Title

Envisioning a new era: Complete genetic information from routine, telomere-to-telomere genomes

Permalink

https://escholarship.org/uc/item/05v1f423

Journal

American Journal of Human Genetics, 110(11)

ISSN

0002-9297

Authors

Miga, Karen H
Eichler, Evan E

Publication Date

2023-11-01

DOI

10.1016/j.ajhg.2023.09.011

Peer reviewed

Envisioning a new era: Complete genetic information from routine, telomere-to-telomere genomes

Karen H. Miga[1,*] and Evan E. Eichler[2,3,*]

Summary

Advances in long-read sequencing and assembly now mean that individual labs can generate phased genomes that are more accurate and more contiguous than the original human reference genome. With declining costs and increasing democratization of technology, we suggest that complete genome assemblies, where both parental haplotypes are phased telomere to telomere, will become standard in human genetics. Soon, even in clinical settings where rigorous sample-handling standards must be met, affected individuals could have reference-grade genomes fully sequenced and assembled in just a few hours given advances in technology, computational processing, and annotation. Complete genetic variant discovery will transform how we map, catalog, and associate variation with human disease and fundamentally change our understanding of the genetic diversity of all humans.

Introduction

We are quickly moving past the celebration of the first release of a complete, telomere-to-telomere (T2T) assembly of a human genome to now anticipating the production release of hundreds, if not thousands, of completely sequenced human genomes. The development of completely sequenced human T2T chromosomes represented a seed change in the field of human genetics because, in principle, all forms of variation could be discovered irrespective of class, frequency, or location. The state of the art up to that point involved the production of reference genomes organized into contigs where gaps corresponded to some of the most highly repetitive and structurally complex regions of the genome. Key to resolving these was the development of long-read sequencing. Long-read sequencing and assembly, while still led by a small number of specialized sequencing centers,[1–3] has made it possible to construct phased assemblies from diploid genomes. Though currently more expensive than short-read sequencing, the costs continue to drop rapidly and throughput is increasing. With the anticipated gains in genome technology, reaching complete T2T-sequenced human genomes is expected to become more accessible and routine. What will this mean for how we map, catalog, and associate variation with human disease? How will access to complete genome sequencing alter the foundation of the future of precision medicine? And how can we broaden access to the benefits of this emerging technology?

We propose that over the next ten years, technological gains in long-read sequencing technologies will deliver routine access to T2T-phased genome assemblies. Ready access and democratization of these complete assemblies from researchers and clinicians around the world will be impactful in expanding variant discovery, understanding genetic diversity, and broadening epigenetic characterization. The potential for T2T genomes to completely capture the full spectrum of human genetic variation is beginning to transform the field of human genetics in terms of both our comprehension of genetic variants and their functional impact and how we represent and associate this variation with human disease.

Access to complete genomes will improve genetic association with human health and disease

At its most fundamental level, human genetics focuses on establishing the link between genotype and phenotype. The discovery of genetic variation, in turn, has always been dependent on advances in technology. Although cytogenetics, microarrays, optical maps, and short-read sequencing (e.g., Illumina and other sequencing platforms) all provided access to different classes and types of genetic variation, no method has been comprehensive in terms of the types, classes, or sizes of variants it can discover. Illumina whole-genome sequence provides, for example, cost-effective and accurate access to single-nucleotide variants (SNVs)[4] for 85%–90% of the genome. Larger genetic variation classes, including specific repeat regions like segmental duplications (and the genes therein), larger mobile elements, centromeres, sites of heterochromatic secondary constriction (i.e., chromosomes 1qh, 9qh, 16qh, and Yqh), and acrocentric portions of human chromosomes, have been largely overlooked in variation studies. This is due to their improper assembly in the human genome and the challenges of unambiguously hybridizing arrays or mapping short-read sequences (<250 bp), which are shorter than the repeats themselves.

[1]Department of Biomolecular Engineering, University of California Santa Cruz, Santa Cruz, CA 95064, USA; [2]Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA 98195, USA; [3]Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195, USA
*Correspondence: khmiga@ucsc.edu (K.H.M.), eee@gs.washington.edu (E.E.E.)
https://doi.org/10.1016/j.ajhg.2023.09.011.

This has translated into a limited understanding of human genetic variation and diversity and underlies some of the missing heritability of disease for Mendelian and complex genetic traits.[5]

Long-read sequencing technologies, such as nanopore sequencing from Oxford Nanopore Technologies (ONT)[6] and single-molecule, real-time (SMRT) sequencing technology from Pacific Biosciences (PacBio),[7] are transforming the field of human genetics. First, they increase the sensitivity for the detection of structural variant changes (insertions, deletions, and inversions >50 bp) by 2- to 3-fold.[8,9] Structural variants are 3-fold more likely to underlie a genome-wide association study (GWAS) signal and 50-fold more likely to affect the expression of a gene than an SNV.[10–13] This means that variants of larger effect are being disproportionately uncovered as genomes are sequenced with long reads. Second, both ONT and PacBio involve the sequencing of native DNA (as opposed to amplified material) where CpG methylation can be readily distinguished. This so-called "fifth base" allows epigenetic differences between individuals and tissue types to be readily identified, leading to the discovery of new disease mechanisms that associate genetic variants with DNA modifications. Finally, both the length and accuracy of current long-read sequencing have meant that phased genome assemblies can now be generated where both the maternal and paternal haplotypes are resolved nearly telomere to telomere.[14] Furthermore, this can now be done in the absence of parental data by assessing methylation status available from long-read sequence data from imprinted loci within the phased genome assembly.[15] This has had the transformational effect that we no longer think of the human genome as 3 Gbp but rather 6 Gbp where all variants are fully sequence resolved and phased with respect to all other genetic differences on that haplotype.

