# Estimation of the amplicon methylation pattern distribution from bisulphite sequencing data

Conrad Burden
Mathematical Sciences Institute
Australian National University
Canberra


Australian National University

Cytosines in a genome, particularly those in the combination CpG, can undergo an epigenetic change called *methylation*


(methylated) (not methylated) (methylated)

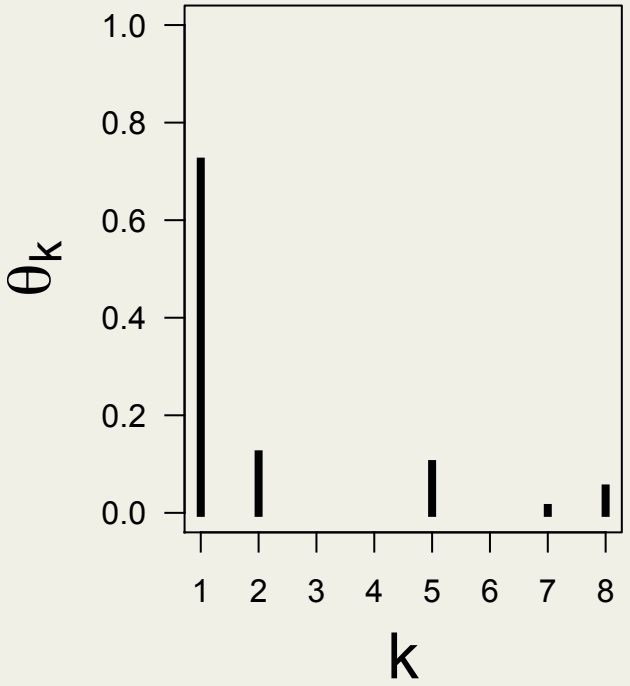
Methylation patterns

- can play a role in cell development
- can play a role in determining phenotype
- can be a response to environmental factors
- can change from cell-type to cell-type in an organism

A population of cells of a given type in a given organism defines a probability distribution over methylation patterns

e.g. for 3 nearby CpG sites

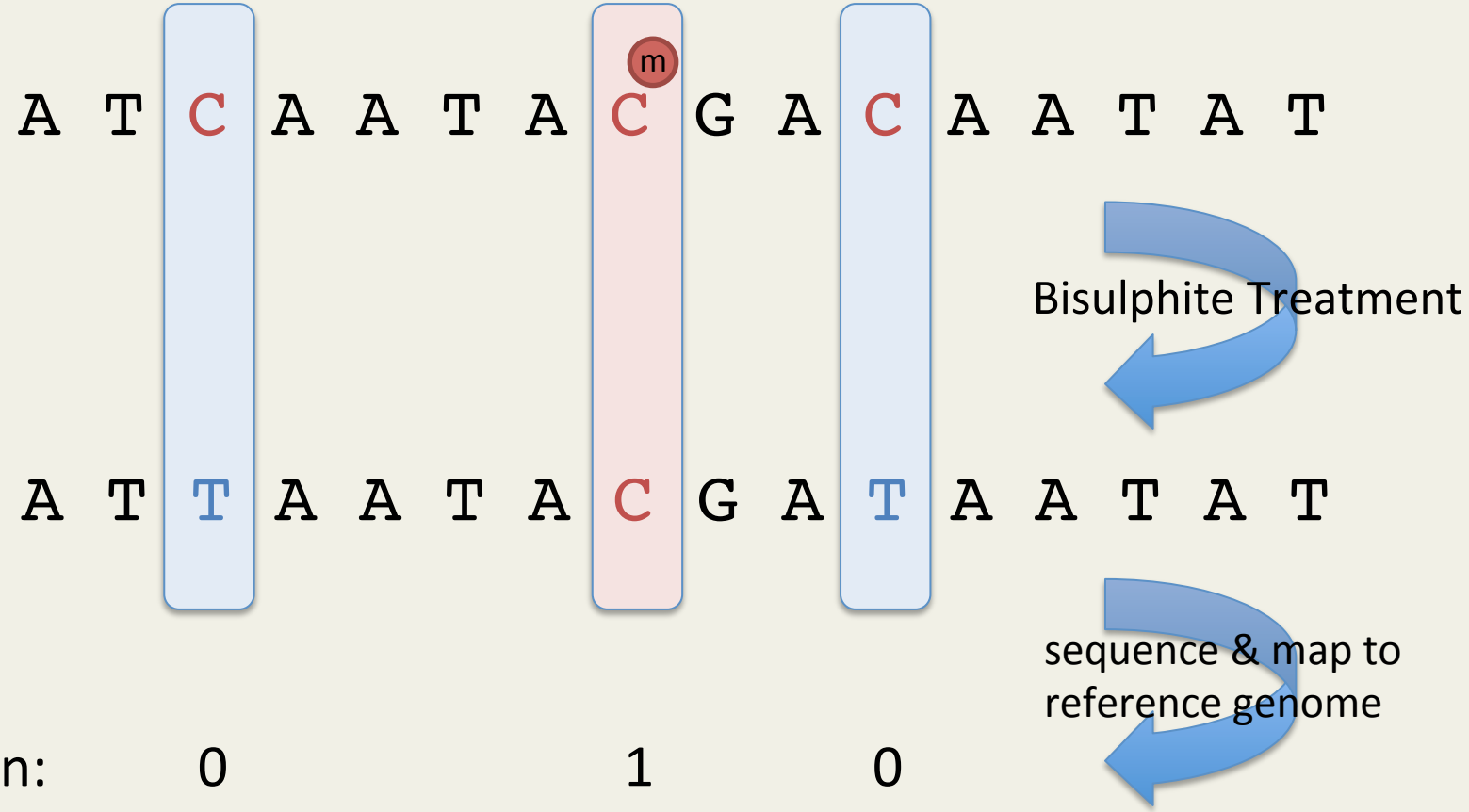
| $k$ | pattern* | $\text{Prob}(K = k) = \theta_k$ |
|---|---|---|
| 1 | 0 0 0 | 0.72 |
| 2 | 0 0 1 | 0.12 |
| 3 | 0 1 0 | 0.00 |
| 4 | 0 1 1 | 0.00 |
| 5 | 1 0 0 | 0.10 |
| 6 | 1 0 1 | 0.00 |
| 7 | 1 1 0 | 0.01 |
| 8 | 1 1 1 | 0.05 |



