Human Pangenome Reference Consortium
Executive Summary: 2019-2022

Population Sampling and Representation for Reference Production: The Human Pangenome Reference Consortium (HPRC) was established in 2019 and aims to generate high-quality reference genome assemblies for at least 350 diploid genomes, with the goal of representing common variation (defined as variants at >1% frequency) in human populations globally. Samples in this project will be selected to improve the representation of human genetic diversity in the human reference genome and provide a richer and more diverse human genomic reference map. A richer human reference map promises to improve our understanding of genomics and our ability to predict, diagnose and treat disease. It can also ensure that the eventual applications of genomic research and precision medicine are effective for all populations equitably. We have assembled a working group of multidisciplinary experts, including legal and ethics experts, to develop a set of principles to guide the project's approach to community engagement, population sampling, and representation.

In the first phase of the project, the HPRC has focused on existing cell lines established from the 1000 Genomes Project (1000GP), which is a deep catalog of human variation in 3,202 participants representing 26 global populations. We selected lines from established 1000GP cell lines housed at the Coriell biorepository that met diversity (ie. genetic and geographical diversity), technical (i.e. less than 5 passages), and regulatory requirements (i.e. consented to open data sharing). In the first two years of our program, we enriched the children of trios to help establish pipelines and best practices to improve phasing and reduce technical artifacts. Male individuals were selected with attention to representing distinct Y-haplogroups. In total, 150 1000GP cell lines were selected for reference production in Years 1-3, which we demonstrate capture >99.9% of single nucleotide variants with >1% allele frequency in the 1000GP dataset. We will continue to select samples to increase our coverage of the remaining common variants from the 1000GP to reach a total of 200 individuals and aim to complete sequence production by mid-Year 4.

We understand that sampling in 1000GP alone is insufficient to reach the goals of this project. There are many geographical regions of the world that are not present in the 1000GP panel, such as West Asia, Oceania, and North Africa. Even in regions that are represented in the 1000GP, the sampling may not fully grasp the population diversity in the region. This can reflect a lack of depth, number of samples, insufficient sampling of diverse geographical groups, or co-existence of culturally endogamous population groups in the region (i.e. in Western and Eastern Africa or in Central or Southern America). Therefore, the second phase of the project is to engage with the broader research community and experts in community engagement for new recruitment to increase diversity. We aim to launch phase 2 recruitment through a partnership with the BioMe biobank in New York City, a highly multi-cultural city and a historical entry point of migration to the US. The BioMe biobank offers >70,000 participants from over 160 countries (>40,000 with available genetic data) who are available to recall and reconsent to participate in our study. New sampling efforts will use similar metrics that we applied for measuring diversity and technical quality established in phase 1 (1000GP), which will enable us to pivot recruitment resources and strategies as we proceed to optimize for genomic diversity. In working with our partners at Coriell, the ELSI team, and the Mt. Sinai Stakeholder Board, we are in a position to begin our recruitment phase and downstream LCL establishment/banking and distribution (we anticipate these samples will enter the data production pipeline mid-to-end of Year 4 and extend into Year 5). We recognize that the broader goal of the second phase is to move beyond solely considering genomic diversity and to evolve a more holistic framework for international engagement. In partnership with GA4GH, we have in parallel launched several efforts to broaden global engagement and outreach.

This working group faces several challenges: We understand that the goal of our program is largely aspirational, as we have an incomplete understanding of the full spectrum of human genetic diversity due to biases in representation in current genomic databases. Our estimates of common single nucleotide variants are currently limited to short reads mapping, and there is an expectation that more than 70% of structural variants are missed in traditional whole-genome sequencing studies. We must expand our survey of genetic diversity outside of 1000GP, including external databases of genetic information (e.g. BioMe 40,000 participants and gnomAD database), and test multiple, well-motivated approaches to ensure we establish a framework for representing diversity. Our efforts will rely on new international engagement, and we will face new challenges with establishing new consents, data production models, and collaborations. Finally, we know that we will need to expand the number of genomes in the next award period to ensure that we are able to expand global partnerships,

improve the representation of large, common structural variants, and by increasing sampling we will represent mildly deleterious mutations (expected at 0.1% frequency).

