npj Genomic Medicine

https://doi.org/10.1038/s41525-026-00555-2

Article in Press

Integrative analysis of in silico predictions and clinical evidence to delineate the capability of HiFi long-read sequencing in paralogous genes

Received: 7 November 2025

Accepted: 17 February 2026

Cite this article as: Kim, S.K., Jang, J., Kim, Y. *et al.* Integrative analysis of in silico predictions and clinical evidence to delineate the capability of HiFi long-read sequencing in paralogous genes. *npj Genom. Med.* (2026). https://doi.org/10.1038/s41525-026-00555-2

Sung Kyung Kim, Joowon Jang, Yeseul Kim, Hobin Sung, Hyesu Lee, Hara Yim, Sung Im Cho, Jee-Soo Lee & Moon-Woo Seong

Integrative analysis of in silico predictions and clinical evidence to delineate the capability of HiFi long-read sequencing in paralogous genes

Sung Kyung Kim[1], Joowon Jang[1], Yeseul Kim[1], Hobin Sung[1], Hyesu Lee[1], Hara Yim[1], Sung Im Cho[1], Jee-Soo Lee[1], Moon-Woo Seong[1,2]

Affiliations

[1]Department of Laboratory Medicine, Seoul National University Hospital, Seoul National University College of Medicine, Seoul, Republic of Korea

[2]Seoul National University Cancer Research Institute, Seoul, Republic of Korea

Corresponding Authors

Moon-Woo Seong, MD, PhD

Department of Laboratory Medicine, Seoul National University Hospital, 101 Daehak-ro, Jongno-gu, Seoul 03080, Republic of Korea

Telephone number: +82-10-5441-8053

Email: mwseong@snu.ac.kr

ORCID iD: https://orcid.org/0000-0003-2954-3677

Jee-Soo Lee, MD, PhD

Department of Laboratory Medicine, Seoul National University Hospital, 101 Daehak-ro, Jongno-gu, Seoul 03080, Republic of Korea

Telephone number: +82-10-3772-9737

Email: leciel85@snu.ac.kr

ORCID iD: https://orcid.org/0000-0002-6633-4631

ORCIDs

Sung Kyung Kim: https://orcid.org/0009-0007-5637-5257

Yeseul Kim: https://orcid.org/0000-0002-6870-9214

Hobin Sung: https://orcid.org/0009-0006-8338-0716

Hyesu Lee: https://orcid.org/0000-0002-7373-6700

Hara Yim: https://orcid.org/0009-0003-5754-2759

Sung Im Cho: https://orcid.org/0000-0002-3819-8046

Jee-Soo Lee: https://orcid.org/0000-0002-6633-4631

Moon-Woo Seong: https://orcid.org/0000-0003-2954-3677

Abstract

Paralogous genes challenge short-read sequencing (SRS) due to high sequence similarity. Although long-read sequencing (LRS) improves resolution, the extent to which it resolves paralogous genes remains unclear. This study evaluates the capability of LRS by integrating in silico mappability-based predictions with clinical data to generate SRS- and LRS-unresolved gene lists, and by assessing whether a paralog-specific phasing (Paraphase) can overcome remaining limitations. Mappability was simulated across read lengths (250 bp to 14 kb) to predict unresolved regions and validated against mapping quality (MQ) from 66 high-fidelity LRS samples. Paraphase was applied to 79 paralog groups. Among 645 medically relevant (MR) genes unresolved by SRS, 419 (65.0%) were predicted to be resolved by LRS, while 226 (35.0%) remained unresolved. These predictions correlated with clinical MQ ($\chi^2 = 92.43$, $p < 2.2\times10^{-16}$; $\kappa = 0.37$), with significant differences between LRS-resolved and LRS-unresolved MR genes ($W = 63{,}656$, $p < 2.2\times10^{-16}$; $r = 0.36$). Paraphase resolved 61 groups (77.2%), providing additional resolution beyond LRS. LRS improves paralogous gene resolution but cannot fully eliminate paralog blind spots. Curated gene lists define boundaries of LRS utility for clinical interpretation, while Paraphase adds complementary resolution, supporting an integrated framework combining predictive modeling with algorithmic strategies.

Keywords

Short-read sequencing, Long-read sequencing, Mappability, Mapping quality, Paralogous genes

INTRODUCTION

Paralogous genes originate from gene-duplication events within the a genome, leading to sequence similarity over evolutionary time[1]. These duplications, ranging from small-scale tandem or segmental to whole-genome duplications, contribute to genomic complexity[2]. While some gene duplicates become inactive, others remain transcriptionally or translationally active in essential biological processes[3,4]. However, despite their biological importance, the presence of highly similar sequences poses substantial challenges for accurate analysis, particularly with next-generation sequencing (NGS), a class of high-throughput sequencing technologies that generate large volumes of DNA reads through massively parallel sequencing reactions.

Short-read sequencing (SRS), a widely adopted NGS approach for clinical genome and exome sequencing, is particularly vulnerable to challenges posed by paralogous genes. Its short read length limits its ability to distinguish between highly similar regions and hinders unique mapping of reads, resulting in insufficient or ambiguous alignment over clinically significant loci. This misalignment may lead to false positives or false negatives, complicating variant interpretation, and often requiring supplementary assays or expert manual review[5].

To evaluate the limitations of SRS in resolving paralogous genes, the concept of mappability was introduced as an alternative to mapping quality (MQ) metrics[6]. MQ is a read-level metric assigned by sequence aligners that reflects the confidence of a given read alignment. MQ values are inherently dependent on the read data, alignment algorithm, and sequencing error profiles. In contrast, mappability is commonly defined as the inverse of the number of genomic loci at which a sequence of a given length appears in the reference genome, thereby reflecting the likelihood that a read of that length can be uniquely mapped. Unlike MQ, which is influenced by multiple factors, mappability offers a sample and platform-independent measure of sequence uniqueness

based solely on a reference genome[7]. Because it can be precomputed under defined read lengths and error thresholds, mappability enables the in silico prediction of theoretically unmappable regions without experimental variability. Mandelker et al. applied a mappability-based metric to catalog genes that are intractable to SRS because of sequence similarity and identified 193 medically relevant (MR) paralogous genes as 'NGS problem genes' on GRCh37 reference[7].

To address the limitations of SRS in resolving complex genomic regions, long-read sequencing (LRS) technologies, such as Pacific Biosciences (PacBio) single-molecule real-time sequencing[8] and Oxford Nanopore sequencing[9], have been developed. LRS platforms produce reads in the kilobase (kb) range, thereby increasing the probability of spanning unique regions across paralogous loci and improving mapping accuracy. Further advances, including the optimization of circular consensus sequencing (CCS), have markedly enhanced LRS sequencing accuracy. PacBio high-fidelity (HiFi) reads have achieved over 99.8% base-calling accuracy, reaching more than 99.9% in recent meta-analyses[10,11].

With the advancement of LRS technologies, efforts have been made to examine their potential in resolving paralogous genes. Stephens and Iyer applied the concept of mappability to estimate the minimum read length required to uniquely map each base and generate a mappability spectrum across read lengths and error rates[12]. While offering the theoretical insight that mappability increases with read length, the study focused solely on this trend without pinpointing specific ambiguous regions, and relied only on SRS data without incorporating LRS to assess whether those regions are actually resolved in practice. Several studies have demonstrated the practical advantages of LRS in resolving paralogous genes, showing substantial gains in coverage and near-complete resolution of unmappable regions by SRS[4,10,13]. However, most previous studies

were limited to reference samples or specific phenotypes and primarily relied on MQ-based metrics.

