UNIVERSITY OF MINNESOTA
Driven to Discover℠

Long-Read Sequencing: A New Frontier for Estimating Biomarkers

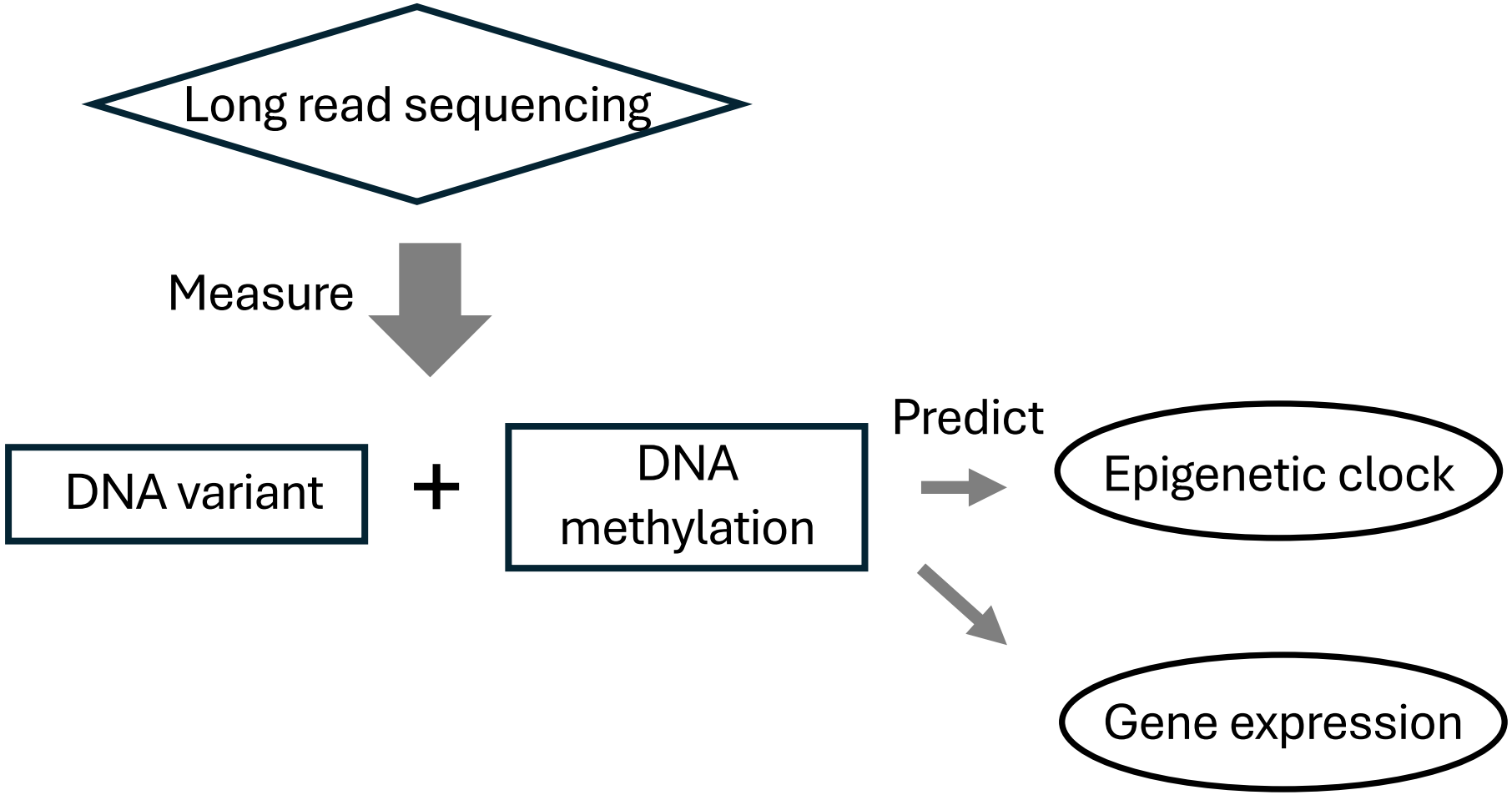
Shijia Zhu, Ph.D
Assistant Professor
Laboratory Medicine and Pathology
University of Minnesota

Disclosure Risk Review:

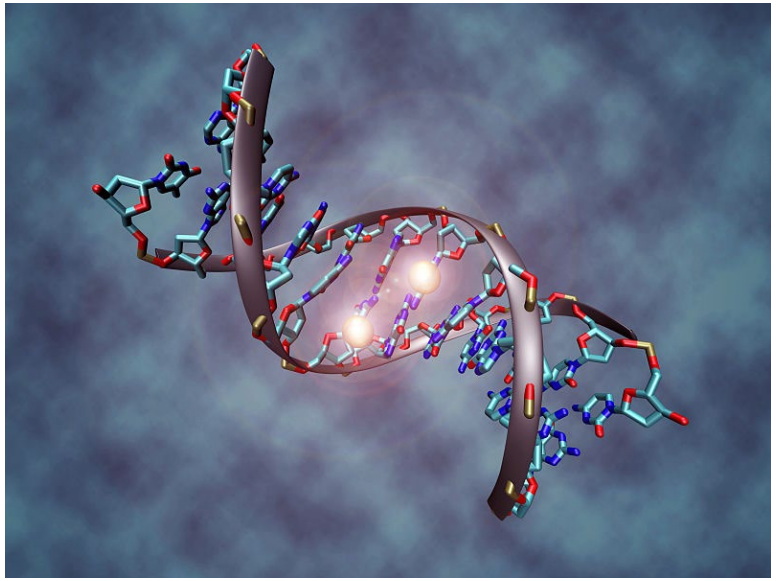
License Number: 21090019

No disclosure risks were identified

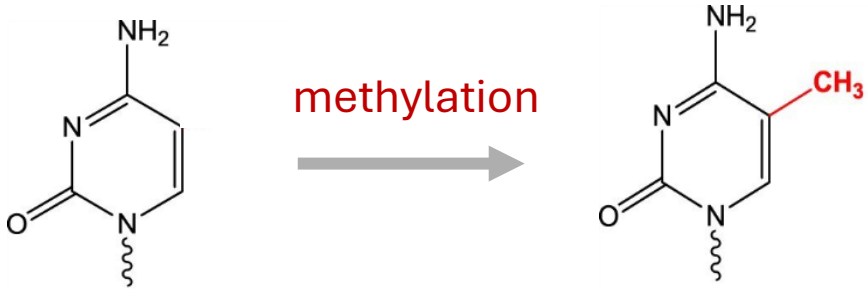
Outlines


Long read sequencing
Measure
DNA variant
+
DNA methylation
Predict
Epigenetic clock
Gene expression

