Improving the reference with a diversity panel of sequence-resolved structural variation

Peter A. Audano¹,², Arvis Sulovari¹,³, Tina A. Graves-Lindsay², Stuart Cantilleari¹, Melanie Sorensen¹, AnneMarie E. Welch¹, Max L. Dougherty¹, Bradley J. Nelson¹, Ankeeta Shah¹, Susan K. Dutcher³, Wesley C. Warren², Vincent Magrini²,³, Sean D. McGrath⁴, Yang I. Li⁵,⁶,⁷, Richard K. Wilson⁴,⁵,⁶, and Evan E. Eichler¹,⁸

Motivation

Structural variants (SVs) are insertions, deletions, and inversions 50 bp or larger. Short-read sequence data is abundant, but it cannot identify most SVs or sequence-resolve them. Long reads (10-20 kbp, PacBio) are capable of resolving 90% of the genome, but a diversity panel of samples was not available until recently.

We sequenced 13 genomes with PacBio and obtained sequence-resolved SVs. With these data, we can better understand SV biology, correct the reference, and make a diverse set of variation accessible to short-read technology.

Variant Discovery

SVs were discovered in our 13 genomes along with with AK1² and KH1³, covering African, Asian, European, American, and S. Asian populations. In the nonredundant set of 99,604 SVs, 2,238 (1.6 Mbp) were shared among all 15 samples indicating errors or extreme minor alleles in GRCh38.

Gene and Regulatory Impact

200 bp missing in FOXO6 shifts the RefSeq annotation out of frame. gnomAD4 contains three common mutations past the insertion (dashed line), but when corrected, two become synonymous variants and one falls into the UTR.

Credit: Max Dougherty

STR and VNTR Enrichment

SVs are enriched in the last 5 Mbp of chromosome arms. This effect is driven by VNTRs, which exhibit a 4.8-fold enrichment (Wilcoxon p = 2.9 × 10⁻⁵). Male meiotic recombination (MR), double-strand breaks (DSB), and de novo mutations (DNM) were most significantly correlated suggesting potential mechanisms of VNTR formation.

Credit: Arvis Sulovari

Building a Patched Reference

By adding sequence-resolved SVs to the reference in alternate contigs, we can improve read mapping over SVs and accurately genotype known variants in short-read samples with a machine learning model. With our patched reference, we called variants in SV insertions and found 2,228 SNPs and indels (GQ 20+) per sample that would be missed by standard pipelines.

Conclusions and Future Work

• A population-level view of SVs supports corrections across the genome
• SVs impact genes, regulatory elements, and their annotations
• SVs are non-randomly distributed over the genome
• VNTRs correlate with double-strand breaks and meiotic recombination
• Sequence-resolution enables genotyping with short reads

We plan to:
• Sequence 50 additional genomes (PacBio, 10X)
• Apply phasing (Phased-SV⁵) to improve SVs and contigs
• Build a pan-genome reference (vg⁶)

References