Transcriptome and genome analysis based on alignment-free protocols

Yunfeng Wang

université PARIS-SACLAY

Transcriptome and Genome Analysis based on Alignment-free Protocols

l'analyse du génome et du transcriptome par des méthodes sans référence

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 577, Structure et Dynamique des Systèmes Vivants (SDSV)

Spécialité de doctorat: Sciences de la vie et de la santé

Unité de recherche: Université Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC), 91198, Gif-sur-Yvette, France.

Référent: Faculté des sciences d'Orsay

Thèse présentée et soutenue à Orsay, le 22 juillet 2021, par

Yunfeng WANG

Composition du jury:

Gaelle Lelandais Professeur, University Paris-Saclay	Président
JiaYin WANG Professeur, Université Jiaotong de Xi'an (Chine)	Rapporteur & Examinateur
Nicolas Gilbert Chercheur, CRCN (U1183), Université de Montpellier	Rapporteur & Examinateur
Aitor Gonzalez Maitre de Conférences, Aix-Marseille Université	Examinateur
Eric BONNET Chercheur, Institut de Biologie François Jacob (CNRGH)	Examinateur

Direction de la thèse:

Daniel GAUTHERET Professeur, Université Paris-Saclay (I2BC)	Directeur de thèse
Yang DU Dr., Annoroad Gene Technology Co., Ltd, Beijing (Chine)	Co-Directeur de thèse

Thèse de doctorat

NNT: 2021UPASL048

Acknowledgement

Herein I guarantee that all content written in this thesis is original and does not contain any illegal reproduction. First of all, I would like to express my deepest and most sincere appreciation to my thesis director Pr. Daniel Gautheret. Before I became Daniel's Ph.D. student, I have little knowledge about the reference-free field. Daniel always patiently and enthusiastically guided me and helped me gradually expand my knowledge. In the process of editing my academic articles and thesis, Daniel provided a lot of insightful comments and critiques, from which I obtained very precious experience. Without Daniel's encouragement and guidance, I would not be able to finish my thesis. It's pity that I did not master French during the past three years. With the help of Daniel, many inconveniences due to language issues have been solved, such as renting a house and applying for a credit card.

My heartfelt thanks are also extended to Dr. Yang DU, who is my co-supervisor and leader in Annoroad. Since I graduated from college in 2015, Dr. Yang DU guided me as my leader on scientific research projects regarding NIPT. Yang not only provided academically valuable suggestions but also helped me a lot in the paper works involved in my contract.

I am truly indebted to members of my laboratory, including Fabrice, Claire, Haoliang, Ha, and Mélina. When I applied for the residence card, Fabrice drove me to the office and help translate for me. Claire, an elegant female investigator in our lab, helps me whenever I asked her for help. She always patiently answered my questions in English considering that I don't understand French. Thanks for your consideration and patience. Haoliang and Ha are the other two Ph.D. students of Daniel. It's a valuable learning opportunity to discuss with them. Mélina, Ha's co-supervisor, also gave me many insightful suggestions on research. It has been a great pleasure to work with all my colleagues.

I would like to thank the Annoroad Gene Technology, which granted me scholarships and facilitated accommodation. Dr. Chongjian Chen was the first to start this

scientific research cooperation project and appointed me to study for a Ph.D. in France. Also, I would like to thank Zhimin Li, the current CEO of Annoroad, who acknowledged the cooperation project and granted the research funding.

I also would like to offer my special thanks to all the members who joined my committee meetings, Pr. Valentina BOEVA, Pr. Chunlong CHEN, and Pr. Gaelle LELANDAIS. Also, I feel grateful to all members of my jury, Pr. JiaYin WANG, Pr. Nicolas Gilbert, Pr. Gaelle LELANDAIS, and Pr. Eric BONNET. Thanks for their comments and corrections for my Ph.D. thesis.

Last but not least, I am profoundly grateful to my parents and friends, who have always been by my side. I sincerely express my thanks to all of them, for their love, support, and encouragement in my life.

Contents

Chapter 1

Thesis Summary

A predominant application of sequencing technologies is the characterization of RNA and DNA variations in any biological samples of interests. Various bioinformatics protocols have been developed for such purpose, and most rely on the same principle, which is the alignment of sequence reads back to a reference genome. This so-called "mapping" step has several limitations. First, it is not adaptable to species with no reference available. Second, even for species with available references, such as humans, the reference is not yet complete with large gaps and unplaced contigs. Third, several regions in a genome are hard to map due to the presence of repeats, especially in the case of short-read sequencing. In this thesis, I present mapping-free protocols for transcriptome and genome analysis. Mapping-free bioinformatic methods capture variations without using a reference. They do not map reads to the reference and thus can capture events in hard-to-map regions or regions absent from the reference genome. Both transcriptomic and genomic variants can be identified in a much shorter running time comparing to other mapping-based methods. We apply our protocols to two important applications in cancer genomics: the discovery of tumor biomarkers via RNA sequencing, in the case of lung adenocarcinoma; and the discovery of somatic mutations, in the case of prostate adenocarcinoma. In the transcriptome analysis, our mapping-free approach led us to identify novel signatures in lung cancer from repeat regions as well as a number of previously un-

reported long-non coding RNA variants. For somatic variant analysis, we developed a new pipeline, 2-kupl, which takes a pair of raw sequencing data from normal and tumor samples, or two evolutionarily related samples like 2 bacterial strains. Using simulated and real-life datasets, we show our pipeline is computationally more efficient than state-of-the-art mapping-based software, with comparable detection accuracy and capability to detect novel variants.

Major contributions

- We developed analysis protocols for sequencing data based on k-mers. Our protocols do not need to map reads to the reference and thus have the capacity of capturing events in the difficult-to-map regions. Both transcriptomic and genomic variants can be identified in a much shorter running time comparing to other mapping-based methods.

- We applied our protocols to two cancer cohorts. The RNA-seq data of Lung Adenocarcinoma patients and DNA-seq data of Prostate Adenocarcinoma patients were thoroughly analyzed with our protocols. Novel transcriptional variants, especially fragments with repeats specific to lung cancer patients, were identified. Recurrent variants and genes specific to prostate cancer were identified.

Findings and Insights

- The current mapping-based protocols highly rely on the reference sequence and thus are not adaptable to species without available reference. Even for species with available references, such as humans, the reference is not yet complete and varies across individuals. Therefore, a mapping-free method without using a reference is an alternative to capture variations.

- We developed two mapping-free protocols handling RNA-seq and DNA-seq data. DE-kupl is designed to capture transcriptomic events specific to a certain phenotype group. 2-kupl is designed to detect variants between matched samples.

- The two protocols were applied to real world datasets originating from lung cancer and prostate cancer patients. The highly consistent pattern identified with DE-kupl using two different lung cancer cohorts indicates the robustness of the method, with true underlying events being captured by DE-kupl. On the other hand, with 2-kupl for DNA-seq analysis, the proposed pipeline is computationally more efficient than state-of-the-art mapping-based software, with comparable detection accuracy and capability to detect novel variants.

Chapter 2

Introduction

2.1 High throughput sequencing

2.1.1 History of High Throughput Sequencing (HTS)

Four bases (adenine (A), cytosine (C), guanine (G), and thymine (T)) constitute the genetic sequence of DNA. Genetic sequences are essential for the survival and reproduction of organisms. Deciphering genetic sequences is critical for our understanding of life. DNA sequencing is the technology to determine the exact order and type of base pairs in a DNA fragment.

The development of Sanger sequencing enabled researchers to study the exact formulation of the human genome, irrespective of its limitation in throughput and laborious technical workaround. The first human reference genome was published in 2001 (Lander et al., 2001). With a typical readout length of 1000 bp, this project took about ten years and nearly $3 billion to complete. Shortly after that, the reference genomes of several model organisms were determined (Waterston and Pachter, 2002; Mikkelsen et al., 2005).

Following the completion of the first human genome, the National Human Genome

Research Institute (NGHRI) created a DNA sequencing technology initiative aimed at reducing the cost of a fully human genome sequencing to 1000 USD (Schloss, 2008). A flurry of **H**igh-**T**hroughput **S**equencing (HTS) technologies emerged, frequented being referred to as **N**ext-**G**eneration **S**equencing (NGS) or Massively Parallel Sequencing (MPS). Compared with Sanger, HTS technologies can sequence hundreds of millions of DNA molecules in parallel, yielding shorter reads of DNA sequence of 50 to a few hundred bases.

To this day, with more advanced Long-Read Single Molecule Real-Time sequencing platforms available from Pacific Biosciences (Rhoads and Au, 2015) and Oxford Nanopore Technologies (Laver et al., 2015), HTS technologies have played an essential role in various research and clinical fields. In the next section, different sequencing technologies and platforms will be described in more detail, covering their limitations and various use case scenarios, which enables the rationale of this research work.

2.1.2 High Throughput Sequencing Technology

Next-**G**eneration **S**equencing (NGS) refers to the deep, high-throughput, parallel DNA sequencing technologies developed around three decades after the Sanger DNA sequencing method (Shendure and Ji, 2008; Sanger et al., 1977). Fueled by technical developments, basic research and market demand, several generations of NGS platforms have arisen since 2005, including Roche 454 pyrosequencing (Rothberg and Leamon, 2008), Illumina/Solexa **S**equencing **B**y **S**ynthesis (SBS) (Chi, 2008) and Ion Semiconductor Sequencing (Rusk, 2011). Comparing to the first-generation Sanger sequencing, NGS generates massive data in a few hours at a significantly reduced cost, thus becoming the top choice for large-scale genomic and transcriptomic studies.

A common first step to most sequencing protocols is library preparation (Van Dijk et al., 2014). DNA or RNA is first isolated and purified from the test sample. For

RNA sequencing experiments, RNA is first converted to cDNA by reverse transcription. The principle NGS workflow is shown in Fig 1. As a critical step in library preparation, sequencing adapters are added to the DNA fragments to make the fragment 'visible' to the sequencing device. To produce sufficient sequencing molecules in the case of low sample input, a pre-amplification step is normally performed on the ligation product as a next step. The amplification step creates theoretically identical copies of the original template, it is partially for this reason duplicated reads are observed in the data readout, which has to be adjusted accordingly with respect to the nature of the amplification method applied. The step also brings in potential thermodynamic bias related to factors like GC content and template size, which should be considered in the correction and quantification of molecular features.

During the sequencing step, libraries of prepared templates are loaded onto a reaction interface, frequently termed as flow cell or chip, and are subsequently processed inside the sequencing device. The identity of each base of a template is read off from sequential images or periodical changes ion current, concurrently for hundreds of millions of templates. Eventually, the whole nucleotide sequence of each library template is consolidated into one single string, a read, and stored in the sequencing output file. The FASTQ file format is universally used to represent raw sequencing data, as a *de facto* standard in the bioinformatics community. This format consists of four lines for each read, including the sequence and quality score of each base along the sequence. The quality scores assigned to each base call are referred to as Phred scores, which correspond to the probability that the sequencer called that base incorrectly.

Thanks to its unbiased nature and dynamic ranges in profiling bulky nucleic acids, NGS today is widely applied to investigate all areas of biology, from molecular biology to genetics, medicine, epidemiology and ecology.

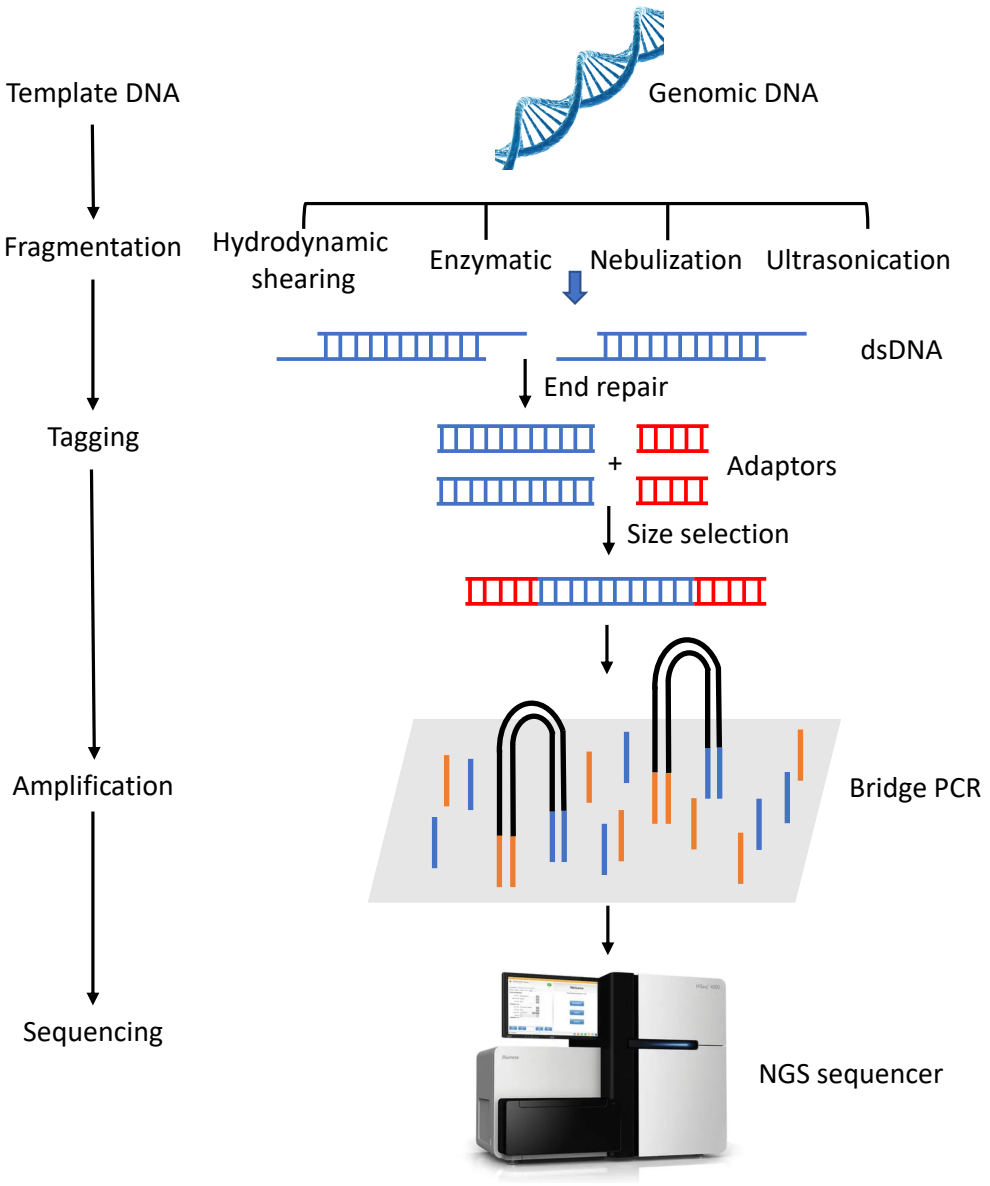


Figure 1: High-throughput sequencing workflow. The schematic shows the main high-throughput sequencing workflow including library preparation and template amplification procedures.

2.1.3 Applications of NGS

Throughout my thesis I have been interested in two major types of NGS technologies, which are generally described as "DNA-seq" and "RNA-seq" and whose main intentions are to characterize genetic variations and RNA expression. There are many other types of NGS analyses such as Methyl-seq for DNA methylation, Chip/ATAC-seq for epigenetics and protein-DNA binding, and high dimensional structural feature (Hi-C) for chromosome 3D conformation. However, I will only focus here on the two technologies at the core of my thesis work: DNA-seq and RNA-seq.

2.1.3.1 DNA Sequencing

DNA-seq technologies are used to detect various genetic alterations, including **S**ingle-**N**ucleotide **V**ariants (SNVs), **In**sertions and **del**etions (INDELS), **S**tructural **V**ariants (SVs) and **C**opy **N**umber **V**ariants (CNVs).

SNVs refer to substitution of a single nucleotide at a specific position in the genome. The term SNV is more frequently used to describe somatic mutations in cancer, while SNPs are germline substitutions and present in the general population. INDELS are classified among small genetic variations, measuring from 1 to 50 bp in length. INDELS are generally considered more deleterious than SNVs because they can lead to frameshift mutations when occurring in coding regions. SVs include various kinds of larger genome variations such as insertions, deletions, duplications, copy number variants, inversions and translocations. The size of a structural variant ranges from 50 bp to several Mb. A CNV is a type of structural variant in which chromosome parts are deleted or duplicated.

These genetic variants can also be divided into two categories, germline and somatic variants according to their biological origins. Germline mutations occur in gametes and can be passed onto offspring. Diseases caused by germline mutations are called

inherited or hereditary disorders. Although germline variants can have negative effects and cause or increase the risk of rare diseases, and even cancer, most are silent and only contribute to the genetic diversity of humans. Somatic mutations occur in non-germinal body cells and cannot be inherited by offspring. They occur from accumulated damages to the genome in individual cells throughout a person's life. Somatic mutations are considered to be the most common cause of cancer. Diseases that occur because of somatic mutations are referred to as sporadic disorders. However, most somatic mutations are silent and do not have pathological consequences. Somatic mutations are valuable for clinical diagnostics and precision medicine given that they are confined in the lesion tissues. Identifying somatic variants in cancer driver genes carries a high clinical therapeutic value. For instance, the BRAF gene is a known proto-oncogene involved in the regulation of key cellular functions. The V600E somatic mutation in BRAF (substitution of valine (V) by glutamic acid (E) at amino acid 600) causes constitutive BRAF activity, a tumor driving event in several cancers, including melanoma (Maldonado et al., 2003), lung cancer (Sánchez-Torres et al., 2013), and colorectal cancer (Li et al., 2006). One drug, Vemurafenib (Bollag et al., 2010) has been shown to be effective for the treatment of patients harboring V600E mutation. Therefore there is a high clinical interest in characterizing patients with this mutation to orient them towards Vemurafenib treatment. There are now drugs for a few dozens of somatic mutations of this kind. Cancer precision medicine seeks to determine such "actionable" somatic mutations in cancer patients in order to direct them towards adapted treatments.

Somatic variants are identified by comparing **W**hole **G**enome **S**equencing (WGS), **W**hole **E**xome **S**equencing (WES) or **T**argeted **S**equencing (TS) data from matched normal and tumor tissues (we will present detailed protocols for this later on). In a clinical setting, only "actionable" genes, for which treatment is available, are of interest. Thus using TS based on a gene panel is more practical.

WGS provides comprehensive genomic information with a uniform read depth, capturing variants either inside or outside coding regions, to explore the entire genetic code of an organism. The most successful clinical application of NGS in routine ge-

netic disease screening is Non-Invasive Prenatal Testing (NIPT). In this application, low depth WGS assists in determining whether the fetus has inherited chromosomal defects from its parents.

TS or panel sequencing, on the other hand, is an alternative that involves a target enrichment step where sequences from specific regions of interest are either directly amplified or captured before sequencing. TS allows to investigate specific areas of the genome more rapidly and cost-effectively than WGS. Clinicians have developed panels containing target genes associated with specific medical questions. WES, is one of the most commonly used forms of TS, which provides coverage restricted to all human exons to investigate only the coding sequences of the genome. Since the exome represents less than 2% of the human genome, it is a cost-effective alternative to WGS for profiling genome-wide variants to study phenotype-genotype relationships in large population studies.

Genetic predisposition plays a substantial role in multiple disorders, including cancers such as breast and ovarian cancer. Genome-wide association studies (GWAS) aims to dig out the causal relationship between specific genetic variations and disorders. These associations help reveal the abnormality of molecular mechanisms leading to complex diseases and provide novel disease causal genes and drug targets. GWAS scans genomes from different individuals and screens genetic markers that can be used to predict the risk of disease. GWAS studies can be performed using SNP arrays, WES or WGS. The major advantage of WGS is the ability of capturing disease-associated loci at any location, including in non-coding regions of the genome. Although variants in non-coding regions are considered less deleterious than in coding regions, they are thought to play important roles in gene expression regulation. A large number of germline variants associated with rare Mendelian disorders such as hearing loss, intellectual disabilities and movement disorders have been captured using WES (Tanaka et al., 2015; Traschütz et al., 2019; Rabbani et al., 2014). Complex disorders such as heart disease, hypertension, diabetes, cancer and many others listed in the **O**nline **M**endelian **I**nheritance in **M**an (OMIM) database are also under investigation using WES (Tetreault et al., 2015).

For instance, so far more than 200 genomic loci harboring common variants associated with breast cancer risk have been identified by GWAS (Michailidou et al., 2017; Zhang et al., 2020a). Among the most well-established genetic predisposing mutations, variants in BRCA1 and BRCA2, which participate in the repair of double-strand DNA breaks by homologous recombination, are responsible for the accumulation of DNA alterations and final genomic instability (Auguste and Leary, 2017). With such causal variants determined, molecular signatures like mutations in BRCA1/2 have been widely used in diagnostic screening panels for various types of cancer in clinical practice (Khatcheressian et al., 2006), especially via targeted sequencing due to the fact that many complex large genomic rearrangements are hard to detect via traditional techniques.

Apart from human genomics, DNAs from other organisms are also being widely studied using HTS. Agrigenomics or agricultural genomics aims at better understanding plant biology and improving crops from the genetic level (Hesse and Höfgen, 2001). In the past decade, NGS has had an essential impact on agrigenomics. Since the publication of the first plant genome in 2000, hundreds of new plant genomes have been sequenced and made available on the NCBI and EBI databases. One of the major contributions of NGS to agrigenomics is genome-based selection. Breeders can now more easily design and implement breeding programs to develop desirable traits such as drought tolerance, disease resistance, and higher yields. Scientists utilize sequencing data for the development of improved crops, enhancing crop productivity, resilience to climate effects, and nutritional quality (Gedil et al., 2016).

Metagenome sequencing is a particular type of WGS applied to DNA of mixed origins (ie. bacterial, plant, animal) extracted from an environmental or medical sample, which is used by microbiologists to evaluate the diversity and abundance of microbial species in a sample. Metagenomics (Thomas et al., 2012), also referred to as environmental and community genomics, is the study of the total genomic content of a microbial community. The term 'meta' implies the purpose of analyzing the mixed collection of DNA or RNA from similar but not identical items. The

total DNA and/or RNA is isolated from a microbial population without prior cultivation. Then DNA is sequenced and compared with previously known reference sequences to identify known species or to discover previously unknown species. The earliest metagenomic studies targeted 16S rRNA genes to genotype and identify the different species within the environment (Janda and Abbott, 2007). Alternatively, shotgun metagenomic sequencing (Sharpton, 2014) indiscriminately sequences genomic DNA from a sample. After reads are assigned to a taxonomic rank using bioinformatics pipelines, a composition profile of the bacterial population can be generated. Shotgun metagenomic sequencing also provides data for analyses beyond taxonomy profiling, such as metabolic pathway analysis.

2.1.3.2 RNA Sequencing

RNA sequencing (RNA-seq) aims at determining the RNA sequence contents in a sample using NGS. Over the past decade, RNA-seq has become an indispensable tool for transcriptome-wide analysis. RNA-seq provides the basic materials (sequence context and quantitative information) to assess different aspects of the transcriptome, including gene and transcript expression, alternative splicing and discovery of new transcripts (Trapnell et al., 2012; Sultan et al., 2008; Robertson et al., 2010; Trapnell et al., 2013). RNA-seq in principle analyzes all RNA types in the transcriptome, including mRNA, lincRNA, snoRNA, miRNA, etc. RNA-seq library preparation differs from DNA-seq as RNAs are highly unstable and need to be converted into cDNA before amplification. RNA-seq libraries can be either strand-specific or non-strand-specific. The strand-specific libraries keep the information about which DNA strand is transcribed. This information is particularly valuable for distinguishing antisense transcripts. Stranded RNA-seq provides a more accurate estimate of transcript expression than non-stranded RNA-seq (Zhao et al., 2015).

In order to detect and quantify mRNA/gene, highly abundant ribosomal RNAs (rRNAs) must be removed from the total RNA before sequencing. There are two

strategies addressing this issue. The first one is to enrich the polyadenylated (polyA) RNA transcripts with oligo (dT) primers (Mortazavi et al., 2008). However, this strategy is not able to capture non-polyA transcripts and partially degraded mRNAs. The other strategy is depletion of highly abundant rRNAs through hybridization capture followed by magnetic bead separation (O'Neil et al., 2013). The rRNA depletion strategy provides more information on non-polyA transcripts and degraded RNAs but costs more than polyA enrichment.

RNA-seq also plays a role in the study of hereditary disorder research. Across a variety of hereditary disorders, more than half of the patients do not receive a genetic diagnosis after WES or TS. The main reason is that part of the detected genetic variations remain of unknown significance (Kremer et al., 2017). With RNA-seq, limitations of the sequential information can be complemented by integrating RNA abundance and RNA sequence, including allele-specific expression and alternative splicing. Genes with expression beyond an expected range are more likely to be causal genes. The genetic causes of such aberrant expression include rare variants in the promoters and enhancers and also inside coding regions. Besides, variants located at splicing sites may induce splicing changes, leading to abnormal transcripts and peptides. Aberrant splicing has long been studied as a major cause of hereditary disorders (Tazi et al., 2009). Nevertheless, detecting aberrant splicing from genetic sequences is difficult because splicing involves a complex set of cis-regulatory elements, some of which are inside intronic regions and are thus not covered by WES or TS (Xiong et al., 2015). Finally, RNA-seq can reflect allele-specific expression, whereby one allele is silenced and the other allele is expressed. When assuming a recessive mode of inheritance, genes with a heterozygous variant identified by WES and WGS are not prioritized. However, allele-specific expression of a heterozygous variant fits the recessive mode of inheritance assumption. Therefore, detection of allele-specific expression can help prioritize heterozygous rare variants. Alterations at the RNA sequence level and the expression levels contribute to at least half of inherited human diseases (Jackson et al., 2018). Therefore, transcriptome variations are also key information for hereditary disease interpretation.

RNA-seq is also widely applied in the study of somatic variations in cancer. The most common application of RNA-seq in cancer is the discovery of gene fusions, which are among the most frequent cancer drivers (Taniue and Akimitsu, 2021). Cancer cells are also characterized by specific gene expression signatures (Golub et al., 1999). RNA-seq is now used to characterize cancer gene expression profiles, but it can also identify new RNA isoforms. Cancer-specific RNA isoforms may be translated and produce neoantigens that are presented on the surface of tumor cells. Standard protocols integrate DNA-seq and RNA-seq to screen neoantigens that are specifically expressed in tumor tissues. Mutated peptides transcribed from these somatic mutations are then submitted to an epitope presentation prediction pipeline (Gopanenko et al., 2020). Neoantigens also derive from transcripts from non-coding regions such as lncRNAs and repeats (Ouspenskaia et al., 2020; Laumont et al., 2018). However, these regions are rarely investigated to discover neoantigens.

With the explosive growth of samples and sequencing output in this era of big data, many challenges have emerged on the computational side, concerning data storage, CPU requirement, replicability, reproducibility, data integration, and interpretation.

2.2 Standard bioinformatics pipelines for NGS analysis

2.2.1 RNA-seq workflows

In this thesis, I will focus on the most common applications of RNA-seq, i.e. gene expression quantification and profiling. A large number of efficient and accurate software have been developed to address this question. The optimal set of tools to use will depend on the specific biological question being explored. However, even within a single application (e.g., finding genes overexpressed in a given condition), different combinations of tools in the workflow can substantially affect biological

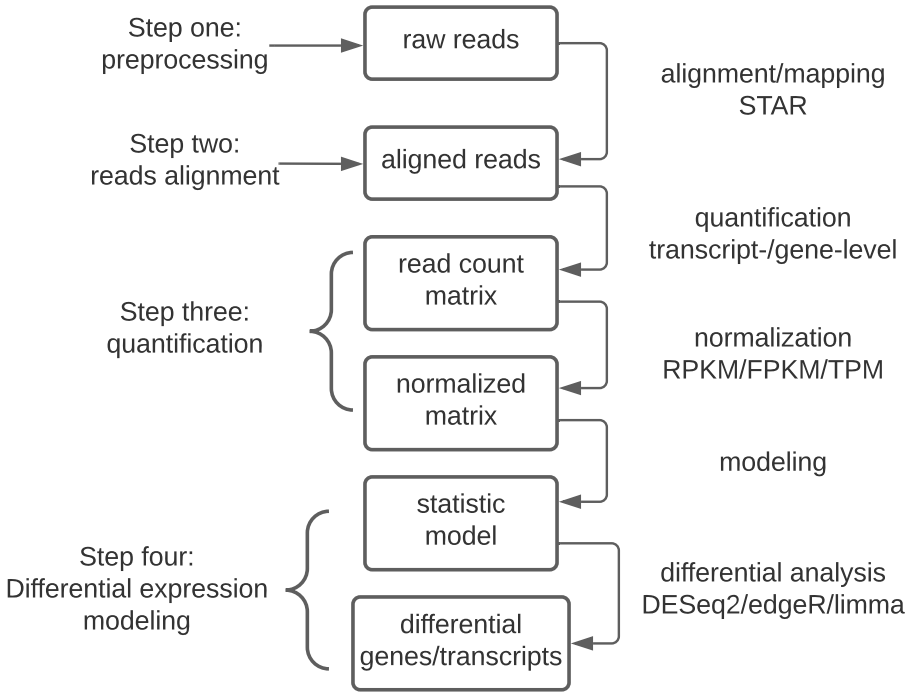


Figure 2: General workflow of RNA-seq analysis. The workflow is composed of four main steps: preprocessing, reads alignment, quantification and statistical modeling

conclusions (Kulkarni and Frommolt, 2017). A typical gene expression profiling workflow comprises four distinct steps described below and summarized in Figure 2.

2.2.1.1 Step one: Preprocessing of raw sequencing reads

A fair quality control assessment and the corresponding preprocessing of raw sequence data are fundamental for optimal downstream RNA-seq data analysis. Quality control for the raw reads involves estimating the read quality, nucleotide distribution, GC content, handling low-quality bases/reads, trimming adapters from raw sequencing reads, filtering unwanted sequences, and contamination test (Li et al., 2015b).

In short read sequencing, the quality of sequential bases steadily declines along the

sequence. The sequencing quality of the library can be measured by summarizing the base qualities of all positions across all reads. Analysts can average the quality of each read and estimate the distribution of all reads quality. There are different base quality encoding systems according to different sequencing platforms. The most commonly used one is the Illumina Phred+33 encoding system (Cock et al., 2010). Under this encoding system, a particular base with a quality score of 30 suggests the probability of a sequencing error at this position is 0.001. Therefore, if most reads have an average quality over 30, it is generally considered of good quality and accumulation of random errors at this scale will not affect the performance of most downstream analyses. Meanwhile, the reads with poor quality are supposed to be eliminated from the raw reads.

Experts often check the sequence content through the nucleotide distribution by sequencing cycle. Ideally, the four nucleotides should have a roughly constant distribution across all cycles. However, in short-read RNA-seq data, nucleotide contents at the beginning of reads are biased, due to library preparation artifacts. In RNA-seq analysis, the random hexamers or transposases are used in the library preparation instead of random primers (Syed et al., 2009). Bias in random hexamers leads to an unstable nucleotide content at the beginning of sequencing reads, which impacts the uniformity of the reads along with expressed transcripts (Hansen et al., 2010). In order to avoid the influence of the bias introduced by random hexamers, two strategies can be considered. One strategy is trimming the first several bases to exclude error-rich stretches of RNA-seq reads (Matvienko et al., 2013; Ashrafi et al., 2012). The other strategy is using specific bias correction models to adjust for the bias (van Gurp et al., 2013; Hansen et al., 2010).

Nucleotide distribution *perse* is not a reliable measure of RNA-seq data quality. For example, if one RNA is overrepresented, then the distribution of bases at each position will to some degree, be affected by the sequence of that RNA. GC content, computed as the percentage of G + C in the data, is also subjected to this problem. GC content varies by species and genomic regions. The GC content of the sequenced data is expected to be an approximation of the reference sequence. However, in

RNA-seq data, highly expressed RNAs are often tissue-specific and vary across samples. The overall GC content is influenced by the total RNA components and turns out to be unstable among samples. When performing total RNA-seq, RNA components include mRNA, rRNA, precursor messenger RNA (pre-mRNA), and several types of non-coding RNA (ncRNA). Thus, the actual GC content depends on the composition of RNA subtypes. Therefore, nucleotide distribution by cycle and the GC content are not suitable parameters to assess sequencing quality for total RNA-seq (Sheng et al., 2017).

Adapter sequences, single-stranded oligonucleotide sequences ligated to RNAs for cDNA synthesis, should be eliminated from reads (Li and Weeks, 2006). Sometimes adapters are also sequenced if the inserted size of the fragment is shorter than the sequencing cycle. Because adapters are artificially introduced and are not part of the organism's transcriptome, these sequences should not be counted or utilized for biological interpretation. Adapter sequences can be trimmed from both ends of the template readout. It is worth noting that trimming makes the reads shorter. So one can filter out the entire read if the average quality is below a certain threshold or the read length is too short. In long RNA-seq data analysis, adapter trimming is unnecessary because the RNA fragments are long, and the adapters are unlikely to be sequenced (Sheng et al., 2017). Even with a partial sequencing of the adapter, the alignment can also be performed because of most aligners' soft clip functionality (Au et al., 2017). Another important quality consideration is the sequencing quality score. Low quality scores may lead to a significant portion of unusable reads. Reads with a large fraction of low-quality bases are generally trimmed or entirely removed from the sequencing data before downstream analysis. Available software for trimming are TagCleaner, Trimmomatic, and cutadapt (Schmieder et al., 2010; Bolger et al., 2014; Martin, 2011).

Bioinformaticians have also developed protocols to encapsulate the aforementioned processes into a single package. For instance, the RSeQC package comprises multiple python and C scripts that comprehensively evaluate the quality of different aspects of RNA-seq data (Wang et al., 2012).

2.2.1.2 Step two: Alignment of sequencing reads

The goal of read alignment is to find out where a read originated from. There are two mapping strategies that consist of mapping to a reference genome or a reference transcriptome. As mapping to a genome can be achieved independently of gene annotation, it allows finding new genes and transcripts. It is worth noting that, in eukaryotes, RNA-seq reads alignment is more challenging than DNA-seq alignment because of RNA splicing. When aligners attempt to match spliced reads to the genome, the beginning part matches to one exon and the last part matches to another exon. So, somehow, the aligner program has to be able to place such read correctly to figure out exactly where exon/intron borders are. This is quite difficult for aligners because the distance between two exons can be thousands of bases long. There are splice site signals, but they are usually too weak to rely on. Therefore, aligners have to be able to cope with those situations.

In eukaryotes, RNA-seq read alignment is achieved using splice-aware alignment tools such as STAR (Dobin et al., 2013). In addition to STAR, RNA-seq alignment has traditionally been accomplished using distinct alignment tools, such as TopHat (Trapnell et al., 2009), MapSplice (Wang et al., 2010b), SOAPSplice (Huang et al., 2011), HISAT (Kim et al., 2015) and GSNAP (Wu et al., 2016). All these tools perform a spliced alignment allowing for gaps in reads spanning exons or exon borders. We will take STAR here as an example of the splice-aware mapping strategy. STAR achieves alignment using three procedures: indexing, seed searching and stitching. STAR first creates genome indexes using suffix arrays. For each read, STAR then searches for the longest sequence that exactly matches one or more locations on the reference genome. These longest matching sequences are called the Maximal Mappable Prefixes (MMPs). The first MMP of the read that is exactly mapped is called 'seed1'. STAR will then search again for only the unmapped part of the read to find the next MMP, which will be seed2. Seed2 will be extended in case no exact matching sequence is found in the unmapped part. STAR uses the indexed genome to efficiently search for the MMPs. Eventually, the seeds are stitched together based

on the best alignment between the read and the genome.

2.2.1.3 Step three: Quantification of gene expression or transcript abundance

The difference between gene-level and transcript-level quantification is that gene-level summarizes read counts over genes while transcript-level summarizes read counts over transcripts. Both gene-level and transcript-level quantification can be assessed based on the alignment against a reference as mentioned above. The counts of mapped reads will be used as a surrogate for quantification.

There are typically two strategies to obtain gene-level quantification. The simplest way is to aggregate raw counts overlapping the features of interest provided in the annotation file (GTF or GFF) containing the genome coordinates of the genes. Using alignment files as input, various approaches like featureCounts (Liao et al., 2014), HTSeq (Anders et al., 2015), or the built-in capability of STAR (Dobin et al., 2013) can be applied to assess how many fragments fall in the genomic region of each gene. Gene-level quantification is simple and fast but suffers from various drawbacks: There is no consensus on the handling of multi-mapping reads. In general, these reads are discarded when summarizing at the gene level. In addition, gene-level quantification is oblivious to potentially important compositional changes that are not represented directly in gene-level read counts (e.g., differential usage of isoforms) (Van den Berge et al., 2019).

The second strategy is to first perform transcript-level quantification and aggregate transcript counts to the gene level. In plants or animals, one gene can have several transcript isoforms that differ by alternative exon usage, transcription start or transcription termination. Transcript-level quantification has remarkable advantages. First, it allows for improved biological interpretation as it potentially captures changes in each transcript usage. Second, it enables more accurate quantification of gene expression (Trapnell et al., 2009, 2010). Third, transcript-level quantification is more appropriate for modeling and correcting technical biases (Roberts et al.,

2011).

More accurate downstream analyses are obtained by appropriately modeling of mean-variance relationship of the feature counts (Pimentel et al., 2017). Despite the advantages mentioned above, transcript-level quantification is more challenging than at the gene-level. If a read comes from an area that is common to several isoforms, one needs to determine which transcript it should be assigned to. The estimated abundance of transcript isoforms depends on the assignment of these shared reads.

Leading algorithms developed to tackle this problem use maximum likelihood (ML) estimate, Bayesian inference or expectation maximization (EM) methods (Zhang et al., 2016; Trapnell et al., 2010). These algorithms assign ambiguously mapped reads to the most likely transcript isoforms. Tools employing such probabilistic approaches can also measure the uncertainty in isoform quantification.

Well-known transcript-level quantification tools include RSEM (Li and Dewey, 2011), CuffLinks (Trapnell et al., 2010) and MMSeq (Turro et al., 2014). These are based on genome alignment, ie. reads are first aligned to a genome and later re-allocated to known genes/transcripts based on transcript annotation. An alternative strategy is 'pseudo-alignment', used by various mapping-free approaches such as Kallisto (Bray et al., 2016), Sailfish (Patro et al., 2014) and Salmon (Patro et al., 2017). These methods are detailed in Chapter 2.3.3.

One should consider factors that might influence the number of reads assigned to a given gene or transcript, such as transcript length, sequencing depth, library preparation (PCR) and *in silico* factors (alignment) (Roberts et al., 2011). The PCR procedure used for cDNA amplification brings additional biases, such as the GC bias and duplicated reads. *In silico* factors represent how pipelines deal with reads that are not uniquely mapped to the genome or transcriptome. Generally, quantified gene or transcript counts need to be normalized to account for differences in read depth, gene length and technical biases (Robinson and Oshlack, 2010). Over the years, researchers have implemented different normalization methods (Abbas-

Aghababazadeh et al., 2018). The simplest one is CPM that represents the count per million reads. When the sample is sent to a facility for sequencing, operators test RNA concentration and load as similar RNA fractions as possible. But read counts per RNA-seq library are never identical and one needs to account for these differences. CPM is basically depth-normalized counts but does not account for the gene length. To address this issue, RPKM and FPKM were proposed. RPKM is a normalization method used for single-end sequencing. RPKM represents the number of reads per kilobase of transcript per million reads of a library. When the RNA-seq data is paired-end, one uses FPKM, where F represents fragments that are one pair of reads. RPKM and FPKM take both read depth and gene length into account. Later on, TPM normalization was developed, which stands for transcripts per million. TPM was proposed as a more accurate measure of transcript expression as it measures the number of transcripts produced by each gene, independently of their length, which is particularly important when genes have multiple isoforms(Trapnell et al., 2013). TPM corrects for the gene length first and then divides by the scaling factor. TPM is most appropriate for comparing the proportion of reads mapped to a gene in different samples (Bedre).

2.2.1.4 Step four: Differential expression modeling

Accurate quantification of the expression level of genes or transcripts enables the identification of genomic features that are expressed differently between conditions. The statistical power of detecting differential expression depends on several factors, such as sample size, expression level, expression fold change, sequencing depth, and dispersion/variability. A statistical model's ability to detect differential expression is better when handling samples of large size due to the high signal-to-noise ratio. High expression level, fold change and sequencing depth make it easier to observe significant differences between conditions. Depending on the experiment condition, one should also pay attention to the dispersion/variability of gene expression, which can be very high in heterogeneous samples such as tumors.

Lastly, two other factors impacting differential expression analysis are independent filtering and multiple testing (Bourgon et al., 2010; Dudoit and Van Der Laan, 2007). As genes or transcripts with low read counts are generally not informative, they can be filtered from the dataset. Multiple testing issues arise when a P-value is computed for a large number of observations. In a nutshell, when one says 'P-values < 0.05 are significant', one means '5% of the time one will report a false positive'. It becomes problematic when multiple genes or transcripts are tested because of the type I error. Then, a false discovery rate (FDR) can be computed, which is defined as the expected proportion of false positives among the declared significant results. The most commonly used multiple testing correction method calculating FDR is the Benjamini–Hochberg (BH) procedure (Bogdan et al., 2008).

Several tools are commonly used for differential expression analysis. Some utilize gene-level expression, whereas others rely on transcript-level estimates. Gene-level tools typically rely on aligned read counts and use generalized linear models (GLM) to evaluate genes' differential expression (Nelder and Wedderburn, 1972). A GLM models each gene's expression as a linear combination of explanatory factors (e.g., Group, time, patient, etc.). GLM allows the expression value distribution to be different from the normal distribution.

