APPLICATIONS OF NEXT-GENERATION SEQUENCING

Genetic variation and the *de novo* assembly of human genomes

Mark J. P. Chaisson[1], Richard K. Wilson[2] and Evan E. Eichler[1,3]

Abstract | The discovery of genetic variation and the assembly of genome sequences are both inextricably linked to advances in DNA-sequencing technology. Short-read massively parallel sequencing has revolutionized our ability to discover genetic variation but is insufficient to generate high-quality genome assemblies or resolve most structural variation. Full resolution of variation is only guaranteed by complete *de novo* assembly of a genome. Here, we review approaches to genome assembly, the nature of gaps or missing sequences, and biases in the assembly process. We describe the challenges of generating a complete *de novo* genome assembly using current technologies and the impact that being able to perfectly sequence the genome would have on understanding human disease and evolution. Finally, we summarize recent technological advances that improve both contiguity and accuracy and emphasize the importance of complete *de novo* assembly as opposed to read mapping as the primary means to understanding the full range of human genetic variation.

Resequencing
Characterizing a sample genome and its associated variation by mapping and aligning sequence reads to a reference genome sequence.

[1]*Department of Genome Sciences, University of Washington, Foege Building S-413A, Box 355065, 3720 15th Ave NE, Seattle, Washington 98195, USA.*
[2]*McDonnell Genome Institute, Department of Medicine, Department of Genetics, Washington University School of Medicine, St. Louis, Missouri 63108, USA.*
[3]*Howard Hughes Medical Institute, University of Washington, Seattle, Washington 98195, USA.*
Correspondence to M.J.P.C. and E.E.E.
e-mails: mchaisso@uw.edu; eee@gs.washington.edu
doi:10.1038/nrg3933
Published online 7 October 2015

Contemporary genetic studies fundamentally require comparing the sequences of individual genomes. The dominant method for this comparison is resequencing, in which random fragments of a genome are obtained and compared to a reference sequence. Such experiments are conducted using instruments for massively parallel sequencing (MPS), in which billions of 100–200-nucleotide sequences may be read by a single instrument in a few days. Although great advances have been made in our knowledge of diversity[1], cancer[2] and genetic disease[3], the genetic information provided by resequencing with current technology is incomplete. There is a lack of sensitivity for detecting small insertions and deletions (indels) and structural variation[1,4,5], there is coverage bias against particularly GC- and AT-rich DNA[6], the phase of mutations over long ranges must be inferred or imputed as opposed to directly observed, and the architecture of large polymorphic copy number variations is incomplete[7–9].

An alternative to resequencing is *de novo* assembly, in which the entire sequence of two haplotypes is resolved from sequence data without comparison to a reference-genome sequence. Although *de novo* assembly is, in principle, complete and therefore the ideal for genetic variation discovery, it is still currently impossible to achieve with data generated by typical MPS resequencing projects[10]. There is evidence that the landscape of sequencing technology is changing in such a way that will ultimately enable more-routine *de novo* assembly of genomes.

In this Review, we first describe the computational challenges of *de novo* assembly and review state-of-the-art *de novo* assembly of human and other mammalian genomes. Next, we discuss the biases involved in detecting sequence variation as a result of incomplete assembly, the implications for biomedicine and the types of variation that may be better accessed with a complete *de novo* assembly. Finally, we review new approaches coupled with advances in sequencing technology that provide additional information that may be used to resolve *de novo* assemblies of human genomes.

Strategies and algorithms for assembling genomes

The goal of *de novo* genome assembly is to determine the sequence of a genome using only randomly sampled sequence fragments, which are typically less than one-millionth the size of a mammalian genome. Most current approaches involve some aspect of a whole-genome shotgun sequencing and assembly (WGSA) strategy, in which random fragments from a genome are sequenced and computationally stitched together to generate sequence contigs and scaffolds[11]. Under ideal conditions (that is, uniformly high sequence coverage and a genome devoid of repetitive sequences), an assembly may be determined with the simple approach of merging reads with maximal overlap. However, it became clear early on that this method is too simplistic to accurately assemble genomes with complex

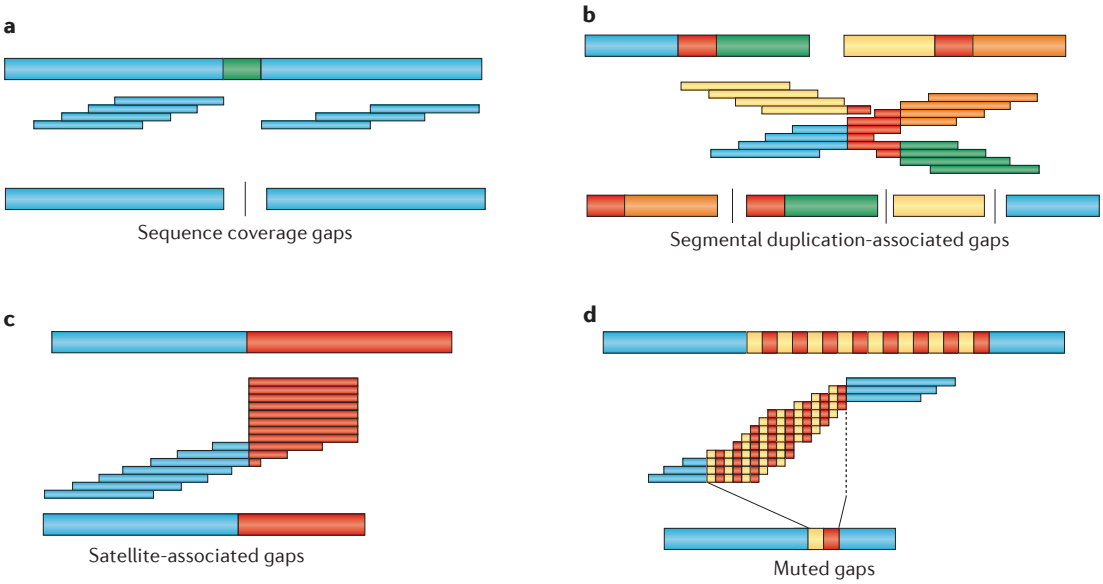


Figure 1 | **Types of genome assembly gaps.** Abstracted images of genome assemblies are illustrated. The genome architecture being resolved is shown at the top of each figure part as thick bars. Repetitive sequences are shown in red. Read overlaps are illustrated below the genome as thin bars (middle of each figure part), with regions overlapping repeats filled as red. The resulting assembly contigs are shown below (bottom of each figure part). Gaps are shown as vertical bars separating contigs to indicate unresolved sequences. **a** | The absence or reduction in sequence reads due to potential amplification or sequencing biases creates 'dropouts', where the assembled sequence is incomplete. **b** | Large segmental duplications of high sequence identity (orange and green) make read overlaps ambiguous, leading to multiple gaps flanking segmental duplications. The effect becomes exacerbated if the duplications are structurally polymorphic in a diploid genome. Long-range sequence information is required to resolve the complete sequence. **c** | Satellite-associated gaps are a special case leading to read 'pileups' due to higher-order tandem arrays of repetitive sequence, and they cannot be resolved using paired-end sequence information. These occur primarily in centromeric, acrocentric and telomeric areas of genomes. **d** | Muted gaps arise when the assembly is contracted relative to the true genome when overlaps are consistent with a smaller representation of the genome. These are often associated with repetitive sequences that cannot be easily amplified and/or are incompatible with cloning and propagation (that is, when they are toxic to *Escherichia coli*), such as simple tandem repeats.

Massively parallel sequencing
(MPS). A general term for a form of DNA sequencing that measures trace signals from millions to hundreds of millions of amplified sequences at once, most frequently referring to sequencing produced by Illumina, Life Technologies and Complete Genomics platforms. Often referred to as next-generation or second-generation sequencing to distinguish it from long-read sequencing approaches (for example, single-molecule sequencing), which are sometimes referred to as third-generation sequencing.

Structural variation
Large insertion, deletion or inversion differences between homologous chromosomes, or translocation differences involving non-homologous chromosomes. Operationally defined as events > 50 bp in size to distinguish from smaller insertion and deletion events.

Coverage bias
Regions with an excess or deficiency in the number of sequence reads originating as a result of platform differences in sequence chemistry, amplification or cloning.

Phase
The assignment of genetic variants or alleles to one of two homologous chromosomes.

De novo assembly
The action of constructing the sequence of a genome from overlapping DNA sequences without guidance from a reference genome.

Haplotypes
Sets of genetic variants or alleles found on the same chromosome that are inherited together until disrupted by recombination.

Whole-genome shotgun sequencing and assembly
(WGSA). The reconstruction of a genome from reads redundantly sampled at random, often with the aid of paired-end sequencing.

Contigs
Continuous (or 'contiguous') sequences produced in a *de novo* assembly, free of any gaps.

organizations. Sequence coverage is almost never uniform, and repetitive sequences of varying length, copy number and sequence complicate this process. This makes the correct merging of sequence reads a nearly impossible task in specific regions of the genome (FIG. 1). The key aspects for *de novo* assembly and repeat resolution are read length, overlap mapping quality and assembly algorithm.

Early genome assembly strategies. Prior to 2007, the sequence and assembly of mammalian genomes was an expensive and time-consuming operation. Although a few groups initially advocated WGSA for mammalian-genome assembly[12], the most-widely used mammalian genomes, human and mouse, were not assembled using this approach. Instead, these assemblies are relatively unique among mammalian genomes in that they were assembled almost entirely using clone-by-clone-based sequencing[13]. Each genome was divided into roughly 200-kb-long overlapping fragments that were cloned into bacterial artificial chromosomes (BACs) and individually assembled. This offered the advantage that BAC sequences that are repetitive within the context of the entire genome are locally unique, thus making gap-free assembly more tractable. As a result, these genomes have become the benchmark for comparison (FIG. 2a). When the result of a *de novo* assembly is a sequence per chromosome without gaps and with 99.99% base-pair accuracy, the assembly is considered complete; otherwise, it is considered a draft. In practice, completeness is considered for only euchromatic portions of the genome, and even the most-recent build of the human genome (GRCh38) contains gaps.

De novo *genome assembly algorithms.* Since 2013, the process of *de novo* assembly of mammalian genomes has shifted from purely WGSA using MPS to assembly with longer sequence reads generated either synthetically or by single-molecule sequencing (SMS). Algorithms for *de novo* assembly have evolved in concert with these technology improvements. The main algorithmic approaches to *de novo* assembly — overlap-layout-consensus (OLC), de Bruijn and, more recently, the string graph[14] — are each based on a separate theoretical graph framework[15]. Below, we describe some of the salient features of *de novo* assembly algorithms and outline specifically their treatment of repeats in the assembly process (FIG. 3).

During OLC assembly, overlaps between all reads are first detected, then contigs are formed by iteratively merging overlapping reads until a read heuristically determined to be at the boundary of a repeat is reached[16–18] (FIG. 3a). Repeats shorter than the minimally expected read overlap are often resolved, implying that genome resolution increases with read length. To account for sequencing errors, imprecise read overlaps are allowed, although this procedure may fragment the assembly even when the genomic repeats are nearly identical. The human genome was constructed primarily using OLC algorithms, and notable OLC-based assembly methods include parallel contig assembly program (PCAP)[18], Arachne[17] and Celera[11].

