Complete resolution of gene/paralog pairs with PacBio HiFi sequencing

Xiao Chen[1], Emily Farrow[2], Isabelle Thiffault[2], Dalia Kasperaviciute[3], Genomics England Research Consortium, Alexander Hoischen[4], Christian Gilissen[4], Tomi Pastinen[2], Michael A Eberle[1]

1. PacBio, Menlo Park, CA, USA. 2. Children's Mercy Kansas City, MO, USA. 3. Genomics England Ltd., London, UK. 4. Radboud University Medical Center, Nijmegen, The Netherlands.

Poster P18.025.A
Contact: xchen@pacb.com

Overview

While whole-genome sequencing allows identification of clinically relevant variants in ~90% of the genome, there exist difficult regions that remain challenging for short read sequencing. Many medically relevant genes fall into these so-called dark regions where accurate analysis is hindered by the presence of highly similar paralogs. High sequence homology promotes unequal crossing over, resulting in frequent copy number variants (CNVs). PacBio HiFi long-read sequencing is ideal for resolving regions with high homology, but informatics methods are still lacking for segmental duplications longer than the HiFi read length. We developed Paraphase[1], a HiFi-based informatics method that accurately genotypes highly homologous genes and applied it to resolve hundreds of homologous regions across the genome.

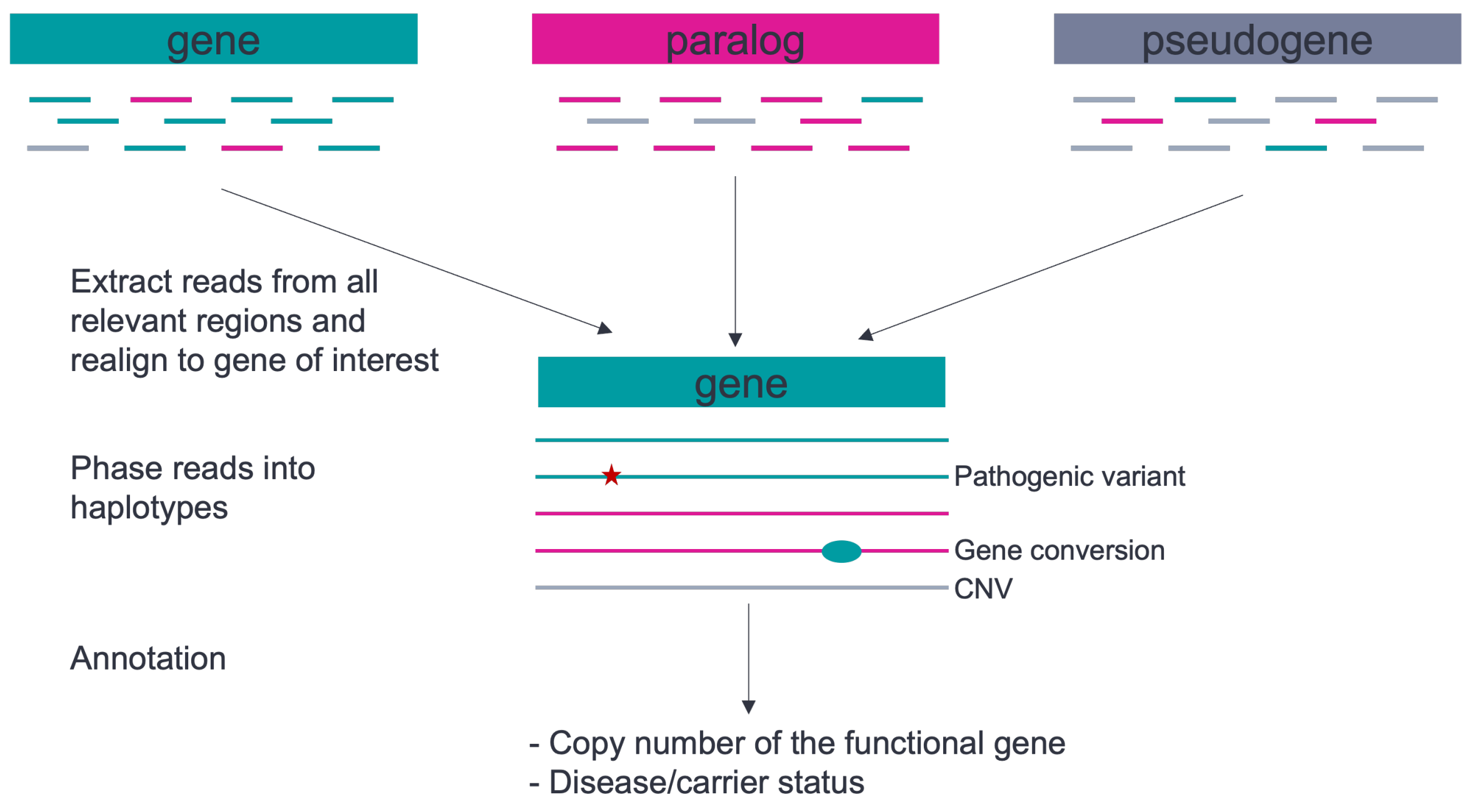
gene
paralog
pseudogene
Extract reads from all relevant regions and realign to gene of interest
gene
Phase reads into haplotypes
Pathogenic variant
Gene conversion
CNV
Annotation
- Copy number of the functional gene
- Disease/carrier status

Figure 1. Paraphase extracts reads from genes of the same family and phases reads into haplotypes. Haplotype annotation enables calling the copy number of the functional gene and the disease/carrier status.

