University of Zurich[UZH]

Institute of Molecular Life Sciences

Epigenomics

- Part 1: Intro to epigenetics
- **Part 2: Computational methods**

Mark D. Robinson, Statistical Genomics, IMLS

Overview of this lecture

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

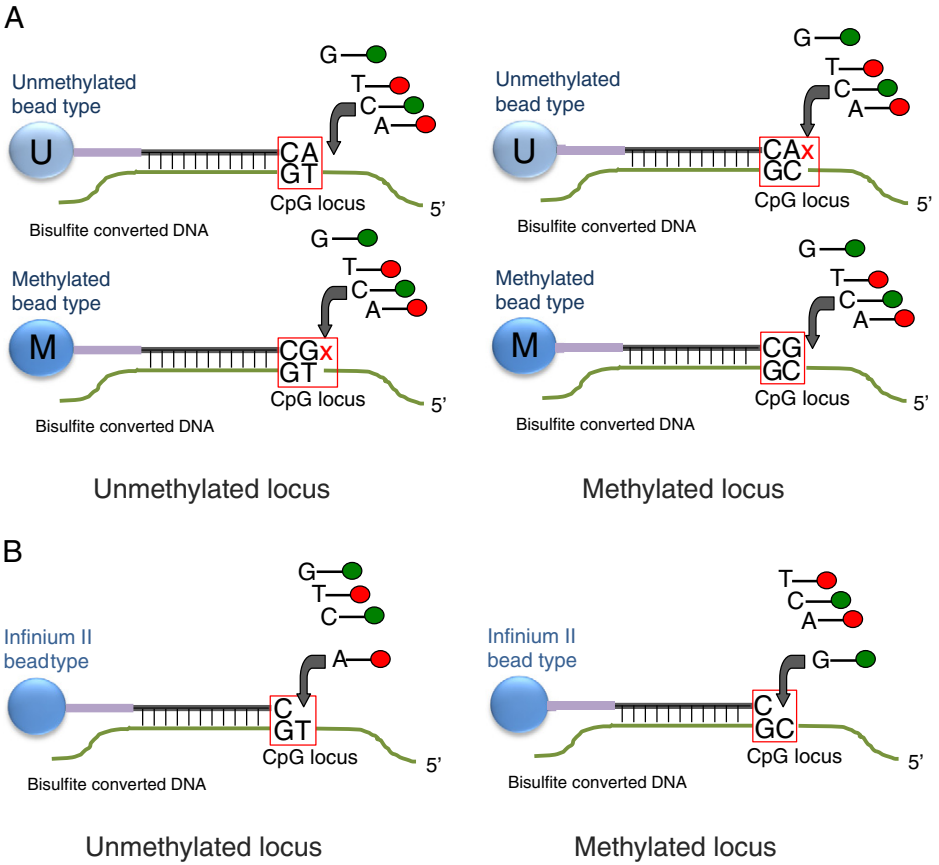
Analysis of 450k arrays

For each CpG site of interest, the array measures signal for methylated (M) and unmethylated (U)

Consensus methylation level (beta value – B) estimated as:

B = M / (M+U+e)

Differential statistics often done, as with 2-colour gene expression microarrays on log(M/U)



Analysis of 450k arrays

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

Normalization issues for different probe types (much current research)

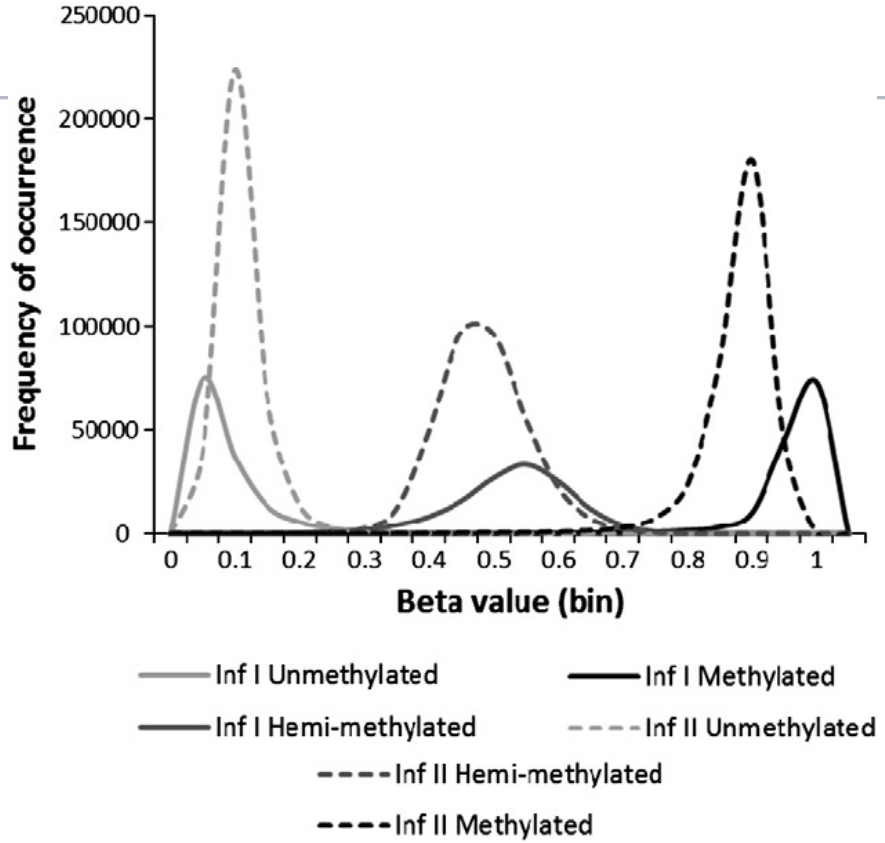


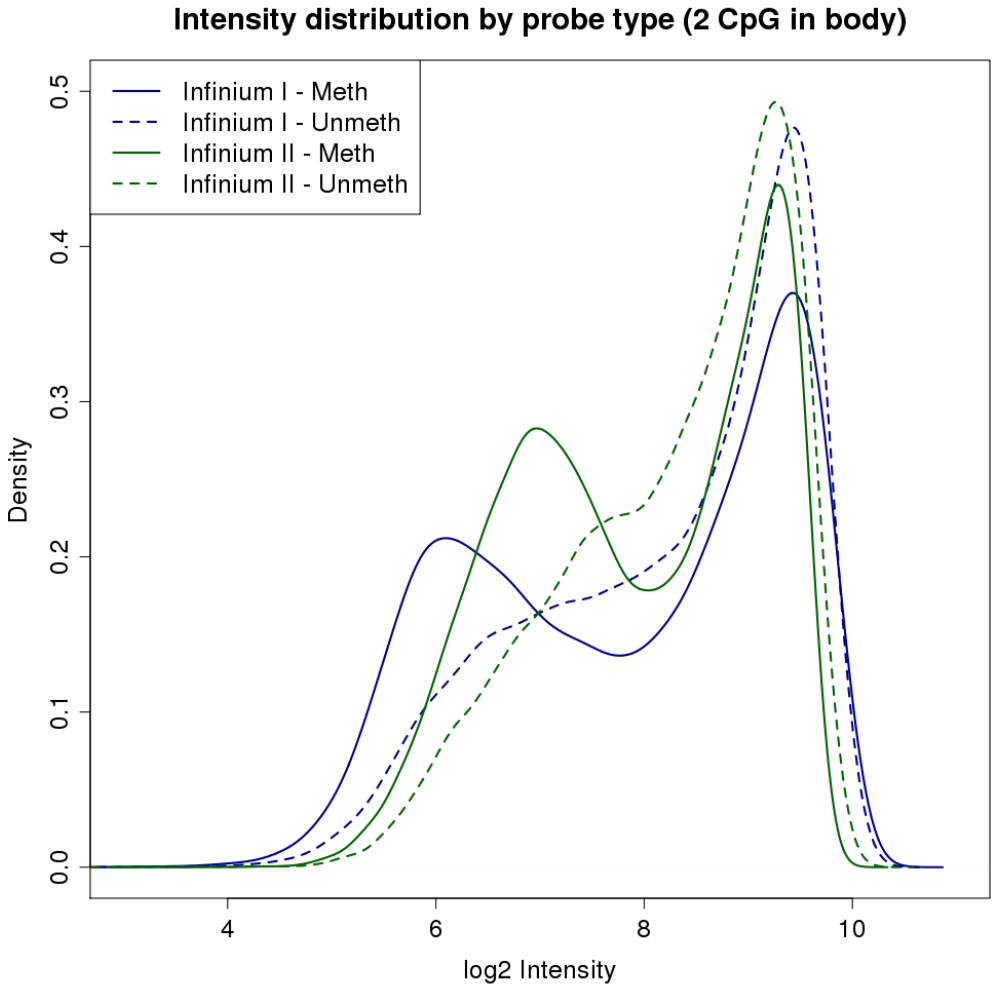
Fig. 3. Distribution of Methylation values for Infinium I and Infinium II loci. Unmethylated (U), Hemi-methylated (H), and Methylated (M) reference standards were created from Coriell genomic DNA sample as discussed in Methods. Note slightly different performance of Infinium I and Infinium II assays in regard to beta value distribution.

Intensity distribution of probes with 2 CpGs

Not only are type I and type II probes distributed very differently, the presence of CpG sites (which can be unmethylated or methylated) can affect the observed signal.

Also, present of SNPs in probe may differentially affect human samples

SWAN: subset within array normalization



Methods for differential methylation

Methods for differential methylation of 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)

charm::dmrFind()

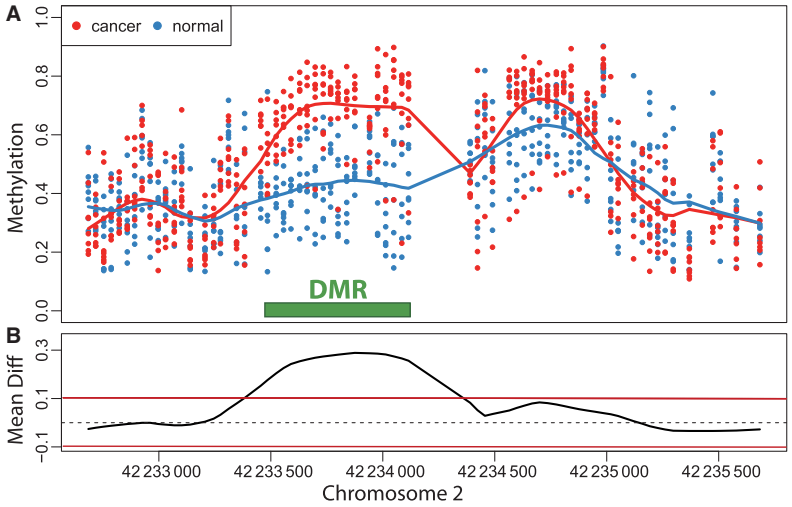


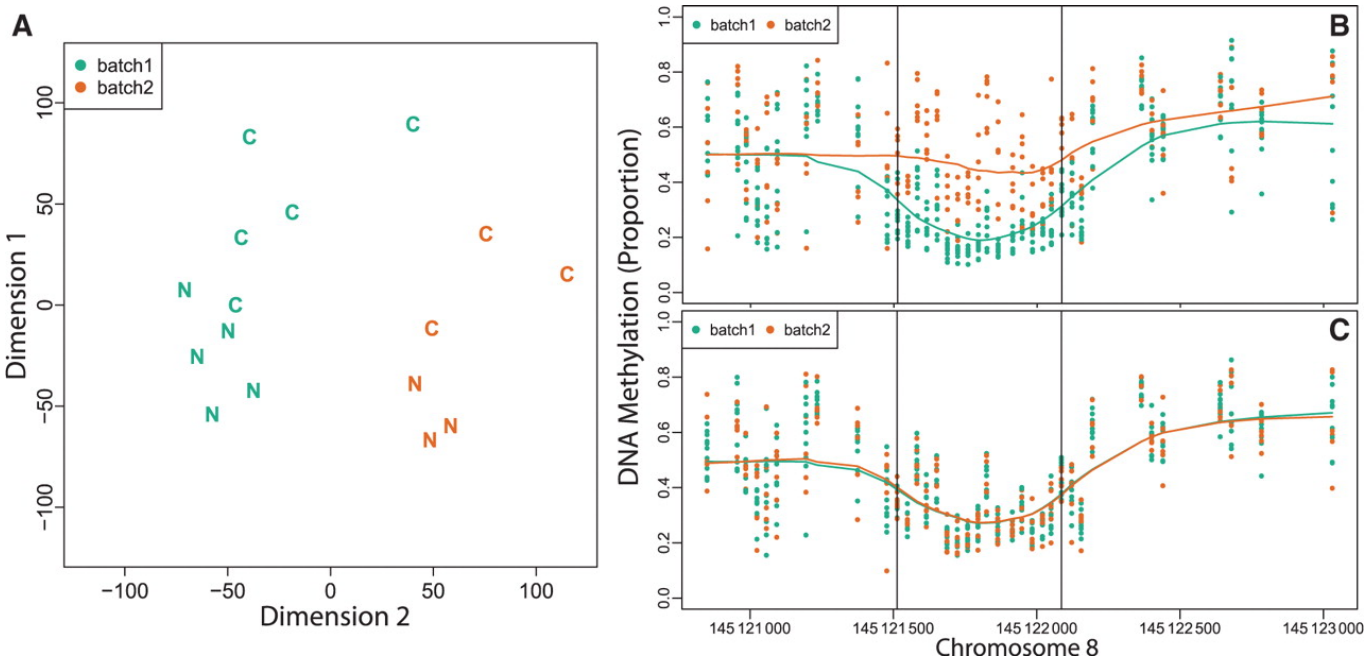
Figure 1 Example of a differentially methylation region (DMR). (A) The points show methylation measurements from the colon cancer dataset plotted against genomic location from illustrative region on chromosome 2. Eight normal and eight cancer samples are shown in this plot and represented by eight blue 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

