From theory to data ready for analysis

EWAS

JON BOHLIN - BIOINFORMATICS - NIPH (FHI)

Outline

- Genomics
- DNA structure and chromatin
- Epigenetics
- Epigenetic data
- Quality control of epigenetic data

Genome

- Enough information to reproduce the organism
- «A chicken is just an egg’s way of making a new egg»
- Genome consist of double stranded DNA = {A,G,C,T} in prokaryotes and eukaryotes but not necessarily in virus
- A=T, G=C across strands, can connect to anything on same strand
- AT/GC same across strands
- Genes are made up of triplets (codons) that code for proteins

Genomics

- EBI: Genomics is the study of whole genomes of organisms, and incorporates elements from genetics. Genomics uses a combination of recombinant DNA, DNA sequencing methods, and bioinformatics to sequence, assemble, and analyse the structure and function of genomes

Genomics

* Human genome 2*3 Gbp - approx similar for mammals

* ~20k genes, 1-2% of genome

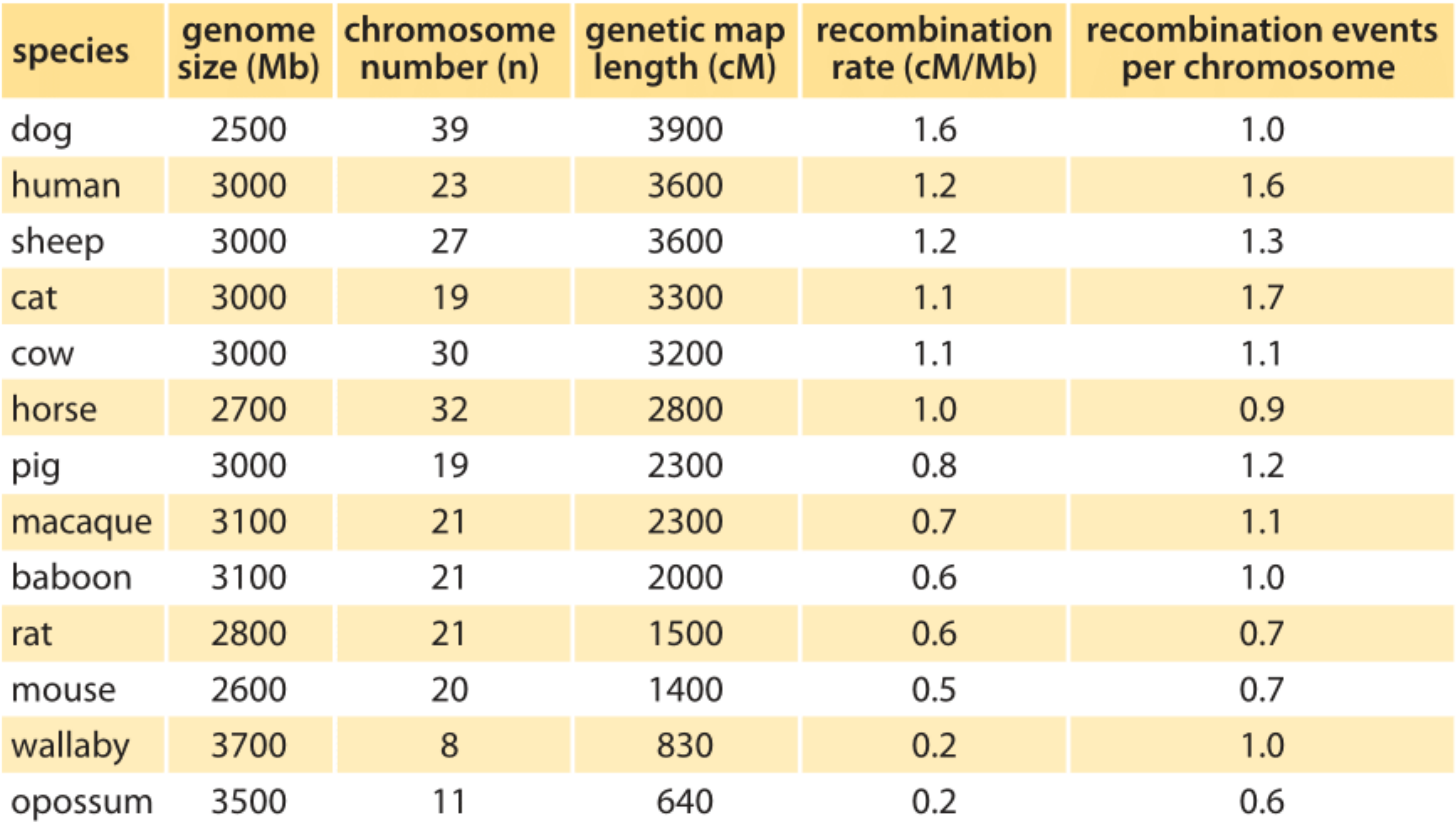
* Sequenced in 2000 - 1 billion $, today 99$

* bacteria 3 Mbp (on average)

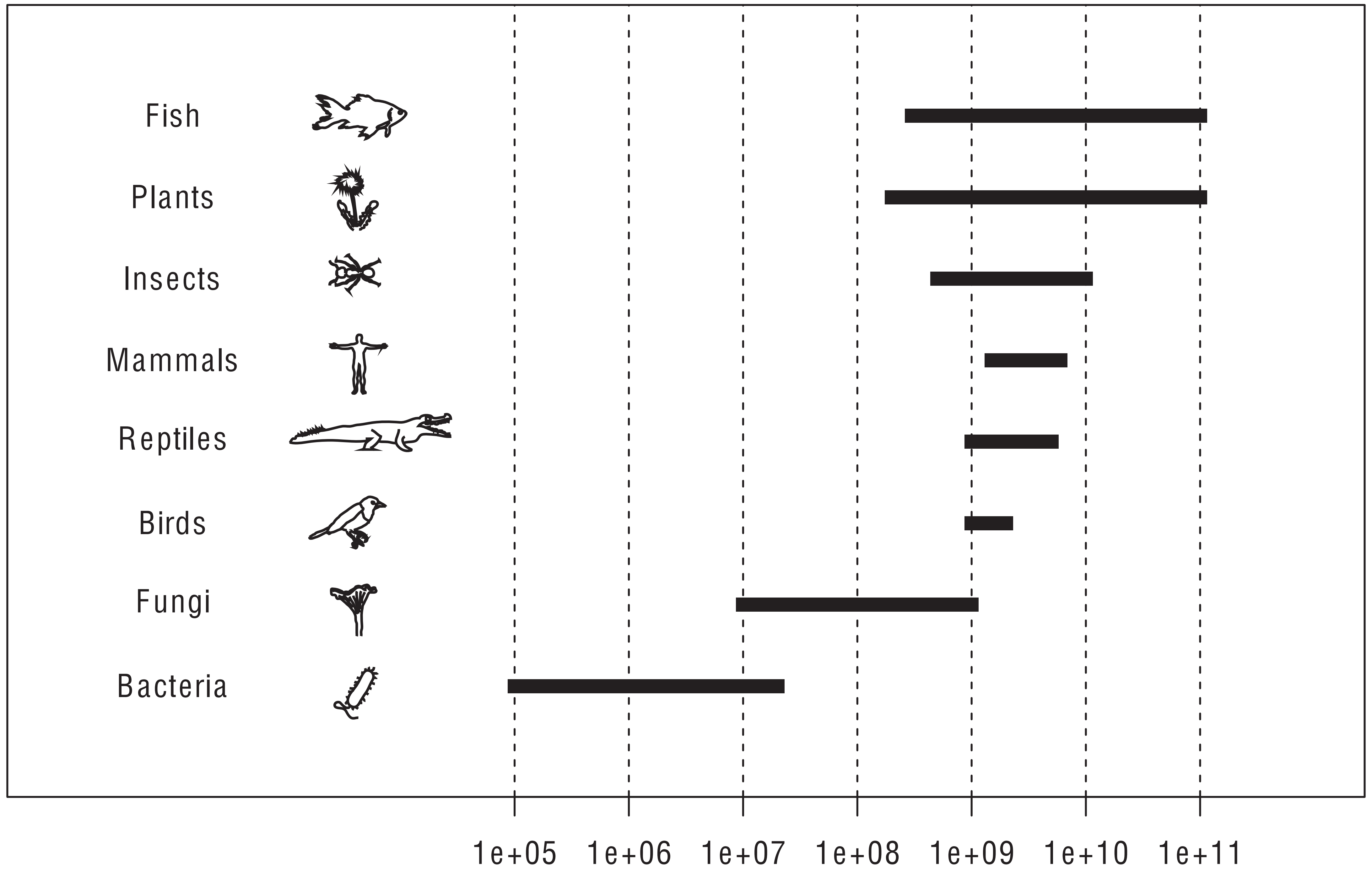
* ~3k genes (on average), 90-95% of genome

* 99% “junk”?

The C-value paradox



species	genome size (Mb)	chromosome number (n)	genetic map length (cM)	recombination rate (cM/Mb)	recombination events per chromosome
dog	2500	39	3900	1.6	1.0
human	3000	23	3600	1.2	1.6
sheep	3000	27	3600	1.2	1.3
cat	3000	19	3300	1.1	1.7
cow	3000	30	3200	1.1	1.1
horse	2700	32	2800	1.0	0.9
pig	3000	19	2300	0.8	1.2
macaque	3100	21	2300	0.7	1.1
baboon	3100	21	2000	0.6	1.0
rat	2800	21	1500	0.6	0.7
mouse	2600	20	1400	0.5	0.7
wallaby	3700	8	830	0.2	1.0
opossum	3500	11	640	0.2	0.6



Mutations and “junk”-DNA

* Approx 37 trillion cells in an adult human body (Bianconi, Ann Hum Biol 2013)

* Cells divide differently, some often (skin) others seldom (nerve/brain)

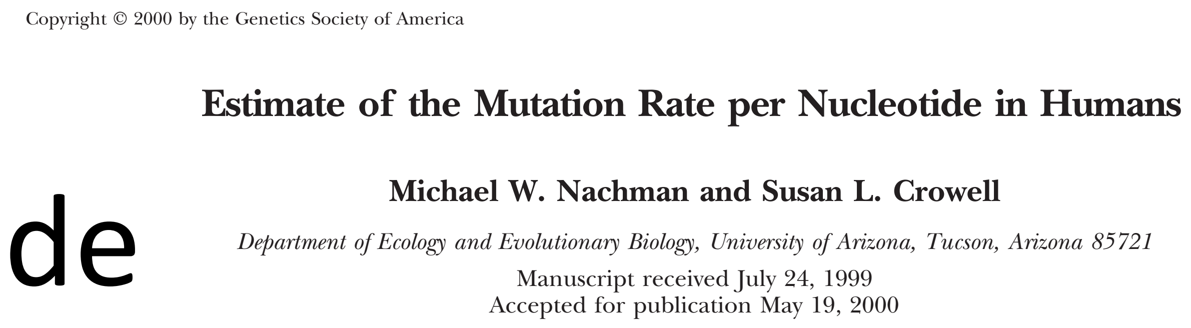
* Approx 2 trillion cell division every day

* DNA mutations w/repair ~ 1 pr 2.5x10^-8 nucleotide

* 150 mutations in every divided cell

* 300 trillion genomic mutations every single day (50000 diploid human genomes!)

