RICE UNIVERSITY
School of Engineering
Department of Computer Science

Accurate and Efficient Computational Approaches for Long-read Alignment and Genome Phasing of Human Genomes

Yilei Fu

Ph.D. Thesis Defense

Department of Computer Science
Rice University

Dissertation Committee:

Dr. Todd J Treangen (Chair)
Dr. Gang Bao
Dr. Vicky Yao
Dr. Fritz J Sedlazeck
Dr. Huw Ogilvie

November 27, 2023

The Human Genome

Year	2000	2009	2013	2021	2022
Ref Genome Name	hg1	hg19/GRCh37	GRCh38	T2T-CHM13	T2T-CHM13+Y
Completeness	Fragmented, 90%	>93%	>95%	Missing Chr Y	100%
Project	**Human Genome Project**	**Genome Reference Consortium**		**T2T Consortium**	

Enhancing Genetic Knowledge

Gene functions, regulations and expressions

Understanding Genetic Diseases

Identifying the genetic contribution to health and diseases

*"**Genetics** is the scientific study of inherited variation. **Human genetics**, then, is the scientific study of inherited human variation."*

-- NIH, National Institutes of Health

Genetic Variants

…CGTCTGGGGGGTATG**C**ACGCGATAGCATTGCGAGACGC
TGGAGCCGGAGCACCCTATGTCGCAGTATC…

…CGTCTGGGGGGTATG**G**ACGCGATAGCATTGCGAGACG
CTGGAGCCGGAGCACCCTATGTCGCAGTATC…

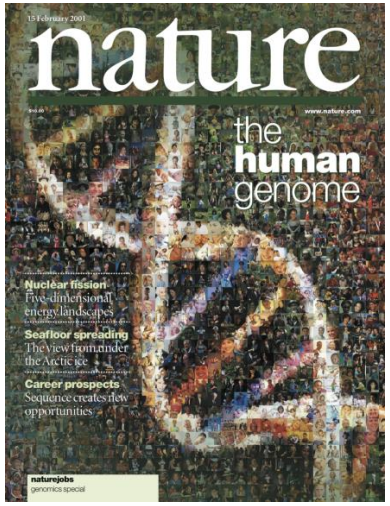
Humans are 99.9% genetically identical!
0.1% of genetic variant leads to:

- Diversities
- Diseases

NIH Curriculum Supplement Series. Understanding Human Genetic Variation https://www.ncbi.nlm.nih.gov/books/NBK20363/

Sequencing Platforms

Year	2000	2009	2013	2021	2022
Ref Genome Name	hg1	hg19/GRCh37	GRCh38	T2T-CHM13	T2T-CHM13+Y
Completeness	Fragmented, 90%	>93%	>95%	Missing Chr Y	100%
Project	**Human Genome Project**	**Genome Reference Consortium**		**T2T Consortium**	



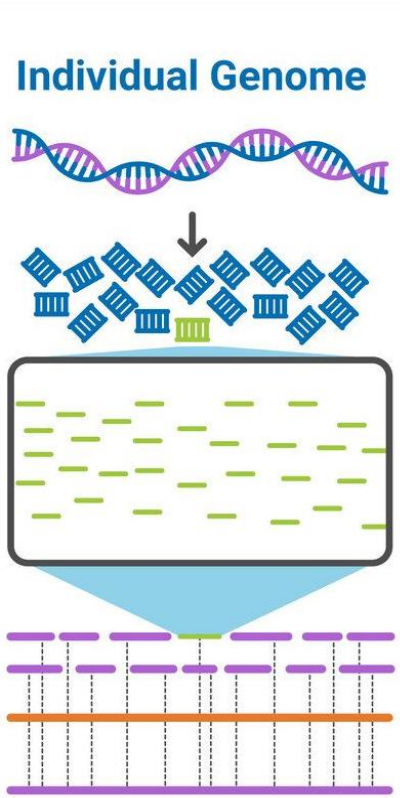
First Generation Sequencing (Sanger)

Second Generation Sequencing (Illumina, etc.)

Third Generation Sequencing (Oxford Nanopore, etc.)

International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature 409, 860–921 (2001).
Sergey Nurk et al. ,The complete sequence of a human genome. Science 376,44-53(2022).
Rhie, A., Nurk, S., Cechova, M. et al. The complete sequence of a human Y chromosome. Nature 621, 344–354 (2023).

Human DNA Sequencing



1. Break genome into small, overlapping fragments
2. **Use sequencers to sequence** millions of sequence **reads**
3. **Align reads** into a reference genome for **variant detection**

	First Gen	Second Gen	Third Gen
Year of born	1977	Mid-2000	2012
Sequencing time for high-quality HG	13 years	Days	<1 day
Cost of HG	>$500 million	$600-2000	$1,000-4,000

The cost of sequencing a human genome. Genome.gov.

Third Generation Sequencing Technologies: **Long-reads**

Long read can capture **structural** and **positional** information in DNA

Article | Published: 30 April 2018

Accurate detection of complex structural variations using single-molecule sequencing

Fritz J. Sedlazeck, Philipp Rescheneder, Moritz Smolka, Han Fang, Maria Nattestad, Arndt von Haeseler & Michael C. Schatz

SPECIAL ISSUE RESEARCH ARTICLE | HUMAN GENOMICS

The complete sequence of a human genome

SERGEY NURK, SERGEY KOREN, ARANG RHIE, MIKKO RAUTIAINEN, ANDREY V. BZIKADZE, ALLA MIKHEENKO, MITCHELL R. VOLLGER, NICOLAS ALTEMOSE, LEV URALSKY, [...], AND ADAM M. PHILLIPPY +90 authors Authors Info & Affiliations

Article | Open access | Published: 16 June 2022

Phasing analysis of lung cancer genomes using a long read sequencer

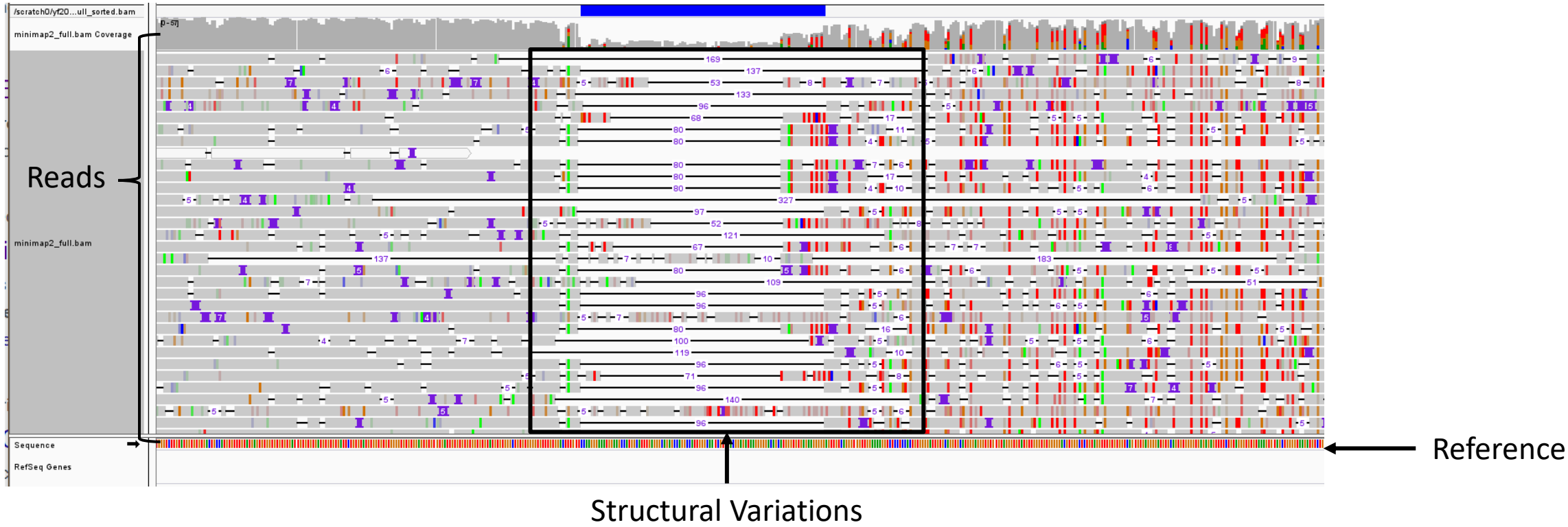
Yoshitaka Sakamoto, Shuhei Miyake, Miho Oka, Akinori Kanai, Yosuke Kawai, Satoi Nagasawa, Yuichi Shiraishi, Katsushi Tokunaga, Takashi Kohno, Masahide Seki, Yutaka Suzuki & Ayako Suzuki

	Second Gen	Third Gen
Read Length	150 bp	Up to **4 Mbp**
Error Rate	<0.1%	1-3%
Advantages	INDEL	Methylation and SV calling
Disadvantages	Too short/Limited in certain regions	Cost

Oxford Nanopore Technologies. https://nanoporetech.com/

How to use sequenced reads to discover the genetic variants?

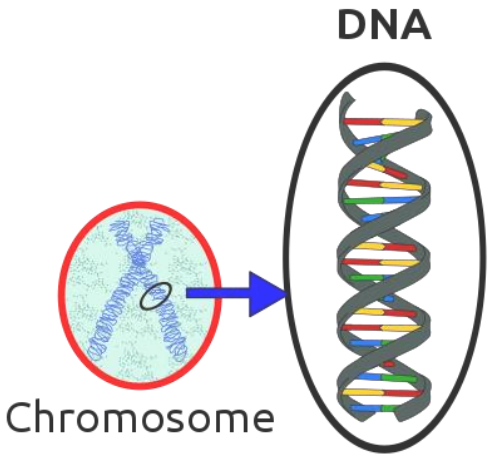
Read Alignment for Variant Detection

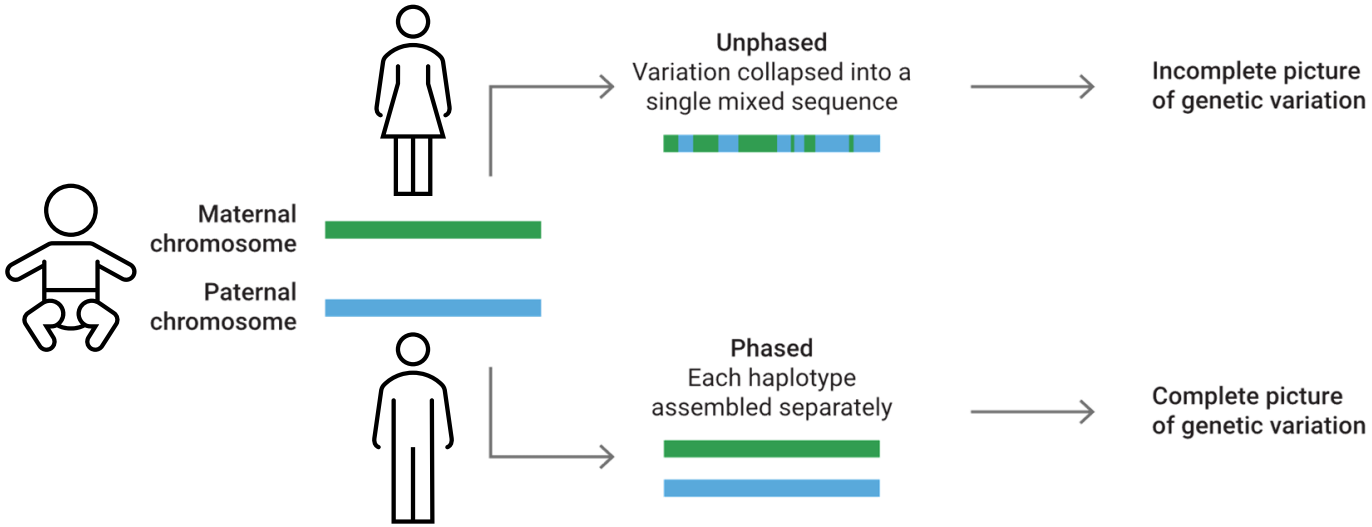


Sample HG002 aligned to human reference GRCh38, location: chr2:240,564,644-240,565,183

Is read alignment enough to decode the human genetic variants?

Variant Phasing



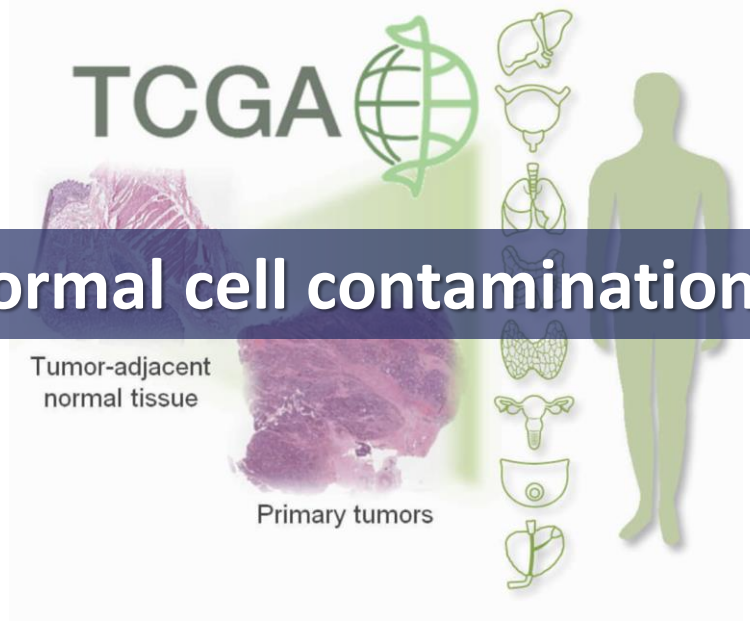


- **Haplotype**: the assignment of a group of variants, they tend to be inherited together.
- **Phasing variants into haplotypes helps us decipher the interaction among variants!**
- 1–5% of human genes are influenced by unbalanced DNA sequence variants (Single Nucleotide Variants - SNVs)
  - Human disorder
  - Disease causing
  - Different phenomena in common diseases

Sequencing 101: ploidy, haplotypes, and phasing — how to get more from your sequencing data https://www.pacb.com/blog/ploidy-haplotypes-and-phasing/

Sample purity affects variant detection and phasing!

Tumor Purity Estimation



Normal cell contaminations

- **Affects allele frequency estimation**
- **Discard variant gain/loss**
- **Affects clinical decisions**

National Cancer Institute. Study Uses Open Data to Analyze "Normal" Tissue Near Tumors https://www.cancer.gov/news-events/cancer-currents-blog/2017/tumor-adjacent-tissue

Research Questions

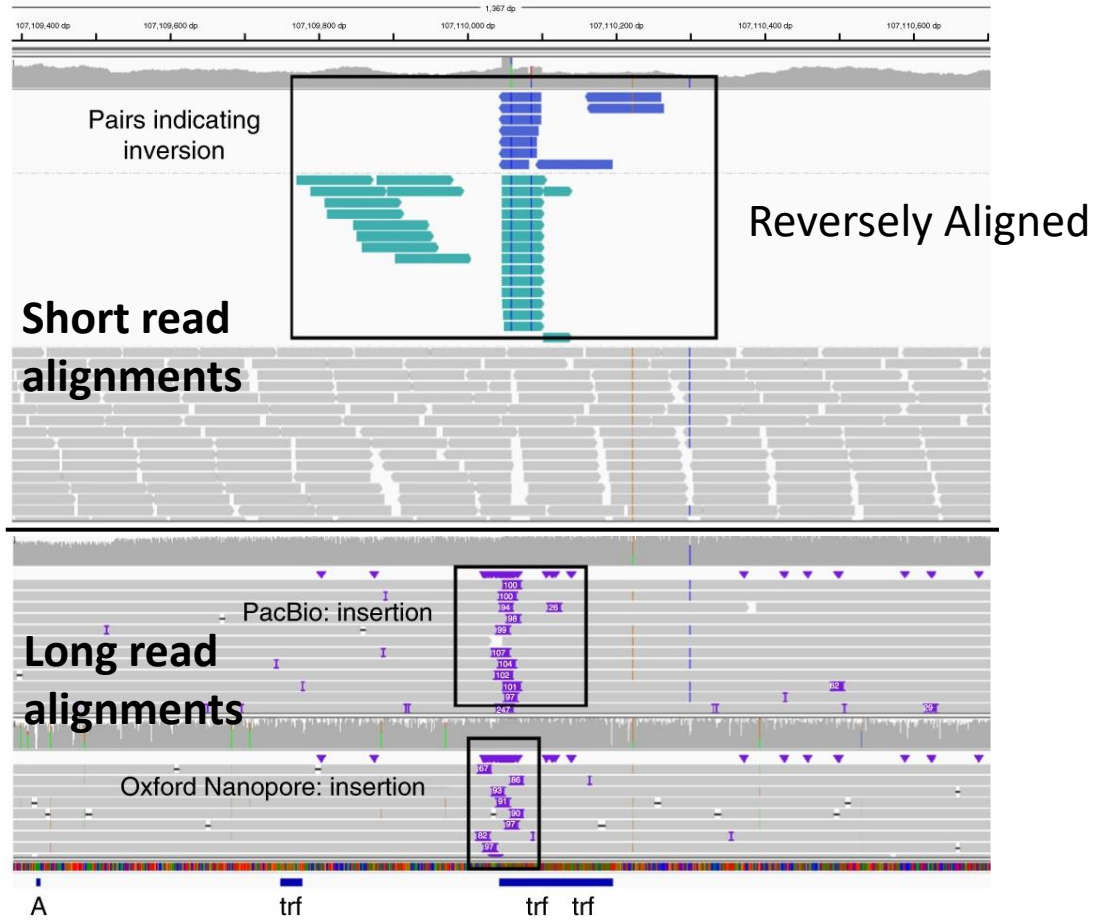
Efficiently and accurately aligning reads to the reference for better variant calling	**Vulcan**
Improve variant phasing	**MethPhaser**
Tumor purity estimation	**MethPhaser-Cancer**

