Understanding DNA Methylation Using Mathematical Models

Lyndsay Kerr

University of Strathclyde
Department of Mathematics and Statistics

Joint work with Duncan Sproul and Ramon Grima (University of Edinburgh)

October 2023

My journey to where I am now...

2006–2012: St Margaret's High School

- Favourite subject was maths–thought this meant I should study accountancy at university;
- Encouraged by teacher to study maths instead.

2012–2016: UG at University of Strathclyde

- Favourite classes were applied analysis and mathematical biology;
- Worked as UG Teaching Assistant during third and fourth years;
- Enjoyed final year research project–encouraged me to do PhD.

My journey to where I am now...

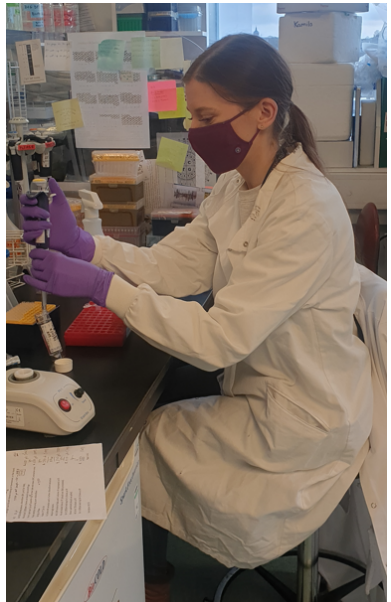
2016–2019: PhD at University of Strathclyde

- Coagulation-fragmentation equations: describes systems of particles that can merge together and break apart;
- Applications in blood clotting, animal groupings, powder production industry;
- I concentrated more on "pure", theoretical side;
- Spent time as teaching assistant for mathbio classes at Strathclyde and at AIMS in South Africa;
- Loved PhD but was motivated to move into more applied biological field...

My journey to where I am now...

2019–2023: Cross-Disciplinary Post-Doctoral Fellowship (XDF) at University of Edinburgh

Aim of 4-year programme:
bring together researchers from different sciences to tackle biomedical problems;



Month 1–2: explore the field of biomedicine

↓

Year 1: Undertake rotation project while getting to grips with (and exploring) the field

↓

Years 2–4: Main 3-year project

My journey to where I am now...

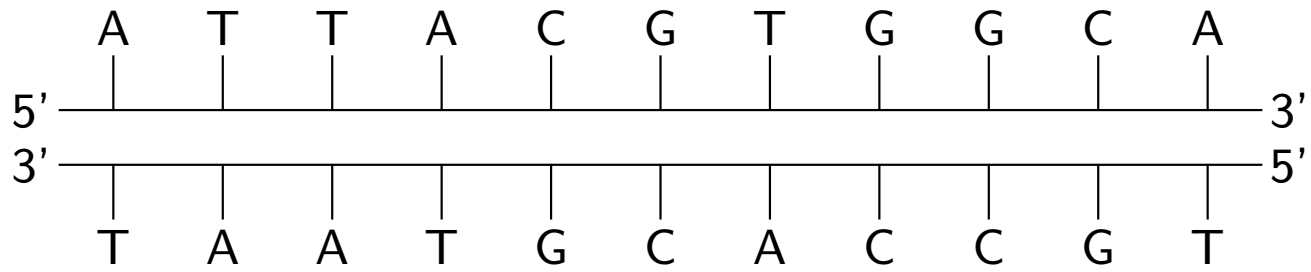
Current: Chancellor's Fellow at University of Strathclyde

- Studying DNA systems using mathematical modelling and analysis of experimental data;
- In particular: how/why do particular changes in DNA occur in disease?
- Time split between research and teaching (more research-focussed to start with);
- Continuing to lecture mathematical biology course at AIMS in South Africa.

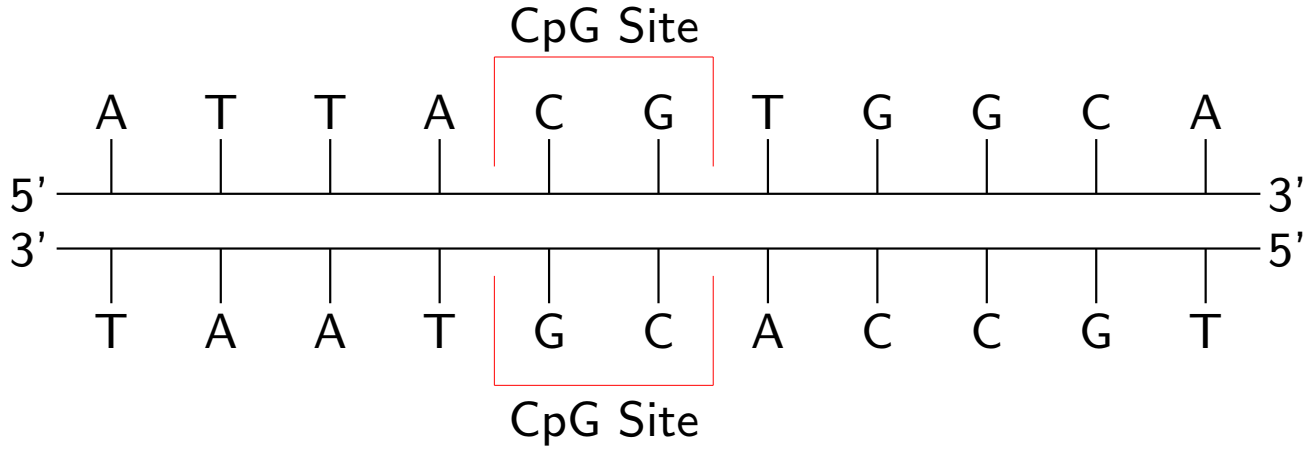
Now onto some research...

DNA is found in every cell in your body and contains your genetic information

- DNA is made up of basic structural units called nucleotides;
- Four types of nucleotide bases in DNA: adenine (A), thymine (T), cytosine (C) and guanine (G);
- A DNA molecule is made up of two "complementary strands" that are linked by weak chemical bonds between A and T nucleotides and between C and G nucleotides.

