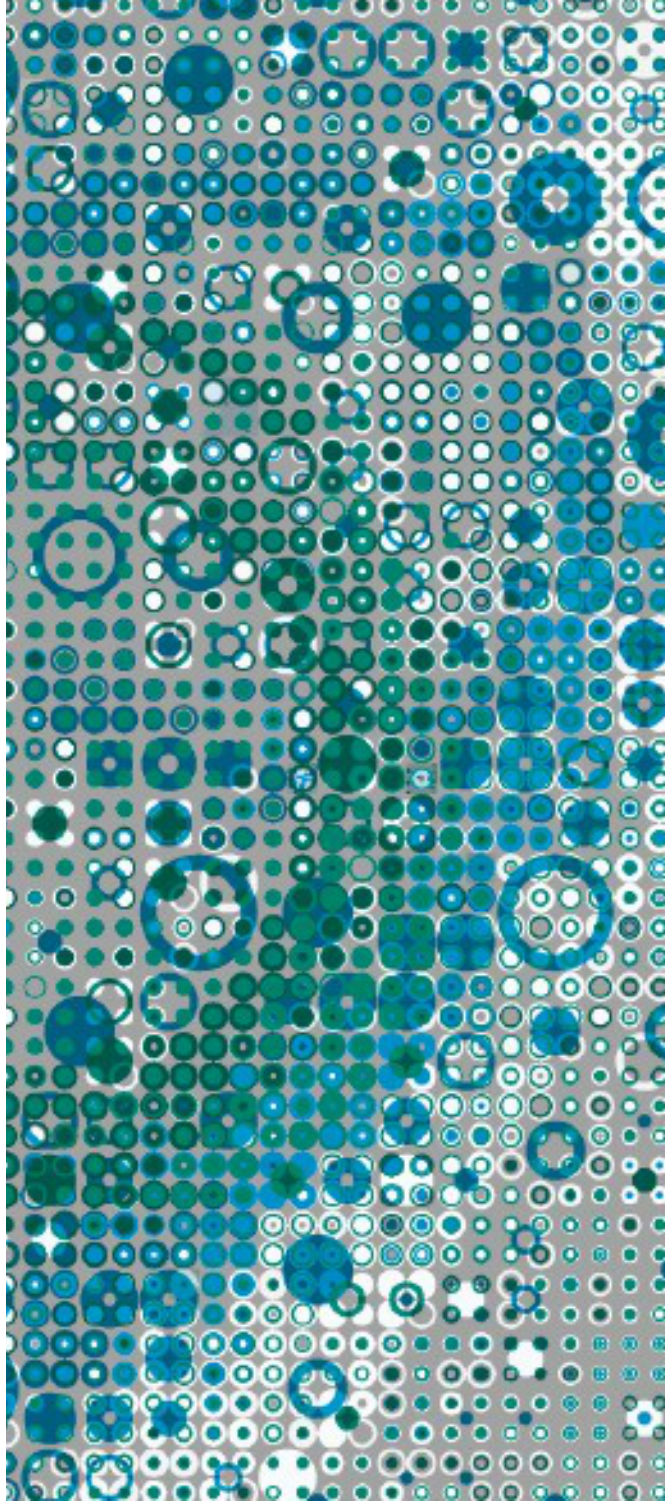
Sources and potential applications of coverage biases in single-cell WGBS-seq

Emma Dann, Master internship
16/04/2018

Hubrecht Institute

Developmental Biology and Stem Cell Research



About me

- From Ferrara, IT
- B.Sc. in Biotechnology - University of Trento, IT:
  - Internship and thesis project at CIBIO, Laboratory of Computational Oncology, Prof. Demichelis
- M.Sc. in Cancer, Stem cells and Developmental biology (Bioinformatics profile) - Utrecht University, NL
  - Major internship at Hubrecht Institute, Van Oudenaarden Lab

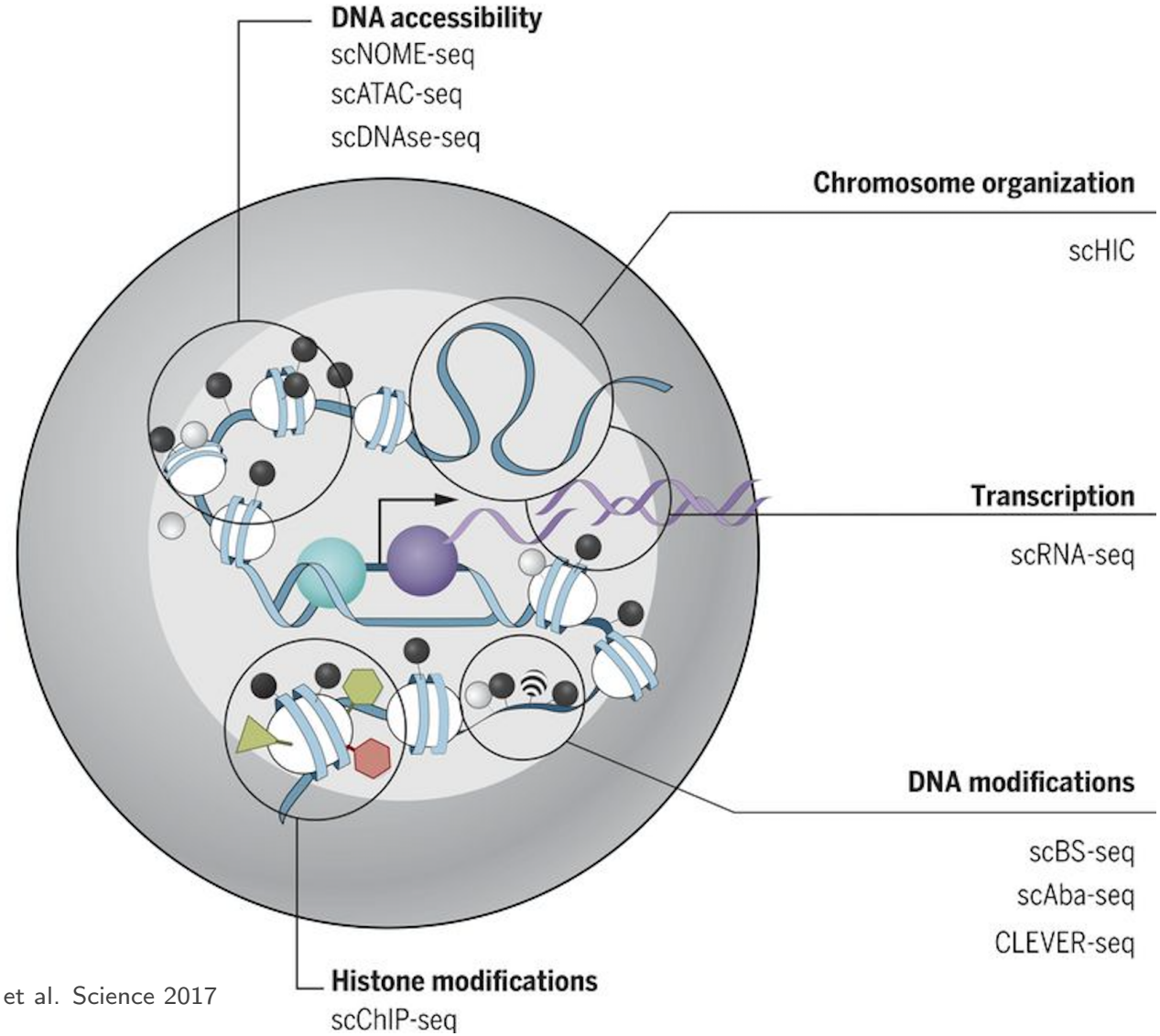
About me

- From Ferrara, IT
- B.Sc. in Biotechnology - University of Trento, IT:
  - Internship and thesis project at CIBIO, Laboratory of Computational Oncology, Prof. Demichelis
- M.Sc. in Cancer, Stem cells and Developmental biology (Bioinformatics profile) - Utrecht University, NL
  - Major internship at Hubrecht Institute, Van Oudenaarden Lab

About me

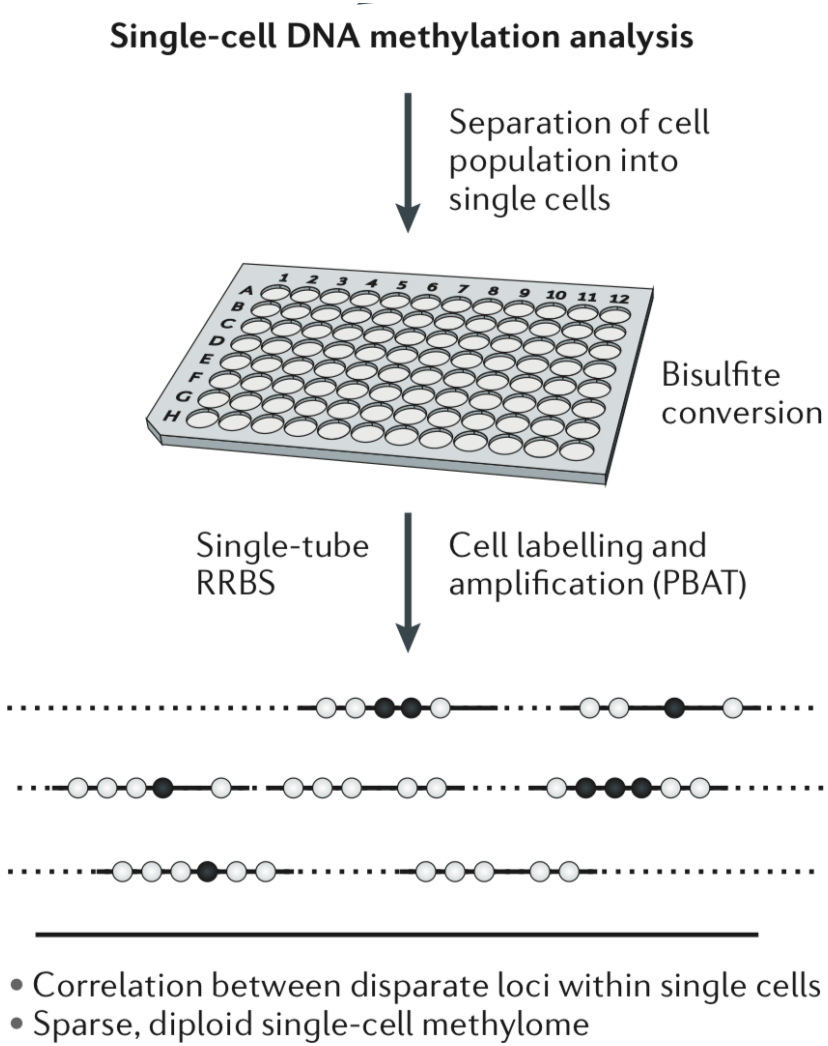
- From Ferrara, IT
- B.Sc. in Biotechnology - University of Trento, IT:
  - Internship and thesis project at CIBIO, Laboratory of Computational Oncology, Prof. Demichelis
- M.Sc. in Cancer, Stem cells and Developmental biology (Bioinformatics profile) - Utrecht University, NL
  - Major internship at Hubrecht Institute, Van Oudenaarden Lab

