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 In the first section I present the published paper I am first author of, together with other unpublished work I lead or independently developed.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

List of Figures

List of Tables

Chapter 1

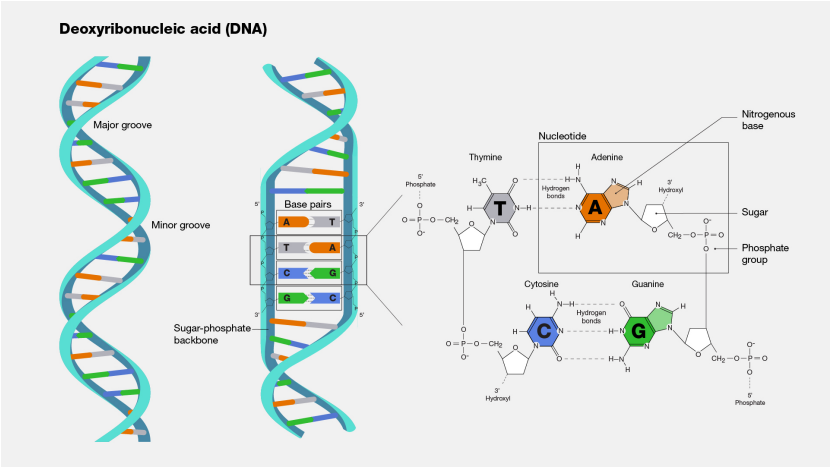
Introduction

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 1.3 *From reads to* k*-mers*.

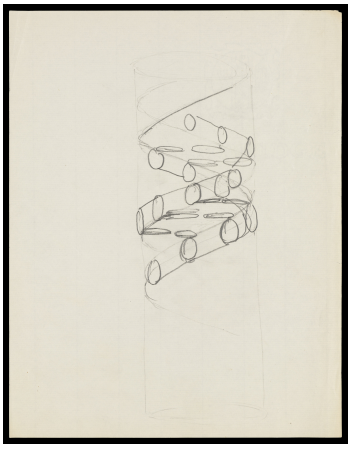
1.1 DNA, genome variation and sequencing data

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 1.1 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. Chromosomes are inherited by the offspring through sexual or asexual reproduction. Humans are diploids, i.e. contain 2 copies of the same chromosome, that receive one copy from the mother (the egg) and one from the father (the sperm). Both the egg and the sperm (the gametes) contain 1 copy of the chromosomes. While mammals are often diploids, other organism can be haploid (one single copy of the chromosome) or polyploid, i.e. have more than 2 copies. For example, the sugar cane plant, the world's most harvested crop by tonnage, can have more than 8 copies of a chromosome, up to 12 [2]. In humans, each nucleus of non-reproductive cells 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



(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 1.1: The DNA molecule.

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.

1.1.1 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: this is genome sequencing, a transformative technology for biology. 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[3]. 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 [4, 5]. Sanger technology through a gel produced the first reference genomes for important organisms. 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.

1.1.1.1 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 [6]. 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 they are too short to assemble into a high-quality *de-novo* complete genome, they are used for *resequencing*, i.e. to be mapped to a reference genome to infer variations from it, to be used, for example, in population variation studies. Additionally, the short length of the fragment makes sequences coming from parts of the genome not in a reference or from complex and/or repetitive regions often impossible to be mapped, loosing all the information associated to them. 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 variations, it is very useful to track some of the reads that would be instead be discarded and finds relevant applications in other fields, like 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 environmental samples sequenced with NGS [7].
Finally, this technology enabled also other kind of sequencing, like STARR-seq, ATAC-seq, ChIP-seq, RNA-seq and others, that enabled to assay regulatory activity in the genome.

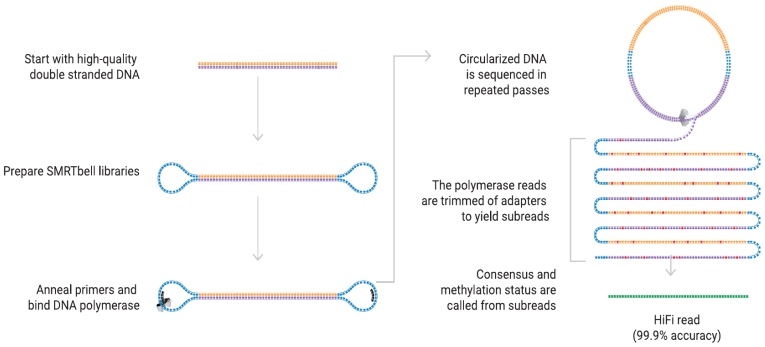
1.1.1.2 Third Generation Sequencing

(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 nucleic acids in their native form. For this reason they are called single molecule technologies.
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