Genetic imputation is, in principle, no longer required to reconstruct haplotypes and previously inaccessible regions of our genome (Figure 1) can now be resolved and accessed for genetic and epigenetic variation. This includes the sequence resolution of complex regions, such as *SMN1* and *SMN2,* important targets for gene therapy associated with spinal muscular atrophy (SMA)[16]; the complete resolution of CGG repeat expansions in the fragile X syndrome gene *FMR1*, as well as the repeat's methylation status; and the precise length and composition of the Kringle IV repeat domain of lipoprotein A, the major genetic risk factor associated with coronary heart disease and stroke risk particularly in individuals of African American descent.[17,18] In addition to these biomedically relevant regions, long thought to be off limits to sequence, assembly, and genetic association, we are now resolving the sequence and variation of other more complex segmental duplications, acrocentric chromosomes,[19–21] and centromeres[22] for the first time. The study of the short arms of the human acrocentric chromosomes offers insight into shared, large homologous regions, including ribosomal DNA repeats and near-identical blocks of segmental duplications and satellite DNA.[19,20,21] High-resolution maps of these regions can reveal pseudo-homologous regions between short arms,[19,20] identifying arrangements that are compatible with crossover in inverted duplications near previously reported breakpoints of Robertsonian translocations,[23] and support new models of ongoing recombination exchange.[21] Additionally, T2T assemblies have opened the door to base-level studies within centromeric regions,[22,24,25] which has allowed the scoring of satellite DNA variants[26] and the precise epigenetic mapping of CpG base modifications[27] and centromere-protein enrichment,[22,24] resulting in detailed models of array evolution at sites of kinetochore assembly.[22,28] Access to high-resolution maps of satellite arrays will advance evolutionary studies of rates of non-homologous crossover and conversion and also require the development of methods[29] to confidently align satellite arrays and perform meaningful variant calling. In doing so, low-resolution labels such as "heteromorphisms," traditionally defined cytogenetically by visible gains and losses of satellite DNA staining, will become enhanced as we adopt a more meaningful definition of the underlying genomic structure necessary to understand centromere identity and function and its role in chromosomal aneuploidies. Overall, the ability to read genomes completely will offer new insight into missing heritability and new models of mutation and establish a new standard of comprehensive genetic association with human health and disease.

Routine access to complete genomes will require another step change in technologies

The cost, throughput, and computational infrastructure to routinely create and analyze T2T-phased genomes are still currently rate limiting. The mechanics of how to do so, however, have largely been solved over the last few years—a combination of more accurate long-read sequencing with extremely long read lengths (>100 kbp) has led to the development of new assembly algorithms that can now phase and assemble most of the human genome if sufficient high-molecular-weight DNA can be generated. Verkko[14] and hifiasm[1] both couple the accuracy of high-fidelity (HiFi) PacBio data[31] to create the backbone of large sequence contigs with the scaffolding potential of ultra-long ONT data[32] to assemble the maternal and paternal complement of chromosome arms and, in some cases, entire chromosomes.[33] Moreover, both long-read sequencing platforms continue to improve throughput, increase accuracy, and reduce costs. The release of the PacBio Revio sequencing platform has been estimated to increase throughput by more than 5-fold, while ONT continues to develop new nanopores and duplex sequencing to improve sequencing accuracy.[34] At present, we estimate the underlying sequencing costs for generating 30-fold sequence coverage from each of the two long-read sequencing platforms to be less than $5,000. If including

