National Advisory Council for Human Genome Research, September 24, 2018
The Human Genome Reference Program
Concept Overview

Background: The Genome Reference Consortium (GRC) provides a resource used by essentially all researchers who need to align experimental or patient genome sequence data. It also serves as a consensus coordinate system for reporting results. The GRC improves the reference, curates and releases updates and new reference builds, develops representations and alignment tools so that the community can use the reference and alternative haplotype information within it, resolves error reports from the community, and performs outreach. The GRC operates as an international consortium, with support provided by NHGRI, as well as the Wellcome Trust (funding the Sanger Center and the European Bioinformatics Institute) and the National Center for Biotechnology Information. A separate NHGRI-funded component is sequencing additional human genomes to very high quality for incorporation into the reference. This effort has produced eight assemblies, with six in progress, and 3-5 more planned from diverse populations (>75% non-European Ancestry). It also produces high quality non-human primate genomes.

On March 1, 2018, NHGRI convened a web meeting of over 65 basic research, clinical, and bioinformatic scientists to discuss scientific opportunities for the genome reference. The meeting addressed key research and resource opportunities for improving the human reference; activities necessary to keep the reference relevant and useful; clinical and research community needs (including education); related resources; and collaborations.[1]

The high-level conclusion of the meeting was that the current version of the human reference does not adequately represent human haplotype variation, that the tools to use existing alternative haplotype information are not well-used, and that there is an opportunity to significantly improve the human reference by developing it into a "Pangenome". One goal of a pangenome reference is to represent essentially all human haplotype variation, implying that any newly sequenced experimental or patient haplotype will be readily alignable to the reference. This would require addition of many more high-quality human genome assemblies chosen to maximize haplotype diversity. This would also require the adoption of better ways of representing the data (e.g., as a genome graph), along with the development of new informatics tools to make use of the new reference.

As a result of these discussions, NHGRI proposes five FOAs (see Table) to re-organize and re-focus its contribution to the genome reference to enable an improved human genome reference for the community, and to foster its long-term sustainability and improvement.

We anticipate that NHGRI's contribution to genome references will be most effective in collaboration with ongoing international efforts (including those of The Wellcome Trust, EBI, NCBI, GA4GH) to build, maintain, and serve references to the worldwide genomics community.

[1] (https://www.genome.gov/pages/research/sequencing/meetings/hgr_webinar_summary_march1_2018.pdf.)

These NHGRI funded concepts propose work only on human; we note that the larger GRC supports references for mouse and zebrafish.

Proposed Components of an NHGRI Human Genome Reference Program (HGRP)

Components/ Mechanism	Description	Rationale	$/year; # awards; #years
Human Genome Reference Center (HGRC) (UM1)	1. Construct and release new versions, receive and resolve error reports, prioritize genomes for addition to references; 2. Implement state of the art reference representations (incl. alternative haplotypes), for community use; 3. Community outreach and training for use of reference, aggregator of tools for use of reference; 4. Logistic and scientific coordinating center for NHGRI GRC program.	Essential function for a human genome reference community resource	$2.5M; 1 award; 5 years
High Quality Reference Genomes (HQRG) (UM1)	Efficient production of additional high quality (haplotype resolved) genomes to add to the Reference Resource.	Existing reference resource does not include sufficient haplotype diversity; improve quality/representation.	$3.5M; 1 award; 5 years
R&D for Reference Representations (U01)	R&D for the next generation of representations, e.g. graph-based and alternatives (e.g. reference-free)	Important for long-term development of reference resources	$1.25M; 3-4 awards; 3 years
R&D for Comprehensive Genome Sequencing (Guide notice to add $1.5M to existing seq tech dev; R01/21; R43/44)	R&D to enable comprehensive (e.g. telomere to telomere) genomes	Improve methods for high quality genomes	$1.5M; 3-4 awards; 4 years
Informatics tools for the Pan-genome (U01)	Development of informatics tools for use of/with next generation representations of references.	Important for long-term; serving diverse scientific communities.	$1.25M; 3-4 awards; 3 years
			TOTAL $10M

National Advisory Council for Human Genome Research, September 24, 2018
Concept Clearance for FOA
HGRP Concept 1: "Human Genome Reference Center"

Purpose: To provide a high-quality Human Genome Reference sequence to the scientific community to enable genome sequence analysis; to integrate the products of, and help coordinate, the other elements of the Human Genome Reference Program (HGRP).

Background: See "Concept Overview". The HGRP will need a central group that works as part of the Genome Reference Consortium (GRC) to maintain and update human reference sequence versions. This group will also work with the HGRP and the larger scientific community to prioritize sample choice and develop quality standards for new high-quality genome assemblies; implement state-of-the-art representations of alternate haplotypes; identify and respond to diverse community needs (e.g., clinical and basic genomics), liaise with other resources that represent human genomic sequence and variation and/or that provide reference resources for human and other organisms. The HGRC is expected to integrate with the GRC and other international efforts that also have responsibility for providing genome references. that

Proposed Scope and Objectives: The HGRC will be the central component of the HGRP. It will:

1. Construct and release new human reference sequence versions (including patches and full updates); incorporate high quality human genome sequence data provided by another program component and from elsewhere if available; receive and resolve error reports.

