Genome-wide profiling of highly similar paralogous genes using HiFi sequencing

Xiao Chen[1*], Daniel Baker[1], Egor Dolzhenko[1], Joseph M Devaney[2], Jessica Noya[2], April S Berlyoung[2], Rhonda Brandon[2], Kathleen S Hruska[2], Lucas Lochovsky[2], Paul Kruszka[2], Scott Newman[2], Emily Farrow[3,4,5], Isabelle Thiffault[3,4,6], Tomi Pastinen[3,4], Dalia Kasperaviciute[7], Christian Gilissen[8,9], Lisenka Vissers[8,9], Alexander Hoischen[8,9,10,11], Seth Berger[12], Eric Vilain[13], Emmanuèle Délot[13], UCI Genomics Research to Elucidate the Genetics of Rare Diseases (UCI GREGoR) Consortium, Michael A Eberle[1*]

1. PacBio, Menlo Park, CA, USA
2. GeneDx, Gaithersburg, MD, USA
3. Genomic Medicine Center, Children's Mercy Kansas City, Kansas City, MO, USA
4. UMKC School of Medicine, University of Missouri Kansas City, Kansas City, MO, USA
5. Department of Pediatrics, Children's Mercy Kansas City, Kansas City, MO, USA
6. Department of Pathology and Laboratory Medicine, Children's Mercy Kansas City, Kansas City, MO, USA
7. Genomics England Ltd., London, UK
8. Department of Human Genetics, Radboud University Medical Center, Nijmegen, The Netherlands
9. Research Institute for Medical Innovation, Radboud University Medical Center, Nijmegen, The Netherlands
10. Radboud Center for Infectious Diseases (RCI), Department of Internal Medicine, Radboud University Medical Center, Nijmegen, The Netherlands.
11. Radboud Expertise Center for Immunodeficiency and Autoinflammation and Radboud Center for Infectious Disease (RCI), Radboud University Medical Center, Nijmegen, The Netherlands
12. Center for Genetics Medicine Research, Children's National Hospital, Washington, DC, USA
13. Institute for Clinical and Translational Science, University of California, Irvine, CA, USA

Supplementary Notes

Comparison of Paraphase calls against HPRC assemblies

We compared Paraphase calls with high quality, phased diploid assemblies in 47 HPRC samples and summarized recall and precision, taking variant calls by the assembly as the truth. Paraphase calls were consistent with the assembly in the majority of paralog groups (defining the assembly as the ground truth, 82.4% of paralog groups have >95% recall and >95% precision) (Supplementary Fig. 1). The reduced precision and recall in some paralog groups are mostly due to errors in the assembly, as described below.

The low precision in some paralog groups appears to be driven by missing gene copies in the assemblies (examples shown in Supplementary Fig. 2, with well-supported Paraphase-called haplotypes missing any matching segment from the assembly). Since these are highly similar paralogous genes, some gene copies may have been collapsed in the assembly, leading to missing copies and incorrect copy number.

The low recall in some paralog groups such as *CDY1*, *SSX4*, *MBD3L2* and *SSX2* is related to the observation that sometimes a segment of a single HiFi read is directly incorporated into the assembly such that the assembly inherits all of the errors (mostly indel errors) from that single read (examples are shown in Supplementary Fig. 3, where the assembly appears to be consistent with a single read but inconsistent with all the other reads overlapping a position). As a result, the assembly calls many more indels than Paraphase, which does not have this problem as it takes the consensus sequence of all the reads that originate from a haplotype. This could be an area for improvement for future assemblers.

This comparison shows that the assembly can be inaccurate especially in regions where different copies of the segmental duplication are highly similar to each other.

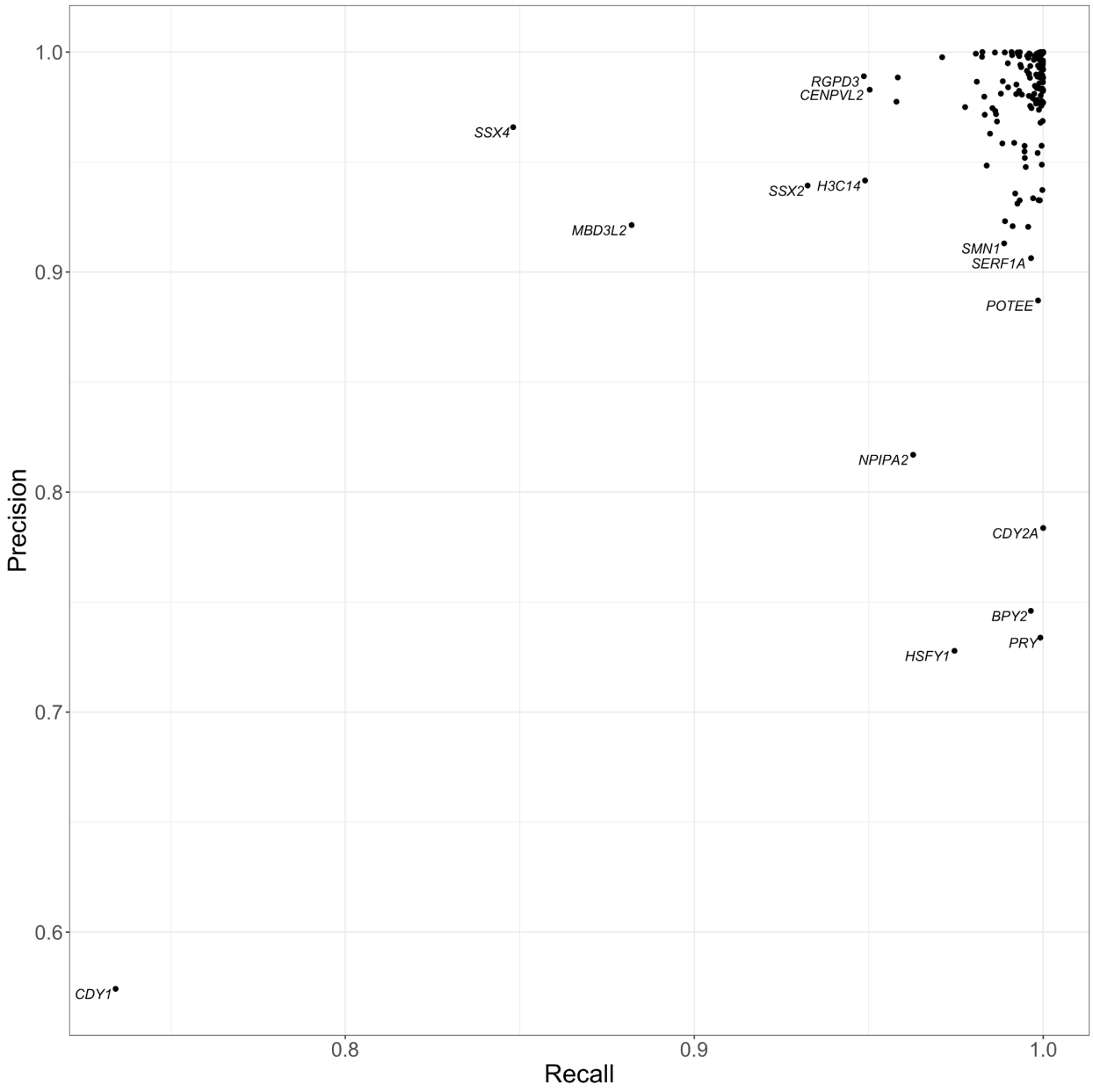
Simulation analysis

We simulated HiFi data at varied read length (5-15 kb), haplotype depth (5-15X) and paralog divergence level (0.01%-0.15%). The simulation result (Supplementary Fig. 4) indicates that at standard WGS parameters (10-15 kb read length and 10-15X haplotype depth) and when paralogs are reasonably diverged from each other (>0.05% divergence, i.e. 1 SNP every 2 kb), Paraphase can achieve high accuracy. Phasing accuracy could be impacted at short read length (5 kb), lower haplotype depth (5X) or low paralog divergence level (0.01%, i.e. 1 SNP every 10 kb). Real world data is of course more complicated, particularly by the fact that paralog divergence can be variable from region to region. There could exist local regions that are very scarce in base differences between haplotypes and are thus more vulnerable to lower read depth and shorter read length.

Supplementary figures

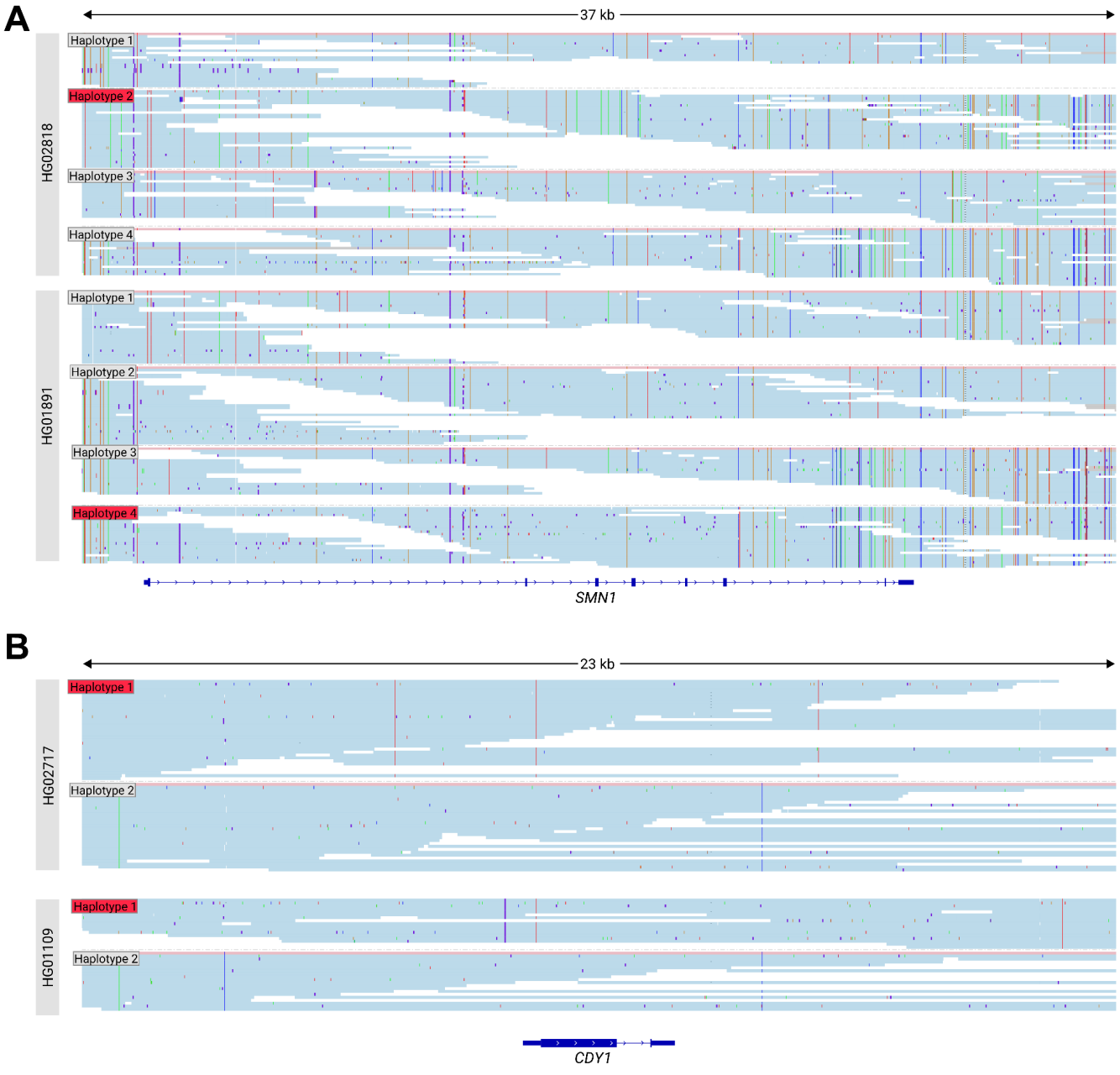
Supplementary Figure 1. Recall and precision of Paraphase variant calls against high quality diploid assemblies in 47 HPRC samples.

Each dot denotes the values of a paralog group.



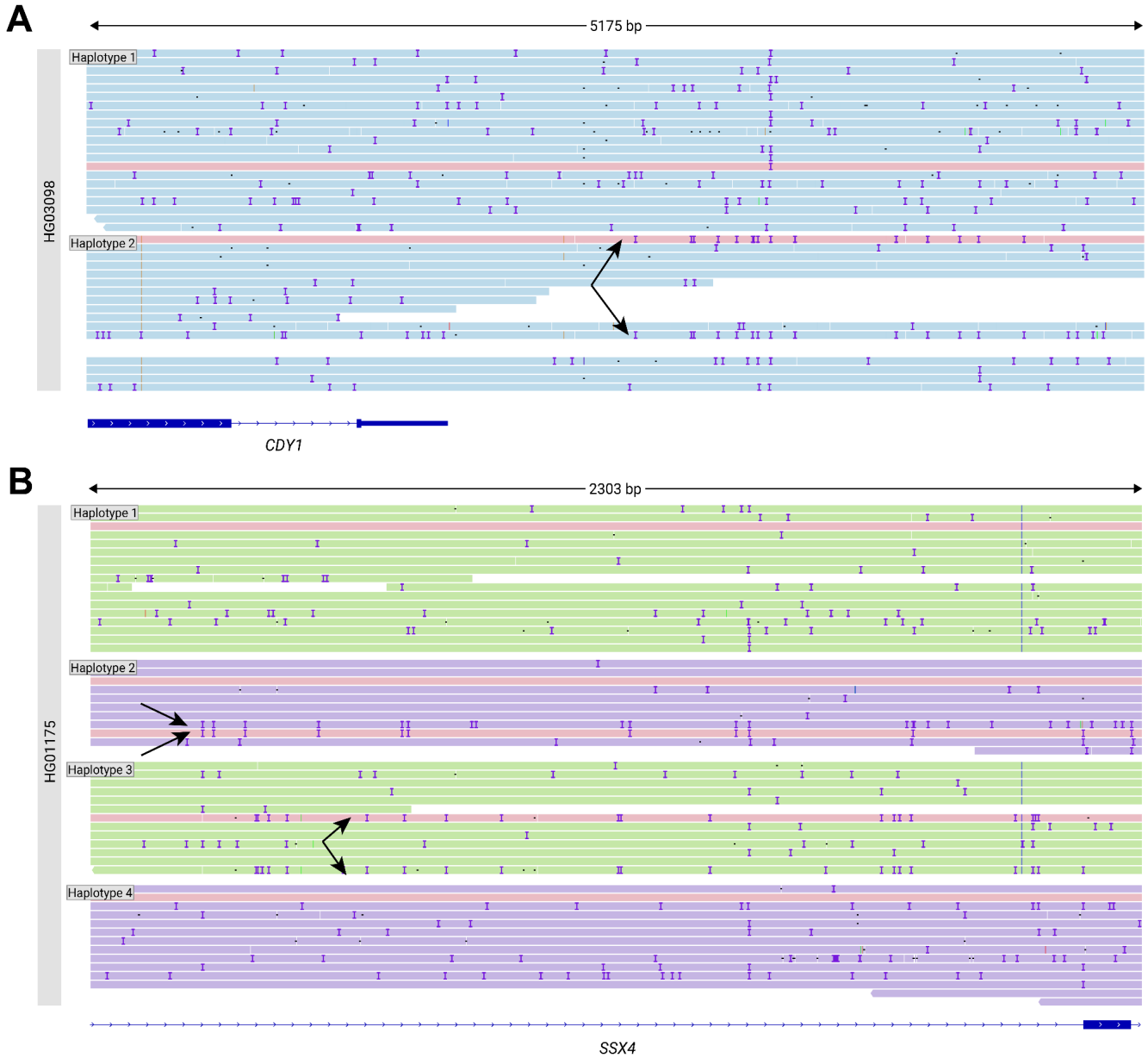
Supplementary Figure 2. Examples of missing gene copies in assemblies, leading to reduced precision of Paraphase in some paralog groups.

