Walter & Eliza Hall Institute of Medical Research
Department of Statistics, UC Berkeley

A New Frontier

Understanding epigenetics through mathematics

Overview

Origin of this talk

What is epigenetics?

Why should we care?

The role of mathematical sciences

Examples

Let's begin.

Origin of this talk

~7 years ago, an epigeneticist paid me
a visit and presented 12 slides entitled:
17dec2007Statistical challenges.ppt

We've been interacting since then, but
the field has grown rapidly.

We need more mathematical scientists to join in!

What is Epigenetics?

ἐπί : Greek, meaning
above, on, over, nearby, upon…

genetics: English, meaning
science of genes, heredity & variation in living organisms

I know this doesn't help a lot, so…

Who will explain to me the difference between genotype and phenotype?

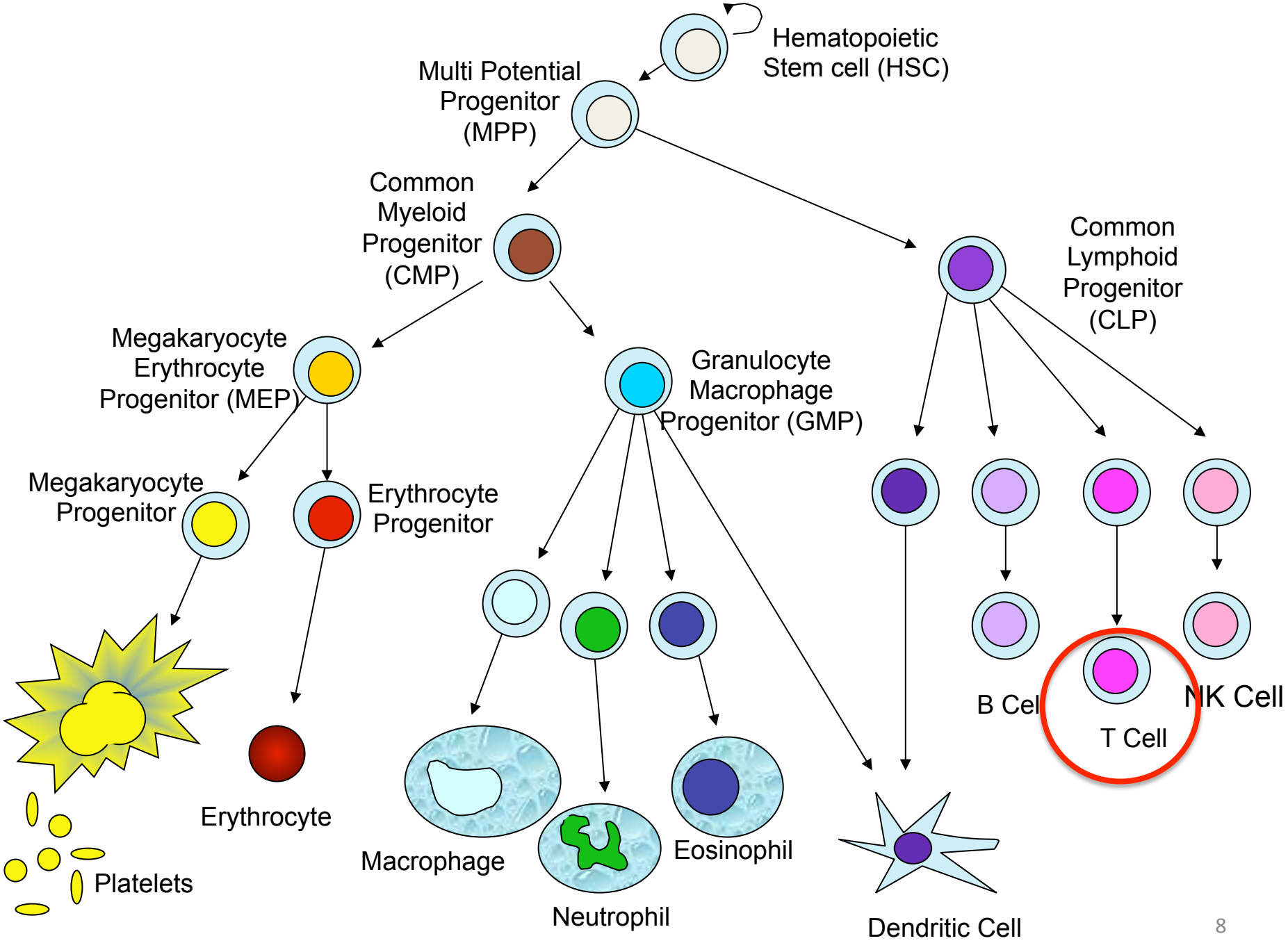
Tortoiseshell cats are ♀, heterozygous for Oo on the X chromosome.

Isogenic A^{vy}/a mice

Flowering of temperate plants after cold periods

Developing queen larvae surrounded by royal jelly

"The best example of epigenetic changes…
is the process of cellular differentiation."


Hematopoietic
Stem cell (HSC)
Multi Potential
Progenitor
(MPP)
Common
Myeloid
Progenitor
(CMP)
Common
Lymphoid
Progenitor
(CLP)
Megakaryocyte
Erythrocyte
Progenitor (MEP)
Granulocyte
Macrophage
Progenitor (GMP)
Megakaryocyte
Progenitor
Erythrocyte
Progenitor
B Cel
T Cell
NK Cell
Erythrocyte
Platelets
Macrophage
Neutrophil
Eosinophil
Dendritic Cell

Prehistoric and historic definitions

"… the branch of biology which studies the causal interactions between genes and their products, which bring the phenotype into being"
Waddington 1942

" Changes in gene expression inherited from cell to cell, not caused by DNA." Holliday, 1996

More or less contemporary definitions

"the structural adaptation of chromosomal regions so as to register, signal or perpetuate altered activity states"

Bird 2007

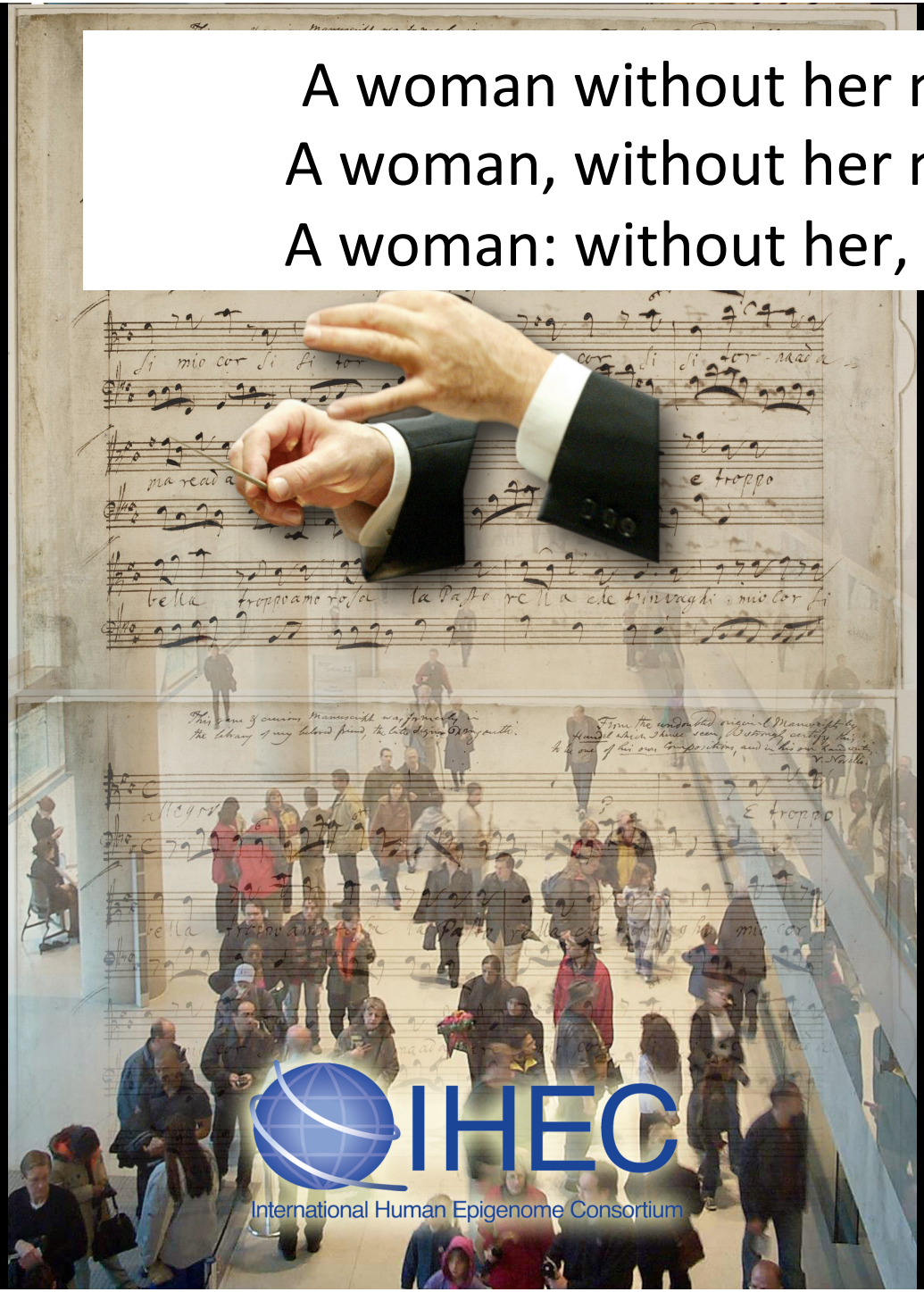
"heritable changes in gene activity and expression (in the progeny of cells or of individuals) and also stable, long-term alterations in the transcriptional potential of a cell that are not necessarily heritable"

www.roadmapepigenomics.org/overview 2008

"phenotypic variation that is not attributable to genetic variation".

Champagne 2010

A woman without her man is nothing.
A woman, without her man, is nothing.
A woman: without her, man is nothing.



Some analogies

score vs orchestra

DNA text vs punctuation

choreography vs dancer

....

compactness vs accessibility

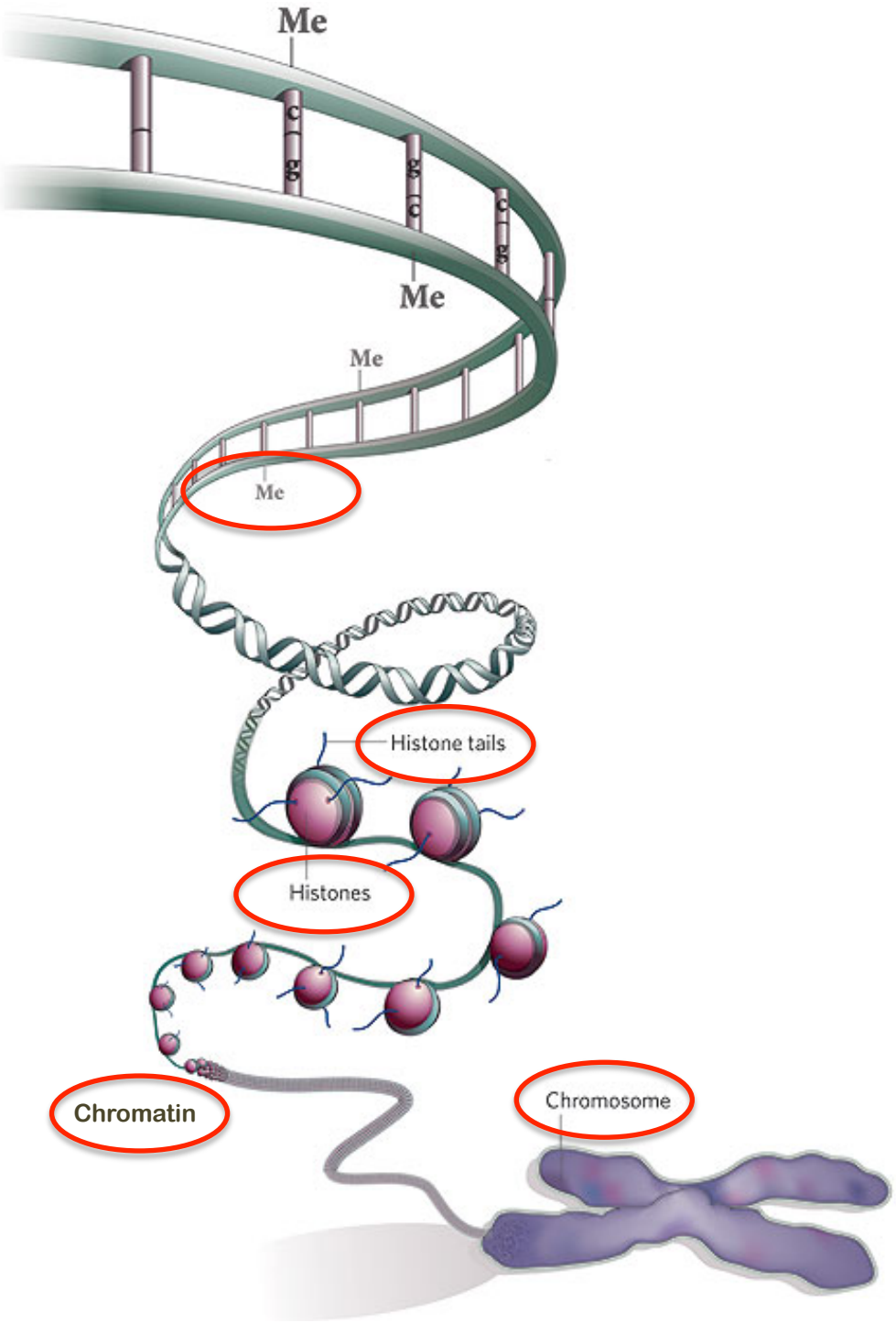
....

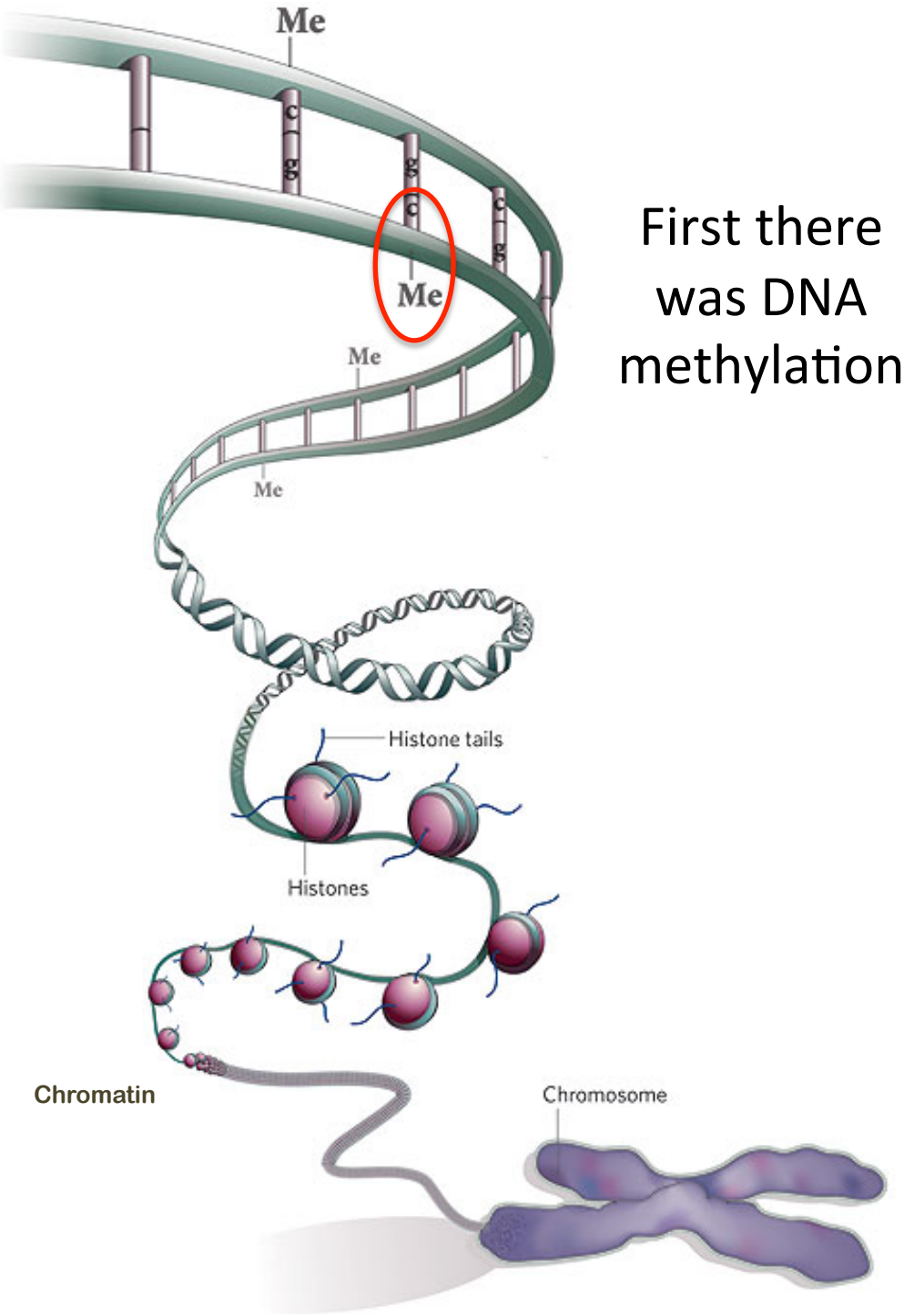
Biology's quantum mechanics

But what *is* Epigenetics?