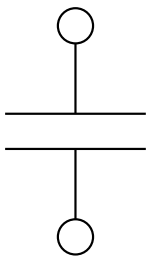


DNA methylation is an epigenetic mark that is primarily found at CpG sites

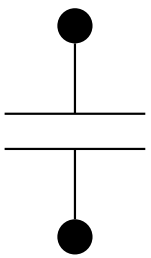


Double-stranded CpG dyad is always in one of three possible states.

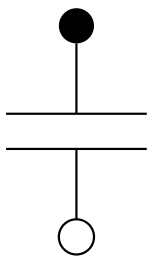
Unmethylated



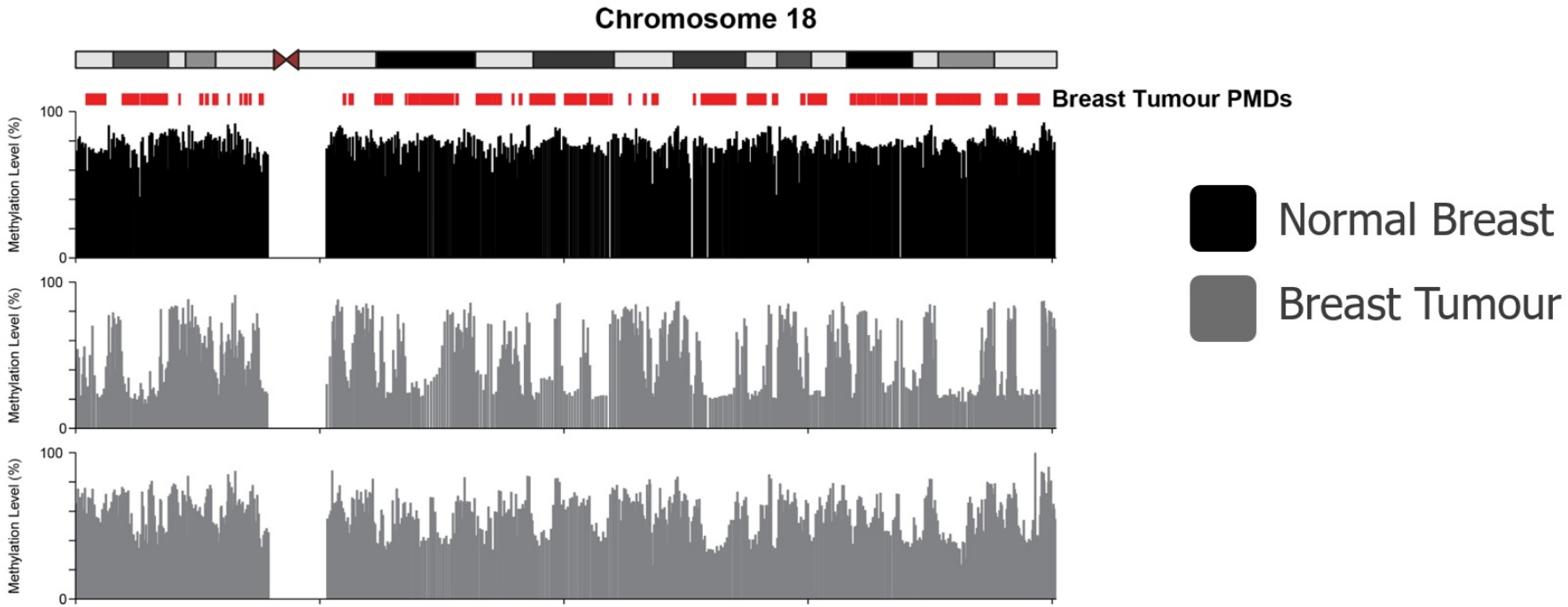
Methylated



Hemimethylated

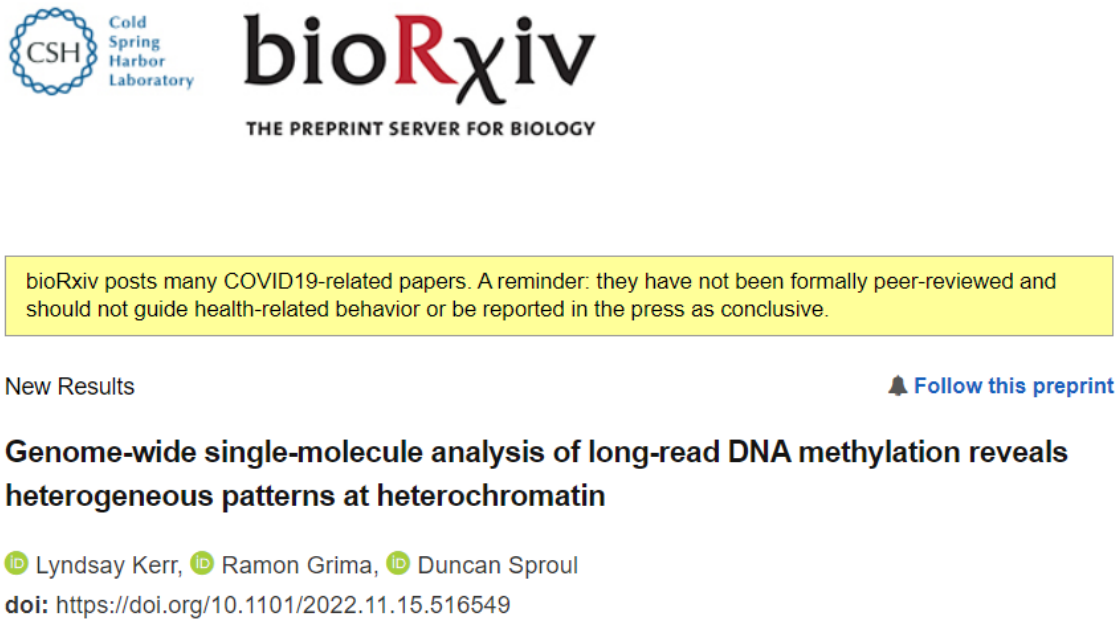


Megabase-scale loss of DNA methylation in cancer has been observed since 1983



- Methylation loss is widely observed in different tumour types.
- Large regions affected by methylation loss are referred to as partially methylated domains (PMDs).

