UNIVERSITÀ
DI SIENA
1240

Department of Molecular and Developmental Medicine
PhD School in Molecular Medicine
Scuola di dottorato in Medicina Molecolare
XXXIII Cycle
Coordinator: Prof. Vincenzo Sorrentino

Unraveling tandem repeat variation in personal genomes with long reads

Academic Discipline: MED/04

Supervisor:
Prof. Alberto Magi

Doctoral Dissertation of:
Dr. Davide Bolognini Davide Bolognini

Academic Year: 2019/2020

Preface

This thesis embraces most of the efforts I put during the last three years as a PhD student.

I have been working under the supervision of professor Alberto Magi, who is also the leader of the research group I am part of. In this time frame I had the wonderful opportunity of being initiated to bioinformatics, which radically changed the way I look at things and led me to discover my natural “thinking outside the box” attitude. I also took part of exciting joint works, with the year II spent at the European Molecular Biology Laboratory (EMBL) being at the first place. The work I am about to discuss in this thesis is not a one-man effort but stems from the collaboration between my home lab, EMBL’s Genomics Core Facility and EMBL’s Genome Biology Unit, which fits seamlessly into the collaborative spirit I was looking for in science.

This research is all about processing (DNA) strings. Clearly, the way I used to process strings has changed a bit since my classical studies. Still, I can remember me trying to identify prefixes and suffixes of Greek and Latin words to better understand their meaning. At the time, I never thought that processing strings would have been that interesting to me.

There is a humongous amount of people that, directly or indirectly, have contributed to my research and, in particular, to this work. Since my first step into the lab, I will not, ever, be thankful enough to Alberto, who, despite my initial skepticism, convinced me to submit that application for the PhD program and trusted me more than I ever did, since the very first moment.

For hosting and supporting my research abroad, I am thankful to Vladimir Benes, Jan O. Korbel and Tobias Rausch, who has been, and still is, an inexhaustible source of inspiration to me.

On the colleagues-side of these acknowledgments, I put all the guys from Alberto’s lab (Roberto Semeraro, Alessandra Mingrino and Gianluca Mattei), the members of the Genomics Core Facility (in particular Jonathan Landry and Jan Provaznik for their meaningful suggestions) and the members of the Genome Biology Unit (in particular Ashley D. Sanders and Hyobin Jeong for our constructive collaborations).

On the friends-side of this list, Marco Cecchi and Simone Romagnoli go first, for being kind of brothers to me.

I have tried to translate in simple words the infinite gratitude I have and will always have to my parents and Iulia for being my fixed point in life.
Obviously, I failed.

Abstract

Tandem repeats are repeated sequences that occur adjacent to each other in the human genome. Due to their prevalence and their association with a number of genetic diseases, there is a rising interest in developing tools for tandem repeat profiling.
Genome-wide discovery approaches are needed to fully understand their roles in health and disease but resolving tandem repeat variation accurately remains a very challenging task. Indeed, while traditional mapping-based and assembly-based approaches using short-read data have severe limitations in the size and type of tandem repeats they can resolve, recent third-generation sequencing technologies provide the long reads required to broaden the scope of detectable tandem repeats but exhibit substantially higher sequencing error rates that complicates repeat resolution.
In order to overcome limitations of prior methods, we developed TRiCoLOR, a freely-available tool for tandem repeat profiling using error-prone long reads from third-generation sequencing technologies.
The method can identify repetitive regions in long-read sequencing data *de novo* and resolve their motif and multiplicity in a haplotype-specific manner. The tool further includes methods to interactively visualize the identified repeats and to trace their Mendelian consistency in pedigrees.
Tested on synthetic data harboring tandem repeat contractions and expansions, TRiCoLOR demonstrates excellent performances and improved precision and recall compared to alternative tools. For real human whole-genome sequencing data, TRiCoLOR achieves high validation rates, suggesting its suitability to identify tandem repeat variation in personal genomes.
Compared to assembly-based approaches for structural variant detection, TRiCoLOR demonstrates capable to resolve tandem repeats in difficult to assemble regions that are prone to misassemblies or incorrect repeat assignments.
TRiCoLOR is open-source and implemented in python 3, with supporting C++ code and bash scripts. The tool is released through GitHub (`https://github.com/davidebolo1993/TRiCoLOR`) and as a docker image (`https://hub.docker.com/r/davidebolo1993/tricolor`), with accompanying documentation.

Contents

List of Figures

List of Tables

List of Acronyms

ApP Alignment Plotter 33

BAC Bacteria Artificial Chromosome 12

BAM Binary Alignment Map 19

CCS Circular Consensus Sequence 10

CLR Continuous Long Read 10

CRT Cyclic Reversible Termination 6

cPAL Combinatorial Probe–Anchor Ligation 5

cPAS Combinatorial Probe–Anchor Synthesis 5

DAG Directed Acyclic Graph 38

ddTTPs Dideoxythymidine Triphosphates 2

DIRS Dictyostelium Intermediate Repeat Sequence 18

dTTPs Deoxythymidine Triphosphates 2

DTW Dynamic Time Warping 30

dNTPs Deoxyribonucleotide Triphosphates 2

emPCR Emulsion PCR 4

F1 F1 score 48

FM-indexes Full-text indexes in Minute space 51

FN	False Negative	48
FP	False Positive	48
FXS	Fragile-X Syndrome	14
GEMs	Gel Bead-In EMulsions	13
HGSVC	Human Genome Structural Variation Consortium	33
HMM	Hidden Markov Model	29
LINEs	Long Interspersed Nuclear Elements	18
LTR	Long Terminal Repeat	18
MSA	Multiple Sequence Alignment	38
MSI	Microsatellite Instability	17
NCRF	Noise Cancelling Repeats Finder	28
NGS	Next Generation Sequencing	1
ONT	Oxford Nanopore Technologies	11
P	Precision	48
PacBio	Pacific Biosciences	10
PCR	Polymerase Chain Reaction	3
Pearson's R	Pearson Correlation Coefficient	49
PLE	Penelope-Like Element	18
POA	Partial Order Alignment	30
R	Recall	48
rDNA	ribosomal DNA	15
REFER	REpeats FindER	32
RegEx	Regular Expression	31
rRNAs	ribosomal RNAs	16

SA Suffix Arrays 51

SAGE SAmple GEnotyper 33

SAM Sequence Alignment/Map 19

satDNA satellite DNA 17

SBL Sequencing By Ligation 4

SBS Sequencing By Synthesis 4

SENSoR Shannon ENtropy ScanneR 32

SGS Second Generation Sequencing 4

SIMD Single Instruction Multiple Data 25

SINEs Short Interspersed Nuclear Elements 18

SLR Synthetic Long Reads 12

SMRT Single-Molecule Real-Time 9

SMS Single Molecule Sequencing 9

SNA Single-Nucleotide Addition 6

SNPs Single-Nucleotide Polimorphisms 15

SNVs Single-Nucleotide Variants 9

STRs Short Tandem Repeats 16

SVs Structural Variants 9

TEs Transposable Elements 17

TP True Positive 48

TRF Tandem Repeats Finder 21

TRiCoLOR Tandem Repeats Caller for LOng Reads 28

TRs Tandem Repeats 15

VNTRs Variable Number of Tandem Repeats 17

ZMWs Zero Mode Waveguides 10

Introduction 1

DNA sequencing evolves quickly. Barely 40 years have passed since the initial sequencing method has been developed in 1977 by Frederick Sanger and colleagues (Sanger et al., 1977). This revolutionary study triggered the improvement of new methods that have provided great opportunities for fast, low-cost and high-throughput DNA sequencing. Indeed, strikingly after the Human Genome Project[1], the time interval between emerging sequencing technologies has been substantially reduced while the amount of sequenced data has grown exponentially (Ari and Arikan, 2016). Considering Sanger sequencing as the first generation, new generations of DNA sequencing have been introduced subsequently and are collectively known as *Next Generation Sequencing* (NGS).

Some of the biggest technical challenges that are associated with NGS are caused by repetitive DNA (Alkan et al., 2011), *i.e.* DNA sequences that are similar or identical to others elsewhere in the genome. From a computational perspective, repeats create ambiguities in alignment and in genome assembly which, in turn, can produce errors when interpreting results.

The sections below describe the main aspects of the different sequencing generations as well as the challenges that are posed by

[1] `https://www.genome.gov/human-genome-project`.

repeats for genome resequencing projects and *de novo* genome assembly.

1.1 Sanger Sequencing: The First Generation

In 1977, Sanger and colleagues announced a new method for determining the nucleotide sequence in DNA, which is nowadays known as Sanger method. The technique was first applied to the DNA of bacteriophage ϕX174 and significantly improved over the plus and minus method from the same authors (Sanger and Coulson, 1975).

The Sanger method takes inspiration from a previous work that demonstrated the inhibitory activity of *Dideoxythymidine Triphosphates* (ddTTPs) on DNA polymerase I. Indeed, ddTTPs lack the $3'$ hydroxyl group needed to form the phosphodiester bond between one nucleotide and the next during DNA strand elongation and hence cause a chain termination reaction when incorporated into the nascent fragment by the DNA polymerase (Atkinson et al., 1969).

If an oligonucleotide primer and single-stranded target DNA are incubated in the presence of a mixture of *Deoxythymidine Triphosphates* (dTTPs) and ddTTPs, as well as the other three *Deoxyribonucleotide Triphosphates* (dNTPs), one of which ^{32}P-radiolabeled, a mixture of fragments having all the same $5'$ and with a ddTTP residue at the $3'$ ends is obtained. When this mixture is fractioned by electophoresis on acrylamide gel, the pattern of bands shows the distribution of dTTPs in the newly synthetized DNA. By using analogous terminators for the other nucleotides in separate incubations and running the samples in parallel on acrylamide gel (*i.e.* one lane for each type of dNTP), a pattern of bands is obtained, from which the entire sequence of the newly synthetized DNA can be deducted (Metzker, 2005).

Significant improvements to the Sanger method have been introduced over the years, including: (1) the development of fluorescent terminator dyes to eliminate the risk caused by the radioisotopes used for labelling; (2) the introduction of thermal-cycle sequencing to reduce the quantity of required input DNA and thermostable polymerases to efficiently and accurately incorporate the terminator dyes into the growing DNA strands; (3) the replacement

of acrylamide gel electophoresis with multichannel capillary electrophoresis powered by automated, refillable and reusable capillaries, and the introduction of electrokinetic sample loading.
Since 1987, the leader in automated Sanger sequencing is Applied Biosystems (nowadays part of Thermo Fisher Scientific). Their sequencers all utilize fluorescent dyes and capillary electrophoresis (from 4 to 48–96 capillaries) and generate 600–1000 bases of accurate sequence (Slatko et al., 2018). Automated Sanger sequencing platforms from Applied Biosystems were successfully utilized in the sequencing of the first human genome (Lander et al., 2001), taking into account 13 years of efforts of the Human Genome Project consortium and with an estimated cost of $2.7 billion.
Although relatively slow and not as cost-effective for high numbers of targets when compared to current NGS standards, the Sanger method remains the most appropriate sequencing strategy for applications where high throughput is not required (*e.g.* verify plasmid constructs or *Polymerase Chain Reaction* (PCR) products). Moreover, Sanger sequencing is currently used to complement NGS in regions that are notorously difficult-to-sequence (*e.g.* GC-rich and low-complexity regions), and to confirm NGS results (Behdad et al., 2015; Mu et al., 2016).

1.2 Second Generation Sequencing

In the 2000s, the concept of DNA sequencing underwent drastic changes. Particularly, the shotgun sequencing strategy[2], which was introduced during the Human Genome Project, laid the foundation for massively parallel sequencing. At the time, the release of the first truly high-throughput sequencing platform by Lynx Therapeutics (later purchased by Illumina) heralded a 50000-fold drop in the cost of human genome sequencing since the Human

[2] In shotgun sequencing, the starting DNA is broken up randomly into many small pieces, sort of in a shotgun fashion, with each of those pieces then sequenced individually. The resulting sequence reads generated from the different pieces are then analyzed by means of dedicated softwares, looking for stretches of sequence from different reads that are identical with one another. When identical regions are identified, they are overlapped with one another, allowing the two sequence reads to be stitched together. This process is repeated over and over and over again, eventually yielding the complete sequence of the starting piece of DNA.

Genome Project and marked the beginning of *Second Generation Sequencing* (SGS)
The major advance offered by SGS is the ability to produce an enormous volume of data, in several cases in excess of one billion short reads per instrument run, as well as to deliver fast and cost-effective genomic informations if compared to sequencing strategies based on the Sanger method (Barba et al., 2013). SGS approaches can be broadly classified into *Sequencing By Ligation* (SBL) and *Sequencing By Synthesis* (SBS) approaches.

On the one hand, in SBL approaches a probe sequence that is bound to a fluorophore hybridizes to a template[3] and is ligated to an adjacent oligonucleotide for imaging. The emission spectrum of the fluorophore indicates the identity of the bases complementary to specific positions within the probe. On the other, in SBS approaches, a polymerase is used and a signal, such as a fluorophore or a change in ionic concentration, identifies the incorporation of a nucleotide into the elongating strand.

Both in SBL and SBS approaches the template is first clonally amplified, usually on a solid surface. Indeed, having many thousands of identical copies of a DNA fragment in a defined area ensures that the signal can be distinguished from background noise. Moreover,the creation of millions of individual SBL/SBS reaction centres (each having its own clonal template), guarantees massive parallelization (Goodwin et al., 2016). Available strategies for clonal amplification of a template are summarized in Figure 1.

The first step of template generation is fragmentation[4] of the sample DNA followed by ligation to a common adaptor set for clonal amplification and sequencing.

In bead-based preparations (Figure 1, panel A), the template is hybridized to bead-bound primers. By means of *Emulsion PCR* (emPCR) the template is amplified so that, at the end, thousands of clonal DNA fragments are immobilized on a single bead. Beads can in turn be distributed onto a glass surface (Jae et al., 2007) or arrayed on a PicoTiterPlate (Leamon et al., 2003).

In solid-state strategies (Figure 1, panel B and C), amplification is achieved directly on a slide. Forward and reverse primers are

[3] DNA fragment that has to be sequenced.

[4] Fragmentation of a large DNA fragment into smaller fragments can be achieved mechanically (*e.g.* by passing the DNA through a narrow passage), by sonication or enzymatically.

covalently bound to the slide surface, either randomly or on a patterned slide (*i.e.* a flow cell), and provide complementary ends to which template can bind.

The only approach that achieves template enrichment in solution is currently the Complete Genomics technology used by the Beijing Genomics Institute (Figure 1, panel D). Here, DNA undergoes an iterative ligation, circularization and cleavage process to create a circular template, also known as rolling circle amplification, which generates up to 20 billion discrete DNA nanoballs that are in turn distributed onto a patterned slide surface containing features that allow a single nanoball to associate with each location (Drmanac et al., 2010).

1.2.1 Sequencing By Ligation

SBL approaches involve the hybridization and ligation of anchor fragments and labelled probes to the template.

In particular, an anchor fragment encodes a known sequence that is complementary to an adaptor sequence on the template and provides a site to initiate ligation. A probe can have either one (*i.e.* one-base-encoded probes) or two (*i.e.* two-base-encoded probes) known bases followed by a series of degenerate bases that drive complementarity between probe and template. After ligation, the template is imaged and the known base or bases in the probe are identified. A new cycle begins after complete removal of the anchor-probe complex or through cleavage to remove the fluorophore and to regenerate the ligation site. Figure 2 illustrates these details.

SBL sequencing platfroms from SOLiD utilize two-base-encoded probes (Figure 2, panel A). Therefore each fluorometric signal represents a dinucleotide. Because there are 16 possible dinucleotide combinations and these cannot be identified with spectrally-resolvable fluorophores, four signals, each representing a subset of four dinucleotide combinations, are used that are further deconvoluted during data analysis (Valouev et al., 2008).

Complete Genomics performs DNA sequencing using *Combinatorial Probe–Anchor Ligation* (cPAL) or *Combinatorial Probe–Anchor Synthesis* (cPAS), which is a modification of cPAL but very few details about this method are available (Fehlmann et al., 2016). In both approaches, hybridizing probes are from a pool of one-base-encoded probes (Figure 2, panel B).

1.2.2 Sequencing By Synthesis

SBS is a term used to describe numerous DNA-polymerase-dependent methods. Following the indications from Goodwin and collegues, SBS methods can be further classified into *Cyclic Reversible Termination* (CRT) and *Single-Nucleotide Addition* (SNA) approaches.

1.2.2.1 Cyclic Reversible Termination

In CRT approaches, terminator molecules that are similar to those used in Sanger sequencing are used, in which the $3'$ hydroxyl group is blocked (Guo et al., 2008). To start the process, the template DNA is primed by a sequence that is complementary to an adaptor region, which initiates polymerase binding. In each cycle, a mixture of all four dNTPs, which are individually labelled and $3'$-blocked, are added. After the incorporation of a single dNTP in each reaction center, unbound dNTPs are washed out and the surface is imaged in order to identify which dNTP was incorporated at each cluster. Fluorophore and blocking group are then removed and a new cycle begins.

Illumina CRT sequencers are currently the globally leading sequencing platforms in the next-generation sequencing market (Jeon et al., 2019). In Illumina platforms, dNTP identification is achieved through total internal reflection fluorescence microscopy using either two or four laser channels. In most Illumina platforms (*e.g.* the HiSeq series), each dNTP is bound to a single fluorophore that is specific to that base type, requiring four different imaging channels, whereas few (*i.e.* NextSeq and MiniSeq) implement a two-fluorophore system.

Qiagen GeneReader uses approximately the same approach used by Illumina sequencers. However, unlike Illumina platforms, GeneReader is intended to be an all-in-one SGS platform, from sample preparation to variant discovery, as it integrates both the QIAcube sample preparation system and the Qiagen Clinical Insight platform for variant analysis.

Figure 3 illustrates Illumina (Figure 3, panel A) and Qiagen (Figure 3, panel B) CRT approaches more in detail.

1.2.2.2 Single-Nucleotide Addition

In SNA approaches, a single signal mark the incorporation of a dNTP into the elongating strand. Thus, each of the four nucleotides must be added iteratively to the sequencing reaction to ensure only one dNTP is responsible for the generated signal. This does not require the dNTPs to be blocked, as the absence of the next nucleotide in the sequencing reaction prevents elongation. However, in homopolymer regions identical dNTPs are added all together and sequence identification relies on a proportional increase in the incorporation signal. Figure 4 summarizes SNA approaches.

The first SNA platform was a 454 pyrosequencing device, distributed by Roche (Figure 4, panel A). This system distributes template-bound beads into a PicoTiterPlate along with beads containing an enzyme cocktail. When a dNTP is incorporeted into a strand, an enzymatic reaction lead to a bioluminescence signal, which is in turn detected by a charge-coupled device camera and traslated into the incorporation of one or more identical dNTPs at a particular bead (Nyrén, 2015).

The Ion Torrent platforms, distributed by Thermo Fisher Scientific (Figure 4, panel B), detect the H^+ ions that are released as each dNTP is incorporated. The resulting change in pH is detected by an integrated complementary metal-oxide-semiconductor and an ion-sensitive field-effect transistor, with pH changes being, theoretically, proportional to the number of nucleotides detected.

Several short-read sequencing platforms exist, each having its own strengths and weaknesses.

SBL approaches by SOLiD and Complete Genomics generate highly accurate data (estimated accuracy is $\sim$ 99.99%), as each base is probed multiple times (Liu et al., 2012). However, there are evidence that all under-estimate AT-rich regions (Rieber et al., 2013), with SOLiD devices displaying some substitution errors and some GC-rich under-representation (Harismendy et al., 2009). Moreover, while Complete Genomics' latest platform[5] extends the length of the reads generated up to 150 bases for paired-end

[5] `https://en.mgitech.cn/products/instruments_info/5/`.

sequencing[6], the maximum read length for SOLiD platforms is just 75 bases, strongly limiting their use for genome assembly and structural variant detection applications.

As mentioned above, SBS platforms from Illumina dominate the short-read sequencing industry, mainly thanks to the large variety of available devices that guarantee a wide range of applications including genomics, transcriptomics and epigenomics (Park, 2009; Wang et al., 2009; Buenrostro et al., 2013; Carless, 2015). The suite of Illumina platforms ranges from the low-throughput MiniSeq to the ultra-high-throughput HiSeq X, with a set of 10 HiSeq X devices being capable to deliver over 18000 human genomes to 30X coverage[7] per year, reducing the cost for a single genome down to \$1000[8]. Although the overall accuracy of Illumina platforms is high (estimated accuracy is $\sim 99.50\%$), they do share with SBL approaches some under-representation in AT-rich and GC-rich regions (Nakamura et al., 2011), as well as a tendency towards substitution errors (Minoche et al., 2011). Among SBS platforms, the Qiagen GeneReader is a clinical device with an explicit focus on cancer gene panels (Darwanto et al., 2017). Although this severely limits its possible applications, it is well optimized within its niche.

SNA approaches offer superior read lengths compared to other short-read sequencers, with reads up to an average of 700 bases for the 454 pyrosequencing devices and 400 bases for the Ion Torrent platforms. Despite the overall error rate is comparable to the other SGS platforms in non-homopolymer regions, homopolymers have proven problematic for these platforms, especially those larger than 6–8 bases (Loman et al., 2012). While 454 pyrosequencing platforms have been unable to compete with the others SGS devices and have been discontinued since 2016, the Ion Torrent

[6] Compared to single-read sequencing, which involves sequencing DNA from only one end, paired-end sequencing allows users to sequence both ends of a fragment. Standard paired-end sequencing provides a pair of reads, 150 bases in length each, that flank a DNA fragment of about 50 bases in length, which is not sequenced.

[7] Per-base coverage is the average number of times a base of a genome is sequenced. The coverage depth of a genome is calculated as the number of bases of all short reads that match a genome divided by the length of this genome.

[8] `https://www.illumina.com/systems/sequencing-platforms/hiseq-x.html?langsel=/us/`.

platforms, thanks to their short runtimes, are currently used for gene-panel sequencing and for point-of-care clinical applications, including transcriptome profiling and splice site identification (Li et al., 2014; Malapelle et al., 2015).

Overall, SGS technologies have become a standard for many applications in basic as well as clinical biology. However, the short length of the reads generated pose several limitations. Indeed, while small variants such as *Single-Nucleotide Variants* (SNVs) and short indels can be accurately detected using SGS platforms, large *Structural Variants* (SVs) are challenging to detect and characterize with such technologies, which is an important issue given the high number of diseases related to SVs (Weischenfeldt et al., 2013). In addition, short reads have a limited capacity to link independent variations on the same nucleic acid molecule, thus not being well suited to discriminate and phase alleles to their respective parental homolog, which is important for many aspects of human genetics (Tewhey et al., 2011). Moreover, it has been shown that, despite the use of sophisticated bioinformatic algorithms, it is often impossible to accurately map, or even assemble, short reads originating from regions harboring repetitive sequences, extreme guanine-cytosine content or sequences with multiple homologous elements within the genome (Mantere et al., 2019).

1.3 Third Generation Sequencing

In the 2010s TGS technologies emerged, which provide reads in excess of several kilobases and allow to overcome limitations of SGS (van Dijk et al., 2018). Among TGS technologies, the *Single Molecule Sequencing* (SMS) and the synthetic approaches can be distinguished. The SMS approaches differ from short-read approaches in that they do not rely on the amplification of DNA fragments nor do they require chemical cycling for each dNTP added. Alternatively, the synthetic approaches do not generate real long-reads; rather, they represent an approach to library preparation that leverages barcodes to allow computational assembly of larger fragments. The SMS approaches are further classified into *Single-Molecule Real-Time* (SMRT)-based and nanopore-based strategies.

1.3.1 Single-Molecule Real-Time Sequencing

In early 2011 *Pacific Biosciences* (PacBio) released the PacBio RS sequencer, based on the SMRT sequencing technology (Eid et al., 2009). This technology uses a closed, circular, single-stranded DNA template, called SMRTbell, which is created by ligating hairpin adaptors to both ends of a double-stranded target DNA molecule (Voskoboynik et al., 2013). A primer and a polymerase are annealed to one of the adaptors, followed by library loading onto a specialized flow cell containing up to 8000000 picolitre wells[9] called *Zero Mode Waveguides* (ZMWs). In each ZMW, a modified DNA polymerase is immobilized at the bottom, where it replicates the target DNA. During the replication process, the incorporation of fluorescently labeled nucleotides produces fluorescence signals upon excitation by a laser and a camera system records the color and duration of the emitted light in real time. An overview of the SMRT sequencing approach is given in panel Aa of Figure 5.

The time between nucleotide incorporations is also recorded, which is delayed when a nucleotide epigenetically modified (*e.g.* 6-methyladenosine) is incorporated, allowing the detection of base modifications. The approach is shown in panels A, B and C of Figure 6.

While initially the average read length was relatively short (∼1500 bases) and the average error rate high (∼13%) (Quail et al., 2012), over recent years the average read length has increased more than tenfold and the introduction of the *Circular Consensus Sequence* (CCS) technology for molecules up to 2 kilobases have strongly improved their overall accuracy (∼ 99.8%) (Wenger et al., 2019). The CCS technology is based on the idea that, as the SMRTbell forms a closed circle, after the polymerase replicates one strand of the target DNA, it can continue using the adaptor and then the other strand as a template. If the lifetime of the polymerase is long enough, both strands can be sequenced multiple times in a single *Continuous Long Read* (CLR). CLR sequences originate from multiple passes and can be split into multiple

[9] `https://www.pacb.com/products-and-services/sequel-system/`.

subsequences by simply recognizing and cutting out the adaptor sequences. A consensus sequence[10] of the subsequences can then be formed (*i.e.* the CCS).

1.3.2 Nanopore Sequencing

The first attempt at using nanopores in a membrane to sequence single-stranded DNA molecules was done at the end of the 1980s (Deamer et al., 2016) but, due to technical limitations, the first successful sequencing results were reported only in 2012 (Manrao et al., 2012). In 2014 *Oxford Nanopore Technologies* (ONT) released the MinION, a pocket-sized sequencing device using nanopores as biosensors (Ip et al., 2015), which lowered the cost of a sequencing run down to 1000$[11].

Nanopore sequencing occurs in a flow cell in which two ionic solution-filled compartments are separated by a membrane with up to 12000[12] individual nanopores incorporated. A costant voltage bias is applied, which generates an ionic currentli through each nanopore and, upon translocation of a DNA molecule, changes in the ionic current can be observed and characterized (Bolognini et al., 2019). The first results demonstrating the feasibility of nanopore sequencing were obtained using α-hemolysin pores (Jetha et al., 2009) but the first real nanopore sequencing results were obtained using the MspA pores (Laszlo et al., 2016) and currently CsgG pores are used (Carter and Hussain, 2017).

After library preparation, where each DNA fragment is end-repaired and ligated to a proper adaptor[13], double-stranded DNA is unwound at the pore, after which one strand passes in and is translated into an actual sequence of bases. An overview of the nanopore-based sequencing strategy is given in panel Ab of Figure 5.

The observed shifts in voltage depend on which part (*i.e.* k-

[10] A DNA consensus sequence is a theoretical representative nucleotide sequence in which each nucleotide is the one which occurs most frequently at that site in the different sequences which occur in nature.

[11] `https://nanoporetech.com/products/minion`

[12] `https://nanoporetech.com/products/promethion`.

[13] DNA–protein complex with a tightly bound helicase enzyme that ensures stepwise movement of the DNA through the pore.

mer[14]) of the DNA molecule flows through the pore at a certain time. Rather than having four possible signals (*i.e.* one for each nucleotide), the sequencing device has thousands (*i.e.* one for each possible k-mer), as it also takes into account signals from epigenetically-modified bases, as shown in panels D, E and F of Figure 6.
In contrast to SMRT sequencing, read length in nanopore sequencing is not limited by the technology itself but rather by the length of the DNA molecules to be sequenced. Thus, by using dedicated protocols, ultra-long reads can been obtained. A major drawback of nanopore sequencing is that the high error rate ($\sim$13%) of the sequenced reads can't be reduced by sequencing the same strand multiple times, as with SMRT sequencing. In order to increase the accuracy, ONT developed a method to sequence both strands of a double-stranded DNA molecule. In this method, called 1D^2 as opposed to the 1D system described above, an adaptor with a specialized sequence promotes the entry of the second strand into the pore after the first strand has passed through. However, a small boost in terms of accuracy[15] comes at the cost of a lower throughput, as both strands of each molecule are sequenced, doubling the consumption of the pore. More interestingly, an approach to mimic CCS from PacBio has been reported, which uses the ϕ29 polymerase to produce a tandem array of copies of the original DNA molecule (Li et al., 2016).