Single-cell epigenomics: a quest for throughput and depth



Adapted from Kelsey et al. Science 2017

Bisulfite sequencing in the single-cell

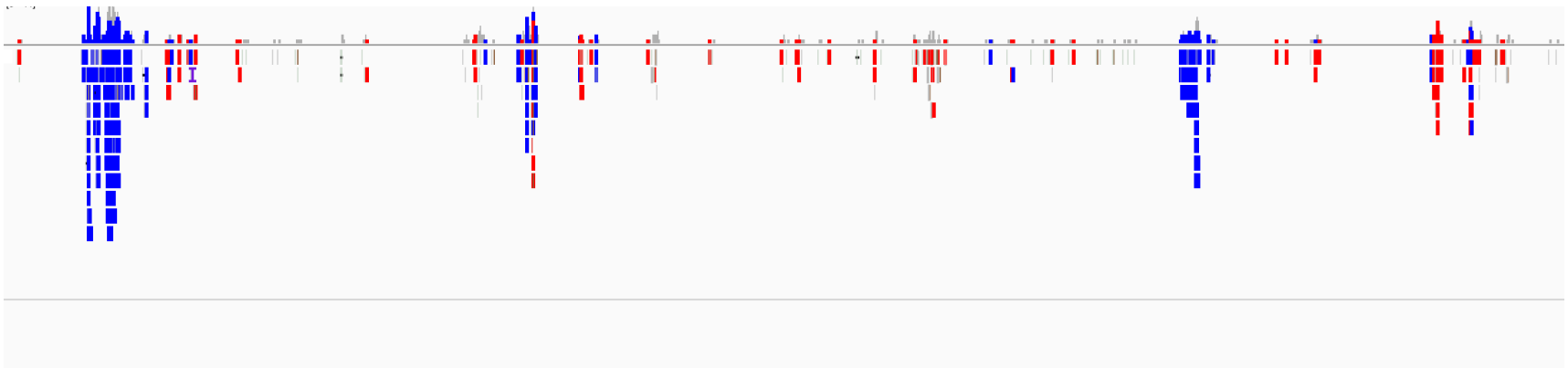


Main issues:

- BS induced fragmentation reduces recovery
- Sparsity of data hinders calling of differential methylation

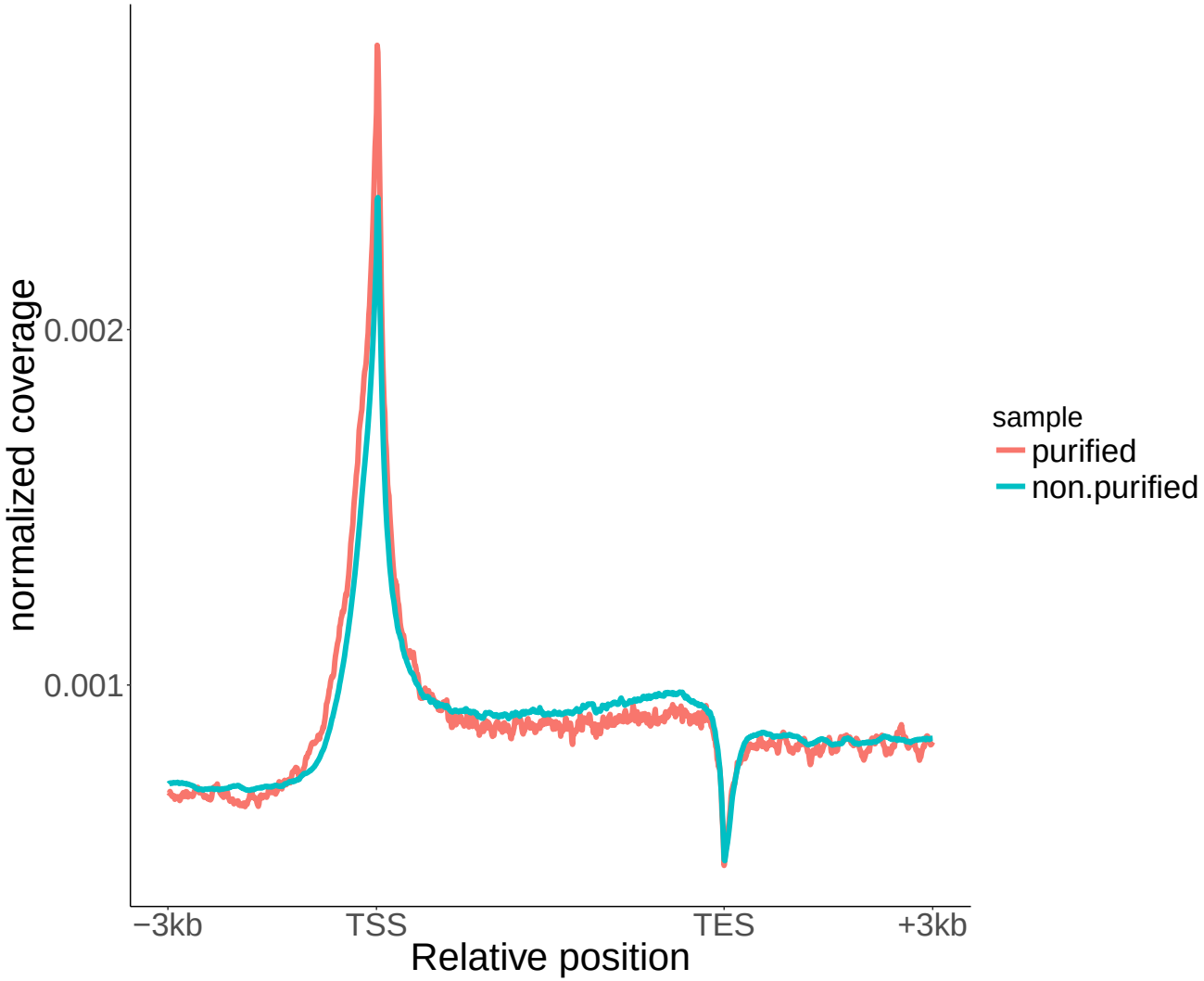
Adapted from Schwartzman, Tanay, Nature Reviews Genetics 2015

