Rapporteur: Taylorlyn Stephan
NHGRI Program Analyst

Human Genome Reference Sequence Opportunities Webinar – March 1, 2018
Executive Summary

Background

On March 1, 2018, over 65 basic research, clinical, and bioinformatic scientists in the genomics community convened for a four-hour web meeting on NHGRI-funded components of the Genome Reference Consortium (GRC). The GRC is an essential resource for the biomedical community and due to recent technological advances, NHGRI staff determined it was time to evaluate current GRC activities and discuss its possible future scope. The web meeting addressed these topics: 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. Several major themes and recommendations emerged during this web meeting.

A Pan-Genome

Many meeting participants expressed the need for a "pan-genome" – a reference genome that represents all human variation. The pan-genome will facilitate alignment for every sequence of interest, with a primary goal that no haplotype will go unaligned. The pan-genome would replace the current reference build, GRCh38, and its multiple alternative paths, as they do not fully represent the diversity of variation and haplotypes in humans. Additional sample collection and sequencing of diverse genomic ancestries were strongly advised for inclusion in the pan-genome. 1000 Genomes Project samples, or samples with the same consent categories and diverse provenance, were regarded as the highest sequencing priority. A corollary request to these additional samples were ones with original blood available for use in genome characterization and benchmarking. While there was discussion on incorporating existing genomes from large-scale sequencing projects, participants agreed that new samples would be easier to incorporate for several reasons: 1) samples from populations missing in GRCh38 could be obtained, 2) they could all be sequenced at the same depth and quality at one or a few institutions in a narrow timeframe, 3) their consents would be standard. For sequencing efforts, a pilot of 50-300 new genomes with haplotype resolution was suggested. There was not a detailed conversation or consensus as to what sequencing technologies or metrics (e.g., 30X sequencing, PacBio long reads, Hi-C) should be used in the pilot. Participants noted that how the pan-genome is displayed (in a graph, linear coordinates, etc.) should not limit its development. If most users can understand and access the data, then the reference format should be a back-end concern. Community members also argued that a pan-genome effort will motivate tool development and use.

Bioinformatic Tools

Several participants argued that bioinformatic tool development for use with the reference has been severely underfunded. In addition, existing tools are difficult to use or find, so future tool development should emphasize ease-of-use for the average biologist and medical scientist. There was an argument that tool development will follow naturally from a pan-genome effort, but there still is a critical lack of bioinformatic tools for current users of builds GRCh37 and GRCh38.

Migration to newer builds is often limited due to lack of tools to transfer annotations from one build to another. One suggestion was for the current GRC to curate a list of existing bioinformatics tools, as well as host educational workshops to train the community on how to use them. In addition, if the GRC can work with larger efforts (*All of Us*, GA4GH), then developers will be more likely to build tools that are backwards compatible with the reference and its alt pathways, as well as look forward to future representations.

Integrating Reference Data with Related Resources

It was argued that the reference sequence is essentially a variant resource, and thus should integrate with existing databases. Allele frequency and haplotype data (in EGA, ESP, ExAC, Bravo, gnomAD, etc.) would be extremely valuable if they were fully compatible with GRCh38 and a pan-genome. For clinical users, the challenge of getting CAP/CLIA approved pipelines updated to GRCh38 was noted as a major reason why most clinicians are still on GRCh37. This highlights why it is essential to bring all users to the current build. While it was unclear if participants argued for a single resource that encompasses the reference, haplotype, variant, and allele frequency data, a push for more integration and leverage of existing resources is clearly a priority for the community.

Education

There is not a set of standards or best practices on how to use the current reference and its tools. This highlighted the lack of existing infrastructure for the community to engage with the GRC. The lack of published standards and best practices leaves communities behind, especially the clinical community. Education and training opportunities should be a priority in current GRC activity and future funding. One suggestion was to hold a training workshop at ACMG tailored for the clinical community. This would allow the GRC to educate, gather feedback, develop clinical best practices, and gain more visibility.

Recommendations and Insights

1. Making the reference easy to use for clinical and basic researchers is key to its adoption and relevance.
2. Additional samples should be collected and sequenced for a future pan-genome. This should be a directed effort to collect samples from diverse populations with consent for broad data sharing (preferably open access) and general research use. One proposal was to sequence a pilot of 300 samples from the 1000 Genomes Project. This sequencing effort samples to high quality diploid resolution for this effort, which could begin now.
    a. The pan-genome should be part of an international, directed effort led by NHGRI and other institutions (e.g., Wellcome Trust, Sanger).
3. Better quality genomes are needed to define structural variation (50 genomes for detecting ≧5% MAFs) and to analyze challenging regions in the genome (e.g., MHC).
4. The GRC should integrate the information in its alterative paths with existing allele frequency data (in ESP, ExAC, Bravo, gnomAD, etc.); alignment data (RefSeq, Ensembl, etc.); interpretation data (ClinVar, HGMD, LOVD, etc.); high quality structural variant and copy number variant data; and phased data.
5. Reference standards, such as pipelines (sequencing, analysis, and informatic), additional sequences, protocols, and secondary materials, are essential for a comprehensive

reference. The GRC should work with CAP/CLIA, GIAB, and other groups to support benchmarking and technical uniformity.

6. Participants argued that there remains a need to continue to fund the current GRC or in the future, a GRC coordinating center. Moving into a pan-genome world would necessitate facilitating education and outreach, tracking tools and other resources, patching technical updates, responding to user inquiries, and coordinating data collection/analysis/release.
7. The GRC should curate a list of existing bioinformatics tools and resources for the community.
    a. This can be done in conjunction with generating reference "best practices" tutorials.
8. Any resources or effort put into the reference needs to be backwards compatible so no user is left behind.
9. Funding for new bioinformatic tools should be a priority.
    a. In addition, funding annotation resources and tools is important for the clinical community and genomic medicine.
10. There is a need to fund educational outreach and community engagement. This could be done through workshops and partnering with stakeholders and other large projects (Genome in a Bottle, *All of Us*).