Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 January 2021 doi:10.20944/preprints202101.0116.v1

Computational methods for chromosome-scale haplotype reconstruction

Shilpa Garg[†]

University of Copenhagen, Copenhagen, Denmark
Harvard Medical School, Boston, MA, USA

[†]e-mail: shilpa.garg2k7@gmail.com

Abstract

High-quality chromosome-scale haplotype sequences— of diploid genomes, polyploid genomes and metagenomes — provide important insights into genetic variation associated with disease and biodiversity. However, whole-genome short read sequencing does not yield haplotype information that spans whole chromosomes directly. Computational assembly of shorter haplotype fragments is required for haplotype reconstruction, which can be challenging owing to limited fragment lengths and high haplotype and repeat variability across genomes. Recent advancements in long-read and chromosome-scale sequencing technologies, alongside computational innovations, are improving the reconstruction of haplotypes at the level of whole chromosomes. Here, we review recent methodological progress in these areas and discuss perspectives that could enable routine high-quality haplotype reconstruction in clinical and evolutionary studies.

[H1] Introduction

Haplotypes are combinations of alleles or SNPs from multiple genetic loci on the same chromosome that are inherited together; the term haplotype can encompass as few as two loci or refer to a whole chromosome (that is, chromosome-scale haplotype). For diploid genomes, a given length of chromosomal DNA will have two haplotypes, one inherited from each parent, whereas several haplotypes exist for any given chromosomal region at the population level or for polyploid genomes. DNA microarrays and short-read sequencing can determine the collection of alleles at genetic loci (that is,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 January 2021 doi:10.20944/preprints202101.0116.v1

genotypes) but provide no information at the level of haplotypes, whether alleles are co-located on the same copy of a chromosome, or which of the parental chromosomes harbours a particular allele. Hence, computational reconstruction of haplotypes, by either read mapping to a reference genome or de novo assembly, is required.

Haplotype information is fundamental for medical and population genetics[1,2], where it is used to study genetic variation associated with human diseases[3,4]. Traditionally, only simple variations, such as SNPs, are studied, with respect to a linear reference sequence, and their association to diseases, for example, two SNPs, rs9494885 and rs2230926 in the TNFAIP3 gene, have known correlation with scleritis disease[5]. However, a collection of individual haplotype sequences (in the form of a pan-genome graph[6], which represents the entire genetic diversity of a population or species) can help to discover highly complex variations such as nested structural variation, inversions and other complex rearrangements[7] (reviewed in [8]). Moreover, as the distribution of *cis*- and *trans*-acting variants between homologous chromosomes, that is, the phase of variants, can affect gene expression, chromosome-scale haplotypes can help study interactions between variants in regulatory elements [4,9,10]. Another important application is in detecting large mosaic chromosomal alterations in cancer genomes[11] to study which haplotype leads to aneuploidy and trace the evolution of somatic mutations. Haplotypes also play important roles in understanding the interplay of evolutionary processes that shape genetic variation, such as recombination, gene conversion, mutation and selection. For example, modification of plant breeding strategies based on evolutionary processes identified through haplotype reconstruction can result in agricultural improvements[12]. Another highly relevant application occurs in the analysis of viral infections[13], where haplotype reconstruction can help to identify drug resistance and virulence factors and aid treatment decisions[14,15].

Despite recent advances, sequencing technologies are limited in their ability to cover repetitive genomic regions to produce telomere-to-telomere haplotypes. Therefore, local (short-range) and genome-wide (long-range) information must be computationally integrated to assemble chromosome-scale haplotypes[16]. The integrative algorithms used for reconstruction must be tuned for the specific genome characteristics of a species, such as genome size, number of haplotypes and repeat or

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 January 2021 doi:10.20944/preprints202101.0116.v1

haplotype variation. Many large-scale sequencing initiatives, such as the Vertebrate Genomes Project[17] , the DNA Zoo project (https://www.dnazoo.org/), Darwin Tree of Life (https://www.darwintreeoflife.org/), the Human Microbiome Project (https://www.hmpdacc.org/), and the Human Pangenome Project (https://humanpangenome.org/), have begun to leverage diverse recent sequencing data types (Table 1) to reconstruct haplotypes for various species. These projects have designed and integrated bioinformatic pipelines in a common platform for large-scale genome analyses[18].

In this Review, we discuss the bioinformatic methods — reference-based, de novo and strain-resolved metagenome assembly — to reconstruct haplotypes in diploids, polyploids and microbial communities. We present the strengths and weaknesses of these methods, alongside examples of their biological applications. Finally, we conclude with perspectives on remaining challenges and future directions, with an emphasis on both the algorithmic and technological progress required to achieve routine high-quality haplotypes for further biological discoveries.

[H1] Evolution of sequencing technologies

Early advancements in sequencing technologies[19], such as next-generation sequencing with read lengths of 150–250 bp and accuracy >99.8%, revolutionized haplotype reconstruction[20–22] and helped to characterize the genomic landscape. However, the fairly short read lengths limit the ability to uniquely span repeats and identify regions of heterozygosity, and these technologies are unable to produce whole-chromosome haplotypes. More recently, developments in long-read sequencing technologies[23,19] have begun to substantially increase the utility and application of haplotype reconstruction.

[H2] Short-range sequence information

Long-read sequencing technologies such as single-molecule real time sequencing from Pacific Biosciences (PacBio) and nanopore sequencing (including ultra-long) from Oxford Nanopore Technologies produce reads on the order of hundreds of kilobases in length, with error rates of 6–10%[24] (Fig. 1). The latest developments of PacBio's HiFi technology can produce reads with an average read length of 15–20 kb at error rates

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 January 2021 doi:10.20944/preprints202101.0116.v1

similar to short-read sequencing (that is, an accuracy of >99%)[25]. These advancements have made it possible to achieve near-complete human haplotypes that include microsatellites, repetitive elements and other complex structural variations[26], which were previously inaccessible. In addition to these 'true' long-read platforms, the Chromium technology from 10x Genomics[25] (10xG) employs genome partitioning and barcoding to generate linked reads that span tens to hundreds of thousands of bases. Finally, new optical mapping instruments from BioNano Genomics[27–29] can rapidly fingerprint megabase segments of a genome, enabling the detection of structural variation at a fairly low cost (Fig. 1). However, the limitation of local sequencing technologies is the inability to uniquely resolve near-perfect repeats above the size of their read length to produce full haplotypes. These limitations have necessitated the development of methods that can resolve haplotypes at the chromosome (or genome-wide) scale.

[H2] Long-range sequence information

Chromosome conformation capture methods such as Hi-C and related chromatin crosslinking protocols produce long-range, mate-pair-like data for short-read sequencing[27] (Fig. 1). Hi-C technology generates chimeric DNA fragments from two interacting chromosomal regions that are covalently linked together. These fragments are sequenced to produce paired-end reads that can span between a few kilobases and tens of megabases in physical distance.

Strand-seq is a recent single-cell sequencing advancement that allows independent sequencing of parental template strands and thereby characterization of individual homologues[30,31]. Specifically, in the presence of bromodeoxyuridine (BrdU) during the DNA replication, sister chromatids contain one original template strand and one newly synthesized, BrdU-incorporated strand. The template strand and its directionality is preserved during the cell division phase that helps to separate the individual homologues.

Because these sequencing methods provide long-range information on genomic structure across centromeres, they can be computationally assembled into chromosome-scale[32] and at low cost. However, these haplotypes contain many gaps, especially in larger repeat regions. This limitation has led to further advancements in

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 January 2021 doi:10.20944/preprints202101.0116.v1

computational approaches for haplotyping (Table 2) , such as the use of a hybrid approach that combines data from long-read and chromosome-scale sequencing technologies.

[H1] Reference-based haplotype reconstruction

When a reference genome is available, haplotype reconstruction of the target sample consists of identifying co-occurring alleles of paternal and maternal copies over variant sites from sequencing data aligned to the reference. The process of obtaining these haplotypes is known as haplotype phasing[2].

Traditionally, reference panels of more than 100,000 individuals (large-scale projects such as UK10K) are genotyped and used to assign, probabilistically, the most likely local phase of the target sample based on the underlying evolutionary model[33–35]. Although inexpensive, this technique requires large sample sizes of more than half a million for accurate phasing and does not apply to rare or *de novo* variants. Moreover, discovered linkages between variants do not necessarily translate to populations beyond the resource[19,36]. This technique is limited in its capacity to produce chromosome-scale haplotypes[19,36]. When genotypes of related individuals are available, haplotypes are inferred based on the Mendelian laws of inheritance, where homozygous sites on parents provide linkage information of allele variants on the offspring[33–35]. The limitation of this technique is to phase variants that are heterozygous in all individuals.

Phasing directly from sequencing reads, that is, the direct observation of two or more variants on a single molecule or in paired reads derived from the same molecule, overcomes the above limitations[37,38]. The process of obtaining haplotypes directly from long-read and chromosome-scale sequencing data of a single individual — as opposed to phasing from genotypes by population inference or genetic analysis of pedigrees — is known as molecular haplotyping[2,27,39]. Molecular haplotyping can produce chromosome-level phasing[40] that is highly accurate as determined by evaluation metrics (Switch error rates and Hamming error rates <1%). In molecular haplotyping, the key challenge is to disambiguate sequencing errors from true genetic variation.

[H2] Diploid phasing

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 January 2021 doi:10.20944/preprints202101.0116.v1

Reconstruction of haplotypes depends on how the heterozygous sites are connected on the chromosome-scale. If there are no reads that connect these sites, then the phasing is fragmented. Thus, the sites must be connected directly or indirectly via sequencing reads to achieve chromosome-scale phasing (Fig. 2). Long- and linked-read sequencing datasets, which span longer segments of heterozygous variants than short reads (Fig. 2), have improved the production of high-quality local phasing segments and the discovery of *de novo* and rare genomic variants.

The most widely used, state-of-the-art phasing methods are WhatsHap[41], HapCut2[42] and ProbHap[43], which generate considerably longer haplotype blocks than short reads, in the order of several megabases in length with a switch error rate of <0.5%. The core aim is to assign all reads to two haplotypes while minimizing the number of sequencing error corrections or flips, also known as the minimum error correction (MEC) problem and weighted minimum letter flip (WMLF) problem[44]. More specifically, the MEC formalism, which is the most widely used, is the process of finding the minimum cost of correcting the sequencing data to partition the read set into two homologues such that the alleles between any two reads in any partition match[41,45]. The MEC formulation is NP-hard[44,46]. In practice, this formulation is solved using computational techniques such as dynamic programming, probabilistic modelling, graph-based optimization and linear programming[47]. To scale these algorithms to human-sized genomes and beyond, a combination of greedy heuristics and dynamic programming is prominent[42,48].

