# Next Generation Sequencing Data Analysis:

# From ChIP-Seq Read Islands to Epigenomics Information

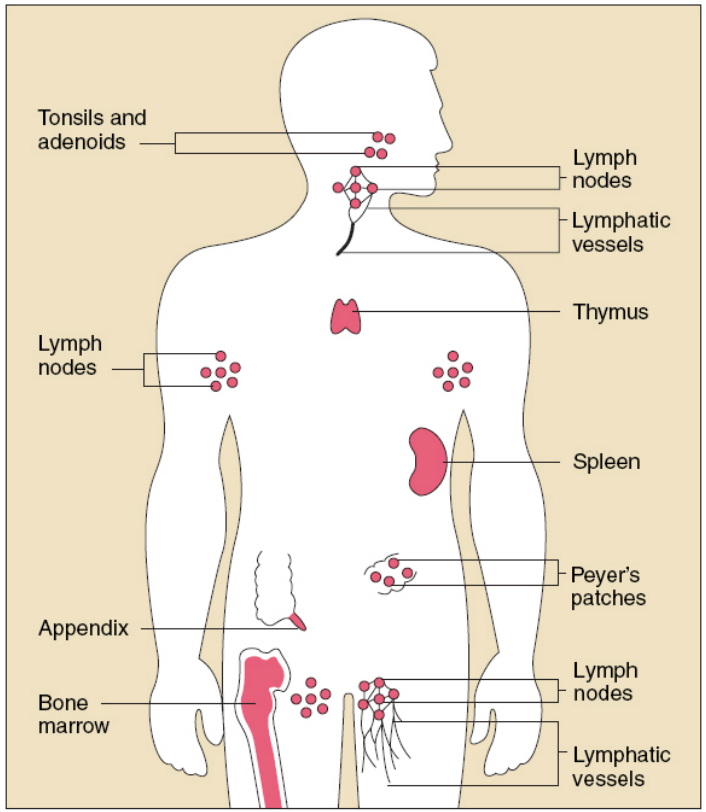
WBS Sommer Seminar, July 4th 2013

M. Jaritz, IMP

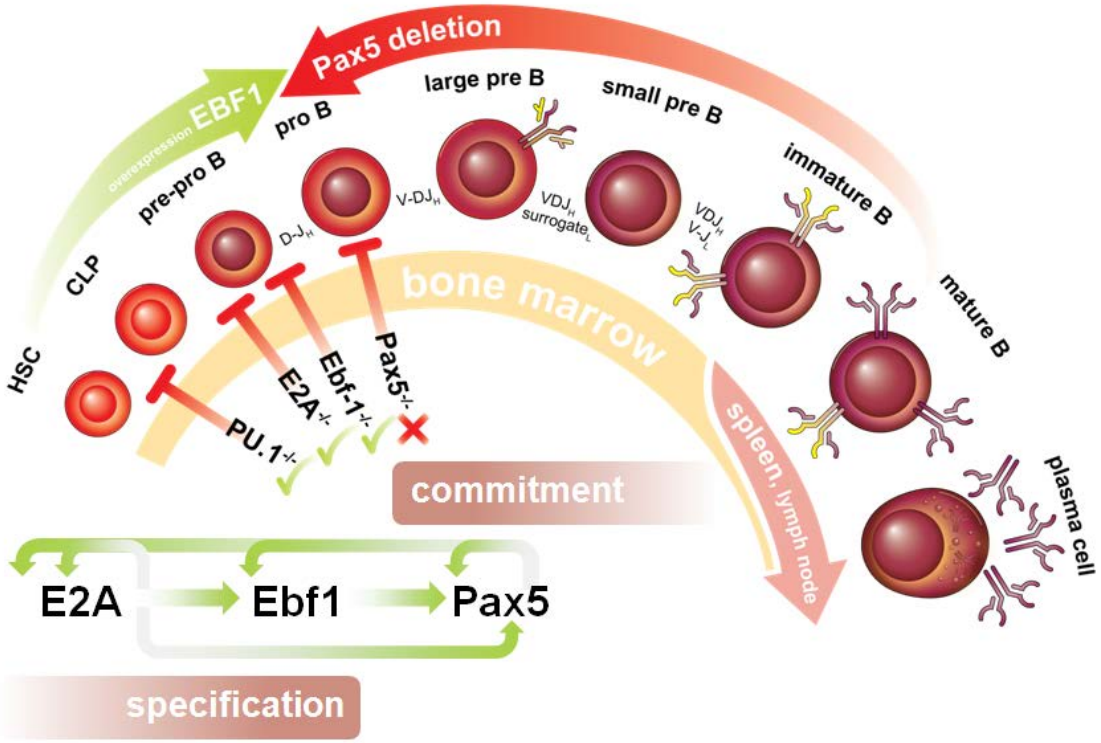
# Outline

- 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

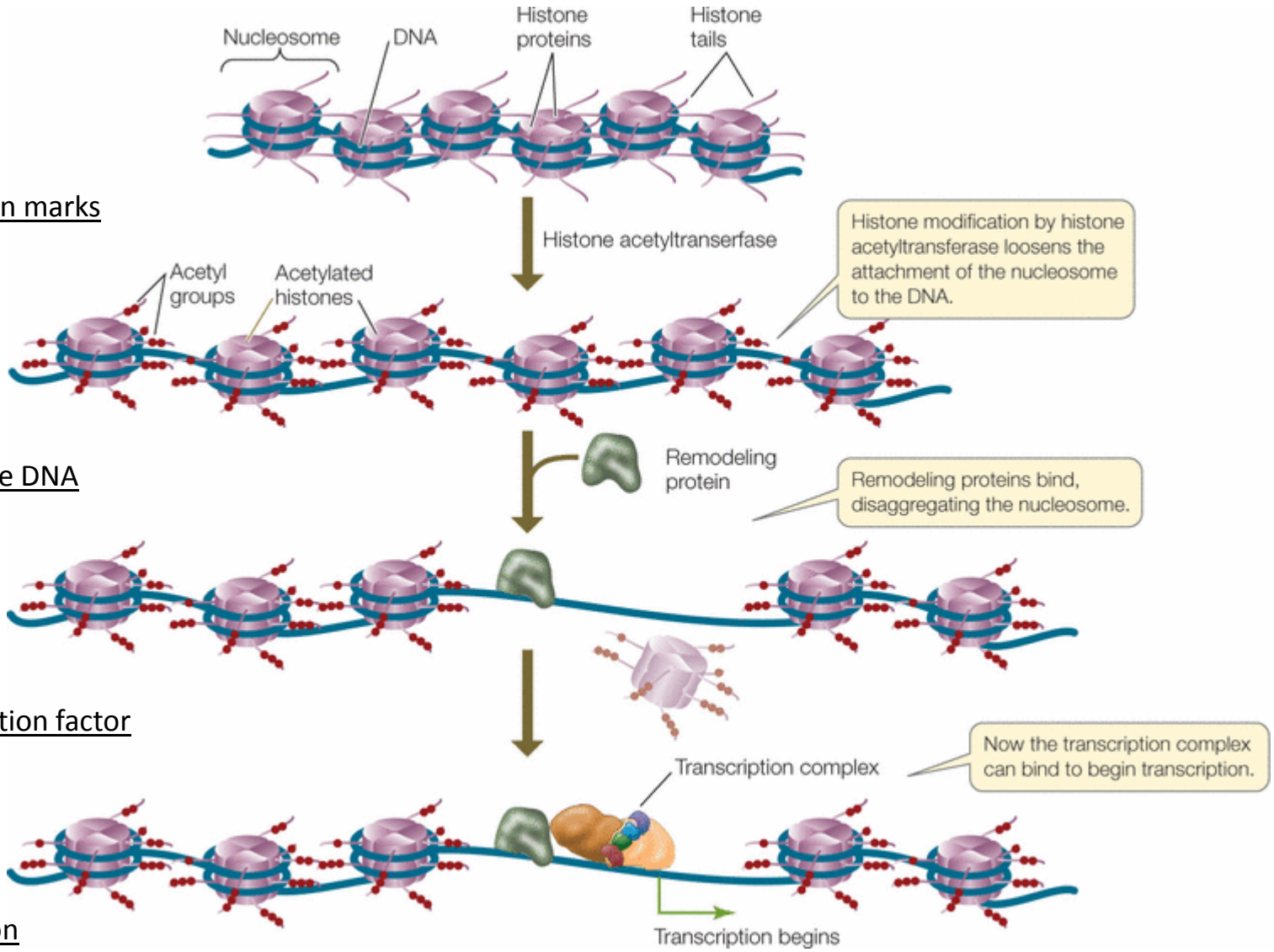
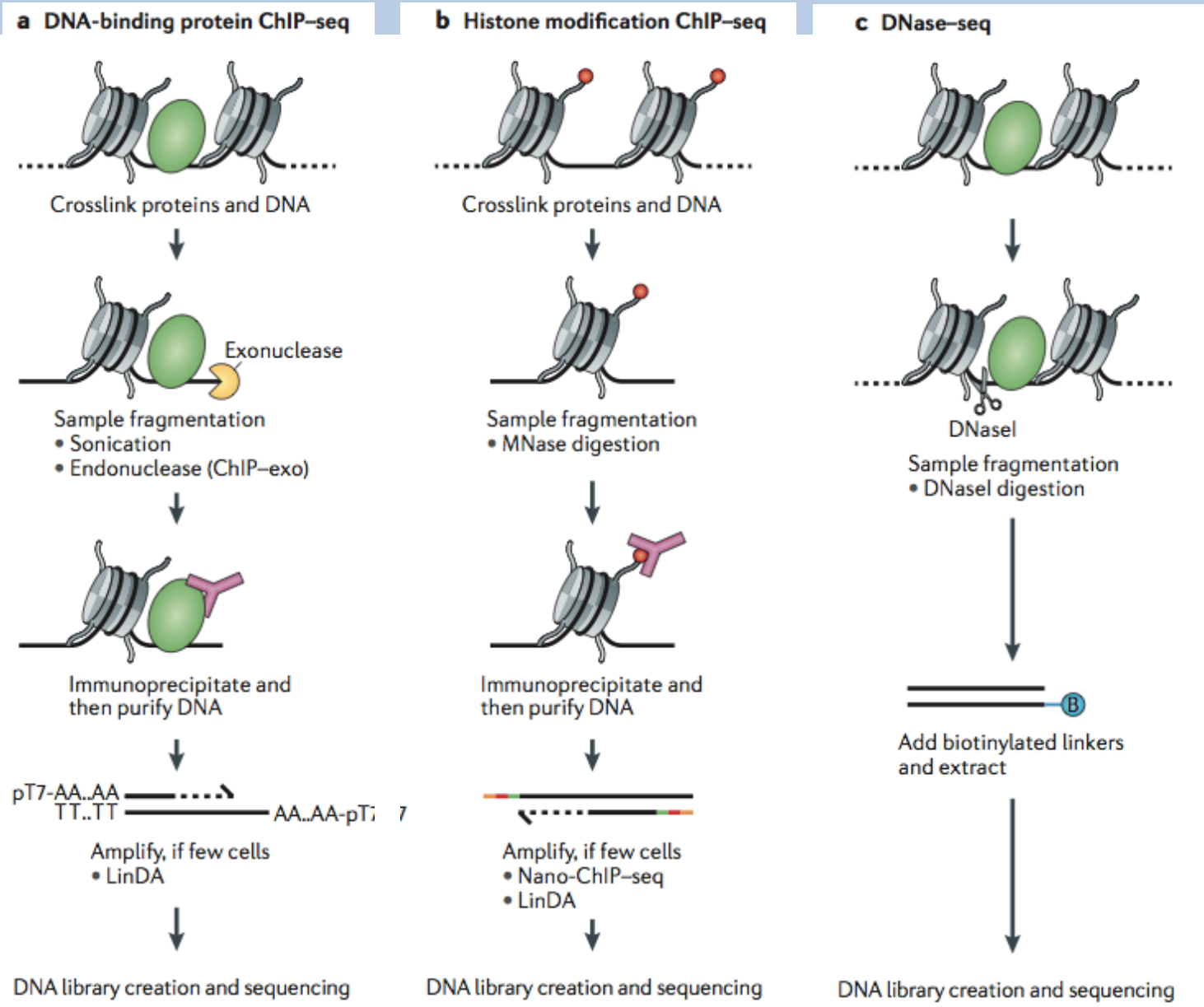



Image: www.studyblue.com

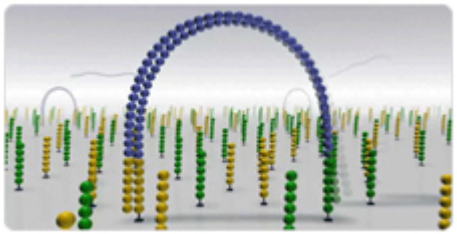
# Sequencing experiment protocols



# Associated Next Generation Sequencing methods

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

sample preparation



