

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

Received: 24 April 2024

Accepted: 21 February 2025

Published online: 08 March 2025

 Check for updates

Xiao Chen¹✉, Daniel Baker¹, Egor Dolzhenko¹, Joseph M. Devaney², Jessica Noya², April S. Berlyoung², Rhonda Brandon², Kathleen S. Hruska², Lucas Lochovsky², Paul Kruszka², Scott Newman², Emily Farrow^{3,4,5}, Isabelle Thiffault^{3,4,6}, Tomi Pastinen^{3,4}, Dalia Kasperaviciute⁷, Christian Gilissen^{8,9}, Lisenka Vissers^{8,9}, Alexander Hoischen^{8,9,10,11}, Seth Berger¹², Eric Vilain¹³, Emmanuèle Délot¹³, UCI Genomics Research to Elucidate the Genetics of Rare diseases (UCI GREGoR) Consortium* & Michael A. Eberle¹✉

Variant calling is hindered in segmental duplications by sequence homology. We developed Paraphase, a HiFi-based informatics method that resolves highly similar genes by phasing all haplotypes of paralogous genes together. We applied Paraphase to 160 long (>10 kb) segmental duplication regions across the human genome with high (>99%) sequence similarity, encoding 316 genes. Analysis across five ancestral populations revealed highly variable copy numbers of these regions. We identified 23 paralog groups with exceptionally low within-group diversity, where extensive gene conversion and unequal crossing over contribute to highly similar gene copies. Furthermore, our analysis of 36 trios identified 7 de novo SNVs and 4 de novo gene conversion events, 2 of which are non-allelic. Finally, we summarized extensive genetic diversity in 9 medically relevant genes previously considered challenging to genotype. Paraphase provides a framework for resolving gene paralogs, enabling accurate testing in medically relevant genes and population-wide studies of previously inaccessible genes.

Population-wide whole-genome sequencing (WGS) studies based on short reads have enabled comprehensive characterization of variants, particularly small variants, in ~90% of the human genome^{1–3}. However, there exist difficult regions and variant classes that remain largely inaccessible to short reads^{4,5}. A large portion of these difficult regions

occur within segmental duplications (SDs)^{6,7}, where high sequence similarity between copies of SDs results in ambiguous mapping of short reads. In addition to difficulty mapping reads within SDs, high sequence similarity promotes unequal crossing over, resulting in hotspots for copy number variants (CNVs), as well as high rates of gene

¹PacBio, Menlo Park, CA, USA. ²GeneDx, Gaithersburg, MD, USA. ³Genomic Medicine Center, Children's Mercy Kansas City, Kansas City, MO, USA. ⁴UMKC School of Medicine, University of Missouri Kansas City, Kansas City, MO, USA. ⁵Department of Pediatrics, Children's Mercy Kansas City, Kansas City, MO, USA. ⁶Department of Pathology and Laboratory Medicine, Children's Mercy Kansas City, Kansas City, MO, USA. ⁷Genomics England Ltd., London, UK. ⁸Department of Human Genetics, Radboud University Medical Center, Nijmegen, The Netherlands. ⁹Research Institute for Medical Innovation, Radboud University Medical Center, Nijmegen, The Netherlands. ¹⁰Radboud Center for Infectious Diseases (RCI), Department of Internal Medicine, Radboud University Medical Center, Nijmegen, The Netherlands. ¹¹Radboud Expertise Center for Immunodeficiency and Autoinflammation and Radboud Center for Infectious Disease (RCI), Radboud University Medical Center, Nijmegen, The Netherlands. ¹²Center for Genetics Medicine Research, Children's National Hospital, Washington, DC, USA. ¹³Institute for Clinical and Translational Science, University of California, Irvine, CA, USA. *A list of authors and their affiliations appears at the end of the paper. ✉e-mail: xchen@pacificbiosciences.com; meberle@pacificbiosciences.com

conversion⁸. These high rates of gene conversion promote sequence exchange between SDs^{9,10}, further increasing the errors in read alignment. While short-read-based computational methods have been developed to improve the genotyping capability and diagnostic yield in segmental duplications^{5,11–15}, comprehensive variant calling in these regions remains a challenge, and SDs have not been studied at the population level by the current high throughput technologies.

Many medically relevant genes fall into SDs where traditional alignment-based analysis has not been demonstrated to reliably detect the full diversity of these regions. For example, spinal muscular atrophy is caused by variants in the *SMN1* gene, which has a highly similar paralog *SMN2*¹⁶. Another disease, 21-Hydroxylase-Deficient Congenital Adrenal Hyperplasia (21-OHD CAH), is caused by variants in the *CYP21A2* gene¹⁷, which resides in a 30 kb tandem repeat called the RCCX module and has a pseudogene *CYP21A1P*. Variants in the *OPN1LW/OPN1MW* gene cluster, which contains 1–5 copies of *OPN1LW* or its paralog *OPN1MW*, cause color vision deficiencies^{18,19}. To date, these medically important SD-encoded genes are studied with multi-step analyses including a combination of low or medium-throughput assays such as multiplex ligation-dependent probe amplification (MLPA), amplicon sequencing, or long-range PCR followed by Sanger sequencing to detect copy number changes or individual variants^{20,21}. These tests are sometimes limited to a few known variants and may be prone to false negatives if the patient has a pathogenic variant that is not part of the test. There remains a need to fully characterize these genes both for research and clinical testing.

Recently, researchers have begun to study SDs using long-read sequencing. High quality phased assemblies have been generated for a number of samples^{22–24} using PacBio HiFi and Oxford Nanopore Technologies (ONT) long reads, revealing the sequences of SDs and providing biological and evolutionary insights^{7,10}. However, SDs with multiple copies of highly similar regions are prone to assembly errors, especially in regions of extended sequence homology^{10,25}. Alternatively, we developed a phasing approach, Paraphase, that identifies haplotypes of genes and their paralogs, and demonstrated its ability to accurately resolve the highly similar *SMN1/SMN2* region²⁶. That study was limited to one difficult region, leaving a need for a genome-wide demonstration.

Here we extended Paraphase to analyze 316 paralogous genes that fall into 160 groups of SD regions across the genome, including many medically relevant genes that were traditionally considered challenging to genotype. Applying Paraphase to 259 individuals from five ancestral populations, we showed the genetic diversity of these regions across populations in copy number (CN) and sequence variation. We note that some of these regions show exceptionally low diversity between genes and paralogs, signaling selective pressures and/or high rates of gene conversion. Finally, we studied the Paraphase derived haplotypes for these paralogous genes in 36 parent-offspring trios and identified 11 de novo events, among which 7 are de novo single nucleotide variants (SNVs) and 4 are consistent with de novo gene conversion events.

Results

Profiling 160 gene-coding paralogous regions with Paraphase

Paraphase resolves highly similar genes by realigning HiFi reads to one, most relevant, gene chosen to represent all copies of the gene and its paralogs. We call this gene the archetype gene. For example, to study *SMN1* and *SMN2*, we realign all of the reads that are aligned to either *SMN1* or *SMN2* to just *SMN1* because that is the fully functional copy. The aligned reads are then phased into haplotypes for variant calling (Fig. 1a). For this study, we identified 160 paralogous regions >10 kb in length with >99% sequence similarity that were found between two and four times in GRCh38 (Supplementary Data 1, also see Methods). These paralogous regions encode 316 genes in total (excluding pseudogenes). In this paper, the term “paralog group” is used to describe a

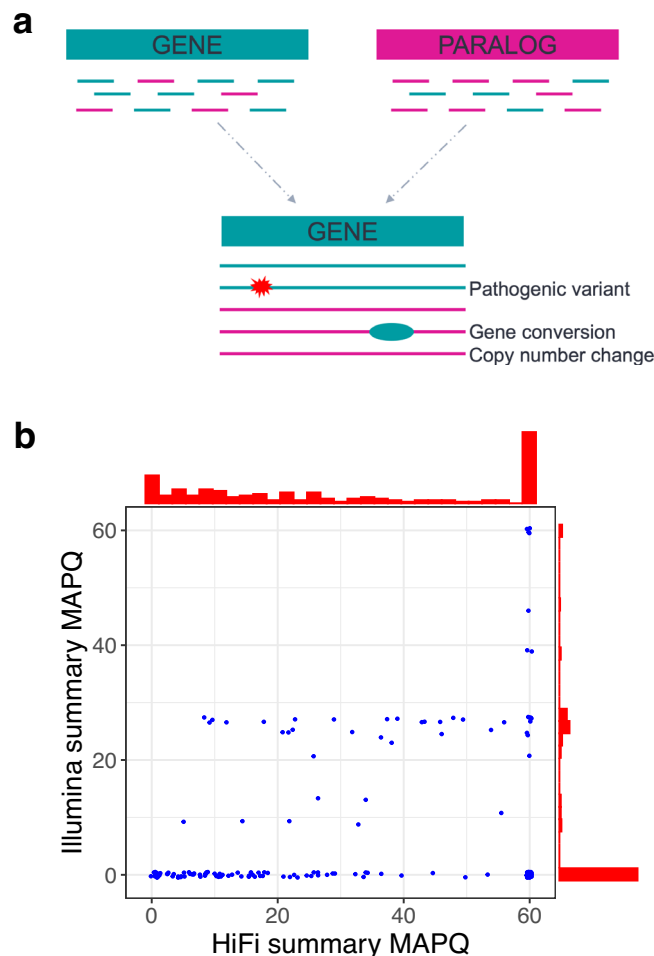


Fig. 1 | Paraphase design and the regions it analyzes. a Paraphase extracts read (short horizontal lines) that align to a paralog group (gene: green, paralog: magenta), realigns to the archetype gene, and phases reads into haplotypes (long horizontal lines). Variant calling is performed on each haplotype. Copy number changes can be identified from the number of haplotypes, e.g. an extra haplotype in this example indicates a copy number gain in the paralog. **b** Comparison of summary MAPQs between HiFi and Illumina WGS data in 160 groups of paralogous regions analyzed by Paraphase, highlighting mapping difficulty in these challenging regions for both short and long reads.

set of genes that are highly similar in sequence and are analyzed by Paraphase as a group.

