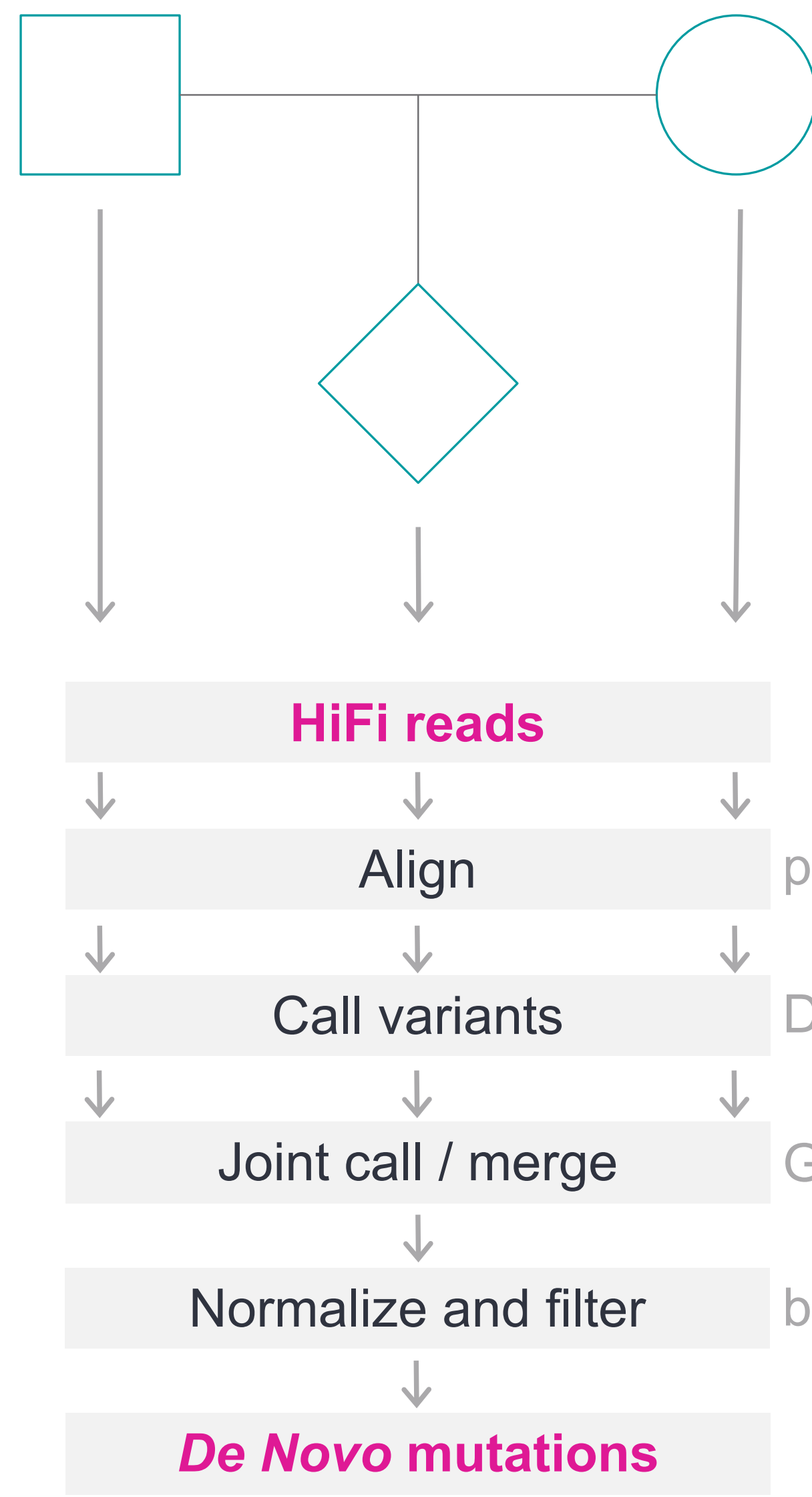
De novo variant detection with HiFi reads

