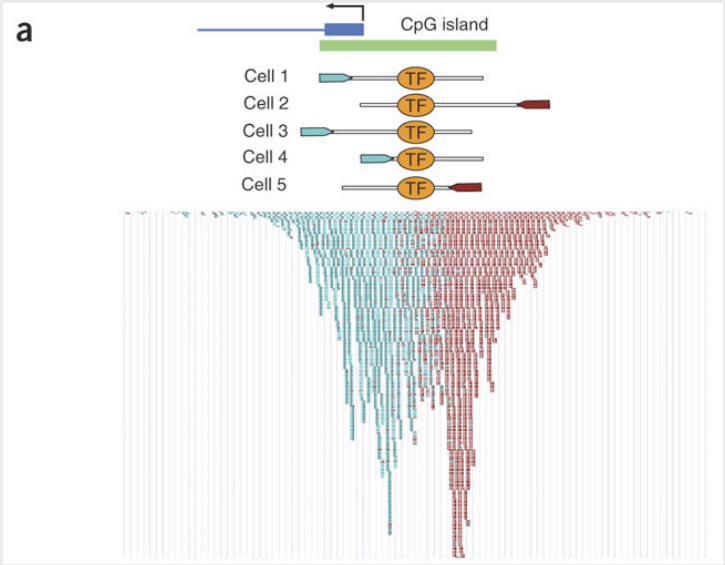
Charles Girardot, Furlong Lab

MACS, CisGenome, SISSRs and other peak calling algorithms: differences and practical use

EMBL

ChIP-Seq signal properties



- Only 5’ ends of ChIPed fragments are sequenced
  - ✓ Shifted read distribution
  - ✓ Expected symmetry between Watson/Crick read distributions

Figure source: Valouev et al. Nat. Methods Sept 2008

Peak finding overview

1. Build strand-specific profiles
   - How (window-scan, KDE…)?
   - Filter duplicates?

2. Combine profiles (shift/extension)
   - Shift/extension estimation?

1. Define enriched regions/peaks
   - Statistics used
   - What boundaries should be reported?
   - What score to use (ratio, p-val, q-val)?
   - Compute/estimate a FDR?

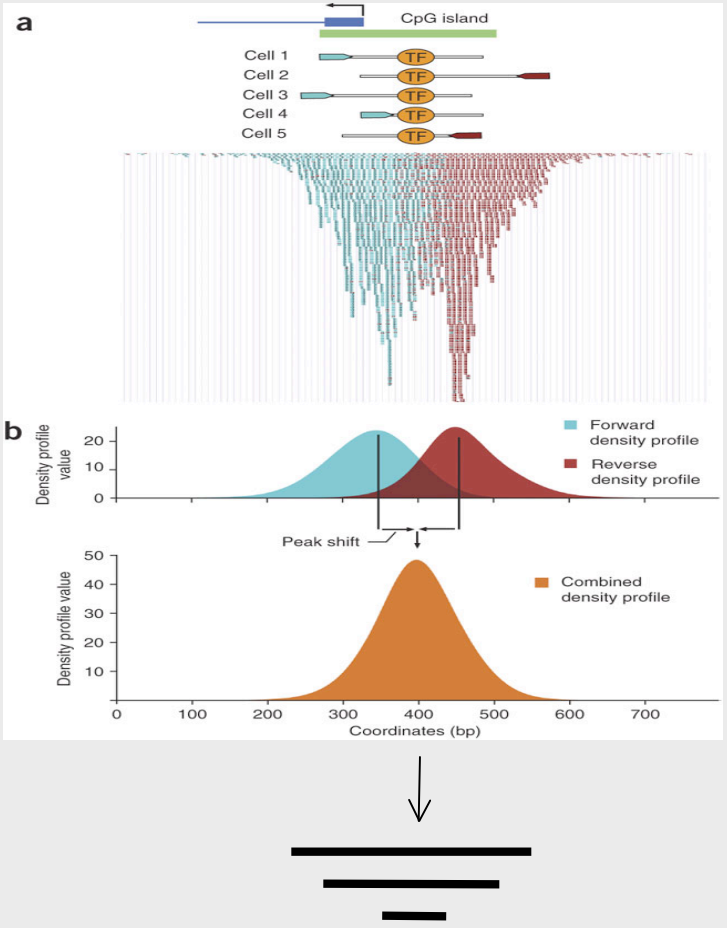


Figure source: Valouev et al. Nat. Methods Sept 2008

EMBL

Main aspects of peak finders

	Profile	Peak criteria[a]	Tag shift	Control data[b]	Rank by	FDR[c]	User input parameters[d]	Artifact filtering: strand-based/ duplicate[e]	Refs.
CisGenome v1.1	Strand-specific window scan	1: Number of reads in window 2: Number of ChIP reads minus control reads in window	Average for highest ranking peak pairs	Conditional binomial used to estimate FDR	Number of reads under peak	1: Negative binomial 2: conditional binomial	Target FDR, optional window width, window interval	Yes / Yes	10
ERANGE v3.1	Tag aggregation	1: Height cutoff High quality peak estimate, per-region estimate, or input	High quality peak estimate, per-region estimate, or input	Used to calculate fold enrichment and optionally *P* values	*P* value	1: None 2: $\frac{\#\ control}{\#\ ChIP}$	Optional peak height, ratio to background	Yes / No	4,18
FindPeaks v3.1.9.2	Aggregation of overlapped tags	Height threshold	Input or estimated	NA	Number of reads under peak	1: Monte Carlo simulation 2: NA	Minimum peak height, subpeak valley depth	Yes / Yes	19
F-Seq v1.82	Kernel density estimation (KDE)	*s* s.d. above KDE for 1: random background, 2: control	Input or estimated	KDE for local background	Peak height	1: None 2: None	Threshold s.d. value, KDE bandwidth	No / No	14
GLITR	Aggregation of overlapped tags	Classification by height and relative enrichment	User input tag extension	Multiply sampled to estimate background class values	Peak height and fold enrichment	2: $\frac{\#\ control}{\#\ ChIP}$	Target FDR, number nearest neighbors for clustering	No / No	17
MACS v1.3.5	Tags shifted then window scan	Local region Poisson *P* value	Estimate from high quality peak pairs	Used for Poisson fit when available	*P* value	1: None 2: $\frac{\#\ control}{\#\ ChIP}$	*P*-value threshold, tag length, mfold for shift estimate	No / Yes	13
PeakSeq	Extended tag aggregation	Local region binomial *P* value	Input tag extension length	Used for significance of sample enrichment with binomial distribution	*q* value	1: Poisson background assumption 2: From binomial for sample plus control	Target FDR	No / No	5
QuEST v2.3	Kernel density estimation	2: Height threshold, background ratio	Mode of local shifts that maximize strand cross-correlation	KDE for enrichment and empirical FDR estimation	*q* value	1: NA 2: $\frac{\#\ control}{\#\ ChIP}$ as a function of profile threshold	KDE bandwidth, peak height, subpeak valley depth, ratio to background	Yes / Yes	9
SICER v1.02	Window scan with gaps allowed	*P* value from random background model, enrichment relative to control	Input	Linearly rescaled for candidate peak rejection and *P* values	*q* value	1: None 2: From Poisson *P* values	Window length, gap size, FDR (with control) or *E*-value (no control)	No / Yes	15
SiSSRs v1.4	Window scan	$N_+ - N_-$ sign change, $N_+ + N_-$ threshold in region[f]	Average nearest paired tag distance	Used to compute fold-enrichment distribution	*P* value	1: Poisson 2: control distribution	1: FDR 1,2: $N_+ + N_-$ threshold	Yes / Yes	11
spp v1.0	Strand specific window scan	Poisson *P* value (paired peaks only)	Maximal strand cross-correlation	Subtracted before peak calling	*P* value	1: Monte Carlo simulation 2: $\frac{\#\ control}{\#\ ChIP}$	Ratio to background	Yes / No	12
USeq v4.2	Window scan	Binomial *P* value	Estimated or user specified	Subtracted before peak calling	*q* value	1, 2: binomial 2: $\frac{\#\ control}{\#\ ChIP}$	Target FDR	No / Yes	20

