Comparison of ChIP-seq peak detectors

Teemu Daniel Laajala

University of Turku

September 18, 2014

My talk will focus on two papers we have published on ChIP-seq peak callers:

Laajala TD, Raghav S, Tuomela S, Lahesmaa R, Aittokallio T, Elo LL. *A practical comparison of methods for detecting transcription factor binding sites in ChIP-seq experiments.* BMC Genomics. 2009 Dec 18;10:618. doi: 10.1186/1471-2164-10-618.

Here, we systematically compared existing peak detectors at the time (2009) in terms of their features and performance on 4 varying data sets.

Elo LL, Kallio A, **Laajala TD**, Hawkins RD, Korpelainen E, Aittokallio T. *Optimized detection of transcription factor-binding sites in ChIP-seq experiments.* Nucleic Acids Res. 2012 Jan;40(1):e1. doi: 10.1093/nar/gkr839.

Here, we proposed a meta-analysis method (called *peakROTS*) based on Elo's ROTS-statistic for optimizing readily available peak detectors.

Research article

Open Access

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]

Address: [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 and [4]Department of Mathematics, University of Turku, FI-20014 Turku, Finland

Email: Teemu D Laajala - tlaajala@cc.hut.fi; Sunil Raghav - sunil.raghav@btk.fi; Soile Tuomela - soile.tuomela@btk.fi; Riitta Lahesmaa - riitta.lahesmaa@btk.fi; Tero Aittokallio - tero.aittokallio@utu.fi; Laura L Elo* - laura.elo@utu.fi

* Corresponding author †Equal contributors

Published: 18 December 2009

BMC Genomics 2009, **10**:618 doi:10.1186/1471-2164-10-618

Received: 5 June 2009
Accepted: 18 December 2009

This article is available from: http://www.biomedcentral.com/1471-2164/10/618

Abstract

Background: Chromatin immunoprecipitation coupled with massively parallel sequencing (ChIP-seq) is increasingly being applied to study transcriptional regulation on a genome-wide scale. While numerous algorithms have recently been proposed for analysing the large ChIP-seq datasets, their relative merits and potential limitations remain unclear in practical applications.

Results: The present study compares the state-of-the-art algorithms for detecting transcription factor binding sites in four diverse ChIP-seq datasets under a variety of practical research settings. First, we demonstrate how the biological conclusions may change dramatically when the different

Table 1: Compared peak detectors

Table 1: Peak detection algorithms investigated in the present study

Algorithm	Availability	Reference	Type	Background model
PeakFinder 2.0.1	http://woldlab.caltech.edu/html/chipseq_peak_finder	[2]	S, C	none
GeneTrack 1.0.1	http://code.google.com/p/genetrack/	[9]	S	none
FindPeaks 3.1.9.2	http://www.bcgsc.ca/platform/bioinfo/software/findpeaks/	[6]	S	uniform
SISSRs v1.4	http://sissrs.rajajothi.com/	[11]	S, C	Poisson/ control sample
QuEST 1.0	http://mendel.stanford.edu/sidowlab/downloads/quest/	[8]	C	control sample
MACS 1.3	http://liulab.dfci.harvard.edu/MACS/	[10]	S, C	local Poisson/ control sample
CisGenome v1	http://www.biostat.jhsph.edu/~hji/cisgenome/	[5]	S, C	negative binomial/ control sample (binomial)
PeakSeq v1.01	http://www.gersteinlab.org/proj/PeakSeq/	[12]	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).

MACS is available in CSC's Chipster-tool, F-seq not included here

Table 2: Datasets

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

Sample	Cell type	Binding motif (Genomatix)	Reads (million)	Reference
NRSF	Jurkat	V$NRSF.01	2.3	[2]
Control	Jurkat	-	1.7	[2]
NRSF mono	Jurkat	V$NRSF.01	5.4	[8]
NRSF poly	Jurkat	V$NRSF.01	8.8	[8]
Control	Jurkat	-	17.4	[8]
FoxA1	MCF7	V$HNF3.01	3.9	[10]
Control	MCF7	-	5.9	[10]
STAT6	Th2 1 h	V$STAT6.01	3.0	Elo *et al.* (unpublished)
STAT6	Th2 4 h	V$STAT6.01	2.7	Elo *et al.* (unpublished)
STAT6	Thp	V$STAT6.01	3.2	Elo *et al.* (unpublished)

