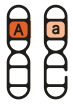
illumina®

APPLICATION NOTE

Deliver haplotype-resolved, copy-aware small variant detection in paralogous regions within a single WGS workflow:

Enabled by Illumina TruPath™ Genome using Multi-Region Joint Detection



Haplotype-resolved variant calling in complex, genomic research relevant paralogous genes

Simplified workflow reduces time to result from weeks to days

Unified genome-wide research workflow that enables assessment of historically difficult-to-characterize genomic regions

Introduction

Many important genes reside in paralogous or highly homologous regions of the genome where standard short-read whole-genome sequencing (WGS) struggles with ambiguous read mapping. At least 5% of the human genome consists of segmental duplications, and it is estimated that 200–500 genes contain regions where high sequence homology is a primary concern for variant calling.[1-3] Genes such as *SMN1/SMN2, PMS2, CYP21A2, STRC, NCF1,* and *CYP2D6* share 98–99.9% sequence identity with their respective paralogs or pseudogenes, complicating accurate read mapping. When reads map ambiguously between a gene and its paralog, conventional variant callers may discard informative reads or report them at incorrect locations, resulting in false negatives and false positives.

Current laboratory workflows address these challenging paralogous genes often through a cascade of orthogonal assays, such as multiplex ligation-dependent probe amplification (MLPA), long-range PCR (LR-PCR), long-read sequencing, or other sequencing methods. Each of these secondary tests adds days to weeks, requires different technologies, and increases operational burden. MLPA and long-range PCR–based approaches, are generally limited to resolving variants within the genomic regions targeted by the assay design and may not provide comprehensive haplotype resolution across larger loci.

TruPath Genome overcomes these barriers by encoding long-range molecular linkage information directly on the flow cell during sequencing. DNA molecules are captured and tagmented in spatially proximal nanowells, preserving physical long-range phasing information between reads derived from the same original DNA molecule. When paired with the DRAGEN™ Multi-Region Joint Detection (MRJD) algorithm, this long-range phasing information enables haplotype-resolved small variant calling across historically difficult-to-characterize genomic regions from standard short reads. Rather than stopping at a partial result and triggering follow-on testing, TruPath Genome is designed to produce a comprehensive genotype characterization directly, within a much shorter timeframe (on the order of days rather than weeks).

Here, we demonstrate that TruPath Genome coupled with MRJD produces haplotype-resolved, copy-aware genotypes across eight genomic-research–relevant paralogous loci in 31 samples. Results were consistent with those obtained using orthogonal research methods, including long-read sequencing approaches and were generated within three days of DNA extraction.

Methods

Sample preparation and sequencing

DNA (350 ng per sample) was added to a library tube strip together with a TruPath Genome reagent (Illumina, Cat. No. 20157406). For sequencing, the prepared library tube strip, two TruPath Genome reagents, and a NovaSeq X Series C8 flow cell were loaded onto the NovaSeq X Plus System (Illumina, Cat. No. 20084804) and sequenced following the user manual under standard run conditions.

Samples were derived from saliva, EDTA blood, and buccal swab sources and were processed using a range of DNA extraction methods encompassing different purification chemistries, including magnetic beads, glass beads, ethanol precipitation, and magnetic disk-based approaches. Both high molecular weight (HMW) and non-HMW DNA samples were included. DNA quality varied substantially across samples. The proportion of DNA fragments >10 kb ranged from approximately 27.8% to 93.3%, while the proportion of fragments >60 kb ranged from <1% to 68.6%, reflecting a broad distribution of input fragment sizes. Although a minimum threshold of 50% of fragments >10 kb is typically recommended, several samples did not meet this criterion yet still demonstrated strong performance in MRJD calling.

All samples were de-identified and used in accordance with the informed consent.

Bioinformatics workflow

After sequencing, the DRAGEN™ Germline pipeline with proximity mode enabled (--enable-proximity=true) was to utilize the proximity information. The GRCh38 (hg38) reference genome was used for all analyses.

