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 with the aid of a “mother”
- 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

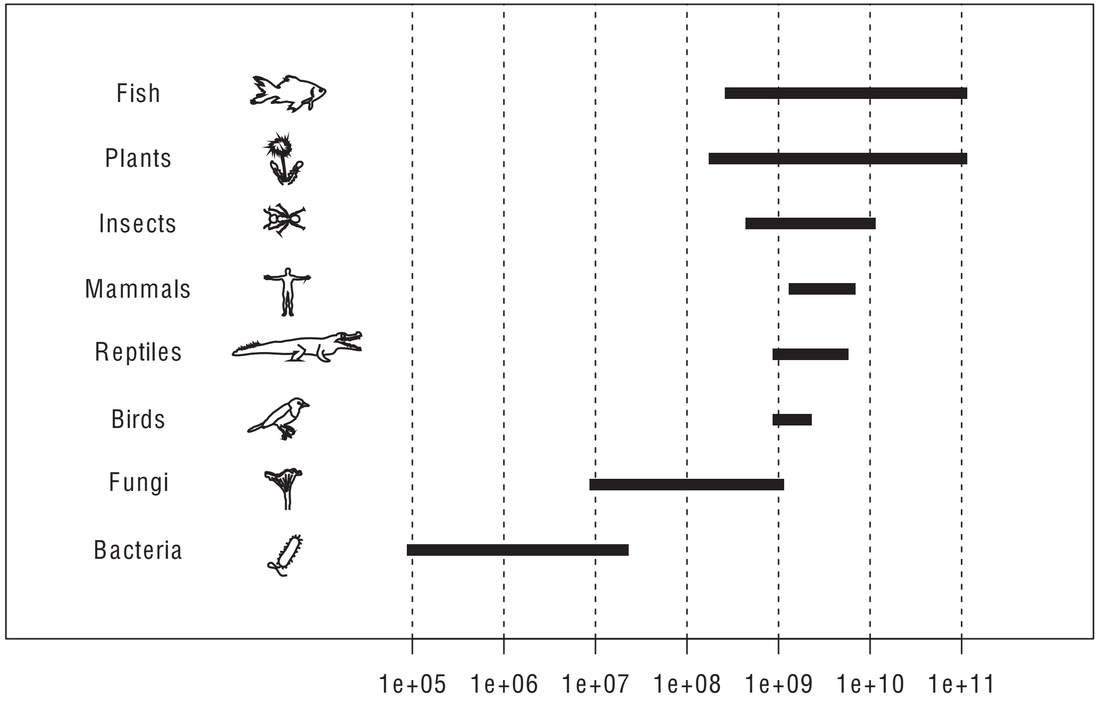
- EBI: Genomics is the study of whole genomes of organisms, and incorporates elements from genetics. Genomics uses a combination of recombinant DNA, DNA sequencing methods, and bioinformatics to sequence, assemble, and analyse the structure and function of genomes

# 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??)

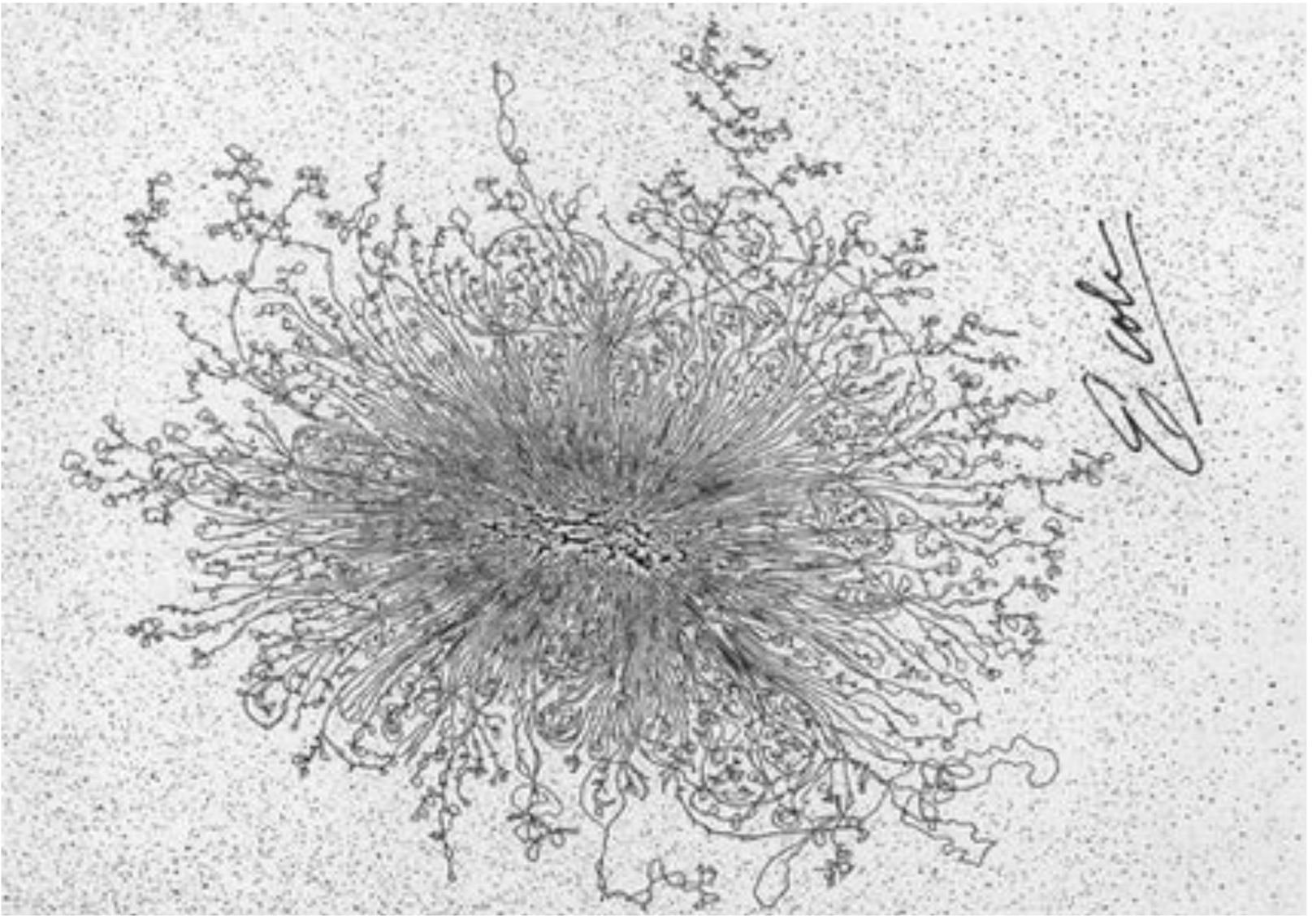


**Estimate of the Mutation Rate per Nucleotide in Humans**

**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





equipment, and to Dr. G. E. R. Deacon and the captain and officers of R.R.S. *Discovery II* for their part in making the observations.

[1] Young, F. B., Gerrard, H., and Jevons, W., *Phil. Mag.*, **40**, 149 (1920).
[2] Longuet-Higgins, M. S., *Mon. Not. Roy. Astro. Soc., Geophys. Supp.*, **5**, 285 (1949).
[3] Von Arx, W. S., Woods Hole Papers in Phys. Oceanog. Meteor., **11** (3) (1950).
[4] Ekman, V. W., *Arkiv. Mat. Astron. Fysik. (Stockholm)*, **2** (11) (1905).

# MOLECULAR STRUCTURE OF NUCLEIC ACIDS

## A Structure for Deoxyribose Nucleic Acid

WE wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biological interest.

A structure for nucleic acid has already been proposed by Pauling and Corey[1]. They kindly made their manuscript available to us in advance of publication. Their model consists of three intertwined chains, with the phosphates near the fibre axis, and the bases on the outside. In our opinion, this structure is unsatisfactory for two reasons: (1) We believe that the material which gives the X-ray diagrams is the salt, not the free acid. Without the acidic hydrogen atoms it is not clear what forces would hold the structure together, especially as the negatively charged phosphates near the axis will repel each other. (2) Some of the van der Waals distances appear to be too small.

Another three-chain structure has also been suggested by Fraser (in the press). In his model the phosphates are on the outside and the bases on the inside, linked together by hydrogen bonds. This structure as described is rather ill-defined, and for this reason we shall not comment on it.

This figure is purely diagrammatic. The two ribbons symbolize the two phosphate—sugar chains, and the horizontal rods the pairs of bases holding the chains together. The vertical line marks the fibre axis

We wish to put forward a radically different structure for the salt of deoxyribose nucleic acid. This structure has two helical chains each coiled round the same axis (see diagram). We have made the usual chemical assumptions, namely, that each chain consists of phosphate diester groups joining β-D-deoxyribofuranose residues with 3′,5′ linkages. The two chains (but not their bases) are related by a dyad perpendicular to the fibre axis. Both chains follow right-handed helices, but owing to the dyad the sequences of the atoms in the two chains run in opposite directions. Each chain loosely resembles Furberg's[2] model No. 1; that is, the bases are on the inside of the helix and the phosphates on the outside. The configuration of the sugar and the atoms near it is close to Furberg's 'standard configuration', the sugar being roughly perpendicular to the attached base. There is a residue on each chain every 3·4 A. in the *z*-direction. We have assumed an angle of 36° between adjacent residues in the same chain, so that the structure repeats after 10 residues on each chain, that is, after 34 A. The distance of a phosphorus atom from the fibre axis is 10 A. As the phosphates are on the outside, cations have easy access to them.

The structure is an open one, and its water content is rather high. At lower water contents we would expect the bases to tilt so that the structure could become more compact.