Pepke et al ; *Nature Methods* ***6, S22 - S32 (2009)***

PeakSeq

ARTICLES

nature biotechnology

PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls

Joel Rozowsky[1], Ghia Euskirchen[2], Raymond K Auerbach[3], Zhengdong D Zhang[1], Theodore Gibson[1], Robert Bjornson[4], Nicholas Carriero[4], Michael Snyder[1,2] & Mark B Gerstein[1,3,4]

PeakSeq

- Sequence tags from certain location are not unique in the genome
- Tags that don't uniquely map are usually discarded

Table 1 Genome mappability fraction

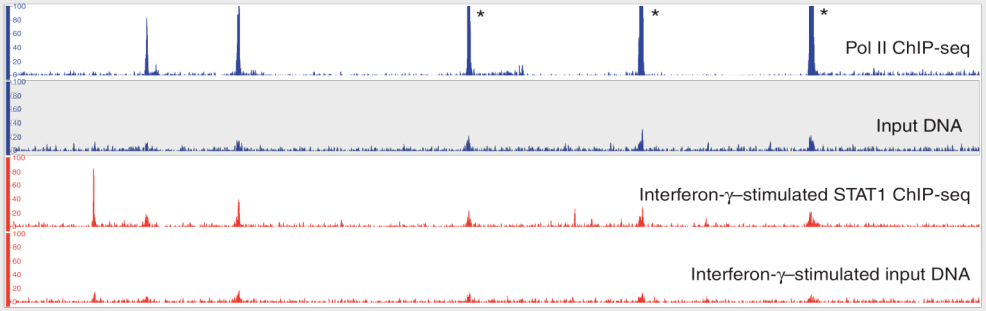
		Nonrepetitive sequence		Mappable sequence	
Organism	Genome size (Mb)	Size (Mb)	Percentage	Size (Mb)	Percentage
Caenorhabditis elegans	100.28	87.01	86.8%	93.26	93.0%
Drosophila melanogaster	168.74	117.45	69.6%	121.40	71.9%
Mus musculus	2,654.91	1,438.61	54.2%	2,150.57	81.0%
Homo sapiens	3,080.44	1,462.69	47.5%	2,451.96	79.6%

For four common model organisms—worm, fruit fly, mouse and human—we have determined the fraction of each genome sequence that is nonrepetitive as well as the fraction that is mappable using 30-nt sequence tags. The genome coverage achievable from genomic tiling arrays corresponds to the nonrepetitive fraction of a genome whereas the mappable coverage is what is achievable by tag-based sequencing approaches. We also determined that as the length of the sequence tags is increased beyond 30, the number of nucleotides in the genomes that are uniquely mappable is 2,452 Mb (79.6%) for 30-nt reads, 2,586 Mb (84.0%) for 40 nt, 2,669 Mb (86.7%) for 50 nt, 2,720 Mb (88.3%) for 60 nt and 2,750 Mb (89.3%) for 70 nt.

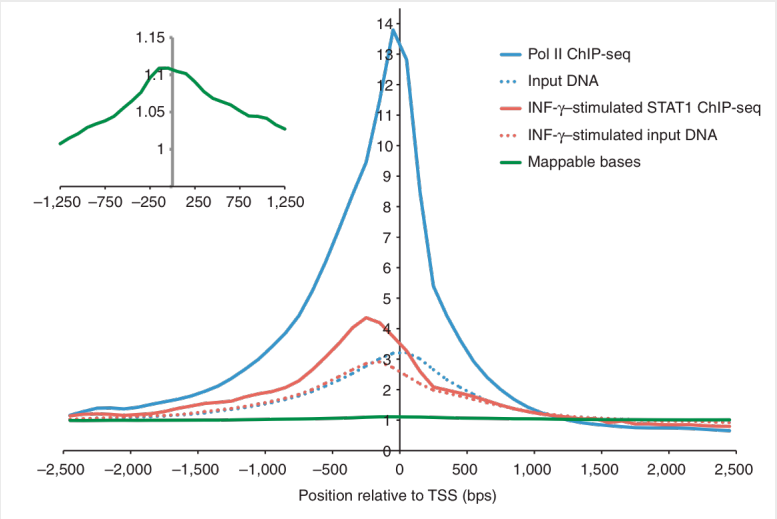
=> The fraction of the "mappable" genome is usually a parameter of peak finders

PeakSeq

- Background models are usually assumed to follow a Poisson statistics
- Unfortunately, the real background results from a multiple effects
  1. Mappability
  2. Chromatin structure (e.g. accessibility/openness)