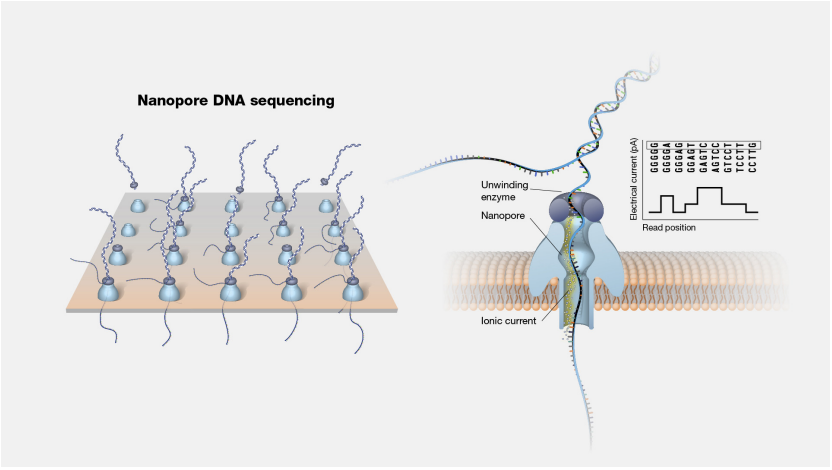
long and accurate reads that enable ultra-fast assembly of human genomes [8] at a cost around $1000 per sequencing reagents kit for a 30X covergae of a human genome.
Oxford Nanopore machines instead provide ultra-long sequences, that are on average longer than the PacBio HiFi ones and can reach up to 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 [9].
Figure 1.2 shows basic schematics of how these two technologies work.

1.2 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.



(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 1.2: Third generation sequencing technologies.

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 1.1, from any sequence S, it is possible to obtain its constituent k-mers. To efficiently extract all k-mers from a sequence, the best approach is to employ a sliding window technique. This is done by identifying the first k-mer at the start of the string and then iteratively shifting the window one position at a time, appending the newly encountered character to the right while removing the leftmost character.
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 table 1.1, 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 which reads there are certain k-mers. More considerations on this are going to be presented in section XXX[BackToSequences].
As presented in section 1.1 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.
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.
k-mers are being used in lots of applications based on NGS short reads while they are less implied on methods for error-prone long reads because using k-mers on one side destroys the long range information provided by reads that span thousands of bases, on the other error-rates higher than NGS would produce too many erroneous k-mers that would be very difficult to correct if not with very deep sequencing, providing additional cost bottlenecks. With Hi-FI reads

and improved quality of nanopore basecalling, it is possible to overcome the error limitation and use k-mers for long reads. One example that uses advanced concepts based on k-mers is the tool `mdbg` that drastically improved assembly of Hi-Fi reads.

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

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

Table 1.1: k-mers with $k = (3, 4)$ being computed from the sequence $S = CTGAACTACA$. $l - k + 1$ k-mers are generated for a total of $(l - k + 1) * k$ bases. While with $k = 3$ the total bases are $8 * 3 = 24$, with $k = 4$ they are instead 28, as larger k encodes more information redundancy.

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

Table 1.2: 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 (by lexicographic order) is underlined.

1.3 Pangenomics, pangenomes and pangenome graphs

1.3.1 Genetic diversity: focus on humans.

The genetic diversity is the variability that exists between organisms at the genetic level, i.e. differences in the information enclosed in their DNA. It is the raw material for biological evolution as, without heritable genetic differences between us, we would not be able to biologically evolve. Here what I will present is valid for humans, as the large part of my work has been with human DNA sequences. Most genetic changes have no effect at all on the individuals carrying them but some can result in phenotypic differences.

