Research Strategy - Component 2, Aims 4-5.

Introduction

The Human Genome Reference Program (HGRP) will produce and disseminate a wide array of online resources for many different sub-communities of users with many different research interests. The known users of the HGRP site include biomedical researchers, clinicians, data scientists, methods developers, analysis commons, and clinical service companies, and we expect an even broader array of new users as the HGRP matures.

As the coordinating center for the HGRP, the Pan-genome Reference Center (PaRC) consortium will coordinate active dissemination of materials, provide information via Web sites and online portals, and create onboarding and training materials for each identified subcommunity of users. We will also provide a customized index of informatics tools based on the bio.tools site, and extend and expand the functionality of bio.tools to meet the needs of the biomedical research community. But this is not enough! **A core challenge for the HGRP is to identify and meet community needs iteratively and organically, throughout the life of the project.** To that end, our consortium will engage in **community building using an open approach** centered around training and online user engagement.

Engagement with other HGRC components: This component (Component 2 of the overall PaRC HGRC application) focuses on broad reach and organic community engagement with the widest possible community of potential users. These are “bottom up” engagement activities. Component 3 (and Aim 7 in particular) is focused on close engagement with larger stakeholders and will be “top down” and strategic in nature. We expect Component 3 to be more closely coordinated with other members of the HGRP, the NHGRI programmatic coordinators, and the NIH overall; this component, Component 2, is focused on information dissemination (Aim 4) and community engagement as the mechanism for needs gathering (Aim 5). In other words, Aim 5 is where we **expect the unexpected** to emerge and be incorporated into the HGRP.

Significance

The HGRP will serve as one nexus of information about the human reference. The coordinating center is responsible for coordinating material dissemination from within the HGRP to the larger community. This involves meeting the needs of a wide range of users, across a wide range of expertise. We divide core activities for this component in two: first, **information dissemination and curation***,* and second, **needs gathering from the community.**

Information dissemination is challenging from two perspectives. One is organization and curation – in particular, the need to maintain and update a comprehensive catalog of materials while also providing opinionated entry points into the materials for different sub-communities of users. The second is maintenance, in which existing materials must be updated, refreshed, and corrected as information changes and new needs emerge.

Needs gathering is key to long-term success of the HGRP, which must be responsive to the needs of the biomedical user community in terms of human genome reference updates, information dissemination, technology development, and training. In our considerable experience with user engagement, community facilitation, training, and outreach, **no plan survives first contact with users.** No matter how well structured or thoroughly architected the overall set of HGRP and HGRC activities are, users will immediately identify

short- and long-term gaps in materials and tools, suggest reprioritization of activities, and otherwise confound the best laid plans. **Community engagement provides a powerful way to organically incorporate user needs into our strategic planning.** However**,** translating diverse user needs into the common reference frame needed for strategic planning is challenging. Moreover, building trust with users takes time, through slowly increasing engagement and response. On the consortium side, scalability and responsiveness of user engagement are critical but hard. Despite these challenges, the rewards for engaging with users, are great: in particular, **the strategy of growing community engagement and iteratively meeting user needs is an important component of the solution to the sociotechnical challenge of driving the adoption of the new reference genome releases.**

Innovation

The challenge of driving adoption of new reference releases is fundamentally a combined social and technical problem that we propose to solve by **explicitly building a community of practice** (Lave, 1991). This framing is in itself innovative and links to a large and diverse literature in shifting practice that is well known in other communities (e.g. the nursing), but has not seen much application in the research technology space (Grudin, 1994). **A community of practice provides rich engagement with many diverse user subcommunities and is a hallmark of many large, successful, open online projects such as the Carpentries, Linux, Python, and R.** But building effective communities of practice involves reciprocation of effort, and hence presents a scalability challenge. We will meet this challenge in two innovative ways: first, we will employ the full-time efforts of a community facilitator who is trained in community engagement through the AAAS Community Engagement Fellows Program, and second, we will leverage automation and the “pull request” model of user-contributed material updates popularized by GitHub. **This will provide low-cost mechanisms that are proven to scale with community size**.

Translating user needs reported from many diverse user subcommunities into a common framework for reporting to the HGRP is also a significant challenge that must be met to achieve success in the long term. Our core method for this will rely on **using training workshops to gather information about user needs**, and in particular in gathering feedback from the instructors themselves after each workshop. This innovative technique helps accurately identify user needs by engaging individuals who can bridge the gap between subcommunities and the HGRP.