Enrichments in reference sample is not randomly placed



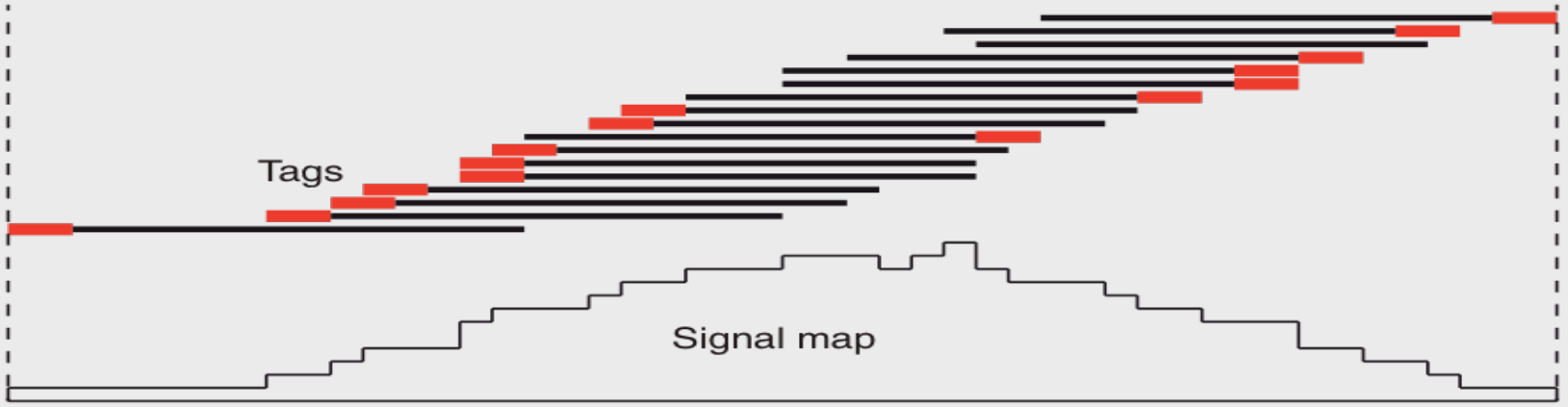
Chromatin structure is the major factor

N.B.: See also Kharchenko et al. Nat biotech 2008

EMBL

PeakSeq

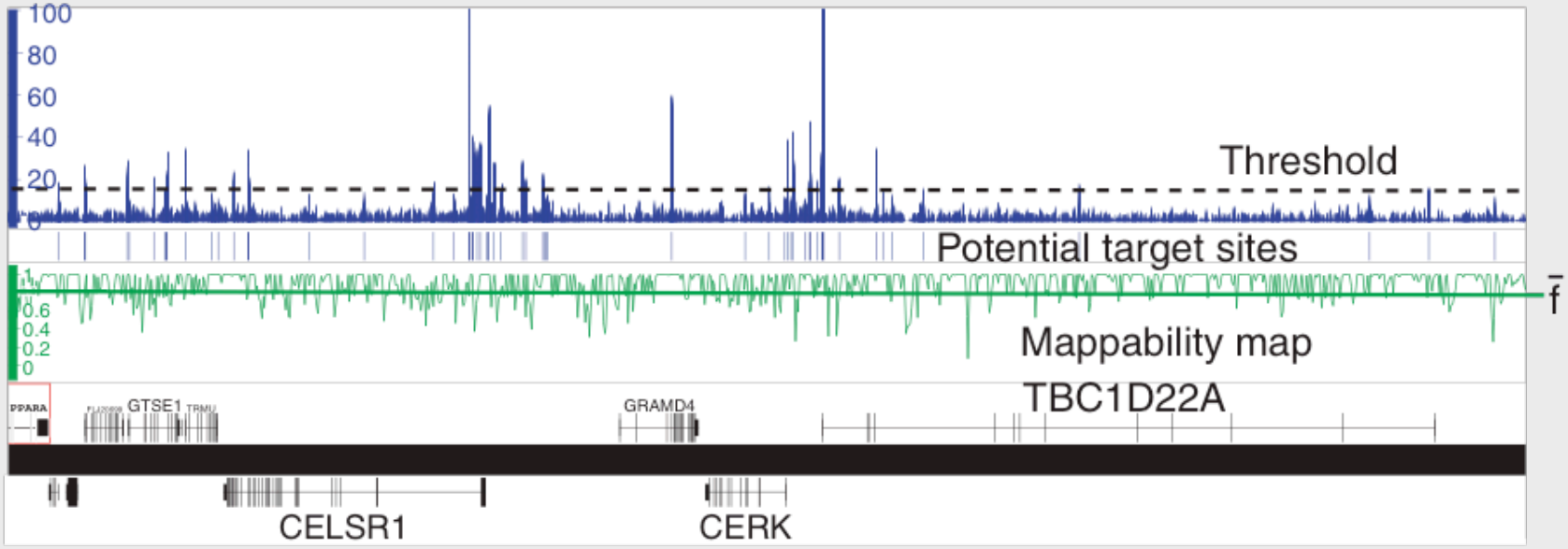
Step 1: Signal map(s) construction



1. Tag extension (user input)
2. Signal map : count for each bp

PeakSeq

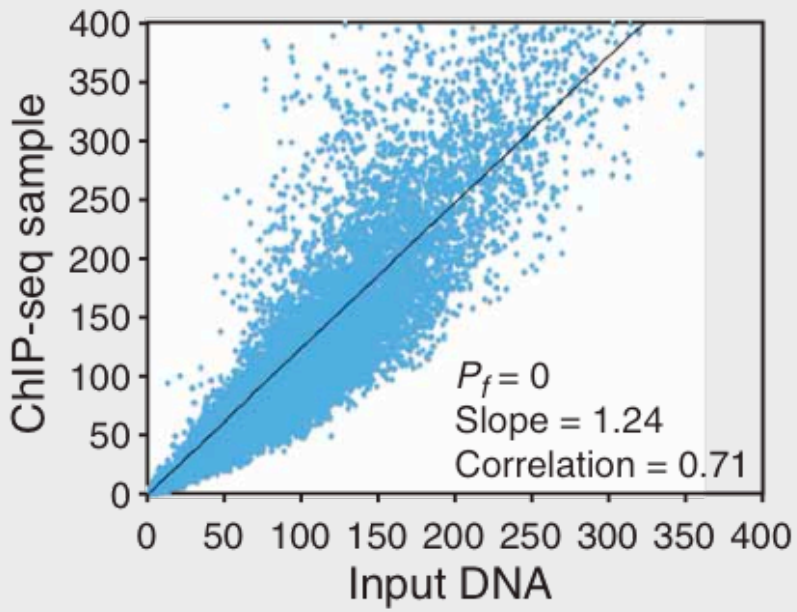
Step 2: Determination of potential regions using simulated bg

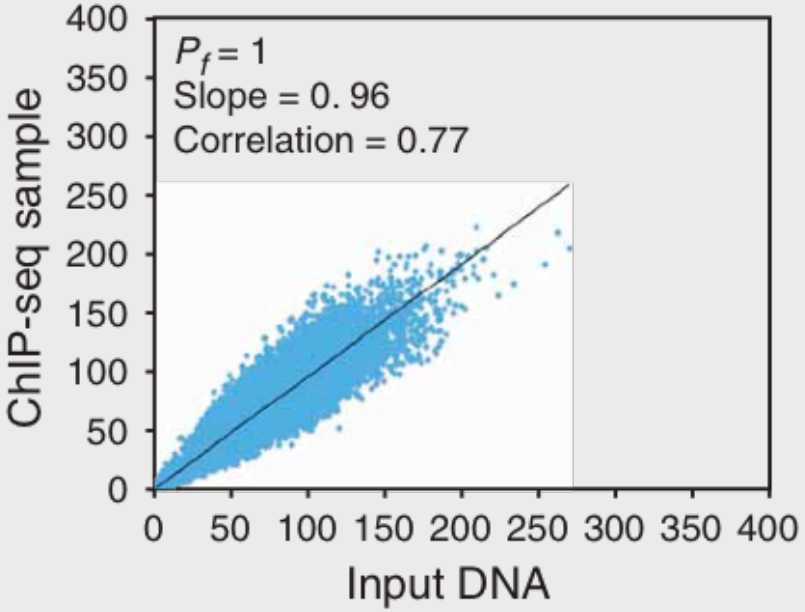


1. Uses Poisson statistics
2. Work per window (1 Mb) and correct signal (of different windows) using mappability maps
3. Given a user-defined target FDR, a threshold is computed
4. Keep regions above threshold

PeakSeq

Step 3: Normalizing control to ChIP-seq sample

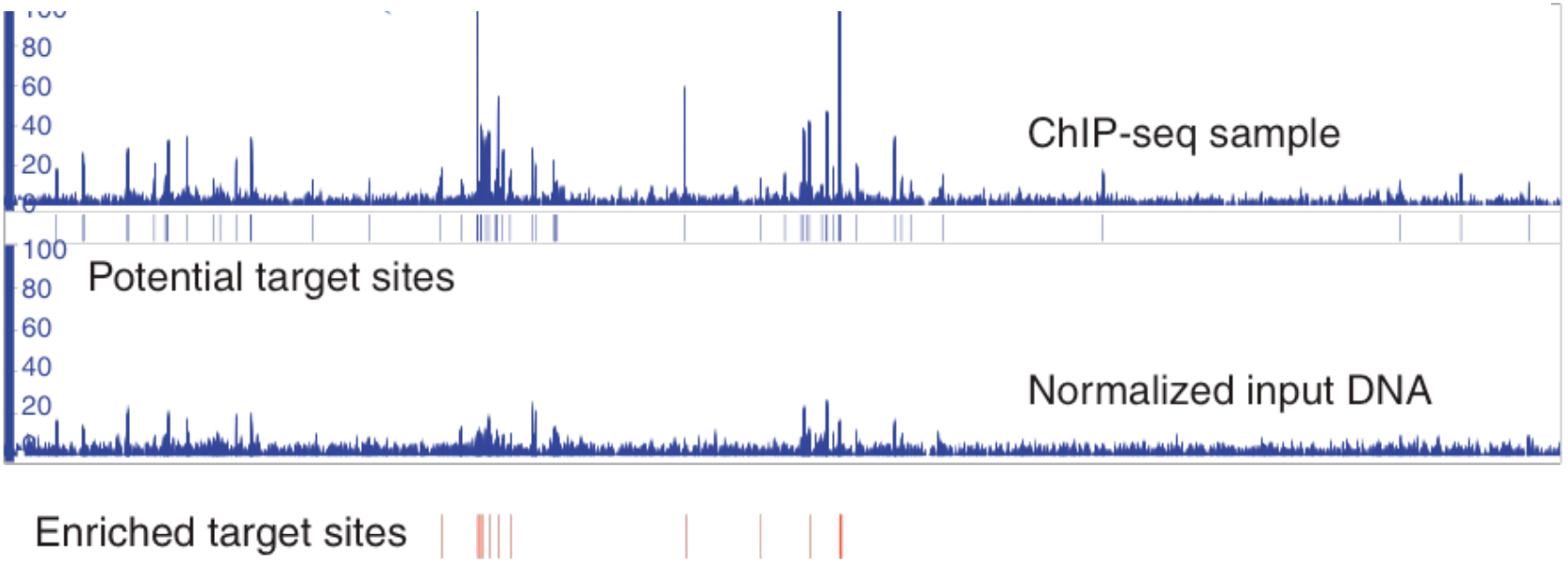




1. Count tags in bins along chromosome for ChIP-seq and reference
2. Correct tag counts using slope of linear regression
3. Pf = fraction (i.e. in [0,1]) of potential peaks to exclude