Vulcan

Improved long-read mapping and structural variant calling via dual-mode alignment

Long-Read Technologies Have Enabled Accurate Detection of Structural Variations

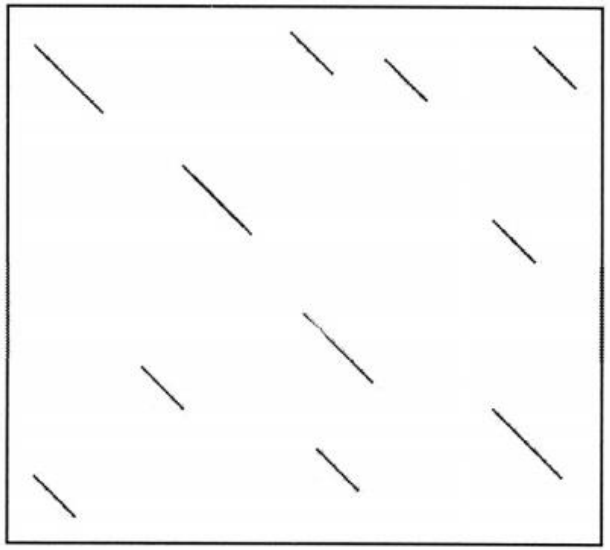
- SVs contribute to
  - Polymorphic variation; pathogenic conditions
  - Large-scale chromosome evolution
  - Human diseases such as cancer, autism, and Alzheimer's.
- Long reads can discover SVs that short reads cannot discover or identify as wrong type
- **The precision of long-read alignment fundamentally affects SV calling accuracy!**



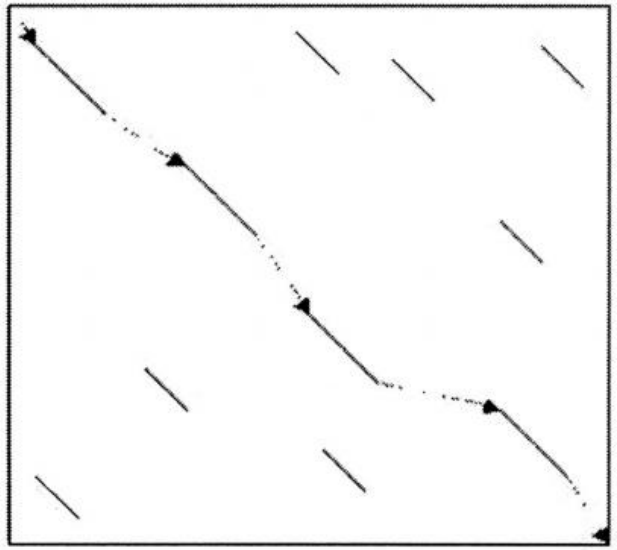
Sedlazeck, Fritz J., et al. "Accurate detection of complex structural variations using single-molecule sequencing." Nature methods 15.6 (2018): 461-468.

Read Alignment

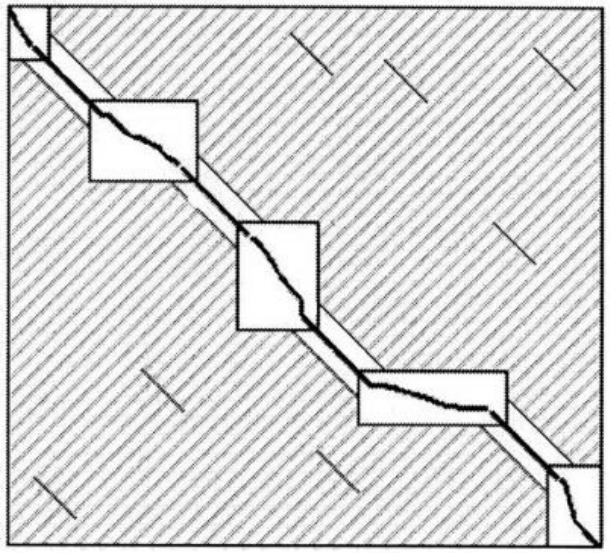
Figure from: Brudno et al. *Genome Research*, 2003



Seeding:
Hash based anchor locating



Chaining:
Search for colinear blocks



Extending:
Pairwise Alignment

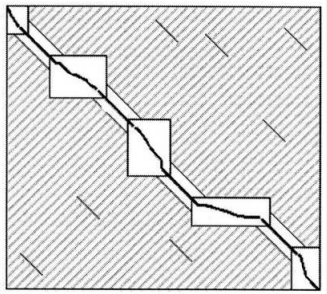
Pairwise Alignment

Smith Waterman Algorithm

Linear Gap Penalty:

- Match Score: +1
- Mismatch Score: -2
- Gap Extension Score: -1

Different scoring schemes lead to different alignment results



S		A_1	A_2	G_3	A_4	T_5	C_6	A_7	C_8
	0	0	0	0	0	0	0	0	0
A_1	0	1	1	0	1	0	0	1	0
A_2	0	1	2	1	1	0	0	1	0
G_3	0	0	1	3	2	1	0	0	0
A_4	0	1	1	2	4	3	2	1	0
A_5	0	1	2	1	3	2	1	3	2
T_6	0	0	1	0	2	4	3	2	1
T_7	0	0	0	0	1	3	2	1	0
T_8	0	0	0	0	0	2	1	0	0
C_9	0	0	0	0	0	1	3	2	1
A_{10}	0	1	1	0	1	0	2	4	3
C_{11}	0	0	0	0	0	0	1	3	5
Score: 5									

Gaps

AAGAATTTCAC

AAGA___TCAC

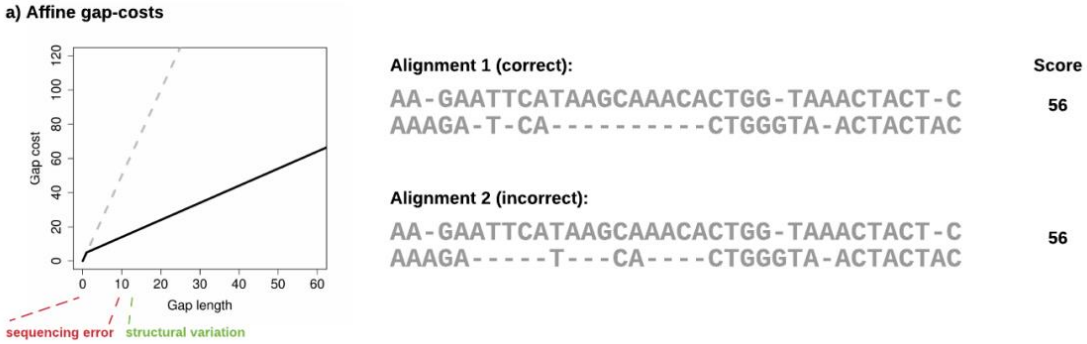
Scoring Scheme: Gap Penalty

- g_O : gap opening penalty
- g_E : gap extension penalty
- g_M : gap matching score
- i : current length of the gap
- g_D : gap decay parameter
- *m* and *n* are the length of two sequences

Affine Gap Penalty (Minimap2)

$$G(i) = gO + gE \times i$$

- Gap penalty linearly increase with gap length
- Time complexity: $\boldsymbol{O(mn)}$

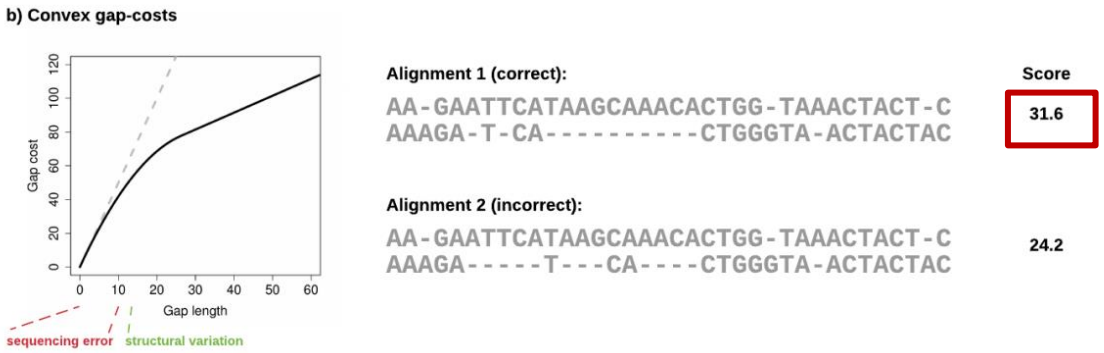


Convex Gap Penalty (NGMLR)

$$G(i) = \begin{cases} g_O, & i = 0 \\ G(i-1) + min\begin{cases} g_M \\ g_E + g_D * (i-1), \end{cases} & i > 0 \end{cases}$$

0 < gap decay < 1

- Gap penalty grows slower with larger gaps
- Time complexity: $\boldsymbol{O(mnlog(m+n))}$



Sedlazeck, F.J., Rescheneder, P., Smolka, M. et al. Accurate detection of complex structural variations using single-molecule sequencing. Nat Methods 15, 461–468 (2018).

Gap Penalty

Alignment 1 and Alignment 2:

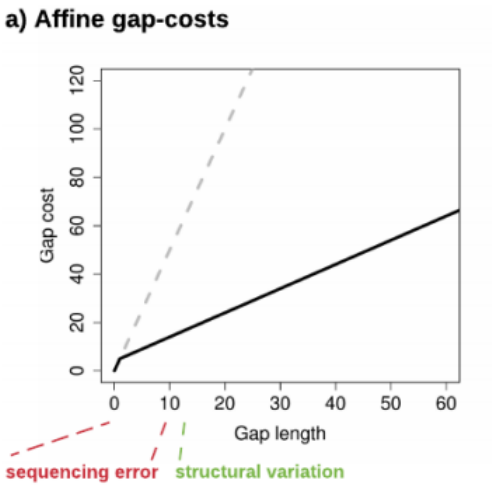
- both 9 gap extensions

Affine Gap Penalty:

- If the number of extended gap is the same, the score is the same
- The gap penalty was outweighed by lots of sequencing errors

Convex Gap Penalty:

- Longer gap, higher score

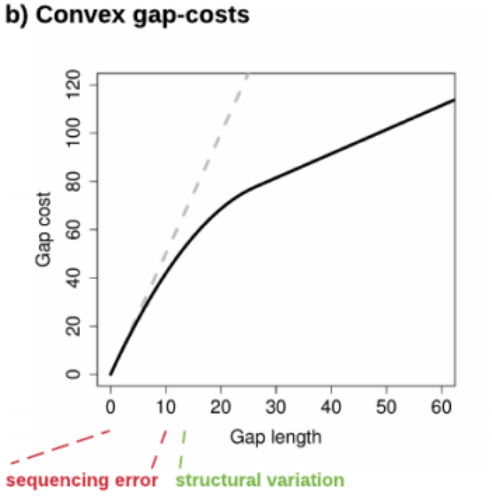


Alignment 1 (correct): Score

AA-GAATTCATAAGCAAACACTGG-TAAACTACT-C 56
AAAGA-T-CA----------CTGGGTA-ACTACTAC

Alignment 2 (incorrect):

AA-GAATTCATAAGCAAACACTGG-TAAACTACT-C 56
AAAGA-----T---CA----CTGGGTA-ACTACTAC



Alignment 1 (correct): Score

AA-GAATTCATAAGCAAACACTGG-TAAACTACT-C 31.6
AAAGA-T-CA----------CTGGGTA-ACTACTAC

Alignment 2 (incorrect):

AA-GAATTCATAAGCAAACACTGG-TAAACTACT-C 24.2
AAAGA-----T---CA----CTGGGTA-ACTACTAC

Sedlazeck, F.J., Rescheneder, P., Smolka, M. et al. Accurate detection of complex structural variations using single-molecule sequencing. Nat Methods 15, 461–468 (2018).

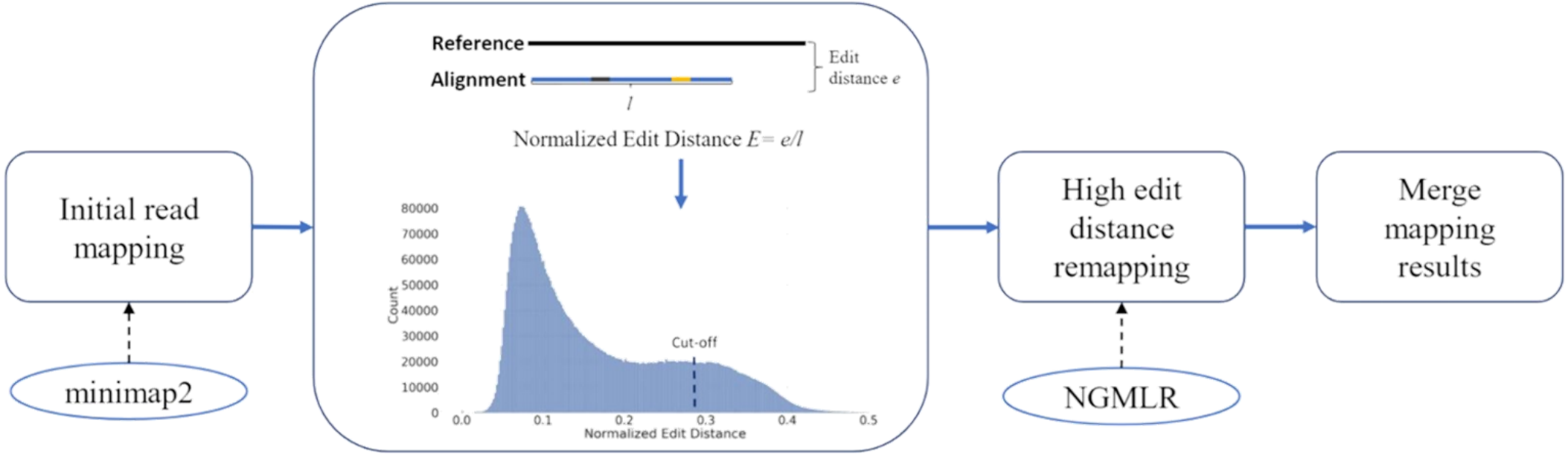
Different Regions May Suit Different Gap Penalties

- Low mutation rate (e.g., House keeping genes)
- High mutation rate (e.g., genes involved in immune responses)
- Highly variable gene across human population (e.g., LPA, Cyp2d6)

Motivation

- **Different human genomic regions suit different read alignment mechanisms**
- More precise read alignments produce better structural variation calling results
- Available methods mostly focus on improving seeding and chaining stage
- **The first method for dual-mode alignment in the extension stage**

Vulcan Pipeline



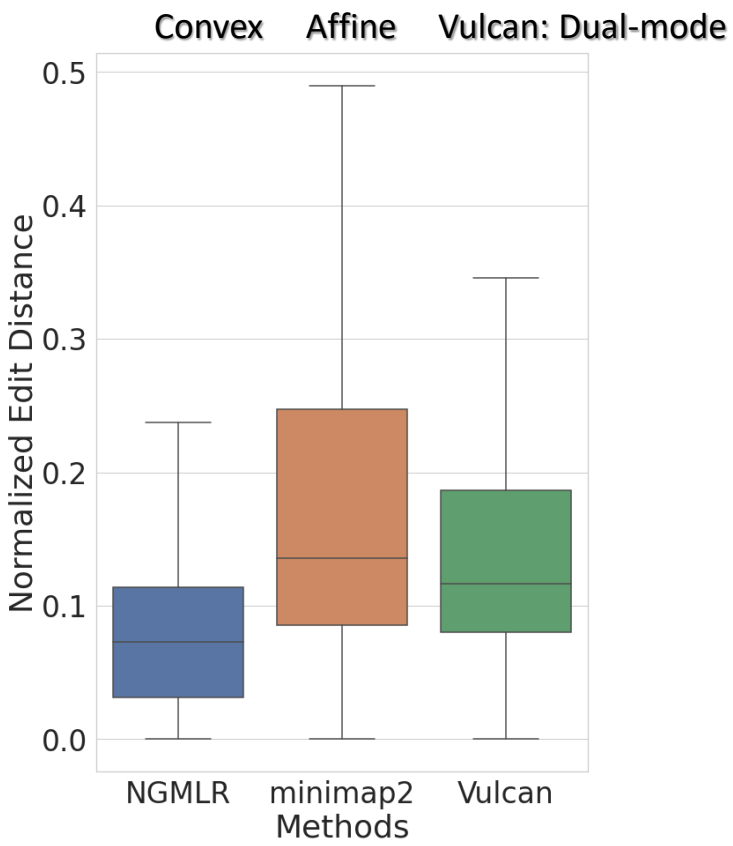
Affine Gap Penalty

Convex Gap Penalty

Fu, Yilei, et al. "Vulcan: Improved long-read mapping and structural variant calling via dual-mode alignment." GigaScience 10.9 (2021): giab063.