ELSI: Constructing a human reference genome that is more globally representative raises a multitude of complex ethical, legal, and social implications (ELSI). These issues are complex and cross disciplines, so they must be thoroughly addressed by scholars with relevant content expertise and ELSI research skills. The research and deliberations that drive ethical decision-making must occur in close collaboration and partnership with the technical and logistical teams of the HPRC to maximize impact. Therefore, in Year 3 we have structured an 'embedded ELSI' model at the HPRC to enable real-time ethics engagement to improve the quality of existing projects and initiatives through the coproduction of knowledge. The relevant areas of expertise of the HPRC-ELSI Working Group members include and are not limited to: Research ethics and informed consent: Pearl O'Rourke, Harvard and Mass General; Jean McEwen, unpaid, formerly at NHGRI; and Jessica Mozersky, Washington University; Law and population health sciences: Shawneequa Callier, George Washington University; and Pilar Ossorio, University of Wisconsin; International data sharing policies and ethical frameworks for genomics: Elena Ghanaim, unpaid, NHGRI; Vasiliki Rahimzadeh, Stanford University; and Mahsa Shabani, Ghent University; Clinical genetics and community-engaged genomics for diversity: Clement Adebamowo, University of Maryland and Human Health and Heredity in Africa (H3Africa); and Leroy Hubert, Invitae; and Social sciences, demography and Indigenous data sovereignty: Joseph Yracheta, Native BioData Consortium; and Desi Small-Rodriguez, University of California, Los Angeles. Additional HPRC-ELSI Working Group members include Karen Miga (University of California, Santa Cruz) and Eimear Kenney (Mount Sinai) from the HPRC leadership team; Lucinda Antonacci-Fulton, Sarah Cody, and Heather Lawson from the Washington University Coordination Center, Julian Lucas from the HPRC data management team at UCSC, and Ann McCartney from the NIH Office of Science Policy. Members of the HPRC-ELSI team attend the HPRC working groups to provide insights into emerging ethical, legal, and regulatory issues in real time. Dr. Lawson serves as a liaison to the Global Alliance for Genomics and Health (GA4GH).

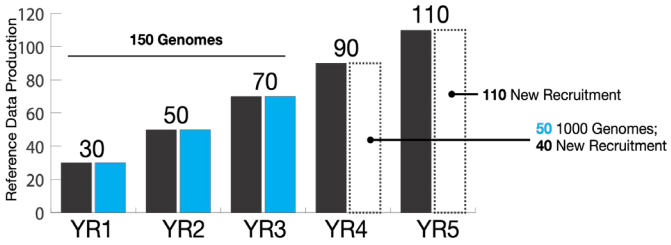
Drs. Yracheta and Small-Rodriguez are working with Nanibaa' Garrison on Indigenous genomic engagement and consultation. HPRC rules embody open science FAIR principles, which are not fully compatible with the CARE (Collective benefit, Authority to control, Responsibility in use, and Ethics) Principles of Indigenous Data Governance. We have organized a focused working group within the ELSI team (Indigenous Outreach) to establish a robust process of engagement with Indigenous academics and tribal organizations such as the Native BioData Consortium (NBDC) and National Congress of American Indians (NCAI) to identify concerns and develop strategies for overcoming barriers to greater inclusion of historically underrepresented groups. Learning from the process that the All of Us program adopted later, our HPRC WG engages with tribal leaders and the NCAI with the goal of building trust, partnerships, and ultimately support for the HPRC in tribal communities.

With a focus on prospective recruitment efforts in Year 4, the ELSI team has led the formal review of the HPRC consent language (ensuring compatibility with 1000GP consent) and assisted in the design and review of outreach materials for prospective recruitment efforts through BioMe. Further, the ELSI team has been key contributors to an HPRC-focused genomic board at BioMe, which is led by Eimear Kenny at MSSM. In parallel, members of the ELSI team have reviewed the use of external IRB, protocols, and documentation.

We have faced challenges with launching an embedded ELSI team in the mid-program (Year 3). The ELSI working group has made important contributions to the HPRC by advising important decision-making that is equitable, and respectful of ethical frameworks, laws, and cultural norms. However, we expect this process to be more efficient in Year 4-5 as new ELSI members are onboarded and become familiar with the various working groups and technical details. Additionally, we understand that engagement with historically underrepresented communities will require respectful forms of engagement and careful study of consent and international policy/law. We envision that the next phase of our project will benefit from having a closer alignment with GA4GH, international legal/policy, and consent.

