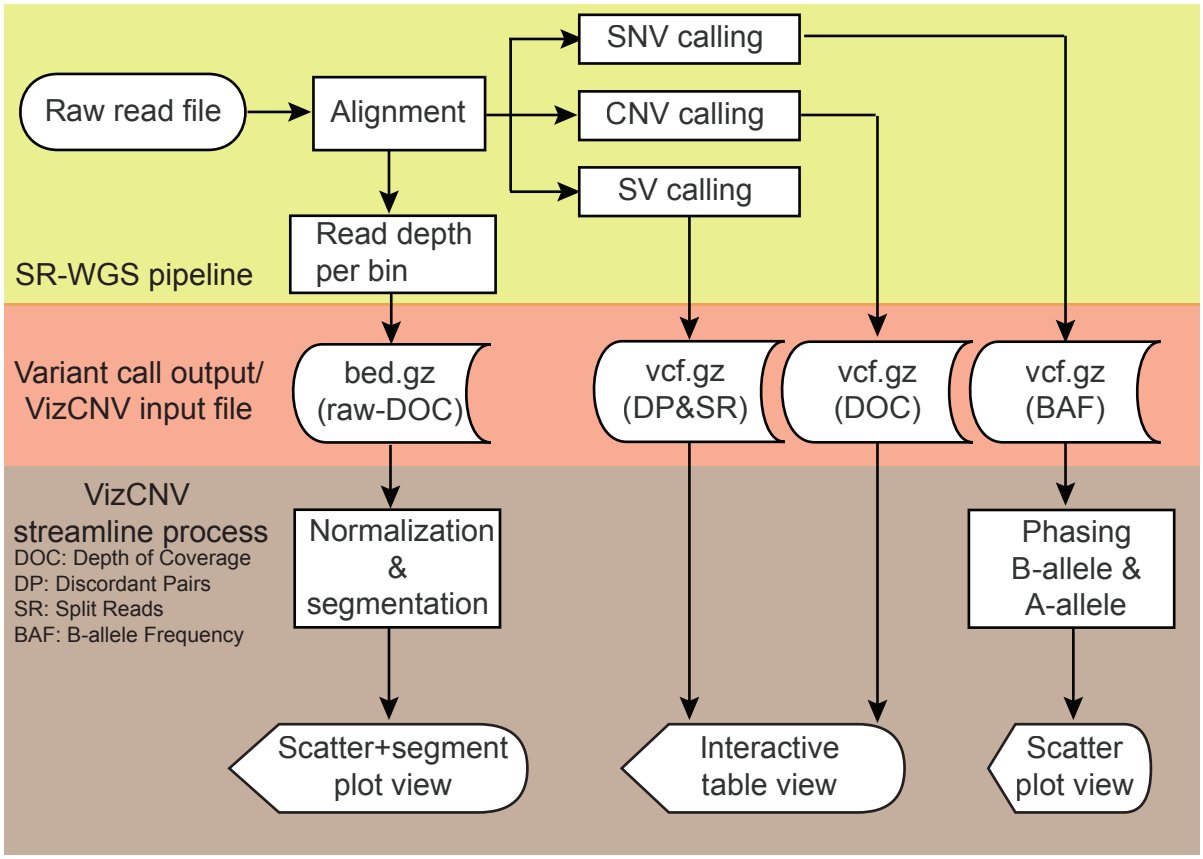
An integrated platform for concurrent structural and single-nucleotide variants improves copy-number detection and reveals pathogenic alleles in undiagnosed Mendelian families

Haowei Du, Ming Yin Lun, Lidiia Gagarina, Jesse D. Bengtsson, Christopher M. Grochowski, Michele G Mehaffey, James Paul Hwang, Shalini N. Jhangiani, Sravya V. Bhamidipati, Donna M. Muzny, M Cecilia Poli, Sebastian Ochoa, Ivan K. Chinn, Anna Lindstrand, Jennifer E. Posey, Richard A Gibbs, Pengfei Liu, James R. Lupski, Claudia M. B. Carvalho

S1

a


Raw read file
Alignment
SNV calling
CNV calling
SV calling
Read depth per bin
SR-WGS pipeline
Variant call output/ VizCNV input file
bed.gz (raw-DOC)
vcf.gz (DP&SR)
vcf.gz (DOC)
vcf.gz (BAF)
VizCNV streamline process
DOC: Depth of Coverage
DP: Discordant Pairs
SR: Split Reads
BAF: B-allele Frequency
Normalization & segmentation
Phasing B-allele & A-allele
Scatter+segment plot view
Interactive table view
Scatter plot view

