

Next Generation Sequencing Data Analysis:

From ChIP-Seq Read Islands to Epigenomics Information

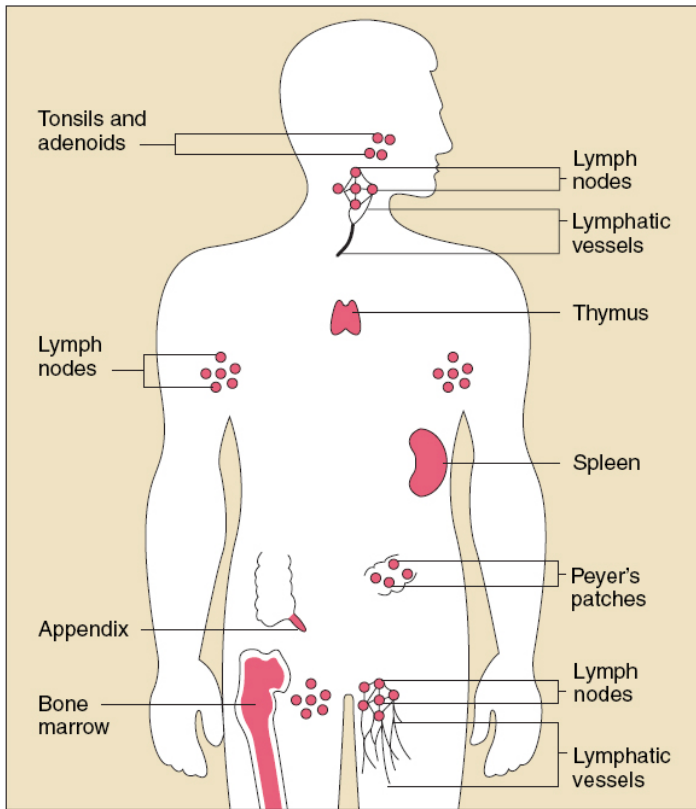
WBS Sommer Seminar, July 4th 2013

M. Jaritz, IMP

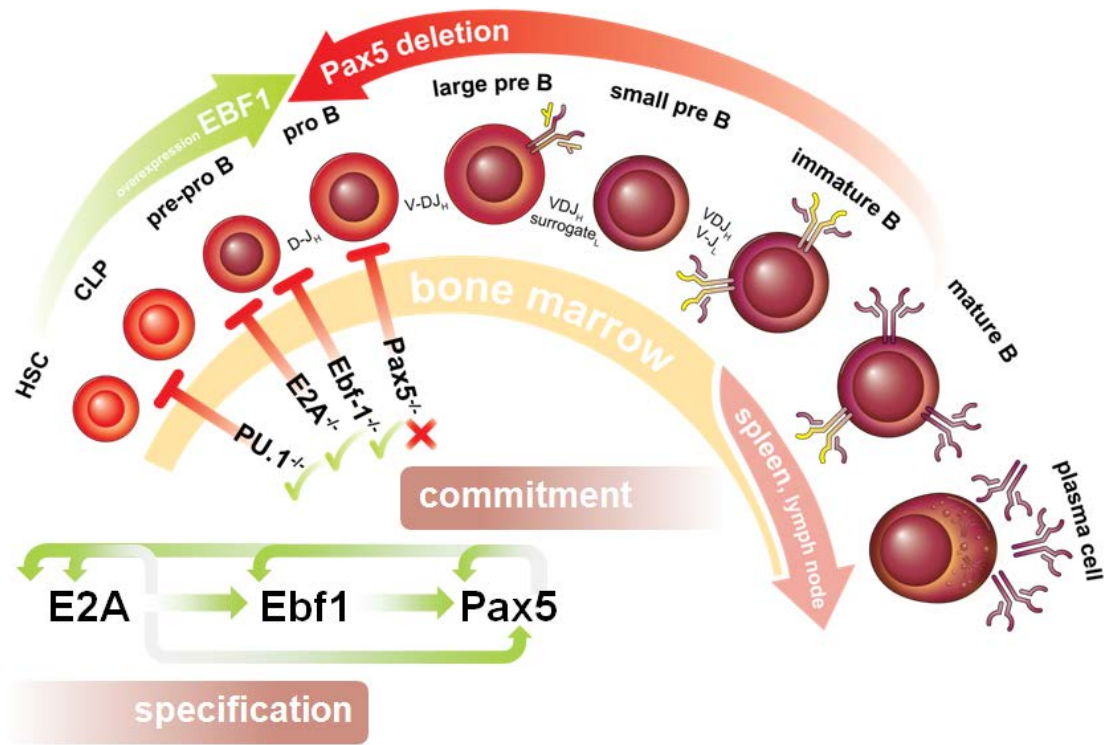
Outline

- Brief intro to experimental techniques
- Next generation sequencing raw results
- Track quality measurements
- Read densities, heatmaps & profiles
- Peak calling & overlaps
- Motif finding

Immune System & B cell development



<http://www.niaid.nih.gov/>



by Bojan Vilagos



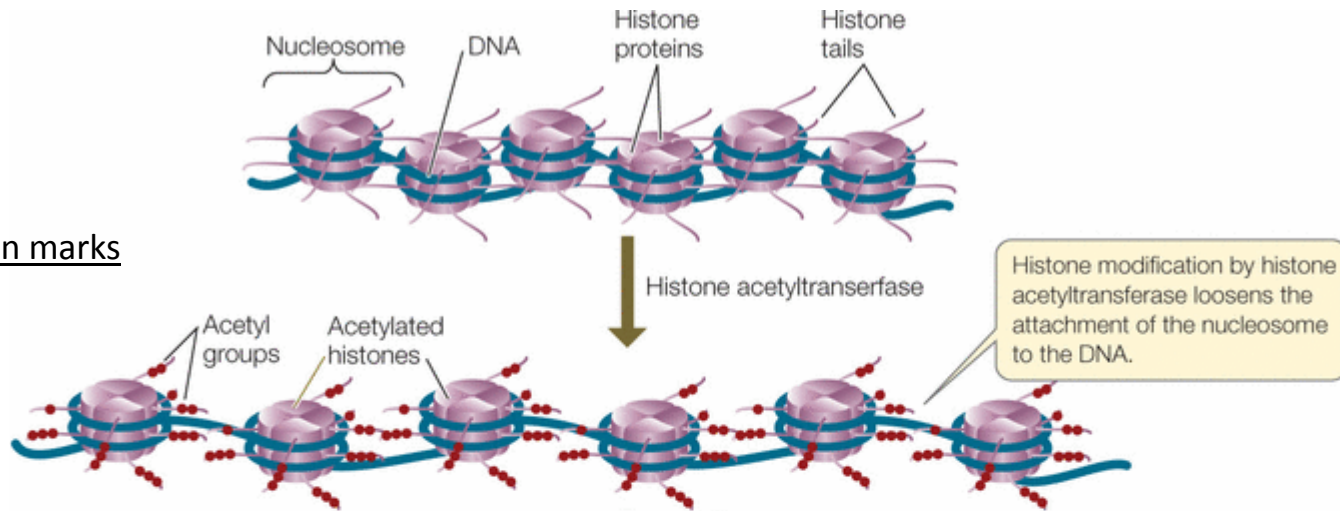
+ site specific recombinase technology (Cre-Lox system) -> KO

Questions: For each cell stage, ...

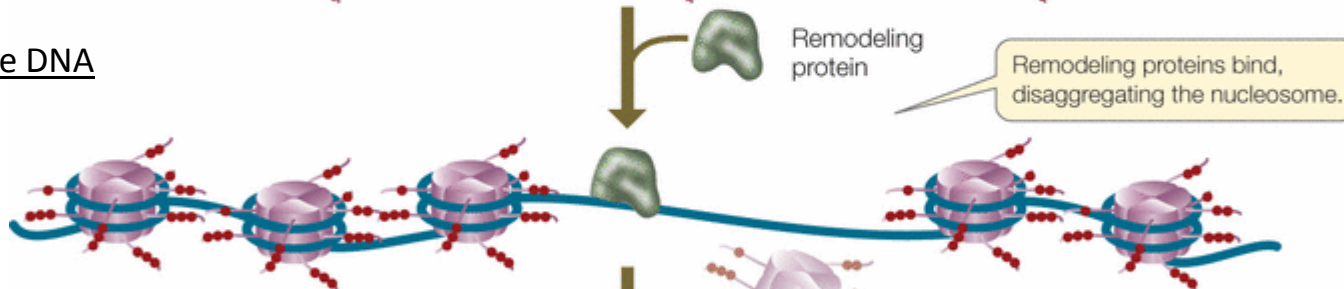
- ... which chromatin marks are associated with each gene?
- ... is the gene's DNA accessible to interacting proteins?
- ... which gene is bound by a transcription factor?
- ... which gene is expressed?
- ... how is the correlation between transcription factor binding, transcription and chromatin state?
- ... how does all this change between cell stages?