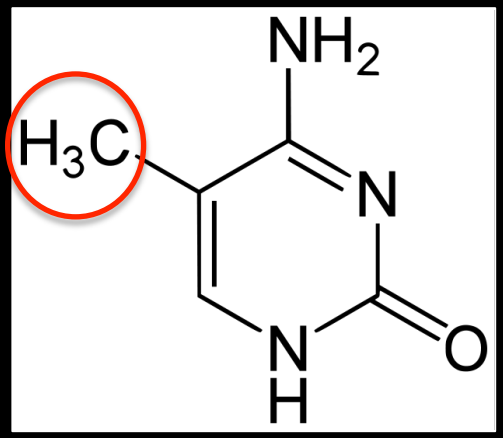
In one sense, it's ultimately about gene expression, what genes can and cannot be expressed in cells.

Let's look at a chromosome.


Me
Me
Me
Me
Histone tails
Histones
Chromatin
Chromosome

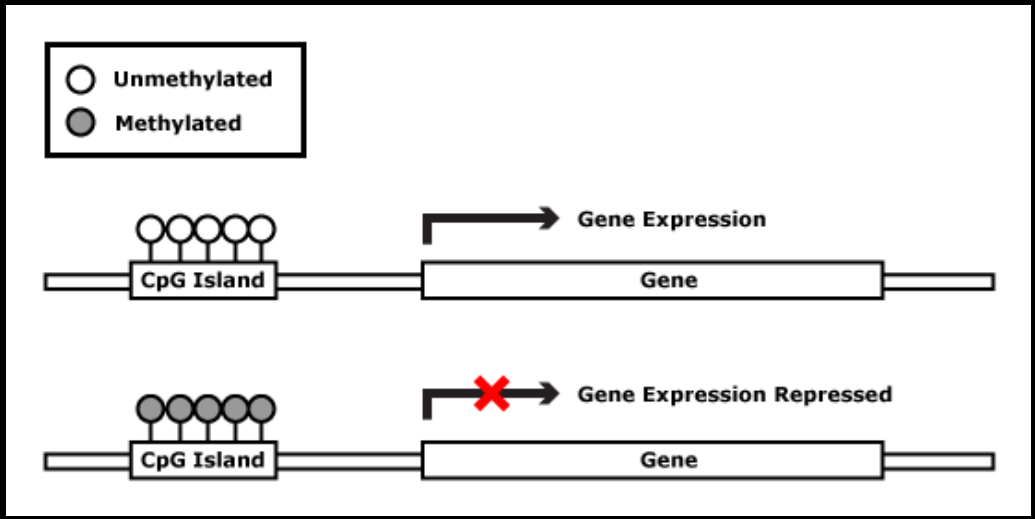


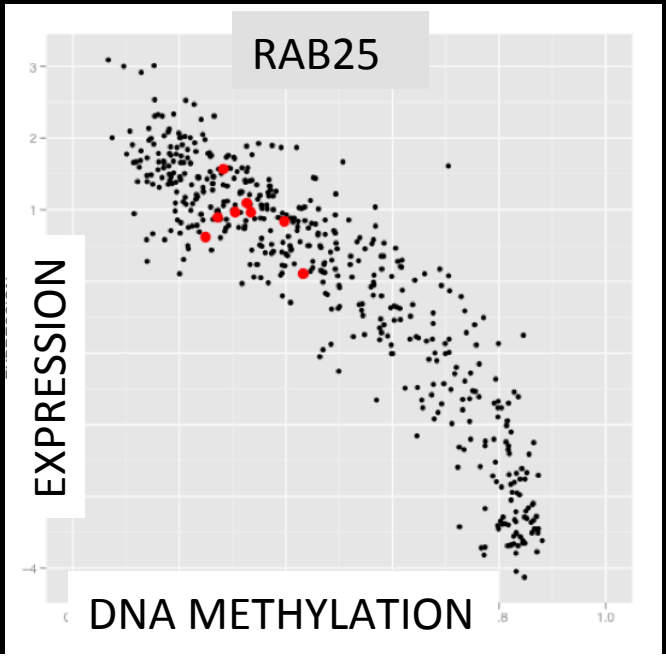
First there was DNA methylation



5-**Me**thyl **C**ytosine

Promoter methylation and gene expression

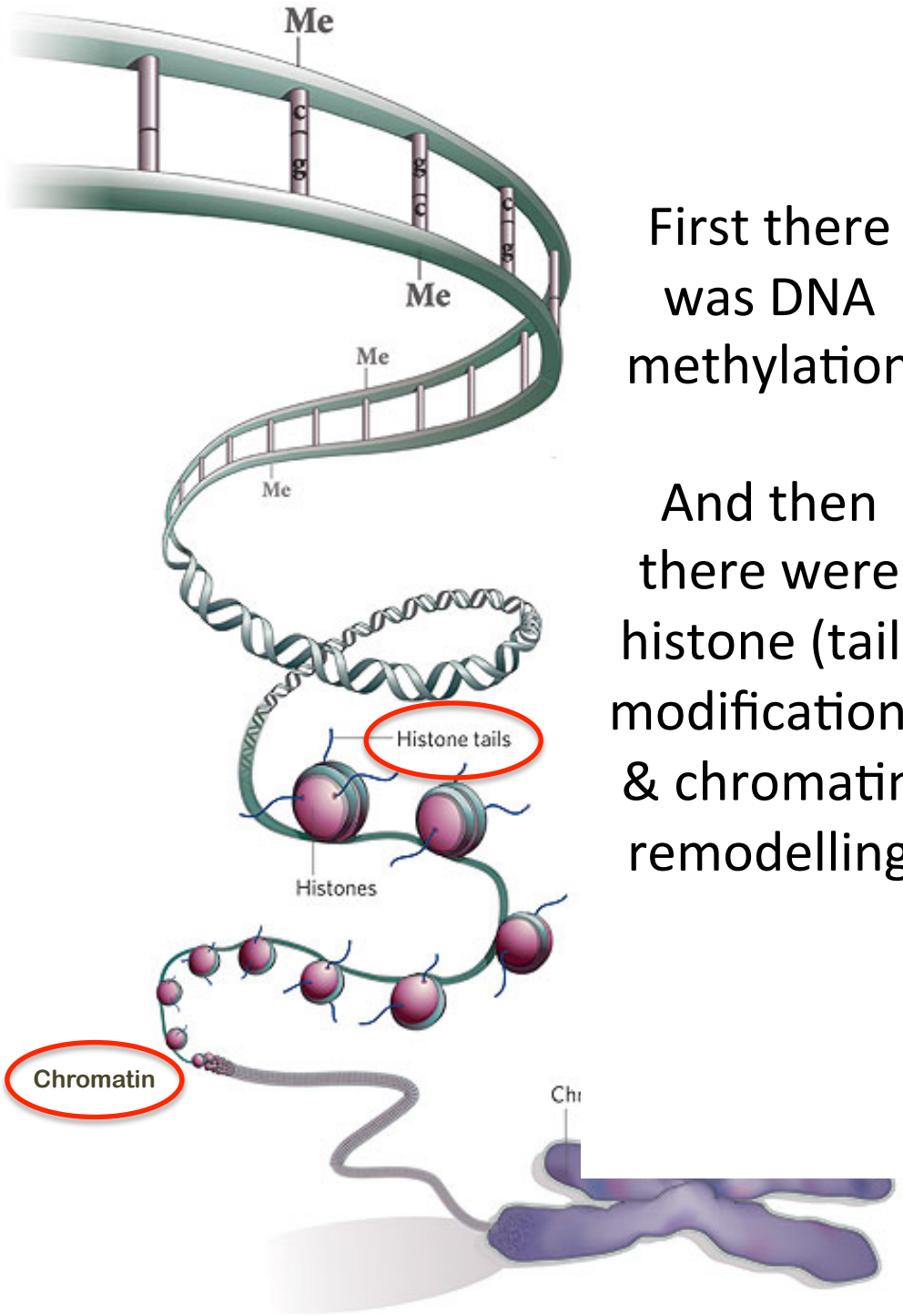




Gene EXPRESSION vs DNA METHYLATION at the promoter of RAB25: 489 Ovarian cancer tumors (+ 8 Fallopian tube samples).

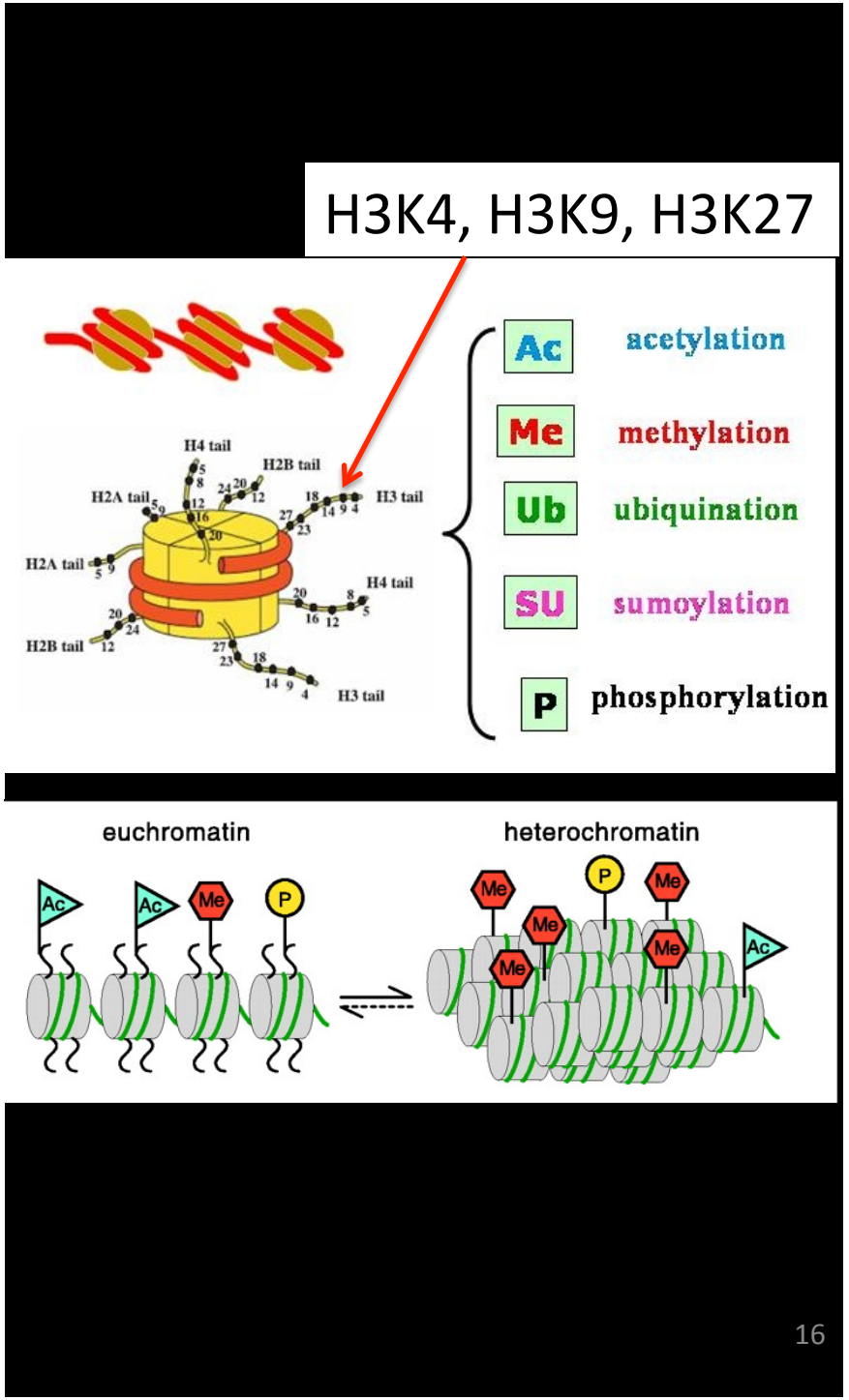
Figure from TCGA, Nature 2011

Not every such plot looks this good!

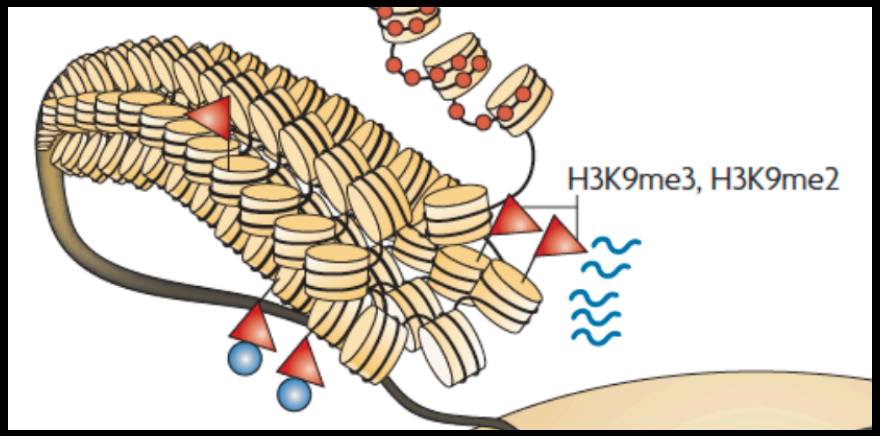


First there was DNA methylation

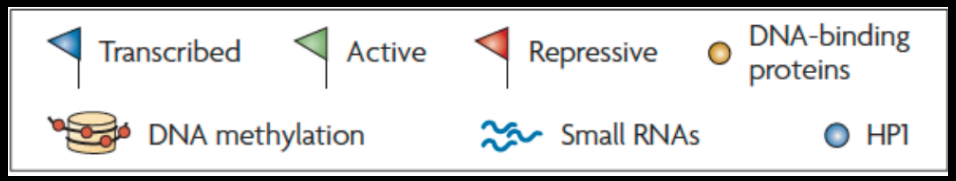
And then there were histone (tail) modifications & chromatin remodelling