b

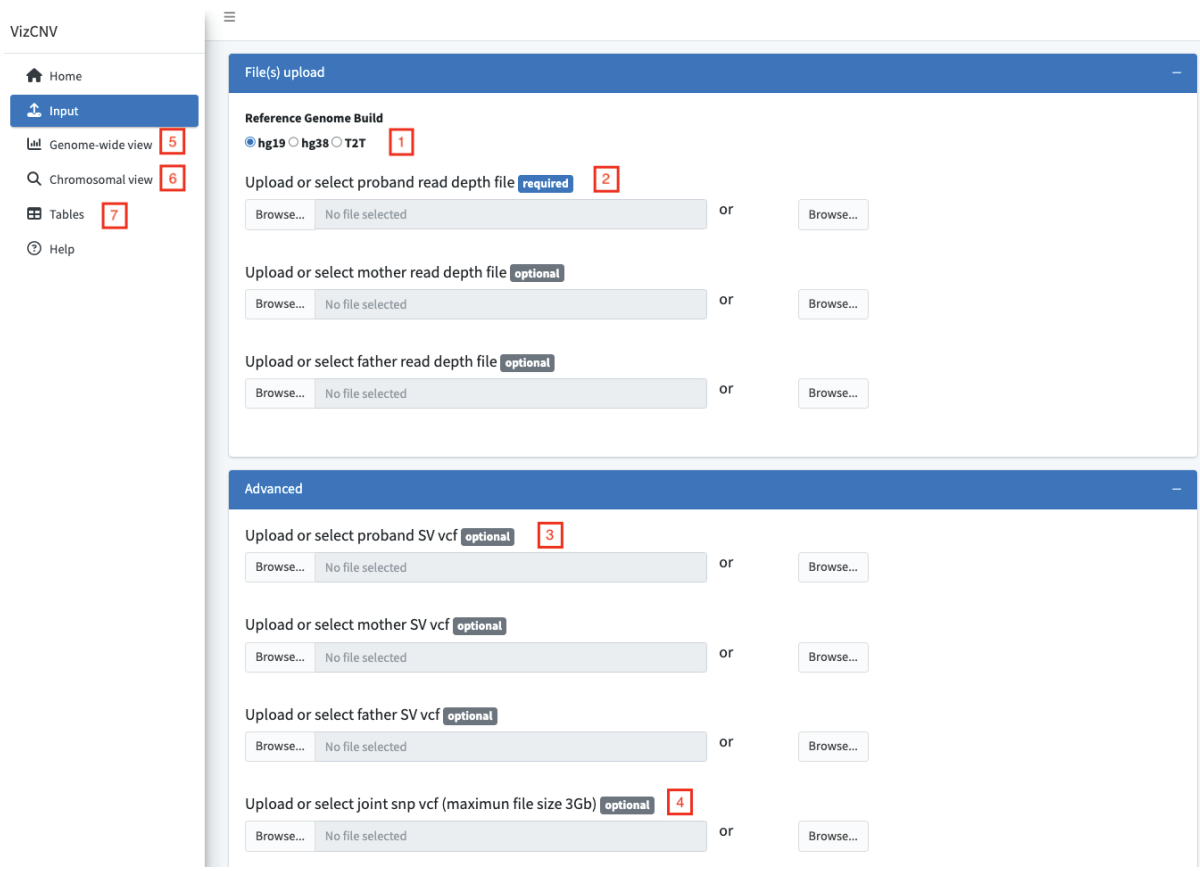
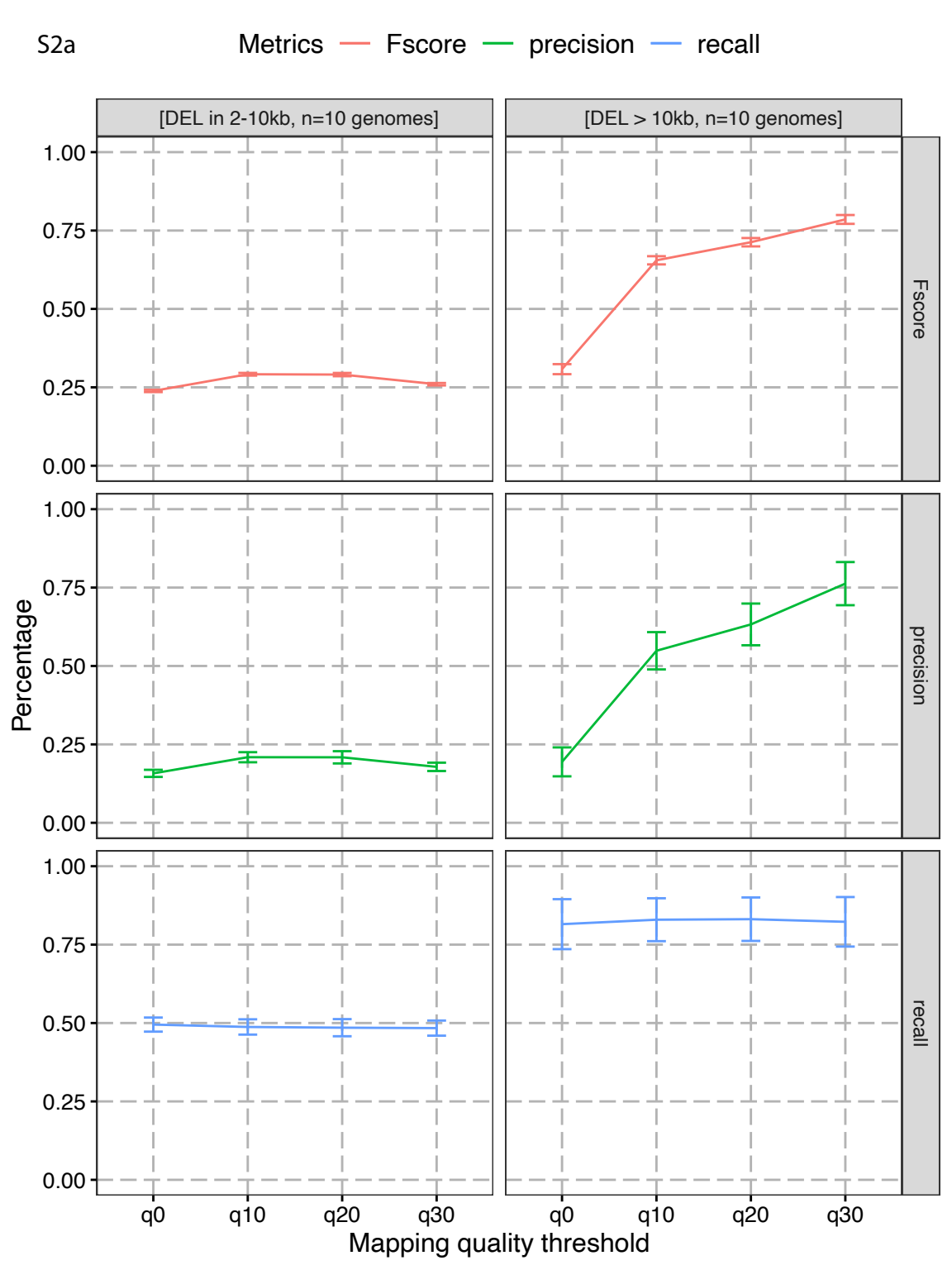
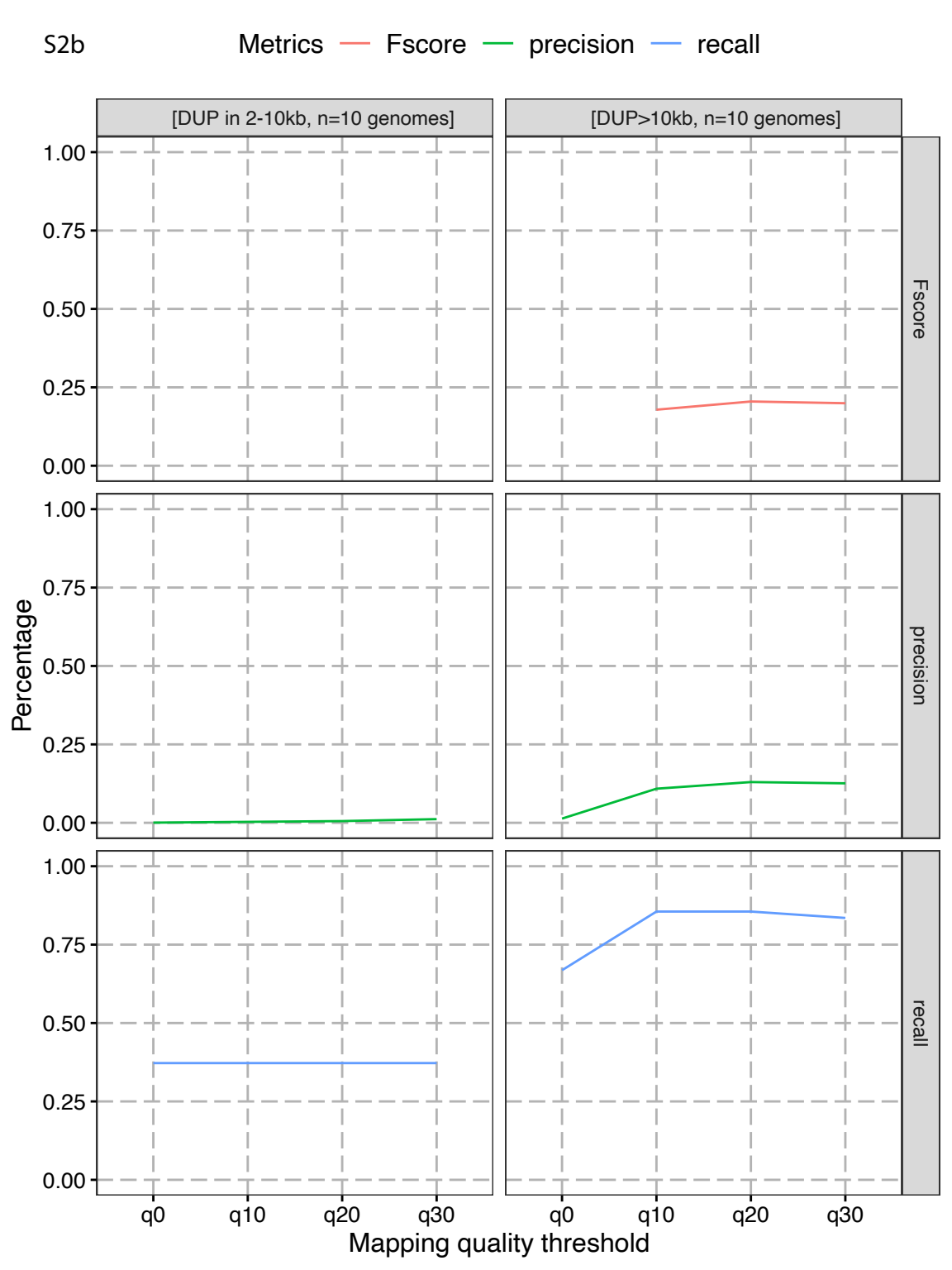

VizCNV
Home
Input
Genome-wide view 5
Chromosomal view 6
Tables 7
Help
File(s) upload
Reference Genome Build
hg19 hg38 T2T 1
Upload or select proband read depth file required 2
Browse... No file selected or Browse...
Upload or select mother read depth file optional
Upload or select father read depth file optional
Advanced
Upload or select proband SV vcf optional 3
Upload or select mother SV vcf optional
Upload or select father SV vcf optional
Upload or select joint snp vcf (maximun file size 3Gb) optional 4

Figure S1. VizCNV framework. a) VizCNV workflow demonstrating the common output files from a standard short-read genome sequencing pipeline incorporated into the VizCNV framework. **b)** VizCNV user interface includes 1. Build-in reference, 2. input for read-depth file, 3. input of proband structure variant call file, 4. joint-called genotype file, 5. genome-wide view which normalizes read-depth data using the full genome data enabling both aneuploidy and chromosomal copy-number variant screening, 6. chromosome view normalizes read-depth data using a single chromosome enabling fast assessment of CNVs, 7. Output tables with annotated CNV and SV information can be downloaded. Visualization of annotated CNVs and SVs for singletons or trios together BAF plots and tools for de novo CNV filtering are accessed in the Chromosomal View window.


