The American Journal of Human Genetics, Volume *109*

Supplemental information

Familial long-read sequencing increases yield of *de novo* mutations

Michelle D. Noyes, William T. Harvey, David Porubsky, Arvis Sulovari, Ruiyang Li, Nicholas R. Rose, Peter A. Audano, Katherine M. Munson, Alexandra P. Lewis, Kendra Hoekzema, Tuomo Mantere, Tina A. Graves-Lindsay, Ashley D. Sanders, Sara Goodwin, Melissa Kramer, Younes Mokrab, Michael C. Zody, Alexander Hoischen, Jan O. Korbel, W. Richard McCombie, and Evan E. Eichler

Supplemental Figures

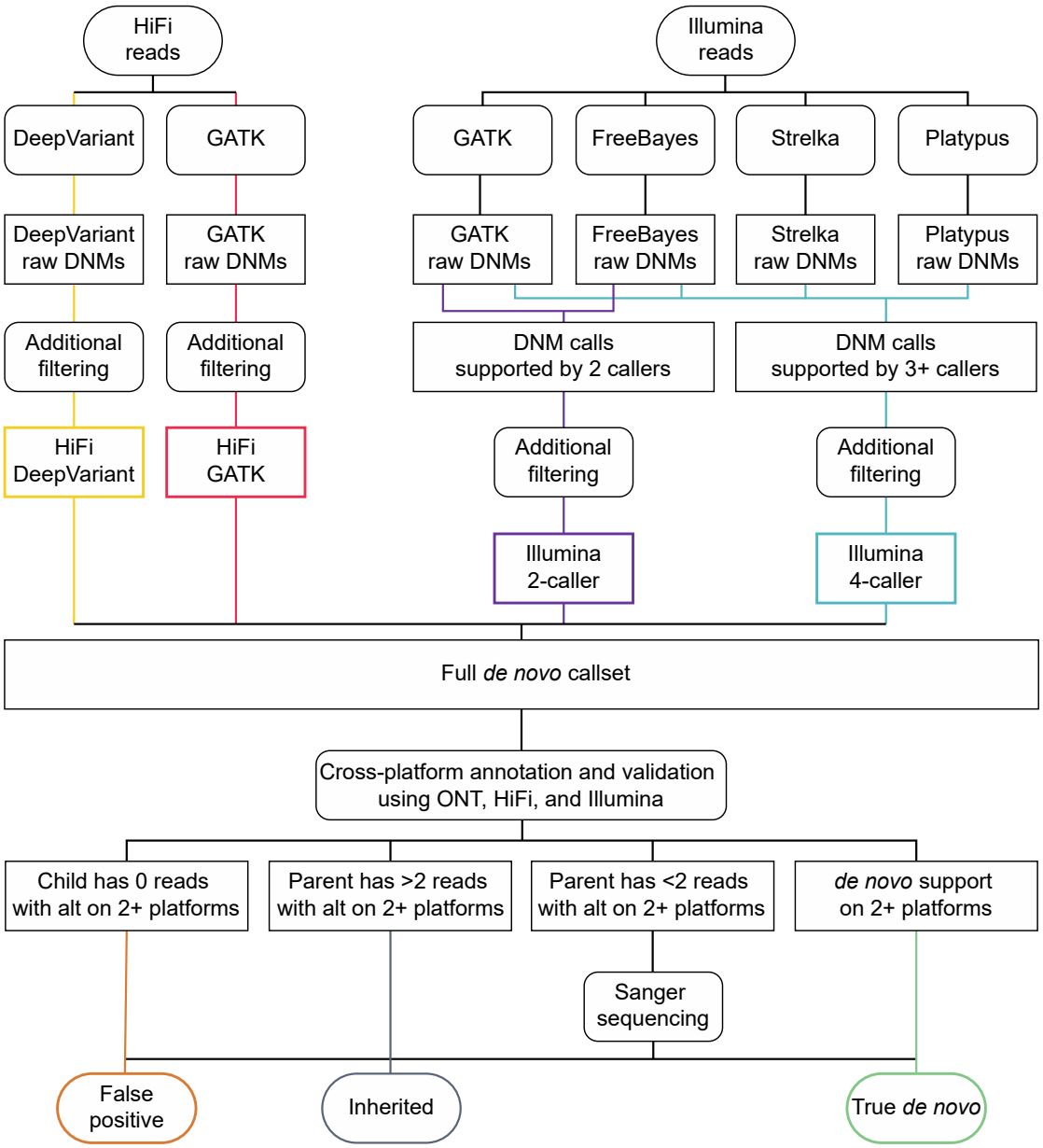


Figure S1. Pipeline for *de novo* SNV and small (<20 bp) indel identification. PacBio HiFi and Illumina reads were used for DNM discovery. Both technologies were used in addition to Oxford Nanopore Technologies (ONT) sequencing data for validation of *de novo* candidate sites.

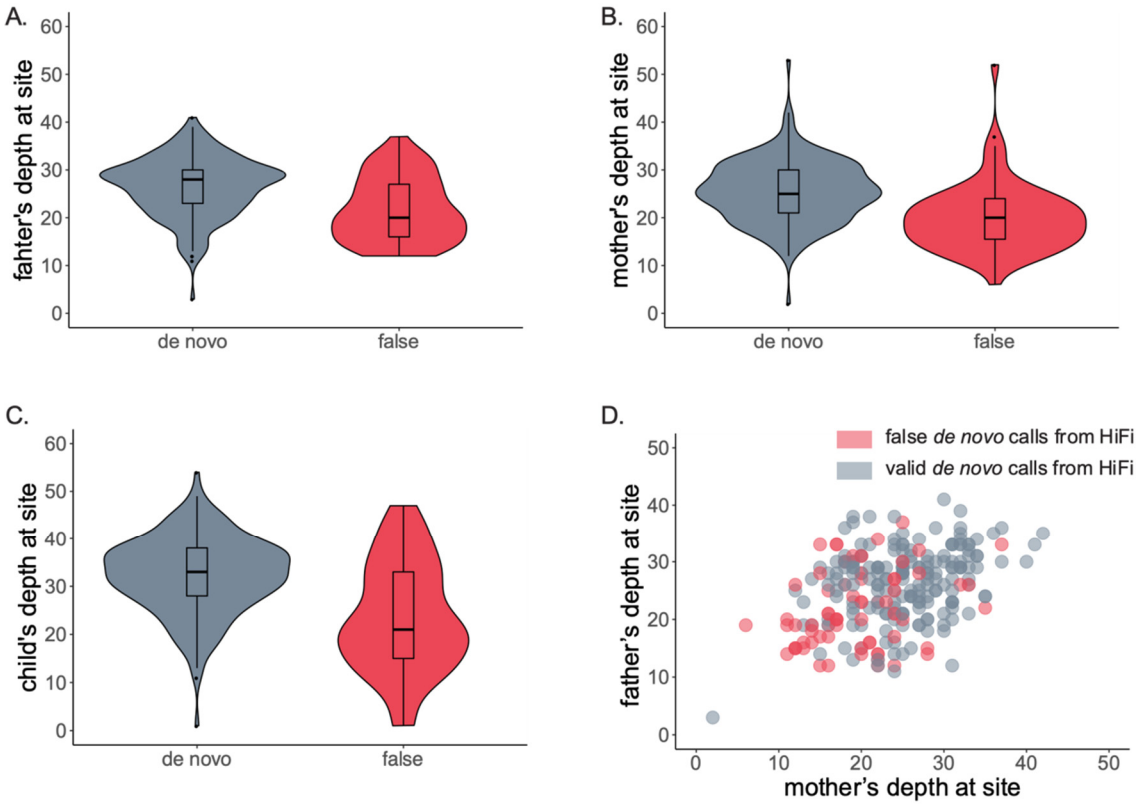


Figure S2. Read depth comparison between true and false calls made by HiFi callers. For all true *de novo* and false calls made by HiFi callers, a comparison of read depth at the site of the call in (**A**) the father, (**B**) the mother, and (**C**) the child with the *de novo* call. In (**D**), the father's depth at the site plotted against the mother's depth, false calls show significantly lower parental read depth than true calls.