Gene expression

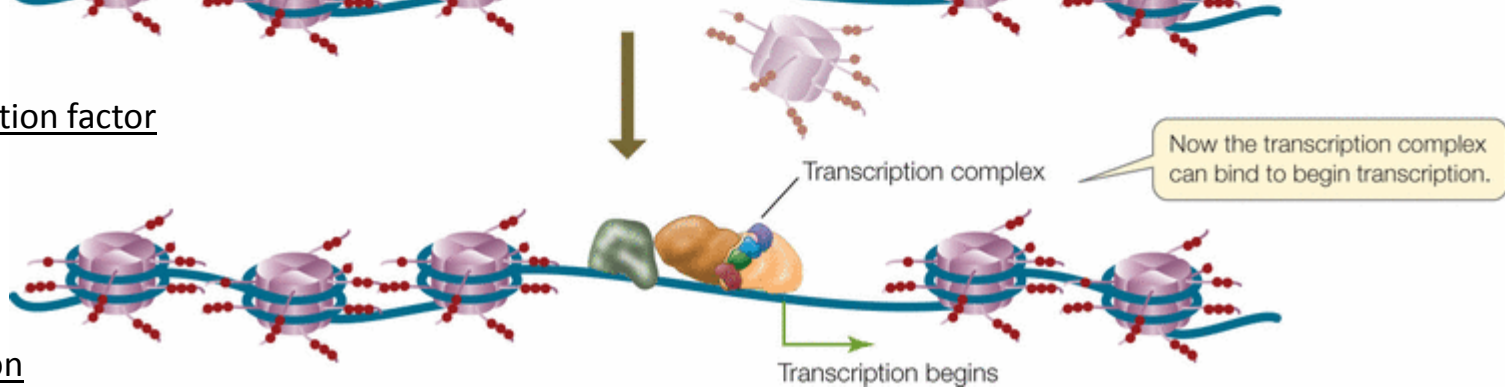
Chromatin marks



Accessible DNA



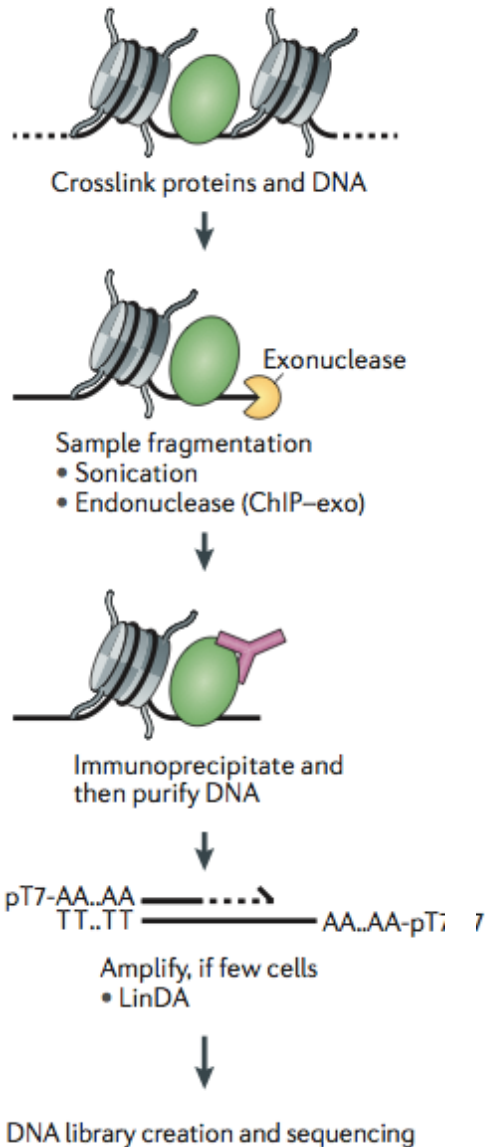
Transcription factor



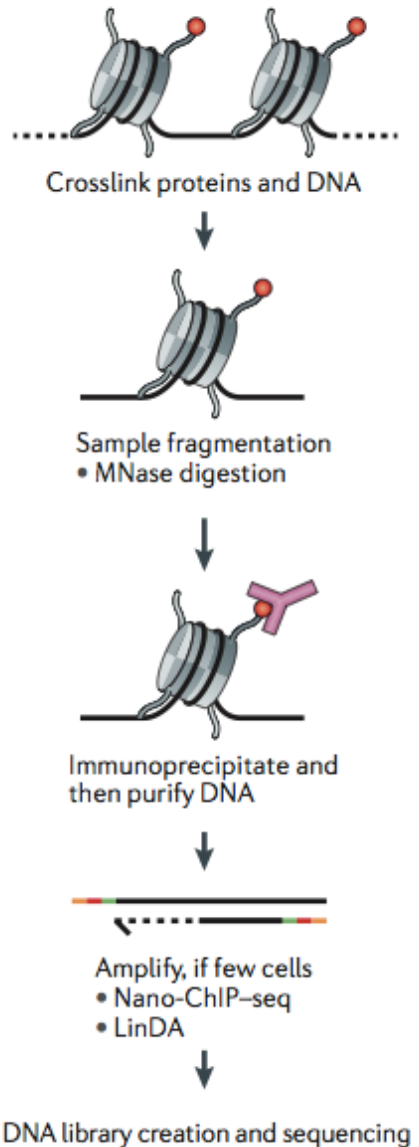
Expression

Sequencing experiment protocols

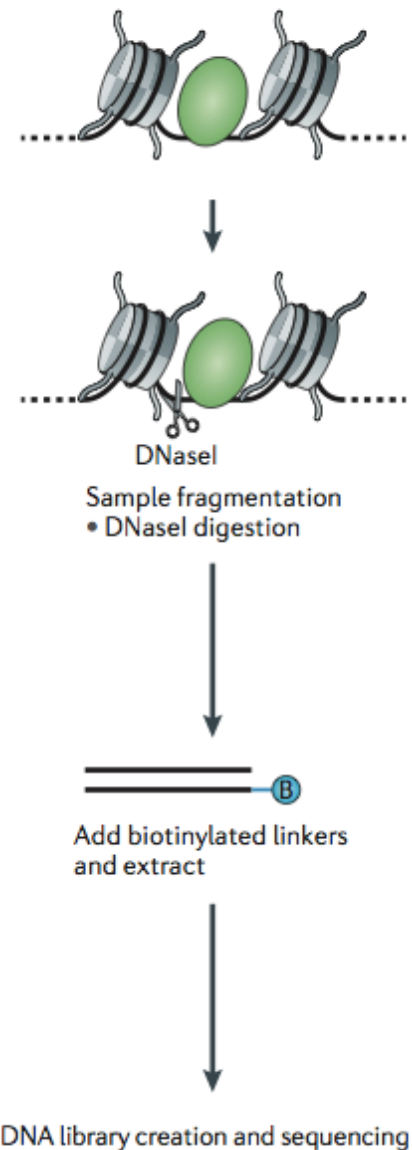
a DNA-binding protein CHIP-seq



b Histone modification CHIP-seq

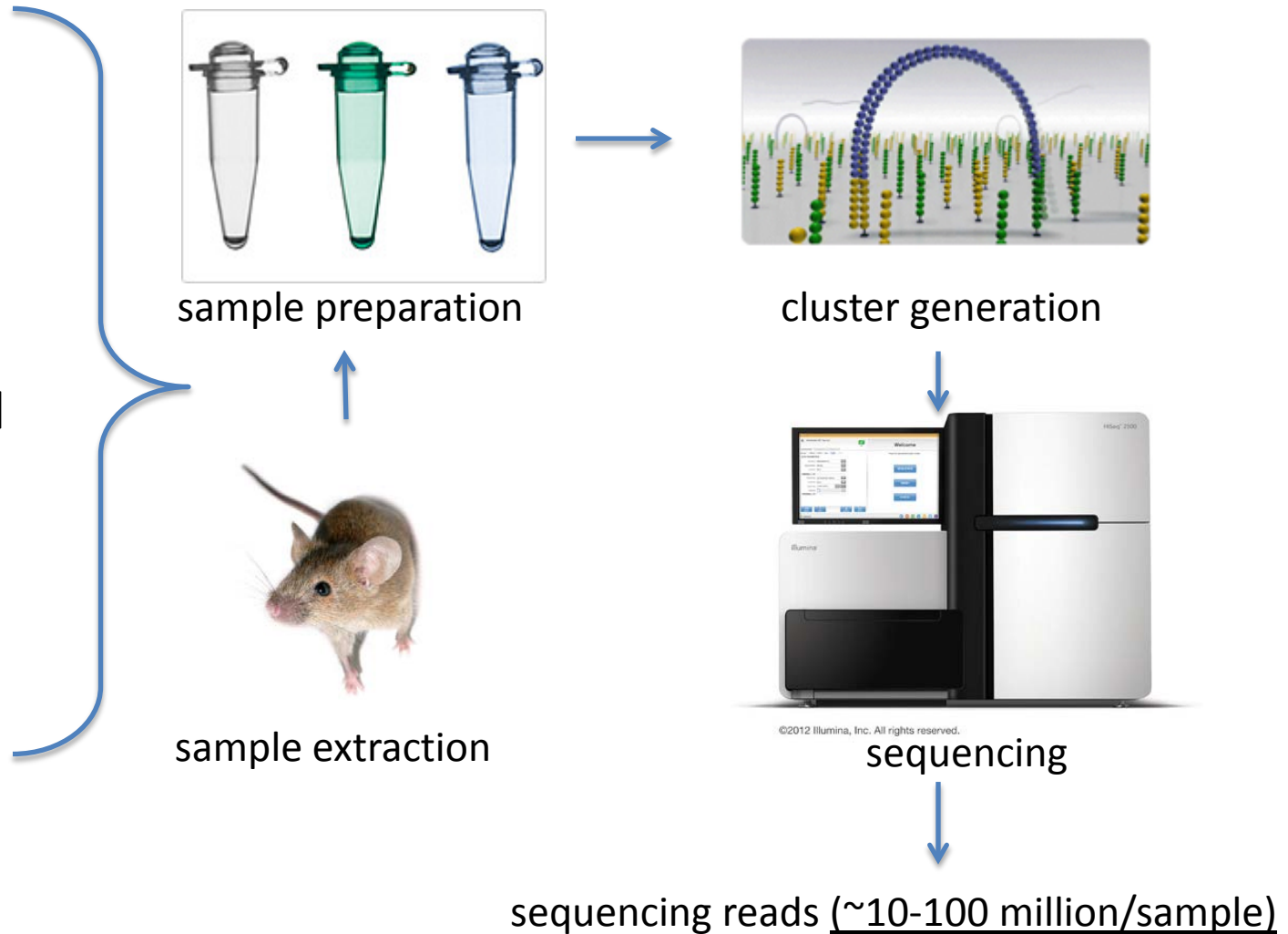


c DNase-seq



Associated Next Generation Sequencing methods

- RNA-Seq
- ChIP-Seq
- DNase-Seq
- GRO-Seq
- CAGE

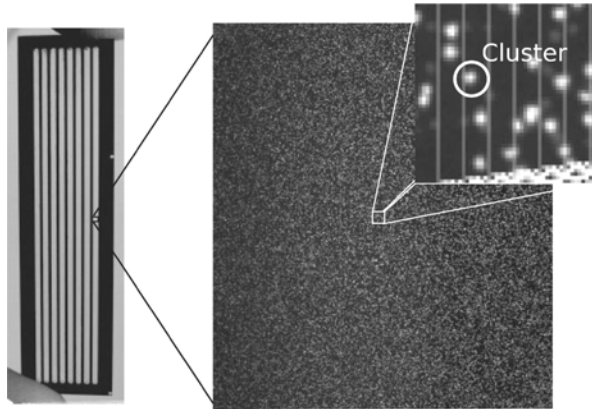


Bioinformatics Pipeline

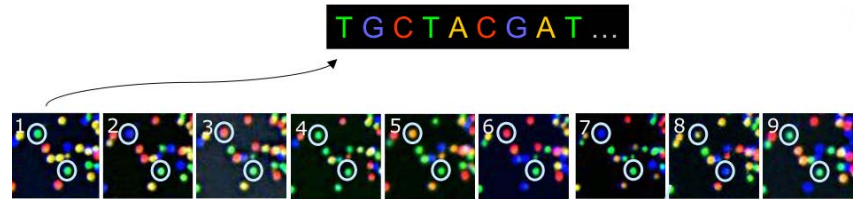
- Image analysis & Base calling
- Quality control (did the sequencing work?)
- Sequence alignment (assembly) + QC
- Read quantification (expressed? bound? chromatin state? open chromatin? ...)
- Genome wide summaries
- Comparison between technical or biological replicates
- Comparison across cell/genotype boundaries

Raw data

~115000 images per flowcell (36 bases)

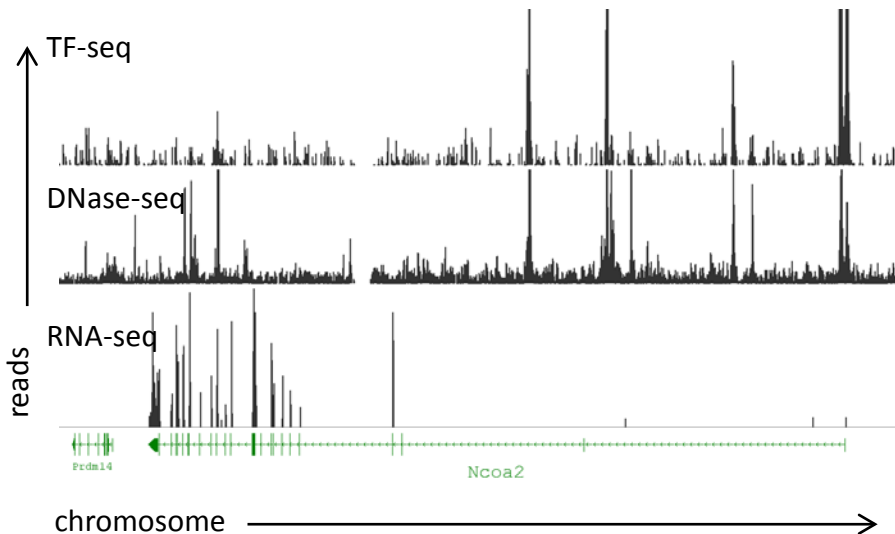


Base calling

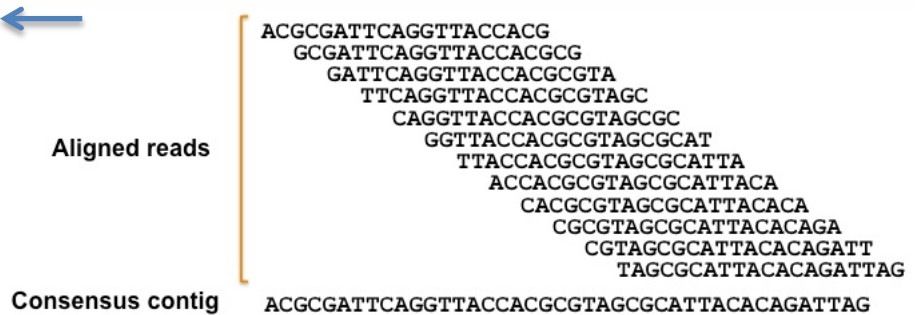


```
@HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
TTAATTGGTAAATAAATCTCCTAATAGCTTAGATN
+HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
efcffffcfeeffcfffffdff`feed]`_Ba_^__[YBBBBBBBB
```