The novel feature of the structure is the manner in which the two chains are held together by the purine and pyrimidine bases. The planes of the bases are perpendicular to the fibre axis. They are joined together in pairs, a single base from one chain being hydrogen-bonded to a single base from the other chain, so that the two lie side by side with identical *z*-co-ordinates. One of the pair must be a purine and the other a pyrimidine for bonding to occur. The hydrogen bonds are made as follows: purine position 1 to pyrimidine position 1; purine position 6 to pyrimidine position 6.

If it is assumed that the bases only occur in the structure in the most plausible tautomeric forms (that is, with the keto rather than the enol configurations) it is found that only specific pairs of bases can bond together. These pairs are: adenine (purine) with thymine (pyrimidine), and guanine (purine) with cytosine (pyrimidine).

In other words, if an adenine forms one member of a pair, on either chain, then on these assumptions the other member must be thymine; similarly for guanine and cytosine. The sequence of bases on a single chain does not appear to be restricted in any way. However, if only specific pairs of bases can be formed, it follows that if the sequence of bases on one chain is given, then the sequence on the other chain is automatically determined.

It has been found experimentally[3,4] that the ratio of the amounts of adenine to thymine, and the ratio of guanine to cytosine, are always very close to unity for deoxyribose nucleic acid.

It is probably impossible to build this structure with a ribose sugar in place of the deoxyribose, as the extra oxygen atom would make too close a van der Waals contact.

The previously published X-ray data[5,6] on deoxyribose nucleic acid are insufficient for a rigorous test of our structure. So far as we can tell, it is roughly compatible with the experimental data, but it must be regarded as unproved until it has been checked against more exact results. Some of these are given in the following communications. We were not aware of the details of the results presented there when we devised our structure, which rests mainly though not entirely on published experimental data and stereochemical arguments.

It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.

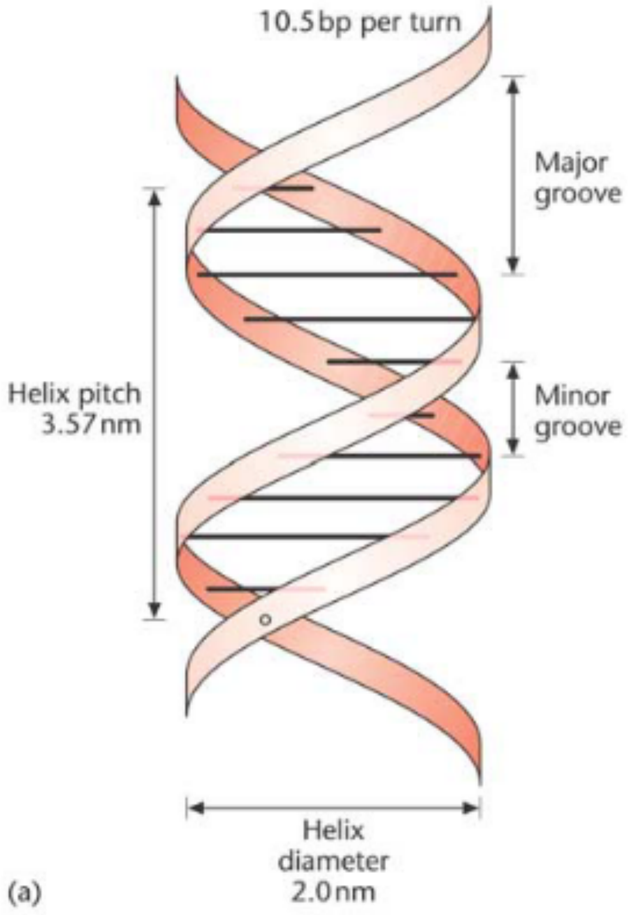
Full details of the structure, including the conditions assumed in building it, together with a set of co-ordinates for the atoms, will be published elsewhere.

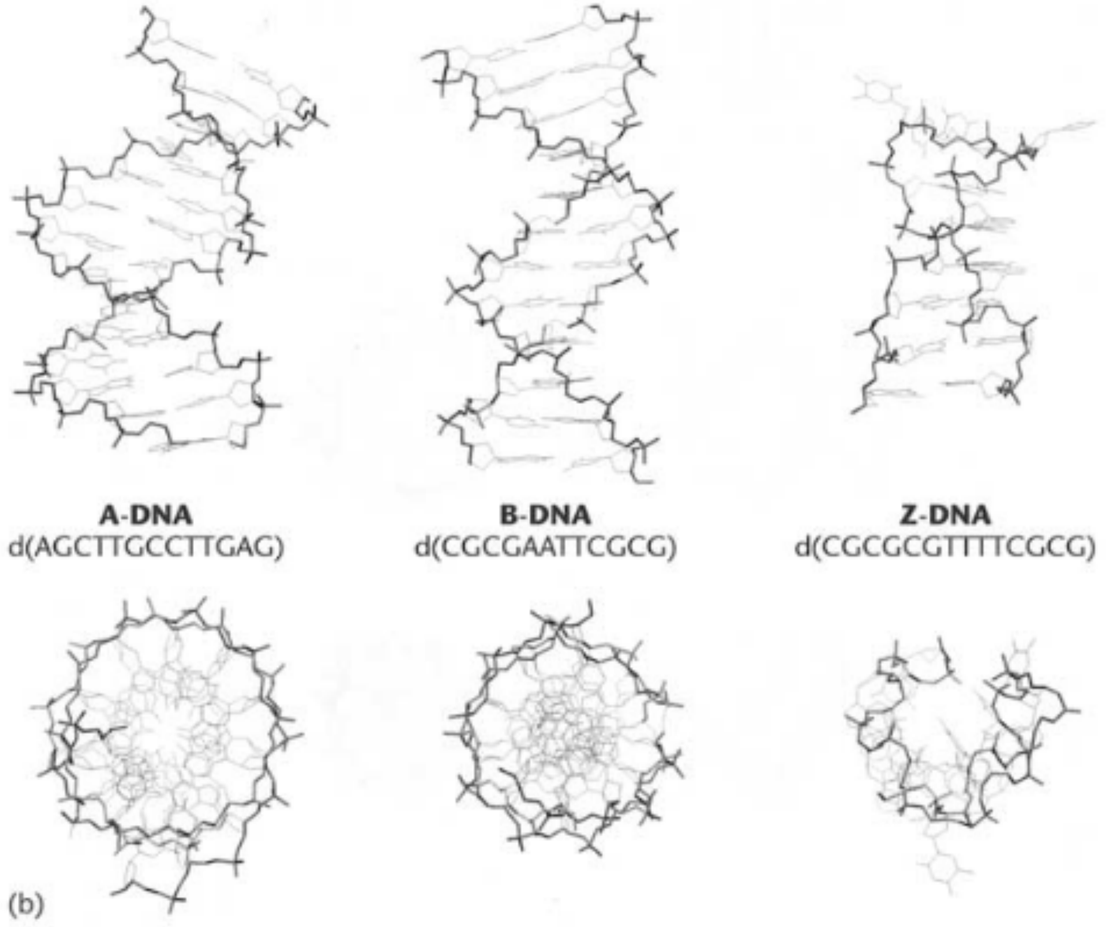
We are much indebted to Dr. Jerry Donohue for constant advice and criticism, especially on interatomic distances. We have also been stimulated by a knowledge of the general nature of the unpublished experimental results and ideas of Dr. M. H. F. Wilkins, Dr. R. E. Franklin and their co-workers at