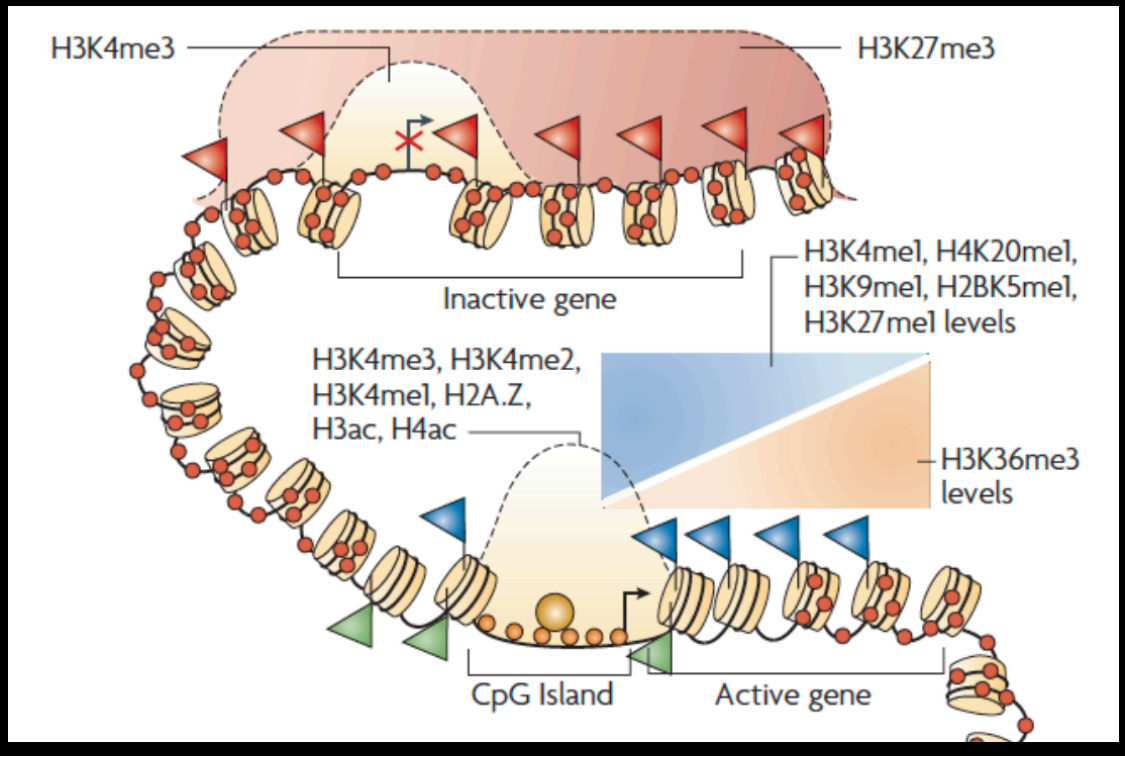


Chromatin state, histone marks & gene expression



Heterochromatin

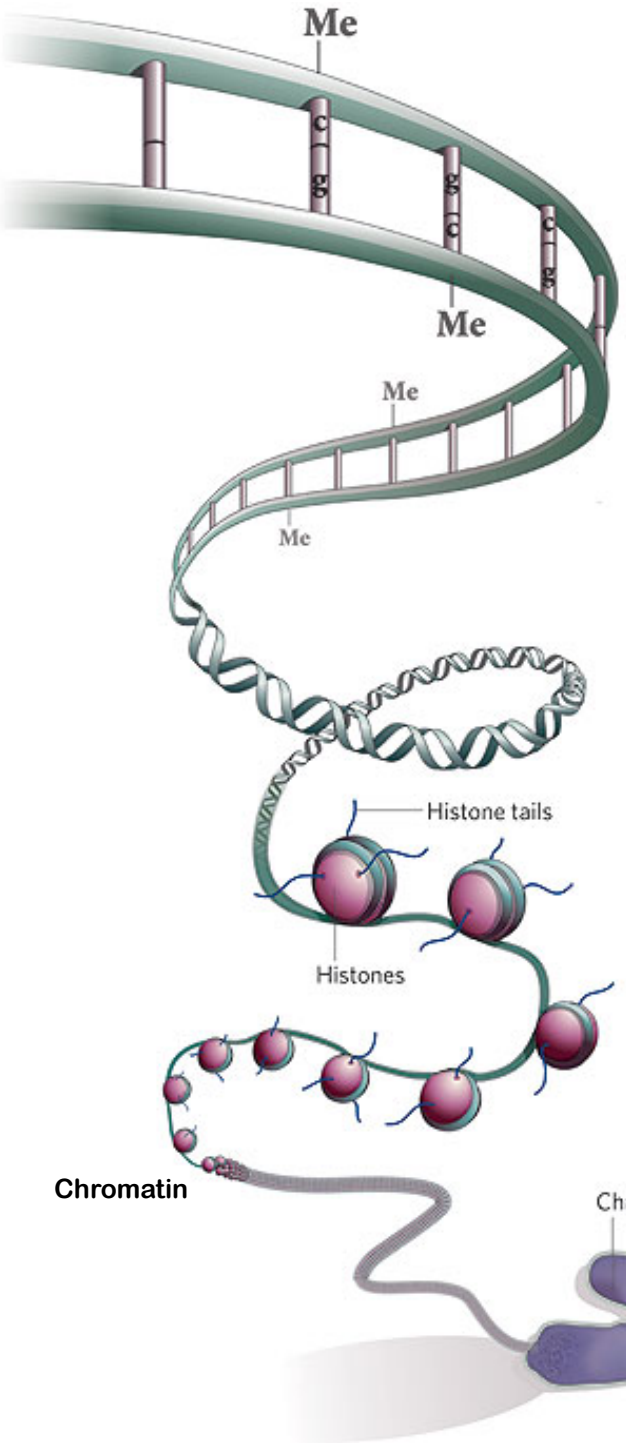




Euchromatin

with modifications permitting or preventing transcription

Schones & Zhao, 2008

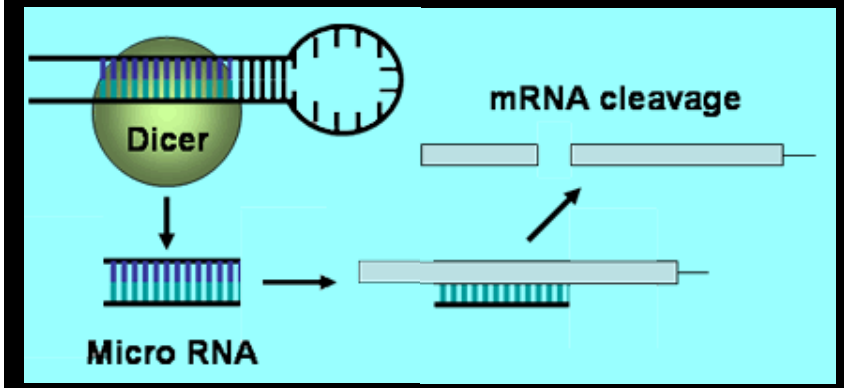


First there was DNA methylation

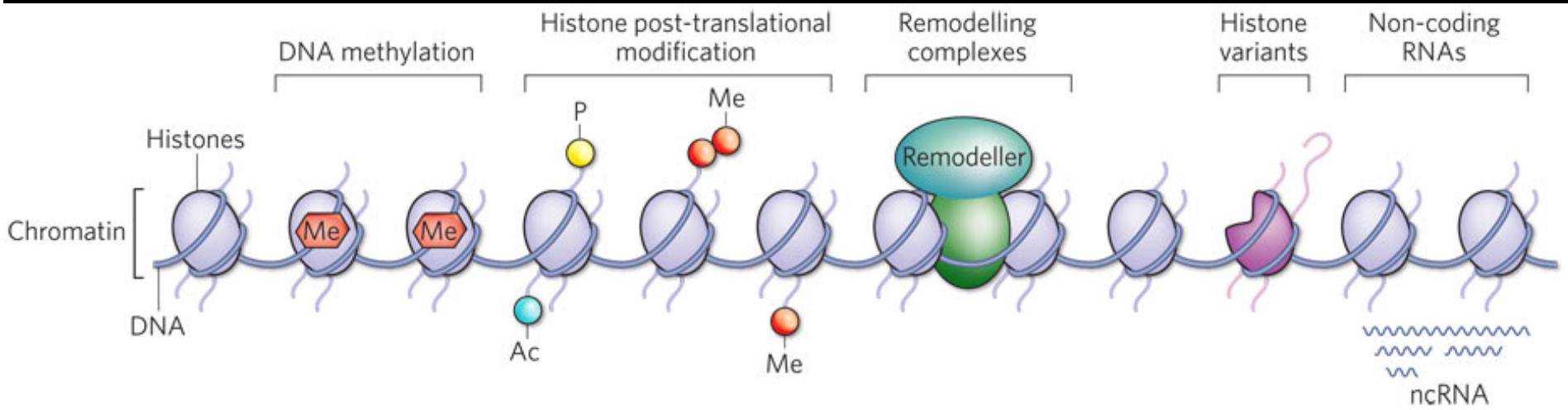
And then there were histone (tail) modifications & chromatin remodelling

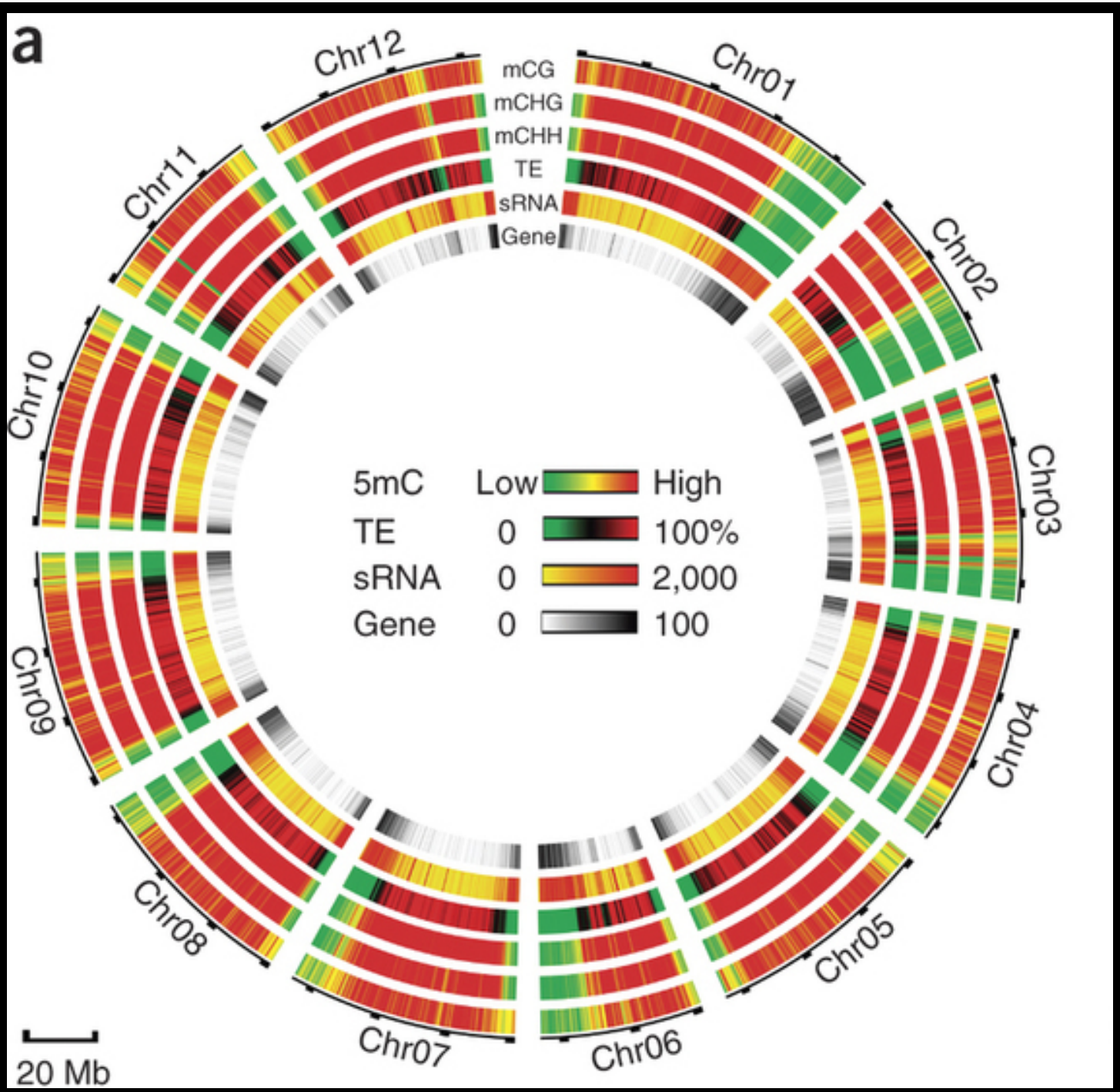
More recently, microRNAs

microRNAs and gene expression



Summary


DNA methylation
Histone post-translational modification
Remodelling complexes
Histone variants
Non-coding RNAs
Histones
Chromatin
DNA
Me
Me
P
Ac
Me
Me
Remodeller
ncRNA



Tomato

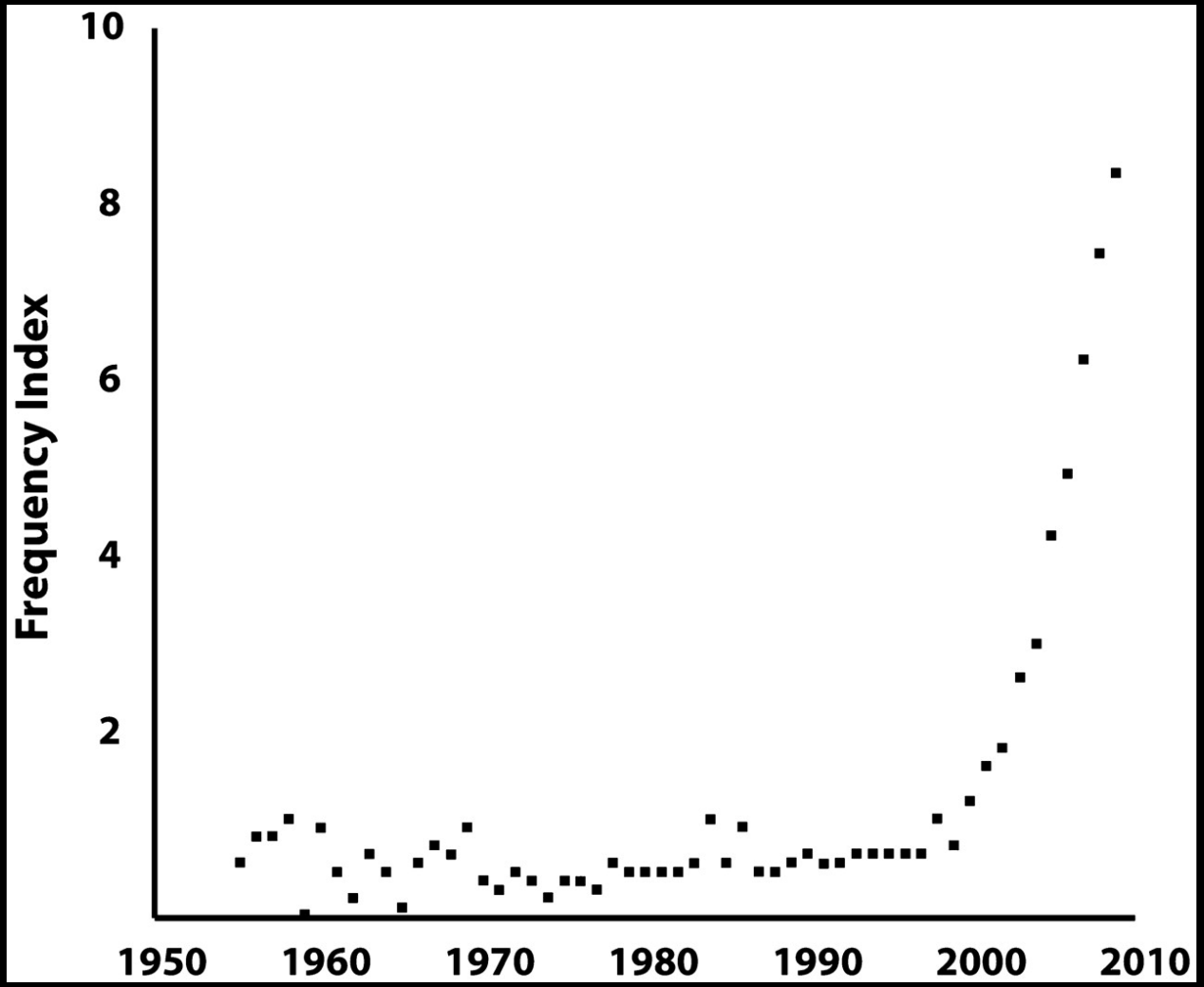
S. lycopersicum

Zhong *et al*,
NBT, 2013

Why do we care?

