ANNUAL REVIEWS

Annual Review of Genomics and Human Genetics

Long-Read DNA Sequencing: Recent Advances and Remaining Challenges

Peter E. Warburton[1,2] and Robert P. Sebra[1,2,3,4]

[1]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA; email: peter.warburton@mssm.edu, robert.sebra@mssm.edu

[2]Center for Advanced Genomics Technology, Icahn School of Medicine at Mount Sinai, New York, NY, USA

[3]Black Family Stem Cell Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA

[4]Icahn Genomics Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA

Annu. Rev. Genom. Hum. Genet. 2023. 24:109–32

First published as a Review in Advance on April 19, 2023

The *Annual Review of Genomics and Human Genetics* is online at genom.annualreviews.org

https://doi.org/10.1146/annurev-genom-101722-103045

Keywords

long-read sequencing, structural variants, repetitive DNA, pathogenic mutations, epigenetic modifications

Abstract

DNA sequencing has revolutionized medicine over recent decades. However, analysis of large structural variation and repetitive DNA, a hallmark of human genomes, has been limited by short-read technology, with read lengths of 100–300 bp. Long-read sequencing (LRS) permits routine sequencing of human DNA fragments tens to hundreds of kilobase pairs in size, using both real-time sequencing by synthesis and nanopore-based direct electronic sequencing. LRS permits analysis of large structural variation and haplotypic phasing in human genomes and has enabled the discovery and characterization of rare pathogenic structural variants and repeat expansions. It has also recently enabled the assembly of a complete, gapless human genome that includes previously intractable regions, such as highly repetitive centromeres and homologous acrocentric short arms. With the addition of protocols for targeted enrichment, direct epigenetic DNA modification detection, and long-range chromatin profiling, LRS promises to launch a new era of understanding of genetic diversity and pathogenic mutations in human populations.

OVERVIEW

DNA sequencing and the identification of disease-causing mutations and human genetic variation has revolutionized medicine over recent decades. Since the first report of a sequenced human genome in 2001 (52), the human reference genome has been the basis for functional annotation and biomedical interpretation, providing an invaluable resource for basic science and clinical genetic diagnostics today. The current human reference genome, hg38, is highly fragmented and contains 349 discontinuous gaps and unresolved portions, and although incremental improvements have been made by way of closing and spanning these gaps, adding alternate genomic loci or patches, and improving the annotations (99), it continues to have at least ~151 Mbp of unknown sequence or gaps, including pericentromeric and subtelomeric regions, acrocentric short arms, and large repeat arrays (81).

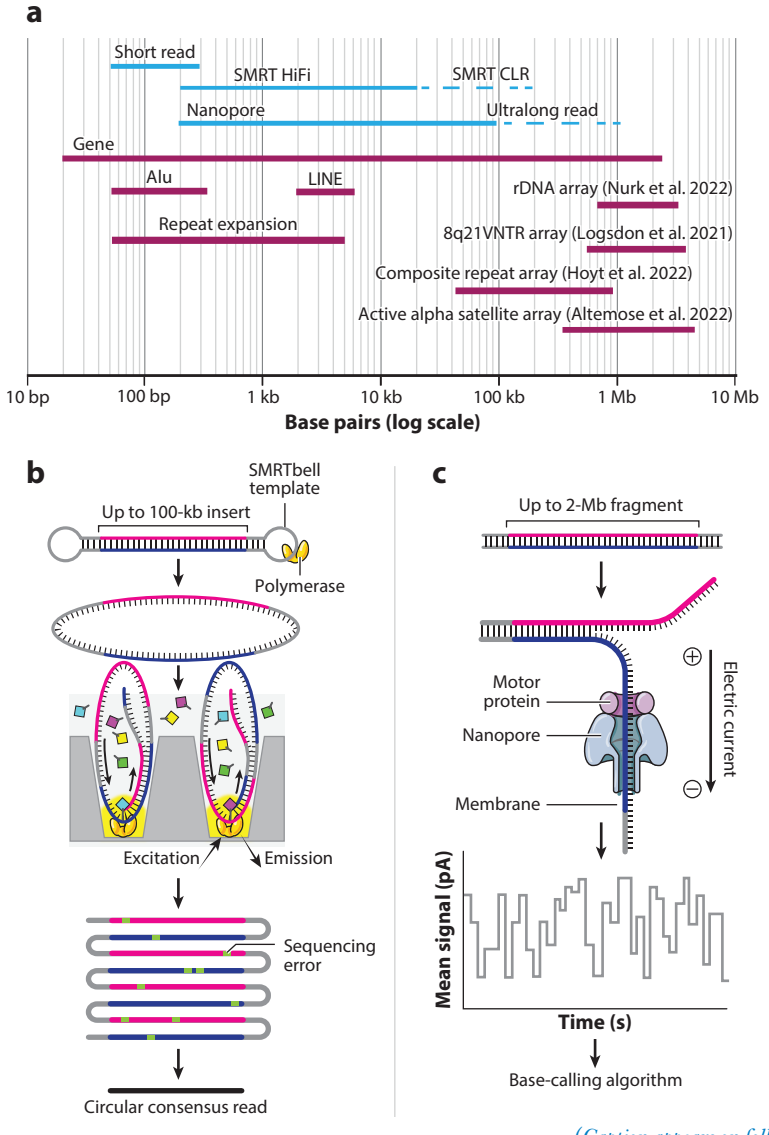
Over time, functional annotation and biomedical gene discovery have relied primarily on comparison with the current reference genome build. The widespread use and relative affordability of this class of next-generation sequencing technologies, which are dominated by short-read sequencing-by-synthesis methods, have led to a large number of scientific discoveries in genetics, evolution, and disease in humans. While extremely useful, the read length of short-read sequencing (SRS), on the order of 100–300 bp per read, can result in the loss of resolution of genomic regions that are not uniquely spanned by overlapping reads of this size. This limitation is particularly true for low-complexity repetitive loci, duplicated regions, tandem arrays, and complex structural variants, which collectively make up the majority of the gaps and missing sequence. Long-read sequencing (LRS) can sequence through DNA fragments that are orders of magnitude longer than the fragments sequenced using SRS (**Figure 1*a***), with highly accurate routine sequencing of 10–20-kbp fragments and lower-accuracy reads up to hundreds of kilobase pairs using single-molecule, real-time (SMRT) sequencing by synthesis. Even longer reads of up to 1–2 Mbp are available using nanopore-based electronic direct sequencing, although these longer reads typically have lower accuracy.

LRS has emerged as critical in overcoming the limitations imposed by SRS technology and has revealed a wide variety of previously underannotated genomic characteristics, such as repeat arrays and structural variants. Long-read technology will continue to facilitate resolution of genomic regions previously considered intractable until we collectively reveal the full spectrum of human genetic variation (58). In this review, we concentrate on LRS technology development and applications and reference a spectrum of studies that have applied LRS data to human genomics and disease, although LRS has also been widely applied to bacterial genomes and other systems, including plants, invertebrates, and other vertebrates (6). Collectively, LRS has provided major advancements in human genomics, from the identification of rare disease-causing structural variants such as insertions, deletions, and inversions to recently using a combination of LRS technologies to provide the first examples of gapless telomere-to-telomere complete human chromosomes (59, 67) and a complete haploid human genome (81).

LONG-READ SEQUENCING TECHNOLOGY

There are currently two major long-read technologies: SMRT sequencing platforms (**Figure 1*b***) and nanopore-based sequencing platforms (**Figure 1*c***). SMRT sequencing (developed primarily by PacBio) is a single-molecule sequencing-by-synthesis technology. High-molecular-weight genomic, double-stranded DNA is size selected and constructed into SMRTbell template libraries with single-stranded closed hairpin adapters ligated to the ends. Primers annealed to the hairpin adapters permit directed DNA polymerization around the single-stranded, topologically circular SMRTbell. Individual SMRTbells are distributed into an array of up to 8 million zero-mode

waveguides (ZMWs) that each contain an immobilized DNA polymerase for real-time sequencing within a restricted observation volume for improved signal-to-noise detection of each individual real-time, fluorescently labeled nucleotide incorporation event. These circular SMRTbell templates enable several passes on the immobilized DNA polymerase, giving continuous long reads (CLRs) that are used to generate subreads that represent multiple reads of the genomic DNA molecule in both directions. Earlier versions of SMRT CLR sequencing used very large insert sizes of up to hundreds of kilobase pairs, and thus the circular template was sequenced only a few times (on the order of 1–5 times, depending on the insert size and collection time), which resulted in a higher error rate and sequencing accuracy of 85–92% and limited the early utility of this technology.



(*Caption appears on following page*)

Figure 1 (*Figure appears on preceding page*)

LRS technologies and relative scales. (*a*) Relative read lengths across sequencing technologies and size ranges of relevant variant structural elements, shown using a log scale. VNTR arrays are from 8q21.2 and show the range of array sizes in the population. (*b*) SMRT HiFi sequencing. This technique uses dumbbell-shaped templates called SMRTbells, which are sequenced as single molecules within an array of ZMWs that each contain a DNA polymerase. An observation volume restricted to the bottom surface of the ZMW improves signal-to-noise detection of each nucleotide incorporation event using fluorescent signals specific for each nucleotide. The circular templates are sequenced multiple times, and a circular consensus sequence is obtained that is used for intramolecular error correction of the random polymerase errors while processing each single pass of the SMRTbell library molecules. (*c*) Nanopore LRS, using Oxford Nanopore Technologies as an example. This technique uses a linear DNA template loaded onto a motor protein and a molecular nanopore. As the DNA unwinds and is driven through the nanopore, characteristic changes in electric current allow the sequence of the resident bases to be determined using a trained base-calling algorithm. Abbreviations: CLR, continuous long read; HiFi, high fidelity; LINE, long interspersed nuclear element; LRS, long-read sequencing; SMRT, single molecule, real time; VNTR, variable number of tandem repeats; ZMW, zero-mode waveguide. Panels *b* and *c* adapted from illustrations by Jill K. Gregory with permission from Mount Sinai Health System; copyright 2022 Mount Sinai Health System.

A major improvement to SMRT sequencing came with the introduction of high-fidelity (HiFi) sequencing, which represents the first LRS technology to provide sequence lengths on the order of 10–20 kbp with an accuracy approaching 99.99% (Q40). Because the genomic DNA fragments are selected to be between 15 and 20 kbp, the SMRTbell can progress through the DNA polymerase multiple times (on the order of 7–12 times or more), providing multiple subreads that are computationally combined via the circular consensus sequencing algorithm to provide highly accurate sequencing of each insert (129) (**Figure 1*b***). HiFi reads have provided improved structural variant discovery and have deciphered some of the underannotated repetitive regions of the genome, such as those found at centromeres and acrocentric short arms (59, 81).