Two samples are shown for the *SMN1* paralog group (A) and two samples are shown for the *CDY1* paralog group (B). HiFi reads are grouped by the haplotypes called by Paraphase. Segments from contigs of the assemblies are shown together with the haplotypes they correspond to and are colored in pink. Each sample has one Paraphase-called haplotype (labeled in red) that does not have any matching segment in the assemblies. In each sample, there are no additional contigs in the assembly that can align to the target region.



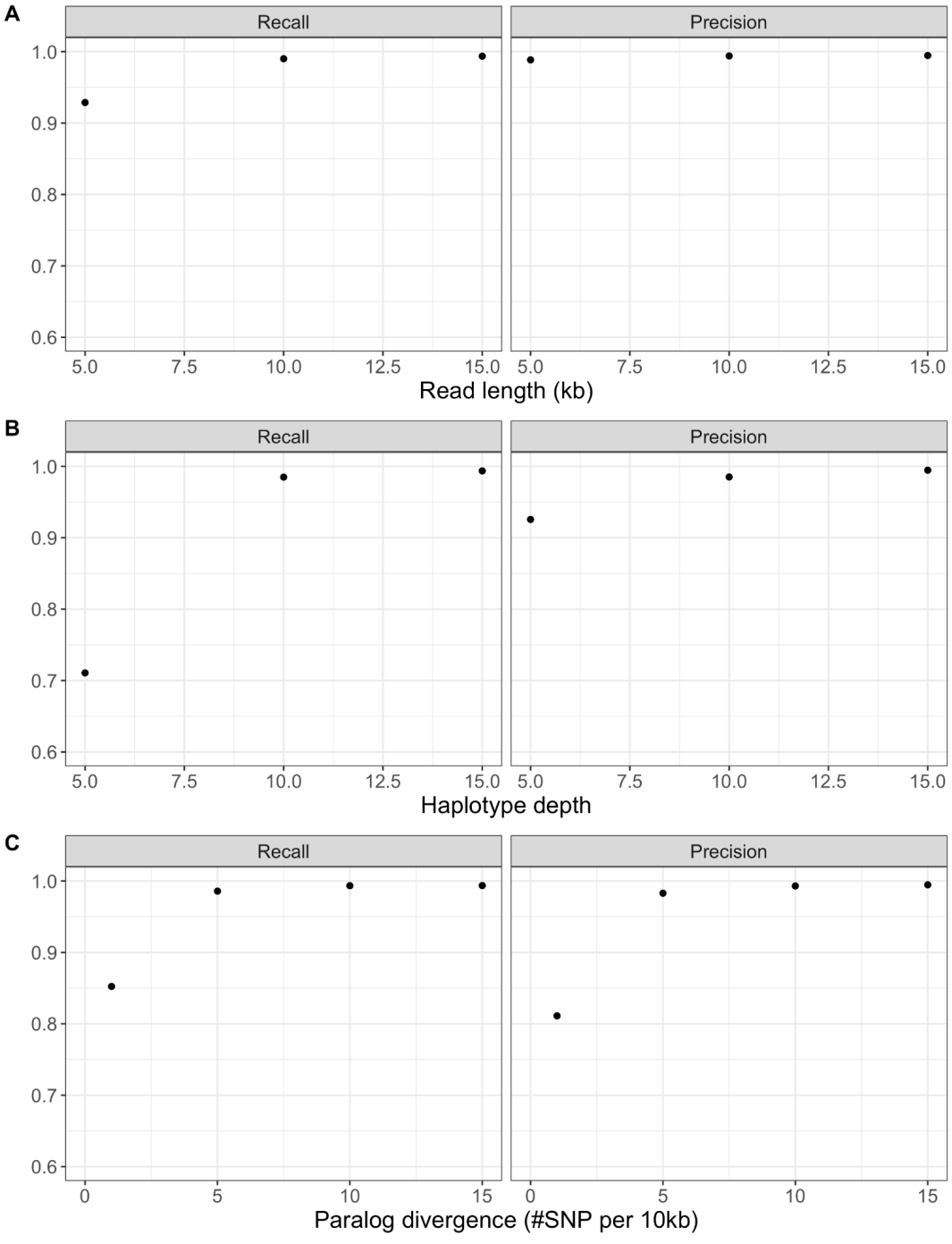
Supplementary Figure 3. Examples of base errors in assemblies, leading to reduced recall of Paraphase in some paralog groups.

One sample is shown for the *CDY1* paralog group (A) and one sample is shown for the *SSX4* paralog group (B). HiFi reads are grouped by the haplotypes called by Paraphase (haplotypes from two alleles are colored in green and purple for *SSX4*). Segments from contigs of the assemblies are shown together with the haplotypes they correspond to and are colored in pink. For each haplotype, Paraphase takes the consensus sequence of all reads that originate from it and it is expected that the consensus matches the assembly. Black arrows indicate contigs (pink) from the assemblies that have incorporated sequencing errors, mostly indels, from a single HiFi read (non-pink). This leads to reduced recall of Paraphase as Paraphase does not call any of those base errors by taking the consensus of a group of reads.



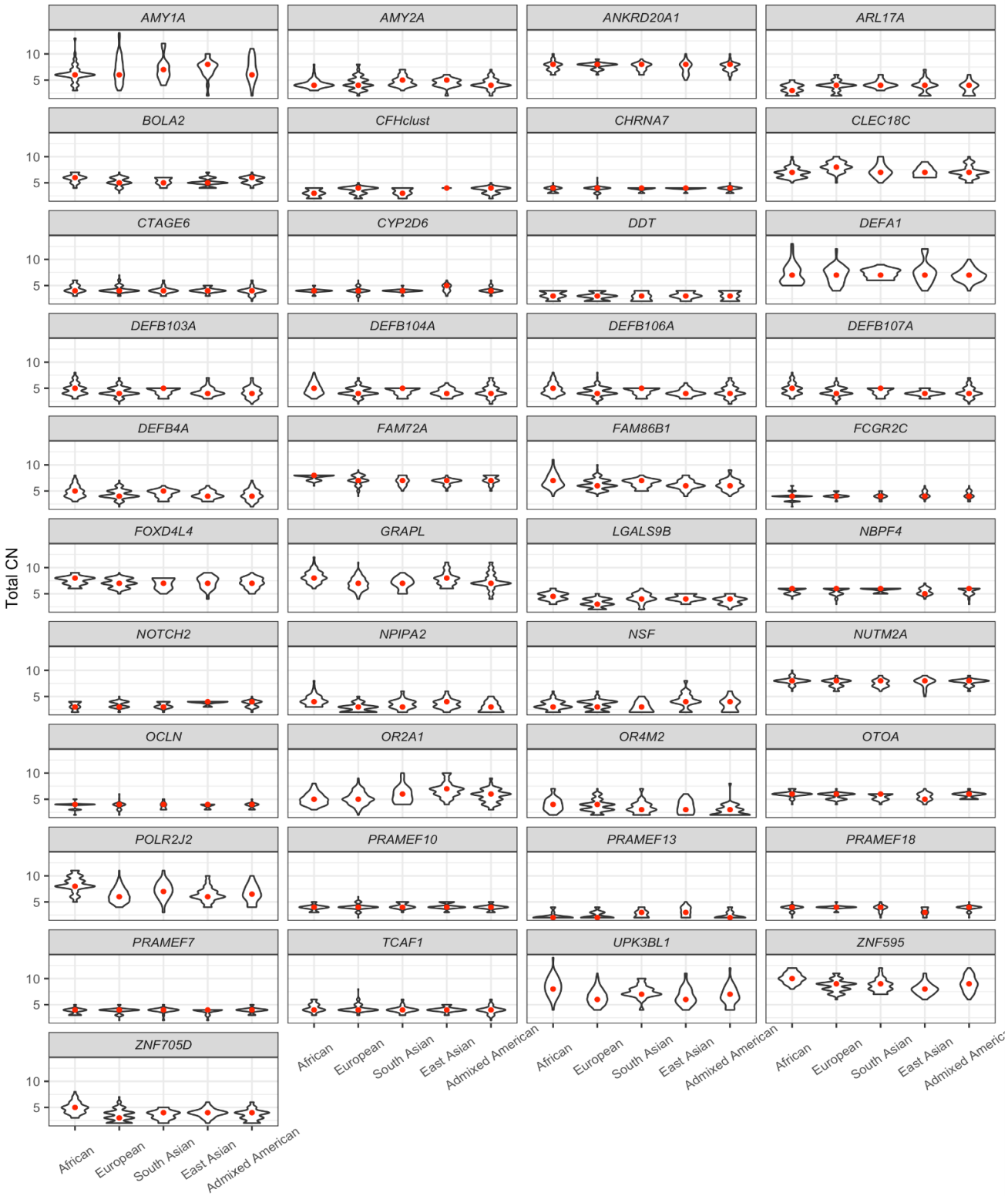
Supplementary Figure 4. Paraphase performance on simulation data of varied read length (baseline 15 kb), haplotype depth (baseline 15X) and paralog divergence (baseline 0.15%).

Paraphase calls were examined for variant recall and precision at varied read lengths of 5 kb, 10 kb and 15 kb (A), varied haplotype depths of 5X, 10X and 15X (B) and varied paralog divergence of 1, 5, 10 and 15 SNPs per 10 kb (0.01%, 0.05%, 0.1% and 0.15%, respectively) (C), while the other two variables are kept at baseline for each panel.



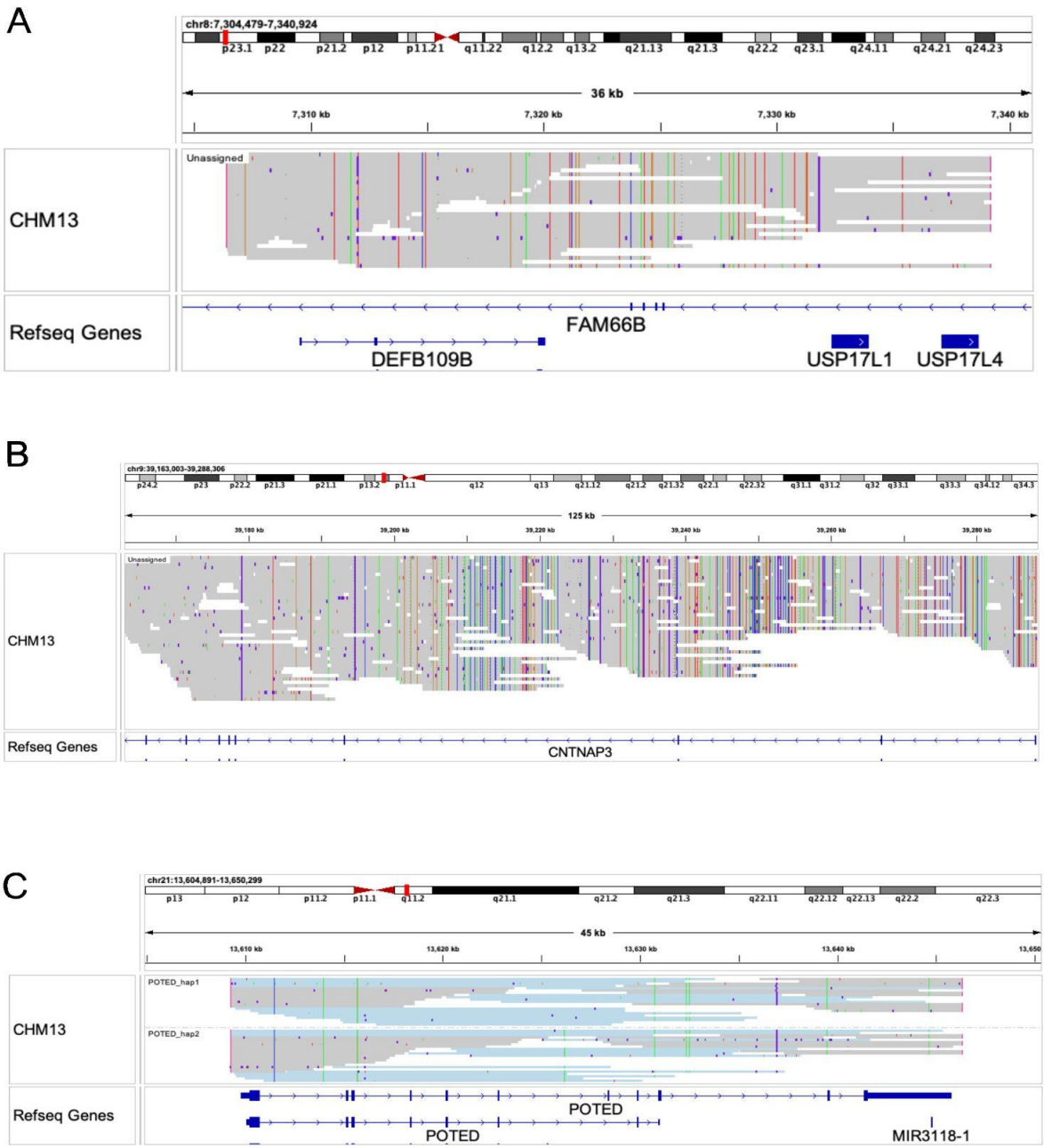
Supplementary Figure 5. Paralog groups with significant copy number (CN) deviations across populations.