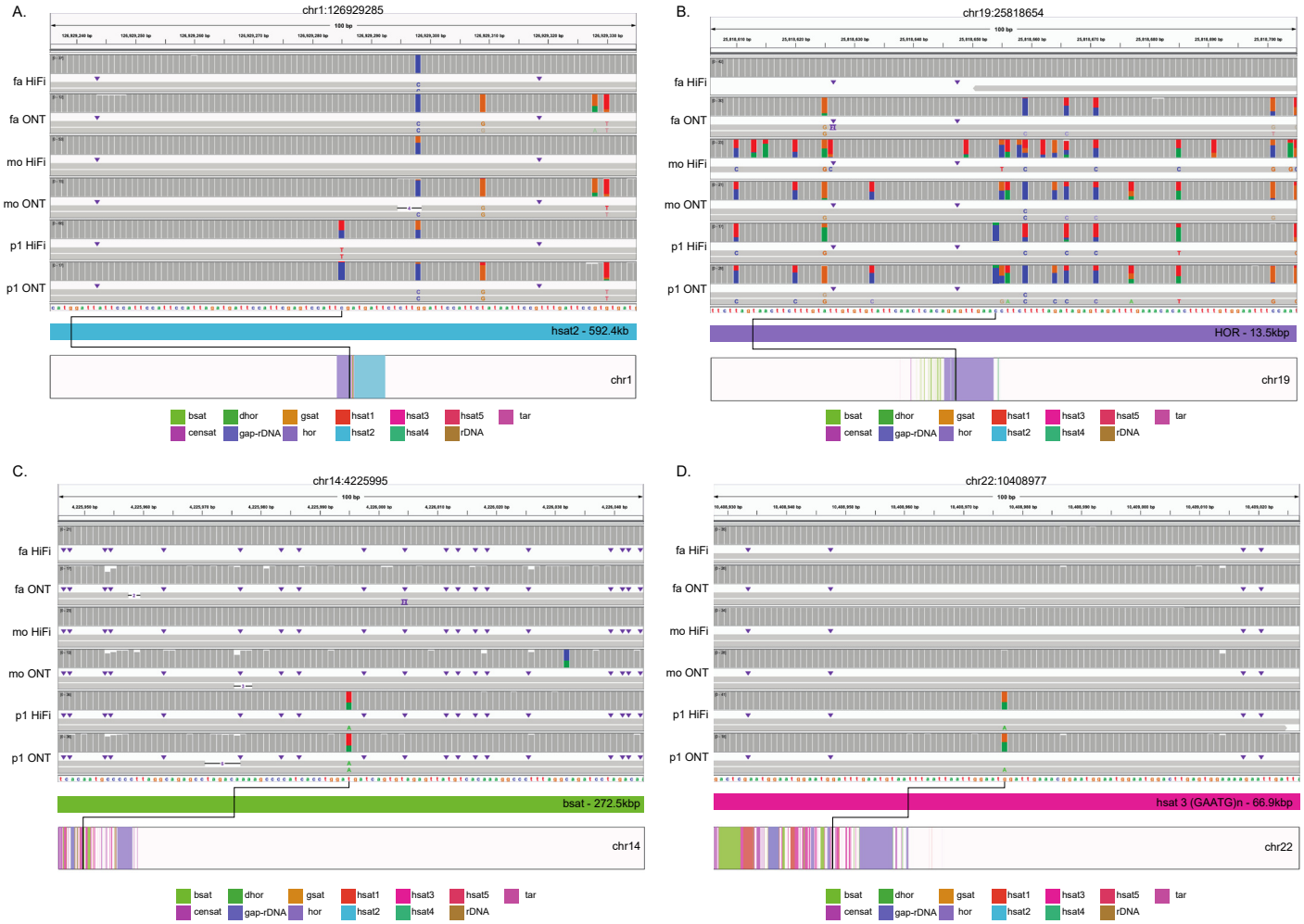


Figure S3. IGV shots of centromeric DNMs. (A-D) IGV shots of PacBio HiFi and ONT reads aligned to centromeric heterochromatic satellite regions. Underneath the reads is the location of the variant in its repetitive context, and below that is the chromosome with centromeric repeats annotated. All DNMs except B (chr19_25818654_C_A) are considered true positive events.

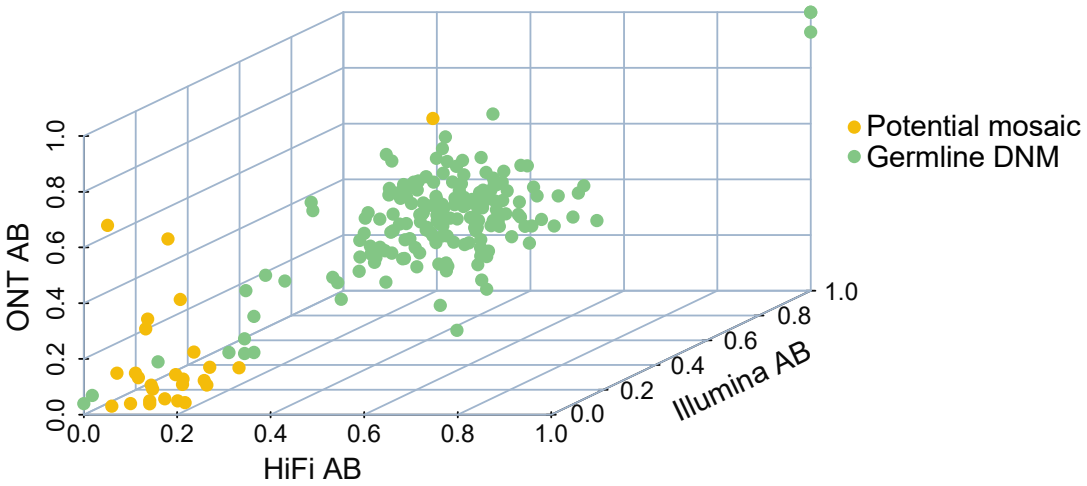


Figure S4. Allele balance in validated DNMs and potential mosaic mutations. The allele balance in the child with the mutation is shown across three sequencing platforms: ONT, PacBio HiFi, and Illumina. The potential mosaic mutations are in yellow, and validated *de novo* SNVs are shown in green. The germline sites that are clustered with the mosaic mutations were only observed in GRCh38-aligned reads and, accordingly, have very low allele balance (AB) in T2T-CHM13 aligned data.

