Genome-scale technologies 2/ Algorithmic and statistical aspects of DNA sequencing

DNase I-seq

Ewa Szczurek

University of Warsaw, MIMUW

szczurek@mimuw.edu.pl

Deoxyribonuclease I (DNase I)

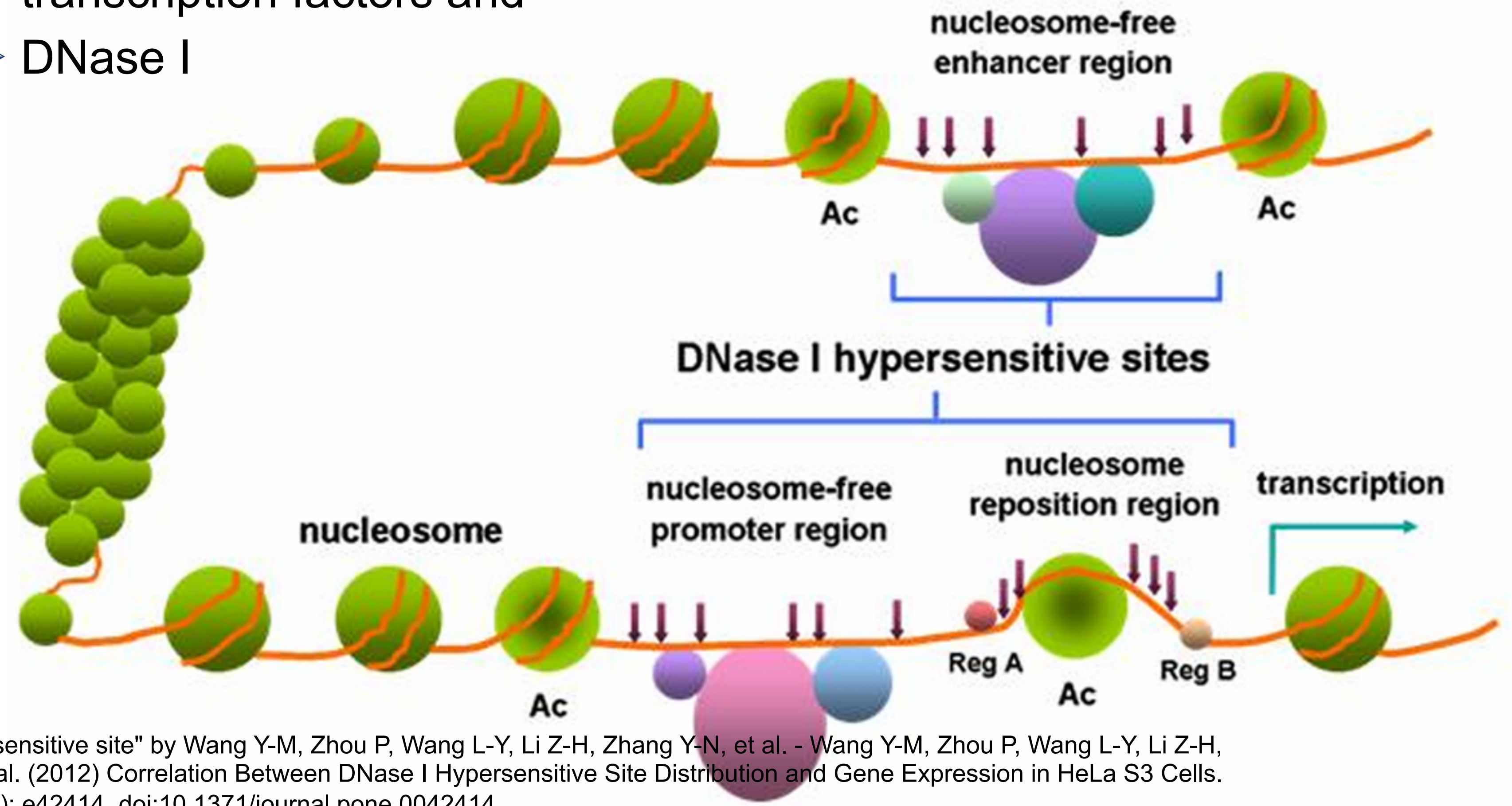
- cleaves DNA adjacent to a pyrimidine nucleotide.
- a waste-management endonuclease
- one of the deoxyribonucleases responsible for DNA fragmentation during apoptosis.
- DNase I *hypersensitive sites* ~
  - open, accessible chromatin;
  - regions of the genome are likely to contain active genes

The project

- http://students.mimuw.edu.pl/~szczurek/TSG2_Project/project.html
- Report deadline: 20.01.2016
- Presentations: 26.01.2016

Deoxyribonuclease I (DNase I) hypersensitive sites

- Short region of chromatin.
- Super sensitivity to Dnase I cleavage
- Nucleosomal structure less compacted
- Increased availability of the DNA to binding by proteins:
  - transcription factors and
  - DNase I



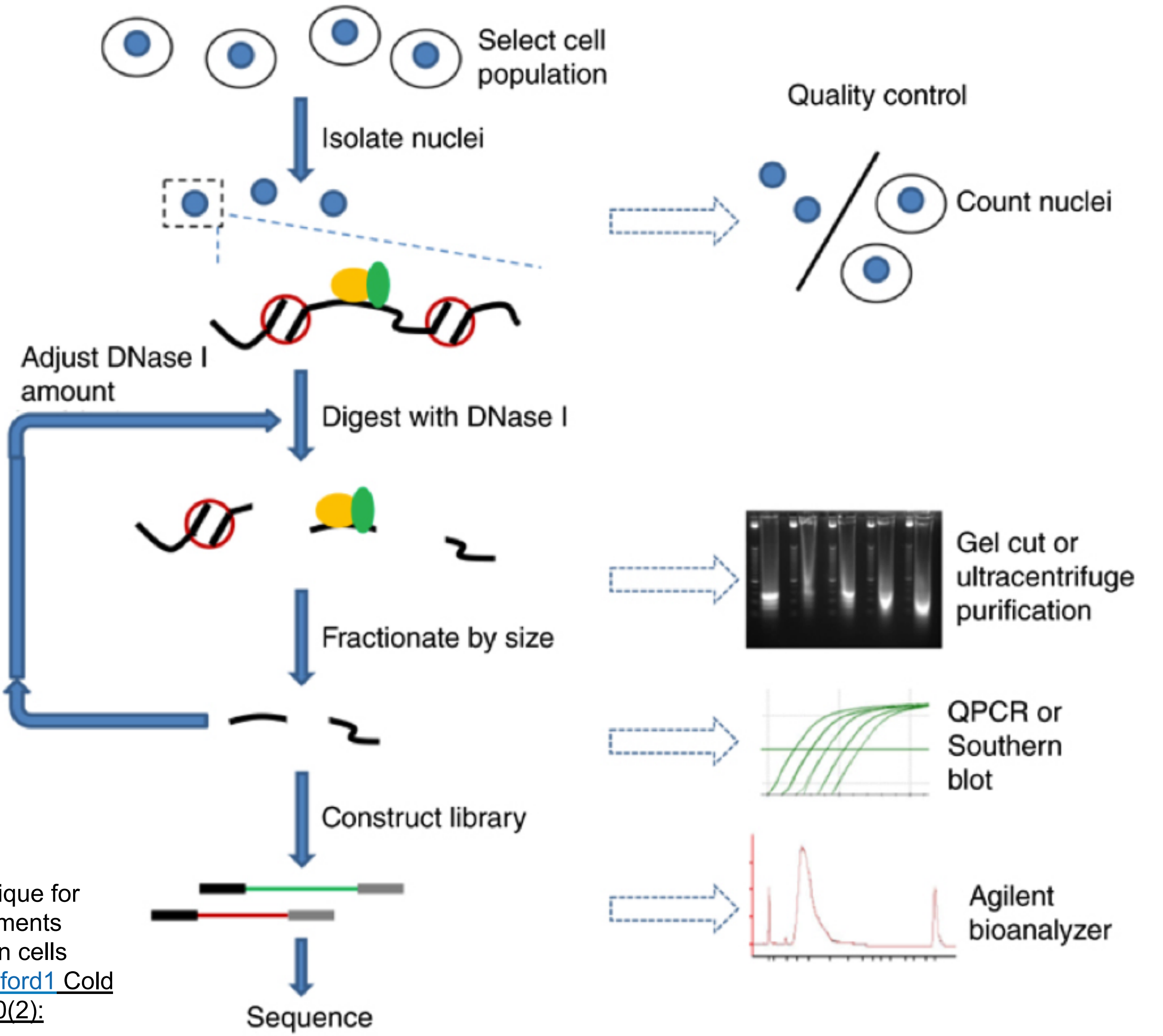
"DNAse hypersensitive site" by Wang Y-M, Zhou P, Wang L-Y, Li Z-H, Zhang Y-N, et al. - Wang Y-M, Zhou P, Wang L-Y, Li Z-H, Zhang Y-N, et al. (2012) Correlation Between DNase I Hypersensitive Site Distribution and Gene Expression in HeLa S3 Cells. PLoS ONE 7(8): e42414. doi:10.1371/journal.pone.0042414.

DNase I hypersensitive sites: location

- Hypersensitive sites (HS) found:
  - On every active gene (often >1 HS per gene)
  - Exclusively on chromatin of cells in which the gene is expressed
  - Before transcription begins, in regions preceding active promoters.

- HS generated as a result of the binding of transcription factors that displace histone octamers.

