Structural variation in All of Us analyzed with long-read sequencing at a scale.

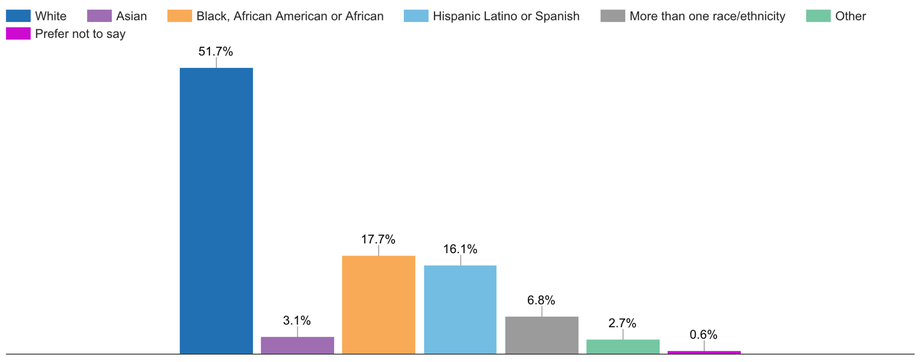
Michal Izydorczyk, Ming Zhu, Medhat Mahmoud, Xinchang Zheng, Ziad M. Khan, Ginger Metcalf, Donna Muzny, Richard Gibbs, Fritz J. Sedlazeck

Abstract

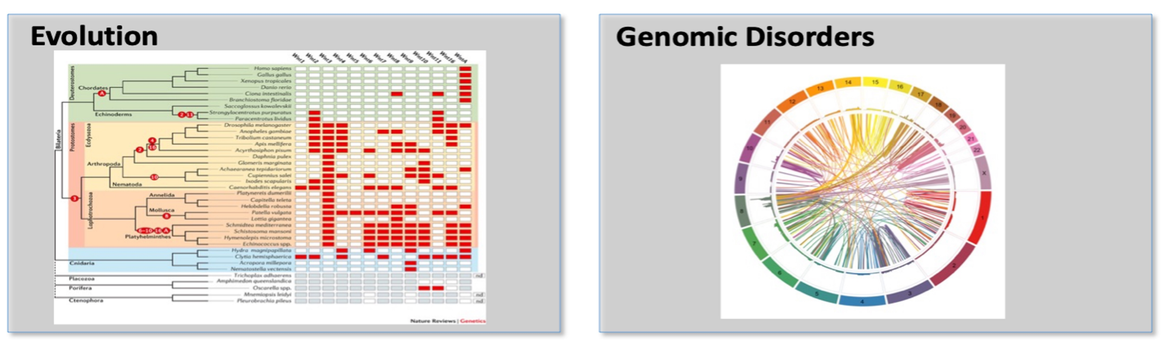
Novel long-read technologies establish scientists to analyze previously inaccessible regions of the Human genome, including 193 medically relevant genes and regions such as centromeres and telomeres. Combined with nationwide initiatives, such as the All of Us research program such technologies can help establish meaningful variants in the human genome on a population level-scale and link them to specific phenotypes/diseases. Novel long-read technologies require different approaches of analysis, therefore we have developed Sniffles structural variant caller to detect variants both on germline and population scale, as well as somatic/mosaic level. Moreover, we are interested in underlying causes of the structural variations and their consequences, including insertional/deleterious mutations or chromosomal rearrangements, which have been observed to be associated with transposable elements. Therefore, our further goal is to characterize transposable elements in the dataset and investigate possible association of specific elements with observed structural variants, which can possibly be linked to alterations within medically relevant genes and other regions associated with numerous diseases.

Study

- **Problem** - bias in the reference genome towards white people samples. Various genomic databases are also white-biased.
- AoU creates most comprehensive and diverse biomedical database in USA.
- HGSC is primarily interested in Hispanic population, sequenced using ONT long-reads.



- We focus on detection of **Structural Variants,** particularly those affecting medically relevant genes, which can impact various biological processes**:**



Methods

Sniffles2 for structural variant analysis.

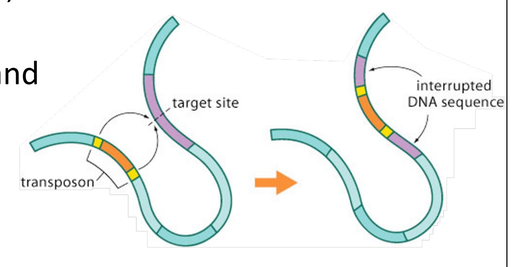
- 2-50x faster than other callers.
- High accuracy.
- Detects large germline SVs.
- Fully genotyped family & population level call sets.
- Somatic/mosaic variants.