More recently, Chen et al. highlighted that LRS still exhibits intrinsic limitations, particularly in long, highly homologous regions where shorter HiFi read lengths or hybrid-capture designs can result in low mapping confidence[14]. Consistent with this limitation, their analysis showed that 44.1% of paralog groups had a summary MQ ≤ 20 and 75.2% contained bases with low MQ, indicating that many paralogous loci remain unresolved even with LRS. To further address these challenges, building on their earlier work for the *SMN1*/*SMN2* locus[15], they applied a paralog-specific phasing tool, Paraphase, to 160 long segmental-duplication regions (> 10 kb, > 99% identity) and characterized paralog-specific haplotypes across populations. However, their work primarily focused on population-level genomic diversity rather than delineating the practical boundaries of paralog resolvability in clinical LRS data.

To address these limitations and extend prior findings, the present study systematically evaluated the resolution of paralogous genes across a range of read lengths (250 bp to 14 kb) using in silico mappability analysis with GRCh38 and validated these predictions against real-world LRS data from clinical samples. Consequently, we generated curated gene lists consisting of SRS and LRS-unresolved exons and genes, derived by integrating mappability predictions, to support clinical application. Furthermore, we explored the utility of Paraphase in resolving the ambiguities that persisted despite LRS. By integrating theoretical, clinical, and algorithmic approaches, this study aimed to identify current barriers to paralog resolution and guide future strategies and provide resources for comprehensive genomic interpretability.

RESULTS

In silico resolution of SRS-unresolved genes by LRS

Mappability-based analysis identified 56,576 exons (2.7% of all exons) as unresolved under SRS, corresponding to 3,745 genes. Of these, 645 were classified as SRS-unresolved MR genes (Table 1). Under the 7 kb LRS simulation, 38,788 exons (1.9%) were unresolved, and 265 genes were identified as 7 kb LRS-unresolved genes. Among the 645 SRS-unresolved MR genes, 387 genes (60.0%) were predicted to be fully mappable with 7 kb LRS and were therefore classified as 7 kb LRS-resolved MR genes. The remaining 258 genes (40.0%) remained unresolved even when the read length was increased to 7 kb. Additionally, seven genes were newly identified as unresolved using the 7 kb LRS parameter. When the read length was extended to 14 kb, which approximates the mean length of the PacBio HiFi reads, 419 of the 645 SRS-unresolved MR genes (65.0%) were identified as 14 kb LRS-resolved MR genes, and 226 genes (35.0%) remained unresolved. An additional 8 genes were newly discovered as unresolved under the 14 kb LRS simulation, resulting in a total of 234 genes classified as 14 kb LRS-unresolved MR genes. A list of SRS- and 14 kb LRS-unresolved exons and genes is provided in Supplementary Data 1 and 2.

To assess the effect of sequencing accuracy on gene resolution, we compared mappability results under varying error rates. At a read length of 7 kb, reducing the mismatch tolerance from 0.1% (Q30) to 0.01% (Q40) led to a marked decrease in the number of unresolved exons and genes, including a reduction in the number of MR genes from 265 to 248, yielding 17 fewer unresolved MR genes (Supplementary Table 1). In contrast, under the 14 kb read length condition, the same change in the error rate resulted in only a single exon and one gene difference, with no change in the number of unresolved MR genes (Supplementary Table 2). These findings suggested that once the read length reaches approximately 14 kb, a length already achieved by current PacBio

HiFi sequencing, further improvements in sequencing accuracy from 0.1% to 0.01% have a minimal effect on resolving additional paralogous genes.

We performed additional simulations using the T2T-CHM13 assembly under the same short-read parameters. When applying this framework to the T2T-CHM13 reference, the number of unresolved MR genes increased markedly from 645 genes to 2,968 genes compared to GRCh38 (Supplementary Table 3). This increase was driven by reduced mappability across regions newly represented or more fully modeled in the T2T-CHM13 assembly, including extended repetitive and centromeric regions.

Validation using long-read WGS data

To assess the clinical relevance of the in silico mappability-based predictions, we analyzed 66 PacBio HiFi whole-genome sequencing (WGS) samples. The average read length of the samples was 14.66 ± 2.54 kb. Among the 645 SRS-unresolved MR genes, 468 (72.6%) had an overall MQ of 60, indicating high-confidence mapping across all samples (Table 2). The remaining 177 genes (27.4%) showed an overall MQ below 60, with a median overall MQ of 8 (range 0-57.5), suggesting persistent mapping ambiguity despite LRS application. To explore variability in sample-level MQs, we plotted the distribution of sample-level MQs for these 177 genes using violin plots (Supplementary Fig. 1). Among these, 114 genes highlighted in red were in silico-predicted LRS-unresolved MR genes.

To quantitatively evaluate this relationship, we dichotomized the genes based on mappability-based prediction (LRS-resolved and -unresolved MR genes) and clinical mapping performance (overall MQ = 60 and < 60). Across simulated read lengths from 7 kb to 14 kb, significant associations were consistently observed between predicted and observed resolution (χ^2 = 87.32–

95.12, $p < 2.2 \times 10^{-16}$), with Cohen's kappa ranging from 0.36 to 0.38 (Fig. 1a). For the 14 kb model, which approximates the mean HiFi read length, the association remained significant ($\chi^2 = 92.43$, $p < 2.2 \times 10^{-16}$) with $\kappa = 0.37$ (95% CI: 0.30–0.45) (Table 2).

To further characterize the differences in mapping performance between the predicted groups, we compared the overall MQ as a continuous variable between 14 kb LRS-resolved and -unresolved MR genes. The overall MQ distributions in both groups deviated significantly from normality, as determined by the Shapiro–Wilk test ($W = 0.437$ [resolved], $W = 0.749$ [unresolved], $p < 2.2 \times 10^{-16}$). Given these non-normal distributions, the Wilcoxon rank-sum test was used to compare the groups. A significant difference in overall MQ was observed between resolved and unresolved MR genes ($W = 63{,}656$, $p < 2.2 \times 10^{-16}$), with a moderate effect size ($r = 0.36$). Density plots showed that the predicted 14 kb LRS-resolved MR genes (blue) were predominantly distributed in the high overall MQ region, indicating reliable mapping of clinical data (Fig. 1b). In contrast, the 14 kb LRS-unresolved MR genes (red) showed a bimodal distribution, with subsets of genes showing either high or low overall MQs. By applying *k*-means clustering to the overall MQs within the 14 kb LRS-unresolved MR genes, two distinct subgroups were identified: high overall MQ and low overall MQ clusters (Fig. 1c). The high overall MQ cluster (pink), containing 129 genes, was characterized by high overall MQs (median = 60, range 35–60), whereas low overall MQ clusters (dark red), containing 97 genes, showed markedly lower overall MQs (median = 6, range 0–32.2). This observation suggested that a subset of genes predicted to be unresolved based on in silico mappability may, in practice, be reliably resolved by LRS, indicating a potential underestimation of LRS performance under real-world conditions.

