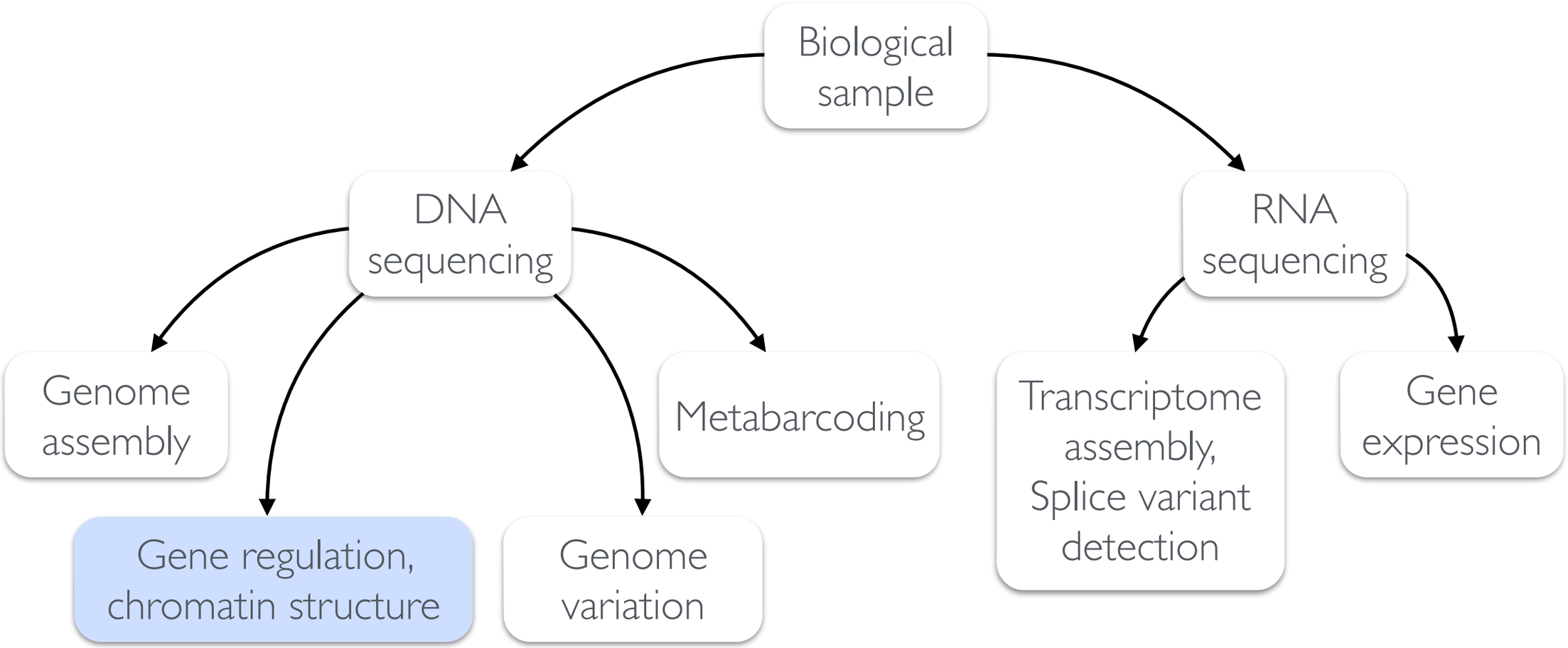
MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

ChIP-seq

Sebastian Schmeier
s.schmeier@gmail.com
http://sschmeier.github.io/bioinf-workshop/
2015

THE ENGINE
OF THE NEW
NEW ZEALAND

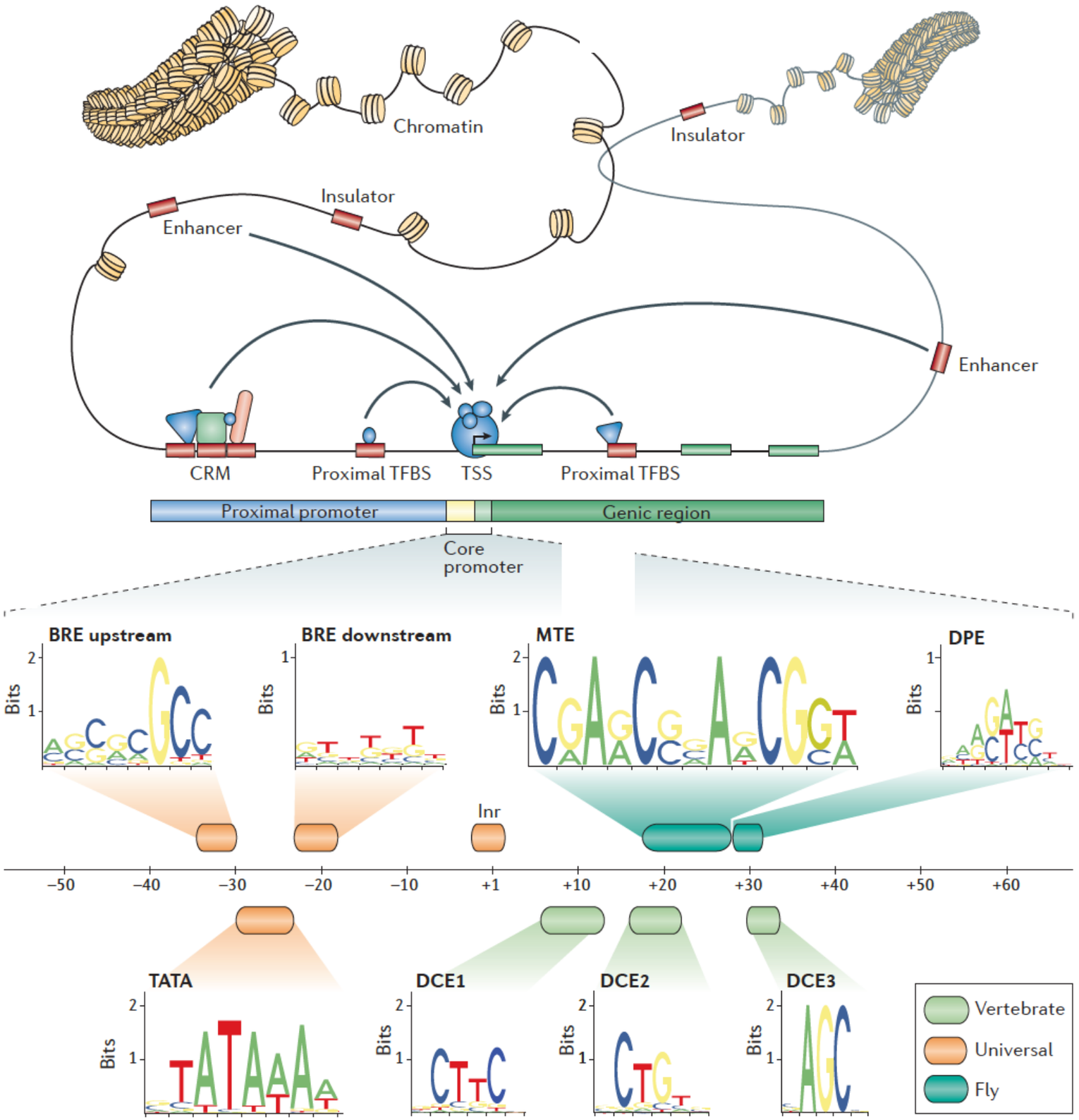
Common analyses overview



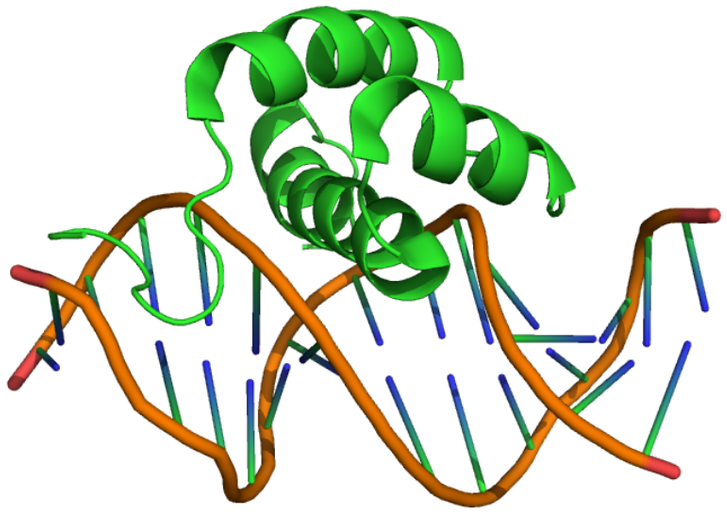
Gene regulation, chromatin structure

- Overview
- How do we analyse it?
  - ChIP
  - Experimental design considerations
  - Data analyses
  - Downstream analyses
  - Methods for higher base-pair resolution