Violin plots for the total copy number of paralog groups across individuals of five ancestral populations. These paralog groups have significant (Chi-squared test, $p<0.05$, with Bonferroni correction) deviations between populations. One archetype gene is selected to represent the name of each paralog group.



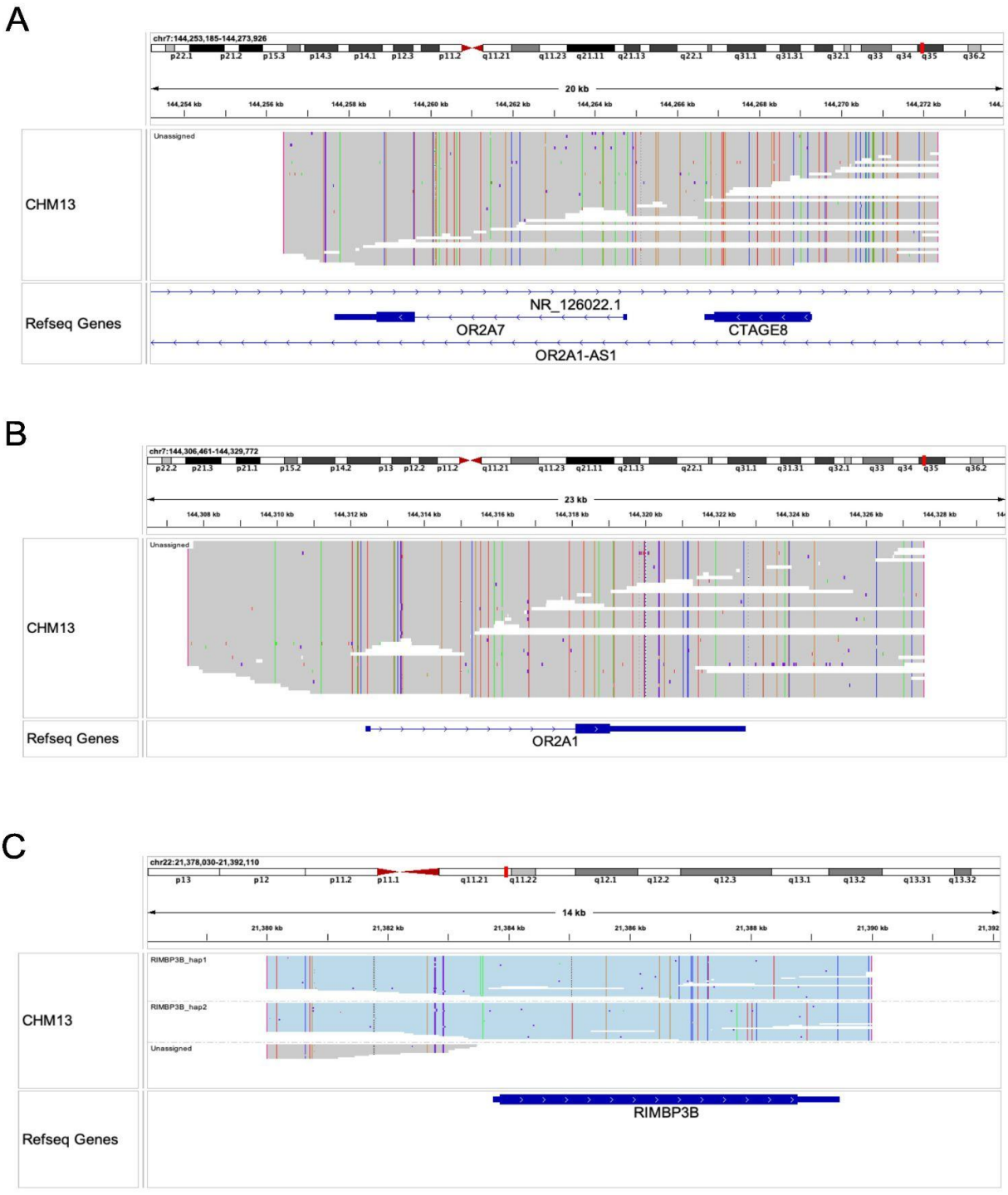
Supplementary Figure 6. Paraphase results with CHM13 data in GRCh38 false duplication regions identified by this study.

These regions are new false duplication regions that we found in GRCh38 (population samples show a consistent copy number of two) and were not previously found by the CHM13 T2T assembly. Paraphase analysis of CHM13 data shows one copy in *DEFB109B* and its SD (A), one copy in *CNTNAP3/CNTNAP3C* (B) and two copies that differ by a single SNP in *POTED* and its SD (C), where the additional haplotype could be due to mosaicism or cell line artifacts, as the coverage of this region is similar to the genome average.



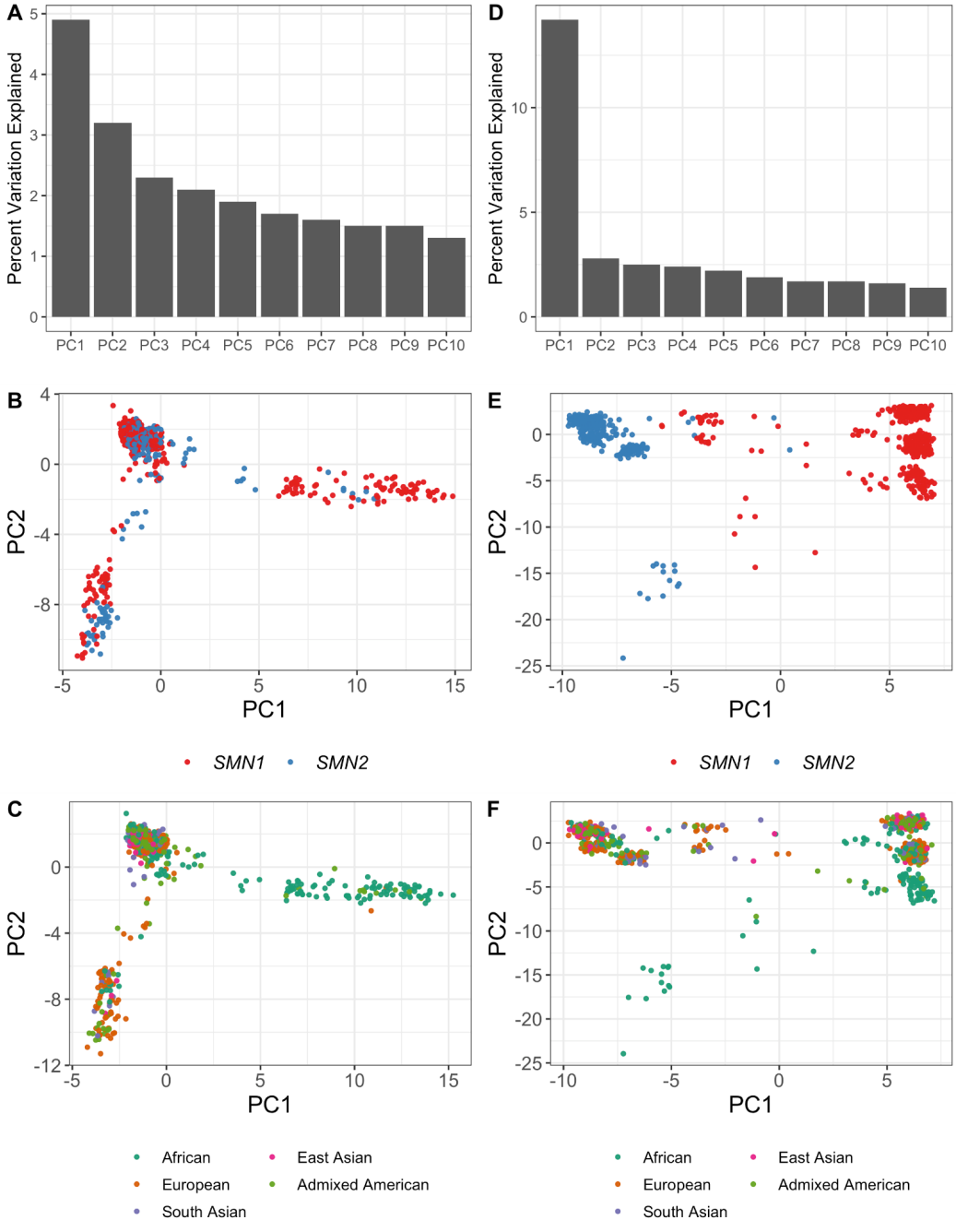
Supplementary Figure 7. Paraphase results with CHM13 data in GRCh38 regions that were previously erroneously assigned to false duplications.

Compared to GRCh38, these regions have one copy less in CHM13 and were thus previously assigned to false duplications in GRCh38 by the CHM13 T2T assembly, but they are truly variable in copy number across populations. Paraphase analysis of CHM13 data shows one copy in *CTAGE8/CTAGE9* (A), one copy in *OR2A1/OR2A42* (B) and two copies in *RIMBP3/RIMBP3B/RIMBP3C* (C).



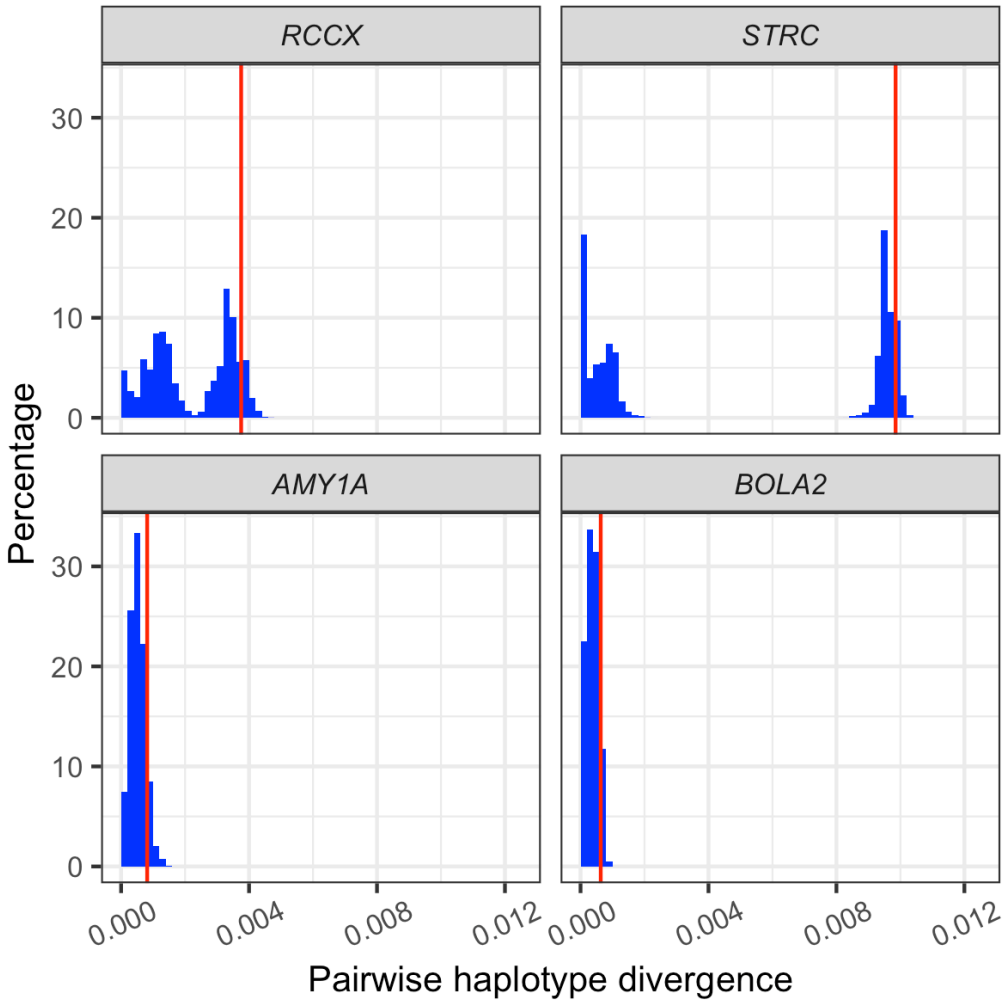
Supplementary Figure 8. Principal component analysis (PCA) of Exons 1-6 and Exons 7-8 sequences in *SMN1/SMN2*.

Exons 1-6 (A-C) sequences are highly similar to each other, with little variation explained by each PC (A). *SMN1* and *SMN2* haplotype sequences do not separate into distinct clusters (B). Instead, the separation on PC1 is by ancestry (C). Exons 7-8 (D-F) are more differentiated, with a much higher level of variation explained by PC1 (D), which separates haplotype sequences by gene (E) rather than ancestry (F).