However, a more challenging algorithmic task is genome-wide molecular phasing, a task that computes combinatorial solutions at chromosome-scale by using long-range sequencing technologies such as Hi-C and Strand-seq (Fig. 2). Computational tools used in practice, such as HapCut2[42] and StrandPhaseR[40], reduce the search space using greedy heuristics based on the MEC formulation. Remarkably, these tools generate haplotype blocks spanning full chromosomes[42]. However, they can typically phase only 50% of variants[42]. In this new era of advancements across technologies, hybrid algorithms that combine different data types at local and genome-wide scale are prominent. For example, WhatsHap[41] and HapCut2[42] both have local as well as chromosome-scale phasing modes. In addition, WhatsHap[41] can

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 January 2021 doi:10.20944/preprints202101.0116.v1

perform family-based phasing, which has been shown to give better results than single-individual approaches in terms of accuracy and phasing completeness[49]. The disadvantage is the unavailability of trio sequencing data for various species.

Hybrid approaches (combining long or 10xG reads with Strand-seq or Hi-C datasets[40]) for single individuals are leading the way into production-level efforts and provide competitive phasing performance at chromosome-scale with hamming error rates <1% and switch error rates <0.5% (Fig. 2). These methods have enabled impressive advances in the production of high-quality chromosome-scale phasing, for example, phasing Ashkenazi and Chinese human genomes[16,50] for a comprehensive SV callset.

Beyond the above bulk sequencing methods, single-cell phasing[51] has recently been used to study single-cell genomic heterogeneity. However, extremely low sequencing coverage (<0.05x per cell) has restricted its use in phasing of large multi-megabase segments in individual cells for genome-scale analysis. Recent single-cell phasing methods such as CHISEL[52], Satas et al.[53] and RCK[54] use probabilistic models at a single-cell level that have the advantage of haplotyping rare alleles, which can be used to determine local relationships in allele-specific somatic aberrations, but cannot phase all variants across the genome. Thus, in the near future, combining single-cell and bulk sequencing approaches for phasing will enable accurate and complete genome-wide characterization of genomic heterogeneity, including rare alleles and cancer genomes.

[H2] Polyploid phasing

In phasing diploid genomes, the haplotypes are complementary: given the genotype data, determining one haplotype sequence directly identifies the other. However, polyploidy is common in plant genomes, and in the case of a *k*-ploid sample, *k*–1 haplotypes need to be computed before the final haplotype can be inferred. For example, there are k! possibilities (instead of two in diploid) to connect a pair of SNPs in the polyploid. A higher number of haplotypes also requires a greater overall sequencing depth, resulting in a larger number of reads per genome to be processed. This

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 January 2021 doi:10.20944/preprints202101.0116.v1

additional complexity requires specialized, highly optimized algorithms to resolve polyploid phasing (Fig. 2).

To solve polyploid phasing problems, the maximum likelihood framework is a common algorithmic strategy. HapTree[38,55] uses the relative likelihood algorithm to identify k-ploidy phasing for first *n* SNPs, that is conditioned on previous *n*-1 SNPs. This approach lays the first theoretical foundation of polyploidy phasing problems. Few works have attempted to formulate the problem using approximate MEC formulations, for example, SDhaP[56] solves approximate MEC using semidefinite programming, and H-PoP[57] partitions the reads into haplotypes by solving a generalization of the MEC problem. However, there is an inherent problem in MEC based methods that it leads to inaccurate phasing as demonstrated by Motazedi *et al*[58].

To address these shortcomings, local phasing methods such as Ranbow[59] follows graph-based algorithms by leveraging allele co-occurrence in overlapping short reads to produce accurate polyploid phasing, but lacks in haplotype block N50 length. An alternative phasing approach, designed specifically for long-read sequencing, is WhatsHap-polyphase[60] (available as part of the diploid phasing tool WhatsHap) that produces accurate phasing (switch error rates <1% and hamming error rates <2%), and better N50 compared to short-read methods. A recently linked-read based method[61]: Hap++ and Hap10 produces slightly more accurate and comparable haplotype block N50 compared to WhatsHap-polyphase, at the cost of efficiency. However, these methods have limitations to produce chromosome-scale haplotypes.

Similar to diploid phasing, additional long-range information can allow for chromosome-scale haplotype reconstruction of polyploid genomes. For example, TriPoly[62] uses family information (parent-offspring trios) to infer haplotypes from either short- or long-read sequencing data. This results in larger haplotype blocks compared to other approaches, in particular in regions with low divergence between haplotypes.

While these methods represent an important step forward in polyploid phasing, new methods that leverage HiFi or alternatively linked-read data can potentially produce accurate and complete haplotyping results for complex repetitive polyploids. Further development of greedy heuristics to combine local and chromosome-scale sequencing data will enable chromosome-scale phasing in large polyploid genomes.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 January 2021 doi:10.20944/preprints202101.0116.v1

[H1] De novo haplotype assembly

De novo genome assembly exploits the overlaps between sequencing reads, without any bias towards reference sequence, and combines them into longer contiguous sequences (in the order of several tens of Mb), commonly referred to as contigs (Box 1). For diploid and polyploid genomes, most standard *de novo* assemblers collapse the haplotypes into a single consensus sequence (Box 1). Reconstructing every individual haplotype from sequencing data instead is known as *de novo* haplotype assembly (Fig. 3), and is even more challenging than consensus generation (*de novo* assembly) due to varying repetitive and heterozygosity rates, noisy sequencing data, chimeric reads, insufficient read length and non-uniform coverage.

In *de novo* haplotype assembly, the major challenge is finding the order of sequencing reads from massive erroneous datasets over the whole-genome level. The brute-force approach is to align all reads to all other reads, where the performance is directly proportional to square of the number of reads. In repetitive regions, finding alignments of reads is even more expensive. For systematic study, overlap-based[63] or de Bruijn graph[64] based techniques are used.

[H2] Diploid haplotype assembly

Algorithms for long-read sequencing are now able to produce megabase contigs for haplotypes and improve the availability of reference-quality genomes for humans and various other eukaryotic organisms[65–67]. This technique has been applied to assemble phased sequences of humans[65,68] (Table 1), diploid potato[69], zebra finch[70], cattle[70] and goat genomes[71]. Broadly, the bioinformatic approaches for diploid assembly fall into three classes: collapsed, semi-collapsed and uncollapsed (Fig. 3).

In collapsed diploid assembly, generic *de novo* assemblers (Box 1) are used to generate a consensus sequence. Subsequently, by using heterozygous SNP information from reads aligned to the consensus sequence, long-read and chromosome-scale sequencing reads are partitioned into haplotype-specific read sets, which are then separately assembled into haplotypes. This technique is used by tools such as DipAsm[68] or Porubsky et al.[72] , resulting in phased contigs of up to several tens of megabases and chromosome-scale phased scaffolds, with haplotype sizes of ~3 Gb

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 January 2021 doi:10.20944/preprints202101.0116.v1

each and overall base quality scores of >Q48. This technique works well for the human genome in regions of low heterozygosity, but fails in highly repetitive and high heterozygosity regions.

Alternatively, the widely used FALCON-unzip[66] method uses a semi-collapsed approach for diploid assembly from long noisy reads, where the initial assembly graph is generated using FALCON and a consensus sequence is generated. Similar to the collapsed approach, reads are partitioned into haplotype-specific sets using SNP information. Phased read information is then used to update the initial assembly graph, and phased contigs (size of about several tens of megabases with <Q48) are reported[73]. These phased contigs are then combined into scaffolds using phase information (>1 Mb) provided by ultra-long nanopore or Hi-C data, as employed by FALCON-Phase[70], producing a chromosome-scale diploid assembly. Similar to the collapsed approach, it works particularly well for human genomes when the heterozygosity rate is low, but fails in regions or genomes with high repeat and heterozygosity rates. However, the most promising uncollapsed approaches overcome these limitations by directly determining haplotype-specific overlaps in the overlap step of graph generation using SNP information from overlapping reads[74]. The core idea is to preserve heterozygosity and repeat information from various data types in the graph space. To achieve this, on every reference read, similar reads from the same haplotype and repeat are detected based on shared alleles at SNP sites and are clustered together. Standard tools use run-length encoding or base-level alignment[75] in the overlap step. Thus, a haplotype and repeat-aware overlap graph is generated with subsequent graph cleaning steps, finally reporting phased contigs.

The recent invention of PacBio HiFi technology has made the diploid assembly process, that includes ordering as well as the phasing in the assembly process, easier[74]. A whole generation of new algorithms based on uncollapsed approaches have become possible due to the availability of accurate long-read data and are implemented in tools such as HiFiasm (https://github.com/chhylp123/hifiasm), HiCanu[75], and SDip[74], producing contigs with lengths of several tens of Mb having base quality scores >Q50, but phased blocks of only a few hundreds of kb. In these systems, the field is moving towards accurate HiFi data using *k*-mer based strategies for haplotype-aware error

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 January 2021 doi:10.20944/preprints202101.0116.v1

correction of phased contigs, which can be completed in a few hours for human-scale genomes. Similar to semi-collapsed approaches, these phased contigs can be combined into phased scaffolds using long-range information to produce a chromosome-scale diploid assembly. For phased scaffolding, one of the largest challenges is the development of computational models that combine both phasing and scaffolding information together, an approach that has yet to be explored.

When trios are available, methods such as TrioCanu[76] search for k-mers from maternal and paternal haplotypes in reads that were sequenced from the child to produce haplotype-specific read sets and then assemble these separately. New methods (HiFiAsm+trio and WHdenovo[77]) use both trio and local phasing information from sequencing data, thus resulting in high-quality phased contigs. Although pedigree-based haplotype assembly allows for improved accuracy as compared to haplotype assembly of individuals, it requires sequencing of three individuals which limit its applications. Moving forward, substantial improvements to the uncollapsed approach using graphs and k-mers from local and chromosome-scale sequencing datasets in single individuals are expected to become routine for chromosome-scale diploid assembly within the next few years.