Jurkat: Human T lymphocite; MCF7: Human BCa; Th: Human T helper

Figure 1: Example calls

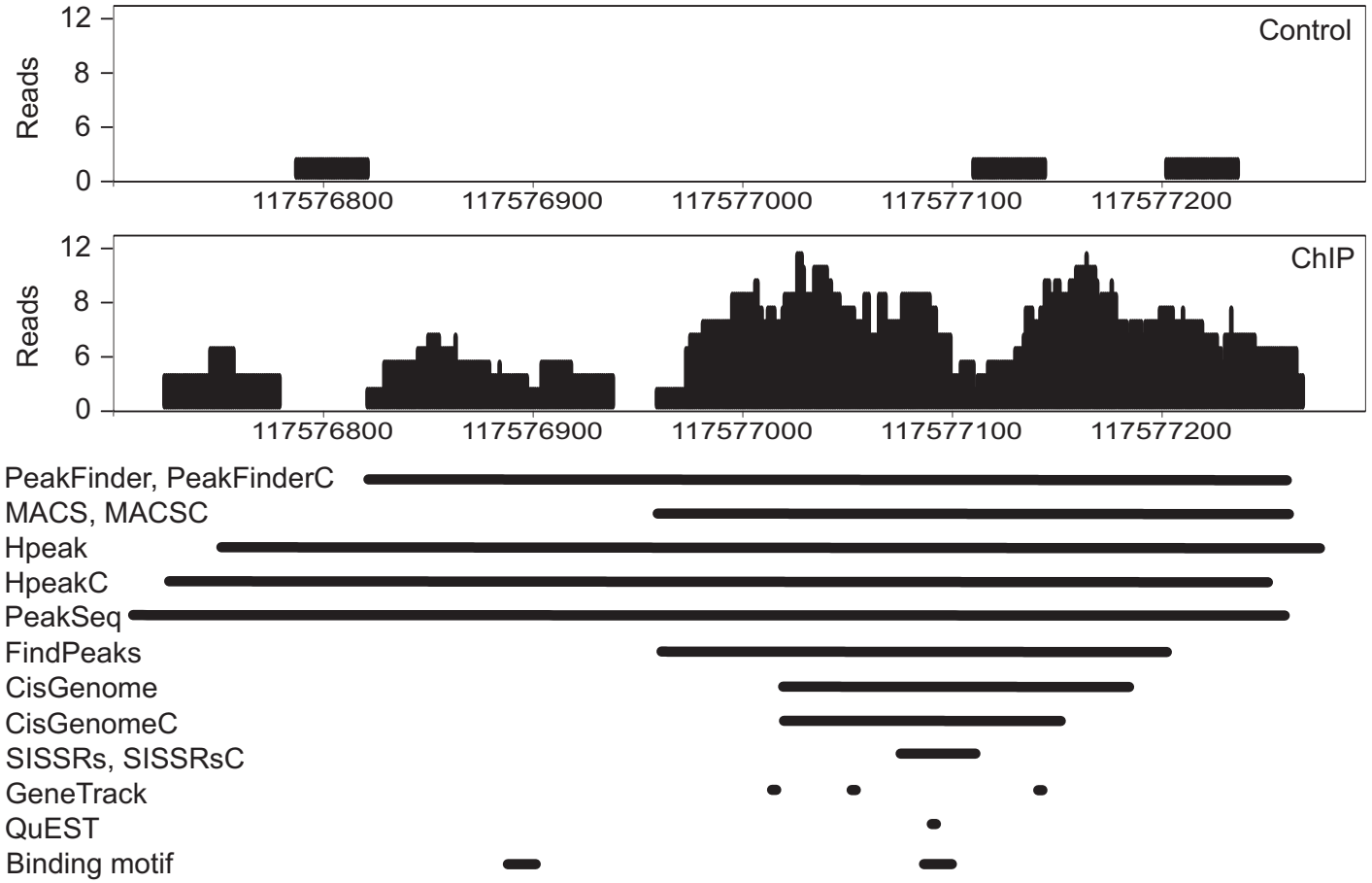


Figure 1
An example region identified as a STAT6 binding site at 1 h after polarization with IL4. The same region was identified as a STAT6 binding site with all the fourteen peak detection approaches applied in the present study. The number of overlapping reads (*y*-axis) is shown at each genomic position (*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.

Figure 2: Number of peaks and overlap (first half)