Vulcan Improves Read Alignment

Edit distance profile in human reads (Sample HG002) alignment result



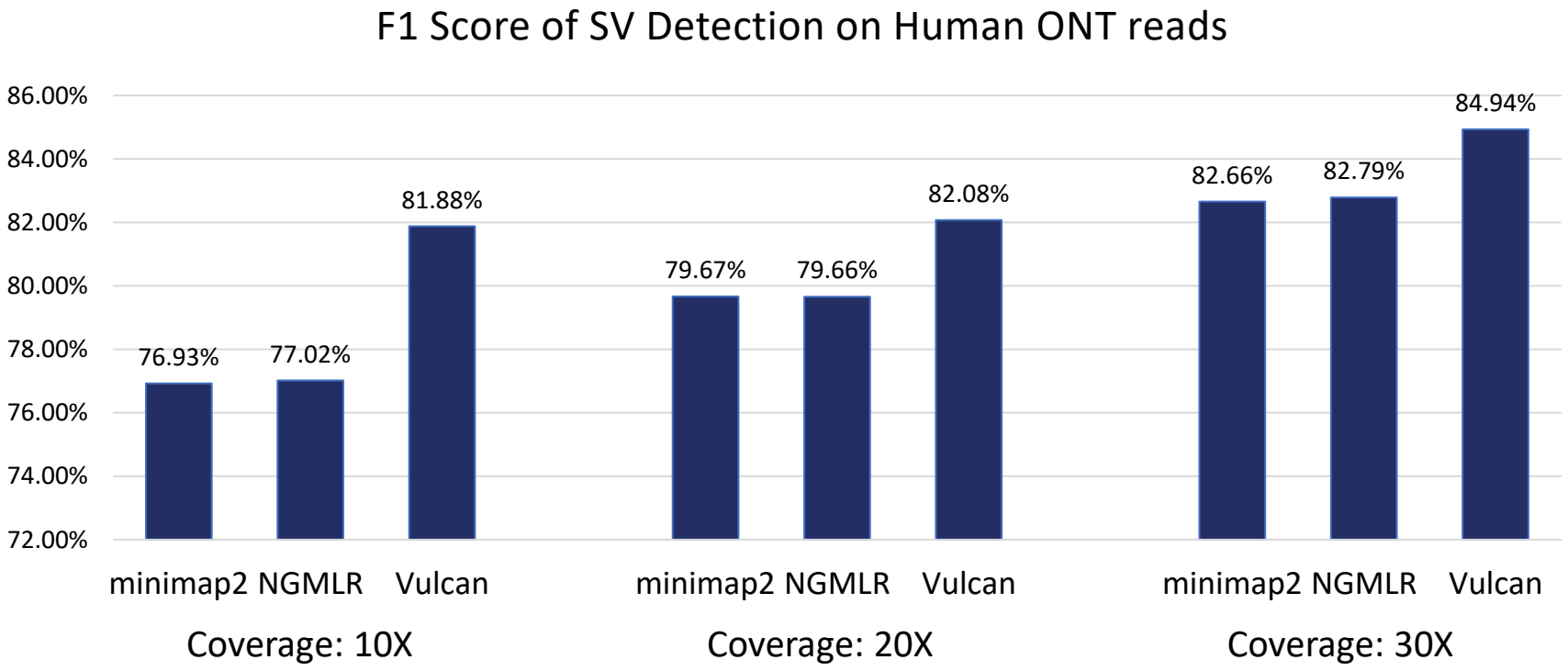
- Edit distance: differences between two sequences
  - An estimation of read alignment quality
- Vulcan achieves an overall smaller edit distance than minimap2 (affine gap penalty method)

Zook JM, Hansen NF, Olson ND, et al (2020) A robust benchmark for detection of germline large deletions and insertions. Nature biotechnology 1–9

Vulcan Improves SV Calling


nature biotechnology
RESOURCE
https://doi.org/10.1038/s41587-020-0538-8
Check for updates
A robust benchmark for detection of germline large deletions and insertions

Evaluation dataset: HG002 sample with robust SV ground truth



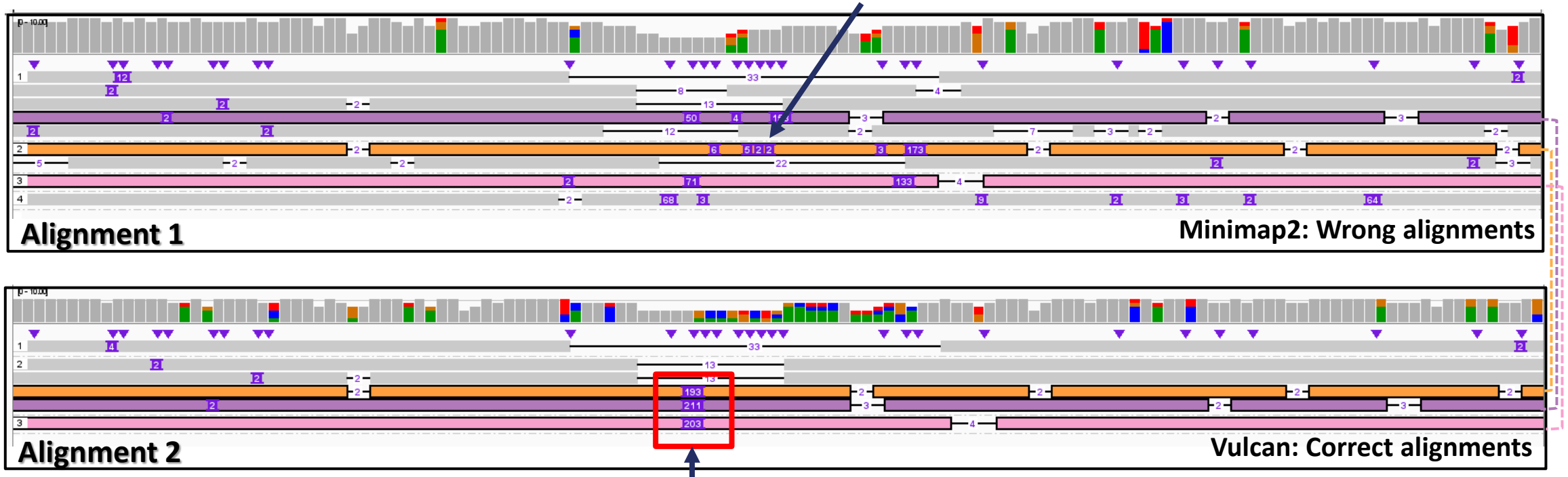
Zook JM, Hansen NF, Olson ND, et al (2020) A robust benchmark for detection of germline large deletions and insertions. Nature biotechnology 1–9

Vulcan Improves SV Calling on *EEIG2* Gene

EEIG2 (FAM102B) is a signature gene of microcystic adenoma, a kind of tumor

- **Same colors:** same reads
- **Number in blue box:** size of insertion

Fragmented

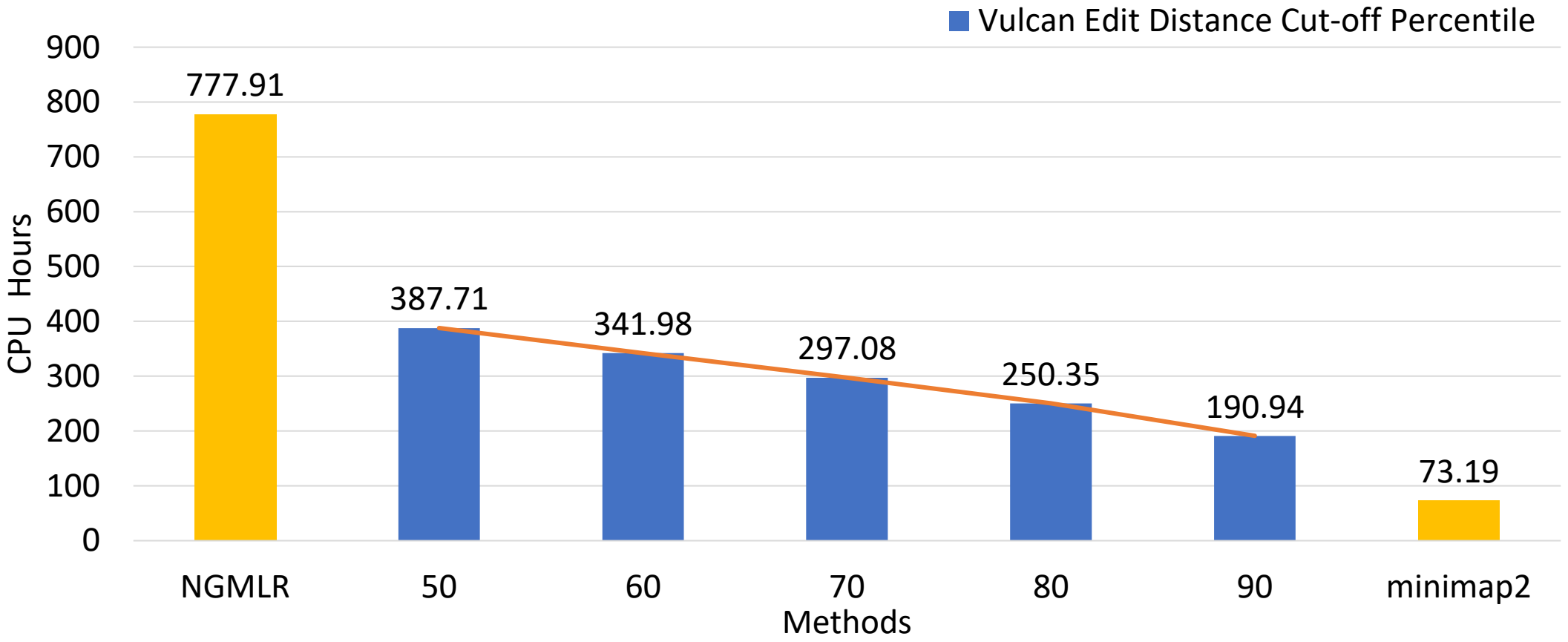


Consolidated structural variation

chr1: 108,567,498–108,567,633

Fu, Yilei, et al. "Vulcan: Improved long-read mapping and structural variant calling via dual-mode alignment." GigaScience 10.9 (2021): giab063.

Vulcan Speed Up



Conclusion

JOURNAL ARTICLE

Vulcan: Improved long-read mapping and structural variant calling via dual-mode alignment

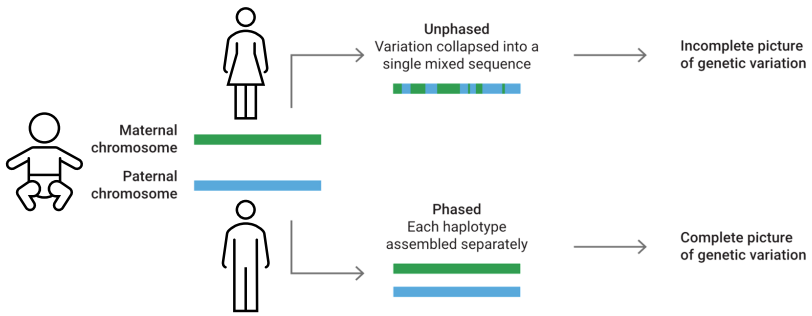
Yilei Fu, Medhat Mahmoud, Viginesh Vaibhav Muraliraman, Fritz J Sedlazeck, Todd J Treangen Author Notes

GigaScience, Volume 10, Issue 9, September 2021, giab063, https://doi.org/10.1093/gigascience/giab063

- Vulcan is the first long read aligner that can **utilize two kinds of gap penalty** to accommodate the varying mutation rate on human genome.
- Improved read mapping accuracy and SV calling accuracy.
- Comparing to the method using convex gap penalty, Vulcan reduces the **time usage**.

Research Questions

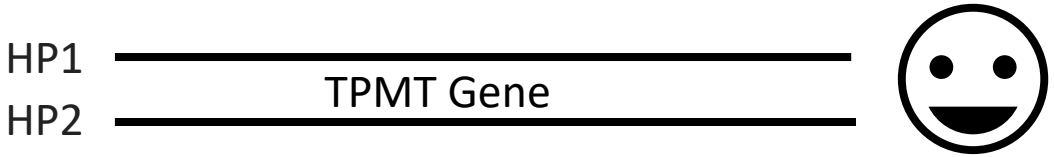
Efficiently and accurately aligning reads to the reference for better variant calling	**Vulcan**
Improve variant phasing	**MethPhaser**
Tumor purity estimation	**MethPhaser-Cancer**



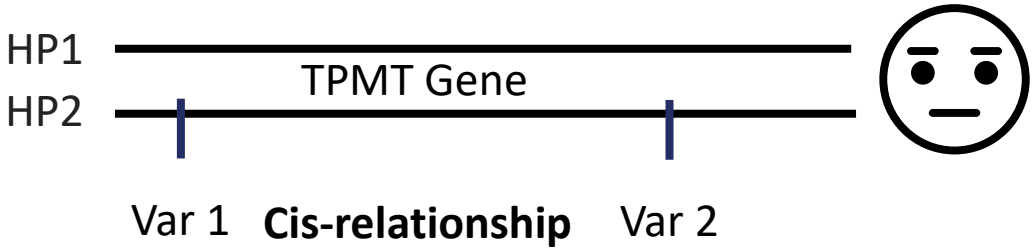
MethPhaser

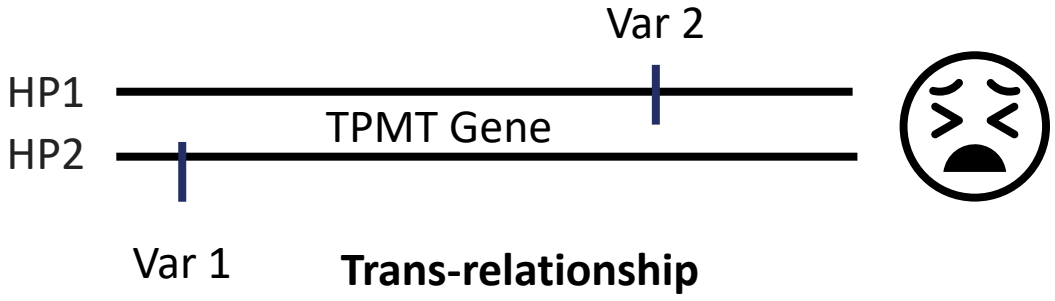
Methylation-based Long-read Haplotype Phasing of Human Genomes

Why Phasing Matters?



