Research Strategy - Component 1, Aims 1-3.

Introduction

The Human Genome Reference Program (HGRP) is charged to advance the human reference to be more inclusive, complete, and accessible. When fully funded, it will consist of several distinct efforts, including developing high-quality sequences for many new genomes through the High Quality Reference Genomes Project (HQRG), the development of new genome representations through multiple U01s (Genome Reference Representations (GRR)), the development of new sequencing technologies and bioinformatics tools, and the provision of an updated Human Genome reference release. The Human Genome Reference Center (HGRC) is tasked with coordinating logistics for the overall HGRP, engaging with the larger biomedical user and stakeholder communities, and maintaining, improving, and managing releases of a new Human Genome reference.

As the coordinating center for the HGRP, the Pan-Genome Reference Center (PaRC) will play a leading role in maintaining, improving, and providing an authoritative Human Genome reference to the larger community. We will maintain a library of materials and reference genomes, and automate the construction and evaluation of new Human Genome reference candidates. Using staged releases of new Human Genome references, we will work to ensure access to tools to use the human Pan-Genome Reference. PaRC will operate with governing principles adopted from the successful Genome in a Bottle Consortium (GIAB) to assure confidence in HGRP products, moving towards **open and transparent validation** of the pan-genome constituent genomes, the synthesis of the pan-genome reference from its constituents, and the user-developed tools that use the pan-genome reference.

Engagement with other components of this grant: This component (Component 1 of the overall PaRC HGRC application) focuses on technical activities around the Human Genome reference. Engagement with the user community is described in Component 2, Aims 4 and 5; coordination of the HGRP is described in Aim 6 (Component 3), and strategic coordination with external stakeholders is described in Aim 7 (Component 3).

Significance

The two primary technical challenges for the Human Genome reference lie in representing human genomic diversity, and in providing a representation of that human diversity in a form that is usable for a wide range of downstream purposes. There are also challenges of completeness, especially of the hard-to-sequence and highly repetitive regions of human genomes, and the need for toolchains to support working with the Human Genome reference.

Even with many high quality genomes that adequately sample human genetic variation, we face the challenge of representing the available genetic variation in a usable way. To be a true reference, the Human Genome reference must simultaneously serve the needs of many downstream users and stakeholders, including biomedical researchers, clinicians, population geneticists, technology developers, and others. A reference should therefore provide a single coordinate system that includes the full set of structural variation present in its constituent genomes. Considerations for downstream usabiility are manifold, but the ideal reference would provide chromosome context for any identified human sequence, enable unambiguous data interpretation at all clinically relevant loci, and introduce no systematic error or bias in genome-wide analyses (Schneider et al., 2017).

The current reference release does not satisfy these considerations. A recent study analyzed structural variants (SVs) for fifteen human genomes and resolved 99,604 common SVs (Audano et al., 2019). Of these SVs, 2,238 were shared among all discovery genomes and 13,053 were present in the majority of the genomes, indicating minor alleles or errors in the current reference. Another recent study found 1842 non-reference unique insertions (NUIs) from the de novo assembly of linked-read WGS data on 17 individuals across five populations concluding that the reference cannot capture much of the genetic diversity across the different continental groups (Wong, Levy-Sakin, & Kwok, 2018). In this study, Africans have the highest abundance of NUIs, while Europeans have the fewest. Another recent study found that the African pan-genome assembled from 910 individuals has as much as 300 megabases of DNA that is entirely missing from GRCh38 (Sherman et al., 2019). A de novo assembly of two Swedish genomes also found about 10 Mb of missing sequences from GRCh38 (Ameur et al., 2018). The authors also showed that the inclusion of missing sequences into the reference improved alignment and variant calling from short-read sequencing data. Up to 1 Mb of the new sequences were assigned to the Y chromosome. Note that the Y chromosome has one of the largest proportions of repeated sequences (> 50%) which is the main reason it has not yet been fully sequenced (Quintana-Murci & Fellous, 2001). Lastly, additional studies show that using ethnicity-matched reference increases accuracy in imputation and risk predictions (Huang et al., 2015; Martin et al., 2017; Nagasaki et al., 2015; van Leeuwen et al., 2015).

Representing this substantial amount of human variation in a single reference is challenging. There has been considerable interest in the notion of developing and providing the Human Reference as a pan-genome, where pan-genome is defined broadly as a *collection* of human genome sequences that can be used jointly as a reference (Ballouz, Dobin, & Gillis, 2019; D. M. Church et al., 2015; Garrison et al., 2018; Sherman et al., 2019). However, there are many detailed considerations for making a pan-genome. These include topology of representation, definition and stability of coordinate system, metadata and provenance, toolchains for mapping reads, interrogating variants, and comparing assemblies; and tools and databases for interpreting variants for a variety of uses. **There is no simple path forward at present for a pan-genome version of the Human Genome reference, in part because there is no mature scientific consensus on the form the pan-genome should take, and few prospective tools available.**

A related major challenge for new releases is that of adoption. Despite the existence of a substantially improved release, GRCh38, and considerable evidence that using GRCh38 increases the sensitivity and specificity of downstream analyses (Chapman, Meynert, Church, Johnson, & Hofmann, 2016; Guo et al., 2017; Pan et al., 2019), the alternative loci in this release have not seen widespread adoption. This is for multiple social, technical, and business reasons, including the lack of visualization and analysis infrastructure, the challenges of lifting over database coordinate systems, and the expense of establishing and validating clinical performance in commercial practice. As such, any proposal for building a new Human Genome reference release must situate itself within the larger context of available visualization, analysis, and interpretation tools, as well as the techno-economic environment of users.