Methods for differential methylation

Batch effects are ever-present

charm::dmrFind()



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

Probe-level methylation → region methylation

i – individual
j – loci

Includes surrogate variable analysis

$$Y_{ij} = \mu(t_j) + \beta(t_j)X_i + \sum_{k=1}^{p} \Upsilon_k(t_j)Z_{i,k} + \sum_{l=1}^{q} a_{l,j}W_{i,l} + \varepsilon_{i,j}$$

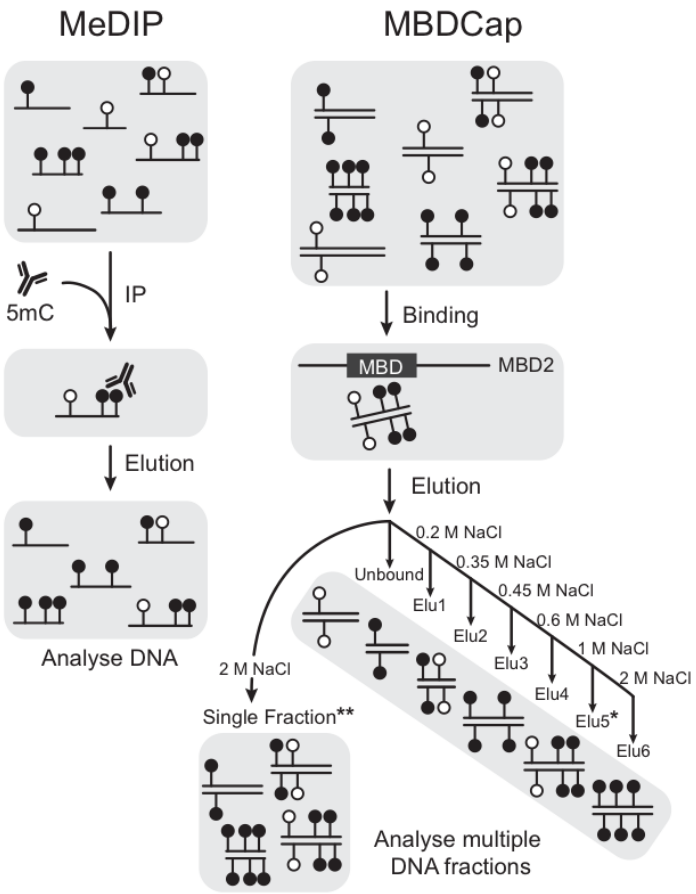
Outcome of interest (e.g. cancer versus normal)

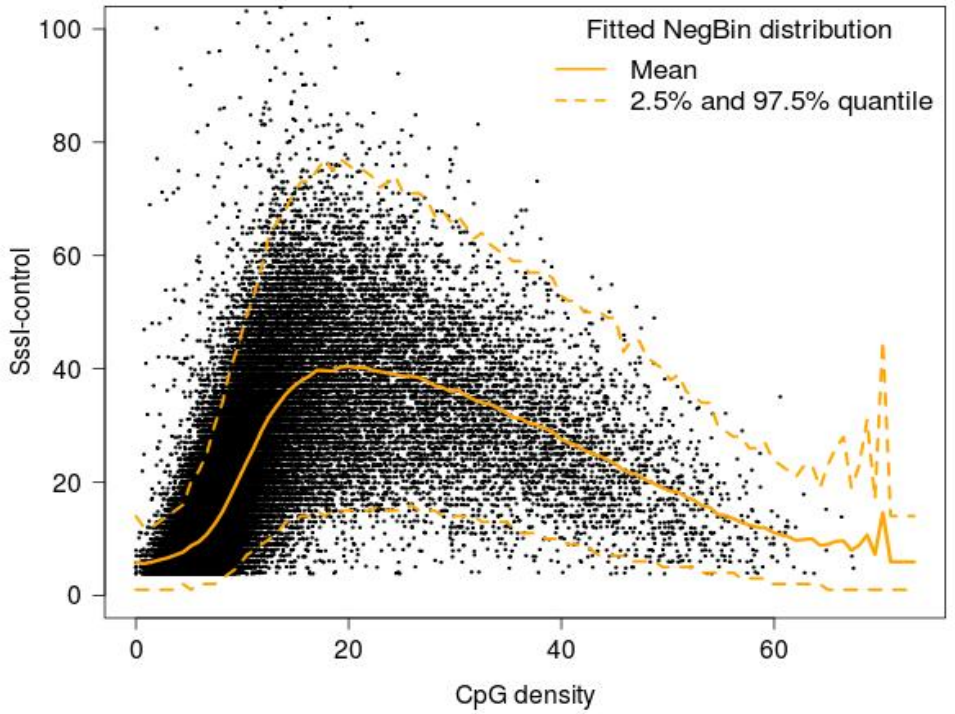
Measured confounders

Unmeasured confounders

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

Methods for affinity enrichment (MeDIP-seq, MBD-seq) DNA methylation data





Efficiency of capture in a fully methylated sample, is strongly associated with CpG density.

BATMAN - Bayesian tool for methylation analysis

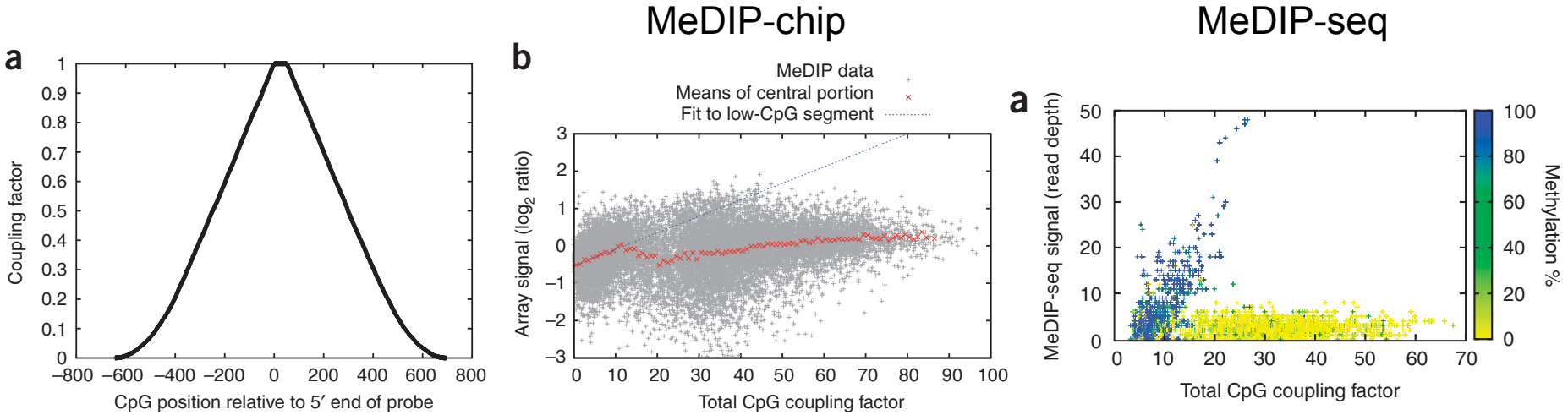


Figure 1 Calibration of the Batman model against MeDIP-chip data. (**a**) Estimated CpG coupling factors for a MeDIP-chip experiment as a function of the distance between a CpG dinucleotide and a microarray probe. (**b**) Plot of array signal against total CpG coupling factor, showing a linear regression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

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.

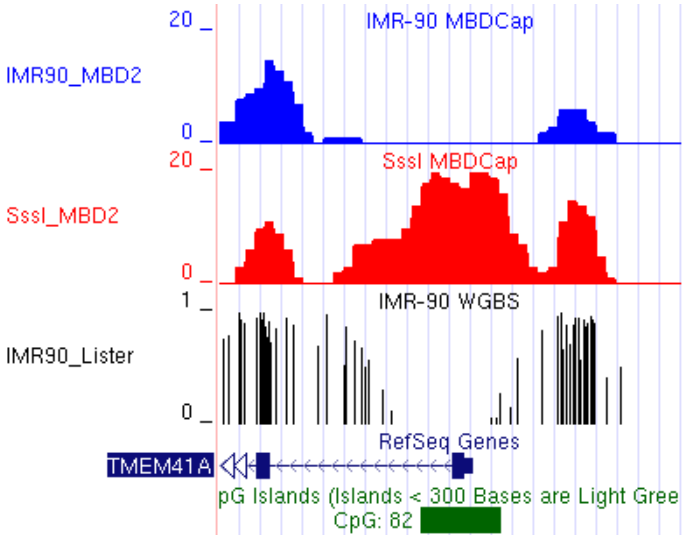
Some potential disadvantages:

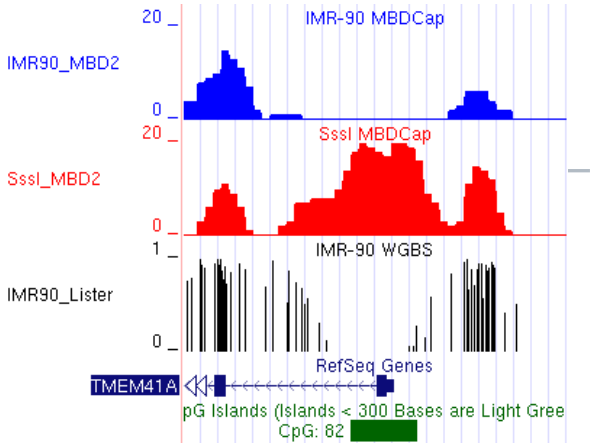
1. No reads = no DNA methylation *or* assay doesn't capture the region
2. MCMC is very computationally intensive (10-15h per chromosome)

Using SssI control to improve estimation

A new method is desired that:

- is computationally light
- uses a control to i) improve estimation; ii) know where the assay is efficient
- can give variance estimates
- account for copy number





Using SssI control to improve estimation

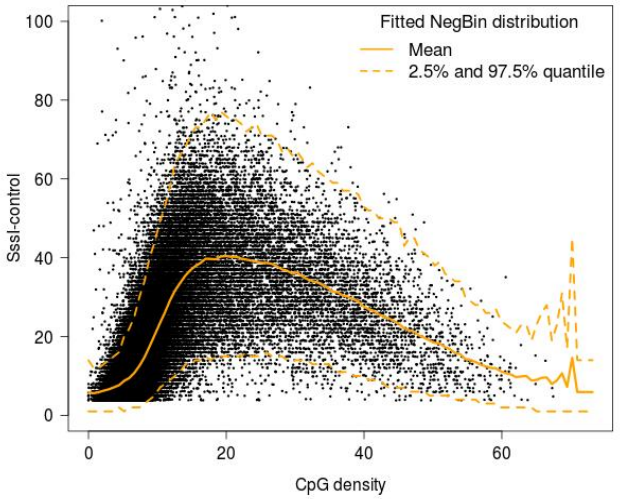
Model

$$y_{i,\text{IMR90}} | \mu_i, \lambda_i \sim \text{Poisson}(\text{const} \times \mu_i \times \lambda_i); \qquad y_{i,\text{SssI}} | \lambda_i \sim \text{Poisson}(\lambda_i)$$

const: offset for the (effective) relative sequencing depth, CNV, etc.