cluster generation

sample extraction


sequencing

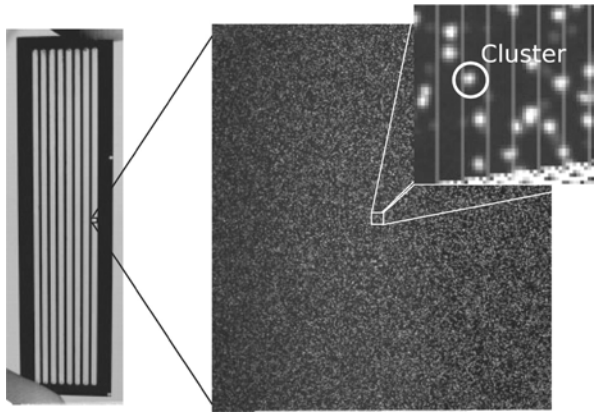
sequencing reads (~10-100 million/sample)

# Bioinformatics Pipeline

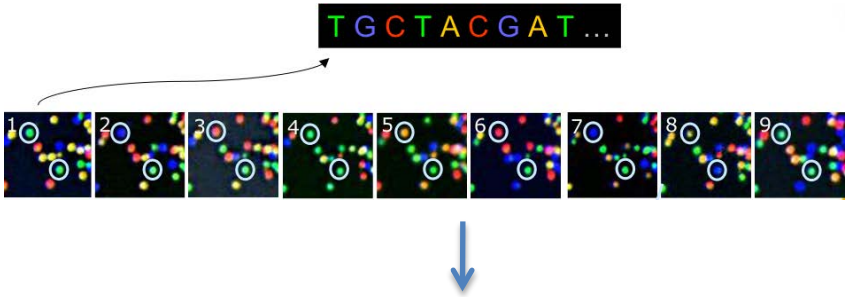
- Image analysis & Base calling
- Quality control (did the sequencing work?)
- Sequence alignment (assembly) + QC
- Read quantification (expressed? bound? chromatin state? open chromatin? ...)
- Genome wide summaries
- Comparison between technical or biological replicates
- Comparison across cell/genotype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

# Raw data

~115000 images per flowcell (36 bases)