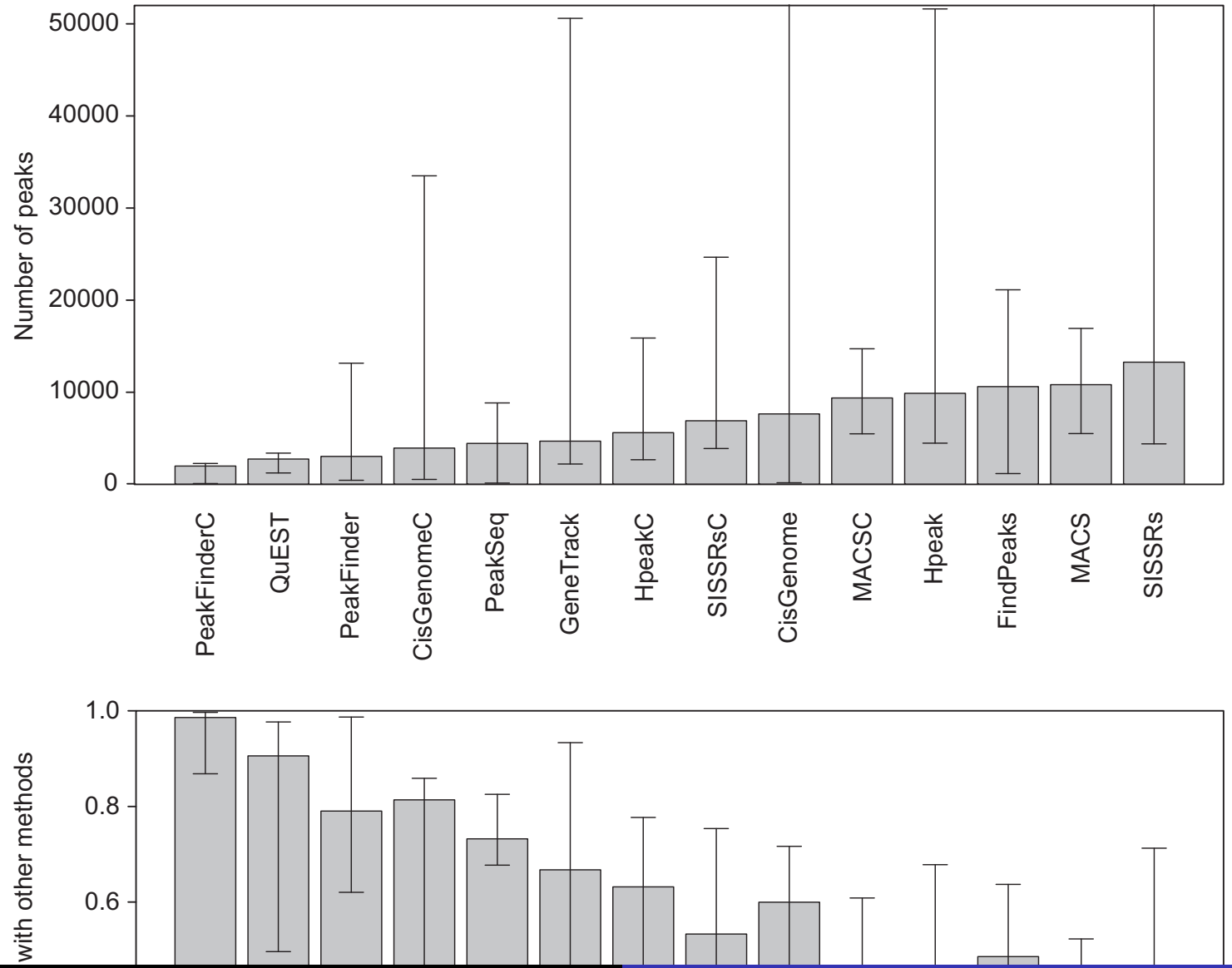


Figure 2: Number of peaks and overlap (continued)