The other major LRS technology is nanopore-based LRS (**Figure 1*c***). A linear double-stranded DNA template, which can be as long as several megabase pairs, is ligated to an adapter that is preloaded with a motor protein and then loaded onto an array of nanopores. The motor protein unwinds the DNA, and a single strand of the molecule is driven through the nanopore by an electric current. The electric current is characteristically disrupted by combinations of 3–7 bp of DNA sequence as the templates pass through individual nanopores, producing a squiggle (**Figure 1*c***) that is translated into DNA sequence based on computational base-calling algorithms (91). Standard nanopore (Oxford Nanopore Technologies) sequencing can sequence molecules that are an order of magnitude longer than those sequenced by SMRT technology, routinely tens to hundreds of kilobase pairs with an accuracy of approximately 87–98%, although improvements in base-calling algorithms and chemistry are increasing this accuracy (91, 124), with current advances achieving 99% or higher accuracy. Nanopore sequencing can also be used to generate ultralong reads, which are generally much larger than 100 kbp and can occasionally be as long as several megabase pairs (46) and are especially useful as genome scaffolds in the assembly of large, difficult-to-sequence regions. Ultralong reads have permitted tremendous access to long repetitive regions, as demonstrated for the Y chromosome centromeric region (47).

Another application is to use nanopores for the electrophoretic loading of ZMW arrays, which greatly increase loading efficiency over the diffusion loading currently used in SMRT sequencing. Thus, hybrid ZMW–nanopore LRS can reduce the DNA input required by SMRT sequencing by orders of magnitude, alongside the reduction of any bias toward loading of smaller molecules into ZMWs (29, 53). An interesting combination of SMRT and nanopore-based technologies is LRS by synthesis using a nanopore-coupled DNA polymerase with distinct polymer-tagged nucleotides, which has allowed real-time electronic DNA sequencing without the need for complex optics and

more straightforward electronic base calling (31, 110). These well-established and emerging LRS platforms exemplify the continuous improvements available for genomics research that provide the opportunity to investigate genome organization and variation at an unprecedented scale.

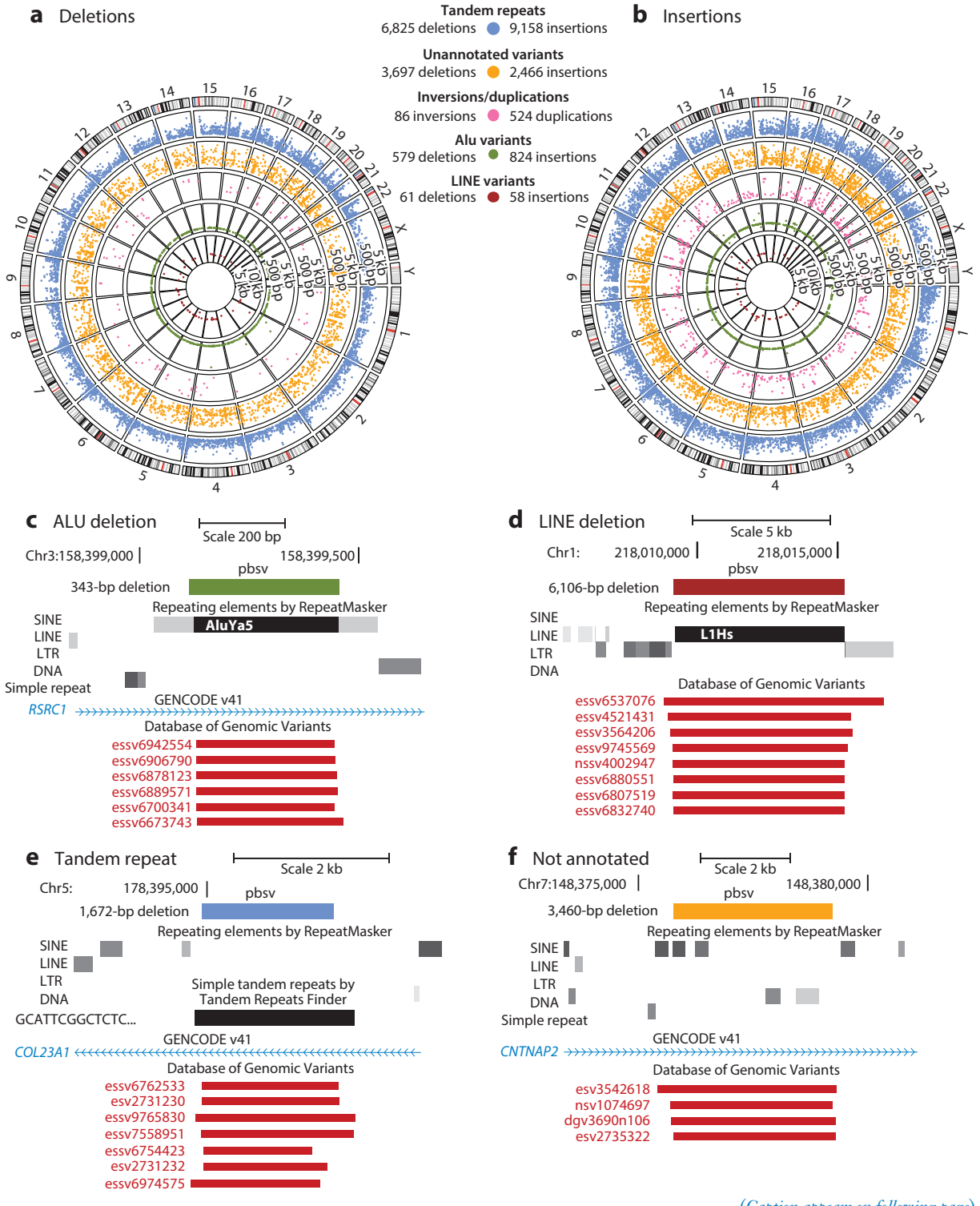
IMPROVING HUMAN GENOME ASSEMBLIES USING LONG-READ SEQUENCING TECHNOLOGIES

The most widely used reference genome (currently hg38) includes data constructed by sequencing a collection of large-insert bacterial artificial chromosome (BAC) clones from several individuals; thus, at any individual genomic location it represents a single human haplotype (58), although an additional 434 patches and alternative haplotypes are available in some variable regions of the genome (21, 78). An important consideration in these reference builds is that BAC cloning leads to an underrepresentation of repetitive sequences and the use of multiple BAC libraries from different individuals can result in mosaicism of haplotypes. While useful, this strategy results in incompatible structural polymorphisms or large repetitive regions flanking many of the unresolved sequence gaps (27).

LRS has been instrumental in improving genome resolution and understanding the full spectrum of genetic variation. Long-read SMRT sequencing has been used along with sequencing-by-synthesis short-read data, genome-mapping scaffolds, and reference-based approaches to improve assembly and understanding of the architecture of a diploid human genome, revealing large structural differences from the reference genome, including 68 Mbp of novel content relative to the NA12878 genome (87). In addition, LRS of trios has been used to analyze structural variation in diploid genomes in a haplotype-resolved manner, resulting in a three-to-sevenfold increase in structural variant detection over most other high-throughput sequencing studies (16, 26, 125). The Genome in a Bottle Consortium has used multiple platforms, including LRS, to establish structural variant benchmarks in multiple reference genomes from diverse populations in order to improve the reliability of variant calls across platforms to provide more robust references for a multitude of haplotypes (125, 136, 137). The major structural variant alleles from 15 human genomes have also been characterized, which has allowed the discovery of many fixed and common structural alleles that were missing from the human reference genome (7). Such work has contributed to the Database of Genomic Variants, which provides an important reference for the position, size, and population frequency of more common structural variations (60).

Long-read SMRT sequencing has also been applied to sequence haploid genomes, revealing a large number of genetic variants and demonstrating the advantage of providing baseline haploid genomes derived from hydatidiform mole cell lines, which retain only a single set of homologous chromosomes due to either fertilization of an enucleated egg or the subsequent loss of the maternal complement postfertilization (45). These hydatidiform mole references therefore represent a functionally haploid equivalent of the human genome lacking allelic variation, in order to remove the complexity of diploidy from sequencing and assembly of genomes (15, 42, 120).

To illustrate the range and depth of variants seen using current LRS data, we used the PacBio structural variant (pbsv) caller to call variants on a publicly available 30-fold coverage NA24385 human HiFi dataset (**Figure 2**). The tracks on the illustrated Circos plots in **Figure 2*a*,*b*** are based on the annotations called by pbsv. Tandem variants (**Figure 2*a*,*b***) are those involving simple repeats, such as those overlapping with Tandem Repeats Finder (12) and/or simple satellite repeats in RepeatMasker (104) (**Figure 2*e***). Other annotations easily observed are the insertion or deletion of transposable elements, such as long interspersed nuclear elements (LINEs) (**Figure 2*d***), which are predominantly around 6 kbp in size (**Figure 2*a*,*b***), and short interspersed nuclear elements (SINEs) (**Figure 2*c***), which are predominantly around 350 bp in size (**Figure 2*a*,*b***). Precise



(*Caption appears on following page*)

Figure 2 (*Figure appears on preceding page*)

Example variation detected using LRS. (*a*,*b*) Circos plots of deletions (panel *a*) and insertions (panel *b*) relative to the hg38 reference, showing variants and annotations ranging from 50 bp to 15 kbp as called by the pbsv tool using SMRT HiFi sequence data (85). Variant coordinates are shown relative to chromosomal location (*outside track*), and variant sizes are shown on a $\log_2$ scale within each Circos track. From the outside to the inside, the colored tracks show variants involving tandem repeats (*blue*), unannotated variants (*orange*), inversions (*pink* in panel *a*), duplications (*pink* in panel *b*), Alu insertions or deletions (*green*), and LINE insertions and deletions (*red*). Counts for each type of variant are shown between the plots. (*c–f*) Examples of deletions relative to the hg38 reference in the UCSC Genome Browser. For each variant, tracks shown from top to bottom include the pbsv variant caller track, RepeatMasker track, GENCODE v41 track (when applicable), simple repeat track (when applicable), and partial data from the Database of Genomic Variants. Panel *c* shows a 343-bp deletion of a full-length AluYa5 element (*green*) in intron 6 of the *RSRC1* gene; panel *d* shows a 6,106-bp deletion of a full-length L1Hs element (*red*); panel *e* shows a 1,672-bp deletion of 19 repeats from a 2,026-bp array of 23 tandem 88-bp G-rich repeats (*blue*); and panel *f* shows a 3,460-bp deletion (*orange*), not further annotated by pbsv, in intron 21 of the *CNTNAP2* gene. Abbreviations: HiFi, high fidelity; LINE, long interspersed nuclear element; LRS, long-read sequencing; LTR, long terminal repeat; pbsv, PacBio structural variant; SINE, short interspersed nuclear element; SMRT, single molecule, real time; UCSC, University of California, Santa Cruz.

insertions or deletions of full-length LINEs and SINEs with little or no surrounding nontransposon sequence illustrate that these are likely to represent transposition events. A precise transposon deletion observed in a demonstrative genome dataset represents a transposition event leading to an insertion in the reference genome relative to the sequenced genome. This is supported by the likelihood that most of these precise insertions and deletions are relatively young elements, such as AluY and L1Hs or L1PA2.