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

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

# DNA structure and chromatin



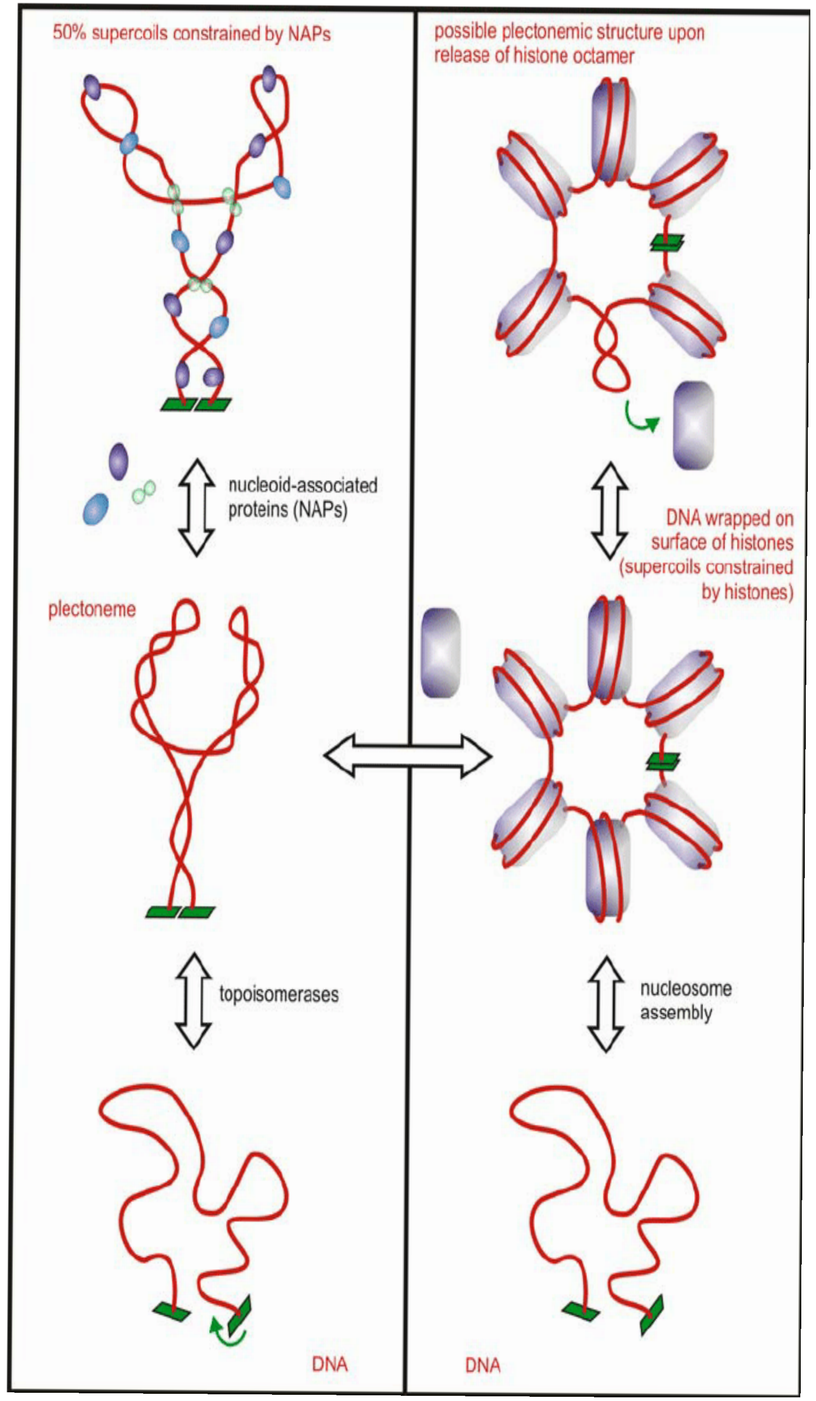





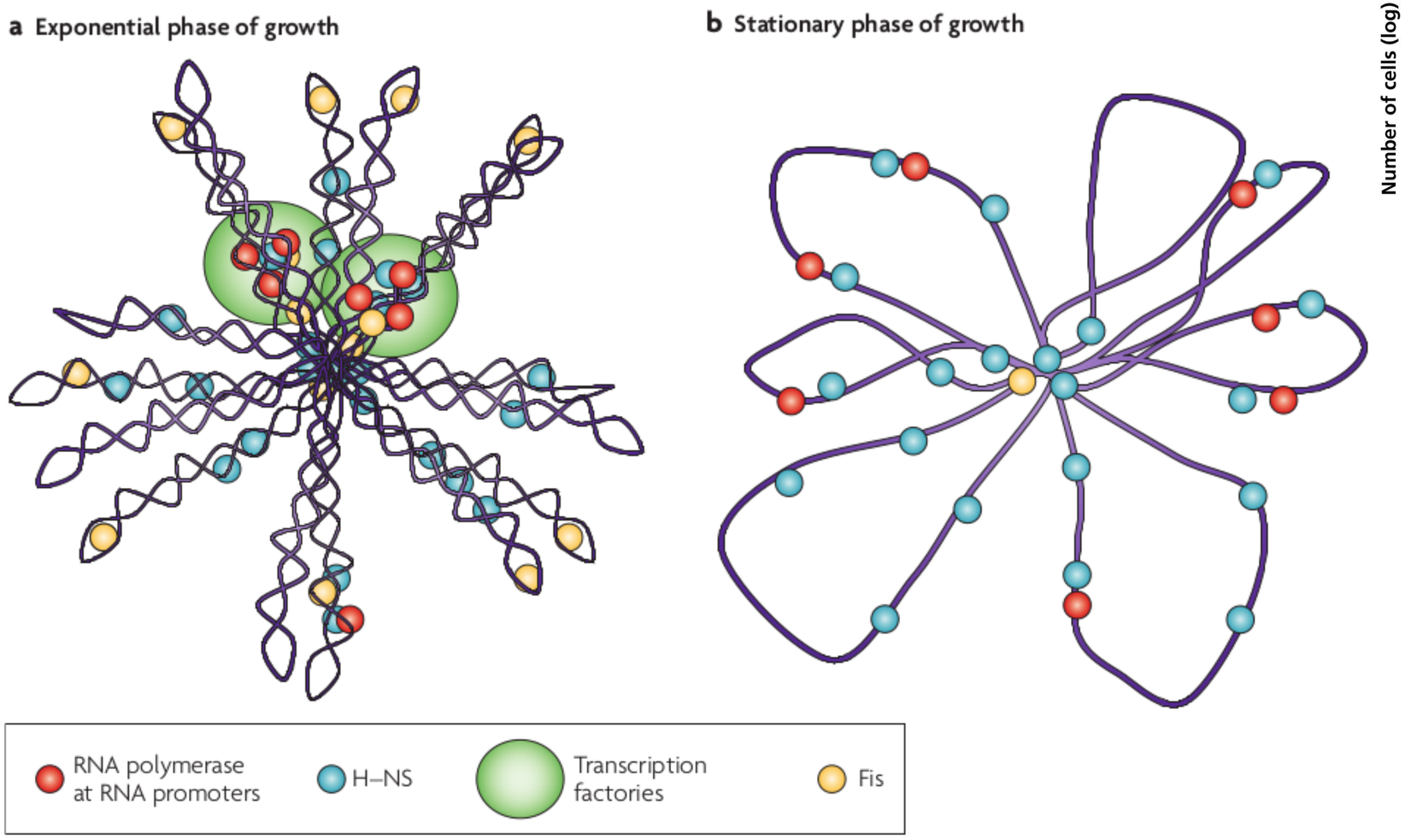
Propeller twist

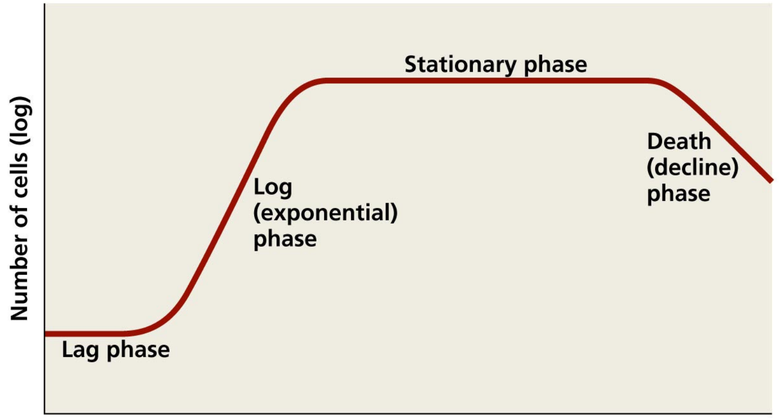

Twist

**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.


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

# DNA structure and transcription





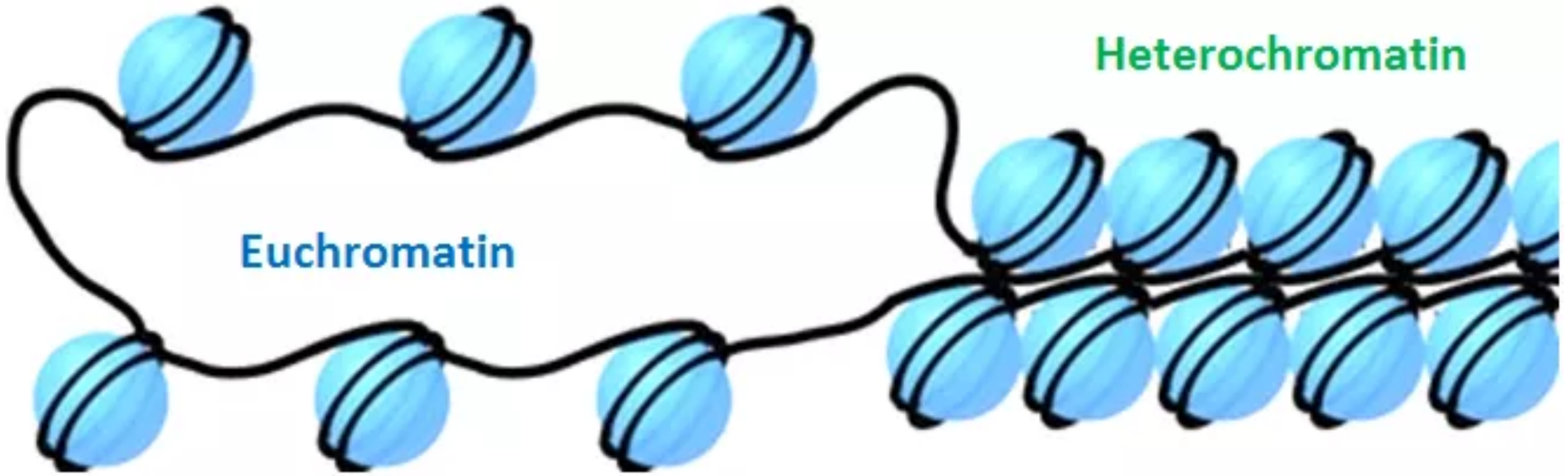

Bacterial nucleoid-associated proteins, nucleoid structure and gene expression

*Shane C. Dillon and Charles J. Dorman*

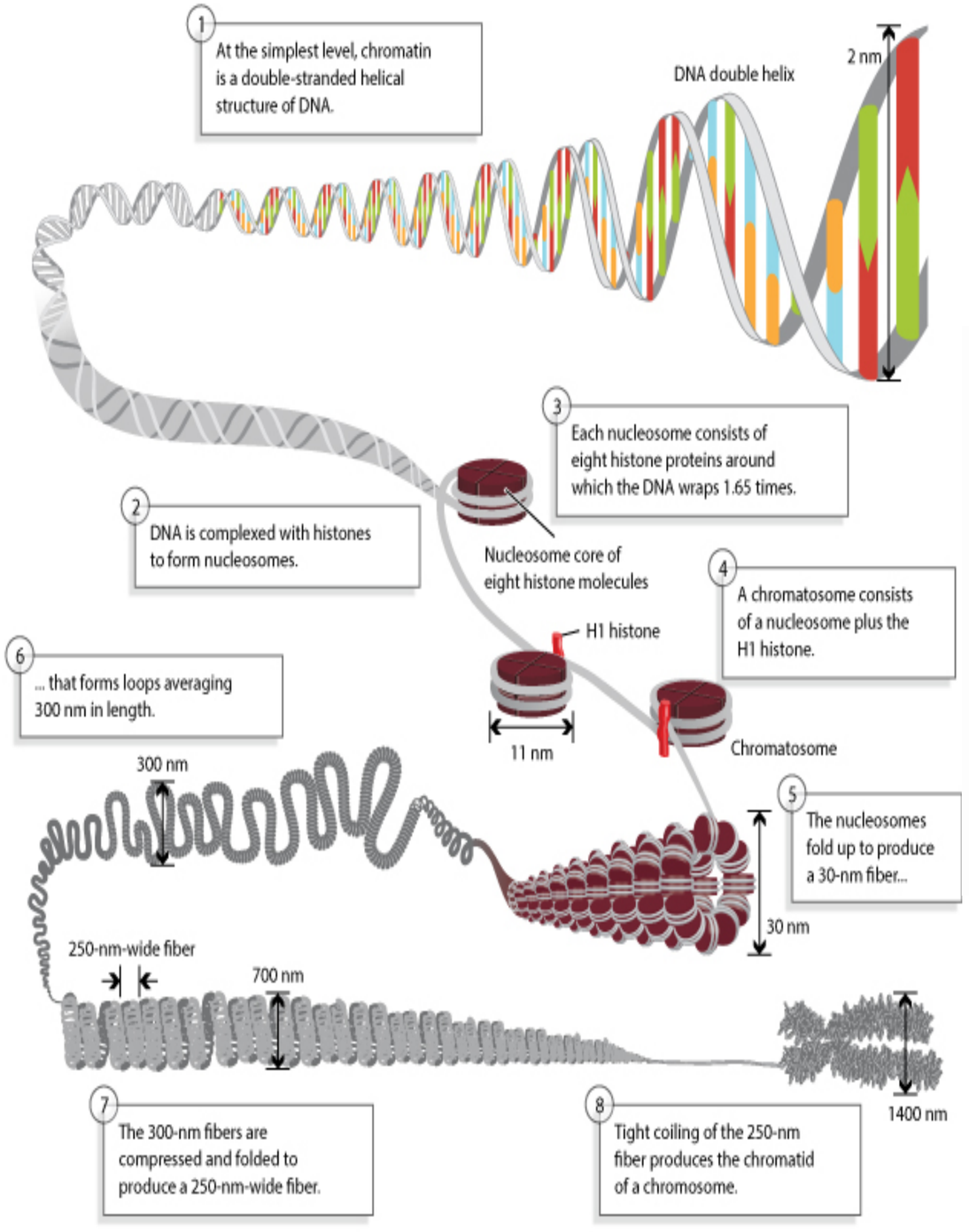
Chromosome organization in bacteria: mechanistic insights into genome structure and function

