Application of genome assembly methods to human and non-human primate genomics

Lukas Kuderna

TESI DOCTORAL UPF / YEAR 2019

DIRECTOR DE LA TESI

Dr. Tomàs Marquès i Bonet,

DEPARTAMENT DE CIÈNCIES EXPERIMENTALS I DE LA VIDA

Universitat
Pompeu Fabra
Barcelona

ACKNOWLEDGEMENTS

Time flies.

I arrived in Barcelona for the second time over 6 years ago. I arrived with the best of intentions, the last of which was to start a PhD. Curiosity and the power of persuasion quickly made me change my mind. The last years have been filled with uncountable experiences of all kinds of colors, magnitudes and directions. These experiences have been created by and shared with numerous people around me, some of which have always been there, many of which have come and gone, some of which I'm lucky enough to still be able to share time with. I'm profoundly grateful to all of you. If you are reading this, there are probably plenty of reasons why your name deserves to be among these lines. Thank you!

Thank you, Tomas, for your continuous trust and the many opportunities you have provided over the last years.

Thank you, Florian, Christa, Max, and the rest of my family for your unconditional support over the last years. Thank you for only asking when I'll come back, but never questioning that I left.

Thank you, Silvia, for everything. Thank you for us, for your endless patience, your empathy, your love. Thank you for making me a better person.

Thanks to the people who have given me food for thought over the last years. But more importantly: Thank you to those who have - quite literally - nurtured me in this period and chosen to share their meal day after day.

ABSTRACT

Genomic analyses are at the center of contemporary biology. These studies heavily rely on reference genome assemblies, yet those are typically highly fragmented. Having accurate representations of complex genomes, or parts thereof, is crucial to study human and primate evolution and disease. Here, we develop and apply new sequencing strategies and technologies to improve reference assemblies. We first explore the combinatorial potential of different datasets to generate a highly improved reference for the chimpanzee, a crucial species for the study of human origins. We are able to close 77% of the over 159.000 remaining gaps in the previous iteration of this species' assembly and increase continuity by more than 750%. We then go on to develop a workflow to assemble the first human Y chromosome of African ancestry, using native flow-sorted chromosomes sequenced on a Nanopore device. We are able to assemble the Y chromosome to a reference grade quality and achieve unprecedented sequence resolution across structurally complex regions. These results open new avenues for comparative studies including the chimpanzee genome or human Y chromosomes.

RESUM

Els anàlisis genòmics són el centre de la biologia contemporània. Aquests estudis depenen molt de l'assemblatge de genomes de referència, tot i que aquets en general estan molt fragmentats. Tenir representacions precises de genomes complexos, o parts d'aquests, és crucial per estudiar les malalties i l'evolució en humans i primats. En els estudis següents, desenvolupem i apliquem noves estratègies i tecnologies de seqüenciació per millorar els assemblatges de referència. En primer lloc, explorem el potencial de combinar diferents conjunts de dades per generar una referència substancialment millorada per al ximpanzé, una espècie crucial per a l'estudi dels orígens humans. Som capaços de tancar el 77% dels més de 159,000 buits que hi havia a la iteració prèvia de l'assemblatge d'aquesta espècie, i augmentar la continuïtat en més del 750%. A continuació, desenvolupem un protocol per assemblar el primer cromosoma Y humà d'ascendència africana, utilitzant cromosomes nadius aïllats per citometria de flux i seqüenciats mitjançant un dispositiu Nanopore. D'aquesta manera, aconseguim assemblar el cromosoma Y a una qualitat de referència i una resolució de seqüències sense precedents en regions estructuralment complexes. Aquests resultats obren noves vies per a estudis comparatius que inclouen el genoma del ximpanzé o els cromosomes Y humans.

Preface

Recent advances in genome sequencing technologies and the ongoing reduction of their cost have taken genome assembly projects from the realm of large international consortia and given them into the hand of individual labs. The methods and strategies for these are being developed at an ever-increasing pace, greatly benefiting biological research. Here, I present results regarding the improvement of the chimpanzee genome assembly on the one hand, and new strategies to assemble human Y chromosomes on the other.

1 INTRODUCTION

Deoxyribonucleic acid (DNA) is the hereditary molecule that contains all necessary information to sustain and reproduce any extant life form known to date. It can be thought of as a string, or a *sequence*, composed of 4 letters called *nucleotides*: A, C, T and G. In a cell, these strings occur paired to another, complementary string in which A pairs with T and C pairs with G to form a so-called double helix (Watson and Crick 1953; Franklin and Gosling 1953; Wilkins, Stokes, and Wilson 1953). We have known these facts for almost 70 years now, yet it has not been until recently that we have developed an adequate methodology to read and analyze DNA at scale. This much-hailed advent of the genomics era has profoundly transformed the way biological questions may be asked and answered. Since its onset, biology has become unimaginable without the analysis of genomes (Goodwin, McPherson, and McCombie 2016).

Much of my work over the last years has focused on how to generate some of the resources upon which biology relies these days, that is: how to appropriately apply current methods, and develop new ones, to reconstruct genomes or its parts *as accurately as possible*. The emphasis here is important! Given sufficient funds, we now have access to a plethora of methods to read genomes, many of which produce results that are a highly accurate picture of reality. Nevertheless: for humans, and almost all other species, there is still not a single genomic representation that is free of some kind of error. It might come as a surprise to some, but it was not even until earlier this year (2019) that the first complete and uninterrupted sequence of a human chromosome has been announced, whose sequence was resolved from one tip to another (Miga et al. 2019). Although this one too is not devoid of errors. It is important to bear in mind that all assemblies, no matter how accurate, are always just a model of the underlying genome.

It was only during my Ph.D. that the generation of whole-genome assemblies from non-model organisms by individual labs has become routine. Since then, the pace at which they were created seemed to steadily increase, and oftentimes the technologies that were chosen at the conception of a project seemed almost dated upon publication of the results. In the following, I will try to provide a summary of some of these methods and their applications.

1.1 A brief overview of genome assembly the human genome

A canonical diploid human genome in a somatic cell is comprised of 46 chromosomes. There are 22 pairs of autosomes as well as two sex chromosomes, which are typically distributed as two X chromosomes in females and one X and one Y chromosome in males. The approximate size of a haploid human genome – that is, taking only into consideration non-redundant chromosomes – is around 3.2 Gigabases (Gb) with individual chromosomal size ranging from ~249 Megabases (Mb) in chromosome 1 to ~57 Mb in chromosome Y (International Human Genome Sequencing Consortium 2001). These chromosomes contain all information necessary to maintain and propagate life. A large part of this information is stored in genes, the ill-defined subunits of DNA that code for a molecule that exhibits some kind of function, for example, a protein. These protein-coding genes make up ~2% of the human genome and most recent estimates place the total number of coding genes in humans at around 20,000 (International Human Genome Sequencing Consortium 2004). Despite not coding for proteins, the remaining 98% of non-coding DNA is not devoid of function. This includes regions that are transcribed into functional RNA such as transfer RNA or ribosomal RNA. There are furthermore several functional DNA elements that are not transcribed, such as regulatory elements, including promotors and enhancers.

Approximately 50% of the human genome is made up of repetitive elements (Treangen and Salzberg 2012). These elements come in a variety of shapes, sizes, and origins, and can be interspersed in the genome, arranged in tandem next to one another, or even be nested within other elements. They are fundamental to understand the makeup of a genome and the challenges in analyzing it. Repeat subunits can be as small as a single base pair in simple repeat sequences and range up to 1 Mb in large segmental duplications, which are recent duplications in the genome that exhibit >90% identity to one another over at least 1 Kb of sequence (Eichler 2001). The most prominent elements in terms of the number of repeat units and proportion of the genome are long interspersed nuclear elements (LINES) and short interspersed nuclear elements (SINES) (Treangen and Salzberg 2012). These two classes alone make up ~36% of the genome and contain the two most numerous repeats in the human genome: Alu and L1 elements. A human genome has well over 1 million copies of Alu elements making it the single most abundant one. Many of them share high identity with each other. They are primate-specific mobile elements of around 300 bp in size whose activity peaked in early primate evolution, but some are still active in the human lineage (Deininger 2011).

L1 elements are another type of repeat that is still active in the human lineage. They occupy around 17% of the human genome, which is the largest proportion of bases of any single repeat element. A single L1 element is around 6000 bp in size (Treangen and Salzberg 2012).

Repeats such as L1 and Alu by themselves do little more than create and insert copies of themselves. However, the insertions and their locations can have broad evolutionary and functional implications by affecting gene structures, disrupting coding sequences or altering expression patterns. Ultimately, these insertions can be responsible for a series of diseases, such as muscular dystrophy, thalassemia, macular degeneration, hemophilia, and several cancers, among many others (Deininger 2011; Hancks and Kazazian 2016; Kazazian and Moran 2017). Having accurate reconstructions of genomes, including proper repeat resolution, is therefore key to understand many biological processes and their impact.

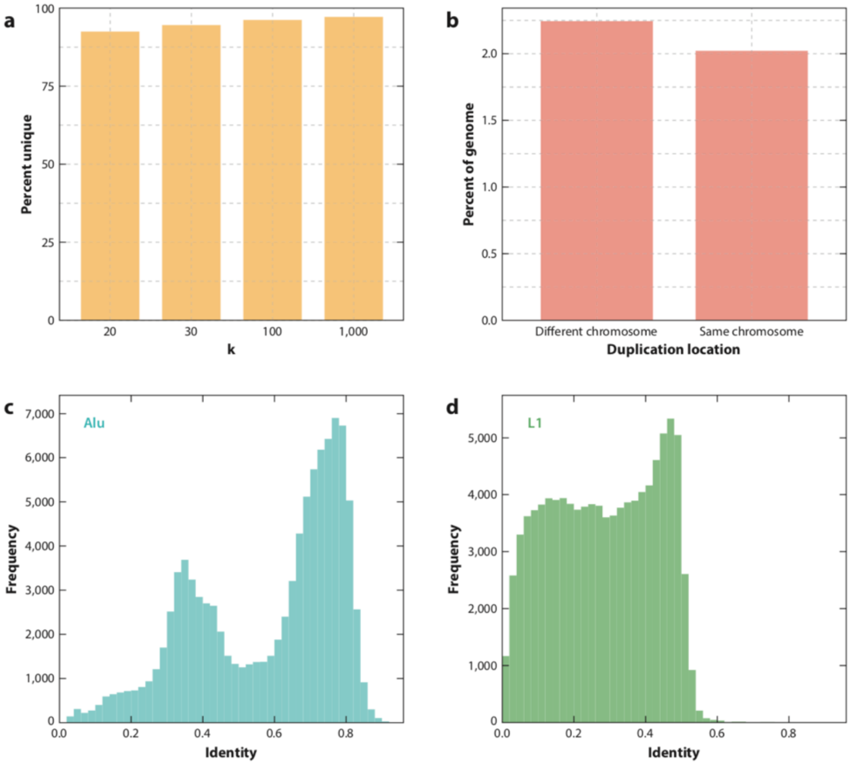


Figure 1: Taken from Rice and Green, 2019. Repeats in the genome confound assembly as exemplified by the human genome (*a*) Percentage of unique k-mers in the human genome. Even at k = 1,000, some k-mers appear multiple times in the genome. (*b*) Percentage of the human genome in segmental duplications (>95% identity, >5 Kb length(*c–d*) Distribution of alignment identities for 100,000 randomly sampled pairs of (*c*) Alu and (*d*) L1 repetitive elements.

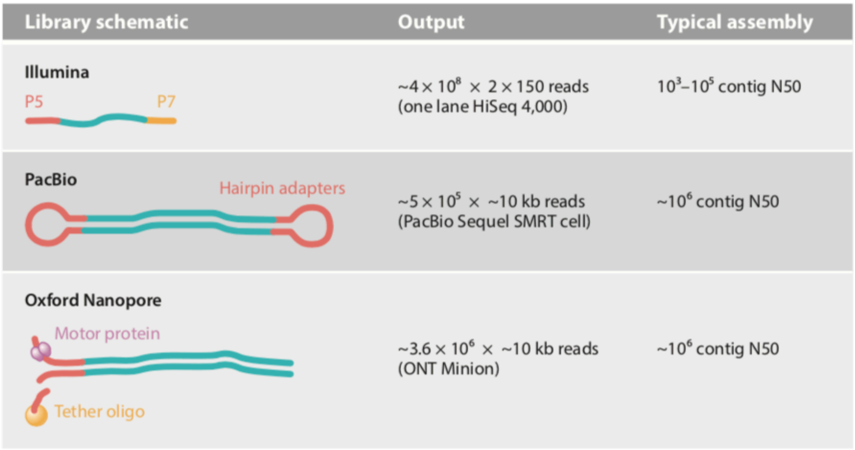
Repetitive elements are at the center of the computational and analytical challenges of genome sequence analysis, as contemporary biological studies fundamental require the availability of accurate genomic sequences. There is currently not a single platform that is capable of reading the sequences of full nuclear chromosomes on a practical level. As we shall see different platforms impose different size limits on the DNA fragments they can detect – the so-called *reads.* Therefore, to analyze the sequences these reads are derived from, they need to be *assembled* to a *reference* genome. Consequently, the central problem genome assembly deals with is the reconstruction of genomic sequences from fragments – reads – which are several orders of magnitude shorter. We shall explore some of the issues associated with this central problem, and how genomics has provided computational and molecular solutions to them.

Assembling a reference genome is usually at the beginning of genomic analyses. If such a reference assembly already exists, a study might opt to *resequence* more individuals to study the differences between them or their *variation*. In this case, reads may be mapped to determine their position relative to the reference and detect the variants. The majority of genomics studies does precisely this, as it is much cheaper to detect variants in a resequencing experiment relative to a reference than to construct a de-novo reference itself. The former usually requires a simple upfront preparation of the DNA and can be sequenced on cost-effective machines, while the latter usually greatly benefits from more complex preparations of the DNA before ideally reading it on less cost-effective machines. The initial assembly of a reference human genome sequence that was released in 2001 took 10 years to create and had a price tag of ~2.7 billion USD, which adjusted for inflation is almost 4 billion USD in 2019. Generating the equivalent amount of data today takes a couple of hours and usually costs less than 1000 USD. Of course, the resulting data of these two numbers are not comparable as we have just established. But even today generating an assembly of comparable quality to the human reference genome with state-of-the-art methods will usually cost ~100 times more than just resequencing (Rice and Green 2019).

As mentioned above, a haploid human genome is approximately 3.2 Gb (that is 3,200,000,000 bases) long. I believe it is important not to lose sight of the scale of this number. We might know certain reads are this long, and a given chromosome has roughly that size, but it is hard to get an intuition for the absolute and relative scale of genomic distances. Imagine the following: If we lay out all chromosomes of a haploid human genome tip to tip, it would span around 2 meters at an average nucleotide length of ~$6 \cdot 10^{-6}$ meters. Now, let us scale that length to the distance between Barcelona and Vienna (or Barcelona to London, if you are more familiar with that route - the results are roughly the same) (Peona, Weissensteiner, and Suh 2018). The two cities are 1350 km apart which equates to around a two-hour flight. At that distance, a single nucleotide would weigh in at around 4 mm in length. The average Illumina read (100 b) would measure around two 2-euro coins back to back. Picture that for a second! If a genome measures the distance between Barcelona and Vienna, we are predominantly reading it with units the size of just two 2-euro coins! A typical PacBio read (10 Kb) would measure the size of a car, and a typical Nanopore read would be around the size of a Bus. Optical- and Hi-C maps would span half the length of a Barcelona metro train. All these objects are of course almost negligibly small when regarded in the context of the scale we are looking at, yet it is the jump in scale from two 2-euro coins to a bus that made the recent dramatic improvement in genome assembly quality possible. It is important not to forget that the scale of units with which we analyze and reconstruct genomes is 4-7 *orders of magnitude* – that is between 10,000 to 10,000,000 times smaller than the actual size of the genome itself.

1.2 Sequencing methods

The dramatic developments in sequence assembly over the last decade are intimately tied to technological advances in genome sequencing, and it is not possible to understand the first without talking about the latter. I will discuss some key platforms that have propelled this progress, each with their own set of advantages and disadvantages, be they technological or economical. The possibilities and limitations of each platform directly translate to the resulting assemblies' quality in most cases. This is either because of direct effects – i.e. the raw data quality and read-length that is produced by the machine, or indirect effects such as necessary or possible types of library preparation associated with the platform (see Figure 2).



Library schematic	Output	Typical assembly
Illumina P5 P7	~4 × 10^8 × 2 × 150 reads (one lane HiSeq 4,000)	10^3–10^5 contig N50
PacBio Hairpin adapters	~5 × 10^5 × ~10 kb reads (PacBio Sequel SMRT cell)	~10^6 contig N50
Oxford Nanopore Motor protein Tether oligo	~3.6 × 10^6 × ~10 kb reads (ONT Minion)	~10^6 contig N50

Figure 2: From Rice and Green, 2019. Overview of libraries, typical output, and typical assembly continuity for Illumina, PacBio and Nanopore.

1.2.1 The first generation: The Sanger era

The early days of genome sequencing were dominated by di-deoxy chain termination sequencing, better known as Sanger sequencing. Originally developed in the 1970ies, it was the most widespread sequencing platform and used almost exclusively for around 40 years (Sanger, Nicklen, and Coulson 1977). Briefly, the Sanger method works by elongating a template strand of DNA with a mixture of ordinary deoxynucleotides (dNTPs) and labeled dideoxynucleotides (ddNTPs) into the elongating DNA chain. The ddNTPSs are readily incorporated into the novel strand by a polymerase but terminate the reaction thereafter as they block any further nucleotides from participating in the reaction. The different elongated DNA strands are then separated by size. As the initial concentrations were chosen in such a way that each position of the template DNA will terminate the reaction, the detection of labeled ddNTPs is then used to reconstruct the original sequence. The Sanger method was originally developed with radioactive ddNTPs to be run on gels where the sequences would correspond to bands after electrophoresis. The key to its success, however, was the possibility to automate the process through the development of Sanger capillary sequencing machines and fluorescently labeled ddNTPs. It was on those machines that the first genomes of multicellular eukaryotes were sequenced, starting with *C. elegans* and the fruit fly as test cases, and leading to the emblematic first mammalian assemblies of human and mouse (The C. elegans Sequencing Consortium 1998; Adams et al. 2000; Mouse Genome

Sequencing Consortium et al. 2002; International Human Genome Sequencing Consortium 2001).

From today's perspective, these early genomic representations were special in that they used comparatively short templates of around 200 Kb that were cloned and propagated in bacteria (bacterial artificial chromosomes, or BACs) in a process called "BAC hierarchical shotgun". To this end, BAC libraries that cover the whole genome were generated, and individual BACs that cover the specific region of interest were identified, pulled out, sequenced and assembled. Due to the high repeat content of mammalian genomes, the assembly of individual BACs is much easier than trying to reconstruct the whole genome at once, as regions that might be repetitive in the context of the whole genome are more likely to be unique in the context of the BAC. The BAC hierarchical shotgun or clone-by-clone approach is laborious and therefore expensive as such. At the time of the initial sequencing of the human genome, automated and parallelized Sanger capillary sequencing of a megabase of DNA took over a day and had a cost estimate of ~1500 USD, which was responsible for the extremely high price tag of the project. Nevertheless, it yielded an assembly that is of stellar quality even by today's standards.

The decision to apply a clone-by-clone strategy was debated at the time of sequencing the human genome with a whole-genome shotgun (WGS) strategy as the alternative (Weber and Myers 1997; P. Green 1997). This approach seeks to sequence and assemble the whole genome at once without prior subdivision. While abundant repetitive elements will lead to a much more fragmented assembly, this strategy was shown to, at least in theory, produce a workable genome assembly. Crucially, it implies a massive cost cut compared to the clone-by-clone approach, which was the main argument of its proponents. Several further mammalian genome assemblies based mainly on Sanger whole genome shotgun data followed over the next years, among them horse, dog, rat, orang-utan, rhesus macaque and gibbon (Rat Genome Sequencing Project Consortium 2004; Lindblad-Toh et al. 2005; Gibbs et al. 2007; Wade et al. 2009; Locke et al. 2011; Carbone et al. 2014). Crucially, among the first Sanger whole-genome shotgun assemblies, and the first non-human primate to be sequenced, was the initial reference of the chimpanzee (The Chimpanzee Sequencing and Analysis Consortium 2005).

The Sanger sequencing machines produced reads between 400-800 bp long, and their assemblies are generally characterized by high base-level accuracy. Given the increased cost of sequencing, the coverage for typical mammalian genomes was around 5-10-fold, which usually resulted in fragmented assemblies. Until today, human and mouse are the only mammals with assemblies based on a BAC hierarchical shotgun.

1.2.2 The second generation: Massively parallel sequencing

In 2005 the technological basis for the first commercially available and viable alternative to Sanger sequencing was introduced: 454 pyrosequencing (Margulies et al. 2005). This platform was subsequently used to sequence a personal genome in 2 months for around 1/100th of the cost of Sanger sequencing (Wheeler et al. 2008). Shortly thereafter, Solexa - the company that developed the technology used today in Illumina sequencers - published the sequencing of an individual human genome to comparatively high coverage (Bentley et al. 2008). Neither of the previous two examples were proper genome assemblies but instead relied on the already produced reference genome to map the reads and detect variants in a process termed resequencing. Nevertheless, they prove the technical feasibility and technological advantage of massively parallel sequencing (MPS) for genome analysis. Illumina reads were initially an order of magnitude shorter than those produced by Sanger platforms - 36 bp compared to almost 1 Kb. The initial difference in size led several people to question the computational feasibility of genome assembly from very short reads. The genome assembly of the giant panda was the first to dismiss these critics and showed it was possible to do so (R. Li, Fan, et al. 2010). This assembly was produced based purely on Illumina data and reached similar continuity to other Sanger WGS assemblies of mammalian genomes, albeit at much deeper coverage. It was the first step towards the democratization of genome assembly over the coming years, as plummeting sequencing costs suddenly brought non-model organism assemblies into the reach of "ordinary" research groups. This resulted in a considerable increase in available reference sequences. The read length has increased since then, currently producing between 150-300 bp reads. Nevertheless, despite the massively increased throughput, short read sequencing aggravated the main issue with genome assemblies: trying to accurately reconstruct repetitive genomic sequences from fragments several orders of magnitude shorter than the sequence of interest (Alkan, Sajjadian, and Eichler 2011). As a consequence, despite the increase in available assemblies, their quality drastically dropped as they often consisted of tens to hundreds of thousands of individual fragments, whose relative order and orientation remained unknown. While even fragmented assemblies can yield valuable insights, they are often limiting factors for certain types of analyses, such as long-range cis-regulation, runs of homozygosity, the study of recombination, genetic association studies or chromosomal evolution (Rice and Green 2019). Fragmented gene models might also conflate analysis regarding, for example, gene number or ortholog detection (Denton et al. 2014)

Despite producing fragmented assemblies, the per-base quality of Illumina is comparatively high, although the platform exhibits a GC-dependent coverage bias (Minoche, Dohm, and Himmelbauer 2011). Single base substitutions are the main error type. To this day, it is far and beyond the most widely used platform.

1.2.3 The third generation: Long-read sequencing

In 2009 Pacific Biosciences (PacBio) introduced the technological foundation of a new platform that allowed the sequencing of over a thousand continuous bases at a time (Eid et al. 2009). This method works by filming the released fluorophores of labeled nucleotides that are synthesized from a template strand by an immobilized polymerase in a nanochannel called a *zero-mode waveguide*. While the read lengths were initially restricted to a couple of thousand base pairs, further development of the platform resulted in an increase currently yielding average read lengths of around 30 Kb. In 2015, the first human genome was sequenced with this platform in a resequencing study, showing large scale structural variation that has been previously inaccessible (Chaisson et al. 2015). Despite producing long reads, this platform showed error rates of up to 17% that consist mainly of small insertions and deletion (Koren and Phillippy 2015). Nevertheless, a first workable assembly of the same data was produced the same year, which showed not only very high continuity but also high accuracy (Berlin et al. 2015). This was achieved by taking advantage of the random distribution of those errors, meaning that sufficiently high coverage can overcome them. Several high-quality assemblies followed in the last years, including all the great apes (Gordon et al. 2016; Kronenberg et al. 2018). The long reads make it possible to span most common repeat elements in vertebrate genomes thus rendering many sites that lead to breaks in assemblies with Illumina trivial to reconstruct. PacBio has since gone on to become the “industry standard” for high-quality assemblies.

More recently, a commercial implementation for nanopore sequencing by Oxford Nanopore Technologies (ONT) has become available and is rapidly evolving. In contrast to most other platforms, these machines do not rely on a polymerase for sequencing. Additionally, they do not require an optical detection system, which usually is the most expensive aspect of a sequencer (Niedringhaus et al. 2011). ONT reads sequences by pulling single-stranded DNA (ssDNA) through an engineered protein nanopore that is embedded in a lipid bilayer and measuring voltage changes across the membrane. To this end, the double-stranded DNA (dsDNA) is placed on one side of the membrane and unwound to form ssDNA. The diameter of the pore only allows for ssDNA to pass through. Once the ssDNA is within the pore, it blocks the ion flow creating characteristic voltage patterns that can be interpreted into nucleotide sequences. Since its introduction nanopore has gone on to produce the longest read lengths of any platform (Payne et al. 2019). There does not seem to be in inherent upper read length limit, as with all other platforms, but rather that the limitation is imposed by the integrity of the loaded DNA. Current average read accuracies are still below PacBio, but more recent estimates go as low as 12% (Rice and Green 2019). These errors are not random but exhibit a certain bias dependent on the GC content of the sequence in question. Nevertheless, the platform is rapidly gaining popularity and the technological and algorithmic ecosystem surrounding it is under intensive and constant development. Recently, the most continuous whole-genome shotgun assembly for human was produced on it with the aid of so-called ultra-long reads, which regularly exceed 100 Kb (Jain, Koren, et al. 2018). Currently, the longest recorded read is > 2 Mb long (Payne et al. 2019). ONT makes certain repeat structures resolvable that were inaccessible with other platforms, such as the first fully resolved human centromere (Jain, Olsen, et al. 2018). Lastly, a combination of different long-read platforms, including both PacBio and Nanopore, is currently being used to produce complete, tip-to-tip assemblies of all human chromosomes (Miga et al. 2019).

