genes

Review

Probably Correct: Rescuing Repeats with Short and Long Reads

Monika Cechova

Genetics and Reproductive Biotechnologies, Veterinary Research Institute, Central European Institute of Technology (CEITEC), 621 00 Brno, Czech Republic; cechova.biomonika@gmail.com

Abstract: Ever since the introduction of high-throughput sequencing following the human genome project, assembling short reads into a reference of sufficient quality posed a significant problem as a large portion of the human genome—estimated 50–69%—is repetitive. As a result, a sizable proportion of sequencing reads is multi-mapping, i.e., without a unique placement in the genome. The two key parameters for whether or not a read is multi-mapping are the read length and genome complexity. Long reads are now able to span difficult, heterochromatic regions, including full centromeres, and characterize chromosomes from "telomere to telomere". Moreover, identical reads or repeat arrays can be differentiated based on their epigenetic marks, such as methylation patterns, aiding in the assembly process. This is despite the fact that long reads still contain a modest percentage of sequencing errors, disorienting the aligners and assemblers both in accuracy and speed. Here, I review the proposed and implemented solutions to the repeat resolution and the multi-mapping read problem, as well as the downstream consequences of reference choice, repeat masking, and proper representation of sex chromosomes. I also consider the forthcoming challenges and solutions with regards to long reads, where we expect the shift from the problem of repeat localization within a single individual to the problem of repeat positioning within pangenomes.

Keywords: repeats; satellite; multi-mapping; reference; long reads

Citation: Cechova, M. Probably Correct: Rescuing Repeats with Short and Long Reads. *Genes* **2021**, *12*, 48. https://doi.org/10.3390/genes12010048

Received: 9 November 2020
Accepted: 24 December 2020
Published: 31 December 2020

1. Introduction

While next-generation sequencing is increasingly used both in research and clinical practice, a subset of sequencing reads is frequently underutilized. These are reads that cannot be uniquely positioned within their respective genomes, and are thus multi-mapping in the chosen reference assembly. They frequently originate from duplicated genes [1], transposable elements [2,3], satellite repeats (e.g., centromeric and telomeric reads) [4], and more generally a heterochromatic portion of the genome [5]. Indeed, an estimated 50–69% of the human genome is repetitive [6,7], as is as much as 80% of the maize genome [8]. Multi-mapping reads are also an unavoidable consequence of segmental and whole-genome duplications [9,10]. Additional sources of nearly identical sequences are allelic variants and haplotypes.

The two most important parameters for whether a read is multi-mapping are the read length and genome complexity, whereas genome complexity can also be defined in terms of the length of repetitive units, i.e., length-sensitive [11]. Importantly, in order to characterize the repeat array, the reads need to be longer than r, where r is the length of the repeat unit. What is the minimal sequencing read length required to capture the repeat array? During PCR design, two ~20bp oligonucleotides can be anchored in the genome. Yet, such a reaction could still yield an unspecified product if one or both primers were positioned in the repetitive regions, requiring an optimization of the PCR reaction. Thus, the repetitiveness of any genome can be defined in terms of read length required to successfully assemble it. The k-mer uniqueness ratio is defined as the percentage of the genome that is covered by unique sequences of length k or longer [12]. For a variety of organisms, a k of at least 50 is required to cover a significant portion of their respective genomes [12]. The mappability, where the genome is divided into windows of size k, and the uniqueness of each window is calculated (i.e., mappability of 0.5 means exactly two identical windows exist in the genome), can be calculated for any *k* (e.g., with the GEM library toolkit), and provided as a genome track. Note that for human, as much as 28.4% of reads are unmappable to the assembled portion of the human genome with the read length 20, while only 2% with the read length 200 [13,14]. These numbers refer to the assembled portion of the human genome, which increased significantly from hg19 to hg38. The assembled portion impacts called variants [15], and is expected to rise further with the assembly by the Telomere-to-Telomere (T2T) consortium.

Along with the technical variation (read length/error rate), there is a substantial biological variability outside of the reference genome sequences, especially in the form of satellite repeats. According to the "satellite library" hypothesis, an initial set of sequences can lead to the vast variability of outcomes, generating changes in sequence and copy number in individuals and populations [16]. Indeed, among human populations, the centromeric array of the X chromosome can vary by an order of magnitude (0.5–5 Mb) [17]. Wei and colleagues showed that repeat clustering did not recapitulate the expected relationships in geographically separated populations of *Drosophila* [18]. In great apes, Cechova and colleagues described vast variability among satellite repeats in great apes [19]. Intriguingly, different sequencing technologies provide different repeat estimates, although they agree qualitatively (abundant versus rare repeats) [19]; I discuss some potential reasons later. Approaches and software for satellite biology are reviewed in [20] and include both short- and long-read solutions.

Because of the intrinsic difficulty of dealing with repetitive parts of the genome, sometimes it might be advantageous to remove the repetitiveness in order to study underlying biological processes, such as cell division. As an example, an artificial, non-repetitive centromeric region was created to study centromere genomics with the use of human artificial chromosomes (HACs) [21]. Last, next-generation sequencing reads, even if not multi-mapping, can fall short of capturing the full repeat variability of individuals [22], especially when compared to a single reference genome (i.e., limited representation of a genome).

2. Reference Genomes Are Inherently Incomplete

Reference genomes represent a simplified, linear representation of the conceivable version of a genome of a given species [23]. Such references are incomplete: even the best representations contain gaps in difficult, heterochromatic parts of the genome [24]. Moreover, as much as 5–10% of the human genome remains poorly characterized [25], and up to tens of percentage points might be completely missing in other organisms, such as birds [26], all while underestimating the copy number of repetitive regions. This is because high-identity regions are often collapsed during the assembly process from short sequencing reads [27] or long erroneous reads. As an example, only <0.1% of GRCh38 reference is composed of repeats HSAT2/3 but as much as 2.6% of read bases are HSAT2/3 [28]. When dealing with such hard-to-assemble regions, it might be advantageous to use "the most likely representation", rather than the reference assembled from any living individual [17]. This idea has been implemented for centromeric arrays [17]. Instead of a multi-megabase gap as in previous human reference genomes, GRCh38 centromeres are composed of these presumed sequences, based on the second-order Markov models of monomer variants. Still, complete assembly is the ultimate solution to repeats [29].

Recently, a newly established Telomere-to-Telomere consortium aimed to assemble human chromosomes in full and present a new (near) complete sequence of a human genome [30,31], including the first complete sequence of the human chromosome X [32]. However, gaps in reference genomes are not the sole reason why reference assemblies are generally incomplete. The variability among individuals of a given species means that the full sequence content simply cannot be captured by looking at a single individual [33–36]. A recent study identified an additional 300 Mbp of sequences (although predominantly HSAT2 and HSAT3) that were not represented in the human reference (GRCh38) but found among 910 individuals of African descent [35]. Another study found 46 Mb in 1000 Swedish individuals [36], complementing a previous study that performed de novo assemblies of two Swedish genomes and revealed as much as 10 Mbps of novel sequence (originating from centromeric and telomeric regions and the chromosome Y), almost one-third of which was different from any sequences present in existing nucleotide databases [37]. The degree to which an individual is represented by a reference also depends on ancestry. The individuals that do not match the ancestry of the particular reference genome built might be misrepresented, leading to false variant calls or uninterpretable GWAS results [38,39]. One of the options is to use ancestry-specific reference builts—an example of which might be the Japanese reference genome [40]. Similar considerations apply to different haplogroups, but, as expected, representing all haplotypes with a contig each leads to a multi-mapping problem (contig from a specific haplotype is referred to as a haplotig). In summary, reference genomes are either incomplete or introduce multi-mapping issues at the allelic, haplogroup, or chromosomal level.