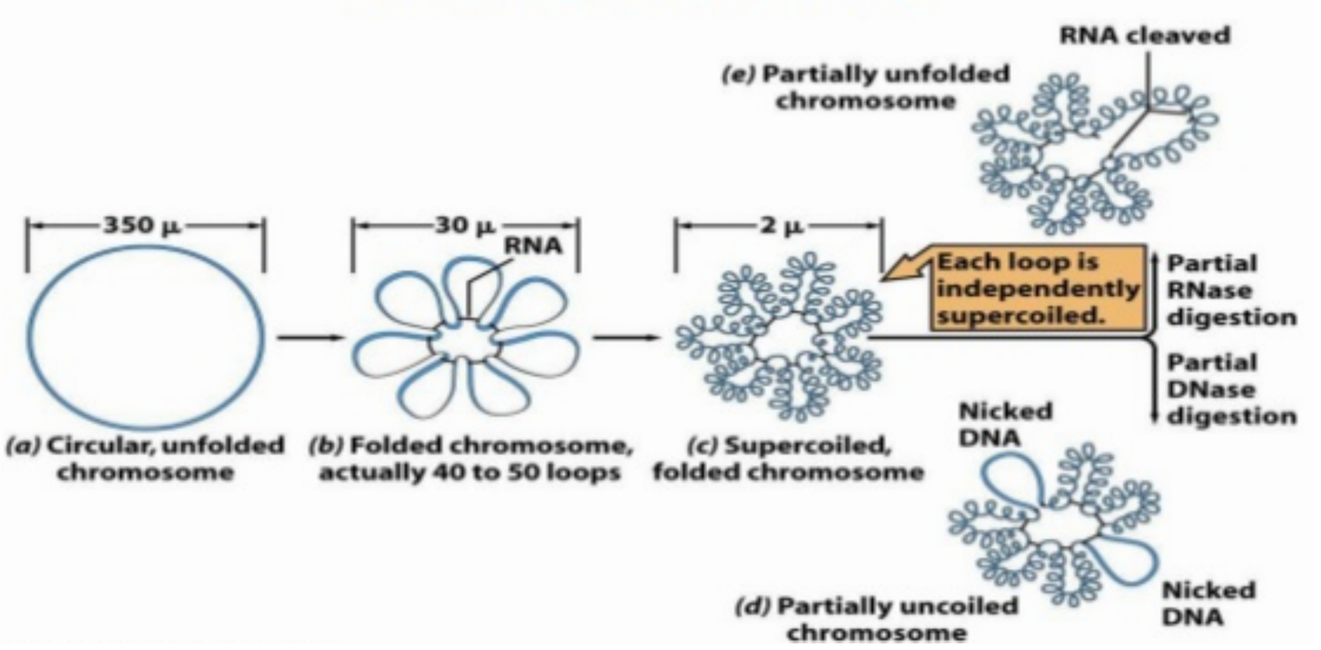
*Remus T. Dame[1,2]*, Fatema-Zahra M. Rashid[1,2] and David C. Grainger[3]**

- Constitutive
- Facultative



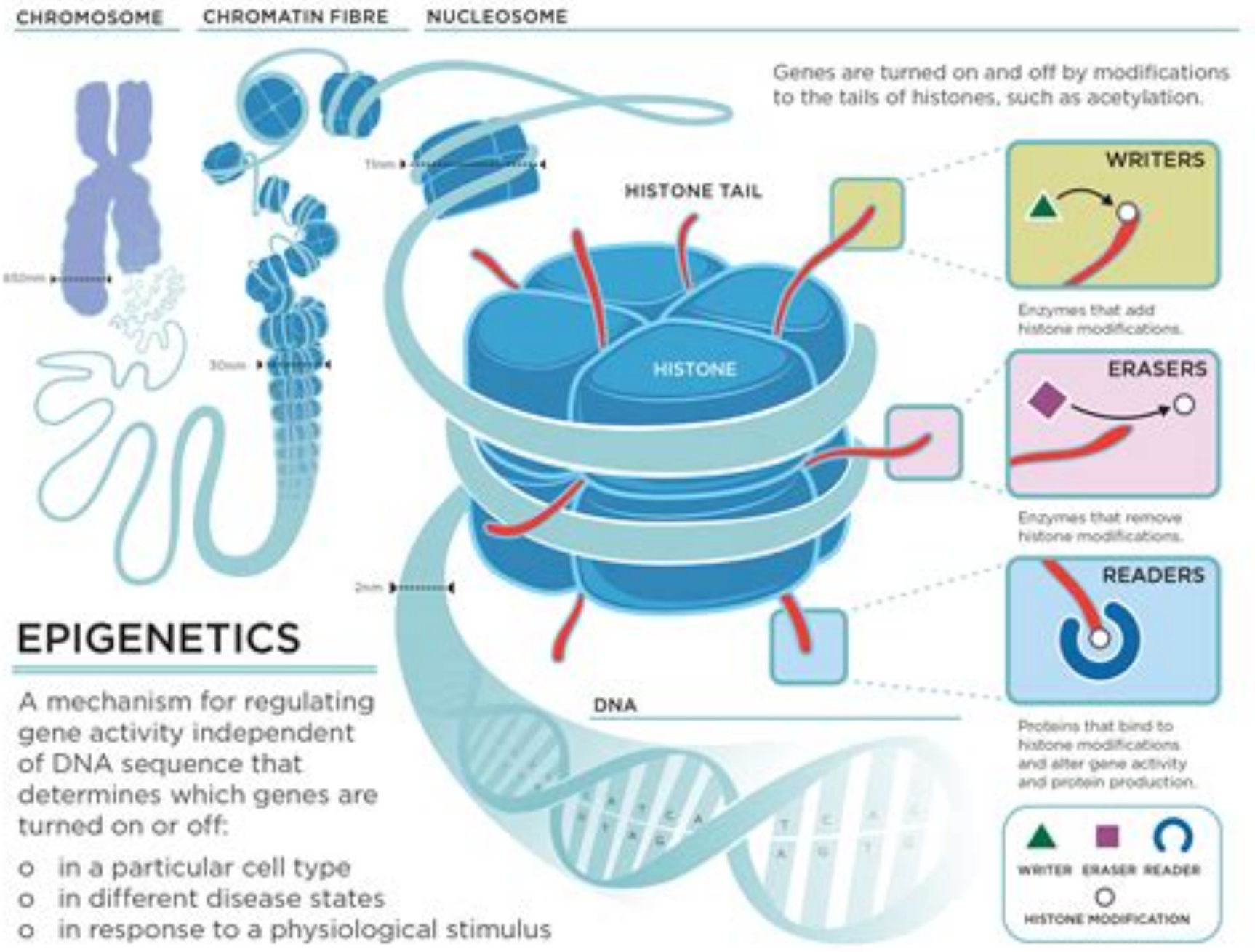

Time


1
At the simplest level, chromatin is a double-stranded helical structure of DNA.
DNA double helix
2 nm
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.
Nucleosome core of eight histone molecules
4
A chromatosome consists of a nucleosome plus the H1 histone.
H1 histone
11 nm
Chromatosome
6
... that forms loops averaging 300 nm in length.
300 nm
5
The nucleosomes fold up to produce a 30-nm fiber...
30 nm
250-nm-wide fiber
700 nm
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.
1400 nm


