Title

The Human Pangenome Project: a global resource to map genomic diversity

Journal

Nature, 604(7906)

ISSN

0028-0836

Authors

Wang, Ting

Antonacci-Fulton, Lucinda

Howe, Kerstin

et al.

Publication Date

2022-04-21

DOI

10.1038/s41586-022-04601-8

Peer reviewed

HHS Public Access

Author manuscript

Nature. Author manuscript; available in PMC 2022 October 20.

Published in final edited form as:

Nature. 2022 April ; 604(7906): 437–446. doi:10.1038/s41586-022-04601-8.

The Human Pangenome Project: a global resource to map genomic diversity

Ting Wang[1,2,3,*], **Lucinda Antonacci-Fulton**[3], **Kerstin Howe**[4], **Heather A. Lawson**[1], **Julian K. Lucas**[5], **Adam M. Phillippy**[6], **Alice B. Popejoy**[7], **Mobin Asri**[5], **Caryn Carson**[1,2,3], **Mark J.P. Chaisson**[8], **Xian Chang**[5], **Robert Cook-Deegan**[9], **Adam L. Felsenfeld**[10], **Robert S. Fulton**[3], **Erik P. Garrison**[11], **Nanibaa' A. Garrison**[12], **Tina A. Graves-Lindsay**[3], **Hanlee Ji**[13], **Eimear E. Kenny**[14,15], **Barbara A. Koenig**[16], **Daofeng Li**[1,2,3], **Tobias Marschall**[17], **Joshua F. McMichael**[3], **Adam M. Novak**[5], **Deepak Purushotham**[1,2,3], **Valerie A. Schneider**[18], **Baergen I. Schultz**[10], **Michael W. Smith**[10], **Heidi J. Sofia**[10], **Tsachy Weissman**[19], **Paul Flicek**[20,*], **Heng Li**[21,22,*], **Karen H. Miga**[5,*], **Benedict Paten**[5,*], **Erich D. Jarvis**[23,24,*], **Ira M. Hall**[25,*], **Evan E. Eichler**[26,27,*], **David Haussler**[5,28,*],

Human Pangenome Reference Consortium[**]

[1]Department of Genetics, Washington University School of Medicine, St. Louis, MO, USA

[2]Edison Family Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St. Louis, MO, USA

[3]McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO

[4]Wellcome Sanger Institute, Cambridge, UK

[5]UC Santa Cruz Genomics Institute, University of California, Santa Cruz, CA, USA

[6]Genome Informatics Section, National Human Genome Research Institute, Bethesda, MD, USA

[7]Epidemiology Division, Department of Public Health Sciences, University of California, Davis, CA, USA

[8]Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, CA, USA

[9]Arizona State University, Barrett & O'Connor Washington Center, Washington DC, USA

[10]National Institutes of Health (NIH)-National Human Genome Research Institute, Bethesda, MD, USA

[11]Department of Genetics, Genomics and Informatics, University of Tennessee Health Science Center, Memphis, TN, USA

[*]corresponding authors Ting Wang (twang@wustl.edu), Paul Flicek (flicek@ebi.ac.uk), Heng Li (hli@jimmy.harvard.edu), Karen H. Miga (khmiga@ucsc.edu), Benedict Paten (bpaten@ucsc.edu), Erich D. Jarvis (ejarvis@rockefeller.edu), Ira M. Hall (ira.hall@yale.edu), Evan E. Eichler (eee@gs.washington.edu), David Haussler (haussler@ucsc.edu).

[**]a full list of members and their affiliations appears in the Supplementary Information.

Author Contributions

All authors contributed to writing the manuscript.

Conflict of interest

None declared.

[12]Institute for Society & Genetics, College of Letters and Science, Institute for Precision Health, Division of General Internal Medicine & Health Services Research, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA

[13]Department of Medicine, Stanford University, School of Medicine, Stanford, CA, USA

[14]Department of Genetics and Genomic Science and Institute for Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY, USA

[15]The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA

[16]Program in Bioethics and Institute for Human Genetics, University of California, San Francisco, San Francisco, CA, USA

[17]Heinrich Heine University, Medical Faculty, Institute for Medical Biometry and Bioinformatics, Düsseldorf, Germany

[18]National Center for Biotechnology Information (NCBI), National Library of Medicine, Bethesda, MD, USA

[19]Department of Electrical Engineering, Stanford University, Stanford, CA, USA

[20]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK

[21]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA

[22]Department of Data Science, Dana-Farber Cancer Institute, Boston, MA, USA

[23]Vertebrate Genome Lab, The Rockefeller University, New York, NY, USA

[24]Laboratory of Neurogenetics of Language, The Rockefeller University, New York, NY, USA

[25]Yale School of Medicine, New Haven, CT, USA

[26]Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA, USA

[27]Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA

[28]Howard Hughes Medical Institute, University of California, Santa Cruz, CA, USA

Preface

The human reference genome is the most widely-used resource in human genetics and is due for a major update. Its current structure is a linear composite of merged haplotypes from more than 20 people, with a single individual comprising most of the sequence. It contains biases and errors within a framework that does not represent global human genomic variation. A high-quality reference with global representation of common variants, including single nucleotide variants (SNVs), structural variants (SVs), and functional elements is needed. The Human Pangenome Reference Consortium (HPRC) aims to create a more sophisticated and complete human reference genome with a graph-based, telomere-to-telomere representation of global genomic diversity. We will leverage innovations in technology, study design, and global partnerships to construct the highest-possible quality human pangenome reference. Our effort will improve data representation

and streamline analyses to enable routine assembly of complete diploid genomes. With attention to ethical frameworks, the human pangenome reference will contain a more accurate and diverse representation of global genomic variation, improve gene-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.

Need for a Complete Human Pangenome Reference

The human reference genome is the foundational open-access resource of modern human genetics and genomics, providing a centralized coordinate system for reporting and comparing results across studies[1–4]. Its release set the bar for genomic data sharing, essential for nearly all human genomics applications, including alignments, variant detection and interpretation, functional annotations, population genetics, and epigenomic analyses. The current human reference (GRCh38.p13) is a mosaic of genomic data assembled from >20 individuals, with ~65% of the sequence contributed by a single individual[5,6]. Dependence on a single, mosaic assembly (which does not represent any one person's sequence) creates reference biases, adversely affecting variant discovery, gene-disease association studies, and the accuracy of genetic analyses[7,8]. More than two decades after the first human genome reference sequences were released, the current reference genome still contains errors, rare structural configurations that do not exist in most human genomes, and gaps in regions that have been difficult to assemble[9,10] because of their repetitive and highly polymorphic nature. The human reference genome, like most technology-driven resources, is overdue for an upgrade[11].

For years the Genome Reference Consortium (GRC) has updated the linear reference by fixing errors, filling in gaps, and adding newly discovered variants[1,4,9,12]. When enough changes accumulate, new builds are generated and released. While this process has served the community well, shortcomings have been identified along the way. Segments of genome sequences sampled from individuals may differ considerably from the reference genome, leading to errors in read mapping to the reference and reducing the accuracy of variant calls[13,14]. Identification of structural variants (SVs; >50bp deletions, insertions, tandem duplications, inversions, and translocations) relies on detecting patterns of discordant read pairs or split read alignments, which in turn depend on the accuracy of read mapping[15,16]. Assembling and detecting these SVs are challenging when the reads are too short to cover long, repetitive regions of the genome[7]. This is because short reads (50–300bp) may be identical and/or overlapping with one another such that it is impossible to determine where they should map. Both the limitations of short reads and reference biases mean that we may have missed >70% of structural variants in traditional whole genome sequencing studies[17,18].

Advances in sequencing technologies and a greater appreciation for the importance of genetic diversity make improving the human reference sequence both timely and practical. First, the development of long-read (>10kb) sequencing technologies has enabled the assembly of large, repeat-rich regions, facilitated phasing and assembly of maternal and

paternal haplotypes, and improved representation of GC-rich regions of the genome that are often missing in short-read assemblies[19–22]. Second, growing recognition of the importance of diversity and inclusion in human genomics[23] has led to widespread calls to improve representation and methods for detecting and presenting global variation.

In this Perspective, we outline the goals, strategies, challenges, and opportunities for the Human Pangenome Reference Consortium (HPRC). We will engage scientists and bioethicists in creating a human pangenome reference and resource that represents genomic diversity across human populations while improving technology for assembly and developing an ecosystem of tools for analyses of graph-based genome sequences. This new reference will maintain essential ties to the original reference for continuity, even as we strive to develop complete and error-free, telomere-to-telomere (T2T) assemblies of all chromosomes of individual human genomes, referred to here as ‘haplotypes’.

Goals and Strategies of the HPRC

A “pangenome” is the collective whole of genomic information for a given species. Originally popularized in the context of highly dynamic bacterial genomes[24], the concept has been adapted to the field of human genomics, in which the full extent of human genomic variation is expected to be much broader than has thus far been revealed. The pangenome data infrastructure depends on the high-throughput production of high-quality, phased haplotypes (segments of a chromosome identified as being maternally or paternally inherited) that improve upon the current human reference genome. Highly accurate and complete haplotype-phased genome assemblies will be organized into a graph-based data structure for the pangenome reference that compresses and indexes information[25–27]. This data structure will contain a coordinate system with a simple, intuitive framework for referring to genomic variants while preserving backward compatibility with GRCh38 and prior linear reference builds. Managing and interpreting these data requires trans-disciplinary collaboration and innovation, focused on the development of novel conceptual frameworks and analytic methods to construct the pangenome infrastructure and tools for downstream analyses and visualization. The HPRC’s goals are laid out in Box 1.