- thiopurine methyltransferase (TPMT) gene - encodes the enzyme that metabolizes thiopurine drugs
- Two variants, rs1800460 and rs1142345 (>8000 bases apart)
  - cis: TPMT*1/*3A diplotype (intermediate metabolizer)
  - trans: TPMT*3B/*3C diplotype (poor metabolizer)
  - *1/*3A diplotype is more common but less severe than *3B/*3C





Haplotype Phasing Methods

		Common Variants	Rare Variants	**Novel Variants**	**Fully Phased**
Population-based Phasing	Based on common variants	√	×	×	×
Trio-based Phasing	Based on parents' variants	√	√	×	√
Long-read-based Phasing	Based on individual's variants	√	√	√	X

Cost Collection effort

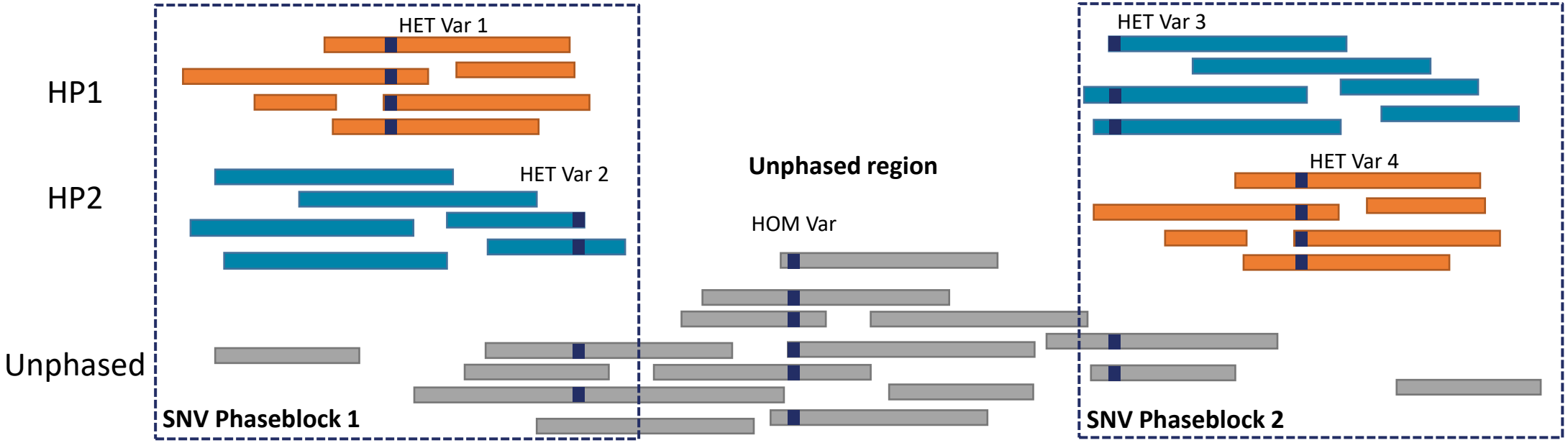
- Common variants: allele frequency > 0.05
- Rare variants: low-frequency but exists on paternal samples
- Novel variants: only exist on the sequenced individual

Long-read Phasing Doesn't Solve Everything

Long-read-based phasing relies on:

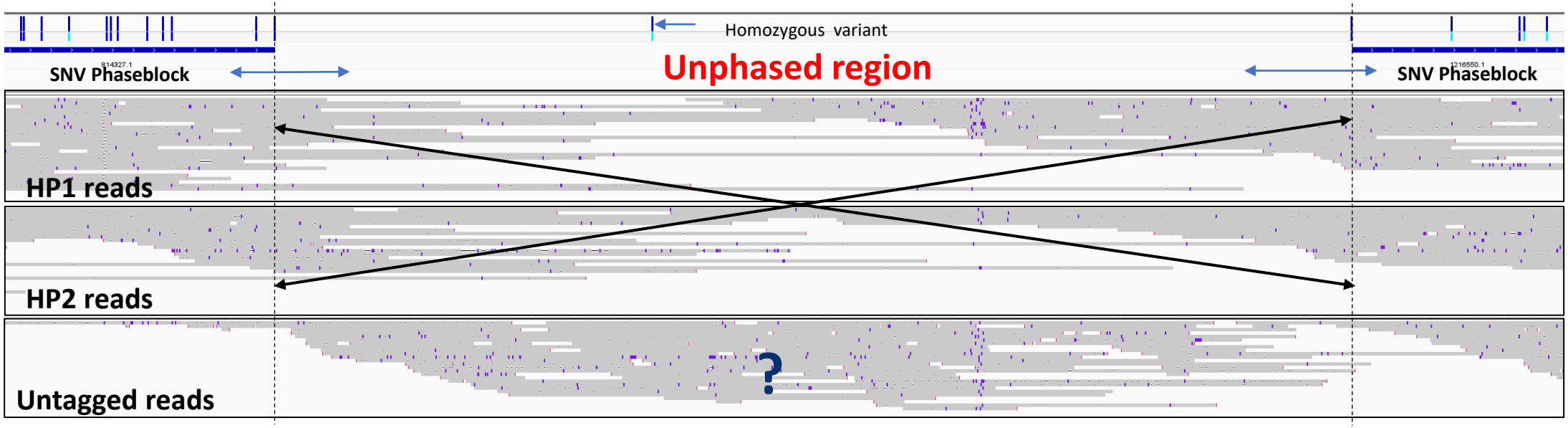
- heterozygous (different on two alleles) variants.
- large read length to connect heterozygous variants

Long-read-based phasing fails when the stretch of the homozygosity longer than read length



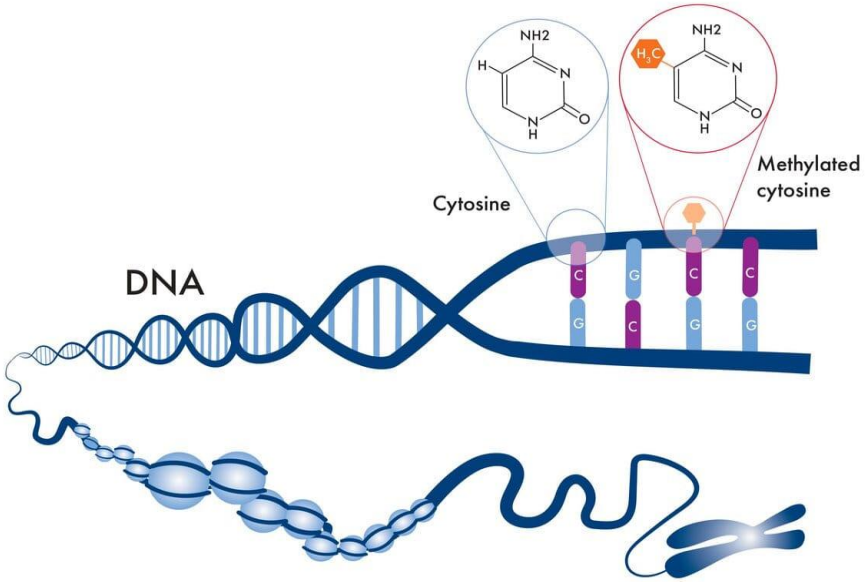
Long-read Phasing Doesn't Solve Everything

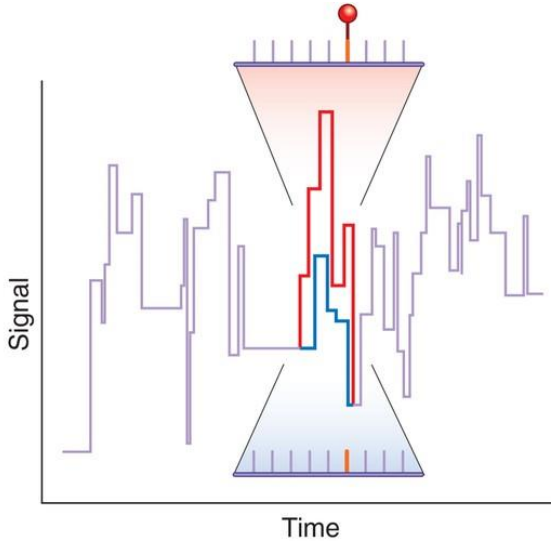
We have 4,518 unphased gaps on the human genome with the state-of-the-art long-read sequencing and phasing methods.



Yilei Fu, Sergey Aganezov, Medhat Mahmoud, John Beaulaurier, Sissel Juul, Todd J. Treangen, Fritz J Sedlazeck, MethPhaser: methylation-based haplotype phasing of human genomes, bioRxiv 2023.05.12.540573;

Methylation: An Epigenetic Signal





...ATCAGATGCTGCGATGGTACCCGCTAGCTACG...

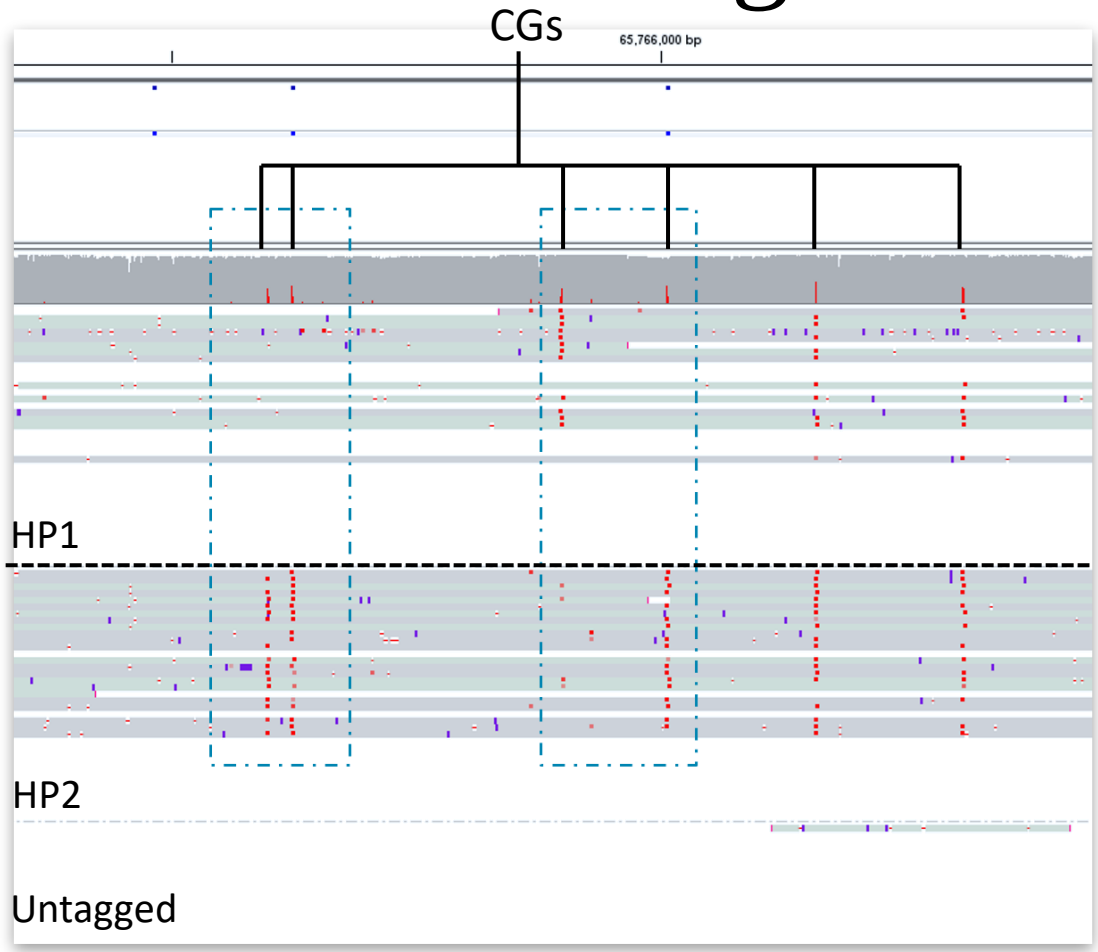
A score showing the probability of each CG on each read being methylated

Only 3rd Gen sequencing technologies can directly report the methylation scores!

Haplotype-specific Methylations Provide Insight for Improving SNV-based Phasing

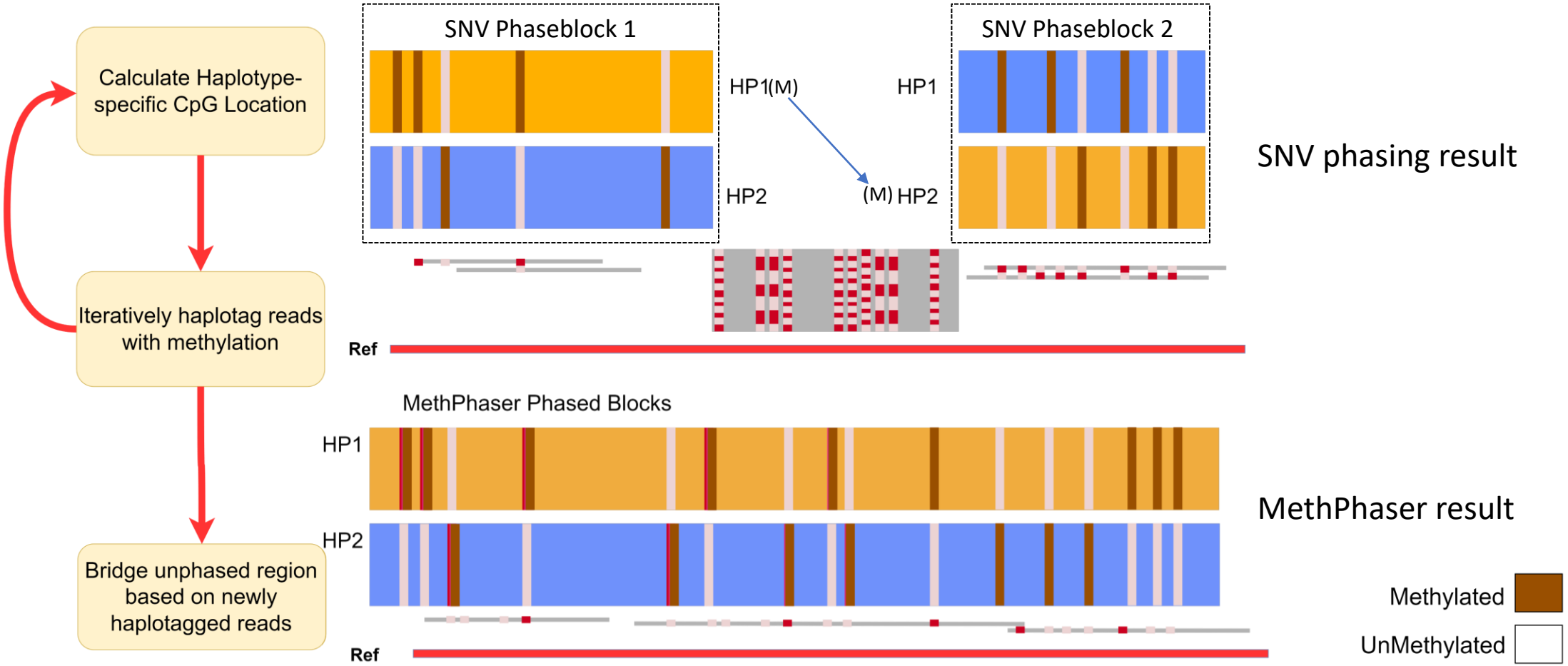
- MethPhaser: Methylation based read haplotagging
- Use haplotype-specific methylation as signal to cluster un-haplotagged reads from SNV based methods

Goal: Fill in the gaps



Methylated
UnMethylated

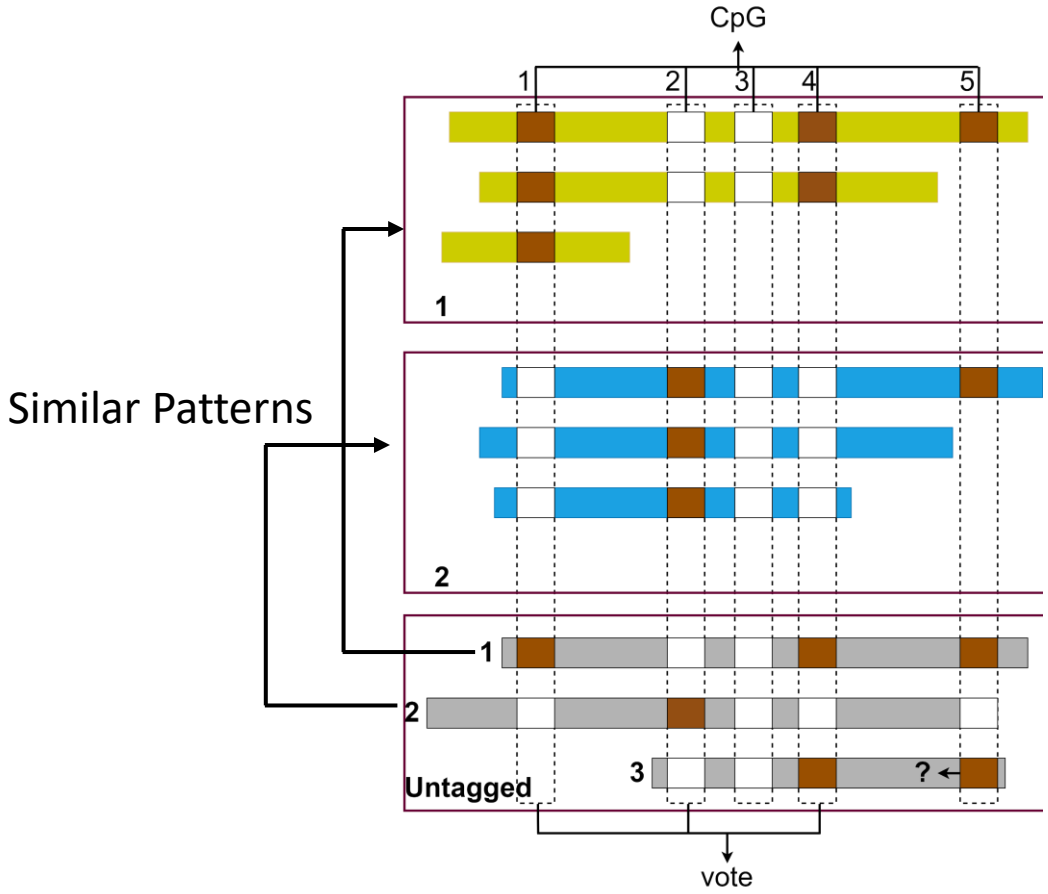
MethPhaser Algorithm – Overview



MethPhaser Algorithm – Overview

- **Build methylation pattern classifiers based on SNV phasing results**
- **Use the classifier to phase reads in unphased region into haplotypes**