We also provide three technical points of innovation. First, PaRC will build and host a metadata browser that will permit users to investigate the composition and provenance of the Human Genome reference. Second, PaRC will provide a personal dashboard (based on our open-source Centillion software) for users to tag HGRP materials and track updates to them. And third, we will work with an existing software catalog, bio.tools (https://bio.tools/), to adopt and extend bio.tools to meet the needs of the HGRP while increasing sustainability through process engineering.

Preliminary Results

We have substantial prior expertise in building and maintaining Web sites for communities and consortia. UC Davis has more than a decade of experience in running highly engaged user training events, between the Carpentries, ANGUS, and one-off workshops, with over 2000 alumni of UC Davis hosted workshops. And we have already developed substantial dashboard functionality as part of the open source Centillion project (https://github.com/dcppc/centillion), described in more detail in Component 3. Because of our decade-long involvement in training, as well as our (Curii and UC Davis) work in coordinating the NIH Data Commons, the

majority of the technical infrastructure described in this proposal already exists, and can easily be redeployed here. (See http://public.nihdatacommons.us for public content related to the NIH Data Commons, and the Use Case Library (http://nih-data-commons.us/use-case-library/) for an example of gathering and integrating diverse user needs.)

Approach

Our approach to community outreach will center around two Aims. First, we will disseminate information to the Human Genome community via an open, online portal (Aim 4). This will include creating and curating an online library of online materials (Aim 4.1), disseminating information through a variety of channels (Aim 4.2), and creating onboarding materials for different subcommunities (Aim 4.3). Second, we will build a community and network of practice around the HGRP (Aim 5). This will include creating a community helpdesk and issue tracker (Aim 5.1), providing and maintaining a library of tools and a gallery of workflows (Aim 5.2), and building and maintaining a collection of training materials (Aim 5.3).

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

The HGRP will need to create, maintain, and support a large collection of formal and informal reference materials, and actively disseminate information via announcements and trainings. We believe that the two major challenges for this Aim will be in ongoing curation and maintenance (addressed in Subaims 4.1 and 4.2), and in maintaining and iterating the onboarding guides for different subcommunities (Subaim 4.3).

Subaim 4.1: Create, maintain and support online materials.

The HGRC will need a website for broadly disseminating information about HGRP activities that is openly accessible. To ensure that information is easy to locate, we will build a single website that will have both public and access-restricted areas, so that all HGRC resources can be accessed from one web portal. Our landing page will contain links to several main areas, including News and Announcements, Frequently Asked Questions (FAQs), link-outs to analysis commons and databases, the Material Library, the Training Library, the Software Tools library, and the Issue Tracker, as well as onboarding guides for various user subcommunities. Access- restricted areas are discussed further in Component 3, Aim 6.

The central challenge of hosting content on a Web site lies in curating and maintaining that content. We will base our central Web site on underlying Markdown content hosted on GitHub, rendered with the MkDocs tool – this is an approach that we used for multiple NIH Data Commons Web sites. This has many advantages over traditional Content Management Systems, including robust user permission system with change management and content tracking via pull requests. This also enables merging of third-party contributions with lightweight moderator review.

We will also provide a personal search, tagging, and monitoring system based on our open source Centillion software, developed by us during the NIH Data Commons Pilot Phase Consortium. Centillion enables users to track changes to documents of interest and provides a flexible tagging capability to build personal document collections. This tagging capability can also be combined with internet search engine optimization to make these documents more readily findable by users. (See Aim 6.3 in Component 2 for more details on Centillion, which can also be used to search *private* documentation collections.)

A core part of our site will be hosting information about the Human Reference genome. We will submit basic information to the public NCBI site, as per current GRC practice, and integrate new assemblies with genome browsers and viewers (UCSC Genome Browser, ENSEMBL, 1000 Genomes Browser, and NCBI Variation viewer).