Statistical models for gene-level analysis include DESeq2 (Love et al., 2014), edgeR (Robinson et al., 2010) and limma (Law et al., 2014), which provide comparable results (Seyednasrollah et al., 2015). The most widely used algorithm, DESeq2, shrinks log fold change estimates toward zero using an empirical Bayes method. DESeq2 also has an outlier detection method to rule out genes with large abundance in only a few samples. edgeR performs an exact test for negative binomial distribution using the likelihood ratio test. A minimum CPM cutoff is typically determined based on the number of genes with FDR lower than 0.05. edgeR recommends filtering out all genes except those with CPM over cutoff in at least two samples. Limma is a linear model for microarray and RNA-seq data. It uses an empirical Bayes method to share information across the genes to provide stable variance estimation. Limma first creates a design matrix using the explanatory variables

and applies this to every gene independently. By default, the Benjamini–Hochberg procedure is used to estimate the FDR.

Tools that model transcript-level expression, such as CuffDiff2 (Trapnell et al., 2013), EBSeq (Leng et al., 2013) and Ballgown (Frazee et al., 2015), tend to be more computationally intensive comparing to gene-level methods and provide more distinctive call sets between different tools (Soneson and Delorenzi, 2013). Cuffdiff2 estimates expression at a transcript-level resolution based on a beta negative binomial model for counts of transcripts. Although Cuffdiff2 performs differential expression analysis at the transcript level, it reports differential expression at the gene level. By default, Cuffdiff2 calculates a normalization factor as DESeq2 to correct sequencing depths and uses the BH procedure to control the FDR. Cuffdiff2 addresses count uncertainty due to ambiguous reads that yield false differential expression signals, especially when genes have similar isoforms. EBSeq was also developed to detect differential expression at the transcript level. This program estimates the posterior likelihoods of differential and similar expression via the empirical Bayesian method. EBSeq uses a median normalization procedure similar to that of DESeq2 to account for the different sequencing depths. A Bayesian FDR estimate is produced. Ballgown extracts the abundance estimates for exons, introns, transcripts or genes, and applies a linear model for differential expression analysis. Ballgown is less computationally demanding than Cuffdiff2 and EBSeq.

2.2.2 DNA-seq workflows

Our focus here will be on WGS and WES pipelines which are most common now in human genomics. Both methods enable the discovery of SNVs, INDELS, CNVs and mutational signatures, which refer to the characteristic combinations of mutations arising from carcinogenesis and normal somatic mutagenesis processes (Alexandrov et al., 2015). Furthermore, WGS and WES can be used to find somatic events when a pair of samples from the same donor is available. With its complete genome coverage, WGS has the extra benefit of enabling variant discovery in non-coding

regions such as introns, UTRs and intergenic regions. Furthermore, WGS provides a better estimation of CNV location and mutational signature due to its denser coverage. Bioinformatics pipelines for WGS and WES (DNA-seq) data analysis generally consist of the common steps detailed below.

2.2.2.1 Step one: Alignment of sequencing reads

The first process in a DNA-seq workflow (Figure 3) involves the alignment of sequencing reads. Alignment is the process of mapping reads to a reference genome. Numerous mapping programs for DNA-seq exist, such as BWA (Li and Durbin, 2009), Bowtie2 (Bray et al., 2016), MAQ (Walsh et al., 2008), Stampy (Lunter and Goodson, 2011), and Novoalign (http://www.novocraft.com). These software use two main algorithms: the hash-based index search and the Burrows-Wheeler Transform (BWT).

Hash-tables allow for a rapid search of sequence words in an index. Hash-based algorithms build a hash table from the NGS reads (MAQ (Walsh et al., 2008), SHRiMP (Rumble et al., 2009) and ZOOM (Lin et al., 2008)), or from the reference genome (SOAPv2 (Li et al., 2009b), GSNAP (Wu et al., 2016), Novoalign (http://novocraft.com/) and PERM (Chen et al., 2009)). After building the hash table the first group of algorithms use the reference genome to scan the hash table of NGS reads while the second group of algorithms uses the set of input reads to scan the hash table of the reference genome.

BWA and Bowtie (Langmead, 2010) use a different technique, the Burrows-Wheeler transform (BWT), which creates an index that enables fast word search while being more space-efficient than hash tables. A BWT first reorders subsequences of the reference genome in a structure called a suffix array. Next, the final BWT index is created and is used for rapid read placement on the genome. An advantage of BWT is memory usage: a BWT can fit the entire human genome in less than two gigabytes of memory. In contrast, MAQ's spaced seed index may require more than fifty gigabytes of memory. Furthermore, hash-based algorithms have a speed

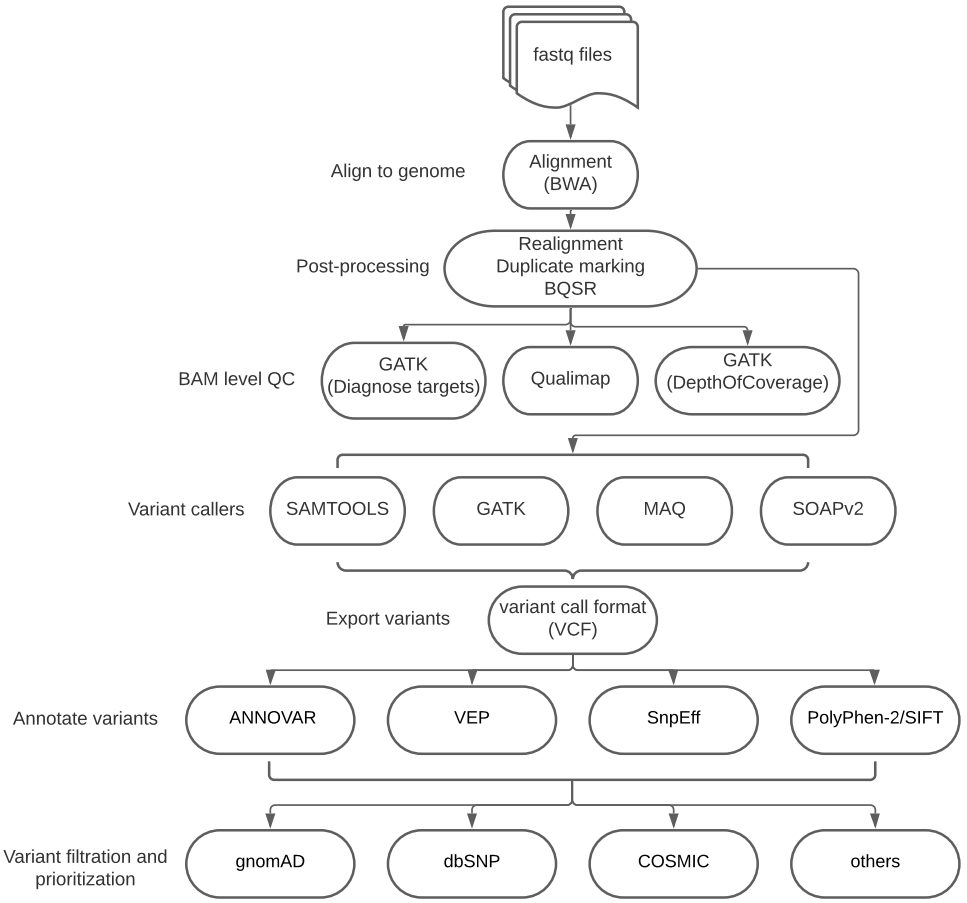


Figure 3: General workflow for DNA-seq analysis. This workflow is compatible with both WES and WGS technologies. The fastq files undergo quality control, mapping to the reference and conversion to alignment files in bam format. Multiple algorithms can be applied to call genomic variants such as SNVs, indels and SVs from the alignment results

disadvantage when applied to large genome size and repetitive regions. A lot of keys in the hash table have thousands of occurrences. Visiting each of them along the genome is quite slow. In contrast, BWT-based algorithms can tell if the read has a perfect match in O(l) time, no matter how long or how repetitive the genome is. Essentially, BWT collapses all the copies of a substring so each sequence can be aligned to all copies, rather than align to each copy as a hash table does.

Factors limiting alignment accuracy include genetic variants, sequencing errors and repetitive sequences. First of all, the matching is often not exact because individuals have variants in their genome relative to the reference genome. Also, the reads contain sequencing errors. Furthermore, there are regions that are not unique meaning that many sequences can appear in multiple parts of the genome. So when aligners try to place the reads from this sort of area, it is difficult to say where they truly come from. The **map**ping **q**uality (MAPQ) is thus introduced to represent the confidence that the read is correctly mapped to the genome. From this stage, aligners produce a certain proportion of low MAPQ alignments. Generally, we expect between 70 and 90% of regular DNA-seq reads to map onto the human genome with a MAPQ of 20, which means the probability of correct alignment is 0.99 (Conesa et al., 2016). The other reads with low MAPQ include reads with sequence variants too complex to be deciphered by aligners, such as large indels (Malde, 2008). When aligning reads to the reference, variants may introduce gaps that allow aligners to match more terms than a gap-less alignment can. To obtain an accurate alignment, aligners will allocate a penalty score to the alignment with gaps. Alignments are discarded when final scores are below a defined threshold. Another major source of low-quality reads is the low complexity regions on the genome (Dozmorov et al., 2015). In some cases, reads overlapping low complexity regions are discarded to increase the quality and reliability of DNA-seq alignment. Unfortunately, biological information located in these low MAPQ regions is then lost.

Alignment of sequence reads to a reference genome results in the generation of a SAM file, which is the universal file format for mapped sequence reads. SAM files can be compressed to BAM files.

2.2.2.2 Step two: Post-alignment processing

Aligned reads are post-processed to guarantee the reliability of alignment results for downstream analysis. Post-processing typically includes duplicate marking, local realignment around known indels, and **b**ase **q**uality **s**core **r**ecalibration (BQSR) (DePristo et al., 2011). This step is not necessary for RNA-seq since the erroneous alignments do not significantly affect gene quantification and differential analysis. However, post-processing is strongly recommended in genome analysis since misaligned reads and low-quality scores have an essential impact on the subsequent variant calling step. False positives may be introduced if such stochastic and systemic errors are not corrected (DePristo et al., 2011).

Duplicated reads are groups of reads that are identical. They are prevalent in both WGS and WES and are believed to derive from the PCR amplification process. The inclusion of duplicated reads introduces erroneous calls since they impact allele frequency estimates.

Local realignment is essential for detecting indels. Reads harboring indels tend to have a higher chance of being aligned incorrectly compared to the reads only with SNVs or without any variant. Studies demonstrated that BWA generated incorrect alignment for over 15% of reads harboring indels (DePristo et al., 2011). Without realignment, these misaligned reads lead to spurious variants. Some aligners (such as Novoalign http://www.novocraft.com/products/novoalign/) and variant callers (GATK HaplotypeCaller) are capable of indel alignment improvement.

After deduplication and indel realignment, BQSR is recommended to improve the accuracy of base quality scores before the variant calling step (Cline et al., 2020). BQSR uses the reference genome of the organism being sequenced and considers any deviation from the reference as a sequencing error. In this way, it can model accuracy dependency on genome regions. For example, when there are a series of identical bases in a read, the sequencer accuracy decreases when calling the next base whatever that base is. This is one of the biases BQSR looks for. Once BQSR

has figured out the biases, it corrects the quality scores to adjust for these biases. BQSR helps generate more accurate quality scores, which leads to more accurate variant calling in DNA-seq.

Popular tools for performing the above post-processing steps are SAMTOOLS (Li et al., 2009a), PICARD (http://picard.sourceforge.net/) and GATK (DePristo et al., 2011).

2.2.2.3 Step three: Variant calling

The variant calling step detects different types of genomic variants, including SNVs, INDELS, CNVs and large SVs. Here it is vital to distinguish somatic from germline variants. Somatic variants are present only in somatic cells and are tissue-specific, while germline variants are inherited mutations presented in the germ cells and are linked to a patient's family history.

A germline variant caller generally has a ploidy-based genotyping algorithm built into it. For instance, one generally expects to see a variant in 50% of reads covering this site when the variant is heterozygous, or 100% when it is homozygous. Therefore, the ratio of reads supporting the alternative allele can be used to call a genotype for a germline variant. However, when handling somatic variants from tumors, the assumption about which ratio to expect with a variant at a position is no longer valid. That is because we have to deal with a whole host of other factors that make somatic variant calling more challenging. First, tumor biopsies are not pure. Normal admixture in the tumor cells leads to underestimation of variant allele fraction (VAF). Second, tumors are generally composed of multiple subclones. The subclonal variants may only occur in any fraction of the cells, meaning that the VAF might vary from 50% to below 1%. Copy number variants and ploidy changes shift the distribution of variant fractions even more.

There are many algorithms for calling variants from a DNA-seq alignment file. The most widely used are SAMTOOLS (Li et al., 2009a), GATK (McKenna et al.,

2010), MAQ (Li et al., 2008), and SOAPv2 (Li et al., 2009b). Variant callers export variants in the variant call format (VCF), which contain variant positions and related statistical information (i.e. quality score, depth, allele frequency, predicted genotype).

Additional analysis steps are required if matched samples are used (i.e., tumor and normal tissues from one individual). WES is widely applied in such cases with the aim to detect in coding sequences somatic variants that impact a particular phenotype. BAM files need to be generated for both matched samples. Software are available that take this input for screening somatic variants specific to one sample. Leading somatic variant calling software include SomaticSniper (Larson et al., 2012), Strelka (Saunders et al., 2012), VarScan 2 (Koboldt et al., 2012), MuTect2 (Cibulskis et al., 2013) and MuSE (Fan et al., 2016).

The performance of variant calling software is affected to varying degrees by several factors (Krøigård et al., 2016). Firstly, somatic variant callers are challenged by the conflicting needs of detecting true low-frequency mutations and avoiding false positives. Each caller algorithm presents its own version of this compromise. Next, the alignment of the sequenced reads significantly influences variant calling accuracy. Variants falling into low complexity regions such as repeats and variants grouped in clusters are more difficult to detect correctly. In such cases, a high sequencing depth results in a higher level of agreement among the different variant callers. Finally, variant callers tend to present a lower agreement in indel calling compared to SNV calling, suggesting that indel calling poses greater challenges to the callers (Krøigård et al., 2016).

2.2.2.4 Step four: Variant annotation

Once variants are identified, they need to be annotated to determine their potential functional impacts on genes. Variant annotation generally involves retrieving information about the locus type, allele frequency and variant type. Researchers usually focus on SNVs and INDELS that occur in coding regions, which account

for 85% of known disease-causing mutations in Mendelian disorders and plenty of modifications in complex somatic disorders (Gilissen et al., 2012). The functional regions include exons, splice sites, and transcription regulatory sites.

The main tools for predicting the consequences of variants are ANNOVAR (Wang et al., 2010a), Ensembl Variant Effect Predictor (VEP) (McLaren et al., 2016), SnpEff (Cingolani et al., 2012), PolyPhen-2 (Adzhubei et al., 2010) and SIFT (Ng and Henikoff, 2003). ANNOVAR is a powerful pipeline that integrates over 4,000 public databases for variant annotation, including 1000-Genomes, dbSNP, and the NCI-60 human tumor cell line panel exome sequencing data. When variants are shared by multiple transcripts, ANNOVAR returns only the most deleterious variant based upon a priority system. VEP determines the effects of regulatory region variants (SNVs, insertions, deletions, CNVs or structural variants) on genes, transcripts and proteins. For variants shared by more than one transcript, VEP lists variant classifications in every transcript. SnpEff annotates variants according to their genomic locations and predicts functional effect using an "interval forest" data structure (Cormen et al., 2009). The interval forest is a hash of interval trees indexed by chromosome, which is used to extract variants that intersect any interval. The intervals are retrieved from a Gene transfer format (GTF) file, which could be downloaded from GENCODE and Ensembl. Each interval tree is composed of nodes. Each node includes a center point and all intervals overlapping the center point. The interval forest makes it possible to perform an efficient interval search. The VCF files containing genomic coordinates of variants are parsed. Each variant queries the interval forest to find intersecting genomic annotations. SnpEff predicts variant effects in various genomic regions (intronic, untranslated, upstream, downstream, splice site, intergenic and coding). In coding regions, effects include synonymous or non-synonymous amino acid substitutions, start/stop codon gains or losses, and frameshifts. PolyPhen-2 and SIFT are two leading machine learning tools for predicting the damaging effects of non-synonymous mutations. PolyPhen-2 uses sequence conservation, structural information and SWISS-PROT annotation to model and score amino acid substitution. SIFT also predicts whether an amino acid

substitution affects protein function, but is based only on sequence conservation.

Several studies have shown that these tools do not always provide consistent results. Comparative analysis showed that the concordance between different annotation tools was lower than 50% (McCarthy et al., 2014). This inconsistency is explained of course by the different scoring methods, but also by the way each tool defines non-coding features. For instance, SnpEff uses 5 kb to define upstream and downstream regions, while ANNOVAR uses 1 kb. In many other cases, variants in the non-coding locations are bucketed into the "ignored" category. Therefore, the dependence of annotation results should be dealt with care in a research context. More work is required to improve variant annotation, especially for clinical use. Careful thought needs to be given before deciding on a tool for variant annotation to achieve reliable interpretations.

2.2.2.5 Step five: Variant filtration and prioritization

In oncology research, germline variants are separated and ruled out from somatic variants based on their presence in the normal tissues. If no normal tissue is used, both germline and somatic variants are processed by the same filtration and prioritization procedure.

Even with strong experimental design, DNA-seq data often predict many more candidates with functional effects than verified experimentally. Genes that are frequently mutated in individuals may turn out unrelated to disease. These genes can be unusually long and thus be more likely to harbor variants or they can be located in highly polymorphic regions of the genome. Moreover, variants of interest are subject to false positives due to contamination artifacts, sequencing errors, incorrect alignment and limitations of scoring models. Assessing candidate variants in the context of existing biological knowledge, taking known molecular functions into account, is an important step in producing a manageable variant set for further study. One powerful screen for variant prioritization is to consider variants significantly related to a phenotype (for instance a disease status) (Broekema et al., 2020). For

this purpose, one can integrate different shreds of evidence, including the variant gene product effect, variant recurrence in the studied population, and gene product function.

A widely used and efficient method to prioritize variants is recurrence analysis. The rationale behind recurrence analysis is that, if a variant occurs in multiple patients independently, it might be a driver variant. If mutations occur in the same gene more often than expected by chance in a given cancer cohort, it is reasonable to postulate the gene is involved in the genesis of this cancer. Different resources are available to estimate variant frequencies. The 1000-Genome project (Siva, 2008) identifies variants with over 1% frequency in the sampled human populations. NHLBI-ESP (Auer et al., 2016) discovers heart, lung, and blood disorder variants at a frequency lower than 1%. The Exome Aggregation Consortium (ExAC) (Karczewski et al., 2017) is a data set that compiles the largest exome sequencing datasets. The cohort includes various diseases besides normal samples. The Genome Aggregation Database (gnomAD) (Karczewski and Francioli, 2017) has aggregated over 15,000 WGS and over 125,000 WES datasets. Analyses of this rich resource discovered different types of variants. The potential functional impacts of these variants are also revealed, which help to identify driver variants and to prioritize therapeutic strategies. Another widely used resource is dbSNP (Sherry et al., 2001) that has been processed uniformly and collects a broad scope of repositories of 'small' genetic variation. COSMIC (Forbes et al., 2006) is a reference set of mutations that have been discovered in cancer genomes. Similar to dbSNP, COSMIC is also a mixed bag of various studies. COSMIC is especially powerful in tackling cancer genomes. COSMIC contains both germline and somatic driver variants, corresponding host genes and cancer types. If a germline or somatic variant is found in COSMIC, it is more likely to be a driver variant.

Once a mutated (or recurrently mutated) gene is identified, one may start to build biological and mechanistic hypotheses of what the mutation may be doing. One way of achieving that is by mining medical variant databases. ClinVar (www.ncbi.nlm.nih.gov/clinvar/), a public archive reporting association be-

tween genomic variants and diseases, classifies variants with clinical significance as disease-causing variants based on supporting evidence. OMIM (http://omim.org) is another excellent resource that integrates expert-curated and experimental verified associations between genes and diseases.

To figure out if a gene is activated or repressed by a specific mutation, one can look at the pathway network context. Gene Ontology (GO) (Consortium et al., 2001), KEGG (Kanehisa et al., 2004) and REACTOME (Fabregat et al., 2018) may help determine whether a set of genes contributes to specific functions or pathways. Besides functional consistency, interactions between candidate genes and known disease genes can also be considered. The STRING database (Szklarczyk et al., 2019) is a powerful resource to address this problem (http://string-db.org), which contains direct (physical) and indirect (functional) interactions between proteins of 5090 organisms.

2.2.3 Limitations of standard RNA-seq and DNA-seq analysis pipelines

In the standard variant calling pipelines, genomic variants are identified by comparing aligned reads to the reference. Reference genomes are produced from a single individual or a limited group of individuals. This single reference does not represent the genomic diversity and polymorphism of a population. This results in reference bias, where reads from polymorphic regions are not handled correctly and are either misaligned or discarded as unmapped reads. Incorrect alignments, in turn, lead to false variant calls.

Several studies have underlined the drawbacks of relying on a reference. Lunter and Goodson (Lunter and Goodson, 2011) pointed out the bias when dealing with INDELS. They tested different aligners such as BWA, MAQ, and Stampy on a heterozygous INDELS. All aligners underestimated indel proportions. Degner and colleagues (Degner et al., 2009) also aligned reads containing heterozygous SNPs

to the reference genome using MAQ and found a significant imbalance between reads with and without the reference allele. Reads containing complex variants are particularly challenging. For instance, structural variants may span more than 100 kb on the genome. To find these events, variant callers rely on detecting patterns of discordant read pairs or split reads, which highly depends on the alignment accuracy. Reads with multiple mutations or gaps are also challenging as their alignment is highly dependent on aligners' penalties for gaps and mismatches.

Most organisms on earth still do not have an available reference genome. Although many species' genomes have been sequenced and made available, these genomes are often incomplete. In humans, for instance, the X chromosome was the only complete human chromosome until 2020 (Miga et al., 2020) and the first complete human genome including centromeres and telomeres has just been released in May 2021 (Nurk et al., 2021) and is not yet used as a reference. The latest official human reference genome (HG38) still contains many unsolved regions with low complexity or repeat sequences (Blaxter, 2010). Pathogenic variants within unsolved regions are missed when using mapping-based methods. Yet, these regions are functional. The recently completed genome introduces nearly 200 million bp of novel sequence containing 115 potentially protein coding genes (Nurk et al., 2021). Studies have proved the association between these regions and various diseases. For instance, the shortening of telomeres induces chromosomal instability and causes cancers (Mathieu et al., 2004).

Another drawback of using one unified reference is ignoring the diversity among populations. Personalized characteristics exist in each individual's genome, and a unified reference genome does not account for this diversity (Sherman et al., 2019). Sherman et al. analyzed 910 deeply sequenced African individuals and discovered almost 10% more DNA than the current human genome. A collection of diverse genomes cannot always cover this issue and is powerless for species without available reference. Using incomplete or unified references will thus lead to the loss of critical genetic information.

Different aligners do not produce the same results. The Alignathon project demonstrated that over 50% of alignments were inconsistent among 13 WGS aligners (Earl et al., 2014). Inconsistencies among aligners result in part from the algorithm and in part from parameters. So far, there is no standard parameter optimization protocol for mapping-based approaches. Neither default parameters nor alternative parameters can guarantee alignments are correct at all sites across entire genomes. Therefore, the quality of downstream analysis relying on the alignments may not always be guaranteed.

Incorrect alignment has a crucial impact on clinical interpretation. For instance, the human leukocyte antigen (HLA) genes have the highest diversity of any region in the genome. It means mapping reads from HLA genes to the correct positions is quite challenging for aligners. Brandt (Brandt et al., 2015) compared HLA genotypes from the 1000-Genomes project and found nearly 20% of SNPs identified by NGS are incorrect. These inaccurate variant-calling results lead to erroneous HLA genotyping, an essential part of the management of autoimmune diseases and organ transplant rejection.

While variation in healthy genomes is a challenge to aligners, the situation is even worse in cancer genomes (Stratton et al., 2009). Genomic instability is the primary characteristic of most cancers (Jackson and Loeb, 1998). Cancer genomes always exhibit significant variation in the number and order of genetic elements due to their high mutational frequency, frequent recombinations, and high heterogeneity. Viral genomes also suffer from similar issues. A virus genome changes due to mutations, horizontal gene transfers, and gene gains/losses (Duffy et al., 2008). As cancer progresses or virus evolves, their genomes become quite different from the reference genome, which of course impacts alignment.

Yet another problem with mapping-based protocols is related to computer resources: standard pipelines are memory- and time-consuming. Typically, indexing the human genome requires nearly 5GB of memory and takes 3 hours of CPU time. Mapping and variant calling of a WGS dataset take about 100 hours of CPU-time with

the leading pipeline. CPU time can be reduced by multi-threading, but at the cost of greater memory consumption.

Researchers have of course noticed the limitations associated with reference genomes and have proposed solutions. One approach is to update the reference genome so that it captures the diversity of the entire population. Mapping and variant calling accuracy improves if reads are aligned to a representative collection of genomes (pan-genome) rather than a single linear genome (Victor et al., 2018; Sherman and Salzberg, 2020). The pan-genome is defined as the combination of genomes, containing all representative variants that occur in a species. By sampling a diverse set of individuals, a generalization of such a representation can be assembled. Efforts have been made to improve the reference through the addition of alternative loci scaffolds and haplotype sequences (Ballouz et al., 2019). The current human genome, GRCh38, includes alternative loci for genomic regions with high diversity. However, using these improved reference genomes is challenging since most aligners still lack the ability to take into account alternative loci. Furthermore, pan-genome sequences are often padded by long stretches of bases that are identical to the primary assembly. This is a problem when using an aligner that is not alternate-locus aware. Reads will get low MAPQ as they can map to the primary assembly and the alternative loci. One is unable to call variants from such regions, as reads that can map to multiple positions are usually dropped by aligners due to the low MAPQ.

Alternatively, a diverse collection of genomes can be represented using pan-genome graphs (Paten et al., 2017; Li et al., 2020), where each individual genome is identified as a path in the graph. Polymorphic regions create bubbles indicating diverse genotypes at the corresponding position in the entire population. A general way to construct a pan-genome graph is to generate a **c**ompacted **D**e **B**ruijn **G**raph (cDBG) composed of different genomes (Beller and Ohlebusch, 2016). The data structure of cDBG is introduced in Section 2.3.3. The *de novo* assembly algorithms use colored cDBG where different colors represent genome paths specific to individual populations (Iqbal et al., 2012).

Finally, the accuracy and reliability of alignments depend on software parameters. These parameters, including substitution matrices, gap penalties and cutoffs for mutation P-values, are either arbitrarily set by users or left in default mode, which in any case impacts alignments (Wong et al., 2008). Despite awareness of this problem, the necessity of optimizing alignment parameters is controversial. Even though parameters can be tweaked to optimize alignments, no one can be sure the "better-looking" alignment is the correct one.

Nevertheless, standard pipelines still have dominant positions in many aspects, particularly in major cancer genomic projects such as Pan-cancer whole-genome (PCAWG) analyses (The et al., 2020; Priestley et al., 2019). Standard pipelines enabled systematic documentation of genetic changes at the whole-genome scale and help discovered cancer drivers. The PCAWG analyses of 2,658 whole-cancer genomes also reported that no drivers were identified in 5% of tumor patients, suggesting that cancer driver discovery conducted by standard pipelines is not complete yet (The et al., 2020).

2.3 Mapping-free approaches

In the previous section, we showed that alignment to a reference limits biological discovery since it does not account for the full diversity of genomes and transcriptomes, and is computationally expensive in most instances. Mapping-free approaches address these limitations. Mapping-free approaches refer to any NGS analysis method that does not rely on the alignment to reference sequences. The main mapping-free approaches are applied in many fields including reads assembly and read contents analysis.

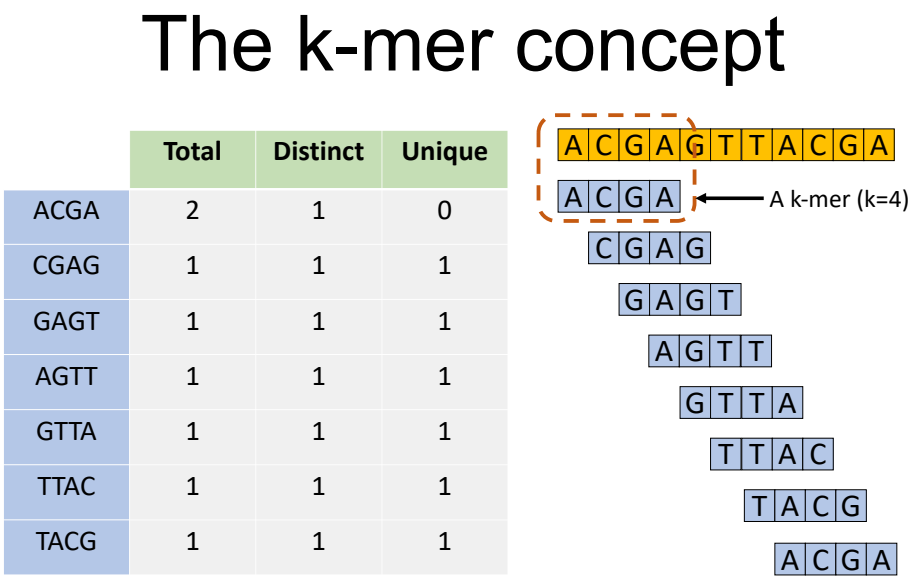


Figure 4: The concept of k-mers. Each k-mer is a substring of a read with the equal length of k.

2.3.1 k-mer approaches

An important class of mapping-free methods relies on the concept of k-mers. A k-mer is a subsequence of fixed size k. For instance, Figure 4 shows a sequence of length 11 decomposed into seven 4-mers. Generally, the next step is to count the k-mers. K-mer counts can be used to quantify expression, estimate copy numbers, or provide statistical support for variant callings. Quantification estimates are much simpler and faster with k-mers than with sequencing reads. K-mer counts can be easily computed and queried from raw sequencing reads instead of aligning to a reference. The k-mer counting procedure is the basis of various mapping-free applications. We introduce the major k-mer counting strategies in the next section.

2.3.2 k-mer counting strategies

While k-mer counting may seem an easy task in the above figure, a genome is definitely much longer and an NGS sequencer output is even longer. Efficient tools have been developed to address k-mer counting from large sequence files. These tools can be roughly classified based on their implemented strategies: sorting, hashing, enhanced suffix array, and memory- or disk-based.

By sorting all k-mers extracted from the dataset, identical k-mers are located at adjacent positions and thus can be easily counted. Tools based on the sorting approach include GenomeTester4 (Kaplinski et al., 2015), Turtle (Roy et al., 2014) and scTurtle (Roy et al., 2014). These tools gather k-mers from the input data and store them in a temporary array. K-mers are counted after sorting. These tools also work slightly differently in terms of memory management. For instance, Turtle compacts identical k-mers in a bucket to free up space. scTurtle also intends to reduce memory usage. It applies a bloom filter to remove all k-mers with a single occurrence before sorting and compaction. A Bloom filter is a space-efficient probabilistic data structure designed to tell whether an element is present in a set in a rapid and memory-efficient way. Bloom filters have much lower space complexity than hash tables (Jiang et al., 2018).

Another k-mer counting strategy involves using a hash table where k-mers are stored as keys, and counts are stored as values. A representative tool in this class is Jellyfish (Marçais and Kingsford, 2011). For each k-mer, Jellyfish first searches for it in the hash table. If the k-mer exists, its count is incremented; otherwise, the k-mer is inserted into the hash table with a count of one. Jellyfish works with multi-threads and writes the hash table to disk when the hash table is full instead of doubling its size in memory. A later, more efficient version of Jellyfish called Jellyfish 2 applies a bloom filter to remove k-mers with a single occurrence.

Suffix arrays (Abouelhoda et al., 2004) are another approach to save memory. This strategy first divides the sequence into smaller distinct partitions, which are further

decomposed into k-mers. The k-mers in each partition are counted based on the longest common prefix constructed from an enhanced suffix array. Finally, a final count is calculated by merging all distinct partitions. A representative tool in this class is Tallymer (Kurtz et al., 2008), which was designed to annotate large plant genomes.

In contrast to these in-memory algorithms, disk-based approaches, of lower memory requirement, were developed to manipulate large genomes such as the human genome. Sequences are divided into k-mers and processed in chunks that are further stored on disk to free up memory. Generally, disk-based approaches first estimate the number of chunks, the size of the hash table based on disk size and the number of k-mers. In the next step, k-mers are assigned to a hash table. Once the hash table is full, k-mers are counted and dumped to disk so the memory is released and new chunks can be loaded iteratively. Disk-based k-mer counting tools include KMC2 (Deorowicz et al., 2015), KMC3 (Kokot et al., 2017), KAnalyze (Audano and Vannberg, 2014) and DSK (Rizk et al., 2013).

K-mer counting is a critical step in mapping-free approaches. A number of bioinformatics problems are addressed based on k-mer counting, including expression quantification, copy number estimation and variant calling. As NGS technologies develop, projects with tens of thousands of samples become commonplace. Counting and storing such a large volume of data is a great challenge with the standard pipelines. Memory- and time-efficient k-mer counting approaches provide opportunities to deal with large-scale genomic datasets with limited memory and disk size.

The data volume of samples sequenced so far is of great challenge to computers' storage and processing capacities. As the next-generation and third-generation sequencing technology develop, the analytic capabilities are confronting a considerable challenge due to the computationally intensive alignment step. Mapping-free approaches significantly improve speed and applicability regarding various analyses, including expression profiling, genetic variant calling, *de novo* genome assembly,

phylogenetic construction, and taxonomic classification in metagenomic studies.

2.3.3 Applications of mapping-free approaches

As an alternative to the main pipeline presented in RNA-seq and DNA-seq, if a reference genome or transcriptome is not available or cannot be trusted due to large variations, sequencing reads can be assembled *de novo*. Assembly tools for DNA-seq reads include ABySS (Simpson et al., 2009) and SOAPdenovo (Simpson and Durbin, 2012). Assembly tools for RNA-seq reads include StringTie (Pertea et al., 2015), Trinity (Grabherr et al., 2011), SPAdes (Bankevich et al., 2012), Trans-ABySS (Robertson et al., 2010), Bridger (Chang et al., 2015) and SOAPdenovo-Trans (Xie et al., 2014). These *de novo* transcript assembly tools are particularly useful not only when the reference is missing or incomplete, but also where aberrant transcripts (for example, in tumour tissue) are of interest. Because *de novo* transcriptome assembly can recover transcripts that are transcribed from regions missing from the known reference (Mittal and McDonald, 2017; Haas et al., 2019). Novel transcripts can be discovered that are crucial in cancer research.

The **D**e **B**ruijn **G**raph (DBG) (Pevzner et al., 2001) data structure is widely used by these assemblers and reduces the computational charge by breaking reads into k-mers (words of size k) and building a directed graph representing overlaps between k-mers (Miller et al., 2010). In this type of sequence management, the k-mers are aligned against each other to obtain contiguous genomic sequences. The DBG is traversed to identify the maximal non-branching paths, which are also called unitigs. Then all unitigs are compacted into a single vertex. Finally, a **c**ompacted **D**e **B**ruijn **G**raph (cDBG) is constructed where nodes are unitigs and edges correspond to (k-1)-overlaps between two nodes sequences (Chikhi et al., 2016; Marchet et al., 2020). Compaction is an important data reduction step in most DBG based algorithms. Once a cDBG is built, reads can be aligned to the graph and depth can be computed for each path. During this procedure, one can also eliminate unsupported paths that have no depth at all and may derive from the misassemblies.

There are important limitations in *de novo* assembly software. Benchmarking studies comparing different assemblers show generally low genome integrity, high degrees of fragmentation and low contig accuracy (Sutton et al., 2019; Limasset et al., 2016) with these software. These limitations are particularly severe in situations of low sequencing depth and the presence of genomic repeats. For instance, the human genome has many repeats longer than NGS reads so that these repetitive regions cannot be correctly resolved. Furthermore, sequencing errors are inevitable during the sequencing process. And any sequencing error may introduce new branches in the assembly process. Genomic repeats and sequencing errors thus increase the complexity of DBG, which impacts memory requirements as it essentially depends on the size of the DBG. Thus, *de novo* assembly is computationally costly and has never been applied to large-scale studies.

A first-class of mapping-free approaches applied in RNA-seq data are so called "pseudo-alignment" or "pseudo-mapping" methods aiming at transcript quantification (see also Chapter 2.2.1.3). These software first create a k-mer index from the transcriptome and then estimate the expression of a read using “pseudo-alignment”. A read’s pseudo-alignment consists of finding the transcript(s) that the read is compatible with. Contrary to normal read alignment, where aligners specify where each read aligns, pseudo-alignment is done via a transcriptome DBG. Each node in the DBG is a k-mer associated with a transcript or set of transcripts, which is described as a k-compatibility class. In other words, a transcript that contains a node k-mer would belong to the k-compatibility class of that node. To find the transcript(s) a read is compatible with, the read is decomposed into k-mers, and k-mers are hashed to different nodes in the DBG. Then we take the intersection of all k-compatibility classes that a read is associated with. Finally, the possible transcript(s) a read comes from are inferred. Grouping reads belonging to the same transcript allows estimating the expression levels of each transcript. This pseudo-alignment-based quantification method is 10-100 times faster than standard alignment-based methods and achieves high consistency with the best performing alignment-based software (Everaert et al., 2017). Even though the reads include

new transcripts coming from fusions, mutations, new splicing or novel genes, quantification of the known genes won't be influenced. An index of known genes is first constructed by Kallisto or Salmon (Bray et al., 2016; Patro et al., 2017) based on the transcriptome and then the reads are assigned to transcripts of known genes based on the index.

Metagenomics has become a primary application of mapping-free approaches. The two software with the highest accuracy and sensitivity for profiling microbial communities are Kraken (Wood and Salzberg, 2014) and CLARK (Ounit and Lonardi, 2016), which work in a manner similar to pseudo-alignment. These methods first construct a taxonomy tree from a pre-computed database and then map k-mers from NGS reads to sequences in the tree, where taxa associated with the read form a pruned subtree of the general taxonomy tree. Taxonomic labels are assigned to individual reads applying a lowest common ancestor (LCA) criterion. This classification procedure can be performed in large datasets with excellent accuracy, even with unknown organisms.

Another mapping-free software for transcriptome analysis is DE-kupl, which was developed in our lab in 2017 (Audoux et al., 2017b). DE-kupl aims at capturing all k-mer variations in raw RNA sequencing data in a differential analysis setup. DE-kupl is composed of five main steps: indexing, filtering, differential expression, k-mer extension and annotation (if a reference genome is provided). First, DE-kupl applies Jellyfish to index and count k-mers. Second, DE-kupl filters out the low abundance k-mers representing potential sequencing errors. Then k-mers that are differentially expressed between two conditions are selected by applying the limma or DESeq2 methods (Law et al., 2014; Love et al., 2014) using a pipeline specially adapted to very large data tables. Selected k-mers are assembled into contigs, which are further annotated by alignment to the reference genome. Since DE-kupl is reference-free, it enables the capture of any novel RNA or RNA isoform present in the data at nucleotide resolution, including unmappable transcripts such as RNAs from repeats and chimeric RNA. Contigs are further annotated and classified into different event categories, such as SNV, splice, intron, polyA, split, repeat,

lincRNA, and unmapped. Audoux et al. validated the reproducibility of DE-kupl using two independent human RNA-seq data sets from the Genotype-Tissue Expression (GTEx) (Lonsdale et al., 2013) and the Human Protein Atlas (HPA) (Uhlén et al., 2015). Nearly 80% of the top differential k-mers identified by DE-kupl were consistent between two datasets. Pinskaya et al identified novel unannotated lncRNAs forming a signature of prostate cancer using DEkupl (Pinskaya et al., 2019). In this thesis, we applied DE-kupl to analyze two lung cancer cohorts and identify lung cancer associated events (see Chapter 3.1).

iMOKA (Lorenzi et al., 2020) is another mapping-free software that enables comprehensive analysis of transcriptome from large cohorts. While DE-kupl finds differential events between two conditions, iMOKA aims at making diagnostic and prognostic classifiers. Some steps in the iMOKA workflow are similar to DE-kupl. In the first step, iMOKA uses KMC3 to index and count k-mers instead of Jellyfish (see Chapter 2.3.2 for differences). Then iMOKA screens k-mers using a Bayes classifier, which evaluates each k-mer individually. In the next step, screened k-mers are assembled into graphs. Component k-mers in a graph are likely generated by the same biological event. Next, iMOKA annotates the k-mer graphs by comparison to a reference genome as in DE-kupl. The final list of highly informative k-mers can be explored via a user interface. Finally, iMOKA provides a random forest classifier that uses filtered k-mer graphs as features. Users can also build a random forest classifier based solely on specific genomic features such as mutations or gene expression. Lorenzi et al applied iMOKA to classify breast cancer subtypes and identify events associated with the response to treatment in ovarian and breast cancer.

KISSPLICE is a reference-free software that extracts AS events from RNA-seq data (Kielbassa—Pavlos et al., 2011). KISSPLICE first constructs a cDBG. Sequence and splicing variations in transcripts generate bubble structures in the cDBG. The KISSPLICE algorithm detects all the bubble patterns in the cDBG. Then the detected candidate bubbles are processed by filtration and classification. Bubbles generated by SNPs exhibit two branches of equal length. Bubbles generated by

AS are characterized by two common sites and a variable part. The common sites are shared by different isoforms and the variable part indicates an AS. Genomic indels generate bubbles with similar branch lengths as bubbles generated by splicing events. Finally, sequencing reads are mapped to each branch of the bubble to estimate the read coherence and depth. Even though KISSPLICE was initially designed for AS detection, it was updated for SNP calling in RNA-seq specifically in the more recent version (Lopez-Maestre et al., 2016). Two new modules of KISSDE and KISSPLICE2REFTRANSCRIPTOME (K2RT) were introduced in the latest version of KISSPLICE. KISSDE finds condition-specific SNPs and K2RT predicts the amino acid change. KISSPLICE is the first approach toward transcriptome-wide association studies in non-model species.

