Research Strategy - Component 3, Aims 6-7.

Introduction

As the coordinating center for the Human Genome Reference Program, the Pan-genome Reference Center (PaRC) will provide logistical support for coordinating these multiple distinct components of the HGRP, including supporting ongoing activities, hosting annual meetings, and hosting online materials. We also propose to convene stakeholders external to the HGRP into a consensus- and standards-building effort. **A crucial part of this effort is to first create an agreed upon governance and consensus building processes**. We will coordinate working groups, develop a "Request for Comments" (RFC) system, and grow a process for reaching consensus around emerging standards.

Engagement with Other HGRP Components: This component (Component 3 of the overall PaRC HGRC application) focuses on strategic engagement with larger stakeholders. Component 2 focuses on community outreach and "bottom up" engagement activities.

Significance

The 2017 paper on the GRCh38 Reference Assembly notes that the "human reference genome assembly plays a central role in nearly all aspects of today's basic and clinical research" (Schneider et al., 2017). The reference genome also plays an important role *outside* of research. The genomics community has matured beyond specialized experts doing basic and clinical research, to a broad set of stakeholders working across basic and applied science, including discovery scientists; technology developers; clinical practitioners; diagnostic device developers; drug developers; health system developers, including regulators and payers; forensic scientists, and genome engineering technology developers. The HGRP needs to ensure that the pan-genome reference is accessible, usable, and beneficial to the larger genomics community.

The Human Genome reference is already a *de facto* standard; it is a shared resource used as a stable data structure to make comparisons against. It has all the hallmarks of a standard: the reference genome is used to compare genome sequences amongst different samples or different conditions within experiments, across different experiments, in different labs, at different times, and between different sequencing technologies. **Comparing observed sequences against the reference *is* the platform for interoperability of human genomics.** Enhancing the *de jure* standardization of this important resource is a critical need: there is a deep history and community of standards as shared global resources for trade, communication, engagement, collaboration, competition, regulation, and innovation. The meta-technology of standards includes international bodies to recognize and coordinate standards development, and formal procedures to assure compatibility, stability, and recognition of standards. The established practices for consensus standards-making are a key element in assuring regulatory oversight and adjudication of conflicts when they arise.

Innovation

Our vision for the HGRP is one in which Consortium materials and issues are open to the entire Consortium, and, where possible, the public. 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 can make products from HGRP sequenced individuals that would be impossible under a 1KHG consent or similar**. In our experience, this open science approach accelerates progress by eliminating barriers to communication, information discovery, and collaboration.

Preliminary Results

Our [Curii, Harvard] previous work with the Archon Genomics XPRIZE competition helped create rules for the validation protocol for "sequencing 100 human genomes within 30 days to an accuracy of 1 error per 1,000,000 bases, with 98% completeness, identification of insertions, deletions and rearrangements, and a complete haplotype, at an audited total cost of $1,000 per genome" (Kedes & Campany, 2011). This work inspired the creation of the Arvados Project to support the informatics infrastructure required to win the competition and influenced future standards for GIAB and CAP. The Arvados Project collaborates closely with several standards efforts, including CWL and the Global Alliance for Genomics and Health (GA4GH).

PaRC will drive social convergence to a new Human Genome reference toolchain via an 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; Schneider et al., 2017)) 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 to guide stakeholder engagement in Aim 7.**

For 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. Of particular note, Dr. Brown led the construction of the NIH Data Commons Use Case Library (http://nih-data-commons.us/use-case-library/) and co-led the Data Commons RFC process (http://public.nihdatacommons.us/RFCtrack/).

Federated search across access-restricted GitHub issues, GitHub repositories, Google drives, and instant messaging platforms emerged as a critical need during the NIH Data Commons Pilot Phase Consortium. In response to this need, Dr. Brown's lab developed and hosted an open source search engine, Centillion (https://github.com/dcppc/centillion). Centillion provides full-text indexing and search, a dashboard for monitoring update times for all indexed content, and can be hosted openly or placed behind a single sign-on

firewall. Centillion is implemented in Python under the MIT license. It provides both interactive Web pages, and a scriptable REST API.