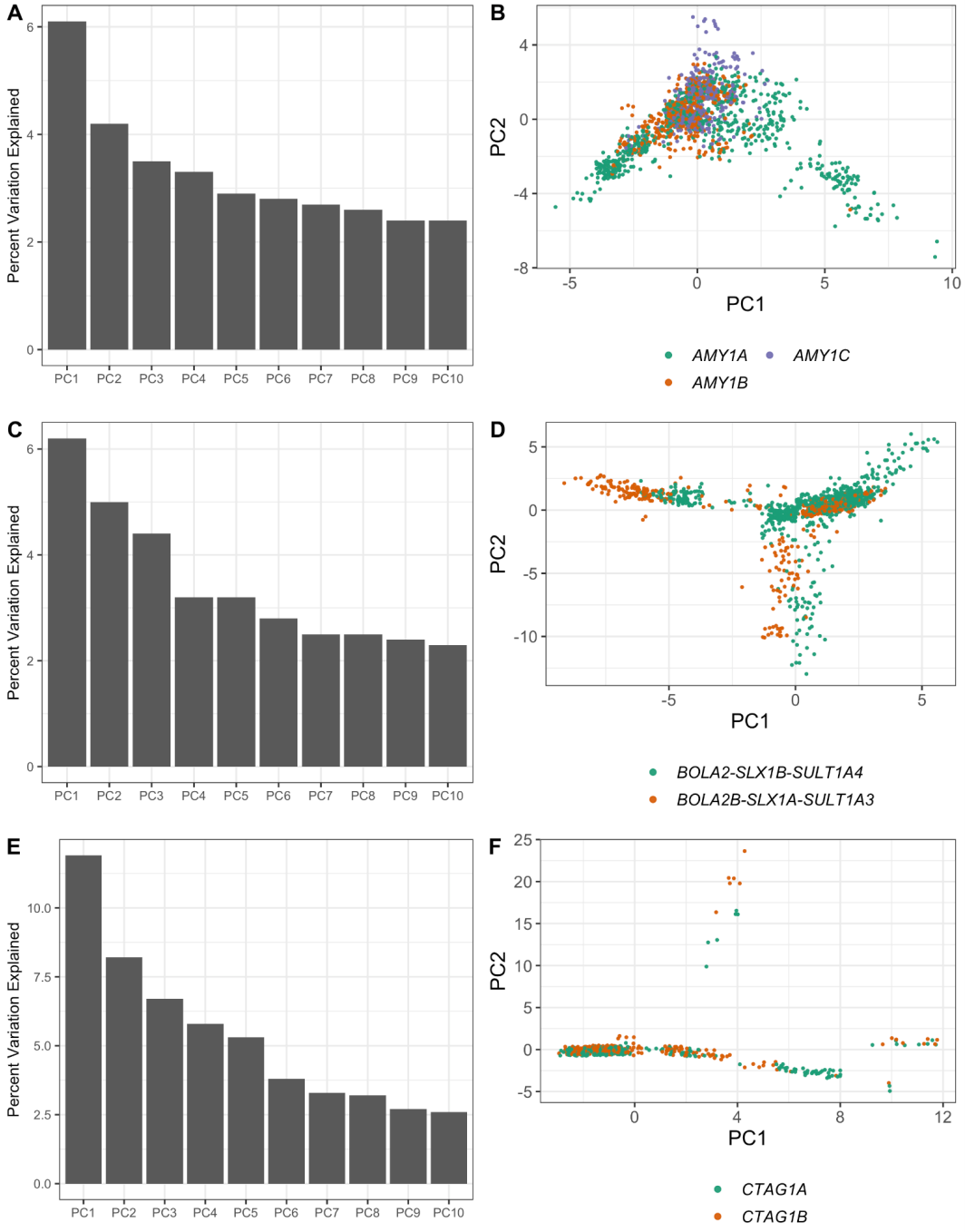
Supplementary Figure 9. Histograms of pairwise sequence divergence for paralog groups with differentiated genes and paralogs (upper panel) vs. paralog groups with low within-group diversity (lower panel).

Red vertical lines represent the 90th percentile. With differentiated genes in the same paralog group (upper panel), we expect to see two distinct peaks representing low gene-gene (and paralog-paralog) sequence divergence and high gene-paralog sequence divergence. Conversely, in paralog groups with low within-group diversity (lower panel), gene-gene (and paralog-paralog) sequence divergence is similar to gene-paralog sequence divergence and they do not form two peaks. RCCX contains *CYP21A2/CYP21A1P*, *C4A/C4B* and *TNXB/TNXA*, genotyped as one region by Paraphase.



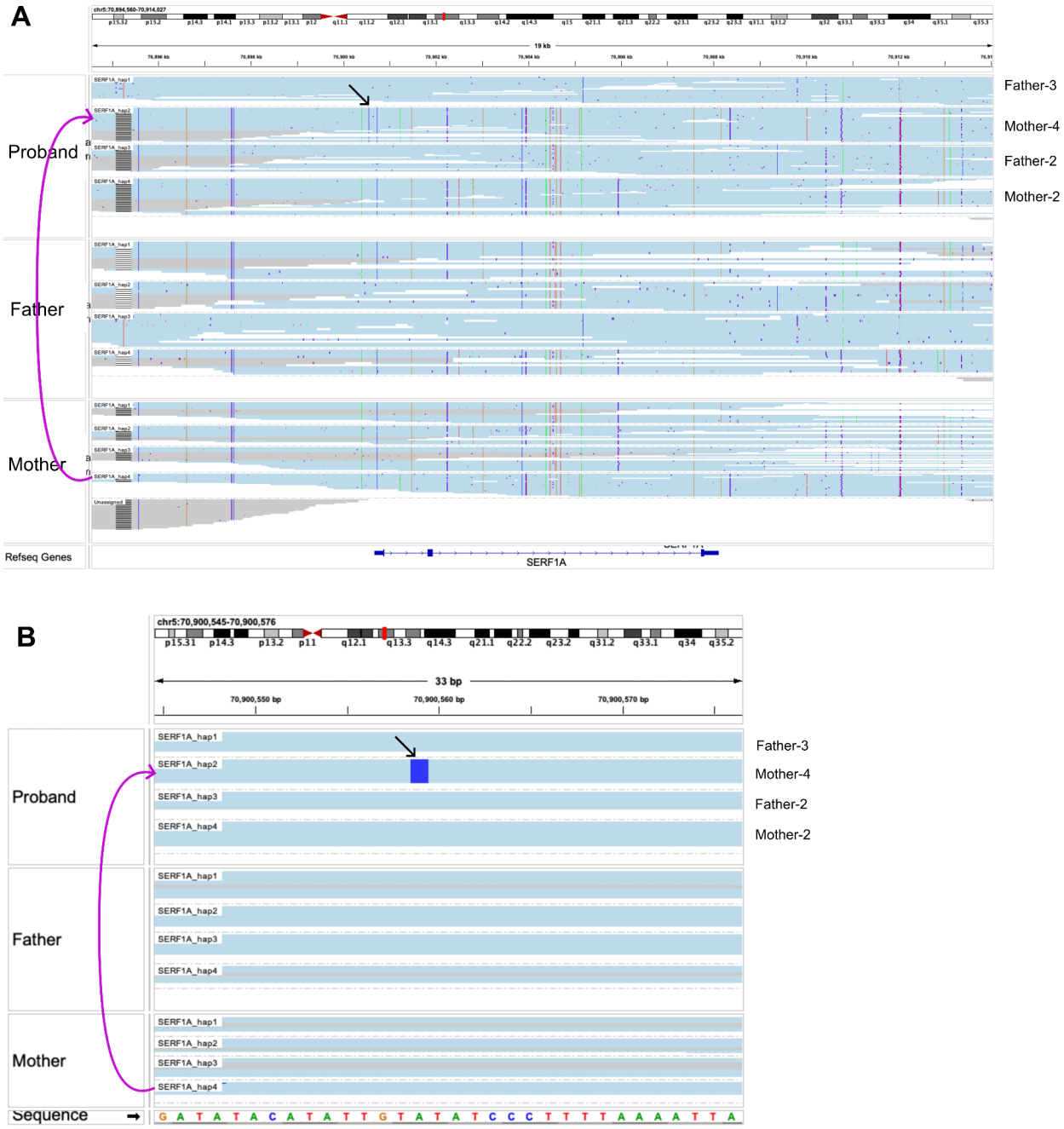
Supplementary Figure 10. Details of PCA results shown in Figure 3B-D.

Each row shows the PCA results (Percent variation explained, left, and PC1 vs. PC2 plot, right) for one paralog group: *AMY1A/AMY1B/AMY1C* (A-B); *BOLA2-SLX1B-SULT1A4/BOLA2B-SLX1A-SULT1A3* (C-D) (three paralog groups in tandem and genotyped as one region by Paraphase); *CTAG1A/CTAG1B* (E-F).



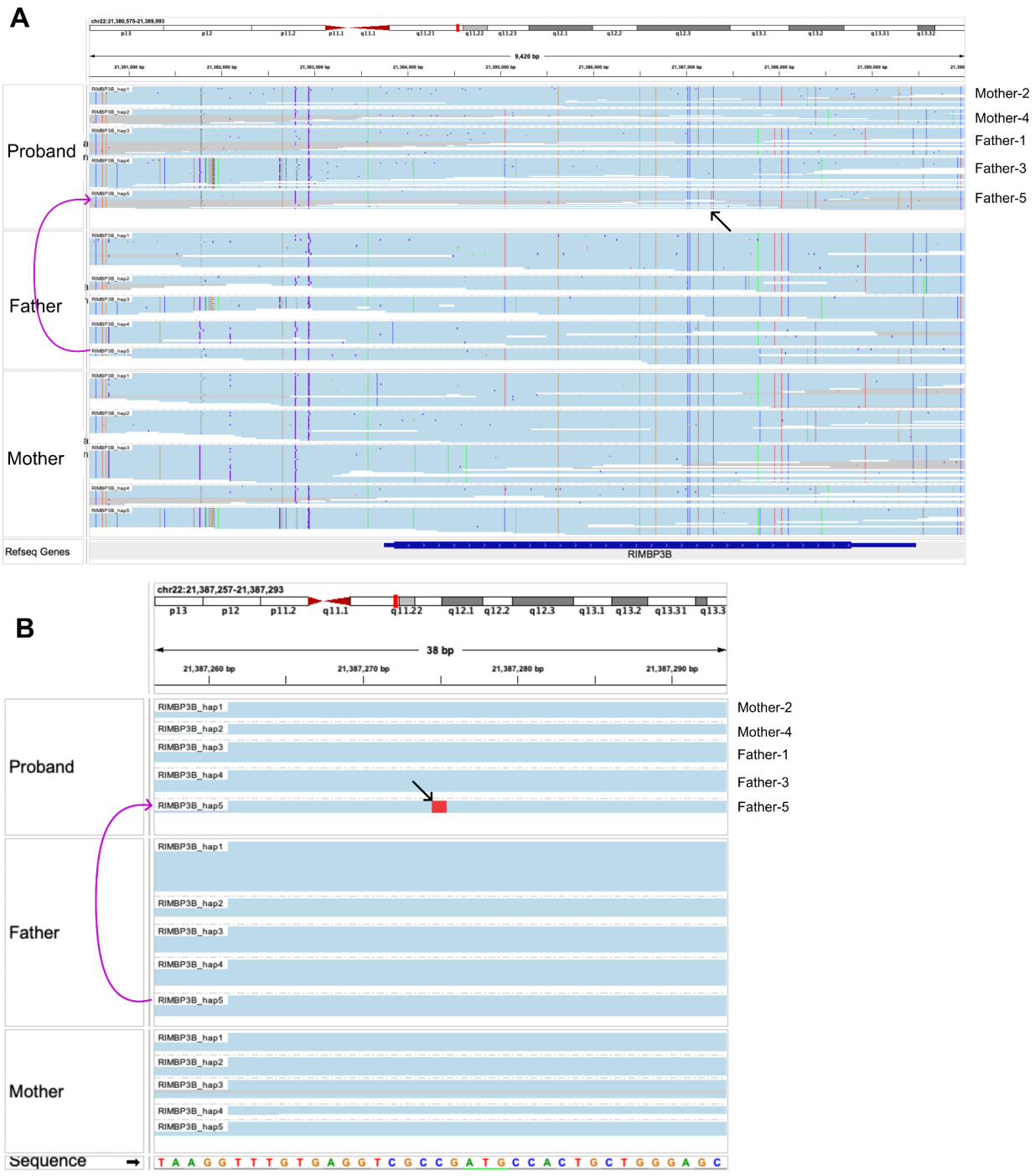
Supplementary Figure 11-1. *De novo* SNV #1.

Variant is intergenic. Variant is marked by the black arrow. B is a close view of the variant.



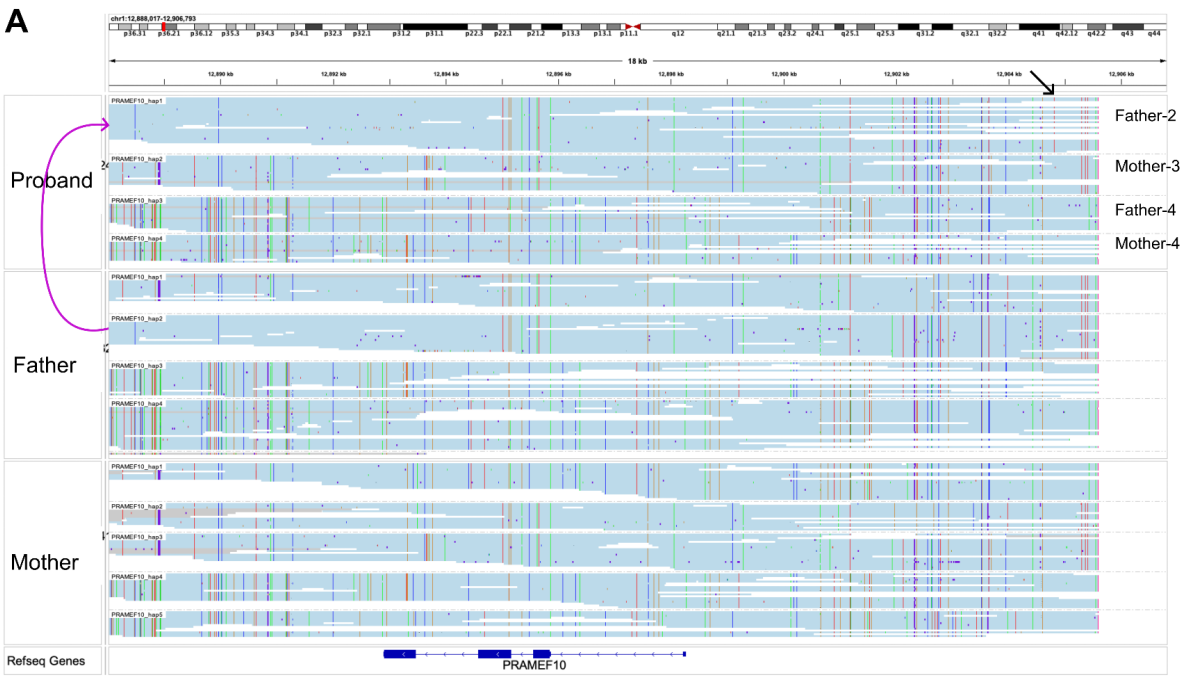
Supplementary Figure 11-2. *De novo* SNV #2.