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

# Epigenetics

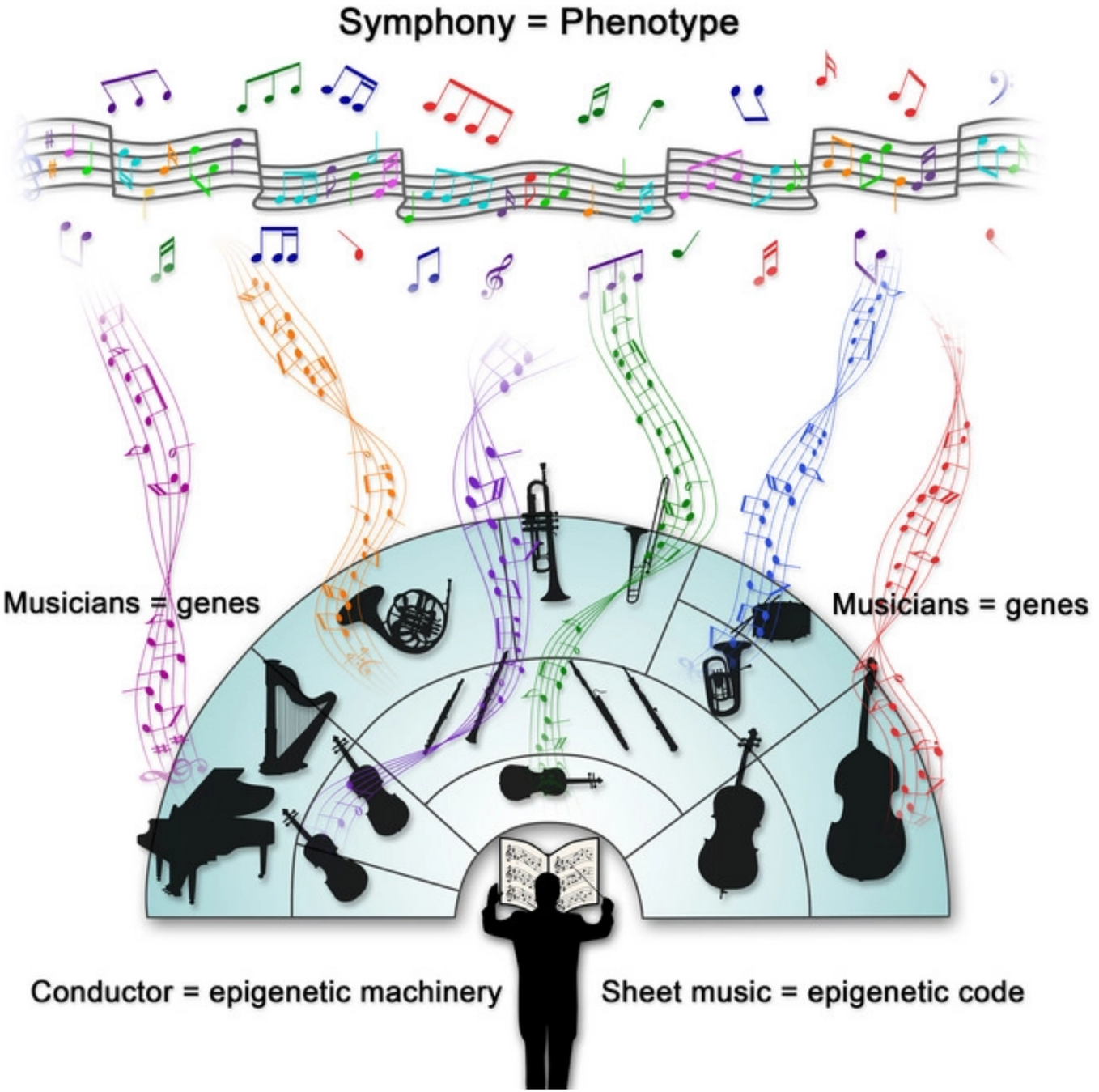
* Reversible, non-nucleotide based genetic changes during course of life (i.e. not caused by “mutations”)
* Examples of epigenetic changes include DNA methylation (DNAm), Histone modification, specific RNA sequences
* Involved in “programming” the different cell types
* Associated with development (controls sex/gender development in mammals)
* Strong environmental exposures have also been shown to affect the epigenome: cancer, smoking, BMI, Folate, ART
* DNA methylation is closely connected to DNA structure




# 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

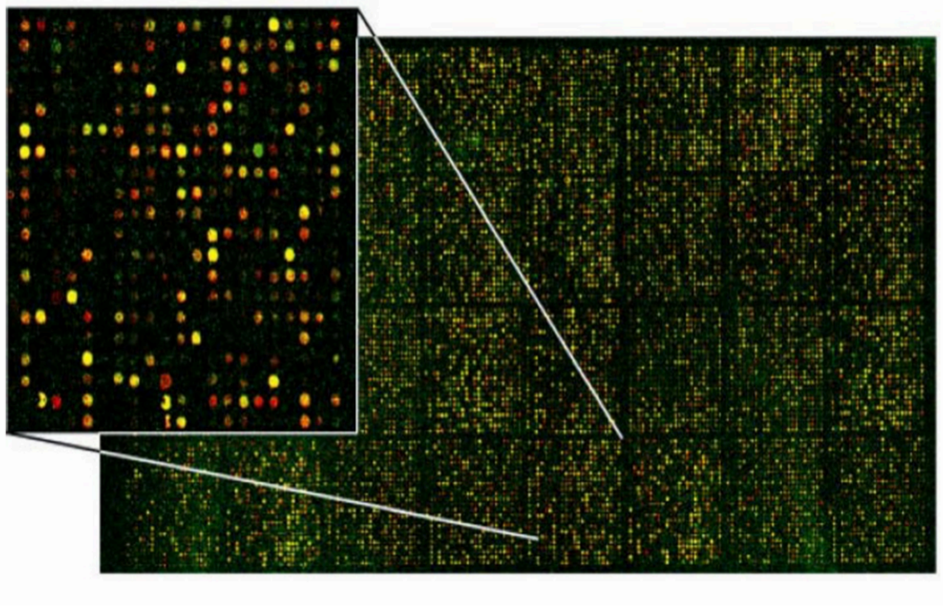

Symphony = Phenotype
Musicians = genes
Musicians = genes
Conductor = epigenetic machinery
Sheet music = epigenetic code

# 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

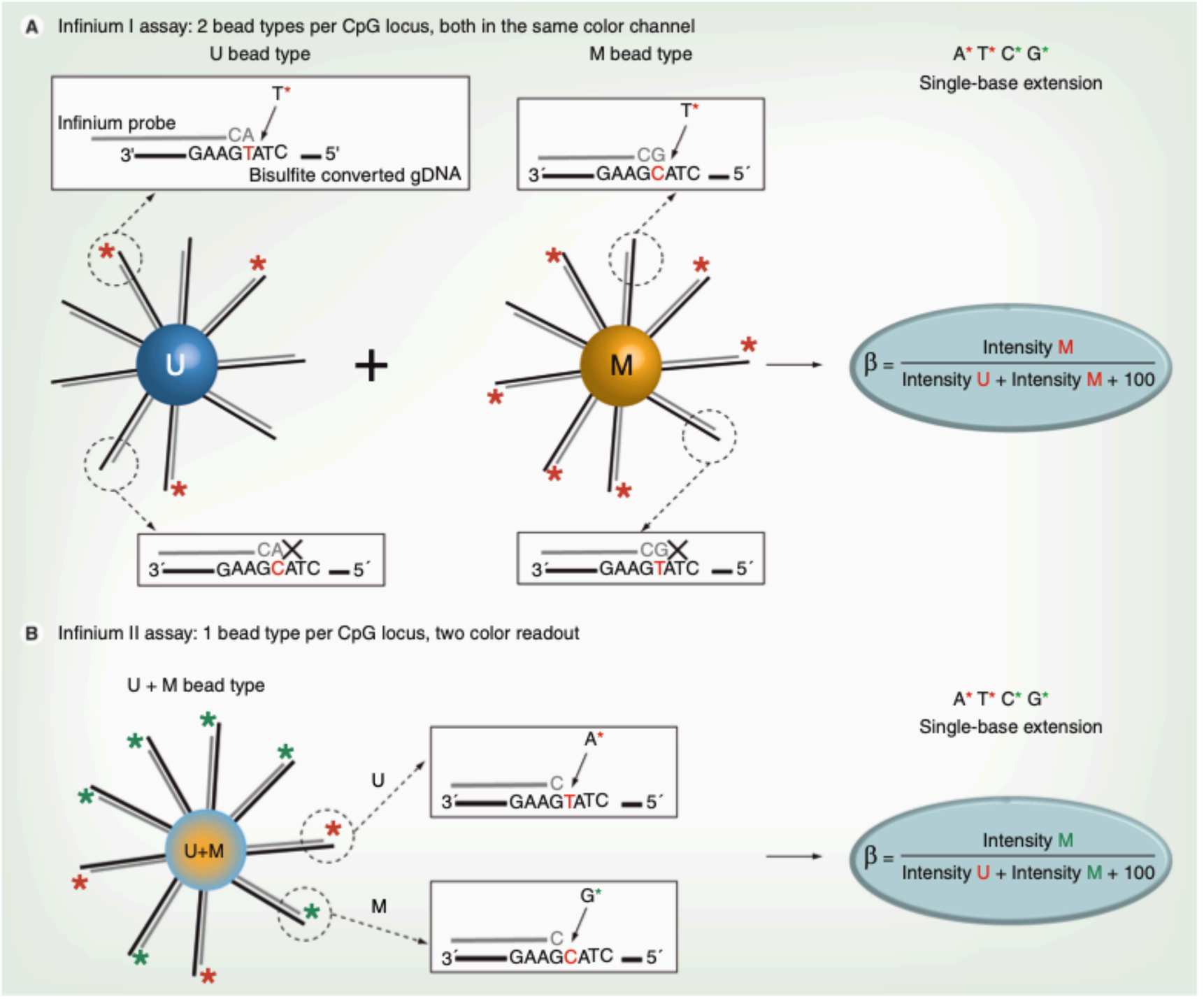
# Methylation platforms

- Illumina (850k “Epic”/450k) are based on «microarray»-technology
- Similar to GWAS but with an added bisulfite convertion step:
  - Uracil - not methylated Cytosine methylated
- Light intensities from hybrization converted to continuos values