Uneven coverage in scBS-seq data

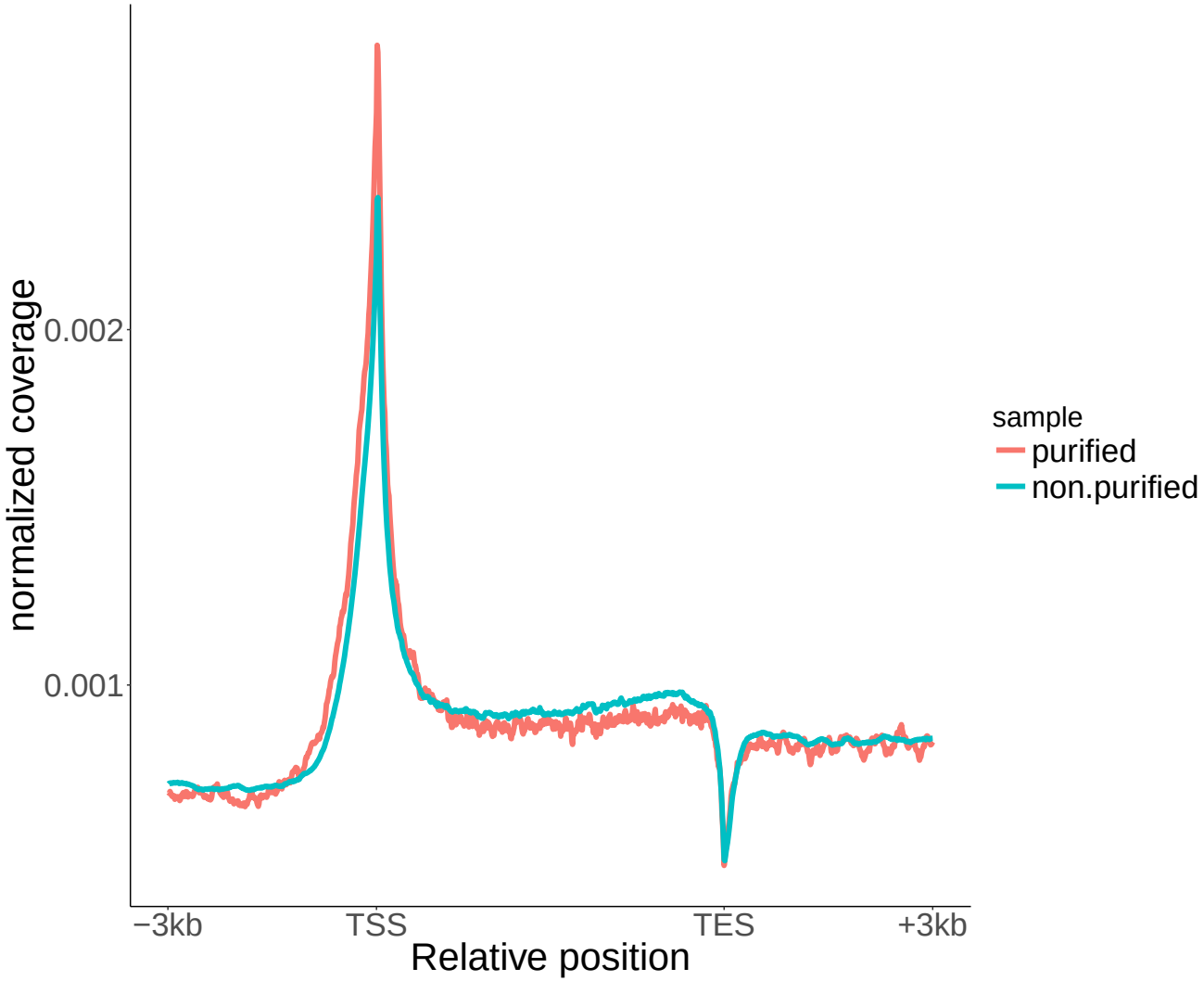


Mouse intestine, 100 cells

Uneven coverage in scBS-seq data

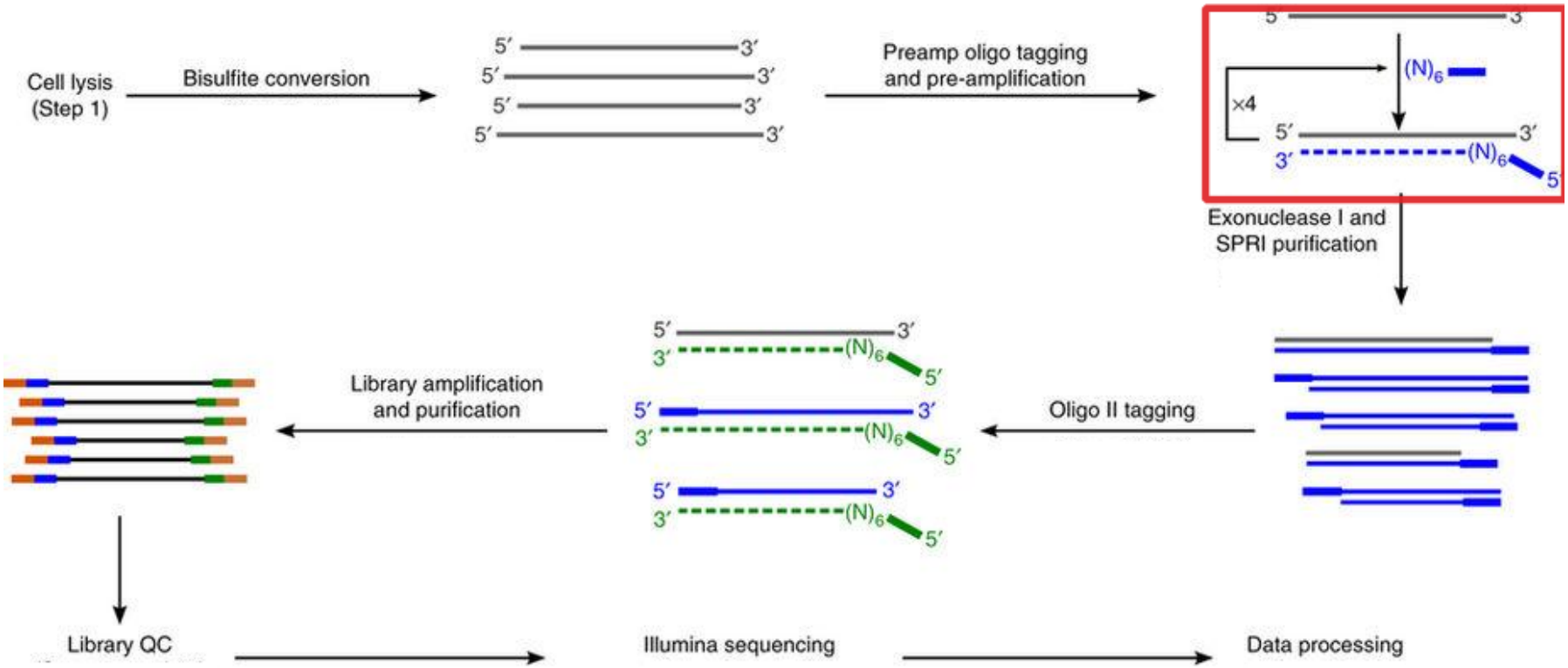


Uneven coverage in scBS-seq data



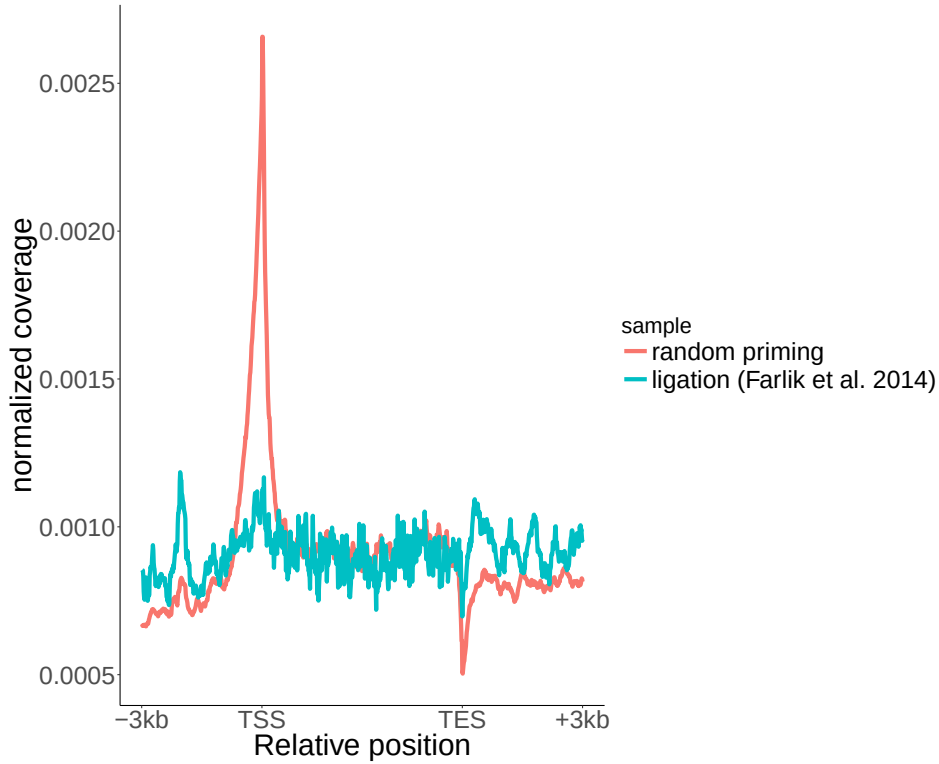
Can we understand (and maybe hack) the source of this coverage bias?

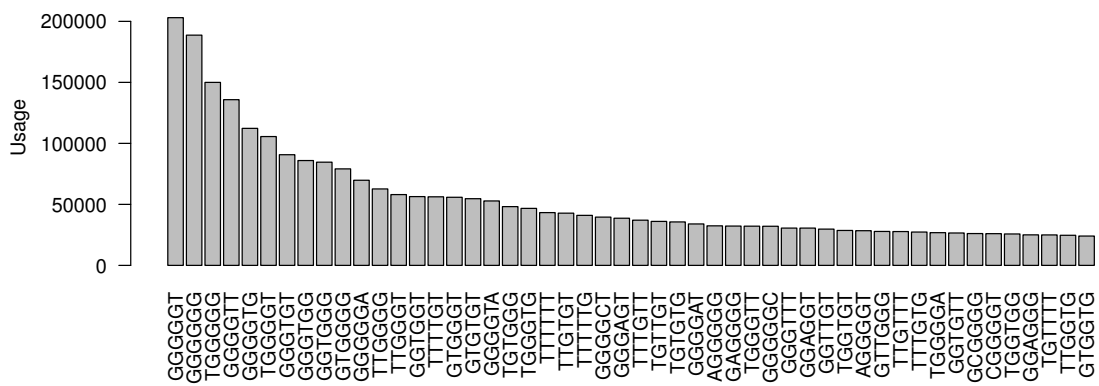
Post-Bisulfite Adaptor Tagging (scPBAT)



Clark et al. Nature Protocols 2017

Random priming as a source of bias





Modelling random hexamer binding

$$p + t \rightleftarrows pt,$$

p = primer
t = free template sequence
pt = bound template sequence

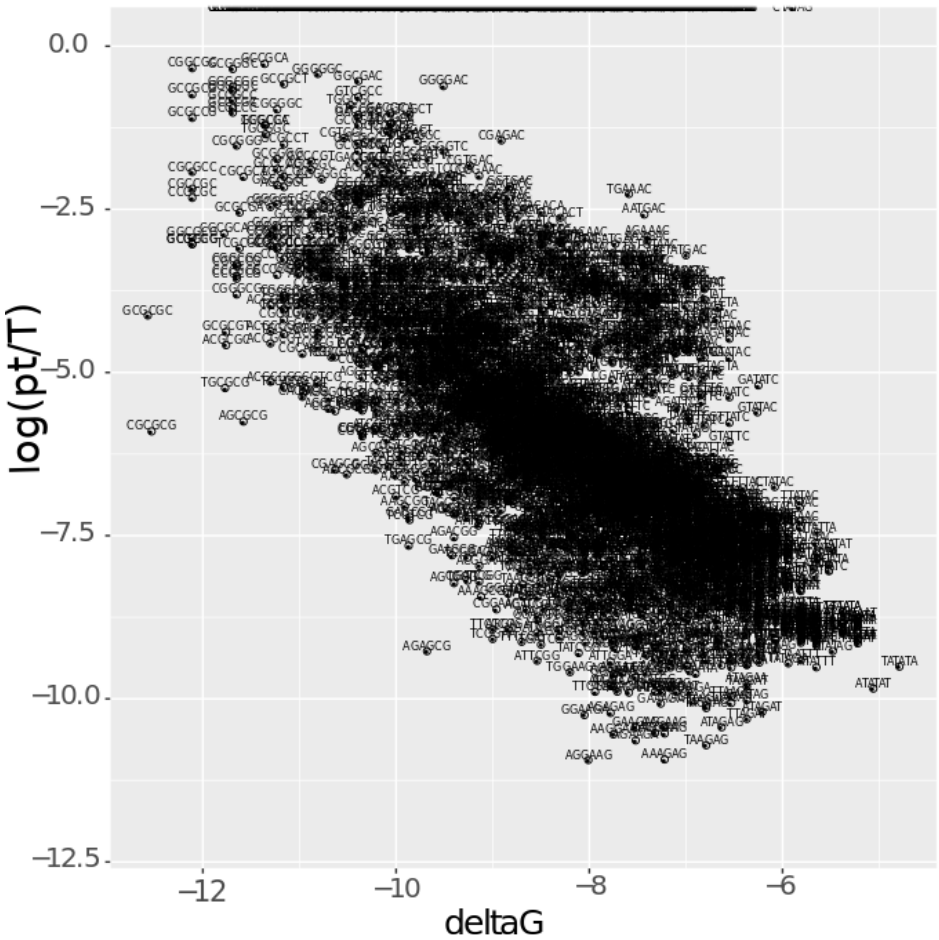
$$\frac{[pt]}{[p][t]} = \exp\left(\beta \Delta G\right).$$