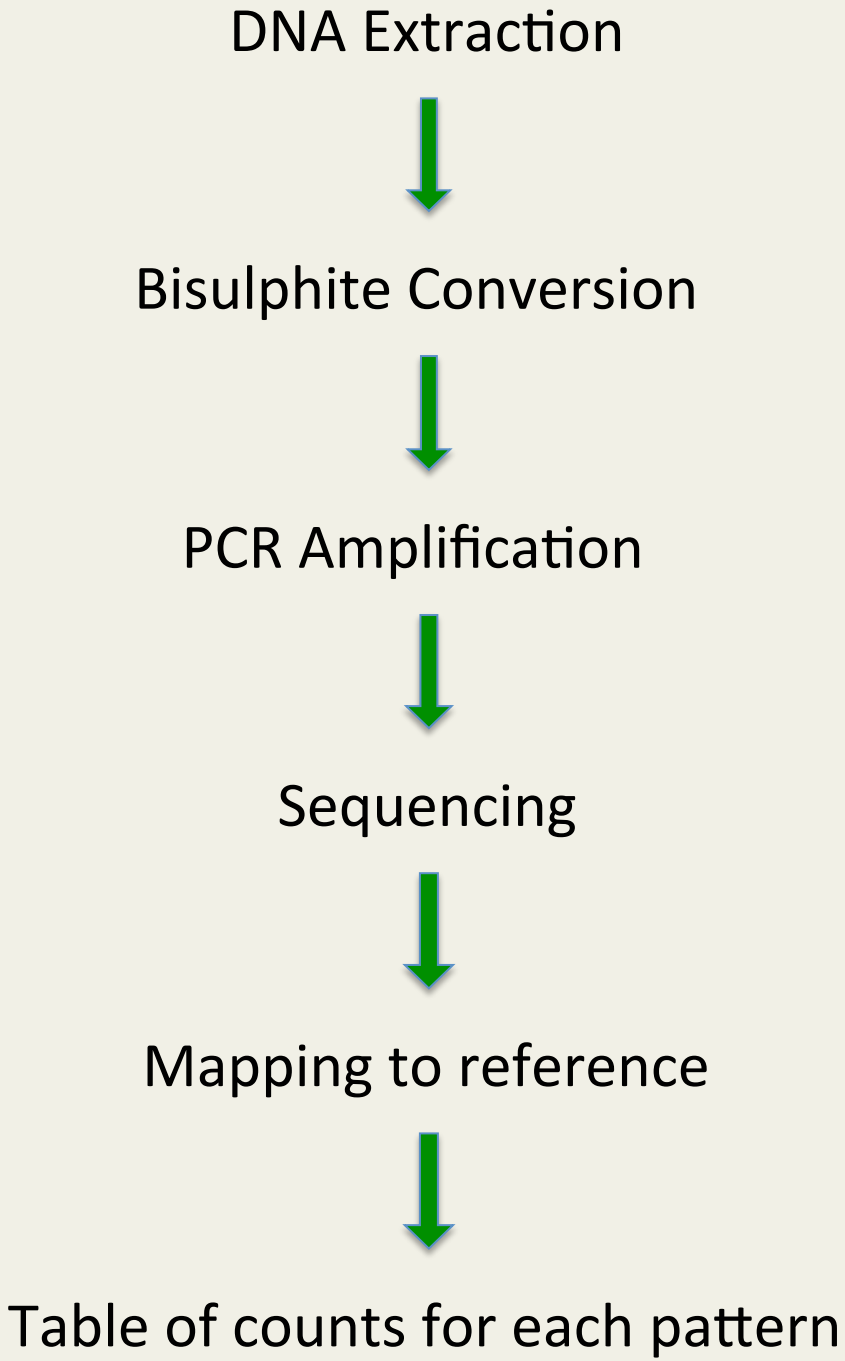

* 0 = unmethyated 1 = methylated

# Methylation patterns are measured directly in the laboratory via

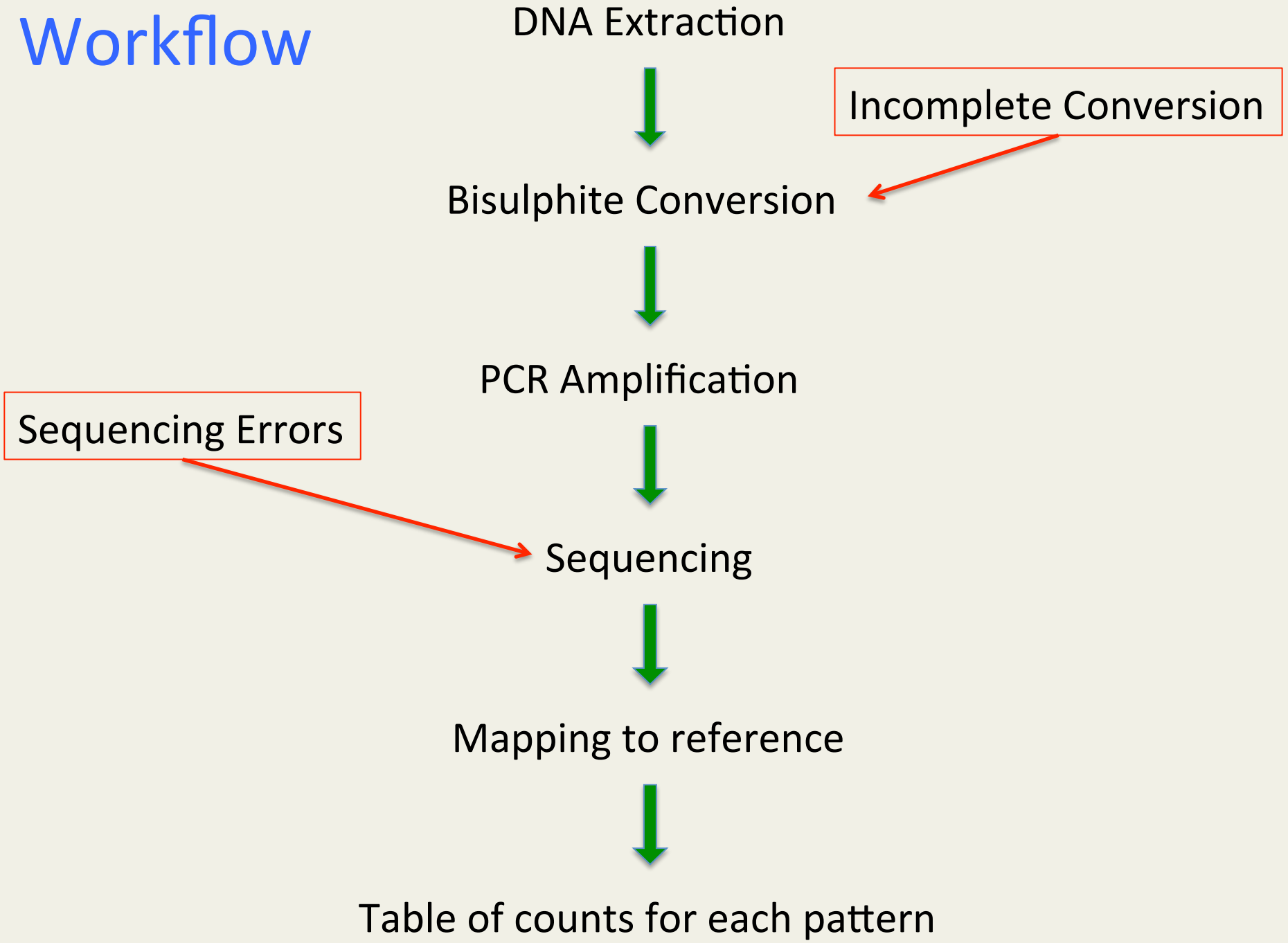
## Bisulphite Conversion



# Workflow


DNA Extraction
Bisulphite Conversion
PCR Amplification
Sequencing
Mapping to reference
Table of counts for each pattern

# Workflow


DNA Extraction
Incomplete Conversion
Bisulphite Conversion
PCR Amplification
Sequencing Errors
Sequencing
Mapping to reference
Table of counts for each pattern