1.2.4 Transitional hybrid assemblies

Although we have treated the different sequencing platforms as discrete assembly entities, several transitional hybrid assemblies between the first and second, and the second and third generation have been reported. One example is the combination of Illumina data to assemble all parts of the genome that are accessible to this platform, and then use low coverage long read data to address the remaining bits. This approach is a tradeoff between the price point a pure long read assembly would provide, and the hindered repeat resolution of a pure short read assembly, and has been applied to several species such as the Mouse Lemur or the Rhesus Macaque (Larsen et al. 2017). Another example is the Olive Baboon for which a previous iteration generated both Sanger and Roche-454 data. Both datasets were recycled for the latest iteration, which additionally included Illumina and PacBio data (Rogers et al. 2019). However, not all hybrid assemblies seek the balance between second and third-generation sequencing. For example, the Tarsier and the Bottlenose Dolphin were assembled with a combination of both Sanger and Illumina data, likely as the project was underway when the latter data type became available (Schmitz et al. 2016).

We will see later on that there are several different steps to an assembly, not all of which deal with the reconstruction of the sequences, but also with their relative order and orientation. To this end, different library types may be created, that can be sequenced on different platforms. Furthermore, current long read assemblies often rely on orthogonal short read data to "polish" remaining systematic errors that hard to overcome. In this sense, there are currently very few assemblies that rely solely on a single platform from start to finish, meaning that most contemporary assemblies are hybrids of different datatypes, from either different sequencing platforms or different library generation types (Chaisson, Wilson, and Eichler 2015).

1.3 The structure of a genome assembly

We have previously established that the fundamental problem in genome assembly is the reconstruction of contiguous sequences from fragments that are orders of magnitude shorter. These contiguous sequences called *contigs* are at the center of a genome assembly. In an ideal world, a genome assembly would produce one contig per chromosome. That, however, is not possible for most organisms as the contig assembly poses a veritable computational challenge in itself. Additionally, the problem is further aggravated by technical limitations and biology: Reads can be error-prone, genomes can be highly repetitive and may have additional allelic diversity of several types, such as single nucleotide variants, short tandem repeats or structural variants that further obscure the reconstruction. We will briefly ignore errors and allelic diversity and focus purely on the fundamental problem of assembly algorithms: how to deal with repetitive regions in the genome.

1.3.1 Contig assembly

Imagine that we would like to read Charles Dickens's “A Tale of Two Cities” [1]. This book starts with the sentence: “It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness […]”. For whatever reason it has been shredded line by line into small pieces of 5 words each:

It was the best of
times, it was the worst
of times, it was the
age of wisdom, it was
the age of foolishness […]

To be able to read the book, we need to find a way to reconstruct the original order of each 5-word fragment or reads. Having only one shredded copy of the book reveals the first problem: Without any overlaps between the fragments (ignoring language semantics for a second) there is no way for us to know which of them originally occurred adjacent to each other. Luckily, we don’t have 1 but 5 shredded copies with 5-word fragments starting at random positions, so we’ll be able to observe each individual word 5 times, as it will be *covered* that many times. The complete 5-word fragments we’ll observe are:

[1] This is a well know teaching example in the context of genome assembly, that has been developed by Michael Schatz at Cold Spring Harbor.

was the age of wisdom,
the age of wisdom, it
times, it was the age
of times, it was the
it was the worst of
was the age of foolishness
It was the best of
the best of times, it
of wisdom, it was the
of times, it was the

times, it was the worst
was the best of times,
was the worst of times,
best of times, it was
the worst of times, it
it was the age of
age of wisdom, it was
worst of times, it was
it was the age of
wisdom, it was the age

Now that we have bridging information between the different fragments and can look for overlaps between them. The simples, most naïve approach is to 'greedily' compare the first fragment with all other ones, retain the longest possible overlap to merge the two fragments. This can be done iteratively for all fragments until all possible overlaps are "used up":

It was the best of
was the best of times,
the best of times, it
best of times, it was
of times, it was the
of times, it was the
times, it was the worst
times, it was the age

The above example quickly demonstrates some limitations of this approach: Ambiguous overlaps due to repeats make it impossible to find the true original sequence and break the next overlap generation. In the above case, the 5-word read “of times, it was the” constitutes a repeat that is not unambiguously resolvable. Furthermore, with this algorithm, it is not possible to know the actual number of repeats, as we don’t know if we are observing a given 5-word read because it occurred more than once in the underlying sequences, or because the same region has been sampled more than once by chance. However, the greedy genome assembly algorithm was widely used in the production of the initial human genome (International Human Genome Sequencing Consortium 2001). This was possible because of the clone-by-clone approach the project deployed. Imagine that instead of needing to reconstruct the whole example above, you can first extract sub-sentences of 12 words that you know belong together, e.g.: “It was the best of times, it was the worst of times”. If you sample 5-word reads from this sub-sentence there is no single repeated one that would make a greedy reconstruction ambiguous. A read that was repetitive in the context of the whole sentence is now unique in the context of the sub-sentence:

It was the best of
was the best of times,
the best of times, it
best of times, it was
of times, it was the
times, it was the worst
it was the worst of
was the worst of times

Similarly, sequences that are repetitive in the context of the whole genome might become unique in the context of a BAC.

The above example is intuitive, but simplistic and ignores much of the technical and biological complexities that assemblers need to take into consideration. We saw that in order to reconstruct the sequences, we need to sample a certain position more than once. The amount of times a given position is sampled is called *coverage* and is a key factor for assembly, as it will determine both the assembly quality, and the cost of data generation. If we assume no biases, the probability of sampling any given position of the genome is quite low and given reads that are far smaller than the genome a large number of them is needed to cover the whole genome. Disregarding biases, this sampling roughly follows a Poisson distribution. If we would like to sequence a genome, we can exploit this fact to calculate the probability of covering a certain proportion of the genome to make sure sufficient data is generated to assemble or analyze it. For a coverage c, the probability of any given base not being covered is $P(X=0)=e^{-c}$. Conversely, the probability of a given base being covered is $P(X>0)=1-e^{-c}$. So, if we would like to cover at least ~99% of our genome of interest, we would need to sequence it to an average of 4.6-fold coverage with ideal data. In practice, the number should be slightly higher to account for sequencing biases, but it is roughly what most Sanger WGS assemblies choose (Z. Li et al. 2012). This calculation also disregards the required overlaps we might need for assembly and assumes perfect data generated from a repeat free genome. To achieve the ~5 X we require to cover ~99% of the genome at least 1 time, we need to produce around 20,000,000 reads of 800 bp. Lander and Waterman provided a mathematical framework which allowed a back of the envelope calculation to estimate the approximate number of fragments a contig assembly would yield (Lander and Waterman 1988). Assuming a coverage $C=N*L/G$, with N representing the number of reads, L the read length and G the genome size, the estimated number of contigs can be calculated as $E_{contigs}=N*e^{-C}$. For the 20,000,000 reads mentioned above needed for approximately 5X coverage at 800 bp reads, this results in an estimated ~135,000 contigs. This number is in good agreement with the empirically observed ones from Sanger WGS assemblies, many of which chose similar parameters.

Greedy reconstructions quickly reach their limitations in the context of whole-genome shotgun assemblies, and different approaches were needed. Two main algorithmic approaches were developed to this end: The Overlap-Layout-Consensus (OLC), and de Bruijn graphs (DBG) (Nagarajan and Pop 2013). Both approaches deal with the issues surrounding repeats encountered above by essentially leaving them out of the resulting assemblies by breaking them into individual contigs if they are not resolvable. OLC and DBG have a common underlying mathematical framework: they model the assembly problem as a graph. This framework deals with nodes, or vertices, that are related by edges, which can be thought of as the reads and their overlaps in this context.

The OLC approach starts by computing overlaps between reads by comparing all against all reads to generate the overlap graph mentioned above, in which a node is generated for each input read and an edge between two nodes corresponds to an overlap. The overlaps can be computed in a number of different ways, and this step is generally very computationally intensive: imagine having to compare all >20,000,000 reads (an absolute bare minimum) mentioned above to each other. The need to compare all-versus-all also means that the burden increases quadratically with the number of input reads. Assemblers usually apply some clever approaches to ease this burden, but it remains challenging regardless. The following step, the graph layout, simplifies this graph by removing all transitively inferable edges, therefore generating what will be the final contigs of the assembly. This means that the graph is "thinned out" and its traversal to emit the contigs is computationally easier. Lastly, a consensus step chooses the most likely base to wash out the remaining errors in the assembly.

The OLC paradigm was developed in the face of relatively low coverage datasets, such as early Sanger assemblies. With the introduction of MPS, these algorithms faced a new obstacle: coverage was not a limiting factor anymore as sequencing became much cheaper. But as read lengths decreased and coverage increased, the absolute number of reads exploded. To assemble a 60X human genome with 36 bp reads, > 5,000,000,000 reads where needed, and constructing an overlap graph for each pair of reads became computationally impractical. To mitigate this issue, the de-Bruijn graph was developed. This assembly paradigm starts counterintuitively by deconstructing the reads into pieces of length k, called *k-mers*. These k-mers are used to populate a graph in which each node is occupied by a k-mer and an edge indicates that two adjacent nodes differ by exactly k-1 letters. Taking the example from above: The read "It_was_the_best_of" can be deconstructed into its constituting 3-mers "It_was_the", "was_the_best", and "the_best_of". The expensive overlap computation can then be replaced by exact matchings of adjacent k-mers in the graph, meaning that reads are not compared all to one another, and the computational burden scales linearly with the amount of input data, as opposed to quadratically. This also implies that the input data for DBG assemblers must have a very low error rate, as the matchings would otherwise fail. Decomposing the input reads into k-mer also implies that the ability to resolve repeats longer than k is initially lost (see figure 3). Furthermore, heuristics to deal with allelic diversity must be implemented as well. It is important to note that the vast majority of assemblers currently produce a haploid representation of the genome, which often chooses one allele at random. This implies that the haplotypes represented within assemblies are not necessarily biologically meaningful. Finally, to emit contigs from the graph, it is traversed trying to visit each node once and breaking the contigs at bifurcations in the graph.

The above descriptions are highly simplified. Many sophisticated algorithmic details are implemented in real life assembler to deal with repeats, error correction or data storage. Crucially though, most assemblers are built around certain data types to be able to deal with the specificities of the platforms. Generally speaking, greedy assemblers were predominantly used for the assembly of BACs. They are not widely used anymore due to their inherently local results. Prominent examples include PHRAP and Tigr (Bastide and McCombie 2007; Sutton et al. 1995). DBG assemblers are designed for short reads with high accuracy, i.e. MPS data such as Illumina or Roche-454. DBG assemblers have completely dominated the analysis of this data type, and thus genome assembly in general in the last years, until the rise of low accuracy long-read platforms. The first assembler to introduce DBG was Euler (Pevzner and Tang 2001). Several others followed, and notable examples include Velvet, SOAPdenovo, Allpaths and Discovar (Zerbino and Birney 2008; R. Li, Zhu, et al. 2010; Butler et al. 2008; Weisenfeld et al. 2014). Lastly, OLC assemblers were initially developed for WGS assembly of low-coverage, high identity Sanger data. The most prominent OLC assembler is the Celera-assembler, originally developed to assemble the WGS human data from the private consortium (Myers et al. 2000; Venter et al. 2001). The same pipeline has been adapted to deal with data from Roche-454 and PacBio and has recently been rewritten and further developed into Canu, a general-purpose long-read assembler (Miller et al. 2008; Berlin et al. 2015; Koren et al. 2017). OLC approaches are at the core of modern long-read assemblers, as DBG cannot deal with the noise inherent to 3rd generation data types. These different approaches mainly differ by their overlapping algorithms that deal with error-prone long reads, and the way they deal with repeats. Further recent examples include Falcon or Flye (Chin et al. 2013; Kolmogorov et al. 2019).

There is no "one size fits all" optimal solution for contig assembly. The assemblers are usually data specific and their choice depends on what genome to assemble with what data. Ultimately, depending on the data type and assembler used a contig assembly usually produces between several 100 to 100,000 fragments. Their relative order and orientation must be determined with additional data.

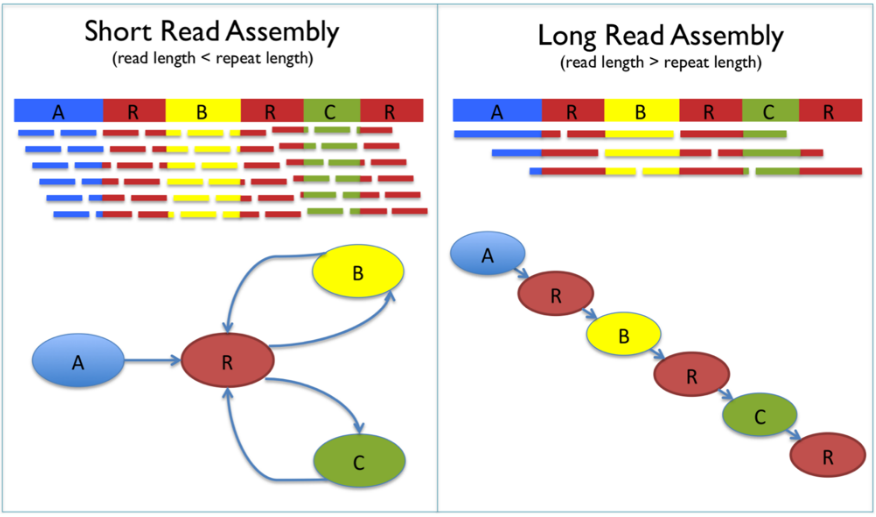


Figure 3: From Lee et al, 2014. Comparison of resulting assembly graphs for toy genome consisting of 3 unique regions (A,B,C) and 3 copies of the repeat R. If the genome is sequenced with reads shorter than the repeat unit (left) the resulting assembly graph is branching and can therefore not be unambiguously resolved. The resulting assembly will be fragmented. If the read length is longer than the repeat unit the resulting graph is linear and a single contig can be emitted.

1.3.2 Scaffold assembly

Scaffolds are sets of ordered an oriented contigs with gaps of approximate size between them. These gaps are usually represented by a series of "N" in an assembly, the IUPAC code for an unknown nucleotide. As the contig assembly produces fragments whose lengths are usually substantially smaller than the chromosomes, some kind of long-range information is needed to stich them together. Traditionally, approaches to creating genetic maps were used such as linkage mapping or radiation hybrid (RH) mapping. Both methods take advantage of the fact that markers that are physically close on a chromosome will co-segregate more often than distant markers. The co-segregation was determined either as the result of recombination for linkage mappings or chromosomal breaks due to radioactivity for RH mapping. The resulting contigs could then be linked to chromosomes via Fluorescent In-Situ Hybridization (FISH). This method anchors fluorescently labeled probes to chromosomes to be detected under a microscope (Rice and Green 2019).

One of the most widely used methods to link contigs together was the possibility to sequence both ends of a fragment of DNA, a process called paired-end sequencing (Kelley et al. 1999). This method was exploited early on, as Sanger sequencing traditionally required the clonal amplification of fragments to be sequenced. To this end, plasmid libraries with insert sizes of several Kb were generated, the extremities of which were sequenced as so-called mate-pairs. BAC-end sequencing was also common for species that had libraries available. These early assemblies greatly benefited from the additional data source that was available for a minor additional cost. Second-generation sequencing did not require these kinds of preparation for sequencing, so creating mate libraries constitutes an additional investment. Nevertheless, creating several different libraries with varying insert sizes was a common strategy to overcome ambiguities in the assembly due to repeats, and to create scaffolds. As an example, the popular DBG assembler Allpaths-LG was designed around a combination of specific libraries, or a "recipe" (Gnerre et al. 2011). This strategy has the benefit of being able to make implicit assumptions for assembly. Specifically, the program was designed around 4 different library types: A fragment library with a 180 bp insert at ~45X and a "short jump" library with an insert of 3 Kb at ~45X. Additionally, a "long jump" library with an insert of ~6 Kb and a "fosmid jump" library with an insert of ~40 Kb could be included at 5 & 1X, respectively. This interplay between experimental and algorithmic design has arguably been very successful (Earl et al. 2011; Salzberg et al. 2012).

One scaffolding strategy that has become increasingly popular is based on proximity ligation and chromosomal conformation capture with a method called Hi-C (Lieberman-Aiden et al. 2009). DNA in a cell needs to be folded to be stored, and this method captures the spatial interactions of this storage process. It takes advantage of the fact that regions that are close to each other in the DNA sequence tend to physically interact more often with each other. Furthermore, sequences that are on the same chromosome interact more frequently with each other than those on different chromosomes, which even holds true for megabase-scale interactions. A Hi-C experiment produces a paired-end sequencing dataset in which each read corresponds to a region of the genome that interacts in space. These facts can be exploited to arrange contigs into chromosome-scale scaffolds. One of the major benefits of chromosomal conformation capture for scaffolding is the comparatively low price point and a broad range of interactions that can be captured. Several experimental and algorithmic solutions, both academic and commercial, are based on these principles (Burton et al. 2013; Putnam et al. 2016; Ghurye et al. 2017).

Another method that is becoming increasingly popular for both contig assembly and scaffolding are linked reads (Weisenfeld et al. 2017). These are short reads that have been barcoded to denote their DNA fragment of origin. One commercial solution is the 10x Genomics chromium system. Here, gel-beads containing around 5 molecules of DNA of around 100 Kb and a specific barcode are created. The DNA is fragmented and the barcode attached. After sequencing, two reads belonging to the original bead can be identified by their common barcode. The idea is analogous to a BAC, where a local complexity reduction eases the assembly problem by removing many possible repetitive overlaps. Similarly, as the DNA molecule of input is ~100 Kb, this long-range information can be exploited for scaffolding. There are also scaffolding solutions that are not based on sequencing, namely optical maps that mark specific sequence patterns with fluorescent dyes. The distance of these patterns together with their identification on the contigs can be used to arrange them. Finally, synteny based scaffolding can be used. To this end, the assembly of a different species is used to anchor the contigs. This praxis was originally applied to the great ape genomes to order them into chromosomes despite poor contig contiguity using the human genome as an anchor. It has the advantage of not requiring the production of additional data but can be confounded by large scale intraspecies structural diversity or karyotypic differences (Rice and Green 2019).

1.3.3 Gaps

Gaps are still an integral part of genome assemblies, and it is important to understand why they occur. In the following, I will summarize the most common reasons for unresolved sequences in assemblies (see figure 4).

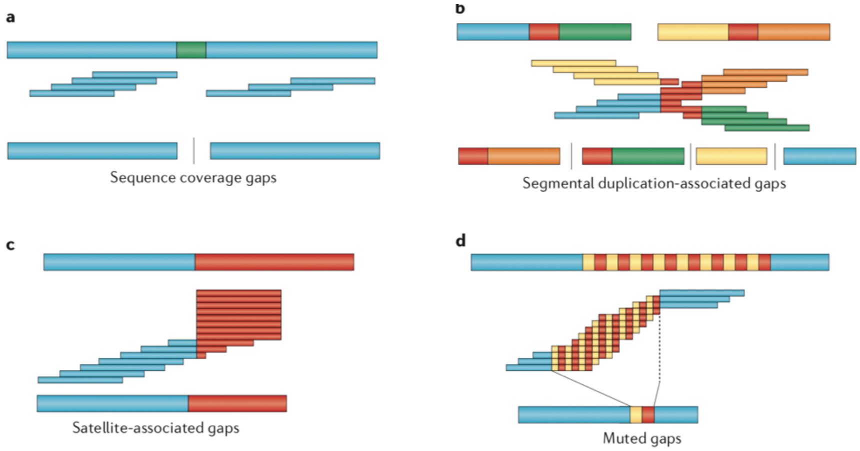

a
Sequence coverage gaps
b
Segmental duplication-associated gaps
c
Satellite-associated gaps
d
Muted gaps

Figure 4: Overview of different types of genome assembly gaps. Regions in colors other than blue represent repetitive elements. From Chaisson et al, 2015

Initially, a common reason for gaps was a lack of sequencing coverage. In the Sanger days, this was mainly because of shallow sequencing coverage. As we have mentioned above, at around 5X sequencing, we expect to find around 130,000 gaps in an assembly, a calculation that is in hand with empirical observations (The Chimpanzee Sequencing and Analysis Consortium 2005). For MPS data, however, sequencing gaps due to lack of coverage is unlikely. Rather, regions with very high or low GC content are underrepresented in this data, which exhibits a GC-dependent coverage bias (Minoche, Dohm, and Himmelbauer 2011; Benjamini and Speed 2012). Contrary to Sanger sequencing gaps, these are hardly mitigable with increased coverage but remain an inherent limitation for platforms like Illumina. The main reason for gaps, however, are different kinds of repeats. The contig assembly strategies we have outlined above usually break contigs at ambiguous positions of the assembly graph and do not necessarily output the corresponding number of repeats as contigs. After scaffolding, the missing sequence between two contigs therefore often corresponds to a common repetitive element, such as a SINE or LINE (Gnerre et al. 2011). While common repeats are not an issue for long-read assemblies, other types of repetitive elements are. Segmental duplications amount to around 3% of the human genome. They are often several Kb in size and copies have a high identity (>95%) to each other, making them particularity difficult to assemble regardless of the underlying technology (Alkan, Sajjadian, and Eichler 2011; Huddleston et al. 2014; Chaisson et al. 2015). Large segmental duplications are usually unresolved in whole-genome shotgun assemblies, with different copies being compressed into a single representation. Similarly, more than 1/3 of the 540 remaining euchromatic gaps in the most recent iteration of the human genome (GRCh38) are flanked by large segmental duplications (Chaisson, Wilson, and Eichler 2015). Another source of gaps is short tandem repeats, microsatellites or centromeres. These regions are composed of a short sequence motif that is repeated over and over again, sometimes up to several Mb. When comparing short read to long-read assemblies, up 80% of the closed gaps consisted of long stretches (5-

10 Kb) of STRs (Chaisson et al. 2015; Chaisson, Wilson, and Eichler 2015; Gordon et al. 2016). Centromeres, on the other hand, are very long stretches of higher-order repeats. It was not until this year (2019) that the first centromere of a human chromosome was fully sequence resolved (Jain, Olsen, et al. 2018).

As we have seen above, genome assemblies are usually a haploid representation (or haploid compressions) of a diploid genome. Variation in genomes is a continuum that ranges from single base substitution to Mb scale structural differences. These structural variants might be interpreted as repeats by the assembler and broken apart, such as for example the HLA region (Raymond et al. 2005). In certain regions, these differences may lead to haploid assembly representations that are biologically meaningless, as the reconstructed haplotype is a mosaic of the two. In this case, the reconstruction is not broken apart, but rather appears to contain no gap when in reality sequence is missing. This situation is called a muted gap (Chaisson, Wilson, and Eichler 2015).

1.3.4 Measures of completeness and quality

Given the lack of a ground truth set, measuring the quality and completeness of an assembly is a tricky undertaking. One of the most frequently reported statistics regards the contiguity of the assembly: the weighted fragment length or N50. This number describes the hypothetical length of a sequence such that 50% of the *assembly* is contained in sequences of at least that size (International Human Genome Sequencing Consortium 2001). It can be reported for contigs or scaffolds. A major limitation of the N50 is that it is not comparable across assemblies, as the assembly length will be specific to each of them. If the genome size of an organism is known, the NG50 statistic reports the analogous sequence length taking into consideration the genome size, instead of the assembly size. While these are useful metrics, they just measure the connectivity of the assembly and don't necessarily make any statement about its quality or error rate. The latter two are particularly hard to assess in the face of genetic variation, as it can be hard to discern an assembly error from a natural variant. Often, structural consistencies, for example, detected based on discordant paired-end sequencing, are curated by hand.

Several algorithms to use sequencing data to "polish assemblies" have been developed, and they are usually part of a standard assembly pipeline. These include platform-specific such as quiver or arrow for PacBio or nanopolish for nanopore (Chin et al. 2013; Simpson et al. 2017). Common tools for polishing with Illumina data are pilon, and a platform-agnostic one is racon (Walker et al. 2014; Vaser et al. 2017). After they are run, the proportion of homozygous alternative high confidence variant calls can serve as a proxy to estimate the base substitution error rate.