[H2] Polyploid haplotype assembly

Polyploid assembly is in principle an immediate extension of diploid assembly; however, an increase in the number of haplotypes inflates the search space dramatically. Using Illumina short-read sequencing reads, POLYTE[78] performs overlap graph-based *de novo* assembly for diploids and polyploids. Since short reads cannot span difficult-to-assemble regions such as long repeats or variant deserts, the haplotype-specific contigs produced by this algorithm remain relatively short, yet highly accurate. Alternatively, linked-read technologies have been used to obtain long, contiguous, polyploid genome assemblies[79]. Long-read based methods such as SDA[80] and SDip[74] have demonstrated their ability to assemble polyploid regions in human genomes several megabases in length (Fig. 3). With some further algorithm engineering, these methods can also be applied to entire polyploid genomes. Further algorithm development for chromosome-scale polyploidy haplotype assembly is

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 January 2021 doi:10.20944/preprints202101.0116.v1

required to exploit the latest sequencing technologies, such as HiFi sequencing, and its combination with other technologies.

[H1] Strain-resolved metagenome assembly

Chromosome-scale haplotype reconstruction is important in understanding long-range interactions, gene order and organization in metagenomes, as well as novel strains/species discovery[89]. The metagenome assembly is the computational process to reconstruct haplotypes from pooled sequencing to identify species or strains. This process is challenging due to multiple levels of variation within and across species[81,82], ranging from <1% to >5%, as well as relatively low per-strain sequencing depth across different datasets[83] that makes distinguishing sequencing errors and true variations hard[84]. Compared to two (and >2) haplotypes in diploids (and polyploids), the goal here is to produce several hundreds of haplotypes in the microbial community. The additional challenges are longer repeats and homologous regions between closely related strains (that is, intergenomic repeats) relative to sequencing read lengths.

The bioinformatic approaches for metagenome assembly are highly related to haplotype reconstruction in diploids and polyploids, as noted by Kolmogorov *et al.*[83] and Nicholls *et al*[84]. In practice, various diploid or polyploid haplotyping approaches are adapted to solve the strain-resolved metagenome assembly problem[83,85,86] and vice versa[78]. Here, we classify two classes of methods: species-level and strain-resolved metagenome assembly — species-level reconstruction aims at constructing a single (consensus) haplotype per species, while strain-resolved assembly aims at every strain of species. Each class of methods can further be distinguished into reference-based (database of species/strains) and *de novo* approaches — similar to diploids and polyploids approaches. The advantage of reference-based approaches is that they are efficient, but they often lead to biases towards the database(s) being used, for example, due to incompleteness of reference databases, as many microbes on Earth remain uncharacterized[87]. This type of reference bias is even more pronounced than the reference bias observed in diploid and polyploid assembly, hence *de novo* algorithms are an essential component of any complete, unbiased analysis of metagenomes.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 January 2021 doi:10.20944/preprints202101.0116.v1

Here, we present a general workflow for metagenome assembly that consists of several steps[88–91] (Fig 4b): (1) *de novo* metagenome assembly to produce contigs or scaffolds; (2) contig binning per genome, either de novo or reference-guided; (3) mapping reads back to individual bins and reassembling each bin; (4) curation of the resulting assembly per bin. In theory, each of these steps can be performed at the species-level or at the strain-level, depending on the goals of the study.

[H2] Short-read metagenome assembly

The commonly used data structures for metagenome assembly are de Bruijn and overlap graphs with special tuning of parameters related to sequencing depth, variations and errors. For example, IDBA-UD[94] is the first metagenome assembly method based on de Bruijn graphs, and LSA[95], a method that uses *k*-mers to identify (partial) bacterial strains in short-read sequencing data with relative abundances as low as 0.00001%. Other de novo approach tools are MEGAHIT[96] and metaSPAdes[97], but these can only produce species-level assemblies. Alternatively, several approaches are based on single-nucleotide variants (SNVs), which are identified using metagenome assemblies or reference databases, or entirely *de novo*---see REF[98] for a detailed review. The major limitation of such approaches is that structural variants are completely ignored. Available methods for SNV-based metagenome assembly with strain-resolution include ConStrains[99] and StrainFinder[88], both of which can trace strain identities across multiple samples (a longitudinal time series). Recently, a Bayesian model for local haplotype reconstruction was proposed in a promising approach called Gretel[87], which is based on a new data structure designed to efficiently store variation across sequencing reads. All of these methods (ConStrains, PathFinder, Gretel) aim at strain-level sensitivity in step 1 (Fig 4b). Another class of methods achieve strain-level sensitivity in step 3 while relying on species-level sensitivity described in steps 1 and 2 (Fig 4b). DESMAN[100] is one such method, which leverages base haplotype frequencies in a Bayesian model. Finally, if strain-level assembly is not achieved in steps 1-3, further curation in step 4 can help to identify intra-species variation[101] (Fig 4b).

These short-read methods take an important step in strain-level metagenome assembly field and are widely used in studying the human microbiome, health and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 January 2021 doi:10.20944/preprints202101.0116.v1

disease[102,103] as well as the biodiversity of marine ecosystems[100]. These methods establish the first step towards producing chromosome-scale metagenome assembly.

[H2] Hybrid metagenome assembly

Local and chromosome-scale sequencing is essential in achieving chromosome-scale, strain-resolved metagenome assemblies. Recently, a hybrid metagenomic assembly approach (OPERA-MS) was proposed, that combines short-read contig assembly with long-read scaffolding and binning to obtain high-quality, strain-resolved metagenomes[107]. OPERA-MS provides an order-of-magnitude improvement in contiguity compared to short-read metagenomic assemblers and a 200% increase compared to generic long-read assemblers. As little as 7x haplotype coverage with long reads was sufficient to obtain megabase N50 genomes[107]. Alternatively, the first long-read metagenome assembler (MetaFlye[85]) proposes the use of local k-mer distributions to identify species of low abundance. MetaFlye can assemble haplotypes with as little as 10x per-haplotype coverage[85], though the extent to which it can distinguish between closely related strains remains to be evaluated. Another approach by Anoton et al.[108] uses long-read assembly (with MetaFlye), followed by assembly curation using short- and long-read data. Yet another approach, MetaMaps[109], offers strain-level long-read binning, but this requires a reference database and therefore complicates discovery of new haplotypes (Fig 4b).

Alternatively, the combination of Hi-C and shotgun sequencing enables chromosome-scale, strain-resolved metagenome assembly through improved clustering of metagenome-assembled contigs at strain level, as well as linking of plasmid sequences to the chromosomes of their hosts[105–107]. Such an approach has recently been used to leverage structural information obtained from Hi-C data of the human gut microbiome to perform strain-level assembly and enable tracking of microbial evolution over time[108].

For complex repetitive metagenomes, HiFi reads, in combination with Hi-C, have the ability to become the strategy of choice to produce complete, strain-level resolved metagenome assemblies in the near future.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 January 2021 doi:10.20944/preprints202101.0116.v1

[H1] Remaining challenges and perspectives

[H2] Repetitive regions

Haplotype reconstruction remains challenging in multi-megabase complex repetitive regions. Despite considerable time and effort, the current version of the human reference genome either contains gaps or is collapsed in these regions without haplotype-level resolution. These regions include tandem repeats[114], segmental duplications[115,116], sex chromosomes (containing complex heterochromatin repeat structures)[117,118], the mitochondrial genome[119], pseudo-autosomal regions[120] (PARs), centromeres (or pericentromeric regions)[12], ribosomal DNAs[121] (or acrocentric regions), and subtelomeric regions[122]. For example, the human genome includes complex satellite arrays of repeats in centromeres. α-satellite DNA contains ~171-bp tandem repeats that are organized into higher-order repeats (HORs), with a single repeat structure reiterated over hundreds or thousands of times with high (>99%) sequence conservation[123]. Some human chromosomes comprise ~3,200 repeats of ~2 kb HORs and ~1,100 repeats of a 1.8 kb HOR unit[124]. The centromere assembly produced by state-of-the-art tools (centroFlye[125], HiCanu[126]) using HiFi and ultra-long nanopore reads is haploid[127]. Humans are diploid and should produce two haplotype sequences in centromeres; however, there are no algorithms, technologies or tools to achieve this goal currently.

Recent developments in long, accurate long reads (Hifi) as well as ultra-long nanopore reads could pave the way for new advancements in finishing centromeric and other highly repetitive regions in humans and polyploids. Computationally, Hifi reads can be decomposed into monomers[128] that are represented in the graph, where monomers are nodes and edges represent the adjacencies of node sequences from reads. In this process, the haplotype variation is also considered in the monomers that can result in a haplotype-aware graph. Through this graph, the ultra-long nanopore reads are anchored to potentially find ordering between repeating units and disentangle the graph. On a complex centromeric region involving >2000 repeat units, the *in situ* information such as chromosome visualization[129] can further be helpful to order the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 January 2021 doi:10.20944/preprints202101.0116.v1

repeating units. Further innovations in user-oriented computational tools may enable exploration of high-resolution haplotypes in these complex centromeric regions.

[H2] Scale

Developments are required to scale haplotype reconstruction efforts to overcome current limitations and enable routine application to more than hundreds of genomes at a time. Such developments require innovations in technologies that are cheaper and easy to use than long-read, HiFi and Hi-C sequencing. Alternatively, a further reduction in sequencing costs of existing technologies will be required to scale up efforts.

With an exponential growth in datasets, the real challenge will be to store and access haplotyping data in an efficient way, which can potentially be achieved by applying massive parallelism (detailed reviews in [130,131]). In addition, cloud-based strategies will be required for storing, accessing and sharing data (for example, https://vgp.github.io/genomeark/). Building a collaborative haplotyping platform that can serve as a repository for data, computational tools and enable exchange of ideas for the scientific community may help to usher in a new era of biological discoveries.

Further integration of datasets using scalable bioinformatics approaches will be important. Innovative algorithm engineering (for example, using sequence sketches instead of full sequences has been shown to vastly reduce storage and memory requirements[132]) could enable production-level integration of datasets for haplotype reconstruction. Beyond engineering efforts, combining reference-based and *de novo* approaches will improve scalability. More specifically, genomes that are similar to known samples can be reconstructed efficiently using reference-based approaches, thus reducing de novo efforts to the remaining highly divergent genomes.