The downstream analysis—such as mapping accuracy, gene expression analysis, and calling of structural variants—are affected by the following: (1) the specific reference genome (that comes in multiple private and public versions) [41], (2) whether or not the genome is repeat masked, and (3) the representation of pseudoautosomal regions (PARs) and alternative haplotypes. The understanding of the reference genome as the representative species genome should be uncoupled from the sequence that is to serve as an alignment reference [23]. First, for typical applications, it might be advantageous not to use alternative haplotigs and to mask large multi-copy sequences such as PARs and a small subset of α-satellites that are artificially identical in the current GRC reference genomes [41]. In this scenario, one must wary that variant calls from these regions might originate from more than one genomic region. However, not including unplaced and unlocalized contigs might force reads from these contigs to be mapped to the chromosomal part of the reference and again lead to false variant calls. On the other hand, aligners typically assign mapping quality 0 to multi-mapping reads, and such reads might be ignored by downstream pipelines.

Sometimes, multiple reference genomes are concatenated and used as a mapping target: good examples are the inclusion (or a lack thereof) of a mitochondrial genome or sequences of spike-in controls. Crucially, even if one is interested only in a single chromosome, sequencing reads still need to be mapped to a full reference genome. This is because if no other chromosomes are offered as a mapping target, the read counts will become overrepresented—as much as one-third of all sequencing reads could map to a single chromosome in the case of the human genome due to repeats; this proportion drops significantly when repeat masking is in place (see Table 1). Second, genes can have repetitive parts (in both exons [42] and introns [43,44]) and intergenic regions can be low-complexity; thus, repeat masking the reference genome will result in an increase in the proportion of unmapped reads (Table 1). Third, some parts of the genomes, e.g., repetitive heterogametic sex chromosomes (chromosome Y in mammals and chromosome W in birds), are often underrepresented. In summary, the particular version of the reference genome must be carefully considered and chosen contingent on the desired application.

Table 1. The comparison in a proportion of mapped reads (%) when using the whole-genome reference, compared to individual chromosomes. Both full and repeat-masked references are contrasted. The SRR622461 dataset of the NA12878 female individual was mapped with bwa mem version 0.7.17-r1188 and default parameters to either unmasked or masked human reference genome hg38. All reads were mapped either to the full reference or to the respective chromosome only.

Chromosome Name	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
mapping proportion [%] (to hg38)	7.6	7.7	6.6	6.4	6.0	5.4	5.0	4.6	3.9	4.3	4.2	4.2
mapping proportion [%] (to itself)	33.8	34.7	32.1	32.4	31.5	30.6	31.5	30.0	30.3	30.7	29.0	29.3
mapping proportion [%] (to masked hg38)	5.2	5.3	4.4	4.1	3.9	3.8	3.4	3.2	2.6	3.0	3.3	2.8
mapping proportion [%] (to masked itself)	6.2	6.5	4.0	4.0	3.7	3.6	4.5	3.0	3.6	4.5	3.1	2.7
Chromosome Name	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**X**	**Y** [1]
mapping proportion [%] (to hg38)	3.2	2.8	2.5	2.7	2.4	2.5	1.6	2.1	1.4	1.2	4.9	0.2
mapping proportion [%] (to itself)	27.7	28.6	27.3	27.7	27.6	26.9	25.0	26.9	25.8	26.0	29.9	23.6
mapping proportion [%] (to masked hg38)	2.2	2.0	1.8	2.0	1.7	1.7	0.9	1.6	0.9	0.8	2.7	0.4
mapping proportion [%] (to masked itself)	2.1	2.0	1.8	3.1	3.1	1.7	1.0	2.7	1.2	2.8	2.6	2.1

[1] Note the spurious hits to the Y chromosome using the female reads.

3. Short Reads

Repetitive regions are hard to resolve and are variable among individuals and technologies: both biological and technical variability is present. If the reads are mapped to such repetitive reference, how should the multi-mapping reads be dealt with? Four main approaches exist: use first position or the “grouped assignment”, uniform assignment, random assignment, and, last, context-dependent distribution (Figure 1).

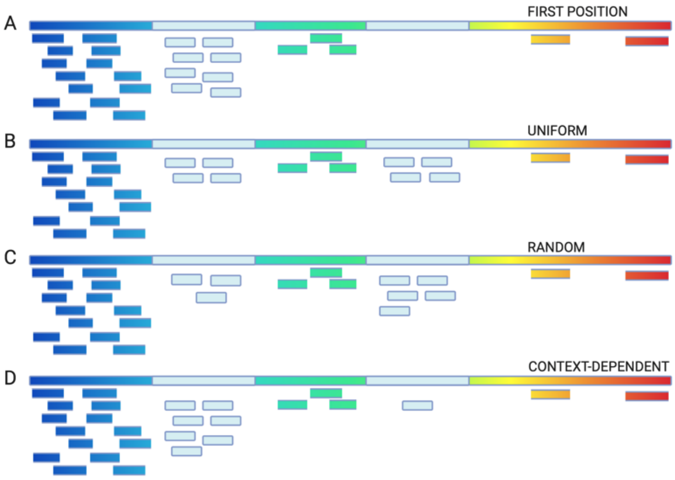

Figure 1. The mapping strategies for the multi-mapping reads. Reference genome (rainbow) with two duplicate regions (light blue). The multi-mapping reads (light blue) are distributed among mappings based on (**A**) the first position or the grouped assignment, (**B**) uniformly among all equally good mappings, (**C**) randomly, or (**D**) context-dependent assignment based on the coverage of neighboring sites. Figure created with BioRender.com.

3.1. Methods of Multi-Mapping Read Assignment

The first approach maps reads to the first possible position out of multiple equally good options (Figure 1A). If no further post-processing is applied, this approach can lead to erroneous downstream conclusions (e.g., in early versions of Tophat all multi-mapping reads were in some instances aligned to the same locus). For example, if for all duplicated genes, only the first annotated gene copy is used as a mapping target, which then biases read counts and an expression value towards the first gene/isoform. Alternatively, all such reads could be annotated as a group: presenting both unique regions and the expression of the "unassigned, multi-mapping group".

The second approach (although rather theoretical) assigns all multi-mapping reads uniformly: the same number of reads is assigned to each of the mapping targets (Figure 1B). While this might not reflect the biological reality, at least it does not favor any particular instance of a repeat.

The third approach represents a naive allocation of multi-mapping reads that randomly separates reads into one of the "equally good" positions (Figure 1C). This is typically a default option in many short-read aligners, including bwa [45], bowtie [46], HiSAT [47], and STAR [48]. In this implementation, multiple runs will yield slightly different results.

The fourth approach represents one of the more sophisticated approaches that aim to gauge which option is more likely biologically.