SMN1 (spinal muscular atrophy)

Biallelic mutations in *SMN1* cause spinal muscular atrophy (SMA), a leading cause of infant death. *SMN1* has a highly similar paralog *SMN2*, differentiated by a single SNP in Exon 7 that marks the functional difference: *SMN1*: c.840C, *SMN2*: c.840T. Newborn screening and carrier screening for SMA are recommended by ACMG. Conventional SMA testing is mainly through copy number testing at c.840. Due to the high sequence similarity, it is challenging to identify pathogenic variants as well as silent carriers (2+0) that carry two copies of *SMN1* on one chromosome and zero copies on the other.

Paraphase phases complete *SMN1/SMN2* haplotypes, determines copy numbers and makes phased variant calls.

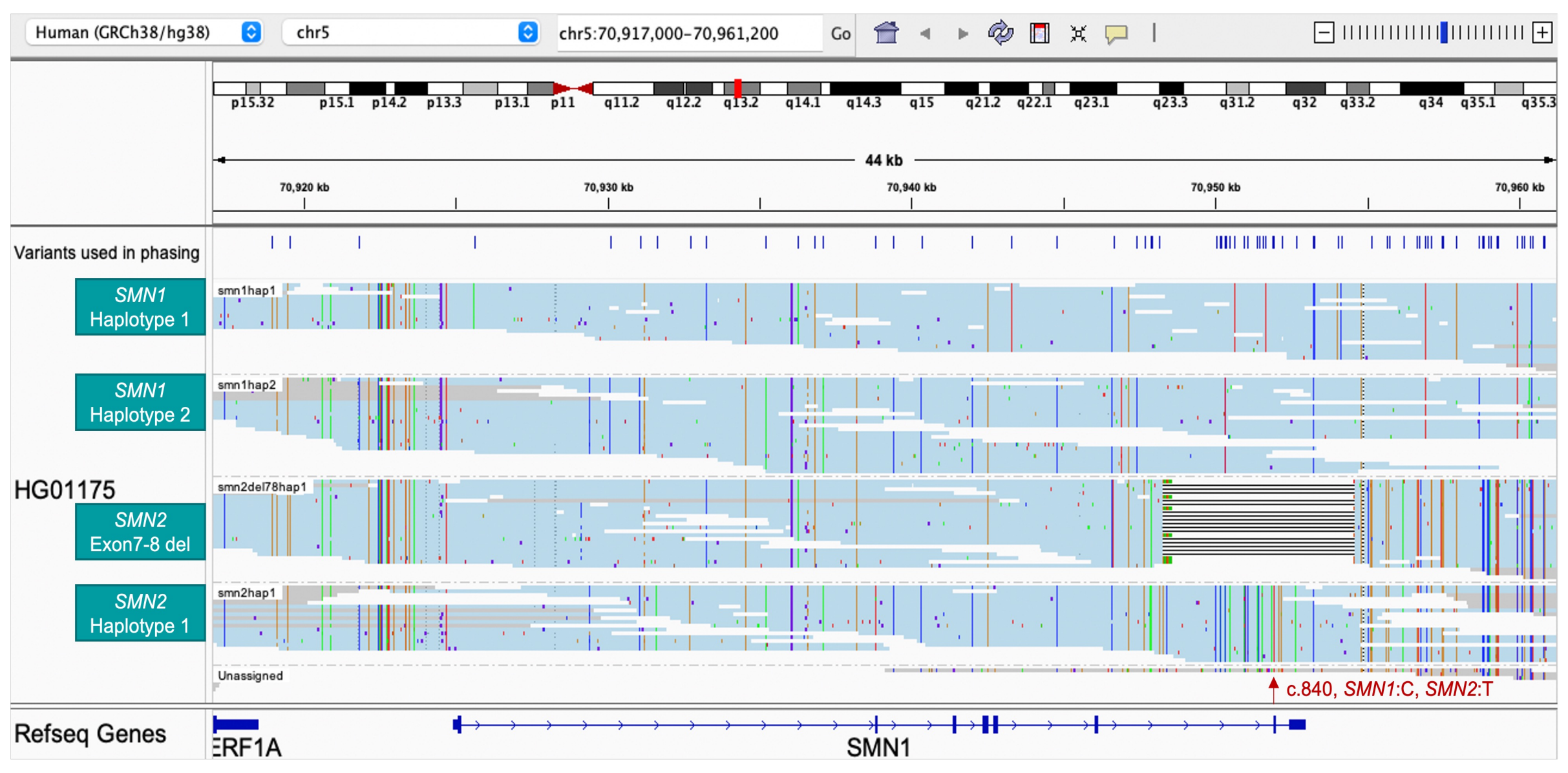

Figure 2. Visualization of *SMN1* and *SMN2* haplotypes identified by Paraphase. Reads are realigned to *SMN1* and grouped by the haplotype they originate from.