* Imagine 90% coding genome in a multicellular organism (but why do protists have such large genomes??)



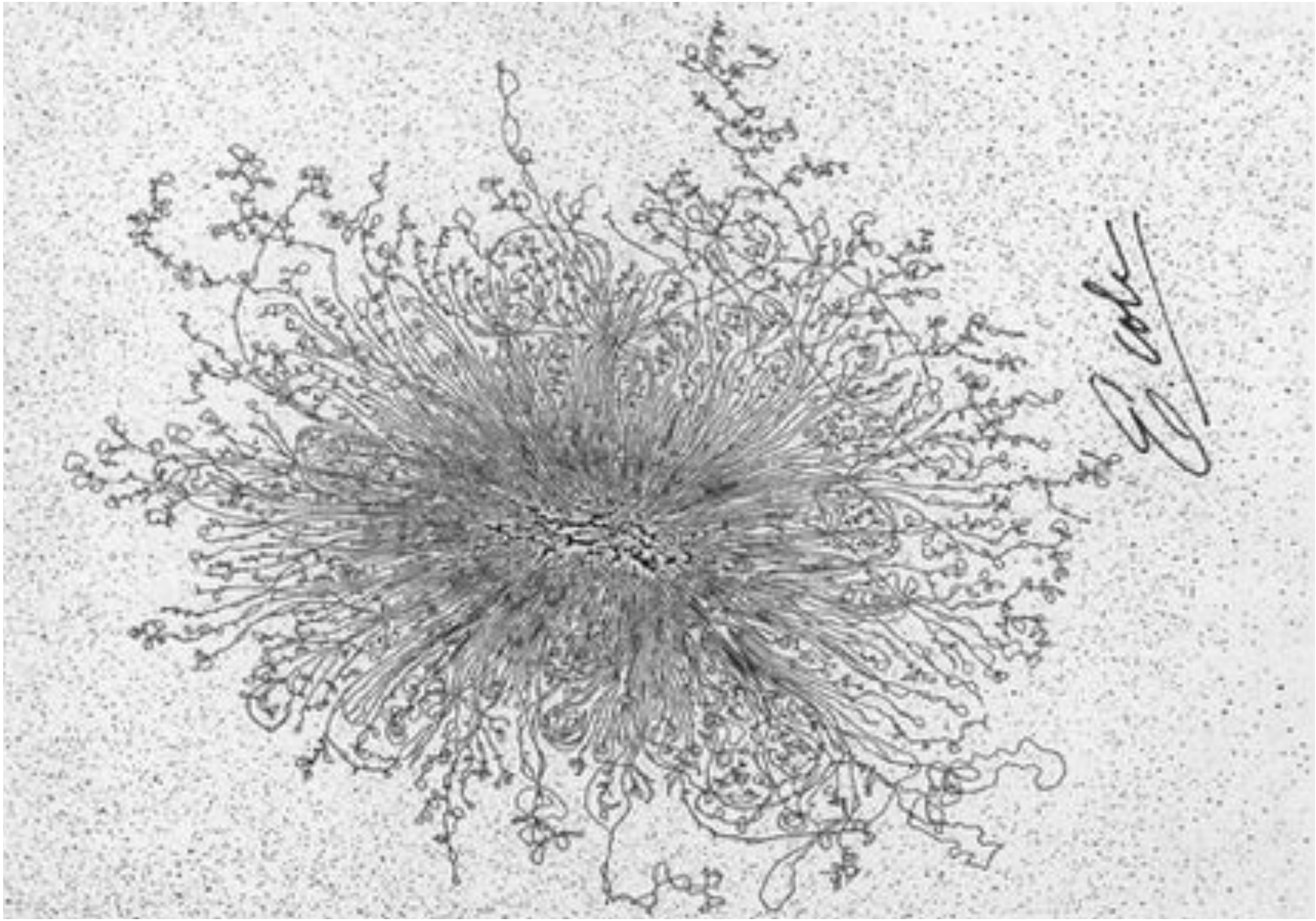
Copyright © 2000 by the Genetics Society of America

Estimate of the Mutation Rate per Nucleotide in Humans

Michael W. Nachman and Susan L. Crowell

Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85721

Manuscript received July 24, 1999
Accepted for publication May 19, 2000



DNA structure and chromatin

DNA Structure: A-, B- and Z-DNA Helix Families

David W Ussery, *Danish Technical University, Lyngby, Denmark*

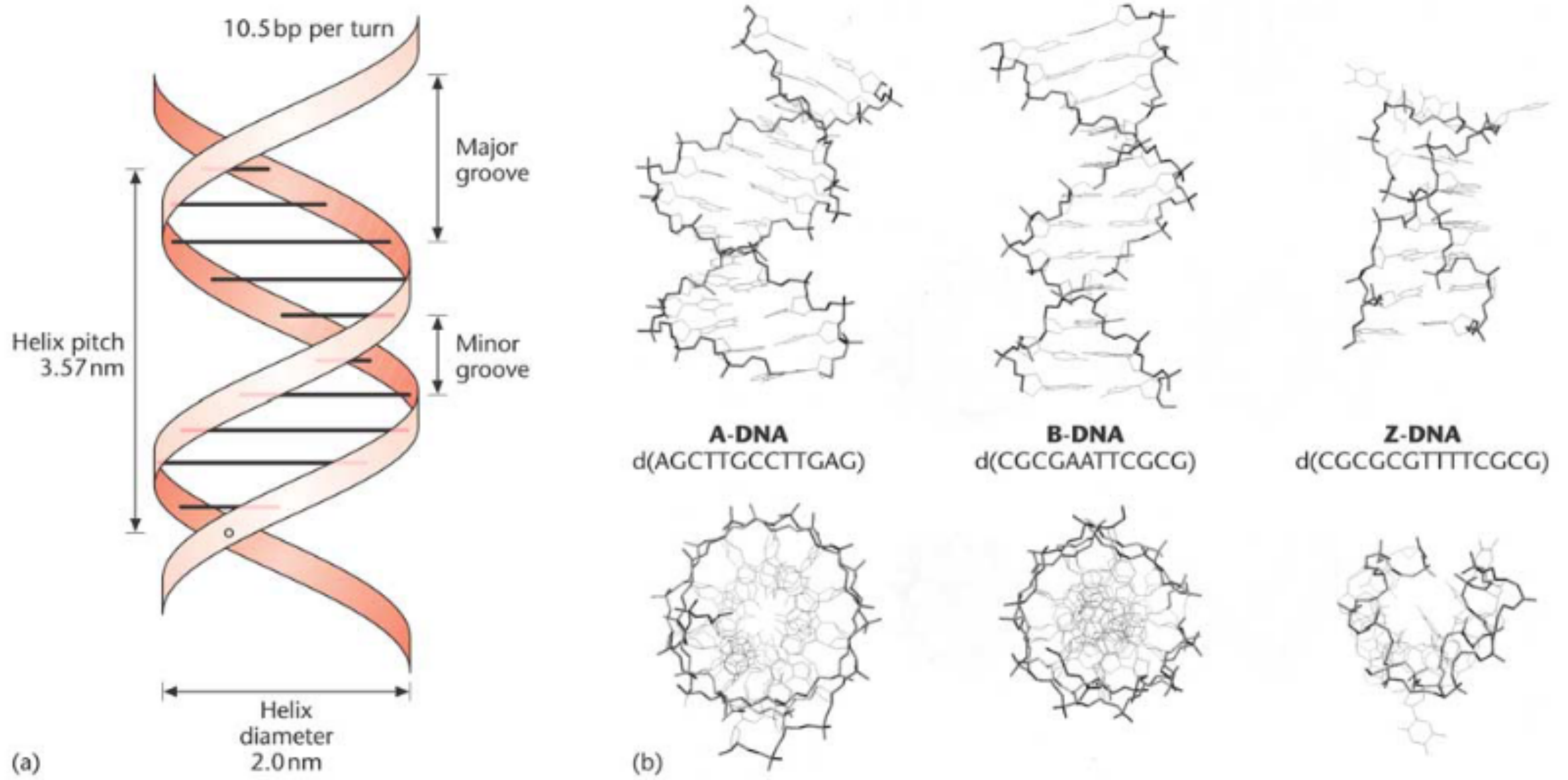
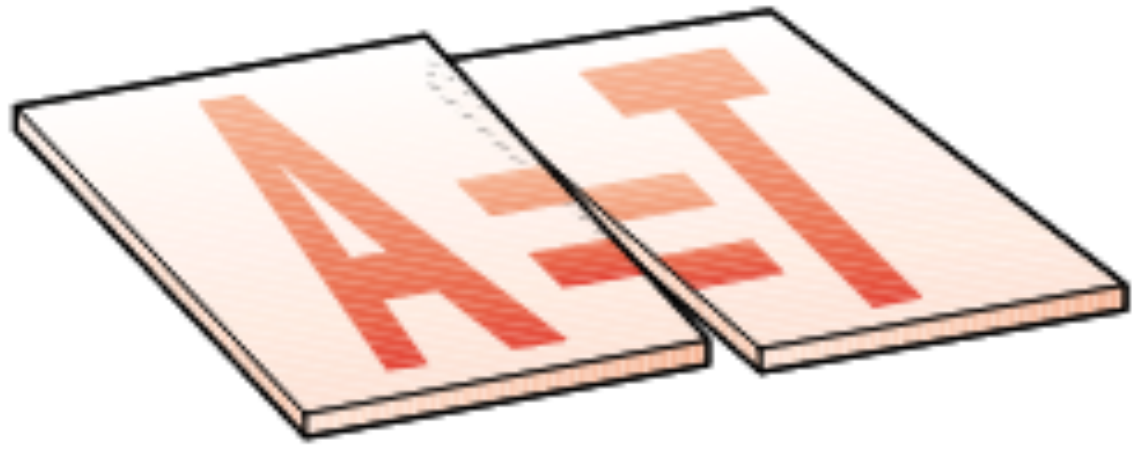
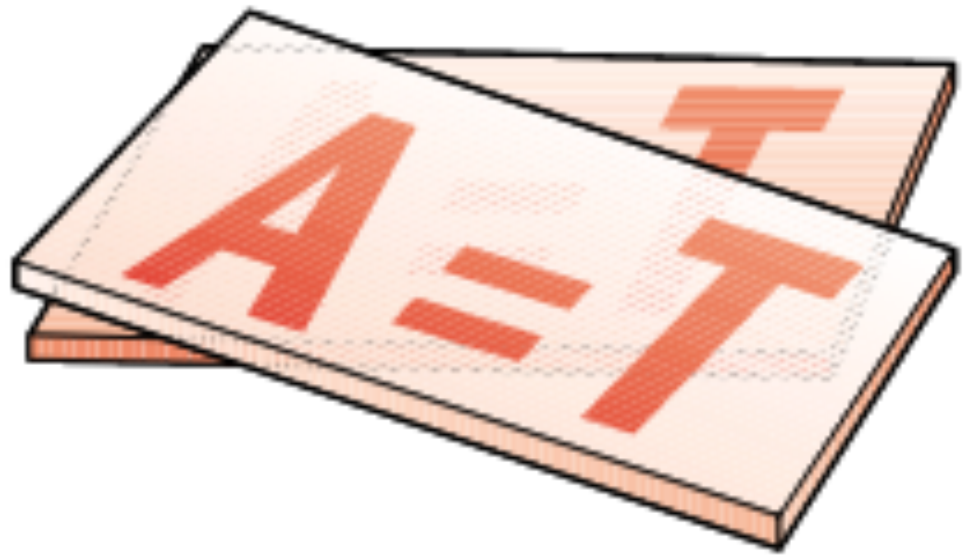


