A practical comparison of methods for detecting transcription factor binding sites in ChIP-seq experiments

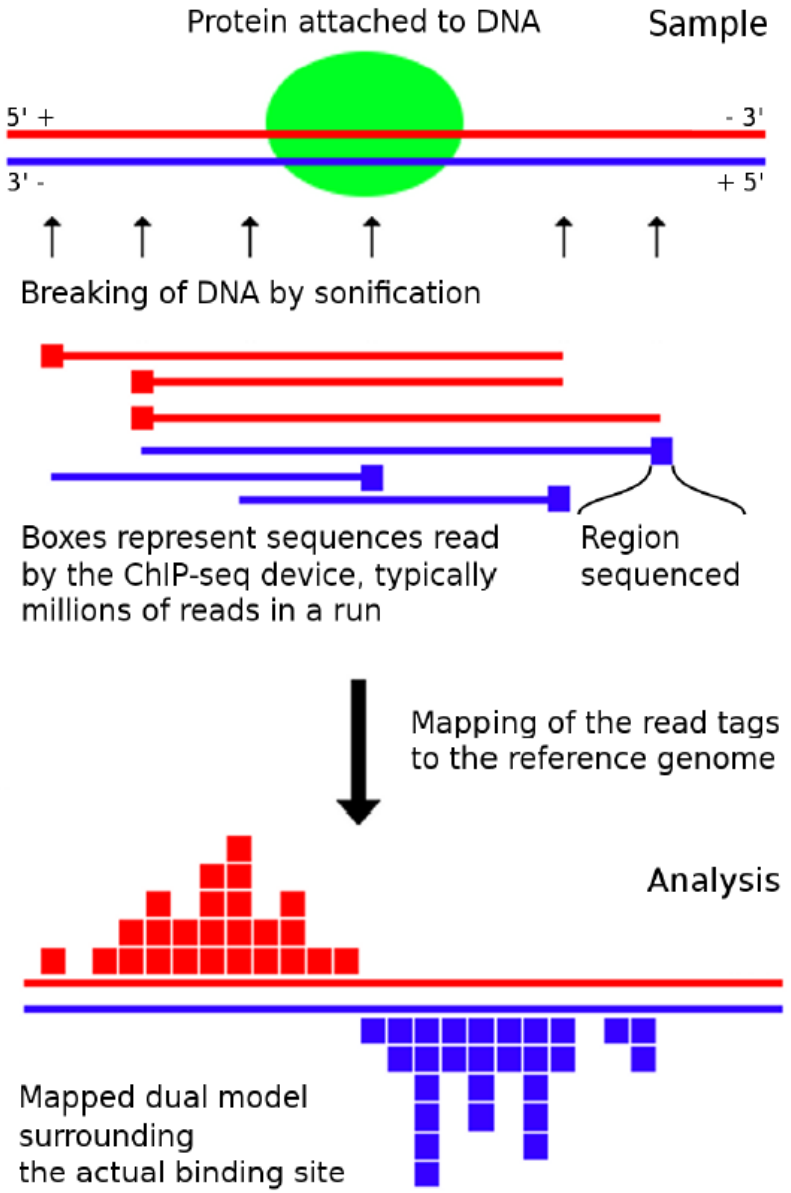
Teemu D Laajala[1], Sunil Raghav[1], Soile Tuomela[1,2], Riitta Lahesmaa[1,3*], Tero Aittokallio[1,4*] and Laura L Elo[1,4]

1 Turku Centre for Biotechnology, FI-20521 Turku, Finland
2 Turku Graduate School of Biomedical Sciences, FI-20520 Turku, Finland
3 Immune Disease Institute, Harvard Medical School, Boston, USA
4 Department of Mathematics, University of Turku, FI-20014 Turku, Finland
* Contributed equally

BMC Genomics 2009, **10**:618

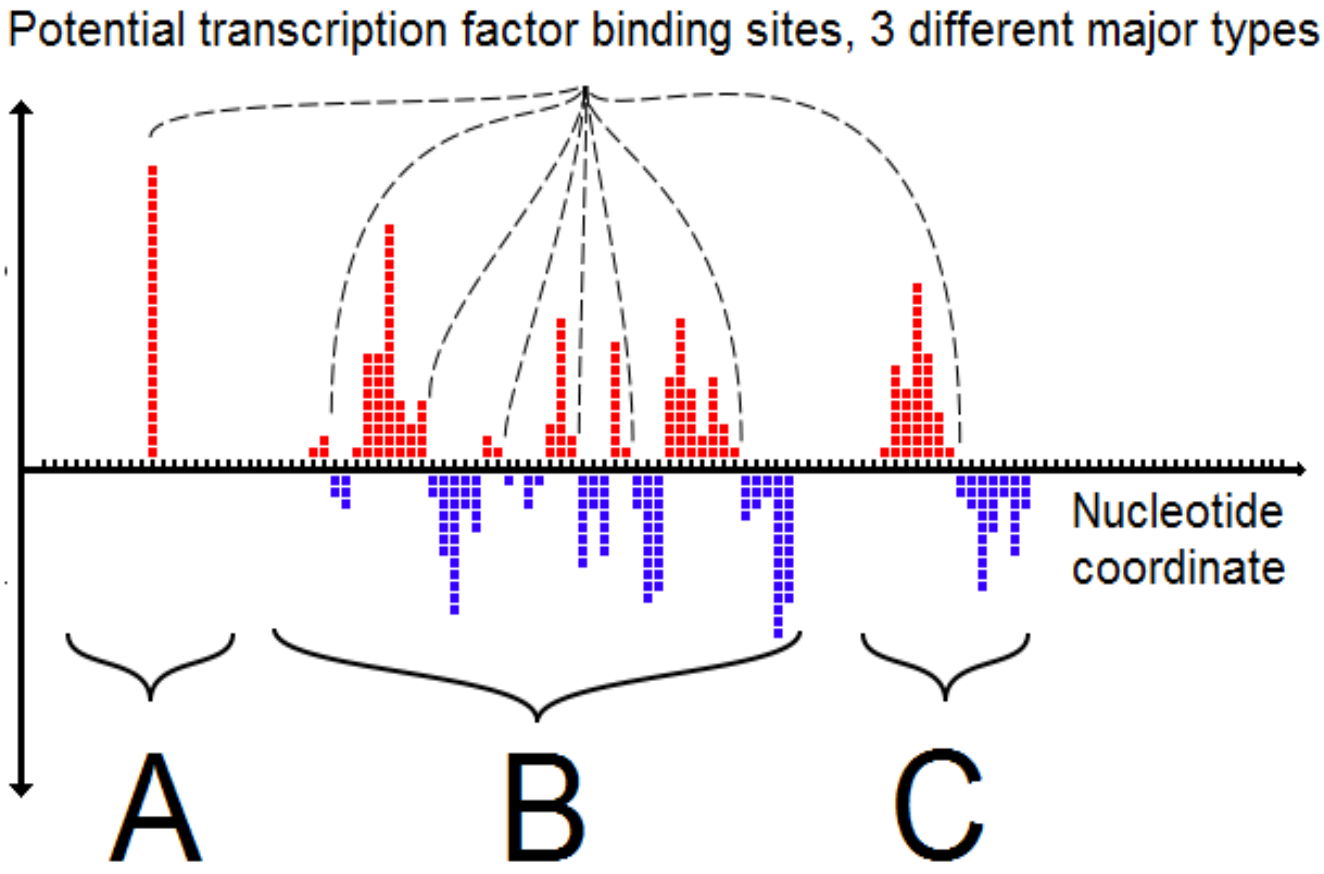
2.6.2010
Conference: Next Generation Sequencing Data Analysis