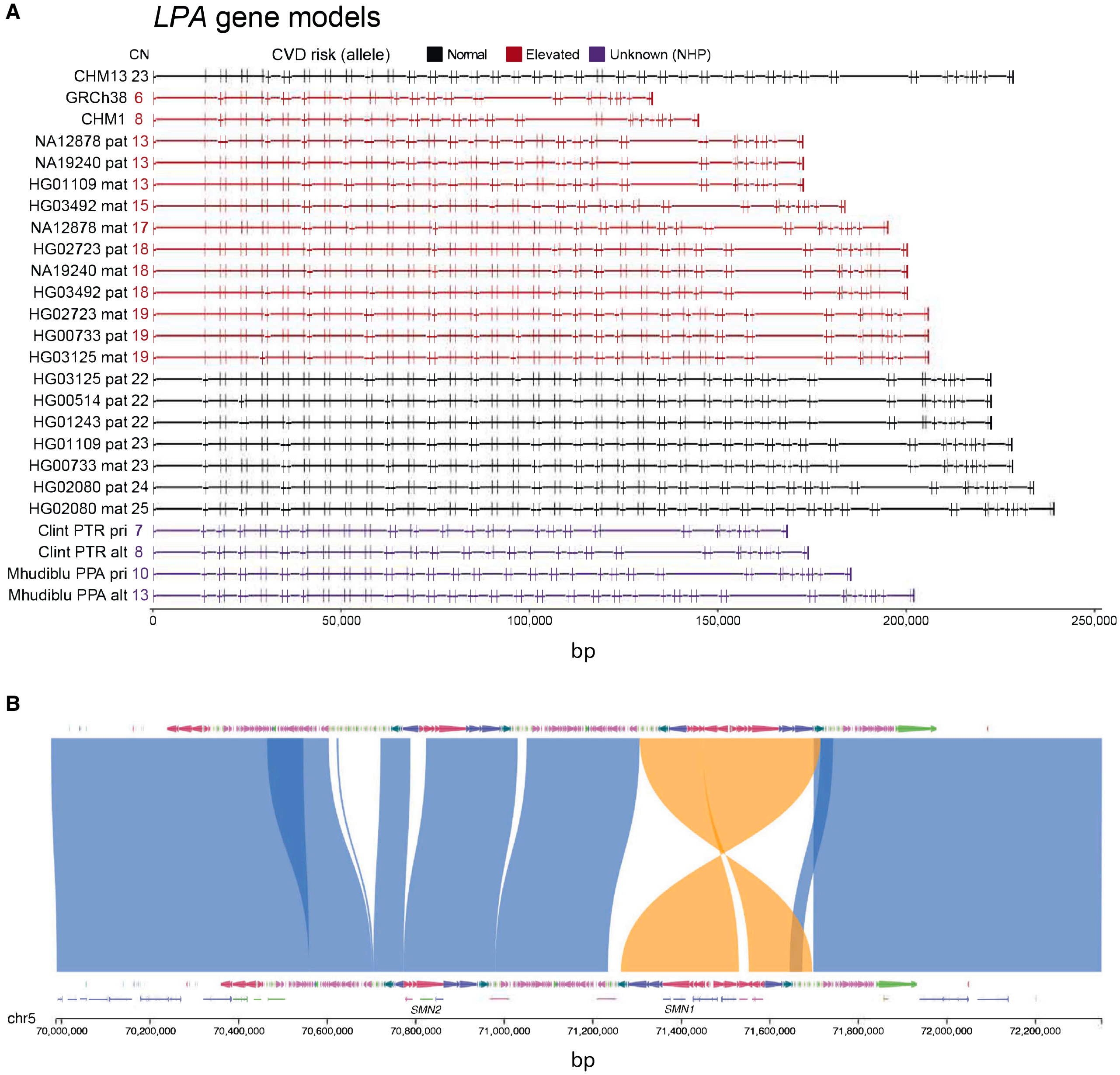


Figure 1. Sequence resolution of complex, biomedically relevant human genetic variation
(A) Sequence resolution of *LPA*, including the region encoding the Kringle IV domain. Gene structure for 20 human haplotypes and 4 chimpanzee haplotypes (PTR and PPA) are shown with haplotypes predicted to be at risk for cardiovascular disease (CVD) and stroke based on the reduced number of Kringle IV domains highlighted in red (adapted from Vollger et al.[19]). CN, copy number; NHP, nonhuman primate; pat, paternal; mat, maternal; pri, primary; alt, alternate.
(B) Structural differences in the organization of the *SMN1* and *SMN2* loci revealed by long-read sequence and assembly of two human haplotypes (adapted from Vollger et al.[30]), which includes an inversion (orange) and ordering of shared segments (blue) across regions enriched with segmental duplications (illustrated as colored arrows with orientation indicated).

labor, computational infrastructure, as well as additional ancillary technologies, such as Hi-C needed for phasing, one can generate nearly complete phased genomes with fewer than 100 gaps for ~$10,000–15,000. It remains to be seen whether best practices for T2T genomes will continue to involve the use of both platforms or if a single platform will emerge. It is clear, however, that costs will drop and throughput is increasing, thus making phased genomes a possibility for basic research pursuits.

Beyond the mechanics, a greater challenge facing human geneticists is to accurately represent the complexity of human genetic variation being uncovered. Unlike the bulk of human genetic variation (i.e., SNVs), which involve the change of a single base from one nucleotide to another, much of the newfound variation being uncovered is more complex in nature involving structural changes, copy number changes of entire genes, interlocus gene conversion, or even the complete evolutionary turnover of large swathes of portions of the genome (e.g., centromeres, acrocentric short arms).[21,22,35] While many tools have been developed in the last couple of years to visualize[36] and understand the evolutionary dynamics of such regions, cataloging these differences between human haplotypes and recording them in the form of a standard VCF is a particular challenge. Placing such variation into the context of human haplotypes, however, is a huge advantage that will be made possible by the emergence of more sophisticated genotyping tools to facilitate genotyping of these and other complex regions of the genome.[37,38] As studies shift from one to hundreds of genomes,[39,40] it is clear that one reference genome is an incompletely sufficient baseline for comparison and understanding the complexity of human variation.

A pangenome reference resource will represent common and shared haplotypes

There is a need to modernize the human reference genome to better reflect the full range of genomic diversity across the globe. In doing so, this new reference resource will contain a more accurate and diverse representation of global genomic variation, improve disease-association studies across populations, expand the scope of genomics research to the most repetitive and polymorphic regions of the genome, and serve as the ultimate genetic resource for future biomedical research and precision medicine. Genetic variations between individuals can impact the effectiveness of medical treatments and influence the risk of developing certain diseases. By creating a more diverse genome reference, we can better understand these variations and develop more personalized and effective medical treatments. Additionally, the diversity of human genomes reflects the complex history of human migration and evolution. By including a more diverse range of genomes in the reference, we can gain a more comprehensive understanding of human history and evolution. To address this challenge, the Human Pangenome Reference Consortium aims to develop and release a new human reference resource that is defined by hundreds of aligned common human haplotypes (i.e., observed with an allele frequency of 1%), known as "pangenome." This new pangenome reference will be an important foundation for identifying and predicting the functional outcomes of variants across diverse populations.