# DNA methylation probes

- Type I (First on the 27K Array): each CpG targeted by two 50bp probes
  - One hybridizes to (M)ethylated site the other to the (U)nmethylated site
  - Assumes that underlying probe CpGs have the same methylation status
- Type II probes one channel…
  - One probe per CpG
  - No assumption that underlying probe CpGs have same methylation status. Each probe can interrogate 3 sites.


A Infinium I assay: 2 bead types per CpG locus, both in the same color channel
U bead type
M bead type
A* T* C* G*
Single-base extension
Infinium probe
CA
3'—GAAGTATC—5'
Bisulfite converted gDNA
T*
CG
3'—GAAGCATC—5'
U
M
CA
3'—GAAGCATC—5'
CG
3'—GAAGTATC—5'
β = Intensity M / Intensity U + Intensity M + 100
B Infinium II assay: 1 bead type per CpG locus, two color readout
U + M bead type
U+M
U
M
A*
C
3'—GAAGTATC—5'
G*
C
3'—GAAGCATC—5'
A* T* C* G*
Single-base extension
β = Intensity M / Intensity U + Intensity M + 100

# Calculating intensities

- $\beta_i = M_i/(M_i+U_i+\alpha)$ - performed during QC
  - $\beta_i = \max(y_{i,methy},0)/(\max(y_{i,unmethy},0)+\max(y_{i,methy},0)+\alpha)$
- logit transform could make analysis more robust, but values are more difficult to interpret
  - $M_i = \log_2((\max(y_{i,methy},0)+\alpha)/\max(y_{i,unmethy},0)+\alpha)$
- Transform back and forth between M and β
  - $\beta_i = 2^{Mi}/(2^{Mi}+1); M_i = \log_2(\beta_i/(1-\beta_i))$

# Illumina array organization 450K

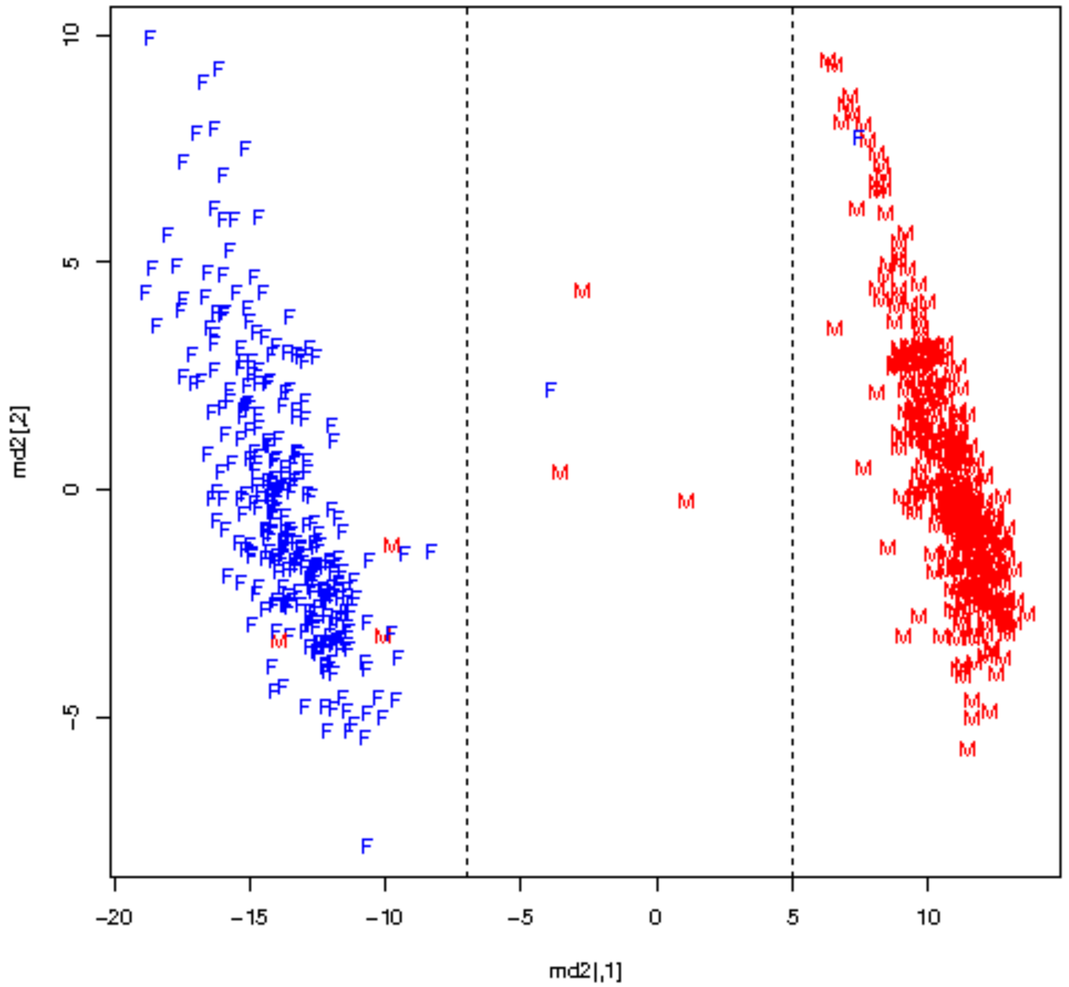
- 450K
  - One observation (1 sample)=one array
  - 12 observations on 1 slide
  - 1 plate max 8 slides (96 arrays)
- EPIC (850K)
  - Each sample one array
  - 8 arrays on a physical slide
  - 8 slide per plate (64 arrays)