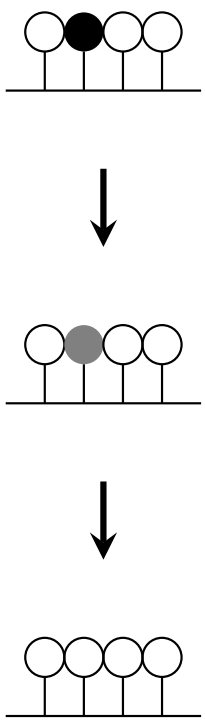
Our recent findings: PMDs are associated with disordered, heterogeneous methylation patterns

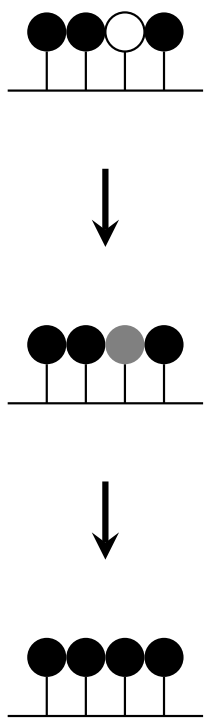


Can mathematical models help us to understand how these regions lose methylation and form disordered methylation patterns?

Studies indicate that CpG sites collaborate with each other in the genome (likely via enzyme recruitment)

○ unmethylated; ● hemimethylated; ● methylated





System of reactions has previously been proposed to describe methylation processes (*Haerter et. al. 2014*)

u: unmethylated; h: hemimethylated; m: methylated

Non-collaborative: $u \xrightarrow{k_1} h$; $h \xrightarrow{k_2} m$; $m \xrightarrow{k_3} h$; $h \xrightarrow{k_4} u$;

Collaborative :

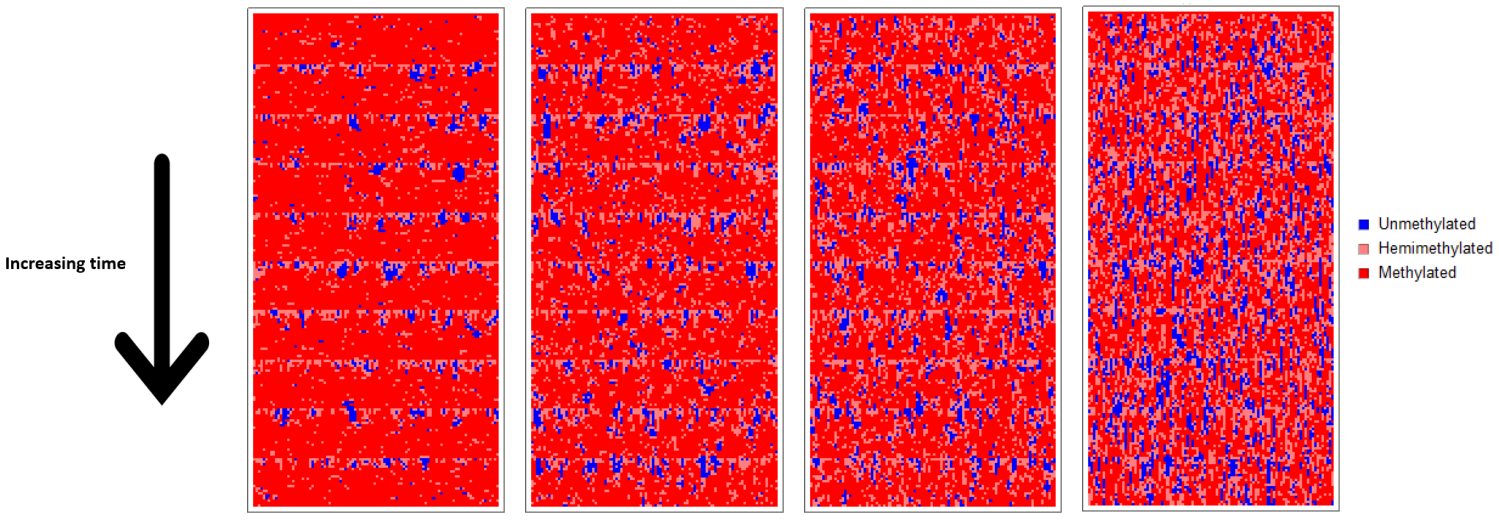
$$u + h \xrightarrow{k_5} h + h; \qquad u + m \xrightarrow{k_6} h + m;$$
$$h + h \xrightarrow{k_7} m + h; \qquad h + m \xrightarrow{k_8} m + m;$$
$$m + h \xrightarrow{k_9} h + h; \qquad m + u \xrightarrow{k_{10}} h + u;$$
$$h + h \xrightarrow{k_{11}} u + h; \qquad h + u \xrightarrow{k_{12}} u + u.$$

Assumption: collaboration can only occur between neighbouring CpGs
$\Longrightarrow$ nearest-neighbour collaborative system

How can varying the reaction rates cause methylation to be lost?

Weakening the collaborativity between CpGs leads to loss of DNA methylation

high collaborativity $\Longrightarrow$ low collaborativity



Result from simulations:

- decreasing collaborativity leads to highly methylated regions losing methylation and "disordered" methylation patterns forming;
- other causes of methylation loss lead to "ordered" unmethylated patterns.

Hypothesis: Could collaborativity strength be lower in PMDs?

Hypothesis: collaborativity strength is lower in PMDs compared to non-PMDs

Why could hypothesis make sense?

CpGs are generally further apart from each other in PMDs compared to non-PMDs;

Strategy to investigate

Infer which collaborativity rates are likely to have produced the patterns in PMDs and in non-PMDs:

- ▷ Predict properties of methylation patterns resulting from different sets of reaction rates.
- ▷ Check which predictions are "most similar" to observed data—i.e. which rates are likely to have produced observed data?


How can we
obtain predictions
for different
reaction rates?

How can we obtain predictions to compare to real data?

- We could use **stochastic simulations** of nearest-neighbour collaborative system...

How can we obtain predictions to compare to real data?

- We could use **stochastic simulations** of nearest-neighbour collaborative system...

but these are **very computationally expensive.**

How can we obtain predictions to compare to real data?

