. M. Armean, A. G. Azov, R. Bennett, *et al.*, "Ensembl 2020," *Nucleic acids research*, vol. 48, no. D1, pp. D682–D688, 2020.

[132] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of molecular biology*, vol. 215, no. 3, pp. 403–410, 1990.

[133] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The protein data bank," *Nucleic acids research*, vol. 28, no. 1, pp. 235–242, 2000.

[134] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers: Original Research on Biomolecules*, vol. 22, no. 12, pp. 2577–2637, 1983.

[135] Z. Wang, F. Zhao, J. Peng, and J. Xu, "Protein 8-class secondary structure prediction using conditional neural fields," in *2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 109–114, IEEE, 2010.

[136] P. Ni and Z. Su, "A map of cis-regulatory modules and constituent transcription factor binding sites in 77.5% regions of the human genome," *bioRxiv*, 2020.

[137] M. S. Taylor, C. P. Ponting, and R. R. Copley, "Occurrence and consequences of coding sequence insertions and deletions in mammalian genomes," *Genome Research*, vol. 14, no. 4, pp. 555–566, 2004.

[138] R. Kim and J.-t. Guo, "Systematic analysis of short internal indels and their impact on protein folding," *BMC structural biology*, vol. 10, no. 1, pp. 1–11, 2010.

[139] E. Pennisi, "Searching for the genome's second code," *Science*, vol. 306, no. 5696, p. 632, 2004.

[140] J. A. Stamatoyannopoulos, M. Snyder, R. Hardison, B. Ren, T. Gingeras, D. M. Gilbert, M. Groudine, M. Bender, R. Kaul, T. Canfield, *et al.*, "An encyclopedia of mouse dna elements (mouse encode)," *Genome biology*, vol. 13, no. 8, pp. 1–5, 2012.

[141] I. Dunham, E. Birney, B. R. Lajoie, A. Sanyal, X. Dong, M. Greven, X. Lin, J. Wang, T. W. Whitfield, J. Zhuang, *et al.*, "An integrated encyclopedia of dna elements in the human genome," 2012.

[142] L. Narlikar and I. Ovcharenko, "Identifying regulatory elements in eukaryotic genomes," *Briefings in functional genomics and proteomics*, vol. 8, no. 4, pp. 215–230, 2009.

[143] N. Nègre, C. D. Brown, L. Ma, C. A. Bristow, S. W. Miller, U. Wagner, P. Kheradpour, M. L. Eaton, P. Loriaux, R. Sealfon, *et al.*, "A cis-regulatory map of the drosophila genome," *Nature*, vol. 471, no. 7339, pp. 527–531, 2011.

[144] Y. Shen, F. Yue, D. F. McCleary, Z. Ye, L. Edsall, S. Kuan, U. Wagner, J. Dixon, L. Lee, V. V. Lobanenkov, *et al.*, "A map of the cis-regulatory sequences in the mouse genome," *Nature*, vol. 488, no. 7409, pp. 116–120, 2012.

[145] N. Neznanov, A. Umezawa, and R. G. Oshima, "A regulatory element within a coding exon modulates keratin 18 gene expression in transgenic mice," *Journal of Biological Chemistry*, vol. 272, no. 44, pp. 27549–27557, 1997.

[146] S. Tümpel, F. Cambronero, C. Sims, R. Krumlauf, and L. M. Wiedemann, "A regulatory module embedded in the coding region of hoxa2 controls expression in rhombomere 2," *Proceedings of the National Academy of Sciences*, vol. 105, no. 51, pp. 20077–20082, 2008.

[147] K. K. Barthel and X. Liu, "A transcriptional enhancer from the coding region of adamts5," *PLoS One*, vol. 3, no. 5, p. e2184, 2008.

[148] X. Dong, P. Navratilova, D. Fredman, Ø. Drivenes, T. S. Becker, and B. Lenhard, "Exonic remnants of whole-genome duplication reveal cis-regulatory function of coding exons," *Nucleic acids research*, vol. 38, no. 4, pp. 1071–1085, 2010.