First, we do,…increasingly

Frequency of articles with *epigenetic* or epigenetics in their title by year, relative to *100 genetics* articles



Haig D Int. J. Epidemiol. 2012;41:13-16

Some early epigenetic phenomena

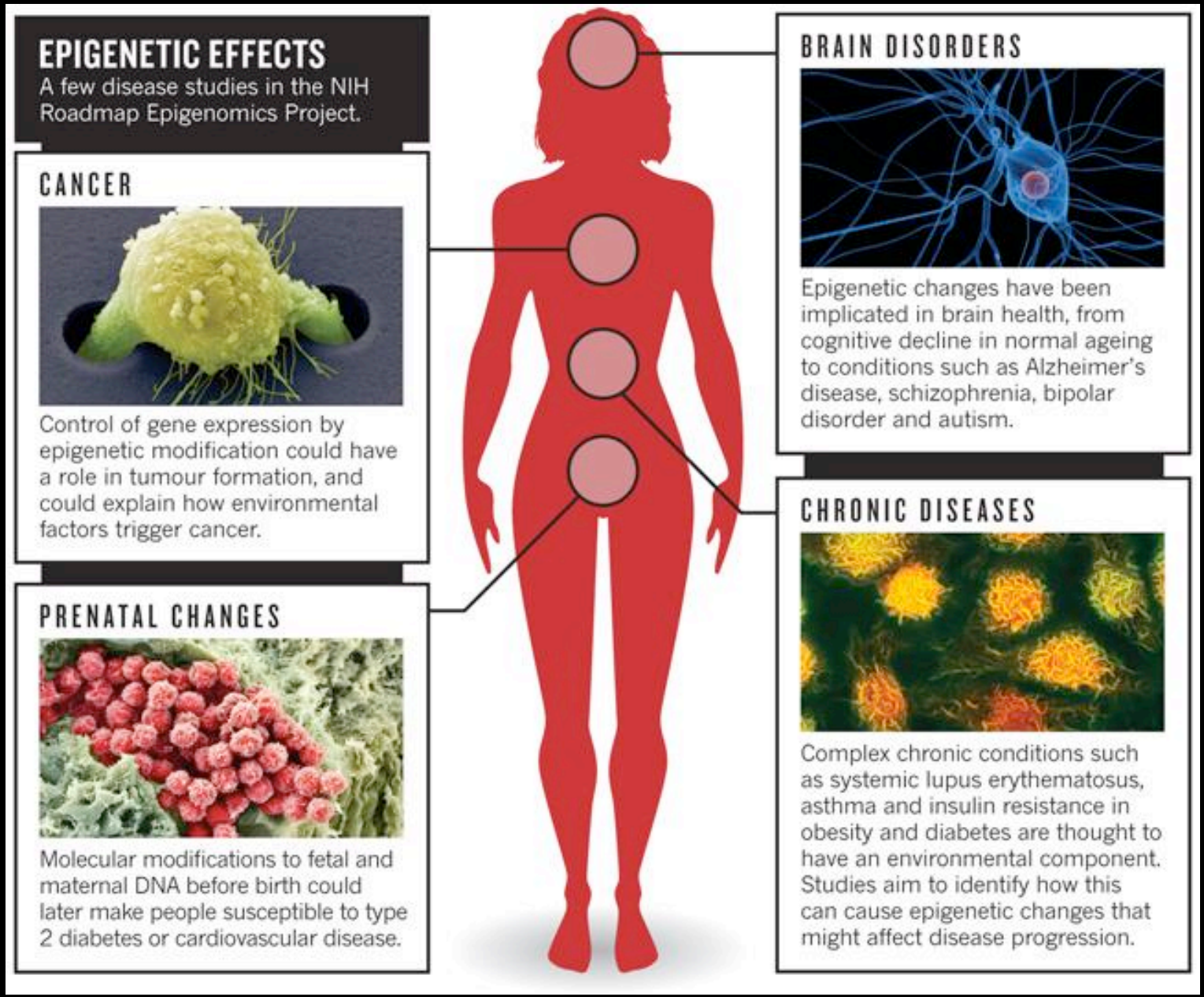
- Position effect variegation (1930s, Drosophila)
- Transposon silencing (Barbara McClintock, 1954) transposons are normally methylated
- X-chromosome inactivation (Mary Lyon, 1961)
- Imprinting, where parental origin of alleles matters, first identified in the 1980s.

More recent epigenetic phenomena

- The role of DNA methylation and epigenetics more generally in stem cells, cancer and aging
- Tissue specificity ≈ histone/chromatin remodelling code for cells
- Cellular memory, lineage-specific silencing

Other epigenetics

- Influence of the environment, e.g. diet (folate, alcohol) on the agouti viable yellow (A^{vy}) mouse; Royal jelly on honeybees; time/temperature on plant flowering (see later); maybe much more
- Emerging possibilities: long-term consequences of environmental exposure; why eating green vegies might protect against cancer,…
- Trans-generational epigenetic phenomena



From: Epigenome effort makes its mark, Nature 2010

The role of mathematical sciences

- Analysing raw epigenomic data, from microarrays and from DNA sequencing – there's a huge amount of this;
- Analysing epigenetic data from experiments or studies, e.g. comparing methylation at specific genes between treated and untreated mice; finding differentially methylated regions,
- Analysing epidemiological data, e.g. as in the Dutch Winter Famine
- Mathematical modelling of epigenetic phenomena

Examples will follow

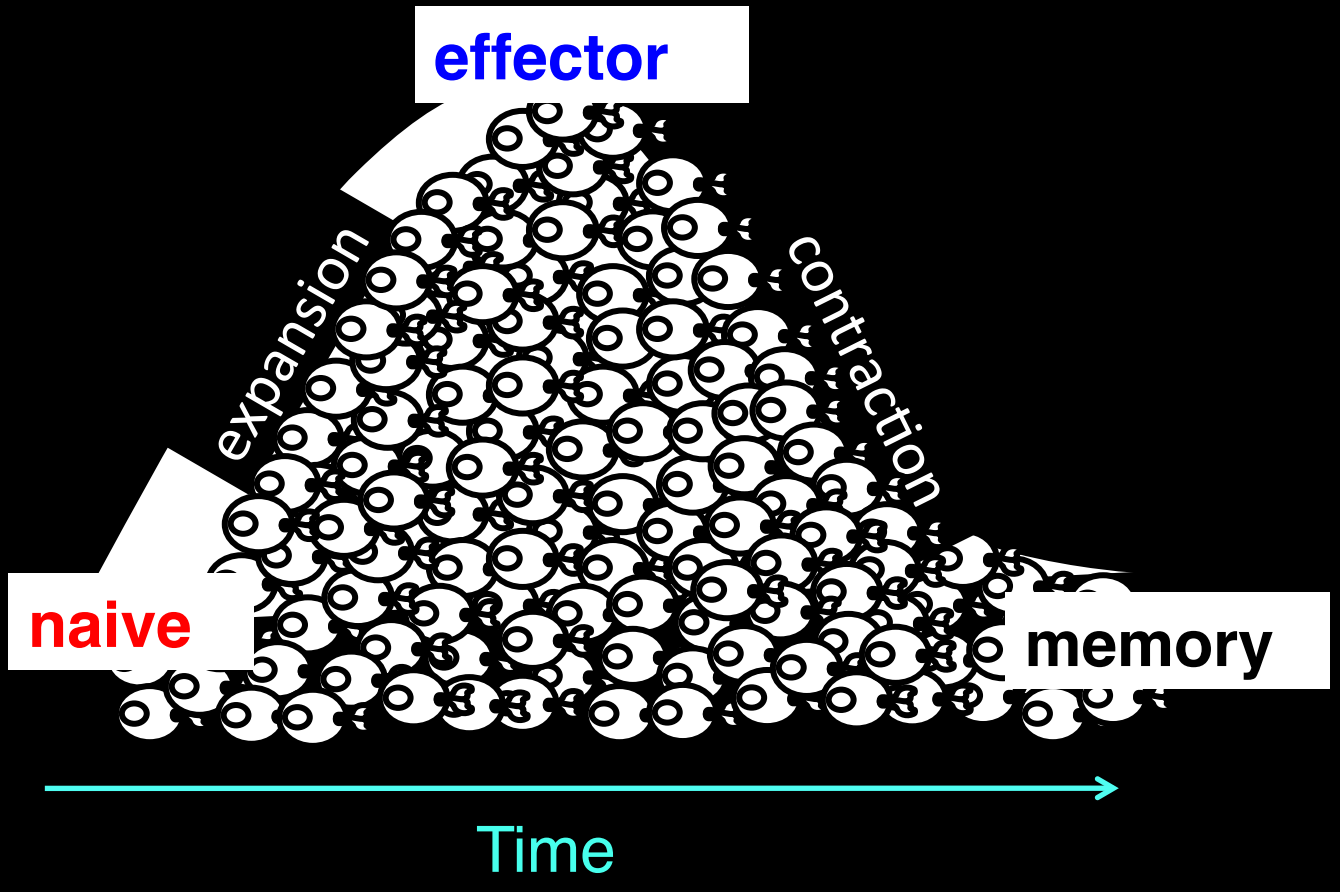
First, Analysing raw epigenomic data

Assaying histone modifications

The 5-year $190M ROADMAP Epigenomics PROJECT of the US NIH is focusing on 261 embryonic stem cell lines, fetal tissue and adult cells and tissues and 39 assays, including ChIP-seq for 30 histone modifications. Other nations and groups are doing similar things, some via the International Human Epigenome Project (IHEC).

This will just scratch the surface, as it's only baseline (molecular) data. All around the world, biomedical researchers are starting to explore the epigenetic dynamics of their favorite biological system, disease or phenomenon. I'll illustrate.

$CD8^+$ (= cytotoxic) T-cell differentiation following infection



Broad goal: a better understanding of immunological memory

H3K4me3 around a gene in CD8+ T-cells

Interferon gamma

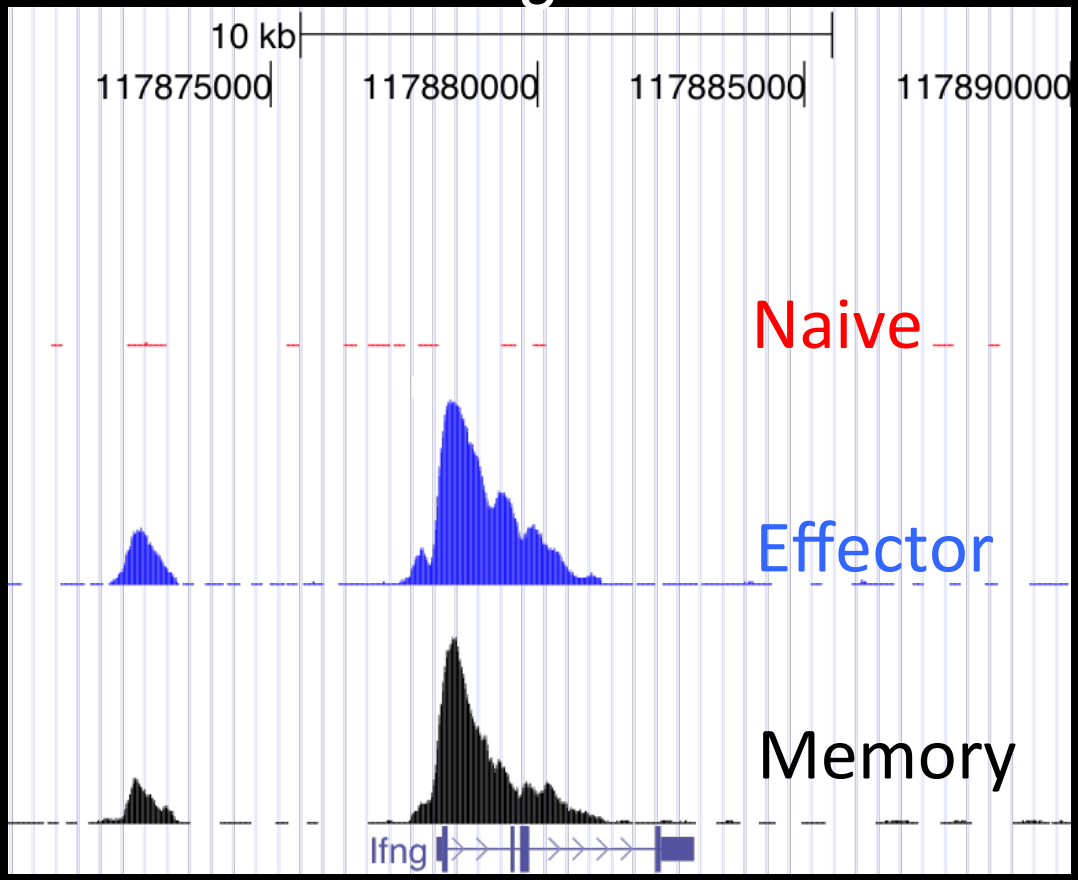


Figure courtesy of B Russ

Sample questions (all "genome-wide")

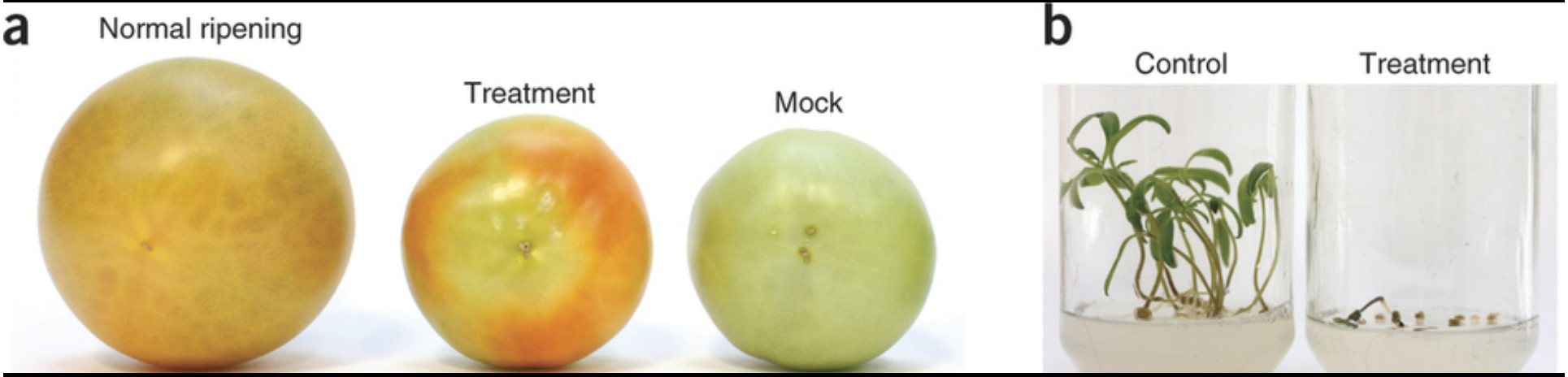
- How do gene expression changes between the different cell types correlate with the histone marks?
- In particular, which marks are present in genes that are up-regulated upon stimulation (in each cell type)?
- Which genes are bivalent, i.e. have both marks in Naïve cells, and how do they resolve, i.e. do they lose H3K27me3 and retain H3K4me3, or vice versa.
- Can we be make quantitative comparisons involving the marks? (Requires careful normalization.)

What makes a memory T-cell?

Analysing epigenetic data from experiments or studies

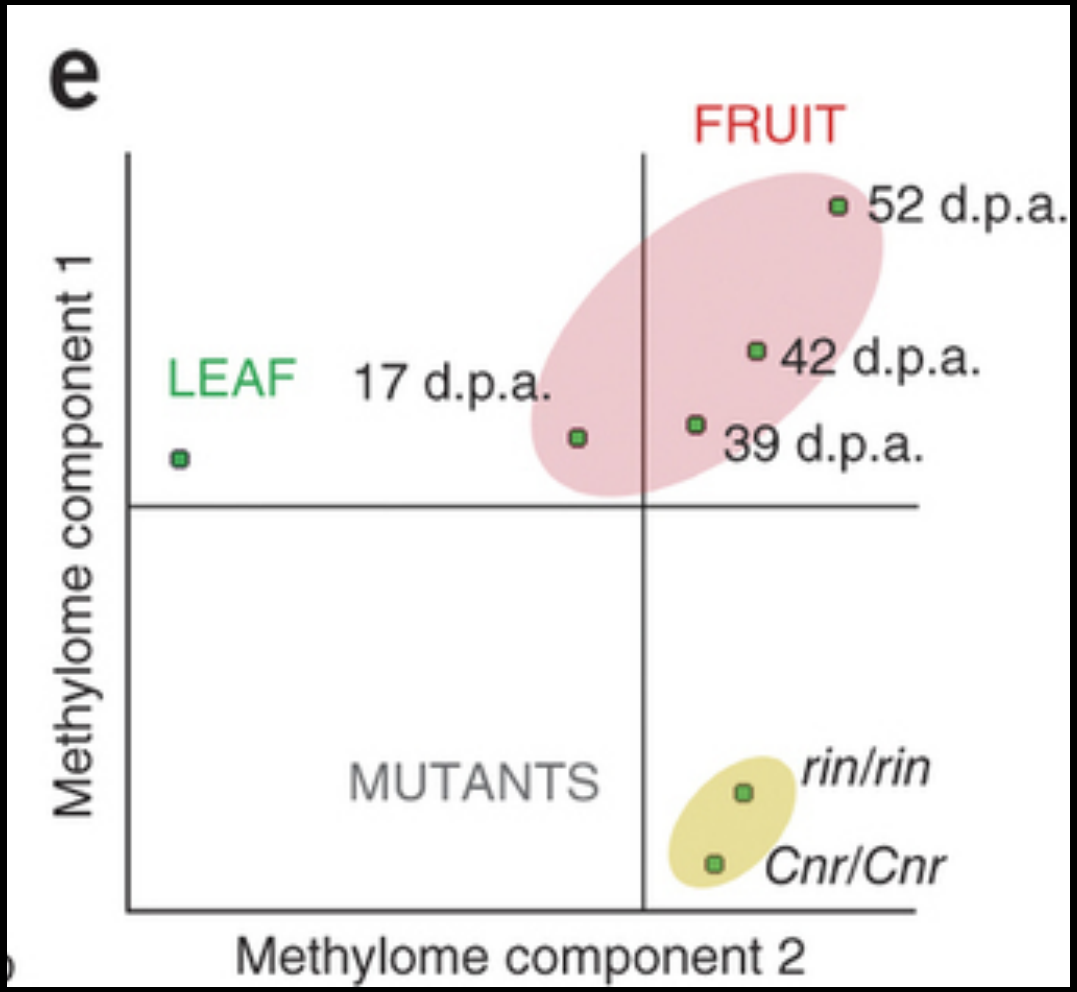
Single-base resolution methylomes of tomato fruit development reveal epigenome modifications associated with ripening

Silin Zhong[1,2,5], Zhangjun Fei[1,3,5], Yun-Ru Chen[1], Yi Zheng[1], Mingyun Huang[1], Julia Vrebalov[1], Ryan McQuinn[1], Nigel Gapper[1], Bao Liu[2], Jenny Xiang[4], Ying Shao[4] & James J Giovannoni[1,3]



Fruit treated with a methylation inhibitor ripen prematurely at 17 dpa (cf 42 dpa normal), but do not contain viable seeds.