Additional types of variants shown are inversions (**Figure 2*a***) and duplications (**Figure 2*b***), both of which are often associated with known segmental duplications (8). Furthermore, there are many insertions and deletions that are not annotated by pbsv (**Figure 2*a*,*b*,*f***), suggesting that they are not associated with any of the above sequence characteristics. We should point out that most of the variants detected by long-read HiFi whole-genome sequencing (WGS) demonstrative data are common and have been compiled in previous databases, such as the Database of Genomic Variants (60) (see **Figure 2*c–f***). However, this database was originally curated using specialized, population-based, and often high-effort applications, including copy-number-variant detection using comparative genomic hybridization or single-nucleotide polymorphism arrays (19), size-selected and fosmid-based end-to-end sequencing (50, 51), and short-read sequencing-by-synthesis techniques such as read depth analysis, read pair analysis based on abnormally mapping read end pairs, or sequence assembly with unmapped read pairs (1, 73). Recently, higher-coverage whole-genome SRS of 602 trios coupled with improved variant-calling algorithms has expanded the discovery of insertions/deletions and structural variants in human populations (14). However, the advent of LRS has greatly simplified the analysis, reliability, and annotation of these structural variants and permitted their identification in individual genomes, especially in regions with repetitive DNA or segmental duplications (135), expanding the combined knowledge base while also informing any individual genome. Furthermore, rare structural variants that may be protective or pathogenic are also amenable to discovery in individual genomes because of LRS technologies that enable diagnostic and risk association legacies previously not achievable with short-read WGS data.

Additional high-resolution technologies have been developed to resolve structural variants. These include long-fragment-read (88) and linked-read sequencing technologies, both of which use SRS to sequence the ends of longer genomic DNA molecules, enabling long-range sequence and haplotype phasing over large segments. With Hi-C technology, protein/DNA complexes are cross-linked in vitro, where DNA within complexes is digested and ligated, which links sequences that are in close proximity in the nucleus. Since the contact frequency is related to distance, usually on the same chromosome, the sequence of the ligated fragments provides long-distance and

haplotypic data that can be used to link contigs and provide long-range sequence information (56, 83). Computational analysis methods have also been developed to use Hi-C data for structural variant calling (22, 121, 123). Optical genome mapping is another technique where high-molecular-weight genomic DNA is captured in a microfluidic chamber and subject to sequence-specific nicking and subsequent fluorescent labeling, providing long-range readout of large genome fragments with preserved genomic order that can permit resolution of up to several megabase pairs when compared with a reference genome map. Optical genome mapping has been used to anchor scaffolds for LRS or SRS WGS assemblies, identify large structural variations, and even detect chromosomal aberrations (39, 131, 134). While neither Hi-C nor optical genome mapping generates single-base-resolved long-read DNA sequence, their use provides scaffolding and long-range sequence information that is critical for accurate genome assembly (55, 81).

TARGETED LONG-READ SEQUENCING APPROACHES AND APPLICATIONS

While whole-genome LRS can be used to detect structural variations and complex sequence regions that are not detectable by short-read technologies, comprehensive LRS of whole human genomes remains prohibitively expensive, especially in a clinical setting. One approach to enhance the utility of LRS is targeted enrichment of specific genomic loci. Targeted approaches permit multiplexing of specific clinically and/or scientifically interesting regions from batches of samples to increase the efficiency of LRS, with the goal of obtaining higher sequence coverage and/or enabling localized assembly or phasing of relevant genomic variants that would otherwise be prohibitively expensive using whole-genome LRS methods. One approach makes use of standard or long-range PCR amplification of the targeted region to generate SMRTbell libraries from the resulting amplicons for higher-throughput LRS. LRS amplicon sequencing has proven valuable for phasing double mutations in *cis* in the *PIK3CA* gene as a regulatory target in breast cancer (118), performing fully phased haplotyping of the *NUDT15* gene (100), investigating *CYP2D6* pharmacogenetic genotypes (89), sequencing the *MUC1* variable number of tandem repeats (VNTR) region in autosomal dominant tubulointerstitial kidney disease (130), and analyzing a 72.8-kbp deletion encompassing *BBS9* (implicated in Bardet–Biedl syndrome) and *RP9* (implicated in retinitis pigmentosa) (93). An additional approach that can target larger regions of the genome uses biotinylated oligonucleotides that are homologous to the region of interest, which can be designed to tile across up to 1 Mbp of DNA or from multiple regions of the genome. The genome is fragmented and ligated to barcoded adapters, PCR preamplified, and annealed to the hybridization probes; the target DNA is purified using streptavidin beads and then PCR amplified again to enrich the yield; and SMRTbell libraries are then sequenced (108). This targeted enrichment approach has been used to investigate hepatitis B virus integrations in hepatitis B patients (116).

However, protocols that include PCR amplification steps pose some technical limitations, especially when amplifying arrays of short tandem GC-rich repeats. Limitations to PCR amplification also include target duplication, size limitations, template switching, and/or loss of epigenetic modifications. Therefore, the Cas9-assisted targeting of chromosome segments (CATCH) approach provides an amplification-free way to target genomic DNA and capture regions spanning hundreds of kilobase pairs (32). CATCH uses Cas9-targeted fragmentation of high-molecular-weight DNA in vitro, followed by separation and purification of the targeted region from the remaining genome using pulsed field gel electrophoresis, followed by LRS. CATCH allows several-hundred-fold enrichment of the targeted region over WGS and has been developed for both SMRT and nanopore-based LRS. CATCH has been used to target the 80-kbp *BRCA1* gene and surrounding regulatory regions on 200-kbp fragments, which were enriched 237-fold and enabled 70-fold

sequencing coverage. Although the low accuracy of the LRS used at the time reduced the ability to confidently call single-nucleotide polymorphisms compared with SRS, the ability to phase haplotypes and examine structural variants underscores the value of the technique (32). CATCH is particularly valuable for unstable short repeat copy expansions and has been used to sequence across an entire expanded 7,941-bp CCCCGG array in the *C9orf72* gene of a patient with amyotrophic lateral sclerosis (25, 36). Myotonic dystrophy type 1 is caused by an unstable CTG expansion up to 4,000 repeats in the *DMPK* locus, which was initially sequenced using PCR amplification and SMRT sequencing (61), but CATCH was subsequently used to improve the accuracy of the repeat expansion analysis (113). Moreover, CATCH and nanopore LRS have recently been used to isolate extrachromosomal DNA from human-derived cancer lines that contained amplified select enhancers or oncogene coding sequences (43).

An alternative approach called targeted LRS uses adaptive sampling by selectively sequencing DNA molecules that reside in a predefined genomic region. After approximately 500 bp of sequence has been called in real time and computationally aligned to a reference genome, if the region is within the predefined region, then the sequencing continues; if this criterion is not met, then the molecule is ejected from the pore by reversing the current, and a new molecule is selected and sequenced. Targeted LRS adaptive sampling has been used to identify previously undetected pathogenic variants in several diseases, including Werner syndrome and pseudohypoparathyroidism (70–72). Nanopore Cas9-targeted sequencing has also been used, where Cas9 guide RNAs cleave a targeted region in the genome, the adapters are ligated, and the library is subjected to nanopore LRS (37). The targeted LRS and nanopore Cas9-targeted sequencing approaches can be combined for further improved enrichment, such as for *CYP2D6–CYP2D7* hybrid allele genotyping (95). Collectively, these enrichment protocols for both SMRT and nanopore LRS allow specific enrichment of genomic regions of interest, which will greatly reduce the per-sample cost and enhance the utility for LRS in local genotyping and phasing of specific single or ensemble structural variants for genetic diagnostics in cases where such rare structural variants are the cause of disease or require downstream validation.

EPIGENOMICS USING LONG-READ SEQUENCING APPROACHES

LRS also has advantages in detecting epigenetic modifications found on native genomic DNA, including chemical nucleotide modifications of cytosine or adenine methylation (38). DNA methylation is most often examined using bisulfite deamination of DNA, where unmethylated cytosines are converted to uracil followed by sequencing and informatic comparison with untreated DNA controls. SMRT bisulfite sequencing has directly sequenced the products of bisulfite deamination up to 1.5 kbp in size using LRS without the need for a cloning step (132). In more recent LRS applications, modified or methylated nucleotides are detected directly in the native genomic DNA because the epigenetic chemical moieties characteristically modify the sequencing signal to provide a significant detection mechanism with sufficient signal to noise against the sequencing baseline. For example, in SMRT LRS, methylated nucleotides perturb the kinetics of base-pair incorporation by the polymerase, resulting in an increased interpulse duration and pulse width of the fluorescent signal between nucleotides as they are sequenced. The methylation status of cytosine and adenine nucleotides are effectively read in real time as the polymerase synthesizes from the SMRTbell template and encounters modified bases (28, 30). The surrounding sequence context affects the changes in interpulse duration and pulse width and is most accurate when multiple circular consensus sequencing passes are achieved, impacting the maximum size that can be sequenced, depending on the epigenetic target (11). 5-Methyl-CpG (5mCpG) is the most common methylation site in mammalian genomes and important in gene regulation and development

(105). 5mCpG detection using SMRT sequencing has been the most explored computationally (80, 114). When electronic nanopore LRS is used, modified bases introduce deviations to the readout squiggle that can be decoded using a pretrained algorithm and hidden Markov models. For example, Nanopolish (90, 101) and additional algorithms have been trained to detect 5mCpG and other modified bases (38).

Bisulfite deamination treatment often results in fragmentation of DNA and thus is less suitable for LRS applications sequencing DNA fragments tens of kilobase pairs in size. Therefore, LRS has been coupled with other chemical treatments that do not cause DNA fragmentation and convert additional modifications that are difficult to detect with altered readout of native DNA. TET enzyme oxidation has been used to convert both 5-methylcytosine (5mC) and 5-hydroxymethylytosine (5hmC) to dihydrouracil, which is recognized as thymine, thus permitting sufficient detection signal for 5hmC (57). Another treatment, called long-read enzymatic modification sequencing, uses selective enzymatic protection of both 5mC and 5hmC before deamination with APOBEC3A (112). Importantly, both of these techniques allow PCR amplification of the treated DNA with preservation of the deamination signal, reducing the required amount of starting material for LRS, although off-target modifications can be possible.

Emerging LRS-based chromatin-profiling methods, which present the opportunity to map chromatin organization across kilobase-length genomic regions and can include repetitive DNA elements such as centromeres and telomeres (23), have also recently been used as an epigenomics tool. Fiber-seq and single-molecule adenine methylated oligonucleosome sequencing assay (SAMOSA) techniques use a nonspecific DNA N6-adenine methyltransferase to reveal the chromatin architecture of kilobase-sized chromatin fibers on the underlying DNA by methylating accessible adenine residues, which are then detected using the changes in interpulse duration in SMRT LRS (3, 109). Fiber-seq has been used to map transcriptional regulatory elements and telomeric chromatin (23, 109). A related nanopore-based technique called nucleosome occupancy and methylome nanopore sequencing (nanoNOMe) uses a GpC (not to be confused with a CpG) methyltransferase to label endogenous GpCs in chromatin fibers and then leverages the Nanopolish base-calling algorithm with additional training to detect GpC methylation from nanopore squiggles. NanoNOMe has the advantage of simultaneously measuring both chromatin accessibility and endogenous 5mCpG; it has been used to examine CTCF binding sites (54) and, more recently, the phasing of gene regulatory and imprinting elements on fragments as large as 116 kbp (10). Using LRS to detect base modifications and chromatin structure provides the critical advantage of allowing mapping and phasing across kilobase-sized genomic alleles, which will continue to provide important functional insight into these epigenetic signatures on chromosomes.

