GeNeXpress

A New and Exciting Era of Genomics: No Region is Beyond Reach

Haseena Sait

Department of Medical Genetics, Sanjay Gandhi Postgraduate Institute of Medical Sciences, Lucknow, India

Correspondence to: Dr A Haseena. *Email:* hasi.flower@gmail.com

Publication of initial drafts of the human genome by Celera Genomics and the International Human Genome Sequencing Consortium in 2001 revolutionized the field of genetics. Still, the current Genome Reference Consortium assembly (GRCh38.p13) contains several unsolvable gaps which include segmental duplications, ribosomal rRNA gene arrays, and satellite arrays. Inability to resolve these gaps was largely due to shortcomings of the existing short read sequencing technologies. Complete telomere-to-telomere reference genome assemblies are necessary to enhance our understanding of the chromosome function, human disease and genomic variation. With availability of low cost, high throughput long read sequencing technologies, the dream of creating complete and gapless human genome assemblies from diploid human genomes was undertaken by scientists from the Telomere-to-Telomereconsortium, an international collaboration of around 30 institutions from all over the world.

Telomere to telomere assembly of a complete human X chromosome (Miga et al., 2020)

To circumvent the complexity of assembling both haplotypes of a diploid genome, an effectively haploid CHM13hTERT cell line (complete hydatiform mole with 46,XX karyotype) which are uniformly homozygous for one set of alleles, was used for sequencing. The X chromosome was first selected for manual finishing and validation, owing to its high continuity in the initial assembly, distinctive and well characterized centromeric alpha satellite array, and disproportionate involvement in Mendelian disease. The high coverage, ultra-long-read nanopore sequencing of CHM13 genome was utilized along with complementary technologies for quality improvement and validation. The centromeric satellite DNA array of size 3.1Mb was reconstructed and 29 unresolved gaps in the current reference were fully resolved, including new sequences from the human pseudoautosomal region. The methylation patterns across complex tandem repeats and satellite arrays were also mapped with the help of ultra-long nanopore data. This manually finished X-chromosome assembly was estimated to be 99.9% accurate except across the largest repeats like DXZ1 satellite array (99.3% accuracy).

Segmental duplications and their variation in a complete human genome (Vollger et al., 2021)

Segmental duplications (SD), the most recent and highly identical sequences, play an important role in disease and evolution. Their length (frequently >100kbp), sequence identity, and extensive structural diversity among human haplotypes hampered the ability to characterize these regions at a genomic level. An analysis was performed in 266 human genomes using long read sequencing technologies. This study showed that 91% of the new T2T-CHM13 SD sequence (68.3Mbp) better represented the human copy number. It was also identified that SDs showed increased single-nucleotide variation diversity when compared to unique regions. Based on methylation signatures, transcription of duplicate genes and 182 novel protein-coding gene candidates were discovered. It was also found that 63% (35.11/55.7Mbp) of acrocentric SDs are distinct from rDNA and satellite sequences. These acrocentric SDs are 1.75-fold longer than other SDs, and are heteromorphic among human chromosomes. The T2T-CHM13 genome was also used to systematically reconstruct the evolution and structural haplotype diversity of biomedically relevant *(LPA,SMN)* and duplicated genes *(TBC1D3,SRGAP2C,ARHGAP11B)* important in the expansion of the human frontal cortex.

Few challenges faced were inability to fully sequence resolve all human haplotypes corresponding to specific duplicated regions using existing technology and accurate representation of these complex forms of human genetic variation, including its functional annotation.

The complete sequence of a human genome (Nurk et al., 2021)

The unfinished or erroneously placed 8% of the human genome primarily comprised the heterochromatin and many other complex regions. These functionally important regions were explored using the single haplotype, denovo strategy using CHM13 cell lines. This cell line essentially being almost homozygous was used to overcome the limitations of the previous GRC's mosaic BAC-based legacy. This group also shifted to a new strategy that leveraged the complementary aspects of PacBio HiFi (20kbp read length with median accuracy of 99.9%) and Oxford ultra-long read sequencing (more than 1 Mbp read length with error rate of 15%) to remove the 20 year old barrier that had hidden 8% of the human genome from sequence based analysis. Targeted validation was done for all these complex regions using various complementary technologies. This new T2T-CHM13 reference included gapless assemblies of all 22 autosomes plus X chromosome, corrected numerous errors of previous assemblies and introduced nearly 200 million bp of novel sequence containing 2,226 paralogous gene copies, 115 of which are predicted to be protein coding. The newly completed regions include all centromeric satellite arrays and the short arms of all five acrocentric chromosomes.

Epigenetic Patterns in a Complete Human Genome (Gershman et al., 2021)

Existing epigenetic studies omitted unassembled and unmappable genomic regions like centromeres, pericentromeres, acrocentric chromosome arms, subtelomeres, segmental duplications and tandem repeats. The new T2T-CHM13 assembly enabled the exploration of full epigenome and enrichment of epigenetic marks was performed using k-mer assisted mapping methods. Base level maps were generated using nanopore sequencing data. A distinctive dip in centromere methylation was observed and was consistent with active sites of kinetochore assembly. Allele specific, long range epigenetic patterns in complex macro-satellite arrays like those involved in X chromosome inactivation were also interrogated using long read sequencing. Single molecular measurements by long reads enabled the clustering of reads based on methylation status alone, which in turn helped in distinguishing epigenetically heterogeneous and homogenous areas. Exploring the epigenome in a larger and more diverse sample set remains a significant challenge owing to difficulty in optimal sequence alignment.

The future of the Human genome – Beginning of the end

To overcome the limitation of CHM13 cell line which lacks a Y chromosome, sequencing and assembly of a heterochromatic and a highly repetitive Y chromosome from HG002 cell line is underway. Telomere-to-telomere assembly of heterozygous diploid genomes is an arduous task that lies ahead. With continued improvement of sequencing and assembly technologies, this is not far from reach. As one genome cannot represent all humanity, the human pan-genome reference will be a key step forward for biomedical research and personalized medicine. With availability of long read sequencing technologies, a new project, the UCSC Human Pangenome Center is underway to create 350 human genomes broadly representative of humanity.

References

1. Gershman A, et al. Epigenetic Patterns in a Complete Human Genome. bioRxiv.2021; https://doi.org/10.1101/2021.05.26.443420.
2. Miga KH, et al. Telomere to telomere assembly of a complete human X chromosome Nature.2020;585:79-84.
3. Nurk S, et al. The complete sequence of a human genome.bioRxiv. 2021;https://doi.org/10.1101/2021.05.26.445798.
4. Vollger MR, et al. Segmental duplications and their variation in a complete human genome. bioRxiv. 2021; https://doi.org/10.1101/2021.05.26.445678.