Population-Scale Discovery of Structural Variants with PacBio SMRT Sequencing

Ralph Vogelsang, Aaron Wenger, Luke Hickey, Armin Töpfer, Yuan Li, Jonas Korlach, Mike Hunkapiller
PacBio, 1305 O'Brien Drive, Menlo Park, CA 94025

Introduction

Structural variants (SVs) – genomic differences ≥50 base pairs – are few by count compared to single nucleotide variants (SNVs) and indels but include most of the base pairs that differ between two humans.

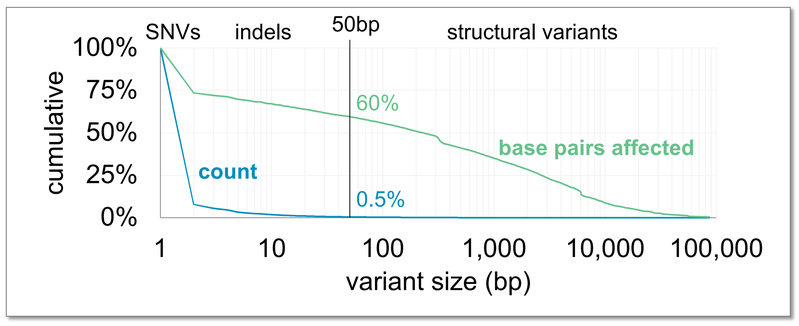
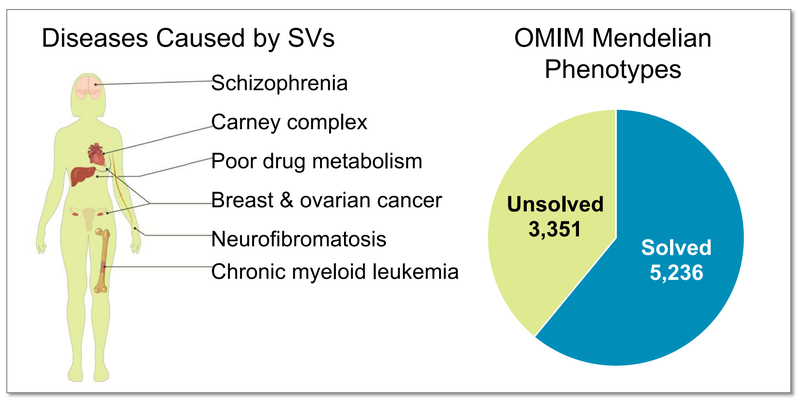


Figure 1. Frequency and size of variants in a human genome – structural variant, indel, and single nucleotide variant calls in HG00733 against GRCh38 from multiple sequencing technologies.[1] 60% of the total base pairs come from the 0.5% of variants that are structural (≥50 bp).

SVs cause rare and common diseases, and contribute to human traits and evolution. SVs are suspected to explain many still unsolved diseases.



Most human SVs are detected only by PacBio long read SMRT Sequencing.

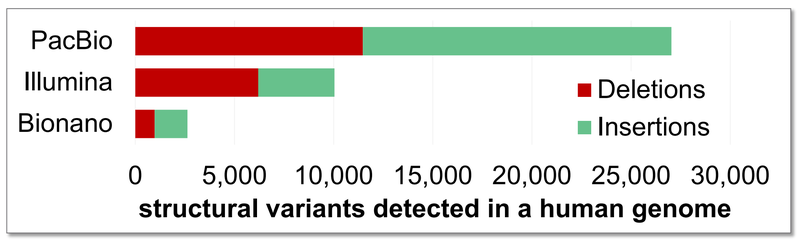


Figure 2. Sensitivity for structural variants in a human genome by technology. Most SVs in HG00733 were detected only by PacBio because many SVs involve repeats that are too large to be spanned by short reads but are too small to be detected by optical mapping.[1]

The 1000 Genomes Project[2,3] applied low-coverage (4-fold) short-read sequencing of many individuals to identify common small variants. This was effective, but studies that use short-read sequencing are blind to most SVs.

Table 1 | Variants discovered by project, type, population and novelty

a Summary of project data including combined exon populations

	Low coverage			
Statistic	CEU	YRI	CHB+JPT	Total
Samples	60	59	60	179
Total raw bases (Gb)	1,402	874	596	2,872
Total mapped bases (Gb)	817	596	468	1,881
Mean mapped depth (×)	4.62	3.42	2.65	3.56
Bases accessed (% of genome)	2.43 Gb (86%)	2.39 Gb (85%)	2.41 Gb (85%)	2.42 Gb (86.0%)
No. of SNPs (% novel)	7,943,827 (33%)	10,938,130 (47%)	6,273,441 (28%)	14,894,361 (54%)
Mean variant SNP sites per individual	2,918,623	3,335,795	2,810,573	3,019,909
No. of indels (% novel)	728,075 (39%)	941,567 (52%)	666,639 (39%)	1,330,158 (57%)

Goal: Population-Scale SV Discovery

Existing population-scale databases of small variants must be extended with PacBio SMRT sequencing to add sensitivity for structural variants. Initial efforts to do so are using the same low-coverage study design as the 1000 Genomes Project.