1.3.1.1 Causes and drivers of genetic diversity in humans

There are two main mechanisms of genetic diversity: the arise of new mutations and the reshuffling of already present genetic material trough recombinations and duplications. Mutations are produced by physical or chemical damage, for example caused by UV radiation, prior cell division or by errors in the DNA replications during cell division. When this occurs in germinal cells they are transmitted to the offspring, while when happening in a somatic cell (not reproductive), the mutation is not transmitted but can instead be responsible for certain type of cancer. In humans, it is estimated that a newborn carries on average 70 point mutations (one nucleotide substitute with another), 15 from the mother and 55 from the father. The amount of mutations is proportional with the age of the person and, more than induced by replication, it is due to not corrected damage.
On top of the mutations, chunks of chromosomes from the mother and the father chromosomes are shuffled to produce new combinations. The effect of this random process produces the differences between siblings with the same biological parents. Recombination is heterogeneous in the DNA and depends on some motifs that promote higher recombination. Finally, recombination is also influenced by the age, mostly of the mother, as older mothers tend to produce offspring with more misplaced recombinations, also causing the well known trisomy 21.
Without diverging too deep into population genetics, it is also important to understand how new variations are conserved, lost or fixed (become prevalent) in a population. These outcomes are driven by two main factors: genetic drift and natural selection.
Genetic drift is a process, given by the randomness in the individuals that reproduce in a specific population. This can contribute to the loss or fixation of some variants just because of randomness and not because they provide an advantage to the individual. Specifically, in populations with small number of reproductive individuals, this can fixate detrimental variants, while in large populations, the large number of individuals buffers the event.
Natual selection, on the other hand, is a mechanism that explains human

evolution: as genetic variations causes the gain or loss of specific phenotipic traits, these traits can confere positive or negative advantage compared to the rest of the individual in a population (fitness). This phenomenon can contribute selecting certain variations in a population by either contribute to the fixation or the loss of a variant. This mechanism explains our species adaptation to nutritional resources, climate and pathogenes: in 10 thousand years a mutation in a gene that conferred the ability to digest milk has almost got fixed in humans, selection on certain genes explains better adaptations to cold or high altitudes and selection in HbS or DARC alleles has helped humans adapt and survive malaria infections[10].

1.3.1.2 Human Genomic variation: types of variants

There are various types of genomic variants: from the shortest, the single-nucleotide variants (SNV) or Single Nucleotide Polimorphism (SNP) when it is present in at least 1% of the population, is the difference of one nucleobase between two individuals. In a specific part of the genome one person can have instead of a cytosine (C) a thymine (T), like for the SNP located 14 thousands bases upstream of the lactase gene that enables the lactase persistence mentioned earlier[11]. A second group of small variants is made of insertions and deletions (called together *indels*): these are events in which it is present or missing a group of less than 50 nucleotides. The number of nucleotides is an arbitrary threshold used to better separate them from other kind of variations. Specific types of indels are the tandem repeats that, as the names suggests, are insertions or deletions of small repeated sequences of DNA. These repetitions usually are one after the other with no other sequence in between [12].

These groups of small variants, shown in figure**??** are the most described,

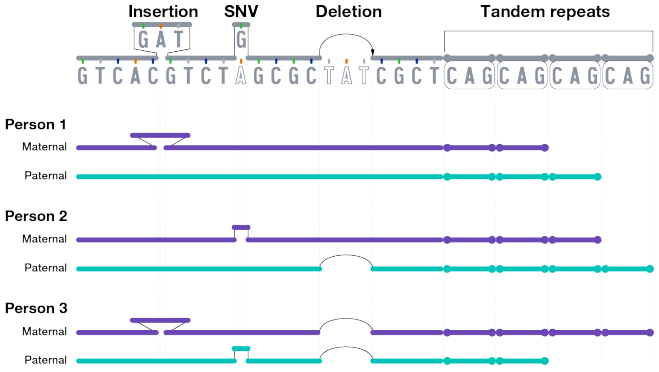


Figure 1.3: Graphic showing the types of small genomic variants [12]

studied and associated with diseases as they were the only one consistently detectable with NGS sequencing. For these reason, studies that tried to associate

genomic variation with diseases commonly used only these kind of variants.
The other kind of variations are the ones that stretch at least 50 nucleobases and that can reach the dimension of large chunks of the chromosomes: they are called structural variations (SVs). These can be indels or tandem repeats with the repeated section longer than 50 nucleotides, accounting for nearly half of all SVs, that take the name of Copy Number Variants (CNVs). Moreover, there are also inversions, in which a chunk of DNA is inverted compared to another and translocations in which pieces of two different chromosomes trade places [12].
Finally, it is important to remember that these kind of variations can be on just