Where is the maths here?



Finding the differentially methylated regions, plus…

Finding differences between 5 mice

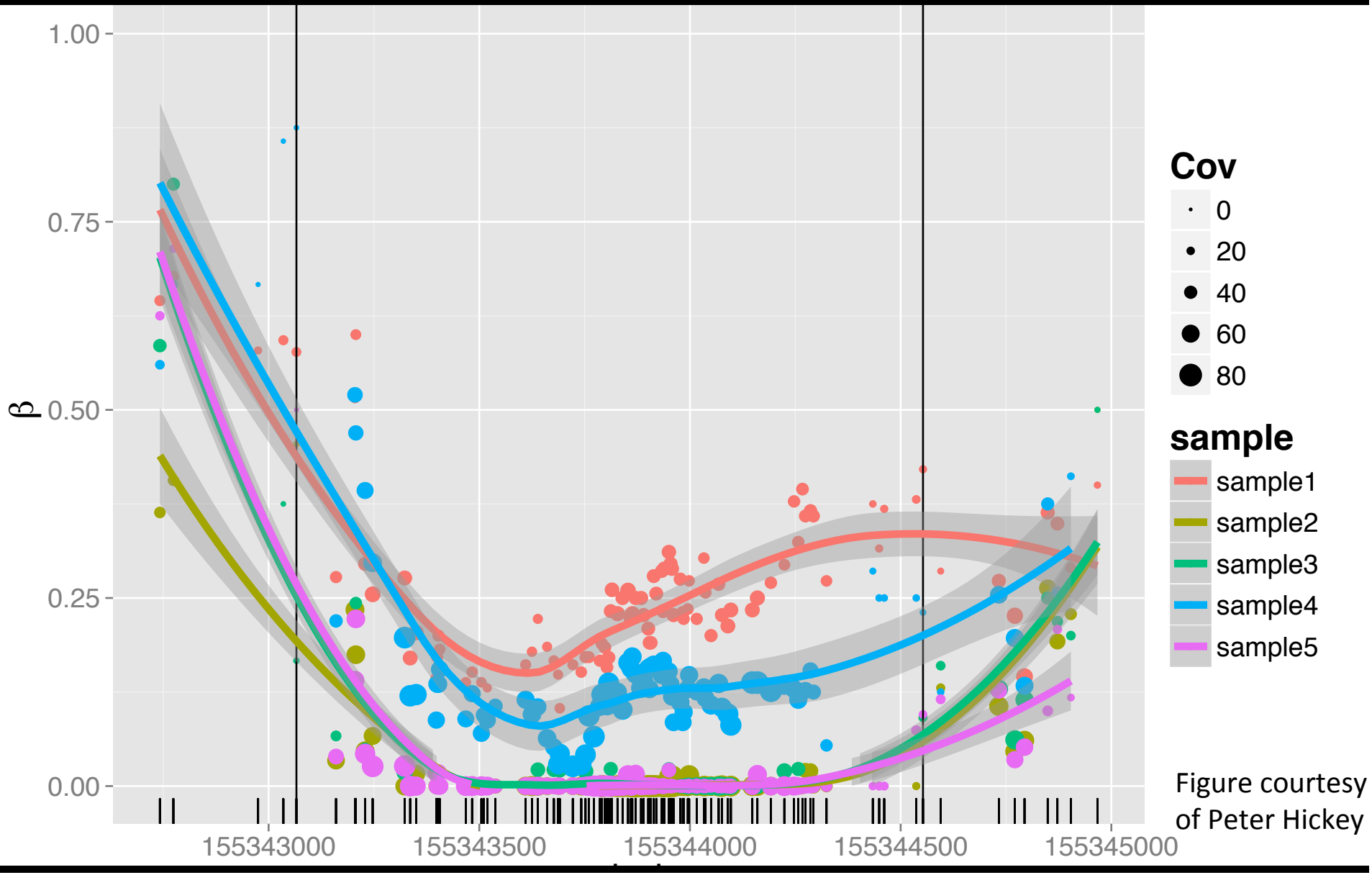
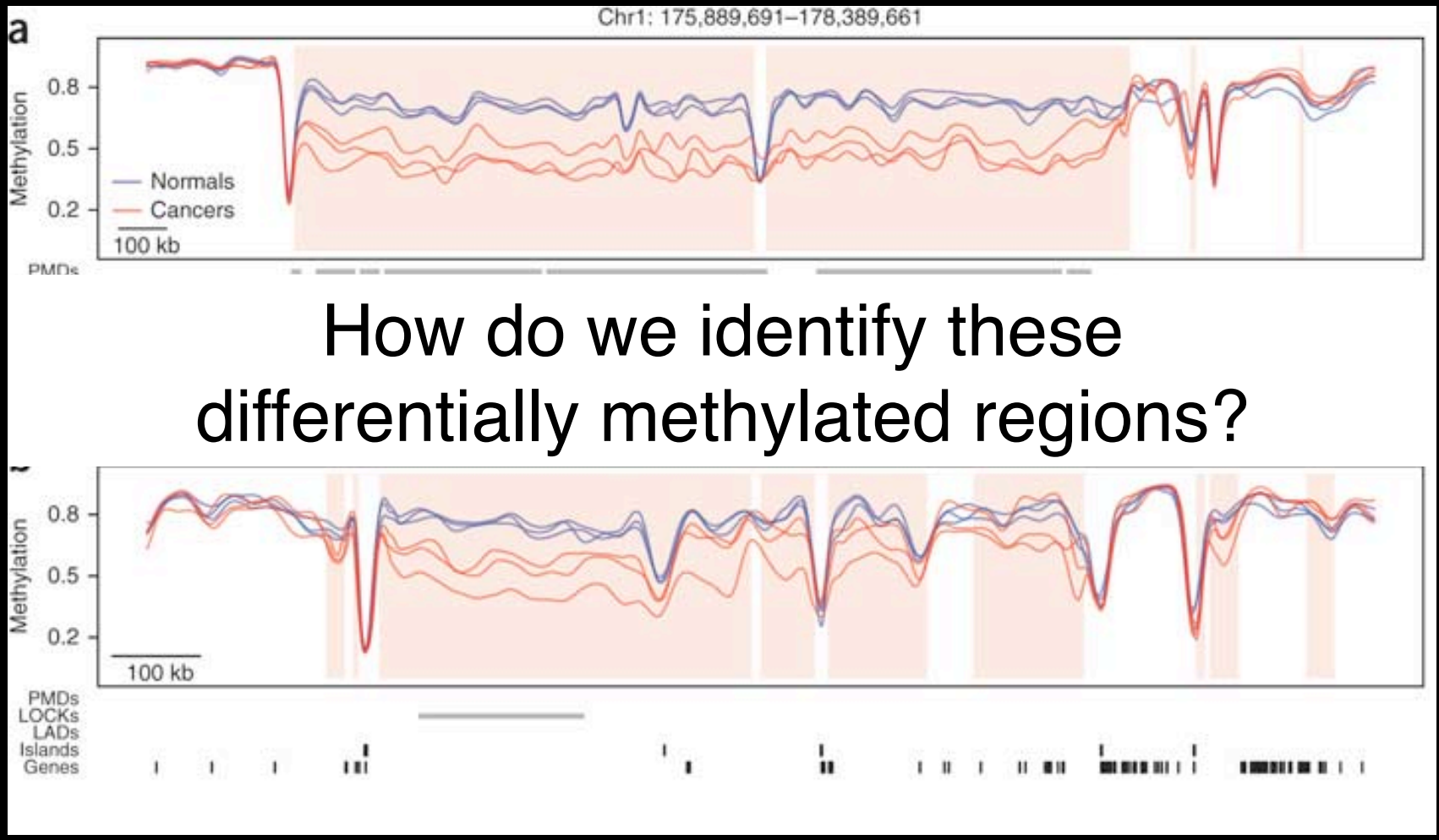


Figure courtesy of Peter Hickey

Human methylation: bisulphite-seq of 3 colon cancers vs 3 (paired) normals



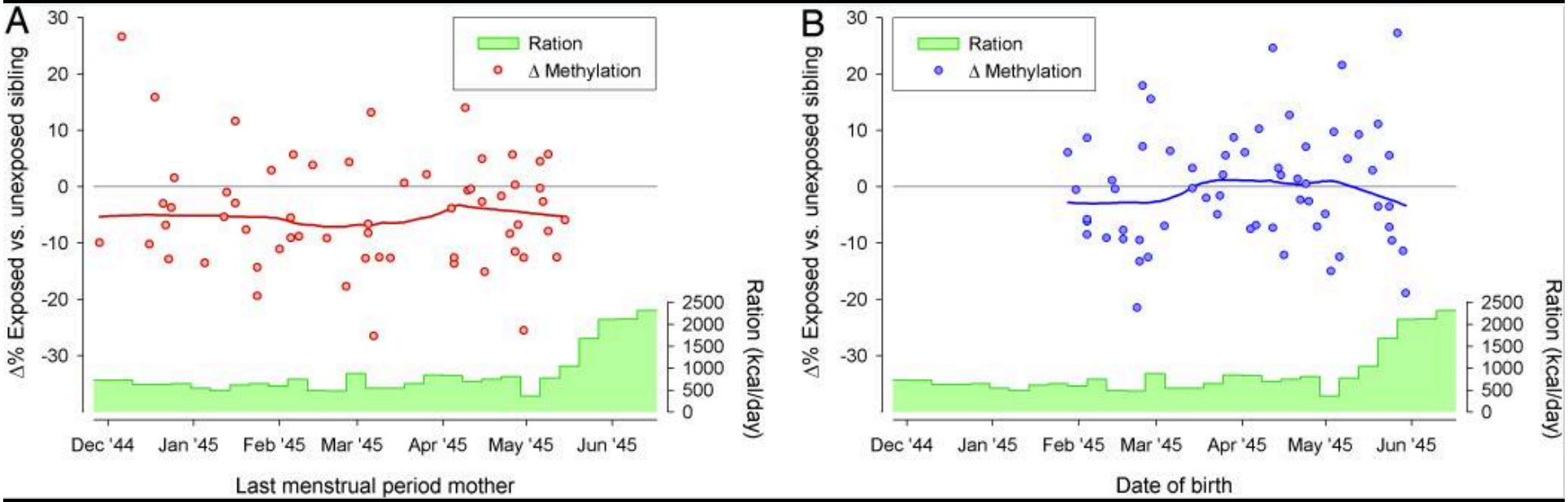
Hansen *et al*, Nat Gen 2011

Analysing epidemiological data

The Dutch Winter Famine (1944-45)

Due to a German food embargo. Registries and health care remained intact, and official food rations were documented.

Methylation differences apparent 60 years later, in a gene (IGF2) which codes for a growth hormone active during gestation



Vertical axis: Difference in methylation at insulin-like growth factor 2 between exposed and unexposed same sex sibs: exposed means "at conception." Effect present at conception, not later in gestation.

More historical data like this has been collected, and new, prospective studies are under way. Pinpointing epigenetic causes will be hard.

There are many challenging statistical questions associated with data like these.

I'd like to finish with a beautiful recent story from the U.K. about flowering.

Stochastic modelling of the system dynamics of vernalization

Work of Angel, Dean, Howard and Song

John Innes Centre, Norwich, UK

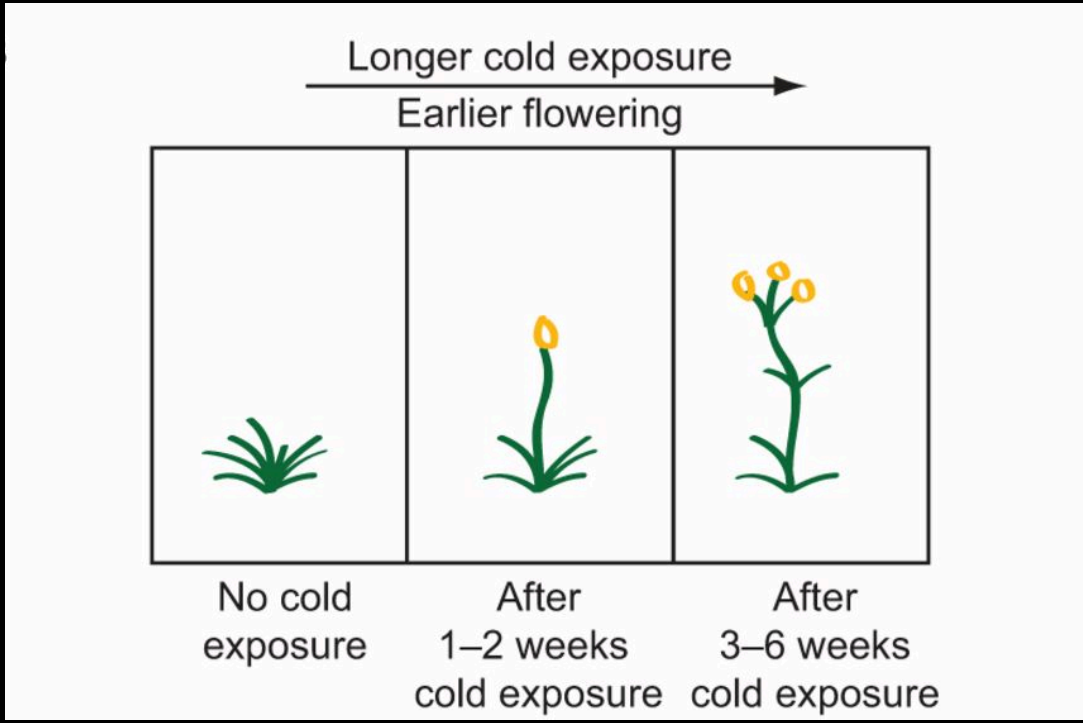
nature 2011 Journal of Cell Science jcs.biologists.org 2012

Please note that I'm simplifying, and that there are always exceptions.

https://www.youtube.com/watch?v=qEqdqXmMULw

Vernalization: promotion of flowering in response to prolonged low temperatures

Plants remember that they have experienced winter. Indeed they know how long winter lasted.