1.3.3 Synthetic Long Reads

Unlike true sequencing platforms, *Synthetic Long Reads* (SLR) technologies rely on a system of barcoding to associate fragments that are sequenced on existing short-read sequencers. Currently, the Illumina SLR sequencing platform and the 10X Genomics emulsion-based system exist, which show similarities with the the earlier BAC-by-BAC sequencing, where a set of overlapping *Bacteria Artificial Chromosome* (BAC) clones is ordered along the chromosomes of a target genome followed by shotgun sequencing

[14] k-mers are subsequences of length k contained within a biological sequence. 5-mers signals are currently registered by ONT sequencers (Lu et al., 2016).

[15] `https://nanoporetech.com/about-us/news/1d-squared-kit-available-store-boost-accuracy-simple-prep`.

of each clone individually (Venter et al., 1996).
With the Illumina SLR system, genomic DNA is sheared into fragments up to 10 kilobases long and ligated to adaptors that are used to denote the extremities of contigs[16] during downstream short-read assembly. These large fragments are then partitioned into a microtiter plate (∼3000 fragments per well) and undergo further shearing and barcodes addition through a tagmentation process[17], with each weel containing a single barcode. The DNA is then pooled and subjected to classical Illumina sequencing followed by local assembly to reconstruct the original long fragments. The Illumina SLR sequencing approach is illustrated in panel Ba of Figure 5. Although still supported, the Illumina kit for SLR sequencing has been recently discontinued[18].
In the 10X Genomics emulsion-based sequencing, DNA fragments of up to ∼100 kilobases are formed and mixed into micelles called *Gel Bead-In EMulsions* (GEMs). Within each GEM, a gel bead dissolves and smaller fragments of DNA are amplified from the original large fragments, each with a barcode identifying the source GEM. Barcoded fragments are then pooled, followed by classical Illumina library preparation and sequencing. The obtained reads are assembled to form a series of anchored fragments that can span up to ∼80 kilobases[19]. Unlike the Illumina system, this approach does not attempt gapless, end-to-end coverage of a single DNA fragment but relies on linked-reads, with dispersed, small fragments that are derived from a single long molecule sharing a communal barcode. Although these fragments leave segments of the original large molecule without any coverage, the gaps are overcome by ensuring that there are many long fragments from the same genomic region in the initial preparation, thus generating a read cloud wherein linked-reads from each long fragment can be stacked, combining their individual coverage into an overall map. An overview of the SLR sequencing strategy using linked-reads is

[16] A contig is a series of overlapping DNA sequences used to make a physical map that reconstructs the original DNA sequence of a chromosome or a region of a chromosome.

[17] Transposon cleaving and tagging of the double-stranded DNA with a universal overhang.

[18] `https://emea.illumina.com/science/technology/next-generation-sequencing/long-read-sequencing.html`.

[19] `https://www.10xgenomics.com/linked-reads/`.

given in panel Bb of Figure 5.

Long-read sequencing methods are frequently used to complement previous short-read strategies in assemblies. A major example is the human genome. Indeed, despite it is considered to be one of the most complete mammalian reference assemblies, more than 160 euchromatic gaps remained after the 1000 Genomes Project (Nothnagel et al., 2011), often enriched for repeated sequences and high GC content (Schmidt and Pearson, 2016). Thanks to SMRT sequencing, most of these were either closed or extended, more than 1 megabase of sequence was added and tens of thousands of structural variants were resolved (Chaisson et al., 2015). SMRT is a great strategy to overcome the low accuracy of SGS in extremely repetitive and GC-rich regions, which is also confirmed by the fact that kilobases-long repeated stretches of CGG implicated in the *Fragile-X Syndrome* (FXS) have been resolved and further characterized using SMRT sequencing (Loomis et al., 2013; Ardui et al., 2017).

Concerning ONT devices, the low throughput of MinION initially limited its use to the sequencing and assembly of small bacterial genomes (Loman et al., 2015). More recently, with the introduction of higher throughput platforms (*i.e.* GridION and PromethION), assemblies of larger genomes have been reported, including human (Jain et al., 2018a). In this study, reads up to 882 kilobases long were obtained. Comparative studies suggest that SMRT and nanopore sequencing perform similarly well for *de novo* genome assembly (Giordano et al., 2017), with the ultra-long nanopore reads enabling the measurement of telomere repeats, which is not possible with the shorter SMRT reads.

Thus, a particular strength of nanopore ultra-long reads is the resolution of extremely long repeated regions that can be resolved with no other technology. In 2004, Rudd and collegues demonstrated that even the most complete human assembly exhibited a lack of centromeric sequences that comprise hundreds or thousands of repeats of α-satellite monomers (Rudd and Willard, 2004). Recently, Jain and collegues succeeded in producing nanopore reads long enough to cover the hundreds of kilobase-long centromeric sequences of the human Y chromosome (Jain et al., 2018b).

Overall, SLR can resolve certain types of repetitive elements (Mc-

Coy et al., 2014) while have difficulties in resolving more tandemly arranged repetitive sequences, as this system relies on the local assembly of short reads. Although less suitable for sequencing highly repetitive regions, SLR approaches are well suited for genome phasing, where the high level of accuracy is clearly an advantage in phasing *Single-Nucleotide Polimorphisms* (SNPs).

1.4 Repetitive DNA

Pioneering work by Britten and Kohne revealed that, in addition to unique sequences, the eukaryotic genomes contain large quantities of repetitive DNA, which was initially classified into moderately or highly repetitive sequences according to their degree of repetitiveness (Britten and Kohne, 1968). Later, the repetitive DNA sequences were grouped according to other criteria such as their organization (tandemly arrayed[20] or dispersed[21]) or their functional role. Although repetitive DNA sequences include several types of protein-coding sequences, most of the repetitive part of the genome was earlier considered *junk DNA* with no known function (López-Flores and Garrido-Ramos, 2012). Today, with many genomes completely sequenced and the background research of more than 40 years, we have ample information on the significance of the repetitive DNA within eukaryotic genomes and concepts are changing.

As shown in Figure 7, approximately 50% of the human genome is comprised of repeats. Among *Tandem Repeats* (TRs) there are both moderately repetitive DNA, such as *ribosomal DNA* (rDNA), and highly repetitive microsatellite, minisatellite and satellite DNA.

rDNA genes are among the best-known examples of multigene families, *i.e.* groups of paralogous genes[22], and encodes the major

[20] DNA repeats that are adjacent to each other and can involve as few as two copies or many thousands of copies.

[21] Identical or nearly identical DNA sequences that are separated by hundreds, thousands or even millions of nucleotides in the source genome.

[22] Class of homologous genes (*i.e.* genes that appear in multiple creatures, because they derive from a common evolutionary ancestor), resulting from one or more duplication events. After duplication, the paralogous genes can keep the same function (*e.g.* the rDNA genes) but can also diverge and develop different functions.

ribosomal RNAs (rRNAs).
Microsatellites are TRs in which the repeat unit contains from 1 to 6 bases, thus being also known as *Short Tandem Repeats* (STRs). Approximately 1 million STR *loci* have been found in both protein-coding and non-coding regions, including regulatory sequences (Liu et al., 2019). Dinucleotides are the foremost type of microsatellite repeats for many species, with the most common dinucleotide repeat type in the human genome being $(CA)_n/(GT)_n$. Microsatellites have a characteristic mutational behavior and their mutation rates are 10 to 100000 times higher than average mutation rates in other parts of the genome (Gemayel et al., 2010). Mutations are mainly due to contractions or expansions in the number of repeat units, caused either by strand-slippage during DNA replication or unequal crossing over. Moreover, mutation rates vary between different microsatellites depending on: (1) the number of repeat units. In particular, the more repeat units, the more unstable the microsatellite, as longer *loci* are more likely to mispair during DNA replication (Lai and Sun, 2003). (2) the repeat purity. Interrupted microsatellite repeats have lower mutation rates than pure repeats, which might be due to a lower rate of mispairing between non-identical repeat units (Shah et al., 2010). (3) the length of the repeat unit. Microsatellite arrays containing longer repeat units evolve faster than those containing shorter units (Chakraborty et al., 1997), probably due to relatively inefficient repair of larger mismatched segments by cell-repair processes. Thanks to these characteristics, microsatellites provide a tool for the estimation of genetic variability within populations and a valuable approach to analysis of parentage. Indeed, their high mutation rates lead to a large number of alleles existing in a single *locus*, so that unrelated individuals are unlikely to share alleles, and they are codominant, which allows for exact genotyping and more precise genetic comparisons between individuals, because heterozygotes can be distinguished from homozygotes (Webster and Reichart, 2005).
In contrast to their historical definition as nonfunctional DNA, microsatellites are currently known to to play a central role both in physiology and pathology. On the one hand, microsatellites are involved in a range of functions such as chromatin organization, regulation of gene activity, recombination, DNA replication, cell cycle, mismatch repair system (Li et al., 2002). On the other, an expansion of the number of repeats located in coding as well as

in untranslated or regulatory regions of specific genes has been identified as the main cause of several neurological diseases, which are further described in Table 1. Moreover, *Microsatellite Instability* (MSI)[23] has been reported in the sporadic colon, gastric, sporadic endometrial and the majority of other cancers, with prognostic and therapeutic implications (Nojadeh et al., 2018).

Minisatellites are defined as TRs with a repeat unit longer than 6 bases (Näslund et al., 2005). Minisatellites can be either monomorphic or polymorphic, with the latter also known as *Variable Number of Tandem Repeats* (VNTRs). Although VNTRs were the first highly polymorphic markers described for the genetic analysis of human traits (Nakamura et al., 1987), they have been soon replaced by microsatellite markers mainly beacuse, while microsatellites are widespread in the genome and easier to clone and characterize, VNTRs are concentrated mostly in the telomeric regions of chromosomes (Vergnaud et al., 1993). More recently, there has been renewed interest in VNTRs, with the realization that they might have important functional roles. For example, it has been shown that VNTRs regulates the expression of specific genes (Michelhaugh et al., 2001) and influence tranlsation efficiency (Nakamura et al., 1998).

Lastly, *satellite DNA* (satDNA) has commonly repeated unit lengths of about 150–180 bases or 300–360 bases and is the main component of the heterochromatin, which is found specifically at pericentromeric and subtelomeric locations of the chromosomes (Garrido-Ramos, 2017). As for microsatellites and minisatellites, in the last few decades results from different studies point to a functional significance of satDNA. These functions include a role in the establishment and maintenance of chromatin states by promoting heterochromatin assembly, influencing gene expression, and contributing to epigenetic regulatory processes, as satellite repeats transcribe and are a source of short interfering RNA molecules (Ugarkovic, 2005).

Among dispersed repeats, *Transposable Elements* (TEs) stand out.

Transposable elements are DNA sequences that are able to move

[23] A unique molecular alteration and hyper-mutable phenotype, which is the result of a defective DNA mismatch repair system, and can be defined as the presence of alternate sized repetitive DNA sequences which are not present in the corresponding germ line DNA.

from one chromosomal position to another within the same genome and are divided into retrotransposons, which are transposed through an RNA intermediate[24], and DNA transposons, which can move without any RNA intermediate. Accordingly to Wicker and collegues, retrotransposons can be further classified into: (1) *Long Terminal Repeat* (LTR) retrotransposons; (2) *Dictyostelium Intermediate Repeat Sequence* (DIRS) retrotransposons; (3) non-LTR retrotransposons or *Long Interspersed Nuclear Elements* (LINEs); (4) *Penelope-Like Element* (PLE) retrotransposons; (5) *Short Interspersed Nuclear Elements* (SINEs) (Wicker et al., 2007). The most obvious effect of the mobility of TEs is the induction of insertional mutations which are a major source of genetic innovation and evolution but have also been found involved in several genetic diseases (Cordaux and Batzer, 2009) and cancer as well (Konkel and Batzer, 2010). In addition, the ectopic recombination between non-allelic homologous elements can generate various types of rearrangements and lead to inversions, deletions, translocations or duplications.

As mentioned above, repeats pose several challenges for both genome resequencing and *de novo* assembly projects using SGS technologies. A more detailed discussion of these challenges as well as of the computational strategies for solving repeat-induced analysis problems with SGS is given as follows, together with an overview of the most recent TR callers for SGS.

1.4.1 Genome Resequencing

Genome resequencing allows researchers to study genetic variation by mapping reads from a sequenced individual to a high-quality reference genome of the same species. Several aligners for short reads are available, some of which are listed in Table 2. A major problem for short-read aligners is trying to decide what to do with reads that map to multiple locations (*i.e.* the multi-reads), such as reads coming from repeated regions. The percentage of short reads that map to a unique location on the human genome is typically reported to be ~80%, although this number varies depending on the read length, the sequencing protocol (*e.g.* the

[24] The RNA is transcribed from the element, then reverse transcribed into a complementary DNA,which is integrated into a new location in the genome.

availability of paired-end reads) and and the sensitivity of the aligner used (Treangen and Salzberg, 2012). However, the repeat content in the human genome is ∼50%. This discrepancy mainly depends on that most repeats are inexact, which implies that they will have a unique best match even if the same sequence occur with slight variations in other locations, as shown in panel A of Figure 8. Assigning reads to the location of their best alignment, is the simplest way to resolve repeats, although it is not always correct. For example, assume that the same read map to two locations, A and B, where the read aligns with one mismatch at A and with one deletion at B. If the aligner considers mismatches more likely to happen than deletions, then it will put the read in location A. However, if the source DNA has a true deletion, then the read would perfectly match position B. This true-to-life problem, that is inherent in the process of aligning reads to a reference genome, is also illustrated in panel B of Figure 8. Indeed, widely-used mappers are mostly based on the Needleman–Wunsch and Smith–Waterman algorithms (Nalbantoğlu, 2014) and each attributes different scores to mismatches, gap opening and gap extending, resulting in different alignments for the same sequences. Another problem comes out when a genome sample is sequenced, but only analysis of the variants that are present in a certain chromosome is required. The most straightforward approach would be to use a short-read aligner to map reads directly to that chromosome, which lead to a large pile up of reads from repetitive regions, because all reads from those repeats would have to go to the same chromosome. In order to avoid this bias, the reads must me mapped against the entire genome and a strategy of random placement of multi-reads to scatter them uniformly across all repeat copies must be applied.

Essentially, aligners have three choices for dealing with multi-reads: (1) ignore multi-reads, meaning that all multi-reads are discarded. This strategy is usually achieved by applying specific filters during the alignment step (*e.g.* by setting the *ambiguous* parameter to *toss* on BBMap (Bushnell, 2014)) or by post-processing the aligned *Sequence Alignment/Map* (SAM)/*Binary Alignment Map* (BAM)[25] file (*e.g.* by filtering on the mapping quality for a SAM/BAM generated with Bowtie2 (Langmead and Salzberg,

[25] `https://samtools.github.io/hts-specs/SAMv1.pdf`.

2012) or by retaining only reads with the *XT:A:U* tag from a SAM/BAM generated with BWA (Li and Durbin, 2009)). This strategy limits analysis to unique regions in the genome, discarding many multigene families as well as all repeats, which might result in biologically important variants being missed. (2) retain the the alignment with the fewest mismatches, *i.e.* the best match. If there are multiple, equally good matches, then an aligner will either choose one at random or report all of them. By default, most aligners use a pseudo-random number generator to choose which read to retain in a set of equally-good choices but filters can be applied to report all reads in the set (*e.g.* by using the *-a* parameter in Bowtie2). This approach is the only one that can provide a reasonable estimate of coverage. (3) report all alignments up to a maximum number, regardless of the total number of alignments found (*e.g.* by setting an upper limit to the *-k* parameter in Bowtie2). Allowing multi-reads to map to all possible positions avoids making a possibly erroneous choice about read placement.
Overall, choosing what alignment strategy to use is of fundamental importance, as it influences downstream tools for variant discovery.

1.4.2 *De Novo* Genome Assembly

Genome assembly algorithms attempt to reconstruct a genome as completely as possible exploiting starting from a set of sequenced reads. As explained in the previos sections, short reads from SGS make assembly extremely difficult in repetitive regions. Indeed, repeats that are longer than the read length create gaps in the assembly and, as a result, genome assemblies based on SGS are much more fragmented than assemblies based on Sanger sequencing (Schatz et al., 2010). In addition to creating gaps, repeats can be erroneously collapsed on top of one another, causing complex misassemblies[26]. However, assemblers that use short reads are available, which are based either on string-overlap graphs (*e.g.* SAGE (Ilie et al., 2014)) or De Bruijn graph (*e.g.* ABySS (Jackman et al., 2017)). Assemblers from both groups create graphs

[26] Assembled regions that contain significantly large variations that are the result of wrong decisions made by the assembly program. These errors can be easily misconstrued as true genetic variation, misleading a range of genomic analyses (Muggli et al., 2015).

from the reads and then traverse these graphs in order to reconstruct the original genome. From a technical perspective, repeats cause branches in these graphs, and assemblers must then make a guess as to which branch to follow, with incorrect guesses creating false joins and erroneous copy numbers. Conservative assemblers break the assembly at branch points, leading to more accurate but fragmented contigs. Some problems assemblers can run into are summarized in Figure 9. A common error is the creation of a rearrangement by joining two chromosomal regions that do not belong near one another (Figure 9, panel A). Even if all the reads align well to the misassembled genome, mate-pair constraints are violated (*i.e.* wrong expected distance and orientation of the paired-end reads). Other common issues are the creation of collapsed repeats, where read alignments remain consistent but mate-pair distances are compressed (Figure 9, panel B), and the creation of collapsed interspersed repeats (Figure 9, panel C). In addition to using mate-pair information from reads that were sequenced in pairs, assemblers exploits statistics on the depth of coverage, which is useful to identify the repeats themselves. Assuming that a genome is uniformly covered, repetitive regions have substantially deeper coverage, which allows the assemblers to identify and process them differently. In particular, repeats are usually assembled after unique regions, and assemblers may require multiple paired-end reads to link a repetitive contig to a unique one.

1.4.3 Short-Read Tandem Repeat Callers

Standard variant-calling pipelines for genome resequencing (*e.g.* Pindel (Ye et al., 2009)) and *de novo* assembly projects (*e.g.* FermiKit (Li, 2012)) classify alterations in repeated regions either as indels or SVs, depending on their size. However, specialized tools that can profile TRs (*i.e.* identify their motif and multiplicity), have been developed over years, mainly for STRs.
Most STR callers require a previous knowledge of the STR *loci* to look for. Defined STR *loci* are available for each release of the human genome[27] and are based on calls from *Tandem Re-*

[27] `http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/hg19.trf.bed.gz`, for the GRCh37 release

peats Finder (TRF) (Benson, 1999). TRF is a robust tool that can detect repeats with pattern size in the range from 1 to 2000 bases from FASTA[28] inputs. The program first uses a set of statistically based criteria to find candidate TRs, then attempts to produce an alignment for each candidate and, if successful, gathers a number of statistics about the alignment (*e.g.* matching probability and indel probability) and the nucleotide sequence (*e.g.* base composition and sequence entropy). TRF is based on Bernoulli distribution. In particular, the tool models alignment of two tandem copies of a pattern of length n by a sequence of n independent Bernoulli trials (*i.e.* coin tosses). Each head in the Bernoulli sequence is interpreted as a match between aligned nucleotides and each tail is a mismatch, an insertion or a deletion. The matching probability represents the average percent identity between the copies, while the indel probability specifies the average percentage of insertions and deletions.
One of the first successful STR profiler for SGS was LobSTR (Gymrek et al., 2012). The algorithm of LobSTR has three steps: (1) scan genomic libraries, flag informative reads that fully encompass known STR *loci*, and characterize their sequence. This procedure relies on a signal processing approach that uses rapid entropy measurements to find informative STR-containing reads, followed by a Fast Fourier Transform to characterize the repeat sequence. In practice, each sequenced read is break into overlapping windows of a fixed length and a fixed nucleotide overlap between consecutive windows. When a read displays a series of windows with entropy below a predefined treshold, then that read is considered informative and further processed through a Fast Fourier Transform to identify the repeat unit size, *k*. The algorithm further determines the actual STR sequence by means of a rolling hash function that records all possible k-mers in the STR region: the most frequently occurring k-mer is set to be the repeat unit of the STR. (2) alignment. The aim of the alignment step is to reveal the identity and the repeat length of a STR-containing read. To this purpose, LobSTR employs a divide-and-conquer approach. It separately anchors the upstream and downstream flanking regions of STR-containing sequence reads, without map-

[28] `https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=BlastHelp`

ping the STR region itself. This procedure identifies the genomic location of the STR and reveals the repeat length by measuring the distance between the flanking regions. (3) allelotyping. The most likely alleles at each STR *locus* are identified by integrating informations from all aligned reads and the expected stutter noise, calculated through a generative approach based on the repeat unit size.

A major drawback of the first TR profilers for SGS is that they are constrained to STR alleles with repeat lengths smaller than the read length employed in the sequencing (*i.e.* ∼150 bases). More recently, TR callers for SGS have emerged which have demonstrated the ability to detect repeat expansions where the expanded allele size is greater than the length of standard short-read sequencing reads and even the read pair fragment length. These tools include ExpansionHunter (Dolzhenko et al., 2019), STRetch (Dashnow et al., 2018), exSTRa (Tankard et al., 2018), TREDPARSE (Tang et al., 2017) and GangSTR (Mousavi et al., 2019). All require paired-end alignments and a catalog of known STRs.

ExpansionHunter extracts STR-containing reads from a given alignment file and realigns them using a graph-based model representing the STR structure at each STR *locus*. In particular, the STR structure is specified using a restricted subset of the regular expression syntax. For example, the HTT repeat region linked to Huntington Disease (Table 1) can be defined through the expression $(CAG)^*CAACAG(CCG)^*$, which means that it harbors variable numbers of the CAG and CCG repeats separated by a CAACAG interruption. Similarly, the FXN repeat region linked to the FRDA corresponds to expression $(A)^*(GAA)^*$ and the ATXN8 repeat region linked to SCA8 corresponds to $(CTA)^*(CTG)^*$. The regular expressions are allowed to contain degenerate base symbols, making it possible to represent certain classes of imperfect DNA repeats where different bases may occur at the same position. Then, ExpansionHunter translates each regular expression into a sequence graph, with nodes that correspond to sequences and directed edges that define how these sequences can be connected together to assemble different alleles. Genotyping of the sequenced individual is performed by analyzing the alignment paths associated with the presence or absence of each constituent allele.

The idea behind STRetch is to construct a modified reference genome containing STR decoy chromosomes that can be used for mapping. STR decoys are sequences that consist of 2000 bases of pure STRs that can be added to any reference genome as additional chromosomes. By mapping the sequenced reads to this modified genome, STRetch identifies all the reads that originate from large STR expansions. As explained above, most aligners have difficulties in accurately mapping reads containing long STRs, which sometimes map to other STR *loci* with the same repeat unit or completely fail to map. While reads with STR lengths similar to the allele length in the reference genome will map to their original STR, reads containing large STR expansions will preferentially align to the STR decoy chromosomes. The reads that map to the STR decoys are then assigned to genomic STR positions. To this purpose, STRetch uses the mapping position of the read at the other end of the DNA fragment to infer from which known STR each read originates. For a given read, if the mate maps within 500 bases of a known STR with the same repeat unit, then the read is assigned to that STR, or to the closest matching if multiple STRs are present. However, some reads may be unassigned in the end, which happens if their mates also map to the STR decoys, are unmapped, or do not map in close proximity to a known STR. Lastly, STRetch compares the number of STR decoy reads assigned to each STRfor a test sample with STR reads from a set of control samples, which provide a median and variance of counts for each locus. The z-score is used to test if the number of reads in the test sample is an outlier compared to control. By working on the assumption that, for a given STR, the number of reads containing the STR is proportional to the length of the repeat in the genome being sequenced, STRetch estimates the size of any detected expansion using the read counts allocated to that STR.

exSTRa is a two-step analysis method: (1) identifies all the reads that map to each STR *locus*. To this purpose, anchor reads, *i.e.* reads that map on or within 800 bases of the STR and have the same STR orientation, are retained. For each anchor read, the anchor mate is then checked and, if it is mapped near the STR or is overlapping the STR, then the pair is taken forward, otherwise is discarded. Remaining anchor-mates have their sequence content matched for the presence of the repeat unit in the correct direction, allowing for the repeat to start at any base of the repeat unit (*e.g.*

if the repeat unit is CAG, the method can also match AGC and GCA). The number of bases found to be part of the repeat unit is counted to derive a repeat-score for that read, with reads where the score is lower than expected in random nucleotide sequences being filtered out. (2) applies an empirical quantile imputation procedure to detect if the number of repeat units identified is an outlier compared to a background set of samples, as outliers are likely to be repeat expansions.

The TREDPARSE workflow involves a number of key steps: (1) determine STRs ploidy. Autosomal STRs are modeled as diploid *loci*, allowing two alleles to be inferred per STR. For STRs on the X chromosome, TREDPARSE infers the gender for the given sample by computing the median read depth on selected unique regions on the Y chromosome. If the median depth on the Y chromosome is less than 1, X chromosome ploidy is 2, 1 otherwise. (2) realignment of reads near STRs. In particulare, TREDPARSE realigns reads that are mapped within a read length from the repeat location and reads that are unmapped but have a mate mapped within a distance of 1000 bases from the repeat location. The goal of the realignment step is to obtain an accurate count of the occurrences of the repeat motifs. Most read mapping methods, when aligning reads to a reference, have a high penalty for long indels, which often results in misalignments. To accommodate long indels, TREDPARSE uses a *Single Instruction Multiple Data* (SIMD) Smith-Waterman algorithm to align STR-containing reads to a series of STR-containing reference sequences that are embedded with a variable number of repeat units. This procedure yields a series of alignments with different scores that are compared to determine the repeat size that corresponds to the highest score. During the alignment, each read is also classified as a prefix read (*i.e.* a read with a flanking sequence of length ≥ 9 to the left of the repeats) or a suffix read (*i.e.* a read with a flanking sequence of length ≥ 9 to the right of the repeats). (3) classification of the reads. On the basis of the alignment informations, reads with both prefix and suffix are classified as spanning reads, and reads with either prefix or suffix but not both are classified as partial reads. Reads that only consist of repeats are repeat-only reads. Distances of paired-end reads are also taken into account for the development of a full-probabilistic model to infer STR size. (4) deployment of a full probabilistic framework. To fully model the uncertainties of observing a set of reads that are generated by a certain repeat size, a probabilistic model is built which uses the read types mentioned above. In particular: (a) for spanning reads, as they show both left and right flanking sequences, inferring the number of repeat units is straightforward as the counted size matches or is close to the true repeat size. A model for the stutter noise based on GC content and on the repeat unit size, similar to that described for LobSTR, is included. (b) partial reads, which do not align all the way across the repeat region and contain only one flanking sequence, have a probability mass function of discrete uniform distribution between a single repeat unit and the true repeat length. Therefore, they only show a lower bound for the number of repeat units of the underlying allele. The inference task is here modeled as the *German tank problem*[29] with replacement. (c) repeat-only reads are possible only when repeat length is the same or longer than a read length. Assuming each read is equally likely to start anywhere in the genome, the expected number of repeat-only reads that fall in a certain region follows a Poisson distribution. (d) the observed distance between the two mate reads gathers additional informations. After inferring the distribution of the distances between all the paired-end reads across the genome, expanded repeats, when mapped to the reference, show a compression of paired-end distances while shortened repeats show an expansion of paired-end distances.

Similarly to TREADPARSE, GangSTR defines four classes of paired-end reads at STR *loci*: (1) enclosing read pairs, which consist of at least one read that contains the entire repeat plus non-repetitive flanking region on either end; (2) spanning read pairs, which originate from a fragment that completely spans the repeat, such that each read in the pair maps on either end of the repeat; (3) flanking read pairs, which contain a read that partially extends into the repetitive sequence of a read; (4) fully repetitive read pairs, which contain at least one read consisting entirely of the repeat motif. Each class provides informations about the length of the repeat in the region, which are integrated into a single joint likelihood framework to find the maximum likelihood diploid genotype and confidence interval on the repeat length.

[29] The problem consists of estimating the maximum of a discrete uniform distribution from sampling without replacement.

Despite the use of complex strategies, short reads from SGS technologies are still largely insufficient for profiling long repetitive DNA segments (Tørresen et al., 2019). TGS offer reads that often encompass the entire TRs and have already proven invaluable for the detection of large structural variants (Mahmoud et al., 2019). However, accurately deciphering TRs from long reads remains a considerable challenge due to their high error rates, especially in low complexity regions.

Aims And Objectives 2

In this dissertation we aim to introduce a novel computational framework capable to profile TRs in long-read alignments without a prior knowledge of their motifs or locations (*i.e. de novo*), namely *Tandem Repeats Caller for LOng Reads* (TRiCoLOR) (Bolognini et al., 2020a).