Approach

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

Subaim 6.1: Facilitate and coordinate ongoing HGRP activities.

From our previous experience with facilitating and coordinating large communities, including many open source projects and several multi-PI/multi-team collaborations like the NIH Data Commons, we expect there to be a number of common needs within the HGRP. In addition to teleconferencing and managing call notes and action items, there will be a need to track and route issues within the HGRP; we will maintain a global issue tracker and individual points of contact for working groups within the HGRP. The HGRP will also need a mailing list and calendaring support, and we will provide a single platform for both (e.g. groups.io). The HGRP will also have many document sharing and collaboration needs, and we will manage access control in Google Docs through Google Groups; these can be done individually and/or on the basis of working groups (see Aim 7.1).

Issue tracking of logistical issues, and routing of external issues, will be done openly within the HGRP using an access-restricted GitHub issue tracker -- GitHub is simple to use, supports highly configurable access and notification settings, and is the *de facto* standard in the data science world. This will also permit flexible interlinking with issues in the community issue tracker maintained in Aim 5.1 (note that GitHub can provide links that are only visible to those with access to the issue). And, finally, we will provide a search system, Centillion, that can search across multiple GitHub issues boards (see Preliminary Results, and Aim 7.3, below).

Teleconferences led by PaRC, e.g. within the HGRP and with external working groups (Subaim 7.1), will adopt the online note taking and facilitation playbook developed by the Carpentries for discussions (https://docs.carpentries.org/topic_folders/instructor_development/instructor_discussion_sessions.html#host-expectations). This involves developing the agenda, configuring and using an online note taking venue, maintaining focus, prioritizing questions, tracking issues, and otherwise facilitating the teleconferences. This facilitation service will also be made available to all working groups as an option. Online notes will be included in the internal Web site and document library and be discoverable via the Centillion search engine (Aim 6.3).

We will create, maintain, and update an HGRP onboarding guide. This will be a critical need given the expected size and complexity of the HGRP consortium and associated activities, as well as the variety of communications modalities and commonly used platforms that will inevitably be used within the consortium. The onboarding guide will list the HGRP teams with administrative and technical contact points, provide an entry point to meeting calendars and teleconference notes, and detail the working groups and their contact points. We will also create and maintain playbooks for HGRC and HGRP operations. These will contain details of ongoing operations, including job duties, daily/weekly/monthly/annual tasks, and a list of infrastructure platforms and contact points for access keys. These playbooks will be critical for any transition efforts after the duration of the initial funding period.

As part of our overall coordination work, we will maintain a code of conduct across the Consortium. We have found this to be critical for maintaining open communication and a culture of participation within consortia. We

will work with the other HGRP PIs and NHGRI program coordinators to define dispute resolution procedures and an escalation policy.

Subaim 6.2: Coordinate and host annual activities.

As part of our coordination and facilitation, we will work closely with the NIH to host an annual meeting of the HGRP, and engage with the programmatic leadership at NIH as well as members of the HGRP around strategic direction. We will help develop a schedule, execute the meeting, facilitate activities during the meeting, organize catering, and provide travel support for external program advisors. Following the meeting we will draft a short report highlighting major discussion points and listing action items.

At this annual meeting we will also present the results of an annual HGRP retrospective. Retrospectives are valuable opportunities to discover and discuss opportunities and challenges within the project. The annual retrospective will cover a number of topics, including progress towards the overall aims of the HGRP, the technical goals of the HGRP, project management activities, interaction with programmatic management, and inter-group communication. We will execute this retrospective in the month leading up to the annual meeting through an iterative process developed during the NIH Data Commons: first, we will gather information from all HGRP participants anonymously, via a form submission; second, we will collate and organize the submitted information into categories; third, we will coordinate multiple teleconferenced discussions of these points; and fourth, we will summarize the discussions in brief document form.