Human Pangenome Reference Consortium
Executive Summary: 2019-2022

Data Production and Assembly: Data production efforts have continued through Year 3, with Pacbio HiFi, ONT ultra-long read sequencing, and Omni-C short read sequencing as primary products. Our effort leverages strong teams and complementary expertise across the four production centers. UCSC developed the early patents to support the Oxford Nanopore Technology and co-authored the initial publications on ultra-long sequencing. Chromosome conformation capture sequencing, (Omni-C) libraries will be produced at UCSC, where Dovetail Genomics Technologies was founded. University of Washington (UW), Washington University, St. Louis (WashU), and Rockefeller University (RU) have led PacBio sequencing production for reference genome projects and long-read variant discovery. We have partnered with the Coriell Institute which ensures quality standards (biobanking quality metrics, cytogenetic assessment, and microarray) and banks pellets of the same passage lot for all sequencing centers.



To meet the goal of 350 genomes, we increase production by 20 genomes every year. To date, all yearly production targets have been met. We perform yearly assembly evaluations to determine optimal coverage and data mix combinations to provide optimal phased assemblies. For example, coverage metrics for HiFi were elevated from a minimum of 30X to 35X coverage for year 3 samples, and Year 4 efforts are planned in a similar staged approach with a coverage model increasing to >45X coverage. This coverage model, combined with the application of Google DeepConsensus, will increase HiFi coverage to >50X which provides a more compatible product for T2T assembly development. The minimum standard for Omni-C has been maintained with a target of two libraries/sample and >30X overall coverage. With significant chemistry and flowcell advances, the ONT team has seen the greatest improvements within their platform, with Year 2 data alignment rates exceeding 96%, N50 read lengths >70kb, average total coverage from 3 flowcells exceeding an average of 85X, and average coverage of molecules >100Kb exceeding >27X, with expectations of equal or greater final results from the Year 3 samples.

The data management team, UCSC, has established a workflow where each center uploads its data to AWS (S3://human-pangenomics), and released data are synchronized with GCP (the AnVIL) and in parallel submitted to SRA and GenBank. Data are publicly available after quality assessment. Assembly workflows and quality assessment pipelines are available through Dockstore (https://dockstore.org/workflows/github.com/human-pangenomics/) and TERRA/AnVIL.

In years 1-3, for the core assembler, we chose Trio-Hifiasm (Cheng et al., 2021) after detailed benchmarking of a large number of alternatives (Jarvis et al., 2022). Trio-Hifiasm uses PacBio HiFi long-reads and parental Illumina short-reads to produce near fully phased contig assemblies. The assembly process, as well as downstream quality control, were organized to ensure a high degree of completeness, contiguity, phasing, and base-level accuracy. The complete assembly pipeline includes steps to remove adaptor and non-human sequence contamination, and to ensure a single mitochondrial assembly per maternal assembly. In our first assembly release, we ran our pipeline on year 1 (30) HPRC samples and an additional set of HPRC+ samples from collaborating projects to create 94 haploid genome assemblies. These assemblies cover more than 99% of the expected sequence and are more than 99% accurate at the structural and base-pair levels.

Towards the goal of producing complete, gapless haplotype assemblies, the HPRC has teamed with the Telomere-to-Telomere (T2T) Consortium to develop improved assembly methods for finishing. Over the past year, the T2T group has finished the first truly complete human haplotype (CHM13) using a combination of PacBio HiFi and Oxford Nanopore ultra-long sequencing. The CHM13 assembly is proving a useful reference for the HPRC pangenome working group, and additional T2T assemblies would help uncover variation in the most repetitive regions of the genome. In year 3, the T2T consortium improved and automated the CHM13 strategy in Verkko (Rautiainen et al., 2022), an iterative, graph-based pipeline for assembling complete, diploid genomes. Verkko begins with a multiplex de Bruijn graph built from long, accurate reads and progressively simplifies this graph via the integration of ultra-long reads and haplotype-specific markers. The

result is a phased, diploid assembly of both haplotypes, with many chromosomes automatically assembled from telomere to telomere. Running Verkko on the HG002 human genome resulted in 20 of 46 diploid chromosomes assembled without gaps at 99.9997% accuracy. We are shifting our production efforts to test a diverse panel of 20 individuals using a production formula of >50x HiFi, >50x ON-UL, and >60x OmniC.