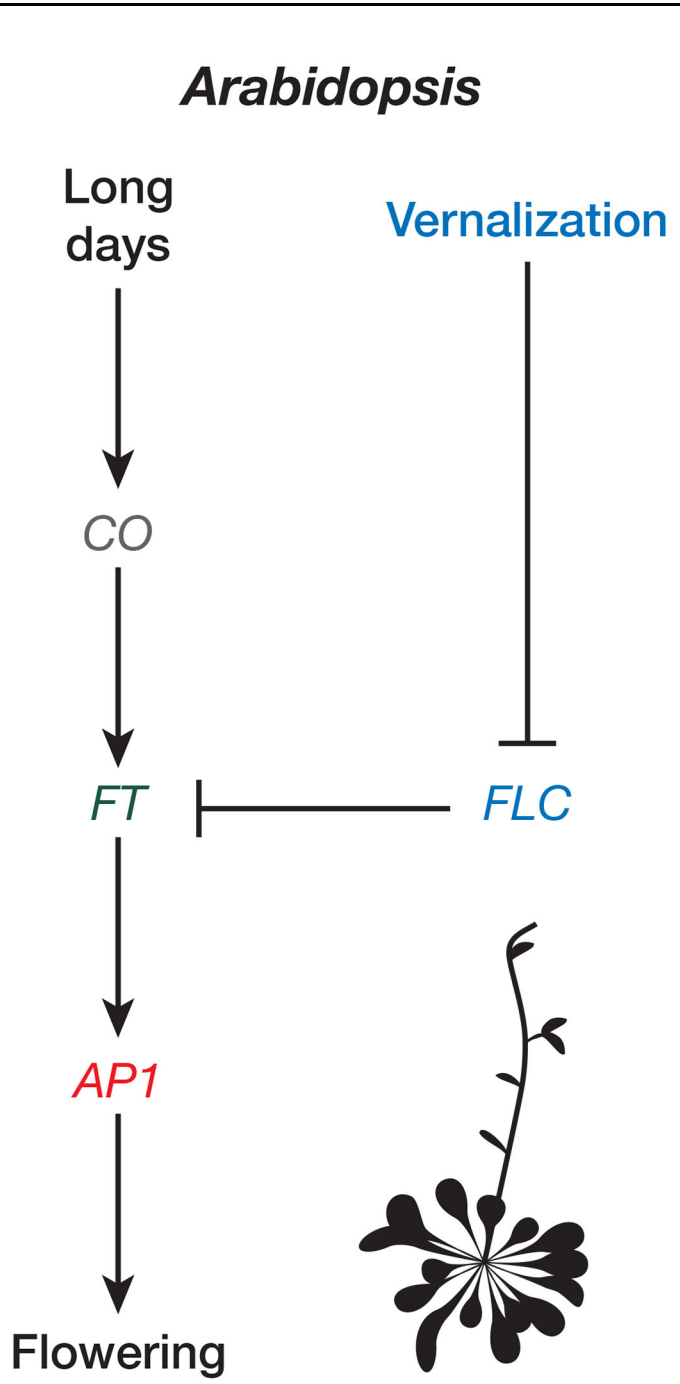


Images a fixed amount of time after cold exposure

CONSTANS

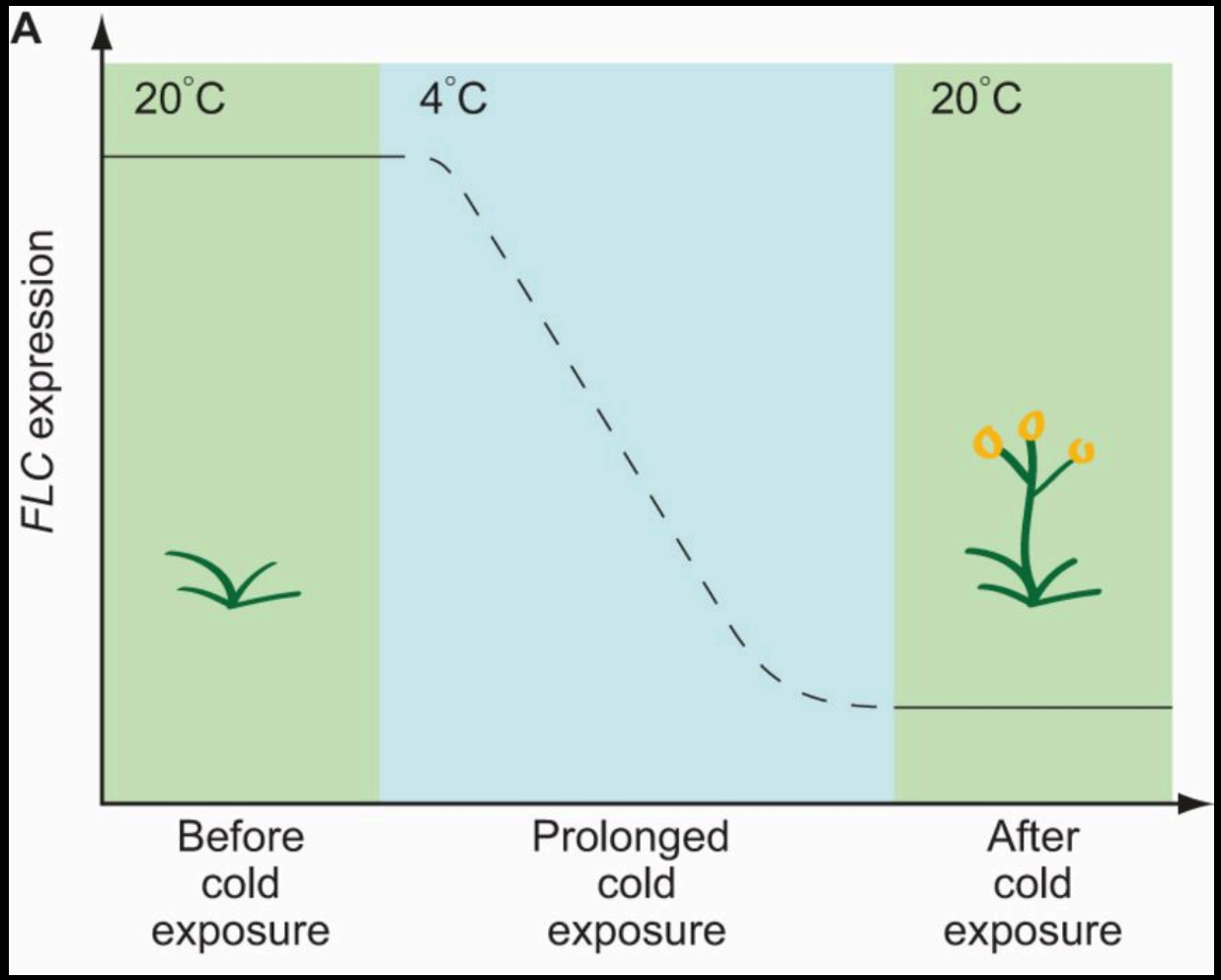
FLOWERING LOCUS T

APETALA1

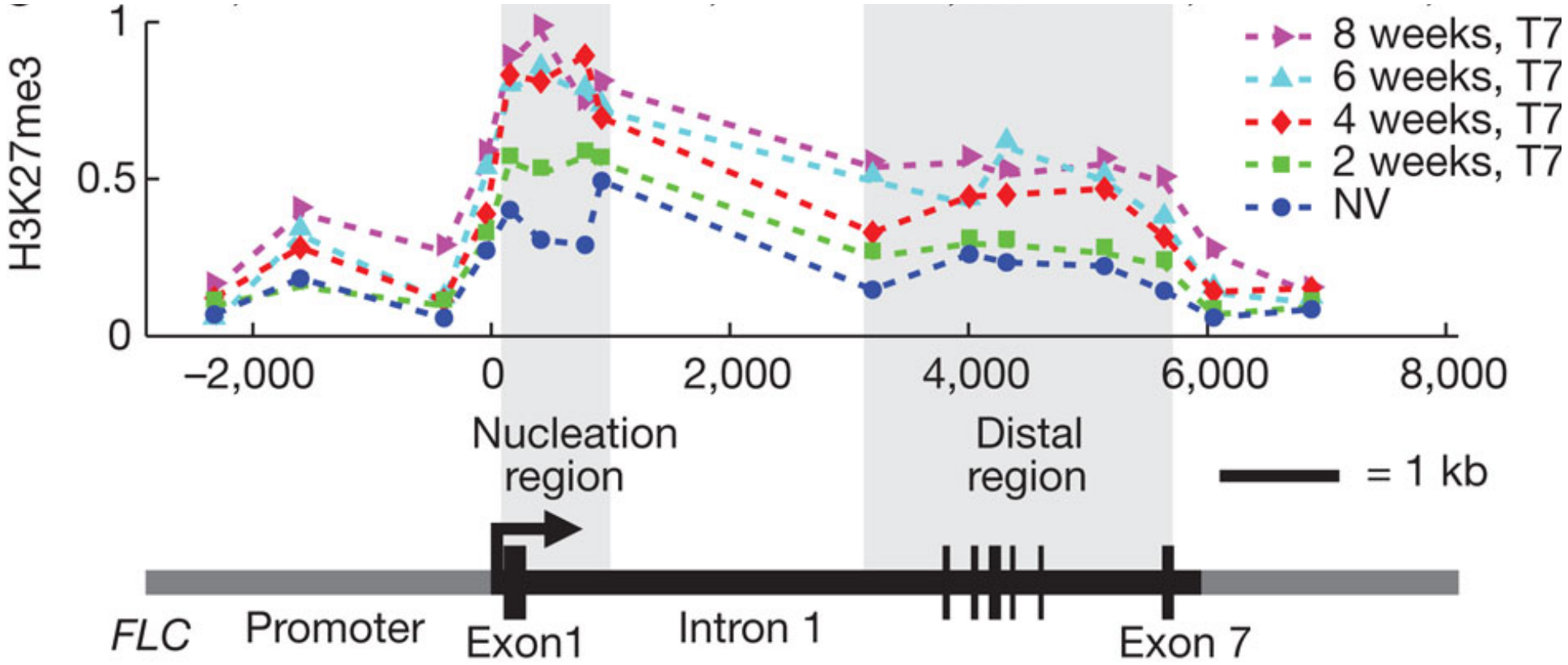


FLOWERING LOCUS C

Expression of Flowering Locus C (*FLC*)



H3K27me3 ChIP experiments.



A Angel *et al. Nature* **000**, 1-4 (2011) doi:10.1038/nature10241

nature

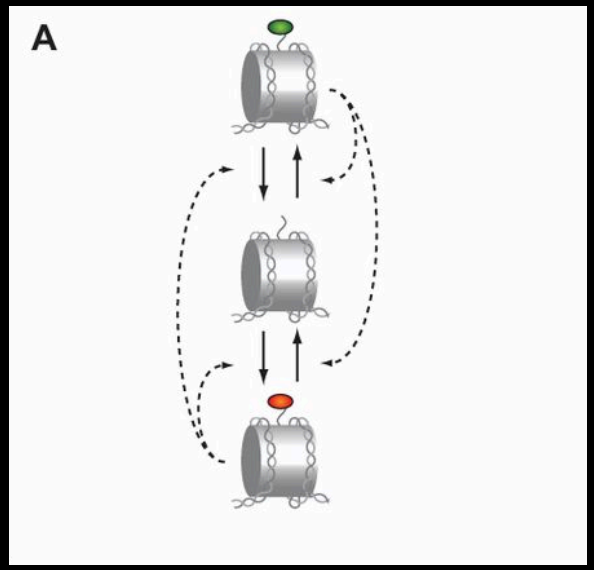
Plants can't count, so how do they measure the duration of winter?

The Innes Centre team showed by stochastic modelling how to establish stable epigenetic silencing (here of *FLC*) at a level that depends quantitatively on the level of a transient stimulus (here duration of cold),

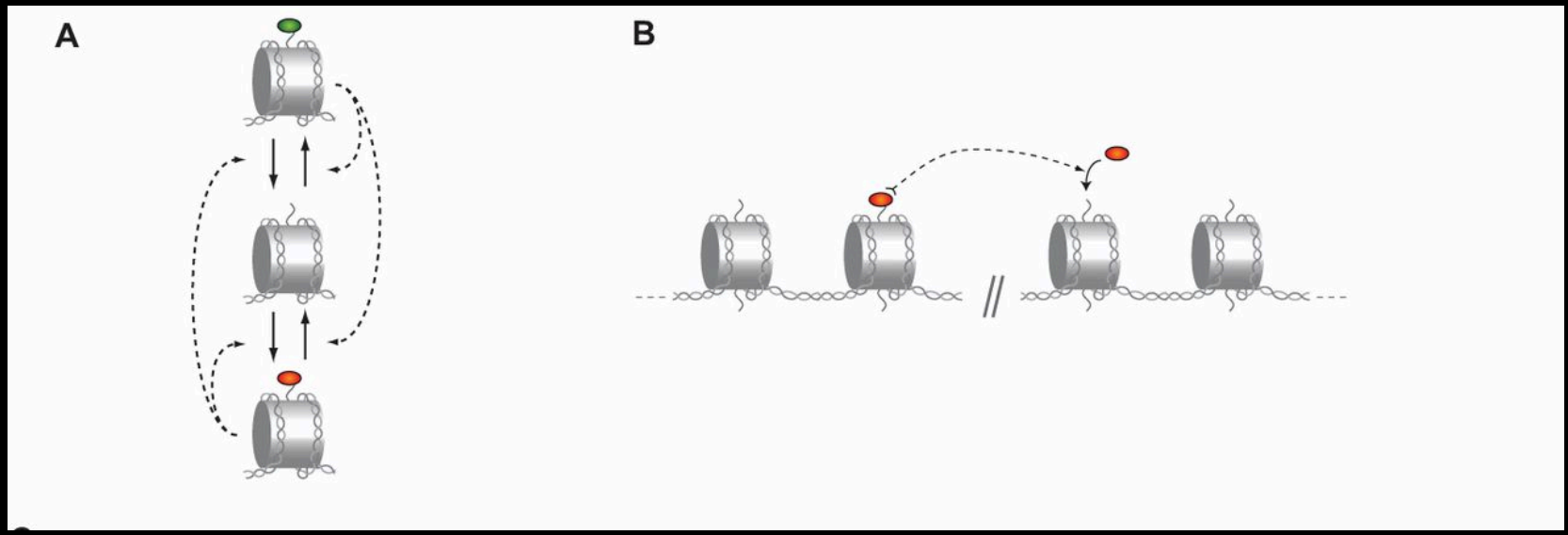
and they

supported their theoretical analysis with the statistical analysis of data from experiments.

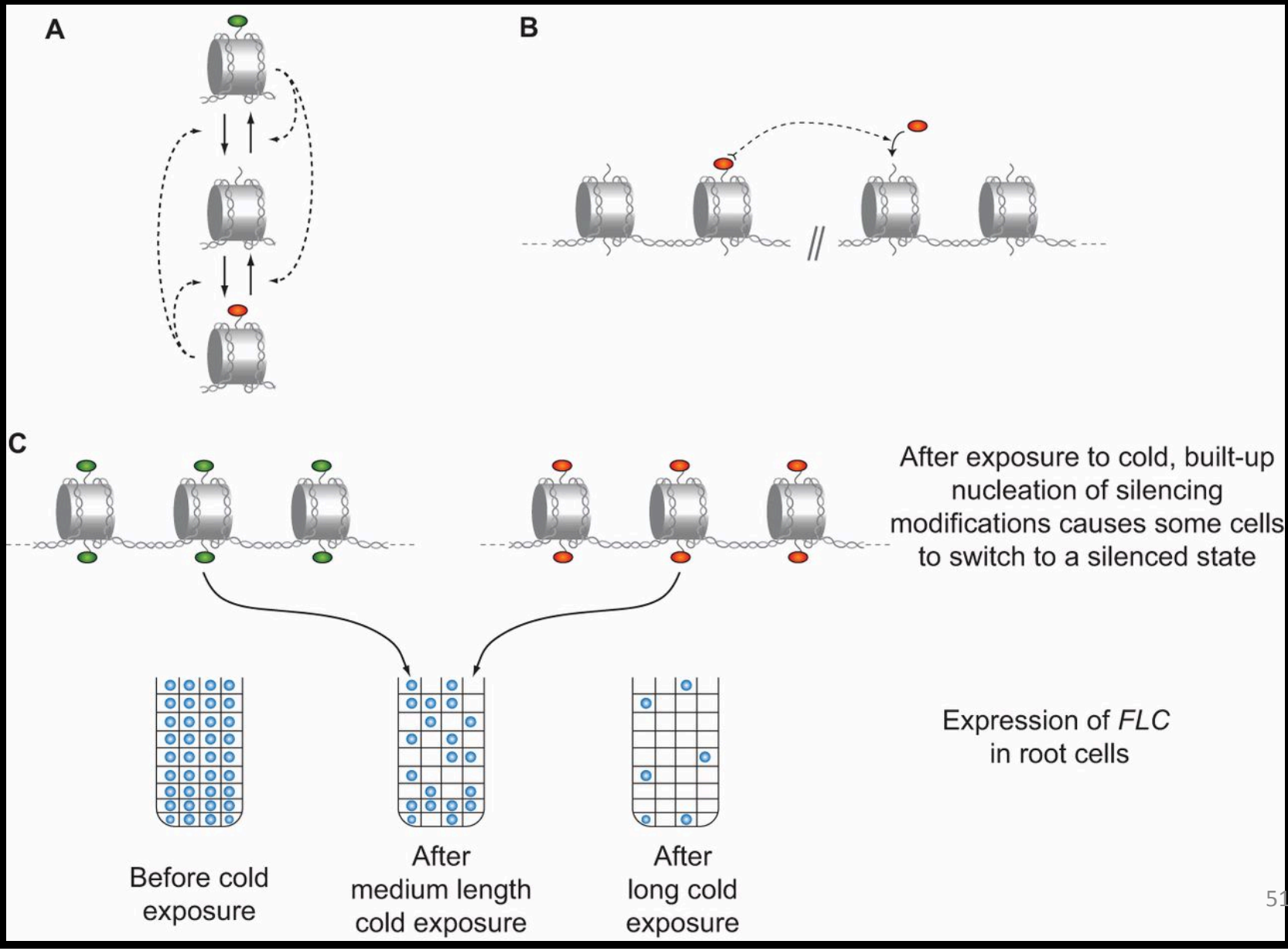
Key modelling principles and the quantitative nature of the vernalization response: repressing *FLC*



Key modelling principles and the quantitative nature of the vernalization response: repressing *FLC*

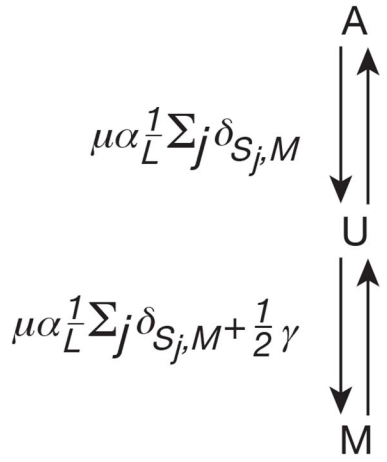


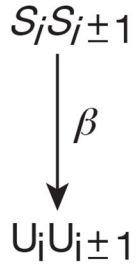
Key modelling principles and the quantitative nature of the vernalization response: repressing *FLC*