Juniper A. Lake, William J. Rowell, Michael A. Eberle
PacBio, 1305 O’Brien Drive, Menlo Park, CA 94025

Methods for detecting *de novo* mutations

Unencumbered by generations of selective pressure, *de novo* mutations (DNMs) are more likely to be deleterious than inherited variants and account for much of the variation driving rare disease. HiFi reads are useful for detecting DNMs because of their accuracy and length. To quantify the number of *de novo* single nucleotide variants (SNVs), small indels (<50 bp), and SVs, we analyzed six trios using HiFi sequence data derived from either cell lines (2 trios) or blood (4 trios).



HiFi reads
Align — pbmm2
Call variants — DeepVariant, pbsv
Joint call / merge — GLnexus, pbsv
Normalize and filter — bcftools
De Novo mutations

PacBio human WGS pipeline

Figure 1. **Pipeline overview.** All members of each trio were sequenced with PacBio HiFi reads to ~30× depth.

- Trios 1–4 sequenced from blood
- Trios 5–6 sequenced from cell lines (HG002 and HG005, respectively)
- All samples sequenced to ~30× coverage and down-sampled for coverage comparisons
- Calls made with CHM13v2.0 were lifted over to GRCh38 for comparison

	SNV	Indel	SV
No missing genotypes	5,635,203 5,636,651	1,303,998 1,306,787	30,142 30,212
Mendelian inconsistency	31,139 27,219	23,355 17,159	57 80
Genotype quality ≥20	**799** **185**	**497** **409**	NA
Child allele balance ≥0.2	NA	NA	**20** **56**

Table 1. **Average variants retained after each filter for cell line trios (top) and blood sample trios (bottom).** To capture only high-quality DNMs, we filtered variants that (1) had a missing genotype for any member of the trio, (2) showed mendelian consistency, (3) possessed a genotype quality score less than 20 in any sample (SNVs or indels) or had an allele balance less than 0.2 in the child (SVs).

Reference genomes GRCh38 and CHM13v2.0 detect distinct DNMs

On average, 55% of SNVs, 31% of indels, and 16% of SVs detected using GRCh38 were also detected with T2T-CHM13v2.0.