Last, reference-free approaches can be applied, such as read clustering. For example, the authors of [49] attempted to characterize satellites directly from short unassembled reads, using clustering and visualizations for which they offered biological interpretations [49].

The choice of the multi-mapping strategy can artificially bias read counts related to specific annotations, yet I believe that the effect sizes of these strategies are largely unknown in the scientific literature.

3.2. Multi-Mapping Reads in RNA-Seq, Chip-Seq, Hi-C, and Exome Sequencing

For RNA-Seq, the most common pipelines deal with multi-mapping reads as follows; HTSeq-count and STAR geneCounts ignore them, while Cufflinks can either split reads equally or use uniquely mapped reads as guidance; the latter option could be problematic for small noncoding RNAs, especially those located in the introns [50]. For a comprehensive review on handling multi-mapping reads in RNASeq datasets, see the work in [50]. In general, it is believed that the expression levels for the genes that contain multi-mapping reads are underestimated [51]; and that hundreds of genes, many of which are relevant for human health, could be affected [51]. One solution is to use group-level expression (for a set of genes) to circumvent the multi-mapping problem [51]. Thus, read counts could be calculated separately for the input genes and merged genes, as was implemented in the tool mmquant, applying a gene clustering strategy [52]. More sophisticated approaches involve hierarchical allocation of reads: first resolving ambiguities among genes, followed by isoforms and individual alleles. A similar approach was implemented in the software MMSEQ [53] or EMASE (Expectation-Maximization for Allele-Specific Expression) [54]. Algorithms implementing EM or Expectation-Maximization are important for assigning multi-mapping reads in RSEM [55], and for pseudoalignments with Kallisto/Salmon [56,57]. Instead of mapping reads to a reference, pseudoalignments estimate which transcripts could have generated them. As short sequencing reads still contain a limited number of sequencing errors, some methods have attempted to extend the array of possible mapping contexts, thus accounting for these errors, in order to identify the most likely mapping in RNA-Seq experiments [58].

Analogous approaches have been used for ChIP-seq datasets; when analyzing transcription binding sites, some regions in the genome are known to bind transcription factors based on the information from the uniquely mapped reads. Thus, one might choose to distribute multi-mapping reads preferentially to those regions (Figure 1D). Zhang and colleagues have proposed models that use local concentrations of directional reads and account for local genome repetitiveness (using whole-genome read mappability profiles) while differentiating between adjacent binding sites [59].

The typical pipelines for a chromosome conformation capture, such as Hi-C, filter out multi-mapping reads, as they are not considered useful for delineating chromatin interactions. However, this disproportionally affects repetitive parts of the genome, such as the Y chromosome, where ampliconic regions span as much as almost half and over half of the male-specific euchromatic proportion in human and chimpanzee, respectively [60]. Ampliconic regions host large inverted repeats, i.e., palindromes, which are then omitted from traditional pipelines. Using a probabilistic approach implemented in the package mHi-C [61], Cechova, Vegesna, and colleagues were able to analyze palindromic arms of human palindromes and found a higher density of chromatin interactions; none of these regions could be analyzed if multi-mapping reads had been excluded [62].

Last, multi-mapping reads also affect whole-genome and exome sequencing data, as well as small RNAs [63], where the effect is especially pronounced due to biologically constrained read lengths, and finally metagenomics, where only subtle differences might exist between various strains [64].

3.3. Repeat Masking and Its Consequences

Repeat masking refers to a process in which the underlying sequence gets marked as repetitive (typically with repetitive parts in lowercase letters: soft-masking) or fully suppressed (typically replaced by Ns/Xs: hard-masking). Soft-masking is relevant whenever the specific part of the sequence is visually inspected (e.g., during primer design or when examining gaps in the assembly), as a repeat-masked sequence might hint why the given region was challenging to analyze. In contrast, hard-masking might be advantageous in specific applications (e.g., the pseudoautosomal region on the chromosome Y is typically hard-masked). Importantly, masking will not only affect heterochromatic parts of the human reference genome but also genes and other regulatory sequences that may carry repeats. In summary, repeat masking (especially hard-masking) will have an effect on all downstream processes that depend on the read mapping, including, but not limited to, variant calling, gene expression, and chromatin capture analysis.

3.4. Sex Chromosomes

In order to build a new reference genome, only the homogametic sex is typically sequenced, as heterogametic sex chromosomes are harder to assemble at any given sequencing coverage (there is always fewer data for a given sex chromosome compared to an autosome) and because chromosomes Y/W tend to be repetitive [65]. The omission of sex chromosomes is especially prominent in the GWAS studies [66–68]. Sex chromosomes require special considerations during the mapping process. The reads from chromosomally male and female individuals will align differently depending on whether the reference even contains the Y/W chromosome [69]. The tool XYalign can identify XX and XY individuals across different experimental conditions (including low-coverage samples and exome sequencing), as well as improve variant calling on sex chromosomes [69]. Moreover, X and Y chromosomes share a pseudoautosomal region, in which homologous sequences have a high degree of identity and still recombine. Accounting for sex chromosomes can increase the number of unique genes identified as differentially expressed between the sexes and to increase expression estimates in the pseudoautosomal region of the X chromosome [70].

4. Long Reads

The biggest promise of long (yet erroneous but see HiFi reads) reads is to span difficult, repetitive, and heterochromatic regions in the genomes or populations of interest [71–74]. Long reads, especially in combination with other orthogonal technologies enhancing the assembly contiguity (such as chromatin captures/Hi-C [75] or optical maps [76]), are now turning near-complete or complete reference genomes into reality [31,32,77]. Additionally, traditionally difficult regions, such as the Major Histocompatibility Complex (MHC), have been recently characterized in detail with PacBio and Nanopore sequencing [73,78–81]. This opens up the possibility to also survey the satellite content, either directly from long reads [82] or from assemblies/references [83,84]. Hundreds of kilobase pairs in read length ensure that many repeat arrays are fully encompassed within the sequencing reads. Indeed, Cechova and colleagues demonstrated that depending on the species, 90–95% and 99% of abundant repeat arrays were fully nested within individual reads in Nanopore and PacBio, respectively [19]. Moreover, the intra-repeat array variability was present: among the 39 most abundant repeats in the great ape genomes, at least 10–25% of all arrays were composed of a mix of different repeated motifs [19]. Such satellites can either be surveyed de novo (when the set of repeats in the genome is unknown; this is difficult due to the intermixing of sequencing errors and rare variants) or by searching for a specific set of repeats when the presumed errors can be "canceled out", as was implemented in the Noise-cancelling repeat finder (NCRF) [82]. Specifically, NCRF can identify long satellite arrays in Nanopore and PacBio reads, notwithstanding their length or error rate. However, the variability in such satellite arrays is challenging to capture in standard file formats such as Variant Call Format (VCF). Comparing VCF files across experiments requires accounting for both the combination of sequencing errors (potentially shifting starting positions) and tandem representation of biologically imperfect repeats. Even in the most simple scenario of small Illumina variants, multiple equivalent representations are possible [85]. The tools for the reconciliation of long tandem arrays, and their comparisons, are currently being developed [86]. Satellite arrays inside the existing assemblies/references can be curated, e.g., with TandemTools, a novel tool for the polishing and quality assessment of extra-long tandem repeats (ETRs) [87]. Specific long-read mappers can be used to align reads to highly repetitive reference sequences [88], while accounting for the allele bias (so that non-reference allele within a repeat does not penalize the alignment) [89]. In summary, long reads are much better equipped to characterize the variable satellite content and to assemble and span difficult, repetitive parts of the genome.