Mapping-free approaches are also used for genomic variant calling. Mapping-free approaches allow direct variant genotyping from sequencing data and are 1-2 orders of magnitude faster than the standard mapping-based pipelines. Improved speed and high accuracy make these mapping-free approaches ideal for clinical use, where large numbers of samples need to be processed in a timely manner. Mapping-free variant calling tools include DISCOSNP (Uricaru et al., 2015), FastGT (Pajuste et al., 2017), LAVA (Shajii et al., 2016) and MICADo (Rudewicz et al., 2016). Here we present DISCOSNP as an example. DISCOSNP can detect isolated SNPs from raw sequencing data. It is composed of two independent modules, KISSNP2 and KISSREADS. KISSNP2 detects putative SNPs based on a DBG. Each SNP generates a bubble structure in the DBG containing couples of paths of length 2k-1. KISSREADS filters out bubbles covered by no read. These bubbles come from situations where a sequence is mapped only by the beginning or end of reads. Such sequence in a DBG path is generated by sets of k-mers that do not pertain to the same read and thus form a chimeric sequence. KISSREADS also adds depth and quality information on the remaining bubbles. DiscoSnp++ (Peterlongo et al., 2017) is an extension of DISCOSNP, which only focuses on isolated SNPs. DiscoSNP++ runs much faster and uses less memory than DISCOSNP. DiscoSnp++ is designed for detecting not only SNPs but also small indels from raw sequencing

data. In Chapter 3.2, I will present our own software for calling somatic variants from raw sequencing data, called "2-kupl" and compare it with DiscoSNP++.

Phylogenomics is another area where mapping-free approaches play an important role. Mapping-free methods enable direct phylogeny construction from raw sequencing data, regardless of genome assembly and alignment. These methods include AAF (Fan et al., 2015), NGS-MC (Ren et al., 2016) and kSNP (Gardner et al., 2015). Mapping-free methods have been widely used to infer phylogenetic relationships among eukaryotes, with resulting trees that were extremely close to trees created from manually curated NCBI taxonomic databases (Criscuolo, 2019).

2.3.4 Limitations of k-mer and mapping-free approaches

K-mer based approaches analyze all sequences in an NGS dataset, including those overlapping complex SNVs, SVs and repeats. From this perspective, these approaches have no false negatives; they retain all biological information. However, it is sometimes difficult to highlight k-mers of biological interest from the noisy background. Taking into account k-mer counts and base quality to distinguish *bona fide* variants from sequencing errors is a common challenge of these software. Rcorrector is a software that aims at correcting sequencing errors from RNA-seq data (Song and Florea, 2015). Rcorrector uses a DBG to compactly represent all trusted k-mers whose occurrences in the input reads exceed a given threshold. Given the non-uniform distribution of RNA-seq data, Rcorrector estimates a local threshold at every position in a read. However, Rcorrector does not take into account the base quality of input reads. Quake is another error correction program that incorporates quality values and rates of specific miscalls to detect and correct sequencing errors in DNA sequencing reads (Kelley et al., 2010). Quake detects many reads potentially containing sequencing errors but it cannot find a valid set of corrections and pinpoint the errors' locations. These reads will lead to underestimated estimation of error probabilities. Adequate sequencing depth is needed for Quake to decide erroneous and genuine k-mers (e.g. >15X). Thus Quake does not make convincing

corrections in low depth regions.

The dissimilarity between two NGS samples calculated based on alignment-free methods is most likely overestimated compared to genome-based calculation (Tang et al., 2019). The overestimation tends to be more severe in NGS datasets with low sequencing depth. In real cases, feature vectors from two NGS samples differ due to the stochastic distribution of reads along the genomes. Therefore, the measured dissimilarity between two NGS samples sampled from the same genome can be greater than zero. The bias introduced by overestimated dissimilarity between NGS samples is a common problem for all alignment-free methods since it results from the intrinsic stochastic distribution of short reads.

Pseudo-mapping RNA-seq quantification also has limitations. A study has compared the accuracy of pseudo-mapping methods Kallisto (Bray et al., 2016) and Salmon (Patro et al., 2017), mapping-based methods HISAT2+featureCounts (Kim et al., 2015) and a customized pipeline TGIRT-map, for the quantification of various transcripts (Wu et al., 2018). The study showed that both mapping-based and pseudo-mapping quantification approaches performed similarly for most protein-coding genes. However, accuracy was lower with Kallisto and Salmon for lowly-expressed genes or small RNAs. Kallisto and Salmon did not perform well for quantifying short genes with abundant biological variations and tended to underestimate their expression relative to mapping-based approaches.

In the field of DNA-seq variant calling, mapping-free methods do not outperform mapping-based methods for variant detection, even though the former were able to detect specific variants, such as SVs (Khorsand and Hormozdiari, 2021). A strength of mapping-based methods is they can use more information to call a variant. Indeed, alignments to references provide information such as genomic positions, variant types, nucleotide/amino acid substitutions and related database records, and potential functional effects. Besides, mapping-based approaches can use haplotype information, such as in the haplotypecaller algorithm used by GATK, which enables strong checks for false positive SNVs.

While mapping-based approaches are still the leading solutions for most biological problems, the development of mapping-free approaches is still in its infancy and holds considerable potential for improvement. Most published articles about mapping-free approaches are evaluated with individually selected and simulated datasets. The absence of well-defined benchmarks covering various genetic events and sequence divergence prevents users from choosing the best tool. In contrast, mapping-based approaches benefit from well-validated benchmarks.

Nevertheless, mapping-free algorithms rapidly extend their application range. Issues of sequencing data processing and storage seem to be particularly well addressed by the mapping-free methods and this is increasingly important with the development of sequencing technologies, as larger sequence datasets are generated and need to be processed and stored. With their higher computational efficiency and capacity to detect a wider diversity of variants, mapping-free approaches are becoming exciting complements to standard mapping-based pipelines.

2.4 Thesis objectives

My general goal in this thesis was to exploit the power of reference-free approaches to discover novel variations in cancer transcriptomes and genomes. We aimed to discover novel events from unannotated and difficult-to-map regions that potentially play causal roles in tumorigenesis. When the thesis started, my host laboratory had recently published DE-kupl, a reference-free software for the discovery of differential events between sets of RNA-seq of experiments (see Chapter 2.3.3). DE-kupl had been used only on small datasets. My first goal was to test its ability to find differential events in a large cancer dataset of several hundreds of samples and to assess the replicability of found events.

We then questioned whether a similar approach could be applied to variant analysis in the context where two samples need to be compared (a normal and a mutated sample). This is a very common problem, yet no reference-free software existed for

this specific question. We also started from the premises of DE-kupl (*i.e.* performing k-mer counts in the samples to be compared) but the following part needed to be completely remodeled since the problem was not anymore quantitative as in RNA-seq analysis, but qualitative, that is to find events that were specific to the mutated/tumor sample. We thus developed a specific pipeline to address this particular aspect. The resulting software, 2-kupl, is presented in Chapter 3.2. Since this is a novel software, we paid special attention to the benchmark part. We involved different datasets from bacteria and humans and different software in the comparisons.

Chapter 3

Results

3.1 The contribution of uncharted RNA sequences to tumor identity in lung adenocarcinoma

Yunfeng WANG[12], Haoliang Xue[1], Marine Aglave[1], Antoine Lainé [1], Mélina Gallopin[1], Daniel Gautheret[1*]

* Correspondence: daniel.gautheret@universite-paris-saclay.fr

[1]Institute for Integrative Biology of the Cell, CEA, CNRS, Université Paris-Saclay, Gif-Sur-Yvette, France.
[2]Annoroad Gene Technology Co., Ltd, Beijing, China

3.1.1 Contribution

Our group previously developed a mapping-free software, DE-kupl, that processes raw sequencing data of RNA-seq and detects transcriptional events associated with a specific phenotype. The objective of this paper was to evaluate its ability to find reproducible differential events in a large-scale dataset. To do so, I compared the

results of differential analyzes performed at gene-level and using DE-kupl in different cancer datasets. I used two lung cancer datasets and one prostate cancer dataset as a control. Part of my development was aimed at comparing DE-kupl contigs found in different datasets. While comparing gene lists from different analyses is trivial, comparing unmapped contigs is not since there is no identifier to link two contigs. Therefore, I constructed a graph-based approach to decide if two contigs correspond to the same local event. If the answer is yes, this contig is considered as shared between two datasets.

The result demonstrates that DE-kupl identifies replicable sets of gene- and contig-level events between cohorts with the same phenotype regardless of sequencing platforms and donor populations. We also investigated in depth transposable elements (TE) and other novel events detected by DE-kupl. Lung cancer subsets with different levels of TE expression, possibly due to genome instability, were identified. Candidate neoantigens were screened from intron and intergenic regions. We also showed the association between various transcriptional events and the overall survival prognosis. In summary, we revealed a diversity of tumor-specific RNAs that could not be identified by standard mapping-based approaches. We proved DE-kupl as a stable and robust method for the exhaustive capture of transcriptional events.

3.1.2 Introduction

Over a period of 20 years, cancer transcriptomics has transformed our understanding of tumor biology and led to improved tools for tumor typing, diagnostic and outcome prediction (Gollub and Prowda, 1999; Parker et al., 2009; Margolin and Lindblom, 2006). While first generation transcriptome analysis was based on DNA microarrays with a focus on protein-coding genes, the current generation relies on RNA-seq data, which promises to deliver a more comprehensive view of gene expression. However, in spite of its potential for transcript discovery, cancer RNA-seq data is still utilized mostly to quantify the expression of annotated genes listed in a reference transcriptome. This ignores a wide array of mRNA isoforms, non-coding

RNAs, endogenous retroelements and transcripts from exogenous viruses and bacteria (Morillon and Gautheret, 2019). The quantity of information left unexploited in non-canonical transcripts remains unknown. A number of studies have started to address this question using publicly available cancer RNA-seq data, focusing on specific transcript classes such as splice variants (Kahles et al., 2018; Vitting-Seerup and Sandelin, 2019), lncRNAs (Iyer et al., 2015), snoRNAs (Gong et al., 2017), bacterial RNA (Ouchenir et al., 2017) or viral RNA (Zapatka et al., 2020). Other neglected sources of RNA diversity are the so-called blacklisted regions of the genome that are too variable or repeated to be properly analyzed by conventional approaches (Amemiya et al., 2019). To our knowledge, no attempt has been made to extract and evaluate at once all this non-standard RNA information from tumor RNA-seq data. We think this approach could be particularly valuable in cancer, since every individual tumor harbors a unique transcriptome that departs from that of normal tissues in multiple, unpredictable ways.

Recently we introduced a computational method, DE-kupl (Audoux et al., 2017b), that performs differential analysis of RNA-seq data at the k-mer level. As this method is reference-free and mapping-free, it identifies any novel RNA or RNA isoform present in the data at nucleotide resolution, including poorly mapped transcripts such as RNAs from repeats and chimeric RNA. Here we set ourselves to evaluate all non-reference events discovered by DE-kupl in a comparison of normal vs. tumor samples using **lu**ng **ad**enocarcinoma (LUAD) as a test case. To mitigate false positives events inherent to gene expression profiling (Ioannidis, 2005; Michiels et al., 2007), we focused on events that were replicated in two independent datasets. This required the development of a dedicated protocol to identify shared events in unmapped RNA sequences. Results revealed a collection of novel tumor-specific unannotated lincRNAs, intron retentions and splicing events. Most strikingly, a collection of endogenous retroelements (EREs) form a major class of tumor defining transcripts. We also identified a subset of events with no expression in normal tissues which could be candidate neoantigens. Finally, we identified a set of transcript variants potentially related to survival. We would like to suggest

DE-kupl as a promising, comprehensive approach to cancer transcript profiling.

3.1.3 Materials and Methods

3.1.3.1 Datasets

LUAD-TCGA: 582 lung RNA-seq samples were downloaded from the TCGA database with permission, including 524 lung adenocarcinoma (LUAD) tissues and 58 adjacent normal tissues (Network et al., 2014). LUAD-SEO: The LUAD RNA-seq dataset of Seo et al. (Seo et al., 2012) was downloaded from the SRA database (accession: ERP001058). This dataset contains fastq files of 87 LUAD and 77 adjacent normal tissues. Only the 77 paired normal and tumor samples were analyzed. PRAD-TCGA: For control, 557 Prostate RNA-seq datasets were downloaded from TCGA with permission, including 505 prostate adenocarcinoma (PRAD) and 52 normal controls (Abeshouse et al., 2015). For the TCGA datasets, raw bam files were converted to fastq format files using Picard tools (version of 2.18.16).

3.1.3.2 DE-kupl pipeline

DE-kupl was applied to the three datasets with the same parameters: in the filtering steps, k-mers with abundance fewer than 5 (min_recurrence_abundance) and present in no more than 10 samples (min_recurrence) were ruled out, and k-mers exactly mapping to the main transcript of each gene were removed as in the original DE-kupl procedure (Audoux et al., 2017b). In order to focus on non-canonical transcripts, we masked all k-mers pertaining to the main transcript of each Gencode gene as in (Audoux et al., 2017b). Normalization factors for k-mer counts were computed as the median of the ratios of sample counts by counts of a pseudo-reference obtained by taking the geometric mean of each k-mer across all samples. In the following, we will use these counts as a proxy to represent the expression of the corresponding RNA fragment.

For DE analysis, the version of DESeq2 available at the time of the experiment was too slow for dealing with hundreds of samples and we found the faster "T-test" option to lack sensibility. Hence we applied Limma Voom (Ritchie et al., 2015) to millions of k-mers using a chunk-based strategy (suppl. methods). This was found to perform 10 times faster than DESeq2. The performances of DESeq2, Limma Voom and T-test for DE evaluation have been evaluated before (De Paepe, 2015). Evaluations of k-mer counts were log-transformed and Limma Voom was used to calculate log fold-changes and P-values. Retention thresholds for log2 fold changes and P-values were 1 and 0.05, respectively. All k-mers passing the filtering process above were merged into contigs and the contig table was saved as output. GC-contents in "up" and "down" contigs in the PRADtcga dataset were verified and did not present any bias (Table S1). High-quality contigs (topctg) were considered as contigs with counts>10 in at least 15% of the smallest class (Normal or Tumor).

Gene-level expression was measured using Kallisto v0.43.02 and Gencode v34 transcripts, followed by summing TPM values of transcripts from the same gene. Gene-level DE analysis was performed using Limma and the same normalization procedure as above. Downstream analyses were conducted using R version 3.5.2. Heatmaps were drawn using the complexHeatmap package (Gu et al., 2016).

3.1.3.3 Shared event identification

Contigs from distinct DE-kupl analyses were decomposed into their constituent k-mer lists and a graph was constructed using the NetworkX Python package (Hagberg et al., 2008), with k-mers as nodes and shared k-mers as edges. Contigs corresponding to the same local event are expected to form a fully connected subgraph or clique (Fig S1). We thus extracted all cliques to identify shared contigs. Hereafter we use the $\cap$ operator to represent contigs shared between two datasets.

3.1.3.4 Contig annotation

A uniform annotation procedure was applied to contigs from each independent analysis (LUADtcga, LUADseo, PRADtcga) and to shared contigs (LUADtcga ∩ LUADseo and LUADtcga ∩ PRADtcga). Initially, differential contigs were mapped and annotated with DE-kupl annotation (https://github.com/Transipedia/dekupl). Briefly, DE-kupl annotation maps contigs to the human genome and reports intronic, exonic or intergenic status, CIGAR string, IDs of mapped or neighboring genes, **D**ifferential **U**sage (DU) status. A new repeat annotation field ("rep_type") was added based on Blast alignments of contigs to the DFAM repeat database (Hubley et al., 2016) (see Suppl. Methods). The results of DEkupl-annot were then loaded into R and submitted to further filtering and classification into event categories. Firstly, a count filter was applied to retain only contigs with a count of 10 in at least 15% of the smallest class (Normal or Tumor). Then a set of criteria was applied to classify contigs into event classes comprising SNV, intronic, splices, split, lincRNA, polyA, repeat and unmapped, as described in Table S3. Since the TCGA datasets in this study are unstranded, antisense events were not called. Differential usage (i.e. the relative change in expression of a local event relative to the expression of the host gene) was also evaluated for each event mapped to an annotated gene. Besides the transcriptional categories, we also produced a new category of "neo", which includes contigs that are only expressed in tumor tissues but silent in normal tissues. All categories of contigs involved in this study were further selected from the topctg.

3.1.3.5 Functional enrichment on intron retention events

Candidate intron retention events were identified based on the DE-kupl DU P-value (computed by comparing the expression or the contig with that of the host gene). Significant pairs of intron retention and host gene were selected. To illustrate the biological functions of all these intron retention events with DU, we performed the Gene Ontology biological process analysis using the clusterProfiler R package (Yu

et al., 2012).

3.1.3.6 Sample clustering based on repeats

We used the K-means algorithm (MacQueen et al., 1967) to cluster LUAD patients into two main subgroups based on the expression of contigs matching AluSx, L1P1_orf2 and L1P3_orf2 repeats. Clusters were then analyzed for enrichment in clinical features, immune infiltration, **T**umor **M**utational **B**urden (TMB) and **C**opy **N**umber **V**ariants (CNVs). Clinical features and immune infiltration were included to analyze the potential differences between clusters. LUAD driver genes were retrieved from the COSMIC Cancer Gene Census (CGC) list (Sondka et al., 2018). Oncoplots were drawn using the maftools R package (Mayakonda et al., 2018). The estimated TMB for each patient was computed using the total number of non-synonymous mutations from the MAF file by 38 that is the estimated size of the whole exome. Thus the unit of TMB is the number of somatic mutations per megabase of interrogated exome sequence. The level 3 CNVs data was downloaded from TCGA, which provides a mean copy number estimate of segments covering the whole genome (inferred from Affy SNP 6.0). We visualized the CNV frequency distribution among 23 human chromosomes using the copynumber R package (Nilsen et al., 2012). The ratio of gain and loss for each patient was estimated by the fraction of segments indicating CNVs. Heatmap representations were produced with ComplexHeatmap (Gu et al., 2016).

3.1.3.7 Survival analysis based on event classes

Since the LUADseo dataset doesn't include survival information, we only performed the survival analysis based on the LUADtcga dataset. The clinical information including overall survival time and status was downloaded from the GDC portal (https://portal.gdc.cancer.gov/projects/TCGA-LUAD). We performed both univariate Cox regression and multivariate Cox regression on each event class to assess the prognosis value of the differential events. Survival analysis was performed using

the survival and survminer R packages (Therneau and Lumley, 2015; Kassambara et al., 2017). Hazard ratios (HR) and P-values were calculated for each contig. Contigs with HR>1 and P-value<0.05 were considered as potential risk factors. For multivariate Cox regression, contigs were initially selected by cox-lasso regression using the glmnet R package (Friedman et al., 2010) applied independently to each contig class. The multivariate model was then constructed using selected. Patients were divided into high and low-risk groups based on the median value of all risk scores for representation in Kaplan–Meier (KM) curves (Kaplan and Meier, 1958).

3.1.3.8 Unsupervised cluster analysis

We applied Principal Component Analysis (PCA) and hierarchical clustering to each event class. PCA analysis was performed with the factoextra R package (Kassambara et al., 2017). Heatmap views were obtained using ComplexHeatmap (Gu et al., 2016).

3.1.3.9 Sequence alignment views

To facilitate event visualization, we created "metabam" alignment files for tumor and normal tissues from each cohort. To this aim, we randomly sampled 1M reads from each fastq file of each subcohort and aligned the aggregated reads to the genome (GRCh38) using STAR (Dobin et al., 2013) with default parameters. BAM files were visualized using Integrative Genomics Viewer (IGV) (Robinson et al., 2011).

3.1.4 Results

3.1.4.1 Gene-level vs contig-level differential events

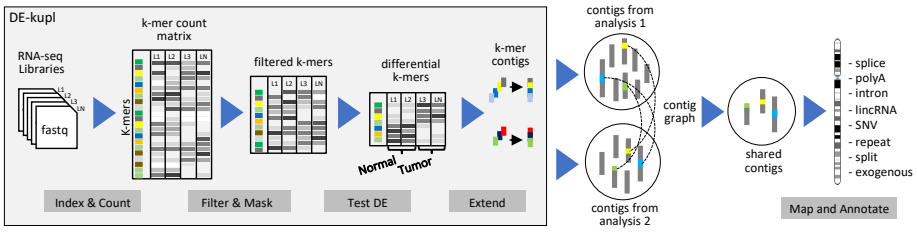
We performed tumor vs. normal differential expression (DE) analysis on two independent Lung adenocarcinoma RNA-seq datasets, from TCGA (LUADtcga) and Seo et al. (LUADseo) and on a prostate adenocarcinoma dataset from TCGA (PRADtcga) as a control. Each dataset was submitted to a conventional, gene-level, DE analysis and a k-mer level DE analysis where all k-mers from annotated genes were first removed and the resulting DE k-mers were assembled into contigs (Fig 1A).

While the number of differential genes in the three comparisons ranged from 6,000 to 9,000, the number of differential k-mers was about a thousand times larger (2 to 12 millions). Assembly of k-mers into contigs reduced this number to about 400,000 differential contigs in each analysis (Fig 1B). For simplification, we will use terms over-/under-expression when referring to contig counts with significant differences as per the DE analysis.

We next compared the DE genes and contigs discovered in independent datasets to identify shared DE events. While this process is trivial for genes, it is not for contigs, since contigs found in each dataset have no standard identifier that could be used to relate them. We thus implemented a graph analysis procedure that identified shared contigs based on their common k-mers (Fig 1A, Fig S1). A final annotation step identified contigs belonging to different categories (repeats, lincRNAs, splice variant, polyadenylation variants, split RNAs, tumor-exclusive or "neo" RNAs) as described in Table S3 and Methods. The numbers of shared elements slightly differ between LUADtcga and LUADseo because a minority of elements are in a 2-to-1 or 1-to-2 relationship in the contig graph. When not specified, numbers of elements will be given for the LUADtcga cohort.

Overall 160,610 differential contigs were shared between the two LUAD analyses

A



B

Study	LUADtcga	LUADseo	PRADtcga
# Normal samples	58	77	52
# Tumor Samples	524	77	506
DE genes	9 058	6 567	6 292
DE k-mers	6 069 791	12 102 802	2 459 131
DE contigs	368 257	449 860	439 062

C

$LUAD_{tcga} \cap LUAD_{seo}$	N	%
Shared contigs	160610	–
High quality	120822	100.0
Multi hit or repeat	50572	44.1
Up	100405	83.1
Down	20417	16.9
Genic	104762	86.6
Intronic	56146	44.9
SNV or indel	54581	32.1

D

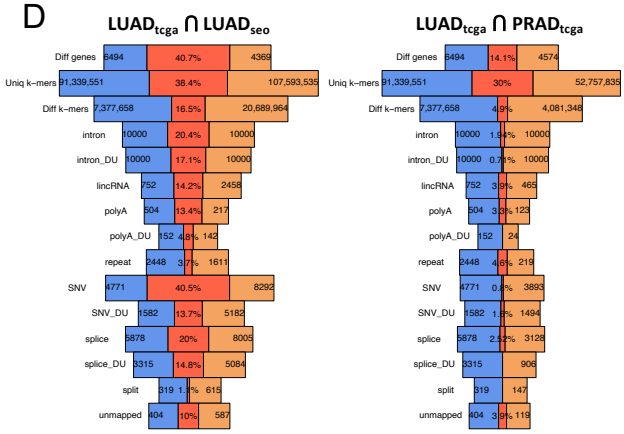


Figure 1: (A). Computational pipeline used to infer differential contigs in each tumor/normal cohort, followed by extraction of shared contigs and annotation. (B). Sizes of RNA-seq cohorts analyzed and numbers of differential events observed. (C). Summary statistics of differential contigs identified as shared between the LUADtcga and LUADseo analyzes. (D). Number of differential genes, k-mers and contigs in each independent analysis and shared between analyzes. On each row, lateral areas represent differential genes/k-mers/contigs found in each independent analysis and the central area represents shared differential genes/k-mers/contigs. Contigs are classified into different annotation groups.

(Fig 1C). Over these, 120,822 contigs were considered of sufficient quality based on counts and occurrence in a minimal number of samples (see Methods). 83% of shared contigs were overexpressed in tumors vs. only 17% underexpressed (Fig 1C).

3.1.4.2 Event replicability

The replicability of differential event was generally lower for k-mer or contigs than for genes. Fig 1D shows the number of DE genes and contigs shared by the two independent LUAD analyzes, with contigs binned by annotation class. About 44% of DE genes (3032 genes) were shared by the two LUAD analyses, compared to an average of 15% for DE contigs (repeats: 3.2%, unmapped RNAs: 10%, alternative polyAs: 13%, lincRNAs: 14%, alternative splices: 20%, retained introns: 20%). Although the ratio of shared events was relatively low for k-mer analysis, it was considerably higher than when comparing two unrelated pathologies (LUADtcga $\cap$ PRADtcga, Fig 1D), and this applied to all event classes. This indicates that, although k-mer based DE events are noisy, a significant subset is replicable in independent studies. Furthermore, we observed a strong correlation between the fold-change value of DE contigs and the likelihood to be shared between cohorts (Fig S2), demonstrating the non-randomness of high scoring, non-reference events.

3.1.4.3 DE contig localization, hypervariable genes

The majority of shared contigs are genic (83%) including intronic (45%) and 32% carry SNVs or INDELS (Fig 2). These characteristics are induced by the initial filter that removed k-mers matching reference transcripts, but retained any intronic or SNV-carrying k-mer. Therefore a large number of SNV and intronic contigs are just "passenger" events of DE genes.

More than 400 genes were matched by 35 or more contigs. We classified these genes into two categories: for 296 genes, most contigs matched introns and were up-regulated in tumors (Fig 2A, B, Table S5). These mostly correspond to the

aforementioned "passenger" events. The second category is composed of 107 genes we refer to as “hypervariable” as they tend to yield a large number of contigs carrying SNVs, INDELS and larger rearrangements (Fig 2A, C, Table S5). The largest class of hypervariable genes are IGK, IGL and IGH immunoglobulin genes, which was expected given their inherent variability due to V(D)J segment recombination and their expression by plasma B-cells which are abundant in the tumor immune infiltrate (Thorsson et al., 2018), hence are seen as up-regulated in tumors. Interestingly, those IG sequence variants are found expressed in different patients and across the two cohorts, suggesting our approach can be used to profile immunoglobulin repertoires, as performed recently with other RNA-seq datasets (Mandric et al., 2020). To evaluate the accuracy of DE-kupl contigs from IG genes, we selected all contigs mapped to one arbitrary IG gene (IGHV: 100 contigs) and aligned them to IGHV contigs from the International ImMunoGeneTics Information System (IMGT) (Lefranc et al., 2009). Ninety out of 100 contigs had significant matches in the corresponding IMGT category extending over 90% of the contig length (Table S6).

Other hypervariable loci were found in surfactant protein (SFTP) and Mucin genes which are known to harbor a high level of polymorphism (Imielinski et al., 2017; Swallow et al., 1987). We observed polymorphism not only in the form of SNPs, but also in the form of splicing variations. Five SFTP genes alone combine over 9000 SNVs and 800 splice sites contigs, while 12 Mucin genes harbour 1324 contigs including 42 splice variants (Fig S3A-B, Table S5). While SFTP contigs were all underexpressed in tumors, Mucin contigs were mostly overexpressed (Table S5). Mucins are immunogenic (Swallow et al., 1987) and are important biomarkers for prognosis (Ning et al., 2020) and drug resistance (Aithal et al., 2018). Therefore the existence of recurrent mucin variants overexpressed in tumors should be taken into account in the development of these biomarkers and therapies. Finally, we also observed hypervariability in CEACAM5 and KR19, two other prognostic biomarkers and/or immunotherapy targets (Wang et al., 2019; Thistlethwaite et al., 2017) (Fig S3C, Table S5).

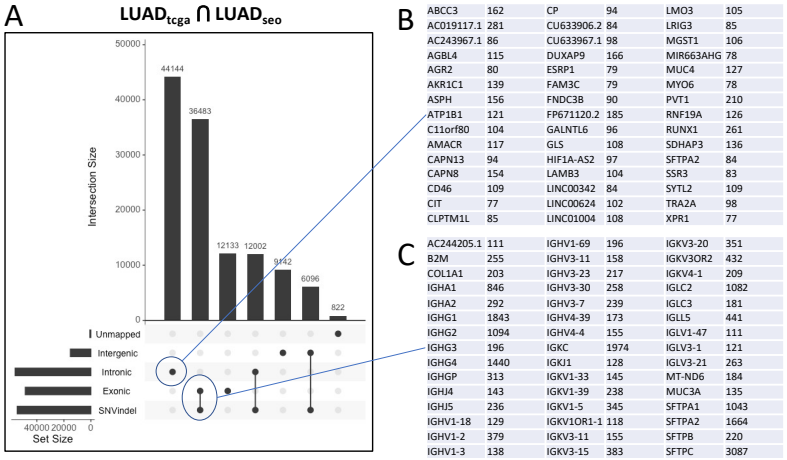


Figure 2: General properties of shared DE contigs in LUAD. (A) UpsetR plot of major contig categories based on mapping location and presence of SNV or INDELS. (B) 45 top genes by number of mapped contigs in the circled intronic category. (C) 45 top genes by number of mapped contigs in the circled exonic+SNVindel category. Numbers of contigs mapped to each gene are indicated.

3.1.4.4 Intron retention and other intronic events

We found intronic contigs with **D**ifferential **U**sage (DU) in 313 host genes, 290 (93%) of which were up-regulated in tumors (Table S4). 70% of the host genes were also up-regulated, thus the apparent overexpression of these intronic sequences may have been confounded by overexpression of host genes. However, 30% of host genes were not overexpressed, and in 103 cases, intron and host gene expressions varied in opposite directions (93 introns up and 10 introns down). Our annotation pipeline did not differentiate intron retentions (as shown for example in Fig S4) from transcription units occurring within introns (example in Fig S5). We observed intron retention events in lung cancer drivers EGFR and MET (Fig S6 and Fig S7). In EGFR, the retained intron was located between exons 18 and 19, just upstream of the principal oncogenic EGFR mutations located in exons 19-21. Intron retention before exon 19 would likely produce a truncated form of EGFR compatible with oncogenic activation. Fig 3A shows the 20 intronic events with the most significant DU P-values. All

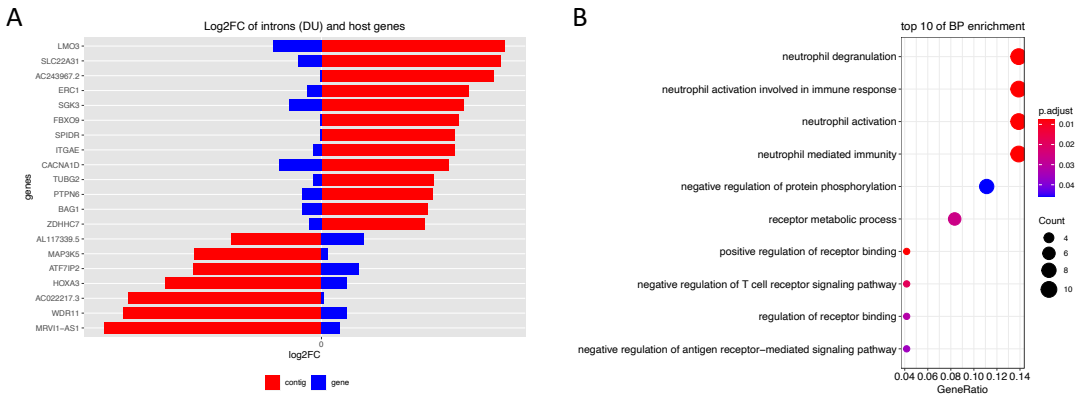


Figure 3: Intronic event analysis. (A) Log2FC values of the top 20 intronic events (DU). Red and blue colors represent the expression fold change of intronic contigs and host genes, respectively. (B) Gene Ontology functional enrichment. Color represents the P-values and size represents the ratio of genes.

show opposite directions of intron and gene expression. Gene Ontology enrichment analysis indicates genes with intronic events are enriched for inflammation and immune response pathways involving neutrophil and T cells. Fig 3B, suggesting these events may come from regulations in the tumor microenvironment rather than in the tumor itself.

3.1.4.5 Novel tumor-specific lincRNAs

The lincRNA category is of particular interest as it contains novel RNAs that do not map any annotated gene. Especially, tumor-specific lincRNAs can be a source of novel biomarkers or neoantigens. Hereafter counts are given for LUADtcga, and in parenthesis for LUADseo. Overall we identified 885 (662) DE lincRNA elements. 83% (63%) were overexpressed in tumors and 73% (74%) were also annotated as repeats. The average length of contigs annotated as lincRNA was 137 nt (189 nt), however actual transcription units were generally longer as most were composed of multiple contigs, as shown in the examples in Fig S8 and Fig S9. No more than one third of the flanking genes of lincRNA contigs were differentially expressed, indi-

cating that lincRNA expression was most often independent from that of flanking genes.

3.1.4.6 Expressed Repeats

The dominant paradigm on endogenous retroelements (EREs) expression is that EREs are mainly expressed in germline and embryonic stem cells while they are repressed in differentiated somatic cells. However recent studies showed expression of EREs in somatic cells is more common and heterogeneous than expected(Larouche et al., 2020). Repeat-containing reads are difficult to analyze by RNA-seq standard pipelines due to ambiguity in the alignment process. We thus questioned whether our alignment-free procedure could help reveal these events.

From the initial set of 50572 contigs annotated as repeats (Fig 1C), we selected a high quality subset of 10341 contigs over 60 bp in size and with expression above a set threshold (see Methods). Of these, 87.7% were up-regulated in tumors and 12.3% were down-regulated (Table S4).

Table 1 shows the distribution of contigs per repeat family. Most repeats correspond to LINE-1 and Alu family sequences. The most frequent repeat overall is L1P1, a LINE-1 of the L1Hs family which is the only retrotransposition-competent EREs in the human genome (Rangwala et al., 2009). L1P1/L1Hs elements, as well as human endogenous retrovirus (HERV) were almost exclusively over-expressed in tumors, suggesting tumor-specific activation of these elements. In contrast, Alu elements, which are often expressed as part of protein coding genes, were either over- or under-expressed in tumors.

To investigate the expression status of various types of repeats from a global perspective, we drew the expression heatmap of the top 60 types of repeats that contribute more contigs among the whole repeat types. For each type of repeats, we first selected the most representative contig with the highest absolute value of log2FC. Then we collected 60 contigs corresponding to 60 repeat types. The heatmap graphs

Table 1: Summary of top 20 repeat types with the most contigs from LUADtcga

rep_type	contigs	Up in tumor	Down in tumor	SNVs	protein_coding	lncRNA
L1P1_orf2	755	754	1	568	233	166
AluSx	455	369	86	250	299	60
FLAM_C_1_143	316	252	64	128	216	38
L1P3_orf2	302	288	14	189	109	82
AluJb	276	227	49	119	189	28
LSU-rRNA_Hsa	264	259	5	249	51	185
AluSz6	174	143	31	74	115	25
AluSp	147	105	42	84	85	18
PRIMA4-int	123	92	31	38	92	8
L1PB_orf2	118	116	2	44	52	26
L1P1_5end	115	115	0	70	39	32
MIR_1_262	111	98	13	10	90	15
AluJr4	110	90	20	47	69	20
AluY	110	79	31	68	59	12
L1HS_5end	109	109	0	65	31	19
L1PREC2_orf2	106	103	3	37	29	32
HERVH	105	96	9	30	19	52
AluSz	104	91	13	61	62	13
L1M3_orf2	89	88	1	5	49	16
AluJr	86	66	20	30	48	14

of LUADseo and LUADtcga can be seen in Fig 4A-B.

To identify repeats most contributing to tumor identity, we submitted the repeat expression matrix to Principal component analysis and ranked contigs according to their contribution to each PCA axis. We selected the top 20 contigs from the first three axis and plotted their expression (Fig 4C-D). The repeats that most contributed to PCA axes were L1P1 and LSU-rRNA. These repeats displayed a highly heterogeneous expression, especially in tumor tissues, delineating clear tumor subgroups with high or low L1P1 and LSU-rRNA expression in both LUAD datasets. LSU-rRNA (rRNA large subunit) had different behavior in LUADtcga (almost always overexpressed in tumors) and LUADseo (either over- or under-expressed in tumors). This inconsistency suggests that LSU-rRNA quantification may be affected by sample or library preparation procedures, without completely excluding a biological origin.

To further investigate repeat-based patient subgroups, we performed clustering of

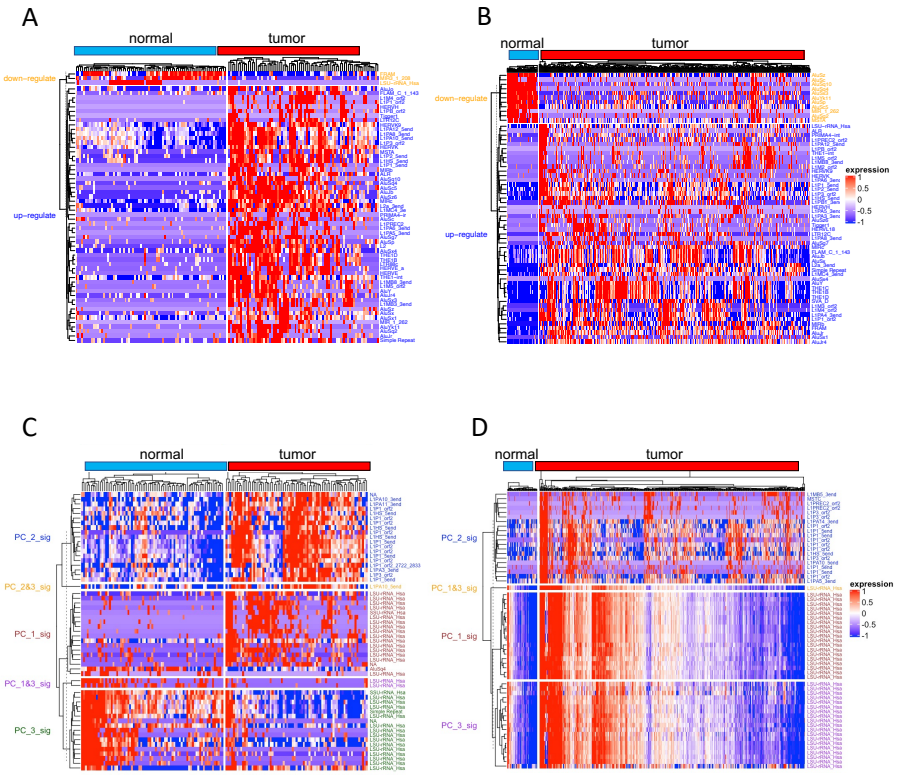


Figure 4: Expression of repeat-containing contigs. Contig expression level is represented from blue (lowest) to red (highest). (A-B) Top up-/down-regulated contig (ranked by fold change) for each repeat type. (C-D) Contigs most contributive to sample clustering. PC_1-3 indicate top contigs from PCA axes 1-3. (A-C) LUADseo dataset. (B-D) LUADtcga dataset.

tumors based on the most frequent repeat elements in Table 1: AluSx, L1P1_orf2, and L1P3_orf2 (as FLAM repeats are a family of Alu-like monomers that give birth to the left arms of the Alu elements, we did not account for FLAM_C_1_143). K-means clustering with k varying from 2 to 4 groups consistently found two major subgroups: subgroup 1 ("repeat-low") displayed generally low expression of Alu and L1 repeats compared to subgroup 2 ("repeat-high") (Fig 5A).

We tested the repeat subgroups for enrichment in clinical and molecular features. We observed no difference between subgroups in terms of age, gender, tumor stage, overall survival (OS), and vital status, but found more smokers in the repeat-high group (Wilcoxon P=0.02). We then assessed the immune cell contents of samples based on gene expression using CIBERSORT. The repeat-high subgroup had lower proportions of dendritic cells, macrophages, mast cells, monocytes and CD4+ T cells and overall immune content than the repeat-low subgroup (Fig 5B).

We further related the two repeat subgroups with somatic variations obtained for TCGA patients Fig 6. Patients in the repeat-high group were more frequently mutated in drivers CSMD3, TP53, PTPRD, PTPRT, GRIN2A, EPHA3, and MB21D2 (Fig 6A, Fisher P<0.05) and had a significantly higher TMB (Wilcoxon P=1.5e-07) (Fig 6B). In addition, patients of the repeat-high group tend to present a higher ratio of CNVs than other patients (Wilcoxon P=5.5e-05 for gain; P=0.019 for loss) (Fig 6C). In summary, "repeat-high" tumors associate with lower immune infiltration, more frequent smoking, and higher genome instability than "repeat-low" tumors.

In addition to the annotated repeats, DE-kupl identified 4762 contigs (4497 up, 265 down) with multiple genome hits but no match in the DFAM repeat database (Table S4, Suppl. files). Notable sources of unannotated repeats are Mucins, immunoglobulins and multicopy gene families such as NBPF and TBC1. These repeats are shared between two cohorts and thus represent robust events of (mostly) overexpressed RNA fragments in tumors that would hardly be noticed in regular RNA-seq analysis due to their low mappability.