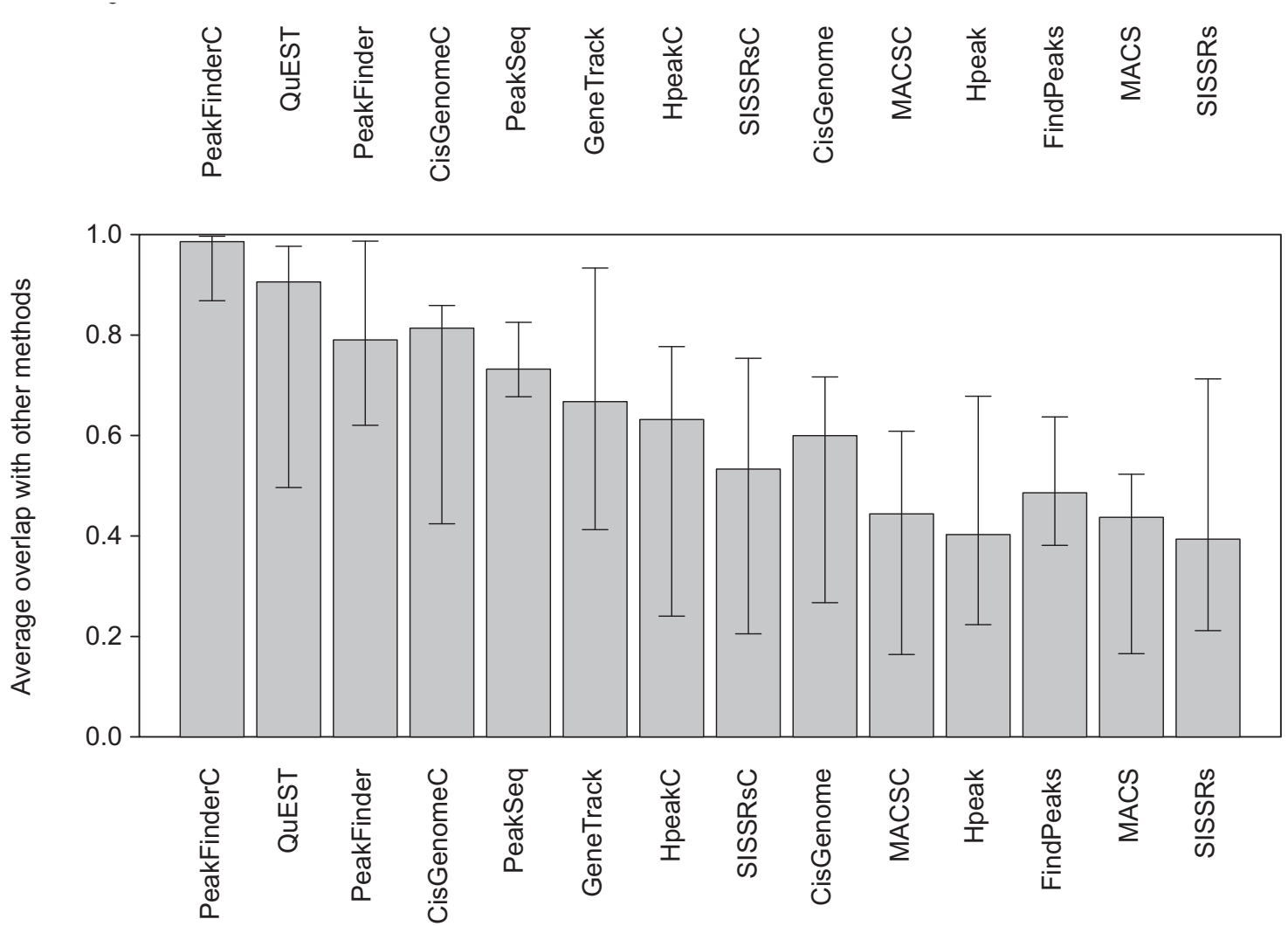


Figure 2
Numbers and overlaps of the detected peaks. The upper panel shows the median number of detected peaks across the different ChIP samples and the corresponding minimum and maximum values (error bars). For the clarity of illustration, the maximum values with SISSRs (78634) and CisGenome (78551) are cut out from the figure. The lower panel illustrates the overlap of the detections with a particular method as compared to all the other methods. The median percentage of overlapping peaks is shown together with the minimum and maximum values (error bars).

Figure 3: Locations of the called peaks

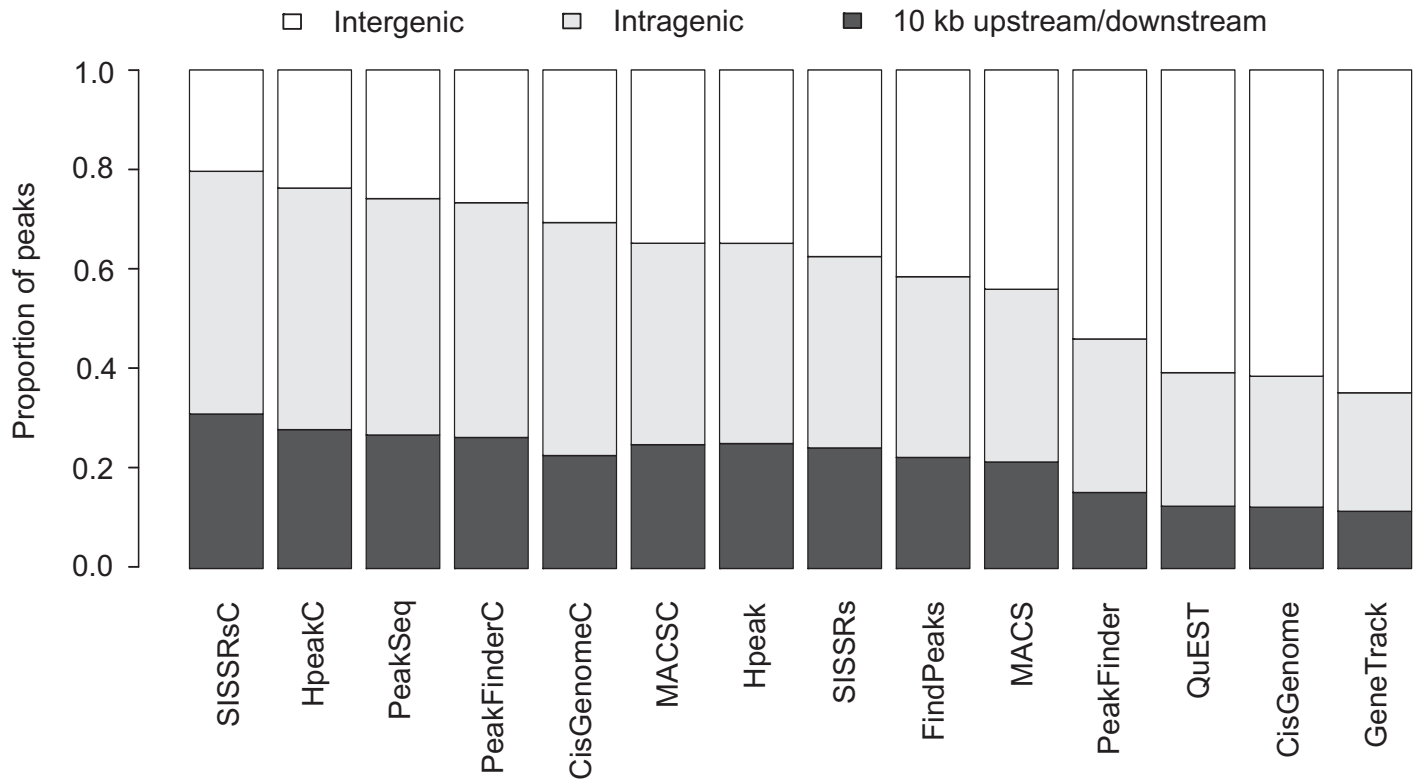


Figure 3
A representative example demonstrating how biological conclusions may change when different algorithms are applied. The physical distribution of the binding sites in the STAT6 data is shown at 1 h after polarization with IL4. The binding sites were divided into three categories: 10 kb upstream/downstream 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.

