Swiss Institute of Bioinformatics

BIOLOGY-INFORMED MULTIOMICS DATA INTEGRATION AND VISUALIZATION

Normalization and Visualization

Deepak Tanwar

June 16-17, 2025

elixir SWITZERLAND

Learning outcomes

- Information from multi-omics datasets
- Heatmap vs EnrichedHeatmap
- Methods for normalizing data in the target regions

ATAC-seq data snapshot

To assess genome-wide chromatin accessibility.

seqnames	start	end	strand	Symbol	distanceTSS	Group1	Group2
chr1	1	70	+	Gene1	5000	11	21
chr1	100	400	-	Gene1	5000	10	100
chr1	200	290	+	Gene3	2000	200	1000
chr2	300	500	-	Gene4	3000	400	1300
chr2	20	100	+	Gene5	20000	20	120
chr3	40	200	-	Gene6	40000	540	40
chr4	15	150	-	Gene7	150	1500	15

ChIP-seq data snapshot

To investigate the interaction between proteins and DNA

seqnames	start	end	strand	Symbol	distanceTSS	Group1	Group2
chr1	11	70	+	Gene1	5000	11	21
chr1	200	400	-	Gene1	5000	10	100
chr1	100	290	+	Gene3	2000	200	1000
chr2	100	500	-	Gene4	3000	400	1300
chr2	50	100	+	Gene5	20000	20	120
chr3	40	200	-	Gene6	40000	540	40
chr4	5	150	-	Gene7	150	1500	15

WGBS data snapshot

To determine the methylation state of the genome.

seqnames	start	end	strand	Symbol	distanceTSS	Group1	Group2
chr1	1	1	+	Gene1	5000	1.0	0.0
chr1	100	100	-	Gene1	5000	0.0	0.0
chr1	200	200	+	Gene3	2000	0.7	1.0
chr2	300	300	-	Gene4	3000	0.4	0.3
chr2	20	20	+	Gene5	20000	0.2	0.1
chr3	40	40	-	Gene6	40000	1.0	1.0
chr4	15	15	-	Gene7	150	0.0	1.0

RNA-seq data snapshot

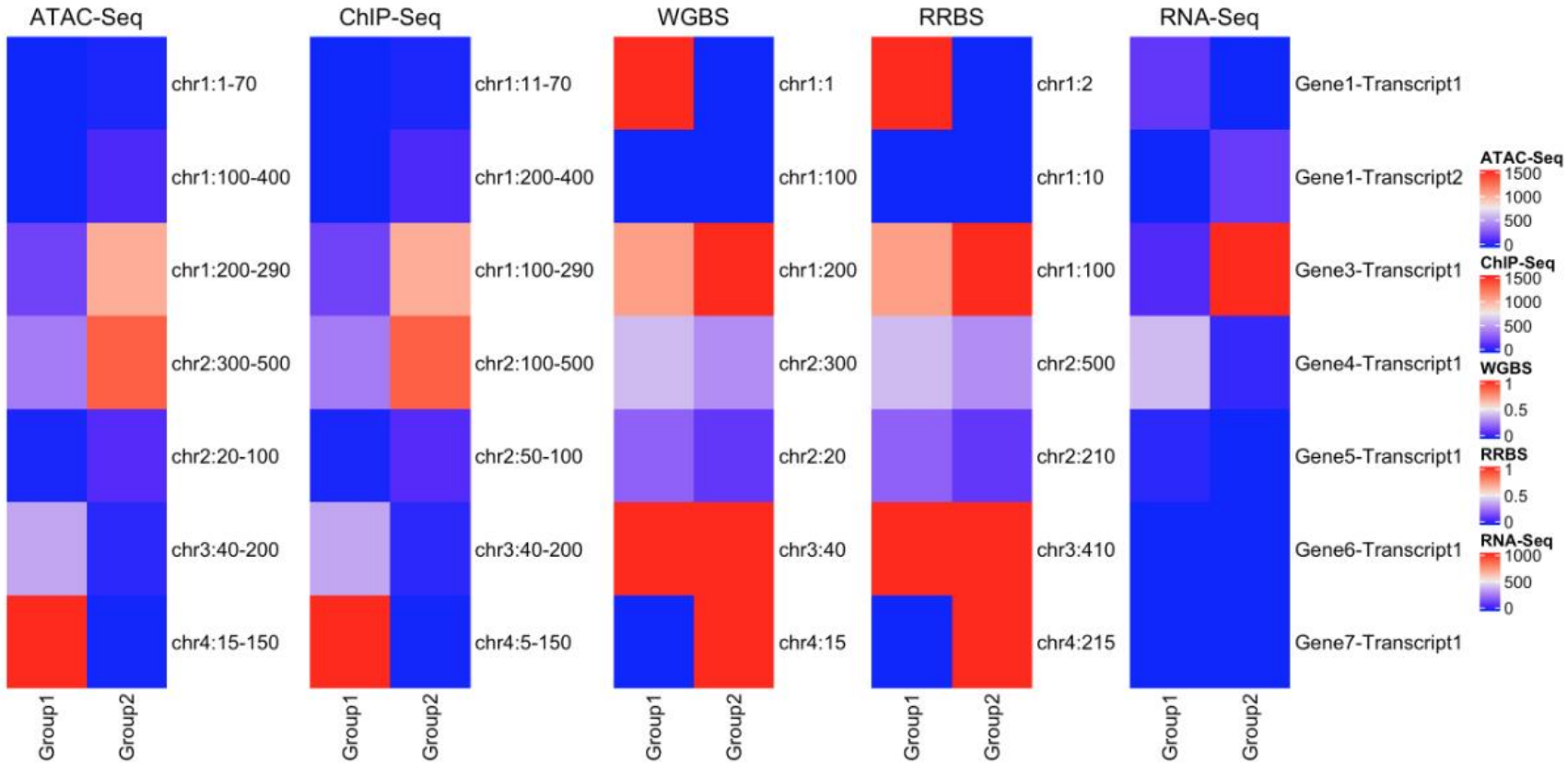
To analyze expression across the transcriptome.