One approach to measure the genome quality is to annotate its genes and look at how many of those that are observed commonly in other species, and may, therefore, be considered universal, are present. Tools like CEGMA or BUSCO deal with this issue (Parra, Bradnam, and Korf 2007; Simão et al. 2015). Finally, orthogonal validation by a different data type or source has often been used as a proxy for an assembly's quality. For example, a comparison to finished BAC sequences which can be regarded as a gold standard. A plethora of orthogonal datasets has been created for NA12878, the "standard" genome which is widely used for benchmarking purpose (Zook et al. 2014). Finally, hydatidiform moles, such as the CHM1 cell line offer an opportunity to remove allelic variation and therefore a major potential pitfall in assessing errors. This fact has been exploited in recent benchmarking studies as well as for assemblies (Huddleston et al. 2014; H. Li et al. 2018)

1.4 The study of primate genomes

Among the primary motivations to study primate genomes is the wish to better understand the origins of our own species (Varki, Geschwind, and Eichler 2008). Cataloging and classifying the genetic differences that make us human requires a comprehensive detection and discovery of all kinds of genetic variants. Initially, genome projects of more distantly related species, such as mouse, focused on aspects of the genome that are conserved across long evolutionary distances by comparing them to human. Primates on the other hand, and hominins particularly, open up the opportunity to study not only what is conserved and shared, but also what differs between them and humans and is specific to either lineage. Among those genetic differences lay the changes that distinguish us from the rest of apes. This qualitative difference in analyses also requires a stronger focus on the assembly's quality to better distinguish a difference from an artifact (Marques-Bonet, Ryder, and Eichler 2009).

Chimpanzees as our closest extant evolutionary relative have been a particular center of attention in this context. The initial sequence of the chimpanzee genome was reported in 2005 (The Chimpanzee Sequencing and Analysis Consortium 2005). Its analysis showed that the average genome-wide divergence rate between the two species was only ~1.06%, albeit with regional differences between chromosomes and variation dependent on sequence context and chromosomal position. This implies that almost 99% of the two species genomes are identical. However, this number only regarded single nucleotide differences, which only constitute a single kind of genetic variation.

Taking into consideration insertions and deletions as well, around 1.5% of either genome was determined to be specific for each lineage. In this context, it is important to note that the chimp was originally assembled with low coverage Sanger data averaging 3.6 X on autosomes, which was estimated to cover around 94% of the genome. While subsequent improvements included several finished BACs and increased the coverage, and thus the quality of the assembly, its fragmentation and lack of representation of certain regions hindered several analyses, especially those assessing variation beyond SNVs. The initial assembly indubitably provided several important insights regarding human-chimpanzee divergence, adaptive substitution rates, or difference in mobile element dynamics. However, several potential key differences fall into regions of rapid structural genomic change, many of which were therefore missed in the initial comparisons, partially owing to the assembly's quality. These include examples of major regulatory changes affecting, among others, human neural development (McLean et al. 2011; Boyd et al. 2015). Additionally, several human specific genes with roles in brain morphology, neuronal count and synapse densities lay within regions of segmental duplications, which are usually inaccessible through draft assemblies (Dennis et al. 2012; Charrier et al. 2012; Florio et al. 2015; Ju et al. 2016, 1). These examples underline the importance of high-quality genomic resources, especially for the chimpanzee which has proved an invaluable resource to many aspects of genomic analysis. Beyond the above importance for the study of human biology and evolution with implications for biomedical research, the chimpanzee genome has also been extensively used as an outgroup for human population genetics and the study of human origins (The 1000 Genomes Project

Consortium 2015; Mallick et al. 2016; R. E. Green et al. 2010; Meyer et al. 2012).

In the context of primates, the study of great apes beyond chimpanzees is also particularly relevant to understand human evolution. Orangutan was the third ape whose genome was released, showing lower rates of structural genome evolution and mobile element insertion (Locke et al. 2011). The gorilla genome followed shortly thereafter, thus providing genomic representations for members of all 4 extant hominid genera (Scally et al. 2012). Its analysis challenged the simplistic view of great ape speciation, by showing that that along 30% of the genome the human is closer to gorilla than to the chimpanzee. Finally, the Bonobo was the last great ape species have its genome sequenced (Prüfer et al. 2012).

Beyond great apes, a set of key primates to be sequenced was identified in the last decade (Marques-Bonet, Ryder, and Eichler 2009). These include species of either biomedical relevance as animal models or at key evolutionary splits. The rhesus macaque was the first primate after the chimp to have its genome sequenced (Gibbs et al. 2007). Its phylogenetic position is at the split between old world monkeys and apes. Additionally, it is of key biomedical importance as an animal model for several diseases such as HIV, Arteriosclerosis or COPD (Phillips et al. 2014). At the split between old-world primates and new world monkey, the marmoset was chosen, which also serves as a primate model for the study of neurodegenerative diseases such as Parkinson, or autism (The Marmoset Genome Sequencing and Analysis Consortium et al. 2014; Zhao, Jiang, and Zhang 2018). Several other species have been included in this "first wave" of primate genome, and new ones are being produced at an ever-increasing rate thanks to the democratization of the assembly process (Warren et al. 2015; Schmitz et al. 2016; Larsen et al. 2017). It is also important to note that several of the aforementioned limitations with respect to assembly's quality have been tackled for some species. Namely gorilla, chimpanzee, orangutan, and macaque all now have high-quality long read assemblies available which constitute a substantial improvement for the resolution of genomic analysis, particularly for structurally challenging regions, facilitating their assessment (Gordon et al. 2016; Kronenberg et al. 2018; He et al. 2019).

While the primary focus and justification to invest in primate genome sequencing have been the study of human evolution and potential biomedical applications, there are important reasons beyond that to study their genetics. Currently, ~60% of the ~500 recognized species are threatened with extinction and ~70% have dwindling population sizes. This situation has been created almost exclusively due to habitat loss because of anthropogenic pressures such as deforestation for agriculture and farming (Estrada et al. 2017). The species survival of all great apes is threatened with varying degrees, with some being among the most threatened of all primates. The partially extremely small population sizes have also left a clear genomic footprint in some of these species (Xue et al. 2015; Nater et al. 2017; Valk et al. 2019). It is furthermore becoming increasingly clear that potential conservational interventions such as genetic rescues should not be undertaken without prior knowledge of the genetic makeup of the population in question (Supple and Shapiro 2018; Robinson et al. 2019). Given the current levels of extinction risk the genetic analysis of primates is particularly pressing. Beyond the satisfaction of our intellectual curiosity and our drive to improve the human condition through research, we humans also hold a unique and clear responsibility towards these animals whose livelihood we are pushing to the verge of extinction - or beyond - through the actions of our own species.

1.5 Peculiarities of Y chromosomes

Primates harbor an X-Y genetic sex-determination system, in which the male is the heterogametic sex[2]. This means that apart from the diploid set of autosomes, females harbor two X chromosomes, and males one X and one Y chromosome. The latter contains the "genetic master-switch" to determine the biological sex of an individual (Sinclair et al. 1990). Sex chromosomes such as X-Y emerged several times independently in animals, and the Y chromosome in primates originated at the root of eutherian mammals (Cortez et al. 2014). In humans, the Y chromosome is the only nuclear chromosome that does not recombine at all over the majority of its length, the so-called male-specific region of the Y chromosome (MSY). Additionally, it only has ½ the effective population size of autosomes and is transmitted exclusively via males. The absence of recombination implies that all loci in the MSY region are in linkage with one another, which means that selection acting upon a single locus will affect the whole chromosome. All these circumstances lead to evolutionary peculiarities (Bachtrog 2013).

The X and Y chromosomes evolved from a pair of ordinary autosomes roughly 200-300 million years ago (Lahn and Page 1999). The process was initiated by a series of large-scale inversion that led to recombinational arrest between the chromosome pairs. Subsequently,

[2] Sex chromosomes have evolved independently many times over in animals, as have X-Y sex determination systems in general. In this chapter, I will talk about primate Y chromosomes as Y chromosomes, without specifying this at each occasion.

the chromosomes diverged, but the common ancestry is still recognizable in their sequence in which the inversions are marked by the presence of distinct evolutionary "strata". Theory predicts that in the absence of recombination Y chromosomes would degenerate (Charlesworth and Charlesworth 2000). This process implies the chromosome-wide decay of functional genes over time, which has even led to the complete loss of the chromosomes in some species (Castillo, Marti, and Bidau 2010). As a result of Y chromosome degeneration, it harbors a small number of genes and a large proportion of repetitive DNA. Despite the considerable gene loss, a core set of essential genes has been conserved in mammals, several of which play ubiquitous "housekeeping" functions (Bellott et al. 2014).

The human Y chromosome was initially assembled and released in 2001 (Skaletsky et al. 2003). Its release showed that it is approximately 52 Mb in size, of which ~23 Mb correspond to resolved euchromatic sequences, which harbor 78 protein-coding genes. For comparison: the ancient homolog of the Y chromosome, the X chromosome, is comprised of around 150 Mb of euchromatic sequence that contains ~800 genes (Ross et al. 2005). A substantial proportion of the few genes left on the Y have functions related exclusively to male biology, and are predominantly transcribed in testis (Lahn and Page 1997).

The initial analysis of the Y chromosomes sequence showed that it is composed of a set of discrete euchromatic sequence classes, each with their own evolutionary history (see figure 5). The X-degenerate region is the reminiscent of the shared ancestry between the X and the Y chromosome. It contains the ubiquitously transcribed single-copy genes of the Y chromosome, and only one single testis-specific gene: The sex-determining master-switch *SRY*. The X-degenerate region holds most of the single copy genes that are ancestral to mammals and retain wide expression throughout tissues beyond testes. These genes are dosage sensitive and their homologs escape X-inactivation in females (Bellott et al. 2014).

The X-transposed region is the result of a massive transposition from the X onto the Y chromosome that happened around 3-4 million ago (i.e. after the split between humans and chimpanzees). It is ~3.4 Mb long and shares ~99% identity with its counterpart on the X chromosome and is therefore considered a segmental duplication. After the transpositions, it was subsequently split up into two discontinuous blocks by an inversion and harbors the lowest gene density of the Y chromosomal sequence classes.

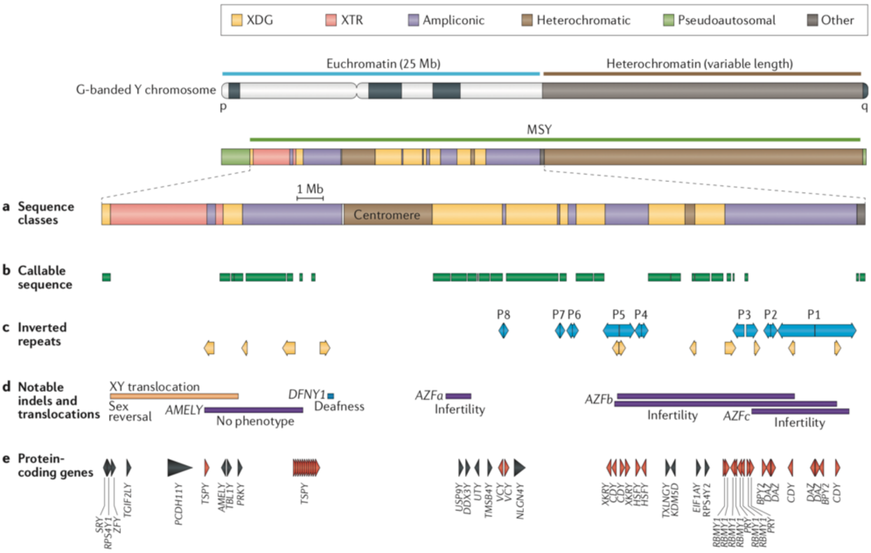


Figure 5: Overview of the structure of the human Y chromosome. Taken from Jobling and Smith, 2017.

The third euchromatic sequence classes are the ampliconic regions. These regions take up over 10 Mb of the MSY and consist of several massive segmental duplications that exhibit > 99.9% identity between their copies. The ampliconic regions are made up of 7 discrete segments and their most pronounced structural feature are 8 massive palindromes, or inverted repeats. Both arms of the palindromes exhibit exceptionally large identity to each other and are between 9 Kb and 1.45 Mb in length. The ampliconic regions are the most gene dense on the Y chromosome, and most of the genes within them show testis-specific expression. Ampliconic genes exhibit high inter-species divergence. In the face of Y chromosomal degeneration, the palindromic structure has been proposed as a solution to maintain identity of the genes laying within them through a process called gene conversion, a non-reciprocal exchange of DNA from one copy to another (Skaletsky et al. 2003). Additionally they show ample evidence of copy number variation in human populations (Skov, Consortium, and Schierup 2017; Lucotte et al. 2018; Vegesna et al. 2019).

Lastly, the pseudo-autosomal regions at the extremities of the Y chromosomes are regions that share homology to the extremities of the X chromosome and are needed for proper chromosomal segregation. These regions of the sex chromosomes behave like autosomes, in that they undergo recombination between the X and the Y chromosome.

The human Y chromosome also harbors massive proportions of heterochromatic sequence. The majority of the long arm is composed of a large heterochromatic expansion whose sequence is unresolved. This heterochromatic sequence amounts to around half the total size of the Y chromosome altogether and is elusive to current sequencing strategies.

The X-degenerate region can be considered "well-behaved" in the context of genome assembly. It is among the parts that would usually assemble well and is also accessible to resequencing experiments (see figure 5) (Jobling and Tyler-Smith 2017). Ampliconic regions, on the other hand, are among the most complex regions of the human genome to reconstruct, as assemblies will usually fail on these large segmental duplications. Additionally, sequencing coverage is proportional to the copy number of the DNA, which led many early sequencing projects to choose female samples to at least get sufficient coverage for the X chromosome (Jobling and Tyler-Smith 2017). For those reasons, very few Y chromosome assemblies for primates (and mammals in general) exist to date. In 2010, the chimpanzee Y chromosome was released (Hughes et al. 2010). Comparisons to the human one showed they are remarkably different, more so than any other chromosome. Their ampliconic regions are 44% larger than humans and contain 19 palindromes, only seven of which are also found in humans. Around 30% of the chimpanzees' MSY region has no counterpart that can be aligned in humans. Despite the ~1% autosomal divergence between the two species, the chimpanzee Y chromosome contains only 2/3 as many functional gene families on the Y chromosome. In 2012, the rhesus Y chromosome was released (Hughes et al. 2012). It is markedly smaller than either human or chimp, with the euchromatic portion accounting for only 11 Mb. The majority of it (9.5 Mb) is taken up by X-degenerate sequences, and only ~500 Kb are ampliconic. This assembly showed that, despite the degeneration of the Y chromosome since its formation, the bulk of genes were lost early on and the overall gene content stayed remarkably stable since.

The human, chimp, and rhesus Y chromosome assemblies were produced with a process called single haplotype iterative mapping and sequencing, or SHIMS. Analogous to BAC hierarchical shotguns, SHIMS is a clone by clone approach that yields a very high-quality assembly. However, given the need for iterative rounds of detection of overlaps between BACs, it is also very time and labor-intensive, and thus expensive. Several other strategies to mitigate the assembly problem for the Y chromosome have been proposed. As mentioned, strategies using whole genome shotgun data usually tend to underrepresent the Y chromosome, or completely omit them by sequencing female samples. (Tomaszkiewicz, Medvedev, and Makova 2017). One possible solution to identify Y chromosomal strategies is to sequence both female and male samples. In this way, a reduction of the sequences that present in the female sample and the male sample aid the identification of Y chromosomal sequences (Cortez et al. 2014). A conceptually different approach is the physical isolation of Y chromosomes prior to sequencing. This not only makes it easier to identify Y chromosomal sequences, but also drastically reduces the number of potential overlaps to the rest of the genome, thus rendering the assembly process far less complex. The isolation can be achieved by flow cytometry, also called flow sorting. In a nutshell, this method works the following way (Doležel et al. 2012): Condensed chromosomes are saturated with intercalating dies that are incorporated into the DNA in amounts that are proportional to the size and GC content of the chromosome. The chromosomes are then suspended in droplets, which can be assessed by an array of lasers.

The intercalants lead to specific absorption spectra, which in theory allow the physical separation of droplets containing the desired chromosomes by a charge, given that they differ sufficiently in size and GC composition from other chromosomes. The separated chromosomes can then be sequenced. This idea has successfully been applied to assemble the gorilla Y chromosome, which revealed that its gene repertoire is more similar to humans than that of the chimp, despite the larger evolutionary distance (Tomaszkiewicz et al. 2016). Flow sorting has also been used in the context of other Y chromosomes, such as Pig and Chimp (Skinner et al. 2016; Hughes et al. 2010). Additionally, the now broad availability of long-read sequencing offers additional avenues to reconstruct complex regions of the genome or chromosome. However, all methods mentioned above still result in fragmented chromosomal representations. Their exact biases are hard to assess given that large structural complexities of the Y chromosome, and not least because they are usually applied to species with previously uncharacterized genomes.

2 OBJECTIVES

1) Create an improved, high-quality reference assembly for the chimpanzee
2) Explore the different contributions of a set of diverse sequencing strategies and library preparation methods to the quality of a new genomic representation for the chimpanzee
3) Establish a workflow to sequence unamplified, flow-sorted chromosomes on a Nanopore sequencing device
4) Apply this workflow to assemble the first high quality genomic representation of a human Y chromosome of African origin.

3 RESULTS

3.1 A 3-Way Hybrid Approach to Generate a New High-Quality Chimpanzee Reference Genome (Pan_tro_3.0)

Kuderna, Lukas F. K., Chad Tomlinson, LaDeana W. Hillier, Annabel Tran, Ian T. Fiddes, Joel Armstrong, Hafid Laayouni, et al. 2017. "A 3-Way Hybrid Approach to Generate a New High-Quality Chimpanzee Reference Genome (Pan_tro_3.0)." *GigaScience* 6 (11). https://doi.org/10.1093/gigascience/gix098.

GigaScience, 6, 2017, 1–6

doi: 10.1093/gigascience/gix098
Advance Access Publication Date: 30 October 2017
Data Note

DATA NOTE

A 3-way hybrid approach to generate a new high-quality chimpanzee reference genome (Pan_tro_3.0)

Lukas F. K. Kuderna[1,2], Chad Tomlinson[3], LaDeana W. Hillier[3], Annabel Tran[4], Ian T. Fiddes[5], Joel Armstrong[5], Hafid Laayouni[1,6], David Gordon[7,8], John Huddleston[7,8], Raquel Garcia Perez[1], Inna Povolotskaya[1], Aitor Serres Armero[1], Jèssica Gómez Garrido[1,2], Daniel Ho[9], Paolo Ribeca[10], Tyler Alioto[1,2], Richard E. Green[11,12], Benedict Paten[5], Arcadi Navarro[1,2,13], Jaume Betranpetit[1], Javier Herrero[4], Evan E. Eichler[7,8], Andrew J. Sharp[9], Lars Feuk[14,*,†], Wesley C. Warren[3,*,†] and Tomas Marques-Bonet[1,2,13,*,†]

[1]Institut de Biologia Evolutiva, (CSIC-Universitat Pompeu Fabra), PRBB, Doctor Aiguader 88, Barcelona, Catalonia 08003, Spain, [2]CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Baldiri i Reixac 4, 08028, Barcelona, Spain, [3]McDonnell Genome Institute, Department of Medicine, Department of Genetics, Washington University School of Medicine, 4444 Forest Park Ave., St. Louis, MO 63108, USA, [4]Bill Lyons Informatics Centre, UCL Cancer Institute, University College London, 72 Huntley Street, London WC1E 6DD, UK, [5]Genomics Institute, University of California Santa Cruz and Howard Hughes Medical Institute, 1156 High Street, Santa Cruz, CA 95064, USA, [6]Bioinformatics Studies, ESCI-UPF, Pg. Pujades 1, 08003, Barcelona, Spain, [7]Department of Genome Sciences, University of Washington School of Medicine, Box 355065, Seattle, WA 98195, USA, [8]Howard Hughes Medical Institute, University of Washington, Box 355065, Seattle, WA 98195, USA, [9]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA, [10]The Pirbright Institute, Ash Road, Pirbright, Woking, GU24 0NF, UK, [11]Department of Biomolecular Engineering, University of California Santa Cruz, 1156 High Street, Santa Cruz, CA 95060, USA, [12]Dovetail Genomics, Santa Cruz, 2161 Delaware Ave., Santa Cruz, CA 95060, USA, [13]Institucio Catalana de Recerca i Estudis Avancats (ICREA), Passeig Lluís Companys 23, Barcelona, Catalonia 08010, Spain and [14]Department of Immunology, Genetics and Pathology, Science for Life Laboratory, Box 815, Uppsala University 751 08 Uppsala, Sweden

Received: 10 November 2016; **Revised:** 2 June 2017; **Accepted:** 8 September 2017

*Corresponding address. Lars Feuk, Box 815, Uppsala University 751 08 Uppsala, Sweden; Tel: +46 18 4714827; Fax: +46 18 558931;
E-mail: lars.feuk@igp.uu.se; Wesley C. Warren, 4444 Forest Park Ave., St. Louis, MO 63108, USA; Tel: +1 314 286-1899; Fax: +1 314 286-1810;
E-mail: wwarren@wustl.edu; Tomas Marques-Bonet, Doctor Aiguader 88, 08003 Barcelona, Spain; Tel: +34 93 316 08 87; Fax: +34 93 316 09 01;
E-mail: tomas.marques@upf.edu
†Equal contribution.

Abstract

The chimpanzee is arguably the most important species for the study of human origins. A key resource for these studies is a high-quality reference genome assembly; however, as with most mammalian genomes, the current iteration of the chimpanzee reference genome assembly is highly fragmented. In the current iteration of the chimpanzee reference genome assembly (Pan_tro_2.1.4), the sequence is scattered across more then 183 000 contigs, incorporating more than 159 000 gaps, with a genome-wide contig N50 of 51 Kbp. In this work, we produce an extensive and diverse array of sequencing datasets to rapidly assemble a new chimpanzee reference that surpasses previous iterations in bases represented and organized in large scaffolds. To this end, we show substantial improvements over the current release of the chimpanzee genome (Pan_tro_2.1.4) by several metrics, such as increased contiguity by >750% and 300% on contigs and scaffolds, respectively, and closure of 77% of gaps in the Pan_tro_2.1.4 assembly gaps spanning >850 Kbp of the novel coding sequence based on RNASeq data. We further report more than 2700 genes that had putatively erroneous frame-shift predictions to human in Pan_tro_2.1.4 and show a substantial increase in the annotation of repetitive elements. We apply a simple 3-way hybrid approach to considerably improve the reference genome assembly for the chimpanzee, providing a valuable resource for the study of human origins. Furthermore, we produce extensive sequencing datasets that are all derived from the same cell line, generating a broad non-human benchmark dataset.

Keywords: chimpanzee reference genome; assembly, genomics

Data Description

Creating a non-human sequencing benchmark dataset

To test the potentially combinatorial power of varied sequencing and mapping strategies, we created several different datasets on different platforms to try to leverage the advantages of each, as the shortcomings of 1 sequencing strategy might be compensated for by another [1]. All datasets are derived from a single male western chimpanzee ("Clint," Coriell identifier S006007), the same individual used to generate the current Chimpanzee genome assembly. We produced ∼120-fold sequence coverage of overlapping 250-bp reads (∼450-bp fragment) on the Illumina HiSeq 2500 platform, offering high accuracy and throughput, but comparatively short reads; ∼9-fold sequence coverage from 43 Pacific Biosciences SMRT-Cells with P5-C3 chemistry on the RSII instrument, offering long reads at lower accuracy; Illumina TruSeq Synthetic long reads at around 2-fold coverage, offering long-range information derived from local assemblies of ∼10-Kb fragments [2]; 1 lane of *in vitro* proximity ligation read pairs (prepared as a Chicago library by Dovetail Genomics) [3] sequenced on the Illumina HiSeq 2000 platform, offering spatial contact information of the chromatin, that can be exploited for scaffolding.

These diverse datasets complement the resources that were already available for the same cell line, namely 6-fold coverage of ABI Sanger capillary reads used for the initial chimpanzee genome assembly, a 100-bp paired Illumina HiSeq data, a fosmid library at 6-fold physical coverage with available end sequences, a Bacterial Artificial Chromosome (BAC) library at 3-fold physical coverage with available end sequences and around 700 finished BACs [4]. Altogether, these data constitute an extensive non-human and non-model organism benchmarking dataset for different sequencing strategies.

Assembly generation

We generated a complete *de novo* assembly for the chimpanzee with a combination of the datasets. At each step of our assembly, we measured increase in contiguity by means of the N50 statistic, which is defined as the length of a contig or scaffold such that 50% of the assembly bases are contained in contigs or scaffolds of at least that length. The starting point of our assembly scaffolding efforts are contigs generated with DISCOVAR *de novo* [5] from 250 bp of paired-end reads. These reads are derived from a 450-bp library, resulting in pairs that overlap over a ∼50-bp region, a feature that is exploited by the assembler. While based on Illumina sequencing, these libraries have recently been shown to produce assemblies superior in contiguity when compared to assemblies derived from conventional Illumina libraries [6]. The DISCOVAR base assembly had a contig N50 of 87 Kbp, and was then scaffolded using proximity ligation read-pairs generated by the Chicago method [3] and sequenced on the Illumina platform. These data increased the scaffold N50 to 26 Mbp. Notably, individual scaffolds exceed lengths of 75 Mbp, and therefore already reach the order of magnitude of full chromosomal arms. We sought to take advantage of these highly contiguous scaffolds and attempt closure of remaining gaps with long-read single-molecule sequences by PacBio using PBJelly (PBJelly, RRID:SCR_012091) [7]. By this means, we filled over 38 000 gaps (or 55%) among all scaffolds, and in so doing increased the contig N50 by over 320% to 283 Kbp when compared to the DISCOVAR base assembly (see Table 1). While we went on to further improve the assembly with additional data (see below), these statistics give an approximation of the contiguity that can be expected for *de novo* assemblies of previously unsequenced species using our 3-way hybrid approach: contigs derived from overlapping 250-bp paired-end reads to scaffold with *in vitro* HiC, and fill remaining gaps with PacBio data. When the contiguity metrics of this intermediate assembly are compared to other representative non-human primate genomes (as annotated by NCBI Refseq category, July 1, 2016; see the Supplementary Data), we observed superior contiguity in contig structure within our assembly compared to all others. The only exception is the gorilla genome, recently assembled from deep (∼75-fold) long-read sequences [8]. However, our stepwise method offers an approach that is considerably cheaper.

