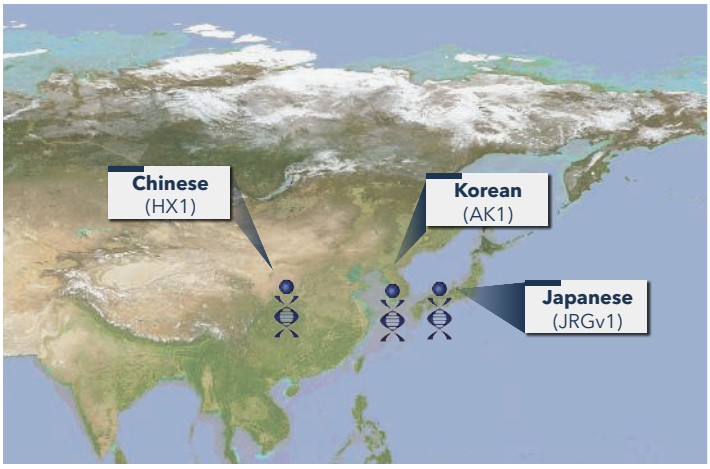
IMPROVING PRECISION MEDICINE STUDIES IN ASIA USING ETHNICITY-SPECIFIC HUMAN REFERENCE GENOMES AND PACBIO® LONG-READ SEQUENCING

PACBIO®

Several new high-quality human genome assemblies produce ethnicity-specific reference sequences and show how scientists can use this genetic information to improve precision medicine studies in Asian sub-populations. These projects demonstrate how long-read SMRT® Sequencing provides robust detection of polymorphic structural variants in clinically relevant gene coding regions and phases variants into haplotypes.

The current human reference genome assembly, GRCh38, was derived from sequencing the DNA of more than 50 ethnically diverse individuals[1]. As such, the haploid reference sequence represents an admixed background of contributing populations and switches from one ethnic haplotype to another at multiple places. This genome provides a useful international coordinate system for uniform annotation of genes, sequence read mapping, and variant calling. However, it does not sufficiently represent the ethnicity-specific haplotype diversity required for Asian sub-population precision medicine studies.

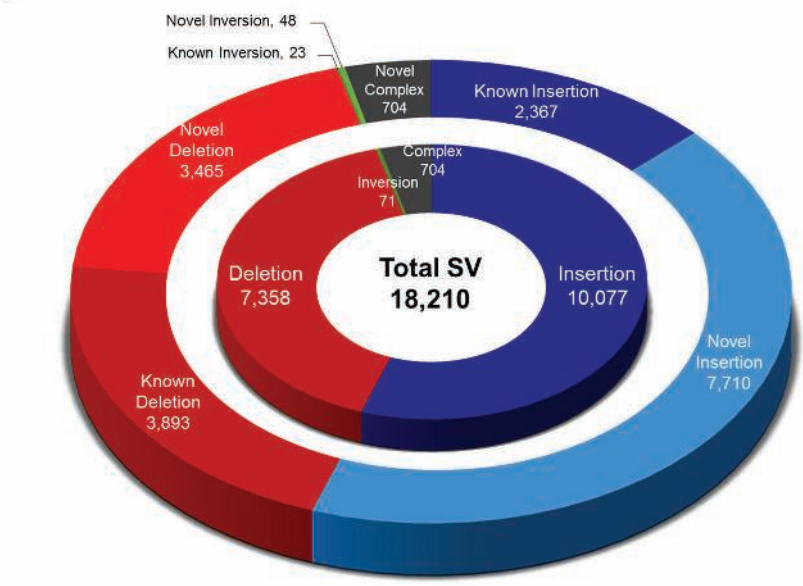


Ethnicity-Specific Asian Human Reference Genomes

While next-generation sequencing (NGS) has been enormously beneficial in the study of human genetic diversity, there are limits to current short-read methods that prevent accurate genotyping of many clinically important genes. One issue is a limitation in the size and type of genetic variants that are discoverable. The limited read lengths produced by these technologies often cannot resolve larger genetic variation types, such as structural variants, making novel allele discovery difficult. Another related issue is incomplete allele representation in the reference genome. Bias originating from an unmatched ethic reference sequence can lead to incorrect mapping and genotype calls for medically relevant genes.