ChIP-seq process



- Wetlab
  - ChIP, sonification etc.
- Sequencing
  - Amplification, reading from ends of strands (typically ~30bp)
- Aligning to reference
  - Straightforward
- Peak detection
  - Different approaches

Proposed types of peak formations



- Type A:

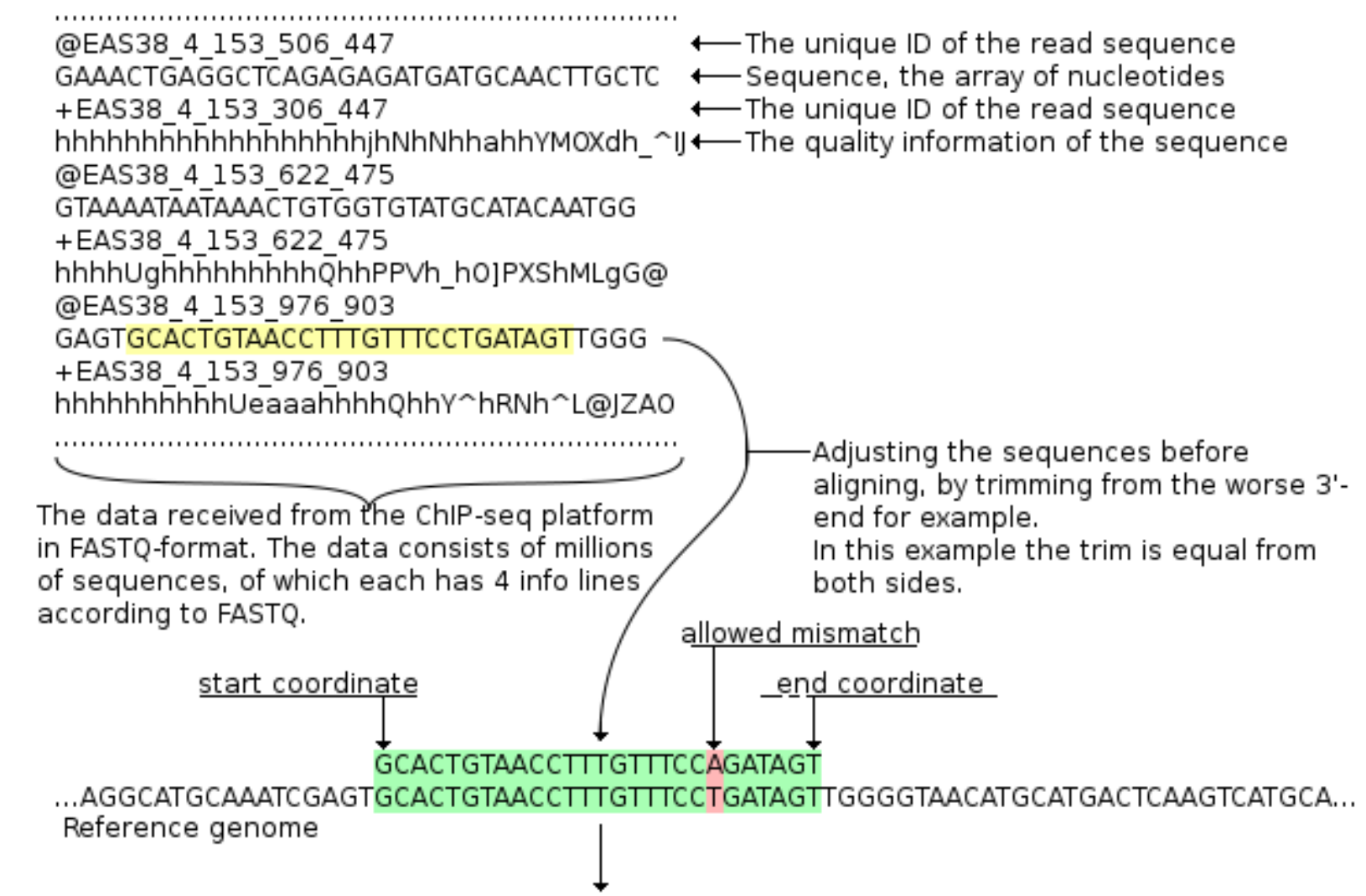
  - A probable artefact

- Type B:

  - Possibly a binding site or several of them

- Type C:
  - The ideal formation

Aligning



When all the sequences from the Illumina/Solexa output have been aligned, we get the aligned data which usually holds information of e.g. start coordinate, end coordinate, mismatch-amount et cetra.