- We could use **stochastic simulations** of nearest-neighbour collaborative system...

  but these are **very computationally expensive.**

- Better to use **mathematical equations** describing the nearest-neighbour collaborative system...

How can we obtain predictions to compare to real data?

- We could use **stochastic simulations** of nearest-neighbour collaborative system...

  but these are **very computationally expensive.**

- Better to use **mathematical equations** describing the nearest-neighbour collaborative system...

  but these are **infeasible to construct** for large systems of CpG sites.

Can we **approximate** the nearest-neighbour collaborative system using a model that can be described by equations?

Is collaborativity strength lower in PMDs compared to non-PMDs?

Aim:

Find a model that can be written in terms of mathematics and provides a good **approximation** to the nearest-neighbour collaborative system...

Step 1: Find this approximation by comparing predictions from models to data simulated using nearest-neighbour collaborative system.

Step 2: Compare predictions obtained from this approximate model to observed data and infer reaction rates.

Interested in examining the system in steady state.

We're Going on a Model Hunt

INTERFACE

royalsocietypublishing.org/journal/rsif

Research

Cluster mean-field theory accurately predicts statistical properties of large-scale DNA methylation patterns

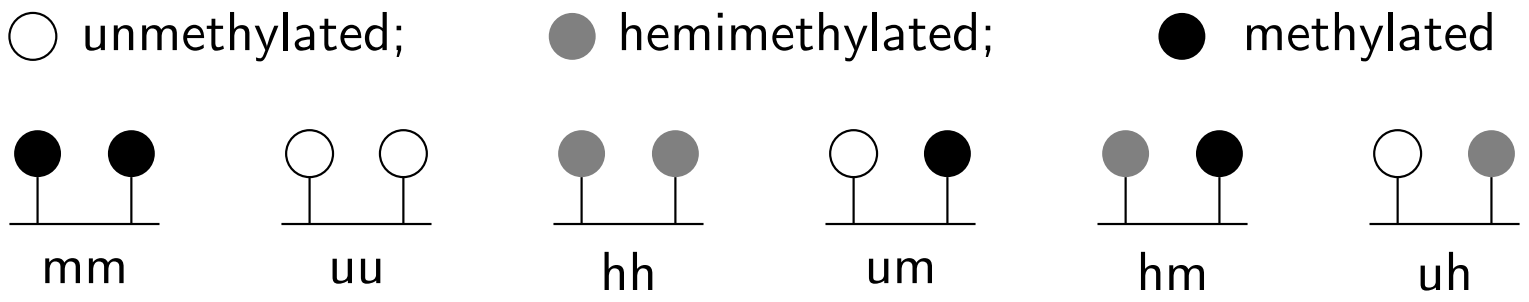
Lyndsay Kerr[1], Duncan Sproul[2] and Ramon Grima[3]

[1]MRC Institute of Genetics and Cancer,[2]MRC Human Genetics Unit and CRUK Edinburgh Centre, Institute of Genetics and Cancer, and[3]School of Biological Sciences, University of Edinburgh, Edinburgh, UK

Can nearest-neighbour collaborativity in large systems be approximated by considering small systems?

Two-site model

There are six possible states for each pair of sites:



$$\frac{dP_{mm}}{dt} = k_2 P_{hm} - 2k_3 P_{mm} + k_8 P_{hm};$$

$$\frac{dP_{uu}}{dt} = -2k_1 P_{uu} + k_4 P_{uh} + k_{12} P_{uh}$$

$$\frac{dP_{hh}}{dt} = k_1 P_{uh} - 2k_2 P_{hh} + k_3 P_{hm} - 2k_4 P_{hh} + k_5 P_{uh} - 2k_7 P_{hh} + k_9 P_{hm} - 2k_{11} P_{hh}$$

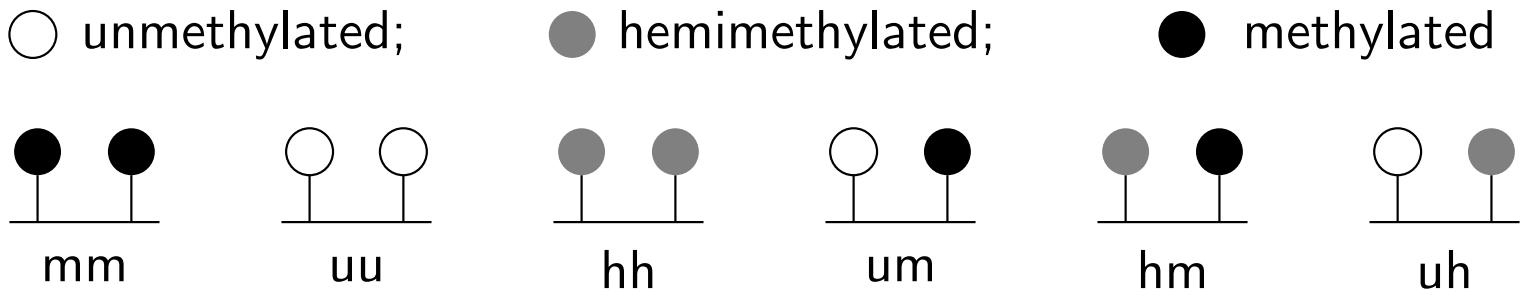
$$\frac{dP_{um}}{dt} = -k_1 P_{um} + k_2 P_{uh} - k_3 P_{um} + k_4 P_{hm} - k_6 P_{um} - k_{10} P_{um}$$

$$\frac{dP_{hm}}{dt} = k_1 P_{um} + 2k_2 P_{hh} - k_2 P_{hm} + 2k_3 P_{mm} - k_3 P_{hm} - k_4 P_{hm} + k_6 P_{um} + 2k_7 P_{hh} - k_8 P_{hm} - k_9 P_{hm}$$