[H2] Validation, benchmarking and annotation

For the final haplotype assemblies of diploid genomes, many high-quality benchmarks are available, and validation is done with standardized evaluation metrics as a standard practice for non-repetitive regions. This is not the case for polyploids, tumours and complex repetitive regions in diploids. Innovations in algorithms (beyond *k*-mer approaches) that improve the capability for assessment and biological validation could

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 January 2021 doi:10.20944/preprints202101.0116.v1

benefit from a public collection of high-quality benchmarks, for example in the form of a community-driven assessment initiative similar to the Critical Assessment of Metagenome Interpretation[83] (CAMI), Assemblathon[133,134] and Genome Assembly Gold-standard Evaluations[135] (GAGE). As the field advances to produce high-quality chromosome-scale phased sequences, the next critical step will be in the development of new gene annotation tools[136] to enable more precise downstream analyses in the coming decade.

[H2] Visualization

Another challenge is the visualization of large-scale haplotyping raw sequencing datasets and haplotype sequences from multiple species. The combination of long haplotype sequences and divergence across or within genomes, and the large diversity of haplotyping data types pose numerous visualization challenges. While a number of tools exist (reviewed in Refs. [137,138]), none can be used to visualize large-scale phased sequences. New visualization techniques will be required that enable abstractions or reductions in data dimensions from multi-scale, multiple data measurements, binary encoding of variations and divergence across haplotypes for visual maps, and discovery of informative patterns in the haplotyping data. Interactive visualization or animation in the chromosome-scale coordinate system can be useful.

[H1] Conclusions

Chromosome-scale haplotype reconstruction has yielded new insights into the genetic underpinnings of disease pathogenesis, evolution and comparative biology. However, due to the limitations of sequencing reads to cover genomic repeats, chromosome-scale haplotype reconstruction using a single technology is not possible. Thus, a combination of long-read (HiFi and ultra-long ONT) and chromosome-scale sequencing (Hi-C) datasets, using integrative algorithms, has become a common strategy to produce haplotypes in diploids, but not polyploids yet.

Improvements in fragment lengths, and combining complementary technologies through innovative algorithms (graphs, *k*-mers and data-driven) will be state-of-the-art to reconstruct high-quality haplotypes with fewer gaps in the near future. Both fragment

 doi:10.20944/preprints202101.0116.v1

accuracy and length — a few megabases size with accuracy of >98% — could be important to finish haplotypes. Major reductions in sequencing and computing costs will be critical to scale efforts to thousands of genomes at a time. In the next decade, algorithmic and technological advances, paired with the incorporation of haplotypes with disparate layers of biological information, could mark a new era of gapless end-to-end haplotypes and further our understanding of complex biological phenomena.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 January 2021 doi:10.20944/preprints202101.0116.v1

Table 1 | Third-generation sequencing initiatives and reference data sets

Initiatives	# samples/#haplotypes	Technologies	Links
Genome in a Bottle[139,140] (GIAB)	2 trios and 1 sample, 6 haplotypes	PacBio, ONT, Illumina, BioNano, Strand-seq, 10xG	ftp://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/
Human Genome Structural Variation Consortium[16] (HGSVC)	>3 trios, >6 haplotypes	PacBio, Illumina, BioNano, Hi-C, Strand-seq, 10xG	https://www.internationalgenome.org/data
Vertebrate Genome Project (VGP; facilitated by Genome 10K), Darwin Tree of Life Project	>100, ongoing haplotyping efforts	10xG, PacBio, Hi-C	https://vgp.github.io/genomeark/
Human Pangenome Project	>10, >20 haplotypes	PacBio, ONT, Hi-C	https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=HPRC/
Earth Biogenome Project (facilitated by Genome 10K)	>10, ongoing haplotyping efforts	PacBio, Hi-C	https://www.earthbiogenome.org/publications
The DNA Zoo project	>10, ongoing haplotyping efforts	Hi-C and WGS	https://www.dnazoo.org/
Japanese Reference Project[141] (1KJPN)	>1, >2 haplotypes	PacBio, Illumina	https://jrg.megabank.tohoku.ac.jp/en
CHM1, CHM13 [127], HX1 [142], PGP-1 [51], AK1[66]	Individual samples, two haplotypes each (except CHM1 and CHM13)	PacBio, ONT, BioNano, Hi-C, Illumina	n/a

Table 2 | Methods and computational tools for haplotype reconstruction

Approach	Tools	Data	Advantages	Disadvantages
Reference-based phasing				

Molecular haplotyping	WhatsHap[41], HapCut2[42] and ProbHap[43]	Long reads such as PacBio, Hi-C of individual	Can phase *de novo* and rare variants	Limitations in complex regions such as centromeres, HLA, etc.
Single-cell phasing	CHISEL[52], Satas et al.[53], RCK[54]	Single-cell short-read	High precision at single-cell, detection of rare alleles	Engineering tricks required to scale to >million cells
Polyploid phasing	HapTree[38], Hap10[61], WhatsHap-polyphase[60], H-PoP[57]	Local phasing	Can phase *de novo* and rare variants	Limitations in repetitive regions and not optimized for ploidy>5
De novo assembly				
Diploid assembly	Falcon Unzip[66], Falcon phase[70]	Long reads and Hi-C of individual	Local phased contigs	No chromosome-scale assembly and computationally expensive
	DipAsm[68], Porubsky et al.[72]	Long reads and Hi-C of individual	Chromosome-scale diploid assembly	Collapsed assembly not suitable for repetitive regions
	HiFiAsm, HiCanu[75], SDip[74]	HiFi reads of individual	High consensus accuracy and continuity	No chromosome-scale assembly
	TrioCanu[76], HiFiAsm+trio, WHdenovo[77]	Long reads of trios	Local phased contigs	Require family information
Polyploid assembly	SDA[80], SDip[74]	Long reads of individual	Local phased contigs	Need to be optimized for whole genomes
	POLYTE[78]	Illumina short reads	Local phased contigs	Doesn't scale well to whole genomes
Strain-resolved metagenome assembly				
De novo (re-)assembly	IDBA-UD[95], DESMAN[101]	Metagenome short reads	No prior knowledge required	Low sensitivity: rare haplotypes can remain undetected
	OPERA-MS[107]	Metagenome using	High continuity	Computationally

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 January 2021 doi:10.20944/preprints202101.0116.v1

		short and long reads		expensive
SNV-based assembly	ConStrains[100], StrainFinder[87], Gretel[86]	Metagenome short reads	Computational efficiency	Assembly accuracy depends on variant calling
Read binning	MetaMaps[109]	Metagenome long reads	Computational efficiency	Accuracy depends on database
Contig binning	ProxiMeta[111], bin3C[112]	Metagenome short reads and Hi-C	Reference-free, ability to link plasmids to host chromosome	Multiple technologies necessary (Hi-C + shotgun sequencing)

Figure 1 | Third-generation sequencing technologies and their characteristics (read length, error rate and scale of information). The read length and scale of information (local versus chromosome-scale) together determine the haplotype range that can be achieved; moving down the schematic this range increases (orange arrow). Sequencing costs per sample increase moving from short-read sequencing down to nanopore sequencing, then decrease again for BioNano and Hi-C (yellow arrows). Similarly, read length and error rate first increase moving down to nanopore sequencing, then decrease again for BioNano and Hi-C (green arrows).

Figure 2. Molecular haplotyping techniques in alignment-based phasing. Individual haplotypes are derived directly from sequencing data of the target sample based on read alignments to the reference genome. Local and chromosome-scale haplotype phasing make use of short- and chromosome-scale sequencing data, respectively; hybrid haplotype phasing combines the two data types.

Figure 3. Haplotype-aware *de novo* assembly. Collapsed assembly approaches identify sequence variants on a consensus assembly (Box 1) and subsequently phase these variants into haplotypes using chromosome-scale data (Hi-C or Strand-seq). Semi-collapsed approaches follow a similar approach, but after phasing, variants in the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 January 2021 doi:10.20944/preprints202101.0116.v1

initial assembly graph are updated and final contigs are produced based on this updated graph. Uncollapsed approaches directly determine haplotype-specific overlaps in local sequencing reads by retaining SNPs and repeat variation in all possible overlaps and construct haplotypes based on the selected overlaps.

Figure 4. Strain-resolved metagenome assembly. a | Given a pooled sequencing sample, the goal of strain-resolved metagenome assembly is to reconstruct all individual microbial strains. **b** | A typical workflow consists of four steps: *de novo* assembly, contig binning, bin-wise re-assembly, and assembly curation. Each step can be performed at the species-level or at the strain-level, as illustrated in the left and middle column, respectively. Some workflows skip the initial *de novo* assembly step and perform strain-resolved binning directly on the sequencing reads, which can be reference-guided (right column).

BOX 1 | Consensus *de novo* assembly

Generic *de novo* assembly approaches produce a consensus assembly, meaning that the different chromosomal copies that make up the genome are collapsed onto a single consensus sequence. The main steps in a standard genome assembly workflow are sequence graph construction, error correction, contig formation, scaffolding, and polishing of the assembled sequences[148]. The most widely used assembly software is Canu[149], FALCON[73], Flye[150], wtdbg2[151] and shasta[152]; we refer the reader to a review by Sedlazeck *et al.*[153] for a literature survey on genome assembly using the latest technologies. Of particular interest to haplotype assembly are assemblers specifically designed for PacBio HiFi datasets, which, for the first time, improved the per-base quality of assemblies dramatically[154] and reduced the need for computational intensity of the error correction step.

After contig construction, the next step is to create scaffolds by ordering and orienting contigs along the chromosomes using chromosome-scale information sources, such as Hi-C data. Scaffolding with the chromosome-scale data types has resulted in chromosome-scale consensus assemblies of human genomes.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 January 2021 doi:10.20944/preprints202101.0116.v1

Due to sequencing errors, the reads often undergo error correction before contigs are formed; this is particularly relevant when using error-prone long-read sequencing technologies. Despite the error correction process, contigs and scaffolds may still be erroneous and thus another round of error correction is performed (now referred to as polishing) using tools such as Racon [155].

Consensus *de novo* assembly approaches can be applied to diploids as well as polyploids, but it is important to realize that all haplotype information is ignored. Nevertheless, the resulting consensus assembly can be used as a first step in the haplotype reconstruction process; this is how so-called collapsed methods operate (see "Diploid haplotype assembly").

Glossary

Allele-specific expression

The phenomenon of unbalanced transcript abundances originating from the different allelic copies.

