Complete resolution of genes with highly homologous gene family members or pseudogenes using long-read PacBio HiFi sequencing

Xiao Chen, Michael Eberle
PacBio, 1305 O'Brien Drive, Menlo Park, CA 94025

Poster P146
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 ten clinically relevant challenging genes. Here we show detailed results in three genomic regions, revealing new population-wide biological and evolutionary insights.

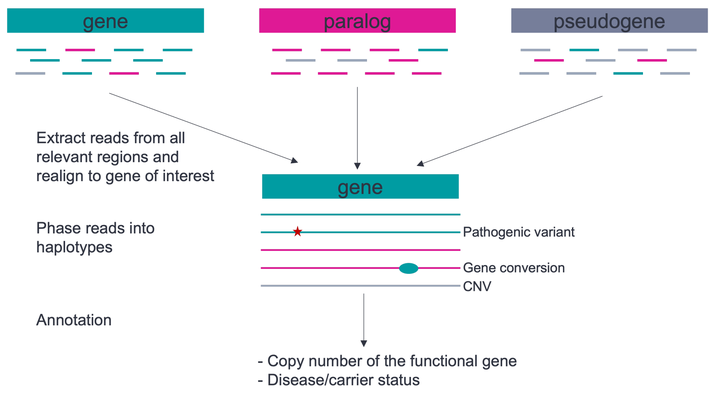

Figure 1. Paraphase extracts reads from genes of the same family and phases reads into haplotypes. Haplotype annotation enables calling the copy number of the functional gene and the disease/carrier status.

SMN1 (spinal muscular atrophy)

-Accurate genotyping of *SMN1*/*SMN2* copy numbers and small variants

Biallelic mutations in *SMN1* causes spinal muscular atrophy (SMA), a leading cause of infant death. *SMN1* has a highly similar paralog *SMN2* (with minimal function), and they are 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 dosage (copy number) testing at c.840. Due to the high sequence similarity, it is challenging to identify other pathogenic variants throughout both genes 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.

- Enables identification of standard carriers
- Calls pathogenic variants other than c.840C>T
- Enables haplotype-based screening of silent carriers (2+0)

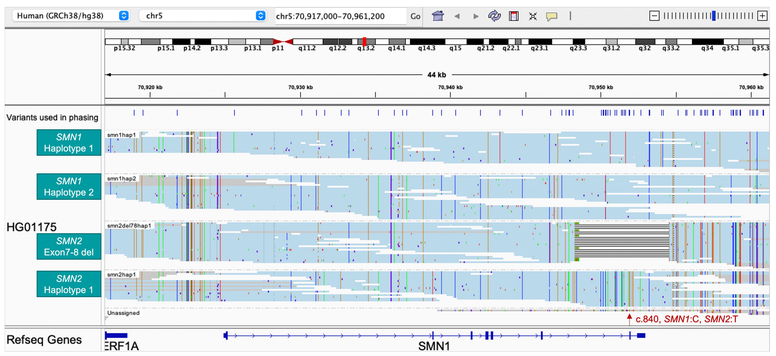

Figure 2. Visualization of *SMN1* and *SMN2* haplotypes identified by Paraphase. Reads are realigned to *SMN1* and grouped by the haplotype they originate from.

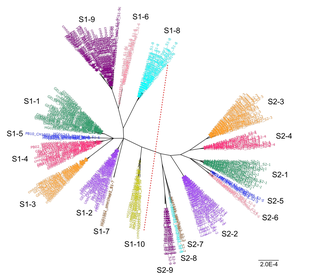

Figure 3. Major *SMN1* (left) and *SMN2* (right) haplogroups separated by the red dotted line.

We analyzed 925 *SMN1* haplotypes and 645 *SMN2* haplotypes from 438 individuals from five ethnic populations and identified ten and nine major *SMN1* and *SMN2* haplogroups, respectively.

