Identify functional patterns in high throughput binding assays

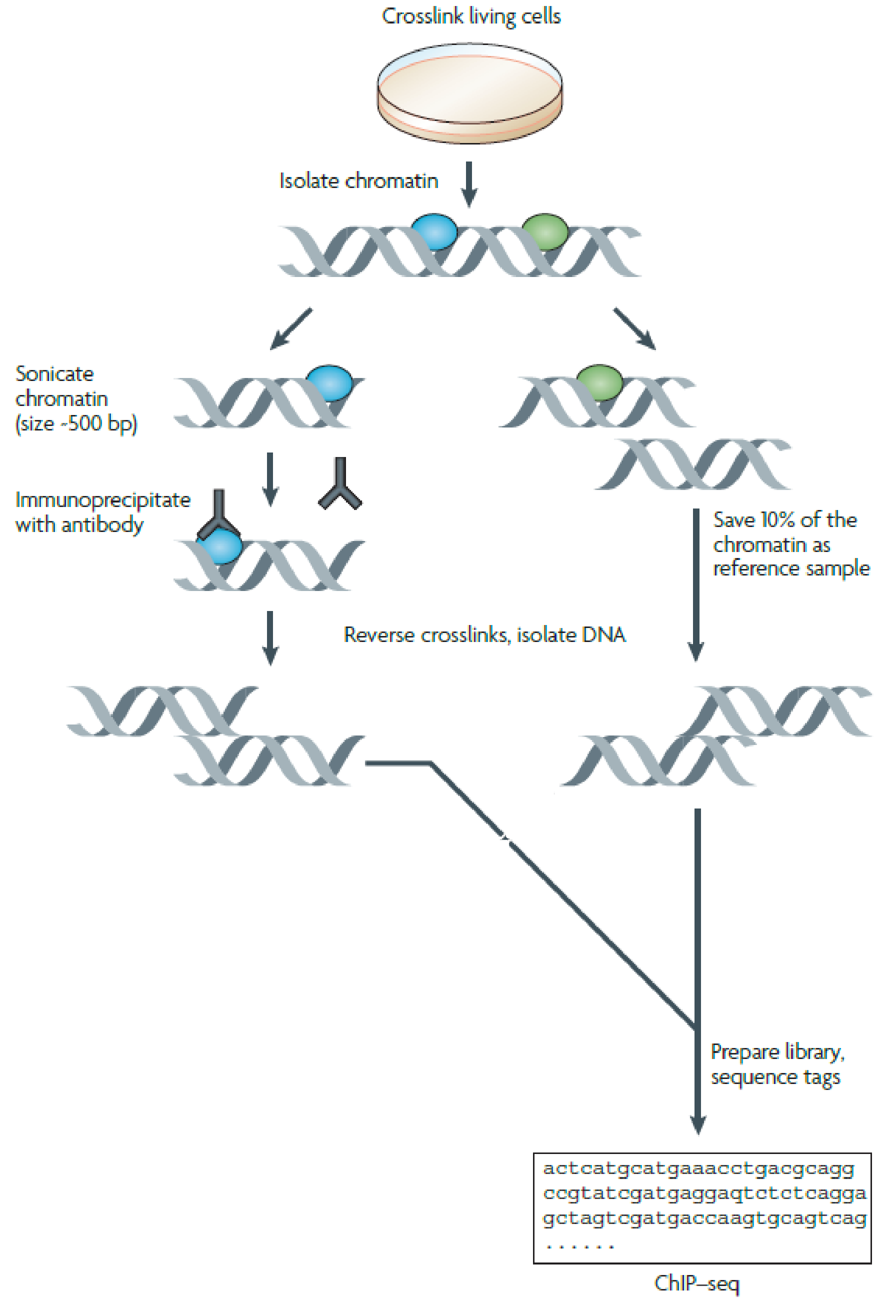
Alex Essebier

Message

- By clustering ChIP-seq peaks we can identify different patterns in transcription factor binding

ChIP-seq experiment

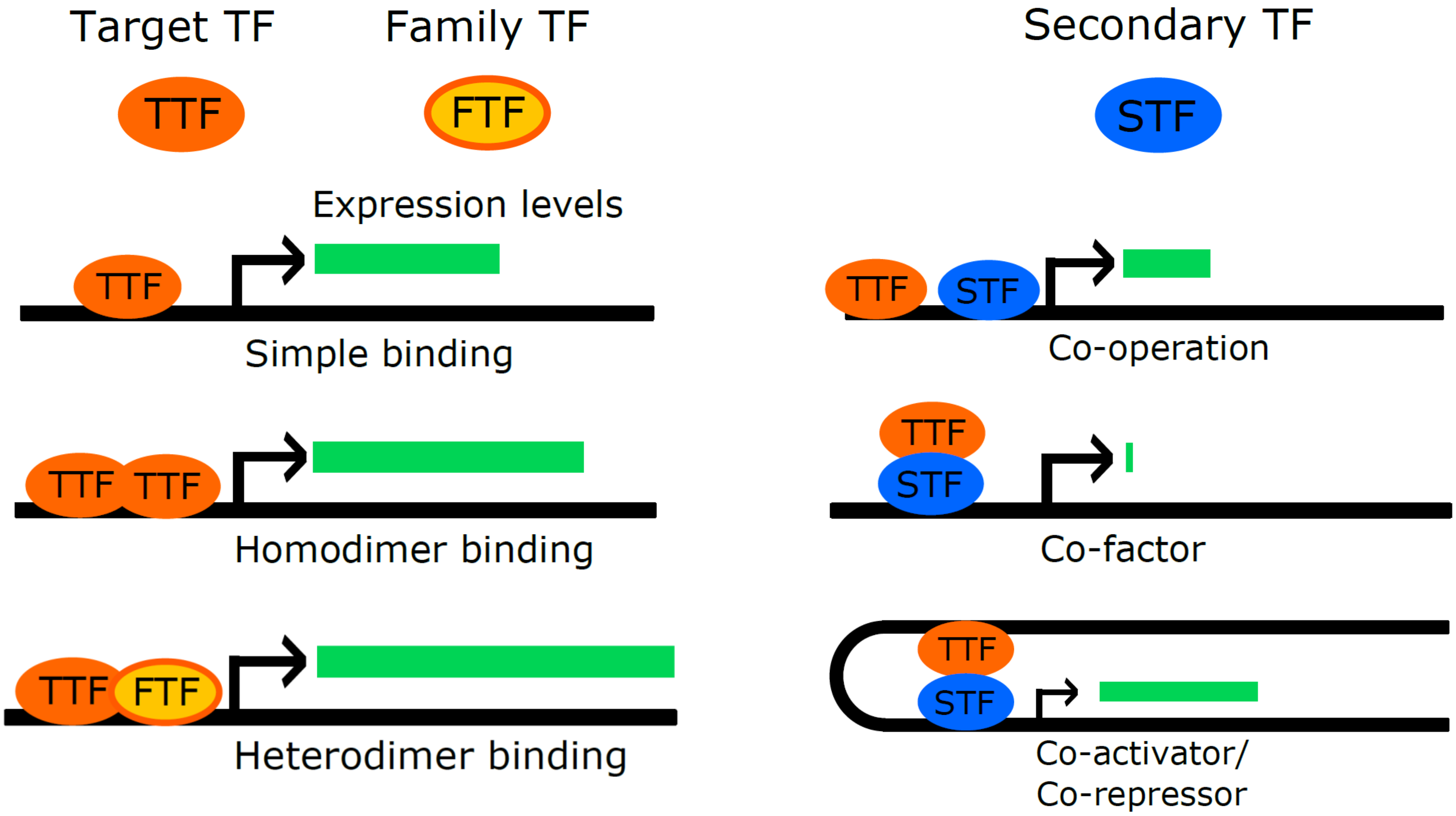
- Chromatin immunoprecipitation followed by sequencing
- To determine where a protein binds the genome
- e.g. for a **single TF** or histone modification


Crosslink living cells
Isolate chromatin
Sonicate chromatin (size ~500 bp)
Immunoprecipitate with antibody
Save 10% of the chromatin as reference sample
Reverse crosslinks, isolate DNA
Prepare library, sequence tags
actcatgcatgaaacctgacgcagg
ccgtatcgatgaggaqtctctcagga
gctagtcgatgaccaagtgcagtcag
......
ChIP–seq

Farnham, P.J. *Nature Reviews Genetics* **10**, 605-616 (September 2009)

Why TFs?

