Swiss TPH
Swiss Tropical and Public Health Institute
Schweizerisches Tropen- und Public Health-Institut
Institut Tropical et de Santé Publique Suisse

Christoph D. Schmid, PhD

Biostatistics and Computational Sciences

Reviews in Computational Biology

Comparing Epigenetic Maps

Computational tasks and aspects of data analysis

2 May 2011

ETHZ, CAB G 56

outline

Swiss TPH

What are epigenetic maps?

'formatting marks' in genome: post-translational modifications of histones, DNA methylation, DNA-binding proteins (or complexes)

How are epigenetic maps produced?

methodology of massively parallel sequencing

mapping to reference genome assembly

infer genomic coordinates of

- peaks (→ binding sites of transcription factors, ...)
- broader epigenetic domains (→ histone modifications)
- single target sites (→ ~ 29M human CpGs for DNA methylation)

How are epigenetic maps analyzed and compared?

quantitative measure to compare characteristics and strengths of signal

Basics of Epigenetics

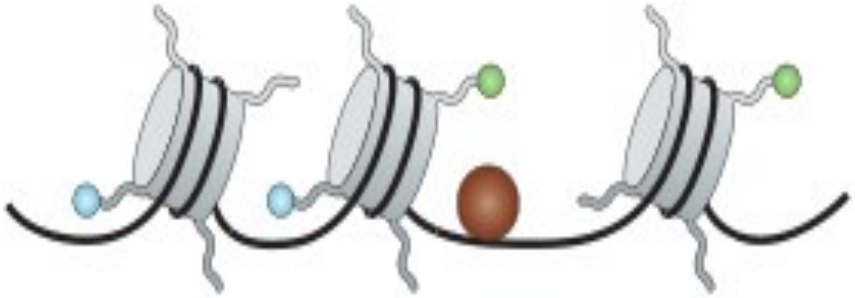
Swiss TPH

Nucleotide sequence:

Thisisawordandthisoneisexpressedwhilethisisrepressed

Epigenetics as 'formatting' of nucleotide sequence:

This is a word and this one is expressed while this is ~~repressed~~

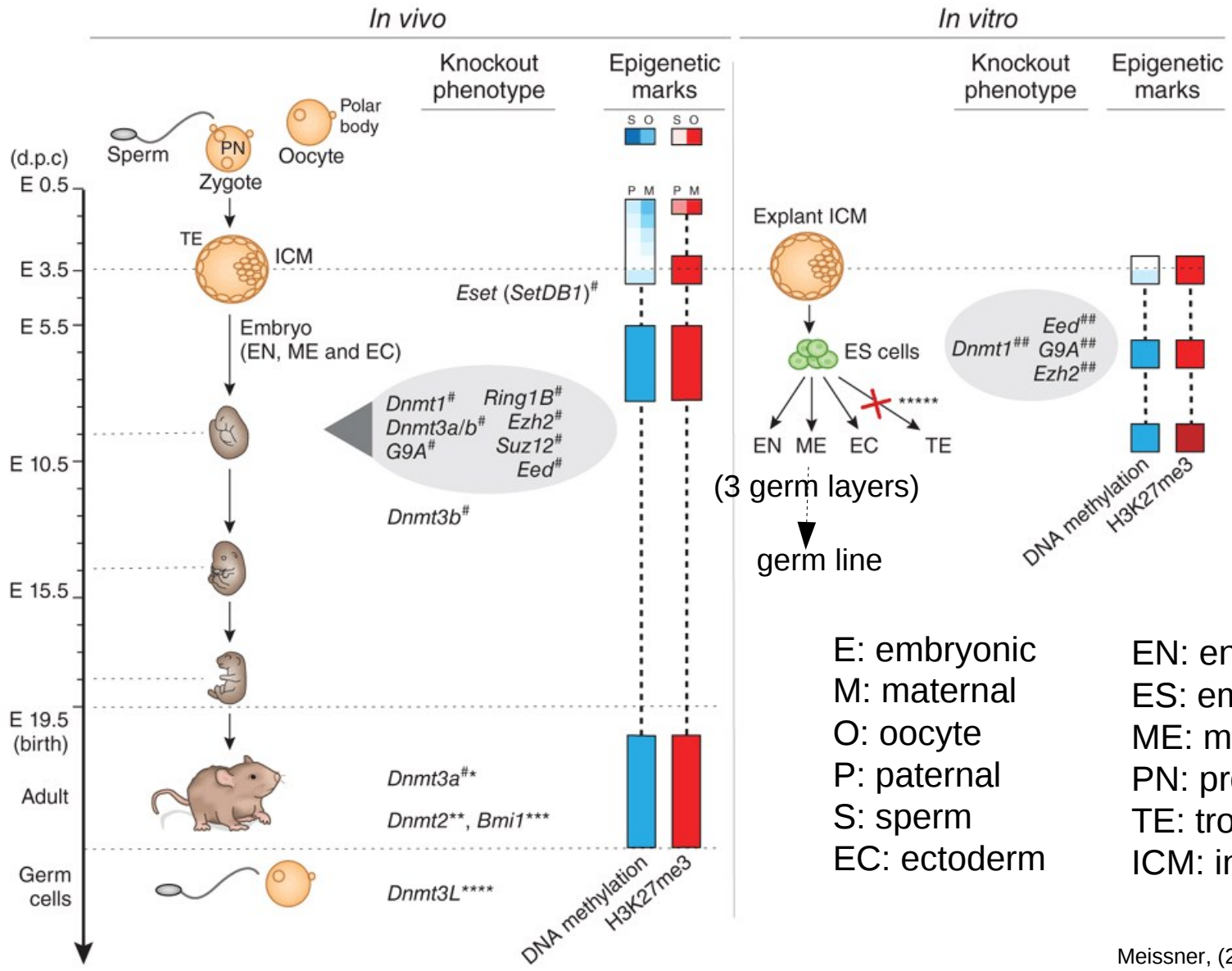


Large variability:
>50 gene copies for human histones
>100 distinct post-translational modifications of histones
DNA methylation

Epigenome at least in part transmitted to daughter cells !

-> 'revival' of Lamarckism!

Epigenetic modifications in pluripotent and differentiated cells



E: embryonic
M: maternal
O: oocyte
P: paternal
S: sperm
EC: ectoderm

EN: endoderm
ES: embryonic stem
ME: mesoderm
PN: pronuclei
TE: trophectoderm.
ICM: inner cell mass

Meissner, (2010) Nat Biotech

Intro ChIP-seq

define genomic sites of antigens associated to DNA

histone modifications

DNA methylation

transcription factors

components of DNA binding complexes

Chromosome conformation capture (3C)

physical location (e.g. via proteins associated to the nuclear membrane)

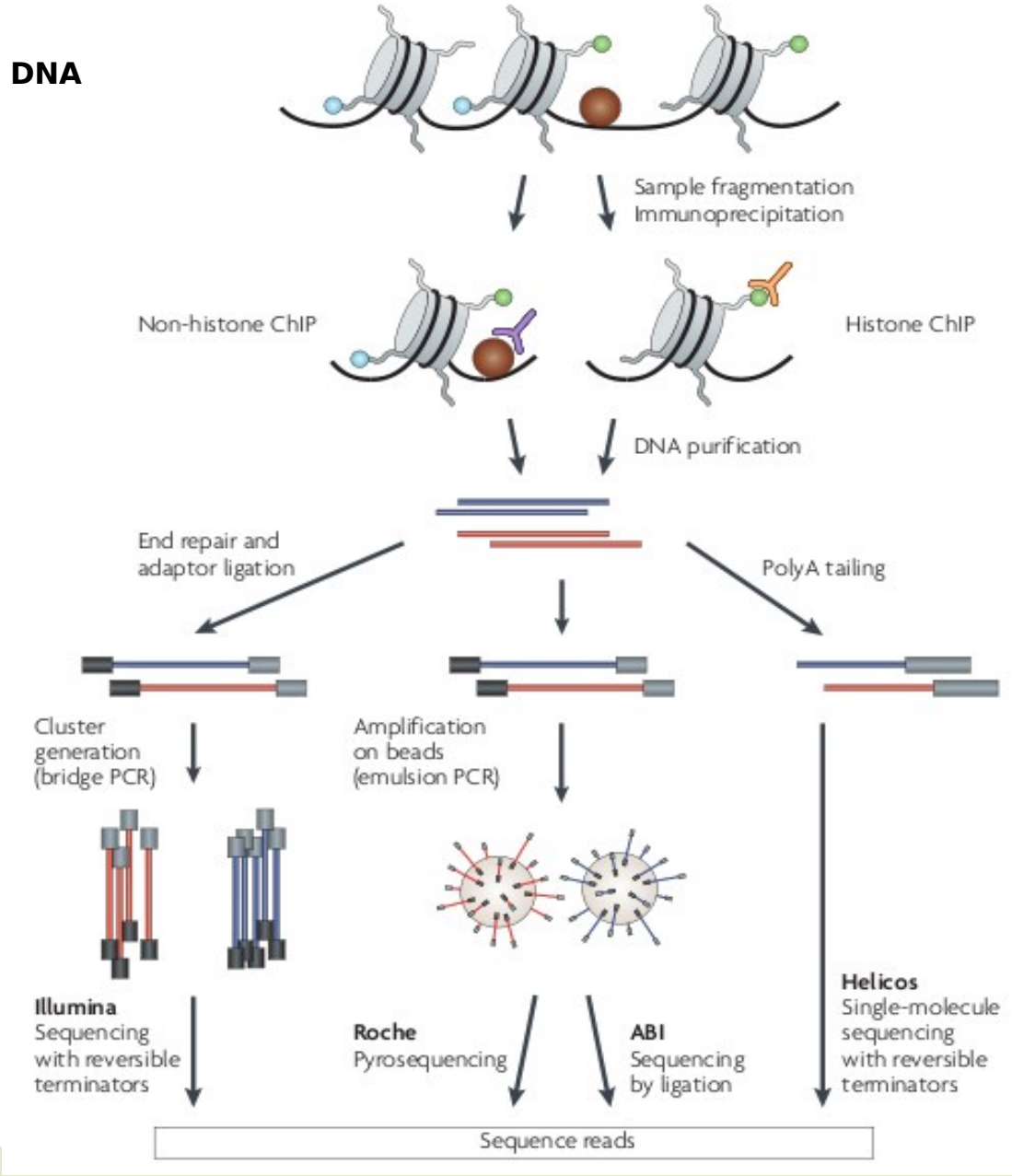
limitation:

Chromatin ImmunoPrecipitation (ChIP)

only enriches for target sequences

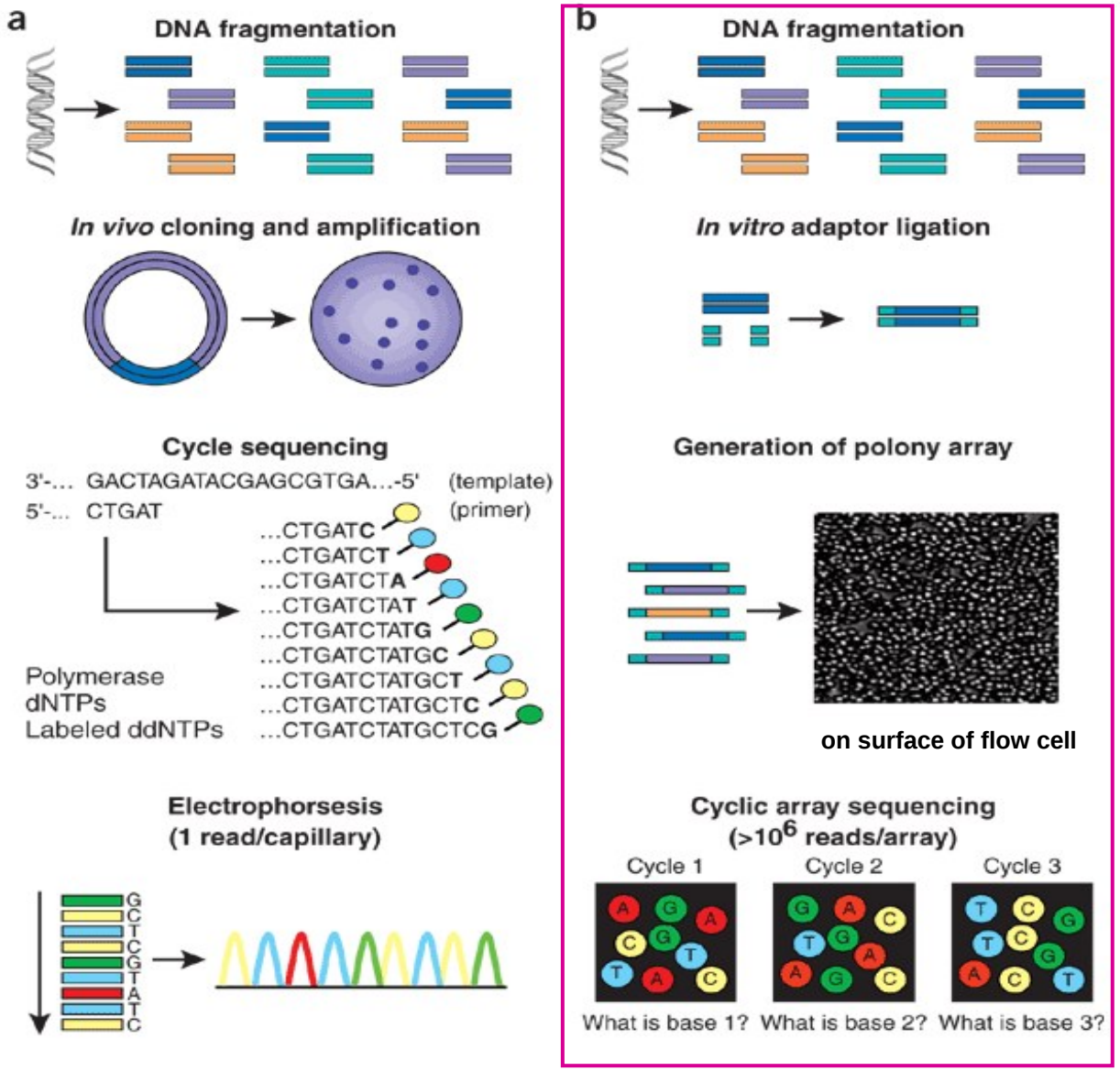
→ considerable level of 'background'

in resulting sequence data!



Park, P.J. (2009) Nat Rev Genet

DNA sequencing to the next level



- Ultra-High-Throughput
- Next-generation
- Deep
- Massively parallel

Sequencing

Jay Shendure & Hanlee Ji
Nature Biotech. 26, 1135 - 1145 (2008)

mapping of reads to genome assembly

Swiss TPH

efficiency problem:

several millions of reads (sequence tags) to ~3 billions of potential positions in human genome

need to allow for mismatches in alignment due to:

sequencing errors

divergences to genomic reference sequences (SNPs)

handling of repetitive sequences, multi-hit reads, ...

- exclude multi-hit reads => signal on repetitive sequences biased
- evenly distribute to all potential loci => introduce biases
- weighting using number of neighboring reads => comp. expensive

Repetitive sequences and ChIP-seq

Swiss TPH

- almost 50% of the human genome is annotated as repetitive sequences (heterochromatin, microsatellites, ...)
  - likely culprit of errors in assembly of genomes
  - large part of genomic diversity may be due to repeats
- not too recent duplications nevertheless with unique subsequences
  - thus potentially identifiable by sequence (while excluded in hybridization-based assays!)
- likely nevertheless source of artifacts !?
- NOT 'junk DNA': reports on binding sites within repeats

Collection of popular mapping tools

Swiss TPH

- Bowtie (Langmead et al. 2009):

very efficient though indexing strategy derived from file compression, supplemented by cloud computing implementation (Crossbow)

- MAQ (Li and Durbin 2009):

slightly more sensitive, ~35x slower than bowtie

- Eland:

hash-based algorithm, proprietary solution and standard of Illumina platform

- BLAT / fetchGWI / tagger (Kent 2002 / Iseli et al. 2007):

efficient perfect match mappers from pre-UHTS era

- BLAST (Altschul et al. 1990):

de facto standard for sequence db search, not suited for short reads (short match length disables to reach significance!)

=> heuristic algorithms: optimal solution not granted and biases not excluded!

→ use **consistent** mapping procedures if comparing maps !

Formats of mapping results

Swiss TPH

Representations of genomic coordinates: chr17;21546235 (hg19)

→ chromosome number and nucleotide position of specific genome assembly:

	remark	approx. file size in MB*
eland	may describe all reads (incl. no or multiple hits) [no reference available]	5800
sga	counts of mapped reads per position [http://ccg.vital-it.ch/chipseq/sga_specs.html]	750
sam	broadly supported, generic format [http://samtools.sourceforge.net/SAM1.pdf]	
bam	binary version of sam format [https://github.com/pezmaster31/bamtools]	
bed	specifies intervals, supported by UCSC [http://genome.ucsc.edu/FAQ/FAQformat#format1]	800
wig	counts of reads within specified bins (→ UCSC) [http://genome.ucsc.edu/goldenPath/help/wiggle.html]	
BigWig	indexed binary format of wig, upload 'on demand' [http://genome.ucsc.edu/FAQ/FAQformat#format6.1]	

*) GEO GSE12782 data set

Why to compare epigenetic maps ?

Swiss TPH

quality control (consistent in replicates, enriched vs. negative control, ...)