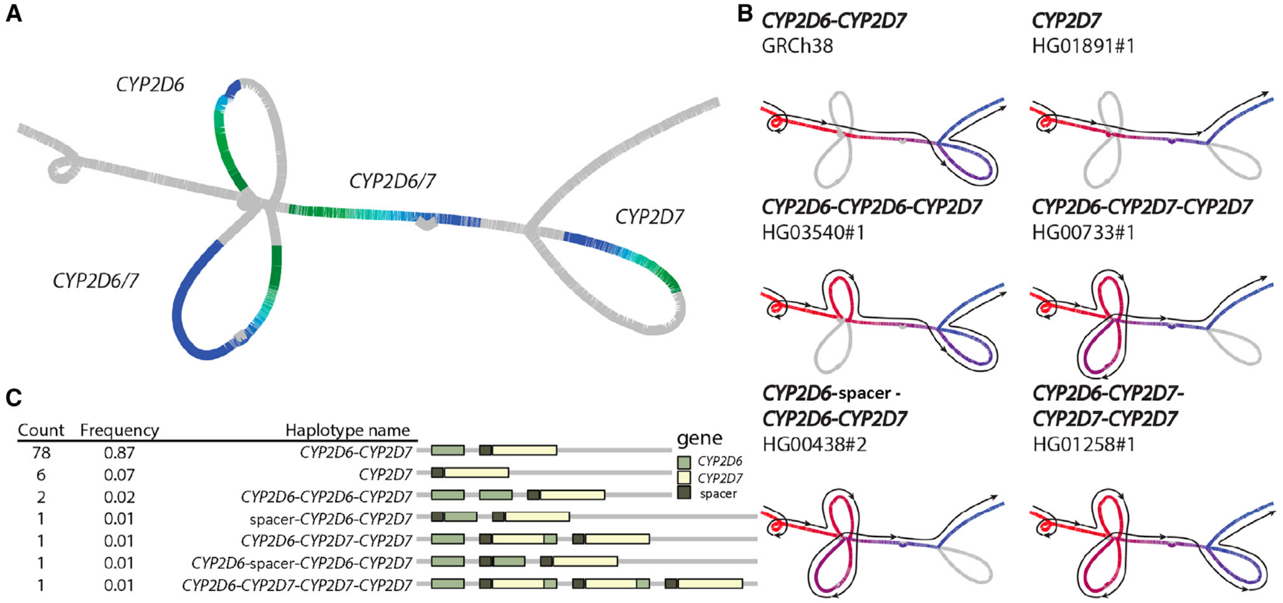
There are two stages to the development of a new human pangenome reference. The first involves the production of a series of nearly complete haplotype-resolved genomes where all genes and genetic variants are represented linearly and the diversity of human genetic variation is adequately surveyed. The second involves the development of methods to represent this diversity where the shared parts are distinguished from those that are variable (Figure 2) and, as a result, all subsequent human genomes can be mapped better than they would to a single reference genome.[41] The goals are simple: eliminate the reference bias introduced by mapping to a single reference and, thereby, improve the mapping and genotyping accuracy of genomes more distantly related to that single reference. A particularly popular approach in the field of computational genomics has been attempting to develop graphical models of the pangenome.[42] While details on construction and visualization vary, multiple sequence alignments essentially undergird pangenome graphs where the variation between any two linear haplotypes is represented as "bubbles" of various sizes and the individual linear haplotype sequences can be reconstructed as a path or a walk through the graph (Figure 2). The first human pangenome graph using 47 human genomes (94 haplotypes) was recently constructed using three different approaches allowing variation in more structurally complex loci (e.g., *HLA*, *RHD*, *CYP2D6,* etc.) to be better represented and characterized.[39] Importantly, the results showed that the pangenomic approach outperformed all other approaches not only for the characterization of structural variation but also for smaller variants, including single-nucleotide polymorphism and indels, significantly decreasing errors and improving accuracy genome-wide. The routine generation of human genomes is thus an essential endeavor, as additional genome data augment the pangenome's robustness and precision, ensuring it remains the most effective tool for mapping, genotyping, and interpreting human genetic variation.

Bold predictions

In this perspective, we have taken the position that added value of complete genetic information will make T2T genomes of individuals an inevitability for both basic and clinical research. Given this, we conclude by making five bold predictions with respect to its impact on our field.

Understanding new mutational mechanisms and their relationship to human health

Routine and comprehensive human genome sequencing and mapping are poised to revolutionize our understanding of mutational mechanisms and their impact on human health. Three such mechanisms are ectopic exchange for acrocentric DNA,[21] recombination processes that drive satellite evolution in centromeric DNA,[28] and interlocus gene conversion of segmental duplications.[30] Acrocentric DNA, which contains repetitive sequences at the ends of chromosomes, is prone to ectopic exchange—a process where non-allelic homologous sequences recombine. This can result in chromosomal rearrangements and has implications for genetic disorders, such as Robertsonian translocations. Similarly, segmental duplications, which are sizable, identical DNA sequences found in different locations within the genome, can undergo interlocus gene conversion, a non-reciprocal transfer of genetic information between the duplicated segments. This process can alter



Count	Frequency	Haplotype name
78	0.87	CYP2D6-CYP2D7
6	0.07	CYP2D7
2	0.02	CYP2D6-CYP2D6-CYP2D7
1	0.01	spacer-CYP2D6-CYP2D7
1	0.01	CYP2D6-CYP2D7-CYP2D7
1	0.01	CYP2D6-spacer-CYP2D6-CYP2D7
1	0.01	CYP2D6-CYP2D7-CYP2D7-CYP2D7

Figure 2. A graph-based representation of structural variation of the *CYP2D6/D7* locus
(A) Genes *CYP2D6* and *CYP2D7* are annotated on a graph-based representation of the locus (indicated as blue, light blue, and green color blocks). Variable portions are indicated as bubbles.
(B and C) Six different paths (starting from red and ending in blue) are depicted (B) based on linear representations (C) of this structurally polymorphic locus from 94 human genomes (adapted from Liao et al.[39]). The count and frequency of each haplotype is shown in the table.

gene function or regulation and change susceptibility to genomic disorders by homogenizing large swathes of sequence to be 100% identical in specific human haplotypes. More recently, additional mutational mechanisms associated with segmental duplications have become apparent from the generation of phased genome assemblies, namely, an increased SNV mutation with a transversion bias of cytosine to guanosine in segmental duplications when compared to unique DNA[30] and recurrent inversion toggling as a result of non-allelic homologous recombination between inverted segmental duplications.[43] In contrast, satellite DNA found enriched in human centromeres evolve extensively through mechanisms of intrachromosomal exchange that result in haplotype-specific variation.[44] The interchromosomal exchange between repeat copies of shared homologs is less frequently observed, and interchromosomal exchange between nonhomologs is expected to be rare.[22,44,45] Changes in the satellite array structure spanning known regions of kinetochore assembly and centromere function can impact chromosome segregation during cell division.[46] Therefore, advancing studies of the mutational mechanisms of centromeric satellite arrays offer new insight into studies of aneuploidy,[47] a condition associated with various disorders such as Down syndrome, Edwards syndrome, and certain cancers.[48,49] As we move toward regular, complete genome mapping, our capacity to detect, study, and understand these events will greatly improve, paving the way for more effective genetic diagnostics and understanding of disease susceptibility.