Gene	Transcript	seqnames	start	end	Group1	Group2
Gene1	Transcript1	chr1	1	1000	100	0
Gene1	Transcript2	chr1	100	12000	0	110
Gene3	Transcript1	chr1	200	500	70	1000
Gene4	Transcript1	chr2	300	900	400	30
Gene5	Transcript1	chr2	20	2000	20	1
Gene6	Transcript1	chr3	40	4000	1	0
Gene7	Transcript1	chr4	15	150	0	0

What is the best way to visualize these data?

These tables have 100's of 1000's of rows

Heatmaps?



How to properly integrate these data?

Let's say, we want to look around 200bp of TSS!

TSS
(target)

	seqnames	position	strand
Gene_1	chr21	13	+
Gene_2	chr21	46	+
Gene_3	chr21	78	+
Gene_4	chr21	10	+
Gene_5	chr21	45	+

How to properly integrate these data?

Let's say, we want to look around 200bp of TSS!

TSS (target)

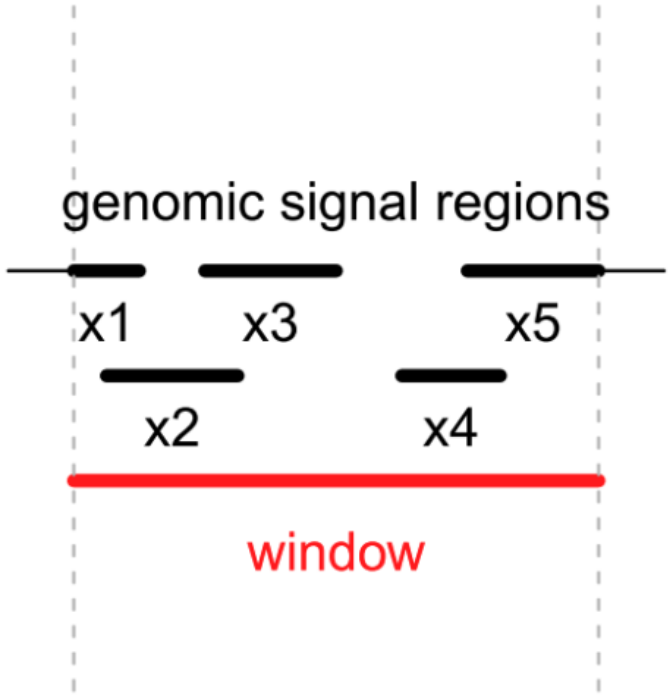
	seqnames	position	strand
Gene_1	chr21	13	+
Gene_2	chr21	46	+
Gene_3	chr21	78	+
Gene_4	chr21	10	+
Gene_5	chr21	45	+

