Research Strategy - Overall Component.

Introduction

The human genome reference serves as a common reference point across human genetics and genomics for investigations into genetic disease and susceptibility, population genetics, and comparative genomics. It is foundational to modern biomedical research and clinical practice, and a public good for the biomedical community. However, the current reference does not adequately represent human genomic diversity: about 70% is from one African-American, 7% is from one East Asian, and the remaining 20+% are of mixed European ancestry (Green et al., 2010). The biomedical community must therefore collectively work toward a more inclusive, complete, and accessible reference.

Evolving the reference depends on both technology – in particular, new approaches to representing human diversity in the reference – and behavior: any new reference sequence must be practically adopted by an incredibly diverse group of actors, including biomedical researchers and data scientists, analysis commons, methods developers, and clinicians. This is at its heart a combined social and technical problem, where new technology is necessary, but insufficient without an attendant social effort to engage downstream consumers and drive adoption. This socio-technical effort should iteratively lower barriers to adoption of a new reference sequence while also providing incentives for this adoption. In particular, key stakeholders will need solid business cases to drive adoption.

We propose to implement a Pan-genome Reference Center (PaRC), to serve as the Human Genome Reference Center. PaRC's mission will be to maintain, improve, and provide the human genome reference, as part of a larger Human Genome Reference Program (HGRP); coordinate the HGRP's internal and external efforts to improve the human genome reference sequence; and engage with the larger biomedical community in a myriad of ways. In addition to driving an **innovative automation and release process** to tackle the technical challenges of representing human diversity in a functional and accessible manner, PaRC will embark on a **standardization effort** that will work to advance the human genome reference towards a globally accepted standard and a **community building effort** that will grow the community and network of practice around the HGRP's work and outputs. This socio-technical effort will drive a more complete and inclusive human reference genome with open and accessible tool-chains, reference materials, and processes so that everyone can benefit from advancements in genome research and precision medicine.

Significance

The Human Genome reference is a *de facto* standard: it is a shared resource used as a stable data structure to make comparisons against. It serves as a reference for comparative genomics, as an addressing scheme for annotation of sequence features and known variants, and as an inference engine for inferring haplotypes from sequencing data. Much of the utility of the human genome reference lies in our ability to properly map both existing and novel data to the reference sequence, enabling downstream analysis of variation in a meaningful context. Reads that cannot be placed on the reference may represent contamination or highly erroneous sequencing data (true negatives), or may represent real human sequences not present in, or significantly divergent from, their corresponding reference sequences (e.g. significant structural variation).

While read mapping can be affected by a number of technological confounding factors (e.g. differences in sample storage and processing, DNA extraction techniques, read length and error profiles of sequencing technology, differences in mapping software), it is also greatly limited by a lack of representation of the true variation across human populations. “Quality and inclusivity” refers to missing or erroneous sequence (Sherman et al. 2019; Audano et al. 2019), missing or incorrect variation, unlinked haplotypes, and unresolved gaps in the reference. The current reference (GRCh38) contains 90% or more of any given individual’s genome (in the sense that that percentage of accurate reads could be mapped to it), but is heavily biased towards European and African American content based on the genomes that went into its construction. Without a high-quality human reference genome inclusive of broad human genetic diversity, new sequences from individuals may go unused or be underused, and allelic variation will only be captured for those that are very similar to the reference. This is particularly problematic for population genetic studies where alternative alleles that differ greatly from the reference will likely be discarded and cause a systematic bias in reported allele frequencies.

With the advent of long-read single-molecule sequencing, high quality linkage sequencing, Hi-C, and optical mapping technologies, among others, we now have an unprecedented ability to generate high quality haplotypes from new samples (Audano et al., 2019; Sherman et al., 2019). Moving forward, it will be essential to identify targeted populations for sequencing and navigate cost-quality tradeoffs in order to ensure a more diverse and representative sequence base for the reference. However, the ability to sequence genomes to high quality does not mean that these genomes are readily accessible to researchers and clinicians. The largest challenge will be to develop a robust system for representing and incorporating the variation (particularly structural variation) in these individual genomes. The importance of this challenge is highlighted in the state of our current reference: although the GRCh38 reference includes ‘alts’, sequences for alternative loci, they are rarely used, both because researchers find them difficult to work with and because few tools support them.