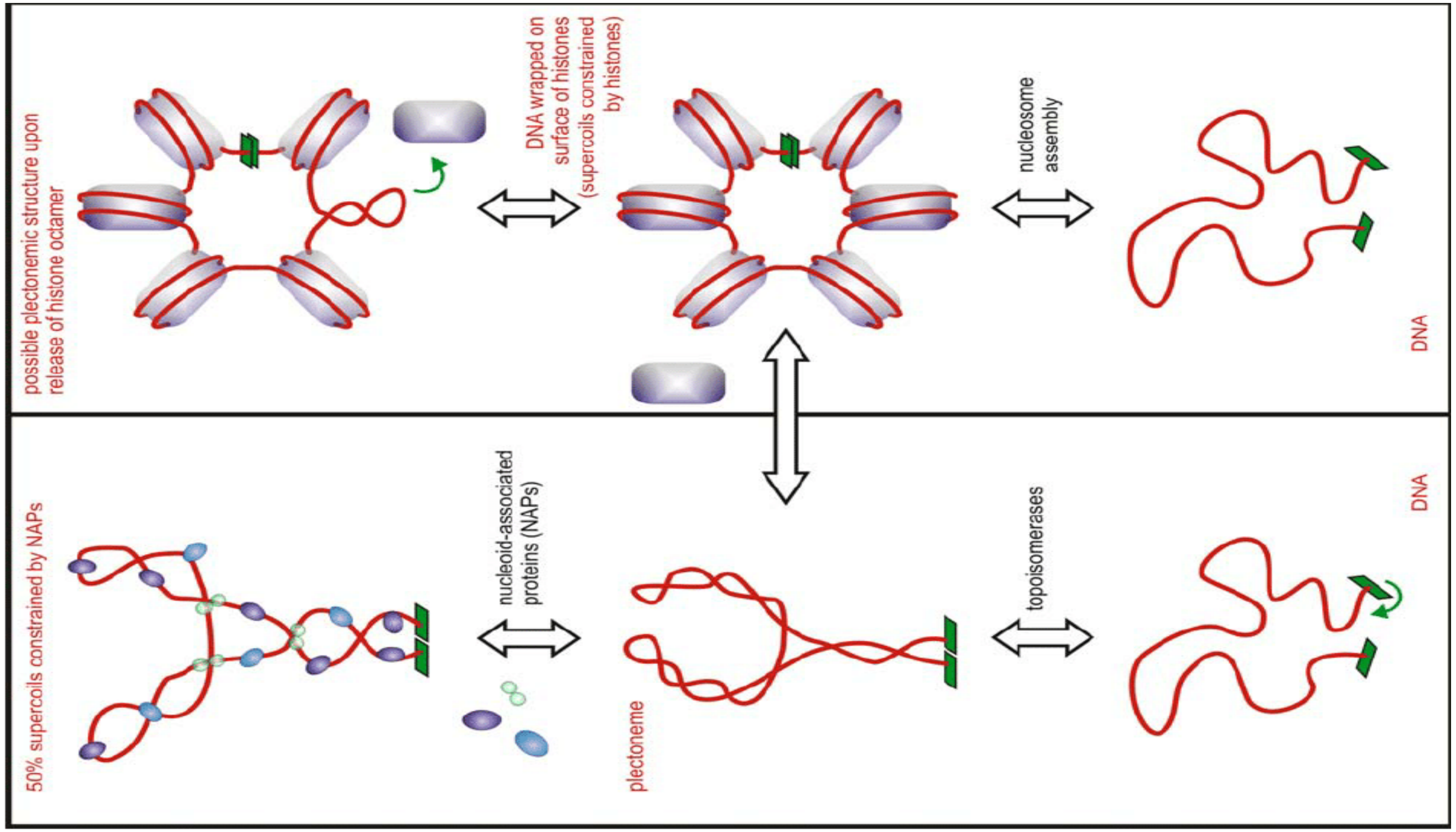
Figure 1 Different views of the DNA helix. (a) The structure of B-DNA as proposed by Watson and Crick in 1953, based on fibre diffraction studies. Modified from Sinden *et al*. (1998). (b) A-, B-and Z-DNA, as seen from the side of the helix (above), and looking down the helix axis (below). The structures were drawn from the crystal structures, using the Cn3D programme, available from the NCBI home page.