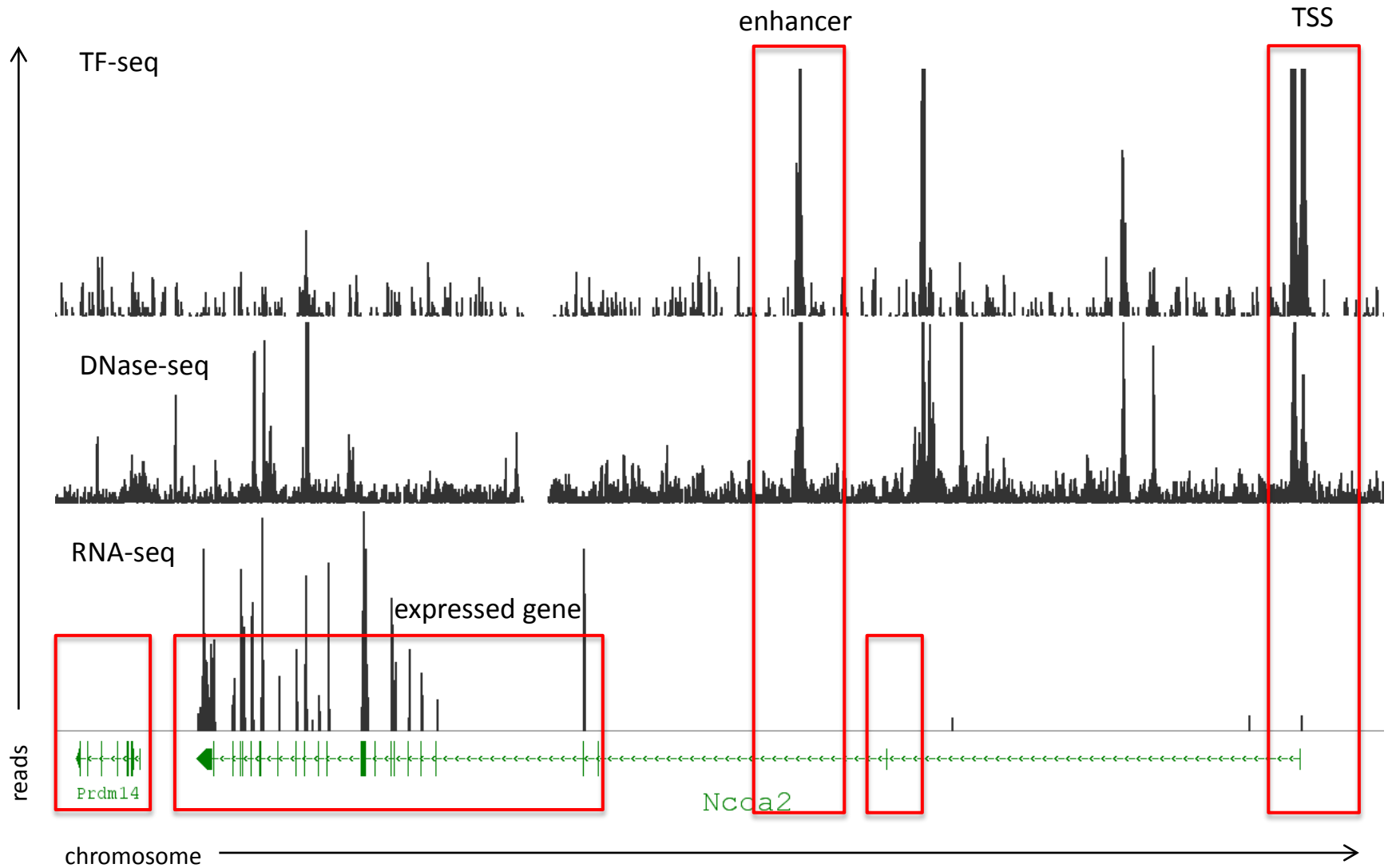
Aligned data - genome browser



Alignment

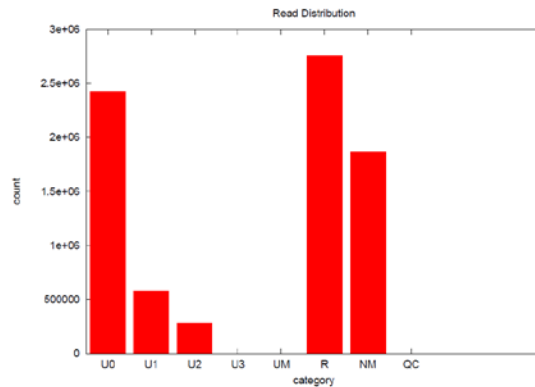


Aligned reads genome view

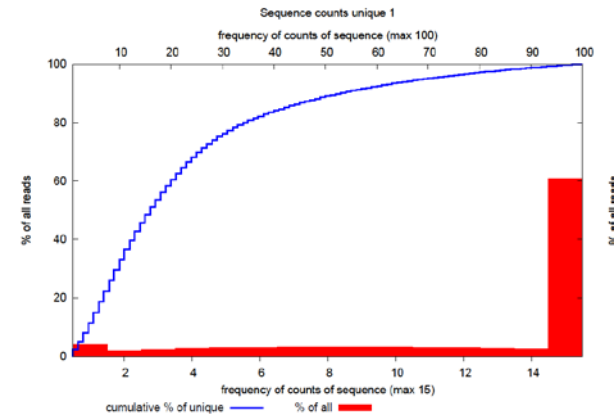


Some important key ChIP-Seq quality check numbers

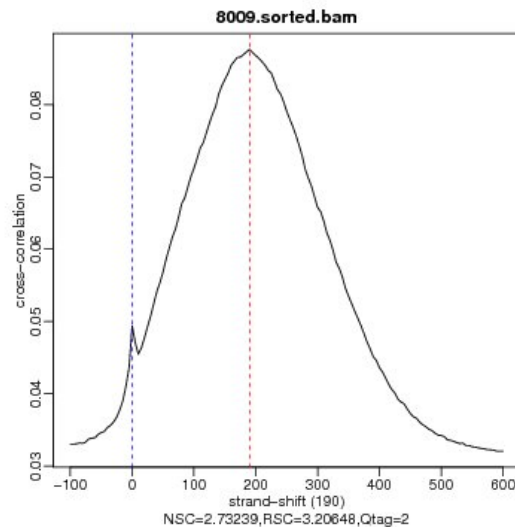
Alignment statistics



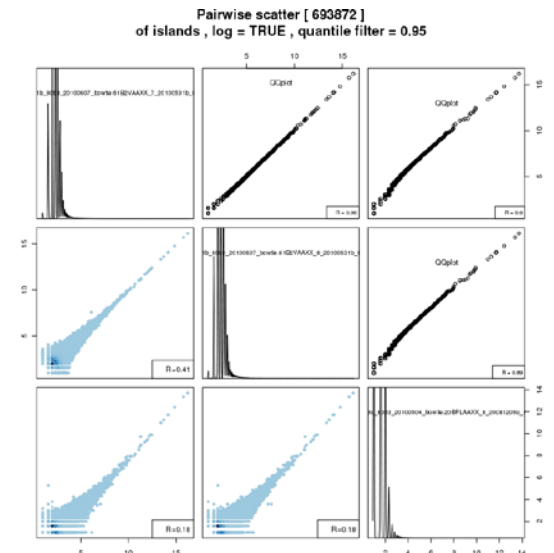
Read duplication



Strand cross correlation



Correlation of technical replicates



Spotfire webplayer & Integrated Genome Browser (IGB)

summary.dxp - TIBCO Spotfire

File Edit View Insert Tools Help

Table Cross_correlation Duplication Read length Page Page (2)

Table

sid	exptype	genotype	celltype	antibody_simple	rlen	alig...	nread...	qua	frip	nsc_log10	rsc_log10	unp	is_merged	img_scatter_b...	img_crosscorr	igb_merged_b...	igb_all_bw_lh	igb_best_bw_lh
11887	ChIP-Seq	Vavcre E2A(F/F) Iκ(Bio/Bio) R26 (BirA/BirA)	E2A def Progenitors	Streptavidin	NA	36M	81.52	1	0.39	0.01	0.00		True			http://127.0.0.1:7... version=mm9&se... 3A%2F% 2Fmammut%2F% 7Eart% 2FProjects%		
11888	ChIP-Seq	Vavcre E2A(F/F) Iκ(Bio/Bio) R26 (BirA/BirA)	E2A def Progenitors	Streptavidin	NA	36M	71.88	1	0.38	0.01	0.00		True			http://127.0.0.1:7... version=mm9&se... 3A%2F% 2Fmammut%2F% 7Eart% 2FProjects%	http://127.0.0.1:7... version=mm9&se... 3A%2F% 2Fmammut%2F% 7Eart% 2FProjects%	
11889	ChIP-Seq	Vavcre Ebf1(F/F) Iκ(Bio/Bio) R26 (BirA/+)	Ebf1 def Progenitors	Streptavidin	NA	36M	53.21	1	0.32	0.01	0.00		True			http://127.0.0.1:7... version=mm9&se... 3A%2F% 2Fmammut%2F% 7Eart% 2FProjects%	http://127.0.0.1:7... version=mm9&se... 3A%2F% 2Fmammut%2F% 7Eart% 2FProjects%	
11889	ChIP-Seq	Vavcre Ebf1(F/F) Iκ(Bio/Bio) R26 (BirA/+)	Ebf1 def Progenitors	Streptavidin	HISeq2000-SR50	36M	29.41	0	0.28	0.01	-0.20	88.57	False			http://127.0.0.1:7... version=mm9&se... 3A%2F% 2Fmammut%2F% 7Eart% 2FProjects%	http://127.0.0.1:7... version=mm9&se... 3A%2F% 2Fmammut%2F% 7Eart% 2FProjects%	
11889	ChIP-Seq	Vavcre Ebf1(F/F) Iκ(Bio/Bio) R26 (BirA/+)	Ebf1 def Progenitors	Streptavidin	HISeq2000-SR50	36M	23.80	0	0.25	0.01	-0.24	89.92	False			http://127.0.0.1:7... version=mm9&se... 3A%2F% 2Fmammut%2F% 7Eart% 2FProjects%	http://127.0.0.1:7... version=mm9&se... 3A%2F% 2Fmammut%2F% 7Eart% 2FProjects%	
11887	ChIP-Seq	Vavcre E2A(F/F) Iκ(Bio/Bio) R26 (BirA/BirA)	E2A def Progenitors	Streptavidin	HISeq2000-SR50	36M	41.48	0	0.32	0.00	-0.25	74.92	False			http://127.0.0.1:7... version=mm9&se... 3A%2F% 2Fmammut%2F% 7Eart% 2FProjects%	http://127.0.0.1:7... version=mm9&se... 3A%2F% 2Fmammut%2F% 7Eart% 2FProjects%	
11887	ChIP-Seq	Vavcre E2A(F/F) Iκ(Bio/Bio) R26 (BirA/BirA)	E2A def Progenitors	Streptavidin	HISeq2000-SR50	36M	40.04	0	0.33	0.00	-0.25	75.56	False			http://127.0.0.1:7... version=mm9&se... 3A%2F% 2Fmammut%2F% 7Eart% 2FProjects%	http://127.0.0.1:7... version=mm9&se... 3A%2F% 2Fmammut%2F% 7Eart% 2FProjects%	