Figure 4: Reproducibility in the NRSF samples

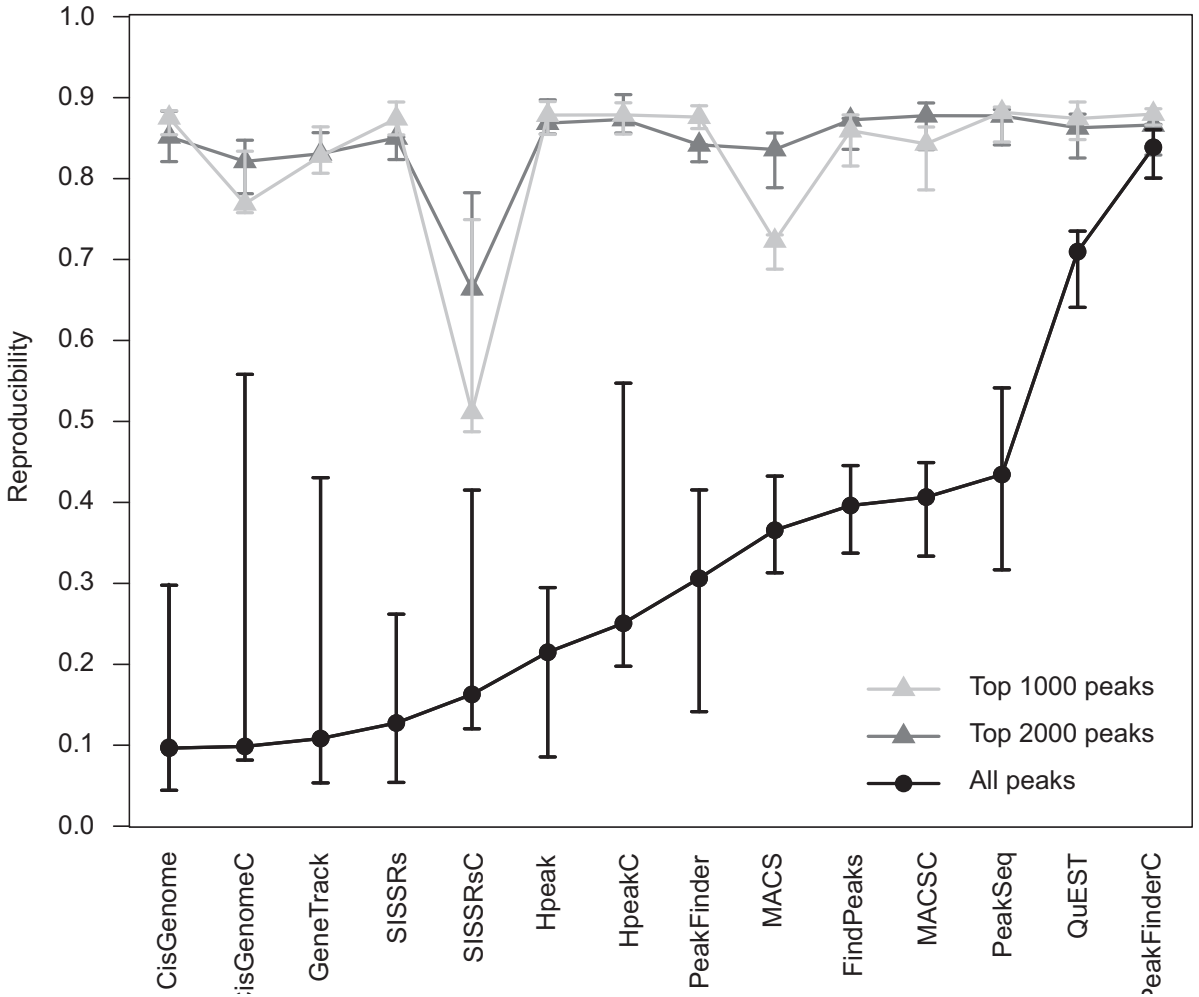


Figure 5: Known motifs (upper; all peaks)