- Important role in gene expression, cell differentiation and homeostasis



Target TF

Family TF

Secondary TF

Expression levels

Simple binding

Homodimer binding

Heterodimer binding

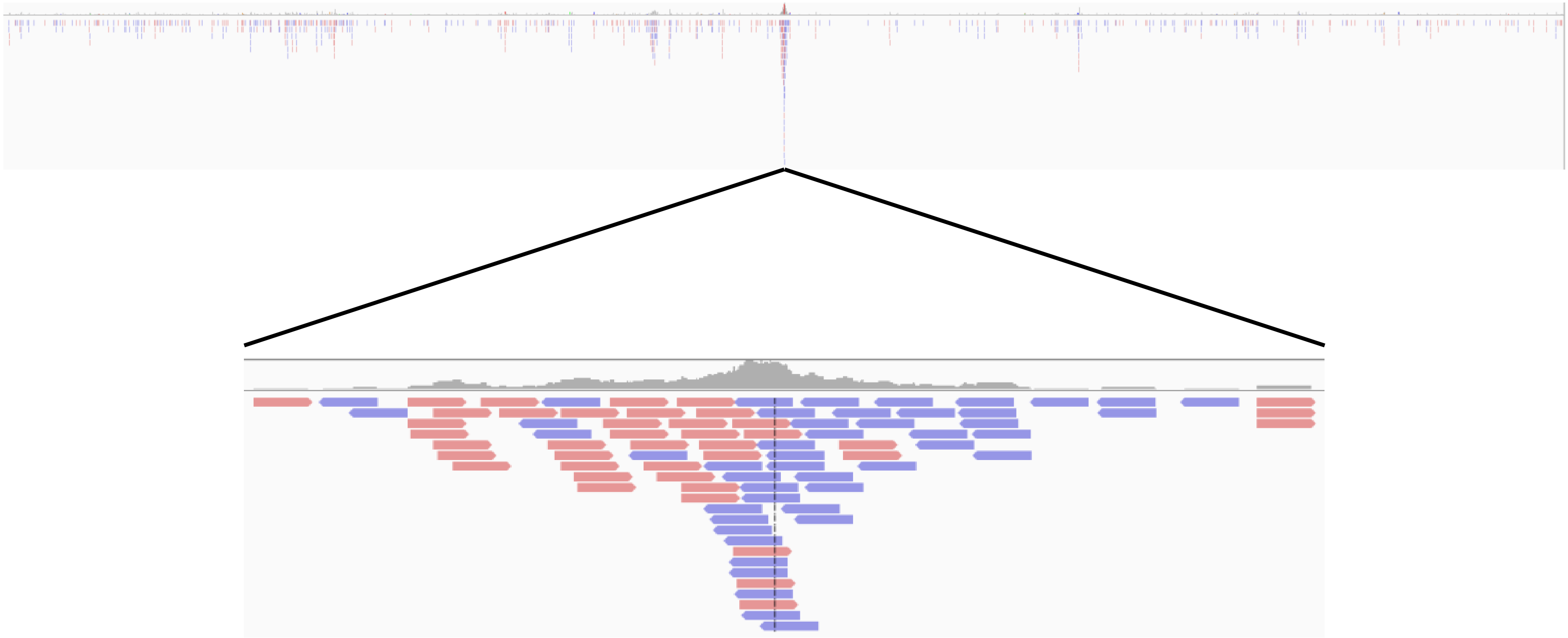
Co-operation

Co-factor

Co-activator/
Co-repressor

Peak calling

- Raw sequencing data
  - Single end reads
  - Red mapped ‘forward’, blue mapped ‘reverse’
  - Distribution across genome