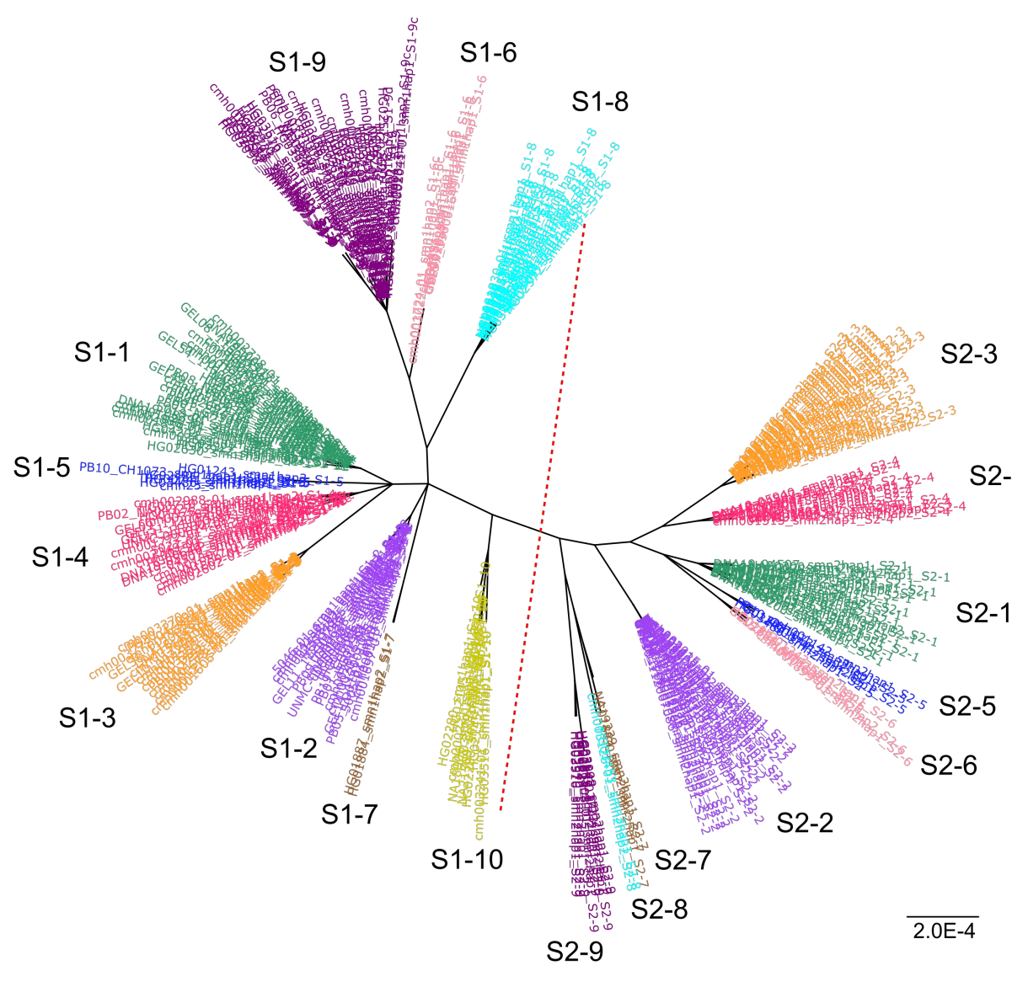

Figure 3. Major *SMN1* (left) and *SMN2* (right) haplogroups identified from population samples, separated by the red dotted line.

We identified a common two-copy *SMN1* allele, S1-8+S1-9d, that comprises two-thirds of two-copy *SMN1* alleles in Africans. Testing positive for S1-8 and S1-9d in an African individual with two copies of *SMN1* gives a silent carrier risk of 88.5%, significantly higher than the currently used marker g.27134T>G (1.7-3.0%)[2,3]. More details can be found in our publication[1].

RCCX module - *C4A/B, CYP21A2, TNXB* (21-hydroxylase-deficient congenital adrenal hyperplasia, Ehlers-Danlos syndrome)

Paraphase resolves this highly copy number variable region[4], including duplications in cis with pathogenic variants that are difficult to genotype by other technologies.

