From complete genomes to pangenomes

Karen H. Miga[1,*]

Highlighting the Distinguished Speakers Symposium on "The Future of Human Genetics and Genomics," this collection of articles is based on presentations at the ASHG 2023 Annual Meeting in Washington, DC, in celebration of all our field has accomplished in the past 75 years, since the founding of ASHG in 1948.

The initial Human Genome Project[1,2] was a landmark achievement, serving as an essential resource for basic and clinical science, as well as for understanding human history, for over two decades. However, it is in need of an upgrade due to missing data, inaccurately assembled regions, and its inability to fully represent and identify sequence variants equitably.[3,4] A single reference map, regardless of its completeness, cannot encapsulate the variation across the human population, leading to biases and ultimately inequity in genomic studies. Recognizing this limitation, the new initiative known as the Human Pangenome Project aims to deliver hundreds of highly accurate and complete genomes.[5] This effort intends to define all bases of each chromosome from telomere to telomere (T2T),[3] ensuring a broader representation of common variants across the human species. Achieving these goals will require the rise of new tools and technology standards for complete genome assemblies and pangenomics,[6–8] which will have broad and lasting impact on genomic research.

The first map of the human genome[1,2] transformed the field of genetic research. Instead of viewing this map simply as a sequence of nucleotides for each chromosome, we can reimagine it as a linear, coordinate-based system that underpins scientific research. Since its debut, scientists worldwide have had the ability to access this map freely, enriching it with information and functional annotations at specific coordinates or sites. As a result, over the years, the human genome has become a fundamental coordinate reference for the global biomedical and basic science research communities. The Human Reference Genome can be compared to Google maps. Individuals do not turn to Google maps to identify specific global positioning system (GPS) data to describe locations around the world; rather, this app is more commonly used to obtain a list of directions and road names to navigate to your friend's house or your favorite coffee shop. Therefore, the impact is not the underlying GPS or sequence information, in the case of the Human Genome Project, but rather the rich metadata or information that is linked to system. For over two decades, the scientific community has invested billions of dollars to add information on top of the coordinate-based genome map. These efforts have led to new discoveries, allowing us to determine if any one position on this map is part of a gene or if a base change or variant at a site contributes to clinical outcomes, and we can track global allele frequencies. In other words, the progress made since the release of the Human Genome Project is not solely based on the sequence itself, but rather on the rich annotations and information collectively layered on top of it. This has made it challenging to shift away from our investment in a single, coordinate-based system.

It is important to emphasize that the Human Reference Genome is a "tool" that we use in genomic research to achieve our best science. Given the rapid pace of technological advancements, like all tools, it requires an upgrade. It is hard to imagine any tool, whether it be a phone, computer, or medical instrument, that hasn't been updated over the past twenty years. Despite its recognized imperfections, there remains a strong dependency on the highly curated reference map initially produced by the Human Genome Project. This map is flawed in that it is incomplete, with an estimated 8% missing or improperly assembled.[3] The majority of our cells contain two genomes, one set of chromosomes inherited from each parent. However, the human genome map depicts only a single haplotype. By combining or merging maternal and paternal components, it fails to accurately represent any naturally occurring genome. And finally, one genome, even if complete and accurate, cannot fully represent the global genetic diversity across our population, and because of that, the reference map is not positioned equitably to call variants, which leads to genetic bias in genomic research.

We are in a place today where it is now possible to generate highly accurate maps that represent complete chromosomes end-to-end or telomere-to-telomere (T2T). In 2022, the T2T Consortium demonstrated that with long-read sequencing technologies and new, innovative genome assembly methods, it is possible to reach a complete and highly accurate map of a human genome that surpasses

[1]Department of Biomolecular Engineering, University of California Santa Cruz, Santa Cruz, CA 95064, USA
*Correspondence: khmiga@ucsc.edu
https://doi.org/10.1016/j.ajhg.2024.05.012.