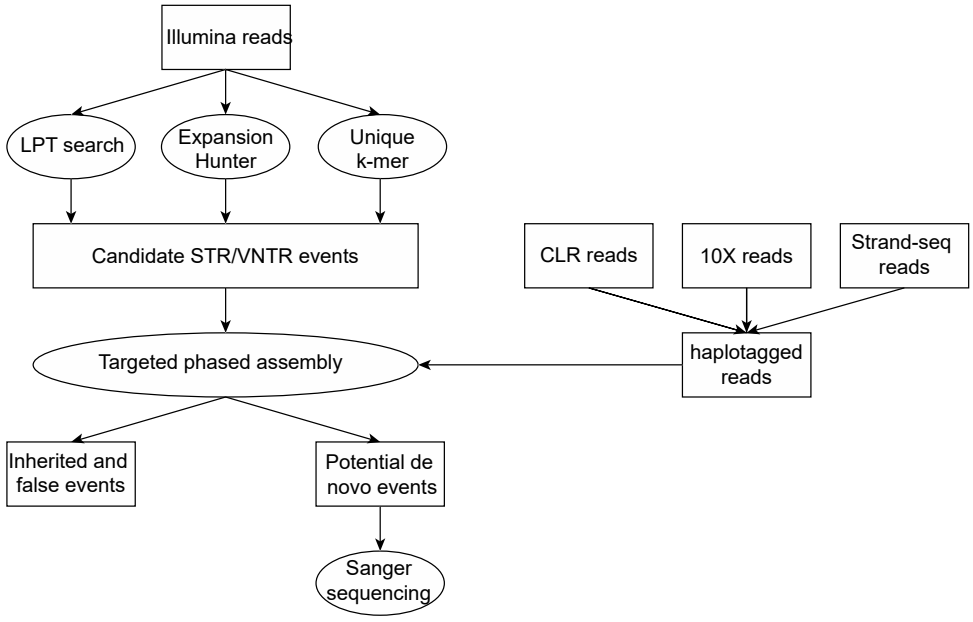


Figure S5. STR and VNTR calling pipeline. This pipeline identifies candidate STR and VNTR mutations in Illumina data. HiFi continuous long-read (CLR) data were assigned haplotypes using Chromium 10X genomic sequencing (10X) and single-cell DNA template strand sequencing (Strand-seq) data. These haplotagged reads were used for targeted phased assembly in order to validate candidate events; all were present in the assemblies.

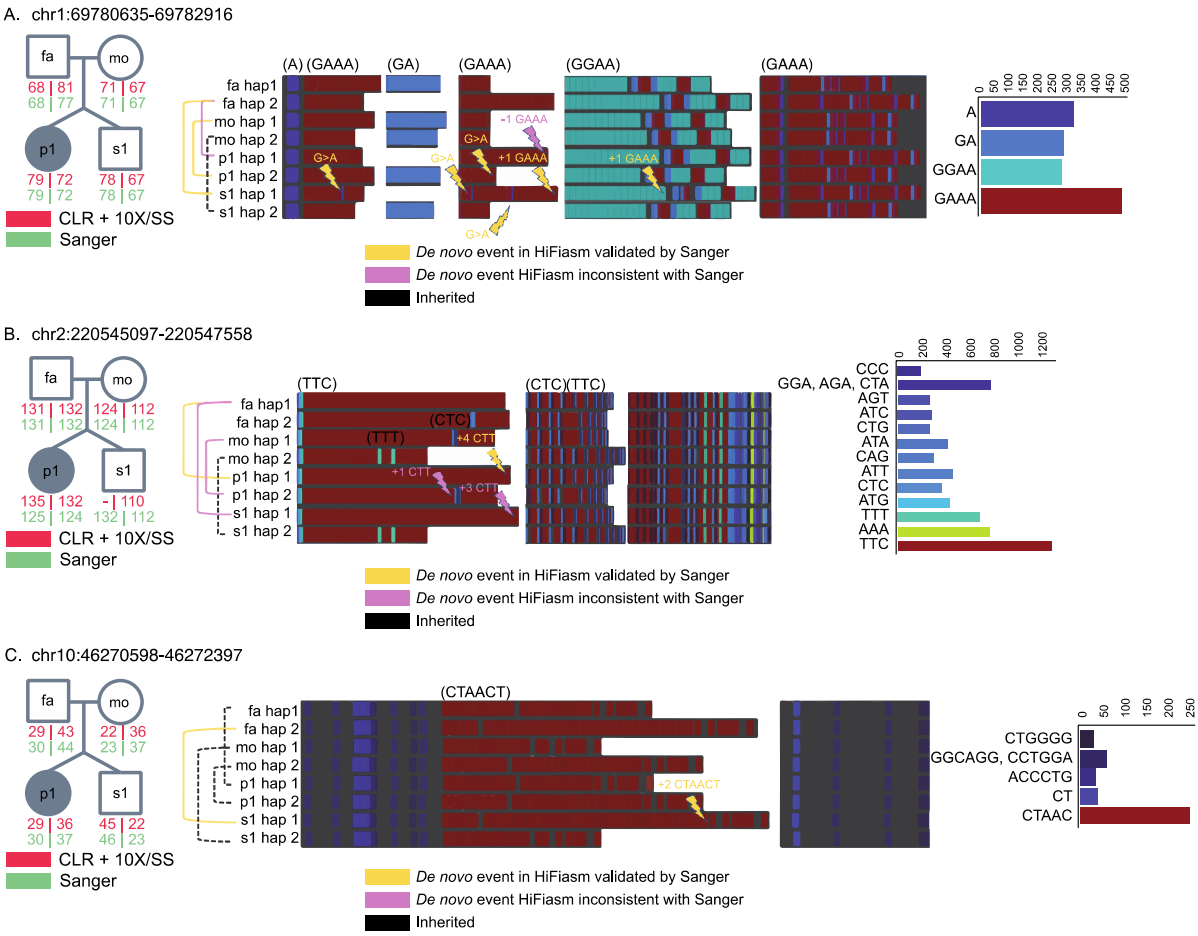
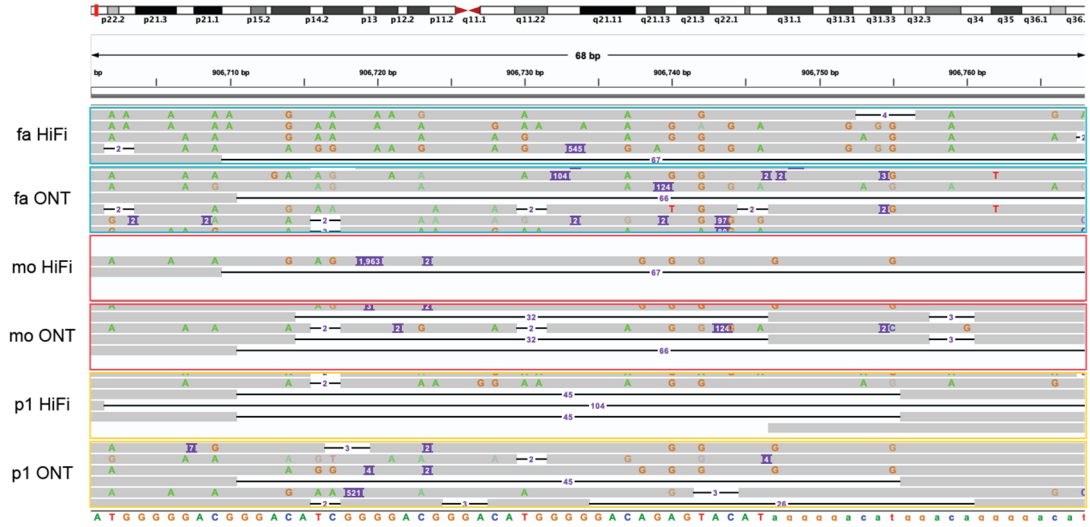
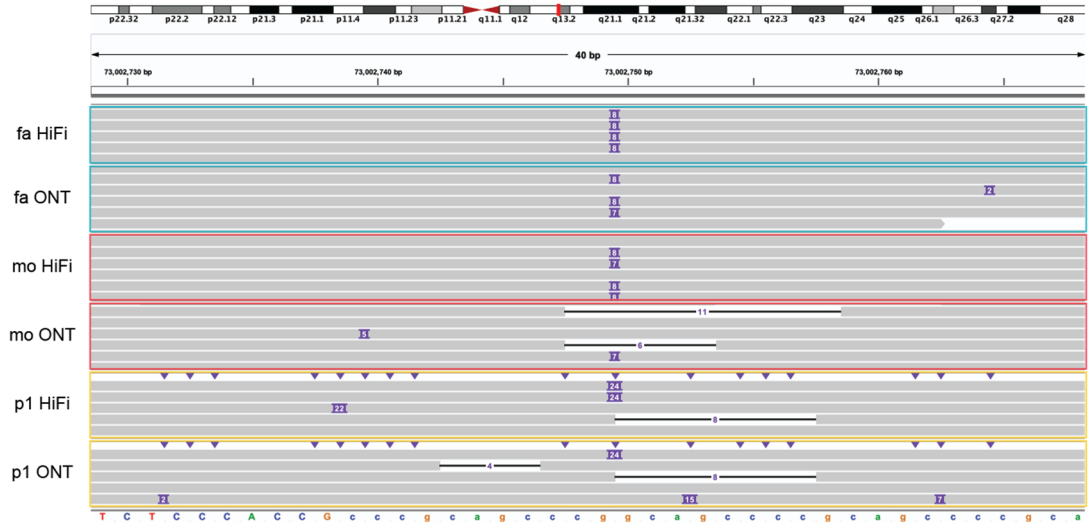