Benchmark-based validation of variant calling

We evaluated variant calling performance within the intersection of SRS-unresolved exons and the Genome in a Bottle (GIAB) HG002 GRCh38 high-confidence benchmark space. Within this evaluation, LRS variant calling using HiFi Revio markedly improved accuracy compared with SRS. For single nucleotide polymorphisms (SNPs), LRS achieved a recall of 0.9940 and precision of 0.9933 (F1=0.9937), whereas SRS showed lower recall (0.9423) despite high precision (0.9872) (F1=0.9642) (Supplementary Table 4). For insertions or deletions (INDELs), LRS also improved performance (recall 0.9881; precision 0.9867; F1=0.9874) relative to SRS (recall 0.9664; precision 0.9922; F1=0.9791).

To assess whether these differences tracked mappability-based predictions, we stratified the evaluation intervals into regions predicted to be resolved versus unresolved by 14 kb LRS (Supplementary Table 5). In the 14 kb LRS-resolved stratum, both platforms performed strongly; however, in the unresolved stratum, SRS exhibited a pronounced loss of recall (SNP recall 0.7579; INDEL recall 0.8108), whereas LRS remained substantially more robust (SNP recall 0.9699; INDEL recall 1.0000), consistent with improved variant detection in low-mappability exons by LRS. LRS performance itself was lower in the 14 kb LRS-unresolved stratum than in the 14 kb LRS-resolved stratum, driven primarily by reduced precision and F1 (SNP precision 0.9623 and F1 0.9661; INDEL precision 0.9268 and F1 0.9620), indicating that these regions remain challenging even for long reads. Depth profiling showed that mean coverage was comparable between datasets (SRS 56.76×; LRS 48.95×), while the fraction of low-depth bases (< 10) was markedly lower in LRS (0.25%) than in SRS (2.81%).

Limits of LRS with increasing read lengths

Despite substantial improvements in the mapping performance with LRS, a subset of genes continued to exhibit low overall MQ. To estimate the extent to which unresolved regions could be further resolved with longer reads, we modeled the relationship between read length and the number of unresolved MR genes using various nonlinear regression models based on mappability-based predictions across read lengths ranging from 250 bp to 14 kb.

Among the models tested, the power–law model showed the best fit (Akaike information criterion (AIC) = 32.83), followed by the exponential (AIC = 37.99), inverse (AIC = 46.62), and logarithmic (AIC = 55.93) models (Fig. 2). The estimated power–law function was $y = 6518.10 \times x^{-0.47} + 164.28$, where y represents the number of unresolved MR genes and x denotes the read length in base pairs. The model demonstrated a marked decline in unresolved MR genes as the read length increased from 250 bp to 7 kb, decreasing from 645 to 265 unresolved MR genes (corresponding to the resolution of 380 genes), followed by a gradual plateau beyond 7–8 kb. The curve asymptotically approached a value of approximately 164 unresolved MR genes, conditional on the GRCh38 reference genome and the mappability and error-rate assumptions used in this study, suggesting a saturation point beyond which further extension of the read length alone offered diminishing returns in resolving additional genes.

Paralog resolution using a paralog-specific phasing tool, Paraphase

To further address the subset of genes that remained unresolved despite the use of LRS, we used Paraphase[14]. From the 160 paralogous gene groups supported by Paraphase, a total of 79 paralogous gene groups were selected from the Paraphase reference, based on either inclusion in the in silico–predicted 14 kb LRS-unresolved MR genes or evidence of poor mapping performance (overall MQ < 60) in clinical WGS data. Among these, 61 gene groups (77.2%)

exhibited a mean phased region fraction (PRF) ≥ 0.95, indicating near-complete resolution of the gene with high-confidence haplotype reconstruction (Fig. 3a). The remaining 18 gene groups showed partial resolution, with mean PRFs ranging from 0.71 to 0.94, suggesting incomplete phasing in some regions. For each gene group, Paraphase reported a copy number corresponding to the total number of phased haplotypes, reflecting the total number of allelic copies across all paralogs within the group (e.g., six copies for a gene family with three paralogs). Representative cases were visualized using the Integrative Genomics Viewer (IGV) to further confirm paralog-specific haplotype separation. Paraphase successfully distinguished between the two *STRC* (HGNC: 16035) and two *STRCP1* (HGNC: 33915) haplotypes, with most reads uniquely assigned to each (Fig. 3b). In contrast, the *SMN1* (HGNC: 11117) group, although successfully phased into two *SMN1* and two *SMN2* (HGNC: 11118) haplotypes, showed a higher proportion of ambiguously aligned reads (grey) than *STRC*, consistent with its lower mean PRF (Fig. 3c). These visualizations support the interpretability of the mean PRF as a practical metric for evaluating paralog resolution.

DISCUSSION

The present study systematically evaluated LRS for resolving paralogous genes that are poorly resolved by SRS, using both in silico predictions and clinical long-read WGS data. Our results demonstrated a statistically significant association between mappability-based computational predictions and empirical outcomes from clinical LRS data, with moderate concordance between predicted and observed resolution, supporting their utility as predictive tools for identifying regions likely to remain unresolved. Although LRS substantially improved mappability across paralogous regions, it did not completely eliminate mapping ambiguities, particularly in regions

with lower MQ, suggesting that longer reads alone do not guarantee complete resolution. Furthermore, the application of a paralog-specific phasing tool enabled additional resolution in these challenging loci, achieving near-complete haplotype reconstruction.

LRS resolved 65.0% of the MR genes predicted to be unresolved by 250 bp reads in the in silico mappability analysis. Consistent with this finding, a comparison of predicted LRS-resolved and -unresolved MR genes with clinical MQ scores from long-read WGS samples revealed a significant association between dichotomous stratification and consistent trends in continuous MQ distributions. Prior studies highlighted the resolving power of LRS in paralogous regions. Wenger et al. (2019) sequenced HG002 with 13.5 kb HiFi reads and fully resolved 152 of 193 MR genes with 'NGS problem exons[10].' Other studies also reported substantial improvements in resolution with LRS. One study increased resolution from 12.1% with Illumina 250 bp reads to 49.5–90.4% depending on LRS platform[4], and another achieved up to 98% resolution of short-read "dead zones" in a subset of phenotype-associated genes[13]. However, these results largely derive from reference genomes (e.g., HG002, HG005) or targeted cohorts and rely on MQ-based proxies, limiting generalizability. In contrast, we incorporated diverse clinical WGS samples and directly related in silico mappability predictions to observed mapping performance, offering a more comprehensive evaluation of LRS utility.

More recent long-read studies have refined the evaluation of difficult genomic regions using variant-level benchmarks. Studies such as Höps et al.[16] and the Genome in a Bottle Consortium's Challenging Medically Relevant Genes efforts[17] demonstrated that HiFi LRS substantially improves the detection of clinically relevant small variants in duplicated and complex regions, extending high-confidence variant calls beyond conventional short-read benchmarks. In parallel, paralog-aware computational approaches have been developed to address mapping ambiguity at

the read and haplotype level. Genome-wide Paraphase enabled high-resolution haplotyping across many paralogous gene groups in multi-ancestry clinical cohorts[14], while alignment-based approaches such as DuploMap leveraged paralogous sequence variants to improve long-read mapping accuracy and variant recovery within segmental duplications[18]. In the present study, we focused on Paraphase because it directly reports haplotype-resolved structure and copy number at the gene-family level, allowing quantitative assessment of whether a paralogous gene group is resolved across samples.