Ideally, one would be able to represent all major human variation in a single reference, refer to locations and variants unambiguously, and easily move between reference versions without requiring users to re-calibrate all of their data for each new release of the genome. A universal pan-genome would allow researchers to recover major haplotypes, map reads to this single reference, and even apply large scale data science techniques such as machine learning and deep learning to the reference to discover correlations with phenotypes. In pursuit of this ideal, multiple competing pan-genome strategies are emerging. As a result, a critical part of our proposal is the production and comprehensive evaluation of *candidate* Human Genome references in the context of downstream toolchains, prediction quality, and clinical implications. The automated evaluation strategy we propose is especially important because the final pan-genome representation format(s) is not determined *a priori* and will instead emerge from the evaluated candidate systems during the course of the HGRP. **Methodological evaluation and validation approaches are critical in order to inform and drive convergence towards a canonical representation format.**

As it stands today, reference improvements represent a nearly insurmountable challenge to the larger biomedical community that relies on having a high quality, accurate, inclusive human reference genome. Individual community members are caught between the social and technical costs of migrating to a new reference that is potentially much more complete, but requires different workflows and coordinate systems and may require migrating years of their own data to new coordinates. Large analysis and visualization infrastructure resources and variant databases must choose how to allocate their support efforts between adding functionality on top of an old reference or migrating to a new reference. Databases of human variation must decide if and when to migrate, and whether to maintain their databases for multiple references. Bioinformatics methods developers must guess about the future applicability of their data structures and algorithms in the context of the reference genome format. These issues have led to stagnation. As bioinformatics technology evolves, it is inevitable that new techniques for representing and searching the reference may render some versions of the reference incompatible with previous software and mapping workflows. The comprehensive and automated reference evaluation strategy we propose will be essential to properly maintaining, evolving, and releasing a pan-genome reference in a manner that enables incorporation of novel sequence content, sequence refinement, and error correction without breaking analysis workflows and user confidence. **Development of a pan-genome reference must learn from current shortcomings and jointly develop a robust system to represent genomic variation alongside a complete bioinformatic toolchain and an engaged community of practice to facilitate adoption, utilization, and stability of the human genome reference.**

Innovation

In this grant, we propose a combination of social and technical solutions that, together, will ensure the HGRP is widely adopted as a foundational element of genomic medicine worldwide. Our core innovations in this proposal are to adopt and adapt techniques from the software development world – in particular, **advanced versioning** and **automation** -- to the maintenance, improvement, and provisioning of the human reference genome; to coordinate the HGRP as an **openly collaborative consortium;** to build a robust **community of practice** for the human reference genome, and to disseminate this practice using open science techniques to ensure maximum applicability; and to lead an international **standardization effort** that will engage with key stakeholders to iterate towards standardized methods of constructing, representing, evaluating, releasing, and using the human reference genome.

At the consortium coordination level, we will **adopt a transparent and open set of processes** in collaboration with HGRP leadership and NHGRI. Drawing on our considerable prior experience with open source, open science, and project/consortium coordination, we will adopt asynchronous and open work practices within the HGRC and HGRP, as well as across the larger community of stakeholders. We (UC Davis and Curii) pioneered an earlier version of this approach when we built the internal community platform and review/evaluation system for the NIH Data Commons Pilot Phase Consortium, and we propose to iterate upon it here. Similarly, leveraging our experience with the Genome in a Bottle initiative where we (Stanford, Curii and Harvard) adopted openly-consented reference material, we will facilitate the use of open consent in the HGRP to ensure that both commercial and non commercial entities could make products from HGRP sequenced individuals that were impossible under a 1KG consent or similar. Our vision for the HGRP is one in which Consortium materials and issues are open to the entire Consortium, and, where possible, public. In our experience, this open science approach accelerates progress by eliminating barriers to communication, information discovery, and collaboration.

