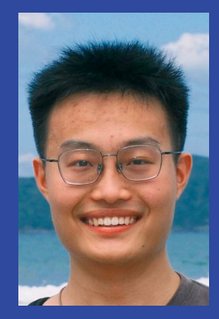
Aggregating Epigenetic Methylation Signals for Improved Long-Read Genetic Variant Calling

Li Junzhe, Zheng Zhenxian, Luo Ruibang

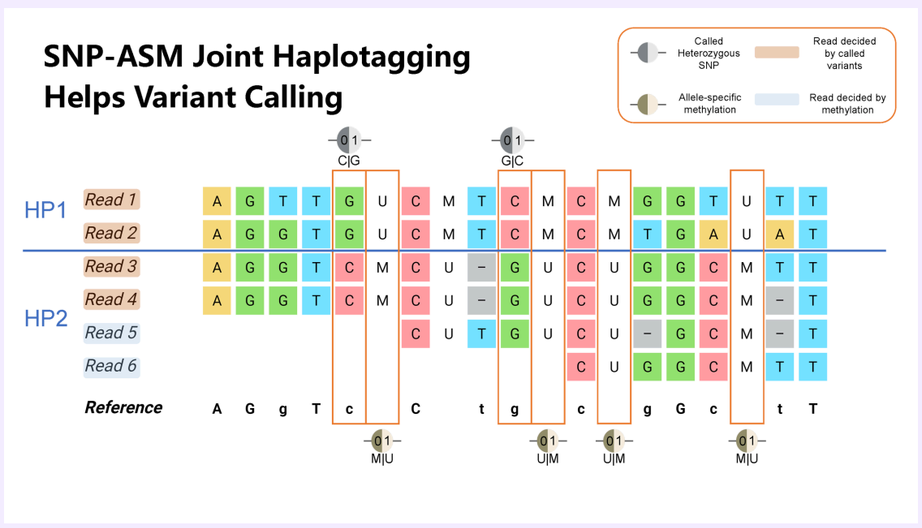
School of Computing and Data Science, The University of Hong Kong, Hong Kong, China



ABSTRACT

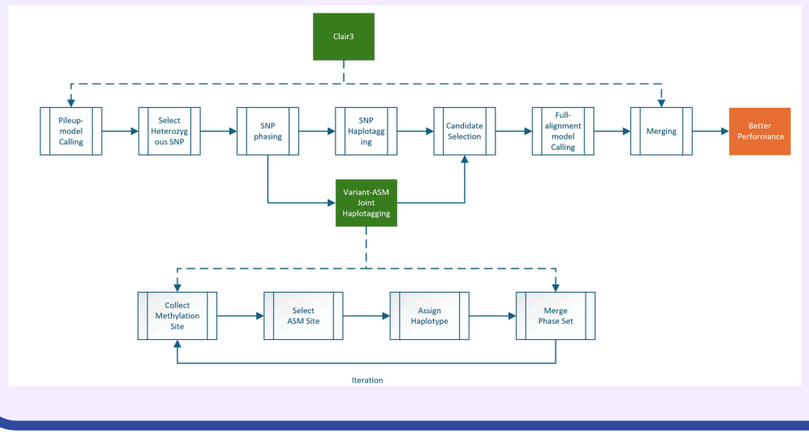
Accurate variant calling is fundamental to genomic analysis. Contemporary haplotype-aware variant callers predominantly rely on heterozygous Single Nucleotide Polymorphisms (SNPs) for read phasing. However, the scarcity of SNPs in homozygous or repetitive genomic regions compromises phasing continuity, resulting in fragmented haplotypes and suboptimal variant calling accuracy. To address this limitation, we introduce a novel Clair3-Methylation workflow that leverages Allele-Specific Methylation (ASM) signals as an orthogonal data source. The proposed algorithm identifies high-confidence ASM sites using Fisher's Exact Test to assign haplotypes to previously unphased reads and concatenates adjacent phase sets based on consistent methylation signatures. Benchmarking on Genome in a Bottle (GIAB) datasets (10x-50x) shows our method increases haplotagged read percentage by ~10% and demonstrates substantial improvements in haplotype contiguity. Notably, the Phase Set N50 on chromosome 20 increased from ~0.9 Mb to ~33 Mb at 50x coverage. These results confirm that integrating epigenetic markers effectively bridges genomic gaps left by SNP-only methods, leading to superior recall and F1-scores in complex genomic contexts.

