Francesco Andreace

Can I decide it or should I put the official one?

Analysis of human pangenome graphs using k-mer-based applications

Thesis submitted for the degree of Philosophiae Doctor

Department of Computational Biology
Insitut Pasteur Paris, Universite' de Paris Cite

Edite doctoral school, Sorbonne Université

2024

Series of dissertations submitted to the
Insitut Pasteur Paris, Universite' de Paris Cite, Sorbonne Universite'
No. 1234

ISSN 1234-5678

Cover: TBD - in PHDUIO.cls.
Print production: Pasteur Paris.

To Sofia, my sweet old cat that lives so well without overthinking.

Preface

This thesis is submitted in partial fulfillment of the requirements for the degree of *Philosophiae Doctor* at Sorbonne Université. The research presented here was conducted at the Insitut Pasteur, under the supervision of Dr. Rayan Chikhi and Dr. Yoann Dufresne. This work was supported by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 956 229 (+Panagaia +Pasteur + INCEPTION).

The thesis is a collection of the different projects I worked on during my stay at Insitut Pasteur. I begin with an small foreword of the research output of my PhD, and a gentle introduction of the scientific concepts needed to understand the rest of the manuscript. In the first section I present the published paper I am first author of, together with other unpublished work I lead or independently developed. In the second section I present other results of my scientific production, with novel elaboration of the work that appeared in the other papers I am co-author and presentation of projects that hava or will be submitted to revision. The common theme is pangenomics and computational methods used to generate and use such models to infer relevant information. This essay ends with a chapter showcasing future perspectives and conclusions.

Acknowledgements

Thanks for spending time reading this.

Francesco Andreace
Paris, September 2024

Contents

Preface iii

Contents v

List of Figures vii

List of Tables ix

Introduction 1

Papers 16

Appendices 17

List of Figures

0.1 The DNA molecule. 2
0.2 Third generation sequencing technologies. 5
0.3 Inter-individual and inter-population variation for 4 primate species. 9
0.4 Genomic difference in chromosome 7 and 16 of 5 primate species. 10
0.5 Spectrum of Human Genetic Variation. 11

List of Tables

0.1 k-mers with $k = 3$ being computed from the sequence $S =' CTGAACTACA'$. with $l = 10$. At the end, $l - k + 1 = 8$ k-mers are generated. 7

0.2 Example of canonical k-mers enumeration and count. Given a set of sequences, for each of them k-mers are computed in a stream. For each of them, on the fly, the reverse complement is computed. Then the ones that are considered canonicals are passed and counted. In the table below, reverse complements are between parenthesis and the canonical between the two is underlined. . . 8

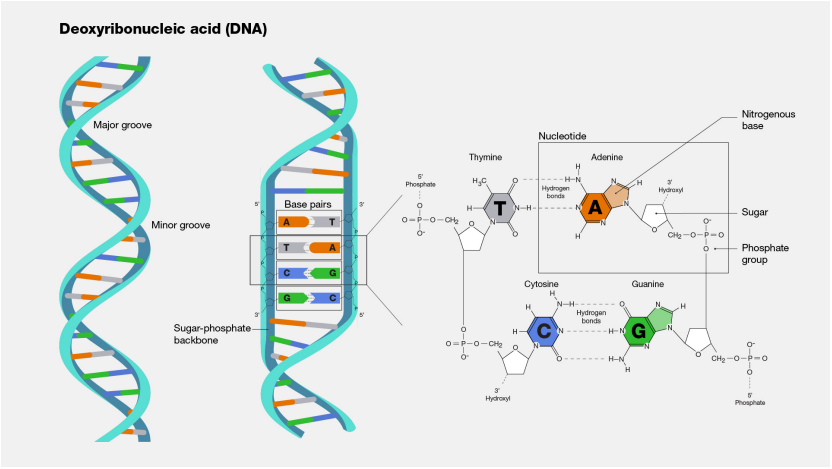
Introduction

A fundamental grasp of the data that is produced by sequencing biological organisms is essential to comprehend the research outlined in this manuscript. If already familiar with DNA sequences, how they are obtained and how they differ between species or individuals, you may proceed to section *From reads to* k-*mers*.