PaRC will also build and host a metadata browser on our website that will permit users to browse the pan-genome, looking at the underlying the overall structure, underlying assemblies, and content sources as it evolves. Annotations will include locations and names of constituent contigs, assemblies and gaps; how regions of the reference map to older references and which regions of the reference don't map; regions that have "problems" (submitted or in-progress); and constitute genomes including details such as sequencing platform(s), creation tools, and cell line availability. Additionally, we will make the variant calls and even available phenotype data for the constitute and validation genomes publically available and queryable. We will also work to make this data made available via the GA4GH Beacon project as mentioned in Aim 7.1.

All materials hosted by us will be FAIRified: we will provide explicit licenses (with CC-BY by default) and DOIs for citation, as well as stable and versioned URLs.

Subaim 4.2: Disseminate information via announcements and training events.

In addition to hosting static content as above, we will engage in a variety of "push" efforts to inform the global biomedical community of updates from the HGRP. We will maintain a variety of public announcement e-mail lists for the HGRP, including high-level newsletters detailing HGRP activities, announcements of conferences and events, reference releases, and other developments. (We do not envision much public discussion taking place via e-mail, because it is not very scalable; see Specific Aim 5, below, for public engagement efforts, and Specific Aim 6 in Component 3 for internal coordination and support of HGRP activities.) These announcements will also be posted on blogs and Twitter, which are common modes of information dissemination in some relevant communities.

Another important method of information dissemination will be webinars and training events around the topics of using reference genome releases, navigating the HGRP resources, and understanding the release roadmap and HGRP roadmap. We will offer quarterly webinars on these topics, and provide reusable slides and recordings of the webinars via the public website, along with free-text searchable transcriptions of the webinars. In-person training events on these topics will be arranged at relevant meetings (e.g. ASHG, HGVS, Human Genetics and Genomics) as well as by invitation.

Subaim 4.3: Create, maintain, and provide onboarding materials for the community.

A key part of this curation overlay will be onboarding guides which help connect users entering the portal with the content, documentation, links out, and training materials they need. Our experience from open online communities (Carpentries, Python, R), as well as the NIH Data Commons Pilot Phase Consortium, is that confusion is an inevitable early part of a truly community oriented process. **Confusion can be masked (but not prevented) by an overly prescriptive process.** Confusion results from the wide variety of materials that will be made available, and the many different community members with distinct and overlapping needs. For example, clinical users may only be interested in exploring analysis interfaces and databases, while biomedical data scientists may wish to find R and Python APIs for cloud execution and associated documentation. Methods developers will need metadata formats and library APIs. New users can be quickly overwhelmed by complex systems for exploring and obtaining data, and the large volumes of information available.

We believe that the best way to address confusion is to grow onboarding materials as new materials are developed, the community grows, and community needs become more sharply defined. A key aspect of this is feedback from community engagement and training (Aim 5). This process is akin to building onboarding guides around “desire paths”, in which prescribed paths are adjusted in response to observed user behavior and needs. Our onboarding guides and FAQs will be regularly updated based on feedback from training events, webinars, and online user engagement via the issue tracker and social media.

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

Gathering user feedback and channeling it to the HGRP is critical for iteratively adjusting the HGRP approach to meet the needs of the larger user community. The most effective and scalable mechanism for doing this is to build a community of practice around the HGRP. This will include creating a community helpdesk and issue tracker (Aim 5.1), providing and maintaining a library of tools and a gallery of workflows (Aim 5.2), and building and maintaining a collection of training materials (Aim 5.3).

Subaim 5.1: Create a community helpdesk and issue tracker to track community issues and feedback.

We will provide a community helpdesk and open issue tracker via GitHub to track community issues and feedback. The helpdesk will serve to route users to the appropriate material, and we envision iteratively incorporating common helpdesk questions into onboarding guides, FAQs, and training materials. The issue tracker will target more technically capable users, and serve as a common repository of issues and needs that to be addressed by the HGRP. Curation and maintenance of both of these will be shared by the Project Manager and Community Facilitator, with automation support from the infrastructure engineer.

One challenge is that of “catastrophic success”, in which there is so much engagement via the helpdesk and the issue tracker that our team cannot curate it all. First, we note that this *is* success, even if catastrophic! (The alternative is irrelevance.) Second, all of our proposed infrastructure (e.g. the GitHub issue tracker) is scriptable, so we can invest in automation as needed to better support scaling. Third, we believe that this can also partly be managed by providing mechanisms for the community to help maintain the issue tracker and documentation (this Aim). **In sum, this type of “catastrophic” success is a common challenge faced by virtually every successful open online project, and we are more experienced than most in managing it.**

Subaim 5.2: Provide and maintain a tool library, and create a tool and workflow gallery.