# We have developed the R Bioconductor Package **MPFE**:

- **M**ethylation **P**atterns **F**requency **E**stimation

- Inputs ($n$ cytosines)
- A table of methylation pattern counts $y_k$, where $1 \leq k \leq 2^n$ labels the $2^n$ different methylation patterns
- The non-conversion rate $\varepsilon$
- The sequencing error rate $\eta$, either global or site-dependent

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

# We have developed the R Bioconductor Package **MPFE**:

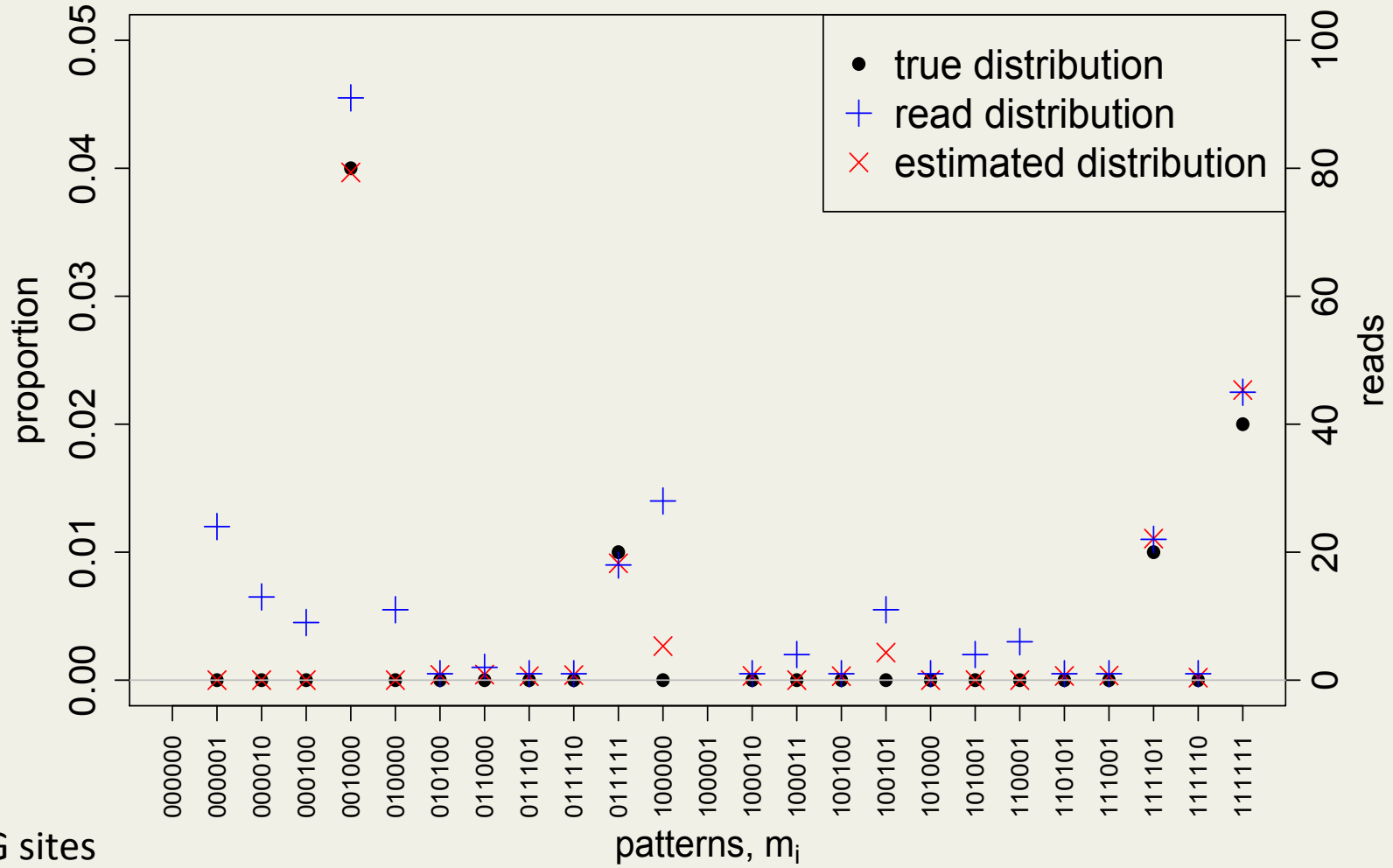
- **M**ethylation **P**atterns **F**requency **E**stimation

- Inputs ($n$ cytosines)
- A table of methylation pattern counts $y_k$, where $1 \leq k \leq 2^n$ labels the $2^n$ different methylation patterns
- The non-conversion rate $\varepsilon$
- The sequencing error rate $\eta$, either global or site-dependent

- Outputs
- A table of patterns and their estimated frequencies $\theta_k$
- List of spurious patterns called
- Plots comparing the observed and estimated frequencies