Assembly methods based on de Bruijn graphs begin, somewhat counter-intuitively, by replacing each read with the set of all-overlapping sequences of a shorter, fixed length (FIG. 3b). The length is often denoted *k*, and the sequences *k*-mers. Contigs are formed by merging *k*-mers appearing adjacently in reads halting at *k*-mers from repeat boundaries (FIG. 3b). This has the cost of requiring highly accurate reads, and it initially discards some of the ability for reads to resolve repeats longer than *k* bases. It has the benefit of not requiring the storage of pairwise overlaps and having a graph structure representative of the repeat structure of the genome. For these reasons, de Bruijn assembly has been favoured for MPS projects in the ALLPATHS[19], SOAPdenovo[20] and ABySS[21] mammalian next-generation sequencing assembly methods, in which little information is lost using *k*-mers, given the shorter read lengths.

The string graph[14] and the related A-Bruijn graph[22] assembly formulations are similar in concept to a de Bruijn graph but have the advantage of not decomposing sequences into *k*-mers but rather taking the full-length of a sequence read. They are produced based on operations of read overlap (FIG. 3c) and the removal of transitively inferred overlaps. There is an open-source implementation of string graph assembly called FALCON and produced by Pacific Biosciences.

The long sequences and high error rate of SMS reads favour overlap-based approaches over de Bruijn graphs, although it is computationally challenging to detect all pairwise overlaps between SMS reads. Initial approaches using read mapping[23] were prohibitively slow, but novel pairwise alignment algorithms[24,25] have enabled mammalian *de novo* assembly with SMS reads using modest computational resources. A final step of consensus calling must be performed for SMS read assembly that incorporates detailed error models of the underlying data, for example with Quiver[23] and Nanocorrect[26].

Diploid genomes present challenges to *de novo* assembly when heterozygous structural variation interrupts read overlap consistency. Assembly methods that assume homozygosity will fragment contigs at the boundaries of such structural variants, and methods that are diploid-aware must select only one of the haplotypes for assembly. An alternative representation of an assembly is the reference graph[22,27] with a foundation in a repeat graph[22], in which every edge represents a sequence and each vertex is at a branching point where alternative haplotypes may be selected or joined back with consensus sequence without structural variation. When available, paired-end sequences may be used to resolve sequences that are not assembled owing to repeats.

Examples of* de novo *genome assembly approaches. The first human *de novo* assembly using MPS was largely for proof-of-concept and was highly fragmented with an N50 length of 1.5 kb[21]. Improvements in MPS-based *de novo* assemblies have arisen partly from an increase in read lengths, and advances in scaffolding have arisen from an increase in the diversity and lengths of paired-end fragment library sizes (TABLE 1). The common 'recipe' for *de novo* assembly using MPS is similar to the multiple-insert-size design originally proposed for WSGA[11]. Adapted for Illumina sequencing (Illumina, Inc.), it uses a combination of high coverage reads from a short-insert library (sequences 200–500 bp long), a lower coverage of a medium-insert library (sequences 2 kb, 5 kb and 10 kb in length) and a sparse coverage of long-insert (40 kb in length) fosmid or jumping libraries[28]. This has been applied to produce draft assemblies of a Yoruban and a Han Chinese individual[19,20]. The addition of long-insert libraries leads to improved scaffolding and an increase in segmental duplication resolution[20] over a previous assembly that used only short- and medium-length inserts. Although additional human genomes have not yet been assembled using this strategy, many mammalian (FIG. 2a) and avian genomes have been generated using MPS for comparative genomics studies[29–31].

More broadly, the set of sequencing technologies and protocols used to assemble the 99 mammalian genomes in GenBank as of February 2015 is diverse: 57 assemblies used only Illumina; 22 were predominantly Sanger-based; 16 mixed Sanger, 454 Sequencing and Illumina technologies; 2 used just 454 Sequencing; 1 used Sequencing by Oligonucleotide Ligation and Detection (SOLiD); and 1 assembly method used Illumina and single-molecule optical mapping[32]. There is little relationship between sequence coverage and contig N50 for published assemblies (FIG. 2a). There is a similar heterogeneity in assembly methods even within the same sequencing methods; Sanger assemblies have used PCAP[18], Arachne[17], Celera[11] and custom BAC assembly[33], whereas genomes assembled predominantly with Illumina reads are roughly balanced between ALLPATHS[19] (31 genomes) and SOAPdenovo[20] (29 genomes). The quality of most genomes is well below the sequence contiguity and accuracy achieved by the mouse and human clone-ordered reference genomes, albeit at far less cost. Few modern genome assemblies exceed an N50 contig of 100 kb (average = 41 kb). This translates into tens to hundreds of thousands of sequence gaps, most of which correspond to various classes of repeat and MPS biases in GC representation (see above). The average percentage of core eukaryotic genes[34] present in these draft genomes is 88%. Despite impressive scaffold N50 lengths, a greater fraction of genomes is not being accurately represented within assemblies (FIG. 2b), gene models and regulatory regions are being missed, and a smaller fraction of genetic variation is being

Scaffolds
Sets of ordered and oriented contigs, with the approximate distances between contigs estimated by traversing paired-end sequences that anchor to different contigs. Scaffolds consist of both sequence contigs and gaps.

Bacterial artificial chromosomes
(BACs). Vectors with an F-plasmid origin of replication used to clonally propagate an organism's DNA (typically 150–250 kb) by transfection into *Escherichia coli*.

Single-molecule sequencing
(SMS). A form of DNA sequencing in which signals are derived from single molecules, frequently referring to sequencing produced by Pacific Biosciences and Oxford Nanopore Technologies platforms.

Paired-end
Two reads sequenced from opposite ends of the same fragment.

N50 length
A statistic in genomics defined as the shortest contig at which half the total length of the assembly is made of contigs of that length or greater. It is commonly used as a metric to summarize the contiguity of an assembly.

Fragment library
A set of DNA fragments of approximately the same length that are paired-end sequenced.

Segmental duplication
When a sequence is represented two or more times in a genome with high sequence identity and did not arise by retrotransposition. Often defined as paralogous sequences that share ≥90% sequence identity and are ≥1 kb in length.

characterized (TABLE 1) relative to clone-based references. As a result, gaps exist within the genomes and arise for various reasons (FIG. 1). We discuss these next in the context of the current human reference genome as well as the draft human genomes generated by MPS.

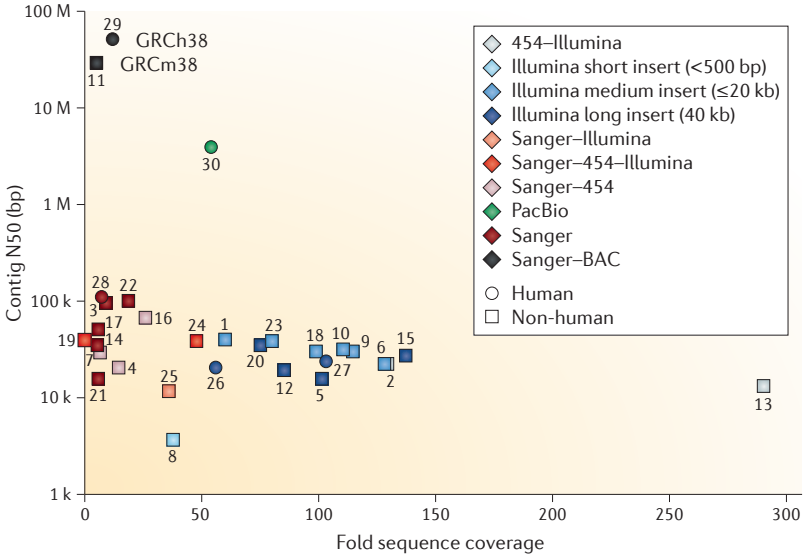
Types of gaps

Sequence-coverage gaps. Sequencing gaps occur, under the simplest condition, where no sequence reads have been sampled for a particular portion of the genome (FIG. 1a). Prior to the introduction of MPS methods, *de novo* assembly projects had limited coverage owing to the high cost of deep WGSA[33,35,36]. With low coverage (less than tenfold sequence coverage), the number of contigs produced in an assembly could be readily estimated using the Lander–Waterman (LW) statistic[37]. For example, a 5.3-fold sequence coverage from Sanger reads of 540 bp is estimated to result in an assembly with 155,906 gaps, which is roughly consistent with the observation of the low-coverage shotgun sequencing of a human genome with 206,552 gaps[36]. Such gaps are usually the easiest to remedy with additional coverage or targeted enrichment and sequencing. In principle, genomes sequenced using MPS methods do not suffer from coverage-based gaps, as the probability that a base is in a gap is vanishingly small ($P < 1.92^{-22}$ at 50× coverage under a Poisson model). Instead, there are platform-dependent sequence compositions that are not well represented, such as the reduced coverage in AT- or GC-rich regions of the genome associated with the MPS platforms[6] (FIG. 2c).

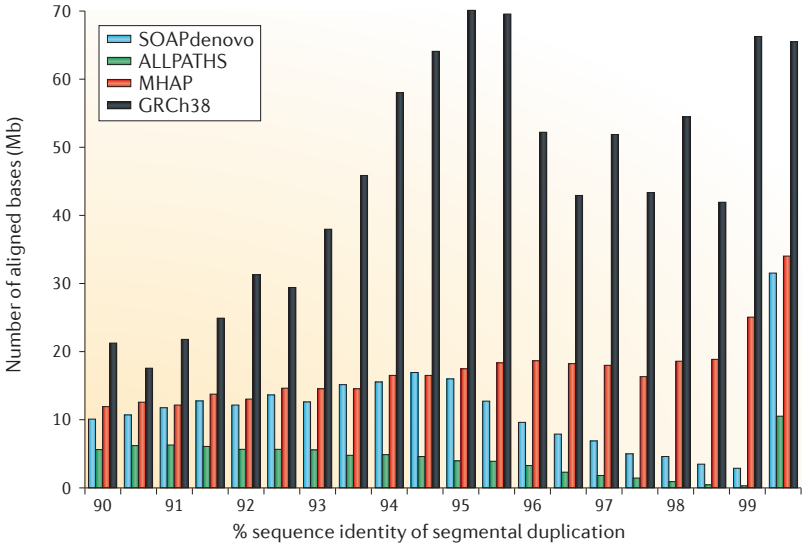
Segmental duplication-associated gaps. Over one-third (206/540) of the euchromatic gaps in the human reference genome (GRCh38) are flanked by large, highly identical segmental duplications. Analyses of human genome assemblies over the past 10 years consistently identified large tracts of duplicated DNA as the most common source of gaps, irrespective of the assembly approach or sequencing platform used[38]. The length (>10 kb), sequence identity (>96%) and structural polymorphism of these regions complicate resolution of this ~3% of the genome. Consider the simple assembly strategy outlined above for a genome that has repeated sequences longer than the length of a read. Most assembly methods iteratively merge overlapping reads into longer sequences called contigs (FIG. 3) and stop merging once the contig is extended to the boundary of a repeat sequence within the genome (FIG. 1b). The assembly process is further complicated by extensive structural polymorphisms, which are enriched more than tenfold in these particular regions[1,39,40]. Different human haplotypes (that is, allelic variation) have been shown to vary by hundreds of kilobases owing to paralogous sequences that are present in different numbers of copies and in different orientations[8,40–42]. Therefore, inadvertent assembly of two structurally diverse haplotypes within these regions is more likely than in unique regions of the genome. This results in an assembly that has no biological meaning (the assembly does not represent either of the haplotypes that are actually present) and cannot be deconvoluted until the incorrect joins are broken and the correct haplotype structures resolved[41,43]. In such regions, a single high-quality reference genome is insufficient, and there have been ongoing efforts to create multiple high-quality alternative reference haplotypes for these regions (FIG. 4A). For example, there are a total of 261 alternative references in the current human reference genome (GRCh38), corresponding primarily to regions of extreme genetic diversity often associated with segmental duplication[41,43].