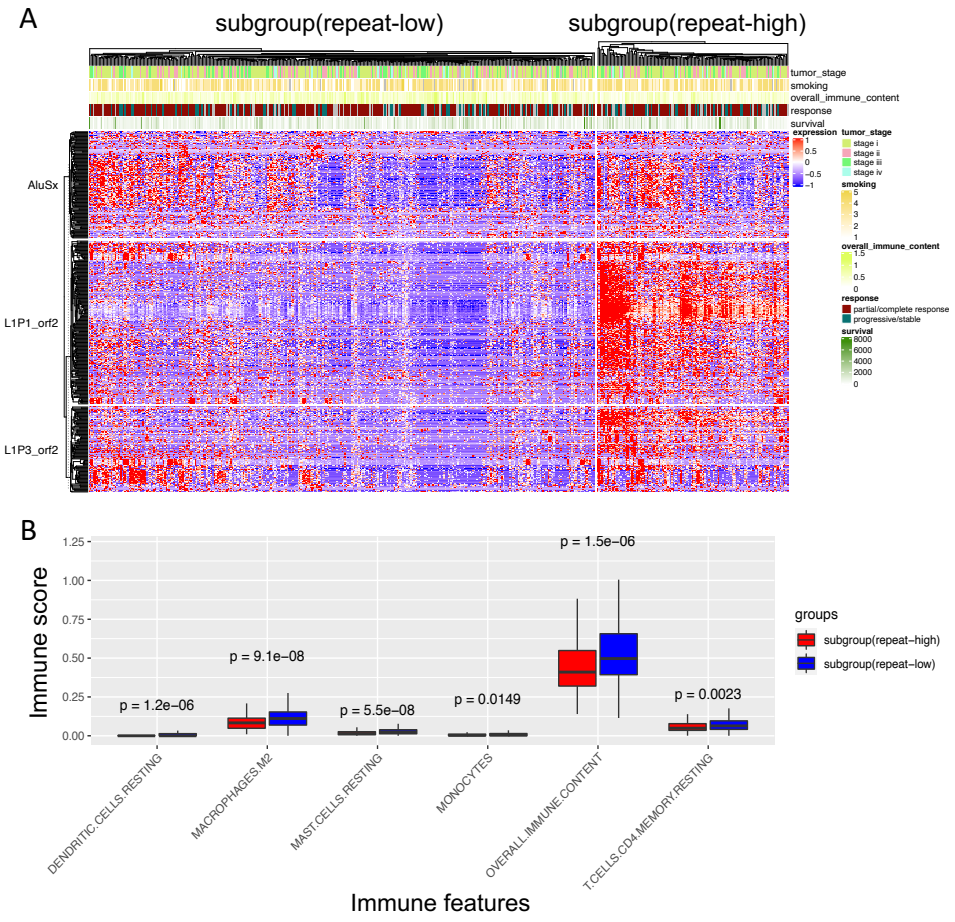


Figure 5: Clustering of TCGA patients into two subgroups based on Alu and L1P1 repeat expression. (A) Heatmap of repeat expression, grouped by Alu and L1 classes. Subgroups were defined by K-means. (B) Variation of immune features between subgroups. The red and blue represent the repeat-high and repeat-low subgroups, respectively. P-values are computed by Wilcoxon test.

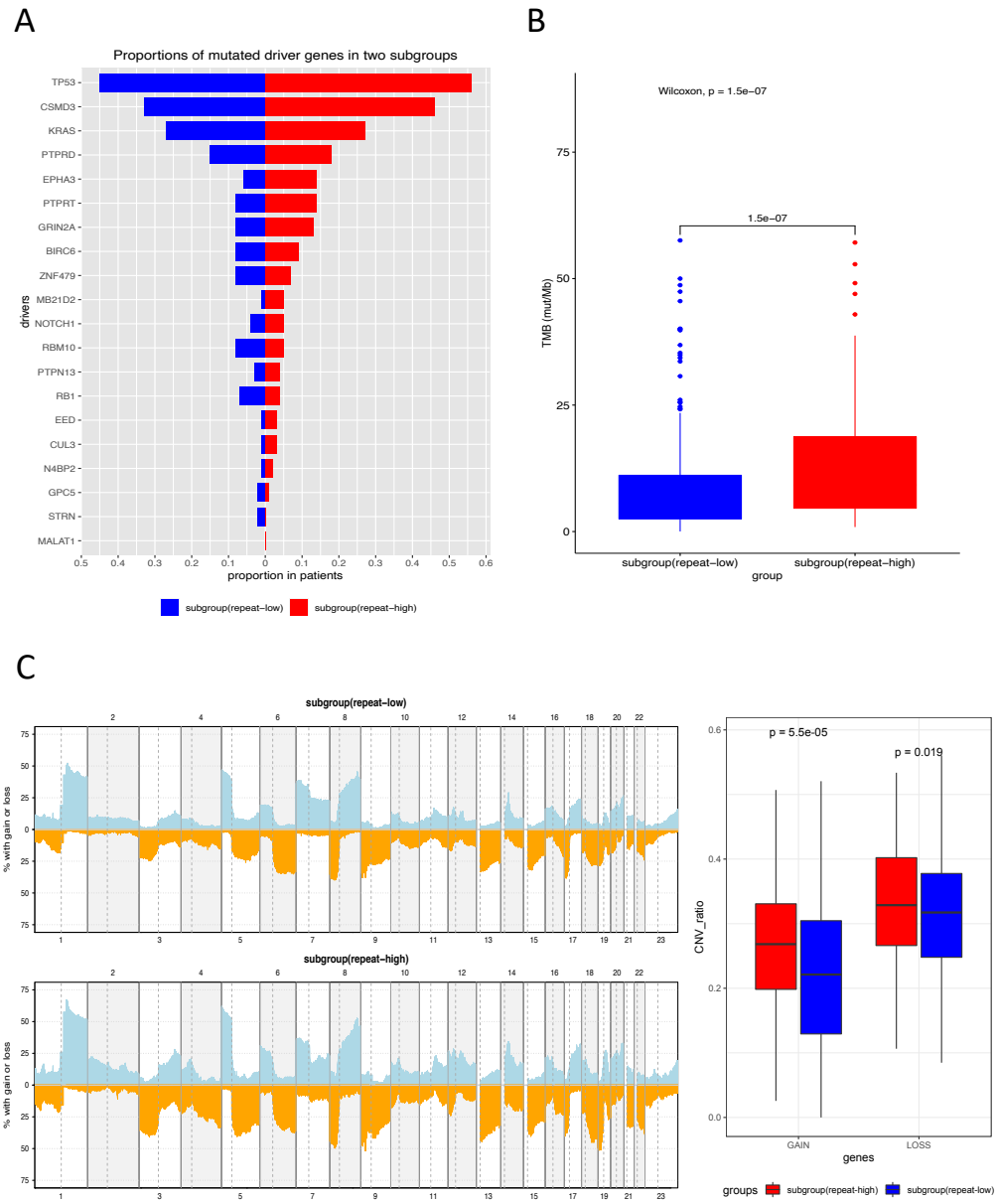


Figure 6: Somatic mutations in the two repeat subgroups in LUADtcga cohort. (A) Fraction of patients with driver mutations for 20 COSMIC LUAD drivers. (B) Mutational burden. Red and blue represent the repeat-high and repeat-low subgroups, respectively. (C) CNV frequency distribution between two subgroups. Lightblue and orange represent amplification and deletion of segments. Red and blue represent repeat-high and repeat-low subgroups.

3.1.4.7 Neoantigen candidates

Genome instability in tumors generates a large amount of transcript alterations (mutations, aberrant mRNAs and lncRNA isoforms, chimeras) which, if translated, can produce peptides recognized as neoantigens by the immune system, triggering antitumor immune response. These novel tumor-specific antigens are the object of active investigation for immunotherapy and tumor vaccine development. Protocols for neoantigen discovery usually start from a list of nonsynonymous somatic mutations identified from WES or WGS libraries and whose expression is confirmed by RNA-seq. Candidate mutated peptides are then submitted to an epitope presentation prediction pipeline (Gopanenko et al., 2020). This protocol predicts neoantigens from annotated and mappable regions. However, neoantigens can be produced from any transcript, including non-coding ones such as lncRNAs and repeats (Ouspenskaia et al., 2020; Laumont et al., 2018). Therefore we thought our reference-free approach could be a good source for such elements.

We considered as neoantigen candidates DE-kupl contigs with zero expression in normal tissues. To focus on candidates shared by several patients, we further requested neoantigen candidates to be expressed in at least 15% of tumor samples. This selected 2375 contigs in the LUADtcga dataset and 1507 in the LUADseo dataset. Candidate neoantigens were mostly found in categories SNVs, introns, repeats and lincRNA (Table 2). There were 469 candidate neoantigen repeats in LUADtcga (hereafter "neo-repeats") (Table S4). Fig 7 shows the expression heatmap of these neo-repeats in the LUADtcga and LUADseo datasets. Expression in normal tissues of the LUADseo cohort was not always zero as this was not a prerequisite. Of 469 neo-repeats in the LUADtcga dataset, only 71 (15%) were also silent in all normal tissues of the LUADseo dataset. We define these contigs which are silent in normal samples in both cohorts as strictly tumoral.

Other major types of neoantigen candidates included SNVs (1235 in LUADtcga, 132 strictly tumoral). Other strictly tumoral contigs included 29 intronic and 50 lincRNA sequences (Table 2). The capacity of some of these strictly tumoral contigs

Table 2: Number of candidate neoantigens in different categories.

category	LUADtcga	LUADseo	shared neos
Repeats	469	599	71
SNVs	1235	1383	132
Introns	405	440	29
lincRNAs	164	260	50

to be translated and processed by the epitope presenting machinery remains to be evaluated.

3.1.4.8 Novel RNA elements as prognostic indicators

To discover new RNA elements associated with prognosis, we obtained overall survival (OS) data for the TCGA cohort and performed univariate Cox regression on shared DE contigs of each class. 45 contigs were significantly related to OS after multiple testing adjustment (Table S7, Table 3). OS-related contigs are mostly enriched in repeats (P=1.419e-05, Fisher's exact test). Four human endogenous retrovirus (HERV) elements are included in the top 10 most significant OS-related repeats. HERV elements were also among the top tumor-specific repeats in Table 1. OS-related repeats also include 21 Alu and L1 family elements (AluSx and L1P3_orf2), but these are of different types than the L1 and Alu elements found to distinguish tumor and normal tissues in our above PCA analysis.

We then performed multivariate Cox regression using sets of contigs selected by lasso regression within each category, with repeats separated into 3 subgroups (Table S8). Kaplan–Meier representations are shown in Fig 8. Models based on annotated and simple repeats had the best prognostic power (log-rank P=2e-16, 2e-13, respectively). The "annotated repeat model" was based on 12 contigs, including 6 L1 and 3 HERV, reinforcing the relevance of these repeats for prognosis.

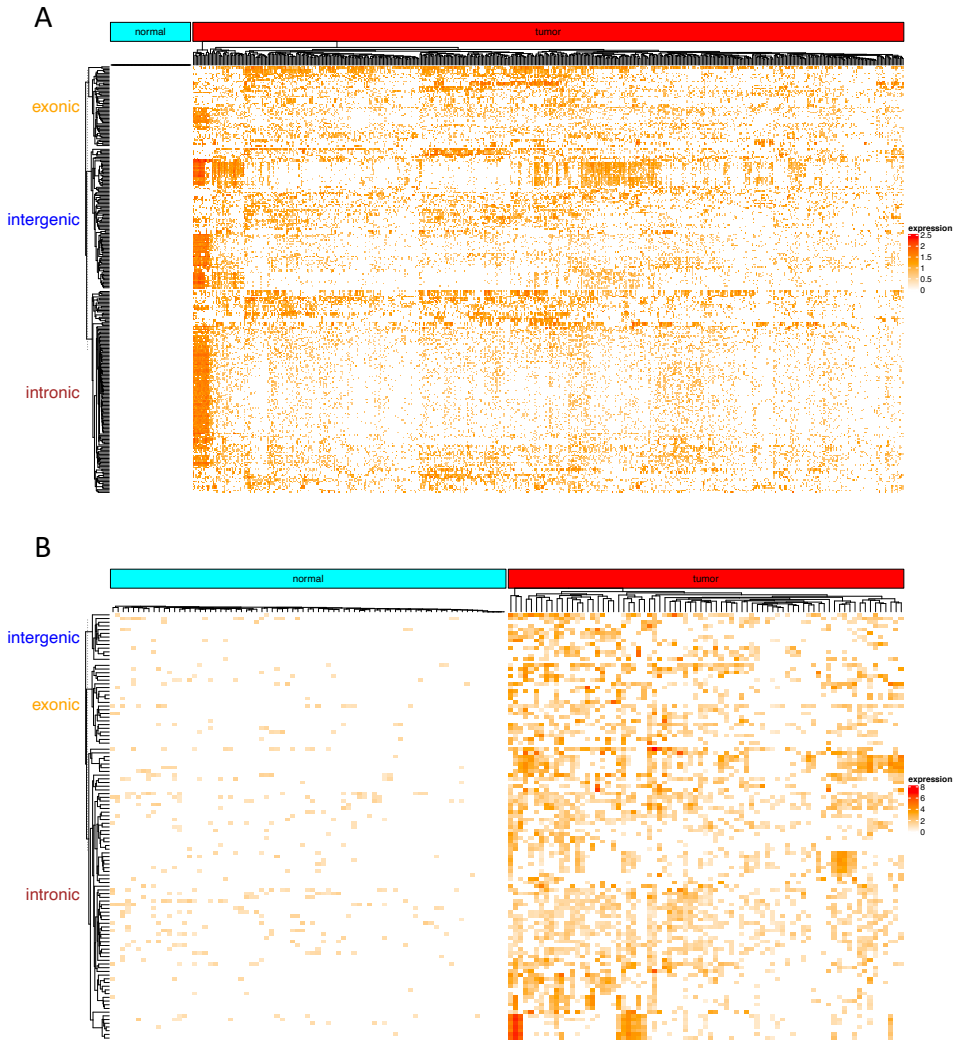


Figure 7: Expression heatmap of candidate neoantigen in repeats. "Neo-repeats" were screened from LUADtcga and validated using LUADseo. (A) Expression of neo-repeats in the LUADtcga dataset (B) Expression of neo-repeats in the LUADseo dataset.

Table 3: Survival significant events in univariate Cox regression

event class	OS-related	hazard ratio>1	hazard ratio<1	Enrichment P-value
repeat	35	16	19	1.419e-05
lincRNA	5	3	2	0.22
SNV	1	0	1	0.37
splice	3	1	2	0.002
split	1	0	1	0.33

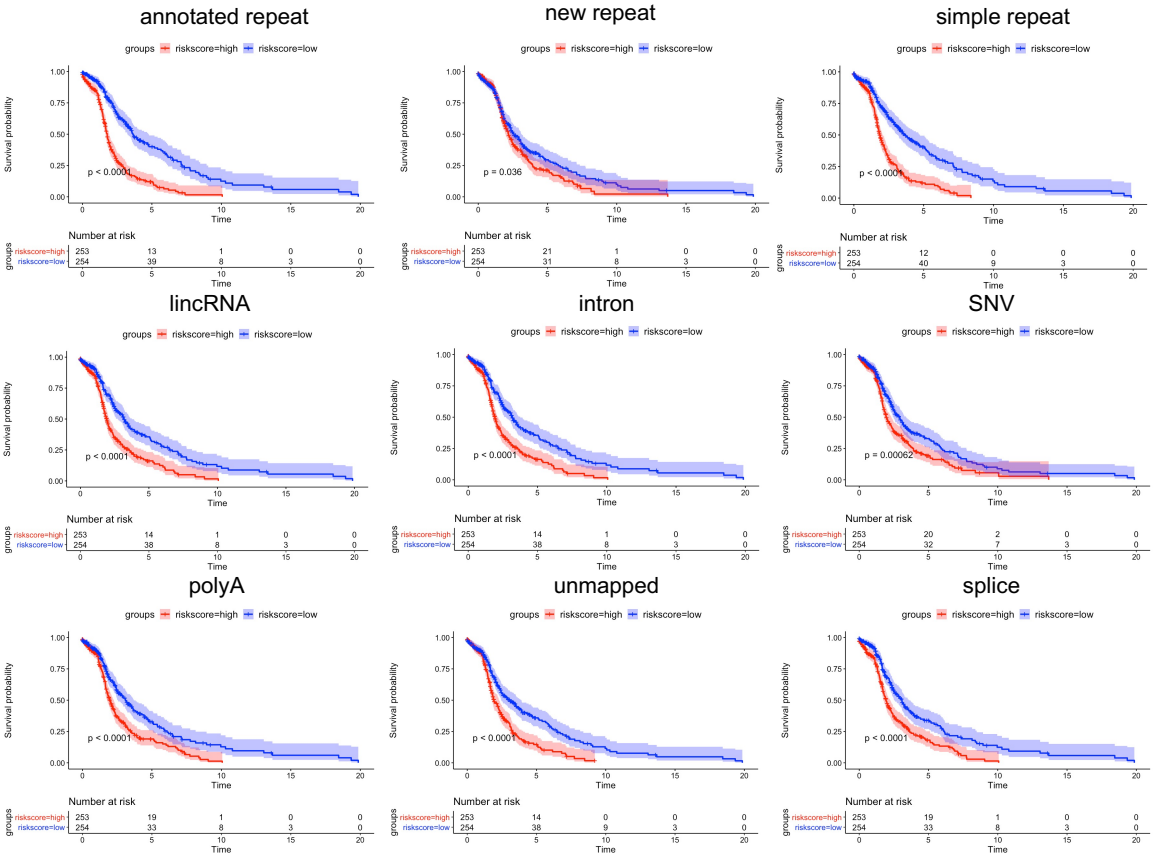


Figure 8: KM curves for multivariate survival models per class of event. Patients in high and low-risk groups are shown in red and blue, respectively. Repeat events were separated into annotated, new and simple repeats. Six categories with more lasso-selected contigs were also included (Table S8). Category "split" is not shown as it contains only one contig after lasso selection.

3.1.4.9 Noise from errors in highly expressed genes

The number of DE contigs in our analysis is at least an order of magnitude higher than the number of differentially expressed genes. One possible reason is DE-kupl identifies local events, which means each contig corresponds to one isolated event, except when two variants occur closely with a distance smaller than k. Therefore, a longer gene composed of multiple exons and introns may contribute to more DE contigs. Another possible reason is the high expression levels of genes. If a gene is highly expressed in tumors, it has a higher chance to introduce more variants that can be detectable by statistical methods. To evaluate this effect, we computed correlations between numbers of contigs and expression of host genes. Fig 9A shows genes with higher expression induce more contigs. However, when considering only shared contigs, the correlation coefficient is strongly reduced (Fig 9B), suggesting shared contigs are substantially less noisy than total contigs.

To figure out which type of event is most affected by highly expressed genes, we selected the 1% most highly expressed genes and grouped their derived contigs into different classes (Fig 9D). We found that novel repeats account for a larger proportion than the other events. The events introduced by highly expressed genes are strongly enriched in the novel repeats with a Fisher exact test P-value of 7.533e-11. The category of SNVs whose ratio is 3.1% in the pie graph is also significantly enriched (P-value = 1.459e-06). Despite a high ratio of 12.2%, the category of annotated repeats is not significantly enriched due to its large number. No matter annotated or novel repeats, they have multiple hits on the genome. Therefore it is reasonable that these repeats are correlated with highly expressed genes. From this perspective, the high expression estimates of genes may also be due to the inclusion of repeat regions, which leads to an over-estimation of gene counts.

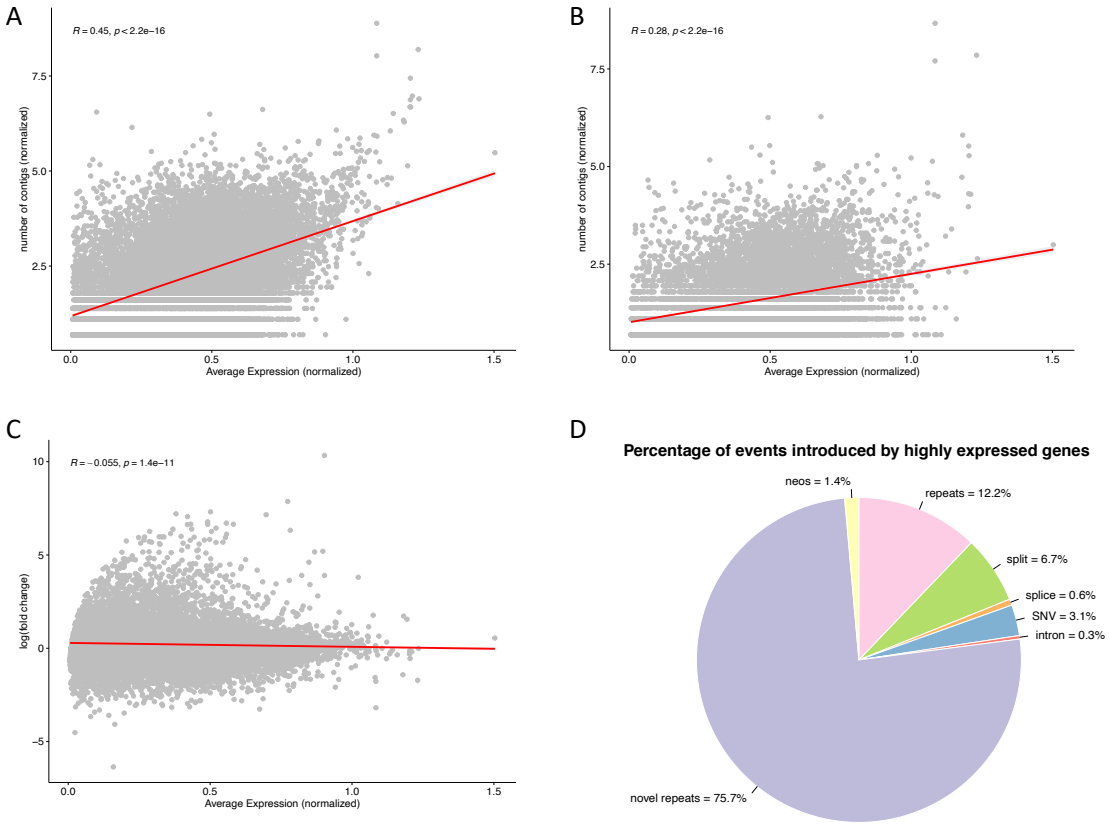


Figure 9: Noise from highly expressed genes. (A) Correlation of gene expression and numbers of contigs in the total contig list (B) Correlation of gene expression and numbers of shared contigs (C) MA-plot showing the correlation between gene expression and fold change. (D) Percentage of contigs contributed by the top 1% highly expressed genes. Gene expression are log- and size-normalized. Red lines show linear regressions. R and P-values are computed with Pearson correlation.

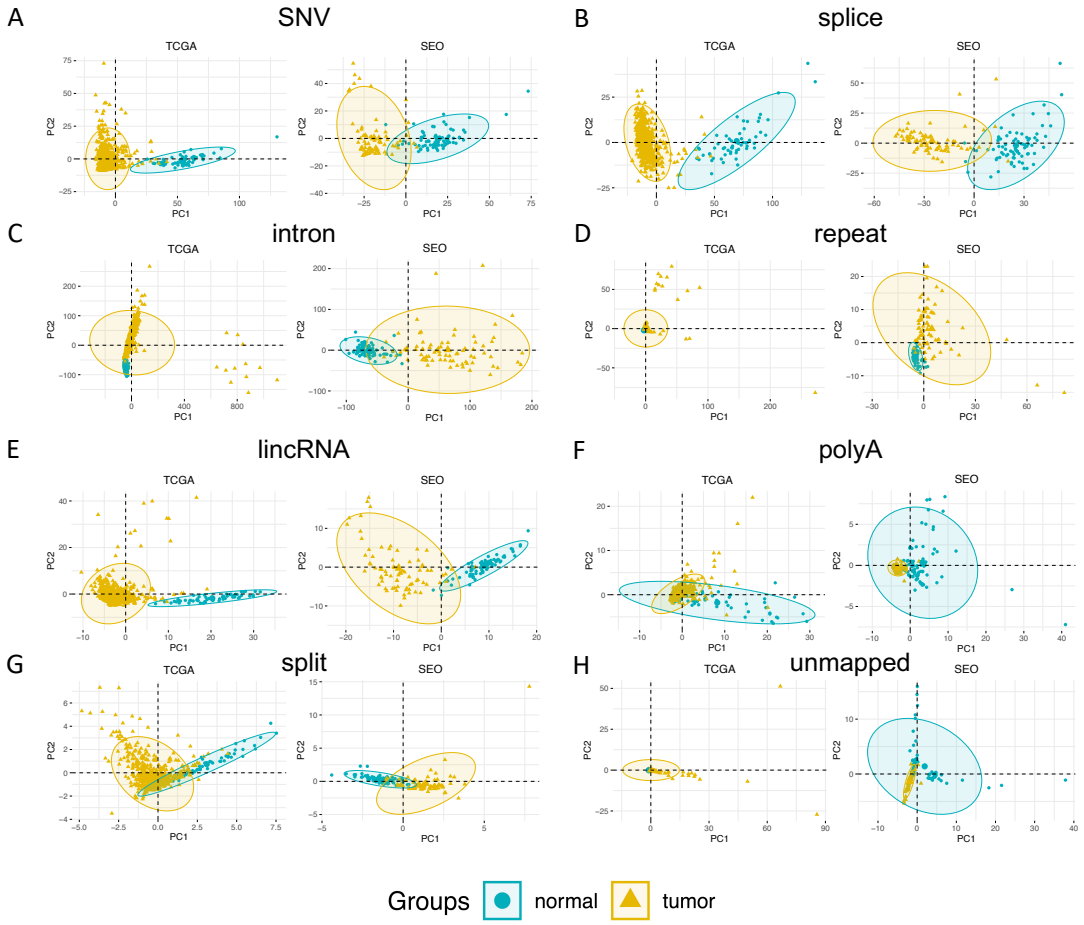


Figure 10: Principal component analysis of samples based on transcriptional events. Each panel represents one transcriptional category. The normal and tumor samples are marked using blue and yellow, respectively. Confidence ellipses were added around each group.

3.1.4.10 Event-based sample clustering

To investigate whether the different event classes could efficiently classify tumor and normal tissue, we performed PCA analysis of samples using each class of contig (Fig 10). Tumor and normal tissues could be clearly separated based on SNV, splice, intron, and lincRNA event classes alone. On the other hand, event classes of repeat, polyA, split and unmapped did not achieve clear separation of tumor and normal tissues. This phenomenon is consistent in both cohorts.

3.1.5 Discussion

Using reference-free analysis of LUAD RNA-seq data, we identified a large set of unannotated RNA variations that were replicated in two independent LUAD cohorts. We classified these variations based on their genomic location, mapping structure and repeat content. In this article we did not systematically analyze all contig classes but focused instead on hypervariable genes, repeats, lincRNAs and intronic elements. Besides these, a number of splice variants, chimeras, exogenous (non-human) sequences were found differentially expressed and could be pursued further.

A defining class of variation involved the expression of endogenous repeats. The expression of L1 and Alu repeats defined two major tumor subgroups. The subgroup with higher L1/Alu expression was associated with more frequent mutations in P53, a higher mutational and copy number burden and a reduced immune cell infiltrate. This is consistent with the previous finding that involved P53 in the control of retrotransposition (Jung et al., 2018) and correlated L1 retrotransposition with a repressed immune environment (Jung et al., 2018; Zhang et al., 2020b). TE mobility can also lead to genome instability. Random TE integrations result in insertional mutagenesis and genomic structural variants including CNVs (Lee et al., 2012).

Besides their capacity to stratify patients, expressed repeats had significant prognostic power. Multivariate signatures composed HERV and L1 expression, or of simple repeat expression separated patients into clear survival groups. HERV expression has been sporadically involved in various cancer types (Bannert et al., 2018), and was recently associated with poor prognosis in colorectal cancer (Golkaram et al., 2021).

A limitation of our approach for TE analysis is that transcripts are not fully assembled and thus the nature of elements, whether expressed as fully functioning retroelements or as part of mRNA or lncRNAs cannot be systematically established. Nonetheless, a fraction of DE contigs are long enough to enable unambiguous map-

ping on the human genome, hence their origin could be further explored, including if coming from novel insertion events.

Another area where reference-free approaches have a high discovery potential is the detection of neoantigens for the development of antitumor therapies and for patient orientation in immunotherapy. We found potential sources of shared neoantigens in repeats, lncRNAs and splice variants of mRNAs. Tumor-specific neoantigens have previously been identified from repeats and supposedly non-coding regions using mapping-based strategies (Smith et al., 2019; Laumont et al., 2018). However, we think our approach has more potential as it collects all events independently of their origin, including from unmappable or profoundly rearranged regions. Therefore we have a better chance to uncover sources of neoepitopes whatever their origin. A next obvious step would be to evaluate all TSA candidates for presentation by the MHC class I complex. We focused here on shared events found in at least 15% of tumors, therefore these candidates are of particular interest since their targeting by antitumor therapy would potentially benefit more patients.

Reference-free analysis has other benefits. First, it is by essence an integrative method as it combines genomic and transcriptomic variation into a single expression matrix that can be analyzed in multiple ways. An attractive application of such matrices is for building predictive models integrating multiple event classes. We (Nguyen et al., 2021) and others (Lorenzi et al., 2020) have initiated this kind of approach with very promising results. Second, reference-free methods could be particularly attractive in meta-transcriptomics projects where RNAs are captured from an environment containing unknown bacterial, archaeal or eukaryotic species. Our protocol guarantees that any RNA that is specific to a sample subset will be captured independently of its origin.

Future developments could use paired normal-tumor samples to systematically collect all tumor-specific events at the patient level, using a simple k-mer count filter instead of differential expression statistics. A recent alternative for patient level variant detection is Mintie (Cmero et al., 2020) which performs *de novo* assembly

of reads from a tumor sample and differential expression analysis of the assembled transcripts against a reference transcriptome.

Declarations

Ethics approval and consent to participate

Not applicable

Consent to publish

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was funded in part by Agence Nationale de la Recherche grant ANR-18-CE45-0020 and by a PhD studentship to YW by Annoroad Technology, Beijing.

Authors' contributions

YW and DG designed the workflow and analyzed the results, YW downloaded and processed the datasets, YW and DG wrote the manuscript, MA and MG assisted in statistical analysis, HX assisted in coding scripts. AL annotated the repeat types.

Acknowledgements

Not applicable

3.1.6 Additional Files

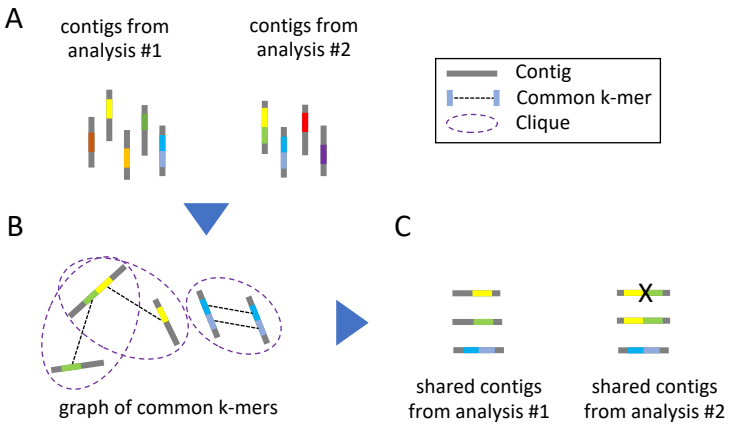


Figure S1: The graph-based protocol detecting shared contigs between TCGA and SEO datasets

Table S1. Nucleotide contents of DE-kupl contigs for the TCGA LUAD dataset

Table S2. Description of event categories extracted from DE-kupl-annot tables

Table S3. General characteristics of contigs shared between LUADtcga and LUADseo

Table S4. Summary statistics for all event categories in contigs shared between LUADtcga and LUADseo

Table S5. Genes with more than 35 mapped contigs (shared LUAD contigs. Colored columns indicate ratio of contigs in said categories)

Table S6. Blast results of 100 contigs mapped to IGHV genes

Table S7. Univariate Cox regression results of all categories

Table S8. Multivariate Cox regression results of all categories

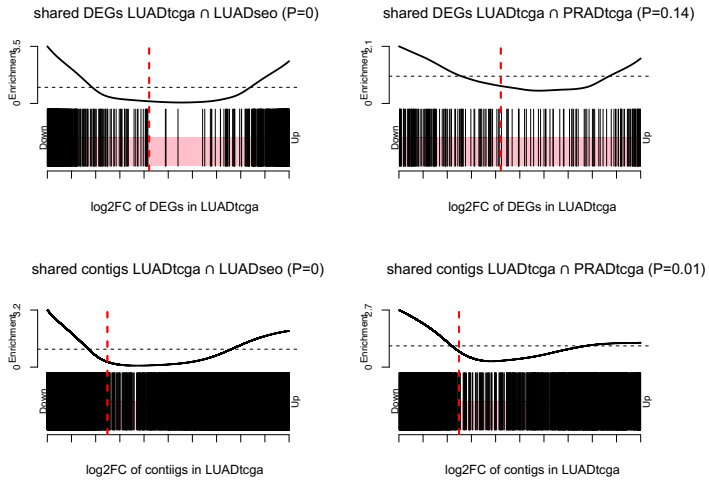


Figure S2: Enrichment analysis of shared DEGs and contigs between TCGA and SEO datasets

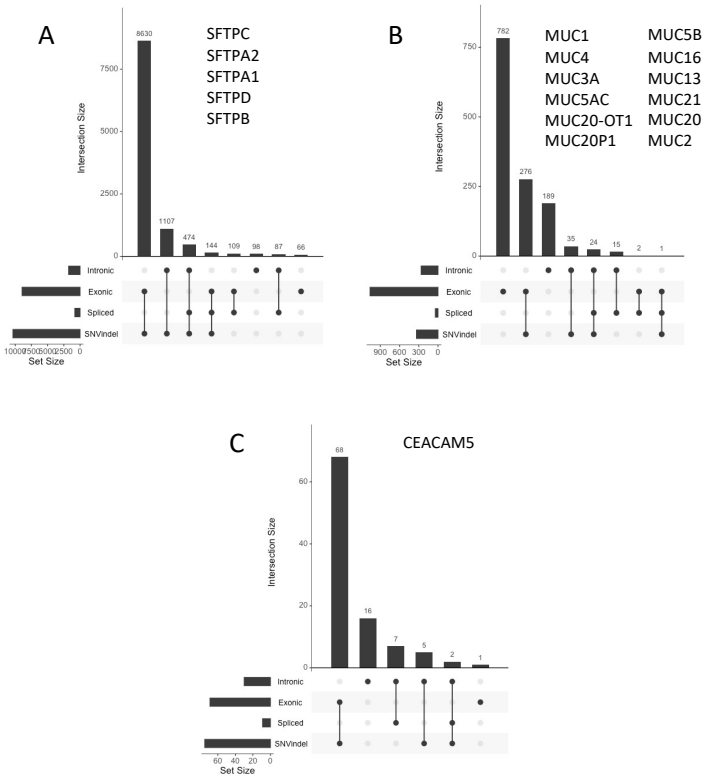


Figure S3: Hypervariable genes in our analysis

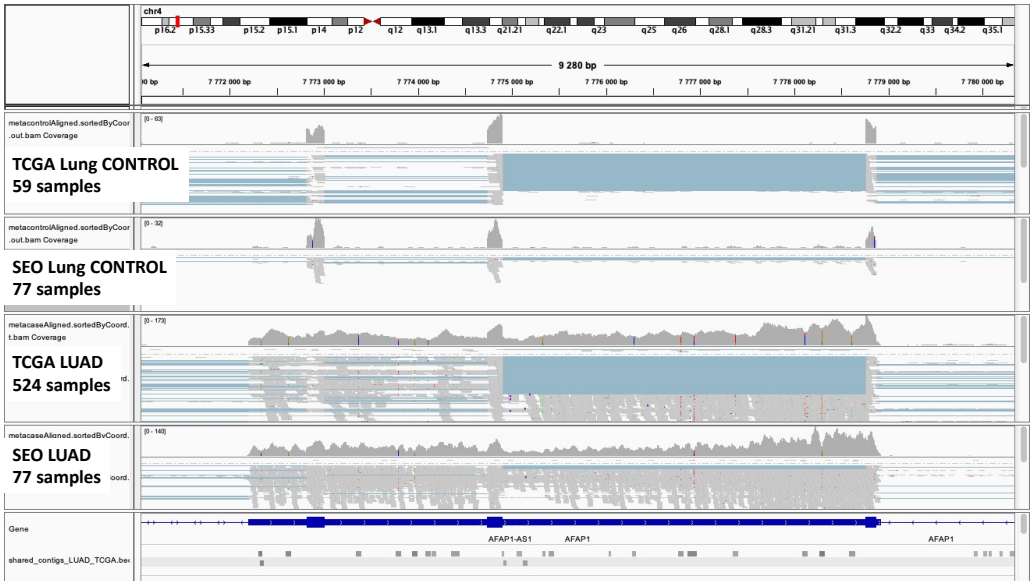


Figure S4: The IGV view of an intron retention in gene AFAP1

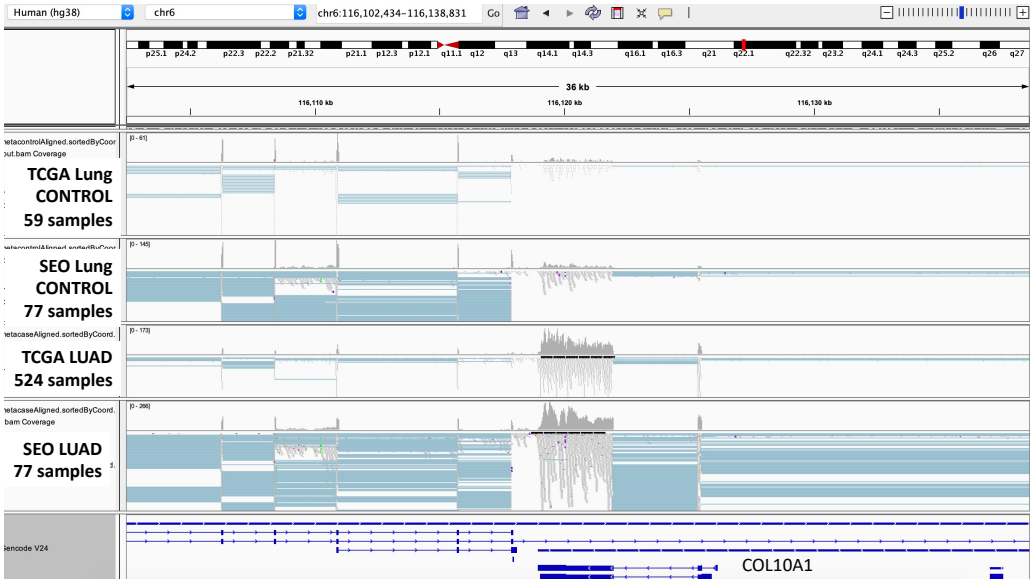


Figure S5: The IGV view of a transcription unit occurring in the intron of gene COL10A1

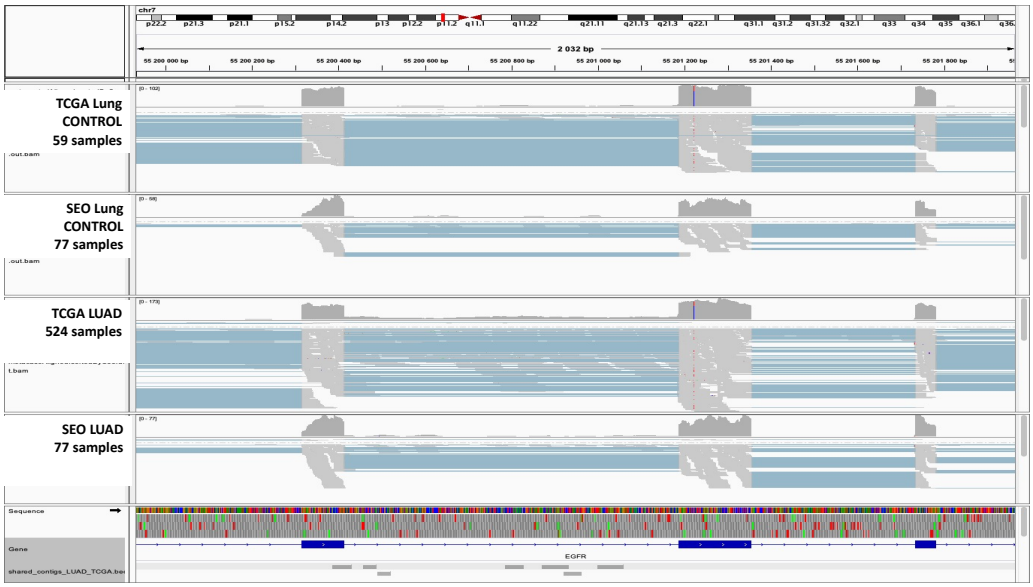


Figure S6: The IGV view of an intron retention in gene EGFR

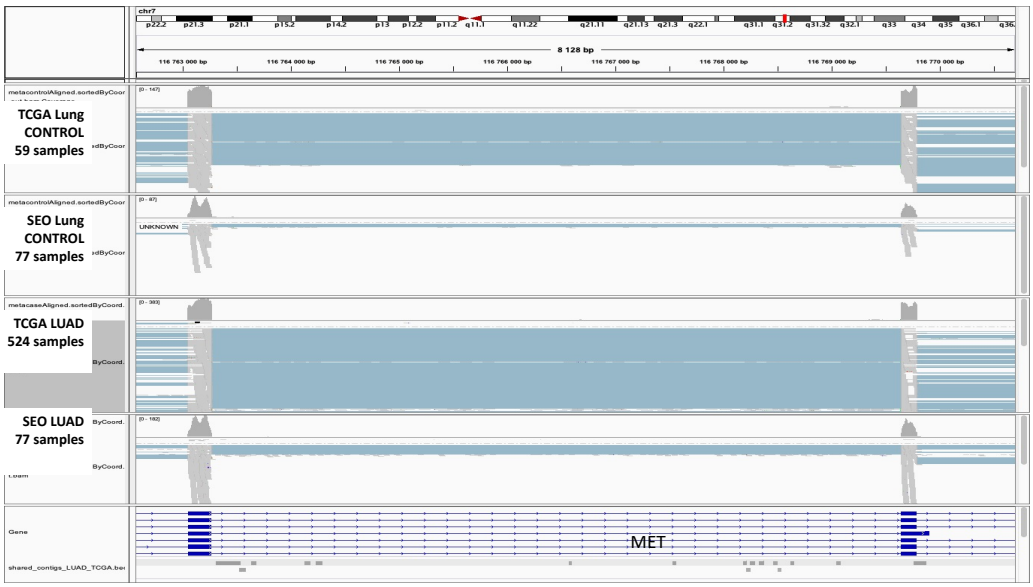


Figure S7: The IGV view of an intron retention in gene MET

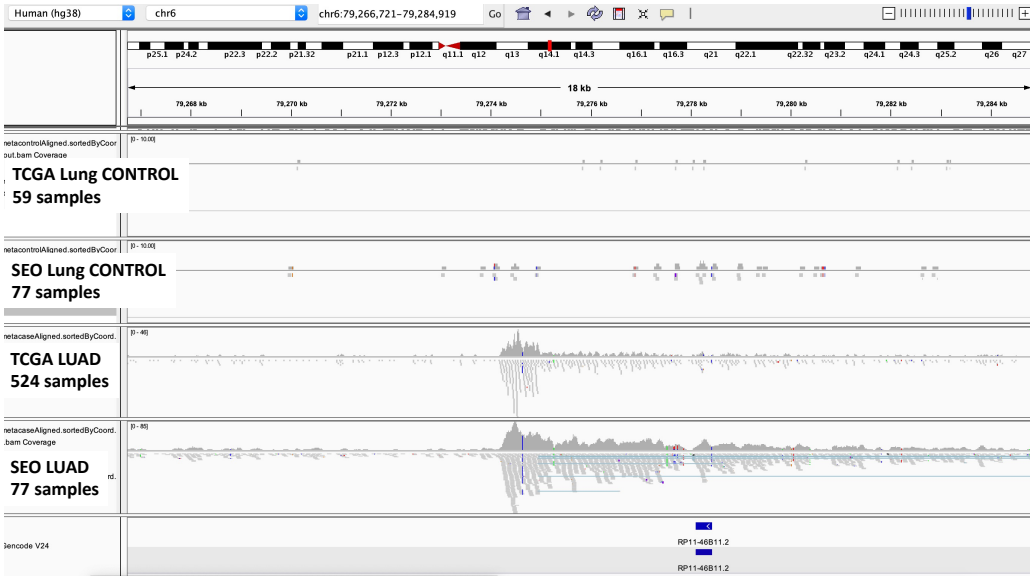


Figure S8: IGV view of a lincRNA element overexpressed in tumors. Each frame shows a metabam file composed of randomly sampled reads corresponding to the subcohort indicated on the left panel

Figure S9: IGV view of a lincRNA element overexpressed in tumors. Each frame shows a metabam file composed of randomly sampled reads corresponding to the subcohort indicated on the left panel

3.2 2-kupl: mapping-free variant detection from DNA-seq data of matched samples

Yunfeng WANG[12], Haoliang[1], Christine POURCEL[1], Yang DU[2], Daniel Gautheret[1*]

* Correspondence: daniel.gautheret@universite-paris-saclay.fr

[1]Institute for Integrative Biology of the Cell, CEA, CNRS, Université Paris-Saclay, Gif-Sur-Yvette, France.
[2]Annoroad Gene Technology Co., Ltd, Beijing, China

3.2.1 Contribution

This article was accepted for publication at BMC Bioinformatics on May 11, 2021. Herein, I developed a mapping-free approach called 2-kupl for calling somatic variants from two matched samples using DNA-seq data. I built the framework of 2-kupl composed of two parallel paths and each module inside. The two paths are aimed at detecting isolated mutations and all other types of variants. The indels, structural variants and multiple mutations are handled by the second path.