Subaim 6.3: Maintain internal Web site and document library.

We will create and maintain a central access-controlled Web site for the HGRP. This site will include the onboarding guide and playbooks developed in Subaim 6.1, as well as any other documents produced during the course of HGRC and HGRP operations. The site will also provide a unified portal to the issue tracker, calendaring system, mailing list archives, and document collaboration system created and supported by Subaim 6.1.

The internal consortium web interface will be access-controlled via a federated single-sign on (SSO) system that supports two-factor authentication. As with the NIH Data Commons, we will use GitHub for this because it provides a free and secure SSO system with a flexible permissions system that can be managed programmatically.

The internal web site will contain links to publications and presentations developed during the program; by default we will ask that these be made available under a CC-BY license. We will make these FAIR by providing infrastructure for hosting and citation via the Open Science Framework (https://osf.io). While we will support access-restricted hosting, we expect the majority of publications and presentations to be public.

In support of the overall Aim 6, we will provide and enhance an internal search engine, Centillion, that indexes and provides search capabilities for all of the access restricted material available to the HGRP through the internal website, including issue tracking and mailing list archives. In addition to search of access-restricted materials, Centillion will support metadata tagging of documents and monitoring of these documents via a personalized dashboard. Here, Centillion fulfills a critical consortium need for keyword and metadata search on access-restricted materials across multiple repositories, as well as providing a personalized dashboard that sorts by modification date of relevant materials.

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

The HGRP will be at the center of a large community of users and stakeholders that rely on the Human Genome reference as the basis of their work. This community will include researchers, clinicians, methods developers, infrastructure efforts, commercial ventures, and certification bodies. We see a critical need to provide avenues for these stakeholders to engage formally with the HGRP and Human Genome reference efforts. In contrast to the broad and relatively unstructured "bottom up" community outreach proposed in Aims 4 and 5 (Component 2), these avenues would be more formal and structured "top down" efforts to engage with the community.

Subaim 7.1: Provide logistical coordination for working groups.

We will nucleate a set of working groups through the HGRP, provide logistical support for these working groups, and enable others to establish working groups. Working groups would not need to be "owned" by the HGRP or PaRC, and we will engage with GA4GH and other collaborative international efforts to reduce overlap in working group focus. This will be our primary collaboration method with representatives from regulatory agencies (FDA, CAP, CLIA), clinical stakeholders (ASHG, CDC, clinical labs), and genomic infrastructure resources (NCBI, ANViL, Stage, NCI Cloud resources, PrecisionFDA), as well as key consortia (HGSV, ENCODE, ClinGen), cross-domain working groups, and standards organizations (GA4GH, GIAB, GRC).

For example, we see a clear need for a working group on Human Genome reference candidate evaluation, with specific reference to our proposed validation efforts in Subaim 2.2 (Component 1). This working group would include sequencing experts, toolchain authors as well as clinical specialists, and would provide a forum for iterating on evaluation criteria. We will also form a working group with the GA4GH to create a driver project around various aspects of creating, maintaining and use of the pan-genome. This would ensure alignment between GA4GH and HGRP, and ensure that standards evolve to meet the needs of pan-genome references. One of the goals would be to provide a Beacon-like service as a computational extension of our metadata and provenance exploration site; Beacon is a federated, web-accessible service that can be queried for information about a specific genomic variant (https://ga4gh-discovery.github.io/categories/beacon.html). Other potential topics include: defining, sharing, and executing our workflows, updating data format standards to handle new reference representations, conforming to or updating variant representations and annotations based on the new reference, and standardizing phenotypes for open consented genomes in the reference library.