Figure 2 | **Sequencing and assembly statistics from different platforms.** **a** | A comparison of sequence coverage versus N50 contig length for 30 mammalian genomes from 25 species deposited into the US National Center for Biotechnology Information (NCBI) genome resource, including 5 human genome assemblies (circles). Colours contrast different sequencing platforms and assembly approaches. GRCh38 (human) and GRCm38 (mouse), generated by Sanger sequencing of bacterial artificial chromosome (BAC) clones represent the highest quality of genome. Genomes are enumerated according to species as follows: 1, *Ailuropoda melanoleuca* GCA_000004335.1; 2, *Bos mutus* GCA_000298355.1; 3, *Bos taurus* GCA_000181335.3; 4, *Felis silvestris catus* GCA_000687225.1; 5, *Ursus maritimus* GCA_000687225.1; 6, *Balaenoptera acutorostrata* GCA_000493695.1; 7, *Callithrix jacchus* GCA_000004665.1; 8, *Daubentonia madagascariensis* GCA_000241425.1; 9, *Lipotes vexillifer* GCA_000442215.1; 10, *Pteropus alecto* GCA_000325575.1; 11 and 12, *Mus musculus* GCA_000001635.6; 13, *Nasalis larvatus* GCA_000772465.1; 14, *Nomascus leucogenys* GCA_000146795.3; 15, *Otolemur garnettii* GCA_000181295.3; 16, *Pan paniscus* GCA_000258655.1; 17, *Pan troglodytes* GCA_000001515.4; 18, *Panthera tigris* GCA_000464555.1; 19, *Papio anubis* GCA_000264685.1; 20, *Physeter macrocephalus* GCA_000472045.1; 21, *Pongo abelii* GCF_000001545.4; 22, *Rattus norvegicus* GCA_000001895.4; 23, *Saimiri boliviensis* GCA_000235385.1; 24, *Tarsius syrichta* GCA_000164805.2; 25, *Tursiops truncatus* GCA_000151865.3; 26–30, *Homo sapiens* (SOAPdenovo, ALLPATHS, HuRef, GRCh38 and MinHash Alignment Process (MHAP), respectively). **b** | The amount of duplicated sequence represented in different genome assemblies, as determined by whole-genome assembly comparison (WGAC)[107], is shown for SOAPdenovo (YH, GenBank GCA_000004845.2), ALLPATHS (NA12878, GenBank GCA_000185165.1) and MHAP (CHM1, GenBank GCA_000772585), as well as for the human reference genome (GRCh38). None of the *de novo* assemblies achieves the amount of duplication content resolved by the clone-based GRCh38 assembly, although the resolution of segmental duplication in massively parallel sequencing (MPS)-based assemblies (SOAPdenovo and ALLPATHS) is reduced compared with that of the single-molecule real-time (SMRT) sequence-based assembly MHAP. **c** | Sequencing read depth is compared to GC composition across the human genome for different platforms: CHM1 Illumina HiSeq (SRP044331), NA12878 Illumina X10 (data from AllSeq) and CHM1 SMRT P5–C3 (SRX533609). (P5–C3 refers to the version of DNA polymerase (P) and chemistry (C) used in the sequencing reaction.) The Illumina bias is decreased in more-modern instruments, whereas the SMRT sequencing coverage is more uniform, with fewer sequence-context gaps. 454, 454 Sequencing; PacBio, Pacific Biosciences.

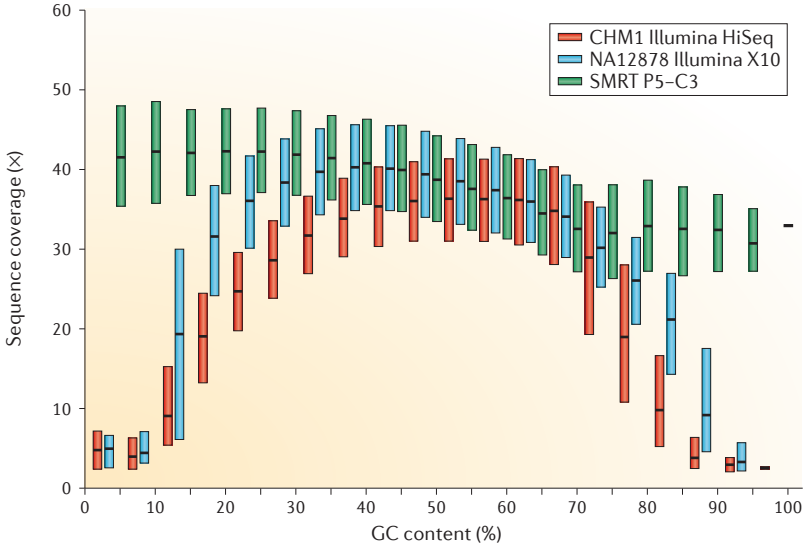
a Properties of mammalian genome assemblies



b Duplicated sequences



c Uniformity of sequence coverage according to GC content



Satellite-associated gaps. In addition to the gaps associated with gene-rich segmental duplication, other forms of repetitive DNA have been found within gap regions (FIG. 1c). These include short and long runs of tandem repeats designated as short tandem repeats (STRs; also known as microsatellites), variable number of tandem repeats (VNTRs; also known as macrosatellites) and Mb-sized centromeric satellite repeats. These sequences are difficult to assemble because the read overlaps are consistent with different copy numbers of the tandem repeat, making it impossible to determine the exact structure. It is possible to estimate the copy number using read depth, but for complete resolution it is necessary to assemble using reads longer than the total satellite length. For example, 80% of the gaps closed by application of the single-molecule real-time (SMRT) long-read sequencing technology consisted of long-degenerate, low-complexity sequences (5–10 kb long) flanked by GC-rich DNA. Many of these gaps occur within the last 5–10-Mb-long telomeric regions of chromosomal arms. Heterochromatic DNA is an extreme case of repeat-associated gaps, in which short satellite sequences are repeated in tandem hundreds of thousands of times. In addition to it being impossible to resolve the sequence structure of the heterochromatic DNA using standard assembly methods, even length estimates are crude approximations owing to the extensive length heterogeneity within a species. As a result, most centromeric, acrocentric and secondary constrictions of chromosomes are not included in standard genome assemblies. More-recent assemblies (GRCh38) have included decoy satellite sequences as a placeholder to improve mapping[44], but this has also resulted in an increase in the number of genome gaps; 212 new centromeric gaps have been added within the decoy sequences, as well as 45 additional gaps adjacent to euchromatic satellite DNA.

Muted gaps. Muted gaps are defined as regions that are inadvertently closed in an assembly but that actually show additional or different sequences in the vast majority of individuals[45] (FIG. 1d). Although some muted gaps may theoretically represent a very rare deletion variant in the individual selected for sequence assembly, most arise as a result of errors in the assembly process or the dependence on clone-based sequencing in which the sequences that are toxic to bacteria were deleted during the cloning propagation process[5]. In the human genome, it has been estimated that there are over 2,600 muted gaps[5]. These gaps frequently associate with complex repeats, tandem duplications (within a few thousand bases) or long, low-complexity STR sequences, which, as described above, are often collapsed or truncated in the assembly process. Although such gaps are not typically annotated in reference genomes, more-recent assemblies produced by the Genome Reference Consortium[46] contain 'black-tag' annotations at sites where there are uncertain merges. These sites have been shown to be enriched 200-fold for muted gaps.

Allelic variation gaps. Some regions of a genome also show extraordinary patterns of allelic variation, often reflecting deep evolutionary coalescence (for example,

Short tandem repeats
(STRs). Tandem repeats in which the individual unit of repetition is less than 10 bp long and varies in length between different individuals in a population.

Variable number of tandem repeats
(VNTR). Any tandem array of repeated sequence motifs that are highly variable in different individuals of a population. Historically, these were originally used in reference to tandem repeats that varied on the scale of thousands of base pairs over the length of the array.

Centromeric
Referring to the primary cytogenetic constriction on metaphase chromosomes where the kinetochore forms and spindle fibre attaches during cell division. In humans the centromere is made up primarily of repetitions of higher-order alpha-satellite DNA.

Heterochromatic DNA
Portions of chromosomes that stain densely, are typically gene poor and are rich in satellite sequences.

Acrocentric
Relating to a type of chromosome in which the centromere maps very close to the short arm. Acrocentric chromosomes in humans are enriched in beta-satellite and ribosomal DNA sequences, which are repeated as hundreds of copies.

Secondary constrictions
A cytogenetic term referring to metaphase chromosome constrictions outside the centromere, typically rich in satellites and used to help identify chromosomes.

Satellite DNA
Highly repetitive DNA composed of thousands to tens of thousands of tandem repeats, usually between 100–300 bp in length, and frequently associated with heterochromatin.

Muted gaps
Regions that have been incorrectly closed in a genome assembly despite additional sequences being present at these sites in the source genome.