Among the 160 paralog groups (Supplementary Data 1), 149 have genes located on the same chromosome, with 16 in tandem (less than 10 kb apart). To quantify the impact of sequence homology on read alignments, we examined the mapping qualities (MAPQs) in these regions in both short-read and long-read sequence data (Fig. 1b, see Methods) with two metrics: base MAPQ (median MAPQ of all reads overlapping a position) and summary MAPQ (median of base MAPQs across all positions of a paralog group). For short-read data, MAPQs are extremely low (76.4% of the paralog groups have a summary MAPQ ≤ 20, and 98.8% of the paralog groups have some bases with a base MAPQ ≤ 20), indicating the difficulty of mapping short reads to these regions. Even for long-read data, 44.1% of the regions have a summary MAPQ ≤ 20 and 75.2% have some bases with a base MAPQ ≤ 20. For long-read data, there are 25 (15.6%) paralog groups where the summary MAPQ is 60 and there are no bases with a base MAPQ ≤ 20. These are either regions where the sequence similarity is high but the homology extends less than the HiFi read length of ~15–20 kb, or regions which have lower sequence similarity but are included in Paraphase for fusion

Table 1 | Validated Paraphase calls in medically relevant genes

Gene	Number of samples	Variants	Number of alleles
CYP21A2	3	Gene deletion	1
		CYP21A1P-CYP21A2 fusion	1
		Pathogenic small variant	4
STRC	2	Gene deletion	3
		Pathogenic small variant	1
SMN1	1	0 copies of <i>SMN1</i> (Absence of c.840 C)	2
PMS2	4	Hybrid <i>PMS2</i> with <i>PMS2CL</i> sequence	1
		Pathogenic SV	3
OPN1LW/OPN1MW	5	Pathogenic small variant	3
		Pathogenic CNV and pathogenic small variant	2
IKBK	1	Pathogenic SV	1
NCF1	3	Gene deletion	4
		Pathogenic small variant	2
CFC1	2	Pathogenic small variant	2
Total	21		30

calling (see Methods). Paraphase analysis can still improve the performance in these high MAPQ regions because: 1) even reads with high MAPQ can be misaligned due to reference genome artifacts, common CNVs, and high rates of gene conversion, 2) gene fusions are hard to detect because split alignments are unlikely to happen in regions of homology and 3) lower MAPQs will be expected in data with shorter read length, such as in HiFi hybrid capture data.

Validation of Paraphase calls

We first validated Paraphase variant calls in 8 medically relevant genes in 21 disease or carrier samples identified using orthogonal methods such as MLPA and Sanger sequencing (Table 1 and Supplementary Data 2, also see Methods). For this validation, Paraphase correctly identified all 30 of the clinical variants in these samples.

We also examined haplotypes called by Paraphase in 36 trios. Among 14,734 full-length haplotypes called in the probands (also requiring full-length haplotypes called in the two parents of each trio), 14,679 (99.6%) agreed exactly with one of the haplotypes observed in the parents. Upon examining the 55 inconsistent cases, 43 (0.29%) are not fully supported by reads and thus determined as Paraphase errors (switch errors or missed haplotypes in the parent). The remaining 12 (0.081%) inconsistent haplotypes are fully supported by reads, and thus are true recombination or de novo events (See “Identification of de novo mutations and gene conversion” section).

In addition, we compared Paraphase variant calls against high quality diploid assemblies in 47 HPRC samples (See Supplementary Notes). 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). To better understand these differences, we manually reviewed the discrepant calls and determined that the reduced precision and recall in some paralog groups is mostly due to errors in the assembly (Supplementary Figs. 2 and 3, also see Supplementary Notes) such as individual base errors and misassembly of the paralog group leading to incorrect copy numbers.

The performance of Paraphase will depend on the sequence divergence between haplotypes, read length, and sequencing depth. We performed a variety of simulated experiments to estimate how these factors impact the accuracy of Paraphase (See Supplementary

Notes). Paraphase maintained high haplotyping accuracy with minimums of 10 kb read length, 10X per-haplotype sequencing depth and 0.05% sequence divergence (Supplementary Fig. 4, also see Supplementary Notes).

Copy number variability of paralog groups

We calculated the distribution of the total CN (defined by the number of unique haplotypes, adjusted by depth) of each paralog group in 259 unrelated individuals across five ancestral populations. We assessed the variability of the total CN by the percentage of individuals having the mode CN. For this study, we say that a paralog group has low CN variability if more than 90% of the individuals have the mode CN value, and medium CN variability if between 80% and 90% of the individuals have the mode CN value. Conversely, a paralog group is defined as having high CN variability if less than 80% of individuals have the mode CN value. Based on these definitions, 79 of the paralog groups have low CN variability, 17 have medium CN variability and 64 have high CN variability (Fig. 2a, Supplementary Data 1). Additionally, 25.6% (41/160) of the paralog groups had significant (Chi-squared test, $p < 0.05$, with Bonferroni correction) deviations between ancestral populations (Supplementary Fig. 5).

The CN variability can give us a general understanding of the population-level “accuracy” of the reference genome (in this case GRCh38). For example, an SD with two paralogous regions would always have a CN of four in our analysis if the reference is correct and generalizes across the population. Likewise, a paralog group where every individual has a CN of two in the population is likely a false SD in the reference. We identified 22 paralog groups where more than 95% of all individuals have a total CN of two (Supplementary Table 1, Fig. 2b). This suggests that duplications are rare in the population for these genes and these SDs could represent errors in the reference genome. Nineteen of these paralog groups overlap regions that were classified as false duplications in GRCh38 based on the CHM13 T2T assembly²⁷. Three of these paralog groups (*DEFB109B* and its SD, *CNTNAP3/CNTNAP3C* and *POTED* and its SD; see Supplementary Table 1) were not identified as false duplications in GRCh38 by the CHM13 T2T assembly, although they are only present once in the CHM13 assembly (Paraphase analysis of CHM13 data shown in Supplementary Fig. 6).

Conversely, we found three paralog groups (*CTAGE8/CTAGE9*, *OR2A1/OR2A42*, and *RIMBP3/RIMBP3B/RIMBP3C*) that are truly CN variable regions in the population (Fig. 2a) but were attributed to false duplications in GRCh38 due to missing genes in the CHM13 T2T assembly²⁷ (Paraphase analysis of CHM13 data shown in Supplementary Fig. 7). Population CN analysis provides a more accurate assessment of false duplications in a reference genome than the analysis of a single individual.

Paralog groups with exceptionally low within-group diversity

Paraphase identified 159,795 haplotypes from the 160 paralog groups in the 259 samples. Extensive gene conversion and unequal crossing over can result in highly similar gene copies that can no longer be separated into different genes based on sequence alone. For example, *SMN1* and *SMN2* are different in sequence in Exons 7–8 but are indistinguishable in Exons 1–6 indicating that gene conversion may be much more common in Exons 1–6 than in Exons 7–8²⁶. Thus, a principal component analysis (PCA) of haplotype sequences in Exons 7–8 can separate *SMN1* haplotypes and *SMN2* haplotypes into distinct clusters, but a PCA of haplotypes in Exons 1–6 does not differentiate the *SMN1* haplotypes from *SMN2* haplotypes (Supplementary Fig. 8).