Table 1: Assembly statistics comparing the previous chimpanzee assembly, our intermediary assembly based on the 3-way hybrid and the finished assembly Pan_tro_3.0

	Pan_tro_2.1.4	3-way hybrid (intermediary)	Pan_tro_3.0
Scaffold N50, bp	8 925 874	26 681 610	26 972 556
Contig N50, bp	50 665	282 774	384 816
Contig N90, bp	7231	41 655	53 112
Assembly length, bp	3 309 577 923	2 992 696 208	3 231 154 112
Assembly length w/o Ns, bp	2 902 338 968	2 990 712 612	3 132 603 062
Scaffolds	24 129	45 000	44 448
Contigs	183 827	76 674	72 226
Gaps	159 698	31 674	26 715

In this context, we defined gaps at stretches of at least 10 consecutive "Ns" in the assembly. Contigs are defined as contiguous stretches of sequence without gaps.

Assembly refinement and comparison to Pan_tro_2.1.4

For the final release of the chimpanzee assembly, we created a reference assembly that leveraged previous resources generated from the same individual [4]. First, we merged in regions from Pan_tro_2.1.4 that were derived from Clint and gapped in our assembly. It is known that Pan_tro_2.1.4 contains sequences from different chimpanzees. To do so, we extracted flanking sequence regions of gaps in our assembly and mapped all to Pan_tro_2.1.4, keeping only unique and concordant mappings that do not span any gaps within Pan_tro_2.1.4, and merged the spanned Pan_tro_2.1.4 sequence in.

To ensure that accuracy was not sacrificed for continuity gains, we utilized various methods to measure error. Given that our assembly likely contained some erroneous links between contigs or misassembled contigs as a result of *de novo* assembly, conformational mapping, or merging mistakes, we first used discordant mapping of fosmid end sequences (~40-Kbp insert size) to identify any large misassemblies. We identified 17 such scaffold errors and manually broke apart each. We also sought to correct any remaining single base substitutions or small indels (<6 bp) with a series of custom mapping and base integration programs (see the Supplementary Data). With the same Illumina data used to generate the DISCOVAR base assembly, we corrected more than 500 000 single base or indel errors. Most of these residual errors are presumably derived from regions where PacBio data were incorporated into the assembly, as this platform is known to have an elevated error rate. As another measure of quality, we produced whole-genome alignments to Pan_tro_2.1.4 and found that our assembly aligns with, on average, 99.9% identity, and the magnitude of remaining differences can thus be reasonably explained by the allelic diversity of western chimpanzees [9].

For our final assembly, named Pan_tro_3.0, we integrated previously available finished clone sequences derived from Clint where possible. Pan_tro_3.0 spans 2.95 Gbp in ordered and oriented chromosomal sequences. An additional 140 Mbp of sequence is assigned to chromosomes, but their order and orientation are unknown, and 123 Mbp remain of unknown chromosomal origin. Pan_tro_3.0 has a genome-wide contig and scaffold N50 of 385 Kbp and 27 Mbp, respectively, constituting an improvement in contiguity over Pan_tro_2.1.4 of 760% and 300%, respectively (see Fig. 1A and Table 1). We observed this increase across all non-finished chromosomes, with the most pronounced effect on the X chromosome (see Fig. 1B). This chromosome shows the highest degree of fragmentation in Pan_tro_2.1.4, likely due to the fact that the effective sequence coverage on the sex chromosomes is only half that of the autosomes, namely around 3-fold in the original assembly. We increased the contig N50 on the X chromosome by 3250% from 13 Kbp to 422 Kbp, thus bringing its contiguity to the range observed on autosomes.

Overall, we decreased the number of contigs by more than 60%, from 183 860 to 72 226, and the number of gaps by 83%, from 156 857 to 26 715. As gap structures between the assemblies may not correspond, we identified filled gaps from Pan_tro_2.1.4 by extracting their flanking regions and mapping them onto Pan_tro_3.0. By keeping only unique and concordant mappings that do not span any gaps in Pan_tro_3.0, we estimate the sequences of 122 943 (77%) gaps to be filled, amounting to 60.3 Mbp of sequence. The majority of these fill sequences are comparably short (see Fig. 1C) and significantly enriched in interspersed genomic repeats, with 58% of them ($P < .0001$, feature permutation test) intersecting with repeats. Of these, around 16 Mbp are fully embedded within fill sequences, corresponding to, amongst others, more than 29 650 novel short interspersed nuclear element (SINE) annotations and 20 888 novel long interspersed nuclear elements (LINE) annotations.

Repeat resolution

Large genomic repeats constitute a major confounding factor in genome assembly and are therefore one of the main reasons for their fragmentation, and thus the assembly repeat representation can be a proxy of its quality. To assess the repeat resolution of interspersed repeats, we masked Pan_tro_3.0 using RepeatMasker (RepeatMasker, RRID:SCR_012954) [10], selecting chimpanzee-specific repeats, resulting in 1.64 Gbp (52.2%) being annotated as repeats. The proportion of repetitive elements is similar in Pan_tro_2.1.4 (50.9%); however, given the large amount of newly resolved sequences, this translates into a substantial increase in annotated repeats. Specifically, we annotate 164 Mbp of novel repeats in Pan_tro_3.0, comprising around 10% of the whole repeat annotation. We observe this increase consistently across all families of interspersed repeats (see Fig. 1D). The increases range as high as 300% for satellite sequences, corresponding to an additional 68.2 Mbp of newly resolved sequence in this category. We also increased the amount of annotated SINE by 27.9 Mbp, including 83 637 additional resolved copies of *Alu* elements. We find the increase in annotations to be negatively correlated with age for *Alu* elements, and thus find the highest increase (8.8%) for the youngest and least divergent subfamily (*Alu*Y), suggesting that common high-identity repeats are now better resolved. We furthermore added 38.2 Mbp of sequence annotated as LINEs to the assembly. We also observed a noteworthy increase in annotated long terminal repeats, adding 15.9 Mbp to this repeat category, corresponding to

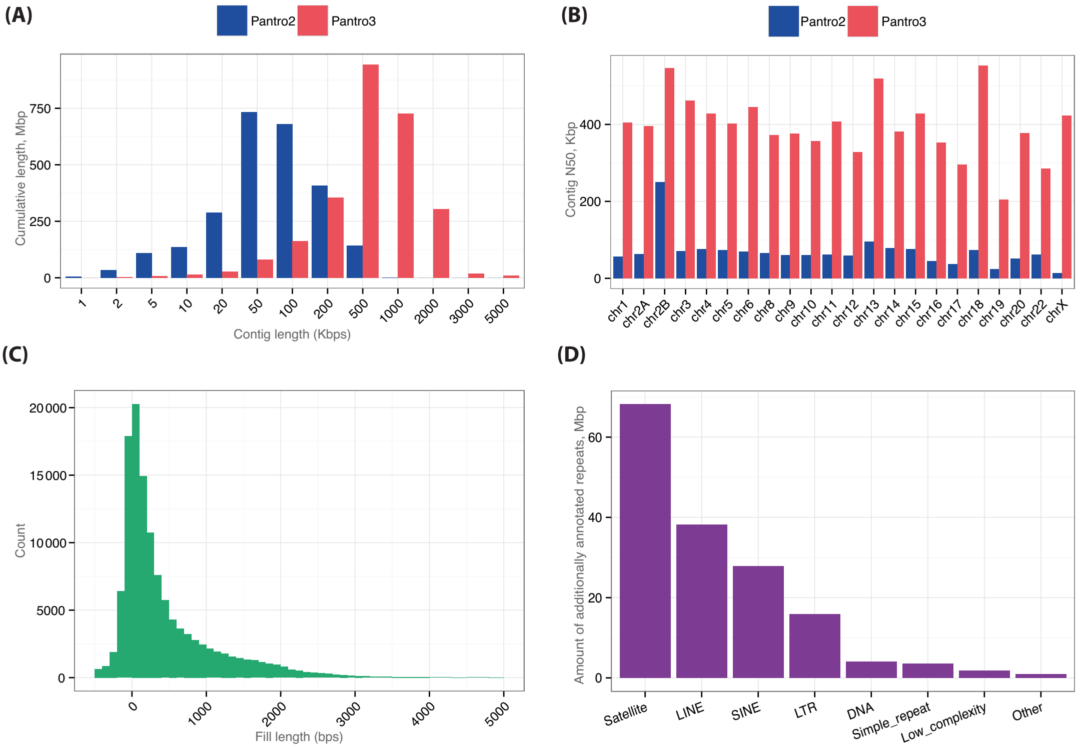


Figure 1: (A) Genome-wide distribution of contig lengths between Pan_tro_2.1.4 and Pan_tro_3.0. The peak for Pan_tro_3.0 is shifted to higher values by an order of magnitude. (B) Increase in contig N50 for all chromosomes that were not finished with clones in Pan_tro_2.1.4 or Pan_tro_3.0. (C) Length distribution of filled gaps in Pan_tro_3. Negative values constitute wrongly separated overlapping contig ends in Pan_tro_2.1.4. (D) Increase in annotated interspersed repeats separated by repeat family.

30 574 additional annotations of endogenous retroviruses in the genome. When comparing all types of interspersed repeats between Pan_tro_2.1.4 and Pan_tro_3.0, we find a median increase of 4.7% of sequence, highlighting that repeat resolution is much improved in Pan_tro_3.0 (see Supplementray Table S4).

Representation of segmental duplications

To analyze the representation of segmental duplications in Pan_tro_3.0, we applied 2 alternative approaches. First, we performed a whole-genome assembly comparison (WGAC) to compare repeat-free sequences of the assembly to itself [11]. This method identifies duplicated sequence in blocks of at least 1 Kbp with 90% identity or higher. Excluding unplaced contigs, we found 140 Mbp of non-redundant duplicated sequence in Pan_tro_3.0 chromosomes, or 4.46% of the non-gap bases in the assembly, results that are consistent with previous read-depth estimates for chimpanzee [12] and analyses of high-quality, finished human genome assemblies (see Supplementary Data S3). Second, we identified duplications by whole-genome shotgun sequence detection (WSSD), which identifies duplications at least 10 Kbp long with over 94% identity by detecting regions of increased read depth compared to known unique regions [13]. We used 31 366 275 Sanger capillary reads derived from Clint, and found 51 Mbp of duplicated sequence meeting these criteria on placed chromosomes, compared to 68 Mbp detected by WGAC.

Genome wide, we discovered 178 245 redundant pairwise alignments corresponding to 388 Mbp of non-redundant sequence greater than 1 Kbp in length and 90% identity (12.39% of the genome sequence excluding gaps) by WGAC, and 63 Mbp of duplicated sequence by WSSD (compared to 284 Mbp WGAC ≥10 Kbp, >94% identity). We then compared Pant_tro_3.0 to the human reference genome assembly GRCh38, an assembly that is based on a BAC hierarchical shotgun assembly strategy and may therefore be considered the gold standard with respect to representation of segmental duplications. We note similar proportions of bases in segmental duplications on chromosomal scaffolds (4.46% in Pan_tro_3.0 vs 5.56% in GRCh38); however, we note an elevated genome-wide rate of bases in duplications when including unplaced and unlocalized scaffolds. This suggests that our assembly includes false-positive paralogous regions (see Supplementary Table S1).

Gene annotation

We produced a new gene annotation based on projections from all human transcripts in the GENCODE annotation V24 set combined with RNA-seq data derived from the brain, heart, liver, and testis from 3 different individuals [14]. To quantify the effect of the underlying sequence on the annotation, we annotated Pan_tro_2.1.4. with the same data. We observed improvements in gene annotation in Pan_tro_3.0 in all considered metrics. We increased the number of recovered consensus gene models for protein coding transcripts by 2.7% and are now able to project and annotate 89.5% of the GENCODE human coding transcripts onto the new assembly. The average coverage of these transcripts within the genome is 98.9%, a gain of 2%. We also observe an increase of 6.6% in transcripts with multiple mappings. We checked for newly resolved exonic sequences in filled gaps with respect to Pan_tro_2.1.4, and find 17 818 exons, amounting to 851 Kbp of non-overlapping sequence, to be fully embedded within them. Altogether, we retrieved models for 77 858 coding transcripts, corresponding to the isoforms of 20 373 coding genes.

We find 5039 human coding transcripts corresponding to 2660 genes with predicted frameshift mutations in Pan_tro_2.1.4 to human, but not in Pan_tro_3.0. Conversely, we find 674 genes with predicted frameshift mutations to human that are present in Pan_tro_3, but not in Pan_tro_2.1.4. Given that both assemblies are mainly based on data from the same individual (with the exception of chromosome 21 and around 28% of chromosome 7 in Pan_tro_2.1.4, which were derived from a different individual), the majority of these predictions constitute either allelic variation or putative sequence errors in Pan_tro_2.1.4.

In summary, we describe a hybrid assembly approach to obtain a more complete *de novo* chimpanzee reference genome assembly, substantially increasing contiguity metrics within it. Our proposed assembly method should be easily applicable to different organisms of similar genomic architecture.

Availability of supporting data

We have corrected several orientation errors in the sequences described in this article. The corrected sequences can be found in the associated Gigascience Database.

Supporting data are available through the *Giga*DB database (*Giga*DB, RRID:SCR_004002) [15]. This whole-genome shotgun project has been deposited at DDBJ/ENA/GenBank under the accession AACZ00000000. The version described in this paper is version AACZ04000000. The assembly is available at https://www.ncbi.nlm.nih.gov/assembly/GCF_000001515.7 and at the UCSC genome browser under the identifier panTro5. The assembly denominated Pan_tro_2.1.4 in the manuscript refers to Pan_troglodytes-2.1.4 with the RefSeq assembly accession number GCF_000001515.6.

Additional file

Kuderna_et_al.SUPPLEMENTARY_resubmission.docx

Abbreviations

bp: base pairs; Kbp: kilo base pairs; Mbp: mega base pairs; indel: insertion-deletion; SINE: short interspersed nuclear element; LINE: long interspersed nuclear element; LTR: long terminal repeat; ERV: endogenous retrovirus; WGAC: whole-genome assembly comparison; WSSD: whole-genome shotgun sequence detection.

Competing interests

EEE is on the Scientific Advisory Board (SAB) of DNAnexus, Inc. REG is the co-founder of Dovetail Genomics.

Funding

J.G.G. is funded by the RED-BIO project of the Spanish National Bioinformatics Institute (INB) under grant number PT13/0001/0044. The INB is funded by the Spanish National Health Institute Carlos III (ISCIII) and the Spanish Ministry of Economy and Competitiveness (MINECO). L.F.K.K. is supported by an FPI fellowship associated with BFU2014-55090-P (FEDER); L.F. is supported by the Swedish Foundation for Strategic Research F06-0045 and the Swedish Research Council; E.E.E. is an investigator of the Howard Hughes Medical Institute. A.J.S. is supported by US National Institutes of Health (NIH) grants DA033660, HG006696, HD073731, and MH097018, and research grant 6-FY13-92 from the March of Dimes. This work was supported, in part, by grants from the NIH (grants R01HG002385 and U24HG009081 to E.E.E., HG007990 and HG007234 to B.P.). T.M.B. is supported by MINECO BFU2014-55090-P (FEDER), BFU2015-7116-ERC, and BFU2015-6215-ERC, Fundacio Zoo Barcelona and Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement de la Generalitat de Catalunya.

Author contributions

T.M.B., W.C.W., and L.F. conceived the study. L.F.K.K., C.T., L.W.H., and R.E.G. produced and analyzed the assembly. I.F., J.A., J.G.G., T.A., B.P., A.T., H.L., J.B., R.G.P., I.P., A.S.A., J.He, P.R., D.H., A.N., and A.J.S. produced, analyzed, and interpreted the assembly and annotations. D.G., J.Hu, and E.E.E. analyzed segmental duplications. T.M.B., W.C.W., and L.F.K.K. wrote the manuscript with input from all authors.

Acknowledgements

We would like to acknowledge Bojan Obradovic and James Richardson for sequence contribution. The authors acknowledge the use of the UCL Legion High Performance Computing Facility (Legion@UCL), and associated support services, in the completion of this work.

References

1. Goodwin S, Mcpherson JD, Mccombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 2016;**17**:333–51.
2. Kuleshov V, Xie D, Chen R et al. Whole-genome haplotyping using long reads and statistical methods. Nat Biotechnol 2014;**32**:261–6.
3. Putnam NH, O'Connell BO, Stites JC et al. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. Genome Res 2016;**26**:342–50.
4. Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 2005;**437**:69–87
5. Weisenfeld NI, Yin S, Sharpe T et al. Comprehensive variation discovery in single human genomes. Nat Genet 2014;**46**:1350–5.
6. Chaisson MJP, Wilson RK, Eichler EE. Genetic variation and the de novo assembly of human genomes. Nat Rev Genet 2015;**16**:627–40.
7. English AC, Richards S, Han Y et al. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. PLoS One 2012;**7**:e47768.

8. Gordon D, Huddleston J, Chaisson MJP et al. Long-read sequence assembly of the gorilla genome. Science 2016; **352**(6281):aae0344.
9. Prado-Martinez J, Sudmant PH, Kidd JM et al. Great ape genetic diversity and population history. Nature. 2013;**499**: 471–5.
10. Smit A, Hubley R, Green P. RepeatMasker Open-3.0. RepeatMasker. 1996. www.repeatmasker.org (27 May 2016, date last accessed).
11. Bailey JA, Yavor AM, Massa HF et al. Segmental duplications: organization and impact within the current human genome project assembly. Genome Res 2001;**11**:1005–17.
12. Cheng Z, Ventura M, She X et al. A genome-wide comparison of recent chimpanzee and human segmental duplications. Nature 2005;**437**:88–93.
13. Bailey JA, Gu Z, Clark RA et al. Recent segmental duplications in the human genome. Science 2002;**297**:1003–7.
14. Ruiz-Orera J, Hernandez-Rodriguez J, Chiva C et al. Origins of de novo genes in human and chimpanzee. PLoS Genet 2015;**11**:e1005721.
15. Kuderna LF, Tomlinson C, Hillier LW et al. High quality chimpanzee reference genome (Pan_tro_3.0) from hybrid assembly approach. GigaScience Database 2017. http://dx.doi.org/10.5524/100327.

3.2 Selective Single Molecule Sequencing and Assembly of a Human Y Chromosome of African Origin

Kuderna, Lukas F. K., Esther Lizano, Eva Julià, Jessica Gomez-Garrido, Aitor Serres-Armero, Martin Kuhlwilm, Regina Antoni Alandes, et al. 2019. “Selective Single Molecule Sequencing and Assembly of a Human Y Chromosome of African Origin.” *Nature Communications* 10 (1): 4. https://doi.org/10.1038/s41467-018-07885-5.

nature
COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-018-07885-5 OPEN

Selective single molecule sequencing and assembly of a human Y chromosome of African origin

Lukas F.K. Kuderna[1], Esther Lizano[1], Eva Julià[2,3], Jessica Gomez-Garrido[4], Aitor Serres-Armero[1], Martin Kuhlwilm[1], Regina Antoni Alandes[4], Marina Alvarez-Estape[1], David Juan[1], Heath Simon[4,5], Tyler Alioto[4,5], Marta Gut[4,5], Ivo Gut[4,5], Mikkel Heide Schierup[6,7], Oscar Fornas[3,5] & Tomas Marques-Bonet[1,4,5,8,9]

Mammalian Y chromosomes are often neglected from genomic analysis. Due to their inherent assembly difficulties, high repeat content, and large ampliconic regions, only a handful of species have their Y chromosome properly characterized. To date, just a single human reference quality Y chromosome, of European ancestry, is available due to a lack of accessible methodology. To facilitate the assembly of such complicated genomic territory, we developed a novel strategy to sequence native, unamplified flow sorted DNA on a MinION nanopore sequencing device. Our approach yields a highly continuous assembly of the first human Y chromosome of African origin. It constitutes a significant improvement over comparable previous methods, increasing continuity by more than 800%. Sequencing native DNA also allows to take advantage of the nanopore signal data to detect epigenetic modifications in situ. This approach is in theory generalizable to any species simplifying the assembly of extremely large and repetitive genomes.

[1] Institut de Biologia Evolutiva, (CSIC-Universitat Pompeu Fabra), PRBB, Doctor Aiguader 88, Barcelona, Catalonia 08003, Spain. [2] Institut Hospital del Mar d'Investigacions Mèdiques (IMIM), Carrer del Doctor Aiguader 88, PRBB Building, Barcelona 08003, Spain. [3] Centre for Genomic Regulation (CRG), The Barcelona Institute for Science and Technology, Carrer del Doctor Aiguader 88, Barcelona 08003, Spain. [4] CNAG-CRG, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Baldiri Reixac 4, Barcelona 08028, Spain. [5] Universitat Pompeu Fabra (UPF), Doctor Aiguader 88, Barcelona 08003, Spain. [6] Bioinformatics Research Center, Aarhus University, C.F. Moellers Alle 8, DK-8000 Aarhus C, Denmark. [7] Department of Bioscience, Aarhus University, Ny Munkegade 116, DK-8000 Aarhus C, Denmark. [8] Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig Lluís Companys 23, Barcelona, Catalonia 08010, Spain. [9] Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona, Edifici ICTA-ICP, c/ Columnes s/n, Cerdanyola del Vallès, Barcelona 08193, Spain. These authors contributed equally: Lukas F.K. Kuderna, Esther Lizano, Oscar Fornas, Tomas Marques-Bonet. Correspondence and requests for materials should be addressed to L.F.K.K. (email: lukas.kuderna@upf.edu) or to E.L. (email: esther.lizano@upf.edu) or to T.M-B. (email: tomas.marques@upf.edu)

Recombinational arrest in the common ancestor of the X and Y chromosomes led to the degeneration and accumulation of large amounts of repetitive DNA on the Y chromosome along its evolutionary trajectory[1]. Furthermore, many sequencing efforts have traditionally chosen female samples, as the hemizygous nature of the sex chromosomes leads to half the effective sequencing coverage on both of them in a male, resulting in inferior genome assemblies[2,3]. Together, these causes have led to an underrepresentation of Y chromosomes in genomic studies and proper characterization of the Y chromosome in only a handful of mammalian species through a time- and labor-intensive clone by clone approach[4–7]. One strategy to reduce the complexity of the assembly problem for the Y chromosome is to isolate it by flow cytometry, thus dramatically reducing the potential amount of overlaps of repetitive regions in the context of the whole genome[3]. Notwithstanding previous efforts, which sought to do this have faced some drawbacks, as the material has been heavily amplified post sorting to increase yield[8]. Whole-genome amplification (WGA) introduces biases that are detrimental to genome assembly, such as unequal sequence coverage and chimera formation, as well as limited fragment length[9]. Moreover, these methods lead to the loss of epigenetic modifications that can now be directly determined from the signal data from nanopore sequencers[10]. Additionally, previous efforts to assemble the Y chromosome purely from flow-sorted material did so using the gorilla[8], a species with a previously uncharacterized Y chromosome, meaning that potential biases in the assembly cannot be detected without a gold standard reference to compare with, such as human. Integrating single-molecule sequencing has been shown to produce far superior whole-genome shotgun (WGS) assemblies than sequencing by synthesis platforms[11–14]. Furthermore, the MinION sequencing platform from Oxford Nanopore Technologies has recently been used to create the most contiguous human WGS assembly to date[15] and to resolve the structure of the human Y-chromosome centromere[16]. To take advantage of these benefits, we developed a protocol to sequence native, unamplified flow-sorted DNA on the MinION sequencing device.