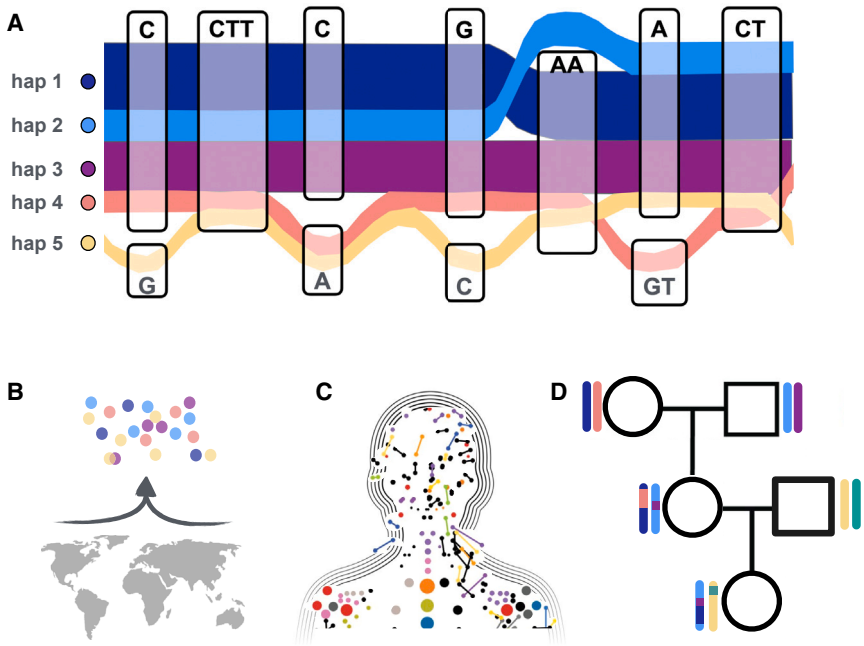

Figure 1. Pangenomic tools and applications
The pangenome is provides a new graph-based data structure (A), in which a collection of high-quality phased assemblies, are aligned representing shared paths, or haplotypes (hap 1–5). Sites where sequences differ, either by variants (e.g., "C" shared by haplotypes 1–4 and a difference in haplotype as observed by a "G," and can represent all differences in the alignment including larger structural variants). These paths in the pangenome can represent rich global genetic variation (B), aligned with the ambitious goal of the international human pangenome project. Additionally the same open-access pangenomic tooling and data structure can be used to study variants across the trillions of genomes that make up an individual (C) or across pedigrees (D) to help improve precision health initiatives in the future. (A) is a modification of a sequence tube graph (crediting Beyer et al.[12]).

the quality of the original reference genome.[3] This work introduced new sequences, or hundreds of millions of coordinates, that had been omitted by the scientific community. We could begin to build information on top of newly characterized paralogs of gene families, previously underrepresented regions mapping to our subtelomere, centromeres, and acrocentric short arms.[3,9,10] These genomic technologies can now be applied to build not one but hundreds of complete and highly accurate genome maps to improve our representation of global genetic variation. Notably, the T2T consortium has released the first diploid T2T genome and is aiming to push this assembly as close as possible to perfection. In other words, we have the technology now to build a much better tool.

Technology will continue to improve in the future, and it is likely that reading complete and accurate assemblies will be economical and accessible to non-experts. Students are now able to routinely isolate DNA from strawberries. This hands-on experiment demystifies the concept of DNA; however, it is conceivable to not only visualize the white, stringy strands the strawberry genome but to also take the next step and generate sequence data and assemble the ~240 million bases that make up the seven pairs of chromosomes of the woodland wild strawberry (*Fragaria vesca*)[11] in a classroom using a hand-held long-read sequencer attached to a laptop. Of course, we are not there yet with human genomes, yet with the rapid gains in technology it is fair to be optimistic. In 2019, The Human Pangenome Reference Consortium (HPRC)[5] used a multi-platform approach and a team of experts to release 30 genomes in our first year. Our sequencing production budget has seen impressive declines in cost, thanks to the many innovative improvements in sequencing technology from the private sector. Further, automated assembly workflows are able to reach near-T2T assemblies now in a day for less than $100. We are entering a pivotal moment where emerging technologies will democratize and standardize the use of complete genome assemblies. This advancement will lead to more comprehensive studies of the functional role of genetic and epigenetic variants and will introduce a new need for data analysis workflows designed to handle large cohorts of T2T genomes.

The HPRC is positioned at the forefront of these emerging genomic technologies, with a goal to replace the use of our current singular reference genome with a new resource defined by hundreds of reference genomes, offering a more accurate representation of global genetic diversity.[5,8] Considering the current model of a single reference genome as a straight line, with bases numbered from 5′ to 3′ to represent the sequence data of each human chromosome, we can liken a pangenome data structure to a subway map (Figure 1A). In this analogy, multiple genomes are aligned together; the points of alignment indicate similarity, while areas of divergence spotlight differences or variants. This new data structure offers an improved representation of sequence data across humanity (Figure 1B), thereby reducing inherent biases in mapping and variant calling. Further, the human pangenome significantly broadens our coordinate system to include millions of newly introduced sequences, potentially providing fresh insights into human health and disease.

It is reasonable to anticipate that advancements in technology will gradually make complete reference genomes more accessible, extending well beyond the scope of the HPRC initiative. Furthermore, the lasting impact of this project is expected to extend beyond merely releasing the human pangenome reference; it will

also include the development and widespread adoption of pangenomic tools and analysis workflows. Specifically, the provision of open-access tools enables researchers to assemble complete genomes, create their own pangenomic databases, and engage in standard analysis workflows—collectively referred to as "pangenomics." This has the potential to significantly transform genetic and genomic research. Over the next decade, we can expect a broadened and extensive application of pangenomic tools and databases, leading to novel discoveries and a new paradigm in genome sciences. This includes variant calling across numerous T2T genomes, new metrics for assessing the quality of pangenomic databases, and potentially the development of federated systems to facilitate information exchange.