Binding dependant on:

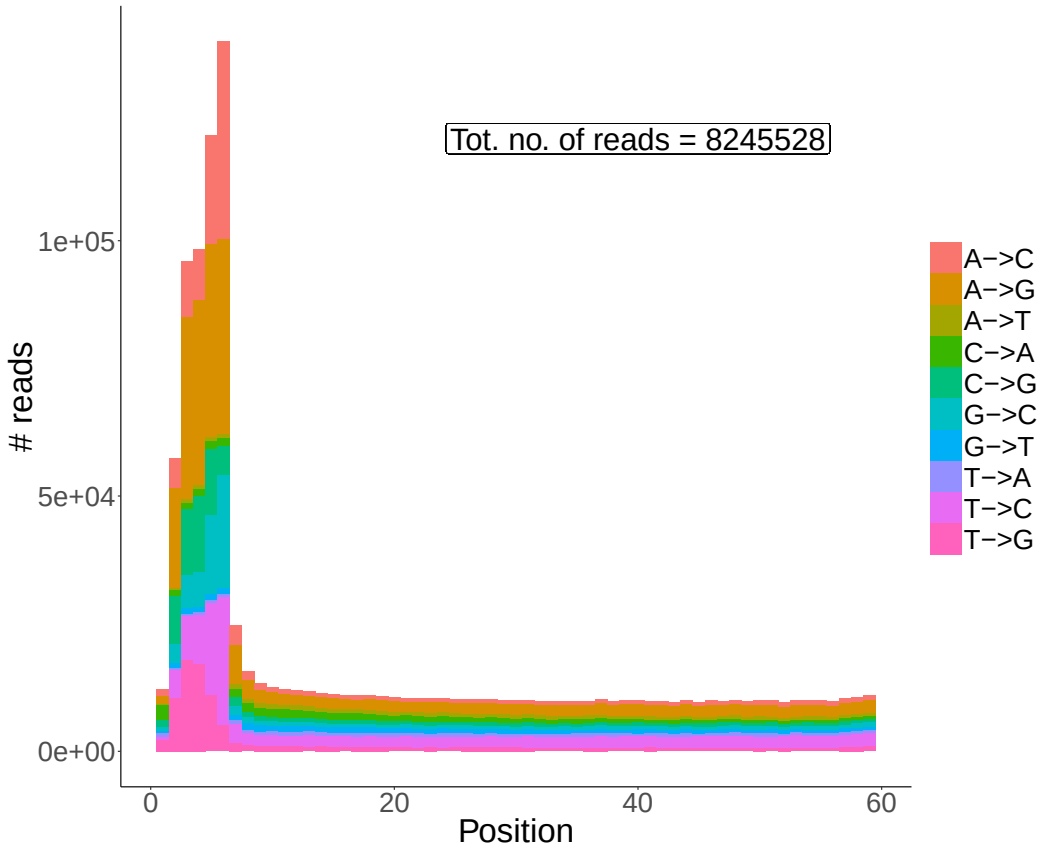
- Total number of template sequences $T = t + pt$
- Binding free energy ΔG

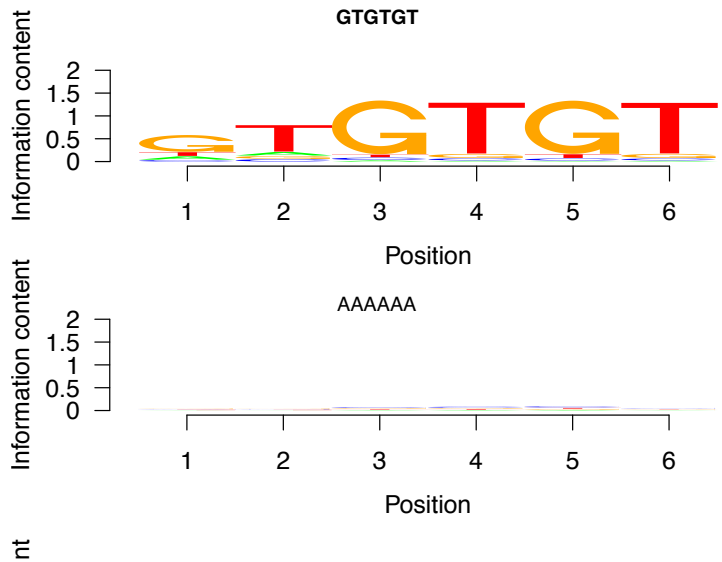
Modelling random hexamer binding

$$\frac{[pt]}{[t]} = [p] \exp(\beta \Delta G).$$



Frequent mismatches at primer binding position in BS seq data





Hexamer binding model 2.0: accounting for all binding events

$$p_i + t_j \rightleftarrows p_i t_j,$$

where $p_i t_j$ is the binding complex. In equilibrium conditions it is satisfied that:

$$\frac{[p_i t_j]}{[p_i][t_j]} = \exp\left(\beta \Delta G_{ij}\right).$$

Hexamer binding model 2.0: accounting for all binding events

$$p_i + t_j \rightleftarrows p_i t_j,$$

where $p_i t_j$ is the binding complex. In equilibrium conditions it is satisfied that:

$$\frac{[p_i t_j]}{[p_i][t_j]} = \exp\left(\beta \Delta G_{ij}\right).$$

How to define coverage?

$$C_j = \sum_i [p_i t_j] = \boxed{T_j \frac{\sum_i [p_i] \exp\left(\beta \Delta G_{ij}\right)}{[1 + \sum_i [p_i] \exp\left(\beta \Delta G_{ij}\right)]}}$$

$T_j = t_j + C_j \longrightarrow$ total abundance of template sequence ;
$[p_i] \longrightarrow$ concentration of primers in the experiments.

Hexamer binding model 2.0: accounting for all binding events

$$p_i + t_j \rightleftarrows p_i t_j,$$

where $p_i t_j$ is the binding complex. In equilibrium conditions it is satisfied that:

$$\frac{[p_i t_j]}{[p_i][t_j]} = \exp\left(\beta \Delta G_{ij}\right).$$

How to define coverage?

$$C_j = \sum_i [p_i t_j] = \boxed{T_j \frac{\sum_i [p_i] \exp\left(\beta \Delta G_{ij}\right)}{[1 + \sum_i [p_i] \exp\left(\beta \Delta G_{ij}\right)]}}$$

$T_j = t_j + C_j \longrightarrow$ total abundance of template sequence ;
$[p_i] \longrightarrow$ concentration of primers in the experiments.

What about ΔG_{ij} ?

Computing ΔG_{ij} from sequencing data

$$\frac{[p_i t_j]}{[p_i][t_j]} = \exp\left(\beta \Delta G_{ij}\right).$$

Reference genome

Aligned reads

Computing ΔG_{ij} from sequencing data

$$\frac{[p_i t_j]}{[p_i][t_j]} = \exp\left(\beta \Delta G_{ij}\right).$$

Reference genome →

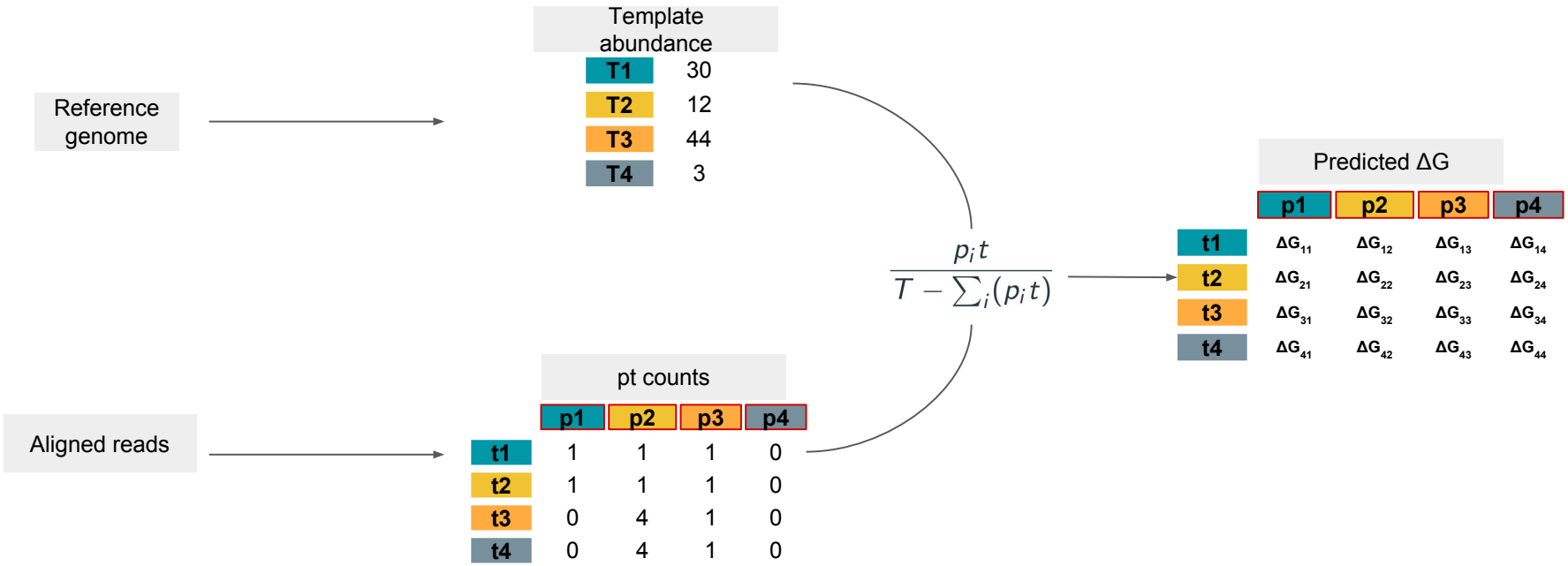
Template abundance	
T1	30
T2	12
T3	44
T4	3

