Epigenomics

- Part 1: Intro to epigenetics
- **Part 2: High-throughput technologies**
- Part 3: Computational methods

Mark D. Robinson, Statistical Genomics, IMLS

Overview of this lecture

- You've seen microarrays and sequencing; here I discuss the epigenomic-specific assays that are upstream of these readouts
  - DNA methylation: enzymatic, chemical, enrichment/affinity capture
  - Sequencing versus microarray; high versus low resolution
  - Chromatin immunoprecipitation, ChIP-exo
  - DNaseI hypersensitivity, total/ribo-/polyA/micro RNA
  - 3C, HiC, etc.

DNA methylation

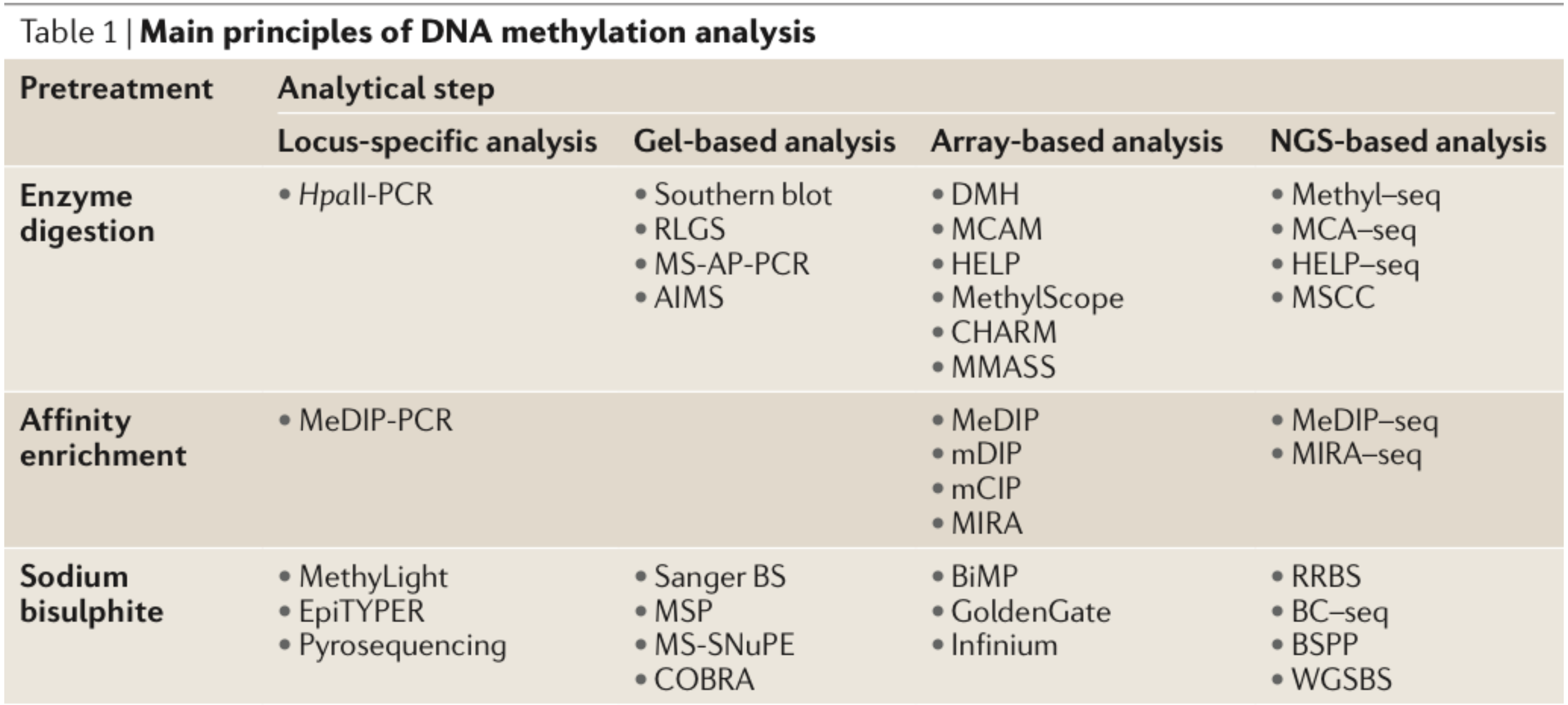


Table 1 | **Main principles of DNA methylation analysis**

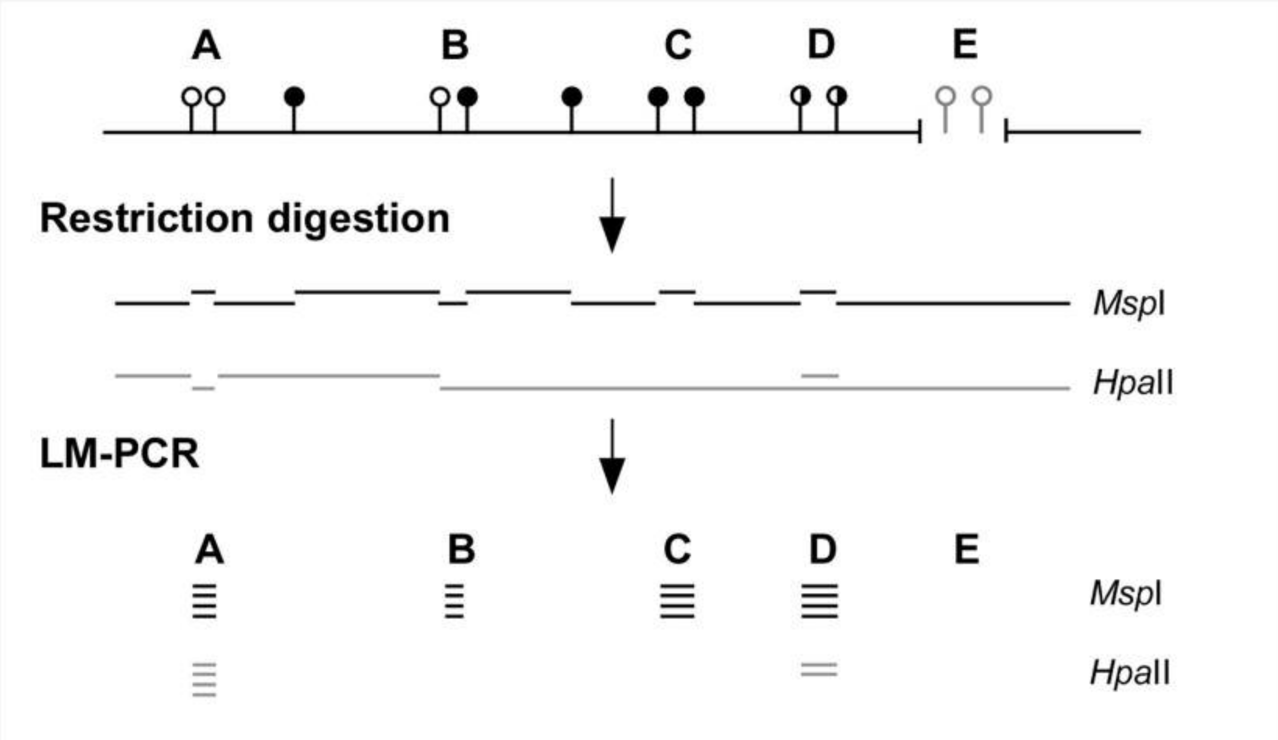
Pretreatment	Analytical step			
	Locus-specific analysis	Gel-based analysis	Array-based analysis	NGS-based analysis
Enzyme digestion	• *Hpa*II-PCR	• Southern blot • RLGS • MS-AP-PCR • AIMS	• DMH • MCAM • HELP • MethylScope • CHARM • MMASS	• Methyl–seq • MCA–seq • HELP–seq • MSCC
Affinity enrichment	• MeDIP-PCR		• MeDIP • mDIP • mCIP • MIRA	• MeDIP–seq • MIRA–seq
Sodium bisulphite	• MethyLight • EpiTYPER • Pyrosequencing	• Sanger BS • MSP • MS-SNuPE • COBRA	• BiMP • GoldenGate • Infinium	• RRBS • BC–seq • BSPP • WGSBS

Direct sequencing

Oxford Nanopore
Pacific Biosciences
etc.

Laird, Nature Reviews Genetics, March 2010

Enzyme digestion example



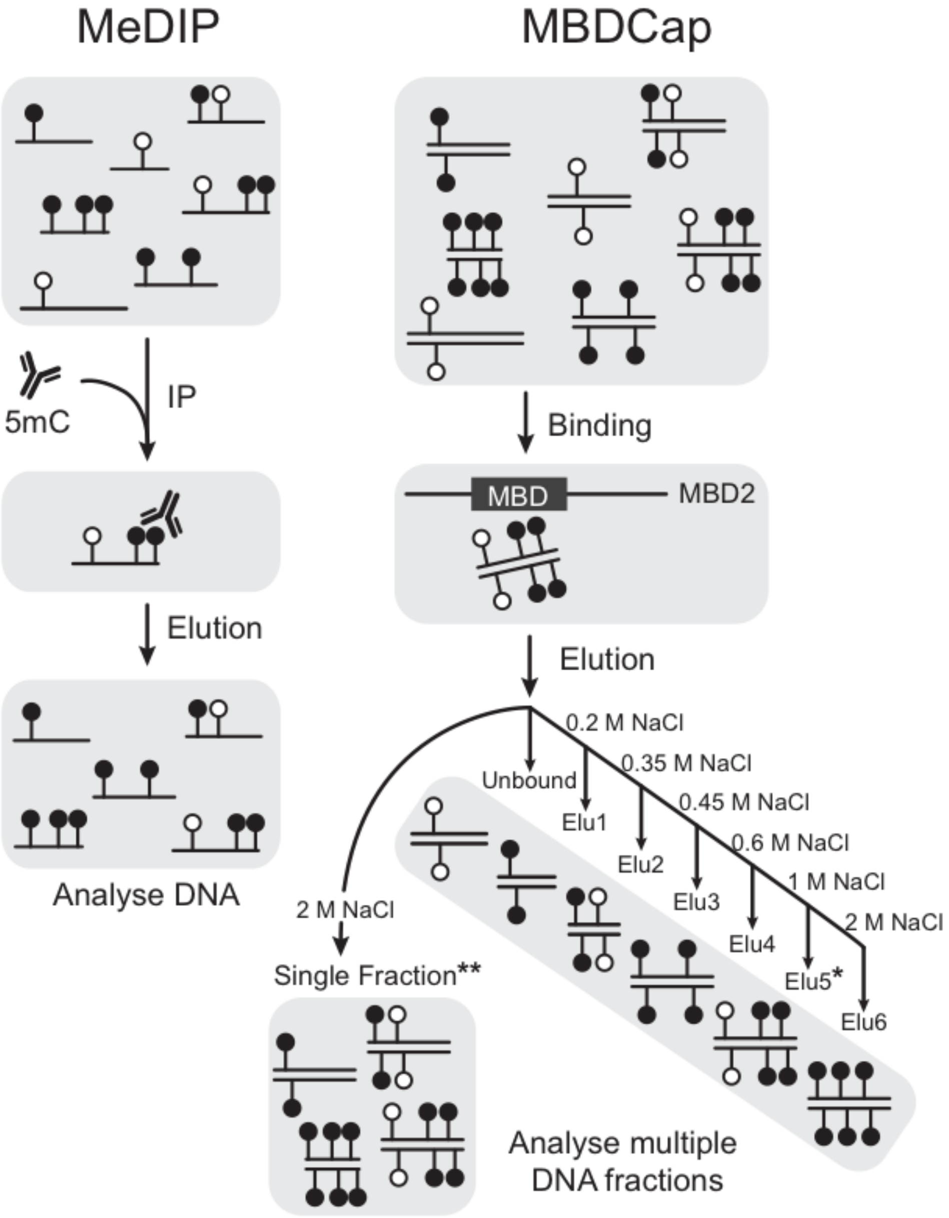
MspI – cuts at CCGG or CCGG sites

HpaII – cuts only at CCGG

CG - unmethylated

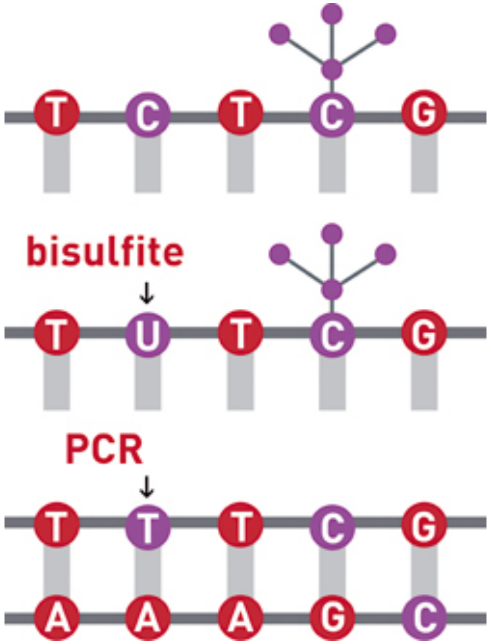
CG - methylated

Affinity capture of methylated DNA



Robinson et al. 2010

Bisulphite sequencing



Sodium bisulphite converts methylated **C**ytosine into **U**racil, which can be read as **T**hymine after PCR

In combination with sequencing (Sanger or NGS), can achieve methylation mapping at single base resolution

Can be nicely combined with genotyping arrays (e.g. Illumina HumanMethylation 450k)

http://www.diagenode.com/en/applications/bisulfite-conversion.php

Bisulphite conversion + “genotyping” array (Illumina HumanMethylaton450)

Unmethylated CpG site

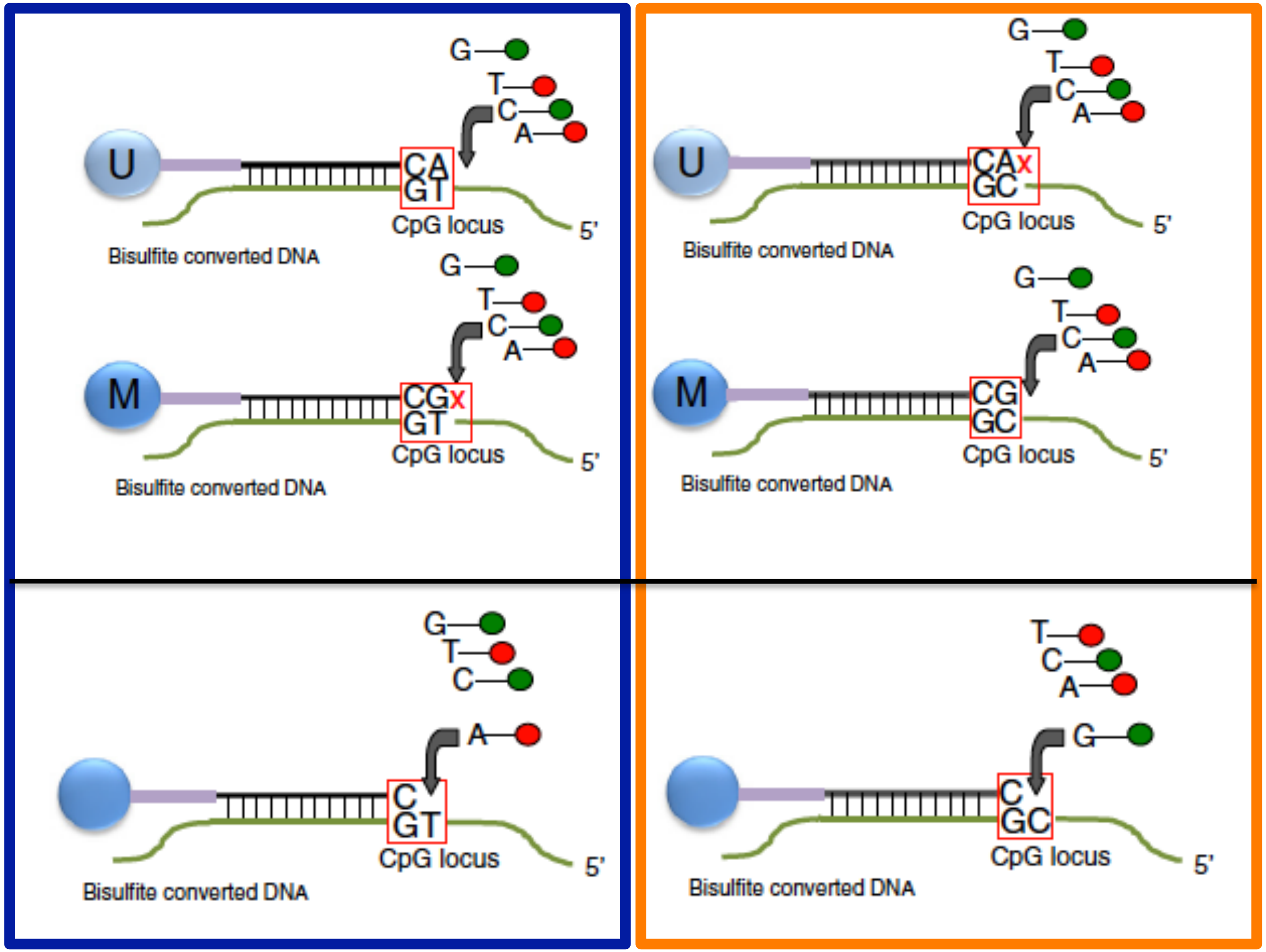
Methylated CpG site

Type I

(2 probes)

Type II

(1 probe)

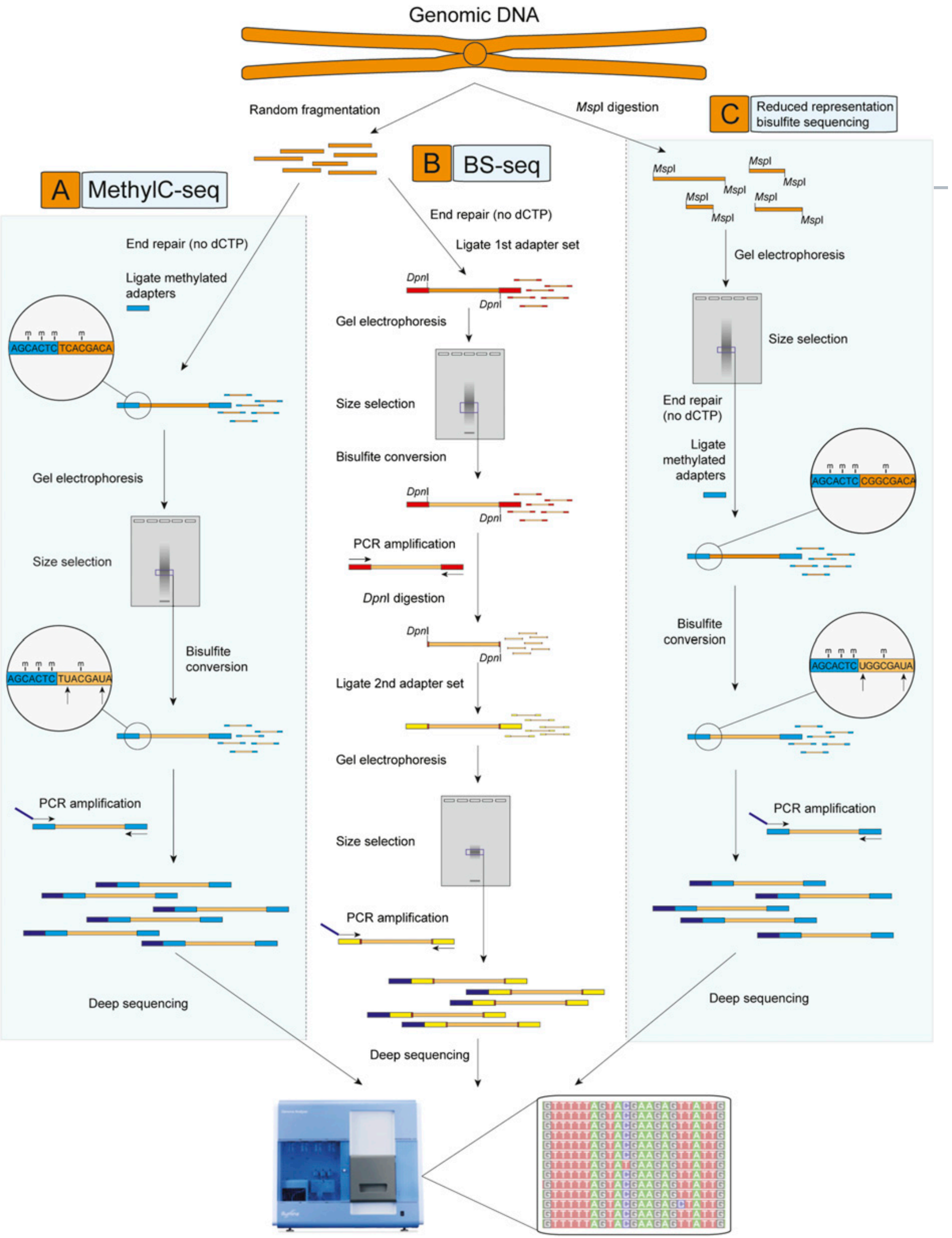


from Bibikova et al. Genomics 2011

University of Zurich^UZH

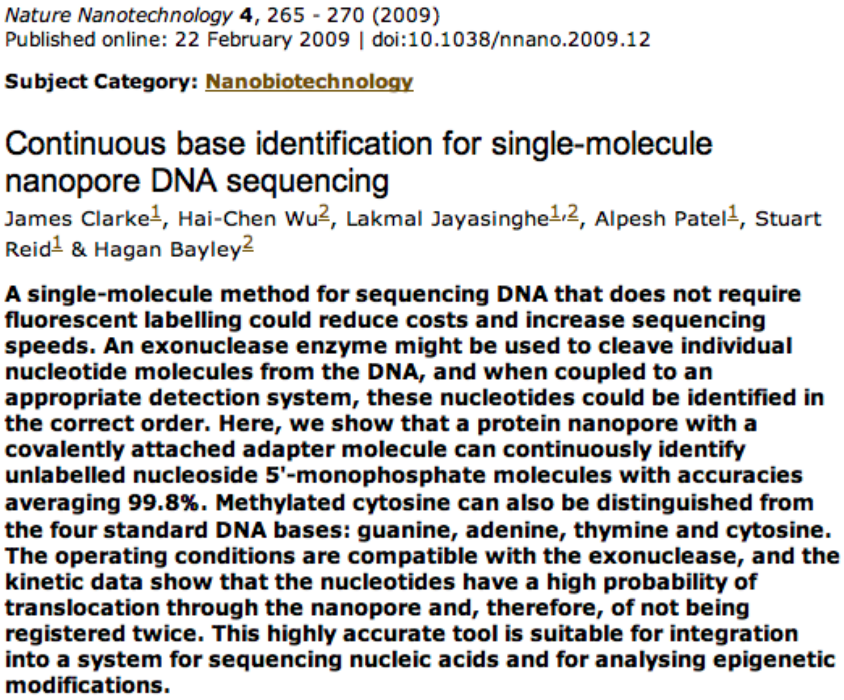
Institute of Molecular Life Sciences

DNAme methods that use bisulphite conversion with NGS



DNA methylation by direct sequencing

Oxford Nanopore, April 2009



Nature Nanotechnology **4**, 265 - 270 (2009)
Published online: 22 February 2009 | doi:10.1038/nnano.2009.12

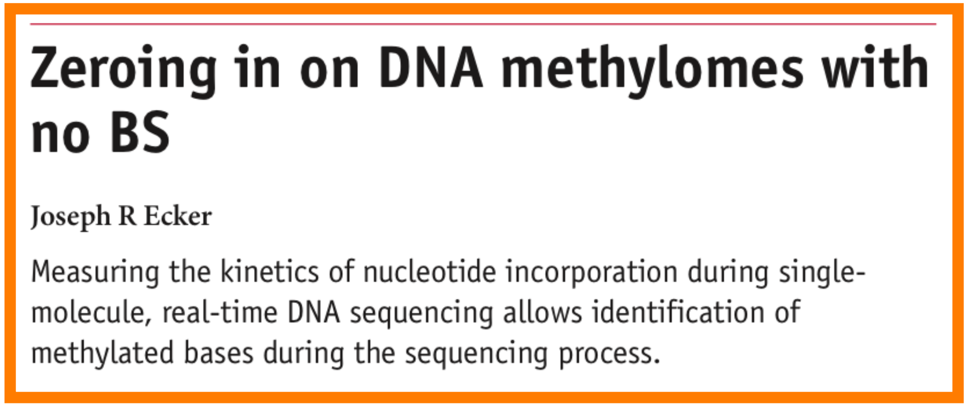
Subject Category: Nanobiotechnology

Continuous base identification for single-molecule nanopore DNA sequencing

James Clarke[1], Hai-Chen Wu[2], Lakmal Jayasinghe[1,2], Alpesh Patel[1], Stuart Reid[1] & Hagan Bayley[2]

A single-molecule method for sequencing DNA that does not require fluorescent labelling could reduce costs and increase sequencing speeds. An exonuclease enzyme might be used to cleave individual nucleotide molecules from the DNA, and when coupled to an appropriate detection system, these nucleotides could be identified in the correct order. Here, we show that a protein nanopore with a covalently attached adapter molecule can continuously identify unlabelled nucleoside 5'-monophosphate molecules with accuracies averaging 99.8%. Methylated cytosine can also be distinguished from the four standard DNA bases: guanine, adenine, thymine and cytosine. The operating conditions are compatible with the exonuclease, and the kinetic data show that the nucleotides have a high probability of translocation through the nanopore and, therefore, of not being registered twice. This highly accurate tool is suitable for integration into a system for sequencing nucleic acids and for analysing epigenetic modifications.

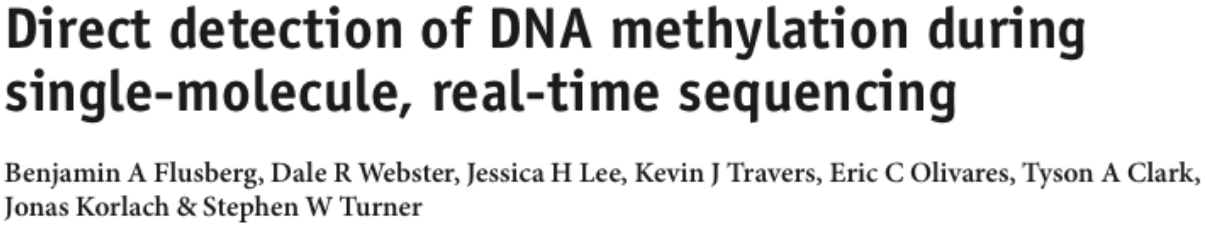
Nature Methods, 1st June 2010



Zeroing in on DNA methylomes with no BS

Joseph R Ecker

Measuring the kinetics of nucleotide incorporation during single-molecule, real-time DNA sequencing allows identification of methylated bases during the sequencing process.



Direct detection of DNA methylation during single-molecule, real-time sequencing

Benjamin A Flusberg, Dale R Webster, Jessica H Lee, Kevin J Travers, Eric C Olivares, Tyson A Clark, Jonas Korlach & Stephen W Turner

Pacific Biosciences, Nature Methods, June 2010

DNA methylation by direct sequencing (Pac Bio)

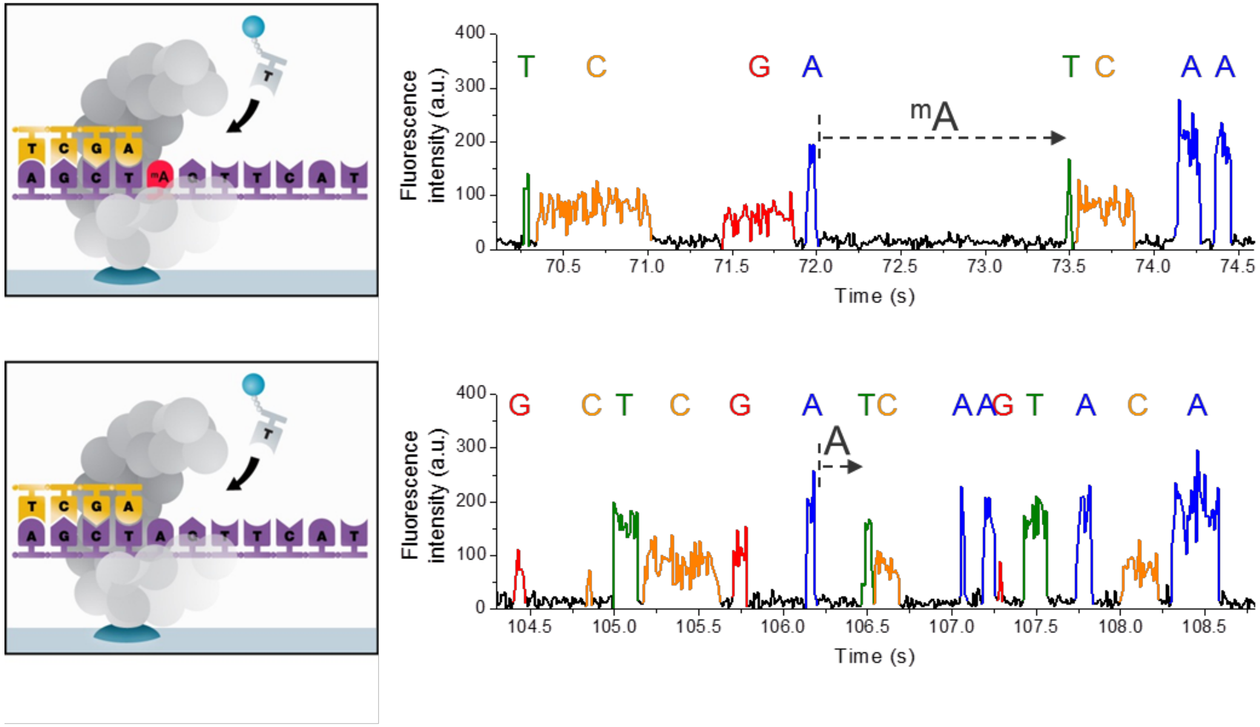
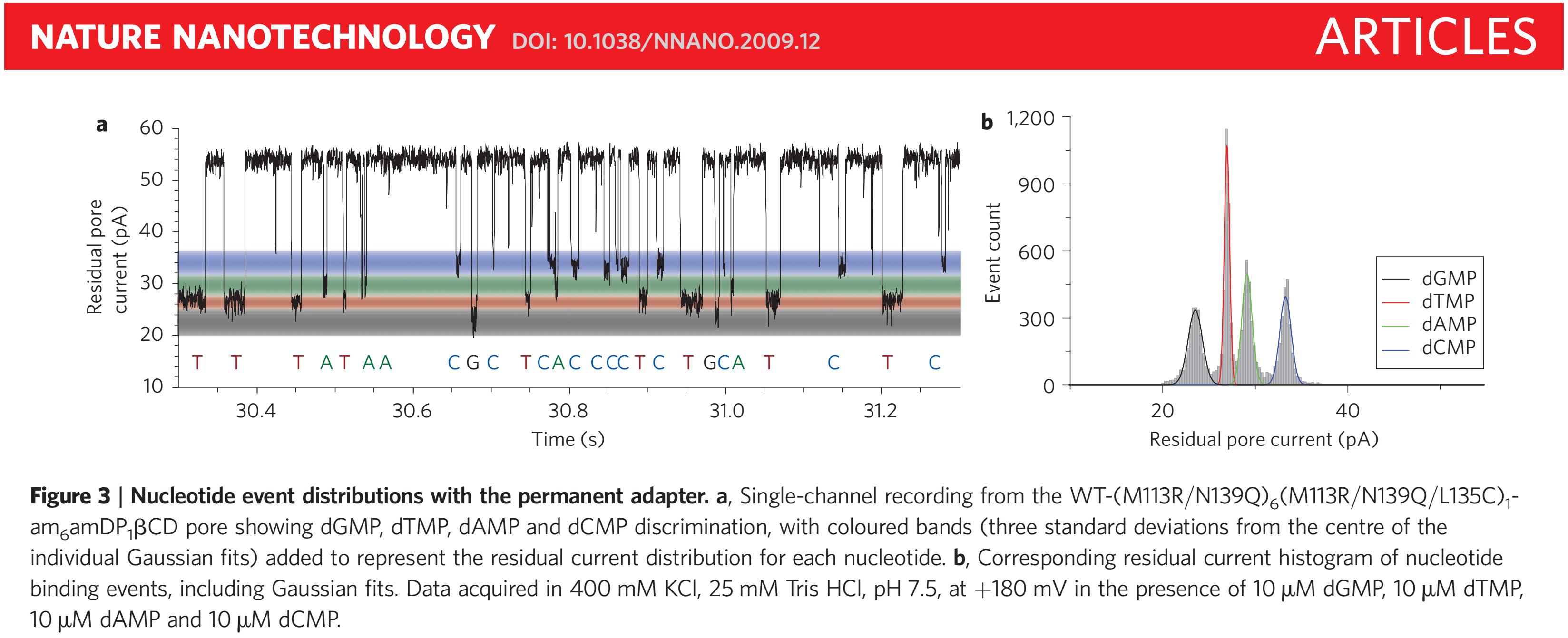


Figure 2. Principle of detecting modified DNA bases during SMRT sequencing. The presence of the modified base in the DNA template (top), shown here for 6-methyladenine, results in a delayed incorporation of the corresponding T nucleotide, i.e. longer interpulse duration (IPD), compared to a control DNA template lacking the modification (bottom).[3]

Pacific Biosciences white paper.