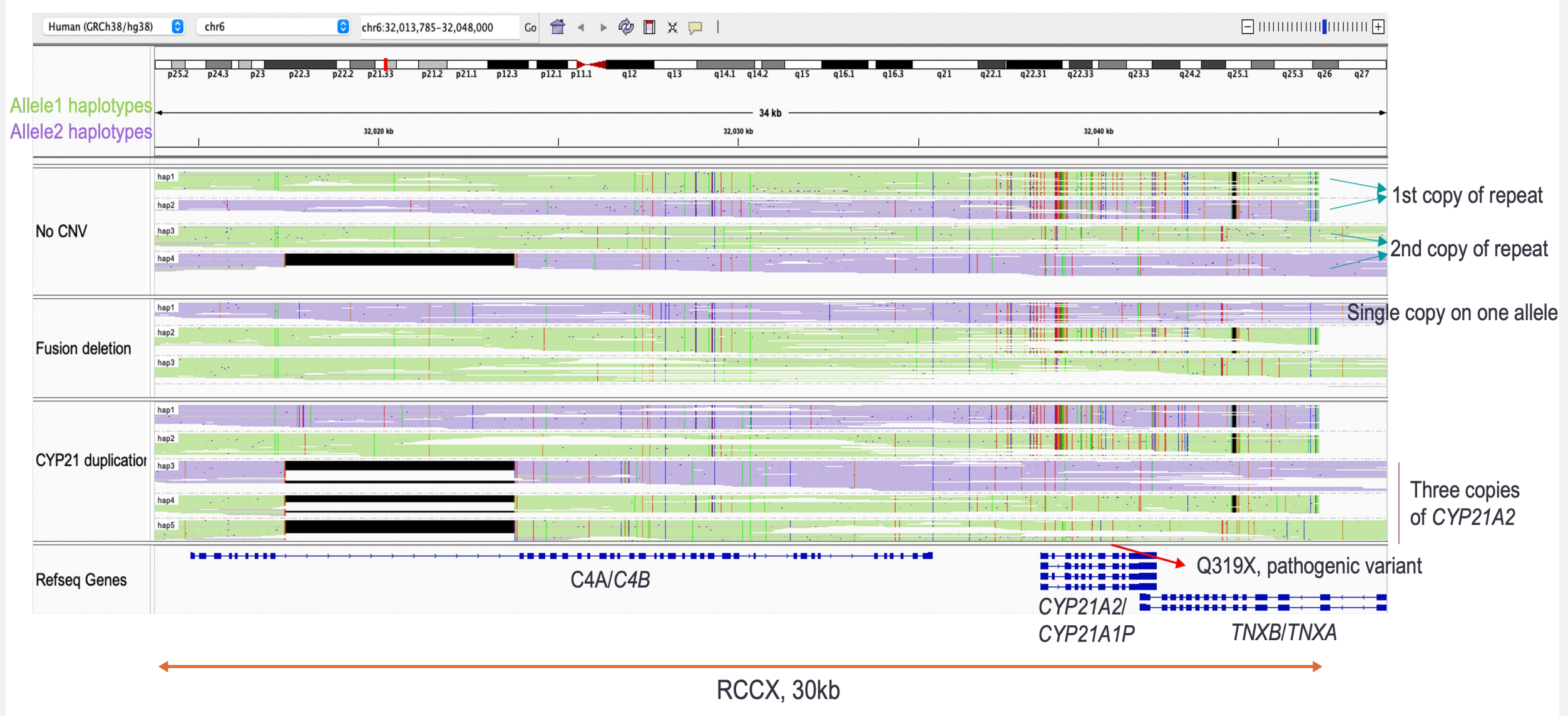

Figure 4. Paraphase resolves haplotypes in a sample with no CNV (top), fusion deletion (*CYP21A1P-CYP21A2* fusion, middle) and RCCX duplication (bottom). Haplotypes of the same color come from the same allele. The bottom sample carries an allele with a wild-type (WT) copy of *CYP21A2* and another copy of *CYP21A2* harboring a pathogenic variant Q319X. This allele is found across populations (1-2% based on our data) and could be wrongly detected as a nonfunctional allele if the additional copy of *CYP21A2* is not properly detected or phased.