Aligned reads →

pt counts	p1	p2	p3	p4
t1	1	1	1	0
t2	1	1	1	0
t3	0	4	1	0
t4	0	4	1	0

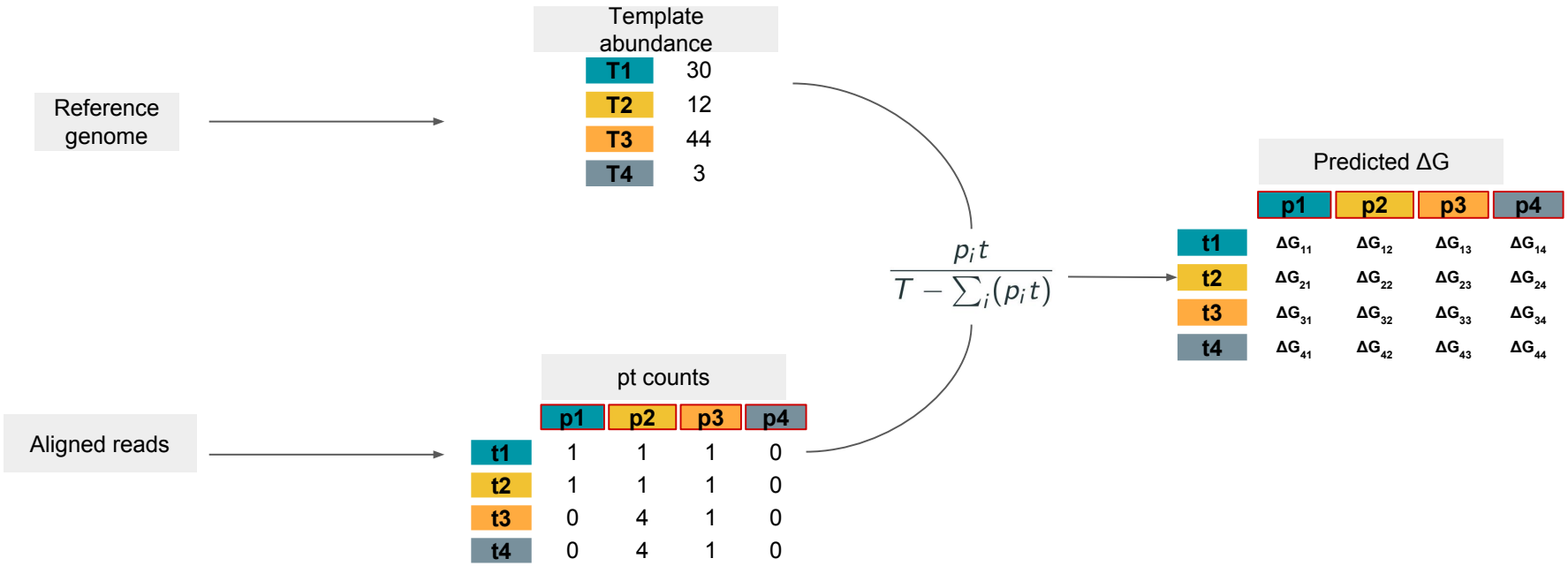
Computing ΔG_{ij} from sequencing data

$$\frac{[p_i t_j]}{[p_i][t_j]} = \exp\left(\beta \Delta G_{ij}\right).$$



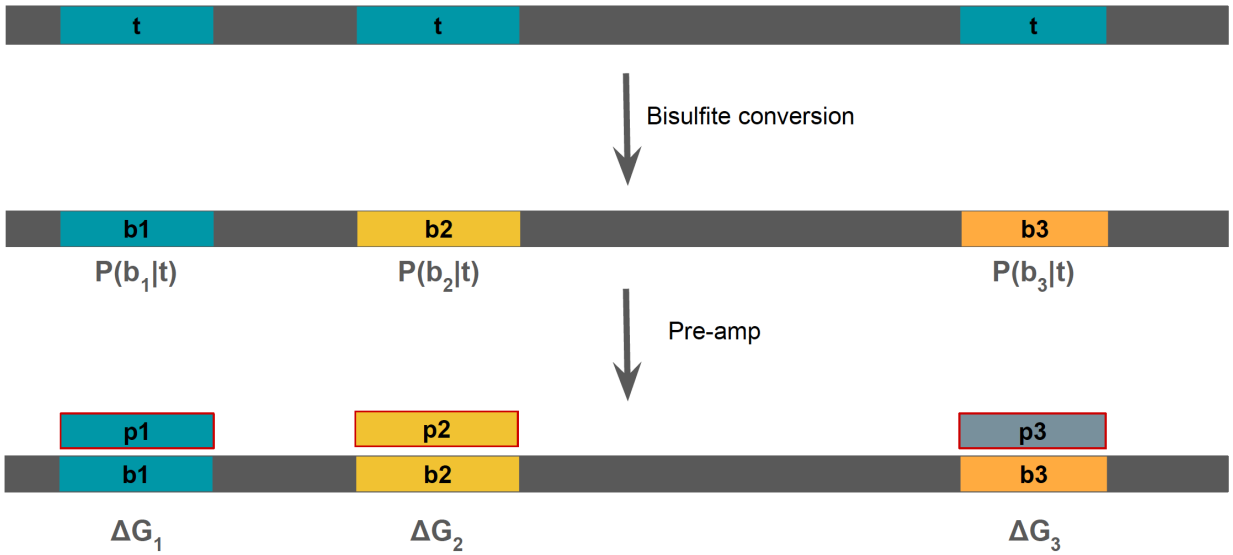
Computing ΔG_{ij} from sequencing data

$$\frac{[p_i t_j]}{[p_i][t_j]} = \exp\left(\beta \Delta G_{ij}\right).$$



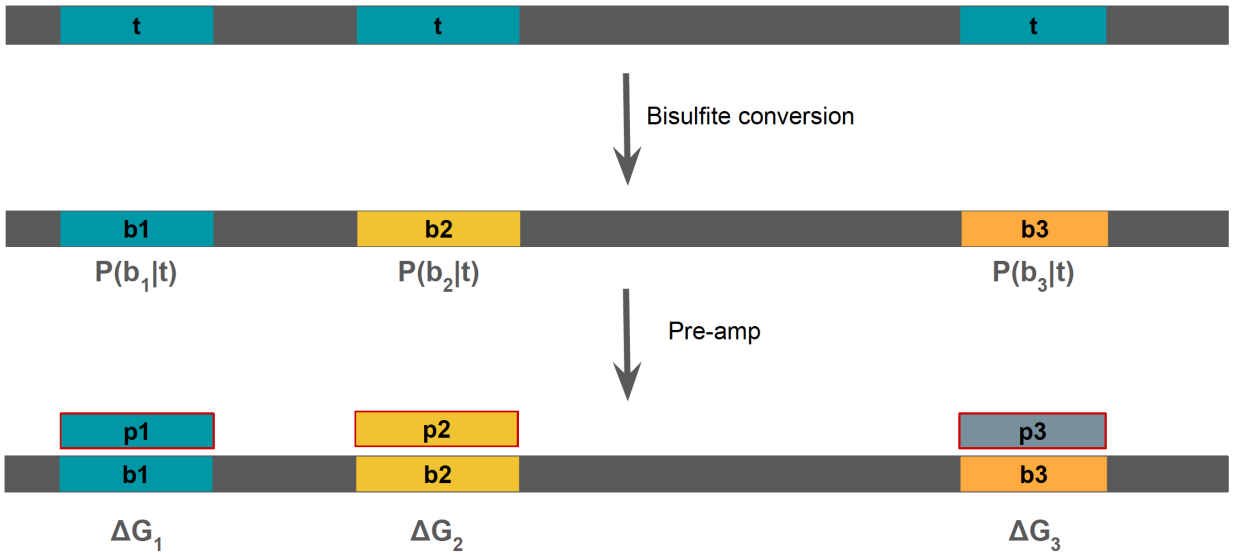
How to use this approach on bisulfite converted genome?

Prediction on BS-seq data



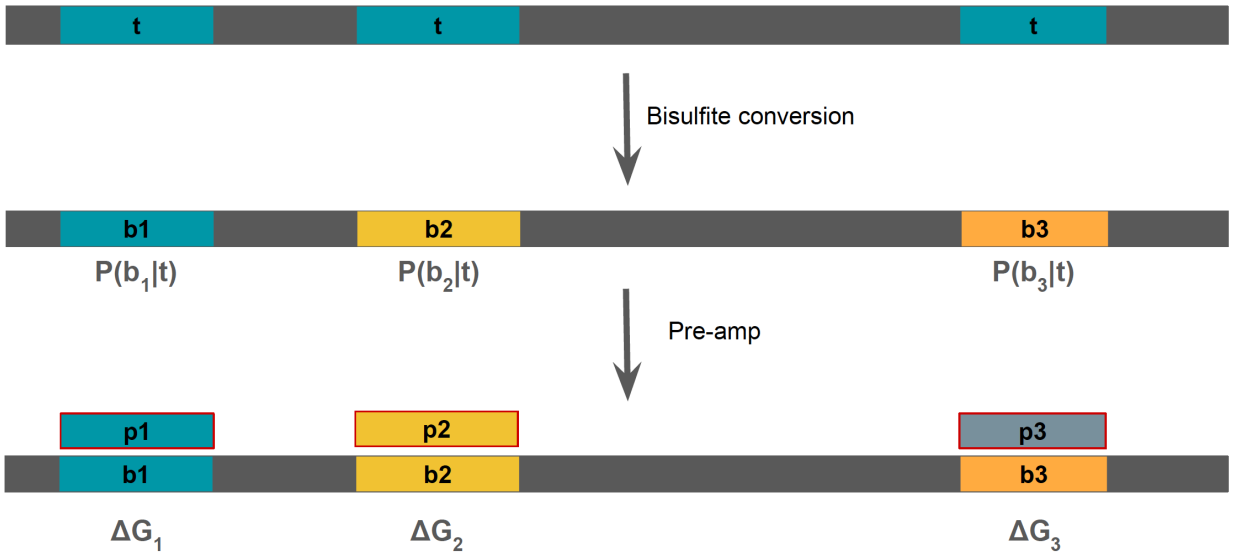
$$[p_i\, t] = [b_1\, p_i] + [b_2\, p_i] + [b_3\, p_i] \ldots$$