To identify paralog groups with low within-group diversity, we developed a metric based on the divergence between individual haplotypes (see Methods). For example, in a paralog group with a gene and a paralog, the gene will evolve independently from the paralog in the absence of gene conversion. This means that the divergence will be lower between two copies of the gene (i.e. gene-gene divergence) or two

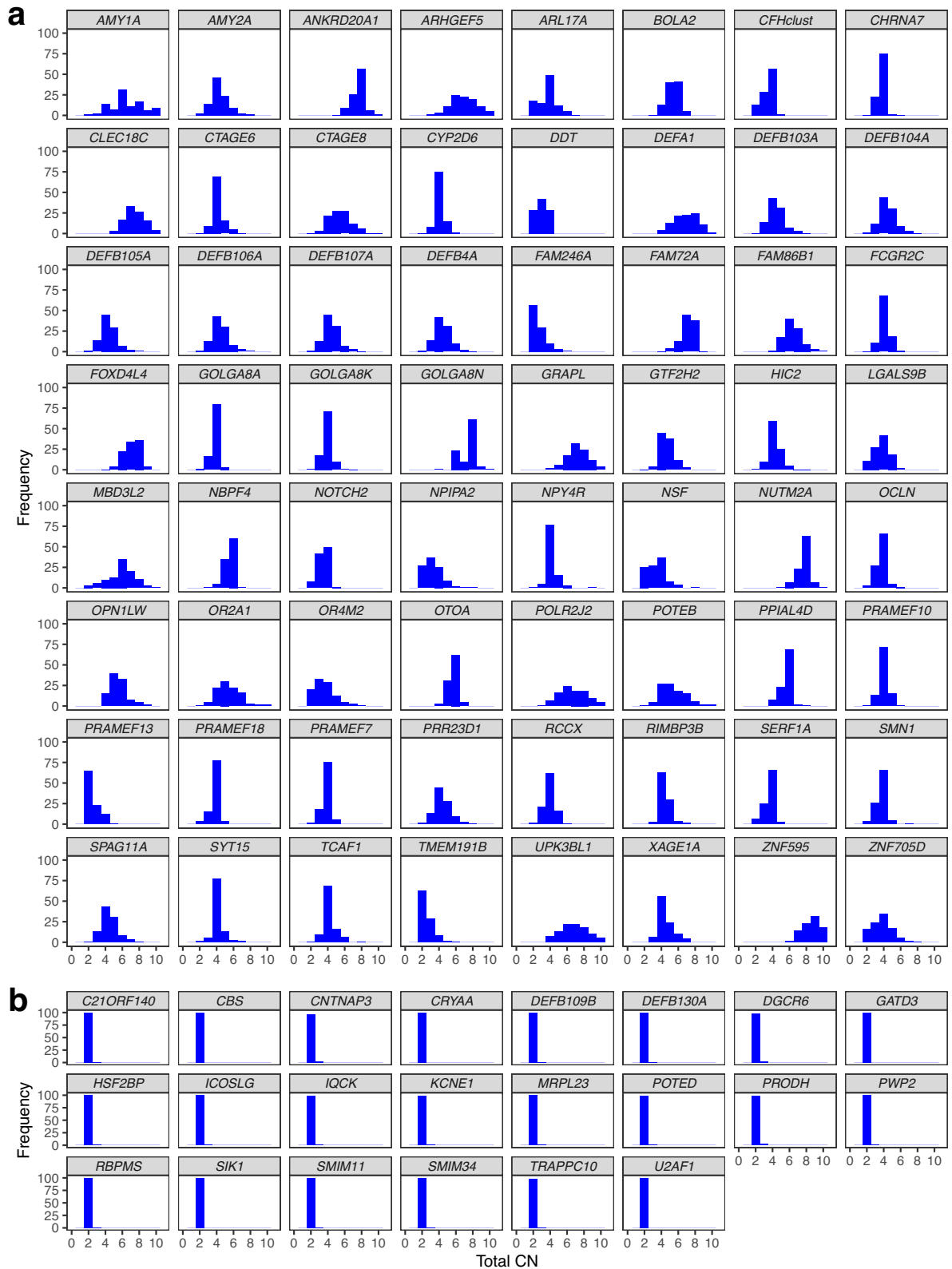


Fig. 2 | Distribution of the total CN of each paralog group across populations. One archetype gene is selected to represent the name of each group. **a** Paralog groups with high CN variability. For the two paralog groups (*OPN1LW* and *XAGE1A*)

located on the X chromosome, only female samples are plotted. **b** False duplication regions in GRCh38, where more than 95% of individuals have a total CN of two.

copies of the paralog in the absence of any selective pressures. Conversely, the gene-paralog divergence will be significantly higher (Supplementary Fig. 9). Increasing rates of gene conversion and unequal crossing over will tend to make the gene more similar to the paralog

and thus drive the gene-paralog divergence down (Supplementary Fig. 9).

We identified 23 paralog groups (termed low-diversity paralog groups) where the within-group sequence divergence is comparable to

Table 2 | Paralog groups with low within-group diversity

	Paralog groups	Palindromic	High CN variability	Human-specific duplication
chrY	<i>BPY2/BPY2B/BPY2C</i>	x		
	<i>CDY1/CDY1B</i>	x		
	<i>CDY2A/CDY2B</i>	x		
	<i>HSFY1/HSFY2</i>	x		
chrX	<i>CENPVL1/CENPVL2</i>	x		
	<i>CTAG1A/CTAG1B</i>	x		
	<i>CXorf49/CXorf49B</i>	x		
	<i>DMRTC1/DMRTC1B</i>	x		
	<i>FAM156A/FAM156B</i>	x		
	<i>MAGED4/MAGED4B</i>	x		
	<i>NXF2/NXF2B</i>	x		
	<i>SSX2/SSX2B</i>	x		
	<i>SSX4/SSX4B</i>	x		
	<i>TCP11X1/TCP11X2</i>	x		
	<i>XAGE1A/XAGE1B</i>	x	x	
Autosomes	<i>SLX1A/SLX1B</i>		x	x
	<i>BOLA2/BOLA2B</i>		x	x
	<i>SULT1A3/SULT1A4</i>		x	x
	<i>NPIPA2/NPIPA3</i>		x	x
	<i>AMY1A/AMY1B/AMY1C</i>		x	x
	<i>SERF1A/SERF1B</i>		x	x
	<i>EIF3C/EIF3CL</i>		x	
	<i>TRIM49D1/TRIM49D2</i>	x		

the general allelic sequence divergence (See Methods). Among these, 4 are on chrY, 11 on chrX, and 8 on autosomes (Table 2). It is often not easy to assign haplotypes of a paralog group to individual genes without prior knowledge of how genes and paralogs differ from each other. However, among the 23 low-diversity paralog groups, there are five where the phased haplotypes extend into non-homologous regions so that we can assign haplotypes to genes based on their flanking sequence: *AMY1A/AMY1B/AMY1C* (Fig. 3a), *CTAG1A/CTAG1B*, *BOLA2/BOLA2B*, *SULT1A3/SULT1A4* and *SLX1A/SLX1B* (*BOLA2/BOLA2B*, *SULT1A3/SULT1A4*, and *SLX1A/SLX1B* are three paralog groups in tandem and genotyped as one region by Paraphase). PCA of the haplotype sequences shows that haplotypes of the different genes of the same group do not form distinct clusters and thus are indistinguishable from each other by sequence alone (Fig. 3b–d, also see Supplementary Fig. 10).

The 23 low-diversity paralog groups show two different patterns in their genomic structure, CN variability and evolutionary history (Table 2). Those on autosomes have high CN variability and many are human-specific duplications (See Discussion). Conversely, low-diversity paralog groups on sex chromosomes mostly have low CN variability, are arranged in palindrome structures and evolutionarily conserved, i.e. all genes are present in other primates where they are also in palindromes^{28,29}. Additionally, there are 3 palindromic paralog groups on chrX where the genes and paralogs are in tandem so Paraphase can identify copies on the same chromosome. In these 3 paralog groups, the gene copies in cis are more similar to each other than those in trans (Fig. 3e), suggesting that gene conversion between arms of palindromes happens more frequently in cis (possibly through forming a hairpin structure) than in trans.

Identification of de novo mutations and gene conversion

In 36 parent-offspring trios we identified 12 events (6 paternal and 6 maternal) where a haplotype in the proband is different from the corresponding haplotype in the parent (Supplementary Figs. 11–12). Eleven of these are de novo events where the proband haplotype

differs from the parent haplotype by one SNV. Among these, 7 are de novo SNVs (not observed in either parent) and 4 are products of gene conversion (observed in a parent but on a different haplotype). Among the gene conversion cases, 2 are non-allelic (an example is shown in Fig. 4), 1 is allelic and 1 could be either allelic or non-allelic. Among the 11 de novo events, 4 are intergenic, 6 are in introns, and 1 is in an exon (synonymous). The remaining case of the 12 events is a hybrid haplotype between two haplotypes from the same parent, which could arise through equal or unequal crossing over (inconclusive without longer range phasing information in the parent due to the high copy number of the paralog group) (Supplementary Fig. 12).

Resolving medically relevant paralogous genes

As a demonstration of how Paraphase can be used to study paralog groups in the population, we examined variant and haplotype frequencies across populations in three known medically relevant paralog groups, *CYP21A2/CYP21A1P*, *PMS2/PMS2CL*, and *OPN1LW/OPN1MW*.