The HPRC functions through multi-disciplinary collaborations, convening cross-institutional and multi-national working groups dedicated to sample collection and consent, population genetic diversity, technology and production, phasing and assembly, approaches to construction of a pangenome reference, resource improvement and maintenance, and resource sharing and outreach (Figure 1). The HPRC has begun the process of engaging international partnerships with the Australian National Centre for Indigenous Genomics (NCIG)[38], the FDA-recognized Clinical Genome Resource (ClinGen)[39], the NIH-funded Human Heredity and Health in Africa (H3Africa) Consortium[40], the Personal Genome Project (PGP)[41], the Vertebrate Genomes Project[7] and the Global Alliance for Genomics and Health (GA4GH)[42]. The HPRC will integrate perspectives from the international scientific community through these collaborators and others yet to be identified to inform the development of HPRC references, methods, and standards.

Inclusion Criteria

In the first phase of the project, HPRC investigators selected individual genomes for high-quality sequencing among existing cell lines established by the 1000 Genomes Project (1KG), which offers a deep catalog of human variation from 26 populations[43]. These cells were originally collected from volunteer donors using consent procedures designed for unrestricted data use, and the cell lines are available in the NHGRI Biorepository at Coriell[44]. The selected cell lines were prioritized based on a combination of criteria, ranging from genetic and geographic diversity of the donors, to availability of relevant parental data (for haplotype-phasing), and limited time in cell culture (to minimize the accumulation of *de novo* mutations).

Individuals were initially identified using clustering/visualization techniques (UMAP clusters generated from 1KG data) and observed allelic diversity (heterozygosity) then selected for inclusion in the first phase of HPRC (N=100). Our inclusion criteria and recruitment strategies are evolving with the project, and we recognize that there are inherent limitations to clustering algorithms and using only the 1KG dataset.

While useful for the first phase of the HPRC, genomes selected from the 1KG data represent a limited scope of geographic and genomic diversity. One reason is that the resource was developed by sampling in 26 geographic locations across the globe, and the discrete number of individuals included from each location limits the amount of genomic variation representing those regions, especially regarding rare variants that are less likely to be observed in small sample sizes. The genomes of individuals sampled from each 1KG location cannot be assumed to harbor sufficient variation to be comprehensive of the genomic diversity in the natural population of the region, let alone to represent an entire continent. Furthermore, 1KG populations were often selected by asking potential study participants questions about their racial, ethnic, or ancestral identities, assigning ancestry based on geographic location, or some combination, which would not necessarily produce a representative sampling of any natural population. Since population descriptors can be inconsistent on clinical forms[45] and are fluid across cultural contexts[46–48], there are many unknown layers of diversity within each geographic sampling cohort of 1KG data.

Because 1KG data are insufficient to support the ambitious sampling and genetic diversity goals of the HPRC, the consortium will include additional genomes from participants identified through the BioMe Biobank at Mount Sinai and a cohort of African Americans recruited by Washington University. Those participants will be invited to contribute to the HPRC, with informed consent that comports with de-identified open data and creation of cells lines at Coriell. In later phases, the HPRC will foster additional domestic and international partnerships to explore additional avenues to broaden diversity and enhance inclusion (Box 2).

Embedded ELSI Scholarship

Most human genomics has been based on individuals of European ancestry, and the datasets available for analyses are thus biased. At present, precision medicine is based on genomic

variation found in populations with primarily European ancestry. Much of the global genetic diversity that contributes to clinical phenotypes is missing from clinical genetic tests. Many ethical, legal, and social challenges arise in efforts to include hitherto excluded populations, communities, or groups.

The HPRC has formed an 'embedded' team of scholars to address Ethical Legal and Social Implications (ELSI) of its work, with expertise at the intersections of genomics with biomedical ethics, law, social sciences, demography, community engagement, and population genetics. The main objective of the HPRC-ELSI team is to identify, investigate, and ultimately offer consortium investigators advice about the issues they face, which must be addressed if the HPRC is to meet its goals. In the embedded model, with ELSI scholars participating in key meetings where decisions are made, investigators can engage these colleagues in discussions that deepen their understanding and appreciation of what is at stake as we seek to improve the human reference genome.

Large-scale human population genetics projects aimed at broadening the diversity in genomic datasets and analyses have often missed the mark in demonstrating respect for individuals and communities. The Human Genome Diversity Project (HGDP) encountered strong opposition three decades ago[50], facing objections that its approach was extractive and its goals benefited scientists and rich country institutions, but did not match the priorities of Indigenous Peoples or people in resource-poor regions who were asked to donate their samples and data. The Havasupai Tribe of northern Arizona sued the Arizona Board of Regents in 2002, when they learned that samples donated for diabetes research were shared with other researchers and re-used for studies of schizophrenia and population origins, to which tribal members did not agree. That case was settled in 2010[51,52], but the impact it had on relations between tribal communities and genomics research has persisted.

More recently, the Wellcome Sanger Institute was criticized for licensing access to data arising from southern African samples, while institutions in Africa asserted terms of informed consent did not permit commercial uses. The NIH was also criticized for inadequate tribal engagement and consultation in the *All of Us* program[53,54]. With a keen awareness of this history, the HPRC is initiating a process to consult with, engage, and genuinely include groups currently not well represented in genomic database. Indigenous scholars have spearheaded a movement for Indigenous data sovereignty[55], for example, including the development of the CARE principles for Indigenous Data Governance[56] to layer onto the FAIR principles that support the open-science approach that the HPRC and similar projects take[57]. The HPRC is reaching out to Indigenous geneticists, leaders, and community members to engage and collaboratively develop a truly global and inclusive reference resource, taking into account FAIR and CARE principles. Furthermore, similar efforts will be made for other diverse populations that the HPRC will work with.

Some groups we seek to engage may develop sampling and sequencing efforts parallel to, rather than directly participating in the HPRC. In developing a state-of-the-art pangenome reference sequence, the HPRC will continue to disseminate standards for accuracy and completeness of sequencing, while emphasizing the importance of ELSI considerations. It is a priority for the HPRC to actively communicate with parallel genome-wide

sequencing efforts to ensure compatibility between efforts, thus enabling integration into a global pangenome reference resource. HPRC is committed to assessing local policies and promoting broad sharing of resources developed through interdisciplinary engagement, scholarship, and innovative technical solutions. The HPRC will establish procedures for navigating potential tensions among its technical, research, and resource-generating objectives with local customs, laws and data sharing policies for the groups within the HPRC as well as for those in the parallel projects.

Initial Data Generation and Release

Technological advances in genomics enable sequencing long repeats, physical mapping to chromosomes, and phasing maternally and paternally inherited haplotypes (Box 3).

For the initial phase of the project, we sequenced a single individual, HG002, whose genomic sequence has been thoroughly characterized by the Genome in a Bottle (GIAB) consortium[62]. We evaluated multiple sequencing technologies and assembly algorithms to identify the optimal combination of platforms and develop an automated pipeline that generated the most complete and accurate genome representation (Figure 2)[63]. We began with the now well-established assumption that long reads (>10kb) yield more complete genome assemblies than short reads alone[7]. The technologies tested included PacBio and/or ONT long reads for generating contigs, 10X Genomics linked reads, Hi-C paired reads, Strand-Seq long reads, and/or Bionano optical maps for scaffolding contigs into chromosomes. This pilot benchmark study created the standards for sequencing technologies and computational methodologies critical to HPRC's success.

We found that the trio approaches using parental short-read sequence data to sort haplotypes of the offspring's long-read data gave the most complete assemblies of each haplotype with the fewest structural errors[63]. Further, all methods attempting to separate haplotype sequences performed much better in generating highly contiguous assemblies than those that merged the consensus between haplotypes into one assembly. The algorithm that gave the highest haplotype separation accuracy for contigs was HiFiasm[64], which incorporates separation of reads of each haplotype into the assembly graph[65]. Generation of contigs were more structurally accurate than scaffolds, where the HPRC identified areas of improvements necessary to prevent contig miss-joins, missed-joins, collapsed repeats, and other structural assembly errors. Based on these findings, an initial set of 47 1KG genomes from parent-offspring trios was assembled with Hifiasm[64], creating high quality diploid contig-only genome assemblies. Going forward, we will further optimize sequencing, assembly, and analysis methods with the goal of creating fully-phased T2T diploid genomes, including repetitive and structurally variable regions such as centromeres, telomeres, and segmental duplications. We anticipate that the high-quality assemblies created in the project will drive tool creation and improvement for diploid genome assembly and QC where new and recently created existing tools (from the T2T assembly of CHM13[22]) are applied to diploid genome assembly.

The first HPRC data release is comprised of the sequencing data from 47 participants – mostly from the 1KG Project (listed and described in Supplemental Table 1). All sequencing data are publicly available and can be downloaded without egress fees from the Amazon

Web Services (AWS) Public Datasets program and can be analyzed with the AWS cloud. Data is also available for analysis within the AnVIL cloud platform, organized as a public workspace[66]. AnVIL is the NHGRI's genomic data science Analysis, Visualization, and Informatics Lab-space that provides a cloud environment for analysis of large genomic datasets, and supports multiple globally used analysis tools including Terra, Bioconductor, Jupyter, and Galaxy[36].