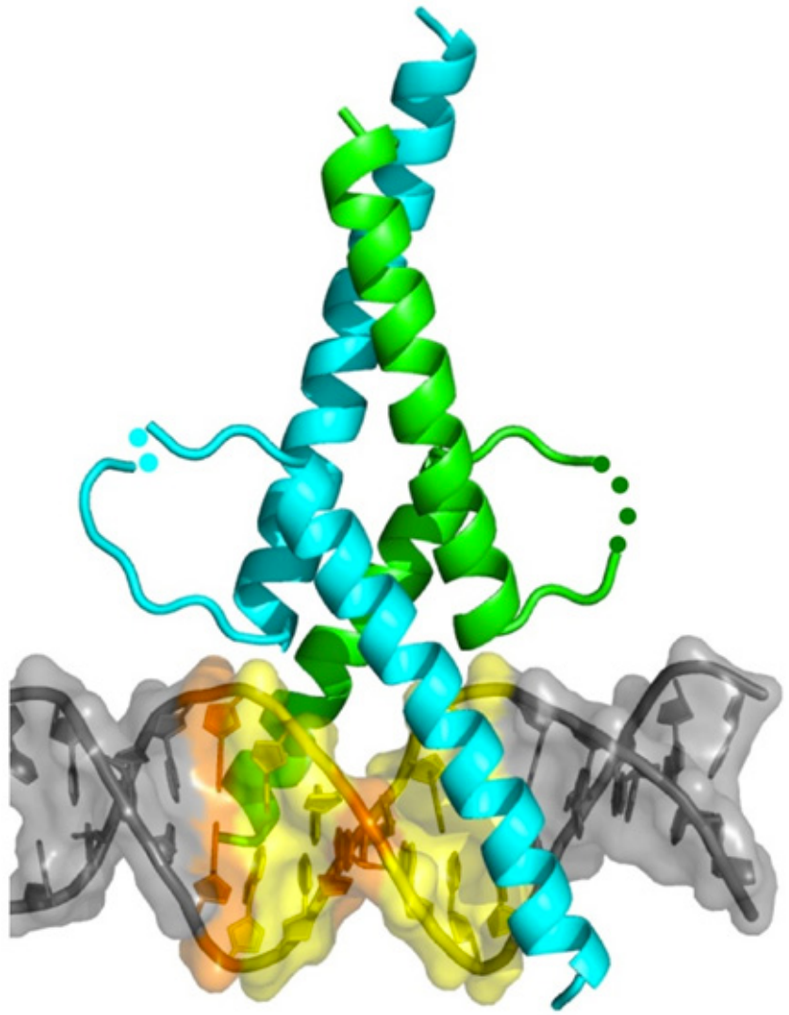
RNA Pol II promoter



Transcription factors



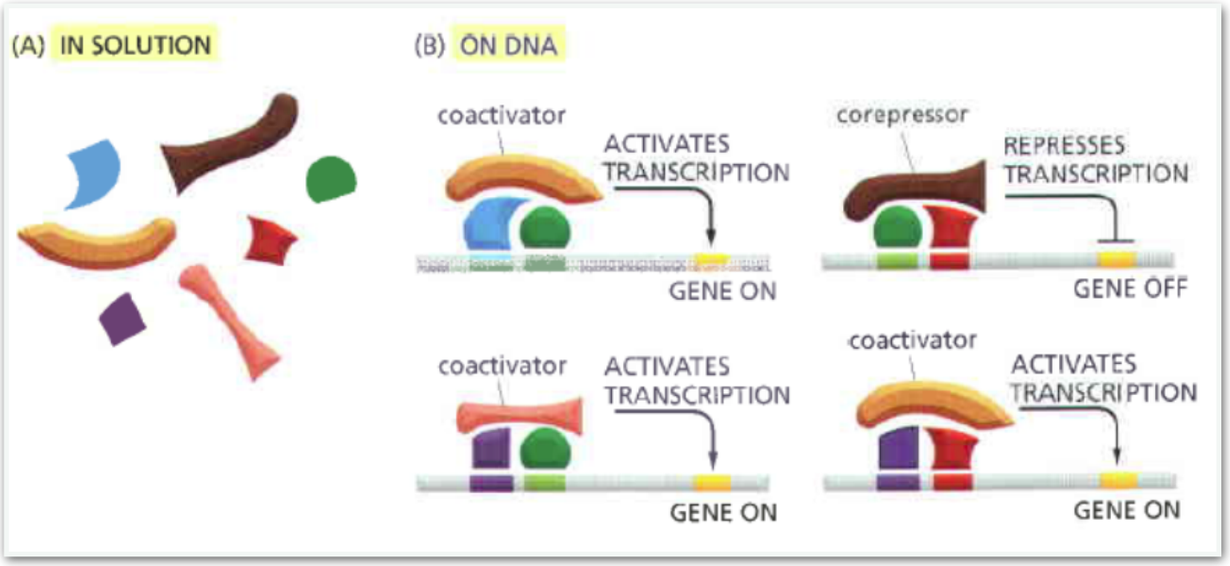
Homeodomain



Zinc-finger

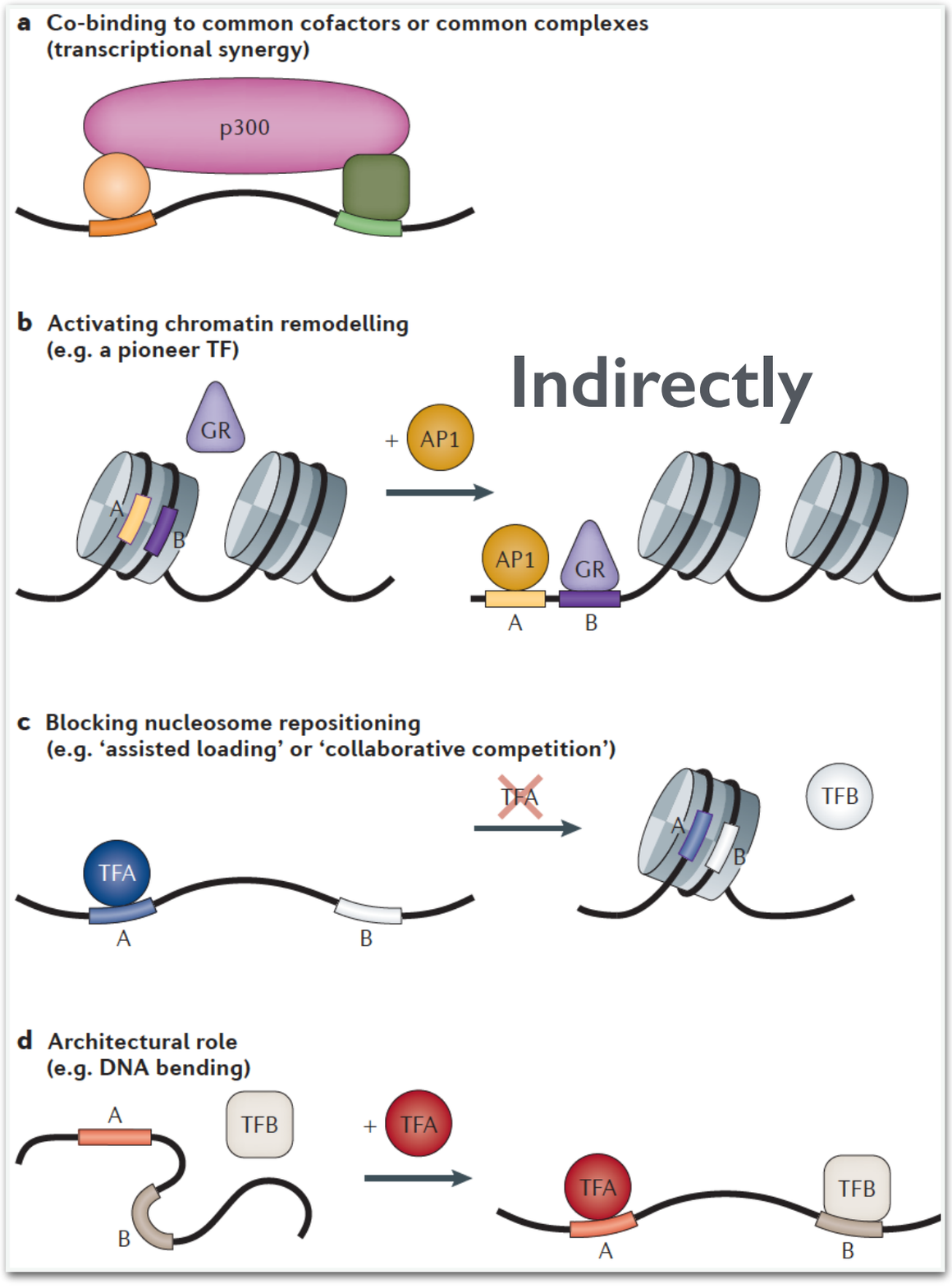
Transcription factors cooperate

Directly



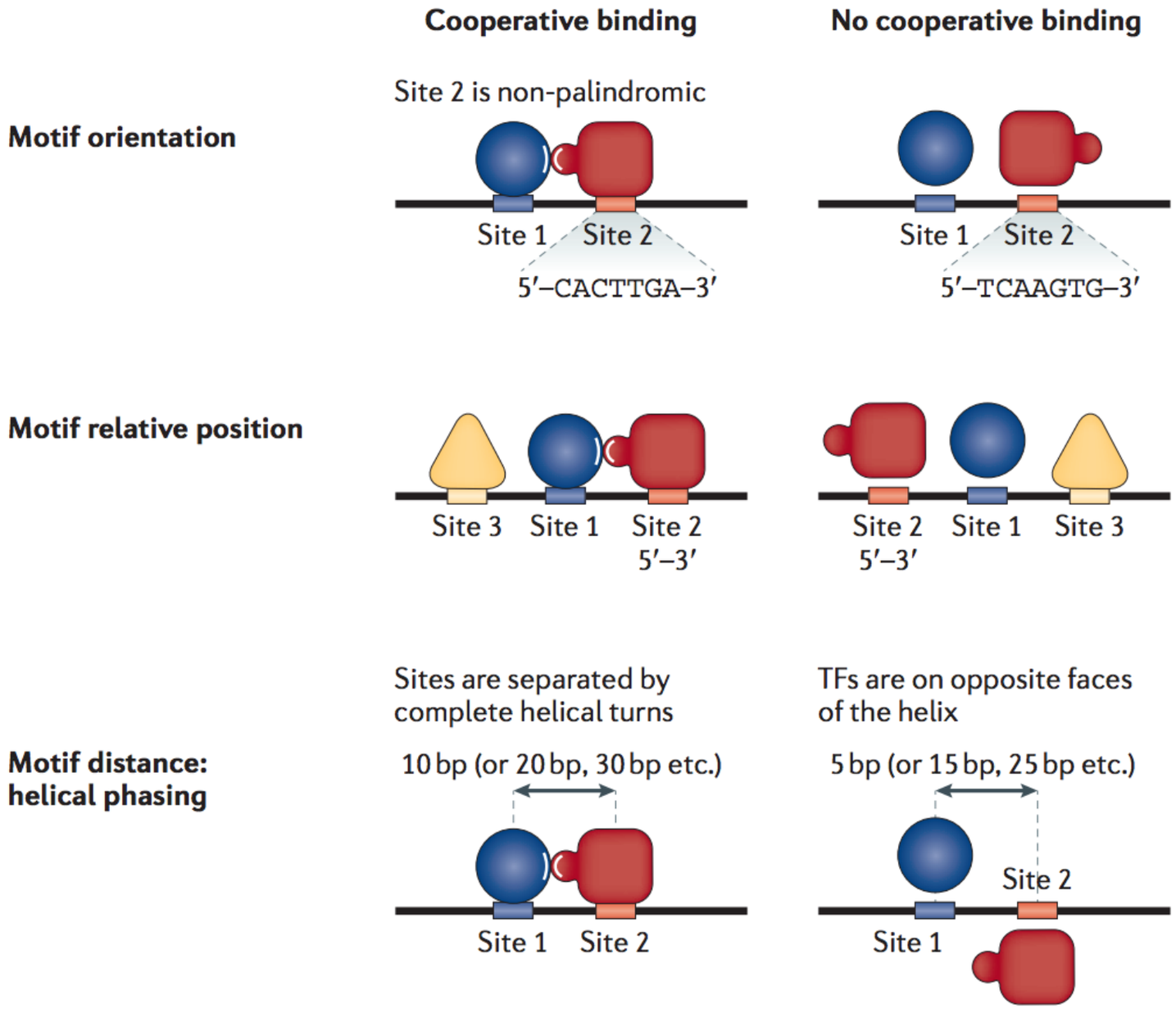
Molecular Biology of the Cell, Bruce Alberts, 5th ed., Page 447

Indirectly



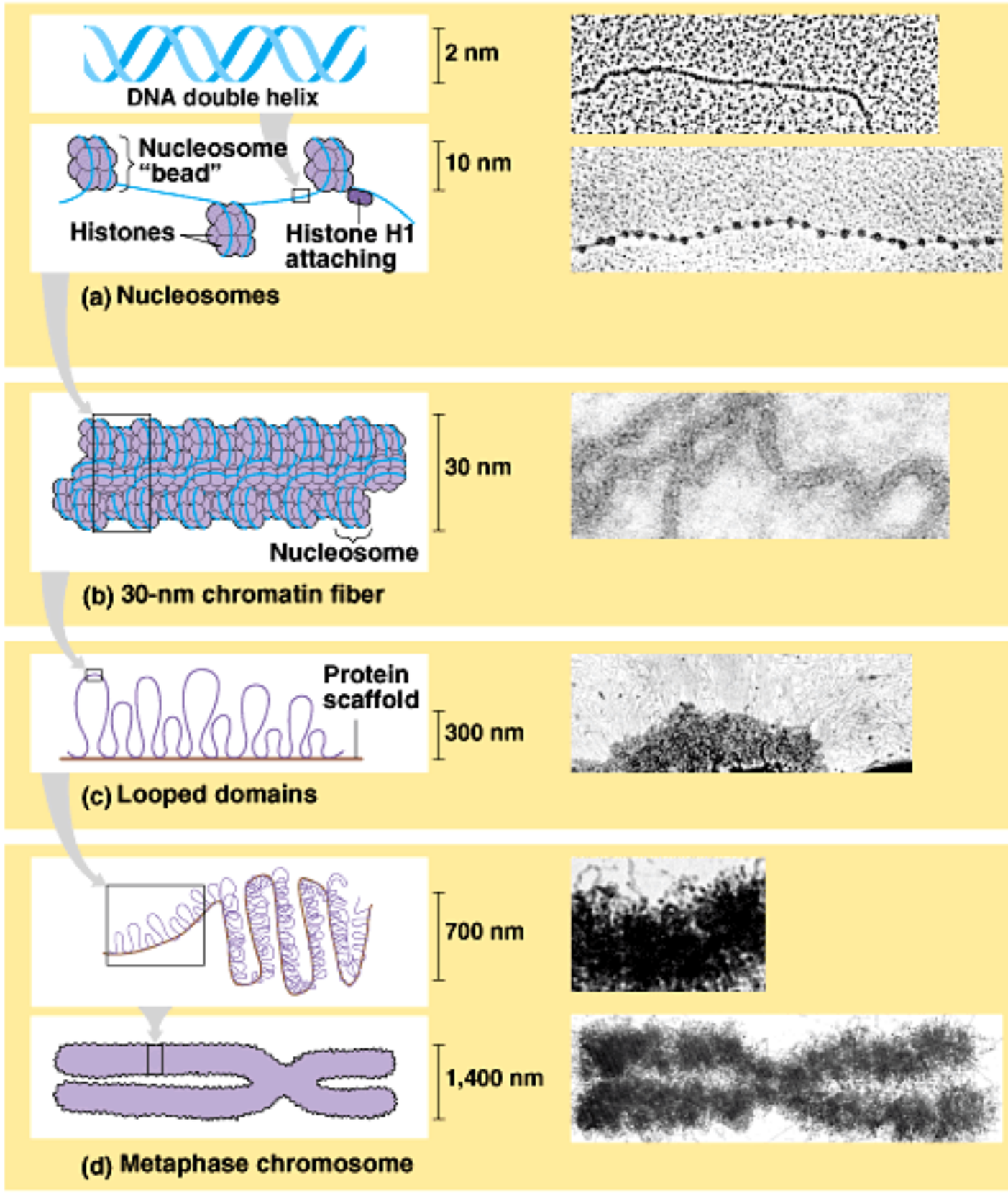
http://www.nature.com/nrg/journal/v13/n9/full/nrg3207.html