In year 4 we will continue to optimize data production, assembly methods, and quality assessment. We anticipate moving away from Trio-hifiasm and using a hybrid assembly method (equal HiFi and ONT-UL), like verkko. We are also in the process of improving our quality assessment and polishing pipelines, onboarding new methods developed from the T2T Consortium. In Year 4 we will address the challenge of assembling genomes without parental information (non-trios). One of the biggest challenges of this grant is the need to scale the production of HiFi sequencing, yet we have not observed an increase in throughput in the first three years, which places constraints on the number of genomes and budget. We also acknowledge that automated assembly and quality assessment protocols will likely advance our assemblies “near T2T” with many regions (e.g. rDNA arrays and large satellites/segmental duplications) needing manual curation and finishing.

Pangenome: The Human Pangenome Reference Consortium (HPRC) released the first draft human pangenome reference, containing 47 phased, diploid assemblies from a cohort of genetically diverse individuals (Liao, Asri, Ebler, *et al* 2022). We used sequence graph representation for pangenomes (Eizenga et al., 2020; Paten et al., 2017) in which nodes correspond to segments of DNA. The process of generating a combined pangenome representation is non-trivial because determining which alignments to include is not always obvious, particularly for recently duplicated and repetitive sequences. Thus, we applied three different graph construction methods: Minigraph (Li et al., 2020), Minigraph-Cactus (MC), and PanGenome Graph Builder (PGGB). The availability of these three models provides us with multiple views into the homology relationships in the pangenome while supporting cross-validation of discovered variation. We included the GRCh38 and T2T-CHM13 assemblies within the pangenomes and three samples were held out from the pangenome graphs to permit their use in benchmarking: HG002, HG005, and NA19240.

The draft pangenome captures known variants and haplotypes, and reveals novel alleles at structurally complex loci. We used the Comparative Annotation Toolkit (CAT) (Fiddes, Armstrong, et al., 2018) to lift-over GENCODE v38 annotations using the MC pangenome graph onto the individual haplotype assemblies. CAT lifted and annotated a median of 99.5% of 86,757 protein-coding transcripts per assembly, almost the same as the Ensembl mapping-based pipeline (a median of 99.4% per assembly). In total, we report that the draft pangenome (MC) adds 119 million base pairs of euchromatic polymorphic sequence and 1,529 gene duplications relative to the existing reference, GRCh38. Roughly 90 million of the additional base pairs derive from structural variation. Using our draft pangenome to analyze short-read data reduces errors when discovering small variants by 34% and boosts the detected structural variants per haplotype by 104% compared to GRCh38-based workflows, and by 34% compared to using previous diversity sets of genome assemblies. Evaluation of benchmarked variant calling using parent-child trios using DeepTrio (Kolesnikov et al., 2021) resulted in better performance relative to DeepVariant across all samples of the Genome in a Bottle (GIAB) and the challenging medically-relevant genes benchmarks.

To create a community resource aiding method development and pangenome-based population genetic analyses, we used Giraffe to align high-coverage short-read data from 3,202 samples of the 1000GP (Byrska-Bishop et al., 2021) to our pangenome graph and DeepVariant to call small variants. Further, the Mendelian consistency computed across 100 trios from those samples was comparable to the one computed across samples from the GIAB truth set, indicating a comparable call set quality. Given that our pangenome-based calls showed superior performance in challenging regions, this call set across the 1000GP cohort now provides the genetics and genomics communities with allele frequency estimates for complex but medically-relevant loci.

The group is now focused on the refinement of methods, with attention to a panel of T2T-X chromosomes to refine methods over highly repetitive regions which were omitted in the first draft pangenome release. The pangenome working group is

also working with the assembly working group to prepare for the next round of genome assemblies that will allow the construction of a pangenome containing hundreds of individual haplotype assemblies from the years 1, 2, and 3 sequencing data.