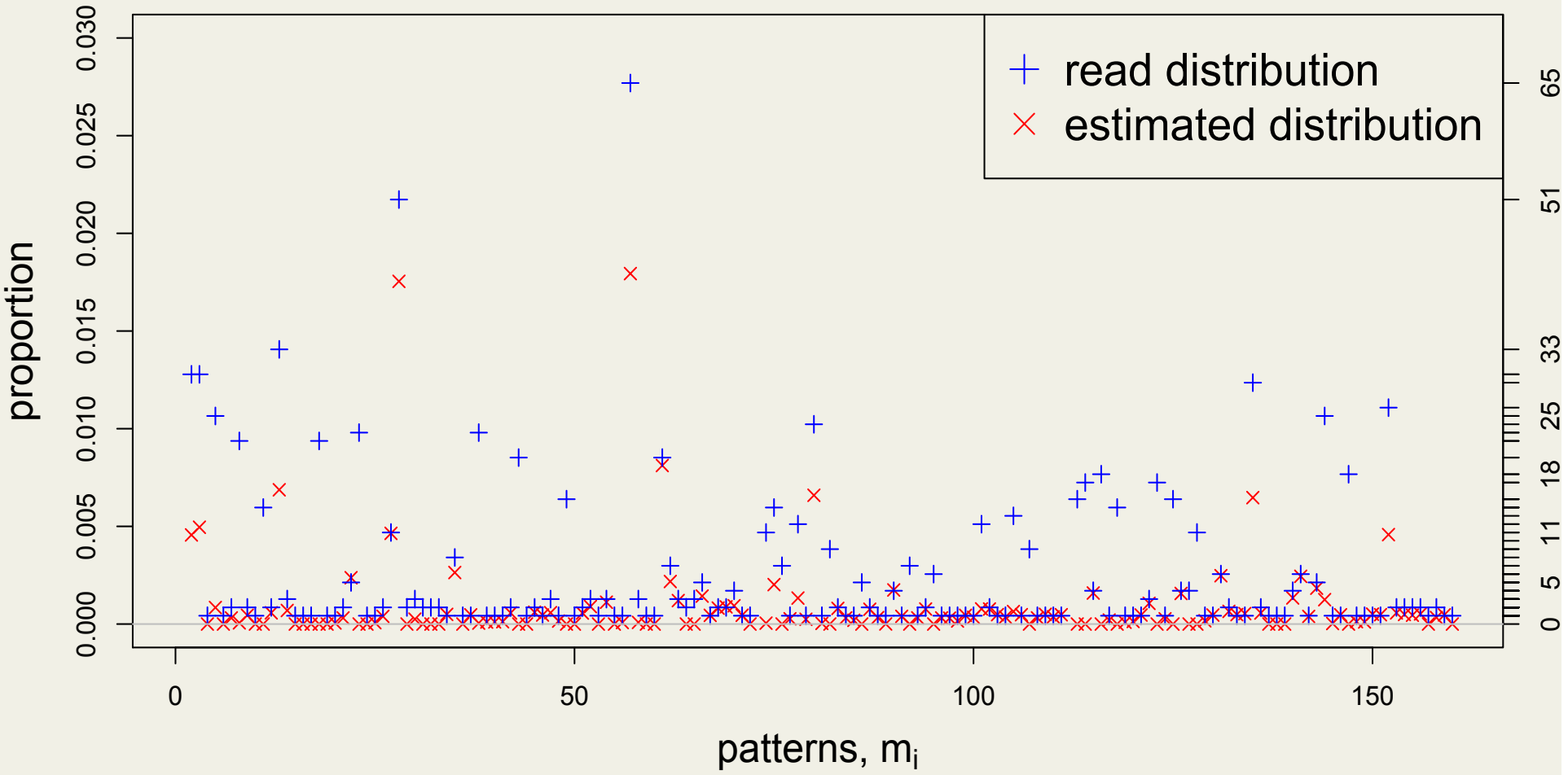
Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

# Synthetic Data




n = 6 CpG sites
total number of reads = 2000
non-conversion rate ε = 0.008
sequencing error rate η = 0.005
25 patterns, 9 are called spurious

# Biological Data (honeybee worker brains – gene GB17113)




n = 14 CpG sites
total number of reads = 2347
non-conversion rate ε = 0.01
sequencing error rate η = 0.02
160 patterns, 47 are called spurious

# How does it work?

Given:

- $n$ CpG sites, $k = 1, 2, ..., 2^n$ possible patterns
- true distribution over patterns to be estimated

$$\Pr(K = k) = \theta_k$$

- non-conversion rate $\varepsilon$ and read error rate $\eta$

The probability of a true pattern $k$ registering as pattern $l$ is

$$\Pr(L = l \mid K = k) = M_{kl}$$

where the $2^n \times 2^n$ matrix $M$ is (after a little bit of algebra)

$$M = \underbrace{E \otimes E \otimes \ldots \otimes E}_{n \text{ times}}, \qquad E = \begin{pmatrix} 1-\varepsilon-\eta+2\varepsilon\eta & \varepsilon+\eta-2\varepsilon\eta \\ \eta & 1-\eta \end{pmatrix}$$