DNase I- Seq



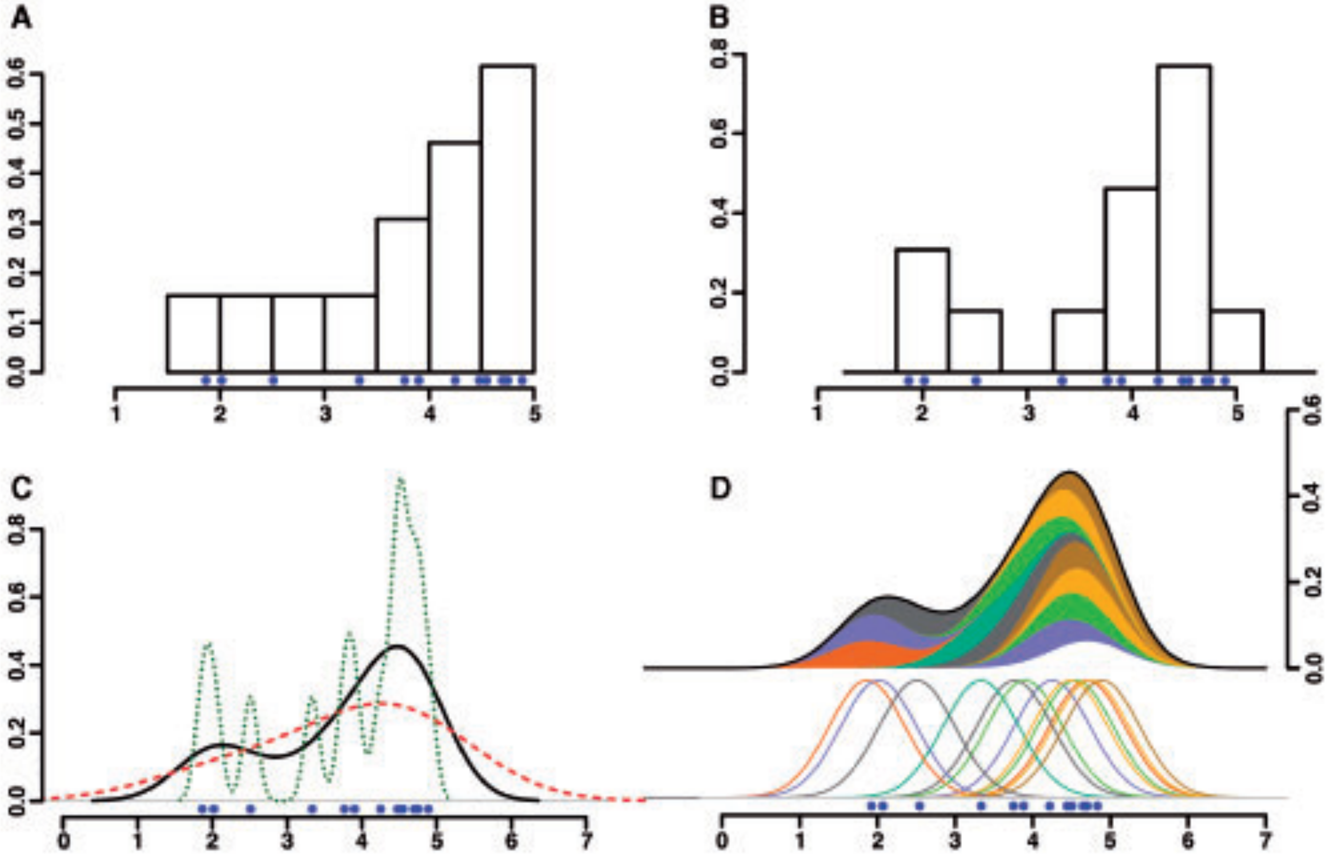
DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells Lingyun Song and Gregory E. Crawford1 Cold Spring Harb Protoc. 2010 Feb; 2010(2): pdb.prot5384.

Dnase I peak calling

- Peaks:
  - Within HS
  - drop of cleavage relative to surrounding

F-seq

- Aim: visually display and summarize tag data in an intuitive way
- generates a continuous tag sequence density estimation
- allowing identification of biologically meaningful sites
- output can be displayed directly in the UCSC Genome Browser.

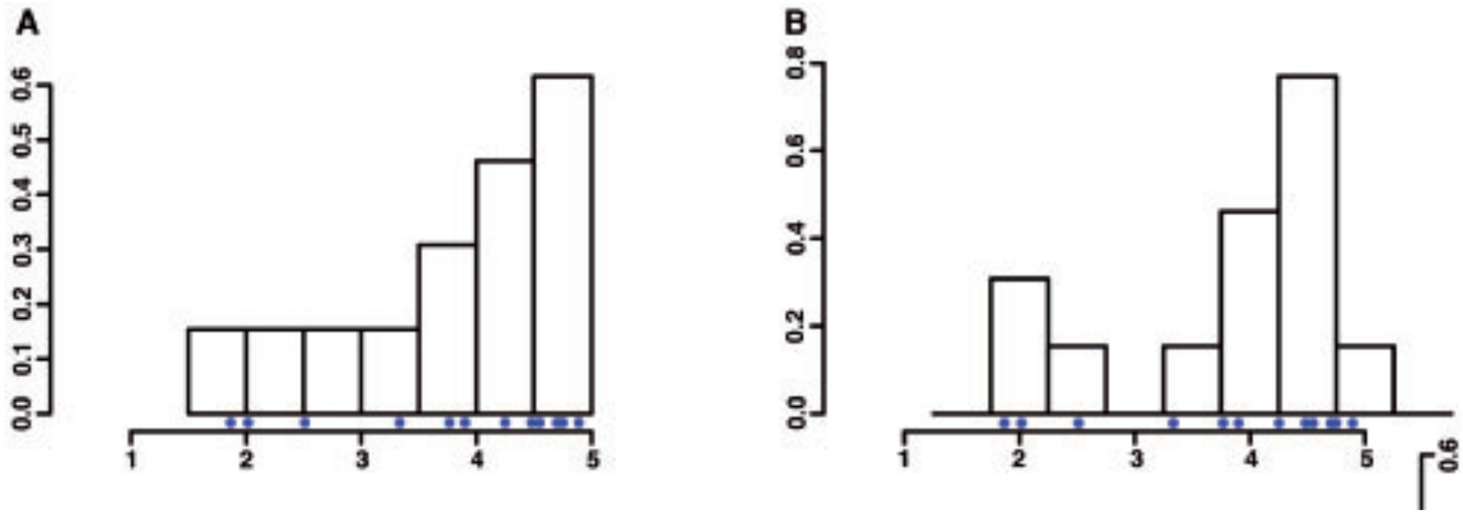


Boyle et al. (2008)

Histogram

- Introduced by Karl Pearson
- Bin (divide) the range of values into
  - consecutive
  - Adjacent
  - (Equal size)
  - non overlapping intervals
- Count how many values end up in each bin

Histograms can be fooled by sparse sequencing data



- Blue dots: sample positions
- Locations of the histogram bins can cause data to look
  - ➢ unimodal (A) or
  - ➢ bimodal (B)
  - ➢ depending on starting positions (here 1.5 or 1.75)

Kernel density estimation

- A non-parametric way to estimate the probability density function of a random variable
- Inference about a population from a sample
- Let *(x_1, …, x_n)* iid samples from a distribution with density *f*
- *Kernel density estimator:*

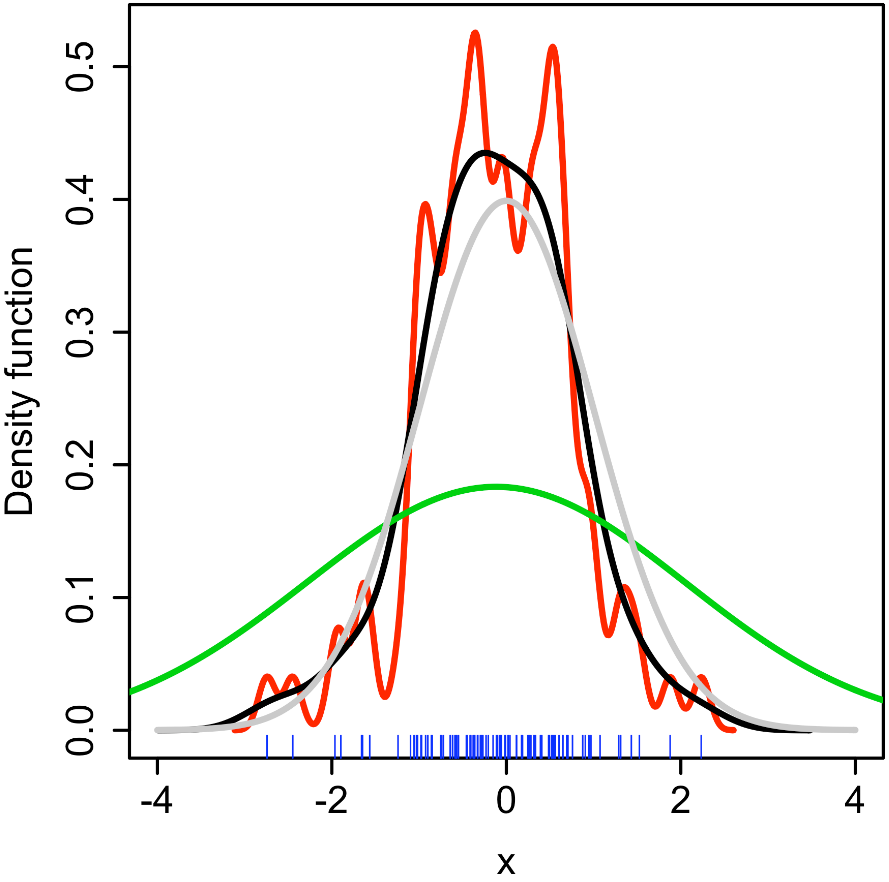
$$\hat{\mathrm{f}}_{\mathrm{h}}(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right),$$

$$K_h(x) = \frac{1}{h} K\left(\frac{x}{h}\right)$$

- *K(•)* - the kernel, a non-negative function that integrates to one and has mean zero
- Popular K(x) = standard normal
- *h* > 0 - a smoothing parameter called the bandwidth.

