Accurate variant calling in paralogous regions using SBX duplex reads

Jagadeeswaran Chandrasekar[1], Taher Mun[2], Mahdi Golkaram[2], Fong Chun Chan[2], Dilmi Perera[2], Alberto Gatto[2], Yanli Hou[2], Daniel Zinder[2], Grant Kingsley[1], McKenna Osentowski[1], Alec Sautter[1], Elizabeth Williams[1], Alan Kimura[1], Alexa Jung[1], Kendall Berg[1], Taylor Lehmann[1], Joanne Leadbetter[1], Majid Babazadeh[1], Melud Nabavi[1], Thomas Reid[1], Brittany Kesic[1], Brent Banasik[1], Marc Prindle[1], Cynthia Cech[1], Megan Freer[1], Anasha Arymann[1], Yui Umezawa[1], Chen Zhao[2], John Mannion[2], Mark Kokoris[1]

[1]Roche Sequencing Solutions, Inc. Seattle, WA, USA
[2]Roche Sequencing Solutions, Inc. Santa Clara, CA, USA

Copies of this poster obtained through QR (Quick Response), and/or text key codes are for personal use only and may not be reproduced without written permission of the authors.

Poster #P15.107.E

Introduction

Copy number variation in paralogous gene families has direct implications for rare disease (e.g., *SMN1*/*SMN2* in spinal muscular atrophy) and pharmacogenomics (e.g., *RHD*/*RHCE*, *CYP2D6*). Because paralogs share high sequence identity (often >95%), standard aligners mismap reads between them and conventional copy number callers cannot resolve which specific paralog is gained or lost. Sequencing By Expansion (SBX) is an ultra-high-throughput, rapid sequencing technology developed by Roche. We present a general-purpose framework for calling both aggregate and paralog-specific copy number from SBX Duplex (SBX-D) data, generalized to any segmental duplication region in the genome. This work is a stepping stone toward building a generalized star-allele caller for challenging genes.

Methods

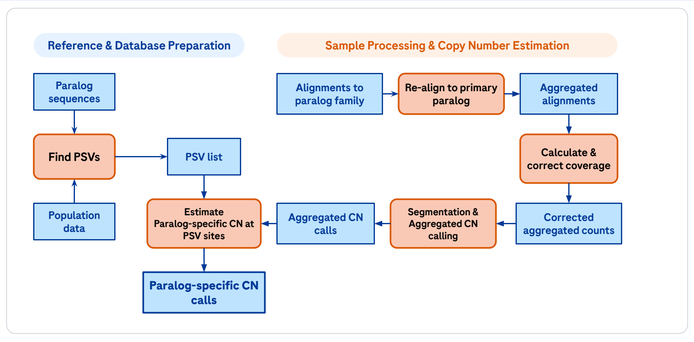

Figure 1. Workflow for generating paralog specific copy number calls

Pipeline Strategy: Reads overlapping any paralog in the family are extracted from the input BAM, excluding unmapped, secondary, QC-failed, duplicate, and supplementary reads (SAM flags 0x704). All recruited reads are then realigned to the reference sequence of the designated primary paralog using BWA-MEM [1], collapsing reads from all paralogs onto a single coordinate system for direct coverage comparison.

Read depth is calculated over the realigned BAM in 500 bp windows. A GC-bias correction model, fitted on a set of normalization regions distributed across the genome known to have little to no copy number variation, is applied to remove systematic coverage variation due to GC content. For each window, the GC-corrected coverage is then normalized against the median coverage of the normalization regions, transforming the signal into a scale where diploid regions center near zero, deletions shift negative, and duplications shift positive.

The normalized signal is segmented in two rounds. First, the signal is segmented using a hidden markov model (HMM) where the hidden states correspond to gaussian distributions centered around distinct copy number states. Aggregate copy number is determined from the resulting segments. Second, at paralogous sequence variant (PSV) positions - sites where paralogs carry fixed single-nucleotide differences (Figure 2) - per-paralog allele fractions are computed and segmented by a second HMM where the emissions correspond to PSV allele frequencies.