Results

Flow sorting and sequencing. We sorted approximately 9,000,000 individual Y chromosomes from a lymphoblastoid cell line (HG02982) from the 1000 Genomes Project, whose haplogroup (A0) represents one of the deepest known splits in humans[17] (see Fig. 1a). Given the large volume in which the chromosomes were sorted, and potential issues with residual dyes that are necessary for the sorting process, we devised a purification protocol to bring the

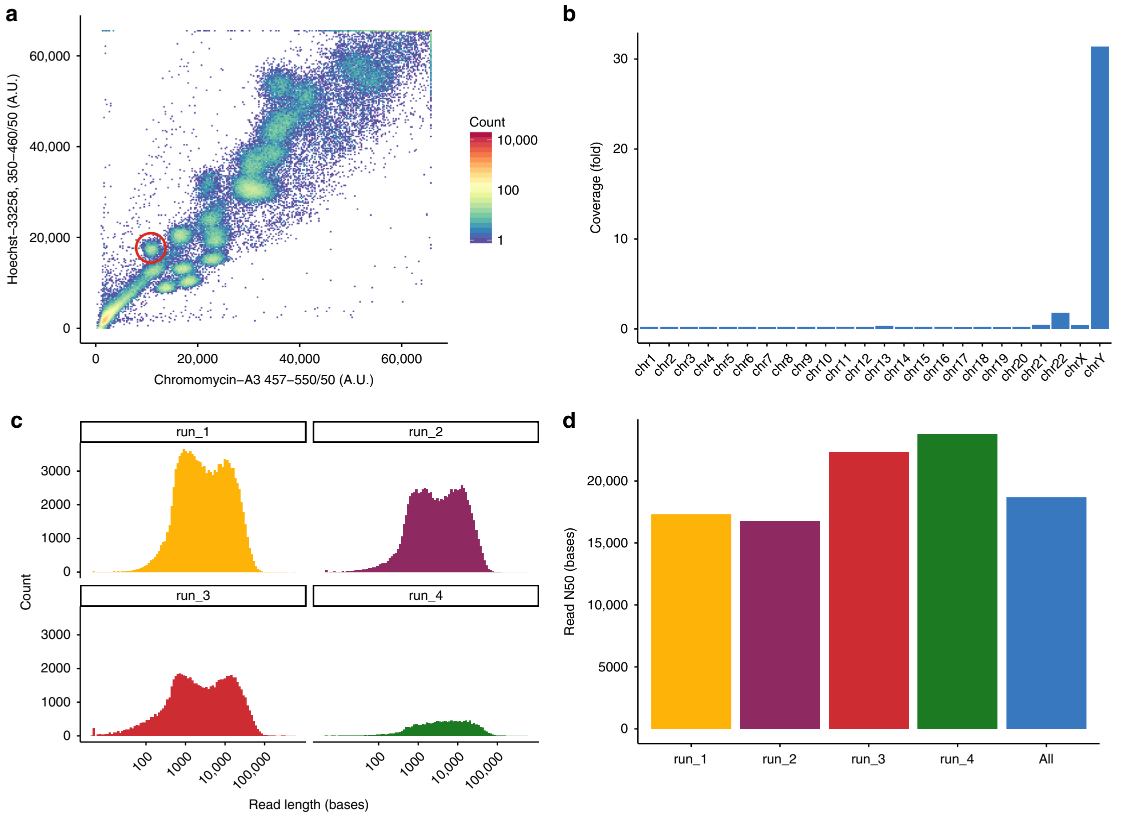


Fig. 1 Flow-sorting and sequencing specificity. **a** Flow-karyogram of a human genome. The different clusters correspond to different chromosomes. The red circle delimits the cluster corresponding to the Y chromosome used for this project. **b** Enrichment specificity of the sequencing data. Sequences on the Y chromosome are ~110-fold enriched compared with WGS sequencing. Chromosome 22 partially co-sorts with Y. All other chromosomes are depleted. **c** Read length (log10 scale) distribution of the four runs. **d** N50 values for all four runs and the combined dataset. Colors in panels **c** and **d** correspond to the different runs. Source data are provided as a Source Data file

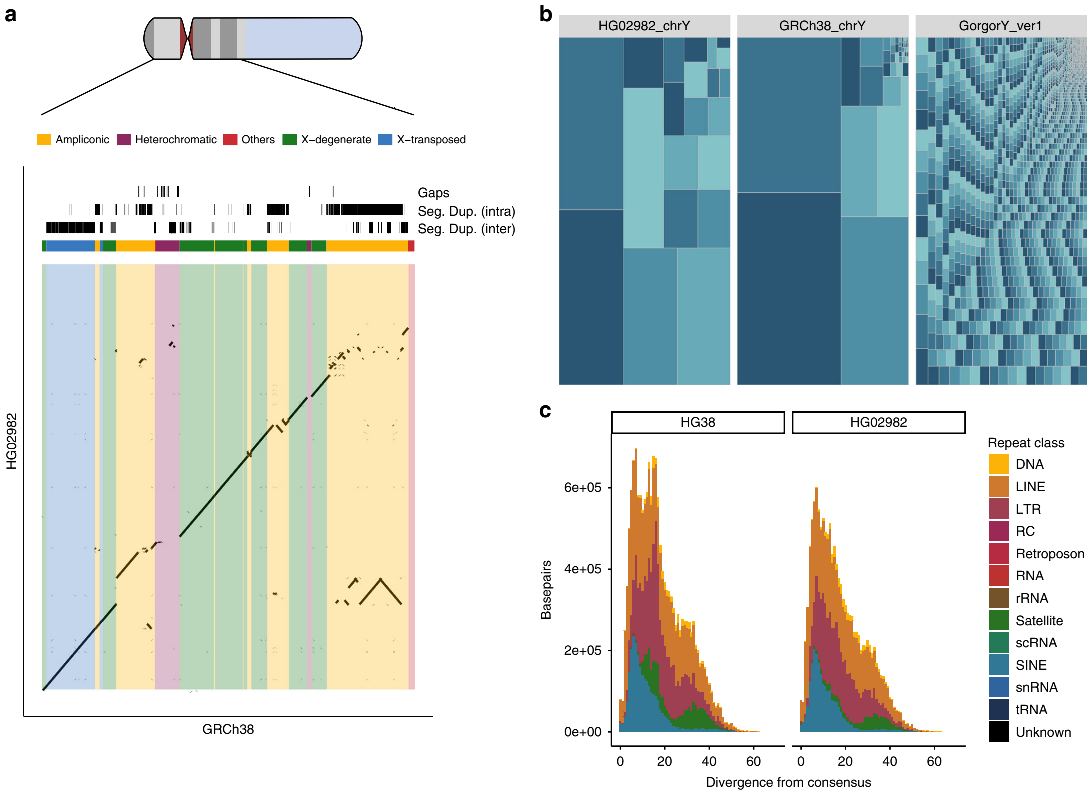


Fig. 2 Chromosome-Y assembly overview and comparisons. **a** Dot-plot comparing the resolved MSY of GRCh38 with HG02982. The reconstruction is highly continuous along most sequence classes, with ampliconic regions showing a higher degree of fragmentation. Seg. Dup. (intra) refers to intra-chromosomal segmental duplications, Seg. Dup. (inter) refers to inter-chromosomal segmental duplications. Altogether, ~50% of the of the Y chromosomes resolved sequence space in GRCh38—> 13 Mb—are annotated as segmental duplications. **b** Treemap comparing the contiguity of HG02982 chrY with GRCh38 chrY and the gorilla Y chromosome by Tomaszkiewicz et al. The size of each rectangle corresponds to the size of a contig within each of the assemblies. Neighboring rectangles are colored differently as a visual aid. **c** Repeat landscape of common, interspersed repeats annotated equally in GRCh38 and HG02982. Common repeats—including very recent ones—are well resolved in HG02982. The exception are satellite sequences, and a population of somewhat divergent (~20%) LTR elements, which are absent in HG02982 (see supplementary Figures 7-9). Source data are provided as a Source Data file

DNA into conditions suitable for sequencing. We ran four Oxford Nanopore MinION flowcells to generate 305,528 reads summing to over 2.3 Gb of data. The yields per flowcell varied considerably from 897.6 Mb to 163.8 Mb (see Fig. 1c and Supplementary Table 2). Sequencing yields were on the lower end of the reported spectrum (75 Mb–5.5 Gb per flowcell in[15]), but read N50 surpassed most of them, ranging from 16.8 to 23.8 kb[15] (see Fig. 1d, supplementary Figures 1–2). Additionally, for the same flow-sorted material we ran an Illumina MiSeq lane for 2 × 300 cycles but including four rounds of PCR amplification. To check the enrichment specificity, we aligned the reads to the human reference genome (GRCh38) and calculated the normalized coverage on each chromosome. Taking into account the size of the Y chromosome and its haploid nature, we find it to be over 110-fold enriched compared with a random sampling from the human genome (see Fig. 1b, Supplementary Figures 3–5, Supplementary Note 2 and Supplementary Data 1–2).

Y-chromosome assembly and comparison with GRCh38. We used the Nanopore data to construct a de novo assembly using Canu[18]. We performed a self-correction by aligning the reads used for assembly and called consensus using Nanopolish[10], correcting a total of 127,809 positions. Finally, the Illumina library served to polish residual errors within the assembly using pilon[19]. By this means, we corrected a further 101,723 single-nucleotide positions and introduced 105,640 small insertions and 6983 small deletions. We also explored further polishing options and found that running one additional round of error correction with racon[20] potentially resolves several remaining errors, despite also introducing additional discordances (see Supplementary Table 4, Supplementary Notes 1,3 and Supplementary Figures 10–13). The final assembly is comprised of 35 contigs, with an N50 of 1.46 Mb amounting to 21.5 Mb of total sequence, in contrast to a contig N50 of 6.91 Mb of the GRCh38 Y-chromosome assembly. Compared with the gorilla Y-chromosome assembly with a contig N50 of 17.95 kb[8], our assembly is two orders of magnitude more contiguous (see Fig. 2b).

The Y chromosome is comprised of a set of discrete sequence classes[4]. To check the completeness of our assembly, we assessed how well each of them is represented. After retaining only single best placements, we were able to align 21.1 Mb, or 98.4% of its

Table 1 Assembly statistics overview

Seq. class	Aln. HG02982 (b)	HG02982 ID SNP (%)	HG02982 ID SNP + InDel(%)	Rec. HG02982 (%)	Aln. NA24385 (b)	Rec. in NA24385 (%)	Len. w/o gaps (b)
Ampliconic	6,146,087	99.91	99.67	62.67	5,242,461	53.46	9,807,089
Heterochromatic	543,005	99.66	99.31	32.77	171,045	10.32	1,656,797
Others	295,160	99.47	99.18	385.59	63,973	83.57	76,547
Pseudo-autosomal	2,219,743	99.58	99.13	78.02	117,626	4.13	2,844,939
X-degenerate	8,537,493	99.95	99.81	98.94	8,238,733	95.48	8,628,904
X-transposed	3,374,011	99.94	99.81	99.21	1,474,610	43.36	3,400,750

Summary of sequence class coverage of HG02982 versus GRCh38, as well as the contigs from NA23385 identified as derived from the Y chromosome. The proportion of recovered sequences and % identity are calculated over the resolved sequences in GRCh38, excluding gaps. There are currently 30.8 Mb of unresolved sequence (represented by the ambiguous base N) in the reference Y chromosome of GRCh38, the vast majority of which belongs to heterochromatin on the q arm

Aln.: aligned bases to GRCh38, ID.: percent identical bases in GRCh38, Rec.: recovered proportion from GRCh38, Len.: length in GRCh38

total length, with 99.9% of identical bases on average (see Fig. 2a). We recovered the full-length (~ 99% of the annotated length in GRCh38) reconstructions of both the X-transposed and the X-degenerate regions. Although the X-degenerate region can be considered a single-copy region due its distant common ancestry with the X chromosome, the X-transposed region emerged only after the split between humans and chimpanzees[21]. The largest sequence class on the Y chromosome is comprised of ampliconic regions, which amount to around 30% of the euchromatic portion and sum to 9.93 Mb. These regions contain eight massive, segmentally duplicated palindromes, all of which share >99.9% identity between their two copies, with the largest one spanning over 2.90 Mb. We find this region to be the most challenging to reconstruct, with fragmented and collapsed sequences, but are nevertheless able to recover 6.14 Mb, or 62.7% of its length in GRCh38. Surprisingly, we recover only 78% of the pseudo-autosomal regions (PARs). We observed a rather steep drop-off in coverage coinciding with the PAR-1 boundary on GRCh38. As we are sequencing native, unamplified DNA, the genomic coverage is directly proportional to the number of copies of the underlying sequenced region[22]. We compared the mapped coverage of our raw data on GRCh38 and find that PAR-1 exhibits only around 72% of the average coverage of the whole chromosome (19.8-fold versus 27.3-fold). We observe the drop-off in coverage to coincide sharply with the PAR-1 boundary (see Supplementary Figure 6). Finally, of the remaining sequence classes, we are able to recover around 32.8% of the resolved heterochromatic regions, and multiple instances of the remaining unclassified sequences (referred to as other; see Table 1 and Supplementary Data 3).

To contrast our approach to a long-read WGS assembly, we assembled the publicly available PacBio dataset from the Ashkenazim son from the Genome in a Bottle Consortium[23], which has a sequencing depth comparable to ours on the sex chromosomes (~ 30X). We identified 193 contigs mapping to the Y chromosome, with an N50 of 213 kb, covering 15.3 Mb, or around 28% less than by our approach. The WGS fails to assemble roughly 56.6% of the X-transposed region and 47% of the ampliconic regions (see Table 1).

Comparative gene annotation. We performed a comparative annotation to check the completeness of our assembly at the gene level. To this end, we projected all Gencode (v. 27, GRCh38) annotations on the Y chromosome onto our assembly and annotated them there. Due to its peculiar evolutionary trajectory, the gene-space on the Y chromosome is degenerated, and any remaining genes can generally be classified into two categories: on one hand there are single-copy genes, which are broadly expressed beyond the testis. On the other, there are multi-copy genes within the ampliconic regions, which are mainly involved in spermatogenesis[24]. We recover the complete gene set of the genes in the male-specific region of the Y chromosome (MSY) region and are therefore able to annotate all single-copy genes. Furthermore, we are able to retrieve at least one member of all multi-copy gene families. For four out of nine of these gene families, we are additionally able to resolve further copies within our assembly (see Supplementary Data 4–5). We also note that four genes (*ASMTL, IL3R, P2RY, SLC25*) from a comparatively short syntenic block of around 200 kb are partially missing from our assembly due to the aforementioned technical challenges in the PAR-1 region. Mapping the raw data onto GRCh38 show that this is an artifact, presumably due to insufficient coverage in this region.

Structural variants. We produced a stringent call set of structural variants (SVs) derived from alignments to GRCh38 using Assemblytics[25]. We detect 347 SVs at least 50 bp in size (931 variants at least 10 bp in size) of which 82 are at least 500-bp long (see Fig. 2c, Supplementary Figures 14–15, Supplementary Table 3). The cumulative length of these variants sums to 184 kb. We observe a 4.8-fold excess number of deletions versus number of insertions, amounting to a twofold excess of bases in deletions versus bases in insertions. Although a deletion bias for nanopore-based assemblies had previously been reported[15], we find the strength of this bias to be decreasing in our analysis, probably reflecting improvements in base-calling accuracy. To check the presence of large-scale copy number variation in multi-copy genes, we additionally determined the chromosome-wide copy number based on a read depth approach using the Illumina data. We find extensive genic copy number variation, with expansions in five of the nine multi-copy genes, when compared with the reference individual. Among these, we find expansions in *RBMY, PRY, BPY2,* and *DAZ*, all members of the AZFc region locus with implications for male fertility. Although these expansions are to some degree represented in our assembly, the precise genomic architecture remains challenging to reconstruct. Due to the high degree of similarity between copies, several of them will be collapsed in the assembly specially in the AZFc region. Finally, to assess concordance with previous studies, we compared our SV calls with those generated by the 1000 Genomes Project, which contains the same cell line used for this study[26]. We manually confirm the presence of all structural three variants called in HG02982 in the 1000 Genomes Project in our data by checking the overlap of calls produced by orthogonal approaches (see Supplementary Figures 16–18).

CpG methylation status. Finally, we called the methylation status of 5-methylcytosines (5-mC) at CpG positions from the Nanopore signal data using a recently developed model implemented in Nanopolish[10]. To assess potential biases on the CpG methylation status introduced by our workflow, we also produced whole-genome bisulfite sequencing data (WGBS) for the same cell line. We calculated the methylation frequency (i.e., the proportion of reads supporting 5-mc at a given CpG) for both datasets. For positions where both datasets have at least 10-fold coverage ($n =$ 4654), we observe a good concordance in the methylation frequency with a Pearson's r of 0.816 (see Supplementary Figure 19 and Supplementary Table 5). Remaining differences might be attributable to differences in sensitivity, variation in the methylation state, or alternative modifications such as 5-hydroxymethylation, which cannot be distinguished from 5-mC by WGBS[10]. Additionally, detecting the 5-mC status on the Y chromosome from long reads in our methodology has the advantage of allowing to interrogate regions that are not accessible to WGBS with short reads, namely the PAR, the X-transposed region, and to some degree the Ampliconic regions. We interrogated the methylation state of CpG 200-bp upstream of the transcription start site (TSS) in protein-coding genes falling within the different sequence classes of the Y chromosome. Genes from the PAR, X-degenerate, and X-transposed regions are expressed throughout the body, whereas Ampliconic genes have testis specific expression[24]. In agreement with these patterns, we find the genes within PAR, X-degenerate, and X-transposed regions to show low degrees of CpG methylation at TSS. Within the Ampliconic regions, the distribution of methylation frequencies of CpGs at TSS shows an overall high degree of methylation and is therefore consistent with the expected downregulation of these genes in lymphoblastoid cells (see Supplementary Figures 20–21). Nevertheless, single-copy resolution is not possible due to potential mapping ambiguities.

Discussion

Here, we report the first successful sequencing and assembly of native, flow-sorted DNA on an Oxford Nanopore sequencing device, without previous amplification. We apply our methodology to assemble the first human Y chromosome of African origin to benchmark our approach. This is arguably the most challenging human chromosome to assemble due to its high repeat and segmental duplication content, and hence a good test-case to explore the possibilities and limitations of this approach. With the exception of bacterial artificial chromosome-based assemblies, we are able to reconstruct the Y chromosome to unprecedented quality in terms of contiguity and sequence class representation. We show that we not only outperform previous efforts that sought to achieve a similar goal of reconstructing Y chromosomes[8], but also accomplish a better reconstruction on all sequence classes than the Y chromosomal sequences derived from a long-read WGS assembly. Additionally, our method is orders of magnitude cheaper than reconstructions from WGS data too, especially considering that twice the desired Y chromosomal target coverage is needed on the autosomes. Given the current developments in sequencing throughput, a single-MinION flowcell should now be sufficient to assemble a whole human Y chromosome. Furthermore, it is becoming clear that the upper read length boundary is only delimited by the integrity of the DNA, suggesting the possibility that complete Y-chromosome assemblies, including full resolution of amplicons, might be possible in the near future. Notwithstanding, some challenges to obtain ultra-long reads from flow-sorted chromosomes are still to be overcome, as sorting sufficient material for this protocol is a substantial endeavor. It also is worth noting that our efforts to sequence the same input material on Pacific Biosciences Sequel platform have been fruitless, presumably due to interference of residual dyes with the sequencers optical detection system. Despite the technical challenges of flow-sorting single chromosomes, the method described here offers the opportunity to take advantage of the benefits of long-range data together with local complexity reduction. Given different chromosomes that are sufficiently distinguishable in terms of size and GC content, immediate applications are either very complex chromosomes, such as the human Y, or extremely large genomes with a very high degree of common repeats, which have long challenged traditional WGS approaches, such as wheat, the loblolly pine, or the axolotl[27–30].

Methods

Chromosome preparation for flow karyotyping. Mitotic chromosomes in suspension were prepared as follows (adapted from[31] with some modifications): the lymphoblastoid cell line HG02982 (purchased from Coriell, cat. no. HG02982) were cultured in RPMI-1640 medium supplemented with 2mM L-glutamine (Invitrogen, ref. 21875-034), 15% fetal bovine serum and antibiotics (penicillin and streptomycin (Invitrogen, ref. 15140-122)) at initial concentration no <150,000 viable cells per ml. Near confluence, cells were subcultured to 50%. After 24 h, the cells were blocked in mitosis by adding Colcemid to the culture (10 μg ml^{-1} demecolcine solution (Gibco, ref. 15210-040)) to a final concentration of 0.1 μg ml^{-1} and incubated for an additional 6–7 h. To swell and stabilize mitotic cells, they were centrifuged 5 min at 300 × g at room temperature. The pellet was slowly resuspended in 10 ml hypotonic solution (Hypotonic solution: 75 mM KCl, 10 mM $MgSO_4$, 0.2 mM spermine, 0.5 mM spermidine. pH 8.0), incubated for 10 min at room temperature. After the incubation in the hypotonic solution, the swollen cells were centrifuged at 300 × g for 5 min. The cell pellet was resuspended in 1.5 ml of ice-cold polyamine isolation buffer (PAB: 15 mM Tris, 2 mM EDTA, 0.5 mM EGTA, 80 mM KCl, 3 mM dithiothreitol, 0.25% Triton X-100, 0.2 mM spermine, 0.5 mM spermidine. pH 8.0) for 20 min to release the chromosomes.

To ensure the integrity of the chromosomes, their morphology was checked before staining them. To this end, the pellet was vigorously vortexed for 30 s to liberate the chromosomes from the mitotic cells. The suspension was filtered through a 35 μm mesh filter and stored at 4 °C until its sorting.

Finally, chromosomes were stained with chromomycin-A3 (Sigma, ref. C2659) and Hoechst 33,258 (Invitrogen, ref. H3569) at a final concentration of 40 μg ml^{-1} and 5 μg ml^{-1}, respectively, in presence of divalent cations (10 mM $MgSO_4$ (Sigma, ref. 60142)). Staining was performed for at least 8 h at 4 °C, to allow the dyes to equilibrate. Before the sample analysis on a cell sorter, potassium citrate was added to a final concentration of 10 mM (Sigma, ref. 89306) to enhance peak resolution in the flow karyotype.

Chromosome sorting. Flow karyotyping for chromosome sorting was performed on BD Influx cell sorter (Becton Dickinson, San Jose, CA), a jet-in-air cell sorter that was selected for its relatively easy manual daily fine-tuning and high-resolution capabilities. Of the five available lasers, only the blue (488 nm laser at 200 mW), deep-blue (457 nm laser at 300 mW), and ultraviolet (355 nm laser at 100 mW) ones were used for flow karyotyping. The setup and performance were optimized using standard 8-peaks Rainbow beads (Sphero^{TM} Rainbow Calibration Particles 3.0–3.4 μm, BD Biosciences, ref. 559123), 1-peak UV beads for UV laser alignment (Alignflow^{TM} Flow Cytometry Alignment 2.7 μm, Molecular Probes, ref. A16502), and 1-peak 457 nm for deep-blue laser alignment (Fluoresbrite^{TM} Plain YG Microspheres 1.0 μm, Polysciences, Inc. ref. 17154) were, respectively, used for 488-blue, 355-UV, and 457-deep-blue optimal laser alignment and instrument fine tuning to obtain the highest resolution of chromosome detection and sorting.

The threshold for chromosome sorting was set triggering in chromomycin-A3 fluorescence on 457 nm laser as primary excitation line and set at approximately 1800 a.u. Then, chromomycin-A3 fluorescence was used as primary fluorescence reference through a light line of 500 LP filter and collected by a 550/50 nm bandpass filter. Hoechst was excited with the UV laser and its fluorescence was collected through a light line of 400 LP filter and by 460/50 BP. All parameters were collected in lineal mode and analyzed with the BD FACS^{TM} Software (v. 1.0.0.0.650, Becton Dickinson, San Jose, CA).

We chose a 100 μm nozzle because we found it to have the best piezoelectric frequency/electronic-noise ratio. The piezoelectric frequency was adjusted at 38.7 KHz. The sample flow rate for chromosome sorting was adjusted at up to 6000 events s^{-1}. The gating strategy for chromosome sorting was simple because only a bi-parametrical dot-plot Hoechst versus chromomycin-A3 fluorescence was used (see Fig. 1a).

Purification and concentration of flow-sorted Y chromosomes. For each of the two rounds of purification, the fractions corresponding to approximately 4.5 M Y chromosomes (~ 500 ng of DNA per aliquot) were divided into 1 ml aliquots with an estimated chromosome count of 400,000, corresponding to a DNA concentration of approximately 0.04 ng μl^{-1}. The approximate total volume per round of purification was around 22.5 ml. Each tube containing the flow-sorted DNA was treated overnight with 10 μl of proteinase K (20 mg ml^{-1}) at 50 °C. After treatment, the buffer was exchanged, and proteinase K, as well as chromomycin-A3 and Hoechst 33,258 removed by dialysis against 1 liter of TE buffer using a Pur-A-Lyzer™ Maxi Dialysis column with a molecular weight cut-off of 50 kDa (Sigma-Aldrich). Dialysis was carried out for 48 h exchanging the buffer every 10–16 h. To reduce the volume after buffer exchange, DNA was transferred into 1.5 ml tubes and concentrated by evaporation in a miVac DNA concentrator (Barnstead GeneVac, Ipswich, UK) up to a volume of approximately 5–10 µl. A final purification step was performed by pooling the concentrated DNA into two tubes and subjecting it to a solid-phase reversible immobilisation (SPRI) bead purification with a 2X ratio (SPRI beads/sample). DNA was eluted in 9 µl of low TE buffer and pooled into one tube. Concentrations were determined by absorbance at 260 nm with a NanoDrop 2000 (Thermo Scientific) and by fluorometric assay with the Qubit 2.0 using the Qubit dsDNA HS kit (Invitrogen) (see Supplementary Table 1).

Sequencing the flow-sorted chromosomes. The purified DNA was prepared for sequencing following the protocol in the Rapid Sequencing kit SQK-RAD002 (ONT, Oxford, UK). Briefly, approximately 200 ng of purified DNA was tagmented for 1 min at 75 °C with the Fragmentation Mix (ONT, Oxford, UK). The Rapid Adapters (ONT, Oxford, UK) were added along with Blunt/TA Ligase Master Mix (NEB, Beverly, MA) and incubated for 30 min at room temperature. The resulting library was combined with Running Buffer with Fuel (ONT, Oxford, UK) and Library Loading Beads (ONT, Oxford, UK) and loaded onto a primed R9.4 SpotOn Flow cell (FLO-MIN106). Sequencing and initial base calling was performed with a MinION Mk1B MinKNOW v1.7.10 software package running for 48 h. Estimates for DNA quantification were based on chromosomal counts with corresponding quantification values from Gribble et al.[31]. The uncertainties in quantification with Qubit 2.0 or NanoDrop are presumed to be due to residual intercalating dyes present within the sample, which interfere with the quantification platforms detection systems, with competition of additional intercalants leading to underestimation on the Qubit 2.0, and the additional presence of aromatic groups leading to overestimation on the NanoDrop.