- how many domains / peaks similar or different between 2 maps?
  - ‘significant’ changes rel. to reference data set ?
  - track events during cellular differentiation pathways

global measure of distance between epigenetic maps ?

mechanistic interactions ↔ positional co-localization

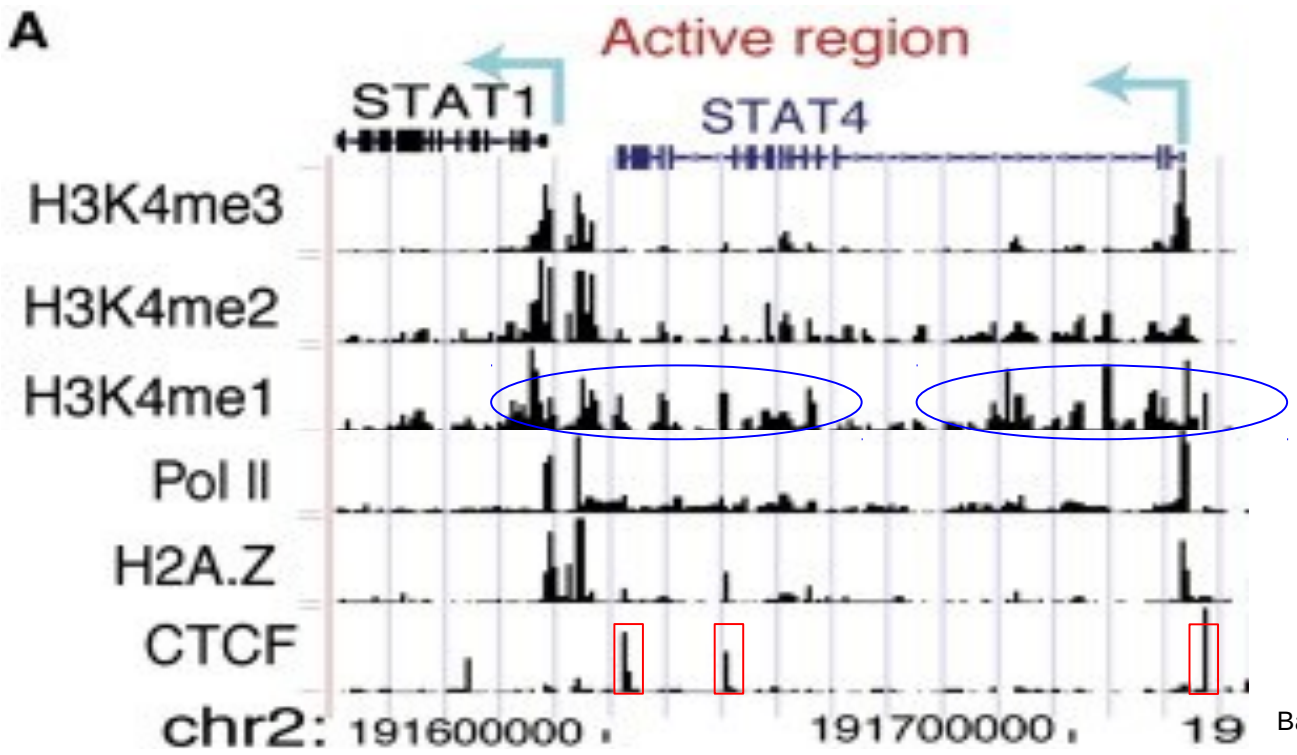
- two (or more) different proteins binding to closely neighboring genomic sites?
- two modifications mutually exclusive?
- infer epigenetic profiles characteristic for a set of sites (genes, enhancers, ...)

find all sites with specific (cooperative) binding patterns?

=> flexible system to allow comparisons of cell types & states,

epigenetic marks, procedures
genome annotation features

Peaks and domains in epigenetic maps



Barski, A. et al. (2007) Cell

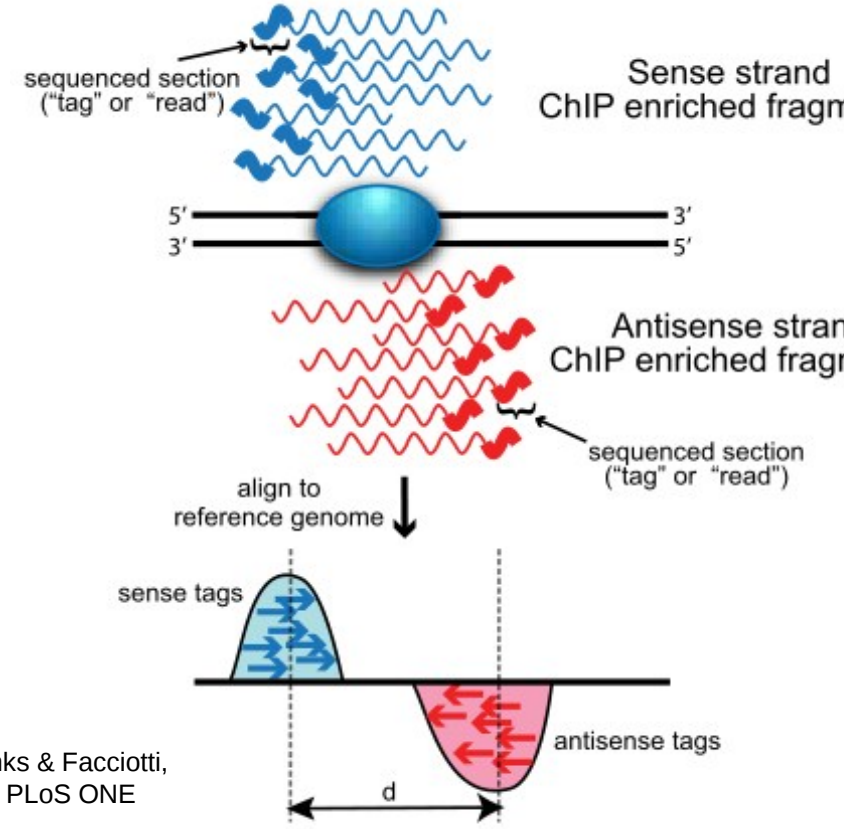
-> represents average over cell population [vs. individual pattern for each cell ?]

-> compare
- many different epigenetic modifications
- in different cell states
- using variable procedures

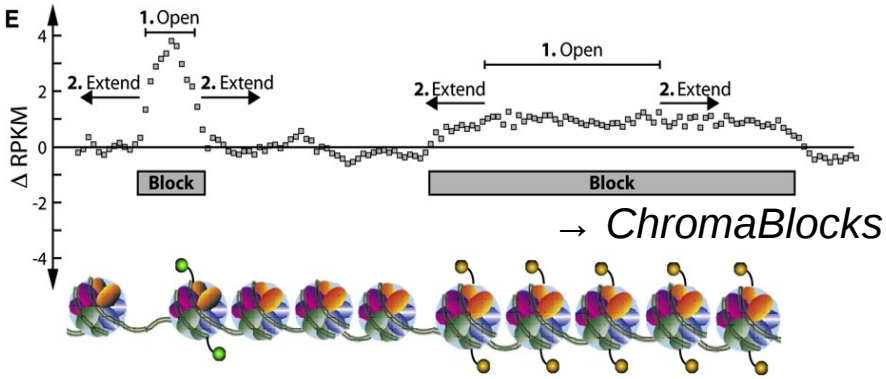
mapped to **1** genome assembly

Infer peaks and domains

peaks (transcription factors) or domains (histone modif., ...)



Wilbanks & Facciotti, (2010) PLoS ONE



→ *ChromaBlocks*

Hawkins, R.D. et al. (2010) Cell Stem Cell

→ intervals of genomic coordinates with signal intensity 'clearly' above background

Approaches to determine thresholds

Swiss TPH

ChIP-seq with a precisely quantifiable output: number of sequence reads

BUT:

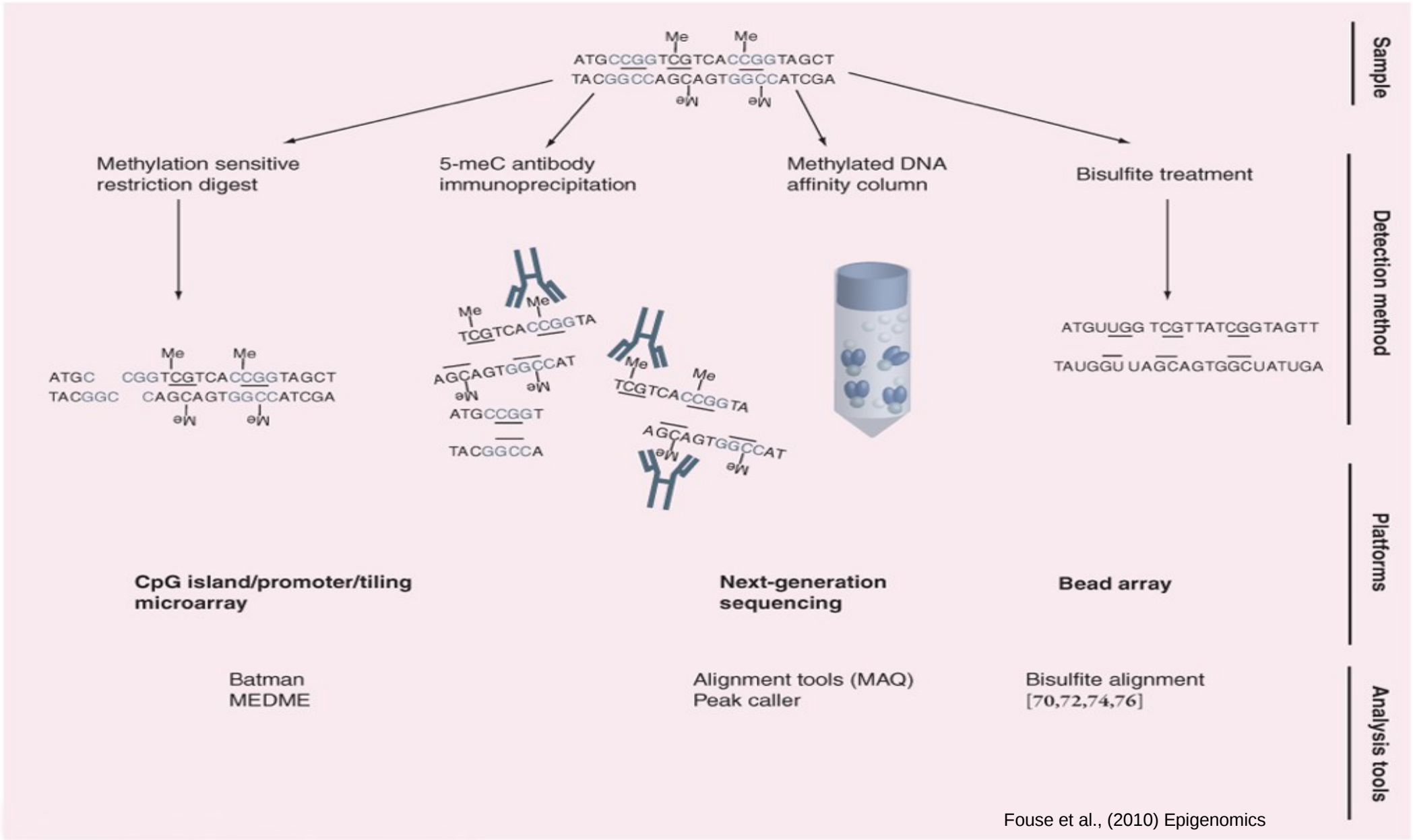
- different ChIP-seq experiments with different total read numbers
→ normalize with total number of reads per experiment

- read densities in input DNA not evenly distributed over genome
→ going local: separate genome into bins to sum reads
→ derive FDRs from data vs. control samples

- even control experiments (input DNA, mock IP, …) with read densities
→ statistical models: binomial, or Poisson distributions and their variants
→ Normalization-Free Significance Analysis (Kowalczyk et al., 2011)

=> if comparing maps, apply **consistent** procedure (large number of different tools and criteria)

Genome-scale DNA methylation analysis



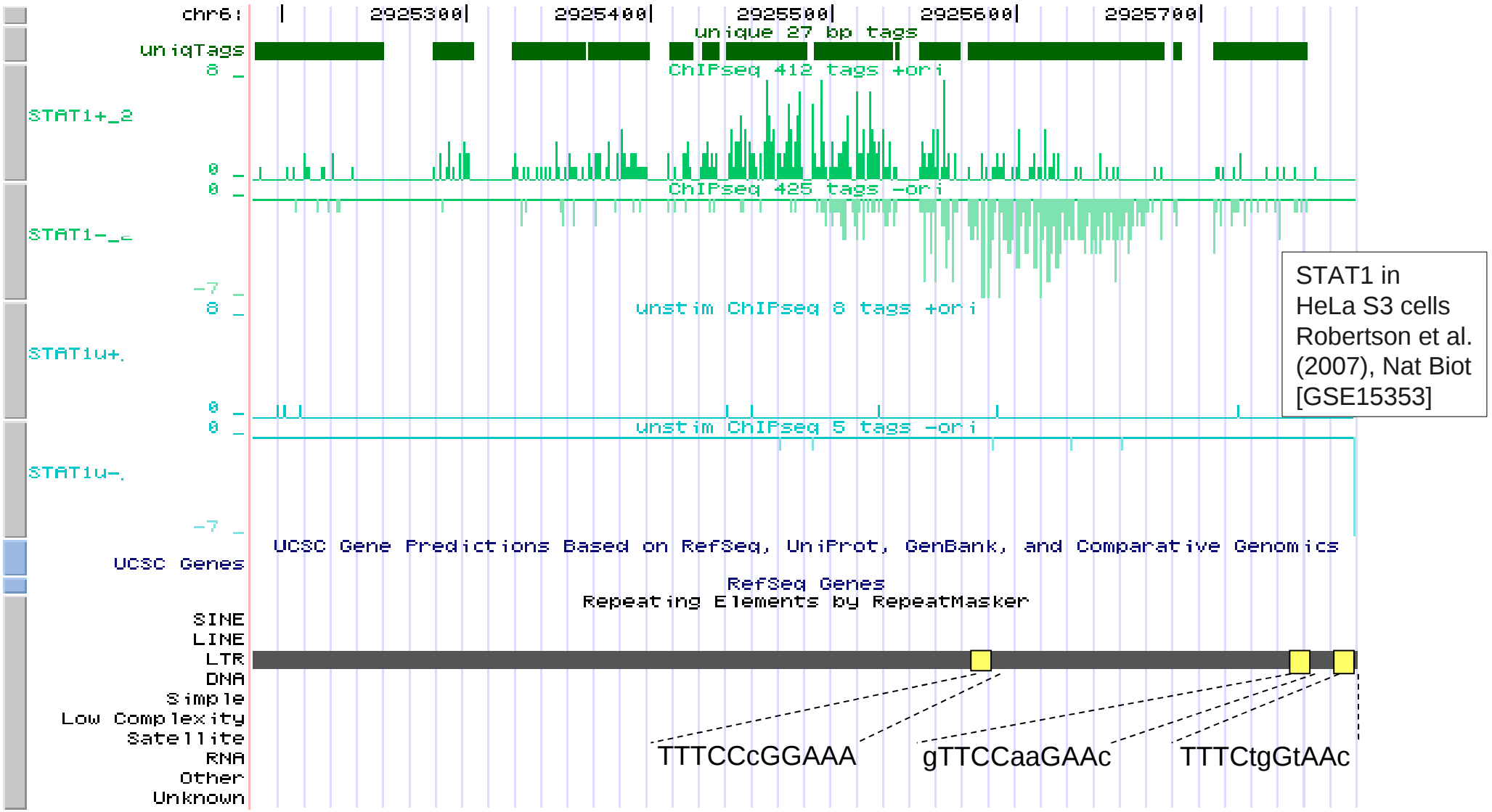
Fouse et al., (2010) Epigenomics

Popular tools to infer domains

Swiss TPH

Name of toolg	Main functionalities & web site	publication
Current standard for ChIP-seq data at UCSC from BROAD & ENCODE		
Igvtools count	fragment densities http://www.broadinstitute.org/software/igv	Robinson et al., 2011
Scripture	discrete intervals, (orig. dev. for RNA-seq !?) http://www.broadinstitute.org/software/scripture/	Gutman et al., 2010
Open source approaches		
MACS	'peak' caller yielding large domains http://liulab.dfci.harvard.edu/MACS/ (see also reviews: Wilbanks & Facciotti, Szalkowski & Schmid	Zhang et al., 2008
RSEG	based on hidden Markov model (HMM) framework http://smithlab.cmb.usc.edu/histone/rseg/	Song & Smith, 2011
ChIP-part	simple partitioning to find very large signal-enriched regions http://ccg.vital-it.ch/chipseq/chip_part.html	in prep.
various		
ChromaBlocks	Signal over-representation in bins, relative to control	Hawkins et al., 2010
Cistrome Analysis pipeline	Galaxy clone at Harvard with additional functionalities	
Various ad-hoc solutions	DNA methylation maps (derived from CHARM, MeDIP; for popular exp. procedures see review by Fouse et al., 2010)	