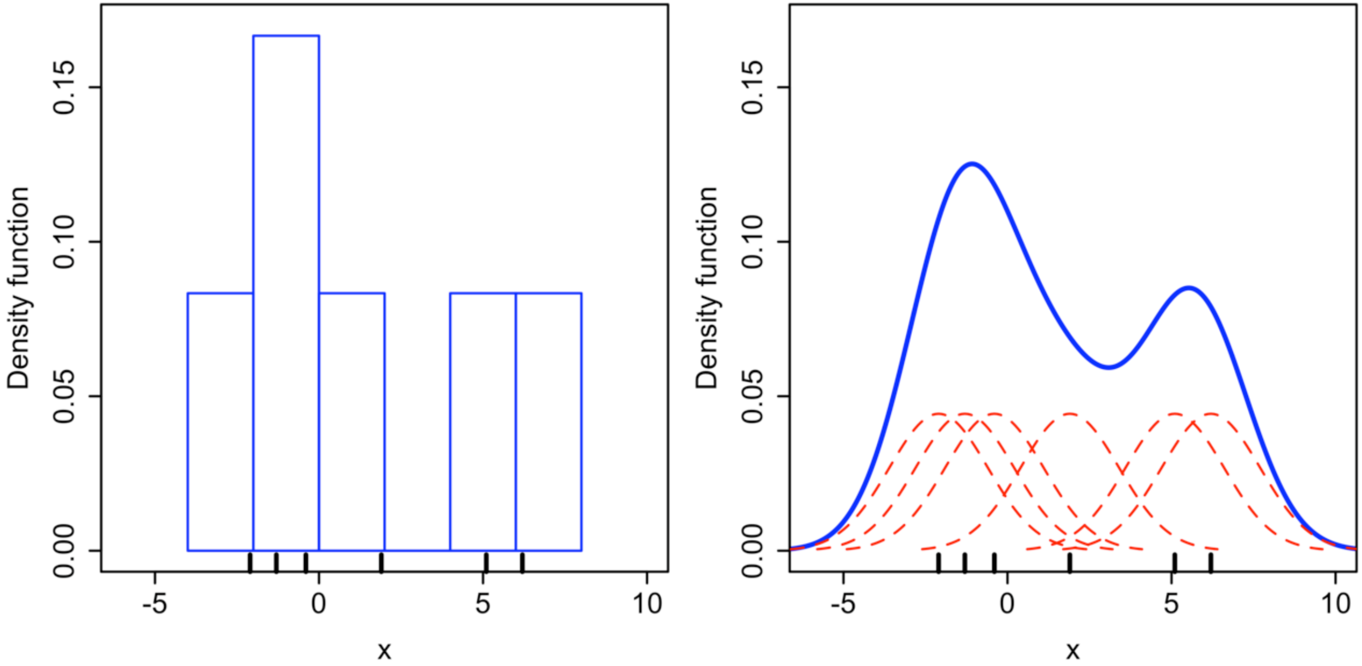
Bandwidth selection

- A random sample of 100 points from a standard normal distribution.
- Grey: true density (standard normal).
- Red: KDE with h=0.05 *undersmoothed*.
- Black: KDE with h=0.337 optimal.
- Green: KDE with h=2 *oversmoothed*.
- Bandwidths chosen to minimize the mean integrated squared err.



Kernels vs histograms

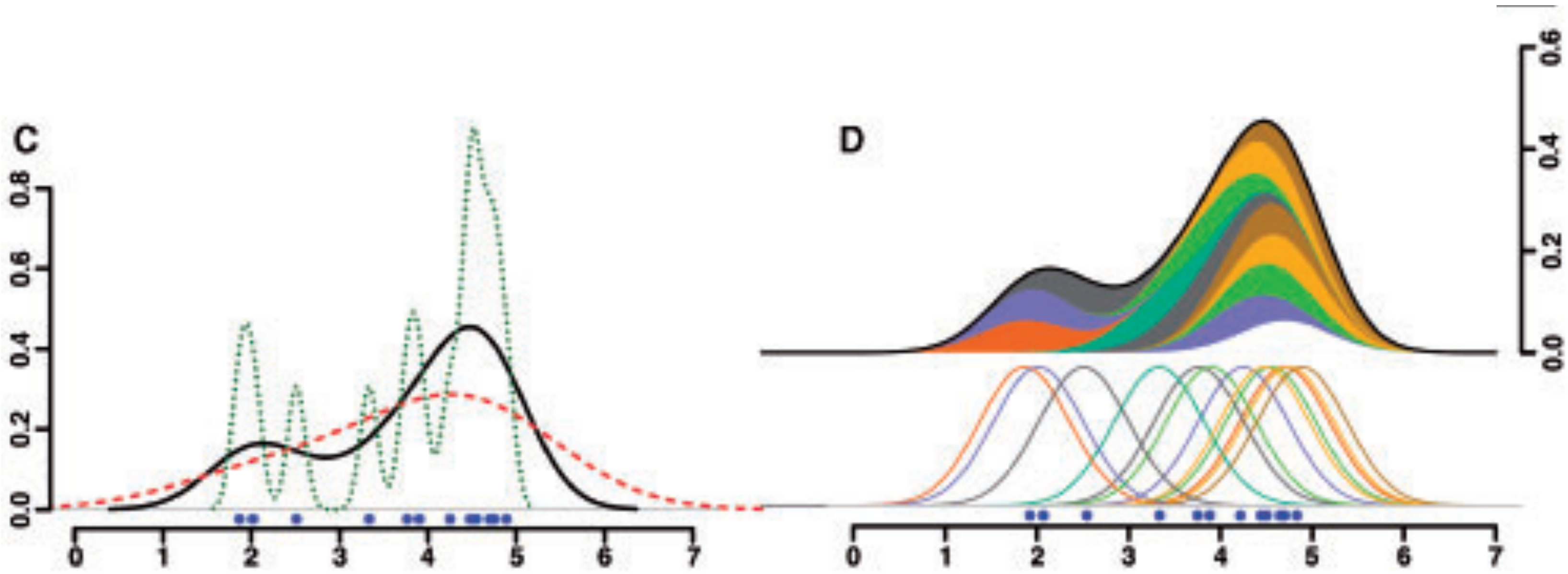
- 6 samples: $x_1 = -2.1$, $x_2 = -1.3$, $x_3 = -0.4$, $x_4 = 1.9$, $x_5 = 5.1$, $x_6 = 6.2$.
- Histogram:
  - ➢ 6 bins width 2
  - ➢ For each data point in a bin, but a box of height 1/12
- Kernel estimate:
  - ➢ For each data point put a normal kernel with var =2.25
  - ➢ Sum the kernels



"Comparison of 1D histogram and KDE" by Drleft (talk) Licensed under CC BY-SA 3.0 via Wikipedia

Bandwidth affects the density estimaiton

- (B) Over and undersmoothing
- (D) Example of how distributions over each point are combined to create the final distribution.
- Each of the samples are represented by Gaussian distributions which are summed to create the final density estimation



F-seq

- n sample points, over chromosome length L
- Gaussian standard kernel estimator with bandwidth b

$$\hat{\rho}(x) = \frac{1}{nb}\sum_{i=1}^{n} K\left(\frac{x - x_i}{b}\right)$$

- User provides feature length (default 600), the larger the smoother
- Use a sliding window *w* to avoid comp. precision problems such that

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{w}{b}\right)^2} > \min(\text{floating point}).$$

F-seq

- Compute a significance threshold, with parameters *k* and *s*

1. Compute an average number of features for window *w* as $n_w = nw/L$.
2. Calculate the kernel density (kd) at a fixed point x_c within w, assuming a random uniform distribution of the n_w features.
3. Repeat (2) *k* times to obtain a distribution of the kd estimates for x_c. For large *k* the kd-es become normally distributed.
4. The threshold is *s* SDs above the mean of this normal distribution.

F-seq

- Input: BED file
- → determine point representatives of aligned sequences
- → Output:
  - a continuous probability wiggle format (http://genome.ucsc.edu/goldenPath/help/wiggle.html) or
  - Discrete-scored regions BED format: where the continuous probability is above the threshold *s* SDs above the background mean.
- → Import into the UCSC Genome Browser (Kent et al., 2002) (http://genome.ucsc.edu).