Innovation

This component of our proposal focuses on the combined **social and technical** aspects of building, releasing, and maintaining a new Human Genome reference. We propose **automated build and evaluation processes** that support a **staged release strategy** for new references. These processes will be provided transparently to the methods development community for evaluation of new reference representations and associated toolchain. We will build new Human Genome reference candidates from a high quality library of genome sequences, and do orthogonal validation with **openly consented, public, open sequencing data** for which

we can also assure long-term access to genomic resources, and which we will consistently refine with innovative sequencing methods.

A key aspect of our approach is to involve a wide array of early adopters, users, and stakeholders in the iterative development of criteria for evaluating new releases, via working groups and a Request for Comments system (described in Aim 7). We will implement a formalized process that moves release candidates through alpha and beta stages before a full final release. This process relies critically on our automated construction and evaluation processes, as well as our community outreach effort (Component 2) and stakeholder engagement. **The strategy outlined above offers maximum flexibility to the HGRP and larger community**, integrates an evaluation of the entire relevant toolchain into each phase, and provides an opportunity to evaluate new potential reference representations in the context of their overall toolchain. **While this is a standard approach in software development, this has yet to be applied explicitly in genomics.**

Last but not least, we propose to coordinate an ongoing HGRP Participant Program, built on the Personal Genome Project and the Open Humans platform. Our goal for this program will be to maintain an ongoing relationship with HGRP participants in order to maximize the opportunity to build on their contributions.

Preliminary Results

The Personal Genome Project (PGP) is a longstanding and highly successful effort to create openly consented human genome resources that has been led by **PaRC Co-PI George Church** (Harvard) since 2005. The PGP has pioneered the Open Consent Model, enabling the world's first and only truly open-access platform for sharing individual human genome, phenotype, and medical data. The consent process educates potential participants on the implications and risks of sharing genomic data, on how the PGP study is structured, and about what they can expect from their participation (Ball et al., 2012; G. M. Church, 2005). Open consent has assisted the creation of the world's first human genome reference material, Genome In A Bottle, for calibration and lab performance measurements (J. M. Zook et al., 2016, 2014; J. Zook et al., 2018)). In 2017, the revision of the Common Rule added support for "broad consent" models governing the future use of data and biospecimens.

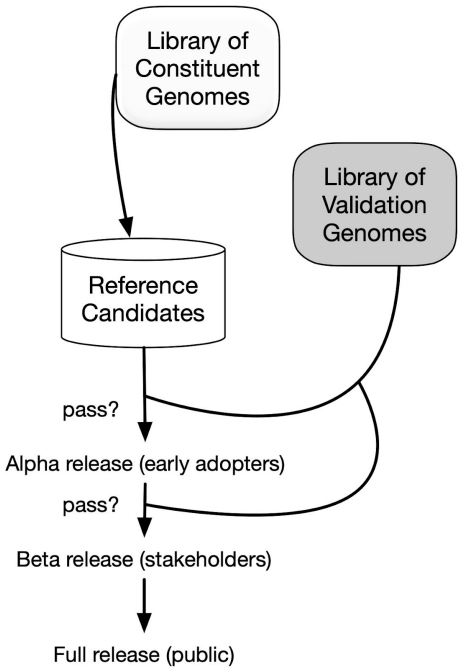
PaRC will drive technical convergence to a new Human Genome reference toolchain via open evaluation and publication of evaluation resources**.** Our strategy is based on the extremely successful Genome in a Bottle (GIAB) project. **PaRC Co-PI Marc Salit leads the Genome in a Bottle Consortium** and is a founder of the Joint Initiative for Metrology in Biology. GIAB has extensive expertise with benchmarking, consensus-based best practices, and standards. GIAB data has been widely used for analytical validation and technology development, optimization, and demonstration. Additionally, GIAB data has proved useful to measure improvements in alignment and variant calling resulting from the use of GRChr38, e.g. (Pan et al., 2019; Schneider et al., 2017), and to test and validate the upgrade of bioinformatics tools to use GRChr38 (Chapman et al., 2016).

PaRC Co-PI George Church is developing novel techniques to sequence the genomic "dark matter" deep within highly repetitive regions. He has been exploring methods that use in situ sequence detection to provide a high-throughput source of physical partitioning information, based on the biological fact that chromosomes often reside in distinct territories in the nucleus. He has been developing in situ chromosome tracing and sequencing methods with collaborators in another grant (NHGRI RM1 HG008525), with the goal of detecting sequence anchors at intervals along each chromosome (telomere to telomere) at the resolution of the imaging method, currently about 500 bp and 20 nm (Nir et al., 2018). In these methods, short sequences on many

chromosomes may be read in situ in many cells at once, either by designing oligonucleotides (“Oligopaints”) that specifically hybridize to them or by enzymatic methods to amplify random segments in situ. Supplementing long reads with in situ chromosome territory information will allow small variations between long repetitive homologous sequences to be mapped to distinct territories, thereby allowing assignment of these reads to e.g. a particular centromere. He is also exploring the use of in vitro stretched DNA methods (Payne et al., 2013) as a way of physically partitioning single molecules. In this approach, long single-molecule segments of chromosomal DNA are stretched on array surfaces and interrogated with probes that can variously identify distinct patterns of AT rich or other sequence characteristics. These sequence patterns can be used to partition long reads and improve assembly of distinct haplotypes or of very long repeats such as centromeric sequences.