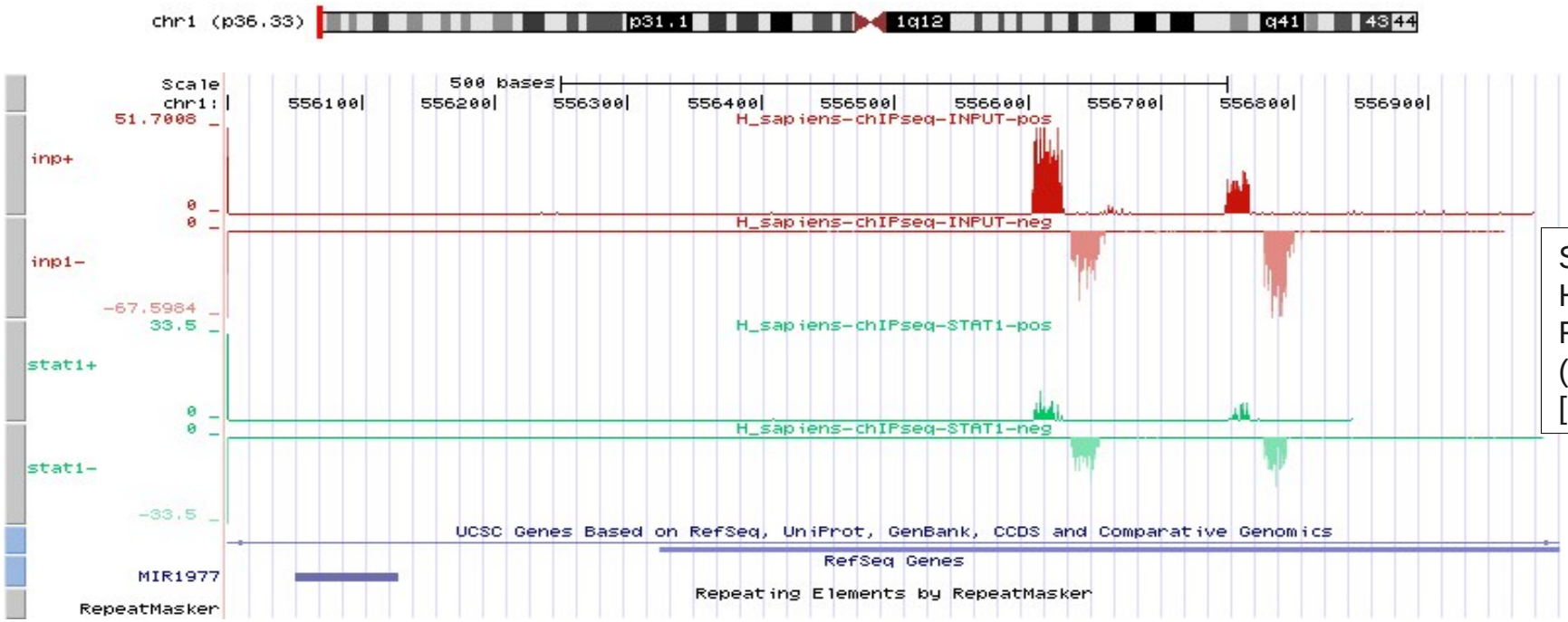
Visualizing single locus



STAT1 in HeLa S3 cells Robertson et al. (2007), Nat Biot [GSE15353]

http://genome.ucsc.edu

Limitations of ChIP controls



STAT1 in HeLa S3 cells Rozowsky et al. (2008), Nat Biot [GSE12782]

- repetitive sequences (telomers) with higher risk of artifacts
- certain variability in efficiencies of precipitation experiments
- considerable differences in 'populations' of sequences in experiment vs. control
- specific loci (i.e. open chromatin):
  - → higher accessibility of binding sites of ie. transcription factors
  - → higher tendency for 'unspecific' background

Viewers of whole genome epigenetic maps

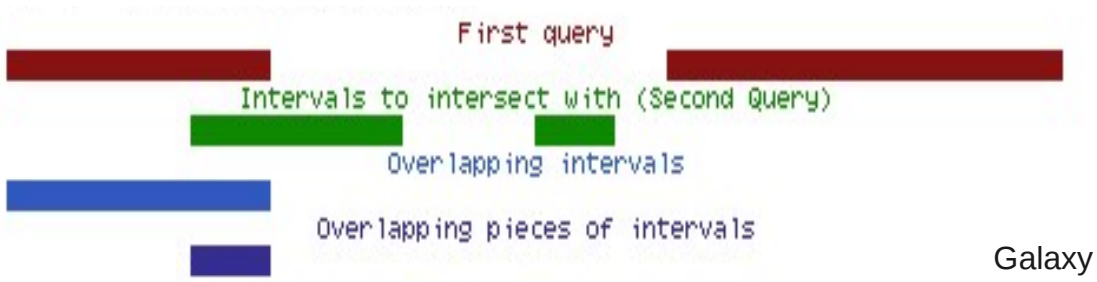
Swiss TPH

Name of tool	Main functionalities	Web site	publication
UCSC genome browser	Linked with Galaxy	genome.ucsc.edu	Kent et al., 2002
Integrative genomics viewer	Based on igvtools	www.broadinstitute.org/softwar	Robinson et al., 2011
Human Epigenome Atlas	maps produced by the NIH Roadmap Epigenomics project	www.epigenomeatlas.org	Bernstein et al., 2010
MethMarker	Includes epigenetic primer-design tool	methmarker.mpi-inf.mpg.de/	Schuffler et al., 2009

Systematic comparisons of epigenetic maps

Swiss TPH

Compute overlaps of genomic intervals:

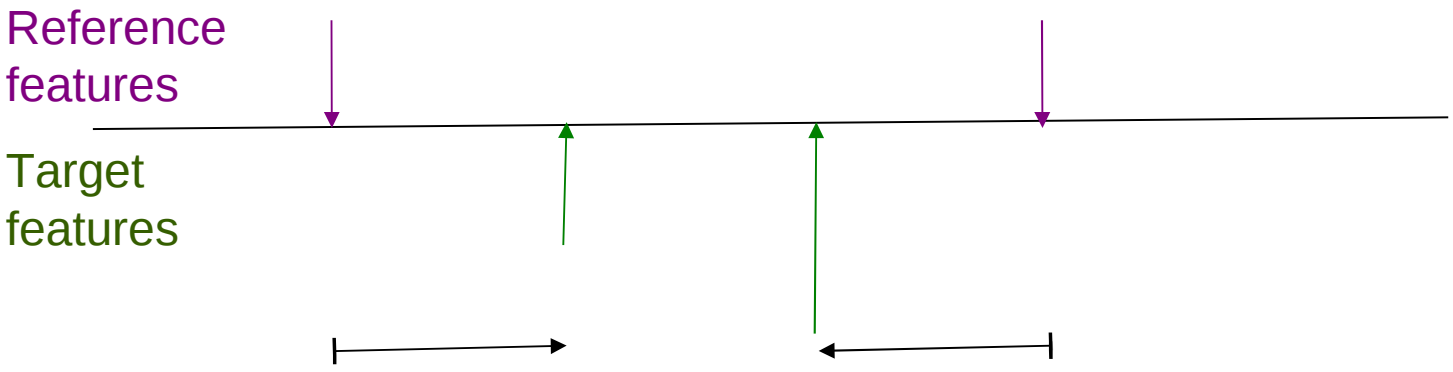


Galaxy →
Operate on Genomic Intervals
Intersect the intervals of two datasets

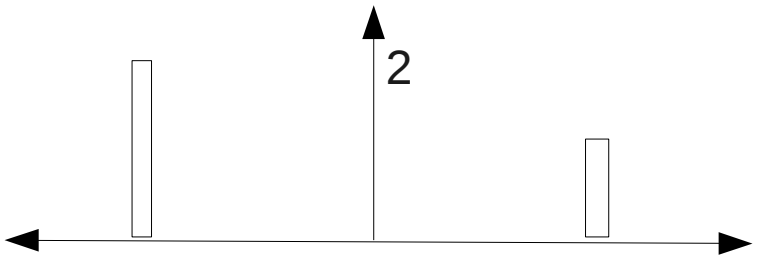
number of common and different features (Venn diagrams)

