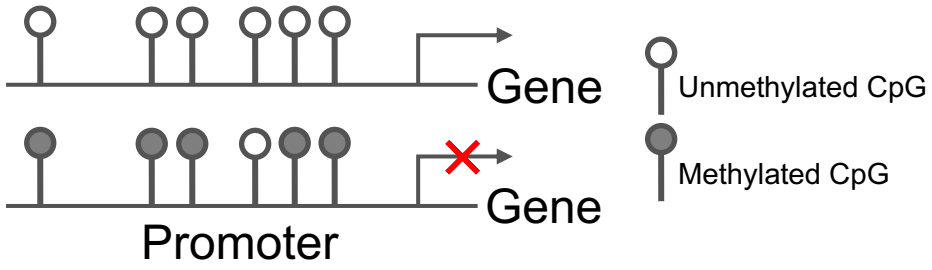
MAX PLANCK INSTITUTE
FOR MOLECULAR GENETICS

Differentially methylated region identification for multi-groups

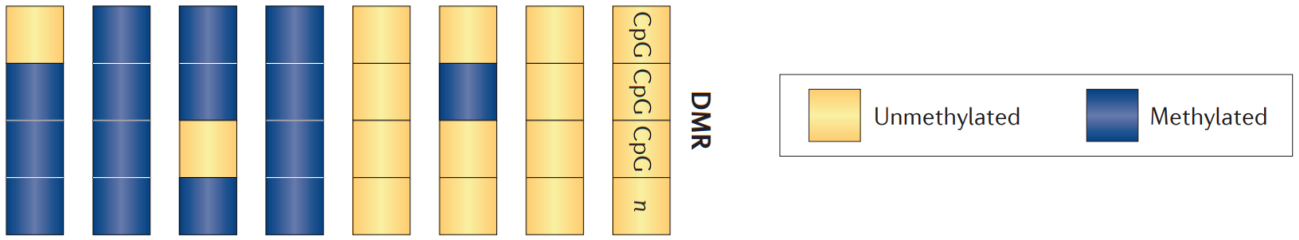
Zhihan Zhu, Kretzmer Lab at MPI for Molecular Genetics
39th TBI Winterseminar in Bled
15.02.2024

Background - Differentially Methylated Region (DMR)

CpG (a Cytosine followed by a Guanine) methylation:



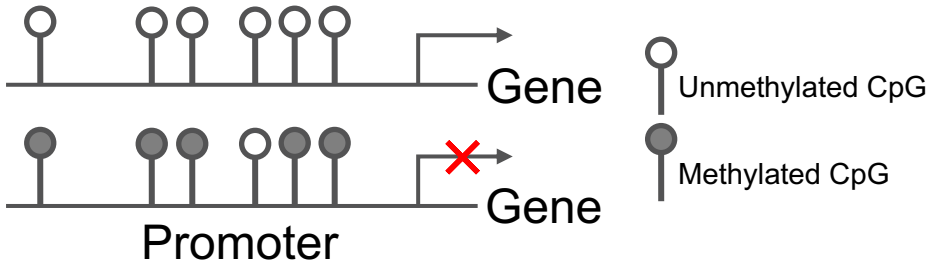
Differentially Methylated Region (DMR):



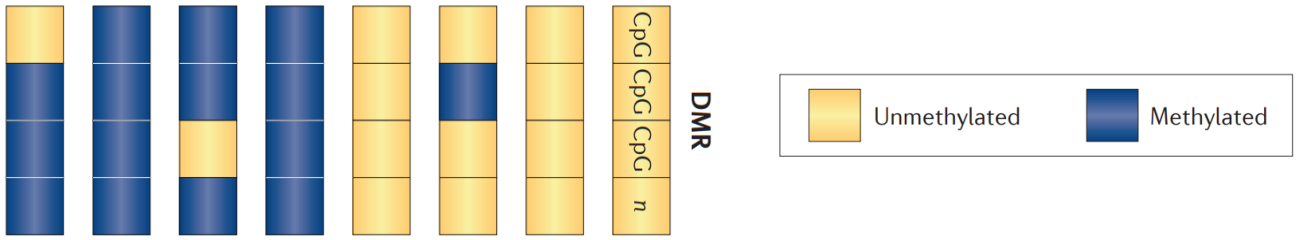
Rakyan, Vardhman K., et al. *Nature Reviews Genetics*. 2011.

Background - Differentially Methylated Region (DMR)

CpG (a Cytosine followed by a Guanine) methylation:



Differentially Methylated Region (DMR):

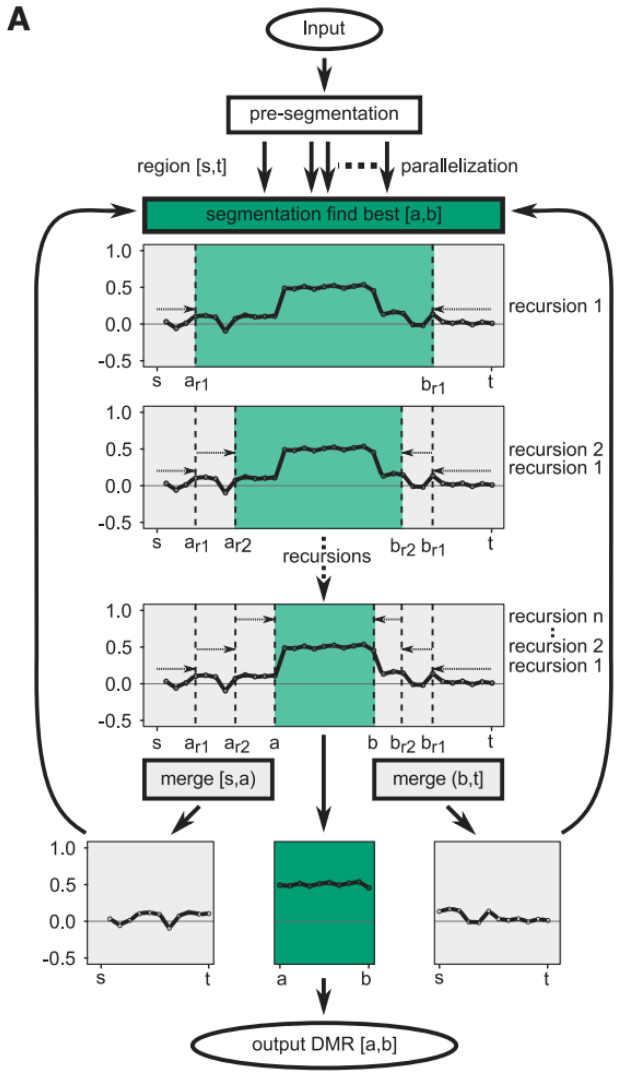


Noisy

Huge #CpGs

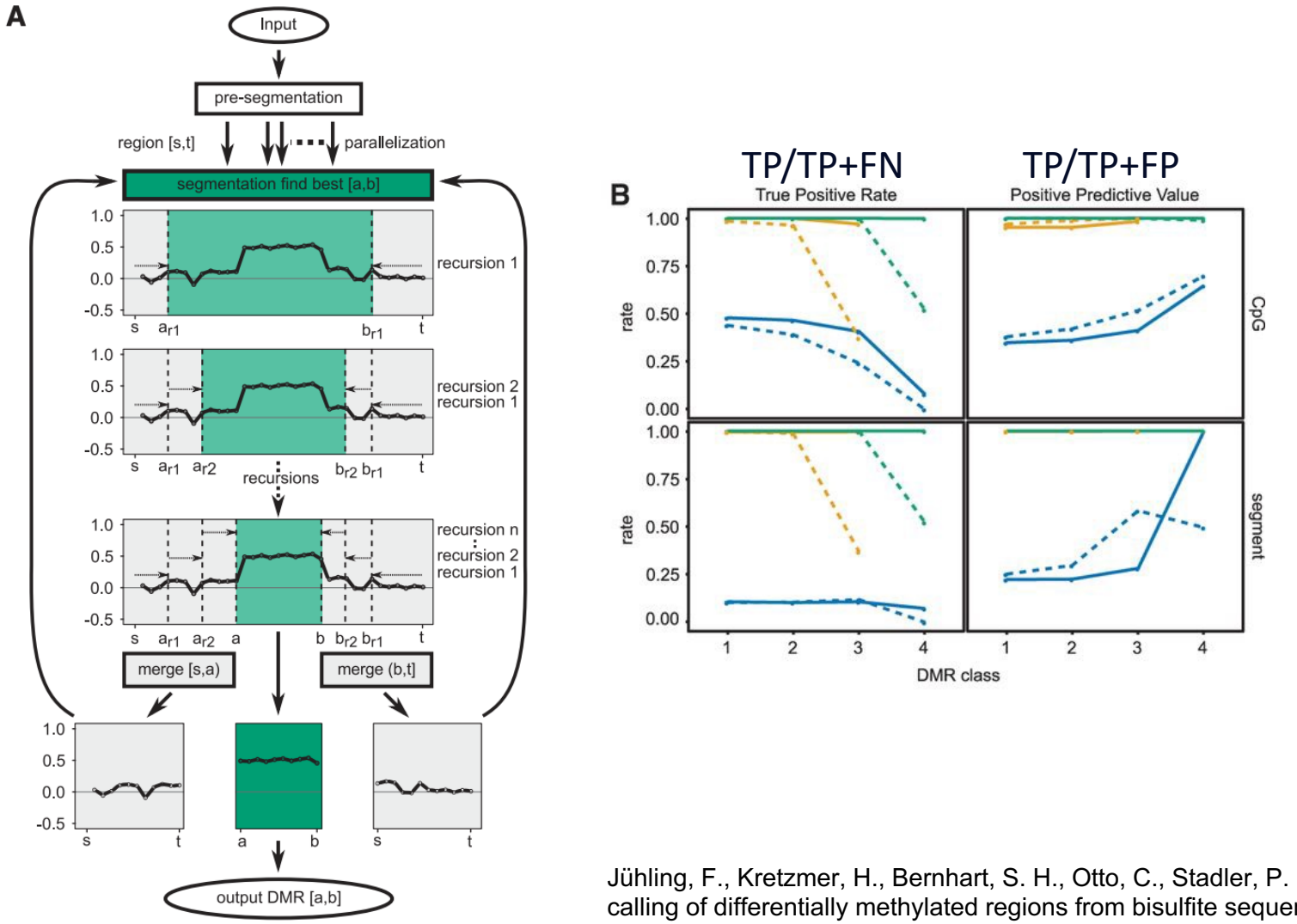
Rakyan, Vardhman K., et al. *Nature Reviews Genetics*. 2011.