Indeed, methods for TR profiling can be broadly classified as reference-based or *de novo* approaches. The former rely on databases of known TRs and look at reads spanning these TRs to call TR alleles. Several strategies used by reference-based TR callers for SGS are described in detail in the *Introduction* chapter. The latter can identify TRs regardless of whether their repeat motif is annotated or not in the reference and are of high interest for annotating newly sequenced genomes (Girgis, 2015). So far, few TR detection methods for long-read sequencing data have been developed, all falling in the reference-based group of tools. Examples include PacmonSTR (Ummat and Bashir, 2014), *Noise Cancelling Repeats Finder* (NCRF) (Harris et al., 2019), TideHunter (Gao et al., 2019) and NanoSatellite (De Roeck et al., 2019).

PacmonSTR is specifically optimized for raw SMRT sequencing data. The tool: (1) selects uniquely mapped reads spanning known TR intervals from an alignment file generated using the BLASR aligner (Chaisson and Tesler, 2012). In the scenario that disparate alignments are found spanning a TR interval, which may occur if a TR allele is highly divergent from the reference allele at that *locus*, multiple unique alignments are merged to provide the initial query seed interval for downstream processing. (2) applies a modified Needleman–Wunsch algorithm to better identify TR boundaries and give initial TR multiplicity estimates; (3) processes the identified TR interval through a pair *Hidden Markov Model* (HMM) to give a more rigorous estimate of the TR multiplicity. Briefly, this approach computes the probability for the sum of all alignment paths between the query and a putative repetitive TR sequence. In this paradigm, the number of TR elements is a random variable and the pair HMM is used to calculate the expected value of this random variable based on an estimated discrete probability mass function. The model structure of the pair HMM takes as input the predicted TR interval and the consensus TR element sequence and models the error modes using matches, deletions and insertions as hidden states. Transition and emission probabilities are generated by BLASR alignments in non-TR regions. (4) groups reads by *locus* and creates clusters of reads based on their estimated TR multiplicities to determine zygosity.

Rather than working on alignment files, NCRF takes as inputs a set of reads in FASTA format, and a motif to look for. Then, the aligner at the core of NCRF, which is based on the Smith-Waterman algorithm with affine gap penalties, finds alignments of the motif to the given DNA sequences. The alignment core utilizes different penalties for insertions and deletions depending on the sequencing technology reads are generated from (*i.e.* ONT or PacBio): therefore, technology-specific scoring parameters are tuned to observed sequencing error profiles, with the dynamic programming recurrence being modified to support a high prevalence of short indels. As a last step, NCRF retains only the high-quality alignments by applying a consensus and an alignment filter. Indeed, in some cases an organism might contain the sought repeat motifs interleaved with other repeat motifs. These additional motifs, are discarded using the consensus filter. Moreover, when two or more similar motifs are searched for, some intervals of a read may align to more than one repeat and the alignment filter groups alignments by those unique to each motif.

Simlarly to NCRF, TideHunter works on sequences in FASTA format but uses a seed-and-chain algorithm to recognize the sought repeat pattern. Briefly, TideHunter collects seeds of long reads, which consist of hash values[1] and locations of the k-mers of these reads. The collected seeds are then sorted by both the hash value and the location, then stored in a hash table. A hit in the table for a tandem repeat is identified for each pair of seeds that have identical hash values and are adjacent to each other in the sorted table, with the distance between the hits (*i.e.* the location distance of two seeds having a tandem repeat hit) being usually close to the true repeat pattern size or its multiples. TideHunter considers all such hits as anchors and attempts to find an optimal chain of colinear anchors using dynamic programming. The optimal chain is expected to consist of anchors that have a hit distance close to the repeat pattern size. Therefore, TideHunter partitions the original long-read into multiple segments based on the optimal chain. A SIMD *Partial Order Alignment* (POA) of these segments is then applied to generate an accurate consensus sequence.

Lastly, NanoSatellite has been designed to call TRs directly on raw ONT squiggle data to circumvent errors introduced by base calling and further downstream alignment processing and is based on a *Dynamic Time Warping* (DTW) algorithm. As the name suggests, DTW is a dynamic programming algorithm, but does not work with actual strings like the Needleman-Wunsch, Smith-Waterman or POA. Rather, DTW attempts to find the optimal alignment between two time series (*e.g.* is frequently used in several pattern recognition applications, such as speech recognition). Thus, DTW can be used to compare the raw current signal generated by a ONT device with a known squiggle for a TR of interest, which can be derived by translating DNA nucleotide sequences to their estimated squiggle patterns.

All these tools have limitations, either because they are technology-specific (PacmonSTR and NanoSatellite) or because they are not intended to be used genome-wide (NCRF, TideHunter and NanoSatellite). Some tools also lack the capability to genotype

[1] A hash value is a string value generated by means of a hash function. A hash function, like the one at `https://gist.githubusercontent.com/MohamedTaha98/ccdf734f13299efb73ff0b12f7ce429f/raw/ab9593d5195a1643388cfc99d03a4fd96a094a5c/djb2%2520hash%2520function.c`, is any function that can be used to map data of arbitrary size to fixed-size values (McKenzie et al., 1990)

the TRs identified (NCRF and TideHunter), a crucial feature in clinical TR profiling applications, and, as mentioned above, none is capable to profile TRs in regions that have previously not been annotated as harboring a TR. TRiCoLOR addresses these shortcomings of existing tools by allowing users to rapidly identify and genotype any TRs from haplotype-resolved long-read alignments, whether from PacBio or ONT. Once low-entropy repetitive regions have been identified in sequenced long reads, TRiCoLOR exploits POA to compute haplotype-specific low-error consensus sequences that are further processed by means of a fast *Regular Expression* (RegEx)-based approximate string matching algorithm to resolve repeat motif and multiplicity of the discovered TRs.
TRiCoLOR's modules and methods are described in detail in the *Methods* chapter. In the *Results* chapter we illustrate the results we got by benchmarking TRiCoLOR on synthetic and real data. Further discussion of TRiCoLOR's features and limitations is given in the *Discussion* chapter.

Methods 3

TRiCoLOR is an open-source framework implemented in Python 3 with supporting C++ code and bash scripts, publicly available at `https://github.com/davidebolo1993/TRiCoLOR`. An overview of its workflow is outlined as follows while detailed informations about its modules are given in the sections below.

The *de novo* identification of repetitive regions from haplotype-resolved long-read alignments in BAM format is achieved using the *Shannon ENtropy ScanneR* (SENSoR) module. This module scans the given haplotypes by chromosome and detects regions having low Shannon entropy DNA content, where repetitive stretches cause low entropy scores. Repetitive regions for which a sufficiently low score has been detected are included in a BED file[1] that is subsequently profiled by TRiCoLOR's *REpeats FindER* (REFER) module. For each haplotype, the REFER module fetches sequencing reads spanning regions in the BED file, building low-error consensus sequences via a SIMD version of the robust POA framework for error-prone long reads. These consensus sequences are then screened by a RegEx-based string matching algorithm: first, the algorithm detects perfectly repeated motifs; then, it looks for nearby, approximate occurrences of the identified motifs; last, it solves nested TRs by means of a N-gram model. The corresponding reference segment is screened using a similar approach and TRs varying between the haplotypes or the reference are genotyped and

[1] `https://m.ensembl.org/info/website/upload/bed.html`

stored in standard VCF/BCF format[2]. The identified TRs can be interactively visualized by the *Alignment Plotter* (ApP) module and for trio sequencing studies, the TRiCoLOR *SAmple GEnotyper* (SAGE) module checks whether parental and child genotypes follow Mendelian segregation laws. This module can also be used to assign the most likely parental genotypes of the TRs identified in the child if parents have been sequenced at low depth, preventing the *de novo* identification of TRs.

3.1 Pre-processing: Phasing

The ability of TRiCoLOR to genotype TRs in diploid samples strictly relies on the *a priori* knowledge of the hapolotype that aligned reads belong to, *i.e.* their phase. Genotyping is the process of determining which genetic variants are present in an individual's genome: a genotype at a given site describes whether both chromosomal copies carry a variant allele, whether only one of them carries it, or whether the variant allele is not present at all. Phasing refers to assigning individual's haplotypes by identifying short variants (single-nucleotide variants and indels) that lie near each other on the same chromosome and are inherited together (Ebler et al., 2019). Short variants can be reliably detected using whole-genome short-read sequencing (Nielsen et al., 2011) but resolving haplotypes with such a technology has limitations because two adjacent heterozygous variants are usually not spanned by a single sequenced fragment to allow a so-called read-backed phasing procedure. Read-backed phasing assembles haplotypes using overlaps between reads that span multiple heterozygous variants but since the heterozygosity ratio of human genomes is comparatively low (Bryc et al., 2013) the average nucleotide distance of heterozgous markers exceeds the short read length. The result is that millions of bases of the reference human genome are not currently reliably genotyped by short reads, primarily in large gaps near the centromeres and short arms of chromosomes. While short reads are unable to uniquely map to these regions, long reads can span into or even across them and have already proven useful for reconstructing haplotypes (Pirola et al., 2016).
Some strategies that exploit long reads to resolve individual's haplotypes, either in combination with other sequencing technologies or alone, have been described in the recent years.
Chaisson and collegues, as part of the *Human Genome Structural Variation Consortium* (HGSVC), were able to phase 3 individuals by applying WhatsHap (Patterson et al., 2015) on Illumina paired-end

[2] `https://samtools.github.io/hts-specs/VCFv4.3.pdf`

reads, Illumina SLR and PacBio reads, StrandPhaseR (Porubský et al., 2016) on Strand-seq data and 10X linked-reads and by integrating these results with traditional trio-based and population-based phasing methods (Loh et al., 2016). By combining a dense, yet local, technology (*e.g.* PacBio or 10X linked-reads) with a chromosome-scale, yet sparse, technology (*e.g.* Illumina paired-end or Strand-seq), they were able to obtain dense and global haplotype blocks (Chaisson et al., 2019).
Guo and collegues described a method for long-read SNV calling and haplotype reconstruction which identifies an exemplar read at each SNV site that best matches nearby reads overlapping the site. Then, the method partitions reads around the site based on similarity to the exemplar at adjacent SNV sites. However, this method is not guaranteed to discover an optimal partitioning of the reads between haplotypes, with authors reporting high false-positive and false-negative discovery rates (Guo et al., 2018). Luo and collegues and Poplin and collegues described methods which uses convolutional neural networks to call variants from long-read data, which they report to achieve high precision and recall scores on PacBio data (Poplin et al., 2018; Luo et al., 2019). Excellent performances on PacBio data were also reported with Longshot, a tool that harnesses SMRT reads to jointly perform SNV detection and haplotyping. To this purpose, Longshot uses the phasing method HapCUT2 (Edge et al., 2017) and, in order to overcome the high error rate of PacBio reads, it utilizes a pair HMM to average over the uncertainty in the local alignments and estimate accurate base quality values that can be used for calculating genotype likelihoods (Edge and Bansal, 2019).
As a proof of concept, we evaluated the performances of different frameworks for generating haplotype-resolved alignments prior to TRiCoLOR. First, we exploited VISOR (Bolognini et al., 2020b), to insert phased single-nucleotide variants from the 1000 Genomes Project (HG00732 sample) on chr20 of the GRCh38 human reference genome and to simulate a final BAM file (~40X coverage) mirroring ONT data, as further described in the *Results* chapter. Then, we applied WhatsHap and LongShot to directly identify candidate single nucleotide variants from the synthetic BAM file, and phase them. We run WhatsHap's *find_ snv_ candidates* module with the *–nanopore* parameter enabled, the *genotype* module, the *phase* module and the *haplotag* module sequentially, using the default parameter settings. We run LongShot with the default parameter settings as well. WhatsHap and LongShot could assign respectively $\sim 68\%$ and $\sim 71\%$ of the synthetic ONT reads to one of the 2 haplotypes and $\sim 89\%$ of all sequenced bases because most unassigned reads are relatively short (as expected). We measured the amount of phasing inconsistencies between the phased single-nucleotide

variants from the 1000 Genomes Project and those from WhatsHap and Longshot by calculating their switch error rates (Choi et al., 2018) using vcftools (Danecek et al., 2011). The calculated switch error rate of WhatsHap was ∼ 1.2%, while the switch error rate of LongShot was ∼ 0.9%
We also evaluated the capability of HapCUT2 and WhatsHap to phase the synthetic ONT BAM file using complementary single-nucleotide variant calls generated by bcftools (Li, 2011) from a short-read alignment (∼40X coverage) simulated with VISOR for the same sample (these calls from bcftools should represent a set of variant calls one can be reasonably confident in). We run the *extractHAIRS* and *hapcut2* commands from the HapCUT2 package and the *phase* and the *haplotag* modules from WhatsHap, using the default parameter settings. Because HapCUT2 does not provide utilities to either tag or split reads by haplotype, we subsequently resolved the 2 haplotypes of the ground truth BAM file using Alfred (Rausch et al., 2019), giving the phased single-nucleotide variants from HapCUT2 as input. WhatsHap and Alfred could assign ∼ 71% of the synthetic ONT reads to one of the 2 haplotypes (∼ 92% of sequenced bases). Using this experimental setting, the calculated switch error rate of WhatsHap was ∼ 0.7% while the switch error rate of HapCut2 was ∼ 0.6%, reflecting the higher-quality of input single-nucleotide variants.
Overall the local phase accuracy of long-read phasing algorithms is high. However, long-read technologies alone are sub-optimal for chromosomal-level phasing, for which they have to be used in combination with a chromosome-scale technology such as Strand-seq or Hi-C, as for the HGSVC data that we exploited in this work. TRiCoLOR can thus be applied to read-backed phased data as well as chromosome-length haplotypes because, as clarified below, it evaluates each tandem repeat locally in a surrounding window, where results are only expected to deteriorate if a rare switch error occurred within a given tandem repeat window.

3.2 TRiCoLOR

TRiCoLOR requires haplotype-resolved long-read alignments as input. It then runs a series of modules to identify and genotype TRs. These modules are described in detail below. An on-line manual containing an in-depth explanation of how to run TRiCoLOR's various modules is also available at `https://davidebolo1993.github.io/tricolordoc/`, together with use case examples.

3.2.1 TRiCoLOR SENSoR

TRiCoLOR can spot repetitive regions in haplotype-resolved BAM files *de novo* through the SENSoR module, which exploits an approach based on the calculation of Shannon entropy (which we refer to as H), similar to the one descibred by Gymrek and collegues for their LobSTR tool.

H was originally devised by Claude Shannon as a measure for order or disorder in strings (Shannon, 1948). Intuitively, a string S with symbols S_i $(i = 0...\lambda)$ from a given alphabet is considered as ordered when it is periodical or when some symbols or substrings occur repeatedly. In constrast, it is considered disordered, when all of its symbols and combinations of symbols occur at equal frequencies. H has been formalized as

$$H = -\sum_i p_i \log p_i$$

where i extends over all symbols of the alphabet, and p_i is the probability that symbol s_i occurs at any position. H is maximal when all symbols occur at equal probability $p_i = 1/\lambda$. The minimum H (*i.e.* $H = 0$) is taken on if one symbol occurs at probability 1, with the others being absent (Schmitt and Herzel, 1997). In the SENSoR module H is calculated as

$$H = -\sum_{x \in X} p_x \log_2 p_x$$

where x is any DNA base from a DNA sequence X and p_x is the probability that $x \in X$ occurs, (*i.e.* the frequency of the DNA base in the DNA sequence). Given this formula, a fully random string results in the maximal H whereas a repetitive string "overuses" certain nucleotides which causes a low H, with perfect homopolymer runs having $H = 0$.

In order to identify the optimal H treshold (which we refer to as Ho) that makes possible to discriminate between repetitive and non repetitive regions in error-prone sequences, we first exploited our read simulator VISOR to generate 1000 synthetic long-read BAM files, with half of the BAM files modelling current sequencing error rates from ONT and the other half from PacBio. The average sequencing error rates and the substitution:insertion:deletion ratios used to generate synthetic ONT and PacBio reads were derived from publicly available datasets that are described in the subsections below.

In particular, we simulated ∼8000 bases-long reads at ∼10X coverage in small regions (∼20000 bases) around known, randomly chosen, TRs from the TR catalog of the GRCh38 human reference genome. We then computed H for all the aligned reads in non-overlapping, sliding windows of 20 bases. Given that the mean length of the TRs from the TR catalog used is ∼40 bases, we chose a window size of 20 bases as this allows to have at least one window encompassed by the TRs picked during the simulations. We empirically set the Ho to be the 2$^{\text{nd}}$ percentile of the H distribution.

Figure 10 shows the negatively-skewed H distributions for the simulated ONT (Figure 10, panel A) and PacBio (Figure 10, panel B) BAM files. For both the technologies, we identified $Ho \sim 1.23$, which allowed to exclude $\sim 98\%$ of the entire alignment information screened, in accordance with the results presented by Gymrek and collegues for their LobSTR tool. Figure 10 also illustrates the read-specific H in non-overlapping, sliding windows of 20 bases for a simulated ONT (Figure 10, panel C) and PacBio (Figure 10, panel D) BAM file including the TR ranging from 48941985 and 48942028 on chr19 of the GRCh38 human reference genome. All the simulated reads have at least one window where H is below the Ho in the region containing the TR, which is highlighted in green, while H is confirmed above the Ho threshold for the other regions.

The entropy-based scanning of aligned reads is provided with TRiCoLOR's SENSoR module. For a given haplotype-resolved long-read alignment, this module scans in parallel the 2 haplotype-specific BAM files and computes, for each sequencing read, its H in non-overlapping, sliding windows of 20 bases, which is the size trained in our simulations. Genomic coordinates of windows in which multiple reads (*i.e.* ≥ 5, by default) support an H drop (*i.e.*, $H \leq 1.23$) are stored and those nearby are merged (*i.e.* those falling within 100 bases intervals, by default). Repetitive regions identified with this approach are eventually outputted in BED format. Since just a few simple calculations are required to retain the informative (*i.e.* repetitive) regions even from massive, whole-genome sequencing data, this module is fairly fast, as further described in the *Result* chapter, and drastically reduces the computational time required by the REFER module, which is described in the subsection below.

3.2.2 TRiCoLOR REFER

TRiCoLOR can profile TRs in haplotype-resolved BAM files through the REFER module, once regions to investigate are provided in proper BED format. This BED file can be generated through the SENSoR module described in the subsection above or can be provided by uses based on prior knowledge of clinically relevant TRs, for instance. For each region in the BED file, TRiCoLOR REFER applies the following strategy.

First, the module fetches from the haplotype-specific BAM files the sequencing reads spanning the selected region and trims them, so that the length of each read is approximately the size of the region. Let $R = [S, E]$ be a region from the BED file, ranging from a start coordinate S to an end coordinate E for a given chromosome. Each sequencing read entirely spanning R is fetched and trimmed so that the actual sequence REFER temporarily stores in FASTA format is that included between S and E, which significantly improves the runtime of the subsequent POA algorithm to generate a consensus sequence. Once the sequencing reads of interests have been fetched and trimmed, TRiCoLOR exploits SPOA, a SIMD version of the robust POA framework, to generate highly-accurate consensus sequences from error-prone long reads (Vaser et al., 2017). Over the past years, several error correction methods have been developed to reduce the sequencing error of long reads. These approaches can be roughly classified into hybrid (*i.e.* involving the use of short reads) (Goodwin et al., 2015) and self, or non-hybrid (*i.e.* using only long reads) (Salmela et al., 2017). Although these methods provide a better per-base accuracy than the raw data, hybrid approaches can bring systematic errors from short reads in long-read data sets and non-hybrid approaches heavily relies on the available coverage. As an alternative approach, one can reconstruct a high-quality template from uncorrected reads *in silico*, by means of a *Multiple Sequence Alignment* (MSA). To this puprose, POA, which is described in detail in two papers from Lee and collegues (Lee et al., 2002; Lee, 2003), performs MSA through a *Directed Acyclic Graph* (DAG), where nodes are individual bases of input sequences, and weighted, directed edges represent whether two bases are neighboring in any of the sequences.

MSA is one of the most important tools in bioinformatics and can be helpful in many circumstances like detecting relations between sequences: in many cases, sequences that undergo MSA are assumed to have an evolutionary relationship, by which they share a linkage and are descended from a common ancestor. MSA is also used to compute a consensus profile for sequences that originate from the same region. A variety of heuristic MSA algorithms exist based on progressive application of pairwise sequence alignment to build up alignments of larger numbers of sequences. For pairwise sequence alignment, a globally optimal solution can be found in $O(L^2)$ time by dynamic programming, where L is the length of the two sequences being aligned. This algorithm can be extended to align N sequences optimally, but requires $O(L^{\mathrm{N}})$ time, with the exponential time required for aligning larger numbers of sequences by dynamic programming being impractical. Therefore, excellent MSA algorithms like Clustal (Thompson et al., 1994), first align all of the sequences pairwise, which results in $N(N-1)/2$ alignments. Then, the scores of these alignments are then used to construct a binary tree of their relationships. Finally, the algorithm builds a MSA in the order dictated by the evolutionary tree: the least diverged sequences are aligned first, resulting in $N/2$ alignment profiles; the $N/2$ alignment profiles are aligned to each other resulting in $N/4$ alignment profiles; and so forth, until all of the sequences have been aligned. Finding the scores and constructing the binary evolutionary tree is $O(NL^2)$ while using the binary tree to build the MSA is $O(L^2logN)$ and can be run in a reasonable amount of time. MSA strategies based on progressive pairwise alignments, however, suffer some major issues: (1) they may find a local minimum either because the guide tree is not correct or because alignment errors that happen early on in the process of building the MSA get locked in; (2) the choice of appropriate alignment parameters, which can cause problems in handling of gaps and insertions; (3) progressive MSA requires aligning pairs of MSAs, to build up larger MSAs. In practice pairwise dynamic programming is not applied directly to align the pairs of MSAs. Instead, progressive alignment relies on reducing each MSA to a 1D-sequence which can be used in pairwise dynamic programming sequence alignment. This reduction of an MSA to a consensus profile inevitably involves loss of information as, while the MSA contains all the information to produce the profile, the profile does not contain all the information needed to reconstruct the original MSA. Artifacts from progressive MSA can be solved by representing the alignment between sequences as a partially ordered graph in which individual sequence letters are represented by nodes, and directed edges are drawn between consecutive letters in each sequence. In POA, a single sequence is simply a linear series of nodes each connected by a single incoming edge and a single outgoing edge: the letters that are aligned and identical are fused as a single node, while the letters that are aligned but not identical are represented as separate nodes that are recorded as being aligned to each other. When letters are fused as one node, the resulting node stores information about all of the individual sequence letters from which it was derived, and their index, making it possible to trace the path of each individual sequence through the alignment. Standard dynamic programming sequence alignment can be extended to work with partial orders, as shown in Figure 11. Standard dynamic programming alignment of two linear sequences can be represented as a 2D matrix, whose two axes correspond to the two sequences. A given point (n, m) in the matrix corresponds to a pair of sequence positions. For a given pair, three basic moves are possible: a diagonal 'alignment' move indicating that n and m are aligned; and horizontal and vertical moves indicating, respectively, n as an insertion relative to m, or m as an insertion relative to n. The set of all possible paths across the 2D matrix constructed from these moves represents all possible alignments of the two sequences allowing the matches or mismatches (*i.e.* diagonal moves) and insertions or deletions (*i.e.* horizontal or vertical moves) (Figure 11, panel A). In POA, one of the linear sequences is replaced by a partial order containing branching, with the 2D matrix bifurcating (*i.e.* generating a new surface) each time a single sequence DAG can align to either sequence stored in the POA. On a given surface, the POA behaves the same as the standard 2D alignment, and the same set of three moves (*i.e.* diagonal, horizontal, vertical) are allowed. At junctions where multiple surfaces fuse, the horizontal and diagonal moves are extended to allow them to go onto any of the incoming surfaces that meet at the junction. Thus for the simplest case where two branches join, the allowed moves are: two diagonal moves (*i.e.* one from each incoming surface), two horizontal moves (*i.e.* one from each incoming surface), only one vertical move, and a start move (Figure 11, panel B). Consensus sequences are obtained from a built POA graph by performing a topological sort and processing the nodes from left to right. Overall, POA has linear time complexity in the number of sequences but most implementations are prohibitively slow for larger data sets. Thus, we integrated into TRiCoLOR SPOA, a SIMD-accelerated POA algorithm, which is inspired by the Rognes and Seeberg Smith-Waterman intra-set parallelization approach (Rognes and Seeberg, 2000) and drastically increases the speed of calculation over non-SIMD versions.

We evaluated the capability of SPOA to reduce the error rate of long-read alignments from Chaisson and collegues. In particular, we applied Alfred to haplotype-resolved BAM files and calculated the error rate of the HG00733, HG00514 and NA19240 individuals, sequenced using platforms from ONT as well as from PacBio. We then applied SPOA to generate consensus sequences from a region on chr20 ranging from 18000000 to 20000000 of the HG00733 individual, using windows of 2000 bases (*i.e.* 1000 windows in total); the consensus sequences formed were aligned to the GRCh38 human reference genome using the long-read aligner minimap2 (Li, 2018) and we derived their error rate as for the original BAM file. For the consensus generation, we exploited the global mode of SPOA with default penalties (*i.e.* matches: +5; mismatches: -4; gap opening: -8; gap extending: -6), which resulted in the lowest consensus error rates. Figure 12 shows the error profiles of the ONT Figure 12, panel A) and PacBio (Figure 12, panel B) alignments for the HG00733, HG00514 and NA19240 individuals, which are also haplotype-resolved (*e.g.* HG00733 is haplotype-resolved in HG00733.h1 and HG00733.h2). For ONT alignments, the mean error rate is ∼11% and the substitution:insertion:deletion ratio is ∼45:25:30; for the PacBio alignments, the mean error rate is ∼13% and the substitution:insertion:deletion ratio is ∼15:50:35. Figure 12 also illustrates how much of the initial error rate of the ONT (Figure 12, panel C) and PacBio (Figure 12, panel D) alignments SPOA can correct. For the regions investigated, the mean error rate of the ONT (∼2.5%, substitution:insertion:deletion ratio ∼31:37:32) and PacBio (∼1.5%, substitution:insertion:deletion ratio ∼17:70:13) consensus sequences is drastically lower compared to the inital error rates derived from the BAM files, *i.e.* ∼11% and ∼13% respectively.

Having the haplotype-specific consensus sequences formed, these are aligned to the reference genome to retrieve their coordinates using the minimap2 aligner, which we found to compare favorably to another widely-used aligner, namely NGMLR (Sedlazeck et al., 2018).

Indeed, in order to identify the best aligner for long sequences, we evaluated the performances of minimap2 and NGMLR on syntethic data. We run minimap2 using the presets *-x map-ont* for the ONT simulations and *-x map-pb* for the PacBio simulations; we run NGMLR using the presets *-x ont* for the ONT simulations and *-x pb* for the PacBio simulations. First, we evaluated the speed of the chosen aligners when aligning an increasing number of long sequences (coverage ∼1X, ∼5x, ∼10X, ∼15X, ∼20x, ∼25X, ∼30X; substitution:insertion:deletion ratio ∼10:60:30; average length of reads ∼8000 bases; accuracy of reads ∼0.90), simulated from a region on chr20 (32000000-62000000) of the GRCh38 human reference genome using PBSIM (Ono et al., 2013). For each simulation (one for each coverage level), we repeated the alignment step 5 times using 6 Intel®Xeon®processors X5460 on an Ubuntu 16.04.6 LTS desktop. As illustrated in Figure 13, minimap2 proved to be ∼6 times faster than NGMLR. Results are shown as mean ± standard deviation (Figure 13, panel A). We further evaluated the accuracy of the chosen aligners when mapping long sequences of increasing length (∼500 bases, ∼1000 bases, ∼5000 bases, ∼10000 bases) and increasing accuracy (∼0.85, ∼0.90, ∼0.95), simulated from the same region on chr20 (32000000-62000000) of the GRCh38 human reference genome with PBSIM. Figure 13 also shows these findings for simulated ONT (substitution:insertion:deletion ratio ∼45:25:30) (Figure 13, panel B) and PacBio (substitution:insertion:deletion ratio ∼15:50:35) (Figure 13, panel C) alignments respectively. We used the ratio between the number of reads mapped in the region chosen for simulating and the total number of reads mapped as a measure of accuracy. The accuracy of both minimap2 and NGMLR increases as the length of the simulated reads increases, approaching ∼1.0 when the length of these reads is $\geq$ 5000. For shorter reads, minimap2 demonstrates an accuracy higher than NGMLR on the ONT simulations and slightly lower on the PacBio simulations. These results led us to choose minimap2 as the default aligner for TRiCoLOR, as it outperformed NGMLR in terms of speed without loosing the comparison in terms of mapping accuracy. We furthermore investigated which preset of the minimap2 aligner performed best for mapping the consensus sequences generated through SPOA (see also Note S3) to the reference genome. Using the simulation schema described above, we generated synthtetic ONT and PacBio alignments from a region on chr1 (100000000-110000000) of the GRCh38 human reference genome, and we exploited SPOA to generate consensus sequences using windows of 1000 bps (10000 windows in total). We aligned the consensus sequences formed back to the chr1 reference sequence using the presets for noisy reads (*i.e. -x map-ont* for the ONT alignments and *-x map-pb* for the PacBio alignments) as well as the presets for the assembly-to-reference alignment (*i.e. -x asm5*, *-x asm10*, *-x asm20*). For all the presets used, minimap2 was able to properly map all the generated consensus sequences to their original location (*e.g.* a consensus sequence generated from the 11th window was properly mapped to the original region chr1:100010000-100011000). Given that the presets evaluated did not influence the mapping accuracy of minimap2, we decided to call the aligner from within TRiCoLOR using the presets for noisy reads.