Hawkins, R.D. et al. (2010) Cell Stem Cell
using VennMaster software

Positional correlation of features

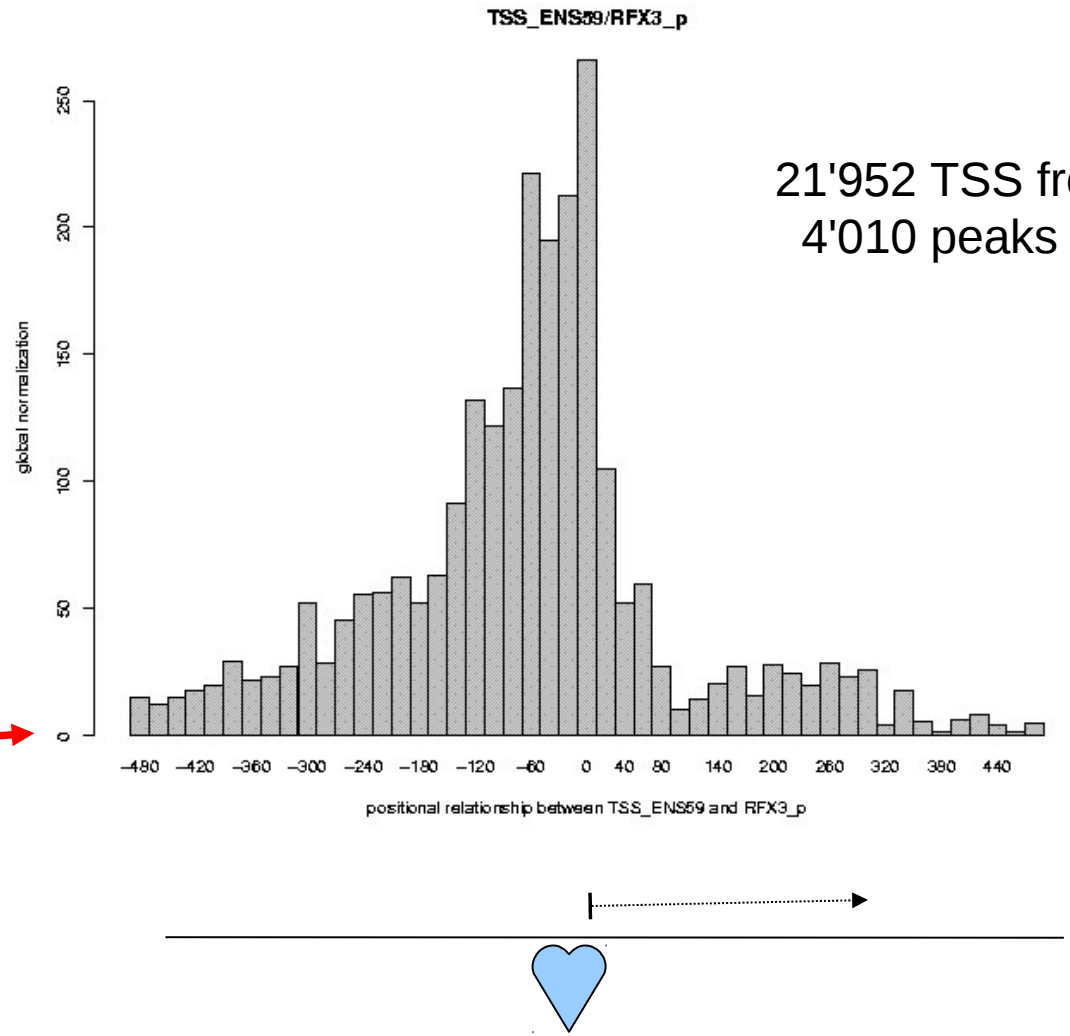


histogram of observed distances



Chipseq web server: http://ccg.vital-it.ch/chipseq/chip_cor.html

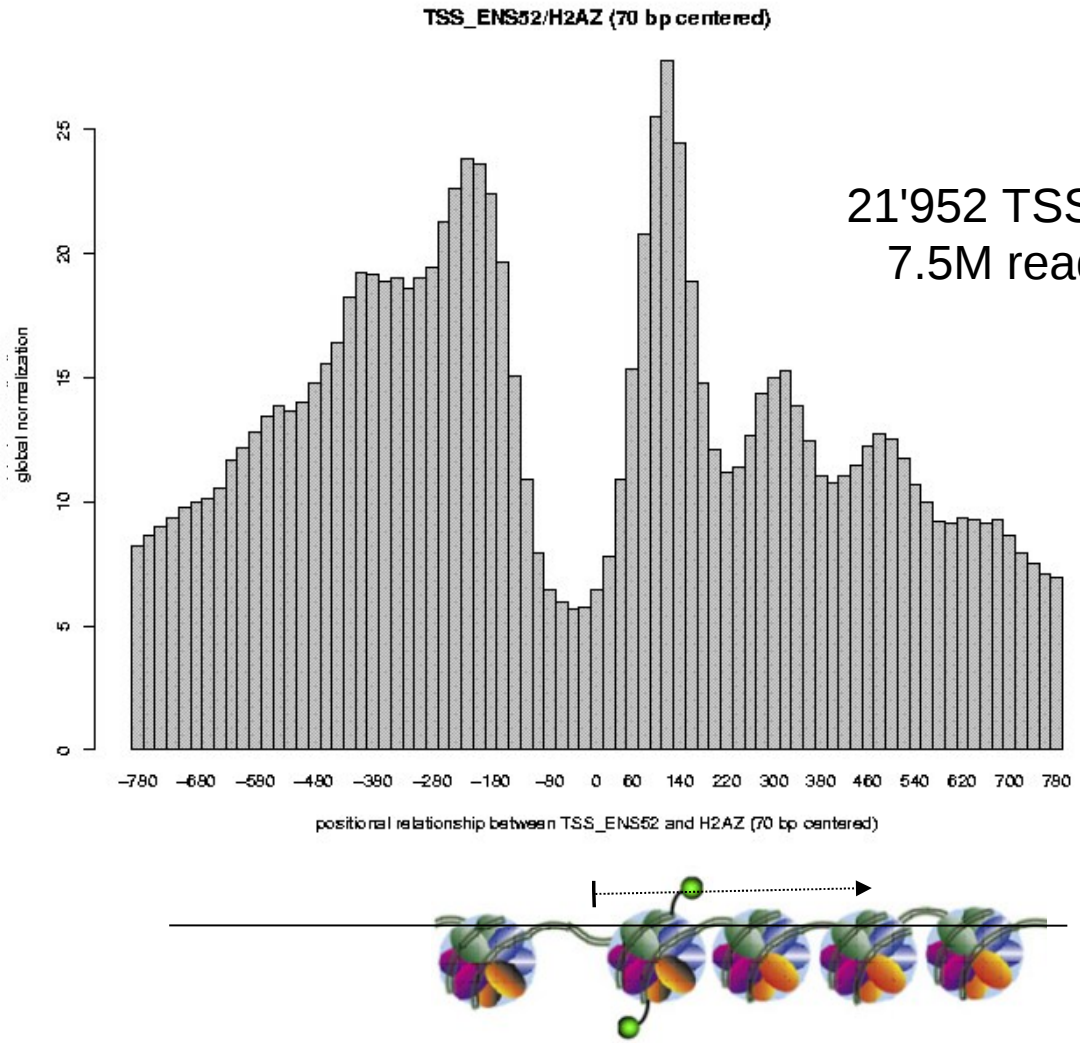
Transcription factor binding at TSS



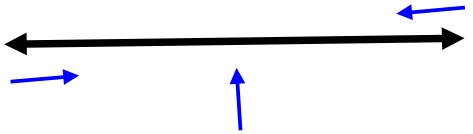
21'952 TSS from ENSEMBL
4'010 peaks from RFX3 ChIP-seq

average over genome = 1

ChIP-Seq Data Reveal Nucleosome Architecture of Human Promoters



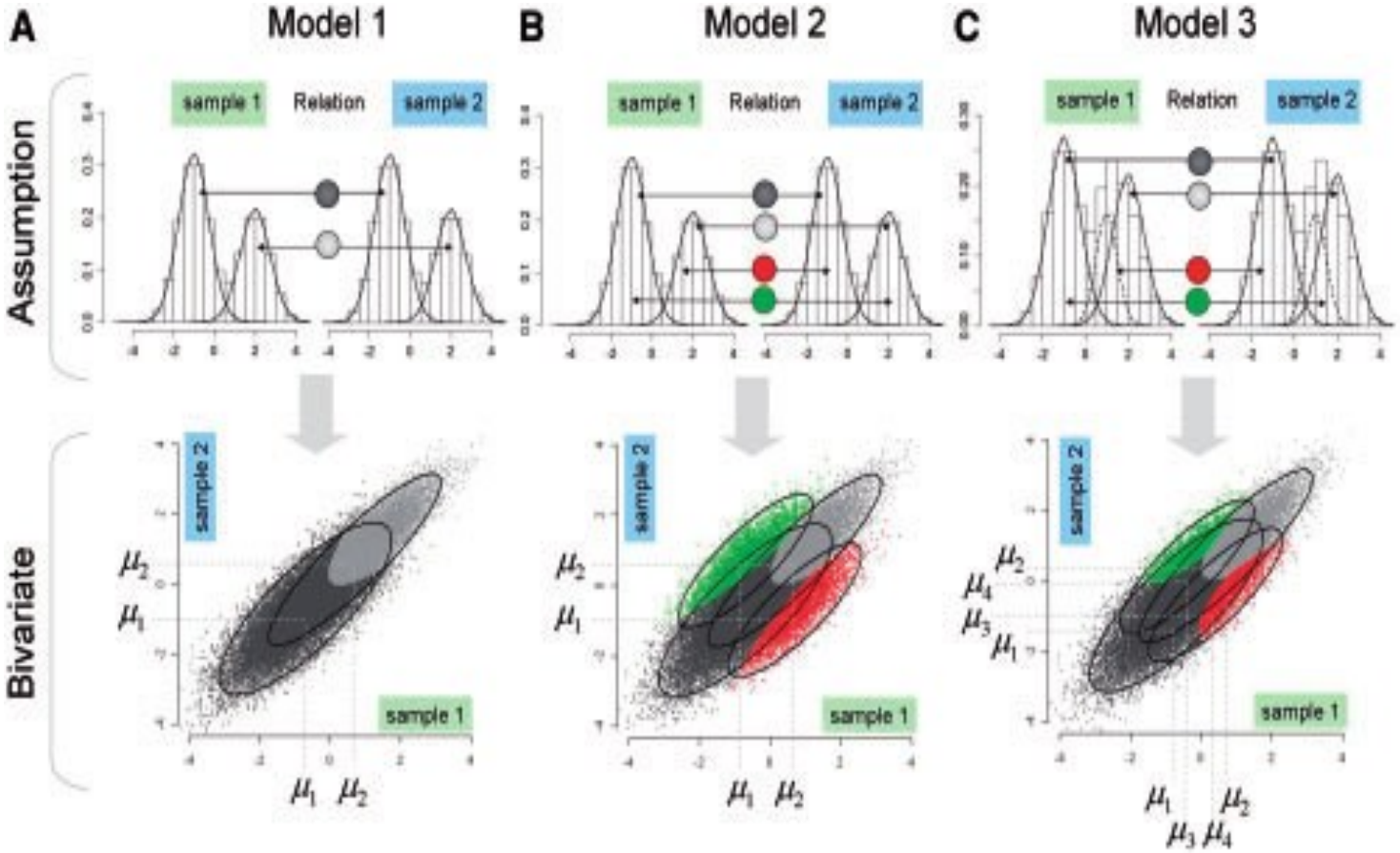
21'952 TSS from ENSEMBL
7.5M reads from H2.AZ ChIP-seq