PeakSeq

Step 4: Scoring enriched target regions relative to control



1. Compute fold enrichment for each candidate (defined in step 2)
2. Compute p-value from binomial distribution
3. Correct for multiple testing and call enriched regions

What have we learned so far

- The size of the mappable genome varies with your tag length
- Background is not accurately modeled by Poisson
  - Use of input DNA is recommended
- The scaling factor between ChIP and input sample is not directly the tag ratio

MACS

Method

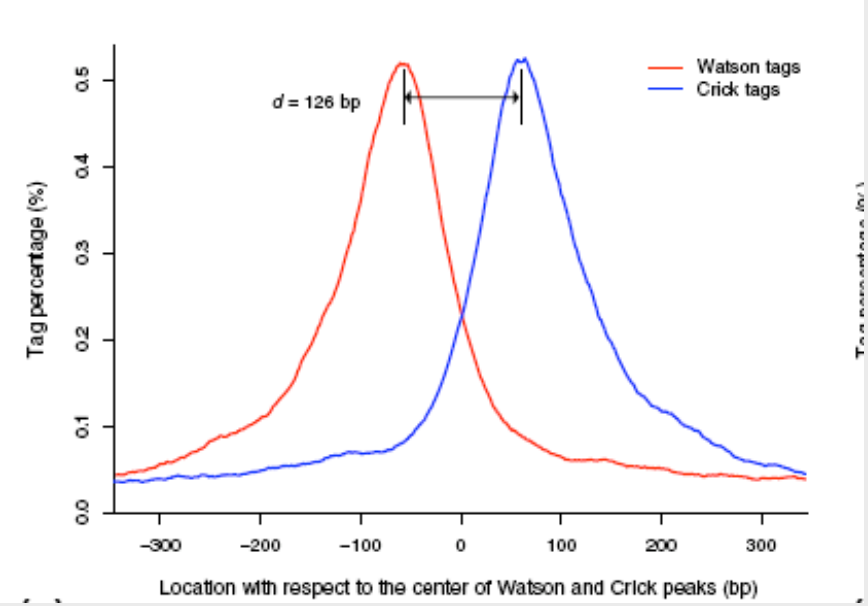
Open Access

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[*]

Genome Biology 2008, **9**:R137 (doi:10.1186/gb-2008-9-9-r137)

- Step 1: Modeling the tag shift



1. Scan genome with a window of user-defined sonication size
2. Keep the best 1000 (or less) peaks having a fold enr. > *mfold* (default 32, relative to random model)
3. Separate Watson/Crick tags
4. Shift size is modeled as the distance between the modes of the Watson and Crick peaks

MACS

- Step 2: Peak detection

1. Shift every tag by d/2
2. Slide a 2d window across the genome to find candidate peaks with significant tag enrichment (according to Poisson distribution, default p-value = 10^{-5})
3. Merge of overlapping peaks
4. Report :
    - fold enrichment for called peaks: ratio between tag counts and expected using Poisson distribution (using input data if provided)
    - Position with highest pile-up is defined as the summit
    - Empiric FDR if control sample is provided (sample swap)

MACS : key aspects

- Adaptive Poisson distribution to model background
  - Usually, this λ is computed once i.e. λ_{BG}
  - Here, they use a dynamic λ_{local} to account for local biases :
    - $\lambda_{local} = \max(\lambda_{BG}, \lambda_{1K}, \lambda_{5K}, \lambda_{10K})$

- Model the tag shift using the bimodal property of ChIP-seq reads using high confidence peaks (fold cutoff)

- Automatic removal of duplicated tags in excess of what is expected given the sequencing depth (using p-val cutoff of 10^{-5} , binomial dist.)
  - ✓ Always check the default setting for duplicates in your peak finder

CisGenome

Nature Biotechnology **26**, 1293 - 1300 (2008)
Published online: 2 November 2008 | doi:10.1038/nbt.1505

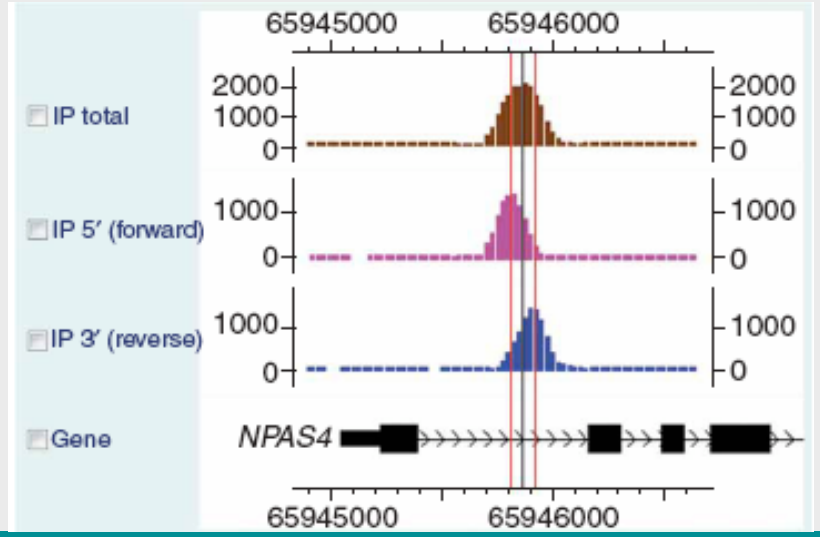
An integrated software system for analyzing ChIP-chip and ChIP-seq data

Hongkai Ji[1], Hui Jiang[2], Wenxiu Ma[3], David S Johnson[4,8], Richard M Myers[5] & Wing H Wong[6,7]

- Two pass algorithm, globally similar to MACS
- First pass:
  - scan similar to MACS (100 bp window) to evaluate DNA fragment length i.e. tag shift value
  - FDR estimation (based on non overlapping window of 100 bp) from following statistics:
    - One sample analysis : based on a negative binomial
    - Two sample analysis : tag count in IP bin evaluated against tag sum n_i (IP+ref) using binomial

CisGenome

- Second pass after tag shift : principle similar to first scan (FDR also recomputed):
  - Overlapping windows below user defined FDR are merged (best FDR is kept). In two sample analysis, the best fold change is also reported
  - Regions that do not exhibit bimodal read distribution (e.g. b/w strands) are filter out (significant strand-specific peak expected)
  - Peak boundaries may be refined using the read distributions (on by default)



EMBL

SISSRs

Published online 6 August 2008

Nucleic Acids Research, 2008, Vol. 36, No. 16 ***5221–5231***
doi:10.1093/nar/gkn488