Filters

Type to search filters

id
1 431
1188

sid
1188

exptype
 ChIP-Seq
 DNaseHS

genotype
Type to search in list
(All) 2 values
Vavcre E2A(F/F) Iκ(Bio/Bio) R26 (BirA/BirA)
Vavcre Ebf1(F/F) Iκ(Bio/Bio) R26 (BirA/+)

celltype
Type to search in list
(All) 2 values
E2A def Progenitors
Ebf1 def Progenitors

antibody
Type to search in list
(All) 1 values
Streptavidin

is_merged
 False
 True

nreads
0 174921053

Chromosomes & (mouse mm9) (ick37) Integrated Genome Browser 7.0.2

File Edit View Tools Tabs Bookmarks Help

Chromosome 4 (mouse mm9) (ick37) Integrated Genome Browser 7.0.2

Selection Info: Click the map below to select annotations

Load Data Load Sequence

ChIP-Seq

11887.sorted.cov.all

11888.sorted.cov.all

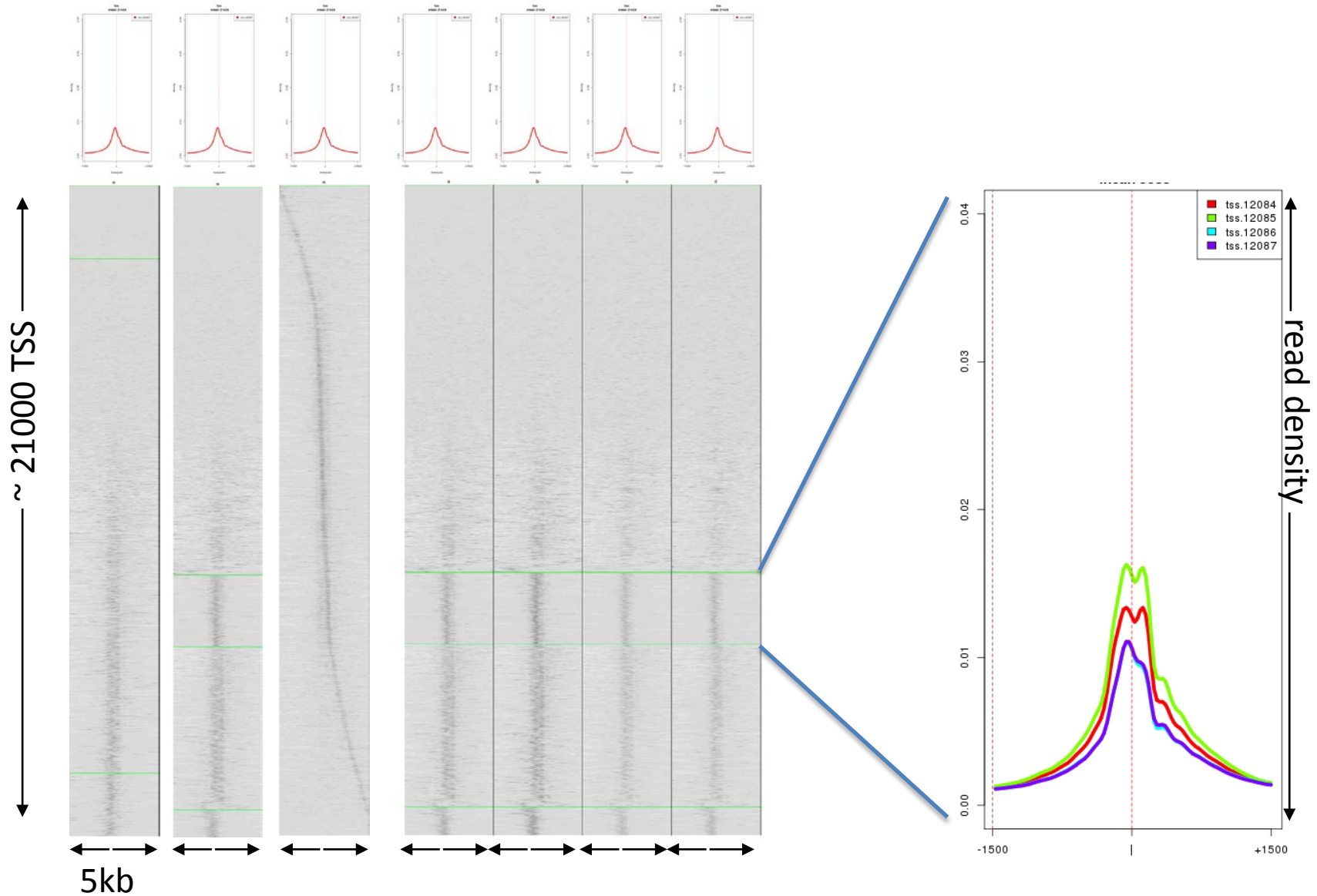
C118FACV_3_11888_CGATGT.sorted.cov.all

C118FACV_4_11888_CGATGT.sorted.cov.all

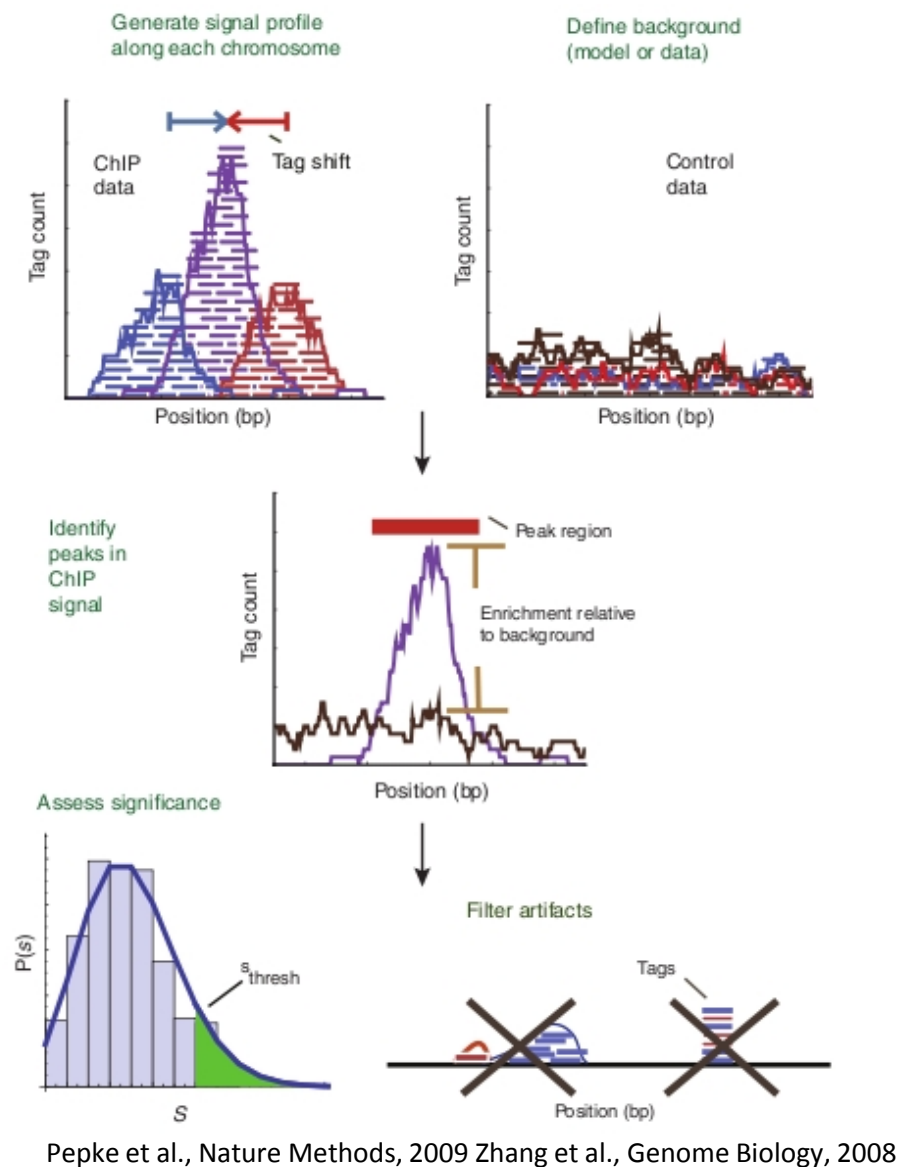
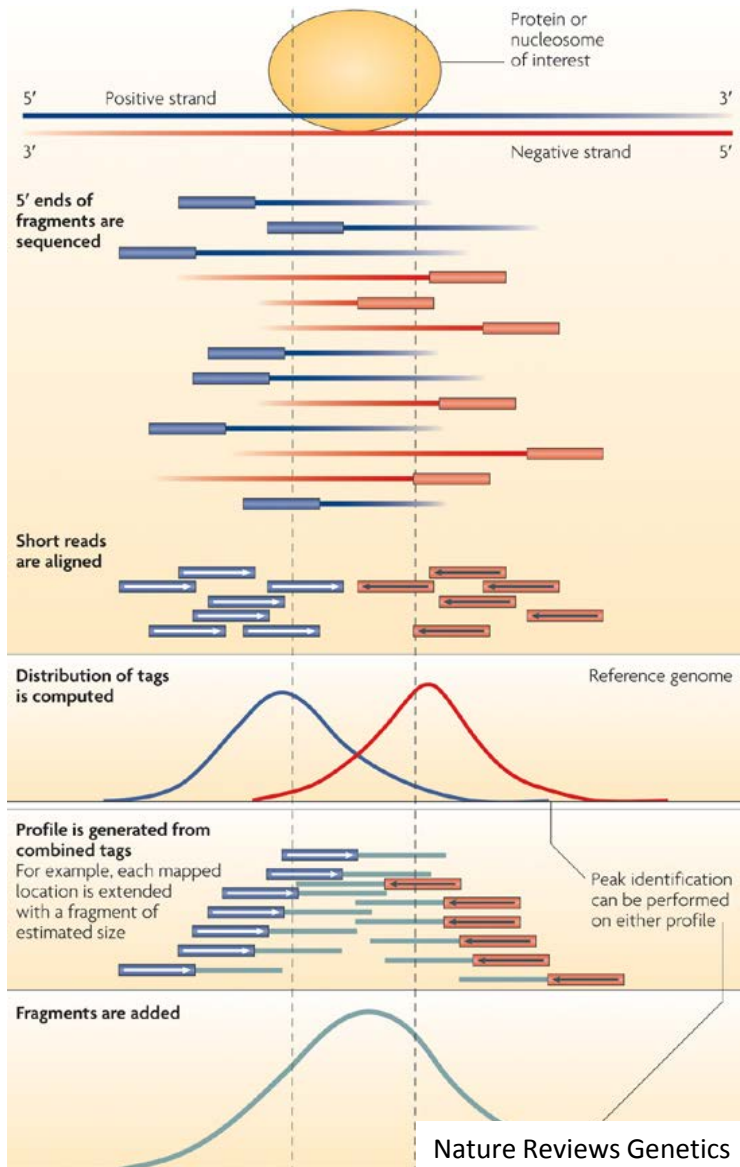
C118FACV_3_11887_ATCACG.sorted.cov.all