Pangenome Reference

We are building a pangenome reference with three complementary parts: 1) the haplotypes: the sequences within the input assemblies; 2) the pangenome alignment: a sequence graph and an efficient embedding of each of the input haplotypes as paths within this graph; and 3) the coordinate system: a backward-compatible coordinate system and set of sequences that make it possible to refer to all variations encoded within the reference equally (Figure 3). The haplotypes provide hundreds of individual representations of the genome, spanning global diversity. Each haplotype assembly will be useful individually as a reference for studying genomic sequences that are divergent from the current human reference assembly. The pangenome alignment represents the homology relationships among the individual assemblies. This canonical alignment will support coordinates translation (liftOver) between the haplotypes and defines the allelic relationships. It will form the substrate for many emerging pangenome tools and pipelines that will improve important genomic workflows, for example, by making genotyping accuracy less dependent on ancestry. The coordinate system provides a global, unambiguous means to refer to all the variations within the pangenome. It makes all the variations within the haplotypes first class objects that can be referred to equally. Ultimately, it will provide a more complete means to refer to variations not contained within the existing linear reference, proving useful for databases and tooling that will build upon the pangenome reference.

Supporting these parts is a new proposed set of file standards[67], notably the rGFA format for representing a pangenome and the GAF format for representing read mappings to a pangenome. We hope these will have an impact on the field similar to how SAM/BAM[68] and VCF[69] formats generated a broad range of interoperable tools that became widely used and accessible. To kick-start this process, we have developed the vg toolkit[70] and minigraph[67], which incorporate downstream tools for graph construction and long- and short-read mapping and genotyping.

We anticipate releasing an alpha pangenome reference based on existing variant calls and assembled contig genomes. Using the proposed incremental coordinate system, we will subsequently release updated graphs incorporating the growing numbers of assemblies.

Variant Detection

A central aim of this research is to document the genetic similarities and differences among the human genomes included in the pangenome reference. Comprehensive variant detection, however, is still a challenge even when high-quality genome assemblies are available. No single data type or bioinformatic approach yet achieves high performance across all variant

classes and genomic regions[58,71]. Therefore, we are pursuing multiple complementary approaches to variant detection using a combination of whole-genome multiple assembly alignment, pairwise assembly-assembly alignment, and traditional reference-based read alignment.

Ideally, we will accomplish variant detection in a single step designed to build pangenome graphs directly from whole-genome, multiple assembly alignments. Genetic variants will be represented naturally as features in the resulting graph because any variant would be captured by the assembly process. This offers a significant advantage, enabling optimal breakpoint reconstruction via joint analysis of all input genomes. Accurate multiple-alignment and graph construction of entire human genomes is extremely challenging, but recent improvements to tools such as minimap2[72], minigraph[67], cactus[73] and pggb[74] make this feasible. However, variant calling errors can still arise from assembly and sequence alignment errors, especially in repetitive regions of the genome. Given this, and the fact that pangenome graph construction tools have not been thoroughly evaluated at scale with real-world data, we are also pursuing the complementary approaches described below.

An alternative approach to multiple-alignments is to map variants from pairwise assembly-assembly alignments. Towards this end, we are using minimap2 and Winnowmap to align each draft assembly to the GRCh38 and T2T-CHM13 references to perform variant detection of SNVs, indels and other SVs. This approach is more straightforward than whole-genome multiple-alignment; however, complications can arise from reference genome effects and the need to merge results across many pairwise comparisons. The exact coordinates of complex and repetitive variants may differ due to alignment ambiguity. To alleviate reference effects, we are mapping variants via pairwise alignment of the two haploid genomes of each individual, enabling detection of natural heterozygous variants within sequences that are missing or poorly represented in GRCh38 and CHM13. Pairwise alignment assembly methods help control for potential errors from the multiple-alignment and graph construction process outlined above; however, they still fail to detect variants that are not captured in the underlying assemblies. We are also running a host of traditional variant callers that rely on alignment of raw reads to the GRCh38 and CHM13 references to control for potential assembly errors. Though limited by reference genome quality and alignment accuracy, these traditional tools are able to capture a subset of variants that are not accurately assembled, and they will serve as a cross-check on newer and less mature assembly-based tools.

In sum, we expect the above methods to capture most genetic variants in genomic regions accessible to current assembly and alignment methods. We will compare our variant calls to published call sets from the 1KG Project[75], HGSVC[76], and GIAB[62] using samples from these projects that are also included in the HPRC references to assess quality. We will evaluate and validate variant calls using independent data types generated by the HPRC but not used for contig assembly – such as ONT, Hi-C, Strand-Seq, and BioNano data – and assess the read-level support for each variant call based on alignment of raw data to assemblies and pangenome graphs.

Achieving comprehensive T2T variant detection across the entire genome will require improved methods for genome assembly, multiple-alignments and graph construction. The development and application of these methods in subsequent years is a major goal of the HPRC, and will help extend the impact of pangenomics to the full spectrum of variant classes.

Pangenome Annotation

Annotation of the current GRCh38 reference includes genes and genomic features, such as repeats, CpG islands, regulatory regions, and ChIP-seq peaks among others. The pangenome reference will have these same utilities and more, including the following:

Genes:

The two primarily used gene sets in genomic analysis are NCBI's RefSeq[77], which exists as independent mRNA definitions, and EMBL-EBI's Ensembl/GENCODE[78], which is built on the GRCh38 reference. The pangenome reference will support both the RefSeq and Ensembl/GENCODE gene sets. We will map both annotations to each haplotype. Specifically, we will evaluate the mapping of the core reference set of human transcriptome data to each haplotype and incorporate putative new genes that are not represented in either RefSeq or Ensembl/GENCODE. Mapping these gene sets in conjunction with other transcriptomic data sets will annotate the pangenome graph. Other tools will support spliced alignments and transcript reconstruction on a mature graphical data structure. We will integrate the results of these approaches into the annotation released for each haplotype, accompanied by a description of whether transcripts are identical by both methods or whether changes were identified, including transcripts that are disabled, duplicated, or missing on a given haplotype. We will also annotate all transcript haplotypes for their global frequency using Haplosaurus tools[79]. We will initially annotate haplotype-by-haplotype, and will explore methods for direct annotation of the pangenome, such as those currently being developed in the GENCODE Consortium. Direct annotation methods simultaneously cover all relevant haplotypes and result in both an annotated genome graph and haplotype-specific annotations. One of the critical use cases of direct annotation of the pangenome will be large transcriptomic datasets aligned directly to the graphical structure that natively annotate it.

Functional Elements and Other Genome Features:

A central goal in biology is to understand how sequence variants affect genome function to influence phenotypes. Genome function includes regulatory regions that influence gene expression, enhancers that modulate expression levels, and the three-dimensional interactions that control chromosome structural organization within a cell. We will use the pangenome reference to annotate such functional information using existing RNA-Seq, MethylC-Seq, and ATAC-Seq datasets from Roadmap Epigenomics, ENCODE, 4D Nucleome (4DN), Genotype-Tissue Expression (GTEx), and Center for Common Disease Genomics (CCDG), among others. This will enhance the functional human genetic variation catalog.

Integrating functional data with the pangenome reference will facilitate the development of toolkits and analysis pipelines that evaluate the impact of genetic variants on complex traits and variation in phenotypes. The HPRC will work with developers to define rules and mechanisms to engage with multimodal 'Big Bio-Data' for both data providers and consumers. We will co-create user-friendly informatics platforms to manage, integrate, visualize, and compare highly heterogeneous datasets in the context of the genetic diversity represented in the pangenome. Box 4 lists available resources for working with pangenome graphs. We will also make all haplotype-by-haplotype annotation methods available in AnVIL so that others can run them to create custom annotation tracks on all or a selected subset of assemblies. These platforms will serve as a foundation for significant clinical datasets and global biobank initiatives that will ultimately improve precision medicine and medical breakthroughs. For example, the NHGRI is establishing the Impact of Genomic Variation on Function (IGVF) Consortium, which aims to develop a framework for systematically understanding the effects of genomic variation on genome function. Data generated by the IGVF will include high-resolution identification and annotation of functional elements and cell type-specific perturbation studies to assess the impact of genomic variants on function. The pangenome will be an important foundation for predicting functional outcomes in these studies.

Data Sharing

To enhance community access and sharing, we will submit sequence data (HiFi, ONT, Hi-C, and others), assemblies, and pangenomes produced by the consortium to AnVIL[36] and INSDC[28]. Data will also be stored and made publicly available on both S3 and Google Cloud Storage. This general model supports future efforts to use cloud-based strategies for biological data analysis spanning multiple centers. Users of various clouds worldwide will know that they are using the same datasets. Data coordination within the Consortium will leverage the established methods in use and constant development since the inception of 1KG more than a decade ago[105,106]. These processes will ensure that we rapidly release data in an organized manner, with proper accessioning of archival datasets, and future traceability of analysis objects and primary data items. Data stored in INSDCs will use BioProjects and BioProject umbrella structures similar to the 1KG and Vertebrate Genome Project[7] to ensure that data are appropriately organized and easily identifiable in the public archives. This approach ensures that sample identifiers are effectively managed via the BioSamples database[107], including metadata provisions, and makes any data generated from the same samples readily tractable. INSDC will archive all reads, assembly data, and other relevant archives will be used, as appropriate for a specific data type. Each haplotype assembly will receive a Genome Collections accession number (GCA_*), which we will version as we make assembly updates. We will address additional data sharing considerations as they arise through our expanded recruitment and sampling efforts to broaden the diverse representation of global variation.