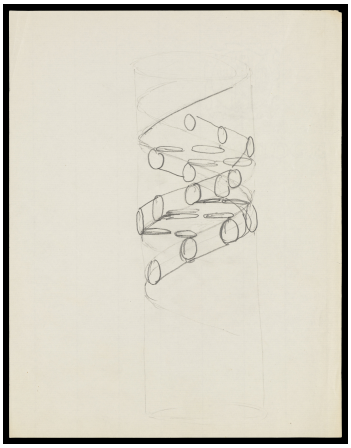
DNA, genome variation and sequencing data

DNA (Deoxyribonucleic Acid) is a complex molecule with a double helix structure that carries the genetic information of an organism. Although its discovery was the result of work by many scientists over nearly 90 years, the currently accepted model was first correctly described by James Watson and Francis Crick in 1953 at Cambridge, UK.
The information DNA carries provides instructions for an organism to develop, survive in the external world, and reproduce. These instructions are encoded as a sequence of monomers called nucleotides. Each nucleotide is composed of a sugar, a phosphate group, and one of four nucleobases: cytosine, guanine, adenine, and thymine. The nucleotides are commonly referred to using the first letter of their nucleobases: A, C, G, and T. In RNA molecules, thymine is replaced by uracil. The nucleotides are linked together in a sugar-phosphate backbone. Hydrogen bonds between complementary nucleotides form the molecule's double-stranded structure, with A pairing with T and C pairing with G bases. This pairing is crucial for DNA replication and protein synthesis. Figure 0.1 shows the structure of the DNA molecule and the nucleotides, with the initial drawing by Francis Crick in 1953.

To fit inside the cell nucleus, DNA is organized in very tight structures. First is coiled around proteins called histones to from a compact structure called chromatin that form loops and is kept in place by other molecules to structure a chromosome. In humans, each nucleus contains 23 chromosome pairs. 22 are the autosomes, i.e. the chromosomes we all have and that are not associated with sex, while the last pair is the sex one that contains or 2 copies of chromosome X for women or 1 X and 1 Y for men. The final part of the chromosome is called telomere while the central one is called centromere and both are regions known to contain a lot of repetitive regions that are very difficult to reconstruct from sequencing.
Finally, there is also the mitochondrial genome, that is not in the nucleus, has circular structure and is mostly inherited maternally.



(a) The DNA molecule and the structure of the nucleotides, the basic piece of information of the DNA. Figure from NIH glossary [1].



(b) The DNA molecule model draw by Francis Crick in 1953.

Figure 0.1: The DNA molecule.

DNA sequencing

In many biological disciplines, studying an organism's genetic information contained in its DNA is crucial. Over the years, researchers have developed various methods and techniques to extract this information from the cell nucleus and to represent it in a useful way. These processes typically involve three main steps. Here I describe them, with many simplifications, to give a brief overview:

Library preparation The first step requires hours long, nontrivial biological manipulation of samples to extract DNA from cell nucleus and purify it without causing damage. This process isolates the genetic material from other

cellular components, like RNA and proteins. The DNA molecule are fragmented into pieces of different length followed by 5′ and 3′ adapter ligation. Some technologies require PCR amplification of fragments, while others don't.

Sequencing Next, specialized machines detect the sequence of nucleotides that compose the extracted DNA pieces. These techniques, called sequencing, use various, most of the time proprietary, technologies to determine the precise order of nucleotides (A, C, G, and T) in a DNA molecule. The raw data output of these machines are sequences of characters that are referred to as sequencing reads or simply reads.

Analysis In this step usually quality control (QC) is performed to remove adapters and too short or low quality reads. Usually the first step after QC is to assemble the sequences together or to provide them as input to a workflow specific for the required application.

The landscape of DNA sequencing has evolved significantly since its inception. In 1977, Frederick Sanger and his colleagues introduced the first widely adopted sequencing method, known as chain termination sequencing or Sanger sequencing[2]. This technique allowed to read the sequence of nucleotides in a DNA molecule for the first time in a reliable and reproducible manner. This was the technique that led to the first sequencing of the mitochondrial DNA and the first ,almost, complete human genome in 2001 [3, 4]. While Sanger sequencing has revolutionized genetic research, it has largely been replaced by more advanced technologies. These newer methods fall into two main categories: Next Generation Sequencing (NGS) and Third Generation Sequencing. These technologies provide significant improvements in terms of speed, cost-effectiveness and data output compared to Sanger sequencing.