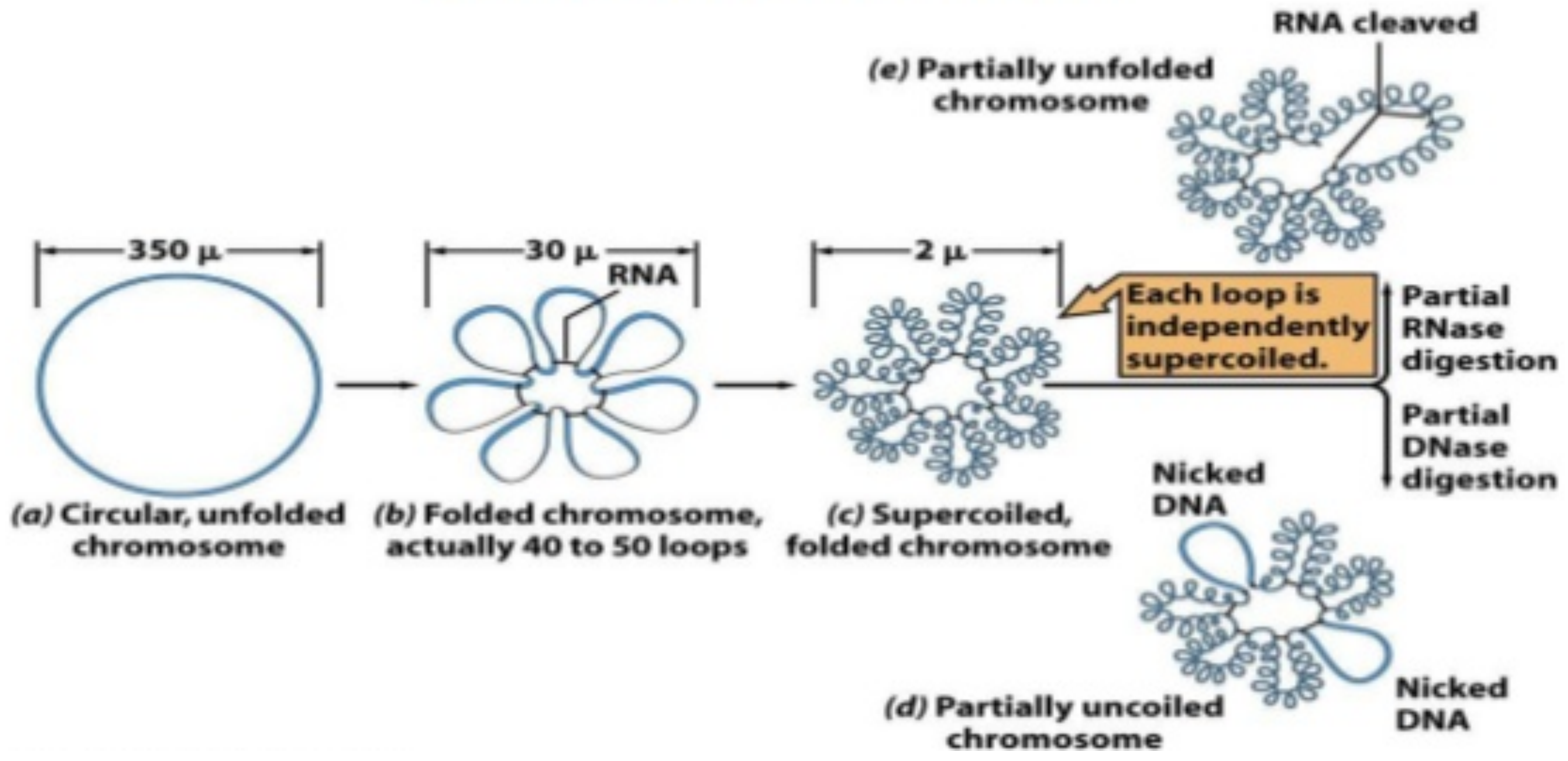


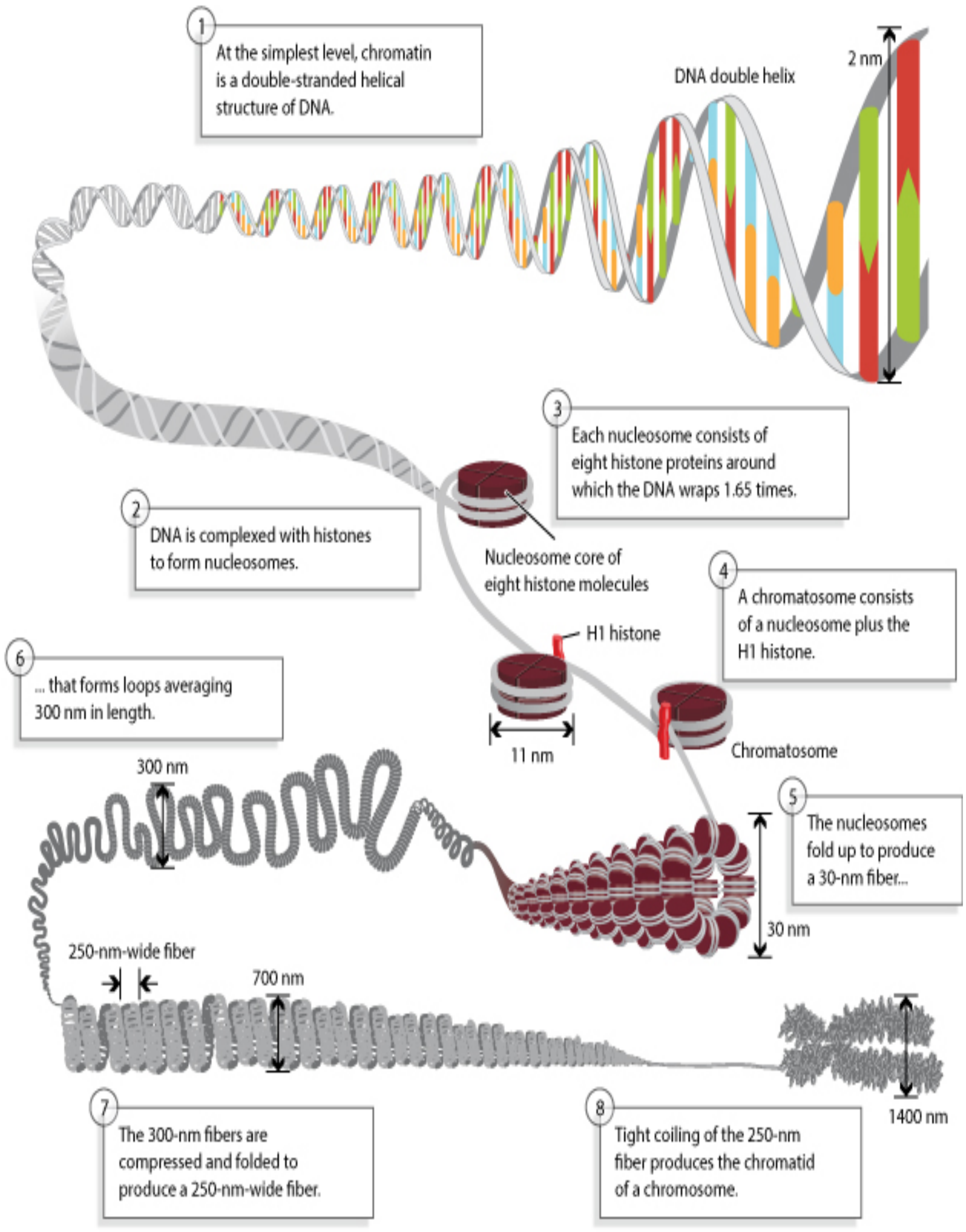
Propeller twist



Twist


possible plectonemic structure upon release of histone octamer
DNA wrapped on surface of histones (supercoils constrained by histones)
nucleosome assembly
DNA
50% supercoils constrained by NAPs
nucleoid-associated proteins (NAPs)
plectoneme
topoisomerases
DNA


RNA cleaved
(e) Partially unfolded chromosome
350 μ
30 μ
RNA
2 μ
Each loop is independently supercoiled.
Partial RNase digestion
Partial DNase digestion
Nicked DNA
(a) Circular, unfolded chromosome
(b) Folded chromosome, actually 40 to 50 loops
(c) Supercoiled, folded chromosome
(d) Partially uncoiled chromosome
Nicked DNA


1 At the simplest level, chromatin is a double-stranded helical structure of DNA.

2 DNA is complexed with histones to form nucleosomes.

3 Each nucleosome consists of eight histone proteins around which the DNA wraps 1.65 times.

4 A chromatosome consists of a nucleosome plus the H1 histone.

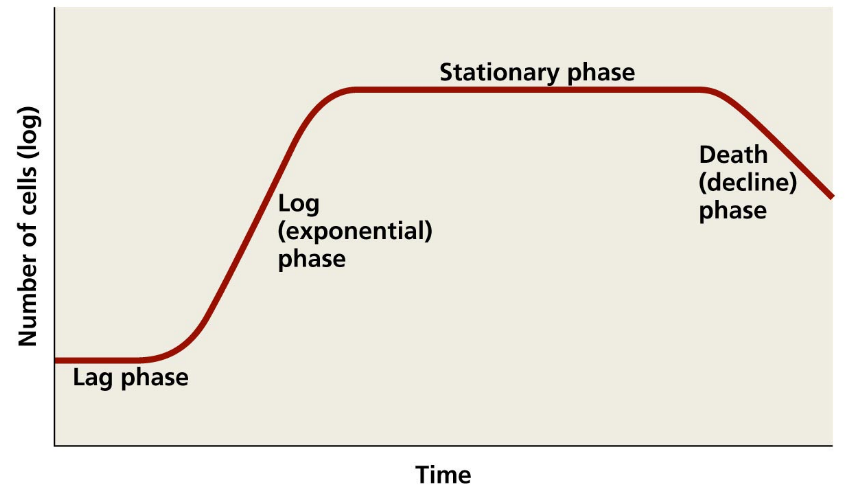
5 The nucleosomes fold up to produce a 30-nm fiber...

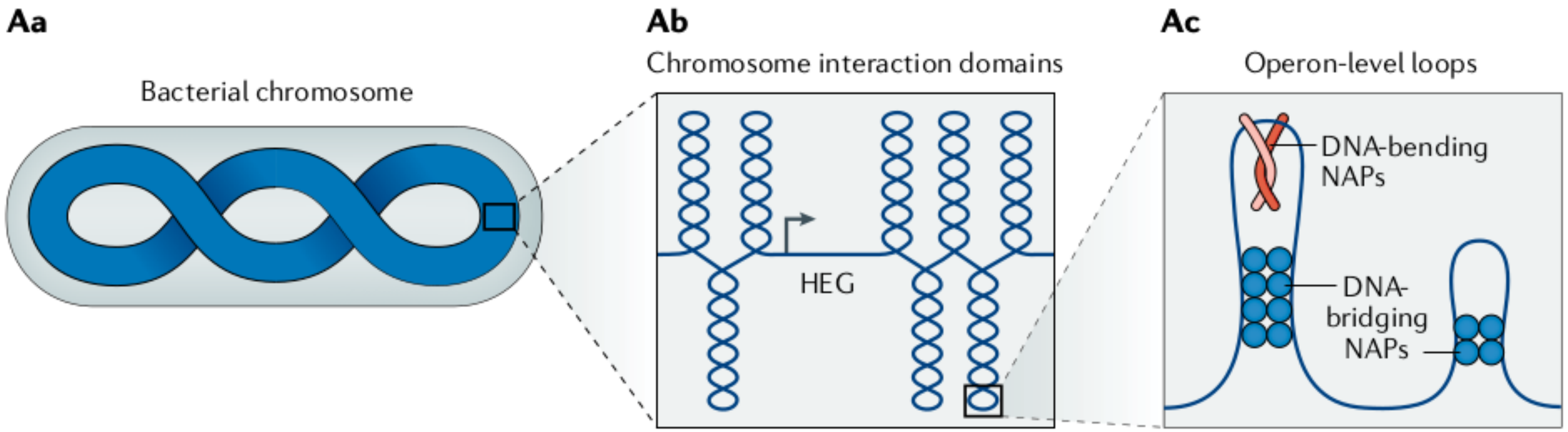
6 ... that forms loops averaging 300 nm in length.

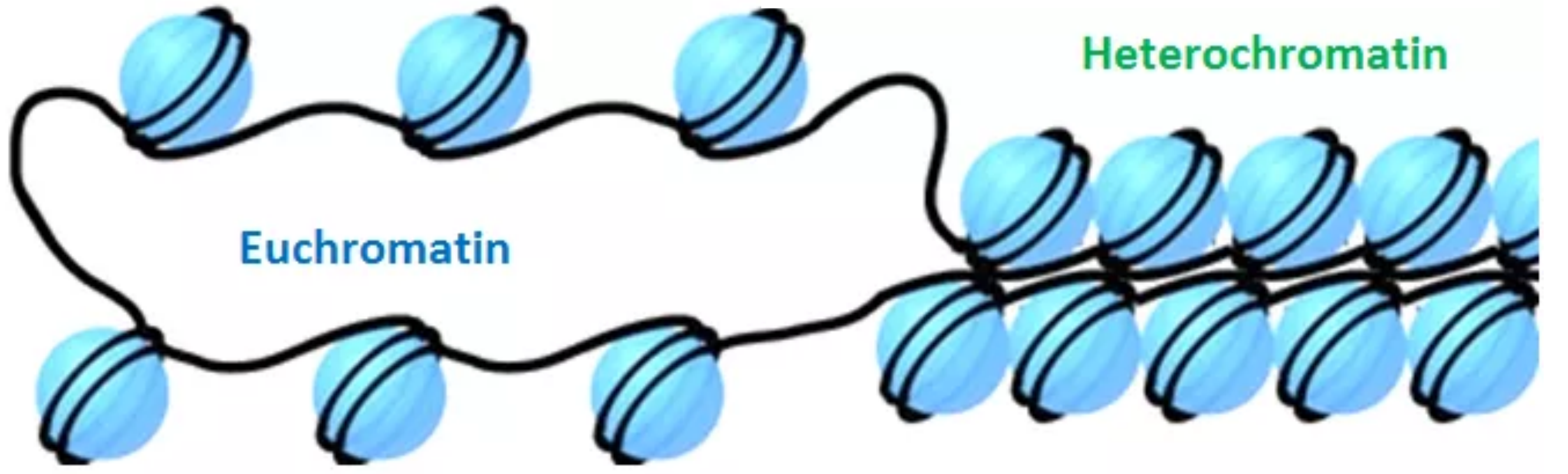
7 The 300-nm fibers are compressed and folded to produce a 250-nm-wide fiber.