These emerging technologies in complete genome assemblies and pangenomics will have an impact in clinical genomics. Highly accurate, complete genome maps of individuals with rare diseases could help to identify causal variants. Additionally, with more economical solutions to reaching complete genomes, it may eventually become standard of care in preventative medicine, meaning that it may not be uncommon for healthy individuals to have their genome as part of their medical records. As we develop from a single cell to trillions, we accumulate minor changes in our genome (Figure 1C), some of which can contribute to cancers, aging, and diseases. This implies that each individual harbors a collection of genomes. While it may sound like science fiction, a personalized genome in the future could realistically be represented by a pangenome. Advances in single-cell T2T genomes offer a glimpse into the spectrum of variation in sampled tissue or blood. Furthermore, the concept of "pedigree pangenomes" is emerging[13] (Figure 1D), where complete genomes from closely related family members provide insights into *de novo* mutations and the tracking of inheritance patterns of phased alleles. In my own lab at UCSC, we are extending these studies by constructing T2T genomes across a three-generation pedigree, enabling us to observe the intact and stable transmission of genetic and epigenetic patterns across some of the genome's most repetitive and complex regions, including the long tandem arrays defining each human centromere. Such efforts could reveal the transmission patterns of the genome's most repetitive and complex regions.

Moving beyond a single individual to billions across the globe allows us to study the transmission of entire haplotypes over time within the context of more complete genomes. The reconstruction of human haplotypes is invaluable, as it enables tracing of evolutionary relationships and understanding of the historical and geographical origins of genetic variations. This, in turn, aids in identifying genetic factors involved in diseases and enhances the accuracy of association studies. We are already uncovering intriguing features within the human genome's complex regions, accessible only through in-depth study of highly accurate chromosome assemblies. It has been established for some years that centromeres, characterized by enriched pericentromeric heterochromatin, exhibit lower recombination rates during meiosis. Consequently, centromeres are marked by large linkage blocks or centromere-spanning haplotypes, which can be passed down intact, similar to Y-haplogroups, across many generations. We have been characterizing and dating these centromere-spanning haplotypes within the assembled genome and categorizing them into distinct groupings. Our findings suggest that some of these haplotypes represent very ancient lineages, possibly inherited from archaic humans.[14]

We can expand the use of these technologies even further to include non-human primates, providing new insight into our evolutionary history, the function and origin of specific genes, and the genetic basis of traits and diseases that are unique to humans. Already the T2T Consortium has released a panel of non-human primate assemblies,[15] with more on the horizon. Living in your genome is the history of our species. Studying a large collection of species is useful for understanding the human genome because it helps to pinpoint which genetic features are unique to humans and which are shared across the Tree of Life, shedding light on the evolutionary processes that have shaped our species. We are already starting to see gains in the emergence of complete genomes and pangenomes in such evolutionary studies, and these technologies will likely recast our comparative genomics studies in the future.

In conclusion, the advent of complete genomes and pangenomics promises to revolutionize genetic and genomic research. In doing so, we can expect to see innovation in our reference genome to improve global genetic diversity, or the representation of our one humanity defined by many genomes. Moreover, this technology is positioned to refine our understanding and resolution of genetic variation in clinical genetics, human histories, and throughout evolutionary biology. Transitioning from complete genomes to pangenomics will fulfill the original genome project's promise, broadening our comprehension of each base in our genomes to better link genetics to biology and function.

References

1. 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.
2. 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.
3. 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.
4. Aganezov, S., Yan, S.M., Soto, D.C., Kirsche, M., Zarate, S., Avdeyev, P., Taylor, D.J., Shafin, K., Shumate, A., Xiao, C., et al. (2022). A complete reference genome improves analysis of human genetic variation. Science *376*, eabl3533.
5. Wang, T., Antonacci-Fulton, L., Howe, K., Lawson, H.A., Lucas, J.K., Phillippy, A.M., Popejoy, A.B., Asri, M., Carson, C., Chaisson, M.J.P., et al. (2022). The Human Pangenome Project: a global resource to map genomic diversity. Nature *604*, 437–446.
6. 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.
7. Computational Pan-Genomics Consortium (2018). Computational pangenomics: status, promises and challenges. Brief. Bioinform. *19*, 118–135.
8. 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.
9. 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.
10. 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.
11. Zhou, Y., Xiong, J., Shu, Z., Dong, C., Gu, T., Sun, P., He, S., Jiang, M., Xia, Z., Xue, J., et al. (2023). The telomere-to-telomere genome of *Fragaria vesca* reveals the genomic evolution of *Fragaria* and the origin of cultivated octoploid strawberry. Hortic. Res. *10*, uhad027.
12. 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.
13. Markello, C., Huang, C., Rodriguez, A., Carroll, A., Chang, P.-C., Eizenga, J., Markello, T., Haussler, D., and Paten, B. (2022). A complete pedigree-based graph workflow for rare candidate variant analysis. Genome Res. *32*, 893–903.
14. 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. https://doi.org/10.7554/eLife.42989.
15. Makova, K.D., Pickett, B.D., Harris, R.S., Hartley, G.A., Cechova, M., Pal, K., Nurk, S., Yoo, D., Li, Q., Hebbar, P., et al. (2023). The Complete Sequence and Comparative Analysis of Ape Sex Chromosomes. bioRxiv. https://doi.org/10.1101/2023.11.30.569198.