S2a
Metrics
Fscore
precision
recall
[DEL in 2-10kb, n=10 genomes]
[DEL > 10kb, n=10 genomes]
Fscore
precision
recall
Percentage
1.00
0.75
0.50
0.25
0.00
q0
q10
q20
q30
Mapping quality threshold


S2b
Metrics
Fscore
precision
recall
[DUP in 2-10kb, n=10 genomes]
[DUP>10kb, n=10 genomes]
Fscore
precision
recall
Percentage
1.00
0.75
0.50
0.25
0.00
q0
q10
q20
q30
Mapping quality threshold

S2c **Caller comparison across size and CNV types**

Metrics: Precision / Recall; Size bins: 2kb–10kb and >10kb

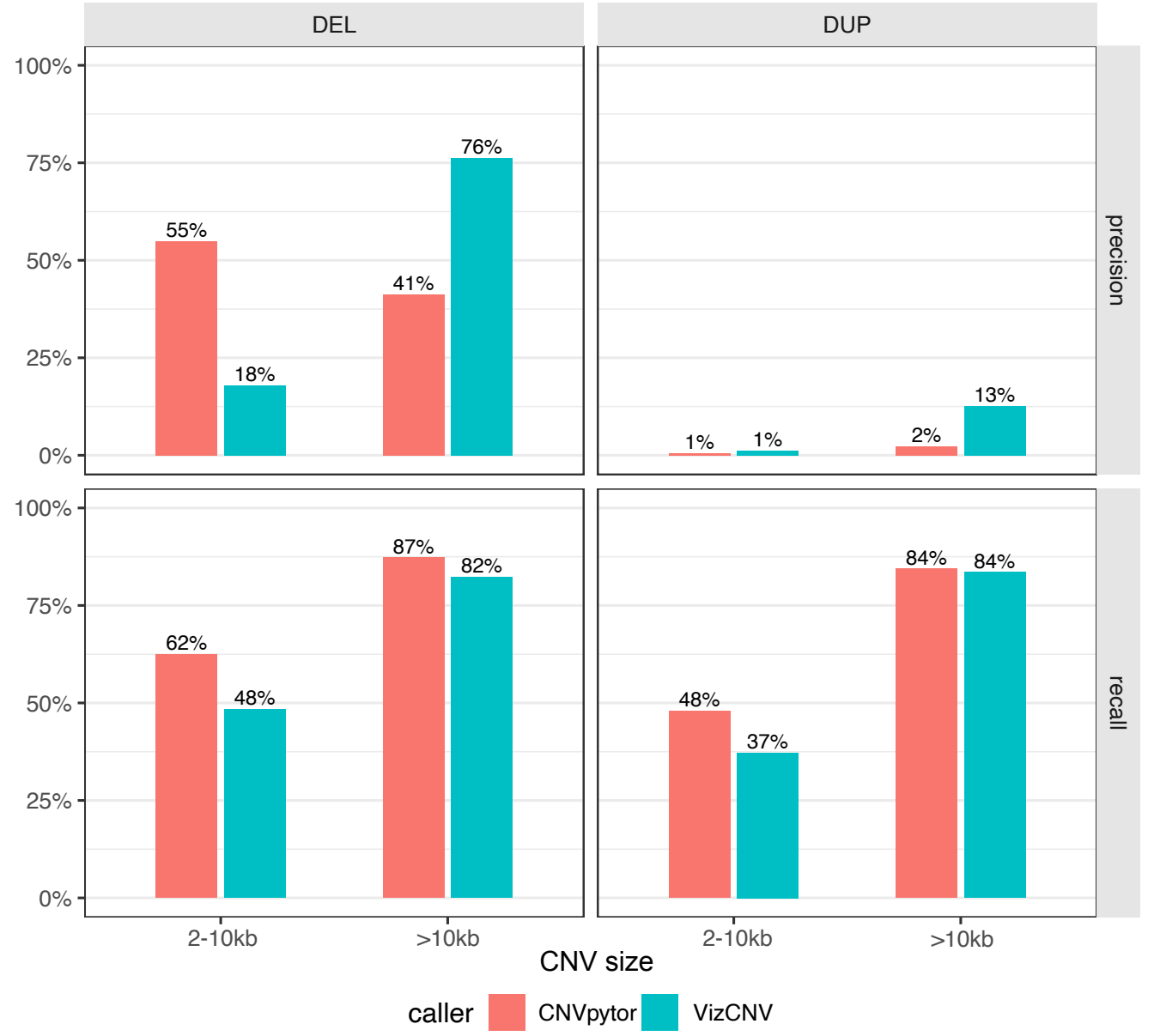
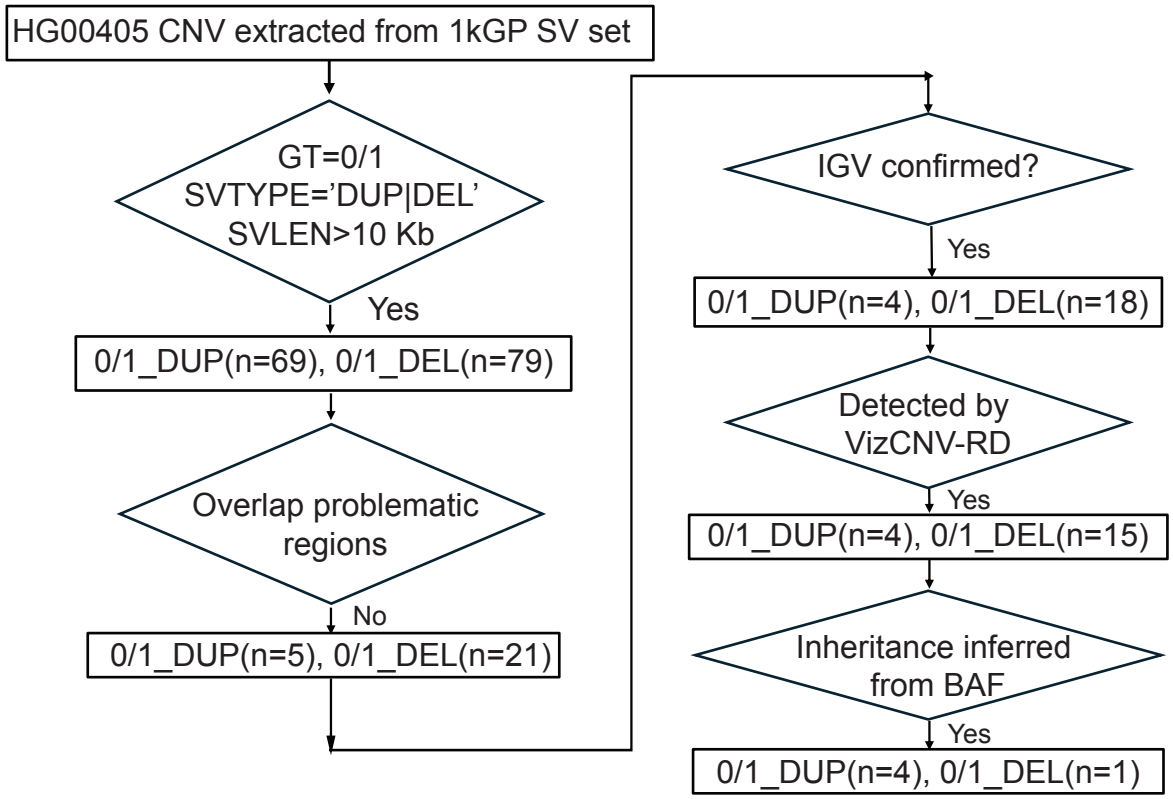


Figure S2. Performance metrics of CNV detection at various mapping quality thresholds across ten samples from 1000 Genomes Project. Metrics include F score (red), precision (green), and recall (blue) calculated separately for for deletions **a**) and duplications **b**) with two size ranges (2-10kb and >10kb). Error bars represent 95% confidence intervals where applicable. **c)** Comparison of CNV detection performance between CNVpytor and VizCNV across CNV types and size categories.