Background - metilene



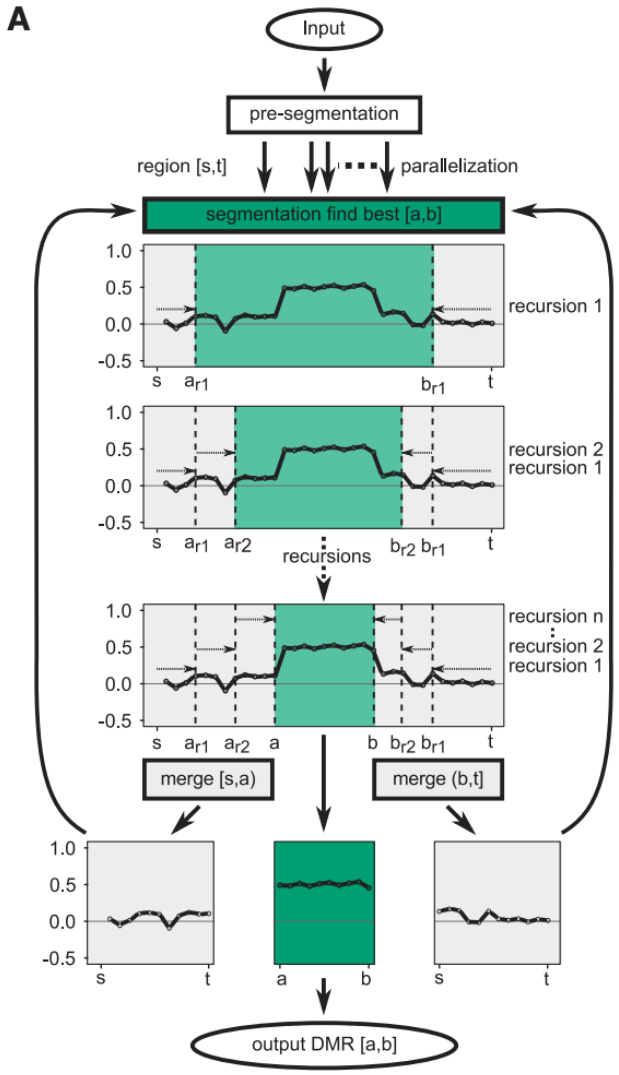
Jühling, F., Kretzmer, H., Bernhart, S. H., Otto, C., Stadler, P. F., & Hoffmann, S. (2016). Metilene: fast and sensitive calling of differentially methylated regions from bisulfite sequencing data. *Genome research*, *26*(2), 256-262.

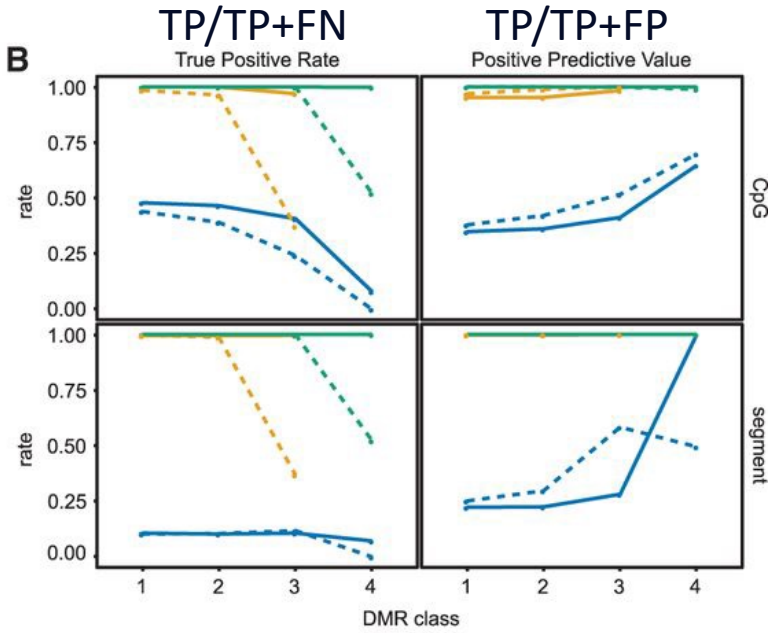
Background - metilene

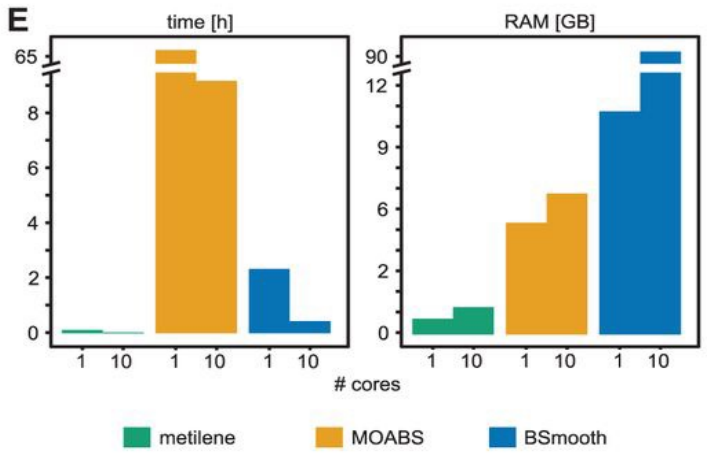


Jühling, F., Kretzmer, H., Bernhart, S. H., Otto, C., Stadler, P. F., & Hoffmann, S. (2016). Metilene: fast and sensitive calling of differentially methylated regions from bisulfite sequencing data. *Genome research*, *26*(2), 256-262.

Background - metilene

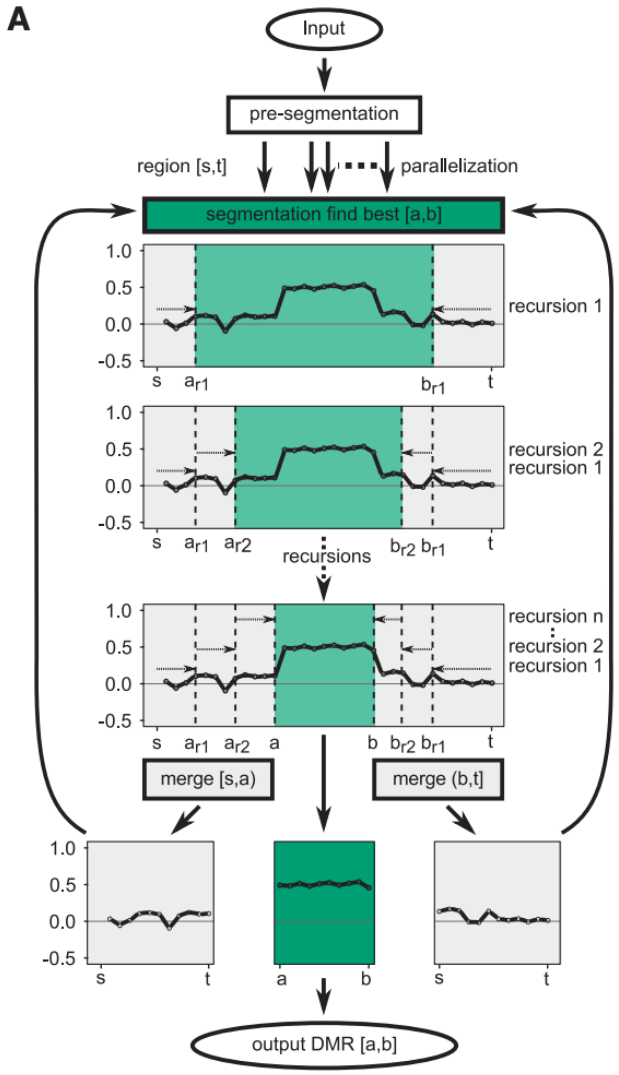


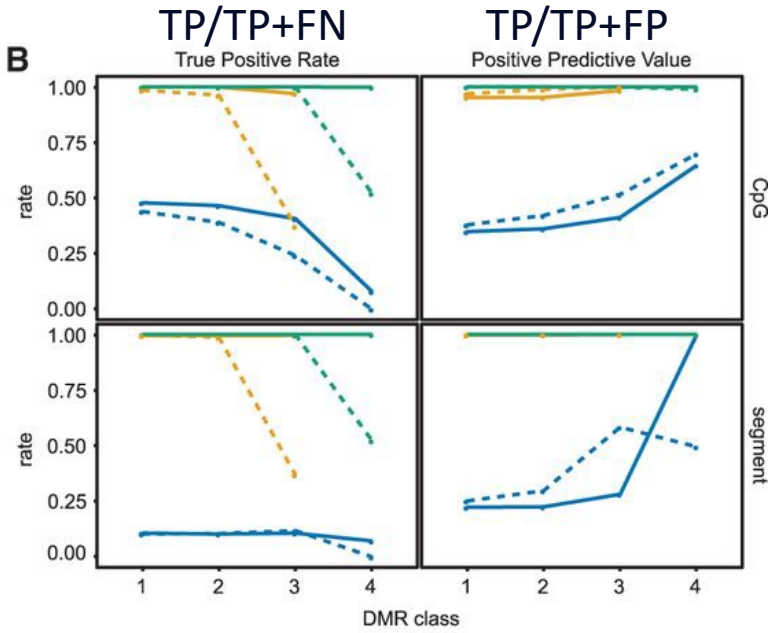


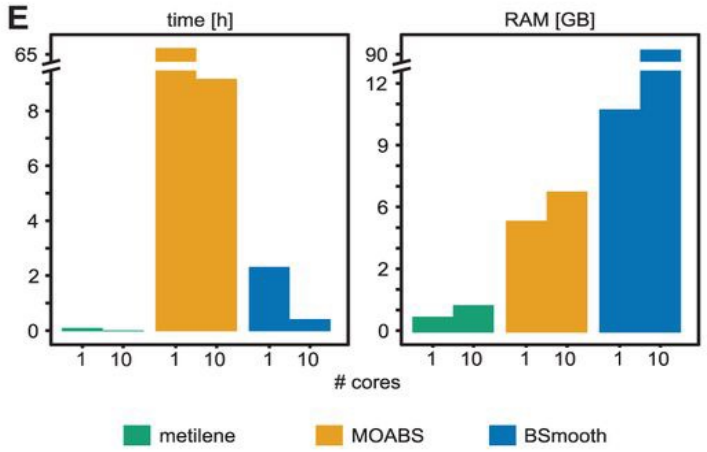


Jühling, F., Kretzmer, H., Bernhart, S. H., Otto, C., Stadler, P. F., & Hoffmann, S. (2016). Metilene: fast and sensitive calling of differentially methylated regions from bisulfite sequencing data. *Genome research*, *26*(2), 256-262.