Evolutionary reconstruction of every base pair of the human genome

With routine access to comprehensive genomes of diverse species across the phylogenetic tree, we are given a unique opportunity to completely sequence the DNA from closely related species (e.g., chimpanzee, gorilla, orangutan, etc.).[50] Given sufficient depth and breadth, it should be possible to reconstruct the ancestral state of each base pair in the human genome.[51] By comparing these complete genomes, we can isolate fixed genetic differences in humans and delineate the evolutionary pathways that make us unique as a species. This form of comparative genomics can identify functionally important regions conserved across species and changes specific to the human lineage. Unveiling these human-specific genetic variants as well as the normal pattern of variation in other primates may provide unprecedented insights into our health and disease. It has been estimated, for example, that there are more than 500 extant nonhuman primate (NHP) species.[52] We anticipate complete T2T sequencing within the next decade of all NHPs and the patterns of naturally occurring "benign" variants observed in these NHP lineages will better inform or classify human mutations associated with disease as has been recently demonstrated.[53]

Additionally, NHP projects matched with increased population surveys of complete human genomes will allow us to readily detect sequences that were introduced through interbreeding between modern humans and archaic hominins.[54–56] Despite the well-established fact that *Homo sapiens* originated in Africa, there is ambiguity about the distinct paths of human evolution and the patterns of migration throughout the continent. The recent model by Ragsdale et al.[57] proposed that *Homo sapiens* descended from several varied African populations, with the first noticeable divergence happening 120,000–135,000 years ago, following extensive genetic intermixing. With T2T genomes we will be able to expand our classification of introgressed sequences across population sampling and representation of new structural variants in repeat-rich, complex regions. Notably, in centromere regions where meiotic recombination is rare, the intact transmission of centromere-spanning haplotypes (or "cenhaps") can reveal large, megabase-sized regions of archaic genetic information to advance our understanding of uncharacterized structural variation in humans.[58] Such expanded catalogs of archaic DNA in T2T assemblies can improve evolutionary models of human prehistory and support studies of introgressed variants in extant humans that have an association with disease.[59–61]

Single-cell genomics to broaden catalogs of somatic mutation and cellular function

Our bodies are a complex collection of multiple genomes, with each cell housing a potentially unique version of our DNA. This somatic variation plays a crucial role in human health and disease, including aging and cancer. As cells divide and differentiate, they can acquire genetic and epigenetic changes—an essential process for the development and diversity of cell functions but also a potential root of diseases when these changes are abnormal. The ability to produce donor-specific T2T-phased genome assemblies of an individual early in life would serve as a guide to chart all subsequent somatic changes that emerge or become clonally expanded as that individual ages. An individual's assembly, thus, could provide a powerful reference to pinpoint new somatic mutations, expression changes, or even methylation differences using single-cell readouts from biopsied material or tumor-matched control material. As single-cell RNA and genome sequencing progress from short to long reads, the precision and accuracy of detecting such changes would become exquisite. It should be noted that significant technological advances in the amplification of single-cell DNA would need to occur first to benefit from long-read sequencing because current protocols require high-molecular-weight DNA, which is difficult to obtain from a single cell. Deep long-read single-molecule sequencing directly from tissue specimens (e.g., microdissection) may prove to be more informative, providing higher-quality data to detect somatic events. Nevertheless, the ability to rapidly determine complete genetic and epigenetic maps across single cells would more generally provide invaluable insight into cellular biology and function. It would enable us to chart the life history of each cell, understand its current state, and predict its future behaviors. Such detailed mapping could uncover new dimensions of biological complexity, guide precise interventions for diseases like cancer, and potentially illuminate the mechanisms of aging at a cellular level. This technological capability is therefore a promising frontier for advancing personalized medicine and our overall understanding of human health.

Reference genomes from affected individuals will serve as a valuable resource for future medical records

Routine access to fully sequenced genomes of affected individuals is poised to significantly improve clinical care. Each person's genomic information can serve as a powerful tool for personalized medicine, allowing healthcare providers to understand an individual's unique genetic makeup and potential susceptibilities to diseases, including how that changes over time due to somatic mutation (see above). In essence, an affected individual's genome becomes a crucial part of their medical record, influencing everything from disease risk assessment to early detection to therapeutic intervention. It could enable more accurate diagnosis, targeted treatments, and even preventative strategies tailored to the person's specific genetic profile. An immediate beneficiary of such knowledge would be pharmacogenomics. Knowledge, for example, of the precise copy number of functional *CYP2D6* and *CYP2D7* (Figure 2) gene copies would guide physicians on the medications and dosages that are likely to be most effective and least harmful based on the affected individual's genetic makeup. In this light, it is noteworthy that large biobank sequencing efforts, such as the *All of Us* program, have recently added long-read sequencing of participants' genomes to the repertoire of sequencing technologies (https://allofus.nih.gov/).

Expanding a global genomics community and information sharing to advance healthcare

The original Human Genome Project was supported by an international team of researchers.[62,63] However, it was a very expensive project that required state-of-the-art laboratory infrastructure and equipment, so not all countries around the world could equally participate.[64,65] Although the genome sequence has been shared openly and without restriction, there continues to be an imbalance in who immediately benefits from having access to a human reference genome—both financially and through new drug development.[66] Therefore, when emphasizing the goal of reaching routine access to T2T genomes, it is important to recognize this current inequity and consider benefit sharing at the global scale. Here, we propose a final bold prediction that genome technologies will work to broaden international access to routine genome sequencing and analyses. This would mean that long-read sequencing technologies must be affordable and readily accessible globally. Sequencing instruments, experimental reagents, and supplies should be optimized in a way to permit long-distance shipping at ambient temperatures. Further, to ensure the broadest benefit sharing to researchers, clinicians, and citizens in the most remote, resource-limited parts of the world, new technologies should work to remove the requirements of expensive laboratory infrastructure, reduce the need for high-speed internet and computational resources for data processing, and alleviate the need for ice or dry ice in protocols. This will mean that genome technologists need to reframe the goal of routine T2T genome sequencing through the lens of "who will have access to this new technology," "who will immediately benefit financially and through improved health outcomes," and "how do we position a new global workforce capable of studying medically actionable variants using these new data."