Figure S6. Identified *de novo* STR events. The structure of the family annotated with the number of STR copies detected in PacBio CLR haplotagged with 10X and Strand-seq data, and the assembled haplotypes for each individual, with variants highlighted by lightning bolts, depicted for an STR event in (**A**) the proband and sibling, (**B**) the proband, and (**C**) the sibling.

A. chr7-906711-DEL-45



B. chrX-73002750-INS-24



C. chrX-65373989-INS-47

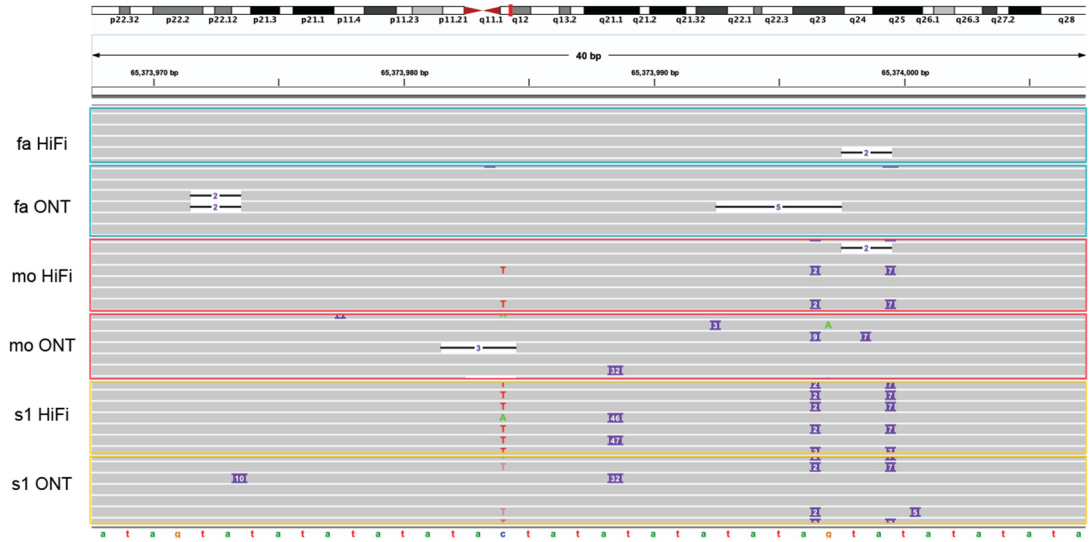


Figure S7. IGV shots of *de novo* 20-50 bp indels. IGV views of indel regions in the father, mother, and affected child across PacBio HiFi and ONT sequencing reads for (**A**) a 45 bp *de novo* deletion in the proband, (**B**) a 16 bp insertion in the proband, and (**C**) a 47 bp insertion in the sibling.

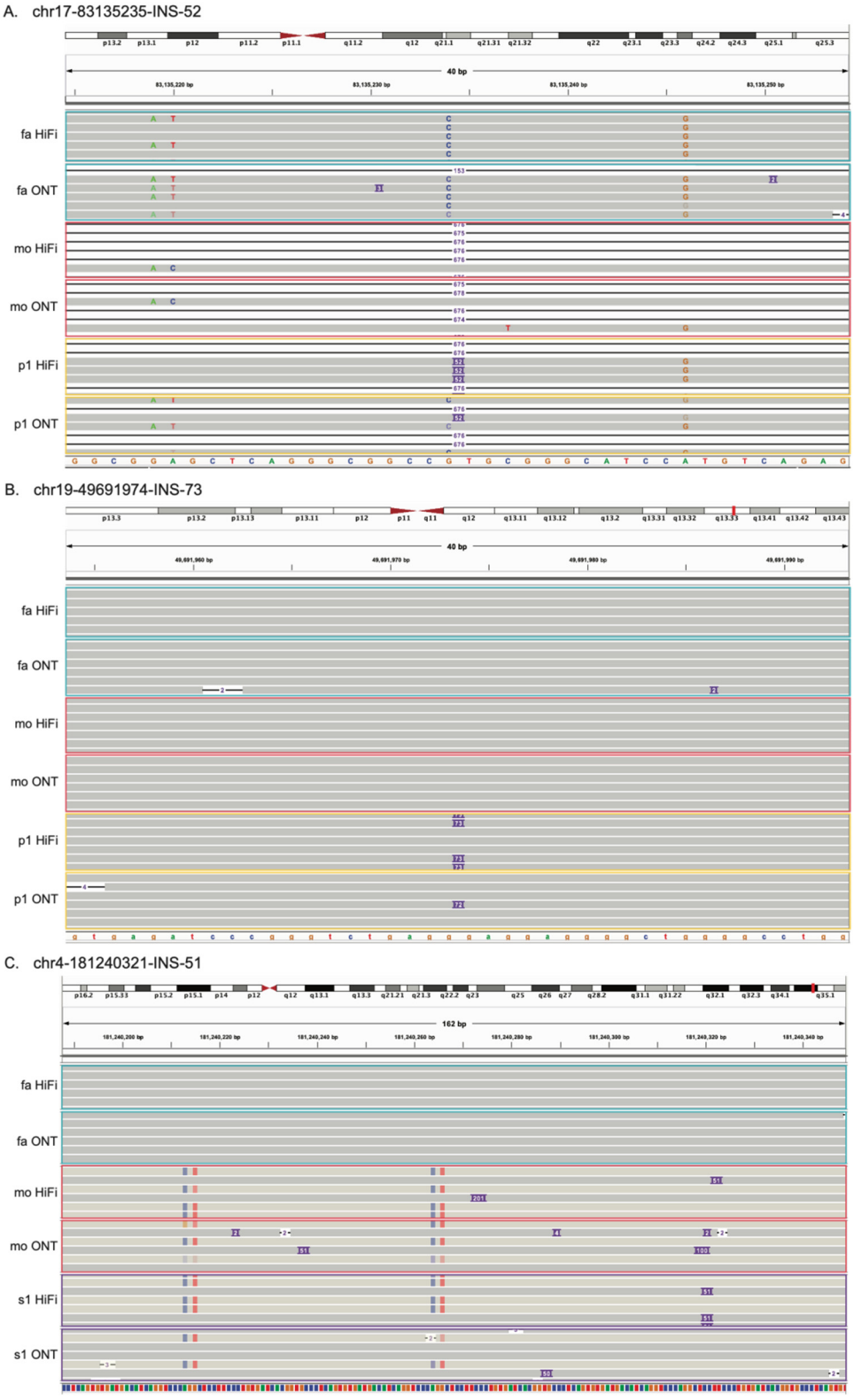
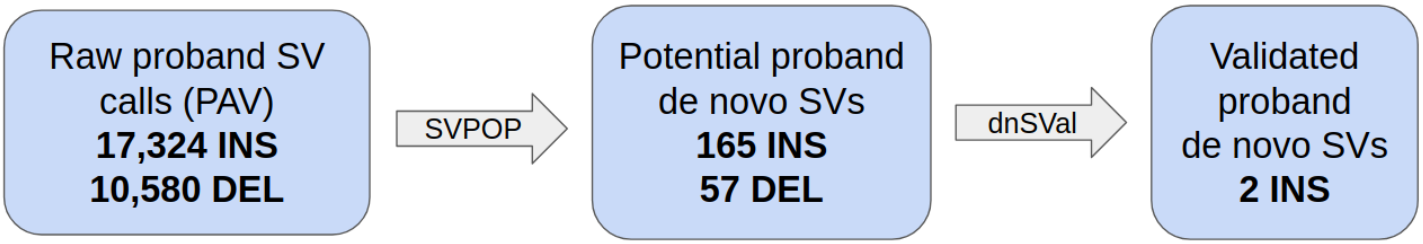