Barcoding

Labelling reads with barcode sequences to identify fragments from the same partition.

Base-level alignment

Position-wise alignment of nucleotides in a pair of sequences.

Chromosome-scale haplotype

Nucleotide sequence spanning a full chromosome for a given homologous copy across centromeres.

Compound heterozygosity

The phenomenon where a combination of recessive alleles for a given locus harboring different mutations together can cause genetic disease.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 January 2021 doi:10.20944/preprints202101.0116.v1

Pan-genome graph

A data structure that contains sequences shared across multiple genomes of species. The differences in genome sequences are also stored.

Diploid

A genome containing two complete sets of chromosomes, one from each parent.

Aneuploidy

Normal human cells contain two chromosomes. Aneuploidy is the phenomenon of increase or decrease in chromosomes in cancer cells compared to normal cells.

Dynamic programming

A mathematical optimization approach where the problem is recursively divided into subproblems whose solutions build towards a global, optimal solution.

Epistatic interactions

Interaction between two different genes contributing to a single phenotype.

Optical mapping

A technique for constructing ordered restriction maps of the whole genome, called optical maps, by locating restriction enzyme sites on an unknown genomic sequence.

Genome partitioning

Using microfluidics to physically separate genomic sequences.

Greedy heuristic

Solving an optimization problem by finding a locally optimal solution.

Heterozygosity rate

Rate of mutations (differences) between haplotypes.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 January 2021 doi:10.20944/preprints202101.0116.v1

k-mer distribution

Frequency distribution of substrings of length *k* from an original sequence

Long-read (or local) sequencing

Reading genomic fragments (reads) that span up to several hundreds of kilobases of the genome.

Long-range promoter-enhancer interaction

Transcriptional enhancers interacting with their target-gene promoters over a considerable genomic distance, affecting gene expression.

Chromosome-scale sequencing

Reading genomic fragments spanning across centromeres, thereby providing information to connect p and q arms over the entire genome. In genomes with centromeres, read fragments spanning at the whole genome level.

Mate-pair sequencing

Generating paired-end (short) reads with particularly long inserts to span a large genomic region.

Haplotype blocks

A genomic regions on the chromosomes that are phased together in a segments

Scaffolds

Scaffold are the sequences produced by ordering the contiguous sequences with their correct orientation.

Variant deserts

Genomic regions with a fewer variants compared to an average

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 January 2021 doi:10.20944/preprints202101.0116.v1

NP hard

The complexity class of decision problems that are intrinsically harder than those that can be solved by a nondeterministic Turing machine in polynomial time.

Sequence sketches

Sketching of genomic sequences is the process of indexing and hashing the data for faster direct access and efficient memory usage

Hamming distance

It is an evaluation metric to compare the binary strings. This metric is used to evaluate the long-range phasing on the chromosome-scale level

Switch error rate

Number of switches between true and alternative haplotypes, relative to the number of variant positions

Base error rate

Number of erroneous bases relative to total assembly length (can evaluate mismatch errors and indel errors separately or jointly)

Phased contig

Contiguous nucleotide sequences that represent a subsequence of a haplotype.

Phased scaffold

Haplotype sequence linking phased contiguous sequences (contigs) originating from the same haplotype, separated by gaps of known length.

Polyploid

A genome containing more than two sets of chromosomes; this is common in plant species.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 January 2021 doi:10.20944/preprints202101.0116.v1

Metagenome

The collection of genomes of many species as well as their strains present in an environmental sample.

Run-length encoding

A form of lossless data compression in which each repetitive sequence is stored as a single repetitive element along with its number of consecutive occurrences, rather than the whole repetitive sequence.

Variant calling

Variant calling is the process of finding genomic variation (mutations) from sequencing data aligned to the reference genome.

References

1. Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of population structure using dense haplotype data. *PLoS Genet.* **8**, e1002453 (2012).
2. Glusman, G., Cox, H. C. & Roach, J. C. Whole-genome haplotyping approaches and genomic medicine. *Genome Med.* **6**, 73 (2014).
3. Mantere, T., Kersten, S. & Hoischen, A. Long-Read Sequencing Emerging in Medical Genetics. *Front. Genet.* **10**, 426 (2019).
4. Tewhey, R., Bansal, V., Torkamani, A., Topol, E. J. & Schork, N. J. The importance of phase information for human genomics. *Nat. Rev. Genet.* **12**, 215–223 (2011).
5. Gao, Y. *et al.* The haplotypes of various TNF related genes associated with scleritis in Chinese Han. *Human Genomics* vol. 14 (2020).
6. Sirén, J., Garrison, E., Novak, A. M., Paten, B. & Durbin, R. Haplotype-aware graph indexes. *Bioinformatics* **36**, 400–407 (2019).
7. Garrison, E. *et al.* Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat. Biotechnol.* **36**, 875–879 (2018).
8. Sherman, R. M. & Salzberg, S. L. Pan-genomics in the human genome era. *Nat. Rev. Genet.* (2020) doi:10.1038/s41576-020-0210-7.
9. Wang, X., Cairns, M. J. & Yan, J. Super-enhancers in transcriptional regulation and genome organization. *Nucleic Acids Res.* **47**, 11481–11496 (2019).
10. Villar, D., Frost, S., Deloukas, P. & Tinker, A. The contribution of non-coding regulatory elements to cardiovascular disease. *Open Biol.* **10**, 200088 (2020).
11. Loh, P.-R. *et al.* Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. *Nature* **559**, 350–355 (2018).

12. Hartley, G. & O'Neill, R. J. Centromere Repeats: Hidden Gems of the Genome. *Genes* **10**, (2019).
13. Dávila-Ramos, S. *et al.* A Review on Viral Metagenomics in Extreme Environments. *Front. Microbiol.* **10**, 2403 (2019).
14. Farci, P. *et al.* Early changes in hepatitis C viral quasispecies during interferon therapy predict the therapeutic outcome. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 3081–3086 (2002).
15. Vignuzzi, M., Stone, J. K., Arnold, J. J., Cameron, C. E. & Andino, R. Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. *Nature* **439**, 344–348 (2006).
16. Chaisson, M. J. P. *et al.* Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.* **10**, 1784 (2019).
17. Scientists, G. 10k C. of & Genome 10K Community of Scientists. Genome 10K: A Proposal to Obtain Whole-Genome Sequence for 10 000 Vertebrate Species. *Journal of Heredity* vol. 100 659–674 (2009).
18. Wadapurkar, R. M. & Vyas, R. Computational analysis of next generation sequencing data and its applications in clinical oncology. *Informatics in Medicine Unlocked* vol. 11 75–82 (2018).
19. Snyder, M. W., Adey, A., Kitzman, J. O. & Shendure, J. Haplotype-resolved genome sequencing: experimental methods and applications. *Nat. Rev. Genet.* **16**, 344–358 (2015).
20. Jarvie, T. Next generation sequencing technologies. *Drug Discovery Today: Technologies* vol. 2 255–260 (2005).

21. Th, A., Attia, T. H. & Saeed, M. A. Next Generation Sequencing Technologies: A Short Review. *Journal of Next Generation Sequencing & Applications* vol. 01 (2015).
22. Stark, R., Grzelak, M. & Hadfield, J. RNA sequencing: the teenage years. *Nat. Rev. Genet.* **20**, 631–656 (2019).
23. Logsdon, G. A., Vollger, M. R. & Eichler, E. E. Long-read human genome sequencing and its applications. *Nat. Rev. Genet.* (2020) doi:10.1038/s41576-020-0236-x.
24. Janitz, K. & Janitz, M. Moving Towards Third-Generation Sequencing Technologies. *Tag-Based Next Generation Sequencing* 323–336 (2012) doi:10.1002/9783527644582.ch20.
25. Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M. & Jaffe, D. B. Direct determination of diploid genome sequences. *Genome Res.* **27**, 757–767 (2017).
26. Wenger, A. M. *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162 (2019).
27. Selvaraj, S., R Dixon, J., Bansal, V. & Ren, B. Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nat. Biotechnol.* **31**, 1111–1118 (2013).
28. Li, M. *et al.* Towards a More Accurate Error Model for BioNano Optical Maps. *Bioinformatics Research and Applications* 67–79 (2016) doi:10.1007/978-3-319-38782-6_6.
29. Weissensteiner, M. H. *et al.* Combination of short-read, long-read, and optical

mapping assemblies reveals large-scale tandem repeat arrays with population genetic implications. *Genome Res.* **27**, 697–708 (2017).

30. Falconer, E. *et al.* DNA template strand sequencing of single-cells maps genomic rearrangements at high resolution. *Nat. Methods* **9**, 1107–1112 (2012).
31. Sanders, A. D., Falconer, E., Hills, M., Spierings, D. C. J. & Lansdorp, P. M. Single-cell template strand sequencing by Strand-seq enables the characterization of individual homologs. *Nat. Protoc.* **12**, 1151–1176 (2017).
32. Porubský, D. *et al.* Direct chromosome-length haplotyping by single-cell sequencing. *Genome Research* vol. 26 1565–1574 (2016).
33. Loh, P.-R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **48**, 1443–1448 (2016).
34. O'Connell, J. *et al.* Haplotype estimation for biobank-scale data sets. *Nat. Genet.* **48**, 817–820 (2016).
35. Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81**, 1084–1097 (2007).
36. Browning, S. R. & Browning, B. L. Haplotype phasing: existing methods and new developments. *Nat. Rev. Genet.* **12**, 703–714 (2011).
37. Fan, H. C., Wang, J., Potanina, A. & Quake, S. R. Whole-genome molecular haplotyping of single cells. *Nat. Biotechnol.* **29**, 51–57 (2011).
38. Berger, E., Yorukoglu, D., Peng, J. & Berger, B. HapTree: a novel Bayesian framework for single individual polyplotyping using NGS data. *PLoS Comput. Biol.* **10**, e1003502 (2014).