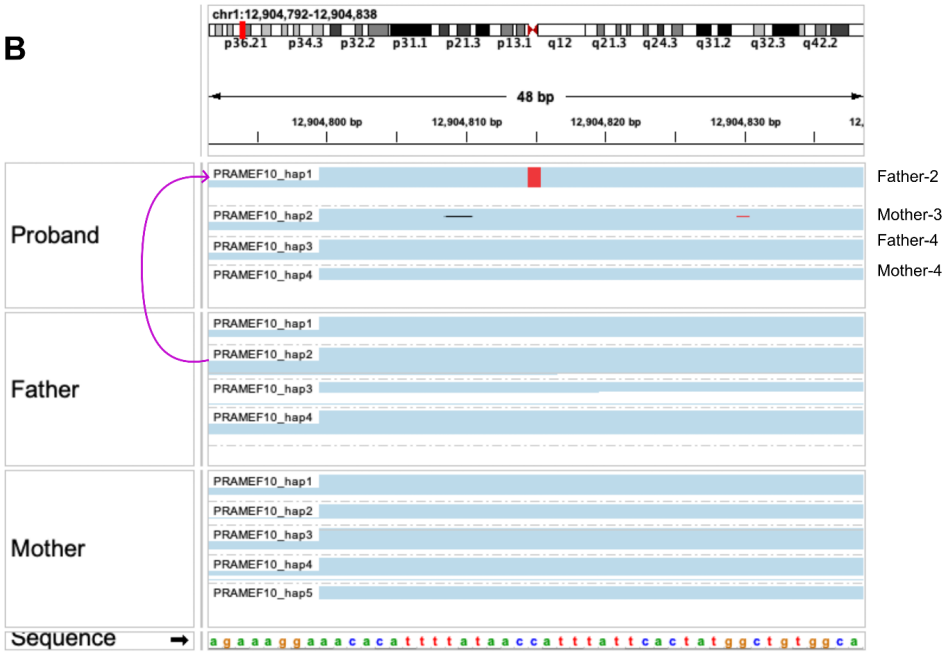
Variant is synonymous. Variant is marked by the black arrow. B is a close view of the variant.



Supplementary Figure 11-3. *De novo* SNV #3.

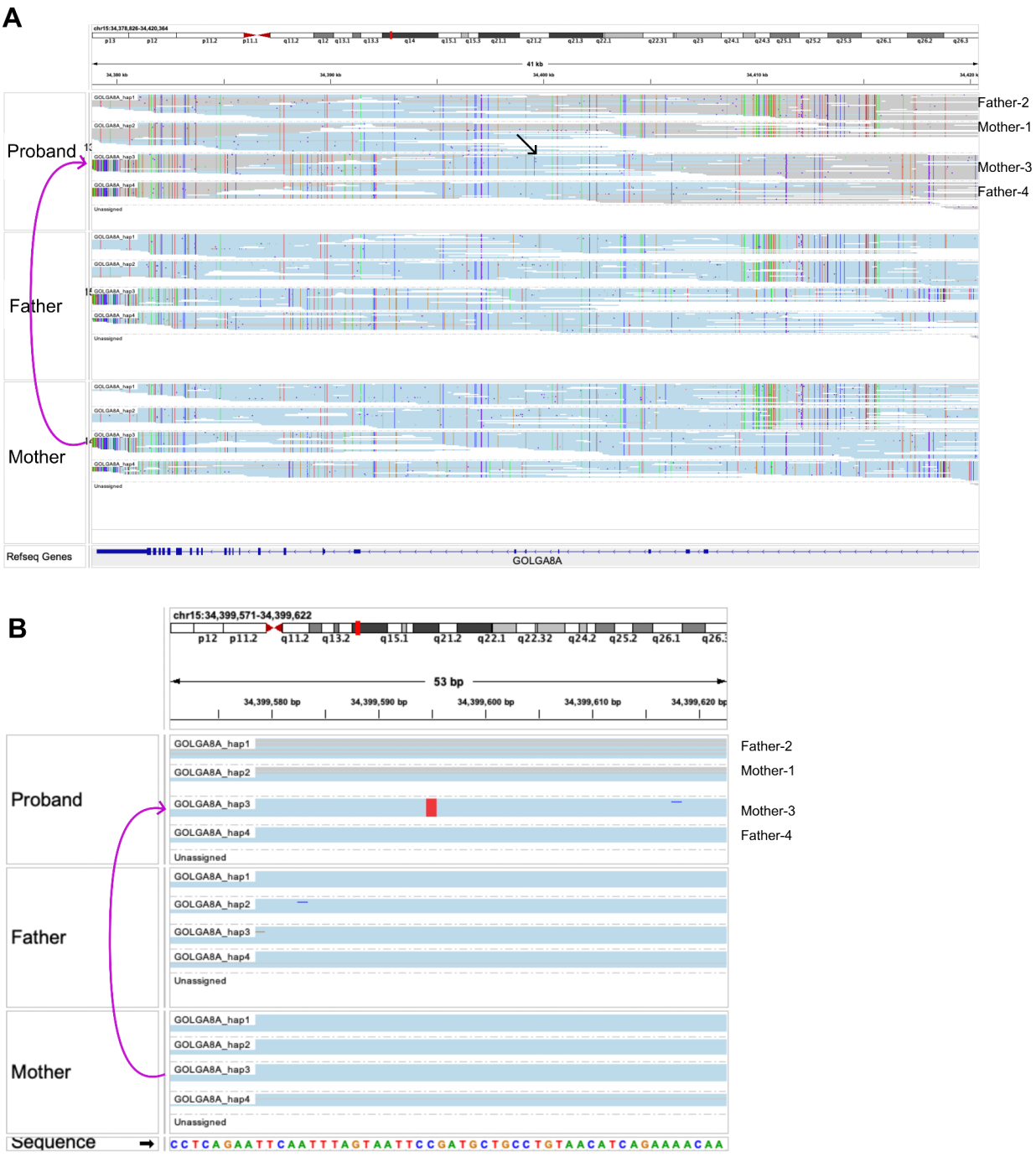
Variant is intergenic. Variant is marked by the black arrow. B is a close view of the variant.





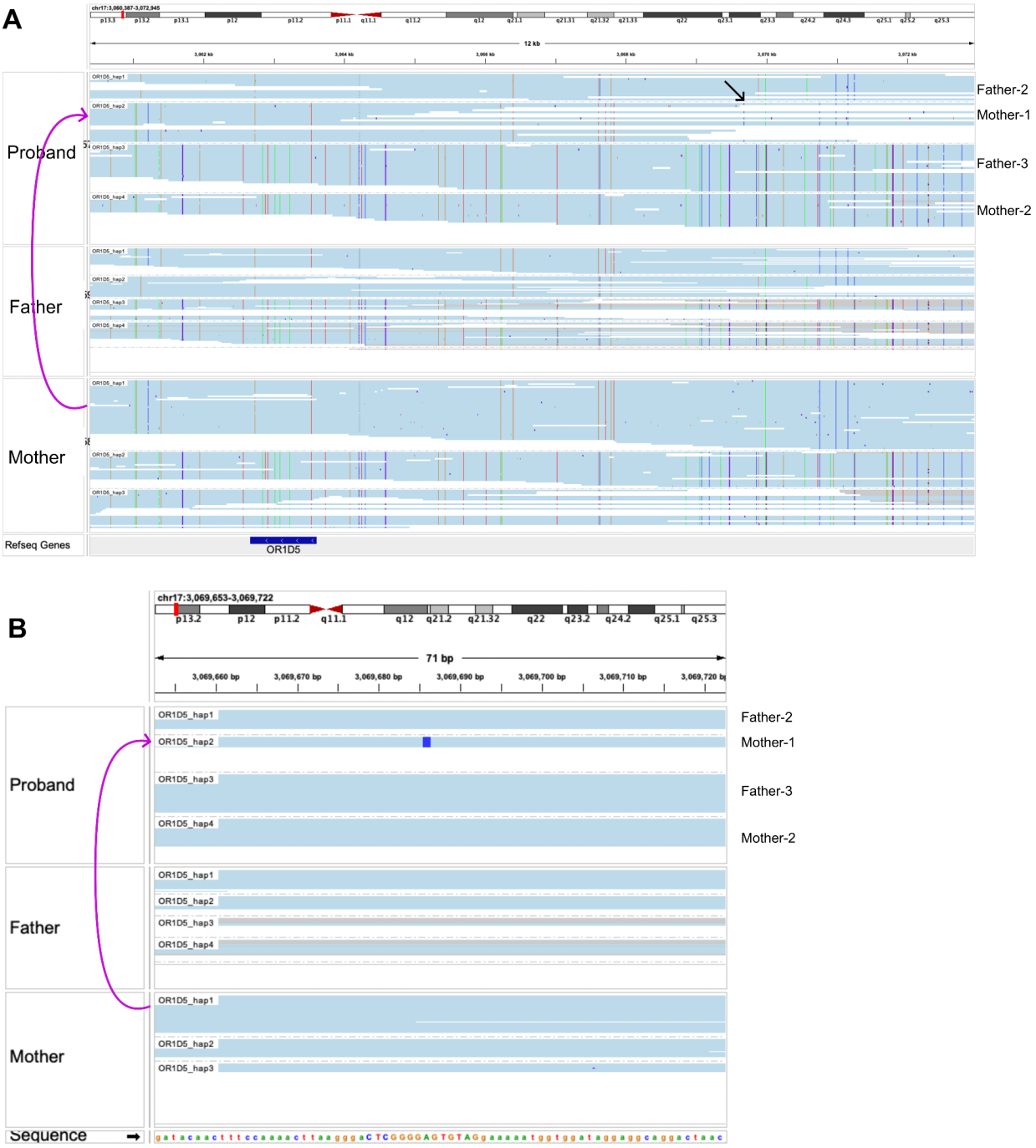
Supplementary Figure 11-4. *De novo* SNV #4.

Variant is intronic. Variant is marked by the black arrow. B is a close view of the variant.



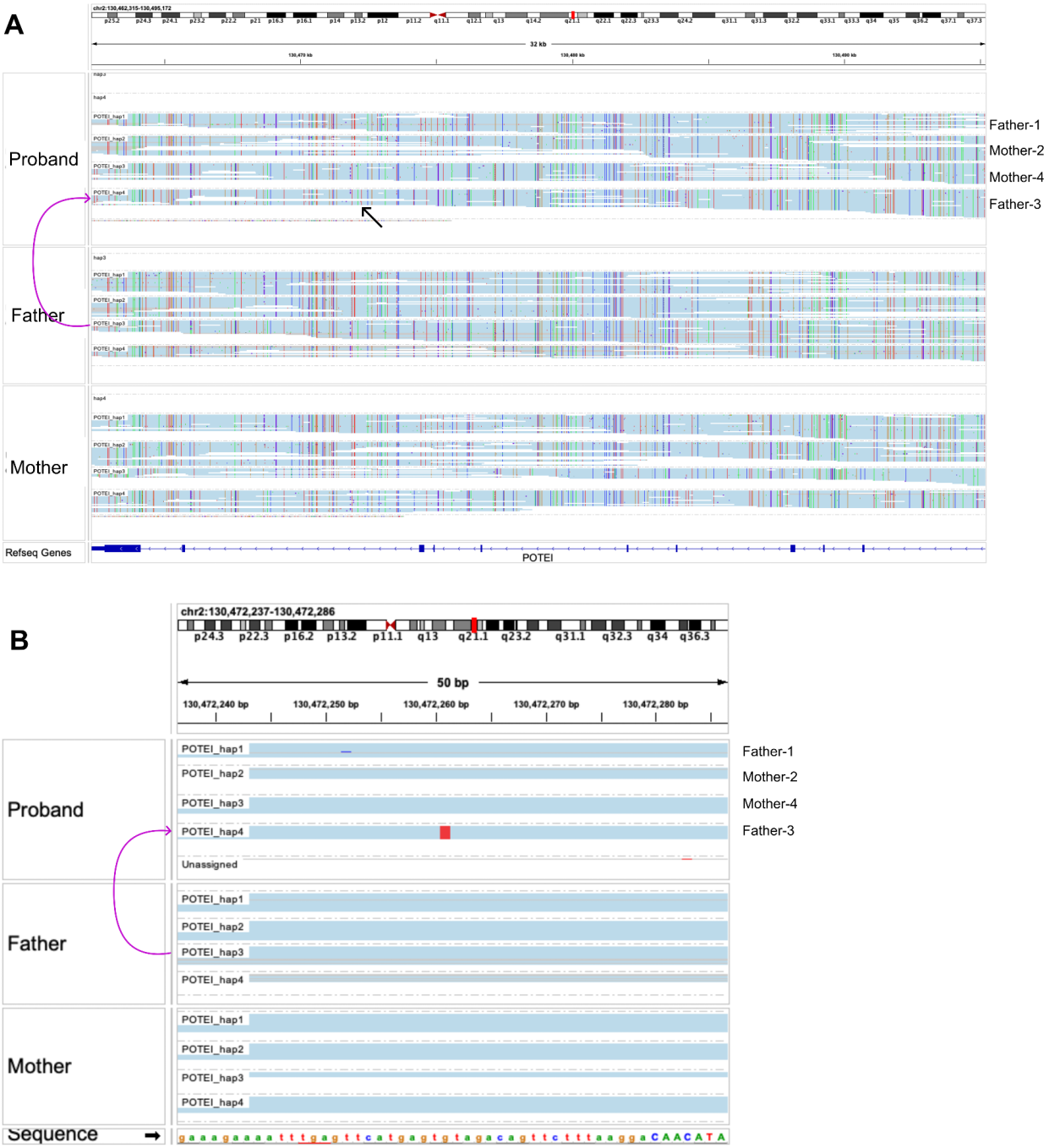
Supplementary Figure 11-5. *De novo* SNV #5.

Variant is intergenic. Variant is marked by the black arrow. B is a close view of the variant.