Next Generation Sequencing

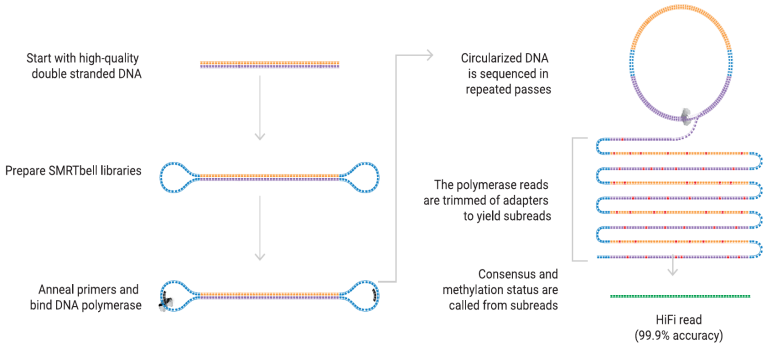
(NGS) derives its name by launching a so-called next generation by revolutionizing sequencing with massive parallelization. This technology has continuously improved since 2005 to yield up to 8 Terabases per single sequencing run, taking it maximum 2 days and dropping the price of, for example, a single individual sequenced per almost 100 dollars [5]. The advancement consists mainly in running many reactions and analysis in parallel to produce millions to billions of reads of a length that varies between 150 and 300 bases. For this reason they take the name of short reads. While a big advantage of this sequencing method is the low error rates, with at least 80% of the bases with less than 1 error in 1000 (i.e. 99.9% accuracy). This technology is mostly dominated by a California biotechnology company called Illumina
The sequence length is the main drawback of this method, as it makes it difficult or impossible to resolve complex large-size DNA variations, where the read won't align to any part of a reference genome to be used to infer from which

part of a chromosome it comes from. The same problem arise for repetitive regions,like centromeres, telomeres or small segmental duplications, that have a length greater than the sequence length, where it is not possible to asses the length or the nature of the repetition. This problem has been partially addressed by the introduction of pair-end sequences, a technology that is now integrated in all Next Generation machines, that sequences both ends of a single DNA fragment and then associate the two reads that come from it, in order to provide more long-range information. Although this method is still not enough to solve complex structural variations or repetitive regions, it is very useful to track some of the reads that would be instead be unused and finds relevant applications in assembly or metagenomics. In fact, I used this property of paired-end reads in one method I developed before the PhD to improve the estimation of different species inside metagenomic samples sequenced with Next Generation Sequencing [6].