Location location location



http://www.nature.com/nrg/journal/v13/n9/full/nrg3207.html

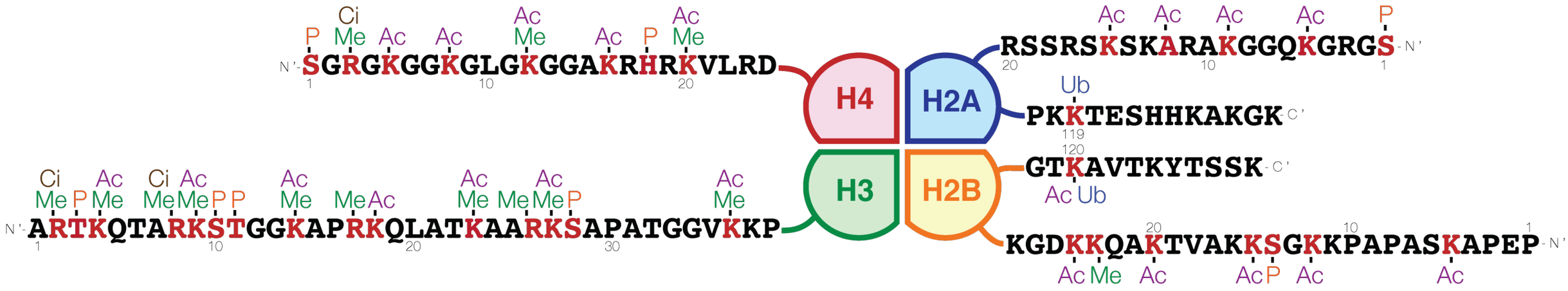
DNA packaging



Copyright © Pearson Education, Inc., publishing as Benjamin Cummings.

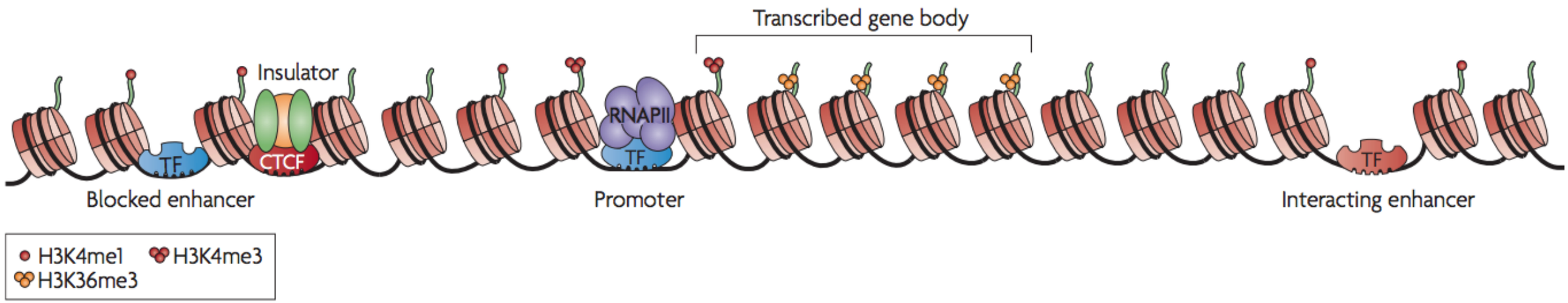
Histone modifications

- Histone modifications specify if DNA is compacted



https://en.wikipedia.org/wiki/Histone#/media/File:Histone_modifications.png

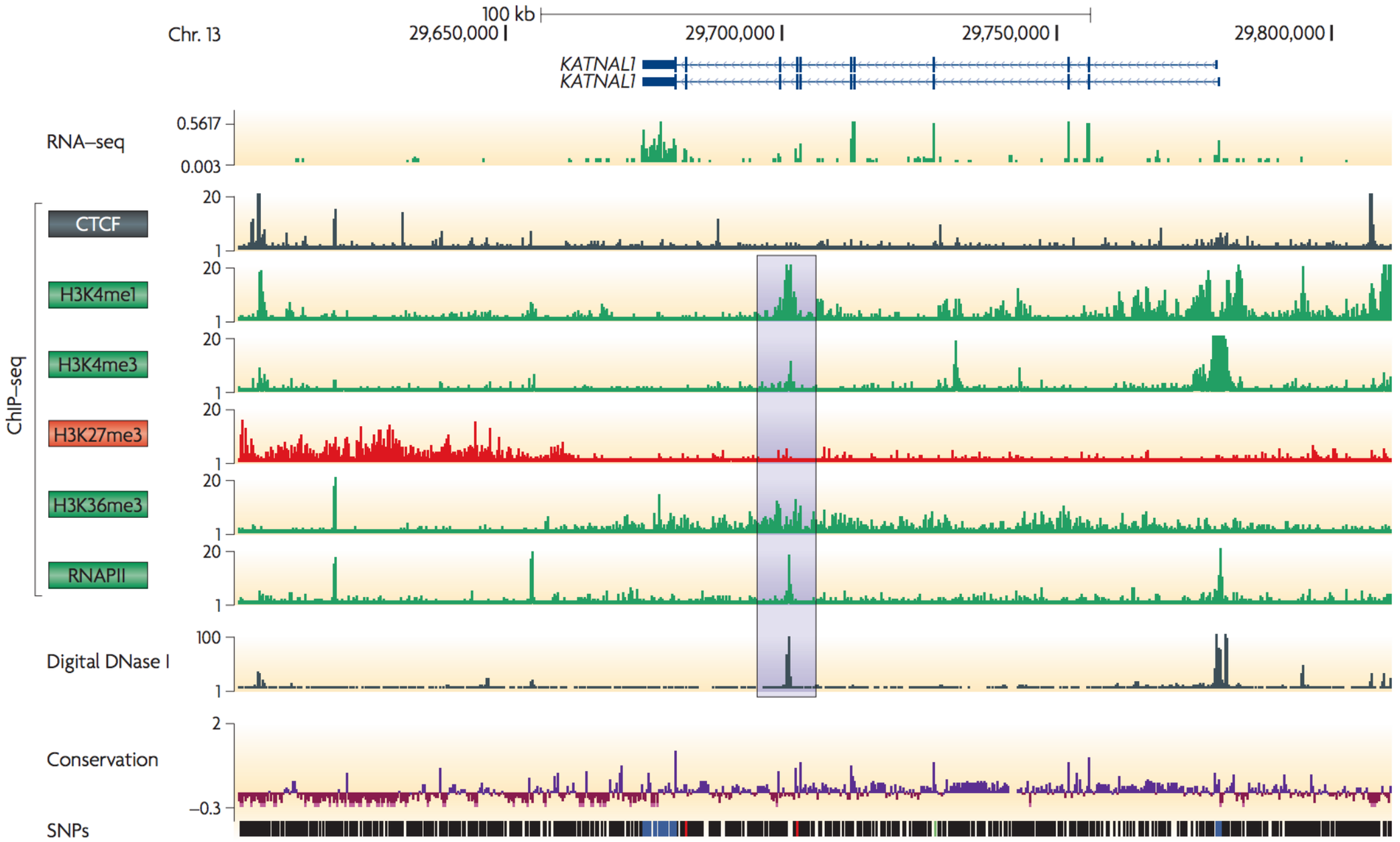
Where are the proteins on the DNA?



H3K4me3: Marks the location of promoters
H3K36me3: Indicates the bodies of transcribed genes and non-coding RNA
H3K4me1: Identifies functional enhancer elements outside of promoters

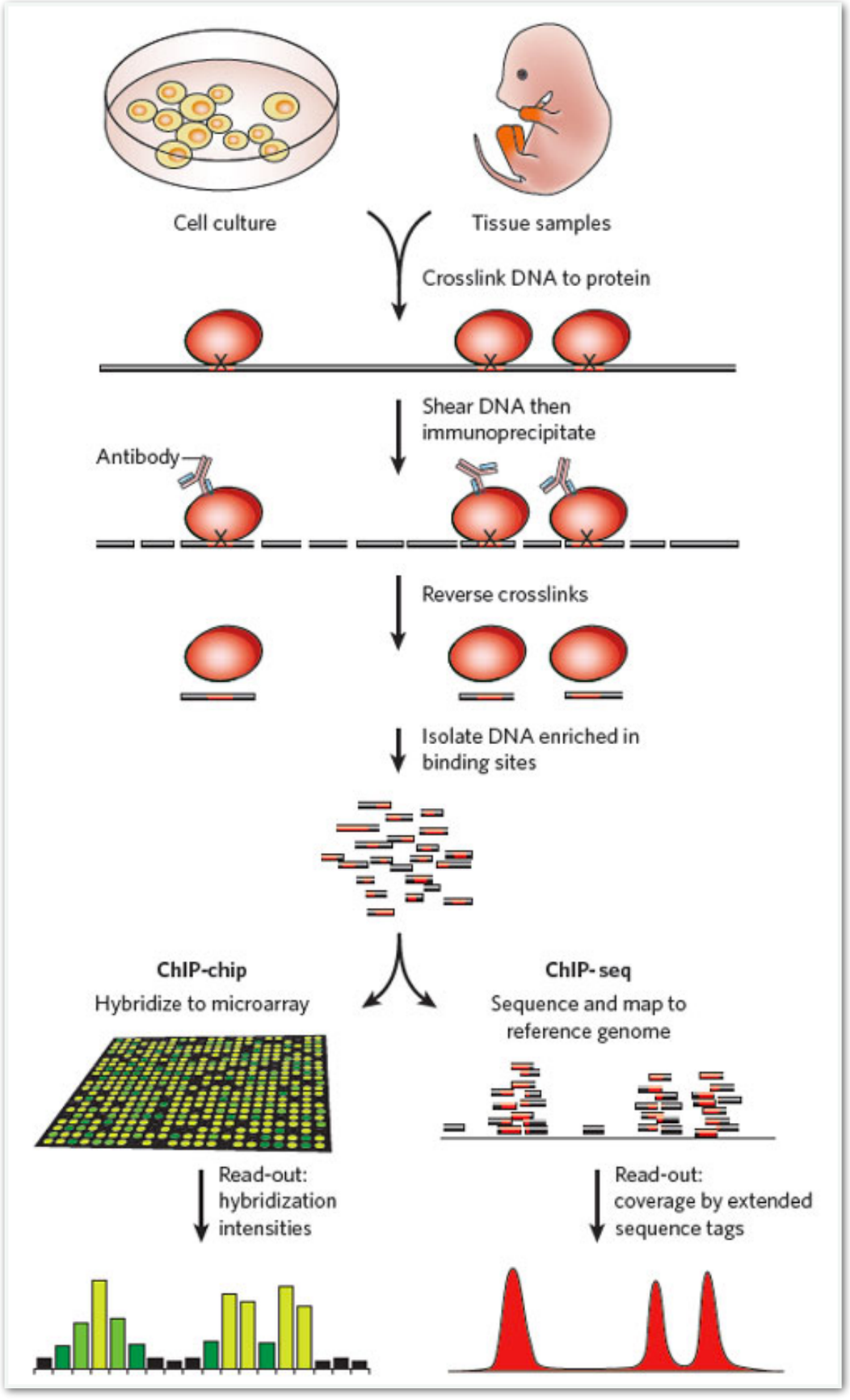
http://www.nature.com/nrg/journal/v11/n7/full/nrg2795.html

Where are the proteins on the DNA?



http://www.nature.com/nrg/journal/v11/n7/full/nrg2795.html

Chromatin immunoprecipitation (ChIP)