DNA methylation by direct sequencing (Oxford Nanopore)

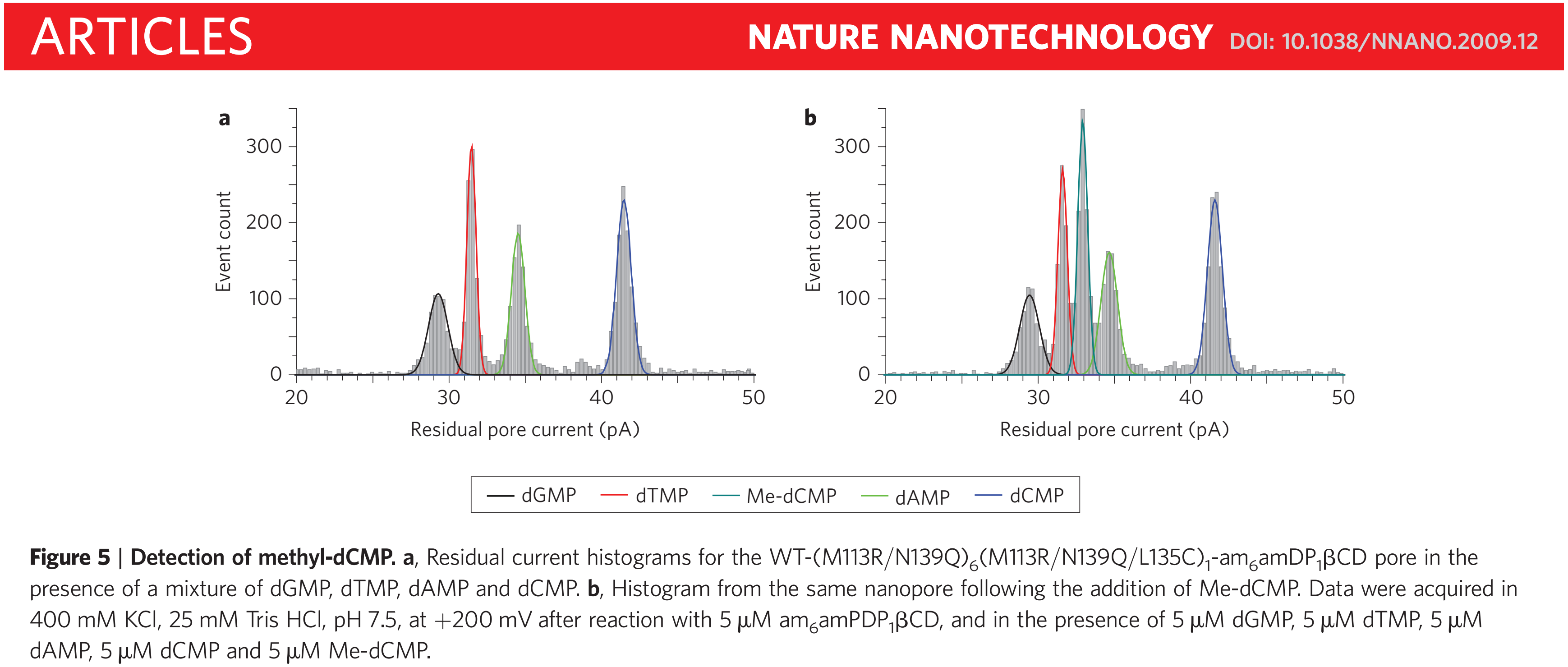


NATURE NANOTECHNOLOGY DOI: 10.1038/NNANO.2009.12 ARTICLES

Figure 3 | Nucleotide event distributions with the permanent adapter. a, Single-channel recording from the WT-(M113R/N139Q)$_6$(M113R/N139Q/L135C)$_1$-am$_6$amDP$_1$βCD pore showing dGMP, dTMP, dAMP and dCMP discrimination, with coloured bands (three standard deviations from the centre of the individual Gaussian fits) added to represent the residual current distribution for each nucleotide. **b**, Corresponding residual current histogram of nucleotide binding events, including Gaussian fits. Data acquired in 400 mM KCl, 25 mM Tris HCl, pH 7.5, at +180 mV in the presence of 10 μM dGMP, 10 μM dTMP, 10 μM dAMP and 10 μM dCMP.

Clarke et al. 2009 Nature Nano

DNA methylation by direct sequencing (Oxford Nanopore)



ARTICLES

NATURE NANOTECHNOLOGY DOI: 10.1038/NNANO.2009.12

Figure 5 | Detection of methyl-dCMP. a, Residual current histograms for the WT-(M113R/N139Q)$_6$(M113R/N139Q/L135C)$_1$-am$_6$amDP$_1$βCD pore in the presence of a mixture of dGMP, dTMP, dAMP and dCMP. **b**, Histogram from the same nanopore following the addition of Me-dCMP. Data were acquired in 400 mM KCl, 25 mM Tris HCl, pH 7.5, at +200 mV after reaction with 5 μM am$_6$amPDP$_1$βCD, and in the presence of 5 μM dGMP, 5 μM dTMP, 5 μM dAMP, 5 μM dCMP and 5 μM Me-dCMP.

Other remarks into DNA methylation data

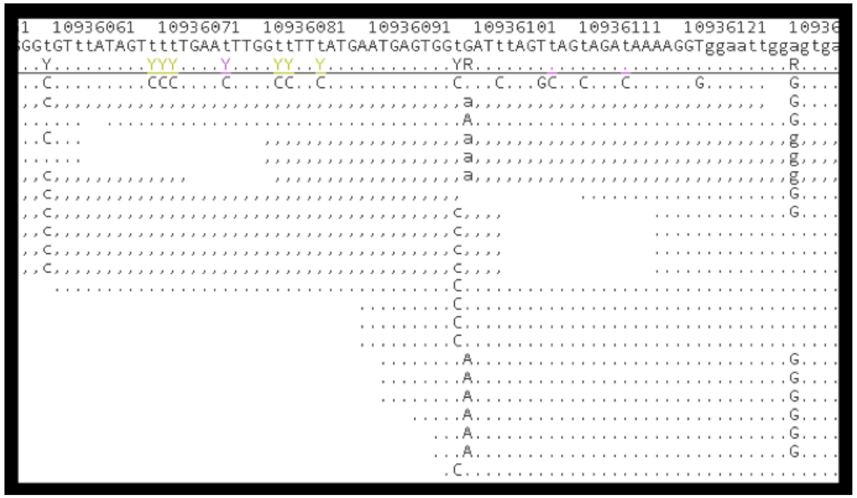
- Whole genome bisulphite sequencing is the most accurate, but expensive and somewhat inefficient
- Performance of affinity capture can vary drastically according to exact specifications of the protocol
- Difficult to compare methods since platforms have different coverage, different resolution

DNAme readouts can be low or high resolution

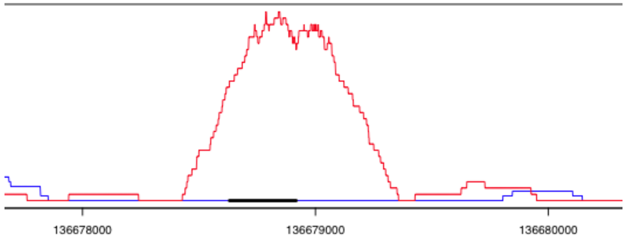
Sequencing: depth of converted reads versus total depth

Microarray: relative intensity of M and U probes

Sequencing: Pileup of reads

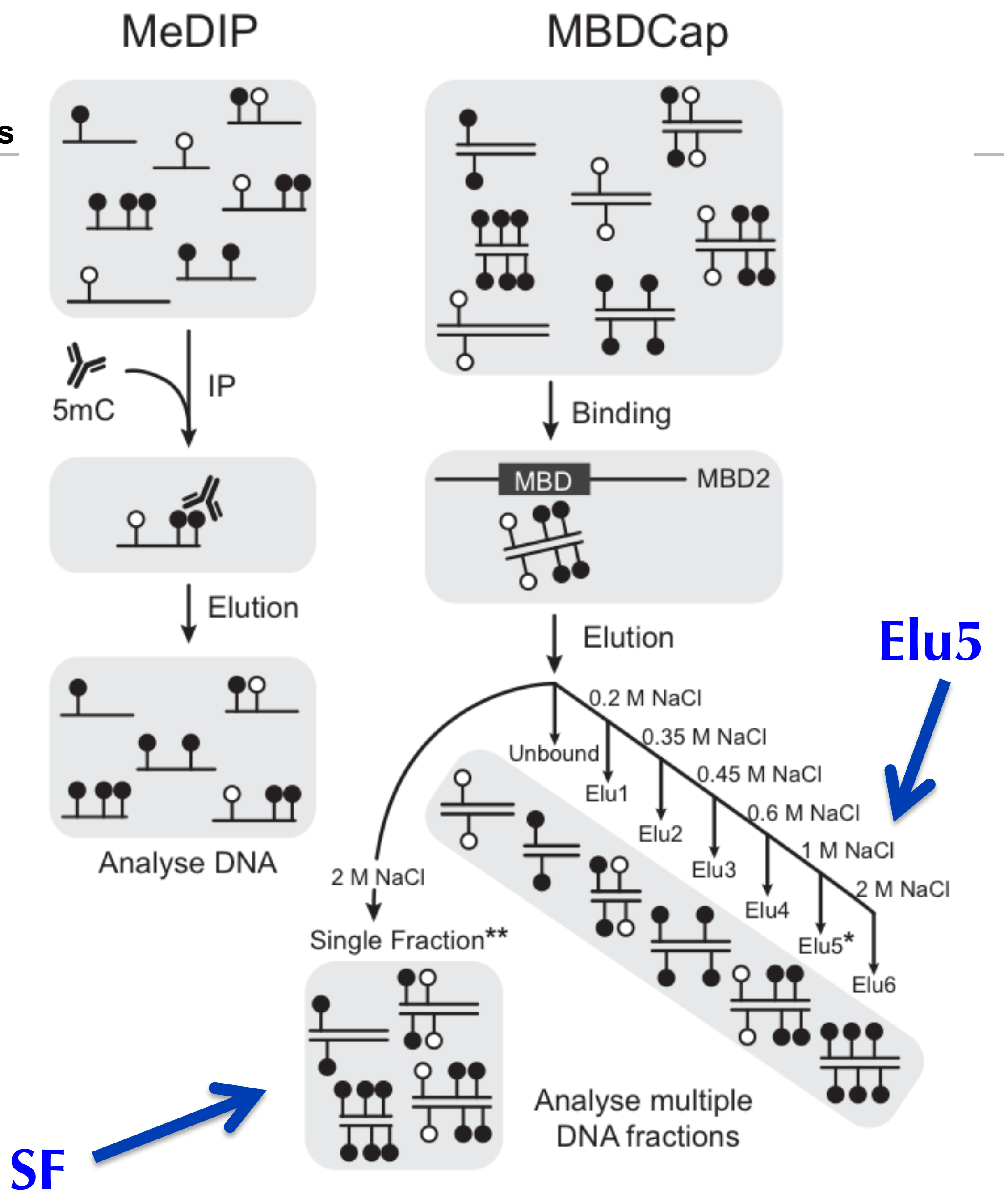


```
1  10936061  10936071  10936081  10936091  10936101  10936111  10936121  10936
GGtGTttATAGTtttTGAAtTTGGttTTtATGAATGAGTGGtGATttAGTtAGtAGAtAAAAGGTggaattggagtga
..Y.........YYY....Y....YY..Y............YR...............................R...
..C.........CCC....C....CC..C............C...C...GC..C...C......G......   G...
,,c,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,a,,,,,,,,,,,,,,,,,,,,,,,,,,,,   G...
......  ..................................A...............................G...
..C...                 ,,,,,,,,,,,,,,,,,,,a,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,g,,,
......                 ,,,,,,,,,,,,,,,,,,,a,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,g,,,
,,c,,,,,,,,,,,,,        ,,,,,,,,,,,,,,,,,,a,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,g,,,
,,c,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,           .....................G...
,,c,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,c,,,,              ..............G...
,,c,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,c,,,,              ..................
,,c,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,c,,,,              ..................
,,c,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,c,,,,              ..................
   ......................................C....................................
                                .........C....................................
                                .........C....................................
                                .........C....................................
                                  ........A...............................G...
                                  ........A...............................G...
                                  ........A...............................G...
                                     .....A...............................G...
                                       ...A...............................G...
                                       ...A...............................G...
                                        .C....................................
```

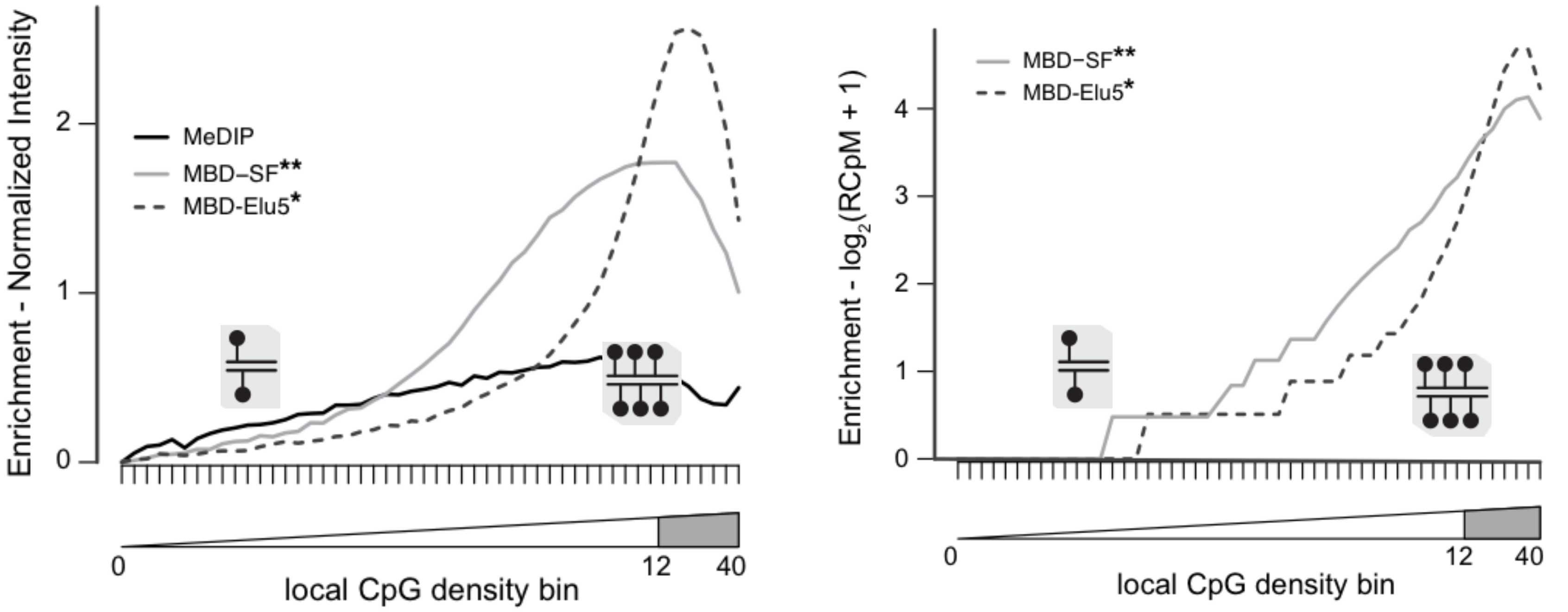


MeDIP versus MBDCap

- •Single stranded versus double stranded
- •MBDCap – elution series by salt gradient

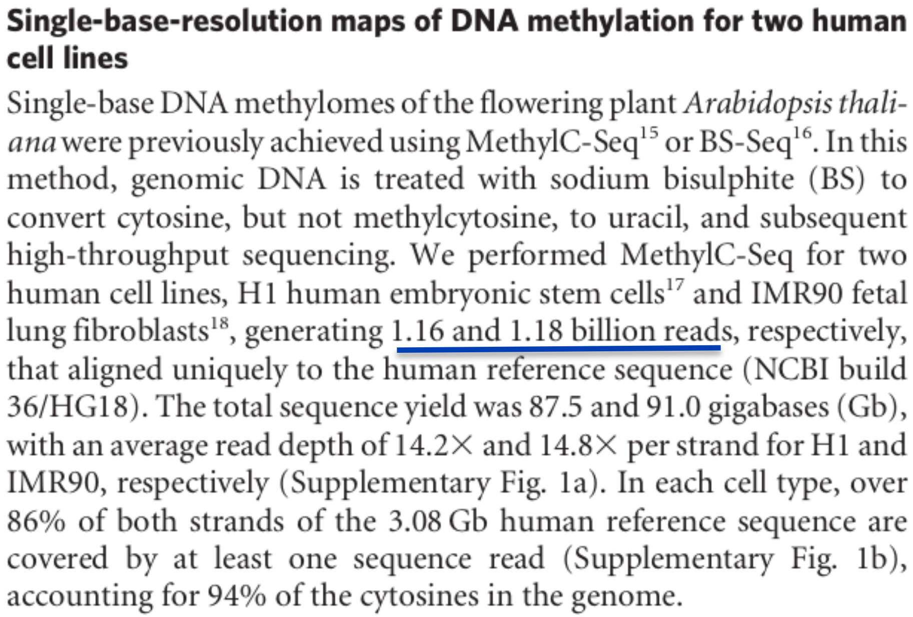


Strength of affinity enrichment is associated with CpG density



Robinson et al. Genome Research 2010

Whole genome BS sequencing can be inefficient



Single-base-resolution maps of DNA methylation for two human cell lines

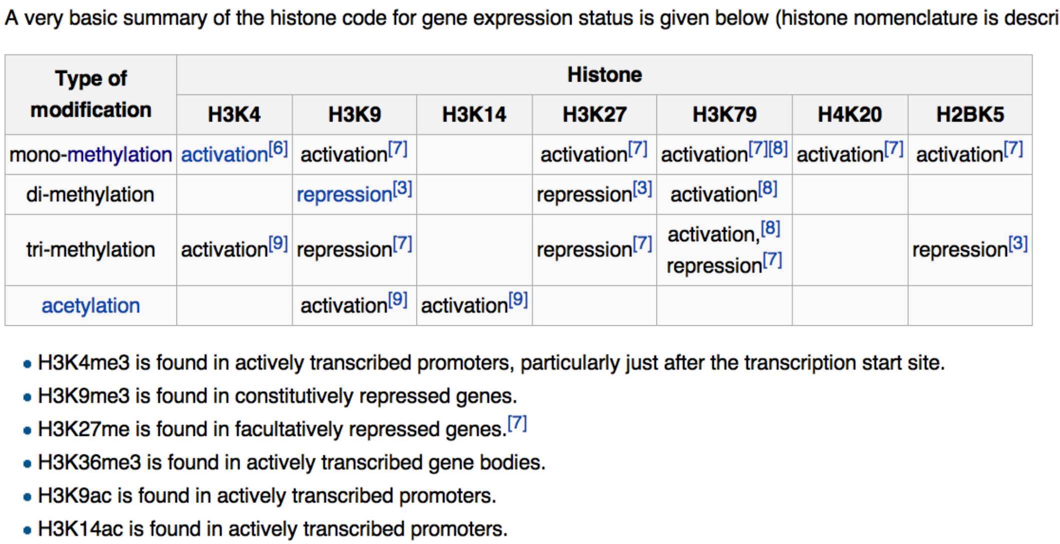
Single-base DNA methylomes of the flowering plant *Arabidopsis thaliana* were previously achieved using MethylC-Seq[15] or BS-Seq[16]. In this method, genomic DNA is treated with sodium bisulphite (BS) to convert cytosine, but not methylcytosine, to uracil, and subsequent high-throughput sequencing. We performed MethylC-Seq for two human cell lines, H1 human embryonic stem cells[17] and IMR90 fetal lung fibroblasts[18], generating 1.16 and 1.18 billion reads, respectively, that aligned uniquely to the human reference sequence (NCBI build 36/HG18). The total sequence yield was 87.5 and 91.0 gigabases (Gb), with an average read depth of 14.2× and 14.8× per strand for H1 and IMR90, respectively (Supplementary Fig. 1a). In each cell type, over 86% of both strands of the 3.08 Gb human reference sequence are covered by at least one sequence read (Supplementary Fig. 1b), accounting for 94% of the cytosines in the genome.

Lister et al. 2009, Nature

Notes re: WGSBS:

1. Mapping is done on BS-converted reads/genome (i.e.3 bases), requires mapping separately to each strand – need longer (paired) reads and high coverage
2. Of the 1.18B reads, approximately 670M (56%) do NOT overlap a CpG site
3. There may be a fair amount of regions that are completely unmethylated

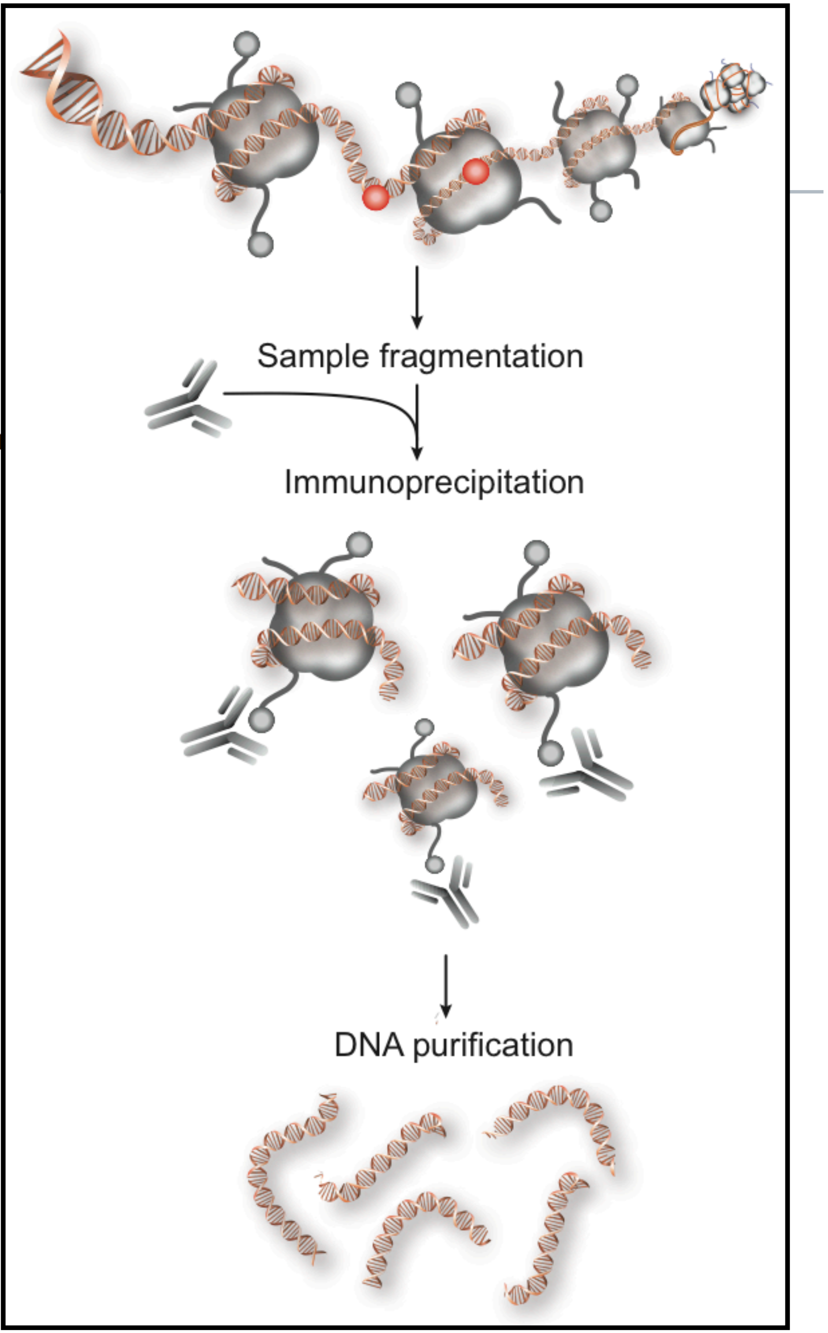
Chromatin immunoprecipitation for protein-DNA interactions



A very basic summary of the histone code for gene expression status is given below (histone nomenclature is descri

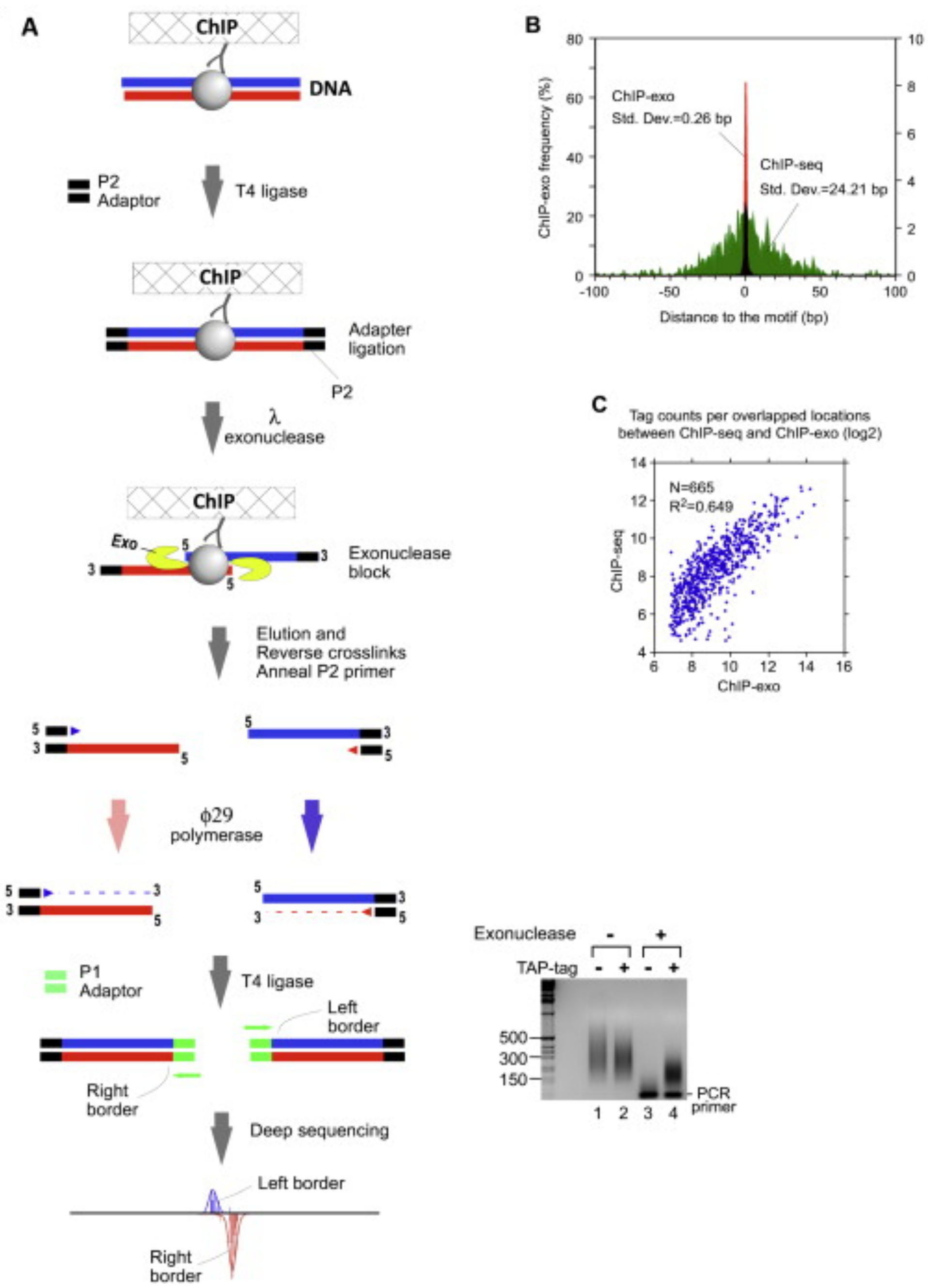
Type of modification	Histone						
	H3K4	H3K9	H3K14	H3K27	H3K79	H4K20	H2BK5
mono-methylation	activation[6]	activation[7]		activation[7]	activation[7][8]	activation[7]	activation[7]
di-methylation		repression[3]		repression[3]	activation[8]		
tri-methylation	activation[9]	repression[7]		repression[7]	activation,[8] repression[7]		repression[3]
acetylation		activation[9]	activation[9]				

- H3K4me3 is found in actively transcribed promoters, particularly just after the transcription start site.
- H3K9me3 is found in constitutively repressed genes.
- H3K27me is found in facultatively repressed genes.[7]
- H3K36me3 is found in actively transcribed gene bodies.
- H3K9ac is found in actively transcribed promoters.
- H3K14ac is found in actively transcribed promoters.