Adoption and Outreach

Achieving widespread international adoption of a pangenome reference will be a challenge[11]. HPRC will design a pragmatic model and transition plan that is simple and

compelling enough to gain traction among researchers and clinical laboratories. Working across scientific and other stakeholder communities, we will foster a new ecosystem of analysis tools. We will maintain and improve the reference, establish scalable bioinformatics methods for resolving errors, improve resolution in difficult-to-resolve genomic regions, and respond to user feedback. Importantly, we envision an integrated pangenome transition plan that involves broad community engagement via outreach and education, from tool developers to end-users. These efforts will create a software ecosystem and expert user base to support the next generation of human genetics. The pangenome reference will provide improved genomic research standards, data sharing, and reproducible cloud-based workflows. Understanding the barriers to adoption will lead to effective outreach and training, ensuring that the pangenome reference resource is widely adopted.

Adoption will ultimately be driven by the creation of a data resource that sustains continued improvement in its accuracy and completeness, enables a range of uses, and improves genomic analyses. We will actively publicize the benefits of using the pangenome. As a starting point for our outreach efforts, we have created a website (humanpangenome.org) to publicize the Consortium. We have also created human pangenome social media accounts that directly connect our Consortium with the end-user community (e.g. @HumanPangenome on Twitter).

To facilitate adoption, we will explore who the user community will be, their needs, and, most importantly, the technical and non-technical barriers they may encounter. Addressing potential obstacles is essential, since we know that adopting an updated version of the linear reference resulted in significant bottlenecks for many laboratories. The cost of switching can be significant, and HPRC is aware that many clinical labs worldwide still use the GRCh37 build from February 2009 for this reason. HPRC will examine how to reduce switching costs and expedite transition. User data will be collected in self-reporting surveys, including user characteristics, location, specific applications, and barriers to adopting a pangenome reference framework.

Creating a coordinate system that builds on GRCh38 and includes both GRCh37 and GRCh38 assemblies is central to user adoption. HPRC will develop training materials that explain the additional sequences included in the pangenome reference coordinates and how these sequences relate to GRCh37/GRCh38. Existing linear-reference tools will continue to work with the expanded pangenome reference coordinate system, and pangenome-based results will be translatable to these existing coordinate systems with improved genotype accuracy.

We will develop liftOver tools that make it easy to go backward from the pangenome reference to GRCh37/GRCh38 when necessary. We already have algorithms for this purpose and demonstration of functionality to predict read mappings from a prototype pangenome to GRCh37/GRCh38. We will precompute all mappings between the previous assemblies and the pangenome and provide these coordinate translation functions with the pangenome reference release. This information should ease the transition of other databases and resources that rely on these coordinates and provide an annotation directly onto the

GRCh37/GRCh38 assemblies in areas where mappings and interpretation on the pangenome are more reliable than current, linear sequence representations.

We will augment the human genome browser's displays to transition to any haplotype assembly in the pangenome reference and display the haplotype alignments. Visualizations will include relevant genetic backgrounds for specific tracks, for example picking the right HLA haplotype for a read mapping track. To ensure we use these tools effectively, we will add detailed information that explains these novel views to our existing training materials and make this information part of our respective workshops.

We have adopted the GA4GH principles and will develop exchange formats analogous to SAM/BAM and utility libraries analogous to htslib/samtools, facilitating the development of transition tools and workflows for the pangenome reference. We will deposit these tools and their guides in the HPRC resource repository. We have also developed a prototype transcript archive that facilitate annotation discovery in GRCh37, GRCh38, CHM13 and the pangenome, and visualize the differences between two transcripts (for example, on two different genomes).

We aim to engage pilot users to obtain feedback about these resources. The HPRC program and related tool developers connected with the community of users will develop new tools that gain additional value from using the pangenome reference rather than linear-reference genome assemblies. We will report on our discoveries in publications and talks, through the blog, webinars, and on the HPRC website and provide educational tools and forums on using and switching to a pangenome reference.

Relevance to Disease Research

We expect that the resources and methods we are developing will profoundly impact studies of the genetic basis of human disease and precision medicine. While we recognize that adoption by the clinical research community will take time, there are three important benefits to using a pangenome reference. First, a more complete reference that incorporates and displays human genetic diversity will produce fewer ambiguous mappings and more accurate copy-number variation analyses throughout the genome when patient samples are sequenced and analyzed[59,108]. This will improve genetic diagnosis and the functional annotation of variants. Second, the resource will enable the discovery of disease risk alleles and previously unobserved rare variants, especially in regions that are inaccessible to standard, short-read sequencing technologies. Studies of unsolved Mendelian genetic disease, for example, have shown that ~25% of "missing" disease variants can be recovered when longer reads are applied and more complex repetitive regions are characterized[109]. Important genetic risk loci such as SMN1/2 (spinal muscular atrophy), LPA (lipoprotein A and coronary heart disease), CYP2D6 (pharmacogenomics), as well as numerous triplet repeat expansion loci are now being sequenced and assembled in large human cohort studies. These studies are revealing the standing pattern of natural genetic variation for loci typically excluded from previous analyses[58,59]. Resolution of these loci by long-read sequencing in even a limited number of human haplotypes improves our ability to genotype them in other patient-derived short read datasets, allowing for the discovery of new genetic

associations, through both GWAS and eQTL methods[58]. Finally, the pangenome approach represents a fundamental change in how human genetic variation is discovered. Instead of simply mapping sequence reads to a reference, we are constructing phased genome assemblies and aligning them to the graph, which in turn will pinpoint all genetic differences both large and small at the base-pair level[26,110]. As long-read sequencing costs fall and pangenome methods evolve[26], we predict patient samples will likely be sequenced using long-read technology to increase sensitivity and accuracy.

Outlook

As we write this Perspective, the world is reeling from the COVID-19 pandemic and the spread of new SARS-Cov-2 variants. Scientists can trace the virus's epidemiology, determine why humans are susceptible[111,112], and determine why some individuals are more susceptible than others[113,114]. The current GRCh38 human reference is one of many resources that made this possible, but we know that it can be improved. Through years of strategic investments in the public and private sectors, we find ourselves with the technologies and methods to build additional references that better represent global human genomic diversity.

The human pangenome reference will collect accurate, haplotype-phased genome assemblies generated by efficient algorithmic innovations, which we anticipate will be widely used by the scientific community. The collection of individual genomes, comprised of sequence information, genomic coordinates, and annotations, will be a critical resource with more accurate representation of human genomic diversity. The original Human Genome Project enabled major advances in human health and genomic medicine[1–4]; it is time to build a more inclusive resource with better representation of human genomic diversity to better serve humanity.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Funding

We would like to thank the National Human Genome Research Institute (NHGRI) for funding multiple components to improve and update the Human Genome Reference Program, which has supported the work represented by this report (1U41HG010972, 1U01HG010971, 1U01HG010961, 1U01HG010973, 1U01HG010963). This work was also supported, in part, by the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health (A.M.P.).

References

1. Lander ES et al. Initial sequencing and analysis of the human genome. Nature 409, 860–921, doi:10.1038/35057062 (2001). [PubMed: 11237011]
2. Venter JC et al. The sequence of the human genome. Science 291, 1304–1351. (2001). [PubMed: 11181995]
3. Gibbs RA The Human Genome Project changed everything. Nat Rev Genet, doi:10.1038/s41576-020-0275-3 (2020).
4. Venter JC et al. The sequence of the human genome. Science 291, 1304–1351, doi:10.1126/science.1058040 (2001). [PubMed: 11181995]