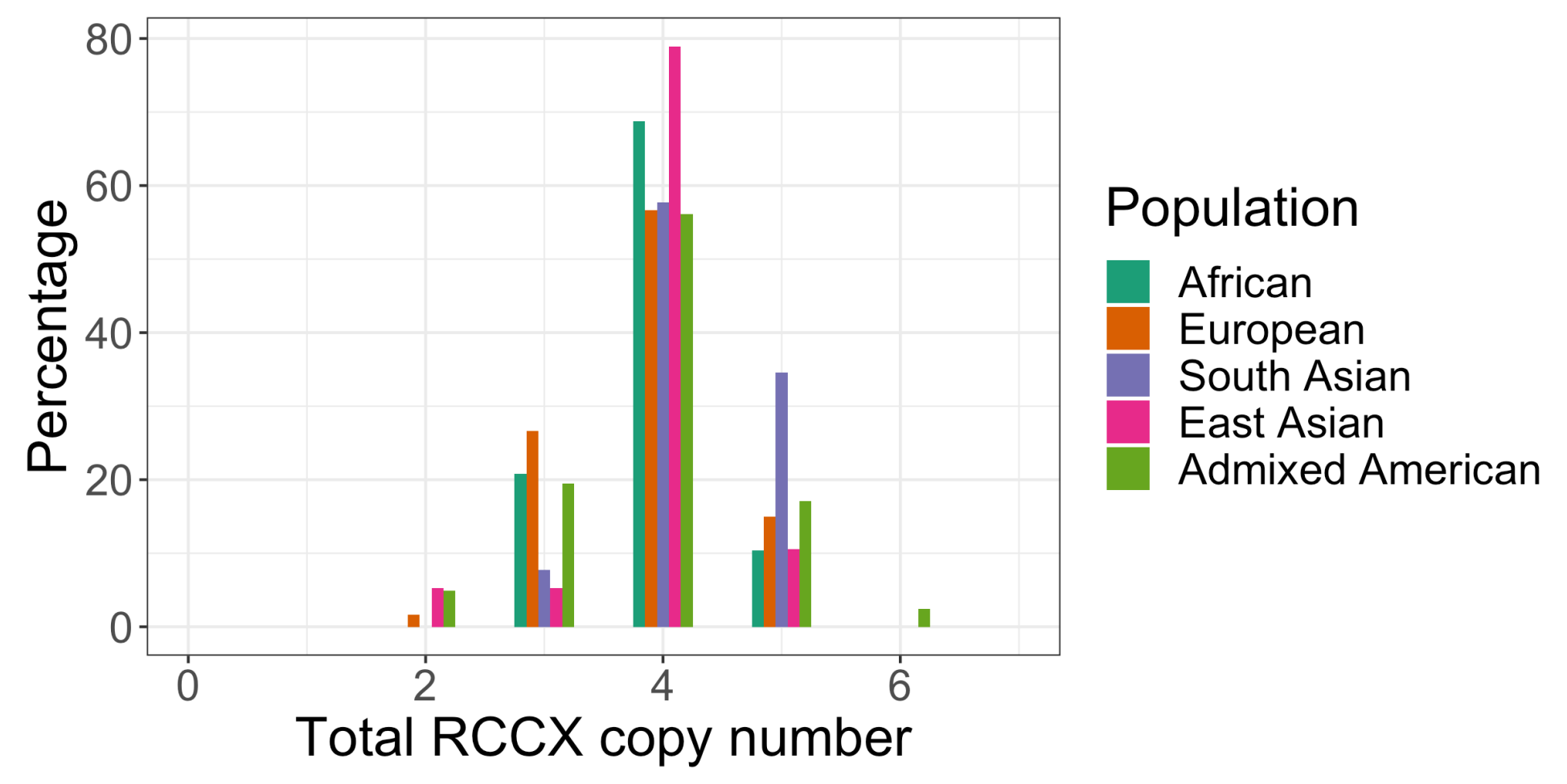

Figure 5. Frequency of total RCCX copy number across populations.

PMS2 (Lynch syndrome)

Paraphase enables variant calling in *PMS2* and characterization of gene conversion between *PMS2* and its pseudogene *PMS2CL*.