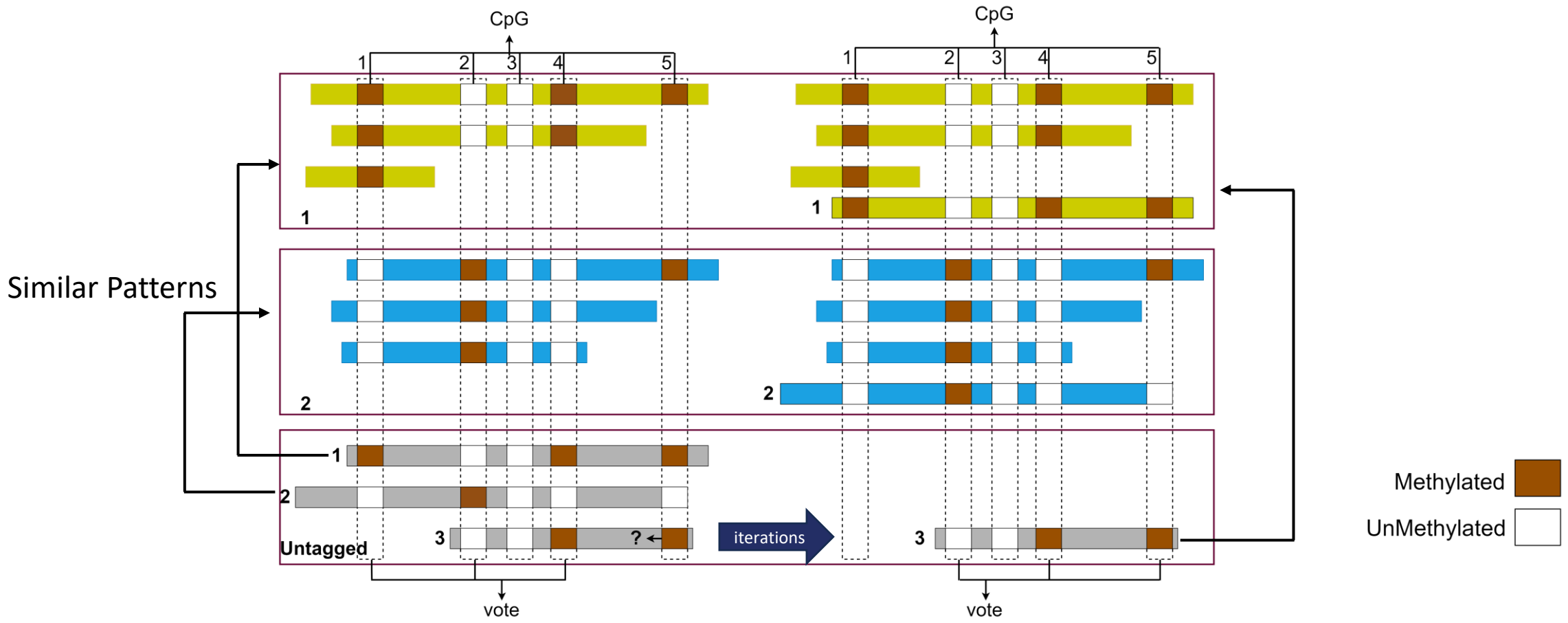
MethPhaser Algorithm – Read Assignment Within the Same Block



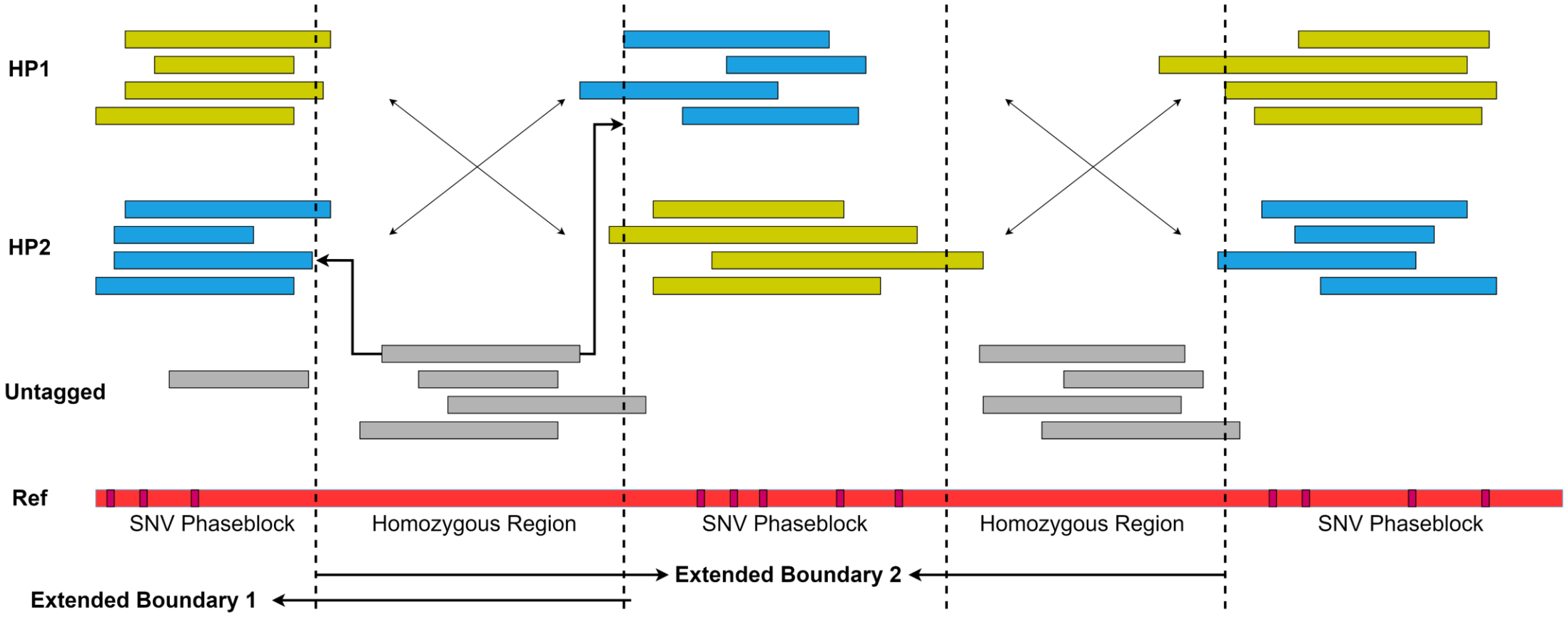
Methylated

UnMethylated

MethPhaser Algorithm – Iterative Read Assignment



MethPhaser Algorithm – Phaseblock Connection



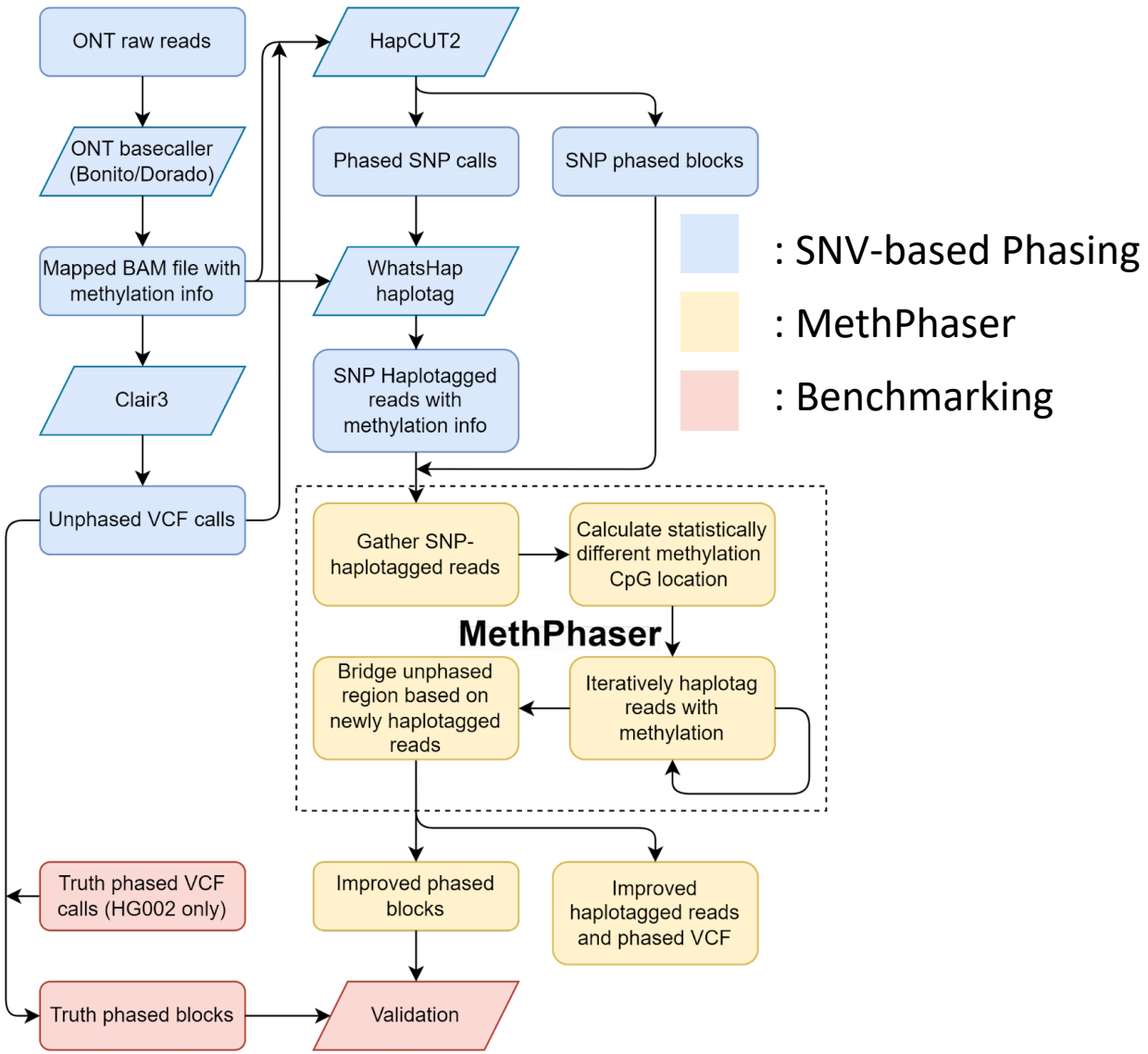
MethPhaser Benchmarking

Evaluation dataset:

- HG002 sample
- Trio-phased with parents' sequencing data
  - **Ground truth**
- Novel variants excluded from the evaluation
- Comparison
  - SNV Phasing
  - SNV Phasing + MethPhaser

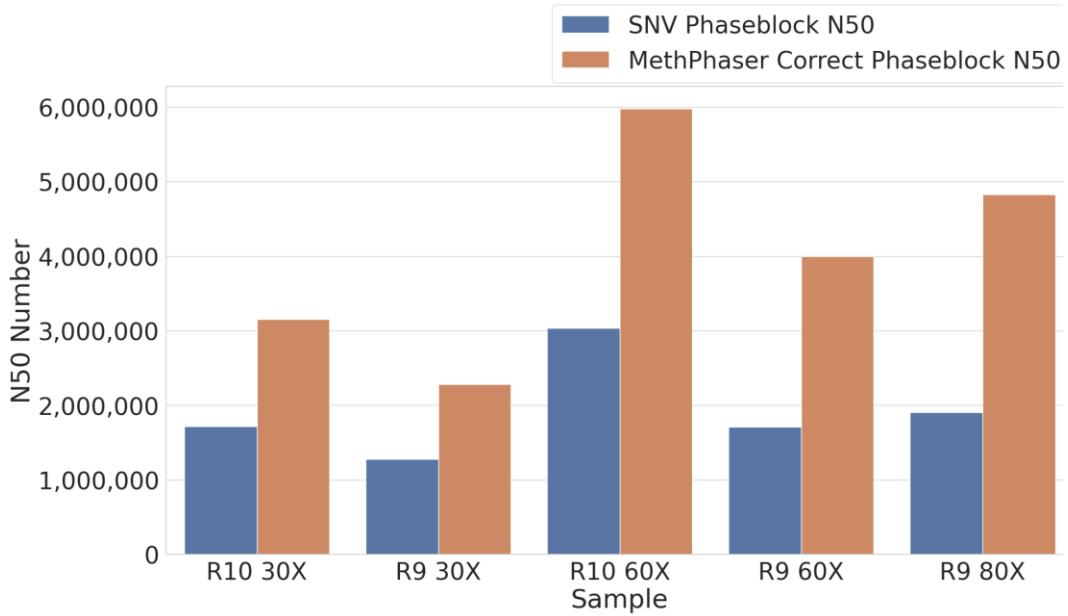
Evaluation criteria:

- N50: reflects the length of phaseblocks
- Switch error: single variant assigned to wrong haplotype
- Flip error: two variants assigned to wrong haplotypes

