



HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI



ChIP-Seq Data Analysis

Teemu Kivioja

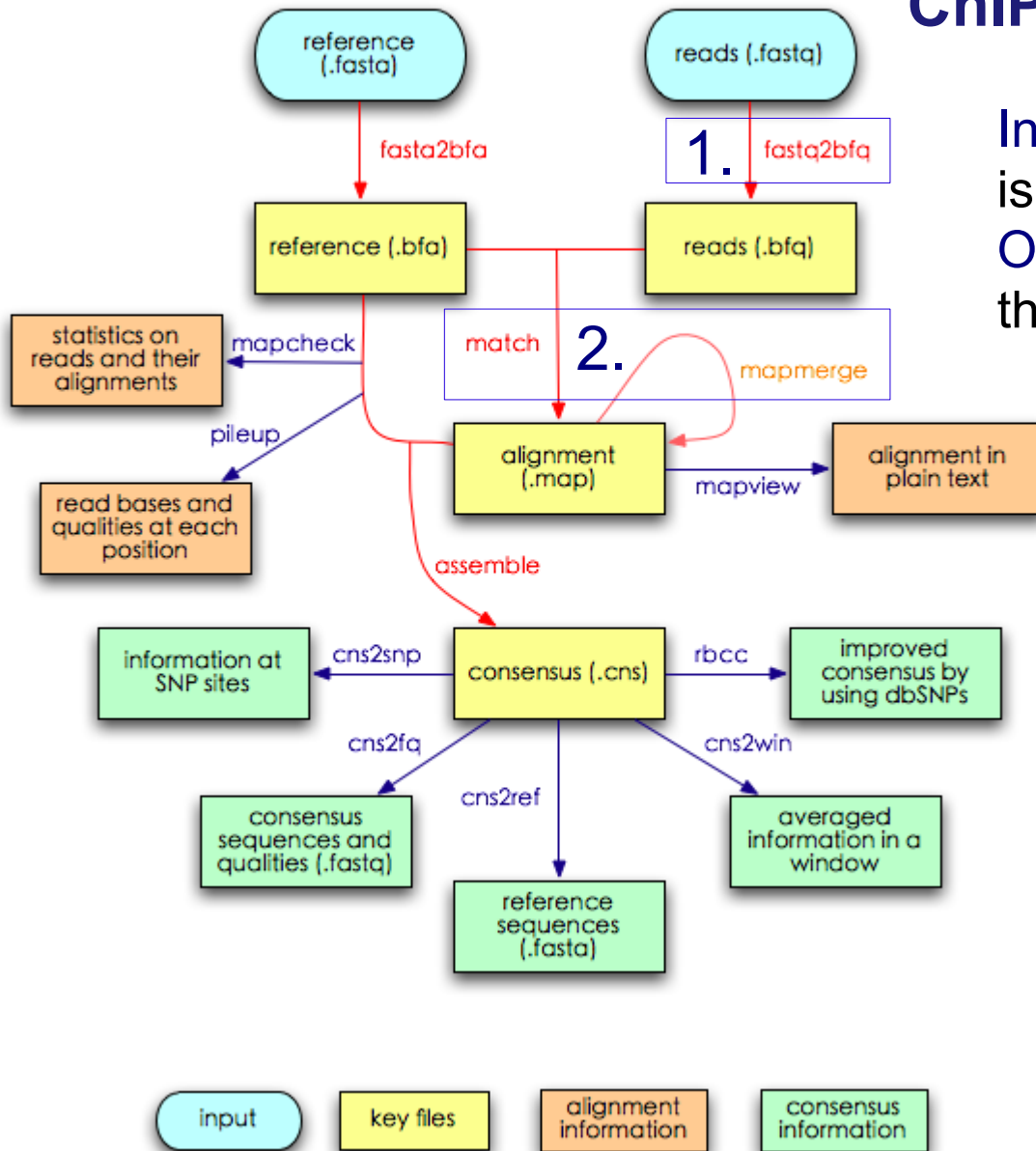
**Taipale lab, Genome-Scale Biology Program, University of Helsinki,
and KTL National Public Health Institute
Department of Computer Science, University of Helsinki**