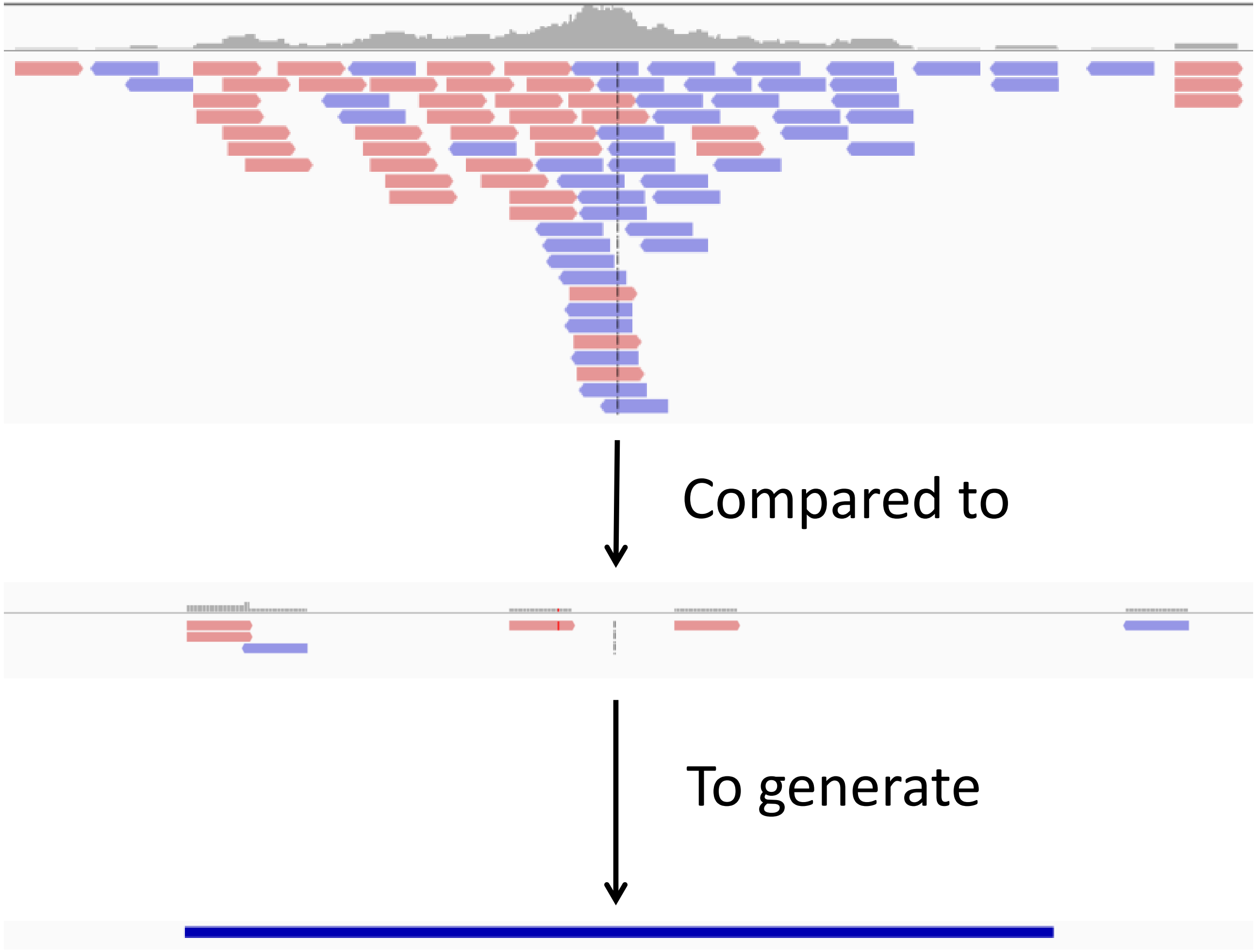


Peak calling

Sample –
exposed to
antibody

Input –
no antibody
exposure

Peak –
with statistical
significance

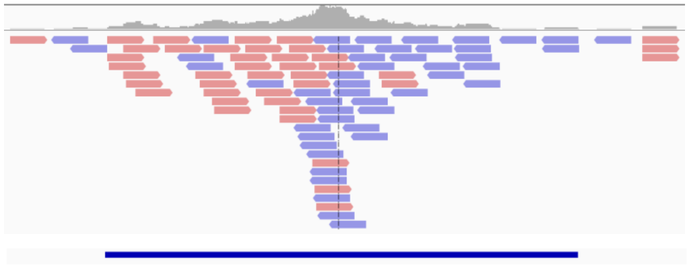


Compared to

To generate

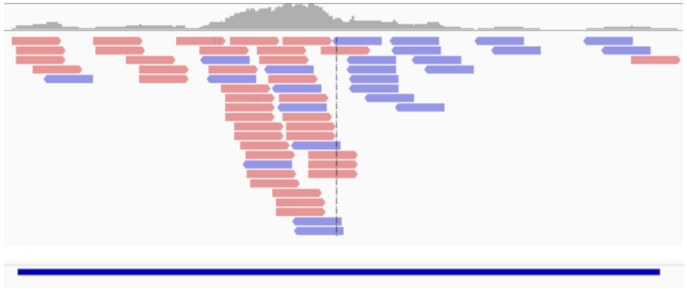
Peak calling

A



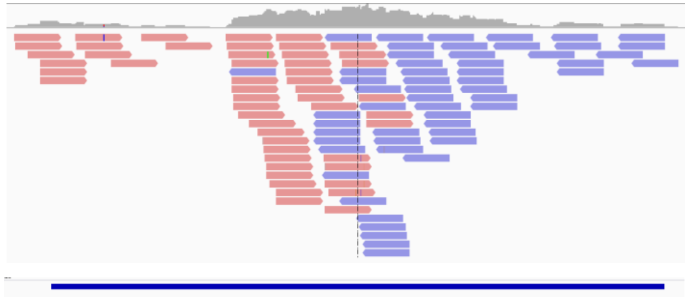
ATTGCC

B



ATTTCC

C

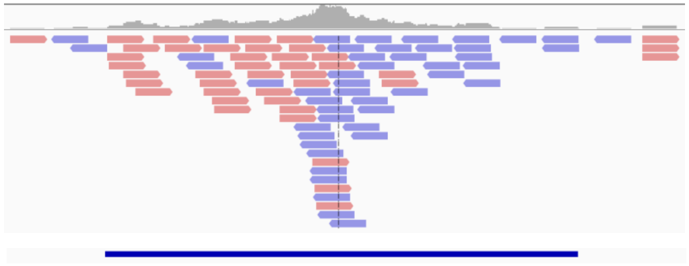


ATACCC

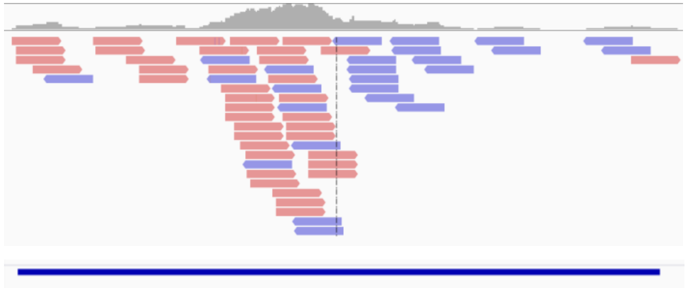
- Peaks have different features within a ChIP-seq experiment

Peak	Width	Enrich-ment	Location	Epigenetic Environment
A	205	298	Distal Intergenic	Insulator
B	162	218	Promoter	Active Promoter
C	194	361	Promoter	Weak Promoter

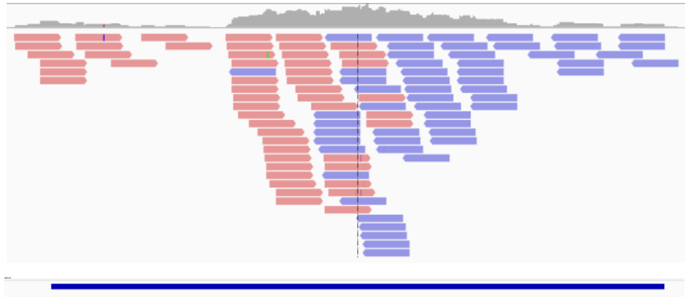
Peak calling



ATTGCC



ATTTCC

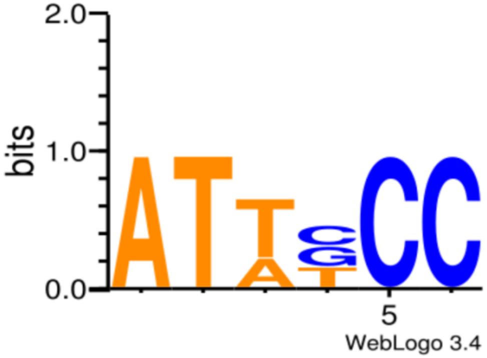


ATACCC

Analyse

Confirm *in vitro* results

Identify consensus motif



Identify target genes of TF

Hypothesis

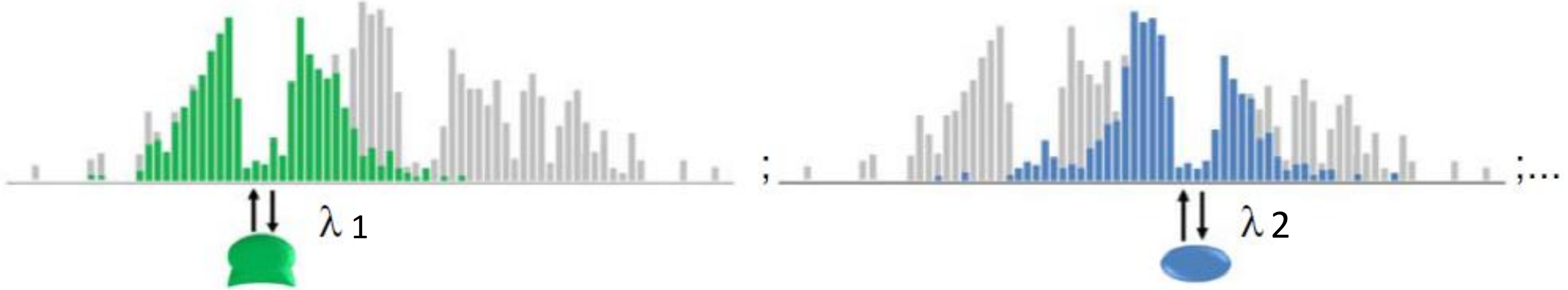
- We propose that ChIP-seq peaks from a TF experiment can be clustered based on their read density or ‘shape’ leading to identification of different binding modes and functional patterns of a TF

Previous use of peak shape

- Differential binding
  - Compare two conditions
  - Compare two TFs
  - Based on read depth

- TF binding from DNase I hypersensitivity

TF binding estimation from modelled DNase I hypersensitivity profiles



Adapted from: Sherwood, R.I., et al. *Nature Biotechnology* **32(2),**171-178 (2014)

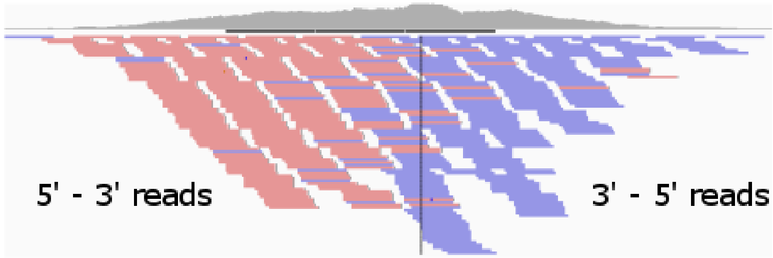
Aims

- Develop a modelling technique to identify functionally relevant clusters, based on ChIP-seq read density, defining TF binding events

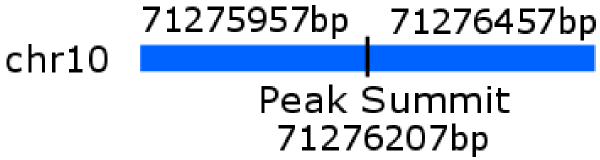
- Identify functional patterns associated with clusters and provide more information about TF binding from ChIP-seq data

Processing peak data

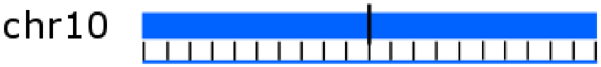
Binding Site aka Peak



Read pile up showing distribution of reads in peak



500bp window around summit



Split window into even segments
Count read depth in each segment

Peak1	5	8	14	19	20	25	30	27	28	23	18	15	6	2

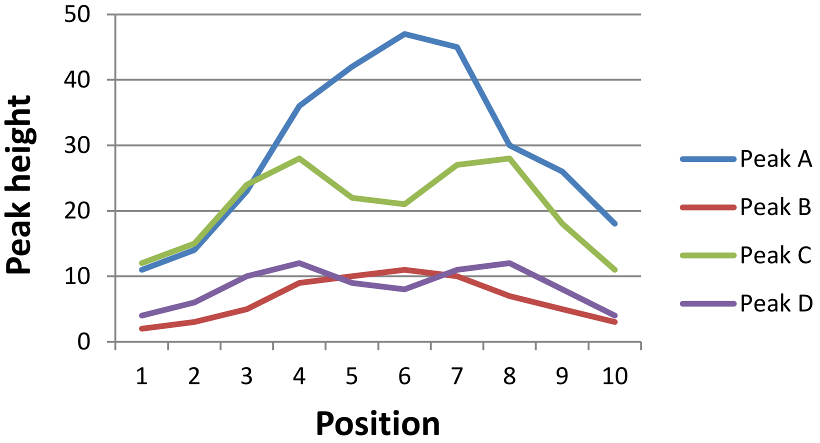
This creates a count vector for each peak with equal numbers of columns