Approach

Our approach to maintaining, improving, and providing the Human Genome reference is spread across Aims 1-3 and diagrammed in Figure 1. First, we will gather and curate a library of high quality genome sequences, with an accompanying validation collection of openly consented materials (Aim 1). Using the library of constituent genomes, we will automate the construction of candidate Human Genome references, evaluate them with the validation collection, and provide tools that enable others to use our evaluation metrics (Aim 2). New Human Genome references will be released using a staged release strategy (Aim 3) that engages with early adopters and evaluators at an alpha stage, then reaches out to infrastructure providers and databases at the beta stage, before a full release is made.



Specific Aim 1: Build and maintain a library of high quality reference genome sequences.

The Human Genome reference must be composed from a suite of high quality genome sequences encompassing as much sequence diversity as possible. These genomes will be produced in the HQRG element of HGRP. **We will supply independent validation data using GIAB and GIAB-like reference samples, reads, assemblies, and call sets.** These validation data will not used to create the pan-genome but will be used to evaluate the utility of the Human Pan-genome reference candidates, along with the performance of downstream toolchains. Additionally, the underlying validation samples can be used by the HQRG to help test and evaluate their methods.

While ideally the Human Genome reference would be built on openly consented materials that permit the widest possible use, this will depend on the principles used to select genomes in the HQRG. We are enthusiastic to work with the HQRG, NHGRI, and other stakeholders to maximize the use of open consent for all elements of the HGRP.

Subaim 1.1: Coordinate the process of choosing new materials for the HGRP library.

We will work with the HGRP and external stakeholders to curate a library of high-quality genome sequences that represent a wide range of human diversity. This library will contain both putative constituent genomes (e.g. HQRG and others) and HGRP-and other project developed validation data sets. No distinction will be made in terms of inclusion in the library, but we plan to use only openly consented materials for validation sets.

Sources of materials

Along with new diverse genomes provided by the HQRG, we can leverage existing high quality raw data, assemblies, and existing cell-lines for incorporation into the HGRP genome library. We will evaluate materials from the McDonnell Genome Institute (MGI) Reference Genomes, Genome in a Bottle Consortium (GIAB), 1000 Genomes Project, Simons Genome Diversity Project, the Personal Genome Project (PGP), and the HGDP-CEPH Human Genome Diversity Cell Line Panel. We will retain full provenance and metadata for each constituent genome.

Evaluation criteria and metrics for inclusion

Evaluation criteria will be focused on metrics related to quality and diversity measurements inspired by the GIAB benchmarking call sets and tools. Ideally, the HQRG will resequence our openly consented validation samples as process controls, and then run the GIAB sequence analysis pipeline. These controls would provide a first set of objective performance measures of HQRG sequencing and assembly processes.

For example, Zimin and Salzberg assembled two human genomes using combinations of short and long read data (pers. communication). They used both the widely-studied NA12878 genome and the son (NIST ID HG002/PGP ID huAA53E0) from the GIAB Ashkenazi trio. For NA12878, they generated Oxford ONT and Illumina reads, and for HG002, they generated ONT, PacBio, and Illumina reads. After assembly, they tiled the polished assembly onto the human chromosomes using GRCh38 as a reference with Nucmer4 (Marçais et al., 2018). Then, chromosome-sized scaffolds were built, oriented, and ordered with the scaffolding tool in MaSuRCA (Zimin et al., 2013). They were able to order and orient about 96% of the sequence from both of the HG002 assemblies and found near-perfect chromosome agreement. It is likely that some amount of sequence will be unique to each new genome, as they demonstrated with the African pan-genome (Sherman et al., 2019). Most of these lineage-specific sequences should occur within contigs when assembled, in which case they will be placed on the human backbone and will appear as insertions with respect to GRCh38. Any sequences that are not linked to chromosome locations are added to the set of unmapped scaffolds. Our proposed validation process would proceed from the mapped scaffolds to do in-depth comparisons of variants as a precursor to inclusion in the library.

Even where variants or assemblies cannot be compared comprehensively due to novel sequence, we can compare the regions of the HQRG genomes that are well sequenced to the high-confidence regions for GIAB samples. If an HQRG genome behaves poorly in those high-confidence regions, more investigation would be needed.

Capturing human genetic diversity: To maximize diversity in the constituent genomes for the pan-genome, we will consider metrics around diversity and work with the HGRP to develop consensus around these metrics. Initially, we can use ancestry estimation tools [e.g. STRUCTURE, ADMIXTURE] or existing ancestry data to include ancestries not well represented in the reference or in current catalog of human genetic variation, for example North African, East African, Oceanian and American ancestries. Additionally, we can produce overall

measurements of the diversity of the genome using techniques such as PCA (Price et al., 2006) or calculating heterozygosity (Gibson, Morton, & Collins, 2006; Samuels et al., 2016) and use those to guide the inclusion of genomes that represent additional content. It is also important to also initially consider metrics around "missing" assemblies that contain sequences that cannot be placed on the current reference genome including which genomes contain missing assemblies and the assemblies' occurrence relative to subpopulations and/or sets of genomes (e.g. a missing assembly occurs in other genomes of the same subpopulation as opposed to only occurring in a single individual or a single family) (Sherman et al., 2019).