Lab: ChIP for determining transcription factor binding sites

1. Protein complexes that contact DNA are cross-linked to their binding sites
2. The chromatin is sheared into short fragments
3. Specific DNA fraction that interacts with the protein (transcription factor, TF) of interest is isolated by immunoprecipitation.
4. A genome-wide readout of the protein binding sites is produced either by
 - Hybridization to a tiling array (Chip-chip) or by
 - End-sequencing millions of different DNA-fragments (Chip-Seq)

Mapass2 Work Flow

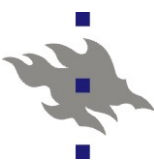


ChIP-Seq analysis work flow

Input: end reads from the isolated fragments

Output: locations in the genome that bound the TF

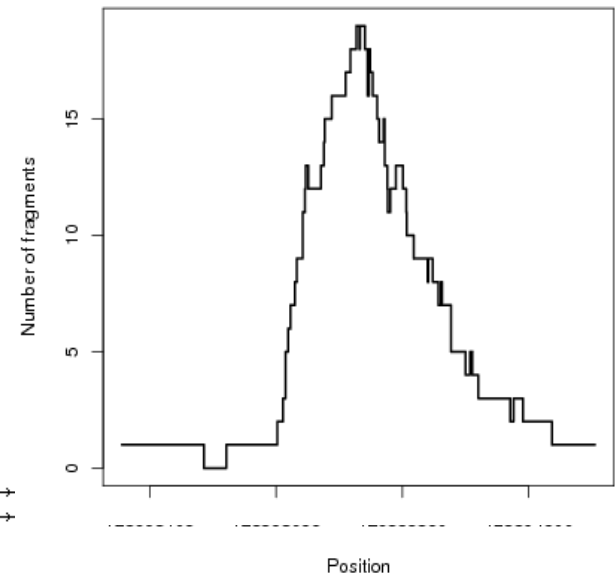
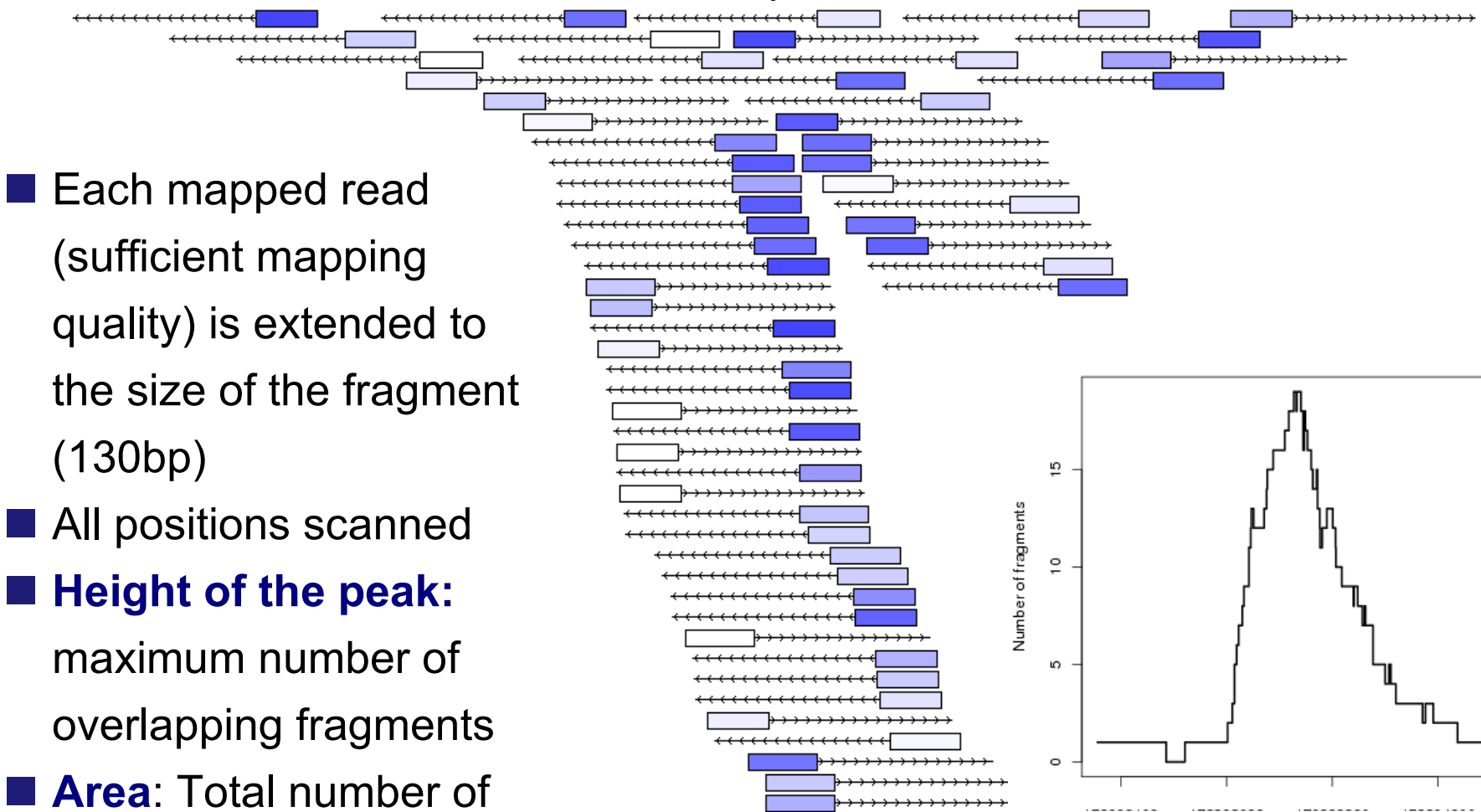
1. Get reads
 2. Align reads to the reference genome and merge lanes
 3. Detect peaks
 4. Compare to IgG control
 5. (Science: visualization, EEL, motif finding ...)
- Maq (Eland) covers steps 1 and 2.
- In-house (MACS?) covers steps 3 and 4.