F-seq on ChIP seq

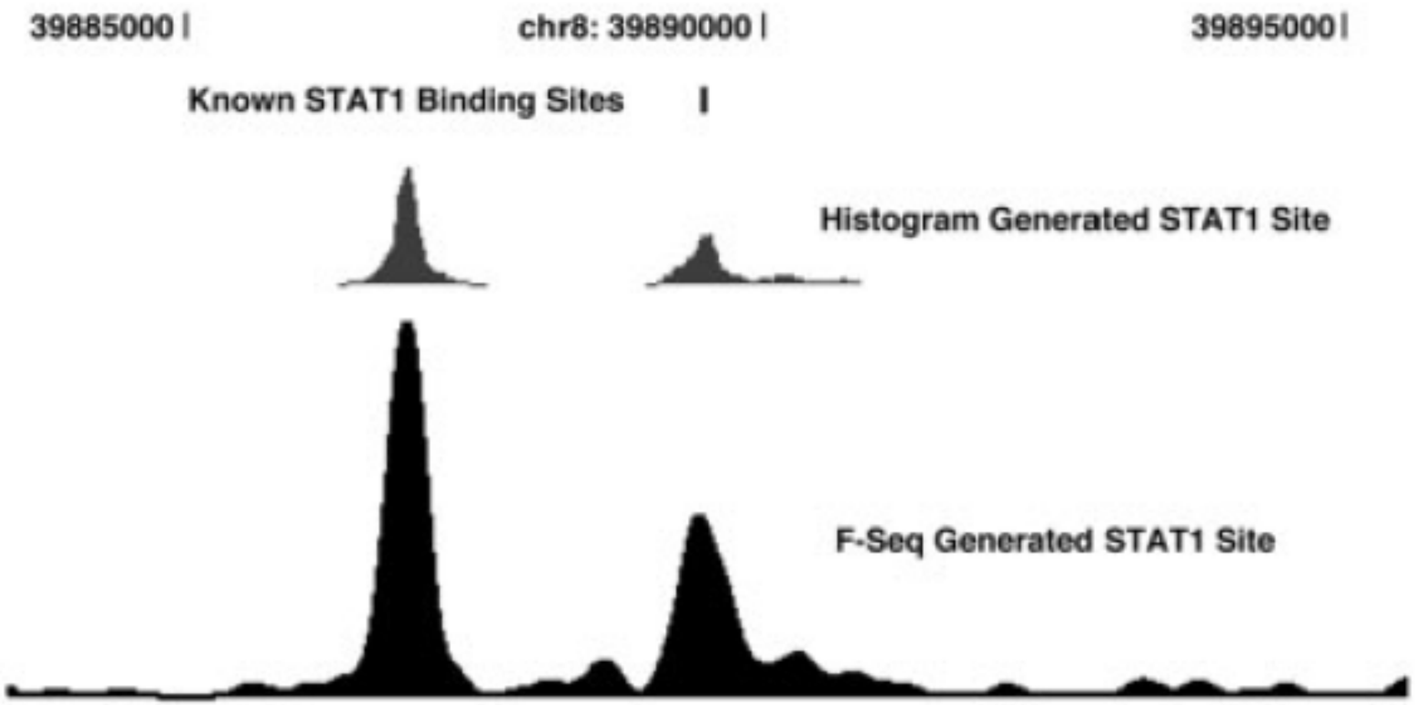
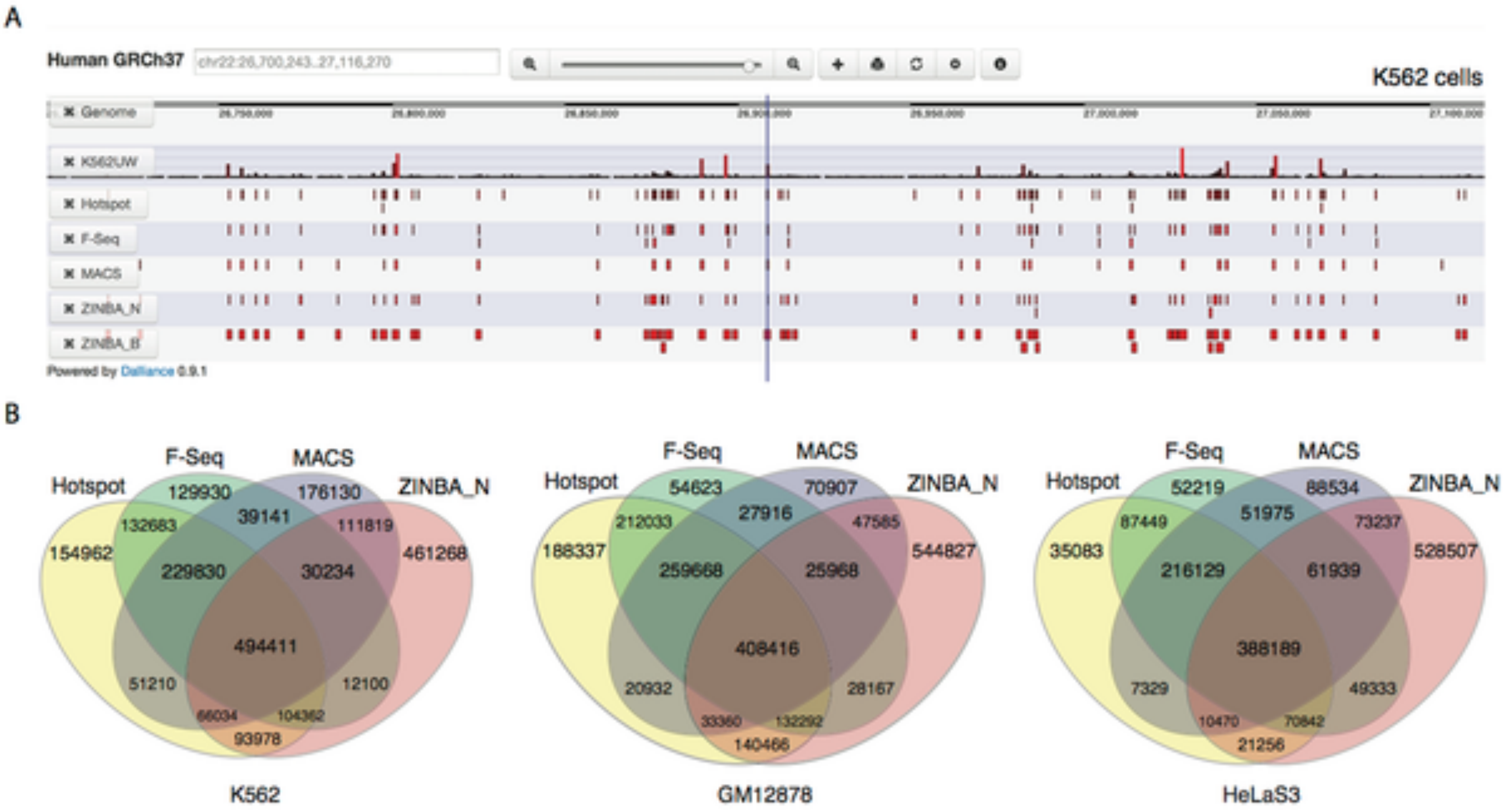


Fig. 2. View of 10 kb region of Chromosome 8 shows an accurate duplication of windowing technique in STAT1 data (Robertson *et al*., 2007). Note that the histogram generated sites from Robertson *et al*. only display sites above a cutoff.

Comparison of DNase I-seq peak callers



Koohy H, Down TA, Spivakov M, Hubbard T (2014) A Comparison of Peak Callers Used for DNase-Seq Data. PLoS ONE 9(5): e96303. doi:10.1371/journal.pone.0096303
http://journals.plos.org/plosone/article?id=info:doi/10.1371/journal.pone.0096303

DNase footprinting assay

- DNA footprinting: investigating the sequence specificity of DNA-binding proteins *in vitro*
- Elucidating gene regulation: binding of regulatory proteins to enhancers, promoters.
- DNase footprinting assay:
  - DNA footprinting technique
  - **Using the fact that a protein bound to DNA will often protect that DNA from enzymatic cleavage**.
  - **Locates protein binding sites**
  - DNase cuts the radioactively end-labeled DNA
  - Gel electrophoresis used to detect the resulting cleavage pattern.

Brenowitz M, Senear DF, Shea MA, Ackers GK (1986). "Quantitative DNase footprint titration: a method for studying protein-DNA interactions". *Methods in Enzymology* **130**: 132–81. doi:10.1016/0076-6879(86)30011-9. PMID 3773731.

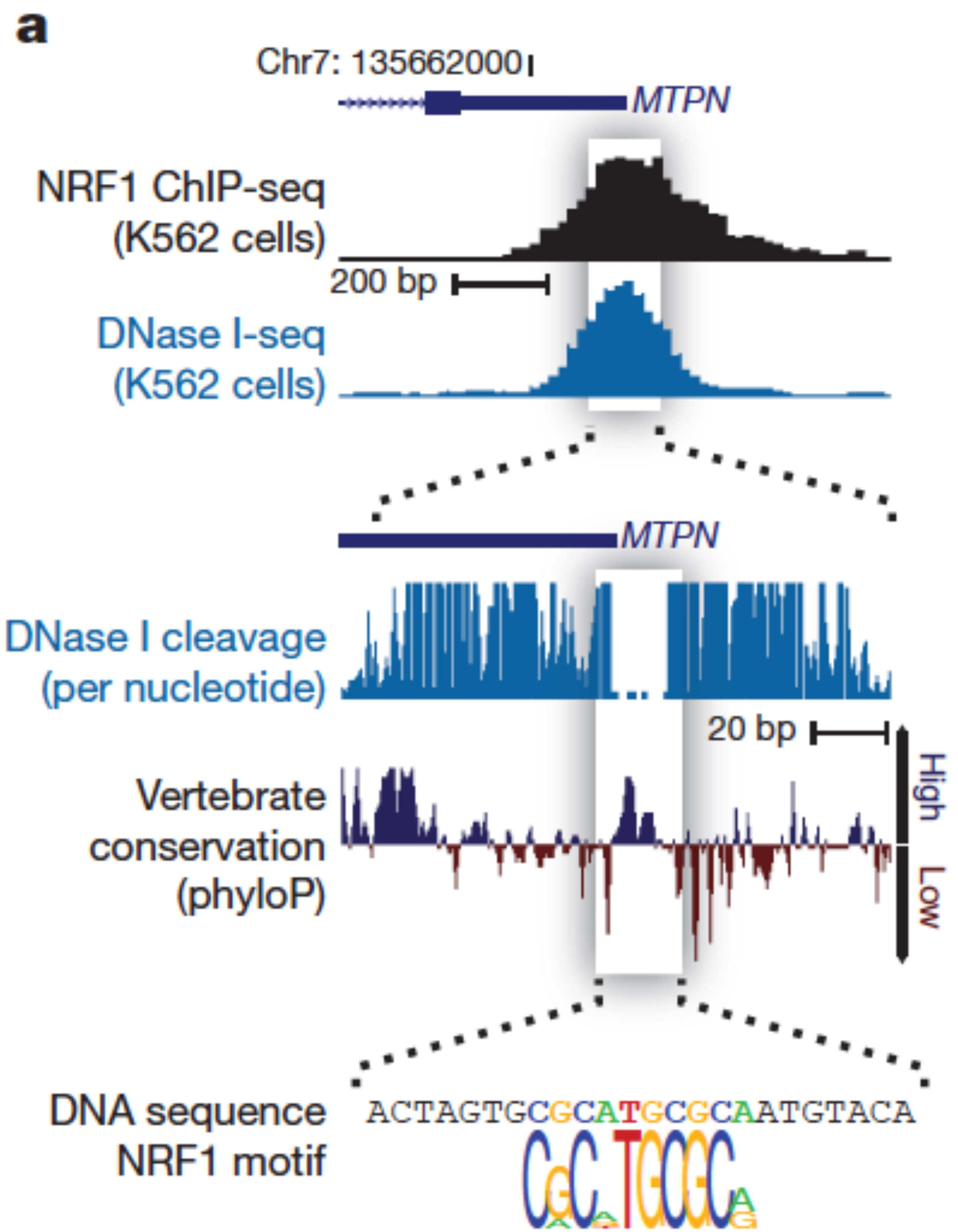
Galas DJ, Schmitz A (Sep 1978). "DNAse footprinting: a simple method for the detection of protein-DNA binding specificity". *Nucleic Acids Research* **5** (9): 3157–70. doi: 10.1093/nar/5.9.3157. PMC 342238. PMID 212715.

DNase I HS footprinting

- Regulatory factor binding to DNA
- → depletion of canonical nucleosomes
- → markedly increased accessibility of the DNA template around the factor binding regions
- This accessibility is manifest as DNase I hypersensitive sites
- Within hypersensitive sites, cleavages accumulate at nucleotides that are *not* protected by protein binding.
- Binding sites detectable provided sufficiently dense local sampling of DNase I cleavage sites.