Once we have an initial subset to consider as constituent genomes, we will track metrics that measure how much the reference genome changes as we add information from those individual genomes. This will allow us to fine-tune which genomes to include and include genomes that add representative new sequences and/or fill gaps. Additionally using this information, we can develop statistics for a given sub-population and use those to estimate how many individuals we would need to sequence and add to capture the variation in that population and adequately represent it.

Development of openly consented validation data sets

In addition to recruiting high quality whole genome sequences for the library from the HQRG and others, we will also generate high quality validation sequencing data sets from groups underrepresented in the library, coordinated with ongoing efforts of the Genome in a Bottle consortium, who plan to moderately expand their collection of authoritatively characterized genomes. Briefly, we will use openly consented materials to develop a large, diverse set of data (e.g. high-depth paired-end short read whole genome sequencing (WGS), long mate-pair WGS, pseudo long read/read clouds WGS, and long read WGS) for the validation library. This data will be used to provide GIAB-like packages of data (free and open data from raw sequencing to high quality call sets) to help validate and test reference candidates (Aim 2.2), demonstrate performance of sequencing methods and tools (Aims 2.2 and 2.3), and port tools to the new reference (Aim 2.3).

We plan to start this effort with Illumina sequencing of up to 10 trios per year (30 x 5 = 150 participants) using the Harvard Personal Genome Project. External funds can be used to created PBMCs and cell lines to be added to the existing Coriell repository. Fibroblast and EBV-transformed lymphocyte cell lines have been established with samples collected from the PGP-10 pilot cohort as well as several other PGP participants; these have been made available through Coriell. Currently the NIGMS HGCR contains >100 PGP samples of various cell types including B-lymphocytes, fibroblasts, and iPSCs. We will use a similar process to collect blood samples and create cell lines from additional PGP participants.

Funds for sequencing at the Church Lab will be balanced between this subaim and targeted sequencing (Subaim 1.3).

Subaim 1.2: Manage consent and provide access to materials.

Creation of the HGRP library will involve human participants whose specimens and genetic information contribute to the library. The inability to anonymize genomic data means that these individuals will have an ongoing relationship with the project -- whether or not this is intended. Furthermore, additional research that builds on the reference genome and validation library may wish to seek new sample donations from reference participants, e.g. for sequencing with new technologies or epigenetic profiling. It is in the interests of the HGRP to assure this capability and develop the human pan-genome reference as an enduring resource. **Coordinating and developing the collection from these open-consented individuals will be part of PaRC's HGRC activities.**

PaRC will organize and engage participants for ongoing availability in new research. We see many reasons for supporting this separately from the HQRG: without it, involvement of the original study team would be necessary for participants to be approached for new specimen collection and other research, and additional activities would need to be reviewed and approved by the study's IRB. Such a process can be especially complex when recruitment is for a different study team. Coordination between two research teams, each with their own IRB oversight, is often prohibitively time-consuming. Furthermore, there is a need to coordinate news sharing, inquiries, and other community needs in an ongoing basis with these participants, separate from HQRG research activities.

Thus, coordination of participants on an ongoing basis is a key logistical and coordination priority for the HGRC. We propose establishing a program that coordinates participants from the original HQRG study, *should they choose to join it*, for the purposes of ongoing communication, community representation, and availability for new research.

Open Humans supports automated methods for online interactions, including depositing data files, online consent and authorization, data management and access, and messaging to participants. The platform also provides privacy-enhancing features, including interaction via random identifiers, and meets US and EU data privacy regulations. Projects conducted in the platform can optionally use web APIs to fully automate engagement with project members.

Open Humans has experience with community coordination and working with various patient and participant communities, including academic projects (American Gut, Personal Genome Project) as well as patient-led communities (Nightscout, OpenAPS). Thus, in addition to providing technical features, the organization is experienced with engaging study teams and IRBs to support participants and research in the platform.

Our team therefore proposes to invite participants to be part of an ongoing HGRP Participant Program, coordinated and managed by the HGRC, and separate to the research studies that engage these participants for sample collections and other research.

Participants who opt to be part of this program will have an additional, ongoing contact point for inquiries related to the activities of the HGRP. They will optionally receive news and updates from the HGRP regarding its work, and optionally be available for contact by research teams, as well as other forms of contact. We will provide a format for interacting with researchers, other participants, and other interested individuals, enabling them to contextualize their ongoing relationship with this research. The format each option takes will be informed by interactions with the HQRP team, their ethics oversight, and members of the team involved in participant engagement. If possible, it will also engage the participants themselves to understand their preferences regarding news, contact, and community features.

Once a design has been agreed upon by these stakeholders, the program will then create informatics infrastructure that supports these features, with user accounts for participants that opt to be in this additional program. We will leverage the existing features of Open Humans, operating as a project within that platform.

Subaim 1.3: Targeted sequencing and novel methods to provide updates and "gap filling".

We expect there to be an ongoing need to improve specific whole genome sequences in the library, e.g. by doing targeted sequencing of challenging regions from one or more constituent or validation genomes. We also expect a variety of sequencing technology improvements to emerge over the next 5 years, nucleated in part by

connected research grant funding. New technologies and targeted sequencing will help resolve current issues with two big sources of assembly error: large repetitive sequences, and complex allelic diversity in the genome.