Clair3 for single nucleotide variant analysis.

- Fast SNV detection for ONT long-reads.
- Utilizes pileup data and deep neural networks for superior accuracy.

RepeatMasker for transposable element detection.

- Popular tool for annotation of repeated elements within DNA sequences.
- Uses *nhmmr* - highly sensitive detection engine, based on hidden Markov models.
- Integrated into HPC workflow establishes fast and scalable annotation of millions of transposable element families defined in Dfam database.



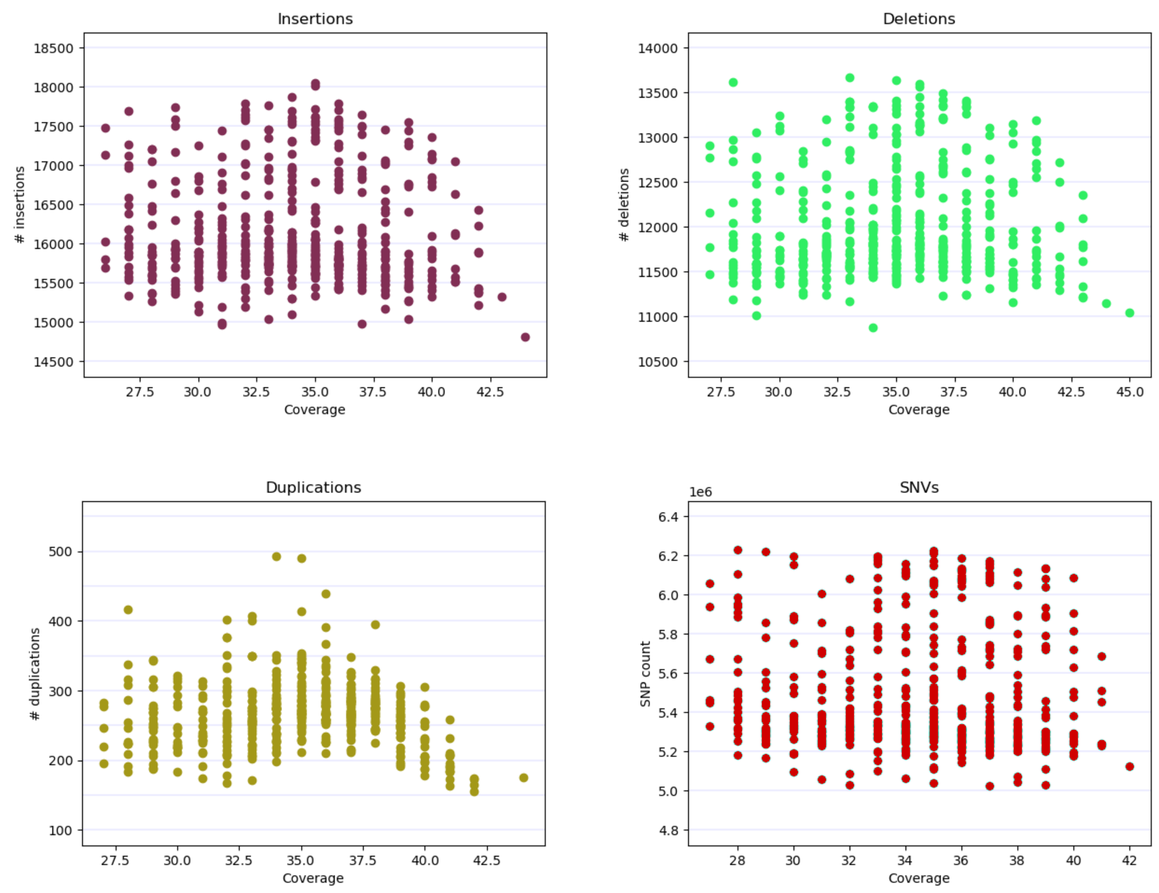
De-novo assembly.

- To avoid reference genome bias and gain insight into difficult regions raw read data is extracted from alignment files and assembled *de-novo*.