39. Zhang, X., Wu, R., Wang, Y., Yu, J. & Tang, H. Unzipping haplotypes in diploid and polyploid genomes. *Comput. Struct. Biotechnol. J.* **18**, 66–72 (2020).
40. Porubsky, D. *et al.* Dense and accurate whole-chromosome haplotyping of individual genomes. *Nature Communications* vol. 8 (2017).
41. Patterson, M. *et al.* WhatsHap: Weighted Haplotype Assembly for Future-Generation Sequencing Reads. *J. Comput. Biol.* **22**, 498–509 (2015).
42. Edge, P., Bafna, V. & Bansal, V. HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Res.* **27**, 801–812 (2017).
43. Kuleshov, V. Probabilistic single-individual haplotyping. *Bioinformatics* **30**, i379–85 (2014).
44. Lippert, R., Schwartz, R., Lancia, G. & Istrail, S. Algorithmic strategies for the single nucleotide polymorphism haplotype assembly problem. *Brief. Bioinform.* **3**, 23–31 (2002).
45. Lancia, G., Bafna, V., Istrail, S., Lippert, R. & Schwartz, R. SNPs Problems, Complexity, and Algorithms. *Algorithms — ESA 2001* 182–193 (2001) doi:10.1007/3-540-44676-1_15.
46. Garg, S. & Mömke, T. A QPTAS for Gapless MEC. in *26th Annual European Symposium on Algorithms (ESA 2018)* 14 (Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 2018).
47. Klau, G. W. & Marschall, T. A Guided Tour to Computational Haplotyping. *Unveiling Dynamics and Complexity* 50–63 (2017) doi:10.1007/978-3-319-58741-7_6.
48. Martin, M. *et al.* WhatsHap: fast and accurate read-based phasing. doi:10.1101/085050.

49. Garg, S., Martin, M. & Marschall, T. Read-based phasing of related individuals. *Bioinformatics* **32**, i234–i242 (2016).

50. Zook, J. M. *et al.* A robust benchmark for detection of germline large deletions and insertions. *Nat. Biotechnol.* **38**, 1347–1355 (2020).

51. Adey, A. C. Haplotype resolution at the single-cell level. *Proceedings of the National Academy of Sciences of the United States of America* vol. 114 12362–12364 (2017).

52. Zaccaria, S. & Raphael, B. J. Characterizing the allele- and haplotype-specific copy number landscape of cancer genomes at single-cell resolution with CHISEL. *bioRxiv* 837195 (2019) doi:10.1101/837195.

53. Satas, G. & Raphael, B. J. Haplotype phasing in single-cell DNA-sequencing data. *Bioinformatics* **34**, i211–i217 (2018).

54. Aganezov, S. & Raphael, B. J. Reconstruction of clone- and haplotype-specific cancer genome karyotypes from bulk tumor samples. *bioRxiv* 560839 (2019) doi:10.1101/560839.

55. Berger, E., Yorukoglu, D. & Berger, B. HapTree-X: An Integrative Bayesian Framework for Haplotype Reconstruction from Transcriptome and Genome Sequencing Data. *Res. Comput. Mol. Biol.* **9029**, 28–29 (2015).

56. Das, S. & Vikalo, H. SDhaP: haplotype assembly for diploids and polyploids via semi-definite programming. *BMC Genomics* **16**, 260 (2015).

57. Xie, M., Wu, Q., Wang, J. & Jiang, T. H-PoP and H-PoPG: heuristic partitioning algorithms for single individual haplotyping of polyploids. *Bioinformatics* **32**, 3735–3744 (2016).

58. Motazedi, E., Finkers, R., Maliepaard, C. & de Ridder, D. Exploiting next-generation sequencing to solve the haplotyping puzzle in polyploids: a simulation study. *Brief. Bioinform.* **19**, 387–403 (2018).
59. Moeinzadeh, M. De novo and haplotype assembly of polyploid genomes. (2019).
60. Schrinner, S. D. *et al.* Haplotype Threading: Accurate Polyploid Phasing from Long Reads. *bioRxiv* 2020.02.04.933523 (2020) doi:10.1101/2020.02.04.933523.
61. Majidian, S., Kahaei, M. H. & de Ridder, D. Hap10: reconstructing accurate and long polyploid haplotypes using linked reads. doi:10.1101/2020.01.08.899013.
62. Motazedi, E. *et al.* TriPoly: haplotype estimation for polyploids using sequencing data of related individuals. *Bioinformatics* **34**, 3864–3872 (2018).
63. Myers, E. W. The fragment assembly string graph. *Bioinformatics* **21 Suppl 2**, ii79–85 (2005).
64. Compeau, P. E. C., Pevzner, P. A. & Tesler, G. How to apply de Bruijn graphs to genome assembly. *Nature Biotechnology* vol. 29 987–991 (2011).
65. Cao, H. *et al.* De novo assembly of a haplotype-resolved human genome. *Nat. Biotechnol.* **33**, 617–622 (2015).
66. Seo, J.-S. *et al.* De novo assembly and phasing of a Korean human genome. *Nature* **538**, 243–247 (2016).
67. Gordon, D. *et al.* Long-read sequence assembly of the gorilla genome. *Science* **352**, aae0344 (2016).
68. Garg, S. *et al.* Efficient chromosome-scale haplotype-resolved assembly of human genomes. *bioRxiv* 810341 (2019) doi:10.1101/810341.
69. Zhou, Q. *et al.* Haplotype-resolved genome analyses of a heterozygous diploid

potato. *Nat. Genet.* **52**, 1018–1023 (2020).

70. Kronenberg, Z. N. *et al.* FALCON-Phase: Integrating PacBio and Hi-C data for phased diploid genomes. *bioRxiv* 327064 (2018) doi:10.1101/327064.
71. Low, W. Y. *et al.* Chromosome-level assembly of the water buffalo genome surpasses human and goat genomes in sequence contiguity. *Nat. Commun.* **10**, 260 (2019).
72. Porubsky, D. *et al.* A fully phased accurate assembly of an individual human genome. doi:10.1101/855049.
73. Chin, C.-S. *et al.* Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* **13**, 1050–1054 (2016).
74. Heller, D., Vingron, M., Church, G., Li, H. & Garg, S. SDip: A novel graph-based approach to haplotype-aware assembly based structural variant calling in targeted segmental duplications sequencing. *bioRxiv* 2020.02.25.964445 (2020) doi:10.1101/2020.02.25.964445.
75. Nurk, S. *et al.* HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. doi:10.1101/2020.03.14.992248.
76. Koren, S. *et al.* De novo assembly of haplotype-resolved genomes with trio binning. *Nat. Biotechnol.* (2018) doi:10.1038/nbt.4277.
77. Garg, S., Aach, J., Li, H., Durbin, R. & Church, G. A haplotype-aware de novo assembly of related individuals using pedigree sequence graph. *Bioinformatics* (2019) doi:10.1093/bioinformatics/btz942.
78. Baaijens, J. A. & Schönhuth, A. Overlap graph-based generation of haplotigs for diploids and polyploids. *Bioinformatics* **35**, 4281–4289 (2019).

79. Ott, A. *et al.* Linked read technology for assembling large complex and polyploid genomes. *BMC Genomics* **19**, 651 (2018).
80. Vollger, M. R. *et al.* Long-read sequence and assembly of segmental duplications. *Nat. Methods* **16**, 88–94 (2019).
81. Lo, I. *et al.* Strain-resolved community proteomics reveals recombining genomes of acidophilic bacteria. *Nature* **446**, 537–541 (2007).
82. Simmons, S. L. *et al.* Population genomic analysis of strain variation in Leptospirillum group II bacteria involved in acid mine drainage formation. *PLoS Biol.* **6**, e177 (2008).
83. Sczyrba, A. *et al.* Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software. *Nat. Methods* **14**, 1063–1071 (2017).
84. King, D. J. *et al.* A Systematic Evaluation of High-Throughput Sequencing Approaches to Identify Low-Frequency Single Nucleotide Variants in Viral Populations. *Viruses* **12**, (2020).
85. Kolmogorov, M. *et al.* metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat. Methods* **17**, 1103–1110 (2020).
86. Nicholls, S. M. *et al.* Recovery of gene haplotypes from a metagenome. doi:10.1101/223404.
87. Smillie, C. S. *et al.* Strain Tracking Reveals the Determinants of Bacterial Engraftment in the Human Gut Following Fecal Microbiota Transplantation. *Cell Host Microbe* **23**, 229–240.e5 (2018).
88. Nicholls, S. M. *et al.* On the complexity of haplotyping a microbial community. doi:10.1101/2020.08.10.244848.

89. Lloyd, K. G., Ladau, J., Steen, A. D., Yin, J. & Crosby, L. Phylogenetically novel uncultured microbial cells dominate Earth microbiomes. doi:10.1101/303602.
90. Lloyd, K. G., Steen, A. D., Ladau, J., Yin, J. & Crosby, L. Phylogenetically Novel Uncultured Microbial Cells Dominate Earth Microbiomes. *mSystems* **3**, (2018).
91. Sangwan, N., Xia, F. & Gilbert, J. A. Recovering complete and draft population genomes from metagenome datasets. *Microbiome* **4**, 8 (2016).
92. Olson, N. D. *et al.* Metagenomic assembly through the lens of validation: recent advances in assessing and improving the quality of genomes assembled from metagenomes. *Brief. Bioinform.* **20**, 1140–1150 (2019).
93. Ayling, M., Clark, M. D. & Leggett, R. M. New approaches for metagenome assembly with short reads. *Brief. Bioinform.* **21**, 584–594 (2020).
94. Vollmers, J., Wiegand, S. & Kaster, A.-K. Comparing and Evaluating Metagenome Assembly Tools from a Microbiologist's Perspective - Not Only Size Matters! *PLoS One* **12**, e0169662 (2017).
95. Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**, 1420–1428 (2012).
96. Cleary, B. *et al.* Detection of low-abundance bacterial strains in metagenomic datasets by eigengenome partitioning. *Nat. Biotechnol.* **33**, 1053–1060 (2015).
97. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* vol. 31 1674–1676 (2015).
98. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new

versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).