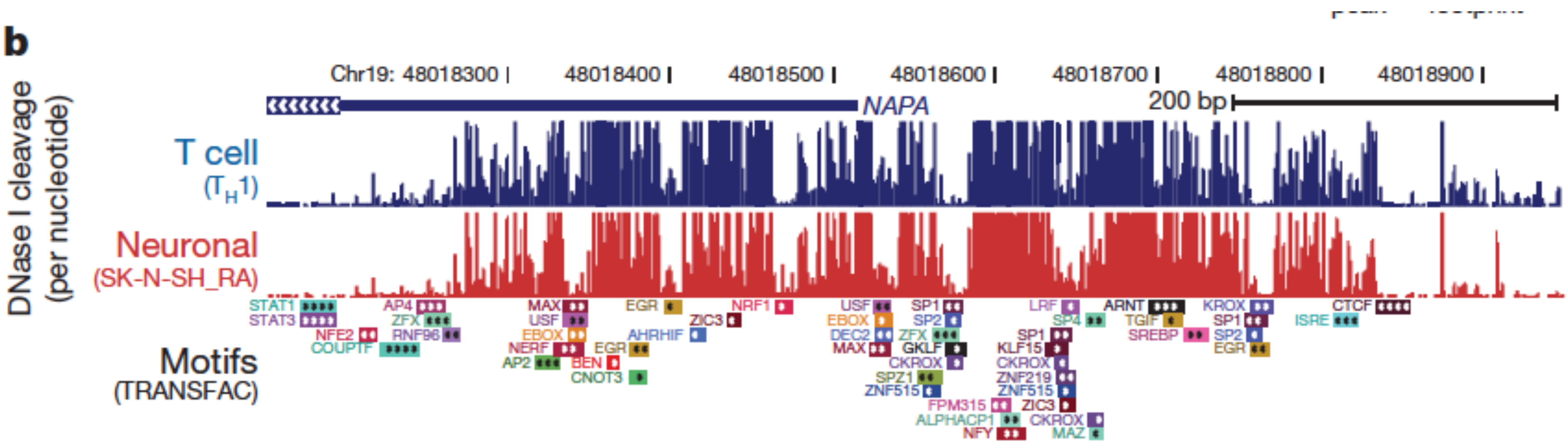
- → DNase I leaves footprints that demarcate transcription factor occupancy at nucleotide resolution

DNase I footprinting



Neph et al., Nature, 2012

Footprints are quantitative markers of factor occupancy

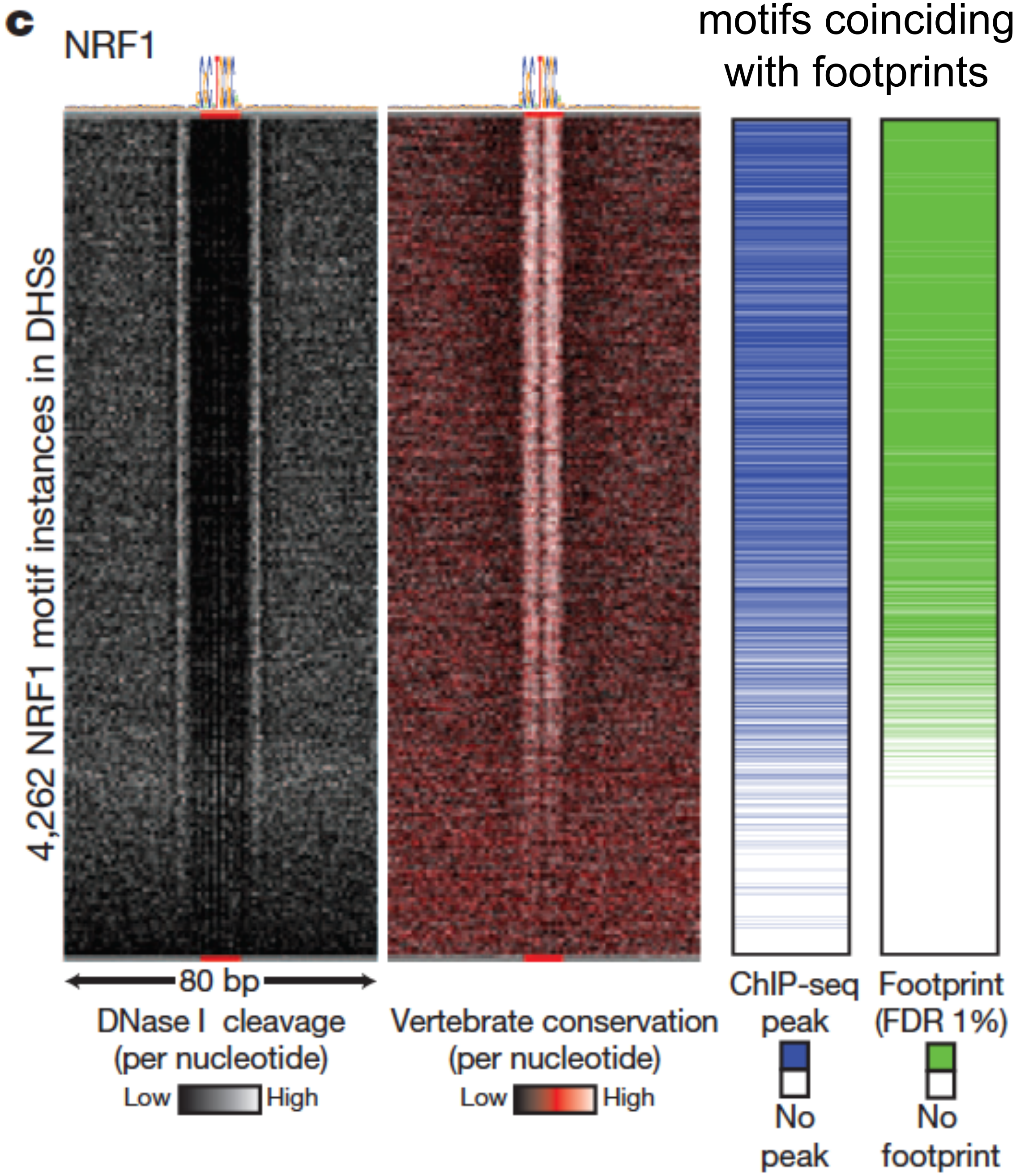


- DNase I cleavage patterns surrounding all 4,262 NRF1 motifs contained within DHSs

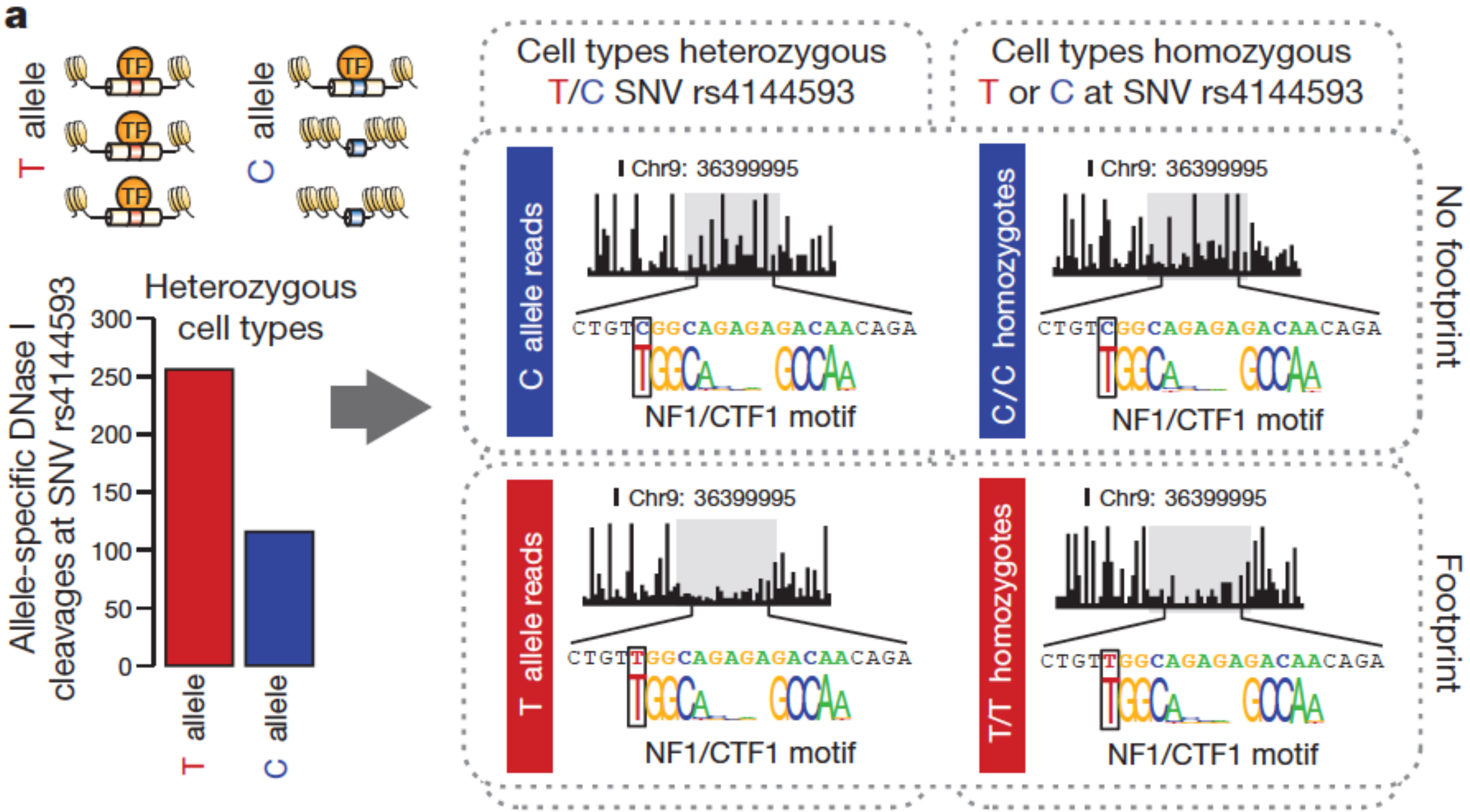
- Ranked by footprint occupancy score (FOS): relating the density of DNase I cleavages within the motif to the flanking regions

- FOS:
  - sequence-specific regulatory factor occupancy
  - evolutionary constraint
  - ChIP-seq signal intensity

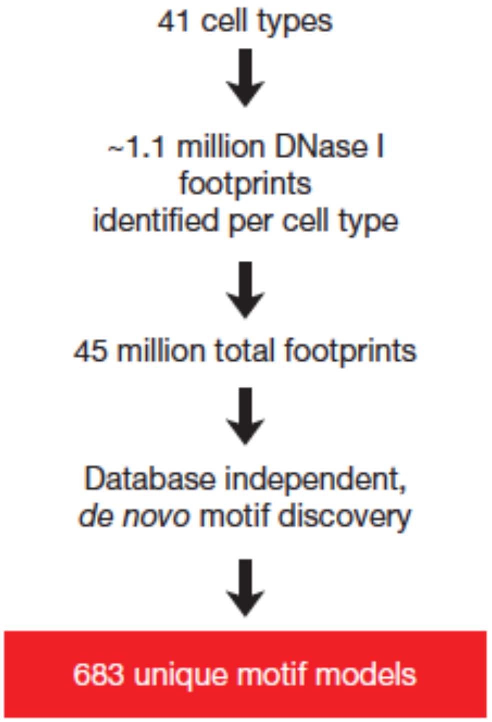
2,351 of 4,262 motifs coinciding with footprints