$$P_{uh} = 1 - P_{mm} - P_{uu} - P_{hh} - P_{um} - P_{hm}.$$

Two-site model

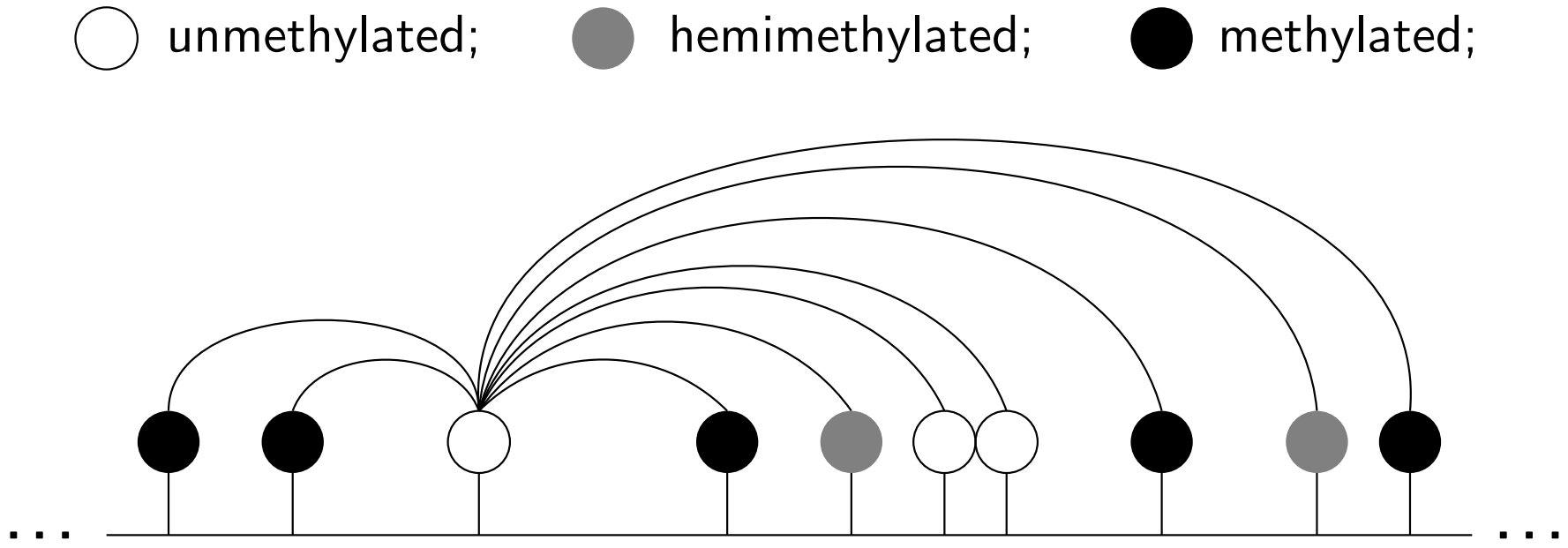
There are six possible states for each pair of sites:



- Can write down equations describing nearest-neighbour collaboration for this small system.
- Statistics obtained from two-site model are poor predictions of statistics associated with large systems.
- Three-site model does not provide much of an improvement.

Perhaps we should instead try a model describing an infinite number of CpG sites...

Mean field (MF) model



The change in state of a CpG depends on the mean of its local environment.

Can write down Chemical Master Equation describing the probability that a site is in each state

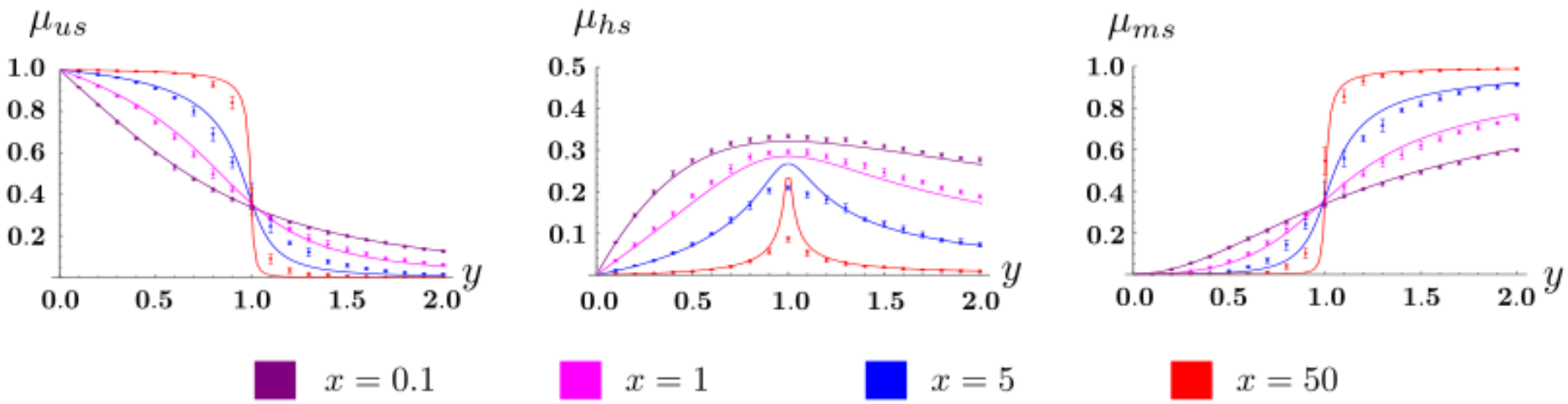
$$\frac{\mathrm{d}P_u}{dt} = -a_3 P_u + a_4 P_h$$
$$\frac{\mathrm{d}P_h}{dt} = a_1\left(1 - P_u - P_h\right) - a_2 P_h + a_3 P_u - a_4 P_h$$
$$P_m = 1 - P_u - P_h.$$

"Effective" reaction rates a_i, $i \in \{1, 2, 3, 4\}$ account for changes in state that occur due to

- non-collaborative reactions;
- collaborative reactions (determined by the mean state of the system rather than nearest-neighbour interactions).