In 2-kupl, I introduced two new concepts that are "cs-kmers" and "ct-kmers". The cs-kmers refer to the case-specific k-mers that are only present in the case sample. A ct-kmer refers to the counterpart k-mer of each cs-kmer with only one mismatch. A minimizer-based hash table is applied to index cs-kmers and match to ct-kmers. The mutant contigs and the corresponding putative references can be obtained through assembling cs-kmers and ct-kmers parallelly. By aligning the mutant contig to the corresponding putative reference, 2-kupl can decide the variant type and relative position without using the reference genome. Meanwhile, statistics such as allele frequency can be estimated based on the counts of cs-kmers and ct-kmers, which is parallelizable.

I also implemented the second path for detecting variants other than single mutations. The cs-kmers corresponding to such variants can be retained and merged to mutant contigs. However, no ct-kmers can be matched due to violation of the "one mismatch" principle. I utilized a thirt-party software of BBDUK to retrieve reads and infer the reference, which is also parallelizable.

I downloaded the simulated datasets to evaluate the performance of 2-kupl. Other mapping-free approaches and standard mapping-based approaches are involved and compared with 2-kupl. The comparison demonstrates that 2-kupl performs better than mapping-free variant callers. 2-kupl also achieves close accuracy as the leading mapping-based variant callers but with much shorter running time. Finally, I applied 2-kupl on large-scale prostate cancer patients to detect recurrent variants/genes and novel events missed by canonical pipelines. The source code of 2-kupl is publically available in the Github repository https://github.com/yunfengwang0317/2-kupl.

3.2.2 Introduction

Searching for genomic variants is a fundamental aspect of medical research, whether in the study of Mendelian diseases or of somatic, cancer-related alterations (Li et al., 2017). While certain variants result in gene dysfunction and disease (MacArthur et al., 2014), others are largely asymptomatic but give rise to neoantigens relevant to immune escape and therapeutic efficacy or treatment (Jiang et al., 2019). Genome variants are also of interest in microbiology to analyze the differences between microbial strains (Shiloach et al., 2010) and reveal mechanisms underlying phenotypes. In this study, we address the problem of finding genomic differences between a matching pair of high throughput **DNA seq**uencing (DNA-seq) datasets from the same individual (human somatic variation) or from two bacterial strains.

Genomic variants include mutations, indels and structural variants (SV). Mutations and indels can alter genes by disrupting the genetic code, while SVs, by pulling

distant regions together or splitting one region into segments, can create chimeric genes or have a broader impact on whole chromosomal regions (Hurles et al., 2008). Variants are typically detected by whole-genome (WGS) or whole-exome (WES) sequencing through comparison with reference sequences. Aligners such as BWA (Li and Durbin, 2009) are first applied to map reads to the reference sequences. The variant calling step then detects differences between mapped reads and the reference. Popular variant callers include MuTect2 (Benjamin et al., 2019), VarScan (Koboldt et al., 2012), somaticsniper (Larson et al., 2012) and MuSE (Fan et al., 2016). Based on variants observed between two sequence samples and a common reference genome, these programs can then infer differences between the two samples (e.g., in MuTect2's somatic mode).

Reference-based variant calling has well-known limitations. Aligners may encounter difficulties while handling reads with low mapping qualities (Li et al., 2008), originating from repeat regions, low complexity regions or complex variants. These reads of low mapping quality are usually discarded. Furthermore, some species have no reliable reference, which is common in microbes (Loeffler et al., 2020).

Alternative approaches to variant calling involve mapping-free protocols (Audano et al., 2018). These methods do not rely on a reference genome and can directly predict variants from the raw fastq file. A typical strategy is to use a de Bruijn graph (DBG) (Compeau et al., 2011). A DBG is constructed using k-mers (subsequences of fixed size k) decomposed from the sequence reads. The occurrence of k-mers harboring a mutant allele and a wild type allele generates a bubble structure in the DBG. Variant callers developed based on DBGs include DiscoSNP++ (Uricaru et al., 2015) and Lancet (Narzisi et al., 2017). DBG-based methods also introduce new issues. First, complex genomic variants and repeats may result in complicated graphs that are difficult to parse (Iqbal et al., 2012). Second, short contigs may be discarded at the post-processing step, where branch pruning may cause many false negatives. Furthermore, sequences assembled by k-mers without variants have little contribution if the purpose is detecting variants. Only reconstructing the active regions spanning the variants is more efficient than considering all k-mers

(Audano et al., 2018). Although it is possible to extend DBG-based methods to SV detection, the lack of sensitivity to local events makes these approaches less suitable for finding variants in ambiguous regions, such as repeats (Heydari et al., 2019). This motivates the need for a method to detect variants in arbitrary genome regions directly from DNA-seq data.

We present 2-kupl, a k-mer-based bioinformatics pipeline that compares matched case and control samples to discover case-specific variants. 2-kupl identifies sequence fragments (contigs) specific to the mutant dataset and their wild-type counterpart in the control dataset. This operation is done without relying on a reference genome. We compare the accuracy and CPU-requirements of 2-kupl with that of other variant calling software using both simulated and real DNA-seq datasets. We analyze the nature of novel variants detected by 2-kupl and potential reasons for their absence in conventional protocols. We also use 2-kupl to detect recurrent variants in **pr**ostate **ad**enocarcinoma (PRAD) WES samples from the TCGA project (Tomczak et al., 2015). Finally, we evaluate 2-kupl precision in bacterial WGS data. Overall, we demonstrate that 2-kupl is a practical and powerful alternative for the discovery of genomic variants in hard-to map regions or species with no reliable reference.

3.2.3 Materials and Methods

3.2.3.1 Outline of 2-kupl pipeline

The general pipeline is presented in Fig 1. The input is composed of DNA-seq data from two matched samples. Samples typically correspond to control/normal/wild-type and a case/tumor/mutant-type. For cancer data, we strongly recommend using as a control of a distant tissue such as white blood cells rather than adjacent normal tissues, as the later can be contaminated by tumor cells and 2-kupl only considers variant sequences that are absent in the control dataset. Sequence types can be either single-end or paired-end sequencing reads. 2-kupl then identifies pairs of case-specific k-mers (cs-kmers) and counterpart k-mers (ct-kmers). 2-kupl

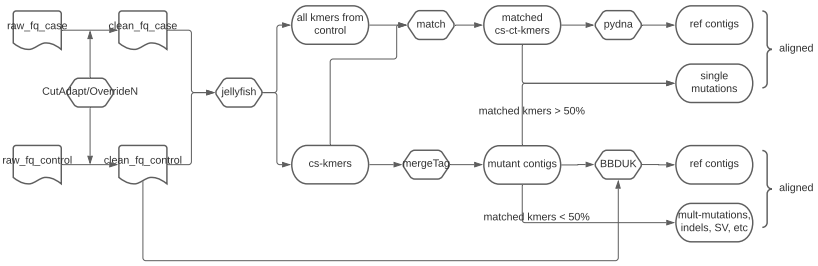


Figure 1: Overall workflow of 2-kupl. This flowchart describes the analysis process of 2-kupl, including the input and output file format and function of each module.

returns predicted variants exclusive to the case sample, including mutations, indels and structural variations. Variant statistics including cs-count, coverage, allele frequency and variant P-value are computed. A variant file and an alignment file are produced. 2-kupl accepts multiple threads and uses 10 threads by default.

2-kupl is developed purely in Python. The main dependencies include Jellyfish (Marçais and Kingsford, 2011) and GSNAP (Wu et al., 2016). Other dependent python libraries and instructions can be found from the Github repository https://github.com/yunfengwang0317/2-kupl

3.2.3.2 Data cleaning

Low quality sequences are trimmed with Cutadapt (Chen et al., 2014) (parameter '–quality-cutoff' = 10). As Cutadapt does not remove low-quality bases within the central part of reads, we implemented an overriding function that replaces each low-quality base (Phred score <10) with N. This procedure is applied to both case and control libraries.

3.2.3.3 k-mer indexing and counting

Jellyfish is used to index and quantify k-mers from both case and control with options k=31 and -C (canonical k-mers). As Jellyfish removes k-mers containing

Ns, none of the low-quality bases is present in the k-mer list. The generated k-mers subsequently undergo two filtering steps. First, k-mers with counts below a user-specified cutoff (default=3) are removed. These low abundance k-mers are assumed to result from sequencing errors or off-target regions in the case of WES data. Second, k-mer lists from case and control are compared and only case-specific k-mers (cs-kmers) are retained.

3.2.3.4 Matching counterparts of cs-kmers

For each cs-kmer harboring a point mutation, there should exist a counterpart k-mer (ct-kmer) from the control dataset with only one base substitution (Hamming distance =1), which can be considered as a product of the wild type sequence. Note that Hamming distance=1 only considers substitutions. Hence single nucleotide insertions and deletions are rejected at this step and will be treated later with unmatched k-mers. Finding the matched ct-kmer for each cs-kmer should allow us to infer the variation without reference sequences. We initially build a hash table where the keys are the continuous 15 bases from each side of cs-kmers. For each 15-bases key, we create a bucket of all k-mers starting or ending with the key. Then we survey the buckets and seek all k-mer pairs with a hamming distance of one in the same bucket. We thus generate all k-mer pairs (ki, kj) with a hamming distance of one. For any pair of k-mers with a Hamming distance of one, if one k-mer comes from the cs-kmer list and the other comes from the control, this pair of k-mers is considered to be matched. Otherwise, we allocate the cs-kmers to the "unmatched k-mers" group. These unmatched k-mers either contain variants of more than one nucleotide (multiple mutations, indels and structural variants) or come from low coverage regions. The schematic workflow is shown in Fig 2.

3.2.3.5 Assembly of cs-kmers into mutant contigs

cs-kmers are assembled into mutant contigs that correspond to variants and their local context. The assembly process is done using the "mergeTag" function from

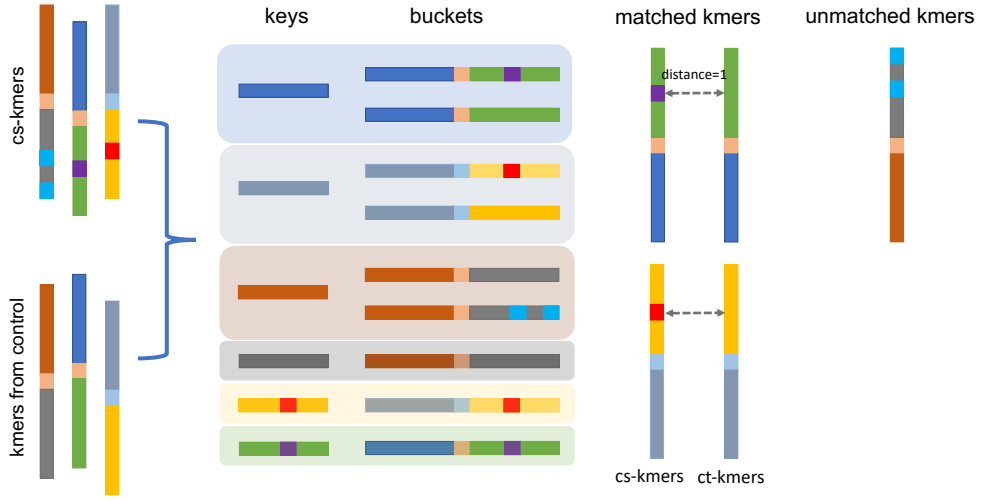


Figure 2: Procedure for matching cs-kmers to ct-kmers. Long rectangles represent one 31-mer. Short rectangles (keys) represent the head or tail 15 bp of a cs-kmer. Color changes indicate sequence differences.

DEkupl (Audoux et al., 2017a) (https://github.com/Transipedia/dekupl). Two k-mers overlapping by k-i bases are merged iteratively with i ranging from 30 to 25 (min_overlap parameter is set to 25 by default). The merging process is interrupted when no k-mers can be added or ambiguity occurs (two different overlapping k-mers are encountered).

3.2.3.6 Inferring reference contigs

We use two distinct procedures for reference sequence determination, depending on whether or not sufficient ct-kmers are available to build a reference contig.

For each mutant contig, if more than half of its component k-mers are matched, all the ct-kmers are merged by the python package pydna (Pereira et al., 2015). The resulting mutant contigs correspond to isolated mutations. Merged contigs produced by ct-kmers can be regarded as putative references. For each pair of mutant and reference contig, we then define two values representing counts of supporting k-mers for the mutant allele (cs-count) and supporting k-mers for both mutant

and reference alleles (coverage). The cs-count is computed from the median k-mer count of cs-kmers and coverage is calculated from the sum of the median count of cs-kmers and ct-kmers. Herein, we select the median count instead of the mean count because mean values are more sensitive to high-count k-mers from repeats or copy number amplification regions.

For mutant contigs in which less than half of the k-mers are paired, we consider that a reference cannot be assembled from paired-kmers. A procedure was implemented to retrieve the reference from the original reads. Reads with at most one mismatch to any k-mer from the mutant contig are retrieved from the control fastq file using BBDUK (Bushnell, 2018). These reads are then assembled by CAP3 (Huang and Madan, 1999). In this way, we can infer the putative reference for each contig and evaluate coverage based on the number of reads retrieved by BBDUK. The cs-kmers in these contigs have no matching ct-kmers and contigs are thus considered to contain multiple mutations, indels and structural variants.

3.2.3.7 Filtering low-quality variants

The cs-count and coverage substantially impact the reliability of events called by 2-kupl. For instance, a sequencing error could be repeatedly generated in a region of high coverage. Besides, sequencing errors may, by chance, be detected as mutations with high allele frequency in low coverage regions. Thus, false positives are introduced due to either high cs-count in high coverage regions or high allele frequency in low coverage regions. However, coverage varies between whole-genome sequencing (WGS) and whole-exome sequencing (WES) data. WGS does not use an upfront enrichment step so it generates a more uniform coverage of the genome. On the other hand, the enrichment steps involved in WES lead to non-uniform coverage, generating coverage 'hot' and 'cold' spots (Wang et al., 2017). 2-kupl provides several criteria for users to evaluate call reliability. A Fisher's exact test P-value is calculated based on the cs-count and coverage in case and matched control libraries for each variation. A Phred quality score is subsequently computed as $-10log_{10}$P.

Users can specify cutoffs for cs-count, coverage, allele frequency and Phred to filter false positives. Default cutoffs for cs-count, coverage, allele frequency and Phred are set to 3, 10, 0.05 and 5, respectively.

3.2.3.8 VCF format export

Events identified by 2-kupl are exported as a variant call format (VCF) file (Danecek et al., 2011). 2-kupl outputs the contig harboring the variation and the corresponding putative reference without the variation for each event. If users provide an available reference, the mutant contig is mapped to this reference using GSNAP (Wu et al., 2016). After the mapping process, actual chromosome and position information are provided in the VCF file. Besides the VCF file, 2-kupl also exports an alignment of each contig and its putative reference obtained using the pairwise2 python package (Cock et al., 2009). Contigs corresponding to indels and structural variants are further mapped to reference by BLAST (McGinnis and Madden, 2004) (default parameters) which we found better suited to fragmented alignments.

3.2.3.9 Comparison with other software

We compared 2-kupl with three other tools. DiscoSNP++ (Uricaru et al., 2015) is designed for detecting SNVs and small indels from fastq files without using reference. DiscoSNP++ first generates a DBG of two matched samples pooled together (Li et al., 2012) and detects variants based on searching bubbles in the graph. The context contigs can be extracted from DBG bubbles that correspond to local variants. As DiscoSNP++ calls variants in each sample rather than specific to one sample, we applied cutoffs to DiscoSNP++ allele frequencies (AF) to extract case-specific calls as found by 2-kupl. After testing multiple combinations, DiscoSNP++ achieved the best performance when AF cutoffs for both case and control samples were set to 0.05. Lancet (Narzisi et al., 2017) relies on localized colored DBG to detect somatic variants in paired samples. K-mers shared by two matched samples or specific to either of them are marked in different colors in the DBG. In this way, Lancet is able

to detect case-specific events. It is worth mentioning that Lancet uses bam format files as input so it also leverages the reference before variant detection. We also compared 2-kupl with the leading reference-based GATK-MuTect2 pipeline (Benjamin et al., 2019). GATK-MuTect2 takes mapped sequence files as input, detects variants based on the reference and compares the variants of two matched samples to identify case-specific variants (somatic mode). Version hg38 of the human genome was used in all reference-based procedures. To make runtime comparisons fair, we took the mapping procedure into account in Lancet and GATK-MuTect2. Alignment was performed using BWA with default parameters. Thus all four protocols started with fastq files. To evaluate the dependency of 2-kupl running time on the number of k-mers, we ignored the part up to k-mer counting. Mapped reads were visualized with the Integrative Genomics Viewer (IGV) (Robinson et al., 2011) 2.6.2 on hg38.

3.2.3.10 Simulated WES analysis

We downloaded simulated WES data from Meng and Chen (Meng and Chen, 2018). This dataset was developed based on the NA12878 pilot genome (Zook et al., 2016) (reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree). The authors used BAM-Surgeon (Ewing et al., 2015) to select genomic loci and introduce random SNV and indel spike-ins, and generated 2x100nt reads WES files at 230X coverage. For our benchmark, we used a tumor sample described by authors as one of the most complicated, NA12878_79_snv_indel_sorted.bam (with four sub-populations, expected variant allele frequency (VAFs) of 0.5, 0.35, 0.2 and 0.1). Picard was used to convert bam files to fastq format files with default parameters. 2-kupl was run using default parameters on pairs of simulated normal-tumor fastq files.

3.2.3.11 Simulated WGS analysis

A simulated WGS dataset containing two matched samples was generated by DWGSM(Li, 2011), with a mean coverage of 50X across available positions. The rates of mutations in case and control group samples were set as 0.0001 and 0, respectively. The fraction of indels in all variants was restricted to 20%. The expected VAF ranged from 0.1 to 0.5. All other parameters were set as default values. Besides the mutations and indels, the simulated WGS dataset also included structural variants including deletions, duplications and translocations longer than 50 bp. DWGSM generates fastq format files that are directly used as input for 2-kupl.

3.2.3.12 TCGA-PRAD data analysis

Matched normal-tumor WES data of 498 patients from TCGA-PRAD (Prostate Adenocarcinoma) (Abeshouse et al., 2015) were retrieved with permission from dbGAP (Tryka et al., 2014). BAM files were converted to paired-ends fastq files using Picard tools with default parameters. 2-kupl somatic variant calls were obtained for each normal/tumor pair using default parameters. Detailed analysis of variant calling was performed on the TCGA-PRAD sample with the highest tumor mutational burden (barcode TCGA-ZG-A9ND).

2-kupl results on the TCGA-PRAD dataset were compared to variant calls downloaded from the GDC portal. Briefly, the GDC portal workflow uses BWA to map reads to the human genome and determines variants with five state of the art variant callers, as described here: https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/. We used the maftools R package (Mayakonda et al., 2018) to retrieve variants predicted using the GATK-MuTect2 pipeline and filtered against a "panel of normals". This mutation dataset is hereafter referred to as the "GDC portal" dataset.

To remove putative germline variants from 2-kupl results, we built a boolean matrix representing the presence of each k-mer in each normal sample. Any k-mer present in at least two normal samples was excluded. Retained recurrent variants were

considered as tumor-specific (Suppl. Table S1). Mutations detected by 2-kupl and absent in the GDC portal variants were considered as 2-kupl specific. To verify whether calls absent in GDC portal variants were not discarded at earlier stages of the GDC portal pipeline, we also retrieved the protected MAF file containing all unfiltered variants called by the MuTect2 workflow.

The oncoplot graph for GDC portal variants (Fig 7B) was drawn using maftools. To obtain recurrently mutated genes by 2-kupl, we aggregated variants belonging to the same gene in 2-kupl results and constructed a gene-level occurrence matrix that was fed to maftools (Fig 7C). Recurrent variants from 2-kupl and the GDC Portal were also compared with a comprehensive prostate cancer dataset from 200 whole-genome sequences and 277 whole-exome sequences from localized prostate tumours(Fraser et al., 2017) (Suppl. Table S2)

Recurrently mutated genes were annotated using a collection of 1404 PRAD-related genes collected from CLINVAR (Landrum et al., 2018), COSMIC (Bamford et al., 2004), DISEASE (Pletscher-Frankild et al., 2015), KEGG (Kanehisa et al., 2017), OMIM (Hamosh et al., 2005), PheGenl (Ramos et al., 2014) and driver predictions by Martincorena et al. and Armenia et al. (Martincorena et al., 2017; Armenia et al., 2018) (Suppl. Table S3).

3.2.3.13 Bacterial genome analysis

We obtained WGS fastq files from the Pseudomonas aeruginosa PAO1Or wild-type strain and 24 phage-tolerant mutants (Latino, 2016). Mutations in the phage-tolerant variants were previously validated by mapping of the WGS raw sequences to the PAO1Or genome (Genbank accession LN871187) and confirmed by PCR amplification and Sanger sequencing. We used one control WGS file and 21 mutant WGS files corresponding to 26 validated variants. Detailed variants (Suppl. Table S4) include seven mutations, 13 small indels and six large deletions longer than 100 bp. 2-kupl was run using default parameters on every mutant WGS file compared to the control WGS file.

3.2.4 Results

3.2.4.1 A novel algorithm for detecting variants between two DNA-seq samples

We developed 2-kupl to predict variants between pairs of matched DNA-seq libraries. Input libraries consist of a "case" and a "control" sample such as a pair of tumor and normal tissues from one patient or a pair of mutant and wild-type bacterial strains. Data can be either WGS or WES. 2-kupl extracts case-specific k-mers (cs-kmers) and matching control k-mers (ct-kmers) corresponding to a putative mutant and reference sequences and merges them into contigs. As 2-kupl begins with a shortlist of cs-kmers, the number of k-mers considered from unaltered regions and non-specific variants is drastically reduced compared with DBG-based methods (see Methods). If a reference genome is provided, 2-kupl can also align contigs to the reference and generate genomic coordinates just like with mapping-based methods.

3.2.4.2 Performance on simulated WES data

We first applied 2-kupl to the detection of somatic mutations in a simulated human cancer WES dataset containing a known number of spliked-in mutations and indels. We compared 2-kupl with three other software, including two mapping-free methods (DiscoSNP++ and Lancet) and the leading mapping-based pipeline GATK-MuTect2. Results are summarized in the first column of Table 7. The number of cs-kmers to process is reduced by nearly 20% after data cleaning by 2-kupl.

88.6% of cs-kmers were matched to ct-kmer, corresponding to predicted point mutations or indels. We evaluated mutations and indel calls by 2-kupl and concurrent methods (Table 1). For mutation calling, 2-kupl performed better than the other mapping-free methods in terms of F1 score (Table 1). Lancet and GATK achieved better recall than 2-kupl, but Lancet also introduced more false positives. 2-kupl had a higher recall for calling indels than DiscoSNP++ and Lancet but

Table 1: Comparison of four approaches on mutations using simulated WES data

mutations	2-kupl	DiscoSNP++	Lancet	GATK-MuTect2
True Positive	581	373	604	689
False Positive	45	3	126	2
False Negative	241	530	218	133
Recall	0.71	0.41	0.73	0.84
FDR	0.07	0.01	0.17	0.003
Precision	0.93	0.99	0.83	0.997
F1 score	0.80	0.58	0.78	0.91

Table 2: Comparison of four approaches on indels using simulated WES data

indels	2-kupl	DiscoSNP++	Lancet	GATK-MuTect2
True Positive	42	29	40	49
False Positive	16	1	44	26
False Negative	39	52	41	32
Recall	0.52	0.36	0.49	0.60
FDR	0.27	0.03	0.52	0.35
Precision	0.72	0.97	0.47	0.65
F1 score	0.60	0.52	0.48	0.63

was outperformed by DiscoSNP++ in FDR and precision (Table 2). Expectedly, GATK-MuTect2 outperformed all mapping-free approaches regardless of variant types. DiscoSNP++ did not perform as well as others in terms of recall ratio due to the different usage. DiscoSNP++ first pooled together two samples and screened case-specific variants afterwards. This procedure contributes to eliminate many false positives but also leads to ignoring some low frequency variants exclusively present in the case sample. Lancet performed well in terms of recall but at a high cost of false positives. As expected, most false positives had few reads containing the alternative allele, which is frequent with Lancet. The high recall and high rate of false positives produced by Lancet are consistent with the conclusions of Meng and Chen (Meng and Chen, 2018). The GATK-MuTect2 pipeline outperformed all mapping-free approaches when calling mutations. The use of a reference sequence and the Haplotype Caller algorithm gives GATK-MuTect2 a clear advantage. Even though 2-kupl got a relatively lower recall than GATK-MuTect2, it had better control of the false positives and got a higher precision when calling indels (Table 2).

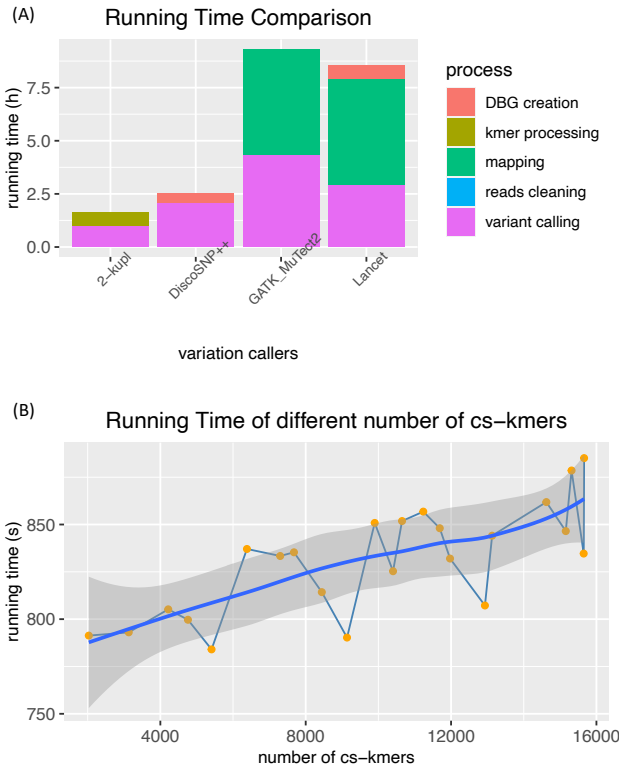


Figure 3: Running time and performance with different types of variants. . (A) Overall running times of four software. The time consumed by each process in four protocols is marked in different colors. (B) Running times of 2-kupl for different numbers of cs-kmers. The line with dots represents the exact running time corresponding to certain number of cs-kmers. The solid line is the fitted line, and the shaded background is the confidence interval.

Another advantage of 2-kupl is the short running time (Fig 3A). 2-kupl took 1.6 hours to analyze the simulated WES data with default parameters. DiscoSNP++ took 2.54 hours to call variants from both case and control samples. Both Lancet and GATK-MuTect2 require prior mapping of reads to the human genome (which takes 3.17 hours), explaining in part their longer runtimes.

To evaluate 2-kupl run time dependency on the number of cs-kmers, we ran 2-kupl on datasets with different numbers of cs-kmers (Fig 3B). Running time increased linearly with the number of cs-kmers. Each additional 10,000 cs-kmers increased the running time by nearly 50 seconds.

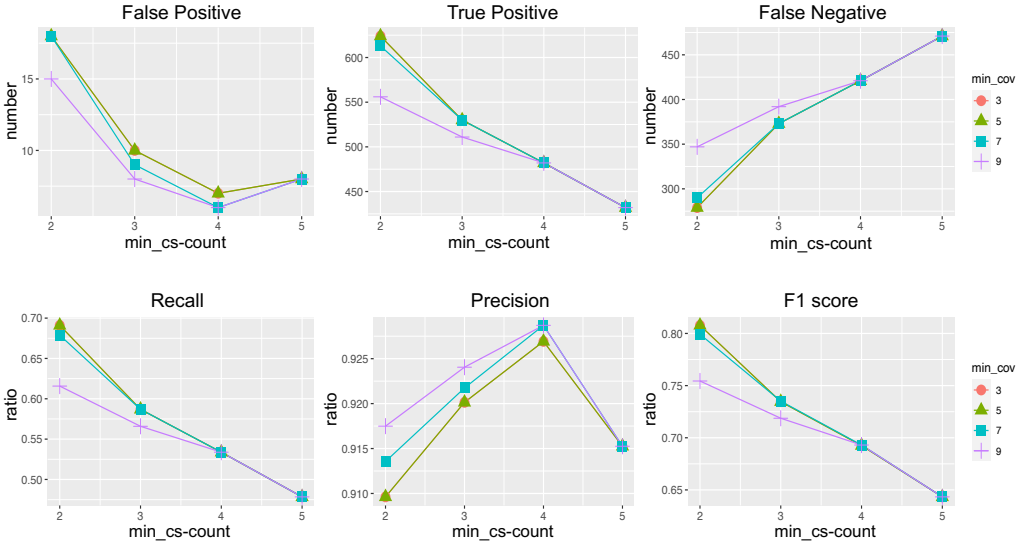


Figure 4: Robustness of 2-kupl using different parameters. The x-axis indicates the min_cs-count parameter and the y-axis represents the corresponding ratio or number. The thresholds of coverage and cs-count are denoted as min_cov and min_cs-count, respectively. The trend lines under different min_cov parameters are represented by four colors.

We estimated the performance of 2-kupl under different parameter combinations. Coverage and cs-count thresholds ('mim_cov' and 'min_cs-count', respectively) were varied from 3 to 9. Results are shown in Fig 4. The min_cs-count parameter was negatively related to recall and positively related to false negatives. The min_cov parameter was inversely related to F1 score, recall, FDR, and true positives. Precision reached an inflection point when min_cs-count was set to 4.

3.2.4.3 Performance on simulated WGS data

We further benchmarked 2-kupl on a simulated WGS dataset with an average read depth of 50X (vs. 230 in WES). For mutation calls, 2-kupl and GATK-MuTect2 achieved the same recall ratio of 0.86 (Table 3). The precision of 2-kupl was slightly lower than GATK-MuTect2 but still above 0.9. For indels, the recall of 2-kupl dropped to 0.82 (Table 4). The false positive call rates of 2-kupl increased with

Table 3: Comparison of 2-kupl and GATK-MuTect2 on mutations using simulated WGS data

mutations	2-kupl	GATK-MuTect2
True Positive	13835	13920
False Positive	1248	30
False Negative	2220	2135
Recall	0.86	0.86
FDR	0.08	0.002
Precision	0.91	0.99
F1 score	0.89	0.93

WGS data relative to WES data due to the lower coverage of WGS. A limitation of 2-kupl is that false signals can not be ruled out by allele frequency in low coverage regions. Also, k-mers may be incorrectly considered as cs-kmers when there is not enough reads covering the locus in the control sample.

The simulated WGS dataset contained 157 SVs (deletions, duplications, and translocations longer than 50bp). Expectedly, GATK-MuTect failed to detect the majority of SVs (Table 5). We thus compared 2-kupl with Delly, a software that finds structural variants based on aligned reads (Rausch et al., 2012). Overall 2-kupl had a slightly lower precision and recall than Delly (Table 5). We investigated 22 SVs missed by Delly and captured by 2-kupl. We found these reads were left unmapped by BWA due to multiple hits in the genome and thus could not be assessed by Delly. An advantage of 2-kupl here is that all k-mers covering SV junctions are kept and assembled regardless of mapping status. Furthermore, 2-kupl is capable of detecting small variants in the same run.

3.2.4.4 Assessing 2-kupl on a real normal-tumor WES dataset

To assess 2-kupl results on actual WES data, we applied 2-kupl on one WES dataset of matched tumor and normal tissues from the TCGA-PRAD dataset. We first compared 2-kupl and GDC portal somatic variant calls (see Methods) on the TCGA patient with the highest tumor mutational burden. The numbers of k-mers, contigs

Table 4: Comparison of 2-kupl and GATK-MuTect2 on indels using simulated WGS data

indels	**2-kupl**	**GATK-MuTect2**
True Positive	3315	3620
False Positive	504	108
False Negative	750	445
Recall	0.82	0.89
FDR	0.13	0.02
Precision	0.84	0.96
F1 score	0.84	0.92

Table 5: Comparison of 2-kupl, GATK-MuTect2 and Delly on structural variants using simulated WGS data

mutations	**2-kupl**	**GATK-MuTect2**	**Delly**
True Positive	133	49	135
False Positive	27	0	16
False Negative	24	108	22
Recall	0.85	0.3	0.86
FDR	0.17	0	0.11
Precision	0.83	1	0.89
F1 score	0.84	0.47	0.88

Table 6: Number of mutations and indels detected by 2-kupl and GDC portal variants

	2-kupl	GDC portal variants	overlap
mutation	3607	3093	319
indel	151	823	8
total	3758	3916	327

Table 7: Number of k-mers and contigs after applying 2-kupl on two matched libraries

	simulated WES	TCGA-ZG-A9ND WES
all k-mers(Tumor/Normal)	465,718,268/465,610,133	184,233,006/177,517,776
raw cs-kmers	23599	393525
cleaned cs-kmers	18439	291350
matched cs-kmers	16914	240360
all contigs	1245	106426
mutations	1026	9901
indels	112	1105
unmapped	0	58
low confidence	107	312

and variants obtained by 2-kupl are shown in the second column of Table 7. Mutation calls by 2-kupl and GDC portal variants are shown in Table 6. Although total call numbers were similar, only 327 calls (9%) were shared by the two approaches, including 319 mutations and 8 indels. Among the variants detected by 2-kupl, 193 (5.13%) mapped to noncoding regions and 101 (2.7%) were annotated as repeats by RepeatMasker (de Koning et al., 2011). 2-kupl also captured 57 (1.5%) unmapped variants. 173 2-kupl variants (4.6%) were mapped to low mappability "blacklist" regions (Amemiya et al., 2019). In spite of the small general overlap of 2-kupl and GDC portal variants, the two methods have a much stronger agreement on high scoring 2-kupl calls (Fig. S1A). Of note, mutation calls obtained on on the same sample by four different mapping-based protocols also show poor consistency (Fig. S1B).

We further analyzed mutations specific to 2-kupl. These calls may have been rejected in GDC portal variants for a number of valid reasons, including low mapping quality, location in short tandem repeats or presence in normal samples. A real

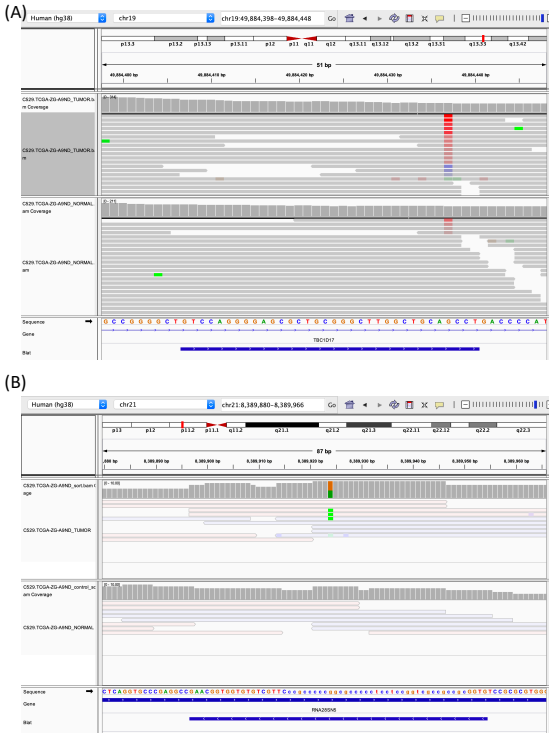


Figure 5: IGV views of variant calls in TCGA-PRAD WES dataset. The two central tracks show aligned reads from the tumor (top) and normal (bottom) WES library. The lower track shows gene annotation and 2-kupl contigs. (A) A likely false-positive call by 2-kupl at a position of low mapping quality (B) A likely true positive within a repeat region. Reads in transparent color have low MAPQ (mapping quality) values (<10).

"miss" by the reference-based pipeline should be recorded only when reads could not possibly be aligned to the genome while they indeed contained a valid mutation.

Fig 5A shows a case of false positives introduced due to artifactual cs-kmers. Generally, k-mers harboring a mutation present in both tumor and normal tissues are supposed to be ruled out. However, erroneous tumor-specific "cs-kmers" can escape the filtering process if the same k-mer in the normal tissue happens to be low quality and is discarded.

Certain 2-kupl specific mutations are possibly true positives discarded by mapping-

based protocols due to their location within a repeat region. Fig 5B shows such a potential somatic mutation. The mutation is located within a ribosomal RNA gene that is repeated multiple times in the genome and further contains a C-rich repeat (represented in lower cases). Reads generated from these repetitive regions are given low MAPQ values by mappers and variants in these regions are then discarded by variant callers.

Among unmapped 2-kupl calls, only one has a Phred score in the top 5% (Fig. S2). The mutant sequence and its inferred reference are shown in Fig. S3. The mutant contig is covered by 0 and 47 reads in the Normal and Tumor sample, respectively while the reference is covered by 88 and 65 reads in the Normal and Tumor sample, respectively (Fig 6). The sequence maps to a centromeric repeat of Chr22, with three mismatches. The mapping procedure would thus miss this highly significant variant.

3.2.4.5 Recurrent mutations in TCGA-PRAD

Recurrence across patients is a powerful criterion for distinguishing drivers from passenger mutations (Pon and Marra, 2015; Greenman et al., 2007; Goncearenco et al., 2017) and has been used to discover drivers and define molecular subtypes of prostate cancer (Barbieri et al., 2012). We applied 2-kupl to each pair of Normal/Tumor samples in the complete PRAD WES dataset (N=498) and identified 3211 recurrent variants (suppl. Table S1). For comparison we retrieved from the GDC portal recurrent variants predicted for the same dataset (GATK-MuTect2 pipeline, see Methods). Among 3734 recurrent variants in the GDC portal, 854 were shared with 2-kupl recurrent variants (suppl. Table S1). We further compared the recurrent variants to a comprehensive dataset of recurrent prostate cancer mutations from Fraser et al. (Fraser et al., 2017) based on 200 whole-genome and 277 whole-exome sequences from multiple sources. Comparisons were restricted to exonic regions. Within the 48 recurrent mutations in exonic regions from Fraser et al, a similar number was shared with 2-kupl or the GDC-portal (22 and 21, respec-

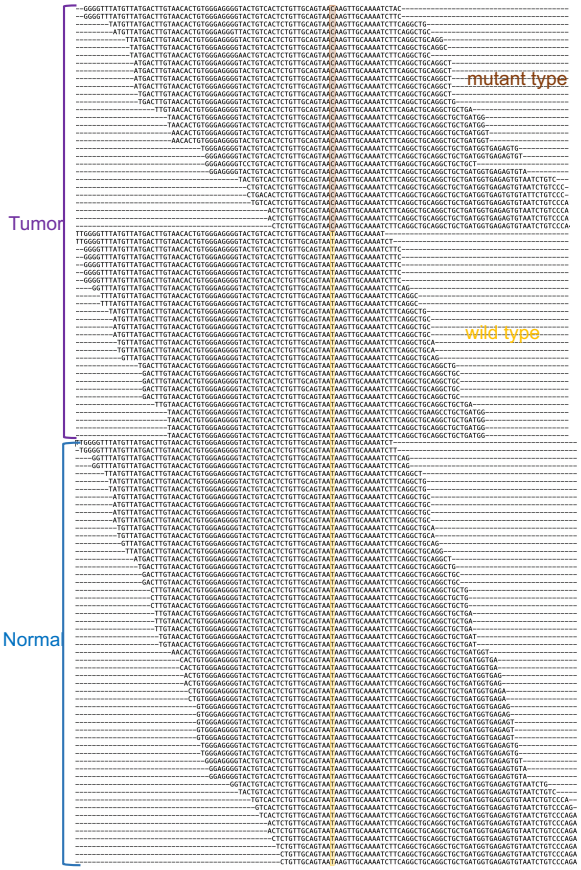


Figure 6: An unmapped somatic variant from a TCGA-PRAD patient. Only reads matching the central k-mer of the tumor-specific variant or its inferred counterpart are shown. Reads from the tumor and normal samples are distinguished. The position of variation is highlighted.

tively) (Suppl. Table S2). Among recurrent mutations specific to 2-kupl, we note the one found at chr14:37592023 within an exon of FOXA1, a putative prostate cancer driver (Martincorena et al., 2017), in three TCGA-PRAD patients.