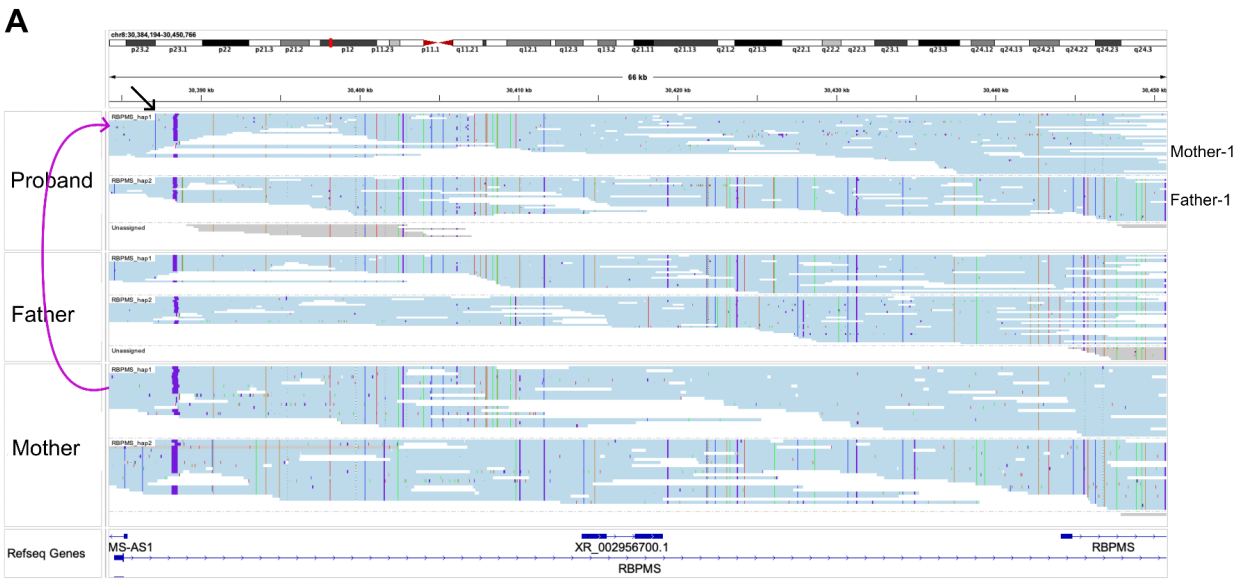
Supplementary Figure 11-6. *De novo* SNV #6.

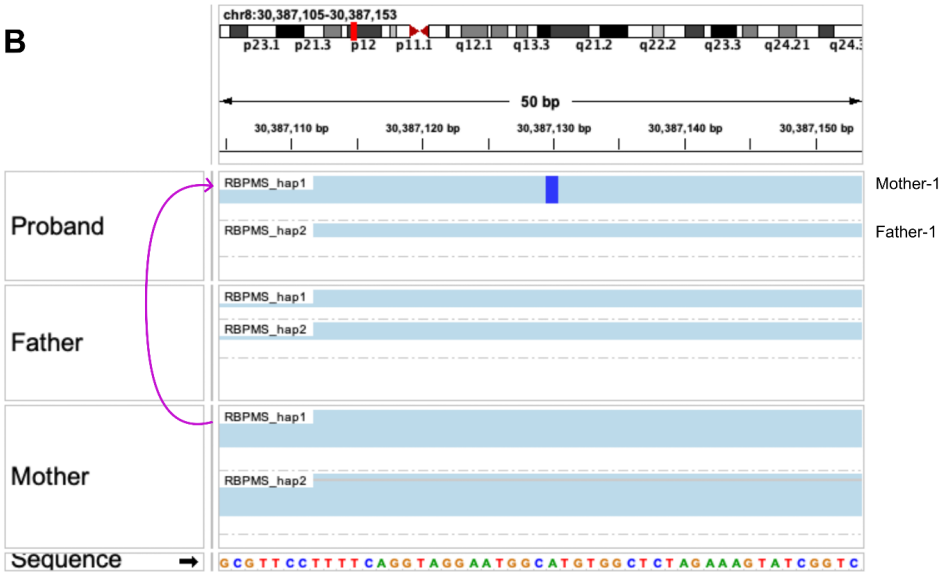
Variant is intronic. Variant is marked by the black arrow. B is a close view of the variant.



Supplementary Figure 11-7. *De novo* SNV #7.

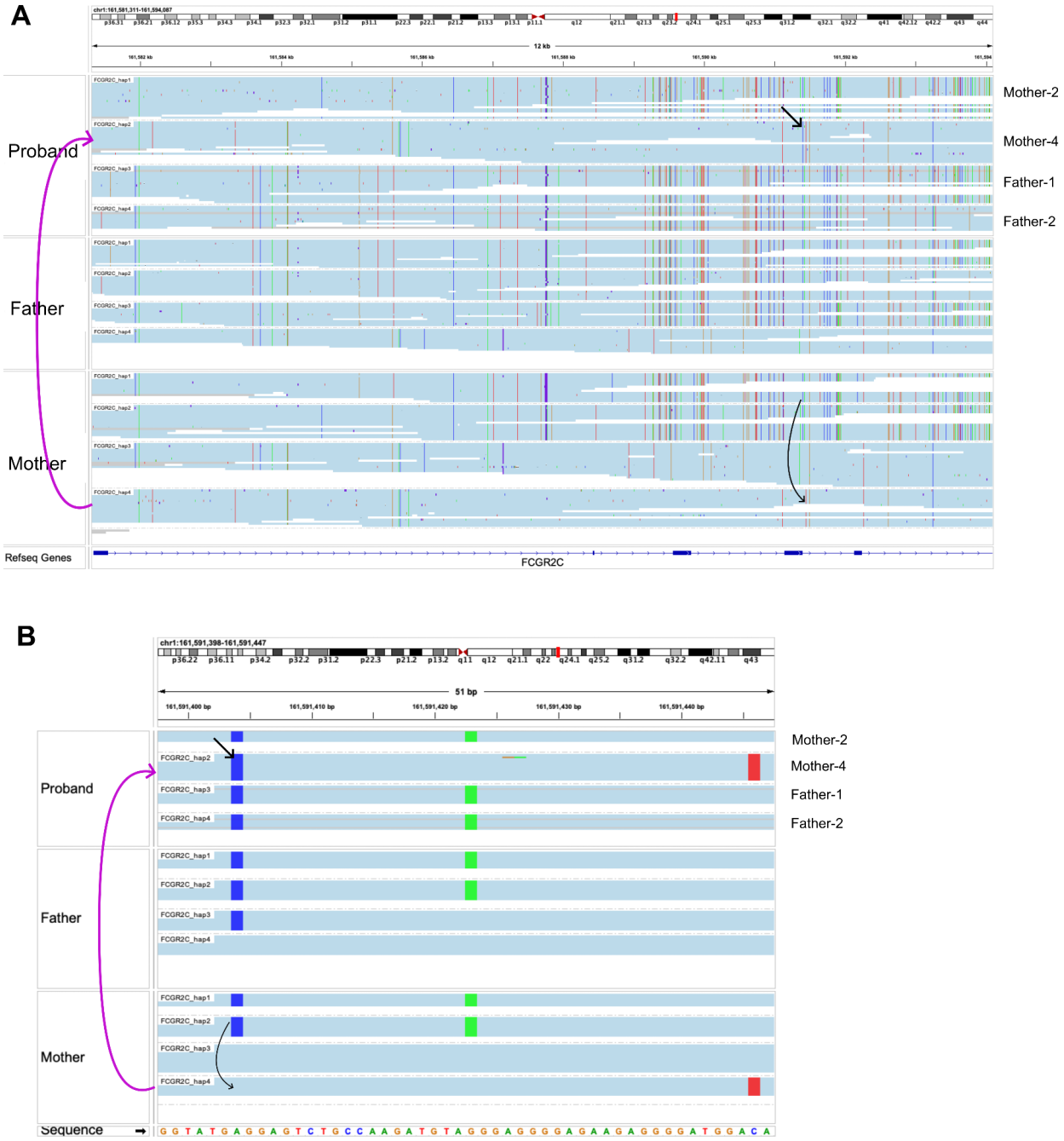
Variant is intronic. Variant is marked by the black arrow. B is a close view of the variant.





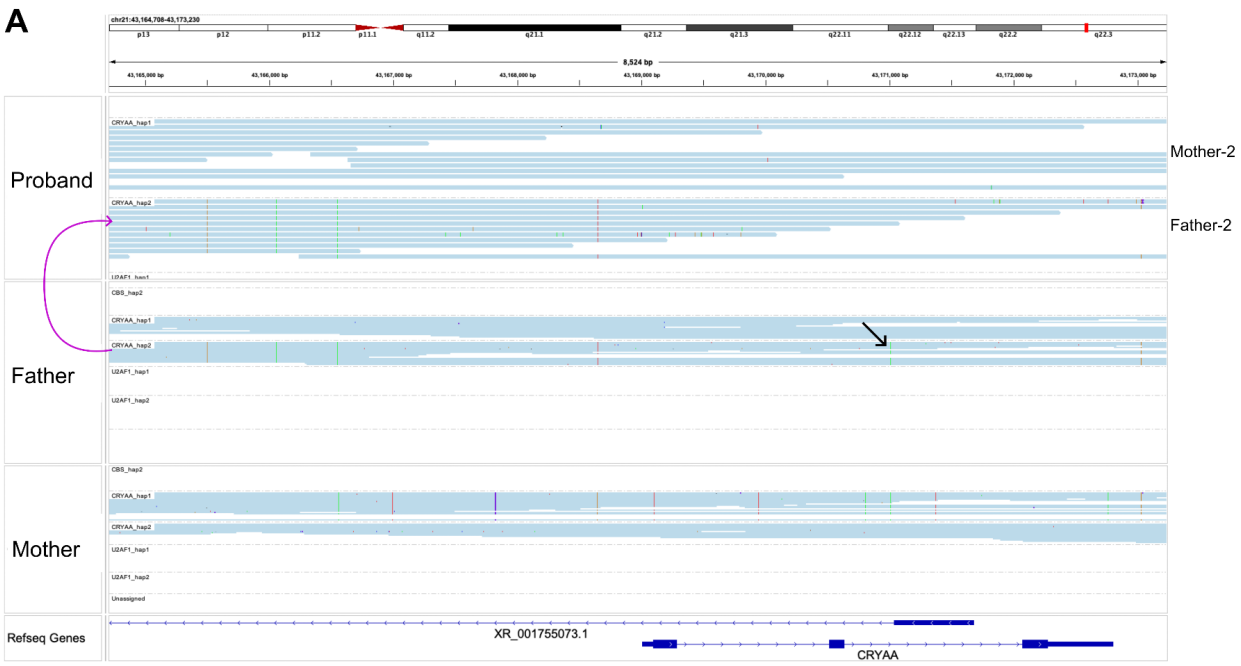
Supplementary Figure 11-8. Non-allelic gene conversion #2 (in addition to Figure 4).

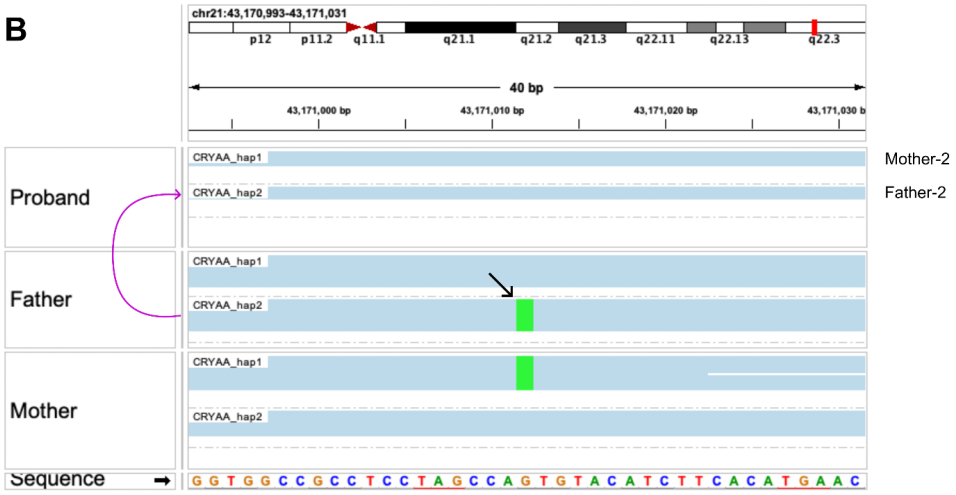
Variant is intronic. Black arrows mark the variant as well as the potential direction of conversion. B is a close view of the variant.



Supplementary Figure 11-9. Allelic gene conversion.