Multi-Region Joint Detection approach

Conventional variant callers rely on the aligner to first uniquely map sequence reads to their genomic location before checking for small differences. This method works well when the read or read pairs resemble only one part of the reference genome, but it struggles when two or more regions match equally well. If a group of reads is mapped with low confidence, a conventional variant caller might ignore the reads, even though they contain useful information. If a read is mismapped (that is, if the primary alignment is not the true source of the read), it can result in variant detection errors.

To address these challenges, a novel computational method called Multi-Region Joint Detection (MRJD) was developed. Instead of considering each region in isolation and genotyping them individually, MRJD considers all locations from which a group of reads may have originated and attempts to detect the underlying sequences jointly.

With TruPath Genome data, MRJD constructs haplotypes by exploiting long-range physical linkage encoded in the proximity information of the sequenced reads. This approach enables highly sensitive germline small variant calling in paralogous regions using standard short-read sequencing data (see link). However, variant calls remain unphased, and variant-to-region assignment can still be ambiguous—an inherent limitation of standard short-read sequencing. The algorithm performs the following steps (Figure 1):

1. **Read collection.** All primary alignments within the paralogous regions of interest are collected, regardless of mapping quality.
2. **Copy number estimation.** Total copy number for each paralogous region set is estimated by comparing read depth across the regions of interest to a panel of pre-selected stable reference regions elsewhere in the genome.
3. **Haplotype construction.** Using the estimated copy number, read sequences, and proximity linking information from TruPath Genome, MRJD constructs all individual copies for each paralogous region set.
4. **Copy assignment.** For non-tandem paralogs, in which the copies are located at distant genomic regions in the reference (e.g., *SMN1* vs. *SMN2*), proximity information is used to assign each reconstructed copy to the genomic locus from which it most likely originated. In contrast, for tandem paralogs, where the copies are located in close proximity in the reference genome (e.g., *CYP21A2* vs. *CYP21A1P*), proximity information instead enables assignment of copies to their respective maternal or paternal haplotypes.
5. **Variant calling.** Small variants are called from the constructed copies and reported together with their assigned genomic regions or haplotypes.

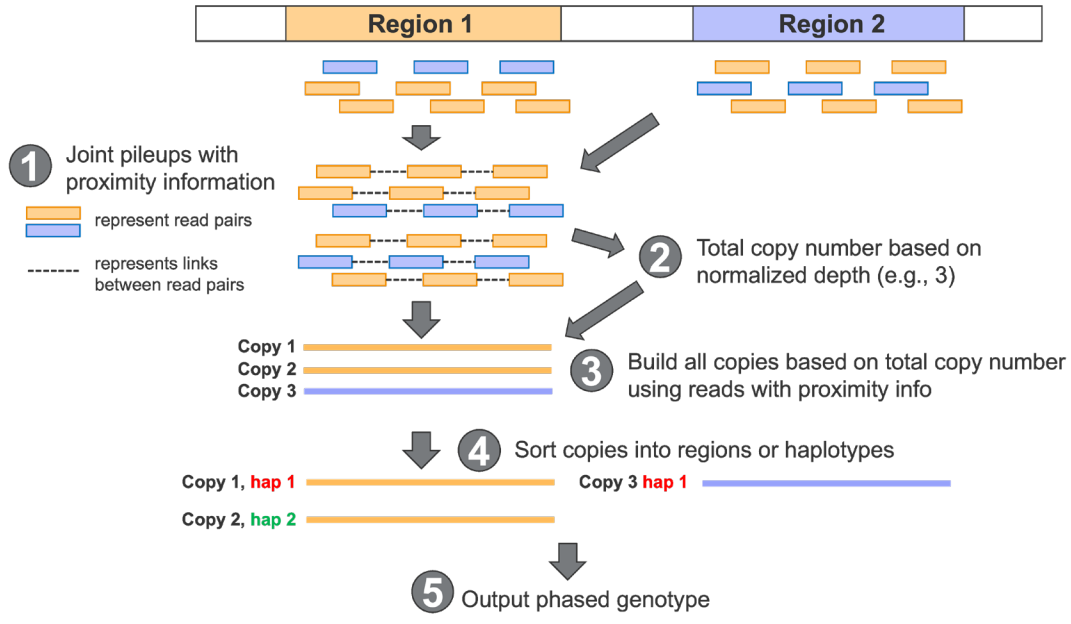


Figure 1: An overview of the MRJD Workflow using TruPath Genome data.

