Building the spectrum of ground truth genetic variation in a four-generation 28-member CEPH family.

Zev N Kronenberg[1], Katherine M Munson[2], David Porubsky[2], Cillian Nolan[1], William J Rowell[1], Brent S Pedersen[3], Cairbre Fanslow[1], Primo Baybayan[1], Nidhi Koundinya[2], William Harvey[2], Kendra Hoekzema[2], Jordan Knuth[2], Gage Garcia[2], Tom Mokveld[1], Egor Dolzhenko[1], Scott Watkins[3], Deborah W Neklason[3], Aaron R Quinlan[3], Lynn B Jorde[3], Evan E Eichler[2] and Michael A Eberle[1]

[1]PacBio, 1305 O'Brien Drive, Menlo Park, CA 94025
[2]University of Washington, Seattle WA, 98195
[3]University of Utah, Salt Lake City, UT, 84114

Largest long-read kinship dataset

Highly accurate long-read sequencing characterizes the full spectrum of genetic variation across the genome, but variant calling software is still catching up to the sequencing technologies. To develop long-read methods for calling difficult variants and variants in the genomic dark regions, it is important to have a comprehensive ground truth dataset for benchmarking. Until now, most benchmarking datasets were primarily built using short-read technologies that are limited to the easily characterized parts of the genome. We are developing a comprehensive truth set by utilizing the power of genetic inheritance within a four-generation family (CEPH pedigree 1463 plus a newly collected fourth generation [**Figure 1**]) characterized with multiple sequencing technologies (PacBio, ONT, Illumina and Strand-seq) from blood-derived DNA. Large kinship pedigrees provide greater power to establish inheritance patterns when compared to trios; allowing us to adjudicate variant calls across the entire genome and not just in well-behaved regions.

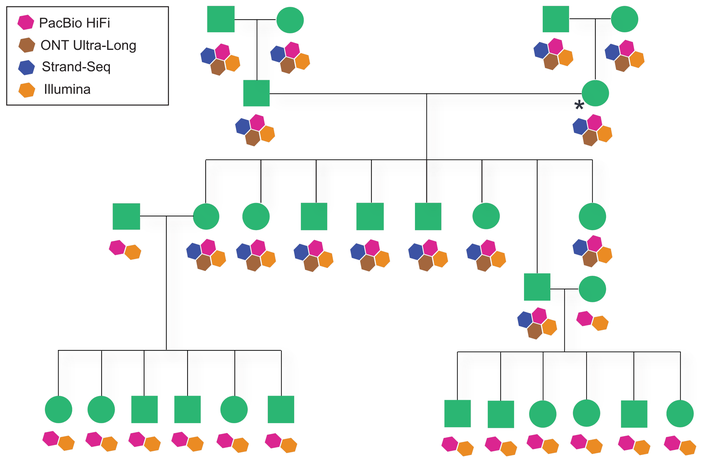

Figure 1. The four generation CEPH 1463 pedigree characterized, in this study, with multiple sequencing technologies. The second-generation female denoted with an asterisks is NA12878, the most sequenced human genome.

Variant calling and pedigree filtering

- An ensemble of variant callers was used to maximize sensitivity of variant detection, summarized in **Figure 2**.
- Patterns of inheritance (**Figure 3**) are used to identify high confidence variants in the second and third generation of the CEPH 1463 pedigree.

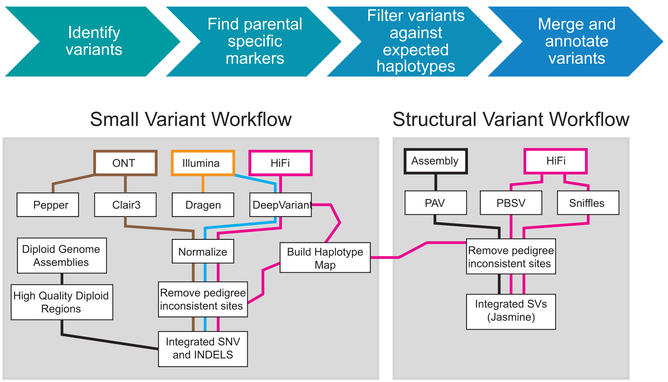

Figure 2. Variant calling, filtering, and phasing workflows for small and large variants.

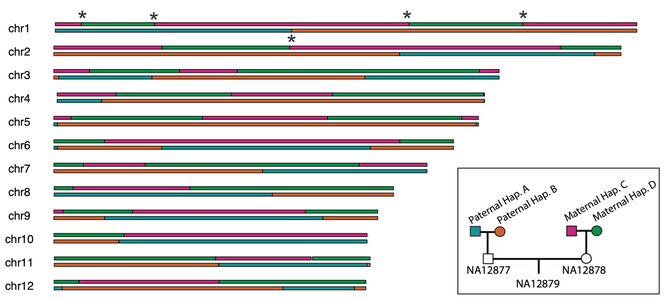

Figure 3. High-resolution haplotype map used for phasing and filtering variant calls in the third generation. Asterisk highlights the recombination sites for chromosome one, in a single individual (NA12879).

Characterizing ground truth variation

- 66-83% of variant calls, in each call set, are pedigree consistent, **Figure 4A.**
- High concordance across call sets after pedigree filtering, **Figure 4B.**
- Our high-confidence regions cover 91.4% of the genome, 8.5% higher than GIAB for NA12878.