Background - metilene







A gap – more than two groups to be compared?

Jühling, F., Kretzmer, H., Bernhart, S. H., Otto, C., Stadler, P. F., & Hoffmann, S. (2016). Metilene: fast and sensitive calling of differentially methylated regions from bisulfite sequencing data. *Genome research*, *26*(2), 256-262.

A gap – more than two groups to be compared?

Multi-groups DMR identification:

We can have ~200 cell types

Cancers can have even more subtypes

…

A gap – more than two groups to be compared?

Multi-groups DMR identification:

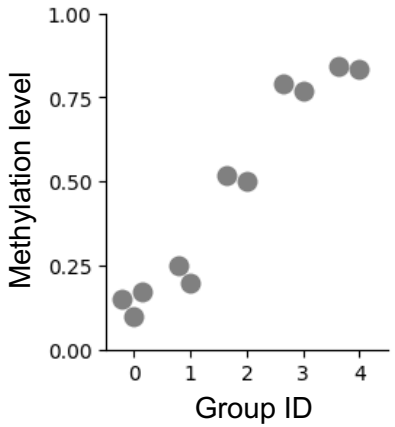
We can have ~200 cell types

Cancers can have even more subtypes

…

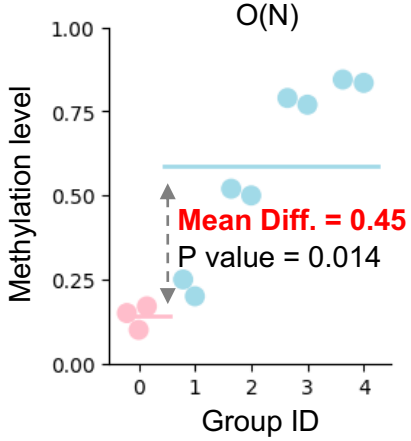
Possible ways of comparing multiple groups – 5-groups example:

Data:

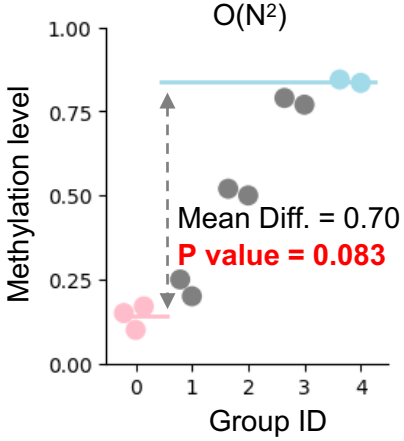


Solutions:

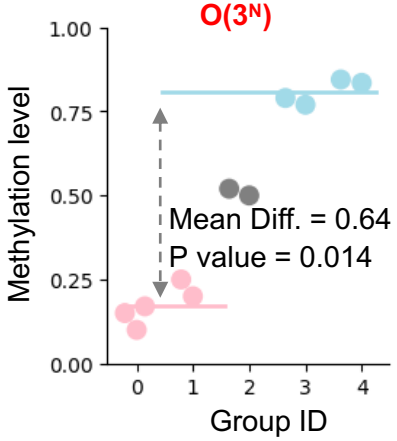
One vs All others
O(N)



One vs One
O(N^2)



All possible comparisons
O(3^N)



$$\frac{3^N - (2\times 2^N - 1)}{2}$$

e.g., N = 20 → 1,742,343,625

Multi-groups metilene – algorithm

Segmentation based on 1 group vs 1 group comparisons
$O(N^2)$

↓

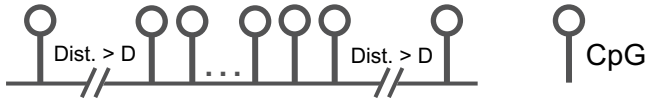
Groups clustering on segments
$O(N)$

↓

Circulation (recursion) based on clustering results
$O(N^2)$

Multi-groups metilene – algorithm

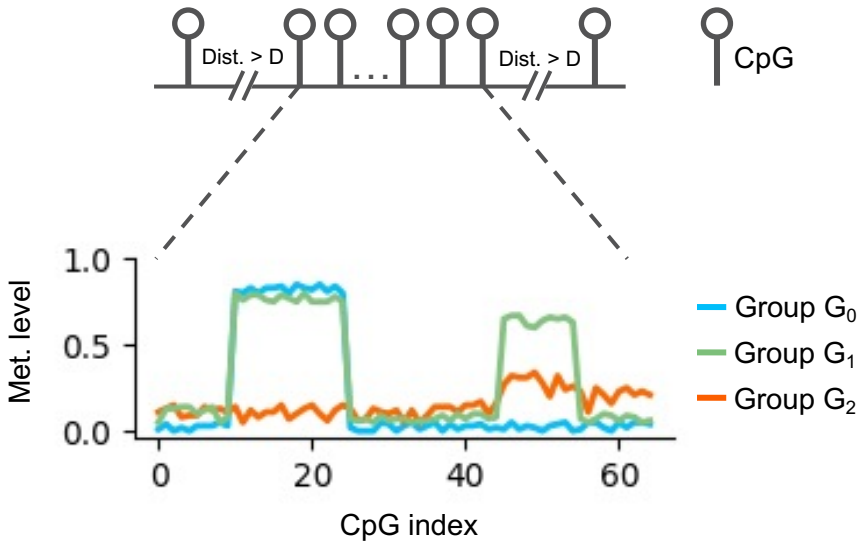
Pre-segmentation



Multi-groups metilene – algorithm

Pre-segmentation

Mean methylation levels M: O(N)

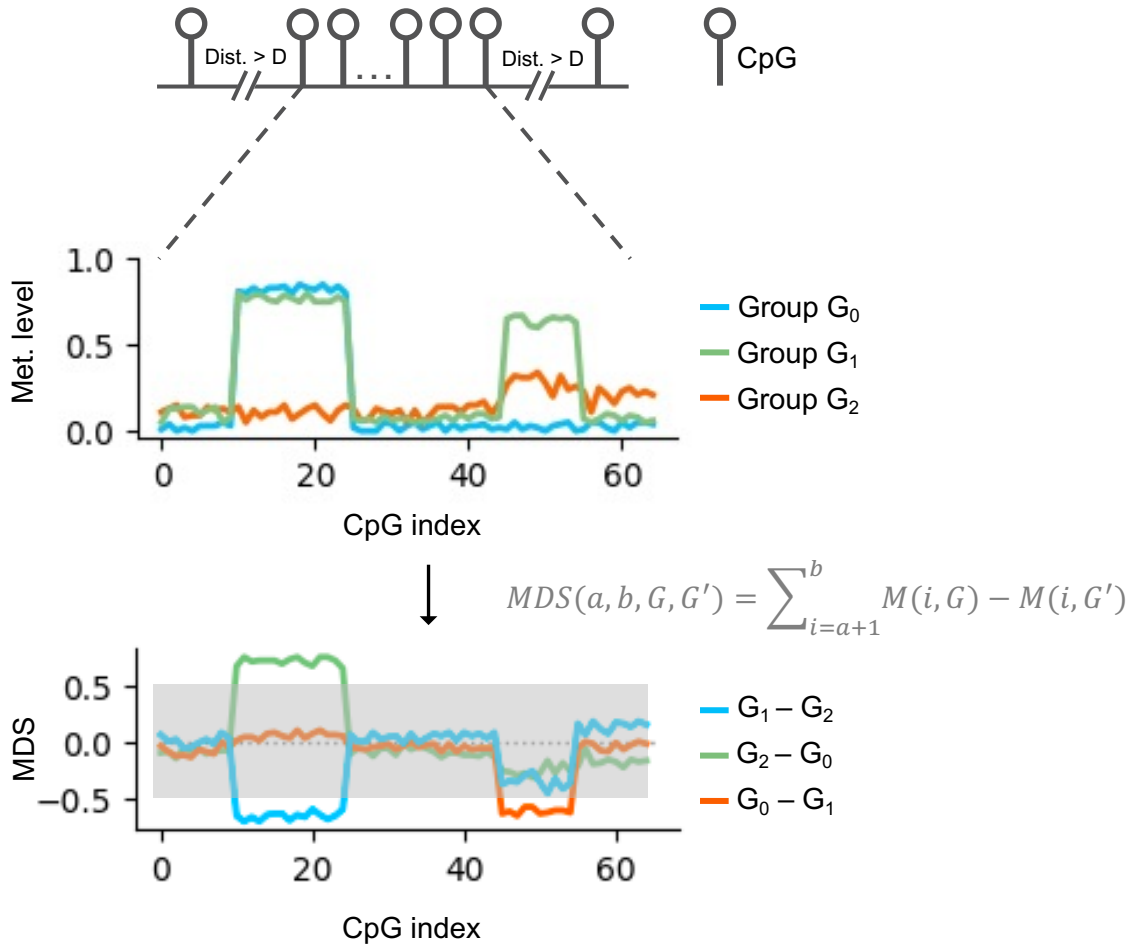


Multi-groups metilene – algorithm

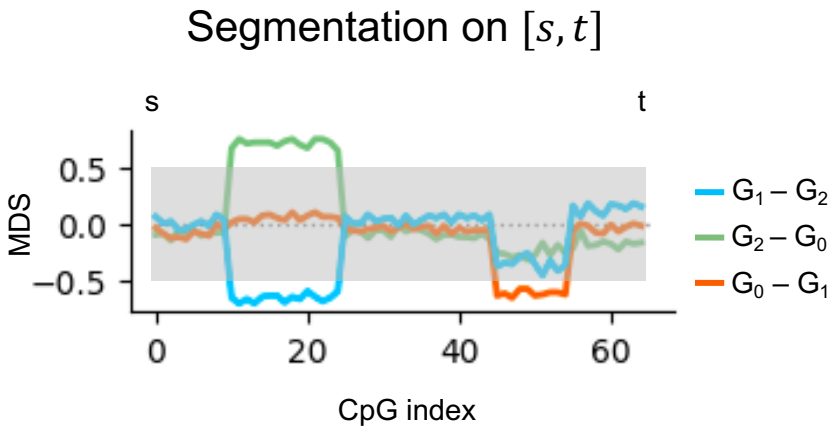
Pre-segmentation

Mean methylation levels M: O(N)

Mean difference signals MDS: O(N^2)