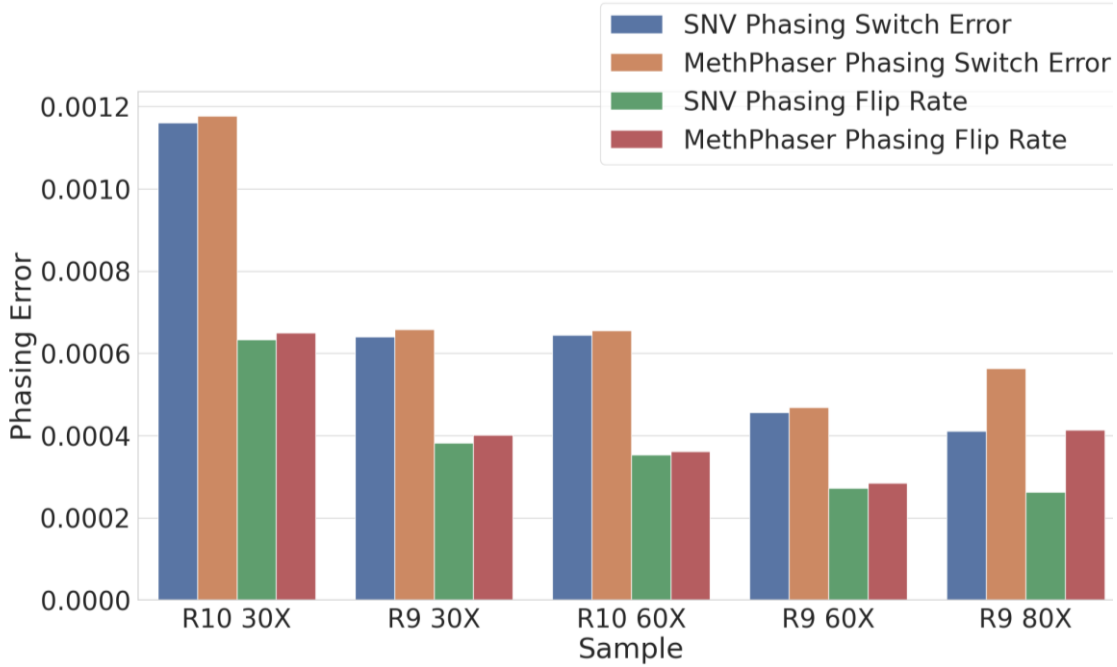


MethPhsaer Significantly Improved Phaseblock Length on HG002 Sample

N50, 1.6-2x increase



N50 improves with only small increases of phasing errors



R9: ONT Flow Cell R9.4.1, R10: ONT Flow Cell R10.4.1; 30X and 60X are coverages

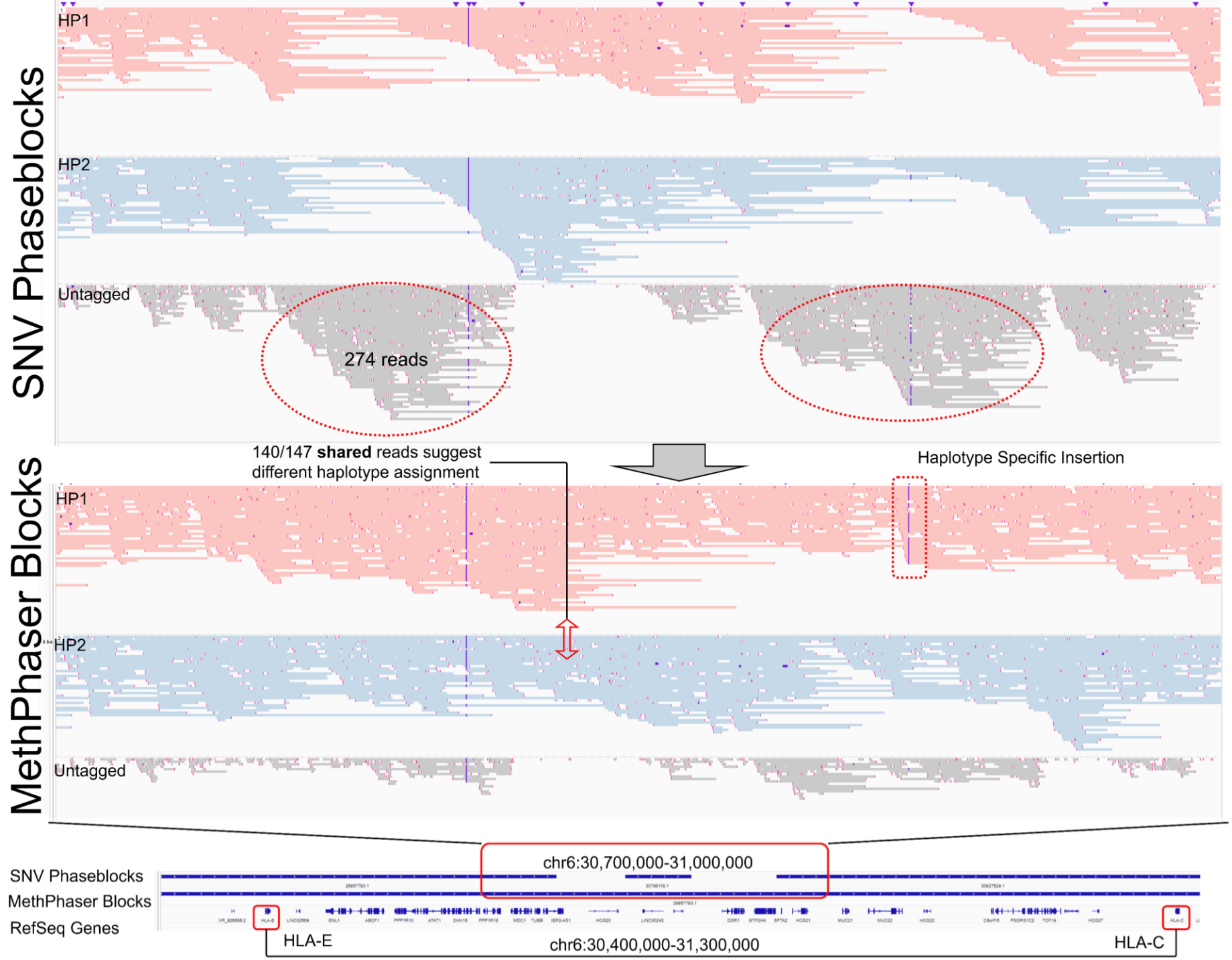
Use case: MethPhaser Connects Medically Relevant Locations


ARTICLES
https://doi.org/10.1038/s41587-021-01158-1
nature biotechnology
Check for updates
Curated variation benchmarks for challenging medically relevant autosomal genes

Method	Phased Medically Relevant Genes Number	Required Block Number
SNV-based	258	160
MethPhaser	265	140

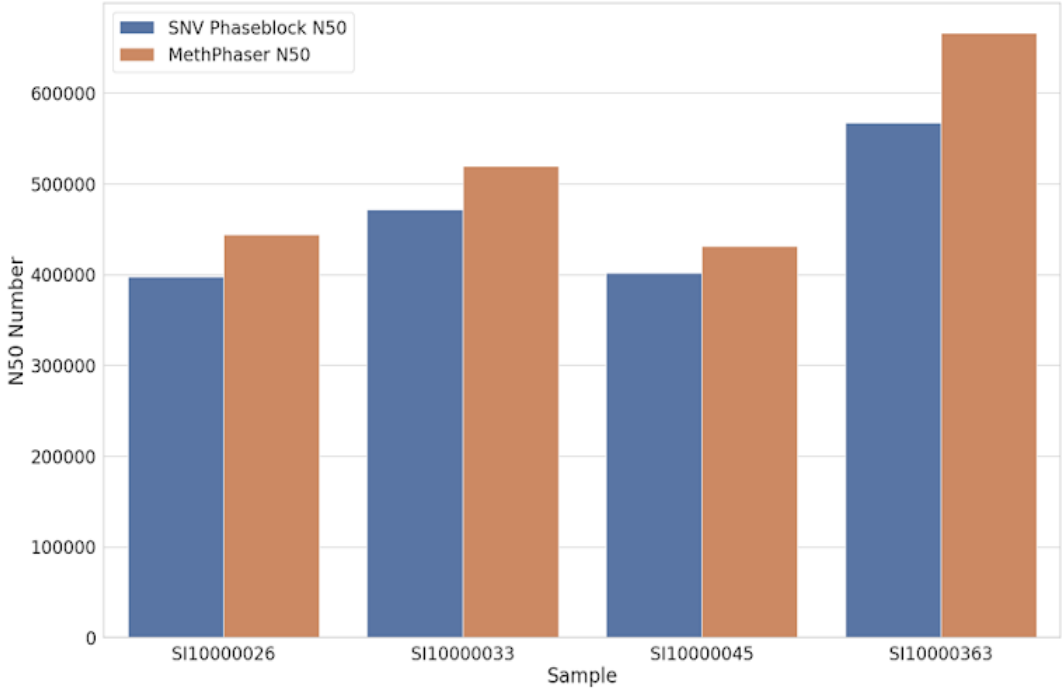
MethPhaser used fewer block to connect more hard-to-resolve medically relevant genes (total 272).

MethPhaser Connects *HLA-E* and *HLA-C* Genes

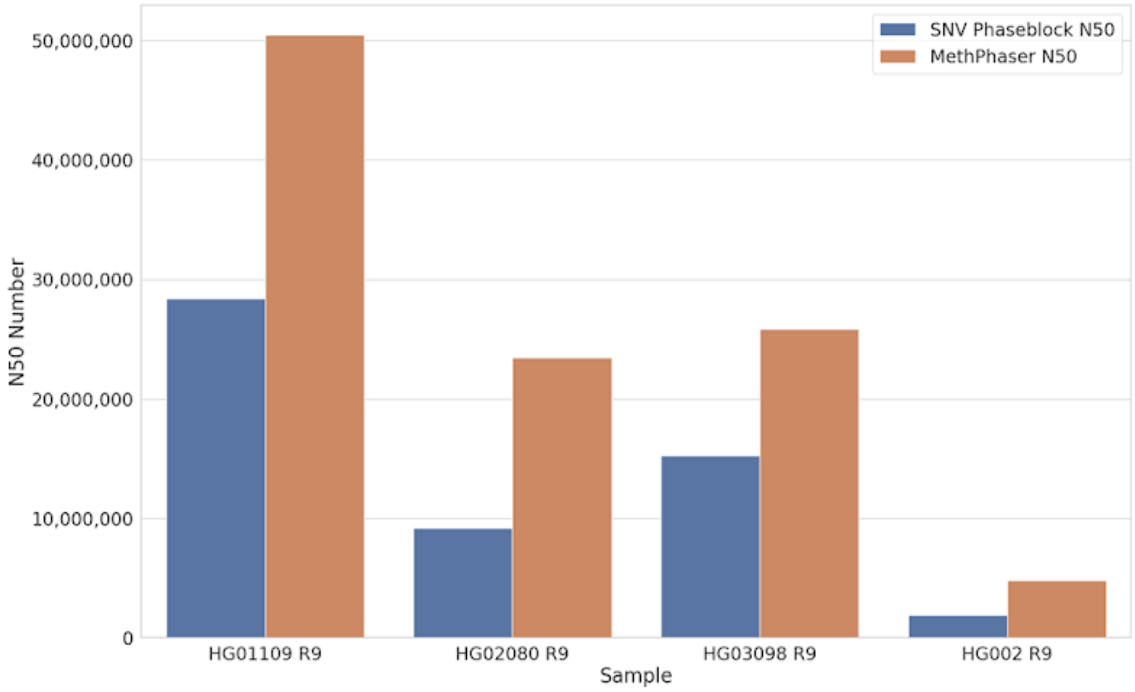


More Tests on Different Ethnic Background and Tissues

Patient Blood Samples



HPRC Samples (Ultra long reads)



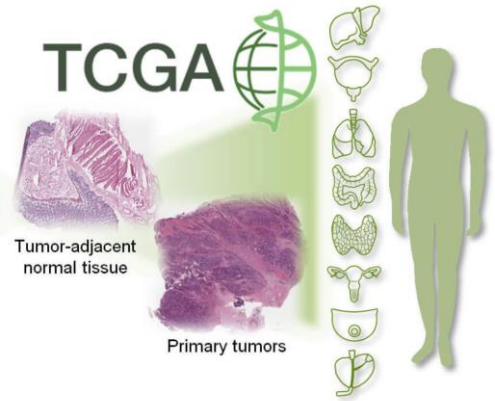
Male, PUR | Female, KHV | Male, MSL | Male, PUR (Regular Long-reads)

Conclusion: Methylation as an extension of SNV phasing

- MethPhaser is the first method that combines long-read epigenomic and genomic variant for genome scale phasing
- MethPhaser achieves 1.5-3X phaseblock N50 length against SNV-based methods on human samples
- MethPhaser rescued previously un-haplotagged reads
- MethPhaser can be directly attached to traditional SNV-based pipeline for great improvement with little cost

Research Questions

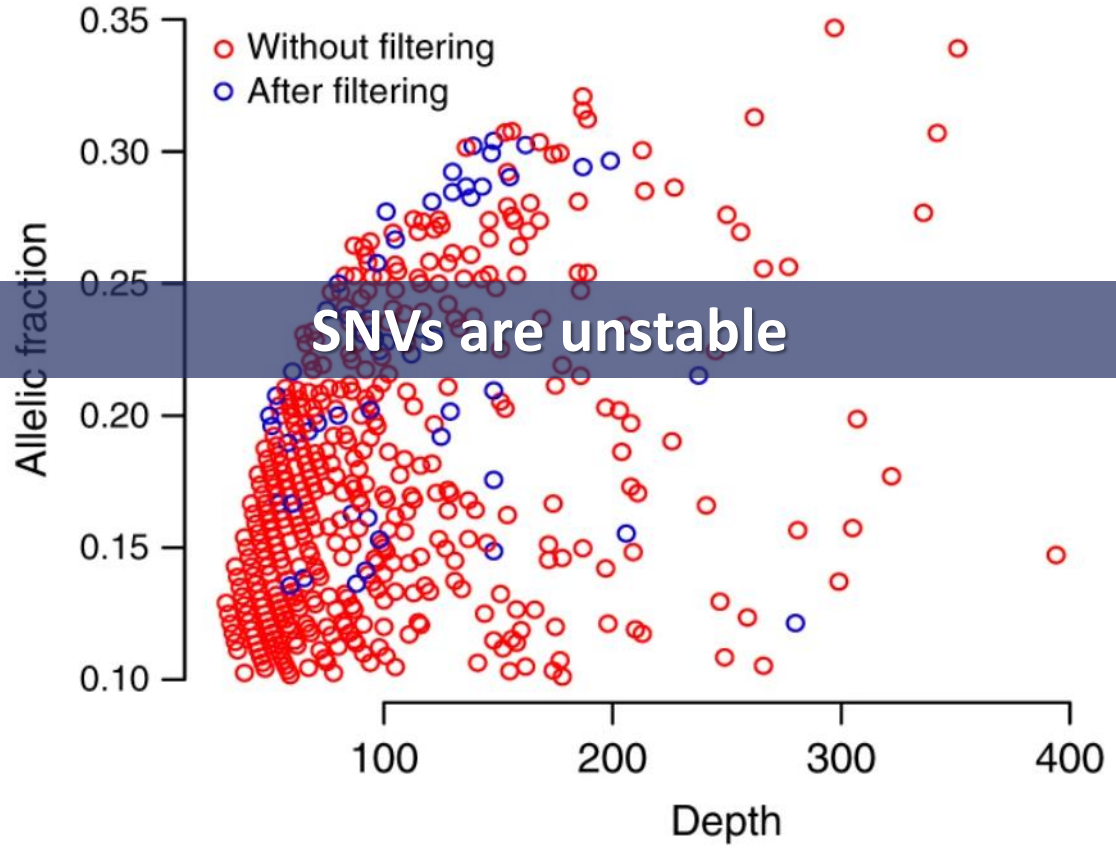
Efficiently and accurately aligning reads to the reference for better variant calling	**Vulcan**
Improve variant phasing	**MethPhaser**
Tumor purity estimation	**MethPhaser-Cancer**