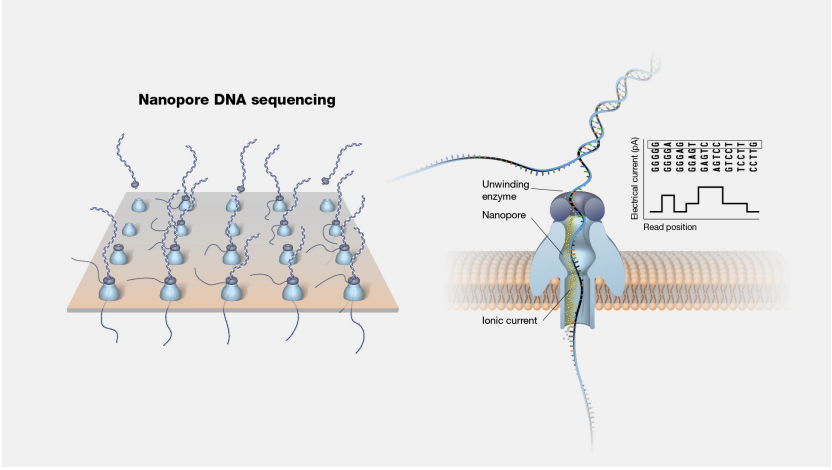
Third Generation Sequencing

(TGS) is the newest technology that uses alternative approaches to NGS, to solve the issues that it currently face due to the short length of the sequences. The main difference relies on the fact that while NGS uses PCR to amplify the small fragments in which DNA is broken into prior to sequencing, these new technologies directly sequence the DNA molecule. Here I will describe the two most important technologies, provided by the two companies that lead this market: a California biotech company called Pacific Biosciences, usually called PacBio, and a Uk based one, called Oxford Nanopore Technologies, or ONT.
PacBio offers "HiFi sequecing" that produces reads long up to 25 thousands bases in length with accuracy comparable to NGS ones. This is achieved by first creating a circularized DNA from high-quality double stranded DNA and then using a DNA polymerase enzyme to read multiple times the same molecule to produce a final consensus sequence with accuracy of around 99.9%. These are long and accurate reads that enable ultra-fast assembly of human genomes [7] at a cost around $1000 per sequencing reagents kit.
Oxford Nanopore machines instead provide ultra-long sequences, that are on average longer than the PacBio HiFi ones and can reach up to the megabase scale (i.e. 100 times longer). The sequencing is done by passing a single-strand DNA molecule trough a tiny nanopore. Each pore is associated to an electrode and a sensor that measure the current that is passing through the pore. As the DNA goes through the pore, the current changes and, thanks to a basecalling algorithm, it is possible to detect the nucleotides by the change in the current. This process is done in parallel across 800-1500 pores.
It is finally important to stress that these two technologies allow the detection of all kind of variations, i.e. small variations as well as large ones and also solve large repetitive regions as they span acorss thousands of bases. Moreover, both these methods allow the direct detection of DNA methylation. This is a chemical mechanism on top of the DNA molecule that regulates gene expression

by recruiting proteins involved in gene repression or by inhibiting the binding of transcription factor(s) to DNA [8].
Figure 0.2 shows basic schematics of how these two technologies work.



(a) Pacific Biosciences Hi-Fi reads generations scheme. Image from PacBio website.



(b) An array of pores sequences multiple molecules in parallel. A dsDNA molecule is split by the helicase enzyme and then a ssDNA sequence slowly gets through the pore for sequencing. Changes in the ionic current is used by a machine learning algorithm to infer the nucleotides of the sequence.

Figure 0.2: Third generation sequencing technologies.

From reads to *k*-mers

The sequences produced by any of the aforementioned technologies are considered as text strings, i.e. successions of characters, like the phrases of this manuscript, in which each character correspond to a nucleotide. These sequences

can therefore being stored in plain text formats, like FASTQ, that preserve basecalling quality information or in others, like FASTA, that retains only the actual sequence. In order to use less space and take advantage of redundancy in the sequencing data, these files are often compressed, using one of the many tools publicly available like `gzip` or `zstd`, by Facebook.
As pair-end short-reads have different features than ultra-long reads or Hi-Fi long reads, most of the tools focus on providing applications for just one single type. In cases like assembling a genome from the reads or calling the variant of the sequenced genome compared to one of reference, however, the information from different sources can be combined to provide superior results. In order to generate high-quality genome assemblies, for example, many consortia, like the Human Pangenome Reference Consortium, use Hi-Fi long reads as bases for assembly plus ultra-long reads as scaffolds to chain together the assemblies into sequences that span from telomere to telomere of a chromosome.
In the work presented in this manuscript, most of the tools will ingest as input or raw sequences (both NGS or TGS) or high-quality, near telomere-to-telomere assemblies. Some of the tools that I have have used and all of the ones I have developed or co-developed transform the input sequences or assemblies into k-mers to produce the desired output.

DNA alphabet The DNA alphabet $\sum$ is composed by the 4 characters that compose the first letter of the nucleobases: A,C, G and T: $\sum = A, C, G, T$,

sequence a biological sequence from $\sum$ is defined as $S =\in \sum^{l}, with|S| = l$, with length l that can be fixed, if originated from NGS, or variable, if originated from TGS.

***k*-mer** a k-mer of S is defined as $k - mer \in \sum^{k}, with|k - mer| = k$ i.e. any valid substring of S of length k.