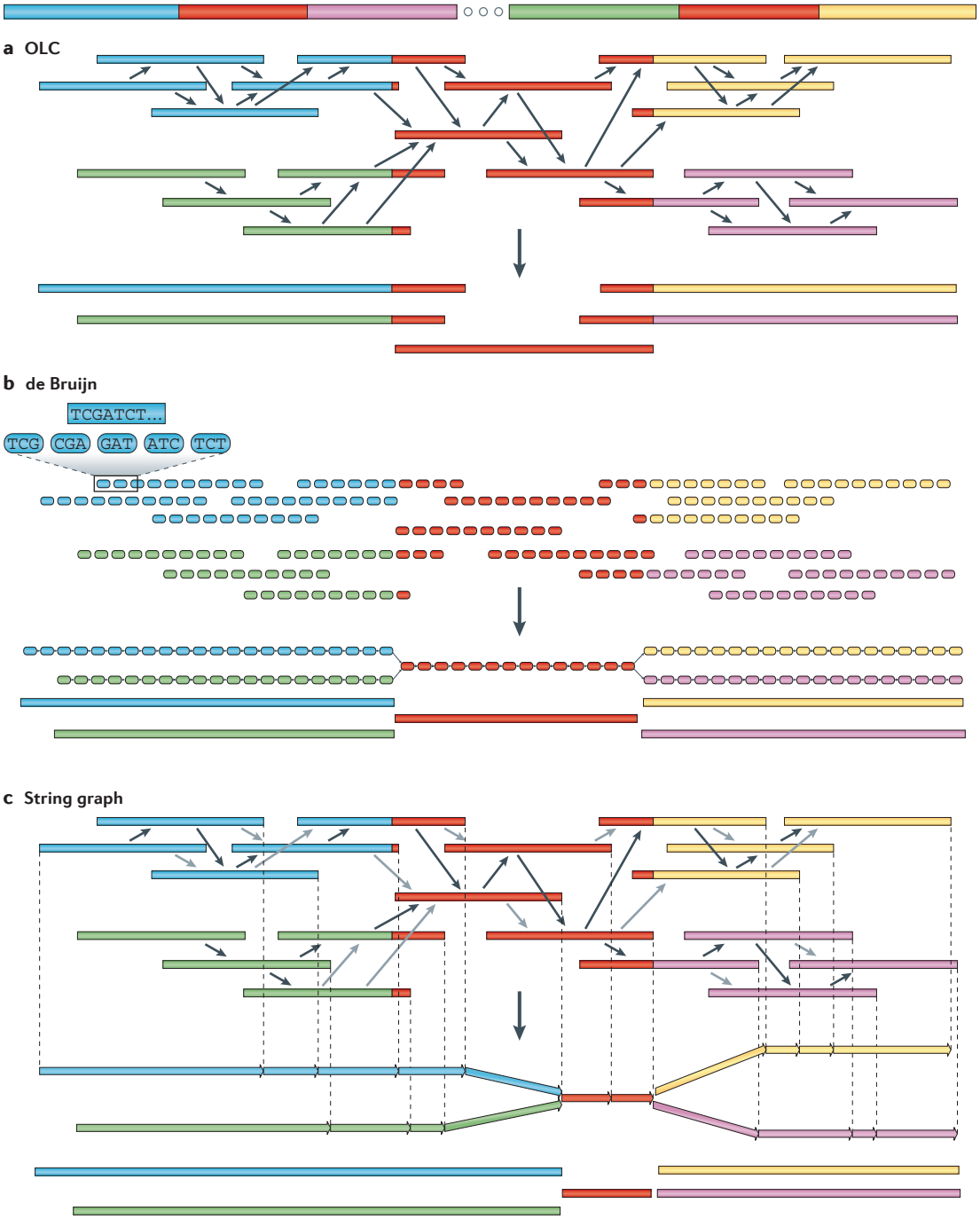


Figure 3 | **Genome assembly algorithms.** A genome schematic is shown at the top with four unique regions (blue, violet, green and yellow) and two copies of a repeated region (red). Three different strategies for genome assembly are outlined below this schematic. **a** | Overlap-layout-consensus (OLC). All pairwise alignments (arrows) between reads (solid bars) are detected. Reads are merged into contigs (below the vertical arrow) until a read at a repeat boundary (split colour bar) is detected, leading to a repeat that is unresolved and collapsed into a single copy. **b** | de Bruijn assembly. Reads are decomposed into overlapping *k*-mers. An example of the decomposition for $k = 3$ nucleotides is shown, although in practice *k* ranges between 31 and 200 nucleotides. Identical *k*-mers are merged and connected by an edge when appearing adjacently in reads. Contigs are formed by merging chains of *k*-mers until repeat boundaries are reached. If a *k*-mer appears in multiple positions (red segment) in the genome, it will fragment assemblies and additional graph operations must be applied to resolve such small repeats. The *k*-mer approach is ideal for short-read data generated by massively parallel sequencing (MPS). **c** | String graph. Alignments that may be transitively inferred from all pairwise alignments are removed (grey arrows). A graph is created with a vertex for the endpoint of every read. Edges are created both for each unaligned interval of a read and for each remaining pairwise overlap. Vertices connect edges that correspond to the reads that overlap. When there is allelic variation, alternative paths in the graph are formed. Not shown, but common to all three algorithms, is the use of read pairs to produce the final assembly product.

Table 1 | **Human genome assemblies**

	ABySS (NA12878)	SOAPdenovo	SOAPdenovo2 (Yan)	ALLPATHS-LG (NA12878)	MHAP (CHM1)	HuRef	DISCOVAR (NA12878)	GRCh38
Accession number	NA	NA	GCA_000004845.2	GCA_000185165.1	GCA_000772585	GCA_000002125.2	NA	NA
Instrument	Illumina	Illumina	Illumina	Illumina	PacBio	Sanger	Illumina	Sanger
Number of scaffolds*	2,760,000	NA	4,197	3,331	NA	4,528	NA	Complete
Scaffold N50 (bp)*	1,499	446,283	22,047,463	12,140,992	NA	NA	NA	NA
Number of contigs‡	4,348,132	3,884,491	239,711	231,194	40,917	65,049	949,302	646
Contig assembly length (bp)	NA	NA	2,851,884,686	2,771,622,777	3,260,000,000	2,782,378,670	3,085,280,835	3,100,000,000
Reference coverage (%)	NA	NA	87.8	81.3	89.7	89.6	90.2	100
Contig N50 (bp)‡	870	1,050	21,011	24,024	3,945,491	111,232	178,303	57,879,411
Percentage exons§	NA	NA	89.9	79.5	93.2	93.7	95.0	100
Percentage UTRs	NA	NA	86.2	86.2	91.5	89.7	92.6	100
Sequence coverage	42	71	56	103.1	54	7.5	69	NA
Largest clone‖	210	10,000	40,000	40,000	NA	200,000	450	NA
Read length¶	42	75	100	101	4,800	540	250	NA

MHAP, MinHash Alignment Process; NA, not applicable; PacBio, Pacific Biosciences; UTRs, untranslated regions. Each genome assembly was downloaded from GenBank and mapped using BLASR back to GRCh38. *Scaffold count and N50 contig length were calculated from published genome assemblies. The scaffold N50 is not applicable to HuRef, as the assembly is reference-guided, nor to the MHAP and DISCOVAR genome assemblies, because neither produces scaffolds. ‡Contig counts and N50 were calculated by removing scaffolding information. §We used the coordinates corresponding to the exons from RefSeq and computed the percentage of exons and UTRs identified based on this annotation. 'Percentage exons' is the percentage of all 181,147 exons in the reference genome completely contained within a contig, with 'percentage UTRs' representing a similar metric for the 97,750 UTRs. ‖The largest clone denotes the greatest mate-pair span used to generate the assembly, which correlates with resolution of duplicated sequences. ¶For MHAP and HuRef, the average read length is reported, whereas for ABySS, SOAPdenovo, SOAPdenovo2 and ALLPATHS multiple read sizes are used and the maximum is reported.

Coalescence
The genealogy of a region of the genome in which all alleles trace back to a common ancestral sequence.

Missing heritability
The observation that only a portion of estimated genetic contribution to disease (for example, heritability of a trait from twin studies) can be explained by our current understanding of genetic variation and its transmission properties.

Exome sequencing
A method for enrichment and targeted sequencing of the protein-coding portions of the genome using massively parallel sequencing.

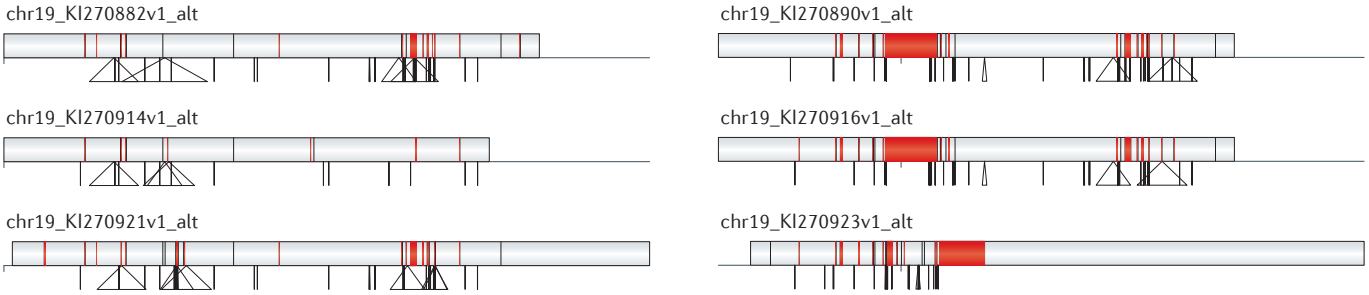
human leukocyte antigen (HLA) or the 17q21.31 polymorphism). Such divergence is frequently missed, especially when diploid genomes are assembled, because a particular haplotype is initially seeded or preferred during the assembly process. In other cases, the similar but not identical allelic haplotypes will appear as repeats to the assembly methods, initiating a contig break at the boundary between similar and diverged sequence. Such regions are not always easily recognized as gaps in the genome (they are a type of muted gap (FIG. 1d) if one haplotype is already resolved). Targeted approaches are usually required to sequence and assemble alternative diverged haplotypes[43,47].

Consequences and impact

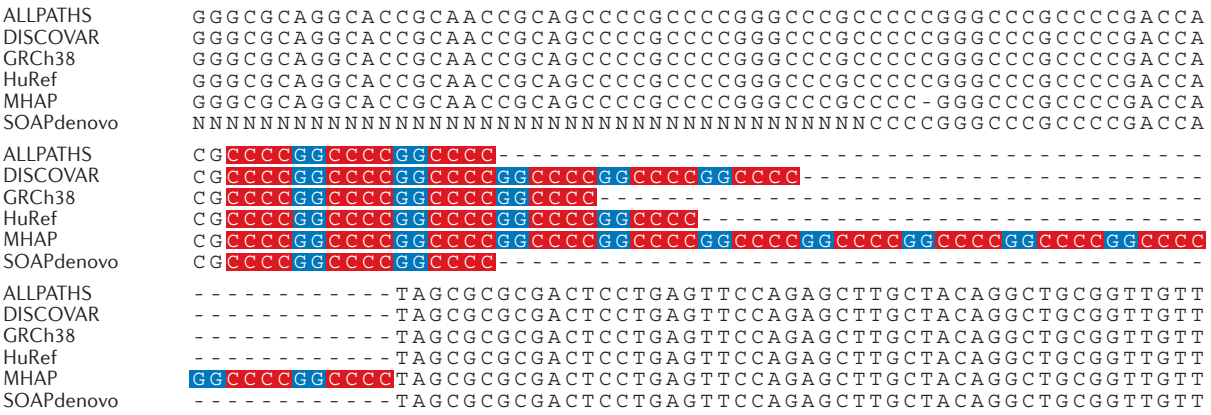
Accurate assembly of genomes has long been recognized as the key to understanding genetic variation. The more accurate the reference genome, the easier it is to map read data and interpret functional importance. A more-complete reference leads to better annotation, less genotyping error genome-wide[1,48] and, concomitantly, a greater likelihood of identifying causal variation associated with human traits. Missing data (in the form of complex genetic variation or sequence gaps) translate into an inability to discover disease-causing mutations and, as a result, contribute to the problem of 'missing heritability' (REF. 49). A few recent examples are illustrative of this issue, even with respect to the high calibre of the human reference genome.