ChIP-Seq data (genomic signal)

seqnames	start	end	strand	Group1
chr21	11	70	+	11
chr21	200	400	-	10
chr21	100	290	+	200
chr21	100	500	-	400
chr21	50	100	+	20

Averaging methods to summarize the signals

Averaging model



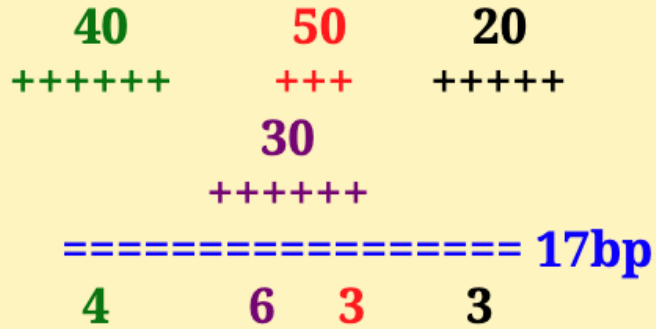
The **red line** represents one window in the target regions or in the flanking regions when normalizing genomic signals to target regions. **Black lines** represent genomic signals that overlap to the given window.

Averaging methods to summarize the signals

absolute

$$v_a = \frac{\sum_{i=1}^{n} x_i}{n}$$

Calculates the mean value from all signal regions regardless of their width.



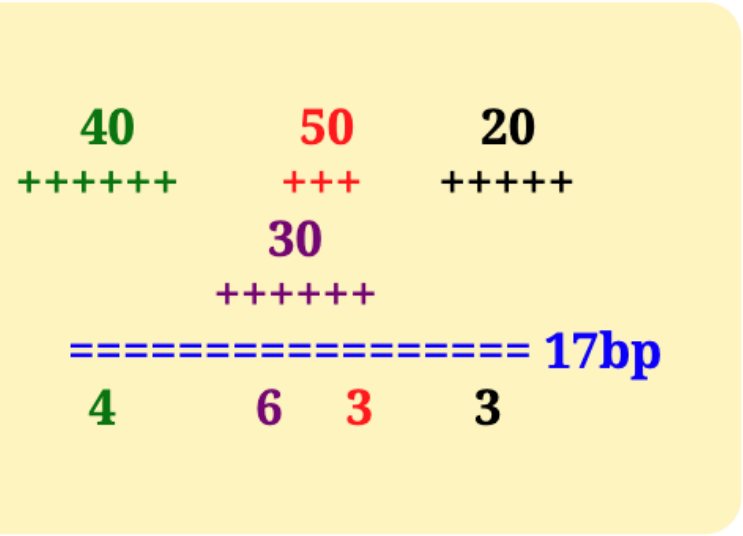
4bp not overlapping to any region

$$v_a = \frac{40 + 30 + 50 + 20}{4}$$

Averaging methods to summarize the signals

weighted

$$v_w = \frac{\sum_{i=1}^{n} x_i w_i}{\sum_{i=1}^{n} w_i}$$



Calculates the mean value from all signal regions weighted by the width of their intersections.

4bp not overlapping to any region

$$v_w = \frac{40\times4 + 30\times6 + 50\times3 + 20\times3}{4 + 6 + 3 + 3}$$

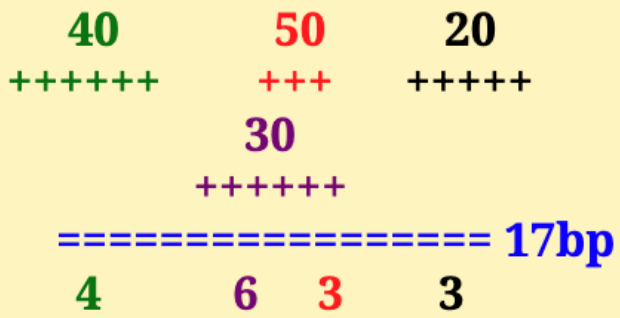
Averaging methods to summarize the signals

WO

$$v_{w0} = \frac{\sum_{i=1}^{n} x_i w_i}{W + W'}$$

Calculates the `weighted mean` between the intersected and un-intersected parts.

$$\left(W = \sum_{i=1}^{n} w_i\right),$$
$$\left(W\prime = \sum_{i=1}^{n} w\prime_i\right)$$