Mapping based analysis

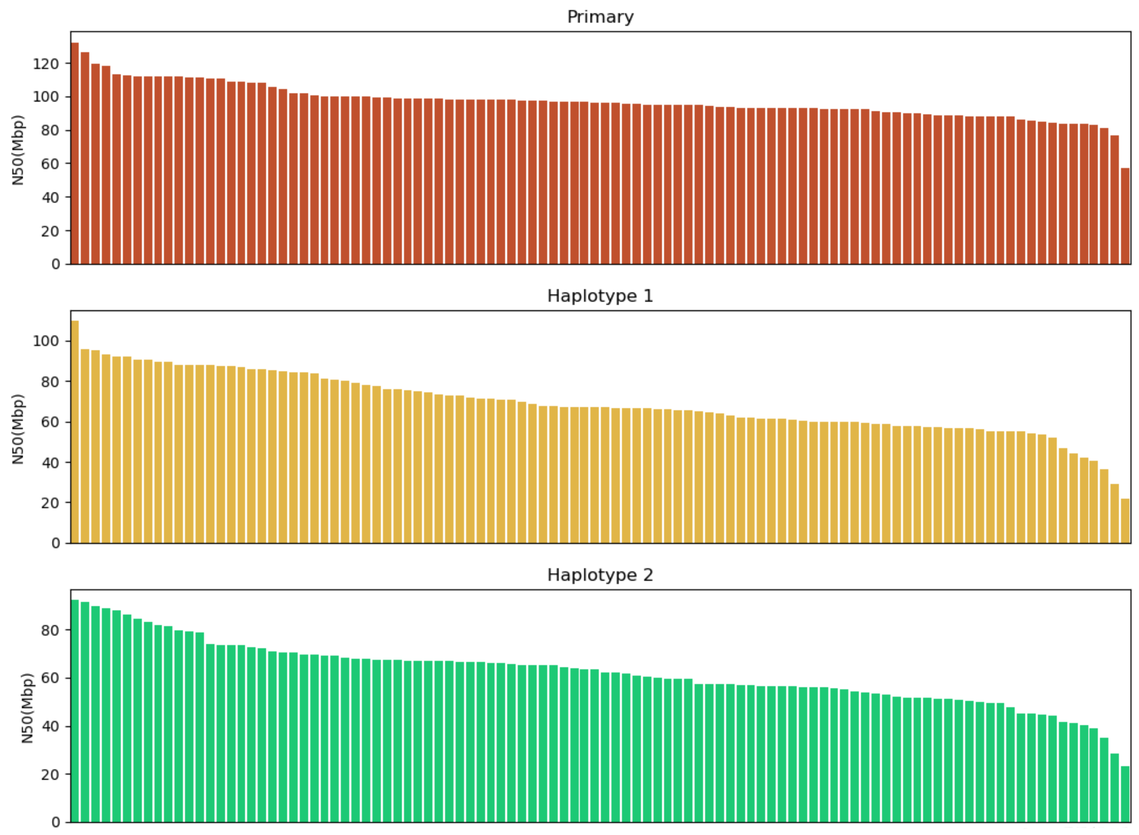
Currently **624 ONT long-read samples** are sequenced and curated at BCM.

The established workflow, utilizing popular bioinformatics tools uncovers structural variants and single nucleotide variants across the tested samples at >30 coverage.



Assembly based analysis

So far, we have created 102 haplotype-resolved assemblies, with mean N50 of 97Mbp, 69Mbp and 62Mbp for primary assembly and haplotypes accordingly.



Further work

- Methylation, STR, CNV analysis.
- Detect transposable elements within mapped SVs.
- Associate the results with metadata available from the AoU project.
- Construct graph genomes based on *de-novo* assemblies.

Acknowledgments & References

All of Us research program is a. US government initiative, run by the National Institutes of Health.

NIH National Institutes of Health
All of Us Research Program

References:

1. Smolka et al 2022, *bioRxiv*; DOI: 10.1101/2022.04.04.487055
2. Mahmoud et al 2021, *Genome Biol*; DOI: 10.1186/s13059-021-02486-w
3. Wagner et al 2022, *Nat Biotechnol*; DOI: 10.1038/s41587-021-01158-1
4. Denny JC et al 2019, *N Engl J Med*; DOI: 10.1056/NEJMsr1809937
5. Zheng et al 2022, *Nat Comput Sci*; DOI: 10.1038/s43588-022-00387-x
6. Hubley R et al 2016, *Nucleic Acids Res*; DOI: 10.1093/nar/gkv1272
7. Cheng, H et al 2021, *Nat Methods*; DOI: 10.1038/s41592-020-01056-5