a


HG00405 CNV extracted from 1kGP SV set
GT=0/1
SVTYPE='DUP|DEL'
SVLEN>10 Kb
Yes
0/1_DUP(n=69), 0/1_DEL(n=79)
Overlap problematic regions
No
0/1_DUP(n=5), 0/1_DEL(n=21)
IGV confirmed?
Yes
0/1_DUP(n=4), 0/1_DEL(n=18)
Detected by VizCNV-RD
Yes
0/1_DUP(n=4), 0/1_DEL(n=15)
Inheritance inferred from BAF
Yes
0/1_DUP(n=4), 0/1_DEL(n=1)

b

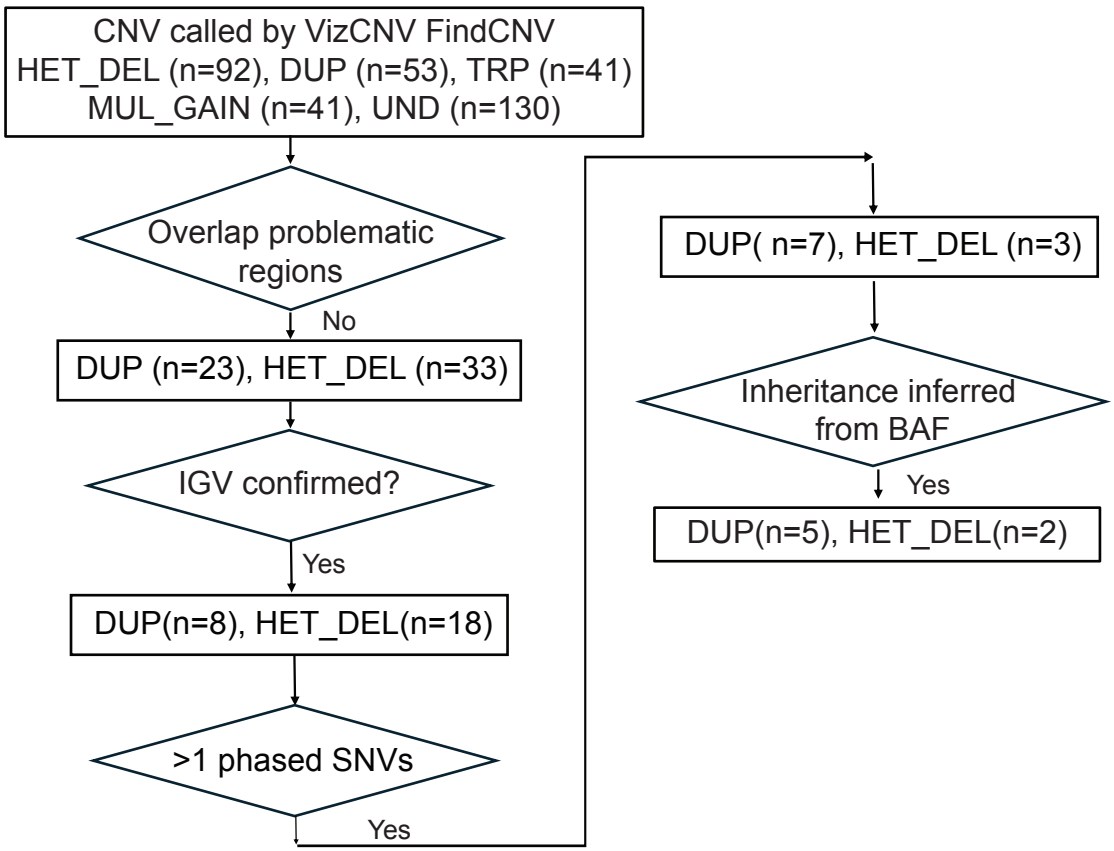

CNV called by VizCNV FindCNV
HET_DEL (n=92), DUP (n=53), TRP (n=41)
MUL_GAIN (n=41), UND (n=130)
Overlap problematic regions
No
DUP (n=23), HET_DEL (n=33)
IGV confirmed?
Yes
DUP(n=8), HET_DEL(n=18)
>1 phased SNVs
Yes
DUP(n=7), HET_DEL (n=3)
Inheritance inferred from BAF
Yes
DUP(n=5), HET_DEL(n=2)

Figure S3. Schematic Representation of Joint BAF and Read Depth Analysis in the Individual HG00405. a) CNVs originally identified and genotyped by structural variant calling from the 1000 Genomes Project report were extracted, focusing specifically on heterozygous CNVs in autosomal regions, and subsequently reanalyzed using the VizCNV FindCNV pipeline. **b**) Autosomal heterozygous deletions and duplications identified by the VizCNV FindCNV pipeline were further validated and analyzed through BAF assessment.

HG00405, HGSV_211297, chr16:78673486-79733774

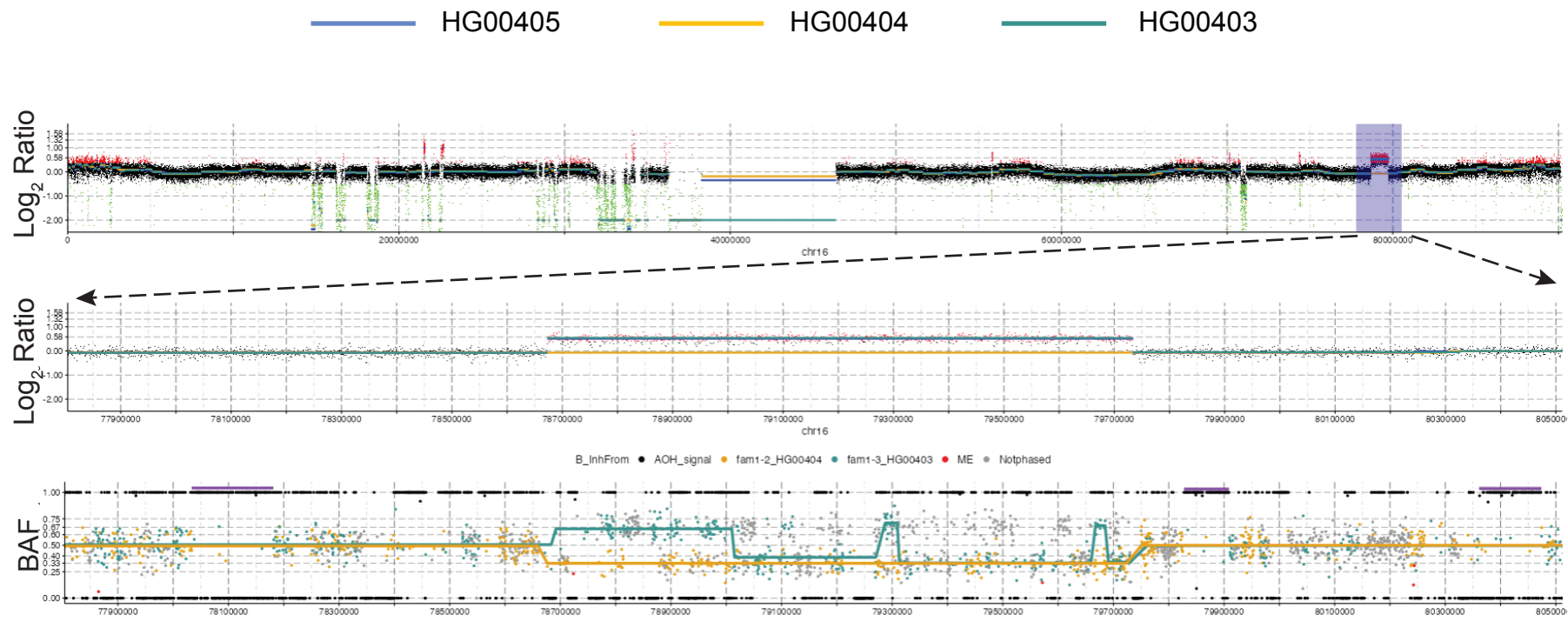
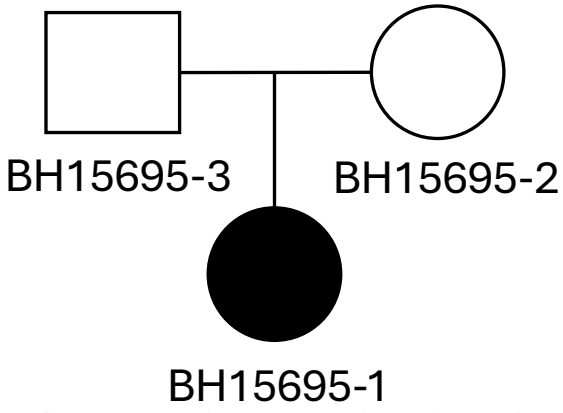


Figure S4. Paternal Inheritance of Duplication with Potential Recombination. The top panel presents a whole chromosome $\log_2$ ratio plot for the child (HG00405), father (HG00403), and mother (HG00404), with a blue rectangle highlighting a ~1 Mb duplication identified in both the child and father. The middle and bottom panels offer a detailed view of the highlighted region, focusing on $\log_2$ ratio and B-allele frequency (BAF), respectively. Colors indicate the origin of the variant: teal for paternal and orange for maternal. The oscillation of maternal and paternal SNP BAF values between 0.33 and 0.66 suggests the presence of three distinct haplotypes rather than two in the duplicated region.