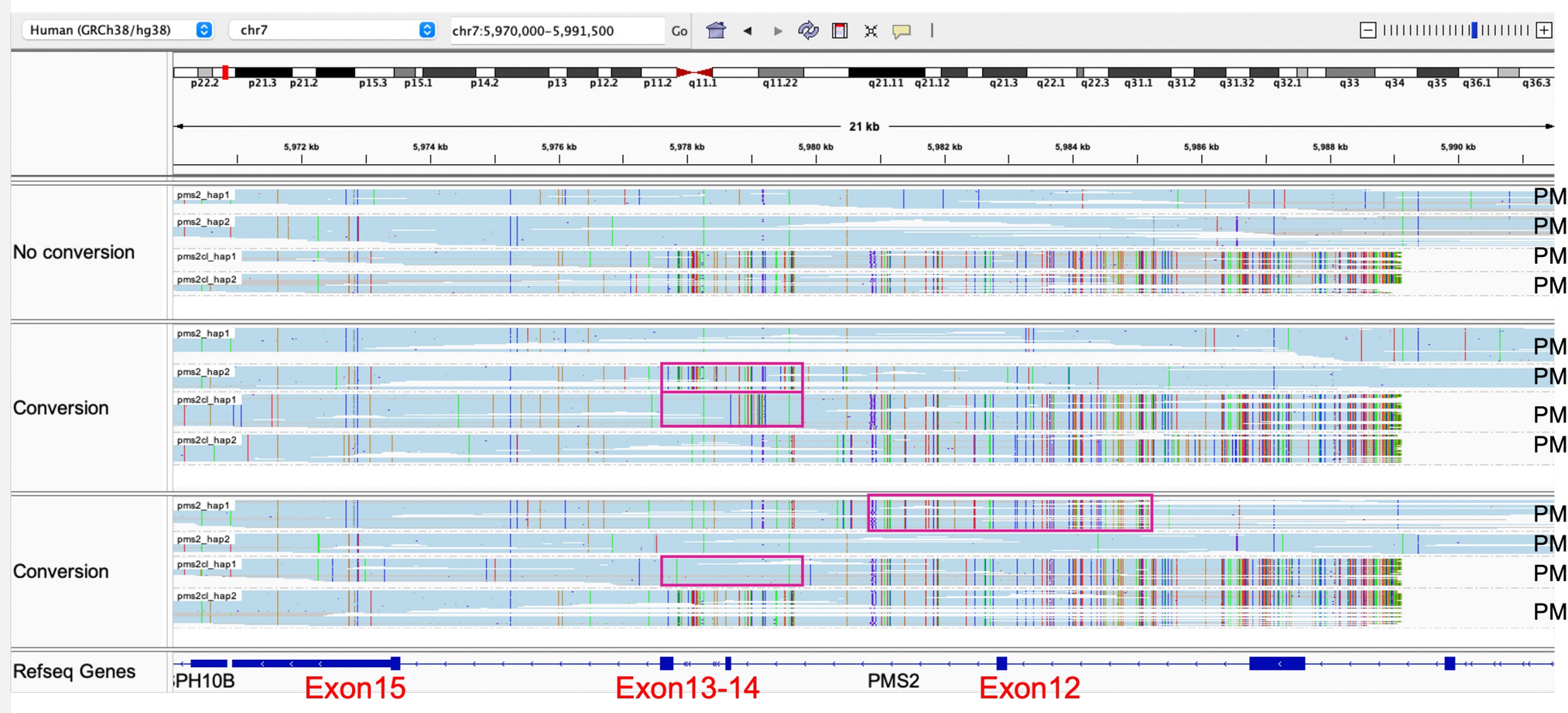

Figure 6. Paraphase resolves haplotypes for *PMS2* and *PMS2CL*. Top panel: a sample with no gene conversion. Middle panel: a sample with a *PMS2* allele converted to *PMS2CL*-like in Exon13-14 (highlighted in the colored boxes), and a *PMS2CL* allele partially converted to *PMS2*-like in Exon13-14. Bottom panel: a sample with a *PMS2* allele converted to *PMS2CL*-like in Exon12, and a *PMS2CL* allele converted to *PMS2*-like in Exon13-14. Exon15 sequences are highly similar (indistinguishable) between *PMS2* and *PMS2CL*.

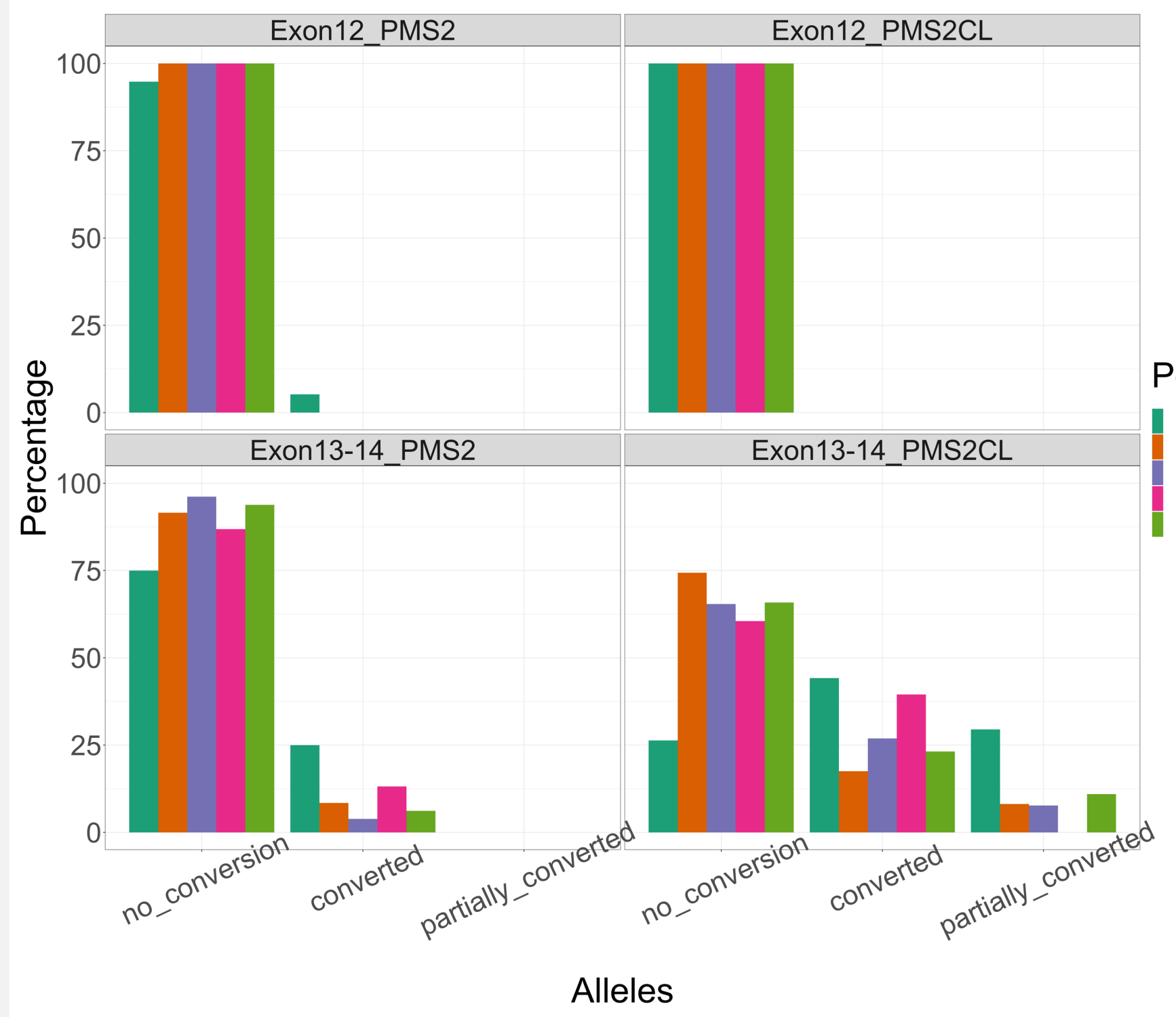

Figure 7. Frequency of gene conversion between *PMS2* and *PMS2CL* across populations in Exon12 and Exon13-14.