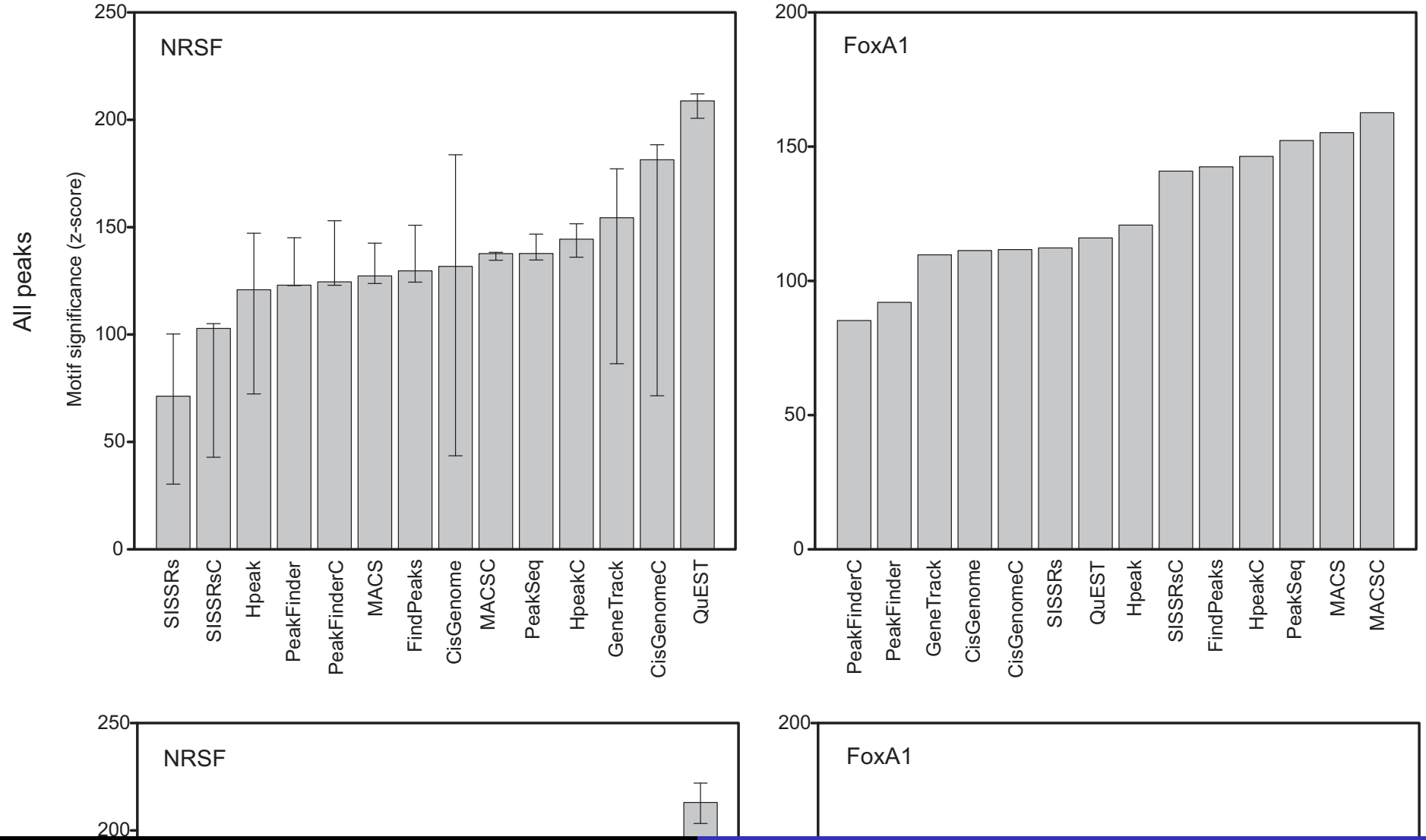


Figure 5: Known motifs (lower; top peaks)

