Concept for a Human Genome Reference Program

Adam Felsenfeld, Heidi Sofia, Mike Smith

September 24, 2018

The **Forefront** of **Genomics**

Human Genome Reference

- Origin in the HGP
- Coordinate system for the genome
- Used for mapping sequence reads – essentially all human genome sequencing using short read technologies
- Critical utility for genomics

Human Genome Reference Program (HGRP): Background

- Genome Reference Consortium maintains the current reference (Human; also mouse and zebrafish)
- NHGRI, Wellcome Trust (through Sanger and EBI), and The National Center for Biotechnology Information (NCBI)
- NHGRI contributions:
  - Serving the human reference (new builds, patches, error reports, outreach, tools)
  - Adding new high-quality human genome sequences
  - Ending in FY18

NIH NHGRI

HGRP Background

New considerations in changing genomics landscape:

- Different and diverse users (basic/clinical; degrees of sophistication)
- Legacy reference “builds” and updating reference versions/adoption
- Learning more about human variation and its relation to different populations/ancestries; need to include in reference to avoid analysis bias

Community Conference Call (March 1, 2018)

- The human reference does not adequately represent human haplotype variation; danger of bias in reference, need more HQ genomes (at least ~300)
- A single linear reference is clearly inadequate; need better ways of representing the information from multiple genomes as a reference, particularly as we add more HQ human genomes
- The current information about alternative versions (“alt paths”) that we actually do now have in the reference is under-used
- Opportunity to significantly improve the human reference by developing it into a “pangenome” to represent much more human variation of the type that is critical for the uses of the reference

Human Genome Reference Program

Five elements:

1. Human Genome Reference Center
2. High-Quality Reference Genomes
3. R&D for Reference Representations
4. R&D for Comprehensive Genome Sequencing
5. Informatics Tools for the Pangenome

1. Human Genome Reference Center (UM1)

- Construct, maintain, and release new reference versions; receive and resolve error reports
- Promote and adopt state-of-the-art representations that include alternative haplotypes; provide basic tools for community use
- Community outreach and training; act as an aggregator of informatics tools created by the community
- Coordinating center. Work with other program members to identify framework for a pangenome implementation

$2.5M/yr; 5 yr; 1 award

NIH NHGRI

2. High Quality Genomes (UM1)

- Produce at least 300 haplotype-resolved human genomes
- Sequence diverse samples (start with 1000 Genomes samples)
- Coordinate (open use) consent/collection of new samples, if needed
- Provide capacity to help resolve error reports, if needed

Technology, cost, quality requirements, optimal number of genomes/needed diversity will be assessed over the course of the project by the program as a whole.

$3.5M/yr; 5 yr; 1 award

NIH NHGRI

3. R&D For Reference Representations (U01)

- Develop, improve, and implement “next generation” reference representations, such as graph genome representations
- R&D on efficient, scalable methods, open source; product that facilitates use and tool development
- Help set benchmarks/standards for a pangenome representation
- May be adopted within the HGRC but also independent and based on common standards

$1.25M/yr; 3 yr; 2-4 awards; also investigator-initiated R01s

NIH NHGRI

4. R&D for Comprehensive Genome Sequencing (Guide Notice; R01/21; R43/44)

- New, or extended, technology for high quality, contiguous "telomere to telomere" phased human genome sequencing
- Includes integration of new technologies with existing approaches
- Generalizable approaches to comprehensively sequence genomes are preferred; focus on specific "difficult" regions allowed
- Metrics at least ten-fold better than state-of-art

$1.5M; 2-4 awards; 4 years (sequencing tech. dev. program)

5. Informatics Tools for the Pangenome (U01)

- Analysis tools that are exemplars for use of emerging, next-generation reference representation
- Diverse applications considered: clinical and basic, e.g., variant detection, admixture analysis, interpretation of variants, etc.
- “Backwards-compatible” tools across reference representations and versions
- New applications or revision/improvement of existing high-value toolkits
- Open science

$1.25M/yr; 2-4 awards; 3 years; also regular invest. init. R01s; SBIR/STTR

Thanks to

Heidi Sofia (Components 3, 5)
Mike Smith (Component 4)
Carolyn Hutter

dgSci, esp: Lisa Brooks, Taylorlyn Stephan

Teri Manolio

Discussion

Human Genome Reference Center	UM1	2.5M/yr	1 award	
High-Quality Reference Genomes	UM1	3.5M/yr	1 award	
R&D for Reference Representations	UM1	1.25M/ yr	2-4 awards	Integrate invest init. R01s/R21s; R43/44 into program as appropriate
R&D for Comprehensive Genome Sequencing	Guide Notice R01/21; R43/44	1.5M/yr	2-4 awards	
Informatics Tools for the Pangenome	UM1	1.25M/ yr	2-4 awards	

$10M/yr total

NIH NHGRI