5. Green RE et al. A draft sequence of the Neandertal genome. Science 328, 710–722, doi:10.1126/science.1188021 (2010). [PubMed: 20448178]
6. Sherman RM & Salzberg SL Pan-genomics in the human genome era. Nat Rev Genet 21, 243–254, doi:10.1038/s41576-020-0210-7 (2020). [PubMed: 32034321]
7. Rhie A et al. Towards complete and error-free genome assemblies of all vertebrate species. Nature 592, 737–746, doi:10.1038/s41586-021-03451-0 (2021). [PubMed: 33911273]
8. Need AC & Goldstein DB Next generation disparities in human genomics: concerns and remedies. Trends Genet 25, 489–494, doi:10.1016/j.tig.2009.09.012 (2009). [PubMed: 19836853]
9. Schneider VA et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome Res 27, 849–864, doi:10.1101/gr.213611.116 (2017). [PubMed: 28396521]
10. Bustamante CD, Burchard EG & De la Vega FM Genomics for the world. Nature 475, 163–165, doi:10.1038/475163a (2011). [PubMed: 21753830]
11. Miga KH & Wang T The Need for a Human Pangenome Reference Sequence. Annu Rev Genomics Hum Genet, doi:10.1146/annurev-genom-120120-081921 (2021).
12. International Human Genome Sequencing, C. Finishing the euchromatic sequence of the human genome. Nature 431, 931–945, doi:10.1038/nature03001 (2004). [PubMed: 15496913]
13. Garrison E et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nat Biotechnol 36, 875–879, doi:10.1038/nbt.4227 (2018). [PubMed: 30125266]
14. Martiniano R, Garrison E, Jones ER, Manica A & Durbin R Removing reference bias and improving indel calling in ancient DNA data analysis by mapping to a sequence variation graph. Genome Biol 21, 250, doi:10.1186/s13059-020-02160-7 (2020). [PubMed: 32943086]
15. Alkan C, Coe BP & Eichler EE Genome structural variation discovery and genotyping. Nat Rev Genet 12, 363–376, doi:10.1038/nrg2958 (2011). [PubMed: 21358748]
16. Sedlazeck FJ et al. Accurate detection of complex structural variations using single-molecule sequencing. Nat Methods 15, 461–468, doi:10.1038/s41592-018-0001-7 (2018). [PubMed: 29713083]
17. Sudmant PH et al. An integrated map of structural variation in 2,504 human genomes. Nature 526, 75–81, doi:10.1038/nature15394 (2015). [PubMed: 26432246]
18. Chaisson MJP et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nat Commun 10, 1784, doi:10.1038/s41467-018-08148-z (2019). [PubMed: 30992455]
19. Li R et al. Building the sequence map of the human pan-genome. Nat Biotechnol 28, 57–63, doi:10.1038/nbt.1596 (2010). [PubMed: 19997067]
20. Miga KH et al. Telomere-to-telomere assembly of a complete human X chromosome. Nature 585, 79–84, doi:10.1038/s41586-020-2547-7 (2020). [PubMed: 32663838]
21. Logsdon GA et al. The structure, function and evolution of a complete human chromosome 8. Nature 593, 101–107, doi:10.1038/s41586-021-03420-7 (2021). [PubMed: 33828295]
22. Nurk S et al. The complete sequence of a human genome. bioRxiv, 2021.2005.2026.445798, doi:10.1101/2021.05.26.445798 (2021).
23. Sirugo G, Williams SM & Tishkoff SA The Missing Diversity in Human Genetic Studies. Cell 177, 26–31, doi:10.1016/j.cell.2019.02.048 (2019). [PubMed: 30901543]
24. Tettelin H et al. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial “pan-genome”. Proc Natl Acad Sci U S A 102, 13950–13955, doi:10.1073/pnas.0506758102 (2005). [PubMed: 16172379]
25. Vernikos G, Medini D, Riley DR & Tettelin H Ten years of pan-genome analyses. Curr Opin Microbiol 23, 148–154, doi:10.1016/j.mib.2014.11.016 (2015). [PubMed: 25483351]
26. Computational Pan-Genomics C Computational pan-genomics: status, promises and challenges. Brief Bioinform 19, 118–135, doi:10.1093/bib/bbw089 (2018). [PubMed: 27769991]
27. Eizenga JM et al. Pangenome Graphs. Annu Rev Genomics Hum Genet 21, 139–162, doi:10.1146/annurev-genom-120219-080406 (2020). [PubMed: 32453966]
28. Arita M, Karsch-Mizrachi I & Cochrane G The international nucleotide sequence database collaboration. Nucleic Acids Res 49, D121–D124, doi:10.1093/nar/gkaa967 (2021). [PubMed: 33166387]

29. Okubo K, Sugawara H, Gojobori T & Tateno Y DDBJ in preparation for overview of research activities behind data submissions. Nucleic Acids Res 34, D6–9, doi:10.1093/nar/gkj111 (2006). [PubMed: 16381940]
30. Kent WJ et al. The human genome browser at UCSC. Genome Res 12, 996–1006, doi:10.1101/gr.229102 (2002). [PubMed: 12045153]
31. Navarro Gonzalez J et al. The UCSC Genome Browser database: 2021 update. Nucleic Acids Res 49, D1046–D1057, doi:10.1093/nar/gkaa1070 (2021). [PubMed: 33221922]
32. Stalker J et al. The Ensembl Web site: mechanics of a genome browser. Genome Res 14, 951–955, doi:10.1101/gr.1863004 (2004). [PubMed: 15123591]
33. Howe KL et al. Ensembl 2021. Nucleic Acids Res 49, D884–D891, doi:10.1093/nar/gkaa942 (2021). [PubMed: 33137190]
34. Zhou X et al. The Human Epigenome Browser at Washington University. Nat Methods 8, 989–990, doi:10.1038/nmeth.1772 (2011). [PubMed: 22127213]
35. Li D, Hsu S, Purushotham D, Sears RL & Wang T WashU Epigenome Browser update 2019. Nucleic Acids Res 47, W158–W165, doi:10.1093/nar/gkz348 (2019). [PubMed: 31165883]
36. Schatz MC et al. Inverting the model of genomics data sharing with the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space (AnVIL). bioRxiv (2021).
37. Popejoy AB & Fullerton SM Genomics is failing on diversity. Nature 538, 161–164, doi:10.1038/538161a (2016). [PubMed: 27734877]
38. National Centre for Indigenous Genomics of Australia, <https://ncig.anu.edu.au/> (
39. Rehm HL et al. ClinGen--the Clinical Genome Resource. N Engl J Med 372, 2235–2242, doi:10.1056/NEJMsr1406261 (2015). [PubMed: 26014595]
40. Human Heredity and Health in Africa (H3Africa), <https://h3africa.org> (
41. The Personal Genome Project, <https://www.personalgenomes.org>(
42. The Global Alliance Genomics and Health (GA4GH), <https://www.ga4gh.org> (
43. Genomes Project, C. et al. A global reference for human genetic variation. Nature 526, 68–74, doi:10.1038/nature15393 (2015). [PubMed: 26432245]
44. Coriell Institute, <https://www.coriell.org> (
45. Popejoy AB et al. The clinical imperative for inclusivity: Race, ethnicity, and ancestry (REA) in genomics. Hum Mutat 39, 1713–1720, doi:10.1002/humu.23644 (2018). [PubMed: 30311373]
46. Popejoy AB et al. Clinical Genetics Lacks Standard Definitions and Protocols for the Collection and Use of Diversity Measures. Am J Hum Genet 107, 72–82, doi:10.1016/j.ajhg.2020.05.005 (2020). [PubMed: 32504544]
47. Bonham VL et al. Physicians' attitudes toward race, genetics, and clinical medicine. Genet Med 11, 279–286, doi:10.1097/GIM.0b013e318195aaf4 (2009). [PubMed: 19265721]
48. Race E & Genetics Working G The use of racial, ethnic, and ancestral categories in human genetics research. Am J Hum Genet 77, 519–532, doi:10.1086/491747 (2005). [PubMed: 16175499]
49. Mills MC & Rahal C A scientometric review of genome-wide association studies. Commun Biol 2, 9, doi:10.1038/s42003-018-0261-x (2019). [PubMed: 30623105]
50. Dodson M & Williamson R Indigenous peoples and the morality of the Human Genome Diversity Project. J Med Ethics 25, 204–208, doi:10.1136/jme.25.2.204 (1999). [PubMed: 10226929]
51. Couzin-Frankel J Ethics. DNA returned to tribe, raising questions about consent. Science 328, 558, doi:10.1126/science.328.5978.558 (2010). [PubMed: 20430983]
52. Dukepoo FC The trouble with the Human Genome Diversity Project. Mol Med Today 4, 242–243, doi:10.1016/s1357-4310(98)01282-9 (1998). [PubMed: 9679240]
53. Fox K The Illusion of Inclusion - The "All of Us" Research Program and Indigenous Peoples' DNA. N Engl J Med 383, 411–413, doi:10.1056/NEJMp1915987 (2020). [PubMed: 32726527]
54. Devaney SA, Malerba L & Manson SM The "All of Us" Program and Indigenous Peoples. N Engl J Med 383, 1892, doi:10.1056/NEJMc2028907 (2020).
55. Hudson M et al. Rights, interests and expectations: Indigenous perspectives on unrestricted access to genomic data. Nat Rev Genet 21, 377–384, doi:10.1038/s41576-020-0228-x (2020). [PubMed: 32251390]