Haplotypes were phased into alleles in 341 pedigrees. We identified a common two-copy *SMN1* allele, S1-8+S1-9d, that comprises two-thirds of two-copy *SMN1* alleles in Africans. Both haplotypes are rarely present as singleton alleles and can serve as a good marker for silent carriers. 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%, which is significantly higher than the currently used marker SNP g.27134T>G (1.7-3.0%)[2,3]. More details can be found in our recent publication[1].

RCCX module (*C4A/B*, *CYP21A2*, *TNXB*)

-Identifying CNVs and pathogenic variants resulting from gene conversion

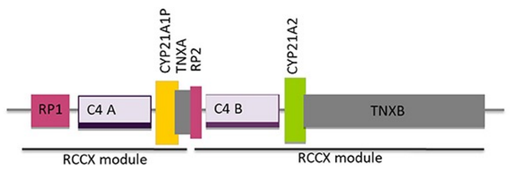

Figure 4. RCCX module, figure taken from Ref 4, is a 30kb tandem repeat with frequent CNVs. It contains four clinically relevant genes: *CYP21A2 (*21-Hydroxylase-Deficient Congenital Adrenal Hyperplasia, pseudogene *CYP21A1P*), *TNXB* (Ehlers-Danlos syndrome, pseudogene *TNXA*) and *C4A*/*C4B* (relevant in autoimmune diseases).

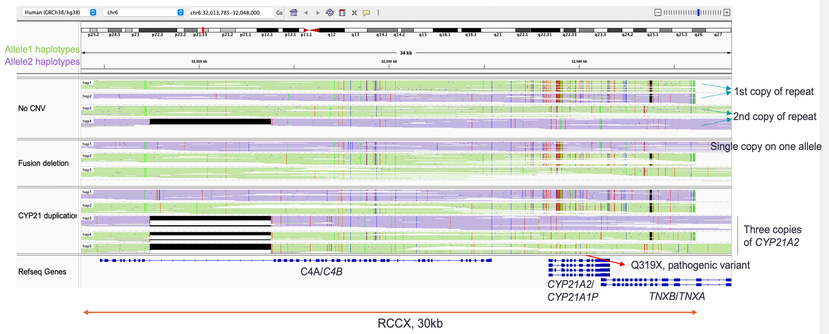

Figure 5. IGV snapshot showing Paraphase resolved haplotypes in a sample with no CNV (top), fusion deletion (*CYP21A1P-CYP21A2* fusion, middle) and RCCX duplication (bottom). The bottom sample carries an allele with a wild-type (WT) copy of *CYP21A2* and another copy of *CYP21A2* harboring a pathogenic variant Q319X. This allele is found across populations, see Table 1 below, and could be wrongly detected as a nonfunctional allele if the additional copy of *CYP21A2* is not properly detected.

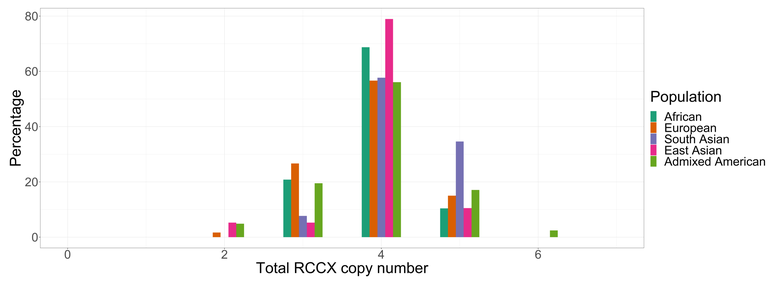

Figure 6. Frequency of the total RCCX copy number across populations.

Table 1. Frequency of CNVs and pathogenic variants in *CYP21A2* across populations.