PATHOGENIC VARIANT DISCOVERY USING LONG-READ SEQUENCING TECHNOLOGIES

The types of structural variation that have been revealed by LRS are usually those not easily resolvable by SRS, such as insertions, deletions, and inversions (135). Although many commonly occurring variants are included in structural variant databases (60), disease-causing pathological structural variants are by definition rare in populations and not easily detected by SRS. Many cases of rare large insertions and deletions relative to the human reference genome have been discovered using both whole-genome LRS and targeted capture methods. Studies have found 375- and 395-kbp deletions in *EYS* in patients with retinitis pigmentosum (97), a 72.8-kbp deletion encompassing *BBS9* and *RP9* in a patient with Bardet–Biedl syndrome (93), a 12.4-kbp deletion in *CLN6* in a patient with progressive myoclonic epilepsy (76), a 7.1-kbp deletion in one allele of *G6PC* (together with a c.326G>A mutation in the other allele) in a patient with glycogen storage

disease type Ia (66), a 2.814-kbp deletion overlapping exons in *PRKAR1a* in a patient with multiple neoplasia and cardiac myxomata (65), and a 4.6-kbp insertion in a family with benign adult familial myoclonus epilepsy (77). Insertions of previously unmapped sequence not present in the reference genome are particularly difficult to detect using SRS data mapping but are often plainly obvious in LRS data because individual and/or ensemble long reads often contain the entire novel inserted sequence in its genomic context. Yet another application for LRS is in human leukocyte antigen (HLA) typing that can span the entirety of the HLA class 1 genes and allow accurate haplotyping in this hypermorphic gene complex (63), which has contributed greatly to the expansion and known diversity of HLA database haplotypes (94).

Another structural variant subtype that is often revealed by LRS is that of repeat expansions, which are otherwise challenging to accurately sequence and characterize using conventional SRS data (74). LRS can often span the entire expansion, which can be on the order of many thousands of base pairs (33). Either WGS or a targeted enrichment approach can produce reads from specific regions known to contain repeats. Repeat expansions are a hallmark of many neurodegenerative diseases, such as expansion of the GGGGCC hexameric repeat in the *c9orf72* gene associated with 5–7% of amyotrophic lateral sclerosis cases (25, 36, 117). Repeat expansions are also common in spinocerebellar ataxia (98), neuronal intranuclear inclusion disease (106), and other diseases (18, 44). Pentanucleotide TTTCA and TTTTA expansions in *SAMD12* have been shown to cause familial cortical myoclonic tremor with epilepsy (133), while ATTCT and ATTCC expansions in *ATXN10* affect disease penetrance in spinocerebellar ataxia type 10 (79). Other pathogenic variants revealed by LRS include insertion of a cytosine into a VNTR region of 60-bp repeats in the *MUC1* gene, which has been identified as a causal mutation of autosomal dominant tubulointerstitial kidney disease (130). Duplications include a 27-bp duplication in the polyalanine tract of the *HoxD13* gene that leads to synpolydactyly (64) and a heterozygous 3.463-kbp duplication, including exon 30 of the *LAMA2* gene, that causes merosin-deficient muscular dystrophy (13). Breakpoints of a balanced chromosomal Xp11.1/20p13 translocation were precisely identified to the base pair using LRS in an individual with developmental disabilities (24). Novel inversions detected using LRS technology include a 5.9-Mbp inversion that disrupted exons 3–79 in the *DMD* gene of a previously undiagnosed male patient (and was also present as a heterozygote in the mother) (13) and a 12-kbp inversion that disrupted the first two exons of *BRPF1* and the last four exons of *CPNE9* on chromosome 3 in a patient with intellectual disability (75). These pathogenic structural variants were found primarily in an unbiased application of LRS. Indeed, the power of LRS to identify the causative genetic variant in rare undiagnosed diseases is rapidly emerging, including recent examples of large-cohort studies being deeply sequenced with SMRT HiFi whole-genome LRS (20, 62). Thus, the value of LRS for rare mutation detection in undiagnosed disease patients cannot be overstated (62). As application of LRS leads to improved reference genomes via the bridging of large gaps and unresolved genomic regions across the range of diverse human haplotypes, the ability to ascertain medically pathogenic variation will likely continuously reveal novel human disease and resilience association in future medicine.

LONG-READ SEQUENCING ENABLES THE FIRST TELOMERE-TO-TELOMERE HUMAN REFERENCE GENOME

A recent advance that demonstrates the true power of LRS is the first complete telomere-to-telomere sequence of several complete chromosomes (59, 67), which was then followed by the first complete human genome sequence (81). For ease of assembly, the Telomere-to-Telomere (T2T) consortium performed LRS data analysis of the haploid CHM13 genome, which is derived from a hydatidiform mole line and therefore represents a functionally haploid equivalent of

the human genome that lacks diploid allelic variation. The use of LRS of a haploid cell line eliminated the limitations of BAC-based assembly used in previously generated reference genomes and the complexity of structural polymorphisms within complete diploid genomes. However, previous technologies used to resolve human references for these genomes (SMRT CLR sequencing and nanopore-based ultralong-read sequencing) collectively had an error rate of greater than 5%, which did not permit assembly of larger, highly homologous tandem arrays. The increased accuracy provided by whole-genome SMRT HiFi circular consensus sequencing reads enabled the sequencing of subtly different but highly homologous tandem arrays. Nanopore-based ultralong reads are also used to span repeat arrays as contiguous scaffolds for anchoring the highly accurate HiFi WGS reads in fully resolved reference generation studies. This strategy was used to generate gapless sequence assemblies of chromosome 8 (59) and the X chromosome (67), including complete centromeres and other tandem arrays of previously unspanned repeats.

The T2T CHM13 haploid assembly was ultimately completed with a hybrid of data from various next-generation sequencing technologies, including 100-fold SRS data, 30-fold whole-genome HiFi LRS data, and 120-fold Oxford Nanopore Technologies–based ultralong nanopore reads, alongside optical genome mapping, Hi-C data, and single-strand sequencing (67, 96). This effort required a significant amount of manual curation by a large team over many months, with different groups focused on each chromosome (48). Thus, despite improvements, a multitude of integrated datasets across various LRS and SRS technologies is currently required to complete even a single haplotypic reference. This limitation in the genomics field will motivate the additional developments that are needed to automate assembly and phasing of diploid human genomes at high quality and at scale. Ultimately, future improvements will be critical for the adoption of LRS data in regulated discovery and validation of clinically relevant variants and a translated understanding of human genetic variation in everyday medical practice.

The members of the T2T consortium were collectively determined to create a complete gapless human genome that contained the large highly repetitive regions of the human genome, including large interspersed arrays of tandem repeats (41), all segmental duplications (119), and, remarkably, the complete span of the megabase-sized centromeric and pericentromeric repeat array in each chromosome (5), as well as the highly homologous short arms of acrocentric chromosomes (35). This research effort added nearly 200 Mbp (~8%) of previously hidden sequence to the human genome (4), including genomic regions representing large gaps that many thought intractable.

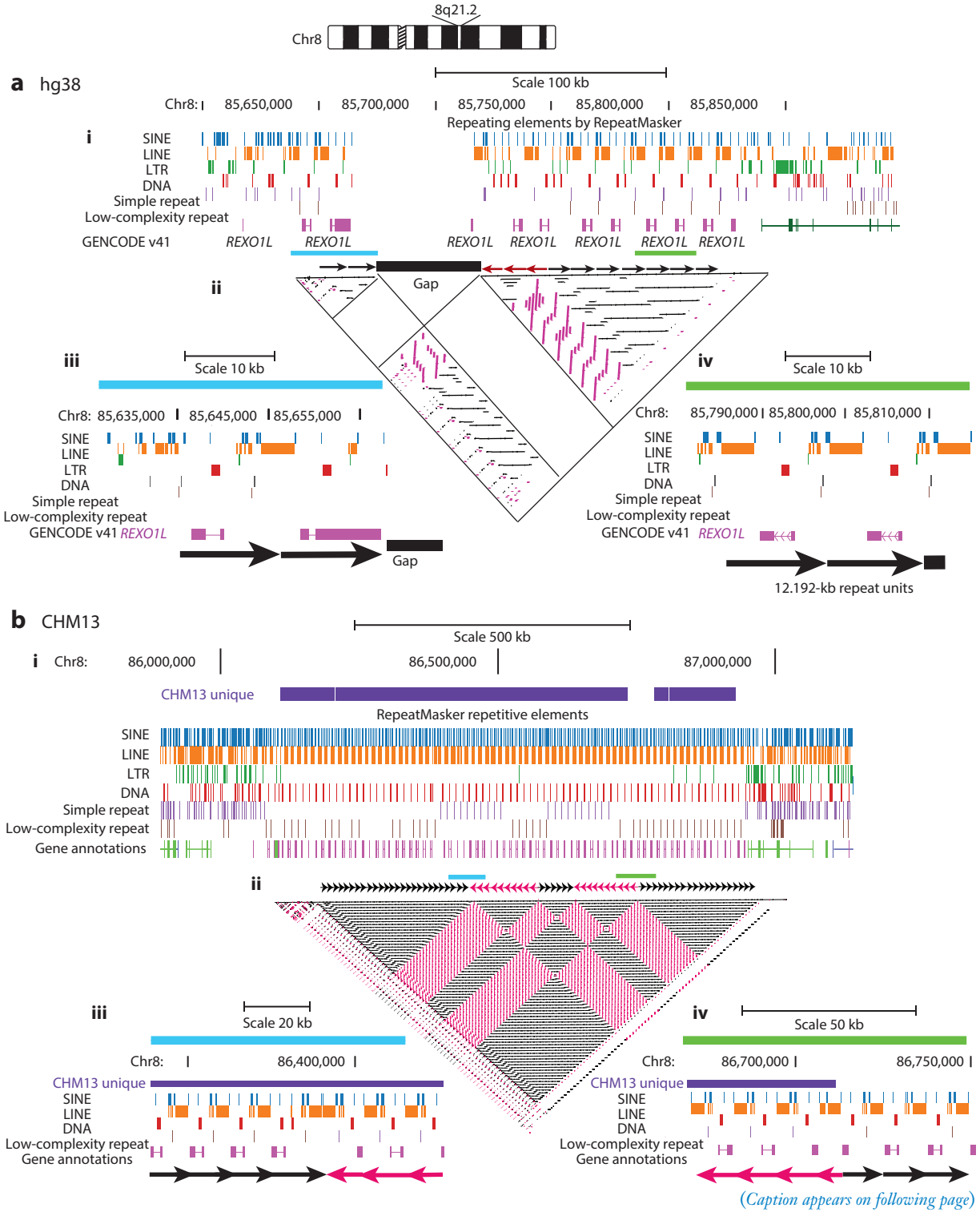
One of the most challenging types of DNA structure to sequence is satellite arrays, which consist of portions of the genome up to several kilobase pairs in size that have been tandemly duplicated into arrays as large as several megabase pairs (128). These arrays are inherently challenging to fully sequence and are poorly represented in reference genomes due to the size and high homology of the repeat units, their instability in BAC libraries from which reference genomes were constructed, and their high variability in repeat copy number and array size between individuals. There are many classes of satellite DNA in the human genome, including satellite, simple, and low-complexity repeats, as well as composite repeats that are more complex. Satellite repeats are generally short tandem copies of several base pairs; for example, classical human satellites 2 and 3 (HSat2 and HSat3), totaling 28.7 and 47.6 Mbp in CHM13, respectively, are derived from the simple repeat $(CATTC)_n$ and constitute the largest contiguous satellite arrays found in the human genome, including a 27.6-Mbp array on chromosome 9 in CHM13 (81). Ten such simple sequence satellite repeat types, including five previously unknown types, were classified in CHM13. Additional tandem DNA classes include ~68-bp beta satellite DNA, found in pericentromeric regions and on the acrocentric short arms, and 171-bp alpha satellite DNA, which is the predominant centromeric class of repetitive DNA.