Variants in *CYP21A2* cause 21-Hydroxylase-Deficient Congenital Adrenal Hyperplasia (21-OHD CAH)¹⁷. *CYP21A2* resides in a 30 kb tandem repeat called the RCCX module that includes its pseudogene, *CYP21A1P*, together with two other pairs of paralogs, *C4A/C4B* and *TNXB/TNXA*^{17,20} (Fig. 5a). This region is susceptible to gene conversion¹⁷, as well as deletions and duplications of the RCCX module resulting in CN changes and disease-causing hybrid genes between *CYP21A2* and *CYP21A1P*. Here, complete haplotype sequences resolved by Paraphase allowed us to fully reveal the genetic diversity in this region. The total CN of RCCX is highly variable across populations (Fig. 5b) with 38.2% of individuals having a CNV. Figure 5a shows examples of samples with various CNs. In addition, we identified a duplication allele (Fig. 5a, bottom panel) that carries a copy of *CYP21A1P*, a copy of *CYP21A2* with a stop-gain variant Q319X, and a second functional copy of *CYP21A2*. We found that this allele is at 1–2% frequency in the populations (Supplementary Table 2) and, without phasing the full region, could be misidentified as a pathogenic allele

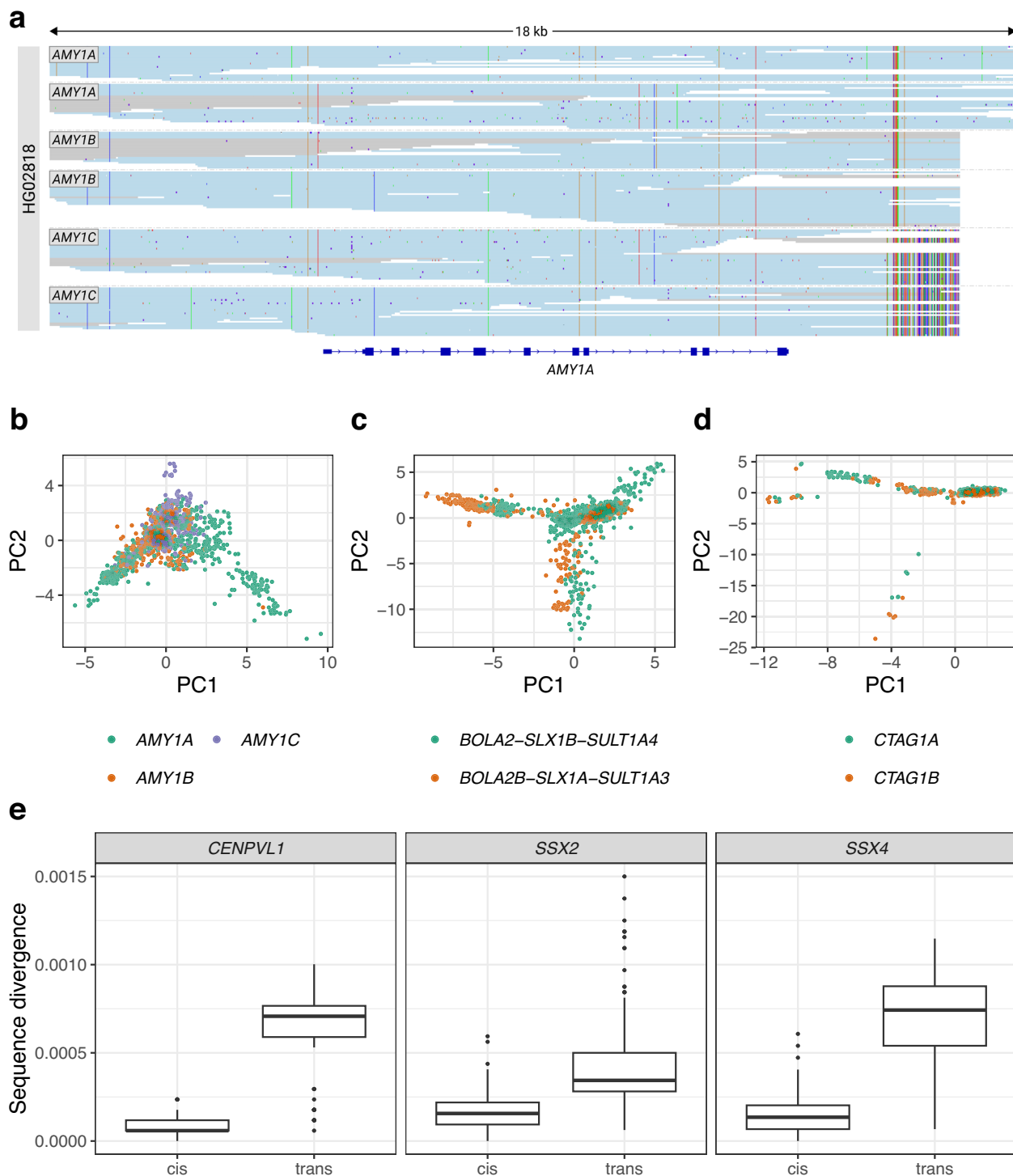


Fig. 3 | Paralog groups with low within-group diversity. a Haplotypes of the *AMY1* paralog group in a sample, realigned to *AMY1A*, showing two copies each of *AMY1A*, *AMY1B* and *AMY1C*. Reads in blue are consistent with a single haplotype. Reads in gray are consistent with more than one possible haplotype, i.e. when two or more haplotypes are identical over a region. The ends of the haplotypes extend into downstream non-homologous regions so we can assign the haplotypes into the three genes. **b–d** PCA of haplotype sequences of the *AMY1A/AMY1B/AMY1C* (**b**), *BOLA2-SLX1B-SULT1A4/BOLA2B-SLX1A-SULT1A3* (three paralog groups in tandem and genotyped as one region by Paraphrase) (**c**) and *CTAG1A/CTAG1B* (**d**). Each dot represents a haplotype in the population. Colors represent different genes in a

paralog group as assigned according to the ending sequences of each haplotype (which extends into non-homologous regions). **e** Sequence divergence between haplotypes in cis vs. trans in three palindromic paralog groups. Within each box-plot, the center lines denote median values; boxes extend from the 25th to the 75th percentile of each group's values; the whiskers extend from the box to the minimum (maximum) value that falls within 1.5 times the interquartile range below (above) the 25th (75th) percentile of each group; dots denote outlier values. One gene is selected to represent the name of each paralog group: *CENPVL1* for *CENPVL1/CENPVL2* (cis $n = 93$, trans $n = 80$), *SSX2* for *SSX2/SSX2B* (cis $n = 117$, trans $n = 163$), *SSX4* for *SSX4/SSX4B* (cis $n = 275$, trans $n = 308$).

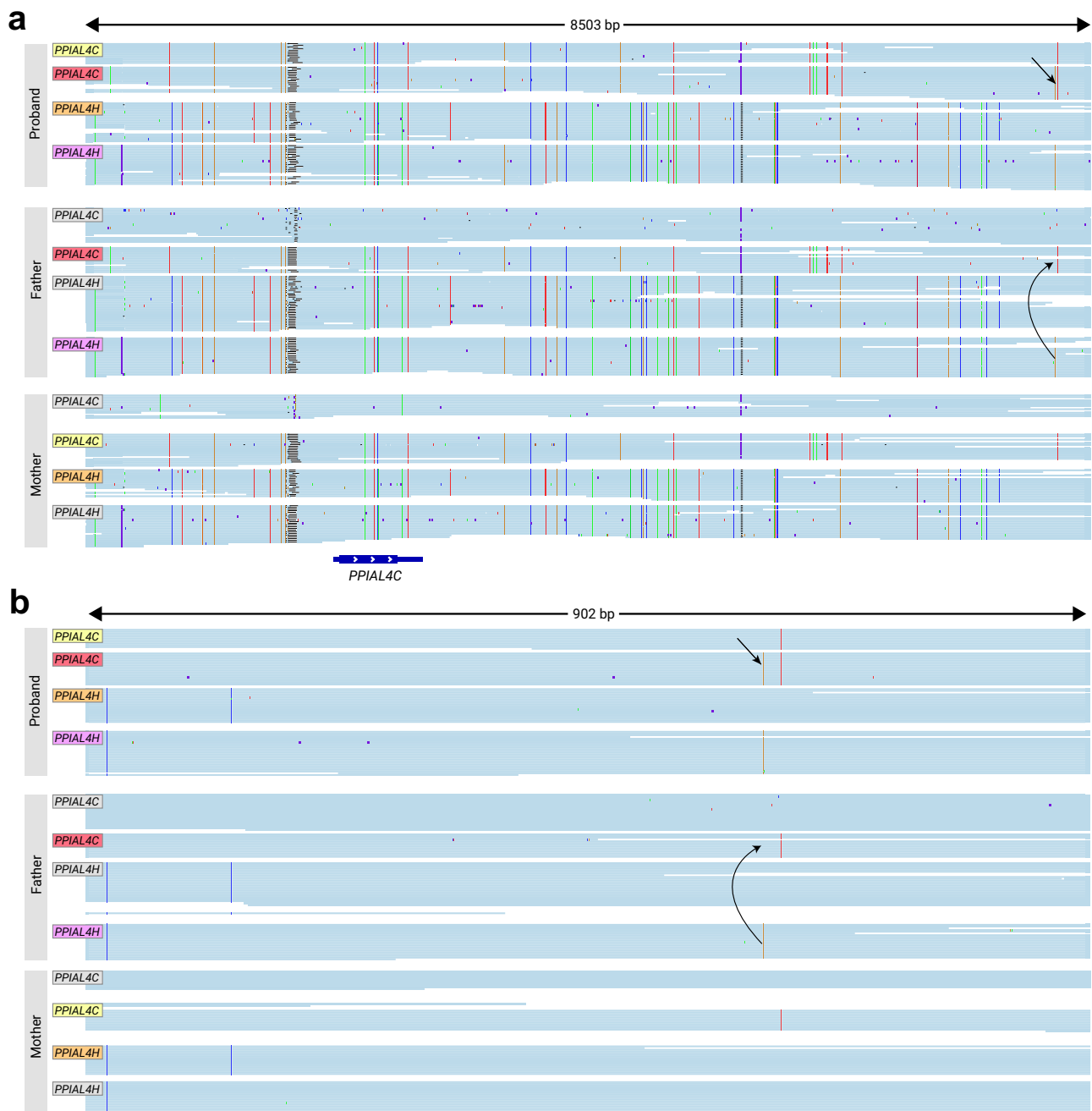


Fig. 4 | De novo non-allelic gene conversion in a trio. a Haplotypes are labeled in different colors in the proband, father, and mother, with matching colors indicating inherited haplotypes (haplotypes not inherited are labeled in gray in the parents). The black arrow denotes the SNV created by non-allelic gene conversion on haplotype 2 (labeled in red) of the proband. It is not present on the inherited haplotype,

haplotype 2 (labeled in red) in the father. Instead, it is present on haplotype 4 (labeled in magenta) of the father, which belongs to the other gene in the paralog group. The curved arrow shows the direction of the gene conversion. **b** Close view of the converted variant.

due to the presence of Q319X. Researchers have previously found that individuals with Q319X frequently have a duplication of *CYP21A2*, which complicates *CYP21A2* testing³⁰. Paraphase can distinguish a *CYP21A2* + *CYP21A2*(Q319X) allele vs. a *CYP21A2*(Q319X) allele.