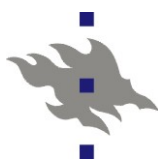
3. Detect peaks: Large number of overlapping fragments indicating possible binding site



- Each mapped read (sufficient mapping quality) is extended to the size of the fragment (130bp)
- All positions scanned
- **Height of the peak:** maximum number of overlapping fragments
- **Area:** Total number of overlapping fragments



Data by Gong-hong Wei



4. Compare peaks to the IgG control: Fragment counts and confidence, challenges

$$p(x) = \frac{e^{-\lambda} \lambda^x}{x!}$$

■ Poisson process?
Rare, independent
events, known rate?

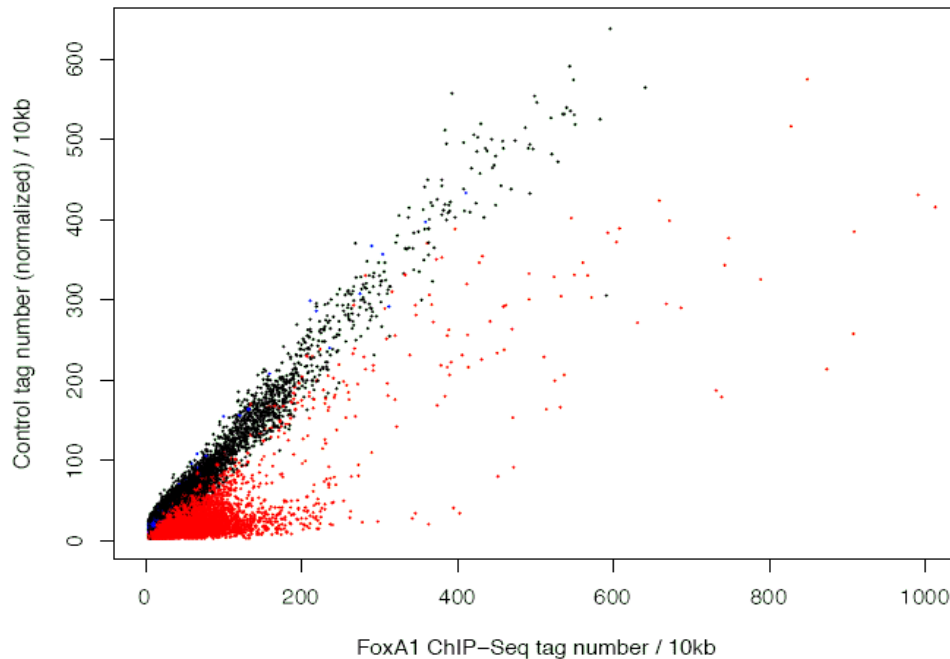


Figure: Zhang et al. *Genome Biol.* 2008, 9:R137

- The height of the peaks varies between samples
 - Antibody quality, TF expression level ...
- The background is not uniform
 - The distribution of the reads in the IgG control does not follow Poisson distribution
 - **ChIP sample and IgG control correlate!**
 - Chromatin structure, DNA-amplification and sequencing bias, genome copy number variation ...