Extending Paraphase to highly homologous genome-wide

In addition to the three genomic regions mentioned above, Paraphase resolves the following clinically relevant genes with high homology.

- *STRC* (hereditary hearing loss and deafness)
- *IKBKG* (Incontinentia Pigmenti)
- *NCF1* (chronic granulomatous disease; Williams syndrome)
- *NEB* (Nemaline myopathy)
- *F8* (intron 22 inversion, Hemophilia A)
- *CFC1* (heterotaxy syndrome)

We are currently extending Paraphase to analyze all highly homologous regions across the genome where this approach is applicable.

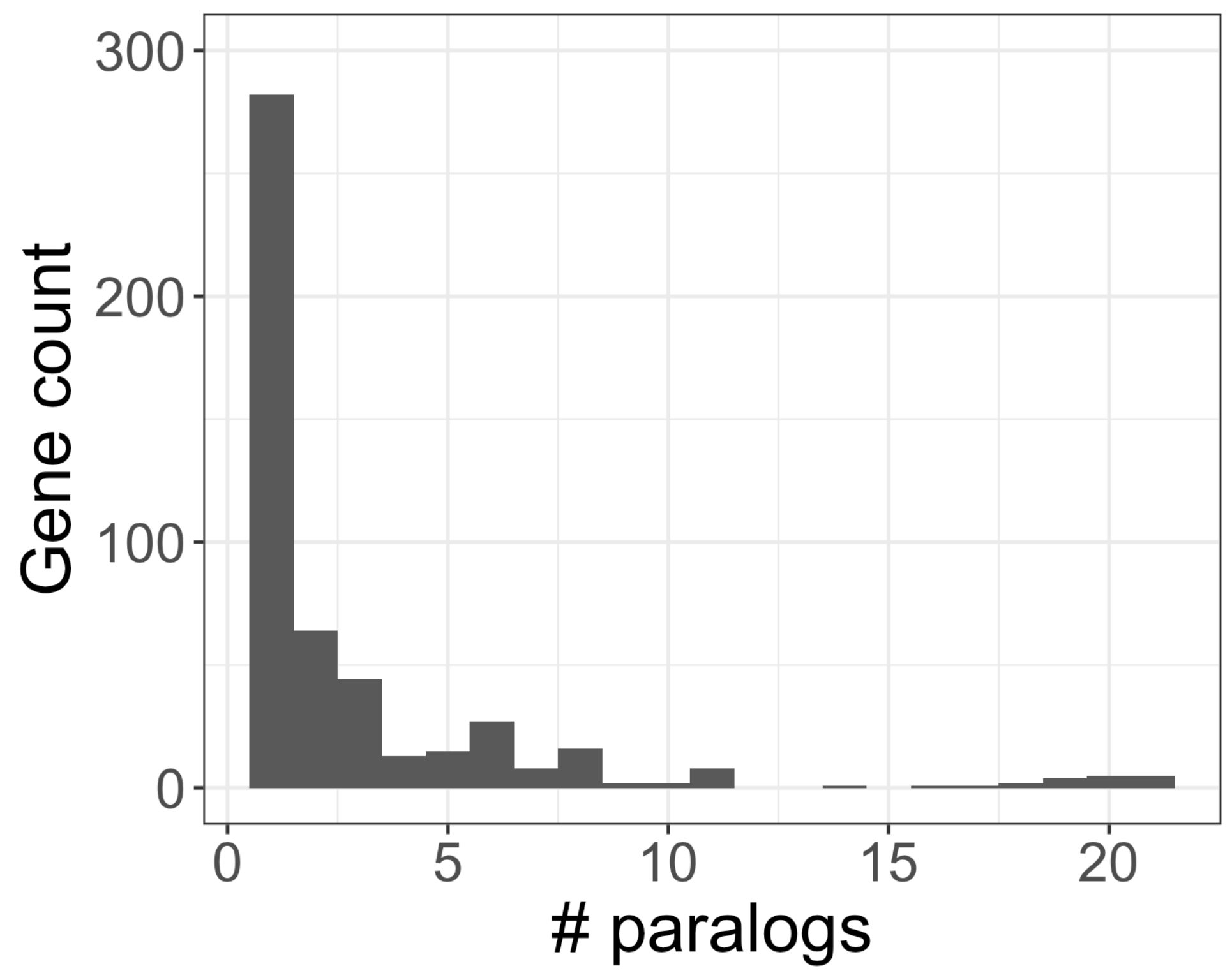

Figure 8. Distribution of the number of paralogs among genes across the genome (genes without any paralog are not plotted). 19394 Ensembl protein coding genes (>=20kb sequences centered on each gene) are aligned against GRCh38. Matches >10kb in length and >99% in sequence similarity are selected as candidate genes with highly homologous paralogs.