We plan to use existing and novel technologies to provide updates to the genome assemblies (from HQRG and others) in the HGRP library. For example, error reports on constituent genomes may need to be evaluated through targeted sequencing, including high resolution genotyping. Sequencing of essentially haploid hydatidiform moles (Steinberg et al., 2014) can provide a way to resolve complex areas in the genome. Challenging-to-sequence “dark” regions may be targeted via novel techniques being developed in the Church lab and elsewhere (see Preliminary Results). For example, we have been exploring methods that improve on flow-sorting chromosomes or BAC libraries (Chu et al., 2017; Jain et al., 2018) by using in situ sequence detection to provide a high-throughput source of physical partitioning. By supplementing long reads with in situ chromosome territory information, we can map small variations between long repetitive homologous sequences to distinct territories, thereby allowing assignment of these reads to e.g. a particular centromere.

Specific Aim 2: Automate construction and evaluation of Human Genome reference candidates.

The goal of the pan-genome is to represent all major human variation in a single reference system, refer to locations and variants unambiguously via a single coordinate system, and easily move between reference versions without requiring users to re-calibrate all of their data for each new release of the genome. The current linear human genome reference (GRCh38) includes ‘alts’, sequences for alternative loci, but they are rarely used, both because researchers find them difficult to work with, and because few tools support them. Development of a pan-genome reference must improve upon this system both to properly represent the variation and to facilitate adoption and utilization. However, the scientific and technical knowledge to construct and evaluate a pan-genome reference is not yet mature. Multiple competing strategies to represent a pan-genome are emerging; these will need to be implemented and evaluated as PaRC develops, disseminates, and maintains a pan-genome reference.

Emerging methods for representation of pan-genome sequences have centered around building genome graphs to facilitate representation of genomic variation (Garrison et al., 2018; Paten, Novak, Eizenga, & Garrison, 2017; Rakocevic et al., 2019). These strategies excel at improved representation of variation, but introduce a number of challenges to adoptions, in particular incompatibility with existing tools and formats and the increased complexity inherent in deviating from a linear-style reference. The significant effort that would be required to transition to a graph-based pan-genome system necessitates implementation, evaluation, and community engagement with the most promising pan-genome strategies that emerge over the course of the HGRP. As critical as the representation itself is, these approaches will need to be evaluated in the full context of an associated toolchain and downstream implications.

Automating both building and evaluating candidate Human Genome references is key to iterative development and refinement, and is a critical technical element of PaRC. No development of a reference of similar complexity has been attempted. An open PaRC-maintained set of resources will enable the HGRP and stakeholder communities to work together to create the missing knowledge and come to consensus on representation and tools. See in particular Specific Aim 2.2, where PaRC’s tool evaluation platform is described.

We will work with experts in each representation to create release candidates from our selection of constituent genomes, and identify how best to fine-tune their representation, while providing metrics, metadata, and

provenance for their representation. **We believe that our role as a coordination center is not to be an expert in each representation but to allow the experts in each to guide us in construction and use, while we document the process through automation and evaluation.**

Subaim 2.1: Automate construction of new Human Genome reference candidates.

We will work with the GRR awardees, Novel Nucleic Acid Sequencing Technology Development (R01, R21, R43/R44) awardees, and other members of the community to automate the construction and evaluation of each new Human Genome reference candidate, execute them on our own compute platform, and make them Findable Accessible Interpretable Re-usable (FAIR) by providing full provenance with version controlled workflows (see (Johnson, Alexander, & Brown, 2018) for an example of this strategy). This will ensure reproducibility of the build process, enable direct comparisons of the outputs, and enable rapid production of new candidates as the genome representation software evolves.

Our preferred workflow automation tool is snakemake, which supports flexible manipulation of inputs, the use of scripts in any programming language, fully Dockerized workflows, job distribution across Slurm and Kubernetes clusters, and export to Common Workflow Language. All construction workflow source code will be made available under the Apache 2 open source license.

Subaim 2.2: Comprehensively evaluate Human Genome reference candidates.

Methodological evaluation and orthogonal validation data sets are critical in order to inform and drive convergence towards a canonical representation format for future releases. As we simultaneously evolve the content, the representation, and the computational workflows for variant calling, we will need to measure and evaluate the impact of these changes continuously, and feed these results back into the HGRP and larger community. Our goal in this subaim is to establish performance metrics and quality benchmarks that can be used to evaluate reference candidates. A set of objective measures will be produced for each candidate to support development, reporting, and decision making purposes.

Ultimately, to be widely adopted the pan-genome and associated toolchain must improve upon the existing reference for a set of community-guided criteria. **These criteria should stretch from read placement to improved clinical applications for all ancestries**. Possible ways to measure this improvement include read placement of previously unplaced reads, the behavior of native validation sets, the existence of verifiable new alleles that cannot be mapped back to previous references, and the extent of continuous genome coverage (gap filling).

For each new reference candidate we will use our validation data to measure the overall process in terms of false and true positive alignments, single nucleotide polymorphisms, small insertion/deletions, copy number and structural variants, and placement of previously unplaced reads from our validation set (see (Chapman et al., 2016; Guo et al., 2017; Pan et al., 2019; Schneider et al., 2017) for detailed results on GRCh38). We will evaluate the impact on clinical applications by examining a subset of important clinical alleles for mapping and variant calling, and report on the effect of clinical interpretation. We will also benchmark the performance of the entire toolchain, to provide objective information on the computational requirements.