4bp not overlapping to any region

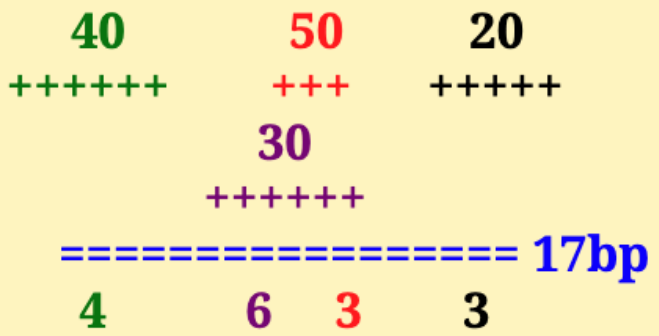
$$v_{w0} = \frac{40\times4 + 30\times6 + 50\times3 + 20\times3}{4 + 6 + 3 + 3 + 4}$$

Averaging methods to summarize the signals

coverage

$$v_c = \frac{\sum_i^n x_i w_i}{L}$$

The mean signal averaged by the width of the window. L is the width of the window itself.



4bp not overlapping to any region

$$v_c = \frac{40\times4 + 30\times6 + 50\times3 + 20\times3}{17}$$

How to use these methods?

Gu *et al. BMC Genomics* (2018) 19:234
https://doi.org/10.1186/s12864-018-4625-x

BMC Genomics

SOFTWARE Open Access

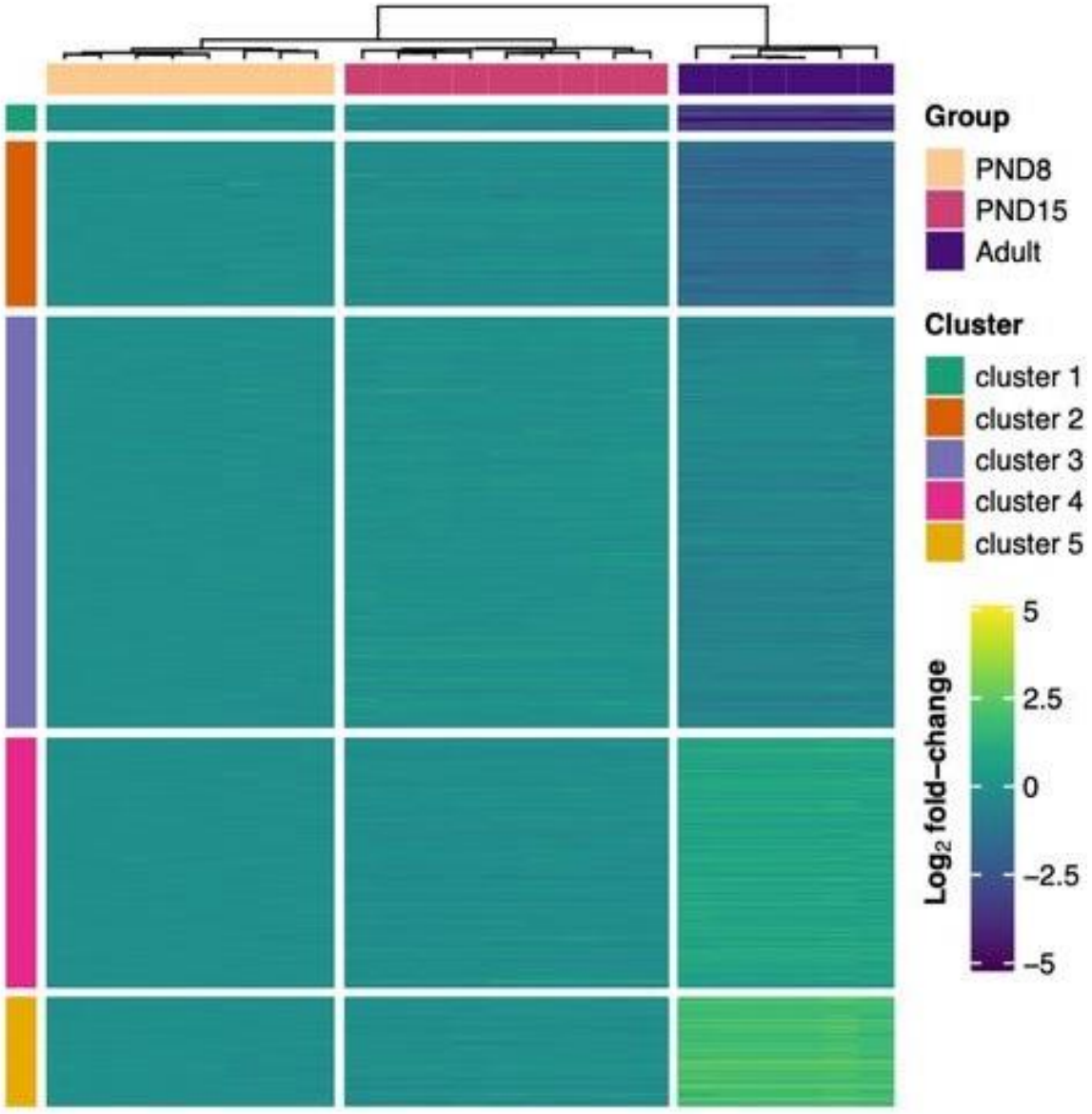
EnrichedHeatmap: an R/Bioconductor package for comprehensive visualization of genomic signal associations