Figure S8. IGV shots of *de novo* SVs. IGV views of SV regions in the father, mother, and affected child across PacBio HiFi and ONT sequencing reads for (**A**) the 52 bp insertion in the proband, (**B**) the 73 bp insertion in the proband, and (**C**) the 51 bp insertion in the sibling.



PAV: https://github.com/EichlerLab/pav
SVPOP: https://github.com/EichlerLab/svpop
dnSVal: https://github.com/EichlerLab/denovo_sv_validation

Figure S9. Overview of automated SV filtering process. From the initial 27,904 *de novo* SV calls made in the proband, automated SVPOP filtering removed all but 232. The dnSVal validation uses subseq and multiple sequence alignment to further filter candidate *de* novo SV calls, resulting in a total of two validated *de novo* events in the proband—the same two that passed the manual filtering process.

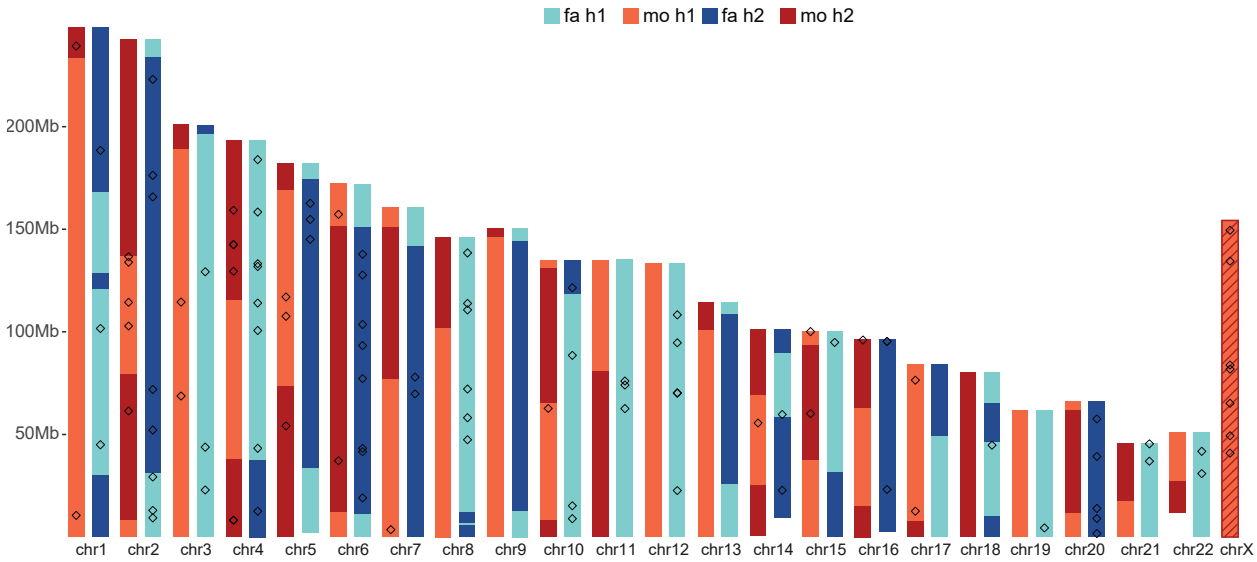


Figure S10. Meiotic crossovers and DNM. A genome-wide overview of detected meiotic recombination breakpoints for the sibling. Inherited segments of maternal homologs (H1-light red, H2-dark red) appear on the left side of each chromosome while inherited segments of paternal homologs (H1-light blue, H2-dark blue) appear on the right side of each chromosome. Recombination breakpoints are visible as changes from H1 to H2 segments and vice versa. Detected DNMs (n=105) that could have been assigned to a single parental homolog are shown as empty boxes over maternal (left) and paternal (right) homologs. This individual is a male meaning that maternal chromosome X does not recombine (empty red box).