Figure S5

a


BH15695-3
BH15695-2
BH15695-1

b

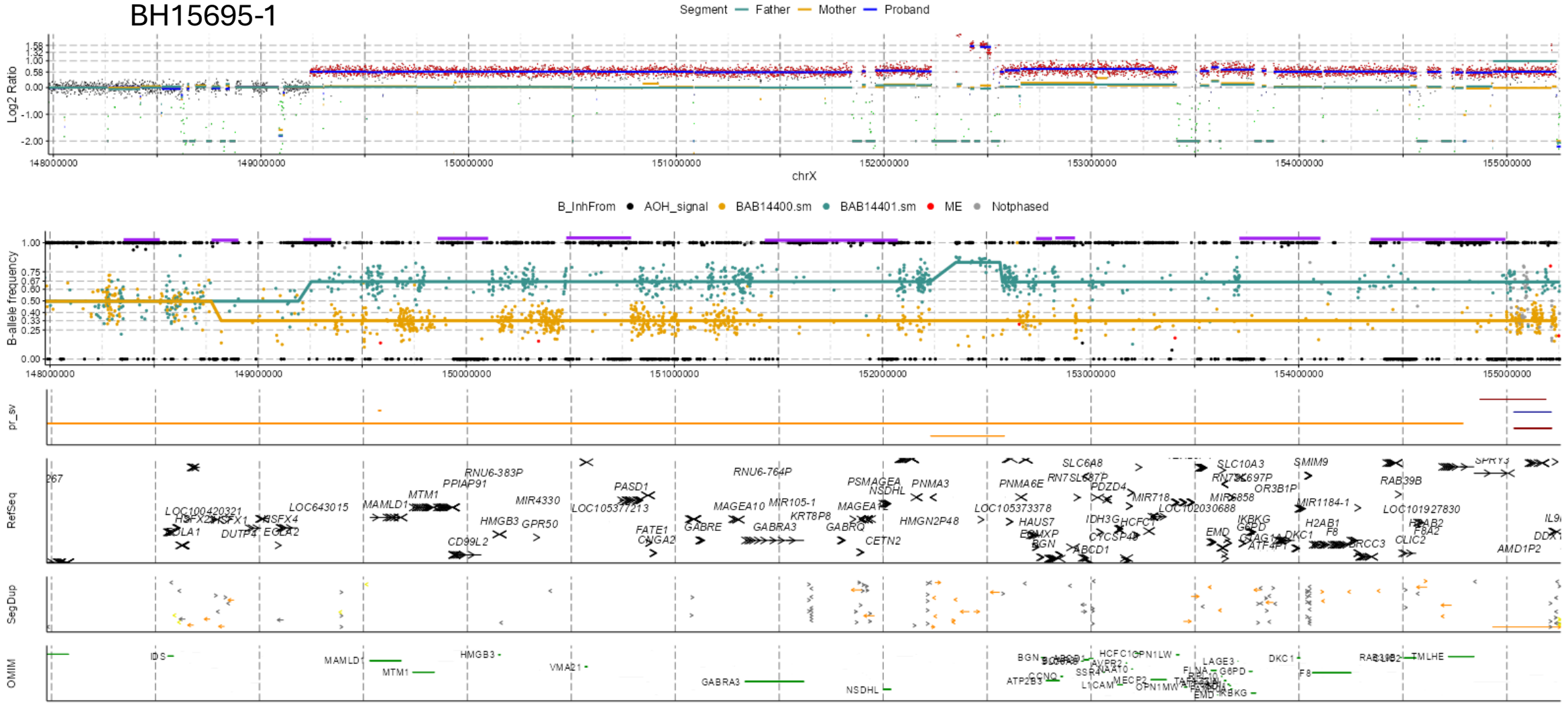
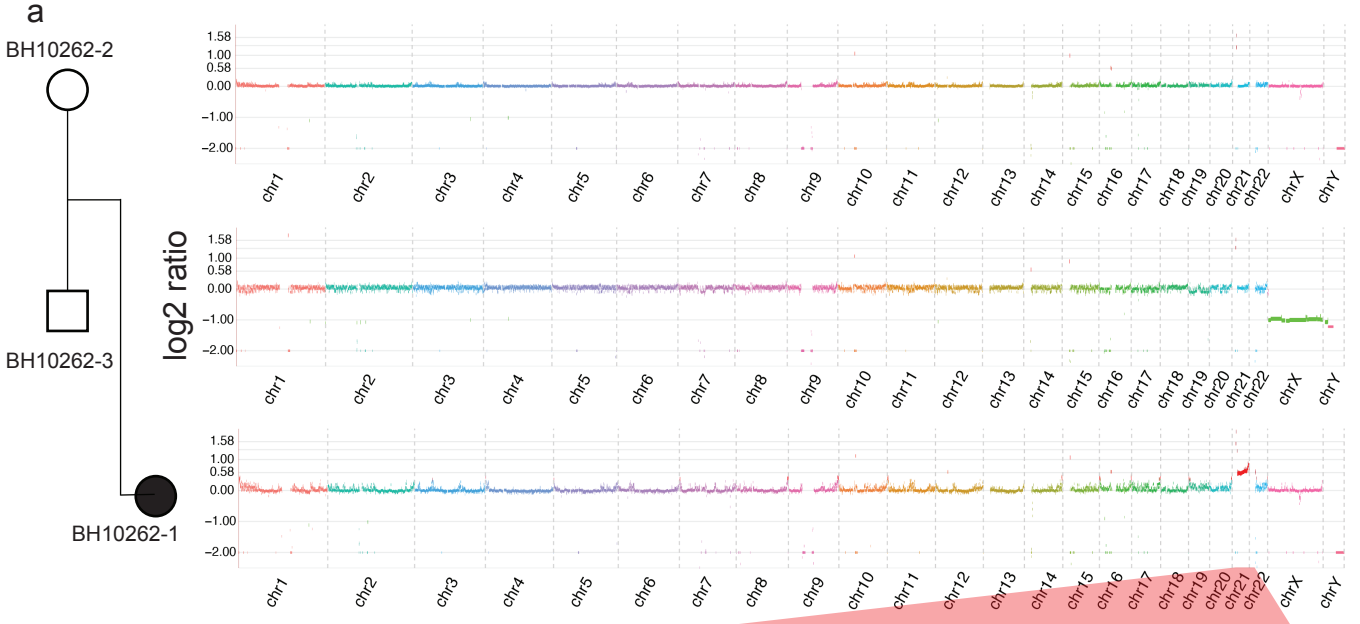

Segment Father Mother Proband
Log2 Ratio
chrX
B_InhFrom AOH_signal BAB14400.sm BAB14401.sm ME Notphased
B-allele frequency
pr_sv
RefSeq
SegDup
OMIM

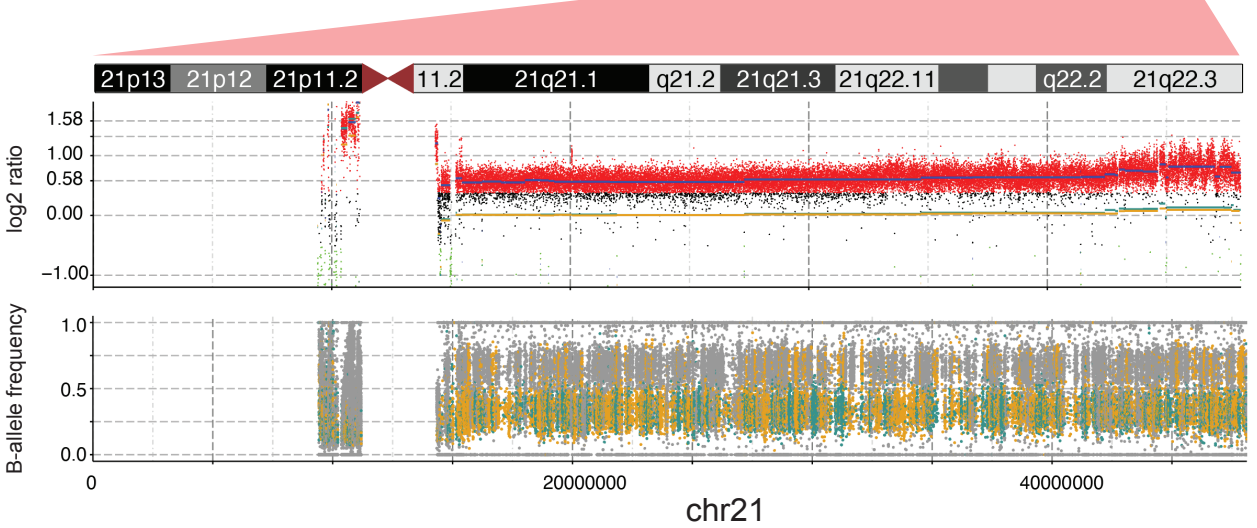
Figure S5. ***De novo*** **CGR in a female proband carrying Xq28 duplication including** ***MECP2*****. a)** Pedigree of the family. **b)** The top panel displays the $\log_2$ ratio segments for the proband (blue), father (teal), and mother (orange), showing a gain in $\log_2$ ratio only detected in the proband. The second panel presents phased B-allele frequencies (BAF) for the same region, with teal indicating paternal and orange indicating maternal origin. Notably, the proband exhibits an BAF of 0.66 for the paternal B-allele and 0.33 for the maternal B-allele, supporting a gain on the paternal X chromosome. Additionally, increased $\log_2$ ratio and BAF within chrX: 152,230,000-152,559,000 indicates higher copy-number suggestive of a CGR event. The third panel displays structural variant (SV) calls from Parliament2, which did not detect the Xq28 duplication likely due to the presence of a CGR (blue: deletion, brown: duplication, orange: inversion). The bottom three panels provide annotation tracks for RefSeq genes, segmental duplications (SegDup), and OMIM genes, respectively.