DNA sequences
G, A, T, C



Chemical modifications to DNA



C

5-mC

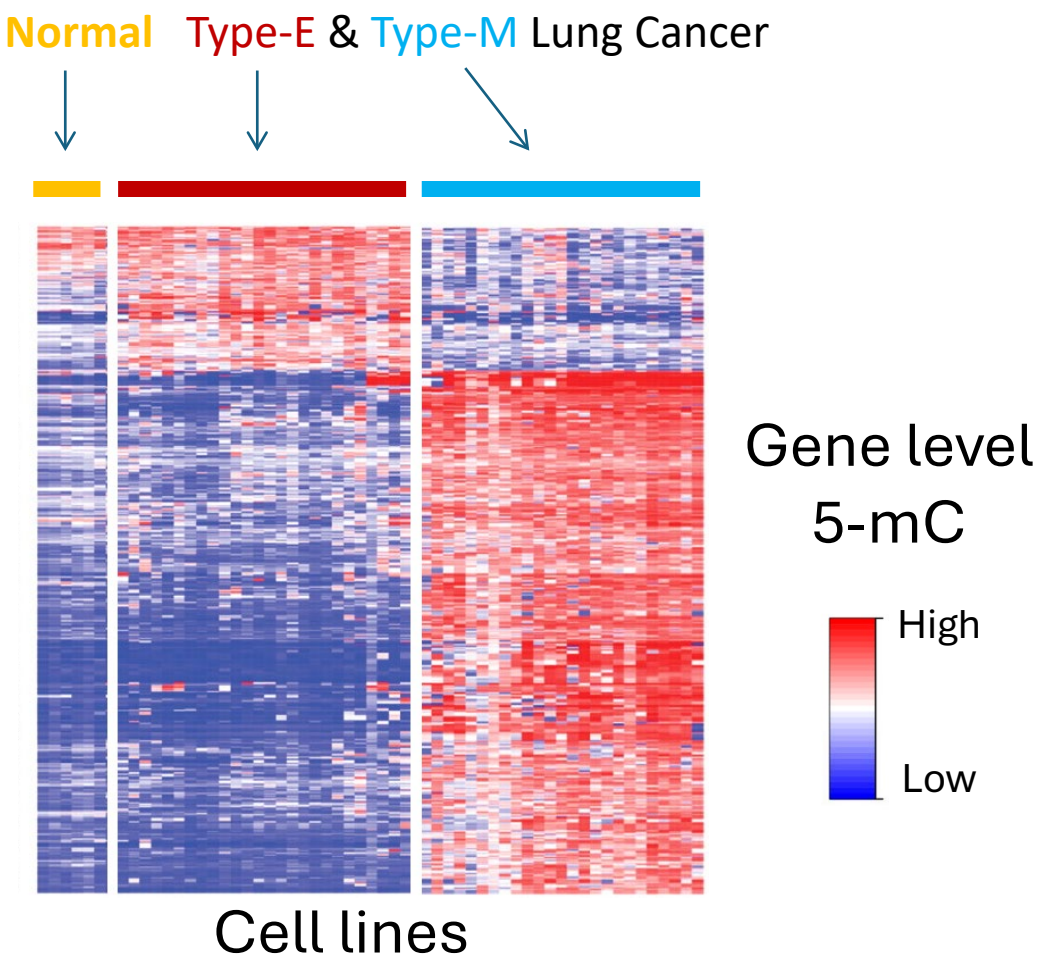
Extensively studied 5-mC

Development

Gene Silencing

© W. P. Armstrong 2000

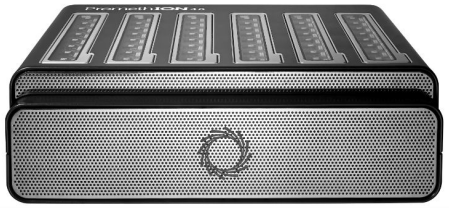
Diseases



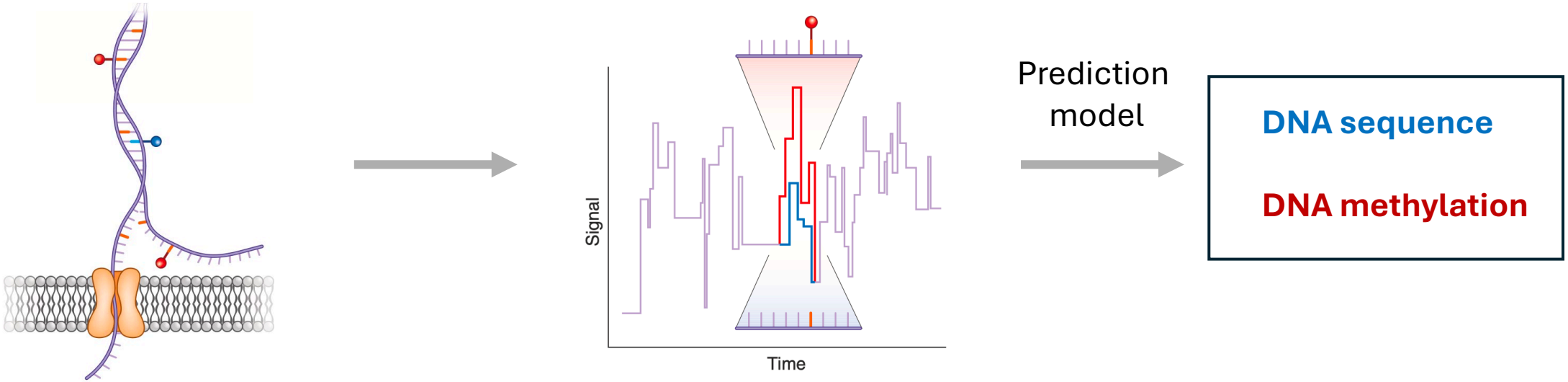
Walter et al. CCR, 2012

3rd Gen Sequencing:
Simultaneous profiling of DNA sequence and methylation

PromethION machine



- **Price:** comparable to Illumina sequencing
- **Throughput:** up to 290 Gb per flow cell
- **Read lengths:** up to >4Mb



(Michael Schatz, Nat Methods, 2017)

Datasets (Long read + EPIC + Gene expression)

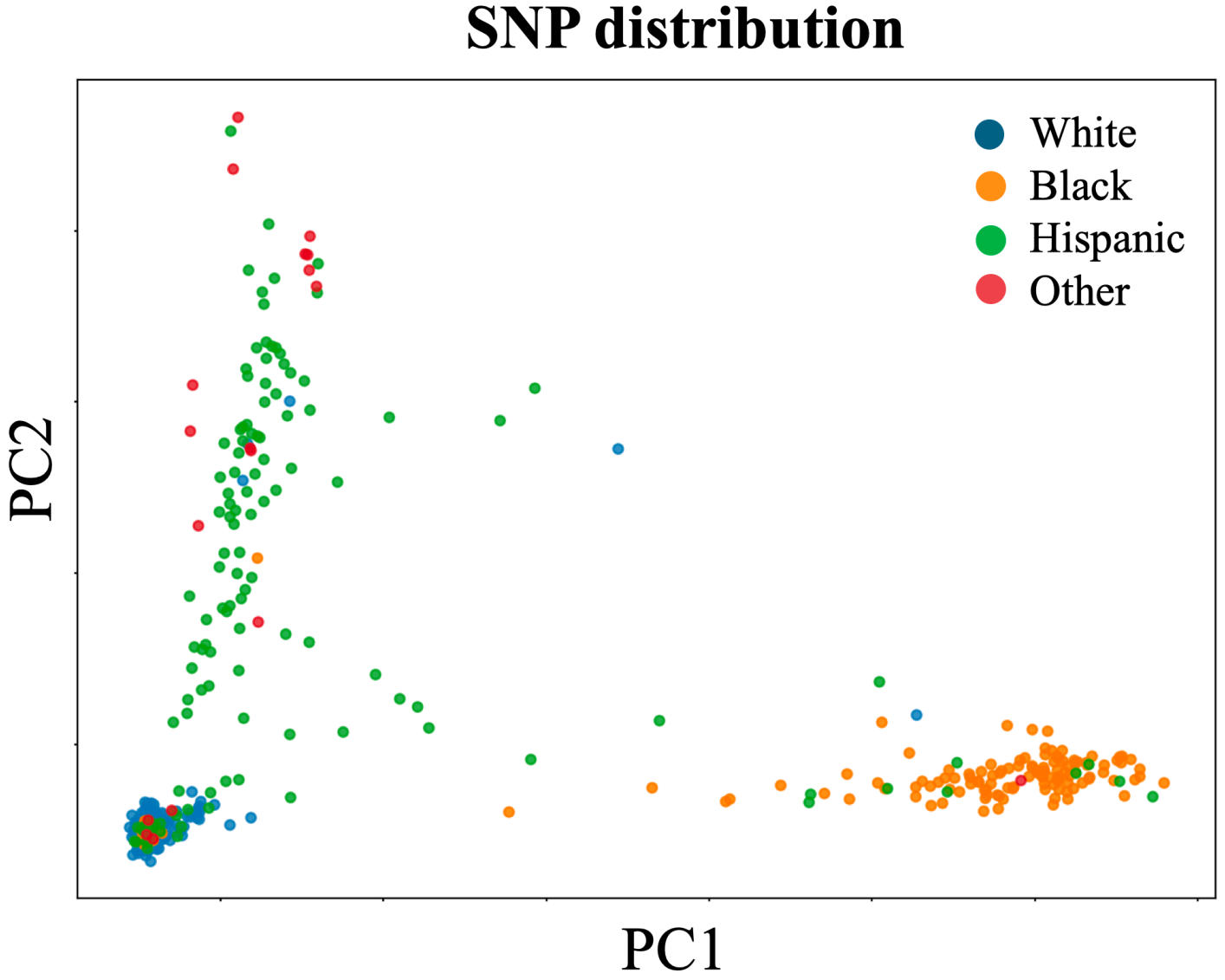
- **Human Retirement Study (HRS):**
  - Nanopore long read data (50 samples)
  - Illumina EPIC v1 data (50 samples + 4,000 samples)
  - Gene expression (50 samples + 4,000 samples)

- **High School and Beyond (HS&B:80):**
  - Nanopore long read data (510 samples)
  - Illumina EPIC v2 data (510 samples)

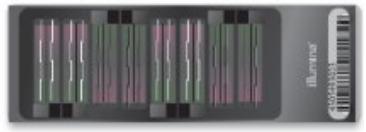
LR variant calling (HS&B:80)

Variant	Subtype	Count
SNP	Indel	2,622,021
	SNV	575,719
SV	Insertion	14,081
	Deletion	9,948
	Duplication	6080

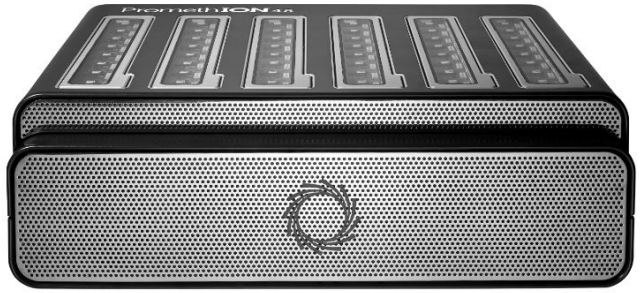
(epi2me/wf-human-variation)



Illumina MethylationEPIC BeadChip



Nanopore PromethION



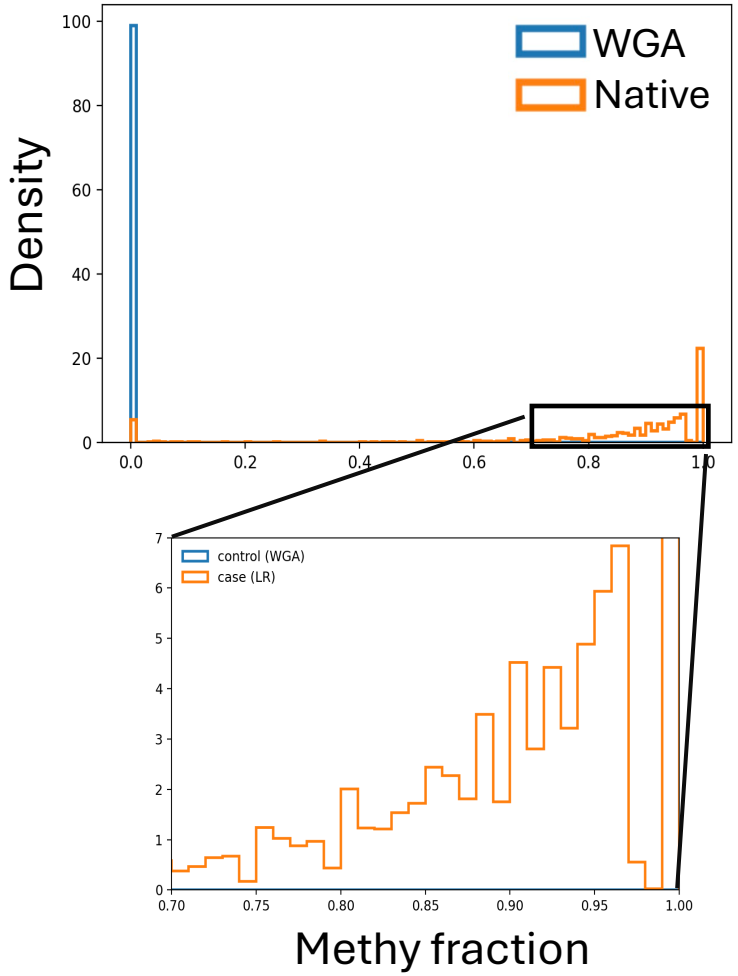
	Illumina methyaltionEPIC v1	Nanopore long read
Detection method	Bisulfite conversion	Ionic current
Information	Only methylation	WGS + methylation
Throughput	>850k (annotated enhancer, promoter, etc)	Whole methylome
Methylation type	5mC	5mC, 5hmC, 6mA

How accurate is the long-read DNA methylation calling?