4.1. Long-Read Sequencing Strategies

The biggest challenge before fully utilizing long reads is their elevated raw error rate, previously (2019) reported at 14.90% and 16.10% in PacBio and Nanopore, respectively, and continuously improving since. The most recent reports suggest 95% (and higher) accuracy for raw Nanopore reads [90], especially due to the improvements in pore design and basecalling algorithms. Raw reads can further be used to build consensus, which is the preferred strategy for PacBio; consensus HiFi reads are both long (>10 kbp), and accurate (>99.9%) [91,92]. This is because HiFi reads require circularizing of the DNA, so that it can be read multiple times over to increase accuracy, rendering them an effective technology for the genome assembly problem (see below).

PacBio and Nanopore differ in their estimates of the repeat copy number and overall repeat content, even for the same individual [19,83] and also present with a strand bias [83]. The repeat content differs even after sequencing reads are subsampled to the common length distribution for both technologies, to account for the potential differences in read lengths [19]. The potential explanations include distinct library preparation protocols, DNA damage, DNA quality (that could potentially decline during prolonged sequencing at room temperature), and non-canonical DNA structures, as reviewed in [19]. Thus, some of the repeats discovered using just one technology might not be confirmed in another, and vice versa.

The multi-mapping problem I described for the short reads has in some ways shifted to larger scales—from arrays spanning a few kilobases we are now able to resolve megabase-long arrays. However, the sequencing errors still disorient the aligners both in accuracy and speed [93]. The alignments are bound to contain some "chance" alignments due to a matching subset of "chance" nucleotides. With the continued progress in basecalling, it is expected that the error rates will continue to decline, and that additional factors (such as epigenetic modifications) will be responsible for the ambiguous calls.

4.2. Differentiating (Nearly) Identical Repeat Arrays

Small variants in otherwise homogeneous repeat arrays are especially useful to aid in the assembly process and can be used to create a tiling path across repeats. For example, these distinct unique markers were reported on average every 2.3 kb in the centromeric satellite array on the X chromosome (DXZ1), with a maximum spacing of 42 kb [32]. Such spacing makes long reads especially useful in delineating repetitive arrays based on their unique markers, and assemblers HiCanu [91] and hifiasm [29] capitalize on this property. However, what if no unique markers are available? Is there a way one could differentiate between identical sequences of the same origin? What if we could, together with the sequence information, also capture the epigenetic modifications associated with these sequencing, and use them to differentiate among otherwise identical sequences to provide an "epigenetic phasing"? This is now possible with both PacBio and Nanopore, for PacBio using kinetic profiles (and recording the pausing of the polymerase) [94] and for Nanopore with electric signals (methylated bases modulate the raw signal) [95]. This can be done either directly [96–98] or through a base conversion in which a matched modified sample is created [99,100]. Therefore, unique markers in combination with epigenetic modification in the long reads are becoming increasingly useful in deciphering long satellite arrays, such as those in centromeres [32].

4.3. Long-Read Assemblies

Long reads, and especially accurate long reads (e.g., those delivered by the consensus HiFi reads from Pacific Biosciences, Menlo Park, CA, USA), are critical not only to improve the assembly quality and contiguity, but also importantly for haplotype phasing (especially in polyploid and allopolyploid genomes). Highly heterozygous genomes, high repeat content, and the presence of segmental duplications all add to the challenge. Several recent algorithms aim to address it. Segmental Duplication Assembler (SDA) [101] enables the partition of the assembly into distinct paralogs, recovering copy-number-variable paralogs that are absent from the human reference genome. To aid phasing, one can make use of the parental genomes (via "trio binning" [102] and derived algorithms). In this approach, each part of an assembly is partitioned into haplotypes (pre-binning strategy) and each haplotype is assembled separately with corresponding reads. This is implemented in HiCanu [91], a modification of the Canu assembler for HiFi reads. In contrast, hifiasm uses graph-binning strategy, allowing the correction of misassigned reads, and attempting to resolve all haplotypes, thus consistently delivering larger assembly contiguity [29]. Ultra-long Oxford Nanopore reads are structurally accurate and can be used to anchor highly accurate assembled HiFi contigs. This strategy was employed to produce a complete assembly of the human chromosome 8 by the T2T consortium [31]. Other strategies require no pedigree information for phasing and combine long reads with Hi-C [103] or single-cell strand sequencing data [104], or make use of several sequencing technologies [105]. Importantly, even if the genome size remains unaffected by the choice of an assembler or assembly parameters, the gene assembly can still be affected, especially when assembling highly heterozygous genomes [106]. This is due to regional sequence expansions or collapses in difficult-to-assemble regions [106]. To conclude, perfect haplotype-resolved assemblies with accurate MHC variants, satellite DNAs, and segmental duplications, all with complete repeat annotations—are now within reach. Last, I expect that after solving the heterochromatin and satellite repeats within a single individual, the focus will shift towards the problem of repeat positioning within pangenomes.

5. Future

A Pan-genome can be defined as a collection of genomic sequences to be analyzed jointly or to be used as a reference [107]. The incorporation of thousands of individuals into a single reference will avoid "reference bias", and mapping reads to such a pan-genome will improve variant calling, especially in regions with a high density of complex variants [107]. While many of the proposed pan-genome implementations represent genomes as graphs with shared and private variants, some of the new approaches have proposed elegant ways of creating pan-genome graphs while preserving linear coordinates [108]. In the future, ultra-long accurate reads, coupled with complete reference pan-genomes, will enable the full understanding of the underlying functional variation hidden in the repetitive parts of the genome. Until then, the considerations outlined in this review, such as reference choice, repeat masking, proper representation of sex chromosomes, and appropriately dealing with multi-mapping reads, will remain essential.

Funding: This work was supported by the Ministry of Education, Youth, and Sports of the Czech Republic under the project CEITEC 2020 (LQ1601).

Institutional Review Board Statement: Ethical review and approval were waived for this study, the 1000 Genomes data is made available according to the Fort Lauderdale Agreement.

Data Availability Statement: The data in Table 1 were generated with the use of the NA12878 individual from the 1000 Genomes Project Consortium [109] and SRR622461 run. Masked and unmasked reference genome hg38 produced by the Genome Reference Consortium was downloaded from the UCSC website.

Acknowledgments: The author would like to thank Matthias Weissensteiner and Marta Tomaszkiewicz for the critical reading of this manuscript.