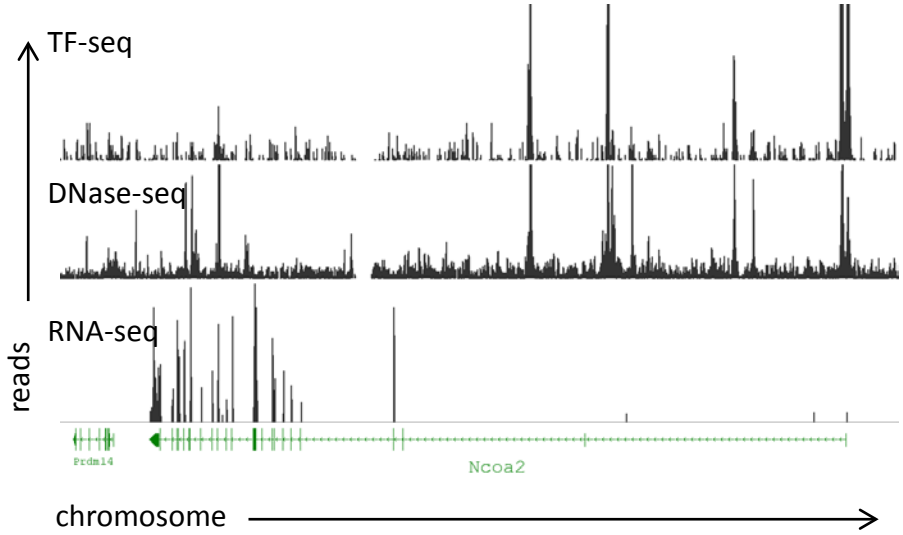

Base calling




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

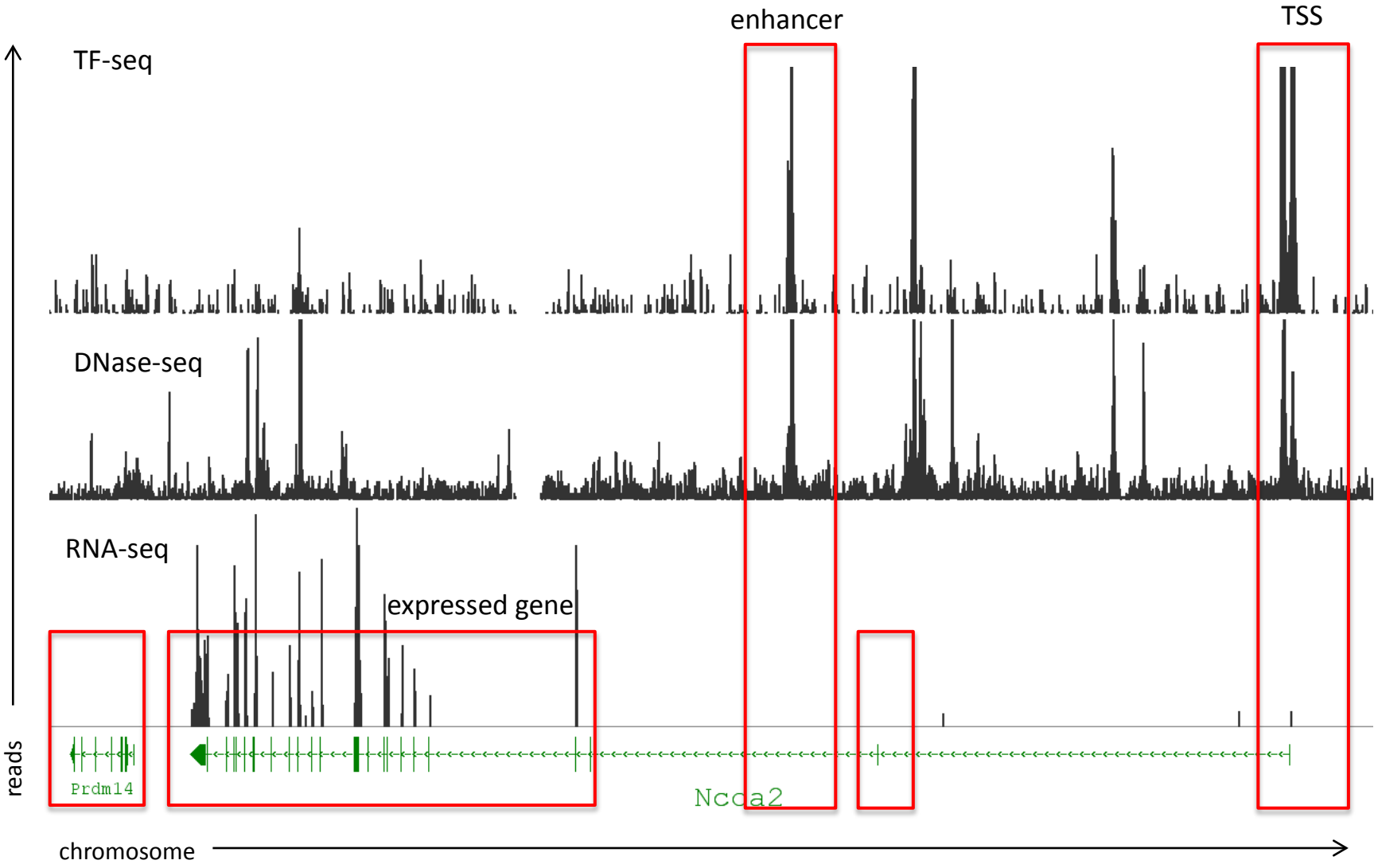
Alignment


Aligned data - genome browser



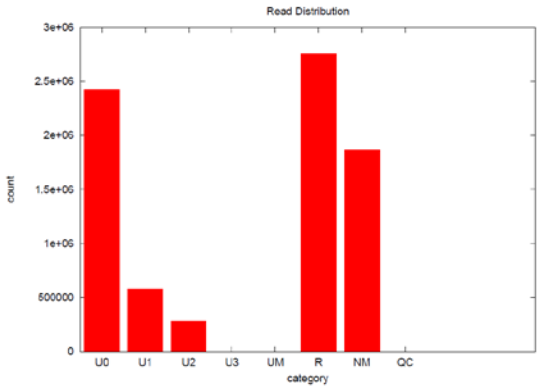

Flowcell images: Whiteford et al., Bioinformatics 2009; Illumina; contig.worldpress.com, flxex 2010