S6

a



b



c

BH10262-3	BH10262-2	BH10262-1	Phased-Baf$_{exp}$
0/0	0/1	0/0/1	0.33
0/1	0/0	0/0/1	0.33
0/0	1/1	0/1/1	0.66
0/1	0/1	0/1/1	0.66
0/1	0/1	0/1/1	0.66

d

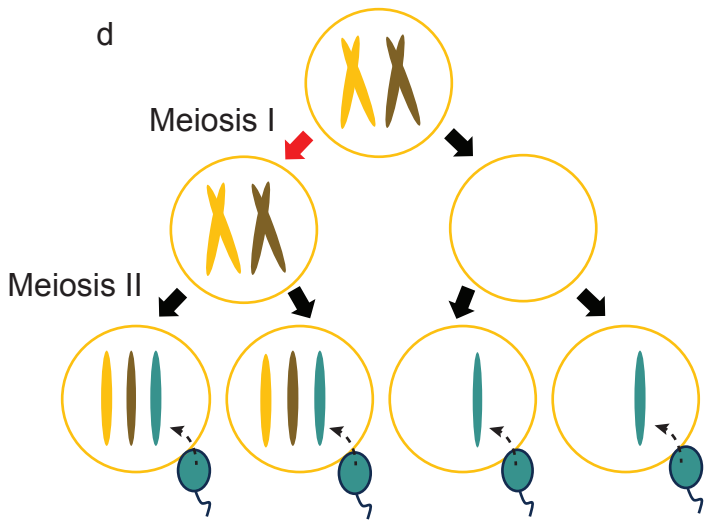
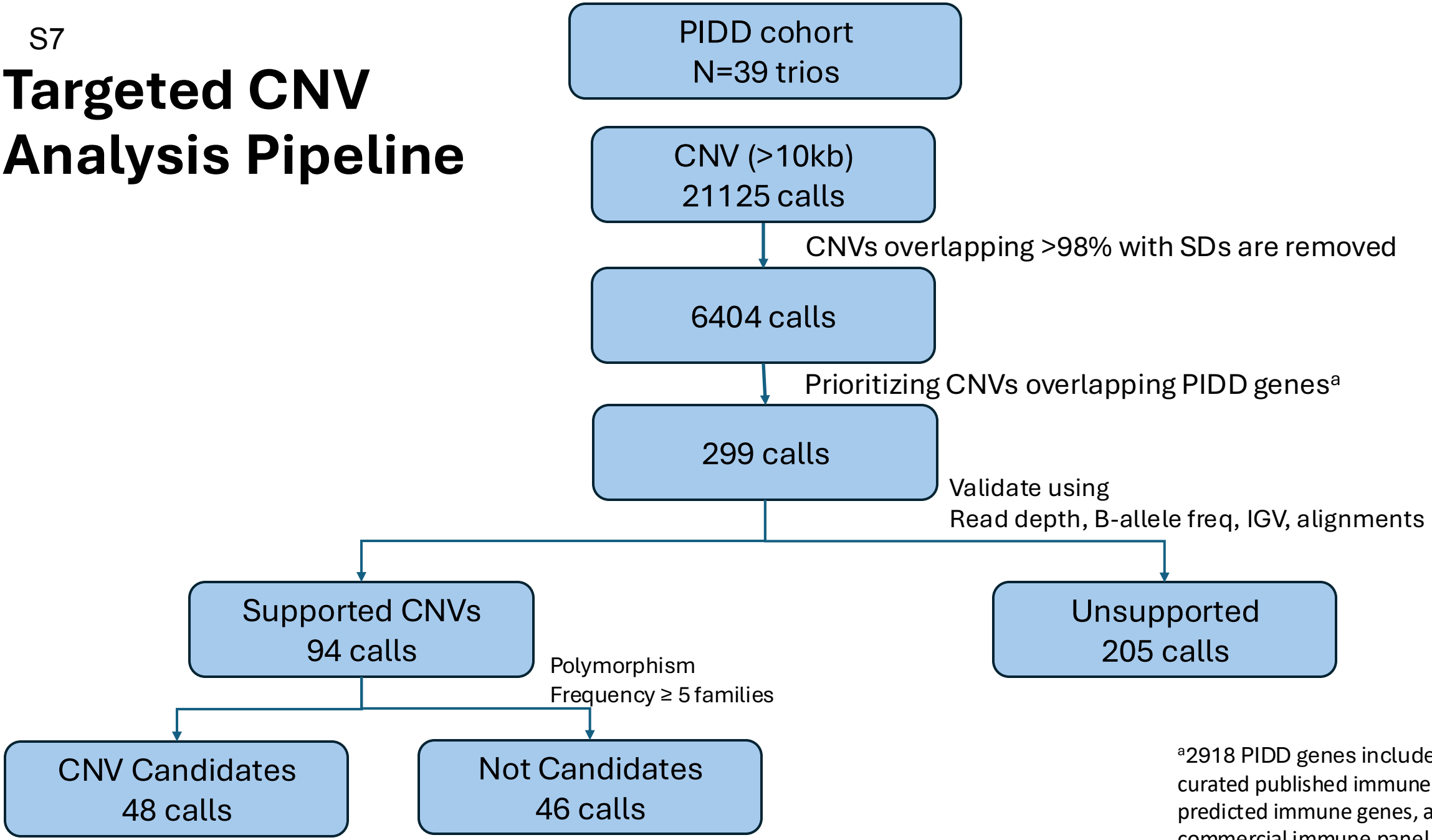


Figure S6. Genomic analysis of a PIDD family with trisomy 21. a) Genome-wide log_2 ratio plot with pedigree showing on the left. **b)** Detailed log_2 ratio and B allele frequency plots for chromosome 21 of the proband, BH10262-1. The top panel shows the log_2 ratio with regions of interest highlighted, and the bottom panel shows phased B allele frequency- color coding indicates allele origins: yellow for from maternal origin (BH10262-2), blue for paternal (BH10262-3), and grey for indeterminate. **c)** Table displays phasing schema for the BAF plot. **d)** Schematic representation of meiotic divisions I and II, illustrating chromosomal behavior and potential segregation patterns that could lead to the observed genetic variations in the family. Chromosome coloring corresponds to the B allele origins in **b) and c).**

S7

Targeted CNV Analysis Pipeline



[a]2918 PIDD genes include curated published immune genes, predicted immune genes, and commercial immune panel genes

Figure S7. Targeted CNV analysis pipeline for the PIDD cohort. Two filters have been used to reduce the number of calls for manual inspection. First, if more than 98% of the call is overlapping with SDs, the call will be filtered out. Second, we only prioritized calls that are overlapping with a predetermined PIDD-relevant gene list. The remaining calls were then manually validated. Finally, polymorphic calls were removed which provided 48 CNV candidates for further analysis.