99. Van Rossum, T., Ferretti, P., Maistrenko, O. M. & Bork, P. Diversity within species: interpreting strains in microbiomes. *Nat. Rev. Microbiol.* **18**, 491–506 (2020).
100. Luo, C. *et al.* ConStrains identifies microbial strains in metagenomic datasets. *Nat. Biotechnol.* **33**, 1045–1052 (2015).
101. Quince, C. *et al.* DESMAN: a new tool for de novo extraction of strains from metagenomes. *Genome Biol.* **18**, 181 (2017).
102. Olm, M. R. *et al.* InStrain enables population genomic analysis from metagenomic data and rigorous detection of identical microbial strains. *bioRxiv* 2020.01.22.915579 (2020) doi:10.1101/2020.01.22.915579.
103. Integrative HMP (iHMP) Research Network Consortium. The Integrative Human Microbiome Project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease. *Cell Host Microbe* **16**, 276–289 (2014).
104. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
105. Karsenti, E. *et al.* A Holistic Approach to Marine Eco-Systems Biology. *PLoS Biology* vol. 9 e1001177 (2011).
106. Sunagawa, S. *et al.* Structure and function of the global ocean microbiome. *Science* **348**, (2015).
107. Bertrand, D. *et al.* Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes. *Nat. Biotechnol.* **37**, 937–944 (2019).
108. Bankevich, A. & Pevzner, P. A. Joint Analysis of Long and Short Reads Enables

Accurate Estimates of Microbiome Complexity. *Cell Syst* **7**, 192–200.e3 (2018).

109. Dilthey, A. T., Jain, C., Koren, S. & Phillippy, A. M. Strain-level metagenomic assignment and compositional estimation for long reads with MetaMaps. *Nature Communications* vol. 10 (2019).
110. Beitel, C. W. *et al.* Strain- and plasmid-level deconvolution of a synthetic metagenome by sequencing proximity ligation products. *PeerJ* **2**, e415 (2014).
111. Press, M. O., Wiser, A. H., Kronenberg, Z. N. & Langford, K. W. Hi-C deconvolution of a human gut microbiome yields high-quality draft genomes and reveals plasmid-genome interactions. *biorxiv* (2017).
112. DeMaere, M. Z. & Darling, A. E. bin3C : Exploiting Hi-C sequencing data to accurately resolve metagenome-assembled genomes (MAGs). doi:10.1101/388355.
113. Yaffe, E. & Relman, D. A. Tracking microbial evolution in the human gut using Hi-C reveals extensive horizontal gene transfer, persistence and adaptation. *Nat Microbiol* **5**, 343–353 (2020).
114. Sulovari, A. *et al.* Human-specific tandem repeat expansion and differential gene expression during primate evolution. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 23243–23253 (2019).
115. Sharp, A. J. *et al.* Segmental Duplications and Copy-Number Variation in the Human Genome. *The American Journal of Human Genetics* vol. 77 78–88 (2005).
116. Liu, R. *et al.* New insights into mammalian sex chromosome structure and evolution using high-quality sequences from bovine X and Y chromosomes. *BMC Genomics* **20**, 1000 (2019).

117. Deshpande, N. & Meller, V. H. Sex chromosome evolution: life, death and repetitive DNA. *Fly* **8**, 197–199 (2014).

118. Miga, K. H. *et al.* Centromere reference models for human chromosomes X and Y satellite arrays. *Genome Research* vol. 24 697–707 (2014).

119. Attimonelli, M. & Calabrese, F. M. Human nuclear mitochondrial sequences (NumtS). *The Human Mitochondrial Genome* 131–143 (2020) doi:10.1016/b978-0-12-819656-4.00006-1.

120. Helena Mangs, A. & Morris, B. J. The Human Pseudoautosomal Region (PAR): Origin, Function and Future. *Curr. Genomics* **8**, 129–136 (2007).

121. Warmerdam, D. O. & Wolthuis, R. M. F. Keeping ribosomal DNA intact: a repeating challenge. *Chromosome Res.* **27**, 57–72 (2019).

122. Shay, J. W. Role of Telomeres and Telomerase in Aging and Cancer. *Cancer Discovery* vol. 6 584–593 (2016).

123. Willard, H. F. & Waye, J. S. Chromosome-specific subsets of human alpha satellite DNA: analysis of sequence divergence within and between chromosomal subsets and evidence for an ancestral pentameric repeat. *J. Mol. Evol.* **25**, 207–214 (1987).

124. Miga, K. H. Centromere studies in the era of 'telomere-to-telomere' genomics. *Experimental Cell Research* vol. 394 112127 (2020).

125. Bzikadze, A. V. & Pevzner, P. A. centroFlye: Assembling Centromeres with Long Error-Prone Reads. 772103 (2019) doi:10.1101/772103.

126. Nurk, S. *et al.* HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res.* (2020) doi:10.1101/gr.263566.120.

127. Miga, K. H. *et al.* Telomere-to-telomere assembly of a complete human X chromosome. *bioRxiv* 735928 (2019) doi:10.1101/735928.

128. Dvorkina, T., Bzikadze, A. V. & Pevzner, P. A. The string decomposition problem and its applications to centromere analysis and assembly. *Bioinformatics* vol. 36 i93–i101 (2020).

129. Nir, G. *et al.* Walking along chromosomes with super-resolution imaging, contact maps, and integrative modeling. *PLoS Genet.* **14**, e1007872 (2018).

130. Kaitoua, A., Gulino, A., Masseroli, M., Pinoli, P. & Ceri, S. Scalable Genomic Data Management System on the Cloud. *2017 International Conference on High Performance Computing & Simulation (HPCS)* (2017) doi:10.1109/hpcs.2017.19.

131. Merelli, I., Pérez-Sánchez, H., Gesing, S. & D'Agostino, D. Managing, analysing, and integrating big data in medical bioinformatics: open problems and future perspectives. *Biomed Res. Int.* **2014**, 134023 (2014).

132. Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132 (2016).

133. Earl, D. *et al.* Assemblathon 1: a competitive assessment of de novo short read assembly methods. *Genome Res.* **21**, 2224–2241 (2011).

134. Bradnam, K. R. *et al.* Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *Gigascience* **2**, 10 (2013).

135. Salzberg, S. L. *et al.* GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Res.* **22**, 557–567 (2012).

136. Shumate, A. *et al.* Assembly and annotation of an Ashkenazi human reference genome. *Genome Biol.* **21**, 129 (2020).

137. Nielsen, C. B., Cantor, M., Dubchak, I., Gordon, D. & Wang, T. Visualizing genomes: techniques and challenges. *Nat. Methods* **7**, S5–S15 (2010).

138. Nusrat, S., Harbig, T. & Gehlenborg, N. Tasks, Techniques, and Tools for Genomic Data Visualization. *Comput. Graph. Forum* **38**, 781–805 (2019).

139. Zook, J. M. *et al.* Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat. Biotechnol.* **32**, 246–251 (2014).

140. Zook, J. M. *et al.* An open resource for accurately benchmarking small variant and reference calls. *Nat. Biotechnol.* **37**, 561–566 (2019).

141. Nagasaki, M. *et al.* Construction of JRG (Japanese reference genome) with single-molecule real-time sequencing. *Hum Genome Var* **6**, 27 (2019).

142. Shi, L. *et al.* Long-read sequencing and de novo assembly of a Chinese genome. *Nat. Commun.* **7**, 12065 (2016).

143. Mikheenko, A., Saveliev, V. & Gurevich, A. MetaQUAST: evaluation of metagenome assemblies. *Bioinformatics* **32**, 1088–1090 (2016).

144. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).

145. Hunt, M. *et al.* REAPR: a universal tool for genome assembly evaluation. *Genome Biol.* **14**, R47 (2013).

146. Mapleson, D., Garcia Accinelli, G., Kettleborough, G., Wright, J. & Clavijo, B. J. KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. *Bioinformatics* **33**, 574–576 (2017).

147. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality and phasing assessment for genome assemblies. *bioRxiv* (2020).

148. Giani, A. M., Gallo, G. R., Gianfranceschi, L. & Formenti, G. Long walk to genomics: History and current approaches to genome sequencing and assembly. *Comput. Struct. Biotechnol. J.* **18**, 9–19 (2020).

149. Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. doi:10.1101/071282.

150. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).

151. Ruan, J. & Li, H. Fast and accurate long-read assembly with wtdbg2. *Nat. Methods* (2019) doi:10.1038/s41592-019-0669-3.

152. Shafin, K. *et al.* Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nature Biotechnology* (2020) doi:10.1038/s41587-020-0503-6.

153. Sedlazeck, F. J., Lee, H., Darby, C. A. & Schatz, M. C. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nat. Rev. Genet.* **19**, 329–346 (2018).

154. Chin, C.-S. & Khalak, A. Human Genome Assembly in 100 Minutes. doi:10.1101/705616.

155. 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).

 doi:10.20944/preprints202101.0116.v1

Acknowledgement

We thank George Church, Heng Li, Jasmijn Baaijens and Michael Baym for inspiring discussions and/or assistance in editing/figure, and their feedback on earlier versions of this manuscript.

Author contributions

S. G. researched the literature and wrote the manuscript.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 January 2021 doi:10.20944/preprints202101.0116.v1

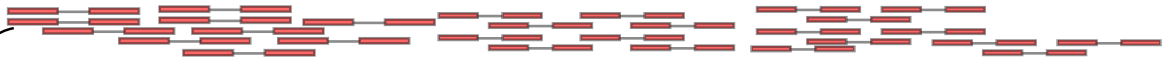
Technologies

Chromosome-scale

centromere

Local sequencing technologies (short-range)

Read length: 250 bp

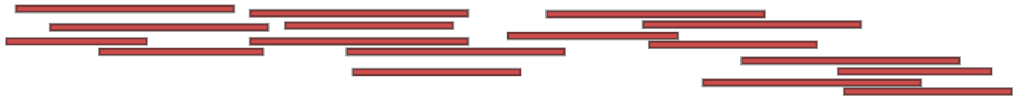


Short-read Illumina

Average read length - 20-30 kb, error rate <15%

Long-read PacBio CLR

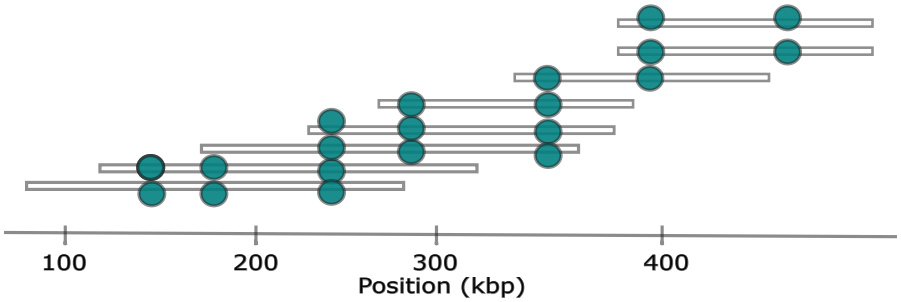
Average read length - 15-20kb, error rate < 1%