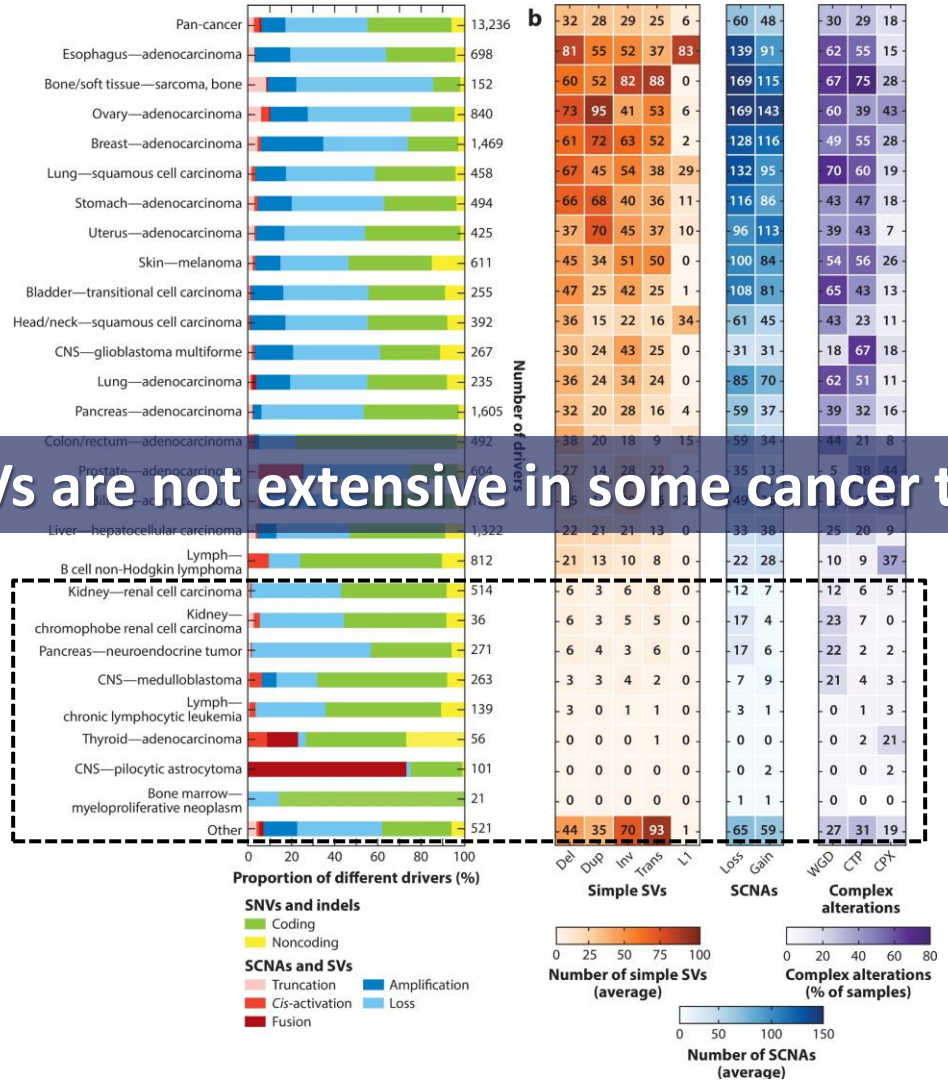
MethPhaser-Cancer

Automated tumor purity estimation and read classification with methylation

SVs and SNVs are not Sufficient for Tumor Purity Estimation



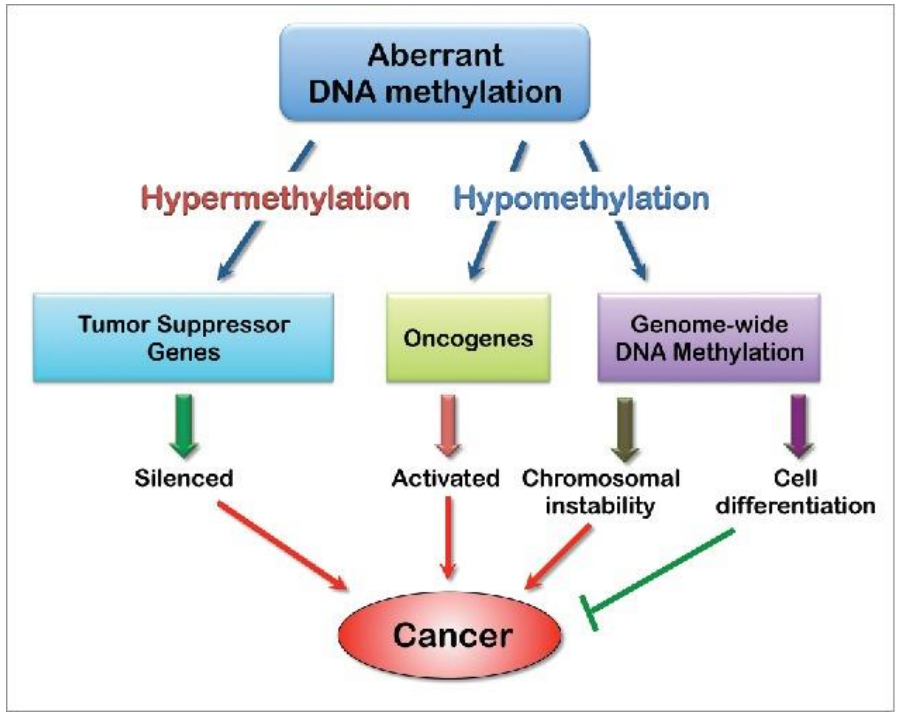
SNVs are unstable

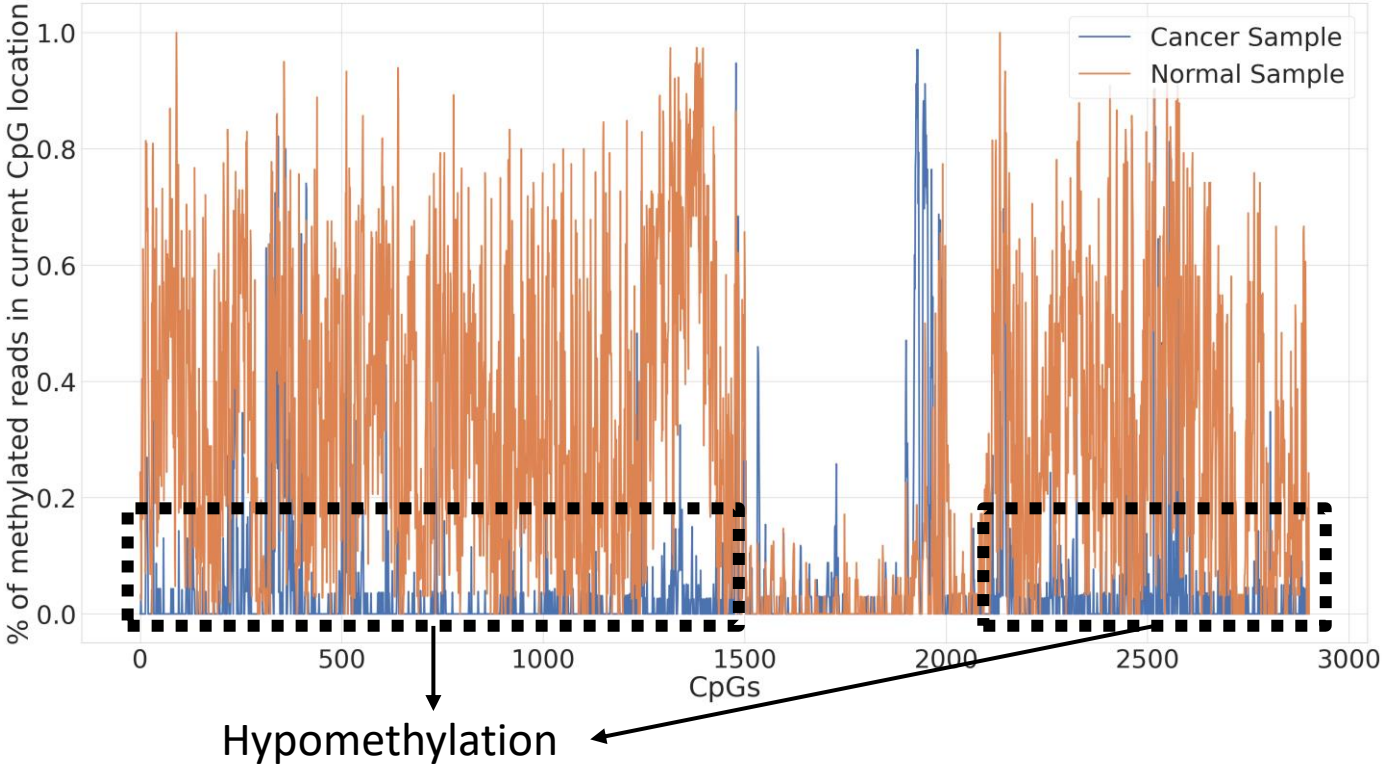


SVs are not extensive in some cancer types

Wei, W., Keogh, M.J., Aryaman, J. et al. Frequency and signature of somatic variants in 1461 human brain exomes. Genet Med 21, 904–912 (2019).
Cosenza MR, Rodriguez-Martin B, Korbel JO. Structural Variation in Cancer: Role, Prevalence, and Mechanisms. Annu Rev Genomics Hum Genet. 2022 Aug 31;23:123-152.

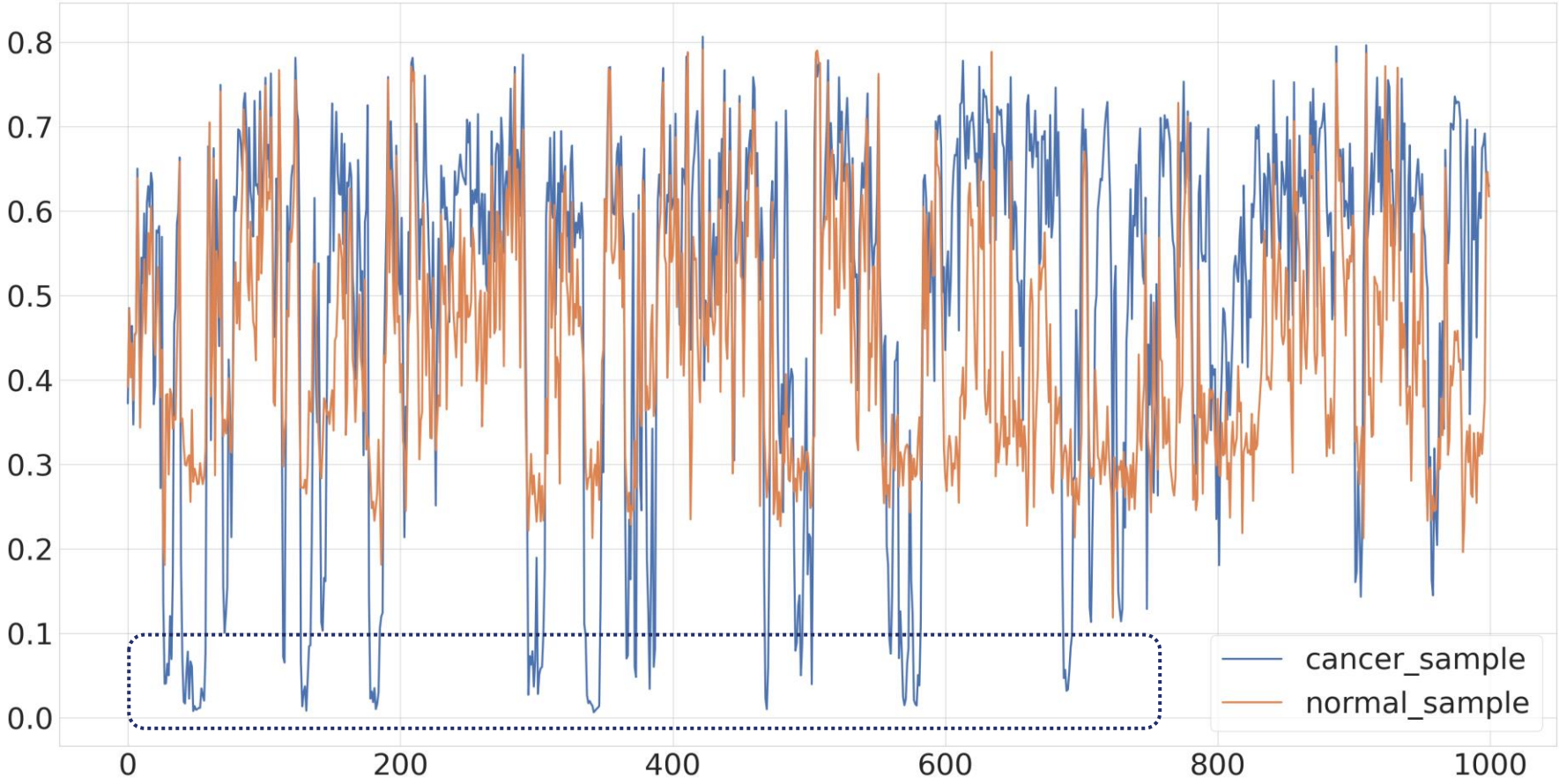
Hypermethylation and Hypomethylation Co-exist in Cancer Cells





Hypomethylation Regions in Tumor Cells

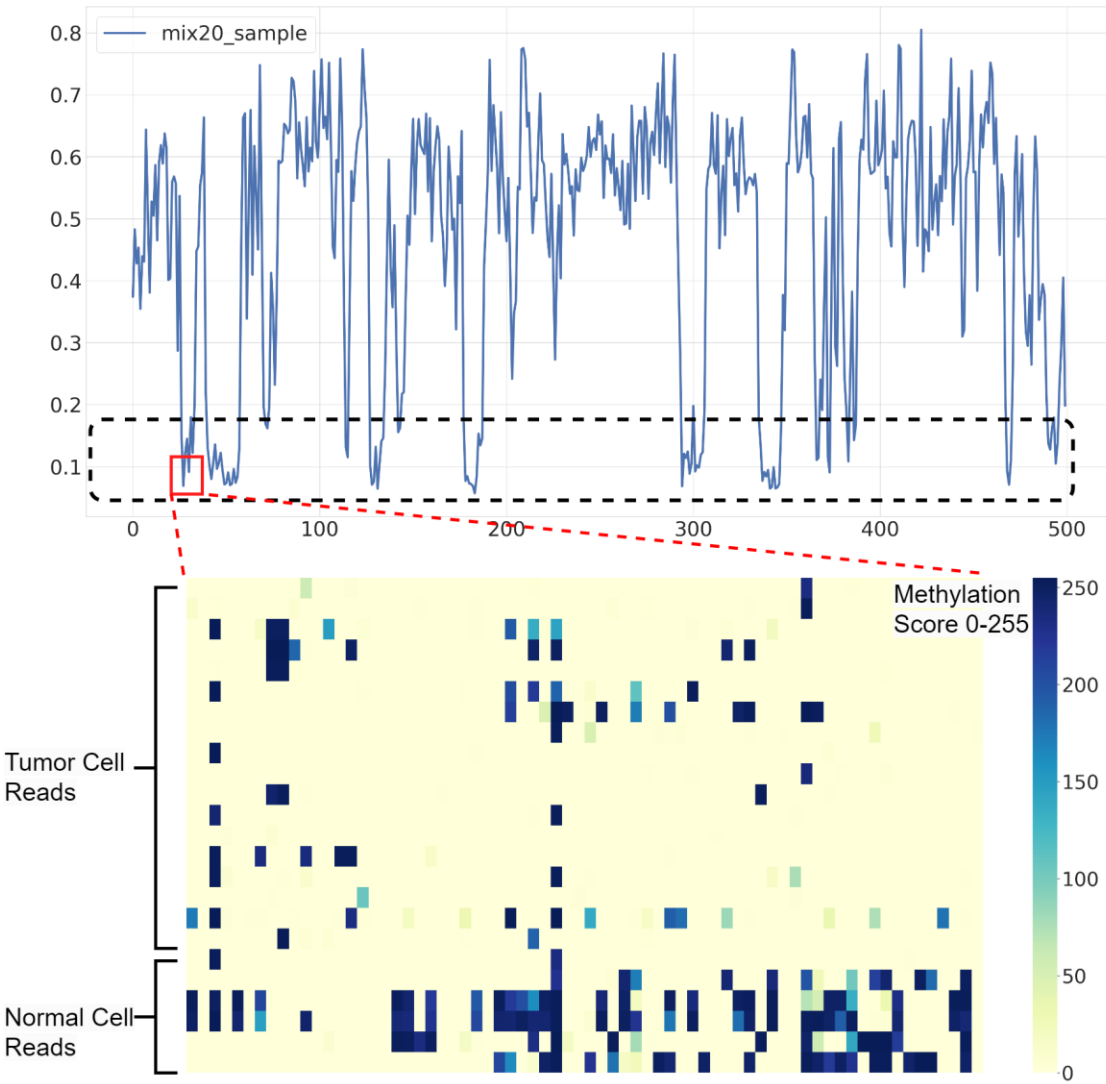
% of methylated reads in current CpG location

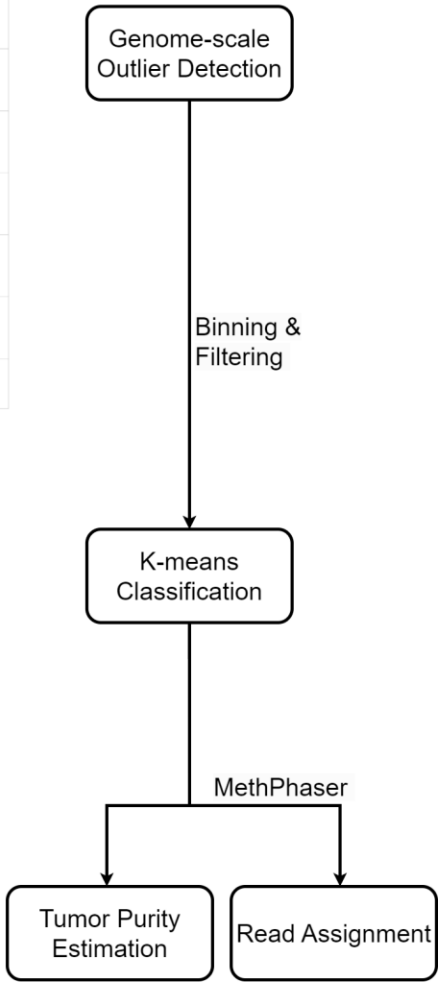


Average methylation level per 100,000 bps, first 100M bps on Chr1

MethPhaser-Cancer Pipeline

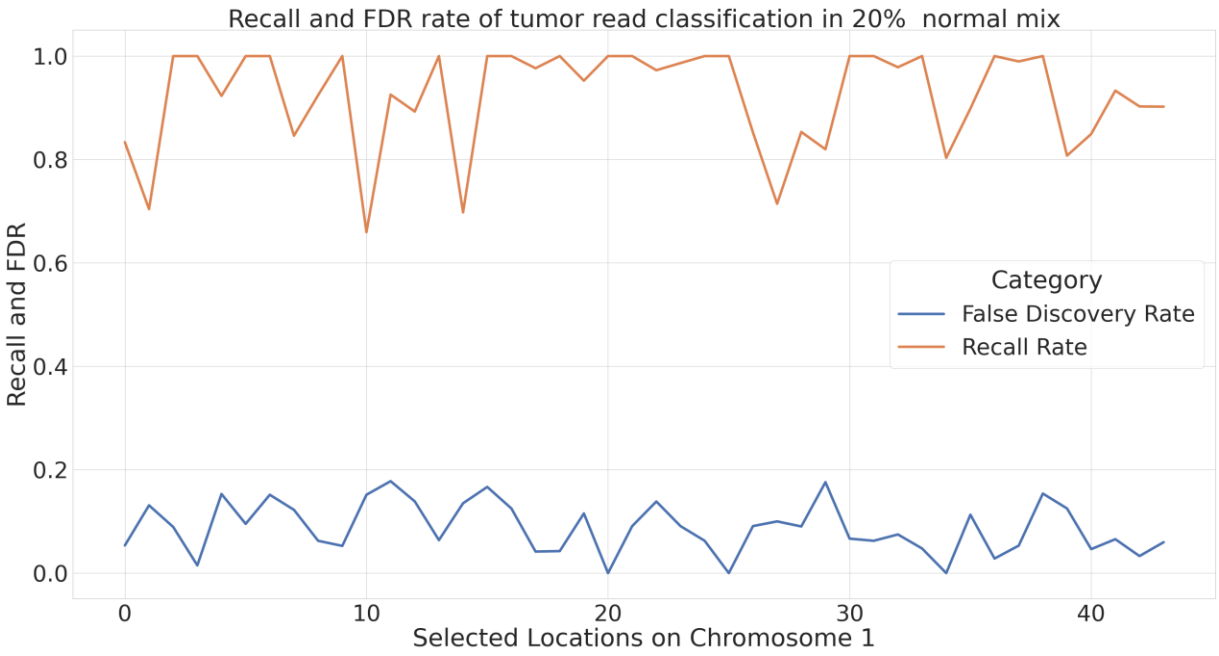
- Simulation with mixing a tumor-normal pair
- Binning the chromosome for outlier (hypomethylation region) detection
- Shrink sparse methylation matrix for k-means classification
- Extending read assignment with MethPhaser algorithm