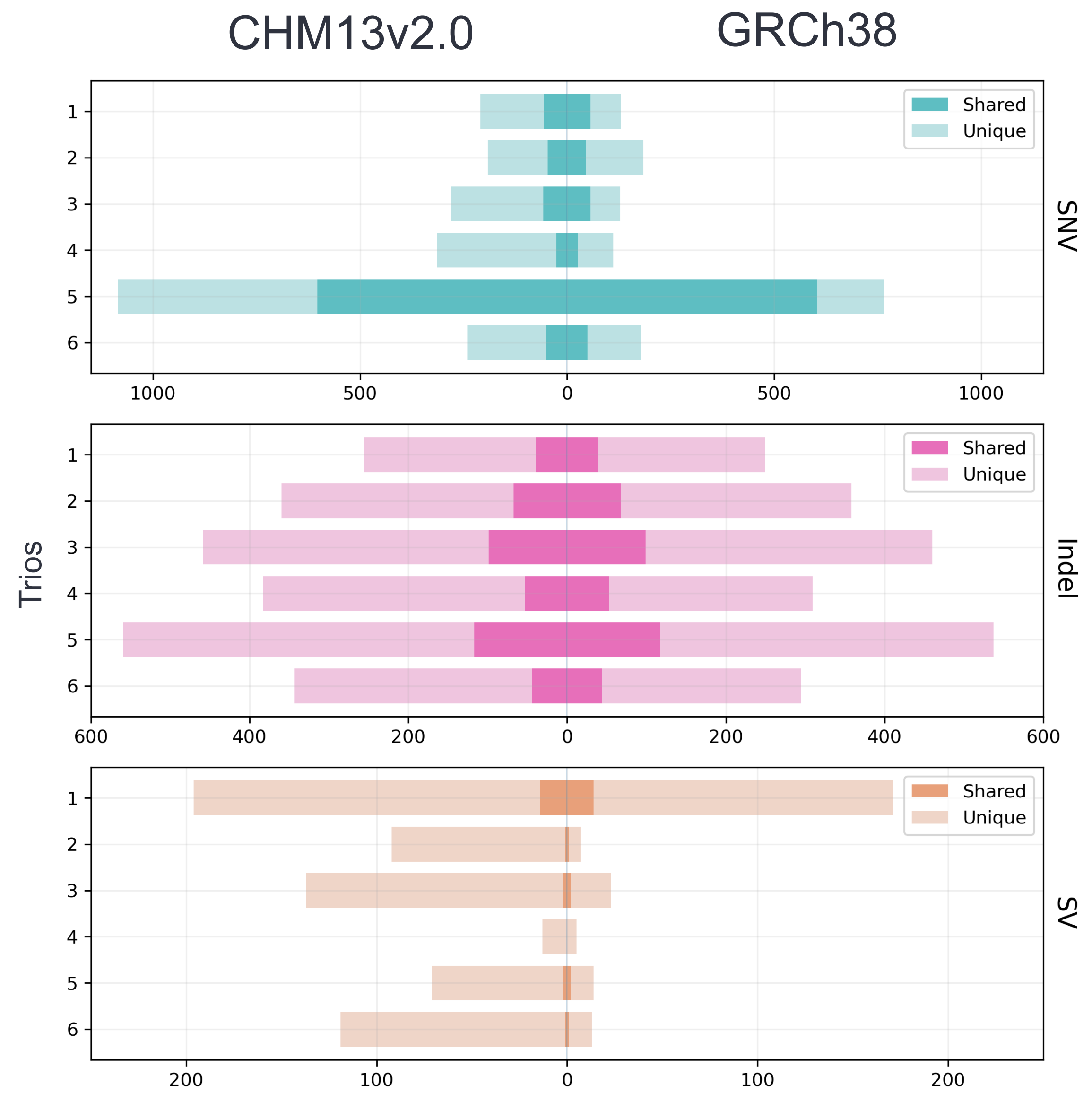


Figure 2. **Overlap of DNMs detected using GRCh38 vs CHM13v2.0.** Choice of reference genome affects what DNMs are detected, especially for SVs.

More DNMs were detected with CHM13 than with GRCh38:

- 55% more SNVs
- 7% more indels
- 468% more SVs

Lower coverage produces many spurious DNMs

Reduced coverage (~10× vs. ~30×) in all samples or just parents produces many false positives (coverage dropout in parents) and false negatives (coverage dropout in child).