Combining all count vectors creates a **Dirichlet distribution** that can be clustered

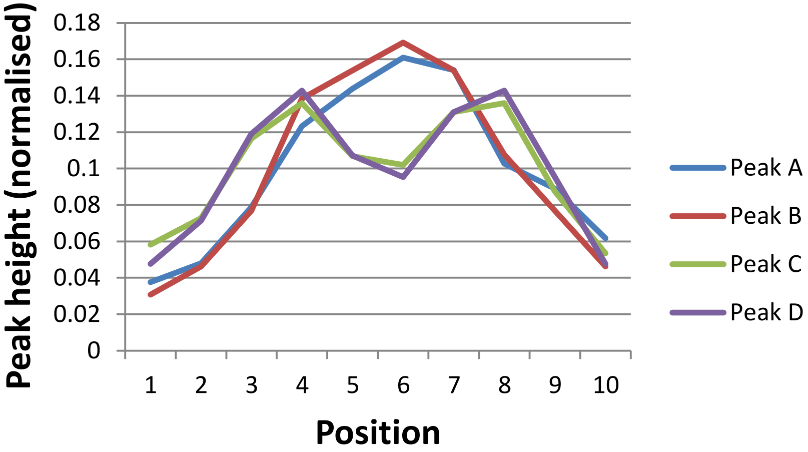
Dirichlet clustering

- Dirichlet distribution – distribution of distributions
- The model is a Dirichlet mixture
- Unsupervised clustering of peaks
- Evidence based clustering using raw counts
- **No normalisation** of data