Accuracy (1): LR prediction model

- **WGA: negative control with no methylation**

5mC



6mA

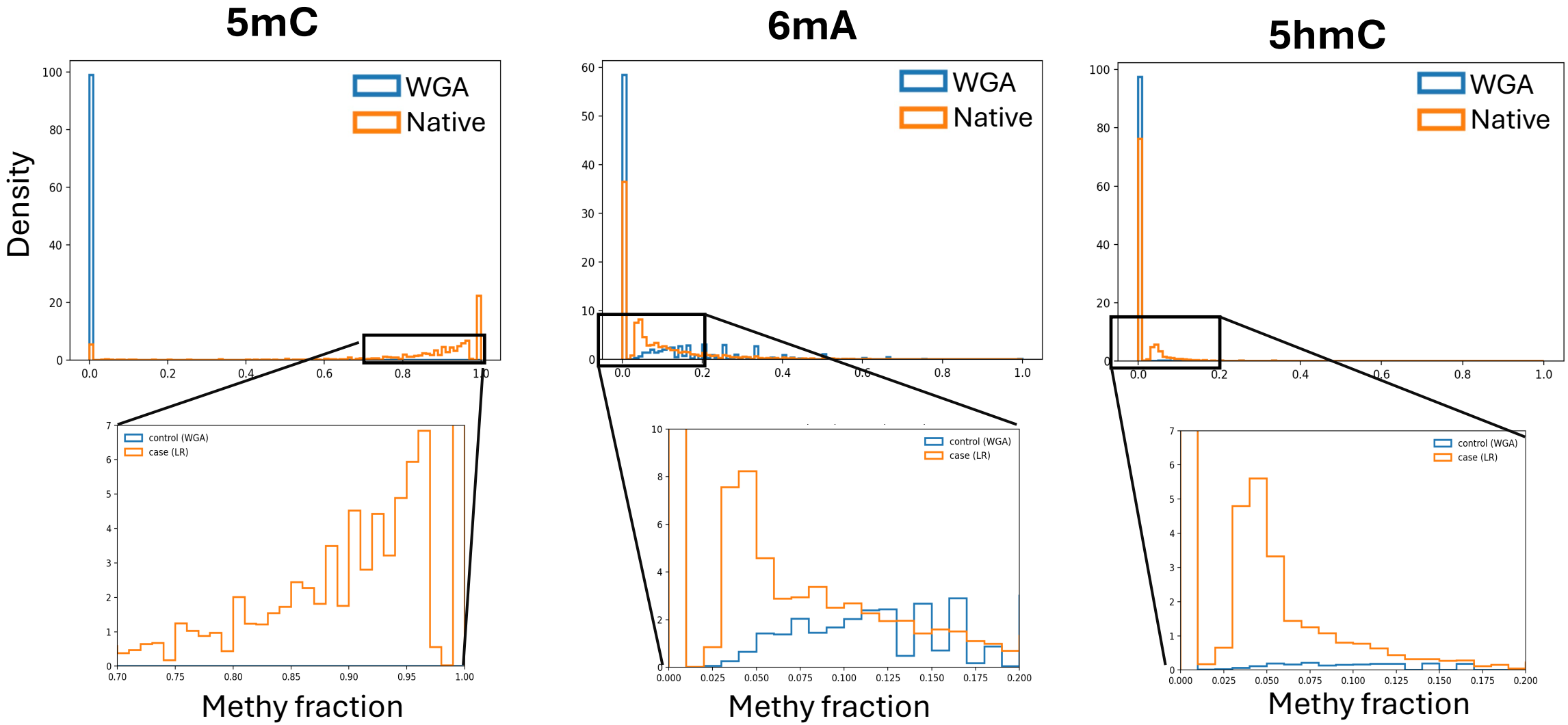
Existence of 6mA in multicellular organisms, including human, is still controversial

5hmC

Low accuracy of 5hmC detection

Accuracy (1): LR prediction model

- **WGA: negative control with no methylation**



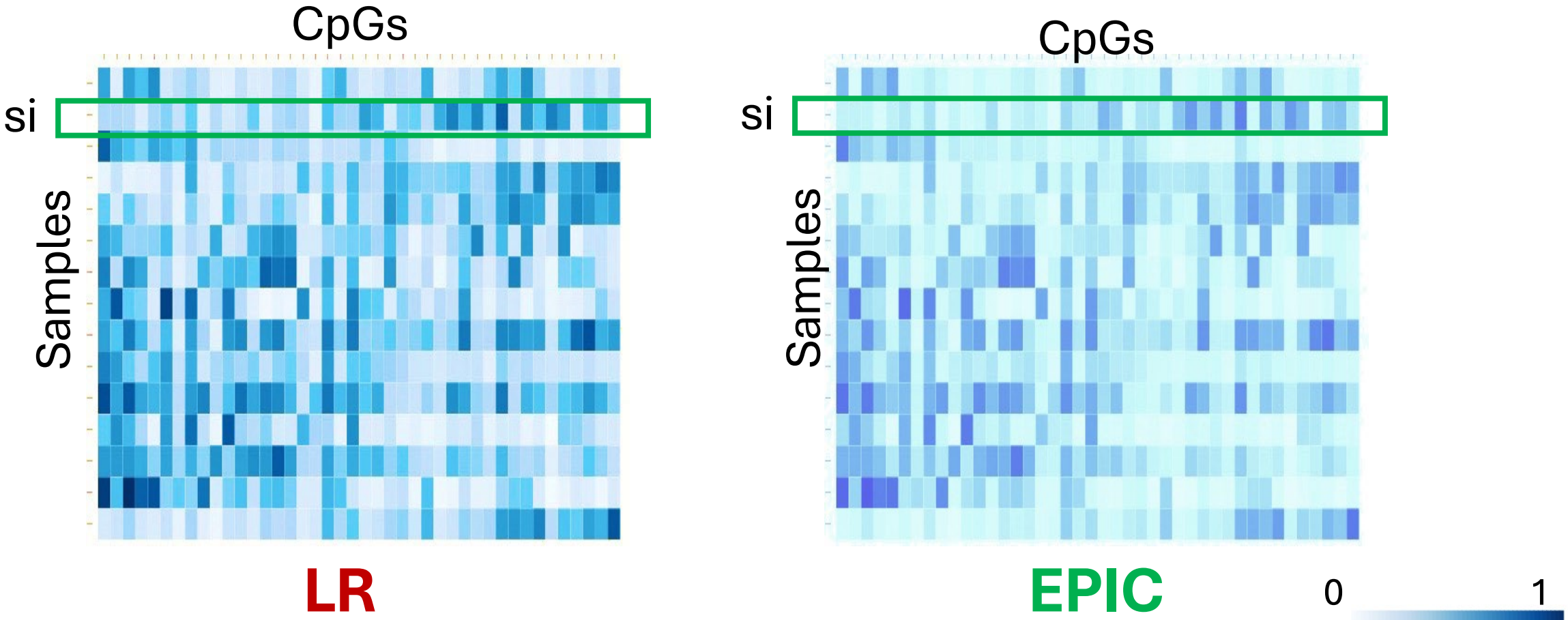
Accuracy (2): Consistency between platforms

Datasets (Long read + EPIC + Gene expression)

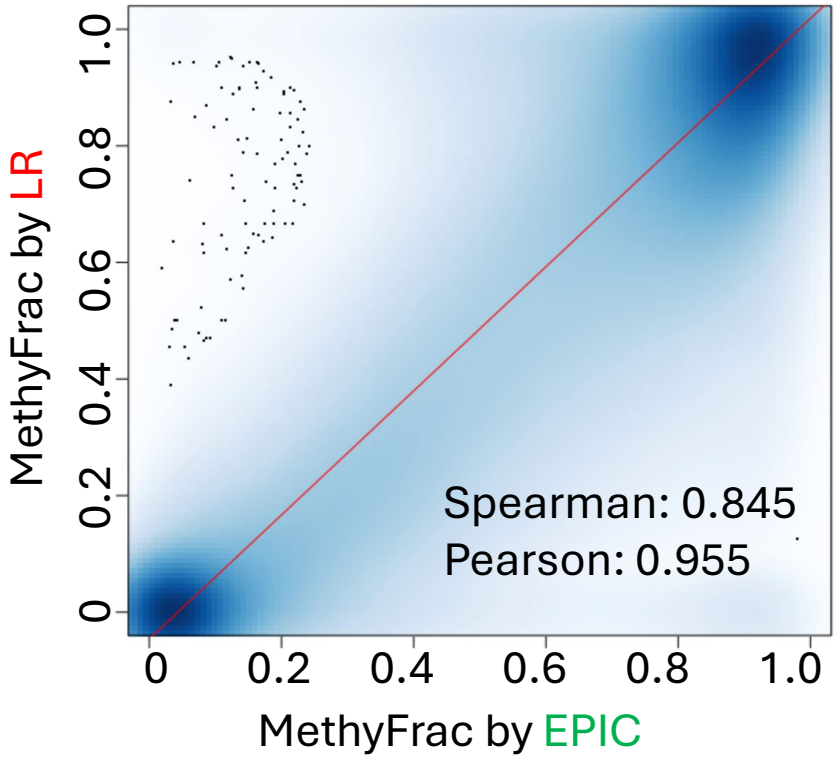
- **Human Retirement Study (HRS):**
  - Nanopore long read data (50 samples)
  - Illumina EPIC v1 data (50 samples + 4,000 samples)
  - Gene expression (50 samples + 4,000 samples)

- **High School and Beyond (HS&B:80):**
  - Nanopore long read data (510 samples)
  - Illumina EPIC v2 data (510 samples)