Pathogenic variants in *PMS2* cause Lynch syndrome³¹. In its last few exons (Exons 12–15), *PMS2* has high sequence similarity to its pseudogene *PMS2CL*, and gene conversion and unequal crossing overs are known to promote sequence exchange between the two genes^{32–34}. We examined the haplotypes of *PMS2* and *PMS2CL* in the population. In Exon 15, the sequences of *PMS2* and *PMS2CL* are indistinguishable from each other, lacking any differentiating variants (Fig. 5c, also see PCA in

Supplementary Fig. 13). For example, a commonly considered *PMS2CL*-specific variant³⁵, NM_000535.7:c.*92dup, is in 72.8% of *PMS2CL* haplotypes and 31.1% of *PMS2* haplotypes. We found that gene conversion happens between *PMS2* and *PMS2CL* occasionally in Exon 12 and frequently in Exons 13–14 (See Methods and Fig. 5c, d). Interestingly, our analysis showed more evidence of gene conversion in individuals of African ancestry, and more than 75% of African *PMS2*/*PMS2CL* haplotypes are partially or fully converted (Fig. 5d).

OPN1LW and its paralog *OPN1MW* are responsible for red-green color vision deficiencies and other vision conditions such as blue cone monochromacy (BCM)¹⁸. The region is arranged in a gene array and

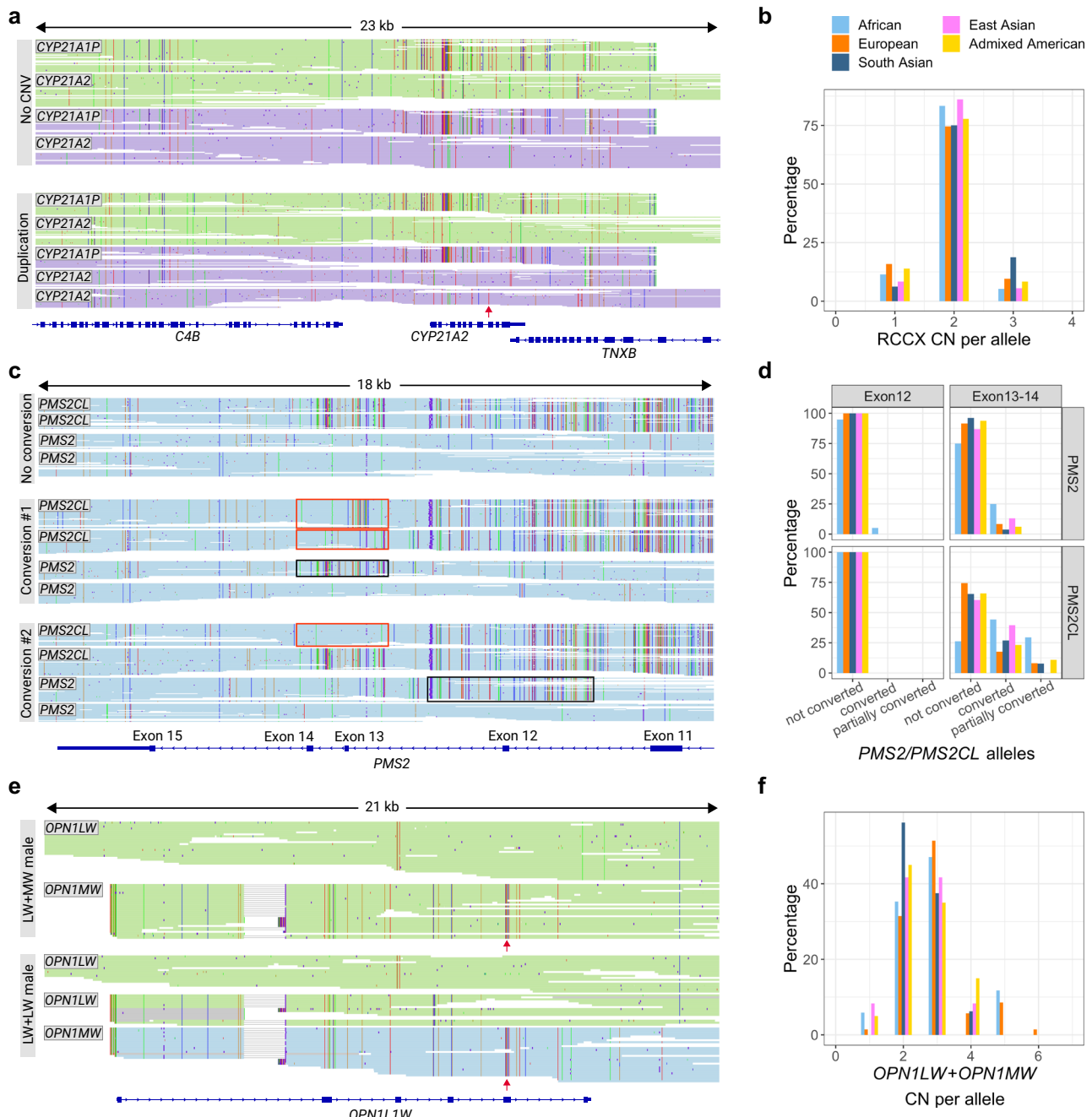


Fig. 5 | Population results in *CYP21A2*, *PMS2* and *OPN1LW/OPN1MW*. **a** Paraphase resolved haplotypes in the RCCX module, realigned to the RCCX copy that encodes *CYP21A2*. Haplotypes of the same color (purple or green) are from the same allele. Longer haplotypes represent the last RCCX copies in the array on each allele, and shorter haplotypes represent remaining copies. Two examples are shown, including a sample with no CNV (top) and a sample with RCCX duplication (bottom), which carries an allele (purple) with a wild-type copy of *CYP21A2* and another copy of *CYP21A2* harboring a pathogenic variant Q319X (red arrow). **b** Frequency of the total RCCX CN per allele across populations. **c** Paraphase resolved haplotypes in *PMS2*/*PMS2CL*, realigned to *PMS2*. Exon numbers are labeled with respect to *PMS2*. Three examples are shown, including a sample with no gene conversion and two samples carrying alleles converted in Exon 12 or Exons 13-14 (conversions in *PMS2*

shown in black boxes and conversions in *PMS2CL* shown in red boxes). **d** Frequency of gene conversions between *PMS2* and *PMS2CL* across populations in Exon 12 and Exons 13-14. **e** Paraphase resolved haplotypes in *OPN1LW/OPN1MW*, realigned to *OPN1LW*. Longer haplotypes represent the first copies of the repeat in the array on each allele, and shorter haplotypes represent remaining copies. The first two copies of the repeat on each allele are colored in green, and the blue color indicates gene copies beyond the second copy in the array, i.e. not expressed. *OPN1LW* and *OPN1MW* are assigned based on variants in Exon 5 (red arrows). Two examples are shown, including a normal allele with a copy of *OPN1LW* followed by a copy of *OPN1MW* (top) and an allele (*deutan*) with a copy of *OPN1LW* followed by a copy of *OPN1MW* (bottom, the third unexpressed copy marked in blue). **f** Distribution of the summed CN of *OPN1LW* and *OPN1MW* per allele across populations.

only the first two genes in the array are expressed¹⁸. Paraphase identifies all copies of the repeat, assigns genes to *OPN1LW* or *OPN1MW*, and identifies the first two copies in the array on each chromosome. Figure 5e shows an allele with one copy each of *OPN1LW* and *OPN1MW* (top panel), and an allele that only has *OPN1LW* in the first two copies

of the array (bottom panel), leading to color vision deficiencies. Our analysis showed that the total CN of this paralog group is highly variable among populations (Fig. 5f). Allele frequencies are summarized in Supplementary Table 3, including alleles that cause color vision deficiencies.