Advanced users (biomedical data scientists, methods developers, analysis infrastructure developers, and database maintainers) will need tools and workflows that work with each version of the human reference as well as its associated metadata. These users are an important aspect of adoption, as they are already working with large data sets and used to technical engagement around tool progression. **We will work with the bio.tools repository** (see Letter of Support) to build and maintain a library of FAIR tools that work with the human reference. As with the current bio.tools site, tools will be hosted by the creators, and we expect most workflows to be hosted on dockstore. We will work with bio.tools to provide an update mechanism based on GitHub PRs, a technique that facilitates decentralized community management of the library. We will also work with bioconda to make as many of the command line tools installable with conda as possible, and facilitate interlinking between bio.tools and bioconda recipes.

We will also provide a **workflow gallery** which will enable users to browse and identify workflows that meet their analysis needs, along with tools and platforms that they can use to run the tools. This workflow gallery will

gather information from bio.tools and dockstore and present a curated set of tools and workflows grouped by analysis type (e.g. variant calling). The gallery will be designed to allow novice genomics users to browse potential research analyses in order to find analyses that might help to answer their research question. Where practicable we will also make use of continuous integration mechanisms on the gallery to verify that tools and workflows hosted there continue to work correctly. As we release new references (Aim 3) we will annotate the library and the gallery with information about which releases and representations the tools and workflows support.

Subaim 5.3: Maintain and update a collection of training materials with help from the community.

The HGRP will face a large number of training needs around reference maintenance and release, including understanding the features of various reference releases (Aims 1-3), navigating the portal developed in Aim 4.1, and adopting new tools and workflows surfaced in Aim 5.2. We will build a scalable and effective training program that also allows us to engage closely with the various user subcommunities. We will combine "push-only" training such as online webinars with a variety of online engaged events, in-person events, and an increasing collection of community-supported training materials.

Our core training approach relies on user engagement, both directly (via invited and opportunistic presentations at conferences; see Aim 4.2) and through a network of volunteer instructors. This is essentially the model pioneered by the data science community and most specifically by the Carpentries. In addition to invited presentations and tutorials, we will run a variety of events that include quarterly "fireside chats" using a discourse chat site, blog and twitter "ask me anythings", issue tracker "bug days", and in-depth in person workshops (e.g. Dr. Brown's annual Data Intensive Biology Summer Institute, and the UW Biostatistics Summer Institutes). We expect these in-person events to be essentially cost neutral except for travel.

We will also provide a small amount of travel support for community members who wish to run training and engagement events at their own community conferences. We will grow a cadre of enthusiastic participants who work with us to organically deliver HGRP training materials at topic-specific conferences, and we will provide them with travel support. (Dr. Brown in particular has tremendous experience and success with this model of training, both with his own workshops at UC Davis and as part of the Carpentries.) Feedback and questions from these events will be entered into the issue tracker, and recurrent issues will be noted for prioritization at the level of the HGRP.

We expect that the first several cohorts of students and instructors will be established users of the current human genome. We will work to extend our initial reach by running outreach sessions at conferences which are likely to host prospective users. Depending on the format of the conferences, we will host in-person trainings, informational seminars, and/or informational booths.

This community, in turn, is what makes the creation of a training program achievable and scalable. In the Carpentries, skilled researchers willingly invest time to participate as trained instructors, and contribute to curriculum development and maintenance. Evidence suggests that these volunteer instructors are motivated to participate because doing so aligns with a variety of personal and professional goals. As this cohort of instructors grows, so too will the breadth of researchers that we can reach. As new instructors with varying specialties are added to the instructor pool, the specialties of the learners who hear about their workshops will shift as well. By adopting this Carpentries-like training program, we will encourage organic growth of our user base and the community as a whole. Further, unlike traditional top-down training initiatives, the only additional costs associated with sustaining such a program involve oversight.

References

Grudin, J. (1994). Computer-supported cooperative work: history and focus. *Computer*, *27*(5), 19–26.

Lave, J. (1991). *Situated Learning: Legitimate Peripheral Participation (Learning in Doing: Social, Cognitive and Computational Perspectives)* (1st edition). Cambridge University Press.