We further compared 2-kupl calls to GDC portal variants at the level of genes (Detailed in Method section). The GDC portal reported 6944 genes mutated in two or more patients, vs. 14137 recurrent genes by 2-kupl. Enrichment analysis shows a good convergence of the most frequently mutated genes by the two methods (Fig 7A). Fig 7B,C show oncoplot views of the top 20 genes according to the GDC portal and 2-kupl, respectively, showing eight shared genes. Both gene lists are contaminated by long (TTN) or highly polymorphic genes (Mucins) whose recurrence is an artifact due to higher mutation counts. Although many software are available to account for those effects (Li et al., 2015a), we purposely analyze the uncorrected list of genes here. Among the top 20 mutated genes by 2-kupl and GDC portal, 7 and 9 genes, respectively, are known prostate cancer-related genes. Among those, UBR4, DNAH5 and LRP1 were only detected by 2-kupl. When considering the top 50 recurrently mutated genes according to 2-kupl and GDC portal, 19 and 23, respectively, are cancer-related. Among those, HSPG2, DNAH3, UBR4, COL6A3, CABIN1, IGF2R, PTPRF, DNAH5, HTT and TRRAP were only detected by 2-kupl.

UBR4 contains 48 2-kupl mutations, more than any other gene. Fig. S4 shows read alignment at this gene for patient TCGA-EJ-7125 who carries the most UBR4 mutations (8/48 mutations). While seven of these mutations are absent in GDC portal variants, all can be visually validated as tumor-specific mutations as per the IGV display (Fig. S4 A-G).

Besides recurrent mutations and indels, we found 20 genes with 43 recurrent structural variants predicted in at least two patients (suppl. Table S1). All these predicted variants can be supported by at least one read from the tumor library. Three recurrent structural variants map to prostate cancer genes SH2B3, ATP10A and FOXA1 (Fig 8). Variants in gene ATP10A and SH2B3 have exactly the same junc-

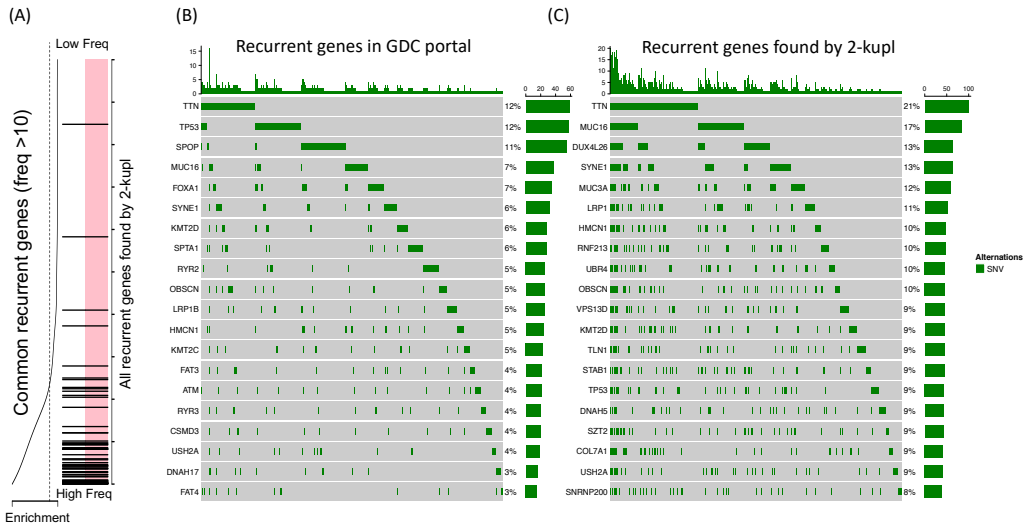


Figure 7: Recurrently mutated genes in the TCGA-PRAD WES dataset. (A) Enrichment analysis of recurrent genes. The vertical bars are the common recurrently mutated genes (altered in at least ten patients) between GDC portal and 2-kupl. The x axis represents the recurrent genes found by 2-kupl sorted by frequency. The smooth curve reflects the degree to which the common genes are overrepresented in the whole 2-kupl recurrent genes. (B) The 20 genes with the highest mutational frequency detected in GDC portal variants. (C) The top 20 recurrent genes with the highest mutational frequency detected by 2-kupl.

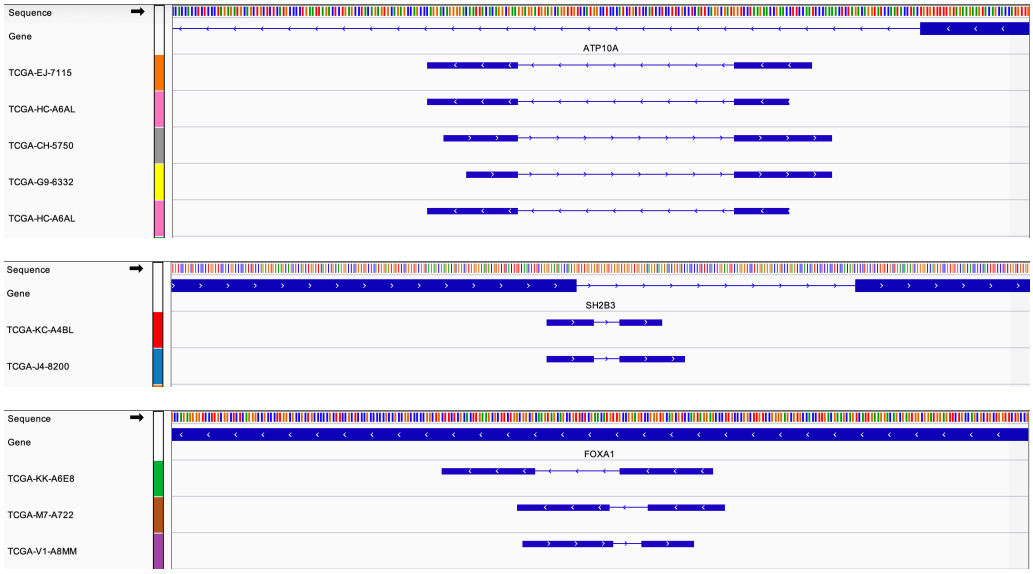


Figure 8: Recurrent structural variants mapping to three prostate cancer genes. In each track, lines represent the genome sequence (top), annotated genes, and variant contigs identified in different patients.

tions in at least two patients. As the three variants in gene FOXA1 impact on the same exon, we grouped them as one same recurrent event despite not representing the exact same variation. All these recurrent structural variants are longer than 10bp. State-of-the-art procedures usually miss such variants at the mapping stage.

3.2.4.6 Performance on bacterial WGS data

2-kupl can be applied to pairwise comparisons of DNA-seq datasets in any species. We present here an application to bacterial whole genome sequences. A frequent problem in bacterial genetics is identifying mutations in strains for which no reliable reference genome is available. We investigated the performance of 2-kupl on 21 DNA-seq datasets from a Pseudomonas aeruginosa strain, in which 26 variants had been previously identified and confirmed by geneticists (see Methods).

About 141 variant contigs were predicted on average for each pair of WT/mutant strains, with an average running time of 10 minutes (Fig 9A,B). Score ranking by 2-kupl and DiscoSNP++ allowed a clear separation of TP from FP (Fig 9C,D). True

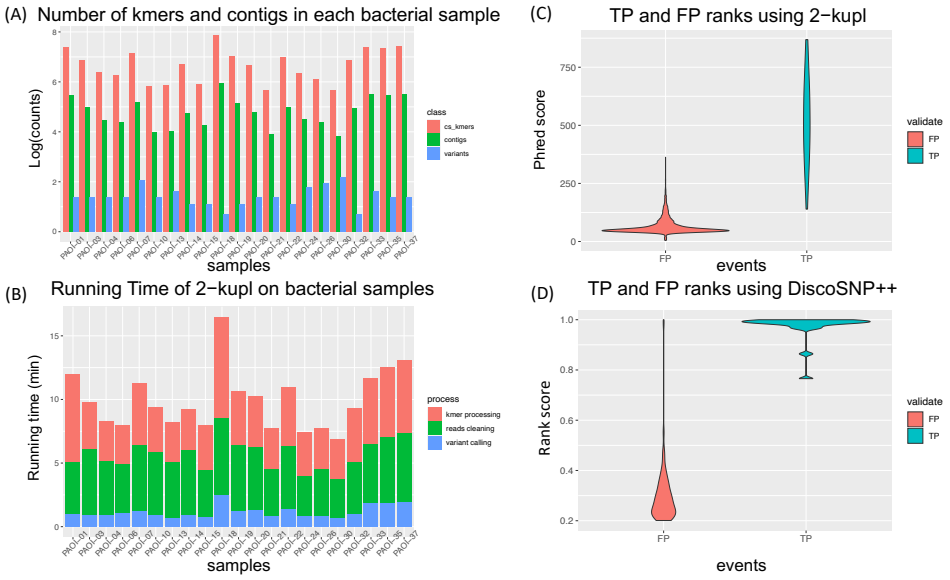


Figure 9: Performance of 2-kupl on bacterial DNA-seq datasets. (A) Number of cs-kmers, contigs and variants are shown for each bacterial sample. (B) Running time of 2-kupl on each sample is shown for different steps. (C) Distribution of Phred scores computed by 2-kupl in TP and FP events. (D) Distribution of DiscoSNP++ score ranks in TP and FP events.

positive calls were ranked first in 19 out of 19 mutant samples by 2-kupl and in 16 out of 16 samples by DiscoSNP++. Compared with Phred scores used in 2-kupl, DiscoSNP++ scales the rank scores from zero to one and thus the true positive variants are more concentrated.

2-kupl could recall all true positive variants, including SNVs and large deletions longer than 100 bp, while DiscoSNP++ missed three large deletions (555 bp, 213 bp and 109 bp, suppl. Table S4). Meanwhile, DiscoSNP++ obtained 129 false positives vs. 45 for 2-kupl (Table 8). Therefore 2-kupl had the best recall and precision on this dataset, especially for large indels.

3.2.5 Discussion

Most variant detection protocols rely on reference genomes. However, even for species with a high-quality reference genome such as humans, depending on a ref-

Table 8: comparison between 2-kupl and DiscoSNP++ on the bacteria DNA-seq data

	2-kupl	**DiscoSNP++**
True Positive	26	23
False Positive	45	129
False Negative	0	3
Recall	1	0.88
FDR	0.64	0.85
Precision	0.36	0.15
F1 score	0.52	0.26

erence is subject to limitations. Genomes contain large numbers of highly variable, repetitive or otherwise unmappable regions, which are unsolvable by short-read sequencing techniques. Hundreds of unsolved regions remain in telomeres and centromeres, also known as 'dark matter' (Blaxter, 2010). The X chromosome is the only complete human chromosome as of today (Miga et al., 2020). Pathogenic variants within these unannotated regions are easily missed by mapping-based approaches due to low mapping quality, especially with low depth in whole-genome sequencing. Furthermore, the human genome varies across individuals and populations and a single reference genome does not account for this diversity (Sherman et al., 2019).

2-kupl is able to detect variants, including mutations, indels and structural variants, without relying on a reference genome. Based on matched DNA-seq data, 2-kupl captures case-specific k-mers and counterpart k-mers (i.e. without the variation) into the same bucket. Sequence contigs harboring a local variation and its putative reference are inferred through the assembly of k-mers in each bucket.

To control artifacts induced by sequencing errors, 2-kupl takes both base quality and coverage into account. The general sequencing error rate in short-read NGS data is larger than 0.1% (Ma et al., 2019). It is worth consuming computing resources and running time to remove these 0.1% artifacts because these sequencing errors result in large numbers of artifactual cs-kmers. To reduce the impact from low-quality bases, we combine Cutadapt and an 'OverrideN' function that flags low quality

bases in the mid part of reads. This significantly reduces the number of cs-kmers and speeds up the computing procedure.

We compared the performance of 2-kupl with that of three competing methods in terms of running time, recall and precision. 2-kupl outperformed mapping-free methods DiscoSNP++ and Lancet in terms of recall or precision but did not reach the performance of the state-of-the-art alignment-based GATK-MuTect2 on human data.

DiscoSNP++ suffers from limitations of DBG data structures in regions with sequencing errors, genomic variants and repeats (Heydari et al., 2019). Efficient solutions searching for bubbles from such complicated structures are still under development. Furthermore, short contigs may be discarded within the post-process, cutting branches, for instance (Medvedev et al., 2011). In our bacterial DNA-seq analysis, DiscoSNP++ missed three validated large deletions.

Lancet has a higher recall ratio than 2-kupl but also introduces more false positives. Furthermore, Lancet missed variants from repetitive regions and is not able to detect fusions from distant regions.

2-kupl has a higher F1 score than DiscoSNP++ and Lancet and performs better in terms of recall ratio or precision than either of them. Expectedly, 2-kupl did not outperform GATK-MuTect2 on WES data. First, GATK-MuTect2 uses a sophisticated Bayesian model to estimate a genotype's likelihood given the observed sequence reads that cover the locus. When GATK-MuTect2 encounters a region showing signs of variation, it discards the existing mapping information and completely reassembles the reads in that region. This allows GATK-MuTect2 to be more accurate when calling regions that are traditionally difficult to call. Despite slightly fewer true positives, 2-kupl also detects fewer false positives than GATK-MuTect2. It is worth mentioning that 2-kupl has the lowest time complexity among the four methods.

By applying 2-kupl to the TCGA-PRAD patients, we were able to detect recur-

rent mutations and indels missed by the GDC portal's GATK-MuTect2 pipeline. Reads in these regions have either low mapping qualities or multiple hits and were discarded in the GDC portal pipeline. Mapping-based methods all suffer from this issue and are powerless when faced with low complexity regions. 2-kupl identified recurrent mutations and recurrently mutated genes in high agreement with GATK-MuTect2. Mutated genes were enriched in PRAD-related genes, some of which specific to 2-kupl. As an example, we visually confirmed multiple 2-kupl-specific mutations in UBR4. Recurrent variants detected from the unmappable regions by 2-kupl provide insights into potential novel somatic variants even though the locus of origin of the contig sometimes cannot be determined.

Standard variant calling pipelines may miss mutations for multiple reasons: low allele frequencies, tumor contamination, ambiguities in short read alignment, inadequate sequencing depth, high GC content, sequencing errors and ambiguities in short read alignment. Different programs are affected by these factors to varying degrees. As a consequence, the mutations called by different pipelines are not consistent (Hwang et al., 2015). 2-kupl is not affected by some of these sources (GC content, alignment artifacts and mappability) and can detect a number of recurrent mutations (ie. potential driver events) that are not found by standard pipelines.

Several natural directions exist for extending 2-kupl. First, 2-kupl lacks sensitivity in detecting structural variants. All cs-kmers covering the junction are retained and extended to contigs. Unfortunately, neither the ct-kmers nor the reads are easily obtained when considering a hamming distance of one. A structural variation can be detected only if enough supporting reads are covering at least one side of the variation. Focusing on the cs-kmers regardless of ct-kmers could address this problem but at the cost of more false positives. A second limitation occurs when control samples are contaminated with tumor cells, which is relatively frequent in tissue biopsies. To address this problem, 2-kupl includes a parameter representing a k-mer count threshold in the control sample. However, a fixed contamination threshold may introduce unwanted non-specific variants. Future works should evaluate probabilistic approaches to address this issue.

3.2.6 Conclusion

In conclusion, the identification of different kinds of variants, using DNA-seq data, remains challenging. The leading protocols developed for DNA-seq highly rely on the reference. In general, the methods that align sequencing data to the reference (mapping-based methods), perform better than do the mapping-free methods. However, 2-kupl can capture events falling into the difficult-to-map regions, and can perform better than other mapping-free protocols. 2-kupl is the fastest tool in the comparison with other methods because the mapping procedure is not included. The high agreement in top ranking variants by 2-kupl and GDC portal variants indicates the capacity of using 2-kupl as an extension and supplementation of the mapping-based methods. New recurrent variants and genes relevant to prostate cancer are captured by 2-kupl.

3.2.7 Additional Files

Table S1. This supplementary table includes recurrent SNVs, SVs and mutated genes identified by 2-kupl.

Table S2. Comparison with the Fraser et al's recurrent PRAD mutations.

Table S3. Prostate cancer related genes collected from various resources.

Table S4. True positive variants in the bacterial WGS data.

Table S5. 2-kupl detected structural variants that are missed by Delly.

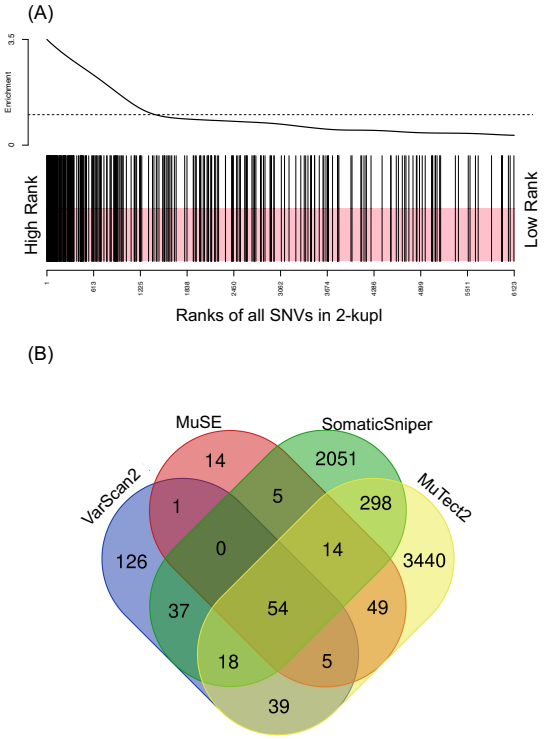


Figure S1: The distribution of shared SNVs in 2kupl and consistency of four mapping-based protocols.

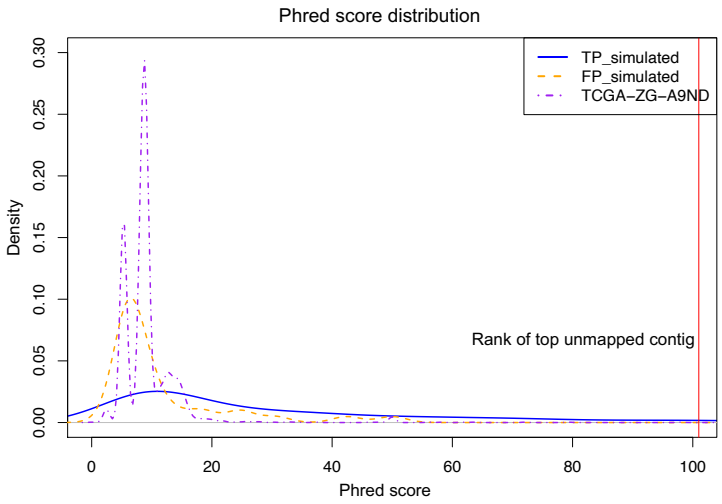


Figure S2: Phred score distribution.

```
GGGAGGGGTACTGTCACTCTGTTGCAGTAACAAGTTGCAAAATCTTCAGGCTGCAGGCTGCT
|||||||||||||||||||||||||||||||.|||||||||||||||||||||||||||||
GGGAGGGGTACTGTCACTCTGTTGCAGTAATAAGTTGCAAAATCTTCAGGCTGCAGGCTGCT
Score=302.5    47       112      0        88
```

Figure S3: Alignment of the mutant contig and inferred reference from one unmapped event.

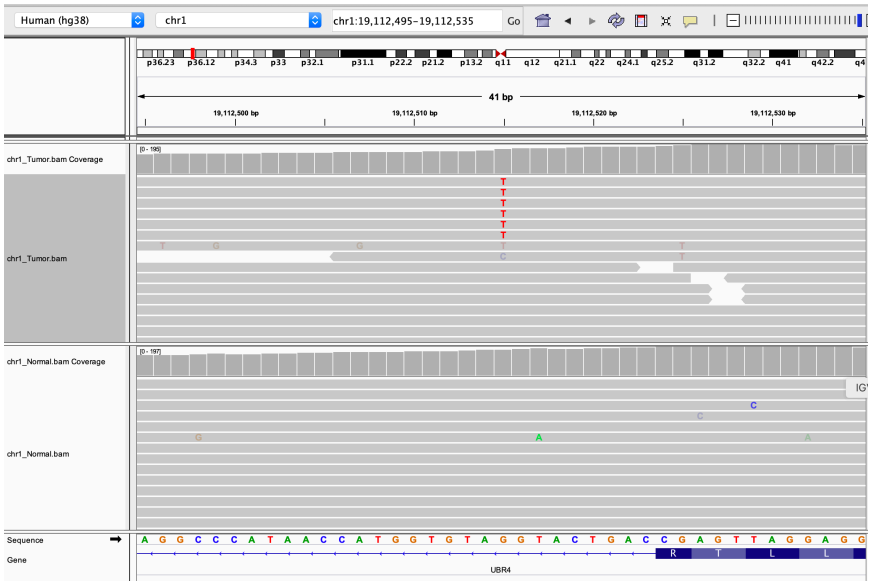


Figure S4: IGV views of UBR4 mutations occurred on patient of TCGA-EJ-7125

Chapter 4

Discussion and Perspectives

4.1 A new stratification of lung cancer patients with potential therapeutic benefits

The advent of precision medicine with targeted therapeutic options with companion diagnostics and immunotherapy is revolutionizing the treatment in oncology, especially in lung cancer. Accurate classification of tumor sub-types is an essential step for personalized treatment of cancer. Many studies have further divided LUAD patients into subgroups based on gene expression, mutation profiles, and immune signatures (Hu et al., 2019; Ding et al., 2020; Xu et al., 2020). To our knowledge, no study has ever used data originated from repetitive regions to define LUAD subgroups. In fact, the complex variation in repetitive regions is a direct indicator of genome instability. For instance, Alu repeats are shown to be associated with many microsatellite instabilities, which plays a critical role in oncogenesis (Arcot et al., 1995). Repeats also impact genomes by inducing small variants, recombinatory events, gene conversion, and abnormal gene expression (Batzer and Deininger, 2002). Therefore, patients with different repeat characteristics may present diverse immune responses or prognosis.

In Chapter 3.1.4.6, we applied DE-kupl to identify differentially expressed RNA elements in two independent LUAD cohorts. We identified a large set of differential RNA elements that were recurrent in both datasets, including SNVs, intronic events, splice events, repeats and others. We classified patients into two subgroups based on the Alu and L1P repeats: the two principal repeat types in the whole set of repeats with differential expression. Despite no significant differences in survival status, gender, age, or tumor stage, we observed differences in immune cell contents between the two subgroups. Patients in the "high repeat" subgroup had lower immune infiltration levels, consistent with previous observations that an immunosuppressed environment favors transposable element activity (Zhang et al., 2020b; Jung et al., 2018). The "high repeat" group also had a higher mutation and CNV burden, suggesting repeat activation may be a cause of genomic instability in LUAD. It is also consistent with the previous conclusions that repeats can cause variations at the transcriptome and genome levels (Lanciano and Cristofari, 2020; Payer and Burns, 2019). Generally, patients with high TMB benefit from immunotherapy. However, we found that the subgroup of LUAD patients with high TMB presents a lower level of immune infiltration. It implies the need of new stratification especially in high TMB patients to indicate their sensitivity to immunotherapy.

The identification of new patient groups with different levels of immune response and genome instability is important for immunotherapy (Litchfield et al., 2021). Therefore we hope our approach could serve as a basis for new patient stratification schemes that would take into account repeats along with more conventional information such as mutation burden, immune checkpoint expression or tumor clonality.

4.2 Candidate neoantigens for vaccine development

The discovery of tumor-specific antigens, or neoantigen, is a key step towards the development of cancer vaccines. Canonical strategies for screening candidate neoantigens consist of several main steps that primarily rely on DNA-seq and RNA-seq data (Gopanenko et al., 2020). First, tumor-specific variants are selected from the

DNA-seq data using variant callers such as GATK. Second, the expression of selected variants in the tumor is verified in the RNA-seq data. Finally, the peptides translated from tumor-specific variants are evaluated for presentation by the MHC class I complex (Gopanenko et al., 2020).

The initial steps of neoantigen prediction rely on alignment to the reference. This strategy can guarantee the security and applicability of candidate neoantigens. However, mapping-based protocols neglect a wide diversity of possible neoantigens. Variants located within unannotated regions of the current human genome are excluded. Variants in low complexity regions such as repeats are difficult to detect by mapping-based callers. Complex variants such as aggregated mutations or structural variants are also challenging to detect because aligners often discard these reads due to the low alignment score. Furthermore, if WES data is used for antigen prediction, variants in intronic regions or noncoding regions are not included. Yet, it has been shown that all these regions can be translated and form neoantigens (Ouspenskaia et al., 2020; Laumont et al., 2018).

In Chapter 3.1.4.7, we identified candidate neoantigen-forming RNAs in exons, introns, lincRNAs and repeats. Exonic candidates are relatively simple to discover by standard alignment-based protocols. To our knowledge, there has not been any research systematically evaluating the potential of intronic, intergenic and repeat regions producing neoantigens. DE-kupl is capable to screen multiple types of tumor-specific events at once. High-quality candidates for neoantigen formation shared by two LUAD datasets were extracted. These tumor-specific events are all recurrently transcribed in at least 15% of LUAD patients, which makes them valuable as a source of "public neoantigens" for vaccine development.

4.3 The potential therapeutic value of novel events as drug targets

Drug target discovery efforts have been increasing over the past decades in response to decreased efficacy due to tolerance of existing drugs. While most drug targets to date have been proteins, a growing diversity of miRNAs and lincRNAs are now considered as potential targets (Ling et al., 2013). However, standard alignment-based methods limit the screening of novel RNAs that could serve as targets. We discovered different types of transcriptional events using DE-kupl (and genetic events using 2-kupl) which cannot be identified by standard methods. Although we have not thoroughly explored the medicinal properties of these events, we hope our work provides new insights for expanding the scope of drug target screening.

4.4 Novel recurrent variants in difficult-to-map regions

There are large numbers of difficult-to-map regions in the genome that are not covered with standard NGS strategies and may contain clinically important variants. These regions include multicopy pseudogenes or other repetitive genomic regions. Addressing these difficult-to-map regions has been a challenging problem to overcome by standard protocols. Since the early characterizations of mobile endogenous retroviruses, numerous repetitive DNA sequences have been discovered to comprise the majority of the human genome. The human genome has domesticated many repetitive elements as regulators of transcription and genome organization (Ishak and De Carvalho, 2020). Cancers usurp repetitive elements to disturb transcriptional networks and promote genome instability. This thesis work found recurrent RNA elements expressed from these difficult-to-map regions, which are transcribed in multiple tumor samples and lowly expressed or even silent in all normal tissues. The occurrence of these events in multiple patients from independent populations

indicates the presence of common unidentified oncogenic mechanisms.

In Chapter 3.2, we identified recurrent somatic variants in prostate adenocarcinoma patients with our newly developed in-house software 2-kupl. 2-kupl achieved high concordance with the standard alignment-based variant callers such as GATK. In addition, 2-kupl detected numerous somatic variants from the difficult-to-map regions, most of which were not detectable by GATK. For instance, 2-kupl identified novel mutations in some key genes, suggesting their mutation rates may be underestimated by standard methods. Furthermore, 2-kupl detected structural variants, in which case reads covering the SV junctions are either unmapped or have multiple hits on the genome and thus are normally discarded by standard variant callers. Dedicated SV callers also miss SVs when their supporting reads are not properly aligned. Still, recurrent SVs hidden in these difficult-to-map regions are potentially disease-related (Levy-Sakin et al., 2019).

There are still a lot of poorly explored regions in the human genome beyond reference annotation. Mapping-free approaches were developed to uncover the genetic information within these regions. However, their accuracy remains generally lower than that of standard protocols. Furthermore, mapping-free approaches are not always computationally cheaper (e.g. *de novo* assembly). Therefore, one should not expect these methods to become a substitute for standard protocols in the short term. However, our results show that reference/alignment-free methods can readily complement standard methods. Integrating alignment-free and alignment-based methods to comprehensively answer biological and clinical questions could bring better understanding of the transcriptome.

4.5 Perspectives

During this thesis, extensive DNA and RNA sequencing data have been analyzed by our proposed alignment free protocols. Among the discoveries, transposable elements are found to be active in tumors and can be used to define patient sub-

groups. Generally, most TEs are supposed to be silent in human somatic cells due to the inhibitory effect of DNA hypermethylation. One possible mechanism of TE activation is the reduced DNA methylation that promotes TE expression. Therefore, looking into the epigenetic alterations in tumors might establish new links to causal factors for TEs activation, such variants could also be used for classification of cancer subtypes and earlier screening of cancer.

One of the most remarkable advantages of k-mer based methods is the excellent computational efficiency for querying and quantification of specific sequence segments. Novel data structures now allow to query large RNA-seq or DNA-seq datasets at the k-mer level in very fast time (Marchet et al., 2020, 2021). An area where such techniques could prove invaluable is non-invasive qualitative and quantitative testing from liquid biopsies for tumor early screening, in which case extremely large volume of sequencing data are needed to capture the pathogenic mutations of ultra low fraction.

The way DNA is packaged inside the nucleus of a cell differs between healthy and tumor cells. The nuclei of healthy cells package DNA in a well-organized manner with strict supervision and repairability. By contrast, nuclei of tumor cells are disorganized and, as a consequence, abnormal DNA is released into the blood. The released cell free DNA carries rich and potentially diagnostic enabling genetic information, which is one of the most active areas in transnational medicine.

A number of large cohort studies, conducted by academic and industrial organizations, such as Grail, are actively pursuing circulating DNA screening for early tumor detection. Using k-mer approaches to investigate the circulating DNA screening specificity of various tumor types or using tissue or subtype specific k-mer signatures to distinguish tumors of different types. In routine clinical diagnostic testing or screening, a panel of pre-selected and clinical validated cancer specific k-mers could be used as a new option for rapid analytical pipeline in liquid biopsies.

Alignment-free protocols are a powerful supplement and extension of standard mapping-based methods. From a biological point of view, integrating both strate-

gies can more comprehensively discover phenotype-related genetic information. In future works, further integration of mapping-based and mapping-free approaches could assist in the systematically and comprehensively discovery of key oncogenic mechanisms.

Résumé en français

Introduction

Historique du séquençage à haut débit (HTS)

Les séquences génétiques composées des bases adenine (A), cytosine (C), guanine (G), et thymine (T) sont le fondement de la reproduction des organismes. Le déchiffrement de ces séquences est essentiel à la compréhension du vivant. Le séquençage de l'ADN est la technologie qui permet de déterminer l'ordre exact et le type des paires de bases dans un fragment d'ADN.

Après l'achèvement du premier génome humain, le National Human Genome Research Institute a lancé une initiative de technologie de séquençage de l'ADN visant à réduire le coût du séquençage d'un génome humain à 1000 USD (Schloss, 2008). Une vague de technologies de séquençage à haut débit (HTS) a émergé, souvent appelée séquençage de nouvelle génération (NGS) ou séquençage massivement parallèle (MPS). Ces technologies peuvent séquencer des centaines de millions de molécules d'ADN en parallèle, produisant des lectures ou "reads" courts de 50 de quelques centaines de bases.

Alimentées par les développements techniques, la recherche fondamentale et la demande du marché, plusieurs générations de plates-formes NGS ont vu le jour depuis 2005. Par rapport au séquençage Sanger de première génération, le NGS génère des données massives en quelques heures à un coût considérablement réduit, devenant

ainsi le premier choix pour études génomiques et transcriptomiques à grande échelle.

Dans mon introduction, je présente les deux technologies NGS les plus utilisées: DNA-seq et RNA-seq. Le DNA-seq est utilisé pour détecter les variations génétiques, telles que les Single-Nucleotide Variants, les insertions et délétions, les variants structuraux et les variants de nombre de copies. Le séquençage d'ARN (RNA-seq) vise à déterminer le contenu en séquences d'ARN d'un échantillon à l'aide de NGS. Au cours de la dernière décennie, le RNA-seq est devenu un outil indispensable pour l'analyse du transcriptome. Le RNA-seq fournit les matériaux de base pour évaluer différents aspects du transcriptome, y compris l'expression des gènes et des transcrits, l'épissage alternatif et la découverte de nouveaux transcrits (Trapnell et al., 2012; Sultan et al., 2008; Robertson et al., 2010; Trapnell et al., 2013).

Les chaînes de traitement ou "pipelines" informatiques utilisés pour l'analyse des données NGS souffrent de plusieurs limitations. Dans les pipelines DNA-seq, les variants génomiques sont identifiés en comparant les reads alignés à une référence. Les reads avec de multiples mutations ou delétions sont difficiles à traiter car leur alignement dépend fortement des aligneurs. De plus, bien que les génomes de nombreuses espèces aient été séquencés et mis à disposition, ces génomes sont souvent incomplets et la plupart des organismes n'ont toujours pas de génome de référence disponible. Un autre inconvénient de l'utilisation d'une référence est d'ignorer la diversité au sein des populations. Des différences importantes existent entre les génomes des individus, et un génome de référence unique ne rend pas compte de cette diversité (Sherman et al., 2019). Enfin différents aligneurs produisent des alignements différents. Par conséquent, la qualité des analyses reposant sur ces alignements ne peut pas être garantie. Un autre problème avec les protocoles basés sur l'alignement est lié aux ressources informatiques : les pipelines standard sont gourmands en mémoire et en temps.

Les chercheurs sont conscients des limites du passage par un génome de référence et ont proposé des solutions. Une approche consiste à incorporer dans le génome de référence la diversité de l'ensemble de la population. Le pan-génome est défini

comme la combinaison de génomes contenant tous les variants représentatifs qui se produisent dans une espèce. La précision des alignemnts et des prédictions de variants s'améliore si les reads sont alignés sur une collection représentative de génomes (pan-génome) plutôt que sur un seul génome linéaire (Victor et al., 2018; Sherman and Salzberg, 2020).

Alternativement, une collection diversifiée de génomes peut être représentée à l'aide de graphes pan-génomiques (Paten et al., 2017; Li et al., 2020), où chaque génome individuel est identifié comme un chemin dans le graphe. Les régions polymorphes créent des bulles indiquant divers génotypes à la position correspondante dans l'ensemble de la population.

Néanmoins, les pipelines standard occupent toujours une position dominante, en particulier dans les grands projets de génomique du cancer tels que le Pan Cancer Analysis of Whole Genome (PCAWG) (The et al., 2020; Priestley et al., 2019). Ces pipelines ont permis d'identifier les changements génétiques à l'échelle du génome entier et de découvrir de nouveaux gènes causatifs ou "driver" du cancer.

Les principaux objectifs de ma thèse sont d'exploiter la puissance des approches sans alignement pour découvrir de nouvelles variations dans les transcriptomes et les génomes du cancer dans des régions difficiles à cartographier ou des régions absentes du génome de référence. Pour atteindre cet objectif, j'ai eu deux projets. Dans le projet sur l'analyse des variants génomiques, j'ai développé un logiciel sans alignement (2-kupl) pour identifier des variants à partir de deux échantillons d'ADN-seq appariés. Dans l'autre projet concernant le transcriptome, j'ai appliqué DE-kupl pour trouver des événements liés au phénotype cancéreux dans deux jeux de données indépendants de cancer du poumon, afin de décrouvrir de nouvelles caractéristiques des ARN tumoraux.

Résultats

La contribution de séquences d'ARN inexplorées à l'identité tumorale dans l'adénocarcinome pulmonaire

Sur une période de 20 ans, la transcriptomique du cancer a transformé notre compréhension de la biologie des tumeurs et a produit des outils performants pour le typage des tumeurs, le diagnostic et la prédiction des résultats (Gollub and Prowda, 1999; Parker et al., 2009; Margolin and Lindblom, 2006). Alors que l'analyse transcriptomique de première génération était basée sur des puces à ADN en mettant l'accent sur les gènes codant pour les protéines, la génération actuelle s'appuie sur les données RNA-seq, qui promettent de fournir une vue plus complète de l'expression des gènes. Cependant, malgré son potentiel de découverte de transcrits, les données de RNA-seq du cancer sont encore utilisées principalement pour quantifier l'expression de gènes annotés répertoriés dans un transcriptome de référence. Cela ignore un large éventail d'isoformes d'ARNm, d'ARN non codants, de rétroéléments endogènes et de transcrits de virus et de bactéries exogènes (Morillon and Gautheret, 2019). La quantité d'informations laissées inexploitées dans les transcriptions non canoniques reste inconnue. Un certain nombre d'études ont commencé à répondre à cette question en utilisant des données RNA-seq de cancer publiquement accessibles, en se concentrant sur des classes de transcrits spécifiques telles que les variantes d'épissage (Kahles et al., 2018; Vitting-Seerup and Sandelin, 2019), les lncRNAs (Iyer et al., 2015), les snoRNAs (Gong et al., 2017), ARN bactériens (Ouchenir et al., 2017) ou ARN viraux (Zapatka et al., 2020). D'autres sources négligées de diversité d'ARN sont les régions dites sur liste noire du génome qui sont trop variables ou répétées pour être correctement analysées par des approches conventionnelles (Amemiya et al., 2019). À notre connaissance, aucune tentative n'a été faite pour extraire et évaluer à la fois toutes ces informations d'ARN non standard à partir des données de RNA-seq tumoral. Nous pensons que cette approche pourrait être particulièrement utile dans le cancer, car chaque tumeur individuelle

abrite un transcriptome unique qui s'écarte de celui des tissus normaux de plusieurs manières imprévisibles.

Récemment, nous avons introduit une méthode de calcul, DE-kupl (Audoux et al., 2017b), qui effectue une analyse différentielle des données RNA-seq au niveau k-mer. Comme cette méthode est sans référence et sans alignement, elle identifie tout nouvel ARN ou isoforme d'ARN présent dans les données à la résolution nucléotidique, y compris les transcrits mal cartographiés tels que les ARN de répétitions et l'ARN chimérique. Les résultats ont révélé une collection de nouveaux lincARN non annotés spécifiques à une tumeur, des rétentions d'intron et des événements d'épissage. Plus frappant encore, une collection de rétroéléments endogènes (ERE) forme une classe majeure de transcrits définissant la tumeur. Nous avons également identifié un sous-ensemble d'événements sans expression dans les tissus normaux qui pourraient être des néoantigènes candidats. Enfin, nous avons identifié un ensemble de variants de transcription potentiellement liés à la survie. Nous aimerions suggérer DE-kupl comme une approche prometteuse et complète pour le profilage des transcrits du cancer.

Nous avons pu montrer qu'une classe de variation déterminante est formée des répétitions endogènes. L'expression des répétitions L1 et Alu définit deux sous-groupes tumoraux majeurs. Le sous-groupe avec une expression L1/Alu plus élevée était associé à des mutations plus fréquentes dans P53, à une charge de mutation et de nombre de copies plus élevée et à un infiltrat de cellules immunitaires réduit. Ceci est cohérent avec la découverte précédente impliquant P53 dans le contrôle de la rétrotransposition (Jung et al., 2018) et corrélant la rétrotransposition L1 avec un environnement immunitaire réprimé (Jung et al., 2018; Zhang et al., 2020b). La mobilité des TE peut également conduire à une instabilité du génome. Les intégrations de TE aléatoires entraînent une mutagenèse insertionnelle et des variations structurelles génomiques, y compris les CNV (Lee et al., 2012).

Outre leur capacité à stratifier les patients, les répétitions exprimées démontrent un pouvoir pronostique important. Des signatures multivariées composées d'expression

de HERV et L1, ou d'une simple expression répétée séparent les patients en groupes de survie clairs. L'expression de HERV a été sporadiquement impliquée dans divers types de cancer (Bannert et al., 2018), et a récemment été associée à un mauvais pronostic dans le cancer colorectal (Golkaram et al., 2021).

Un autre domaine où les approches sans référence ont un potentiel de découverte élevé est la détection de néo-antigènes pour le développement de thérapies antitumorales et pour l'orientation des patients en immunothérapie. Nous avons trouvé des sources potentielles de néoantigènes partagés dans les répétitions, les ARNnc et les variantes d'épissage des ARNm. Des néoantigènes spécifiques de la tumeur ont déjà été identifiés à partir de répétitions et de régions supposées non codantes à l'aide de stratégies basées sur la cartographie (Smith et al., 2019; Laumont et al., 2018). Cependant, nous pensons que notre approche a plus de potentiel car elle collecte tous les événements indépendamment de leur origine, y compris des régions non cartographiables ou profondément réarrangées. Par conséquent, nous avons une meilleure chance de découvrir des sources de néoépitopes quelle que soit leur origine.