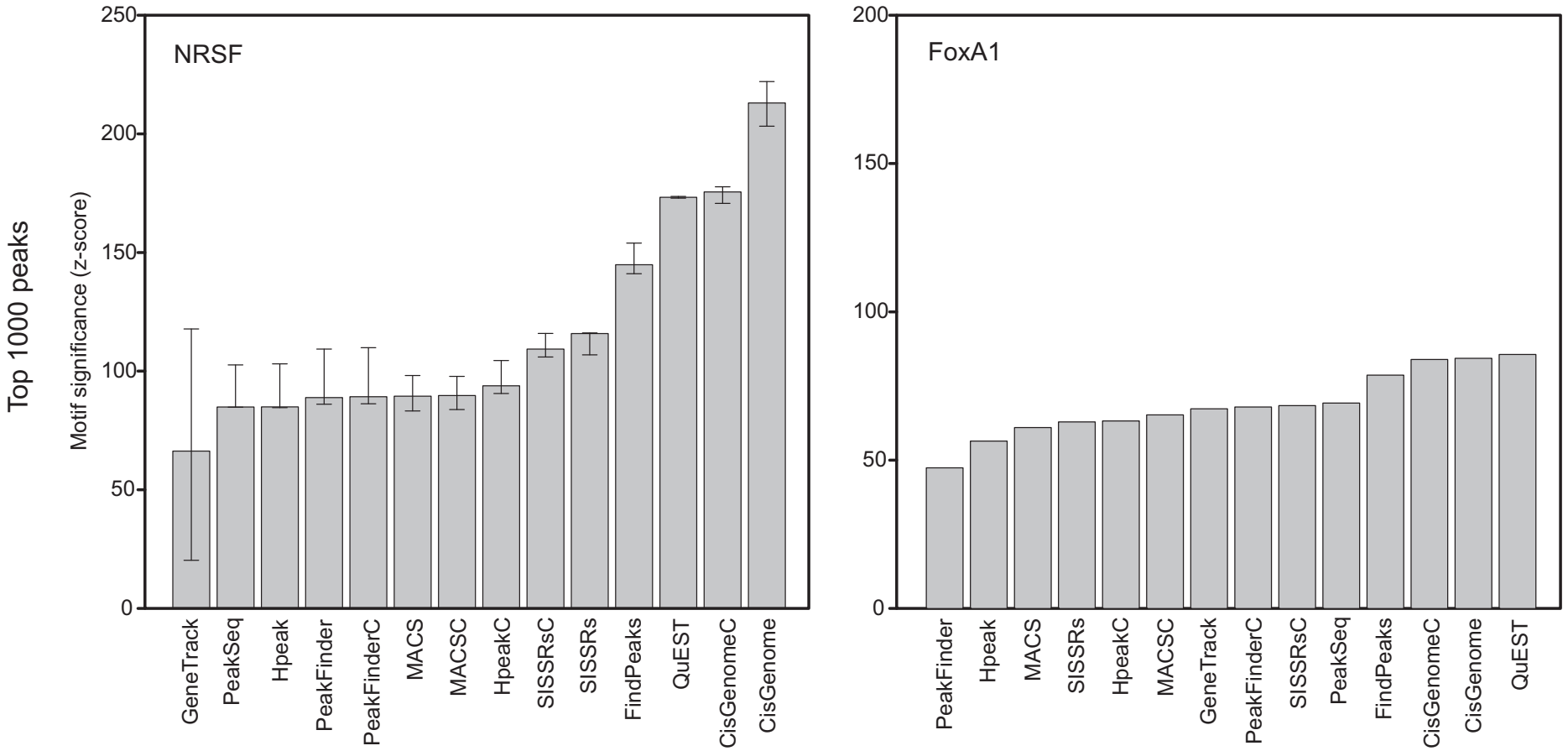


Figure 5

External validation of the predicted binding sites using binding motifs. The sign
peaks with the corresponding high-scoring sequence motifs (see Table 2 for the motif id
atix RegionMiner software separately for each transcription factor (columns), either wh

Summary of Laajala et al.

Laajala et al.

- Null distribution of reads is typically assumed to be poisson or negative binomial in absence of a control sample
- Number of called peaks may change radically over methods (default parameters)
- Inclusion of a biological control sample tended to work better than using an idealized null distribution
- Less peaks & reproducibility vs. More peaks & explorative novelty value? (analogy to $p < 0.05$ threshold)

Next, peakROTS

One method $\neq$ one possible result. Tuning the parameters for a method may radically change its results, so we explored options in the parameter space in our next paper.

Optimized detection of transcription factor-binding sites in ChIP-seq experiments

Laura L. Elo[1,2,*], Aleksi Kallio[3], Teemu D. Laajala[1,2], R. David Hawkins[2,4,5], Eija Korpelainen[3] and Tero Aittokallio[1,2,6,*]

[1]Department of Mathematics, University of Turku, FI-20014 Turku, [2]Turku Centre for Biotechnology, FI-20520 Turku, [3]CSC - IT Center for Science Ltd., FI-02101 Espoo, Finland, [4]Department of Medicine, Division of Medical Genetics, [5]Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA and [6]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, FI-00014 Helsinki, Finland

Received June 14, 2011; Revised September 2, 2011; Accepted September 20, 2011

ABSTRACT

We developed a computational procedure for optimizing the binding site detections in a given ChIP-seq experiment by maximizing their reproducibility under bootstrap sampling. We demonstrate how the procedure can improve the detection accuracies beyond those obtained with the default settings of popular peak calling software, or inform the user whether the peak detection results are compromised, circumventing the need for arbitrary re-iterative peak calling under varying parameter settings. The generic, open-source implementation is easily extendable to accommodate additional features and to promote its widespread application in future ChIP-seq studies. The peakROTS R-package and user guide are freely available at http://www.nic.funet.fi/pub/sci/molbio/peakROTS.

analysis. We have recently demonstrated that the choice of the software package may considerably affect the biological conclusions made from the ChIP-seq data (7), calling into question the validity of the binding site detections unless they are carefully confirmed in independent qPCR experiments. Another practical challenge is to decide whether the data is similar enough to those on which a specific peak calling algorithm was tuned to, in order to justify the use of its default parameters (6). However, even among the same type of data, variability in data quality may necessitate using various parameter settings (8). Accordingly, with the fixed default parameter settings, the choice of the best package is strongly dependent on the ChIP-seq data under analysis, making the selection between the different packages and optimization of their performance for a given data a challenging task (7,9,10).