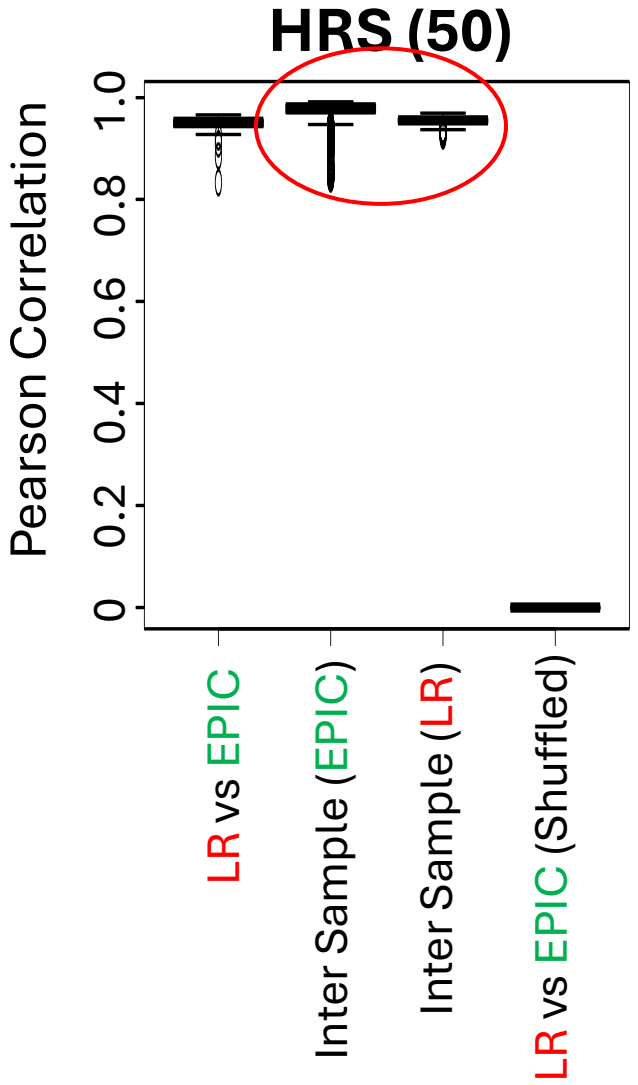
5mC platform comparison (Sample-level)



5mC platform comparison (Sample-level)

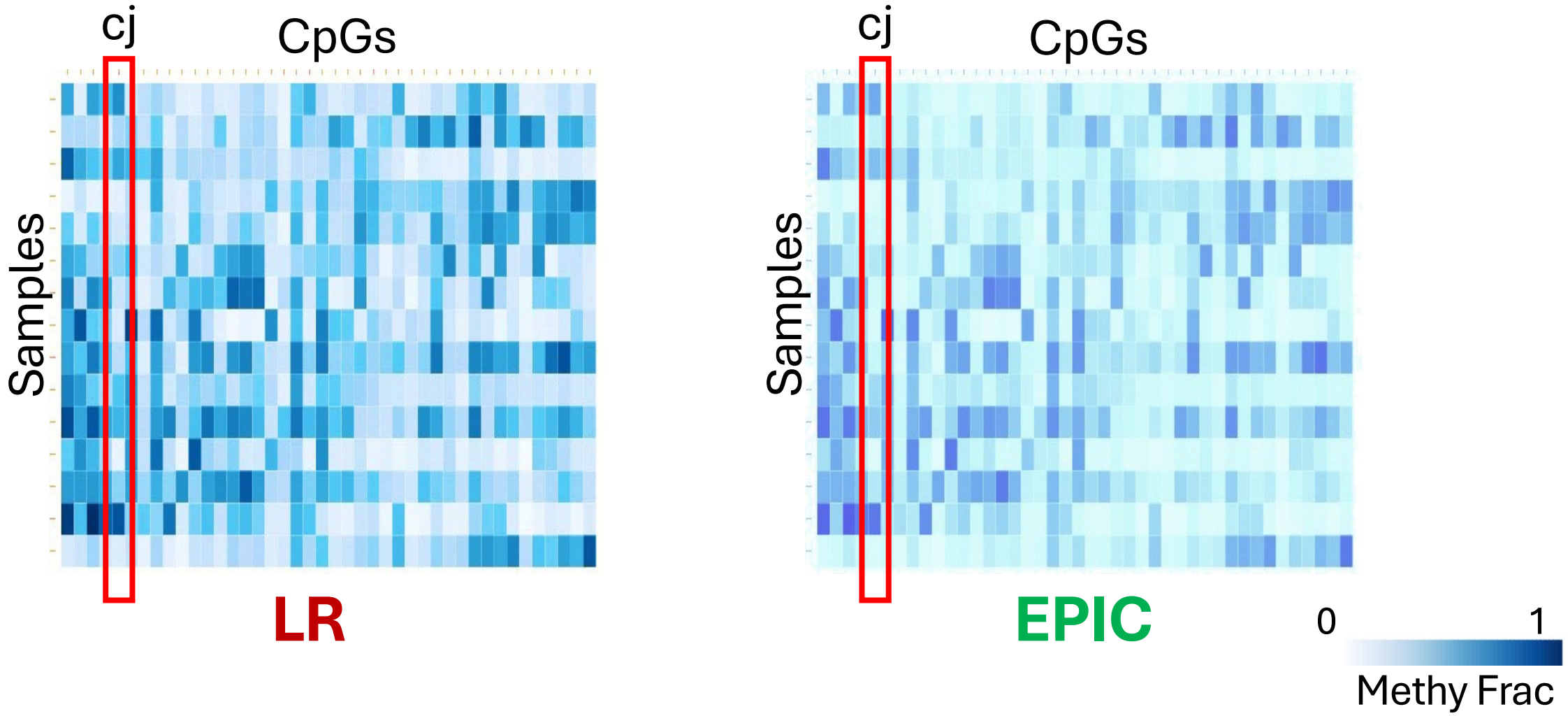


~830k overlap between LR and EPIC



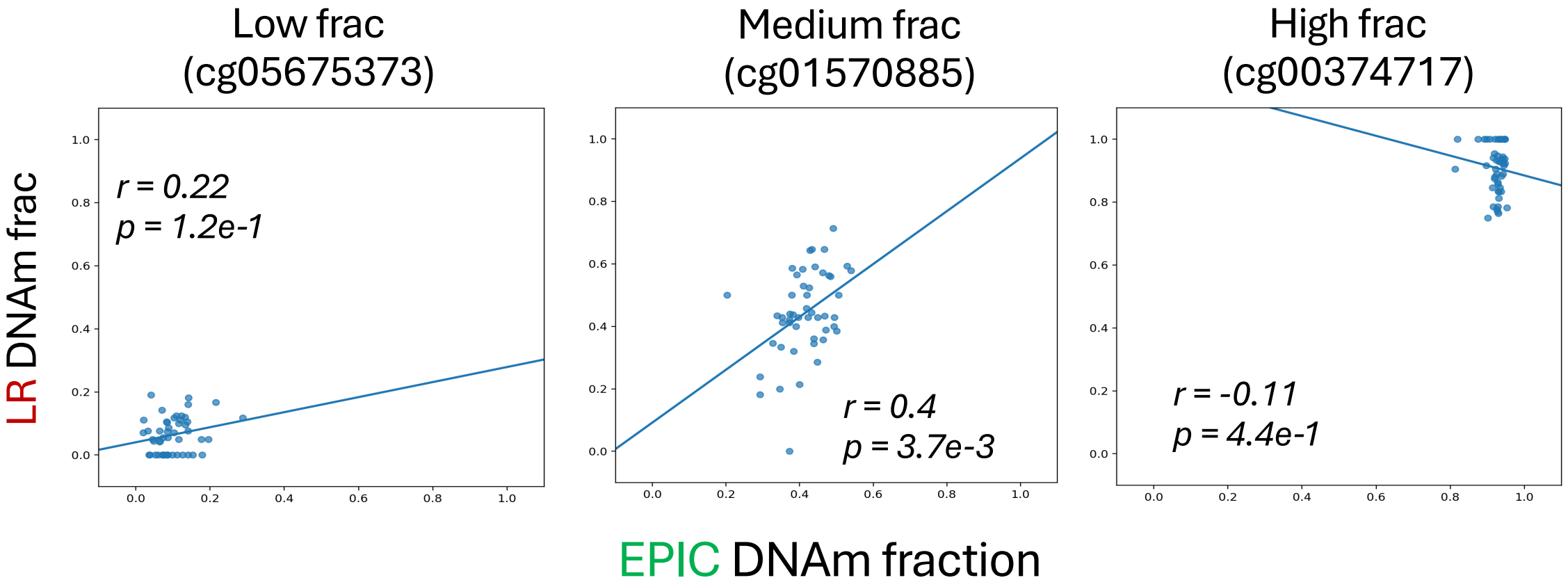
Correlation of each sample Meth fraction between LR and EPIC across all CpG sites

5mC platform comparison (CpG-level)

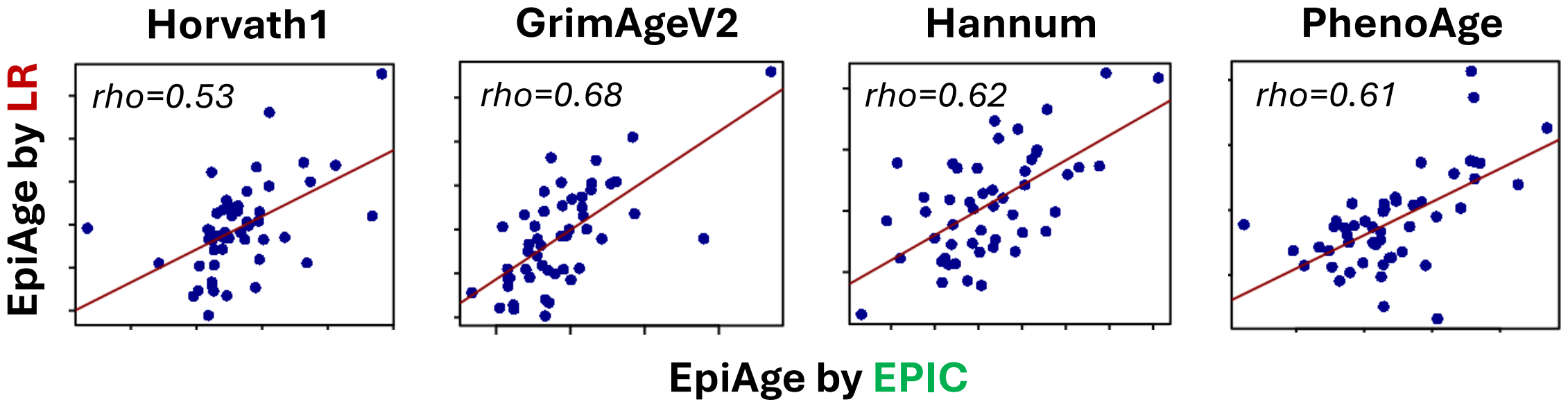


5mC platform comparison (CpG-level)