In addition to the three paralog groups described above, we also summarized population results for other medically relevant genes, including *SMN1/SMN2* (spinal muscular atrophy¹⁶, Supplementary Fig. 14), *STRC* (hereditary hearing loss and deafness³⁶, Supplementary Fig. 15), *HBA1/HBA2* (Alpha thalassemia³⁷, Supplementary Table 4), *IKBK* (Incontinentia Pigmenti³⁸, Supplementary Table 5), the *CFH* gene cluster (*CFH/CFHR1/CFHR2/CFHR3/CFHR4*) (atypical hemolytic uremic syndrome³⁹ and age-related macular degeneration^{40,41}, Supplementary Table 6) and *GBA* (Gaucher and Parkinson's disease^{42,43}, Supplementary Table 7). Together, we identified medically relevant variants in at least one of these 9 paralog groups in 75%, 49.6%, 45.8%, 52.2%, and 17.4% of individuals of African, European, Admixed American, South Asian and East Asian ancestries, respectively.

Discussion

In this paper, we applied Paraphase to 160 segmental duplication regions where large (>10 kb) regions of high (>99%) sequence similarity exist between genes and their paralogs. By phasing reads from the same paralog group together, Paraphase can recover misaligned reads and correctly resolve genes together with their highly similar paralogs/pseudogenes. This method enables high-throughput CN detection and genotyping of SD-encoded genes with only HiFi data at standard WGS depth (30X).

An important benefit of the paralog group-centered approach is that it is not influenced by the CN difference between an individual and the reference. This approach can work even when the CN of a paralog group does not agree with the reference genome in most individuals of the population, such as in the case of false segmental duplications in GRCh38. In addition, by calling variants against the same reference gene within a paralog group, Paraphase outputs gene copies that can be easily compared against each other, allowing us to perform within-group divergence analysis, as well as to detect de novo mutations including gene conversion events between paralogs. Our analysis of 36 trios identified 7 de novo SNVs and 4 de novo gene conversion events, demonstrating the power of long-read sequencing in detecting de novo variations^{44,45}, particularly in previously inaccessible regions of the genome.

Among regions analyzed by Paraphase, we observed that paralog groups on sex chromosomes are more CN invariable (93.1% of paralog groups on sex chromosomes have low CN variability vs. 39.7% on autosomes) and have drastically lower within-group diversity (median pairwise haplotype divergence 0.00033 on sex chromosomes vs. 0.00187 on autosomes, p -value 4.179e-11). This could be related to the fact that most paralog groups on sex chromosomes are arranged in palindrome structures (86.2% vs. 16.8% on autosomes). Unequal crossing-overs between arms of a palindrome results in inversions and do not change the copy number. Arm-to-arm gene conversion is known to occur frequently to prevent sex chromosomes from accumulating deleterious mutations in the absence of homologous chromosomes^{29,46}, and could contribute to the low within-group diversity.

We identified 23 paralog groups with extremely low within-group diversity (Table 2), where genes and their paralogs are as similar as alleles from the same gene. Consistent with most other paralog groups on sex chromosomes, the low-diversity paralog groups on sex chromosomes are all arranged in palindromes and mostly have low CN variability. For these paralog groups, both the genes and the palindrome structure are evolutionarily conserved in other primates. The low-diversity paralog groups on autosomes, however, are not arranged in palindrome structure and mostly have high CN variability. Interestingly, many of these paralog groups are duplicated exclusively in the human lineage, with positive selection detected, e.g. *AMY1A/AMY1B/AMY1C*⁴⁷, *BOLA2/BOLA2B*^{48,49} and *SULT1A3/SULT1A4*⁵⁰. It is possible that recent duplication and positive selection and/or gene conversion could play a role in the evolution of these genes,

preventing sequence divergence and maintaining an elevated gene dosage in humans. Beyond paralog groups with low within-group diversity throughout the entire gene body, one future direction is to extend this analysis to identify local low-diversity regions resulting from gene conversions, such as the gene conversion found in Exons 1-6 of *SMN1/SMN2* (Supplementary Fig. 8), and Exon 15 of *PMS2* (Fig. 5c, Supplementary Fig. 13).

The SD-encoded genes presented in this paper were previously inaccessible to population-wide genomic analyses and hence are largely missing from variant annotation databases such as gnomAD³, creating hurdles in variant interpretation. Here we provide a database (<https://zenodo.org/doi/10.5281/zenodo.10909886>) of variant allele frequencies collected from the population samples used in this paper. This annotation resource can be further expanded as more HiFi data are generated and analyzed with Paraphase.

One limitation of Paraphase is that currently it focuses on paralog groups with 2-4 paralogous genes in GRCh38 and does not include other highly similar genes with even higher CNs. This excludes 79 genes that fall into SDs in our analysis. Nevertheless, Paraphase can be customized to analyze user-specific regions, allowing new targets to be added in the future.

Paraphase, combined with HiFi long reads, provides a single framework for resolving paralogous genes. In medically important genes challenged by pseudogenes or paralogs, Paraphase helps enable more accurate testing to detect pathogenic variants, thus bringing us one step closer to consolidating the numerous currently offered genetic tests into a single test. Furthermore, in previously inaccessible and less studied genes, population-wide sequencing-based analysis with Paraphase will facilitate the discovery of novel gene-disease associations.

Methods

Paraphase: HiFi-based caller for highly similar paralogous genes

Paraphase is designed to work with both PacBio HiFi WGS and target enrichment data. Paraphase resolves a group of highly similar genes by extracting HiFi reads aligned to any member of the paralog group, realigning them to the archetype gene, and phasing them into haplotypes, followed by variant calling on each haplotype²⁶ (Fig. 1a). Briefly, haplotype phasing is achieved by identifying variant sites in the target region and reducing reads to just the bases at variant sites, followed by graph-based assembly of simplified reads into haplotypes. Reads are then assigned to the haplotypes that they correspond to. For each haplotype, variant calling (fully phased) is done by taking the consensus sequence at each position across reads that have been assigned to the same haplotype and reporting the base differences between the consensus and the reference. Realigning all reads from all genes of the same paralog group to one gene bypasses the error-prone process of aligning reads to multiple similar regions. This framework enables all copies of the paralog group, including genes and their paralogs or pseudogenes, to be examined for variants and annotated for functional status.

When two paralogous regions are in tandem, Paraphase uses read-based phasing to further phase gene haplotypes into alleles, i.e. gene copies on the same chromosome, by grouping haplotypes that have an overlapping set of supporting reads. For example, for the RCCX module demonstrated in Fig. 5a, reads are grouped by the haplotypes they originate from and haplotypes of the same color (green or purple) represent those from the same allele.

Gene fusions between paralogs are called by detecting haplotypes whose flanking regions (upstream and downstream of the paralogous region) are consistent with two different genes. Fusion breakpoints are called by detecting a switch in bases at the paralogous sequence variant (PSV) sites that have been carefully curated previously.

Within Paraphase, there are a few gene-specific callers for medically relevant genes. These callers use gene-specific information during analysis, for example, known sequence differences between genes and

paralogs/pseudogenes. In addition, these callers produce gene-specific output information such as hybrid gene structures and known pathogenic variants, including large difficult-to-call structural variants.

For a single WGS sample, across 160 paralog groups, Paraphase requires 4 Gb memory and typically completes analysis in 90 minutes (1 thread) or 15 minutes (16 threads). All results presented in this paper were generated using Paraphase V3.1.2.

Genome-wide identification of highly similar genes for analysis by Paraphase

We extracted 19,394 Ensembl protein-coding genes (>20 kb sequences centered on each gene, adding flanking sequences for shorter genes) and aligned them against GRCh38 (ALT contigs excluded, pseudoautosomal regions (PARs) masked) with blastn⁵¹ (Version 2.16.0+, command “-outfmt = ‘6 std qlen slen’ -word_size 50 -max_hsp 30”). We selected genes that had alignment matches >10 kb in length and >99% in sequence similarity as candidate paralog groups. Genes that do not meet the criteria are considered sufficiently different from the rest of the genome, and they are likely to have no alignment problems and should be genotyped correctly by the standard HiFi workflow. The majority of genes have zero paralogs, and the remaining ones vary in the number of paralogs (Supplementary Fig. 16). Among genes with three or fewer paralogs, which represent the majority of genes with paralogs, we incorporated 155 groups of regions into Paraphase. In addition, we included genes impacted by shorter homology or lower sequence similarity, where gene fusions are highly medically relevant yet difficult to call by conventional SV callers due to homology, including *HBA1/HBA2* (Alpha thalassemia), *GBA1/GBAP1* (Gaucher and Parkinson’s disease), *CYP2D6/CYP2D7* (pharmacogenomics), *CYP11B1/CYP11B2* (Glucocorticoid-remediable aldosteronism) and *CFH/CFHRI/CFHR2/CFHR3/CFHR4* (atypical hemolytic uremic syndrome and age-related macular degeneration). In total, Paraphase analyzes 160 groups of regions (Supplementary Data 1), which encode 316 genes in total (pseudogenes are not included).

Calculation of base MAPQ and summary MAPQ

We selected 20 samples from five ancestral populations with both Illumina (data downloaded from the 1000 Genomes Project¹) and HiFi WGS data available to assess alignment MAPQs. For each base position, we calculated a “base MAPQ”, defined as the median values across the MAPQs of all reads from all 20 samples aligned to the position. We then calculated a “summary MAPQ” for each paralog group, defined as the median value of the base MAPQ values across all base positions of that paralog group.