How ASM Helps Identifying Variants



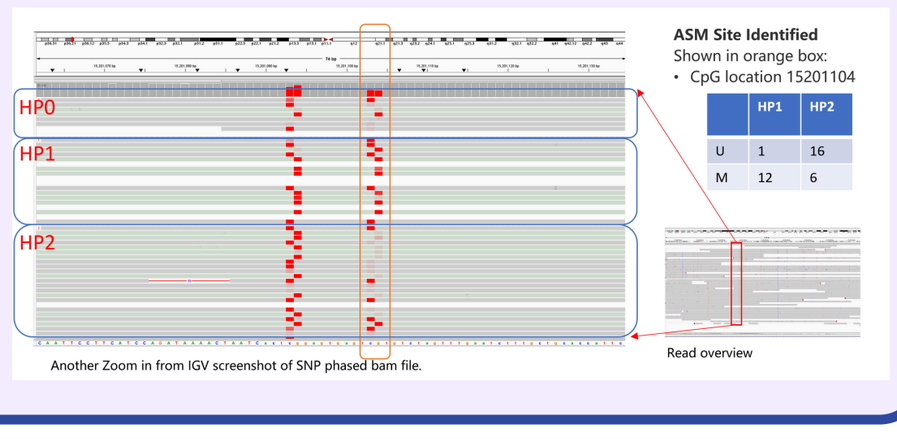
METHODOLOGY

Clair-Methylation Framework



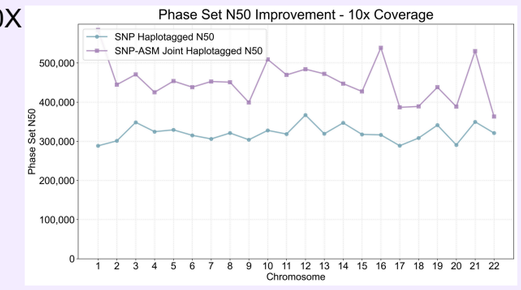
- Data Input: ONT BAM files with MM/ML methylation tags.
- ASM Identification: Scans CpG sites; uses Fisher's Exact Test to find significant methylation-haplotype associations.
- Haplotype Assignment: Unphased reads covering ASM sites are assigned haplotypes based on methylation status on ASM sites.
- Phase Set concatenation: Adjacent phase sets are concatenated if they share consistent ASM signatures, bridging genomic gaps.