- Three representative CpGs from Horvath clock

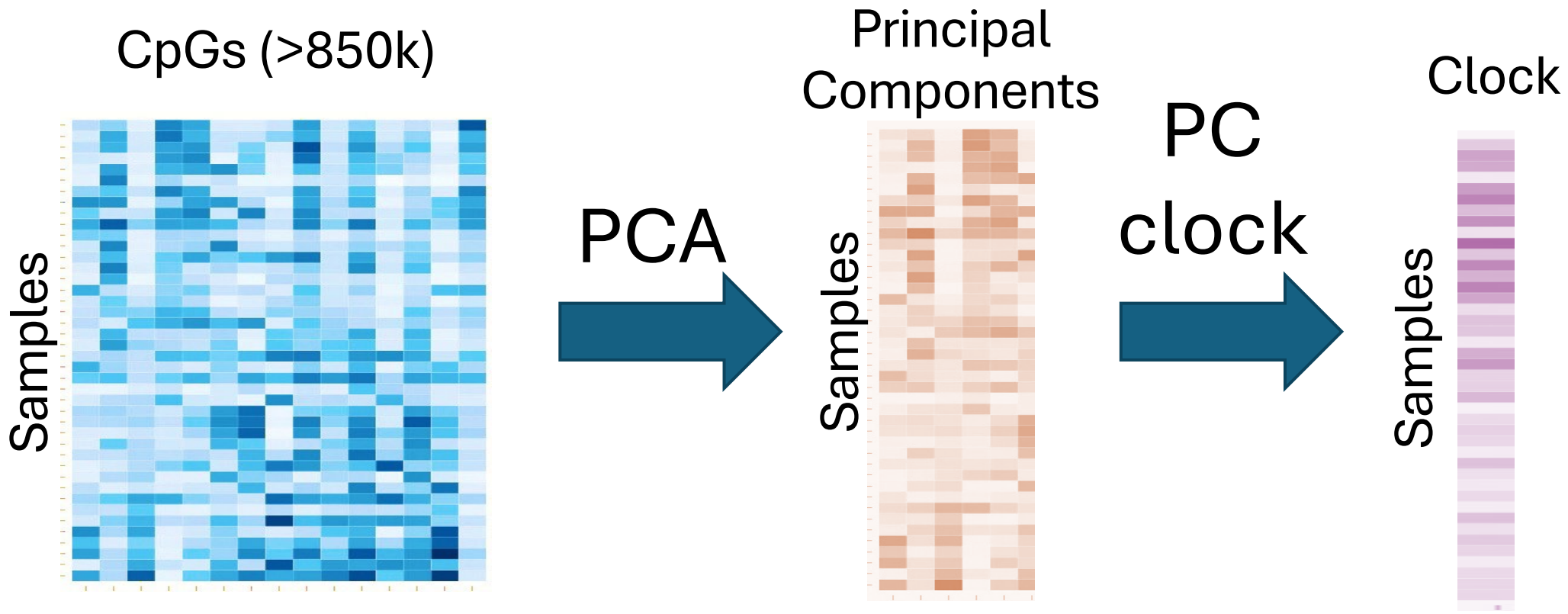


CpG-level inconsistency impacts the downstream analysis



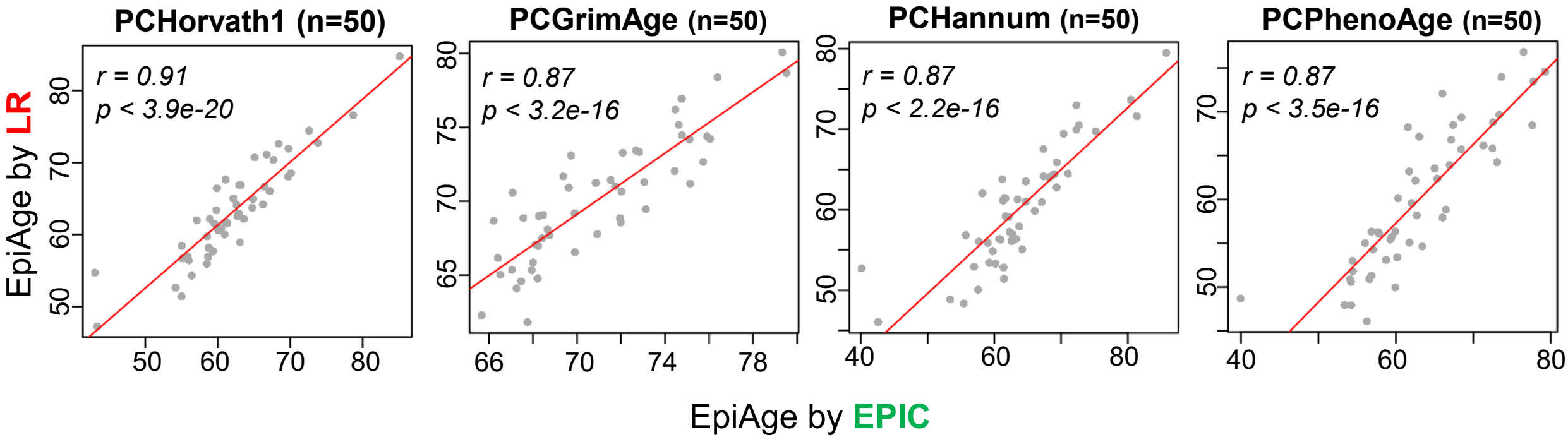
How to improve the consistency?

PC clock improves the consistency

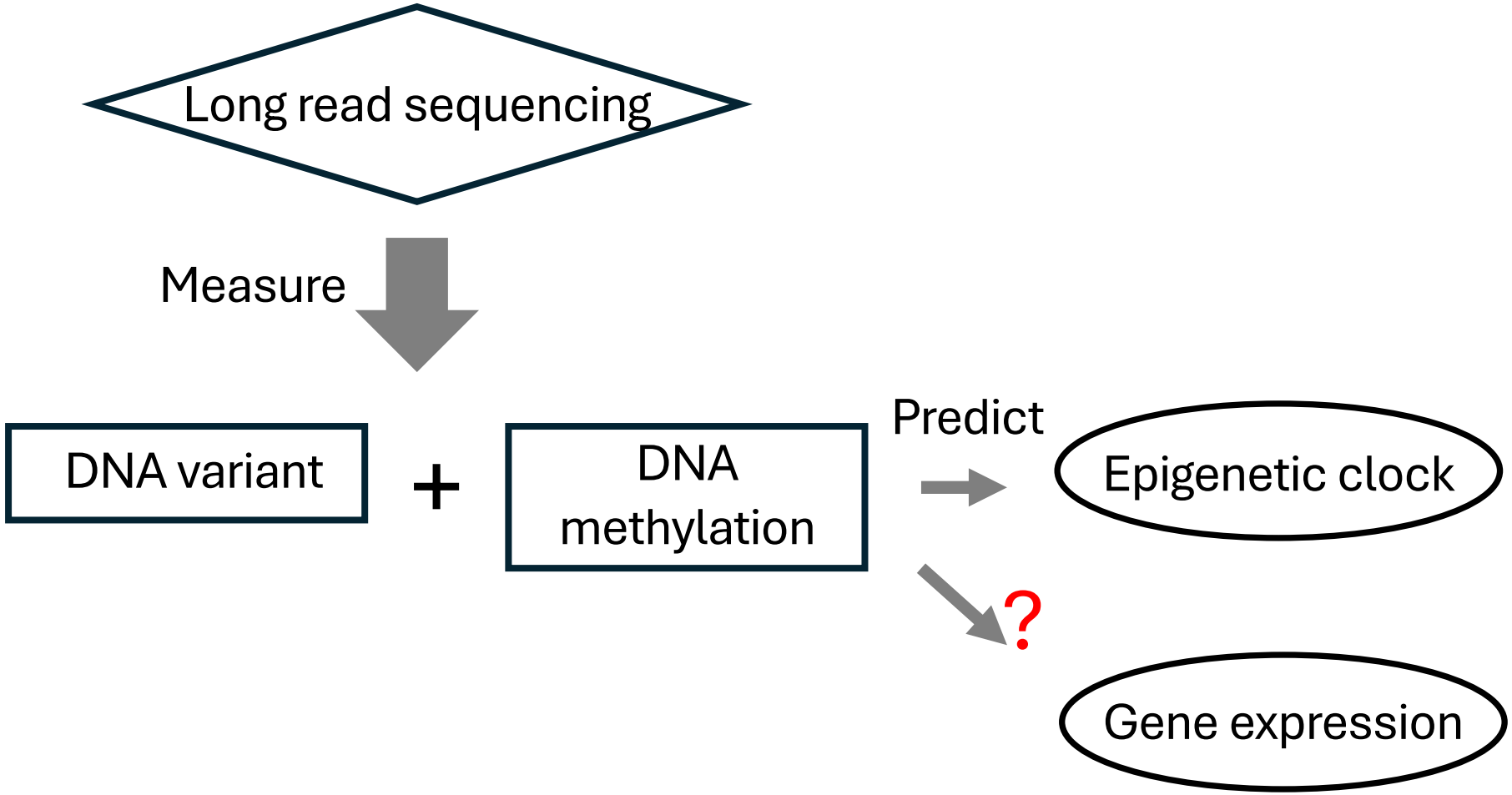


(Higgins-Chen, et al., Nature Aging, 2022)