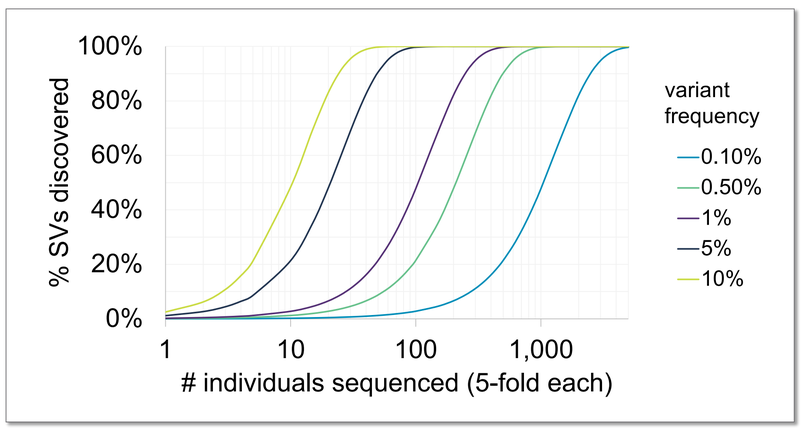


Figure 3. Power to discover structural variants with joint calling, by cohort size. In an idealized population with no substructure, the power to detect SVs of different population frequencies depends on the number of individuals sequenced. Common variant discovery saturates with few individuals. Rarer variants require larger cohorts. (pacb.com/calculator-structural-variation)

Methods

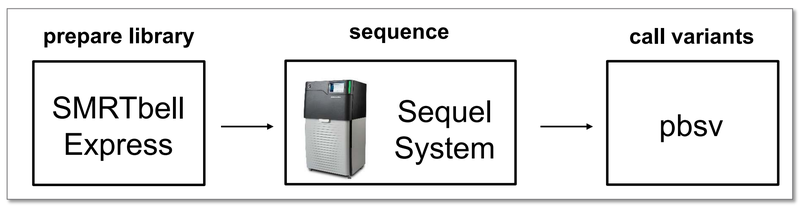


Figure 4. Workflow to detect SVs from PacBio long reads.