Zuguang Gu[1,2*], Roland Eils[1,2,3], Matthias Schlesner[1] and Naveed Ishaque[1,2]

Which method to be used for different data?

Assay	Signal Representation	Recommended `mean_mode`	Key Considerations
ATAC-seq	Peaks (binary)	`"coverage"`	Shows fraction of window covered by peaks.
	Coverage (numeric)	`"w0"`	Averages signal intensity, weighted by overlap. Good for BigWig data.
ChIP-seq	Peaks (binary)	`"coverage"`	Shows fraction of window covered by peaks.
	Coverage (numeric)	`"w0"`	Averages signal intensity, weighted by overlap. Good for BigWig data. Set `empty_value = 0` (or background).
WGBS	CpG methylation (numeric)	`"absolute"`	Averages methylation values of individual CpG sites within the window. **Crucial to set** `value_column`. **Highly recommend** `smooth = TRUE` **and** `empty_value = NA` for better visualization due to sparse CpG distribution.

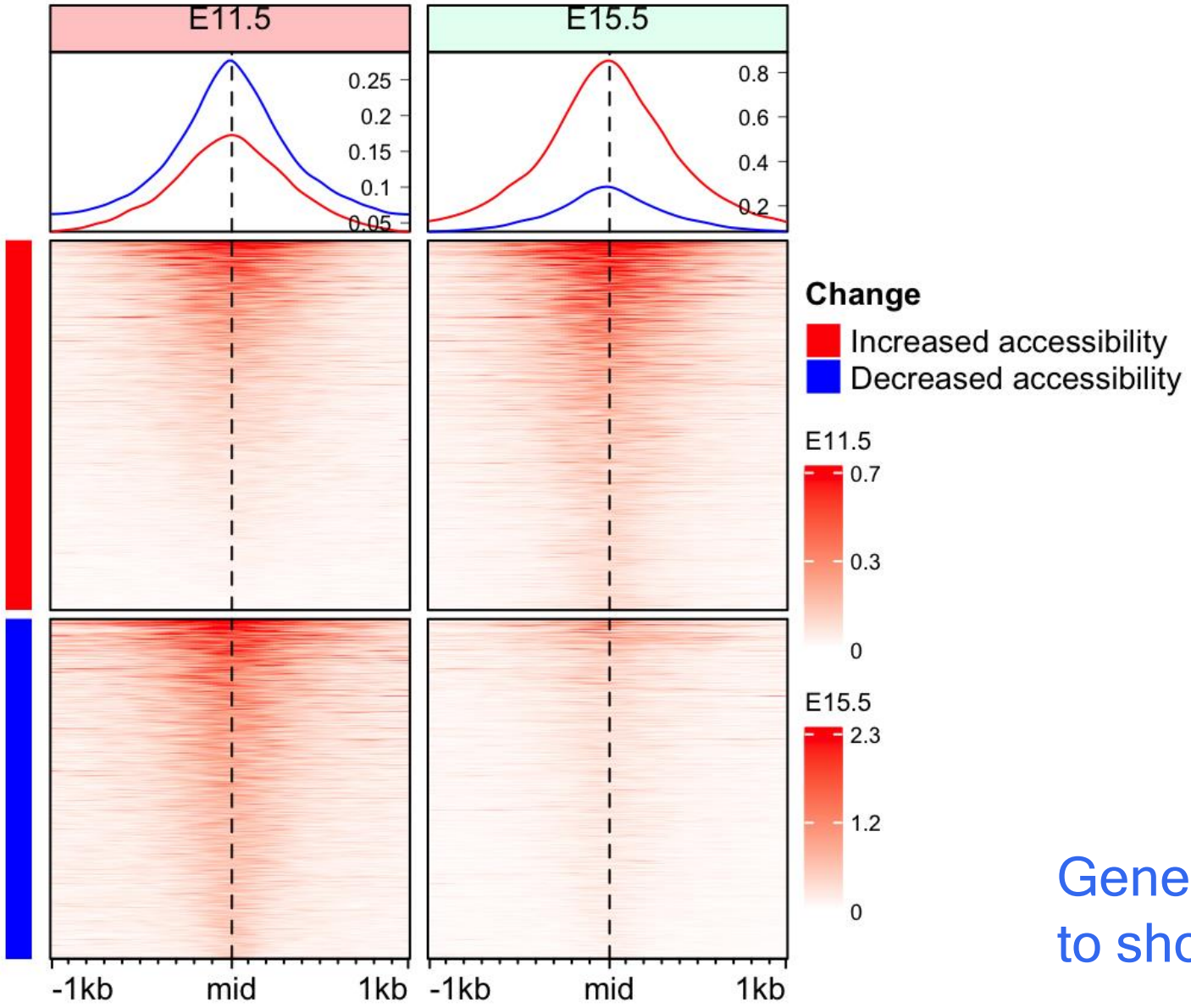
Understanding Heatmap



Heatmap

- Matrix
- Color
- Rows
- Rows annotation
- Columns
- Columns annotation
- Color scale
- Clustering
- Rows and columns order