Unresolved repeats. The region associated with medullary cystic kidney disease type 1 (MCKD1) was identified in the early 2000s[50], but the causative mutation for this straightforward Mendelian disease eluded discovery for several years[51]. Copy number variant analysis, exome sequencing, targeted capture and characterization of noncoding sequences failed to identify the causal variant. A VNTR region containing repeats of 60 bp in length composed primarily of GC-rich DNA (80%) was associated with a candidate gene encoding mucin 1 (*MUC1*)[51]. Southern blot analysis revealed that the VNTR was substantially underrepresented in the human reference genome compared to what had been observed in the normal human population. Targeted cloning, sequencing and *de novo* assembly was necessary to resolve the sequence structure of this particular region; when

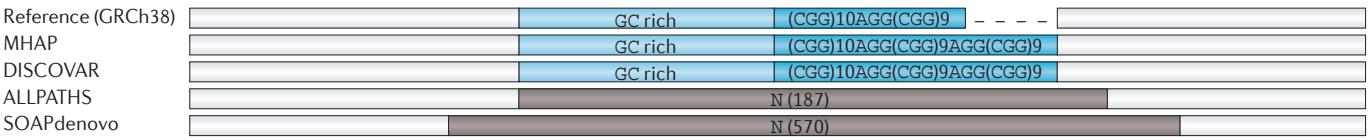
A KIR alternative haplotypes



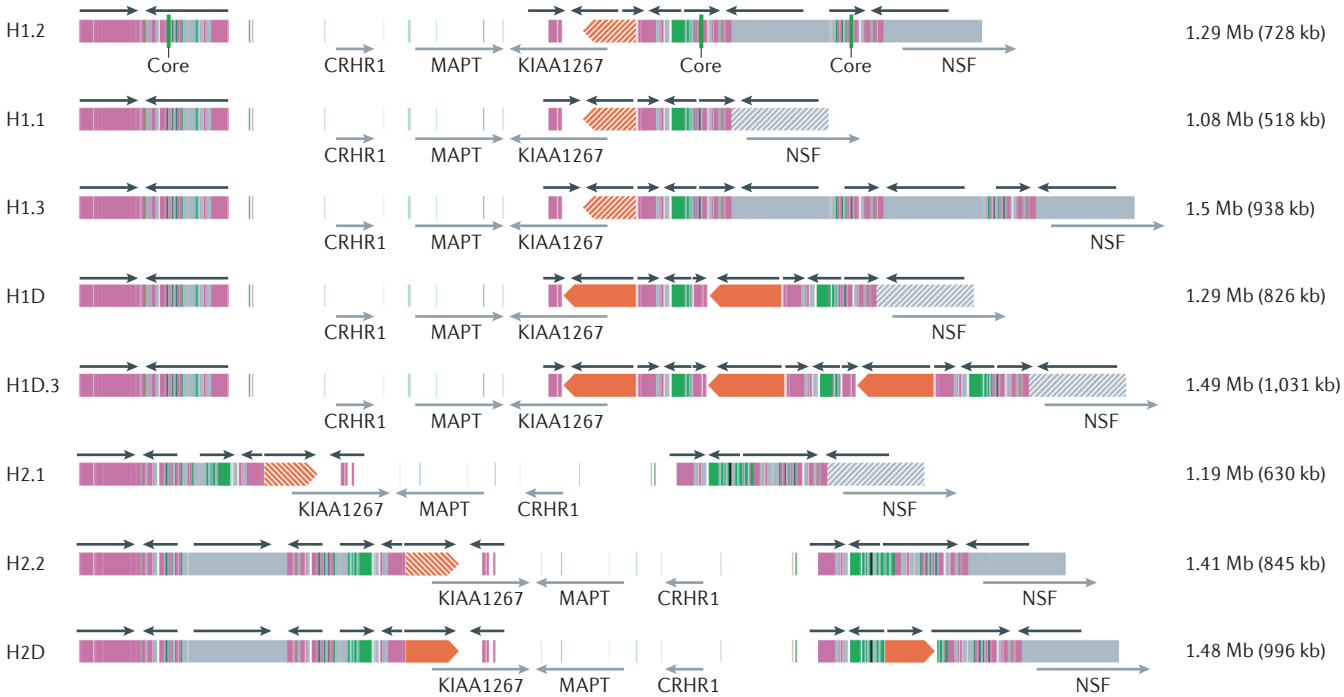
Ba ***C9ORF72***



Bb ***FMR1***



C Chromosome 17q21.31 alternative haplotypes



◀ Figure 4 | **Assembly of complex regions of human genetic variation. A** | Six alternative haplotypes (GRCh38) in the KIR region (chromosome 19q13.42), assembled and sequenced using fosmid clones. The span of each haplotype with respect to the reference genome is denoted by the large rectangle, with the reference genome length being shown as a horizontal line below the rectangle. Deletions (red) are shown within the rectangle and insertions as triangles below, with the base of each triangle representing the length of the insertion. **B** | A comparison of two GC-rich disease-causing loci, chromosome 9 open reading frame 72 (*C9ORF72*; which causes frontotemporal dementia (FTD) and amyotrophic lateral sclerosis (ALS)) and fragile X mental retardation 1 (*FMR1*; which causes fragile X syndrome), in different genome assemblies. The sequence motif associated with the ALS *C9ORF72* hexanucleotide repeat (red and blue) is partially resolved in all assemblies except for SOAPdenovo, in which the flanking 3′ region, which contains a divergent repeat motif and interspersed adenine nucleotides, is incomplete (**Ba**). The *FMR1* trinucleotide repeat associated with fragile X syndrome is resolved by the DISCOVAR and MinHash Alignment Process (MHAP) assemblies (**Bb**). **C** | Eight different genomic structures associated with direct (H1) and inverted (H2) haplotypes of a gene-rich region on chromosome 17q21.31. The overall length of this region varies from 1.08 Mb to 1.50 Mb owing to variation in segmental duplication content (shown as coloured bars, with the total length given in brackets on the right). The H2D configuration is the only configuration that has large, highly identical duplications in a direct orientation predisposing to deletions that cause Koolen–de Vries syndrome. CRHR1, corticotropin releasing hormone receptor 1; MAPT, microtubule-associated protein tau; NSF, *N*-ethylmaleimide-sensitive fusion protein. Part **C** reproduced from REF. 7, Nature Publishing Group.

the region had been properly assembled, comparisons between individuals with and those without the disease revealed that the disease is associated with a single cytosine insertion within a stretch of seven cytosines (at positions 53–59) in a protein-coding region of *MUC1*. The resulting frameshift leads to a premature stop codon and hence a truncated mutant MUC1 protein. This truncated MUC1 is defective due to its lack of functional domains and underlies MCKD1 in >60% of families. In this case, both incomplete assembly and short-read mapping biases common to MPS data complicated discovery.

A similar cautionary tale of incomplete genome assembly and variation assessment was revealed for the discovery of the mutations responsible for amyotrophic lateral sclerosis (ALS)[52,53] — the most common cause of adult-onset motor neuron disease in the human population. Manual realignment of chromosome flow-sorted sequencing data, as opposed to routine variant-calling algorithms, identified an expanded GGGGCC hexanucleotide repeat located within the first intron of the chromosome 9 open reading frame 72 (*C9ORF72*) transcript for an estimated 40% of familial ALS cases[52,53]. The *C9ORF72* repeat is probably condensed in the ALLPATHS (NA12878) and HuRef assemblies, but is completely unresolved in the SOAPdenovo assembly (FIG. 4Ba).

GC-rich repetitive sequences are frequently missing or underrepresented from most modern genome assemblies. For example, the CGG trinucleotide sequence within the 5′-untranslated region of the fragile X mental retardation 1 (*FMR1*) gene is expanded to 200–1,000 copies in individuals with fragile X syndrome[54], but the sequence of unaffected individuals, which occurs in 5–55 CGG trinucleotide copies, is missing from assemblies that used reads with lengths of up to 101 bp (ALLPATHS and SOAPdenovo). This locus is resolved in an assembly of NA12878 using 250-bp-long MPS sequences assembled with DISCOVAR. Interestingly, the *FMR1* CGG repeat is shown to be further expanded in both the SMS-based MinHash Alignment Process (MHAP) assembly[54] and the DISCOVAR Illumina assembly (FIG. 4Bb), suggesting that normal allele lengths, but not pathogenic repeat sizes, can be sequenced and assembled accurately.

In the case of facioscapulohumeral muscular dystrophy[55], the genetic basis of the disease was not fully understood until the repeat-rich region was accurately assembled and contraction of the D4Z4 VNTR repeat in conjunction with a point mutation within duplicated DNA was identified in patients with the disease. This particular constellation of genomic features leads to the stable production of a toxic double homeobox 4 (*DUX4*) transcript and took more than 10 years to deduce, in large part because of difficulties in assembly and in assessing genetic variation in this region of chromosome 4q35.

Unannotated genes. In some cases, functional elements, including entire genes, still await discovery in new human genomes. The discovery of mutations associated with thyrotoxic hypokalaemic periodic catalysis required the cloning and sequencing of inward rectifier potassium channel 18 (*KCNJ18*; also known as *KIR2.6*), first by PCR and later in large-insert BAC clones. Variation in this gene, which is ~99% identical to *KCNJ12*, was originally thought to represent allelic variation of *KCNJ12* until the presence of the duplicated copy (*KCNJ18*) could be definitively proven and mutation of this associated with the phenotype[56].

Missing sequence. It is currently estimated that 5–40 Mb of euchromatic sequences are absent from a given human reference genome owing to structural polymorphism[57–59] and that an additional 125–150 Mb of gene-rich regions of the genome are inaccessible to standard variation analyses[1]. High-quality complete *de novo* sequence assembly would provide access to these regions and the inherent allelic variation within, without having to infer the composition of these regions from structural variation methods. Many of these regions show evidence of association with complex diseases, such as obesity[60], lupus[61] and infectious diseases[62]. In the absence of complete sequence structure, the findings remain difficult to interpret or are often mired in controversy because of associations for or against a particular phenotype (for example, C-C motif chemokine 3-like 1 (*CCL3L1*) and HIV infection[63], and α-amylase 1 (*AMY1*) and obesity[60,64]). The frequency of recurrent mutation and the size and complexity of structural variation further limits imputation by a flanking-tag single-nucleotide polymorphism[1,39]. For such regions of the genome, genetic variation discovery is synonymous with high-quality sequence and assembly.

Characterizing normal genetic variation. In addition to disease, the inability to sequence and assemble portions of the genome limits our understanding of normal human genetic variation and evolution. The HLA region is perhaps the best-studied example in which alternative haplotypes differ by as much as 10%–15% between humans, but additional examples of hidden genetic variation, including

novel paralogous copies, have emerged as a result of high-quality *de novo* sequencing and assembly of alternative haplotypes. For example, sequencing of 4 different haplotypes of the chromosome 15q13.3 region harbouring the gene neuronal acetylcholine receptor subunit alpha 7 (*CHRNA7*) showed that individuals differ by more than 1.5 Mb over this 3-Mb region, primarily with respect to duplicated genes[41]. These differences have important consequences, such as causing specific haplotypes to be predisposed to, or protected from, recurrent microdeletions that are associated with disease. Some complex structural haplotypes have risen to high allele frequencies in populations, consistent with their role in adaptive evolution. For example, the 17q21.31 inversion polymorphism, haplotype H2 (REFS 7,8) (FIG. 4C), has been associated with increased fecundity and overall global recombination[65] but, paradoxically, an increased risk of Koolen–de Vries syndrome[66]. Recently, novel human-specific duplicate genes — such as SLIT–ROBO Rho GTPase-activating protein 2C (*SRGAP2C*), which is important for neuronal migration and dendrite density — were discovered as missing from the human reference genome[9,67]. Local BAC-based assemblies resolved the sequence and structure of three distinct copies of *SRGAP2C* that emerged during the human lineage of evolution, adding 379,665 bp of new sequence completely absent from the human reference and correcting 559,693 bp of incorrectly mapped sequence. This recently filled gap in *SRGAP2C* is even more remarkable as it is now considered a potential explanation — along with another human-specific gene, Rho GTPase-activating protein 11B (*ARHGAP11B*)[68] — for the expansion of the human neocortex since humans and australopithecines diverged 2–3 million years ago.

Bioinformatics and technology advances

Recent algorithmic development in *de novo* assembly of MPS reads has focused on reducing computational space requirements[69–71], more-sensitive variant detection[72,73], single-cell sequencing[74] and metagenomics[75]. Novel methods for resolving repeat structures with paired ends have only been demonstrated on smaller genomes[76]. The methods commonly used to produce mammalian assemblies — ABySS, SOAPdenovo and ALLPATHS-LG — have improved assembly results through incremental development[19,20] and more-optimized use of library insert sizes, which have improved scaffold N50 statistics. Bioinformatics approaches are intrinsically linked to advances in sequencing technology. Newer approaches that take advantage of longer short-reads (for example, DISCOVAR *de novo* for 250-bp Illumina reads) hold considerable promise for improving contig lengths and enhancing the discovery of genetic variation[72].