Evidence based clustering



Shape of read counts



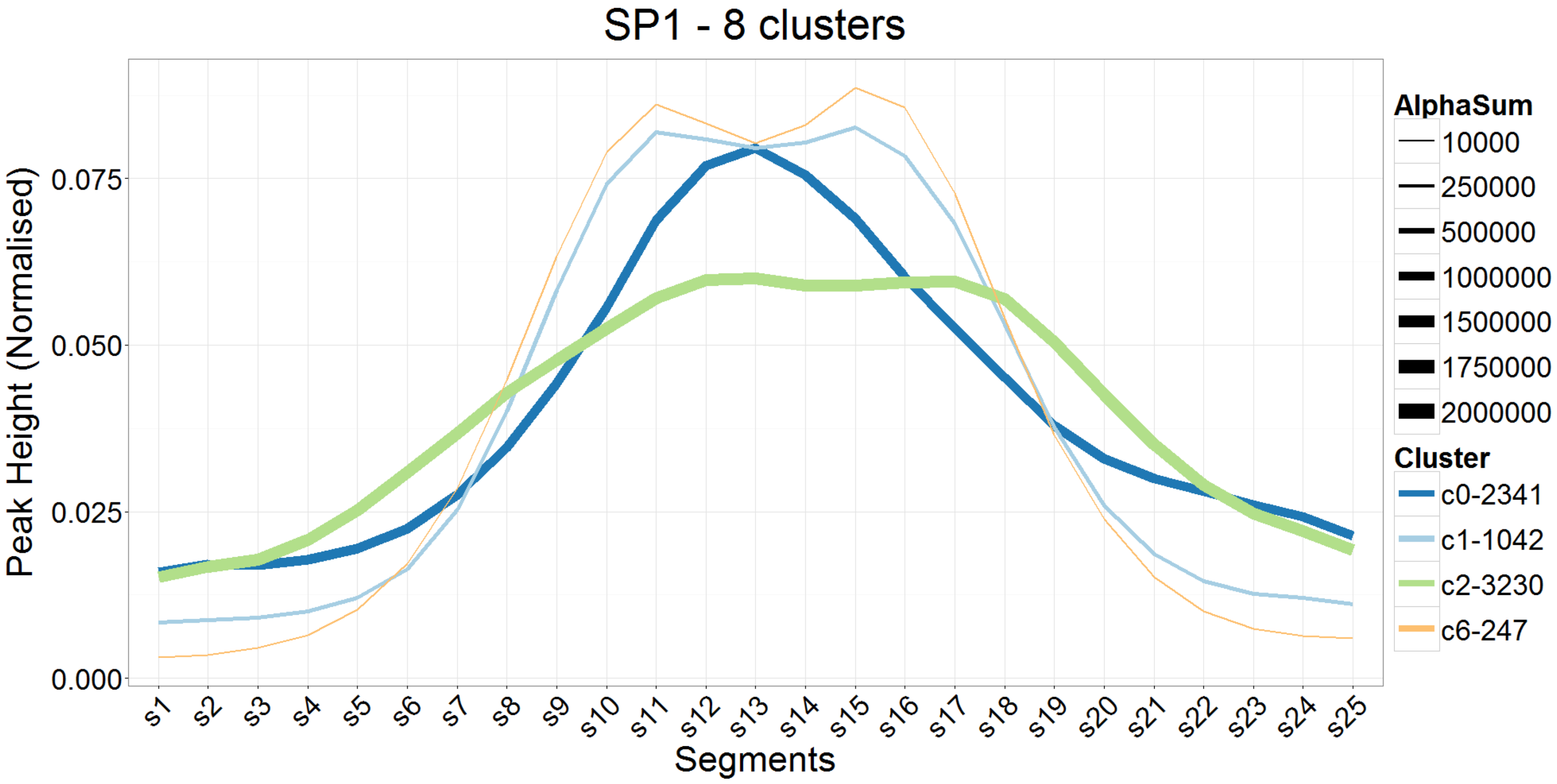
Shape of normalised read counts

K-means	
Peak	Cluster
A	1
B	1
C	2
D	2

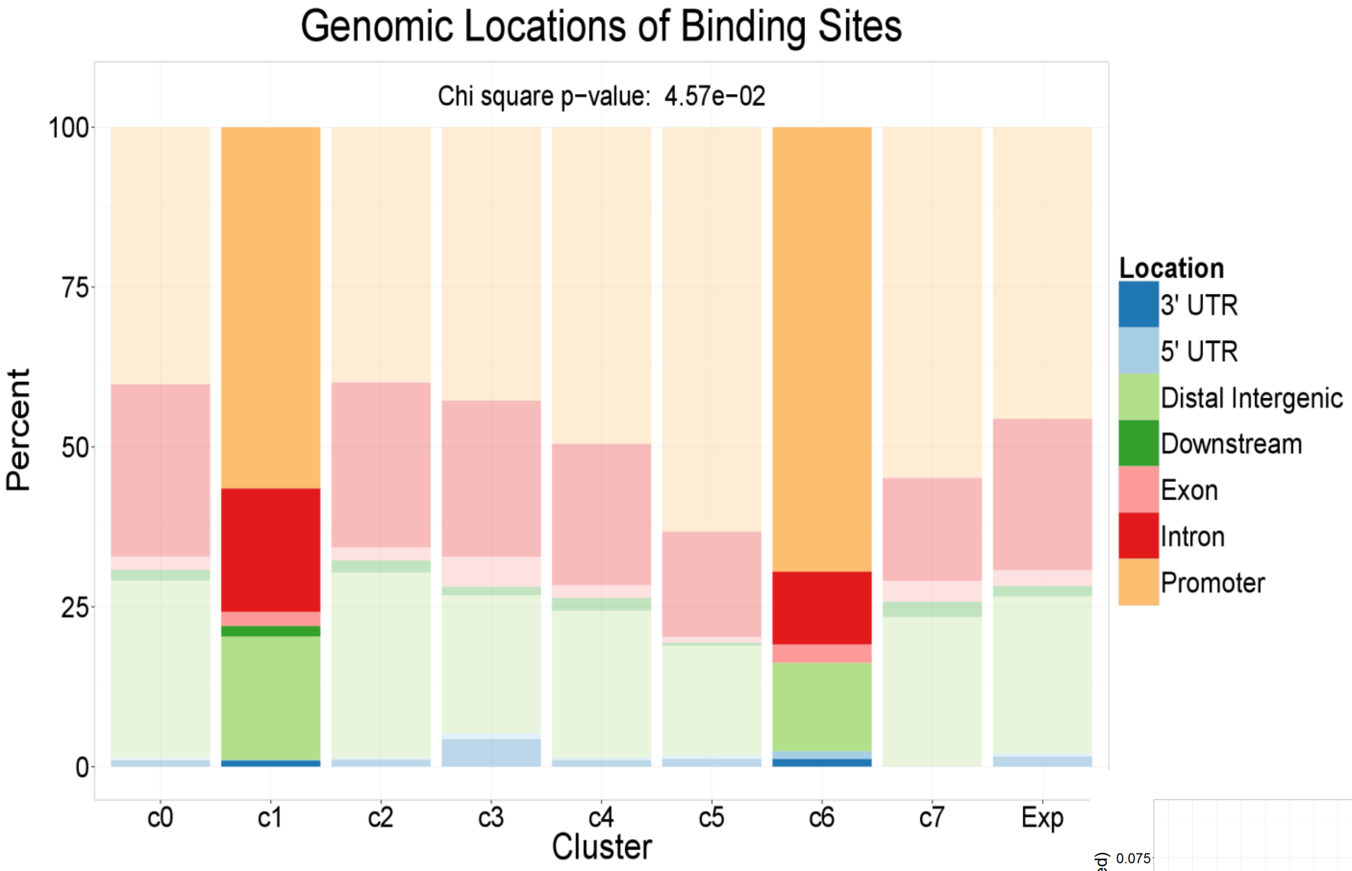
Dirichlet	
Peak	Cluster
A	1
B	2
C	3
D	4

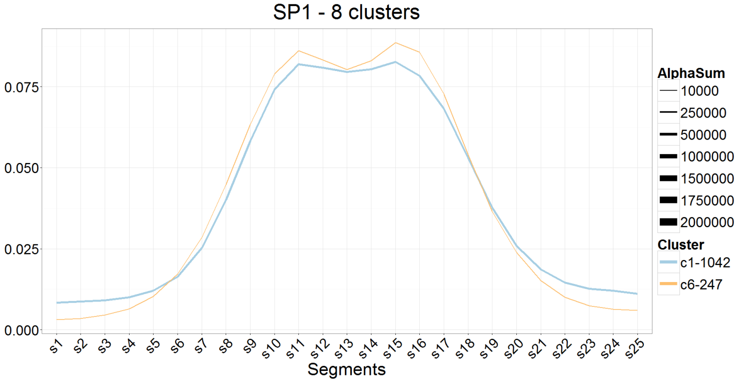
- Read depth is key and can be masked by normalisation
- Dirichlet approach does not require normalisation

Clustering example – SP1

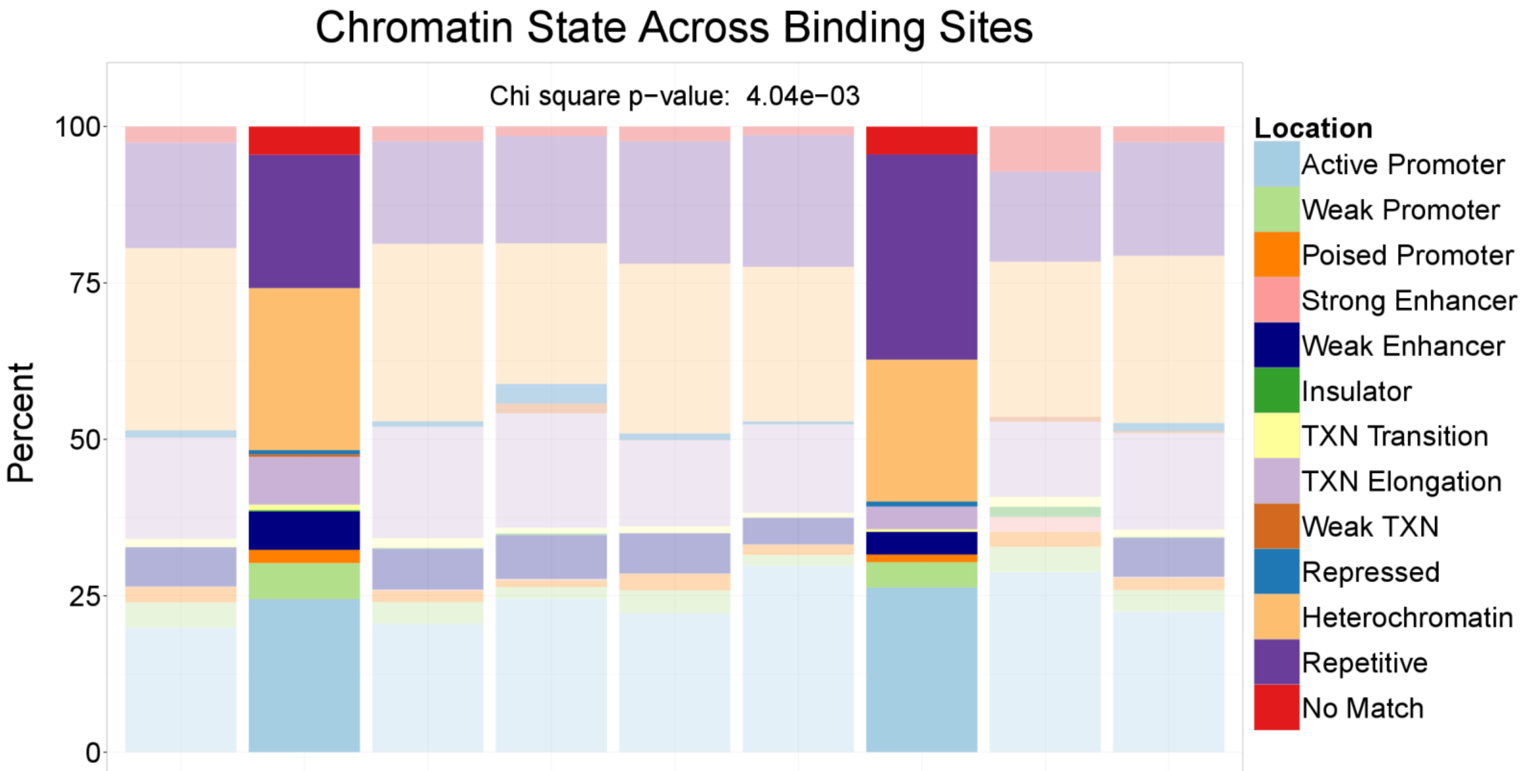


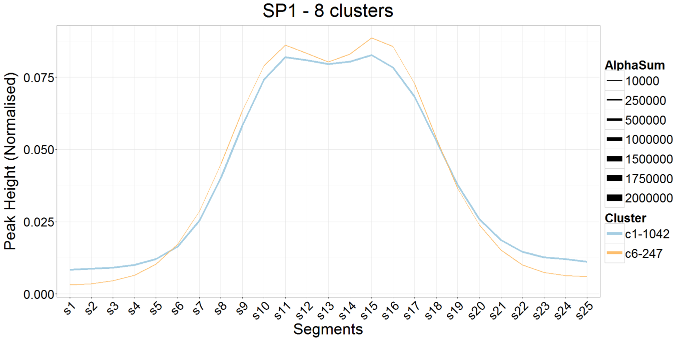
Genomic location



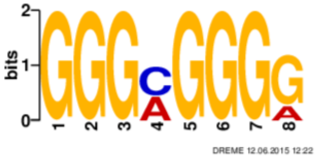
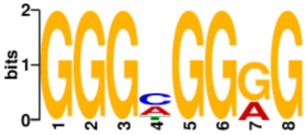