- A way to detect what DNA a certain protein is binding to
- You cross-link the sample, and fragment the DNA into pieces
- Immunoprecipitate using an antibody to your protein of interest
- Reverse the cross-links, and isolate the DNA
- PCR to see if your DNA is there
- You can also sequence the DNA and map to a reference genome (**ChIP-seq**) or hybridise it to a microarray (**ChIP-chip**)

http://www.nature.com/nature/journal/v461/n7261/full/nature08451.html

Chromatin immunoprecipitation (ChIP)

chip vs. seq

	ChIP-chip	ChIP-seq
Resolution	Array-specific	High - single nucleotide
Coverage	Limited by sequences on the array	Limited by "alignability" of reads to the genome, increases with read length
Repeat elements	Masked out	Many can be covered (40% of human genome is repetitive but 80% is uniquely mappable)
Cost	400-800$ per array (1-6M probes), multiple arrays needed for human genome	Around 1000$ per lane; 20-30M reads
Source of noise	Cross hybridization	Sequencing bias, GC bias, sequencing error
Amount of ChIP DNA required	High, few micrograms	Low 10-50ng
Dynamic range	Lower detection limit and saturation at high signal	Not limited
Multiplexing	Not possible	Possible

Chromatin immunoprecipitation (ChIP)

Lab procedures

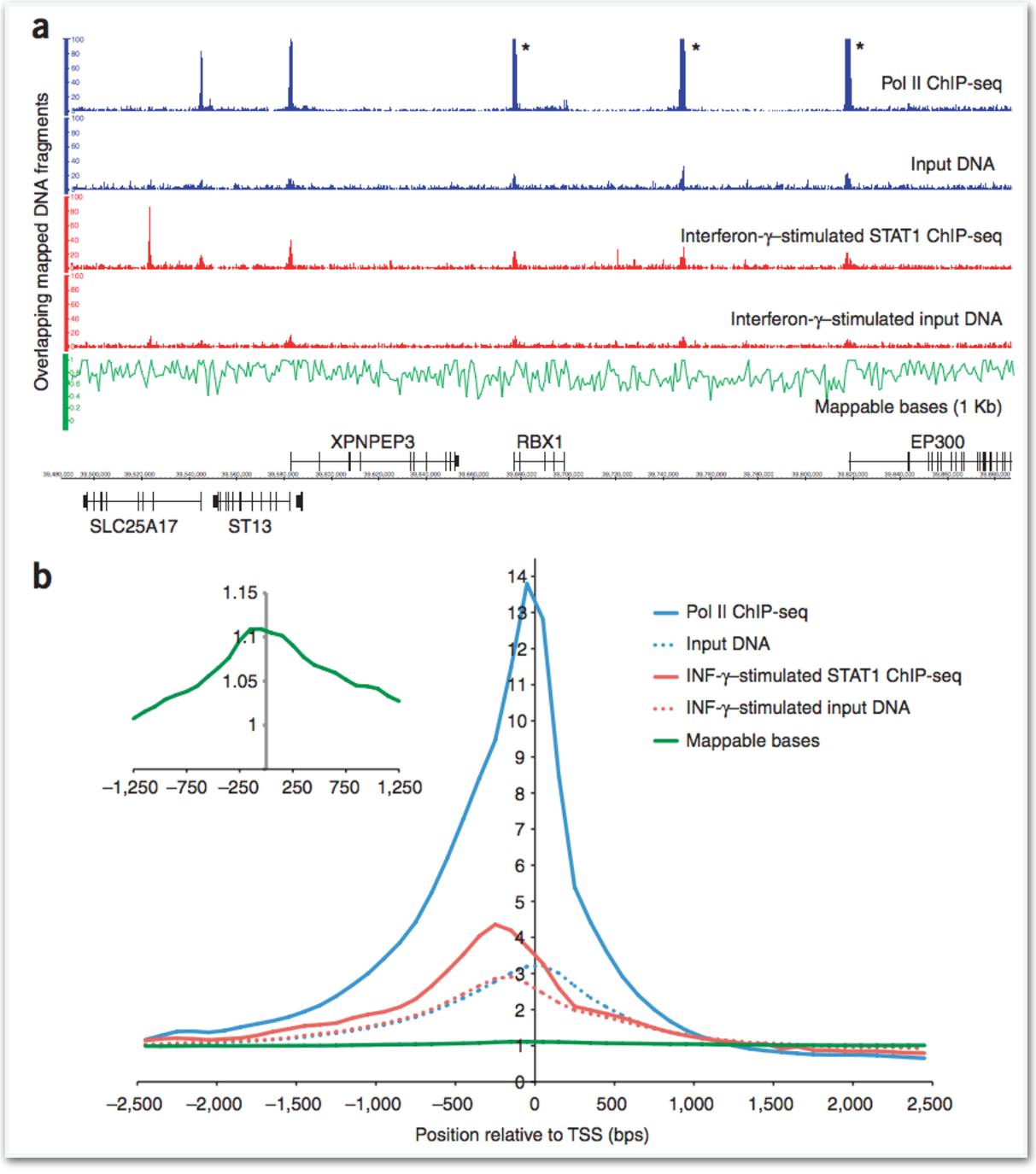
	Transcription factor binding location	Map nucleosome positions or histone modifications
Crosslinking	Formaldehyde	Usually not
Fragmentation	Sonication (200-600bp)	MNase treatment
Immunoprecipitation	Antibody specific to protein of interest	Antibody specific for histone modification

Experimental Design

- **Antibody quality**
  - A sensitive and specific antibody will give a high level of enrichment
  - Limited efficiency of antibody is the main reason for a failed ChIP-seq experiments
  - Check your antibody ahead if possible, e.g. Western blotting to check the reactivity of the antibody with unmodified and non-histone proteins.

Chromatin immunoprecipitation (ChIP)

Experimental Design



- **Control experiments**
  - Open chromatin regions are fragmented more easily than closed regions.
  - Uneven distribution of sequence tags across the genome
- Some fraction of the peaks in the ChIP-seq signal map for a TF might be due to the nature of the chromatin structure in regions of open chromatin
- ***A ChIP-seq peak should be compared with the same region in a matched control***

http://www.nature.com/nbt/journal/v27/n1/full/nbt.1518.html

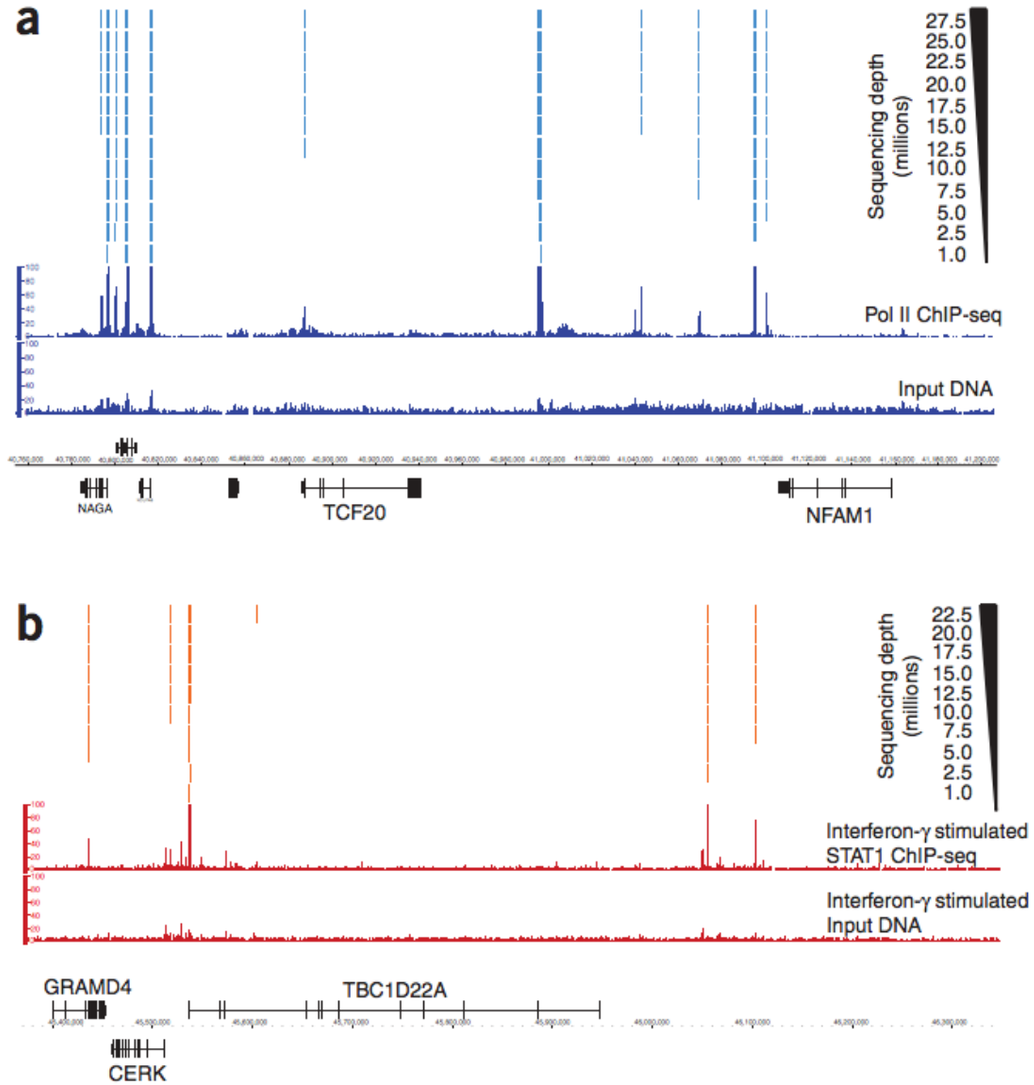
Experimental Design

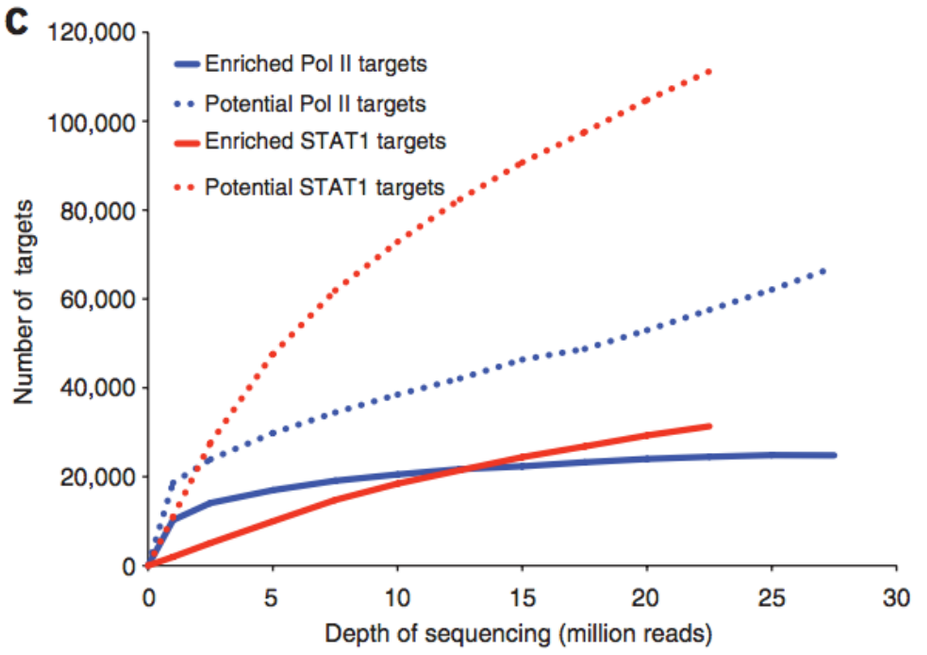
- **Control type**
  - *Input DNA*
  - *Mock IP* - DNA obtained from IP without antibody
    - Very little material can be pulled down leading to inconsistent results of multiple mock IPs.
  - *Nonspecific IP* - using an antibody against a protein that is not known to be involved in DNA binding
  - There is no consensus on which is the most appropriate
  - *Sequencing a control can be avoided when looking at:*
    - time points
    - differential binding pattern between conditions

Chromatin immunoprecipitation (ChIP)