Building on recent advances in LRS and paralog-aware analysis, our study addresses a gap left by prior work that has primarily evaluated paralog resolution at the level of individual variants, loci, or selected gene families. While these studies convincingly demonstrated that HiFi LRS improves mapping and variant detection in duplicated regions, they did not provide a genome-wide, quantitative view of how many MR paralogous genes transition from unresolved to resolved states as sequencing strategies change. By integrating in silico mappability-based predictions with empirical MQ derived from clinical HiFi WGS data, we systematically trace the resolution landscape of MR paralogous genes as sequencing moves from SRS to LRS. This approach enables us to explicitly delineate not only which paralogous genes benefit from increased read length, but also which genes remain refractory to resolution under current read-length and error-rate regimes—thereby defining practical and theoretical boundaries of LRS utility that are not captured by variant-level benchmarks alone.

While predicted LRS-resolved genes generally displayed high MQ in clinical datasets, a subset predicted as unresolved also achieved unexpectedly high MQ, indicating a prediction–observation mismatch. This discrepancy implies that mappability-based predictions may underestimate LRS's resolving power. Several factors may have contributed to these findings.

First, our mappability framework classifies a gene as unresolved if even a single exon has low mappability; however, in practice, long HiFi reads may span such exons and enable unique alignments using adjacent high-mappability regions, thereby achieving high MQ. Second, unlike the static sequence-based calculation used in mappability-based predictions, long-read aligners, such as minimap2, employ dynamic heuristic-driven algorithms that integrate multiple anchors along a read to construct coherent alignment chains based on sequence context and error profiles[19]. This chaining-based approach allows long-read mappers to tolerate local ambiguity and leverage broader sequence context, often rescuing alignments in regions of low theoretical mappability. In contrast, short-read mappers are more sensitive to local sequence similarity and multi-mapping, leading to different biases in paralogous regions[20]. Therefore, curated gene lists should be interpreted as conservative indicators rather than as definitive thresholds. They serve as safeguards to identify potentially problematic genes where additional validation or complementary approaches may be warranted, especially in clinical settings where diagnostic accuracy is critical.

To better understand the potential limitations of LRS in resolving paralogous genes, we examined how the number of unresolved genes changed as the simulated read lengths increased from 250 bp to 14 kb. Our analysis revealed a power-law relationship between read length and the number of unresolved genes, consistent with previous studies that employed different approaches to estimate mappability[12,21]. Notably, the reduction in the unresolved gene count plateaued beyond 7–8 kb, indicating diminishing returns with further increases in read length. These findings suggest that simply increasing the read length provides only limited additional benefits in resolving paralogous regions, especially considering the potential compromise between longer reads and sequencing accuracy. This saturation behavior is not merely a technical

limitation of current LRS platforms, but also reflects intrinsic biological properties of paralogous regions in the human genome. Many paralogous genes reside within long segmental duplications or evolutionarily recent duplication blocks characterized by extremely high sequence identity and limited divergence. In such regions, extending read length beyond a certain threshold does not necessarily introduce additional unique sequence context, particularly when duplicated segments themselves exceed typical read lengths or are accompanied by gene conversion events. As a result, even long and highly accurate reads may remain insufficient to unambiguously resolve paralogous loci. These genomic architecture-driven constraints help explain why unresolved gene counts plateau despite increasing read length, highlighting that paralog resolution is fundamentally bounded not only by sequencing technology, but also by the repetitive structure and evolutionary history of the genome.

Interestingly, we observed the emergence of new unresolved genes with longer simulated read lengths. This unexpected result was probably because of the increased absolute number of mismatches in longer reads under fixed error rates. This suggests that longer reads do not uniformly improve mappability across all loci and, in some cases, may introduce novel mapping ambiguities.

Sequencing accuracy is also a critical factor for resolution. Our study demonstrated that reducing the error rate from Q30 to Q40 substantially improved mappability at 7 kb read lengths, but had a negligible benefit at 14 kb. This observation is consistent with the mappability spectrum analysis by Stephens and Iyer, who showed that once reads exceed the size of dominant repetitive structures, improvements in sequencing accuracy alone do not substantially increase the proportion of uniquely mappable regions[12]. As current LRS technologies have already achieved read lengths of approximately 14 kb and error rates below 0.1%, further improvements in these

parameters alone are unlikely to yield marked gains in paralog resolution. These results highlighted the limitations of relying solely on longer or more accurate reads to resolve paralogous regions. Instead, future efforts should increasingly focus on complementary strategies, such as Paraphase, which can resolve sequence ambiguities, even in regions that remain challenging under optimal LRS conditions.

To evaluate the practical utility of such complementary strategies, we applied Paraphase, which successfully resolved 77.2% of the paralog gene groups with high-confidence haplotype reconstruction. This finding demonstrates the potential of paralog-aware tools in complementing LRS by resolving regions where conventional alignment remains ambiguous. Nonetheless, 22.8% of the gene groups exhibited only partial resolution, suggesting that even with advanced phasing strategies, certain paralogous loci remain technically challenging. Furthermore, our analysis was limited to paralogous genes currently supported by Paraphase, which focuses on paralog groups with 2–4 copies in GRCh38[14]. Consequently, several high-copy paralogous genes, despite being predicted to be unresolved or showing low MQ, were not included in the analysis, reflecting an additional limitation of the current tool. These results highlight the need for continued methodological development that can accommodate broader paralog complexity, particularly for clinically relevant genes, where ambiguity can affect diagnostic reliability.

The substantial increase in unresolved MR genes observed when using the T2T-CHM13 reference highlights a property of mappability-based analyses: unresolved gene counts do not solely reflect sequencing technology limitations, but are also strongly influenced by the structure and repetitiveness of the reference genome itself. T2T-CHM13 provides a more complete representation of highly repetitive genomic regions, including centromeres and long segmental duplications, which were partially absent or simplified in GRCh38. As a consequence, regions

that appeared uniquely mappable under GRCh38 may become non-unique when assessed against a more complete reference. This result should not be interpreted as evidence that T2T-CHM13 inherently worsens gene resolution in practice. Rather, it underscores that increased reference completeness can expose latent ambiguities that were previously masked by incomplete assemblies. Similar trends have been reported in prior work comparing GRCh37 and GRCh38, in which GRCh38 exhibited approximately twofold more non-uniquely mappable coordinates across all read lengths at a 1% error rate[12]. In this context, the increase in unresolved genes reflects a more realistic representation of genomic complexity rather than a regression in sequencing performance.

The present study has limitations. First, our in silico predictions were based on the GEM library tool, which assumes uniform error profiles and perfect alignment. Although this provides a conservative and reproducible framework, it cannot capture non-uniform errors, context-dependent alignment, or the presence of true biological variation such as SNVs, indels, and structural variants. Also, the mismatch thresholds used in the simulations represent simplified abstractions rather than full sequencing and alignment pipelines. Given these considerations, predictions can serve as a cautious screening layer for flagging regions that may be problematic under suboptimal conditions. Second, our clinical evaluation relied on MQ, which does not directly measure downstream variant calling accuracy. Nevertheless, MQ remains a practical metric for large-scale evaluation of paralogous regions, and future studies incorporating variant-level metrics are needed to fully validate functional resolution. Third, Paraphase was limited to paralogous gene groups currently supported by the tool; many other clinically relevant paralogs—especially with high copy numbers or complex structures—were not assessed.