56. Carroll SR, Herczog E, Hudson M, Russell K & Stall S Operationalizing the CARE and FAIR Principles for Indigenous data futures. Sci Data 8, 108, doi:10.1038/s41597-021-00892-0 (2021). [PubMed: 33863927]
57. Wilkinson MD et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018, doi:10.1038/sdata.2016.18 (2016). [PubMed: 26978244]
58. Ebert P et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. Science 372, doi:10.1126/science.abf7117 (2021).
59. Vollger MR et al. Segmental duplications and their variation in a complete human genome. bioRxiv, 2021.2005.2026.445678, doi:10.1101/2021.05.26.445678 (2021).
60. Lieberman-Aiden E et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326, 289–293, doi:10.1126/science.1181369 (2009). [PubMed: 19815776]
61. Ulahannan N et al. Nanopore sequencing of DNA concatemers reveals higher-order features of chromatin structure. bioRxiv, 833590, doi:10.1101/833590 (2019).
62. Genome in a Bottle, <https://www.nist.gov/programs-projects/genome-bottle> (
63. HG002_Data_Freeze_v1.0, <https://github.com/human-pangenomics/HG002_Data_Freeze_v1.0> (
64. Cheng H, Concepcion GT, Feng X, Zhang H & Li H Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods 18, 170–175, doi:10.1038/s41592-020-01056-5 (2021). [PubMed: 33526886]
65. Nurk S et al. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. Genome Res 30, 1291–1305, doi:10.1101/gr.263566.120 (2020). [PubMed: 32801147]
66. AnVIL-HPRC, <https://anvil.terra.bio/#workspaces/anvil-datastorage/AnVIL_HPRC> (
67. Li H, Feng X & Chu C The design and construction of reference pangenome graphs with minigraph. Genome Biol 21, 265, doi:10.1186/s13059-020-02168-z (2020). [PubMed: 33066802]
68. Li H et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079, doi:10.1093/bioinformatics/btp352 (2009). [PubMed: 19505943]
69. Danecek P et al. The variant call format and VCFtools. Bioinformatics 27, 2156–2158, doi:10.1093/bioinformatics/btr330 (2011). [PubMed: 21653522]
70. Rosen Y, Eizenga J & Paten B Modelling haplotypes with respect to reference cohort variation graphs. Bioinformatics 33, i118–i123, doi:10.1093/bioinformatics/btx236 (2017). [PubMed: 28881971]
71. Abel HJ et al. Mapping and characterization of structural variation in 17,795 human genomes. Nature 583, 83–89, doi:10.1038/s41586-020-2371-0 (2020). [PubMed: 32460305]
72. Li H Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34, 3094–3100, doi:10.1093/bioinformatics/bty191 (2018). [PubMed: 29750242]
73. Paten B et al. Cactus: Algorithms for genome multiple sequence alignment. Genome Res 21, 1512–1528, doi:10.1101/gr.123356.111 (2011). [PubMed: 21665927]
74. Pangenome Graph Builder, <https://github.com/pangenome/pggb> (
75. 1000 Genome Project, <https://www.internationalgenome.org> (
76. The Human Genome Structural Variation Consortium, <https://www.internationalgenome.org/human-genome-structural-variation-consortium/> (
77. O'Leary NA et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 44, D733–745, doi:10.1093/nar/gkv1189 (2016). [PubMed: 26553804]
78. Frankish A et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res 47, D766–D773, doi:10.1093/nar/gky955 (2019). [PubMed: 30357393]
79. Spooner W et al. Haplosaurus computes protein haplotypes for use in precision drug design. Nat Commun 9, 4128, doi:10.1038/s41467-018-06542-1 (2018). [PubMed: 30297836]
80. Liu B, Guo H, Brudno M & Wang Y deBGA: read alignment with de Bruijn graph-based seed and extension. Bioinformatics 32, 3224–3232, doi:10.1093/bioinformatics/btw371 (2016). [PubMed: 27378303]

81. Limasset A, Cazaux B, Rivals E & Peterlongo P Read mapping on de Bruijn graphs. BMC Bioinformatics 17, 237, doi:10.1186/s12859-016-1103-9 (2016). [PubMed: 27306641]
82. Heydari M, Miclotte G, Van de Peer Y & Fostier J BrownieAligner: accurate alignment of Illumina sequencing data to de Bruijn graphs. BMC Bioinformatics 19, 311, doi:10.1186/s12859-018-2319-7 (2018). [PubMed: 30180801]
83. GenomeMapper, <https://www.1001genomes.org/software/genomemapper_graph.html> (
84. Kim D, Paggi JM, Park C, Bennett C & Salzberg SL Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol 37, 907–915, doi:10.1038/s41587-019-0201-4 (2019). [PubMed: 31375807]
85. Hickey G et al. Genotyping structural variants in pangenome graphs using the vg toolkit. Genome Biol 21, 35, doi:10.1186/s13059-020-1941-7 (2020). [PubMed: 32051000]
86. Rautiainen M & Marschall T GraphAligner: rapid and versatile sequence-to-graph alignment. Genome Biol 21, 253, doi:10.1186/s13059-020-02157-2 (2020). [PubMed: 32972461]
87. GRAF by SevenBridges, <https://www.sevenbridges.com/graf/> (
88. PaSGAL, <https://github.com/ParBLiSS/PaSGAL> (
89. Dvorkina T, Antipov D, Korobeynikov A & Nurk S SPAligner: alignment of long diverged molecular sequences to assembly graphs. BMC Bioinformatics 21, 306, doi:10.1186/s12859-020-03590-7 (2020). [PubMed: 32703258]
90. Mokveld T, Linthorst J, Al-Ars Z, Holstege H & Reinders M CHOP: haplotype-aware path indexing in population graphs. Genome Biol 21, 65, doi:10.1186/s13059-020-01963-y (2020). [PubMed: 32160922]
91. Ghaffaari A & Marschall T Fully-sensitive seed finding in sequence graphs using a hybrid index. Bioinformatics 35, i81–i89, doi:10.1093/bioinformatics/btz341 (2019). [PubMed: 31510650]
92. Wick RR, Schultz MB, Zobel J & Holt KE Bandage: interactive visualization of de novo genome assemblies. Bioinformatics 31, 3350–3352, doi:10.1093/bioinformatics/btv383 (2015). [PubMed: 26099265]
93. Gonnella G, Niehus N & Kurtz S GfaViz: flexible and interactive visualization of GFA sequence graphs. Bioinformatics 35, 2853–2855, doi:10.1093/bioinformatics/bty1046 (2019). [PubMed: 30596893]
94. Kunyavskaya O & Prjibelski AD SGTK: a toolkit for visualization and assessment of scaffold graphs. Bioinformatics 35, 2303–2305, doi:10.1093/bioinformatics/bty956 (2019). [PubMed: 30475983]
95. Mikheenko A & Kolmogorov M Assembly Graph Browser: interactive visualization of assembly graphs. Bioinformatics 35, 3476–3478, doi:10.1093/bioinformatics/btz072 (2019). [PubMed: 30715194]
96. Beyer W et al. Sequence tube maps: making graph genomes intuitive to commuters. Bioinformatics 35, 5318–5320, doi:10.1093/bioinformatics/btz597 (2019). [PubMed: 31368484]
97. Yokoyama TT, Sakamoto Y, Seki M, Suzuki Y & Kasahara M MoMI-G: modular multi-scale integrated genome graph browser. BMC Bioinformatics 20, 548, doi:10.1186/s12859-019-3145-2 (2019). [PubMed: 31690272]
98. ODGI, <https://github.com/pangenome/odgi> (
99. Shlemov A & Korobeynikov A in Algorithms for Computational Biology. (eds Holmes Ian, Martín-Vide Carlos, & Vega-Rodríguez Miguel A.) 80–94 (Springer International Publishing).
100. Ebler J et al. Pangenome-based genome inference. bioRxiv, 2020.2011.2011.378133, doi:10.1101/2020.11.11.378133 (2020).
101. Leggett RM et al. Identifying and classifying trait linked polymorphisms in non-reference species by walking coloured de bruijn graphs. PLoS One 8, e60058, doi:10.1371/journal.pone.0060058 (2013). [PubMed: 23536903]
102. BayesTyper, <https://github.com/bioinformatics-centre/BayesTyper> (
103. Chen S et al. Paragraph: a graph-based structural variant genotyper for short-read sequence data. Genome Biol 20, 291, doi:10.1186/s13059-019-1909-7 (2019). [PubMed: 31856913]

104. Eggertsson HP et al. GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. Nat Commun 10, 5402, doi:10.1038/s41467-019-13341-9 (2019). [PubMed: 31776332]

105. Clarke L et al. The 1000 Genomes Project: data management and community access. Nat Methods 9, 459–462, doi:10.1038/nmeth.1974 (2012). [PubMed: 22543379]

106. Clarke L et al. The international Genome sample resource (IGSR): A worldwide collection of genome variation incorporating the 1000 Genomes Project data. Nucleic Acids Res 45, D854–D859, doi:10.1093/nar/gkw829 (2017). [PubMed: 27638885]

107. Courtot M et al. BioSamples database: an updated sample metadata hub. Nucleic Acids Res 47, D1172–D1178, doi:10.1093/nar/gky1061 (2019). [PubMed: 30407529]

108. Aganezov S et al. A complete reference genome improves analysis of human genetic variation. bioRxiv, 2021.2007.2012.452063, doi:10.1101/2021.07.12.452063 (2021).

109. Miller DE et al. Targeted long-read sequencing identifies missing disease-causing variation. Am J Hum Genet, doi:10.1016/j.ajhg.2021.06.006 (2021).

