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.

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

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 80% of the total base pairs come from the 0.5% of variants that are structural (≥50 bp).

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 ability to detect SVs of different population frequencies depends on the number of individuals sequenced. Common variant discovery saturates with few individuals. Rarer variant require larger cohorts. (pacb.com/calculator-structural-variation)

Methods

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

Conclusions

- 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