Peak detection approaches

- 3 examples of different approaches:
  - *CisGenome* uses a fixed width window which is shifted through the genome to find enriched regions
  - *FindPeaks* first finds local maxima of aligned tags and then examines the nearby locations to find the peak region
  - *Hpeak* uses a Hidden Markov Model to examine read accumulation through two states: binding sites and background

Comparison of different methods

- 4 different datasets: Three publicly available and one in-house dataset
  - In-house dataset compensates for the fact that some of the peak finders are trained with these public datasets
- 9 different peak detection methods: *PeakFinder*, *GeneTrack*, *FindPeaks*, *SISSRs*, *QuEST*, *MACS*, *CisGenome*, *PeakSeq* and *Hpeak*
- Peak detection methods were applied with default or suggested parameters for ChIP-seq analysis

Table 2: ChIP-seq samples analysed in the present study

Sample	Cell type	Binding motif (Genomatix)	Reads (million)
NRSF	Jurkat	V$NRSF.01	2.3
Control	Jurkat	-	1.7
NRSF mono	Jurkat	V$NRSF.01	5.4
NRSF poly	Jurkat	V$NRSF.01	8.8
Control	Jurkat	-	17.4
FoxA1	MCF7	V$HNF3.01	3.9
Control	MCF7	-	5.9
STAT6	Th2 1 h	V$STAT6.01	3.0
STAT6	Th2 4 h	V$STAT6.01	2.7
STAT6	Thp	V$STAT6.01	3.2

Validation

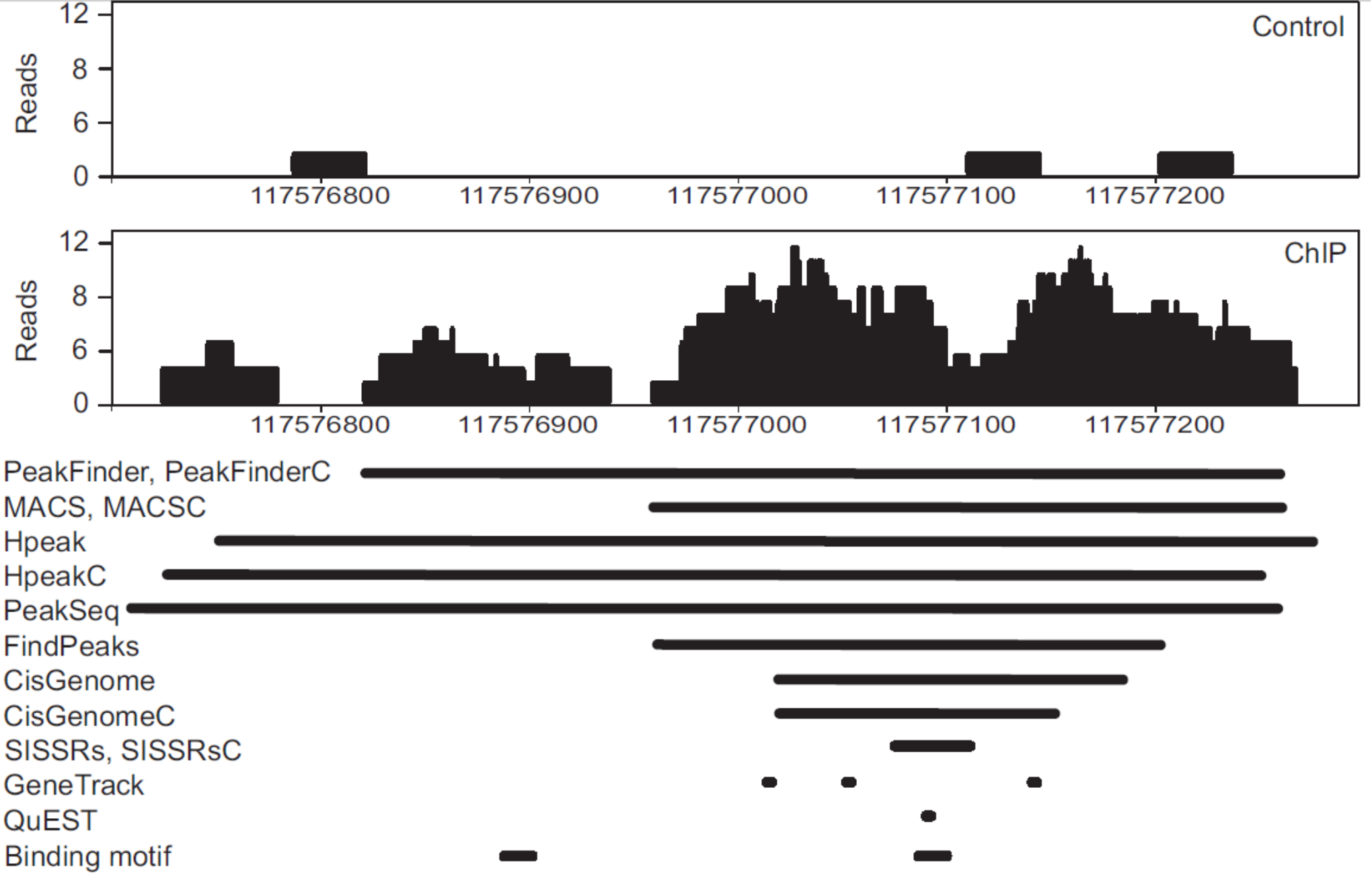
- To assess the results from the peak finders
  - The reproducibility over the 3 NRSF-samples was used as an internal validation of peak detections
  - External validation was retrieved for binding motifs from Genomatix RegionMiner service for each transcription factor separately
  - Motifs and binding regions were confirmed with independent qPCR experiments

Control vs No control

- Some peak detection methods require a control sample in order to be applicable, e.g. QuEST
- Few methods cannot incorporate control data, e.g. GeneTrack
- Most of the methods offer two variants depending on if there is a control sample available
- *Peak detection methods systematically had better results when applied with a control sample*