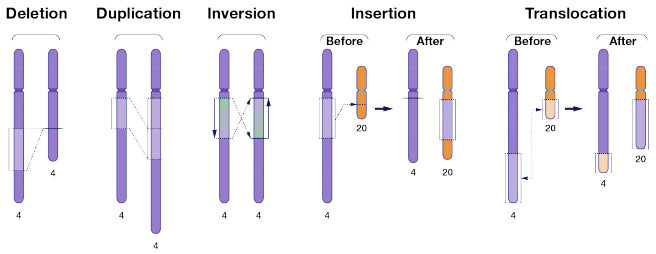


Figure 1.4: Graphic showing the types of large genomic variants [12]

one haplotype (copy of the chromosome) or on both: this distinguish between heterozygous and homozygous alleles.

1.3.2 The importance of studying genomic diversity in populations context

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 between different populations of the same species (inter-population): figure 1.5 shows the percentage of inter-individual variation for four close primates. Different species can differ in the amount of genetic variation that is s As discussed before, differences in DNA are given by having a different nucleotide at the same place (SNV), indels and large and complex variations, up to Megabases, 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 [10]. 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 1.6 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 functions.

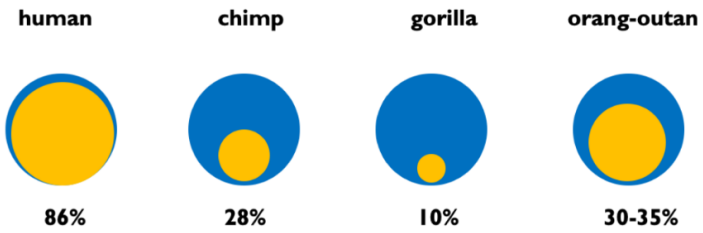


Figure 1.5: Share of inter-individual (yellow) and inter-population (blue) diversity for four different primates. While for humans the majority of the diversity is within populations, for other primates it is between populations. This shows how Humans are more mixed than other primates. Percentage shows the inter-individual variation share [10].

1.3.3 The premises for Pangenomics

There are a number of factors that must be taken into consideration to understand one one side the need for a new paradigm and on the other side the conditions that lead to its development. Here I will briefly expose some of them before diving into pangenomics approaches and methods.

1.3.3.1 A single linear genome for all analyses

Since the first complete genome sequences have been available in the late '90, all analysis based on sequencing data depended upon the use of a single linear reference genome, i.e. the best version of the genome available for any species. This reference sequence can or originate from the genome of a single organism or be a patch and consensus of multiple available genomes of the same species. It purpose is to use it infer information from newly, less refined, genomes that are being sequenced. 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 representation: this is valid for eukaryotes and even more for

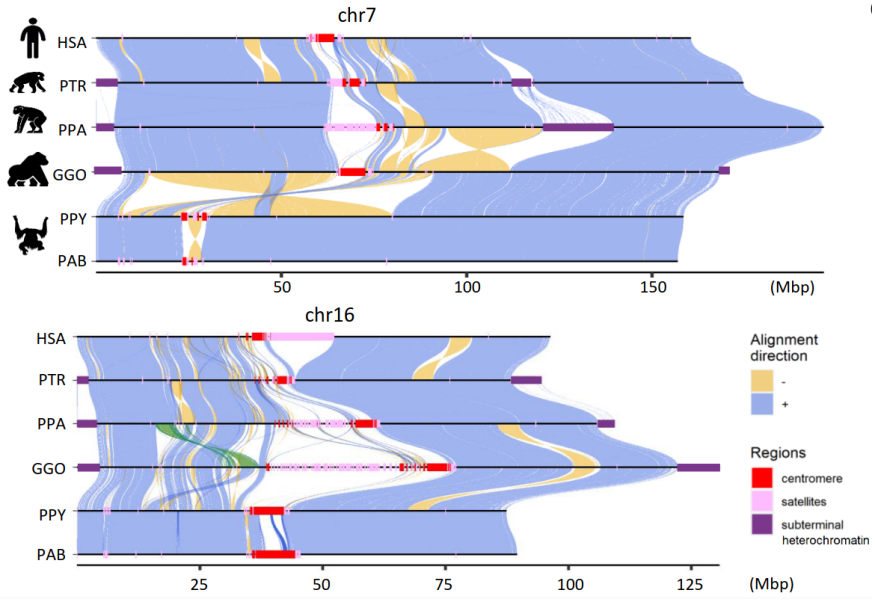


Figure 1.6: A comparative ape alignment of human (HSA) chromosomes 7 and 16 with chimpanzee (PTR), bonobo (PPA), gorilla (GGO),Bornean and Sumatran orangutans (PPY and PAB). The image on the top shows most of the chromosome 7 is conserved except for large inversions happening between the species. The image below shows complex inversions in chromosome 16. Image taken from 'Complete sequencing of ape genomes' [13].