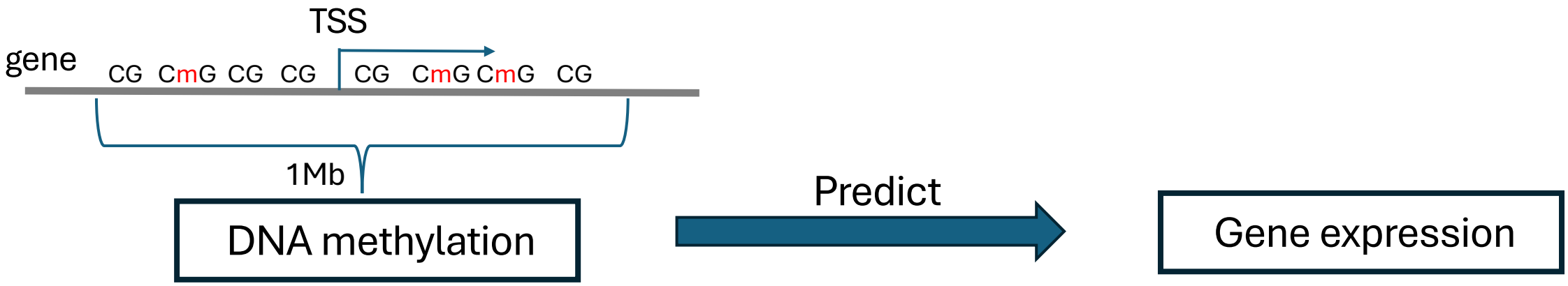
PC clock improves the consistency



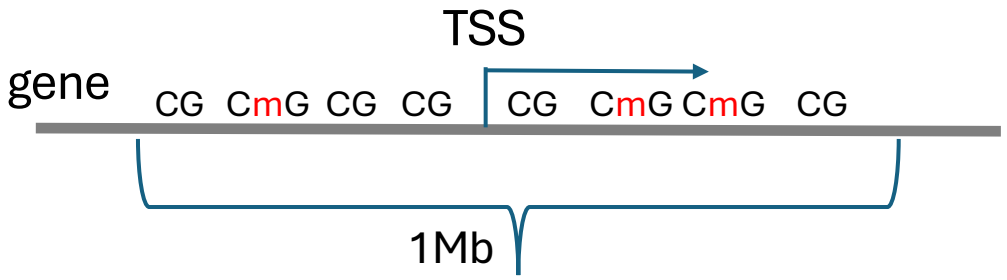
Summary (1)


Long read sequencing
Measure
DNA variant
+
DNA methylation
Predict
Epigenetic clock
?
Gene expression

Transcriptomics prediction



Transcriptomics prediction



DNA methylation

PCA

DNA methylation
Principal components

Predict

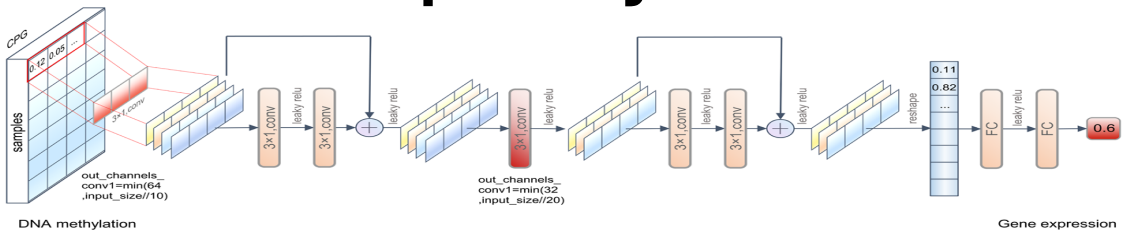
Gene expression

Training dataset

(~4,000 HRS samples)

- **Exposure**: DNAm by EPIC
- **Outcome**: Gene expr by RNAseq

DeepMethyGene

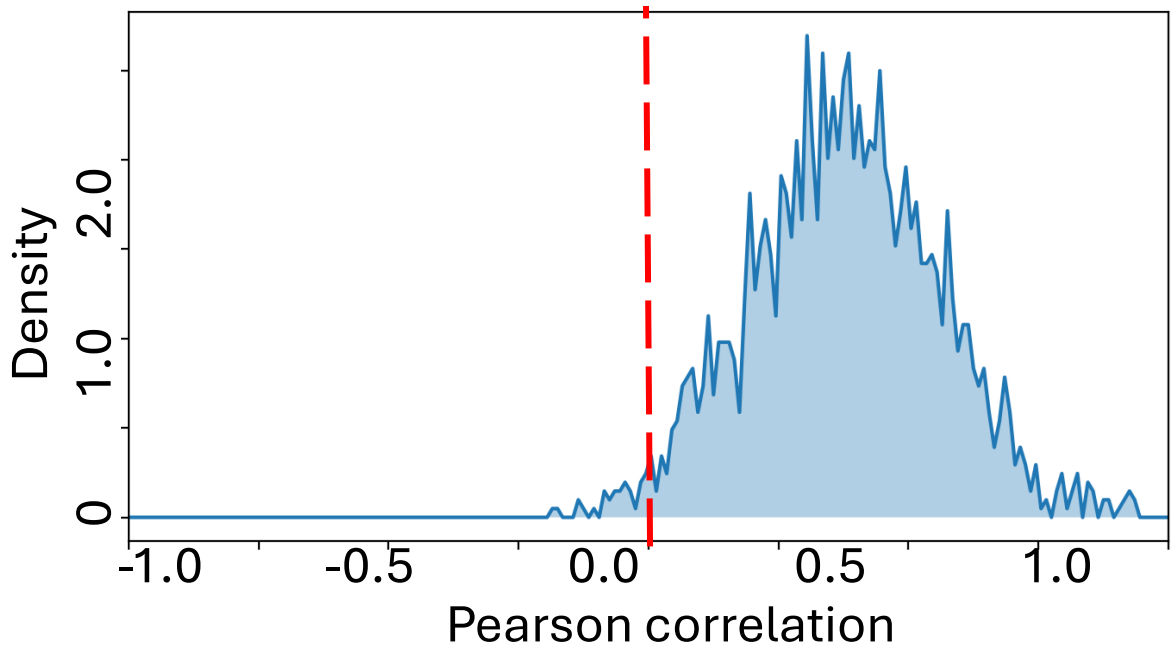


An adaptive recursive convolutional neural network model (Yan, et al., BMC Bioinformatics, 2025)

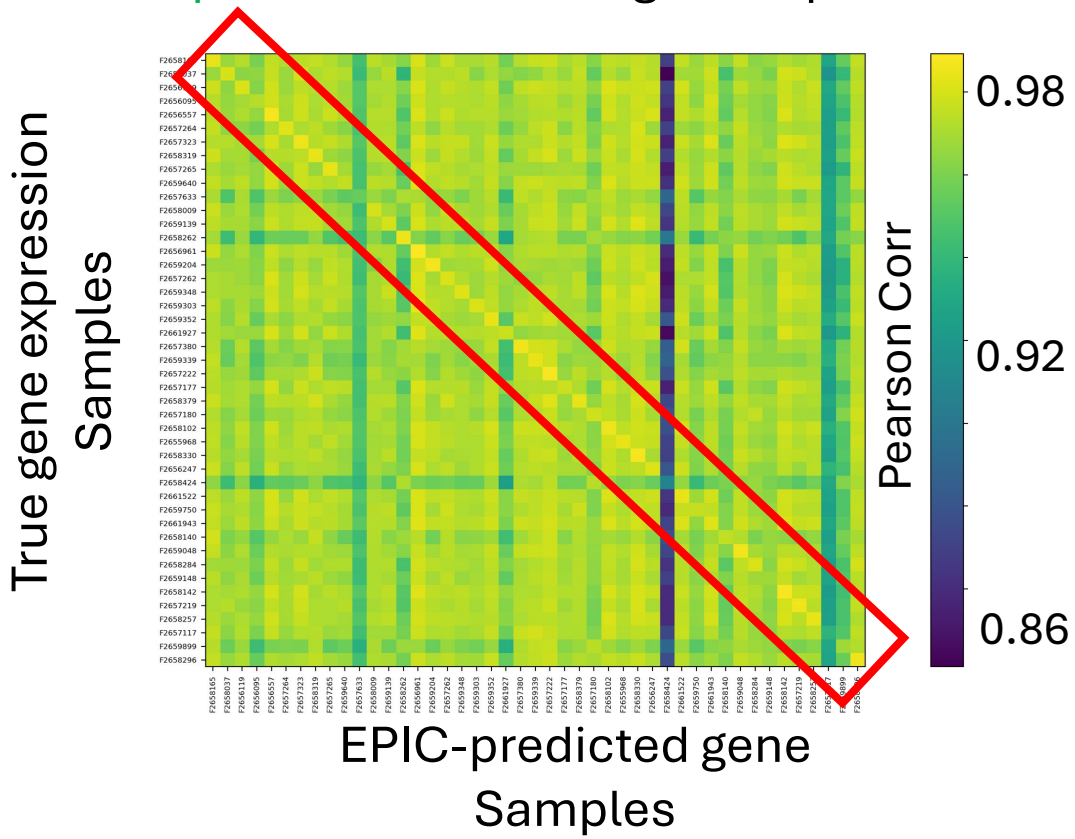
Transcriptomics prediction (same platform testing)

- **Trained on EPIC (~4,000 samples) → tested on EPIC (50 samples)**

Per-gene correlation between EPIC-predicted and **true** gene expression (>16,000 genes)