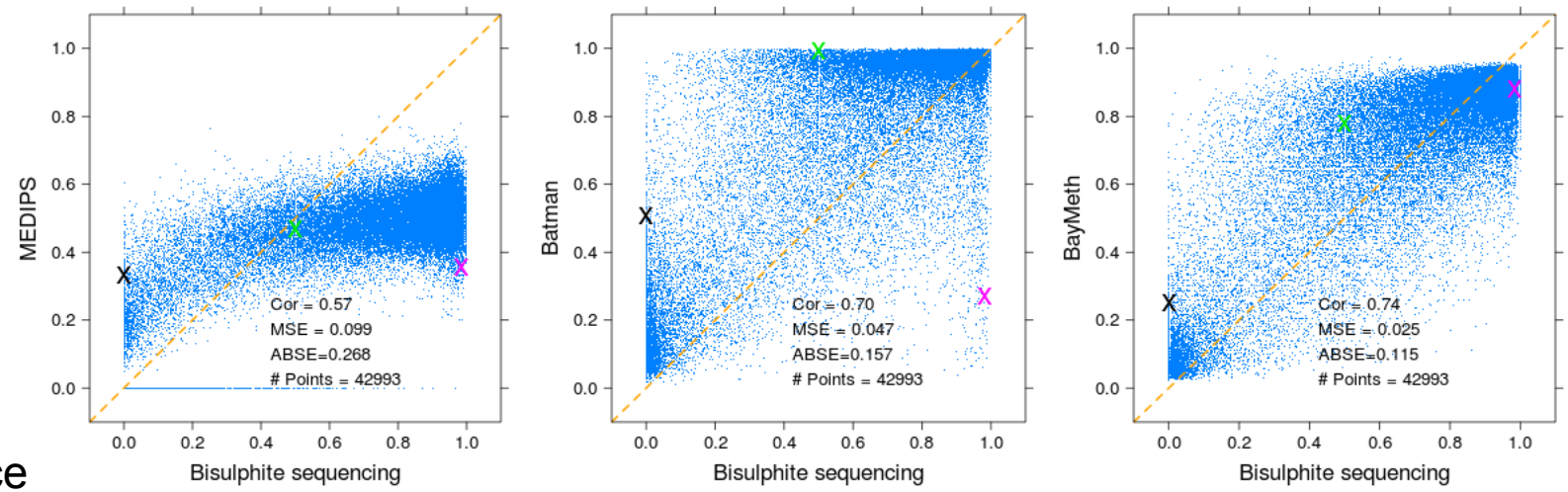
λ_i: region-specific read density, and

μ_i: the regional methylation level (Parameter of interest)



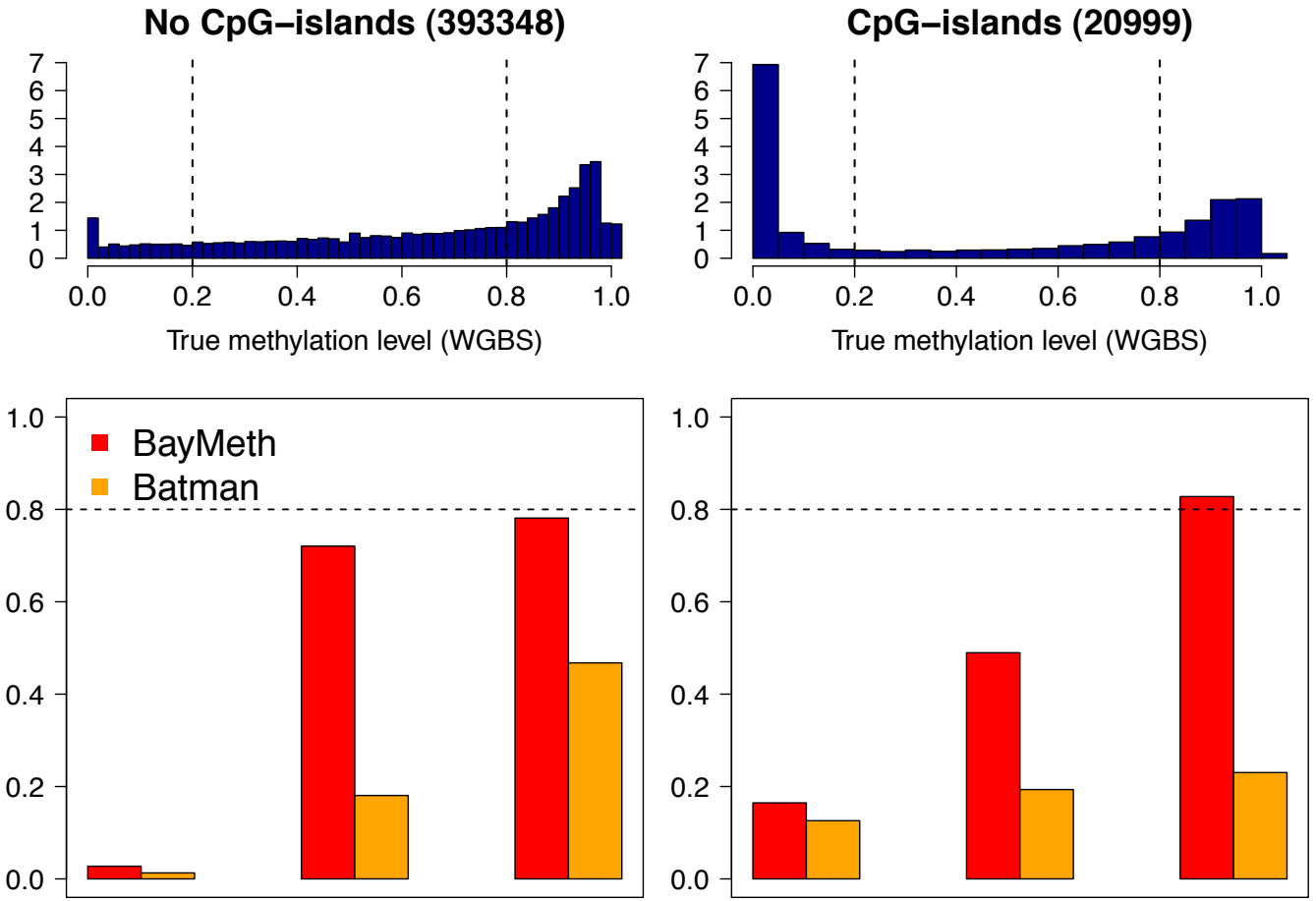
Using SssI control to improve estimation

Better overall prediction performance



Using SssI control to improve estimation

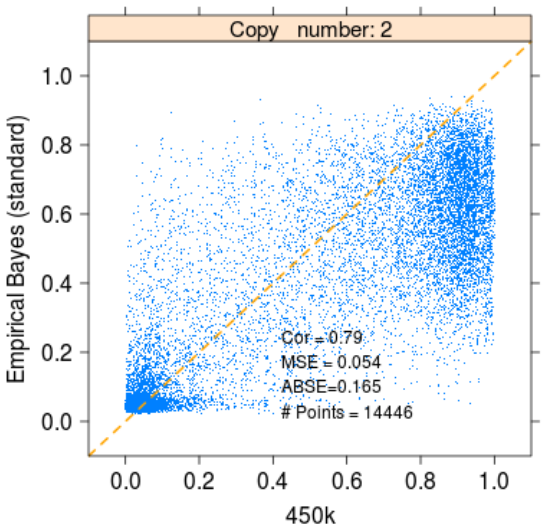
Meaningful variance estimates



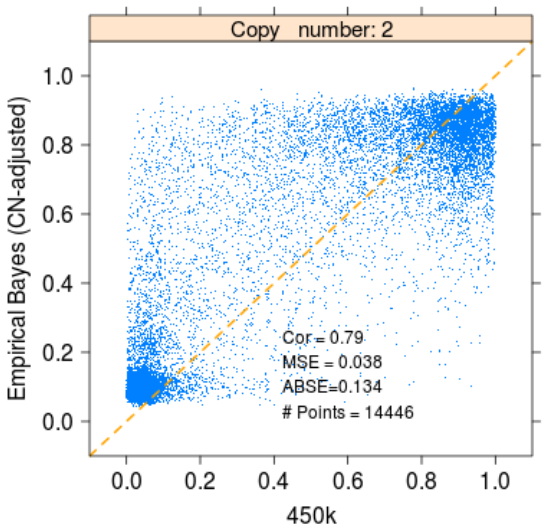
Using Sssl control to improve estimation

Can improve even further by integrating CNV information

Unadjusted



Adjusted



Pipelines: sequencing reads to data analysis for ChIP-seq

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

Downstream informatic analyses are specific to the scientific question.

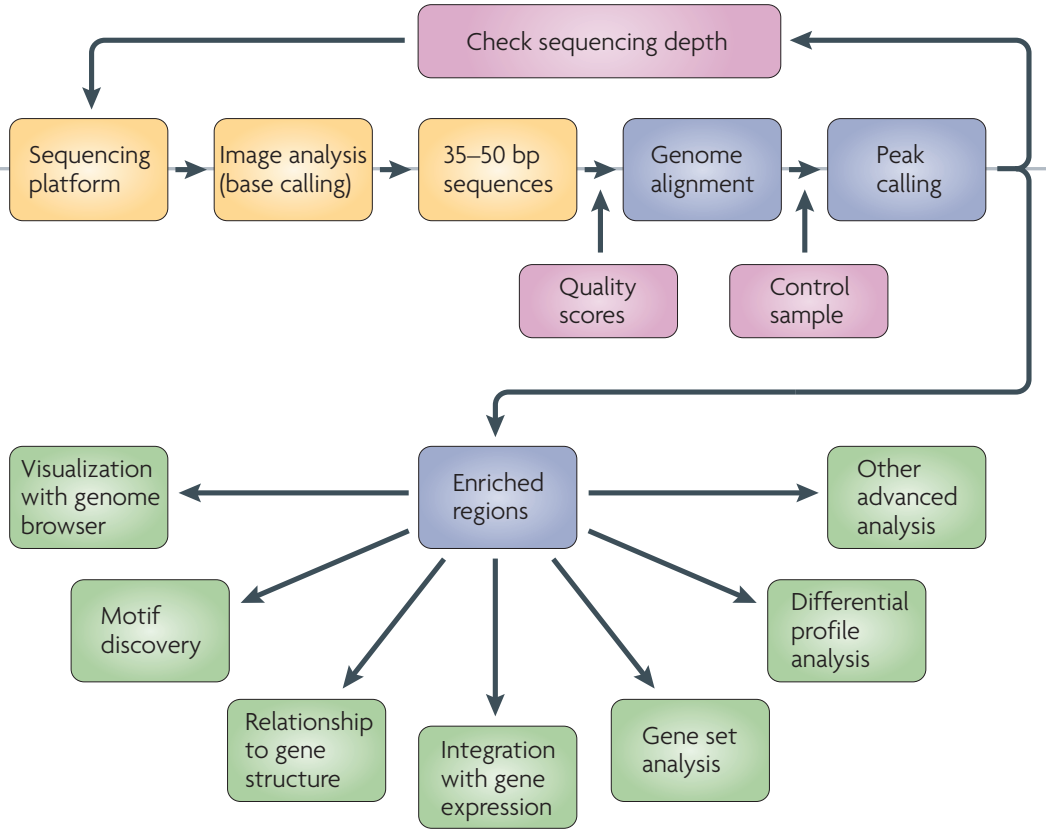


Figure 4 | **Overview of ChIP–seq analysis.** The raw data for chromatin immunoprecipitation followed by sequencing (ChIP–seq) analysis are images from the next-generation sequencing platform (top left). A base caller converts the image data to sequence tags, which are then aligned to the genome. On 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.

Repitools

- Exploratory analysis and visualizations for {ChIP/MBD/MeDIP}-{chip/seq}
- Statistical analyses – promoter-centric gene set tests, differential region finding

BIOINFORMATICS APPLICATIONS NOTE

Vol. 26 no. 13 2010, pages 1662–1663
doi:10.1093/bioinformatics/btq247

Genome analysis

Advance Access publication May 10, 2010

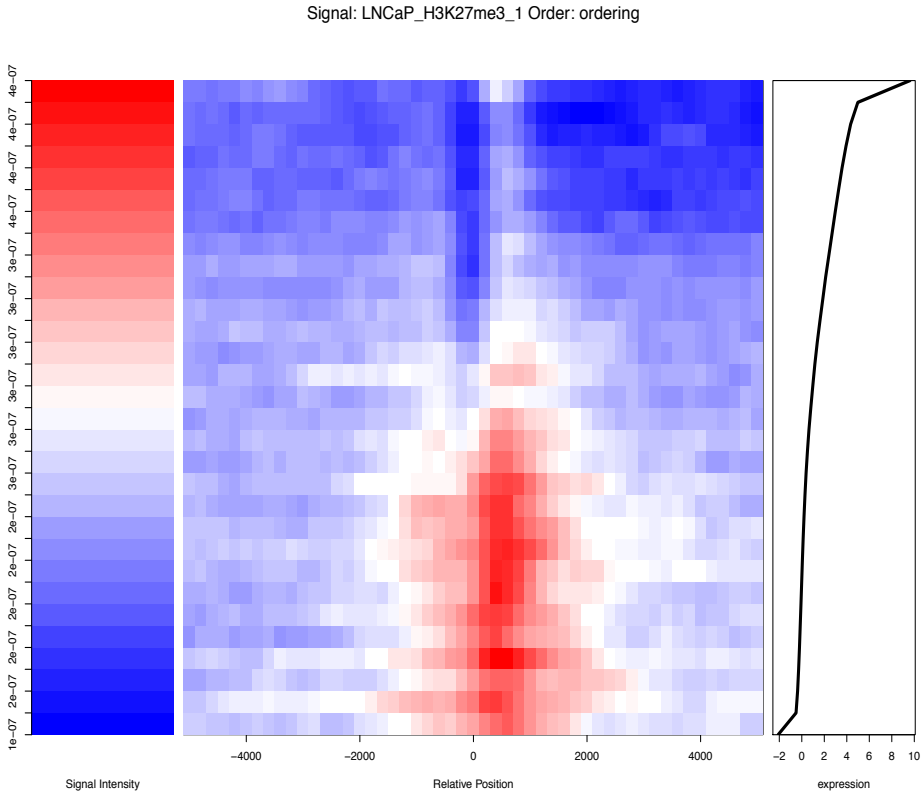
Repitools: an R package for the analysis of enrichment-based epigenomic data