The reference-aligned low-error consensus sequences are then screened by a RegEx-based approximate string matching algorithm to identify TRs.

A RegEx is a pattern in which the rules for matching text are written in form of metacharacters, quantifiers or plain text. For instance, the well-known RegEx metacharacter called Kleene Star (*), that derives its name from the American mathematician Stephen Cole Kleene who invented the RegEx strategy, means to “match the preceding character zero to many times”. RegEx expressions are an efficient and hence popular way to search for repeats of a certain size and a large number of patterns (Merkel and Gemmell, 2008) and a variety of TR callers that implements a RegEx-based search engine have been released in the past, including MsatFinder[3], SSRIT (Temnykh et al., 2001) and MISA (Thiel et al., 2003).

The RegEx-based string matching algorithm of TRiCoLOR has three processing steps: (1) identifying motifs (motifs of length $\leq$ 6 bases, by default) that are perfectly repeated a minimum number of times (5, by default). A simplified version of the Python function used by TRiCoLOR to find perfect repeated motifs in consensus strings is shown below.

[3] `http://web.archive.org/web/20071026090642/http://www.genomics.ceh.ac.uk/msatfinder/`

```
import re
def regexfinder(consensus,motif=6,size=5,overlapping=False):
  seen=set()
  if not overlapping:
    regex=r'(.+?)\1{'+str(size-1)+r',}'
  else:
    regex=r'(?=(.+?)\1{'+str(size-1)+r',})'
  r=re.compile(regex)
  for match in r.finditer(consensus):
    m=match.group(1)
    if len(m) <= motif:
        seen.add(m)
  return seen
```

Specifically, the first part (*(.+?)*), which is also known as capturing group, matches any character, except for line terminators, between one and unlimited times, as few times as possible, expanding only when needed. This expression, which has a lazy behaviour, allows to identify the shortest repeated motif, whatever its length is. The second part (*\1{5,}*) matches the same text as most recently matched by the previous capturing group between 5 (the default value) and unlimited times, as many times as possible, giving back as needed. This expression, which has a greedy behaviour, allows to identify the longest repeated stretch of the motif chosen by the capturing group. By default, the algorithm does not look for overlapping repetitions (*i.e.* it keeps only one between a repeated AT and a repeated TA) but this behaviour can be changed by enabling the *–overlapping* parameter, which expands the previous RegEx by using the positive lookahead structure *(?=...)*. Lookahead and lookbehind structures, collectively known as lookaraound structures, are zero-length assertions, that match characters but then give up the match, returning only the result (*i.e.* match or no match), without consuming characters. In the original

TRiCoLOR function, when looking for a known repeated motif, the RegEx built can be further adjusted to identify only motifs of a predefined length. (2) looking for approximate repetitions of the motifs identified. Consensus sequences are quite accurate but still contain sparse errors interrupting TRs: therefore, the algorithm allows up to a certain number of insertions, deletions or mismatches (*i.e.*, a certain edit distance) between repeated motifs (maximum edit distance 1, by default). (3) where appropriate, resolving overlapping approximate repetitions by way of a N-gram model.

The N-grams are strings containing N words: for instance, a bigram is a two-word sequence of words and a trigram is a three-word sequence of words. N-gram models estimate the probability of a word given some history, *i.e.* knowing all the words preceding. However, instead of computing the probability of a word given its entire history, one can approximate the history by just the last few words. The bigram model, in particular, approximates the probability of a word given all the previous words by using only the conditional probability of the word preceding. The assumption that the probability of a word depends only on the previous word is called a Markov assumption. Markov models are the class of probabilistic models that assume we can predict the probability of some future unit without looking too far into the past. This concept is implemented in TRiCoLOR REFER, which estimates how many times in the consensus string the repetitive motifs that are found to overlap are preceded by theirselves. In practice, when multiple approximately repeated motifs are found to overlap, the N-gram model we use favors the motif (*i.e.* the N-gram) that more frequently is found to repeat itself perfectly.

Together with the haplotype-specific consensus sequences, the corresponding reference is screened in a similar manner, with few differences being noteworthies: (1) the algorithm assumes the reference does not contain errors and does not look for approximate repetitions of the motifs identified; (2) among overlapping repetitions, the longest repeat is taken and no N-gram model is required.

TRs (those $\geq$ 50 bases, by default) varying between the haplotypes or the reference are eventually stored in BCF-compliant format. TRiCoLOR REFER also stores in the output folder several BED files describing the TRs identified (both for the reference and each haplotype) and haplotype-specific BAM files containing the aligned consensus sequences. These additional files can be given to the dedicated module for interactive visualization of the identified TRs, which is described below.

3.2.3 TRiCoLOR ApP

The TRs profiled using TRiCoLOR REFER can be interactively visualized through the ApP module. This module takes as inputs the BED and the BAM files generated by TRiCoLOR REFER together with an additional BED file describing one ore more regions to plot. For each region, TRiCoLOR ApP produces a static HTML page based on the plotly graphing library[4], illustrating the alignment between the reference and the individual's haplotypes at single base resolution and highlighting the TRs found.

As a proof of concept, in Figures 14A to 14C we showed how to browse the HTML file generated by TRiCoLOR ApP for a TR (dinucleotide TG repeated 15 times) on chromosome 17 (64240234-64240263) of the human GRCh38 reference genome for which we simulated a small expansion (dinucleotide TG repeated 22 times) on haplotype 1 using VISOR.

Figure 14A shows the home screen of the HTML. Top left buttons allow users to highlight repetitions found in the reference and in the individuals' haplotypes. In addition to the TR of interest, we found in the browsed region a stretch of 28 As (64239876-64239903), which is also highlighted in the screenshot. To further resolve a TR, users can zoom into a certain area (*e.g.*, the area we selected with the red rectangle).

Figure 14B shows the sequences of the reference and the individual's haplotypes at higher magnification. Each dot corresponds to a single nucleotide and users can check the base composition of each molecule by simply scrolling across the alignment. Deletions of one or more bases in the haplotypes can be seen as gaps in their sequences while insertions are represented by multiple dots sharing the same coordinate (thus, closer than expected). We highlighted with a red rectangle the insertion of TG bases found on haplotype 1.

Figure 14C shows a further magnification of the extended TG repetition (TG repeated 7 times more, the area we selected with the red rectangle) that we simulated on haplotype 1. The inserted TGs share the same coordinate and are missing in the reference sequence.

3.2.4 TRiCoLOR SAGE

Detection of genotyping errors is a necessary step to minimize false results in genetic analysis, and is particularly important when the rate of genotyping errors is high, as has been reported for high-throughput sequence data (Zhi et al., 2012). In pedigrees, assigned genotypes can

[4] `https://plotly.com`

be either Mendelian consistent or Mendelian inconsistent. A Mendelian inconsistent genotyping error is an error that is detected because the observed genotypes are not consistent with the transmission pattern as specified by the Mendel's First Law (Cheung et al., 2014). When a marker is flagged as Mendelian inconsistent this marker most likely has either a genotyping error or a *de novo* mutation.

TRiCoLOR allows users to check for genotype consistency in the TRs profiled from a child when haplotype-resolved long-read alignments for both parents are also available but, for instance, have been sequenced at low depth, which prevents the *de novo* identification of TRs. This is achieved through the SAGE module.

Using the same approach described for TRiCoLOR REFER, for each parent TRiCoLOR SAGE forms haplotype-specific consensus alignments around the TRs listed in the BCF file produced for the child. Then, the module checks whether the parental TRs are more similar (*i.e.*, have a lower edit distance) to the reference or to the TR identified in the child and assigns them the most likely genotype. Knowing the genotype of both parents, the module eventually flags each TR as Mendelian consistent or inconsistent with the *–mendel* parameter enabled.

For instance, assuming that the module is tracing the Mendelian inheritance of a heterozygous TR expansion (*i.e.*, its genotype is 0|1) in the child, if both haplotypes of the first parent do not contain the expansion (*i.e.*, the TR genotype is 0|0) and one haplotype of the second parent does contain the expansion (*i.e.*, the TR genotype is 0|1), then the child TR is considered Mendelian consistent. On the contrary, if both haplotypes of the first parent do not contain the expansion (*i.e.*, the TR genotype is 0|0) and neither do the haplotypes of the second parent (*i.e.*, the TR genotype is 0|0), then the child TR is considered Mendelian inconsistent.

TRiCoLOR SAGE stores results in a multi-sample BCF file that contains the genotypes for the index child and his or her parents.

Results 4

We benchmarked TRiCoLOR using both synthetic data generated with VISOR and real, publicly available data from the HGSVC. A detailed explanation of our benchmarking procedure is available in the sections below. Briefly, we applied VISOR to generate synthetic BAM files containing contractions and expansions of known TRs and we evaluated the capability of TRiCoLOR to correctly predict the simulated repeat motif and multiplicity. TRiCoLOR shows excellent performances in all the simulations we have run and consistently outperformed NCFR. For real data, we applied TRiCoLOR to discover and genotype TRs *de novo* on three individuals from the HGSVC, namely HG00733, HG00514 and NA19240[1]. We verified the generated calls using compressed data structures, built on high-quality Illumina sequences, achieving on average ∼84% validation rates. We also estimated a ∼80% Mendelian consistency rate of TR genotypes for the HGSVC trio HG00731, HG00732 and HG0733. Among the Mendelian consistent calls generated by TRiCoLOR for the HG00733 individual, we identified 32 long TRs that were missing in the corresponding HGSVC callset. For these TRs, we manually inspected the assembly to identify the cause of these apparent discrepancies. Most TRs were properly assembled but not called. However, ∼25% of the inspected alleles were not correctly resolved by the HGSVC, which suggests that mapping-based and assembly-based approaches can be complementary for TR detection using long reads.

[1] `https://www.internationalgenome.org/data-portal/data-collection/structural-variation`

4.1 TRiCoLOR on synthetic data

We used our read simulator VISOR to generate synthetic ONT and PacBio alignments exhibiting contractions and expansions of TRs.
First, we simulated haplotype-resolved ONT and PacBio BAM files (the average length of simulated reads was set to 8000 bases based on statistics derived from recent ONT sequencing runs; the substitution:insertion:deletion ratio was set to ∼45:25:30 for the synthetic ONT reads and to ∼15:50:35 for the synthetic PacBio reads, in accordance with findings in the *Methods* chapter) exhibiting variable error rates (accuracy of reads ∼0.85, ∼0.90 and ∼0.95) and depth of coverage (haplotype-specific depth of coverage 5X-10X and 10X-20X), with each BAM file harboring a heterozygous contraction or expansion of a known, randomly chosen, TR from the TR catalog of the GRCh38 human reference genome. At this stage, we simulated small TR contractions/expansions (*i.e.* 7 motifs on average), in order to evaluate the capability of our method to spot even minor changes in the TR multiplicity of the 2 haplotypes. For each group, we simulated 200 haplotype-resolved BAM files. Then, we evaluated the performances of TRiCoLOR in terms of *Precision* (P), *Recall* (R) and *F1 score* (F1).
In the classification task, P is the proportion of positive identifications that is actually correct, R is the proportion of actual positives that is identified correctly and F1 is the harmonic mean of the P and R. P, R and F1 are defined as:

$$P = \frac{TP}{TP + FP}$$

$$R = \frac{TP}{TP + FN}$$

$$F1 = 2 * \frac{P * R}{P + R}$$

A *True Positive* (TP) is a call for which the number of repetitions identified by TRiCoLOR matches the number of repetitions in the haplotype containing the TR contraction/expansion.
A *False Positive* (FP) is a call for which the number of repetitions identified by TRiCoLOR does not match the number of repetitions in the unaltered haplotype.
A *False Negative* (FN) is a call for which the number of repetitions identified by TRiCoLOR does not match the number of repetitions in the haplotype containing the TR contraction/expansion.
In particular, P, R and F1 values were calculated allowing no discrepancies, 1 discrepancy or 2 discrepancies between the number of repeated

motifs in the ground truth and the number of repeated motifs predicted by TRiCoLOR. Figure 15 shows these findings for synthetic TR contractions (Figure 15, panel A) and expansions (Figure 15, panel B). TRiCoLOR demonstrated high P and R in all the simulated groups: our method always achieved an F1 close to 1 when allowing a single-motif discrepancy between simulated and predicted TRs and hit P ~1 and R ~1 when allowing up to 2 motif discrepancies. For both contractions and expansions the F1 depends on the coverage and input read accuracy as expected. In all the simulated TR contractions and expansions, TRiCoLOR was also able to properly identify the correct repeated motif, few times shifted (*e.g.*, a repeated TG instead of a repeated GT). Figure 16 illustrates these findings for the same simulated groups of Figure 15, averaged over the different accuracy levels.
Furthermore, as a proof of concept, we compared TRiCoLOR to a TR caller for long reads recently published, namely NCRF. Using the same approach described above, we simulated 100 ONT and 100 PacBio BAM files (accuracy of reads ~0.90, depth of coverage for each haplotype 5X-10X) harboring small TR contractions/expansions and we run both TRiCoLOR and NCRF on these data. As NCRF cannot deal with BAM input, we slightly modified TRiCoLOR to store in FASTA format the sequences used for the consensus computation step, which could be processed through NCRF. Specifically, we run TRiCoLOR REFER with the *-m* parameter set to the length of the repeated motif (*e.g.*, *-m 2* for a GT repetition), the *–precisemotif* parameter enabled and the *–readstype* parameter set to *ONT* for ONT simulations and to *PB* for the PacBio simulations. We run NCRF using the authors' README suggestions[2], with *–scoring* parameter set to *nanopore* for ONT simulations and to *pacbio* for PacBio simulations; we adjusted the *–minlength* parameter accordingly to the length of the TRs simulated and we averaged the number of repetitions found by NCRF in each FASTA. Figure 17 shows the correlation results between the number of TRs in the ground truth and the number of TRs predicted by TRiCoLOR and NCRF for the simulated TR contractions (Figure 17, panel A) and expansions (Figure 17, panel B). For both TR contractions and expansions, TRiCoLOR got excellent *Pearson Correlation Coefficient* (Pearson's R) scores (R = 0.97 for contractions and R = 0.86 for expansions), outperforming NCRF (R = 0.87 for contractions and R = 0.74 for expansions). We next evaluated exceptionally long TR expansions because these have been implicated in several neurological disorders. For instance, the common Fragile-X Syndrome is related to a CGG-repeat usually consisting of $\leq$55 repeated motifs that expands

[2] `https://github.com/makovalab-psu/NoiseCancellingRepeatFinder/tree/master/tutorial`

to $\geq$ 200 repeated motifs. Following the simulation schema described above, we generated 100 ONT and 100 PacBio synthetic BAM files harboring TRs expanded by 200 motifs and we run both TRiCoLOR and NCRF on these data. Figure 18 shows the correlation results between the number of TRs in the ground truth and the number of TRs predicted by TRiCoLOR and NCRF for the simulated long TR expansions. As above, TRiCoLOR achieved the best R score (R = 0.73), outperforming NCRF (R = 0.53).

4.2 TRiCoLOR on real data

We applied TRiCoLOR to call TRs *de novo* on publicly available ONT and PacBio human whole-genome sequencing data from the HGSVC project, which have applied several genomics assays to three trios of individuals from the 1000 Genomes Project: the Yoruban trio (*i.e.* NA19238, NA19239 and NA19240); the Puerto Rican trio (*i.e.* HG00731, HG00732 and HG00733) and the Southern Han Chinese trio (*i.e.* HG00512, HG00513, HG00514). In particular, we used the ONT sequencing data for the HG00514 (Han Chinese, son), HG00733 (Puerto Rican, son) and NA19240 (Yoruban Nigerian, son) and the PacBio sequencing data for HG00731 (Puerto Rican, father), HG00732 (Puerto Rican, mother) and HG00733.

We aligned the ONT FASTQ files of the three sons[3] to the human GRCh38 reference genome using minimap2 and we merged the chromosome-specific PacBio alignments[4] of the Puerto Rican trio using samtools (Li et al., 2009). We then split the ONT and PacBio alignments by haplotype with Alfred, using phased single-nucleotide variants from the HGSVC project[5]. We calculated the coverage of the initial and the haplotype-resolved BAM files using mosdepth (Pedersen and Quinlan, 2018). For all the ONT samples, we identified an initial ~20X coverage (HG00733 ~21X, HG00514 ~23X and NA19240 ~24X), slightly reduced after splitting by haplotype due to some unassigned reads (HG00733 ~8X, HG00514 ~9X and NA19240 ~10X for each haplotype), which are likely to originate from autozygous regions and low mappability regions such as segmentally duplicated and heterochromatic regions (Porubsky et al., 2019). For the PacBio samples, we identified a ~42X coverage

[3] `http://ftp.ebi.ac.uk/1000g/ftp/data_collections/hgsv_sv_discovery/working/20181210_ONT_rebasecalled`

[4] `http://ftp.ebi.ac.uk/1000g/ftp/data_collections/hgsv_sv_discovery/working/20180102_pacbio_blasr_reheader`

[5] `http://ftp.ebi.ac.uk/1000g/ftp/data_collections/hgsv_sv_discovery/working/20170323_Strand-seq_phased_FB%2BGATK_VCFs`

for HG00733 and ∼21X coverage for HG00731 and HG00732, reduced after splitting the data by haplotype (HG00733 ∼14X, HG00731 and HG00732 ∼8X for each haplotype).

We then run TRiCoLOR SENSoR using the default parameter settings on the HG00733 (ONT and PacBio), HG00514 and NA19240 individuals. Using an Ubuntu 16.04.6 LTS desktop with Intel®Xeon®processors X5460, the module took ∼4 hours to scan the ONT samples and ∼8 hours to scan the PacBio sample, which reflects the higher coverage available for PacBio. For the HG00733, HG00514 and NA19240 ONT individuals the module identified ∼160000, ∼190000 and ∼260000 low-entropy regions, which were reduced to ∼70000, ∼100000 and ∼160000 respectively after filtering for regions with average coverage > 8. For the HG00733 PacBio individual the module identified ∼380000 low-entropy regions, which were reduced to ∼150000, after filtering for regions with average coverage > 10. For HG00733, ∼97% of the low-entropy regions originally identified in the ONT individual overlapped those in the PacBio one; due to the different coverage distributions, this percentage was reduced to ∼31% after filtering.

We run TRiCoLOR REFER on the samples processed by TRiCoLOR SENSoR using the default parameter settings. With 7 processors on our Ubuntu desktop, the module took ∼10-12 hours to profile TRs on the ONT individuals and ∼14 hours to profile TRs on the PacBio individual. We calculated the number of TRs properly called by TRiCoLOR using an alignment-free validation approach, as this does not require a pre-existent ground truth generated by either a mapping-based or an assembly-based TR caller. Indeed, available TR callsets are mostly based on short reads and for the reasons explained in the *Introduction* chapter, these are often inaccurate. Based on the idea from Dolle and colleagues (Dolle et al., 2017), we built searchable *Full-text indexes in Minute space* (FM-indexes)(Ferragina and Manzini, 2000) both for the GRCh38 human reference genome FASTA file and the high-quality Illumina FASTQ files of the HG00733, HG00514 and NA19240 individuals[6].

Given a pattern P and a text T, FM-indexes support the counting (*i.e.* the number of occurrences of P in T) and locating (*i.e.* the occurrence positions of P in T) operations. FM-indexes were first proposed to emulate classical *Suffix Arrays* (SA). However, while providing similar searching functionalities, FM-indexes require less space than SA, as they are based on the Burrows Wheeler Transform data structure (Burrows and Wheeler, 1994). A major benefit of FM-indexes, when compared to other reference-free approaches (*e.g.* those based on de

[6] `http://ftp.ebi.ac.uk/1000g/ftp/data_collections/hgsv_sv_discovery/illumina_wgs.sequence.index`

Bruijn graphs), is that the construction process does not constrain the length of possible queries.

For each individuals' variant identified by TRiCoLOR REFER, the FM-index-based validation algorithm: (1) checks if the variant sequence appears at any position in the reference FM-index: if so, using the consensus BAM files stored by TRiCoLOR REFER, the variant sequence is extended by 1 base to the left and 1 base to the right and step 1 is repeated; if not, the variant sequence is unique and the algorithm proceeds to the next step; (2) checks if the unique variant appears at any position in the corresponding Illumina FM-index: if so, the variant is considered a valid call; if not, the variant is considered an invalid call. Taking into account possible errors both in the consensus sequences generated by TRiCoLOR and in the Illumina sequences, we counted as valid calls also unique variants that are found in the Illumina FM indexes with up to 2 bases discrepancies (*i.e.* their edit distance is ≤ 2). Limited by the length of the available Illumina sequences, using this approach we could not validate variant TRs longer than 124 bases. As shown in Figure 19, we got high validation ratios (*i.e.* ratios between the valid calls and the number of calls that could be assessed using short reads): ∼82% for HG00733 (ONT and PacBio), ∼85% for HG00514 and ∼86% for NA19240.

We eventually run TRiCoLOR SAGE on the Puerto Rican PacBio trio HG00731, HG00732 and HG00733, with the default parameter settings and the *–mendel* parameter enabled to check the Mendelian consistency of the TRs identified in HG00733. With 7 processors on our Ubuntu desktop, the module took ∼2 hours to complete the analysis. Filtering for variants differing from the reference for at least 10 bases and for multi-allelic variants differing from each other by the same distance, we identified ∼80% of Mendelian consistent TRs, which is low compared to trio-based single-nucleotide variant and indels Mendelian consistency rates, but above reported genotype agreement rates for SVs in repetitive regions, such as inversions mediated by inverted repeats (Giner-Delgado et al., 2019).

Among the Mendelian consistent calls generated by TRiCoLOR on the HG00733 PacBio individual, we identified 32 long TRs (≥ 150 bases) that were absent in the HGSVC ground truth for the same individual[7]. In order to identify the cause of these apparent discrepancies, we aligned the HG00733 phased contigs from HGSVC[8] to the GRCh38 human

[7] `http://ftp.ebi.ac.uk/1000g/ftp/data_collections/hgsv_sv_discovery/working/20180627_PanTechnologyIntegrationSet/HG00733.merged_nonredundant.vcf`

[8] `http://ftp.ebi.ac.uk/1000g/ftp/data_collections/hgsv_sv_discovery/working/20180227_PhasedSVGenomes`

reference genome with minimap2, using the assembly-to-reference alignment mode (-*x asm5*) and the parameters suggested by QUAST-LG (*–mask-level 0.9 –min-occ 200 -g 2500 –score-N 2*) (Mikheenko et al., 2018) and we manually inspected the discordant TRs in the aligned contigs using IGV (Thorvaldsdóttir et al., 2013). As shown in Table 3, out of 58 non-reference TR alleles identified by TRiCoLOR, we could visually confirm 42 (∼75%) of them in the assembly, which means that both TRiCoLOR and the HGSVC predicted the same variant type (*i.e.* deletion or insertion) and the predicted variant size is roughly similar (*i.e.* the difference does not exceed 50 bases). However, for the other 16 variants (∼25%), the assembly either did not contain the allele predicted by TRiCoLOR (*i.e.* the variant type is discordant or the predicted variant size differs more than 50 bases) or did not cover the investigated region, which suggests that mapping-based and assembly-based approaches can be complementary for TR detection using long reads.

Discussion 5

In this dissertation, we described TRiCoLOR, a tool capable to resolve TR motifs and their multiplicity in long read sequencing data sets. Long reads are in theory ideally suited to investigate TRs, which often cannot be adequately profiled with short reads if TRs exceed the sequencing reads in length. However, the high error profiles of ONT and PacBio reads complicate repeat resolution.

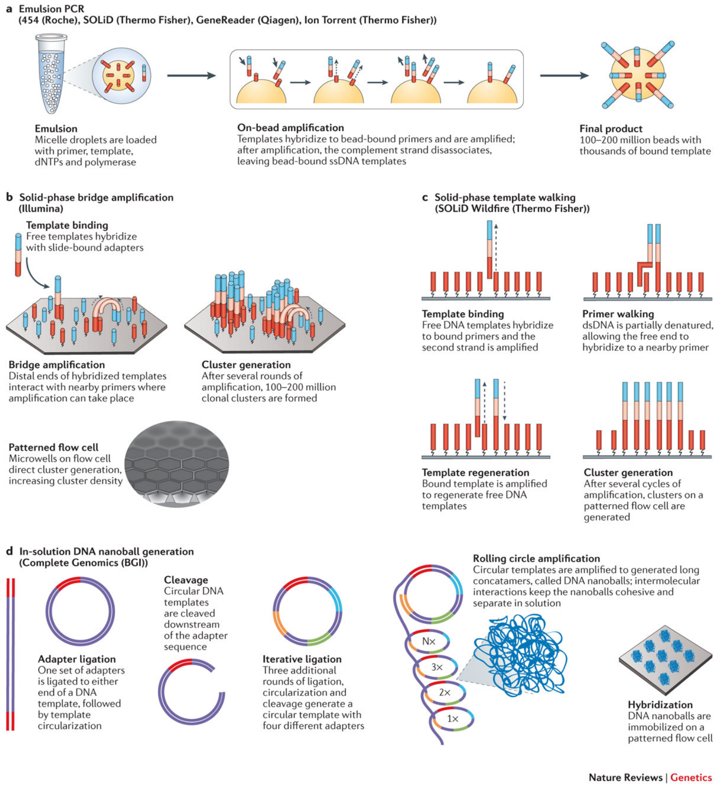
TRiCoLOR is a comprehensive TR caller for long reads that supports the *de novo* identification of TRs in whole-genome sequencing data. TRiCoLOR profiles TRs through an efficient POA algorithm combined with a RegEx-based string matching search, facilitating a robust and accurate discovery of the full spectrum of expanded and contracted TRs in personal genomes.

In comparison to previous tools, TRiCoLOR works with ONT and PacBio data seamlessly. TRiCoLOR also identifies TRs *de novo* and does not require *a priori* knowledge of annotated TR regions. The unique combination of features for genome-wide, *de novo* discovery and genotyping of TRs in ONT and PacBio data is to the best of our knowledge unmet by any other TR caller for long-read data. Besides the detection of TRs, TRiCoLOR visualizes TRs in their haplotype context and it can infer parental genotypes using low-coverage parental sequencing data.

TRiCoLOR has been designed for diploid organisms and future work includes extending its feature set to polyploid species and haploid chromosomes (e.g.human Y chromosome). As a mapping-based approach, TRiCoLOR cannot identify repeats in unassembled regions of the genome (*e.g.*, human centromeres and telomeres). Furthermore, the entropy threshold and window size for the *de novo* identification of repetitive stretches that we empirically estimated is well-suited for short repeated motifs (2-3 bps) but may need adjustments for long motifs of higher nucleotide complexity. Lastly, by default TRiCoLOR profiles TRs with motif lengths $\leq$ 6 bps (also known as micro-satellites), excluding those with motif lengths $\geq$7 bps (also known as mini-satellites), which are less abundant in diploid organisms (Richard et al., 2008). The RegEx algorithm can be also tuned to profile mini-satellites (*i.e.*, by extending the *–size* parameter) but TRiCoLOR has been extensively applied so far only to micro-satellites.

Given these limitations, future work will focus on extending TRiCoLOR to other ploidies, broadening the size spectrum of detectable repeat motif lengths and taking advantage of improved sequencing read accuracy (*e.g.*, high-fidelity long reads from PacBio). The latter directly improves the RegEx-based identification of repeats employed by TRiCoLOR and we thus believe TRiCoLOR is well-suited to characterize the TR landscape in present and future long-read data sets, making it an instrumental tool to robustly decipher the multiplicity of TRs in repeat-mediated clinical disorders.