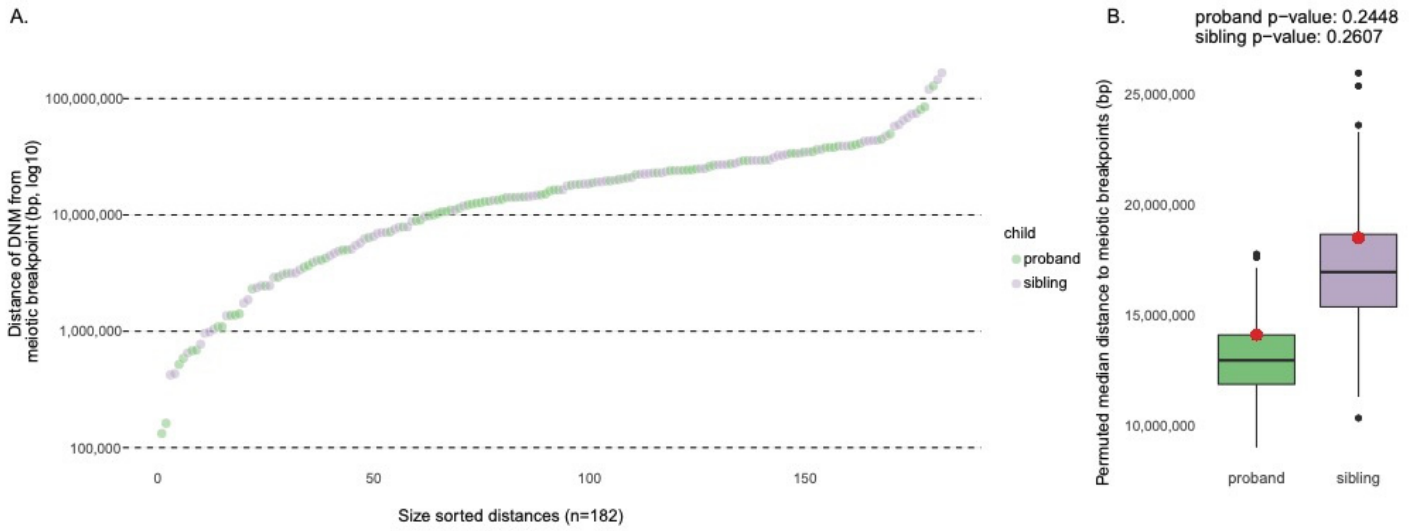


Figure S11. Meiotic recombination distance to DNMs. (**A**) Sorted distances of DNMs to the closest meiotic breakpoint reported for both proband (green) and sibling (purple). (**B**). An enrichment analysis comparing observed median distance of DNMs to meiotic breakpoints in comparison to permuted meiotic breakpoints (1000 permutations) separately for proband- and sibling-specific DNMs.

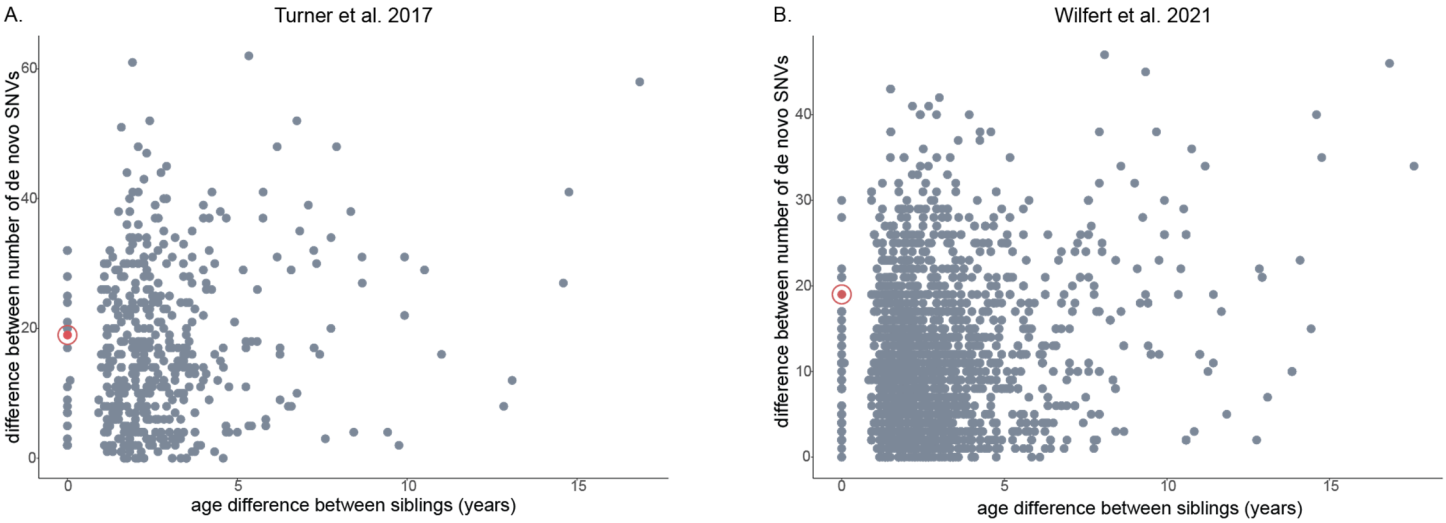


Figure S12. Intersibling *de novo* mutation difference. The differences in the DNM count are compared between proband and sibling as a function of the age difference between the siblings based on two previous studies. Red dots indicate Illumina-based estimates of intersibling DNM difference for the family studied here (SSC14455).

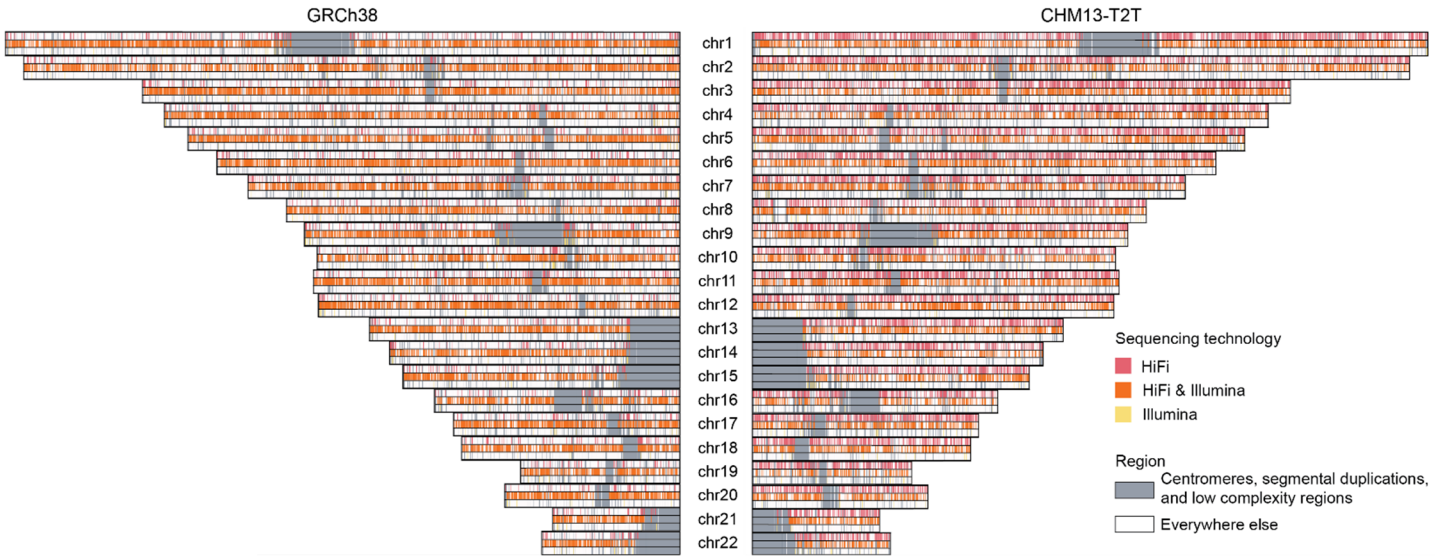


Figure S13. Distribution of rare inherited SNVs. Variant calls were made using GATK on GRCh38- (left) and T2T-CHM13- (right) aligned reads. Inherited variants were identified using a modification of the *de novo* pipeline to select all variants with genotype 0/1 or 1/1 in exactly one parent, and 0/1 in at least one child. Inherited candidates were filtered for depth >10 in both parents and children, genotype quality >25 in both parents and children, and allele balance >0.2 in all individuals with the variant. Variants were annotated with VEP for their frequency in gnomAD, and all variants with allele frequency less than 0.1% were classified as rare. Any variant that was confirmed to be present in the child's ONT data was retained for the final callset.