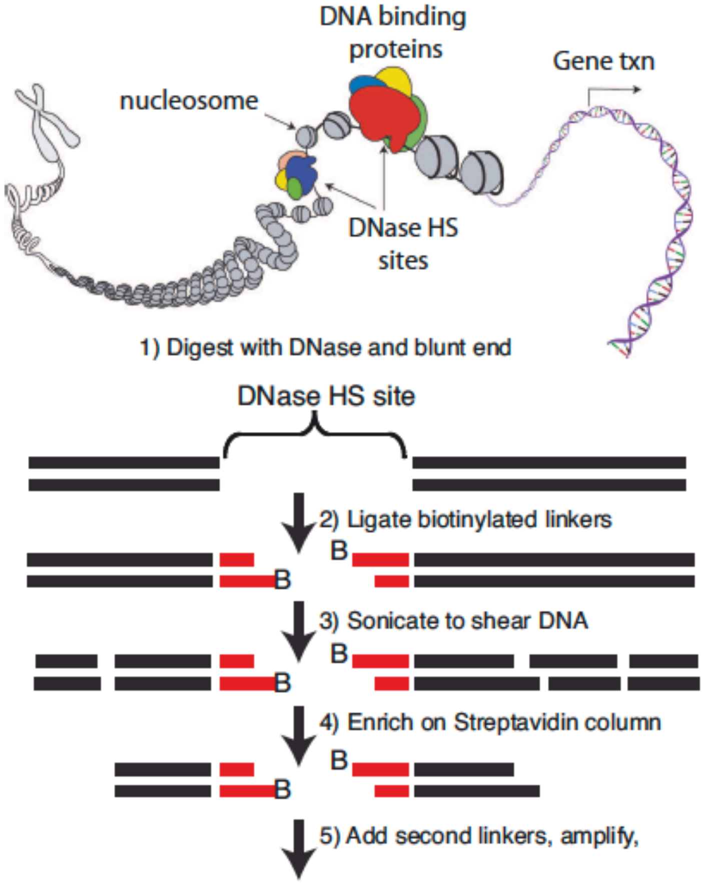


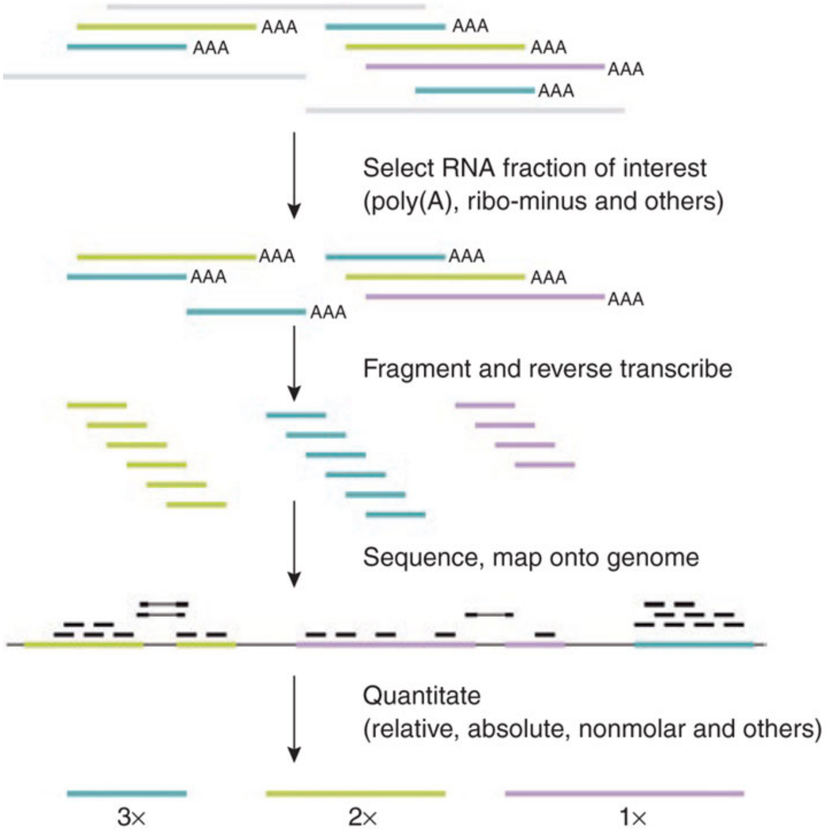
ChIP-exo

ChIP DNA is treated with a 5′ to 3′ exonuclease while still present within the immunoprecipitate.

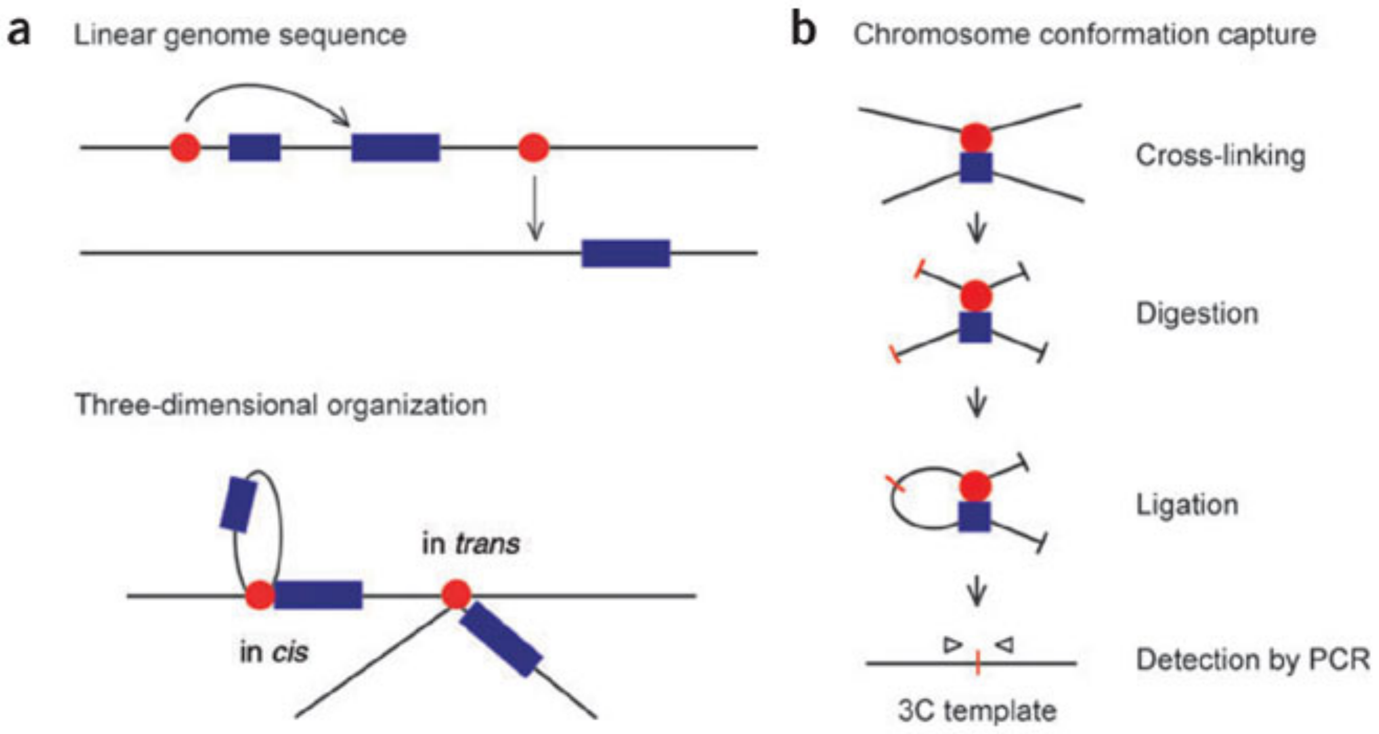


Techniques: DNaseI, RNA-seq


DNA binding proteins
Gene txn
nucleosome
DNase HS sites
1) Digest with DNase and blunt end
DNase HS site
2) Ligate biotinylated linkers
B
B
3) Sonicate to shear DNA
B
B
4) Enrich on Streptavidin column
B
B
5) Add second linkers, amplify,

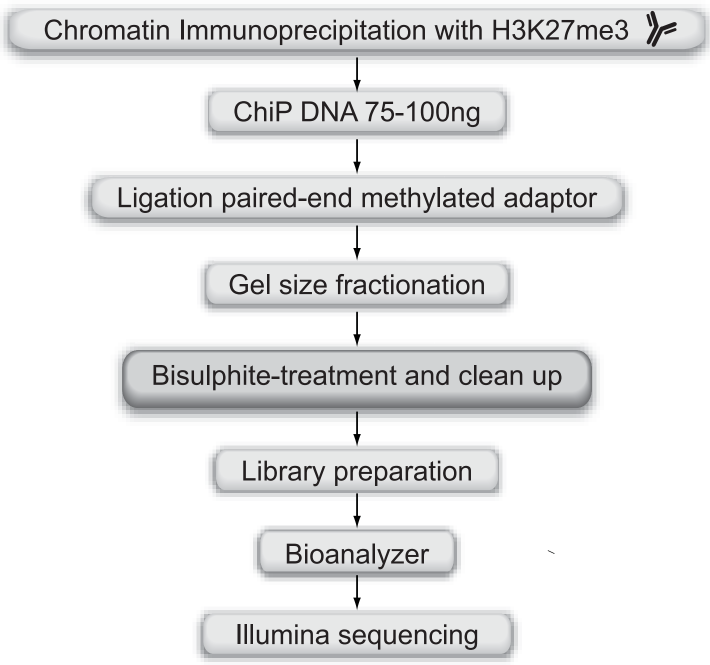

AAA
AAA
AAA
AAA
AAA
AAA
Select RNA fraction of interest
(poly(A), ribo-minus and others)
AAA
AAA
AAA
AAA
AAA
AAA
Fragment and reverse transcribe
Sequence, map onto genome
Quantitate
(relative, absolute, nonmolar and others)
3×
2×
1×

Higher-order chromatin structure



Combinations: ChIP-BS-seq

A few tricks on the technical side to facilitate this.



Chromatin Immunoprecipitation with H3K27me3

ChiP DNA 75-100ng

Ligation paired-end methylated adaptor

Gel size fractionation

Bisulphite-treatment and clean up

Library preparation

Bioanalyzer

Illumina sequencing

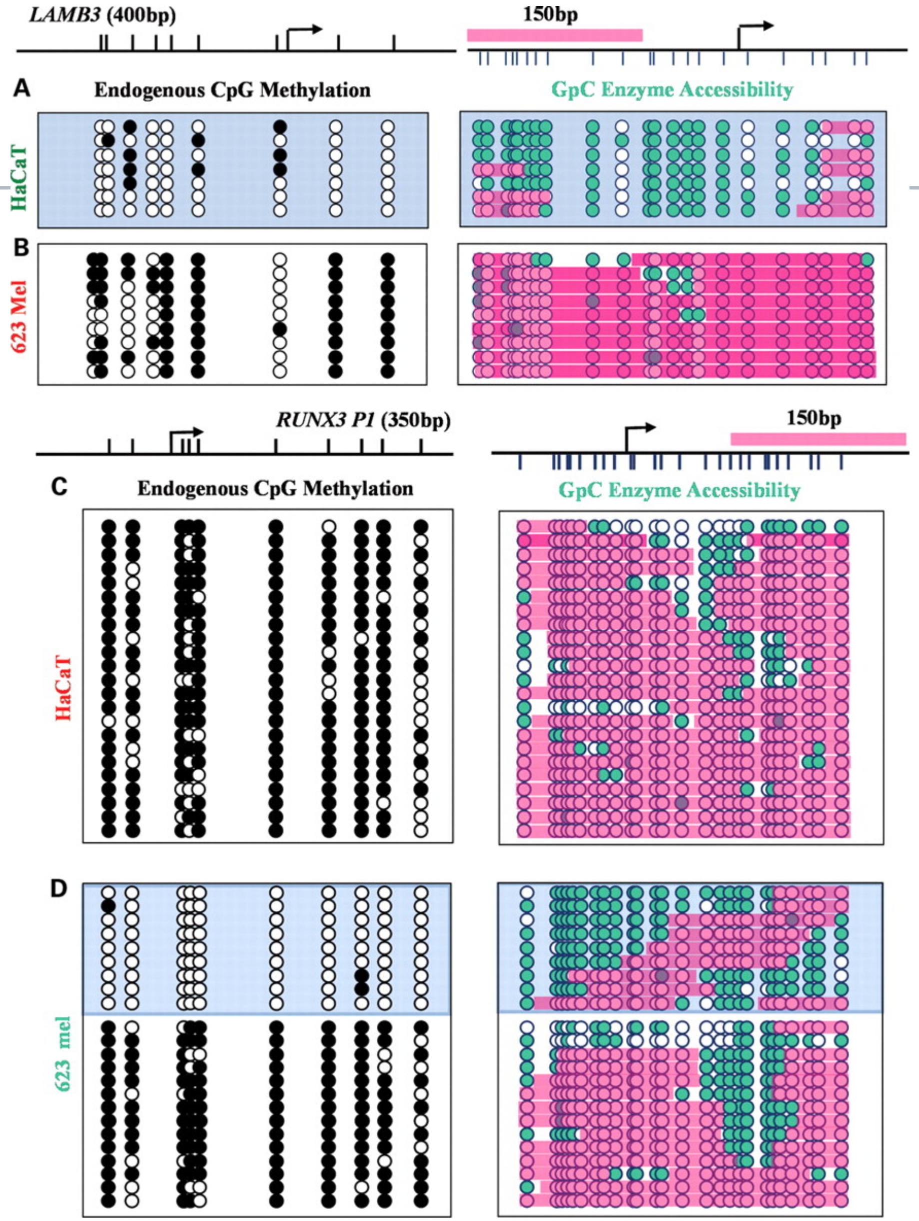
Statham*, Robinson* et al., Genome Research, 2012

Combinations: NOME-seq

M.CviPI enzyme is used to methylate GpC sites **not bound by nucleosomes**

Both GpC methylation and CpG methylation can be readout (on the same clone) after bisulphite treatment

Pink: nucleosome-bound (not methylated by M.CviPI)
Green: accessible



Remarks: Allele-specific epigenetics, cell populations

- A couple key points to recognize:
  - Typically, MBD-seq/ChIP-seq/etc. are analyzing populations of cells (e.g. patient tumours that may contain normal cell types as well) – so we are really studying the population average! So called “bulk analysis”
  - In some instances, we may be able to combine the information we get from genome sequencing (e.g. SNPs) to partition transcription and epigenetic factors by allele

Technical limitation in the amount of DNA need to create library and sequence

- We often want to know about several factors on a single population of cells – requires a lot of DNA/RNA
- New technologies (e.g. sequencing small amounts / amplification) are trying to address this
- Patient (e.g. tumour sample) cell population purity?

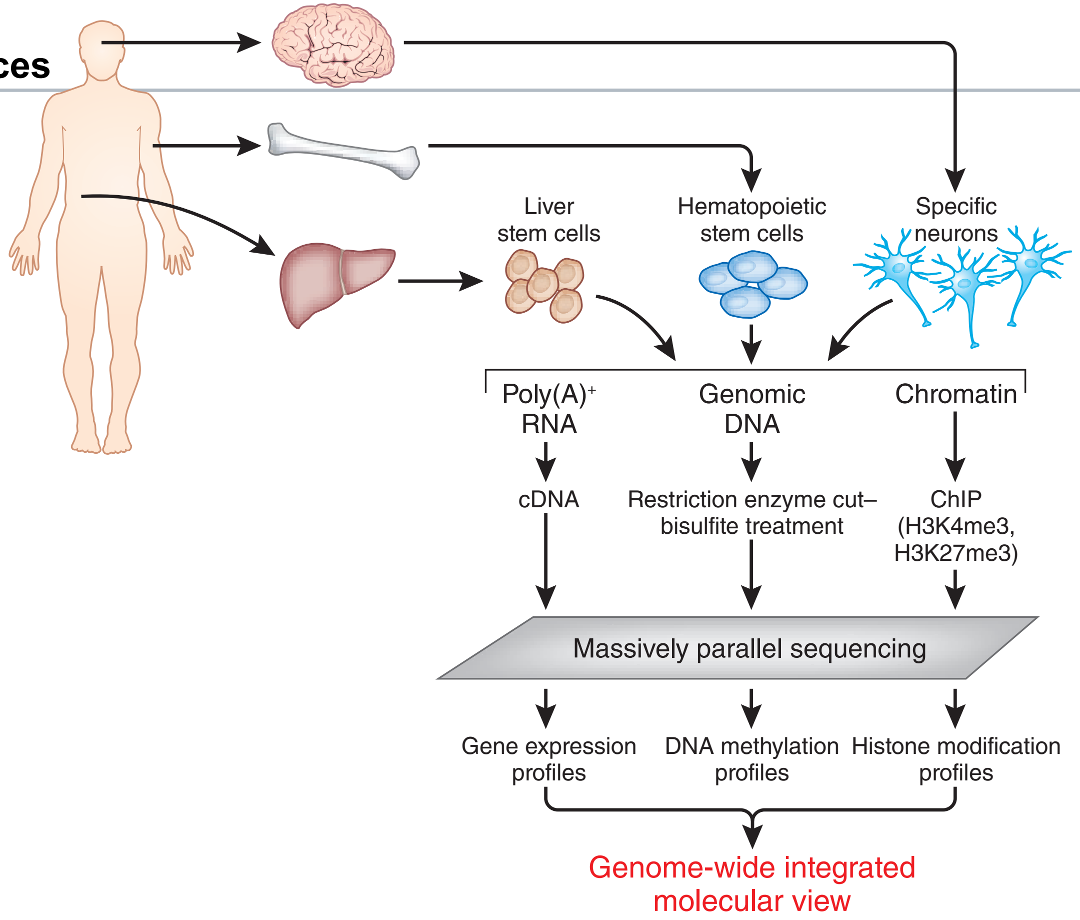
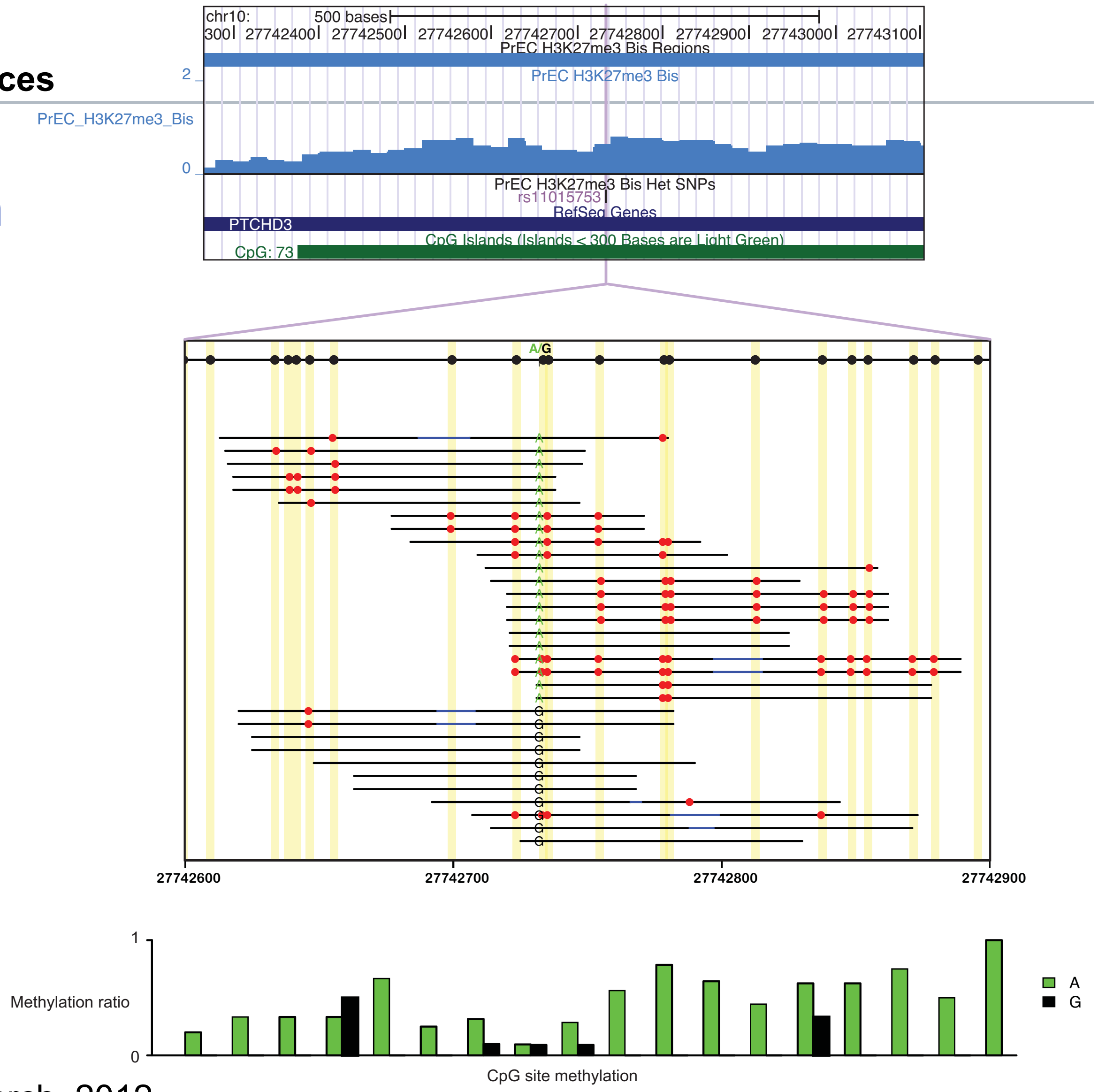


Figure 1 | Schematic flow chart of experimental design. Rare cell types are isolated from specific organs and used for RNA and DNA preparation, and ChIP. Combining gene expression, DNA methylation and histone modification profiles gives an integrated view of the epigenome.

Allele-specific methylation

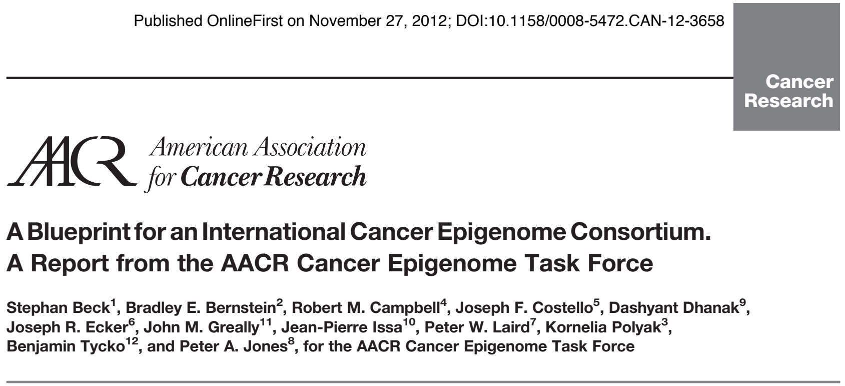
- Biologically, what affect does this have?
- How prominent is this?



Statham*, Robinson* et al., Genome Research, 2012

Era of big data is upon us

- ENCODE - Encyclopedia Of DNA Elements ("to identify all functional elements in the human genome sequence") – [Funny aside next slide]
- BLUEPRINT – "apply highly sophisticated functional genomics analysis on a clearly defined set of primarily human samples from healthy and diseased individuals and to provide at least 100 reference epigenomes to the scientific community"
- IHEC – "aims to coordinate epigenome mapping for a broad spectrum of human cell types and a wide range of developmental stages."
- ICGC – "To obtain a comprehensive description of genomic, transcriptomic and epigenomic changes in 50 different tumor types and/or subtypes which are of clinical and societal importance across the globe"
- TCGA – "systematically explore the entire spectrum of genomic changes involved in more than 20 types of human cancer."
- Nucleosome4D/4DCellFate



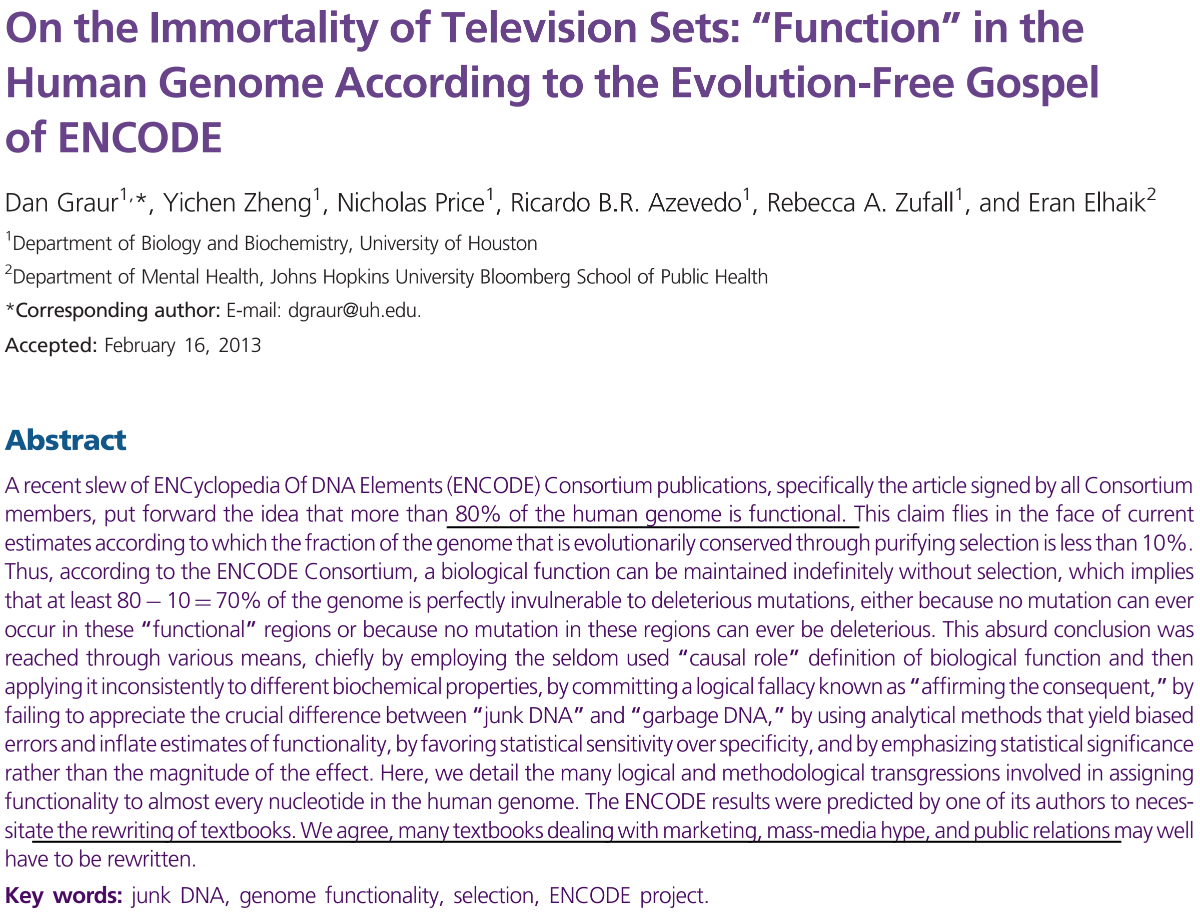
Published OnlineFirst on November 27, 2012; DOI:10.1158/0008-5472.CAN-12-3658

Cancer Research

AACR American Association for Cancer Research

A Blueprint for an International Cancer Epigenome Consortium. A Report from the AACR Cancer Epigenome Task Force

Stephan Beck[1], Bradley E. Bernstein[2], Robert M. Campbell[4], Joseph F. Costello[5], Dashyant Dhanak[9], Joseph R. Ecker[6], John M. Greally[11], Jean-Pierre Issa[10], Peter W. Laird[7], Kornelia Polyak[3], Benjamin Tycko[12], and Peter A. Jones[8], for the AACR Cancer Epigenome Task Force



On the Immortality of Television Sets: “Function” in the Human Genome According to the Evolution-Free Gospel of ENCODE

Dan Graur[1,*], Yichen Zheng[1], Nicholas Price[1], Ricardo B.R. Azevedo[1], Rebecca A. Zufall[1], and Eran Elhaik[2]

[1]Department of Biology and Biochemistry, University of Houston

[2]Department of Mental Health, Johns Hopkins University Bloomberg School of Public Health

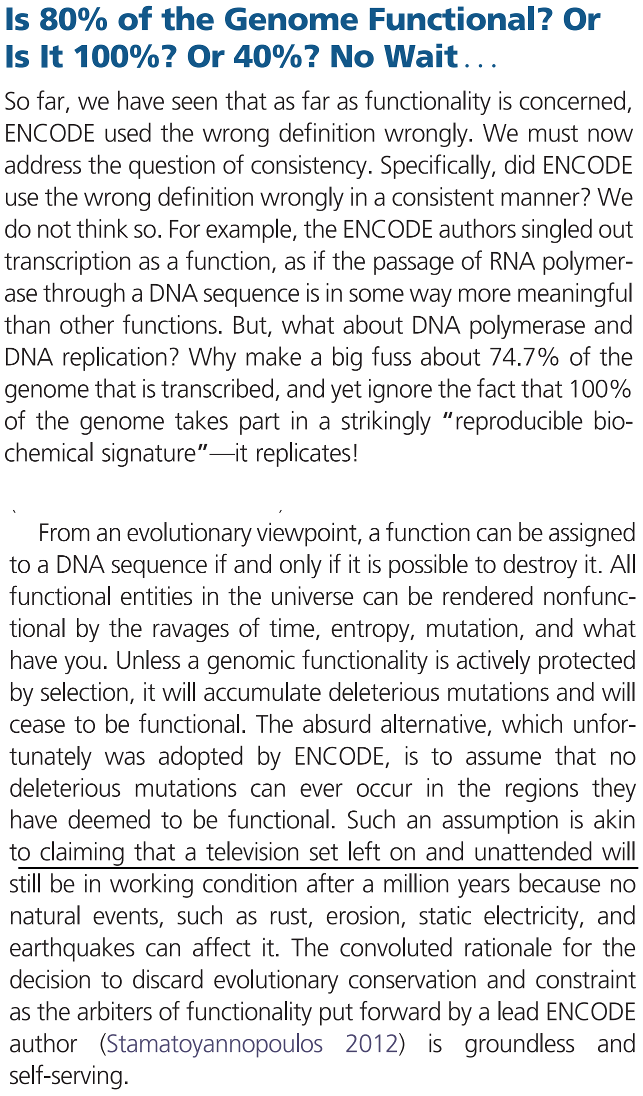
*Corresponding author: E-mail: dgraur@uh.edu.