Footprints harbour functional SNVs



De novo motif finding



41 cell types

~1.1 million DNase I footprints identified per cell type

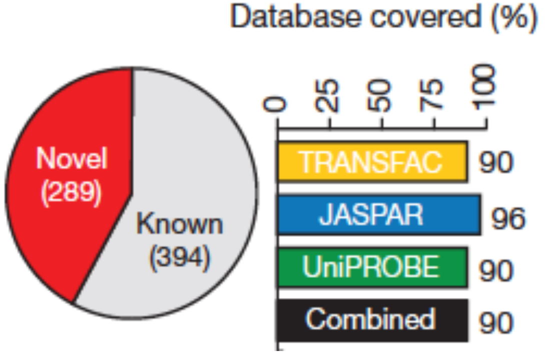
45 million total footprints

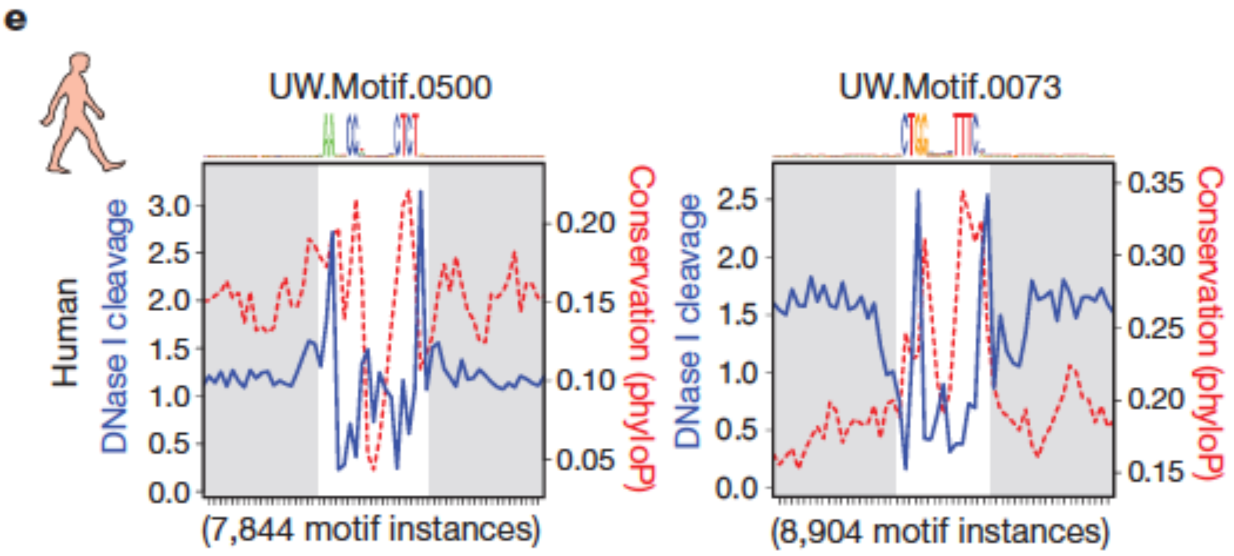
Database independent, *de novo* motif discovery

683 unique motif models

b

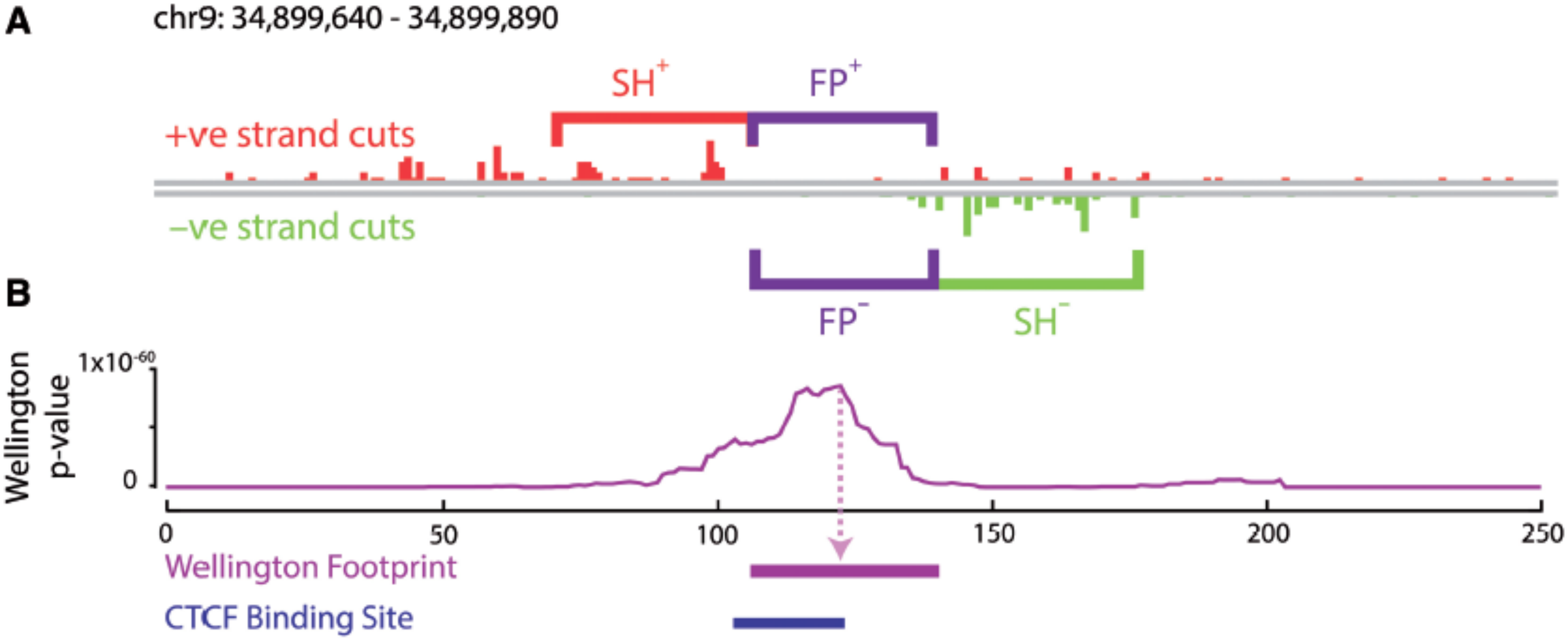
Annotation of 683 *de novo* motif models





The Wellington algorithm

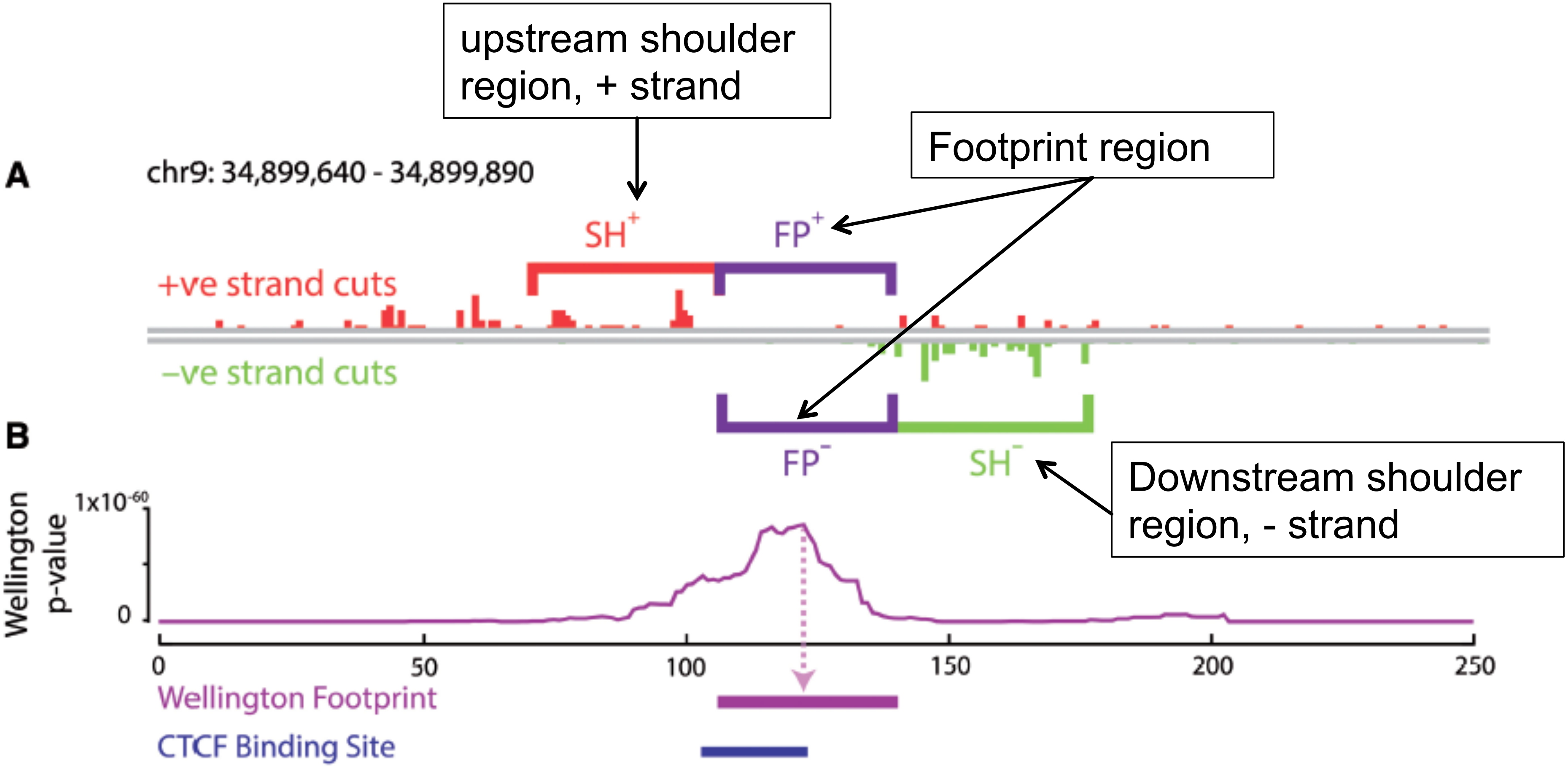
- Detects Protein–DNA binding sites as
  - Short sites within DNase I HS
  - with depletion of cuts
  - compared with a large number of cuts in the surrounding region