One of our most important innovations is to treat the challenge of evolving the human reference as a combined social and technical challenge, and more specifically to **grow a community of practice** around using the human reference as it evolves and improves, and distribute this practice via a **network of practice**. We will do this by building a community, onboarding community members into current practice, engaging with community members through trainings, meetings, online forums, and social media, aggregating feedback from this engagement into informal and formal reports, resolving confusion with updated onboarding and training materials where possible, and working with community members to generate draft RFCs for submission to the larger HGRP and standardization efforts. By harnessing community feedback in this bottom up way, we will be able to organically grow and adapt the set of materials and products produced by the HGRP to meet the needs of this self-identified user community; more, we can focus the energy and enthusiasm of this community towards producing materials, products, and software. This technique is a core aspect of successful open online communities.

This bottom-up community effort will be combined with focused engagement between the HGRP and external stakeholders with the goal of producing formal standards for constituent genome quality, build processes, sequencing quality, and toolchain consistency. Our overall coordination approach will rely on automating routine tasks and infrastructure needs as much as possible (particularly in Aims 1-3). We will use GitHub for consortium coordination and material change tracking and maintenance, permitting a flexible set of graded options for consortium engagement. Another key innovative aspect of our approach is the open source Centillion search engine prototyped during the NIH Data Commons, which enables the **indexing, search, annotation, and dashboard viewing of access-restricted content**. Centillion will provide the HGRP and larger standardization efforts with search capabilities across internal static content (e.g. PDFs), collaboratively evolving documentation (e.g. Google Docs), issue trackers (e.g. GitHub as well as custom issue trackers), collaboration platforms (e.g. Slack and Discourse), and calendaring.

Much as open standards, open data and free and open source software implementations helped drive the innovation that powered the world wide web, we will bring these to the HGRP. As a consortium we will create a future in which everyone can benefit from precision medicine driven by a more complete and inclusive human reference genome and its adjacent reference materials, tool-chains and processes.

Preliminary Results

PaRC will drive social and technical convergence to a new Human Genome reference toolchain via open evaluation and iterative consensus building. 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, standards and leading workshops. The Genome-In-A-Bottle (GIAB) cell lines are some of the world's most widely sequenced and best characterised cell lines currently available. 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) and test and validate the upgrade of bioinformatics tools to use GRChr38 (Chapman, Meynert, Church, Johnson, & Hofmann, 2016). To establish best practices for using GIAB genomes for benchmarking, GIAB works with the Global Alliance for Genomics and Health Benchmarking Team (Krusche et al., 2019). **We will use the GIAB expertise in consensus building, providing validation data and working external standards groups to guide our Human Genome reference candidate quality evaluations in Component 1 and our stakeholder engagement in Component 3.**

Our community outreach (Component 2) and logistical coordination strategies (Component 3, Aim 6) are based on our extensive experience with open source projects, open online communities such as the Carpentries, a decade of organizing training workshops at UC Davis, and most recently **on the coordinating effort UC Davis co-lead for the NIH Data Commons Pilot Phase Consortium** (November 2017-November 2018). Success in these projects relies on asynchronous engagement with diverse communities, extensive technical infrastructure for open collaborative work, and iterative refinement of social strategies to identify and meet a broad range of user needs. With the NIH Data Commons, Dr. Brown’s team created and adapted the collaboration infrastructure used to coordinate the work of over 500 participants to meet the needs of many diverse stakeholders in the fast-paced initial phase. **UC Davis reviewed and refined over 80 major deliverables, ran almost a hundred teleconferences, and coordinated and produced multiple large-team collaborative outputs using the sociotechnical strategies outlined in Components 2 and 3**. Public outputs of the NIH Data Commons are available at http://public.nihdatacommons.us.