A third class of tandemly repeated DNA is composite elements, which by definition contain at least three types of other repetitive elements, such as transposable elements, satellite repeats, or composite subunits. The T2T consortium described the repeat unit and array sizes for 19 of these composite elements, as well as their distinct epigenetic and transcriptional patterns, in the CHM13 genome. Most of these events are found in a single array, and eight contain protein-coding annotations (41). While the repeat unit of many of these composite satellites had been described more than a decade ago (128), several were found on either side of polymorphic gaps in the hg38 reference genome. With the application of recent LRS and assembly protocols developed by the T2T consortium, the exact size, copy number, repeat organization, and position of these arrays have been determined for CHM13 and other genomes. A polymorphic composite repeat on chromosome 8q21, referred to as the VNTR repeat (41), is found in a large array of ~12-kbp repeats, each of which includes the *REXO1L* gene and several types of interspersed repeats detected by RepeatMasker (128) (**Figure 3**). This repeat array can be visible cytogenetically (107, 115). In the hg38 reference genome, several repeats are visible spanning a gap (**Figure 3*a***), while in CHM13, this array has been fully sequenced and shows a complex arrangement of 73 repeats spanning almost 1 Mbp of DNA (59) (**Figure 3*b***). Genomic features like composite arrays likely represent the product of natural chromosomal evolution processes, such as unequal crossing over. Nonetheless, their functional significance remains unclear, especially given that the 8q21 VNTR composite repeat array is the site of ectopic neocentromere formation (40). The ability to sequence across and precisely locate these arrays may provide the means for their targeted deletion using CRISPR technology, to assess their effect on chromosome stability, homologous pairing, and epigenetics.

Moreover, the short arms of human acrocentric chromosomes 13, 14, 15, 21, and 22 are conspicuously missing from the recent human reference genome hg38. These arms contain the 45-kbp tandemly repeated rDNA genes, and during interphase they cluster in the nucleolus, where the rDNA genes are expressed. The large amount of repetitive and satellite DNA, along with the high degree of homology between acrocentric short arms, made them extremely challenging to sequence and assemble. With the exception of a small amount of pericentromeric p-arm material, these regions were largely excluded from reference genomes. With the advent of the approaches developed by the T2T consortium, these regions have now been sequenced for each acrocentric chromosome, which in CHM13 range in size from ~10 to 17 Mbp. Although variable in size, each acrocentric short arm follows a similar pattern, with inverted segmental duplications and arrays of acrocentric, HSat3, beta satellite, and HSat1 repeats surrounding the rDNA arrays, most of which are also seen in other regions of the genome. The organizational similarity between acrocentric satellite repeats is likely a result of the dynamic evolution of intra- and interchromosomal exchanges of acrocentric short arms in the nucleus (81).

Satellite arrays, including active alpha satellite arrays, HSat3 on chromosome 9, and beta satellite arrays on chromosome 1 and in the acrocentric regions, contain many inversions in CHM13 that were confirmed in other sample genomes (5) (**Figure 3**). Inversions can form extruded cruciform structures that may mimic regions of paired homologs, and such DNA that is both tandemly repeated and has inversions could form complex structures that may play a role in sister chromatid or homologous chromosome pairing (127). As more genomes are sequenced from telomere to telomere, generalizable patterns of organization may emerge that could be candidates for experimental functional testing using high-throughput CRISPR-based screening approaches. For example, a systematic dissection of the repeats on acrocentric short arms using CRISPR technology may reveal nucleolus-targeting sequences.

Human centromeres are characterized by megabase-sized arrays of tandemly repeated alpha satellite DNA, which were previously represented by large gaps or reference models (68) in



(Caption appears on following page)

Figure 3 (*Figure appears on preceding page*)

Gap filling using LRS at the 8q21.2 VNTR composite repeat. (*a*, *i*) UCSC Genome Browser hg38 chr8:85,600,000–85,900,000 (300 kbp), showing the region containing the VNTR composite repeat and an arbitrarily sized 50-kbp gap. (*ii*) Dot plot showing the repeat structure in the region. Black lines indicate tandem orientation, and red lines indicate inverted orientation. (*iii*) Detail of several 12-kbp repeats in the area indicated by the light blue line between subpanels *i* and *ii*, which border the gap. (*iv*) Detail of several repeats in the area indicated by the green line between subpanels *i* and *ii*, distal to the gap. (*b*, *i*) UCSC Genome Browser CHM13 chr8:86,000,000–87,000,000 (1 Mbp) showing the region containing the VNTR composite repeat, which has 73 copies of a 12-kbp repeat. CHM13 unique is additional sequence in the T2T assembly relative to hg38. (*ii*) Dot plot of the VNTR repeat array. Black lines indicate tandem orientation, and red lines indicate inverted orientation. (*iii*) Detail of several 12-kbp repeats and an inversion in the area indicated by the light blue line between subpanels *i* and *ii*. (*iv*) Detail of several 12-kbp repeats and an inversion in the area indicated by the green line between subpanels *i* and *ii*. Abbreviations: LINE, long interspersed nuclear element; LRS, long-read sequencing; LTR, long terminal repeat; SINE, short interspersed nuclear element; T2T, Telomere-to-Telomere; UCSC, University of California, Santa Cruz; VNTR, variable number of tandem repeats.

the human reference assembly. The lack of a linear assembly across human centromeres has inhibited the ability to describe sequence and epigenetic elements that provide centromere function (41). The sequencing across complete arrays of human centromeric alpha satellite arrays using the methods developed by the T2T consortium has revealed a high degree of substructure across the centromeric regions, alongside epigenetic activity that may regulate function. Human centromeres are characterized by layered expansions of alpha satellite repeat arrays, with the active kinetochore associated mainly with the most recent array. In addition, several unexpected types of organization have been observed in CHM13 centromeres, such as inversions, splitting of homogeneous alpha satellite arrays by insertions of other satellite DNAs, and deletions, which were confirmed to be polymorphic in 16 additional draft diploid assemblies, demonstrating that ongoing evolutionary forces are at work shaping the DNA at human centromeres (5, 69).

Centromeres are the critical chromosomal locus responsible for the correct segregation of chromosomes, which is epigenetically determined by the presence of specialized chromatin containing the histone H3 analog CENP-A (49, 111, 126). Ultralong nanopore reads were used to investigate the DNA methylation status across the recently emerged active alpha satellite arrays at centromeres in CHM13 (35, 67). This has revealed that most centromeres contain a small region of tens to hundreds of kilobase pairs of decreased methylation, called a centromere dip region (CDR), within the otherwise heavily methylated active alpha satellite array. The use of native chromatin immunoprecipitation or the cleavage under targets and release using nuclease (CUT&RUN) chromatin-mapping technique (103) showed that this methyl dip region is the location of the CENP-A chromatin (5, 35). Thus, this hypomethylated CDR represents a second epigenetic mark, in addition to the presence of CENP-A, that appears to be important for centromere formation and/or maintenance. To rule out that the CDR was a consequence of CHM13 being a cell type from early human development, these CDRs were confirmed in human reference genome HG002, a diploid terminally differentiated lymphoblastoid line. At least some chromosomes in some individuals showed exceptions, such as multiple CDRs or that CENP-A did not occupy the most recently emerged array of alpha satellite DNA. Studying these exceptions and variation in centromere positioning across many samples, across families, and across different tissues from the same individuals will eventually reveal the extent of this epigenetic plasticity in centromere localization (5). The position and size of the CDR within and between chromosomes may have profound implications for centromere function and chromosomal mitotic and meiotic stability (35).

The ability to generate de novo centromeres from transfected DNA sequence, permitting the construction of artificial human chromosomes, has long been a goal of centromere research. The discovery that the CDR domain of hypomethylated alpha satellite DNA is the true location of the CENP-A chromatin and kinetochore may suggest additional approaches to create artificial

chromosomes, perhaps by transfection of unmethylated or unmethylatable alpha satellite DNA constructs. If the CENP-A chromatin can be epigenetically established on the unmethylated DNA before it becomes methylated, then it will likely propagate and potentially lead to de novo centromere and thus artificial chromosome formation. Thus, the LRS genome-wide sequencing and associated epigenetic analyses across human centromeres have provided some important surprises that open many avenues of centromere structure and chromosome stability research and illuminate the power of LRS approaches in these otherwise perplexing regions of the human genome.

THE FUTURE PANGENOME

The Human Pangenome Reference Consortium (HPRC) has the goal of creating a collection of human diploid reference genomes that represents the true diversity of human genetics. This daunting task will be critical for biomedical research, where recognition of the importance of diversity inclusion in medical genomic studies is growing (102, 137). Pangenomics efforts will require a large team of experts in population genetics, computational biology, ethics, and especially hybrid and long-read genome sequencing technologies (69). A major goal of the current pangenome project is to provide complete telomere-to-telomere genomes using the sequencing technology and computational framework developed for the T2T genome assemblies, as these gapless assemblies will be the new standard in our complete understanding of variation in population-level genomics. The realization of a complete, albeit haploid, human genome provides a framework for better understanding some previously unknown gaps and difficult-to-sequence regions such as centromeres, acrocentric short arms, segmental duplications, and satellite arrays, and many findings have been confirmed in additional genomic data (81). However, the very nature of a pangenome project suggests that we will continuously discover additional unknowns and potentially novel genomic structure and organization that will reveal that the full range of human genomic diversity is much broader than previously anticipated (122).