Conflicts of Interest: The author declares no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Lallemand, T.; Leduc, M.; Landès, C.; Rizzon, C.; Lerat, E. An overview of duplicated gene detection methods: Why the duplication mechanism has to be accounted for in their choice. *Genes* **2020**, *11*, 1046. [CrossRef] [PubMed]
2. Lerat, E. Identifying repeats and transposable elements in sequenced genomes: How to find your way through the dense forest of programs. *Heredity* **2010**, *104*, 520–533. [CrossRef] [PubMed]
3. Kojima, K.K. Human transposable elements in Repbase: Genomic footprints from fish to humans. *Mob. DNA* **2018**, *9*, 2. [CrossRef] [PubMed]
4. Miga, K.H. Centromere studies in the era of "telomere-to-telomere"genomics. *Exp. Cell Res.* **2020**, *394*, 112127. [CrossRef]
5. Chaisson, M.J.P.; Huddleston, J.; Dennis, M.Y.; Sudmant, P.H.; Malig, M.; Hormozdiari, F.; Antonacci, F.; Surti, U.; Sandstrom, R.; Boitano, M.; et al. Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **2015**, *517*, 608–611. [CrossRef]
6. de Koning, A.P.J.; Gu, W.; Castoe, T.A.; Batzer, M.A.; Pollock, D.D. Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet.* **2011**, *7*, e1002384. [CrossRef]
7. Lander, E.S.; Linton, L.M.; Birren, B.; Nusbaum, C.; Zody, M.C.; Baldwin, J.; Devon, K.; Dewar, K.; Doyle, M.; FitzHugh, W.; et al. Initial sequencing and analysis of the human genome. *Nature* **2001**, *409*, 860–921.
8. Haberer, G.; Kamal, N.; Bauer, E.; Gundlach, H.; Fischer, I.; Seidel, M.A.; Spannagl, M.; Marcon, C.; Ruban, A.; Urbany, C.; et al. European maize genomes highlight intraspecies variation in repeat and gene content. *Nat. Genet.* **2020**, *52*, 950–957. [CrossRef]
9. Singh, P.P.; Affeldt, S.; Malaguti, G.; Isambert, H. Human dominant disease genes are enriched in paralogs originating from whole genome duplication. *PLoS Comput. Biol.* **2014**, *10*, e1003754. [CrossRef]
10. Sharp, A.J.; Locke, D.P.; McGrath, S.D.; Cheng, Z.; Bailey, J.A.; Vallente, R.U.; Pertz, L.M.; Clark, R.A.; Schwartz, S.; Segraves, R.; et al. Segmental duplications and copy-number variation in the human genome. *Am. J. Hum. Genet.* **2005**, *77*, 78–88. [CrossRef]
11. Phan, V.; Gao, S.; Tran, Q.; Vo, N.S. How genome complexity can explain the difficulty of aligning reads to genomes. *BMC Bioinform.* **2015**, *16*, S3. [CrossRef] [PubMed]
12. Schatz, M.C.; Delcher, A.L.; Salzberg, S.L. Assembly of large genomes using second-generation sequencing. *Genome Res.* **2010**, *20*, 1165–1173. [CrossRef] [PubMed]
13. Li, W.; Freudenberg, J.; Miramontes, P. Diminishing return for increased Mappability with longer sequencing reads: Implications of the k-mer distributions in the human genome. *BMC Bioinform.* **2014**, *15*, 2. [CrossRef] [PubMed]
14. Li, W.; Freudenberg, J. Mappability and read length. *Front. Genet.* **2014**, *5*, 381. [CrossRef] [PubMed]
15. Pan, B.; Kusko, R.; Xiao, W.; Zheng, Y.; Liu, Z.; Xiao, C.; Sakkiah, S.; Guo, W.; Gong, P.; Zhang, C.; et al. Similarities and differences between variants called with human reference genome HG19 or HG38. *BMC Bioinform.* **2019**, *20*, 101. [CrossRef]
16. Ugarković, Ð.; Plohl, M. Variation in satellite DNA profiles—Causes and effects. *EMBO J.* **2002**, *21*, 5955–5959. [CrossRef]
17. Miga, K.H.; Newton, Y.; Jain, M.; Altemose, N.; Willard, H.F.; Kent, W.J. Centromere reference models for human chromosomes X and Y satellite arrays. *Genome Res.* **2014**, *24*, 697–707. [CrossRef]
18. Wei, K.H.-C.; Grenier, J.K.; Barbash, D.A.; Clark, A.G. Correlated variation and population differentiation in satellite DNA abundance among lines of Drosophila melanogaster. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 18793–18798. [CrossRef]