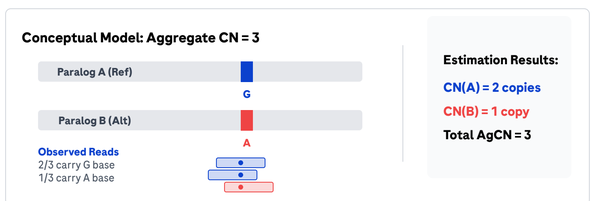

Figure 2: Visualization of how paralogous sequence variants (PSVs) can help determine paralog-specific copy number by assigning reads to one or the other paralog.

Methods, continued

Paralog-specific copy numbers are assigned using a multinomial likelihood model. For a given aggregate CN, all possible paralog-specific combinations are evaluated and the combination that best fits the observed per-paralog allele fractions is selected.

PSV discovery: Paralogous sequence variants are identified through an automated workflow that aligns paralog sequences, extracts fixed single-nucleotide differences, and filters candidates using gnomAD [2] population variation data to retain only the most confident distinguishing positions.

Output: Results are output as a standard bgzipped, indexed VCF with per-paralog copy number calls, segment coordinates, and quality metrics, enabling integration with downstream workflows.

Simulations: We validated paralog-specific copy number (PSCN) calling accuracy for two gene families with known variations caused by homology –RHD/RHCE and SMN1/SMN2–using synthetic diploid genomes. For each locus, we assembled haplotype sequences from GRCh38 representing the full spectrum of known structural variants: 36 RHD alleles across 7 categories (deletions, duplications, pseudogenes, and RHD-CE/CE-D hybrid genes) and 8 SMN alleles across 8 categories (deletions, duplications, gene conversions, and compound events). Breakpoints were placed at multiple positions within 3 categories of intervals (span 25%, 50%, 75% of surrounding introns), and haplotypes were paired to generate 206 RHD/RHCE and 37 SMN1/SMN2 diploid genotypes in heterozygous and homozygous configurations. SBX reads were simulated at 42× depth using pbsim3 [3] using an SBX-derived error profile and aligned to GRCh38 with BWA-MEM [2]. The caller was run on the resulting BAM files.

Real Data: The caller was used to detect the presence of RHD/RHCE and SMN1/SMN2 copy number variants from SBX-D sequencing runs for GIAB samples HG001-HG007 as well as the reference cell lines HG00154-HG00157, NA18548, NA19703, and NA12761. The calls for these samples were compared against orthogonal methods [4] [5]

Results

Simulated Data: The caller was evaluated on simulated hybrid genes and gene conversions per exon against expected copy numbers. Exons without paralog-specific variants (PSVs) were excluded. For SMN, the caller achieved 100% PSCN accuracy across 37 genotypes on PSV-containing exons (2–9) (Table 1). For RHD/RHCE, accuracy was 99.1% across 117 genotypes on exons 1–9 (1,044/1,053 exon-calls correct) (Table 2). 5 errors occurred at exon 6 in hybrid genotypes; a narrow adjacent intron (1,635 bp) meant some recombination breakpoint positions fell within the exon's 500 bp analysis interval, preventing resolution. The 4 remaining errors occurred at exon 1, coverage was undercounted to its proximity to a non-homologous region to RHCE.

Category	Genotypes	Exons	% Exons with Correct Aggregate CN	% Exons with Correct Paralog-Specific CN
SMN1 Deletion	6	48	100%	100%
SMN2 Deletion	6	48	100%	100%
SMN1 Duplication	6	48	100%	100%
SMN2 Duplication	6	48	100%	100%
SMN1+NAIP Deletion	3	24	100%	100%
SMN1 Del + SMN2 Dup	3	24	100%	100%
SMN1→SMN2 Conversion	3	24	100%	100%
SMN2→SMN1 Conversion	3	24	100%	100%

Table 1. SMN1/SMN2 simulation results. Simulations are broken down into variant category, number of genotypes per category, total number of exons across all genotypes, and the percent of correct aggregate and paralog-specific copy number calls across all exons

Category	Genotypes	Exons	% Exons with Correct Aggregate CN	% Exons with Correct Paralog-Specific CN
RHD Deletion	6	54	96.30%	100.00%
RHD-CE Hybrid	72	648	99.70%	99.10%
RHD-CE Hybrid (Partial)	12	108	100.00%	100.00%
RHCE-D Hybrid	18	162	100.00%	99.40%
RHD Partial Deletion	3	27	100.00%	100.00%
RHD Duplication	6	54	94.40%	96.30%

Table 2. RHD/RHCE simulation results. Simulations are broken down into variant category, number of genotypes per category, total number of exons across all genotypes, and the percent of correct aggregate and paralog-specific copy number calls across all exons.

