LUKE HICKEY
SENIOR DIRECTOR FOR HUMAN BIOMEDICAL APPLICATIONS
PACBIO

LUKE HAS SERVED ON THE COMMITTEE FOR THE NIST GENOME IN A BOTTLE CONSORTIUM, AND THE TEAM SUPPORTING THE 1,000 GENOMES PROJECT STRUCTURAL VARIATION WORKING GROUP, AS WELL AS THE HUMAN GENOME REFERENCE CONSORTIUM.

HUNTING STRUCTURAL VARIANTS: POPULATION BY POPULATION

HOW ARE POPULATION AND ETHNIC-GROUP SPECIFIC EFFORTS AROUND THE GLOBE WORKING TO CATALOGUE VARIANTS IN UNDERREPRESENTED GENOMIC DATABASES?

Until recently, most population-scale genome sequencing studies have focused on identifying single nucleotide variants (SNVs) to explore genetic differences between individuals. Like so many SNV-based genome-wide association studies, however, these efforts have had difficulty identifying causative genetic mechanisms underlying most complex functions.

More and more, the genomics community has realised that structural variation is likely responsible for many of the traits and phenotypes that scientists have not been able to attribute to SNVs. This class of variants, defined as genetic differences of 50 bp or larger, accounts for most of the DNA sequence differences between any two people. Structural variants (SVs) are also already known to cause many common and rare diseases including ALS, schizophrenia, leukemia, Carney complex, and Huntington's disease.

Despite the importance of SVs, these larger variants have been understudied and underreported compared to their single-nucleotide counterparts. One reason is that they remain difficult to detect. Their length often means they cannot be fully spanned using short sequencing reads. They also often occur in highly repetitive or GC-rich regions of the genome, making them challenging targets. As such, this class of human genetic variation has remained vastly under-explored in global populations and is now ripe for discovery.

In recent studies, researchers have overcome this challenge by making use of new long-read sequencing methods. Using average reads longer than 12,000 base pairs, these studies describe a five to 10-fold increase in the number of SVs detected in a single human genome, putting the total number at around 20,000 to 30,000 SVs, accounting for about 10 Mb of unique genomic sequence.

"FOR MANY GROUPS, THE FIRST STEP IN CATALOGING STRUCTURAL VARIANTS ACROSS POPULATIONS IS TO ESTABLISH REFERENCE-GRADE GENOME ASSEMBLIES FOR THE POPULATION OF INTEREST."

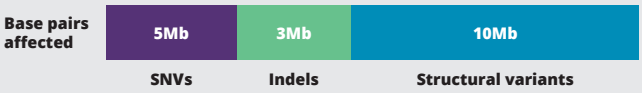
With these long-read sequencing advances, scientists around the world are now intent on discovering population-specific structural variants, particularly those that contribute to disease or drug response. From ethnic reference genomes to gold standard SV call-sets and large-scale SV mapping projects, these studies represent a new frontier in characterising and cataloging human genetic variation. Ultimately, novel genetic findings from such studies will be key contributors to the discovery of disease genes and mechanisms, bringing novel targets for drug discovery and development pipelines. (See fig. 1)

Detecting Structural Variants

Structural variations occurs in many flavours within and between human genomes, including insertions, deletions, translocations, inversions, copy number variation, repeat expansions, duplications, and more. The size and frequency of this class of variants is also diverse, with the majority ranging from 50 bp to 6,000 bp, with short insertions and deletions (collectively called indels) comprising the vast majority of events. (See fig. 2)

Despite the diversity of SVs, the one constant has been their elusiveness for scientists attempting to resolve them using short-read sequencing platforms. These platforms typically generate reads of 200 bp to 300 bp, making it difficult to map across longer variants, or map uniquely in or near repetitive sequence where SVs often occur. An illustrative example of this comes from repeat expansion disorders, including fragile X syndrome and many ataxias, which are caused by pathogenic insertion SVs occurring in repeats[1]. →

Figure 1. Variation between two human genomes, by number of base pairs affected.