Then the probability a read registers as pattern number $l$ is

$$\Pr(L=l)=\sum_{k=1}^{2^n}\Pr(K=k)\Pr(L=l\,|\,K=k)=\sum_{k=1}^{2^n}\theta_k M_{kl}$$

The distribution of read counts $Y_1, Y_2, \ldots, Y_{2\hat{}n}$ for patterns $l = 1, \ldots, 2^n$ out of a total of $N$ reads is a multinomial distribution:

$$\Pr\left(\mathbf{Y}=\mathbf{y}\,|\,\theta\right)=\frac{N!}{y_1!y_2!\ldots y_{2^n}!}\prod_{l=1}^{2^n}\left(\sum_{k=1}^{2^n}\theta_k M_{kl}\right)^{y_l}$$

... which enables a maximum likelihood estimate of the underlying distribution $\theta$ from

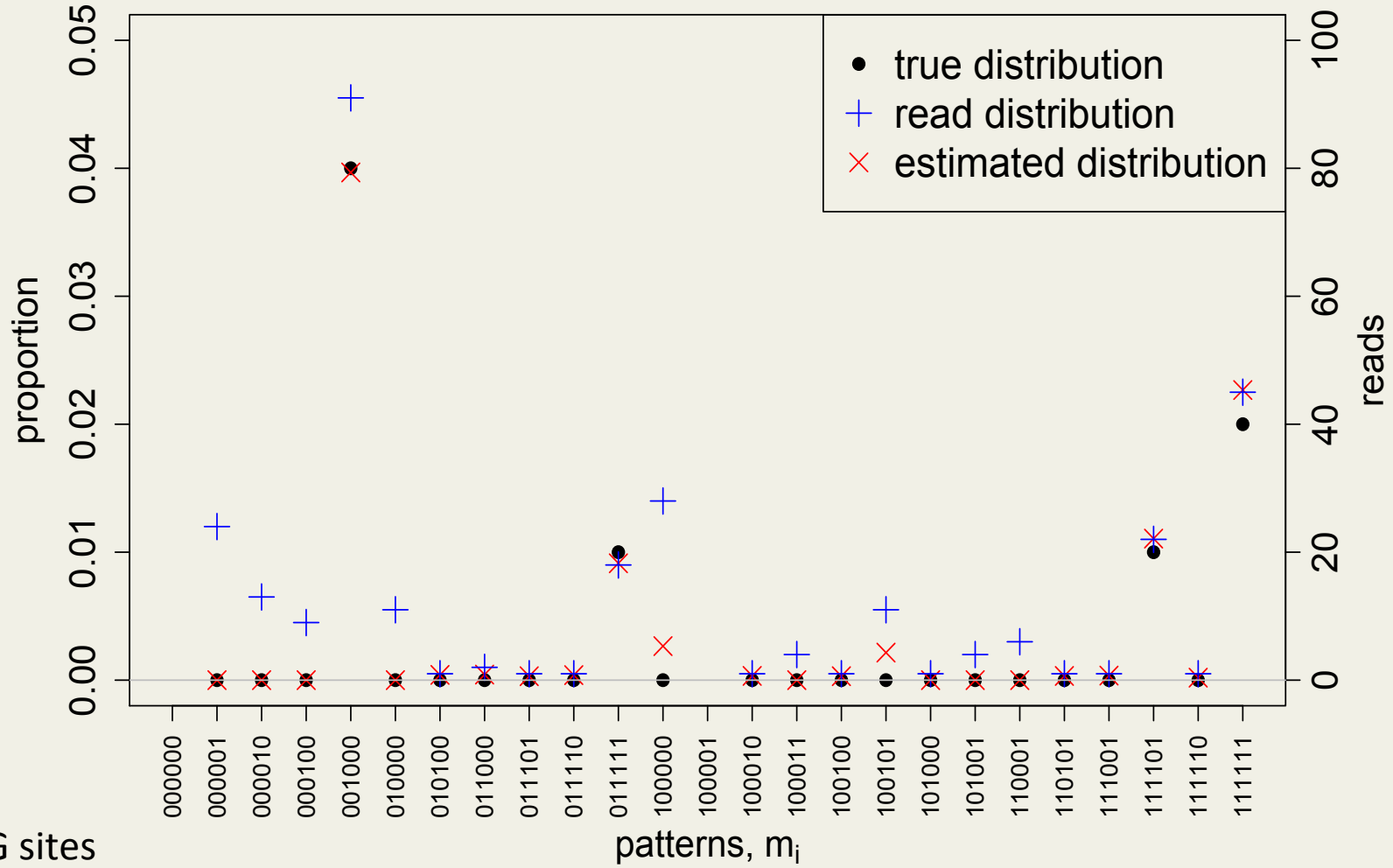
$$L\left(\theta\middle|\mathbf{Y}=\mathbf{y}\right)=\log\left(\Pr\left(\mathbf{Y}=\mathbf{y}\middle|\theta\right)\right)\propto\sum_{l=1}^{2^n}y_k\log\left(\sum_{k=1}^{2^n}\theta_k M_{kl}\right)$$