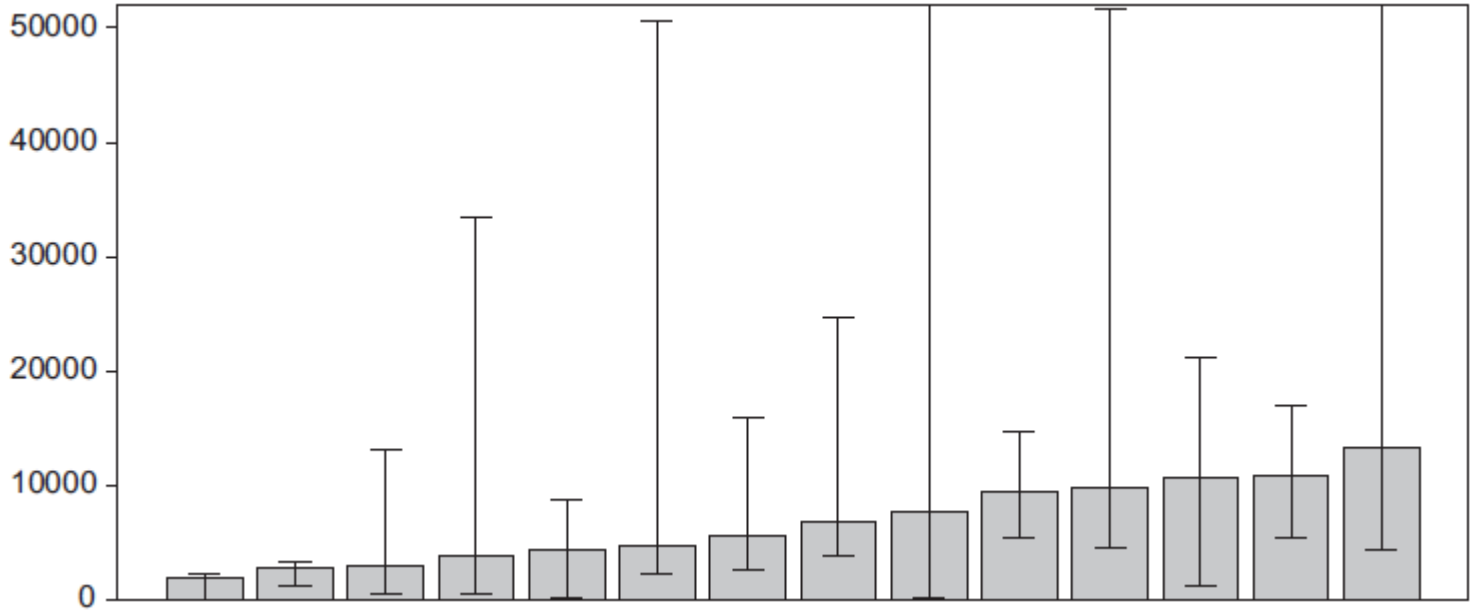
Algorithm	Availability	Type	Background model
PeakFinder 2.0.1	http://woldlab.caltech.edu/html/chipseq_peak_finder	S, C	none
GeneTrack 1.0.1	http://code.google.com/p/genetrack/	S	none
FindPeaks 3.1.9.2	http://www.bcgsc.ca/platform/bioinfo/software/findpeaks/	S	uniform
SISSRs v1.4	http://sissrs.rajajothi.com/	S, C	Poisson/ control sample
QuEST 1.0	http://mendel.stanford.edu/sidowlab/downloads/quest/	C	control sample
MACS 1.3	http://liulab.dfci.harvard.edu/MACS/	S, C	local Poisson/ control sample
CisGenome v1	http://www.biostat.jhsph.edu/~hji/cisgenome/	S, C	negative binomial/ control sample (binomial)
PeakSeq v1.01	http://www.gersteinlab.org/proj/PeakSeq/	C	local Poisson and control sample (binomial)
Hpeak 1.1	http://www.sph.umich.edu/csg/qin/HPeak/	S, C	hidden Markov model

The column Type indicates whether the method is applicable to a single sample analysis (S) or a two-sample analysis involving a control sample (C).