Because of the predominance of short-read sequencing, most human genome studies to date have reported just a small fraction of the structural variation that really exists in our DNA. It was only when scientists began using long-read sequencing to interrogate human genomes that these previously undetected variants began coming out of the woodwork. In a recent study from the Human Genome Structural Variation Consortium, for instance, scientists determined that long-read sequencing routinely discovered seven times more structural variants compared to short-read sequencing, even in thoroughly characterised data sets such as the 1000 Genomes Project cohort[2]. The scope of this initial long-read study was limited to three ethnically diverse parent-child trios (Han Chinese, Puerto Rican, and Yoruban Nigerian). However, it demonstrates that the majority of common structural variation in global populations is still yet to be discovered.

Following the success of these early SV mapping projects, scientists have also begun to explore the sensitivity of long-read sequencing to detect causative pathogenic SVs in rare disease cases. Rare diseases are individually rare, but collectively common and estimated to affect 10% of the population. Despite our knowledge that around 80% of rare disease are genetic in origin, the solve-rate using short-read sequencing methods remains stubbornly low, between 25% and 40%[3].

In a recent publication from scientists led by Euan Ashley, a patient whose debilitating syndrome had defied diagnosis for 20 years was finally determined to have Carney complex, a rare disease caused in this case by a structural variant[4]. The SV, a 2.2 kb deletion in the PRKAR1A gene, was discovered with long-read sequencing after short-read sequencing failed to provide answers.

As the community works to build a comprehensive database of human structural variation, efforts are underway to find variants that may be common in specific sub-populations. Just as understanding allele frequency has been a critical component of interpreting the functional importance of SNVs, so too will this kind of information help scientists and clinicians figure out whether a particular structural variant is likely to be causative of disease in a specific individual.

Reference Genomes

For many groups, the first step in cataloging structural variants across populations is to establish reference-grade genome assemblies for the population of interest. One recent study used a form of long-read sequencing known as single-molecule, real-time (SMRT) sequencing and complementary technologies to generate a high-quality de novo genome assembly for an individual of Chinese descent. Compared to the latest human reference genome, the Chinese genome assembly has nearly 13 Mb of novel sequence[5]. Analysis of the genome found a sizable number of structural variants, with 20,000 insertions and deletions alone. Nearly 50 indels located in exons were determined to be specific to the Chinese assembly, with potentially significant implications for biological function in the broader population. (See fig. 3)

The results from this initial study were so enlightening that a second group in China announced plans to build a robust database to further map common structural variants in the population by sequencing 1,000 Chinese individuals with SMRT sequencing. Scientists expect that their database will include several disease types and will provide new insight into the effects of population-specific variants, ultimately improving the delivery of precision medicine.

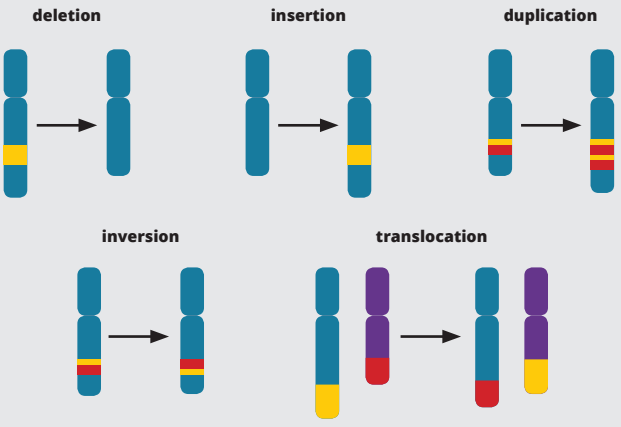
Similarly, studies have produced reference genomes for Korean and Japanese individuals using long-read sequencing to capture structural variation. The Korean reference includes fully phased chromosomes and is so contiguous that 90% of the genome is represented in just 91 scaffolds[6]. Scientists detected more than 18,000 structural variants, of which two-thirds had never been previously identified. Almost 10% of the variants were specific to people of Asian descent. Among the variants was a CYP2D6 duplication, determined in a validation effort to be clinically relevant. In the two Japanese reference genomes assembled, scientists found more than 6 Mb of sequence that was absent from the human reference genome[7]. The novel sequence was attributed to more than 9,600 insertions, including variants common to the Japanese population as well as specific to the individual.