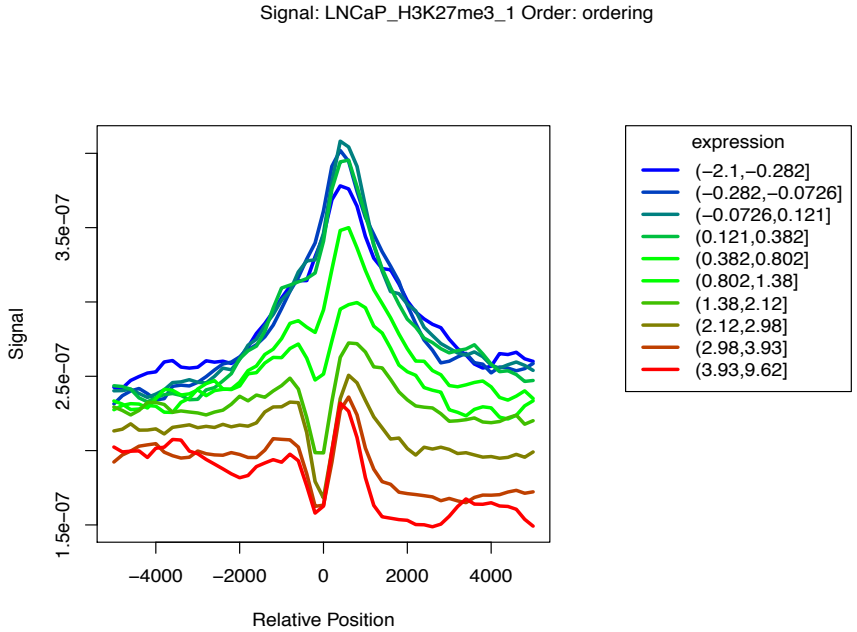
Aaron L. Statham[1], Dario Strbenac[1], Marcel W. Coolen[1], Clare Stirzaker[1], Susan J. Clark[1,2] and Mark D. Robinson[1,3]*

[1]Epigenetics Laboratory, Cancer Program, Garvan Institute of Medical Research, 384 Victoria Street, Darlinghurst, NSW 2010, [2]St Vincent's Clinical School, The University of New South Wales, NSW 2052 and [3]Bioinformatics Division, The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, Victoria 3052, Australia

Associate Editor: John Quackenbush

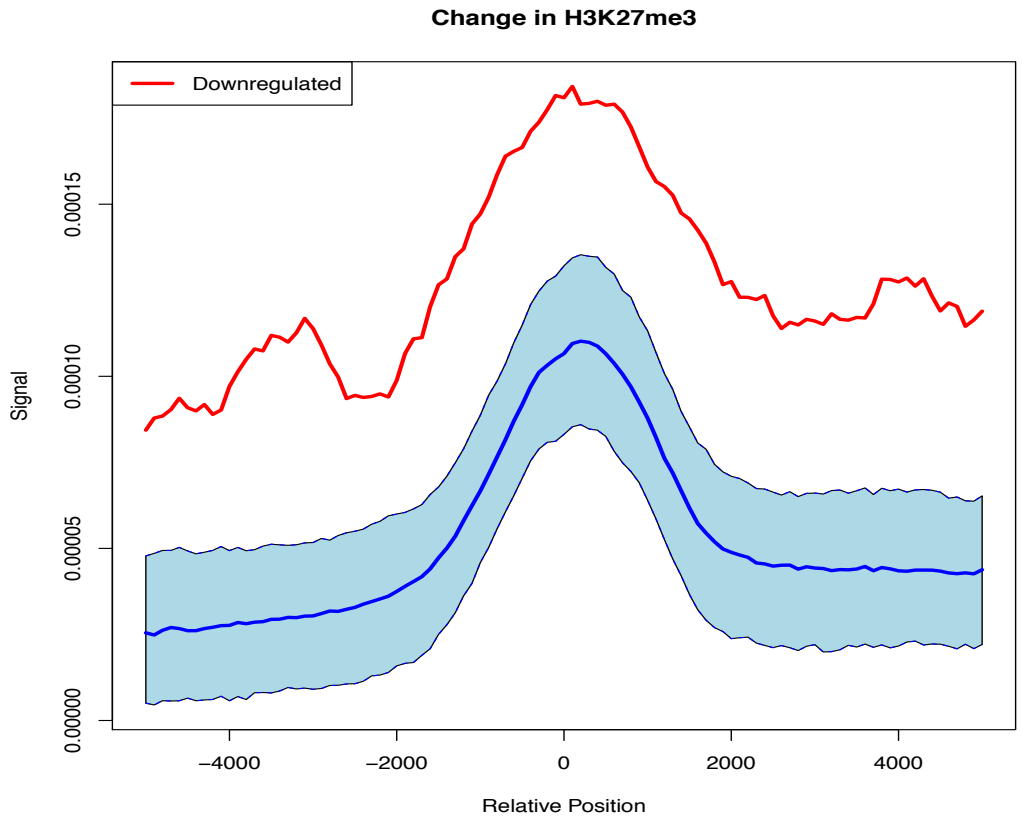
binPlots()





Input: set of reads **or** tiling array data set + gene expression
(can also do Δread density, relate to Δexpression)

profilePlots()



Input: set of reads **or** tiling array data set + gene set
(can also do Δread density)

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

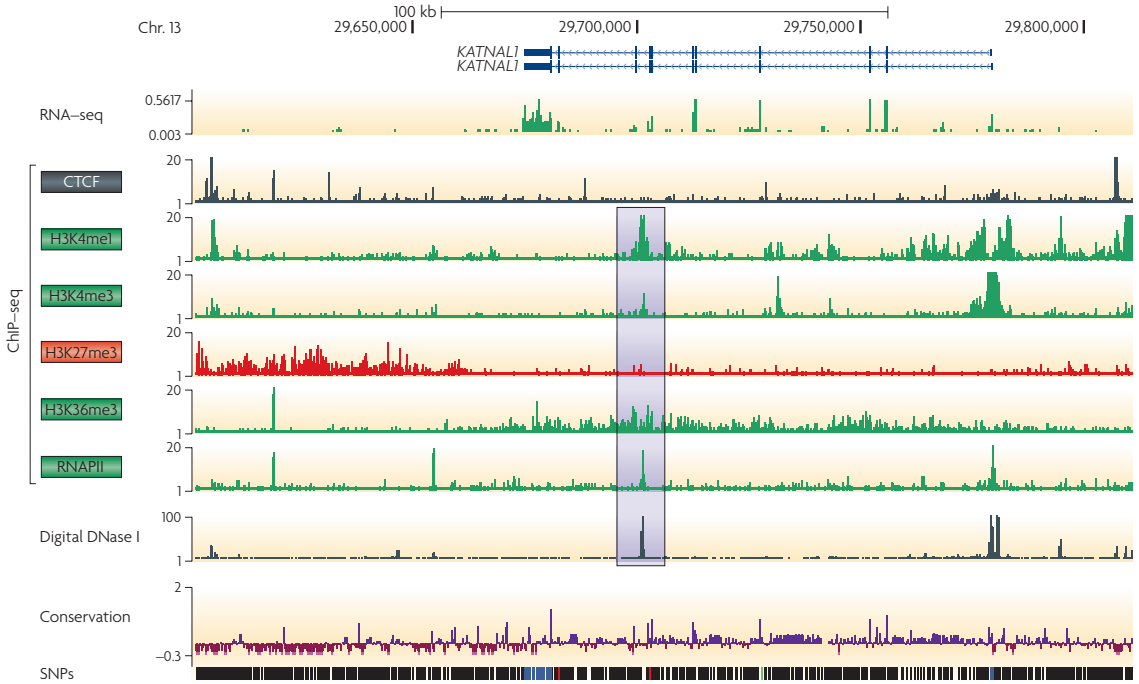


Figure 3 | **Data visualization.** The University of California-Santa Cruz (UCSC) Genome Browser is a tool for viewing genomic data sets. A vast amount of data is available for viewing through this browser. This example from the browser shows numerous data types in K562 cells from the ENCODE Consortium. 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.

ChIP-seq programs

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

Wilbanks and Facciotti (2010) PLoS ONE

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

Peak/region detection for ChIP-seq data

MACS: model-based analysis of ChIP-seq data

Analysis notes:

Adjustment for strandedness of reads

Window-based, simple Poisson model with a region-specific rate estimated from control

FDR control

With the current genome coverage of most ChIP-Seq experiments, tag distribution along the genome could be modeled by a Poisson distribution [7]. The advantage of this model is that one parameter, λ_{BG}, can capture both the mean and the variance of the distribution. After MACS shifts every tag by *d*/*2*, it slides *2d* windows across the genome to find candidate peaks with a significant tag enrichment (Poisson distribution *p*-value based on λ_{BG}, default 10^{-5}). Overlapping enriched peaks are merged, and each tag position is extended *d* bases from its center. The location with the highest fragment pileup, hereafter referred to as the *summit*, is predicted as the precise binding location.

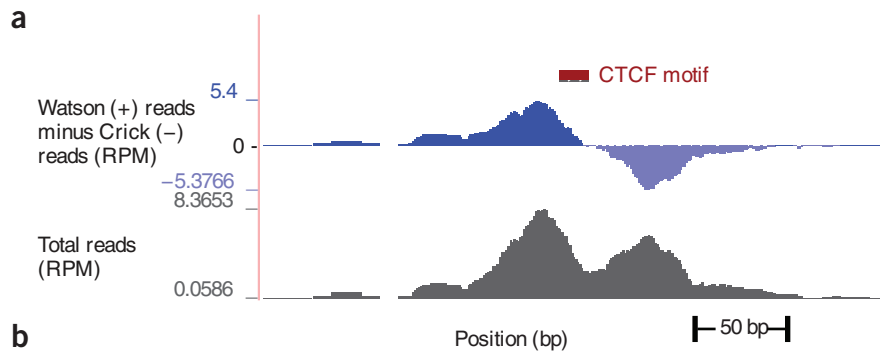
In the control samples, we often observe tag distributions with local fluctuations and biases. For example, at the FoxA1 candidate peak locations, tag counts are well correlated between ChIP and control samples (Figure 1c,d). Many possible sources for these biases include local chromatin structure, DNA 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:

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

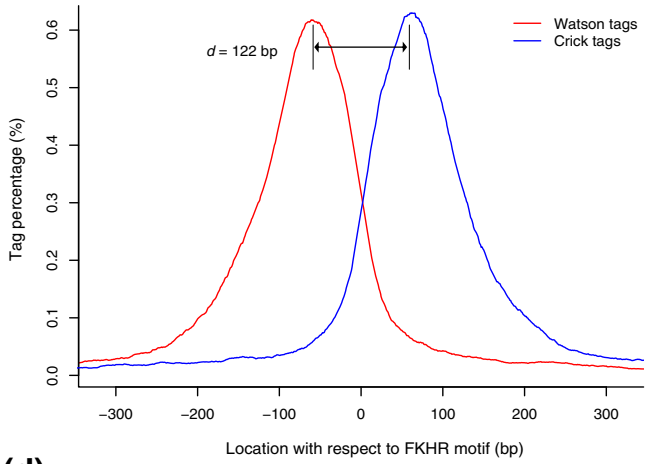
Peak/region detection for ChIP-seq data

MACS: model-based analysis of ChIP-seq data

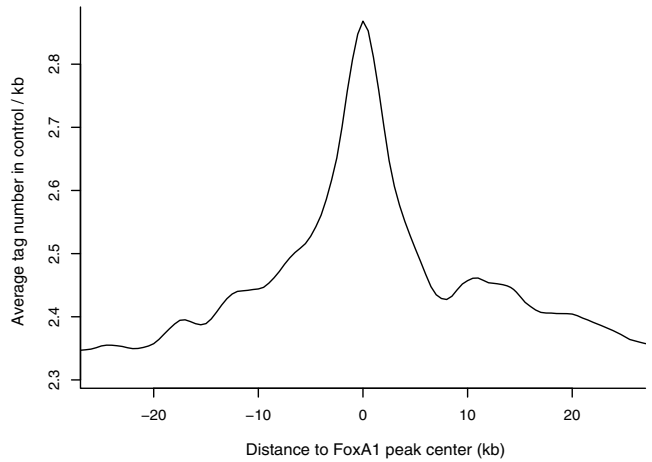
Accounting for strandedness of the reads



(b)



(d)



(f)

Downstream analysis

GREAT predicts functions of *cis*-regulatory regions.

1. **Input:** A set of Genomic Regions (such as transcription factor binding events identified by ChIP-Seq).

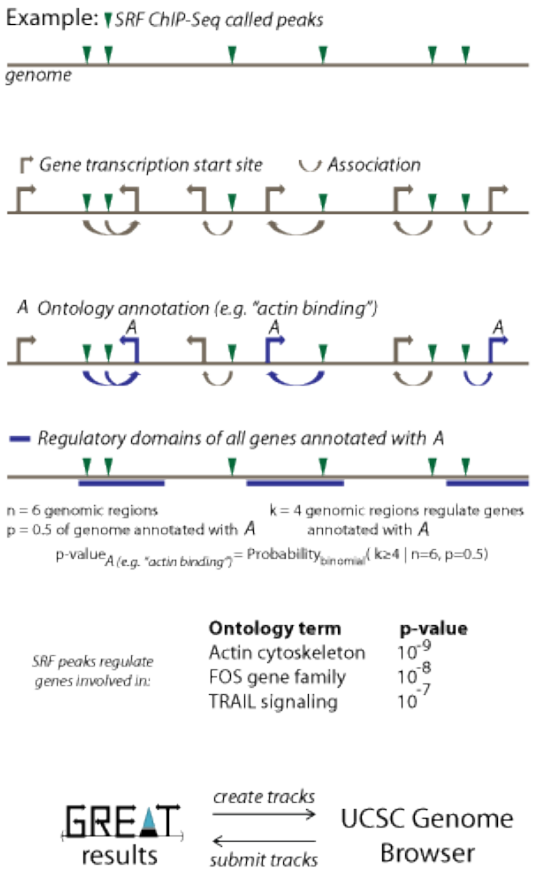
2. GREAT associates both proximal and distal input Genomic Regions with their putative target genes.