2. Implement, in the context of providing the reference, state-of-the-art representations that include alternative haplotypes. Provide basic tools for community use of the reference.

3. Provide community outreach and training for use of the reference; act as an aggregator of informatics tools created by the community for use of the reference.

4. Be the logistical and scientific coordinating center for the NHGRI HGRP; working to implement a human "pangenome" reference in a way that maximizes its value to the community.

Relationship to Ongoing Activities: NHGRI previously funded a U41 on "Improving the Human Reference Genome Resource". Funding for that effort ended in FY16, and the award has been in a no-cost extension. We do not fund other activities equivalent to the HGRC.

Mechanism of Support: UM1 (Cooperative agreement)

Funds Anticipated: $2.5M/year; 1 award; 5 years; starting in FY19.

National Advisory Council for Human Genome Research, September 24, 2018
Concept Clearance for FOA
HGRP Concept 2: "High Quality Reference Genomes"

Purpose: To produce very high quality, haplotype-resolved human genome sequence assemblies to add to the Human Genome Reference Resource.

Background: See "Concept Overview". The current human genome reference does not adequately represent human haplotype diversity. Participants on the March 1 conference call recommended adding at least 300 diploid genomes to the reference for inclusion in a "pangenome" resource, to facilitate studies across a wide swath of genomics, including: clinical studies, population studies, studies of variation, genomic basis of disease, and analyses of individual patient genomes. To achieve appropriate haplotype diversity, samples should be selected from ancestrally diverse populations, informed by population genetics and by community needs for basic genomics, variant discovery and clinical applications. Failure to add sufficient haplotype representation will limit future analyses, with consequences that may include under-representation of diverse populations in basic and clinical research. Additional high-quality genomes also will illuminate previously under-sampled genomic features, such as: repeats, centromeric regions, and structural variation.

High quality "long-read" data are becoming easier to obtain and potentially less expensive. NHGRI is initially pursuing this recommendation through two existing grantees who produce high-quality genomes—the grantees have previously produced 14 genomes and are aiming to produce ~50 more with FY18 supplemental funding. The proposed FOA will focus this effort on NHGRI's more refined goals as part of the new HGRP program organization for a human reference resource. This FOA will also open the effort to competition and peer review.

Proposed Scope and Objectives: Produce up to 350 haplotype-resolved human genomes using diverse samples consented for full data release. Awardee will initially use 1000 Genomes Project samples. Awardee will also be expected coordinate the consent and collection of new samples if they are needed to obtain appropriate haplotype diversity. Sample selection and prioritization will be done in cooperation with the Human Genome Reference Resource Center. This component may also provide capacity to help resolve error reports received by the Human Genome Reference Resource Center. The technology, cost, required quality, and added value of new genomes will be assessed by the program over the course of the project; funding, or specific uses of the capacity, may be adjusted accordingly.

Relationship to Ongoing Activities: NHGRI currently funds one award (Kwok, through 2022, $444K total costs/yr) includes production of high-quality human genome assemblies. Other related efforts mentioned above end in 2018.

Mechanism of Support: UM1 (Cooperative agreement)

Funds Anticipated: $3.5M/year; 1 award; 5 years; starting in FY19.

National Advisory Council for Human Genome Research, September 24, 2018
Concept Clearance for FOA
HGRP Concept 3: "R&D for Reference Representations"

Purpose: To support research and development for a next-generation genome reference representation that can capture all human genome variation.

Background: See "Concept Overview". During the March 2018 web meeting, participants strongly converged on the need for a "pangenome" reference that faithfully and usefully represents sequences from diverse haplotypes and populations. Currently the linear genome reference has been extended to include alternative pathways ("alt paths") that are added on top of standard reference sequence to show some population-based variation. However, there is a recognized need for more advanced representations, such as a graph representation. A full graph representation is expected to provide better ways to understand haplotype diversity and complex variation such as structural variation, copy number variation, and repetitive sequences. Graph methods may also support efficient genome data management by compressing out redundancy in population-based sequences and enable use of network algorithms developed in other domains.

Proposed Scope and Objectives: While the concept of the graph representation of the pangenome is well-established, further research and development is needed to refine and implement it to be usable as a practical human genome reference. Work is needed to demonstrate efficiency, scalability, computational speed, ease of use, adoption, and ability to foster analysis tool development for a wide range of purposes. In addition, other next-generation representations other than graph representations may emerge. This FOA would fund multiple projects that will together help set benchmarks and standards in this domain for a pangenome representation. A primary requirement is to adhere a high level of open science including open-source tools, standards, and specifications to enable this core resource to be integrated in the larger community and support outside contributions.