Expanding Paraphase reference panels or developing additional tools will be necessary to fully evaluate these unresolved loci.

In conclusion, LRS markedly improves paralog resolution over SRS, yet a subset of genes remains challenging. Our findings show complementary roles for predictive mappability modeling and paralog-aware phasing in approaching comprehensive resolution. The tiered curated gene list offers a practical reference for clinical interpretation and method development targeting difficult loci. Future efforts should focus not only on increasing read length or sequencing accuracy but also on refining algorithms to enhance diagnostic applicability in clinical genomics.

METHODS

Reference genome preparation and mappability calculation

The GRCh38 reference genome file (hg38.fa.gz) was downloaded from the UCSC Genome Browser (https://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/) and processed by excluding alternate contigs and masking pseudoautosomal regions to avoid redundancy. Mappability analysis was performed using the GEM library tool (2013-04-06 modified version) (https://sourceforge.net/projects/gemlibrary/files/). A GEM index was generated using gem-indexer (build 1.423), and mappability scores were computed using gem-mappability (build 1.315).

To simulate SRS, mappability was computed using gem-mappability with options -l 250 -m 5, corresponding to 250 bp reads and up to five mismatches (i.e., 98% identity), a threshold where variant detection remains generally reliable. This read length was chosen to represent SRS practice, as 250 bp paired-end reads are widely used in clinical and research settings, including

Illumina NovaSeq platforms, and have been adopted in a prior mappability-based analysis of short-read limitations in paralogous regions[7]. The command used was: gem-mappability -I hg38_ALTexcluded_PARmasked.gem -l 250 -m 5 -o hg38_l250_m5. For the LRS simulations, based on a mean HiFi read length of 13.5 ± 1.2 kb[10], we initially set the read length to 7 kb (-l 7000), which is approximately half the mean read length. Simulations were subsequently extended to 8, 10, 12, and 14 kb to evaluate the effect of increased read lengths on mappability. The mismatch threshold (-m) was set to read length divided by 1,000 (e.g., -l 7000 -m 7), reflecting 99.9% accuracy of HiFi reads (e.g., gem-mappability -I hg38_ALTexcluded_PARmasked.gem -l 7000 -m 7 -o hg38_l7000_m7)[11].

Output .mappability files were converted into .wig using gem-2-wig and to .bed using BEDOPS (v2.4.41) wig2bed[22], with the coordinates adjusted to 1-based format for bedtools compatibility. To assess the impact of reference genome completeness on short-read mappability, we additionally performed mappability simulations using the T2T-CHM13 reference genome. The T2T-CHM13 assembly (hs1.fa.gz) was downloaded from the UCSC Genome Browser (https://hgdownload.soe.ucsc.edu/goldenPath/hs1/bigZips/). SRS mappability was computed using the same parameters as for GRCh38 (read length 250 bp, mismatch threshold corresponding to 98% identity) to ensure comparability between assemblies.

Exon annotation and unresolved gene identification

Exon annotations were retrieved from the ncbiRefSeq table (RefSeq All track) in the UCSC Table Browser. These annotations were then extended by ± 65 bp using an AWK command without restricting overlaps between adjacent exons, which ensured the potential splicing effects. The final coordinates were adjusted to 1-based format. These were intersected with the

mappability .bed files using bedtools intersect -wa -wb with bedtools (v2.31.1) (https://github.com/arq5x/bedtools2)[23]. Exons were treated as unstranded, as mappability is strand-independent. Exons were defined as unresolved if (1) ≥ 90% of bases had mappability scores < 1, a threshold adopted from the prior mappability study to capture exons that are predominantly non-unique[7], or (2) the maximum contiguous bases with mappability scores < 1 exceeded the simulated read length (7, 8, 10, 12, or 14 kb), indicating that no single read could uniquely anchor across the paralogous region. These metrics were computed using custom scripts in R (v4.4.3).

Genes were defined as unresolved if they contained at least one unresolved exon. To refine these to clinically relevant targets, we established a list of MR genes: 4,773 MR genes defined by Mandelker et al.[7]; 8,277 ClinVar genes with at least one allele reported as pathogenic or likely pathogenic (https://www.ncbi.nlm.nih.gov/clinvar/); 2,398 from ClinGen with moderate, strong, or definitive evidence (https://clinicalgenome.org/); and 4,760 from Genomics England PanelApp with high or moderate evidence (https://panelapp.genomicsengland.co.uk/). All gene sets were merged and deduplicated, resulting in 9,838 MR genes (Supplementary Data 3). These MR genes were intersected with the unresolved gene set to generate unresolved MR gene lists.

MQ analysis of clinical data

Whole-blood samples were collected from 69 Korean participants (40 males and 29 females, as recorded in the electronic medical records), between June 2024 and April 2025, for diagnostic or preimplantation genetic testing. High-molecular weight (HMW) genomic DNA was extracted from 300 μL of whole blood stored at −80°C using the Wizard® HMW DNA Extraction Kit (Promega). Library preparation was performed using the PacBio SMRTbell® Prep Kit 3.0. Short

DNA fragments were removed using the PacBio Short Read Eliminator (SRE) Kit, and 4 μg of DNA was sheared into 15–20 kb fragments using a Covaris g-TUBE. After cleanup with SMRTbell® beads, 46 μL eluate underwent DNA repair, A-tailing, and ligation with uniquely indexed SMRTbell adapters. Post-ligation products were treated with nucleases and size-selected with diluted AMPure PB beads to enrich fragments > 5 kb. The final library was prepared using the Revio® SPRQ™ Polymerase Kit. Sequencing was performed using the PacBio Revio® platform. HiFi reads were generated using CCS v8.2.0. Three of the 69 Hi-Fi WGS samples were excluded because of poor DNA quality or low sequencing throughput, resulting in 66 high-quality samples for analysis. HiFi reads were aligned to GRCh38 using pbmm2 (v1.16.0) (https://github.com/PacificBiosciences/pbmm2) with the following command: pbmm2 align hg38.fa.mmi Input.ccs.bam Output.bam --sort --preset HIFI --sample Sample_Name -J 20 --bam-index BAI.

For each SRS-unresolved MR gene (i.e., MR genes previously identified as unresolved by 250 bp-based mappability analysis), the MQ scores of all reads mapped to exons were extracted using bedtools intersect with -bed -wa -wb options. For each gene in each sample, the sample-level MQ was calculated as the median MQ of the exon-aligned reads. To obtain a summary measure across the cohort, the overall MQ for each gene was defined as the median of its sample-level MQs across all 66 samples, with an overall MQ of 60 considered high-confidence mapping.

All procedures involving human participants were approved by the Institutional Review Board (IRB) of Seoul National University Hospital (IRB No. 1102-102-357), and informed consent was obtained from all participants in accordance with the Declaration of Helsinki.