Evaluated loci

We evaluated TruPath Genome with MRJD across 31 samples spanning eight relevant paralogous loci (Table 1). These regions have 98.2–99.9% sequence homology and sizes vary from 18 to 357 kb, representing a broad range of genomic complexities including recombination, gene conversion, and copy number changes.

Table 1: Summary of pathogenic variant detection across eight paralogous loci using TruPath Genome with MRJD. *CFC1*, *IKBKG*, and *HYDIN* are not yet supported in TruPath Genome with MRJD in DRAGEN v4.5 but are planned for inclusion in a future release. In DRAGEN v4.5, additional loci supported by TruPath Genome with MRJD include *CYP2D6*, *TNXB* (high-homology region), *CYP11B1*, *CFHR1–4*, and *USP18*.

Gene	**Variant types**	**Number of expected events observed with TruPath Genome**
SMN1/SMN2	*SMN1* copy number loss	9/9
PMS2	Del/Dup, small variant	7/7
CYP21A2	Recombination, small variant	5/5
STRC	*STRC* copy number loss, small variant	3/3
NCF1	*NCF1* copy number loss	3/3
CFC1	Small variant	1/1
IKBKG	Exon 4-10 del	2/2
HYDIN	Small variant	1/1

Complete list of genes supported in TruPath Genome with MRJD in DRAGEN v4.5

Results

The following gene-specific case examples provide examples of TruPath Genome's performance advantages.

Spinal muscular atrophy (*SMN1/SMN2*)—9 cases

TruPath Genome produced fully phased *SMN1* and *SMN2* copies and accurately quantified *SMN1* copy number in both homozygous (n=4) and heterozygous (n=5) deletion backgrounds. Each copy was individually reconstructed and assigned to *SMN1* or *SMN2*, providing complete sequence for each copy and allelic configuration that is not achievable with standard short-read methods or MLPA alone (Figure 2).

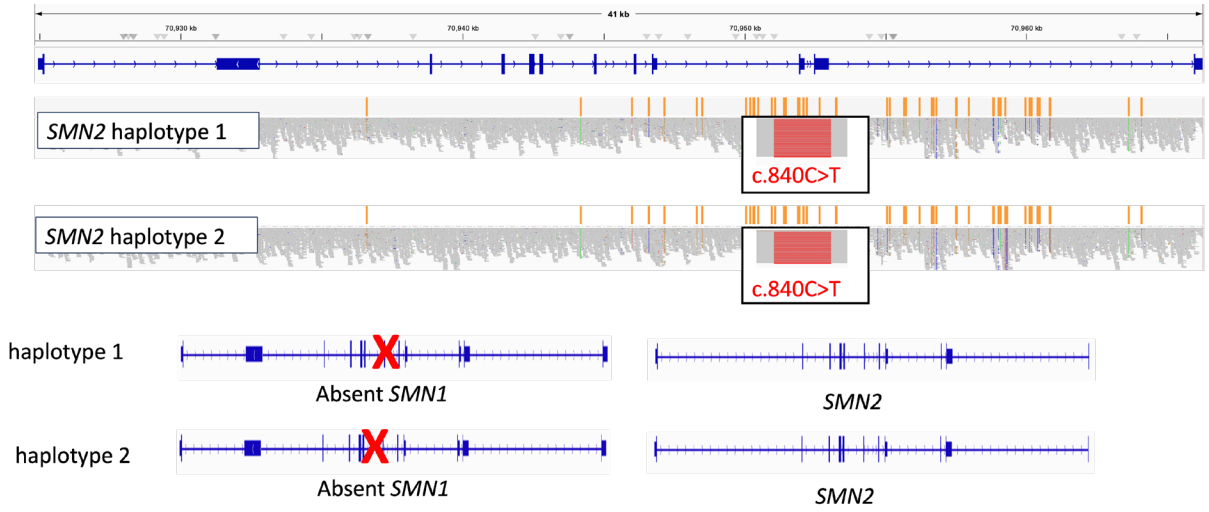


Figure 2: Haplotype-resolved detection of homozygous *SMN1* deletion in a spinal muscular atrophy case.