Sequencing-based scaffolding methods. More-substantial gains in assembly quality have been made through novel applications of MPS and developments of long-range sequencing technologies. To improve assemblies using existing MPS methods, sequencing techniques have focused on compartmentalizing sequence data into bins based on proximity intervals ranging from 10 kb to 1,000 kb in length[77]. For example, the ligating adjacent chromatin enables scaffolding *in situ* (LACHESIS) method[78], leverages the observation that most reads sequenced by high-throughput chromatin-interaction sequencing (Hi-C) experiments map to chromatin intervals of ~1 Mb. As a result, this method has been used to significantly enhance the scaffolding of human MPS data at the chromosomal scale. As expected, erroneous scaffolds are highly enriched for segmental duplications (6.4-fold) and simple tandem repeats (2.9-fold). However, Hi-C read counts do not correlate well over shorter genomic distances (<1 Mb), hence contigs are difficult to more-precisely order[78], requiring further algorithmic improvement. Thus, local complex repeat architectures cannot be readily resolved at the sequence level with this method, although the method does show promise for phasing[79]. A complementary technology, contiguity-preserving transposition sequencing (CPT-seq), uses a transposase-mediated barcoding strategy to generate low-coverage (0.05–0.10×) sequencing from midrange (80-kb-long) fragments[80] that may be used in scaffolding. However, the sequences read from each fragment are unordered, making it similarly difficult to resolve complicated repeat structures.

Dilution pool sequencing. Another approach shared by several methods is dilution pool sequencing[81], in which pools of low-concentration, high-molecular-weight DNA are separately barcoded and sequenced in aggregate; the number of sequences associated with the same barcode is sufficiently low that they do not overlap allelically. Low-level random amplification of DNA from each pool significantly improves haplotype phasing[82–84], while greater levels of amplification give sufficient coverage for preassembly of reads from each dilution pool into longer synthetic reads[85]. This method has been adapted by Illumina as TruSeq and applied to synthetic reads averaging 4.39 kb from *Drosophila melanogaster* to build an assembly with an N50 of 69.7 kb[86]. There were 3,524 gaps in this assembly, of which 93% were due to decreased synthetic read coverage attributed to AT-rich regions (29.7%), consistent with PCR bias, and were enriched 2.3-fold for transposable elements and 10.4-fold for simple tandem repeats.

Optical mapping technologies. Complementing the long-range sequencing approaches have been advances in long-range mapping technologies that are enhancing the quality of draft *de novo* assemblies and highlighting potential inconsistencies and larger forms of structural variation. For example, high-throughput optical mapping technologies[87] have now been successfully commercialized that operate by imaging individual DNA molecules at fluorescently labelled restriction sites (such as products created by OpGen, Inc. and BioNano Genomics). Signals are typically read at 1–9 kb intervals, and consensus-assembled optical maps are produced from overlapping restriction patterns of large molecules (>100 kb). These methods have been applied to scaffolding prokaryotic genomes[88], discovery of complex structural variation, BAC assembly[89] and mammalian genome assembly[32]. The Irys system, as an example, was recently used to generate optical maps of fragments ≥150 kb to resolve

one of the most duplication-rich regions (chromosome 1q21) of the human genome[90]. These approaches provide accurate scaffolding information for contigs that may be uniquely aligned to the optical map. The methods require the isolation of high-molecular-weight DNA, indicating a resurgence in the need for expertise in pulsed-field gel electrophoresis (PFGE) DNA isolation or the development of new high-molecular-weight DNA isolation methods.

SMS. Developments in nanoscale sensing in zero-mode waveguides[91] and nanopores[92] hold considerable promise for revolutionizing the field of *de novo* genome assemblies. These technologies have enabled routine sequencing from single molecules without cloning and amplification of DNA. Importantly, the trace signal does not attenuate with read length, unlike in MPS methods that sequence from amplified DNA, allowing read lengths that are 1 to 2 orders of magnitude longer and that have fewer sequencing biases[6] (FIG. 2c) and the potential to be limited only by the length of intact DNA after sample preparation. For example, the SMRT P6–C4 chemistry now produces sequence reads of >10 kb from genomic libraries that are >20 kb in length. Sequence error with this technology is more random in nature

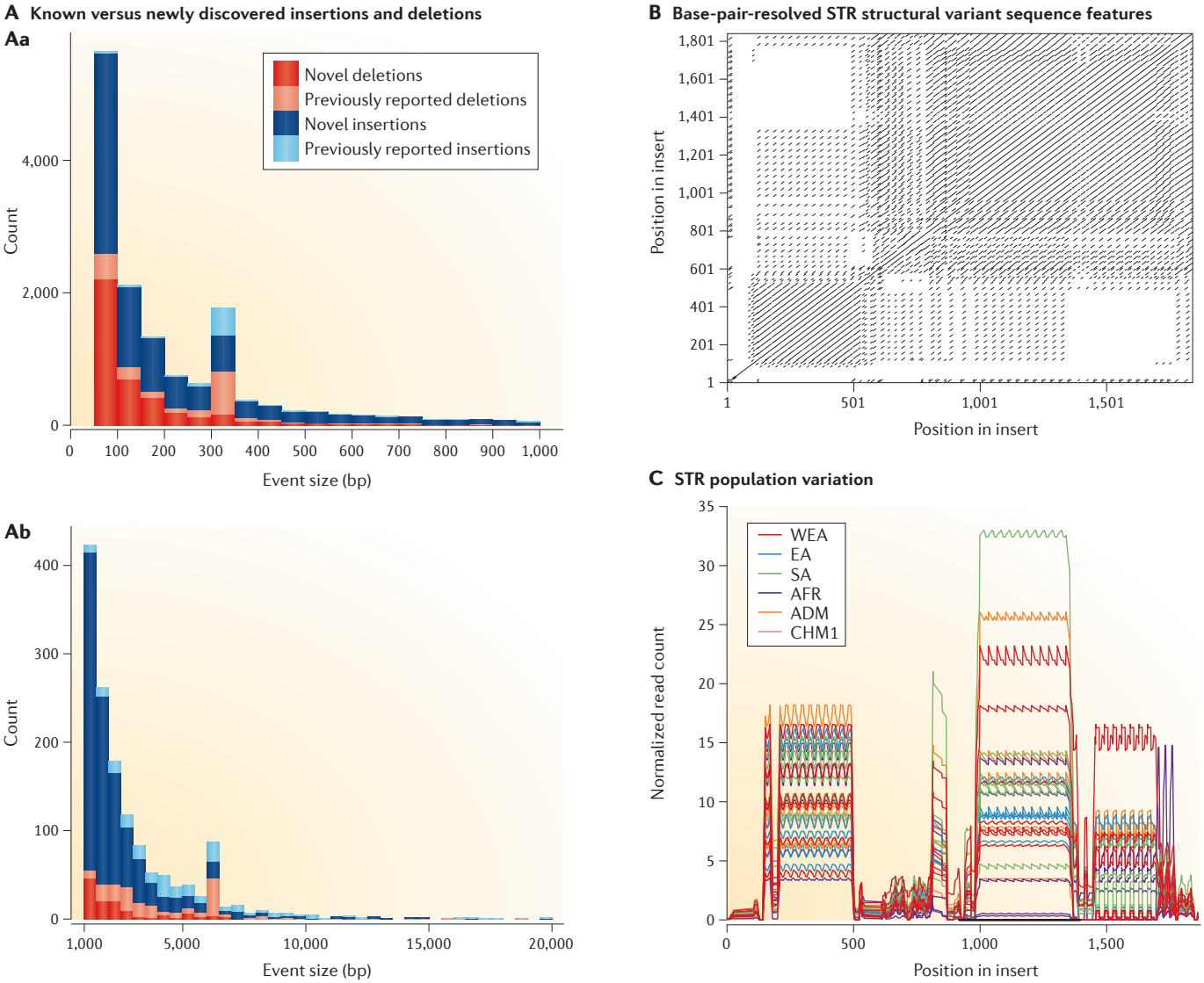


Figure 5 | **Human genetic variation detected with local assembly of single molecules.** **A** | Deletions (red and pink) and insertions (dark and light blue) resolved at base-pair resolution in the genome from the CHM1 cell line through local assembly of the single-molecule real-time (SMRT) reads for events less than 1 kb (**Aa**) and greater than 1 kb (**Ab**)[5]. Copy number variants found in previous studies[4,57,108] are in lighter shades, with roughly 85% of events being unique to the CHM1 results. **B** | An example of a 1.7-kb short tandem repeat (STR) insertion event (represented in a self dot plot) not detected by Illumina resequencing of CHM1 but detected and assembled by SMRT reads. **C** | This STR insertion contains uniquely identifying 30 bp sequences that, once sequence resolved, may be used to genotype the presence of the insertion in genomes sequenced using Illumina technology. Normalized read depth serves as a proxy for estimating variability of STR length and demonstrates that the STR is highly variable in diverse populations (shown for Western Eurasian (WEA), East Asian (EA), South Asian (SA), African (AFR) and admixed (ADM) individuals). Figure adapted from REF. 5, Nature Publishing Group.

Box 1 | **New human reference genomes and standards for genetic variation**

With the realization that one reference human genome would be insufficient for variation discovery and precision medicine, several initiatives have been launched in the past couple of years to create new human reference genomes and/or sets of highly curated and validated genetic variants for benchmarking of new technologies.

Genome in a Bottle

Genome in a Bottle (GIAB) is a US National Institute of Standards and Technology (NIST) public–private consortium of investigators whose goal is to develop reference standards, methods and datasets for genome analysis. It is using multiple technologies to create standard variant callsets for which the sensitivity and specificity of new variant calling methods developed for clinical sequencing may be evaluated. Data and reference material from one European genome (NA12878) and a father, mother and son Ashkenazim trio (GM24149, GM24143 and GM24385, respectively) have been made available. Additional reference genomes will help tune methods to produce more accurate calls facilitating the implementation of whole-genome sequencing in clinical practice.

Platinum and gold genomes

The goal of this US National Human Genome Research Institute (NHGRI) initiative is to generate new, high-quality reference genomes that are as accurate as ('gold'), or are more complete and accurate than ('platinum'), the current human reference genome. Target platinum genomes include two complete hydatidiform moles that are devoid of allelic variation (CHM1 and CHM13) for which draft assemblies are already available using single-molecule real-time (SMRT) technology (for example, GCA_000772585.3 and GCA_001015355.1). In addition, five additional diploid gold genomes have been selected as continental references from Europe, Asia, America and Africa (two samples from Africa). The strategy involves deep genome sequencing using multiple sequence platforms along with targeted bacterial artificial chromosome (BAC)-based sequencing of complex regions to generate haplotype-resolved genome assemblies[105].