Real Data: The paralog CN caller was able to correctly call a SMN2 deletion from SBX-D sequencing data for the samples HG006 (Figure 3) and HG00154, and neutral copy numbers for SMN1/SMN2 for the rest of the GIAB and reference cell line samples. The caller also successfully called exon 2 RHCE->RHD gene conversions for all GIAB samples, NA18548, and NA19703. HG002, in particular, contains has a homozygous conversion for exon 2 (Figure 3). The caller also successfully detected RHD deletions for HG001, NA18548 and NA19703, as well as a RHD->RHCE gene conversion spanning exon 3 to exon 9 for NA18548. and a homozygous RHD deletion for NA12761.

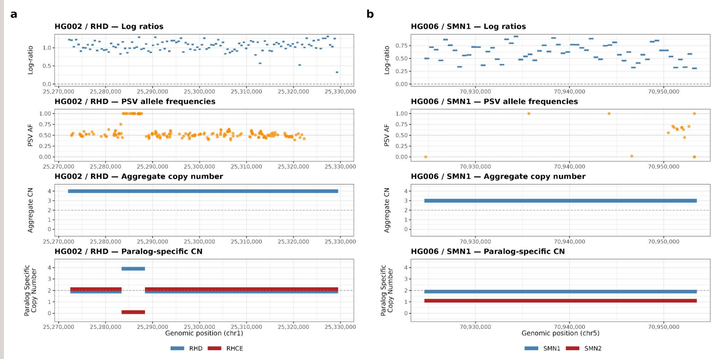

Figure 3. Log-ratios, PSV allele frequencies, aggregate copy number, and paralog-specific copy number showing a) a homozygous RHCE->RHD exon 2 conversion in HG002 and b) a heterozygous SMN2 deletion in HG006

Conclusion

We developed a general-purpose framework for paralog-specific copy number calling from SBX sequencing data, applicable to any segmental duplication region in the genome. The framework resolves both aggregate and per-paralog copy number by combining coverage-based segmentation with paralogous sequence variant allele frequencies. The approach was evaluated on the SMN (SMN1/SMN2) and RH (RHD/RHCE) paralog families and is potentially exentable to other gene families such as *CYP2D6*, *PMS2*, and *STRC*. Results are delivered in standard VCF format for seamless integration with downstream pipelines. This work is the first stepping stone toward a generalized star-allele caller tailored for SBX data. We aim to leverage paralog-specific copy number to call paralog-specific SNVs, enabling star-allele determination for challenging genes such as *CYP2D6* and *RHD*/*RHCE*.

References

[1] Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv. 2013. doi: 10.48550/arXiv.1303.3997
[2] Guez J, et al. Integrating 730,947 exome sequences with clinical literature improves gene discovery. medRxiv [Preprint]. 2026 Mar 25. doi: 10.64898/2026.03.23.26349081.
[3] Ono Y, et al. PBSIM3: a simulator for all types of PacBio and ONT long reads. *NAR Genomics and Bioinformatics*, 2022 Dec 4;4(4). doi: /10.1093/nargab/lqac092
[4] Chen X, et al. Spinal muscular atrophy diagnosis and carrier screening from genome sequencing data. Genet Med. 2020 May;22(5):945-953. doi: 10.1038/s41436-020-0754-0.
[5] Nuttle X, et, al. Rapid and accurate large-scale genotyping of duplicated genes and discovery of interlocus gene conversions. Nat Methods. 2013 Sep;10(9):903-9. doi: 10.1038/nmeth.2572.

The sequencing by expansion (SBX) technology is in development and not commercially available. The content of this material reflects current study results and/or design goals. For Research Use Only. Not for use in diagnostic procedures.