19. Cechova, M.; Harris, R.S.; Tomaszkiewicz, M.; Arbeithuber, B.; Chiaromonte, F.; Makova, K.D. High satellite repeat turnover in great apes studied with short- and long-read technologies. *Mol. Biol. Evol.* **2019**, *36*. [CrossRef]
20. Lower, S.S.; McGurk, M.P.; Clark, A.G.; Barbash, D.A. Satellite DNA evolution: Old ideas, new approaches. *Curr. Opin. Genet. Dev.* **2018**, *49*, 70–78. [CrossRef]
21. Logsdon, G.A.; Gambogi, C.W.; Liskovykh, M.A.; Barrey, E.J.; Larionov, V.; Miga, K.H.; Heun, P.; Black, B.E. Human artificial chromosomes that bypass centromeric DNA. *Cell* **2019**, *178*, 624–639.e19. [CrossRef] [PubMed]
22. Miga, K.H. Centromeric satellite DNAs: Hidden sequence variation in the human population. *Genes* **2019**, *10*, 352. [CrossRef] [PubMed]
23. Schröder, J.; Girirajan, S.; Papenfuss, A.T.; Medvedev, P. Improving the power of structural variation detection by augmenting the reference. *PLoS ONE* **2015**, *10*, e0136771. [CrossRef] [PubMed]
24. Zhao, T.; Duan, Z.; Genchev, G.Z.; Lu, H. Closing human reference genome gaps: Identifying and characterizing gap-closing sequences. *G3* **2020**, *10*, 2801–2809. [CrossRef]
25. Altemose, N.; Miga, K.H.; Maggioni, M.; Willard, H.F. Genomic characterization of large heterochromatic gaps in the human genome assembly. *PLoS Comput. Biol.* **2014**, *10*, e1003628. [CrossRef]
26. Peona, V.; Weissensteiner, M.H.; Suh, A. How complete are "complete" genome assemblies? An avian perspective. *Mol. Ecol. Resour.* **2018**, *18*, 1188–1195. [CrossRef]
27. Salzberg, S.L.; Yorke, J.A. Beware of mis-assembled genomes. *Bioinformatics* **2005**, *21*, 4320–4321. [CrossRef]
28. Li, H. Identifying centromeric satellites with dna-brnn. *Bioinformatics* **2019**, *35*, 4408–4410. [CrossRef]
29. Cheng, H.; Concepcion, G.T.; Feng, X.; Zhang, H.; Li, H. Haplotype-resolved de novo assembly with phased assembly graphs. *arXiv* **2020**, arXiv:2008.01237v1.
30. GIS. The (Near) Complete Sequence of a Human Genome. Available online: https://genomeinformatics.github.io/CHM13v1/ (accessed on 25 October 2020).
31. Logsdon, G.A.; Vollger, M.R.; Hsieh, P.; Mao, Y.; Liskovykh, M.A.; Koren, S.; Nurk, S.; Mercuri, L.; Dishuck, P.C.; Rhie, A.; et al. The structure, function, and evolution of a complete human chromosome 8. *bioRxiv* **2020**. [CrossRef]
32. Miga, K.H.; Koren, S.; Rhie, A.; Vollger, M.R.; Gershman, A.; Bzikadze, A.; Brooks, S.; Howe, E.; Porubsky, D.; Logsdon, G.A.; et al. Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **2020**, *585*, 79–84. [CrossRef] [PubMed]
33. Liu, Y.; Koyutürk, M.; Maxwell, S.; Xiang, M.; Veigl, M.; Cooper, R.S.; Tayo, B.O.; Li, L.; LaFramboise, T.; Wang, Z.; et al. Discovery of common sequences absent in the human reference genome using pooled samples from next generation sequencing. *BMC Genom.* **2014**, *15*, 685. [CrossRef] [PubMed]
34. Li, R.; Tian, X.; Yang, P.; Fan, Y.; Li, M.; Zheng, H.; Wang, X.; Jiang, Y. Recovery of non-reference sequences missing from the human reference genome. *BMC Genom.* **2019**, *20*, 746. [CrossRef]
35. Sherman, R.M.; Forman, J.; Antonescu, V.; Puiu, D.; Daya, M.; Rafaels, N.; Boorgula, M.P.; Chavan, S.; Vergara, C.; Ortega, V.E.; et al. Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nat. Genet.* **2019**, *51*, 30–35. [CrossRef] [PubMed]
36. Eisfeldt, J.; Mårtensson, G.; Ameur, A.; Nilsson, D.; Lindstrand, A. Discovery of novel sequences in 1.000 Swedish genomes. *Mol. Biol. Evol.* **2020**, *37*, 18–30. [CrossRef] [PubMed]
37. Ameur, A.; Che, H.; Martin, M.; Bunikis, I.; Dahlberg, J.; Höijer, I.; Häggqvist, S.; Vezzi, F.; Nordlund, J.; Olason, P.; et al. De novo assembly of two Swedish genomes reveals missing segments from the human GRCh38 reference and improves variant calling of population-scale sequencing data. *Genes* **2018**, *9*, 486. [CrossRef]
38. Tian, C.; Gregersen, P.K.; Seldin, M.F. Accounting for ancestry: Population substructure and genome-wide association studies. *Hum. Mol. Genet.* **2008**, *17*, R143–R150. [CrossRef]
39. Martin, A.R.; Kanai, M.; Kamatani, Y.; Okada, Y.; Neale, B.M.; Daly, M.J. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **2019**, *51*, 584–591. [CrossRef]
40. Nagasaki, M.; Kuroki, Y.; Shibata, T.F.; Katsuoka, F.; Mimori, T.; Kawai, Y.; Minegishi, N.; Hozawa, A.; Kuriyama, S.; Suzuki, Y.; et al. Construction of JRG (Japanese reference genome) with single-molecule real-time sequencing. *Hum. Genome Var.* **2019**, *6*, 27. [CrossRef]
41. Li, H. Which Human Reference Genome to Use? Available online: https://lh3.github.io/2017/11/13/which-human-reference-genome-to-use (accessed on 14 October 2020).
42. Song, S.; Huang, Q.; Guo, J.; Li-Ling, J.; Chen, X.; Ma, F. Comparative component analysis of exons with different splicing frequencies. *PLoS ONE* **2009**, *4*, e5387. [CrossRef]
43. Liang, D.; Wilusz, J.E. Short intronic repeat sequences facilitate circular RNA production. *Genes Dev.* **2014**, *28*, 2233–2247. [CrossRef] [PubMed]
44. Lozada-Chávez, I.; Stadler, P.F.; Prohaska, S.J. Genome-wide features of introns are evolutionary decoupled among themselves and from genome size throughout Eukarya. *bioRxiv* **2018**. [CrossRef]
45. Li, H.; Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **2009**, *25*, 1754–1760. [CrossRef] [PubMed]
46. Langmead, B. Aligning short sequencing reads with Bowtie. *Curr. Protoc. Bioinform.* **2010**, *11*. [CrossRef] [PubMed]
47. Kim, D.; Langmead, B.; Salzberg, S.L. HISAT: A fast spliced aligner with low memory requirements. *Nat. Methods* **2015**, *12*, 357–360. [CrossRef] [PubMed]
48. Dobin, A.; Davis, C.A.; Schlesinger, F.; Drenkow, J.; Zaleski, C.; Jha, S.; Batut, P.; Chaisson, M.; Gingeras, T.R. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **2013**, *29*, 15–21. [CrossRef] [PubMed]