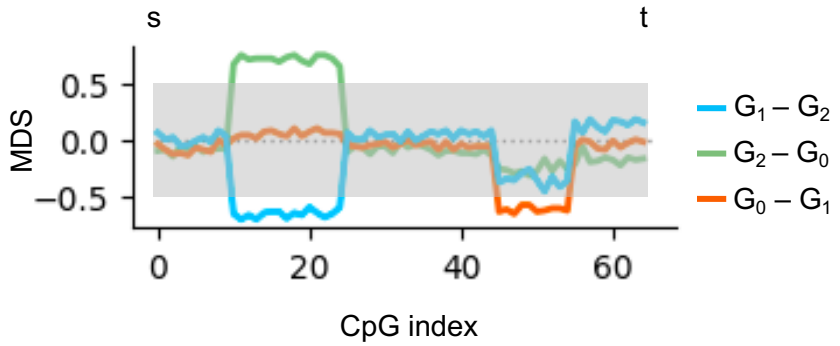


Multi-groups metilene – algorithm



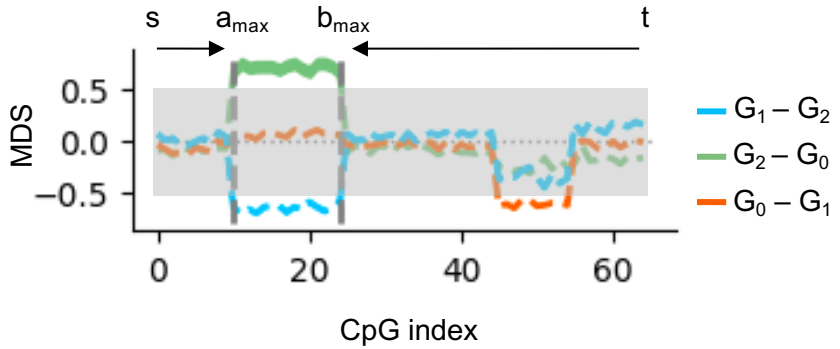
Multi-groups metilene – algorithm

Segmentation on $[s, t]$



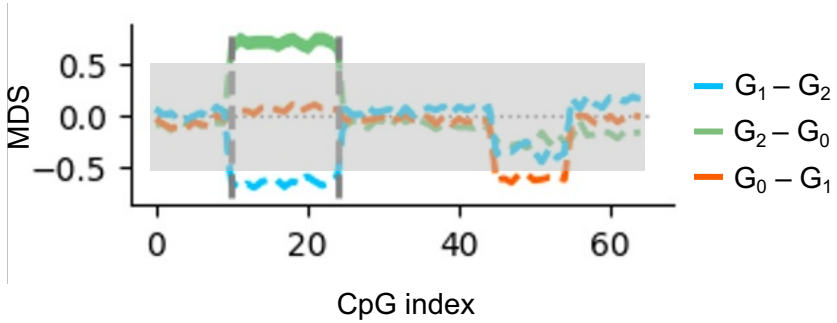
$$Z(s,t,a,b,G,G') = \frac{\left[|MDS(a,b,G,G')| - \frac{b-a}{t-s} \cdot |MDS(s,t,G,G')|\right]^2}{(b-a)\left[1 - \frac{b-a}{t-s}\right]}$$

$$a_{max}, b_{max}, G_{max}, G'_{max} = \underset{s \le a < b \le t, G, G'}{argmax} Z(s,t,a,b,G,G')$$



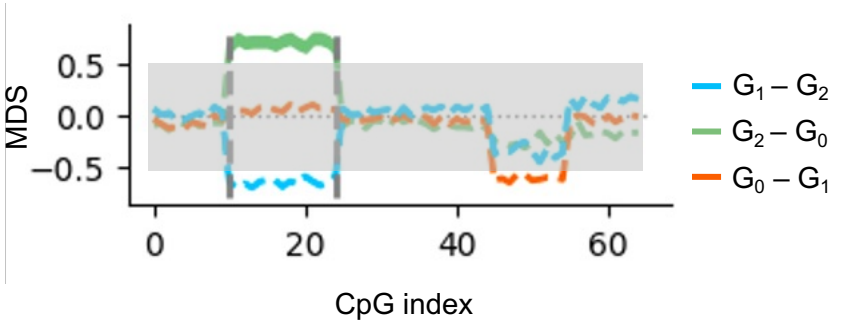
Multi-groups metilene – algorithm

Clustering based on G_{max}, G'_{max} :

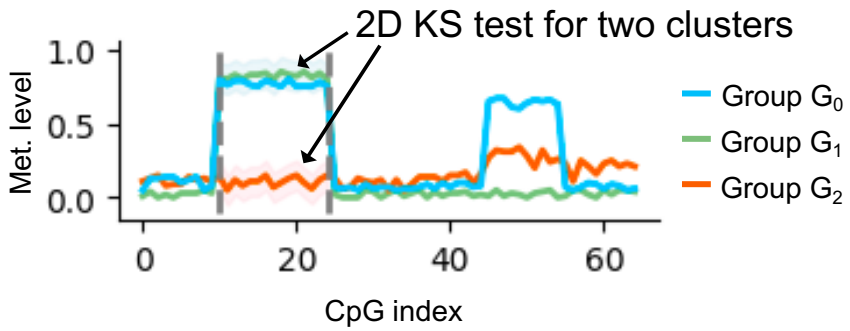


Multi-groups metilene – algorithm

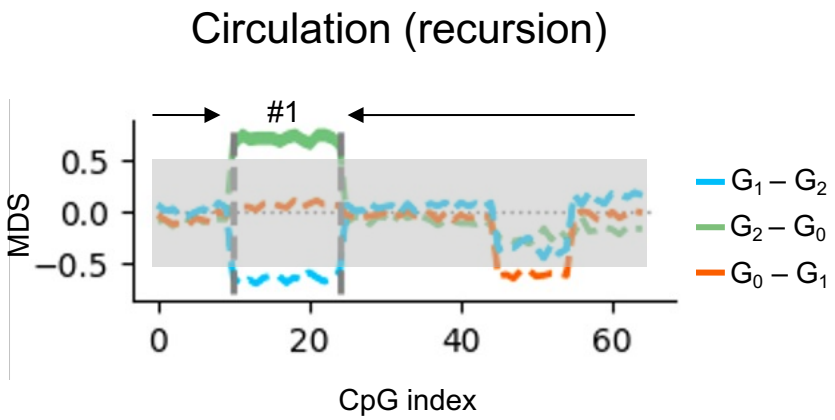
Clustering based on G_{max}, G'_{max} :



$$Distance(a,b,n,m) = \frac{\sum_{i=a+1}^{b}(|M(i,n) - M(i,m)| < \varepsilon)}{(b\ -a)}$$

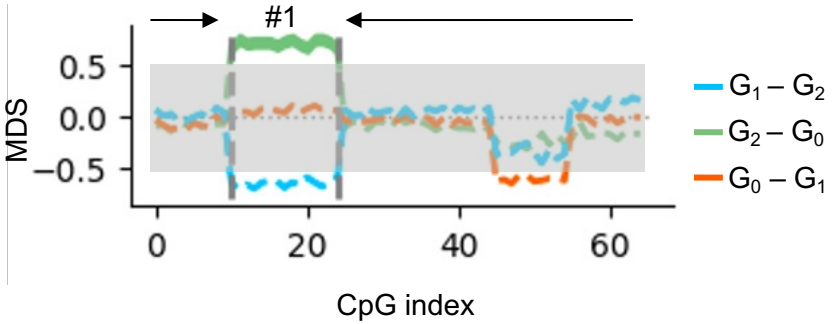


Multi-groups metilene – algorithm

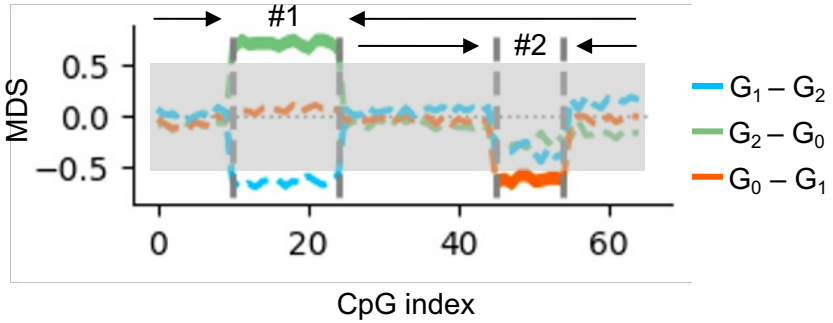


Multi-groups metilene – algorithm

Circulation (recursion)



- # CpGs $> \omega$ and
- $\exists P(a, b, A', B') < P(s, t, A, B),\ s \leq a < b \leq t$



Multi-groups metilene – De novo mode

DMR identification between **TWO** groups

Multi-groups metilene – De novo mode

DMR identification between **TWO** groups

DMR identification among **>=3** groups

Multi-groups metilene – De novo mode

DMR identification between **TWO** groups

DMR identification among **>=3** groups

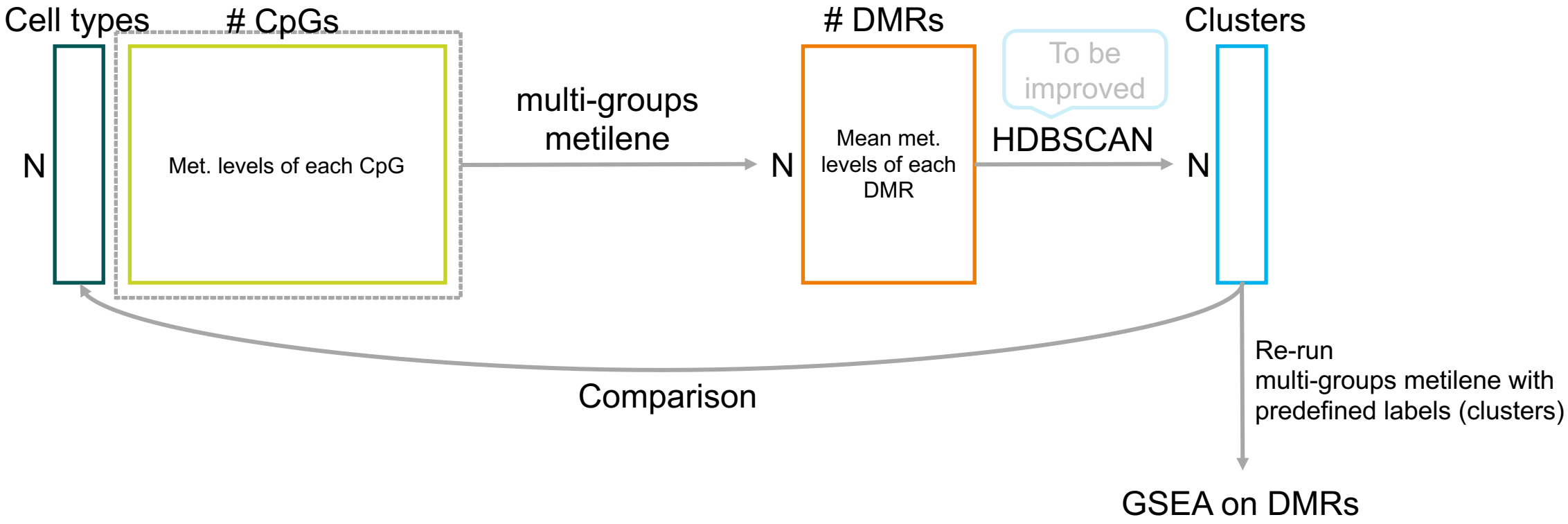
DMR identification without group information (label) - De novo mode
(One group one sample)