Figures 6



a Emulsion PCR
(454 (Roche), SOLiD (Thermo Fisher), GeneReader (Qiagen), Ion Torrent (Thermo Fisher))

Emulsion
Micelle droplets are loaded with primer, template, dNTPs and polymerase

On-bead amplification
Templates hybridize to bead-bound primers and are amplified; after amplification, the complement strand disassociates, leaving bead-bound ssDNA templates

Final product
100–200 million beads with thousands of bound template

b Solid-phase bridge amplification
(Illumina)

Template binding
Free templates hybridize with slide-bound adapters

Bridge amplification
Distal ends of hybridized templates interact with nearby primers where amplification can take place

Cluster generation
After several rounds of amplification, 100–200 million clonal clusters are formed

Patterned flow cell
Microwells on flow cell direct cluster generation, increasing cluster density

c Solid-phase template walking
(SOLiD Wildfire (Thermo Fisher))

Template binding
Free DNA templates hybridize to bound primers and the second strand is amplified

Primer walking
dsDNA is partially denatured, allowing the free end to hybridize to a nearby primer

Template regeneration
Bound template is amplified to regenerate free DNA templates

Cluster generation
After several cycles of amplification, clusters on a patterned flow cell are generated

d In-solution DNA nanoball generation
(Complete Genomics (BGI))

Adapter ligation
One set of adapters is ligated to either end of a DNA template, followed by template circularization

Cleavage
Circular DNA templates are cleaved downstream of the adapter sequence

Iterative ligation
Three additional rounds of ligation, circularization and cleavage generate a circular template with four different adapters

Rolling circle amplification
Circular templates are amplified to generated long concatamers, called DNA nanoballs; intermolecular interactions keep the nanoballs cohesive and separate in solution

Hybridization
DNA nanoballs are immobilized on a patterned flow cell

Figure 1: **Template amplification strategies**. Different strategies used to generate clonal DNA template populations: bead-based generation (**a**), solid-state generation (**b**,**c**), DNA nanoball generation (**d**). Figure is taken from *Coming of age: ten years of next-generation sequencing technologies* (Goodwin et al., 2016).

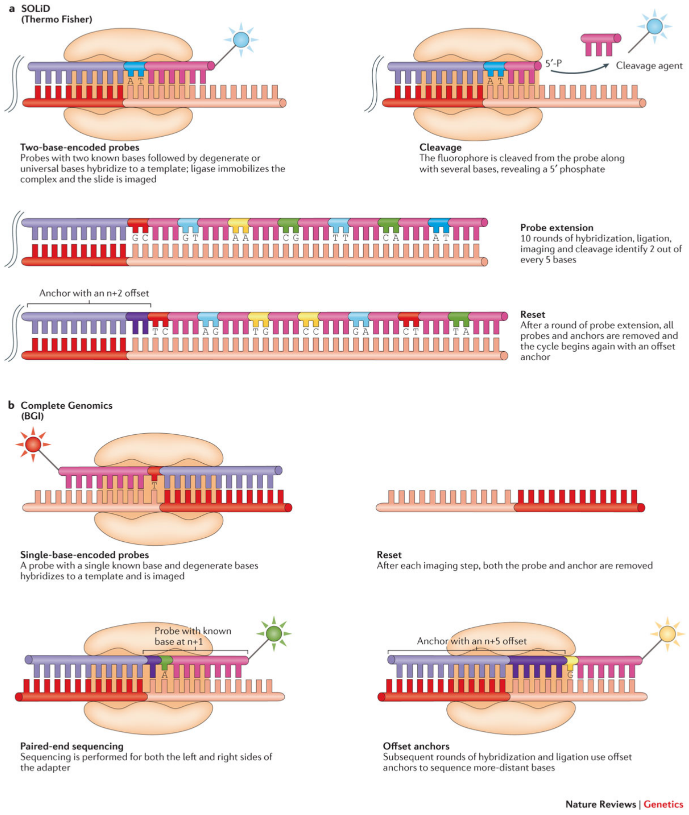
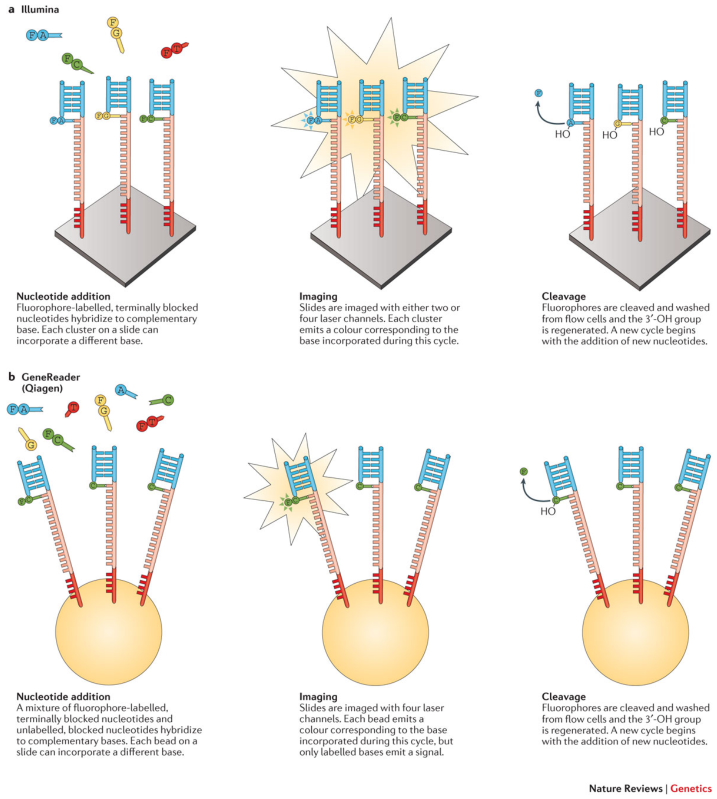

a SOLiD
(Thermo Fisher)
A T
5′-P
Cleavage agent
Two-base-encoded probes
Probes with two known bases followed by degenerate or universal bases hybridize to a template; ligase immobilizes the complex and the slide is imaged
Cleavage
The fluorophore is cleaved from the probe along with several bases, revealing a 5′ phosphate
G C
G T
A A
C G
T T
C A
A T
Probe extension
10 rounds of hybridization, ligation, imaging and cleavage identify 2 out of every 5 bases
Anchor with an n+2 offset
T C
A G
T G
C C
G A
C T
T A
Reset
After a round of probe extension, all probes and anchors are removed and the cycle begins again with an offset anchor
b Complete Genomics
(BGI)
T
Single-base-encoded probes
A probe with a single known base and degenerate bases hybridizes to a template and is imaged
Reset
After each imaging step, both the probe and anchor are removed
Probe with known base at n+1
A
Paired-end sequencing
Sequencing is performed for both the left and right sides of the adapter
Anchor with an n+5 offset
G
Offset anchors
Subsequent rounds of hybridization and ligation use offset anchors to sequence more-distant bases
Nature Reviews | Genetics

Figure 2: `SBL methods`. Summary of the SBL approaches by SOLiD (**a**) and Complete Genomics (**b**). Figure is taken from *Coming of age: ten years of next-generation sequencing technologies* (Goodwin et al., 2016).



Nucleotide addition
Fluorophore-labelled, terminally blocked nucleotides hybridize to complementary base. Each cluster on a slide can incorporate a different base.

Imaging
Slides are imaged with either two or four laser channels. Each cluster emits a colour corresponding to the base incorporated during this cycle.

Cleavage
Fluorophores are cleaved and washed from flow cells and the 3′-OH group is regenerated. A new cycle begins with the addition of new nucleotides.

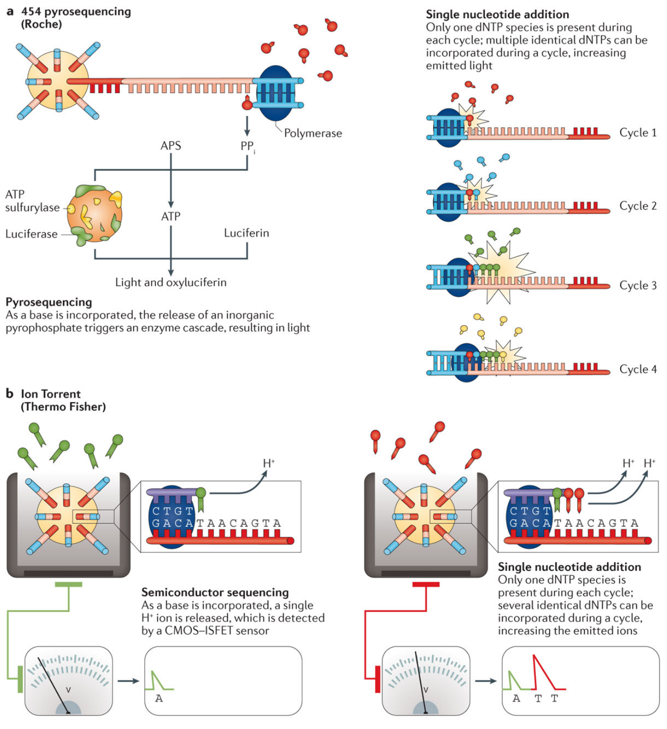
Nucleotide addition
A mixture of fluorophore-labelled, terminally blocked nucleotides and unlabelled, blocked nucleotides hybridize to complementary bases. Each bead on a slide can incorporate a different base.

Imaging
Slides are imaged with four laser channels. Each bead emits a colour corresponding to the base incorporated during this cycle, but only labelled bases emit a signal.

Cleavage
Fluorophores are cleaved and washed from flow cells and the 3′-OH group is regenerated. A new cycle begins with the addition of new nucleotides.

Nature Reviews | Genetics

Figure 3: `SBS methods: CRT approaches`. Summary of the CRT approaches by Illumina (**a**) and Qiagen (**b**). Figure is taken from *Coming of age: ten years of next-generation sequencing technologies* (Goodwin et al., 2016).



Pyrosequencing
As a base is incorporated, the release of an inorganic pyrophosphate triggers an enzyme cascade, resulting in light

Single nucleotide addition
Only one dNTP species is present during each cycle; multiple identical dNTPs can be incorporated during a cycle, increasing emitted light

Semiconductor sequencing
As a base is incorporated, a single H$^+$ ion is released, which is detected by a CMOS–ISFET sensor

Single nucleotide addition
Only one dNTP species is present during each cycle; several identical dNTPs can be incorporated during a cycle, increasing the emitted ions

Nature Reviews | Genetics

Figure 4: `SBS methods: SNA approaches`. Summary of the SNA approaches by Roche (**a**) and Thermo Fisher Scientific (**b**). Figure is taken from *Coming of age: ten years of next-generation sequencing technologies* (Goodwin et al., 2016).

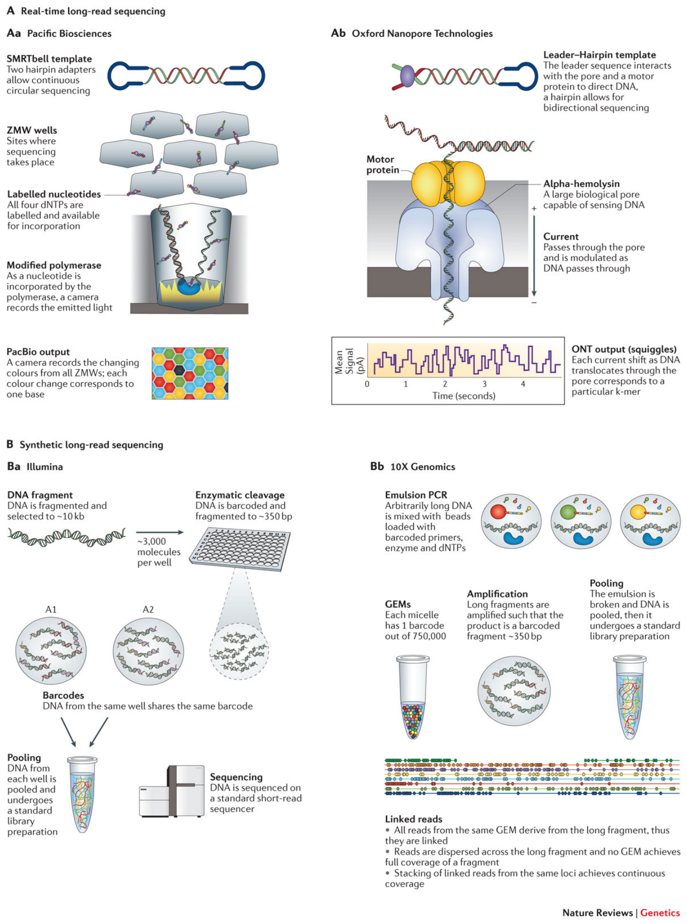

A Real-time long-read sequencing
Aa Pacific Biosciences
SMRTbell template
Two hairpin adapters allow continuous circular sequencing
ZMW wells
Sites where sequencing takes place
Labelled nucleotides
All four dNTPs are labelled and available for incorporation
Modified polymerase
As a nucleotide is incorporated by the polymerase, a camera records the emitted light
PacBio output
A camera records the changing colours from all ZMWs; each colour change corresponds to one base
Ab Oxford Nanopore Technologies
Leader–Hairpin template
The leader sequence interacts with the pore and a motor protein to direct DNA, a hairpin allows for bidirectional sequencing
Motor protein
Alpha-hemolysin
A large biological pore capable of sensing DNA
+
Current
Passes through the pore and is modulated as DNA passes through
−
Mean Signal (pA)
0
1
2
3
4
Time (seconds)
ONT output (squiggles)
Each current shift as DNA translocates through the pore corresponds to a particular k-mer
B Synthetic long-read sequencing
Ba Illumina
DNA fragment
DNA is fragmented and selected to ~10 kb
~3,000 molecules per well
Enzymatic cleavage
DNA is barcoded and fragmented to ~350 bp
A1
A2
Barcodes
DNA from the same well shares the same barcode
Pooling
DNA from each well is pooled and undergoes a standard library preparation
Sequencing
DNA is sequenced on a standard short-read sequencer
Bb 10X Genomics
Emulsion PCR
Arbitrarily long DNA is mixed with beads loaded with barcoded primers, enzyme and dNTPs
GEMs
Each micelle has 1 barcode out of 750,000
Amplification
Long fragments are amplified such that the product is a barcoded fragment ~350 bp
Pooling
The emulsion is broken and DNA is pooled, then it undergoes a standard library preparation
Linked reads
• All reads from the same GEM derive from the long fragment, thus they are linked
• Reads are dispersed across the long fragment and no GEM achieves full coverage of a fragment
• Stacking of linked reads from the same loci achieves continuous coverage
Nature Reviews | Genetics

Figure 5: Long-read sequencing approaches. Different strategies used to generate long reads: SMRT sequencing by PacBio (**Aa**), nanopore sequencing by ONT (**Ab**), Synthetic long-read sequencing by Illumina (**Ba**) and 10X Genomics (**Bb**). Figure is taken from *Coming of age: ten years of next-generation sequencing technologies* (Goodwin et al., 2016).

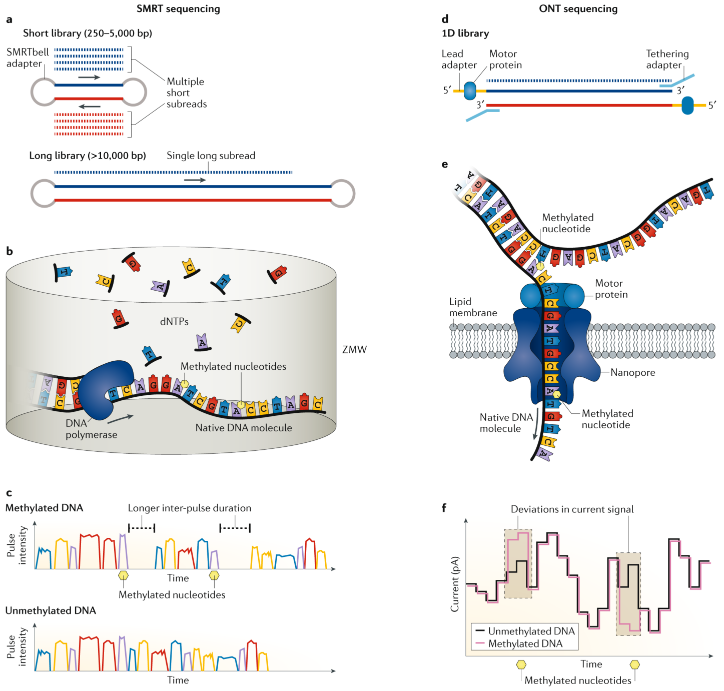
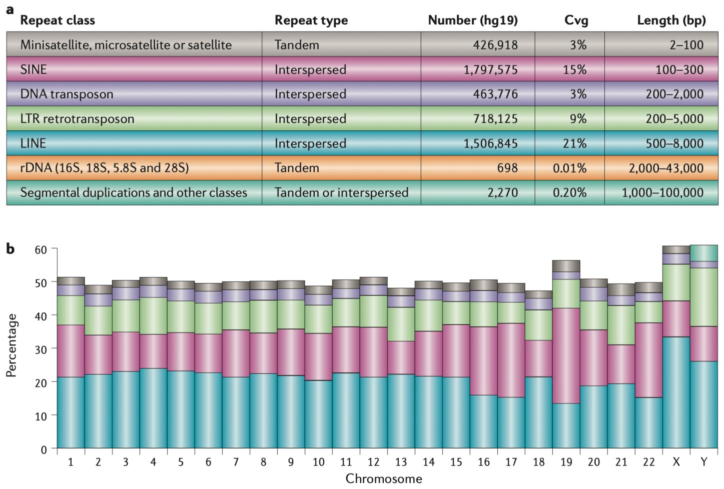


Figure 6: `Detection of base modifications with SMS`. Different strategies used to identify nucleotides epigenetically mofified using SMS: SMRT sequencing by PacBio (**a,b,c**) and nanopore sequencing by ONT (**d,e,f**). Figure is taken from *Deciphering bacterial epigenomes using modern sequencing technologies* (Beaulaurier et al., 2019).



a

Repeat class	Repeat type	Number (hg19)	Cvg	Length (bp)
Minisatellite, microsatellite or satellite	Tandem	426,918	3%	2–100
SINE	Interspersed	1,797,575	15%	100–300
DNA transposon	Interspersed	463,776	3%	200–2,000
LTR retrotransposon	Interspersed	718,125	9%	200–5,000
LINE	Interspersed	1,506,845	21%	500–8,000
rDNA (16S, 18S, 5.8S and 28S)	Tandem	698	0.01%	2,000–43,000
Segmental duplications and other classes	Tandem or interspersed	2,270	0.20%	1,000–100,000

Figure 7: **Repetitive DNA in the human genome**. Named classes of repeats in the human genome, along with their pattern of occurrence, the percentage of the genome that is covered by each repeat class and the approximate upper and lower bounds on their lengths (**a**). The percentage of each chromosome covered by each repeat class is also shown (**b**). Data are based on the RepeatMasker annotation on release hg19 of the human genome (`http://www.repeatmasker.org`). Figure is taken from *Repetitive DNA and next-generation sequencing: computational challenges and solutions* (Treangen and Salzberg, 2012).

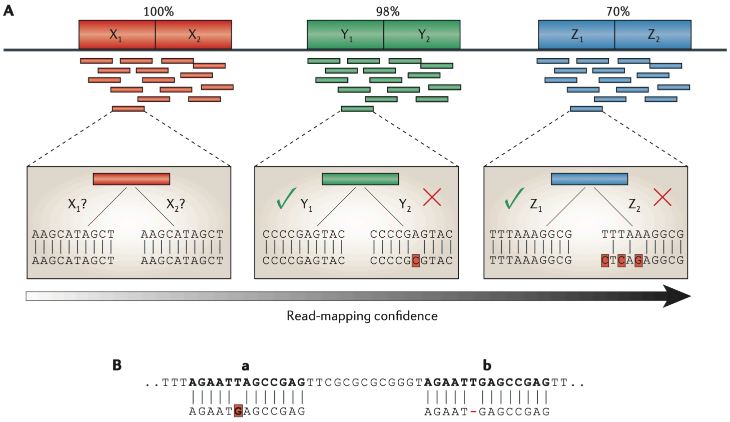


Figure 8: `Ambiguities in read mapping`. As the difference between two copies of a repeat increases, the confidence in any read placement within the repeat increases as well (**A**). When a read maps equally well to two different locations, this is assigned to either the first or the second depending on the score given by the aligner to mismatches and gaps (**B**). Figure is taken from *Repetitive DNA and next-generation sequencing: computational challenges and solutions* (Treangen and Salzberg, 2012).

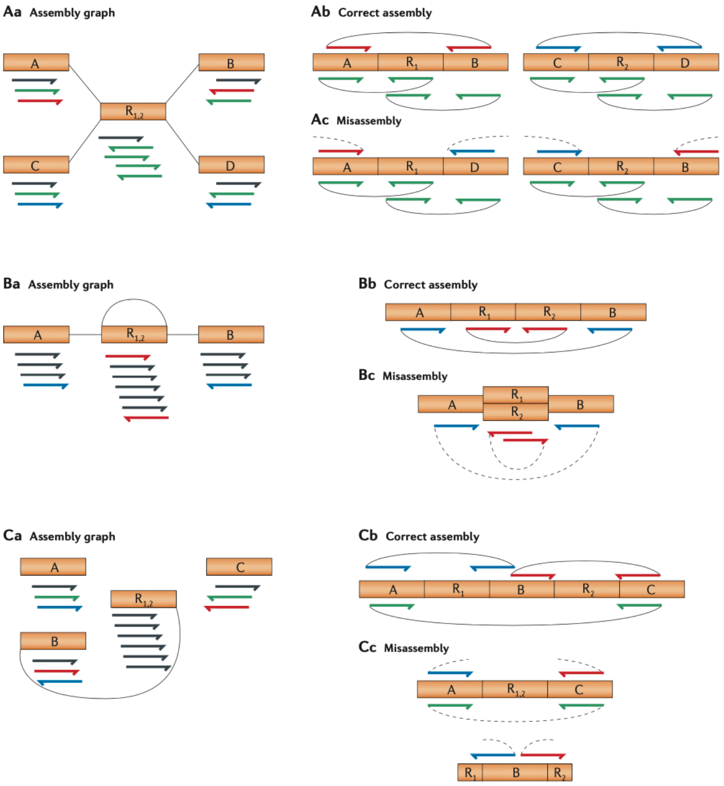


FIGURE 9: `Assembly errors caused by repeats`. Different assembly errors caused by repeats: a rearrangement error (**A**), a collapsed repeat (**B**) and a collapsed interspersed repeat (**C**). Figure is taken from *Repetitive DNA and next-generation sequencing: computational challenges and solutions* (Treangen and Salzberg, 2012).

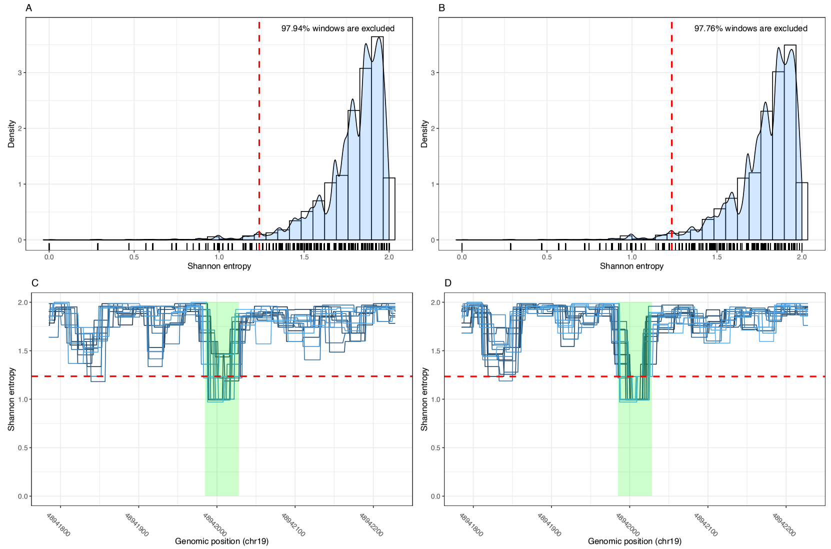


Figure 10: `Shannon entropy of tandem repeats`. Negatively- skewed distribution of Shannon entropy in simulated ONT (**A**) and PacBio (**B**) BAM files. A Shannon entropy value of ~ 1.23 allows to exclude $\sim 98\%$ of the alignment informations screened. Read-specific Shannon entropy in simulated ONT (**C**) and PacBio (**D**) BAM files. All the reads show a Shannon entropy drop below the ~ 1.23 threshold in the repeated region. Figure is taken from *TRiCoLOR: tandem repeat profiling using whole-genome long-read sequencing data*(Bolognini et al., 2020a).

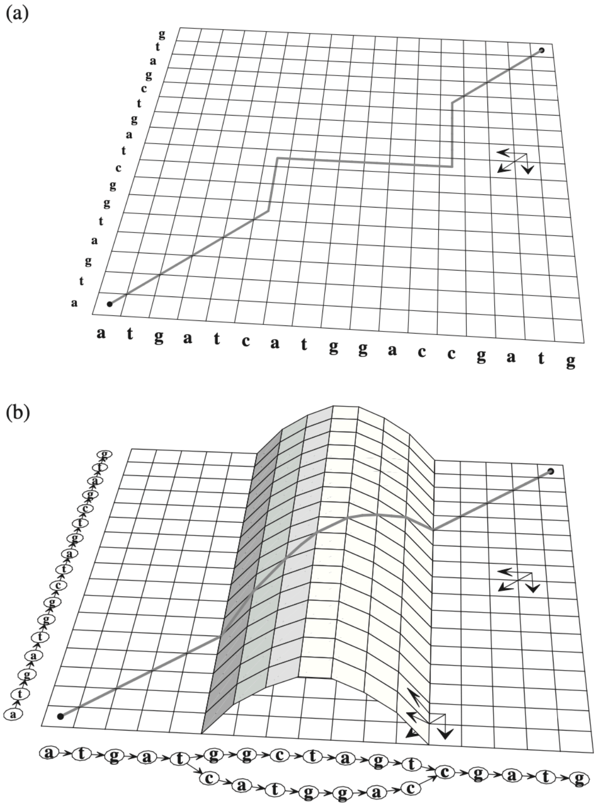

(a)
g t a g c t g a t c g g t a g t a
a t g a t c a t g g a c c g a t g
(b)

FIGURE 11: `Dynamic programming matrix for partial orders`. Dynamic programming matrix for Needleman–Wunsch sequence alignment algorithm (**a**) and for the POA algorithm (**b**), with the optimal alignment paths shown. Figure is taken from *Multiple sequence alignment using partial order graphs* (Lee et al., 2002).

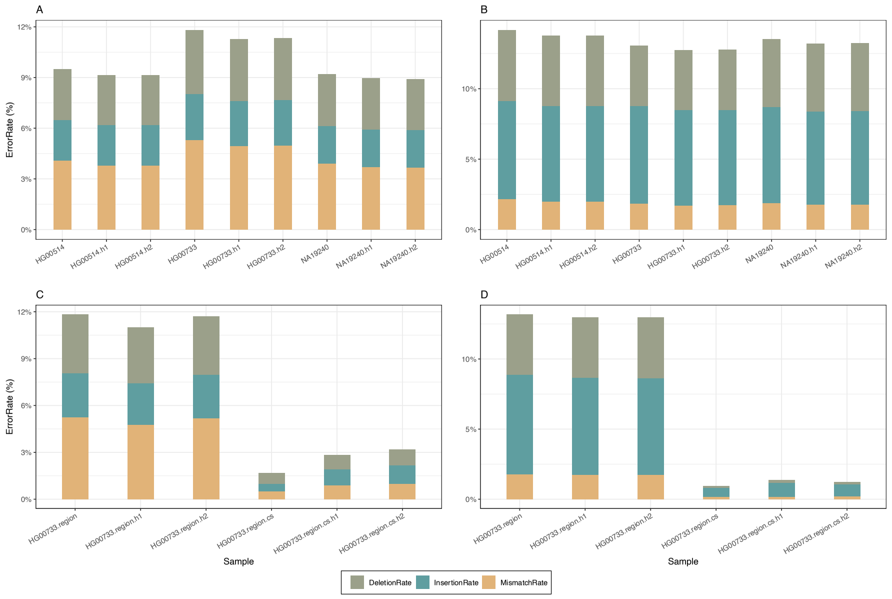


Figure 12: `Error profiles of ONT and PacBio reads`. Error profiles of real ONT (**A**, **C**) and PacBio (**B**, **D**) alignments from the HG00733, HG00514 and NA19240 HGSVC individuals. SPOA can correct most of the initial error rates of the original alignments (**C**, **D**). Haplotype-resolved alignments for each individual are indicated with the "h1" or "h2" suffixes. Figure is taken from *TRiCoLOR: tandem repeat profiling using whole-genome long-read sequencing data*(Bolognini et al., 2020a).