C118FACV_4_11887_ATCACG.sorted.cov.all

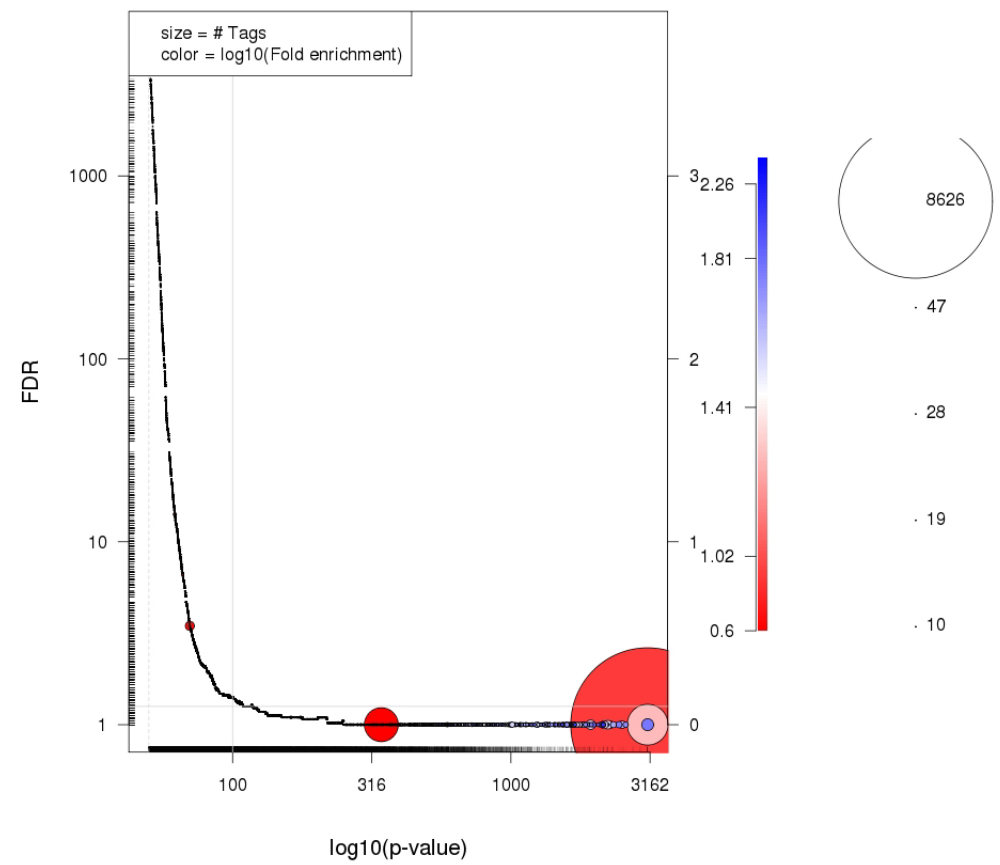
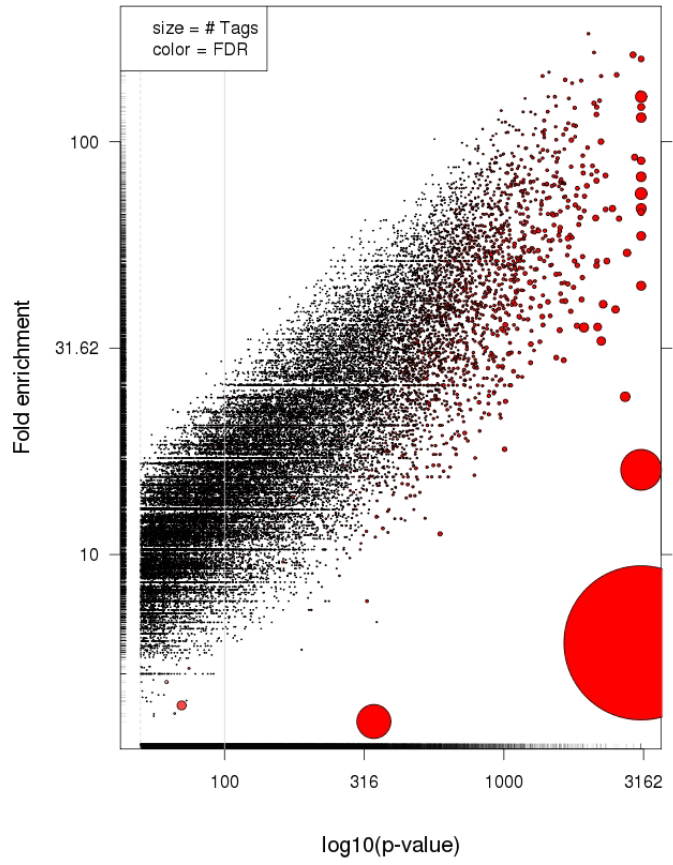
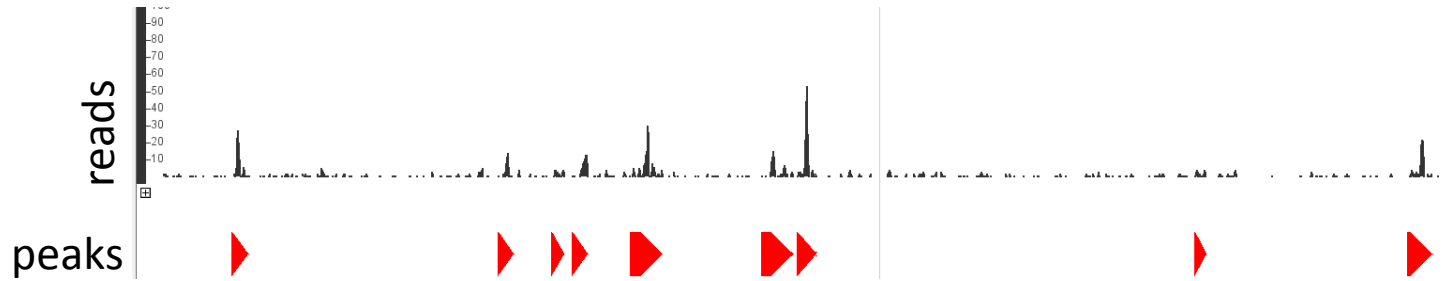
TSS Read density across the genome