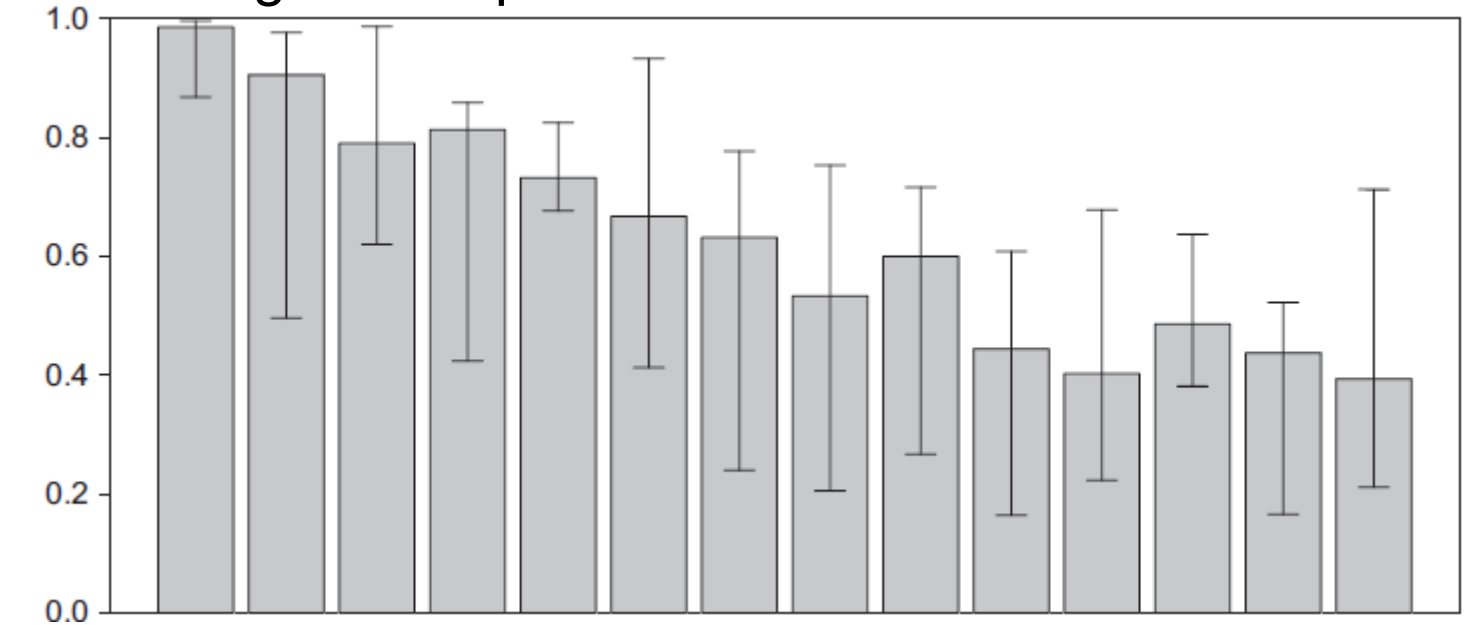


An example region identified as a STAT6 binding site at 1 h after polarization with IL4. The same region was identified as a STAT6 binding site with all the fourteen peak detection approaches applied in the present study. The number of overlapping reads (*y*-axis) is shown at each genomic position (*x*-axis). The horizontal bars below the profile illustrate the detected binding regions, as well as the high-scoring STAT6 binding motifs as determined using the Genomatix MatInspector tool.

Number of peaks detected



Average overlap with other methods



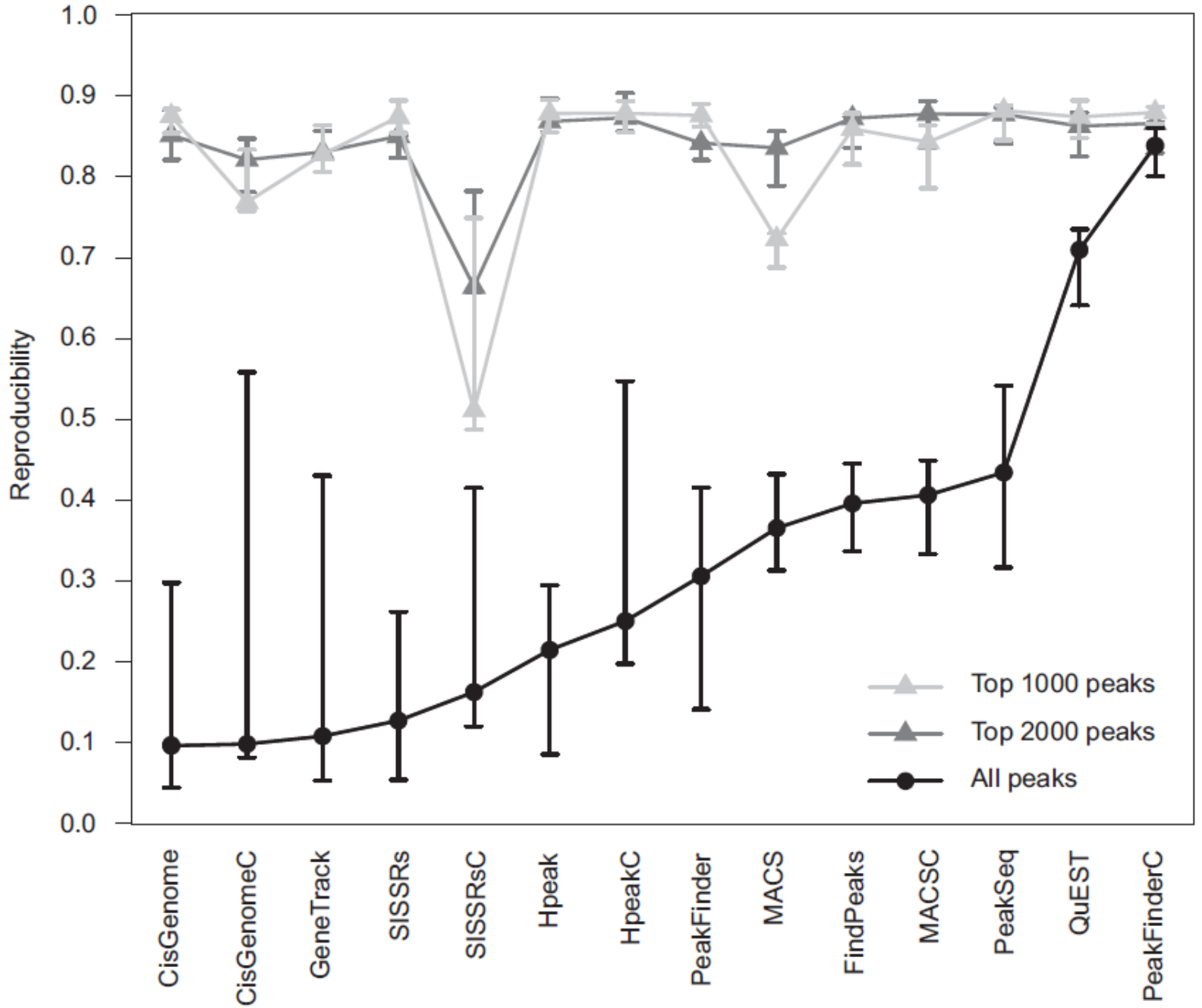
Methods
(from left to right):

- PeakFinder (Control)
- QuEST
- PeakFinder
- CisGenome (Control)
- PeakSeq
- GeneTrack
- Hpeak (Control)
- SISSRs (Control)
- CisGenome
- MACS (Control)
- Hpeak
- FindPeaks
- MACS
- SISSRs