HG002 benchmark-based validation of variant calling in SRS-unresolved exons

To provide real-world validation of variant calling performance in regions predicted to be unresolved by SRS, we benchmarked variant calls against the GIAB (https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/) HG002 truth set on GRCh38 (NIST v4.2.1). We downloaded the HG002 benchmark VCF (HG002_GRCh38_1_22_v4.2.1_benchmark.vcf.gz) and the corresponding high-confidence bed (HG002_GRCh38_1_22_v4.2.1_benchmark_noinconsistent.bed) and restricted all analyses to autosomes (chr1–chr22). SRS-unresolved exons were merged and intersected with merged GIAB high-confidence intervals to define the evaluation regions using bedtools.

Using HG002 BAM files aligned to GRCh38, we evaluated a short-read Illumina 2×250 bp dataset (HG002.GRCh38.2x250.bam) and a long-read PacBio HiFi Revio dataset (HG002_PacBio-HiFi-Revio_20231031_48x_GRCh38-GIABv3.bam).

DeepVariant (v1.9.0) (https://github.com/google/deepvariant) was used to generate VCFs for each dataset, and benchmarking against the GIAB truth VCF was conducted using hap.py within the evaluation bed files. We reported recall, precision, and F1-score separately for SNPs and INDELs. To assess whether accuracy differences were associated with mappability-based predictions, we additionally stratified the evaluation bed files into regions expected to be resolved versus unresolved under 14kb LRS and repeated hap.py benchmarking within each stratum. Sequencing depth across the evaluation regions was summarized using mosdepth (v0.3.6) (https://github.com/brentp/mosdepth) to compute bp-weighted mean depth and the fraction of bases with depth < 10× for SRS and LRS.

Statistical analysis of mappability-based predictions and clinical data

Statistical analysis was performed in R (v4.4.3). Genes were dichotomized based on predicted LRS resolvability and the observed overall MQ and compared using chi-square tests (stats). Agreement was measured using Cohen's kappa. The overall MQ was also evaluated as a continuous variable. The Shapiro–Wilk test (rstatix) assessed normality, and the Wilcoxon rank-sum test (stats) was used to compare distributions between LRS-resolved and -unresolved MR genes. *K*-means clustering (stats) was used to evaluate within-group heterogeneity.

To model the relationship between read length and the number of unresolved MR genes, nonlinear regressions (logarithmic, exponential, inverse, and power-law) were fitted using nls function (stats) and compared using AIC scores.

Paralog resolution using Paraphase

Paraphase involves paralog-specific haplotype reconstruction by extracting reads from paralogous loci and realigning them to an archetype gene. It then phases the realigned reads into haplotypes to distinguish between paralogous genes[14]. HiFi reads from 66 samples were aligned to GRCh38 with alternate contigs excluded (https://github.com/PacificBiosciences/reference_genomes), as recommended by Paraphase, using pbmm2. We applied Paraphase (v3.3.1) (https://github.com/PacificBiosciences/paraphase) to 160 paralogous gene groups and quantified the resolution using a PRF, which is defined as the proportion of a gene resolved on a haplotype. Specifically, the PRF was calculated as the ratio of the total length of the haplotype-resolved region, defined by the start and end boundary coordinates reported by Paraphase for each haplotype, to the gene length. For haplotypes marked as 5′ or 3′ truncated, the gene length was adjusted by replacing the truncated side with the nearest observed clip position. Haplotype-level PRFs were extracted for each haplotype within a sample

and averaged across haplotypes to obtain a sample-level PRF for each gene group. The mean PRF for each gene group was then calculated by averaging sample-level PRFs across the cohort. Copy number estimates reported by Paraphase were analyzed separately to characterize the total number of allelic copies within each paralogous gene group.

DECLARATION

Data Availability

The GRCh38 reference genome FASTA file used for mappability analysis is available from the UCSC Genome Browser (https://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/). The GRCh38 version with excluded alternate contigs was downloaded from the Pacific Biosciences GitHub repository (https://github.com/PacificBiosciences/reference_genomes). The GEM library tool was downloaded from SourceForge (https://sourceforge.net/projects/gemlibrary/files/). The code for Paraphase is available on GitHub (https://github.com/PacificBiosciences/paraphase). Owing to ethical constraints and the sensitive nature of clinical genomic data, full BAM files from patients cannot be made publicly available. The data are available following review and approval by the corresponding author's institution and under appropriate data use agreements upon request.

Acknowledgements

The present study received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author Contributions

Conceptualization: S.K., J.L, M.S.; Data curation: H.S., H.L., H.I., S.C.; Formal analysis: S.K., J.L.; Methodology: S.K., J.J., Y.K., J.L., M.S.; Supervision: J.L., M.S.; Visualization: S.K.; Writing-original draft: S.K.; Writing-review & editing: all authors. All authors reviewed and approved the final manuscript.

Competing Interest

The authors declare no competing financial or non-financial interests.

References

1 Koonin, E. V. Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet* **39**, 309-338 (2005). https://doi.org/10.1146/annurev.genet.39.073003.114725

2 Kuzmin, E., Taylor, J. S. & Boone, C. Retention of duplicated genes in evolution. *Trends Genet* **38**, 59-72 (2022). https://doi.org/10.1016/j.tig.2021.06.016

3 Drobek, M. Paralogous genes involved in embryonic development: lessons from the eye and other tissues. *Genes (Basel)* **13** (2022). https://doi.org/10.3390/genes13112082

4 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). https://doi.org/10.1186/s13059-019-1707-2

5 Olivucci, G. *et al.* Long read sequencing on its way to the routine diagnostics of genetic diseases. *Front Genet* **15**, 1374860 (2024). https://doi.org/10.3389/fgene.2024.1374860

6 Derrien, T. *et al.* Fast computation and applications of genome mappability. *PLoS One* **7**, e30377 (2012). https://doi.org/10.1371/journal.pone.0030377

7 Mandelker, D. *et al.* Navigating highly homologous genes in a molecular diagnostic setting: a resource for clinical next-generation sequencing. *Genet Med* **18**, 1282-1289 (2016). https://doi.org/10.1038/gim.2016.58

8 Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133-138 (2009). https://doi.org/10.1126/science.1162986

9 Mikheyev, A. S. & Tin, M. M. A first look at the Oxford Nanopore MinION sequencer. *Mol Ecol Resour* **14**, 1097-1102 (2014). https://doi.org/10.1111/1755-0998.12324

10 Wenger, A. M. *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol* **37**, 1155-1162 (2019).

https://doi.org/10.1038/s41587-019-0217-9

11 Logsdon, G. A., Vollger, M. R. & Eichler, E. E. Long-read human genome sequencing and its applications. *Nat Rev Genet* **21**, 597-614 (2020). https://doi.org/10.1038/s41576-020-0236-x

12 Stephens, Z. D. & Iyer, R. K. Measuring the Mappability Spectrum of Reference Genome Assemblies. Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, 47-52 (2018). https://doi.org/10.1145/3233547.3233582

13 Sanford Kobayashi, E. *et al.* Approaches to long-read sequencing in a clinical setting to improve diagnostic rate. *Sci Rep* **12**, 16945 (2022). https://doi.org/10.1038/s41598-022-20113-x