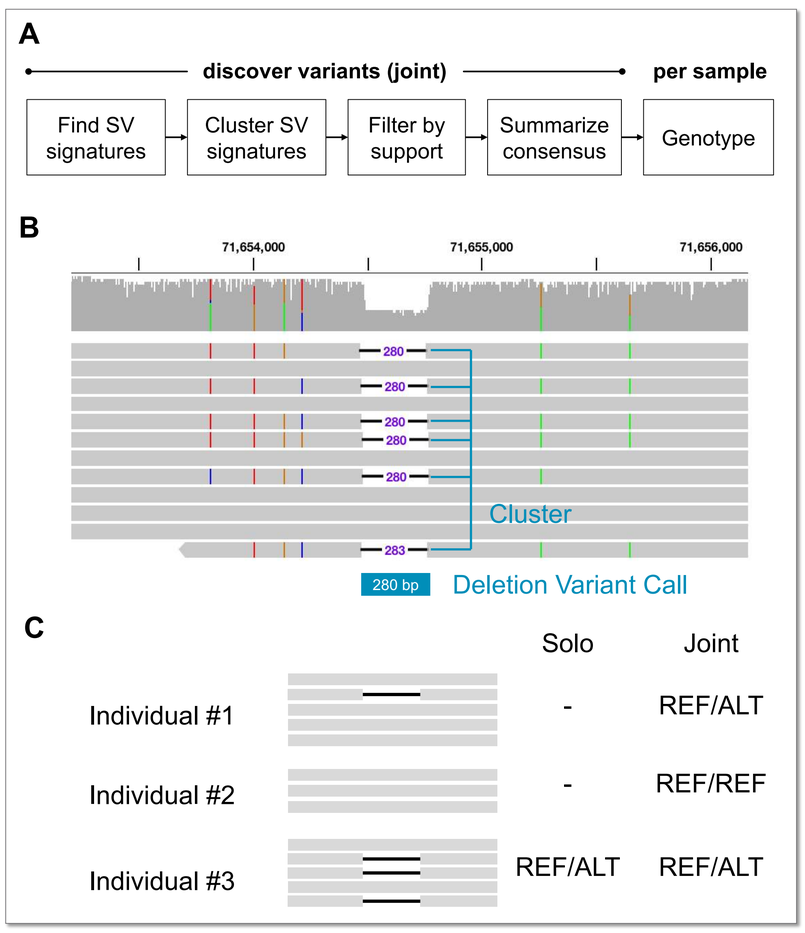
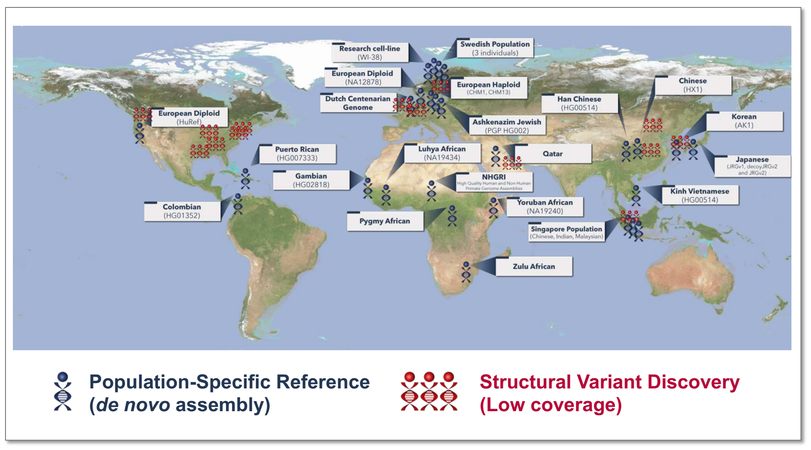
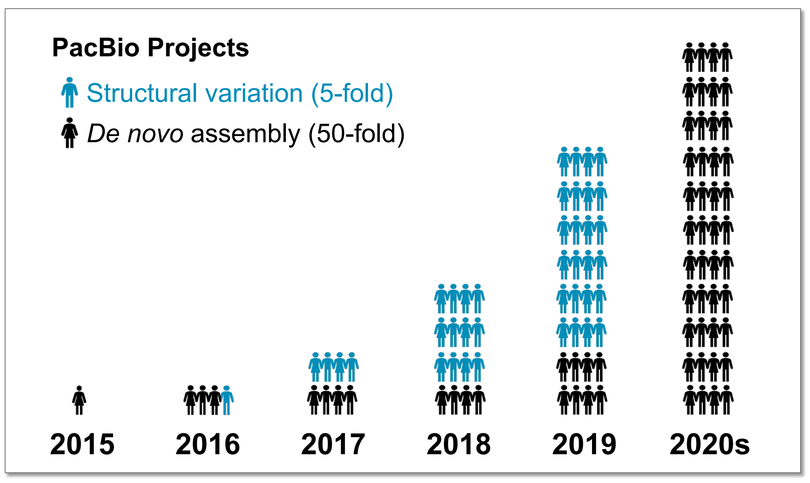


Figure 5. pbsv joint calling identifies structural variants simultaneously in multiple samples. (A) Joint calling pools reads from samples to discover variants then genotypes per sample, increasing sensitivity at low coverage. (B) To call structural variants, pbsv identifies large deletion or insertion signatures, clusters nearby signatures, and summarizes into a call. (C) Joint calling uses support in the population to call variants in each sample.

Active and Future Projects



Numerous efforts across the globe are applying PacBio long-read sequencing to build population-specific reference genomes and to characterize structural variation, which is largely missed by short-read sequencing. These **databases will empower rare and common disease studies for SVs**, just as ExAC has done for small variants.



Conclusion

- Variant databases built with short-read sequencing miss most SVs, and thus most of the variant base pairs, in the human population.
- Progress in library prep, throughput, and variant calling support the application of low-coverage PacBio sequencing for population-scale SV discovery.
- Active projects are building databases of common SVs to support studies of rare and common diseases.

References

1. Chaisson MJ, et al. (2017). Multi-platform discovery of haplotype-resolved structural variation in human genomes. *bioRxiv*. doi:10.1101/193144.
2. 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature*. 526(7571):68-74.
3. 1000 Genomes Project Consortium. (2010). A map of human genome variation from population-scale sequencing. *Nature*. 467(7319):1061-73.
4. Telenti A, et al. (2016). Deep sequencing of 10,000 human genomes. *PNAS*. 113(42):11901-6.

Thank you to David Scherer for poster production support.