3. GREAT uses gene Annotations from numerous ontologies to associate genomic regions with annotations.

4. GREAT calculates statistical Enrichments for associations between Genomic Regions and Annotations.

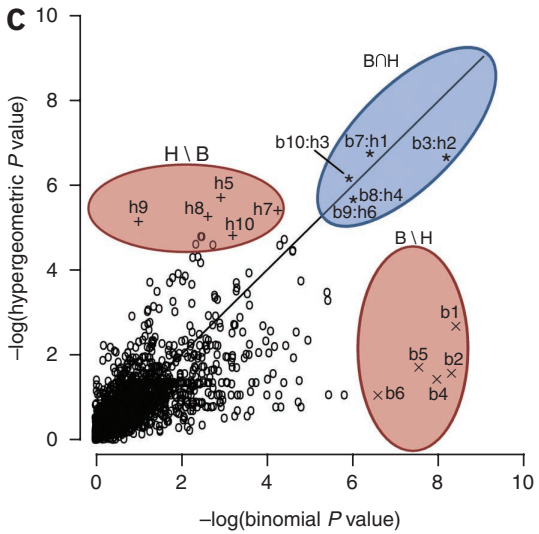
5. **Output:** Annotation terms that are significantly associated with the set of input Genomic Regions.

6. Users can create UCSC custom tracks from term-enriched subsets of Genomic Regions. Any track can be directly submitted to GREAT from the UCSC Table Browser.



McLean et al. (2010) Nature Biotech

Downstream analysis



Binomial-based test

McLean et al. (2010) Nature Biotech

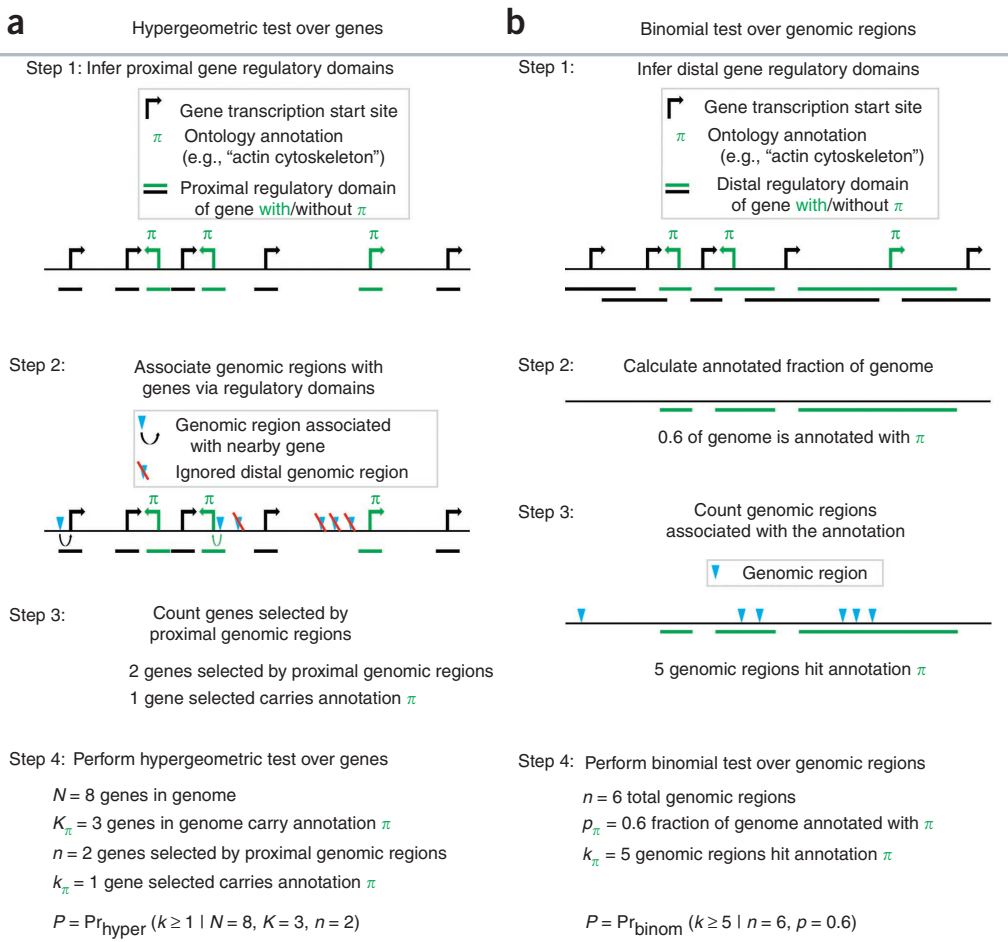
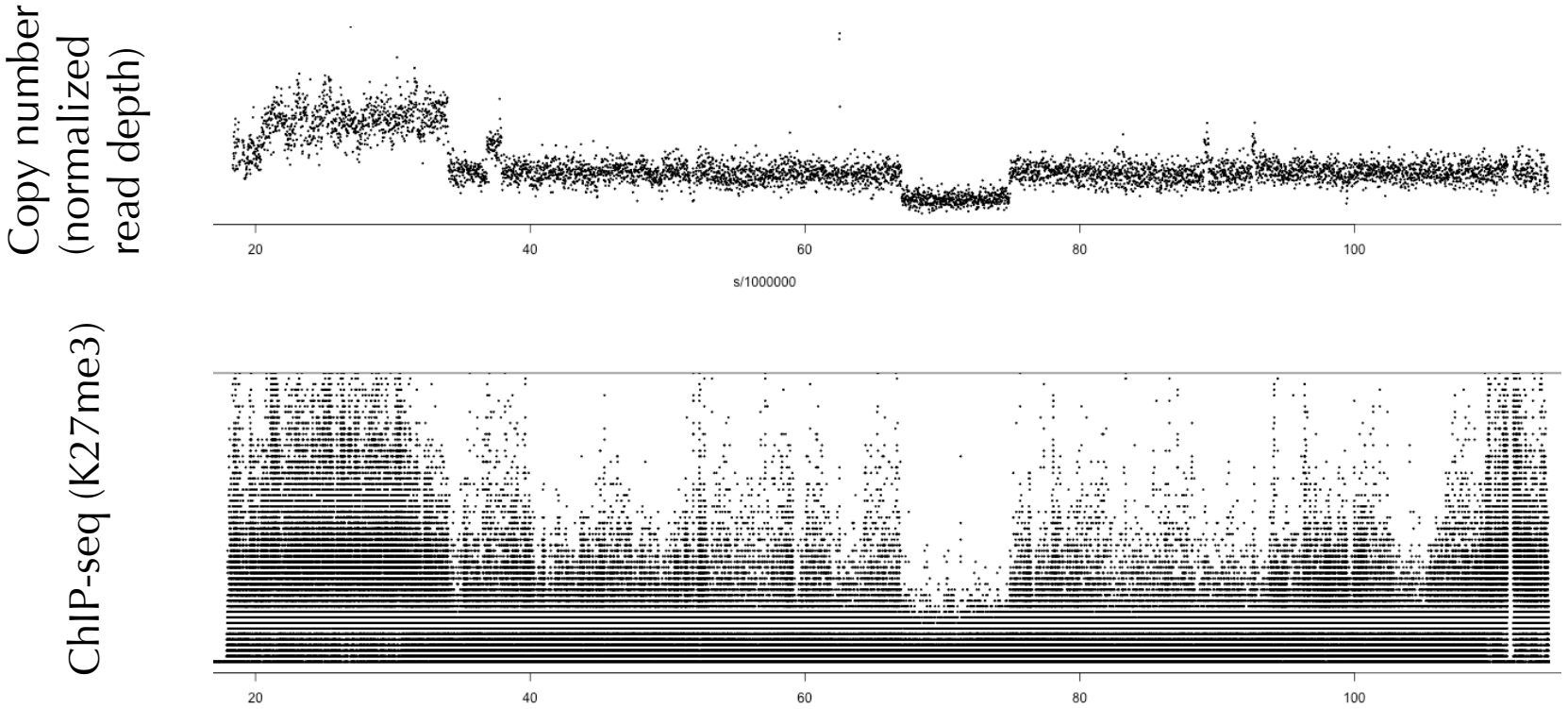


Figure 1 Enrichment analysis of a set of *cis*-regulatory regions. (**a**) The current prevailing methodology associates only proximal binding events with genes and performs a gene-list test of functional enrichments using tools originally designed for microarray analysis. (**b**) GREAT's binomial approach over genomic regions uses the total fraction of the genome associated with a given ontology term (green bar) as the expected fraction of input regions associated with the term by chance.

ChIP-seq signal = biology (copy number, enrichment) + technical effects

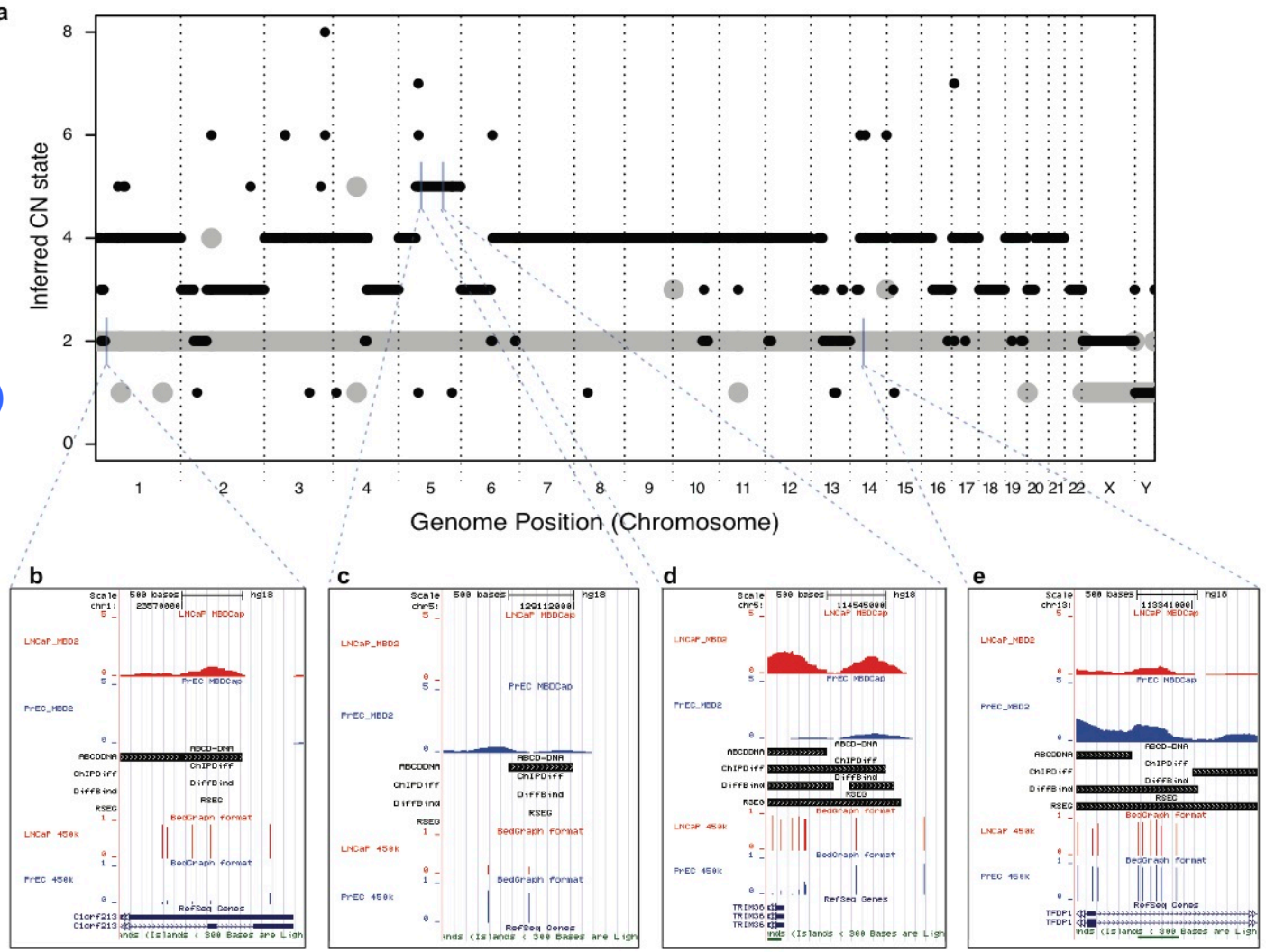


Differential analysis of ChIP-seq is sensitive to CNV

Various tools available:

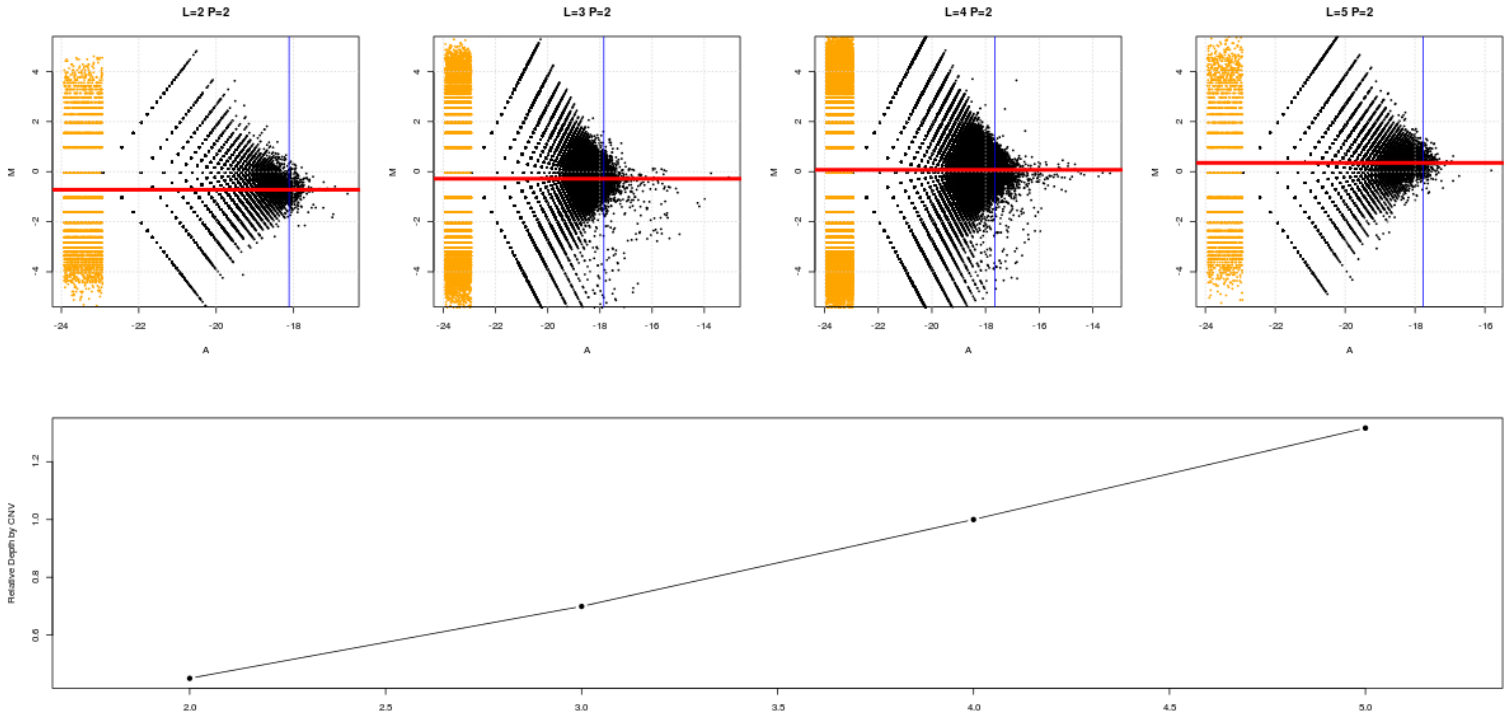
1. ChIPDiff
2. DiffBind (BioC)
3. RSEG
4. "ABCD-DNA" (BioC)

CNV represents a source of FPs and FNs.



Differential ChIP-seq using count-based inferential machinery used in RNA-seq

With an additional step to normalize for CNV
MA-plots by CNV state (L=cancer, P=normal)



What does ABCD-DNA (Affinity-Based Copy-number-aware differential analysis of quantitative DNA-seq) do?

A general framework for CNV-aware differential QDNA-seq analyses

1. Generate read counts at regions of interest (e.g. at detected peaks, tiled regions genome-wide, or proximal to transcription starts);
2. Estimate copy number offsets from an external data source
3. Estimate normalization offsets based on CNV-neutral loci
4. Perform differential analysis of count data (e.g. using edgeR) using offsets.

More details

We model the logarithm of expected value of Y_{ij} as follows:

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

O_{ij} is an r x n matrix of offsets that match the count matrix

X is an r x k matrix that captures the experimental design (conditions, covariates)

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

O_{ij} can be decomposed into **$\log(CN_{ij})$ + $\log(1\ D_j)$** where CN_{ij} is a matrix of offsets for copy number and D_j represents sample-specific offset vector that effectively represents depth of sequencing

Beyond 1D

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 (or affected by) several factors

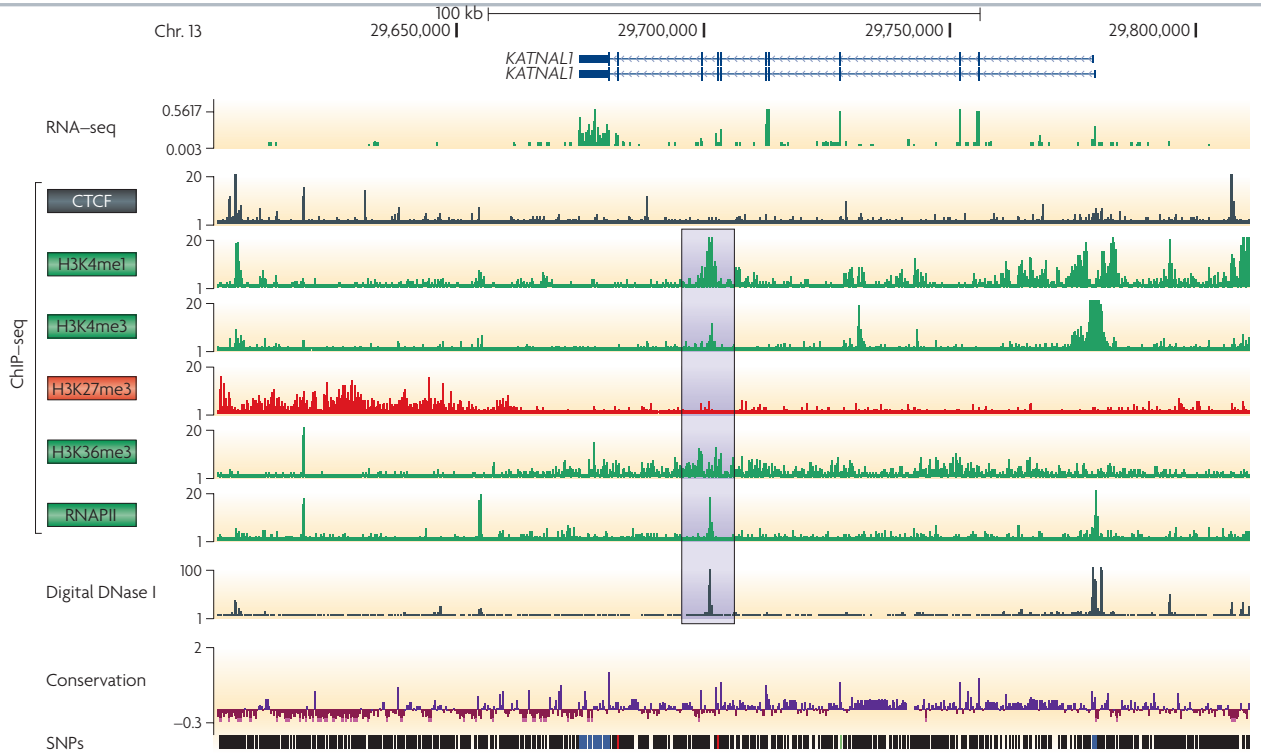


Figure 3 | **Data visualization.** The University of California-Santa Cruz (UCSC) Genome Browser is a tool for viewing genomic data sets. A vast amount of data is available for viewing through this browser. This example from the browser shows numerous data types in K562 cells from the ENCODE Consortium. 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.

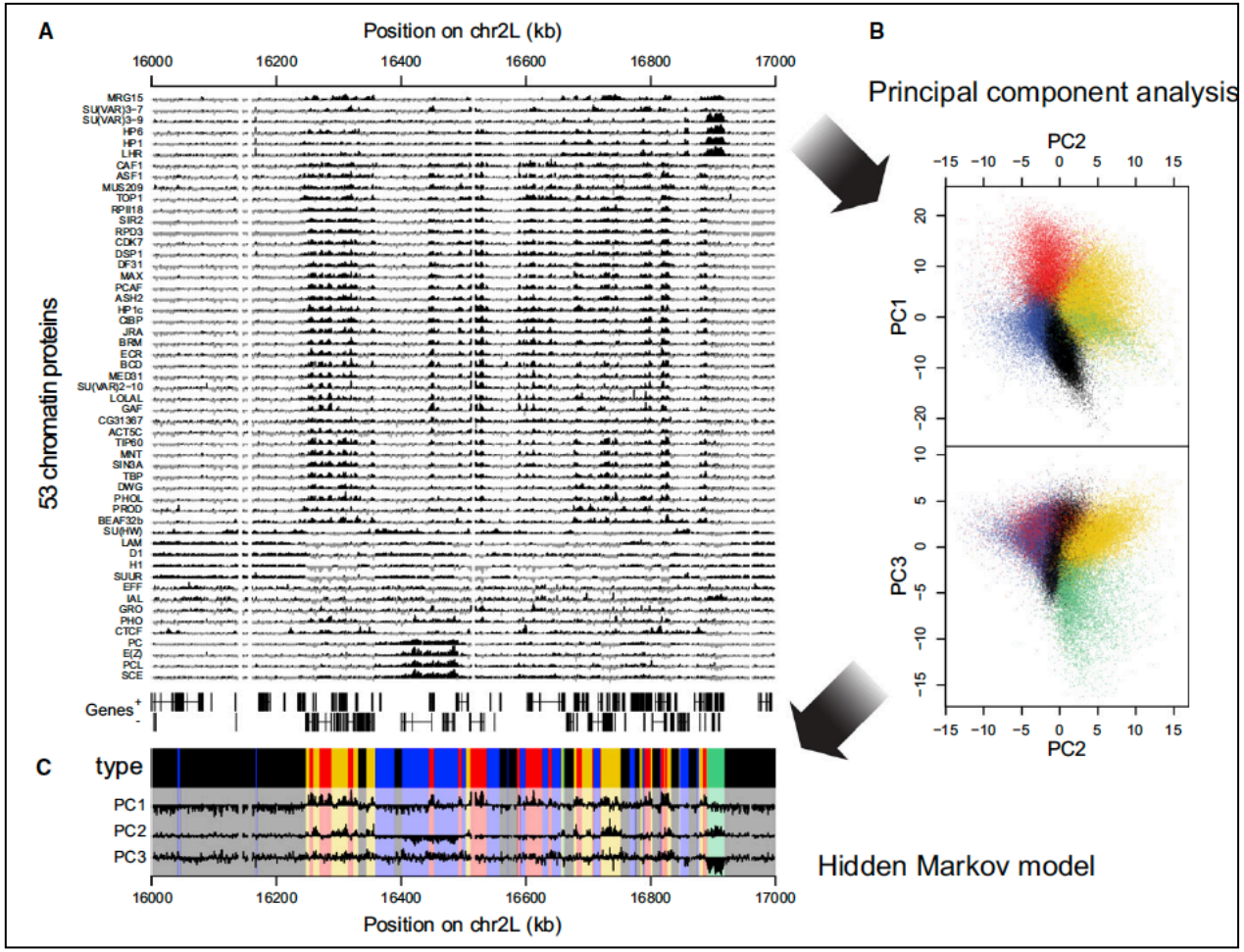
Exploratory analyses

53 chromatin factors (ChIP-seq)

Compression to 3 principal components

Learn HMM

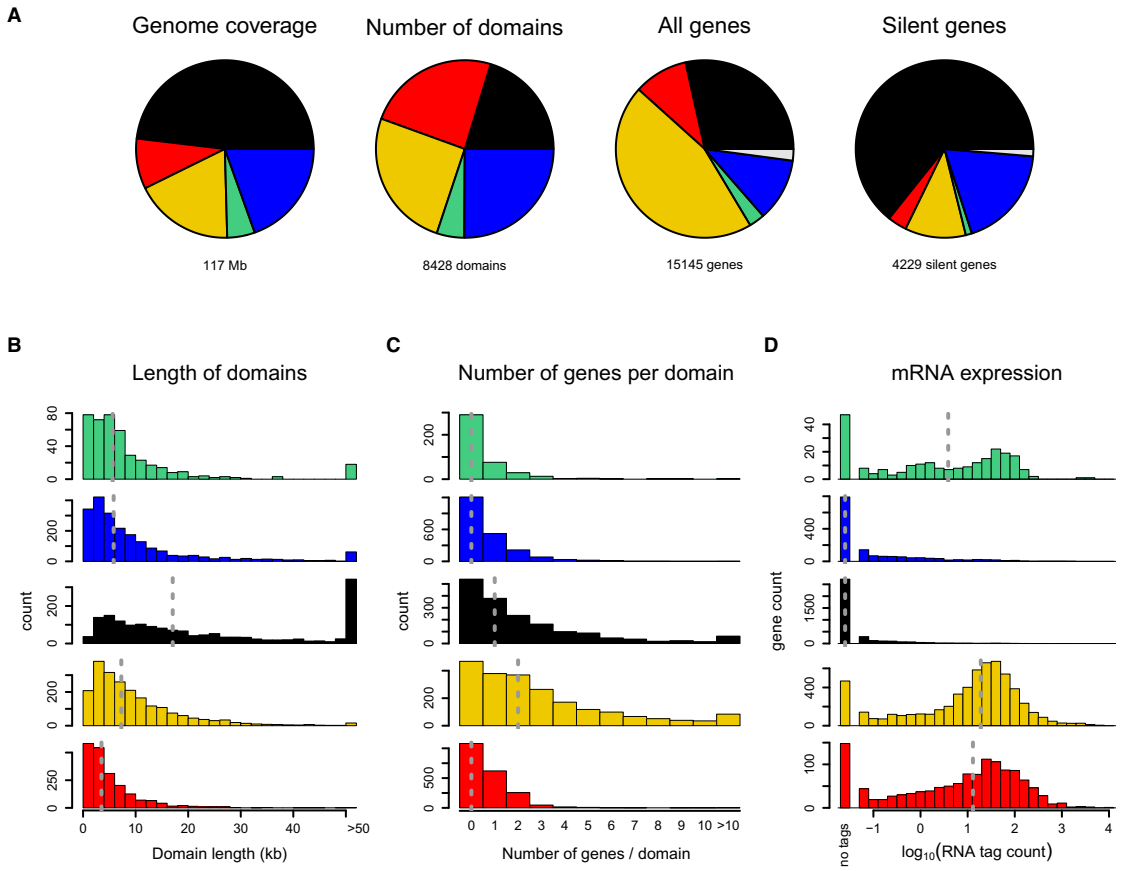
Every region of the genome partitioned into 5 “states” (here, assigned a colour)