Concluding remarks

We are moving into an era where complete genomic information will be an important driver for precision medicine and clinical care. Therefore, we need to ensure that the ability to read our genome is a common global technology and is accessible to all. The promise of genome medicine should be a fundamental human right. This bold prediction, we believe, is not purely science fiction as we have seen early indications that the global community can rally and prioritize access to sequencing technology in efforts to combat COVID via variant tracing during the pandemic.[67] We recognize, however, that global access is an ideal situation and one which will require dealing more directly with both social and economic inequities that still plague humankind first and foremost. Ultimately, the bases of a human genome sequence is a universal language, and it is time that we work together to make sure that this new technology is globally accessible to advance the promise of precision medicine and to improve health outcomes around the world for its citizens.

Acknowledgments

We would like to acknowledge the valuable contributions of Shuangjia Liu, Mitchell Vollger, and Phil Dishuck for figure assistance and Tonia Brown for assistance in editing and preparing this manuscript. This work was supported, in part, by US National Institutes of Health (NIH) grants R01HG010169 and U24HG007497 to E.E.E. and U01HG010971 and R01HG011274 to K.H.M. E.E.E. is an investigator of the Howard Hughes Medical Institute.

Declaration of interests

E.E.E. is a scientific advisory board (SAB) member of Variant Bio, Inc.

References

1. Cheng, H., Concepcion, G.T., Feng, X., Zhang, H., and Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat. Methods *18*, 170–175.
2. Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., and Phillippy, A.M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. *27*, 722–736.
3. Porubsky, D., Ebert, P., Audano, P.A., Vollger, M.R., Harvey, W.T., Marijon, P., Ebler, J., Munson, K.M., Sorensen, M., Sulovari, A., et al. (2021). Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads. Nat. Biotechnol. *39*, 302–308.
4. Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. Nature *456*, 53–59.
5. Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. Nature *461*, 747–753.
6. Deamer, D., Akeson, M., and Branton, D. (2016). Three decades of nanopore sequencing. Nat. Biotechnol. *34*, 518–524.
7. Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., et al. (2009). Real-time DNA sequencing from single polymerase molecules. Science *323*, 133–138.
8. Chaisson, M.J.P., Huddleston, J., Dennis, M.Y., Sudmant, P.H., Malig, M., Hormozdiari, F., Antonacci, F., Surti, U., Sandstrom, R., Boitano, M., et al. (2015). Resolving the complexity of the human genome using single-molecule sequencing. Nature *517*, 608–611.
9. Chaisson, M.J.P., Sanders, A.D., Zhao, X., Malhotra, A., Porubsky, D., Rausch, T., Gardner, E.J., Rodriguez, O.L., Guo, L., Collins, R.L., et al. (2019). Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nat. Commun. *10*, 1784.
10. Conrad, D.F., and Hurles, M.E. (2007). The population genetics of structural variation. Nat. Genet. *39*, S30–S36.
11. Pang, A.W., MacDonald, J.R., Pinto, D., Wei, J., Rafiq, M.A., Conrad, D.F., Park, H., Hurles, M.E., Lee, C., Venter, J.C., et al. (2010). Towards a comprehensive structural variation map of an individual human genome. Genome Biol. *11*, R52.
12. Sudmant, P.H., Rausch, T., Gardner, E.J., Handsaker, R.E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Fritz, M.H.-Y., et al. (2015). An integrated map of structural variation in 2,504 human genomes. Nature *526*, 75–81.
13. Abel, H.J., Larson, D.E., Regier, A.A., Chiang, C., Das, I., Kanchi, K.L., Layer, R.M., Neale, B.M., Salerno, W.J., Reeves, C., et al. (2020). Mapping and characterization of structural variation in 17,795 human genomes. Nature *583*, 83–89.
14. Rautiainen, M., Nurk, S., Walenz, B.P., Logsdon, G.A., Porubsky, D., Rhie, A., Eichler, E.E., Phillippy, A.M., and Koren, S. (2023). Telomere-to-telomere assembly of diploid chromosomes with Verkko. Nat. Biotechnol. https://doi.org/10.1038/s41587–023 – 01662–01666.
15. Akbari, V., Hanlon, V.C.T., O'Neill, K., Lefebvre, L., Schrader, K.A., Lansdorp, P.M., and Jones, S.J.M. (2023). Parent-of-origin detection and chromosome-scale haplotyping using long-read DNA methylation sequencing and Strand-seq. Cell Genom. *3*, 100233.
16. Lefebvre, S., Bürglen, L., Reboullet, S., Clermont, O., Burlet, P., Viollet, L., Benichou, B., Cruaud, C., Millasseau, P., and Zeviani, M. (1995). Identification and characterization of a spinal muscular atrophy-determining gene. Cell *80*, 155–165.
17. Emerging Risk Factors Collaboration, Erqou, S., Kaptoge, S., Perry, P.L., Di Angelantonio, E., Thompson, A., White, I.R.,

Marcovina, S.M., Collins, R., Thompson, S.G., and Danesh, J. (2009). Lipoprotein(a) concentration and the risk of coronary heart disease, stroke, and nonvascular mortality. JAMA *302*, 412–423.