This evaluation strategy will control the process of building the Human Genome reference, as well as evaluating at least one entire toolchain for variant calling. Differences in call sets from one reference release to the next release candidate will be evaluated to ensure new references perform equally or better than previous releases. This will result in the reliable identification of changes, improvements, and errors in each version of

the reference, along with **at least one fully functioning and benchmarked variant calling toolchain for each Human Genome reference candidate**. This process can also be reliably used to help identify errors in new sequencing and bioinformatics processes applied to new reference representations (explored in Aim 2.3).

As with the construction of reference candidates, we will incrementally automate our evaluation processes, put them in version control, and provide full provenance and traceability for each benchmark run. An important part of the evaluation will be human input from curators in the HGRC, the HGRP, and the larger community. We will provide entry points for feedback and involvement in the benchmarking and evaluation process, which will be added to our automated processes as practicable; we expect there to be considerable interest in joining the evaluation working group (see Subaim 7.1).

We note that GIAB-like high-confidence call sets tend to exclude more difficult types of variation and regions of the genome. Evaluation of new references or variant-calling methods against these standards may find lower accuracy in high-confidence regions even when the method produces much higher accuracy in difficult regions. We will take these biases into account in our metrics and evolve our evaluation process to progressively include more difficult types of variation.

All workflow source code for evaluation will be made available under the Apache 2 open source license.

Subaim 2.3: Provide resources for validation of toolchains that operate with the pan-genome reference.

Driving adoption of new Human Genome reference releases is a major focus of PaRC, and a key criterion for adoption is the availability of validated toolchains for working with the new reference. This includes liftover tools for mapping alleles from one reference to another, the ability to map reads and call variants, and the use of standard file formats (VCF, GFF, BED). As part of our staged release strategy (Aim 3), and our engagement with the user and stakeholder communities (Aim 7), **we will work to systematize and release the evaluation system in Subaim 2.2 for others** so that as much of our evaluation pipeline (data and code) as possible is available to toolchain authors.

We expect that our processes will be broadly usable by methods developers, and will feed into both experimental and bioinformatic methods development within the HGRP, as well as the larger NHGRI, and genomics communities.

Specific Aim 3: Define and implement a staged release strategy.

We propose a staged release strategy that maximizes opportunities for community engagement and feedback by proceeding through a series of alpha and beta releases, in line with our community outreach plans (Component 2). The criteria for moving from release candidate to alpha and beta releases, as well as a full release, would be developed iteratively in concert with the HGRP and the larger stakeholder community via working groups and Request for Comment (RFC) processes (Aim 7).

As part of our multifaceted approach to community and stakeholder engagement (see Aims 4, 5, and 7 for details), we will work with our partners to develop a consensus release roadmap. In terms of actual releases, we propose a conservative strategy: for the first major release we would plan to produce a version that includes significantly more human diversity but in a linear form that is usable by extant software, and then plan a more avant garde release closer to the end of the initial 5 year period. (See proposed release timeline at bottom.)

Subaim 3.1: Define a staged release strategy.

As per Subaim 2.2, each release candidate will be benchmarked with validation data. A summary of the results, as well as the results themselves, will be provided publicly to enable human-guided and machine evaluation of each new release. We will also grow a set of analyses tailored to different stakeholder needs so that different users can build an understanding of the impact that the release would have on their overall scientific process.

Once we have one or more candidates for a new release based on our evaluation metrics, we would enter the candidates into an alpha release stage. Specific criteria for this will be developed in conjunction with stakeholders (see Aim 7 for stakeholder engagement) but will include the existence of a toolchain for variant calling, a draft liftover implementation for translating between coordinate systems, some level of improvement in read mapping, and minimal impact on clinical implications. Alpha releases will be made available to an early adopter community recruited in Component 2, for the purpose of toolchain development, evaluation, and comparison. Iterative engagement with these communities is key to understanding the drawbacks and challenges of a new release, as well as gaining buy-in for new reference formats; these criteria would be added to RFC drafts.

Alpha releases would move into a beta release stage once the major concerns of the alpha release community have been collated and addressed, to the extent possible. The specific criteria will again be developed in conjunction with the larger community, but would include maturity of toolchain and liftover implementations. These criteria will also be iteratively formalized in draft RFCs.

Our beta release partners would be recruited from amongst the community of stakeholders. In particular, we see infrastructure stakeholders such as genome browsers, databases, and analysis commons as key partners in any transition from the current GRCh38. We will work with these partners to ensure some level of ecosystem support for the full release is made, with formalized criteria emerging through the RFC process.

Subaim 3.2: Release new references with metadata and informatics tools.

As part of our release process, we will formalize the current versioning scheme in which major versions (39, 40) contain additional genomes representing human diversity (Aim 1), while minor versions (39.1, 39.2) are iterative improvements of that major version that do not add new constituent genomes. We will plan for a time-based release system, in which new versions are released quarterly or semi-annually, with version numbers adjusted according to the content guidelines. We anticipate regular minor versions with few major versions (see below, and Proposed Timeline).

Each reference will be released to the public as rapidly as possible, both to GenBank and also to the NHGRI Human Genome Reference Center (HGRC) and the NHGRI AnVIL (Analysis,Visualization, and Informatics Lab-Space) data repositories. We may also provide them via a variety of other portals, including InterPlanetary File System (IPFS), Dat and an Arvados collection. Releases will be accompanied by metadata files containing provenance and evaluation information.