Formal description of the model

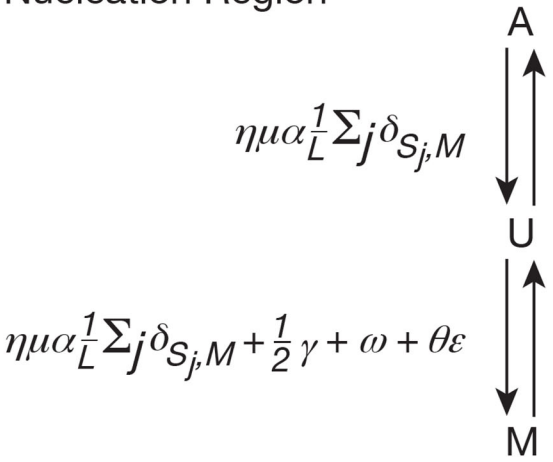
a Histone away from Nucleation Region

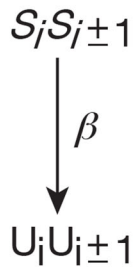




(± 1 represents other histone in same nucleosome)

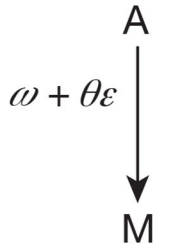
b Histone in Nucleation Region





c Competency to Nucleate

In cold, $\theta \rightarrow 1$, with probability $\frac{Ct^h}{K+t^h}$ per sweep (t is time in days)
In warm, $\theta \rightarrow 0$, with probability v per sweep

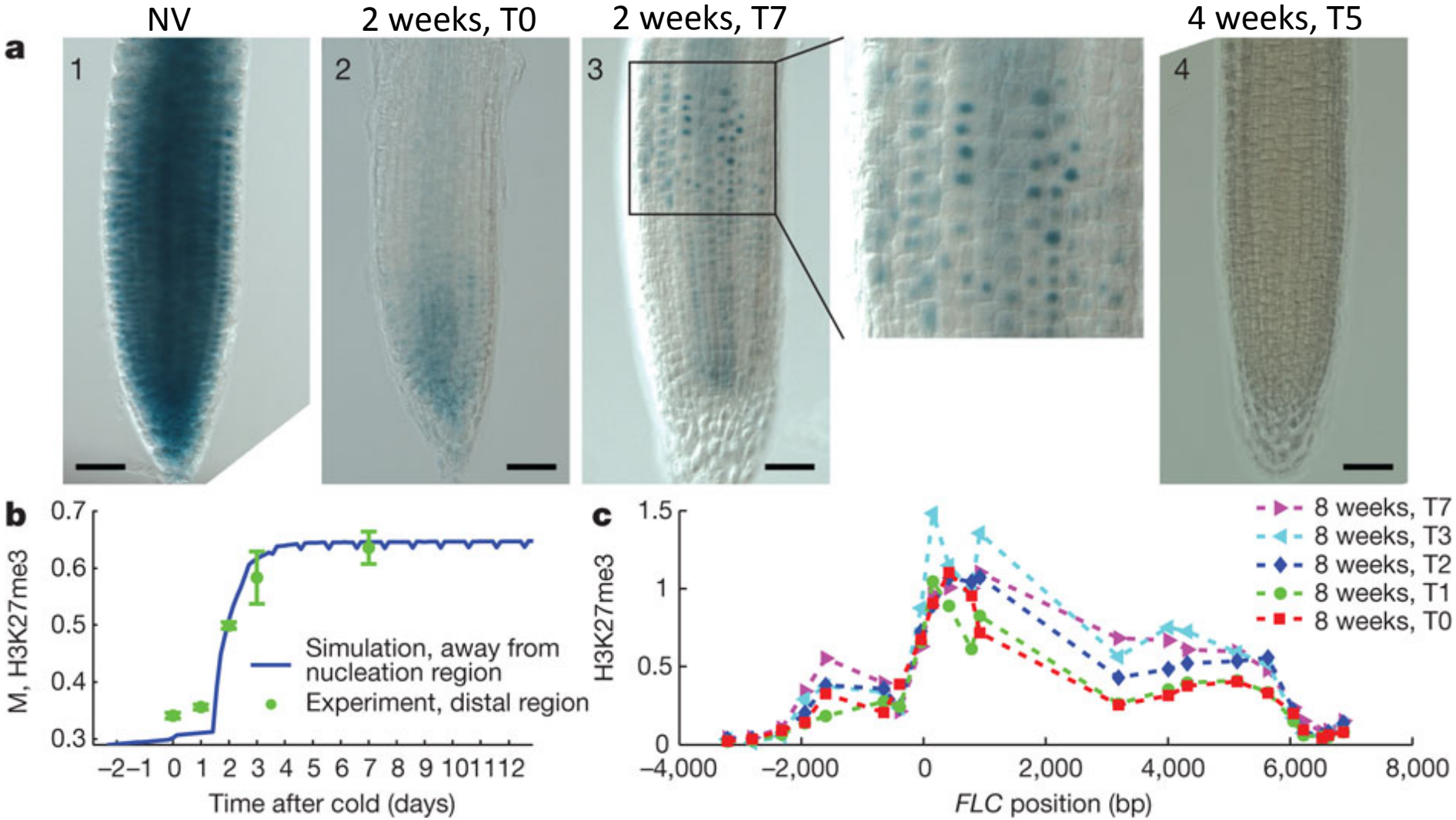


d DNA replication

Once per day (week) in warm (cold) $S_iS_{i+1} \rightarrow U_iU_{i+1}$ with probability 0.5 for i = 1, 3, 5,.....

Validating model predictions

Dark = *FLC* expression



Summary

- Epigenetics is pretty interesting (and a bit weird)
- It's early days yet, but epigenetics will increase in importance, as it affects everything
- There's a lot of data now, and much more coming
- There are plenty of opportunities for mathematical modelling and statistical analysis in conjunction with experiments, and
- Plenty of observational data on humans requiring careful statistical analysis

Thanks to many, especially

WEHI
Peter Hickey
Moshe Olshansky

Uni Melb
Stephen Turner
Brendan Russ

MCRI
Alicia Oshlack
Jovana Maksimovic

LaTrobe Uni
Emma Whitelaw
Harry Oey

Garvan Institute
Sue Clark

Uni Zurich
Mark Robinson

Johns Hopkins
Rafael Irizarry

Uni So Cal Hui Shen

Who won the 2012 Nobel Prize for Physiology or Medicine?

John B. Gurdon

Shinya Yamanaka

And for what work was the award given?

for the discovery that mature cells can
be reprogrammed to become pluripotent.

Three “old” papers by Sir John Gurdon

DNA demethylation is necessary for the epigenetic reprogramming of somatic cell nuclei

2004

Characterization of somatic cell nuclear reprogramming by oocytes in which a linker histone is required for pluripotency gene reactivation

2010

Histone variant macroH2A confers resistance to nuclear reprogramming

2011

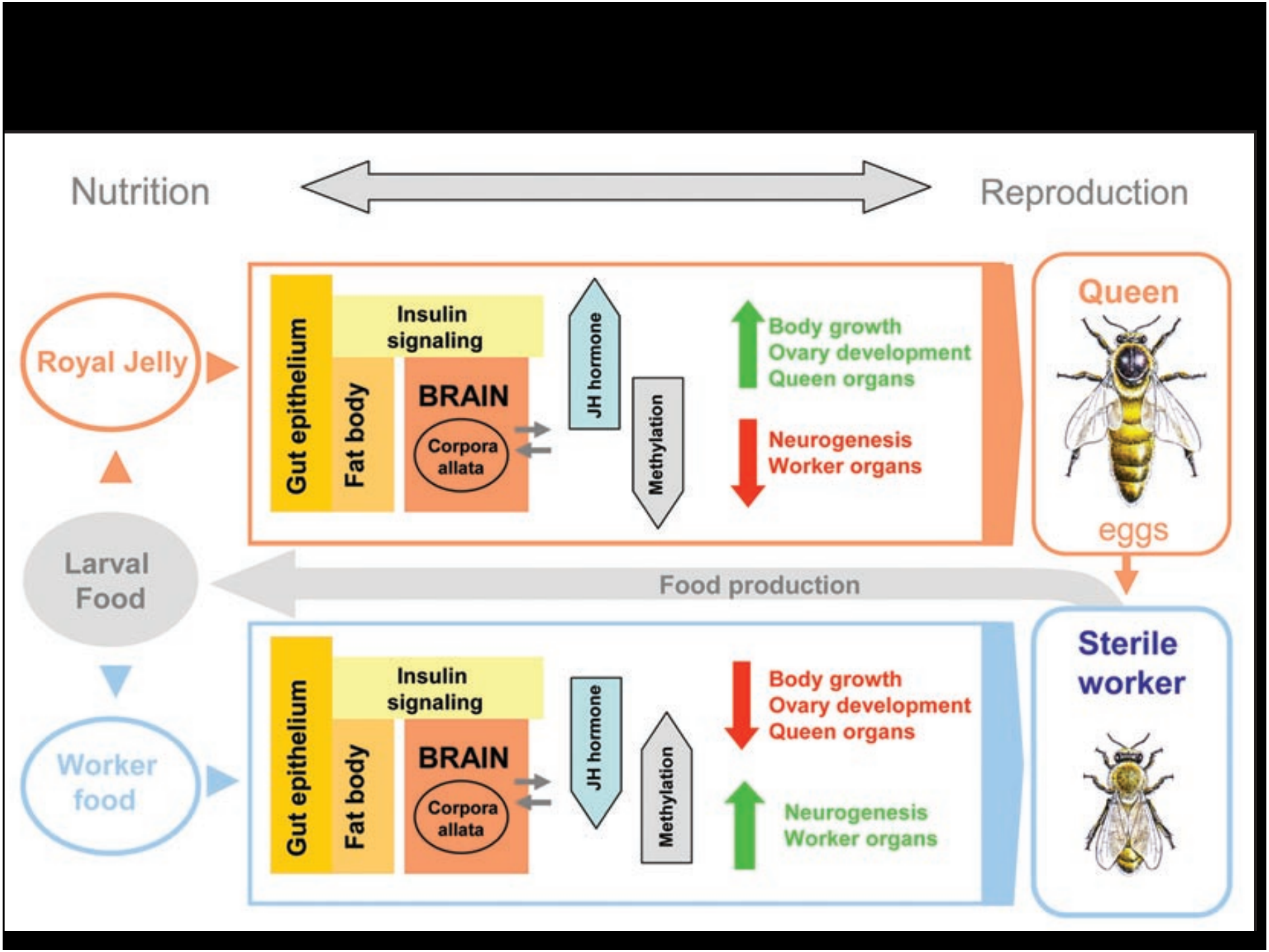
From the website of Shinya Yamanaka Institute for Integrated Cell-Material Sciences, Kyoto University

We are also working to change the **epigenetic status** in cancer cells using **reprogramming** technology, thereby making differences between genetic abnormality and **epigenetic status** in cancer cells. Through the analysis of the biological behaviors of these reprogrammed cancer cells, we seek the significance of epigenetic abnormality in carcinogenesis. Our goal is to find out the original epigenetic abnormality which causes the cancer through an analysis of epigenetic changes in the reprogrammed cancer cells and to develop a new "epigenetic cancer therapy" which resets the epigenetic state in cancer cells.

A hint of the "histone code"

Modification\Histone	H3K4	H3K27
mono-methylation	activation	activation
di-methylation		repression
tri-methylation	activation	repression
acetylation		activation

General effect of some histone modifications around genes.
Nothing is simple: some genes can have opposing signals, the above interpretation is not universal. Outcomes may depend on modifications we haven't measured or don't know.


Nutrition
Reproduction
Royal Jelly
Gut epithelium
Insulin signaling
Fat body
BRAIN
Corpora allata
JH hormone
Methylation
Body growth
Ovary development
Queen organs
Neurogenesis
Worker organs
Queen
eggs
Larval Food
Food production
Worker food
Gut epithelium
Insulin signaling
Fat body
BRAIN
Corpora allata
JH hormone
Methylation
Body growth
Ovary development
Queen organs
Neurogenesis
Worker organs
Sterile worker

Four eminent Australian epigeneticists

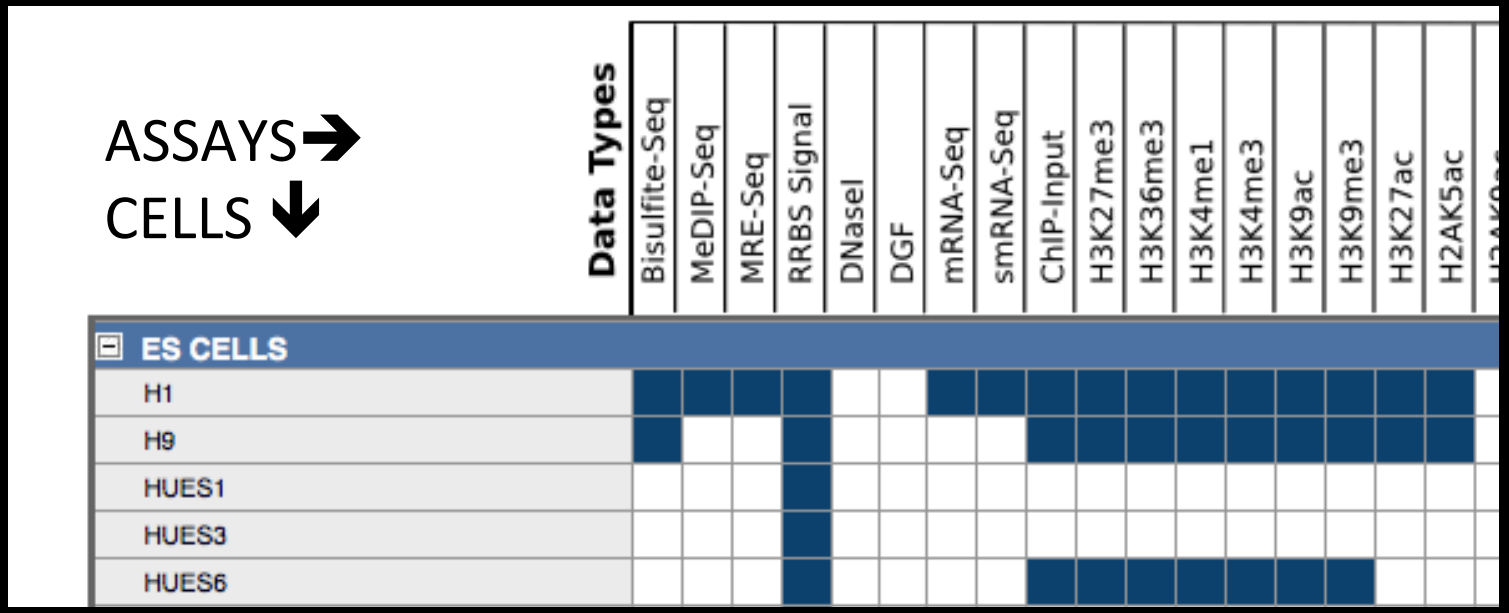
Marianne Frommer
CpG islands
bisulphite sequencing

Emma Whitelaw
Epigenetics in mammals

Sue Clark
Cancer epigenetics
Long-range epigenetic silencing

Jean Finnegan
Flowering and epigenetics

ROADMAP data



40 columns

261 rows

This will all be sequence data. Most other groups will do likewise, with the exception of methylation microarrays.

What are some statistical challenges?

There are many, ranging from low-level analysis of assays to deciphering the histone code.

Some statistical issues with DNA methylation

- $\beta = M/(M+U)$ or $\gamma = \log M/U$?
- QC, batch removal, normalization of methylation arrays, see also next slide
- Identifying differential methylation at a probe or CpG (bearing in mind cellular, e.g. tumor heterogeneity)
- Identifying regions of differential methylation along a single methylome, between two single methylomes, or between two sets of methylome data (see next + 1)
- Association between methylation and other variables, such as disease, mutations, gender, age…

Where is this sort of study heading?