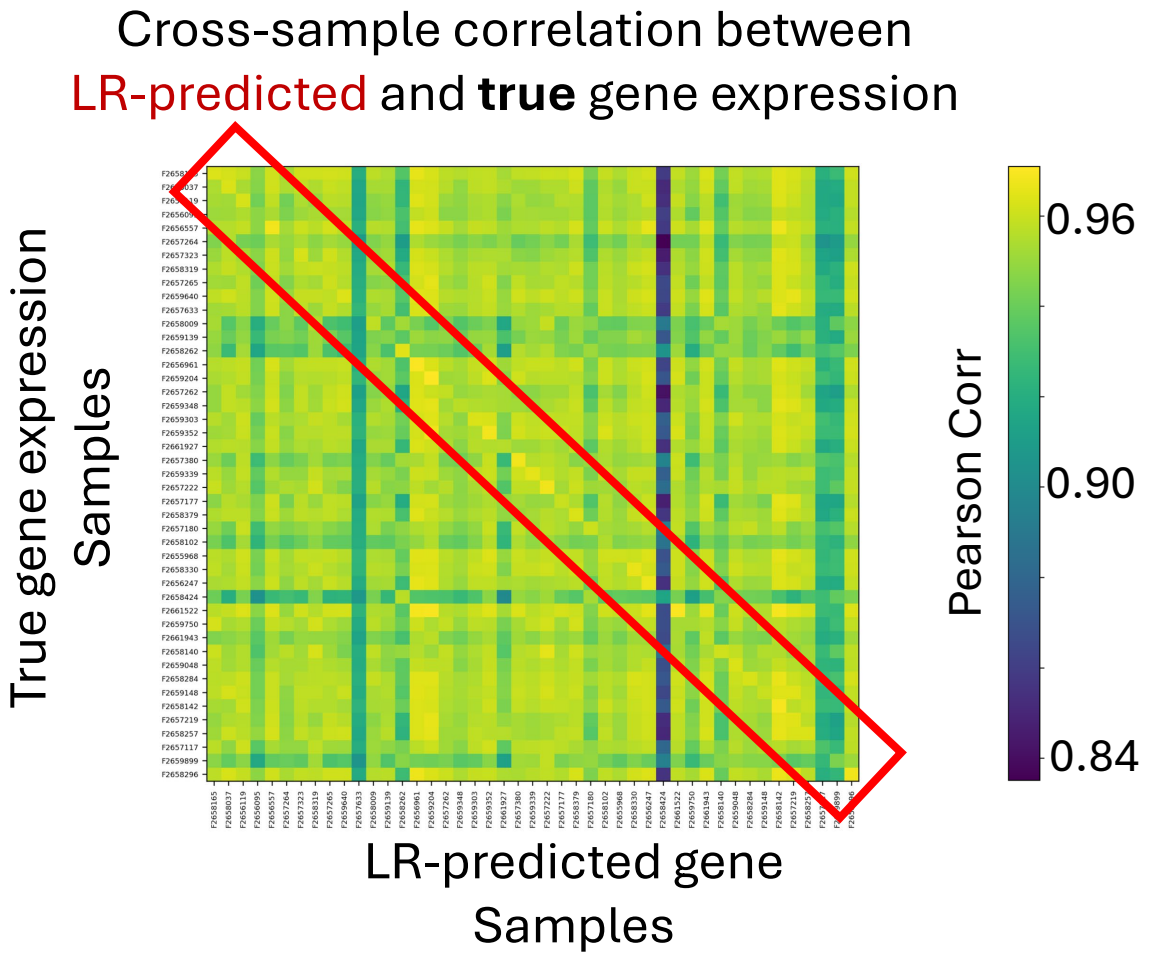


Cross-sample correlation between EPIC-predicted and **true** gene expression

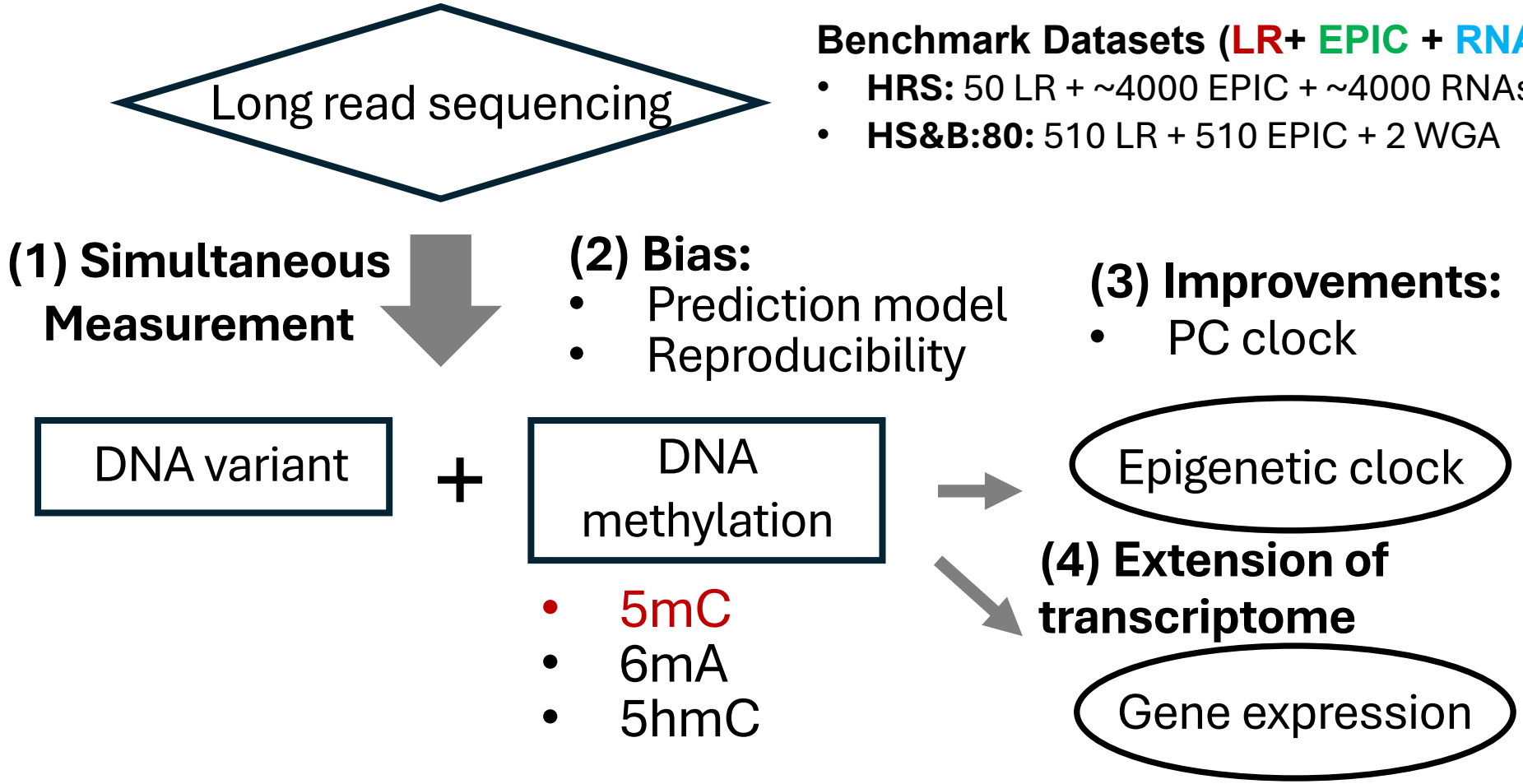


Transcriptomics prediction (platform transferring)

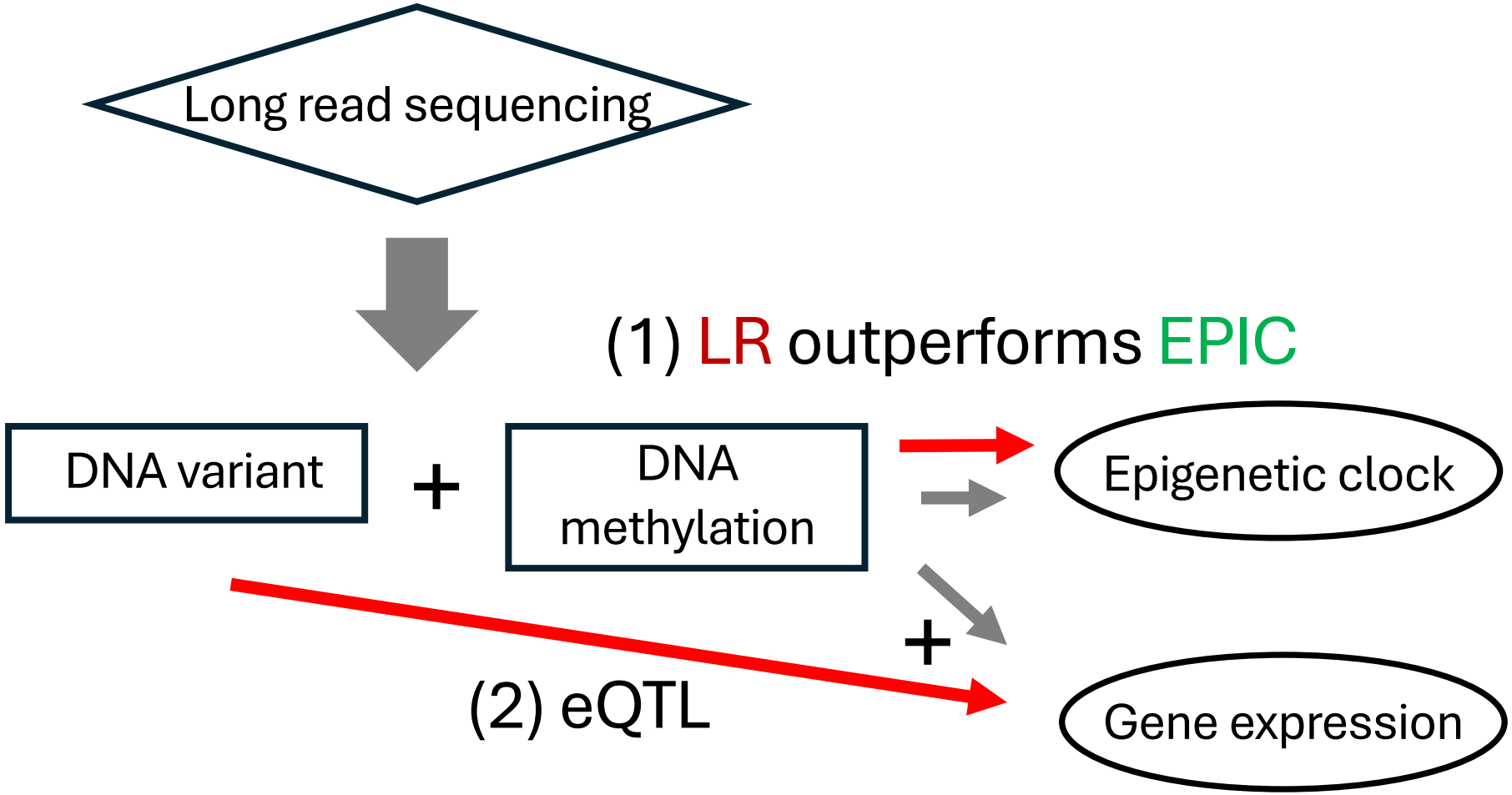
- **Trained on EPIC (~4,000 samples) → tested on LR (50 samples)**



Summary (2)


Long read sequencing
Benchmark Datasets (LR+ EPIC + RNAseq)
• HRS: 50 LR + ~4000 EPIC + ~4000 RNAseq
• HS&B:80: 510 LR + 510 EPIC + 2 WGA
(1) Simultaneous Measurement
(2) Bias:
• Prediction model
• Reproducibility
(3) Improvements:
• PC clock
DNA variant
+
DNA methylation
• 5mC
• 6mA
• 5hmC
Epigenetic clock
(4) Extension of transcriptome
Gene expression