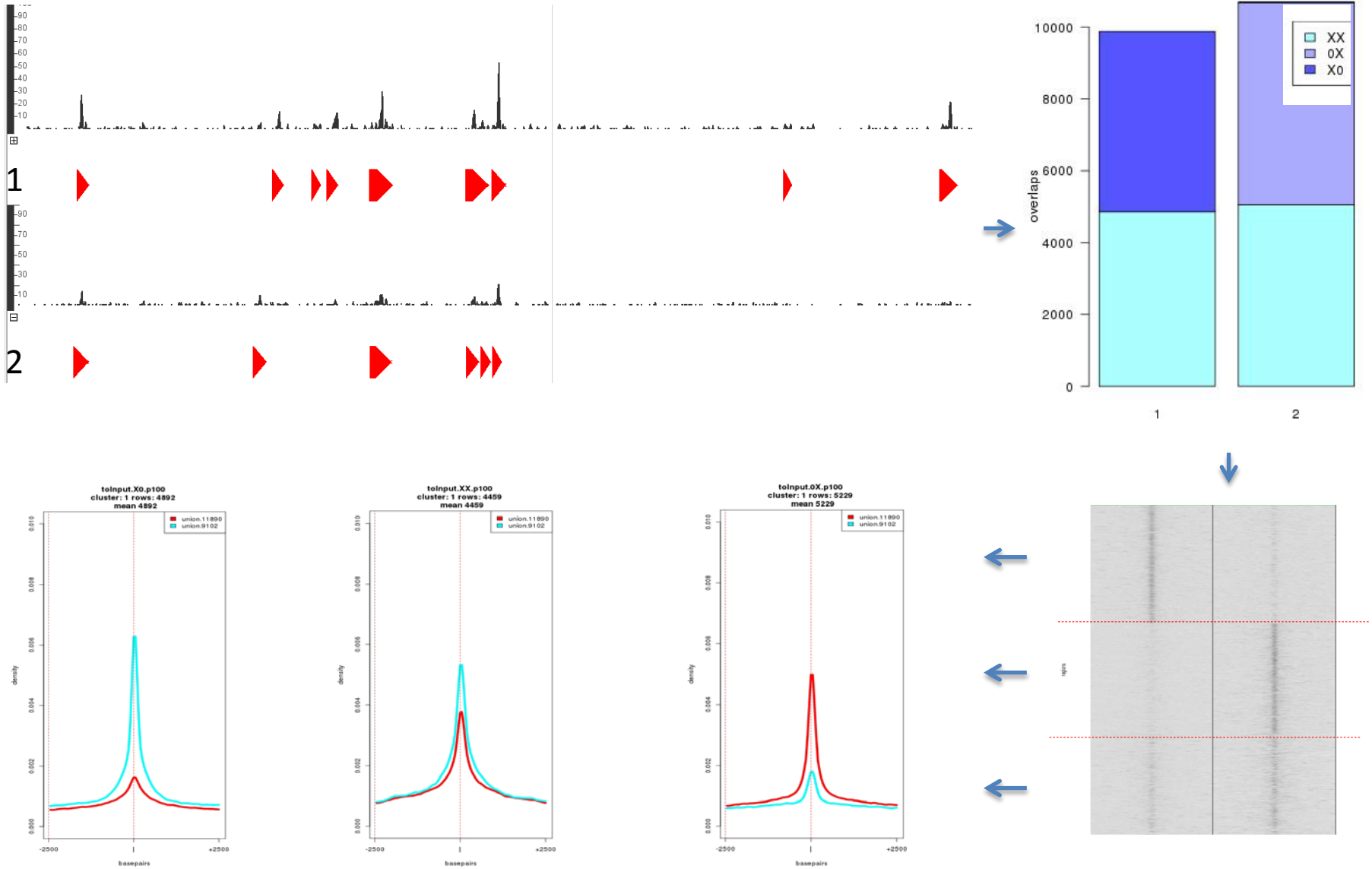
How to find piles of read densities: "peak calling"