Cell

Comparative Epigenomic Analysis of Murine and Human Adipogenesis

Tarjei S. Mikkelsen,[1,4] Zhao Xu,[1,2,4] Xiaolan Zhang,[1] Li Wang,[1] Jeffrey M. Gimble,[3] Eric S. Lander,[1] and Evan D. Rosen[1,2,*]
[1]Broad Institute, 7 Cambridge Center, Cambridge, MA 02142, USA
[2]Division of Endocrinology and Metabolism, Beth Israel Deaconess Medical Center, 330 Brookline Avenue, Boston, MA 02115, USA
[3]Stem Cell Biology Laboratory, Pennington Biomedical Research Center, Louisiana University System, 6400 Perkins Road, Baton Rouge, LA 70808, USA
[4]These authors contributed equally to this work
*Correspondence: erosen@bidmc.harvard.edu
DOI 10.1016/j.cell.2010.09.006

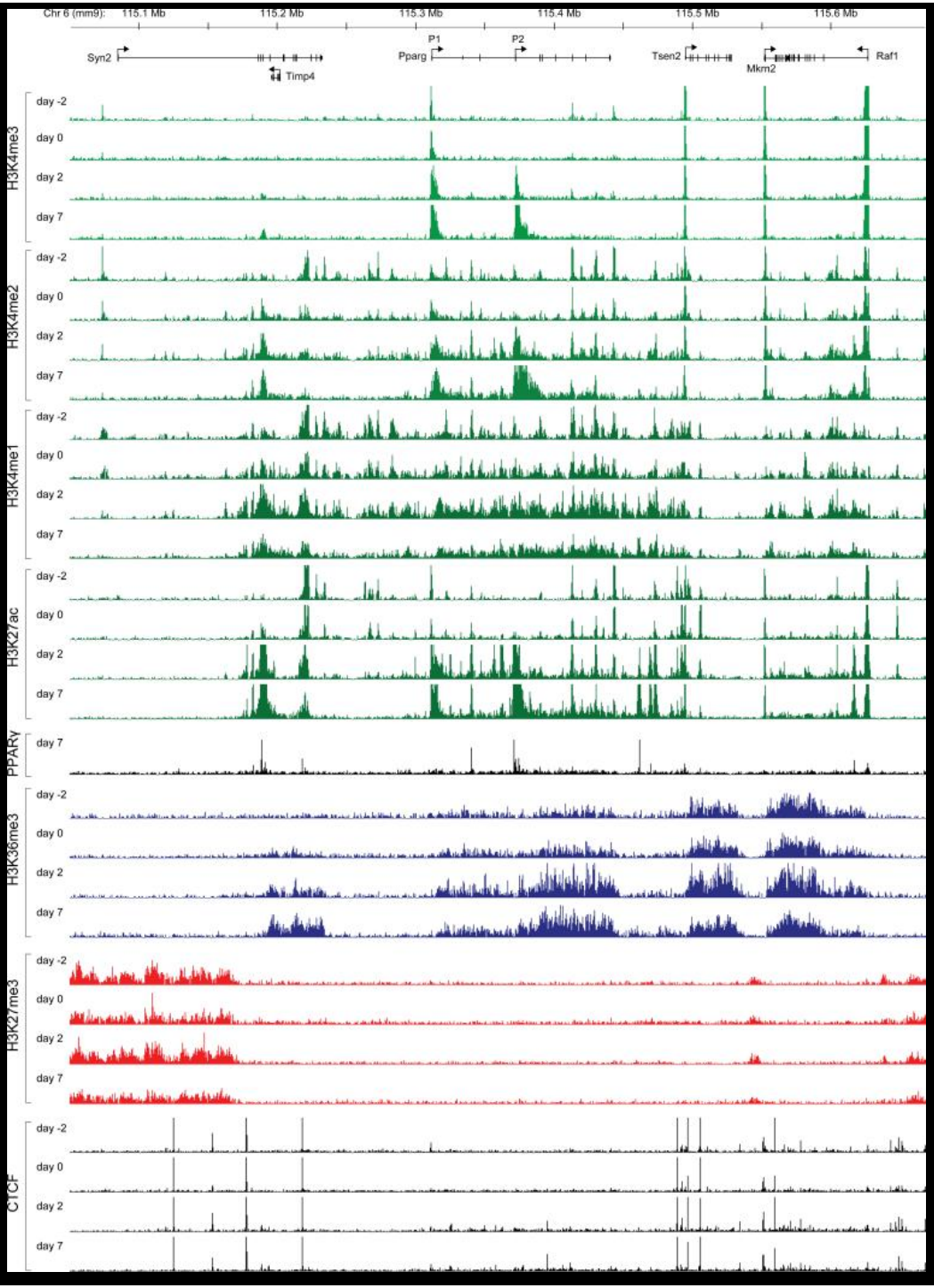


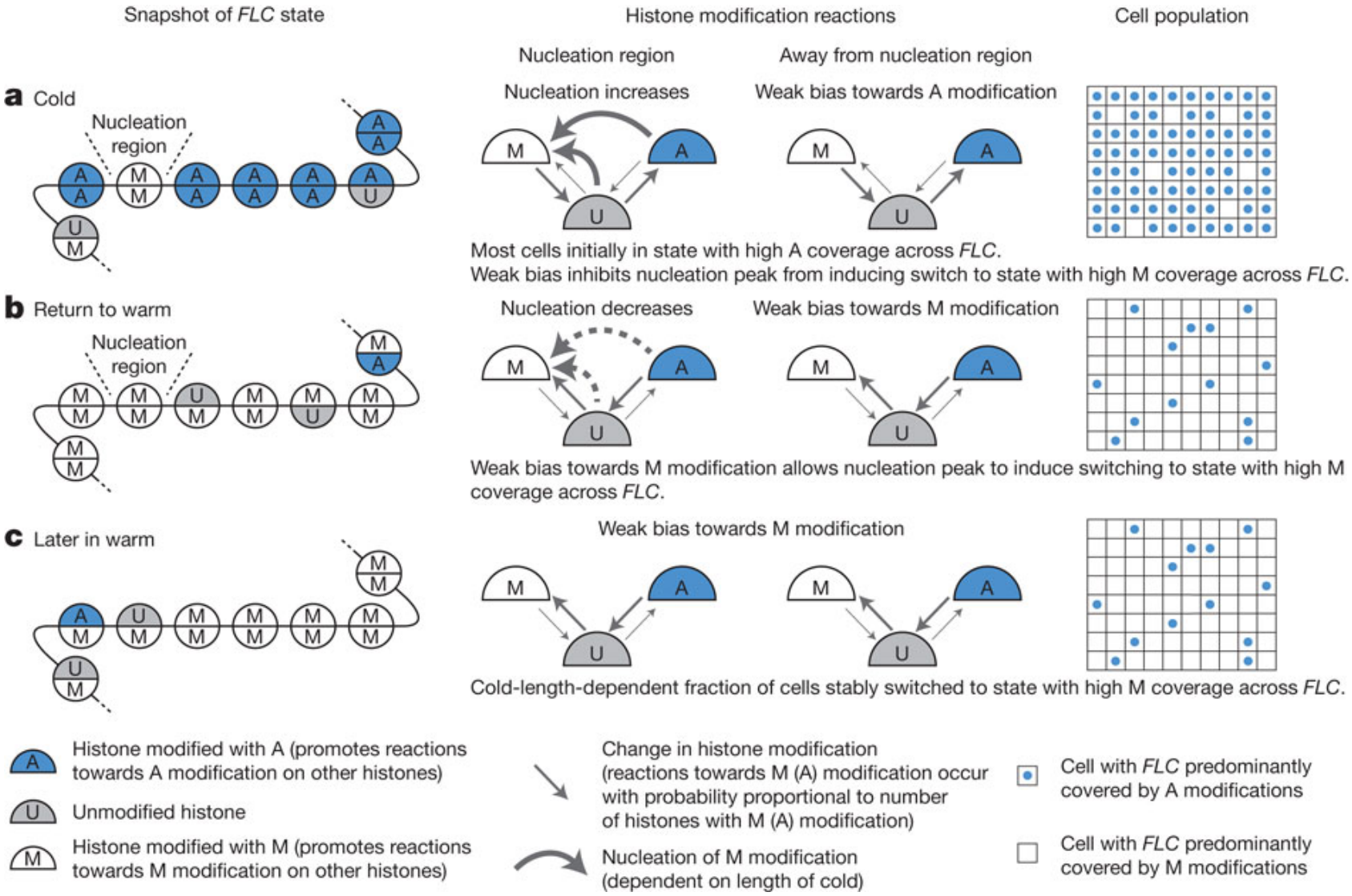
Figure 1 from Mikkelsen *et al*, **Cell** 2010:
6 histone marks,
1 transcription factor,
1 insulator
3 time points

Challenges in figuring this out

- The huge number of combinations of different histone modifications and chromatin remodellers
- The confounding effect of sequence dependent gene regulation
- The necessarily limited amount of data
- The biological truth that everything interacts with everything, to some extent, here methylation, histone modifications, chromatin remodelling and small RNAs (and probably more to come).


Worth1000.com

Schematic outline of mathematical model for *FLC* silencing.



A Angel *et al. Nature* **000**, 1-4 (2011) doi:10.1038/nature10241

nature

What are the data?

An explosion of data is emerging that will make the gene expression microarray wave over the last 15 years look like a little splash.

Many assays, but few underlying platforms

- DNA Methylation
- Histone modification (many: H3 alone has over 80, many more combinations) microRNA abundance
- DNase I hypersensitivity to measure nucleosome occupancy

ChIP-seq =

Chromatin ImmunoPrecipitation + sequencing

- mRNA abundance (= gene expression levels)
- Transcription factor binding sites (+ other DNABP sites)

Platforms: PCR, gels, microarrays, DNA sequencing

H3K4me3 around a gene in CD8+ T cells

Pile-up plots from a ChIP-seq assay

Ornithine decarboxylase antizyme 1

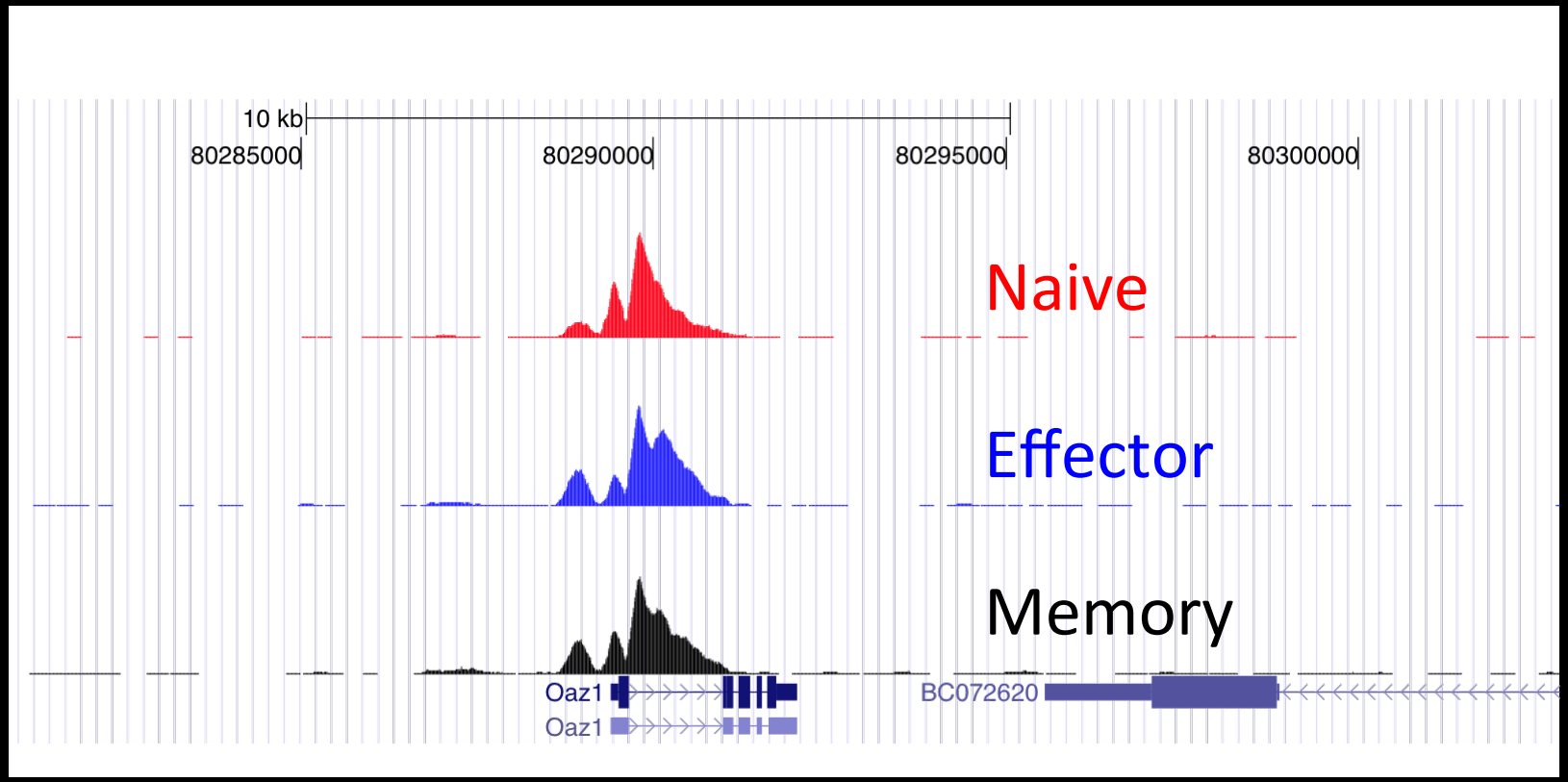
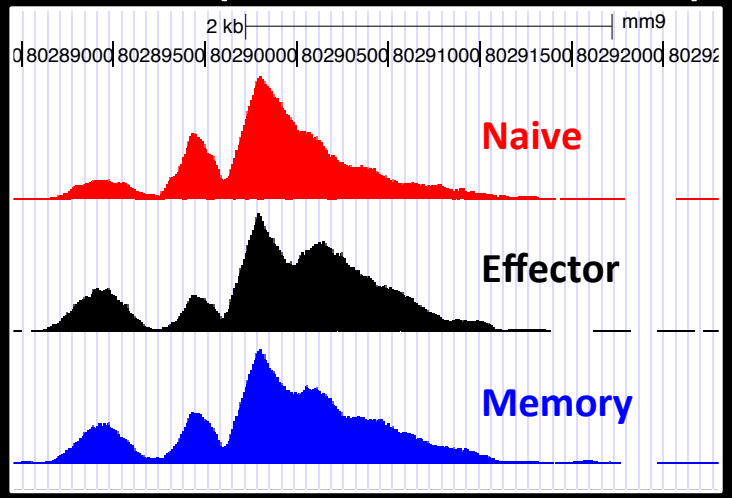


Figure courtesy of B Russ

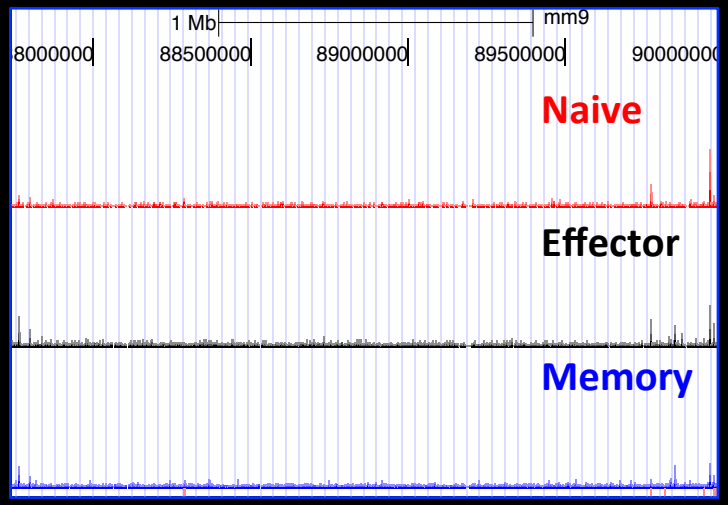
Epigenetic modifications within active and repressed loci during CD8+ T cell differentiation

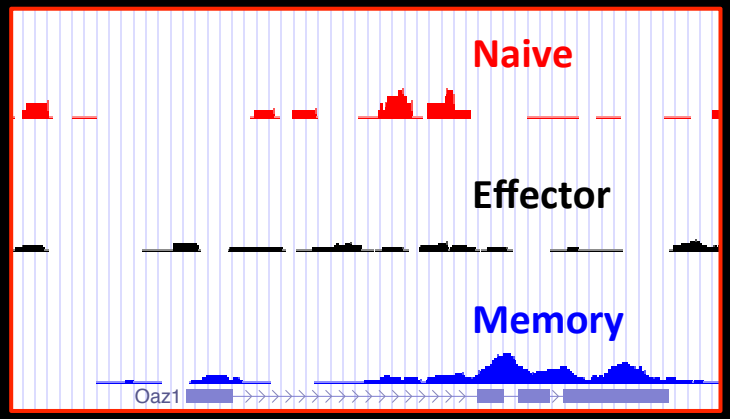
OAZ1 (active in T cells)



H3K4me3

Keratin (repressed in T cells)





H3K27me3