To this end, we introduce here an adaptive procedure, which provides the user with an informed means to optimally adjust the parameters of a given software package to the intrinsic properties of each ChIP-seq data set sep-

Sampling and ROTS

Idea: If multiple rounds of a subset of random reads are picked from a sample, analysis methods should yield consistent results nevertheless.
Thus, we performed multiple rounds of boostrapping (sampling with replacement) for a dataset, and computed ROTS for top k peak lists:

$$Z_{k,\alpha} = \frac{R_{k,\alpha} - R^0_{k,\alpha}}{s_{k,\alpha}}$$

- $R_{k,\alpha}$ is the reproducibility of pairs of case-control samples that have been bootstrapped from the actual dataset
- $R^0_{k,\alpha}$ is the null reproducibility; combine case and control to a single sample and compute reproducibilities over multiple bootstrapped pairs of randomized samples
- $s_{k,\alpha}$ is the standard deviation of bootstrapped reproducibility

$\rightarrow$ choose the highest $Z_{k,\alpha}$ of multiple candidates

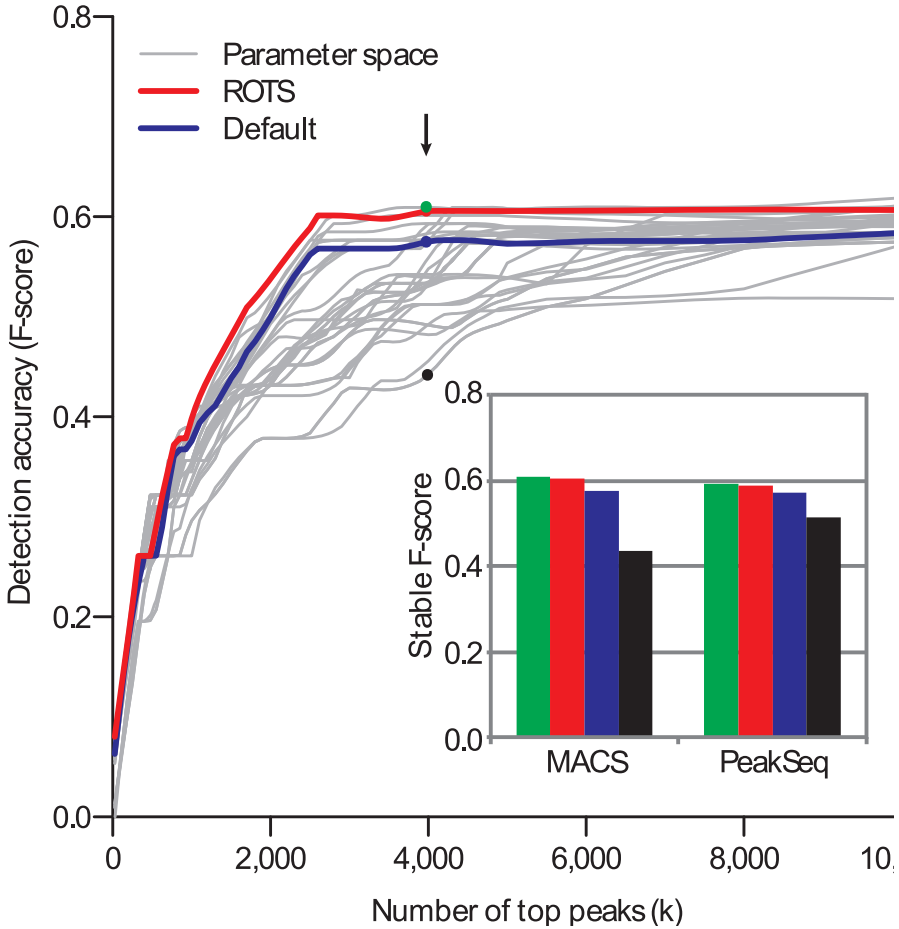
Parameter space

For example, MACS uses tag shifting and windowing to scan chromosome regions and a dynamic Poisson distribution to model the background signal. Over a lattice of possible parameter candidates, we computed $Z_{k,\alpha}$ for methods *MACS* and *PeakSeq* (former covered in Laajala et al.):

shift size	band width
1	100
1	300
1	500
5	100
...	...
200	500

The final choice of a tuned method should be the one that yielded the highest $Z_{k,\alpha}$. In general, our approach yielded as good as or better results than if method was run only with the default parameter values.

Figure 1: Exploring the parameter space (qPCR validated ChIP data)



$F = 2PR/(P + R)$; P = Precision, R = Recall

Pros and cons of the peakROTS approach

Pros

- Meta-analysis that improves any currently available method
- The sampling + exploring parameter space principle is generalizable
- An objective method for tuning in optimal parameters for a method ...

Cons

- ... well not entirely objective. The choice of parameter space is subjective, as we cannot test all possible parameter combinations with infinite precision
- Requires extra computational effort (e.g. CSC cluster)

Take home message

- There are multiple available methods, which vary in software platforms, underlying assumptions, and practical applicability
- Choice of a method affects the subsequent conclusions
- The optimal choice is data and application specific
- Less peaks $\rightarrow$ more conservative results and better reproducibility
- More peaks $\rightarrow$ more candidates (and perhaps novel discoveries from less prominent phenomena?)
- Reproducibility based sampling can help you optimize your analysis approach when the 'correct' answer is unknown
- Simplified: Test multiple possible ways to run a method, and choose the one that yields the most consistent results

Thanks!

Time for questions