S8

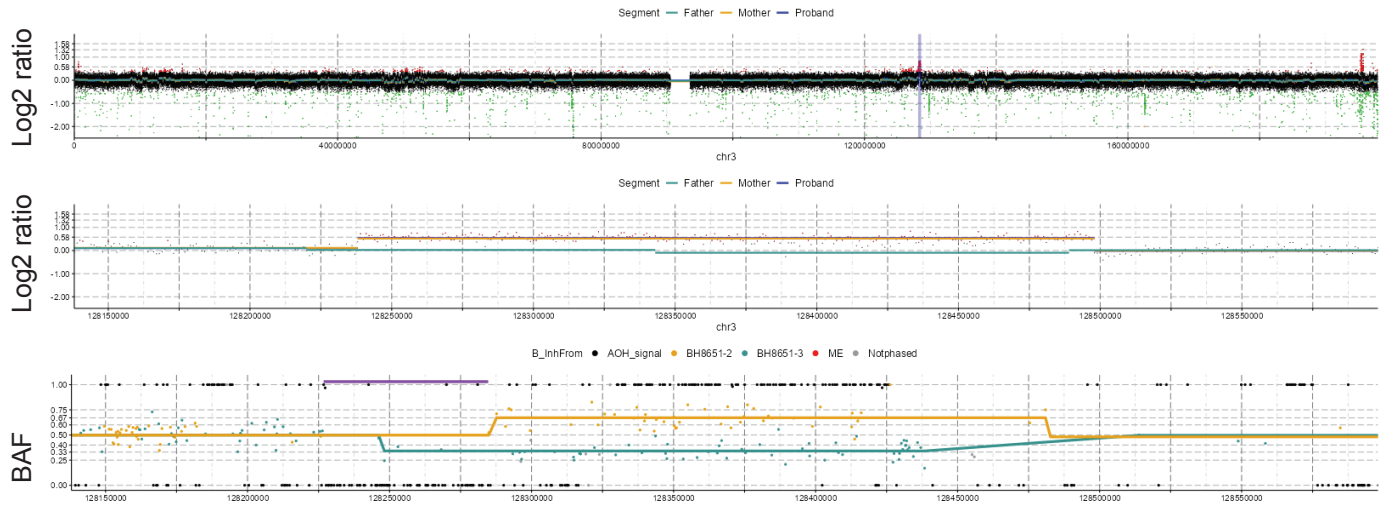
a

BH8651-1

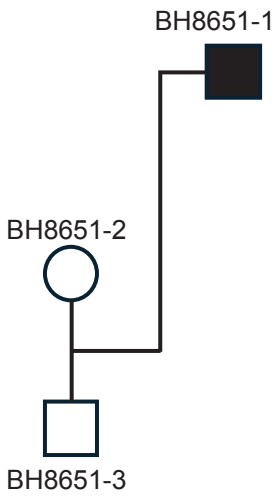
BH8651-2

BH8651-3

ROH



b



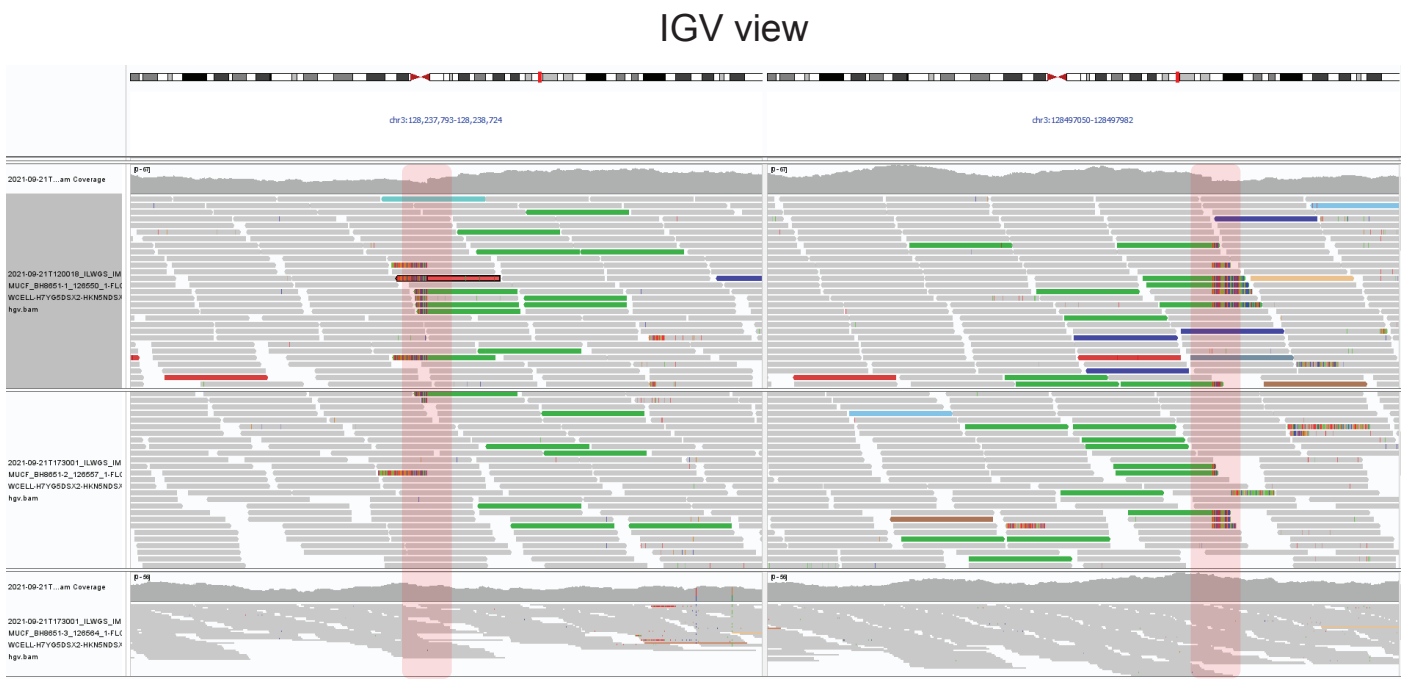
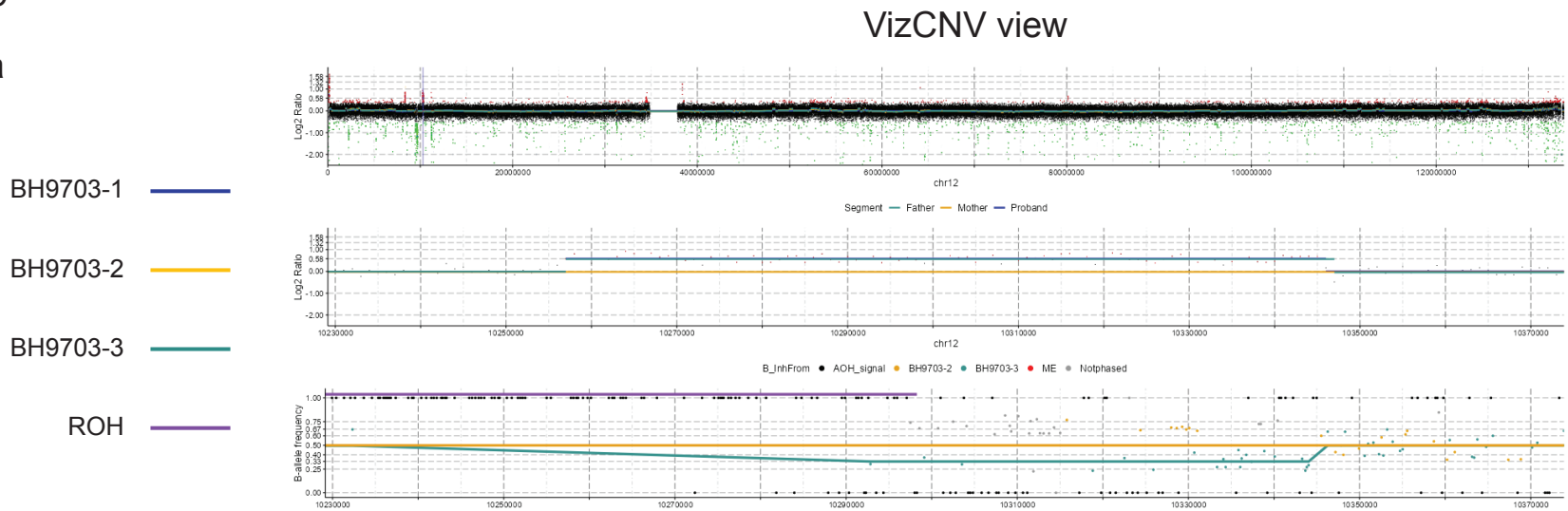


Figure S8. Representative Duplication with Consistent BAF and log_2 Analysis. a) The top panel displays the log_2 ratio of chromosome 3. The middle panel provides a zoomed-in view of the duplication region, while the bottom panel shows the BAF, with lines representing segmented BAF values. Colors represent segments from family members: blue for the proband, teal for the father, and yellow for the mother. Purple segments denote region with ROH call. **b**) The left side shows the pedigree, and the right side presents the IGV view of the duplication. Highlighted regions point to changes in read depth and several discordant read pairs suggesting a tandem duplication in both the proband and the mother. This case presents a maternally inherited duplication.