subject to the important constraints

$$\sum_{k=1}^{2^n}\theta_k=1,\qquad\theta_k\geq 0.$$

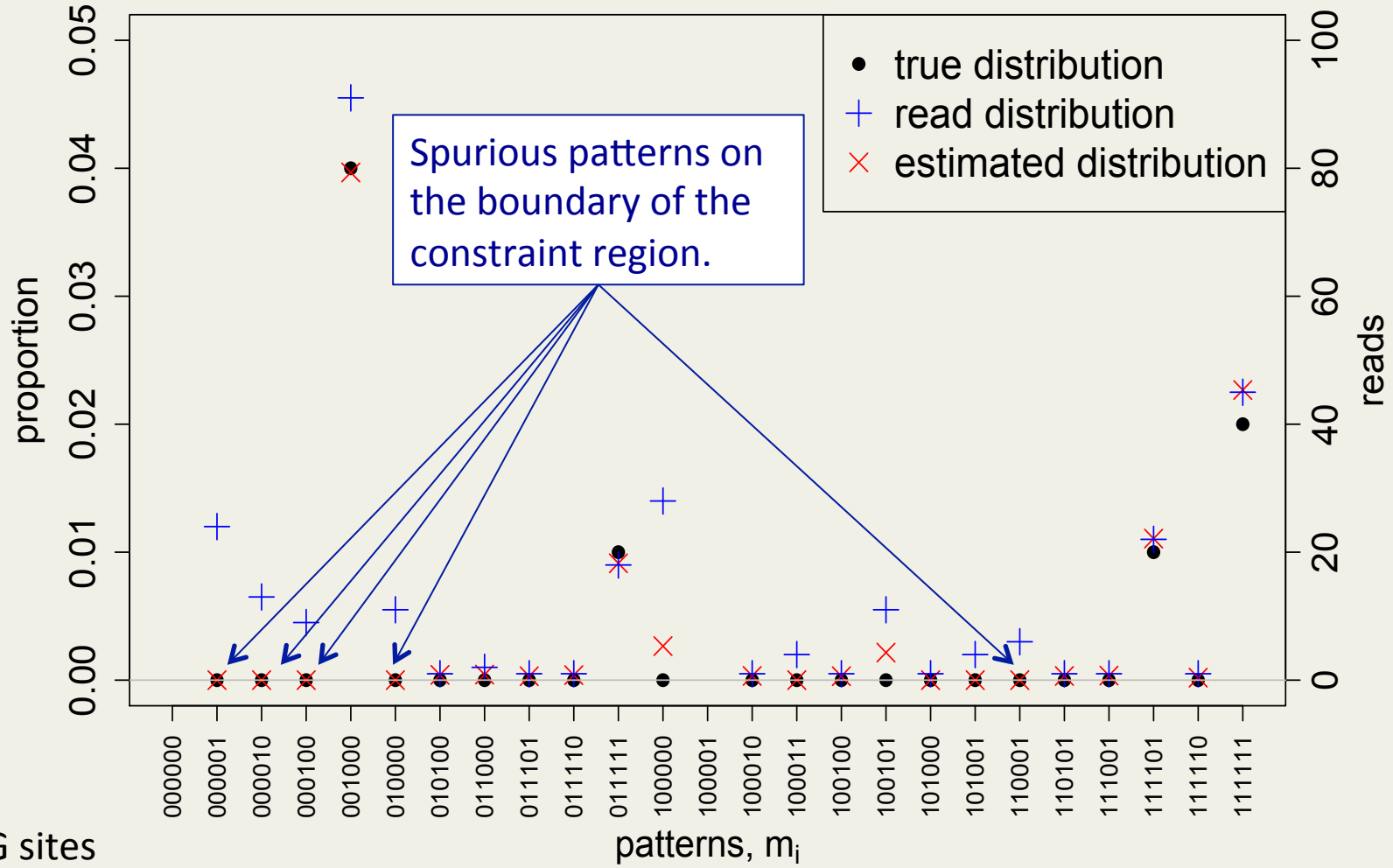
- Implemented in R using the function constrOptim()
- For realistic data the estimate of $\theta_k$ is generally on the boundary of the constraint $\theta_k\geq 0$. I.e. there are many 'observed' patterns which turn out to be spurious.