14 Chen, X. *et al.* Genome-wide profiling of highly similar paralogous genes using HiFi sequencing. *Nat Commun* **16**, 2340 (2025). https://doi.org/10.1038/s41467-025-57505-2

15 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). https://doi.org/10.1016/j.ajhg.2023.01.001

16 Hops, W. *et al.* HiFi long-read genomes for difficult-to-detect, clinically relevant variants. *Am J Hum Genet* **112**, 450-456 (2025). https://doi.org/10.1016/j.ajhg.2024.12.013

17 Wagner, J. *et al.* Benchmarking challenging small variants with linked and long reads. *Cell Genom* **2** (2022). https://doi.org/10.1016/j.xgen.2022.100128

18 Prodanov, T. & Bansal, V. Sensitive alignment using paralogous sequence variants improves long-read mapping and variant calling in segmental duplications. *Nucleic Acids Res* **48**, e114 (2020). https://doi.org/10.1093/nar/gkaa829

19 Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100 (2018). https://doi.org/10.1093/bioinformatics/bty191

20 Sahlin, K., Baudeau, T., Cazaux, B. & Marchet, C. A survey of mapping algorithms in the long-reads era. *Genome Biol* **24**, 133 (2023). https://doi.org/10.1186/s13059-023-02972-3

21 Li, W. & Freudenberg, J. Mappability and read length. *Front Genet* **5**, 381 (2014). https://doi.org/10.3389/fgene.2014.00381

22 Neph, S. *et al.* BEDOPS: high-performance genomic feature operations. *Bioinformatics* **28**, 1919-1920 (2012). https://doi.org/10.1093/bioinformatics/bts277

23 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010). https://doi.org/10.1093/bioinformatics/btq033

Tables

Table 1 Unresolved exons and genes by mappability-based prediction.

	SRS	LRS				
Parameter	250 bp	7 kb	8 kb	10 kb	12 kb	14 kb
Length (-l)	250	7000	8000	10000	12000	14000
Mismatch (-m)	5	7	8	10	12	14
Unresolved exons (% of all exons)	56576 (2.7%)	38788 (1.9%)	38585 (1.8%)	36771 (1.8%)	34401 (1.6%)	33167 (1.6%)
Unresolved genes	3745	2048	2012	1945	1875	1819
Unresolved MR genes	645	265	255	250	244	234
Resolved MR genes	-	387	396	402	409	419
Newly identified unresolved MR genes	-	7	6	7	8	8

This table summarizes the number of unresolved exons and genes identified using mappability-based prediction across simulated read lengths ranging from 250 bp to 14 kb. For each parameter, it shows the total number of unresolved exons, genes, and MR genes, as well as the number of MR genes resolved with longer reads and newly identified unresolved genes that emerged under LRS compared with 250 bp SRS.

Table 2 Association between mappability-based prediction of 14 kb LRS resolution and overall MQ of clinical data.

	Overall MQ < 60	Overall MQ = 60	Total
14 kb LRS-unresolved MR genes[a]	114	112	226
14 kb LRS-resolved MR genes	63	356	419
SRS-unresolved MR genes	177	468	645

This table shows the association between the in silico prediction of the 14 kb LRS resolution and the overall MQ from LRS clinical data. The SRS-unresolved MR genes were grouped into 14 kb LRS-unresolved or -resolved MR genes based on mappability analysis and dichotomized by overall MQ (< 60 vs. 60).

[a]Only SRS-unresolved MR genes were included. The eight genes that were newly identified as 14 kb LRS-unresolved MR genes were excluded.

Figure and figure legends

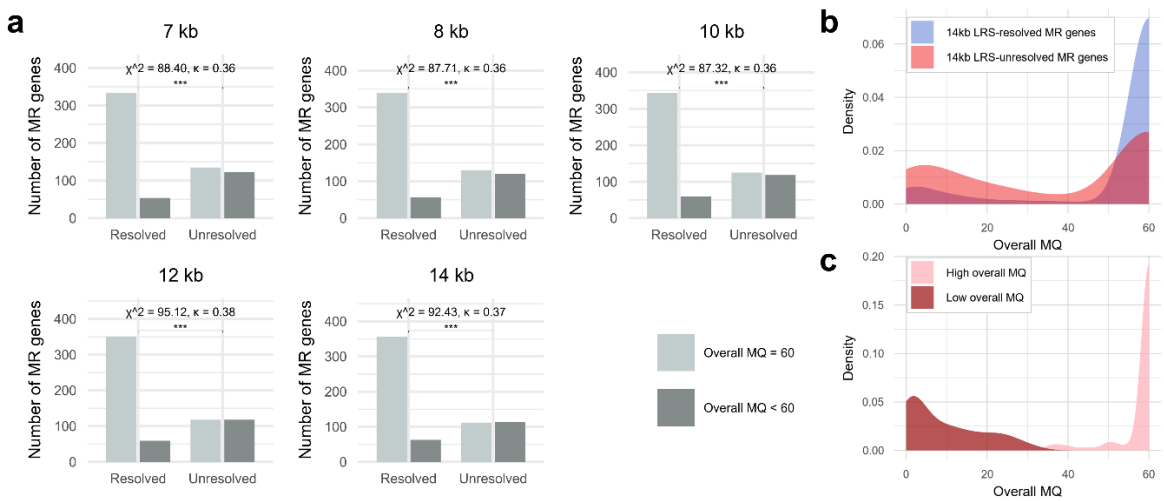

Fig. 1 Assessing MQ of resolved and unresolved MR genes in PacBio WGS.

a Bar plots show the number of MR genes predicted as resolved or unresolved under each simulated read length (7 kb–14 kb). Genes were stratified by their overall MQ in clinical LRS data (overall MQ = 60, grey; overall MQ < 60, dark grey). **b** Distribution of the overall MQ of 14 kb LRS-resolved and -unresolved MR genes**.** Density plots showing the distribution of the overall MQ for MR genes predicted to be resolved or unresolved by 14 kb LRS. 14 kb LRS-resolved MR genes were enriched in the high-MQ region, while 14 kb LRS-unresolved MR genes showed a bimodal pattern. **c** *K*-means clustering of 14 kb LRS-unresolved MR genes based on the overall MQ**.** Density plots show the two clusters identified within the 14 kb LRS-unresolved MR genes using *k*-means clustering. The high overall MQ cluster represents genes with consistently high mapping quality across samples, whereas the low overall MQ cluster includes genes with markedly lower MQ.