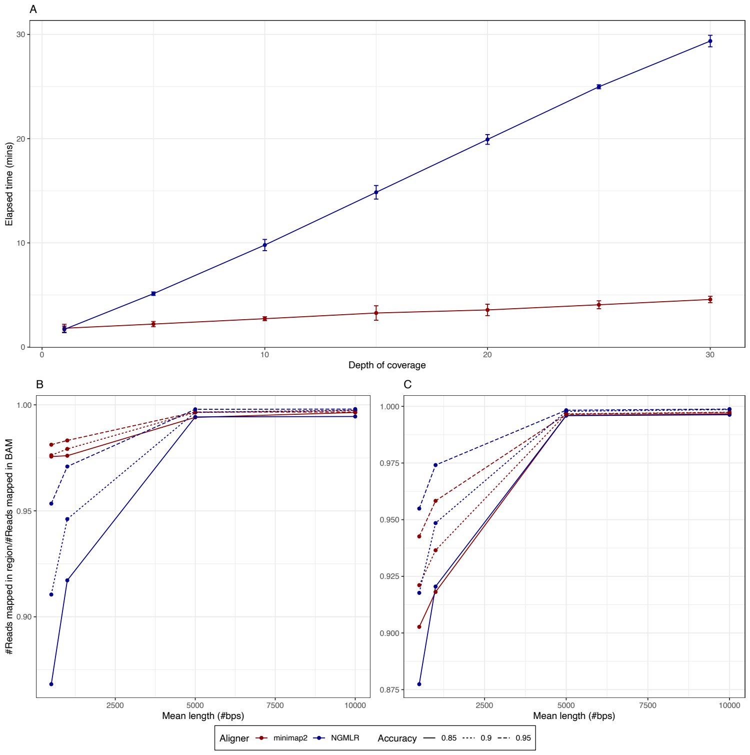


FIGURE 13: `Performances of minimap2 and NGMLR`. Minimap2 (red line) and NGMLR (blue line)'s speed (y-axis) when aligning increasing number of reads (x-axis). Minimap2 and NGMLR's accuracy performances (y-axis) on synthetic ONT (**B**) and PacBio (**C**) reads having different average lengths (x-axis). Figure is taken from *TRiCoLOR: tandem repeat profiling using whole-genome long-read sequencing data*(Bolognini et al., 2020a).

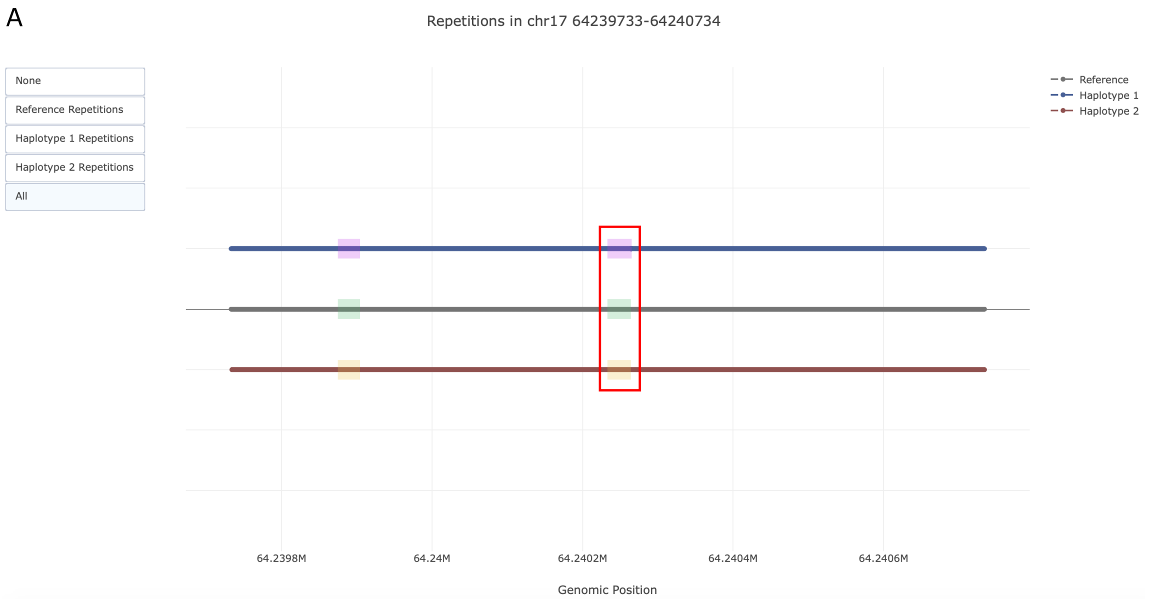


Figure 14: `HTML file from TRiCoLOR ApP`. Home page of the HTML file generated by TRiCoLOR REFER. Figure is taken from *TRiCoLOR: tandem repeat profiling using whole-genome long-read sequencing data*(Bolognini et al., 2020a).

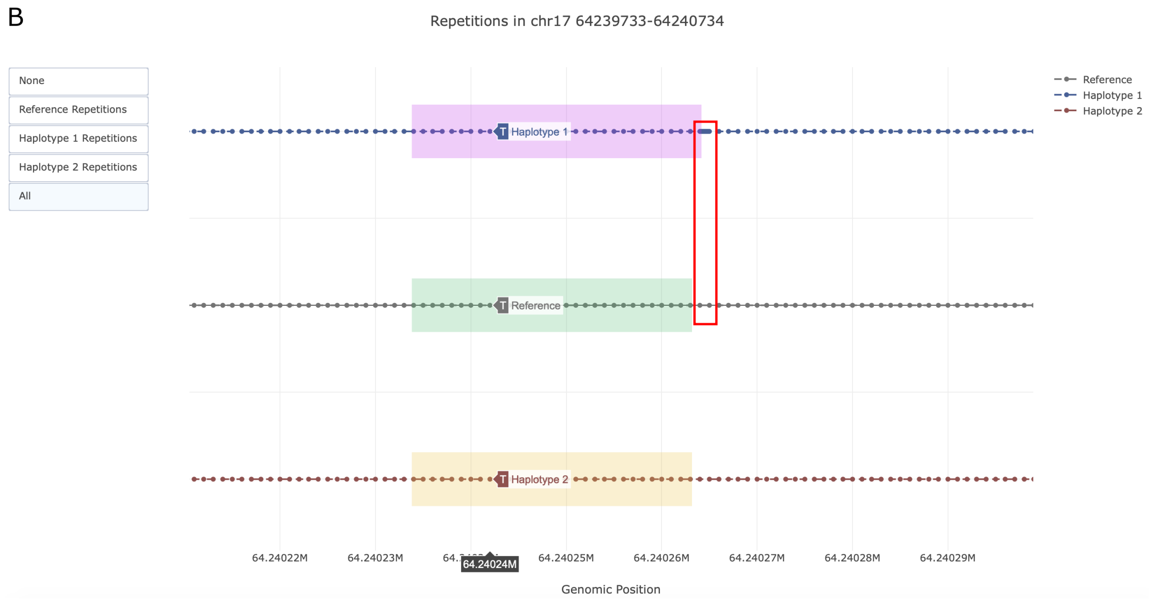


Figure 14: `HTML file from TRiCoLOR ApP`. Higher magnification of the HTML file generated by TRiCoLOR REFER. Figure is taken from *TRiCoLOR: tandem repeat profiling using whole-genome long-read sequencing data*(Bolognini et al., 2020a).

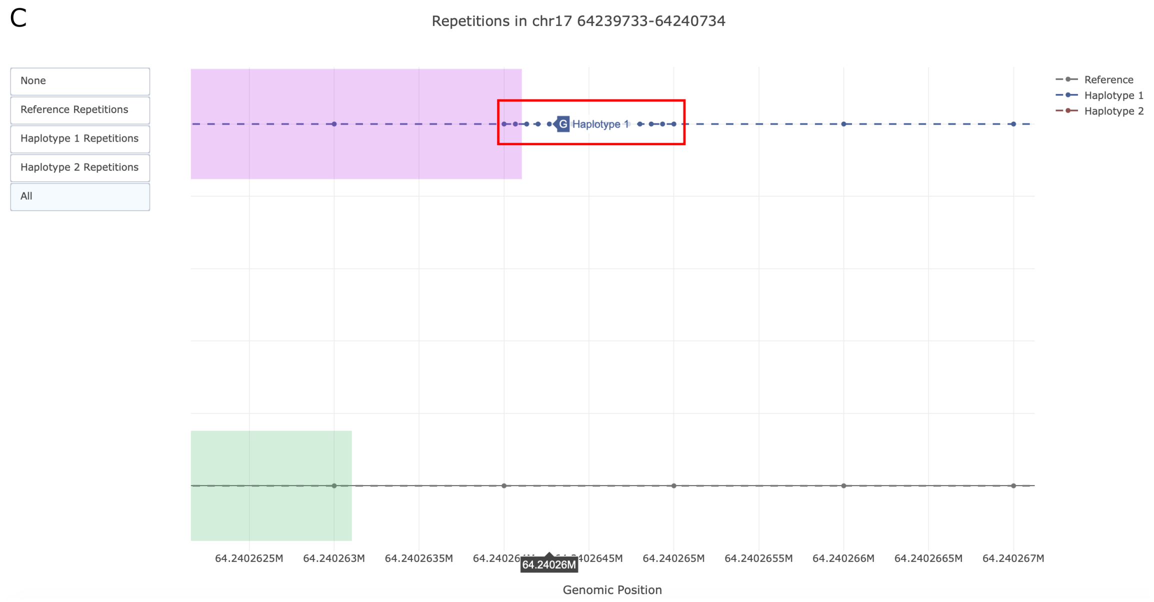


Figure 14: `HTML file from TRiCoLOR ApP`. Further magnification of the HTML file generated by TRiCoLOR REFER. Figure is taken from *TRiCoLOR: tandem repeat profiling using whole-genome long-read sequencing data*(Bolognini et al., 2020a).

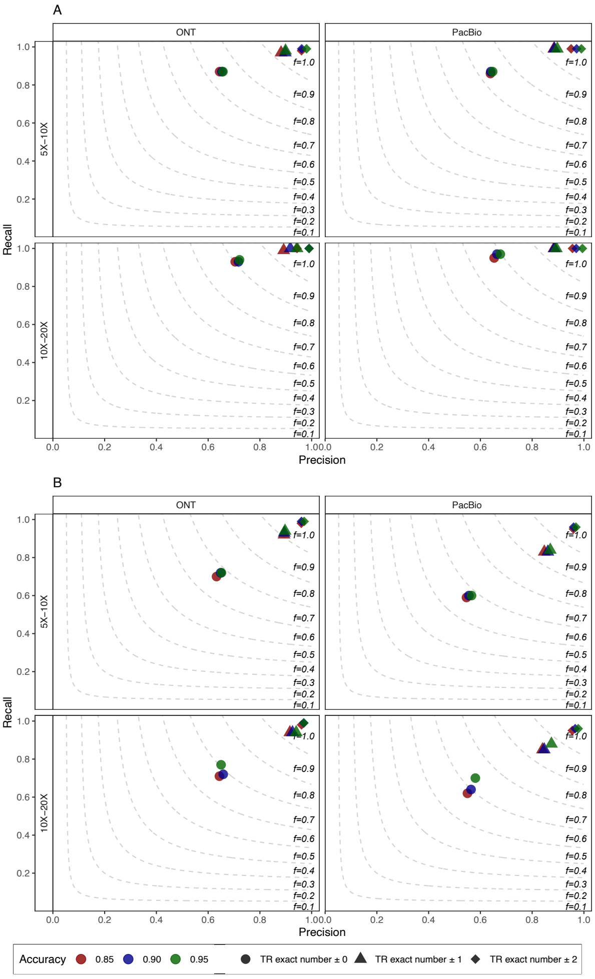


FIGURE 15: P, R and F1 of TRiCoLOR. TRiCoLOR's P (x-axis), R (y-axis) and F1 (dashed lines) on syntethic TR contractions (**A**) and expansions (**B**). ONT and PB reads exhibit variable error rates (accuracy ∼0.85, red; accuracy ∼0.90, blue; accuracy ∼0.95, green) and were simulated using variable haplotype-specific depth of coverage. P, R and F1 were calculated allowing no motif discrepancies (circle symbol), 1 motif discrepancy (triangle symbol) or 2 motif discrepancies (rhombus symbol) between TRiCoLOR's predictions and the number of repeated motifs in the ground truth. Figure is adapted from *TRiCoLOR: tandem repeat profiling using whole-genome long-read sequencing data*(Bolognini et al., 2020a).

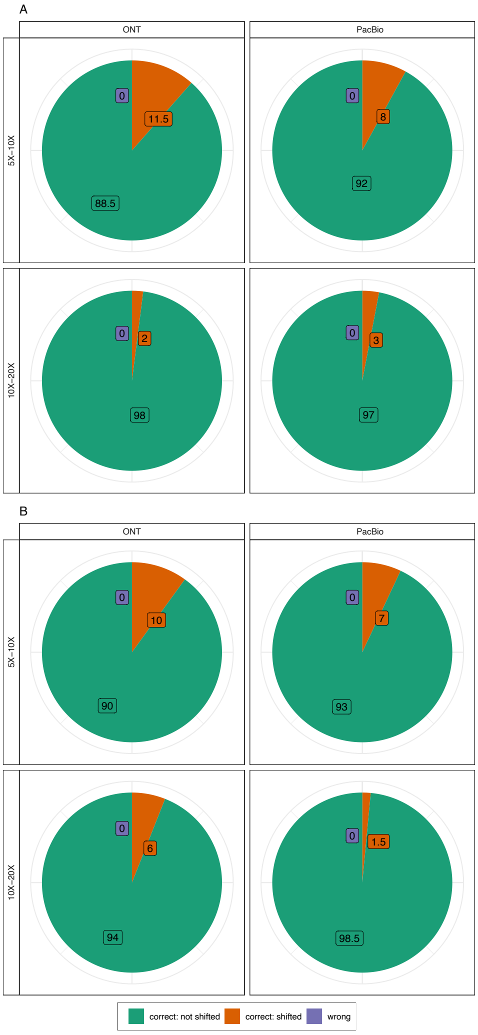


Figure 16: `Accuracy of TRiCoLOR on motif prediction`. The pie charts contain the percentage of the motifs correctly (not shifted, green; shifted, orange) and wrongly (violet) predicted by TRiCoLOR on synthetic data from Figure 15. Figure is adapted from *TRiCoLOR: tandem repeat profiling using whole-genome long-read sequencing data*(Bolognini et al., 2020a).

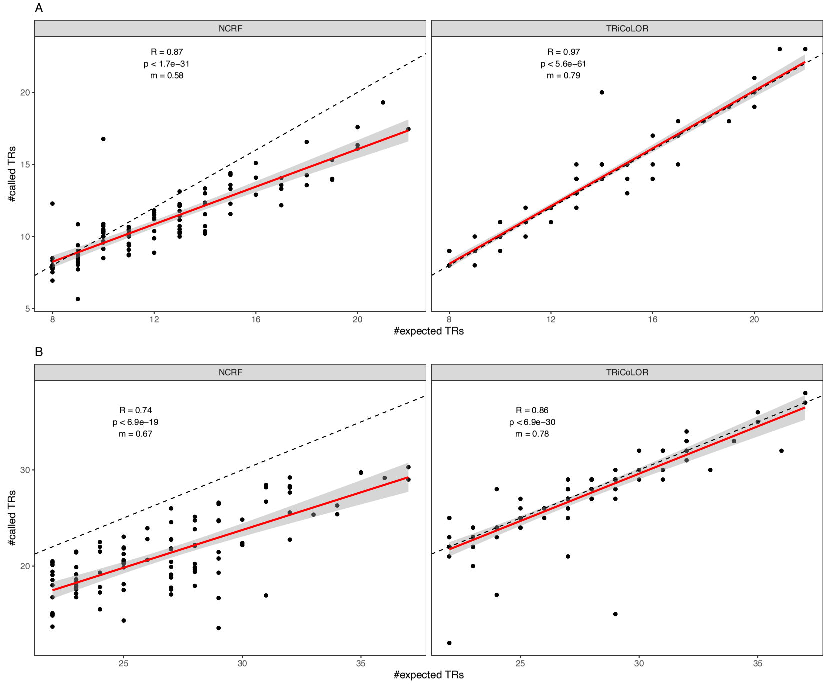


FIGURE 17: `Correlation results of TRiCoLOR and NCRF on short repeats`. Correlation results between the number of repeated motifs in the ground truth (x-axis) and the number of repeated motifs predicted by TRiCoLOR and NCRF (y-axis) for syntethic TR contractions (**A**) and expansions (**B**). Each dot represents the synthetic contraction/expansion of a single TR. R is the Pearson's correlation coefficient, p is the p-value of the linear regression analysis, m is the slope of the regression line and the dashed line is the bisector of the first quadrant angle that marks the perfect correspondence between expected and predicted number of TRs. Figure is taken from *TRiCoLOR: tandem repeat profiling using whole-genome long-read sequencing data* (Bolognini et al., 2020a).

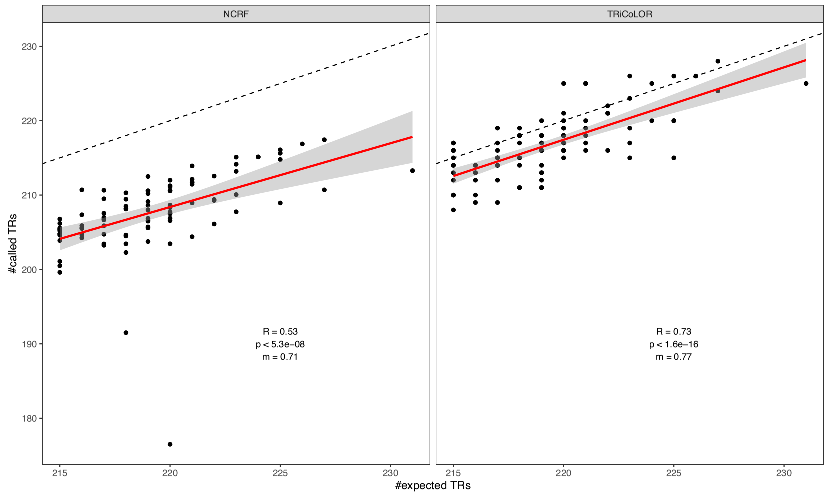


Figure 18: `Correlation results of TRiCoLOR and NCRF on long repeats`. Correlation results between the number of repeated motifs in the ground truth (x-axis) and the number of repeated motifs predicted by TRiCoLOR and NCRF (y-axis) for long syntethic TR expansions. Each dot represents the synthetic contraction/expansion of a single TR. R is the Pearson's coefficient, p is the p-value of the linear regression analysis, m is the slope of the regression line and the dashed line is the bisector of the first quadrant angle that marks the perfect correspondence between expected and predicted number of TRs. Figure is taken from *TRiCoLOR: tandem repeat profiling using whole-genome long-read sequencing data*(Bolognini et al., 2020a).

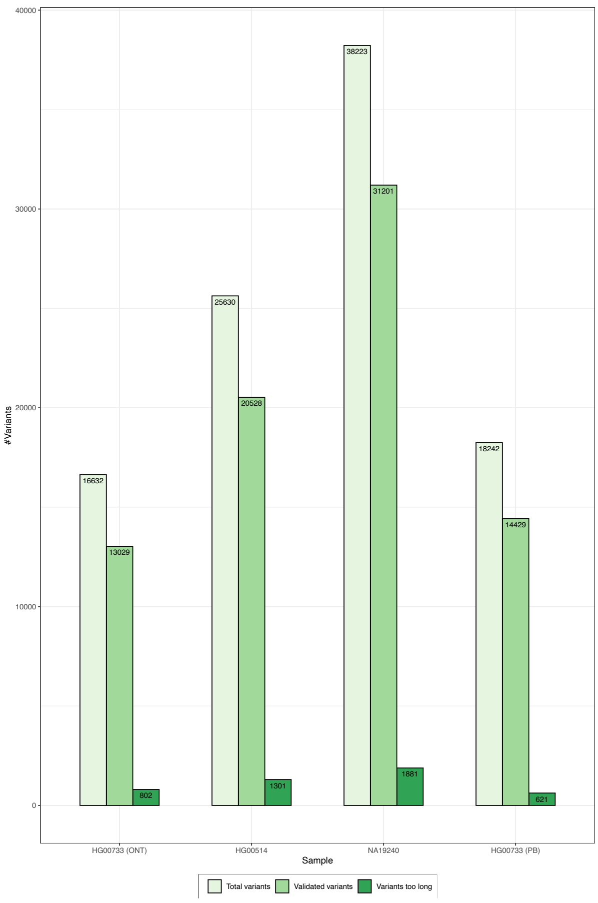


FIGURE 19: `Validation ratios`. Reference-free validation ratios of TRiCoLOR calls on the HGSVC individuals HG00733 (ONT and PacBio), HG00514 and NA19240 (total calls, light green; valid calls, green; variants too long to be validated, dark green). Figure is taken from *TRiCoLOR: tandem repeat profiling using whole-genome long-read sequencing data*(Bolognini et al., 2020a).

Tables 7

Disease	Symbol	OMIM	Gene	Cytogenetic location	Start GRCh37	Repeat Motif	Normal Range	Expansion Range	Reference Repeat Number
Huntington Disease	HD	143100	HTT	4p16.3	chr4:3076604	CAG	6–34	36–100+	21.3
Kennedy Disease	SBMA	313200	AR	Xq12	chrX:66765159	CAG	9–35	38–62	33.3
Spinocerebellar Ataxia 1	SCA1	164400	ATXN1	6p23	chr6:16327865	CAG	6–38	39–82	30.3
Spinocerebellar Ataxia 2	SCA2	183090	ATXN2	12q24	chr12:112036754	CAG	15–24	32–200	23.3
Machado-Joseph Disease	SCA3	109150	ATXN3	14q32.1	chr14:92537355	CAG	13–46	61-84	14
Spinocerebellar Ataxia 6	SCA6	183086	CACNA1A	19p13	chr19:13318673	CAG	4–7	21-33	13.3
Spinocerebellar Ataxia 7	SCA7	164500	ATXN7	3p14.1	chr3:63898361	CAG	4-35	37-306	32
Spinocerebellar Ataxia 17	SCA17	607136	TBP	6q27	chr6:170870995	CAG	7–34	49-88	19.7
Dentatorubral-Pallidoluysian Atrophy	DRPLA	125370	DRPLA/ATN1	12p13.31	chr12:7045880	CAG	7-34	49-88	19.7
Huntington Disease-Like 2	HDL2	606438	JPH3	16q24.3	chr16:87637889	CTG	7–28	66–78	15.3
Fragile-X Site A	FRAXA	300624	FMR1	Xq27.3	chrX:146993555	CGG	6–54	200–1000+	25
Fragile-X Site E	FRAXE	309548	FMR2	Xq28	chrX:147582159	CCG	4–39	200–900	15.3
Myotonic Dystrophy 1	DM1	160900	DMPK	19q13	chr19:46273463	CTG	5-37	50–10000	20.7
Friedreich Ataxia	FRDA	229300	FXN	9q13	chr9:71652201	GAA	6-32	200-1700	6.7
Myotonic Dystrophy 2	DM2	602668	ZNF9/CNBP	3q21.3	chr3:128891420	CCTG	10-26	65-11000	20.8
Frontotemporal Dementia	FTDALS1	105550	C9orf72	9p21	chr9:27573483	GGGGCC	2–19	250–1600	10.8
Spinocerebellar Ataxia 36	SCA36	614153	NOP56	20p13	chr20:2633379	GGCCTG	3–8	1500-2500	7.2
Spinocerebellar Ataxia 10	SCA10	603516	ATXN10	22q13.31	chr22:46191235	ATTCT	10-20	500-4500	14
Myoclonic Epilepsy of Unverricht and Lundborg	EPM1	254800	CSTB	21q22.3	chr21:45196324	CCCCGCCCCGCG	2-3	40-80	3.1
Spinocerebellar Ataxia 12	SCA12	604326	PPP2R2B	5q32	chr5:146258291	CAG	7-45	55-78	10.7
Spinocerebellar Ataxia 8	SCA8	608768	ATXN8OS/ATXN8	13q21	chr13:70713516	CTG	16–34	74+	15.3
Spinocerebellar Ataxia 31	SCA31	117210	BEAN1/TK2	16q21	chr16:66524302	TGGAA	0	2500-3800	0
Spinocerebellar Ataxia 37	SCA37	615945	DAB1	1p32.2	chr1:57832716	ATTTC	0	31-75	0
Familial Adult Myoclonic Epilepsy 1	FAME1	601068	SAMD12	8q24	chr8:119379055	TTTCA	0	440-3680	0
Fuchs Endothelial Corneal Dystrophy 3	FECD3	613267	TCF4	18q21.2	chr18:53253385	CTG	10–40	50-150+	25.3
Oculopharyngeal Muscular Dystrophy	OPMD	164300	PABPN1	14q11.2	chr14:23790682	GCG	6-7	8-13	6.7
Early Infantile Epileptic Encephalopathy 1	EIEE1	308350	ARX	Xp21.3	chrX:25031771	GCG	7-12	17-20	14.7

TABLE 1: `Microsatellite` *`loci`* `involved in neurological disorders`. The table contains detailed informations on microsatellite *loci* involved in neurological disorders associated with repeat expansions. Inserted repeats do not appear in the reference at the respective *locus*, thus having repeat number 0. Table is adapted from *Recent advances in the detection of repeat expansions with short-read next-generation sequencing* (Bahlo etal., 2018).

Scope	Program	Website	Reference
Short-read Alignment	BWA	https://github.com/lh3/bwa	(Li and Durbin, 2009)
	Bowtie2	https://github.com/BenLangmead/bowtie2	(Langmead and Salzberg, 2012)
	BBMap	https://sourceforge.net/projects/bbmap/	(Bushnell, 2014)
Short-read Assembly	SAGE	https://github.com/lucian-ilie/SAGE2	(Ilie etal., 2014)
	ABySS	https://github.com/bcgsc/abyss	(Jackman etal., 2017)
Short-read Variant Calling	Pindel	http://gmt.genome.wustl.edu/packages/pindel/	(Ye etal., 2009)
	FermiKit	https://github.com/lh3/fermikit	(Li, 2012)
Reference Tandem Repeats Calling	TRF	https://tandem.bu.edu/trf/trf.html	(Benson, 1999)
	MsatFinder	http://web.archive.org/web/20071026090642/http://www.genomics.ceh.ac.uk/msatfinder/	
	SSRIT	https://archive.gramene.org/db/markers/ssrtool	(Temnykh etal., 2001)
	MISA	https://webblast.ipk-gatersleben.de/misa/	(Thiel etal., 2003)
Short-read Tandem Repeats Calling	LobSTR	http://lobstr.teamerlich.org	(Gymrek etal., 2012)
	ExpansionHunter	https://github.com/Illumina/ExpansionHunter	(Dolzhenko etal., 2019)
	STRetch	https://github.com/Oshlack/STRetch	(Dashnow etal., 2018)
	exSTRa	https://github.com/bahlolab/exSTRa	(Tankard etal., 2018)
	TREDPARSE	https://github.com/humanlongevity/tredparse	(Tang etal., 2017)
	GangSTR	https://github.com/gymreklab/GangSTR	(Mousavi etal., 2019)
Long-read Tandem Repeats Calling	PacmonSTR	https://github.com/alibashir/pacmonstr	(Ummat and Bashir, 2014)
	NCRF	https://github.com/makovalab-psu/NoiseCancellingRepeatFinder	(Harris etal., 2019)
	TideHunter	https://github.com/Xinglab/TideHunter	(Gao etal., 2019)
	NanoSatellite	https://github.com/arnederoeck/NanoSatellite	(De Roeck etal., 2019)
	TRiCoLOR	https://github.com/davidebolo1993/TRiCoLOR	

TABLE 2: `Repeats-related tools`. The table summarizes the tools cited in this dissertation that are inherent with the discovery of repeats in genomes.