[149] M. Li, H. Zhao, J. Wei, J. Zhang, and Y. Hong, "Medaka vasa gene has an exonic enhancer for germline expression," *Gene*, vol. 555, no. 2, pp. 403–408, 2015.

[150] S. Bhatia and D. A. Kleinjan, "Disruption of long-range gene regulation in human genetic disease: a kaleidoscope of general principles, diverse mechanisms and unique phenotypic consequences," *Human genetics*, vol. 133, no. 7, pp. 815–845, 2014.

[151] N. Tayebi, A. Jamsheer, R. Flöttmann, A. Sowinska-Seidler, S. C. Doelken, B. Oehl-Jaschkowitz, W. Hülsemann, R. Habenicht, E. Klopocki, S. Mundlos, *et al.*, "Deletions of exons with regulatory activity at the dync1i1 locus are associated with split-hand/split-foot malformation: array cgh screening of 134 unrelated families," *Orphanet journal of rare diseases*, vol. 9, no. 1, p. 108, 2014.

[152] H. L. Allen, R. Caswell, W. Xie, X. Xu, C. Wragg, P. D. Turnpenny, C. L. Turner, M. N. Weedon, and S. Ellard, "Next generation sequencing of chromosomal rearrangements in patients with split-hand/split-foot malformation provides evidence for dync1i1 exonic enhancers of dlx5/6 expression in humans," *Journal of medical genetics*, vol. 51, no. 4, pp. 264–267, 2014.

[153] N. Ahituv, "Exonic enhancers: proceed with caution in exome and genome sequencing studies," *Genome medicine*, vol. 8, no. 1, pp. 1–3, 2016.

[154] V. K. Yadav, K. S. Smith, C. Flinders, S. M. Mumenthaler, and S. De, "Significance of duon mutations in cancer genomes," *Scientific reports*, vol. 6, no. 1, pp. 1–9, 2016.

[155] J. C. Kwasnieski, C. Fiore, H. G. Chaudhari, and B. A. Cohen, "High-throughput functional testing of encode segmentation predictions," *Genome research*, vol. 24, no. 10, pp. 1595–1602, 2014.

[156] D. Kleftogiannis, P. Kalnis, and V. B. Bajic, "Deep: a general computational framework for predicting enhancers," *Nucleic acids research*, vol. 43, no. 1, pp. e6–e6, 2015.

[157] N. Dogan, W. Wu, C. S. Morrissey, K.-B. Chen, A. Stonestrom, M. Long, C. A. Keller, Y. Cheng, D. Jain, A. Visel, *et al.*, "Occupancy by key transcription factors is a more accurate predictor of enhancer activity than histone modifications or chromatin accessibility," *Epigenetics & chromatin*, vol. 8, no. 1, p. 16, 2015.

[158] H. Arbel, S. Basu, W. W. Fisher, A. S. Hammonds, K. H. Wan, S. Park, R. Weiszmann, B. W. Booth, S. V. Keranen, C. Henriquez, *et al.*, "Exploiting regulatory heterogeneity to systematically identify enhancers with high accuracy," *Proceedings of the National Academy of Sciences*, vol. 116, no. 3, pp. 900–908, 2019.

[159] R. Zheng, C. Wan, S. Mei, Q. Qin, Q. Wu, H. Sun, C.-H. Chen, M. Brown, X. Zhang, C. A. Meyer, *et al.*, "Cistrome data browser: expanded datasets and new tools for gene regulatory analysis," *Nucleic acids research*, vol. 47, no. D1, pp. D729–D735, 2019.

[160] Y. Huang, B. Niu, Y. Gao, L. Fu, and W. Li, "Cd-hit suite: a web server for clustering and comparing biological sequences," *Bioinformatics*, vol. 26, no. 5, pp. 680–682, 2010.

[161] S. Wang, W. Li, S. Liu, and J. Xu, "Raptorx-property: a web server for protein structure property prediction," *Nucleic acids research*, vol. 44, no. W1, pp. W430–W435, 2016.