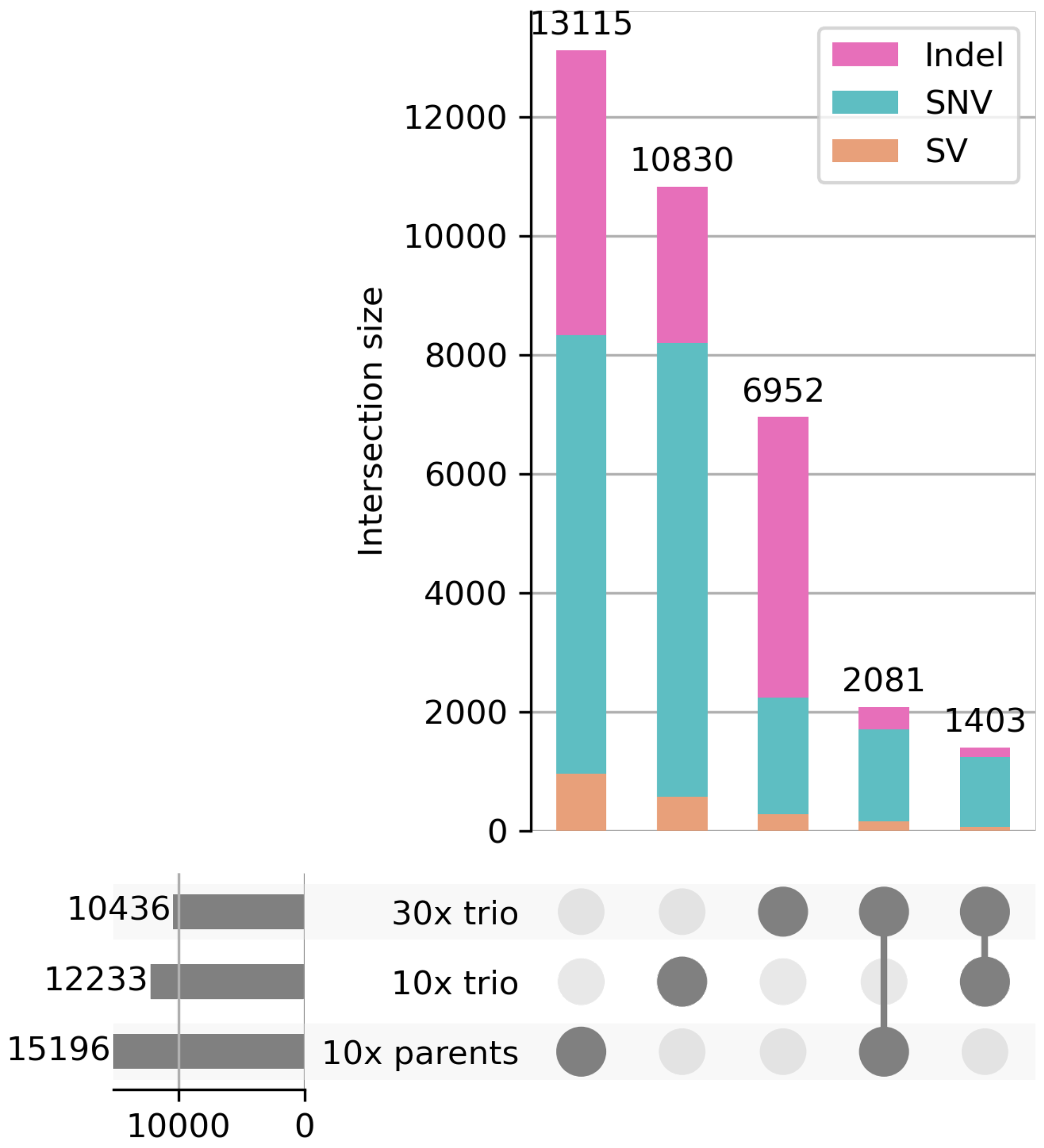


Figure 3. **Intersection of DNMs detected in 30× trios, 10× trios, and trios with 10× parents and 30× child.** Counts reflect all trios combined. Rows show DNMs detected in each set and columns show intersections between sets.

Characterizing DNMs by type and reference annotation

Most SNVs were non-coding and outside of repetitive regions. In contrast, most indels and structural variants overlapped at least 50% with tandem repeats, homopolymers, or segmental duplications. Insertions outnumbered deletions across variant sizes, with 1 bp insertions being the most common.