bacteria, that tend to be very diverse even in the same strain. The goal would therefore be to find a representation that provides more genetic material of a single species by intelligently combines the information from genome of multiple organisms and their differences.

1.3.3.2 A quantity and quality revolution

In the last few years we are witnessing a new revolution in sequencing. As the price of sequencing is lowering more than 2x per year, from 1$/basepair to $10^{-}7$/basepair[14], new scientific discoveries and technological advances are leading to a remarkable increase of quality, in term of per-base error rate, and throughput of TGS. This means that in the next future we will dispose of a rich wealth of high quality sequencing information to produce hundreds or thousands of new first grade assemblies of large eukaryotic genomes.

For example, the history of complete human genome assemblies clearly exposes how much more high quality genomes it is now possible to generate. The Human Genome Project took 13 years to produce its result [15] and the absence of long reads with low error rate made it impossible to automatically resolve repetitive regions like telomeres and centromeres [16], producing a reference only

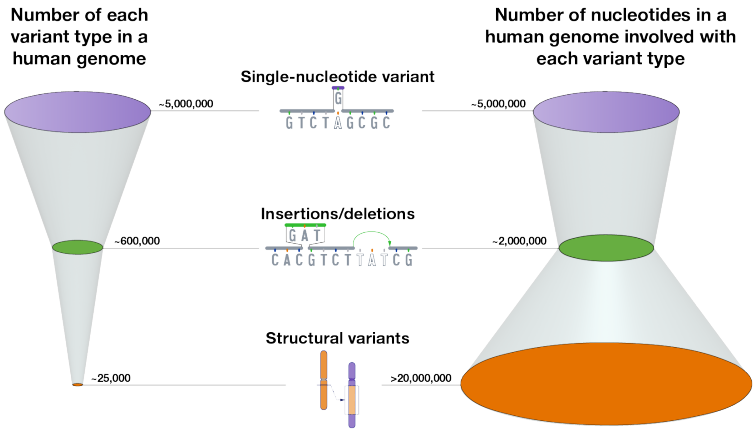


Figure 1.7: Spectrum of Human Genetic Variation. While SNPs are the most common variation event, their impact in the total amount of bases in a genome is 4 times smaller than the one of Structural Variations, that are 200 times less frequent. This shows the great need to consider SVs in genomic analysis and not to stop at the SNP/indel level.

92% complete [17]. This problem was only solved in 2022 with a new, reference genome that did not have any gaps or unresolved regions, from the telomere to the other telomere of each chromosome [17]. Now, many consortia are producing increasingly more genomes to a level comparable to the one produced in 2022. 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 of very high quality.

Finally, it is important to understand the quantity of biological information produced. As shown in table 1.3, the number of base pairs sequenced has more than doubled each year since 1995. As this is faster than the famous Moore's law on computing power, it is becoming evident that a new paradigm is needed to store and analyze such wealth of data. Public repositories, like Sequence Read Archive (SRA) and European Nucleotide Archive (ENA), are rapidly increasing the number of samples being sequenced and rendered publicly available to everyone, with tens of billions of millions of basepairs from genomic samples, as shown in figure 1.8. Other repositories of genomic data with associated medical metadata, like the UK biobank that comprises around 500 thousands individuals, are also emerging. These conditions are pushing the adoption of novel methods to process and analyze genomes.

year	genome(s)	base pairs
1995	Bacterium	$2 * 10^6$
2001	Mammal	$3 * 10^9$
2021	1M genomes	$3 * 10^{15}$

Table 1.3: Base pairs had a 10^9 increase in less than 30 years. As 10^9 2^{30} (from $\log_2(10^9) = 29.9$), the base pairs have more than doubled each year[14].

