Joint Calling and PacBio SMRT Sequencing for Indel and Structural Variant Detection in Populations

Aaron M. Wenger, Armin Töpfer, Yuan Li, Luke Hickey, Jonas Korlach
PacBio, 1305 O'Brien Drive, Menlo Park, CA 94025

Structural Variants (SVs) and Indels

Most of the base pairs that differ between two human genomes are in indels 1-49 base pairs and in SVs, differences ≥50 base pairs. Short-read sequencing has limited sensitivity for indels and SVs, while PacBio SMRT Sequencing comprehensively detects variants of all sizes.

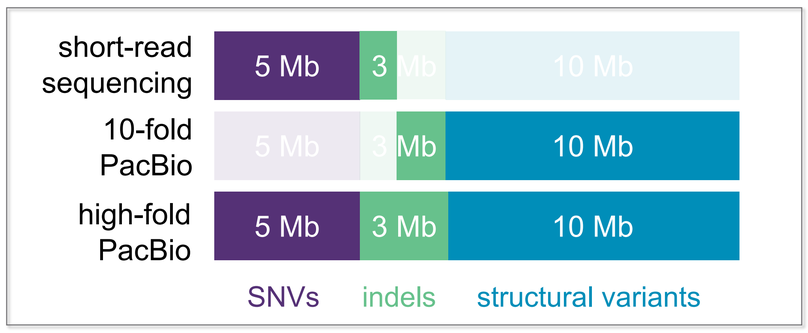

Figure 1. Variation in a typical human genome compared to the GRCh38 reference.[1]

Indels and SVs identified first by PacBio SMRT Sequencing contribute to human disease and evolution.

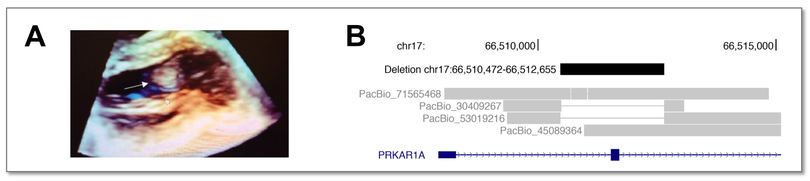

Figure 2. PacBio SMRT Sequencing identifies the causative mutation in Carney complex[2]. Short-read whole genome sequencing failed to identify the causative mutation in an individual with **(A)** cardiac myxomata characteristic of Carney complex. **(B)** PacBio identified the causative mutation, a heterozygous 2.2 kb deletion in *PRKAR1A*.

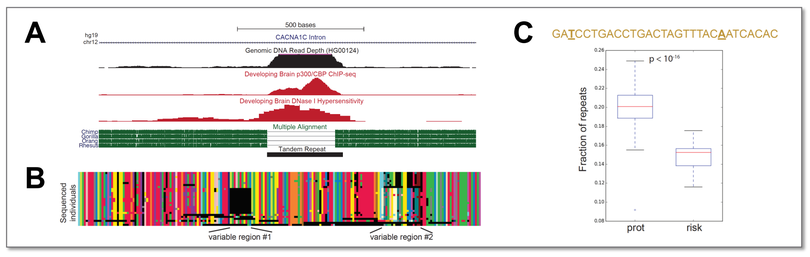

Figure 3. PacBio SMRT Sequencing of a tandem repeat linked to bipolar disorder and schizophrenia[3]. A 100 kb interval intronic to *CACNA1C* has been linked to psychiatric disorders. The region contains **(A)** a human specific 30-mer tandem repeat with **(B)** variation in length and sequence content in the human population. **(C)** Sequence variants in the repeat impact the enhancer function of the repeat array and are strongly associated with disease risk.

Current Population-Scale Variant Databases Lack SVs

The 1000 Genomes Project and ExAC databases comprehensively catalog small variants but largely miss the indels and SVs to which short-read sequencing is blind.

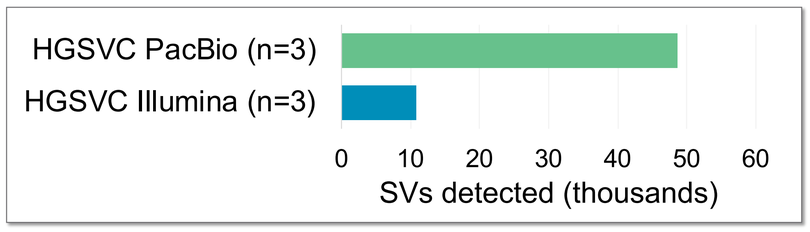

Figure 4. The Human Genome SV Consortium (HGSVC) identified 4.5× more SVs with PacBio sequencing than with Illumina sequencing of the same three samples.[1]

pbsv: Joint Variant Caller for Structural Variants and PacBio Reads

To support use of large variants in disease and association studies, it is necessary to perform population-scale surveys with a technology effective at detecting indels and SVs, such as PacBio SMRT Sequencing. The pbsv variant caller aims to provide a workflow for these studies that is similar to the ExAC GATK workflow.