Epigenetic environment

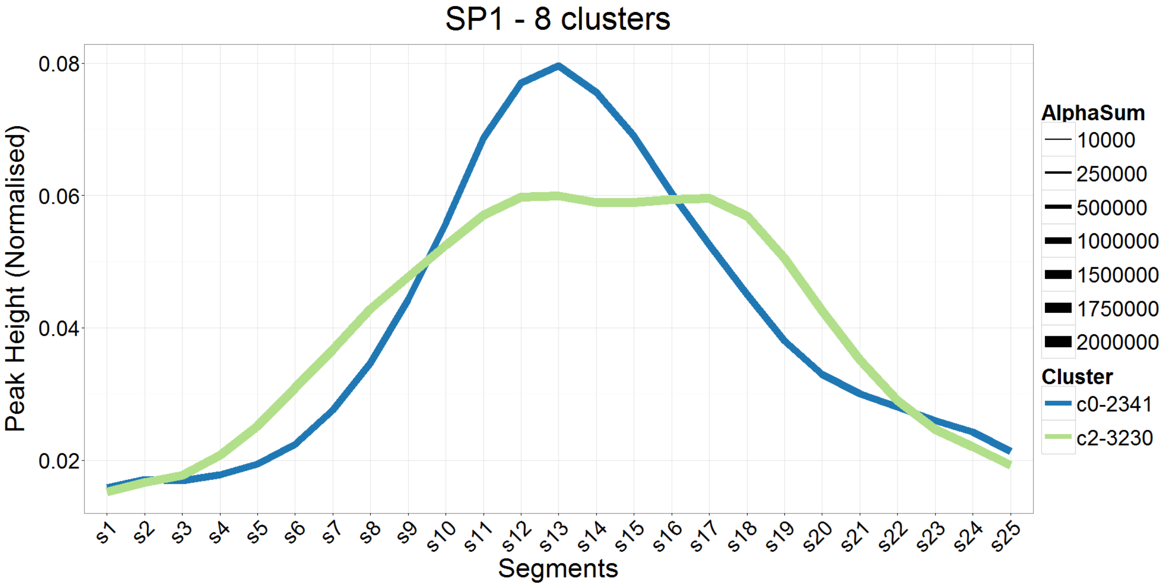




Consensus motifs

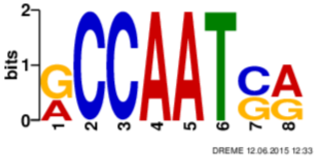
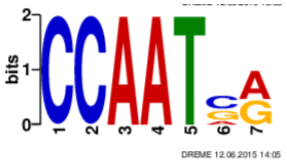
Cluster	Motif
0	
2	

- SP1 motif
- Differentiating feature in c0 and c2 is binding affinity or read depth

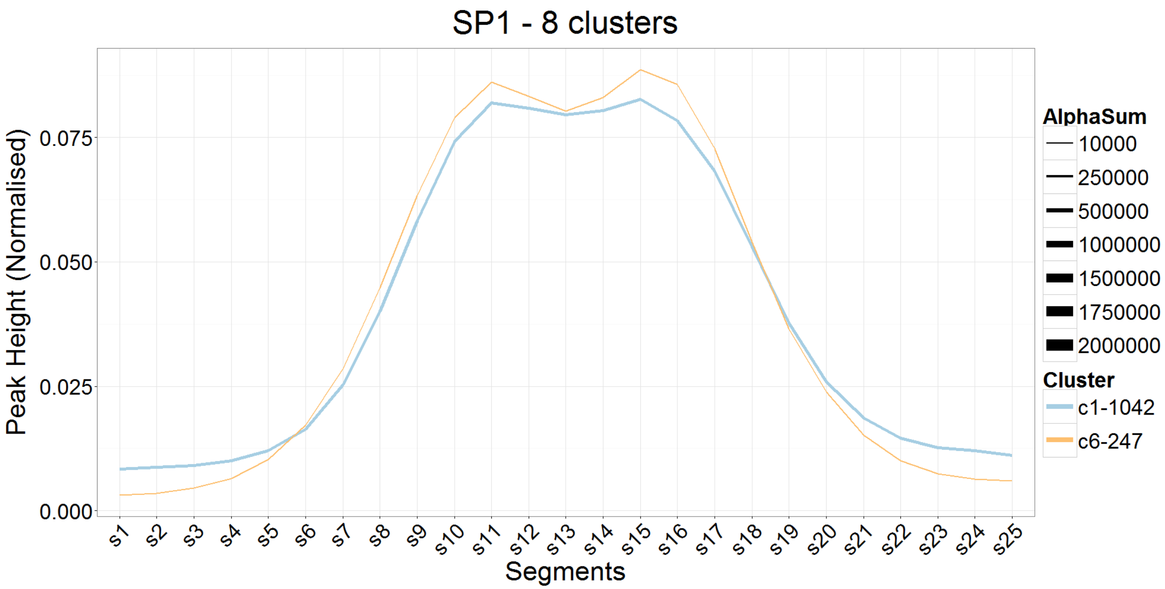


SP1 - 8 clusters

Consensus motifs

Cluster	Motif
1	 bits 2 1 0 G A CCAAT CA GG 1 2 3 4 5 6 7 8 DREME 12.06.2015 12:33
6	 bits 2 1 0 CCAAT C G A G 1 2 3 4 5 6 7 DREME 12.06.2015 14:05

- NFY motif
- Known interaction between two TFs
- A bimodal peak shape indicates increased NFY binding



Applications

- Explore TF families by comparing clustering outcomes
- Explore TF dimers using clustering in combination with *in vitro* sequence data
- Explore cooperative interactions

Summary

- We successfully clustered ChIP-seq peaks based on their shape, density and magnitude then demonstrated how each cluster contains unique, biologically relevant, features

Thanks

Supervisor

Mikael Bodén

Bodén Group

Ralph Patrick

Tim O’Connor

Julian Zaugg

Gabe Foley

Piper Group

Michael Piper

Rostlab (TUM)