Piper et al. (2013)

The Wellington algorithm

- Detects Protein–DNA binding sites as
  - Short sites within DNase I HS
  - with depletion of cuts
  - compared with a large number of cuts in the surrounding region



upstream shoulder region, + strand

Footprint region

Downstream shoulder region, - strand

Piper et al. (2013)

The Wellington algorithm

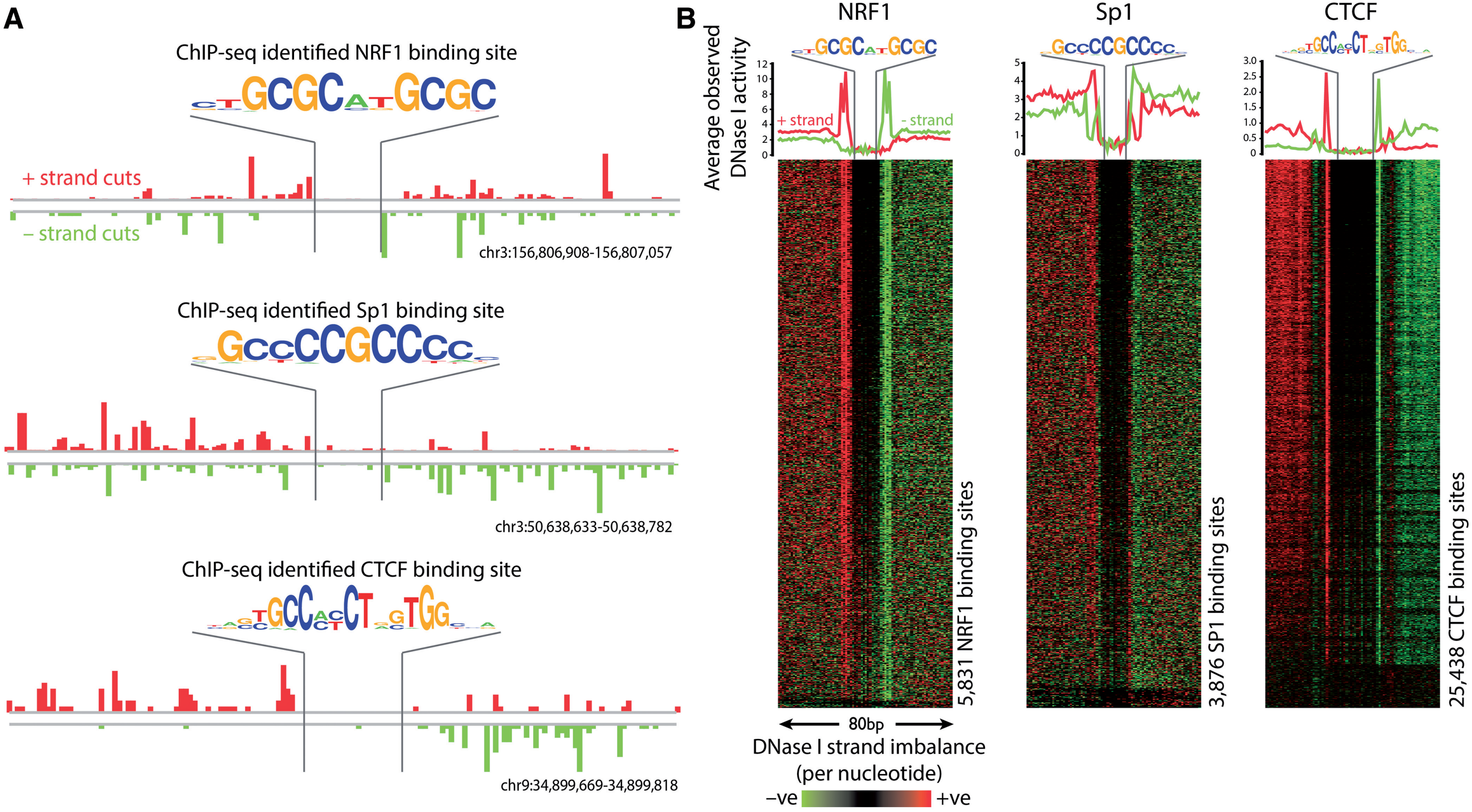
- FP^+ : # cuts on the forward reference strand inside the possible footprint
- SH^+: in the upstream shoulder region on the forward reference strand
- FP^+ : on the backward reference strand inside the possible footprint
- SH^+: in the downstream shoulder region on the backward strand
- l_{FP} : the length (in base pairs) of the possible footprint
- l_{SH}: the length (in base pairs) of the shoulder region

- Test each strand separately
- Binomial test: null hypothesis is that the number of reads is proportional to the region length:
  - Let F[k, n, p]: the binomial cumulative distribution function (the probability of achieving at least k out of n successes with the probability of each success being p)

$$\text{P-value}=\{1 - F[FP^+, FP^+ + SH^+,\ l_{FP}/(l_{FP}+l_{SH})]\ \}\{1-F[FP^-, FP^- + SH^-, l_{FP}/(l_{FP}+l_{SH})]\}$$

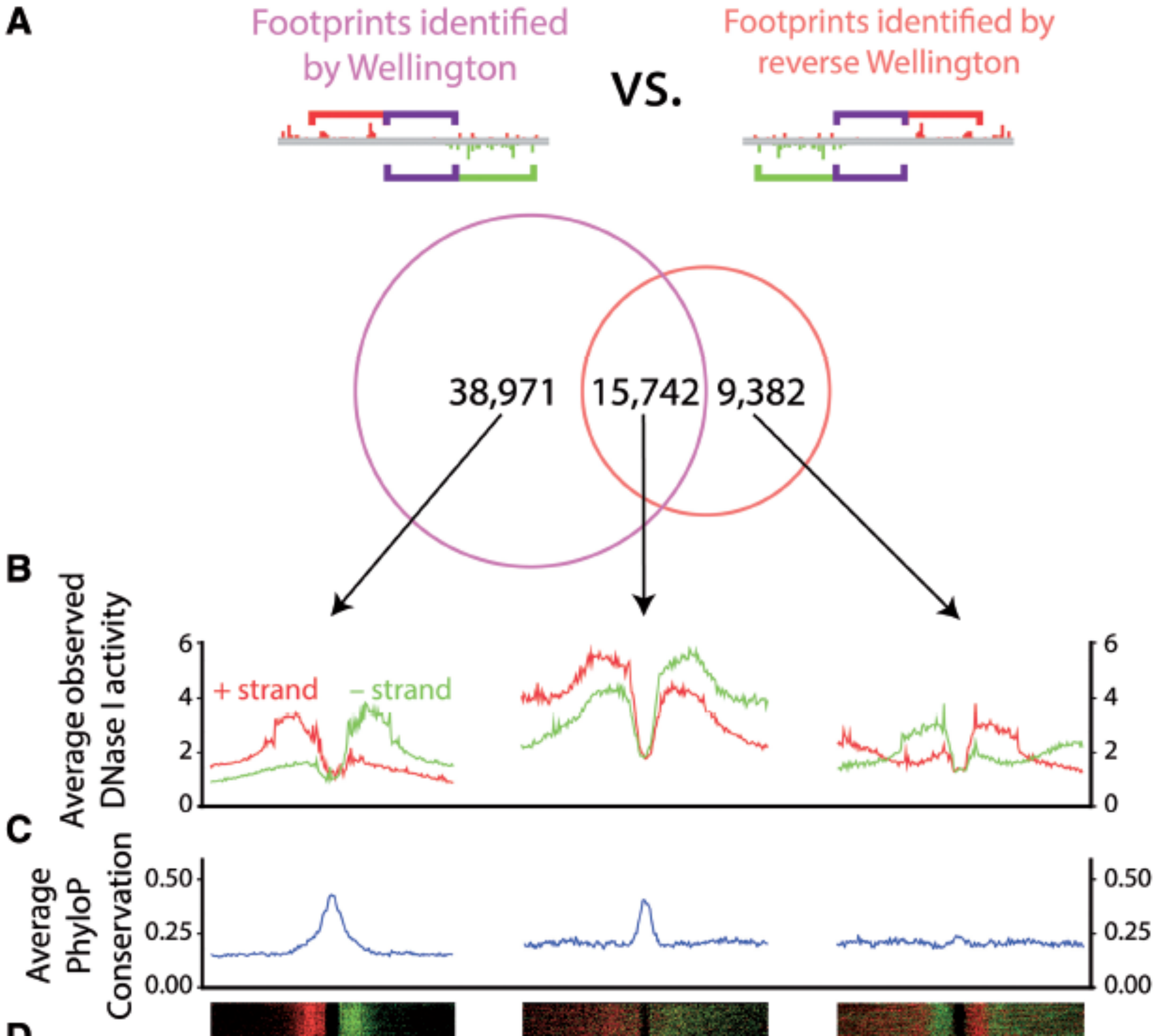
Strand imbalance improves TF binding localization

- Large numbers of sequencing fragments align to
  - the + strand upstream of the protein–DNA binding site and
  - the - strand downstream of the protein–DNA binding site