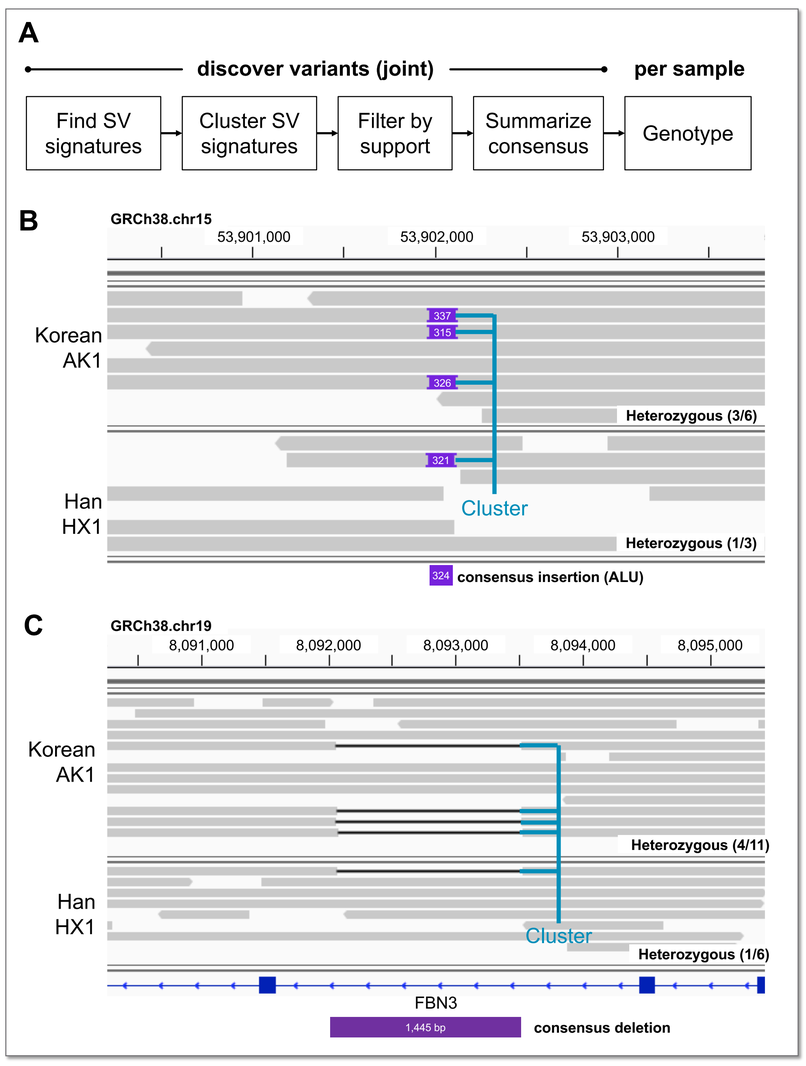

Figure 5. The pbsv workflow (A) separates variant discovery, which is performed jointly over all samples, and genotyping, which is per sample. Joint discovery uses reads from one sample to support variant calls in another, which increases sensitivity, particularly with the low coverage that is typical in population studies. **(B)** A heterozygous ALU insertion is called from one read in HX1[4] using support from AK1[5]. **(C)** A similar deletion locus where AK1 rescues HX1.

Cohort of 20 Human Samples

Public PacBio datasets for 20 human samples were gathered for joint analysis. The coverage ranges from 4-fold to 82-fold.

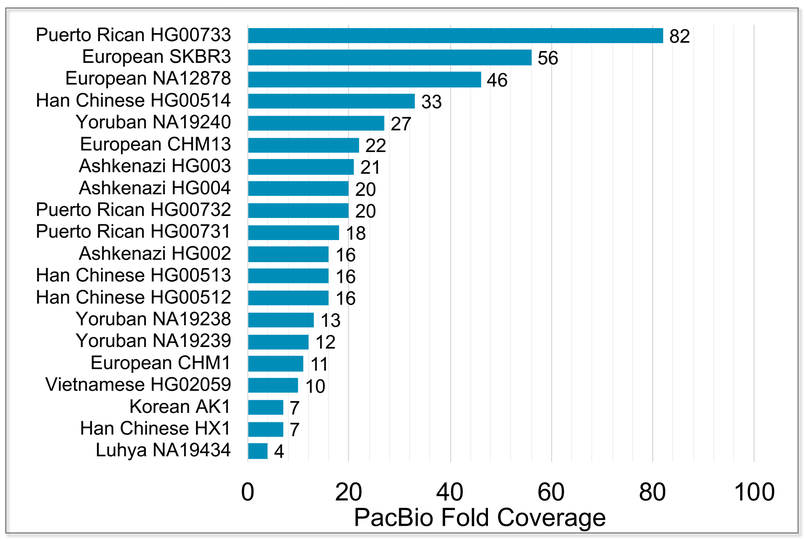

Figure 6. PacBio fold coverage in 20 human samples.