Schmid and Bucher, (2007) Cell
Original data: Barski et al., (2007) Cell

Compare signal intensities

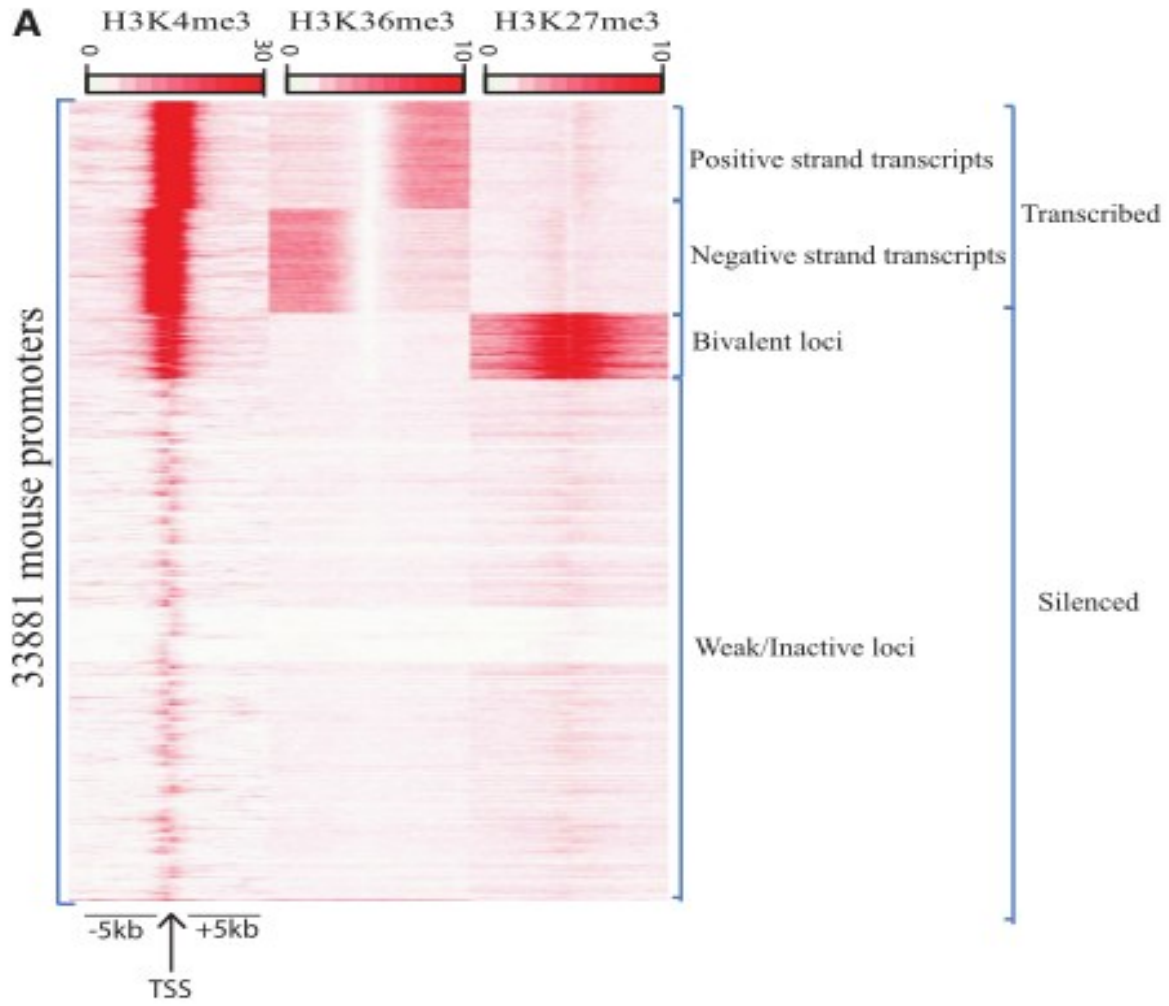
Signal intensities frequently not really 'on-off'
→ identify loci with specific profile of multiple epigenetic marks



Johannes et al.,
(2010) Bioinformatics

parametric classification approach (EM parameter estimation)
+ flexible and extensible to more dimensions (samples)
- based on R and limited accessibility to application-oriented users

Clustering of loci based on epigenetics



seqMINER | Ye et al., (2010) NAR

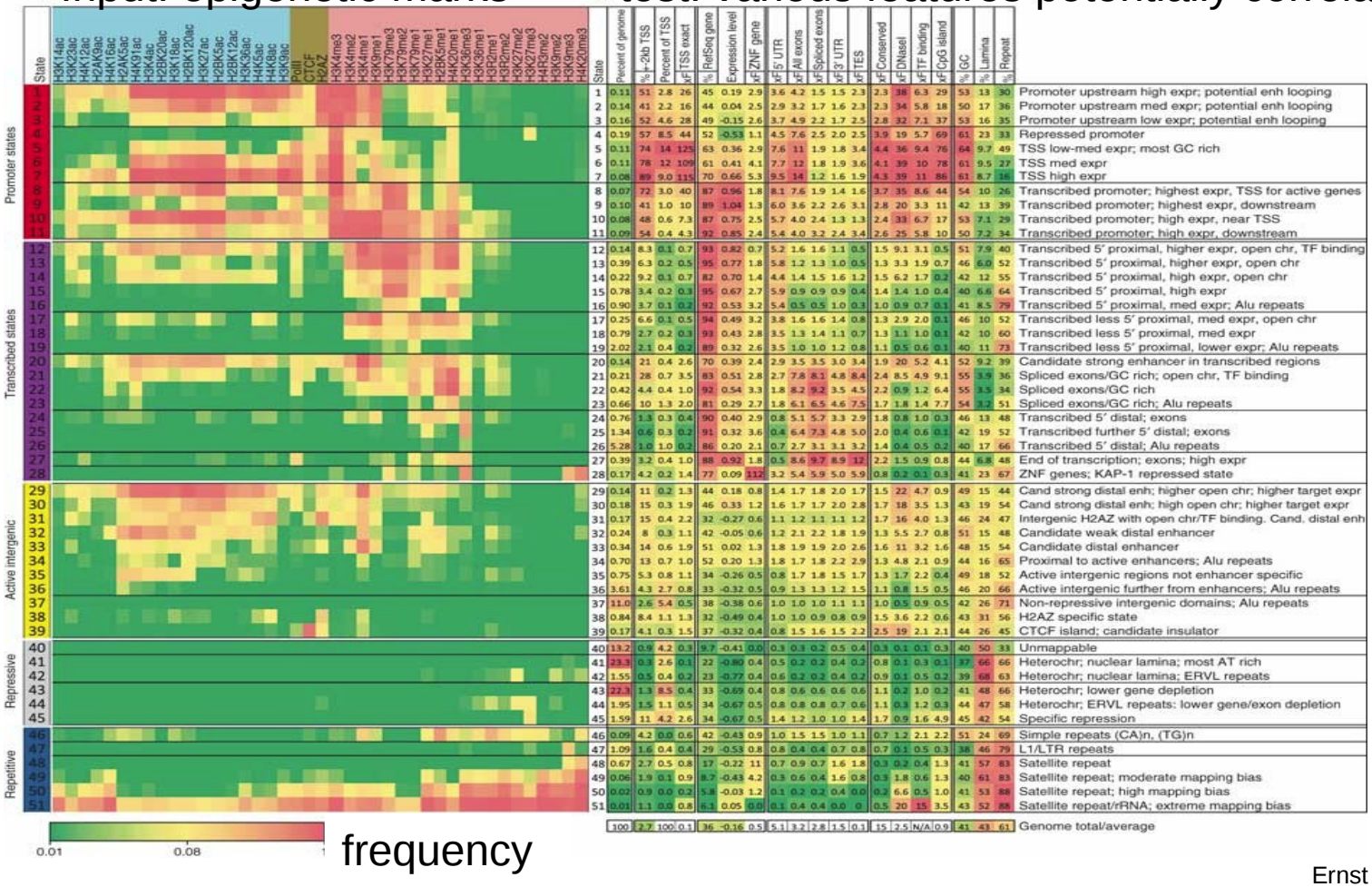
extensive normalization of data and k-means clustering
+ open source code
- limited flexibility