Two important examples are found in the major histocompatibility complex (MHC) region harboring the human leukocyte antigen (HLA) genes and the cytochrome P450-2D6 (*CYP2D6*) gene. The enzyme encoded by the *CYP2D6* gene is responsible for metabolizing 25% of all commonly used drugs, and has more than 100 known allelic variations. These variations include deletions, duplications, and other structural rearrangements, as well as a highly homologous pseudogene. MHC proteins encoded by the HLA gene complex are responsible for regulation of the immune system. Allelic variation in these genes are known to contribute to multiple diseases including infectious disease, autoimmune disorders, and cancer.

To improve the accuracy of human genetic studies, multiple groups across Asia are now using SMRT Sequencing technology from PacBio to assemble high-quality, ethnicity-specific reference genomes that better represent haplotypes common in their regional populations. These assemblies provide the ability to discover novel gene alleles, find common and rare variants ranging from single nucleotide polymorphisms to larger structural variants, and resolve and genotype allele haplotypes within a population. Here, we look at examples from Korea, China, and Japan to demonstrate the value of this approach.



Number of SVs in a Korean human genome by direct comparison between AK1 assembly and GRCh37 International reference genome[2]

Korea

Scientists from Seoul National University, Macrogen, and other institutions produced a *de novo* genome assembly for a Korean individual, publishing their results in *Nature*[2]. The project primarily relied on SMRT Sequencing to generate the assembly, fully phase all chromosomes, and perform detailed analyses of structural variation (SV) influencing clinically significant gene regions.

Standard short-read sequencing approaches could not have accomplished this kind of high-quality genomic resource. "Simple alignment of short reads to a reference genome cannot be used to investigate the full range of structural variation and phased diploid architecture, which are important for precision medicine," the scientists reported. "By contrast, the single molecule real-time (SMRT) sequencing platform produces long reads that can resolve repetitive structures effectively."

The contiguity of the assembly allowed the team to delve deeply into structural variation, identifying more than 18,000 variants – nearly 12,000 of which had never been reported before. Almost half of insertions detected had significant variability in frequency across populations, while nearly 10 percent of them were specific to people of Asian descent.

After constructing separate assemblies for each haplotype to more accurately represent the diploid genome, the scientists examined haplotypes of the HLA genes and *CYP2D6* gene in detail. They found an allele-specific duplication of the *CYP2D6* gene that was clinically relevant. "This result demonstrates that *de novo* assembly-based phasing has advantages in resolving challenging hypervariable regions and could be used further for pharmacogenomics," the team reported.

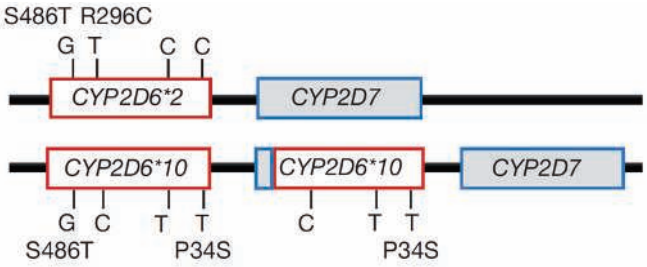
In a related study, scientists from Mount Sinai School of Medicine in New York City describe similar success for novel *CYP2D6* allele discovery by combining long-range PCR for target enrichment with PacBio long-read sequencing. This long-read targeted sequencing approach, published in *Human Mutation*[3], offers an economical way to genotype highly polymorphic regions across large study cohorts.

China

Scientists at Jinan University, the University of Southern California, and other organizations in China and the United States generated a *de novo* genome assembly and transcriptome of a Chinese individual using long-read SMRT Sequencing. The work was published in *Nature Communications*[4].

For the genome assembly, the team sequenced DNA from an anonymous Chinese individual (HX1), producing a 2.93 Gb genome with a contig N50 of 8.3 Mb. Consensus concordance for the assembly was 99.73%, matching the accuracy of the well-known NA12878 genome assembly. The HX1 *de novo* genome also revealed 12.8 Mb of novel reference sequence unique to the Chinese individual, further demonstrating the value of generating ethnicity-specific human reference genomes.