A total estimated amount of 100 ng of Y chromosome was fragmented on a Covaris ultrasonicator with settings targeting fragments of 450 bp. The library was prepared using NEBNext Ultra II DNA Library Prep Kit (New England BioLabs) following the manufacturer's instructions, including four cycles of PCR amplification. Agilent BioAnalyzer High-Sensitivity DNA Kit was used to determine the size distribution and molarity. The library was sequenced on an Illumina MiSeq using the v3 kit and 600 cycles resulting in 300-bp paired-end reads.

Assembly, error correction, and polishing. The initial base-calls (MinKNOW 1.7.10 using Albacore 1.1) from the Nanopore data were assembled with Canu (v 1.6)[18] without previous read separation of reads deriving from different chromosomes and assuming a chromosome size of 52 Mb. The following parameters were used:

```
  canu -p HG02982 -d HG02982_canu genomeSize=52
m overlapper=mhap utgReAlign=true -nanopore-
raw raw_data/HG02982/all.joint.fastq
```

The 2.3 Gb of input data resulted in 25X of error corrected reads for assembly, assuming a chromosome size of 52 Mb. The data assembled into 35 contigs, which where self-corrected using the Nanopore input reads. To this end, we re-performed base calling from the fast5 files using Albacore (v 2.1, available from the nanopore user community) to be used for variant calling with Nanopolish (v. 0.8.4, https://github.com/jts/nanopolish, 11 December 2017).

```
  read_fast5_basecaller.py -f FLO-MIN106 -k
SQK-RAD002 -i input_folder -s outout_folder -t 8
-o fastq,fast5 -q 10000000 -n 100000 --disable_
pings
```

We indexed the reads to be used with Nanopolish:

```
  nanopolish index -f fast5.fofn reads.joint.
fastq
```

The reads were mapped onto the raw assembly using bwa mem (v. 0.7.120)[32] with the additional flag -x ont2d and the mappings merged and sorted with samtools (v. 1.5):

```
  bwa mem -x ont2d HG02982_canu.uncorrected.
fasta reads.joint.fastq | samtools sort -o
reads.joint.mappings.bam -T tmp -
```

The mappings were fed to Nanopolish and corrected in chunks of 50 kb using the helper script "nanopolish_makerange.py" included in the Nanopolish package. Variants were called using "nanopolish variants –consensus" with the optional flag "--min-candidate-frequency 0.1".

```
  nanopolish_makerange.py HG02982_canu.
uncorrected.fasta | xargs -i echo nanopolish
variants --consensus selfcorrected.{}.fa -w {}
-r reads.joint.fastq -b reads.joint.mappings.
bam -g HG02982_canu.uncorrected.fasta -t 4
--min-candidate-frequency 0.1 | sh
```

By this means, we corrected 127,801 positions in the initial assembly. The self-corrected assembly was further polished with the Illumina library. To this end, we trimmed the Illumina reads to get rid of any adapters in the sequences using trimgalore (v 3.7, https://github.com/FelixKrueger/TrimGalore).

```
  trim_galore --fastqc --paired --retain_
unpaired gzip pair1.fastq pair2.fastq
```

The trimmed reads were mapped with BWA mem (v.0.7.12)[32] in paired-end mode and the mappings converted to a sorted bam files using samtools sort. PCR duplicates were removed with Picardtools (v. 2.8.2, https://broadinstitute.github.io/picard/).

```
  bwa mem HG02982_canu.selfcorrected.fasta
reads.p1.fastq reads.p2.fastq | samtools sort
-o reads.paired.mappings.bam -T tmp -;
  java -jar picard.jar MarkDuplicates I=reads.
paired.mappings.bam O=reads.paired.mappings.
markdup.bam M=reads.paired.mappings.markdup.
bam
```

Polishing was performed with Pilon (v 1.22)[19], resulting in 132,336 residual errors being corrected.

```
  java -Xmx96G -jar pilon-1.22.jar --threads 12
--genome HG02982_canu.selfcorrected.fasta
--frags reads.paired.mappings.markdup.bam
--output HG02982_canu.selfcorrected.pileon
--outdir pilon_corrections --changes --vcf
--tracks --fix all
```

To run racon (v 1.3.1, see Supplementary Table 4), we mapped the Illumina reads onto the polished reference with bwa, sorted the alignments with samtools and removed duplicates as described above. The resulting alignments were provided to racon as an input:

```
  racon -u -t 12 reads.fastq mappings.sam
HG02982_chrY_v1.fasta
```

Variant calls. For variant calls, the Illumina data were mapped onto the GRCh38 or the HG02982 assembly, respectively, and processed the same way as detailed above. Variants were called using GATKs Haplotype Caller with the following optional flags: "--genotyping-mode DISCOVERY --sample-ploidy 1".

```
  java -jar gatk-package-4.0.0.0-local.jar
HaplotypeCaller -R reference.fa -I mappings.
bam --genotyping-mode DISCOVERY -O variants.
vcf
```

Repeat annotations. Repeat annotations were performed using RepeatMasker (v. 4.0.7) with rmblastn v. 2.6.0+ as the engine. To be comparable, the annotations for both the HG02982, as well as the GRCh38 assembly were performed the same way. We used the RepBase-20170127 as the repeatmasker database, and Homo sapiens as the query species. Divergence of the repeat annotations to their consensus was calculated using the "calcDivergenceFromAlign.pl" utility included in the RepeatMasker package.

```
  RepeatMasker -e ncbi -pa 12 -s -species human
-no_is -noisy -dir ./outDir -a -gff -u reference.
fa
```

Whole-genome alignments. Whole-genome alignments to GRCh38 were produced using last (v. 914) with the following parameters as suggested by the developer for highly similar genomes for indexing and alignments:

```
  lastdb -uNEAR -R01 index reference.fa
  lastal -e25 -v -q3 -j4 index query.fa >
mappings.maf
```

Single best placements of query sequences were retained using the "last-split" script included in the last alignment package. Alignments were filtered for a maximum mismap probability of 10e–5. The alignments were converted to psl format for further processing.

Comparison with WGS PacBio data. The PacBio data from the Ashkenazim Son (Coriel ID NA24385) produced by the genome in a bottle consortium was also assembled using Canu (v. 1.6) using default assembly parameters and assuming a genome size of 3.2 Gb:

```
  canu -p NA24385 -d NA24385_canu genomeSize=3.
2g -pacbio-raw data/fastq/*fastq.gz
gridOptionsExecutive='--mem-per-cpu=16g
--cpus-per-task=2'
```

After genome assembly, we performed a whole-genome alignment to GRChg38 and retained single best placements as mentioned above. To identify contigs belonging to the Y chromosome, we performed the following filtering steps: for contigs, which have local best placements on a chromosome different than the Y, we filtered out those whose proportion of mapped bases is higher on a sequence from the reference assembly different from the Y chromosome. Additionally, we filtered out any alignments with a mismap probability higher than 10e–5. By this means, we retained 184 contigs mapping 15,308,468 base pairs on the Y chromosome (see Supplementary Data 6)

SV calls. SVs were called with assemblytics[25]. To this end, we produced whole-genome alignments using nucmer from the Mummer package (v. 3.22)[33]. The resulting delta file was passed to assemblytics, with the required unique anchor length set to 10000 bp.

```
  nucmer -maxmatch -l 100 -c 500 GRCh38.chrY.fa
HG02982_chrY_v1.fasta -prefix HG02982_vs_HG38
```

```
  Assemblytics HG02982_vs_HG38_.delta
HG02982_chrY_v1.vs.hg38_10kanchor.50kmax
10000 bin/Assemblytics/
```

Read depth duplication detection. We estimated absolute copy number with a depth of coverage approach using the Illumina data[22]. We masked all common repeats as identified by RepeatMasker (see above) and tandem repeat finder. We created non-overlapping 36-mers of the raw reads, which were mapped onto the assembly using GEM (v 2)[34] allowing for a divergence of up to 5%. The read depth was calculated in non-overlapping windows of 1 kb of non-repetitive sequence. After correcting for GC content using mrCanavar (v. 0.51), we normalized by the mean read depth. To assign a copy number to each gene, we calculated the median copy number of all windows intersecting a gene. For the hg38 Y chromosome, a set of custom single-copy regions needed to be provided to the CN caller as calibration. These regions were inferred by subtracting the reference WGAC (whole-genome assembly comparison, UCSC track genomic superdups) segmental duplication track from the whole Y chromosome and keeping only stretches of single-copy sequence longer than 2 kb.

Gene annotation. The annotation of the HG02982 assembly was performed by trying to assign the genes present in the Y-chromosome annotation of GRCh38 gencode version 27. For this purpose, we downloaded the gff3, the transcript sequences and the protein sequences that corresponded to the Y-chromosome annotation and performed transcript and protein mappings with GMAP (v. 20170317[35]) and exonerate (v. 2.2.0[36]), respectively. Additionally, a numeric index was assigned to each gene in the HG38 Y chromosome according to the order in the chromosome. Next, we combined all the data (transcript mappings, protein mappings and gene synteny) with an in-house script (available at https://doi.org/10.6084/m9.figshare.7359065.v1) to locate each gene in our assembly and assign parts of the assembly to their corresponding region in the Y chromosome of GRCh38. After following the strategy mentioned above for all the genes, we took a closer look to the protein-coding genes, by manually checking some of the mappings in order to determine possible errors in the sequence caused by the Nanopore reads that could introduce frameshifts or internal stop codons in the aminoacidic sequence.

Illumina WGBS sequencing and methylation calls. Two micrograms of genomic DNA from a lymphoblastoid cell line (HG02982) were spiked with unmethylated bacteriophage λ DNA (5 ng of λ DNA per microgram of genomic DNA; Promega) and with methylated T7 phage DNA (5 ng of T7 DNA per microgram of genomic DNA). The DNA was sheared to 50–500 bp in size using Covaris LE220 ultrasonicator, and fragments of 150–300 bp were size-selected using AMPure XP beads (Agencourt Bioscience). The libraries were constructed using the KAPA Library Preparation Kit with no PCR Library Amplification/Illumina series (Roche-Kapa Biosystems) together with the NEXTFLEX® Bisulfite-Seq Barcodes (Perkin Elmer). After adaptor ligation, the DNA was treated with sodium bisulfite using the EpiTect Bisulfite kit (Qiagen) following the manufacturer's instructions. Enrichment for adaptor-ligated DNA was carried out through seven PCR cycles using KAPA HiFi HotStart Uracil+ReadyMix PCR 2x Kit (Roche-Kapa Biosystems). Library quality was monitored using the Agilent 2100 Bioanalyzer DNA 7500 assay, and the library concentration was estimated using quantitative PCR using the KAPA Library Quantification Kit for Illumina® Platforms, v1.14 (Roche-Kapa Biosystems).

Paired-end DNA sequencing (2×101 bp) of the converted libraries was performed using the HiSeq 2500 (Illumina) following the manufacturer’s protocol with HiSeq Control Software (HCS) 2.2.68. Primary data analysis, image analysis, base calling, and quality scoring of the run, was processed using the manufacturer’s software Real Time Analysis (RTA 1.18.66.3) and followed by generation of FASTQ sequence files by CASAVA.

We used the gemBS pipeline[37] using the default parameters to perform the analysis. The reference genome used for the alignment was GRCh38. Methylated and unmethylated cytosine conversion rates were determined from spiked-in bacteriophage DNA (fully methylated phage T7 and unmethylated phage lambda). The under and over conversion rates for the sample were <1 and ~ 5%, respectively. Only uniquely mapping reads were retained for downstream analysis. The comparison with the Nanopore calls was performed for all canonical CpG sites on the Y chromosome where there was sequencing data available from both experiments. The comparison took account of the variable precision of the methylation estimates due to variation in sequencing coverage between sites so that low-coverage sites did not affect the comparison.

Nanopore methylation calls. The methylation status was called using Nanopolish[10] as suggested by the developers. To this end, we aligned the Nanopore reads to GRCh38 with minimap2[38] and sorted with samtools (v 1.5). The calls were performed in 200 kb windows.

```
  minimap2 -a -x map-ont chrY.fa joint_reads.
fastq | samtools sort -T tmp -o joint_reads.
mappings.bam
  samtools index joint_reads.mappings.bam
  nanopolish call-methylation -v --progress -t
8 -r joint_reads.fastq -b joint_reads.
mappings.bam -g chrY.fa -w"chrY:$start-$stop"
> methylation_calls.tsv
```

Finally, we calculated the methylation frequency and log-likelihood ratios of methylation at each position:

```
  calculate_methylation_frequency.py -i
methylation_calls.tsv
```

We filtered out any position with <10 reads in either the WGBS or the Nanopore data. Additionally, any position with a log-likelihood ratio of <2.5 in the Nanopore data were also excluded.

Code availability. The custom script used for the gene annotation has been deposited at Figshare at https://doi.org/10.6084/m9.figshare.7359065.v1.

Data availability

All raw sequencing data for this study have been deposited at the European Nucleotide Archive (ENA) under the study accession PRJEB28143. The assembly is deposited at the ENA under the accession ULGL01000000. The WGS assembly for the NA24385 individual, the repeat masker tracks for the GRCh38 chrY and HG02982 assemblies, and the methylation calls from the Illumina WGBS and the Nanopore data are deposited at Figshare at https://doi.org/10.6084/m9.figshare.7358480.v1. The source data underlying Figs. 1a-d and 2a-c are provided as a Source Data File. A Reporting Summary for this Article is available as a Supplementary Information file. All other relevant data are available upon request.

Received: 29 June 2018 Accepted: 2 December 2018
Published online: 02 January 2019

References

1. Charlesworth, B. & Charlesworth, D. The degeneration of Y chromosomes. *Philos. Trans. R. Soc. B Biol. Sci.* **355**, 1563–1572 (2000).
2. Hughes, J. F. & Page, D. C. The biology and evolution of mammalian Y chromosomes. *Annu. Rev. Genet.* **49**, 507–527 (2015).
3. Tomaszkiewicz, M., Medvedev, P. & Makova, K. D. Y and W chromosome assemblies: approaches and discoveries. *Trends Genet.* **33**, 266–282 (2017).
4. Skaletsky, H. et al. The male-specific region of the human Y chromosome is a mosic of discrete sequence classes. *Nature* **423**, 825–837 (2003).
5. Hughes, J. F. et al. Chimpanzee and human y chromosomes are remarkably divergent in structure and gene content. *Nature* **463**, 536–539 (2010).
6. Hughes, J. F. et al. Strict evolutionary conservation followed rapid gene loss on human and rhesus y chromosomes. *Nature* **483**, 82–87 (2012).
7. Soh, Y. Q. S. et al. Sequencing the mouse y chromosome reveals convergent gene acquisition and amplification on both sex chromosomes. *Cell* **159**, 800–813 (2014).
8. Tomaszkiewicz, M. et al. A time- and cost-effective strategy to sequence mammalian Y chromosomes: an application to the de novo assembly of gorilla Y. *Genome Res.* **26**, 530–540 (2016).
9. Zhang, K. et al. Sequencing genomes from single cells by polymerase cloning. *Nat. Biotechnol.* **24**, 680–686 (2006).
10. Simpson, J. T. et al. Detecting DNA cytosine methylation using nanopore sequencing. *Nat. Methods* **14**, 407–410 (2017).
11. Berlin, K. et al. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat. Biotechnol.* **33**, 623–630 (2015).
12. Chaisson, M. J. P. et al. Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**, 608–611 (2015).
13. Kuderna, L. F. K. et al. A 3-way hybrid approach to generate a new high-quality chimpanzee reference genome (Pan tro 3.0). *Gigascience* **6**, gix098 (2017).
14. Bickhart, D. M. et al. Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nat. Genet.* **49**, 643–650 (2017).
15. Jain, M. et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* **36**, 338–345 (2018).
16. Jain, M. et al. Linear assembly of a human centromere on the Y chromosome. *Nat. Biotechnol.* **36**, 321–323 (2018).
17. Auton, A. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
18. Koren, S. et al. Canu : scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2016).
19. Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* https://doi.org/10.1371/journal.pone.0112963 (2014).
20. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).

21. Page, D. C., Harper, M. E., Love, J. & Botstein, D. Occurrence of a transposition from the X-chromosome long arm to the Y-chromosome short arm during human evolution. *Nature* **311**, 119–123 (1984).
22. Alkan, C. et al. Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat. Genet.* **41**, 1061–1067 (2009).
23. Zook, J. M.et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci. Data*. **3**, 160025 (2016).
24. Hughes, J. F. & Rozen, S. Genomics and genetics of human and primate Y chromosomes. *Annu. Rev. Genom. Hum. Genet.* **13**, 83–108 (2012).
25. Nattestad, M. & Schatz, M. C. Assemblytics: a web analytics tool for the detection of variants from an assembly. *Bioinformatics* **32**, 3021–3023 (2016).
26. Poznik, G. D. et al. Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences. *Nat. Genet.* **48**, 593–599 (2016).
27. Lukaszewski, A. J. et al. A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. *Science* **345**, 6194 (2014).
28. Zimin, A. V. et al. The first near-complete assembly of the hexaploid bread wheat genome, *Triticum aestivum*. *Gigascience* **6**, gix097 (2017).
29. Neale, D. B.et al. Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. *Genome Biol.* **15**, R59 (2014).
30. Nowoshilow, S. et al. The axolotl genome and the evolution of key tissue formation regulators. *Nature* **554**, 50–55 (2018).
31. Gribble, S. M., Ng, B. L., Prigmore, E., Fitzgerald, T. & Carter, N. P. Array painting: a protocol for the rapid analysis of aberrant chromosomes using DNA microarrays. *Nat. Protoc.* **4**, 1722–1736 (2009).
32. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
33. Kurtz, S. et al. Versatile and open software for comparing large genomes. *Genome. Biol.* **5**, R12 (2004).
34. Marco-Sola, S., Sammeth, M., Guigó, R. & Ribeca, P. The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat. Methods* **9**, 1185–1188 (2012).
35. Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005).
36. Slater, G. S. C. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinforma.* **6**, 31 (2005).
37. Merkel, A. et al. gemBS: high throughput processing for DNA methylation data from bisulfite sequencing. *Bioinformatics* https://doi.org/10.1093/bioinformatics/bty690 (2018).
38. Li, H. Genome analysis Minimap2 : pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

Acknowledgements

This study was supported by the Spanish Ministry of Economy and Competitiveness with Proyectos de I+D “Excelencia” y Proyectos de I+D+I “Retos Investigación” BFU2014-55090-P awarded to T.M.-B. and O.F., Centro de Excelencia Severo Ochoa 2013–2017 and Centro de Excelencia Maria de Maeztu 2016–2019. We acknowledge the support from the CERCA Programme of the Generalitat de Catalunya, institutional support from the Spanish Ministry of Economy, Industry and Competitiveness (MEIC) through the Instituto de Salud Carlos III, from the Generalitat de Catalunya through the Departament de Salut and Departament d’Empresa i Coneixement, and co-financing by the Spanish Ministry of Economy, Industry and Competitiveness (MEIC) with funds from the European Regional Development Fund (ERDF) corresponding to the 2014–2020 Smart Growth Operating Program. L.F.K.K. is supported by an FPI fellowship associated with BFU2014-55090-P (MINECO/FEDER, UE). M.K. is supported by a Deutsche Forschungsgemeinschaft (DFG) fellowship (KU 3467/1-1). T.M.-B. is supported by BFU2017-86471-P (MINECO/FEDER, UE), U01 MH106874 grant, Howard Hughes International Early Career, Obra Social “La Caixa” and Secretaria d’Universitats i Recerca del Departament d’Economia i Coneixement de la Generalitat de Catalunya. D.J. is supported by a Juan de la Cierva fellowship (FJCI-2016-29558) from MICINN.

Author contributions

T.M.-B. conceived the study, L.F.K.K., J.G.-G., A.S.A., M.K., S.H., D.J., and T.A. performed computational analysis, E.L. developed and performed the purification protocol, E.J. and O.F. cultured cells and performed the flow cytometry, R.A.A., M.A.-E., M.G., I. G., and M.H.S. prepared materials and/or performed the sequencing. L.F.K.K., E.L., and T.M.-B. wrote the manuscript with input from all authors. All authors approved the manuscript.

Additional information

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-018-07885-5.

Competing interests: The authors declare no competing interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Journal peer review information: *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

4 DISCUSSION

The constant development of sequencing technologies and ongoing fall of their prices have enabled their application by “ordinary” research groups. The incredible pace with which methods for reference assemblies are being developed also implies that approaches that might have been state of the art when I started my Ph.D. are dated by now, as the appropriate choice of methods seems to be quickly changing for a given budget.

We have produced a substantially improved reference assembly for the chimpanzee, a key species also for human genetics. We have done so by incorporating diverse sequencing strategies, namely high coverage short reads, low coverage long reads and chromosomal conformation maps, among others. By doing so, we’ve substantially increased continuity, repeat resolution and improved the gene annotation. We have furthermore improved the representation of segmental duplications compared to previous iterations. We have shown the combinatorial power of short and long reads, with specific benefits of each data type. Our assembly has been rendered obsolete not long after its release by the publication of pure long read assemblies for all great apes, which offer a greatly improved overall sequence representation (Kronenberg et al. 2018). Although long read assemblies offer far superior quality, they come at a cost that might be prohibitive for some groups. In this context, our 3-way hybrid offers a potential alternative at a much lower price point. However, methods like 10X-genomics have enabled relatively high continuity for short read assemblies for a comparatively low price tag, and offer additional phasing information (Marks et al. 2019). Additionally, the per base cost of Nanopore sequence production is approaching that of Illumina, making long read assemblies more accessible.

Despite the broadened accessibility of genome assemblies, there is still a lack of primates for which these resources are available. Its increase, specifically for high quality assemblies, will allow plenty new insights into their phylogenetic relationship, chromosomal evolution, lineage specific adaptations and the overall role of structural variation in primate evolution. Currently, large scale consortia such as Genome 10K and the Earth BioGenome Project are underway to create genomic resources for a vast number of species which will allow to answer these questions (Koepfli, Paten, and O'Brien 2015; Lewin et al. 2018). However, there are still further limitations to the current way assemblies are approached beyond having a reference for a given species. In my opinion, among the biggest current shortcomings is their haploid, linear representation. Traditionally, assemblies have been presented as haploid compressions of (in the case of human and other primates) diploid genomes. This representation usually chooses one of the two alleles at random for each position, which might lead to biologically meaningless haplotypes in several cases. This issue is further aggravated by the fact, that the human reference genome is a mosaic of several different individuals. In the case of single nucleotide variants, their detection might not be strongly influenced by the haplotype represented in the reference. However, there are plenty regions of rampant structural diversity in the human genome in which the inclusion of one haplotype will prevent the detection of others

(Chaisson, Wilson, and Eichler 2015). The most recent iteration of the human reference assembly includes several so-called alternative haplotypes in regions of complex structural diversity, such as the MHC, KIR or the 17q21.31 region. These alternative haplotypes, however, exist free from genomic context as parallel, linear representations to the reference haplotype. There is furthermore a glaring lack of tools to deal with them, which in practice has led to low usage rates because of a lack of adaption by the community. Genomic research is limiting itself in this context, as assessable genomic regions heavily rely on accurate reference representation. As long as the genomics community is stuck to linear haploid reference representations of a single – or like in the case of the human genome "pseudosingle" – individual. Given the quality of assemblies that are possible today, the clear way forward is the assembly of a rich panel of several human individuals with diverse ancestry to include and represent human diversity and variation on all levels in a reference graph, rather than a single reference assembly.

Flow sorting chromosomes for is not a new idea, and the technique has been available for several decades now (Gray et al. 1975). Subsequently, it has been applied to several genomics projects. For example, the initial human Y chromosome assembly used BAC libraries that were derived from DNA enriched in Y chromosomes by flow sorting. The wheat genome has an estimated size of 17 Gb, and its initial assembly was also derived from flow sorted material (The International Wheat Genome Sequencing Consortium (IWGSC) 2014). Lastly, the Gorilla Y chromosome has also recently been assembled based on flow sorted DNA. Of the three examples, the

Gorilla Y chromosome project is the only one that also created a long-read sequencing dataset. Crucially though, our work is the first to sequence native DNA. All the aforementioned works have either used their material to create clone libraries, or amplified it post sorting. Excluding amplification avoids biases such as uneven coverage, chimeras, or restricted read length due to restricted fragment size from the amplified material. It is therefore arguably the first work in which we can fully take advantage of long read sequencing and isolation by flow sorting. The read lengths we are able to achieve are certainly lower than what is currently possible with Nanopore, especially in the context of ultra-long reads (Jain, Koren, et al. 2018). This might be due to many handling steps in the preparation. As the read length only depends on the integrity of the DNA, forcing it through a pressurized system such as a flow cytometer might not be beneficial. They have, however, proven greatly useful to generate a high-quality assembly and with a read N50 of around 18 Kb are clearly above what the restricted fragment lengths of amplified DNA can achieve. Additionally, the Nanopore signal data can not only be used to detect canonical bases, but rather all kinds of covalent modifications od the DNA as long as training datasets exists. Currently models for 5-methylcytosine (5mC) and N6-methyladenosine (6mA) have been published (Simpson et al. 2017; Liu et al. 2019).