49. Novák, P.; Ávila Robledillo, L.; Koblížková, A.; Vrbová, I.; Neumann, P.; Macas, J. TAREAN: A computational tool for identification and characterization of satellite DNA from unassembled short reads. *Nucleic Acids Res.* **2017**, *45*, e111. [CrossRef]
50. Deschamps-Francoeur, G.; Simoneau, J.; Scott, M.S. Handling multi-mapped reads in RNA-seq. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 1569–1576. [CrossRef]
51. Robert, C.; Watson, M. Errors in RNA-Seq quantification affect genes of relevance to human disease. *Genome Biol.* **2015**, *16*, 177. [CrossRef]
52. Zytnicki, M. mmquant: How to count multi-mapping reads? *BMC Bioinform.* **2017**, *18*, 411. [CrossRef]
53. Turro, E.; Su, S.-Y.; Gonçalves, Â.; Coin, L.J.M.; Richardson, S.; Lewin, A. Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. *Genome Biol.* **2011**, *12*, R13. [CrossRef] [PubMed]
54. Raghupathy, N.; Choi, K.; Vincent, M.J.; Beane, G.L.; Sheppard, K.S.; Munger, S.C.; Korstanje, R.; Pardo-Manual de Villena, F.; Churchill, G.A. Hierarchical analysis of RNA-seq reads improves the accuracy of allele-specific expression. *Bioinformatics* **2018**, *34*, 2177–2184. [CrossRef] [PubMed]
55. Li, B.; Dewey, C.N. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinform.* **2011**, *12*, 323. [CrossRef]
56. Bray, N.L.; Pimentel, H.; Melsted, P.; Pachter, L. Erratum: Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **2016**, *34*, 888. [CrossRef]
57. Patro, R.; Duggal, G.; Love, M.I.; Irizarry, R.A.; Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **2017**, *14*, 417–419. [CrossRef]
58. Bonfert, T.; Csaba, G.; Zimmer, R.; Friedel, C.C. A context-based approach to identify the most likely mapping for RNA-seq experiments. *BMC Bioinform.* **2012**, *13*, S9. [CrossRef]
59. Zhang, X.; Robertson, G.; Krzywinski, M.; Ning, K.; Droit, A.; Jones, S.; Gottardo, R. PICS: Probabilistic inference for ChIP-seq. *Biometrics* **2011**, *67*, 151–163. [CrossRef]
60. Hughes, J.F.; Skaletsky, H.; Pyntikova, T.; Graves, T.A.; van Daalen, S.K.M.; Minx, P.J.; Fulton, R.S.; McGrath, S.D.; Locke, D.P.; Friedman, C.; et al. Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content. *Nature* **2010**, *463*, 536–539. [CrossRef]
61. Zheng, Y.; Ay, F.; Keles, S. Generative modeling of multi-mapping reads with mHi-C advances analysis of Hi-C studies. *eLife* **2019**, *8*, e38070. [CrossRef]
62. Cechova, M.; Vegesna, R.; Tomaszkiewicz, M.; Harris, R.S.; Chen, D.; Rangavittal, S.; Medvedev, P.; Makova, K.D. Dynamic evolution of great ape Y chromosomes. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 26273–26280. [CrossRef]
63. Johnson, N.R.; Yeoh, J.M.; Coruh, C.; Axtell, M.J. Improved placement of multi-mapping small RNAs. *G3* **2016**, *6*, 2103–2111. [CrossRef] [PubMed]
64. Nielsen, H.B.; Almeida, M.; Juncker, A.S.; Rasmussen, S.; Li, J.; Sunagawa, S.; Plichta, D.R.; Gautier, L.; Pedersen, A.G.; Le Chatelier, E.; et al. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat. Biotechnol.* **2014**, *32*, 822–828. [CrossRef] [PubMed]
65. Tomaszkiewicz, M.; Medvedev, P.; Makova, K.D. Y and W chromosome assemblies: Approaches and discoveries. *Trends Genet.* **2017**, *33*, 266–282. [CrossRef] [PubMed]
66. Clayton, D.G. Sex chromosomes and genetic association studies. *Genome Med.* **2009**, *1*, 110. [CrossRef]
67. Anonymous. Accounting for sex in the genome. *Nat. Med.* **2017**, *23*, 1243. [CrossRef]
68. König, I.R.; Loley, C.; Erdmann, J.; Ziegler, A. How to include chromosome X in your genome-wide association study. *Genet. Epidemiol.* **2014**, *38*, 97–103. [CrossRef]
69. Webster, T.H.; Couse, M.; Grande, B.M.; Karlins, E.; Phung, T.N.; Richmond, P.A.; Whitford, W.; Wilson, M.A. Identifying, understanding, and correcting technical artifacts on the sex chromosomes in next-generation sequencing data. *Gigascience* **2019**, *8*. [CrossRef]
70. Olney, K.C.; Brotman, S.M.; Andrews, J.P.; Valverde-Vesling, V.A.; Wilson, M.A. Reference genome and transcriptome informed by the sex chromosome complement of the sample increase ability to detect sex differences in gene expression from RNA-Seq data. *Biol. Sex Differ.* **2020**, *11*, 42. [CrossRef]
71. Wick, R.R.; Holt, K.E. Benchmarking of long-read assemblers for prokaryote whole genome sequencing. *F1000Research* **2019**, *8*, 2138. [CrossRef]
72. Jain, M.; Olsen, H.E.; Turner, D.J.; Stoddart, D.; Bulazel, K.V.; Paten, B.; Haussler, D.; Willard, H.F.; Akeson, M.; Miga, K.H. Linear assembly of a human Y chromosome centromere. *Nat. Biotechnol.* **2018**, *36*, 321. [CrossRef]
73. Jain, M.; Koren, S.; Miga, K.H.; Quick, J.; Rand, A.C.; Sasani, T.A.; Tyson, J.R.; Beggs, A.D.; Dilthey, A.T.; Fiddes, I.T.; et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* **2018**, *36*, 338–345. [CrossRef] [PubMed]
74. Vollger, M.R.; Logsdon, G.A.; Audano, P.A.; Sulovari, A.; Porubsky, D.; Peluso, P.; Wenger, A.M.; Concepcion, G.T.; Kronenberg, Z.N.; Munson, K.M.; et al. Improved assembly and variant detection of a haploid human genome using single-molecule, high-fidelity long reads. *Ann. Hum. Genet.* **2020**, *84*, 125–140. [CrossRef] [PubMed]
75. Dudchenko, O.; Batra, S.S.; Omer, A.D.; Nyquist, S.K.; Hoeger, M.; Durand, N.C.; Shamim, M.S.; Machol, I.; Lander, E.S.; Aiden, A.P.; et al. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* **2017**, *356*, 92–95. [CrossRef] [PubMed]
76. Howe, K.; Wood, J.M.D. Using optical mapping data for the improvement of vertebrate genome assemblies. *GigaScience* **2015**, *4*, 10. [CrossRef] [PubMed]