Bar is median with error bar indicating maximum and minimum

Amount of peaks detected

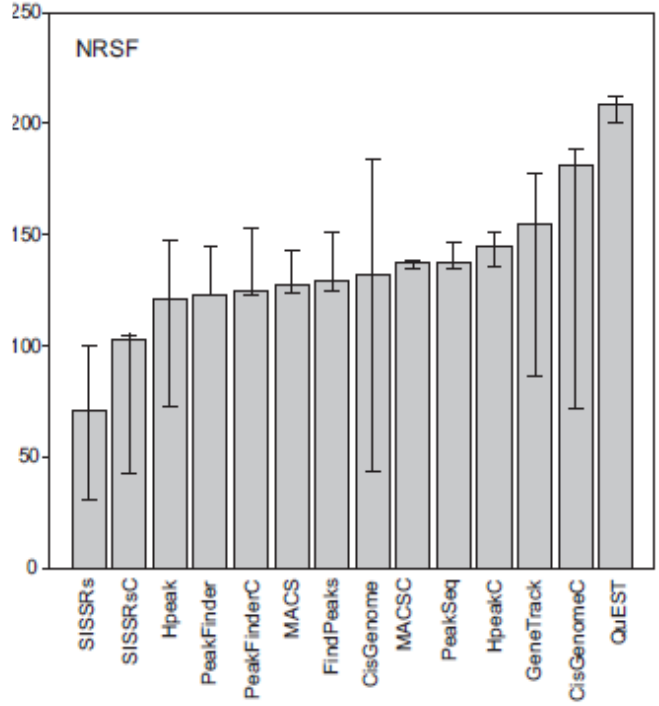
- It should be considered whether we're interested in
  - Finding as many candidate binding sites as possible
    - A lot of variation among the performance of different methods, one reason is the high amount of peaks detected
  - More interested in few, more exact (narrow region) candidate sites
    - When considering only a smaller subset of the best found candidate regions, all the methods performed fairly well in all datasets

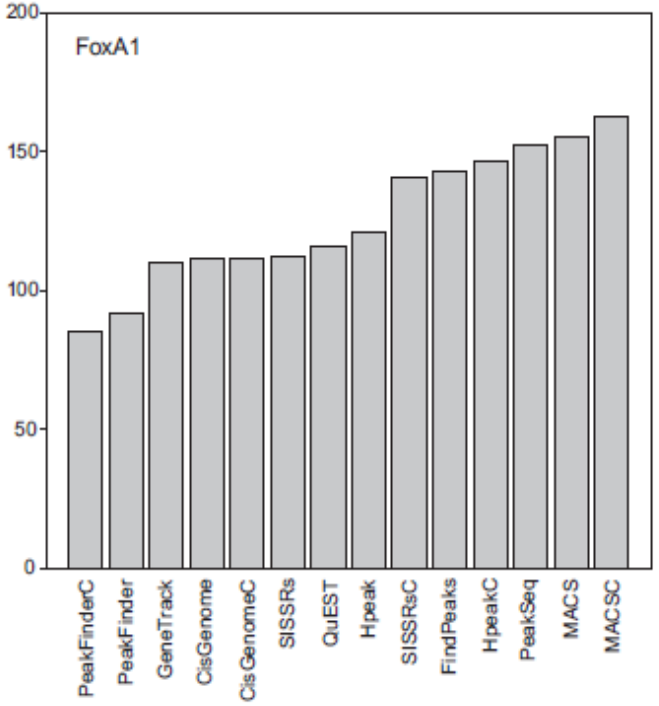


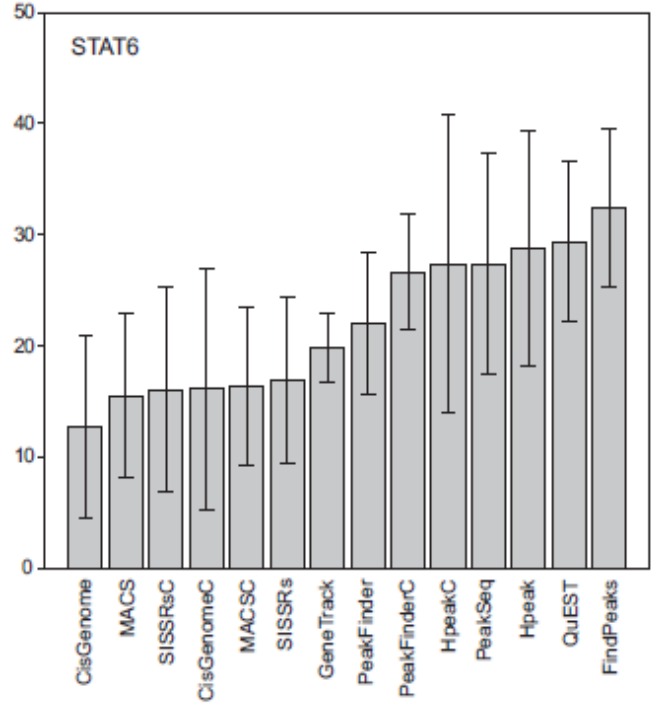
Reproducibility of the detections across the three NRSF samples. With each method, the reproducibility was determined by first creating a union set of the detected regions and then assessing which of these regions were specific to only one of the samples under comparison and which were detected in both samples. The median reproducibility is shown together with the minimum and maximum values (error bars).

External validation of the predicted binding sites using binding motifs (Motif significance, z-score)