Understanding EnrichedHeatmap

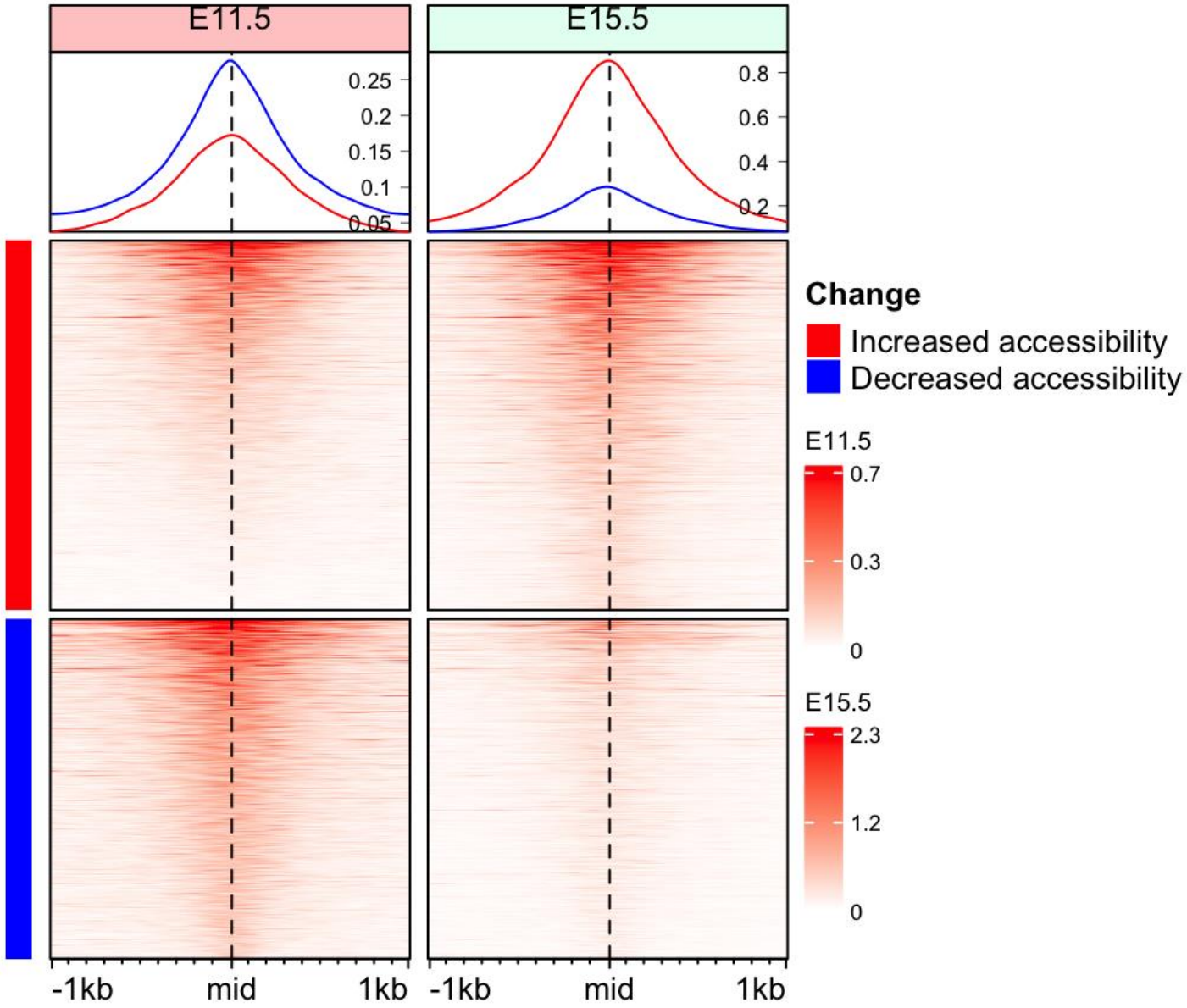


EnrichedHeatmap

- Matrix
- Color
- Rows
- Rows annotation
- Profile plot
- Columns title
- Color scale
- Rows order
- Axis