77. Hoang, P.T.N.; Fiebig, A.; Novák, P.; Macas, J.; Cao, H.X.; Stepanenko, A.; Chen, G.; Borisjuk, N.; Scholz, U.; Schubert, I. Chromosome-scale genome assembly for the duckweed Spirodela intermedia, integrating cytogenetic maps, PacBio and Oxford Nanopore libraries. *Sci. Rep.* **2020**, *10*, 19230. [CrossRef]
78. Suzuki, S.; Ranade, S.; Osaki, K.; Ito, S.; Shigenari, A.; Ohnuki, Y.; Oka, A.; Masuya, A.; Harting, J.; Baybayan, P.; et al. Reference grade characterization of polymorphisms in full-length HLA class I and II genes with short-read sequencing on the ION PGM system and long-reads generated by single molecule, real-time sequencing on the PacBio platform. *Front. Immunol.* **2018**, *9*, 2294. [CrossRef]
79. Turner, T.R.; Hayhurst, J.D.; Hayward, D.R.; Bultitude, W.P.; Barker, D.J.; Robinson, J.; Madrigal, J.A.; Mayor, N.P.; Marsh, S.G.E. Single molecule real-time DNA sequencing of HLA genes at ultra-high resolution from 126 international HLA and immunogenetics workshop cell lines. *Hladnikia* **2018**, *91*, 88–101. [CrossRef]
80. Albrecht, V.; Zweiniger, C.; Surendranath, V.; Lang, K.; Schöfl, G.; Dahl, A.; Winkler, S.; Lange, V.; Böhme, I.; Schmidt, A.H. Dual redundant sequencing strategy: Full-length gene characterisation of 1056 novel and confirmatory HLA alleles. *Hladnikia* **2017**, *90*, 79–87. [CrossRef]
81. Chin, C.-S.; Wagner, J.; Zeng, Q.; Garrison, E.; Garg, S.; Fungtammasan, A.; Rautiainen, M.; Aganezov, S.; Kirsche, M.; Zarate, S.; et al. A diploid assembly-based benchmark for variants in the major histocompatibility complex. *Nat. Commun.* **2020**, *11*, 4794. [CrossRef]
82. Harris, R.S.; Cechova, M.; Makova, K.D. Noise-cancelling repeat finder: Uncovering tandem repeats in error-prone long-read sequencing data. *Bioinformatics* **2019**, *35*, 4809–4811. [CrossRef]
83. Mitsuhashi, S.; Frith, M.C.; Mizuguchi, T.; Miyatake, S.; Toyota, T.; Adachi, H.; Oma, Y.; Kino, Y.; Mitsuhashi, H.; Matsumoto, N. Tandem-genotypes: Robust detection of tandem repeat expansions from long DNA reads. *Genome Biol.* **2019**, *20*, 58. [CrossRef] [PubMed]
84. Ummat, A.; Bashir, A. Resolving complex tandem repeats with long reads. *Bioinformatics* **2014**, *30*, 3491–3498. [CrossRef] [PubMed]
85. Sun, C.; Medvedev, P. VarMatch: Robust matching of small variant datasets using flexible scoring schemes. *Bioinformatics* **2017**, *33*, 1301–1308. [CrossRef] [PubMed]
86. Mousavi, N.; Margoliash, J.; Pusarla, N.; Saini, S.; Yanicky, R.; Gymrek, M. TRTools: A toolkit for genome-wide analysis of tandem repeats. *Bioinformatics* **2020**. [CrossRef]
87. Mikheenko, A.; Bzikadze, A.V.; Gurevich, A.; Miga, K.H.; Pevzner, P.A. TandemTools: Mapping long reads and assessing/improving assembly quality in extra-long tandem repeats. *Bioinformatics* **2020**, *36*, i75–i83. [CrossRef]
88. Jain, C.; Rhie, A.; Zhang, H.; Chu, C.; Walenz, B.P.; Koren, S.; Phillippy, A.M. Weighted minimizer sampling improves long read mapping. *Bioinformatics* **2020**, *36*, i111–i118. [CrossRef]
89. Jain, C.; Rhie, A.; Hansen, N.; Koren, S.; Phillippy, A.M. A long read mapping method for highly repetitive reference sequences. *Cold Spring Harb. Lab.* **2020**, *2020*, 363887.
90. Nanopore Technologies. R10.3: The Newest Nanopore for High Accuracy Nanopore Sequencing. Available online: https://nanoporetech.com/about-us/news/r103-newest-nanopore-high-accuracy-nanopore-sequencing-now-available-store (accessed on 5 November 2020).
91. Nurk, S.; Walenz, B.P.; Rhie, A.; Vollger, M.R.; Logsdon, G.A.; Grothe, R.; Miga, K.H.; Eichler, E.E.; Phillippy, A.M.; Koren, S. HiCanu: Accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res.* **2020**, *30*, 1291–1305. [CrossRef]
92. Wenger, A.M.; Peluso, P.; Rowell, W.J.; Chang, P.-C.; Hall, R.J.; Concepcion, G.T.; Ebler, J.; Fungtammasan, A.; Kolesnikov, A.; Olson, N.D.; et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **2019**, *37*, 1155–1162. [CrossRef]
93. Salari, F.; Zare-Mirakabad, F.; Sadeghi, M.; Rokni-Zadeh, H. Assessing the impact of exact reads on reducing the error rate of read mapping. *BMC Bioinform.* **2018**, *19*, 406. [CrossRef]
94. Mondo, S.J.; Dannebaum, R.O.; Kuo, R.C.; Louie, K.B.; Bewick, A.J.; LaButti, K.; Haridas, S.; Kuo, A.; Salamov, A.; Ahrendt, S.R.; et al. Widespread adenine N6-methylation of active genes in fungi. *Nat. Genet.* **2017**, *49*, 964–968. [CrossRef] [PubMed]
95. Ding, H.; Bailey, A.D.; Jain, M.; Olsen, H.; Paten, B. Gaussian mixture model-based unsupervised nucleotide modification number detection using nanopore-sequencing readouts. *Bioinformatics* **2020**, *8*, 4928–4934. [CrossRef] [PubMed]
96. Beaulaurier, J.; Zhu, S.; Deikus, G.; Mogno, I.; Zhang, X.-S.; Davis-Richardson, A.; Canepa, R.; Triplett, E.W.; Faith, J.J.; Sebra, R.; et al. Metagenomic binning and association of plasmids with bacterial host genomes using DNA methylation. *Nat. Biotechnol.* **2018**, *36*, 61–69. [CrossRef] [PubMed]
97. Schatz, M.C. Nanopore sequencing meets epigenetics. *Nat. Methods* **2017**, *14*, 347–348. [CrossRef] [PubMed]
98. Schreiber, J.; Wescoe, Z.L.; Abu-Shumays, R.; Vivian, J.T.; Baatar, B.; Karplus, K.; Akeson, M. Error rates for nanopore discrimination among cytosine, methylcytosine, and hydroxymethylcytosine along individual DNA strands. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 18910–18915. [CrossRef] [PubMed]
99. Liu, Y.; Cheng, J.; Siejka-Zielińska, P.; Weldon, C.; Roberts, H.; Lopopolo, M.; Magri, A.; D'Arienzo, V.; Harris, J.M.; McKeating, J.A.; et al. Accurate targeted long-read DNA methylation and hydroxymethylation sequencing with TAPS. *Genome Biol.* **2020**, *21*, 54. [CrossRef]
100. Liu, Q.; Georgieva, D.C.; Egli, D.; Wang, K. NanoMod: A computational tool to detect DNA modifications using Nanopore long-read sequencing data. *BMC Genom.* **2019**, *20*, 78. [CrossRef]
101. Vollger, M.R.; Dishuck, P.C.; Sorensen, M.; Welch, A.E.; Dang, V.; Dougherty, M.L.; Graves-Lindsay, T.A.; Wilson, R.K.; Chaisson, M.J.P.; Eichler, E.E. Long-read sequence and assembly of segmental duplications. *Nat. Methods* **2019**, *16*, 88–94. [CrossRef]
102. Koren, S.; Rhie, A.; Walenz, B.P.; Dilthey, A.T.; Bickhart, D.M.; Kingan, S.B.; Hiendleder, S.; Williams, J.L.; Smith, T.P.L.; Phillippy, A.M. De novo assembly of haplotype-resolved genomes with trio binning. *Nat. Biotechnol.* **2018**, *36*, 1174–1182. [CrossRef]

103. Garg, S.; Fungtammasan, A.; Carroll, A.; Chou, M.; Schmitt, A.; Zhou, X.; Mac, S.; Peluso, P.; Hatas, E.; Ghurye, J.; et al. Chromosome-scale, haplotype-resolved assembly of human genomes. *Nat. Biotechnol.* **2020**. [CrossRef]
104. Porubsky, D.; Ebert, P.; Audano, P.A.; Vollger, M.R.; Harvey, W.T.; Marijon, P.; Ebler, J.; Munson, K.M.; Sorensen, M.; Sulovari, A.; et al. Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads. *Nat. Biotechnol.* **2020**. [CrossRef] [PubMed]
105. Di Genova, A.; Buena-Atienza, E.; Ossowski, S.; Sagot, M.-F. Efficient hybrid de novo assembly of human genomes with WENGAN. *Nat. Biotechnol.* **2020**. [CrossRef]
106. Asalone, K.C.; Ryan, K.M.; Yamadi, M.; Cohen, A.L.; Farmer, W.G.; George, D.J.; Joppert, C.; Kim, K.; Mughal, M.F.; Said, R.; et al. Regional sequence expansion or collapse in heterozygous genome assemblies. *PLoS Comput. Biol.* **2020**, *16*, e1008104. [CrossRef] [PubMed]
107. The Computational Pan-Genomics Consortium. Computational pan-genomics: Status, promises and challenges. *Brief. Bioinform.* **2018**, *19*, 118–135.
108. Li, H.; Feng, X.; Chu, C. The design and construction of reference pangenome graphs with minigraph. *Genome Biol.* **2020**, *21*, 265. [CrossRef] [PubMed]
109. The 1000 Genomes Project Consortium; Auton, A.; Brooks, L.D.; Durbin, R.M.; Garrison, E.P.; Kang, H.M.; Korbel, J.O.; Marchini, J.L.; McCarthy, S.; McVean, G.A.; et al. A global reference for human genetic variation. *Nature* **2015**, *526*, 68–74. [CrossRef] [PubMed]