MF model improves upon two-site model but decreases in accuracy as collaborativity strength increases

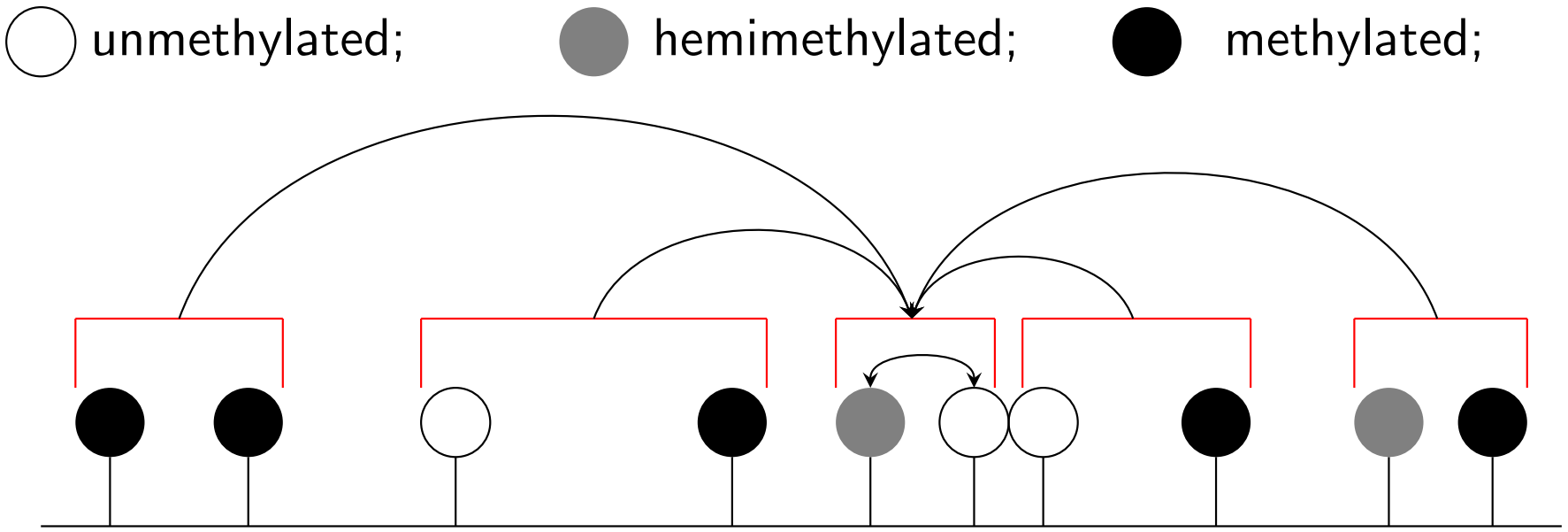
x: collaborativity strength; y: methylation strength



solid line: MF model; data points/error bars: simulations.

What will happen if we combine the two-site model with the MF model?

Distinct pairs MF model

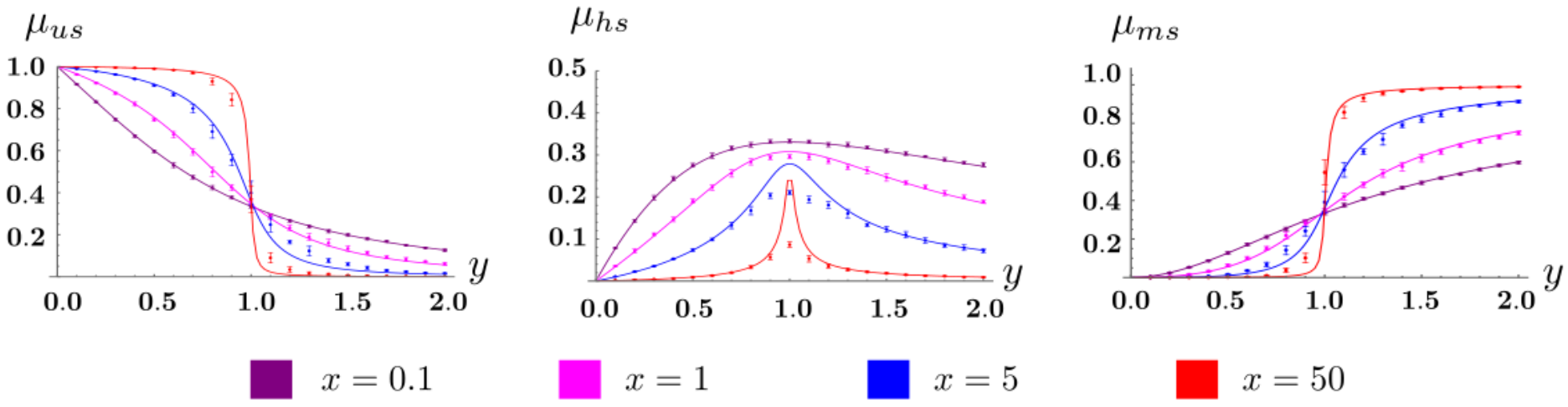


Split genome into distinct pairs: each CpG belongs to only one pair.

CpGs within a pair can interact and other interactions are approximated by considering the probability that two paired states are adjacent.

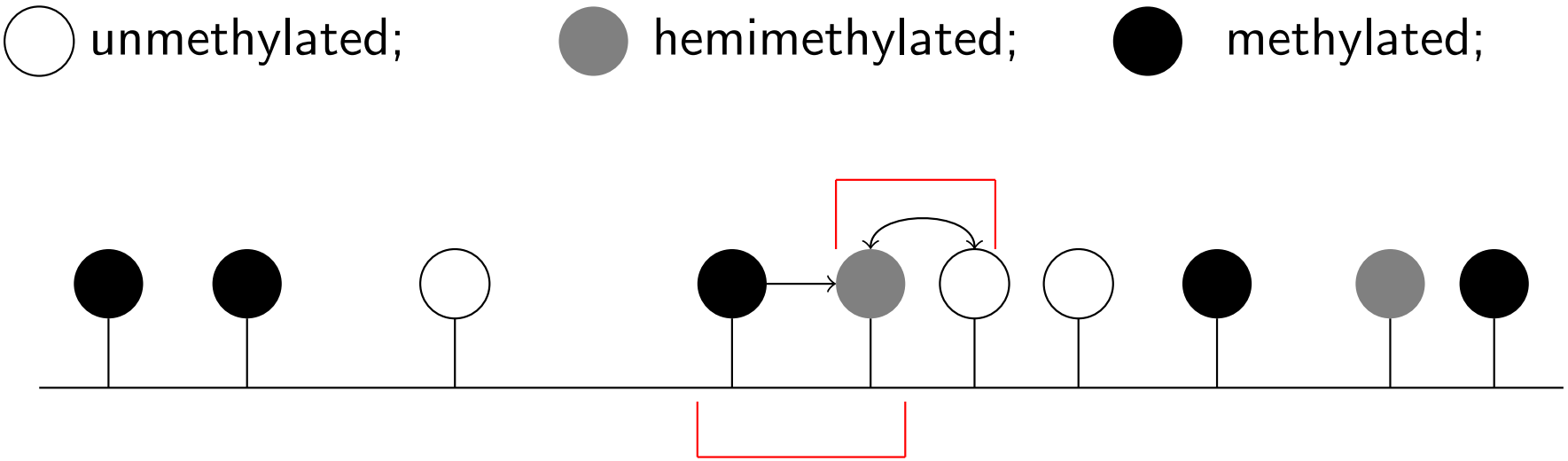
Distinct pairs MF model is a slight improvement of original MF model but still struggles for large x