# 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**

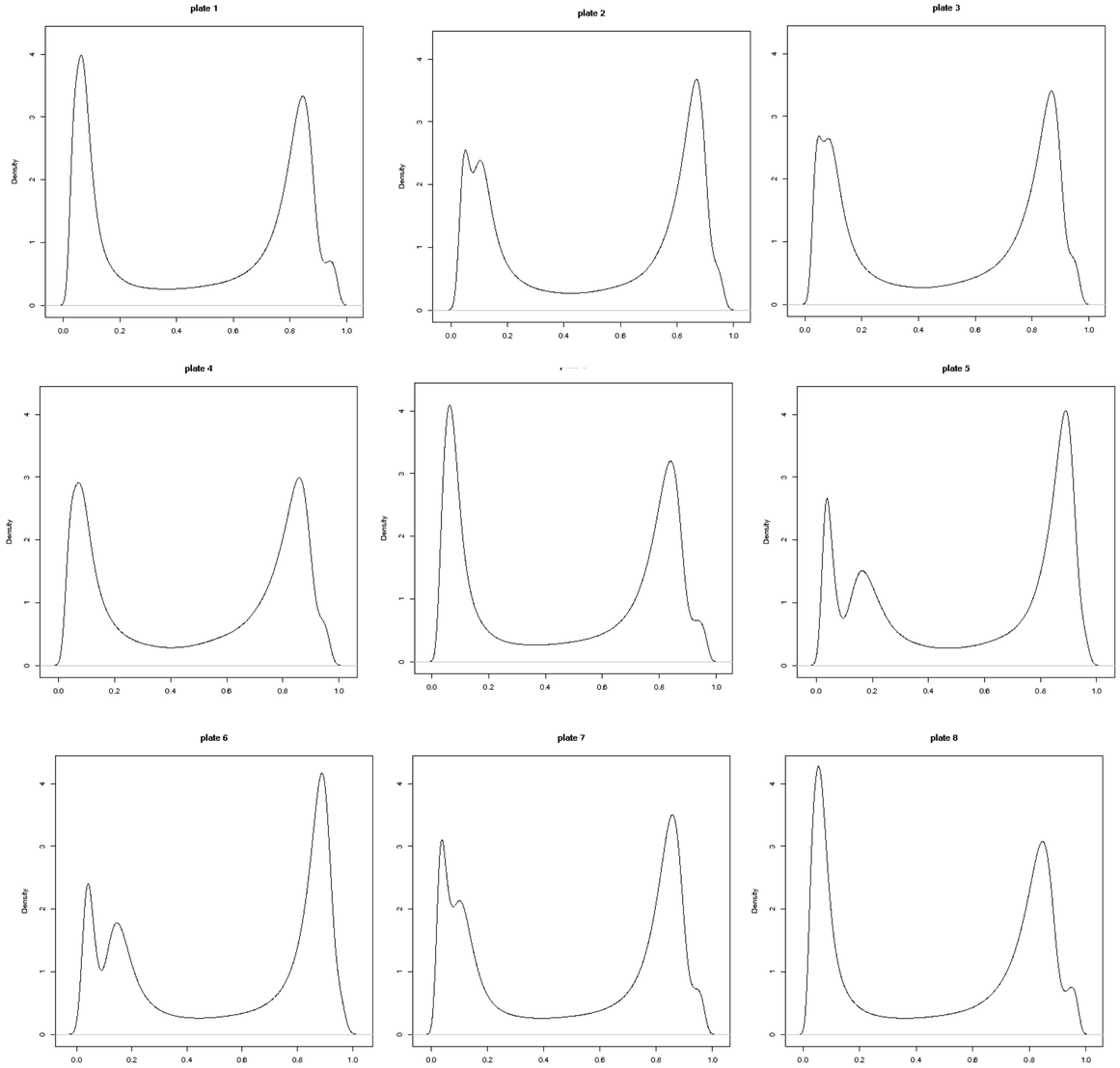
# MDS plot to evaluate sex outliers

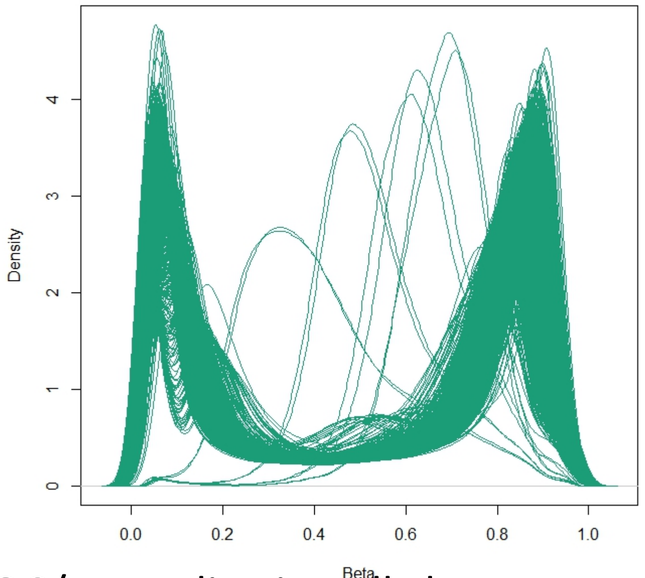


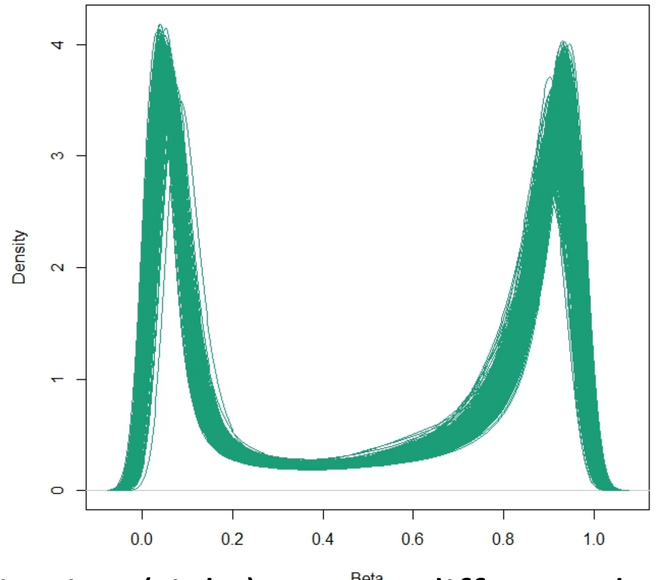

8 outliers

# Betas by Plate

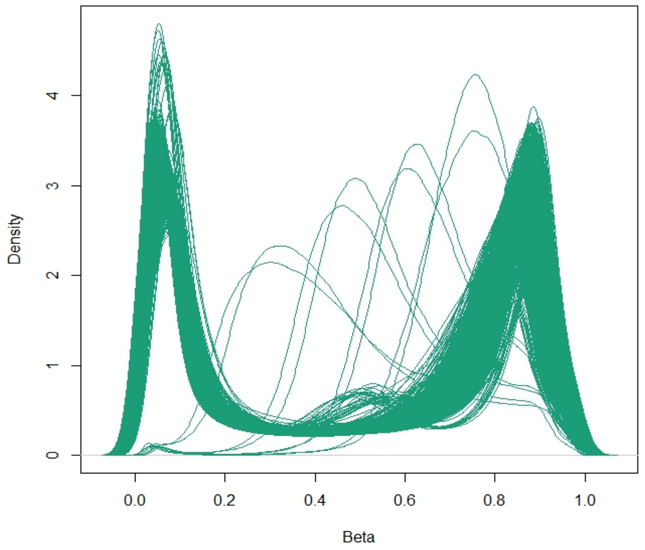
| Plate # | N run | N passed QC | % passed |
|---|---|---|---|
| 1 | 96 | 92 | 96% |
| 2 | 96 | 69 | 72% |
| 3 | 96 | 80 | 83% |
| 4 | 96 | 87 | 91% |
| 5 | 96 | 67 | 70% |
| 6 | 96 | 83 | 86% |
| 7 | 96 | 90 | 94% |
| 8 | 96 | 88 | 92% |
| 9 | 96 | 69 | 72% |
| Total | 864 | 725 | |

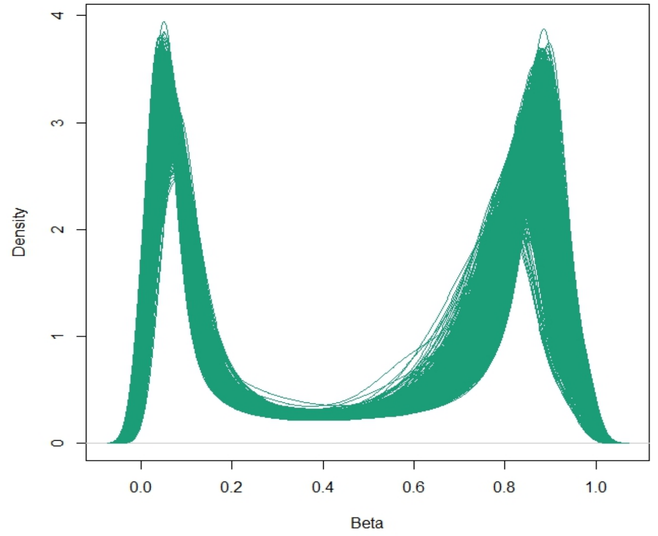




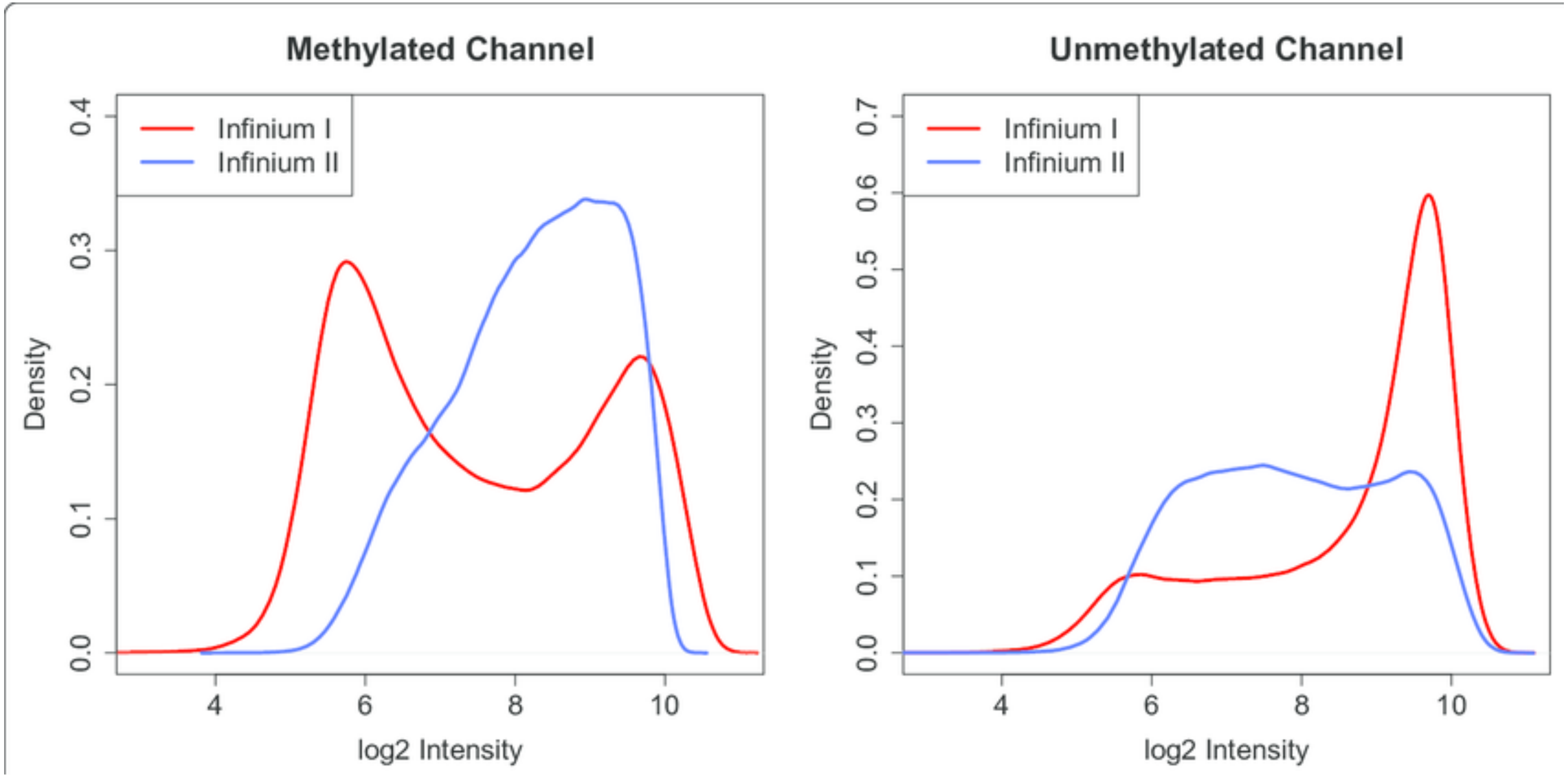




No QC/normalization all chromosomes (left), QC/normalization (right) on two different datasets