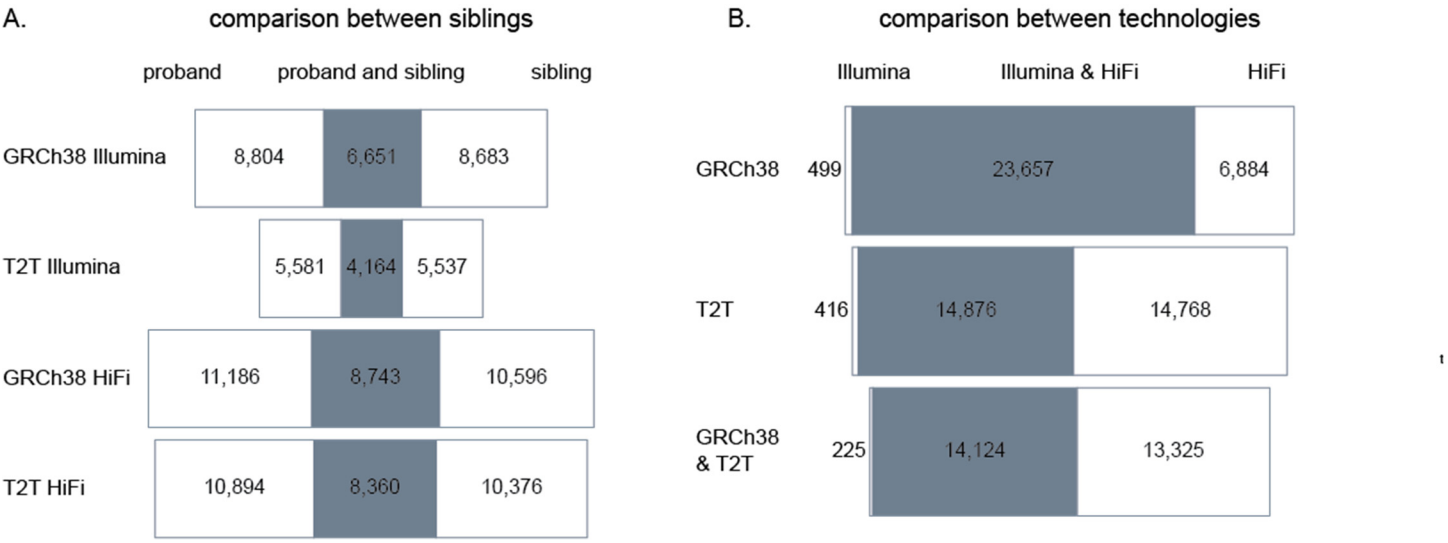


Figure S14. Inherited variants by child and technology. ONT-validated rare inherited SNV (<0.1% frequency in gnomAD) discovery comparing (**A**) proband and sibling callsets generated by Illumina or HiFi reads aligned to the GRCh38 or T2T-CHM13 references or (**B**) comparing discovery based on use of different sequencing technologies. In (**B**), sites identified in the same sample(s) were considered to be common to both callsets, whereas the same site identified in different samples would be considered unique to each callset. Gray bars in the Venn diagram represent shared SNVs based on platform or assembly while white bars represent SNVs unique to each. Inherited callsets generated using HiFi were larger than their Illumina counterparts, and nearly every site in the short-read callsets was also present in long-read callsets. The number of novel inherited T2T-CHM13 calls was fewer than GRCh38 callsets for both short and long reads - driven in part by a failure to liftover T2T-CHM13 to GRCh38 genomic coordinates.

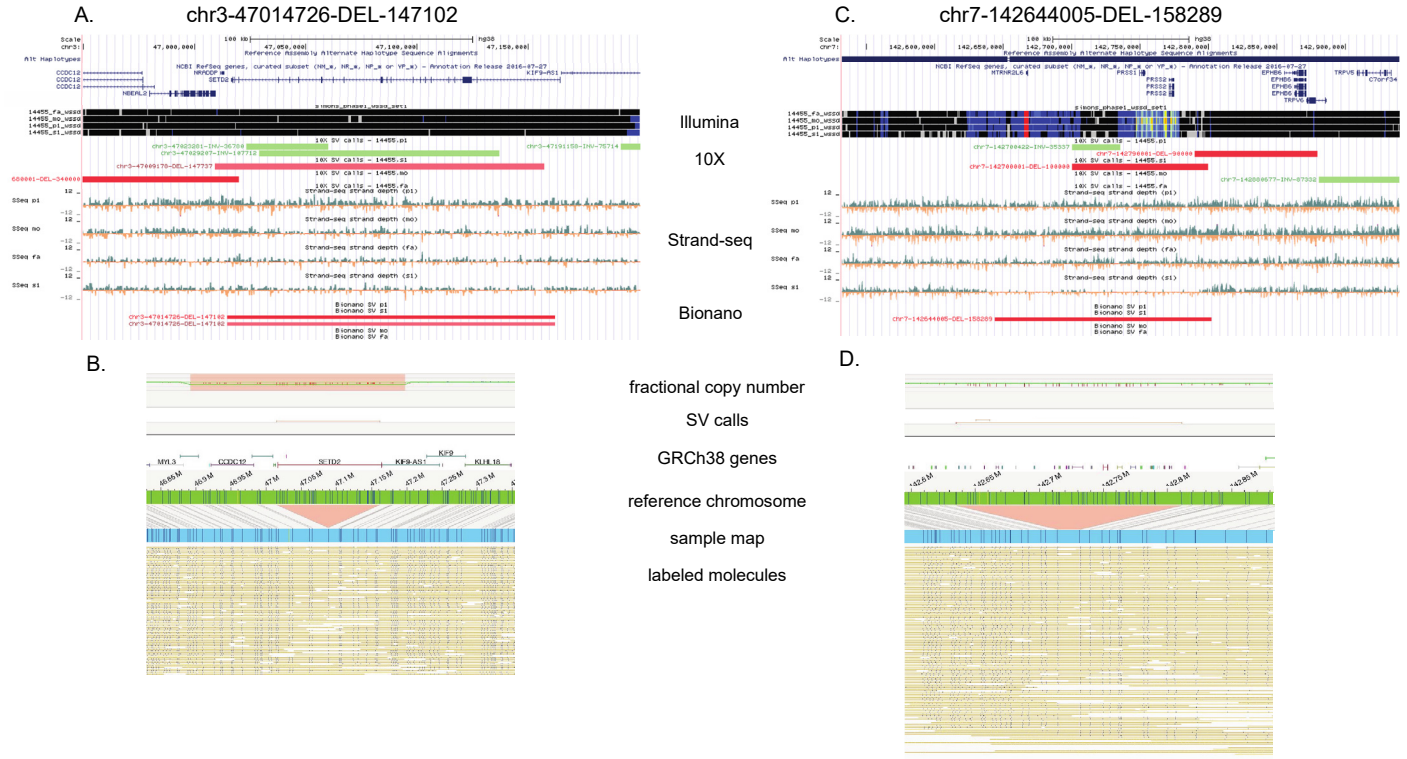


Figure S15. Support for cell line artifacts. (**A**) chr3-47014726-DEL-147102. Large sibling *SETD2* deletion cell line artifact discovered by 10X (LCL derived) and Bionano (LCL derived), but not supported by Strand-seq (LCL derived) or WSSD (blood derived). (**B**) Bionano support for *SETD2* deletion. The top bar represents the fractional copy number (green line is copy number two); the second bar is SV calls based on the assembly map. The red lines compare the green reference region to the sample map (the label pattern observed over the region). The bottom represents all the labeled molecules over the region. (**C**) chr7-142644005-DEL-158289. Large deletion over *PRSS1* and *PRSS2* is likely a cell line artifact with 10X, Bionano, and Strand-seq support. (**D**) Bionano support for *PRSS1* and *PRSS2* deletion, in the same format as B.