Fisher's Exact Test: Construct contingency table for ASM site selection

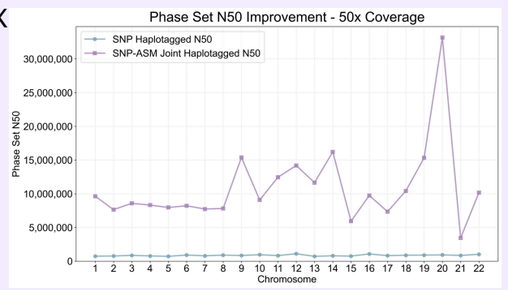


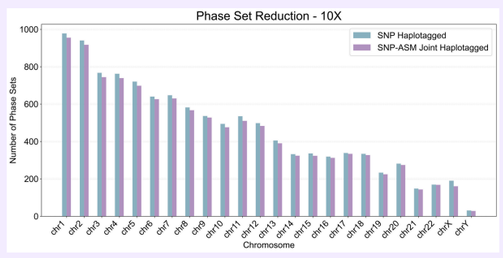
RESULTS - PHASING PERFORMANCE

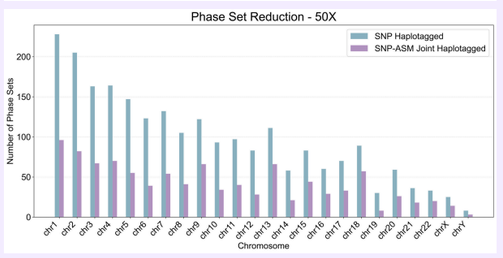
10X

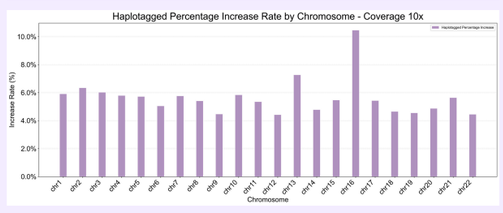


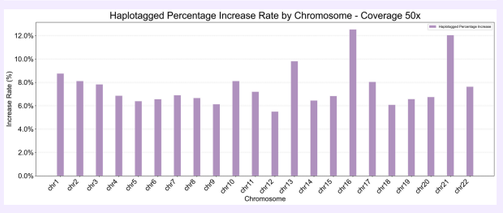
50X









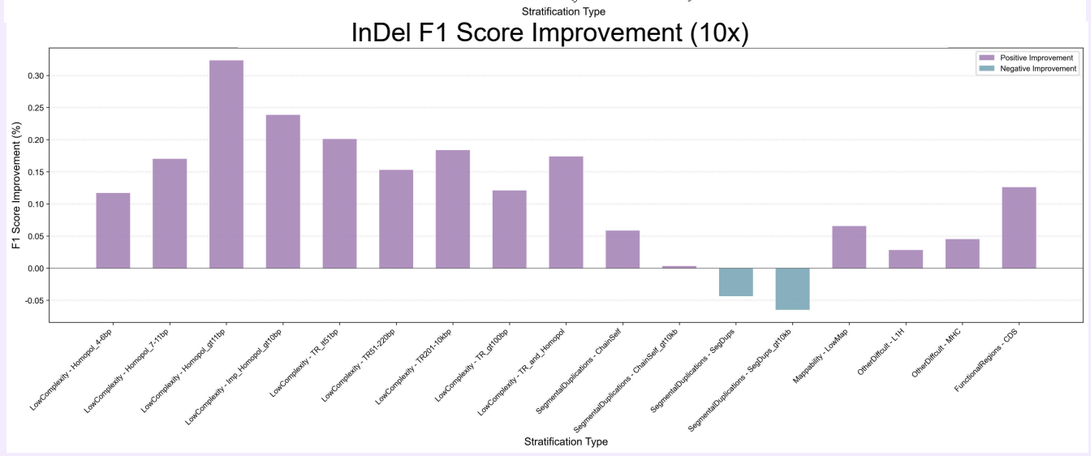


- Fragmentation Reduced: The proposed method successfully merges fragmented blocks, evidenced by the widespread reduction in phase set counts (Middle Row) and a consistent 5-12.5% increase in haplotagged reads percentage across chromosomes (Bottom Row).
- Continuity Surge: This merging leads to massive continuity gains. Notably, the Phase Set N50 (Top Row) on chromosome 20, 50x surged from ~0.9 Mb (baseline) to ~33 Mb (ours).
- Deeper sequencing amplifies the power of ASM.
- Conclusion: ASM signals effectively bridge gaps between fragmented phase blocks.

RESULTS - VARIANT CALLING PERFORMANCE

- Accuracy: Reduced FN by ~10,000 variants (at 30x) compared to origin Clair3.
- Difficult Regions: Achieved positive F1-score gains in Low Complexity and Tandem Repeat (TR) regions.
- Significance: Improved phasing results in better Recall without sacrificing precision.





Platform	Dataset	Caller	SNP						Total					
			Precision	Recall	F1-score	TP	FP	FN	Precision	Recall	F1-score	TP	FP	FN
ONT 5kHz	HG003 20x	Clair3	0.994576	0.991975	0.993274	3,300,793	18,010	26,702	0.985264	0.962068	0.973528	3,686,641	55,319	145,355
		Clair3 with implemented methylation	0.994131	0.99405	0.99409	3,307,695	19,535	19,800	0.984064	0.964438	0.974152	3,695,723	60,043	136,273
ONT 5kHz	HG003 30x	Clair3	0.997397	0.994162	0.995777	3,308,068	8,638	19,427	0.990325	0.969156	0.979627	3,713,803	36,412	118,193
		Clair3 with implemented methylation	0.997419	0.996669	0.997044	3,316,411	8,584	11,084	0.989737	0.971772	0.980672	3,723,825	38,754	108,171
ONT 5kHz	HG003 40x	Clair3	0.998218	0.994715	0.996464	3,309,910	5,911	17,585	0.992041	0.97272	0.982286	3,727,459	30,019	104,537
		Clair3 with implemented methylation	0.998102	0.997213	0.997657	3,318,221	6,314	17,585	0.991536	0.975356	0.983379	3,737,561	32,029	94,435
ONT 5kHz	HG003 50x	Clair3	0.998509	0.995477	0.996991	3,312,444	4,947	15,051	0.992768	0.975674	0.984147	3,738,780	27,346	93,216
		Clair3 with implemented methylation	0.998563	0.997469	0.998016	3,319,072	4,778	8,423	0.992496	0.977809	0.985098	3,746,960	28,444	85,036

CONCLUSION

- **Multi-omic Proof-of-Concept:** We validated a multi-omic approach, demonstrating that epigenetic signals (methylation) can successfully substitute and complement for missing genetic markers (SNPs) in variant calling.
- **Breakthrough:** Established a rigorous pipeline that leverages ASM sites as orthogonal markers to improve phasing and variant calling in difficult genomic regions.
- **Future Work:** 1. Refine ASM selection logic for complex regions (e.g., Segmental Duplications). 2. Move towards a fully end-to-end epigenome-aware neural network. 3. Utilize methylation signals to identify and correct alignment artifacts (misalignments).

HKU CDS BIOAI LAB

REFERENCES

- J. Rosenskiet al., "Atlas of imprinted and allele-specific DNA methylation in the human body," Nature Communications, vol. 16, no. 1, Mar. 2025, doi: https://doi.org/10.1038/s41467-025-57433-1.
- Y. Fu et al., "MethPhaser:methylation-based long-read haplotype phasing of human genomes," Nature Communications, vol. 15, no. 1, Jun. 2024, doi: https://doi.org/10.1038/s41467-024-49588-0.
- Z. Zheng,S. Li, J. Su, A. Leung, T.-W. Lam, and R. Luo, "Symphonizing pileup and full-alignment for deep learning-based long-read variant calling," vol. 2, no. 12, pp. 797–803, Dec. 2022, doi: https://doi.org/10.1038/s43588-022-00387-x.