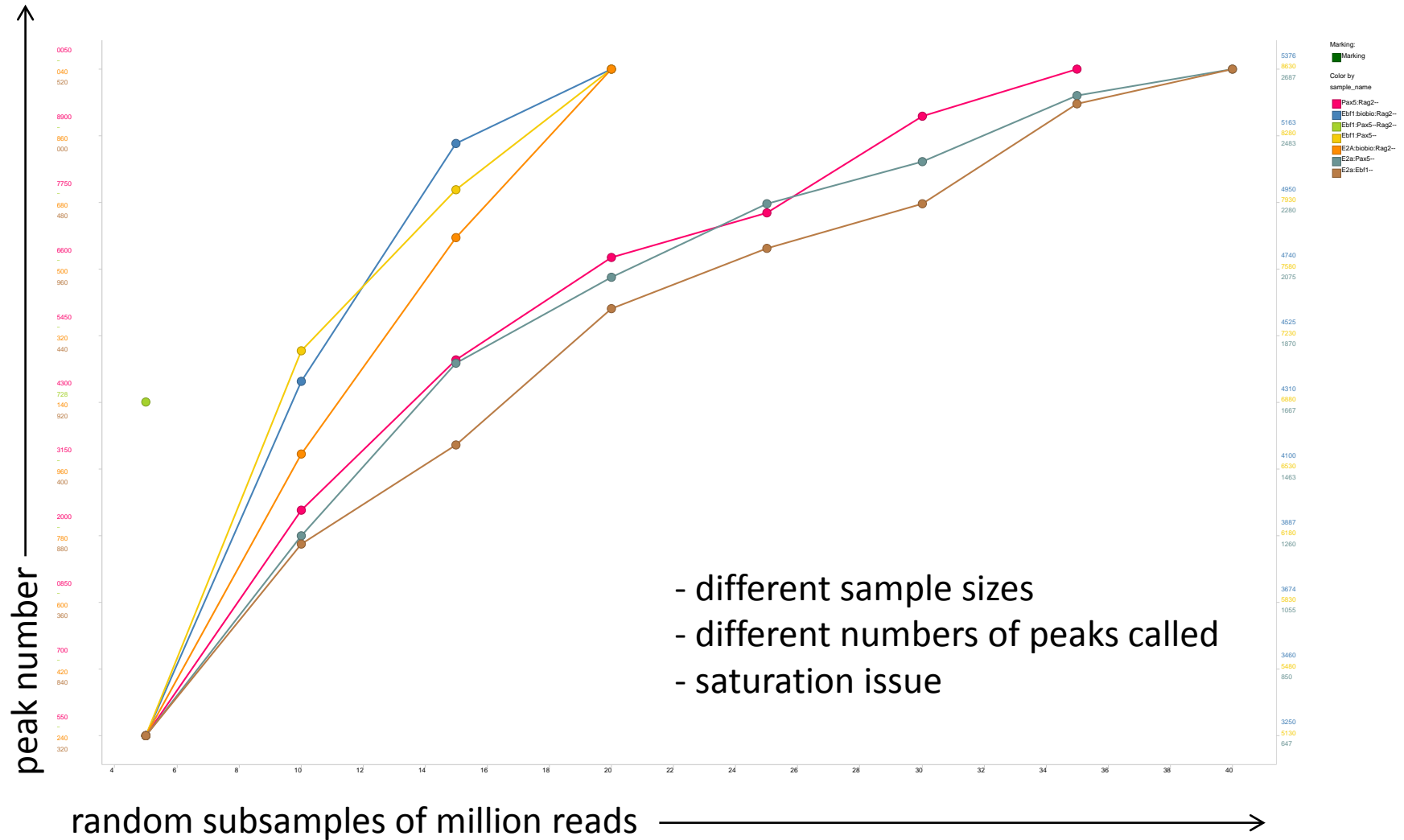
Peak calling results



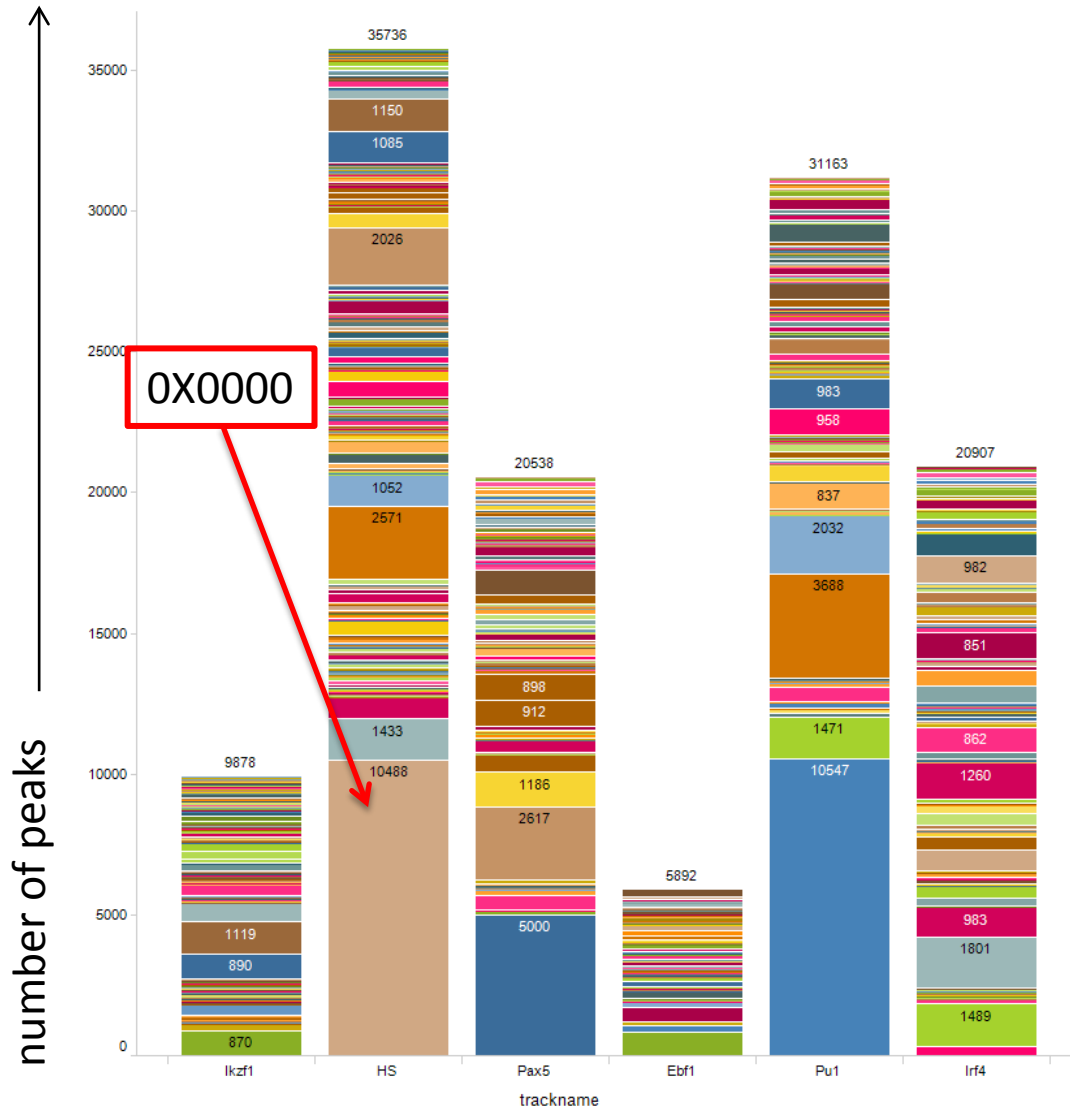
Overlap of two track's peaks



Peak calling & sample size



Peak overlaps of multiple tracks

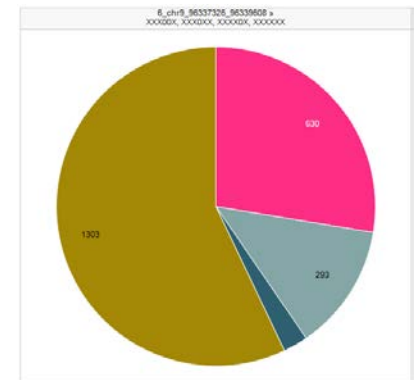
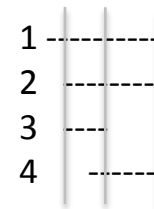


Marking:
■ Marking

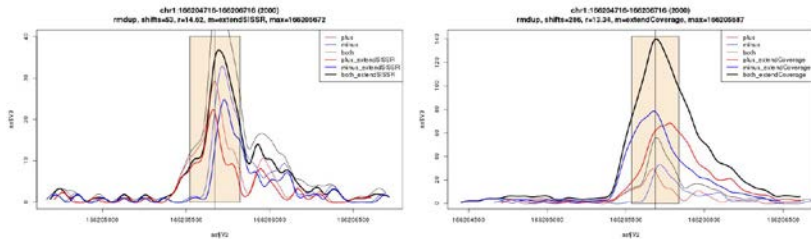
Color by
code_strict_per_peak

00000X
0000X0
0000XX
0000XX, 000X0X, 000XXX
0000XX, 000X0X, 00X0XX, 00X0XX, 00XXXX
0000XX, 000XXX
0000XX, 000X0, 000XXX
0000XX, 000XXX
0000XX, 000XXX, 00X00X, 00X0XX, 00XXXX
0000XX, 000XXX, 00X0X0, 00X0XX, 00XXXX
0000XX, 000XXX, 00X0X0, 00XXXX, 00XXXX
0000XX, 000XXX, 00X0XX, 00XXXX
0000XX, 000XXX, 00XXXX
0000XX, 000XXX, 00XXXX
0000XX, 000XXX, 00XXXX, 0X00XX, 0X0XXX
0000XX, 000XXX, 0X00X0, 0X0XXX, 0X0XXX, 0XXXXX
0000XX, 000XXX, 0X00XX, 0X0XXX
0000XX, 000XXX, 0X00XX, 0XXXXX
0000XX, 00X00X, 00X0XX
0000XX, 00X0X0
0000XX, 00X0X0, 00X0XX
0000XX, 00X0X0, 00X0XX, 00XXXX, 00XXXX
0000XX, 00X0X0, 00X0XX, 0X00XX
0000XX, 00X0X0, 00X0XX, 0X00XX
0000XX, 00X0X0, 0X00XX, 0X00X0, 0X0XXX, 0XXXXX
0000XX, 00X0XX
0000XX, 00X0XX, 00XXXX
0000XX, 00X0XX, 00XXXX, 0X00X0, 0X0XXX, 0XXXXX
0000XX, 00X0XX, 00XXXX, 0XXXXX
0000XX, 00X0XX, 0X000X, 0X00XX, 0X00XX

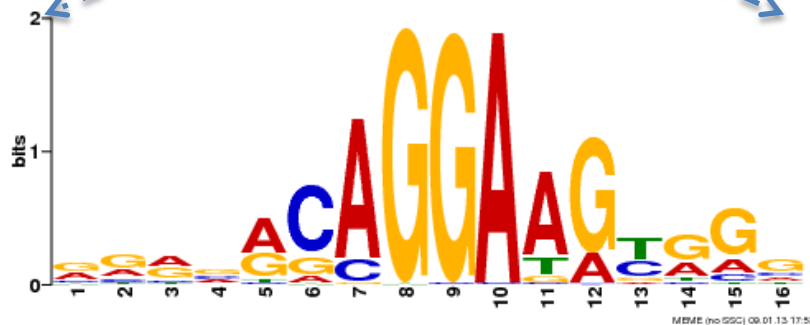
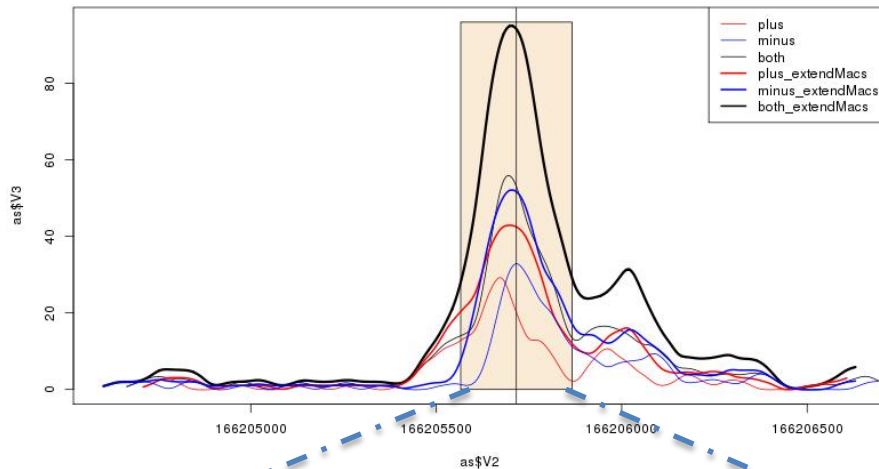
overlap
example



Finding the most common sequence motif in a peak



chr1:166204716-166206716 (2000)
 rmdup, shifts=137, r=16.19, m=extendMacs, max=166205716



For each peak:

- get all reads, remove duplicates
- shift + extend reads (3 methods)
- get location and read count of summit (peak max)
- get ratio of reads around peaks over rest of peak (“peakiness”)
- recenter the peak
- extract associated peak sequences (masked/unmasked)
