Refining the Platinum Pedigree truth set with genome assemblies and Aardvark benchmarking

Abstract # 2051F

Zev Kronenberg[1], James M. Holt[1], Tom Mokveld[1], Egor Dolzhenko[1], Christopher T. Saunders[1], Michael A. Eberle[1], The Platinum Pedigree Consortium[2]

1. PacBio, 1305 O'Brien Drive, Menlo Park, CA 94025; 2. https://github.com/Platinum-Pedigree-Consortium

Accurate benchmarking is essential for advancing sequencing technologies

- High-quality genomic standards are essential for developing bioinformatics software and validating clinical pipelines. Conservative, short-read–based benchmarks slow progress by excluding the most challenging regions of the human genome.
- Through integration of CEPH 1463 genome assemblies and variant calls, we have expanded the small-variant truth set by 10 to 15% and produced the first NA12878 structural-variant and tandem-repeat truth sets.
- To support the growing use of assemblies in benchmarking, we developed *Aardvark*, a tool for assessing sequence-level accuracy. It reports precision, recall, and F1 scores in both variant- and sequence-based modes, evaluating how many bases are correct to measure regional accuracy.

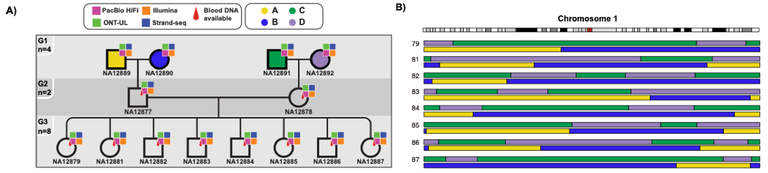

Figure 1. Comprehensive haplotype map of CEPH 1463. A) The first three generations of the CEPH 1463 pedigree used to validate genetic variation through haplotype transmission. Beyond the ten family members shown, sequencing has been extended to a fourth generation, yielding 23 fully consented samples. All living individuals were sequenced from blood to avoid cell line artifacts. **B)** Ideogram of founder haplotypes (yellow, blue, green, purple) tracked using long-read single nucleotide variants (SNVs).

Integrating multiple technologies to discover and validate genetic variation

- Small variants were integrated across multiple callers and technologies (**Figure 2**).
- Genome assembly was used to adjudicate conflicts between callers in pedigree consistent assembled regions.
- Structural variants were called from the long-read technologies and genome assembly.
- We developed a new tool, *Aardvark*, to simplify and streamline variant merging.

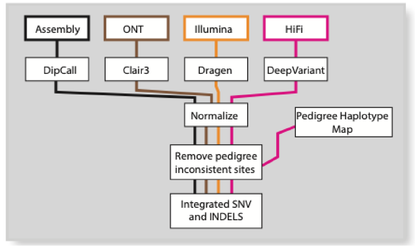

Figure 2. The small variant truth set workflow. Variants were filtered based on the pedigree haplotype map (**Figure 1B**).

https://github.com/Platinum-Pedigree-Consortium/Platinum-Pedigree-Inheritance

Three truth sets capturing 26.3 Mb of NA12878 variation

The Platinum Pedigree benchmark dataset (v1.2, **Figure 3**) introduces the first structural variant and tandem repeat truth sets for NA12878. These resources have been tested with standard benchmarking tools, including *Hap.py* and *TruVari* (**Figure 4**), and are now publicly available through Amazon Open Data and PacBio.

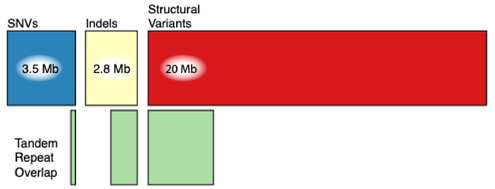

Figure 3. Total length of genetic variation in the NA12878 truth sets. Tandem repeats overlap multiple categories and are displayed to scale relative to other classes of variation.