Accepted: February 16, 2013

Abstract

A recent slew of ENCyclopedia Of DNA Elements (ENCODE) Consortium publications, specifically the article signed by all Consortium members, put forward the idea that more than 80% of the human genome is functional. This claim flies in the face of current estimates according to which the fraction of the genome that is evolutionarily conserved through purifying selection is less than 10%. Thus, according to the ENCODE Consortium, a biological function can be maintained indefinitely without selection, which implies that at least $80 - 10 = 70\%$ of the genome is perfectly invulnerable to deleterious mutations, either because no mutation can ever occur in these “functional” regions or because no mutation in these regions can ever be deleterious. This absurd conclusion was reached through various means, chiefly by employing the seldom used “causal role” definition of biological function and then applying it inconsistently to different biochemical properties, by committing a logical fallacy known as “affirming the consequent,” by failing to appreciate the crucial difference between “junk DNA” and “garbage DNA,” by using analytical methods that yield biased errors and inflate estimates of functionality, by favoring statistical sensitivity over specificity, and by emphasizing statistical significance rather than the magnitude of the effect. Here, we detail the many logical and methodological transgressions involved in assigning functionality to almost every nucleotide in the human genome. The ENCODE results were predicted by one of its authors to necessitate the rewriting of textbooks. We agree, many textbooks dealing with marketing, mass-media hype, and public relations may well have to be rewritten.

Key words: junk DNA, genome functionality, selection, ENCODE project.



Is 80% of the Genome Functional? Or Is It 100%? Or 40%? No Wait . . .

So far, we have seen that as far as functionality is concerned, ENCODE used the wrong definition wrongly. We must now address the question of consistency. Specifically, did ENCODE use the wrong definition wrongly in a consistent manner? We do not think so. For example, the ENCODE authors singled out transcription as a function, as if the passage of RNA polymerase through a DNA sequence is in some way more meaningful than other functions. But, what about DNA polymerase and DNA replication? Why make a big fuss about 74.7% of the genome that is transcribed, and yet ignore the fact that 100% of the genome takes part in a strikingly “reproducible biochemical signature”—it replicates!

From an evolutionary viewpoint, a function can be assigned to a DNA sequence if and only if it is possible to destroy it. All functional entities in the universe can be rendered nonfunctional by the ravages of time, entropy, mutation, and what have you. Unless a genomic functionality is actively protected by selection, it will accumulate deleterious mutations and will cease to be functional. The absurd alternative, which unfortunately was adopted by ENCODE, is to assume that no deleterious mutations can ever occur in the regions they have deemed to be functional. Such an assumption is akin to claiming that a television set left on and unattended will still be in working condition after a million years because no natural events, such as rust, erosion, static electricity, and earthquakes can affect it. The convoluted rationale for the decision to discard evolutionary conservation and constraint as the arbiters of functionality put forward by a lead ENCODE author (Stamatoyannopoulos 2012) is groundless and self-serving.

Epigenomics

- Part 1: Intro to epigenetics
- Part 2: High-throughput technologies
- **Part 3: Computational methods**

Mark D. Robinson, Statistical Genomics, IMLS

Overview of this part

- **Goal**: highlight where informatics approaches are being used, insights into (*a subset of*) bioinformatics research related to epigenomics
- Methods for individual platforms
  - DNA methylation
    - (BS-microarray) Illumina 450k array
    - (Affinity capture) BATMAN + new methods
  - Peak/region detection
    - MACS
  - Copy number and MBD/ChIP-seq
- Methods for integrating multiple data types
  - ChromHMM, Segway, ENCODE SOM “donuts”
  - Clustering - Repitools

University of Zurich^UZH

Institute of Molecular Life Sciences

Bisulphite conversion + "genotyping" array (Illumina HumanMethylaton450)

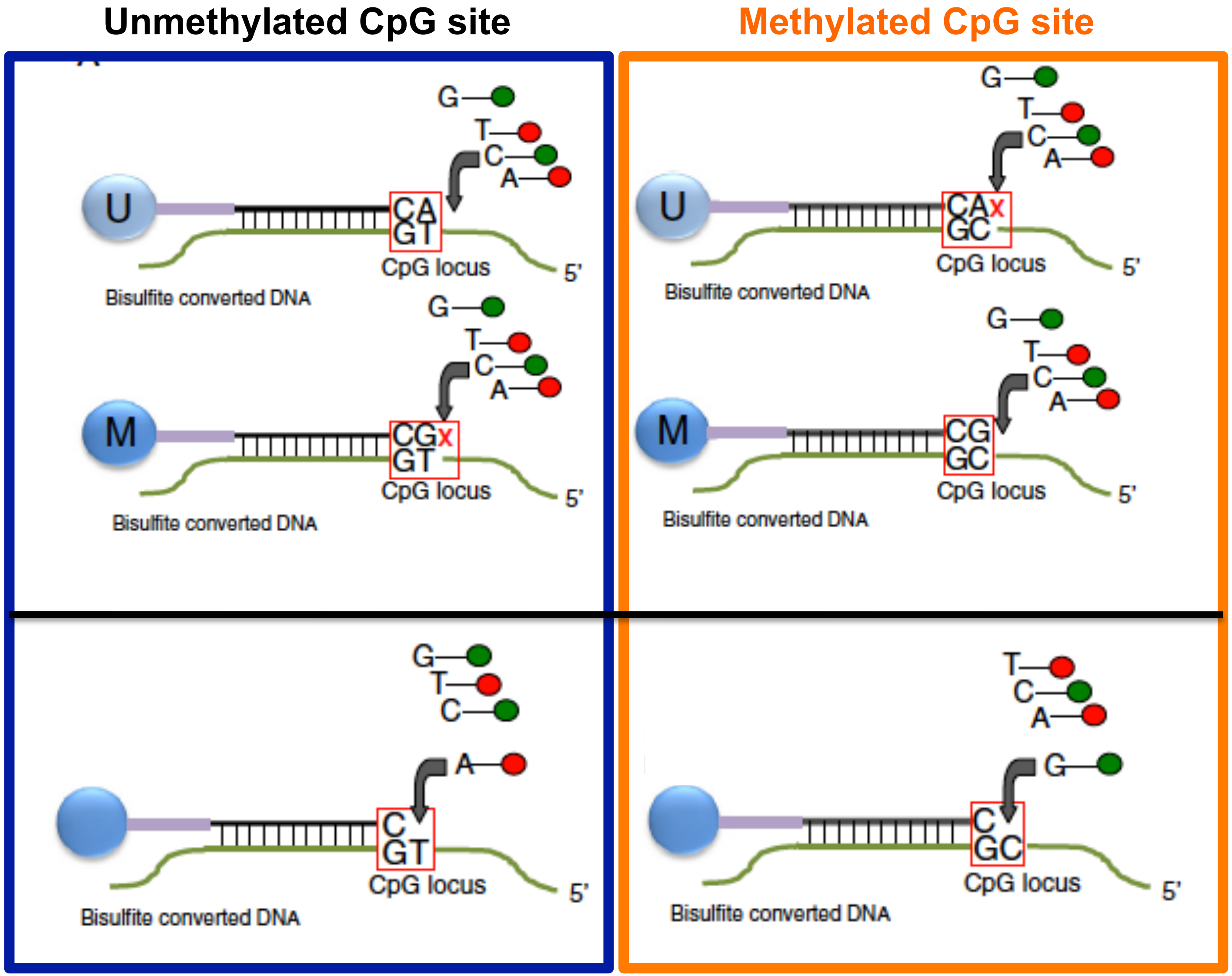
Type I

(2 probes)

beta = M / (M+U+e)

Type II

(1 probe)

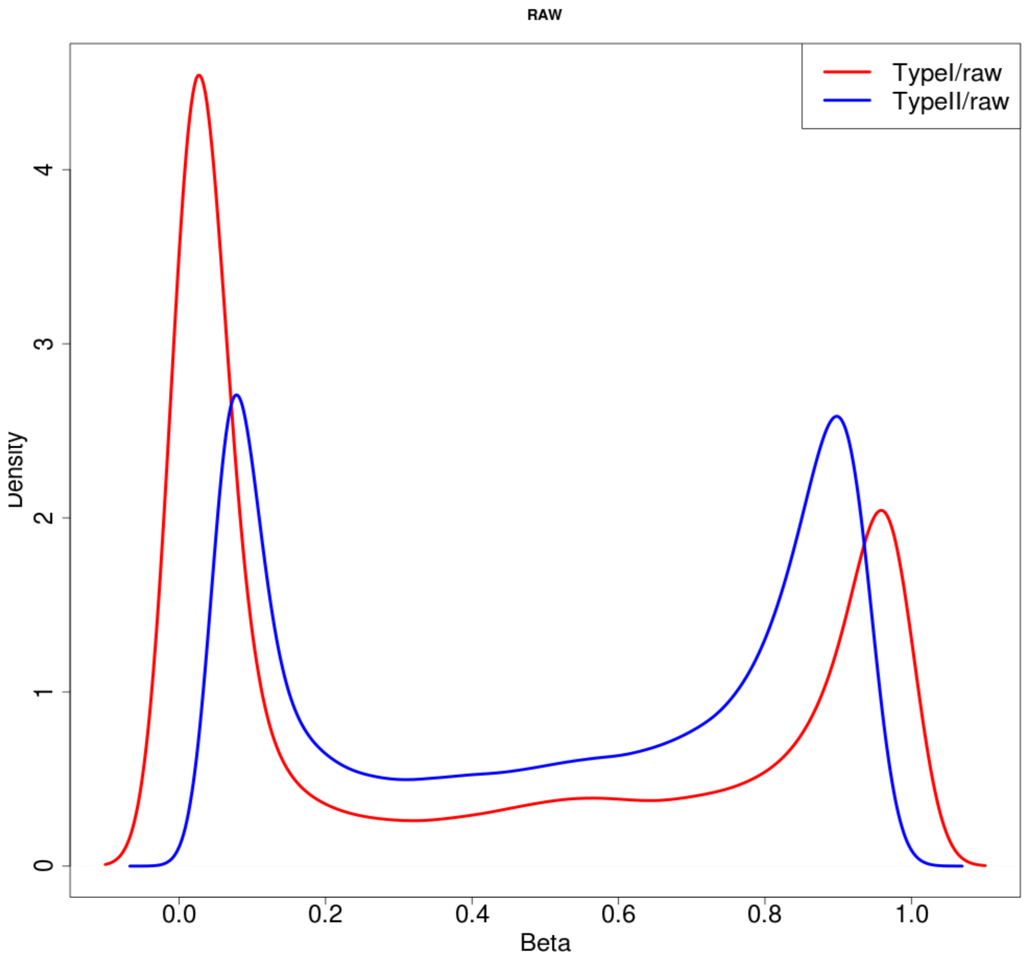


from Bibikova et al. Genomics 2011

450k arrays: probe-type bias

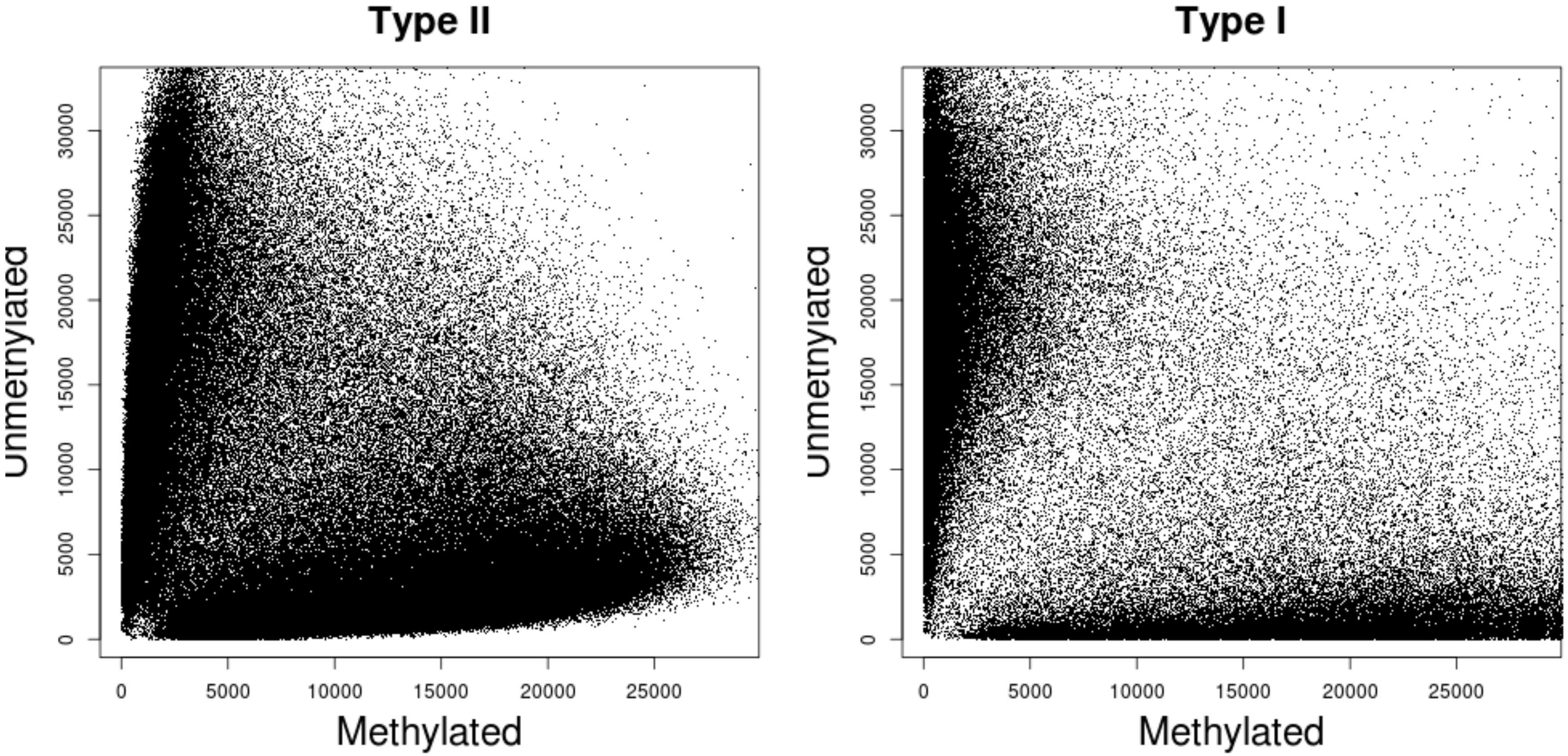
Overall, very good correspondence between 450k platform and others (e.g. BS-seq)

Normalization issues for different probe types (current research)



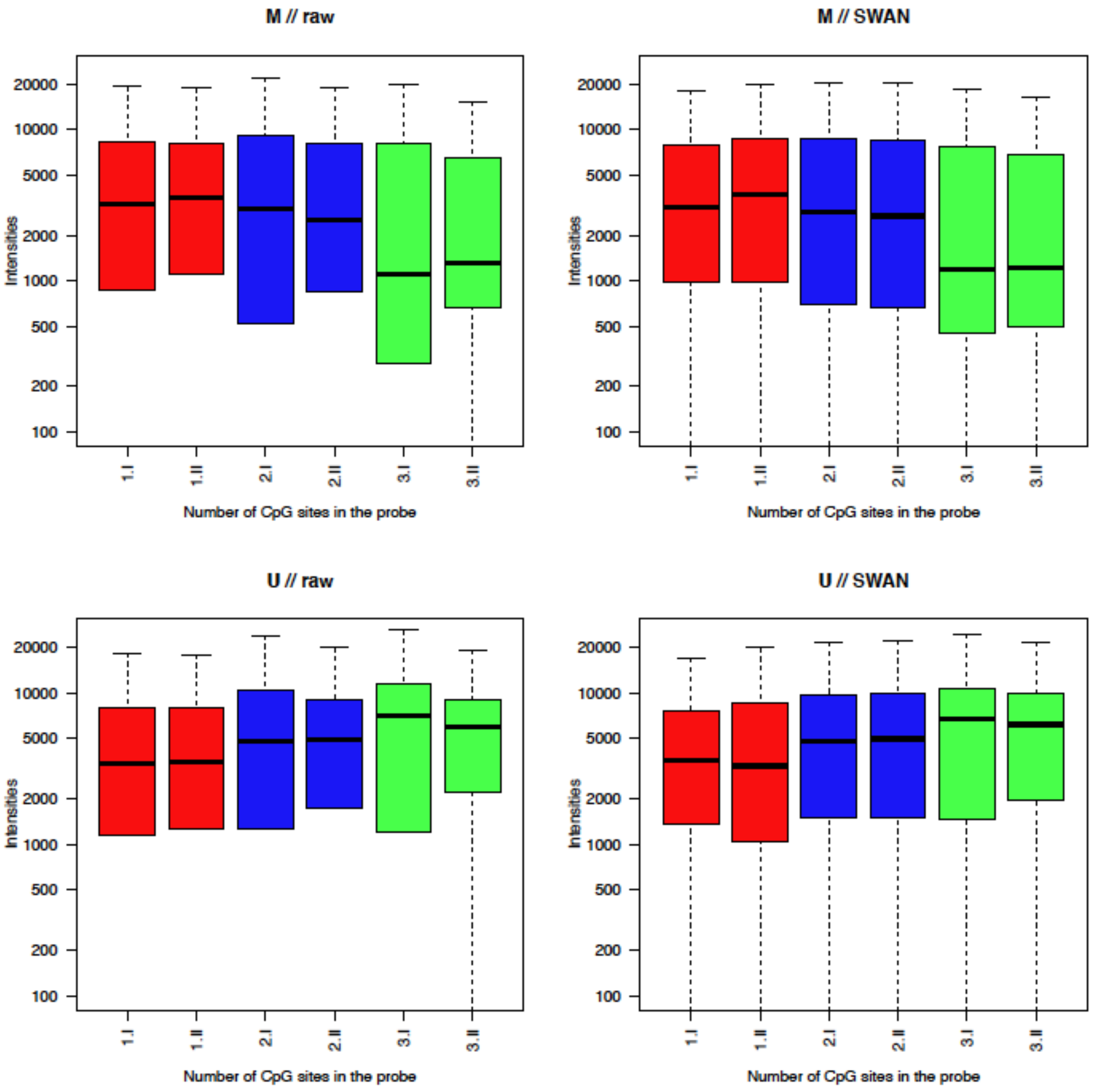
450k array data

Very different behaviour of Type I and Type II probes

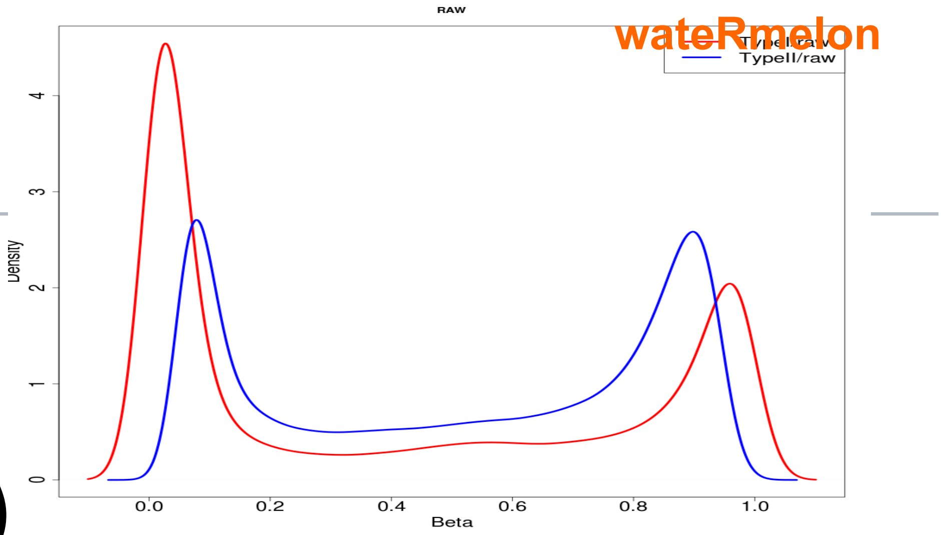


SWAN

(Maksimovic et al. 2012)
- SWAN: *Subset-quantile Within Array Normalization*
- quantile normalization based on the number of CpG sites
- outcome: makes Infinium I and II beta values distributions more similar



BMIQ



(Teschendorff et al. 2013)

- *Beta-mixture quantile normalization method*

- start from raw **beta** values

- fitting the three-state beta mixture model to the type I and type II probes separately

$$p(\beta^t) = \pi^t_U B(\beta|a^t_U, b^t_U) + \pi^t_H B(\beta|a^t_H, b^t_H) + \pi^t_M B(\beta|a^t_M, b^t_M)$$

charm/bumphunter

Methods for differential methylation

Methods for differential methylated **sites** use: i) log-ratios of methylated to unmethylated signal (450k array); ii) difference in binomials (BS-seq)

Methods are in active development for going from differentially methylated sites to differentially methylated **regions** (e.g. bump hunting).

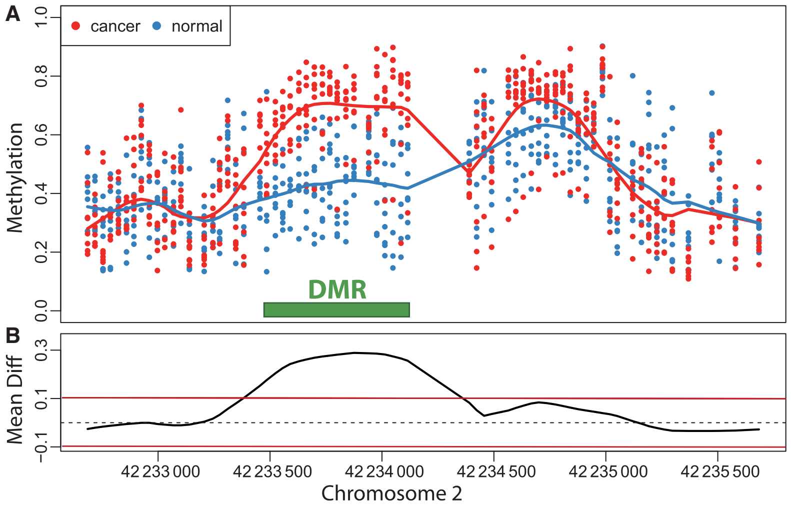
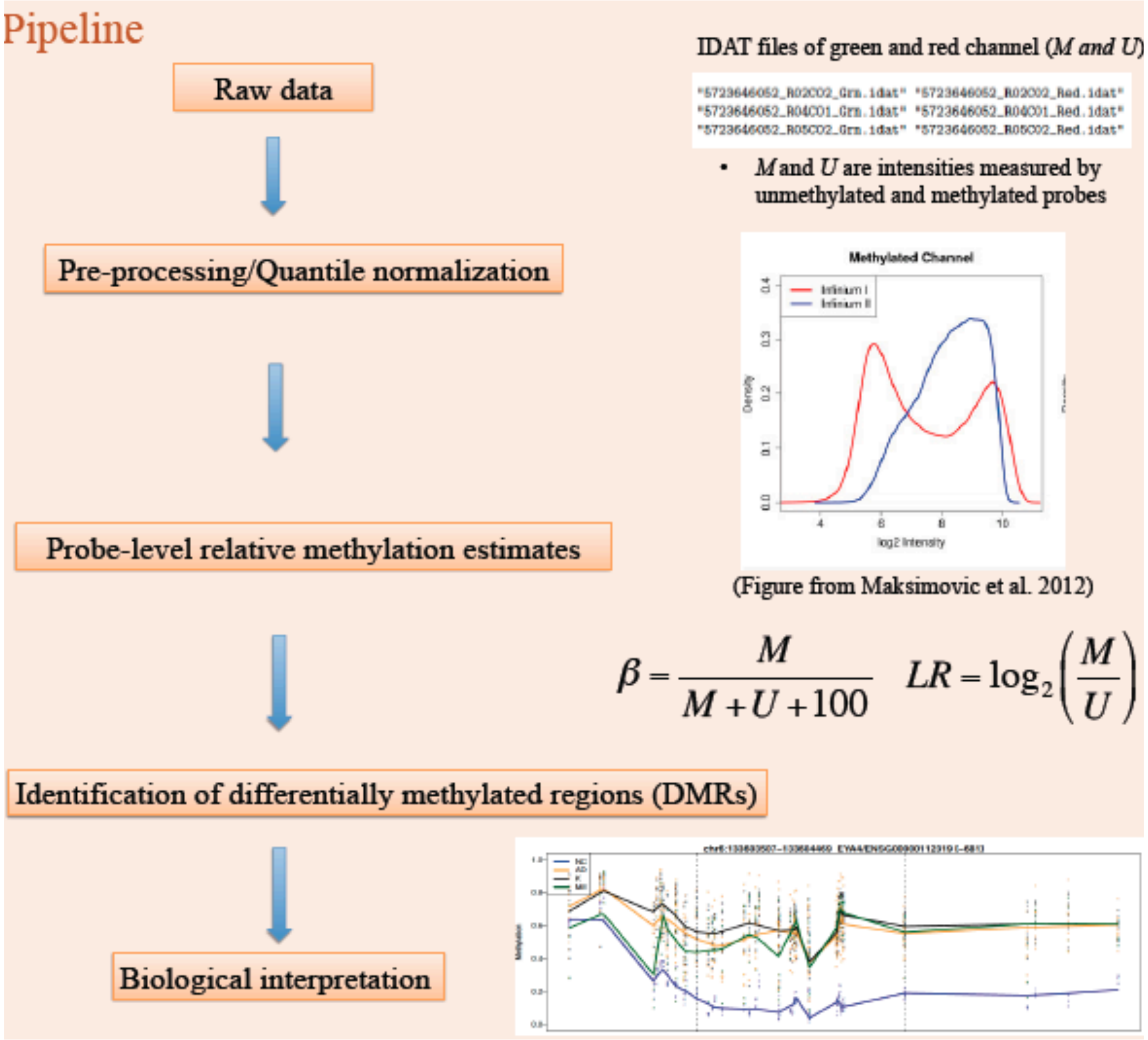


Figure 1 Example of a differentially methylation region (DMR). (A) The points show methylation measurements from the colon cancer dataset plotted against genomic location from illustrative region on chromosome 2. Eight normal and eight cancer samples are shown in this plot and represented by eight blue points and eight red points at each genomic location for which measurements were available. The curves represent the smooth estimate of the population-level methylation profiles for cancer (red) and normal (blue) samples. The green bar represents a region known to be a cancer DMR.[20] (B) The black curve is an estimate of the population-level difference between normal and cancer. We expect the curve to vary due to measurement error and biological variation but to rarely exceed a certain threshold, for example those represented by the red horizontal lines. Candidate DMRs are defined as the regions for which this black curve is outside these boundaries. Note that the DMR manifests as a *bump* in the black curve

Jaffe et al. (2012) Int. Journal of Epidemiology

Mechanics of DMR finding



Pipeline

Raw data

Pre-processing/Quantile normalization

Probe-level relative methylation estimates

Identification of differentially methylated regions (DMRs)

Biological interpretation

IDAT files of green and red channel (*M and U*)

```
"5723646052_R02C02_Grn.idat" "5723646052_R02C02_Red.idat"
"5723646052_R04C01_Grn.idat" "5723646052_R04C01_Red.idat"
"5723646052_R05C02_Grn.idat" "5723646052_R05C02_Red.idat"
```

- *M* and *U* are intensities measured by unmethylated and methylated probes

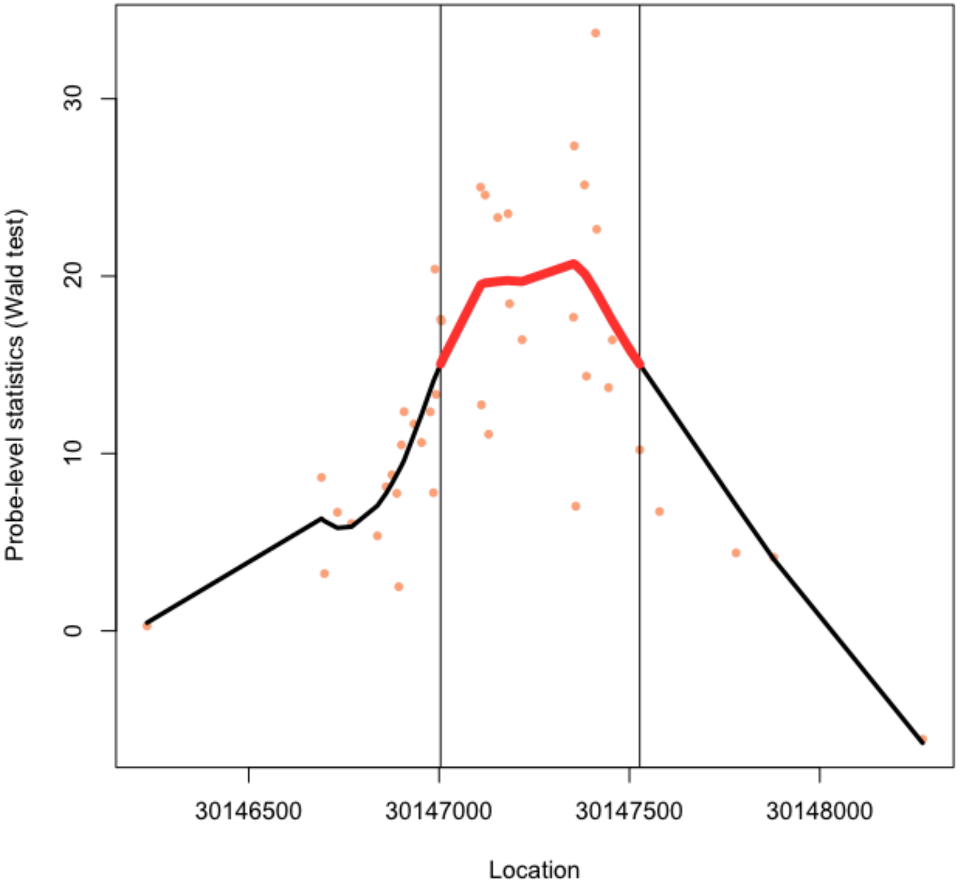
(Figure from Maksimovic et al. 2012)

$$\beta = \frac{M}{M+U+100} \qquad LR = \log_2\left(\frac{M}{U}\right)$$

Mechanics of DMR finding: charm/bumphunter package

Steps:

1. Get normalized data
2. For each probe (CpG site), calculate (differential) statistics at each probe
3. Apply a smoothing technique to these statistics
4. Set threshold and call regions as those that persist beyond threshold



BATMAN - Bayesian tool for methylation analysis

MeDIP-chip

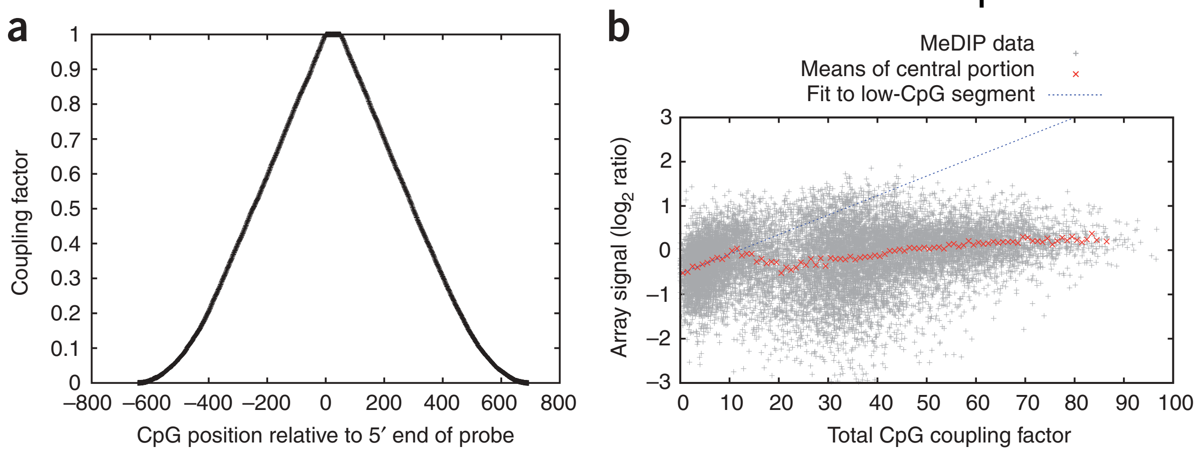
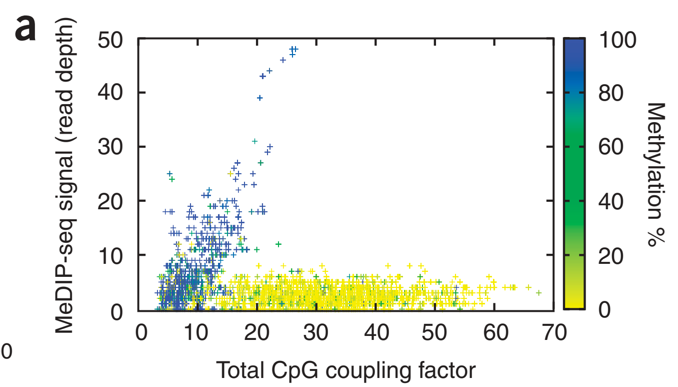


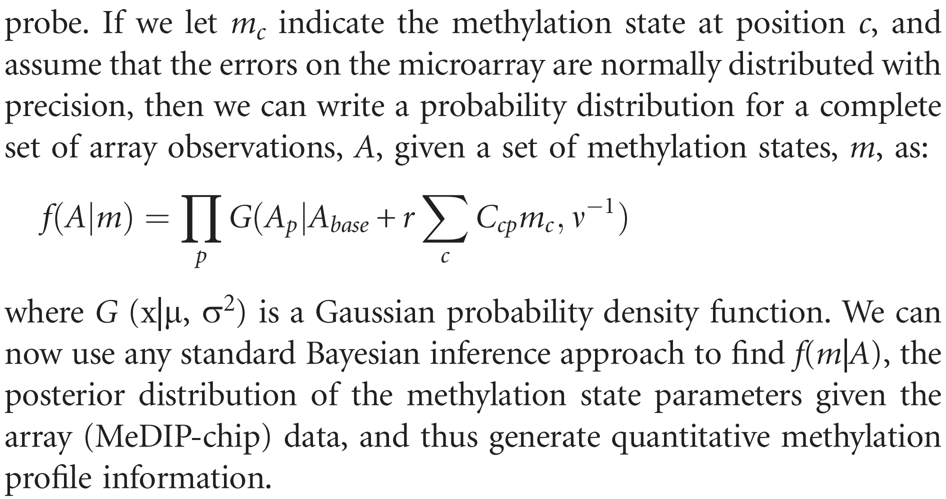
Figure 1 Calibration of the Batman model against MeDIP-chip data. (**a**) Estimated CpG coupling factors for a MeDIP-chip experiment as a function of the distance between a CpG dinucleotide and a microarray probe. (**b**) Plot of array signal against total CpG coupling factor, showing a linear regression fit to the low-CpG portion, as used in the Batman calibration step. This plot shows all data from one array on chromosome 6.