x: collaborativity strength; y: methylation strength



solid line: distinct pairs MF model; data points/error bars: simulations.

Overlapping pairs MF model

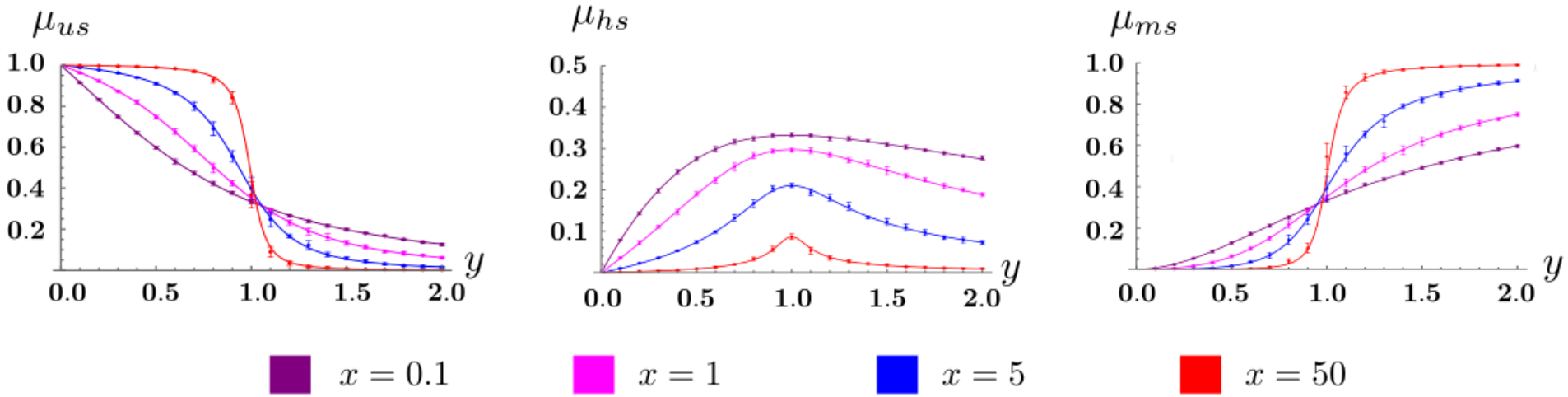


Split genome into overlapping pairs of CpGs: each CpG is in two pairs.

CpGs within a pair can interact and other interactions are approximated by considering the probabilities that certain states "overlap".

Overlapping pairs MF model does very well at approximating nearest-neighbour collaborative system

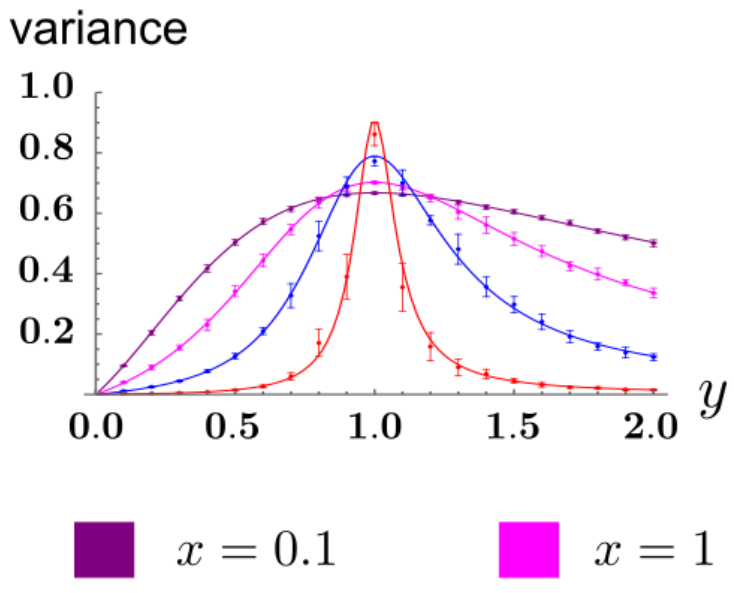
x: collaborativity strength; y: methylation strength

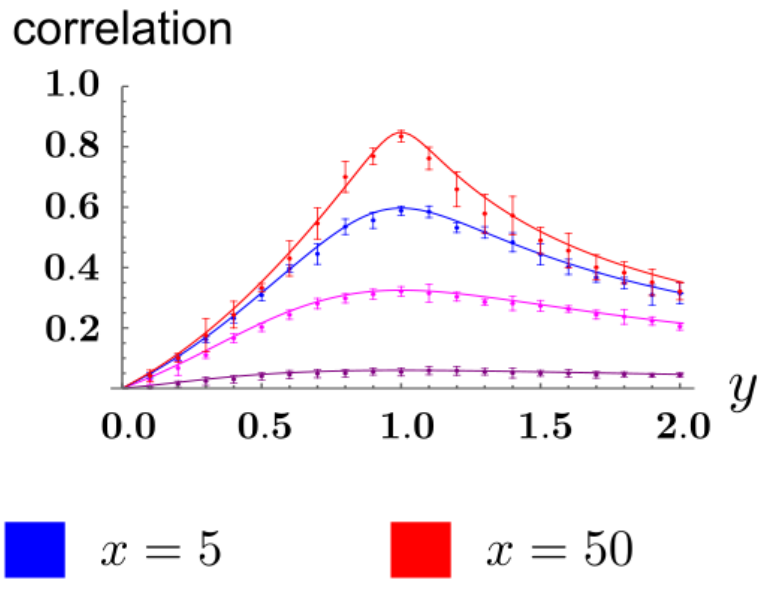


solid line: overlapping pairs MF model; data points/error bars: simulations.