Although the LRS technical approaches used for the T2T consortium can certainly be applied to diploid genomes, a major challenge will be the efficient haploid-aware assembly of these diploid genomes, especially in the difficult-to-sequence satellite and repetitive regions of the genome (122). Powerful graph-based approaches were used for the assembly of the complete T2T genome (9, 17, 34, 81, 82, 92). The HPRC team determined which combination of genome sequencing and automated assembly approaches gave the most accurate diploid genome assemblies and provided the highest-quality haploid-aware complete diploid assembly of HG002 (48). These assemblies were not quite telomere to telomere, with some missing assembly and unmapped contigs in centromeres and telomeric repeats, thus requiring additional development in these regions. The HPRC will initially use select samples from the 1000 Genomes Project, which provides a catalog of human variation from 26 populations that were consented for unrestricted data usage (2), but additional samples will be included from other sources to support the ambitious sampling and diversity goals of the HPRC, especially for rare variants (122). The initial goal of the HPRC is to produce high-quality genomes from 350 individuals (resulting in 700 diverse haplotypes) selected to maximize global diversity on which to build the foundation for the complete pangenome in the future. One important aspect of the HPRC is to address and incorporate research on ethical, legal, and social implications in order to include divergent and Indigenous populations and obtain correct permissions and inclusion of the populations both in the science and in the use and interpretation of the sequence data.

The realization of a complete pangenome dataset could revolutionize human genomics and understanding of variation and diversity on a scale not seen since the completion of the first human reference genome in 2001. While LRS technologies continue to improve the resolution of

genomic features discovered and cataloged in the HPRC efforts, consideration of a global data federation mechanism that promotes more seamless consent, sharing, and integration of diverse genomics data will be necessary to empower the amount of data necessary to fully compile and translate the true power of the future pangenome database.

REMAINING CHALLENGES

LRS has provided the ability to generate DNA sequence data with read lengths routinely in the tens to hundreds of kilobase pairs and as high as 1 Mbp using single-molecule approaches at various accuracies. Some primary challenges remain in improving read accuracy, variable throughput, and cost, which collectively still impede the routine utility of LRS in a high-throughput clinical setting. LRS throughput is highly sensitive to molecular damage during library preparation, which requires high-quality, unnicked high-molecular-weight DNA. Accurate size selection of high-molecular-weight DNA for library preparation is important as LRS technology can be biased toward sequencing of smaller, more rapidly diffusing molecules, which can lead to uneven read lengths due to loading biases. Furthermore, improvement of the read accuracy in some LRS approaches is clearly important for accurate diploid assembly as well as detecting rare pathogenic variants.

The manufacturers of both major LRS platforms have recently announced technical innovations that purport to overcome some major limitations, although at the time of writing, neither has been used in a peer-reviewed publication. PacBio has announced the Revio system (86), which increases the number of ZMWs per SMRT cell from 8 million to 25 million, with four independent 25-million-ZMW arrays on each machine and the run times reduced to 24 hours. This theoretically increases the throughput of SMRT LRS up to 12.5 times with a greatly reduced cost, making highly accurate human HiFi WGS genomes and highly multiplexed comprehensive long-read diagnostic panels widely available for basic research and potentially routine clinical applications. Alternatively, Oxford Nanopore Technologies has announced the K14 chemistry (84) with an updated enzyme and new nanopore design that routinely improves the sequencing accuracy to Q20 (99%), with options to reach Q24 (99.6%) or higher, partially overcoming the low-accuracy limitation of nanopore LRS. Thus, the emerging demand for higher-accuracy, higher-throughput LRS, as well as the competition between the two major LRS platforms, is continuously driving innovations to overcome technological limitations.

CONCLUDING REMARKS

LRS has emerged as a critical tool for the discovery of the myriad of complex variation in human genomes. It has driven the next generation of complete human reference genomes, which include previously intractable regions such as polymorphic gaps and large arrays of repetitive elements, especially at centromeres and acrocentric short arms. It also provides a straightforward method to discover rare pathogenic structural variants in basic and clinical research, and as costs decrease and throughput and accuracy increase, it is beginning to be adopted in regulated clinical settings. The high-resolution, fully phased long-read genomic data continue to motivate a more extensive database of pathogenic variants that assists in disease subtyping through novel genetic associations. Coupled with the ability to phase variants alongside epigenetic modifications and chromatin elements across large kilobase-sized regions, LRS will continue to provide unprecedented insight into chromatin structure, gene regulation, and disease state across the human genome. Additional developments in LRS-based, full-length, isoform-resolved bulk and single-cell transcriptomics integrated with whole-genome LRS and epigenetic data collectively advance the potential for

biomarker discovery on the human, tissue, or even cellular level and inform downstream functional genomics validation studies.

While LRS has transformed our understanding of the human genome, the LRS pangenome is the future of personalized medicine. The pangenome is only realized once individuals and clinicians have access to high-resolution next-generation genomes that will generate a data legacy where healthy and pathogenic associations will be unlocked as they are discovered and validated with time. Given the wide degree of complex variation among individual diploid genomes and across diverse human populations, LRS will very likely continue to be used independently or coupled with complementary hybrid scaffolding technologies for closing assembly gaps or revealing novel genomic content. Recent momentum to hurdle physical and computational LRS technology challenges will accelerate the eventual ability to produce a de novo assembled genome for any individual, without the need for a reference genome, permitting unprecedented understanding of personalized human genetics and genomics. We should also consider ways to build more extensive databases of whole-genome LRS and SRS data to federate diverse genomics datasets. As the pangenomics database grows, near-real-time functional associations will require both adoption of deep learning algorithms and community guidelines on how to expedite the translation of vast discoveries enabled by LRS-based omics technologies. Ultimately, the pace of LRS data generation and the diversity of novel variants uncovered continue to enable a clearer database of the human condition. The future will require this genomic resolution at the individual level to provide a truly personalized framework that unlocks information in time from a single data source for each individual condition.

DISCLOSURE STATEMENT

R.P.S. is a paid consultant with Sema4, Stamford, Connecticut, USA.

ACKNOWLEDGMENTS

The authors thank members of the Center for Advanced Genomics Technology for helpful discussions and Aishwarya Mandava for help with the Circos plots in **Figure 2**.

LITERATURE CITED

1. 1000 Genomes Proj. Consort. 2010. A map of human genome variation from population-scale sequencing. *Nature* 467:1061–73. Corrigendum. 2011. *Nature* 473:544
2. 1000 Genomes Proj. Consort. 2015. A global reference for human genetic variation. *Nature* 526:68–74
3. Abdulhay NJ, McNally CP, Hsieh LJ, Kasinathan S, Keith A, et al. 2020. Massively multiplex single-molecule oligonucleosome footprinting. *eLife* 9:e59404
4. Aganezov S, Yan SM, Soto DC, Kirsche M, Zarate S, et al. 2022. A complete reference genome improves analysis of human genetic variation. *Science* 376:eabl3533
5. Altemose N, Logsdon GA, Bzikadze AV, Sidhwani P, Langley SA, et al. 2022. Complete genomic and epigenetic maps of human centromeres. *Science* 376:eabl4178
6. Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. 2020. Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.* 21:30
7. Audano PA, Sulovari A, Graves-Lindsay TA, Cantsilieris S, Sorensen M, et al. 2019. Characterizing the major structural variant alleles of the human genome. *Cell* 176:663–75.e19
8. Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, et al. 2002. Recent segmental duplications in the human genome. *Science* 297:1003–7
9. Bankevich A, Bzikadze AV, Kolmogorov M, Antipov D, Pevzner PA. 2022. Multiplex de Bruijn graphs enable genome assembly from long, high-fidelity reads. *Nat. Biotechnol.* 40:1075–81

10. Battaglia S, Dong K, Wu J, Chen Z, Najm FJ, et al. 2022. Long-range phasing of dynamic, tissue-specific and allele-specific regulatory elements. *Nat. Genet.* 54:1504–13
11. Beaulaurier J, Zhang XS, Zhu S, Sebra R, Rosenbluh C, et al. 2015. Single molecule-level detection and long read-based phasing of epigenetic variations in bacterial methylomes. *Nat. Commun.* 6:7438
12. Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* 27:573–80
13. Bruels CC, Littel HR, Daugherty AL, Stafki S, Estrella EA, et al. 2022. Diagnostic capabilities of nanopore long-read sequencing in muscular dystrophy. *Ann. Clin. Transl. Neurol.* 9:1302–9
14. Byrska-Bishop M, Evani US, Zhao X, Basile AO, Abel HJ, et al. 2022. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* 185:3426–40.e19
15. Chaisson MJP, Huddleston J, Dennis MY, Sudmant PH, Malig M, et al. 2015. Resolving the complexity of the human genome using single-molecule sequencing. *Nature* 517:608–11
16. Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, et al. 2019. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.* 10:1784
17. Cheng H, Concepcion GT, Feng X, Zhang H, Li H. 2021. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* 18:170–75
18. Chintalaphani SR, Pineda SS, Deveson IW, Kumar KR. 2021. An update on the neurological short tandem repeat expansion disorders and the emergence of long-read sequencing diagnostics. *Acta Neuropathol. Commun.* 9:98
19. Coe BP, Witherspoon K, Rosenfeld JA, van Bon BW, Vulto-van Silfhout AT, et al. 2014. Refining analyses of copy number variation identifies specific genes associated with developmental delay. *Nat. Genet.* 46:1063–71
20. Cohen ASA, Farrow EG, Abdelmoity AT, Alaimo JT, Amudhavalli SM, et al. 2022. Genomic answers for children: dynamic analyses of >1000 pediatric rare disease genomes. *Genet. Med.* 24:1336–48
21. Dennis MY, Harshman L, Nelson BJ, Penn O, Cantsilieris S, et al. 2017. The evolution and population diversity of human-specific segmental duplications. *Nat. Ecol. Evol.* 1:69
22. Dixon JR, Xu J, Dileep V, Zhan Y, Song F, et al. 2018. Integrative detection and analysis of structural variation in cancer genomes. *Nat. Genet.* 50:1388–98
23. Dubocanin D, Sedeno Cortes AE, Ranchalis J, Real T, Mallory B, Stergachis AB. 2022. Single-molecule architecture and heterogeneity of human telomeric DNA and chromatin. bioRxiv 2022.05.09.491186. **https://doi.org/10.1101/2022.05.09.491186**
24. Dutta UR, Rao SN, Pidugu VK, VS V, Bhattacherjee A, et al. 2019. Breakpoint mapping of a novel de novo translocation t(X;20)(q11.1;p13) by positional cloning and long read sequencing. *Genomics* 111:1108–14
25. Ebbert MTW, Farrugia SL, Sens JP, Jansen-West K, Gendron TF, et al. 2018. Long-read sequencing across the C9orf72 'GGGGCC' repeat expansion: implications for clinical use and genetic discovery efforts in human disease. *Mol. Neurodegener.* 13:46
26. Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, et al. 2021. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* 372:eabf7117
27. Eichler EE, Clark RA, She X. 2004. An assessment of the sequence gaps: unfinished business in a finished human genome. *Nat. Rev. Genet.* 5:345–54
28. Fang G, Munera D, Friedman DI, Mandlik A, Chao MC, et al. 2012. Genome-wide mapping of methylated adenine residues in pathogenic *Escherichia coli* using single-molecule real-time sequencing. *Nat. Biotechnol.* 30:1232–39
29. Farhangdoust F, Cheng F, Liang W, Liu Y, Wanunu M. 2022. Rapid identification of DNA fragments through direct sequencing with electro-optical zero-mode waveguides. *Adv. Mater.* 34:e2108479
30. Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, et al. 2010. Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat. Methods* 7:461–65
31. Fuller CW, Kumar S, Porel M, Chien M, Bibillo A, et al. 2016. Real-time single-molecule electronic DNA sequencing by synthesis using polymer-tagged nucleotides on a nanopore array. *PNAS* 113:5233–38