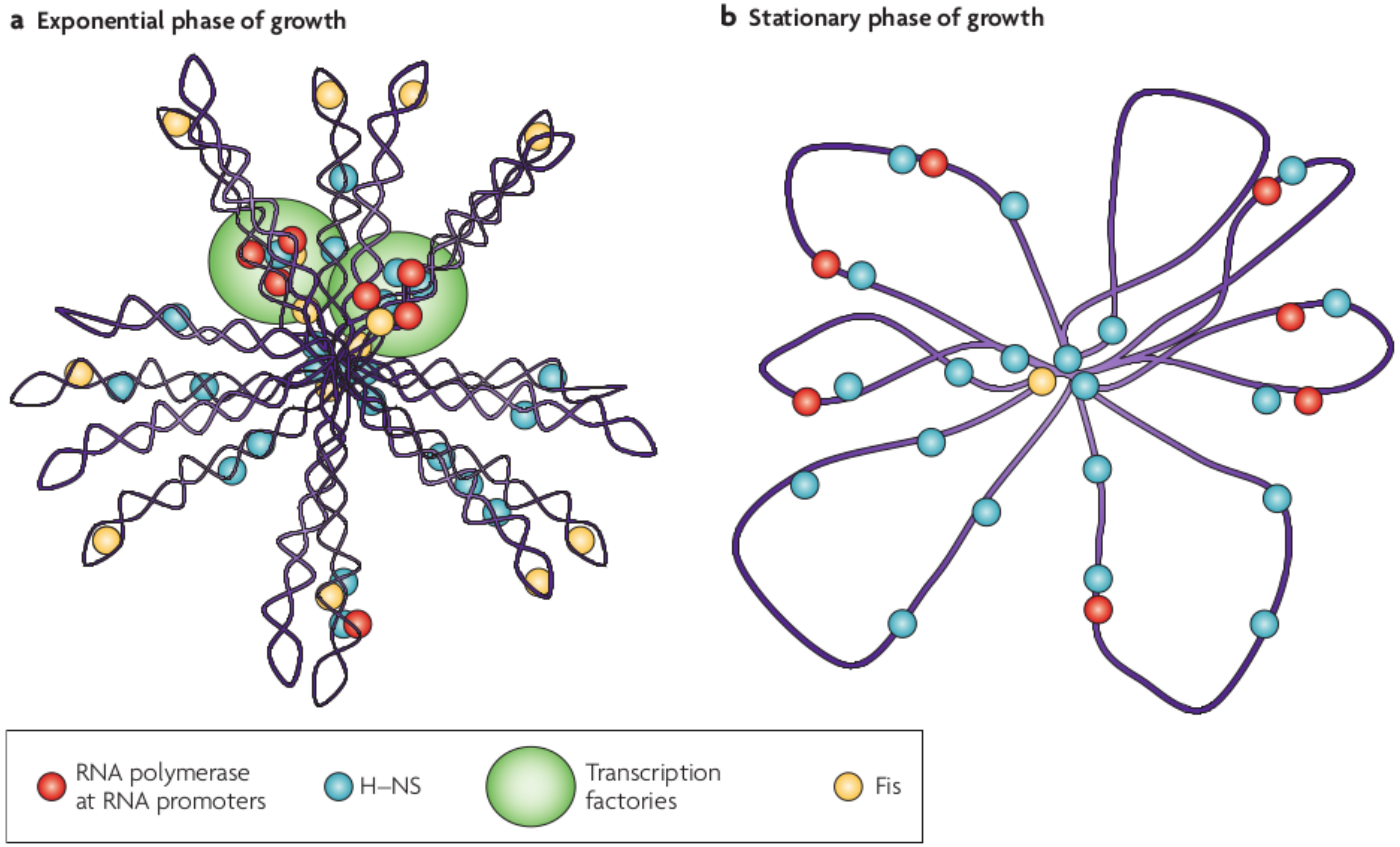
8 Tight coiling of the 250-nm fiber produces the chromatid of a chromosome.

DNA structure and transcription

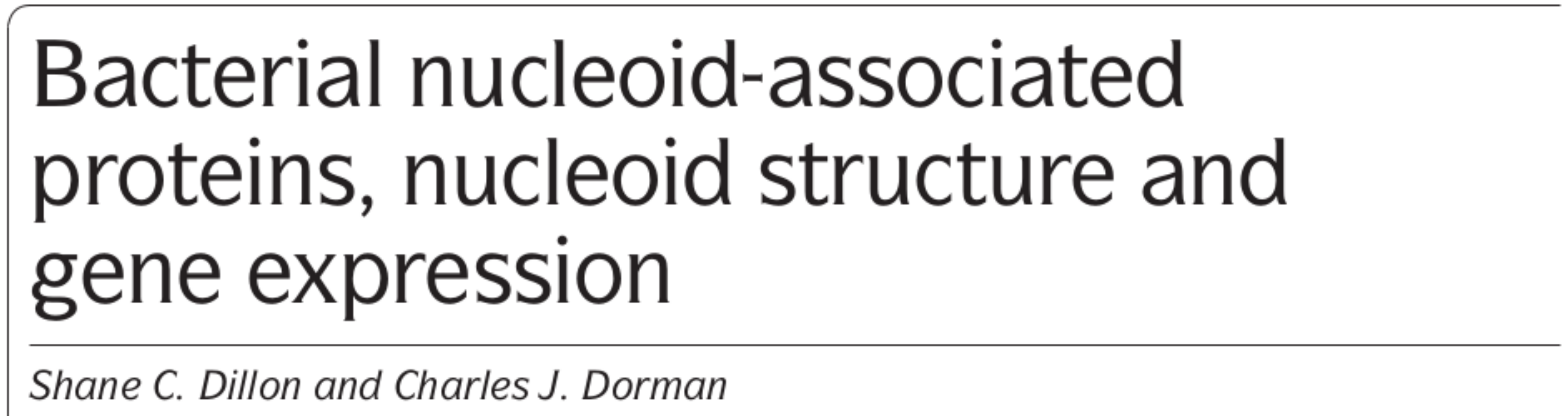






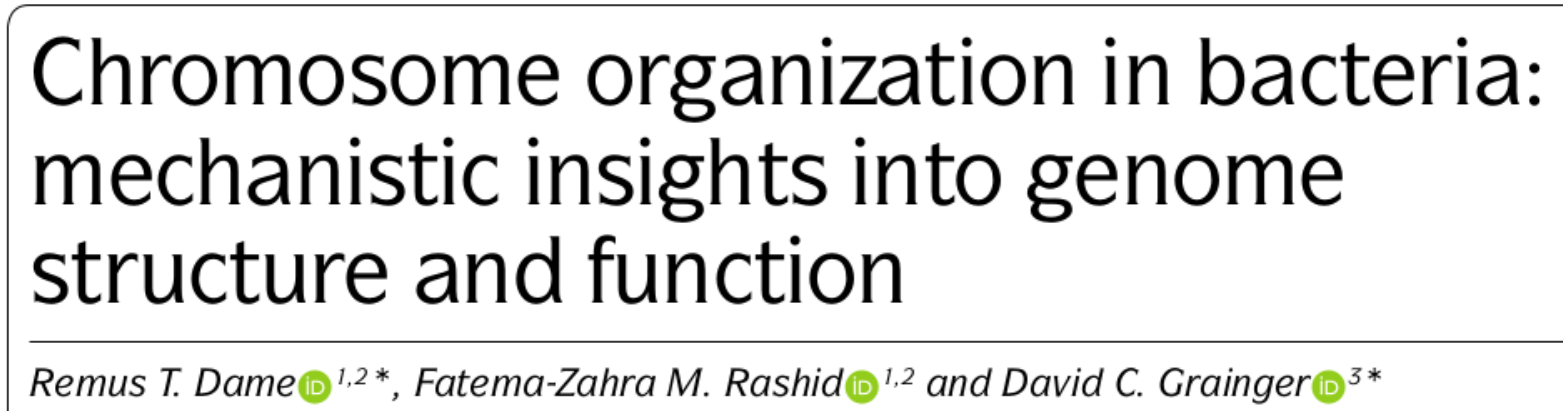


a Exponential phase of growth
b Stationary phase of growth
RNA polymerase at RNA promoters
H–NS
Transcription factories
Fis



Bacterial nucleoid-associated proteins, nucleoid structure and gene expression

Shane C. Dillon and Charles J. Dorman



Chromosome organization in bacteria: mechanistic insights into genome structure and function

Remus T. Dame[1,2], Fatema-Zahra M. Rashid[1,2] and David C. Grainger[3]**

Epigenetics

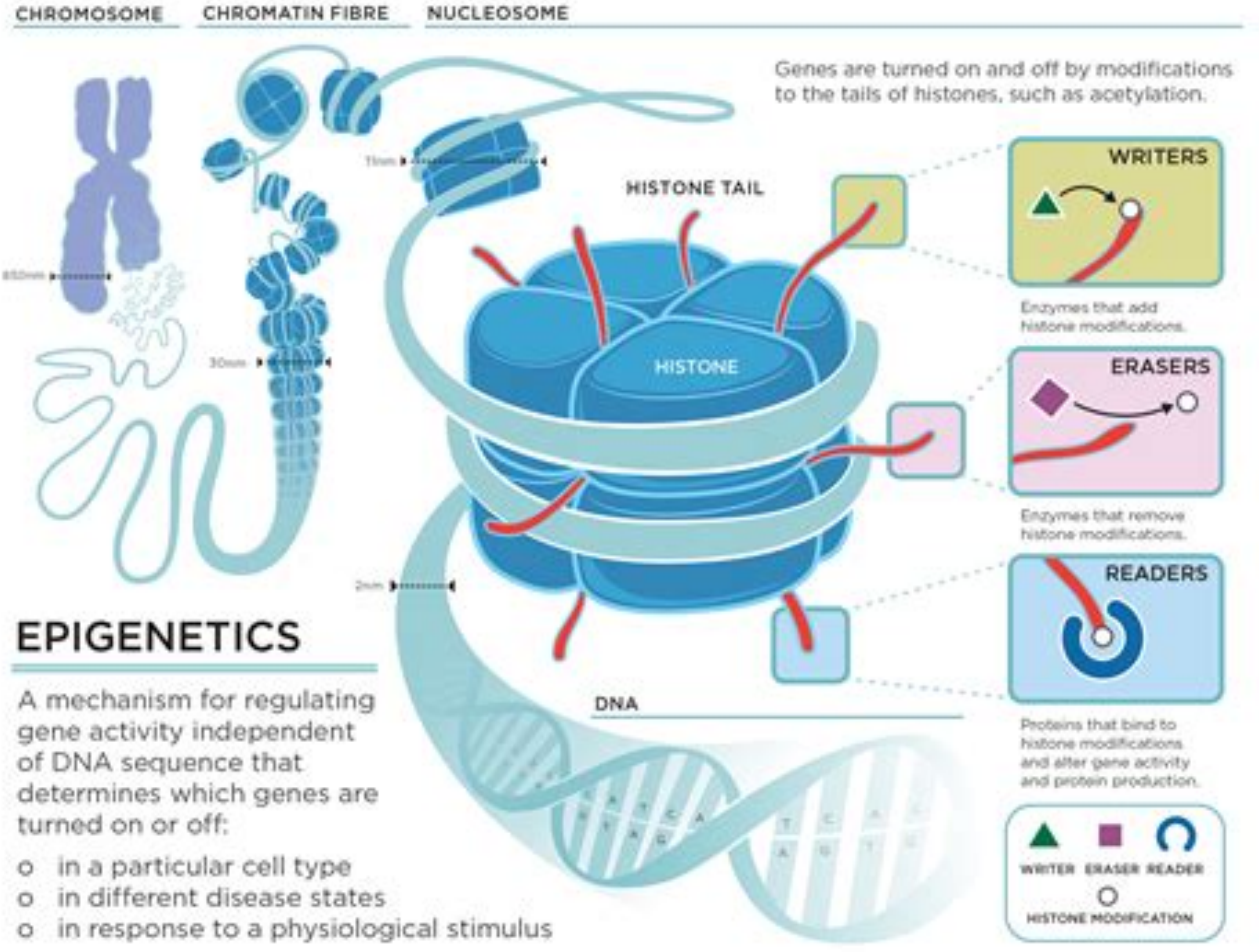
* Reversible, non-nucleotide based genetic changes during course of life (i.e. not caused by “mutations”)

* Examples of epigenetic changes include DNA methylation (DNAm), Histone modification, specific RNA sequences

* Involved in “programming” the different cell types

* Associated with development (controls sex/gender development in mammals)

* Strong environmental exposures have also been shown to affect the epigenome: cancer, smoking, BMI, Folate, ART



Genes are turned on and off by modifications to the tails of histones, such as acetylation.