Multi-groups metilene – application

DMR identification (de novo mode) on N=21 WGBS samples from pancreas

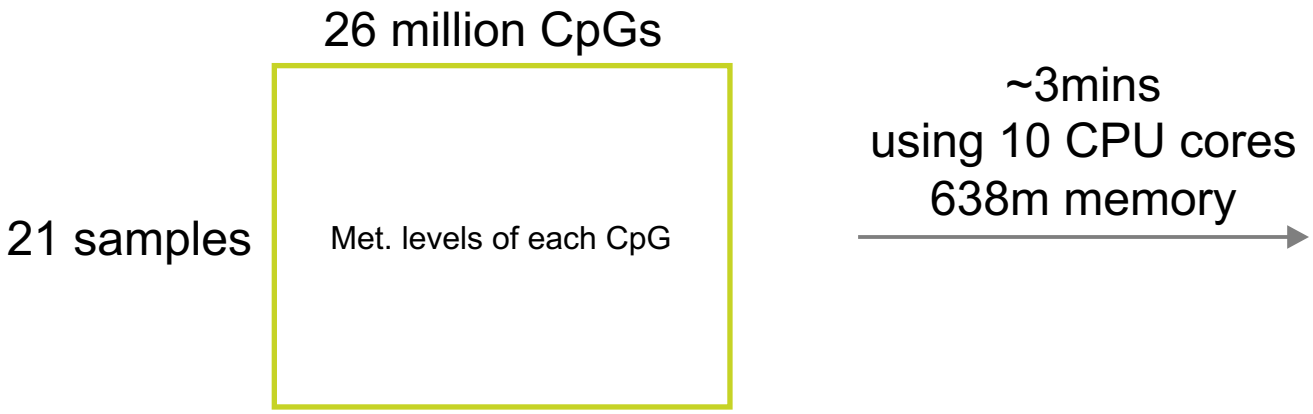
Loyfer, Netanel, et al. *Nature*. 2023.