Human Genome Structural Variation Consortium

The Human Genome Structural Variation Consortium (HGSVC) is a large consortium of investigators (including members of the 1,000 Genomes Project) focused on increasing the sensitivity and specificity of structural variation detection. One of its goals is to deeply characterize the pattern and inheritance of structural variation in two-parent, one-child trios. Using a battery of sequencing and orthogonal mapping technologies, nine genomes are currently being characterized for structural variation, including Yoruban, Puerto Rican and Han Chinese trios. An exhaustive analysis and validation of structural variation in these three families will enhance the development of structural variant discovery and genotyping algorithms[106].

than it is for MPS, largely due to the small observable signal from nanoscale devices and non-redundant sampling. Mapping SMRT sequencing reads followed by local assembly has revealed that ~50% of the current gaps in the human genome build 37 can now be resolved or reduced[5]. More importantly, a comparison of one human genome (from a haploid hydatidiform mole, the CHM1 cell line) to the human reference suggests that more than 23,000 structural variants and longer indels, corresponding to 10.6 Mb of sequence, can now be resolved at the single-base-pair level (FIG. 5A,B). In contrast with methods that indirectly detect structural variation with discordant paired-end mapping[93,94] and read depth[95], base-pair resolution generates alternative reference sequences that allow genotyping analyses using other MPS datasets (FIG. 5C). Remarkably, 85% of insertions and 65% of deletions between 50 bp and 1,000 bp in length are novel, when compared to structural variant maps produced for more than 1,000 genomes sequenced and analysed using MPS technologies[4,96]. A large, common inversion polymorphism in 15q13.3 mediated by large (>50 kb) inverted duplications[97] that is detectable in CHM1 with BAC end mapping was not detected in the SMRT sequence alignments, indicating a limitation of the local alignment approach for detecting large structural variation. This suggests that SMS technologies have the potential to provide a more-comprehensive assessment of the full range of human genetic variation through local alignment, but that the complete assessment of variation will rely on complete *de novo* assembly.

However, current commercially available instruments for SMS (from Oxford Nanopore Technologies and Pacific Biosciences,) have a much higher sequencing error rate than other MPS technologies, such as Illumina (~15%)[98,99]. Although this can be overcome by higher sequencing coverage (>30×), this translates into a much higher cost per base than other next-generation sequencing methods. Initial applications to assembling SMS reads have been used on human BACs[100], bacterial genome assemblies[23,101–103] and eukaryotic genome assemblies through hybrid approaches that couple MPS and SMRT sequencing technologies[102,104]. Efficient methods to detect all pairwise overlaps between reads[24,25] have enabled mammalian assembly from SMS reads alone. A draft human genome based solely on *de novo* assembly of SMRT reads from the CHM1 genome (GenBank accession number GCA_000772585) demonstrates an N50 of 4.5 Mb[24], which is 20-fold greater than the initial draft sequence assemblies of the human genome using MPS. This draft closed 51 gaps in the current human reference genome. Analysis of the segmental duplication content of this assembly reveals that the structure of ~200 kb of flanking sequence for 5,543 out of 7,177 segmental duplication regions is resolved, although this is biased towards the resolution of shorter segmental duplication regions, as only 24.8% (41.4 of 167 Mb) of segmentally duplicated bases are resolved. This represents a substantial improvement when compared to some of the early MPS *de novo* assemblies, in which >90% of duplications were misassembled[38]. However, the sequences of gaps are

enriched tenfold for segmental duplications, indicating the inability to resolve larger segmental duplications with current SMS read lengths. Furthermore, the assembly is based on a haploid sample, and the ability to accurately resolve complex regions in diploid genomes has not yet been demonstrated by long-read technology. This is the critical next step in the application of this technology.

Conclusions and perspective

The ability to sequence genomes more deeply in a rapid and cost-effective manner has not yet translated into improved *de novo* assemblies of genomes. Although sequencing genomes is routine, their assembly is not. Generating a high-quality genome remains a more-expensive proposition that requires considerably more time and effort to do well. As a result, some of the most genetically diverse and complex regions of the human genome are not understood, and large swathes of structural genetic variation remain undiscovered. If a comprehensive understanding of genetic variation is the key to resolving the missing heritability paradox, cost and sequence coverage are not the only considerations.

Advances in long-read sequencing and single-molecule mapping technologies in the past few years have now made it possible for individual laboratories to consider the possibility of generating high-quality *de novo* assemblies of new genomes. Using these technologies, efforts are underway to create new vertebrate reference genomes and to produce additional human genomes with quality on par with or exceeding the quality of the human reference ('gold' and 'platinum' human genome assemblies). Other efforts are focused on applying these technological advances to understand the full range of human genetic variation more completely as standards for clinical sequencing (BOX 1). Efforts are being pursued to assess genetic variation using complete *de novo* assembly as well as mapping and local re-assembly, as each may be complementary. For example, in the MHAP assembly of CHM1 (REF. 24), 8 gaps in the human reference were closed that were not closed by local assembly[5]; similarly, local assembly closed 25 gaps not resolved by whole-genome *de novo* assembly.

Although it is clear that multiple human reference genomes corresponding to different continental or ethnic groups will emerge over the next few years, continued advances in sequencing technology are essential for moving this field forwards. Substantial improvements in sequence read length (>100 kb) as well as reductions in cost and gains in long-read throughput (by a factor of 10) are required before comprehensive and accurate *de novo* assemblies of human genomes become routine. Until that time, comprehensive *de novo* assembly of genomes will still depend on hybrid approaches that leverage large-insert clones, mixed whole-genome sequence datasets and long-range sequence data. Nevertheless, the routine and accurate *de novo* assembly of the complete 6-Gb-long diploid human genome should represent a milestone not only for the genomics community but also for the clinical interpretation of human genetic disease. *De novo* assembly of human genomes, as opposed to alignment to a reference, promises a more accurate and comprehensive understanding of human genetic variation, consistent with the vision of precision medicine.

1. Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
2. Weinstein, J. N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).
3. Ng, S. B. *et al.* Exome sequencing identifies the cause of a mendelian disorder. *Nat. Genet.* **42**, 30–35 (2010).
4. Mills, R. E. *et al.* Mapping copy number variation by population-scale genome sequencing. *Nature* **470**, 59–65 (2011).
5. Chaisson, M. J. P. *et al.* Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**, 608–611 (2015).
Long-read sequencing paired with local assembly reveals structural variation and closes or extends ~50% of the gaps in the reference human genome.
6. Ross, M. G. *et al.* Characterizing and measuring bias in sequence data. *Genome Biol.* **14**, R51 (2013).
7. Steinberg, K. M. *et al.* Structural diversity and African origin of the 17q21.31 inversion polymorphism. *Nat. Genet.* **44**, 872–880 (2012).
High-quality sequencing of the 17q21.31 region reveals a complex haplotype polymorphic region in which certain structural haplotypes predispose for disease.
8. Boettger, L. M., Handsaker, R. E., Zody, M. C. & McCarroll, S. A. Structural haplotypes and recent evolution of the human 17q21.31 region. *Nat. Genet.* **44**, 881–885 (2012).
Uses population genetics to infer the architecture and evolutionary history of chromosome 17q21.31 haplotypes. References 7 and 8 show a rapid rise of a particular inverted haplotype in European and Middle Eastern individuals that is consistent with adaptive selection.
9. Dennis, M. Y. *et al.* Evolution of human-specific neural *SRGAP2* genes by incomplete segmental duplication. *Cell* **149**, 912–922 (2012).
Shows that genes potentially responsible for unique aspects of human neuronal development were missing from the reference human genome, highlighting the importance of focusing on obtaining higher-quality reference sequences.
10. Motahari, A. S., Bresler, G. & Tse, D. N. C. Information theory of DNA shotgun sequencing. *IEEE Trans. Inf. Theory* **59**, 6273–6289 (2013).
11. Myers, E. W. *et al.* A whole-genome assembly of *Drosophila*. *Science* **287**, 2196–2204 (2000).
12. Weber, J. L. & Myers, E. W. Human whole-genome shotgun sequencing. *Genome Res.* **7**, 401–409 (1997).
13. Church, D. M. *et al.* Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLoS Biol.* **7**, e1000112 (2009).
14. Myers, E. W. The fragment assembly string graph. *Bioinformatics* **21**, ii79–ii85 (2005).
15. Nagarajan, N. & Pop, M. Sequence assembly demystified. *Nat. Rev. Genet.* **14**, 157–167 (2013).
A review of algorithmic details of fragment assembly.
16. Myers, E. W. Toward simplifying and accurately formulating fragment assembly. *J. Comput. Biol.* **2**, 275–290 (1995).
17. Batzoglou, S. *et al.* ARACHNE: a whole-genome shotgun assembler. *Genome Res.* **12**, 177–189 (2002).
18. Huang, X., Wang, J., Aluru, S., Yang, S.-P. & Hillier, L. PCAP: a whole-genome assembly program. *Genome Res.* **13**, 2164–2170 (2003).
19. Gnerre, S. *et al.* High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl Acad. Sci. USA* **108**, 1513–1518 (2011).
20. Luo, R. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *GigaScience* **1**, 18 (2012).
21. Simpson, J. T. *et al.* ABySS: a parallel assembler for short read sequence data. *Genome Res.* **19**, 1117–1123 (2009).
22. Pevzner, P. A., Tang, H. & Tesler, G. *De novo* repeat classification and fragment assembly. *Genome Res.* **14**, 1786–1796 (2004).
23. Chin, C. S. *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10**, 563–569 (2013).
Describes the method of correcting sequencing error in long SMRT sequences with short SMRT sequences so that they may be assembled using the Celera assembler and consensus called with the Quiver method.
24. Berlin, K. *et al.* Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat. Biotechnol.* **33**, 623–630 (2015).
Introduces one of the first SMS assemblers. Draft genomes on par with the original human draft sequence may be efficiently assembled with SMS reads.
25. Myers, G. in *Algorithms in Bioinformatics* (eds Raphael, B. & Tang, J.) 52–67 (Springer, 2014).
26. Loman, N. J., Quick, J. & Simpson, J. T. A complete bacterial genome assembled *de novo* using only nanopore sequencing data. *Nat. Methods* **12**, 733–735 (2015).
27. Dilthey, A., Cox, C., Iqbal, Z., Nelson, M. R. & McVean, G. Improved genome inference in the MHC using a population reference graph. *Nat. Genet.* **47**, 682–688 (2015).
The first practical study using a graphical representation of the genome to encode the structural diversity of the major histocompatibility complex region.
28. Williams, L. J. *et al.* Paired-end sequencing of Fosmid libraries by Illumina. *Genome Res.* **22**, 2241–2249 (2012).
29. Yim, H. S. *et al.* Minke whale genome and aquatic adaptation in cetaceans. *Nat. Genet.* **46**, 88–92 (2014).
30. Parker, J. *et al.* Genome-wide signatures of convergent evolution in echolocating mammals. *Nature* **502**, 228–231 (2013).
31. Zhang, G. *et al.* Comparative genomics reveals insights into avian genome evolution and adaptation. *Science* **346**, 1311–1320 (2014).