In an analysis of structural variants, the team found about 20,000 insertions and deletions, including 49 potentially functional variants that overlap a RefSeq exon. Half of all structural variants classified as short tandem repeats or mobile elements. Short repetitive structural variants such as simple sequence repeats (SSRs) and



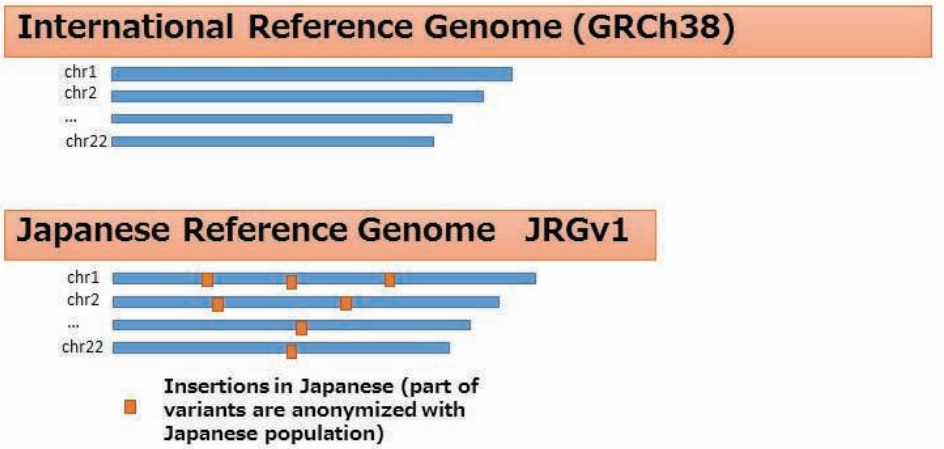
Phased haplotypes of the CYP2D6 and CYP2D7 alleles in the AK1 Korean human reference genome[2]

short tandem repeats (STRs) are known to cause many clinical disorders, including well-studied repeat expansion diseases such as fragile X syndrome, Huntington's disease, ataxias, and ALS. The ability to study these structural variants has been limited in current precision medicine studies using short-read NGS due to GC bias and read lengths too short to span pathogenic repeat expansions and other structural variants.

"The HX1 *de novo* genome also revealed 12.8 Mb of novel reference sequence unique to the Chinese individual, further demonstrating the value of generating ethnicity-specific human reference genomes."

Japan

At Tohoku University in Japan, the Tohoku Medical Megabank Organization (ToMMo) was established to help facilitate reconstruction after the devastating 2011 earthquake. As part of its mission, the group has produced a *de novo* Japanese reference genome (JRGv1) with SMRT Sequencing. The final assembly was publicly released in April 2016[5]. Scientists will use the information to begin cataloging structural variants in the Japanese population, gain support for future genome projects, and promote genome science in the country.



Comparison of international reference genome (GRCh38) and Japanese reference genome (JRGv1)

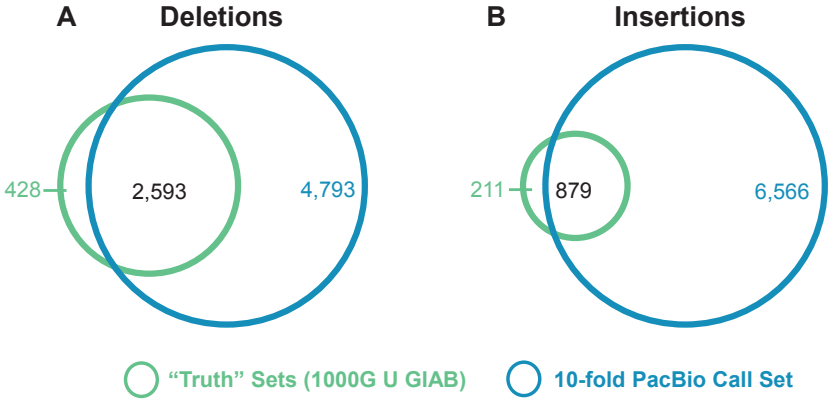
The scientists chose PacBio long-read sequencing for its ability to resolve structural variants. In a comparison with the GRCh38 reference, their genome assembly identified approximately 3,500 novel insertion sequence regions consisting of about 2.5 Mb.

This new reference will enhance ongoing precision medicine initiatives at ToMMo and other organizations. For example, one study using short-read sequencing to analyze genetic variation across 150,000 residents in the Miyagi and Iwate Prefectures found both rare and common single nucleotide variations but could not accurately identify structural variants in the participants' genomes. This was due in part to difficulty with aligning short sequencing reads used in the study (324 base pairs on average) to GRCh38. The new JRGv1 genome assembly will make it much easier to comprehensively catalog structural variation by providing ethnic-specific haplotypes to improve short-read alignments and SV genotype calling. The improved haplotype reference sequence will also allow scientists to expand on precision medicine programs with targeted capture techniques designed to find specific variants in alleles of clinical interest.