For each release, we will provide workflows and tutorials that demonstrate how to make use of the release. (See Aims 4 and 5 in Community Outreach, Component 2.) We will link out to associated visualization platforms and analysis commons for each release, and update the links as more partners and databases adopt each release. (Component 2.)

We will work to identify unmet software needs (e.g. representation and release-specific liftover technologies for coordinate translation) and fill those needs that we can. We will produce a substantial amount of metadata and workflow tooling as part of this Component, and this will be made publicly available and supported for use by external consumers. (See Component 2.) We will also provide sequence search and metadata browser tools across the releases that enable users to find versions of human genome reference that contain specific input genomes and haplotypes. (See Subaim 4.1, Component 2.)

Subaim 3.3: Track problems with Human Genome reference releases and maintain a patch library.

For each reference release, starting with GRCh38, we will resolve errors and gaps in the reference as part of maintenance of the reference release. We will also maintain a computational list of reported errors and mis-assemblies in the reference and provide a point of contact for reporting new concerns.

Several approaches can be used to characterize more difficult regions of the genomes used for the HQRG. For example, pedigree information can be used to filter out likely errors in difficult regions, particularly in homopolymers, and phased pedigree-based calls may be used to supplement current high-confidence calls. Long reads and linked reads can enable high-confidence calls in difficult-to-map regions of the genome.

The list of computational errors will be provided publicly in an issue tracker, and for each verified issue we will create an evaluation metric that ensures that future reference releases do not have that error. This growing list of automated evaluation metrics will be applied to each version of the reference, past and future, and the results reported on a public website in both human-readable and computationally readable forms; this will enable the development of automated quality evaluation, exploration, reporting and summarization tools by us, and by others.

Local computational errors will be fixed via a patch process (see Subaim 1.2), in which the transformation of one version of the reference to another will be specified declaratively. New minor versions and releases of the reference (see Aims 2 and 3) will be created through iterations of these automated processes, enabling traceability. These patches will be provided publicly as well. In general, smaller updates will be fed back into the reference release in the form of patches, guided largely by the “tried and true” format used by the GRC (https://www.ncbi.nlm.nih.gov/grc/help/patches/). Initially, patches will be marked as either “fix” patches (changes to existing assembly sequence) or “novel” patches (new alternate loci to the assembly). We expect most maintenance patches will fall into the “fix” category where “novel” patches will most likely be generated within the improvement workflow but could be created in maintenance. As the reference genome format evolves, we will consider the possible introduction of new types of patches or a new taxonomy if the need arises. Larger updates, in particular those that change the underlying structure (which we loosely term topology) of the reference or the sequence coordinates, will result in a new major release of a genome as described in Subaim 3.1.

Proposed Timeline

Year 2 - Release GRCh 39.0 candidates based on existing genome sequences.
Year 3 - Release GRCh 39.0 final.
Year 4 - Release GRCh 40.0 candidates based on HQRG genomes.
Year 5 - Release GRCh 40.0 final.

References

Ameur, A., Che, H., Martin, M., Bunikis, I., Dahlberg, J., Höijer, I., … Gyllensten, U. (2018). De Novo Assembly of Two Swedish Genomes Reveals Missing Segments from the Human GRCh38 Reference and Improves Variant Calling of Population-Scale Sequencing Data. *Genes*, *9*(10). https://doi.org/10.3390/genes9100486

Audano, P. A., Sulovari, A., Graves-Lindsay, T. A., Cantsilieris, S., Sorensen, M., Welch, A. E., … Eichler, E. E. (2019). Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell*, *176*(3), 663–675.e19.

Ball, M. P., Thakuria, J. V., Zaranek, A. W., Clegg, T., Rosenbaum, A. M., Wu, X., … Church, G. M. (2012). A public resource facilitating clinical use of genomes. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(30), 11920–11927.

Ballouz, S., Dobin, A., & Gillis, J. (2019). *Is it time to change the reference genome? bioRxiv*. https://doi.org/10.1101/533166

Chapman, B., Meynert, A., Church, D., Johnson, J., & Hofmann, O. (2016). Abstract 3636: Improved clinical variant calling and HLA genotyping with GRCh38. *Cancer Research*, *76*(14 Supplement), 3636–3636.

Church, D. M., Schneider, V. A., Steinberg, K. M., Schatz, M. C., Quinlan, A. R., Chin, C.-S., … Flicek, P. (2015). Extending reference assembly models. *Genome Biology*, *16*, 13.

Church, G. M. (2005). The personal genome project. *Molecular Systems Biology*, *1*, 2005.0030.

Chu, W. K., Edge, P., Lee, H. S., Bansal, V., Bafna, V., Huang, X., & Zhang, K. (2017). Ultraaccurate genome sequencing and haplotyping of single human cells. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(47), 12512–12517.

Garrison, E., Sirén, J., Novak, A. M., Hickey, G., Eizenga, J. M., Dawson, E. T., … Durbin, R. (2018). Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature Biotechnology*, *36*(9), 875–879.

Gibson, J., Morton, N. E., & Collins, A. (2006). Extended tracts of homozygosity in outbred human populations. *Human Molecular Genetics*, *15*(5), 789–795.

Guo, Y., Dai, Y., Yu, H., Zhao, S., Samuels, D. C., & Shyr, Y. (2017). Improvements and impacts of GRCh38

human reference on high throughput sequencing data analysis. *Genomics*, *109*(2), 83–90.