Experimental Design

- **Sequencing depth**
- More prominent peaks are identified with fewer reads, whereas weaker peaks require greater depth
- Number of putative target regions continues to increase significantly as a function of sequencing depth





d

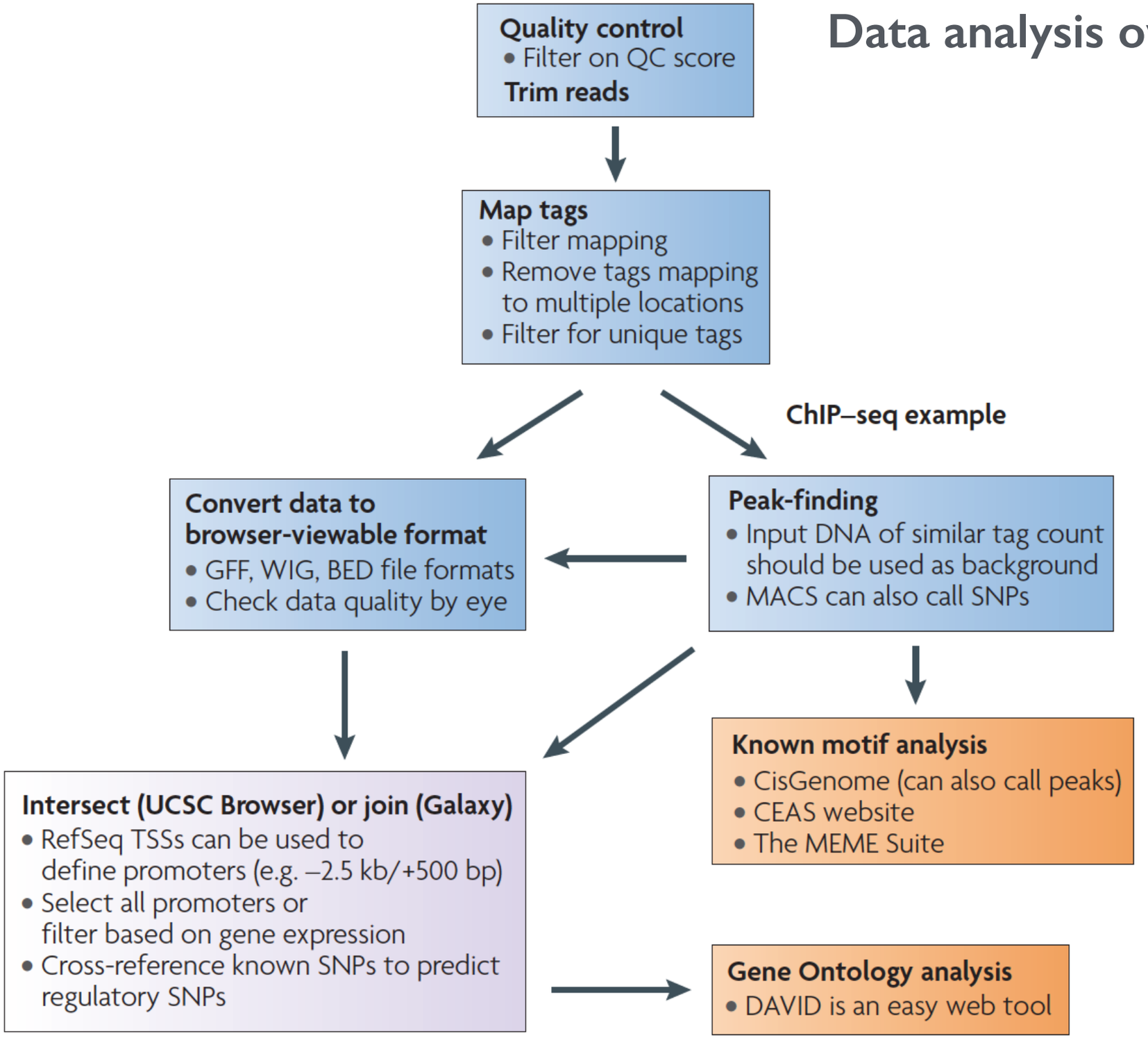
Number of replicas	Number of targets	True positives	False positives	False negatives	Sensitivity	Positive predictive value
1	20,902	18,833	2,069	5,906	0.761	0.901
2	20,733	19,257	1,476	5,482	0.778	0.929
3	20,328	19,126	1,202	5,613	0.773	0.941

http://www.nature.com/nbt/journal/v27/n1/full/nbt.1518.html

Experimental Design

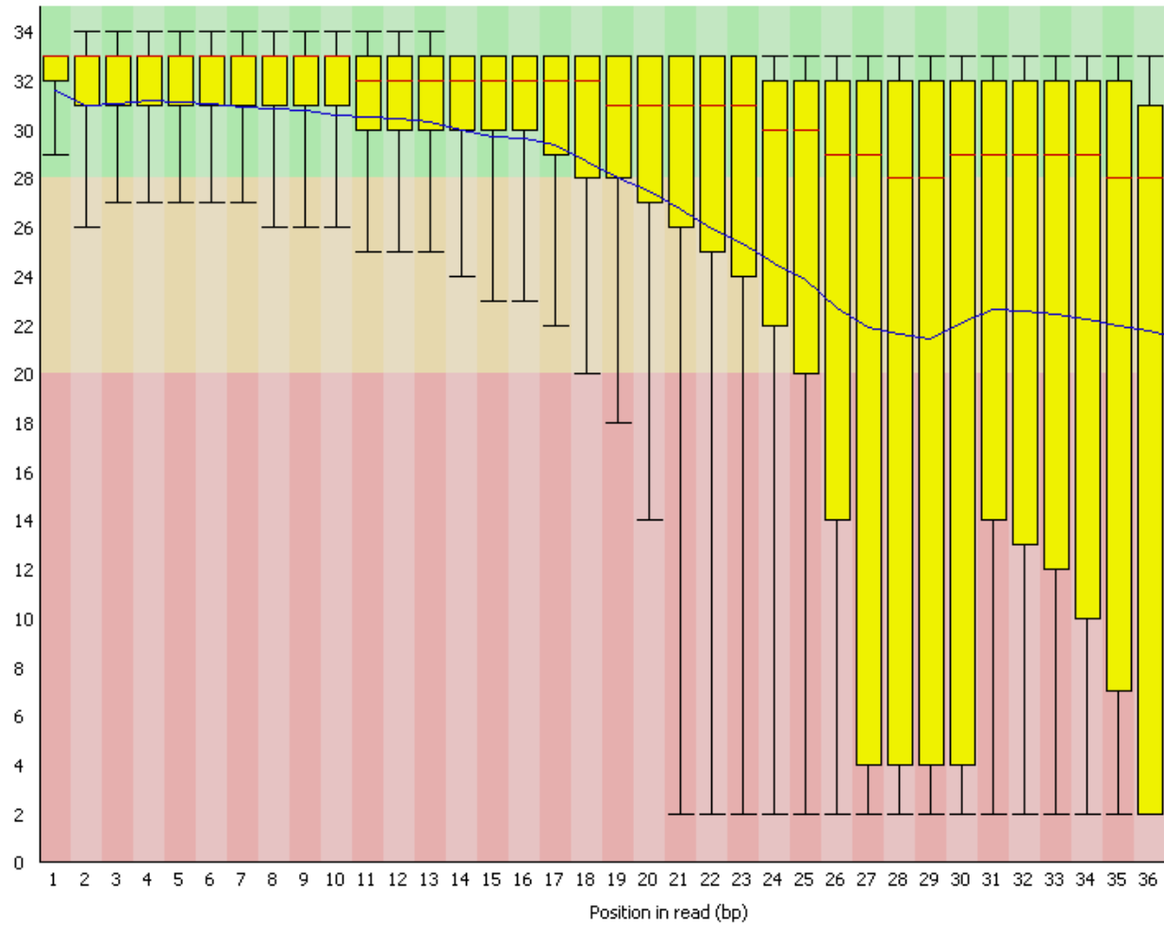
- **Multiplexing**
  - Number of reads per run continue to increase
  - The ability to sequence multiple samples (e.g. DNA from several ChIP experiments) at the same time becomes important, especially for small genomes
  - Different barcode adaptors are ligated to different samples
  - After sequencing reads are separated according to barcodes
- **Paired-end sequencing**
  - Reads are sequenced from both ends
  - Increase "mappability" - especially in repetitive regions
  - Costs more as single end reads
  - For ChIP-seq, usually not worth the extra cost, unless you have a specific interest in repeat regions

Data analysis overview



http://www.nature.com/nrg/journal/v11/n7/full/nrg2795.html

- Very short reads
- Use dynamic trimming to get rid of bad read ends
  - e.g. FastQC / SolexaQA++



http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Mapping reads

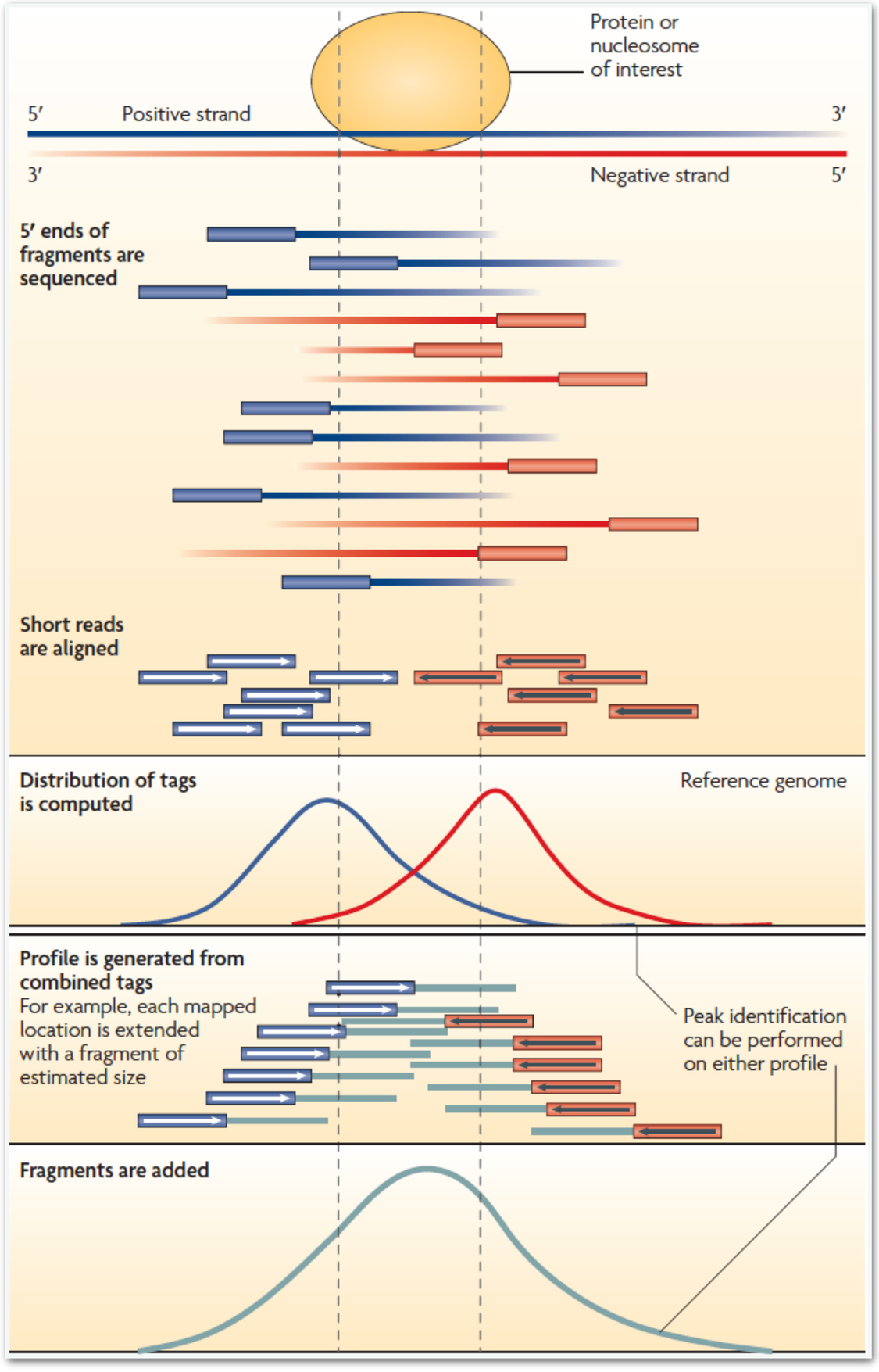
- Not all of the genome is 'available' for mapping
- Align your reads to the unmasked genome
- For ChIP-seq, usually short reads are used (36bp)
- Limited gain in using longer reads