Among genes with three or fewer paralogs, we incorporated into Paraphase 169 groups of segmental duplications, which encode 326 genes in total (pseudogenes are not included).

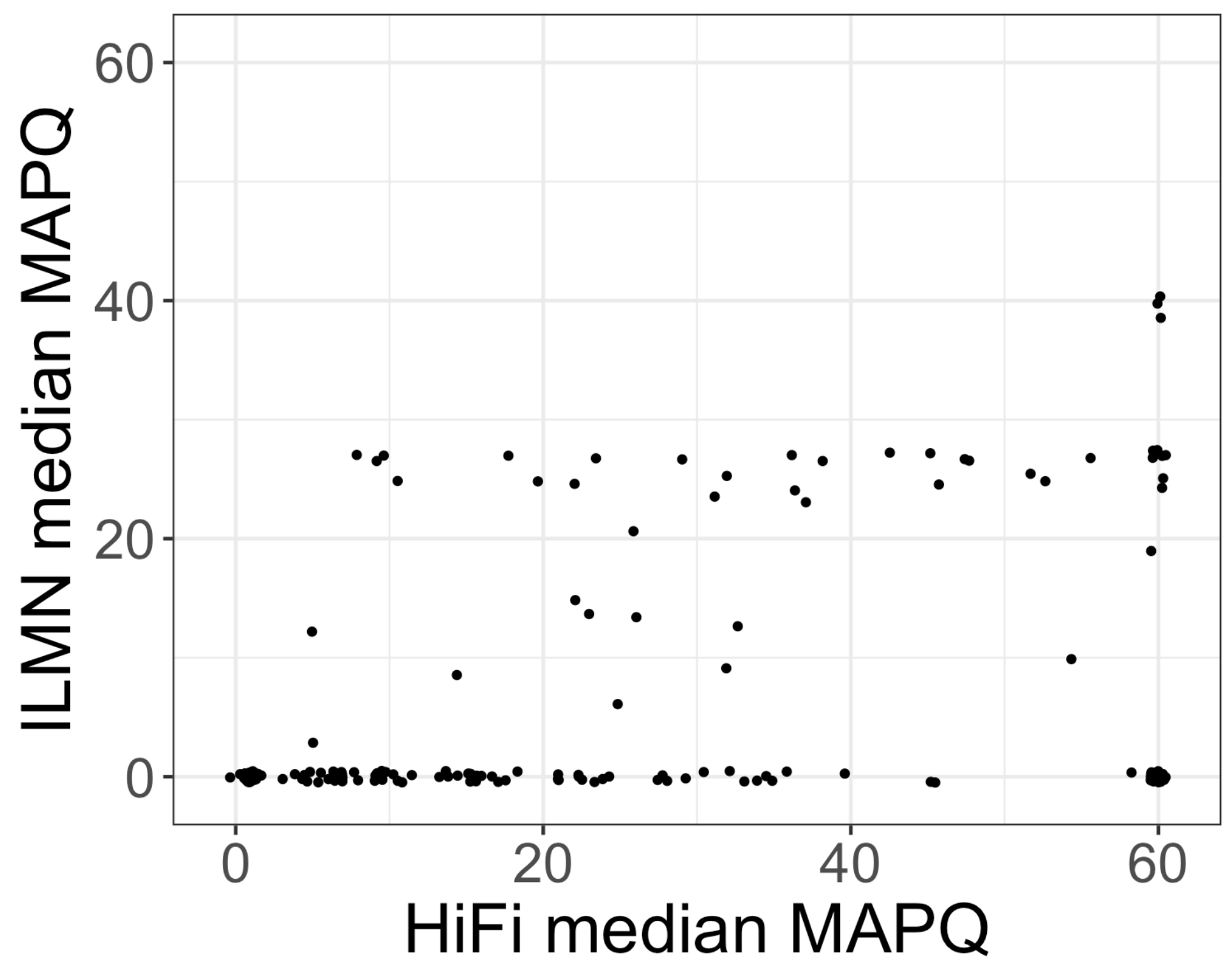

Figure 9. Comparison of median MAPQ between HiFi and Illumina WGS data in genes analyzed by Paraphase, highlighting mapping difficulty in these challenging regions for both short and long reads.

Conclusion

Paraphase, combined with highly accurate long reads, provides a single framework for resolving highly homologous genes, enabling better variant calling and novel gene-disease association discoveries in previously inaccessible genes.

References

1. Chen, et al. *Am. J. Hum. Genet.* 2023
2. Luo et al. *Genet. Med.* 2014
3. Chen et al. *Genet. Med.* 2020
4. Pignatelli et al. *Front. Endocrinol.* 2019

Paraphase

Conflict of interests: XC and MAE are employees of PacBio