MeDIP-seq



Down et al. Nature Biotech 2008

BATMAN - Bayesian tool for methylation analysis



probe. If we let m_c indicate the methylation state at position c, and assume that the errors on the microarray are normally distributed with precision, then we can write a probability distribution for a complete set of array observations, A, given a set of methylation states, m, as:

$$f(A|m) = \prod_p G(A_p|A_{base} + r\sum_c C_{cp}m_c, v^{-1})$$

where G (x|μ, σ^2) is a Gaussian probability density function. We can now use any standard Bayesian inference approach to find $f(m|A)$, the posterior distribution of the methylation state parameters given the array (MeDIP-chip) data, and thus generate quantitative methylation profile information.

Same assumptions for MeDIP-chip (continous) can be applied to MeDIP-seq (count) and work quite well.

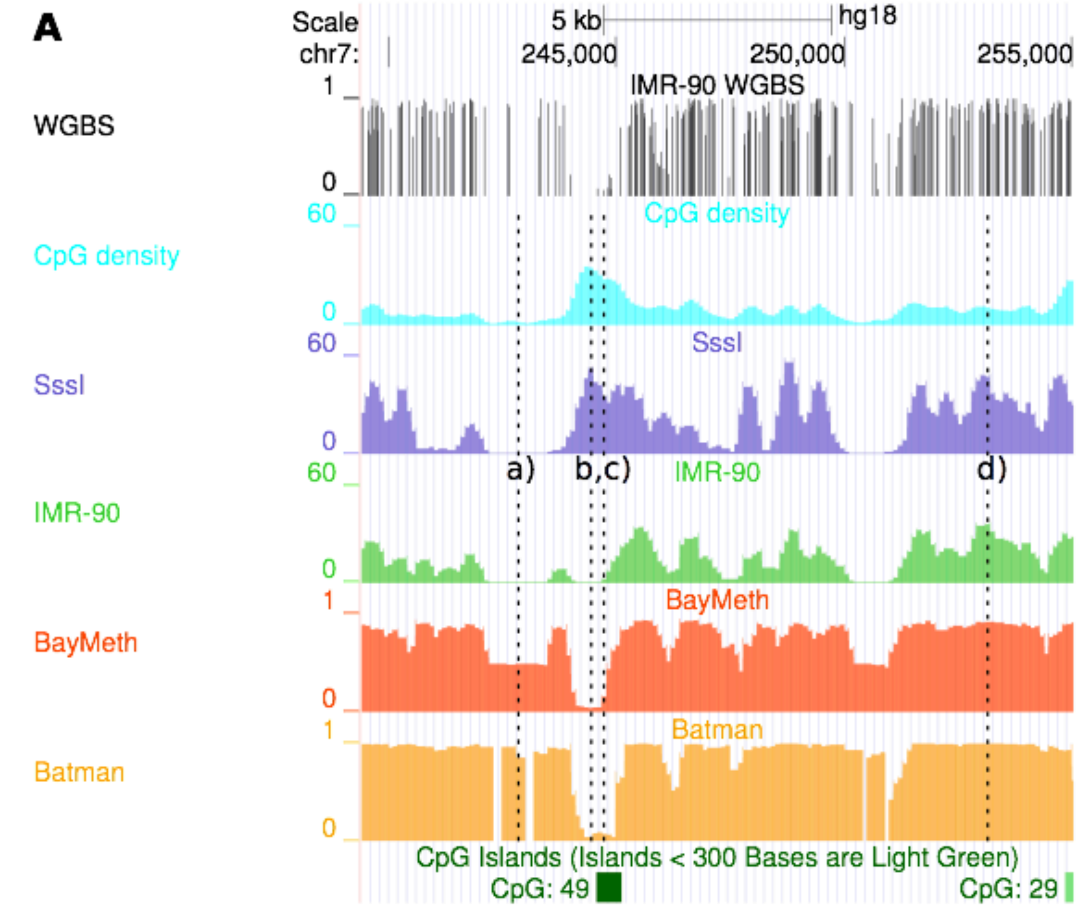
Some potential disadvantages:

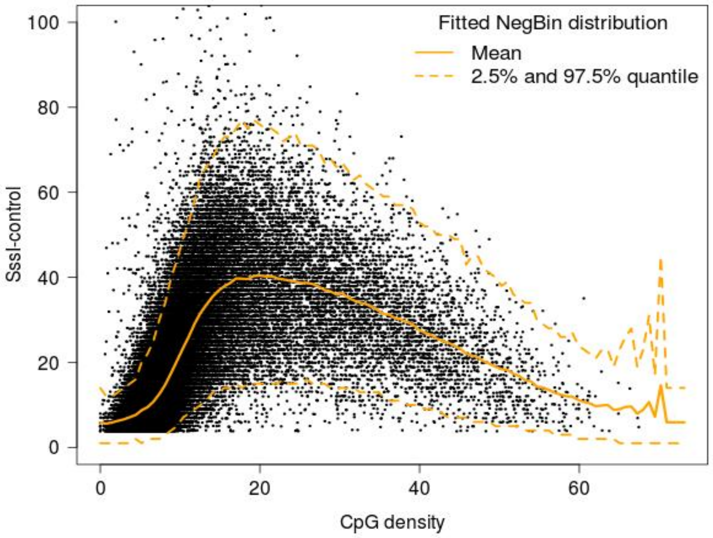
1. No reads = no DNA methylation *or* assay doesn't capture the region
2. MCMC is very computationally intensive (10-15h per chromosome)

MeDIP/MBD-seq: Count-based analyses using fully methylated control

Repitools

Figure 2 - Example data tracks for IMR-90 chromosome 7





Riebler et al. 2013, in revision

Model formulation: BayMeth

We consider genomic regions $i = 1, \ldots, n$ and define

- $y_{i,S}$: Number of reads for the sample of interest.
- $y_{i,C}$: Number of reads for the *SssI-control.*

Then,

$$y_{i,S} | \mu_i, \lambda_i \sim \text{Poisson}(f \times \mu_i \times \lambda_i); \qquad y_{i,C} | \lambda_i \sim \text{Poisson}(\lambda_i)$$

f: offset

λ_i: region-specific read density, and

μ_i: the regional methylation level (Main parameter of interest)

An analytic estimator!

$$\boxed{\mathrm{p}(\mu_i | y_{i1}, y_{i2}) = \frac{\mu_i^{y_{i1}}}{W} \left(1 - \frac{E(1-\mu_i)}{\beta + 1 + E}\right)^{-(\alpha + y_{i1} + y_{i2})}}$$

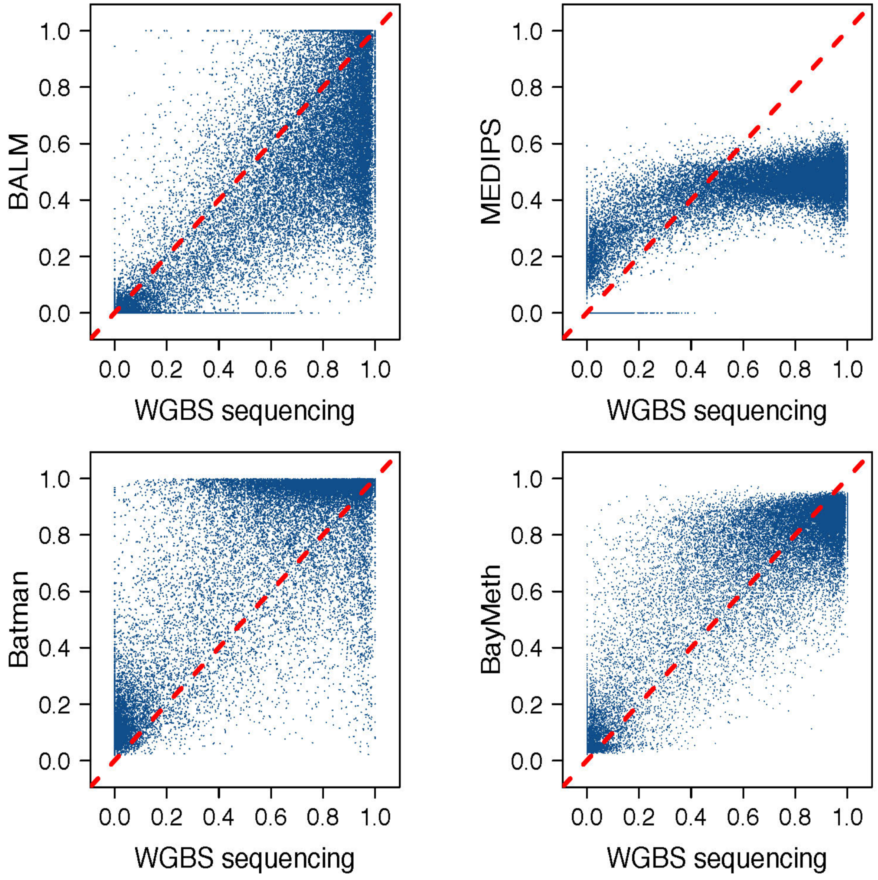
Model extension

We propose to adjust for this bias by including a second offset:

$$y_{i,\text{LNCaP}} | \mu_i, \lambda_i \sim \text{Poisson}(f \times \frac{cn_i}{4} \times \mu_i \times \lambda_i); \quad y_{i,C} | \lambda_i \sim \text{Poisson}(\lambda_i)$$

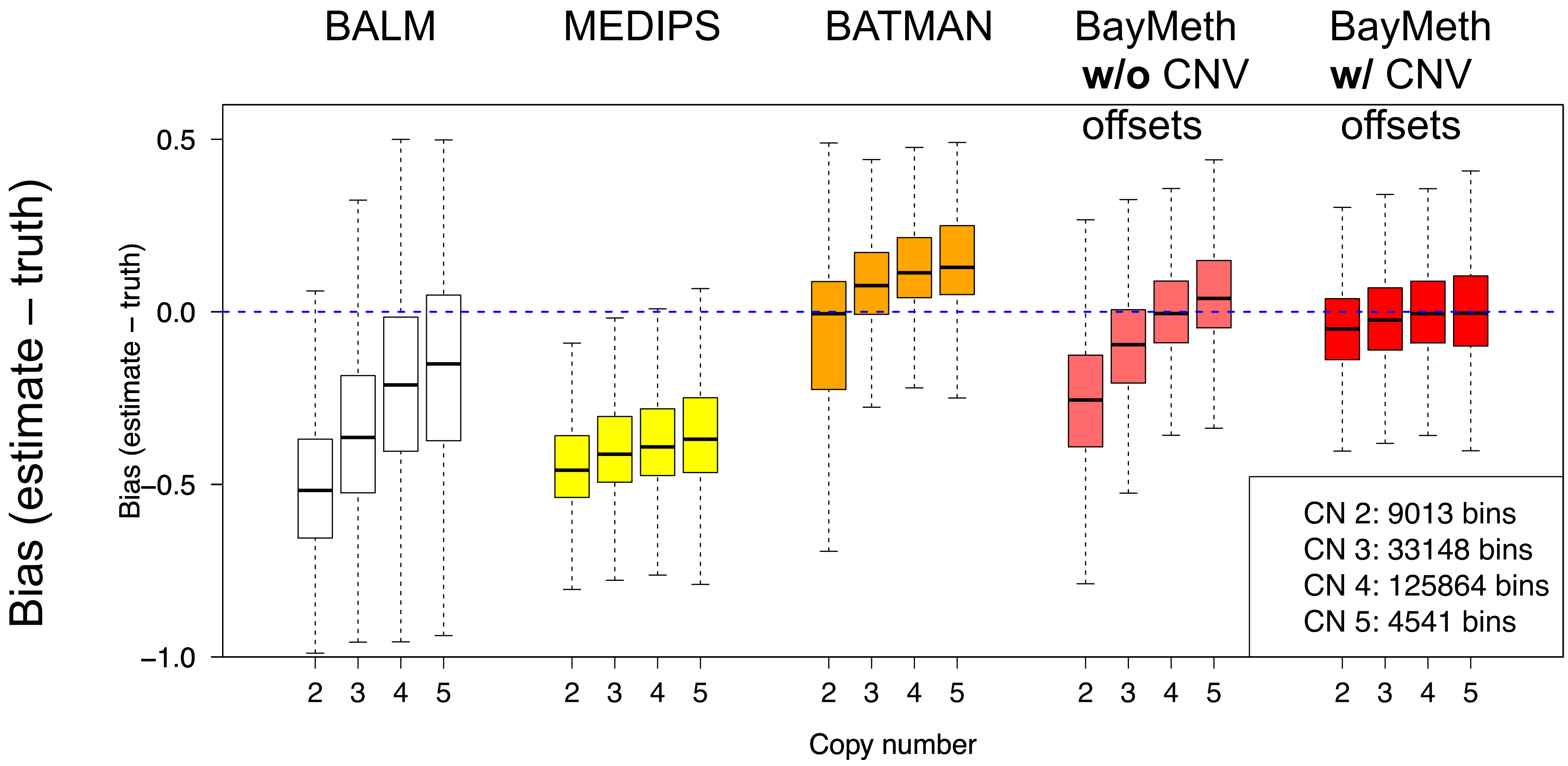
Riebler et al. 2013, in revision

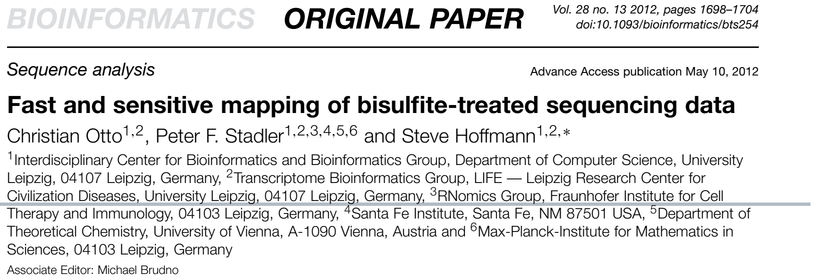
Improvements can be made



Our new method: **BayMeth**

Estimation bias, by CNV state





BIOINFORMATICS ORIGINAL PAPER Vol. 28 no. 13 2012, pages 1698–1704
doi:10.1093/bioinformatics/bts254

Sequence analysis Advance Access publication May 10, 2012

Fast and sensitive mapping of bisulfite-treated sequencing data

Christian Otto[1,2], Peter F. Stadler[1,2,3,4,5,6] and Steve Hoffmann[1,2,*]

[1]Interdisciplinary Center for Bioinformatics and Bioinformatics Group, Department of Computer Science, University Leipzig, 04107 Leipzig, Germany, [2]Transcriptome Bioinformatics Group, LIFE — Leipzig Research Center for Civilization Diseases, University Leipzig, 04107 Leipzig, Germany, [3]RNomics Group, Fraunhofer Institute for Cell Therapy and Immunology, 04103 Leipzig, Germany, [4]Santa Fe Institute, Santa Fe, NM 87501 USA, [5]Department of Theoretical Chemistry, University of Vienna, A-1090 Vienna, Austria and [6]Max-Planck-Institute for Mathematics in Sciences, 04103 Leipzig, Germany

Associate Editor: Michael Brudno

Bisulphite sequencing analyses: mapping

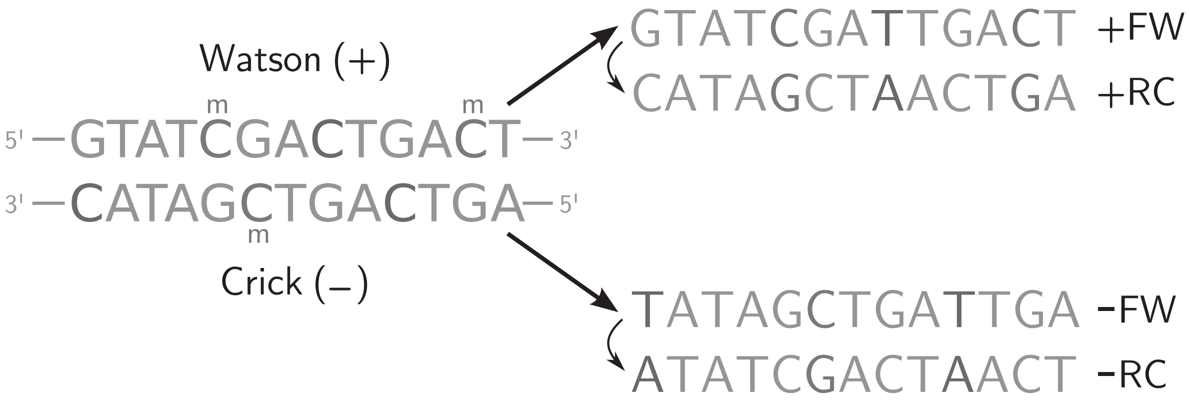
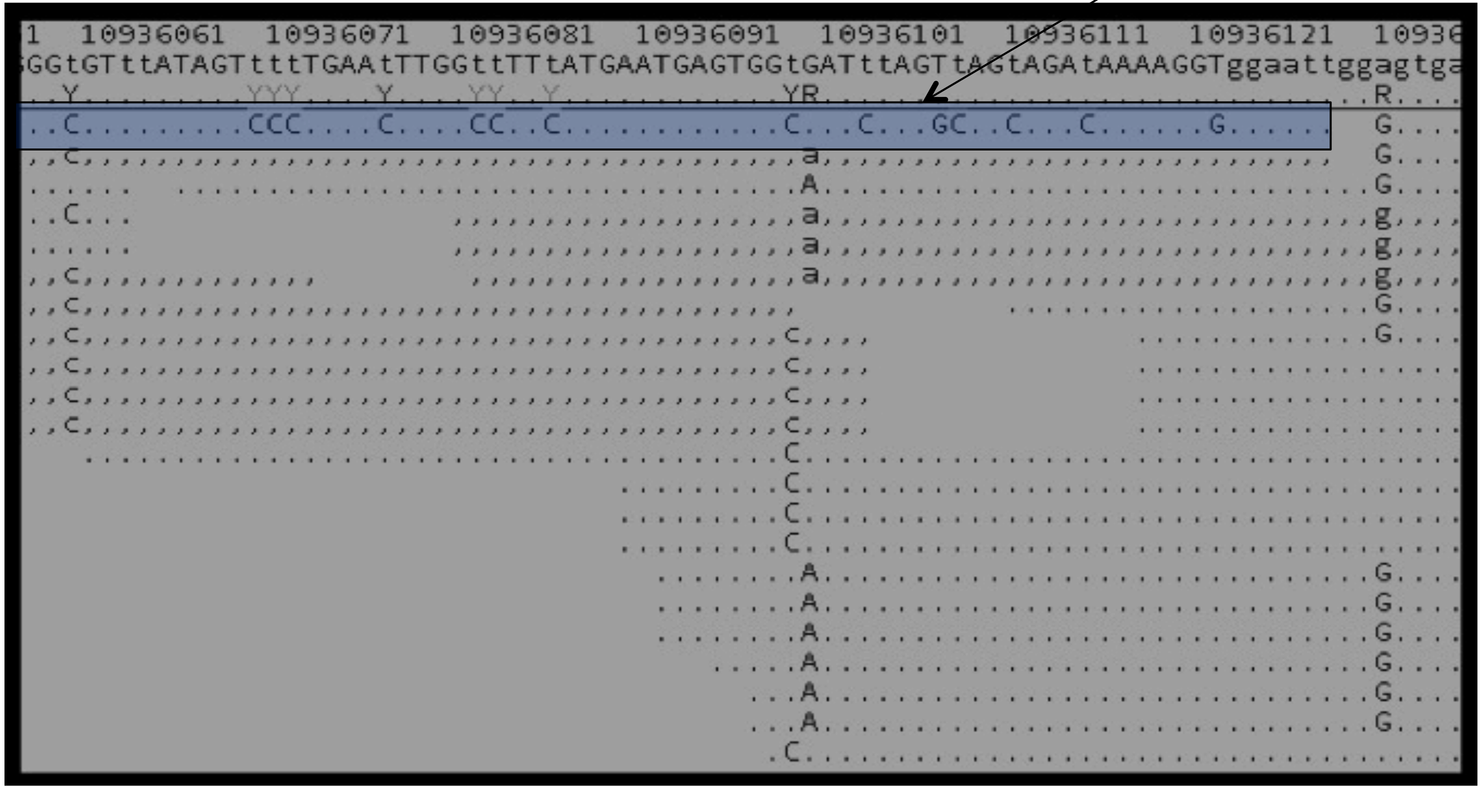


Fig. 1. Possible read types (+FW, +RC, −FW and −RC) in bisulfite sequencing protocols. Methylated and unmethylated cytosines in the genomic sequence (left) are coloured in red and blue, respectively, and positions in the read sequences (right) derived from genomic cytosines are coloured correspondingly. Note that the intermediate conversion of unmethylated cytosines into uracils after bisulfite treatment is omitted

Analysis of BS-seq data

Unconverted sequence

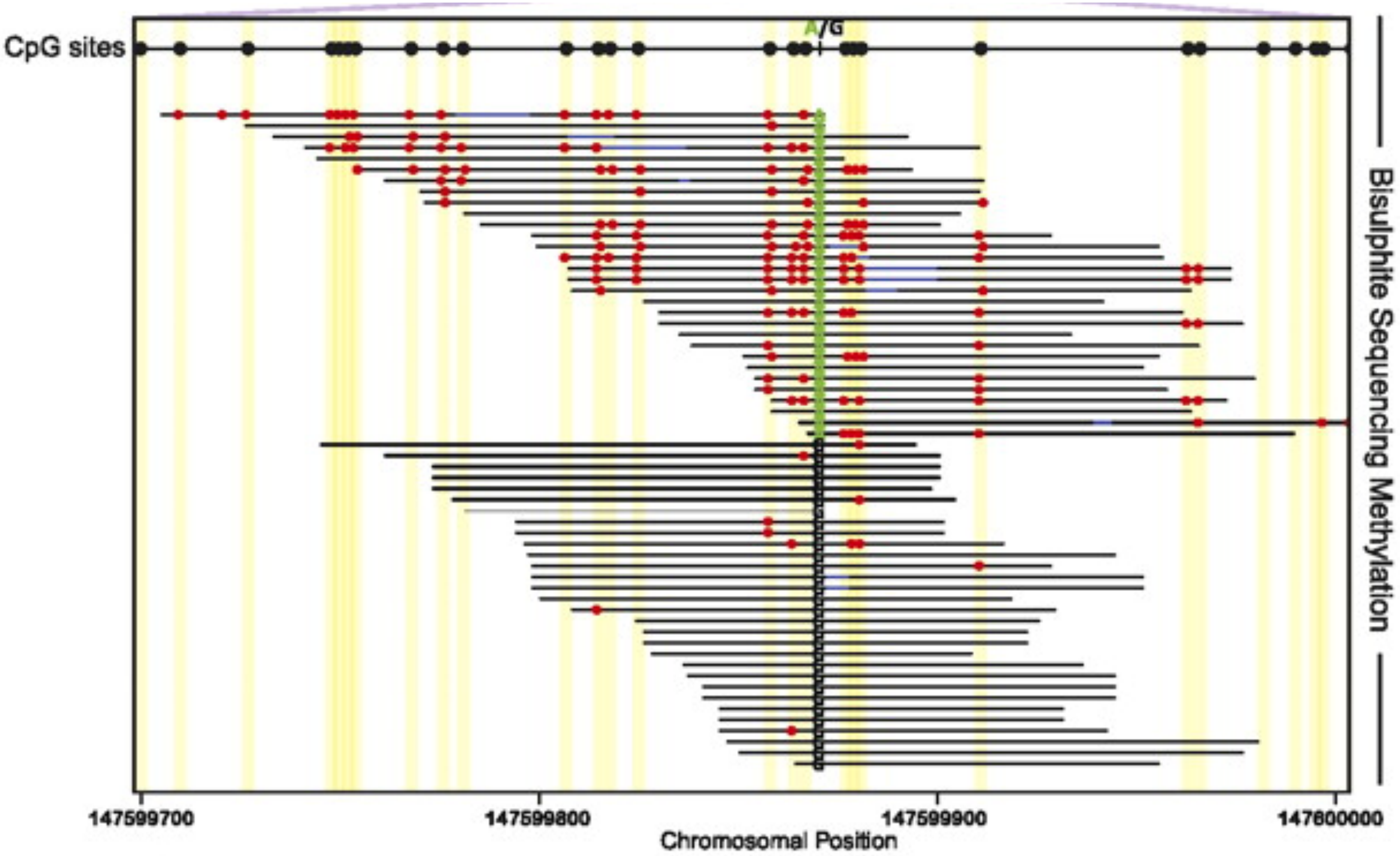


```
1  10936061  10936071  10936081  10936091  10936101  10936111  10936121  10936
GGtGTttATAGTtttTGAAtTTGGtTTtATGAATGAGTGGtGATttAGTtAGtAGAtAAAAGGTggaattggagtga
..Y.........YYY....Y....YY..Y...........YR..............................R....
..C.........CCC....C....CC..C...........C...C...GC..C...C......G.....   G....
,,C,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,a,,,,,,,,,,,,,,,,,,,,,,,,,,,   G....
......  .................................A..............................G....
..C...                 ,,,,,,,,,,,,,,,,,,a,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,g,,,,
......                 ,,,,,,,,,,,,,,,,,,a,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,g,,,,
,,C,,,,,,,,,,,,,        ,,,,,,,,,,,,,,,,,a,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,g,,,,
,,C,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,           ....................G....
,,C,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,C,,,,              .............G....
,,C,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,C,,,,              ..................
,,C,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,C,,,,              ..................
,,C,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,C,,,,              ..................
   .....................................C....................................
                                ........C....................................
                                ........C....................................
                                ........C....................................
                                 ........A..............................G....
                                 ........A..............................G....
                                 ........A..............................G....
                                    .....A..............................G....
                                      ...A..............................G....
                                      ...A..............................G....
                                       .C....................................
```

BS-converted reference.
t = converted C

Statham, Robinson et al.
2012 Genome Research

Alternative visualization of BS-seq data

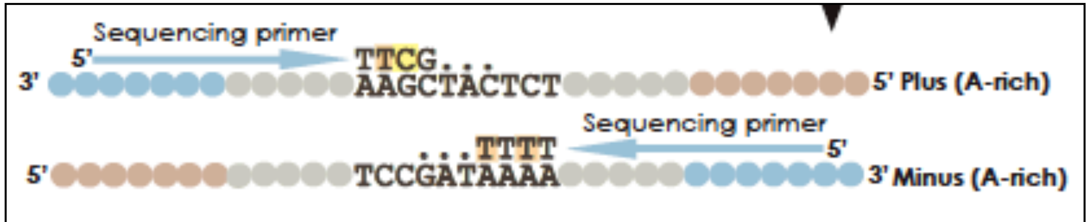


Genotypes with BS-seq data

	Positive Strand	Negative Strand
Reference genome:	TCCGATGAGA	TCTCATCGGA
Add optional methylation:	TC**CG**ATGAGA	TCTCAT**CG**GA
Actual read:	TTCGATGAGA	TTTTATCGGA

Rule:

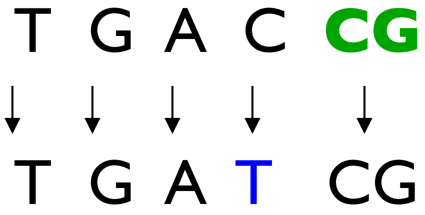
T G A C **CG**
↓ ↓ ↓ ↓ ↓
T G A T CG



Sequence recovered strands on Illumina GA2, yeilding T-rich reads.