Approach

Here, we describe our seven Specific Aims in brief, together with their cross-cutting interactions.

We start by giving an overview of our highly integrated technical and social strategy for maintaining the current human reference representation with small patches and corrections, while supporting a transition to a pan-genome representation over the duration of the 5 years of the proposed project. Portions of this strategy are distributed among the Specific Aims that follow, with the key technical activities in Component 1, and the key social and coordinating activities in Components 2 and 3, respectively.

Specific Aims 1-3 address the core technical activities of PaRC in maintaining, improving, and providing the Human Reference Genome, and are described in detail in Component 1. These aims include our technical plan and timeline for transitioning to a pan-genome reference representation.

Specific Aims 4 and 5 provide support for and feedback mechanisms from the research and clinical communities using the Human Reference Genome, and are described in detail in Component 2. These aims detail the social strategies we will employ to support transition to a pan-genome reference representation, including training and incentivization.

Specific Aims 5 and 6 describe PaRC's role in coordination of the HGRP and our planned engagement with stakeholders in standards-building activities, and are detailed in Component 3. These aims include our strategy to coordinate with the HQRG and GRR activities in the HGRP, our rationale for prioritization of specific external partners, and our plan to coordinate more broadly with other US and international efforts.

After the Specific Aims, we describe project management and reporting for the HGRC, summarize our "Value Add" products across all of the Aims, and propose a timeline for HGRC and HGRP activities.

Proposed Strategy and Tactics for Transitioning the Human Reference Genome to a Pan-Genome

Many communities and stakeholders depend on the human reference genome as both a coordinate reference and a basis for variant analysis. Any attempt to transition from GRCh38 to a more accurate and inclusive pan-genome in later releases must confront the question of how to transition practice most effectively; this involves not just releasing an updated reference, but providing a functional toolchain to make use of the new reference, and ensuring that there several infrastructure providers offering services based on it. The software toolchain here is key, but in our experience it takes a minimum of 3-5 years to produce robust new software functionality, iterate it to meet user experience needs at an acceptable level, and produce guides, documentation and tutorials that support the shift in practice. In a research ecosystem of this size, it will take more time because the early stage utility of a new reference will be minimal until databases, analysis commons, and partner projects can update their codebase to work on the new system.

To address this, in Specific Aims 1-7, we propose a process that offers the maximum flexibility to the HGRP and larger community, lays out clear communication around content and iterative improvements, integrating an evaluation of the entire relevant toolchain into each phase, and providing an opportunity to evaluate new potential reference representations in the context of their overall toolchain.

Briefly, we will work with the HGRP and partners to

(1) Provide a new and better representation, with improved content (Aims 1-3);
(2) Release early stage technology (representation and associated toolchains) that can be used by early adopters (Aim 2 and 4);
(3) Iterate to make the technology more robust in response to feedback (Aims 2 and 5);
(4) Produce benchmarks and metrics that demonstrate the improved quality of outcomes with the new reference (Aim 2.2); and
(5) Engage with infrastructure providers and databases to include the new reference in their offerings (Aims 3, 5 and 7).

The early adopters community will have access to alpha and beta builds of the genome, as well as release candidates; we do not propose to restrict this membership in any way.

While this process will seem slow at first, we believe steady iterative process towards a more inclusive reference genome and format will let us reach a tipping point within the next 3-5 years, at which time we expect the incentives to begin transitioning a broad set of tools and databases to the new reference will be clear, and there will be a sufficiently robust path for those who wish to do so.

Specific Aims

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

Here we focus on the construction of a library of genomes that is of both high quality and, collectively, encompasses known human diversity. Some of these genomes will be used for constructing a new Human Genome reference, while others will be used as a validation set. A key part of this Aim is the coordination of a HGRP Participant Program, that will build connections with individuals who have contributed to this library.