As shown in table 0.1, from any sequence S, it is possible to obtain its constituent k-mers. The length k of a k-mer is an arbitrary value, that is usually chosen depending on the kind of sequences used (cannot have $k > n$), the characteristics of the data that is used (is it from a single organism, a collection of the same species, a collection of different organisms) and on the disk or memory space that is available for computation or storage (as in figure XXX, the longer the k, the more space is used by repetitive characters). A more detailed explanation of these considerations will be provided in section XXX[QF].
As it is possible to retrieve k-mers from a single read, it is trivial to extend this property to any set of reads, for example produced by a single sequencing run of a sample. This means that a set of k-mers is equivalent to the set of reads it is obtained from. In order to characterize this transformation as lossless, i.e. without any loss of information, an association from each k-mer to the read(s) it comes from would be needed. In most of the cases this is not useful and k-mers are obtained from reads without remembering from which reads do they come from. In other, specific, applications it might instead be needed to know in

Position	1	2	3	4	5	6	7	8	9	10
Sequece S	C	T	G	A	A	C	T	A	C	A
k-mers	C	T	G							
		T	G	A						
			G	A	A					
				A	A	C				
					A	C	T			
						C	T	A		
							T	A	C	
								A	C	A

Table 0.1: k-mers with $k = 3$ being computed from the sequence $S =' CTGAACTACA'$. with $l = 10$. At the end, $l - k + 1 = 8$ k-mers are generated.

which reads there are certain k-mers. More considerations on this are going to be presented in section XXX[BackToSequences].
As presented in section the DNA is double-stranded, with A bases are paired with T ones, while C bases are paired with G ones, also called complements. If a k-mer appears in a sequence, in the other strand of the molecule there would be what is called its reverse complement. This is the spelling of the k-mer from the end to the beginning, substituting each base with its complement. For example if in one strand there appear the sequence $ACGT$, on the other strand it would spell $TGCA$.
When enumerating k-mers from a sequence or when storing them, only "canonical" k-mers are kept: this means that for each k-mer produced from a sequence, its reverse-complement is computed and only the one that is considered smaller by a certain property is kept. For example, if the lexicographic order is used, the k-mer (with $k = 4$) $ACGT$ is lexicographically smaller than $TGCA$ so when either of the two is seen, only the first is kept.
Finally, a classic operation that is done when enumerating k-mers from sequences is to keep track of how many times each canonical k-mer appears in the set of sequences. This is called k-mer counting and finds important applications in many genomic disciplines like metagenomics or transcriptomics.

Pangenomics, pangenomes and pangenome graphs

Genomic diversity in populations

The Human genome contains more than 3 billions base pairs and contains probably more than 20 thousands protein coding genes, i.e. specific parts of the DNA that serve as blueprint for proteins. The rest is non-coding, i.e. is not a gene but can serve as regulatory element, like enhancers, promoters and silencers or as other conserved, functional element.
DNA differs between individuals of the same population (inter-individual) and

Table 0.2: Example of canonical k-mers enumeration and count. Given a set of sequences, for each of them k-mers are computed in a stream. For each of them, on the fly, the reverse complement is computed. Then the ones that are considered canonicals are passed and counted.
In the table below, reverse complements are between parenthesis and the canonical between the two is underlined.

Sequence id	sequence
seq1	ACATCA
seq2	CTTCAG
seq3	TACAGC
seq4	GCTTAC

Sequence id	seq1	seq2	seq3	seq4
k-mers	ACA (TGT)	CTT (AAG)	TAC (GTA)	GCT (AGC)
	CAT (ATG)	TTC (GAA)	ACA (TGT)	CTT (AAG)
	ATC (GAT)	TCA (TGA)	CAG (CTG)	TTA (TAA)
	TCA (TGA)	CAG (CTG)	AGC (GCT)	TAC (GTA)

oredered caonical k-mer	count
AAG	2
ACA	2
AGC	2
ATG	1
ATC	1
CAG	2
GAA	1
GTA	2
TAA	1
TCA	2

between different populations of the same species (inter-population): figure 0.3 shows these differences for four close primates. Differences in DNA are given by having a different nucleotide at the same place (also called Single Nucleotide Polymorphism or SNP), small insertions or deletions of bases (also known as indels) and large and complex variations, up to Megabases, (also called Structural Variations, or SVs) that can produce different counts of copies or different ordering of a same region.
On average, each human carries around 10 thousands amino-acid altering mutations, 300-400 gene disruption events (like stop, splice and indels) affecting 200-300 genes and is heterozygous at 50-100 mutations associated with an inherited disorder [9]. This is due genomic variability, that is assessed at Finally, even when close species share a large portion of genetic material, structural changes that rearrange the same material in different order or invert it, contribute to meaningful changes. In figure 0.4 it is shown how the chromosome 7 and 16 of some primates, even if very similar, differs in terms of organization. These large structure rearrangements are thus fundamental to understand the biology of organisms. It is This is because the variation in DNA is produced by two main mechanisms: mutations and recombination.