32. Gabrieli T, Sharim H, Fridman D, Arbib N, Michaeli Y, Ebenstein Y. 2018. Selective nanopore sequencing of human BRCA1 by Cas9-assisted targeting of chromosome segments (CATCH). *Nucleic Acids Res.* 46:e87
33. Gall-Duncan T, Sato N, Yuen RKC, Pearson CE. 2022. Advancing genomic technologies and clinical awareness accelerates discovery of disease-associated tandem repeat sequences. *Genome Res.* 32:1–27
34. Garrison E, Siren J, Novak AM, Hickey G, Eizenga JM, et al. 2018. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat. Biotechnol.* 36:875–79
35. Gershman A, Sauria MEG, Guitart X, Vollger MR, Hook PW, et al. 2022. Epigenetic patterns in a complete human genome. *Science* 376:eabj5089
36. Giesselmann P, Brandl B, Raimondeau E, Bowen R, Rohrandt C, et al. 2019. Analysis of short tandem repeat expansions and their methylation state with nanopore sequencing. *Nat. Biotechnol.* 37:1478–81
37. Gilpatrick T, Lee I, Graham JE, Raimondeau E, Bowen R, et al. 2020. Targeted nanopore sequencing with Cas9-guided adapter ligation. *Nat. Biotechnol.* 38:433–38
38. Gouil Q, Keniry A. 2019. Latest techniques to study DNA methylation. *Essays Biochem.* 63:639–48
39. Hanlon VCT, Lansdorp PM, Guryev V. 2022. A survey of current methods to detect and genotype inversions. *Hum. Mutat.* 43:1576–89
40. Hasson D, Alonso A, Cheung F, Tepperberg JH, Papenhausen PR, et al. 2011. Formation of novel CENP-A domains on tandem repetitive DNA and across chromosome breakpoints on human chromosome 8q21 neocentromeres. *Chromosoma* 120:621–32
41. Hoyt SJ, Storer JM, Hartley GA, Grady PGS, Gershman A, et al. 2022. From telomere to telomere: the transcriptional and epigenetic state of human repeat elements. *Science* 376:eabk3112
42. Huddleston J, Chaisson MJP, Steinberg KM, Warren W, Hoekzema K, et al. 2017. Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res.* 27:677–85
43. Hung KL, Luebeck J, Dehkordi SR, Colon CI, Li R, et al. 2022. Targeted profiling of human extrachromosomal DNA by CRISPR-CATCH. *Nat. Genet.* 54:1746–54
44. Ishiura H, Shibata S, Yoshimura J, Suzuki Y, Qu W, et al. 2019. Noncoding CGG repeat expansions in neuronal intranuclear inclusion disease, oculopharyngodistal myopathy and an overlapping disease. *Nat. Genet.* 51:1222–32
45. Jacobs PA, Wilson CM, Sprenkle JA, Rosenshein NB, Migeon BR. 1980. Mechanism of origin of complete hydatidiform moles. *Nature* 286:714–16
46. Jain M, Koren S, Miga KH, Quick J, Rand AC, et al. 2018. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* 36:338–45
47. Jain M, Olsen HE, Turner DJ, Stoddart D, Bulazel KV, et al. 2018. Linear assembly of a human centromere on the Y chromosome. *Nat. Biotechnol.* 36:321–23
48. Jarvis ED, Formenti G, Rhie A, Guarracino A, Yang C, et al. 2022. Semi-automated assembly of high-quality diploid human reference genomes. *Nature* 611:519–31
49. Karpen GH, Allshire RC. 1997. The case for epigenetic effects on centromere identity and function. *Trends Genet.* 13:489–96
50. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, et al. 2008. Mapping and sequencing of structural variation from eight human genomes. *Nature* 453:56–64
51. Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, et al. 2007. Paired-end mapping reveals extensive structural variation in the human genome. *Science* 318:420–26
52. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* 409:860–921
53. Larkin J, Henley RY, Jadhav V, Korlach J, Wanunu M. 2017. Length-independent DNA packing into nanopore zero-mode waveguides for low-input DNA sequencing. *Nat. Nanotechnol.* 12:1169–75
54. Lee I, Razaghi R, Gilpatrick T, Molnar M, Gershman A, et al. 2020. Simultaneous profiling of chromatin accessibility and methylation on human cell lines with nanopore sequencing. *Nat. Methods* 17:1191–99
55. Levy-Sakin M, Pastor S, Mostovoy Y, Li L, Leung AKY, et al. 2019. Genome maps across 26 human populations reveal population-specific patterns of structural variation. *Nat. Commun.* 10:1025
56. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326:289–93

57. Liu Y, Cheng J, Siejka-Zielinska P, Weldon C, Roberts H, et al. 2020. Accurate targeted long-read DNA methylation and hydroxymethylation sequencing with TAPS. *Genome Biol.* 21:54
58. Logsdon GA, Vollger MR, Eichler EE. 2020. Long-read human genome sequencing and its applications. *Nat. Rev. Genet.* 21:597–614
59. Logsdon GA, Vollger MR, Hsieh P, Mao Y, Liskovykh MA, et al. 2021. The structure, function and evolution of a complete human chromosome 8. *Nature* 593:101–7
60. MacDonald JR, Ziman R, Yuen RK, Feuk L, Scherer SW. 2014. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res.* 42:D986–92
61. Mangin A, de Pontual L, Tsai YC, Monteil L, Nizon M, et al. 2021. Robust detection of somatic mosaicism and repeat interruptions by long-read targeted sequencing in myotonic dystrophy type 1. *Int. J. Mol. Sci.* 22:2616
62. Marwaha S, Knowles JW, Ashley EA. 2022. A guide for the diagnosis of rare and undiagnosed disease: beyond the exome. *Genome Med.* 14:23
63. Mayor NP, Robinson J, McWhinnie AJ, Ranade S, Eng K, et al. 2015. HLA typing for the next generation. *PLOS ONE* 10:e0127153
64. Melas M, Kautto EA, Franklin SJ, Mori M, McBride KL, et al. 2022. Long-read whole genome sequencing reveals HOXD13 alterations in synpolydactyly. *Hum. Mutat.* 43:189–99
65. Merker JD, Wenger AM, Sneddon T, Grove M, Zappala Z, et al. 2018. Long-read genome sequencing identifies causal structural variation in a Mendelian disease. *Genet. Med.* 20:159–63
66. Miao H, Zhou J, Yang Q, Liang F, Wang D, et al. 2018. Long-read sequencing identified a causal structural variant in an exome-negative case and enabled preimplantation genetic diagnosis. *Hereditas* 155:32
67. Miga KH, Koren S, Rhie A, Vollger MR, Gershman A, et al. 2020. Telomere-to-telomere assembly of a complete human X chromosome. *Nature* 585:79–84
68. Miga KH, Newton Y, Jain M, Altemose N, Willard HF, Kent WJ. 2014. Centromere reference models for human chromosomes X and Y satellite arrays. *Genome Res.* 24:697–707
69. Miga KH, Wang T. 2021. The need for a human pangenome reference sequence. *Annu. Rev. Genom. Hum. Genet.* 22:81–102
70. Miller DE, Hanna P, Galey M, Reyes M, Linglart A, et al. 2022. Targeted long-read sequencing identifies a retrotransposon insertion as a cause of altered *GNAS* exon A/B methylation in a family with autosomal dominant pseudohypoparathyroidism type 1b (PHP1B). *J. Bone Miner. Res.* 37:1711–19
71. Miller DE, Lee L, Galey M, Kandhaya-Pillai R, Tischkowitz M, et al. 2022. Targeted long-read sequencing identifies missing pathogenic variants in unsolved Werner syndrome cases. *J. Med. Genet.* 59:1087–94
72. Miller DE, Sulovari A, Wang T, Loucks H, Hoekzema K, et al. 2021. Targeted long-read sequencing identifies missing disease-causing variation. *Am. J. Hum. Genet.* 108:1436–49
73. Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, et al. 2011. Mapping copy number variation by population-scale genome sequencing. *Nature* 470:59–65
74. Mitsuhashi S, Frith MC, Mizuguchi T, Miyatake S, Toyota T, et al. 2019. Tandem-genotypes: robust detection of tandem repeat expansions from long DNA reads. *Genome Biol.* 20:58
75. Mizuguchi T, Okamoto N, Yanagihara K, Miyatake S, Uchiyama Y, et al. 2021. Pathogenic 12-kb copy-neutral inversion in syndromic intellectual disability identified by high-fidelity long-read sequencing. *Genomics* 113:1044–53
76. Mizuguchi T, Suzuki T, Abe C, Umemura A, Tokunaga K, et al. 2019. A 12-kb structural variation in progressive myoclonic epilepsy was newly identified by long-read whole-genome sequencing. *J. Hum. Genet.* 64:359–68
77. Mizuguchi T, Toyota T, Adachi H, Miyake N, Matsumoto N, Miyatake S. 2019. Detecting a long insertion variant in *SAMD12* by SMRT sequencing: implications of long-read whole-genome sequencing for repeat expansion diseases. *J. Hum. Genet.* 64:191–97
78. Morales J, Pujar S, Loveland JE, Astashyn A, Bennett R, et al. 2022. A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. *Nature* 604:310–15