Prediction on BS-seq data



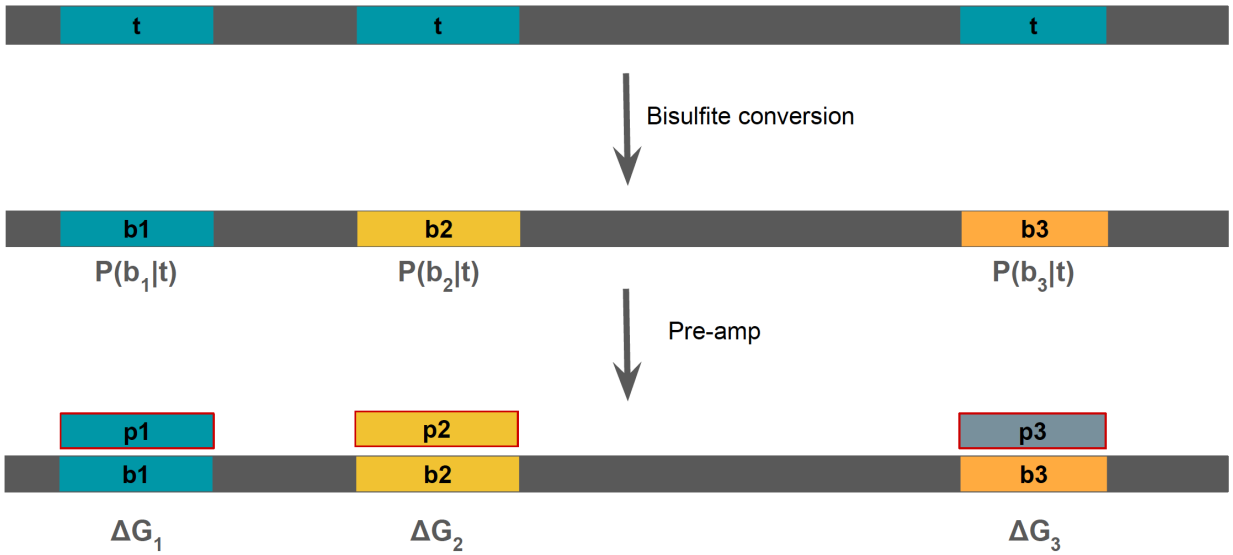
$$[p_i\, t] = [b_1\, p_i] + [b_2\, p_i] + [b_3\, p_i] \ldots$$
$$= [b_1][p_i] \exp(\beta \Delta G) + [b_2][p_i] \exp(\beta \Delta G) + \ldots$$

Prediction on BS-seq data



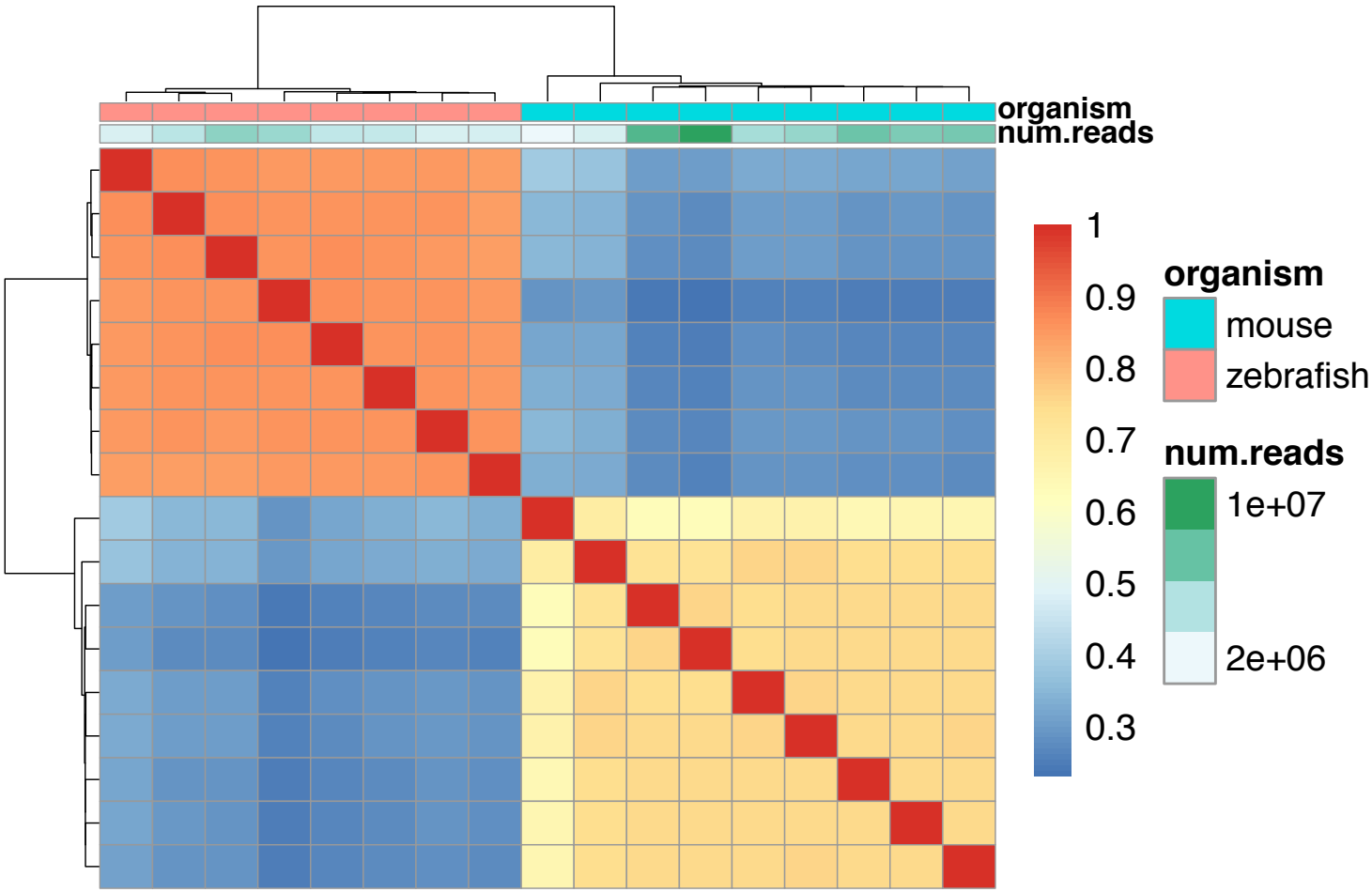
$$
\begin{aligned}
[p_i\, t] &= [b_1 p_i] + [b_2 p_i] + [b_3 p_i] \ldots \\
&= [b_1][p_i] \exp(\beta \Delta G) + [b_2][p_i] \exp(\beta \Delta G) + \ldots \\
&= [t][p_i] \sum_n \exp(\beta \Delta G)\, P(b_n|t)
\end{aligned}
$$

Prediction on BS-seq data



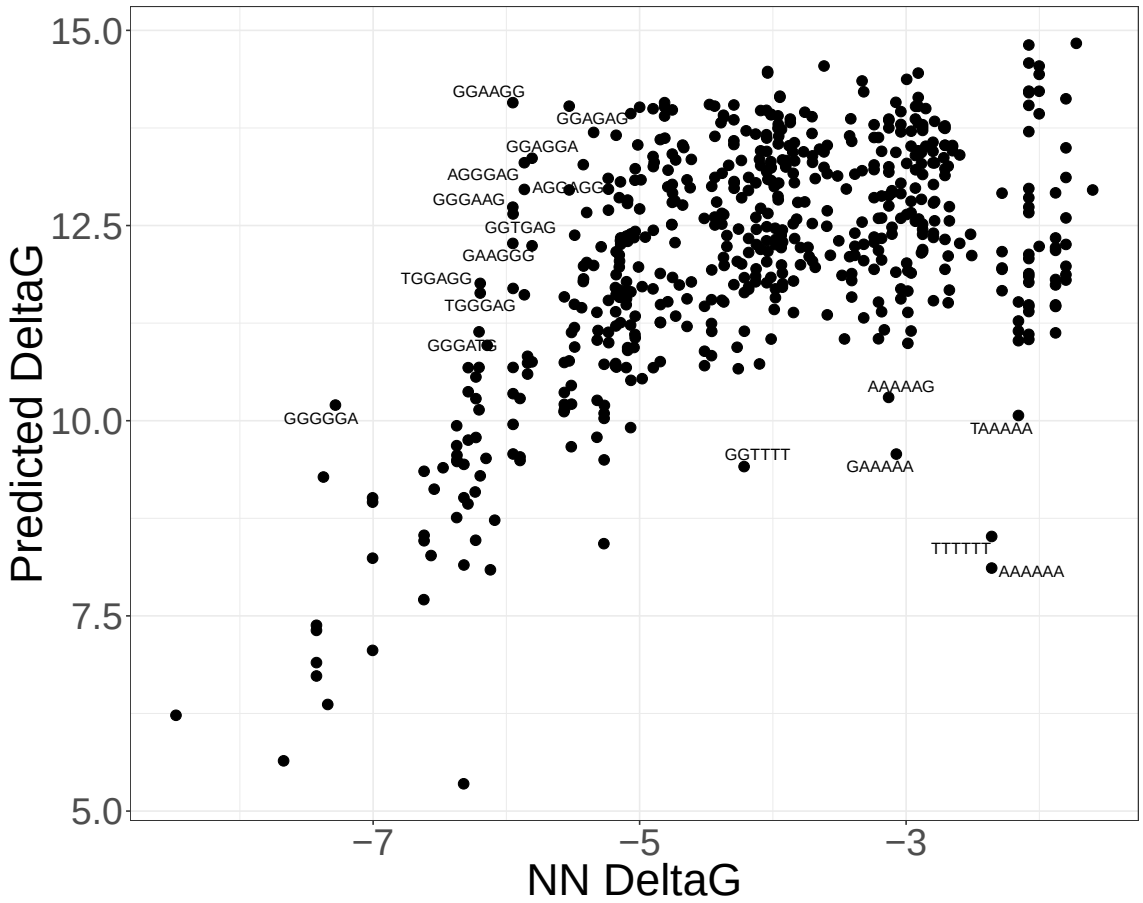
$$
\begin{aligned}
[p_i\, t] &= [b_1 p_i] + [b_2 p_i] + [b_3 p_i] \ldots \\
&= [b_1][p_i] \exp(\beta \Delta G) + [b_2][p_i] \exp(\beta \Delta G) + \ldots \\
&= [t][p_i] \sum_n \exp(\beta \Delta G)\, P(b_n|t) \\
&= [t][p_i] \exp(\beta \Delta F)
\end{aligned}
$$