Validation against clinical samples with known variants

Validation samples were collected from 21 clinical samples (disease or carrier samples) with 30 pathogenic variants in 8 disease-causing genes that were previously validated by orthogonal methods⁵², such as MLPA and Sanger sequencing (Table 1 and Supplementary Data 2). Data (standard 30X HiFi WGS) for these clinical samples were collected from Radboud University Medical Center, GeneDx, and Genomics Research to Elucidate the Genetics of Rare diseases (GREGoR) Consortium. In addition, we used 36 trios to examine the consistency of haplotypes called in probands vs. parents. Among the 36 trio, 8 were collected from Radboud University Medical Center⁴⁵, 10 were from the 100,000 Genomes Project and 18 were from Genomics Research to Elucidate the Genetics of Rare diseases (GREGoR) Consortium.

Comparison against assemblies

We compared Paraphase variant calls against high-quality diploid assemblies in 47 HPRC samples²³. Contigs from assemblies were aligned with Minimap2⁵³ (Version 2.26-r1175, command “-x asm5”) against Paraphase target regions to identify haplotypes of each paralog group in the assembly. Variants in the assembly were called by

identifying base differences from the reference with a custom Python script. Segments from contigs were matched against Paraphase-called haplotypes by matching variants. Assembly-based variant calls were compared against Paraphase variant calls among matching haplotypes between Paraphase and the assembly. For this analysis, the assembly was defined as the ground truth. Thus, for Paraphase-called haplotypes that do not have a matching segment in the assembly, all variants called on those haplotypes were considered false positives by Paraphase. For segments in the assembly that do not have a matching Paraphase haplotype, all variants on those segments were considered false negatives by Paraphase.

Simulation analysis

We conducted simulation experiments to assess haplotyping accuracy when varying read length, haplotype depth and paralog divergence. Reference sequences were extracted from GRCh38 corresponding to Paraphase target regions (plus 50 kb flanking sequences). Five paralogous copies of each gene were simulated at different divergence levels (0.01%, 0.05%, 0.1%, and 0.15%) with Mutation-Simulator⁵⁴ (Version 3.0.2, command “-sn” for specified divergence levels). HiFi reads of different lengths (5 kb, 10 kb, and 15 kb) and depths (5X, 10X and 15X haplotype depth) were then simulated from those simulated gene copies with a workflow that consists of PBSIM3⁵⁵ (Version 3.0.4, command “--strategy wgs --method qshmm --qshmm QSHMM-RSIL.model --pass-num 8”), which simulated CLR reads, and the ccs software (Version 6.4.0, default parameters, <https://github.com/PacificBiosciences/ccs>), which generated consensus HiFi reads. For each target region, Paraphase was run on simulated data, with reads from all paralogous copies merged. Paraphase variant calls were compared against variants simulated by Mutation-Simulator. Precision and recall were calculated in the same way as described above in the “Comparison against assemblies” section.

Population samples

For frequency calculations, we used HiFi WGS data from 259 unrelated individuals from five ancestral populations (113 Europeans, 52 Africans, 48 Admixed Americans, 23 South Asians, and 23 East Asians), collected from the Human Pangenome Reference Consortium (HPRC)^{22,23}, the 100,000 Genomes Project, Radboud University Medical Center⁴⁵, Genomics Research to Elucidate the Genetics of Rare diseases (GREGoR) Consortium, and Genomic Answers for Kids (GA4K) at Children’s Mercy Kansas City.

Paralog groups with low within-group diversity

We searched for paralog groups within which the haplotype diversity is comparable to the general sequence diversity between alleles of the same gene. To profile the average allelic sequence divergence, we used Paraphase to phase haplotypes sequences for 600 randomly selected genes (400 on autosomes and 200 on chrX) that fall outside of SDs, i.e. each individual is expected to have two copies of a gene when there is no CNV, in the same set of 259 individuals. For each of the randomly selected genes, we calculated pairwise divergence values between any two haplotypes of the same gene among individuals of the same ancestral population. Between a pair of any two haplotypes, sequence divergence was calculated by dividing the number of base differences (SNVs only) by the length of the region. Focusing only on SNVs in non-homopolymer regions, 90% of the pairwise sequence divergence values among haplotypes were lower than 0.00156 for autosomal genes (Supplementary Fig. 17), and for chrX genes, 90% of the sequence divergence among haplotypes were lower than 0.00101 (Supplementary Fig. 17), consistent with a lower mutation rate on chrX⁵⁶. We then calculated pairwise sequence divergence values between any two haplotypes of the same paralog group targeted by Paraphase. To obtain candidate paralog groups where the within-group sequence divergence is as low as the general allelic diversity, we required that 90% of the

pairwise sequence divergence values among haplotypes of the same paralog group to be lower than 0.00156 and 0.00101 for autosomal and sex chromosome paralog groups, respectively. For both randomly selected genes and Paraphase paralog groups, only haplotypes from the same ancestral populations were compared for pairwise divergence calculation. For paralog groups reported in Table 2, we further filtered out paralog groups where the homology does not span the entire gene, i.e. partial paralogs. Principal component analysis (PCA) within a paralog group was conducted on SNV sites identified across all haplotypes of the paralog group using the `prcomp` function in R.

PMS2 gene conversion calling

PMS2 and *PMS2CL* haplotype sequences in Exon 12 region (chr7:5,981,000–5,985,000, GRCh38) and Exons 13–14 region (chr7:5,977,000–5,980,000) are separated into two main groups (*PMS2*-like and *PMS2CL*-like) based on the PCA (Supplementary Fig. 13). Variants (called against the *PMS2* reference sequence) that are present in >95% of the *PMS2CL*-like group and <5% of the *PMS2*-like group are selected as signature sites for calling gene conversion. Gene conversion is called when a *PMS2CL* haplotype has <20% of the signature variants or when a *PMS2* haplotype has >80% of the signature variants. A partial conversion in Exons 13–14 is a special haplotype common in the population, called based on a predefined subset of the signature variants (Fig. 5c, middle panel, first haplotype).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

HiFi WGS data for HPRC samples are documented in <https://github.com/orgs/human-pangenomics/repositories> and can be downloaded from <https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=working/> for analysis.

Code availability

Paraphase is implemented in Python and is freely available for download from GitHub (<https://github.com/PacificBiosciences/paraphase>). It can also be installed via Conda or PyPI.

References

1. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
2. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
3. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
4. Mandelker, D. et al. Navigating highly homologous genes in a molecular diagnostic setting: a resource for clinical next-generation sequencing. *Genet. Med. J. Am. Coll. Med. Genet.* **18**, 1282–1289 (2016).
5. Ebbert, M. T. W. et al. Systematic analysis of dark and camouflaged genes reveals disease-relevant genes hiding in plain sight. *Genome Biol.* **20**, 97 (2019).
6. Bailey, J. A., Yavor, A. M., Massa, H. F., Trask, B. J. & Eichler, E. E. Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res* **11**, 1005–1017 (2001).
7. Vollger, M. R. et al. Segmental duplications and their variation in a complete human genome. *Science* **376**, eabj6965 (2022).
8. Chen, J.-M., Cooper, D. N., Chuzhanova, N., Férec, C. & Patrinos, G. P. Gene conversion: mechanisms, evolution and human disease. *Nat. Rev. Genet.* **8**, 762–775 (2007).
9. Dumont, B. L. Interlocus gene conversion explains at least 2.7 % of single nucleotide variants in human segmental duplications. *BMC Genomics* **16**, 1–11 (2015).
10. Vollger, M. R. et al. Increased mutation and gene conversion within human segmental duplications. *Nature* **617**, 325–334 (2023).
11. Alkan, C. et al. Personalized copy-number and segmental duplication maps using next-generation sequencing. *Nat. Genet.* **41**, 1061 (2009).
12. Sudmant, P. H. et al. Diversity of human copy number variation and multicopy genes. *Science* **330**, 641–646 (2010).
13. Handsaker, R. E. et al. Large multiallelic copy number variations in humans. *Nat. Genet.* **47**, 296–303 (2015).
14. Prodanov, T. & Bansal, V. Robust and accurate estimation of paralog-specific copy number for duplicated genes using whole-genome sequencing. *Nat. Commun.* **13**, 3221 (2022).
15. Steyaert, W. et al. Systematic analysis of paralogous regions in 41,755 exomes uncovers clinically relevant variation. *Nat. Commun.* **14**, 6845 (2023).
16. Lunn, M. R. & Wang, C. H. Spinal muscular atrophy. *Lancet Lond. Engl.* **371**, 2120–2133 (2008).
17. Merke, D. P. & Auchus, R. J. Congenital adrenal hyperplasia due to 21-hydroxylase deficiency. *N. Engl. J. Med.* **383**, 1248–1261 (2020).
18. Neitz, J. & Neitz, M. The genetics of normal and defective color vision. *Vision Res.* **51**, 633–651 (2011).
19. Neitz, M. & Neitz, J. Intermixing the *OPN1LW* and *OPN1MW* Genes Disrupts the Exonic Splicing Code Causing an Array of Vision Disorders. *Genes* **12**, 1180 (2021).
20. Pignatelli, D. et al. The complexities in genotyping of congenital adrenal hyperplasia: 21-hydroxylase deficiency. *Front. Endocrinol.* **10**, 432 (2019).
21. Haer-Wigman, L. et al. Diagnostic analysis of the highly complex *OPN1LW/OPN1MW* gene cluster using long-read sequencing and MLPA. *Npj Genomic. Med.* **7**, 1–10 (2022).
22. Wang, T. et al. The Human Pangenome Project: a global resource to map genomic diversity. *Nature* **604**, 437–446 (2022).
23. Liao, W.-W. et al. A draft human pangenome reference. *Nature* **617**, 312–324 (2023).
24. Gao, Y. et al. A pangenome reference of 36 Chinese populations. *Nature* **619**, 112–121 (2023).
25. Porubsky, D. et al. Gaps and complex structurally variant loci in phased genome assemblies. *Genome Res.* **33**, 496 (2023).
26. Chen, X. et al. Comprehensive *SMN1* and *SMN2* profiling for spinal muscular atrophy analysis using long-read PacBio HiFi sequencing. *Am. J. Hum. Genet.* **110**, 240–250 (2023).
27. Nurk, S. et al. The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
28. Trombetta, B. & Cruciani, F. Y chromosome palindromes and gene conversion. *Hum. Genet.* **136**, 605–619 (2017).
29. Jackson, E. K. et al. Large palindromes on the primate X Chromosome are preserved by natural selection. *Genome Res.* **31**, 1337–1352 (2021).
30. Lekarev, O. et al. Erroneous prenatal diagnosis of congenital adrenal hyperplasia owing to a duplication of the *CYP21A2* gene. *J. Perinatol. J. Calif. Perinat. Assoc.* **33**, 76–78 (2013).
31. ten Broeke, S. W. et al. Cancer risks for *PMS2*-Associated lynch syndrome. *J. Clin. Oncol.* **36**, 2961–2968 (2018).
32. Hayward, B. E. et al. Extensive gene conversion at the *PMS2* DNA mismatch repair locus. *Hum. Mutat.* **28**, 424–430 (2007).
33. van der Klift, H. M. et al. Quantification of sequence exchange events between *PMS2* and *PMS2CL* provides a basis for improved mutation scanning of lynch syndrome patients. *Hum. Mutat.* **31**, 578–587 (2010).
34. Ganster, C. et al. Functional *PMS2* hybrid alleles containing a pseudogene-specific missense variant trace back to a single ancient intrachromosomal recombination event. *Hum. Mutat.* **31**, 552–560 (2010).
35. Wernstedt, A. et al. Improved multiplex ligation-dependent probe amplification analysis identifies a deleterious *pms2* allele generated