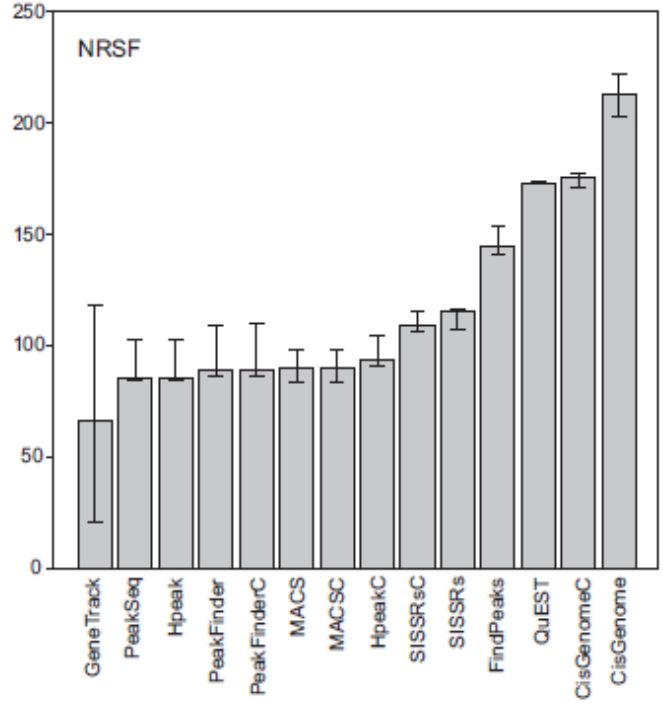
All peaks

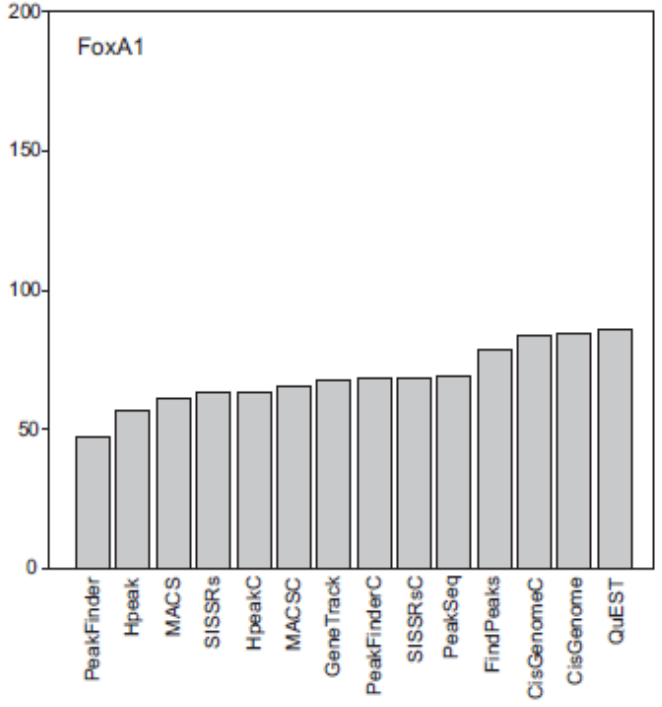


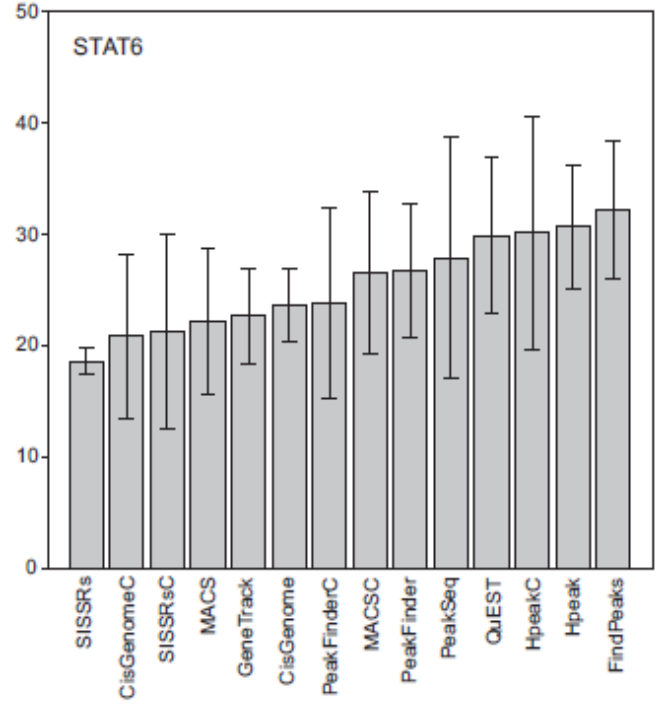




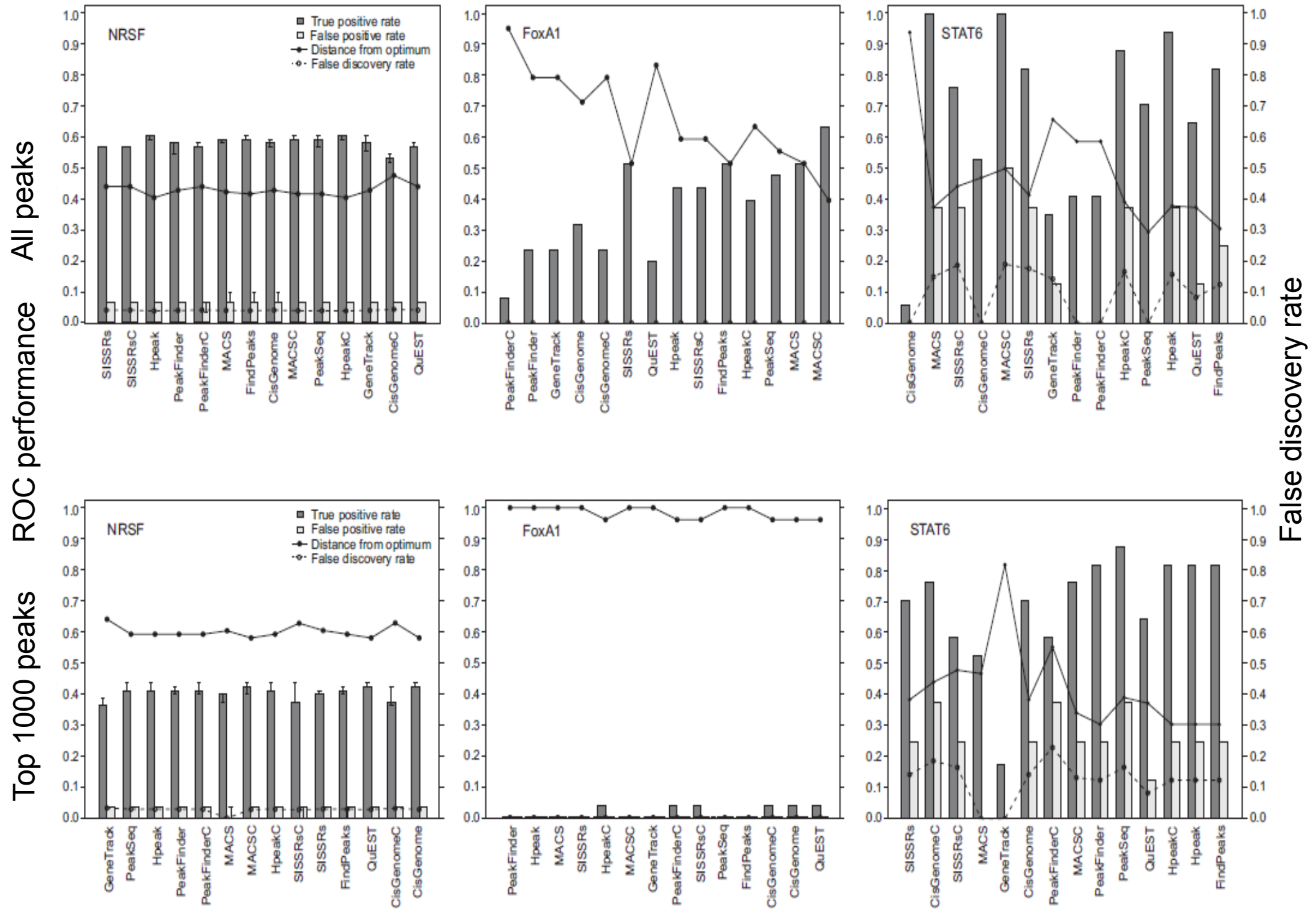
Top 1000 peaks





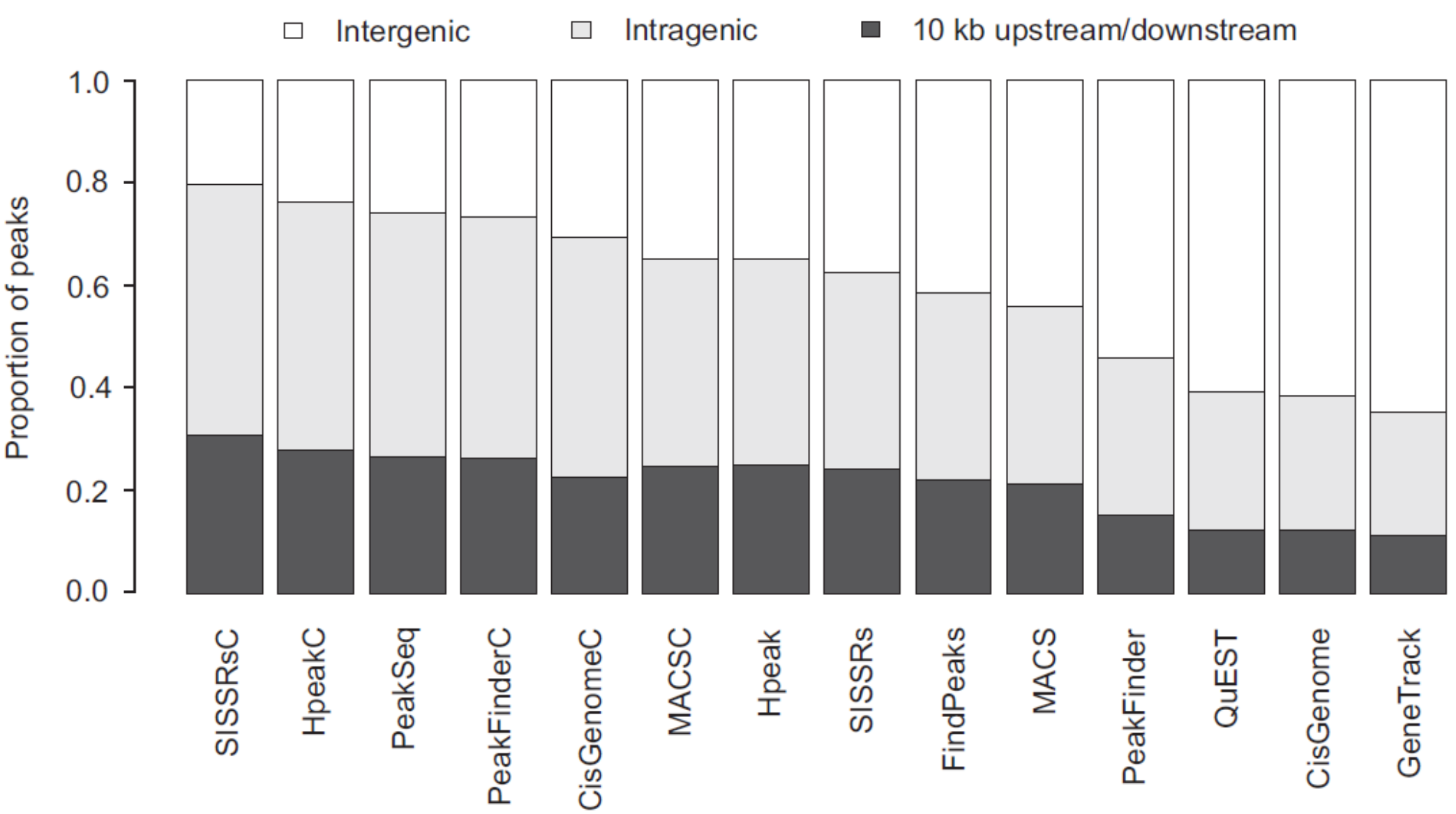


External validation of the predicted binding sites using qPCR



Assessment

- The best choice for a peak detector depended strongly on the dataset under analysis
  - QuEST performed well in the NRSF data
  - MACS was a better choice for the FoxA1 data
  - FindPeaks showed best performance on the STAT6 data
- How to choose a suitable threshold to select out the peaks that indicate the true binding sites
  - Still the method has to be able to find enough candidates to find the regions (e.g. PeakFinder had very high reproducibility, but low motif significance z-score)



A representative example demonstrating how biological conclusions may change when different algorithms are applied. The physical distribution of the binding sites in the STAT6 data is shown at 1 h after polarization with IL4. The binding sites were divided into three categories: 10 kb upstream/downstream of a transcription start/end site, within a gene (intragenic), or over 10 kb from a gene (intergenic). The proportion of binding sites in each category is indicated by the colours.

Conclusion

- Most prominent peaks were typically detected robustly
  - Reproducibilities with the top 1000 and 2000 peaks were typically 80%-90%
- All the algorithms identified binding sites with highly significant overlap with the corresponding known sequence motif
  - The external Genomatix binding site assessment
- However the choice of a peak detector can alter the final biological conclusions
  - Proportion of inter-/intragenic and up-/downstream peaks
  - ROC-curve and FDR performance

Summary

- Peak detection is the stage in ChIP-seq data analysis that has the strongest impact on conclusions made in the end
- All the different methods perform well when only the best subset of peaks is considered, but problems occur when we're trying to detect all possible binding sites
- For the time being no algorithm (of the ones presented here) clearly outperforms the others
- Future problem: how to select the optimal algorithm and its parameters for a given dataset