We know surprisingly (maybe even embarrassingly) little about structural evolution of Y chromosomes in the primate lineage. This is clearly owed to technical limitations in assembling them, and our efforts have provided a solution to mitigate this situation. With the currently ongoing advancements in genome assembly, particularly surrounding the developments of long reads, I could foresee the physical isolation becoming obsolete in the future, however, at the state of the art we do show that there is vastly better sequences class representation in our approach than in an assembly based on whole-genome shotgun data. It is also worth mentioning, that these insights were only possible because we used a human Y chromosome to benchmark this method, allowing a comparison to a reference assembly of gold-standard quality. It is quite possible to imagine that whole-genome shotgun assemblies will surpass the quality of our isolation assemblies, and the price tag for sequencing the whole genome becomes so low that the upfront cost for running the flow cytometer does not make economic sense. However, there are plenty of animals and plants with genome sizes well beyond those typically observed in primates, for which no genomic resources are available. For example, the genome of the Axolotl was recently assembled (Nowoshilow et al. 2018). It has an estimated size of 32 Gb, or 10 times that of human. Despite being a pure PacBio assembly, the very high repeat content of the genome led the assembly to have a contig N50 of only 216 Kb, while human assemblies based on the same data type typically exhibit contig N50 in the double digit megabase range. Plants and animal genomes of up to 130 Gb have been reported, and their assembly will certainly constitute a veritable computational challenge (Pellicer, Fay, and Leitch 2010; Metcalfe et al. 2012). For these cases, isolation by flow sorting with subsequent long read sequencing is likely to yield substantial benefits for genome assembly. However, genome assembly is probably going to remain a quickly moving target for some time, so any kind of prediction should be taken with precaution.

The ampliconic regions remain a major challenge for Y chromosome analysis, as do many other, smaller, segmental duplications. Structural variation of several ampliconic genes is associated with some cases of diseases, mainly in the context of male fertility (Hughes and Page 2015). There is also widespread copy number variation of ampliconic genes that stratify by different human populations, and the incidence of these events is high (Lucotte et al. 2018; Teitz et al. 2018). The functional impact of these variants is not yet fully clear, with evidence showing that for a given gene, copy number variation does not heavily influence gene expressions, but genes with high levels of copy number variation generally show higher overall expression levels (Lucotte et al. 2018; Vegesna et al. 2019). Additionally, there might be selection for the reference ampliconic haplotype (Teitz et al. 2018). Crucially though, all studies mentioned above are based on indirect observations of the variants, based on the reference haplotype. They use a read depth approach to infer the copy number of a given regions but have no way of deducing the structure of the chromosome. Sequencing reference panels of individuals from different populations to create structural variant calls with long reads has proved invaluable by greatly increasing their detection rate (Audano et al. 2019). Additionally, there is ongoing developments to characterize large segmental duplications from these datasets (Vollger et al. 2019). Ultimately, the direct comparison of sequence resolved assemblies will be the gold-standard method to answer these questions.

5 FULL LIST OF COMMUNICATIONS

1. Dobrynin P, Liu S, Tamazian G, et al. Genomic legacy of the African cheetah, Acinonyx jubatus. Genome Biology. 2015;16(1):277. doi:10.1186/s13059-015-0837-4

2. Manuel M de, Kuhlwilm M, Frandsen P, et al. Chimpanzee genomic diversity reveals ancient admixture with bonobos. Science. 2016;354(6311):477-481. doi:10.1126/science.aag2602

3. Librado P, Gamba C, Gaunitz C, et al. Ancient genomic changes associated with domestication of the horse. Science. 2017;356(6336):442-445. doi:10.1126/science.aam5298

4. Gopalakrishnan S, Samaniego Castruita JA, Sinding M-HS, et al. The wolf reference genome sequence (Canis lupus lupus) and its implications for Canis spp. population genomics. BMC Genomics. 2017;18(1):495. doi:10.1186/s12864-017-3883-3

5. Mak SST, Gopalakrishnan S, Carøe C, et al. Comparative performance of the BGISEQ-500 vs Illumina HiSeq2500 sequencing platforms for palaeogenomic sequencing. Gigascience. 2017;6(8). doi:10.1093/gigascience/gix049

6. Kuderna LFK, Tomlinson C, Hillier LW, et al. A 3-way hybrid approach to generate a new high-quality chimpanzee reference

genome (Pan_tro_3.0). Gigascience. 2017;6(11). doi:10.1093/gigascience/gix098

7. Warren WC, Kuderna L, Alexander A, et al. The Novel Evolution of the Sperm Whale Genome. Genome Biology and Evolution. 2017;9(12):3260-3264. doi:10.1093/gbe/evx187

8. Serres-Armero A, Povolotskaya IS, Quilez J, et al. Similar genomic proportions of copy number variation within gray wolves and modern dog breeds inferred from whole genome sequencing. BMC Genomics. 2017;18(1):977. doi:10.1186/s12864-017-4318-x

9. Warren WC, García-Pérez R, Xu S, et al. Clonal polymorphism and high heterozygosity in the celibate genome of the Amazon molly. Nature Ecology & Evolution. 2018;2(4):669. doi:10.1038/s41559-018-0473-y

10. Gopalakrishnan S, Sinding M-HS, Ramos-Madrigal J, et al. Interspecific Gene Flow Shaped the Evolution of the Genus Canis. Current Biology. 2018;28(21):3441-3449.e5. doi:10.1016/j.cub.2018.08.041

11. Quesada V, Freitas-Rodríguez S, Miller J, et al. Giant tortoise genomes provide insights into longevity and age-related disease. Nature Ecology & Evolution. 2019;3(1):87. doi:10.1038/s41559-018-0733-x

12. Kuderna LFK, Lizano E, Julià E, et al. Selective single molecule sequencing and assembly of a human Y chromosome of African origin. Nature Communications. 2019;10(1):4. doi:10.1038/s41467-018-07885-5

13. Lorente-Galdos B, Lao O, Serra-Vidal G, et al. Whole-genome sequence analysis of a Pan African set of samples reveals archaic gene flow from an extinct basal population of modern humans into sub-Saharan populations. Genome Biology. 2019;20(1):77. doi:10.1186/s13059-019-1684-5

14. Tollis M, Robbins J, Webb AE, et al. Return to the Sea, Get Huge, Beat Cancer: An Analysis of Cetacean Genomes Including an Assembly for the Humpback Whale (Megaptera novaeangliae). Mol Biol Evol. 2019;36(8):1746-1763. doi:10.1093/molbev/msz099

15. Fages A, Hanghøj K, Khan N, et al. Tracking Five Millennia of Horse Management with Extensive Ancient Genome Time Series. Cell. 2019;177(6):1419-1435.e31. doi:10.1016/j.cell.2019.03.049

6 BIBLIOGRAPHY

Adams, Mark D., Susan E. Celniker, Robert A. Holt, Cheryl A. Evans, Jeannine D. Gocayne, Peter G. Amanatides, Steven E. Scherer, et al. 2000. "The Genome Sequence of Drosophila Melanogaster." *Science* 287 (5461): 2185–95. https://doi.org/10.1126/science.287.5461.2185.

Alkan, Can, Saba Sajjadian, and Evan E. Eichler. 2011. "Limitations of Next-Generation Genome Sequence Assembly." *Nature Methods* 8 (1): 61–65. https://doi.org/10.1038/nmeth.1527.

Audano, Peter A., Arvis Sulovari, Tina A. Graves-Lindsay, Stuart Cantsilieris, Melanie Sorensen, AnneMarie E. Welch, Max L. Dougherty, et al. 2019. "Characterizing the Major Structural Variant Alleles of the Human Genome." *Cell* 176 (3): 663-675.e19. https://doi.org/10.1016/j.cell.2018.12.019.

Bachtrog, Doris. 2013. "Y-Chromosome Evolution: Emerging Insights into Processes of Y-Chromosome Degeneration." *Nature Reviews Genetics* 14 (2): 113–24. https://doi.org/10.1038/nrg3366.

Bastide, Melissa de la, and W. Richard McCombie. 2007. "Assembling Genomic DNA Sequences with PHRAP." *Current Protocols in Bioinformatics* 17 (1): 11.4.1-11.4.15. https://doi.org/10.1002/0471250953.bi1104s17.

Bellott, Daniel W., Jennifer F. Hughes, Helen Skaletsky, Laura G. Brown, Tatyana Pyntikova, Ting-Jan Cho, Natalia Koutseva, et al. 2014. "Mammalian Y Chromosomes Retain Widely Expressed Dosage-Sensitive Regulators." *Nature* 508 (7497): 494–99. https://doi.org/10.1038/nature13206.

Benjamini, Yuval, and Terence P. Speed. 2012. "Summarizing and Correcting the GC Content Bias in High-Throughput Sequencing." *Nucleic Acids Research* 40 (10): e72–e72. https://doi.org/10.1093/nar/gks001.

Bentley, David R., Shankar Balasubramanian, Harold P. Swerdlow, Geoffrey P. Smith, John Milton, Clive G. Brown, Kevin P. Hall, et al. 2008. "Accurate Whole Human Genome Sequencing Using Reversible Terminator Chemistry." *Nature* 456 (7218): 53–59. https://doi.org/10.1038/nature07517.

Berlin, Konstantin, Sergey Koren, Chen-Shan Chin, James P. Drake, Jane M. Landolin, and Adam M. Phillippy. 2015. "Assembling

Large Genomes with Single-Molecule Sequencing and Locality-Sensitive Hashing." *Nature Biotechnology* 33 (6): 623–30. https://doi.org/10.1038/nbt.3238.

Boyd, J. Lomax, Stephanie L. Skove, Jeremy P. Rouanet, Louis-Jan Pilaz, Tristan Bepler, Raluca Gordân, Gregory A. Wray, and Debra L. Silver. 2015. "Human-Chimpanzee Differences in a FZD8 Enhancer Alter Cell-Cycle Dynamics in the Developing Neocortex." *Current Biology: CB* 25 (6): 772–79. https://doi.org/10.1016/j.cub.2015.01.041.

Burton, Joshua N., Andrew Adey, Rupali P. Patwardhan, Ruolan Qiu, Jacob O. Kitzman, and Jay Shendure. 2013. "Chromosome-Scale Scaffolding of de Novo Genome Assemblies Based on Chromatin Interactions." *Nature Biotechnology* 31 (12): 1119–25. https://doi.org/10.1038/nbt.2727.

Butler, Jonathan, Iain MacCallum, Michael Kleber, Ilya A. Shlyakhter, Matthew K. Belmonte, Eric S. Lander, Chad Nusbaum, and David B. Jaffe. 2008. "ALLPATHS: De Novo Assembly of Whole-Genome Shotgun Microreads." *Genome Research* 18 (5): 810–20. https://doi.org/10.1101/gr.7337908.

Carbone, Lucia, R. Alan Harris, Sante Gnerre, Krishna R. Veeramah, Belen Lorente-Galdos, John Huddleston, Thomas J. Meyer, et al. 2014. "Gibbon Genome and the Fast Karyotype Evolution of Small Apes." *Nature* 513 (7517): 195–201. https://doi.org/10.1038/nature13679.

Castillo, Elio Rodrigo, Dardo Andrea Marti, and Claudio Juan Bidau. 2010. "Sex and Neo-Sex Chromosomes in Orthoptera: A Review*." *Journal of Orthoptera Research* 19 (2): 213–31. https://doi.org/10.1665/034.019.0207.

Chaisson, Mark J. P., John Huddleston, Megan Y. Dennis, Peter H. Sudmant, Maika Malig, Fereydoun Hormozdiari, Francesca Antonacci, et al. 2015. "Resolving the Complexity of the Human Genome Using Single-Molecule Sequencing." *Nature* 517 (7536): 608–11. https://doi.org/10.1038/nature13907.

Chaisson, Mark J. P., Richard K. Wilson, and Evan E. Eichler. 2015. "Genetic Variation and the *de Novo* Assembly of Human Genomes." *Nature Reviews Genetics* 16 (11): 627–40. https://doi.org/10.1038/nrg3933.

Charlesworth, B, and D Charlesworth. 2000. "The Degeneration of Y Chromosomes." *Philosophical Transactions of the Royal Society B: Biological Sciences* 355 (1403): 1563–72.

Charrier, Cécile, Kaumudi Joshi, Jaeda Coutinho-Budd, Ji-Eun Kim, Nelle Lambert, Jacqueline de Marchena, Wei-Lin Jin, et al. 2012. "Inhibition of SRGAP2 Function by Its Human-Specific Paralogs Induces Neoteny during Spine Maturation." *Cell* 149 (4): 923–35. https://doi.org/10.1016/j.cell.2012.03.034.

Chin, Chen-Shan, David H. Alexander, Patrick Marks, Aaron A. Klammer, James Drake, Cheryl Heiner, Alicia Clum, et al. 2013. "Nonhybrid, Finished Microbial Genome Assemblies from Long-Read SMRT Sequencing Data." *Nature Methods* 10 (6): 563–69. https://doi.org/10.1038/nmeth.2474.

Cortez, Diego, Ray Marin, Deborah Toledo-Flores, Laure Froidevaux, Angélica Liechti, Paul D. Waters, Frank Grützner, and Henrik Kaessmann. 2014. "Origins and Functional Evolution of Y Chromosomes across Mammals." *Nature* 508 (7497): 488–93. https://doi.org/10.1038/nature13151.

Deininger, Prescott. 2011. "Alu Elements: Know the SINEs." *Genome Biology* 12 (12): 236. https://doi.org/10.1186/gb-2011-12-12-236.

Dennis, Megan Y., Xander Nuttle, Peter H. Sudmant, Francesca Antonacci, Tina A. Graves, Mikhail Nefedov, Jill A. Rosenfeld, et al. 2012. "Evolution of Human-Specific Neural SRGAP2 Genes by Incomplete Segmental Duplication." *Cell* 149 (4): 912–22. https://doi.org/10.1016/j.cell.2012.03.033.

Denton, James F., Jose Lugo-Martinez, Abraham E. Tucker, Daniel R. Schrider, Wesley C. Warren, and Matthew W. Hahn. 2014. "Extensive Error in the Number of Genes Inferred from Draft Genome Assemblies." *PLOS Computational Biology* 10 (12): e1003998. https://doi.org/10.1371/journal.pcbi.1003998.

Doležel, Jaroslav, Jan Vrána, Jan Šafář, Jan Bartoš, Marie Kubaláková, and Hana Šimková. 2012. "Chromosomes in the Flow to Simplify Genome Analysis." *Functional & Integrative Genomics* 12 (3): 397–416. https://doi.org/10.1007/s10142-012-0293-0.

Earl, Dent, Keith Bradnam, John St John, Aaron Darling, Dawei Lin, Joseph Fass, Hung On Ken Yu, et al. 2011. "Assemblathon 1: A Competitive Assessment of de Novo Short Read Assembly Methods." *Genome Research* 21 (12): 2224–41. https://doi.org/10.1101/gr.126599.111.

Eichler, Evan E. 2001. "Recent Duplication, Domain Accretion and the Dynamic Mutation of the Human Genome." *Trends in*

Genetics 17 (11): 661–69. https://doi.org/10.1016/S0168-9525(01)02492-1.

Eid, John, Adrian Fehr, Jeremy Gray, Khai Luong, John Lyle, Geoff Otto, Paul Peluso, et al. 2009. "Real-Time DNA Sequencing from Single Polymerase Molecules." *Science* 323 (5910): 133–38. https://doi.org/10.1126/science.1162986.

Estrada, Alejandro, Paul A. Garber, Anthony B. Rylands, Christian Roos, Eduardo Fernandez-Duque, Anthony Di Fiore, K. Anne-Isola Nekaris, et al. 2017. "Impending Extinction Crisis of the World's Primates: Why Primates Matter." *Science Advances* 3 (1): e1600946. https://doi.org/10.1126/sciadv.1600946.

Florio, Marta, Mareike Albert, Elena Taverna, Takashi Namba, Holger Brandl, Eric Lewitus, Christiane Haffner, et al. 2015. "Human-Specific Gene ARHGAP11B Promotes Basal Progenitor Amplification and Neocortex Expansion." *Science* 347 (6229): 1465–70. https://doi.org/10.1126/science.aaa1975.

Franklin, Rosalind E., and R. G. Gosling. 1953. "Molecular Configuration in Sodium Thymonucleate." *Nature* 171 (4356): 740–41. https://doi.org/10.1038/171740a0.

Ghurye, Jay, Mihai Pop, Sergey Koren, Derek Bickhart, and Chen-Shan Chin. 2017. "Scaffolding of Long Read Assemblies Using Long Range Contact Information." *BMC Genomics* 18 (1): 527. https://doi.org/10.1186/s12864-017-3879-z.

Gibbs, Richard A., Jeffrey Rogers, Michael G. Katze, Roger Bumgarner, George M. Weinstock, Elaine R. Mardis, Karin A. Remington, et al. 2007. "Evolutionary and Biomedical Insights from the Rhesus Macaque Genome." *Science* 316 (5822): 222–34. https://doi.org/10.1126/science.1139247.

Gnerre, Sante, Iain MacCallum, Dariusz Przybylski, Filipe J. Ribeiro, Joshua N. Burton, Bruce J. Walker, Ted Sharpe, et al. 2011. "High-Quality Draft Assemblies of Mammalian Genomes from Massively Parallel Sequence Data." *Proceedings of the National Academy of Sciences* 108 (4): 1513–18. https://doi.org/10.1073/pnas.1017351108.

Goodwin, Sara, John D. McPherson, and W. Richard McCombie. 2016. "Coming of Age: Ten Years of next-Generation Sequencing Technologies." *Nature Reviews Genetics* 17 (6): 333–51. https://doi.org/10.1038/nrg.2016.49.

Gordon, David, John Huddleston, Mark J. P. Chaisson, Christopher M. Hill, Zev N. Kronenberg, Katherine M. Munson, Maika

Malig, et al. 2016. "Long-Read Sequence Assembly of the Gorilla Genome." *Science* 352 (6281): aae0344. https://doi.org/10.1126/science.aae0344.

Gray, J. W., A. V. Carrano, L. L. Steinmetz, M. A. Van Dilla, D. H. Moore, B. H. Mayall, and M. L. Mendelsohn. 1975. "Chromosome Measurement and Sorting by Flow Systems." *Proceedings of the National Academy of Sciences of the United States of America* 72 (4): 1231–34. https://doi.org/10.1073/pnas.72.4.1231.

Green, Philip. 1997. "Against a Whole-Genome Shotgun." *Genome Research* 7 (5): 410–17. https://doi.org/10.1101/gr.7.5.410.

Green, Richard E., Johannes Krause, Adrian W. Briggs, Tomislav Maricic, Udo Stenzel, Martin Kircher, Nick Patterson, et al. 2010. "A Draft Sequence of the Neandertal Genome." *Science* 328 (5979): 710–22. https://doi.org/10.1126/science.1188021.

Hancks, Dustin C., and Haig H. Kazazian. 2016. "Roles for Retrotransposon Insertions in Human Disease." *Mobile DNA* 7 (May). https://doi.org/10.1186/s13100-016-0065-9.

He, Yaoxi, Xin Luo, Bin Zhou, Ting Hu, Xiaoyu Meng, Peter A. Audano, Zev N. Kronenberg, et al. 2019. "Long-Read Assembly of the Chinese Rhesus Macaque Genome and Identification of Ape-Specific Structural Variants." *Nature Communications* 10 (1): 1–14. https://doi.org/10.1038/s41467-019-12174-w.

Huddleston, John, Swati Ranade, Maika Malig, Francesca Antonacci, Mark Chaisson, Lawrence Hon, Peter H. Sudmant, et al. 2014. "Reconstructing Complex Regions of Genomes Using Long-Read Sequencing Technology." *Genome Research* 24 (4): 688–96. https://doi.org/10.1101/gr.168450.113.

Hughes, Jennifer F., and David C. Page. 2015. "The Biology and Evolution of Mammalian Y Chromosomes." *Annual Review of Genetics* 49: 507–27. https://doi.org/10.1146/annurev-genet-112414-055311.

Hughes, Jennifer F., Helen Skaletsky, Laura G. Brown, Tatyana Pyntikova, Tina Graves, Robert S. Fulton, Shannon Dugan, et al. 2012. "Strict Evolutionary Conservation Followed Rapid Gene Loss on Human and Rhesus Y Chromosomes." *Nature* 483 (7387): 82–86. https://doi.org/10.1038/nature10843.

Hughes, Jennifer F., Helen Skaletsky, Tatyana Pyntikova, Tina A. Graves, Saskia K. M. van Daalen, Patrick J. Minx, Robert S. Fulton, et al. 2010. "Chimpanzee and Human Y

Chromosomes Are Remarkably Divergent in Structure and Gene Content." *Nature* 463 (7280): 536–39. https://doi.org/10.1038/nature08700.

International Human Genome Sequencing Consortium. 2001. "Initial Sequencing and Analysis of the Human Genome." *Nature* 409 (6822): 860–921. https://doi.org/10.1038/35057062.

———. 2004. "Finishing the Euchromatic Sequence of the Human Genome." *Nature* 431 (7011): 931–45. https://doi.org/10.1038/nature03001.

Jain, Miten, Sergey Koren, Karen H. Miga, Josh Quick, Arthur C. Rand, Thomas A. Sasani, John R. Tyson, et al. 2018. "Nanopore Sequencing and Assembly of a Human Genome with Ultra-Long Reads." *Nature Biotechnology* 36 (4): 338–45. https://doi.org/10.1038/nbt.4060.

Jain, Miten, Hugh E. Olsen, Daniel J. Turner, David Stoddart, Kira V. Bulazel, Benedict Paten, David Haussler, Huntington F. Willard, Mark Akeson, and Karen H. Miga. 2018. "Linear Assembly of a Human Centromere on the Y Chromosome." *Nature Biotechnology* 36 (4): 321–23. https://doi.org/10.1038/nbt.4109.

Jobling, Mark A., and Chris Tyler-Smith. 2017. "Human Y-Chromosome Variation in the Genome-Sequencing Era." *Nature Reviews Genetics* 18 (8): 485–97. https://doi.org/10.1038/nrg.2017.36.

Ju, Xiang-Chun, Qiong-Qiong Hou, Ai-Li Sheng, Kong-Yan Wu, Yang Zhou, Ying Jin, Tieqiao Wen, Zhengang Yang, Xiaoqun Wang, and Zhen-Ge Luo. 2016. "The Hominoid-Specific Gene TBC1D3 Promotes Generation of Basal Neural Progenitors and Induces Cortical Folding in Mice." Edited by Joseph G Gleeson. *ELife* 5 (August): e18197. https://doi.org/10.7554/eLife.18197.

Kazazian, Haig H., and John V. Moran. 2017. "Mobile DNA in Health and Disease." *The New England Journal of Medicine* 377 (4): 361–70. https://doi.org/10.1056/NEJMra1510092.

Kelley, J M, C E Field, M B Craven, D Bocskai, U J Kim, S D Rounsley, and M D Adams. 1999. "High Throughput Direct End Sequencing of BAC Clones." *Nucleic Acids Research* 27 (6): 1539–46.

Koepfli, Klaus-Peter, Benedict Paten, and Stephen J. O'Brien. 2015. "The Genome 10K Project: A Way Forward." *Annual Review of*

Animal Biosciences 3: 57–111. https://doi.org/10.1146/annurev-animal-090414-014900.

Kolmogorov, Mikhail, Jeffrey Yuan, Yu Lin, and Pavel A. Pevzner. 2019. "Assembly of Long, Error-Prone Reads Using Repeat Graphs." *Nature Biotechnology* 37 (5): 540–46. https://doi.org/10.1038/s41587-019-0072-8.

Koren, Sergey, and Adam M Phillippy. 2015. "One Chromosome, One Contig: Complete Microbial Genomes from Long-Read Sequencing and Assembly." *Current Opinion in Microbiology*, Host–microbe interactions: bacteria • Genomics, 23 (February): 110–20. https://doi.org/10.1016/j.mib.2014.11.014.

Koren, Sergey, Brian P. Walenz, Konstantin Berlin, Jason R. Miller, Nicholas H. Bergman, and Adam M. Phillippy. 2017. "Canu: Scalable and Accurate Long-Read Assembly via Adaptive k-Mer Weighting and Repeat Separation." *Genome Research*, March, gr.215087.116. https://doi.org/10.1101/gr.215087.116.

Kronenberg, Zev N., Ian T. Fiddes, David Gordon, Shwetha Murali, Stuart Cantsilieris, Olivia S. Meyerson, Jason G. Underwood, et al. 2018. "High-Resolution Comparative Analysis of Great Ape Genomes." *Science* 360 (6393): eaar6343. https://doi.org/10.1126/science.aar6343.

Lahn, Bruce T., and David C. Page. 1997. "Functional Coherence of the Human Y Chromosome." *Science* 278 (5338): 675–80. https://doi.org/10.1126/science.278.5338.675.

———. 1999. "Four Evolutionary Strata on the Human X Chromosome." *Science* 286 (5441): 964–67. https://doi.org/10.1126/science.286.5441.964.

Lander, Eric S., and Michael S. Waterman. 1988. "Genomic Mapping by Fingerprinting Random Clones: A Mathematical Analysis." *Genomics* 2 (3): 231–39. https://doi.org/10.1016/0888-7543(88)90007-9.

Larsen, Peter A., R. Alan Harris, Yue Liu, Shwetha C. Murali, C. Ryan Campbell, Adam D. Brown, Beth A. Sullivan, et al. 2017. "Hybrid de Novo Genome Assembly and Centromere Characterization of the Gray Mouse Lemur (Microcebus Murinus)." *BMC Biology* 15 (1): 110. https://doi.org/10.1186/s12915-017-0439-6.