18. Forbang, N.I., Criqui, M.H., Allison, M.A., Ix, J.H., Steffen, B.T., Cushman, M., and Tsai, M.Y. (2016). Sex and ethnic differences in the associations between lipoprotein(a) and peripheral arterial disease in the Multi-Ethnic Study of Atherosclerosis. J. Vasc. Surg. *63*, 453–458.
19. Vollger, M.R., Guitart, X., Dishuck, P.C., Mercuri, L., Harvey, W.T., Gershman, A., Diekhans, M., Sulovari, A., Munson, K.M., Lewis, A.P., et al. (2022). Segmental duplications and their variation in a complete human genome. Science *376*, eabj6965.
20. Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A.V., Mikheenko, A., Vollger, M.R., Altemose, N., Uralsky, L., Gershman, A., et al. (2022). The complete sequence of a human genome. Science *376*, 44–53.
21. Guarracino, A., Buonaiuto, S., de Lima, L.G., Potapova, T., Rhie, A., Koren, S., Rubinstein, B., Fischer, C., Gerton, J.L., et al.; Human Pangenome Reference Consortium (2023). Recombination between heterologous human acrocentric chromosomes. Nature *617*, 335–343.
22. Altemose, N., Logsdon, G.A., Bzikadze, A.V., Sidhwani, P., Langley, S.A., Caldas, G.V., Hoyt, S.J., Uralsky, L., Ryabov, F.D., Shew, C.J., et al. (2022). Complete genomic and epigenetic maps of human centromeres. Science *376*, eabl4178.
23. Jarmuz-Szymczak, M., Janiszewska, J., Szyfter, K., and Shaffer, L.G. (2014). Narrowing the localization of the region breakpoint in most frequent Robertsonian translocations. Chromosome Res. *22*, 517–532.
24. Logsdon, G.A., Vollger, M.R., Hsieh, P., Mao, Y., Liskovykh, M.A., Koren, S., Nurk, S., Mercuri, L., Dishuck, P.C., Rhie, A., et al. (2021). The structure, function and evolution of a complete human chromosome 8. Nature *593*, 101–107.
25. Miga, K.H., Koren, S., Rhie, A., Vollger, M.R., Gershman, A., Bzikadze, A., Brooks, S., Howe, E., Porubsky, D., Logsdon, G.A., et al. (2020). Telomere-to-telomere assembly of a complete human X chromosome. Nature *585*, 79–84.
26. Kunyavskaya, O., Dvorkina, T., Bzikadze, A.V., Alexandrov, I.A., and Pevzner, P.A. (2022). Automated annotation of human centromeres with HORmon. Genome Res. *32*, 1137–1151.
27. Gershman, A., Sauria, M.E.G., Guitart, X., Vollger, M.R., Hook, P.W., Hoyt, S.J., Jain, M., Shumate, A., Razaghi, R., Koren, S., et al. (2022). Epigenetic patterns in a complete human genome. Science *376*, eabj5089.
28. Miga, K.H., and Alexandrov, I.A. (2021). Variation and Evolution of Human Centromeres: A Field Guide and Perspective. Annu. Rev. Genet. *55*, 583–602.
29. Bzikadze, A.V., and Pevzner, P.A. (2023). UniAligner: a parameter-free framework for fast sequence alignment. Nat. Methods *20*, 1346–1354. https://doi.org/10.1038/s41592–023 – 01970–01974.
30. Vollger, M.R., Dishuck, P.C., Harvey, W.T., DeWitt, W.S., Guitart, X., Goldberg, M.E., Rozanski, A.N., Lucas, J., Asri, M., et al.; Human Pangenome Reference Consortium (2023). Increased mutation and gene conversion within human segmental duplications. Nature *617*, 325–334.
31. Wenger, A.M., Peluso, P., Rowell, W.J., Chang, P.-C., Hall, R.J., Concepcion, G.T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N.D., et al. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat. Biotechnol. *37*, 1155–1162.
32. Jain, M., Koren, S., Miga, K.H., Quick, J., Rand, A.C., Sasani, T.A., Tyson, J.R., Beggs, A.D., Dilthey, A.T., Fiddes, I.T., et al. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat. Biotechnol. *36*, 338–345.
33. Jarvis, E.D., Formenti, G., Rhie, A., Guarracino, A., Yang, C., Wood, J., Tracey, A., Thibaud-Nissen, F., Vollger, M.R., Porubsky, D., et al. (2022). Semi-automated assembly of high-quality diploid human reference genomes. Nature *611*, 519–531.
34. MacKenzie, M., and Argyropoulos, C. (2023). An Introduction to Nanopore Sequencing: Past, Present, and Future Considerations. Micromachines *14*, 459.
35. Logsdon, G.A., Rozanski, A.N., Ryabov, F., Potapova, T., Shepelev, V.A., Mao, Y., Rautiainen, M., Koren, S., Nurk, S., Porubsky, D., et al. (2023). The variation and evolution of complete human centromeres. Preprint at bioRxiv. https://doi.org/10.1101/2023.05.30.542849.
36. Beyer, W., Novak, A.M., Hickey, G., Chan, J., Tan, V., Paten, B., and Zerbino, D.R. (2019). Sequence tube maps: making graph genomes intuitive to commuters. Bioinformatics *35*, 5318–5320.
37. Ebler, J., Ebert, P., Clarke, W.E., Rausch, T., Audano, P.A., Houwaart, T., Mao, Y., Korbel, J.O., Eichler, E.E., Zody, M.C., et al. (2022). Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. Nat. Genet. *54*, 518–525.
38. Sirén, J., Monlong, J., Chang, X., Novak, A.M., Eizenga, J.M., Markello, C., Sibbesen, J.A., Hickey, G., Chang, P.-C., Carroll, A., et al. (2021). Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. Science *374*, abg8871.
39. Liao, W.-W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G., Lu, S., Lucas, J.K., Monlong, J., Abel, H.J., et al. (2023). A draft human pangenome reference. Nature *617*, 312–324.
40. Ebert, P., Audano, P.A., Zhu, Q., Rodriguez-Martin, B., Porubsky, D., Bonder, M.J., Sulovari, A., Ebler, J., Zhou, W., Serra Mari, R., et al. (2021). Haplotype-resolved diverse human genomes and integrated analysis of structural variation. Science *372*, eabf7117.
41. Computational Pan-Genomics Consortium (2018). Computational pan-genomics: status, promises and challenges. Brief. Bioinform. *19*, 118–135.
42. Eizenga, J.M., Novak, A.M., Sibbesen, J.A., Heumos, S., Ghaffaari, A., Hickey, G., Chang, X., Seaman, J.D., Rounthwaite, R., Ebler, J., et al. (2020). Pangenome Graphs. Annu. Rev. Genomics Hum. Genet. *21*, 139–162.
43. Porubsky, D., Höps, W., Ashraf, H., Hsieh, P., Rodriguez-Martin, B., Yilmaz, F., Ebler, J., Hallast, P., Maria Maggiolini, F.A., Harvey, W.T., et al. (2022). Recurrent inversion polymorphisms in humans associate with genetic instability and genomic disorders. Cell *185*, 1986–2005.e26.
44. Warburton, P.E., and Willard, H.F. (1995). Interhomologue sequence variation of alpha satellite DNA from human chromosome 17: evidence for concerted evolution along haplotypic lineages. J. Mol. Evol. *41*, 1006–1015.
45. Durfy, S.J., and Willard, H.F. (1989). Patterns of intra- and interarray sequence variation in alpha satellite from the human X chromosome: evidence for short-range homogenization of tandemly repeated DNA sequences. Genomics *5*, 810–821.
46. Barra, V., and Fachinetti, D. (2018). The dark side of centromeres: types, causes and consequences of structural abnormalities implicating centromeric DNA. Nat. Commun. *9*, 4340.