Ongoing works



Acknowledgements

Laboratory Medicine and Pathology
University of Minnesota

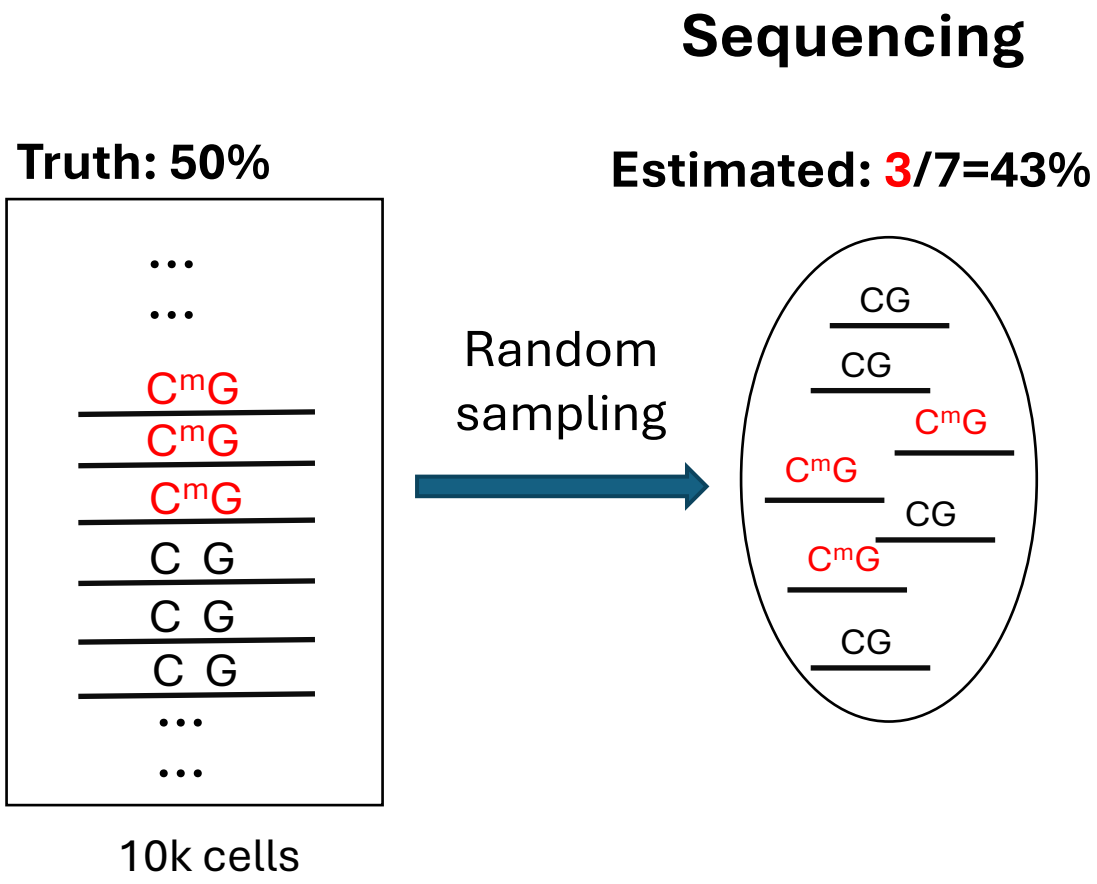
ARDL
Bharat Thyagarajan
Shannon Sulllivan
Stefani Thomas

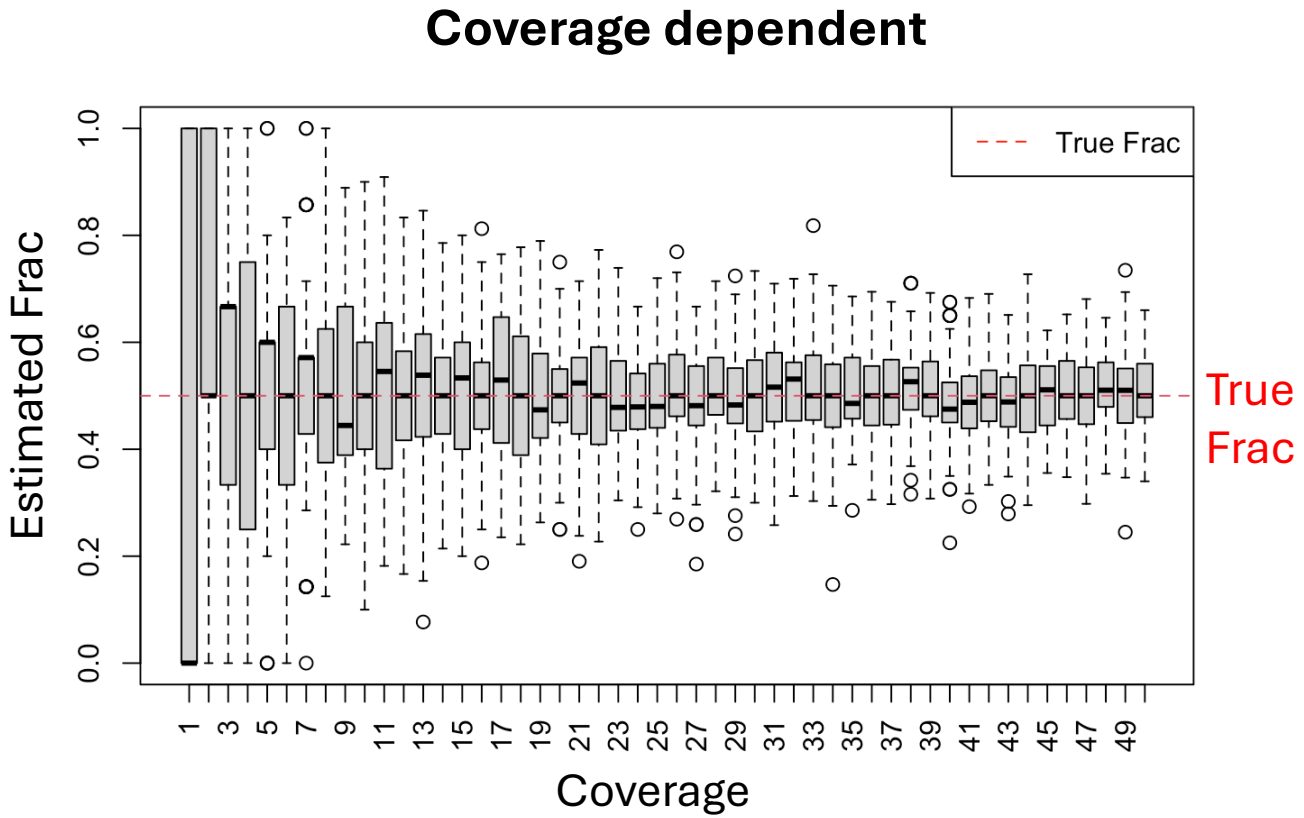
Akshat Rawat
Alec Victorsen
Aidan Ellison

High School and Beyond Study
John Robert Warren
Eric Grodsky
Chandra Muller
Adam Brickman
Jennifer Manly
Mateo Farina

Human Retirement Study
Eilleen Crimmins
Jessica Faul

Bias (2): LR read coverage

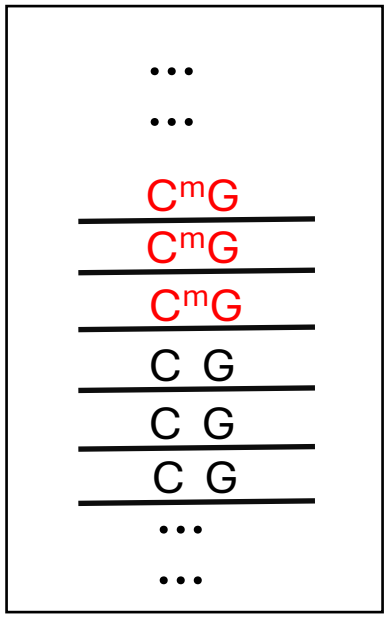




Bias (2): LR read coverage

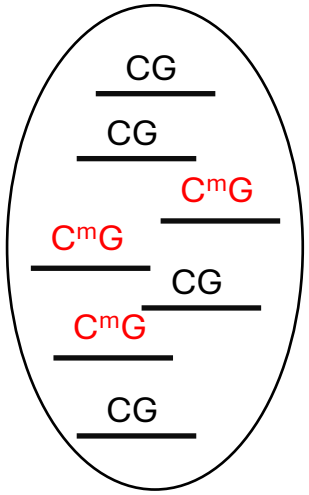
Sequencing

Truth: 50%



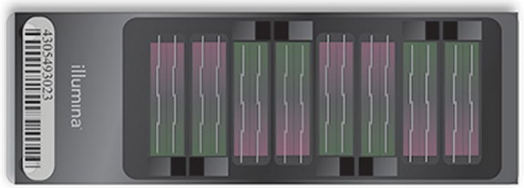
Random sampling

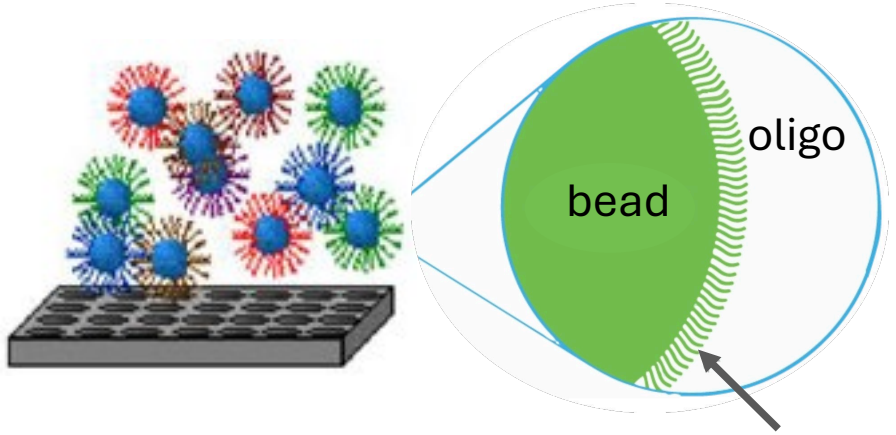
Estimated: 3/7=43%



VS.

BeadChip





Thousands of specific oligonucleotide probe to target a unique CpG site.