Moreover, genetic diversity is driven by two main factors: genetic drift and natural selection. Genomic duplication followed by adaptive mutation is considered one of the primary forces for evolution of new function

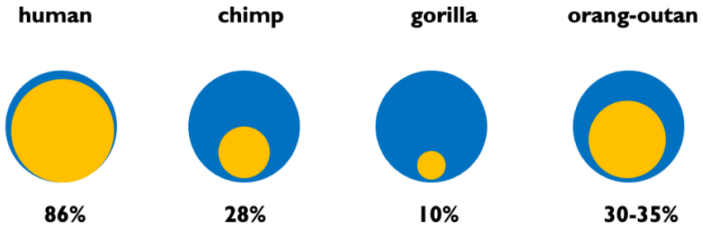


Figure 0.3: Share of inter-individual (yellow) and inter-population (blue) diversity for four different primates. While for humans the majority of the diversity is within populations, for other primates it is between populations. This shows how Humans are more mixed than other primates. Percentage shows the inter-individual variation share [9].

The premisis for human pangenomics

A LINEAR REFERENCE FOR ALL GENOMIC ANALYSES

Since the beginning of genomics, all analysis based on sequencing data depended upon the use of a single linear reference genome, i.e.

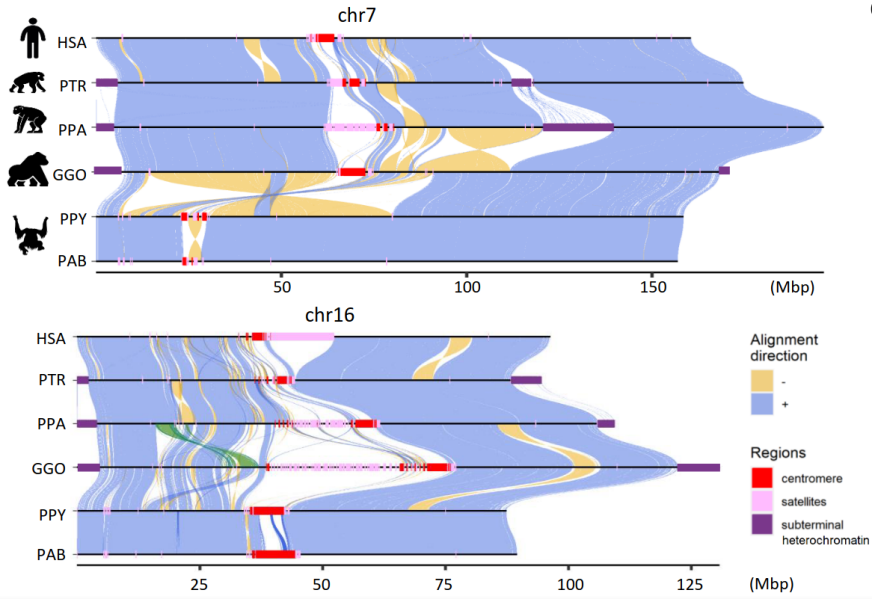


Figure 0.4: A comparative ape alignment of human (HSA) chromosomes 7 and 16 with chimpanzee (PTR), bonobo (PPA), gorilla (GGO),Bornean and Sumatran orangutans (PPY and PAB). The image on the top shows most of the chromosome 7 is conserved except for large inversions happening between the species. The image below shows complex inversions in chromosome 16. Image taken from 'Complete sequencing of ape genomes' [10].

the best assembled genome available for a species, to extract useful information from the DNA. We now know that this approach is suboptimal in a wide range of applications as a lot of genetic material of the species cannot be present in a single linear refererence: this is valid for eukariotes and even more for bacteria. **A SEQUENCING REVOLUTION**