Lynch syndrome (PMS2)—7 cases

TruPath Genome resolved all haplotypes across the high-homology *PMS2/PMS2CL* region and correctly assigned pathogenic variants to *PMS2* versus *PMS2CL* in all seven cases. Two cases demonstrated particular advantages over existing methods:

- A tandem duplication in *PMS2CL* that was non-informative via MLPA was fully resolved by TruPath Genome.
- TruPath Genome detected a homozygous *PMS2CL* deletion and a heterozygous gene conversion from *PMS2CL* into the *PMS2* high-homology region—a configuration consistent with prior characterization of the sample. By contrast, MLPA reported only a total copy number of two and could not distinguish *PMS2* from *PMS2CL* (Figure 3).

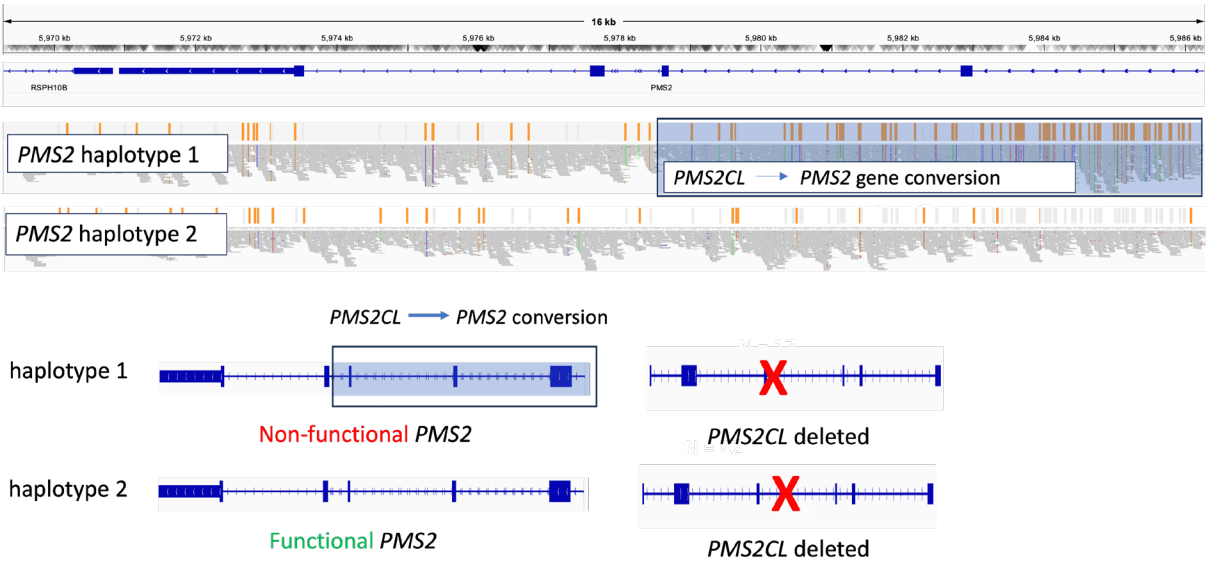


Figure 3: Resolution of *PMS2CL*-to-*PMS2* gene conversion and homozygous *PMS2CL* deletion in a Lynch syndrome case.

Congenital adrenal hyperplasia (CYP21A2)—5 cases

For all congenital adrenal hyperplasia (CAH) cases, TruPath Genome reconstructed all *CYP21A2*, *CYP21A1P*, and fusion copies, and grouped each copy into maternal or paternal haplotypes, which is far beyond the resolution of MLPA and LR-PCR (Figure 4). In one previously unresolved affected proband, TruPath Genome detected two *CYP21A1P* copies and one *CYP21A2* copy on one haplotype, and two *CYP21A2* copies on the other—all carrying pathogenic variants.