Genotypes with BS-seq data

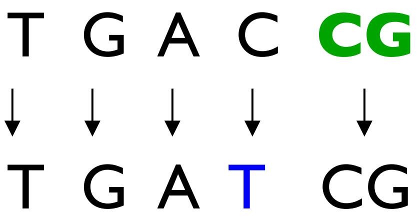
	Positive Strand	Negative Strand
Reference:	TCCGATGAGA	TCTCATCGGA
What if the genome was:	GCCGATGAGA	TCTCATCGGC
	CCCGATGAGA	TCTCATCGGG
	ACCGATGAGA	TCTCATCGGT
Actual read:	TTCGATGAGA	TTTTATCGGA
	GCCGATGAGA	TTTTATCGGT
	TCCGATGAGA	TTTTATCGGG
	ACCGATGAGA	TTTTATCGGT



You can reconcile the ambiguity with the read from the opposite strand.

Genotypes with BS-seq data

	Positive Strand	Negative Strand
Reference:	TCCGATGAGA	TCTCATCGGA
What if the genome was:	TGCGGTGAGA TACGCTGAGA TTCGTTGAGA	TCTCACCGCA TCTCAGCGTA TCTCAACGAA
Actual read:	TTCGATGAGA	TTTTATCGGA
	TGCGGTGAGA TACGTTGAGA TTCGTTGAGA	TTTTATCGTA TTTTAGCGTA TTTTAACGAA


T G A C CG
T G A T CG

Again, you can reconcile the ambiguity with the read from the opposite strand.

Genotype information from BS-seq data

We don’t always get allele information from both strands … i.e. when the methylation base call interferes with the SNP base call

Ref	Alt	Genotype	Info from	Ref (+) read as:	Ref(-) read as:	Alt (+) read as:	Alt (-) read as:
A	C	A/C	Both	A	A	C or T	C
A	G	A/G	+	A	n/a	G	n/a
A	T	A/T	Both	A	A	T	T
C	A	A/C	Both	C or T	C	A	A
C	G	C/G	Both	C or T	C	G	G or A
C	T	C/T	-	n/a	C	n/a	T
G	A	A/G	+	G	n/a	A	n/a
G	C	C/G	Both	G	G or A	C or T	C
G	T	G/T	Both	G	G or A	T	T
T	A	A/T	Both	T	T	A	A
T	C	C/T	-	n/a	T	n/a	C
T	G	G/T	Both	T	T	G	G or A

Pipelines: sequencing reads to data analysis

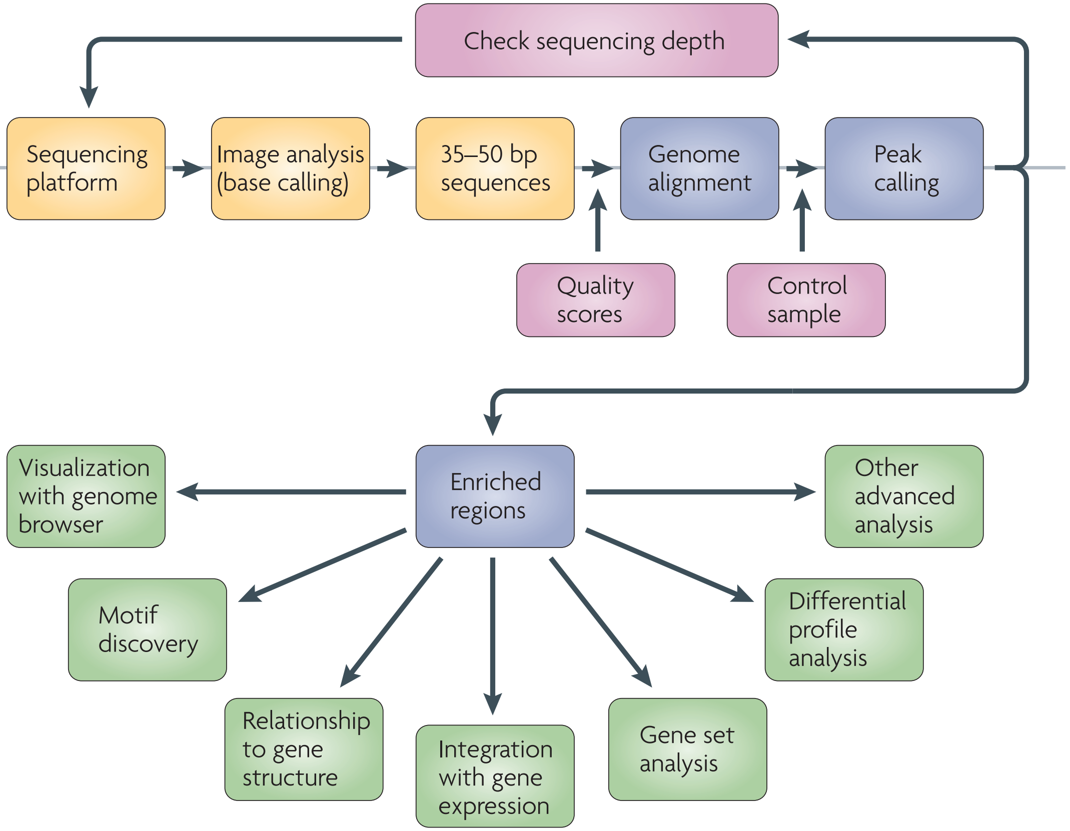


Figure 4 | **Overview of ChIP–seq analysis.** The raw data for chromatin immunoprecipitation followed by sequencing (ChIP–seq) analysis are images from the next-generation sequencing platform (top left). A base caller converts the image data to sequence tags, which are then aligned to the genome. On some platforms, they are aligned with the aid of quality scores that indicate the reliability of each base call. Peak calling, using data from the ChIP profile and a control profile (which is usually created from input DNA), generates a list of enriched regions that are ordered by false discovery rate as a statistical measure. Subsequently, the profiles of enriched regions are viewed with a browser and various advanced analyses are performed.

Many sequencing experiments have some common initial preprocessing elements (e.g. read mapping); microarray experiments – normalization.

Downstream informatic analyses are catered to the scientific question.

Peter J. Park Nature Reviews Genetics 2009

ChIP-seq for TFs versus ChIP-seq for histone modifications

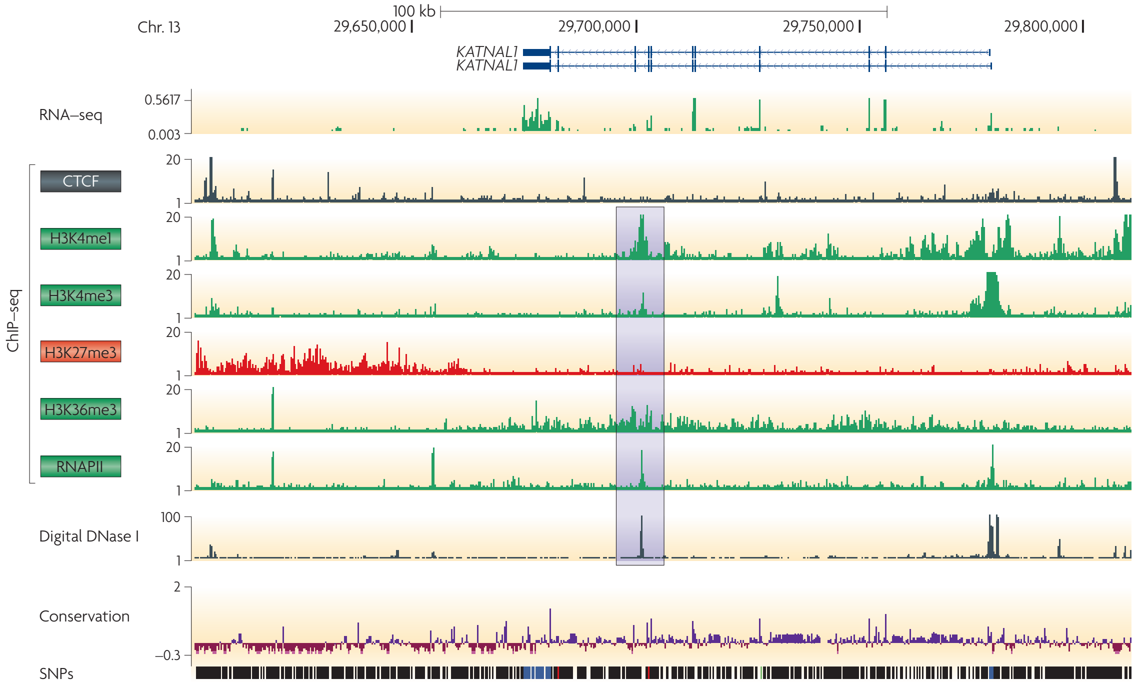


Figure 3 | **Data visualization.** The University of California-Santa Cruz (UCSC) Genome Browser is a tool for viewing genomic data sets. A vast amount of data is available for viewing through this browser. This example from the browser shows numerous data types in K562 cells from the ENCODE Consortium. A random gene was selected — katanin p60 subunit A-like 1 (*KATNAL1*) — that shows several points that can be identified by using this tool. The promoter has a typical chromatin structure (a peak of histone 3 lysine 4 trimethylation (H3K4me3) between the bimodal peaks of H3K4me1), is bound by RNA polymerase II (RNAPII) and is DNase hypersensitive. The gene is transcribed, as indicated by RNA sequencing (RNA–seq) data, as well as H3K36me3 localization. The gene lies between two CCCTC-binding factor (CTCF)-bound sites that could be tested for insulator activity. An intronic H3K4me1 peak (highlighted) predicts an enhancer element, corroborated by the DNase I hypersensitivity site peak. There is a broad repressive domain of H3K27me3 downstream, which could have an open chromatin structure in another cell type.

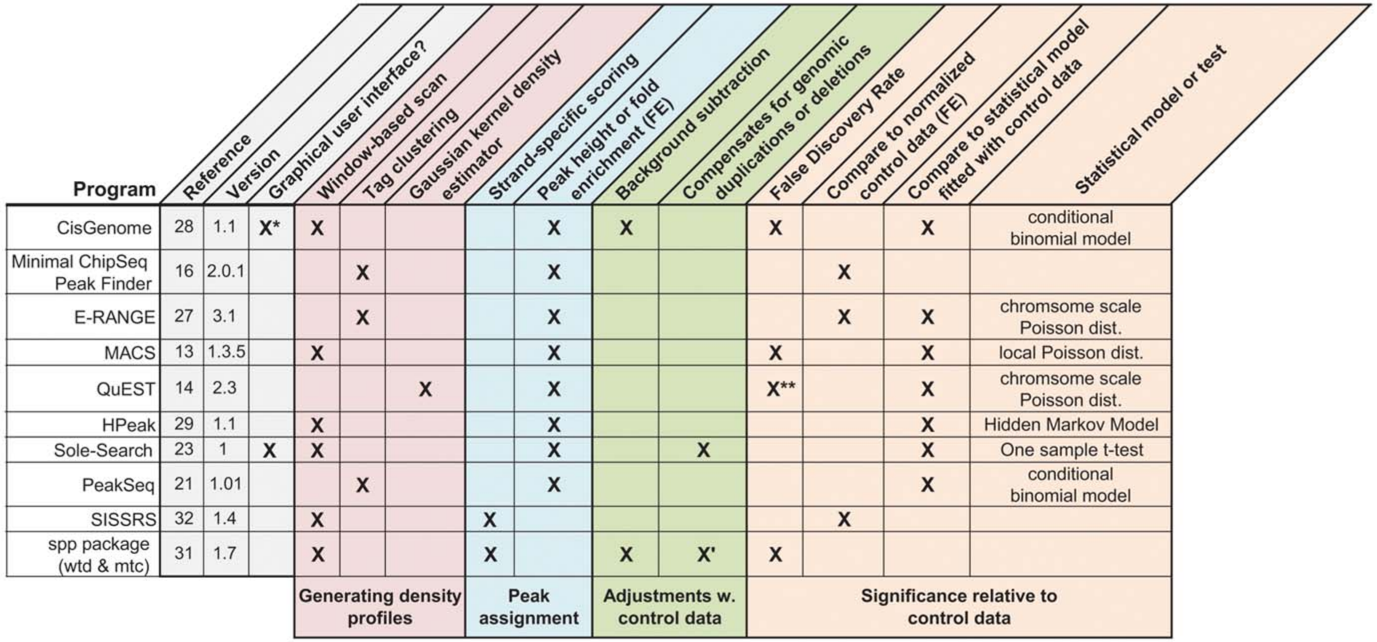
APPLICATIONS OF NEXT-GENERATION SEQUENCING

Next-generation genomics: an integrative approach

R. David Hawkins, Gary C. Hon* and Bing Ren*

ChIP-seq programs

Wilbanks and Facciotti (2010) PLoS ONE



Program	Reference	Version	Graphical user interface?	Window-based scan	Tag clustering	Gaussian kernel density estimator	Strand-specific scoring	Peak height or fold enrichment (FE)	Background subtraction	Compensates for genomic duplications or deletions	False Discovery Rate	Compare to normalized control data (FE)	Compare to statistical model fitted with control data	Statistical model or test
CisGenome	28	1.1	X*	X				X	X		X		X	conditional binomial model
Minimal ChipSeq Peak Finder	16	2.0.1			X			X				X		
E-RANGE	27	3.1			X			X				X	X	chromsome scale Poisson dist.
MACS	13	1.3.5		X				X			X		X	local Poisson dist.
QuEST	14	2.3				X		X			X**		X	chromsome scale Poisson dist.
HPeak	29	1.1		X				X					X	Hidden Markov Model
Sole-Search	23	1	X	X				X		X			X	One sample t-test
PeakSeq	21	1.01			X			X					X	conditional binomial model
SISSRS	32	1.4		X			X					X		
spp package (wtd & mtc)	31	1.7		X			X		X	X'	X			
				Generating density profiles			Peak assignment		Adjustments w. control data		Significance relative to control data			

X* = Windows-only GUI or cross-platform command line interface

X** = optional if sufficient data is available to split control data

X' = method exludes putative duplicated regions, no treatment of deletions

Figure 2. ChIP-seq peak calling programs selected for evaluation. Open-source programs capable of using control data were selected for testing based on the diversity of their algorithmic approaches and general usability. The common features present in different algorithms are summarized, and grouped by their role in the peak calling procedure (colored blocks). Programs are categorized by the features they use (Xs) to call peaks from ChIP-seq data. The version of the program evaluated in this analysis is shown for each program, as the feature lists can change with program updates.

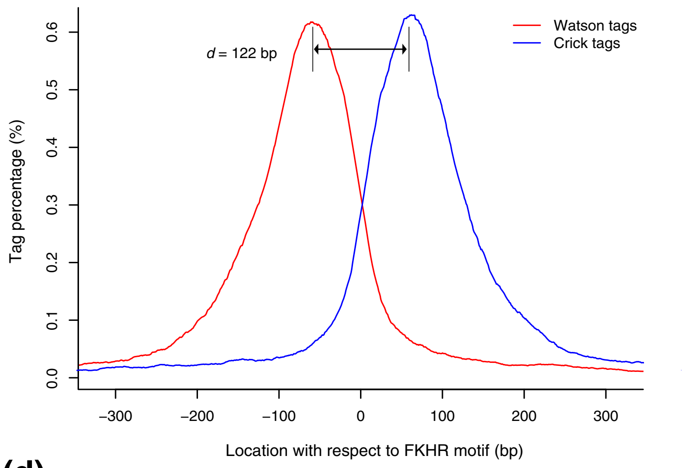
doi:10.1371/journal.pone.0011471.g002

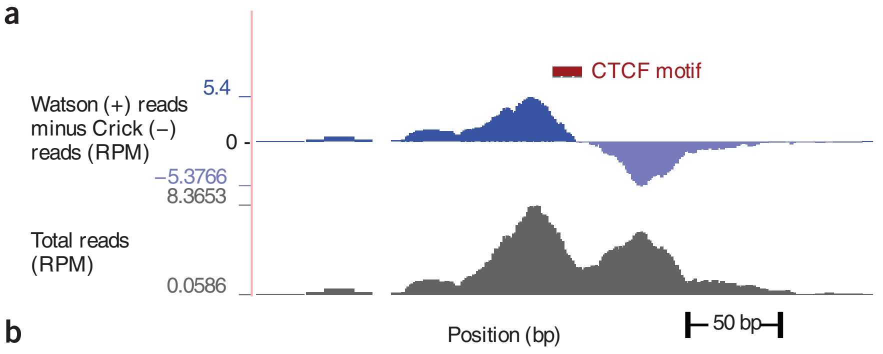
(b)

Peak/region detection for ChIP-seq data

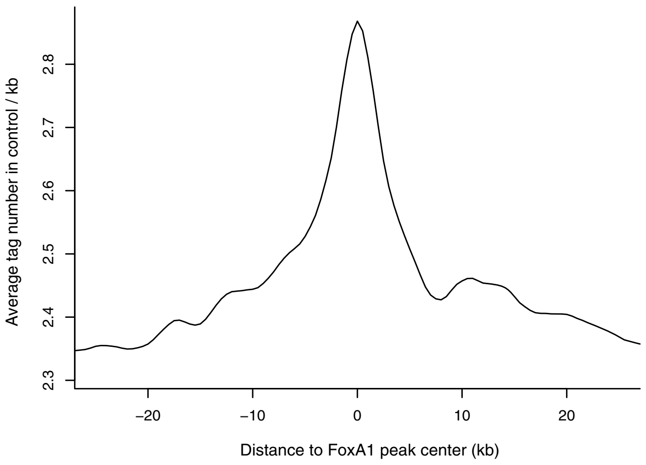
MACS: model-based analysis of ChIP-seq data

Accounting for strandedness of the reads





(d)

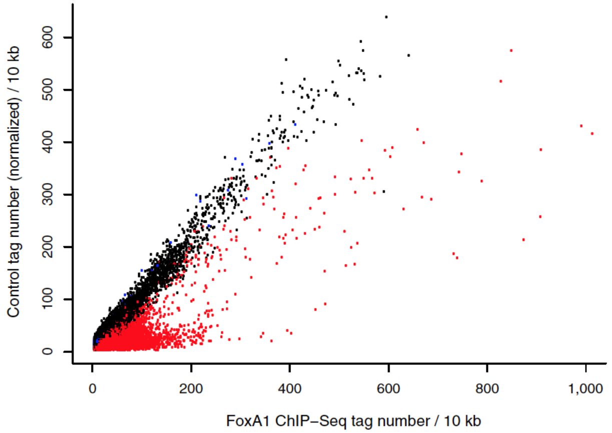


(f)

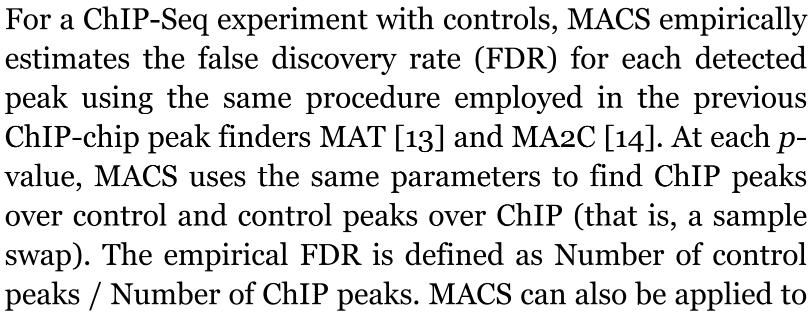
MACS – model-based analysis of ChIP-seq

Simple algorithm:

1. Estimate average fragment size ‘d’
2. Adjust reads by d/2
3. From control sample, estimate local background (if control sample used)
4. For each window, calculate Poisson P-value (probability of more extreme than local rate)
5. Estimate empirical FDR



$$\lambda_{\text{local}} = \max(\lambda_{\text{BG}}, [\lambda_{\text{1k}},]\ \lambda_{\text{5k}}, \lambda_{\text{10k}})$$


For a ChIP-Seq experiment with controls, MACS empirically estimates the false discovery rate (FDR) for each detected peak using the same procedure employed in the previous ChIP-chip peak finders MAT [13] and MA2C [14]. At each *p*-value, MACS uses the same parameters to find ChIP peaks over control and control peaks over ChIP (that is, a sample swap). The empirical FDR is defined as Number of control peaks / Number of ChIP peaks. MACS can also be applied to

BayesPeak:

$$\begin{aligned} Y_t^+, Y_{t+1}^- \mid Z_t = 0 \sim & \quad \text{Poisson}(\lambda_0 \gamma^{w_t}) \\ Y_t^+, Y_{t+1}^- \mid Z_t = 1,2,3 \sim & \quad \text{Poisson}((\lambda_0 + \lambda_1)\gamma^{w_t}) \\ \lambda_0 \sim & \quad \Gamma(\alpha_0, \beta_0) \\ \lambda_1 \sim & \quad \Gamma(\alpha_1, \beta_1) \end{aligned}$$

$$Z_t = \begin{cases} 0 & \text{if} \quad (S_t, S_{t+1}) = (0,0) \\ 1 & \text{if} \quad (S_t, S_{t+1}) = (0,1) \\ 2 & \text{if} \quad (S_t, S_{t+1}) = (1,0) \\ 3 & \text{if} \quad (S_t, S_{t+1}) = (1,1) \end{cases}$$

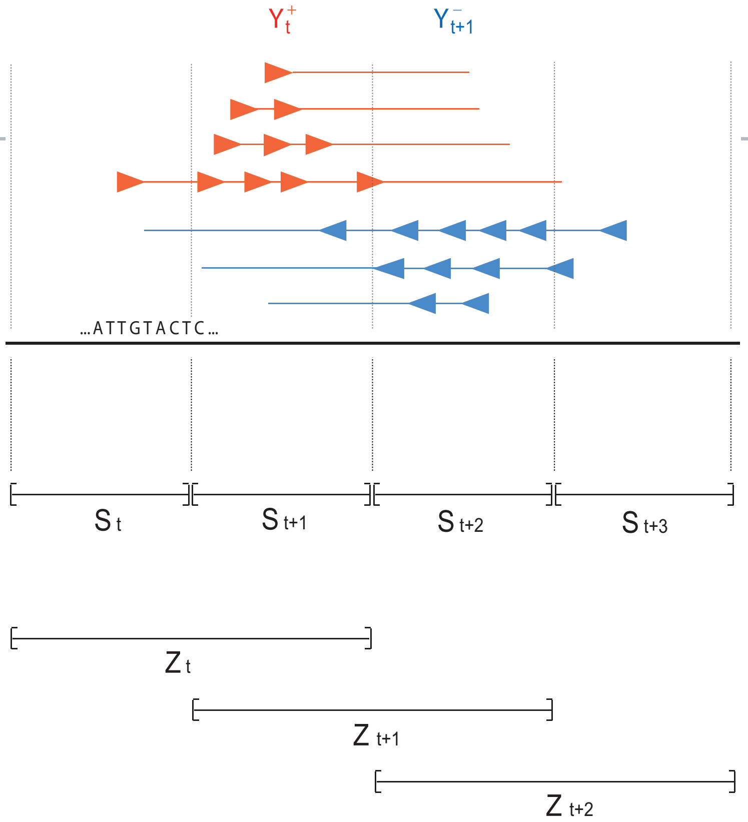


Figure I

Illustration of the model. This figure shows how the reads (arrows) on the forward and reverse strand, indicated by red and blue respectively, are counted as Y_t^+ and Y_{t+1}^- and depend on the nature of the underlying regions *t* and *t* + I when their full length is taken into consideration. Moreover, this figure shows how each Z_t state corresponds to the underlying ones S_t and S_{t+1}.

BayesPeak models +/- strands directly

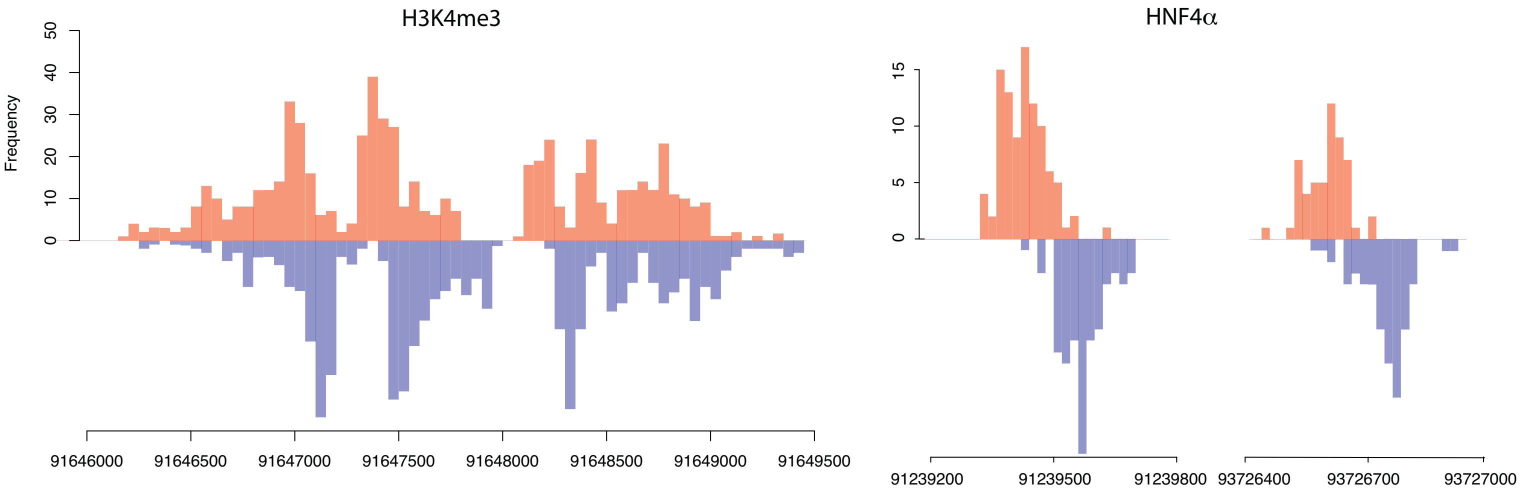
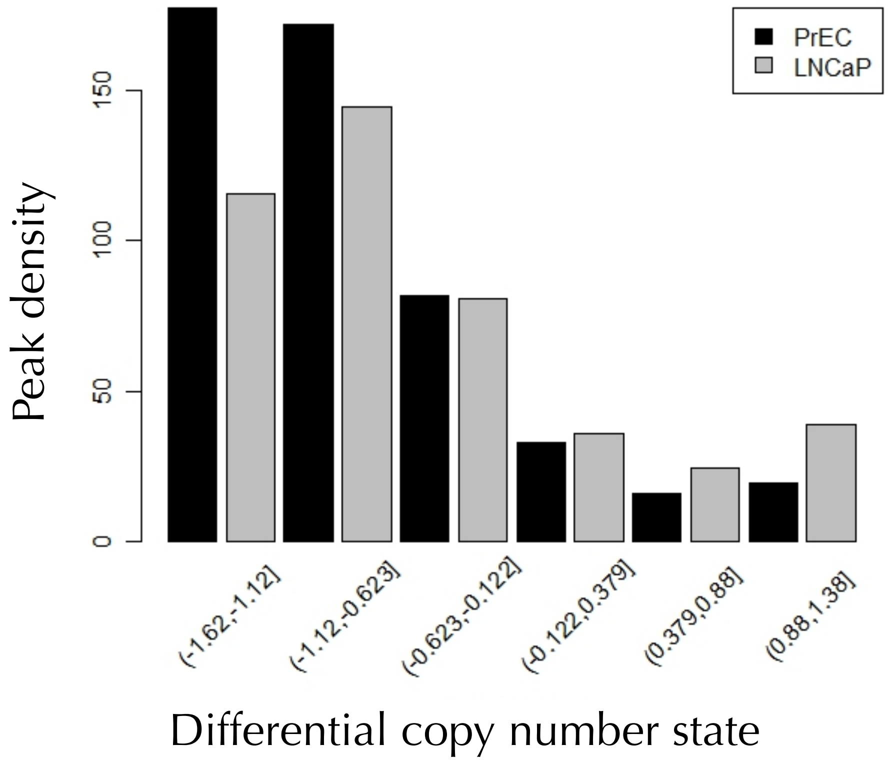


Figure 3

A closer view of some HeK4me3 and HNF4α peaks. These histograms present the counts of the 5' ends of the reads from the H3K4me3 and the HNF4α data, forming peaks on the forward (red) and reverse (blue) strand. The offset between them shows how the enclosed area corresponds to an enriched region. The plots are on a different scale to show the density of reads clearly and highlight the difference between the peaks formed by a histone mark and a transcription factor.

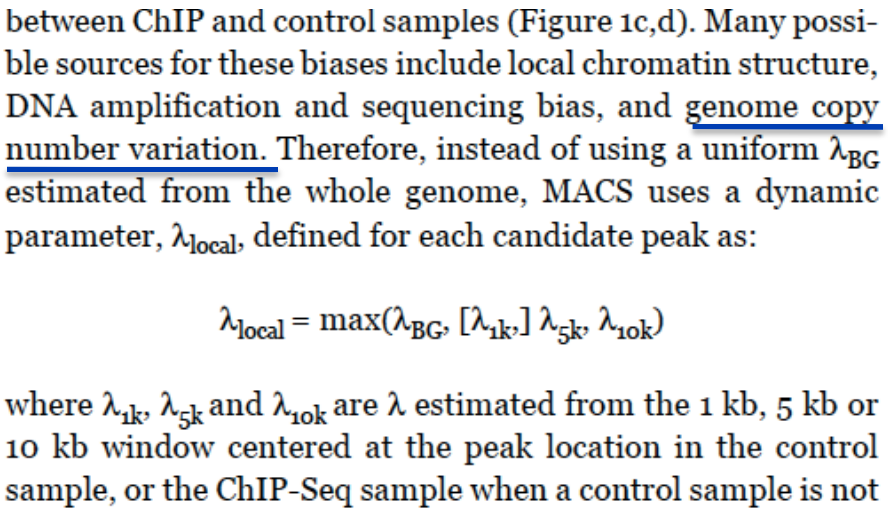
Copy number on region finding



For a peak caller (generally):
more reads = more peaks.