Right now we are witnessing a real revolution in the sequencing. As the price is significantly lowering, also thanks to competition of new companies entering the market, new scientific discoveries and

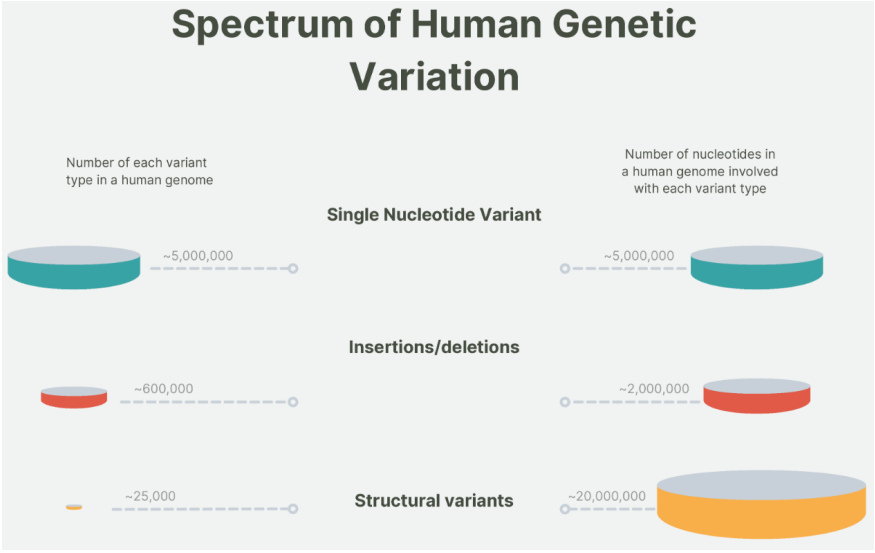


Figure 0.5: Spectrum of Human Genetic Variation. While SNPs are the most common variation Event, their impact in the total amount of bases in a genome is 4 times smaller than the one of Structural Variations, that are 200 times less. Image produced form Evan Eichler slides that are not public.

technological advances are leading to a remarkable increase of quality, in term of per-base error rate, and throughput. This means than right now we dispose of a rich wealth of high quality sequencing information to produce hundreds or thousands of new first grade assemblies.

A QUALITY REVOLUTION This limitation at the beginning was not solvable due to the scarcity of high quality assembled genomes as the technologies of sequencing and computational tools were not mature enough. For example, the Human Genome Project took 13 years to produce

its result [11] and the absence of long reads with decent error rate made it impossible to automatically resolve repetitive regions like telomers and centromeres [12], producing a reference only 92% complete [13]. This problem was only solved in 2022 [13]. At the same time, many consortia are producing increasingly more genomes to a level comparable to the T2T consortium. For example, the HPRC, i.e. the Human Pangenome Reference Consortium released 47 new human genomes (92 haplotypes) in 2021 and has recently released other 153 genomes to a total of 400 haplotypes. The ability to produce such high quality data for human genomes is the main driver of the

Right now we are witnessing a real revolution in the sequencing. As the price is significantly lowering, also thanks to competition of new companies entering the market, new scientific discoveries and technological advances are leading to a remarkable increase of quality, in term of per-base error rate, and throughput. This means than right now we dispose of a rich wealth of high quality sequencing information to produce hundreds or thousands of new first grade assemblies. This progress lead to a shift in paradigm with increasing effort from the

scientific community to propose new methods to analyse one or multiple genomes: not anymore by comparing it against a single reference sequence but against a comprehensive representation of the species.
This novel way to overcome the limits of "linear genomic" and consider all the variation in a single species is called pangenomics.
Various efforts are being made on producing reference pangenomes of yeasts, bacterias, plants and animals, including humans. In order to do so, new tools to construct and then analyse and use such representations are being developed. It is important here to notice, as it will be stressed in the next sections and chapters, that construction is just the first step and that is very important to understand and work on which are the operations that can be succesfully performed by these representations.

Pangenomics

Pangenomes

Pangenome Graphs

Graphs

De Bruijn Graphs

Colored and Compacted De Bruijn Graphs

Variation Graphs

Outline

Papers

Appendices