Lewin, Harris A., Gene E. Robinson, W. John Kress, William J. Baker, Jonathan Coddington, Keith A. Crandall, Richard Durbin, et al. 2018. "Earth BioGenome Project: Sequencing Life for the Future of Life." *Proceedings of the National Academy of Sciences* 115 (17): 4325–33. https://doi.org/10.1073/pnas.1720115115.

Li, Heng, Jonathan M Bloom, Yossi Farjoun, Mark Fleharty, Laura Gauthier, Benjamin Neale, and Daniel MacArthur. 2018. "A Synthetic-Diploid Benchmark for Accurate Variant Calling Evaluation." *Nature Methods* 15 (8): 595–97. https://doi.org/10.1038/s41592-018-0054-7.

Li, Ruiqiang, Wei Fan, Geng Tian, Hongmei Zhu, Lin He, Jing Cai, Quanfei Huang, et al. 2010. "The Sequence and *de Novo* Assembly of the Giant Panda Genome." *Nature* 463 (7279): 311–17. https://doi.org/10.1038/nature08696.

Li, Ruiqiang, Hongmei Zhu, Jue Ruan, Wubin Qian, Xiaodong Fang, Zhongbin Shi, Yingrui Li, et al. 2010. "De Novo Assembly of Human Genomes with Massively Parallel Short Read Sequencing." *Genome Research* 20 (2): 265–72. https://doi.org/10.1101/gr.097261.109.

Li, Zhenyu, Yanxiang Chen, Desheng Mu, Jianying Yuan, Yujian Shi, Hao Zhang, Jun Gan, et al. 2012. "Comparison of the Two Major Classes of Assembly Algorithms: Overlap–Layout–Consensus and de-Bruijn-Graph." *Briefings in Functional Genomics* 11 (1): 25–37. https://doi.org/10.1093/bfgp/elr035.

Lieberman-Aiden, Erez, Nynke L. van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, et al. 2009. "Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome." *Science* 326 (5950): 289–93. https://doi.org/10.1126/science.1181369.

Lindblad-Toh, Kerstin, Claire M. Wade, Tarjei S. Mikkelsen, Elinor K. Karlsson, David B. Jaffe, Michael Kamal, Michele Clamp, et al. 2005. "Genome Sequence, Comparative Analysis and Haplotype Structure of the Domestic Dog." *Nature* 438 (7069): 803–19. https://doi.org/10.1038/nature04338.

Liu, Qian, Li Fang, Guoliang Yu, Depeng Wang, Chuan-Le Xiao, and Kai Wang. 2019. "Detection of DNA Base Modifications by Deep Recurrent Neural Network on Oxford Nanopore Sequencing Data." *Nature Communications* 10 (1): 1–11. https://doi.org/10.1038/s41467-019-10168-2.

Locke, Devin P., LaDeana W. Hillier, Wesley C. Warren, Kim C. Worley, Lynne V. Nazareth, Donna M. Muzny, Shiaw-Pyng Yang, et al. 2011. "Comparative and Demographic Analysis of

Orang-Utan Genomes." *Nature* 469 (7331): 529–33. https://doi.org/10.1038/nature09687.

Lucotte, Elise A., Laurits Skov, Jacob Malte Jensen, Moisès Coll Macià, Kasper Munch, and Mikkel H. Schierup. 2018. "Dynamic Copy Number Evolution of X- and Y-Linked Ampliconic Genes in Human Populations." *Genetics* 209 (3): 907–20. https://doi.org/10.1534/genetics.118.300826.

Mallick, Swapan, Heng Li, Mark Lipson, Iain Mathieson, Melissa Gymrek, Fernando Racimo, Mengyao Zhao, et al. 2016. "The Simons Genome Diversity Project: 300 Genomes from 142 Diverse Populations." *Nature* 538 (7624): 201–6. https://doi.org/10.1038/nature18964.

Margulies, Marcel, Michael Egholm, William E. Altman, Said Attiya, Joel S. Bader, Lisa A. Bemben, Jan Berka, et al. 2005. "Genome Sequencing in Microfabricated High-Density Picolitre Reactors." *Nature* 437 (7057): 376–80. https://doi.org/10.1038/nature03959.

Marks, Patrick, Sarah Garcia, Alvaro Martinez Barrio, Kamila Belhocine, Jorge Bernate, Rajiv Bharadwaj, Keith Bjornson, et al. 2019. "Resolving the Full Spectrum of Human Genome Variation Using Linked-Reads." *Genome Research* 29 (4): 635–45. https://doi.org/10.1101/gr.234443.118.

Marques-Bonet, Tomas, Oliver A. Ryder, and Evan E. Eichler. 2009. "Sequencing Primate Genomes: What Have We Learned?" *Annual Review of Genomics and Human Genetics* 10 (1): 355–86. https://doi.org/10.1146/annurev.genom.9.081307.164420.

McLean, Cory Y., Philip L. Reno, Alex A. Pollen, Abraham I. Bassan, Terence D. Capellini, Catherine Guenther, Vahan B. Indjeian, et al. 2011. "Human-Specific Loss of Regulatory DNA and the Evolution of Human-Specific Traits." *Nature* 471 (7337): 216–19. https://doi.org/10.1038/nature09774.

Metcalfe, Cushla J., Jonathan Filée, Isabelle Germon, Jean Joss, and Didier Casane. 2012. "Evolution of the Australian Lungfish (Neoceratodus Forsteri) Genome: A Major Role for CR1 and L2 LINE Elements." *Molecular Biology and Evolution* 29 (11): 3529–39. https://doi.org/10.1093/molbev/mss159.

Meyer, Matthias, Martin Kircher, Marie-Theres Gansauge, Heng Li, Fernando Racimo, Swapan Mallick, Joshua G. Schraiber, et al. 2012. "A High-Coverage Genome Sequence from an Archaic Denisovan Individual." *Science* 338 (6104): 222–26. https://doi.org/10.1126/science.1224344.

Miga, Karen H., Sergey Koren, Arang Rhie, Mitchell R. Vollger, Ariel Gershman, Andrey Bzikadze, Shelise Brooks, et al. 2019. "Telomere-to-Telomere Assembly of a Complete Human X Chromosome." *BioRxiv*, August, 735928. https://doi.org/10.1101/735928.

Miller, Jason R., Arthur L. Delcher, Sergey Koren, Eli Venter, Brian P. Walenz, Anushka Brownley, Justin Johnson, Kelvin Li, Clark Mobarry, and Granger Sutton. 2008. "Aggressive Assembly of Pyrosequencing Reads with Mates." *Bioinformatics* 24 (24): 2818–24. https://doi.org/10.1093/bioinformatics/btn548.

Minoche, André E., Juliane C. Dohm, and Heinz Himmelbauer. 2011. "Evaluation of Genomic High-Throughput Sequencing Data Generated on Illumina HiSeq and Genome Analyzer Systems." *Genome Biology* 12 (11): R112. https://doi.org/10.1186/gb-2011-12-11-r112.

Mouse Genome Sequencing Consortium, Robert H. Waterston, Kerstin Lindblad-Toh, Ewan Birney, Jane Rogers, Josep F. Abril, Pankaj Agarwal, et al. 2002. "Initial Sequencing and Comparative Analysis of the Mouse Genome." *Nature* 420 (6915): 520–62. https://doi.org/10.1038/nature01262.

Myers, Eugene W., Granger G. Sutton, Art L. Delcher, Ian M. Dew, Dan P. Fasulo, Michael J. Flanigan, Saul A. Kravitz, et al. 2000. "A Whole-Genome Assembly of Drosophila." *Science* 287 (5461): 2196–2204. https://doi.org/10.1126/science.287.5461.2196.

Nagarajan, Niranjan, and Mihai Pop. 2013. "Sequence Assembly Demystified." *Nature Reviews Genetics* 14 (3): 157–67. https://doi.org/10.1038/nrg3367.

Nater, Alexander, Maja P. Mattle-Greminger, Anton Nurcahyo, Matthew G. Nowak, Marc de Manuel, Tariq Desai, Colin Groves, et al. 2017. "Morphometric, Behavioral, and Genomic Evidence for a New Orangutan Species." *Current Biology* 27 (22): 3487-3498.e10. https://doi.org/10.1016/j.cub.2017.09.047.

Niedringhaus, Thomas P., Denitsa Milanova, Matthew B. Kerby, Michael P. Snyder, and Annelise E. Barron. 2011. "Landscape of Next-Generation Sequencing Technologies." *Analytical Chemistry* 83 (12): 4327–41. https://doi.org/10.1021/ac2010857.

Nowoshilow, Sergej, Siegfried Schloissnig, Ji-Feng Fei, Andreas Dahl, Andy W. C. Pang, Martin Pippel, Sylke Winkler, et al. 2018.

"The Axolotl Genome and the Evolution of Key Tissue Formation Regulators." *Nature* 554 (7690): 50–55. https://doi.org/10.1038/nature25458.

Parra, Genis, Keith Bradnam, and Ian Korf. 2007. "CEGMA: A Pipeline to Accurately Annotate Core Genes in Eukaryotic Genomes." *Bioinformatics* 23 (9): 1061–67. https://doi.org/10.1093/bioinformatics/btm071.

Payne, Alexander, Nadine Holmes, Vardhman Rakyan, and Matthew Loose. 2019. "BulkVis: A Graphical Viewer for Oxford Nanopore Bulk FAST5 Files." *Bioinformatics* 35 (13): 2193–98. https://doi.org/10.1093/bioinformatics/bty841.

Pellicer, Jaume, Michael F. Fay, and Ilia J. Leitch. 2010. "The Largest Eukaryotic Genome of Them All?" *Botanical Journal of the Linnean Society* 164 (1): 10–15. https://doi.org/10.1111/j.1095-8339.2010.01072.x.

Peona, Valentina, Matthias H. Weissensteiner, and Alexander Suh. 2018. "How Complete Are 'Complete' Genome Assemblies?—An Avian Perspective." *Molecular Ecology Resources* 18 (6): 1188–95. https://doi.org/10.1111/1755-0998.12933.

Pevzner, Pavel A., and Haixu Tang. 2001. "Fragment Assembly with Double-Barreled Data." *Bioinformatics* 17 (suppl_1): S225–33. https://doi.org/10.1093/bioinformatics/17.suppl_1.S225.

Phillips, Kimberley A., Karen L. Bales, John P. Capitanio, Alan Conley, Paul W. Czoty, Bert A. 't Hart, William D. Hopkins, et al. 2014. "Why Primate Models Matter." *American Journal of Primatology* 76 (9): 801–27. https://doi.org/10.1002/ajp.22281.

Prüfer, Kay, Kasper Munch, Ines Hellmann, Keiko Akagi, Jason R. Miller, Brian Walenz, Sergey Koren, et al. 2012. "The Bonobo Genome Compared with the Chimpanzee and Human Genomes." *Nature* 486 (7404): 527–31. https://doi.org/10.1038/nature11128.

Putnam, Nicholas H., Brendan L. O'Connell, Jonathan C. Stites, Brandon J. Rice, Marco Blanchette, Robert Calef, Christopher J. Troll, et al. 2016. "Chromosome-Scale Shotgun Assembly Using an in Vitro Method for Long-Range Linkage." *Genome Research* 26 (3): 342–50. https://doi.org/10.1101/gr.193474.115.

Rat Genome Sequencing Project Consortium. 2004. "Genome Sequence of the Brown Norway Rat Yields Insights into

Mammalian Evolution." *Nature* 428 (6982): 493–521. https://doi.org/10.1038/nature02426.

Raymond, Christopher K., Arnold Kas, Marcia Paddock, Ruolan Qiu, Yang Zhou, Sandhya Subramanian, Jean Chang, et al. 2005. "Ancient Haplotypes of the HLA Class II Region." *Genome Research* 15 (9): 1250–57. https://doi.org/10.1101/gr.3554305.

Rice, Edward S., and Richard E. Green. 2019. "New Approaches for Genome Assembly and Scaffolding." *Annual Review of Animal Biosciences* 7 (1): 17–40. https://doi.org/10.1146/annurev-animal-020518-115344.

Robinson, Jacqueline A., Jannikke Räikkönen, Leah M. Vucetich, John A. Vucetich, Rolf O. Peterson, Kirk E. Lohmueller, and Robert K. Wayne. 2019. "Genomic Signatures of Extensive Inbreeding in Isle Royale Wolves, a Population on the Threshold of Extinction." *Science Advances* 5 (5): eaau0757. https://doi.org/10.1126/sciadv.aau0757.

Rogers, Jeffrey, Muthuswamy Raveendran, R. Alan Harris, Thomas Mailund, Kalle Leppälä, Georgios Athanasiadis, Mikkel Heide Schierup, et al. 2019. "The Comparative Genomics and Complex Population History of Papio Baboons." *Science Advances* 5 (1): eaau6947. https://doi.org/10.1126/sciadv.aau6947.

Ross, Mark T., Darren V. Grafham, Alison J. Coffey, Steven Scherer, Kirsten McLay, Donna Muzny, Matthias Platzer, et al. 2005. "The DNA Sequence of the Human X Chromosome." *Nature* 434 (7031): 325–37. https://doi.org/10.1038/nature03440.

Salzberg, Steven L., Adam M. Phillippy, Aleksey Zimin, Daniela Puiu, Tanja Magoc, Sergey Koren, Todd J. Treangen, et al. 2012. "GAGE: A Critical Evaluation of Genome Assemblies and Assembly Algorithms." *Genome Research* 22 (3): 557–67. https://doi.org/10.1101/gr.131383.111.

Sanger, F., S. Nicklen, and A. R. Coulson. 1977. "DNA Sequencing with Chain-Terminating Inhibitors." *Proceedings of the National Academy of Sciences of the United States of America* 74 (12): 5463–67.

Scally, Aylwyn, Julien Y. Dutheil, LaDeana W. Hillier, Gregory E. Jordan, Ian Goodhead, Javier Herrero, Asger Hobolth, et al. 2012. "Insights into Hominid Evolution from the Gorilla Genome Sequence." *Nature* 483 (7388): 169–75. https://doi.org/10.1038/nature10842.

Schmitz, Jürgen, Angela Noll, Carsten A. Raabe, Gennady Churakov, Reinhard Voss, Martin Kiefmann, Timofey Rozhdestvensky, et al. 2016. "Genome Sequence of the Basal Haplorrhine Primate Tarsius Syrichta Reveals Unusual Insertions." *Nature Communications* 7 (1): 1–11. https://doi.org/10.1038/ncomms12997.

Simão, Felipe A., Robert M. Waterhouse, Panagiotis Ioannidis, Evgenia V. Kriventseva, and Evgeny M. Zdobnov. 2015. "BUSCO: Assessing Genome Assembly and Annotation Completeness with Single-Copy Orthologs." *Bioinformatics (Oxford, England)* 31 (19): 3210–12. https://doi.org/10.1093/bioinformatics/btv351.

Simpson, Jared T., Rachael E. Workman, P. C. Zuzarte, Matei David, L. J. Dursi, and Winston Timp. 2017. "Detecting DNA Cytosine Methylation Using Nanopore Sequencing." *Nature Methods* 14 (4): 407–10. https://doi.org/10.1038/nmeth.4184.

Sinclair, A. H., P. Berta, M. S. Palmer, J. R. Hawkins, B. L. Griffiths, M. J. Smith, J. W. Foster, A. M. Frischauf, R. Lovell-Badge, and P. N. Goodfellow. 1990. "A Gene from the Human Sex-Determining Region Encodes a Protein with Homology to a Conserved DNA-Binding Motif." *Nature* 346 (6281): 240–44. https://doi.org/10.1038/346240a0.

Skaletsky, Helen, Tomoko Kuroda-Kawaguchi, Patrick J. Minx, Holland S. Cordum, LaDeana Hillier, Laura G. Brown, Sjoerd Repping, et al. 2003. "The Male-Specific Region of the Human Y Chromosome Is a Mosaic of Discrete Sequence Classes." *Nature* 423 (6942): 825–37. https://doi.org/10.1038/nature01722.

Skinner, Benjamin M., Carole A. Sargent, Carol Churcher, Toby Hunt, Javier Herrero, Jane E. Loveland, Matt Dunn, et al. 2016. "The Pig X and Y Chromosomes: Structure, Sequence, and Evolution." *Genome Research* 26 (1): 130–39. https://doi.org/10.1101/gr.188839.114.

Skov, Laurits, The Danish Pan Genome Consortium, and Mikkel Heide Schierup. 2017. "Analysis of 62 Hybrid Assembled Human Y Chromosomes Exposes Rapid Structural Changes and High Rates of Gene Conversion." *PLOS Genetics* 13 (8): e1006834. https://doi.org/10.1371/journal.pgen.1006834.

Supple, Megan A., and Beth Shapiro. 2018. "Conservation of Biodiversity in the Genomics Era." *Genome Biology* 19 (1): 131. https://doi.org/10.1186/s13059-018-1520-3.

Sutton, Granger G., Owen White, Mark D. Adams, and Anthony R. Kerlavage. 1995. "TIGR Assembler: A New Tool for Assembling Large Shotgun Sequencing Projects." *Genome Science and Technology* 1 (1): 9–19. https://doi.org/10.1089/gst.1995.1.9.

Teitz, Levi S., Tatyana Pyntikova, Helen Skaletsky, and David C. Page. 2018. "Selection Has Countered High Mutability to Preserve the Ancestral Copy Number of Y Chromosome Amplicons in Diverse Human Lineages." *The American Journal of Human Genetics* 103 (2): 261–75. https://doi.org/10.1016/j.ajhg.2018.07.007.

The 1000 Genomes Project Consortium. 2015. "A Global Reference for Human Genetic Variation." *Nature* 526 (7571): 68–74. https://doi.org/10.1038/nature15393.

The C. elegans Sequencing Consortium. 1998. "Genome Sequence of the Nematode C. Elegans: A Platform for Investigating Biology." *Science* 282 (5396): 2012–18. https://doi.org/10.1126/science.282.5396.2012.

The Chimpanzee Sequencing and Analysis Consortium. 2005. "Initial Sequence of the Chimpanzee Genome and Comparison with the Human Genome." *Nature* 437 (7055): 69–87. https://doi.org/10.1038/nature04072.

The International Wheat Genome Sequencing Consortium (IWGSC). 2014. "A Chromosome-Based Draft Sequence of the Hexaploid Bread Wheat (Triticum Aestivum) Genome." *Science* 345 (6194): 1251788. https://doi.org/10.1126/science.1251788.

The Marmoset Genome Sequencing and Analysis Consortium, Kim C. Worley, Wesley C. Warren, Jeffrey Rogers, Devin Locke, Donna M. Muzny, Elaine R. Mardis, et al. 2014. "The Common Marmoset Genome Provides Insight into Primate Biology and Evolution." *Nature Genetics* 46 (8): 850–57. https://doi.org/10.1038/ng.3042.

Tomaszkiewicz, Marta, Paul Medvedev, and Kateryna D. Makova. 2017. "Y and W Chromosome Assemblies: Approaches and Discoveries." *Trends in Genetics* 33 (4): 266–82. https://doi.org/10.1016/j.tig.2017.01.008.

Tomaszkiewicz, Marta, Samarth Rangavittal, Monika Cechova, Rebeca Campos Sanchez, Howard W. Fescemyer, Robert Harris, Danling Ye, et al. 2016. "A Time- and Cost-Effective Strategy to Sequence Mammalian Y Chromosomes: An Application to the de Novo Assembly of Gorilla Y." *Genome Research* 26 (4): 530–40. https://doi.org/10.1101/gr.199448.115.

Treangen, Todd J., and Steven L. Salzberg. 2012. "Repetitive DNA and Next-Generation Sequencing: Computational Challenges and Solutions." *Nature Reviews Genetics* 13 (1): 36–46. https://doi.org/10.1038/nrg3117.

Valk, Tom van der, David Díez-del-Molino, Tomas Marques-Bonet, Katerina Guschanski, and Love Dalén. 2019. "Historical Genomes Reveal the Genomic Consequences of Recent Population Decline in Eastern Gorillas." *Current Biology* 29 (1): 165-170.e6. https://doi.org/10.1016/j.cub.2018.11.055.

Varki, Ajit, Daniel H. Geschwind, and Evan E. Eichler. 2008. "Human Uniqueness: Genome Interactions with Environment, Behaviour and Culture." *Nature Reviews Genetics* 9 (10): 749–63. https://doi.org/10.1038/nrg2428.

Vaser, Robert, Ivan Sovic, Niranjan Nagarajan, and Mile Sikic. 2017. "Fast and Accurate de Novo Genome Assembly from Long Uncorrected Reads." *Genome Research*, January, gr.214270.116. https://doi.org/10.1101/gr.214270.116.

Vegesna, Rahulsimham, Marta Tomaszkiewicz, Paul Medvedev, and Kateryna D. Makova. 2019. "Dosage Regulation, and Variation in Gene Expression and Copy Number of Human Y Chromosome Ampliconic Genes." *PLOS Genetics* 15 (9): e1008369. https://doi.org/10.1371/journal.pgen.1008369.

Venter, J. Craig, Mark D. Adams, Eugene W. Myers, Peter W. Li, Richard J. Mural, Granger G. Sutton, Hamilton O. Smith, et al. 2001. "The Sequence of the Human Genome." *Science* 291 (5507): 1304–51. https://doi.org/10.1126/science.1058040.

Vollger, Mitchell R., Philip C. Dishuck, Melanie Sorensen, AnneMarie E. Welch, Vy Dang, Max L. Dougherty, Tina A. Graves-Lindsay, Richard K. Wilson, Mark J. P. Chaisson, and Evan E. Eichler. 2019. "Long-Read Sequence and Assembly of Segmental Duplications." *Nature Methods* 16 (1): 88–94. https://doi.org/10.1038/s41592-018-0236-3.

Wade, C. M., E. Giulotto, S. Sigurdsson, M. Zoli, S. Gnerre, F. Imsland, T. L. Lear, et al. 2009. "Genome Sequence, Comparative Analysis, and Population Genetics of the Domestic Horse." *Science* 326 (5954): 865–67. https://doi.org/10.1126/science.1178158.

Walker, Bruce J., Thomas Abeel, Terrance Shea, Margaret Priest, Amr Abouelliel, Sharadha Sakthikumar, Christina A. Cuomo, et al.

2014. “Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement.” *PLOS ONE* 9 (11): e112963. https://doi.org/10.1371/journal.pone.0112963.

Warren, Wesley C., Anna J. Jasinska, Raquel Garcia-perez, Hannes Svardal, Chad Tomlinson, Mariano Rocchi, Nicoletta Archidiacono, et al. 2015. “The Genome of the Vervet (Chlorocebus Aethiops Sabaeus).” *Genome Research*, September, gr.192922.115. https://doi.org/10.1101/gr.192922.115.

Watson, J. D., and F. H. C. Crick. 1953. “Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid.” *Nature* 171 (4356): 737–38. https://doi.org/10.1038/171737a0.

Weber, James L., and Eugene W. Myers. 1997. “Human Whole-Genome Shotgun Sequencing.” *Genome Research* 7 (5): 401–9. https://doi.org/10.1101/gr.7.5.401.

Weisenfeld, Neil I., Vijay Kumar, Preyas Shah, Deanna M. Church, and David B. Jaffe. 2017. “Direct Determination of Diploid Genome Sequences.” *Genome Research* 27 (5): 757–67. https://doi.org/10.1101/gr.214874.116.

Weisenfeld, Neil I., Shuangye Yin, Ted Sharpe, Bayo Lau, Ryan Hegarty, Laurie Holmes, Brian Sogoloff, et al. 2014. “Comprehensive Variation Discovery in Single Human Genomes.” *Nature Genetics* 46 (12): 1350–55. https://doi.org/10.1038/ng.3121.

Wheeler, David A., Maithreyan Srinivasan, Michael Egholm, Yufeng Shen, Lei Chen, Amy McGuire, Wen He, et al. 2008. “The Complete Genome of an Individual by Massively Parallel DNA Sequencing.” *Nature* 452 (7189): 872–76. https://doi.org/10.1038/nature06884.

Wilkins, M. H. F., A. R. Stokes, and H. R. Wilson. 1953. “Molecular Structure of Nucleic Acids: Molecular Structure of Deoxypentose Nucleic Acids.” *Nature* 171 (4356): 738–40. https://doi.org/10.1038/171738a0.

Xue, Yali, Javier Prado-Martinez, Peter H. Sudmant, Vagheesh Narasimhan, Qasim Ayub, Michal Szpak, Peter Frandsen, et al. 2015. “Mountain Gorilla Genomes Reveal the Impact of Long-Term Population Decline and Inbreeding.” *Science* 348 (6231): 242–45. https://doi.org/10.1126/science.aaa3952.

Zerbino, Daniel R., and Ewan Birney. 2008. “Velvet: Algorithms for de Novo Short Read Assembly Using de Bruijn Graphs.”

Genome Research 18 (5): 821–29. https://doi.org/10.1101/gr.074492.107.

Zhao, Hui, Yong-Hui Jiang, and Yong Q. Zhang. 2018. "Modeling Autism in Non-Human Primates: Opportunities and Challenges." *Autism Research : Official Journal of the International Society for Autism Research* 11 (5): 686–94. https://doi.org/10.1002/aur.1945.

Zook, Justin M., Brad Chapman, Jason Wang, David Mittelman, Oliver Hofmann, Winston Hide, and Marc Salit. 2014. "Integrating Human Sequence Data Sets Provides a Resource of Benchmark SNP and Indel Genotype Calls." *Nature Biotechnology* 32 (3): 246–51. https://doi.org/10.1038/nbt.2835.