# Aligned reads genome view

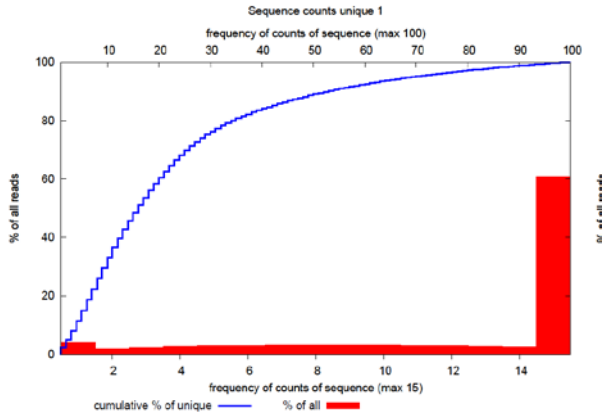


# Some important key ChIP-Seq quality check numbers

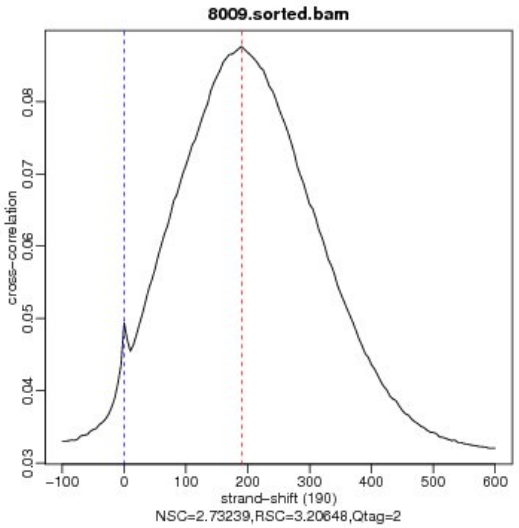
## Alignment statistics



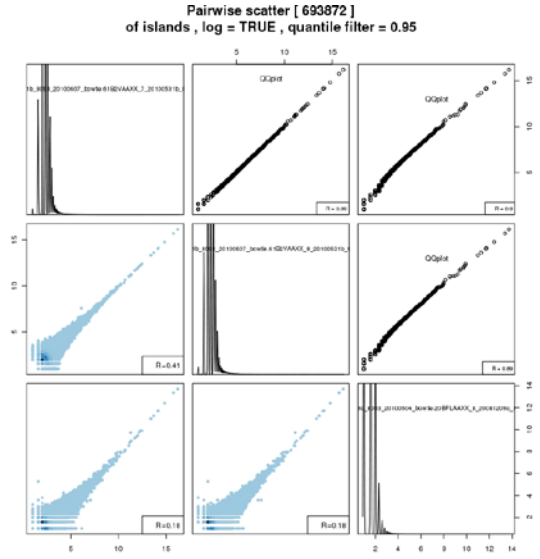