Validating ΔF: consistency between experiments



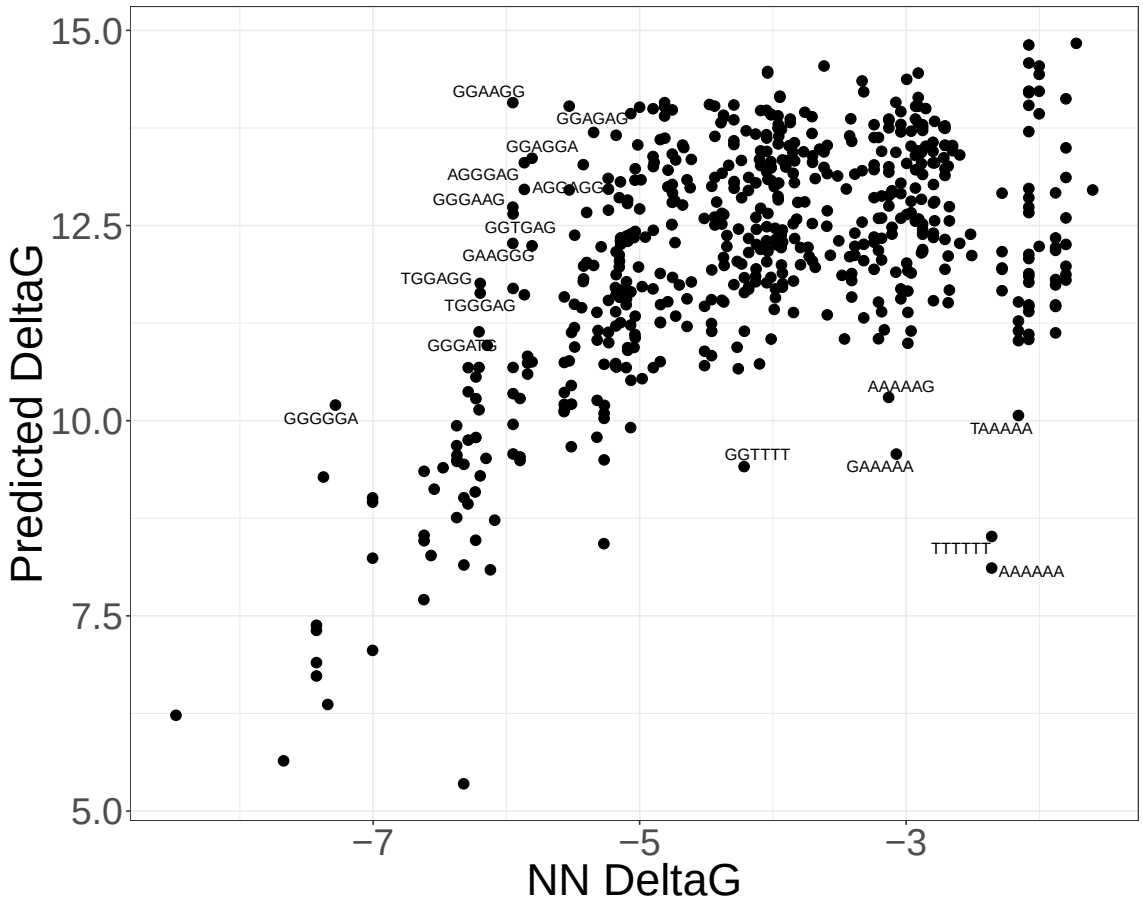
Validating ΔF: consistency with tabulated free-binding energy values

Hexamers with no C $\longrightarrow \Delta F \equiv \Delta G$



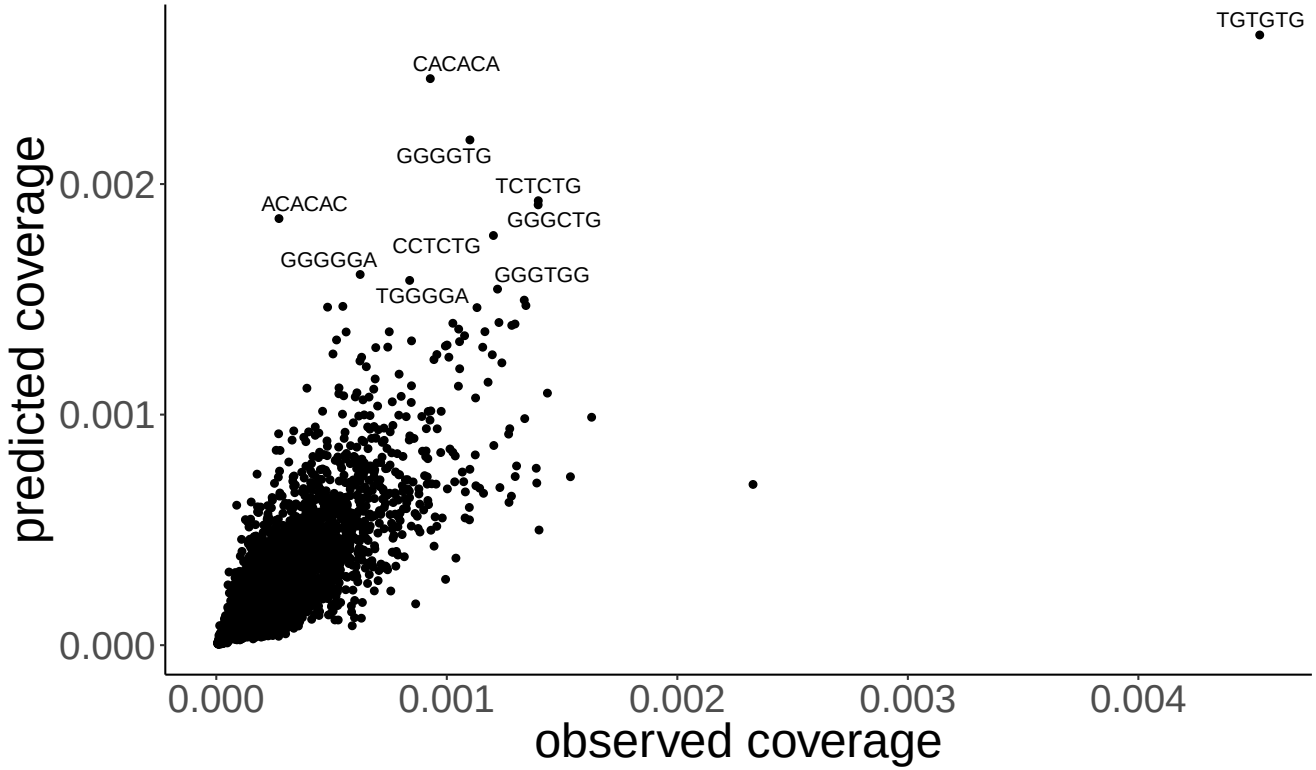
Validating ΔF: consistency with tabulated free-binding energy values

Hexamers with no C $\longrightarrow \Delta F \equiv \Delta G$

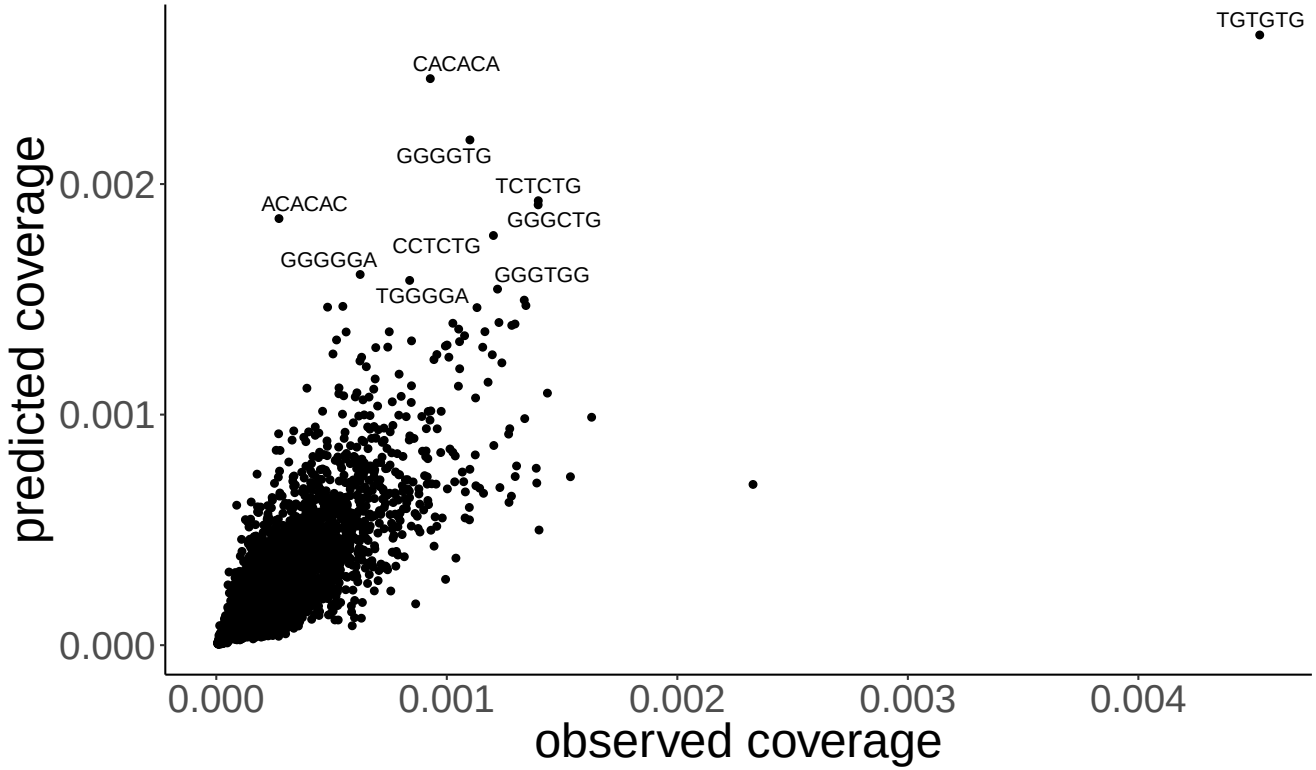


Ongoing: generating data on non-converted samples for validation of ΔG prediction