# Synthetic Data




n = 6 CpG sites
total number of reads = 2000
non-conversion rate ε = 0.008
sequencing error rate η = 0.005
25 patterns, 9 are called spurious

# Synthetic Data




n = 6 CpG sites
total number of reads = 2000
non-conversion rate ε = 0.008
sequencing error rate η = 0.005
25 patterns, 9 are called spurious

Lin *et al. BMC Bioinformatics* (2015) 16:145
DOI 10.1186/s12859-015-0600-6

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

# Estimation of the methylation pattern distribution from deep sequencing data

Peijie Lin[1], Sylvain Forêt[2], Susan R Wilson[1,3] and Conrad J Burden[1*]

**Abstract**

**Background:** Bisulphite sequencing enables the detection of cytosine methylation. The sequence of the methylation states of cytosines on any given read forms a methylation pattern that carries substantially more information than merely studying the average methylation level at individual positions. In order to understand better the complexity of DNA methylation landscapes in biological samples, it is important to study the diversity of these methylation patterns. However, the accurate quantification of methylation patterns is subject to sequencing errors and spurious signals due to incomplete bisulphite conversion of cytosines.

**Results:** A statistical model is developed which accounts for the distribution of DNA methylation patterns at any given locus. The model incorporates the effects of sequencing errors and spurious reads, and enables estimation of the true underlying distribution of methylation patterns.

**Conclusions:** Calculation of the estimated distribution over methylation patterns is implemented in the R Bioconductor package `MPFE`. Source code and documentation of the package are also available for download at http://bioconductor.org/packages/3.0/bioc/html/MPFE.html.

**Keywords:** DNA methylation, Bisulfite sequencing, DNA methylation patterns, Epiallele

## Background

Epigenetic regulations are involved in a broad range of biological processes, including development, tissue homeostasis, learning and memory, as well as various diseases such as obesity and cancer [1-3].

DNA methylation is one of the best studied epigenetic molecular mechanisms. It consists of the addition of a methyl group to the cytosine residues (C) of a DNA molecule. In animals, DNA methylation usually takes place in the CpG context: cytosines followed by a guanine (G) residue.

DNA methylation modulates gene expression through

The diverse and subtle effects of DNA methylation enable a given genome to produce different phenotypic outputs as part of a developmental program or in response to environmental factors. This has fundamental implications at the organismal level, where DNA methylation plays an important role in phenotypic plasticity [6]. This is also important at the cellular level to create diverse cell types, tissues and organs all based on the same genome. DNA methylation patterns can thus change from one cell type to another or within a cell under different conditions [7].

The diversity of methylation patterns in a sample can be

Lin *et al. BMC Bioinformatics* (2015) 16:145
DOI 10.1186/s12859-015-0600-6

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

# Estimation of the methylation pattern distribution from deep sequencing data

Peijie Lin[1], Sylvain Forêt[2], Susan R Wilson[1,3] and Conrad J Burden[1*]

**Abstract**

**Background:** Bis
methylation states
information than
the complexity of
methylation patte
and spurious sign

**Results:** A statisti
given locus. The
the true underlyin

**Conclusions:** Ca
Bioconductor pac
http://bioconduct

**Keywords:** DNA

## Background

Epigenetic regulati
biological processes
ostasis, learning an
such as obesity and

DNA methylatio
molecular mechan
methyl group to t
molecule. In anim
place in the CpG c
(G) residue.

DNA methylatio

www.nature.com/scientificreports

SCIENTIFIC REPORTS

OPEN

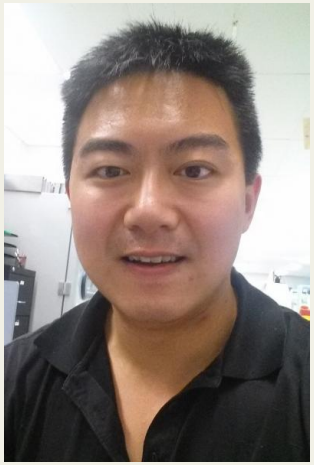
# EGFR gene methylation is not involved in Royalactin controlled phenotypic polymorphism in honey bees

Received: 18 June 2015
Accepted: 17 August 2015
Published: 11 September 2015

R. Kucharski, S. Foret & R. Maleszka

# Acknowledgements

Sylvain Forêt
RSB, ANU



Paul Lin
UNSW

Susan Wilson
ANU & UNSW

We thank Prof. Maleszka for sharing sequencing data.

ARC Grants DP120101422 and DE130101450
NHMRC Grant NHMRC525453

Australian National University

Australian National University


# Lecturer or Senior Lecturer

**Job no:** 507316
**Work type:** Continuing
**Location:** Canberra / ACT
**Categories:** Academic

Apply now

Existing a
Logi

search for jobs

e.g 'Administration', 'IT'

## Categories

Academic (1)

## Work type

Continuing (1)

## Locations

Canberra / ACT (1

**Classification:** Academic Levels B or C

**Salary package:**

Level B $91,541 - $104,254 plus 17% superannuation

Level C $110,610 - $123,325 plus 17% superannuation

**Term:** Continuing

**Position overview**

The Mathematical Sciences Institute at the Australian National University is seeking to invigorate its research and teaching profile in the areas of statistics, probability, stochastic analysis, mathematical finance and/or biomathematics/biostatistics. We wish to fill several continuing positions at the Academic Level B and/or Level C (which equates to the position of Associate Professor within the United States of America). Up to 3 full time positions may be awarded.