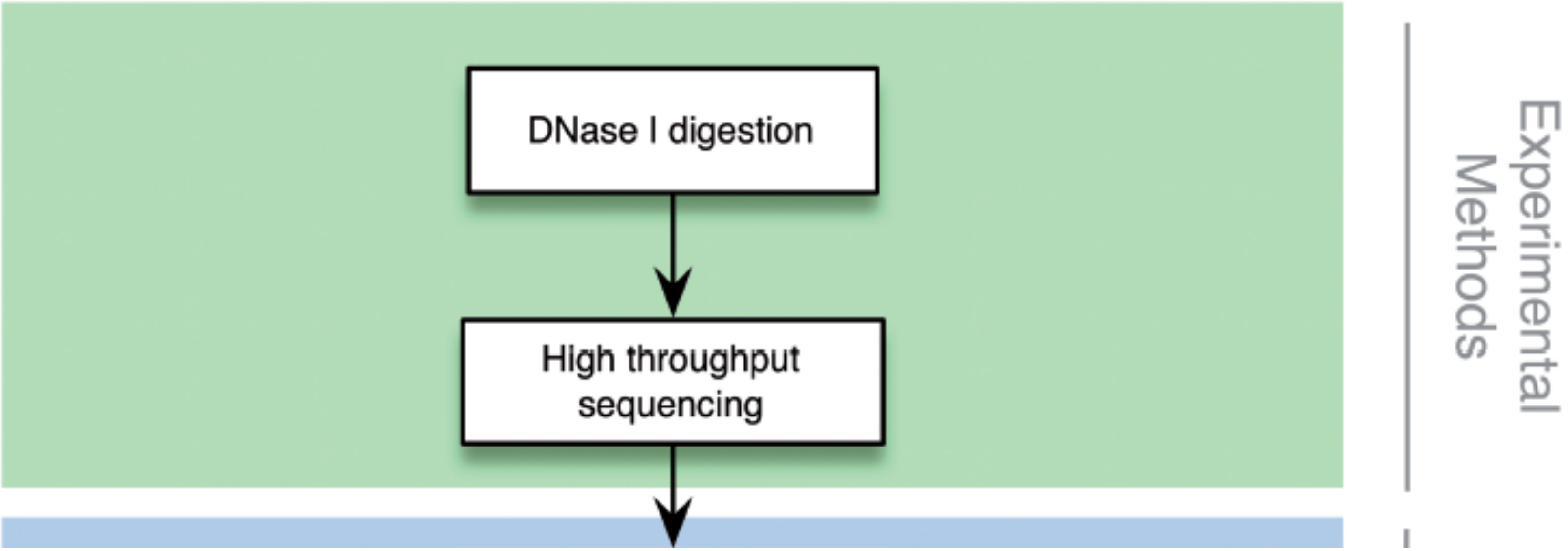


Strand imbalance improves TF binding localization

- Repeated using reversed imbalance (testing FP^+ vs SH downstream on the + strand, and FP^- vs SH upstream on the -strand)
- Lower evolutionary conservation

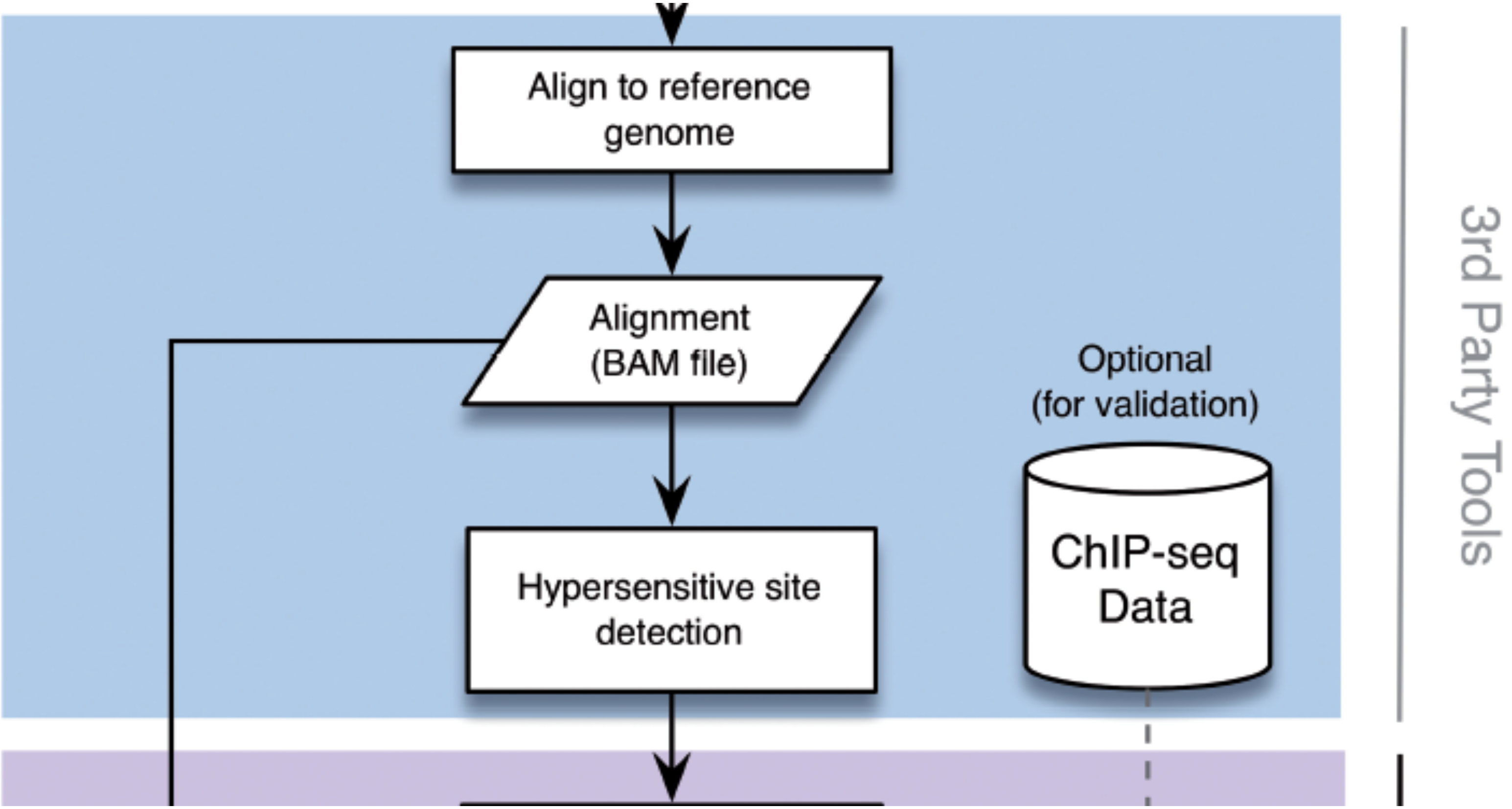


PyDNAse

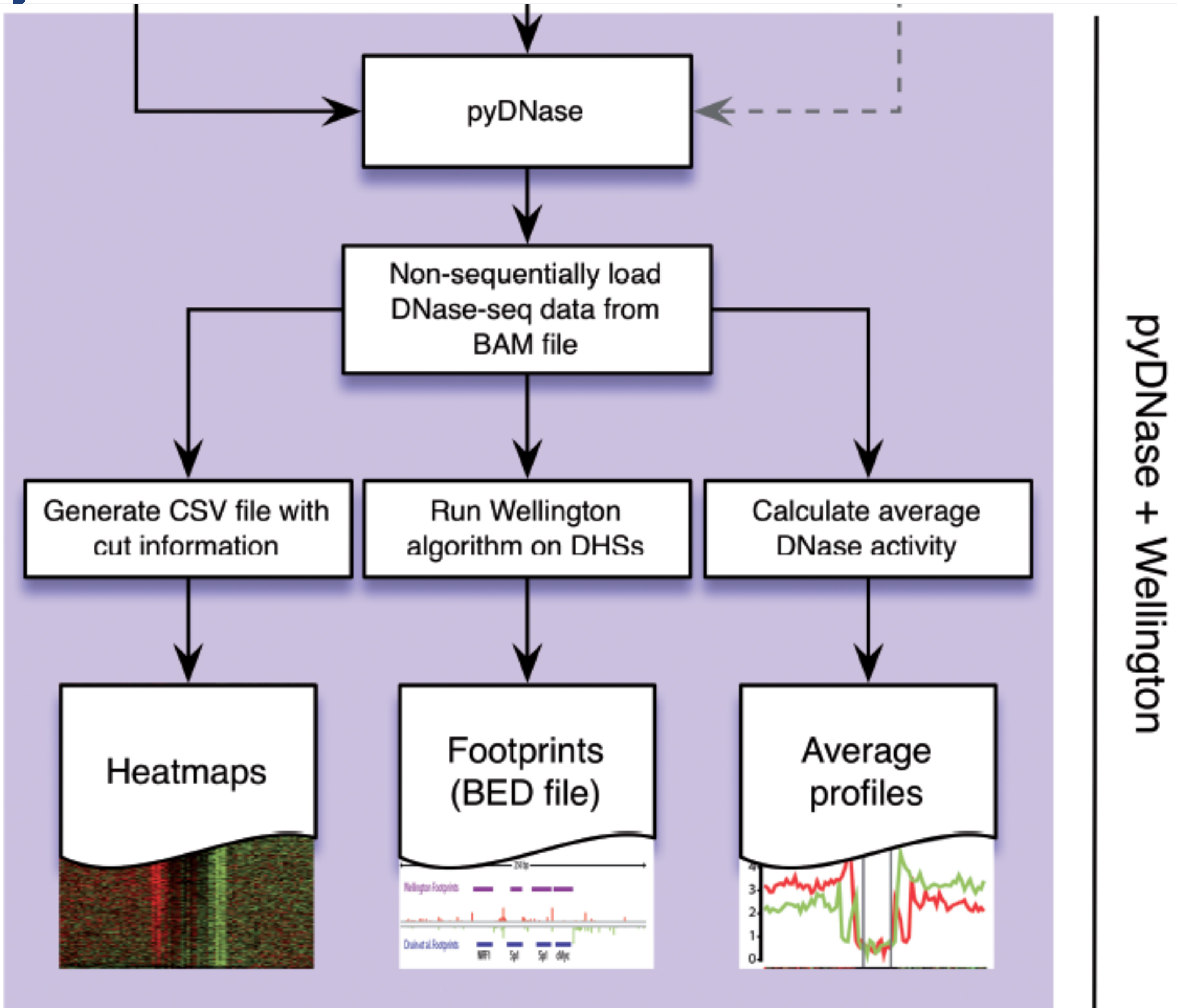


DNase I digestion
High throughput sequencing
Experimental Methods

PyDNAse


Align to reference genome
Alignment (BAM file)
Optional (for validation)
ChIP-seq Data
Hypersensitive site detection
3rd Party Tools

PyDNAse



pyDNase

Non-sequentially load DNase-seq data from BAM file

Generate CSV file with cut information

Run Wellington algorithm on DHSs

Calculate average DNase activity

Heatmaps

Footprints (BED file)

Average profiles

pyDNase + Wellington

Bibliography

- Hesselberth et al. *Global mapping of protein-DNA interactions in vivo by digital genomic footprinting.* Nat Methods. 2009 April ; 6(4): 283–289. doi:10.1038/nmeth.1313.

- Neph et al. *An expansive human cis-regulatory lexicon encoded in transcription factor footprints.* Nature 489:83-90, 2012

- Piper et al. *Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data.* Nucleic Acids Research, 2013, Vol. 41, No. 21 e201

- Boyle et al. *F-Seq: a feature density estimator for high-throughput sequence tags.* Bioinformatics Vol. 24 no. 21 2008, pages 2537–2538