Representations developed by this effort will be candidates for adoption within the Human Genome Reference Center (see "Concept Overview"). However, successful efforts should also be independent and based on common standards. This would allow for the filed to evolve to allow for broader efforts, such as the development of community specific references.

Relationship to Ongoing Activities: The NIH BD2K initiative previously funded a Center of Excellence to David Haussler (1U54HG007990) that included some work in this area. NHGRI will continue to accept R01 and R21 applications in this area submitted outside the auspices of this FOA that adhere to open-science principles.

Mechanism of Support: U01 (Cooperative agreement)

Funds Anticipated: $1.25 M/year; 2-3 awards; 3 years; starting in FY20.

National Advisory Council for Human Genome Research, September 24, 2018
Concept Clearance for FOA
HGRP Concept 4: "R&D for Comprehensive Genome Sequencing"

Purpose: Develop technologies for complete de novo sequencing of phased diploid human genomes.

Background: See "Concept Overview". During the March 2018 web meeting participants discussed the need for human references based on complete phased human genome sequences. They agreed that technologies available today are on the cusp of completely phasing and resolving previously unsolved regions of the genome. The attendees also stressed the need for further technology development to address particularly difficult regions of the genome that are generally large and composed of identical or nearly identical repeats, and not adequately resolved comprehensively with sequencing methods in use today. The participants coalesced around the need for methods to achieve telomere to telomere sequencing of phased human genomes.

Proposed Scope and Objectives: Generate new, or significantly extend existing, DNA sequencing technology to achieve contiguous "telomere to telomere" phased human genome sequencing of high quality. This includes integration of new or emerging technologies with existing approaches and data types, and a significant, but not standalone, informatics component. Efforts may focus on specific difficult regions, but generalizable approaches that comprehensively sequence genomes are preferred. The objective will be genomes with high contiguity, completeness, and phasing with metrics at least ten-fold better than what is produced today using the most specialized technologies and methods. Development of these improvements for reference-quality methods is also expected to catalyze and enable improvements in routine genome sequencing.

Coordination between this and other program elements in the HGRP will facilitate generation of the best quality references possible.

Relationship to Ongoing Activities: One NHGRI award develops methods for high-quality assemblies, with some data production (funded FY19 – FY22). The Novel Nucleic Acid Sequencing Technology Development R01 and companion R21 and R43/R44 announcements already encourage applications in this area.

Mechanism of Support: Guide notice pointing to existing Novel Nucleic Acid Sequencing Technology Development RFAs for R01, R21 and R43/44.

Funds Anticipated:
R01/R21: Additional $1.5M/year; 3-4 awards; 4 years; starting in FY20.
R43/R44: Additional $1.3M/year dedicated to Small Business Awards; 2-5 awards; 1-2 years; starting in FY20-FY22.

National Advisory Council for Human Genome Research, September 24, 2018
Concept Clearance for FOA
HGRP Concept 5: "Informatics Tools for the Pan-genome"

Purpose: To develop informatics tools that can apply the new pangenome representation for analysis and enable use of the high-quality genome reference by clinical and basic researchers.

Background: See "Concept Overview". Making the genome reference easily usable for clinical and basic researchers is key to its adoption and relevance in the field. Participants on the March 2018 web meeting described a gap in bioinformatics tools that are available for the current reference, and noted that existing tools can be difficult to use or find, especially ones that use existing haplotype variation information. The pangenome is expected to support new types of analysis and enable questions to be addressed that previously have been out of reach. Future tool development should both emphasize the needs of biological and clinical scientists and democratize access to the pangenome reference so that the field develops evenly.

Proposed Scope and Objectives: We seek development of analysis tools for the pangenome reference that are also user-friendly for communities of biological and clinical scientists. Tools should be exemplars that can both take advantage of the emerging, next-generation, pangenome reference representations and also gain adoption for scientific impact. Given that communities of researchers will be using older reference representations and builds for some time, we also value development of tools that are "backwards-compatible" across new and older reference representations and versions. The work sought here can include new applications or revision/improvement of existing high-value toolkits to use the new reference representation. It is expected that tools will be developed with high engineering standards in an open-science paradigm and adhere to community standards. This investment is intended to ensure that the impact of the reference is fully realized in the community. Diverse applications will be considered: clinical and basic, covering different aspect of genome analysis that rely on the genome reference, e.g., variant detection, admixture analysis, interpretation of patient variants. Awardees will be expected to participate in the research network to provide feedback on the use of the pangenome resource.

Relationship to Ongoing Activities: NHGRI recently issued a PAR for a broad range of research efforts in computational genomics, data science, statistics and bioinformatics, which includes development of novel informatics tools. NHGRI will accept R01 applications in this area submitted outside the auspices of this proposed HGRP FOA.

Mechanism of Support: U01 (Cooperative agreement)

Funds Anticipated: $1.25 M/year; 3-4 awards; 3 years; starting in FY20.