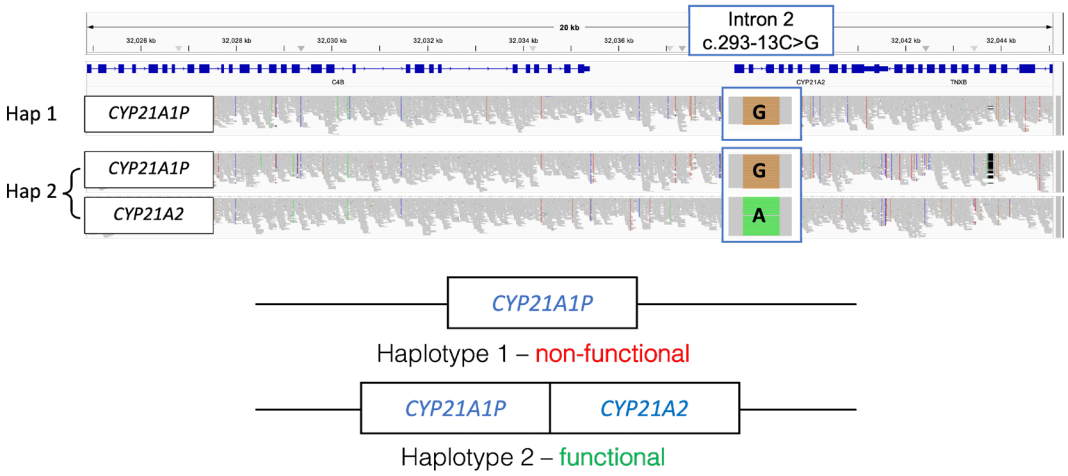


Figure 4: MRJD reconstructed the *CYP21A2* locus architecture and assigned copies to parental haplotypes.

Hearing loss (STRC)—1 case

In a hearing loss case, the phenotype indicated an affected individual, yet the SOC assay reported a heterozygote based on a copy-number deletion of *STRC*. TruPath Genome detected a complete deletion of one *STRC* allele and a pathogenic small variant in the other, confirming biallelic *STRC* disruption (Figure 5). This case illustrates how compound heterozygous disruption involving both a structural and a sequence variant can be missed by assays that assess copy number alone.

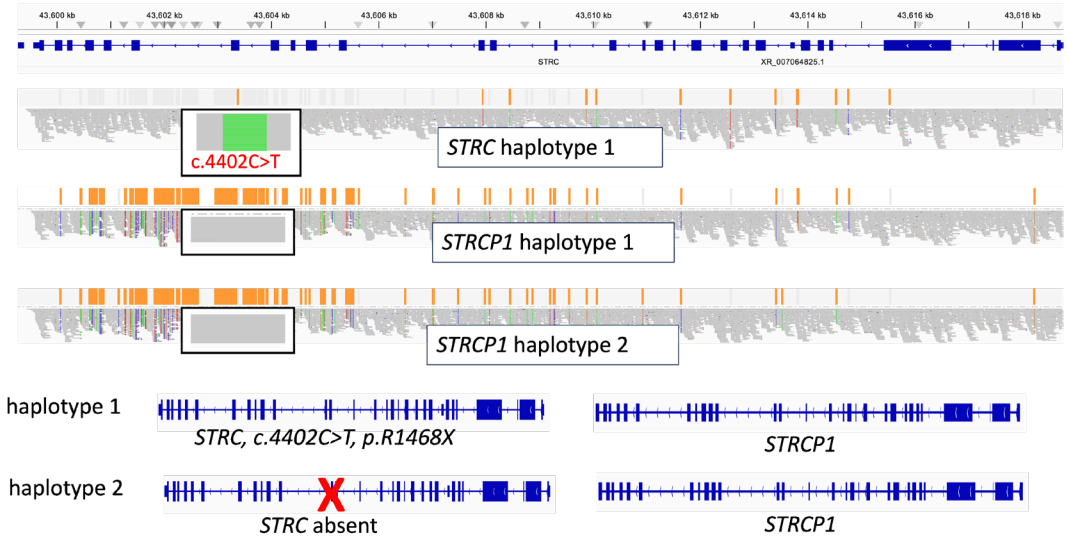


Figure 5: Detection of biallelic *STRC* disruption in a hearing loss case.