For clinical users, the challenge of updating CAP/CLIA approved pipelines is a major reason for staying on GRCh37. A working group involving representatives from the FDA, CAP, CLIA, AMP, ClinVar, and ClinGen, bioinformaticians, and clinical geneticists, could meet to discuss challenges faced by teams updating clinical pipelines. They could establish criteria for improvements that would promote the adoption of new references, and best practices for updating pipelines and clinical databases, as well as suggest improvements to existing toolchains to help with pipeline maintenance.

Working groups would be open to participation by anyone. The output of working groups could include use cases, reference implementations, and draft documents to be entered into our proposed "Request for Comments" system (Subaim 7.2). All materials developed in the working groups would be licensed CC-BY.

Subaim 7.2: Develop a "Request for Comments" system and process.

The Request for Comments (RFC) system is a key standards development, dissemination, and peer review effort underlying the Internet. RFCs are essentially memoranda describing methods, behaviors, research, or innovations, and are submitted to convey concepts and information to an engineering community. Internet Drafts (early stage RFCs) and RFCs differ from publications in that they are often engineering focused, are written openly and iteratively, and developed in response to ongoing peer review. Internet Drafts, RFCs, and other efforts such as Python Enhancement Proposals serve as key artifacts in consensus building, decision-making processes, and standardization efforts. RFCs also document this process and supply a transparent archive of the discussions and decisions that lead to an implementation.

Over the next five years, we propose to stand up an RFC process with multiple stages, where documents would move from Drafts to Proposals as they mature, and in some cases move forward to consideration for Standards (Aim 7.3). Drafts would primarily emerge from working groups, and be opened for comment to the broader community. Drafts would be iterated upon in response to community questions during an open-ended process, until the working group feels that they have adequately addressed all comments, at which time they would be submitted as full Proposals or left as final Drafts. The overall process would be managed by Dr. Salit's group in coordination with the other PIs.

Subaim 7.3: Create governance and consensus building processes for a standards effort.

While genomics touches nearly all biomedical research, expanding to broader commercial and regulated applications needs to take the well-established path of using common, recognized, formal standards. Common standards are developed and adopted across many technical and scientific domains using recognized international standards bodies.

We will bring the best practices of standards-making to the HGRP to be sure that the reference pan-genome has recognition and stakeholder support as a global standard for human genomics. This will include partnership and collaboration and coordination with the Genome Reference Consortium, the GA4GH, the Genome in a Bottle Consortium, industry groups like the Medical Device Innovation Consortium, regulators like the FDA, clinical organizations like AMP and ACMG, and research organizations like ASHG. PaRC will support all elements of the HGRP to function as a standards development platform. Of note, there is a desire amongst the International Bureau of Weights and Measures (BIPM) stakeholders to adopt programs assuring global comparability of biological measurements; at the recent General Conference on Weights and Measures, a historic event where the conference voted to adopt a new SI that establishes a new definition of the kilogram, there was a featured discussion of "biological measurements as the frontier of metrology." PaRC will bring the human pan-genome reference to the BIPM and its members to explore adoption of the pan-genome as a global standard. PaRC's JIMB team will work with its principal funding agency, the US National Institute of Standardsand Technology to reach the BIPM and its stakeholders, partnering with BIPM to host an interlab study to evaluate technical performance across labs and economies amongst its membership.

PaRC will engage with a host of global, regional, and national standards bodies, including the ISO TC 276 Biotechnology Technical Committee, and the International Laboratory Accreditation Council (ILAC). These standards developers and accreditors have influence on market and regulatory acceptance and recognition of the validity of measurement results. Results tied to recognized standards are understood to be reproducible and valid, which leads to acceptance and confident use. Our human pan-genome reference is an on-ramp to mature, global expansion of human genomics.

References

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.

Kedes, L., & Campany, G. (2011). The new date, new format, new goals and new sponsor of the Archon Genomics X PRIZE competition. *Nature Genetics*, *43*(11), 1055–1058.

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.

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.