Variant Calls in Cohort of 20 Human Samples

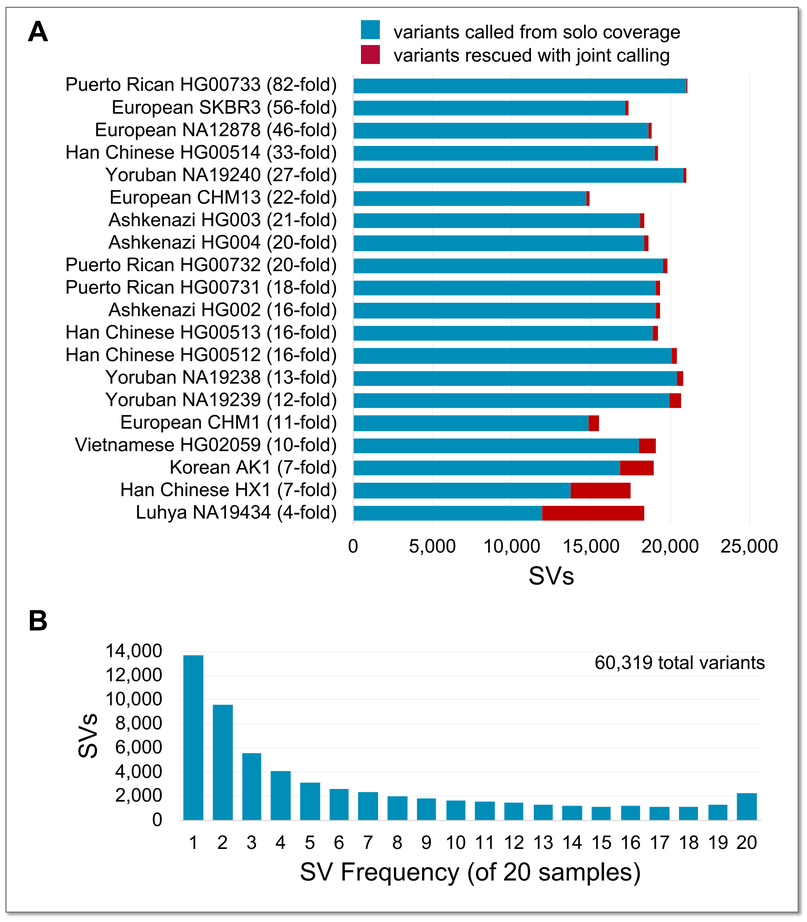

Figure 7. SVs in human cohort (A) Around 20,000 SVs are detected per sample. Joint calling boosts low-coverage samples. **(B)** The modal variant is private but most are shared, with the average frequency being 6.3 of 20 samples.

Active and Future Projects

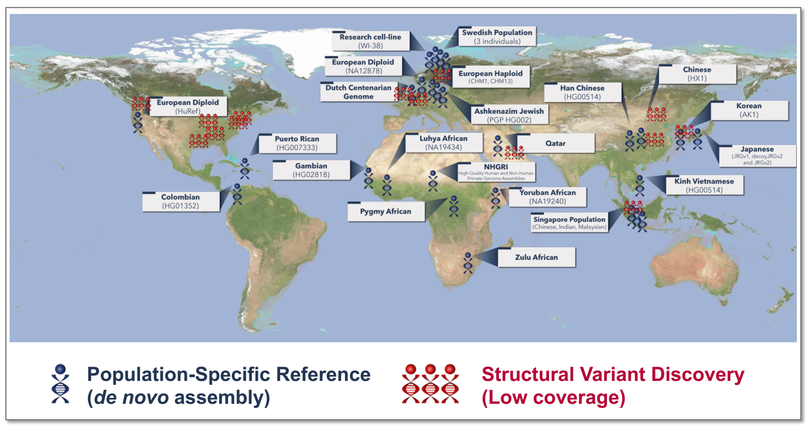

Figure 8. Global sequencing projects using PacBio.

Conclusions

- Variant databases built with short-read sequencing miss most SVs.
- pbsv provides a scalable, effective structural variant caller for human cohorts.
- Active projects are using PacBio SMRT Sequencing to build databases of SVs to support disease studies.

References

1. Chaisson MJ, et al. (2017). Multi-platform discovery of haplotype-resolved structural variation in human genomes. *bioRxiv*. doi:10.1101/193144.
2. Merker JD, et al. (2017). Long-read genome sequencing identifies causal structural variation in a Mendelian disease. *Genet Med*. 20(1):159-163.
3. Song JHT, et al. (2018). Characterization of a Human-Specific Tandem Repeat Associated with Bipolar Disorder and Schizophrenia. *Am J Hum Genet*. 103(3):421-430.
4. Shi L, et al. (2016). Long-read sequencing and de novo assembly of a Chinese genome. *Nat Commun*. 7:12065.
5. Seo JS, et al. (2016). De novo assembly and phasing of a Korean human genome. *Nature*. 538(7624):243-247

Thank you to David Scherer, Kristin Robertshaw, and Pamela Bentley Mills for poster production support.