Enzymes that add histone modifications.

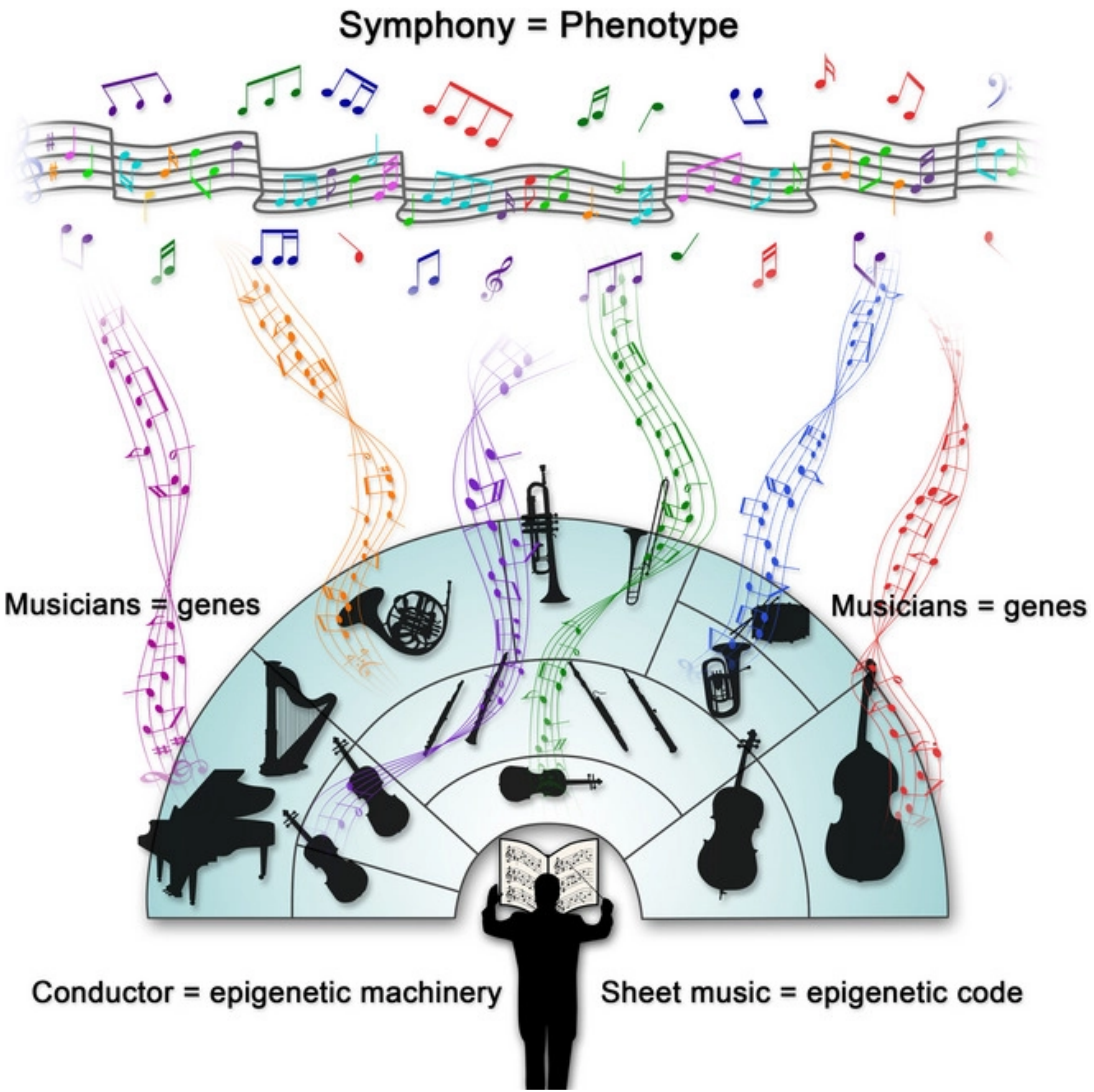
Enzymes that remove histone modifications.

Proteins that bind to histone modifications and alter gene activity and protein production.

EPIGENETICS

A mechanism for regulating gene activity independent of DNA sequence that determines which genes are turned on or off:

- in a particular cell type
- in different disease states
- in response to a physiological stimulus


Symphony = Phenotype
Musicians = genes
Musicians = genes
Conductor = epigenetic machinery
Sheet music = epigenetic code

Epigenetic data

* DNA methylation occurs mostly at Cytosine in CpG dinucleotides but other variants do exit

* Rudimentary form also in bacteria

* Often leads to C→T mutation if not attended

* Approx 2x28 mill CG dinucleotides in the human genome

* …but not all seem to be methylated

Methylation platforms

- Illumina (850k “Epic”/450k) are based on «microarray»-technology
- hybridisation with a methylated base=green light
- No hybridisation gives red light
- Intensities vary for both light

DNA methylation probes

- Red/green intensity signals converted to Methylated and Unmethylated signals
- Two probe types: Type I probes two channels, Type II probes one channel…
- …CpGs close together and Type II probes may correlate
- β_i=max($y_{i,methy}$,0)/(max($y_{i,unmethy}$,0)+max($y_{i,methy}$,0)+α) (performed during QC)
- M_i=log$_2$((max($y_{i,methy}$,0)+α)/max($y_{i,unmethy}$,0)+α) (logit transform could make analysis more robust, but values are more difficult to interpret)

- $\beta_i=2^{Mi}/(2^{Mi}+1);M_i=\log_2(\beta_i/(1-\beta_i))$

Illumina 450k nomenclature

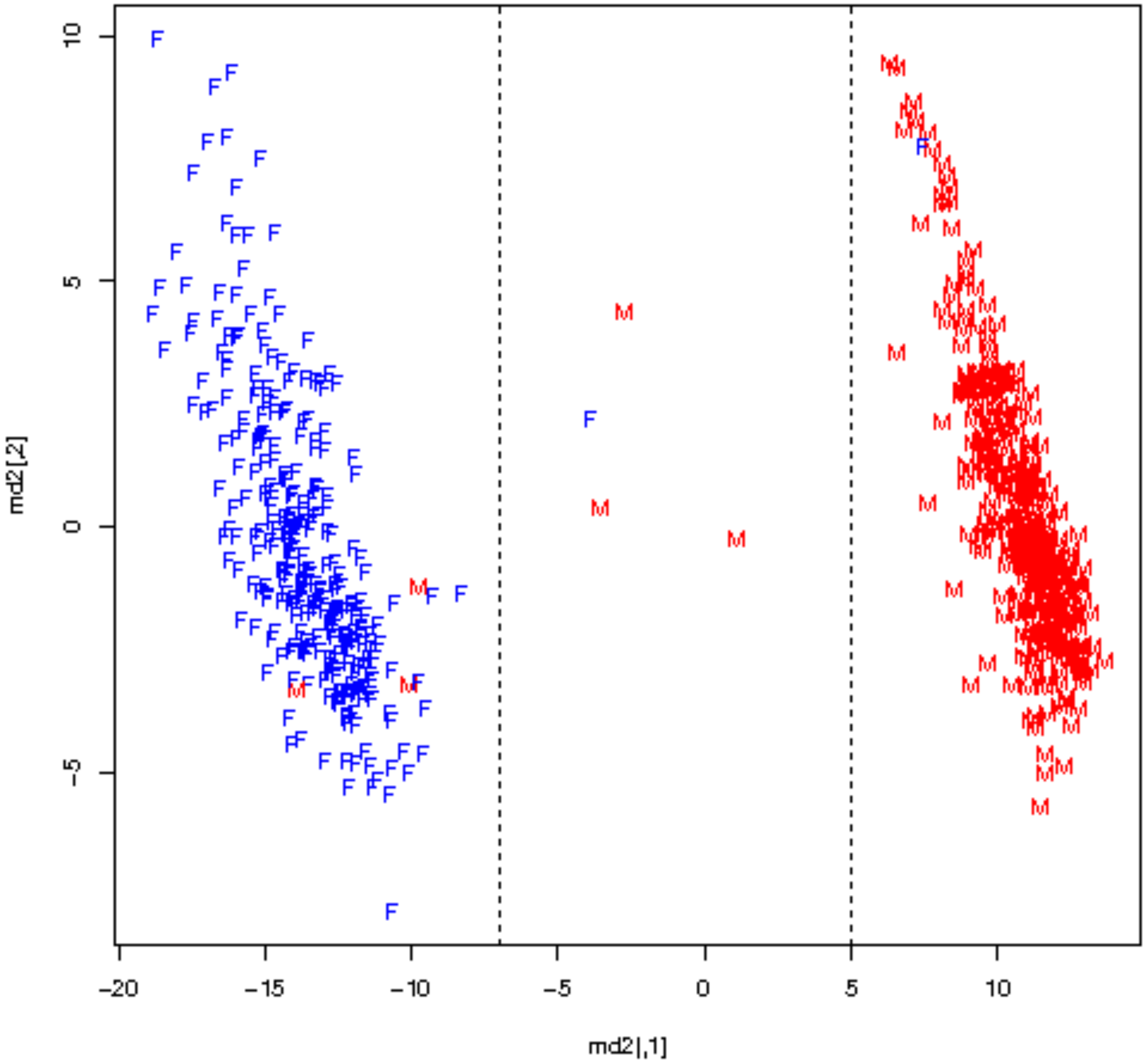
- One observation (1 sample)=one array (450k methylation sites)
- 12 observations (8 for EPIC) on 1 slide
- 1 plate max 8 slides (96 arrays)

QC - Workflow –from start to finish

- Quality control
  - Removal of bad samples
  - Removal of bad probes
  - Removal of SNP based probes
  - Removal of inserted control probes
  - Removal of gender-issues
- Normalization
  - Correct for technical bias
  - Correct for technology-specific features
  - Type I/II probes (adjustment for red/green intensity)