LNCaP = cancer
PrEC = normal

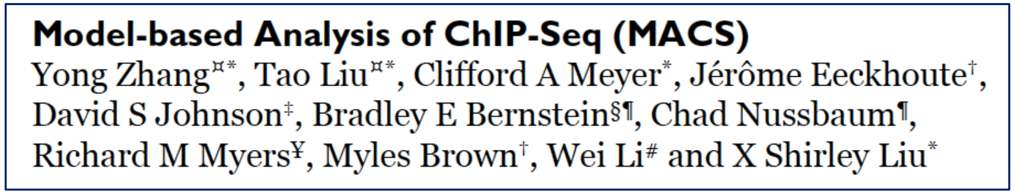
MACS, run with control (input) sample



between ChIP and control samples (Figure 1c,d). Many possible sources for these biases include local chromatin structure, DNA amplification and sequencing bias, and genome copy number variation. Therefore, instead of using a uniform λ_{BG} estimated from the whole genome, MACS uses a dynamic parameter, λ_{local}, defined for each candidate peak as:

$$\lambda_{local} = \max(\lambda_{BG}, [\lambda_{1k},]\ \lambda_{5k}, \lambda_{10k})$$

where λ_{1k}, λ_{5k} and λ_{10k} are λ estimated from the 1 kb, 5 kb or 10 kb window centered at the peak location in the control sample, or the ChIP-Seq sample when a control sample is not

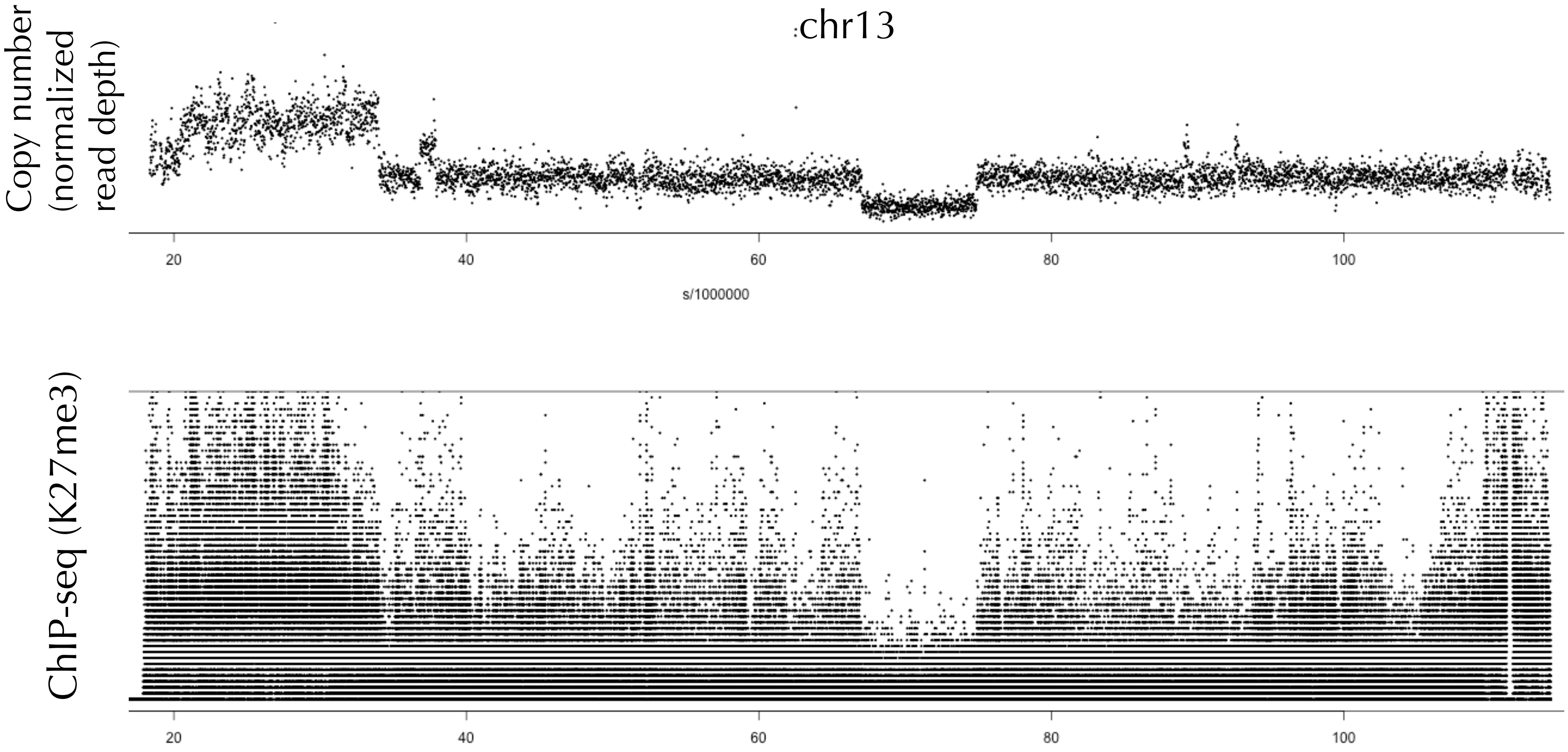


Model-based Analysis of ChIP-Seq (MACS)

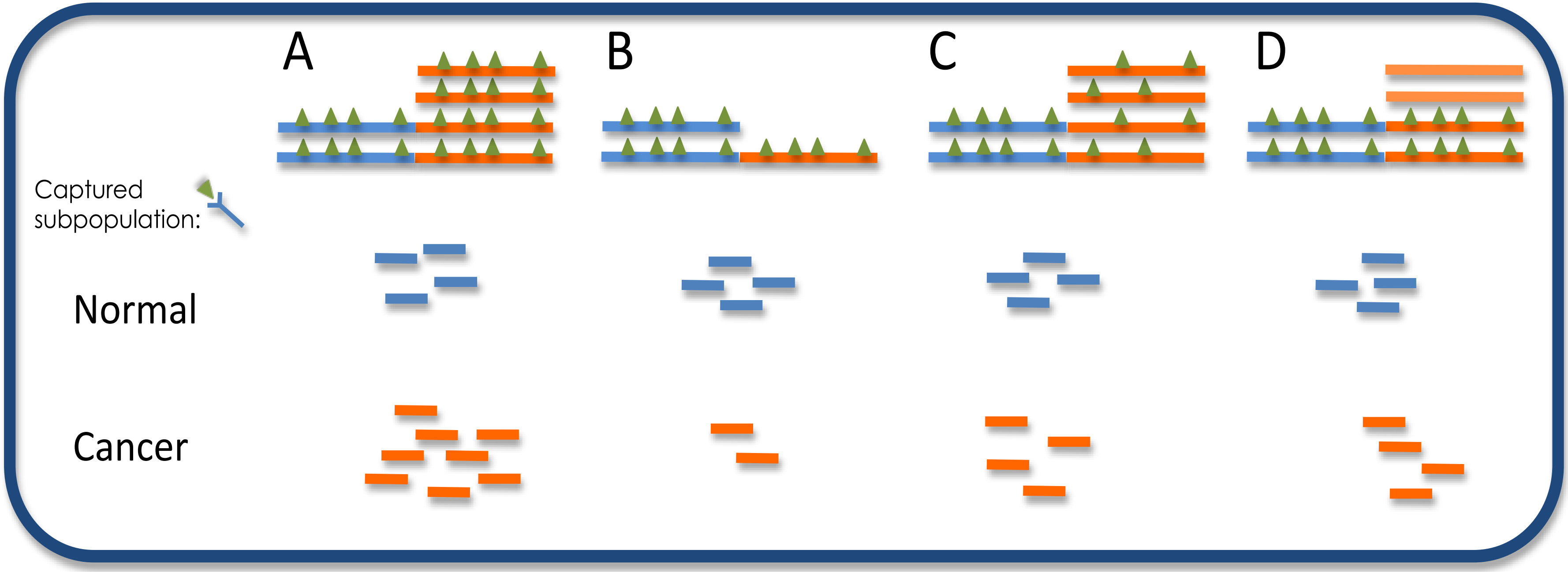
Yong Zhang[¤*], Tao Liu[¤*], Clifford A Meyer[*], Jérôme Eeckhoute[†], David S Johnson[‡], Bradley E Bernstein[§¶], Chad Nussbaum[¶], Richard M Myers[¥], Myles Brown[†], Wei Li[#] and X Shirley Liu[*]

QDNA-seq signal

= biology (copy number, enrichment) + technical effects

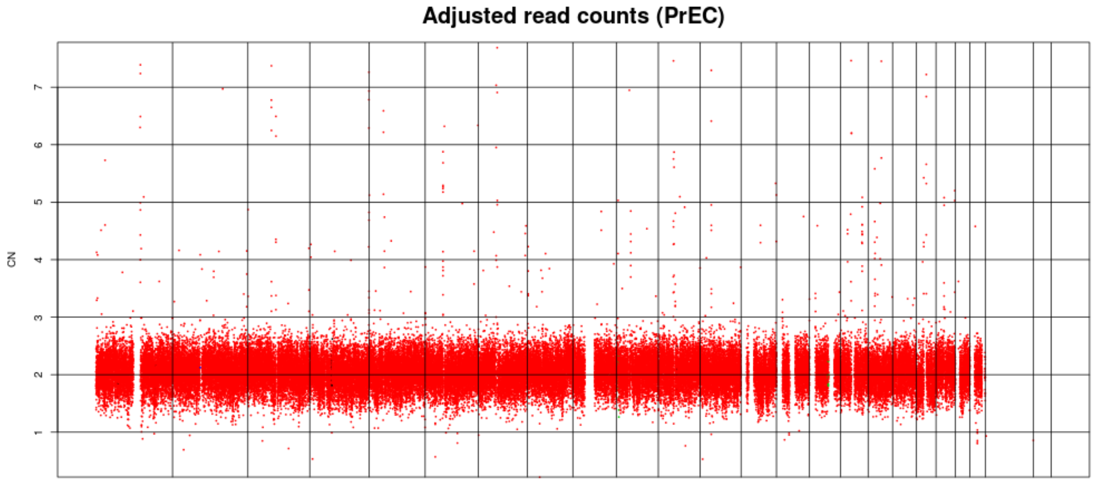


CNV affects differential comparisons: various scenarios

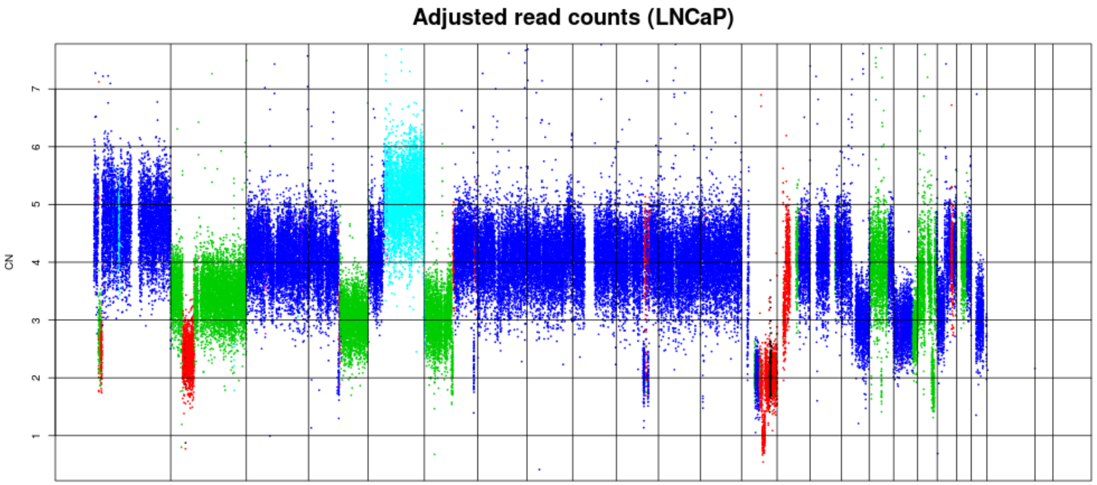


Test case: Compare cancer and normal epigenomes, considering changes in CNV

Read depths of gDNA sequencing, coloured by CNV calls by Affymetrix SNP 6.0 data

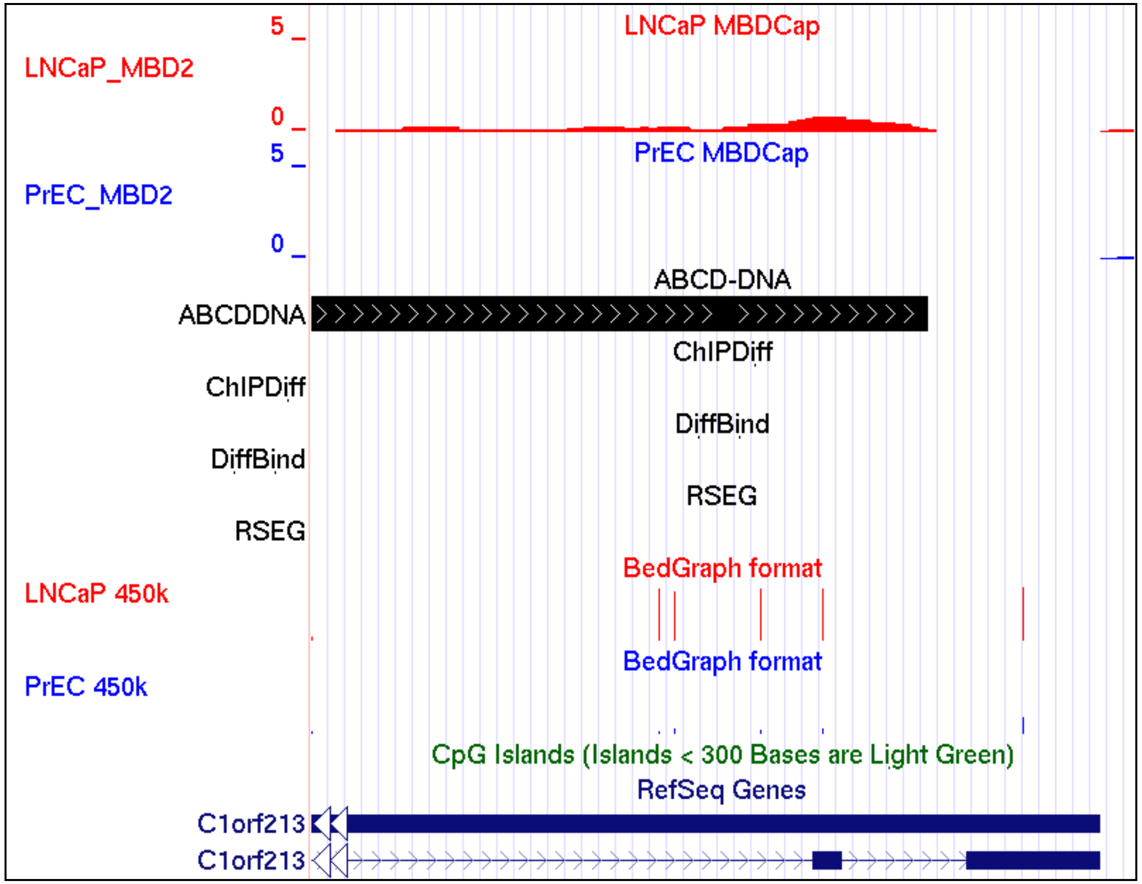


PrEC
(prostate epithelial cells)
Normal copy number



LNCaP
(prostate cancer cells)
primarily 4 copies, many variations

This region has 2 copies in normal PrEC cells and 2 copies in prostate cancer LNCaP cells (We normalize LNCaP=4 to PrEC=2, so this is effectively a net ***loss*** of copy number)

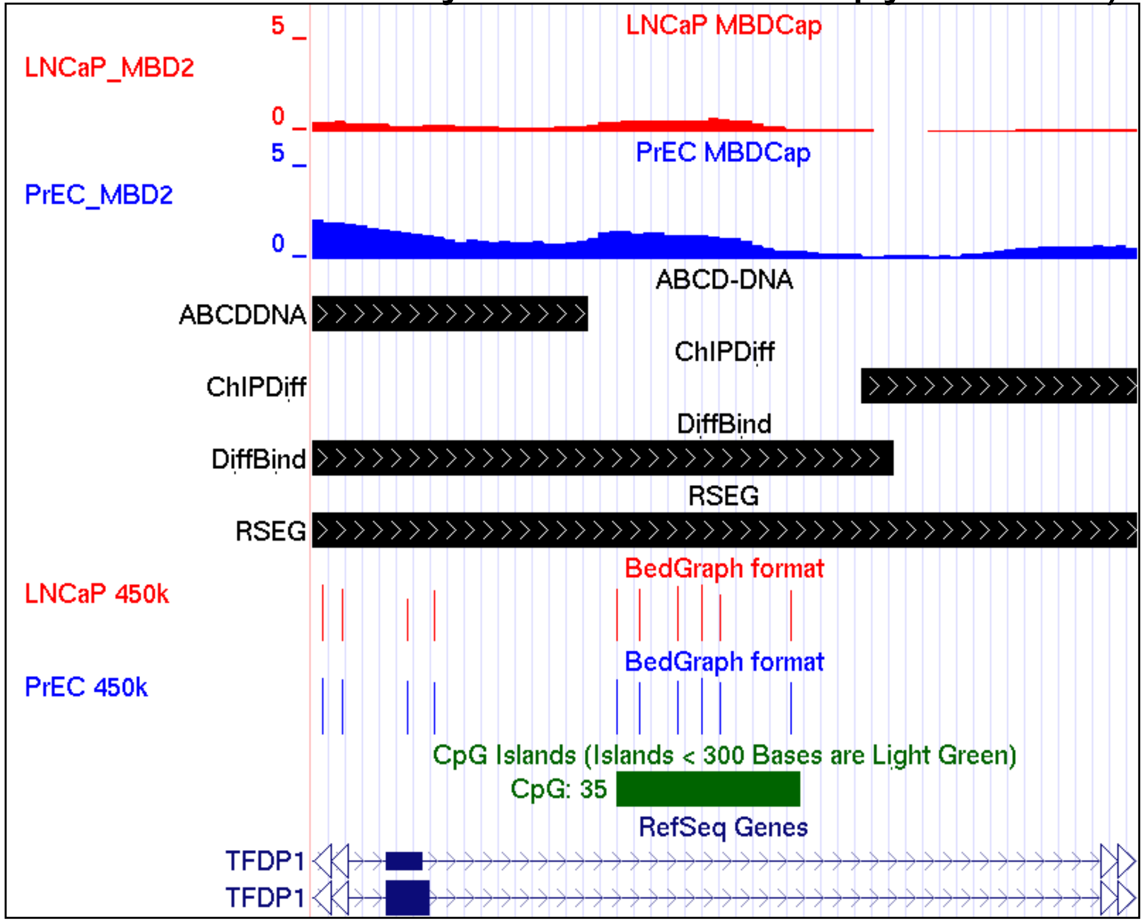


Existing tools: False *negative* due to CN "loss"

"Truth" from Reference dataset

Robinson et al. 2012 Genome Research

This region has 2 copies in normal PrEC cells and 2 copies in prostate cancer LNCaP cells (We normalize LNCaP=4 to PrEC=2, so this is effectively a net ***loss*** of copy number)



Existing tools: False *positive* due to CN "loss"

"Truth" from Reference dataset

Robinson et al. 2012 Genome Research

Statistical details of ABCD-DNA: offsets

We model read densities, $Y_{ij,}$ in a generalized linear model:

$$\log(E[Y_{ij}]) = O_{ij} + B_i X$$

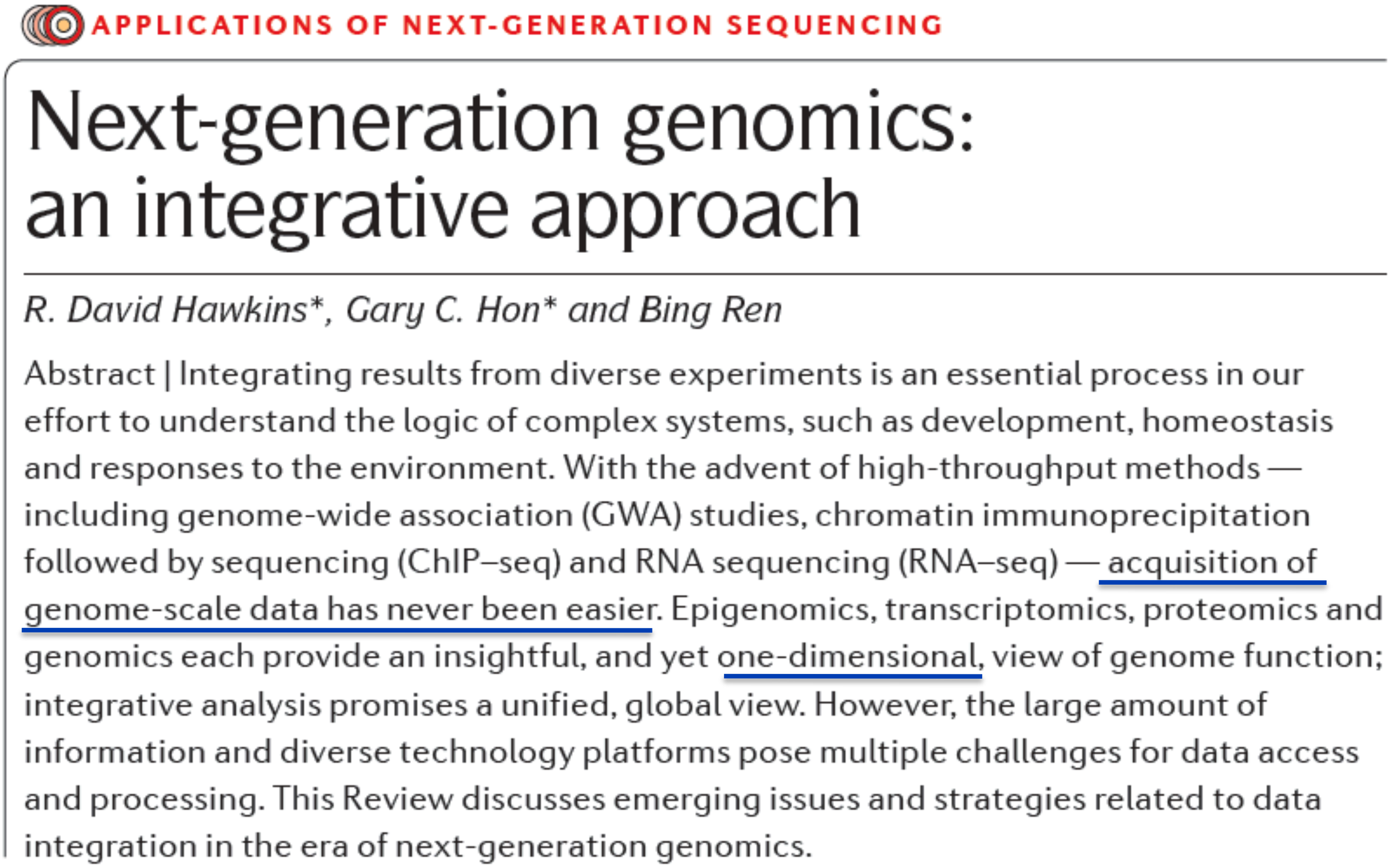
O_{ij} is an r x n matrix of **offsets**

X is an k x n **design matrix**

B_i is a r x k matrix of region-specific **coefficients**

O_{ij} can be decomposed into $\boxed{\log(CN_{ij})} + \boxed{\log(1\ D_j)}$

Using independent data (e.g. SNP array, gDNA-seq) to estimate offsets



APPLICATIONS OF NEXT-GENERATION SEQUENCING

Next-generation genomics: an integrative approach

R. David Hawkins*, Gary C. Hon* and Bing Ren

Abstract | Integrating results from diverse experiments is an essential process in our effort to understand the logic of complex systems, such as development, homeostasis and responses to the environment. With the advent of high-throughput methods — including genome-wide association (GWA) studies, chromatin immunoprecipitation followed by sequencing (ChIP–seq) and RNA sequencing (RNA–seq) — acquisition of genome-scale data has never been easier. Epigenomics, transcriptomics, proteomics and genomics each provide an insightful, and yet one-dimensional, view of genome function; integrative analysis promises a unified, global view. However, the large amount of information and diverse technology platforms pose multiple challenges for data access and processing. This Review discusses emerging issues and strategies related to data integration in the era of next-generation genomics.

Hawkins et al. (2010) Nature Reviews Genetics

Expression outcome is related to several factors

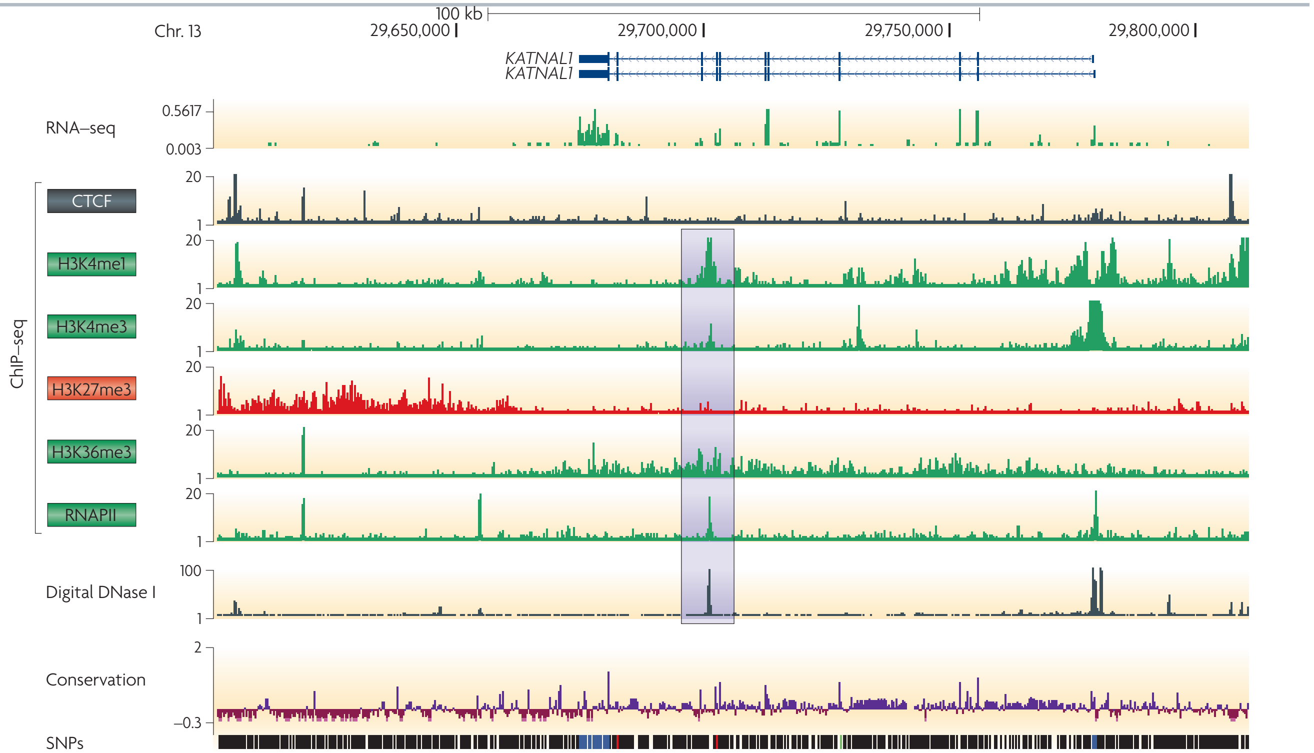
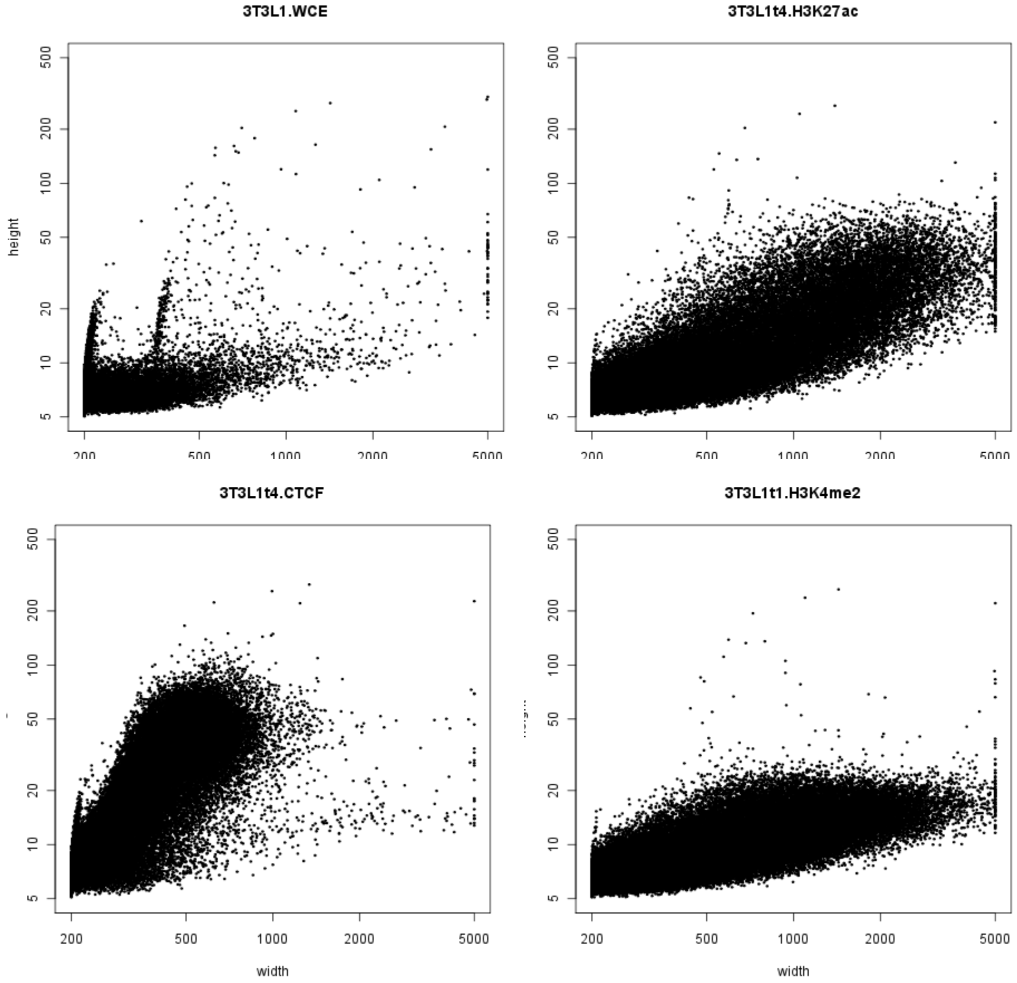


Figure 3 | **Data visualization.** The University of California-Santa Cruz (UCSC) Genome Browser is a tool for viewing genomic data sets. A vast amount of data is available for viewing through this browser. This example from the browser shows numerous data types in K562 cells from the ENCODE Consortium. A random gene was selected — katanin p60 subunit A-like 1 (*KATNAL1*) — that shows several points that can be identified by using this tool. The promoter has a typical chromatin structure (a peak of histone 3 lysine 4 trimethylation (H3K4me3) between the bimodal peaks of H3K4me1), is bound by RNA polymerase II (RNAPII) and is DNase hypersensitive. The gene is transcribed, as indicated by RNA sequencing (RNA–seq) data, as well as H3K36me3 localization. The gene lies between two CCCTC-binding factor (CTCF)-bound sites that could be tested for insulator activity. An intronic H3K4me1 peak (highlighted) predicts an enhancer element, corroborated by the DNase I hypersensitivity site peak. There is a broad repressive domain of H3K27me3 downstream, which could have an open chromatin structure in another cell type.