## Read duplication



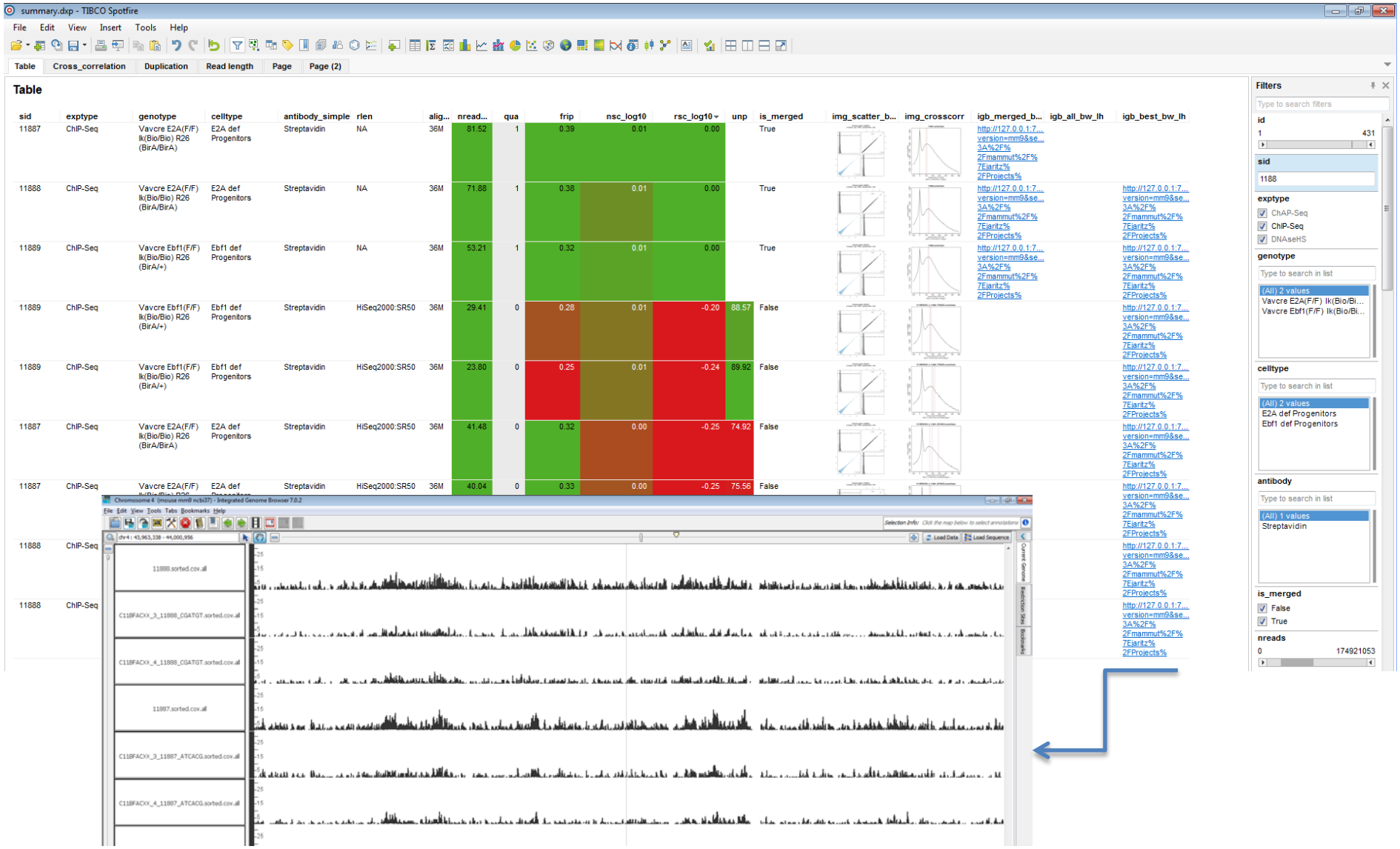

## Strand cross correlation



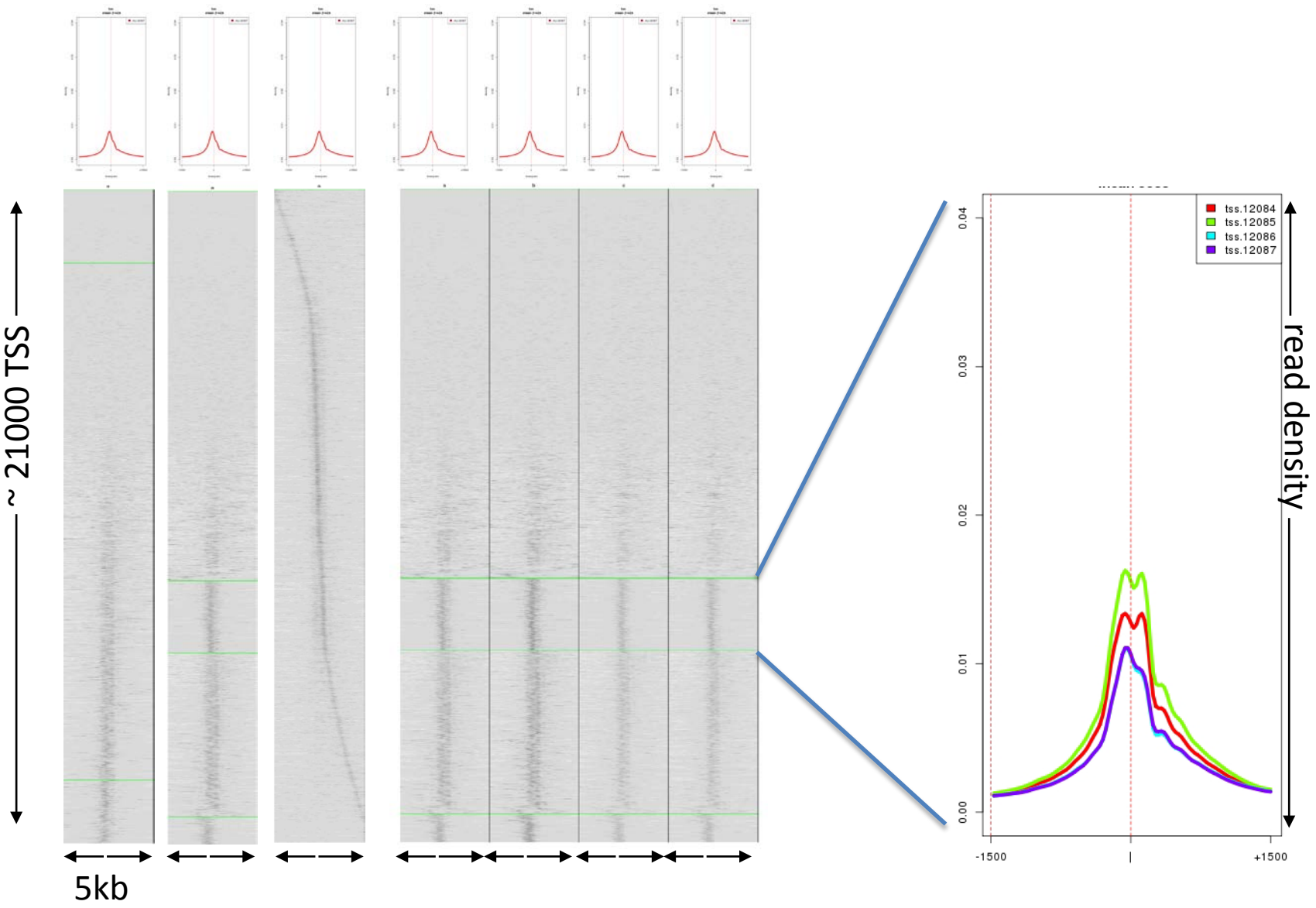

## Correlation of technical replicates



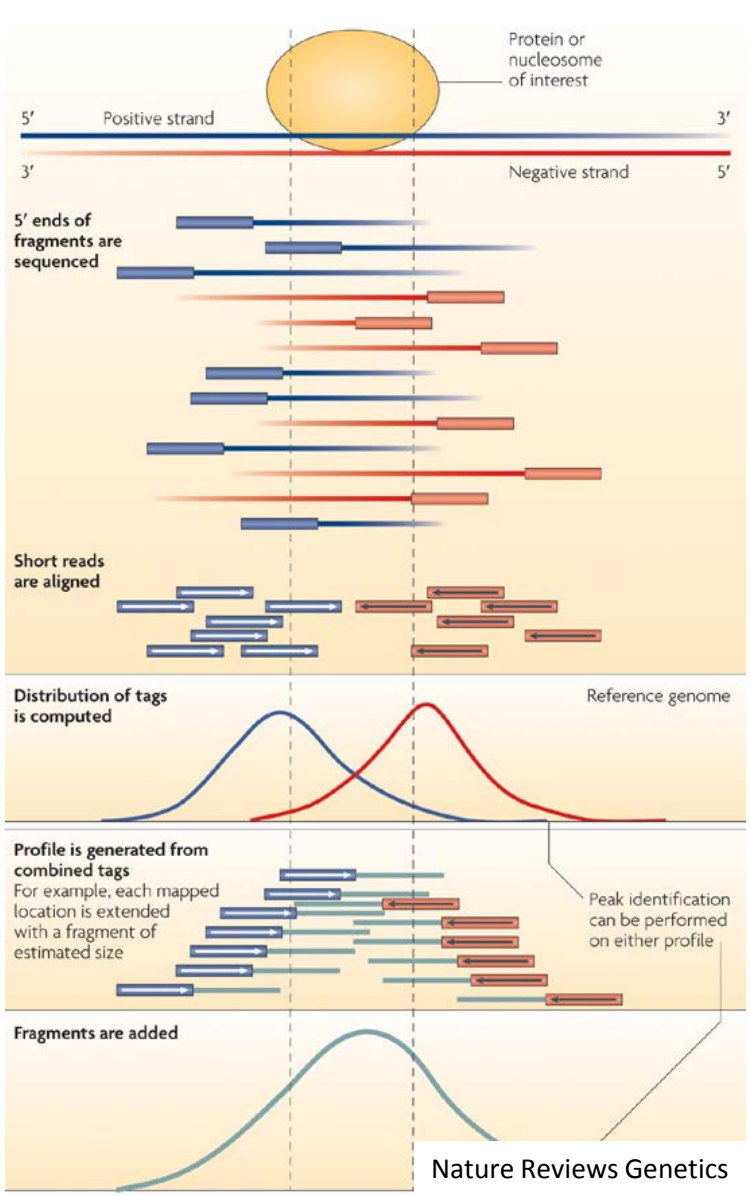
# Spotfire webplayer & Integrated Genome Browser (IGB)