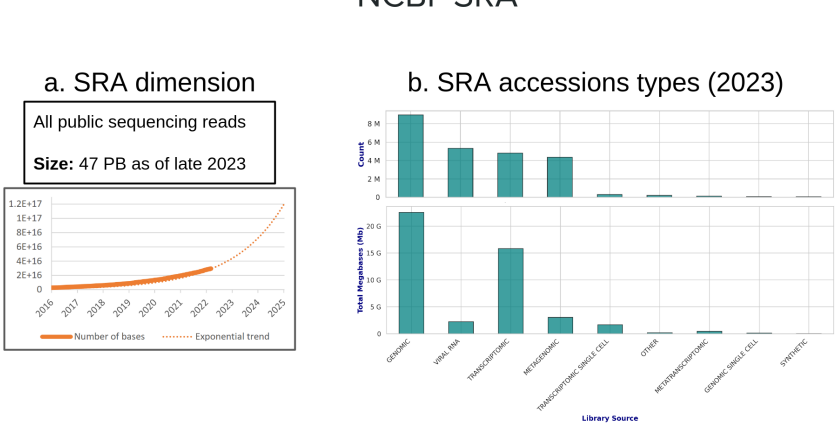


Figure 1.8: a) The size in PetaBytes (peta = 10^{15}); b) The type of data in the SRA database shows the vast amount of genomic data available. Image made from Rayan Chikhi's slides.

1.3.3.3 The need to better understand difference between genomes

The ability to produce such good data is the main enabler of increasing efforts from the scientific community to propose new methods to analyze genomes: not anymore by comparing new genomes against a single good reference sequence but by comparing it in a comprehensive representation of the species.
Moreover, as new high quality sequences and assembled genomes are available, complex and/or highly repetitive regions can be now represented also for new genomes therefore enabling comparison between the ones of different genomes. This is very important as up until these improvement in sequencing and assemblies arrived, analyses were mostly blind to large, complex and/or repetitive structural variations. As we now know that these are the ones responsible for most of the difference between human genomes, new proposed approaches should provide new and better tools to understand, represent and analyze such variations.

1.3.4 Pangenomics

Pangenomics is therefore a rapidly evolving field in genomics that aims to capture and analyze the full genetic diversity within a species or a group of closely related species. It does not rely on a single, linear reference genome, but on comparing any genome with a group of other similar ones, as it seeks to represent all genetic variations and structural differences across multiple organisms. It leverages the availability of large collections of high quality assemblies of many species to overcome the observational bias of using a single haplotype as reference for a whole population.
It was first conceptualized for bacterial genomes, and at gene level, without considering non coding regions. This was moslty due to the fact that bacterias share genes between each other, generating high diversity in the gene repertoir between organisms of the same species or strain. The first proposed pangenome model had a subdivision between a core genome, made by genes present in all individuals of a species, and a dispensable or accessory genome, with genes present in some, but not all, individuals.
This definition would then extend to a more general model that would consider variations at the nucleotide level to contain all variations in a set of genomes.

1.3.4.1 Pangenomes

A pangenome can be therefore be considered any collection of genomic sequences to be analyzed jointly or to be used as reference. This definition provides two important concepts for the rest of the studies provided in this manuscript:

Model the pangenome is not a well-defined structure or model but can be from a simple collection of sequences to complex data structures. This means that different approaches are developed and used depending on the application of interest;

scope a pangenome can be either used as:

- a new reference for a specific species to be used for analyses in a similar way as linear genome. This means that a large consortium would be producing a representation that is accepted as new standard. For the Human genome this is done by the HPRC consotrium as the T2T consortium produced the best-quality linear reference genome [17].
- a different model that can be used to study a set of genomes, without needing *a priori* to use a reference. This model can find applications in population variation studies.

On these premises, it is important to highlight that the high quality assemblies being produce by different consortia, that can be human or mammals or bacterial, constitute the most basic and usable pangenome. By having fully resolved the centromeres of 2 human genomes, with one of them being the reference produced by the T2T consortium, it is possible to detect small-scale and large-scale

centromere variations, something that was never possible before [18]. By having high quality assemblies of various apes, it is possible to reconstruct complex and large variations and rearrangements compared to the human genome that could not be detected before [19].

1.3.4.2 Pangenome Graphs

1.4 Graphs

1.4.1 De Bruijn Graphs

1.4.1.1 Colored and Compacted De Bruijn Graphs

1.4.2 Variation Graphs

1.5 Outline

Papers

Appendices