- select **300** “best” peaks, according to “peakiness” score (sort and min deviation of shift method results (max location) ≥ 100)
- meme-chip (meme [de novo], dreme [de novo], mast [scan], tomtom [motif compare], centrmo [centralenriched])
- search full peak sequence set with obtained motifs (mast)
- also for random(shuffled sequences)

Bioinformatics pipeline & outlook

- Build a database of all our tracks (resource)
- Compute quality values of ChIP-Seq tracks
- Call peaks (TF/HS/Chromatin)
 - calculate peak saturation
- Assign peaks to genes
- Find motifs (where applicable)
- Compare (biological) replicates
- Perform data analysis/mining
 - Identify hotspots
 - Find interesting Heatmaps, densities, clusters ...

Data != Publications

2008-2013

888 samples

234 sequencing requests

606 “lanes”

126 geno types

55 cell types



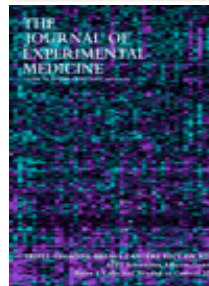
Revilla-I-Domingo R, Bilic I, Vilagos B, Tagoh H, Ebert A, Tamir IM, Smeenk L, Trupke J, Sommer A, Jaritz M, Busslinger M. EMBO J. 2012



Vilagos B, Hoffmann M, Souabni A, Sun Q, Werner B, Medvedovic J, Bilic I, Minnich M, Axelsson E, Jaritz M, Busslinger M. J Exp Med. 2012



McManus S, Ebert A, Salvagiotto G, Medvedovic J, Sun Q, Tamir I, Jaritz M, Tagoh H, Busslinger M. EMBO J. 2011



Ebert A, McManus S, Tagoh H, Medvedovic J, Salvagiotto G, Novatchkova M, Tamir I, Sommer A, Jaritz M, Busslinger M. Immunity. 2011

Acknowledgements

- Meinrad Busslinger & lab members, IMP
- NGS Bioinformatics office
 - Elin Axelsson, Malgorzata Goiser, Roman Stocsits
- Campus Science Support Facilities
 - Andreas Sommer (NGS)
 - Bioinformatics: Ido Tamir, Heinz Ekker (NGS)
 - Andras Aszodi (SCC)
- Bioinformatics Services
 - Wolfgang Lugmayr (linux cluster)
- Open source tools (R, NGS tools, peak callers, Galaxy ...)
- Commercial tools: Spotfire, Ingenuity