Long-read PacBio HiFi

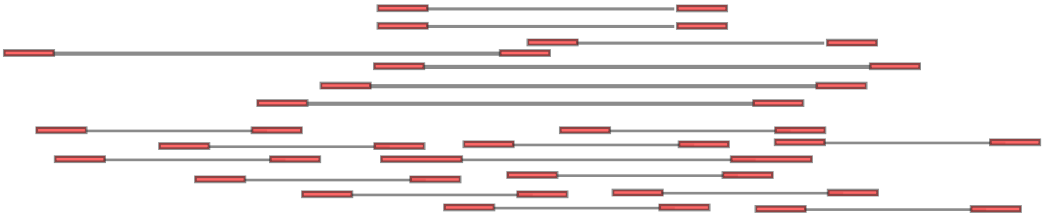
Average read length - 100 kb, error rate <6%

Ultra long-read Oxford nanopore



BioNano

Chromsome-scale (long-range)



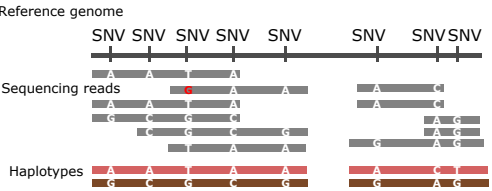
Hi-C

Sequencing cost

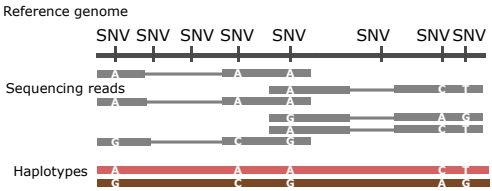
Haplotype range

Sequencing length and error-rate

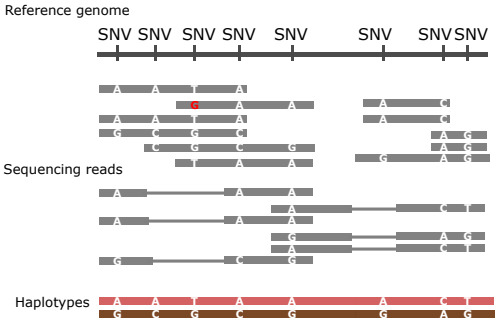
Diploid genomes



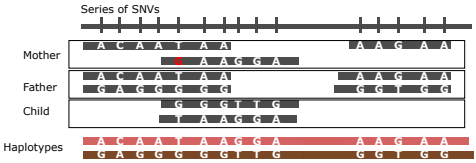
Local haplotype phasing



Chromosome-scale haplotype phasing

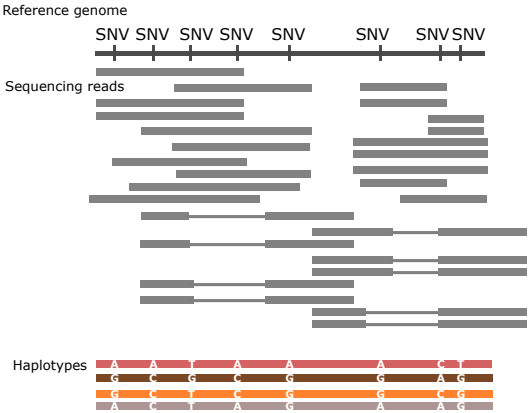


Hybrid haplotype phasing



Trio haplotype phasing

Polyploid genomes (more than two haplotypes)



Hybrid polyploid phasing

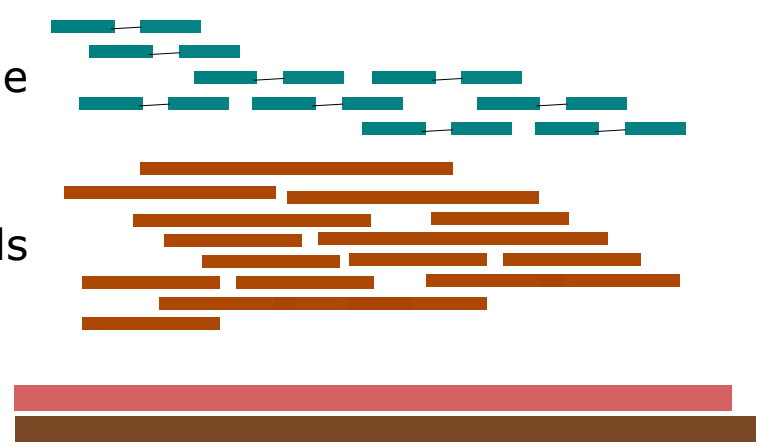
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 January 2021 doi:10.20944/preprints202101.0116.v1

Diploid genomes

Collapsed haplotypes

Long-range

Long-reads



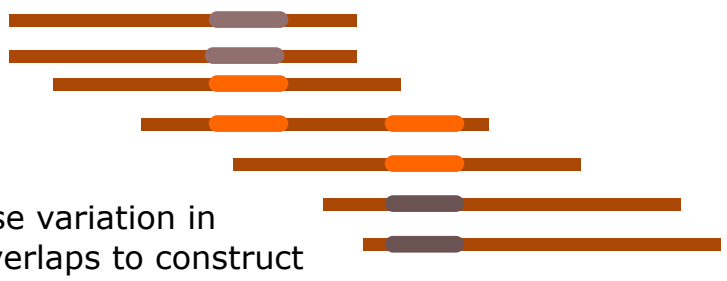
Collapsed assembly
(similar to hybrid reference-based)

Overlap graph
using long reads

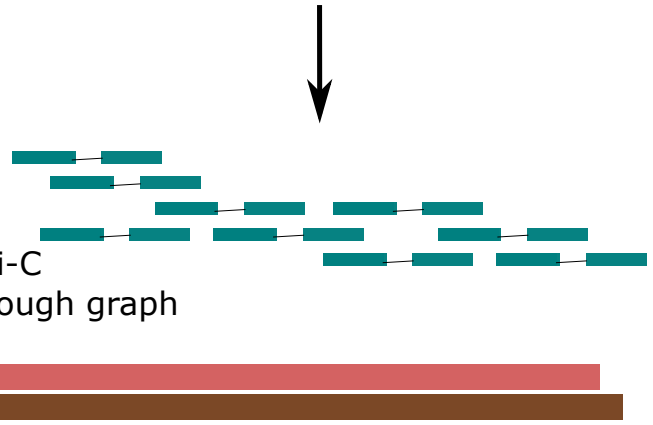
Update overlap graph using phased long-reads

Use Hi-C reads to scaffold phased overlap graph

Semi-collapsed assembly
(hybrid approach)



Use variation in
overlaps to construct
haplotype-resolved graph

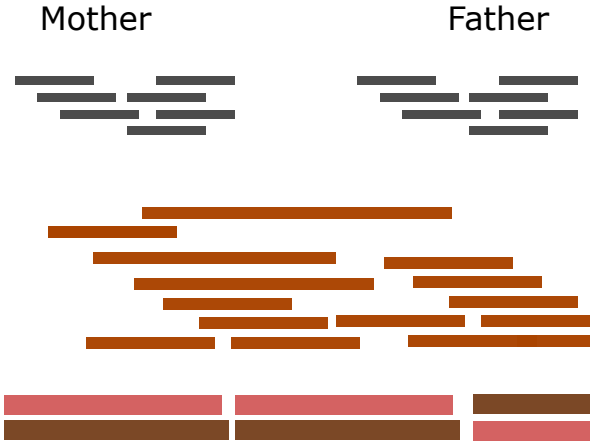


Thread Hi-C
reads through graph

Hybrid uncollapsed assembly

Mother

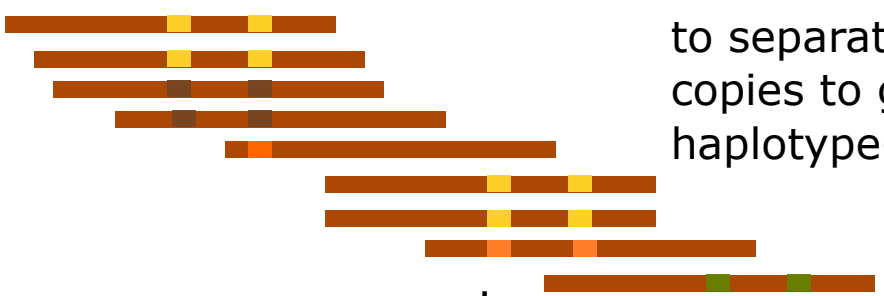
Father



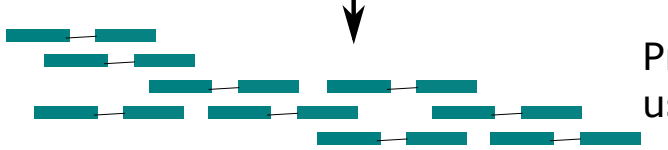
Use k-mers from
parents and
local phasing from
reads in child

Trio long read assembly

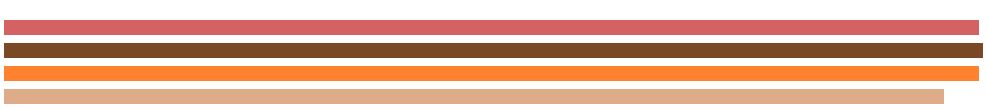
Polyploid genomes



Use variation in overlaps
to separate reads belonging to different
copies to generate
haplotype-aware overlap graph

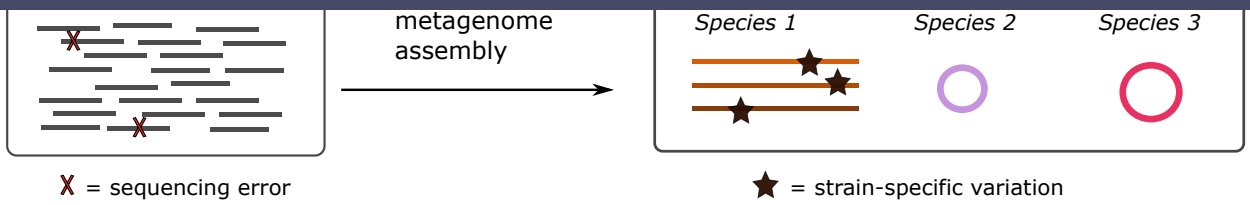


Produce scaffolds on the graph
using Hi-C



Hybrid uncollapsed assembly

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 January 2021 doi:10.20944/preprints20



b