Burkhard Rost

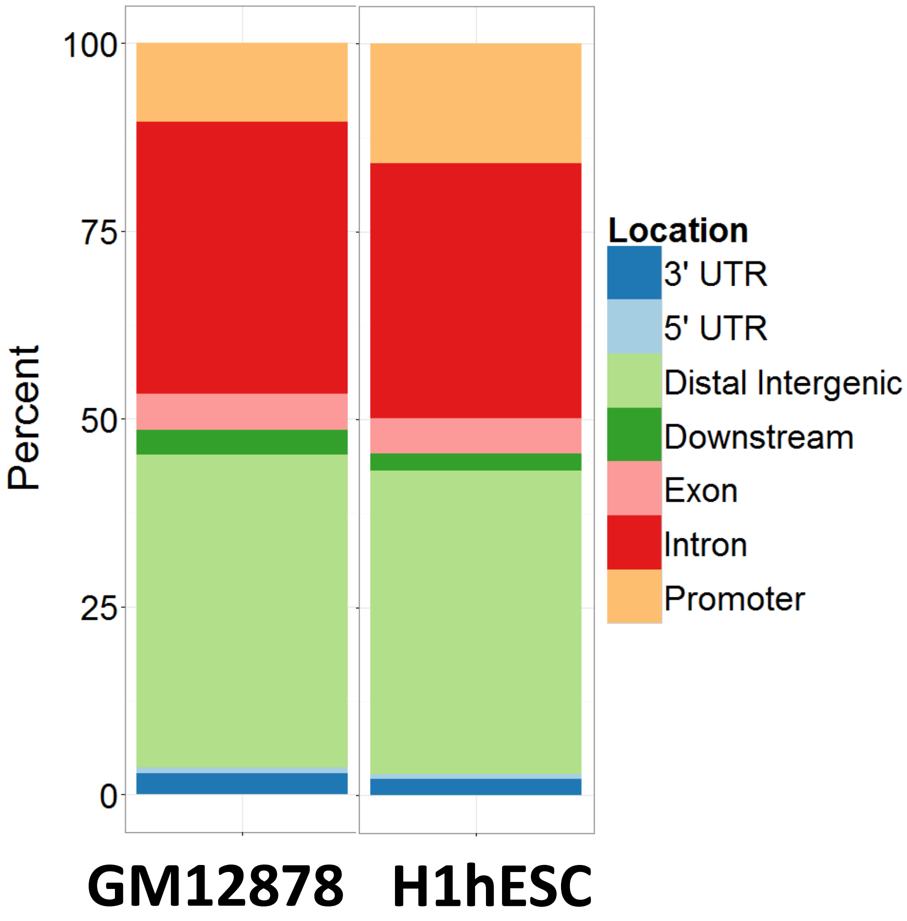
Tatyana Goldberg

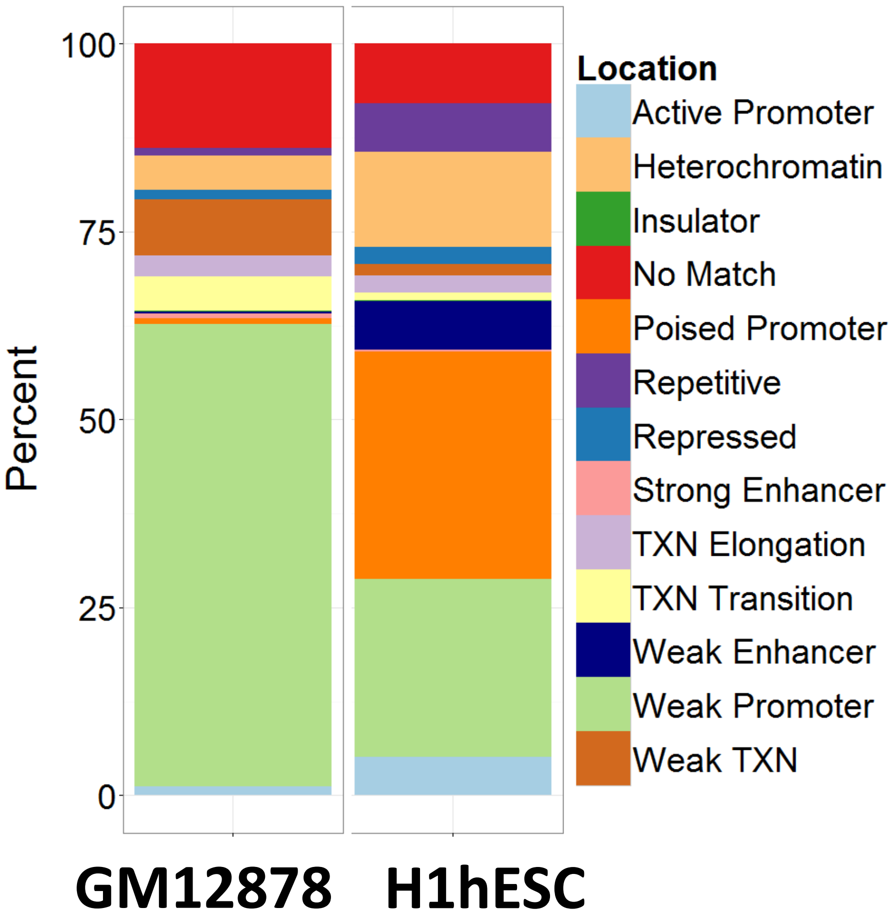
THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

qbi queensland brain institute

ROSTLAB.

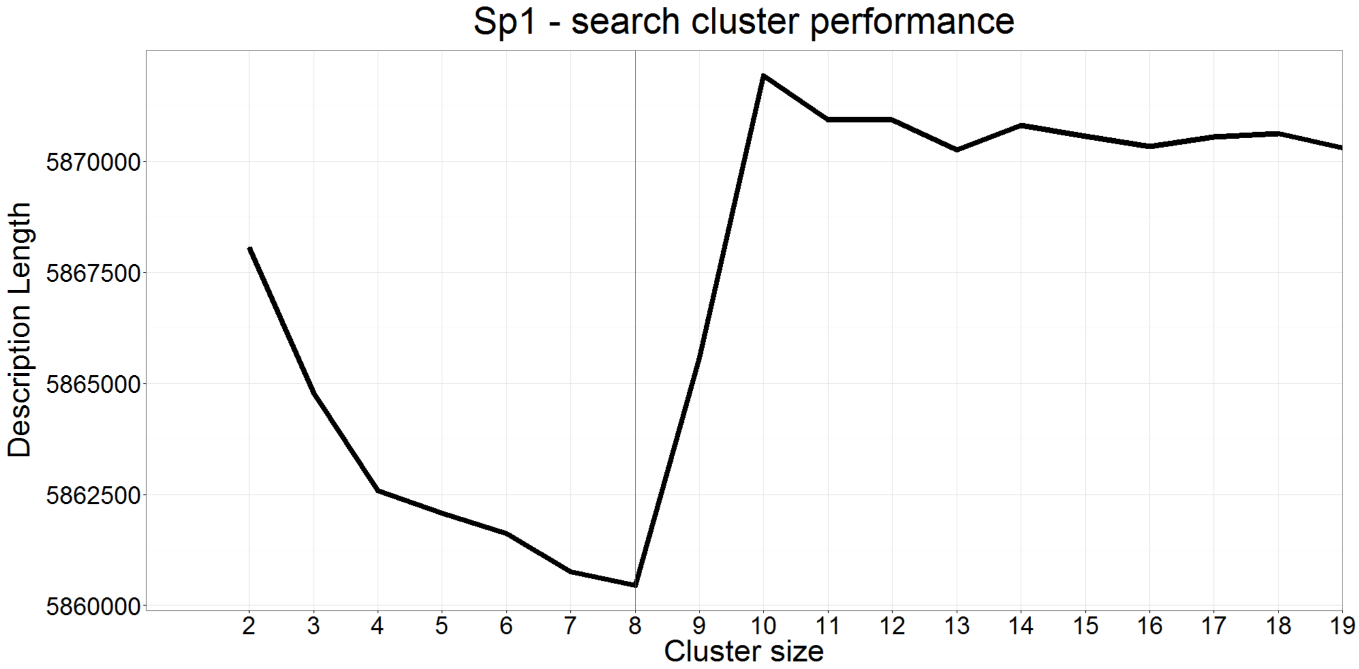
Cell type

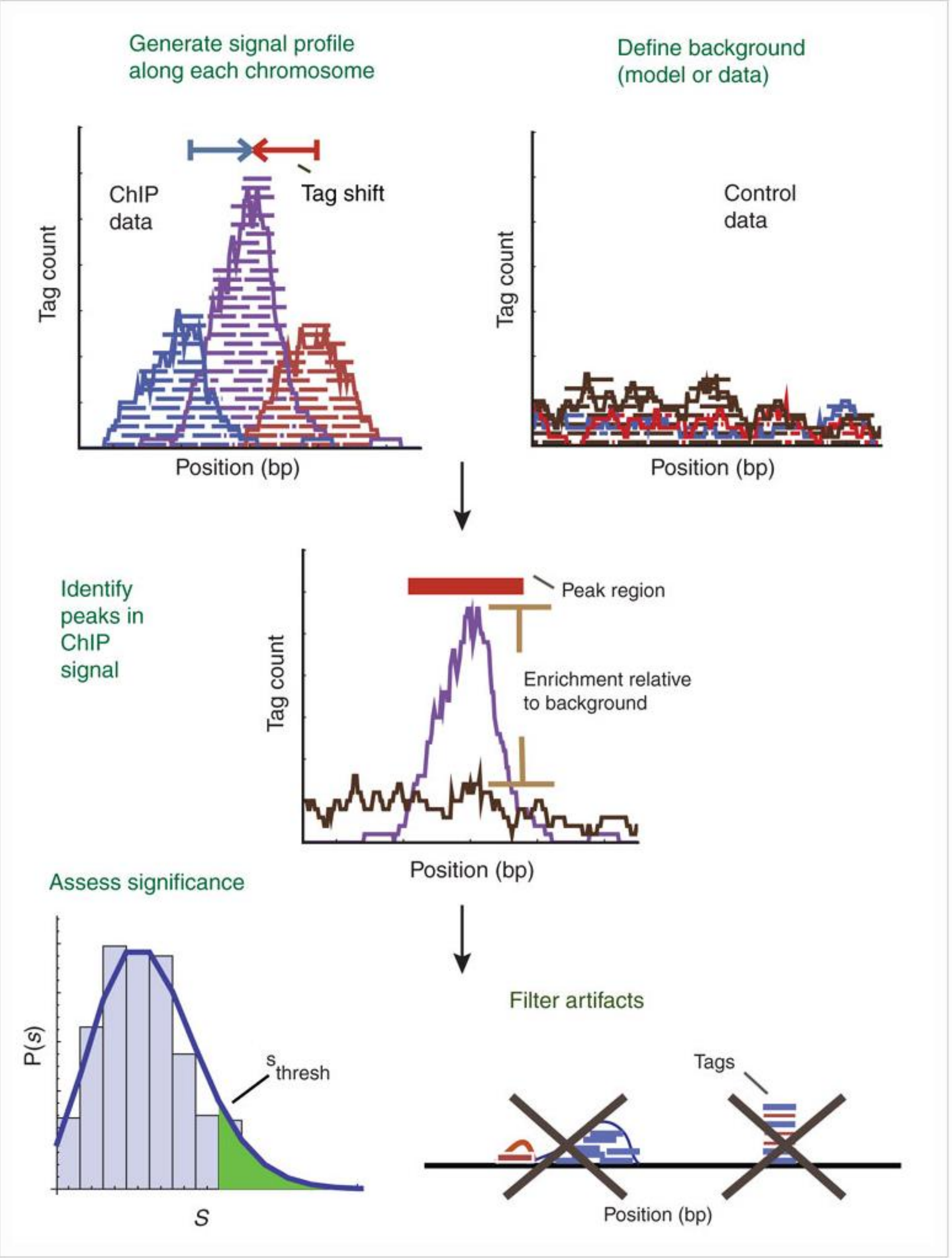

Percent
100
75
50
25
0
Location
3' UTR
5' UTR
Distal Intergenic
Downstream
Exon
Intron
Promoter
GM12878
H1hESC