Supplemental Tables

Table S1. All candidate STR and VNTR events.

Chr	Position	Type	Motif	Calling approach	Fa repeat copies	Mo repeat copies	P1 repeat copies	S1 repeat copies	Sanger validation
Chr1	69781685	STR	AAAG	LPT	68 | 77	67 | 71	72 | 79	67 | 78	True positive
Chr1	103525498	VNTR	ACGGCGGGGC GGGGCGC	ExpansionHunter Denovo	9 | 14	11 | 11	11 | 15	11 | 14	True positive
Chr2	220546187	STR	TTC	LPT	131 | 132	112 | 124	124 | 125	112 | 132	True positive
Chr10	46271376	STR	CTAACT	30-mer	30 | 44	23 | 37	30 | 37	23 | 46	True positive
Chr2	133895459	VNTR	AAGAGAGAGGG GAGG	ExpansionHunter Denovo	12 | 18	6 | 18	17 | 18	12 | 17	Inherited
Chr3	111780258	STR	AAG	LPT	66 | 67	69 | 71	67 | 69	67 | 71	Inherited
Chr5	38824434	STR	CCACCA	30-mer	18 | 21	18 | 19	18 | 19	19 | 21	Inherited
Chr13	44142133	STR	CTCGG	LPT	18 | 25	18 | NA	18 | 25	18 | 18	Inherited
Chr17	51831668	STR	AGC	LPT	22 | 23	19 | 22	19 | 22	19 | 23	Inherited
Chr1	101657855	STR	TTC	LPT	NA	NA	NA	NA	Not supported
Chr7	84690930	STR	GAA	LPT	79 | NA	71 | NA	NA	71| 72	Not supported
Chr12	111257196	VNTR	AAGAAGTGGGA GGG	ExpansionHunter Denovo	33 | 37	36 | 36	37 | NA	37 | NA	Not supported
Chr14	44005178	STR	AGA	LPT	NA	71 | NA	76 | 77	76 | 77	Not supported
Chr14	99927754	STR	GTG	LPT	NA	NA	NA	NA	Not supported
Chr16	20041114	STR	AGGAG	LPT	NA	NA	NA	NA	Not supported

Events were identified by one of three approaches: a custom k-mer based approach (30-mer), an approach based on the longest pure tandem repeat (LPT), and the tool ExpansionHunter Denovo.

Table S2. 20-50 bp *de novo* calls in the proband and sibling.

Child	Chr	Position	Indel type	Length	HiFi/ONT Validation
14455.p1	chr7	906710	DEL	45	**potential de novo**
14455.p1	chr7	132561922	DEL	28	false positive
14455.p1	chr11	82293621	DEL	24	inherited from mom
14455.p1	chr16	14928921	DEL	27	inherited from mom
14455.p1	chr16	61425944	DEL	35	false positive
14455.p1	chr19	15778633	DEL	40	false positive
14455.p1	chr21	39583858	DEL	33	inherited from dad
14455.p1	chr2	90033846	INS	21	false positive
14455.p1	chr2	114396849	INS	44	inherited from mom
14455.p1	chr6	95411818	INS	29	inherited from mom
14455.p1	chr8	57949332	INS	44	inherited from mom
14455.p1	chr9	71816230	INS	43	false positive
14455.p1	chr9	137350578	INS	33	inherited
14455.p1	chr12	3978627	INS	29	false positive
14455.p1	chr13	84254692	INS	25	inherited from dad
14455.p1	chr15	99927049	INS	22	false positive
14455.p1	chr18	32773926	INS	32	inherited from mom
14455.p1	chr20	31890154	INS	23	inherited
14455.p1	chr22	25093859	INS	38	false positive
14455.p1	chr22	28867115	INS	25	false positive
14455.p1	chrX	73002749	INS	24	**potential de novo**
14455.s1	chr8	67056650	DEL	24	inherited from mom
14455.s1	chr10	128976652	DEL	32	inherited from dad
14455.s1	chrX	40557339	DEL	48	inherited from mom
14455.s1	chr1	19020904	INS	45	inherited from mom
14455.s1	chr4	182833096	INS	42	inherited from mom
14455.s1	chr5	29815038	INS	28	inherited from mom
14455.s1	chr8	108938105	INS	44	inherited from mom
14455.s1	chr8	128825889	INS	32	false positive
14455.s1	chr8	143218041	INS	21	inherited
14455.s1	chr9	42090783	INS	44	inherited from mom
14455.s1	chr10	131625735	INS	27	inherited from mom
14455.s1	chr11	33614584	INS	21	inherited from dad
14455.s1	chr12	76955213	INS	24	inherited from mom
14455.s1	chr13	87253488	INS	23	inherited from dad
14455.s1	chr17	71886772	INS	37	inherited
14455.s1	chr18	22204288	INS	48	inherited
14455.s1	chrX	65373988	INS	47	**potential de novo**
14455.s1	chrY	1735802	INS	34	no read data

All 39 20-50 bp indel calls identified using assembly-driven variant discovery.

Table S3. *De novo* SVs identified by Bionano Genomics.

Child	Chr	Position	SV type	Length	Population Frequency (%)
14455.p1	chr3	90544141	DEL	4,656	2.5
14455.p1	chr14	105847409	DEL	1,461	1
14455.p1	chr2	89995213	INS	812	9.3
14455.s1	chr3	47014725	DEL	147,102	0
14455.s1	chr7	142635453	DEL	158,289	0
14455.s1	chr9	66003864	DEL	905	0
14455.s1	chr21	8706195	DEL	442,533	9.8
14455.s1	chr19	8729966	INS	114,606	37.7

Summary of *de novo* SVs detected by Bionano analysis of the proband and unaffected sibling generated from cell line DNA. Population frequency is determined by Bionano controls.

Table S4. Crossover events in the proband and sibling. (See attached excel spreadsheet)
All 451 crossover events identified in the proband and the sibling from Strand-seq data.

Table S5. Distribution of DNMs by variant class.

Mutation type	Affected child	Count
VNTR >50 bp (SV INS)	14455.p1	2
	14455.s1	0
VNTR <50 bp	14455.p1	1
	14455.s1	0
STR	14455.p1	4
	14455.s1	3
indel	14455.p1	5
	14455.s1	9
SNV	14455.p1	81
	14455.s1	97

The number of mutations identified in each category - variable number tandem repeat (VNTR) expansions >50 bp (also represent SV insertions) and <50 bp, short tandem repeat (STR) expansions, short indels <20 bp (indels), and SNVs.

Table S6. Master table of DNMs. (See attached excel spreadsheet)
Every validated *de novo* event identified by this study, including STR/VNTR expansions, indels, SNVs, and potential mosaic mutations.