L'analyse sans référence présente d'autres avantages. Premièrement, il s'agit par essence d'une méthode intégrative car elle combine la variation génomique et transcriptomique en une seule matrice d'expression qui peut être analysée de plusieurs manières. Une application intéressante de telles matrices est la construction de modèles prédictifs intégrant plusieurs classes d'événements. Nous (Nguyen et al., 2021) et d'autres (Lorenzi et al., 2020) avons initié ce type d'approche avec des résultats très prometteurs. Deuxièmement, les méthodes sans référence pourraient être particulièrement intéressantes dans les projets de méta-transcriptomique où les ARN sont capturés dans un environnement contenant des espèces bactériennes, archéennes ou eucaryotes inconnues. Notre protocole garantit que tout ARN spécifique à un sous-ensemble d'échantillons sera capturé indépendamment de son origine.

2-kupl: détection de variants sans cartographie à partir des données DNA-seq des échantillons appariés

La recherche de variants génomiques est un aspect fondamental de la recherche médicale, que ce soit dans l'étude des maladies mendéliennes ou des altérations somatiques liées au cancer (Li et al., 2017). Alors que certaines variations entraînent un dysfonctionnement génétique et une maladie (MacArthur et al., 2014), d'autres sont en grande partie asymptomatiques mais donnent lieu à des néoantigènes pertinents pour l'échappement immunitaire et l'efficacité thérapeutique ou le traitement (Jiang et al., 2019). Les variations du génome présentent également un intérêt en microbiologie pour analyser les différences entre les souches microbiennes (Shiloach et al., 2010) et révéler les mécanismes sous-jacents aux phénotypes. Dans cette étude, nous abordons le problème de trouver des différences génomiques entre une paire de jeux de séquences DNA-seq à haut débit provenant du même individu (variation somatique humaine) ou de deux souches bactériennes.

Les variations génomiques comprennent les mutations, les indels et les variations structurelles (SV). Les mutations et les indels peuvent altérer les gènes en perturbant le code génétique, tandis que les SV, en rapprochant des régions distantes ou en divisant une région en segments, peuvent créer des gènes chimériques ou avoir un impact plus large sur des régions chromosomiques entières (Hurles et al., 2008). Les variants sont généralement détectés par séquençage du génome entier (WGS) ou de l'exome entier (WES) par comparaison avec des séquences de référence. Des aligneurs tels que BWA (Li and Durbin, 2009) sont d'abord appliqués pour aligner les reads aux séquences de référence. L'étape de détection de variant identifie alors les différences entre les reads alignés et la référence. Les outls de détection de variants les plus utilisés sont MuTect2 (Benjamin et al., 2019), VarScan (Koboldt et al., 2012), somaticsniper (Larson et al., 2012) et MuSE (Fan et al., 2016). Sur la base des variations observées entre deux échantillons de séquence et un génome de référence commun, ces programmes peuvent déduire des différences entre les deux échantillons (par exemple, avec le mode somatique de MuTect2).

L'La détection de variants basé sur des références a des limitations bien connues. Les aligneurs peuvent rencontrer des difficultés lors de la gestion des reads avec de faibles qualités d'alignement (Li et al., 2008), provenant de régions répétées, de régions de faible complexité ou de variantes complexes. Ces reads de faible qualité d'alignement sont généralement ignorés. De plus, certaines espèces n'ont pas de référence fiable, ce qui est courant chez les microbes (Loeffler et al., 2020).

Nous présentons 2-kupl, un pipeline bioinformatique basé sur les k-mer qui compare des échantillons cas / contrôle appariés pour découvrir des variations spécifiques au cas. 2-kupl identifie les fragments de séquence (contigs) spécifiques à un jeu de données (cas) et absents du jeu de données de contrôle. Cette opération se fait sans s'appuyer sur un génome de référence. Nous comparons la précision et les besoins CPU de 2-kupl avec celles d'autres logiciels de détection de variantes en utilisant des jeux de données DNA-seq simulés et réels. Nous analysons la nature des nouvelles variations détectées par 2-kupl et les raisons potentielles de leur absence dans les protocoles conventionnels. Nous utilisons également 2-kupl pour détecter des variations récurrentes dans les données de séqeune d'exome d'adénocarcinome de la prostate (PRAD) du projet TCGA (Tomczak et al., 2015). Enfin, nous évaluons la précision de 2-kupl dans les données bactériennes WGS. Dans l'ensemble, nous démontrons que 2-kupl est une alternative pratique et puissante pour la découverte de variations génomiques dans des régions difficiles à cartographier ou des espèces sans référence fiable.

En conclusion, l'identification de différents types de variants, à l'aide de données DNA-seq reste un défi. Les principaux protocoles développés pour le DNA-seq s'appuient fortement sur la référence. En général, les méthodes qui alignent les données de séquençage sur une référence fonctionnent mieux que les méthodes sans alignement. Cependant, 2-kupl peut capturer des événements tombant dans les régions difficiles à mapper et peut fonctionner mieux que d'autres protocoles sans alignement. 2-kupl est l'outil le plus rapide dans notre comparaison avec d'autres méthodes, car il économise la procédure d'alignement. La forte concordance entre les variations de haut score prédites par 2-kupl et des prédictions obtenues sur le portail

officiel GDC indique la capacité d'utiliser 2-kupl comme extension et complément des méthodes conventionnelle. De nouvelles ou des variations récurrentes et de nouveaux gènes pertinents pour le cancer de la prostate sont capturés par 2-kupl.

Discussion

L'avènement de la médecine de précision avec ses options thérapeutiques ciblées, et celui de l'immunothérapie révolutionnent actuellement le traitement en oncologie, notamment dans le cancer du poumon.

La classification précise des sous-types tumoraux est une étape essentielle pour le traitement personnalisé du cancer. De nombreuses études ont subdivisé les patients LUAD en sous-groupes en fonction de l'expression des gènes, des profils de mutation et des signatures immunitaires (Hu et al., 2019; Ding et al., 2020; Xu et al., 2020). À notre connaissance, aucune étude n'avait jamais utilisé des données provenant de régions répétitives pour définir des sous-groupes LUAD. En effet, la variation complexe des régions répétitives est un indicateur direct de l'instabilité du génome. Par exemple, les répétitions Alu sont associées à de nombreuses instabilités microsatellites, qui jouent un rôle essentiel dans l'oncogenèse (Arcot et al., 1995). Les répétitions ont également un impact sur les génomes en induisant de petites variations, des événements de recombinaison, des conversions géniques et une expression génique anormale (Batzer and Deininger, 2002). Par conséquent, les patients présentant des caractéristiques de répétition différentes peuvent présenter des réponses immunitaires ou des pronostics divers.

Dans le chapitre 3.1.4.6, nous avons appliqué DE-kupl pour identifier des éléments d'ARN exprimés de manière différentielle dans deux cohortes LUAD indépendantes. Nous avons identifié un grand nombre d'éléments d'ARN différentiels qui étaient récurrents dans les deux jeux de données, dont des SNV, des événements introniques, des événements d'épissage, des répétitions, etc. Nous avons classé les patients en deux sous-groupes sur la base des répétitions Alu et L1P : les deux principaux

types de répétitions dans l'ensemble des répétitions avec expression différentielle. Malgré l'absence de différences significatives dans la survie, le sexe, l'âge ou le stade de la tumeur, nous avons observé des différences dans le contenu en cellules immunitaires entre les deux sous-groupes. Les patients du sous-groupe "à répétition élevée" présentaient des niveaux d'infiltration immunitaire plus faibles, ce qui est cohérent avec les observations précédentes selon lesquelles un environnement immunodéprimé favorise l'activation des éléments transposables (Zhang et al., 2020b; Jung et al., 2018). Le groupe "à répétition élevée" présentait également une charge de mutation et de CNV plus élevée, ce qui suggère que l'activation répétée peut être une cause d'instabilité génomique dans LUAD. Ceci est également cohérent avec les conclusions précédentes selon lesquelles les répétitions peuvent provoquer des variations aux niveaux du transcriptome et du génome (Lanciano and Cristofari, 2020; Payer and Burns, 2019). Généralement, les patients avec une charge mutationnelle (TMB) élevée sont plus aptes à bénéficier d'une immunothérapie. Cependant, nous avons constaté que le sous-groupe de patients LUAD avec une TMB élevée présente un niveau d'infiltration immunitaire plus faible. Cela implique la nécessité d'une nouvelle stratification en particulier chez les patients à TMB haute, pour prédire leur sensibilité à l'immunothérapie.

La découverte d'antigènes spécifiques de tumeurs, ou néo-antigènes, est une étape clé vers le développement de vaccins contre le cancer. Les stratégies canoniques de criblage des néo-antigènes candidats consistent en plusieurs grandes étapes qui reposent sur les données DNA-seq et RNA-seq (Gopanenko et al., 2020). Les premières étapes de la prédiction des néoantigènes reposent sur un alignement sur le génome de référence. Cependant, ces protocoles classiques négligent une grande diversité de néoantigènes possibles.

Dans le chapitre 3.1.4.7, nous avons identifié des ARN candidats à la production de néoantigènes dans les exons, les introns, les lincARN et les répétitions. À notre connaissance, il n'y existait pas auparavant de travaux évaluant aussi systématiquement que le notre le potentiel des régions introniques, intergéniques et répétées à la production de néoantigènes.

Il existe un grand nombre de régions difficiles à analyser dans le génome humain et qui ne sont par conséquent pas couvertes par les stratégies NGS standard alors qu'elle peuvent contenir des variations cliniquement importantes. Ces régions comprennent notamment des pseudogènes multicopies, télomères, centromère ou d'autres régions génomiques répétitives. Aborder ces régions difficiles à cartographier est un problème difficile à surmonter avec les protocoles standard. Ce travail de thèse a trouvé des éléments d'ARN récurrents exprimés à partir de ces régions. De plus, ces ARN sont transcrits dans de multiples échantillons de tumeurs et faiblement exprimés voire silencieux dans tous les tissus normaux. La survenue de ces événements chez plusieurs patients de populations indépendantes indique la présence de mécanismes oncogènes communs non identifiés.

Dans le chapitre 3.2, nous avons identifié des variations somatiques récurrentes chez les patients atteints d'adénocarcinome de la prostate avec notre nouveau logiciel 2-kupl. 2-kupl atteint une concordance élevée avec les logiciels de détection de variants standard basés sur l'alignement, tels que GATK-Mutect.

Notre programme 2-kupl détecte de nombreuses variations somatiques dans les régions difficiles à cartographier, dont la plupart ne sont pas détectables par GATK-Mutect. Par exemple, 2-kupl identifie de nouvelles mutations dans certains gènes de cancer, suggérant que leurs taux de mutation pourrait être sous-estimé par les méthodes standard. De plus, 2-kupl détecte des variations structurales. Avec les méthodes classiques de détection de variants, les reads couvrant les jonctions de ces variations ne sont pas correctment alignés ou ont plusieurs alignements possibles sur le génome et sont donc rejetés par les logiciels de détection de variants. Les logiciels spécialisés dans la détection de variants structuraux manquent également les variants lorsque les reads de support ne sont pas correctement alignés. Pourtant, les variations structurales récurrentes cachés dans ces régions difficiles à analyser sont potentiellement liées à la maladie (Levy-Sakin et al., 2019).

Il reste encore beaucoup de régions mal explorées dans le génome humain au-delà des régions annotées dans les bases de référence. Des approches sans alignement ont

ici été développées pour éclairer l'information génétique présente dans ces régions. Cependant, leur précision reste généralement inférieure à celle des protocoles standards. De plus, dans l'offre générale de logiciels de bioinformatique, les approches sans alignement sont souvent perrçues comme coûteuses en calcul (par exemple, l'assemblage *de novo* est très couteux). Il ne faut donc pas s'attendre à ce que ces méthodes se substituent à court terme aux protocoles standards. Cependant, nos résultats montrent que les méthodes sans référence/alignement peuvent facilement compléter les méthodes standard. L'intégration de méthodes sans alignement et basées sur l'alignement pour répondre de manière exhaustive aux questions biologiques et cliniques pourrait à terme apporter une meilleure compréhension du génome et du transcriptome.

Bibliography

F. Abbas-Aghababazadeh, Q. Li, and B. L. Fridley. Comparison of normalization approaches for gene expression studies completed with high-throughput sequencing. PloS one, 13(10):e0206312, 2018.

A. Abeshouse, J. Ahn, R. Akbani, A. Ally, S. Amin, C. D. Andry, M. Annala, A. Aprikian, J. Armenia, A. Arora, et al. The molecular taxonomy of primary prostate cancer. Cell, 163(4):1011–1025, 2015.

M. I. Abouelhoda, S. Kurtz, and E. Ohlebusch. Replacing suffix trees with enhanced suffix arrays. Journal of discrete algorithms, 2(1):53–86, 2004.

I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, and S. R. Sunyaev. A method and server for predicting damaging missense mutations. Nature methods, 7(4):248–249, 2010.

A. Aithal, S. Rauth, P. Kshirsagar, A. Shah, I. Lakshmanan, W. M. Junker, M. Jain, M. P. Ponnusamy, and S. K. Batra. Muc16 as a novel target for cancer therapy. Expert opinion on therapeutic targets, 22(8):675–686, 2018.

L. B. Alexandrov, P. H. Jones, D. C. Wedge, J. E. Sale, P. J. Campbell, S. Nik-Zainal, and M. R. Stratton. Clock-like mutational processes in human somatic cells. Nature genetics, 47(12):1402, 2015.

H. M. Amemiya, A. Kundaje, and A. P. Boyle. The encode blacklist: identification of problematic regions of the genome. Scientific reports, 9(1):1–5, 2019.

S. Anders, P. T. Pyl, and W. Huber. Htseq—a python framework to work with high-throughput sequencing data. Bioinformatics, 31(2):166–169, 2015.

S. S. Arcot, Z. Wang, J. L. Weber, P. L. Deininger, and M. A. Batzer. Alu repeats: a source for the genesis of primate microsatellites. Genomics, 29(1):136–144, 1995.

J. Armenia, S. A. Wankowicz, D. Liu, J. Gao, R. Kundra, E. Reznik, W. K. Chatila, D. Chakravarty, G. C. Han, I. Coleman, et al. The long tail of oncogenic drivers in prostate cancer. Nature genetics, 50(5):645–651, 2018.

H. Ashrafi, T. Hill, K. Stoffel, A. Kozik, J. Yao, S. R. Chin-Wo, and A. Van Deynze. De novo assembly of the pepper transcriptome (capsicum annuum): a benchmark for in silico discovery of snps, ssrs and candidate genes. BMC genomics, 13(1): 1–15, 2012.

C. H. Au, D. N. Ho, A. Kwong, T. L. Chan, and E. S. Ma. Bamclipper: removing primers from alignments to minimize false-negative mutations in amplicon next-generation sequencing. Scientific reports, 7(1):1–7, 2017.

P. Audano and F. Vannberg. Kanalyze: a fast versatile pipelined k-mer toolkit. Bioinformatics, 30(14):2070–2072, 2014.

P. A. Audano, S. Ravishankar, and F. O. Vannberg. Mapping-free variant calling using haplotype reconstruction from k-mer frequencies. Bioinformatics, 34(10): 1659–1665, 2018.

J. Audoux, N. Philippe, R. Chikhi, M. Salson, M. Gallopin, M. Gabriel, J. Le Coz, T. Commes, and D. Gautheret. Exhaustive capture of biological variation in rna-seq data through k-mer decomposition. BioRxiv, page 122937, 2017a.

J. Audoux, N. Philippe, R. Chikhi, M. Salson, M. Gallopin, M. Gabriel, J. Le Coz, E. Drouineau, T. Commes, and D. Gautheret. De-kupl: exhaustive capture of biological variation in rna-seq data through k-mer decomposition. Genome biology, 18(1):1–15, 2017b.

P. L. Auer, A. P. Reiner, G. Wang, H. M. Kang, G. R. Abecasis, D. Altshuler, M. J. Bamshad, D. A. Nickerson, R. P. Tracy, S. S. Rich, et al. Guidelines for large-scale sequence-based complex trait association studies: lessons learned from

the nhlbi exome sequencing project. The American Journal of Human Genetics, 99(4):791–801, 2016.

A. Auguste and A. Leary. Abnormalities of dna repair and gynecological cancers. Bulletin du cancer, 104(11):971–980, 2017.

S. Ballouz, A. Dobin, and J. A. Gillis. Is it time to change the reference genome? Genome biology, 20(1):1–9, 2019.

S. Bamford, E. Dawson, S. Forbes, J. Clements, R. Pettett, A. Dogan, A. Flanagan, J. Teague, P. A. Futreal, M. R. Stratton, et al. The cosmic (catalogue of somatic mutations in cancer) database and website. British journal of cancer, 91(2): 355–358, 2004.

A. Bankevich, S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, S. I. Nikolenko, S. Pham, A. D. Prjibelski, et al. Spades: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of computational biology, 19(5):455–477, 2012.

N. Bannert, H. Hofmann, A. Block, and O. Hohn. Hervs new role in cancer: from accused perpetrators to cheerful protectors. Frontiers in microbiology, 9:178, 2018.

C. E. Barbieri, S. C. Baca, M. S. Lawrence, F. Demichelis, M. Blattner, J.-P. Theurillat, T. A. White, P. Stojanov, E. Van Allen, N. Stransky, et al. Exome sequencing identifies recurrent spop, foxa1 and med12 mutations in prostate cancer. Nature genetics, 44(6):685–689, 2012.

M. A. Batzer and P. L. Deininger. Alu repeats and human genomic diversity. Nature reviews genetics, 3(5):370–379, 2002.

R. Bedre. Gene expression units explained: Rpm, rpkm, fpkm, tpm, deseq, tmm, scnorm, getmm, and combat-seq. https://www.reneshbedre.com/blog/expression$_u nits.html$.

T. Beller and E. Ohlebusch. A representation of a compressed de bruijn graph for pan-genome analysis that enables search. Algorithms for Molecular Biology, 11 (1):1–17, 2016.

D. Benjamin, T. Sato, K. Cibulskis, G. Getz, C. Stewart, and L. Lichtenstein. Calling somatic snvs and indels with mutect2. BioRxiv, page 861054, 2019.

M. Blaxter. Revealing the dark matter of the genome. Science, 330(6012):1758–1759, 2010.

M. Bogdan, J. K. Ghosh, S. T. Tokdar, et al. A comparison of the benjamini-hochberg procedure with some bayesian rules for multiple testing. In Beyond parametrics in interdisciplinary research: Festschrift in honor of Professor Pranab K. Sen, pages 211–230. Institute of Mathematical Statistics, 2008.

A. M. Bolger, M. Lohse, and B. Usadel. Trimmomatic: a flexible trimmer for illumina sequence data. Bioinformatics, 30(15):2114–2120, 2014.

G. Bollag, P. Hirth, J. Tsai, J. Zhang, P. N. Ibrahim, H. Cho, W. Spevak, C. Zhang, Y. Zhang, G. Habets, et al. Clinical efficacy of a raf inhibitor needs broad target blockade in braf-mutant melanoma. Nature, 467(7315):596–599, 2010.

R. Bourgon, R. Gentleman, and W. Huber. Independent filtering increases detection power for high-throughput experiments. Proceedings of the National Academy of Sciences, 107(21):9546–9551, 2010.

D. Y. Brandt, V. R. Aguiar, B. D. Bitarello, K. Nunes, J. Goudet, and D. Meyer. Mapping bias overestimates reference allele frequencies at the hla genes in the 1000 genomes project phase i data. G3: Genes, Genomes, Genetics, 5(5):931–941, 2015.

N. L. Bray, H. Pimentel, P. Melsted, and L. Pachter. Near-optimal probabilistic rna-seq quantification. Nature biotechnology, 34(5):525–527, 2016.

R. Broekema, O. Bakker, and I. Jonkers. A practical view of fine-mapping and gene prioritization in the post-genome-wide association era. Open biology, 10(1): 190221, 2020.

B. Bushnell. Bbmap. https://sourceforge. net/projects/bbmap, 2018.

Z. Chang, G. Li, J. Liu, Y. Zhang, C. Ashby, D. Liu, C. L. Cramer, and X. Huang. Bridger: a new framework for de novo transcriptome assembly using rna-seq data. Genome biology, 16(1):1–10, 2015.

C. Chen, S. S. Khaleel, H. Huang, and C. H. Wu. Software for pre-processing illumina next-generation sequencing short read sequences. Source code for biology and medicine, 9(1):1–11, 2014.

Y. Chen, T. Souaiaia, and T. Chen. Perm: efficient mapping of short sequencing reads with periodic full sensitive spaced seeds. Bioinformatics, 25(19):2514–2521, 2009.

K. R. Chi. The year of sequencing, 2008.

R. Chikhi, A. Limasset, and P. Medvedev. Compacting de bruijn graphs from sequencing data quickly and in low memory. Bioinformatics, 32(12):i201–i208, 2016.

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, 31(3):213–219, 2013.

P. Cingolani, A. Platts, L. L. Wang, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Lu, and D. M. Ruden. A program for annotating and predicting the effects of single nucleotide polymorphisms, snpeff: Snps in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. Fly, 6(2):80–92, 2012.

E. Cline, N. Wisittipanit, T. Boongoen, E. Chukeatirote, D. Struss, and A. Eungwanichayapant. Recalibration of mapping quality scores in illumina short-read alignments improves snp detection results in low-coverage sequencing data. PeerJ, 8:e10501, 2020.

M. Cmero, B. Schmidt, I. J. Majewski, P. G. Ekert, A. Oshlack, and N. M. Davidson. Mintie: identifying novel structural and splice variants in transcriptomes using rna-seq data. bioRxiv, 2020.

P. J. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. Bioinformatics, 25 (11):1422–1423, 2009.

P. J. Cock, C. J. Fields, N. Goto, M. L. Heuer, and P. M. Rice. The sanger fastq file format for sequences with quality scores, and the solexa/illumina fastq variants. Nucleic acids research, 38(6):1767–1771, 2010.

P. E. Compeau, P. A. Pevzner, and G. Tesler. How to apply de bruijn graphs to genome assembly. Nature biotechnology, 29(11):987–991, 2011.

A. Conesa, P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M. W. Szcześniak, D. J. Gaffney, L. L. Elo, X. Zhang, et al. A survey of best practices for rna-seq data analysis. Genome biology, 17(1):1–19, 2016.

G. O. Consortium et al. Creating the gene ontology resource: design and implementation. Genome research, 11(8):1425–1433, 2001.

T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. Introduction to algorithms. MIT press, 2009.

A. Criscuolo. A fast alignment-free bioinformatics procedure to infer accurate distance-based phylogenetic trees from genome assemblies. Research Ideas and Outcomes, 5:e36178, 2019.

P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, et al. The variant call format and vcftools. Bioinformatics, 27(15):2156–2158, 2011.

A. J. de Koning, W. Gu, T. A. Castoe, M. A. Batzer, and D. D. Pollock. Repetitive elements may comprise over two-thirds of the human genome. PLoS Genet, 7 (12):e1002384, 2011.

K. De Paepe. Comparison of methods for differential gene expression using rna-seq data. 2015.

J. F. Degner, J. C. Marioni, A. A. Pai, J. K. Pickrell, E. Nkadori, Y. Gilad, and J. K. Pritchard. Effect of read-mapping biases on detecting allele-specific expression from rna-sequencing data. Bioinformatics, 25(24):3207–3212, 2009.

S. Deorowicz, M. Kokot, S. Grabowski, and A. Debudaj-Grabysz. Kmc 2: fast and resource-frugal k-mer counting. Bioinformatics, 31(10):1569–1576, 2015.

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, 43(5):491, 2011.

Y. Ding, L. Zhang, L. Guo, C. Wu, J. Zhou, Y. Zhou, J. Ma, X. Li, P. Ji, M. Wang, et al. Comparative study on the mutational profile of adenocarcinoma and squamous cell carcinoma predominant histologic subtypes in chinese non-small cell lung cancer patients. Thoracic cancer, 11(1):103–112, 2020.

A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras. Star: ultrafast universal rna-seq aligner. Bioinformatics, 29(1):15–21, 2013.

M. G. Dozmorov, I. Adrianto, C. B. Giles, E. Glass, S. B. Glenn, C. Montgomery, K. L. Sivils, L. E. Olson, T. Iwayama, W. M. Freeman, et al. Detrimental effects of duplicate reads and low complexity regions on rna-and chip-seq data. In BMC bioinformatics, volume 16, pages 1–11. BioMed Central, 2015.

S. Dudoit and M. J. Van Der Laan. Multiple testing procedures with applications to genomics. Springer Science & Business Media, 2007.

S. Duffy, L. A. Shackelton, and E. C. Holmes. Rates of evolutionary change in viruses: patterns and determinants. Nature Reviews Genetics, 9(4):267–276, 2008.

D. Earl, N. Nguyen, G. Hickey, R. S. Harris, S. Fitzgerald, K. Beal, I. Seledtsov, V. Molodtsov, B. J. Raney, H. Clawson, et al. Alignathon: a competitive assessment of whole-genome alignment methods. Genome research, 24(12):2077–2089, 2014.

C. Everaert, M. Luypaert, J. L. Maag, Q. X. Cheng, M. E. Dinger, J. Hellemans, and P. Mestdagh. Benchmarking of rna-sequencing analysis workflows using whole-transcriptome rt-qpcr expression data. Scientific reports, 7(1):1–11, 2017.

A. D. Ewing, K. E. Houlahan, Y. Hu, K. Ellrott, C. Caloian, T. N. Yamaguchi, J. C. Bare, C. P'ng, D. Waggott, V. Y. Sabelnykova, et al. Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. Nature methods, 12(7):623–630, 2015.

A. Fabregat, S. Jupe, L. Matthews, K. Sidiropoulos, M. Gillespie, P. Garapati, R. Haw, B. Jassal, F. Korninger, B. May, et al. The reactome pathway knowledgebase. Nucleic acids research, 46(D1):D649–D655, 2018.

H. Fan, A. R. Ives, Y. Surget-Groba, and C. H. Cannon. An assembly and alignment-free method of phylogeny reconstruction from next-generation sequencing data. BMC genomics, 16(1):1–18, 2015.

Y. Fan, L. Xi, D. S. Hughes, J. Zhang, J. Zhang, P. A. Futreal, D. A. Wheeler, and W. Wang. Muse: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. Genome biology, 17(1):1–11, 2016.

S. Forbes, J. Clements, E. Dawson, S. Bamford, T. Webb, A. Dogan, A. Flanagan, J. Teague, R. Wooster, P. Futreal, et al. Cosmic 2005. British journal of cancer, 94(2):318–322, 2006.

M. Fraser, V. Y. Sabelnykova, T. N. Yamaguchi, L. E. Heisler, J. Livingstone, V. Huang, Y.-J. Shiah, F. Yousif, X. Lin, A. P. Masella, et al. Genomic hallmarks of localized, non-indolent prostate cancer. Nature, 541(7637):359–364, 2017.

A. C. Frazee, G. Pertea, A. E. Jaffe, B. Langmead, S. L. Salzberg, and J. T. Leek. Ballgown bridges the gap between transcriptome assembly and expression analysis. Nature biotechnology, 33(3):243–246, 2015.

J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. Journal of statistical software, 33(1):1, 2010.

S. N. Gardner, T. Slezak, and B. G. Hall. ksnp3. 0: Snp detection and phylogenetic analysis of genomes without genome alignment or reference genome. Bioinformatics, 31(17):2877–2878, 2015.

M. Gedil, M. Ferguson, G. Girma, A. Gisel, L. Stavolone, and I. Rabbi. Perspectives on the application of next-generation sequencing to the improvement of africa's staple food crops. In Next Generation Sequencing–Advances, Applications and Challenges, pages 287–321. InTechOpen, 2016.

C. Gilissen, A. Hoischen, H. G. Brunner, and J. A. Veltman. Disease gene identification strategies for exome sequencing. European Journal of Human Genetics, 20(5):490–497, 2012.

M. Golkaram, M. L. Salmans, S. Kaplan, R. Vijayaraghavan, M. Martins, N. Khan, C. Garbutt, A. Wise, J. Yao, S. Casimiro, et al. Hervs establish a distinct molecular subtype in stage ii/iii colorectal cancer with poor outcome. NPJ genomic medicine, 6(1):1–11, 2021.

M. J. Gollub and J. C. Prowda. Primary melanoma of the esophagus: radiologic and clinical findings in six patients. Radiology, 213(1):97–100, 1999.

T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. science, 286(5439):531–537, 1999.

A. Goncearenco, S. L. Rager, M. Li, Q.-X. Sang, I. B. Rogozin, and A. R.

Panchenko. Exploring background mutational processes to decipher cancer genetic heterogeneity. Nucleic acids research, 45(W1):W514–W522, 2017.

J. Gong, Y. Li, C.-j. Liu, Y. Xiang, C. Li, Y. Ye, Z. Zhang, D. H. Hawke, P. K. Park, L. Diao, et al. A pan-cancer analysis of the expression and clinical relevance of small nucleolar rnas in human cancer. Cell reports, 21(7):1968–1981, 2017.

A. V. Gopanenko, E. N. Kosobokova, and V. S. Kosorukov. Main strategies for the identification of neoantigens. Cancers, 12(10):2879, 2020.

M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, et al. Trinity: reconstructing a full-length transcriptome without a genome from rna-seq data. Nature biotechnology, 29(7):644, 2011.

C. Greenman, P. Stephens, R. Smith, G. L. Dalgliesh, C. Hunter, G. Bignell, H. Davies, J. Teague, A. Butler, C. Stevens, et al. Patterns of somatic mutation in human cancer genomes. Nature, 446(7132):153–158, 2007.

Z. Gu, R. Eils, and M. Schlesner. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics, 32(18):2847–2849, 2016.

B. J. Haas, A. Dobin, B. Li, N. Stransky, N. Pochet, and A. Regev. Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. Genome biology, 20(1):1–16, 2019.

A. Hagberg, P. Swart, and D. S Chult. Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick. Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders. Nucleic acids research, 33(suppl_1):D514–D517, 2005.

K. D. Hansen, S. E. Brenner, and S. Dudoit. Biases in illumina transcriptome sequencing caused by random hexamer priming. Nucleic acids research, 38(12): e131–e131, 2010.

H. Hesse and R. Höfgen. Application of genomics in agriculture. In Molecular analysis of plant adaptation to the environment, pages 61–79. Springer, 2001.

M. Heydari, G. Miclotte, Y. Van de Peer, and J. Fostier. Illumina error correction near highly repetitive dna regions improves de novo genome assembly. BMC bioinformatics, 20(1):1–13, 2019.

F. Hu, Y. Zhou, Q. Wang, Z. Yang, Y. Shi, and Q. Chi. Gene expression classification of lung adenocarcinoma into molecular subtypes. IEEE/ACM transactions on computational biology and bioinformatics, 17(4):1187–1197, 2019.

S. Huang, J. Zhang, R. Li, W. Zhang, Z. He, T.-W. Lam, Z. Peng, and S.-M. Yiu. Soapsplice: genome-wide ab initio detection of splice junctions from rna-seq data. Frontiers in genetics, 2:46, 2011.

X. Huang and A. Madan. Cap3: A dna sequence assembly program. Genome research, 9(9):868–877, 1999.

R. Hubley, R. D. Finn, J. Clements, S. R. Eddy, T. A. Jones, W. Bao, A. F. Smit, and T. J. Wheeler. The dfam database of repetitive dna families. Nucleic acids research, 44(D1):D81–D89, 2016.

M. E. Hurles, E. T. Dermitzakis, and C. Tyler-Smith. The functional impact of structural variation in humans. Trends in Genetics, 24(5):238–245, 2008.

S. Hwang, E. Kim, I. Lee, and E. M. Marcotte. Systematic comparison of variant calling pipelines using gold standard personal exome variants. Scientific reports, 5(1):1–8, 2015.

M. Imielinski, G. Guo, and M. Meyerson. Insertions and deletions target lineage-defining genes in human cancers. Cell, 168(3):460–472, 2017.

J. P. Ioannidis. Microarrays and molecular research: noise discovery? Lancet (London, England), 365(9458):454–455, 2005.

Z. Iqbal, M. Caccamo, I. Turner, P. Flicek, and G. McVean. De novo assembly and genotyping of variants using colored de bruijn graphs. Nature genetics, 44(2): 226–232, 2012.

C. A. Ishak and D. D. De Carvalho. Reactivation of endogenous retroelements in cancer development and therapy. Annual Review of Cancer Biology, 4:159–176, 2020.

M. K. Iyer, Y. S. Niknafs, R. Malik, U. Singhal, A. Sahu, Y. Hosono, T. R. Barrette, J. R. Prensner, J. R. Evans, S. Zhao, et al. The landscape of long noncoding rnas in the human transcriptome. Nature genetics, 47(3):199–208, 2015.

A. L. Jackson and L. A. Loeb. The mutation rate and cancer. Genetics, 148(4): 1483–1490, 1998.

M. Jackson, L. Marks, G. H. May, and J. B. Wilson. The genetic basis of disease. Essays in biochemistry, 62(5):643–723, 2018.

J. M. Janda and S. L. Abbott. 16s rrna gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. Journal of clinical microbiology, 45(9):2761–2764, 2007.

M. Jiang, C. Zhao, Z. Mo, and J. Wen. An improved algorithm based on bloom filter and its application in bar code recognition and processing. EURASIP Journal on Image and Video Processing, 2018(1):1–12, 2018.

T. Jiang, T. Shi, H. Zhang, J. Hu, Y. Song, J. Wei, S. Ren, and C. Zhou. Tumor neoantigens: from basic research to clinical applications. Journal of hematology & oncology, 12(1):1–13, 2019.

H. Jung, J. K. Choi, and E. A. Lee. Immune signatures correlate with l1 retrotransposition in gastrointestinal cancers. Genome research, 28(8):1136–1146, 2018.

A. Kahles, K.-V. Lehmann, N. C. Toussaint, M. Hüser, S. G. Stark, T. Sachsenberg, O. Stegle, O. Kohlbacher, C. Sander, S. J. Caesar-Johnson, et al. Comprehensive analysis of alternative splicing across tumors from 8,705 patients. Cancer cell, 34 (2):211–224, 2018.

M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, and M. Hattori. The kegg resource for deciphering the genome. Nucleic acids research, 32(suppl_1):D277–D280, 2004.

M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima. Kegg: new perspectives on genomes, pathways, diseases and drugs. Nucleic acids research, 45(D1):D353–D361, 2017.

E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. Journal of the American statistical association, 53(282):457–481, 1958.

L. Kaplinski, M. Lepamets, and M. Remm. Genometester4: a toolkit for performing basic set operations-union, intersection and complement on k-mer lists. Gigascience, 4(1):s13742–015, 2015.

K. Karczewski and L. Francioli. The genome aggregation database (gnomad). MacArthur Lab, 2017.

K. J. Karczewski, B. Weisburd, B. Thomas, M. Solomonson, D. M. Ruderfer, D. Kavanagh, T. Hamamsy, M. Lek, K. E. Samocha, B. B. Cummings, et al. The exac browser: displaying reference data information from over 60 000 exomes. Nucleic acids research, 45(D1):D840–D845, 2017.

A. Kassambara, M. Kosinski, P. Biecek, and S. Fabian. Package 'survminer'. Drawing Survival Curves using 'ggplot2'.(R package version 0.3. 1.), 2017.

D. R. Kelley, M. C. Schatz, and S. L. Salzberg. Quake: quality-aware detection and correction of sequencing errors. Genome biology, 11(11):1–13, 2010.

A. C. Khatcheressian, James L Wolff, E. Smith, Thomas J Grunfeld, V. G. Muss, Hyman B Vogel, M. R. Halberg, Francine Somerfield, and N. E. Davidson. American society of clinical oncology 2006 update of the breast cancer follow-up and

management guidelines in the adjuvant setting. Journal of Clinical Oncology, 24 (31):5091–5097, 2006.

P. Khorsand and F. Hormozdiari. Nebula: Ultra-efficient mapping-free structural variant genotyper. Nucleic Acids Research, 49(8):e47–e47, 2021.

G. A. S. Kielbassa—Pavlos, A. C. Uricaru—Marie, F. S. Peterlongo, and V. Lacroix. kissplice: de-novo calling alternative splicing events from rna-seq data. 2011.

D. Kim, B. Langmead, and S. L. Salzberg. Hisat: a fast spliced aligner with low memory requirements. Nature methods, 12(4):357–360, 2015.

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, 22(3):568–576, 2012.

M. Kokot, M. Długosz, and S. Deorowicz. Kmc 3: counting and manipulating k-mer statistics. Bioinformatics, 33(17):2759–2761, 2017.

L. S. Kremer, D. M. Bader, C. Mertes, R. Kopajtich, G. Pichler, A. Iuso, T. B. Haack, E. Graf, T. Schwarzmayr, C. Terrile, et al. Genetic diagnosis of mendelian disorders via rna sequencing. Nature communications, 8(1):1–11, 2017.

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, 11(3):e0151664, 2016.

P. Kulkarni and P. Frommolt. Challenges in the setup of large-scale next-generation sequencing analysis workflows. Computational and structural biotechnology journal, 15:471–477, 2017.

S. Kurtz, A. Narechania, J. C. Stein, and D. Ware. A new method to compute k-mer frequencies and its application to annotate large repetitive plant genomes. BMC genomics, 9(1):1–18, 2008.

S. Lanciano and G. Cristofari. Measuring and interpreting transposable element expression. Nature Reviews Genetics, 21(12):721–736, 2020.

E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, et al. Initial sequencing and analysis of the human genome. 2001.

M. J. Landrum, J. M. Lee, M. Benson, G. R. Brown, C. Chao, S. Chitipiralla, B. Gu, J. Hart, D. Hoffman, W. Jang, et al. Clinvar: improving access to variant interpretations and supporting evidence. Nucleic acids research, 46(D1):D1062–D1067, 2018.

B. Langmead. Aligning short sequencing reads with bowtie. Current protocols in bioinformatics, 32(1):11–7, 2010.

J.-D. Larouche, A. Trofimov, L. Hesnard, G. Ehx, Q. Zhao, K. Vincent, C. Durette, P. Gendron, J.-P. Laverdure, É. Bonneil, et al. Widespread and tissue-specific expression of endogenous retroelements in human somatic tissues. Genome medicine, 12:1–16, 2020.

D. E. Larson, C. C. Harris, K. Chen, D. C. Koboldt, T. E. Abbott, D. J. Dooling, T. J. Ley, E. R. Mardis, R. K. Wilson, and L. Ding. Somaticsniper: identification of somatic point mutations in whole genome sequencing data. Bioinformatics, 28 (3):311–317, 2012.

L. Latino. Pseudolysogeny and sequential mutations build multiresistance to virulent bacteriophages in Pseudomonas aeruginosa. PhD thesis, Université Paris-Saclay, 2016.

C. M. Laumont, K. Vincent, L. Hesnard, É. Audemard, É. Bonneil, J.-P. Laverdure, P. Gendron, M. Courcelles, M.-P. Hardy, C. Côté, et al. Noncoding regions are the main source of targetable tumor-specific antigens. Science translational medicine, 10(470), 2018.

T. Laver, J. Harrison, P. OÂ¡Â¯neill, K. Moore, A. Farbos, K. Paszkiewicz, and

D. J. Studholme. Assessing the performance of the oxford nanopore technologies minion. Biomolecular detection and quantification, 3:1–8, 2015.

C. W. Law, Y. Chen, W. Shi, and G. K. Smyth. voom: Precision weights unlock linear model analysis tools for rna-seq read counts. Genome biology, 15(2):1–17, 2014.

E. Lee, R. Iskow, L. Yang, O. Gokcumen, P. Haseley, L. J. Luquette, J. G. Lohr, C. C. Harris, L. Ding, R. K. Wilson, et al. Landscape of somatic retrotransposition in human cancers. Science, 337(6097):967–971, 2012.

M.-P. Lefranc, V. Giudicelli, C. Ginestoux, J. Jabado-Michaloud, G. Folch, F. Bellahcene, Y. Wu, E. Gemrot, X. Brochet, J. Lane, et al. Imgt®, the international immunogenetics information system®. Nucleic acids research, 37(suppl_1): D1006–D1012, 2009.

N. Leng, J. A. Dawson, J. A. Thomson, V. Ruotti, A. I. Rissman, B. M. Smits, J. D. Haag, M. N. Gould, R. M. Stewart, and C. Kendziorski. Ebseq: an empirical bayes hierarchical model for inference in rna-seq experiments. Bioinformatics, 29 (8):1035–1043, 2013.

M. Levy-Sakin, S. Pastor, Y. Mostovoy, L. Li, A. K. Leung, J. McCaffrey, E. Young, E. T. Lam, A. R. Hastie, K. H. Wong, et al. Genome maps across 26 human populations reveal population-specific patterns of structural variation. Nature communications, 10(1):1–14, 2019.

B. Li and C. N. Dewey. Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. BMC bioinformatics, 12(1):1–16, 2011.

H. Li. wgsim-read simulator for next generation sequencing. Github repository, 2011.

H. Li and R. Durbin. Fast and accurate short read alignment with burrows–wheeler transform. bioinformatics, 25(14):1754–1760, 2009.

H. Li, J. Ruan, and R. Durbin. Mapping short dna sequencing reads and calling variants using mapping quality scores. Genome research, 18(11):1851–1858, 2008.

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, 25(16):2078–2079, 2009a.

H. Li, X. Feng, and C. Chu. The design and construction of reference pangenome graphs with minigraph. Genome biology, 21(1):1–19, 2020.

J. Li, D. Drubay, S. Michiels, and D. Gautheret. Mining the coding and non-coding genome for cancer drivers. Cancer letters, 369(2):307–315, 2015a.

M. M. Li, M. Datto, E. J. Duncavage, S. Kulkarni, N. I. Lindeman, S. Roy, A. M. Tsimberidou, C. L. Vnencak-Jones, D. J. Wolff, A. Younes, et al. Standards and guidelines for the interpretation and reporting of sequence variants in cancer: a joint consensus recommendation of the association for molecular pathology, american society of clinical oncology, and college of american pathologists. The Journal of molecular diagnostics, 19(1):4–23, 2017.

R. Li, C. Yu, Y. Li, T.-W. Lam, S.-M. Yiu, K. Kristiansen, and J. Wang. Soap2: an improved ultrafast tool for short read alignment. Bioinformatics, 25(15):1966–1967, 2009b.

T. W. Li and K. M. Weeks. Structure-independent and quantitative ligation of single-stranded dna. Analytical biochemistry, 349(2):242–246, 2006.