- 6 cell types: endothelium, acinar, alpha, beta, delta, duct



Multi-groups metilene – application

Multi-groups metilene with (default) parameters -

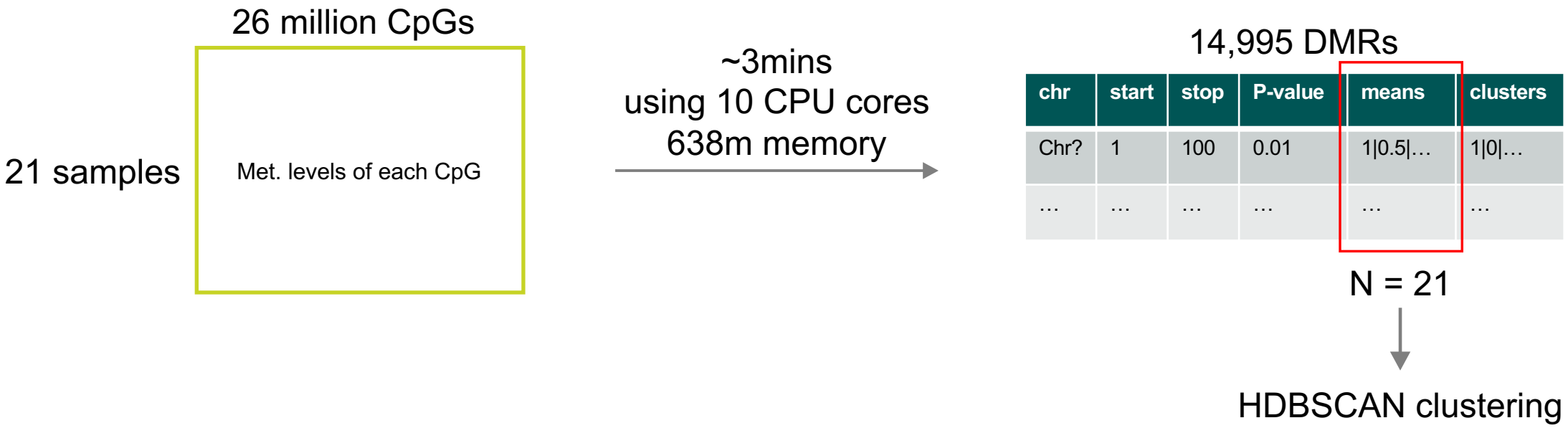


14,995 DMRs

chr	start	stop	P-value	means	clusters
Chr?	1	100	0.01	1|0.5|…	1|0|…
…	…	…	…	…	…

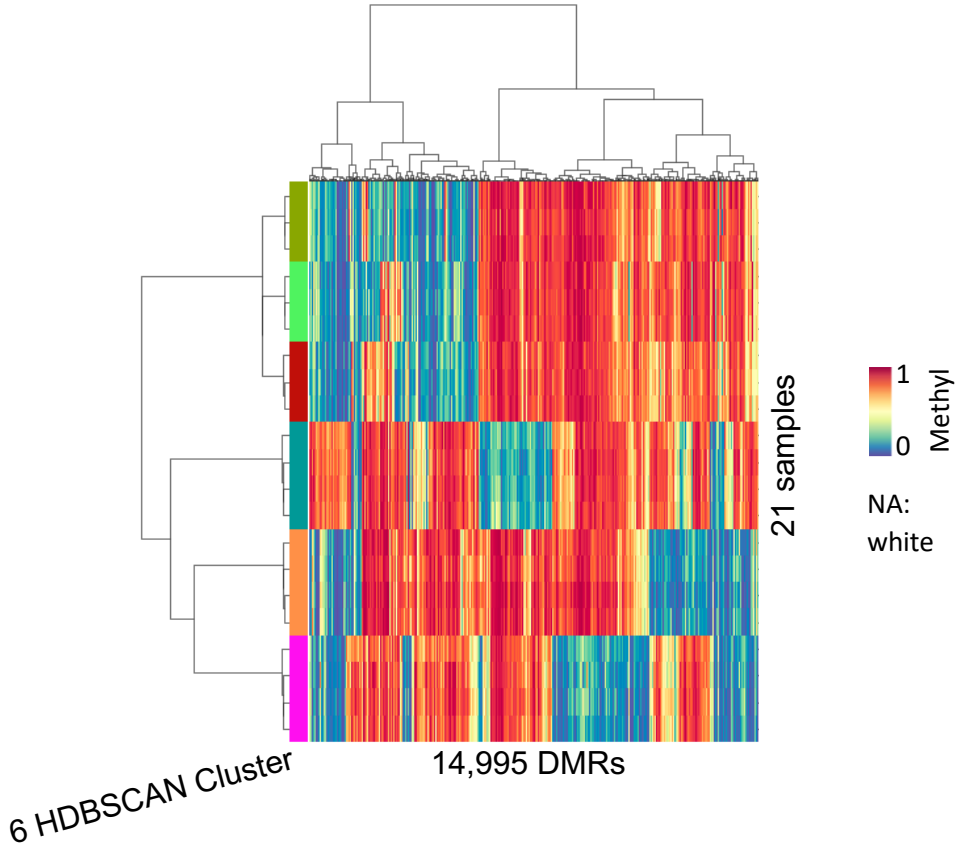
Multi-groups metilene – application

Multi-groups metilene with (default) parameters -



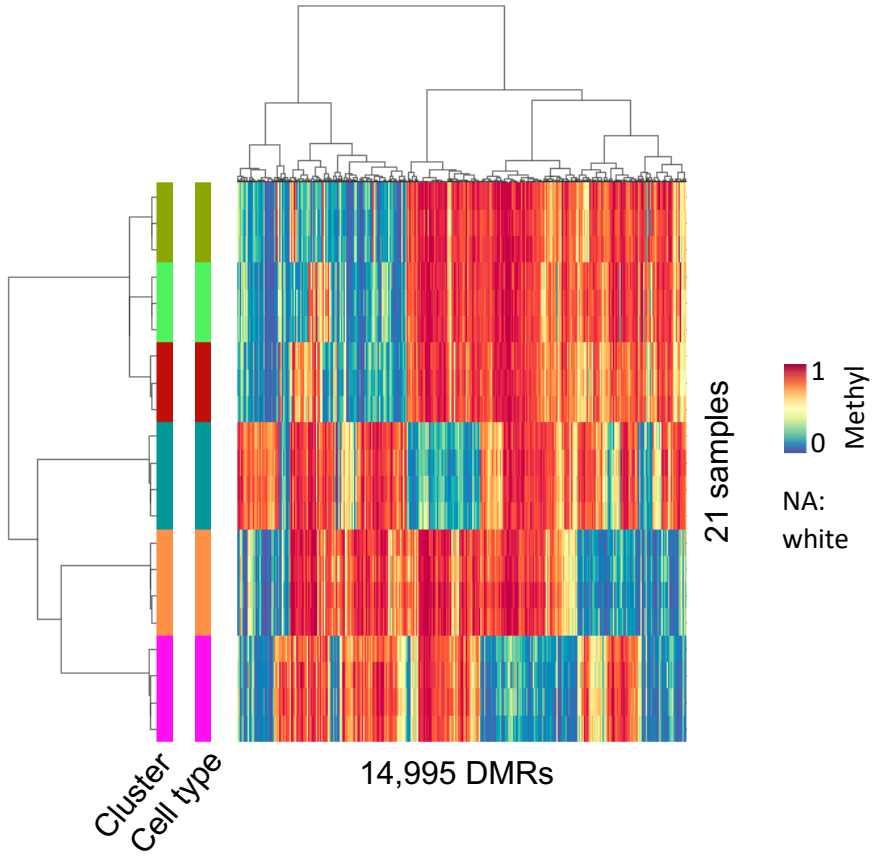
Multi-groups metilene – application

HDBSCAN clustering on mean methylation levels of DMRs



Multi-groups metilene – application

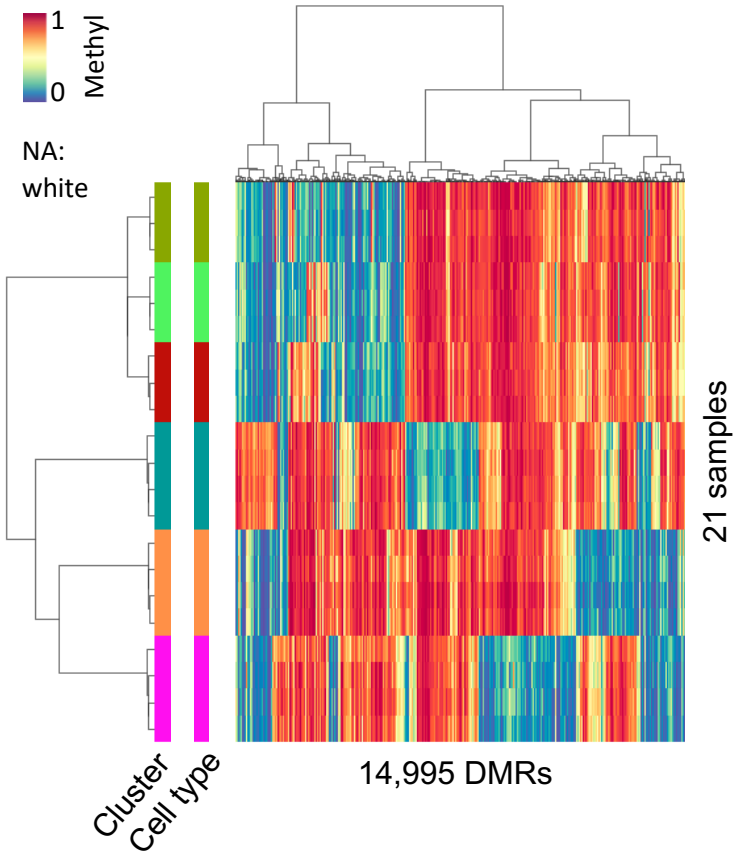
HDBSCAN clustering on mean methylation levels of DMRs

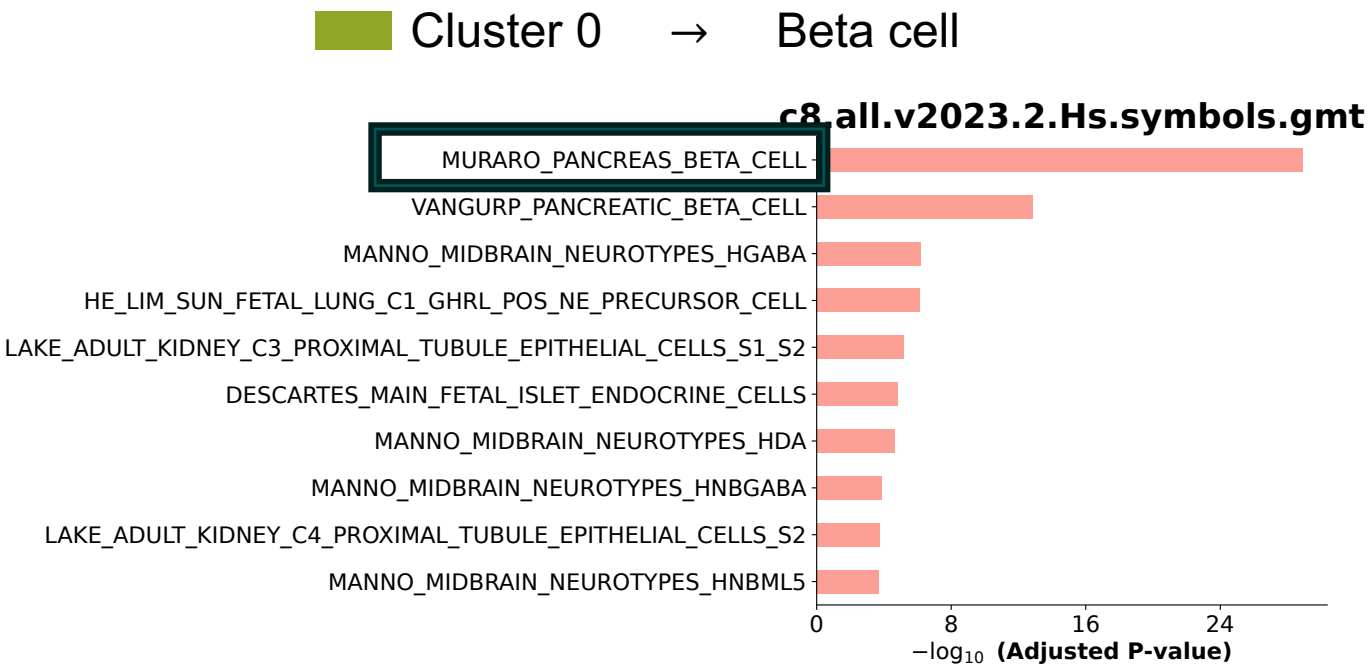


- Cluster 0 → Beta cell
- Cluster 1 → Alpha cell
- Cluster 2 → Delta cell
- Cluster 3 → Endothelium cell
- Cluster 4 → Duct cell
- Cluster 5 → Acinar cell

Multi-groups metilene – application

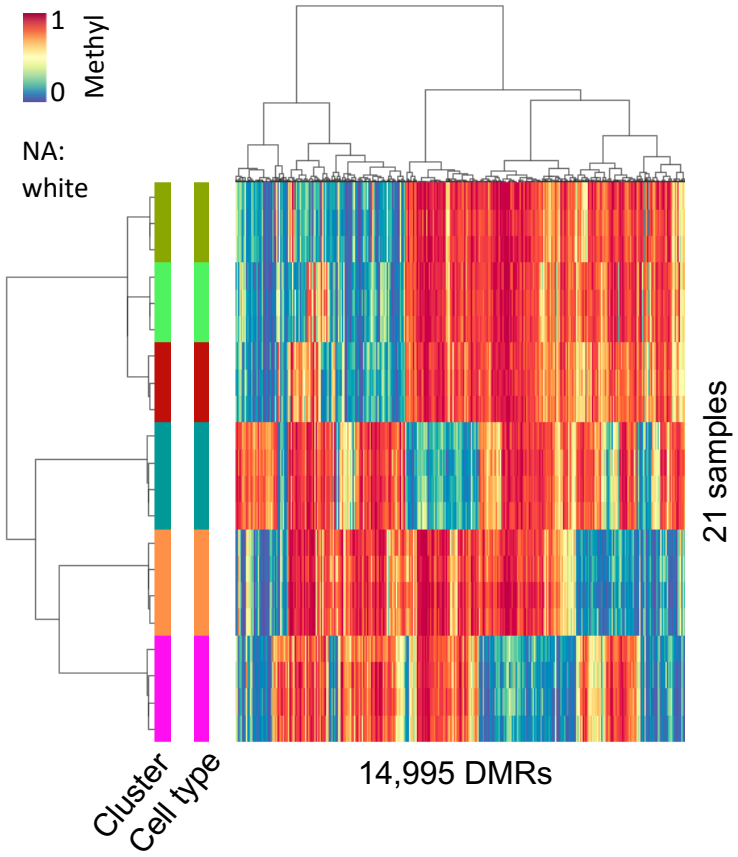
GSEA on cell type-specific expression gene sets from single-cell RNA sequencing:

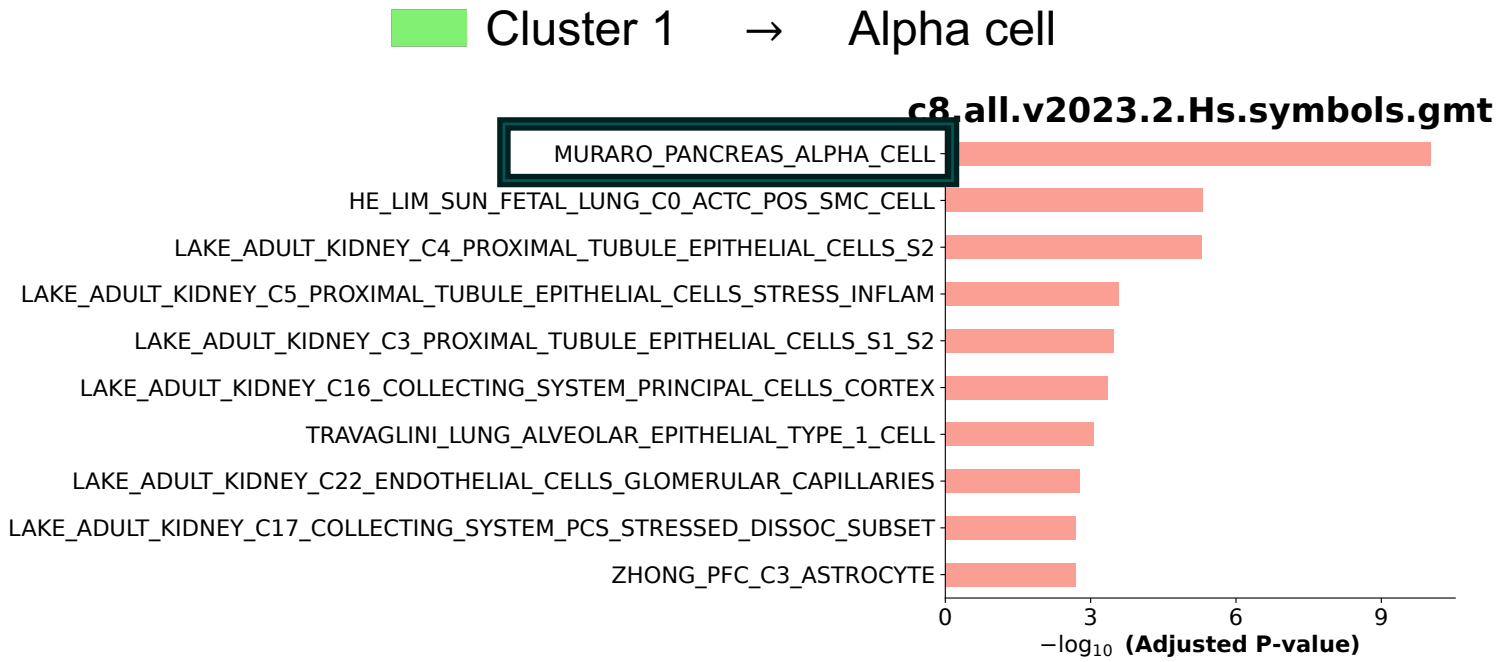




Multi-groups metilene – application

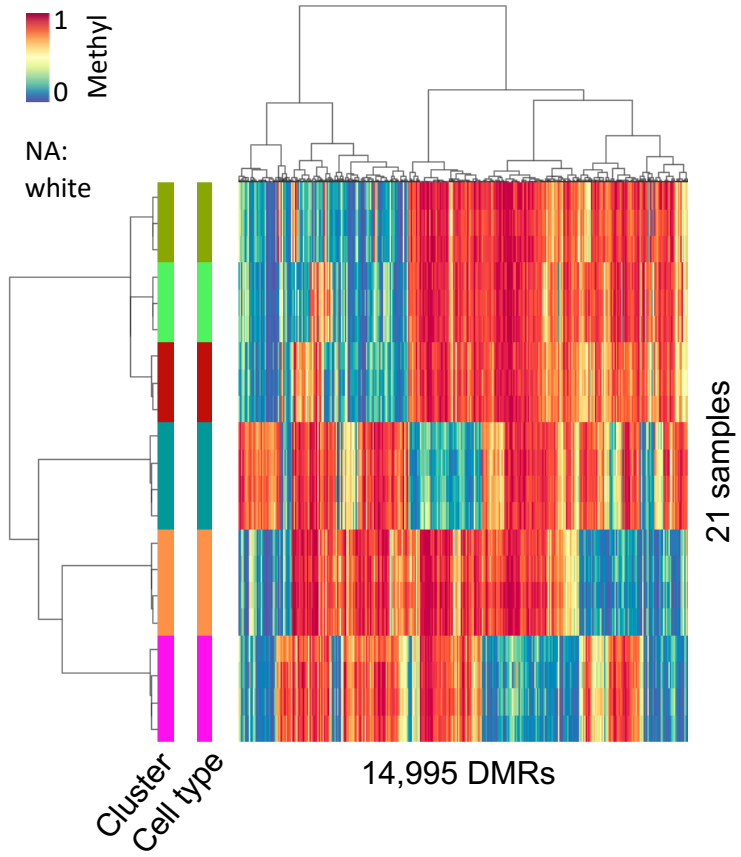
GSEA on cell type-specific expression gene sets from single-cell RNA sequencing:

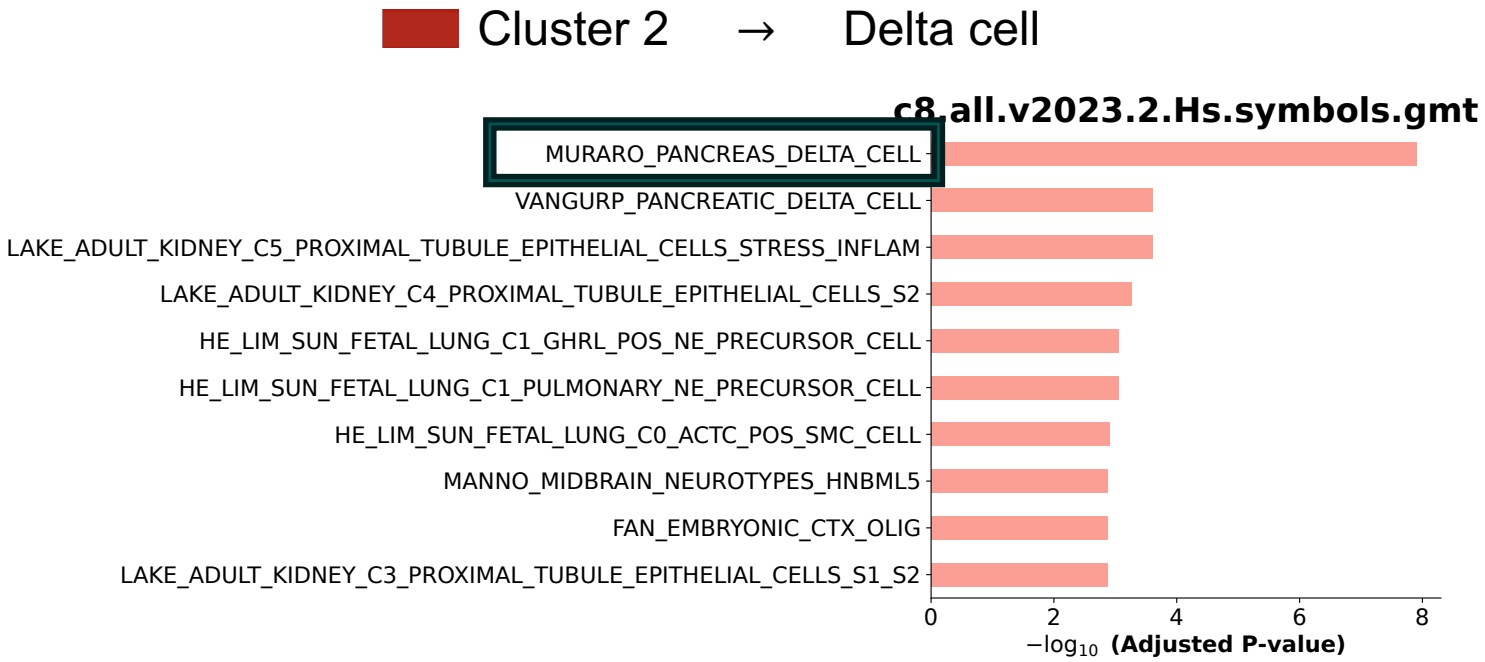




Multi-groups metilene – application

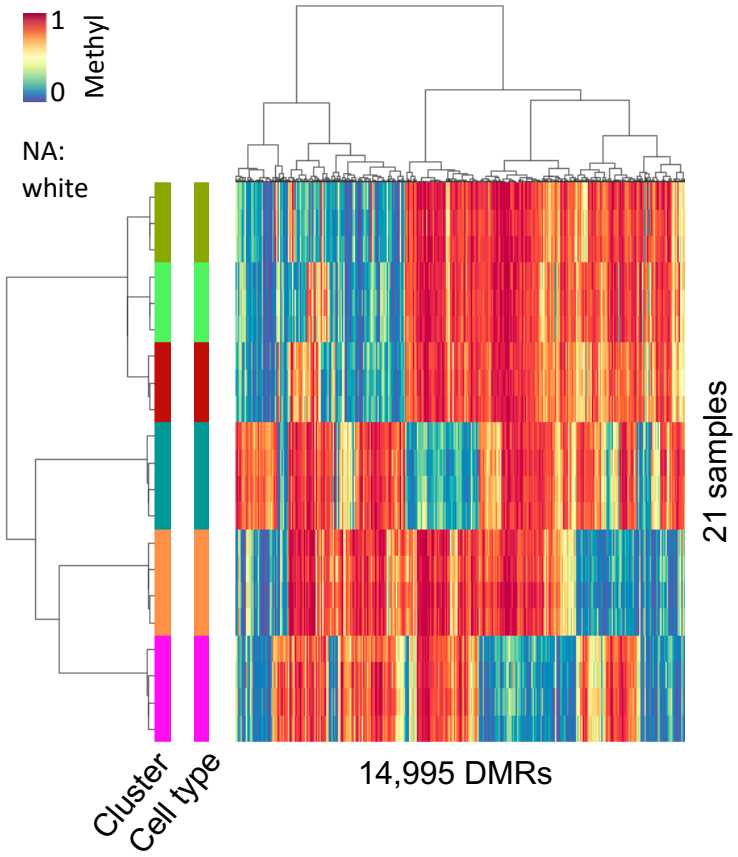
GSEA on cell type-specific expression gene sets from single-cell RNA sequencing:

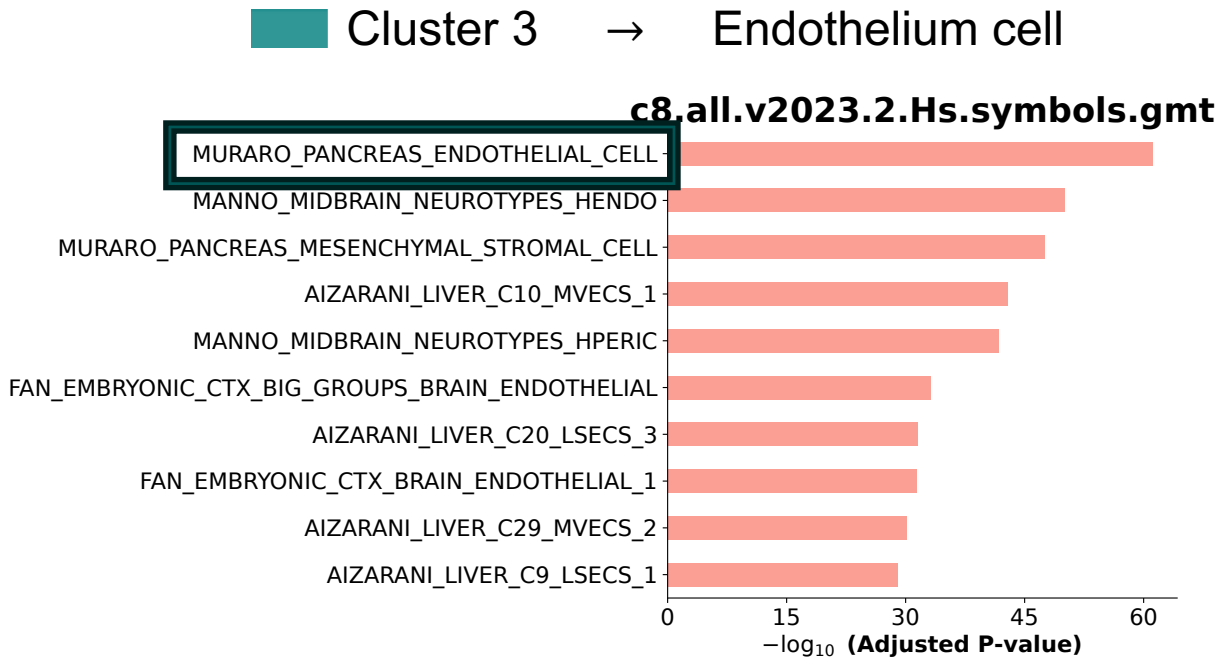




Multi-groups metilene – application

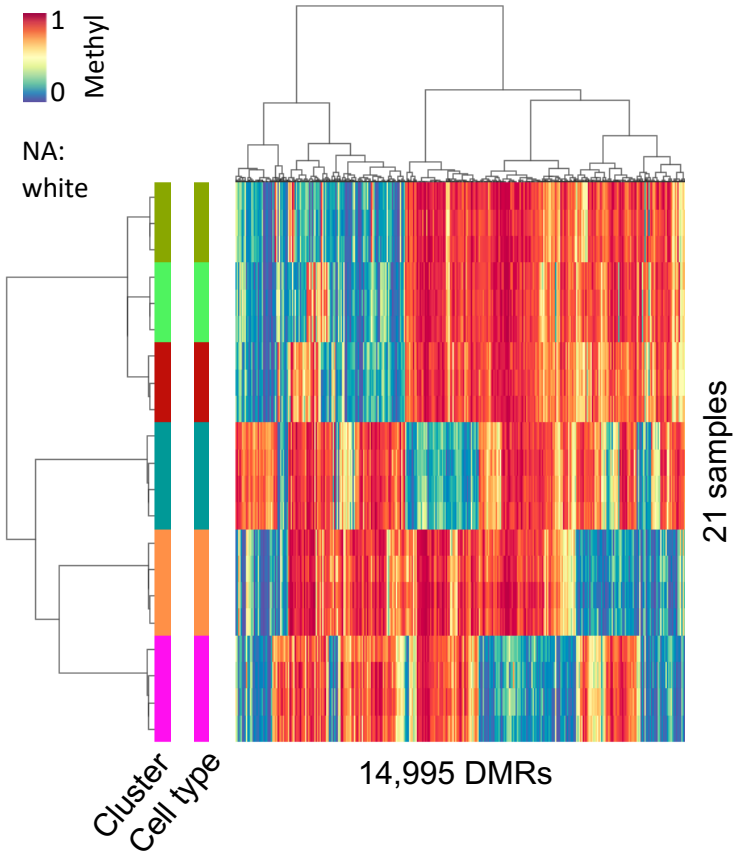
GSEA on cell type-specific expression gene sets from single-cell RNA sequencing:

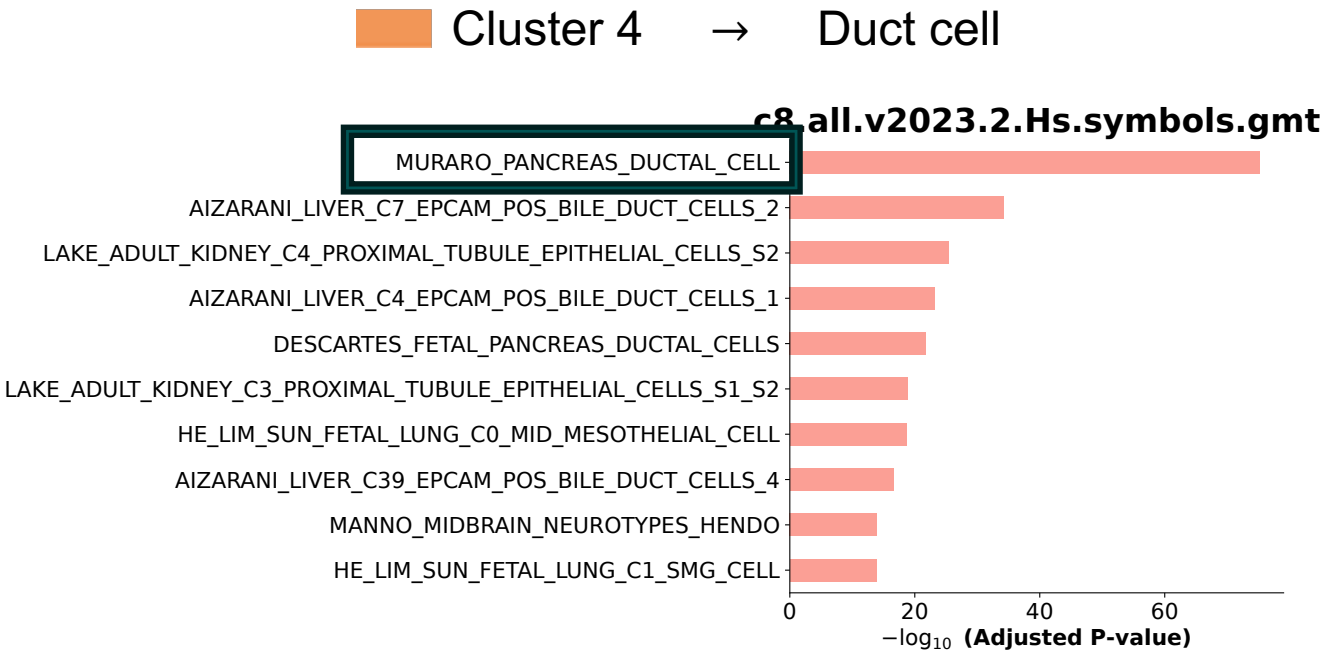




Multi-groups metilene – application

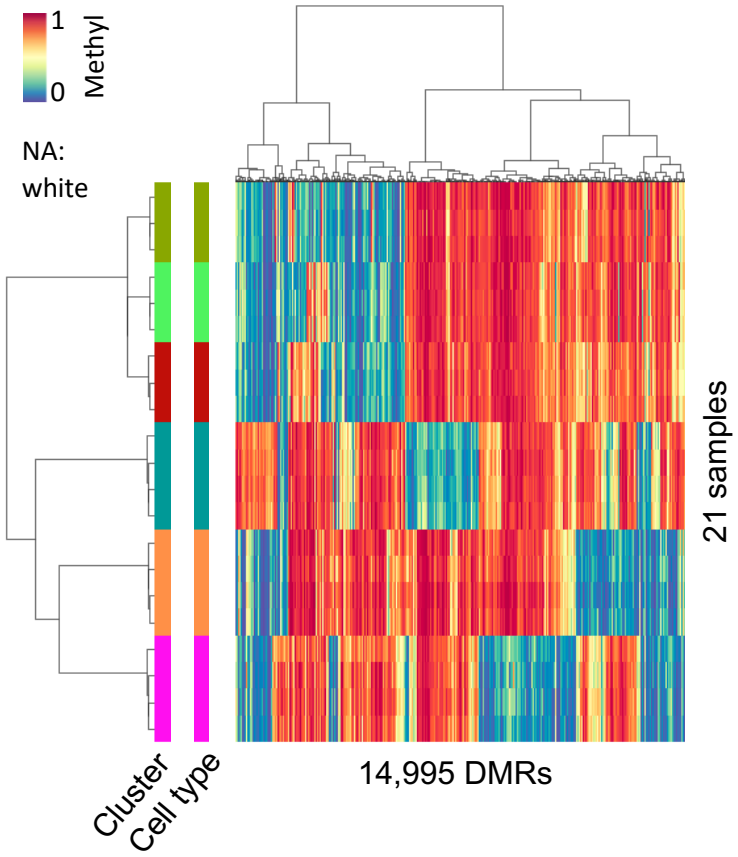
GSEA on cell type-specific expression gene sets from single-cell RNA sequencing:

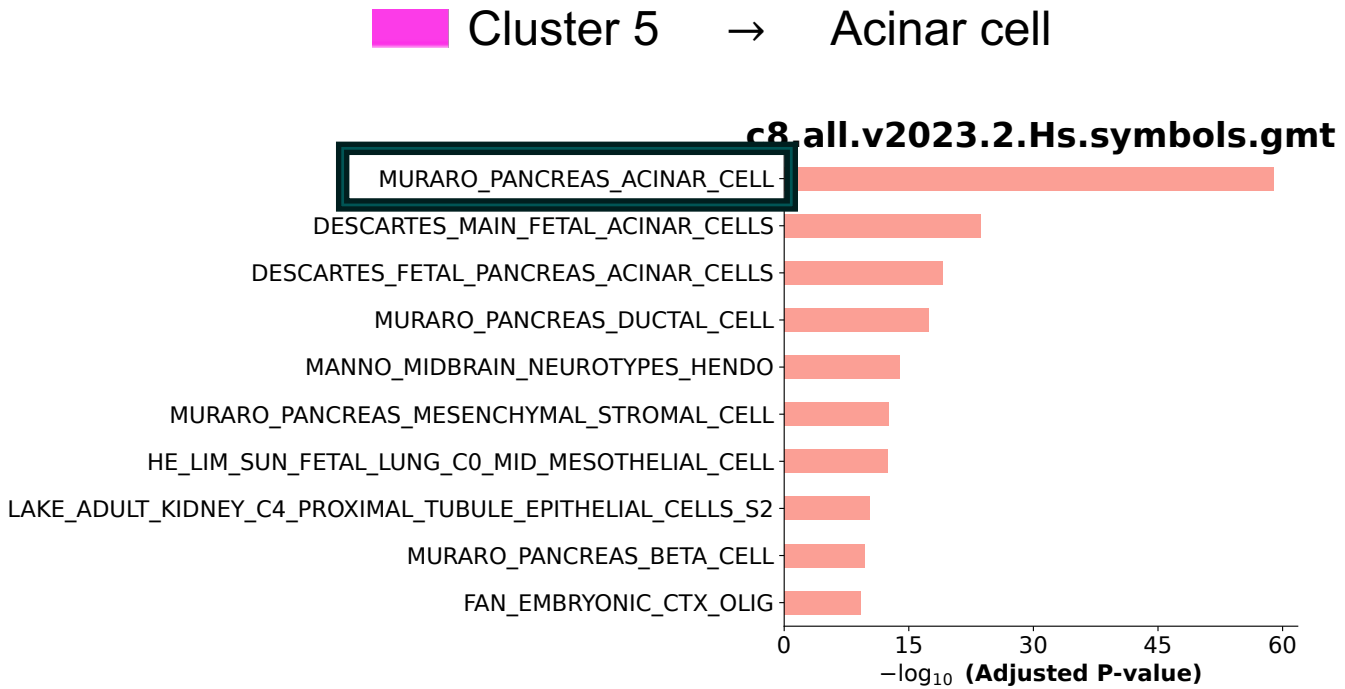




Multi-groups metilene – application

GSEA on cell type-specific expression gene sets from single-cell RNA sequencing:





Summary

- Multi-groups metilene: integrating segmentation and clustering leads to fast and accurate DMR identification among multiple groups.

- DMRs identified by multi-groups metilene (de novo mode) could be used to find biological meaningful clusters.

- Downstream clustering (currently HDBSCAN) might be further improved.

Acknowledgement

Supervision:

Helene Kretzmer

Steve Hoffmann (Leibniz Institute on Aging)

Lab members:

Sara Hetzel

Rosaria Tornisiello

Mara Steiger

Isabelle Kraus

Thank you!