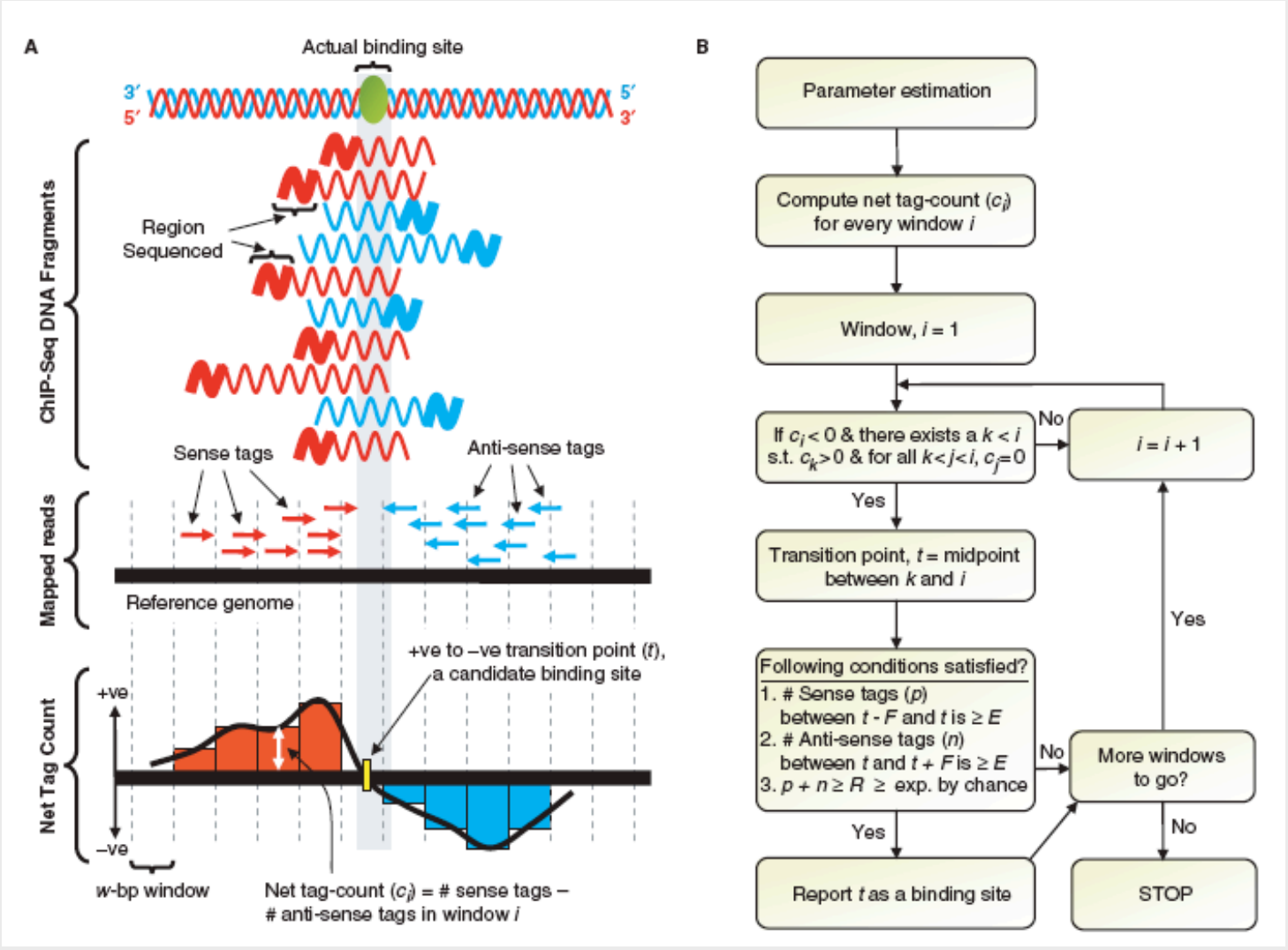
Genome-wide identification of *in vivo* protein–DNA binding sites from ChIP-Seq data

Raja Jothi, Suresh Cuddapah, Artem Barski, Kairong Cui and Keji Zhao*

Laboratory of Molecular Immunology, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, MD 20894, USA

- DNA fragment length estimated from the data
- No tag shift / extension
- FDR estimate from Poisson model or from reference dataset
- Reports TFBS location estimation i.e. very small region