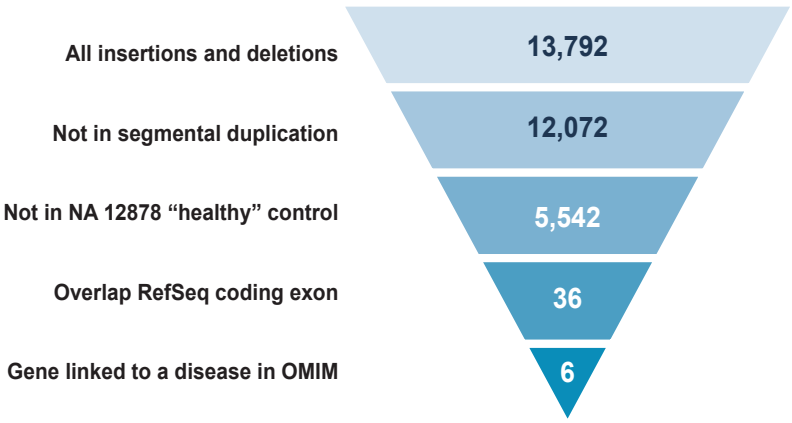
Low Coverage Long-read Whole Genome Sequencing

In other recent studies, scientists have reported success with low-coverage SMRT Sequencing of individual human genomes (typically about 10-fold). These re-sequencing studies do not produce a reference-grade *de novo* genome assembly, but reveal the majority of structural variation present in an individual genome, complete with allele-specific haplotype phasing.

By applying these methods to the NA12878 diploid human genome, scientists discovered 7,386 deletions (≥50 bp) and 7,445 insertions (≥50 bp), representing a five-fold improvement in sensitivity for structural variants previously found with short-read technologies[6].



NA12878 structural variant call set generated using low-coverage SMRT Sequencing on the Sequel™ System[6]



Pathogenic SV Discovery using Low Coverage Long Read WGS - Carney Complex Case Study[7]

In a real-world application of this method, scientists from Stanford University used the Sequel System to sequence an individual with an undiagnosed rare disease. Short-read sequencing had failed to provide any insights, but with low coverage, long-read whole genome sequencing the scientists discovered a novel pathogenic heterozygous deletion of about 2 kb in the first coding exon of the *PRKAR1A* gene. The structural variant was confirmed with Sanger sequencing and classified as causative for Carney complex, a rare Mendelian disease[7].

Conclusion

Through improved haplotype resolution and structural variant discovery, PacBio long-read sequencing is aiding in the discovery of novel, medically relevant alleles within Asian sub-populations. Emerging SMRT Sequencing applications, such as targeted sequencing and low coverage long-read whole genome sequencing for structural variation, are reducing study costs and increasing the addressable sample sizes. This data will help further resolve clinically important, ethnicity-specific alleles and structural variants that are less common, thereby accelerating the use of precision medicine in Asian sub-populations. As our understanding of structural variation and allelic diversity improves, precision medicine studies will yield more and more actionable findings.

References

1. Graves-Lindsay T, et al. "Reference Genomes Improvement" Project. *The Elizabeth H. and James S. McDonnell III Genome Institute at Washington University,* 2016. Web. 31 Jan. 2017.
2. Seo JS, et al. (2016) *De novo* assembly and phasing of a Korean human genome. *Nature*. 538(7624), 243-247.
3. Qiao W, et al. (2016) Long-read Single-Molecule Real-Time (SMRT) full gene sequencing of cytochrome P450-2D6 (*CYP2D6*). *Human Mutation*. 37(3), 315-323.
4. Shi L, et al. (2016) Long-read sequencing and *de novo* assembly of a Chinese genome. *Nature Communications*. 7, 12065.
5. Tohoku Medical Megabank Project. "Japanese Reference Genome: JRGv1." Tohoku University and Iwate Medical University, 25 Apr. 2016. Web. 31 Jan. 2017.
6. Wenger A, et al. "Identifying Structural Variants in NA12878 from Low-Fold Coverage Sequencing on the PacBio Sequel System." Web blog post. PacBio Blog. PacBio, 19 Oct. 2016.
7. Merker J, et al. (2016) Long-read whole genome sequencing identifies causal structural variation in a Mendelian disease. *bioRxiv*. doi:10.1101/090985.

PN: CS119-021317