S9

a


VizCNV view
BH9703-1
BH9703-2
BH9703-3
ROH
Log2 Ratio
chr12
Segment Father Mother Proband
B_InhFrom AOH_signal BH9703-2 BH9703-3 ME Notphased
B-allele frequency

b

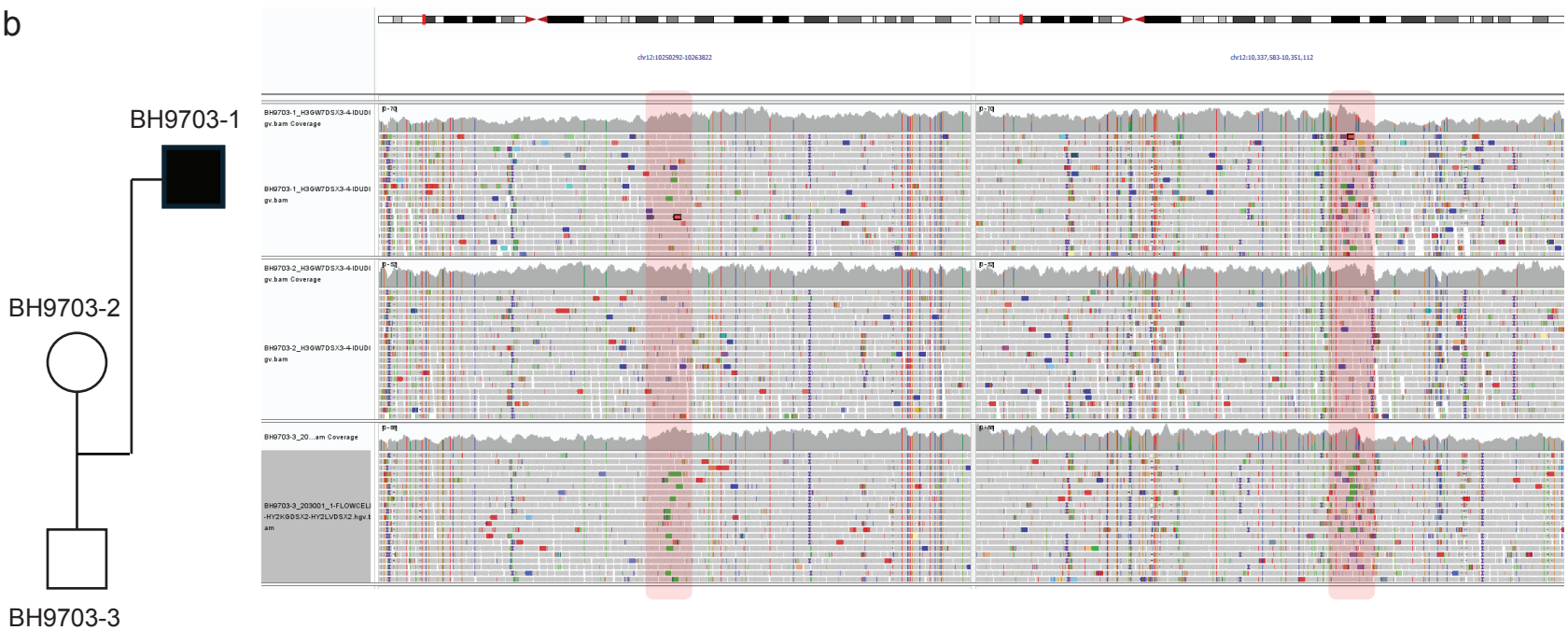

IGV view
BH9703-1
BH9703-2
BH9703-3

Figure S9. Paternally Inherited Duplication with inconsistent BAF origin. a) The top panel displays the $\log_2$ ratio of chromosome 12. The middle panel provides a zoomed-in view of the duplication region, and the bottom panel shows BAF plot, with lines representing segmented BAF values. Colors denote individuals from the family: blue for the proband, teal for the father, and yellow for the mother. **b**) The left side shows the pedigree, and the right side presents the IGV view of the duplication. Highlighted regions indicate changes in read depth and several discordant read pairs, consistent with a tandem duplication in both proband and father results although BAF distribution suggest a maternal inheritance.

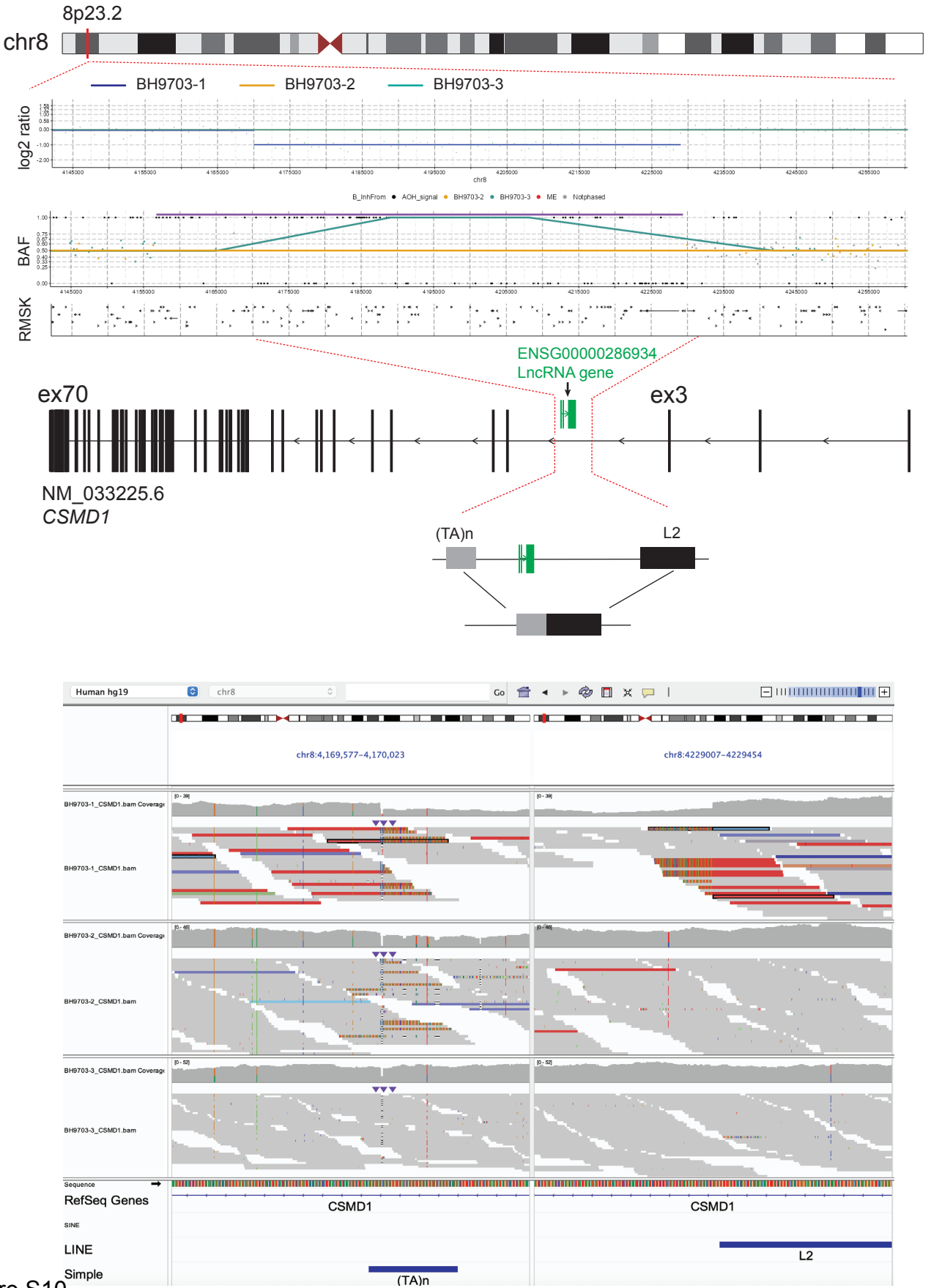

8p23.2
chr8
a
BH9703-1
BH9703-2
BH9703-3
log2 ratio
BAF
RMSK
B_InhFrom
AOH_signal
BH9703-2
BH9703-3
ME
Notphased
ENSG00000286934
LncRNA gene
b
ex70
ex3
NM_033225.6
CSMD1
(TA)n
L2
c
Human hg19
chr8
Go
chr8:4,169,577-4,170,023
chr8:4229007-4229454
BH9703-1_CSMD1.bam Coverage
BH9703-1_CSMD1.bam
BH9703-2_CSMD1.bam Coverage
BH9703-2_CSMD1.bam
BH9703-3_CSMD1.bam Coverage
BH9703-3_CSMD1.bam
Sequence
RefSeq Genes
CSMD1
CSMD1
SINE
LINE
L2
Simple
(TA)n

Figure S10

Figure S10. *De Novo* Deletion of LncRNA in Proband BH9703-1. a) Upper panel illustrates the ideogram of chromosome 8. The lower panel presents VizCNV output, displaying $\log_2$ ratio and B-allele frequency plots that identify a *de novo* deletion impacting *CSMD1*. **b)** The upper panel displays the canonical *CSMD1* transcript along with a smaller noncoding gene (highlighted in green), which is antisense to *CSMD1*. The deletion, within intron 3 of *CSMD1*, does not disrupt any exons but completely remove the entire long non-coding RNA gene. The diagram also provides a detailed view of the deletion's genomic structure. Note breakpoint junction involves a simple repeat and LINE element. **c)** IGV visualization of the deletion's breakpoint junction of SR-WGS.