- by recombination with crossover between *PMS2* and *PMS2CL*. *Genes. Chromosomes Cancer* **51**, 819–831 (2012).
36. Yokota, Y. et al. Frequency and clinical features of hearing loss caused by *STRC* deletions. *Sci. Rep.* **9**, 4408 (2019).
 37. Farashi, S. & Hartevel, C. L. Molecular basis of α -thalassaemia. *Blood Cells. Mol. Dis.* **70**, 43–53 (2018).
 38. Cammarata-Scalisi, F., Fusco, F. & Ursini, M. V. Incontinentia pigmenti. *Actas Dermosifiliogr* **110**, 273–278 (2019).
 39. Zipfel, P. F. et al. Deletion of complement factor H-related genes *CFHR1* and *CFHR3* is associated with atypical hemolytic uremic syndrome. *PLoS Genet* **3**, e41 (2007).
 40. Klein, R. J. et al. Complement factor H polymorphism in age-related macular degeneration. *Science* **308**, 385 (2005).
 41. Hughes, A. E. et al. A common *CFH* haplotype, with deletion of *CFHR1* and *CFHR3*, is associated with lower risk of age-related macular degeneration. *Nat. Genet.* **38**, 1173–1177 (2006).
 42. Hruska, K. S., LaMarca, M. E., Scott, C. R. & Sidransky, E. Gaucher disease: mutation and polymorphism spectrum in the glucocerebrosidase gene (*GBA*). *Hum. Mutat.* **29**, 567–583 (2008).
 43. Sidransky, E. et al. Multicenter analysis of glucocerebrosidase mutations in Parkinson's disease. *N. Engl. J. Med.* **361**, 1651–1661 (2009).
 44. Noyes, M. D. et al. Familial long-read sequencing increases yield of de novo mutations. *Am. J. Hum. Genet.* **109**, 631–646 (2022).
 45. Kucuk, E. et al. Comprehensive de novo mutation discovery with HiFi long-read sequencing. *Genome Med.* **15**, 34 (2023).
 46. Swanepoel, C. M., Gerlinger, E. R. & Mueller, J. L. Large X-linked palindromes undergo arm-to-arm gene conversion across mouse lineages. *Mol. Biol. Evol.* **37**, 1979–1985 (2020).
 47. Perry, G. H. et al. Diet and the evolution of human amylase gene copy number variation. *Nat. Genet.* **39**, 1256–1260 (2007).
 48. Nuttle, X. et al. Emergence of a *Homo sapiens*-specific gene family and chromosome 16p11.2 CNV susceptibility. *Nature* **536**, 205–209 (2016).
 49. Giannuzzi, G. et al. The human-specific *BOLA2* duplication modifies iron homeostasis and anemia predisposition in chromosome 16p11.2 autism individuals. *Am. J. Hum. Genet.* **105**, 947–958 (2019).
 50. Butcher, N. J. et al. Sulfotransferase 1A3/4 copy number variation is associated with neurodegenerative disease. *Pharmacogenomics J* **18**, 209–214 (2018).
 51. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
 52. Höps, W. et al. HiFi long-read genomes for difficult-to-detect, clinically relevant variants. *Am. J. Hum. Genet.* **112**, 450–456 (2025).
 53. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
 54. Kühl, M. A., Stich, B. & Ries, D. C. Mutation-Simulator: fine-grained simulation of random mutations in any genome. *Bioinform. Oxf. Engl* **37**, 568–569 (2021).
 55. Ono, Y., Hamada, M. & Asai, K. PBSIM3: a simulator for all types of PacBio and ONT long reads. *NAR Genomics Bioinforma* **4**, lqac092 (2022).
 56. Hodgkinson, A. & Eyre-Walker, A. Variation in the mutation rate across mammalian genomes. *Nat. Rev. Genet.* **12**, 756–766 (2011).

Acknowledgements

We thank John Belmont and Evan Eichler for their valuable comments for the manuscript. We acknowledge colleagues from the diagnostic division of the Human Genetics Department from Radboudumc (Genome Diagnostics Nijmegen) and the Radboud Genomics Technology Center for their technical assistance and the library preparation and sequencing of all samples provided from Radboudumc, in particular Ronny Derks, Amber den Ouden, Janneke Weiss, and Lonneke Haer-Wigman. We

thank the Human Pangenome Reference Center (HPRC) for generating and releasing the HiFi WGS data. We thank the Genomic Answers for Kids (GA4K) program at Children's Mercy Kansas City for generating the GA4K HiFi sequencing data. This research was made possible through access to the 100,000 Genomes Project data in the National Genomic Research Library, which is managed by Genomics England Limited (a wholly owned company of the Department of Health and Social Care). The National Genomic Research Library holds data provided by patients and collected by the NHS as part of their care and data collected as part of their participation in research. The National Genomic Research Library is funded by the National Institute for Health Research and NHS England. The Wellcome Trust, Cancer Research UK and the Medical Research Council have also funded research infrastructure.

Author contributions

X.C. and M.A.E. designed the study; X.C. implemented the software and performed the analyses; X.C. and M.A.E. wrote the paper; D.B., Egor D., J.M.D., J.N., A.S.B., R.B., K.S.H., L.L., P.K., S.N., E.F., I.T., T.P., D.K., C.G., L.V., A.H., S.B., E.V., and Emmanuèle D. provided samples, data and feedback for the paper.

Competing interests

X.C., D.B., Egor D., and M.A.E. are employees of PacBio. J.M.D., J.N., A.S.B., R.B., K.S.H., L.L., P.K. and S.N. are employees of GeneDx. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-025-57505-2>.

Correspondence and requests for materials should be addressed to Xiao Chen or Michael A. Eberle.

Peer review information *Nature Communications* thanks Matteo Chiara, Mark Ebbert and Timofey Prodanov for their contribution to the peer review of this work. A peer review file is available.


Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025

UCI Genomics Research to Elucidate the Genetics of Rare diseases (UCI GREGoR) Consortium

Andrea Cohen¹², Arthur Ko¹², Georgia Pitsava¹³, Lightning Auriga¹³, Ya Cui¹³, Changrui Xiao¹³, Charles Hadley King¹³, Ivan de Dios¹³, Jonathan LoTempio¹³, Miguel Almalvez¹³, Vincent Fusaro¹³, Wei Li¹³, Jamie Fraser¹², Meghan Delaney¹², Leandros Boukas¹², Bojan Losic¹⁴, Esther Yun-Hua Hsiao¹⁴, John Harting¹⁴, Krista Bluske¹⁴, Rachid Karam¹⁴, Sami Belhadj¹⁴, Seth Berger ¹², Eric Vilain¹³ & Emmanuèle Délot¹³

¹⁴Ambry Genetics, Aliso Viejo, CA, USA.