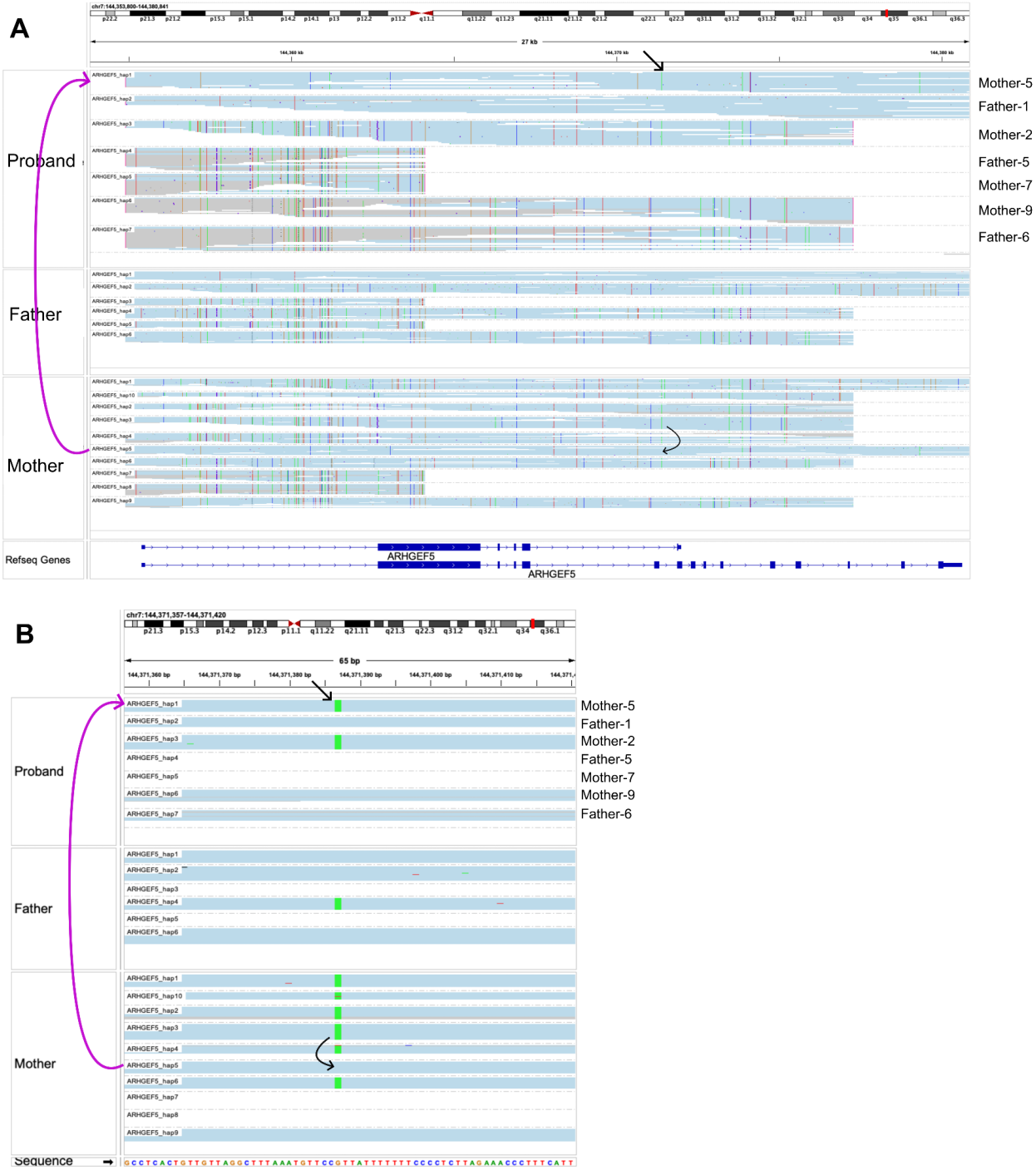
Variant is intronic. Variant is marked by the black arrow. B is a close view of the variant.





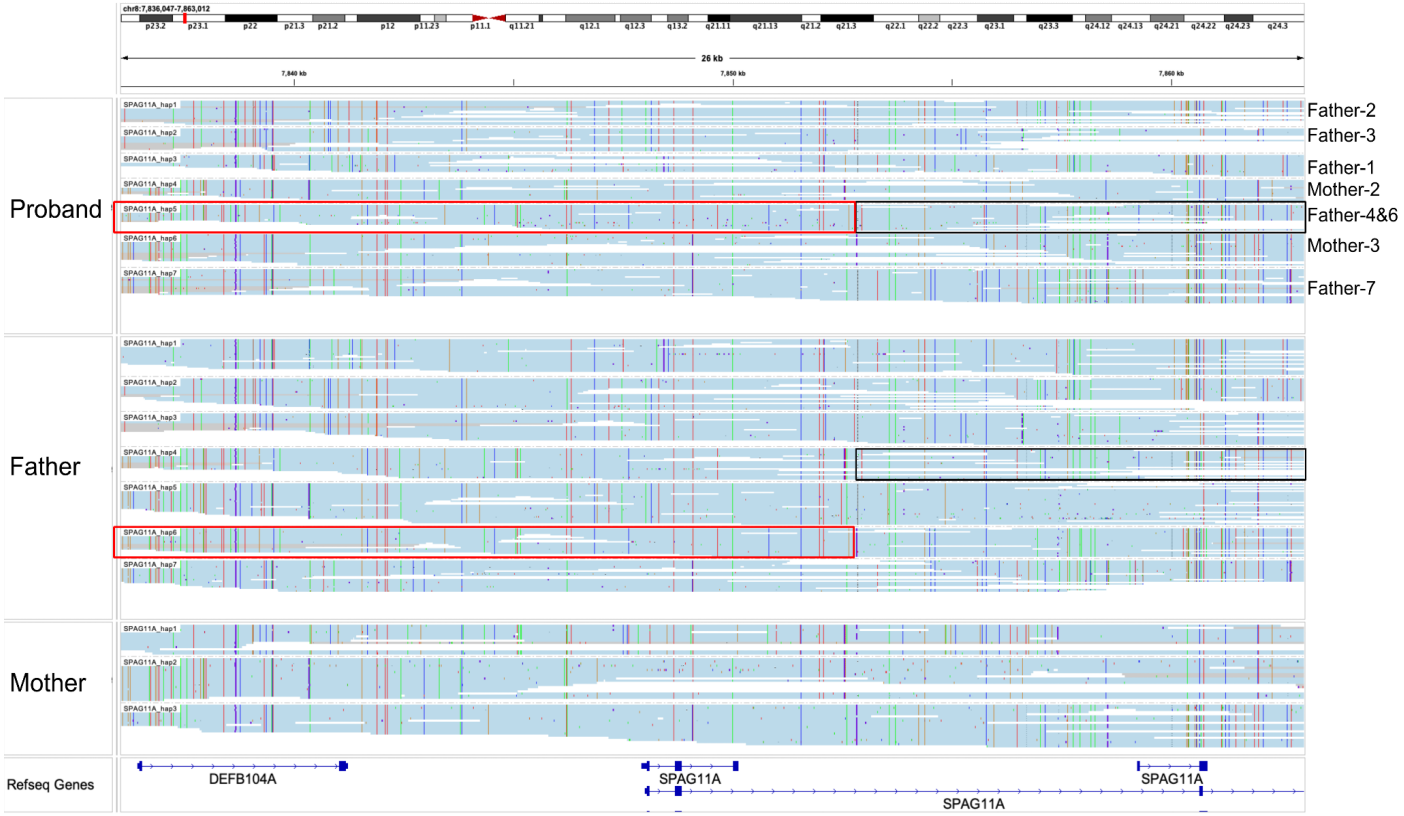
Supplementary Figure 11-10. Gene conversion allelic/non-allelic.

Variant is intronic. Black arrows mark the variant as well as the potential direction of conversion. B is a close view of the variant.



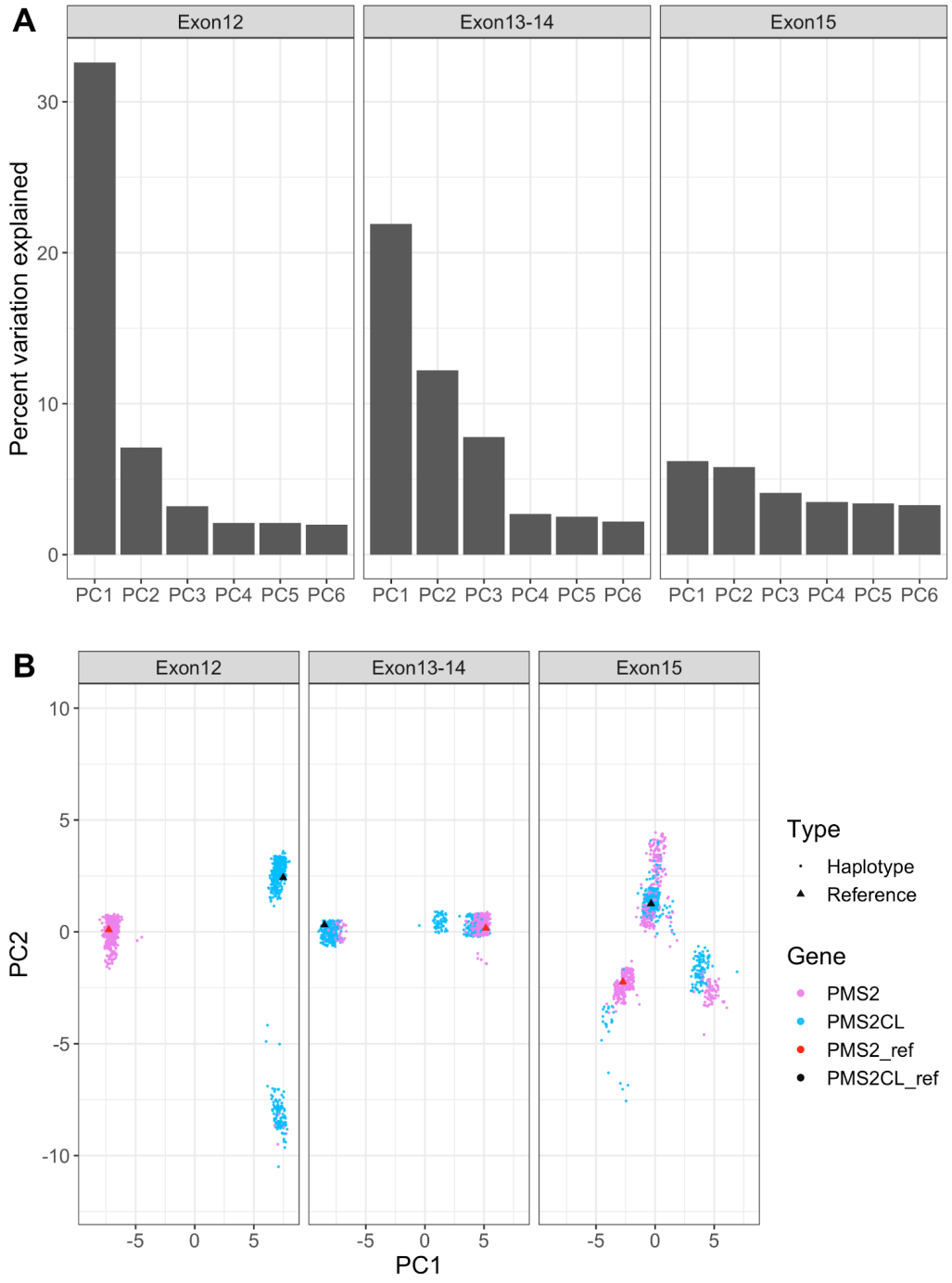
Supplementary Figure 12. Hybrid haplotype due to equal or unequal crossing over.

The red and black boxes mark two parts of the hybrid haplotype and their origin in the father.

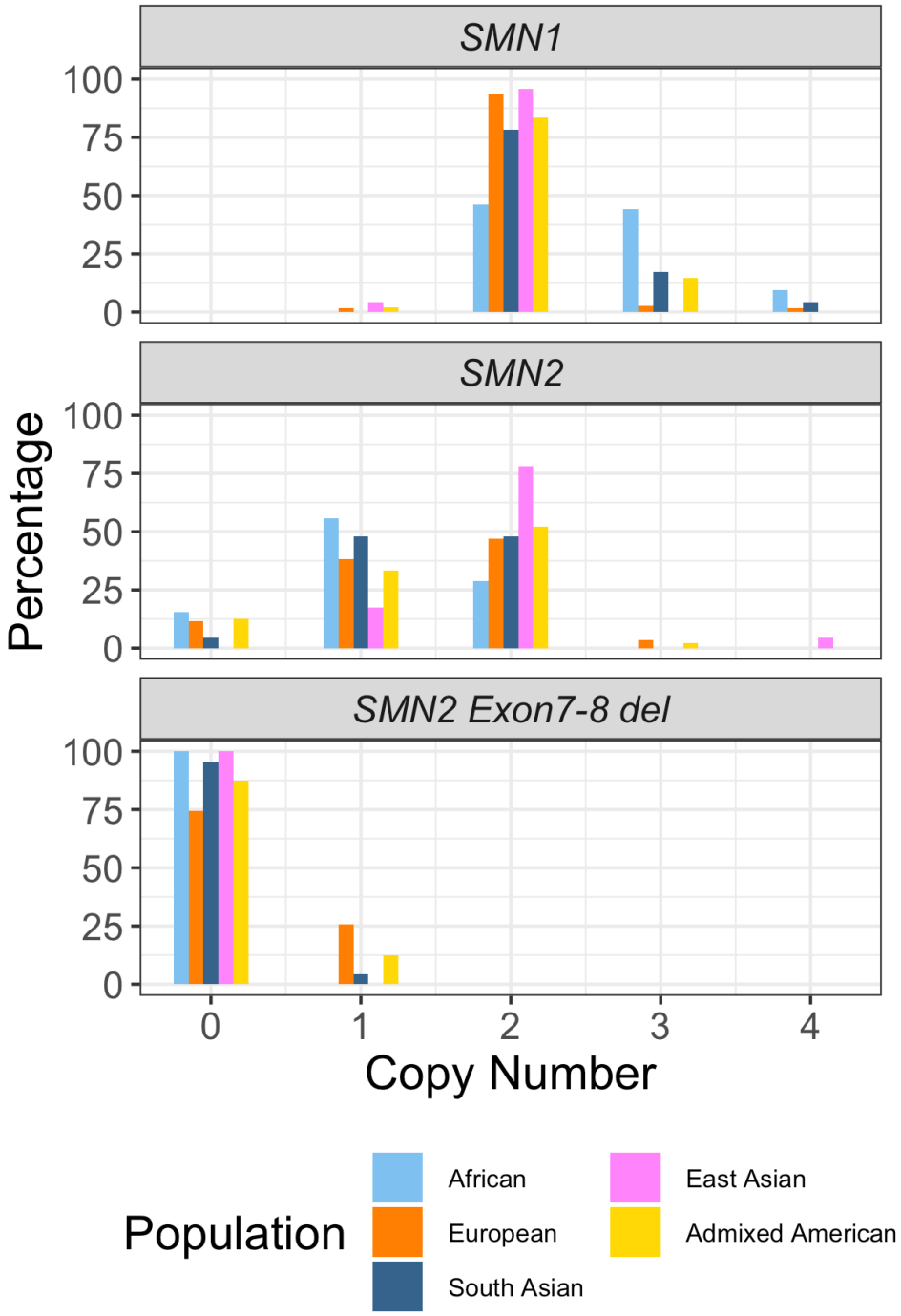


Supplementary Figure 13. PCA of Exon 12, Exons 13-14 and Exon 15 sequences in *PMS2/PMS2CL*.