Predicted coverage from ΔF

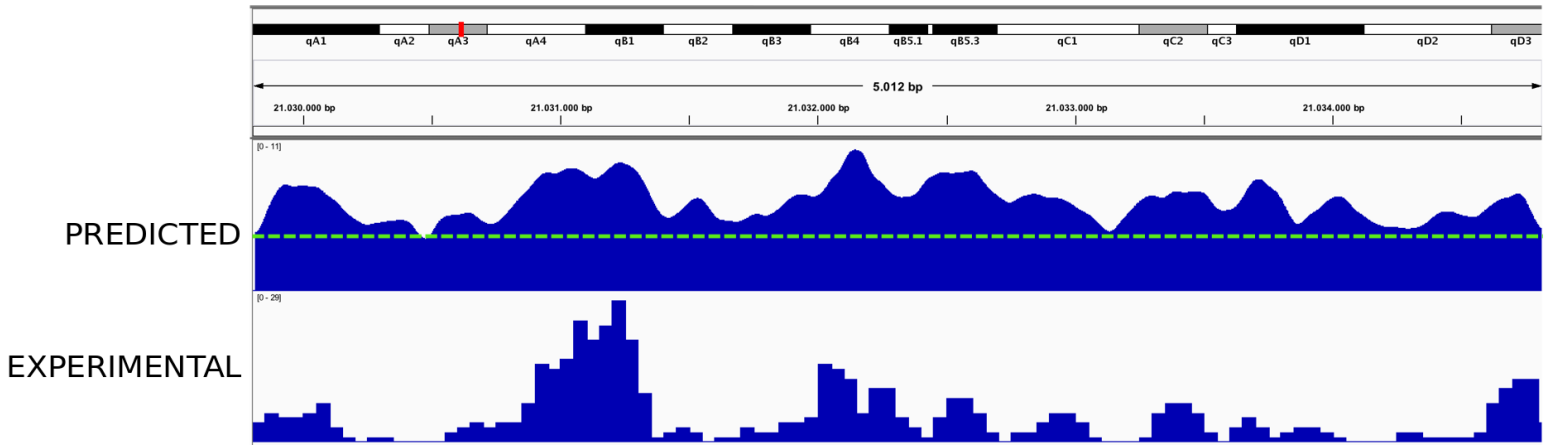


Predicted coverage from ΔF

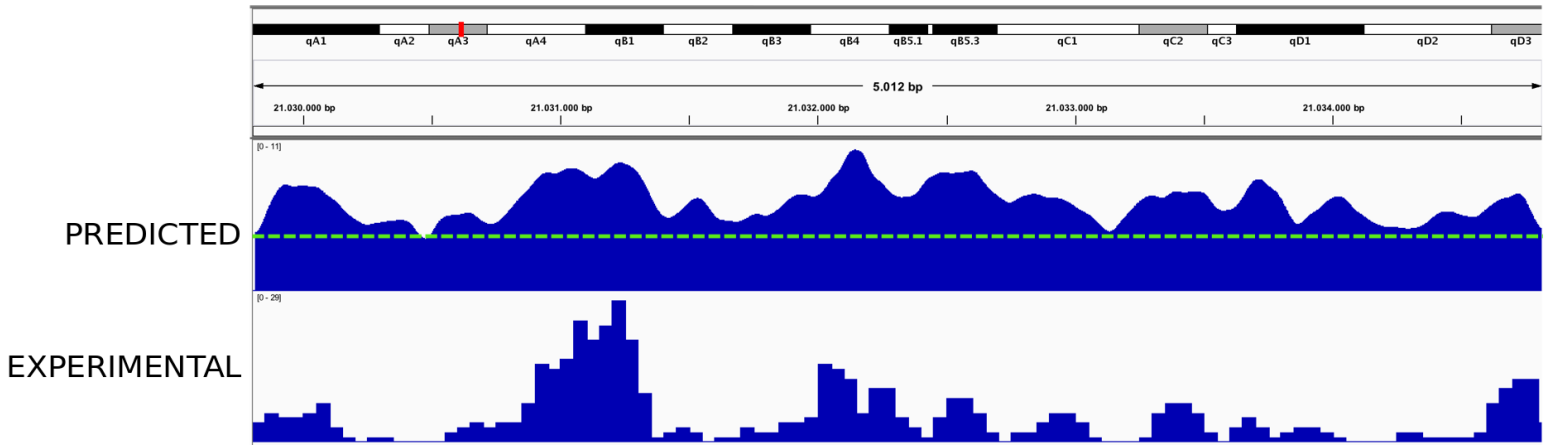


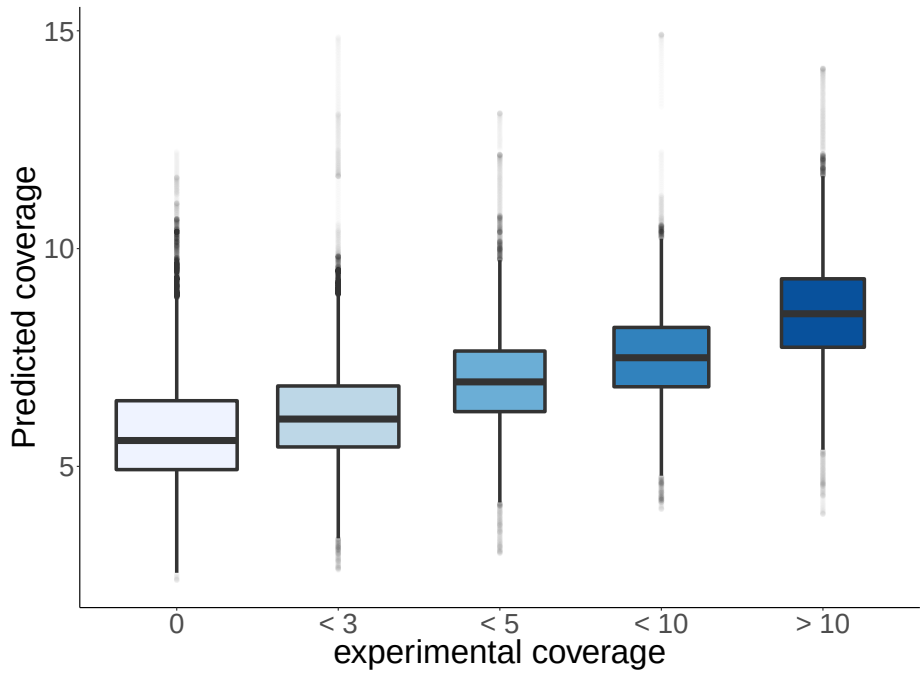
Does binding explain the spatial genome-wide distribution of coverage?

Positional information from binding probability



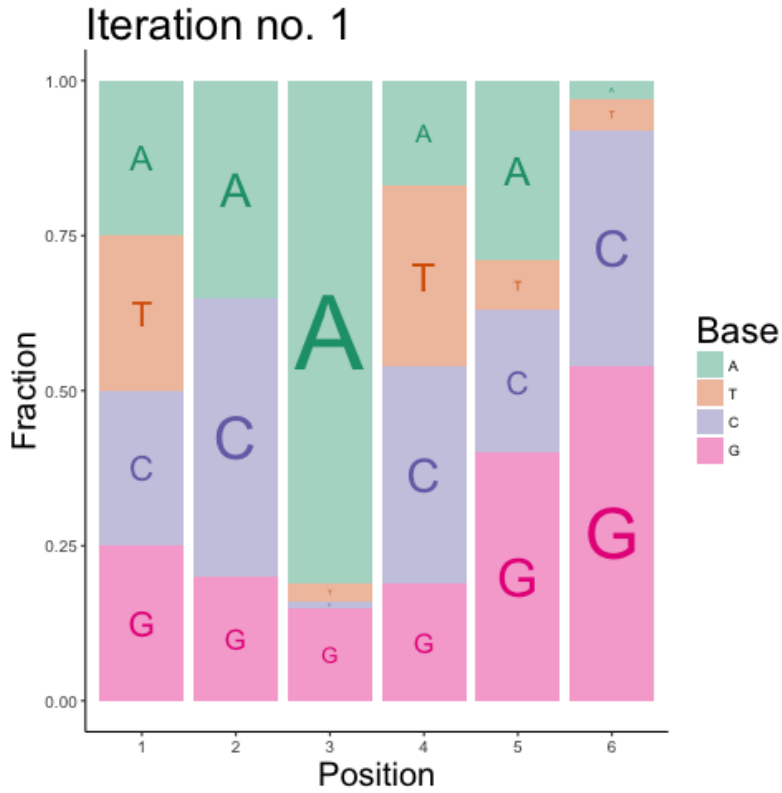
Positional information from binding probability





Next steps: optimizing the primer pool

- Altering the coverage by changing primer concentration
- Finding the optimal primer composition matrix (Differential evolution algorithm)



Acknowledgements

- **Alexander Van Oudenaarden**
- **Christoph Geisenberger**
- **Anna Alemany Arias**
- Juan Pedraza
- Anna Van Oudenaarden
- Maya Sen
- Lennart Kaester
- Buys de Barbanson
- the AvO lab