		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%

http://www.nature.com/nbt/journal/v27/n1/full/nbt.1518.html

Mapping challenges

- Enormous amount of reads to align
- Done against large genome - needs pre-indexing structure and large memory
- Has to be fast and memory efficient
- Shorter read length
- Mismatches
- Repetitive regions
- Multi-mapping reads



http://www.nature.com/nrg/journal/v10/n10/abs/nrg2641.html

Chromatin immunoprecipitation (ChIP)

Peak calling

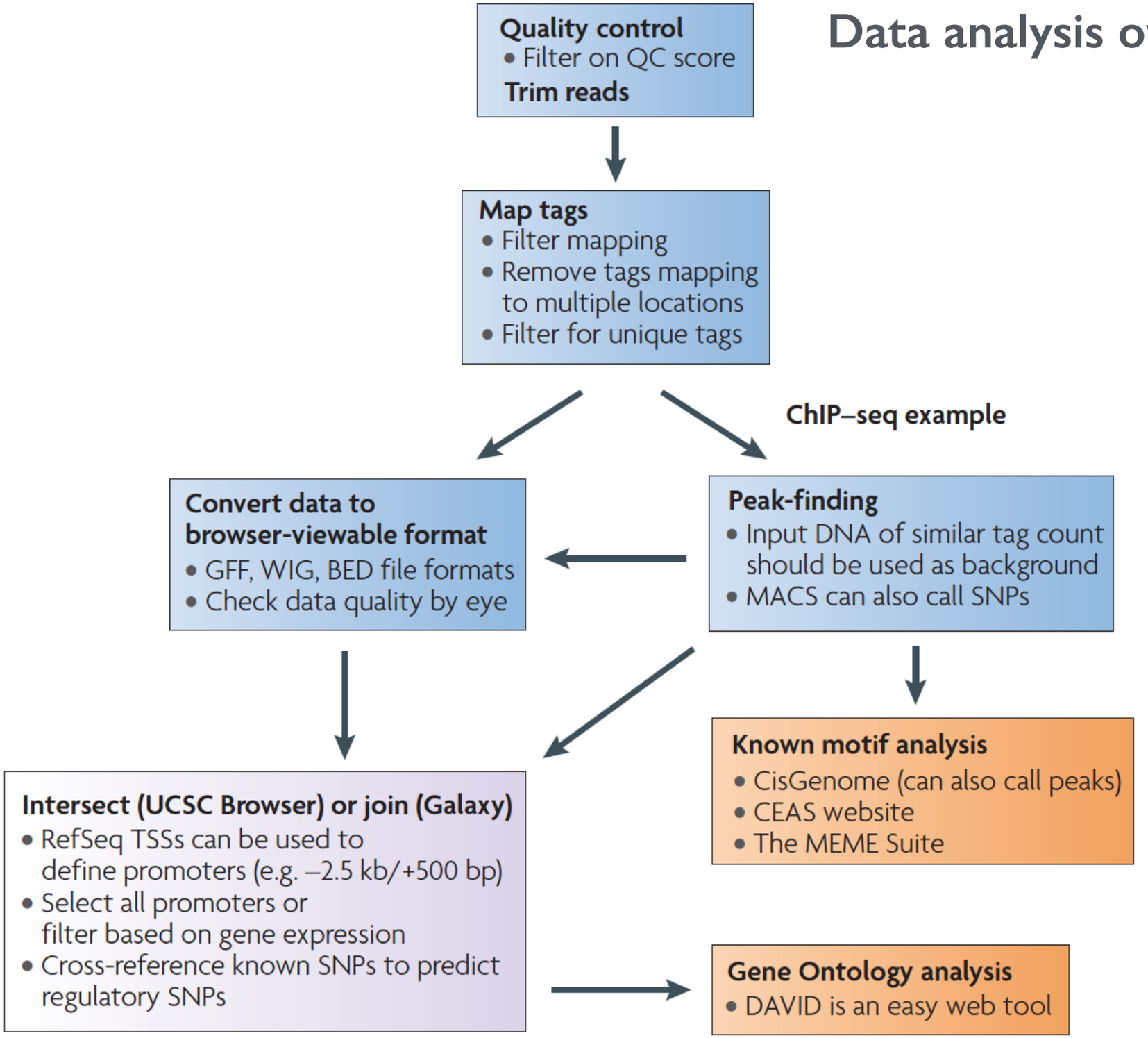
- Basic - regions are scored by the number of tags in a window of a given size.
- Each region is assessed by enrichment over control and minimum tag density.
- Advanced - take advantage of the directionality of the reads.

Challenges

- Adjust for sequence alignability - regions that contain repetitive elements have different expected tag count
- Different ChIP-seq applications produce different type of peaks. Most current tools have been designed to detect sharp peaks (TF binding, histone modifications at regulatory elements)

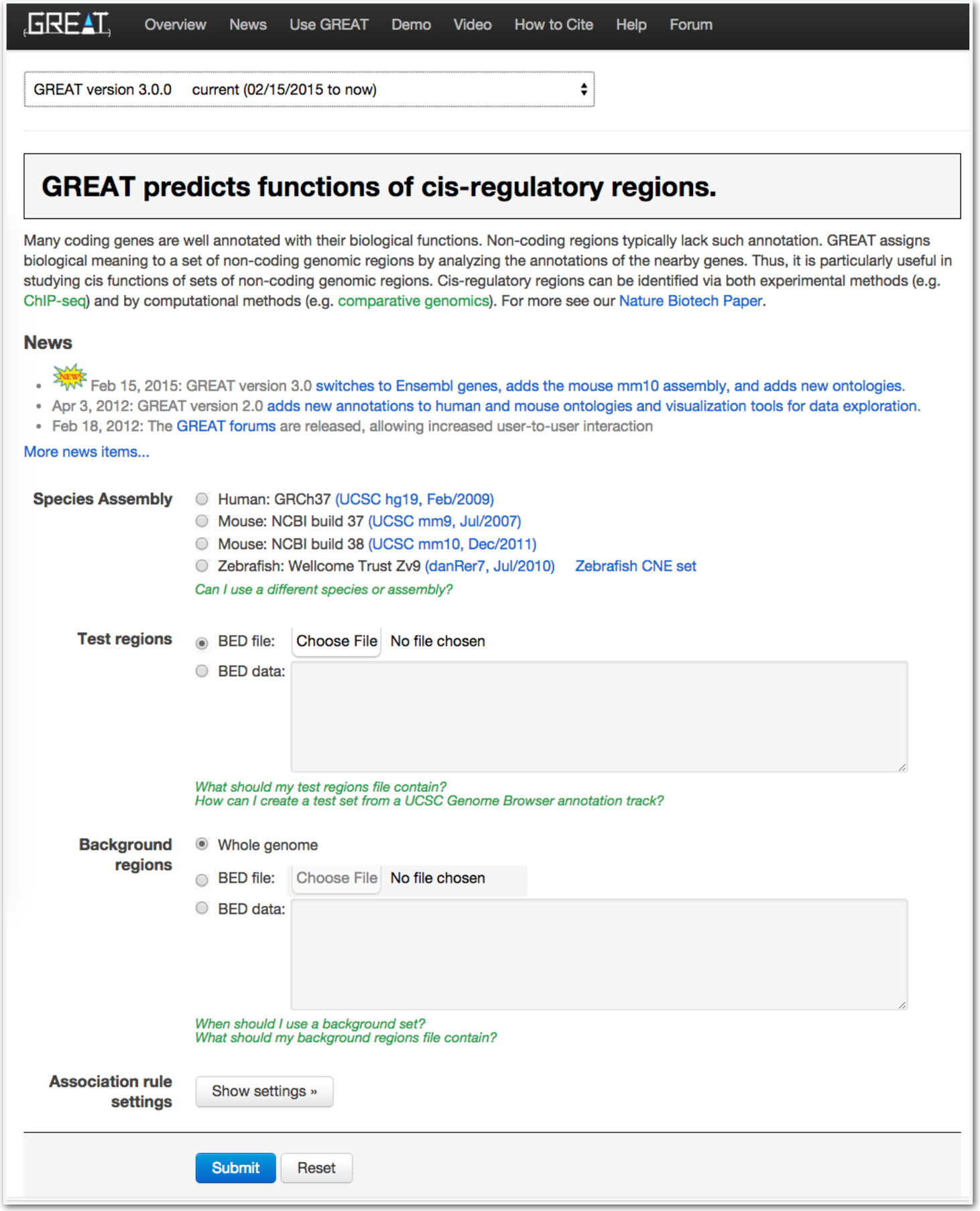
Chromatin immunoprecipitation (ChIP)

Data analysis overview



http://www.nature.com/nrg/journal/v11/n7/full/nrg2795.html

GREAT



http://bejerano.stanford.edu/great/public/html/

- Find some function based on the genes that are close to your peak regions

GREAT

GREAT Overview News Use GREAT Demo Video How to Cite Help Forum Bejerano Lab, Star

GREAT version 3.0.0 current (02/15/2015 to now)

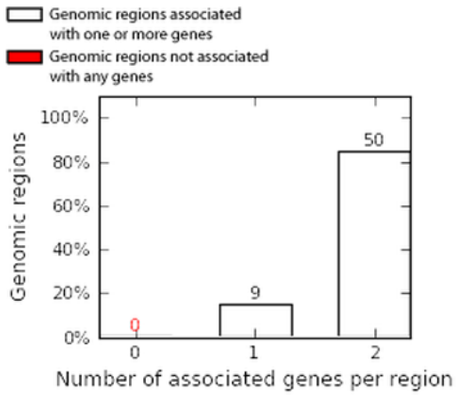
Job Description

Region-Gene Association Graphs

What do these graphs illustrate?

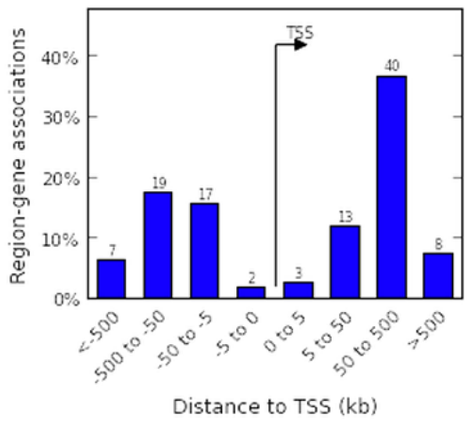
Number of associated genes per region

Download as PDF.



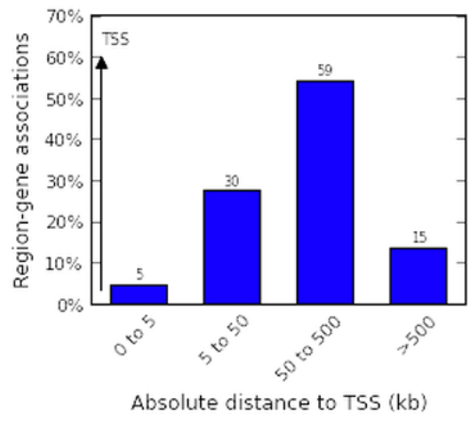
Binned by orientation and distance to TSS

Download as PDF.



Binned by absolute distance to TSS

Download as PDF.