*** NOTE!! QC takes time ,not so easy as it looks, DOCUMENT EVERYTHING and SAVE EVERY CHANGE, PIs often impatient**

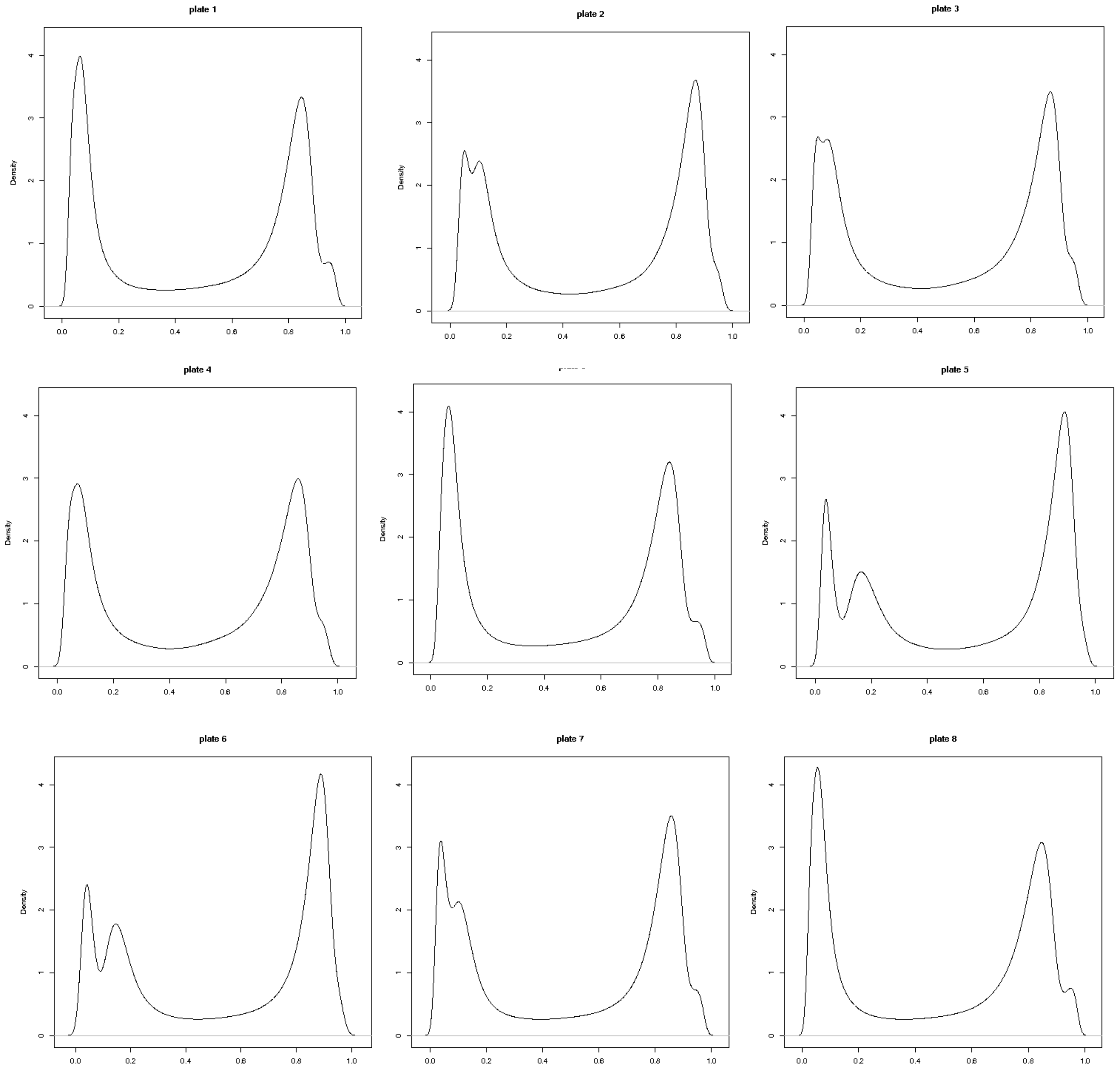
MDS plot to evaluate sex outliers



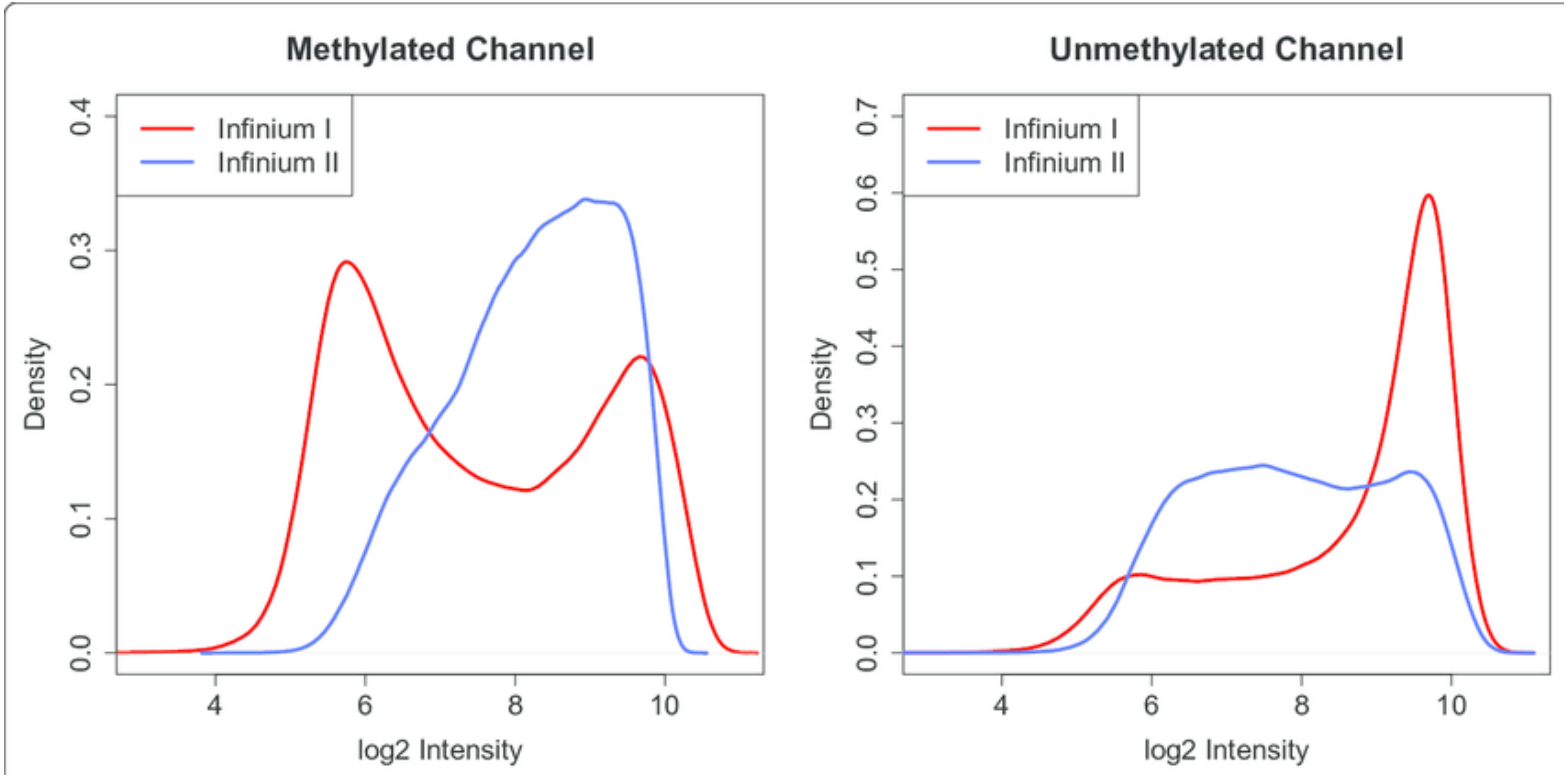
8 outliers

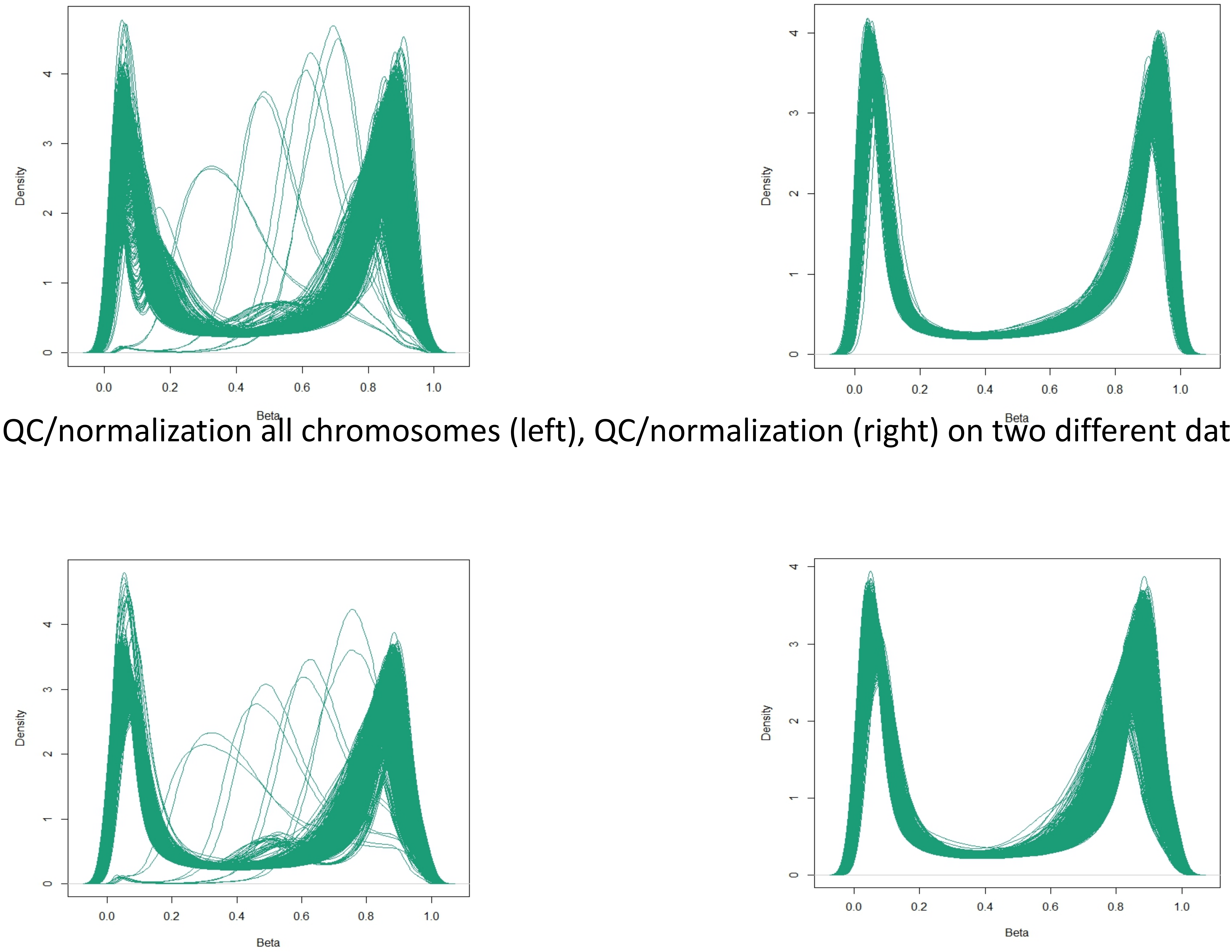
Betas by Plate

Plate #	N run	N passed QC	% passed
1	96	92	96%
2	96	69	72%
3	96	80	83%
4	96	87	91%
5	96	67	70%
6	96	83	86%
7	96	90	94%
8	96	88	92%
9	96	69	72%
Total	864	725	