[162] E. V. Davydov, D. L. Goode, M. Sirota, G. M. Cooper, A. Sidow, and S. Batzoglou, "Identifying a high fraction of the human genome to be under selective constraint using gerp++," *PLoS Comput Biol*, vol. 6, no. 12, p. e1001025, 2010.

[163] K. S. Pollard, M. J. Hubisz, K. R. Rosenbloom, and A. Siepel, "Detection of nonneutral substitution rates on mammalian phylogenies," *Genome research*, vol. 20, no. 1, pp. 110–121, 2010.

[164] J. Casper, A. S. Zweig, C. Villarreal, C. Tyner, M. L. Speir, K. R. Rosenbloom, B. J. Raney, C. M. Lee, B. T. Lee, D. Karolchik, *et al.*, "The ucsc genome browser database: 2018 update," *Nucleic acids research*, vol. 46, no. D1, pp. D762–D769, 2018.

[165] I. H. G. S. Consortium *et al.*, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, pp. 860–921, 2001.

[166] H. Long, W. Sung, S. Kucukyildirim, E. Williams, S. F. Miller, W. Guo, C. Patterson, C. Gregory, C. Strauss, C. Stone, *et al.*, "Evolutionary determinants of genome-wide nucleotide composition," *Nature ecology & evolution*, vol. 2, no. 2, pp. 237–240, 2018.

[167] G. M. Cooper, D. L. Goode, S. B. Ng, A. Sidow, M. J. Bamshad, J. Shendure, and D. A. Nickerson, "Single-nucleotide evolutionary constraint scores highlight disease-causing mutations," *Nature methods*, vol. 7, no. 4, pp. 250–251, 2010.

[168] A. G. Hinnebusch, I. P. Ivanov, and N. Sonenberg, "Translational control by 5â^2−*untranslatedregionsofeukaryoticmrnas*,″ *Science*, *vol*. 352, *no*. 6292, *pp*. 1413−−1416, 2016.

[169] C. Mayr, "What are 3'utrs doing?," *Cold Spring Harbor perspectives in biology*, vol. 11, no. 10, p. a034728, 2019.

[170] J. Dekker, K. Rippe, M. Dekker, and N. Kleckner, "Capturing chromosome conformation," *science*, vol. 295, no. 5558, pp. 1306–1311, 2002.

[171] J.-M. Belton, R. P. McCord, J. H. Gibcus, N. Naumova, Y. Zhan, and J. Dekker, "Hi–c: a comprehensive technique to capture the conformation of genomes," *Methods*, vol. 58, no. 3, pp. 268–276, 2012.

[172] W. De Laat and D. Duboule, "Topology of mammalian developmental enhancers and their regulatory landscapes," *Nature*, vol. 502, no. 7472, pp. 499–506, 2013.

[173] E. E. Furlong and M. Levine, "Developmental enhancers and chromosome topology," *Science*, vol. 361, no. 6409, pp. 1341–1345, 2018.

[174] G. Li, X. Ruan, R. K. Auerbach, K. S. Sandhu, M. Zheng, P. Wang, H. M. Poh, Y. Goh, J. Lim, J. Zhang, *et al.*, "Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation," *Cell*, vol. 148, no. 1-2, pp. 84–98, 2012.

[175] S. Buratowski, "Progression through the rna polymerase ii ctd cycle," *Molecular cell*, vol. 36, no. 4, pp. 541–546, 2009.

[176] E. F. Cáceres and L. D. Hurst, "The evolution, impact and properties of exonic splice enhancers," *Genome biology*, vol. 14, no. 12, p. R143, 2013.

[177] A. Pavesi, A. Vianelli, N. Chirico, Y. Bao, O. Blinkova, R. Belshaw, A. Firth, and D. Karlin, "Overlapping genes and the proteins they encode differ significantly in their sequence composition from non-overlapping genes," *PloS one*, vol. 13, no. 10, p. e0202513, 2018.

[178] Q. Ning, Y. Li, Z. Wang, S. Zhou, H. Sun, and G. Yu, "The evolution and expression pattern of human overlapping lncrna and protein-coding gene pairs," *Scientific reports*, vol. 7, p. 42775, 2017.