Filion et al. Cell 2010

Exploratory analyses

"Colours" are reflective of various features



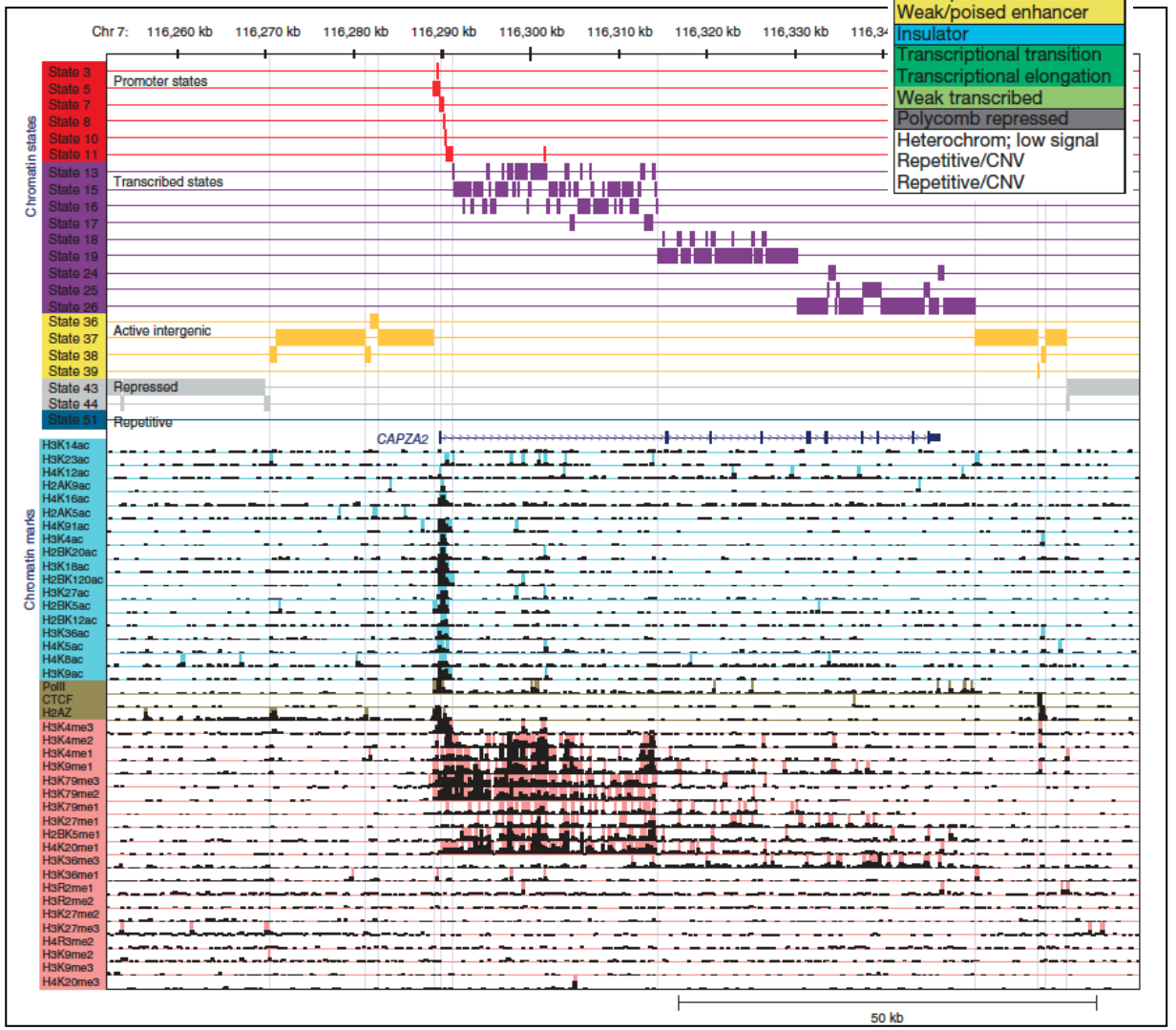
Filion et al. Cell 2010

Exploratory analyses

No compression

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

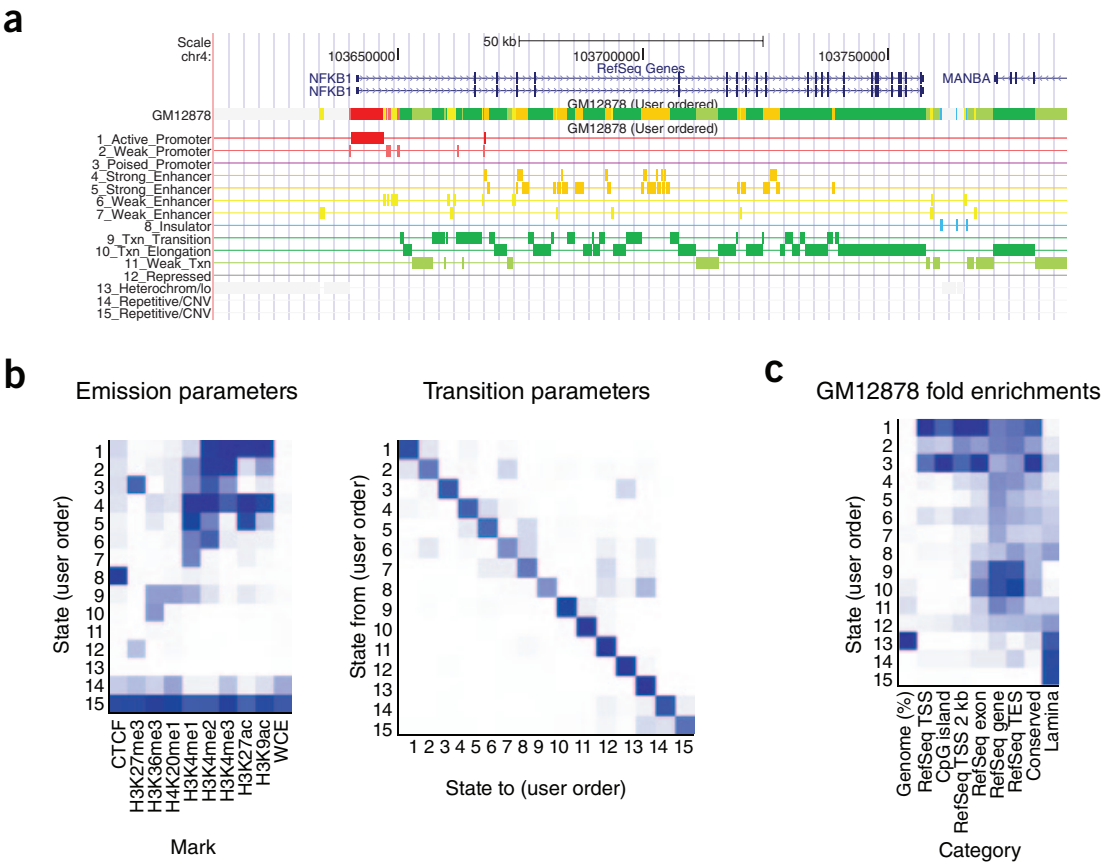
Multivariate HMM is trained; genome is partitioned into 15 states



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

Segway

Dynamic Bayesian Network

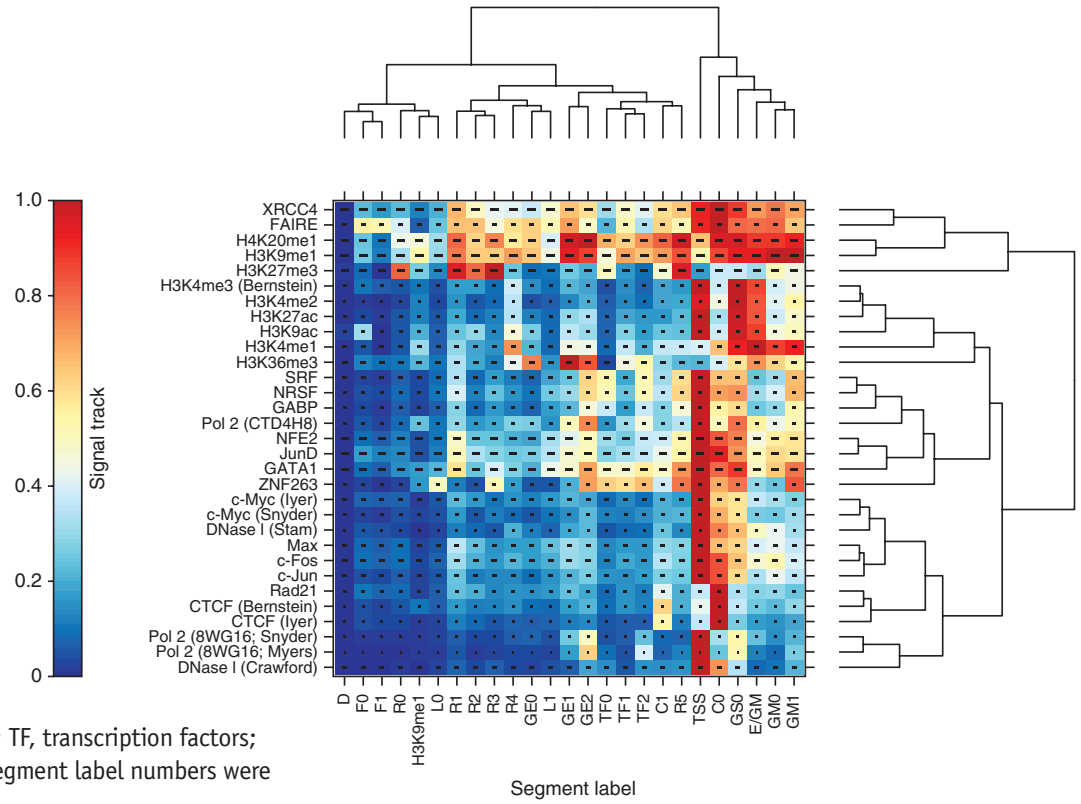
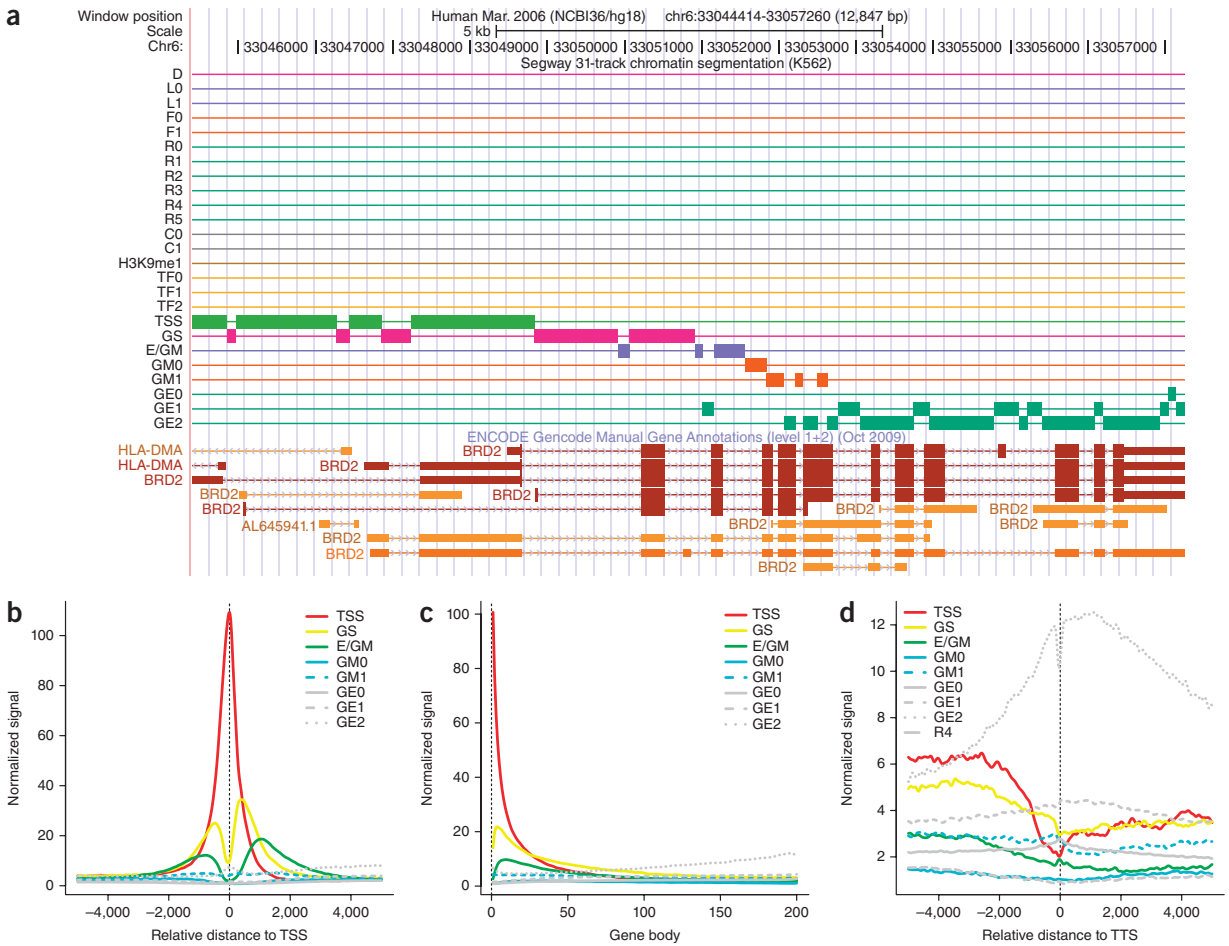


Figure 1 | Heat map of discovered Gaussian parameters in an unsupervised 25-label segmentation trained on 31 tracks of histone modification, transcription-factor binding and open chromatin signal data in 1% of the human genome. Row labels include last names of the principal investigator in whose laboratory data were generated, when assays were conducted in multiple laboratories (Stam, Stamatoyannopoulos). Each row contains parameters for one signal track, and each column contains parameters for one segment label. Within each row, we did an affine transformation, such that the largest mean was 1 and the smallest 0. The color in each cell indicates the transformed mean parameter μ according to the color bar on the left. The width of the black inner boxes is proportional to the square root of the variance parameter σ^2, after multiplying by the linear factor used in the transformation of μ. Dendrogram show a hierarchical clustering by both rows and columns. Functional categories manually assigned to segment labels: D, dead; F, FAIRE; R, repression; H3K9me1, histone 3 lysine 9 onomethylation; L, low; GE, gene end; TF, transcription factors; C, CTCF; GS, gene start; E, enhancer; GM, gene middle; segment label numbers were assigned arbitrarily.