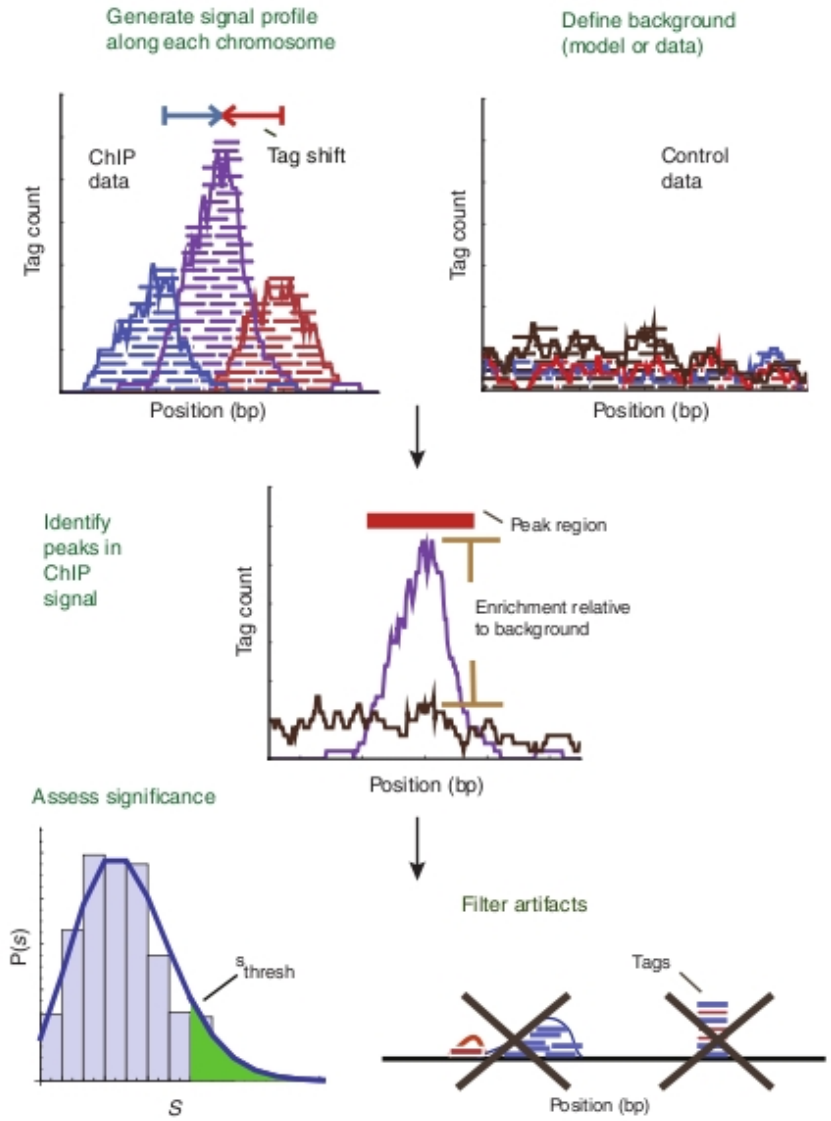
# TSS Read density across the genome



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

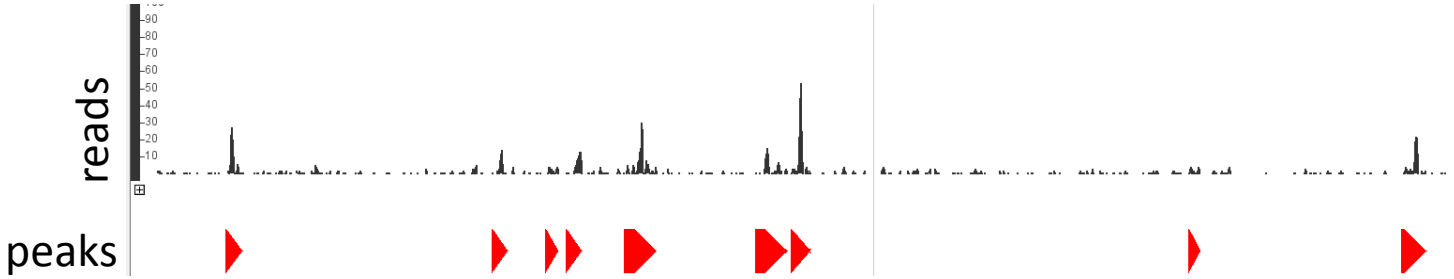


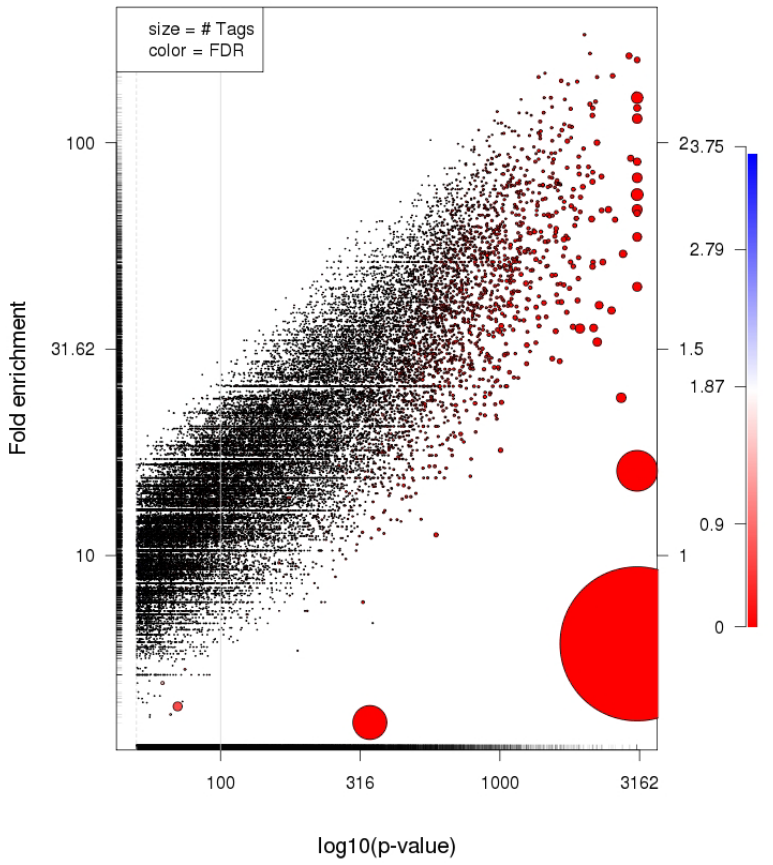

Nature Reviews Genetics

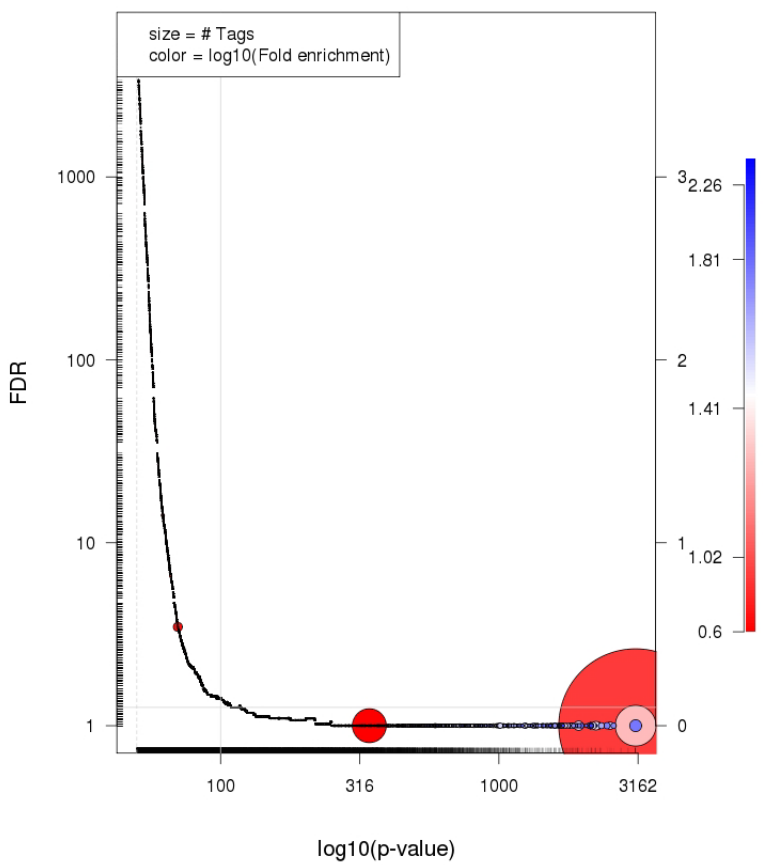



Pepke et al., Nature Methods, 2009 Zhang et al., Genome Biology, 2008

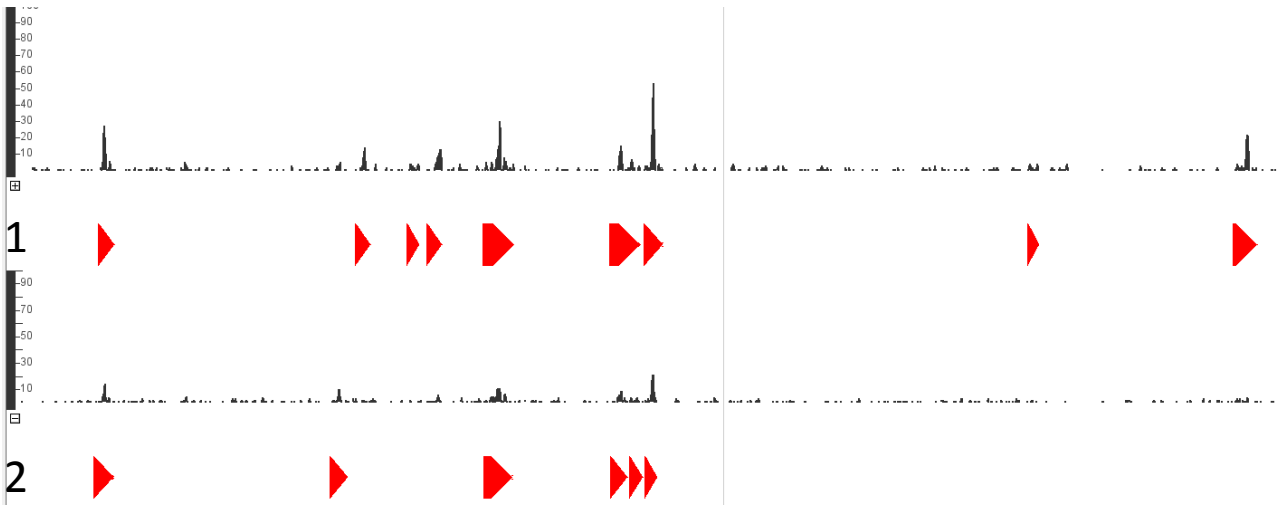
# Peak calling results

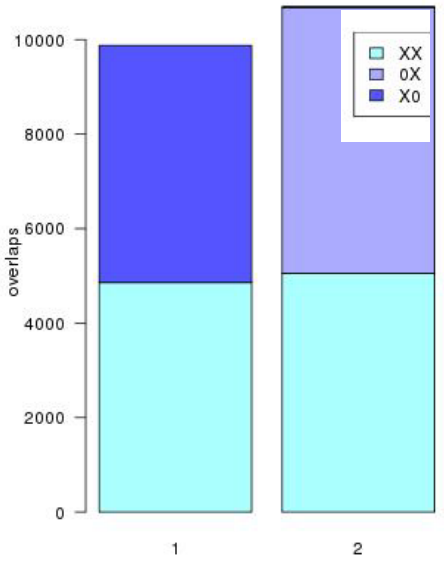


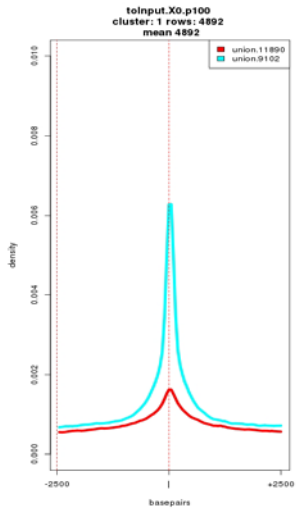