Other major improvements in reference assemblies have come from consortium approaches, such as the Genome Reference Consortium (GRC), caretaker of the official, international human reference genome, and the Genome in a Bottle (GIAB) consortium. The GRC has invested significant resources to improve assemblies for a number of ethnic groups. By incorporating SMRT sequencing along with other technologies, they have now released high-quality assemblies for individuals of Han Chinese, Puerto Rican, European, Yoruban, Colombian, and Gambian descent[8]. Boosting representation of each ethnic group has benefits not only for members of that group, but also for people who are more closely related to these groups than to the groups most represented in public genetic databases, which tend to skew European. Similarly, the GIAB team has sequenced and made publicly available several new reference-grade assemblies for family trios. These currently include an Ashkenazim Jewish trio and a Han Chinese trio, with plans to expand to additional groups in the future.

Gold Standard

All of these reference assemblies serve as a solid foundation on which to build our understanding of structural variants, but there is another important element needed: carefully validated

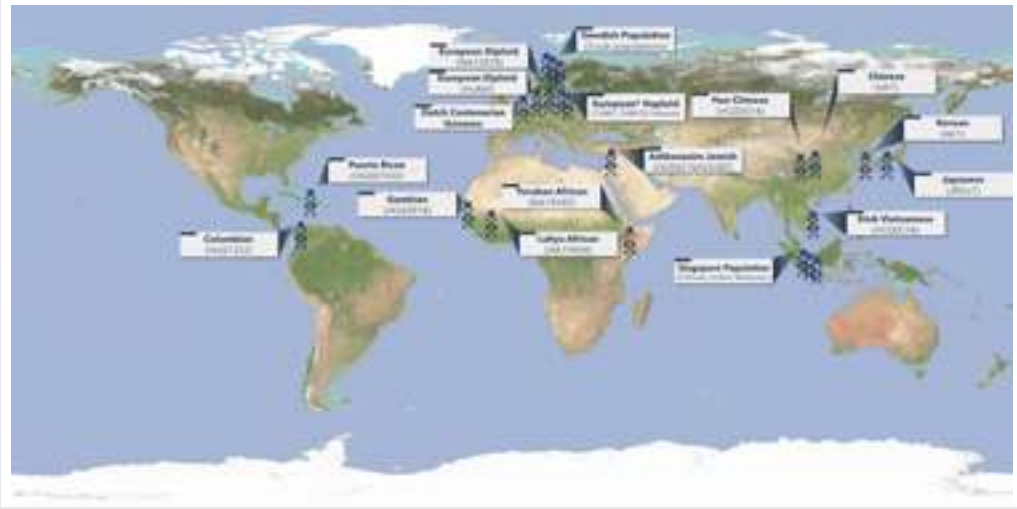
Figure 2. Examples of Structural Variation



sets of variants. Like the truth sets established by GIAB for SNVs, or large public databases such as gnomAD that serve this purpose, the genome interpretation community will need a gold standard database of structural variants to help verify that results generated from any given project are accurate.

In the Human Genome Structural Variation Consortium project, recently reported in a bioRxiv preprint, a large team of scientists performed a detailed analysis of structural variation in the human genome. Using a number of approaches, including SMRT sequencing, short-read sequencing, optical mapping, scaffolding services, and more, they characterised family trios from people of Han Chinese, Puerto Rican, and Yoruban descent. These particular ethnic groups were chosen for their range of genetic diversity: low in Han Chinese, high in Yoruban, and mixed in Puerto Rican. Each human genome was assembled to extremely high quality and phased for accurate diploid representation.