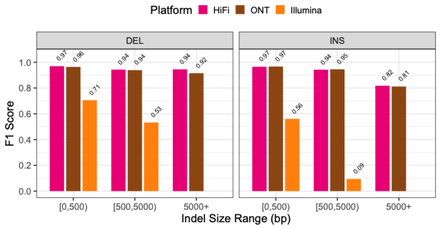

Figure 4. Benchmarking NA12878 structural variants with TruVari. The bar plot shows F1 scores across size ranges, separated into insertions and deletions. Overall F1 scores are 0.9611 for HiFi (*sawfish),* 0.9599 for ONT (*Sniffles2*), and 0.582 for Illumina (Dragen SVs).

Improving benchmarking tools with sequence accuracy measures

- Standard measures of variant calling accuracy rely on variant counts (variant counts: VC) which can lead to errors(**Figure 5**).
- With long-read technologies we can sequence resolve complex variation, which is difficult to measure in a count framework.
- We have built a new tool, *Aardvark*, to measure sequence level accuracy (basepair: BP) (**Figure 6**).

Aardvark variant comparison tool

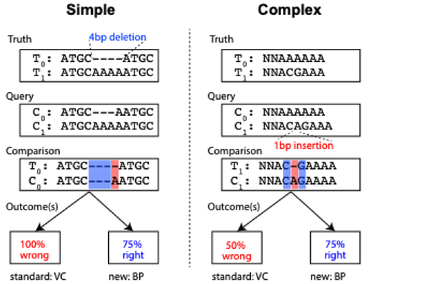

Figure 5. Comparing variant counts and basepair-based measures of accuracy. In the simple example, variant counting ignores partial calls, while in the complex example it underestimates correct calls due to counting complexity.

Basepair benchmarking with Aardvark

- Different technologies and alignment parameters can produce very different variant calls even if the underlying sequence is identical, which is naturally accounted for with basepair scoring.
- Basepair analysis is not impacted by how the number of variants (e.g. an MNP vs two SNPs count differently).
- The ground truth is increasingly becoming a full genome assembly, and we should be comparing against the haplotypes.

	HiFi	ONT	Illumina
SNV F1	**99.7**	99.5	99.5
Indel F1	**97.5**	93.7	96.0
Total F1	**99.1**	97.3	98.4

Table 1. Overall small variant basepair 35x coverage measured with *Aardvark*.

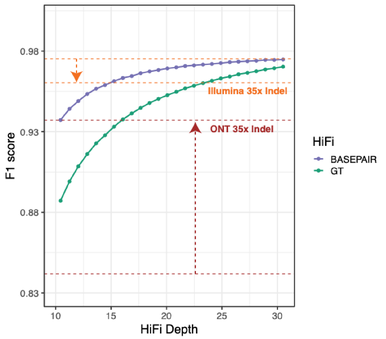

Figure 6. Basepair accuracy (F1) for Indels comparing HiFi, ONT, and Illumina. The arrows indicate the relative change of Indel accuracy when moving from genotype counts to basepair counts. HiFi data are titrated across depths, genotype counts and basepair scoring schemes. Short reads often miss larger indels, leading to lower basepair scores.

Conclusion

- The latest release of the Platinum Pedigree truth set spans the full spectrum of variation (SNVs, indels, tandem repeats, and structural variants) and is further improved by the integration of genome assemblies.
- As the field shifts toward sequence-resolved haplotypes, new methods for merging and scoring are needed to capture complex variation in difficult regions. *Aardvark* fills this gap with its basepair-based accuracy measure.

References

1. David Porubsky et. al (2025).Human De novo mutation rates from a four generation Pedigree reference. *Nature*
2. Zev Kronenberg et. al (2025). The Platinum Pedigree: A long-read benchmark for genetic variants reference. *Nature Methods*

Check out these other posters!: 4109W (Jonathan Belyeu), 4050F (Matt Holt), and 4104F (Xiao Chen)

Acknowledgements

The authors would like to thank the Amazon Open Data program for hosting our project.