47. Dumont, M., Gamba, R., Gestraud, P., Klaasen, S., Worrall, J.T., De Vries, S.G., Boudreau, V., Salinas-Luypaert, C., Maddox, P.S., Lens, S.M., et al. (2020). Human chromosome-specific aneuploidy is influenced by DNA-dependent centromeric features. EMBO J. *39*, e102924.
48. Hurst, L.D. (2022). Selfish centromeres and the wastefulness of human reproduction. PLoS Biol. *20*, e3001671.
49. Sen, S., Dodamani, A., and Nambiar, M. (2023). Emerging mechanisms and roles of meiotic crossover repression at centromeres. Curr. Top. Dev. Biol. *151*, 155–190.
50. Pääbo, S. (2003). The mosaic that is our genome. Nature *421*, 409–412.
51. Kronenberg, Z.N., Fiddes, I.T., Gordon, D., Murali, S., Cantsilieris, S., Meyerson, O.S., Underwood, J.G., Nelson, B.J., Chaisson, M.J.P., Dougherty, M.L., et al. (2018). High-resolution comparative analysis of great ape genomes. Science *360*, eaar6343.
52. Rylands, A.B., and Mittermeier, R.A. (2021). In Primate names: Working Taxonomic List of Primates of the IUCN SSC Primate Specialist Group, P.,K.B.5 Strier, ed. (Routledge).
53. Gao, H., Hamp, T., Ede, J., Schraiber, J.G., McRae, J., Singer-Berk, M., Yang, Y., Dietrich, A.S.D., Fiziev, P.P., Kuderna, L.F.K., et al. (2023). The landscape of tolerated genetic variation in humans and primates. Science *380*, eabn8153.
54. Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P.H., de Filippo, C., et al. (2014). The complete genome sequence of a Neanderthal from the Altai Mountains. Nature *505*, 43–49.
55. Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M.H.-Y., et al. (2010). A draft sequence of the Neandertal genome. Science *328*, 710–722.
56. Meyer, M., Kircher, M., Gansauge, M.-T., Li, H., Racimo, F., Mallick, S., Schraiber, J.G., Jay, F., Prüfer, K., de Filippo, C., et al. (2012). A high-coverage genome sequence from an archaic Denisovan individual. Science *338*, 222–226.
57. Ragsdale, A.P., Weaver, T.D., Atkinson, E.G., Hoal, E.G., Möller, M., Henn, B.M., and Gravel, S. (2023). A weakly structured stem for human origins in Africa. Nature *617*, 755–763.
58. Langley, S.A., Miga, K.H., Karpen, G.H., and Langley, C.H. (2019). Haplotypes spanning centromeric regions reveal persistence of large blocks of archaic DNA. Elife *8*, e42989.
59. Zeberg, H., and Pääbo, S. (2020). The major genetic risk factor for severe COVID-19 is inherited from Neanderthals. Nature *587*, 610–612.
60. Racimo, F., Sankararaman, S., Nielsen, R., and Huerta-Sánchez, E. (2015). Evidence for archaic adaptive introgression in humans. Nat. Rev. Genet. *16*, 359–371.
61. Skoglund, P., and Mathieson, I. (2018). Ancient Genomics of Modern Humans: The First Decade. Annu. Rev. Genomics Hum. Genet. *19*, 381–404.
62. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.
63. Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. (2001). The sequence of the human genome. Science *291*, 1304–1351.
64. Koski, C.A. (2005). The Human Genome Project: an examination of its challenge to the technological imperative. New Genet. Soc. *24*, 265–281.
65. Elrod-Erickson, M.J., and Ford, W.F. (2000). Economic Implications of the Human Genome Project. Bus. Econ. *35*, 57.
66. Knoppers, B.M. (2000). Population genetics and benefit sharing. Community Genet. *3*, 212–214.
67. Brito, A.F., Semenova, E., Dudas, G., Hassler, G.W., Kalinich, C.C., Kraemer, M.U.G., Ho, J., Tegally, H., Githinji, G., Agoti, C.N., et al. (2022). Global disparities in SARS-CoV-2 genomic surveillance. Nat. Commun. *13*, 7003.