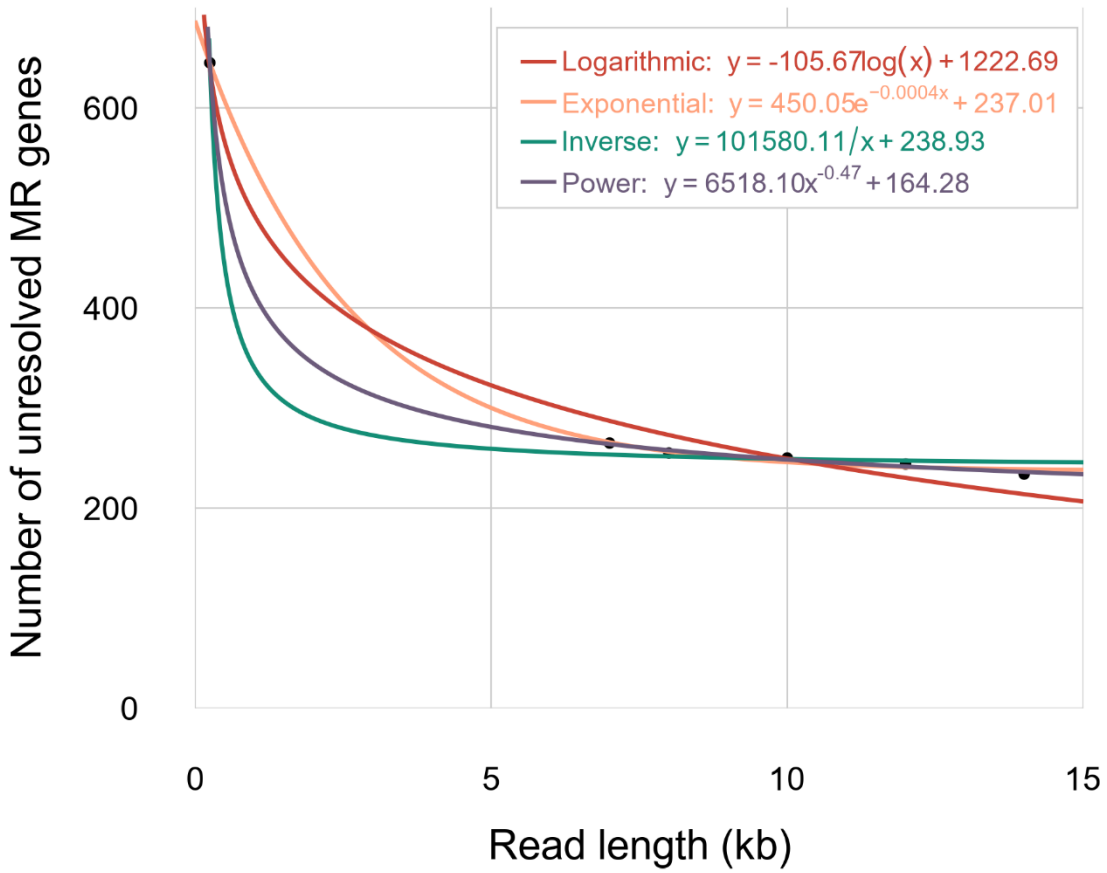

Fig. 2 Modeling of the relationship between read length and the number of unresolved MR genes.

The plot shows the predicted number of unresolved MR genes across read lengths based on nonlinear regression modeling. Among the models tested, the power-law model provided the best fit, indicating the rapid resolution of MR genes with increasing read length up to 7 kb, followed by a plateau beyond 7–8 kb.

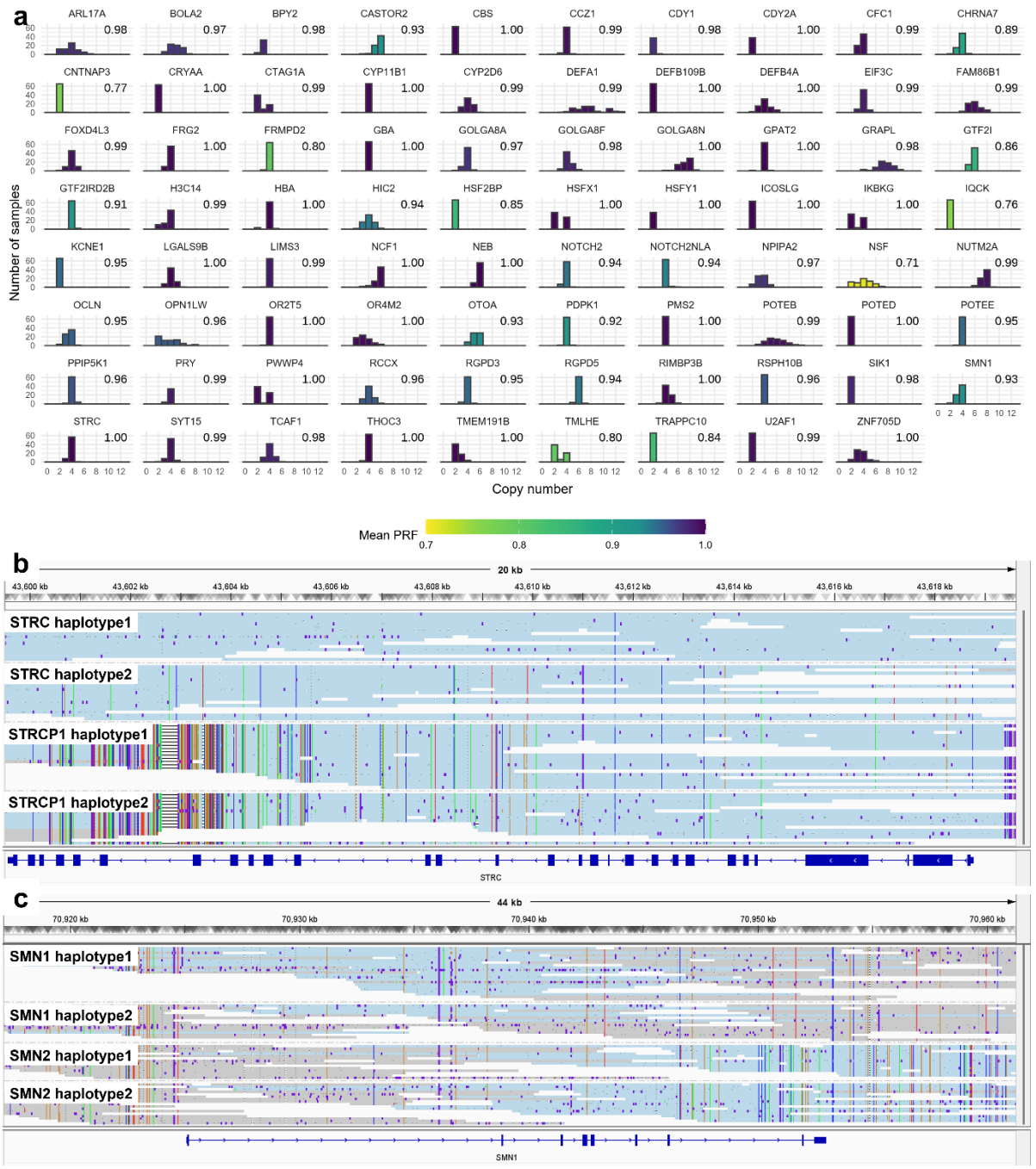

Fig. 3 Paralog resolution analysis using Paraphase.

a Mean PRF values and copy numbers of 79 paralogous gene groups. Each histogram displays the copy number detected in Paraphase for the corresponding gene group. The histogram bars are colored based on the mean PRF value for each group, which is shown as a numerical value in the

upper-right corner of each histogram. **b** IGV visualization of the *STRC* gene group. Paraphase successfully separated *STRC* and *STRCP1* haplotypes, with most reads clearly assigned to each gene. **c** IGV visualization of the *SMN1* gene group. While *SMN1* and *SMN2* haplotypes were distinguished, a greater proportion of ambiguously mapped reads was observed compared to *STRC*, reflecting lower PRF values.