Segway



labels: D, dead; F, FAIRE; R, repression; H3K9me1, histone 3 lysine 9 onomethylation; L, low; GE, gene end; TF, transcription factors; C, CTCF; GS, gene start; E, enhancer; GM, gene middle; segment label numbers were assigned arbitrarily.

Exploratory analysis: clustering combined epigenomic profiles at feature (promoter/gene) level

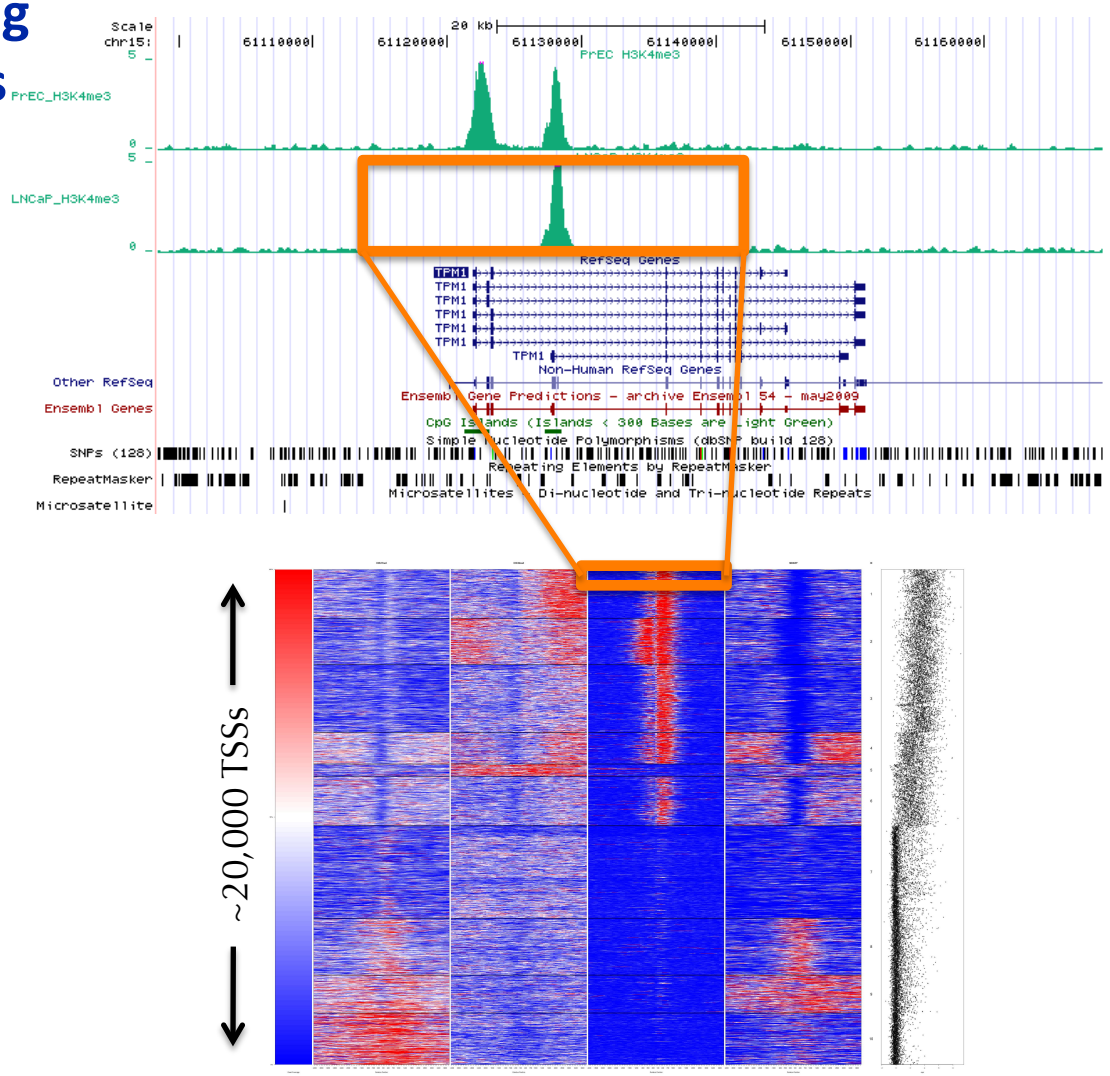
Calculate coverage around features of interest (here, TSSs)

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

Overlay expression, order clusters by median

Available in Bioconductor (Repitools)

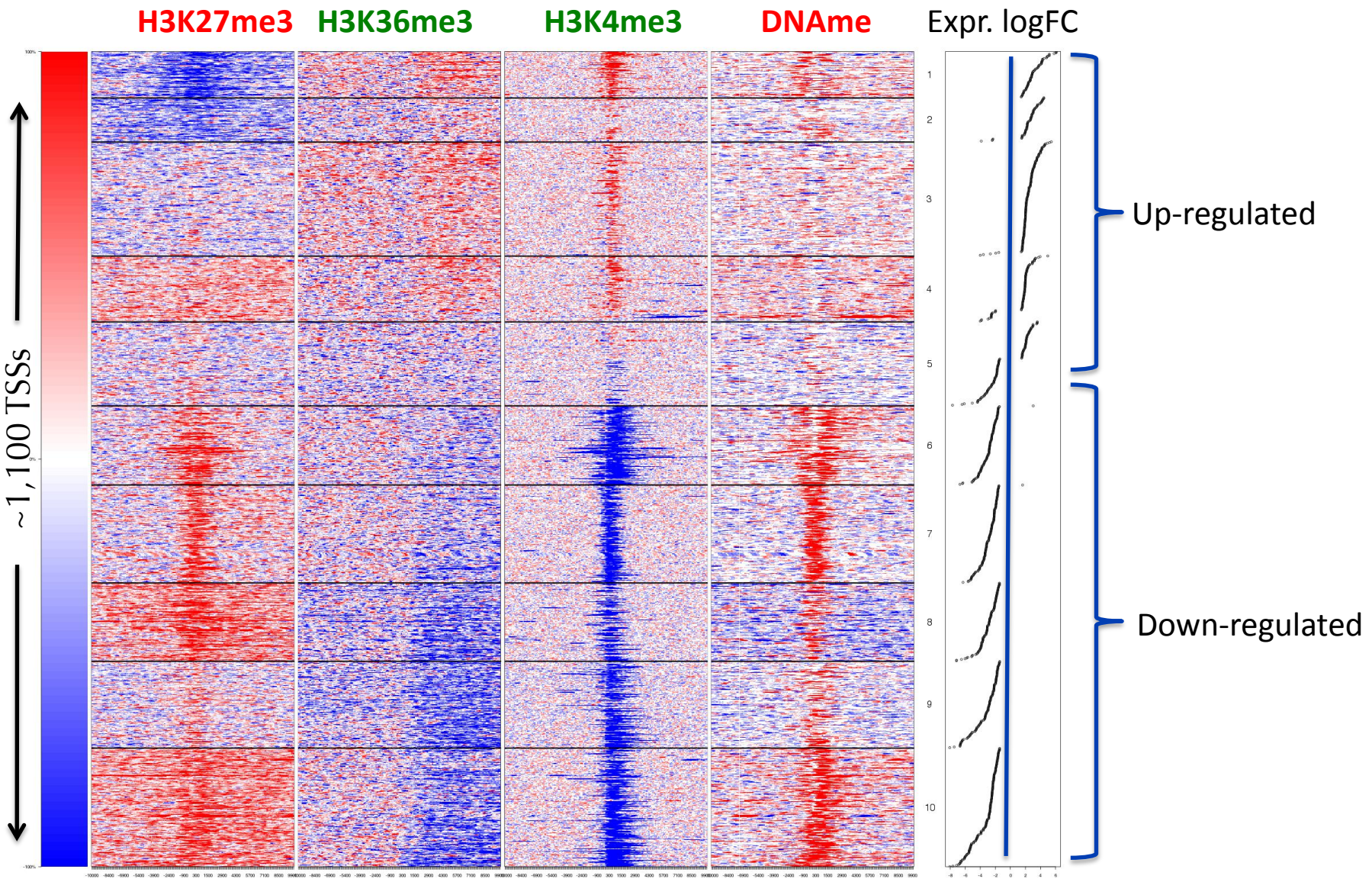
featureScores() to collect information, clusterPlots() to plot



University of Zurich[UZH]

Institute of Molecular Life Sciences

Clustering changes (just DE genes)



H3K27me3 profiles along a gene

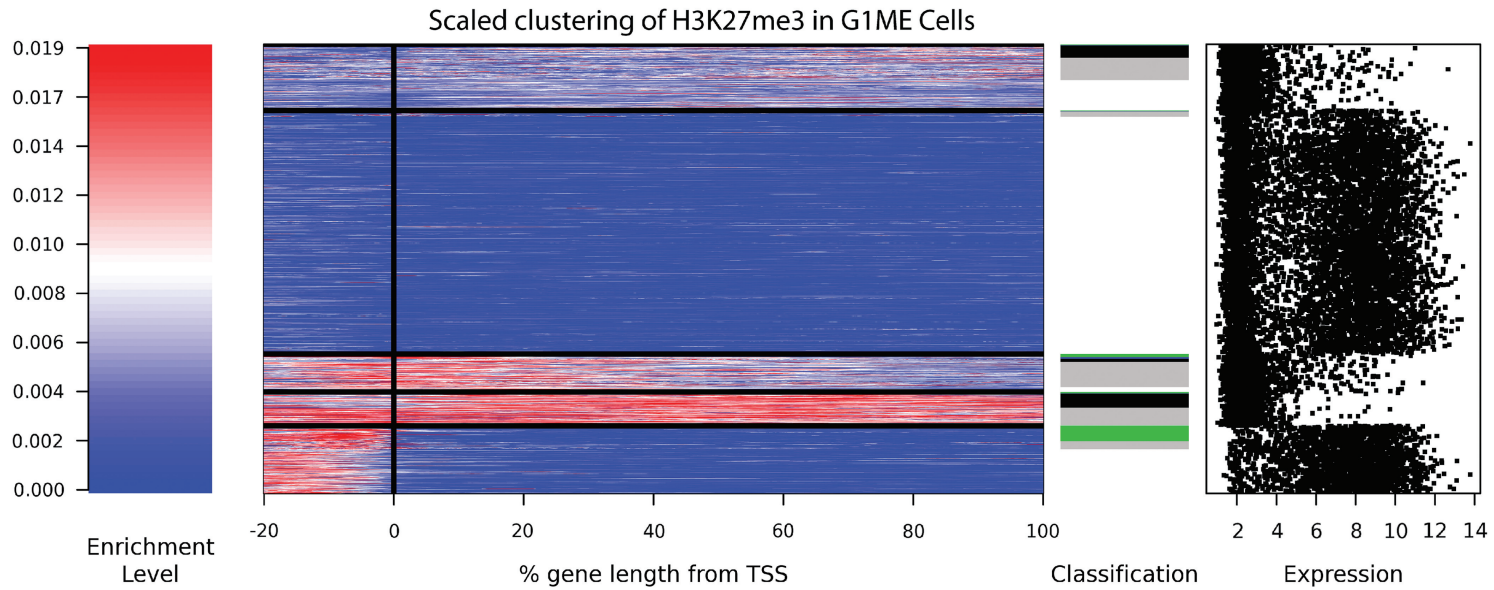


Figure 5. *K*-means clustering of genic H3K27me3 profiles in G1ME cells. The signal intensity is shown as a spectrogram, with red reflecting a high enrichment signal and blue reflecting no signal. All genes were scaled to have the same length, and position relative to the TSS is shown in percentage terms. Genes were sorted first by cluster, then by classification (black: broad; green: promoter; blue: TSS; grey: marked but unclassified). The expression level of all genes is shown on the far right. Additional cluster profiles are provided for the other cell types (Supplementary Figure S8).

Young et al. 2011