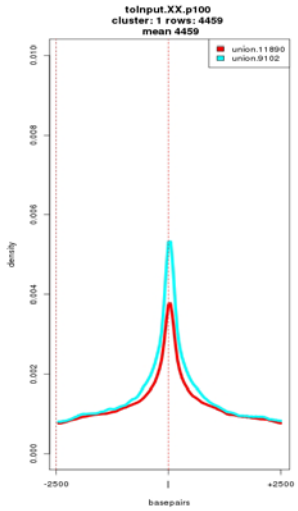



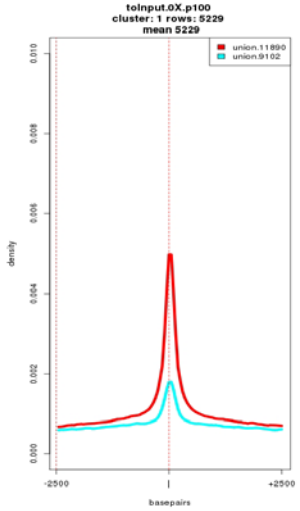

# Overlap of two track's peaks

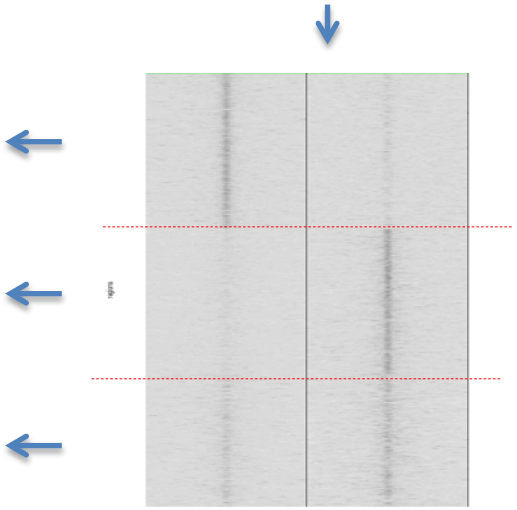


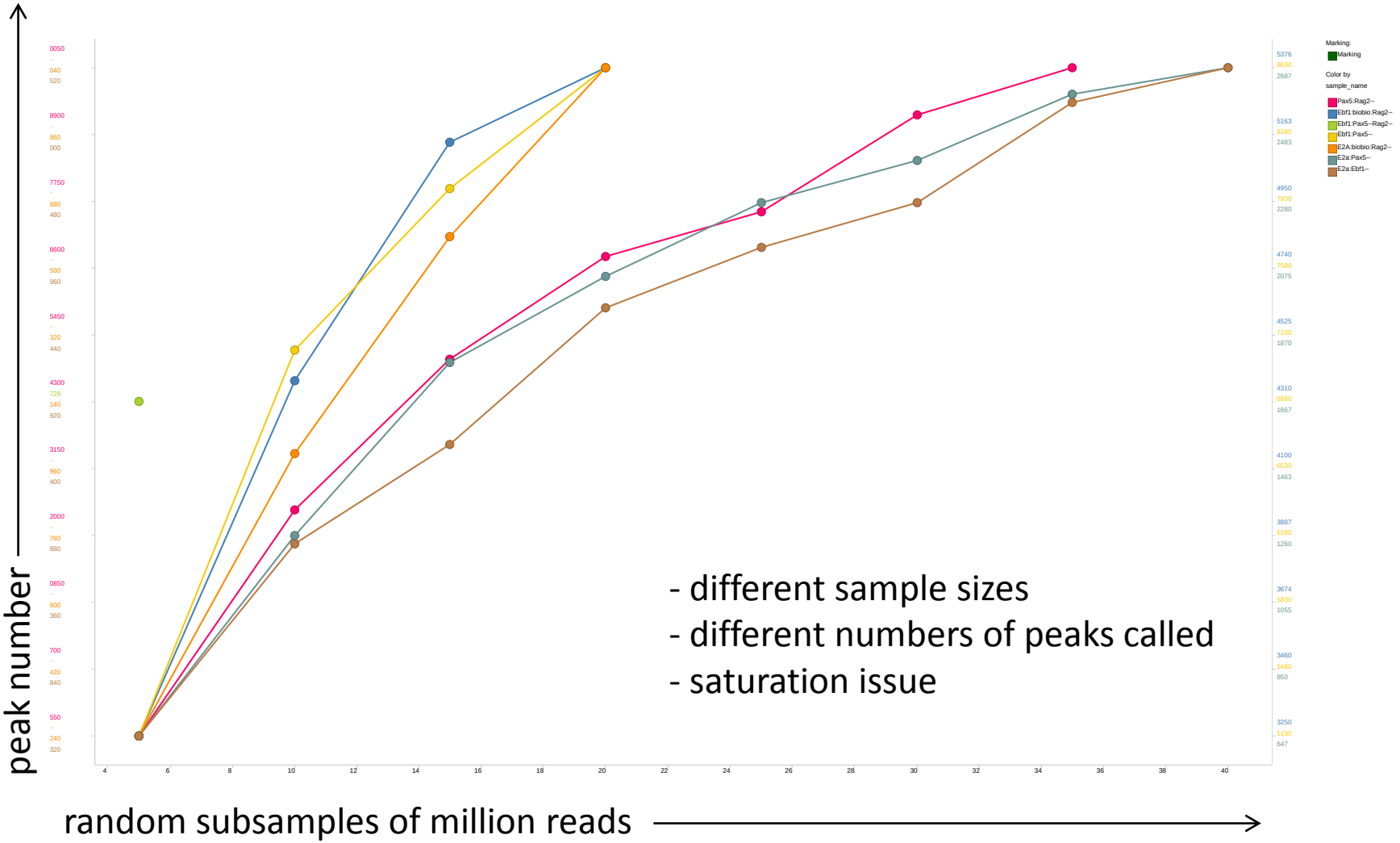



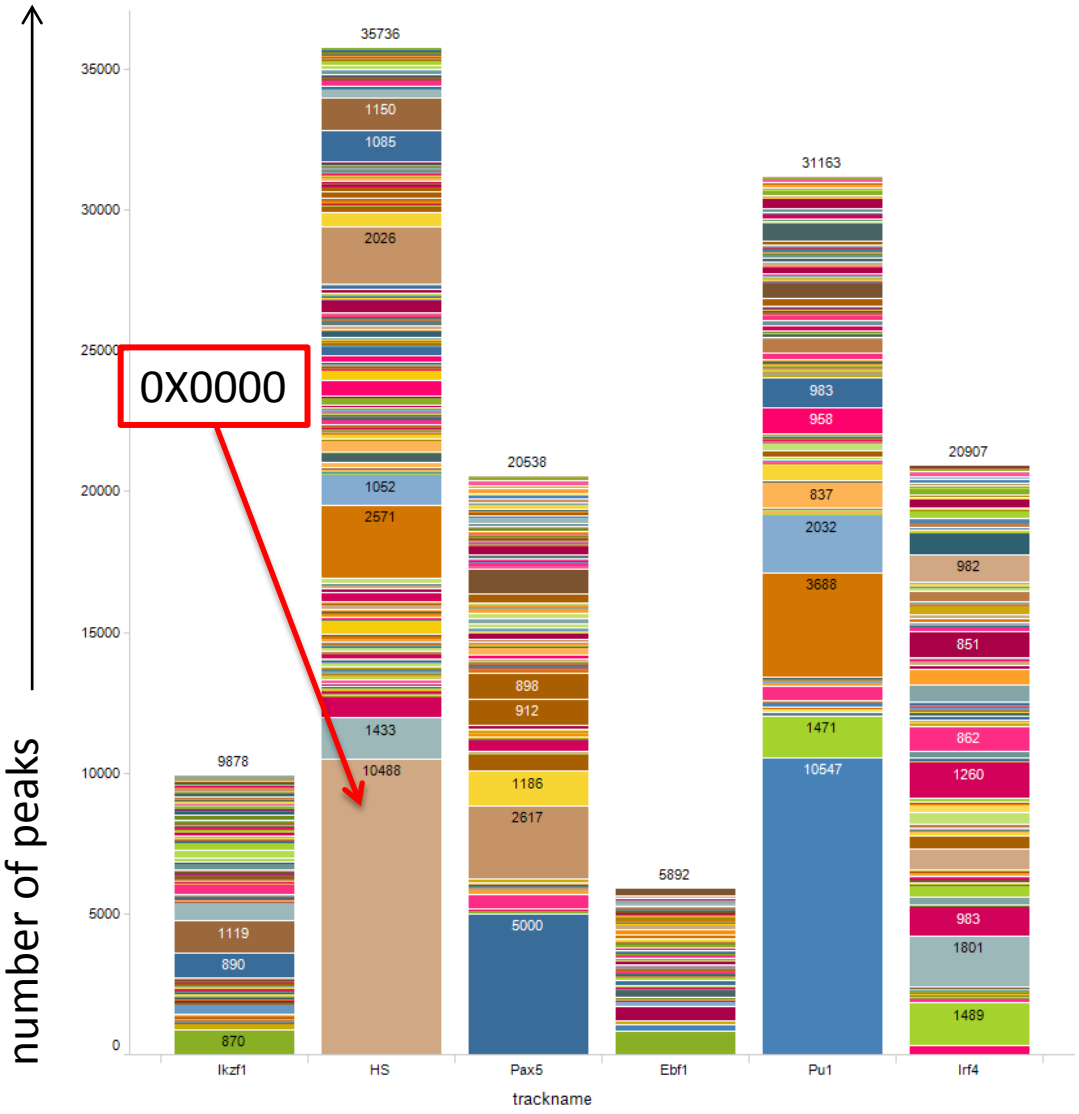



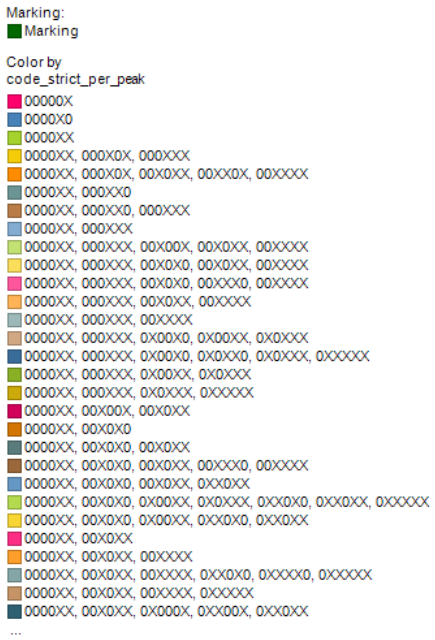



# Peak calling & sample size

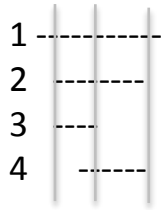


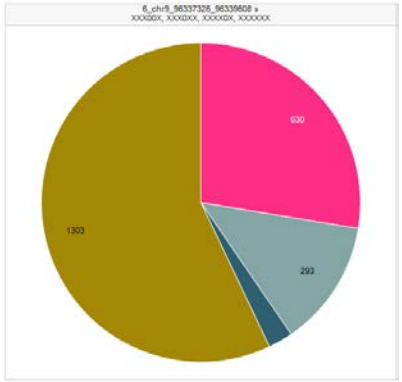
# Peak overlaps of multiple tracks



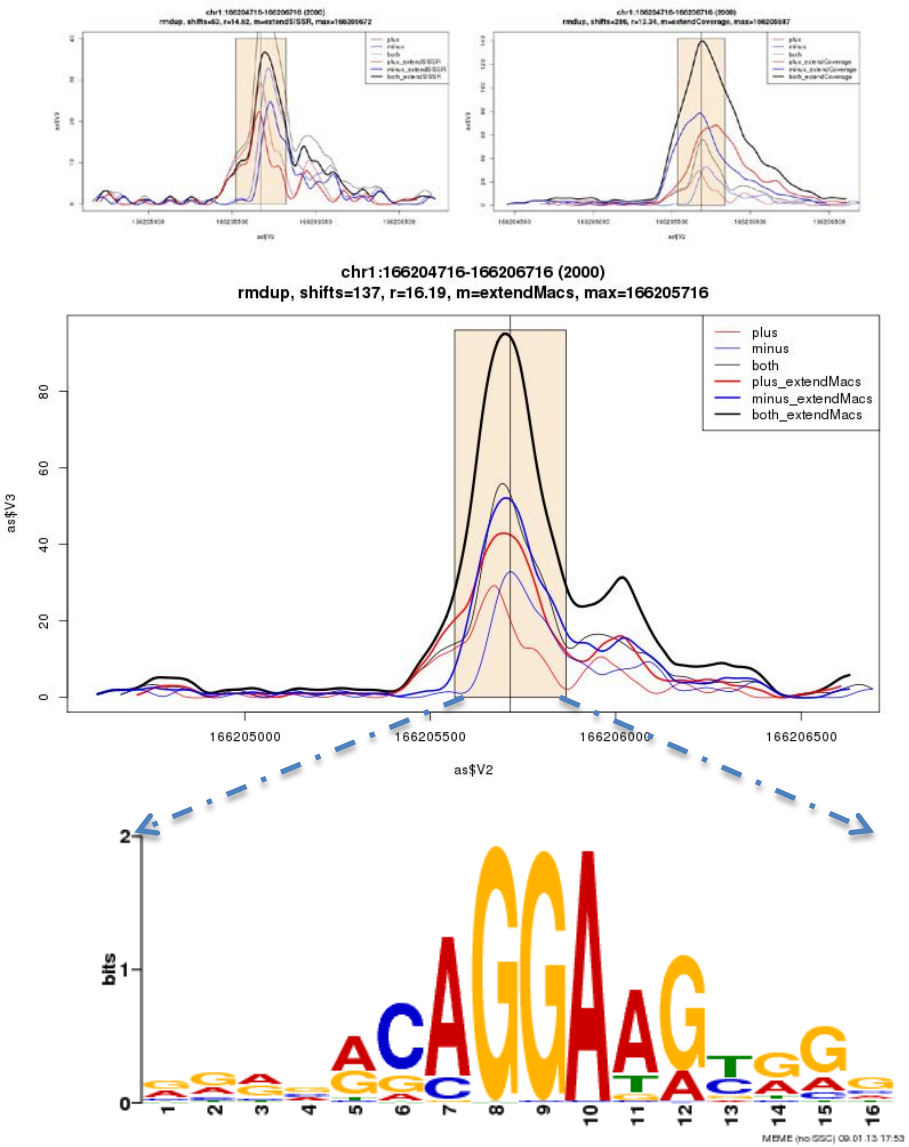




overlap example



# Finding the most common sequence motif in a peak




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], centrimo [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

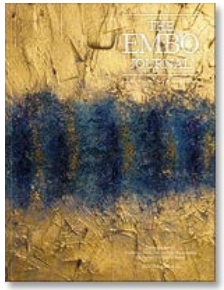
888 samples

234 sequencing requests

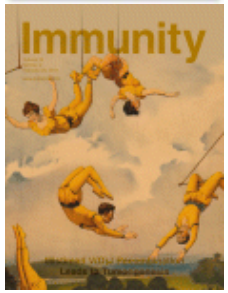
606 “lanes”

126 geno types

55 cell types

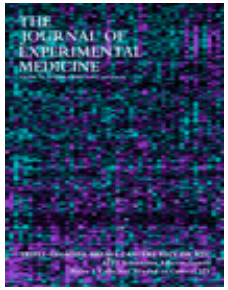


Revilla-I-Domingo R, Bilic I, Vilagos B, Tagoh H, Ebert A, Tamir IM, Smeenk L, Trupke J, Sommer A, Jaritz M, Busslinger M. EMBO J. 2012




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