Epigenetic patterns, profiles, states

Swiss TPH

Input: epigenetic marks test: various features potentially correlating

51 distinct chromatin states



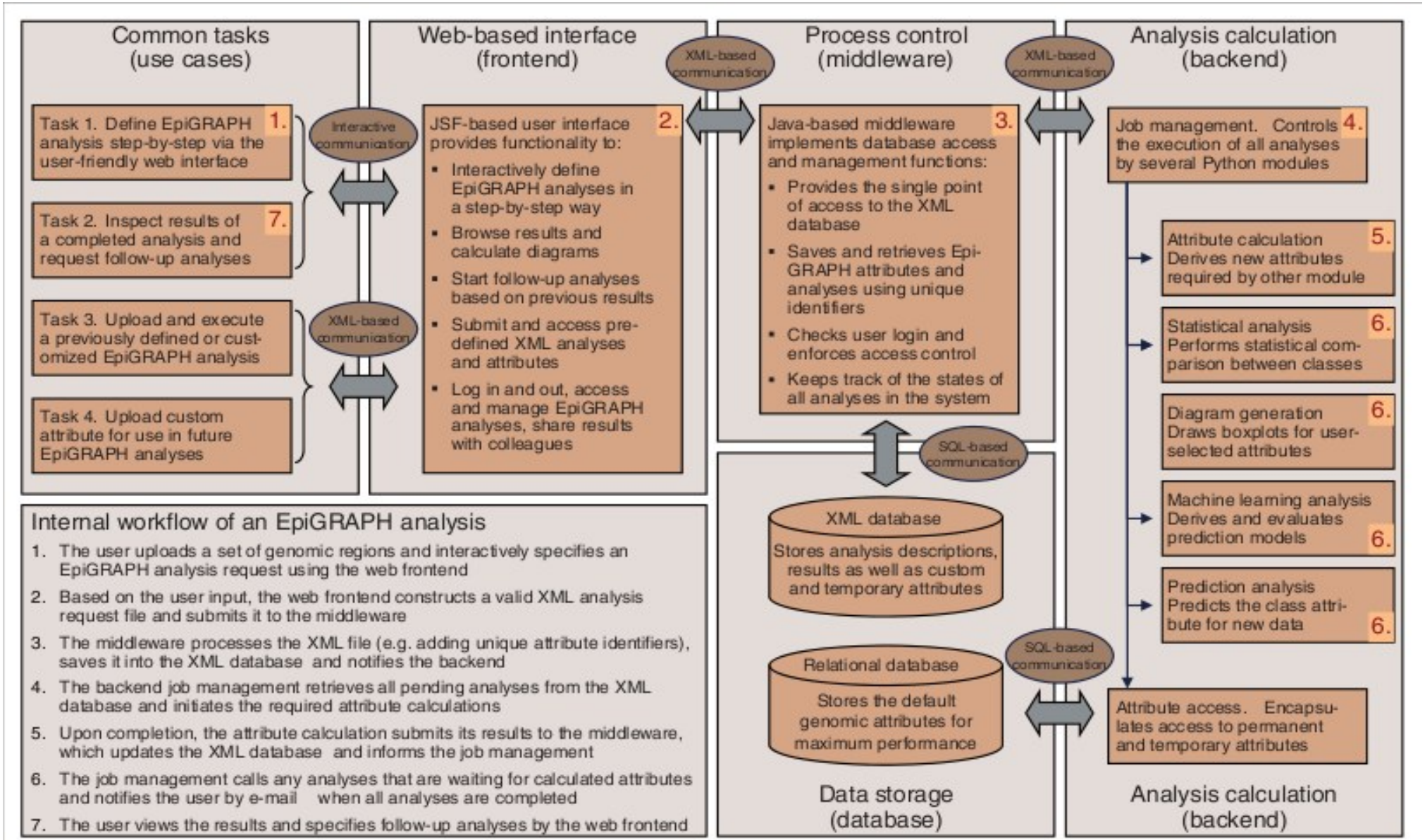
Ernst & Kellis, (2010) Nat Biotech

based on multivariate Hidden Markov Models

+ comprehensive approach potentially including all data sources

- approach not compacted in freely accessible code for application to user data

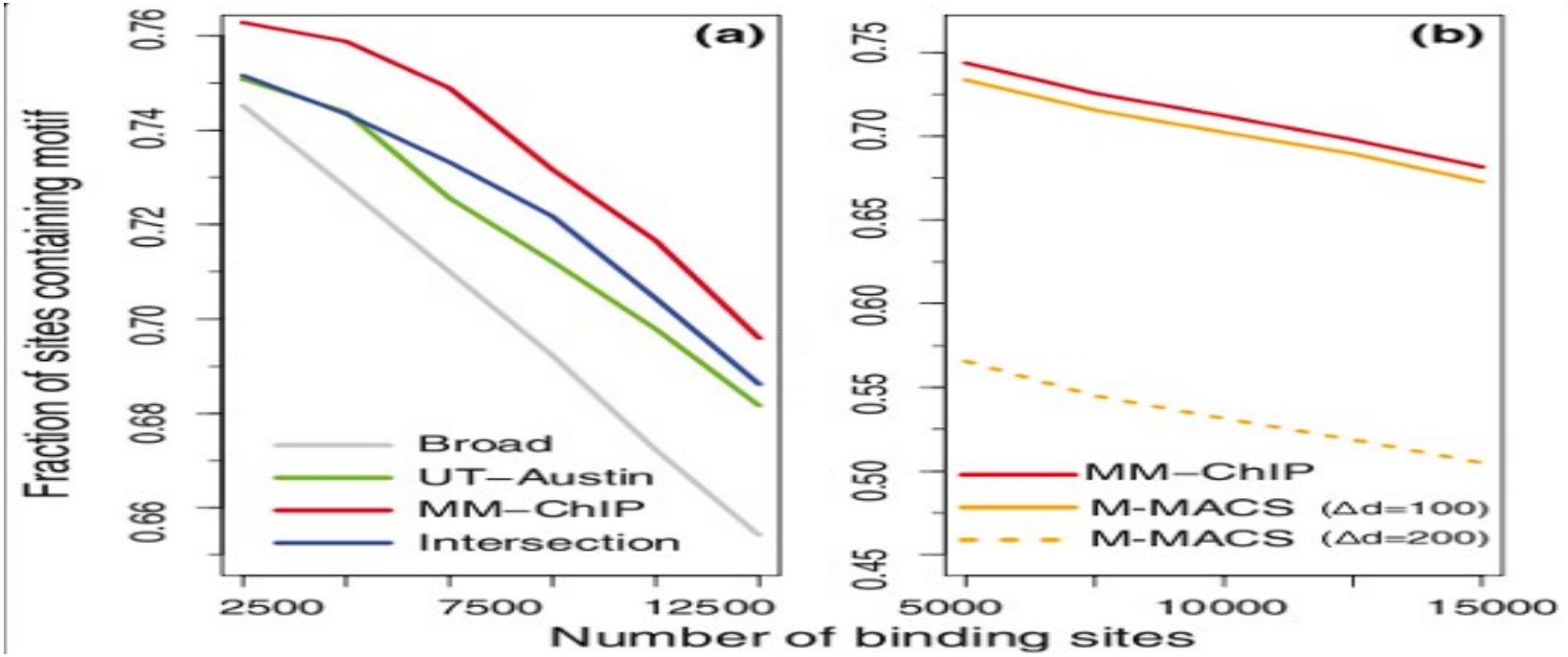
EpiGRAPH: user-friendly software for statistical analysis and prediction of (epi)genomic data



Bock et al., (2009) Genome Biology

MM-ChIP enables integrative analysis of ChIP-chip or ChIP-seq data

Example: assess content of (known!) binding sites in different data sets



Chen et al., (2011) Genome Biology
Motif content as evaluation criteria used previously elsewhere

Model-based Meta-analysis of ChIP data

- assess effect of increasing stringency
- extensible to any measure for 'true positive' (gene expression, ...)
- do not vary >1 parameter (method, lab, cell type, analysis procedure, ...)

Accessible approaches to compare epigenetic maps

Swiss TPH

Name of tool	Main functionalities	Web site	publication
GALAXY	Close integration with UCSC genome browser	main.g2.bx.psu.edu/	Goecks et al., 2010
EpiGRAPH	web toolkits, customizable work-flows based on xml database	epigraph.mpi-inf.mpg.de/Web	Bock et al., 2009
Chipseq web server	positional correlations, peak caller and partitioning (domains),	ccg.vital-it.ch/chipseq/	in prep.
seqMINER	Clustering based on epigenetic signature	bips.u-strasbg.fr/seqminer/	Ye et al., 2010
mm-chip	Integrative analysis of data sets from ChIP-chip and ChIP-seq	liulab.dfci.harvard.edu/MM-Ch	Chen et al., 2011
	R packages in bioconductor [computations in native R not overly efficient]		
-	parametric classification	[*R source code*]	Johannes et al., 2010
ChIPpeakAnno	association of enriched regions → genome annotations from BioMart	via www.bioconductor.org	Zhu et al., 2010
Repitools	toolbox to visualize epigenomic data	repitools.r-forge.r-project.org/	Statham et al., 2010

Summary

Swiss TPH

State of the art in the analysis of epigenetic maps

intensive data production
→ few established procedures for data analysis

Many specialized ad-hoc solutions
→ growing number of tools
with partially overlapping functionalities

Biologists typically lack IT skills
- 'user friendliness' vs. transparency and versatility
- web servers relieve installation and maintenance of local computational setups
  - Data transfer rate over internet limited

Suggestions for review

- Identify latest (and possibly overlooked) publications on topic (extremely dynamic ...)
- Potential structure of review

  -> Summarize currently proposed solutions

  -> inform reader on advantages and limitations

  -> Eventually provide outlook on future development