SISSRs



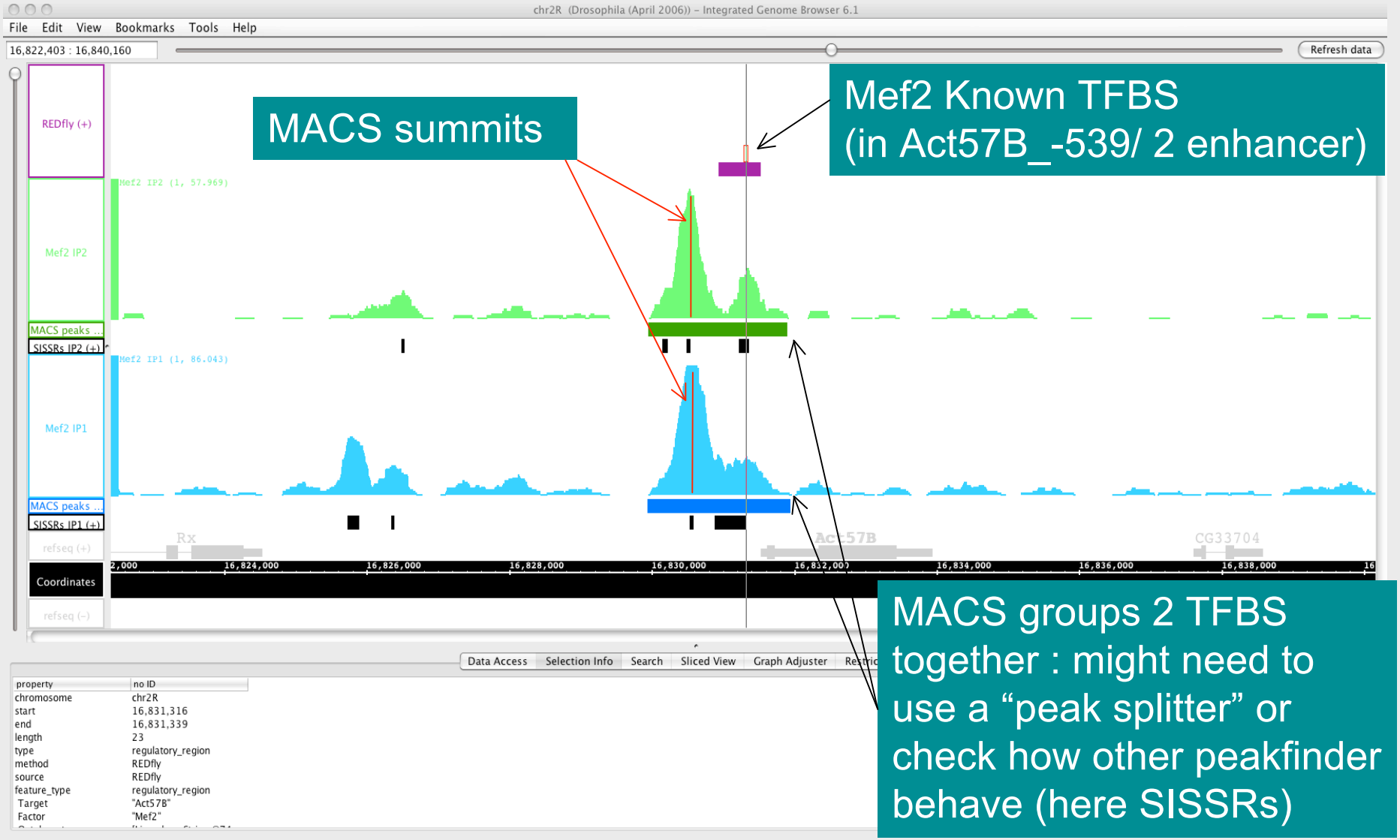
Jothi et al ; *NAR* ***36, 16 (2008)***

EMBL

What have we learned more

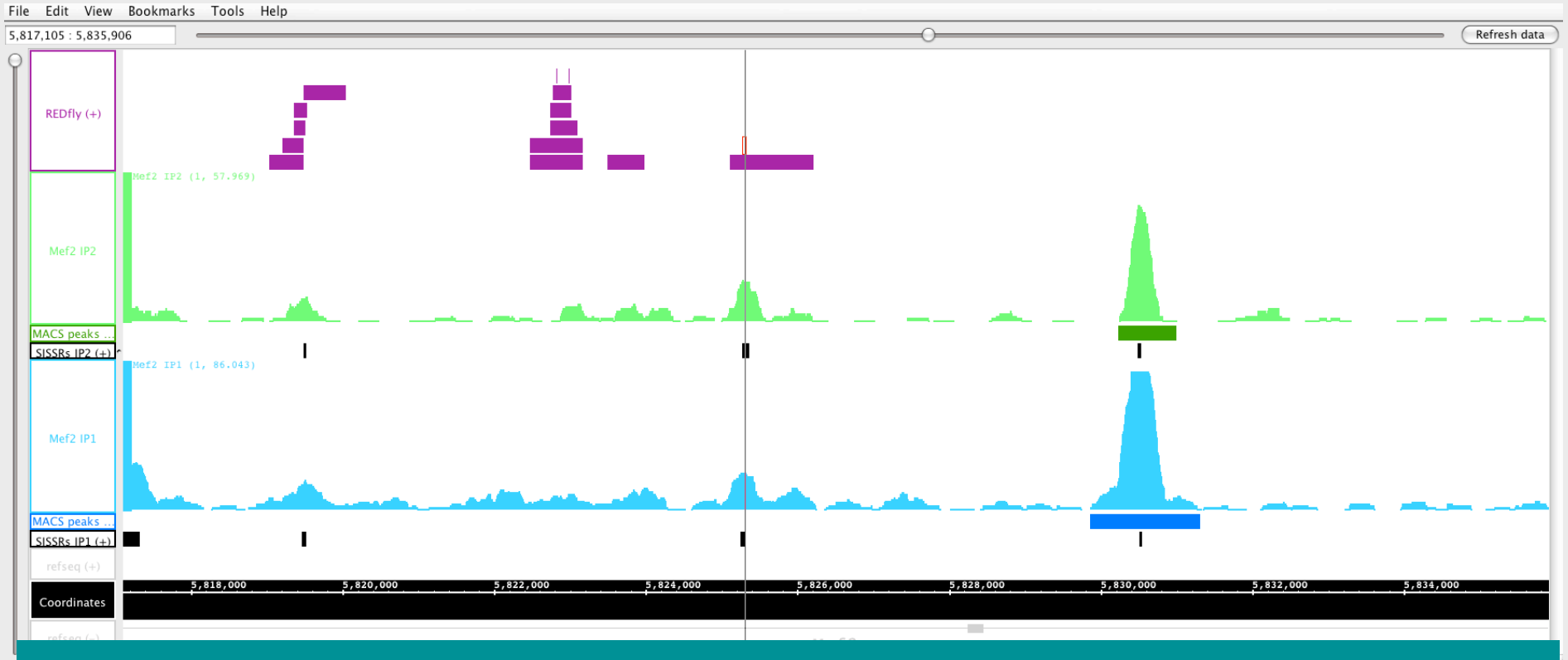
- The size of the mappable genome varies with your tag length
- Background is not accurately modeled by Poisson
  - Use of input DNA is recommended
  - If no input is available, favor methods using negative binomial (or local poisson)
- The scaling factor between ChIP and input sample is not the tag ratio
- Fragment length can be estimated from top peaks or given as input
- Usually duplicate reads are filtered, a gentler approach might be better or no filtering (?)
- Enrichment is usually reported, sometimes with FDR/q-value ; methods vary

Peaks vs enriched regions (TF ChIP-Seq)



EMBL

IGB – Another Mef2 known TFBS

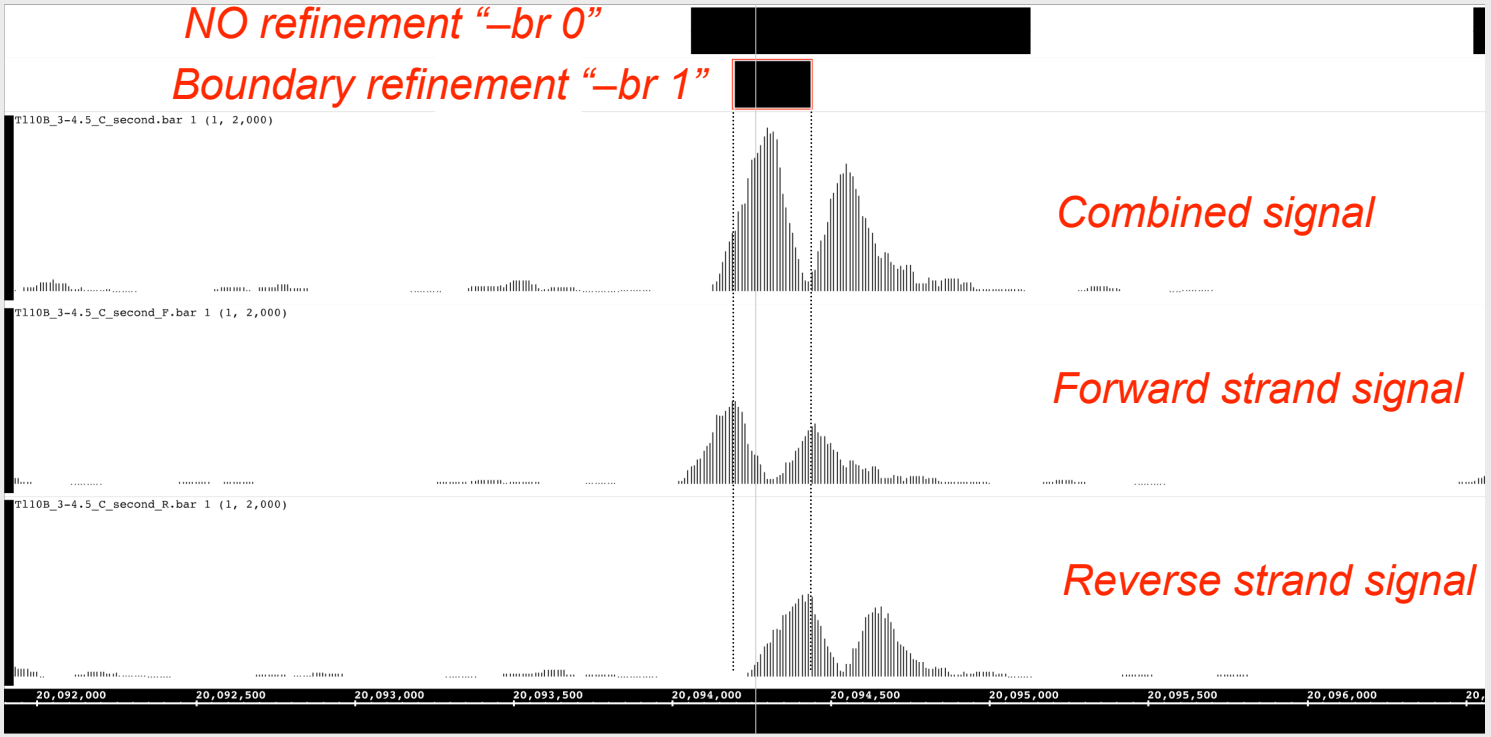


- Check different parameter settings together with positive controls
- Visualize to get a feel

EMBL

Some options might look great…

-br option in cisgenome hts_peakdetectorv2* tool



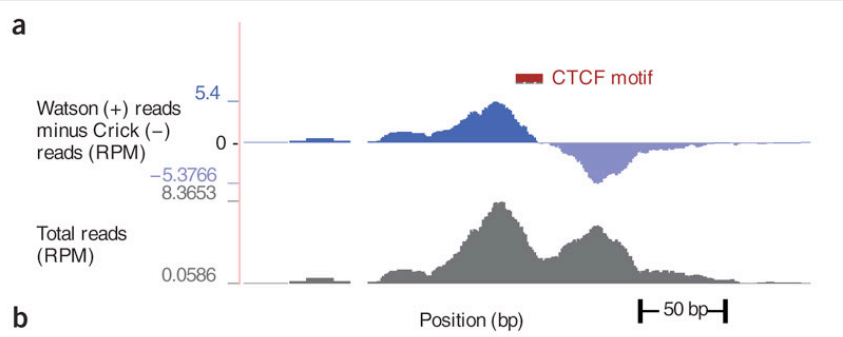
Asking for boundary refinement may cause loss of peaks:
peak finders assumes a single peak is in the region…