CHROMOSOME	START	END	HGSVC ASSEMBLY*	TRiCoLOR CALL*
chr1	23703657	23703893	DEL;INS	DEL;INS
chr1	223672571	223672681	INS;INS	INS;INS
chr10	69539376	69539572	INS;INS	INS;INS
chr11	79190887	79191145	**REF**;**REF**	**DEL**;**INS**
chr11	128436913	128437081	INS;INS	INS;INS
chr14	84276747	84276903	**REF**;DEL	**INS**;DEL
chr15	70364402	70364587	INS;**NA**	INS;**INS**
chr16	3529535	3529854	REF;**DEL**	LC;**INS**
chr17	27525992	27526118	INS;INS	INS;INS
chr18	44544809	44545037	INS;INS	INS;INS
chr18	59081301	59081379	INS;**INS**	INS;**INS**
chr18	71198388	71198450	REF;**NA**	REF;**INS**
chr2	160426201	160426342	INS;INS	INS;INS
chr2	211860947	211861156	DEL;**NA**	DEL;**INS**
chr21	35063465	35063588	INS;INS	INS;INS
chr22	46174187	46174274	REF;INS	REF;INS
chr3	13856835	13857013	DEL;INS	DEL;INS
chr4	13807826	13807982	REF;**REF**	REF;**INS**
chr4	18837113	18837320	INS;DEL	INS;DEL
chr4	81637241	81637408	DEL;DEL	DEL;DEL
chr5	54513584	54513735	**REF**;INS	**INS**;INS
chr6	25450910	25450975	REF;**INS**	REF;**INS**
chr6	55543085	55543393	INS;INS	INS;INS
chr6	106945844	106946002	DEL;**DEL**	DEL;**INS**
chr7	38610247	38610412	**NA**;DEL	**INS**;DEL
chr7	71847696	71847865	INS;INS	INS;INS
chr7	109663557	109663744	INS;DEL	INS;DEL
chr7	131933466	131933651	INS;INS	INS;INS
chr9	82850174	82850347	DEL;DEL	DEL;DEL
chr9	91622218	91622365	**NA**;NA	**INS**;REF
chr9	91634814	91634973	**NA**;**NA**	**DEL**;**INS**
chr9	116632126	116632280	INS;INS	INS;INS

* DEL indicates a deletion; INS indicates an insertion; REF indicates a reference allele; NA indicates that the region is not covered by the assembly or mis-assembled; LC indicates that TRiCoLOR could not generate a consensus sequence for the allele due to the low coverage in the region. The 2 alleles are separated by a semicolon. Differing alleles are highlighted in bold.

TABLE 3: Comparison of TRiCoLOR and HGSVC calls. Comparison between TRiCoLOR's mapping-based and HGSVC's assembly-based approaches for Mendelian consistent long TRs identified by TRiCoLOR on the HG0733 PacBio individual.

Bibliography

Can Alkan, Bradley P. Coe, and Evan E. Eichler. Genome structural variation discovery and genotyping, 2011. ISSN 14710056.

Simon Ardui, Valerie Race, Alena Zablotskaya, Matthew S. Hestand, Hilde Van Esch, Koenraad Devriendt, Gert Matthijs, and Joris R. Vermeesch. Detecting AGG Interruptions in Male and Female FMR1 Premutation Carriers by Single-Molecule Sequencing. *Human Mutation*, 2017. ISSN 10981004. doi: 10.1002/humu.23150.

Şule Ari and Muzaffer Arikan. Next-generation sequencing: Advantages, disadvantages, and future. In *Plant Omics: Trends and Applications*. 2016. ISBN 9783319317038. doi: 10.1007/978-3-319-31703-8_5.

Maurice R. Atkinson, Murray P. Deutscher, Arthur Kornberg, Alan F. Russell, and J. G. Moffatt. Enzymatic Synthesis of Deoxyribonucleic Acid. XXXIV. Termination of Chain Growth by a 2′,3′-Dideoxyribonucleotide. *Biochemistry*, 1969. ISSN 15204995. doi: 10.1021/bi00840a037.

Melanie Bahlo, Mark F. Bennett, Peter Degorski, Rick M. Tankard, Martin B. Delatycki, and Paul J. Lockhart. Recent advances in the detection of repeat expansions with short-read next-generation sequencing [version 1; referees: 3 approved], 2018. ISSN 1759796X.

Marina Barba, Henryk Czosnek, and Ahmed Hadidi. Historical perspective, development and applications of next-generation sequencing in plant virology. *Viruses*, 2013. ISSN 19994915. doi: 10.3390/v6010106.

John Beaulaurier, Eric E. Schadt, and Gang Fang. Deciphering bacterial epigenomes using modern sequencing technologies, 2019. ISSN 14710064.

Amir Behdad, Helmut C. Weigelin, Kojo S.J. Elenitoba-Johnson, and Bryan L. Betz. A clinical grade sequencing-based assay for CEBPA mutation testing: Report of a large series of myeloid neoplasms. *Journal of Molecular Diagnostics*, 2015. ISSN 19437811. doi: 10.1016/j.jmoldx.2014.09.007.

Gary Benson. Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Research*, 1999. ISSN 03051048. doi: 10.1093/nar/27.2.573.

Davide Bolognini, Niccolò Bartalucci, Alessandra Mingrino, Alessandro Maria Vannucchi, and Alberto Magi. NANOR: A user-friendly R package to analyze and compare nanopore sequencing data. *PLoS ONE*, 2019. ISSN 19326203. doi: 10.1371/journal.pone.0216471.

Davide Bolognini, Alberto Magi, Vladimir Benes, Jan O Korbel, and Tobias Rausch. TRiCoLOR: tandem repeat profiling using whole-genome long-read sequencing data. *GigaScience*, 9(10), 10 2020a. ISSN 2047-217X. doi: 10.1093/gigascience/giaa101. URL `https://doi.org/10.1093/gigascience/giaa101`. giaa101.

Davide Bolognini, Ashley Sanders, Jan O. Korbel, Alberto Magi, Vladimir Benes, and Tobias Rausch. VISOR: A versatile haplotype-aware structural variant simulator for short-and long-read sequencing. *Bioinformatics*, 2020b. ISSN 14602059. doi: 10.1093/bioinformatics/btz719.

R. J. Britten and D. E. Kohne. Repeated sequences in DNA, 1968. ISSN 00368075.

Katarzyna Bryc, Nick Patterson, and David Reich. A novel approach to estimating heterozygosity from low-coverage genome sequence. *Genetics*, 2013. ISSN 00166731. doi: 10.1534/genetics.113.154500.

Jason D. Buenrostro, Paul G. Giresi, Lisa C. Zaba, Howard Y. Chang, and William J. Greenleaf. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, 2013. ISSN 15487091. doi: 10.1038/nmeth.2688.

M Burrows and Dj Wheeler. A block-sorting lossless data compression algorithm. *Algorithm, Data Compression*, 1994. ISSN 15708667. doi: 10.1.1.37.6774.

Brian Bushnell. BBMap: a fast, accurate, splice-aware aligner. Technical report, 2014.

Melanie A. Carless. Determination of DNA methylation levels using illumina humanmethylation450 beadchips. In *Chromatin Protocols: Third Edition*. 2015. ISBN 9781493924745. doi: 10.1007/978-1-4939-2474-5_10.

Jean Michel Carter and Shobbir Hussain. Robust long-read native DNA sequencing using the ONT CsgG Nanopore system. *Wellcome Open Research*, 2017. ISSN 2398502X. doi: 10.12688/wellcomeopenres.11246.1.

Mark J. Chaisson and Glenn Tesler. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): Application and theory. *BMC Bioinformatics*, 2012. ISSN 14712105. doi: 10.1186/1471-2105-13-238.

Mark J.P. Chaisson, John Huddleston, Megan Y. Dennis, Peter H. Sudmant, Maika Malig, Fereydoun Hormozdiari, Francesca Antonacci, Urvashi Surti, Richard Sandstrom, Matthew Boitano, Jane M. Landolin, John A. Stamatoyannopoulos, Michael W. Hunkapiller, Jonas Korlach, and Evan E. Eichler. Resolving the complexity of the human genome using single-molecule sequencing. *Nature*, 2015. ISSN 14764687. doi: 10.1038/nature13907.

Mark J.P. Chaisson, Ashley D. Sanders, Xuefang Zhao, Ankit Malhotra, David Porubsky, Tobias Rausch, Eugene J. Gardner, Oscar L. Rodriguez, Li Guo, Ryan L. Collins, Xian Fan, Jia Wen, Robert E. Handsaker, Susan Fairley, Zev N. Kronenberg, Xiangmeng Kong, Fereydoun Hormozdiari, Dillon Lee, Aaron M. Wenger, Alex R. Hastie, Danny Antaki, Thomas Anantharaman, Peter A. Audano, Harrison Brand, Stuart Cantsilieris, Han Cao, Eliza Cerveira, Chong Chen, Xintong Chen, Chen Shan Chin, Zechen Chong, Nelson T. Chuang, Christine C. Lambert, Deanna M. Church, Laura Clarke, Andrew Farrell, Joey Flores, Timur Galeev, David U. Gorkin, Madhusudan Gujral, Victor Guryev, William Haynes Heaton, Jonas Korlach, Sushant Kumar, Jee Young Kwon, Ernest T. Lam, Jong Eun Lee, Joyce Lee, Wan Ping Lee, Sau Peng Lee, Shantao Li, Patrick Marks, Karine Viaud-Martinez, Sascha Meiers, Katherine M. Munson, Fabio C.P. Navarro, Bradley J. Nelson, Conor Nodzak, Amina Noor, Sofia Kyriazopoulou-Panagiotopoulou, Andy W.C. Pang, Yunjiang Qiu, Gabriel Rosanio, Mallory Ryan, Adrian Stütz, Diana C.J. Spierings, Alistair Ward, Anne Marie E. Welch, Ming Xiao, Wei Xu, Chengsheng Zhang, Qihui Zhu, Xiangqun Zheng-Bradley, Ernesto Lowy, Sergei Yakneen, Steven McCarroll, Goo Jun, Li Ding, Chong Lek Koh, Bing Ren, Paul Flicek, Ken Chen, Mark B. Gerstein, Pui Yan Kwok, Peter M. Lansdorp, Gabor T. Marth, Jonathan Sebat, Xinghua Shi, Ali Bashir, Kai Ye, Scott E. Devine, Michael E. Talkowski, Ryan E. Mills, Tobias Marschall, Jan O. Korbel, Evan E. Eichler, and Charles Lee. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nature Communications*, 2019. ISSN 20411723. doi: 10.1038/s41467-018-08148-z.

Ranajit Chakraborty, Marek Kimmel, David N. Stivers, Leslea J. Davison, and Ranjan Deka. Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. *Proceedings of the National Academy of Sciences of the United States of America*, 1997. ISSN 00278424. doi: 10.1073/pnas.94.3.1041.

Charles Y.K. Cheung, Elizabeth A. Thompson, and Ellen M. Wijsman. Detection of mendelian consistent genotyping errors in pedigrees. *Genetic Epidemiology*, 2014. ISSN 10982272. doi: 10.1002/gepi.21806.

Yongwook Choi, Agnes P. Chan, Ewen Kirkness, Amalio Telenti, and Nicholas J. Schork. Comparison of phasing strategies for whole human genomes. *PLoS Genetics*, 2018. ISSN 15537404. doi: 10.1371/journal.pgen.1007308.

Richard Cordaux and Mark A. Batzer. The impact of retrotransposons on human genome evolution, 2009. ISSN 14710056.

Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert E. Handsaker, Gerton Lunter, Gabor T. Marth, Stephen T. Sherry, Gilean McVean, and Richard Durbin. The variant call format and VCFtools. *Bioinformatics*, 2011. ISSN 13674803. doi: 10.1093/bioinformatics/btr330.

Agus Darwanto, Anne Mette Hein, Sascha Strauss, Yi Kong, Andrew Sheridan, Dan Richards, Eric Lader, Monika Ngowe, Timothy Pelletier, Danielle Adams, Austin Ricker, Nishit Patel, Andreas Kühne, Simon Hughes, Dan Shiffman, Dirk Zimmermann, Kai te Kaat, and Thomas Rothmann. Use of the QIAGEN GeneReader NGS system for detection of KRAS mutations, validated by the QIAGEN Therascreen PCR kit and alternative NGS platform. *BMC Cancer*, 2017. ISSN 14712407. doi: 10.1186/s12885-017-3328-z.

Harriet Dashnow, Monkol Lek, Belinda Phipson, Andreas Halman, Simon Sadedin, Andrew Lonsdale, Mark Davis, Phillipa Lamont, Joshua S. Clayton, Nigel G. Laing, Daniel G. MacArthur, and Alicia Oshlack. STRetch: Detecting and discovering pathogenic short tandem repeat expansions. *Genome Biology*, 2018. ISSN 1474760X. doi: 10.1186/s13059-018-1505-2.

Arne De Roeck, Wouter De Coster, Liene Bossaerts, Rita Cacace, Tim De Pooter, Jasper Van Dongen, Svenn D'Hert, Peter De Rijk, Mojca Strazisar, Christine Van Broeckhoven, and Kristel Sleegers. NanoSatellite: Accurate characterization of expanded tandem repeat length and sequence through whole genome long-read sequencing on PromethION. *Genome Biology*, 2019. ISSN 1474760X. doi: 10.1186/s13059-019-1856-3.

David Deamer, Mark Akeson, and Daniel Branton. Three decades of nanopore sequencing, 2016. ISSN 15461696.

Dirk D. Dolle, Zhicheng Liu, Matthew Cotten, Jared T. Simpson, Zamin Iqbal, Richard Durbin, Shane A. McCarthy, and Thomas M. Keane. Using reference-free compressed data structures to analyze sequencing reads from thousands of human genomes. *Genome Research*, 2017. ISSN 15495469. doi: 10.1101/gr.211748.116.

Egor Dolzhenko, Viraj Deshpande, Felix Schlesinger, Peter Krusche, Roman Petrovski, Sai Chen, Dorothea Emig-Agius, Andrew Gross, Giuseppe Narzisi, Brett Bowman, Konrad Scheffler, Joke J.F.A. Van Vugt, Courtney French, Alba Sanchis-Juan, Kristina Ibáñez, Arianna Tucci, Bryan R. Lajoie, Jan H. Veldink, F. Lucy Raymond, Ryan J. Taft, David R. Bentley, Michael A. Eberle, and Inanc Birol. ExpansionHunter: A sequence-graph-based tool to analyze variation in short tandem repeat regions. *Bioinformatics*, 2019. ISSN 14602059. doi: 10.1093/bioinformatics/btz431.

Radoje Drmanac, Andrew B. Sparks, Matthew J. Callow, Aaron L. Halpern, Norman L. Burns, Bahram G. Kermani, Paolo Carnevali, Igor Nazarenko, Geoffrey B. Nilsen, George Yeung, Fredrik Dahl, Andres Fernandez, Bryan Staker, Krishna P. Pant, Jonathan Baccash, Adam P. Borcherding, Anushka Brownley, Ryan Cedeno, Linsu Chen, Dan Chernikoff, Alex Cheung, Razvan Chirita, Benjamin Curson, Jessica C. Ebert, Coleen R. Hacker, Robert Hartlage, Brian Huser, Steve Huang, Yuan Jiang, Vitali Karpinchyk, Mark Koenig, Calvin Kong, Tom Landers, Catherine Le, Jia Liu, Celeste E. McBride, Matt Morenzoni, Robert E. Morey, Karl Mutch, Helena Perazich, Kimberly Perry, Brock A. Peters, Joe Peterson, Charit L. Pethiyagoda, Kaliprasad Pothuraju, Claudia Richter, Abraham M. Rosenbaum, Shaunak Roy, Jay Shafto, Uladzislau Sharanhovich, Karen W. Shannon, Conrad G. Sheppy, Michel Sun, Joseph V. Thakuria, Anne Tran, Dylan Vu, Alexander Wait Zaranek, Xiaodi Wu, Snezana Drmanac, Arnold R. Oliphant, William C. Banyai, Bruce Martin, Dennis G. Ballinger, George M. Church, and Clifford A. Reid. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science*, 2010. ISSN 00368075. doi: 10.1126/science.1181498.

Jana Ebler, Marina Haukness, Trevor Pesout, Tobias Marschall, and Benedict Paten. Haplotype-aware diplotyping from noisy long reads. *Genome Biology*, 2019. ISSN 1474760X. doi: 10.1186/s13059-019-1709-0.

Peter Edge and Vikas Bansal. Longshot enables accurate variant calling in diploid genomes from single-molecule long read sequencing. *Nature Communications*, 2019. ISSN 20411723. doi: 10.1038/s41467-019-12493-y.

Peter Edge, Vineet Bafna, and Vikas Bansal. HapCUT2: Robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Research*, 2017. ISSN 15495469. doi: 10.1101/gr.213462.116.

John Eid, Adrian Fehr, Jeremy Gray, Khai Luong, John Lyle, Geoff Otto, Paul Peluso, David Rank, Primo Baybayan, Brad Bettman, Arkadiusz Bibillo, Keith Bjornson, Bidhan Chaudhuri, Frederick Christians, Ronald Cicero, Sonya Clark, Ravindra Dalal, Alex DeWinter, John Dixon, Mathieu Foquet, Alfred Gaertner, Paul Hardenbol, Cheryl Heiner, Kevin Hester, David Holden, Gregory Kearns, Xiangxu Kong, Ronald Kuse, Yves Lacroix, Steven Lin, Paul Lundquist, Congcong Ma, Patrick Marks, Mark Maxham, Devon Murphy, Insil Park, Thang Pham, Michael Phillips, Joy Roy, Robert Sebra, Gene Shen, Jon Sorenson, Austin Tomaney, Kevin Travers, Mark Trulson, John Vieceli, Jeffrey Wegener, Dawn Wu, Alicia Yang, Denis Zaccarin, Peter Zhao, Frank Zhong, Jonas Korlach, and Stephen Turner. Real-time DNA sequencing from single polymerase molecules. *Science*, 2009. ISSN 00368075. doi: 10.1126/science.1162986.

Tobias Fehlmann, Stefanie Reinheimer, Chunyu Geng, Xiaoshan Su, Snezana Drmanac, Andrei Alexeev, Chunyan Zhang, Christina Backes, Nicole Ludwig, Martin Hart, Dan An, Zhenzhen Zhu, Chongjun Xu, Ao Chen, Ming Ni, Jian Liu, Yuxiang Li, Matthew Poulter, Yongping Li, Cord Stähler, Radoje Drmanac, Xun Xu, Eckart Meese, and Andreas Keller. cPAS-based sequencing on the BGISEQ-500 to explore small non-coding RNAs. *Clinical Epigenetics*, 2016. ISSN 18687083. doi: 10.1186/s13148-016-0287-1.

Paolo Ferragina and Giovanni Manzini. Opportunistic data structures with applications. In *Annual Symposium on Foundations of Computer Science - Proceedings*, 2000. doi: 10.1109/sfcs.2000.892127.

Yan Gao, Bo Liu, Yadong Wang, and Yi Xing. TideHunter: Efficient and sensitive tandem repeat detection from noisy long-reads using seed-and-chain. In *Bioinformatics*, 2019. doi: 10.1093/bioinformatics/btz376.

Manuel A. Garrido-Ramos. Satellite DNA: An evolving topic, 2017. ISSN 20734425.

Rita Gemayel, Marcelo D. Vinces, Matthieu Legendre, and Kevin J. Verstrepen. Variable Tandem Repeats Accelerate Evolution of Coding and Regulatory Sequences. *Annual Review of Genetics*, 2010. ISSN 0066-4197. doi: 10.1146/annurev-genet-072610-155046.

Carla Giner-Delgado, Sergi Villatoro, Jon Lerga-Jaso, Magdalena Gayà-Vidal, Meritxell Oliva, David Castellano, Lorena Pantano, Bárbara D. Bitarello, David Izquierdo, Isaac Noguera, Iñigo Olalde, Alejandra Delprat, Antoine Blancher, Carles Lalueza-Fox, Tõnu Esko, Paul F. O'Reilly, Aida M. Andrés, Luca Ferretti, Marta Puig, and Mario Cáceres. Evolutionary and functional impact of common polymorphic inversions in the human genome. *Nature Communications*, 2019. ISSN 20411723. doi: 10.1038/s41467-019-12173-x.

Francesca Giordano, Louise Aigrain, Michael A. Quail, Paul Coupland, James K. Bonfield, Robert M. Davies, German Tischler, David K. Jackson, Thomas M. Keane, Jing Li, Jia Xing Yue, Gianni Liti, Richard Durbin, and Zemin Ning. De novo yeast genome assemblies from MinION, PacBio and MiSeq platforms. *Scientific Reports*, 2017. ISSN 20452322. doi: 10.1038/s41598-017-03996-z.

Hani Z. Girgis. Red: An intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale. *BMC Bioinformatics*, 2015. ISSN 14712105. doi: 10.1186/s12859-015-0654-5.

Sara Goodwin, James Gurtowski, Scott Ethe-Sayers, Panchajanya Deshpande, Michael C. Schatz, and W. Richard McCombie. Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. *Genome Research*, 2015. ISSN 15495469. doi: 10.1101/gr.191395.115.

Sara Goodwin, John D. McPherson, and W. Richard McCombie. Coming of age: Ten years of next-generation sequencing technologies, 2016. ISSN 14710064.

Fei Guo, Dan Wang, and Lusheng Wang. Progressive approach for SNP calling and haplotype assembly using single molecular sequencing data. *Bioinformatics*, 2018. ISSN 14602059. doi: 10.1093/bioinformatics/bty059.

Jia Guo, Ning Xu, Zengmin Li, Shenglong Zhang, Jian Wu, Hyun Kim Dae, Sano Marma Mong, Qinglin Meng, Huanyan Cao, Xiaoxu Li, Shundi Shi, Lin Yu, Sergey Kalachikov, James J. Russo, Nicholas J. Turro, and Jingyue Ju. Four-color DNA sequencing with 3′-O-modified nucleotide reversible terminators and chemically cleavable fluorescent dideoxynucleotides. *Proceedings of the National Academy of Sciences of the United States of America*, 2008. ISSN 00278424. doi: 10.1073/pnas.0804023105.

Melissa Gymrek, David Golan, Saharon Rosset, and Yaniv Erlich. lobSTR: A short tandem repeat profiler for personal genomes. *Genome Research*, 2012. ISSN 10889051. doi: 10.1101/gr.135780.111.

Olivier Harismendy, Pauline C. Ng, Robert L. Strausberg, Xiaoyun Wang, Timothy B. Stockwell, Karen Y. Beeson, Nicholas J. Schork, Sarah S. Murray, Eric J. Topol, Samuel Levy, and Kelly A. Frazer. Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biology*, 2009. ISSN 14747596. doi: 10.1186/gb-2009-10-3-r32.

Robert S. Harris, Monika Cechova, Kateryna D. Makova, and Inanc Birol. Noise-cancelling repeat finder: Uncovering tandem repeats in error-prone long-read sequencing data. *Bioinformatics*, 2019. ISSN 14602059. doi: 10.1093/bioinformatics/btz484.

Lucian Ilie, Bahlul Haider, Michael Molnar, and Roberto Solis-Oba. SAGE: String-overlap Assembly of GEnomes. *BMC Bioinformatics*, 2014. ISSN 14712105. doi: 10.1186/1471-2105-15-302.

Camilla L.C. Ip, Matthew Loose, John R. Tyson, Mariateresa de Cesare, Bonnie L. Brown, Miten Jain, Richard M. Leggett, David A. Eccles, Vadim Zalunin, John M. Urban, Paolo Piazza, Rory J. Bowden, Benedict Paten, Solomon Mwaigwisya, Elizabeth M. Batty, Jared T. Simpson, Terrance P. Snutch, Ewan Birney, David Buck, Sara Goodwin, Hans J. Jansen, Justin O'Grady, and Hugh E. Olsen. MinION Analysis and Reference Consortium: Phase 1 data release and analysis. *F1000Research*, 2015. ISSN 1759796X. doi: 10.12688/f1000research.7201.1.

Shaun D. Jackman, Benjamin P. Vandervalk, Hamid Mohamadi, Justin Chu, Sarah Yeo, S. Austin Hammond, Golnaz Jahesh, Hamza Khan, Lauren Coombe, Rene L. Warren, and Inanc Birol. ABySS 2.0:

Resource-efficient assembly of large genomes using a Bloom filter. *Genome Research*, 2017. ISSN 15495469. doi: 10.1101/gr.214346.116.

Bum Kim Jae, Gregory J. Porreca, Lei Song, Steven C. Greenway, Joshua M. Gorham, George M. Church, Christine E. Seidman, and J. G. Seidman. Polony multiplex analysis of gene expression (PMAGE) in mouse hypertrophic cardiomyopathy. *Science*, 2007. ISSN 00368075. doi: 10.1126/science.1137325.

Miten Jain, Sergey Koren, Karen H. Miga, Josh Quick, Arthur C. Rand, Thomas A. Sasani, John R. Tyson, Andrew D. Beggs, Alexander T. Dilthey, Ian T. Fiddes, Sunir Malla, Hannah Marriott, Tom Nieto, Justin O'Grady, Hugh E. Olsen, Brent S. Pedersen, Arang Rhie, Hollian Richardson, Aaron R. Quinlan, Terrance P. Snutch, Louise Tee, Benedict Paten, Adam M. Phillippy, Jared T. Simpson, Nicholas J. Loman, and Matthew Loose. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology*, 2018a. ISSN 15461696. doi: 10.1038/nbt.4060.

Miten Jain, Hugh E. Olsen, Daniel J. Turner, David Stoddart, Kira V. Bulazel, Benedict Paten, David Haussler, Huntington F. Willard, Mark Akeson, and Karen H. Miga. Linear assembly of a human centromere on the y chromosome. *Nature Biotechnology*, 2018b. ISSN 15461696. doi: 10.1038/nbt.4109.

Sol A. Jeon, Jong Lyul Park, Jong Hwan Kim, Jeong Hwan Kim, Yong Sung Kim, Jin Cheon Kim, and Seon Young Kim. Comparison of the MGISEQ-2000 and illumina hiseq 4000 sequencing platforms for RNA sequencing. *Genomics and Informatics*, 2019. ISSN 22340742. doi: 10.5808/GI.2019.17.3.e32.

Nahid N. Jetha, Matthew Wiggin, and Andre Marziali. Forming an alpha-hemolysin nanopore for single-molecule analysis. *Methods in molecular biology (Clifton, N.J.)*, 2009. ISSN 10643745. doi: 10.1007/978-1-59745-483-4_9.

Miriam K. Konkel and Mark A. Batzer. A mobile threat to genome stability: The impact of non-LTR retrotransposons upon the human genome, 2010. ISSN 1044579X.

Yinglei Lai and Fengzhu Sun. The Relationship between Microsatellite Slippage Mutation Rate and the Number of Repeat Units. *Molecular Biology and Evolution*, 2003. ISSN 07374038. doi: 10.1093/molbev/msg228.