32. Dong, Y. *et al.* Sequencing and automated whole-genome optical mapping of the genome of a domestic goat (*Capra hircus*). *Nat. Biotechnol.* **31**, 135–141 (2013).
33. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
34. Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061–1067 (2007).
35. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
36. Istrail, S. *et al.* Whole-genome shotgun assembly and comparison of human genome assemblies. *Proc. Natl Acad. Sci. USA* **101**, 1916–1921 (2004).
37. Lander, E. S. & Waterman, M. S. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* **2**, 231–239 (1988).
38. Alkan, C., Sajjadian, S. & Eichler, E. E. Limitations of next-generation genome sequence assembly. *Nat. Methods* **8**, 61–65 (2011).
39. Sharp, A. J. *et al.* Segmental duplications and copy-number variation in the human genome. *Am. J. Hum. Genet.* **77**, 78–88 (2005).
40. Sudmant, P. H. *et al.* Diversity of human copy number variation and multicopy genes. *Science* **330**, 641–646 (2010).
41. Antonacci, F. *et al.* Palindromic *GOLGA8* core duplicons promote chromosome 15q13.3 microdeletion and evolutionary instability. *Nat. Genet.* **46**, 1293–1302 (2014).
42. Pyo, C. W. *et al.* Recombinant structures expand and contract inter and intragenic diversification at the *KIR* locus. *BMC Genomics* **14**, 89 (2013).
43. Zody, M. C. *et al.* Evolutionary toggling of the *MAPT* 17q21.31 inversion region. *Nat. Genet.* **40**, 1076–1083 (2008).
44. Altemose, N., Miga, K. H., Maggioni, M. & Willard, H. F. Genomic characterization of large heterochromatic gaps in the human genome assembly. *PLoS Comput. Biol.* **10**, e1003628 (2014).
45. Eichler, E. E. Recent duplication, domain accretion and the dynamic mutation of the human genome. *Trends Genet.* **17**, 661–669 (2001).
46. Church, D. M. *et al.* Modernizing reference genome assemblies. *PLoS Biol.* **9**, e1001091 (2011).
47. Raymond, C. K. *et al.* Ancient haplotypes of the HLA Class II region. *Genome Res.* **15**, 1250–1257 (2005).
48. Li, H. Towards better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* **30**, 2843–2851 (2014).
49. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
50. Fuchshuber, A. *et al.* Refinement of the gene locus for autosomal dominant medullary cystic kidney disease type 1 (MCKD1) and construction of a physical and partial transcriptional map of the region. *Genomics* **72**, 278–284 (2001).
51. Kirby, A. *et al.* Mutations causing medullary cystic kidney disease type 1 lie in a large VNTR in *MUC1* missed by massively parallel sequencing. *Nat. Genet.* **45**, 299–303 (2013).
52. Renton, A. E. *et al.* A hexanucleotide repeat expansion in *C9ORF72* is the cause of chromosome 9p21-linked ALS-FTD. *Neuron* **72**, 257–268 (2011).
53. DeJesus-Hernandez, M. *et al.* Expanded GGGGCC hexanucleotide repeat in noncoding region of *C9ORF72* causes chromosome 9p-linked FTD and ALS. *Neuron* **72**, 245–256 (2011).
54. Eichler, E. E. *et al.* Haplotype and interspersion analysis of the *FMR1* CGG repeat identifies two different mutational pathways for the origin of the fragile X syndrome. *Hum. Mol. Genet.* **5**, 319–330 (1996).
55. Lemmers, R. J. *et al.* Digenic inheritance of an *SMCHD1* mutation and an FSHD-permissive *D4Z4* allele causes facioscapulohumeral muscular dystrophy type 2. *Nat. Genet.* **44**, 1370–1374 (2012).
56. Ryan, D. P. *et al.* Mutations in potassium channel Kir2.6 cause susceptibility to thyrotoxic hypokalemic periodic paralysis. *Cell* **140**, 88–98 (2010).
57. Kidd, J. M. *et al.* A human genome structural variation sequencing resource reveals insights into mutational mechanisms. *Cell* **143**, 837–847 (2010).
58. Genovese, G. *et al.* Using population admixture to help complete maps of the human genome. *Nat. Genet.* **45**, 406–414 (2013).
59. Li, R. *et al.* Building the sequence map of the human pan-genome. *Nat. Biotechnol.* **28**, 57–63 (2010).
Describes how the draft assembly of a personal genome using MPS uncovered 19–40 Mb of sequence missing from the reference.
60. Falchi, M. *et al.* Low copy number of the salivary amylase gene predisposes to obesity. *Nat. Genet.* **46**, 492–497 (2014).
61. Yang, Y. *et al.* Gene copy-number variation and associated polymorphisms of complement component C4 in human systemic lupus erythematosus (SLE): low copy number is a risk factor for and high copy number is a protective factor against SLE susceptibility in European Americans. *Am. J. Hum. Genet.* **80**, 1037–1054 (2007).
62. Shen, S., Pyo, C. W., Vu, Q., Wang, R. & Geraghty, D. E. The essential detail: the genetics and genomics of the primate immune response. *ILAR J.* **54**, 181–195 (2013).
63. Hollox, E. J. & Hoh, B. P. Human gene copy number variation and infectious disease. *Hum. Genet.* **133**, 1217–1233 (2014).
64. Usher, C. L. *et al.* Structural forms of the human amylase locus and their relationships to SNPs, haplotypes and obesity. *Nat. Genet.* **47**, 921–925 (2015).
65. Stefansson, H. *et al.* A common inversion under selection in Europeans. *Nat. Genet.* **37**, 129–137 (2005).
66. Koolen, D. A. *et al.* A new chromosome 17q21.31 microdeletion syndrome associated with a common inversion polymorphism. *Nat. Genet.* **38**, 999–1001 (2006).
67. Charrier, C. *et al.* Inhibition of SRGAP2 function by its human-specific paralogs induces neoteny during spine maturation. *Cell* **149**, 923–935 (2012).
68. Florio, M. *et al.* Human-specific gene *ARHGAP11B* promotes basal progenitor amplification and neocortex expansion. *Science* **347**, 1465–1470 (2015).
69. Chikhi, R. & Rizk, G. Space-efficient and exact de Bruijn graph representation based on a Bloom filter. *Algorithms Mol. Biol.* **8**, 22 (2013).
70. Simpson, J. T. & Durbin, R. Efficient *de novo* assembly of large genomes using compressed data structures. *Genome Res.* **22**, 549–556 (2012).
71. Zimin, A. V. *et al.* The MaSuRCA genome assembler. *Bioinformatics* **29**, 2669–2677 (2013).
72. Weisenfeld, N. I. *et al.* Comprehensive variation discovery in single human genomes. *Nat. Genet.* **46**, 1350–1355 (2014).
Shows that MPS deduces more variation than do resequencing methods.
73. Li, H. Exploring single-sample SNP and INDEL calling with whole-genome *de novo* assembly. *Bioinformatics* **28**, 1838–1844 (2012).
74. Nurk, S. *et al.* Assembling single-cell genomes and mini-metagenomes from chimeric MDA products. *J. Comput. Biol.* **20**, 714–737 (2013).
75. Howe, A. C. *et al.* Tackling soil diversity with the assembly of large, complex metagenomes. *Proc. Natl Acad. Sci. USA* **111**, 4904–4909 (2014).
76. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
77. Adey, A. *et al. In vitro*, long-range sequence information for *de novo* genome assembly via transposase contiguity. *Genome Res.* **24**, 2041–2049 (2014).
78. Burton, J. N. *et al.* Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. *Nat. Biotechnol.* **31**, 1119–1125 (2013).
79. Selvaraj, S., Dixon, J. R., Bansal, V. & Ren, B. Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nat. Biotechnol.* **31**, 1111–1118 (2013).
80. Amini, S. *et al.* Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and combinatorial indexing. *Nat. Genet.* **46**, 1343–1349 (2014).
81. 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).
82. Kaper, F. *et al.* Whole-genome haplotyping by dilution, amplification, and sequencing. *Proc. Natl Acad. Sci. USA* **110**, 5552–5557 (2013).
83. Kuleshov, V. *et al.* Whole-genome haplotyping using long reads and statistical methods. *Nat. Biotechnol.* **32**, 261–266 (2014).
84. Kitzman, J. O. *et al.* Haplotype-resolved genome sequencing of a Gujarati Indian individual. *Nat. Biotechnol.* **29**, 59–63 (2011).
85. Voskoboynik, A. *et al.* The genome sequence of the colonial chordate, *Botryllus schlosseri*. *eLife* **2**, e00569 (2013).
86. McCoy, R. C. *et al.* Illumina TruSeq synthetic long-reads empower *de novo* assembly and resolve complex, highly-repetitive transposable elements. *PLoS ONE* **9**, e106689 (2014).
87. Schwartz, D. C. *et al.* Ordered restriction maps of *Saccharomyces cerevisiae* chromosomes constructed by optical mapping. *Science* **262**, 110–114 (1993).
88. Onmus-Leone, F. *et al.* Enhanced *de novo* assembly of high throughput pyrosequencing data using whole genome mapping. *PLoS ONE* **8**, e61762 (2013).
89. Lam, E. T. *et al.* Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nat. Biotechnol.* **30**, 771–776 (2012).
90. O'Bleness, M. *et al.* Finished sequence and assembly of the DUF1220-rich 1q21 region using a haploid human genome. *BMC Genomics* **15**, 387 (2014).
91. Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133–138 (2009).
92. Rosenstein, J. K., Wanunu, M., Merchant, C. A., Drndic, M. & Shepard, K. L. Integrated nanopore sensing platform with sub-microsecond temporal resolution. *Nat. Methods* **9**, 487–492 (2012).
93. Hormozdiari, F., Alkan, C., Eichler, E. E. & Sahinalp, S. C. Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Res.* **19**, 1270–1278 (2009).
94. Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012).
95. Handsaker, R. E., Korn, J. M., Nemesh, J. & McCarroll, S. A. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat. Genet.* **43**, 269–276 (2011).
96. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* http://www.dx.doi.org/10.1038/nature15393 (2015).
97. Sharp, A. J. *et al.* A recurrent 15q13.3 microdeletion syndrome associated with mental retardation and seizures. *Nat. Genet.* **40**, 322–328 (2008).
98. Chaisson, M. J. & Tesler, G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* **13**, 238 (2012).
99. Quick, J., Quinlan, A. R. & Loman, N. J. A reference bacterial genome dataset generated on the MinION™ portable single-molecule nanopore sequencer. *GigaScience* **3**, 22 (2014).
100. Huddleston, J. *et al.* Reconstructing complex regions of genomes using long-read sequencing technology. *Genome Res.* **24**, 688–696 (2014).
101. Bashir, A. *et al.* A hybrid approach for the automated finishing of bacterial genomes. *Nat. Biotechnol.* **30**, 701–707 (2012).
102. Koren, S. *et al.* Hybrid error correction and *de novo* assembly of single-molecule sequencing reads. *Nat. Biotechnol.* **30**, 693–700 (2012).
103. Prjibelski, A. D. *et al.* ExSPAnder: a universal repeat resolver for DNA fragment assembly. *Bioinformatics* **30**, i293–301 (2014).
104. English, A. C. *et al.* Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS ONE* **7**, e47768 (2012).
105. Callaway, E. 'Platinum' genome takes on disease. *Nature* **515**, 323 (2014).
106. Human Genome Structural Variation Consortium. The phase 3 structural variant dataset. *1000 Genomes* [online], http://www.1000genomes.org/phase-3-structural-variant-dataset (2015).
107. Bailey, J. A., Yavor, A. M., Massa, H. F., Trask, B. J. & Eichler, E. E. Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res.* **11**, 1005–1017 (2001).
108. Conrad, D. F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704–712 (2010).

Acknowledgements
The authors thank T. Brown for assistance in editing this manuscript. This work was supported, in part, by a US National Institutes of Health grant (2R01HG002385) to E.E.E.. E.E.E. is an investigator of the Howard Hughes Medical Institute.

Competing interests statement
The authors declare competing interests: see Web version for details.

FURTHER INFORMATION
AllSeq: http://allseq.com/x-ten-test-data
FALCON: www.github.com/PacificBiosciences/falcon
GenBank: http://www.ncbi.nlm.nih.gov/genbank/
Genome in a Bottle: https://sites.stanford.edu/abms/giab
US National Center for Biotechnology Information (NCBI) genome resource: http://www.ncbi.nlm.nih.gov/genome

ALL LINKS ARE ACTIVE IN THE ONLINE PDF