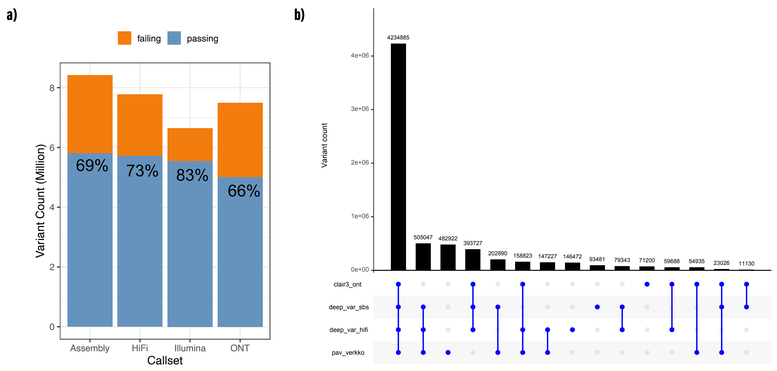

Figure 4. Counts of pedigree consistent/inconsistent variants broken down by call set. A) Small variant counts in the second/third generation of the pedigree. The fraction of pedigree filtering passing and failing are depicted with blue and orange, respectively. **B)** The intersections of the technologies/callers used in this study. Most small variant calls are shared across all variant callers (~4 million variants).

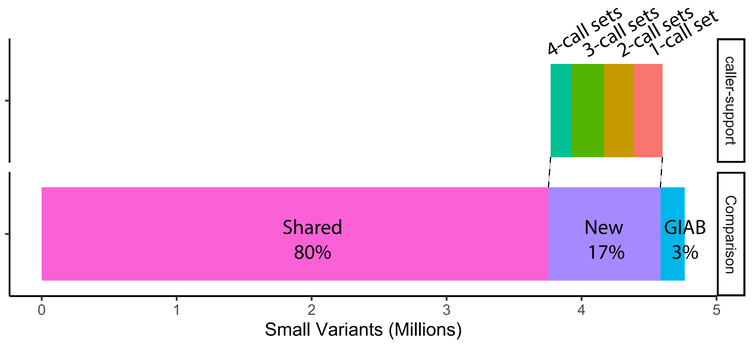

Figure 5. Comparison between our small variant truth set and GIAB 4.2 NA12878. The bottom bar chart shows the intersection and unique calls for GIAB and our truth set. The top bar chart annotates the number of call sets that agree for each new variant call.

Compared to GIAB, our truth set contains 17% more small variants **(Figure 5)**

73% of variants unique to our NA12878 truth set are supported by more than one calling method

Our truth set has 174,462 and 473,690 more indels and SNVs compared to GIAB

Benchmarking variant callers against the extended NA12878 truth dataset

- Using the Hap.py framework we have characterized the recall of three variant call sets (**Table 1**).
- The HiFi DeepVariant call set had the highest recall for single nucleotide variants (SNVs).
- Depth titrations show that 10-fold HiFi coverage results in better SNP recall compared to 30-fold Illumina data (**Figure 6**).

Technology/Software	Recall SNV	Recall INDEL
HiFi DeepVariant	0.985	0.952
Illumina DeepVariant	0.935	0.914
ONT Clair3	0.969	0.733

Table 1. Recall statistics for NA12878 using our new truth dataset.

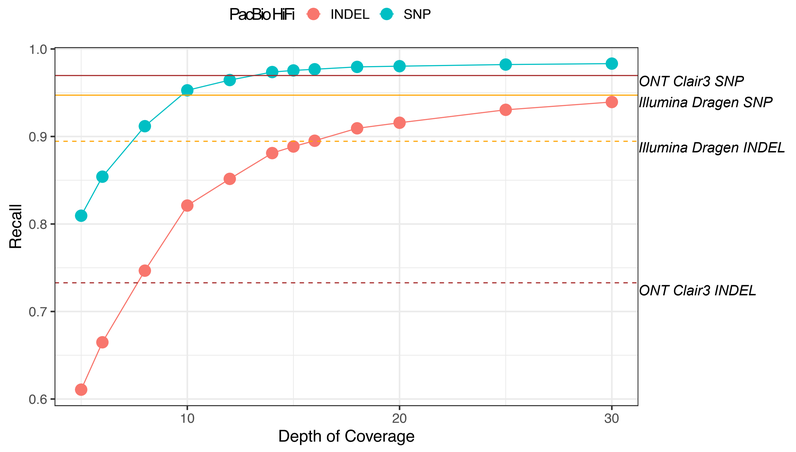

Figure 6. HiFi DeepVariant recall across titrated depth. SNPs are colored cyan and INDELs are colored salmon. The horizontal lines represent the max recall for Illumina and ONT data.

Conclusions

- Over 4 Million variants in NA12878 have been validated using patterns of Mendelian inheritance across different technologies and variant callers.
- The new NA12878 truth set extends the number of validated variants by 17% and characterized 8.5% more of the genome compared to earlier truth sets.
- Extending the benchmarking truth set beyond the "easy" regions of the genome will highlight areas where variant callers could improve.
- Low coverage (10-fold) HiFi data provides better SNP recall compared to ~30-fold Illumina coverage.

References

- Majidian, S., Agustinho, D.P., Chin, CS. *et al.* Genomic variant benchmark: if you cannot measure it, you cannot improve it. *Genome Biol* **24**, 221 (2023). https://doi.org/10.1186/s13059-023-03061-1
- Eberle MA, *et al*. A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. *Genome Res*. 2017;27(1):157-164. https://doi:10.1101/gr.210500.116