Chronic granulomatous disease (NCF1)—3 cases

TruPath Genome built all *NCF1* and pseudogene copies across the three cases. In one case, all six copies carried the pathogenic NM_000265.7(NCF1):c.75_76del variant allele suggesting origin via gene conversion rather than independent occurrence of the variant allele (Figure 6).

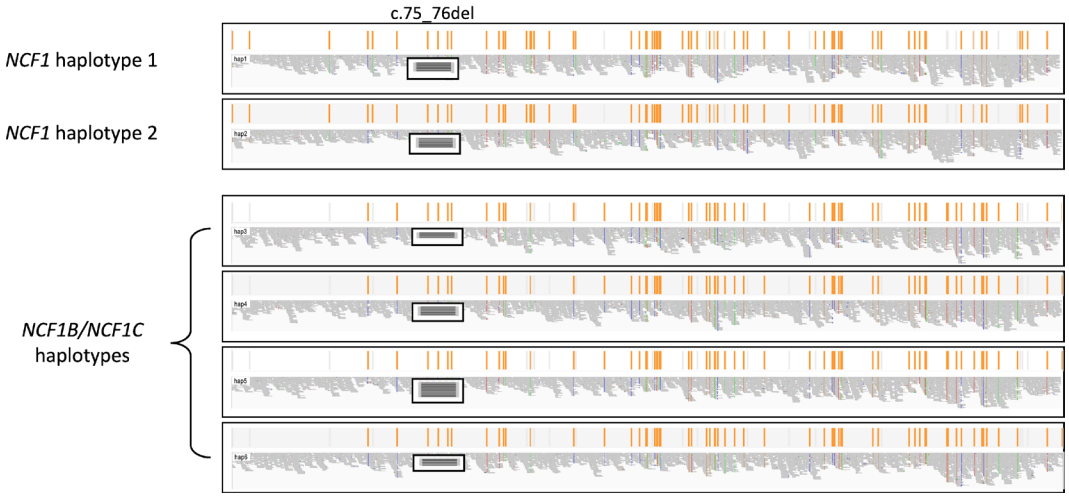


Figure 6: Detection of homozygous *NCF1* loss and putative gene conversion event.

Conclusion

TruPath Genome combined with MRJD deliver haplotype-resolved, copy-aware small variant detection in paralogous regions within a single WGS workflow. Across 31 samples over eight paralogous loci, this approach resolved known pathogenic variants in all cases, produced phased locus-wide haplotypes, and clarified allelic configuration, gene-pseudogene assignment, and CNV context. These results support TruPath Genome with MRJD as a unified, genome-wide assay that brings historically difficult-to-characterize genomic regions into routine genomic research, providing complementary genomic characterization within a single genome sequencing workflow, consolidating multiple research analyses into a single workflow and their associated complexity, cost, and time to result.

Learn more

TruPath Genome technical note

TruPath Genome flyer

TruPath Genome product guide

References

1. Mandelker, Diana, et al. "Navigating highly homologous genes in a molecular diagnostic setting: a resource for clinical next-generation sequencing." *Genetics in Medicine* 18.12 (2016): 1282-1289.
2. Ebbert, Mark TW, et al. "Systematic analysis of dark and camouflaged genes reveals disease-relevant genes hiding in plain sight." *Genome biology* 20.1 (2019): 97.
3. Wagner, Justin, et al. "Curated variation benchmarks for challenging medically relevant autosomal genes." *Nature biotechnology* 40.5 (2022): 672-680.

1.800.809.4566 toll-free (US) | +1.858.202.4566 tel
techsupport@illumina.com | www.illumina.com

M-GL-04321