[179] M. F. Lin, P. Kheradpour, S. Washietl, B. J. Parker, J. S. Pedersen, and M. Kellis, "Locating protein-coding sequences under selection for additional, overlapping functions in 29 mammalian genomes," *Genome research*, vol. 21, no. 11, pp. 1916–1928, 2011.

[180] R. Savisaar and L. D. Hurst, "Purifying selection on exonic splice enhancers in intronless genes," *Molecular biology and evolution*, vol. 33, no. 6, pp. 1396–1418, 2016.

[181] R. Savisaar and L. D. Hurst, "Exonic splice regulation imposes strong selection at synonymous sites," *Genome research*, vol. 28, no. 10, pp. 1442–1454, 2018.

[182] R. Savisaar and L. D. Hurst, "Estimating the prevalence of functional exonic splice regulatory information," *Human genetics*, vol. 136, no. 9, pp. 1059–1078, 2017.

[183] C. F. Mugal, C. C. Weber, and H. Ellegren, "Gc-biased gene conversion links the recombination landscape and demography to genomic base composition: Gc-biased gene conversion drives genomic base composition across a wide range of species," *Bioessays*, vol. 37, no. 12, pp. 1317–1326, 2015.

[184] F. Pouyet, S. Aeschbacher, A. Thiéry, and L. Excoffier, "Background selection and biased gene conversion affect more than 95% of the human genome and bias demographic inferences," *Elife*, vol. 7, p. e36317, 2018.

[185] A. G. Baltz, M. Munschauer, B. Schwanhäusser, A. Vasile, Y. Murakawa, M. Schueler, N. Youngs, D. Penfold-Brown, K. Drew, M. Milek, *et al.*, "The mrna-bound proteome and its global occupancy profile on protein-coding transcripts," *Molecular cell*, vol. 46, no. 5, pp. 674–690, 2012.

[186] D. Dominguez, P. Freese, M. S. Alexis, A. Su, M. Hochman, T. Palden, C. Bazile, N. J. Lambert, E. L. Van Nostrand, G. A. Pratt, *et al.*, "Sequence, structure, and context preferences of human rna binding proteins," *Molecular cell*, vol. 70, no. 5, pp. 854–867, 2018.

[187] D. Ray, H. Kazan, K. B. Cook, M. T. Weirauch, H. S. Najafabadi, X. Li, S. Gueroussov, M. Albu, H. Zheng, A. Yang, *et al.*, "A compendium of rna-binding motifs for decoding gene regulation," *Nature*, vol. 499, no. 7457, pp. 172–177, 2013.

[188] J. Brennecke, A. Stark, R. B. Russell, and S. M. Cohen, "Principles of microrna–target recognition," *PLoS biol*, vol. 3, no. 3, p. e85, 2005.

[189] K. Leppek, R. Das, and M. Barna, "Functional 5'utr mrna structures in eukaryotic translation regulation and how to find them," *Nature reviews Molecular cell biology*, vol. 19, no. 3, p. 158, 2018.

[190] A. Piovesan, M. Caracausi, F. Antonaros, M. C. Pelleri, and L. Vitale, "Genebase 1.1: a tool to summarize data from ncbi gene datasets and its application to an update of human gene statistics," *Database*, vol. 2016, 2016.

[191] M. J. Rowley, M. H. Nichols, X. Lyu, M. Ando-Kuri, I. S. M. Rivera, K. Hermetz, P. Wang, Y. Ruan, and V. G. Corces, "Evolutionarily conserved principles predict 3d chromatin organization," *Molecular cell*, vol. 67, no. 5, pp. 837–852, 2017.

[192] D. U. Gorkin, D. Leung, and B. Ren, "The 3d genome in transcriptional regulation and pluripotency," *Cell stem cell*, vol. 14, no. 6, pp. 762–775, 2014.

[193] F. Sun, C. Chronis, M. Kronenberg, X.-F. Chen, T. Su, F. D. Lay, K. Plath, S. K. Kurdistani, and M. F. Carey, "Promoter-enhancer communication occurs primarily within insulated neighborhoods," *Molecular cell*, vol. 73, no. 2, pp. 250–263, 2019.