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

Both the construction and the evaluation of Human Genome reference candidate genomes should be automated to a significant extent, in order to ensure provenance of workflows and establish trust in validation. An important product of this Aim will be validation resources that can be used by methods developers building their own toolchains.

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

Once we have converged on a potential new reference, we will release it via a staged process that engages early adopters and infrastructure stakeholders to ensure that the new release meets the needs of the community and is supported by existing infrastructure.

Specific Aim 4: Provide an open online portal for the Human Genome community.

The HGRC is responsible for hosting and dissemination of a wide variety of materials; here we describe our approach to hosting and dissemination of materials. We will also maintain a set of onboarding guides for communities that wish to make use of these resources.

Specific Aim 5: Build a community of practice around the HGRP.

Building a community of practice in using new Human Genome reference releases is key to driving adoption by users. Here we describe a strategy for building a community of practice that includes user engagement, the building of a tool library and workflow gallery, and a community focused training program.

Specific Aim 6: Coordinate logistics for the NHGRI Human Genome Reference Program.

The HGRC will serve as the overall logistics coordinating center for the HGRP. We describe our plan to host and maintain technical infrastructure, our proposed annual meeting coordination, and our internal Web site and document library.

Specific Aim 7: Convene global stakeholders in a consensus-building process to move towards standardization of the Human Genome reference.

We will create and host working groups, implement a “Request for Comment” process, and begin working to identify the many stakeholders who are interested in a standardization effort.

Project management

UC Davis (Brown, PI) will provide the central point of contact with NHGRI program officers, manage reporting, and create and maintain the internal and external coordination infrastructure. Process automation, release management, and related efforts will be led by UC Davis. Overall project management will be handled by this team, in collaboration with the Project Manager in Curii who will maintain and track internal deadlines.

Curii (Zaranek, PI) will work with the HQRG and Open Humans to coordinate choice of samples, manage consent, and otherwise participate in building and maintaining the library of genomic materials. They will also support compute, workflows, and cloud orchestration through their CWL-enabled Arvados open source platform.

Stanford (Salit, PI) will develop metrics and evaluate reference quality for the build and evaluation process, and will drive the standardization process and engagement with external stakeholders.

Harvard (Church) will provide sequencing capacity for directed sequencing, as well as managing incremental genome improvements.

Our overall budget reflects our expected emphases and division of labor. Approximately 45% of PaRC’s budget is in Component 1, with the remainder split evenly between Components 2 and 3. UC Davis and Curii are split evenly across Components 1, 2, and 3, Harvard is focused in Component 1, and Stanford is split across Components 1 and 3.

Day to day operations will be coordinated by the lead project manager at Davis, in consultation with Curii. The leads for each team will form an executive council that will work closely with NHGRI representatives, external advisors, and so on.

Collectively, we have a history of working together and within larger open communities. More specifically, Curii and UC Davis collaborated on the coordination of the NIH Data Commons. Harvard, Curii, and Stanford collaborate closely through the Personal Genome Project. All of us are committed to open science and open collaborative practices and have substantial experience in distance collaborations. We will coordinate regular in person meetings in Boston (Harvard, Curii) and at Genome in a Bottle meetings.

Value add products, and how we will release them to the community

We will significantly add to the already openly consented materials available through the Personal Genome Project and Genome in a Bottle, and provide a coordination platform for additional collaboration and participation requests as part of the HGRP Participant Project (Aim 1).

The automated Human Genome reference construction and evaluation scripts and processes produced in Aim 3 will be broadly valuable to the community. They will be made available on GitHub and Zenodo.

Our metadata browser and Web site built as part of Aim 4 will be openly available, with full content on GitHub.

Our additions to bio.tools (Aim 5) will be made available through the existing bio.tools repository. The workflow gallery produced in Aim 5 will be publicly hosted via a GitHub-based site.