Outreach: HPRC leadership submitted a genome reference perspective article, which was published in Nature in April 2022 (Wang et al 2022). The report details the state of the linear reference and plans for the human pangenome reference. Members of the HPRC community give presentations at scientific meetings to raise awareness about the pangenome reference, and we continue our outreach efforts through our social media accounts. We have supported HPRC workshops at meetings (TERRA/AnVIL HPRC workflows at ASHG). This is a hands-on effort to show how to use the data that has been produced to this point. We have issued a YouTube tutorial to support this training: https://www.youtube.com/playlist?list=PL6aYJ_0zJ4uA_sbziZu84uas3k0iPIKBx. The HPRC steering committee has developed a process of inviting and approving associate members to engage other scientific community members. HPRC has adopted the GA4GH principles, and as noted above, the coordinating center has added a new position focused on outreach activities in conjunction with GA4GH. HPRC leadership has contacted H3Africa and the GEnome Medical alliance Japan (GEM Japan) to discuss ways our groups can work together. We have consulted with the GRC, many of whom are members of the HPRC, about the transition of support for the human reference genomes to create a single-facing entity. To achieve this aim, HPRC leadership has discussed the best way to support the user community. HPRC leadership has also engaged clinical community members to learn how the pangenome can improve clinical results and the barriers we need to understand that might impede adoption. We will discuss additional plans for the engagement of the clinical community at the meeting.

Just recently the genome graphs from the data generated in the first year have been released and are now publicly available. These are the first graph assemblies ever submitted to the INSDC databases (NCBI, ENA, and DDBJ). In order to get these data sets submitted, there were a few challenges given that this is a new data type to the INSDC databases. File formats were discussed, data exchange within the INSDC, updates, and history tracking within these new file formats were all considered in order to make the decision on which format was best. It was decided that the assembly graphs will be in a GFA format which is a standardized format that is fairly small, so should be relatively easy to store. Once the GFA format was decided upon, members of the group started working on testing the submission of this data type. Another detail that needed to be worked out was what metadata would be submitted along with the GFA graphs. There currently is a metadata proposal circulating within the group to determine the final set of data needed to add to the submission to make these graphs as useful as possible. The HPRC year one graph can be found using the umbrella BioProject accession, PRJNA850430. Separate graphs were created using GRCh38 and CHM13 as the backbone of the graphs. Currently, GRCh38 is used by many as the primary reference source, but CHM13, a complete haploid sequence, has more recently been released and is a promising reference source as well.

The coordinating center has organized an internal group focused on annotation. This effort is aligned with the Improvement and Maintenance working group. In the future, we will generate data to examine the function of all genes within the genome for the subset of the HPRC 1,000 samples. The benefit of utilizing these samples is that we can identify the structural variants present in each genome and then compare them to the functional data based on the DNA sequence. From this, we will learn how specific changes within a gene affect the way the gene functions.

Challenges include ramping outreach efforts now that the assembly graphs have been generated, there needs to be a focus on tool development to work with these data sets. This work consists of annotating the graphs, learning the best way to update the graphs, and learning about the needs of the clinical community. There is also a need for the HPRC and the GRC to provide clarity to the user community regarding the support and contact points for the reference community.

International Outreach: The Human Pangenome Project aims to achieve an inclusive pangenome reference resource with global support to improve clinical genomics and biomedical research, provide training and outreach, and share

resources, standards, and workflows. The ultimate goal is to form an international alliance of genomics partners and provide accessible, high-quality genomes that represent the diversity of the human population. We will represent these genomes as a pangenome reference. This effort is funded by a supplement awarded to the Coordinating Center to support Dr. Heather Lawson who is taking the lead on International Outreach. Dr. Lawson has engaged with multiple GA4GH driver projects, including H3Africa and AMED/Gem Japan. She has also had discussions with members of the Genome Institute of Singapore, Sidra Medicine in Qatar, and Xi'an Jiaotong University in China. Members of these institutions have applied for HPRC associate membership. In July 2022, Dr. Lawson organized a special session devoted to international outreach at the International Genome Graph Symposium conference that was held in Monte Verita, Switzerland. During this special session representatives from the above-mentioned institutions as well as from GA4GH gave presentations. Dr. Lawson has integrated with the GA4GH community and regularly attends working group meetings of the Regulatory and Ethics and Genomic Knowledge Standards workstreams. Additionally, she meets monthly with GA4GH leadership to discuss progress. Her current efforts are focused on organizing a Working Group that will include both HPRC and GA4GH members, and will focus on defining international partnerships, establishing ground rules for international participation, and ensuring all issues of consent and regional governance, as well as data and technical standards, are considered.