EMBL

Peak Splitting

- Window based detection (MACS, CisGenome,…) will report unique regions encompassing several binding sites
- A post processing to split regions into multiple peaks is needed
- PeakSplitter developed by Mali Salmon in EBI
- The new beta version of MACS integrate PeakSplitter
- Tools like SISSR and QuEST implement a different approach (detect summit then extend)

The different types of ChIP-seq signal

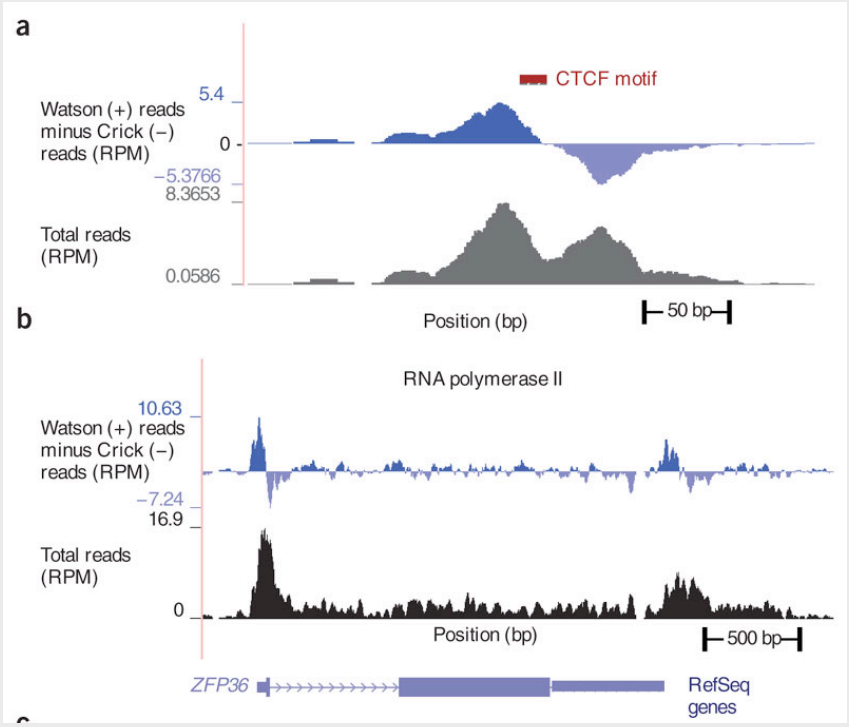


1. Proteins binding DNA in a site-specific fashion

=> Narrow peaks, hundreds of bp wide

Pepke et al ; *Nature Methods* ***6, S22 - S32 (2009)***

The different types of ChIP-seq signal



1. Proteins binding DNA in a site-specific fashion

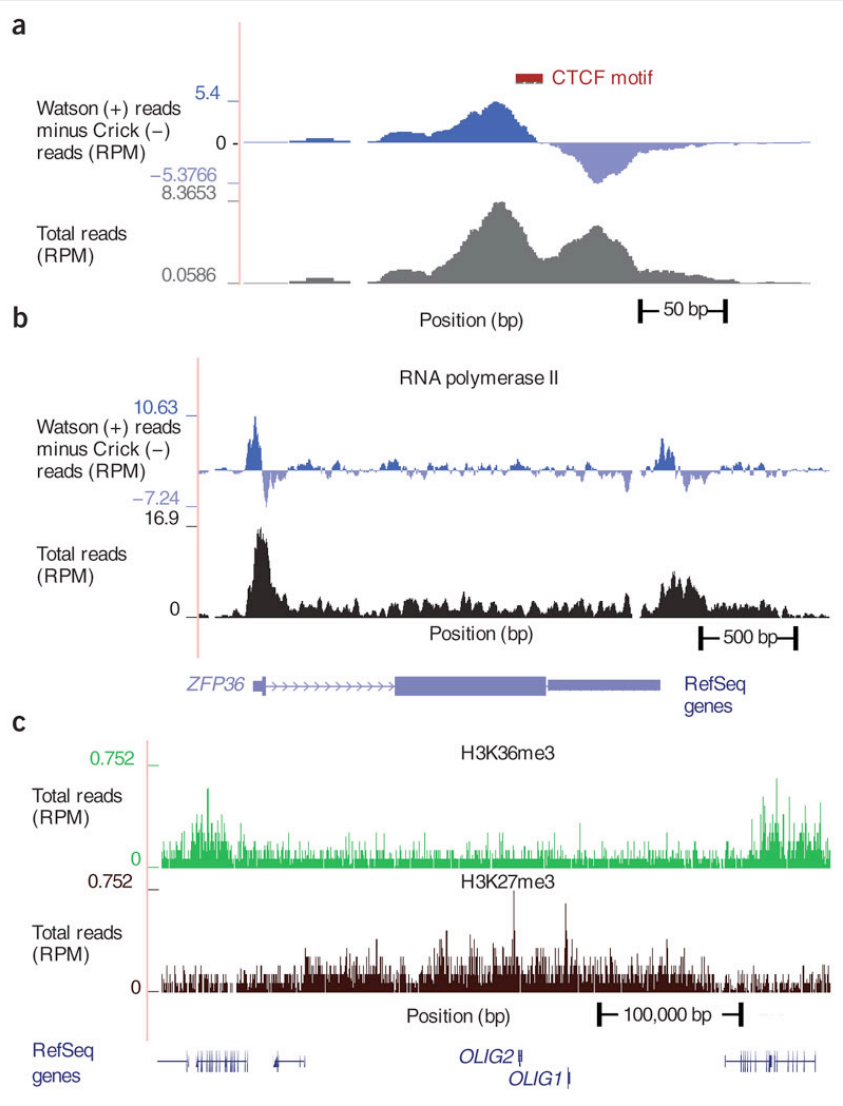
2. RNA Pol II like signal

=> Mixture of strong binding (at TSS) and broader enrichment over several Kb (active transcription)