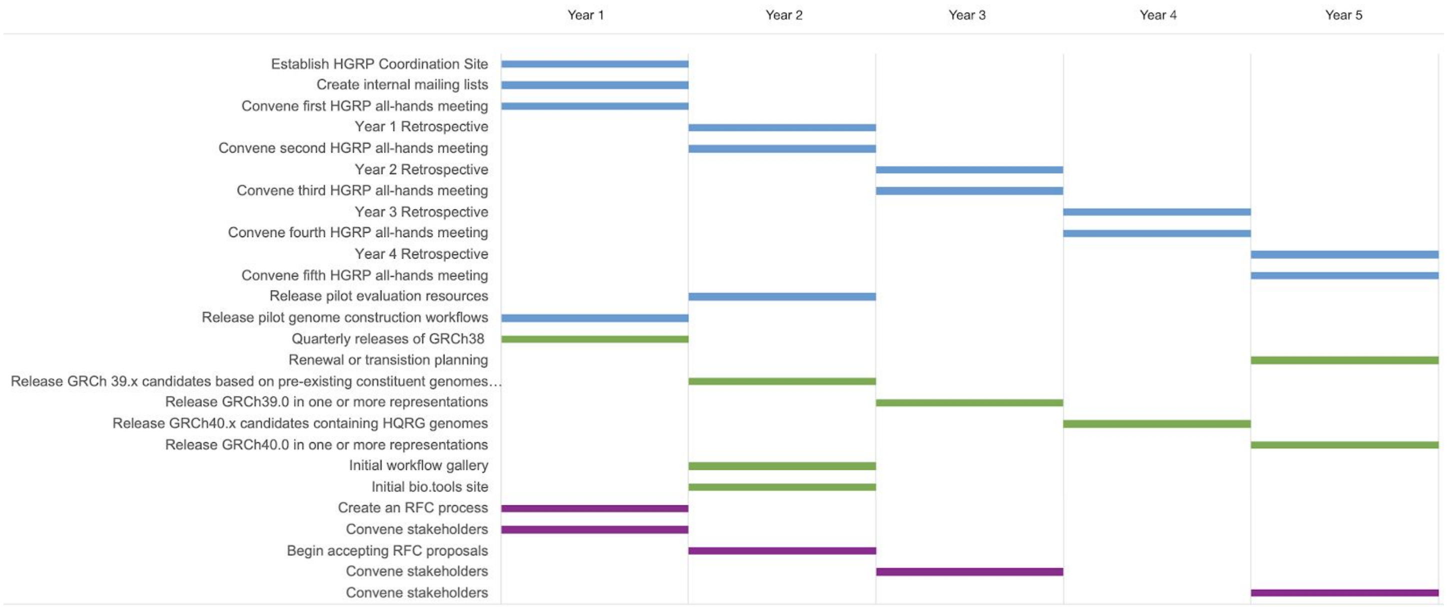
The training materials generated in Component 2 will be hosted on GitHub, as is our current practice.

The Request for Comments process will generate valuable documents that will be useful for longer than the duration of this project (Aim 7.2). These will be made openly available on osf.io.

Playbooks developed in Aim 6 will be made available through the access-restricted HGRP site.

All infrastructure software developed as part of PaRC will be made available under an Apache 2 license, and all materials for which we hold copyright will be available under CC-BY. The Centillion search engine is already open source under BSD 3.

Proposed Timeline



References

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.

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.

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., … Pääbo, S. (2010). A draft sequence of the Neandertal genome. *Science*, *328*(5979), 710–722.

Krusche, P., Trigg, L., Boutros, P. C., Mason, C. E., De La Vega, F. M., Moore, B. L., … Global Alliance for Genomics and Health Benchmarking Team. (2019). Best practices for benchmarking germline small-variant calls in human genomes. *Nature Biotechnology*. https://doi.org/10.1038/s41587-019-0054-x

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.

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.