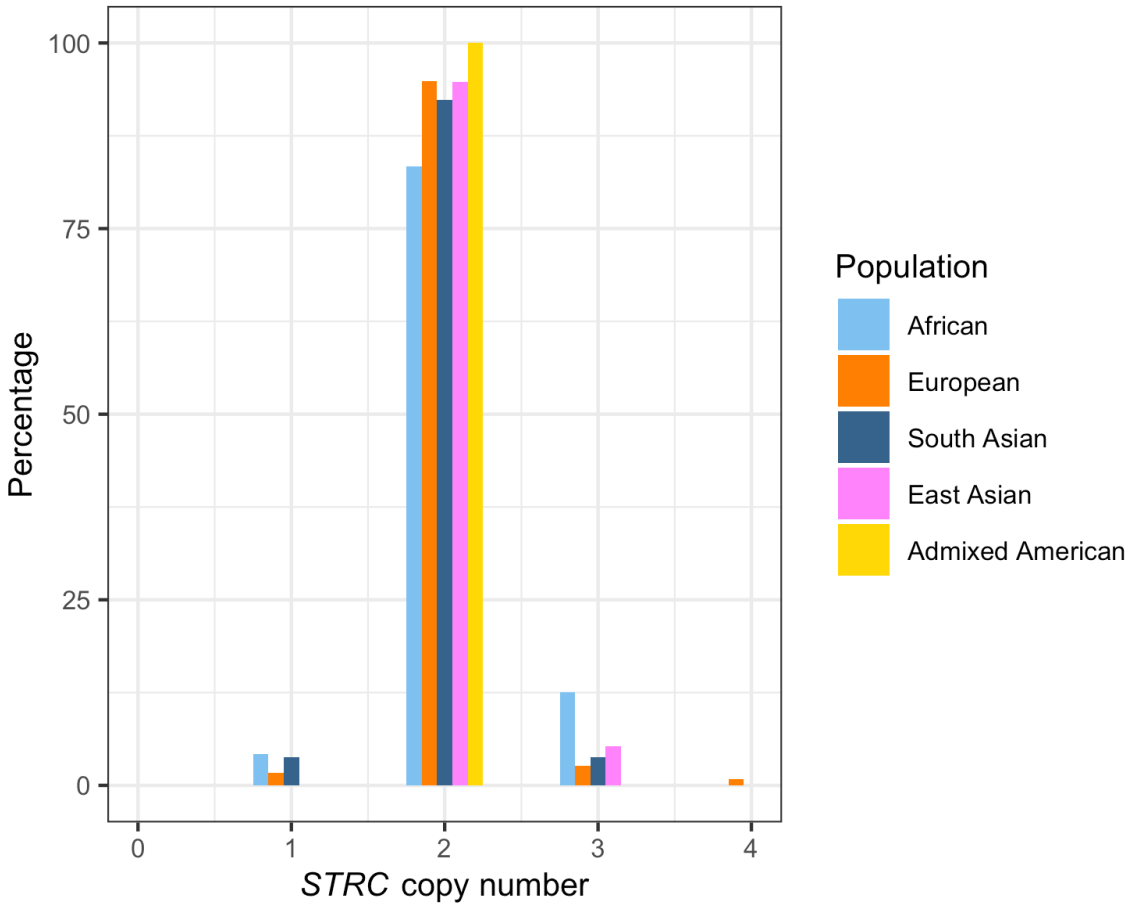
Percent variation explained (A) and PC1 vs. PC2 plots (B), where each dot is a haplotype in the population (pink or blue dots) or the reference sequence (red or black triangles). *PMS2* and *PMS2CL* are separated and distinguishable in Exon 12 and Exons 13-14, with occasional gene conversions. *PMS2* and *PMS2CL* are indistinguishable, with fully overlapping haplotypes in Exon 15.



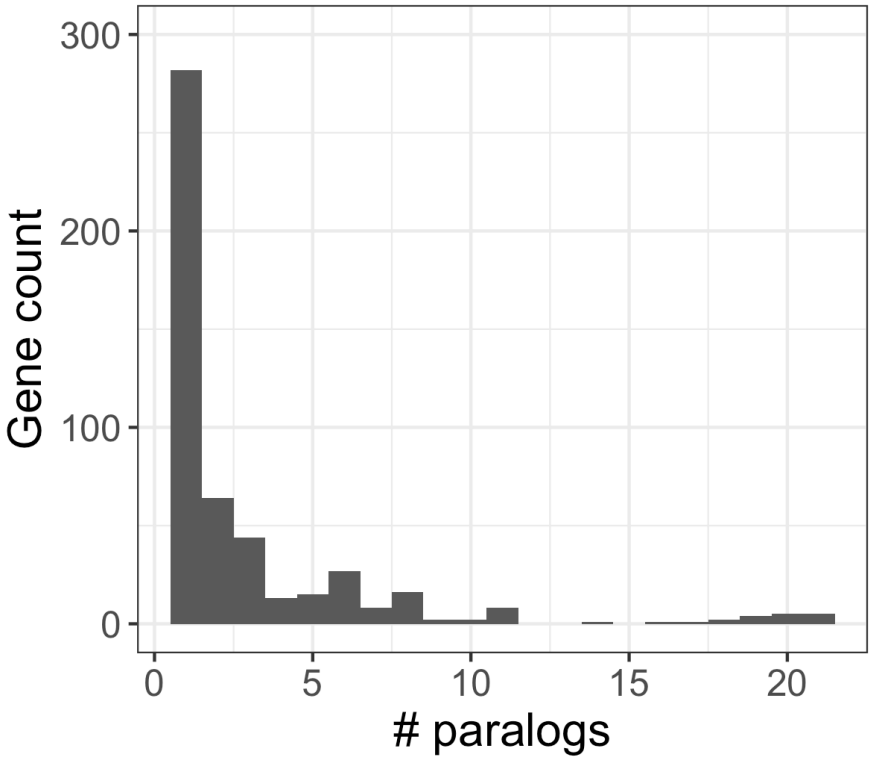
Supplementary Figure 14. Distribution of the copy number of *SMN1*, *SMN2* and *SMNΔ7-8* (deletion of Exons 7-8) across populations.



Supplementary Figure 15. Distribution of the copy number of *STRC* across populations.

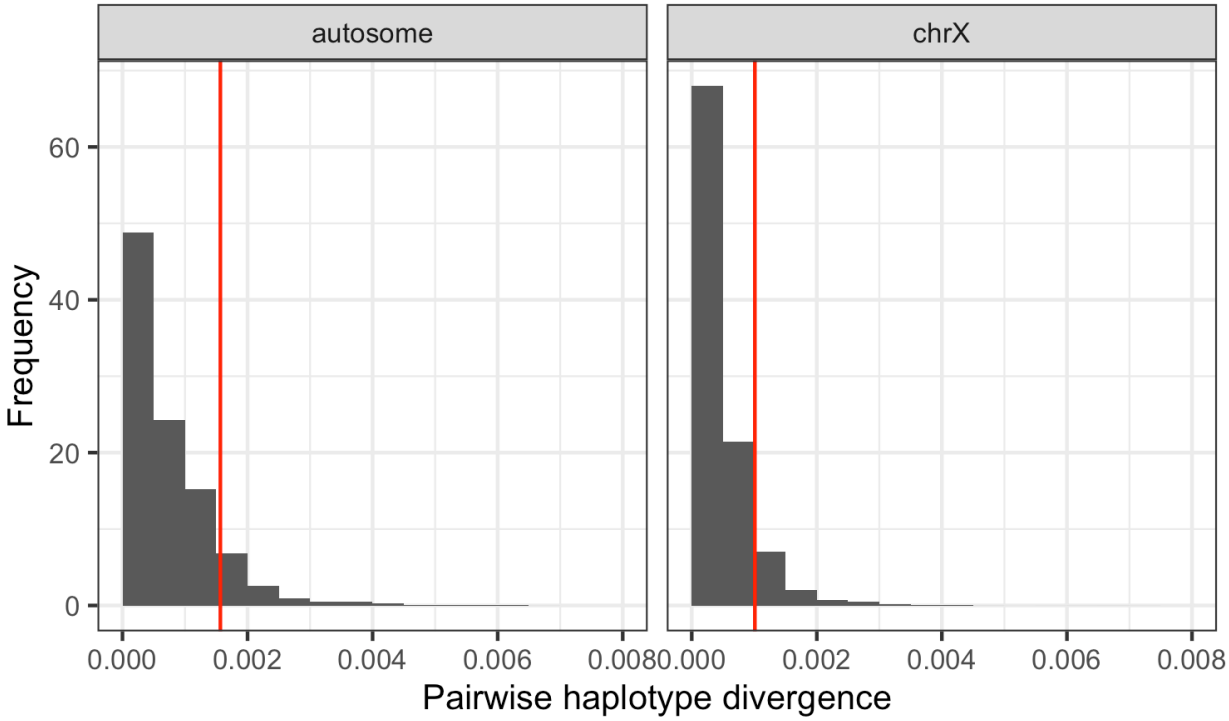


Supplementary Figure 16. Distribution of the number of highly similar paralogs for genes in SDs.



Supplementary Figure 17. Distributions of allelic sequence divergence for genes without paralogs

Distribution of pairwise sequence divergence between any two haplotypes of the same gene among 400 randomly selected autosomal genes and 200 randomly selected chrX genes. Red vertical lines represent the 90th percentile.



Supplementary Tables

Supplementary Table 1. Genes that fall in SDs in GRCh38 but have a consistent CN of two across populations.

Genes	Coordinate in GRCh38	Other coordinate in GRCh38	% samples with CN>2	Previously identified by CHM13 T2T assembly
DEFB130A/ DEFB130B	chr8:12304640-12323365	chr8:12058076-12076797	0.39%	x
RBPMS	chr8:30384041-30448023	chrX:37085009-37099262	0.39%	x
DEFB109B	chr8:7306410-7339183	chr8:7971187-8003955	0.78%	
CNTNAP3/ CNTNAP3C	chr9:39165433-39288814	chr9:61330217-61453530	2.7%	
MRPL23	chr11:1944688-1961966	chr11_KI270721v1_random:1-17145	0.39%	x
IQCK	chr16:19760000-19830000	chr13:86252980-86269524	1.16%	x
POTED	chr21:13609278-13646288	chrUn_GL000213v1:103302-140211	1.54%	
SMIM11	chr21:34374240-34392329	chr21:7743711-7761805	0.39%	x
C21orf140	chr21:34390593-34406217	chr21:7760069-7775696	0.39%	x
SMIM34	chr21:34411334-34431333	chr21:7781322-7801320	0.39%	x
KCNE1	chr21:34446191-34495759	chr21:7816178-7865746	1.16%	x
CBS	chr21:43052692-43076335	chr21:6444370-6468013	0	x
U2AF1	chr21:43090264-43110263	chr21:6481931-6501941	0	x
CRYAA	chr21:43160907-43173805	chr21:6552585-6565485	0	x
SIK1	chr21:43410808-43430807	chr21:6107453-6127452	0	x
HSF2BP	chr21:43528687-43659988	chr21:5966594-6009880	0.77%	x
TRAPPC10	chr21:44011809-44107052	chr21:5155001-5166246	1.17%	x
PWP2	chr21:44106900-44131681	chr21:5130371-5155153	0.39%	x
GATD3	chr21:44129698-44147864	chr21:5114188-5132354	0	x
ICOSLG	chr21:44221968-44241967	chr21:5021530-5041695	0.77%	x
DGCR6	chr22:18899205-18919204	chr22_KI270734v1_random:124478-144502	1.96%	x
PRODH	chr22:18912282-18936793	chr22_KI270734v1_random:137587-162092	1.93%	x

Supplementary Table 2. *CYP21A2* allele frequencies

Allele	**African**	**European**	**South Asian**	**East Asian**	**Admixed American**
WT (*CYP21A2*, *CYP21A1P*)	80.2%	73.5%	74.0%	84.2%	78.9%
CYP21A1P deletion	11.5%	15.0%	6.0%	7.9%	13.2%
CYP21A1P duplication	3.1%	6.6%	16.0%	5.3%	6.6%
CYP21A2 duplication	1.0%				
CYP21A2(WT), *CYP21A2*(Q319X), *CYP21A1P*	1.0%	2.2%	2.0%		1.3%
Alleles carrying pathogenic *CYP21A2* variants, listed below	3.1%	2.5%	2.0%	2.6%	
M240K	1.0%				
Q319X		0.4%	2.0%		
V282L	2.1%	1.3%			
G111Vfs		0.4%			
Fusion deletion		0.4%			
Large conversion				2.6%	

Supplementary Table 3. *OPN1LW/OPN1MW* allele frequencies

Allele (first two copies in the array)	**African**	**European**	**South Asian**	**East Asian**	**Admixed American**	**Phenotype**
OPN1LW+OPN1MW	96.2%	94.0%	96.3%	95.2%	93.9%	Normal
OPN1LW	1.3%			4.8%	1.5%	Deutan
OPN1LW+OPN1LW	2.5%	3.4%	3.7%		1.5%	
OPN1MW					1.5%	Protan
OPN1MW+OPN1MW		2.6%			1.5%	

Supplementary Table 4. *HBA1/HBA2* allele frequencies

Allele	**African**	**European**	**South Asian**	**East Asian**	**Admixed American**
-a	11.5%	1.7%	5.9%	2.6%	3.7%
aa	88.5%	95.8%	92.2%	97.4%	93.9%
aaa		2.5%	2.0%		2.4%

Supplementary Table 5. Frequencies of the deletion/duplication of the 11.7 kb region (Exons 4-10) in *IKBKG/IKBKGP1.*

Gene	Allele	African	European	South Asian	East Asian	Admixed American
IKBKG	11.7 kb deletion					
	11.7 kb duplication		0.6%			
IKBKGP1	11.7 kb deletion		0.6%			3.2%
	11.7 kb duplication	5.1%	1.2%	2.8%		

Supplementary Table 6. Deletion/Duplication frequencies in the *CFH* gene cluster.

SV (coordinates on chr1)	African	European	South Asian	East Asian	Admixed American
196748683_196833366_deletion	1.9%				
196760029_196844710_deletion	36.5%	20.2%	26.1%		18.8%
196792578_196917466_deletion		0.4%			
196813817_196935714_duplication		0.9%			
196815503_196937401_deletion	2.9%	0.4%			
196815503_196937401_duplication					1.0%

Supplementary Table 7. Deletion/Duplication frequencies in the *GBA/GBAP1* region.

SV (coordinates on chr1)	African	European	South Asian	East Asian	Admixed American
155212923_155233549_deletion		0.4%			
155214276_155234903_duplication	5.8%				
155215100_155235727_duplication			2.2%		1.0%