Allele	European (N=240)	East Asian (N=38)	South Asian (N=52)	Admixed American (N=82)	African (N=96)
WT (*CYP21A2*, *CYP21A1P*)	73.5%	84.2%	74.0%	78.4%	78.7%
CYP21A1P deletion	15.2%	7.9%	6.0%	13.5%	11.7%
CYP21A1P duplication	6.1%	5.3%	16.0%	6.8%	4.3%
CYP21A2 duplication	0.4%				1.1%
CYP21A2(WT), *CYP21A2*(Q319X), *CYP21A1P*	2.2%		2.0%	1.4%	1.1%
Alleles carrying pathogenic *CYP21A2* variants	2.6%	2.6%	2.0%		3.2%

PMS2 (Lynch syndrome)

-Characterizing frequent gene conversion between *PMS2* and *PMS2CL*

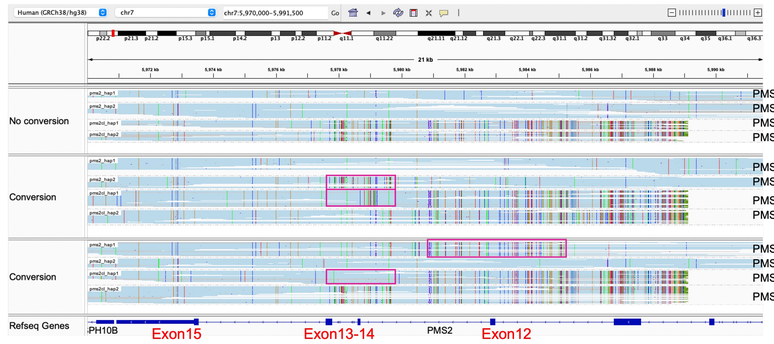

Figure 7. IGV snapshot showing Paraphase resolved haplotypes for *PMS2* and its pseudogene *PMS2CL*. The top panel shows a sample with no gene conversion. The middle panel shows a sample with a *PMS2* allele converted to *PMS2CL*-like in Exon13-14, and a *PMS2CL* allele partially converted to *PMS2*-like in Exon13-14. The bottom panel shows 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 indistinguishable between *PMS2* and *PMS2CL*, also see Figure 8.

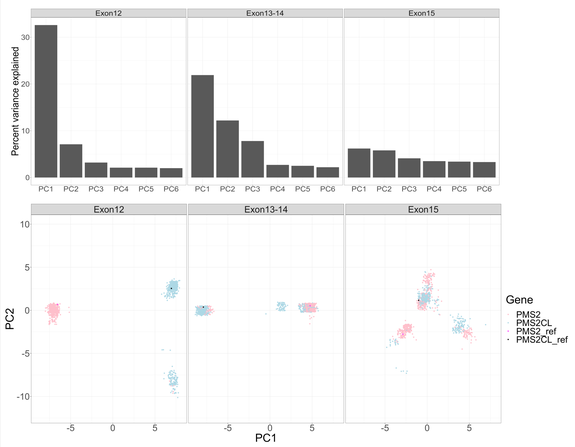

Figure 8. Principal component analysis revealed different gene conversion patterns in *PMS2*, showing occasional conversion in Exon12, frequent bidirectional conversion in Exon13-14, and highly homogeneous sequences in Exon15. *PMS2* and *PMS2CL* haplotypes were analyzed together with the *PMS2* and *PMS2CL* reference sequences, shown as two darker dots.

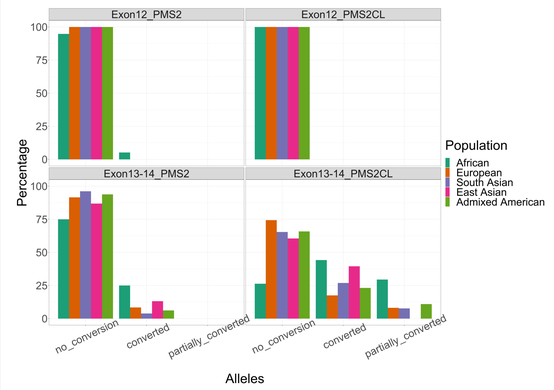

Figure 9. Frequency of gene conversion between *PMS2* and *PMS2CL* across populations.

Conclusion

Paraphase provides a single framework for resolving highly homologous genes. Paraphase works for more clinically important genes listed below and is being extended to a generalized genome-wide caller for paralogs.

- *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)

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