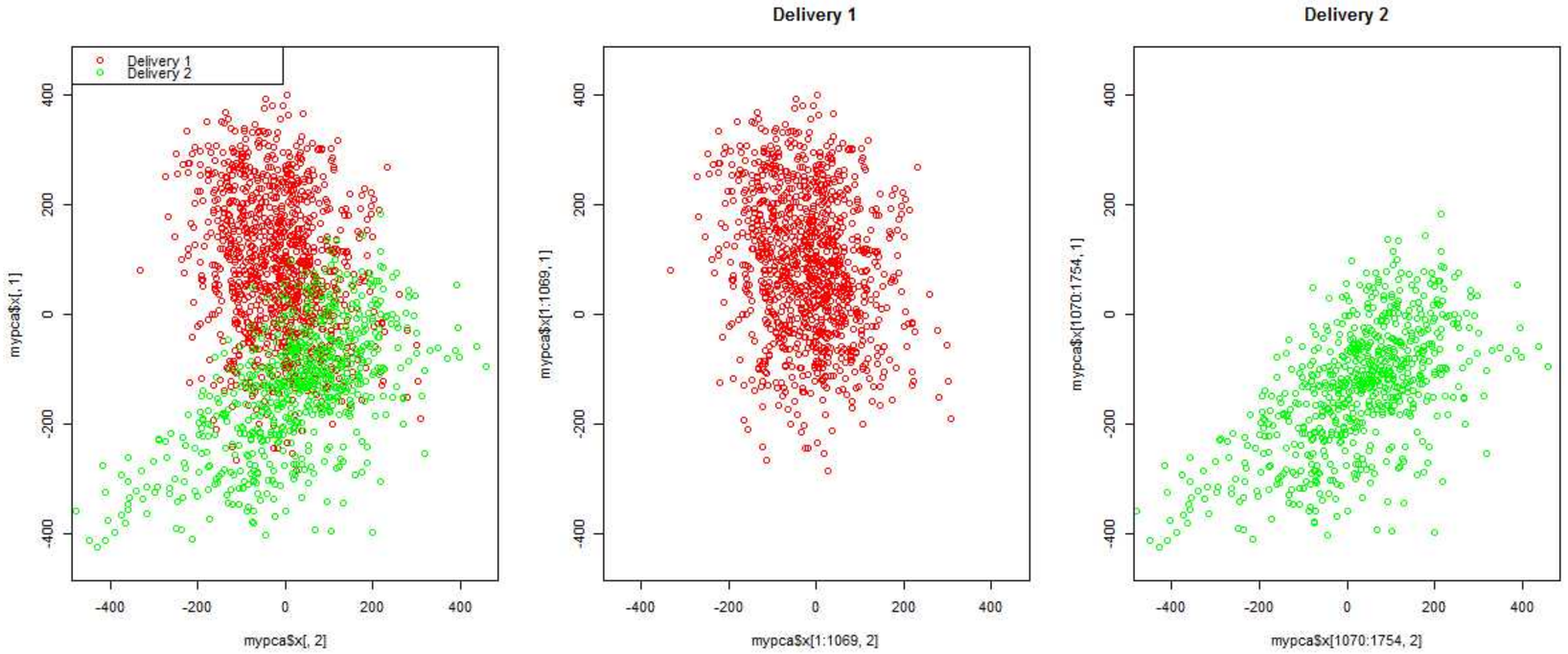



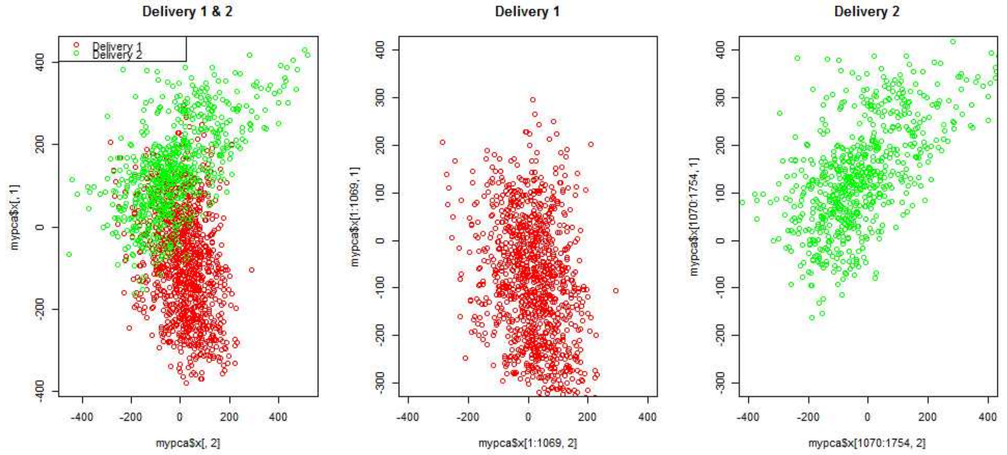
# Last but not least, probe correction



# Dataset (batch) correction

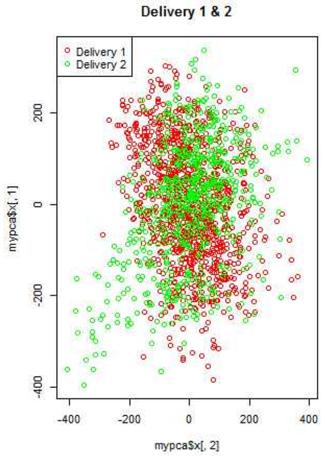
- Necessary when combining 2 or more datasets
- Colored wrt dataset (batch), 2 pictures,
- PCA of dataset 1 and dataset 2 before ComBat, all chromosomes

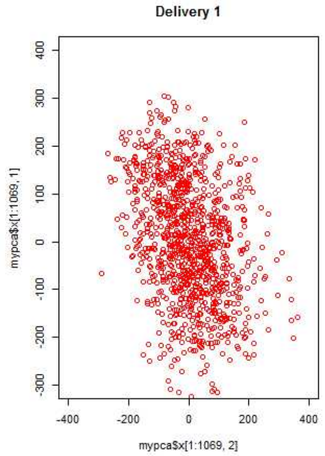


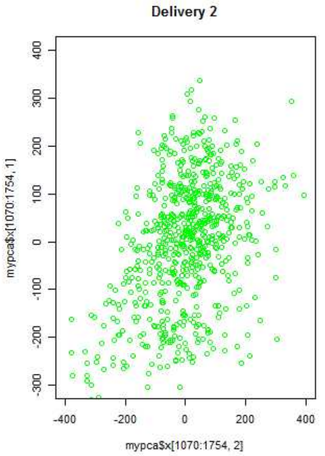

Delivery 1 & 2
Delivery 1
Delivery 2
mypca$x[, 1]
mypca$x[, 2]

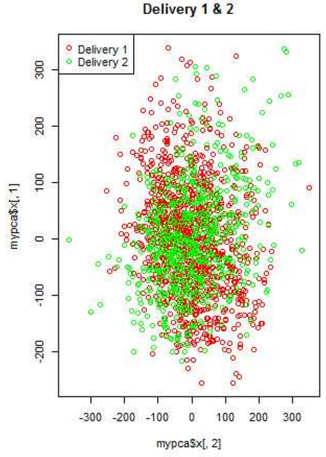

Delivery 1
mypca$x[1:1069, 1]
mypca$x[1:1069, 2]

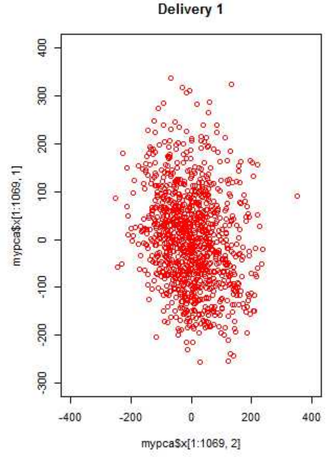

Delivery 2
mypca$x[1070:1754, 1]
mypca$x[1070:1754, 2]

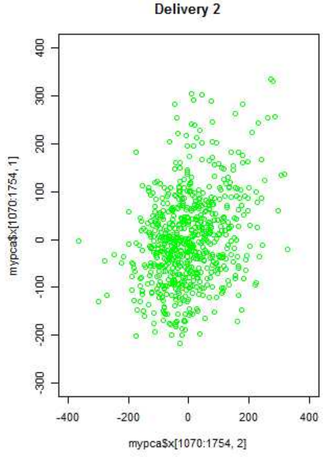

Delivery 1 & 2
Delivery 1
Delivery 2
mypca$x[, 1]
mypca$x[, 2]


Delivery 1
mypca$x[1:1069, 1]
mypca$x[1:1069, 2]


Delivery 2
mypca$x[1070:1754, 1]
mypca$x[1070:1754, 2]


Delivery 1 & 2
Delivery 1
Delivery 2
mypca$x[, 1]
mypca$x[, 2]


Delivery 1
mypca$x[1:1069, 1]
mypca$x[1:1069, 2]


Delivery 2
mypca$x[1070:1754, 1]
mypca$x[1070:1754, 2]

# 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*