4. Compare peaks to the IgG control: Fragment counts and confidence, solutions?

- The probability of observing x fragments in ChIP sample and y fragments in IgG control by chance?

- Challenges

- Often more ChIP reads than IgG reads: e.g. 2:1
- Low counts: e.g. 6/1 vs. 60/10 reads in ChIP/IgG (2:1)
- Is y a good local estimate of expected number of fragments λ ?

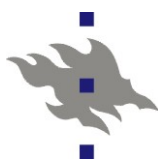
- Zhang et al. Model-Based ... *Genome Biol.* 2008, 9:R137

- Linear scaling of total read counts
- Estimate λ as max from windows of different sizes in IgG

- Audic, Claverie. The significance of digital gene expression profiles, *Genome Res.* 1997, 7, 986-995.

- Integrates over unknown λ given the observed count y
- Takes into account the uncertainty associated to low counts and the different number of ChIP sample and IgG reads

$$p(X) = \frac{e^{-\lambda} \lambda^x}{x!}$$



4. Compare peaks to IgG control: False discovery rate?

Number of FoxA1 peaks for FDR 1%

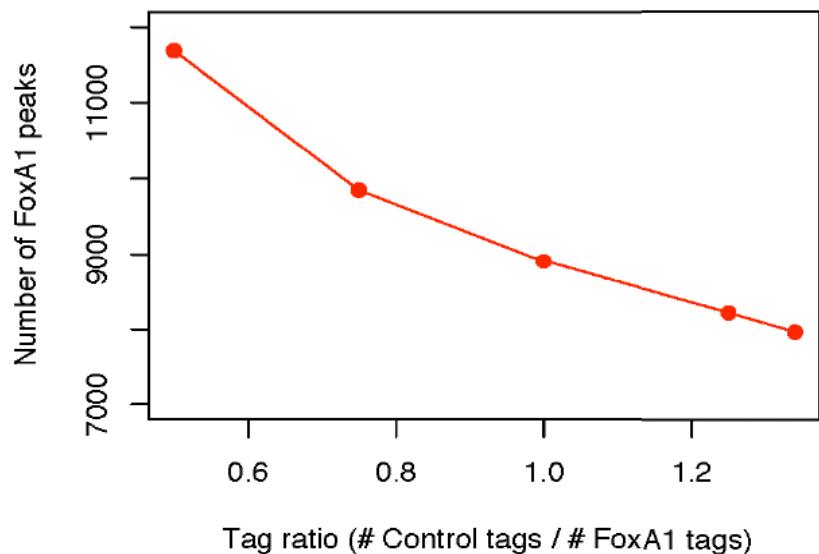


Figure: Zhang et al. *Genome biology* 2008, 9:R137

- Testing every base in the genome!
- For given p-value, how many false positives do we take?
- Sample swap (Zhang et al. *Genome Biol.* 2008)?
 - $FDR = \#(\text{IgG peaks}) / \#(\text{ChIP peaks})$
 - Problematic when the number of ChIP and IgG reads is different
 - $\#(\text{IgG reads}) \ll \#(\text{ChIP reads})$ and scaling \Rightarrow uncertainty about λ amplified



Conclusions

- Method and tool development still needed
- Open questions?
 - How much to sequence
 - Sequence effects, saturation ...
 - Resolution, number of binding sites in a peak...



Acknowledgements

■ Taipale lab

■ **Jussi Taipale**

■ Mikael Björklund

■ Martin Bonke

■ Lin Feng

■ Outi Hallikas

■ Song-Ping Li

■ Arttu Jolma

■ Sini Miettinen

■ Ritva Nurmi

■ Minna Taipale

■ **Mikko Turunen**

■ **Gong-hong Wei**

■ Jian Yan