Global Controls

Global Export NEW *Which data is exported by each option?*

GO Molecular Function (4 terms)

Global controls

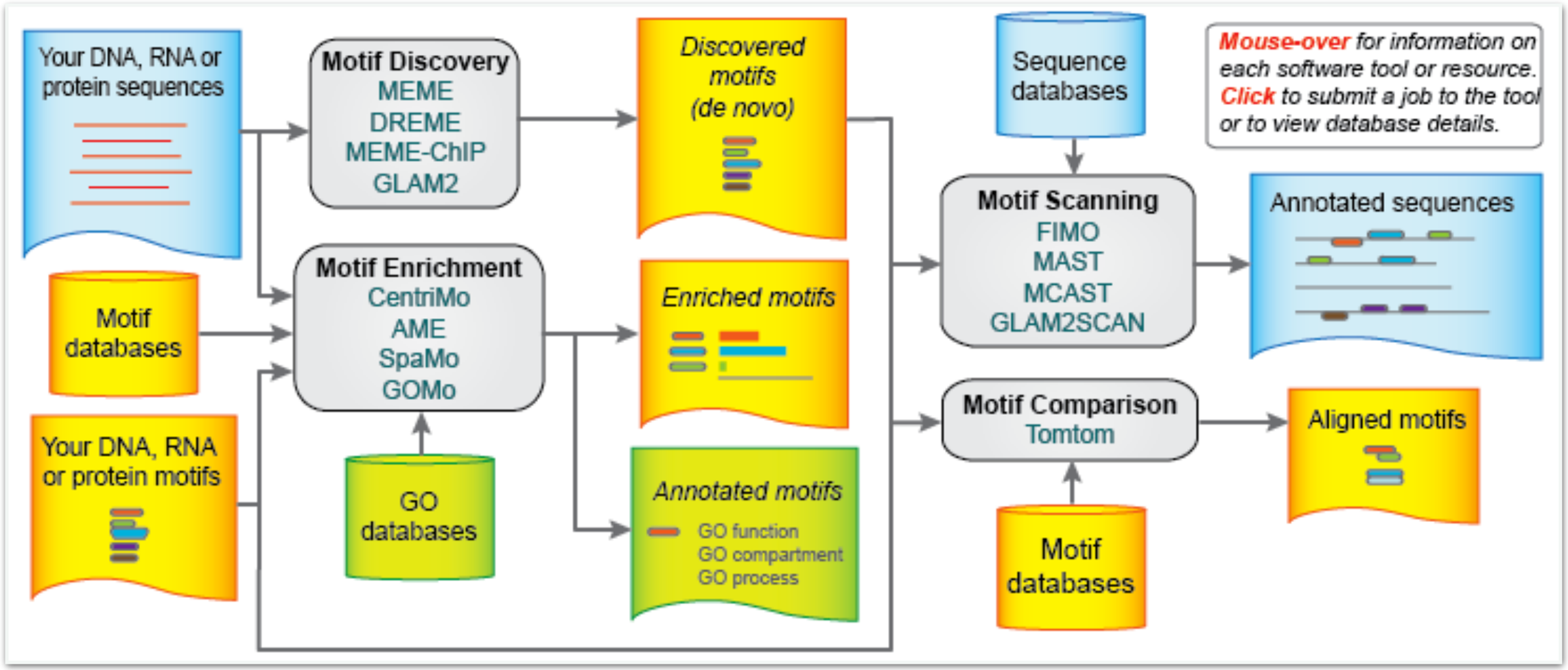
Table controls: Export Shown top rows in this table: 20 Set Term annotation count: Min: 1 Max: Inf Set Visualize this table: NEW [select one]

Term Name	Binom Rank	Binom Raw P-Value	*Binom FDR Q-Val*	Binom Fold Enrichment	Binom Observed Region Hits	Binom Region Set Coverage	Hyper Rank	*Hyper FDR Q-Val*	Hyper Fold Enrichment	Hyper Observed Gene Hits	Hyper Total Genes	Hyper Gene Set Coverage
aromatase activity	1	**8.6047e-8**	**2.9970e-4**	47.2131	5	8.47%	1	**1.2756e-5**	45.2479	6	36	7.69%
oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, reduced flavin or flavoprotein as one donor, and incorporation of one atom of oxygen	3	**2.6078e-7**	**3.0277e-4**	37.6685	5	8.47%	2	**8.9811e-5**	29.6168	6	55	7.69%
monooxygenase activity	15	**1.6488e-4**	**3.8284e-2**	9.7911	5	8.47%	3	**1.6399e-2**	11.4713	6	142	7.69%
heme binding	18	**2.2576e-4**	**4.3684e-2**	9.1443	5	8.47%	4	**4.1243e-2**	9.2552	6	176	7.69%

http://bejerano.stanford.edu/great/public/html/

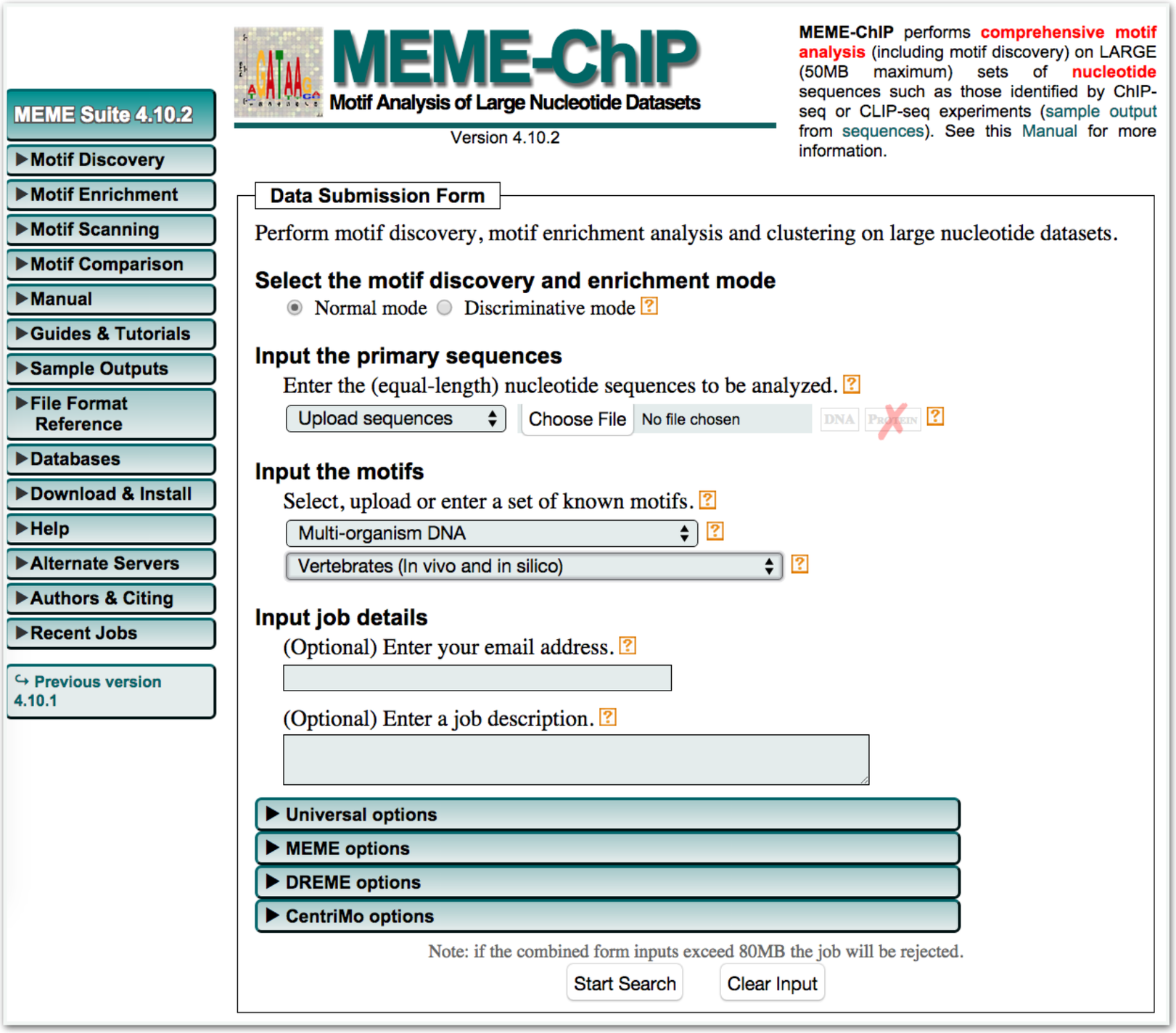
Find those motifs

- MEME suite



Find those motifs

- MEME suite
  - MEME-ChIP



http://meme-suite.org/

Find those motifs

- MEME suite
  - MEME-ChIP

Name	Alt. Name	Preview	Matches	List
1	MEME		1	UP00082_2 (Zfp187_secondary)
2	MEME		0	
3	MEME		2	MA0080.3 (Spi1), UP00031_2 (Zbtb3_secondary)

TARGET DATABASES Previous Next Top

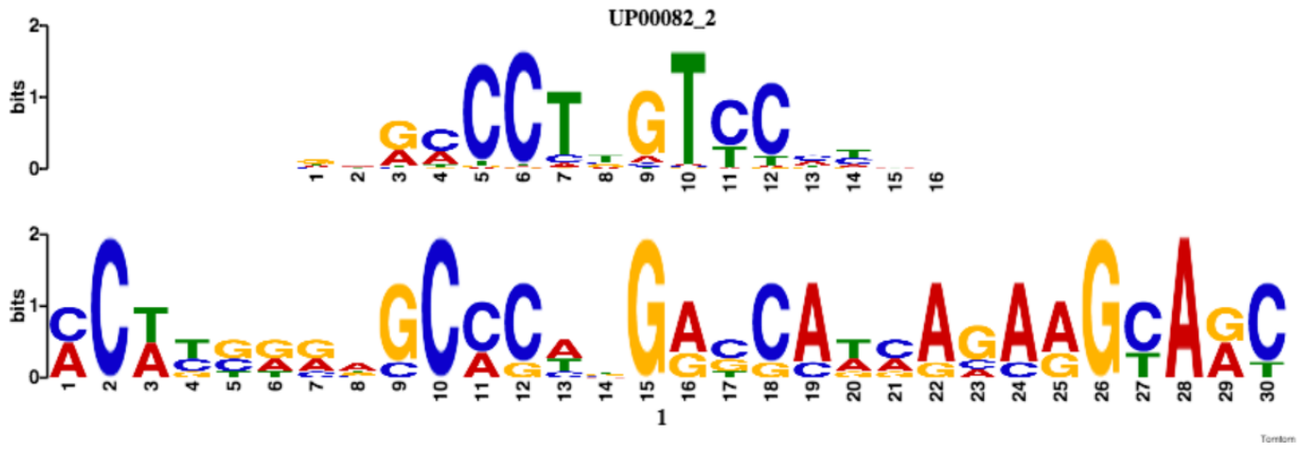
Database	Number of Motifs	Motifs Matched
jolma2013.meme	843	0
JASPAR_CORE_2014_vertebrates.meme	205	1
uniprobe_mouse.meme	386	2

MATCHES TO QUERY MOTIF 1 (MEME) Previous Next Top

Summary

Name	UP00082_2
Alt. Name	Zfp187_secondary
Database	uniprobe_mouse.meme
***p*-value**	0.000514725
***E*-value**	0.738115
***q*-value**	1
Overlap	16
Offset	-6
Orientation	Normal

Alignment

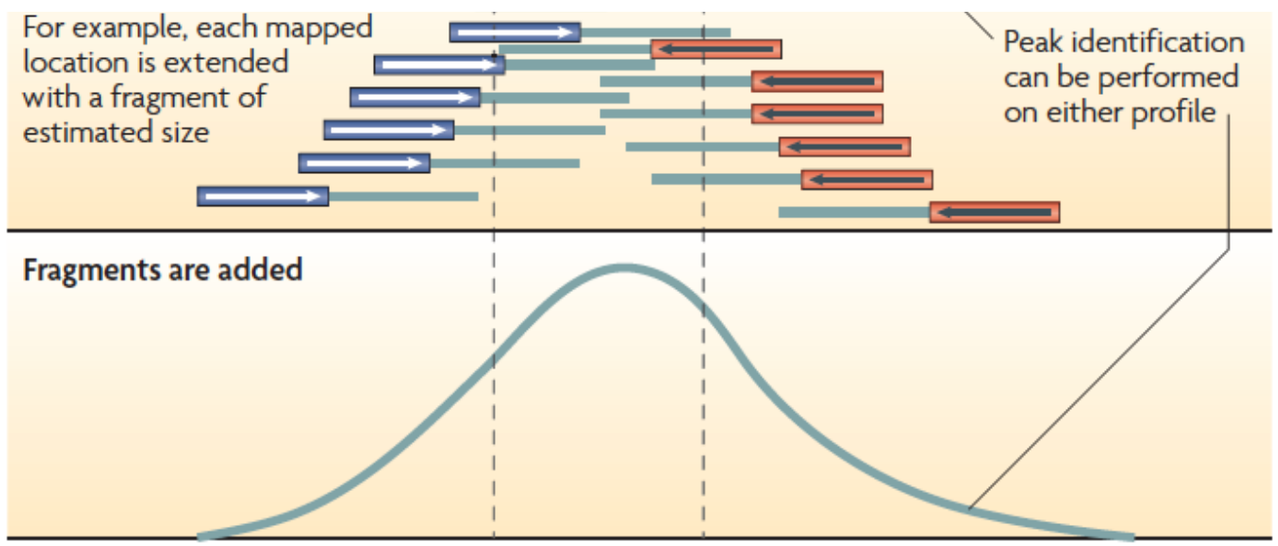


Create custom LOGO [Query Top]

http://meme-suite.org/

Base-pair resolution digital epigenome profiling

- The resolution of most epigenomic techniques, e.g. traditional ChIP-seq, is not high enough.
- It has been limited by the methods used to prepare chromatin.
- Traditional ChIP-seq gives you peak ***regions***