Heights and widths of "peaks" across ChIP-seq datasets



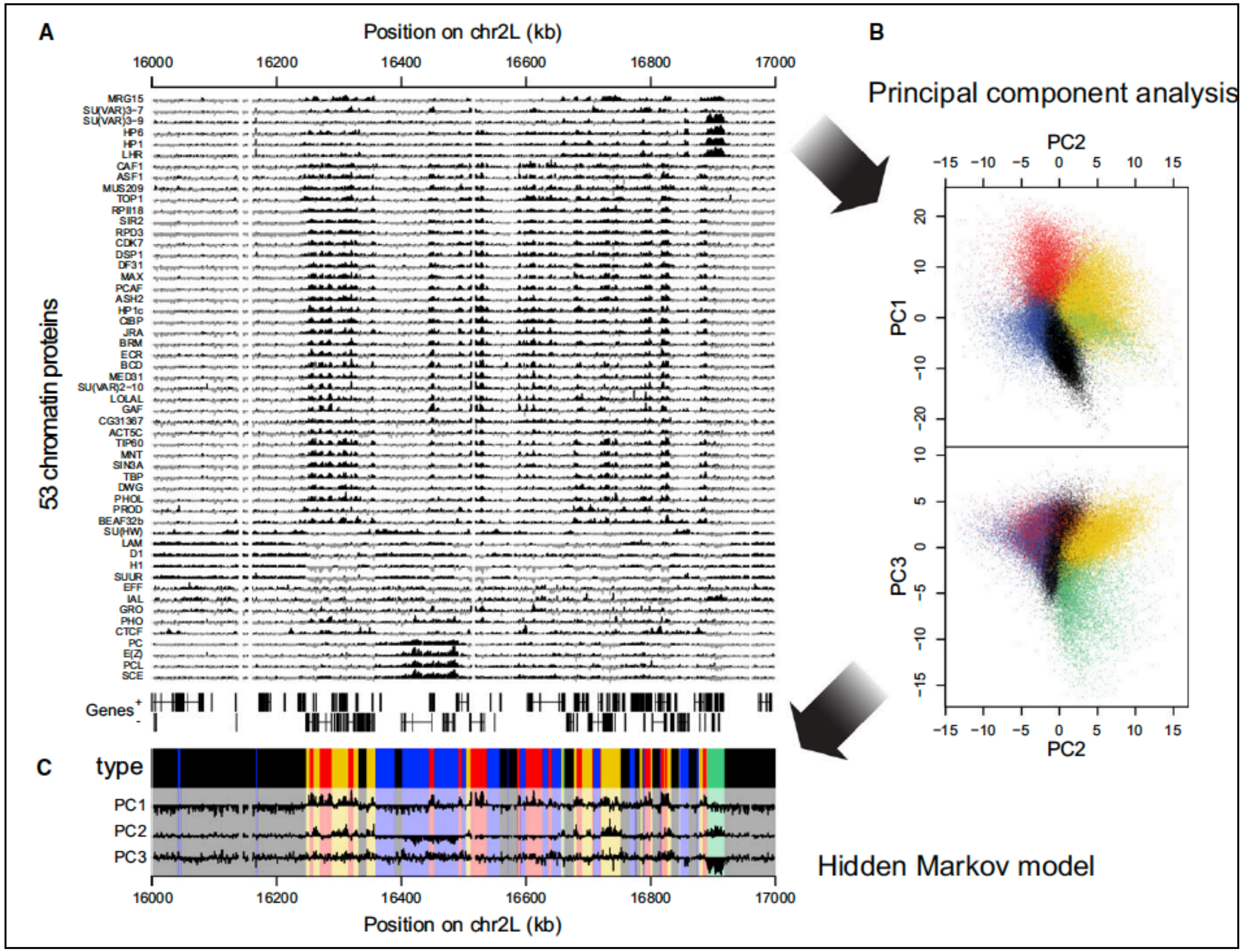
Exploratory analyses

53 chromatin factors (ChIP-seq)

Compression to 3 principal components

Learn HMM

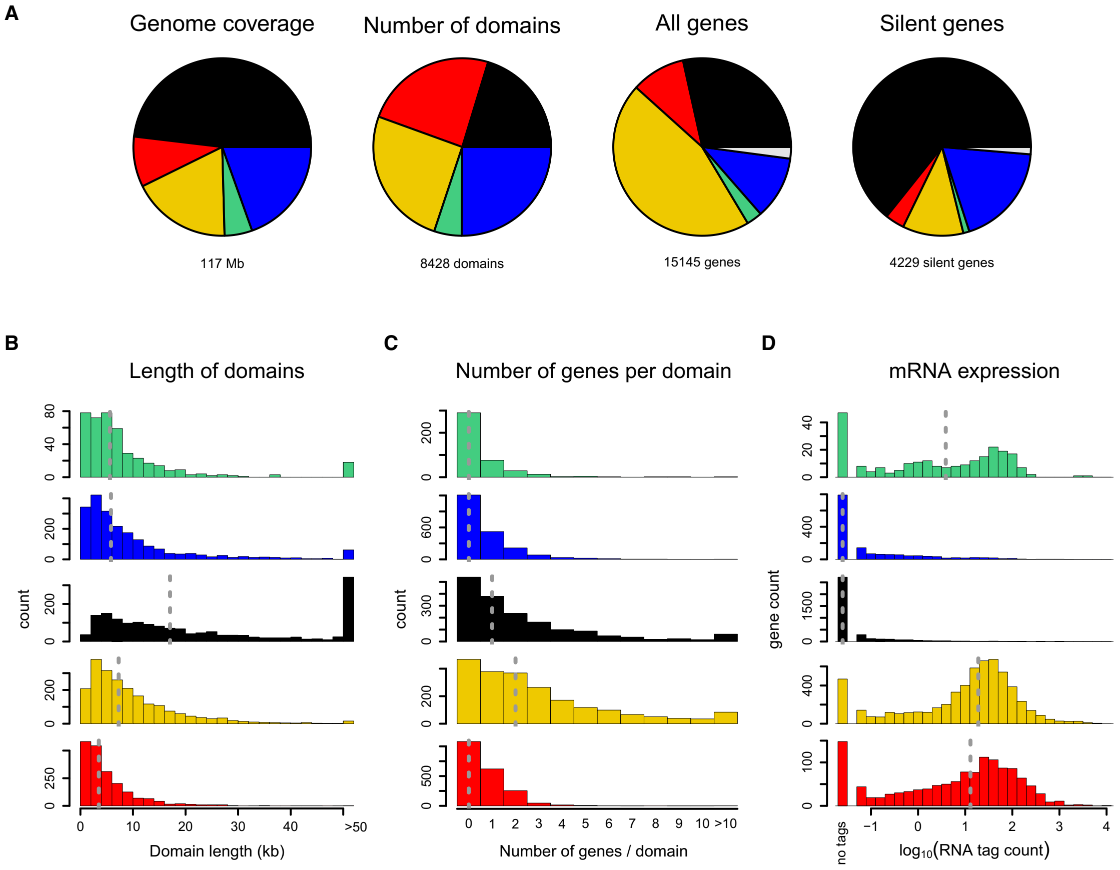
Every region of the genome partitioned into 5 “states” (here, assigned a colour)



Filion et al. Cell 2010

Exploratory analyses

“Colours” are reflective of various features



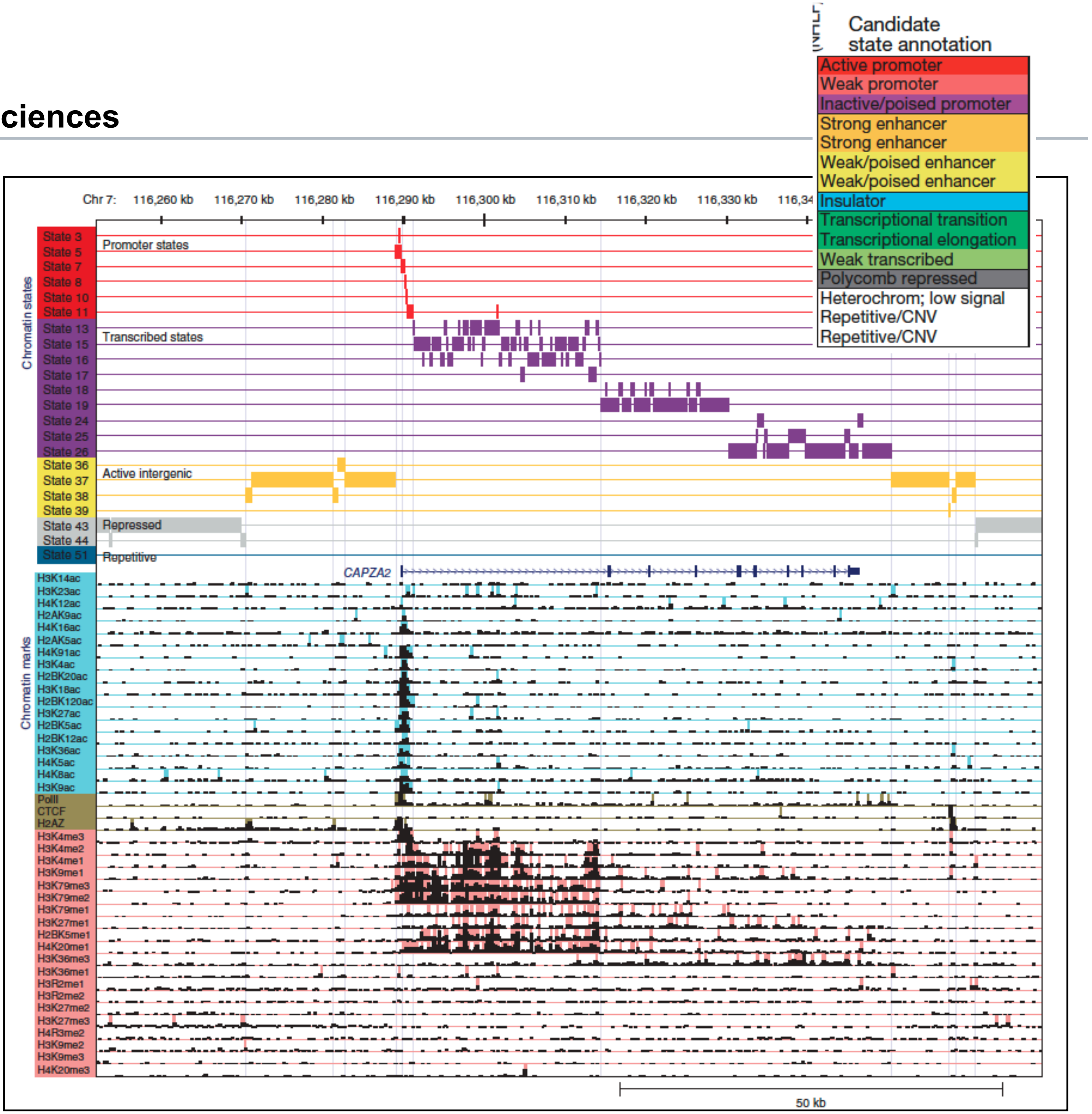
Filion et al. Cell 2010

Exploratory analyses

No compression

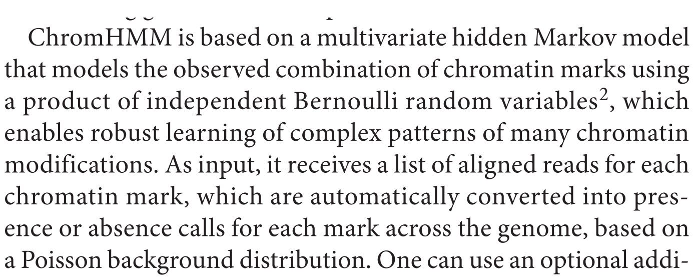
Every 200bp region of the genome is binarized based on a background model

Multivariate HMM is trained; genome is partitioned into 15 states

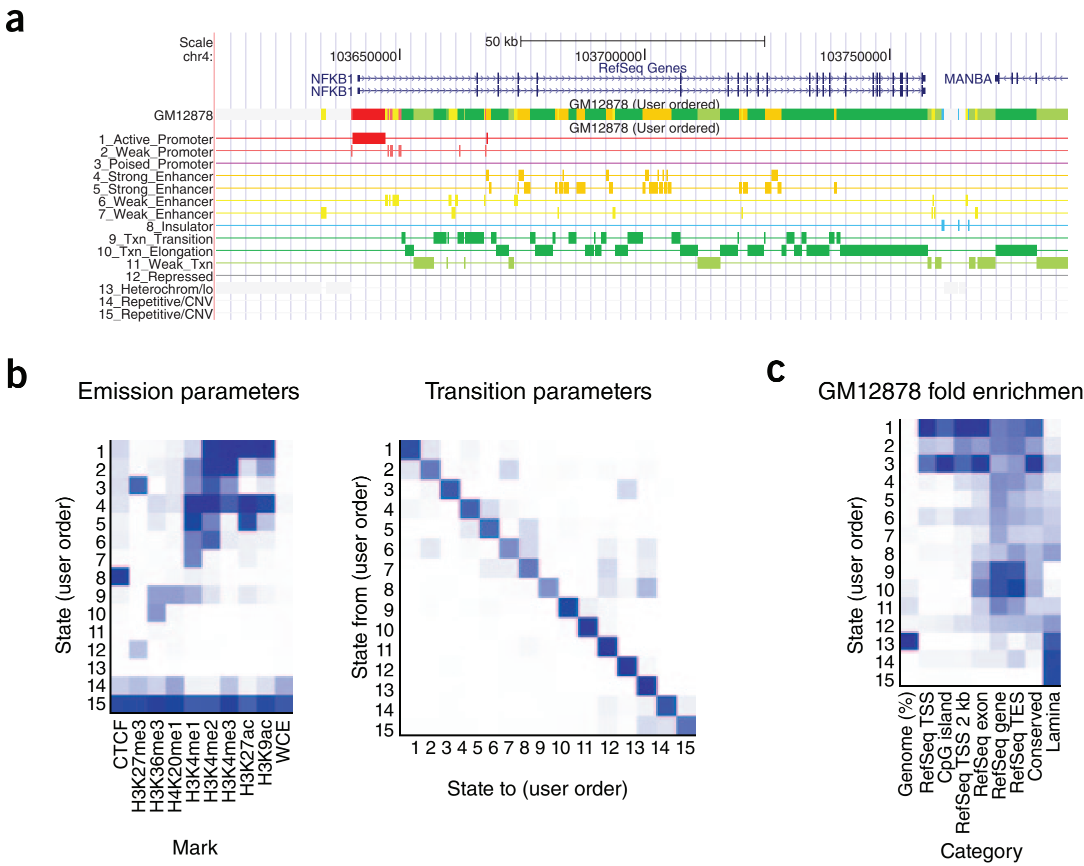


Ernst et al., Nature 2010
Ernst and Kellis, Nature Biotech 2010

ChromHMM

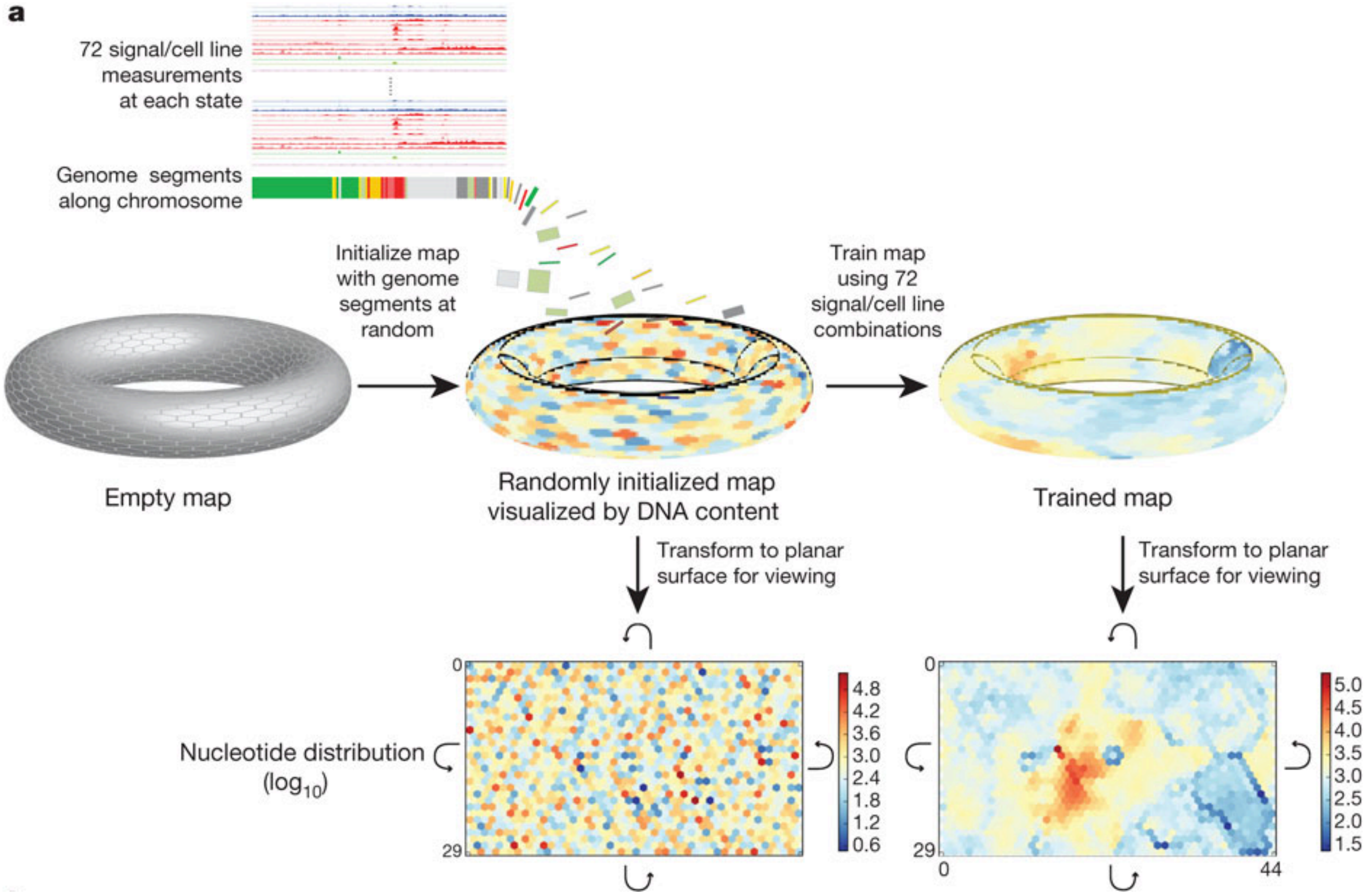


ChromHMM is based on a multivariate hidden Markov model that models the observed combination of chromatin marks using a product of independent Bernoulli random variables[2], which enables robust learning of complex patterns of many chromatin modifications. As input, it receives a list of aligned reads for each chromatin mark, which are automatically converted into presence or absence calls for each mark across the genome, based on a Poisson background distribution. One can use an optional addi-



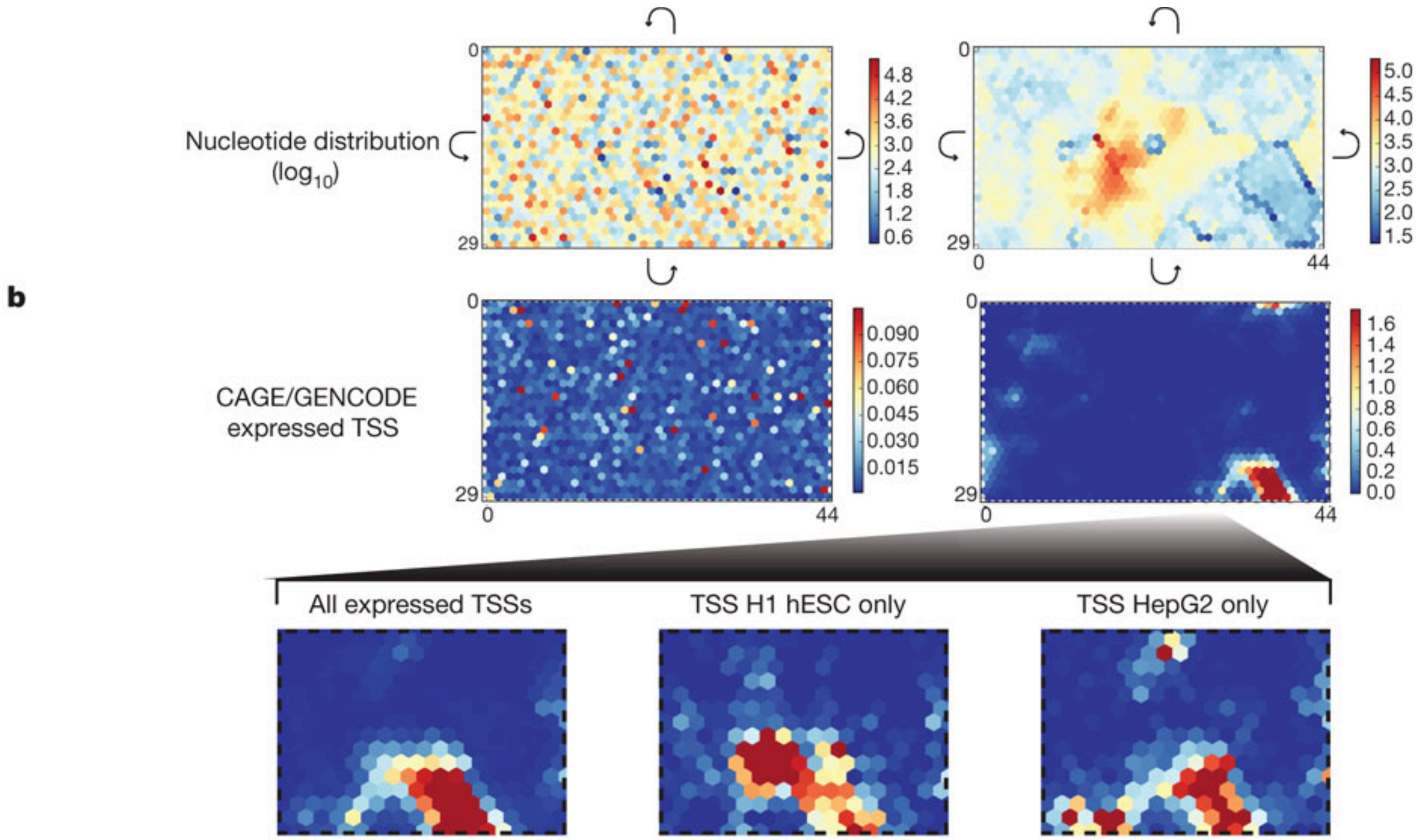
Ernst and Kellis, Nature Methods (March) 2012

Self-organizing map “compression”


a
72 signal/cell line measurements at each state
Genome segments along chromosome
Initialize map with genome segments at random
Train map using 72 signal/cell line combinations
Empty map
Randomly initialized map visualized by DNA content
Trained map
Transform to planar surface for viewing
Transform to planar surface for viewing
Nucleotide distribution (log10)
0
29
4.8
4.2
3.6
3.0
2.4
1.8
1.2
0.6
0
29
0
44
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5

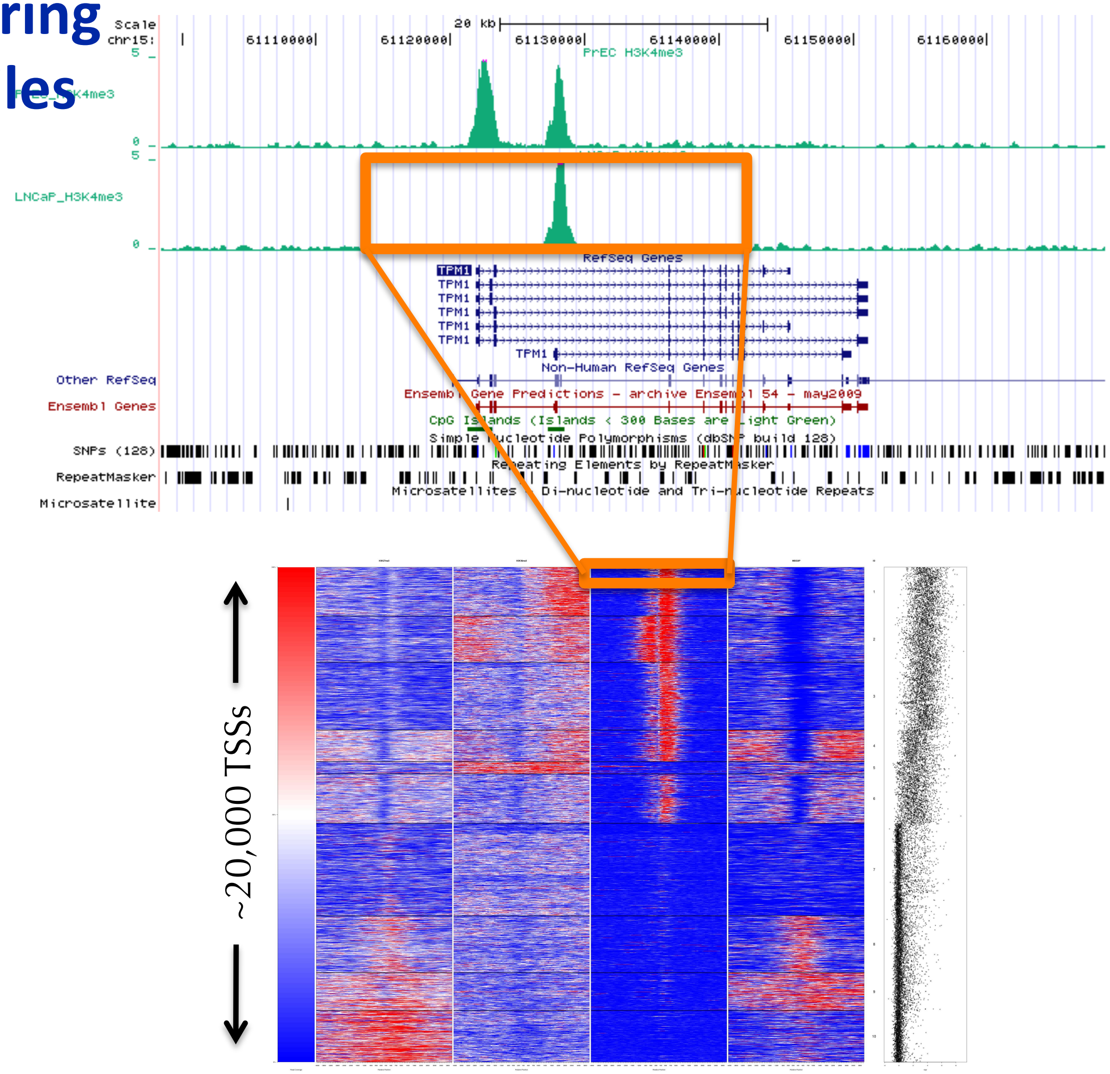
Dunham et al. 2012 Nature

SOM to other features



Dunham et al.
2012 Nature

Exploratory analysis: clustering combined epigenomic profiles



Calculate coverage around features of interest (here, TSSs)

Cluster collective epigenomeic signal using k-means, display as heatmap/line, order clusters by expression

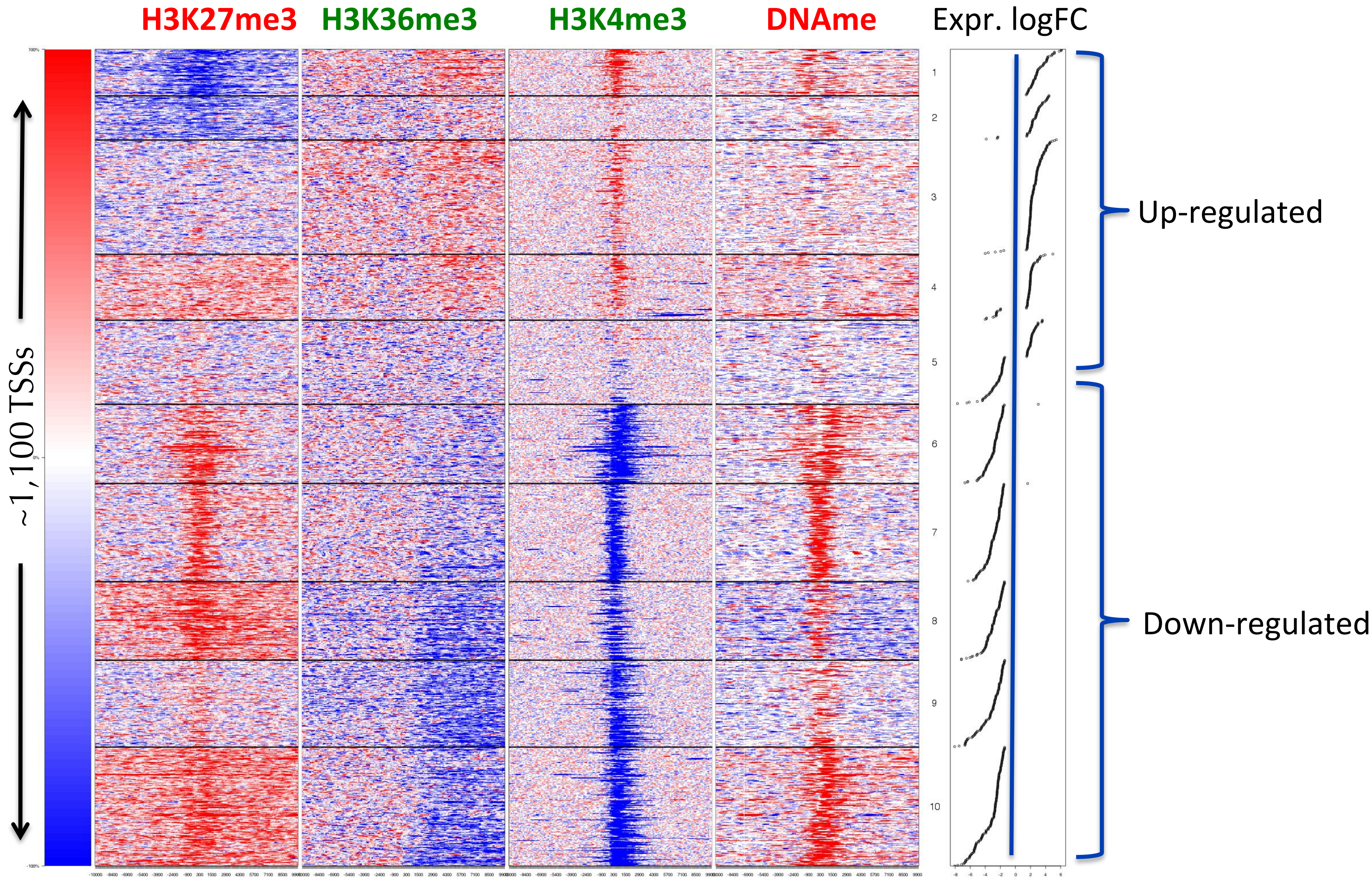
Overlay expression, order clusters by median

Available in Repitools package*

*Statham et al. Bioinformatics 2010

Clustering changes (just DE genes)

Repitools


H3K27me3
H3K36me3
H3K4me3
DNAme
Expr. logFC
~1,100 TSSs
Up-regulated
Down-regulated