Conference Report: Representing Ethnic Diversity for Precision Medicine

Luke Hickey[1], Benedict Paten[2], Robert Sebra[3], Valerie Schneider[4], and Justin Zook[5]

1 PacBio
2 University of California, Santa Cruz
3 Icahn School of Medicine at Mount Sinai
4 National Center for Biotechnology Information
5 Genome in a Bottle Consortium, National Institute of Standards and Technology

There is a pressing need to better represent ethnic diversity with genomic resources — and to do so in a way that maximizes utility for people working with the Genome Reference Consortium's human reference genome. That was the theme of a panel discussion among the authors at the Precision Medicine Leaders Summit, held in San Diego in August 2017. The session, entitled "Minority Report: Ethnic Diversity and the Real Promise for Precision Medicine," included discussion about better representing the human genome; addressing population bias in existing databases; evolving technology; data sharing; and achieving precision medicine for all individuals. Here, we summarize the key discussion points from the panel.

Representing the Human Genome

For all the effort and resources that have gone into it, the human reference genome is still an imperfect representation of what our genome really looks like. It was initially conceived as a linear, haploid genome, and that has been the foundation for any number of genome analysis tools purpose-built to mine it. With the latest build, GRCh38, we now have multiple representations of hundreds of genes, including regions that are highly diverse such as the MHC complex and KIR genes. Those alternative sequences are presented as snippets aligned to the chromosome, but this is not an ideal approach as we discover variants that look quite different from one person to another, such as a structural variant that may be an inversion in one individual and a deletion in another.

Graph-based methods allow for a more accurate and natural representation of the genetic diversity of humans in a single reference assembly. Today, the alternative sequences in GRCh38 are not used by the community nearly as often as the linear parts of the reference, which may in part be due to the challenge of processing information in this context. A graph is a more intuitive way to think about divergent sequences; it could more easily show all possible variants in real, rather than consensus, haplotypes. A major challenge in implementing such an approach, however, is its incompatibility with the existing reference and current bioinformatics tools built for a linear representation. Even small updates to the human reference genome are problematic for the community when they change coordinates in the genome, so we must consider the ramifications of any significant structural changes as we evaluate better ways to represent the human genome.

Despite that challenge, we must continue to evolve the human reference. We have learned much from it already, but there is still a long way to go to fully understand our genomes and the universe of sequence diversity they encompass.

Increasing Ethnic Diversity in Genome Resources

Any scientist who has tried to study a large cohort of individuals is familiar with the ethnic disparity that exists in curating variants for populations underrepresented in current genomic databases. African Americans, for instance, are more likely to have variants interpreted as having unknown significance than people of European descent. Interpreting variation across populations, and implementing that information for drug trials or diagnostic development, will be essential for moving precision medicine forward more universally.

The human reference genome was built with DNA sequences collected from about 50 different people, but for logistical reasons about 70 percent of the final reference genome came from one individual, a male of admixed African and European descent. The reference is a mosaic, primarily representing DNA characteristic of people with European ancestry. Variant interpretation for people of other ancestries would be significantly improved with more genomic data from as many ethnic groups as possible.

Many populations for which we lack data are from less advantaged countries or regions, where people may not be as informed about the value of genetic data. They may also be from groups that have historically been treated poorly by the research community, making them reluctant to take part in new studies. Figuring out how to reach out and encourage participation is a non-trivial matter; we must overcome fears about how data will be used in order to achieve the population coverage needed for precision medicine.

For scientists involved in genome interpretation, it is important to keep considerations about ethnic diversity in mind when analyzing genomic data. Representing different types of ancestry is critical even during assessment of a new interpretation workflow; performance in variant calling cannot be fully evaluated without ensuring that it works across genomes of various ancestries.

The community has not yet reached consensus on how best to collect and represent ethnic diversity for optimal use. We must have representations of genetic diversity, but will we be better served by having multiple references or a single reference that captures as much variation as possible?

Evolving Technology

Strides in accurately representing the human genome have been made with the availability of new technologies, such as long-read sequencing, that are complementary to short-read or Sanger sequencing tools. There is still dark matter in the human genome — regions of the genome that are intractable to short-read technologies — that is now being characterized with these newer approaches. Finishing centromeres and telomeres, for instance, is an important goal.

These new technologies have been very effective at uncovering variation missed by short-read sequencing, such as structural variants. The current reference genome is excellent at helping scientists identify SNPs from short-read data, but it is much more challenging to identify

structural variants, which can have major phenotypic effects. Initiatives such as the Genome in a Bottle Consortium, for instance, are integrating complementary short, linked, and long read sequencing, as well as optical and nanopore mapping, to characterize complex regions and establish high-confidence benchmark variant calls. A current GRC project at the McDonnell Genome Institute involves using complementary technologies to sequence individuals from various populations, with the goal of feeding that information into the reference genome. Other population-specific approaches have generated important reference-grade genomes for Korean, Chinese, and other ethnic groups. These assemblies have tremendous value to scientists or clinicians who must interpret variants in patients of these ethnicities.

Ultimately, technology is selected based on what must be detected, but utilizing complementary tools offers the greatest likelihood of identifying everything from SNPs to complex structural variants.

Data Sharing

While data sharing was a small part of the discussion, the authors agreed that it is the single biggest challenge for practicing precision medicine today. Sharing data across international boundaries, or across population studies with different consent rules, is exceedingly difficult. Making information broadly useful for the whole community — and for patients everywhere — would have enormous benefit for society. Ultimately, it will be necessary to link phenotypic data to genomic data as well, and this introduces a new data-sharing hurdle: protected patient information.

As a community, we must address this issue by improving our consent materials and making data as open and accessible as we possibly can.

Precision Medicine for All

Improved reference genomes and increased representation of ethnic diversity will ultimately power next-generation diagnostics, targeted therapies, and accurate variant interpretation. From a clinical perspective, the big win will come from making it easy to type variants that are currently too difficult to implement for patient care. Translating the raw data and making it accessible for clinical use, however, remains a challenge.

Early successes have come in pharmacogenetics and HLA typing, where long-read sequencing of CYP2D6 alleles and the HLA locus has enabled the discovery of novel variants as well as more accurate representation and phasing of known variants. Understanding drug metabolism through CYP2D6 alleles, for example, will be important for stratifying patients into populations most likely to benefit from a particular medication. Also, long-read sequencing has been integral for detecting the repeat expansion associated with fragile X syndrome, which has implications for both diagnosis and prognosis.

It is exciting to see precision medicine taking its first steps, but concerns about unequal benefit for different populations are well-founded. In order to make the promise of precision medicine

available to everyone, we must make significant progress in representing the genetic variation in all ethnic groups.

Acknowledgements
Certain commercial equipment, instruments, or materials are identified to adequately specify experimental conditions or reported results. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the equipment, instruments, or materials identified are necessarily the best available for the purpose.