Generally, replicates are merged into one to show average differences

Understanding EnrichedHeatmap

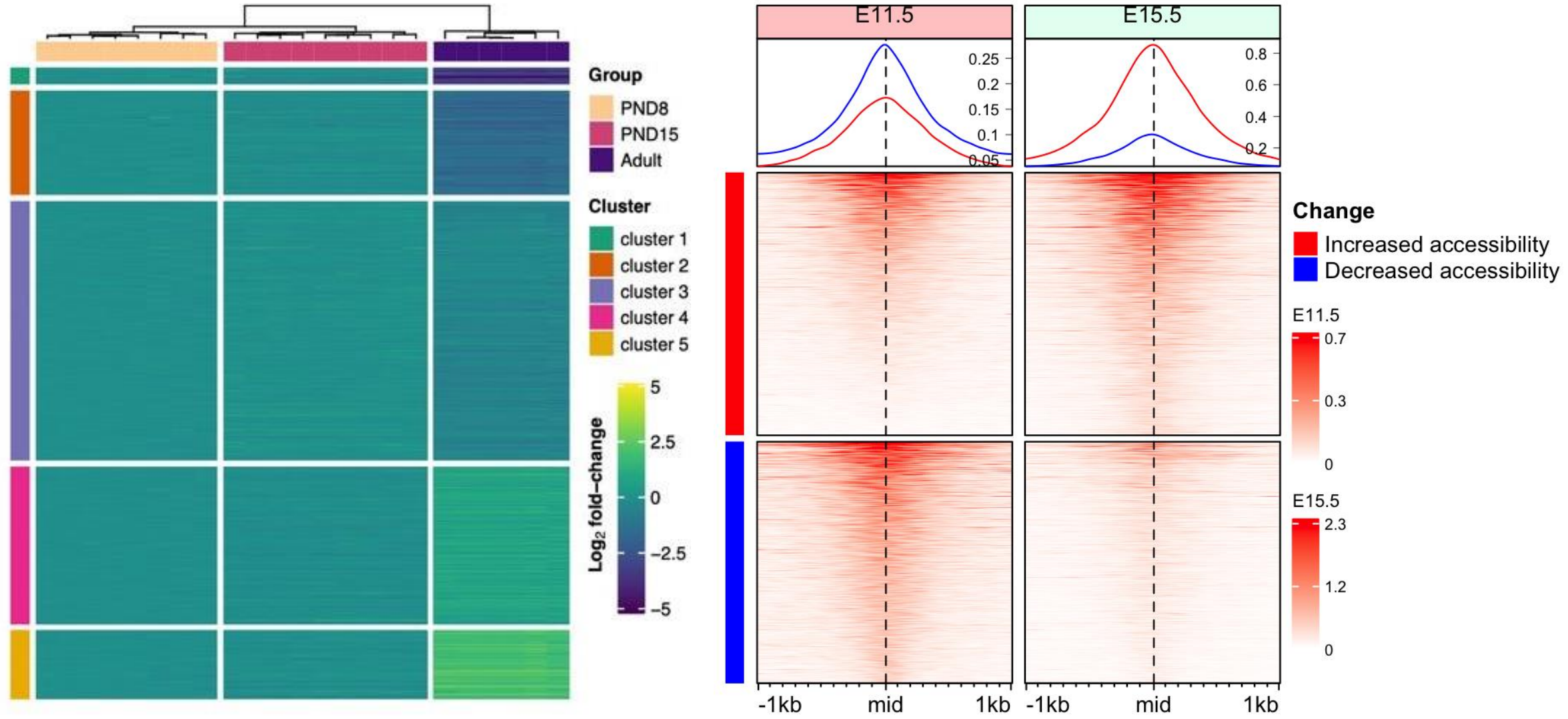


E15.5

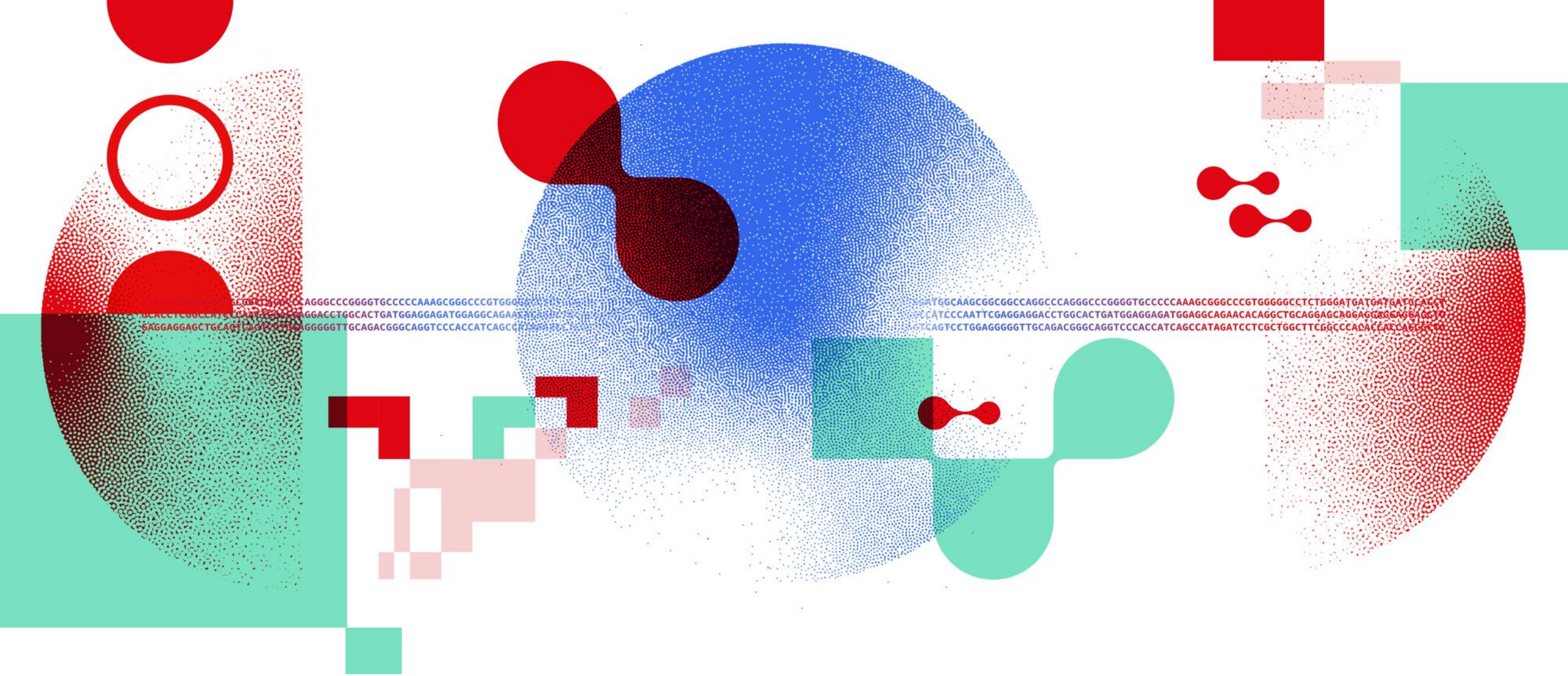
- Color scale: 0 – 2.3
- Profile plot scale: 0-0.8

Why?

Best way to represent complex data. Other ways?



Exercise 2 & 3



Thank you

DATA SCIENTISTS FOR LIFE

sib.swiss

elixir SWITZERLAND

SIB