Last but not least, probe correction

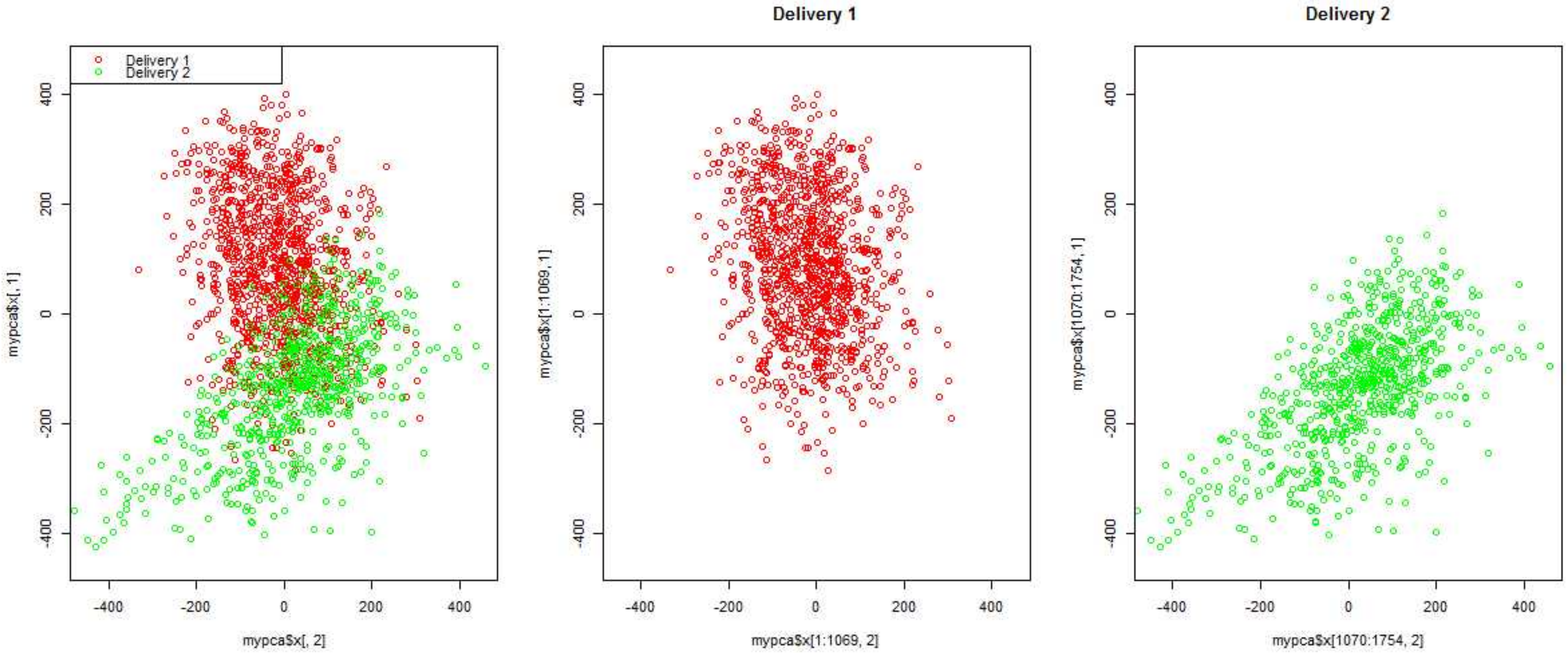


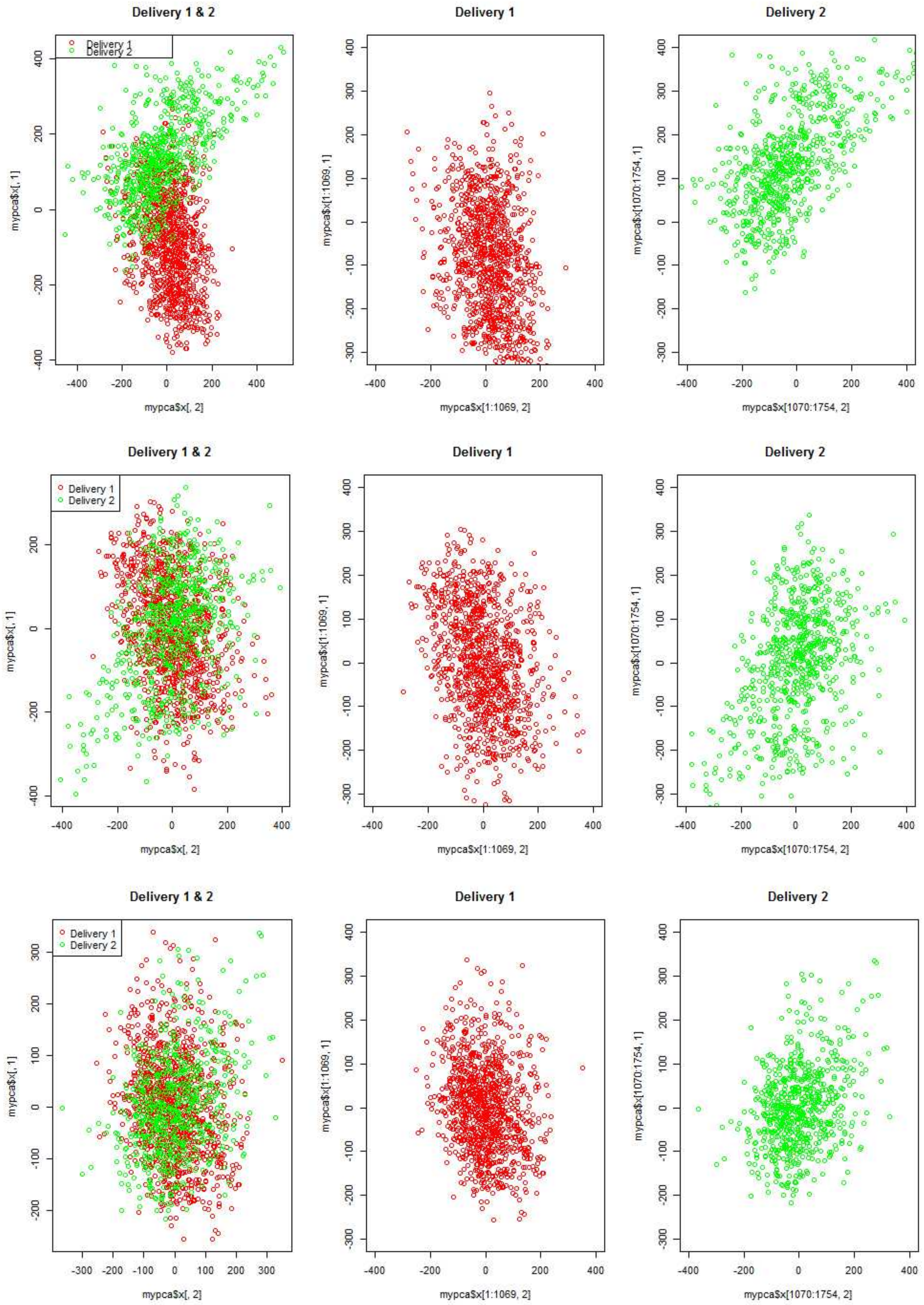


No QC/normalization all chromosomes (left), QC/normalization (right) on two different datasets

Dataset (batch) correction

- Necessary when combining 2 or more datasets
- Colored wrt dataset (batch), 2 pictures,
- PCA of dataset 1 and dataset 2 before ComBat, all chromosomes




Delivery 1 & 2
Delivery 1
Delivery 2
mypca$x[, 1]
mypca$x[, 2]
mypca$x[1:1069, 1]
mypca$x[1:1069, 2]
mypca$x[1070:1754, 1]
mypca$x[1070:1754, 2]
-400
-300
-200
-100
0
100
200
300
400
Delivery 1 & 2
Delivery 1
Delivery 2
mypca$x[, 1]
mypca$x[, 2]
mypca$x[1:1069, 1]
mypca$x[1:1069, 2]
mypca$x[1070:1754, 1]
mypca$x[1070:1754, 2]
Delivery 1 & 2
Delivery 1
Delivery 2
mypca$x[, 1]
mypca$x[, 2]
mypca$x[1:1069, 1]
mypca$x[1:1069, 2]
mypca$x[1070:1754, 1]
mypca$x[1070:1754, 2]
-300
-200
-100
0
100
200
300

Papers that will get you going with pre-processing and QC

- RnBeads 2.0: comprehensive analysis of DNA methylation data: Fabian Müller, Michael Scherer, Yassen Assenov3*†, Pavlo Lutsik, Jörn Walter, Thomas Lengauer and Christoph Bock
- Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi: *Fortin JP, Triche TJ Jr, Hansen KD.*
- Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays: *Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, Irizarry RA.*
- A data-driven approach to preprocessing Illumina 450K methylation array data: *Pidsley R, Y Wong CC, Volta M, Lunnon K, Mill J, Schalkwyk LC.* (wateRmellon package)
- A framework for analyzing DNA methylation data from Illumina Infinium HumanMethylation450 BeadChip.: *Wang Z, Wu X, Wang Y.*
- A systematic assessment of normalization approaches for the Infinium 450K methylation platform: *Michael C Wu Bonnie R Joubert Pei-fen Kuan Siri E Håberg Wenche Nystad Shyamal D Peddada and Stephanie J London*
- quantro: a data-driven approach to guide the choice of an appropriate normalization method: *Hicks SC, Irizarry RA*