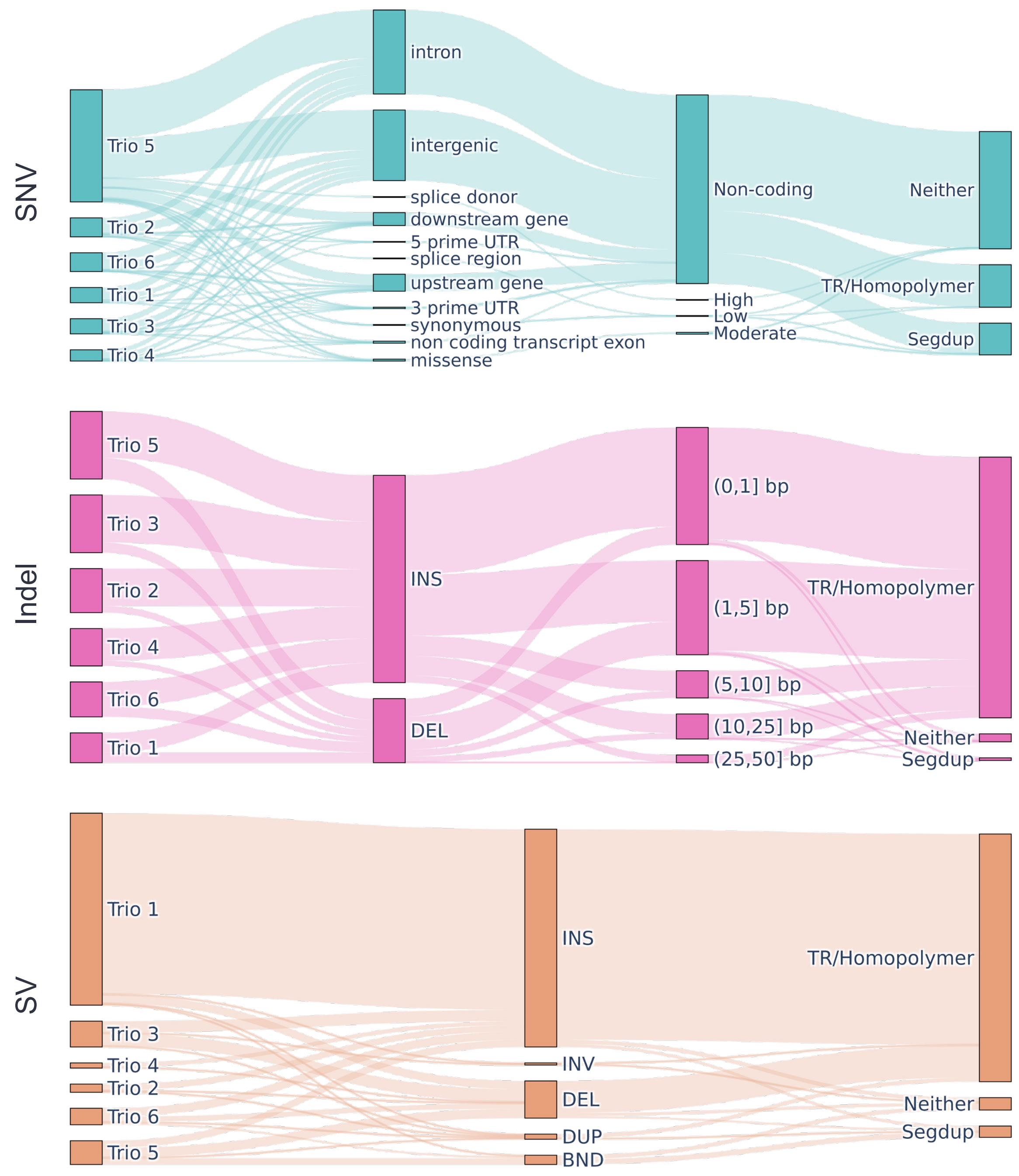


Figure 4. **Breakdown of DNMs by cohort, type, and reference annotation.** INS=insertion, DEL=deletion, BND=breakend, INV=inversion, DUP=duplication, TR=tandem repeat, Segdup=segmental duplication. Variants were classified as TR/Homopolymer or Segdup if they overlapped at least 50% with reference genome regions annotated as such. Overlap with TR/Homopolymer regions was considered before overlap with segmental duplications.

Conclusions

- In primary blood samples, we detected an average of 650 DNMs, most of which were insertions of 5 bp or less
- Most indels and structural variants are found in tandem repeats or homopolymers, where sequencing errors are more likely, and alignment is more difficult
- Genome reference has a substantial impact on the DNMs detected, especially for SVs
- Reduced coverage in parents (a common practice in rare disease studies) increases false positive and false negative DNMs

Acknowledgements

The authors would like to thank Christine Lambert, Primo Baybayan, and Shreyasee Chakraborty for sequencing these samples.