Overlapping pairs MF model does very well at approximating nearest-neighbour collaborative system

x: collaborativity strength; y: methylation strength



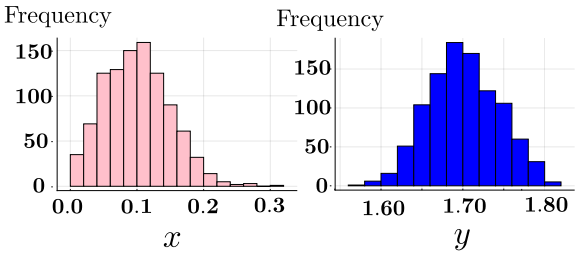


solid line: overlapping pairs MF model; data points/error bars: simulations.

Overlapping pairs MF model can be used to infer parameters for nearest-neighbour collaborative system

(a)

True Parameter Values: $x = 0.1, y = 1.7$



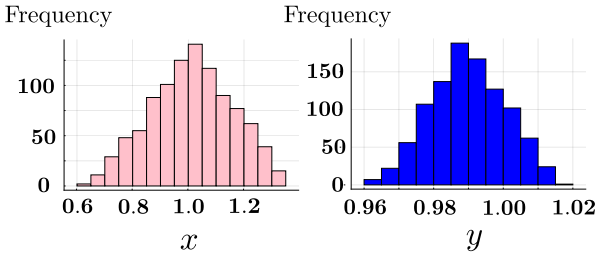
Inferred Parameter Values (95% CIs):

$x = 0.101\ (0.015, 0.198)$,

$y = 1.701\ (1.621, 1.787)$

(b)

True Parameter Values: $x = 1, y = 1$



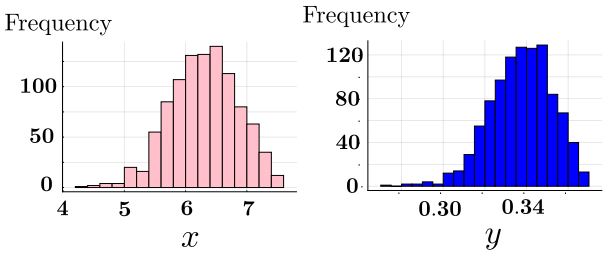
Inferred Parameter Values (95% CIs):

$x = 1.011\ (0.724, 1.283)$,

$y = 0.990\ (0.969, 1.010)$

(c)

True Parameter Values: $x = 5, y = 0.3$



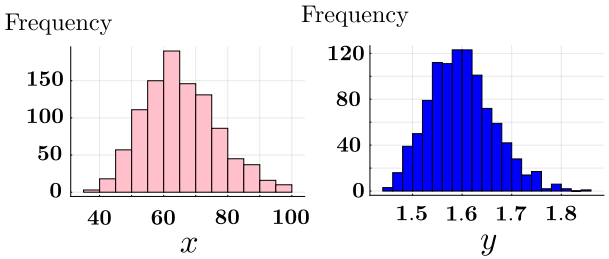
Inferred Parameter Values (95% CIs):

$x = 6.302\ (5.166, 7.302)$,

$y = 0.338\ (0.305, 0.363)$

(d)

True Parameter Values: $x = 50, y = 1.7$



Inferred Parameter Values (95% CIs):

$x = 65.347\ (45.327, 90.057)$,

$y = 1.599\ (1.486, 1.741)$

Future work: combine the data analysis and modelling

- High heterogeneity of patterns observed within PMDs supports hypothesis that methylation loss arises due to collaborativity breakdown;
- Investigate further by inferring collaborativity strength in different genomic regions;
- Is collaborativity lower in PMDs compared to non-PMDs?
- Infer parameters from different types of experimental dataset to identify possible mechanisms behind collaborativity;
- Examine DNA methylation using different types of mathematical models.

Acknowledgements

Grima Group

Sproul Lab

XDFs; XDF Directors; Arek Welman

THE UNIVERSITY OF EDINBURGH

MRC Medical Research Council

Chemical Master Equation for the Distinct/Overlapping Pairs MF Model

$$
\begin{aligned}
\frac{\mathrm{d}P_{mm}}{dt} &= -a_1 P_{mm} + a_8 P_{hm}, \\
\frac{\mathrm{d}P_{uu}}{dt} &= -a_2 P_{uu} + a_{11}\left(1 - P_{mm} - P_{uu} - P_{hh} - P_{um} - P_{hm}\right), \\
\frac{\mathrm{d}P_{hh}}{dt} &= -(a_3 + a_4) P_{hh} + a_9 P_{hm} + a_{10}\left(1 - P_{mm} - P_{uu} - P_{hh} - P_{um} - P_{hm}\right), \\
\frac{\mathrm{d}P_{um}}{dt} &= -(a_5 + a_6) P_{um} + a_7 P_{hm} + a_{12}\left(1 - P_{mm} - P_{uu} - P_{hh} - P_{um} - P_{hm}\right), \\
\frac{\mathrm{d}P_{hm}}{dt} &= a_1 P_{mm} + a_4 P_{hh} + a_5 P_{um} - (a_7 + a_8 + a_9) P_{hm}, \\
P_{uh} &= 1 - P_{mm} - P_{uu} - P_{hh} - P_{um} - P_{hm}.
\end{aligned}
$$

"Effective" reaction rates a_i, $i \in \{1, 2, \ldots, 12\}$ account for changes in state that occur due to

- non-collaborative reactions;
- collaborative reactions (between the two CpGs in the same pair);
- collaborative reactions (due to interaction between adjacent pairs in distinct pairs model or due to interaction between overlapping pairs in overlapping model).