Huang, J., Howie, B., McCarthy, S., Memari, Y., Walter, K., Min, J. L., … Soranzo, N. (2015). Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nature Communications*, *6*, 8111.

Jain, M., Olsen, H. E., Turner, D. J., Stoddart, D., Bulazel, K. V., Paten, B., … Miga, K. H. (2018). Linear assembly of a human centromere on the Y chromosome. *Nature Biotechnology*, *36*(4), 321–323.

Johnson, L. K., Alexander, H., & Brown, C. T. (2018). Re-assembly, quality evaluation, and annotation of 678 microbial eukaryotic reference transcriptomes. *GigaScience*. https://doi.org/10.1093/gigascience/giy158

Marçais, G., Delcher, A. L., Phillippy, A. M., Coston, R., Salzberg, S. L., & Zimin, A. (2018). MUMmer4: A fast and versatile genome alignment system. *PLoS Computational Biology*, *14*(1), e1005944.

Martin, A. R., Gignoux, C. R., Walters, R. K., Wojcik, G. L., Neale, B. M., Gravel, S., … Kenny, E. E. (2017). Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *American Journal of Human Genetics*, *100*(4), 635–649.

Nagasaki, M., Yasuda, J., Katsuoka, F., Nariai, N., Kojima, K., Kawai, Y., … Yamamoto, M. (2015). Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. *Nature Communications*, *6*, 8018.

Nir, G., Farabella, I., Pérez Estrada, C., Ebeling, C. G., Beliveau, B. J., Sasaki, H. M., … Wu, C.-T. (2018). Walking along chromosomes with super-resolution imaging, contact maps, and integrative modeling. *PLoS Genetics*, *14*(12), e1007872.

Pan, B., Kusko, R., Xiao, W., Zheng, Y., Liu, Z., Xiao, C., … Hong, H. (2019). Similarities and differences between variants called with human reference genome HG19 or HG38. *BMC Bioinformatics*, *20*(Suppl 2), 101.

Paten, B., Novak, A. M., Eizenga, J. M., & Garrison, E. (2017). Genome graphs and the evolution of genome inference. *Genome Research*, *27*(5), 665–676.

Payne, A. C., Andregg, M., Kemmish, K., Hamalainen, M., Bowell, C., Bleloch, A., … Andregg, W. (2013). Molecular threading: mechanical extraction, stretching and placement of DNA molecules from a liquid-air interface. *PloS One*, *8*(7), e69058.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, *38*(8), 904–909.

Quintana-Murci, L., & Fellous, M. (2001). The Human Y Chromosome: The Biological Role of a “Functional Wasteland.” *Journal of Biomedicine & Biotechnology*, *1*(1), 18–24.

Rakocevic, G., Semenyuk, V., Lee, W.-P., Spencer, J., Browning, J., Johnson, I. J., … Kural, D. (2019). Fast and accurate genomic analyses using genome graphs. *Nature Genetics*, *51*(2), 354–362.

Samuels, D. C., Wang, J., Ye, F., He, J., Levinson, R. T., Sheng, Q., … Guo, Y. (2016). Heterozygosity Ratio, a Robust Global Genomic Measure of Autozygosity and Its Association with Height and Disease Risk. *Genetics*, *204*(3), 893–904.

Schneider, V. A., Graves-Lindsay, T., Howe, K., Bouk, N., Chen, H.-C., Kitts, P. A., … Church, D. M. (2017). Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Research*, *27*(5), 849–864.

Sherman, R. M., Forman, J., Antonescu, V., Puiu, D., Daya, M., Rafaels, N., … Salzberg, S. L. (2019). Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nature Genetics*, *51*(1), 30–35.

Steinberg, K. M., Schneider, V. A., Graves-Lindsay, T. A., Fulton, R. S., Agarwala, R., Huddleston, J., … Wilson, R. K. (2014). Single haplotype assembly of the human genome from a hydatidiform mole. *Genome Research*, *24*(12), 2066–2076.

van Leeuwen, E. M., Karssen, L. C., Deelen, J., Isaacs, A., Medina-Gomez, C., Mbarek, H., … van Duijn, C. M. (2015). Genome of The Netherlands population-specific imputations identify an ABCA6 variant associated with cholesterol levels. *Nature Communications*, *6*, 6065.

Wong, K. H. Y., Levy-Sakin, M., & Kwok, P.-Y. (2018). De novo human genome assemblies reveal spectrum of alternative haplotypes in diverse populations. *Nature Communications*, *9*(1), 3040.

Zimin, A. V., Marçais, G., Puiu, D., Roberts, M., Salzberg, S. L., & Yorke, J. A. (2013). The MaSuRCA genome assembler. *Bioinformatics* , *29*(21), 2669–2677.

Zook, J. M., Catoe, D., McDaniel, J., Vang, L., Spies, N., Sidow, A., … Salit, M. (2016). Extensive sequencing

of seven human genomes to characterize benchmark reference materials. *Scientific Data*, *3*, 160025.

Zook, J., McDaniel, J., Parikh, H., Heaton, H., Irvine, S. A., Trigg, L., … Genome in a Bottle Consortium. (2018). *Reproducible integration of multiple sequencing datasets to form high-confidence SNP, indel, and reference calls for five human genome reference materials*. *bioRxiv*. https://doi.org/10.1101/281006

Zook, J. M., Chapman, B., Wang, J., Mittelman, D., Hofmann, O., Hide, W., & Salit, M. (2014). Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nature Biotechnology*, *32*(3), 246–251.