Pepke et al ; *Nature Methods* ***6, S22 - S32 (2009)***

EMBL

The different types of ChIP-seq signal



1. Proteins binding DNA in a site-specific fashion

2. RNA Pol II signal

3. Chromatin marks

H3K4me3,active promoters

H3K36me3, active genes

H3K27me3, repressed regions

=> Enrichment from nucleosome size domain to several hundreds of Kb

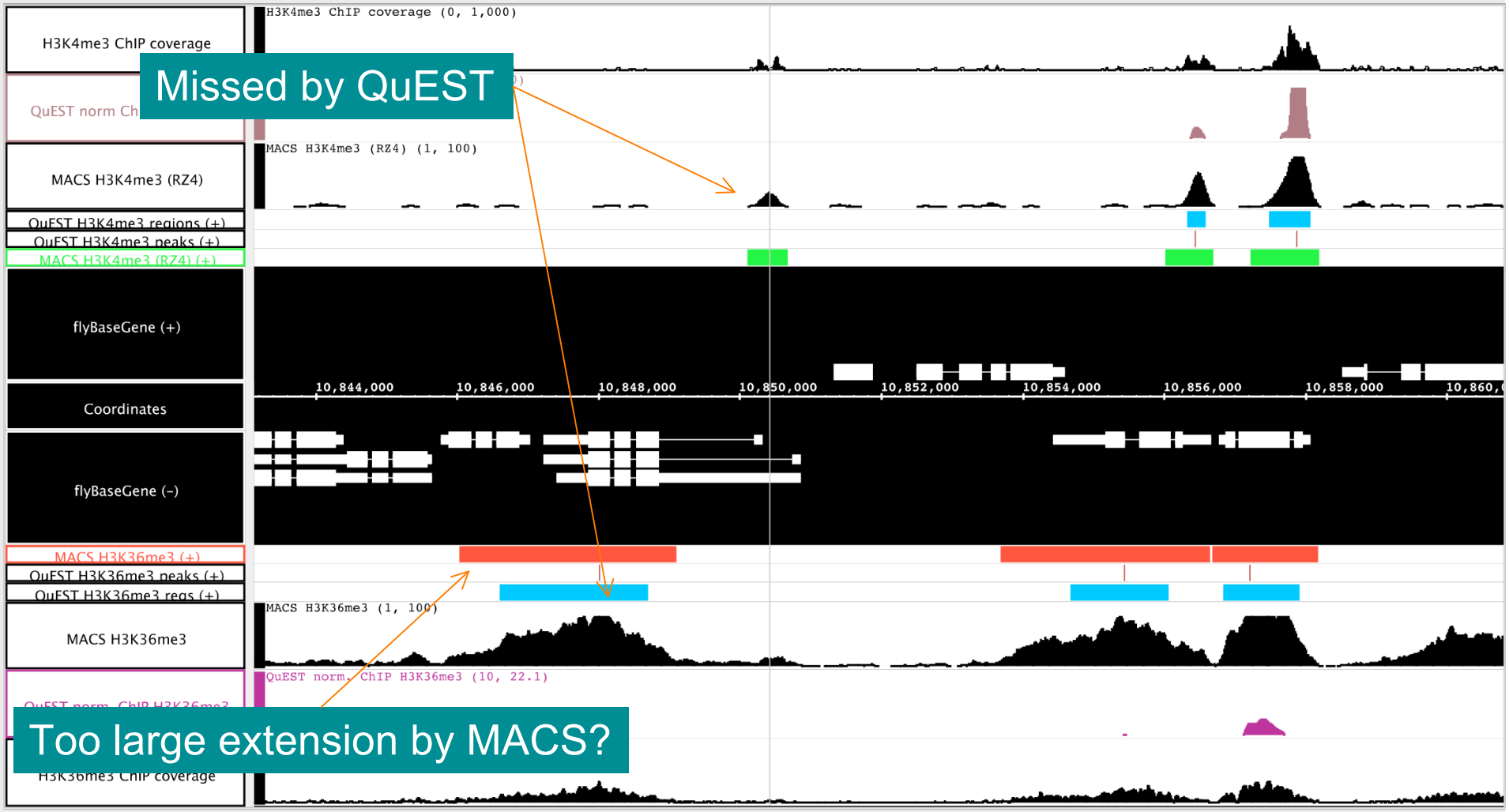
EMBL

Example of Histone marks

- Two marks, at same dev. Stage:
  - H3K4me3 : active promoters (~ short mark)
  - H3K36me3 : active genes (~ long mark)

  => Good test case b/c one should see both marks at active genes

- Analyzed with (in progress):
  - SISSRs failed at finding anything
  - CisGenome also (still investigating the pb)
  - Will show MACS and QuEST results

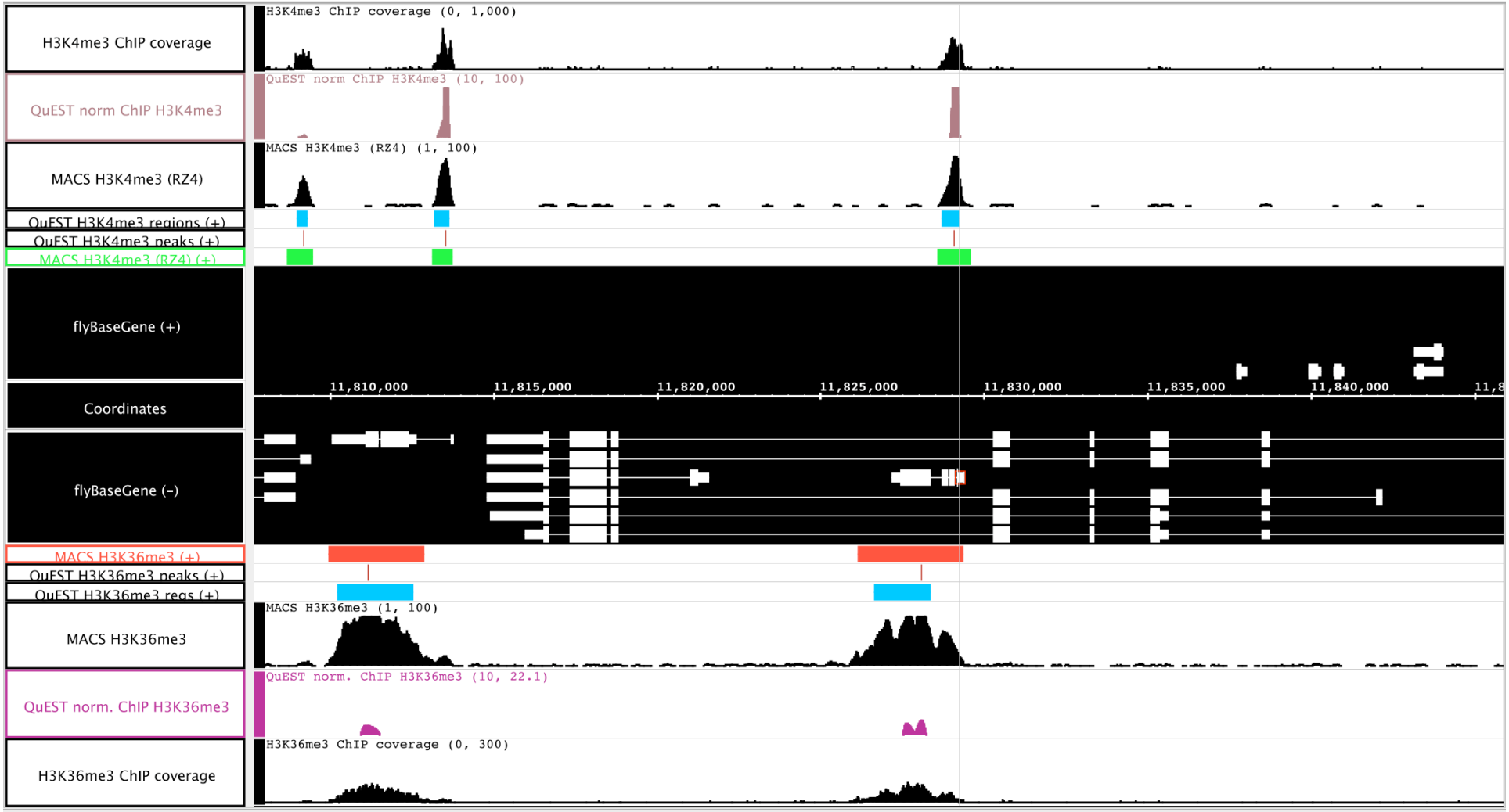


Good agreement between the tools

EMBL

MACS suggests more active promoters and genes: predictions correlate
=> Is QuEST too stringent?

EMBL



Detection of gene within gene example
=> Would you trust this with only one of the two marks?

EMBL

Which one to use?

- You might want to run different tools and check how they behave on your datasets
- Do you have reference sample or not?
- Detection method should be adapted to signal type i.e. SISSR certainly has a too strong peak assumption for (long) histone marks?
- Laajala et al compared results with different peak finders – using TF signal only (*BMC Genomics* 2009, **10:**618)

Visualization is important

- Assessment of the data quality e.g. positive controls, background
- Determine cutoffs (looking at positive controls)
- Compare peak finders outputs
- Integration of data / co-visualization
  - Your brain catches aspects that computers can’t : hypothesis generation.

Thanks !

You

Eileen Furlong

Robert Zinzen / Stefan Bonn

Nicolas Delhomme

Ismael Padioleau

Martina Braun

Furlong Lab

GeneCore