W. Q. Li, K. Kawakami, A. Ruszkiewicz, G. Bennett, J. Moore, and B. Iacopetta. Braf mutations are associated with distinctive clinical, pathological and molecular features of colorectal cancer independently of microsatellite instability status. Molecular cancer, 5(1):1–6, 2006.

X. Li, A. Nair, S. Wang, and L. Wang. Quality control of rna-seq experiments. In RNA Bioinformatics, pages 137–146. Springer, 2015b.

Z. Li, Y. Chen, D. Mu, J. Yuan, Y. Shi, H. Zhang, J. Gan, N. Li, X. Hu, B. Liu, et al. Comparison of the two major classes of assembly algorithms: overlap–layout–consensus and de-bruijn-graph. Briefings in functional genomics, 11(1): 25–37, 2012.

Y. Liao, G. K. Smyth, and W. Shi. featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics, 30(7): 923–930, 2014.

A. Limasset, B. Cazaux, E. Rivals, and P. Peterlongo. Read mapping on de bruijn graphs. BMC bioinformatics, 17(1):1–12, 2016.

H. Lin, Z. Zhang, M. Q. Zhang, B. Ma, and M. Li. Zoom! zillions of oligos mapped. Bioinformatics, 24(21):2431–2437, 2008.

H. Ling, M. Fabbri, and G. A. Calin. Micrornas and other non-coding rnas as targets for anticancer drug development. Nature reviews Drug discovery, 12(11): 847–865, 2013.

K. Litchfield, J. L. Reading, C. Puttick, K. Thakkar, C. Abbosh, R. Bentham, T. B. Watkins, R. Rosenthal, D. Biswas, A. Rowan, et al. Meta-analysis of tumor-and t cell-intrinsic mechanisms of sensitization to checkpoint inhibition. Cell, 184(3): 596–614, 2021.

C. Loeffler, A. Karlsberg, L. S. Martin, E. Eskin, D. Koslicki, and S. Mangul. Improving the usability and comprehensiveness of microbial databases. BMC biology, 18:1–6, 2020.

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, 45(6):580–585, 2013.

H. Lopez-Maestre, L. Brinza, C. Marchet, J. Kielbassa, S. Bastien, M. Boutigny, D. Monnin, A. E. Filali, C. M. Carareto, C. Vieira, et al. Snp calling from rna-seq data without a reference genome: identification, quantification, differential

analysis and impact on the protein sequence. Nucleic Acids Research, 44(19): e148–e148, 2016.

C. Lorenzi, S. Barriere, J.-P. Villemin, L. D. Bretones, A. Mancheron, and W. Ritchie. imoka: k-mer based software to analyze large collections of sequencing data. Genome Biology, 21(1):1–19, 2020.

M. Love, S. Anders, and W. Huber. Differential analysis of count data–the deseq2 package. Genome Biol, 15(550):10–1186, 2014.

G. Lunter and M. Goodson. Stampy: a statistical algorithm for sensitive and fast mapping of illumina sequence reads. Genome research, 21(6):936–939, 2011.

X. Ma, Y. Shao, L. Tian, D. A. Flasch, H. L. Mulder, M. N. Edmonson, Y. Liu, X. Chen, S. Newman, J. Nakitandwe, et al. Analysis of error profiles in deep next-generation sequencing data. Genome biology, 20(1):1–15, 2019.

D. MacArthur, T. Manolio, D. Dimmock, H. Rehm, J. Shendure, G. Abecasis, D. Adams, R. Altman, S. Antonarakis, E. Ashley, et al. Guidelines for investigating causality of sequence variants in human disease. Nature, 508(7497):469–476, 2014.

J. MacQueen et al. Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, volume 1, pages 281–297. Oakland, CA, USA, 1967.

K. Malde. The effect of sequence quality on sequence alignment. Bioinformatics, 24(7):897–900, 2008.

J. L. Maldonado, J. Fridlyand, H. Patel, A. N. Jain, K. Busam, T. Kageshita, T. Ono, D. G. Albertson, D. Pinkel, and B. C. Bastian. Determinants of braf mutations in primary melanomas. Journal of the National Cancer Institute, 95 (24):1878–1890, 2003.

I. Mandric, J. Rotman, H. T. Yang, N. Strauli, D. J. Montoya, W. Van Der Wey, J. R. Ronas, B. Statz, D. Yao, V. Petrova, et al. Profiling immunoglobulin

repertoires across multiple human tissues using rna sequencing. Nature communications, 11(1):1–14, 2020.

G. Marçais and C. Kingsford. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics, 27(6):764–770, 2011.

C. Marchet, Z. Iqbal, D. Gautheret, M. Salson, and R. Chikhi. Reindeer: efficient indexing of k-mer presence and abundance in sequencing datasets. Bioinformatics, 36(Supplement_1):i177–i185, 2020.

C. Marchet, C. Boucher, S. J. Puglisi, P. Medvedev, M. Salson, and R. Chikhi. Data structures based on k-mers for querying large collections of sequencing data sets. Genome Research, 31(1):1–12, 2021.

S. Margolin and A. Lindblom. Familial breast cancer, underlying genes, and clinical implications: a review. Critical Reviews™ in Oncogenesis, 12(1-2), 2006.

M. Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. journal, 17(1):10–12, 2011.

I. Martincorena, K. M. Raine, M. Gerstung, K. J. Dawson, K. Haase, P. Van Loo, H. Davies, M. R. Stratton, and P. J. Campbell. Universal patterns of selection in cancer and somatic tissues. Cell, 171(5):1029–1041, 2017.

N. Mathieu, L. Pirzio, M.-A. Freulet-Marrière, C. Desmaze, and L. Sabatier. Telomeres and chromosomal instability. Cellular and Molecular Life Sciences CMLS, 61(6):641–656, 2004.

M. Matvienko, A. Kozik, L. Froenicke, D. Lavelle, B. Martineau, B. Perroud, and R. Michelmore. Consequences of normalizing transcriptomic and genomic libraries of plant genomes using a duplex-specific nuclease and tetramethylammonium chloride. PLoS One, 8(2):e55913, 2013.

A. Mayakonda, D.-C. Lin, Y. Assenov, C. Plass, and H. P. Koeffler. Maftools: efficient and comprehensive analysis of somatic variants in cancer. Genome research, 28(11):1747–1756, 2018.

D. J. McCarthy, P. Humburg, A. Kanapin, M. A. Rivas, K. Gaulton, J.-B. Cazier, and P. Donnelly. Choice of transcripts and software has a large effect on variant annotation. Genome medicine, 6(3):1–16, 2014.

S. McGinnis and T. L. Madden. Blast: at the core of a powerful and diverse set of sequence analysis tools. Nucleic acids research, 32(suppl_2):W20–W25, 2004.

A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, et al. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. Genome research, 20(9):1297–1303, 2010.

W. McLaren, L. Gil, S. E. Hunt, H. S. Riat, G. R. Ritchie, A. Thormann, P. Flicek, and F. Cunningham. The ensembl variant effect predictor. Genome biology, 17 (1):1–14, 2016.

P. Medvedev, S. Pham, M. Chaisson, G. Tesler, and P. Pevzner. Paired de bruijn graphs: a novel approach for incorporating mate pair information into genome assemblers. Journal of Computational Biology, 18(11):1625–1634, 2011.

J. Meng and Y.-P. P. Chen. A database of simulated tumor genomes towards accurate detection of somatic small variants in cancer. PloS one, 13(8):e0202982, 2018.

K. Michailidou, S. Lindström, J. Dennis, J. Beesley, S. Hui, S. Kar, A. Lemaçon, P. Soucy, D. Glubb, A. Rostamianfar, et al. Association analysis identifies 65 new breast cancer risk loci. Nature, 551(7678):92–94, 2017.

S. Michiels, S. Koscielny, T. Boulet, and C. Hill. Gene expression profiling in cancer research. Bulletin du cancer, 94(11):976–980, 2007.

K. H. Miga, S. Koren, A. Rhie, M. R. Vollger, A. Gershman, A. Bzikadze, S. Brooks, E. Howe, D. Porubsky, G. A. Logsdon, et al. Telomere-to-telomere assembly of a complete human x chromosome. Nature, 585(7823):79–84, 2020.

T. Mikkelsen, L. Hillier, E. Eichler, M. Zody, D. Jaffe, S.-P. Yang, W. Enard, I. Hellmann, K. Lindblad-Toh, T. Altheide, et al. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature, 437(7055):69–87, 2005.

J. R. Miller, S. Koren, and G. Sutton. Assembly algorithms for next-generation sequencing data. Genomics, 95(6):315–327, 2010.

V. K. Mittal and J. F. McDonald. De novo assembly and characterization of breast cancer transcriptomes identifies large numbers of novel fusion-gene transcripts of potential functional significance. BMC medical genomics, 10(1):1–20, 2017.

A. Morillon and D. Gautheret. Bridging the gap between reference and real transcriptomes. Genome biology, 20(1):1–7, 2019.

A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold. Mapping and quantifying mammalian transcriptomes by rna-seq. Nature methods, 5(7): 621–628, 2008.

G. Narzisi, A. Corvelo, K. Arora, E. A. Bergmann, M. Shah, R. Musunuri, A.-K. Emde, N. Robine, V. Vacic, and M. C. Zody. Lancet: genome-wide somatic variant calling using localized colored debruijn graphs. bioRxiv, page 196311, 2017.

J. A. Nelder and R. W. Wedderburn. Generalized linear models. Journal of the Royal Statistical Society: Series A (General), 135(3):370–384, 1972.

C. G. A. R. Network et al. Comprehensive molecular profiling of lung adenocarcinoma. Nature, 511(7511):543, 2014.

P. C. Ng and S. Henikoff. Sift: Predicting amino acid changes that affect protein function. Nucleic acids research, 31(13):3812–3814, 2003.

H. T. Nguyen, H. Xue, V. Firlej, Y. Ponty, M. Gallopin, and D. Gautheret. Reference-free transcriptome signatures for prostate cancer prognosis. BMC cancer, 21(1):1–12, 2021.

G. Nilsen, K. Liestøl, P. Van Loo, H. K. M. Vollan, M. B. Eide, O. M. Rueda, S.-F. Chin, R. Russell, L. O. Baumbusch, C. Caldas, et al. Copynumber: efficient algorithms for single-and multi-track copy number segmentation. BMC genomics, 13(1):1–16, 2012.

Y. Ning, H. Zheng, Y. Zhan, S. Liu, H. Zang, J. Luo, Q. Wen, S. Fan, et al. Comprehensive analysis of the mechanism and treatment significance of mucins in lung cancer. Journal of Experimental & Clinical Cancer Research, 39(1):1–10, 2020.

S. Nurk, S. Koren, A. Rhie, M. Rautiainen, A. V. Bzikadze, A. Mikheenko, M. R. Vollger, N. Altemose, L. Uralsky, A. Gershman, S. Aganezov, S. J. Hoyt, M. Diekhans, G. A. Logsdon, M. Alonge, S. E. Antonarakis, M. Borchers, G. G. Bouffard, S. Y. Brooks, G. V. Caldas, H. Cheng, C.-S. Chin, W. Chow, L. G. de Lima, P. C. Dishuck, R. Durbin, T. Dvorkina, I. T. Fiddes, G. Formenti, R. S. Fulton, A. Fungtammasan, E. Garrison, P. G. Grady, T. A. Graves-Lindsay, I. M. Hall, N. F. Hansen, G. A. Hartley, M. Haukness, K. Howe, M. W. Hunkapiller, C. Jain, M. Jain, E. D. Jarvis, P. Kerpedjiev, M. Kirsche, M. Kolmogorov, J. Korlach, M. Kremitzki, H. Li, V. V. Maduro, T. Marschall, A. M. McCartney, J. McDaniel, D. E. Miller, J. C. Mullikin, E. W. Myers, N. D. Olson, B. Paten, P. Peluso, P. A. Pevzner, D. Porubsky, T. Potapova, E. I. Rogaev, J. A. Rosenfeld, S. L. Salzberg, V. A. Schneider, F. J. Sedlazeck, K. Shafin, C. J. Shew, A. Shumate, Y. Sims, A. F. A. Smit, D. C. Soto, I. Sović, J. M. Storer, A. Streets, B. A. Sullivan, F. Thibaud-Nissen, J. Torrance, J. Wagner, B. P. Walenz, A. Wenger, J. M. D. Wood, C. Xiao, S. M. Yan, A. C. Young, S. Zarate, U. Surti, R. C. McCoy, M. Y. Dennis, I. A. Alexandrov, J. L. Gerton, R. J. O'Neill, W. Timp, J. M. Zook, M. C. Schatz, E. E. Eichler, K. H. Miga, and A. M. Phillippy. The complete sequence of a human genome. bioRxiv, 2021. doi: 10.1101/2021.05.26.445798. URL https://www.biorxiv.org/content/early/2021/05/27/2021.05.26.445798.

D. O'Neil, H. Glowatz, and M. Schlumpberger. Ribosomal rna depletion for efficient use of rna-seq capacity. Current protocols in molecular biology, 103(1):4–19, 2013.

L. Ouchenir, C. Renaud, S. Khan, A. Bitnun, A.-A. Boisvert, J. McDonald, J. Bowes, J. Brophy, M. Barton, J. Ting, et al. The epidemiology, management, and outcomes of bacterial meningitis in infants. Pediatrics, 140(1), 2017.

R. Ounit and S. Lonardi. Higher classification sensitivity of short metagenomic reads with clark-s. Bioinformatics, 32(24):3823–3825, 2016.

T. Ouspenskaia, T. Law, K. R. Clauser, S. Klaeger, S. Sarkizova, F. Aguet, B. Li, E. Christian, B. A. Knisbacher, P. M. Le, et al. Thousands of novel unannotated proteins expand the mhc i immunopeptidome in cancer. bioRxiv, 2020.

F.-D. Pajuste, L. Kaplinski, M. Möls, T. Puurand, M. Lepamets, and M. Remm. Fastgt: an alignment-free method for calling common snvs directly from raw sequencing reads. Scientific reports, 7(1):1–10, 2017.

J. S. Parker, M. Mullins, M. C. Cheang, S. Leung, D. Voduc, T. Vickery, S. Davies, C. Fauron, X. He, Z. Hu, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. Journal of clinical oncology, 27(8):1160, 2009.

B. Paten, A. M. Novak, J. M. Eizenga, and E. Garrison. Genome graphs and the evolution of genome inference. Genome research, 27(5):665–676, 2017.

R. Patro, S. M. Mount, and C. Kingsford. Sailfish enables alignment-free isoform quantification from rna-seq reads using lightweight algorithms. Nature biotechnology, 32(5):462–464, 2014.

R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford. Salmon provides fast and bias-aware quantification of transcript expression. Nature methods, 14 (4):417–419, 2017.

L. M. Payer and K. H. Burns. Transposable elements in human genetic disease. Nature Reviews Genetics, 20(12):760–772, 2019.

F. Pereira, F. Azevedo, Â. Carvalho, G. F. Ribeiro, M. W. Budde, and B. Johansson. Pydna: a simulation and documentation tool for dna assembly strategies using python. BMC bioinformatics, 16(1):1–10, 2015.

M. Pertea, G. M. Pertea, C. M. Antonescu, T.-C. Chang, J. T. Mendell, and S. L. Salzberg. Stringtie enables improved reconstruction of a transcriptome from rna-seq reads. Nature biotechnology, 33(3):290–295, 2015.

P. Peterlongo, C. Riou, E. Drezen, and C. Lemaitre. Discosnp++: de novo detection of small variants from raw unassembled read set (s). BioRxiv, page 209965, 2017.

P. A. Pevzner, H. Tang, and M. S. Waterman. An eulerian path approach to dna fragment assembly. Proceedings of the national academy of sciences, 98(17): 9748–9753, 2001.

H. Pimentel, N. L. Bray, S. Puente, P. Melsted, and L. Pachter. Differential analysis of rna-seq incorporating quantification uncertainty. Nature methods, 14(7):687, 2017.

M. Pinskaya, Z. Saci, M. Gallopin, M. Gabriel, H. T. Nguyen, V. Firlej, M. Descrimes, A. Rapinat, D. Gentien, A. de La Taille, et al. Reference-free transcriptome exploration reveals novel rnas for prostate cancer diagnosis. Life science alliance, 2(6), 2019.

S. Pletscher-Frankild, A. Pallejà, K. Tsafou, J. X. Binder, and L. J. Jensen. Diseases: Text mining and data integration of disease–gene associations. Methods, 74:83–89, 2015.

J. R. Pon and M. A. Marra. Driver and passenger mutations in cancer. Annual Review of Pathology: Mechanisms of Disease, 10:25–50, 2015.

P. Priestley, J. Baber, M. P. Lolkema, N. Steeghs, E. de Bruijn, C. Shale, K. Duyvesteyn, S. Haidari, A. van Hoeck, W. Onstenk, et al. Pan-cancer whole-genome analyses of metastatic solid tumours. Nature, 575(7781):210–216, 2019.

B. Rabbani, M. Tekin, and N. Mahdieh. The promise of whole-exome sequencing in medical genetics. Journal of human genetics, 59(1):5–15, 2014.

E. M. Ramos, D. Hoffman, H. A. Junkins, D. Maglott, L. Phan, S. T. Sherry, M. Feolo, and L. A. Hindorff. Phenotype–genotype integrator (phegeni): synthesizing

genome-wide association study (gwas) data with existing genomic resources. European Journal of Human Genetics, 22(1):144–147, 2014.

S. H. Rangwala, L. Zhang, and H. H. Kazazian. Many line1 elements contribute to the transcriptome of human somatic cells. Genome biology, 10(9):1–18, 2009.

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

J. Ren, K. Song, M. Deng, G. Reinert, C. H. Cannon, and F. Sun. Inference of markovian properties of molecular sequences from ngs data and applications to comparative genomics. Bioinformatics, 32(7):993–1000, 2016.

A. Rhoads and K. F. Au. Pacbio sequencing and its applications. Genomics, proteomics & bioinformatics, 13(5):278–289, 2015.

M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. Nucleic acids research, 43(7):e47–e47, 2015.

G. Rizk, D. Lavenier, and R. Chikhi. Dsk: k-mer counting with very low memory usage. Bioinformatics, 29(5):652–653, 2013.

A. Roberts, C. Trapnell, J. Donaghey, J. L. Rinn, and L. Pachter. Improving rna-seq expression estimates by correcting for fragment bias. Genome biology, 12(3): 1–14, 2011.

G. Robertson, J. Schein, R. Chiu, R. Corbett, M. Field, S. D. Jackman, K. Mungall, S. Lee, H. M. Okada, J. Q. Qian, et al. De novo assembly and analysis of rna-seq data. Nature methods, 7(11):909–912, 2010.

J. T. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander, G. Getz, and J. P. Mesirov. Integrative genomics viewer. Nature biotechnology, 29(1):24–26, 2011.

M. D. Robinson and A. Oshlack. A scaling normalization method for differential expression analysis of rna-seq data. Genome biology, 11(3):1–9, 2010.

M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 26(1):139–140, 2010.

J. M. Rothberg and J. H. Leamon. The development and impact of 454 sequencing. Nature biotechnology, 26(10):1117–1124, 2008.

R. S. Roy, D. Bhattacharya, and A. Schliep. Turtle: Identifying frequent k-mers with cache-efficient algorithms. Bioinformatics, 30(14):1950–1957, 2014.

J. Rudewicz, H. Soueidan, R. Uricaru, H. Bonnefoi, R. Iggo, J. Bergh, and M. Nikolski. Micado–looking for mutations in targeted pacbio cancer data: an alignment-free method. Frontiers in genetics, 7:214, 2016.

S. M. Rumble, P. Lacroute, A. V. Dalca, M. Fiume, A. Sidow, and M. Brudno. Shrimp: accurate mapping of short color-space reads. PLoS Comput Biol, 5(5): e1000386, 2009.

N. Rusk. Torrents of sequence. Nature Methods, 8(1):44–44, 2011.

J. M. Sánchez-Torres, S. Viteri, M. A. Molina, and R. Rosell. Braf mutant non-small cell lung cancer and treatment with braf inhibitors. Translational lung cancer research, 2(3):244, 2013.

F. Sanger, S. Nicklen, and A. R. Coulson. Dna sequencing with chain-terminating inhibitors. Proceedings of the national academy of sciences, 74(12):5463–5467, 1977.

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, 28(14):1811–1817, 2012.

J. A. Schloss. How to get genomes at one ten-thousandth the cost. Nature biotechnology, 26(10):1113–1115, 2008.

R. Schmieder, Y. W. Lim, F. Rohwer, and R. Edwards. Tagcleaner: Identification and removal of tag sequences from genomic and metagenomic datasets. BMC bioinformatics, 11(1):1–14, 2010.

J.-S. Seo, Y. S. Ju, W.-C. Lee, J.-Y. Shin, J. K. Lee, T. Bleazard, J. Lee, Y. J. Jung, J.-O. Kim, J.-Y. Shin, et al. The transcriptional landscape and mutational profile of lung adenocarcinoma. Genome research, 22(11):2109–2119, 2012.

F. Seyednasrollah, A. Laiho, and L. L. Elo. Comparison of software packages for detecting differential expression in rna-seq studies. Briefings in bioinformatics, 16(1):59–70, 2015.

A. Shajii, D. Yorukoglu, Y. William Yu, and B. Berger. Fast genotyping of known snps through approximate k-mer matching. Bioinformatics, 32(17):i538–i544, 2016.

T. J. Sharpton. An introduction to the analysis of shotgun metagenomic data. Frontiers in plant science, 5:209, 2014.

J. Shendure and H. Ji. Next-generation dna sequencing. Nature biotechnology, 26 (10):1135–1145, 2008.

Q. Sheng, K. Vickers, S. Zhao, J. Wang, D. C. Samuels, O. Koues, Y. Shyr, and Y. Guo. Multi-perspective quality control of illumina rna sequencing data analysis. Briefings in functional genomics, 16(4):194–204, 2017.

R. M. Sherman and S. L. Salzberg. Pan-genomics in the human genome era. Nature Reviews Genetics, 21(4):243–254, 2020.

R. M. Sherman, J. Forman, V. Antonescu, D. Puiu, M. Daya, N. Rafaels, M. P. Boorgula, S. Chavan, C. Vergara, V. E. Ortega, et al. Assembly of a pan-genome from deep sequencing of 910 humans of african descent. Nature genetics, 51(1): 30–35, 2019.

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, 29(1):308–311, 2001.

J. Shiloach, S. Reshamwala, S. B. Noronha, and A. Negrete. Analyzing metabolic variations in different bacterial strains, historical perspectives and current trends–example e. coli. Current opinion in biotechnology, 21(1):21–26, 2010.

J. T. Simpson and R. Durbin. Efficient de novo assembly of large genomes using compressed data structures. Genome research, 22(3):549–556, 2012.

J. T. Simpson, K. Wong, S. D. Jackman, J. E. Schein, S. J. Jones, and I. Birol. Abyss: a parallel assembler for short read sequence data. Genome research, 19 (6):1117–1123, 2009.

N. Siva. 1000 genomes project, 2008.

C. C. Smith, S. R. Selitsky, S. Chai, P. M. Armistead, B. G. Vincent, and J. S. Serody. Alternative tumour-specific antigens. Nature Reviews Cancer, 19(8): 465–478, 2019.

Z. Sondka, S. Bamford, C. G. Cole, S. A. Ward, I. Dunham, and S. A. Forbes. The cosmic cancer gene census: describing genetic dysfunction across all human cancers. Nature Reviews Cancer, 18(11):696–705, 2018.

C. Soneson and M. Delorenzi. A comparison of methods for differential expression analysis of rna-seq data. BMC bioinformatics, 14(1):1–18, 2013.

L. Song and L. Florea. Rcorrector: efficient and accurate error correction for illumina rna-seq reads. GigaScience, 4(1):s13742–015, 2015.

M. R. Stratton, P. J. Campbell, and P. A. Futreal. The cancer genome. Nature, 458(7239):719–724, 2009.

M. Sultan, M. H. Schulz, H. Richard, A. Magen, A. Klingenhoff, M. Scherf, M. Seifert, T. Borodina, A. Soldatov, D. Parkhomchuk, et al. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science, 321(5891):956–960, 2008.

T. D. Sutton, A. G. Clooney, F. J. Ryan, R. P. Ross, and C. Hill. Choice of assembly software has a critical impact on virome characterisation. Microbiome, 7(1):1–15, 2019.

D. M. Swallow, S. Gendler, B. Griffiths, G. Corney, J. Taylor-Papadimitriou, and M. E. Bramwell. The human tumour-associated epithelial mucins are coded by an expressed hypervariable gene locus pum. Nature, 328(6125):82–84, 1987.

F. Syed, H. Grunenwald, and N. Caruccio. Next-generation sequencing library preparation: simultaneous fragmentation and tagging using in vitro transposition. Nature Methods, 6(11):i–ii, 2009.

D. Szklarczyk, A. L. Gable, D. Lyon, A. Junge, S. Wyder, J. Huerta-Cepas, M. Simonovic, N. T. Doncheva, J. H. Morris, P. Bork, et al. String v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic acids research, 47(D1):D607–D613, 2019.

A. J. Tanaka, M. T. Cho, F. Millan, J. Juusola, K. Retterer, C. Joshi, D. Niyazov, A. Garnica, E. Gratz, M. Deardorff, et al. Mutations in spata5 are associated with microcephaly, intellectual disability, seizures, and hearing loss. The American Journal of Human Genetics, 97(3):457–464, 2015.

K. Tang, J. Ren, and F. Sun. Afann: bias adjustment for alignment-free sequence comparison based on sequencing data using neural network regression. Genome biology, 20(1):1–17, 2019.

K. Taniue and N. Akimitsu. Fusion genes and rnas in cancer development. Non-coding RNA, 7(1):10, 2021.

J. Tazi, N. Bakkour, and S. Stamm. Alternative splicing and disease. Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease, 1792(1):14–26, 2009.

M. Tetreault, E. Bareke, J. Nadaf, N. Alirezaie, and J. Majewski. Whole-exome sequencing as a diagnostic tool: current challenges and future opportunities. Expert review of molecular diagnostics, 15(6):749–760, 2015.

I. The, T. P.-C. A. of Whole, G. Consortium, et al. Pan-cancer analysis of whole genomes. Nature, 578(7793):82, 2020.

T. M. Therneau and T. Lumley. Package 'survival'. R Top Doc, 128(10):28–33, 2015.

F. C. Thistlethwaite, D. E. Gilham, R. D. Guest, D. G. Rothwell, M. Pillai, D. J. Burt, A. J. Byatte, N. Kirillova, J. W. Valle, S. K. Sharma, et al. The clinical efficacy of first-generation carcinoembryonic antigen (ceacam5)-specific car t cells is limited by poor persistence and transient pre-conditioning-dependent respiratory toxicity. Cancer Immunology, Immunotherapy, 66(11):1425–1436, 2017.

T. Thomas, J. Gilbert, and F. Meyer. Metagenomics-a guide from sampling to data analysis. Microbial informatics and experimentation, 2(1):1–12, 2012.

V. Thorsson, D. L. Gibbs, S. D. Brown, D. Wolf, D. S. Bortone, T.-H. O. Yang, E. Porta-Pardo, G. F. Gao, C. L. Plaisier, J. A. Eddy, et al. The immune landscape of cancer. Immunity, 48(4):812–830, 2018.

K. Tomczak, P. Czerwińska, and M. Wiznerowicz. The cancer genome atlas (tcga): an immeasurable source of knowledge. Contemporary oncology, 19(1A):A68, 2015.

C. Trapnell, L. Pachter, and S. L. Salzberg. Tophat: discovering splice junctions with rna-seq. Bioinformatics, 25(9):1105–1111, 2009.

C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. Van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter. Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature biotechnology, 28(5):511–515, 2010.

C. Trapnell, A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, H. Pimentel, S. L. Salzberg, J. L. Rinn, and L. Pachter. Differential gene and transcript expression analysis of rna-seq experiments with tophat and cufflinks. Nature protocols, 7 (3):562, 2012.

C. Trapnell, D. G. Hendrickson, M. Sauvageau, L. Goff, J. L. Rinn, and L. Pachter. Differential analysis of gene regulation at transcript resolution with rna-seq. Nature biotechnology, 31(1):46–53, 2013.

A. Traschütz, J. van Gaalen, M. Oosterloo, M. Vreeburg, E.-J. Kamsteeg, N. Deininger, O. Rieß, M. Reimold, T. Haack, L. Schöls, et al. The movement disorder spectrum of sca21 (atx-tmem240): 3 novel families and systematic review of the literature. Parkinsonism & related disorders, 62:215–220, 2019.

K. A. Tryka, L. Hao, A. Sturcke, Y. Jin, Z. Y. Wang, L. Ziyabari, M. Lee, N. Popova, N. Sharopova, M. Kimura, et al. Ncbi0304s database of genotypes and phenotypes: dbgap. Nucleic acids research, 42(D1):D975–D979, 2014.

E. Turro, W. J. Astle, and S. Tavare. Flexible analysis of rna-seq data using mixed effects models. Bioinformatics, 30(2):180–188, 2014.

M. Uhlén, L. Fagerberg, B. M. Hallström, C. Lindskog, P. Oksvold, A. Mardinoglu, Å. Sivertsson, C. Kampf, E. Sjöstedt, A. Asplund, et al. Tissue-based map of the human proteome. Science, 347(6220), 2015.

R. Uricaru, G. Rizk, V. Lacroix, E. Quillery, O. Plantard, R. Chikhi, C. Lemaitre, and P. Peterlongo. Reference-free detection of isolated snps. Nucleic acids research, 43(2):e11–e11, 2015.

K. Van den Berge, K. M. Hembach, C. Soneson, S. Tiberi, L. Clement, M. I. Love, R. Patro, and M. D. Robinson. Rna sequencing data: hitchhiker's guide to expression analysis. 2019.

E. L. Van Dijk, Y. Jaszczyszyn, and C. Thermes. Library preparation methods for next-generation sequencing: tone down the bias. Experimental cell research, 322 (1):12–20, 2014.

T. P. van Gurp, L. M. McIntyre, and K. J. Verhoeven. Consistent errors in first strand cdna due to random hexamer mispriming. PloS one, 8(12):e85583, 2013.

Victor, T. Guryev, A. Marschall, F. Schonhuth, K. Vandin, and Ye. Computational pan-genomics: status, promises and challenges. Briefings in bioinformatics, 19 (1):118–135, 2018.

K. Vitting-Seerup and A. Sandelin. Isoformswitchanalyzer: analysis of changes in genome-wide patterns of alternative splicing and its functional consequences. Bioinformatics, 35(21):4469–4471, 2019.

T. Walsh, J. M. McClellan, S. E. McCarthy, A. M. Addington, S. B. Pierce, G. M. Cooper, A. S. Nord, M. Kusenda, D. Malhotra, A. Bhandari, et al. Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. science, 320(5875):539–543, 2008.

K. Wang, M. Li, and H. Hakonarson. Annovar: functional annotation of genetic variants from high-throughput sequencing data. Nucleic acids research, 38(16): e164–e164, 2010a.

K. Wang, D. Singh, Z. Zeng, S. J. Coleman, Y. Huang, G. L. Savich, X. He, P. Mieczkowski, S. A. Grimm, C. M. Perou, et al. Mapsplice: accurate mapping of rna-seq reads for splice junction discovery. Nucleic acids research, 38(18): e178–e178, 2010b.

L. Wang, S. Wang, and W. Li. Rseqc: quality control of rna-seq experiments. Bioinformatics, 28(16):2184–2185, 2012.

Q. Wang, C. S. Shashikant, M. Jensen, N. S. Altman, and S. Girirajan. Novel metrics to measure coverage in whole exome sequencing datasets reveal local and global non-uniformity. Scientific reports, 7(1):1–11, 2017.

X.-M. Wang, Z. Zhang, L.-H. Pan, X.-C. Cao, and C. Xiao. Krt19 and ceacam5 mrna-marked circulated tumor cells indicate unfavorable prognosis of breast cancer patients. Breast cancer research and treatment, 174(2):375–385, 2019.

R. H. Waterston and L. Pachter. Initial sequencing and comparative analysis of the mouse genome. Nature, 420(6915):520–562, 2002.

K. M. Wong, M. A. Suchard, and J. P. Huelsenbeck. Alignment uncertainty and genomic analysis. Science, 319(5862):473–476, 2008.

D. E. Wood and S. L. Salzberg. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome biology, 15(3):1–12, 2014.

D. C. Wu, J. Yao, K. S. Ho, A. M. Lambowitz, and C. O. Wilke. Limitations of alignment-free tools in total rna-seq quantification. BMC genomics, 19(1):1–14, 2018.

T. D. Wu, J. Reeder, M. Lawrence, G. Becker, and M. J. Brauer. Gmap and gsnap for genomic sequence alignment: enhancements to speed, accuracy, and functionality. In Statistical genomics, pages 283–334. Springer, 2016.

Y. Xie, G. Wu, J. Tang, R. Luo, J. Patterson, S. Liu, W. Huang, G. He, S. Gu, S. Li, et al. Soapdenovo-trans: de novo transcriptome assembly with short rna-seq reads. Bioinformatics, 30(12):1660–1666, 2014.

H. Y. Xiong, B. Alipanahi, L. J. Lee, H. Bretschneider, D. Merico, R. K. Yuen, Y. Hua, S. Gueroussov, H. S. Najafabadi, T. R. Hughes, et al. The human splicing code reveals new insights into the genetic determinants of disease. Science, 347 (6218), 2015.

F. Xu, J.-x. Chen, X.-b. Yang, X.-b. Hong, Z.-x. Li, L. Lin, and Y.-s. Chen. Analysis of lung adenocarcinoma subtypes based on immune signatures identifies clinical implications for cancer therapy. Molecular Therapy-Oncolytics, 17:241–249, 2020.

G. Yu, L.-G. Wang, Y. Han, and Q.-Y. He. clusterprofiler: an r package for comparing biological themes among gene clusters. Omics: a journal of integrative biology, 16(5):284–287, 2012.

M. Zapatka, I. Borozan, D. S. Brewer, M. Iskar, A. Grundhoff, M. Alawi, N. Desai, H. Sültmann, H. Moch, C. S. Cooper, et al. The landscape of viral associations in human cancers. Nature genetics, 52(3):320–330, 2020.

C. Zhang, B. Zhang, M. Vincent, and S. Zhao. Bioinformatics tools for rna-seq gene and isoform quantification. Next Generat. Sequenc. & Applic, 3:140, 2016.

H. Zhang, T. U. Ahearn, J. Lecarpentier, D. Barnes, J. Beesley, G. Qi, X. Jiang, T. A. O'Mara, N. Zhao, M. K. Bolla, et al. Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses. Nature genetics, pages 1–10, 2020a.

X. Zhang, R. Zhang, and J. Yu. New understanding of the relevant role of line-1 retrotransposition in human disease and immune modulation. Frontiers in Cell and Developmental Biology, 8:657, 2020b.

S. Zhao, Y. Zhang, W. Gordon, J. Quan, H. Xi, S. Du, D. von Schack, and B. Zhang. Comparison of stranded and non-stranded rna-seq transcriptome profiling and investigation of gene overlap. BMC genomics, 16(1):1–14, 2015.

J. M. Zook, D. Catoe, J. McDaniel, L. Vang, N. Spies, A. Sidow, Z. Weng, Y. Liu, C. E. Mason, N. Alexander, et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. Scientific data, 3(1):1–26, 2016.

List of Figures

List of Tables

Acronyms

AS **A**lternative **S**plicings. 44, 45

BQSR **b**ase **q**uality **s**core **r**ecalibration. 27, 28

cDBG **c**ompacted **D**e **B**ruijn **G**raph. 36, 41, 44

CNVs **C**opy **N**umber **V**ariants. 8, 23, 28, 56, 67, 77

DBG **D**e **B**ruijn **G**raph. 41, 42, 45, 46, 87, 93, 112

DE **E**ifferential **E**xpression. 54, 58, 60, 62, 74, 172

DNA-seq **DNA seq**uencing. 8, 12, 14, 18, 24, 26, 28, 31, 41, 47, 86

DU **D**ifferential **U**sage. 55, 62

EREs endogenous retroelements. 52, 64

HTS **H**igh-**T**hroughput **S**equencing. 5, 11

INDELS **In**sertions and **del**etions. 8, 23, 28, 29, 33, 60, 61, 62, 172

LUAD **lu**ng **ad**enocarcinoma. 52, 53, 56, 58, 60, 65, 69, 173

MAPQ **map**ping **q**uality. 26, 36

NGS **N**ext-**G**eneration **S**equencing. 5, 6, 8, 9, 11, 12, 24, 35, 37, 39, 40, 42, 43, 46, 47

OMIM **O**nline **M**endelian **I**nheritance in **M**an. 10, 33

PRAD **pr**ostate **ad**enocarcinoma. 88, 95, 96, 101, 105, 107, 112, 113

RNA-seq **RNA seq**uencing. 8, 12, 13, 14, 15, 16, 17, 18, 21, 22, 27, 41, 42, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 58, 59, 61, 64, 70, 171, 172

SBS **S**equencing **B**y **S**ynthesis. 5

SNPs **S**ingle-**N**ucleotide **P**olymorphisms. 8, 33, 35, 44, 45, 61

SNVs **S**ingle-**N**ucleotide **V**ariants. 8, 23, 25, 27, 28, 29, 46, 47, 60, 61, 70, 71, 93, 110, 115, 171, 176

SVs **S**tructural **V**ariants. 8, 28, 46, 47, 101, 121

TCGA **T**he **C**ancer **G**enome **A**tlas. 53, 55, 56, 58, 68, 80, 81, 173, 174

TMB **T**umor **M**utational **B**urden. 56, 67

TS **T**argeted **S**equencing. 9, 10, 13

WES **W**hole **E**xome **S**equencing. 9, 10, 13, 23, 24, 25, 27, 29, 32, 70, 88, 90, 92, 94, 95, 97, 98, 99, 101, 103, 104, 105, 108, 112, 171, 175, 176, 177

WGS **W**hole **G**enome **S**equencing. 9, 10, 11, 13, 23, 24, 25, 27, 32, 35, 70, 88, 92, 95, 96, 97, 100, 101, 102, 171, 177

université PARIS-SACLAY
ÉCOLE DOCTORALE
Structure et dynamique des systèmes vivants (SDSV)

Titre : l'analyse du génome et du transcriptome par des méthodes sans référence

Mots clés : Cancer, Bioinformatique, Transcriptome, Génome, Alignement de séquences

Résumé : Les comportements et les phénotypes des animaux sont en partie intégrés dans les molécules génétiques de la vie: l'ARN et l'ADN. En déchiffrant les informations cachées dans ces molécules, nous pouvons lever un voile sur les mystères de la biologie. Le séquençage de nouvelle génération (NGS) est un outil puissant pour décoder les molécules d'ADN et d'ARN à très grande échelle. Le NGS a considérablement élargi notre compréhension de tous les domaines de la biologie, de la biologie moléculaire à la génétique, la médecine, l'écologie et l'épidémiologie. Une pierre angulaire de l'analyse des données NGS est la comparaison avec un génome de référence. Bien que les scientifiques utilisent un génome de référence par espèce, la croissance explosive de la production de séquençage a remis en question ce point de vue en montrant que les séquences réelles d'ADN et d'ARN sont beaucoup plus diversifiées.

Dans cette thèse, nous proposons de nouveaux protocoles bioinformatiques pour l'analyse NGS qui ne reposent pas sur une référence. Nos projets visent à exploiter la puissance des approches sans alignement pour découvrir de nouvelles variations dans les transcriptomes et les génomes du cancer dans des régions difficiles à cartographier ou des régions absentes du génome de référence. Nous avons appliqué cette stratégie pour découvrir de nouveaux événements liés au phénotype à partir de cohortes de cancer à grande échelle. Du point de vue de l'analyse du génome, nous avons découvert de nouvelles variantes récurrentes de patients atteints de cancer de la prostate. Sur la base de l'analyse du transcriptome, nous avons découvert des événements de non-référence avec une haute réplicabilité.

Nous démontrons qu'un grand nombre de nouveaux événements pertinents pour les maladies peuvent être découverts sans alignement. Ces nouveaux événements non référencés ne nécessitent pas de connaissance a priori du génome humain ou du transcriptome et présentent des valeurs pronostiques significatives et un potentiel de production de néoantigènes. De plus, ces nouveaux événements non référencés impliqués dans le risque de cancer pourraient orienter les biologistes vers de nouveaux mécanismes d'oncogenèse.

université PARIS-SACLAY

ÉCOLE DOCTORALE
Structure et dynamique des systèmes vivants (SDSV)

Title : Transcriptome and Genome Analysis based on Alignment-free Protocols

Keywords : Cancer, Bioinformatic, Transcriptome, Genome, Sequence alignment

Summary : Animal's behaviors and phenotypes are in part embedded in the genetic molecules of life: RNA and DNA. By deciphering the information hidden in these molecules, we can peek into the mysteries of biology. Next generation sequencing (NGS) was developed as a powerful tool for decoding DNA and RNA molecules on a very large scale. NGS has considerably broadened our understanding of all areas of biology, from molecular biology to genetics, medicine, ecology and epidemiology. A cornerstone of NGS data analysis is the comparison with a reference genome. Although scientists use one reference genome per species, the explosive growth of sequencing output has challenged this view by showing that actual DNA and RNA sequences are much more diverse.

In this thesis, we propose new bioinformatics protocols for NGS analysis that do not rely on a reference. Our projects aim to exploit the power of alignment-free approaches to discover novel variations in cancer transcriptomes and genomes in hard-to-map regions or regions absent from the reference genome. We applied this strategy to uncover novel phenotype-related events from large-scale cancer cohorts. From the perspective of genome analysis, we uncovered novel recurrent variants from prostate cancer patients. Based on transcriptome analysis, we discovered non-reference events with high replicability.

We demonstrate that a large number of novel events relevant to diseases can be discovered in the manner of alignment-free. These novel non-reference events do not require a priori knowledge of the human genome or transcriptome and present significant prognostic values and potential to produce neoantigens. In addition, these novel non-reference events involved in cancer risk may orient biologists towards new oncogenesis mechanisms.