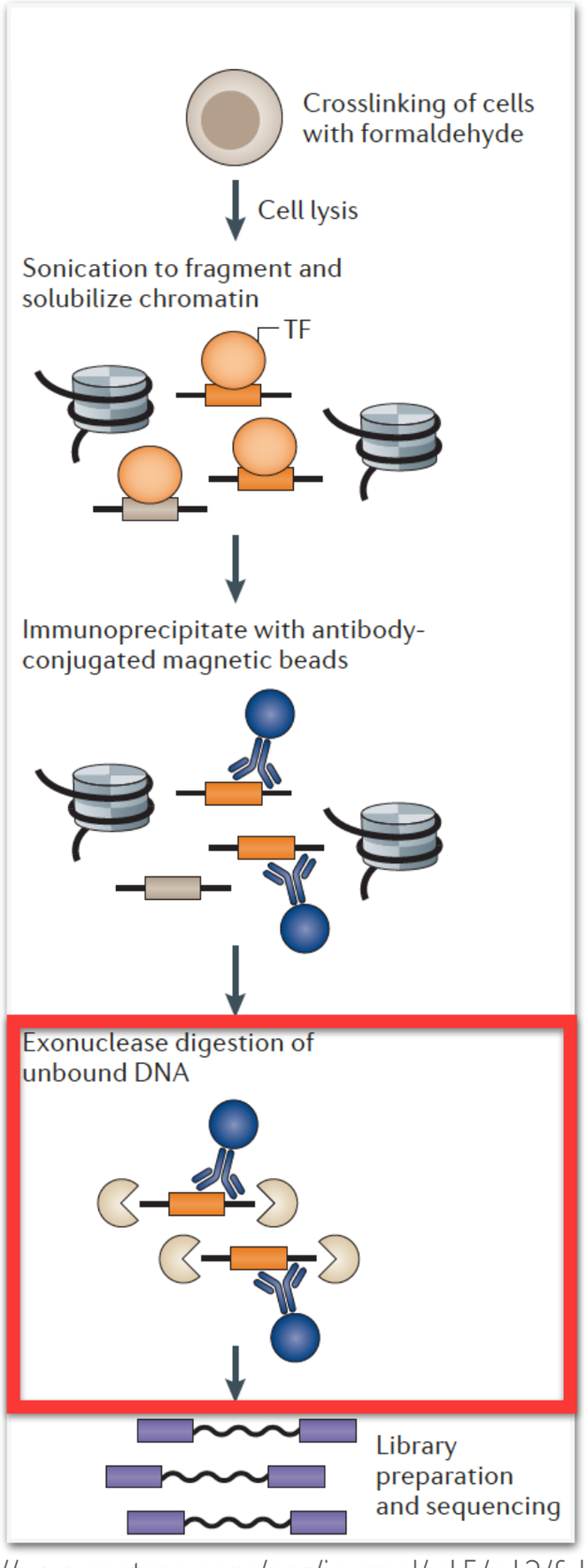


http://www.nature.com/nrg/journal/v10/n10/abs/nrg2641.html

Base-pair resolution digital epigenome profiling

But, base-pair resolution is essential because…

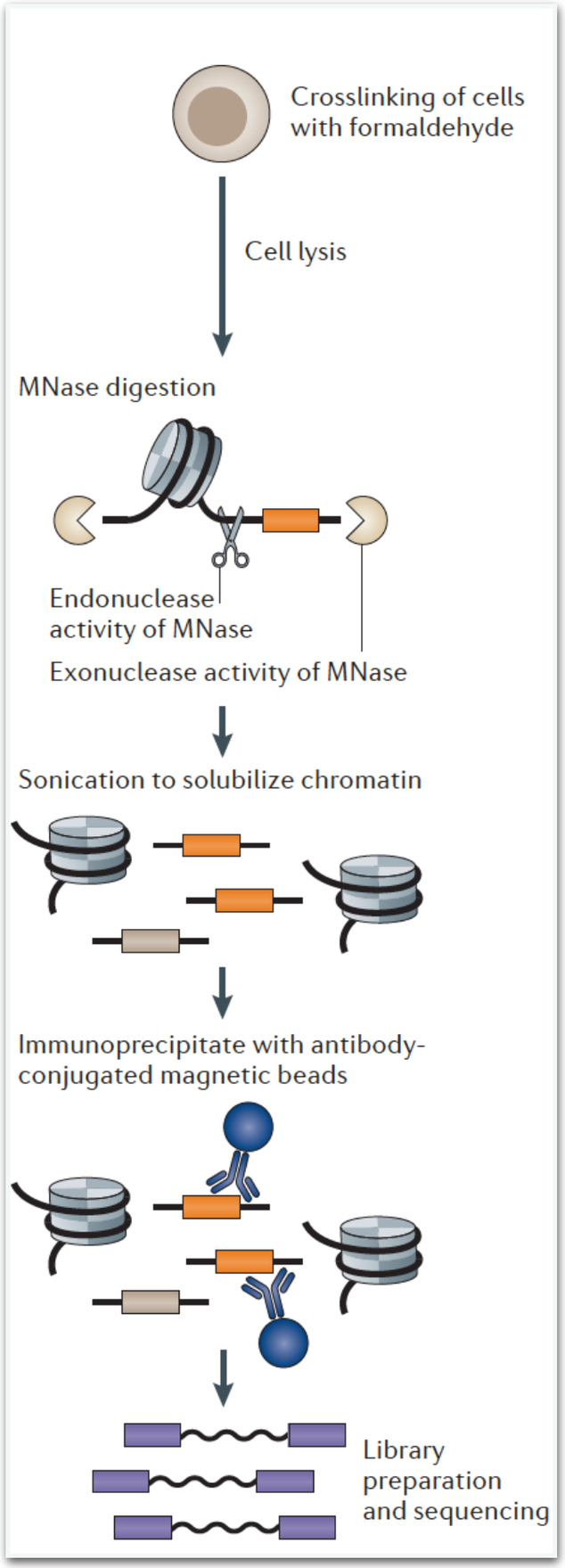
- single-base shifts in nucleosome positioning can alter chromatin structure
- the precise DNA sequence bound by a TF is important to deduce binding site motifs
- it allows the identification of single binding sites within a cluster of closely spaced sites



ChIP-exo

- Crosslinked cells are sonicated to fragment and solubilise chromatin.
- ChIP is carried out with an antibody directed against the protein of interest.
- *Immunoprecipitated DNA is digested with exonucleases to remove DNA that is not protected by the protein*
- Resected DNA is then purified and sequenced
- Less straightforward than standard ChIP–seq

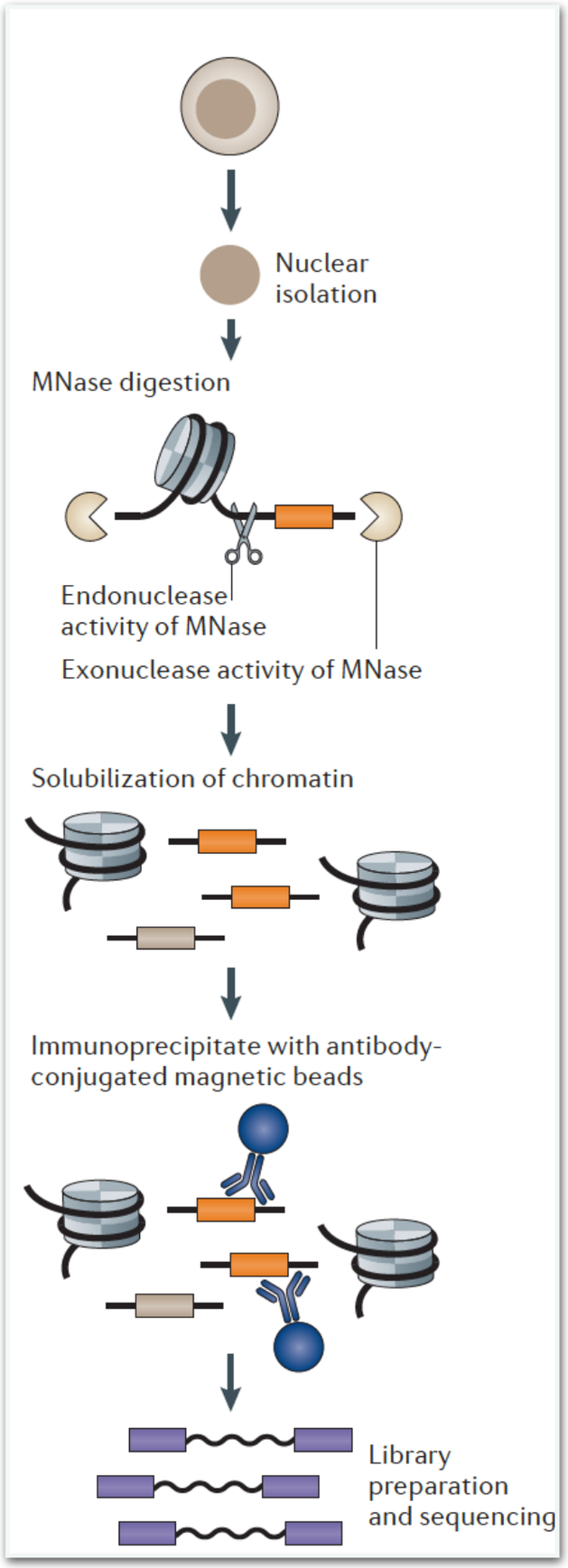
http://www.nature.com/nrg/journal/v15/n12/full/nrg3798.html



High-resolution X-ChIP

- Crosslinked cells are lysed and chromatin is digested with micrococcal nuclease (MNase).
- Chromatin is then sonicated to improve solubility.
- The use of both MNase and sonication results in near-complete solubilisation of some chromatin-bound proteins, making it especially useful for large complexes that resist solubilisation
- An antibody directed against the protein of interest is then used to immunoprecipitate DNA, which is then purified and sequenced

http://www.nature.com/nrg/journal/v15/n12/full/nrg3798.html



ORGANIC

- In the *occupied regions* of genomes from affinity-purified *naturally isolated chromatin* (ORGANIC) method, soluble chromatin extracted from MNase treatment of nuclei is used as an input to ChIP
- The input sample provides a genome-wide footprinting of factors
- ChIP pulldown provides a factor-specific map in a single experiment using a simple library preparation protocol
- Highly-specific and identifies more binding sites with *consensus motifs* than previous X-ChIP studies

http://www.nature.com/nrg/journal/v15/n12/full/nrg3798.html

X-ChIP vs. native ChIP

- Native ChIP is performed *without* crosslinking.
- It is usually applied to nucleosomes
  - The assumption being that the wrapping of DNA around histones *precludes* rearrangement during chromatin preparation and immunoprecipitation
- It is often assumed that native ChIP is unsuitable for profiling non-histone proteins owing to potential rearrangement.
  - Can be addressed by ORGANIC
- Solubility of proteins of interest can also be an issue with native ChIP
  - As harsh detergents and sonication are not used, recovery might be lower than for X–ChIP, especially for large complexes, in which case high-resolution X-ChIP is preferred

ChIP-exo	• Base-pair resolution • High input requirement
High-resolution X-ChIP–seq	• Base-pair resolution • Complete solubilization of certain chromatin proteins
ORGANIC	• Base-pair resolution • No crosslinking artefacts • Potentially poor solubilization of proteins

http://www.nature.com/nrg/journal/v15/n12/full/nrg3798.html

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

References

Lenhard B et al., Metazoan promoters: emerging characteristics and insights into transcriptional regulation, Nature Reviews Genetics 2012

Zentner GE, Henikoff S. High-resolution digital profiling of the epigenome. Nat Rev Genet. 2014

Hawkins RD1, Hon GC, Ren B. Next-generation genomics: an integrative approach. Nat Rev Genet. 2010

Rozowsky J1 et al. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. Nat Biotechnol. 2009

Sebastian Schmeier
s.schmeier@gmail.com
http://sschmeier.github.io/bioinf-workshop/

THE ENGINE
OF THE NEW
NEW ZEALAND