110. Logsdon GA, Vollger MR & Eichler EE Long-read human genome sequencing and its applications. Nat Rev Genet 21, 597–614, doi:10.1038/s41576-020-0236-x (2020). [PubMed: 32504078]

111. Kim D et al. The Architecture of SARS-CoV-2 Transcriptome. Cell 181, 914–921 e910, doi:10.1016/j.cell.2020.04.011 (2020). [PubMed: 32330414]

112. Zhou P et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579, 270–273, doi:10.1038/s41586-020-2012-7 (2020). [PubMed: 32015507]

113. Toh C & Brody JP Evaluation of a genetic risk score for severity of COVID-19 using human chromosomal-scale length variation. Hum Genomics 14, 36, doi:10.1186/s40246-020-00288-y (2020). [PubMed: 33036646]

114. Zeberg H & Paabo S The major genetic risk factor for severe COVID-19 is inherited from Neanderthals. Nature 587, 610–612, doi:10.1038/s41586-020-2818-3 (2020). [PubMed: 32998156]

Key References

- [61] Cheng H, Concepcion GT, Feng X, Zhang H & Li H Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods 18, 170–175, doi:10.1038/s41592-020-01056-5 (2021). [PubMed: 33526886] [61] o Hifiasm is a haplotype-resolved assembler specifically designed for PacBio HiFi reads, that aims to represent haplotype information in a phased assembly graph.
- [22] Nurk S et al. The complete sequence of a human genome. bioRxiv, 2021.2005.2026.445798, doi:10.1101/2021.05.26.445798 (2021).[22] o The first complete genome assembly issued from the Telomere-to-telomere (T2T) Consortium, which closed all remaining gaps in the GRCh38 including all acrocentric short arms, segmental duplications, and human centromeric regions.
- [20] Miga KH et al. Telomere-to-telomere assembly of a complete human X chromosome. Nature 585, 79–84, doi:10.1038/s41586-020-2547-7 (2020). [PubMed: 32663838] [20] o The sequence of the first complete human chromosome.
- [64] Li H, Feng X & Chu C The design and construction of reference pangenome graphs with minigraph. Genome Biol 21, 265, doi:10.1186/s13059-020-02168-z (2020). [PubMed: 33066802] [64] o Minigraph toolkit used to efficiently construct a pangenome graph, useful for mapping and for constructing graphs encoding structural variation.
- [70] Paten B et al. Cactus: Algorithms for genome multiple sequence alignment. Genome Res 21, 1512–1528, doi:10.1101/gr.123356.111 (2011). [PubMed: 21665927] [70] o Cactus is a highly accurate, reference-free multiple genome alignment program useful to study general rearrangement and copy number variation.
- [13] Garrison E et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nat Biotechnol 36, 875–879, doi:10.1038/nbt.4227 (2018). [PubMed: 30125266] [13] o A model for representing genomes aims to improve read mapping by representing genetic variation in the reference.

- [36] Schatz MC et al. Inverting the model of genomics data sharing with the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space (AnVIL). bioRxiv (2021).[36] o The AnVIL platform provides scalable solutions for genomic data access, analysis, and education.
- [104] Aganezov S et al. A complete reference genome improves analysis of human genetic variation. bioRxiv, 2021.2007.2012.452063, doi:10.1101/2021.07.12.452063 (2021).[104] o Demonstrates the importance of complete, T2T genomes in novel variant discovery and offering major improvements of variant calls within clinically relevant genes.
- [55] Ebert P et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. Science 372, doi:10.1126/science.abf7117 (2021).[55] o Use of Long-read data from 64 human genomes to predict structural variants and the patterns of variation across diverse populations.
- [37] Popejoy AB & Fullerton SM Genomics is failing on diversity. Nature 538, 161–164, doi:10.1038/538161a (2016). [PubMed: 27734877] [37] o Analysis of sample descriptions included in the GWAS Catalog indicates that some populations are still underrepresented and left behind in studies of genomic medicine.
- [10] Bustamante CD, Burchard EG & De la Vega FM Genomics for the world. Nature 475, 163–165, doi:10.1038/475163a (2011). [PubMed: 21753830] [10] o Emphasizes the importance of reference data from ancestral and diverse genomes, while stating that researcher invest time and money into education and outreach to explain why studying global (and local) health is so important.

Box 1: Goals of the HPRC

- Identify individuals representing diverse genomic and biogeographic backgrounds to include in the pangenome reference, with at least 350 reference quality haplotype-phased human diploid genomes (700 haplotypes total).
- Integrate ethical, legal, and social implications (ELSI) scholarship in the development of recommended policies and protocols for inclusion, data acquisition, and stewardship from study recruitment to publication of findings.
- Prioritize the use of long-read and long-range technologies and assemblies with haplotype-aware algorithms to generate the highest quality phased genomes possible.
- Create methods to finish diploid genomes from telomere-to-telomere across complex regions, closing gaps and ensuring hard-to-measure variants are identified.
- Foster an ecosystem of pangenome reference tools to facilitate the annotation of genes and other genomic features.
- Implement an iterative design-development-engagement process to understand and respond to user community needs.
- Develop communication strategies that will assure understanding of the pangenome reference resource, including the community's ability to fix and report errors.
- Enable appropriately controlled access to data through genomics platforms such as the International Nucleotide Sequence Database Collaboration (INSDC)[28,29], the National Center for Biotechnology Information (NCBI), the UCSC Genome Browser[30,31], Ensembl[32,33], the WashU Epigenome Browser[34,35] and NHGRI's cloud-based analysis platform, AnVIL[36,37].
- Foster an international human pangenome reference alliance that actively engages the diverse populations it seeks to represent.

Box 2: Commitment to Diversity and Inclusion

There are many aspects of diversity to consider for broad inclusion, and the first step is to assess current gaps in diversity. Researchers have demonstrated a lack of diversity in genomics research using biogeographic ancestry groupings at the continental level, as well as socio-cultural categories such as racial and ethnic identities[49]. It is important to distinguish biological and sociocultural diversity, since sociocultural labels are not derived from genotype data, and vice versa. Due to gaps in genomic sampling worldwide, the distribution of allelic variants appears to be correlated with continental-level biogeographic 'ancestry'. However, variants are rarely unique to a single biogeographic 'population', and factors such as effective population size, founder events, and genetic drift are responsible for differences in allele frequencies between human groups.

Capturing the full range of human genomic diversity is a daunting task: some gaps are understood and predictable, but we also face *unknown* unknowns. The initial HPRC dataset cannot be comprehensive of global genomic variation, but it can set a foundation to build upon. The HPRC will produce high-quality genome data for 350 individuals (700 haploid genomes) selected to maximize global representation within the logistical constraints of the initial HPRC efforts. Strategic partnerships with organizations such as the GA4GH and H3Africa are underway, which we anticipate will help facilitate international engagement and broaden our understanding of the cultural, ethical, legal, social, and political considerations of the HPRC. However, further partnerships will be needed to include populations that are under-represented or entirely missing from current data resources. HPRC actively welcomes additional partners and collaborators to join us in rising to this challenge.

Box 3: Sequencing and Assembly

Notable improvements in long-read technologies have resulted in complete chromosome assemblies[20–22] and have demonstrated the ability to broaden variant analysis to span large, complex human SVs[58]. Use of highly accurate consensus reads (99.9%, or Q30) of moderate length (e.g. 10–20 kb), such as high fidelity (HiFi) reads from Pacific Biosciences (PacBio), routinely resolves long tandem repeats, or satellite arrays, and large segmental duplications[20,21,59]. In parallel, the nanopore-based sequencing platform (Oxford Nanopore Technologies) offers long-read data that routinely generate substantial coverage of reads that are 100s of kb in length (or "ultra-long" data, UL) with an increasing number of reported reads greater than one million bases. Like the HiFi data, UL data is used to close large and persistent assembly gaps, including in human centromeres, subtelomeric regions, and large segmental duplications[22,59]. Further, chromosome conformation capture methods produce long-range data for both short-read (Hi-C[60]) and long-read sequencing (Pore-C[61]). Such chromatin crosslinking protocols generate chimeric DNA fragments from interacting chromosomal regions that are covalently linked together. These ligated DNA molecules are sequenced to help determine phasing and spatial organization at the level of an entire chromosome. With continuing gains in both HiFi read length and nanopore single read base-level quality, and improved methods for the use of chromosome conformation capture methods to guide phased haplotype assembly, we are entering into a new era of routine complete chromosome-level assemblies[20–22,37]. In collaboration with the T2T Consortium, which aims to use long-read sequencing and cutting-edge algorithmic approaches to close the hundreds of gaps persisting in the human reference genome[22], the HPRC will generate accurate assemblies of entire chromosomes. These assemblies will empower us to characterize variations in large, repeat-rich regions that have historically been out of reach for standard genetic analysis and interpretation.