Eric S. Lander, Lauren M. Linton, Bruce Birren, Chad Nusbaum, Michael C. Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William Fitzhugh, Roel Funke, Diane Gage, Katrina Harris, Andrew Heaford, John Howland, Lisa Kann, Jessica Lehoczky, Rosie Levine, Paul McEwan, Kevin McKernan, James Meldrim, Jill P. Mesirov, Cher Miranda, William Morris, Jerome Naylor, Christina Raymond, Mark Rosetti, Ralph Santos, Andrew Sheridan, Carrie Sougnez, Nicole Stange-Thomann, Nikola Stojanovic, Aravind Subramanian, Dudley Wyman, Jane Rogers, John Sulston, Rachael Ainscough, Stephan Beck, David Bentley, John Burton, Christopher Clee, Nigel Carter, Alan Coulson, Rebecca Deadman, Panos Deloukas, Andrew Dunham, Ian Dunham, Richard Durbin, Lisa French, Darren Grafham, Simon Gregory, Tim Hubbard, Sean Humphray, Adrienne Hunt, Matthew Jones, Christine Lloyd, Amanda McMurray, Lucy Matthews, Simon Mercer, Sarah Milne, James C. Mullikin, Andrew Mungall, Robert Plumb, Mark Ross, Ratna Shownkeen, Sarah Sims, Robert H. Waterston, Richard K. Wilson, Ladeana W. Hillier, John D. McPherson, Marco A. Marra, Elaine R. Mardis, Lucinda A. Fulton, Asif T. Chinwalla, Kymberlie H. Pepin, Warren R. Gish, Stephanie L. Chissoe, Michael C. Wendl, Kim D. Delehaunty, Tracie L. Miner, Andrew Delehaunty, Jason B. Kramer, Lisa L. Cook, Robert S. Fulton, Douglas L. Johnson, Patrick J. Minx, Sandra W. Clifton, Trevor Hawkins, Elbert Branscomb, Paul Predki, Paul Richardson, Sarah Wenning, Tom Slezak, Norman Doggett, Jan Fang Cheng, Anne Olsen, Susan Lucas, Christopher Elkin, Edward Uberbacher, Marvin Frazier, Richard A. Gibbs, Donna M. Muzny, Steven E. Scherer, John B. Bouck, Erica J. Sodergren, Kim C. Worley, Catherine M. Rives, James H. Gorrell, Michael L. Metzker, Susan L. Naylor, Raju S. Kucherlapati, David L. Nelson, George M. Weinstock, Yoshiyuki Sakaki, Asao Fujiyama, Masahira Hattori, Tetsushi Yada, Atsushi Toyoda, Takehiko Itoh, Chiharu Kawagoe, Hidemi Watanabe, Yasushi Totoki, Todd Taylor, Jean Weissenbach, Roland Heilig, William Saurin, Francois Artiguenave, Philippe Brottier, Thomas Bruls, Eric Pelletier, Catherine Robert, Patrick Wincker, André Rosenthal, Matthias Platzer, Gerald Nyakatura, Stefan Taudien, Andreas Rump, Douglas R. Smith, Lynn Doucette-Stamm, Marc Rubenfield, Keith Weinstock, Mei Lee Hong, Joann Dubois, Huanming Yang, Jun Yu, Jian Wang, Guyang Huang, Jun Gu, Leroy Hood, Lee Rowen, Anup Madan, Shizen Qin, Ronald W. Davis, Nancy A. Federspiel, A. Pia Abola, Michael J. Proctor, Bruce A. Roe, Feng Chen, Huaqin Pan, Juliane Ramser, Hans Lehrach, Richard Reinhardt, W. Richard McCombie, Melissa De La Bastide, Neilay Dedhia, Helmut Blöcker, Klaus Hornischer, Gabriele Nordsiek, Richa Agarwala, L. Aravind, Jeffrey A. Bailey, Alex Bateman, Serafim Batzoglou, Ewan Birney, Peer Bork, Daniel G. Brown, Christopher B. Burge, Lorenzo Cerutti, Hsiu Chuan Chen, Deanna Church, Michele Clamp, Richard R. Copley, Tobias Doerks, Sean R. Eddy, Evan E. Eichler, Terrence S. Furey, James Galagan, James G.R. Gilbert, Cyrus Harmon, Yoshihide Hayashizaki, David Haussler, Henning Hermjakob, Karsten Hokamp, Wonhee Jang, L. Steven Johnson, Thomas A. Jones, Simon Kasif, Arek Kaspryzk, Scot Kennedy, W. James Kent, Paul Kitts, Eugene V. Koonin, Ian Korf, David Kulp, Doron Lancet, Todd M. Lowe, Aoife McLysaght, Tarjei Mikkelsen, John V. Moran, Nicola Mulder, Victor J. Pollara, Chris P. Ponting, Greg Schuler, Jörg Schultz, Guy Slater, Arian F.A. Smit, Elia Stupka, Joseph Szustakowki, Danielle Thierry-Mieg, Jean Thierry-Mieg, Lukas Wagner, John Wallis, Raymond Wheeler, Alan Williams, Yuri I. Wolf, Kenneth H. Wolfe, Shiaw Pyng Yang, Ru Fang Yeh, Francis Collins, Mark S. Guyer, Jane Peterson, Adam Felsenfeld, Kris A. Wetterstrand, Richard M. Myers, Jeremy Schmutz, Mark Dickson, Jane Grimwood, David R. Cox, Maynard V. Olson, Rajinder Kaul, Christopher Raymond, Nobuyoshi Shimizu, Kazuhiko Kawasaki, Shinsei Minoshima, Glen A. Evans, Maria Athanasiou, Roger Schultz, Aristides Patrinos, and Michael J. Morgan. Initial sequencing and analysis of the human genome. *Nature*, 2001. ISSN 00280836. doi: 10.1038/35057062.

Ben Langmead and Steven L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 2012. ISSN 15487091. doi: 10.1038/nmeth.1923.

Andrew H. Laszlo, Ian M. Derrington, and Jens H. Gundlach. MspA nanopore as a single-molecule tool: From sequencing to SPRNT, 2016. ISSN 10959130.

John H. Leamon, William L. Lee, Karrie R. Tartaro, Janna R. Lanza, Gary J. Sarkis, Alex D. DeWinter, Jan Berka, and Kenton L. Lohman. A massively parallel PicoTiterPlate™ based platform for discrete picoliter-scale polymerase chain reactions. *Electrophoresis*, 2003. ISSN 01730835. doi: 10.1002/elps.200305646.

Christopher Lee. Generating consensus sequences from partial order multiple sequence alignment graphs. *Bioinformatics*, 2003. ISSN 13674803. doi: 10.1093/bioinformatics/btg109.

Christopher Lee, Catherine Grasso, and Mark F. Sharlow. Multiple sequence alignment using partial order graphs. *Bioinformatics*, 2002. ISSN 13674803. doi: 10.1093/bioinformatics/18.3.452.

Chenhao Li, Kern Rei Chng, Esther Jia Hui Boey, Amanda Hui Qi Ng, Andreas Wilm, and Niranjan Nagarajan. INC-Seq: Accurate single molecule reads using nanopore sequencing. *GigaScience*, 2016. ISSN 2047217X. doi: 10.1186/s13742-016-0140-7.

Heng Li. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 2011. ISSN 13674803. doi: 10.1093/bioinformatics/btr509.

Heng Li. Exploring single-sample snp and indel calling with whole-genome de novo assembly. *Bioinformatics*, 2012. ISSN 13674803. doi: 10.1093/bioinformatics/bts280.

Heng Li. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, 2018. ISSN 14602059. doi: 10.1093/bioinformatics/bty191.

Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 2009. ISSN 13674803. doi: 10.1093/bioinformatics/btp324.

Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 2009. ISSN 13674803. doi: 10.1093/bioinformatics/btp352.

Sheng Li, Scott W. Tighe, Charles M. Nicolet, Deborah Grove, Shawn Levy, William Farmerie, Agnes Viale, Chris Wright, Peter A. Schweitzer, Yuan Gao, Dewey Kim, Joe Boland, Belynda Hicks, Ryan Kim, Sagar Chhangawala, Nadereh Jafari, Nalini Raghavachari, Jorge Gandara, Natàlia Garcia-Reyero, Cynthia Hendrickson, David Roberson, Jeffrey A. Rosenfeld, Todd Smith, Jason G. Underwood, May Wang, Paul Zumbo, Don A. Baldwin, George S. Grills, and Christopher E. Mason. Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study. *Nature Biotechnology*, 2014. ISSN 15461696. doi: 10.1038/nbt.2972.

You-Chun Li, Abraham B. Korol, Tzion Fahima, Avigdor Beiles, and Eviatar Nevo. Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. *Molecular Ecology*, 11(12):2453–2465, 2002. doi: 10.1046/j.1365-294X.2002.01643.x. URL `https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1365-294X.2002.01643.x`.

Lin Liu, Yinhu Li, Siliang Li, Ni Hu, Yimin He, Ray Pong, Danni Lin, Lihua Lu, and Maggie Law. Comparison of next-generation sequencing systems, 2012. ISSN 11107243.

Xin Liu, Dandan Feng, Xueyun Huo, Xiaoqin Xiao, and Zhenwen Chen. Association of intron microsatellite status and exon mutational profiles of TP53 in human colorectal cancer. *Experimental and Therapeutic Medicine*, 2019. ISSN 1792-0981. doi: 10.3892/etm.2019.8095.

Po Ru Loh, Petr Danecek, Pier Francesco Palamara, Christian Fuchsberger, Yakir A. Reshef, Hilary K. Finucane, Sebastian Schoenherr, Lukas Forer, Shane McCarthy, Goncalo R. Abecasis, Richard Durbin, and Alkes L. Price. Reference-based phasing using the Haplotype Reference Consortium panel. *Nature Genetics*, 2016. ISSN 15461718. doi: 10.1038/ng.3679.

Nicholas J. Loman, Raju V. Misra, Timothy J. Dallman, Chrystala Constantinidou, Saheer E. Gharbia, John Wain, and Mark J. Pallen. Performance comparison of benchtop high-throughput sequencing platforms. *Nature Biotechnology*, 2012. ISSN 10870156. doi: 10.1038/nbt.2198.

Nicholas J. Loman, Joshua Quick, and Jared T. Simpson. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nature Methods*, 2015. ISSN 15487105. doi: 10.1038/nmeth.3444.

Erick W. Loomis, John S. Eid, Paul Peluso, Jun Yin, Luke Hickey, David Rank, Sarah McCalmon, Randi J. Hagerman, Flora Tassone, and Paul J. Hagerman. Sequencing the unsequenceable: Expanded CGG-repeat alleles of the fragile X gene. *Genome Research*, 2013. ISSN 10889051. doi: 10.1101/gr.141705.112.

I. López-Flores and M. A. Garrido-Ramos. The repetitive DNA content of eukaryotic genomes. *Genome Dynamics*, 2012. ISSN 16609263. doi: 10.1159/000337118.

Hengyun Lu, Francesca Giordano, and Zemin Ning. Oxford Nanopore MinION Sequencing and Genome Assembly, 2016. ISSN 22103244.

Ruibang Luo, Fritz J. Sedlazeck, Tak Wah Lam, and Michael C. Schatz. A multi-task convolutional deep neural network for variant calling in single molecule sequencing. *Nature Communications*, 2019. ISSN 20411723. doi: 10.1038/s41467-019-09025-z.

Medhat Mahmoud, Nastassia Gobet, Diana Ivette Cruz-Dávalos, Ninon Mounier, Christophe Dessimoz, and Fritz J. Sedlazeck. Structural variant calling: The long and the short of it, 2019. ISSN 1474760X.

Umberto Malapelle, Elena Vigliar, Roberta Sgariglia, Claudio Bellevicine, Lorenzo Colarossi, Domenico Vitale, Pierlorenzo Pallante, and Giancarlo Troncone. Ion Torrent next-generation sequencing for routine identification of clinically relevant mutations in colorectal cancer patients. *Journal of Clinical Pathology*, 2015. ISSN 14724146. doi: 10.1136/jclinpath-2014-202691.

Elizabeth A. Manrao, Ian M. Derrington, Andrew H. Laszlo, Kyle W. Langford, Matthew K. Hopper, Nathaniel Gillgren, Mikhail Pavlenok, Michael Niederweis, and Jens H. Gundlach. Reading DNA at single-nucleotide resolution with a mutant MspA nanopore and phi29 DNA polymerase. *Nature Biotechnology*, 2012. ISSN 10870156. doi: 10.1038/nbt.2171.

Tuomo Mantere, Simone Kersten, and Alexander Hoischen. Long-read sequencing emerging in medical genetics, 2019. ISSN 16648021.

Rajiv C. McCoy, Ryan W. Taylor, Timothy A. Blauwkamp, Joanna L. Kelley, Michael Kertesz, Dmitry Pushkarev, Dmitri A. Petrov, and Anna Sophie Fiston-Lavier. Illumina TruSeq synthetic long-reads empower de novo assembly and resolve complex, highly-repetitive transposable elements. *PLoS ONE*, 2014. ISSN 19326203. doi: 10.1371/journal.pone.0106689.

B. J. McKenzie, R. Harries, and T. Bell. Selecting a hashing algorithm. *Software: Practice and Experience*, 1990. ISSN 1097024X. doi: 10.1002/spe.4380200207.

Angelika Merkel and Neil Gemmell. Detecting short tandem repeats from genome data: Opening the software black box. *Briefings in Bioinformatics*, 2008. ISSN 14675463. doi: 10.1093/bib/bbn028.

Michael L. Metzker. Emerging technologies in DNA sequencing, 2005. ISSN 10889051.

Sharon K. Michelhaugh, Carolyn Fiskerstrand, Elizabeth Lovejoy, Michael J. Bannon, and John P. Quinn. The dopamine transporter gene (SLC6A3) variable number of tandem repeats domain enhances transcription in dopamine neurons. *Journal of Neurochemistry*, 2001. ISSN 00223042. doi: 10.1046/j.1471-4159.2001.00647.x.

Alla Mikheenko, Andrey Prjibelski, Vladislav Saveliev, Dmitry Antipov, and Alexey Gurevich. Versatile genome assembly evaluation with QUAST-LG. In *Bioinformatics*, 2018. doi: 10.1093/bioinformatics/bty266.

André E. Minoche, Juliane C. Dohm, and Heinz Himmelbauer. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. *Genome Biology*, 2011. ISSN 14747596. doi: 10.1186/gb-2011-12-11-r112.

Nima Mousavi, Sharona Shleizer-Burko, Richard Yanicky, and Melissa Gymrek. Profiling the genome-wide landscape of tandem repeat expansions. *Nucleic acids research*, 2019. ISSN 13624962. doi: 10.1093/nar/gkz501.

Wenbo Mu, Hsiao Mei Lu, Jefferey Chen, Shuwei Li, and Aaron M. Elliott. Sanger Confirmation Is Required to Achieve Optimal Sensitivity and Specificity in Next-Generation Sequencing Panel Testing. *Journal of Molecular Diagnostics*, 2016. ISSN 19437811. doi: 10.1016/j.jmoldx.2016.07.006.

Martin D. Muggli, Simon J. Puglisi, Roy Ronen, and Christina Boucher. Misassembly detection using paired-end sequence reads and optical mapping data. In *Bioinformatics*, 2015. doi: 10.1093/bioinformatics/btv262.

Kensuke Nakamura, Taku Oshima, Takuya Morimoto, Shun Ikeda, Hirofumi Yoshikawa, Yuh Shiwa, Shu Ishikawa, Margaret C. Linak, Aki Hirai, Hiroki Takahashi, Md Altaf-Ul-Amin, Naotake Ogasawara, and Shigehiko Kanaya. Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Research*, 2011. ISSN 03051048. doi: 10.1093/nar/gkr344.

Yusuke Nakamura, Mark Leppert, Peter O'Connell, Roger Wolff, Tom Holm, Melanie Culver, Cindy Martin, Esther Fujimoto, Mark Hoff, Erika Kumlin, and Ray White. Variable number of tandem repeat (VNTR) markers for human gene mapping. *Science*, 1987. ISSN 00368075. doi: 10.1126/science.3029872.

Yusuke Nakamura, Kumiko Koyama, and Mieko Matsushima. VNTR (variable number of tandem repeat) sequences as transcriptional, translational, or functional regulators. *Journal of Human Genetics*, 1998. ISSN 14345161. doi: 10.1007/s100380050059.

Ö. Ufuk Nalbantoğlu. *Dynamic Programming*, pages 3–27. Humana Press, Totowa, NJ, 2014. ISBN 978-1-62703-646-7. doi:

10.1007/978-1-62703-646-7_1. URL https://doi.org/10.1007/978-1-62703-646-7_1.

Karl Näslund, Peter Saetre, Jenny Von Salomé, Tomas F. Bergström, Niclas Jareborg, and Elena Jazin. Genome-wide prediction of human VNTRs. *Genomics*, 2005. ISSN 08887543. doi: 10.1016/j.ygeno.2004.10.009.

Rasmus Nielsen, Joshua S. Paul, Anders Albrechtsen, and Yun S. Song. Genotype and SNP calling from next-generation sequencing data, 2011. ISSN 14710056.

Jafar Nouri Nojadeh, Shahin Behrouz Sharif, and Ebrahim Sakhinia. Microsatellite instability in colorectal cancer, 2018. ISSN 16112156.

Michael Nothnagel, Alexander Herrmann, Andreas Wolf, Stefan Schreiber, Matthias Platzer, Reiner Siebert, Michael Krawczak, and Jochen Hampe. Technology-specific error signatures in the 1000 Genomes Project data. *Human Genetics*, 2011. ISSN 03406717. doi: 10.1007/s00439-011-0971-3.

Pål Nyrén. The history of Pyrosequencing®. *Methods in Molecular Biology*, 2015. ISSN 10643745. doi: 10.1007/978-1-4939-2715-9_1.

Yukiteru Ono, Kiyoshi Asai, and Michiaki Hamada. PBSIM: PacBio reads simulator - Toward accurate genome assembly. *Bioinformatics*, 2013. ISSN 13674803. doi: 10.1093/bioinformatics/bts649.

Peter J. Park. ChIP-seq: Advantages and challenges of a maturing technology, 2009. ISSN 14710056.

Murray D. Patterson, Tobias Marschall, Nadia Pisanti, Leo Van Iersel, Leen Stougie, Gunnar W. Klau, and Alexander Schönhuth. WhatsHap: Weighted Haplotype Assembly for Future-Generation Sequencing Reads. *Journal of Computational Biology*, 2015. ISSN 10665277. doi: 10.1089/cmb.2014.0157.

Brent S. Pedersen and Aaron R. Quinlan. Mosdepth: Quick coverage calculation for genomes and exomes. *Bioinformatics*, 2018. ISSN 14602059. doi: 10.1093/bioinformatics/btx699.

Yuri Pirola, Simone Zaccaria, Riccardo Dondi, Gunnar W. Klau, Nadia Pisanti, and Paola Bonizzoni. HapCol: Accurate and memory-efficient haplotype assembly from long reads. *Bioinformatics*, 2016. ISSN 14602059. doi: 10.1093/bioinformatics/btv495.

Ryan Poplin, Pi Chuan Chang, David Alexander, Scott Schwartz, Thomas Colthurst, Alexander Ku, Dan Newburger, Jojo Dijamco, Nam Nguyen, Pegah T. Afshar, Sam S. Gross, Lizzie Dorfman, Cory Y. McLean, and Mark A. Depristo. A universal snp and small-indel variant caller using deep neural networks. *Nature Biotechnology*, 2018. ISSN 15461696. doi: 10.1038/nbt.4235.

David Porubský, Ashley D. Sanders, Niek Van Wietmarschen, Ester Falconer, Mark Hills, Diana C.J. Spierings, Marianna R. Bevova, Victor Guryev, and Peter M. Lansdorp. Direct chromosome-length haplotyping by single-cell sequencing. *Genome Research*, 2016. ISSN 15495469. doi: 10.1101/gr.209841.116.

David Porubsky, Peter Ebert, Peter A Audano, Mitchell R Vollger, William T Harvey, Katherine M Munson, Melanie Sorensen, Arvis Sulovari, Marina Haukness, Maryam Ghareghani, Peter M Lansdorp, Benedict Paten, Scott E Devine, Ashley D Sanders, Charles Lee, Mark J P Chaisson, Jan O Korbel, Evan E Eichler, and Tobias Marschall. A fully phased accurate assembly of an individual human genome. *bioRxiv*, 2019. doi: 10.1101/855049.

Michael A. Quail, Miriam Smith, Paul Coupland, Thomas D. Otto, Simon R. Harris, Thomas R. Connor, Anna Bertoni, Harold P. Swerdlow, and Yong Gu. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, 2012. ISSN 14712164. doi: 10.1186/1471-2164-13-341.

Tobias Rausch, Markus Hsi-Yang Fritz, Jan O. Korbel, and Vladimir Benes. Alfred: Interactive multi-sample BAM alignment statistics, feature counting and feature annotation for long- and short-read sequencing. *Bioinformatics*, 2019. ISSN 14602059. doi: 10.1093/bioinformatics/bty1007.

Guy-Franck Richard, Alix Kerrest, and Bernard Dujon. Comparative Genomics and Molecular Dynamics of DNA Repeats in Eukaryotes. *Microbiology and Molecular Biology Reviews*, 2008. ISSN 1092-2172. doi: 10.1128/mmbr.00011-08.

Nora Rieber, Marc Zapatka, Bärbel Lasitschka, David Jones, Paul Northcott, Barbara Hutter, Natalie Jäger, Marcel Kool, Michael Taylor, Peter Lichter, Stefan Pfister, Stephan Wolf, Benedikt Brors, and Roland Eils. Coverage Bias and Sensitivity of Variant Calling for Four Whole-genome Sequencing Technologies. *PLoS ONE*, 2013. ISSN 19326203. doi: 10.1371/journal.pone.0066621.

T. Rognes and E. Seeberg. Six-fold speed-up of Smith-Waterman sequence database searches using parallel processing on common microprocessors. *Bioinformatics*, 2000. ISSN 13674803. doi: 10.1093/bioinformatics/16.8.699.

M. Katharine Rudd and Huntington F. Willard. Analysis of the centromeric regions of the human genome assembly, 2004. ISSN 01689525.

Leena Salmela, Riku Walve, Eric Rivals, Esko Ukkonen, and Cenk Sahinalp. Accurate self-correction of errors in long reads using de Bruijn graphs. *Bioinformatics*, 2017. ISSN 14602059. doi: 10.1093/bioinformatics/btw321.

F. Sanger and A. R. Coulson. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*, 1975. ISSN 00222836. doi: 10.1016/0022-2836(75)90213-2.

F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 1977. ISSN 00278424. doi: 10.1073/pnas.74.12.5463.

Michael C. Schatz, Arthur L. Delcher, and Steven L. Salzberg. Assembly of large genomes using second-generation sequencing, 2010. ISSN 10889051.

Monika H.M. Schmidt and Christopher E. Pearson. Disease-associated repeat instability and mismatch repair, 2016. ISSN 15687856.

Armin O. Schmitt and Hanspeter Herzel. Estimating the entropy of DNA sequences. *Journal of Theoretical Biology*, 1997. ISSN 00225193. doi: 10.1006/jtbi.1997.0493.

Fritz J. Sedlazeck, Philipp Rescheneder, Moritz Smolka, Han Fang, Maria Nattestad, Arndt Von Haeseler, and Michael C. Schatz. Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods*, 2018. ISSN 15487105. doi: 10.1038/s41592-018-0001-7.

Sandeep N. Shah, Suzanne E. Hile, and Kristin A. Eckert. Defective mismatch repair, microsatellite mutation bias, and variability in clinical cancer phenotypes, 2010. ISSN 00085472.

C. E. Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal*, 1948. ISSN 15387305. doi: 10.1002/j.1538-7305.1948.tb01338.x.

Barton E. Slatko, Andrew F. Gardner, and Frederick M. Ausubel. Overview of Next-Generation Sequencing Technologies. *Current Protocols in Molecular Biology*, 122(1), 2018. ISSN 19343647. doi: 10.1002/cpmb.59.

Haibao Tang, Ewen F. Kirkness, Christoph Lippert, William H. Biggs, Martin Fabani, Ernesto Guzman, Smriti Ramakrishnan, Victor Lavrenko, Boyko Kakaradov, Claire Hou, Barry Hicks, David Heckerman, Franz J. Och, C. Thomas Caskey, J. Craig Venter, and Amalio Telenti. Profiling of Short-Tandem-Repeat Disease Alleles in 12,632 Human Whole Genomes. *American Journal of Human Genetics*, 2017. ISSN 15376605. doi: 10.1016/j.ajhg.2017.09.013.

Rick M. Tankard, Mark F. Bennett, Peter Degorski, Martin B. Delatycki, Paul J. Lockhart, and Melanie Bahlo. Detecting Expansions of Tandem Repeats in Cohorts Sequenced with Short-Read Sequencing Data. *American Journal of Human Genetics*, 2018. ISSN 15376605. doi: 10.1016/j.ajhg.2018.10.015.

Svetlana Temnykh, Genevieve DeClerck, Angelika Lukashova, Leonard Lipovich, Samuel Cartinhour, and Susan McCouch. Computational and experimental analysis of microsatellites in rice (Oryza sativa L.): Frequency, length variation, transposon associations, and genetic marker potential. *Genome Research*, 2001. ISSN 10889051. doi: 10.1101/gr.184001.

Ryan Tewhey, Vikas Bansal, Ali Torkamani, Eric J. Topol, and Nicholas J. Schork. The importance of phase information for human genomics, 2011. ISSN 14710056.

T. Thiel, W. Michalek, R. K. Varshney, and A. Graner. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.). *Theoretical and Applied Genetics*, 2003. ISSN 00405752. doi: 10.1007/s00122-002-1031-0.

Julie D. Thompson, Desmond G. Higgins, and Toby J. Gibson. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 1994. ISSN 03051048. doi: 10.1093/nar/22.22.4673.

Helga Thorvaldsdóttir, James T. Robinson, and Jill P. Mesirov. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, 2013. ISSN 14675463. doi: 10.1093/bib/bbs017.

Ole K. Tørresen, Bastiaan Star, Pablo Mier, Miguel A. Andrade-Navarro, Alex Bateman, Patryk Jarnot, Aleksandra Gruca, Marcin Grynberg, Andrey V. Kajava, Vasilis J. Promponas, Maria Anisimova, Kjetill S. Jakobsen, and Dirk Linke. Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. *Nucleic acids research*, 2019. ISSN 13624962. doi: 10.1093/nar/gkz841.

Todd J. Treangen and Steven L. Salzberg. Repetitive DNA and next-generation sequencing: Computational challenges and solutions, 2012. ISSN 14710056.

Durdica Ugarkovic. Functional elements residing within satellite DNAs, 2005. ISSN 1469221X.

Ajay Ummat and Ali Bashir. Resolving complex tandem repeats with long reads. *Bioinformatics*, 2014. ISSN 14602059. doi: 10.1093/bioinformatics/btu437.

Anton Valouev, Jeffrey Ichikawa, Thaisan Tonthat, Jeremy Stuart, Swati Ranade, Heather Peckham, Kathy Zeng, Joel A. Malek, Gina Costa, Kevin McKernan, Arend Sidow, Andrew Fire, and Steven M. Johnson. A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. *Genome Research*, 2008. ISSN 10889051. doi: 10.1101/gr.076463.108.

Erwin L. van Dijk, Yan Jaszczyszyn, Delphine Naquin, and Claude Thermes. The Third Revolution in Sequencing Technology, 2018. ISSN 13624555.

Robert Vaser, Ivan Sović, Niranjan Nagarajan, and Mile Šikić. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Research*, 2017. ISSN 15495469. doi: 10.1101/gr.214270.116.

J. C. Venter, H. O. Smith, and L. Hood. A new strategy for genome sequencing, 1996. ISSN 00280836.

G. Vergnaud, D. Gauguier, J. J. Schott, D. Lepetit, V. Lauthier, D. Mariat, and J. Buard. Detection, cloning, and distribution of minisatellites in some mammalian genomes. *EXS*, 1993. ISSN 1023294X. doi: 10.1007/978-3-0348-8583-6_4.

Ayelet Voskoboynik, Norma F. Neff, Debashis Sahoo, Aaron M. Newman, Dmitry Pushkarev, Winston Koh, Benedetto Passarelli, H. Christina Fan, Gary L. Mantalas, Karla J. Palmeri, Katherine J. Ishizuka, Carmela Gissi, Francesca Griggio, Rachel Ben-Shlomo, Daniel M. Corey, Lolita Penland, Richard A. White, Irving L. Weissman, and Stephen R. Quake. The genome sequence of the colonial chordate, Botryllus schlosseri. *eLife*, 2013. ISSN 2050084X. doi: 10.7554/eLife.00569.

Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: A revolutionary tool for transcriptomics, 2009. ISSN 14710056.

Michael S. Webster and Letitia Reichart. Use of microsatellites for parentage and kinship analyses in animals. *Methods in Enzymology*, 2005. ISSN 00766879. doi: 10.1016/S0076-6879(05)95014-3.

Joachim Weischenfeldt, Orsolya Symmons, François Spitz, and Jan O. Korbel. Phenotypic impact of genomic structural variation: Insights from and for human disease, 2013. ISSN 14710056.

Aaron M. Wenger, Paul Peluso, William J. Rowell, Pi Chuan Chang, Richard J. Hall, Gregory T. Concepcion, Jana Ebler, Arkarachai Fungtammasan, Alexey Kolesnikov, Nathan D. Olson, Armin Töpfer, Michael Alonge, Medhat Mahmoud, Yufeng Qian, Chen Shan Chin, Adam M. Phillippy, Michael C. Schatz, Gene Myers, Mark A. DePristo, Jue Ruan, Tobias Marschall, Fritz J. Sedlazeck, Justin M. Zook, Heng Li, Sergey Koren, Andrew Carroll, David R. Rank, and Michael W. Hunkapiller. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*, 2019. ISSN 15461696. doi: 10.1038/s41587-019-0217-9.

Thomas Wicker, François Sabot, Aurélie Hua-Van, Jeffrey L. Bennetzen, Pierre Capy, Boulos Chalhoub, Andrew Flavell, Philippe Leroy, Michele Morgante, Olivier Panaud, Etienne Paux, Phillip SanMiguel, and Alan H. Schulman. A unified classification system for eukaryotic transposable elements, 2007. ISSN 14710064.

Kai Ye, Marcel H. Schulz, Quan Long, Rolf Apweiler, and Zemin Ning. Pindel: A pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, 2009. ISSN 13674803. doi: 10.1093/bioinformatics/btp394.

Degui Zhi, Jihua Wu, Nianjun Liu, and Kui Zhang. Genotype calling from next-generation sequencing data using haplotype information of reads. *Bioinformatics*, 2012. ISSN 13674803. doi: 10.1093/bioinformatics/bts047.