Percent
100
75
50
25
0
Location
Active Promoter
Heterochromatin
Insulator
No Match
Poised Promoter
Repetitive
Repressed
Strong Enhancer
TXN Elongation
TXN Transition
Weak Enhancer
Weak Promoter
Weak TXN
GM12878
H1hESC

Cluster size optimisation

- Minimum description length (MDL)
- Description length (DL)
  - A measure of information content and model complexity
- Larger models will always be more complex



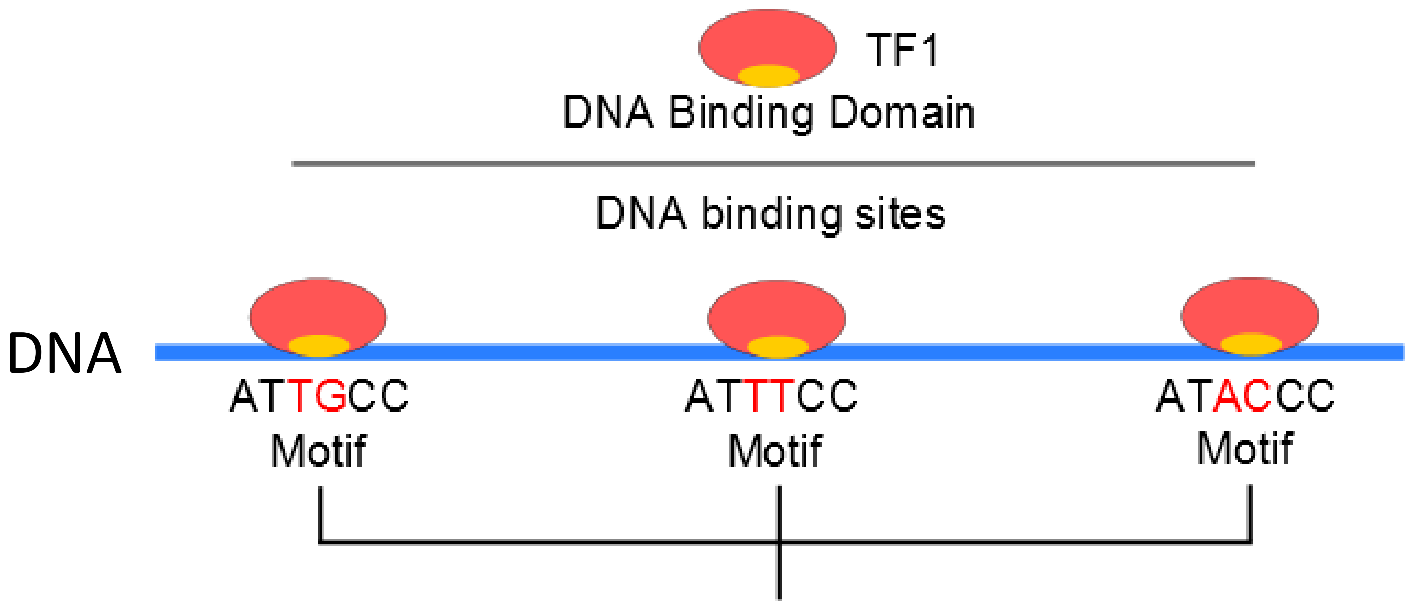


ChIP-seq peak calling

- Shift reads on both strands to find peak
- Compare to control reads
- Identify significant hits according to a threshold
- Remove potential artefacts

Computation for ChIP-seq and RNA-seq studies
Shirley Pepke, Barbara Wold & Ali Mortazavi
Nature Methods **6**, S22 - S32 (2009) Published online: 15 October 2009
doi:10.1038/nmeth.1371

DNA binding sites



TF1

DNA Binding Domain

DNA binding sites

DNA

ATTGCC
Motif

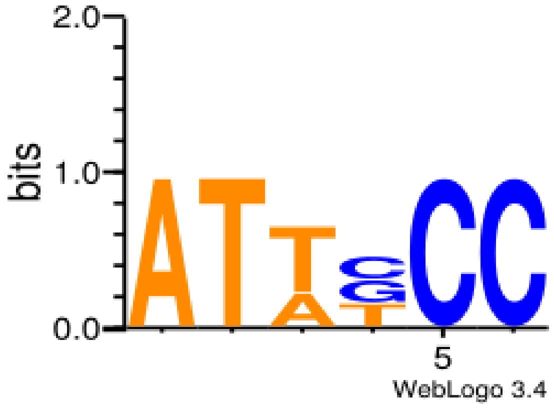
ATTTCC
Motif

ATACCC
Motif

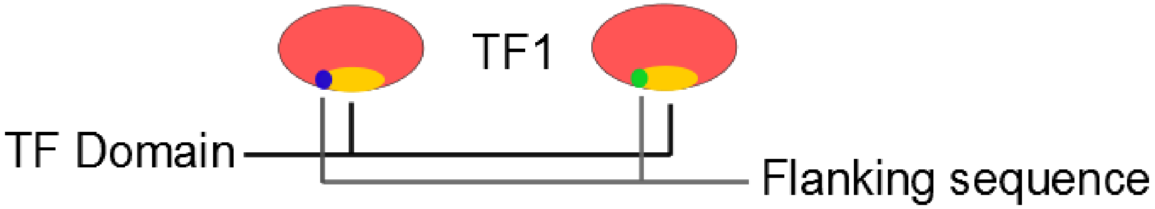
Position Weight Matrix (PWM)

Position	A	T	C	G
1	1	0	0	0
2	0	1	0	0
3	0.33	0.66	0	0
4	0	0.33	0.33	0.33
5	0	0	1	0
6	0	0	1	0

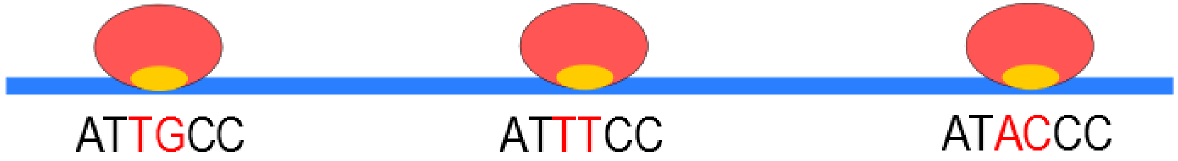
Consensus Motif



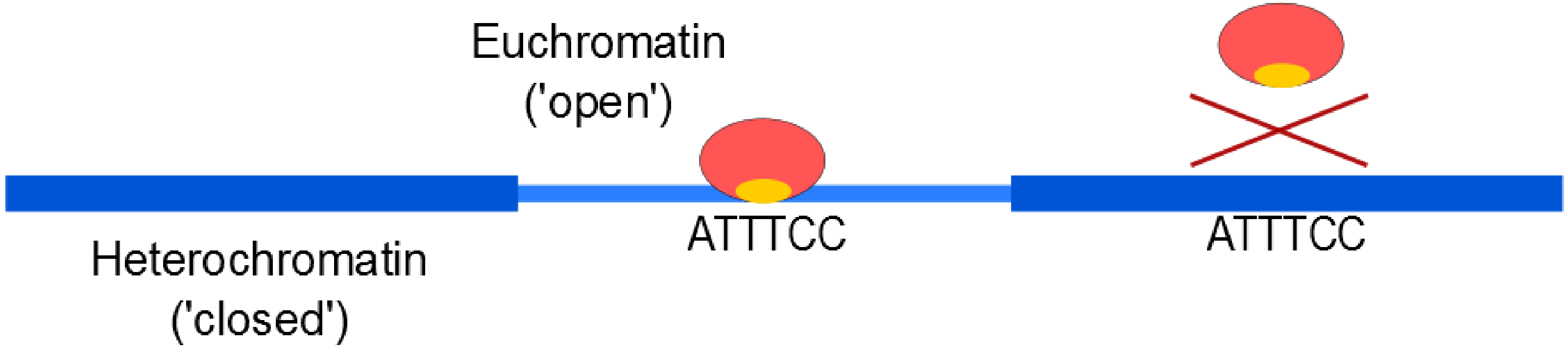
Variation in flanking sequence of TF domains



Sequence polymorphism in target DNA



Site accessibility within chromatin landscape



Minimum description length (MDL) Principle

- Calculate model complexity
- Calculate smallest data description length (DDL)
- Total DL = sum of complexity and DDL
- Plot total DL as number of clusters increases and search for global minima
- Global minima = optimal number of clusters