Figure 3. Map of Global Population-Specific Reference Genome Projects



Since the work was done on genomes collected for the 1000 Genomes Project, it was expected that they had been thoroughly characterised already. However, investigators found more than 31,000 structural variants in each genome, seven times the number found in previous analyses using short-read sequencing. They also discovered more than 150 inversions in each genome and some large chromosomal rearrangements, most of which had not been detected before. The team noted that SMRT sequencing delivered improved access to structural variants from 50 bp to 500 bp and better sequence resolution for insertions of all sizes; they estimate that 77% of insertions are typically missed by variant-calling algorithms working with short-read data. "Using ~15 kbp reads at an average of 40-fold sequence coverage per child, we have been able to span areas of the genome that were previously opaque and discover three to fourfold more structural variation when compared to short-read sequencing platforms," the scientists reported in the preprint publication.

Meanwhile, scientists involved with the GIAB consortium have been working to expand their SNV-based truth sets with comparable information for structural variants. This kind of resource has been a major goal for the biomedical community. Late last year, scientists from the University of Washington and the McDonnell Genome Institute used SMRT sequencing with two haploid human cell lines

> MORE AND MORE, THE GENOMICS COMMUNITY HAS REALISED THAT STRUCTURAL VARIATION IS LIKELY RESPONSIBLE FOR MANY OF THE TRAITS AND PHENOTYPES THAT SCIENTISTS HAVE NOT BEEN ABLE TO ATTRIBUTE TO SNVs."

to "build a verifiable gold standard for human genetic variation," they reported[9]. They cataloged several types of structural variation, from just 2 bp to as long as 28 kb. Their analysis of some 460,000 variants revealed that nearly 90% of them had not been detected by the 1000 Genomes Project, even though scientists determined that most of the variants are common among the human population. "The theoretical amount of genetic variation in a single human diploid genome far exceeds expectations established by previous whole-genome studies," they concluded.

Moving Forward

Now that it is technically possible to detect structural variation routinely and robustly in human genomes, there is a veritable gold rush to discover as many of these variants as possible. Scientists around the world are racing to fill the catalogs of structural variants; by accomplishing this in diverse populations, the community is generating resources that will be crucial for interpreting the biological and clinical relevance of these elements. Only when we recognise the full spectrum of human genetic variation will we be able to improve healthcare through precision medicine. ■

References:

1. Ashley EA. Towards precision medicine. Nat Rev Genet. 2016 Aug 16;17(9):507-22. doi: 10.1038/nrg.2016.86.
2. Chaisson MJP, Sanders AD, Zhao X, Malhotra A, et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. bioRxiv preprint. doi: 10.1101/193144
3. Rare Diseases: Facts and Figures (webpage). Global Genes. https://globalgenes.org/rare-diseases-facts-statistics/
4. Merker JD, Wenger AM, Sneddon T, Grove M, et al. Long-read genome sequencing identifies causal structural variation in a Mendelian disease. Genetics in Medicine (2017). doi:10.1038/gim.2017.86
5. Shi L, Guo Y, Dong C, Huddleston J, et al. Long-read sequencing and de novo assembly of a Chinese genome. Nat Commun. 2016 Jun 30;7:12065. doi: 10.1038/ncomms12065.
6. Seo J, Rhie A, Kim J, Lee S, et al. De novo assembly and phasing of a Korean human genome. Nature 538, 243–247 (13 October 2016). doi:10.1038/nature20098
7. Japanese Reference Genome: The portal site (webpage). Tohoku University. https://jrg.megabank.tohoku.ac.jp/en/
8. Project: Reference Genomes Improvement (webpage). McDonnell Genome Institute. http://genome.wustl.edu/projects/detail/reference-genomes-improvement/
9. Huddleston J, Chaisson MJP, Steinberg KM, Warren W, et al. Discovery and genotyping of structural variation from long-read haploid genome sequence data. Genome Res. 2017 May;27(5):677-685. doi: 10.1101/gr.214007.116.