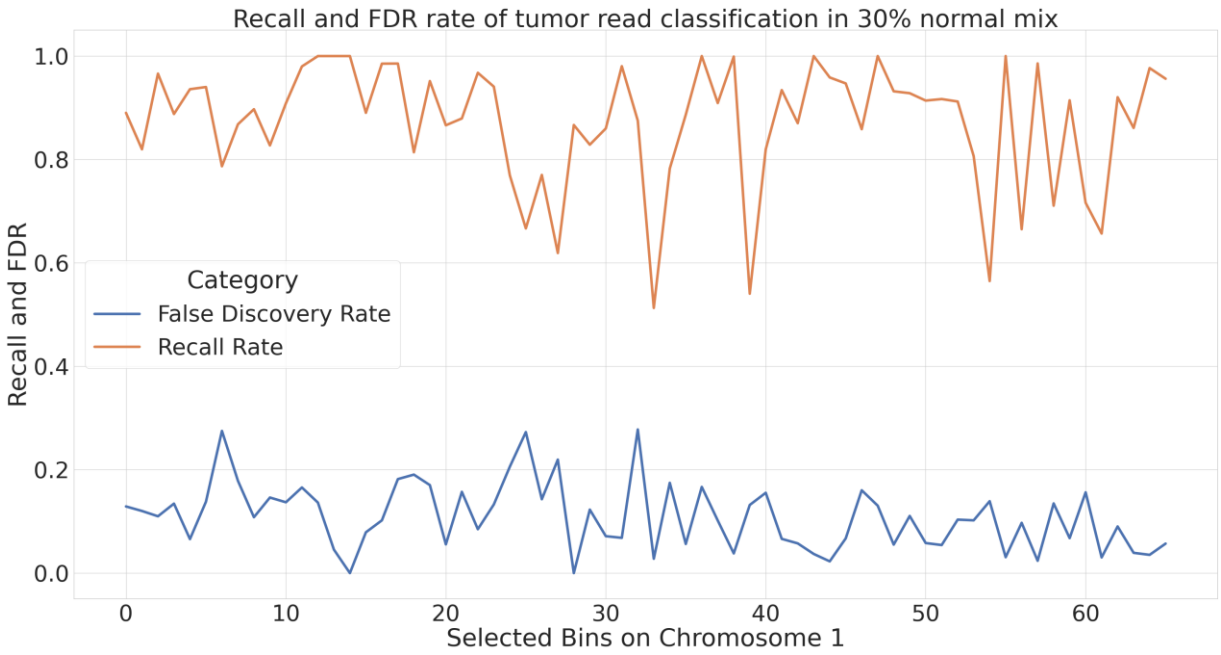




K-means Read Classification

Evaluation dataset: Mixing a tumor-normal pair the knowledge of reads' source

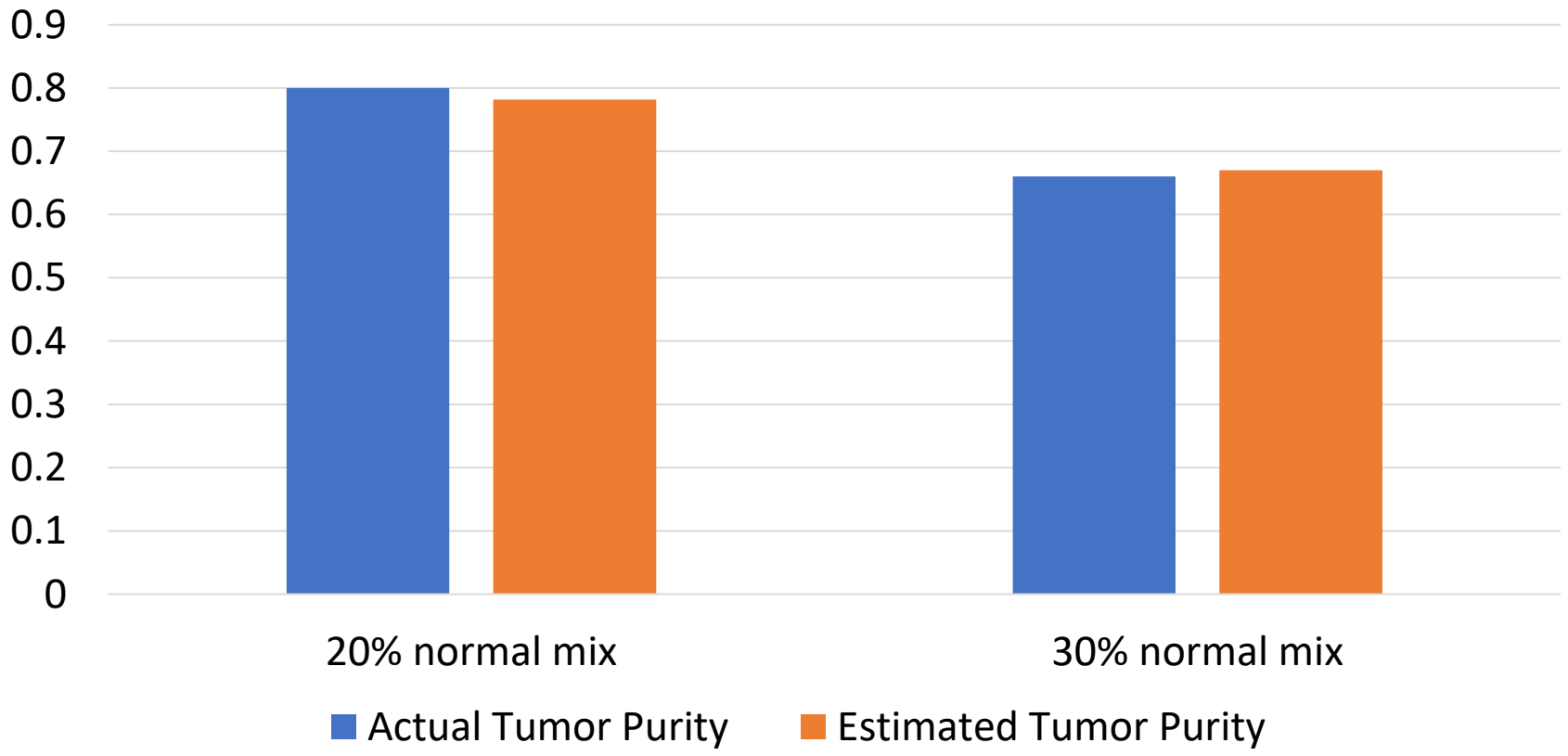




Recall: % of tumor reads classified as tumor reads
False Discovery Rate (FDR): % of normal reads classified as tumor reads

Tumor Purity Estimation

Evaluation dataset: Mixing a tumor-normal pair the knowledge of reads' source



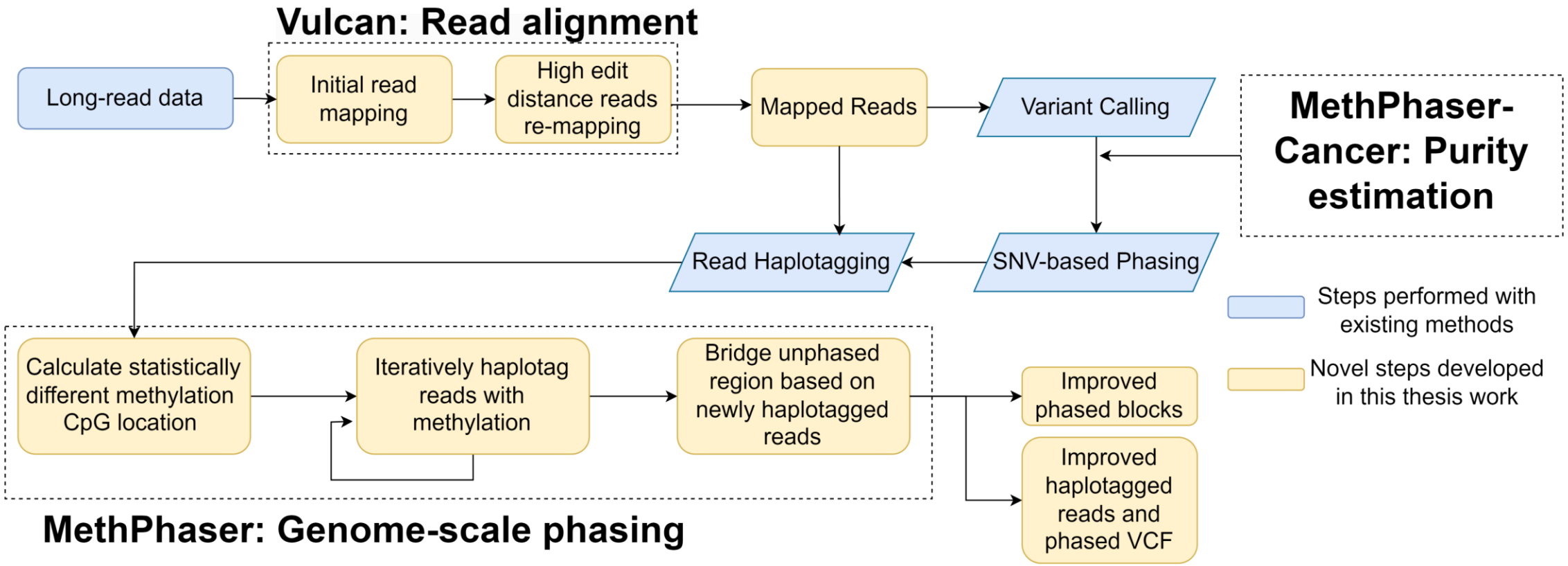
Conclusion

- MethPhaser-Cancer is the first method that automatically estimate the tumor purity with long-read methylation signals
- MethPhaser-Cancer is the first method that can accurately classify long reads in selected regions into two samples
- Future: Analysis on real patient tumor samples

Conclusions

Accurate and Efficient Computational Approaches for Long-read Alignment and Genome Phasing of Human Genomes

Advanced Long-read Analysis Protocol



Publishments

First/Co-first:

- **Yilei Fu**, Medhat Mahmoud, Viginesh Vaibhav Muraliraman, Fritz J Sedlazeck, Todd J Treangen, **Vulcan: Improved long-read mapping and structural variant calling via dual-mode alignment**, *GigaScience*, Volume 10, Issue 9, September 2021, giab063, https://doi.org/10.1093/gigascience/giab063
- **Yilei Fu**, Sergey Aganezov, Medhat Mahmoud, John Beaulaurier, Sissel Juul, Todd J. Treangen, Fritz J Sedlazeck, **MethPhaser: methylation-based haplotype phasing of human genomes**, bioRxiv 2023.05.12.540573; doi: https://doi.org/10.1101/2023.05.12.540573
- Esther G. Lou, **Yilei Fu**, Qi Wang, Todd J. Treangen, Lauren B. Stadler, **Sensitivity and consistency of long- and short-read metagenomics and epicPCR for the detection of antibiotic resistance genes and their bacterial hosts in wastewater**, medRxiv 2023.08.08.23293828; doi: https://doi.org/10.1101/2023.08.08.23293828

Publishments

- Comprehensive analysis and accurate quantification of unintended large gene modifications induced by CRISPR-Cas9 gene editing, *Science Advances*.
- Olivar: fully automated and variant aware primer design for multiplex tiled amplicon sequencing of pathogen genomes, *biorxiv.*
- The third international hackathon for applying insights into large-scale genomic composition to use cases in a wide range of organisms, *F1000Research.*
- KOMB: K-core based de novo characterization of copy number variation in microbiomes, *Computational and Structural Biotechnology Journal.*
- Methods developed during the first National Center for Biotechnology Information Structural Variation Codeathon at Baylor College of Medicine, *F1000Research.*
- The World Ahead: Exploring the Impact of Long-Read Sequencing on Microbiome Analysis, *In prep, submitted to Nature Method.*

Acknowledgements

Committee Members:

Dr. Todd J Treangen (Chair)

Dr. Gang Bao

Dr. Vicky Yao

Dr. Fritz J Sedlazeck

Dr. Huw Ogilvie

Advisor:

Dr. Todd J Treangen

Special Thanks:

Dr. Fritz J Sedlazeck

Funding:

RICE KEN KENNEDY INSTITUTE
AI, Data, and Computing for Global Impact

Collaborators:

Bao lab at Rice BIOE

RICE UNIVERSITY
Department of Bioengineering

Sedlazeck lab at BCM HGSC

Baylor College of Medicine
HGSC
HUMAN GENOME SEQUENCING CENTER

Stadler lab at Rice CEE

RICE UNIVERSITY
Civil and Environmental Engineering

Oxford Nanopore Technology

Oxford NANOPORE Technologies

Treangen Lab Members:

Dr. Michael Nute

Dr. Advait Balaji

Dr. Qi Wang

Kristen Curry

Eddy Huang

Bryce Kille

Natalie Kokroko

Yunxi Liu

Felix Quintana

Nicolae Sapoval

Michael Wang

R. Matt Barnett

Family & Friends

Questions

Thanks for listening!

Long-read Technology Brings Global Methylation Catalog

	Microarray	Whole Genome Bisulfite Sequencing	Long-read Sequencing
Sequencing region	Selected regions	Whole genome	Whole genome
Read length	Only report methylation	Short read mostly	Long-reads
Difficulty	Design probe	Bisulfite treatment	N/A
Price	$	$$	$$

HG002

GM26105 iPSC from B-Lymphocyte

Description: PERSONAL GENOME PROJECT
Affected: Unknown
Sex: Male
Age: 45 YR (At Sampling)

Overview | Characterizations | Phenotypic Data | Publications | Culture Protocols

Repository	NIGMS Human Genetic Cell Repository
Subcollection	Apparently healthy iPSCs Apparently Healthy Collection PIGI Consented Sample
Protocols	Protocol PDF
Cell Type	Stem cell
Cell Subtype	Induced pluripotent stem cell
Transformant	Reprogrammed (Episomal)
Sample Source	iPSC from B-Lymphocyte
Race	White
Family Member	1
Family History	N
Relation to Proband	proband
ISCN	46,XY[24].arr(1-22)x2,(X,Y)x1
Species	Homo sapiens
Common Name	Human

Tumor-normal Pair

COLO 829

CRL-1974 ™

99/100 Bioz Stars 215 Product Citations

Product category	Human cells
Organism	*Homo sapiens*, human
Cell type	fibroblast
Morphology	fibroblast
Tissue	Skin
Disease	Melanoma
Applications	3D cell culture
Product format	Frozen
Storage conditions	Vapor phase of liquid nitrogen

COLO 829BL

CRL-1980 ™

COLO 829BL is a B lymphoblast cell line that was isolated from the peripheral blood of a 45-year-old White male. The transformed B cells are from the same patient as the COLO 829 malignant melanoma cell line (ATCC CRL-1974) and as such offers a paired human normal and tumor cell line for comparative studies. This cell line was deposited by G.E. Moore. It can be used in cancer and immunology research.

99/100 Bioz Stars 248 Product Citations

Product category	Human cells
Organism	*Homo sapiens*, human
Cell type	B lymphoblast
Morphology	lymphoblast
Tissue	Peripheral blood
Disease	Normal
Applications	3D cell culture Immunology
Product format	Frozen
Storage conditions	Vapor phase of liquid nitrogen

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2873040/pdf/nihms171925.pdf

Box 1

A summary of some of the salient spatial–temporal features of cancer DNA epigenetics

At the level of several hundred base pairs: a high degree of site-to-site dependence

- A neighboring-sites model rather that an independent-sites or context-dependent model best describes methylation changes in regions that are not under strong cancer-selection pressure.
- Sometimes hypomethylated and hypermethylated CpG dyads are neighbors.
- Some CpG dyads persist as hemimethylated sites.

At the level of genes, gene clusters or other large regions: long-range coupling

- Selection pressures during tumorigenesis help shape DNA methylation patterns, for example, homogeneous hypermethylation of promoters of genes whose silencing facilitates cancer formation.
- There are regions of long-range epigenetic changes containing either:
  - –Clusters of hypermethylated CpG islands;
  - –Pockets of hypomethylation; or
  - –Long tandem repeats with overall hyper- or hypomethylation.

Snapshots of a specific time & place

- The detection of cancer-linked DNA methylation changes depends on: the method of analysis and choice of DNA sequences to be analyzed, the tumor specimen, tumor subregion, amount of ‘contaminating’ normal cells, the stage of carcinogenesis and the use of an appropriate normal tissue for comparison.
- Usually only DNA sequences with an intermediate level of methylation in normal tissues will reveal both hypo- and hyper-methylation.