Box 4: Pangenome Graph Tools

Graph Building

- minigraph[67]
- PGGB[74]

Graph Aligners

- deBGA[80]
- BGREAT[81]
- BrownieAligner[82]
- GenomeMapper[83]
- HISAT2[84]
- VG[85]
- GraphAligner[86]
- GRAF[87] ▫
- PaSGAL[88]
- SPAligner[89]

Graph Indexing

- CHOP[90]
- PSI[91]

Graph Visualization

- Bandage[92]
- GfaViz[93]
- SGTK[94]
- AGB[95]
- Sequence Tube Map[96]
- MoMI-G[97]
- VG view[85]
- VG vis[85]
- ODGI viz[98]

Gene Prediction

- Path Racer[99]

Variant Detection

- PanGenie[100]
- Cortex with Bubbleparse[101]
- BayesTyper[102]
- Paragraph[103]
- GraphTyper2[104]
- VG[85]

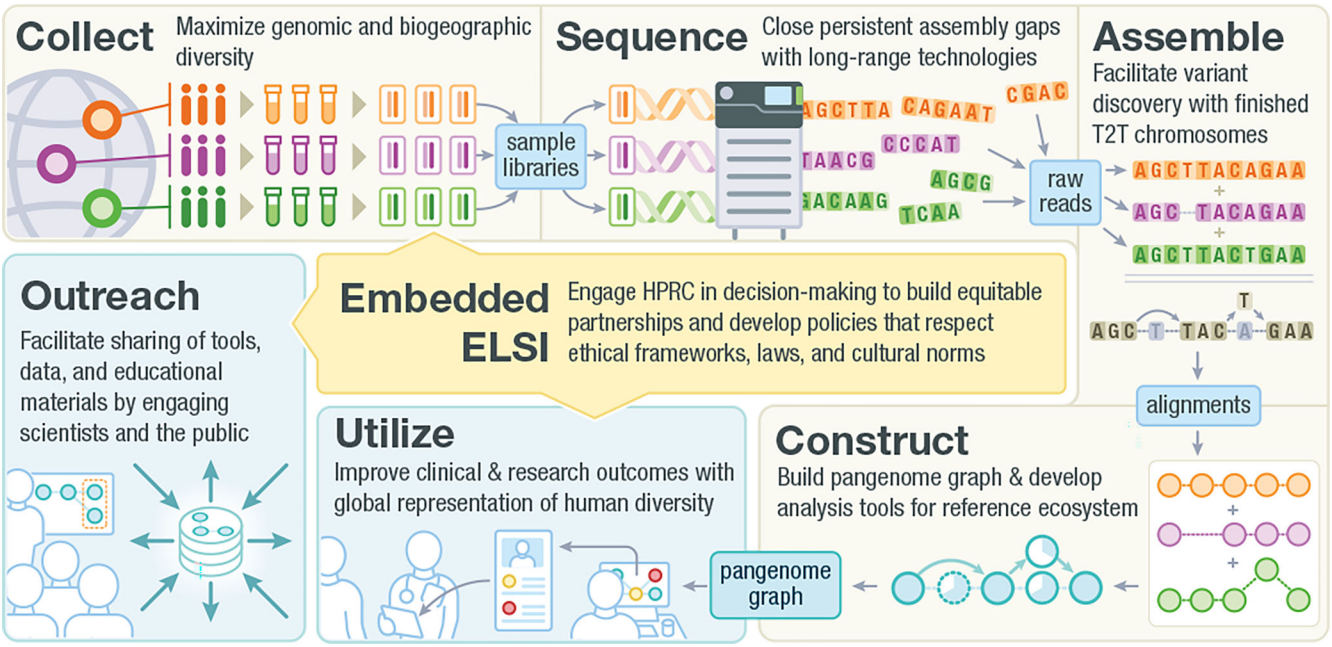

Figure 1: The Human Pangenome Reference Consortium.

This diagram provides an overview of the HPRC's several components. **Collect**: I,000 Genomes samples jump-start the project and will be followed by additional samples collected through community engagement and recruitment. Sample selection efforts will ensure the graph-based reference captures global human genomic diversity. **Sequence**: Long-read and long-range technologies are used to generate genome graphs and bridge gaps in difficult-to-assemble genomic regions. **Assemble**: Telomere-to-telomere finished diploid genomes will foster variant discovery, especially in complex, difficult to assemble genomic regions. **Construct**: Scalable bioinformatics approaches assemble, QC, call variants, and benchmark graph assembly accuracy. The graph is annotated with gene descriptions and transcriptome data, making it more accessible and interpretable. **Utilize**: Collaboration across scientific and stakeholder communities will create a new ecosystem of analysis tools. Clinical applications and research use will involve analysis, validation, interpretation, and publication of results. **Outreach**: Members of the HPRC Outreach community engage and educate the user community and broadly share all genomic products and informatics platforms. **ELSI**: ELSI scholars will develop selection processes and policy frameworks that meet investigator needs while respecting research partner autonomy and cultural norms.

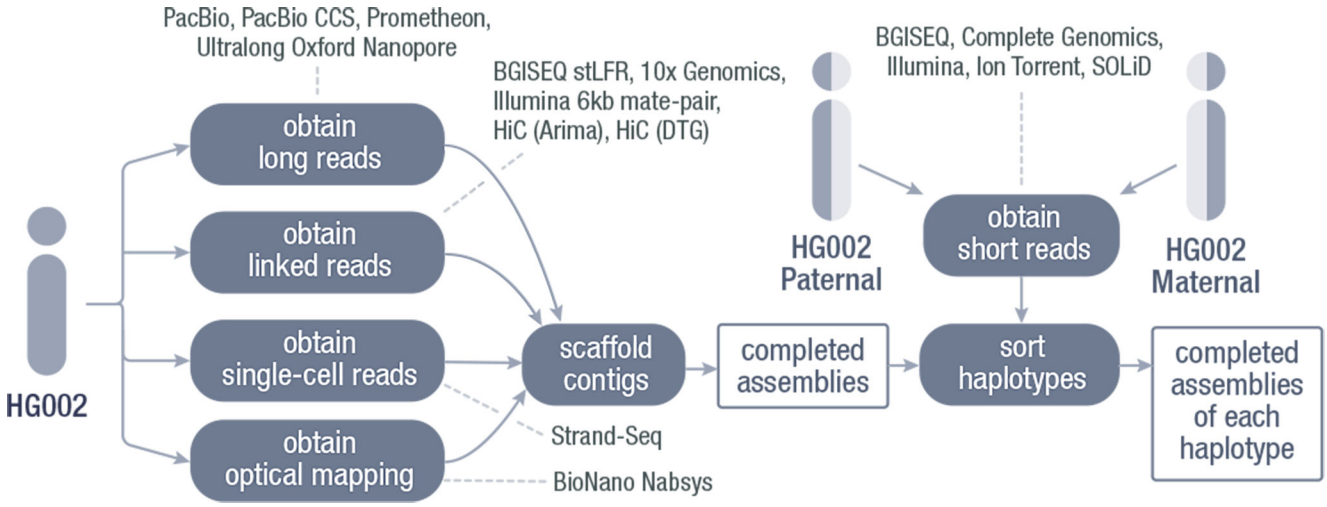

Figure 2: Standards were developed through a pilot benchmark study of one individual.

Multiple long-read and long-range technologies and computational methods were evaluated to develop the combination of platforms and an automated pipeline that provides the most complete and accurate genome graph.

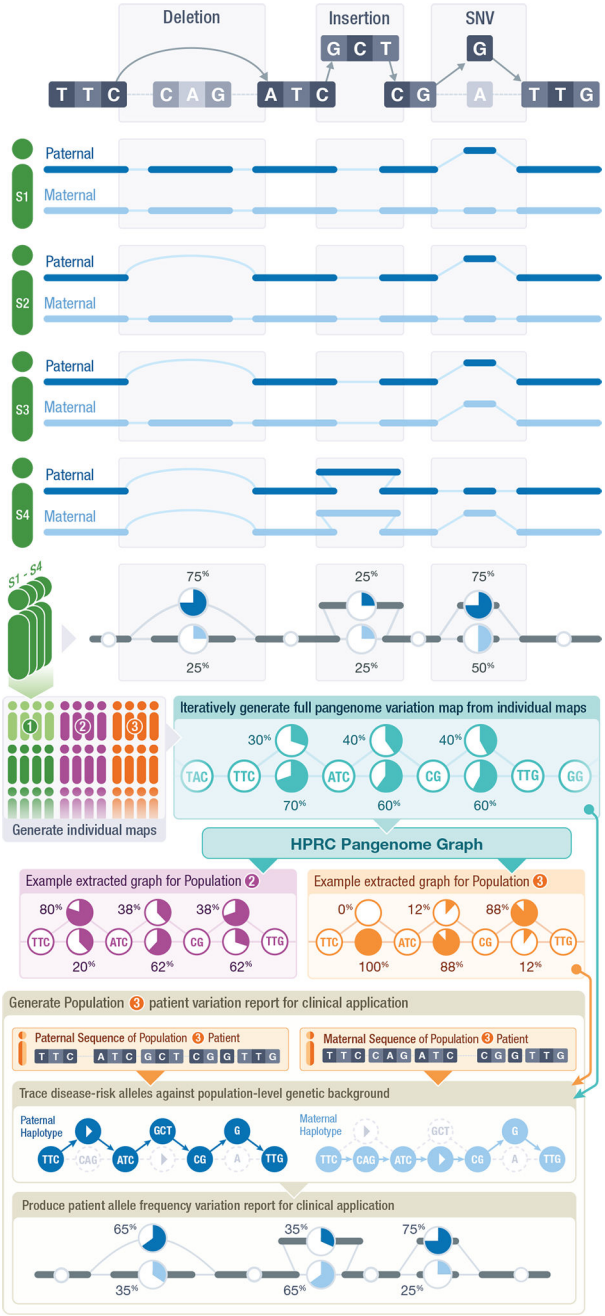

Figure 3: The Human Pangenome Reference.

Graph-aware mappers can be used to genotype samples by directly mapping against the graph. This simplified example shows how to create a pangenome graph for four people and calculate the allele frequency of three variants. Iterating through each individual produces the graph structure, which improves as new genomes are added. Genomic data is arranged into a sequence variation map based on edges. Alternative haplotypes are depicted as alternate pathways across the graph, with the edges being the primary data-bearing elements. The pangenome reference catalogs genomic variation and allows for population- scale analysis thanks to its graph structure. Tracing a path through the network and connecting

sequences at access edges yields haplotypes for individuals. For clinical interpretation, allele frequencies are reported.