79. Morato Torres CA, Zafar F, Tsai YC, Vazquez JP, Gallagher MD, et al. 2022. ATTCT and ATTCC repeat expansions in the ATXN10 gene affect disease penetrance of spinocerebellar ataxia type 10. *HGG Adv*. 3:100137
80. Ni P, Zhong Z, Xu J, Huang N, Zhang J, et al. 2023. DNA 5-methylcytosine detection and methylation phasing using PacBio circular consensus sequencing. bioRxiv 2022.02.26.482074. **https://doi.org/10.1101/2022.02.26.482074**
81. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, et al. 2022. The complete sequence of a human genome. *Science* 376:44–53
82. Nurk S, Walenz BP, Rhie A, Vollger MR, Logsdon GA, et al. 2020. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res*. 30:1291–305
83. Olivares-Chauvet P, Mukamel Z, Lifshitz A, Schwartzman O, Elkayam NO, et al. 2016. Capturing pairwise and multi-way chromosomal conformations using chromosomal walks. *Nature* 540:296–300
84. Oxford Nanopore Technol. 2022. The power of Q20+ chemistry. *Oxford Nanopore Technologies*. **https://nanoporetech.com/q20plus-chemistry**
85. PacBio. 2022. pbsv. *GitHub*. **https://github.com/PacificBiosciences/pbsv**
86. PacBio. 2022. Revio system. *PacBio*. **https://www.pacb.com/revio**
87. Pendleton M, Sebra R, Pang AW, Ummat A, Franzen O, et al. 2015. Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat. Methods* 12:780–86
88. Peters BA, Kermani BG, Sparks AB, Alferov O, Hong P, et al. 2012. Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature* 487:190–95
89. Qiao W, Yang Y, Sebra R, Mendiratta G, Gaedigk A, et al. 2016. Long-read single molecule real-time full gene sequencing of cytochrome P450-2D6. *Hum. Mutat.* 37:315–23
90. Rand AC, Jain M, Eizenga JM, Musselman-Brown A, Olsen HE, et al. 2017. Mapping DNA methylation with high-throughput nanopore sequencing. *Nat. Methods* 14:411–13
91. Rang FJ, Kloosterman WP, de Ridder J. 2018. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome Biol*. 19:90
92. Rautiainen M, Marschall T. 2020. GraphAligner: rapid and versatile sequence-to-graph alignment. *Genome Biol*. 21:253
93. Reiner J, Pisani L, Qiao W, Singh R, Yang Y, et al. 2018. Cytogenomic identification and long-read single molecule real-time (SMRT) sequencing of a *Bardet-Biedl Syndrome 9* (*BBS9*) deletion. *npj Genom. Med.* 3:3
94. Robinson J, Barker DJ, Georgiou X, Cooper MA, Flicek P, Marsh SGE. 2020. IPD-IMGT/HLA Database. *Nucleic Acids Res*. 48:D948–55
95. Rubben K, Tilleman L, Deserranno K, Tytgat O, Deforce D, Van Nieuwerburgh F. 2022. Cas9 targeted nanopore sequencing with enhanced variant calling improves *CYP2D6-CYP2D7* hybrid allele genotyping. *PLOS Genet*. 18:e1010176
96. Sanders AD, Falconer E, Hills M, Spierings DCJ, Lansdorp PM. 2017. Single-cell template strand sequencing by Strand-seq enables the characterization of individual homologs. *Nat. Protoc.* 12:1151–76
97. Sano Y, Koyanagi Y, Wong JH, Murakami Y, Fujiwara K, et al. 2022. Likely pathogenic structural variants in genetically unsolved patients with retinitis pigmentosa revealed by long-read sequencing. *J. Med. Genet.* 59:1133–38
98. Sato N, Amino T, Kobayashi K, Asakawa S, Ishiguro T, et al. 2009. Spinocerebellar ataxia type 31 is associated with "inserted" penta-nucleotide repeats containing $(TGGAA)_n$. *Am. J. Hum. Genet.* 85:544–57
99. Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen HC, et al. 2017. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* 27:849–64
100. Scott ER, Yang Y, Botton MR, Seki Y, Hoshitsuki K, et al. 2022. Long-read HiFi sequencing of *NUDT15*: phased full-gene haplotyping and pharmacogenomic allele discovery. *Hum. Mutat.* 43:1557–66
101. Simpson JT, Workman RE, Zuzarte PC, David M, Dursi LJ, Timp W. 2017. Detecting DNA cytosine methylation using nanopore sequencing. *Nat. Methods* 14:407–10

102. Sirugo G, Williams SM, Tishkoff SA. 2019. The missing diversity in human genetic studies. *Cell* 177:26–31
103. Skene PJ, Henikoff S. 2017. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *eLife* 6:e21856
104. Smit AF. 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr. Opin. Genet. Dev.* 9:657–63
105. Smith ZD, Meissner A. 2013. DNA methylation: roles in mammalian development. *Nat. Rev. Genet.* 14:204–20
106. Sone J, Mitsuhashi S, Fujita A, Mizuguchi T, Hamanaka K, et al. 2019. Long-read sequencing identifies GGC repeat expansions in *NOTCH2NLC* associated with neuronal intranuclear inclusion disease. *Nat. Genet.* 51:1215–21
107. Song XH, Hsu HK, Su MT, Chang TS, Su PY, et al. 2017. Euchromatic variants of 8q21.2 in twins. *Taiwan J. Obstet. Gynecol.* 56:227–29
108. Steiert TA, Fuss J, Juzenas S, Wittig M, Hoeppner MP, et al. 2022. High-throughput method for the hybridisation-based targeted enrichment of long genomic fragments for PacBio third-generation sequencing. *NAR Genom. Bioinform.* 4:lqac051
109. Stergachis AB, Debo BM, Haugen E, Churchman LS, Stamatoyannopoulos JA. 2020. Single-molecule regulatory architectures captured by chromatin fiber sequencing. *Science* 368:1449–54
110. Stranges PB, Palla M, Kalachikov S, Nivala J, Dorwart M, et al. 2016. Design and characterization of a nanopore-coupled polymerase for single-molecule DNA sequencing by synthesis on an electrode array. *PNAS* 113:E6749–56
111. Sullivan BA, Karpen GH. 2004. Centromeric chromatin exhibits a histone modification pattern that is distinct from both euchromatin and heterochromatin. *Nat. Struct. Mol. Biol.* 11:1076–83
112. Sun Z, Vaisvila R, Hussong LM, Yan B, Baum C, et al. 2021. Nondestructive enzymatic deamination enables single-molecule long-read amplicon sequencing for the determination of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Genome Res.* 31:291–300
113. Tsai YC, de Pontual L, Heiner C, Stojkovic T, Furling D, et al. 2022. Identification of a CCG-enriched expanded allele in DM1 patients using amplification-free long-read sequencing. *J. Mol. Diagn.* 24:1143–54
114. Tse OYO, Jiang P, Cheng SH, Peng W, Shang H, et al. 2021. Genome-wide detection of cytosine methylation by single molecule real-time sequencing. *PNAS* 118:e2019768118
115. Tyson C, Sharp AJ, Hrynchak M, Yong SL, Hollox EJ, et al. 2014. Expansion of a 12-kb VNTR containing the *REXO1L1* gene cluster underlies the microscopically visible euchromatic variant of 8q21.2. *Eur. J. Hum. Genet.* 22:458–63
116. van Buuren N, Ramirez R, Soulette C, Suri V, Han D, et al. 2022. Targeted long-read sequencing reveals clonally expanded HBV-associated chromosomal translocations in patients with chronic hepatitis B. *JHEP Rep.* 4:100449
117. van der Ende EL, Jackson JL, White A, Seelaar H, van Blitterswijk M, Van Swieten JC. 2021. Unravelling the clinical spectrum and the role of repeat length in *C9ORF72* repeat expansions. *J. Neurol. Neurosurg. Psychiatry* 92:502–9
118. Vasan N, Razavi P, Johnson JL, Shao H, Shah H, et al. 2019. Double *PIK3CA* mutations in cis increase oncogenicity and sensitivity to PI3Kα inhibitors. *Science* 366:714–23
119. Vollger MR, Guitart X, Dishuck PC, Mercuri L, Harvey WT, et al. 2022. Segmental duplications and their variation in a complete human genome. *Science* 376:eabj6965
120. Vollger MR, Logsdon GA, Audano PA, Sulovari A, Porubsky D, et al. 2020. Improved assembly and variant detection of a haploid human genome using single-molecule, high-fidelity long reads. *Ann. Hum. Genet.* 84:125–40
121. Wang S, Lee S, Chu C, Jain D, Kerpedjiev P, et al. 2020. HiNT: a computational method for detecting copy number variations and translocations from Hi-C data. *Genome Biol.* 21:73
122. Wang T, Antonacci-Fulton L, Howe K, Lawson HA, Lucas JK, et al. 2022. The Human Pangenome Project: a global resource to map genomic diversity. *Nature* 604:437–46
123. Wang X, Luan Y, Yue F. 2022. EagleC: a deep-learning framework for detecting a full range of structural variations from bulk and single-cell contact maps. *Sci. Adv.* 8:eabn9215

124. Wang Y, Zhao Y, Bollas A, Wang Y, Au KF. 2021. Nanopore sequencing technology, bioinformatics and applications. *Nat. Biotechnol.* 39:1348–65
125. Wang YC, Olson ND, Deikus G, Shah H, Wenger AM, et al. 2019. High-coverage, long-read sequencing of Han Chinese trio reference samples. *Sci. Data* 6:91
126. Warburton PE. 2001. Epigenetic analysis of kinetochore assembly on variant human centromeres. *Trends Genet*. 17:243–47
127. Warburton PE, Giordano J, Cheung F, Gelfand Y, Benson G. 2004. Inverted repeat structure of the human genome: the X-chromosome contains a preponderance of large, highly homologous inverted repeats that contain testes genes. *Genome Res*. 14:1861–69
128. Warburton PE, Hasson D, Guillem F, Lescale C, Jin X, Abrusan G. 2008. Analysis of the largest tandemly repeated DNA families in the human genome. *BMC Genom*. 9:533
129. Wenger AM, Peluso P, Rowell WJ, Chang PC, Hall RJ, et al. 2019. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* 37:1155–62
130. Wenzel A, Altmueller J, Ekici AB, Popp B, Stueber K, et al. 2018. Single molecule real time sequencing in ADTKD-*MUC1* allows complete assembly of the VNTR and exact positioning of causative mutations. *Sci. Rep.* 8:4170
131. Yang H, Garcia-Manero G, Sasaki K, Montalban-Bravo G, Tang Z, et al. 2022. High-resolution structural variant profiling of myelodysplastic syndromes by optical genome mapping uncovers cryptic aberrations of prognostic and therapeutic significance. *Leukemia* 36:2306–16
132. Yang Y, Sebra R, Pullman BS, Qiao W, Peter I, et al. 2015. Quantitative and multiplexed DNA methylation analysis using long-read single-molecule real-time bisulfite sequencing (SMRT-BS). *BMC Genom*. 16:350
133. Zeng S, Zhang MY, Wang XJ, Hu ZM, Li JC, et al. 2019. Long-read sequencing identified intronic repeat expansions in *SAMD12* from Chinese pedigrees affected with familial cortical myoclonic tremor with epilepsy. *J. Med. Genet.* 56:265–70
134. Zhang S, Pei Z, Lei C, Zhu S, Deng K, et al. 2023. Detection of cryptic balanced chromosomal rearrangements using high-resolution optical genome mapping. *J. Med. Genet.* 60:274–84
135. Zhao X, Collins RL, Lee WP, Weber AM, Jun Y, et al. 2021. Expectations and blind spots for structural variation detection from long-read assemblies and short-read genome sequencing technologies. *Am. J. Hum. Genet.* 108:919–28
136. Zook JM, Catoe D, McDaniel J, Vang L, Spies N, et al. 2016. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci. Data* 3:160025
137. Zook JM, Hansen NF, Olson ND, Chapman L, Mullikin JC, et al. 2020. A robust benchmark for detection of germline large deletions and insertions. *Nat. Biotechnol.* 38:1347–55

Annual Review of Genomics and Human Genetics

Volume 24, 2023

Contents

Errata

An online log of corrections to *Annual Review of Genomics and Human Genetics* articles may be found at http://www.annualreviews.org/errata/genom