Differential ChIP-Seq analysis

Oleg Shpynov

JetBrains Biolabs

February 26, 2017

ENCODE guidelines summary[2]

- Read depth $\geq$ 10 mln reads
- 2 biological replicates
- NFR[1]
- **FRiP**
- Strand cross-correlation
- NSC for sharp histone modifications
- IDR

Consensus of 3 peak callers comparison papers:

- Narrow peaks: MACS2 or SPP or PeakSeg
- Broad and mixed: SICER or ZINBA or PeakSeg or RSEG

[1] This will fail ENCODE guideline, because of new protocol

[2] `http://encodeproject.org/ENCODE/experiment_guidelines.html`

IDR - irreproducible discovery rate

Given a set of peak calls for a pair of replicate data sets, the peaks can be ranked based on a criterion of significance, such as the P-value, the q-value, the ChIP-to-input enrichment, or the read coverage for each peak. The most significant peaks, which are likely to be genuine signals, are expected to have high consistency between replicates, whereas peaks with low significance, which are more likely to be noise, are expected to have low consistency.

A **major advantage** of IDR is that it can be used to establish a stable threshold for called peaks that is more consistent across laboratories, antibodies, and analysis protocols (e.g., peak callers) than are FDR measures (different for each tool).[3]

[3] A caution in applying IDR is that it is dominated by the weakest replicate. Additional qc is required

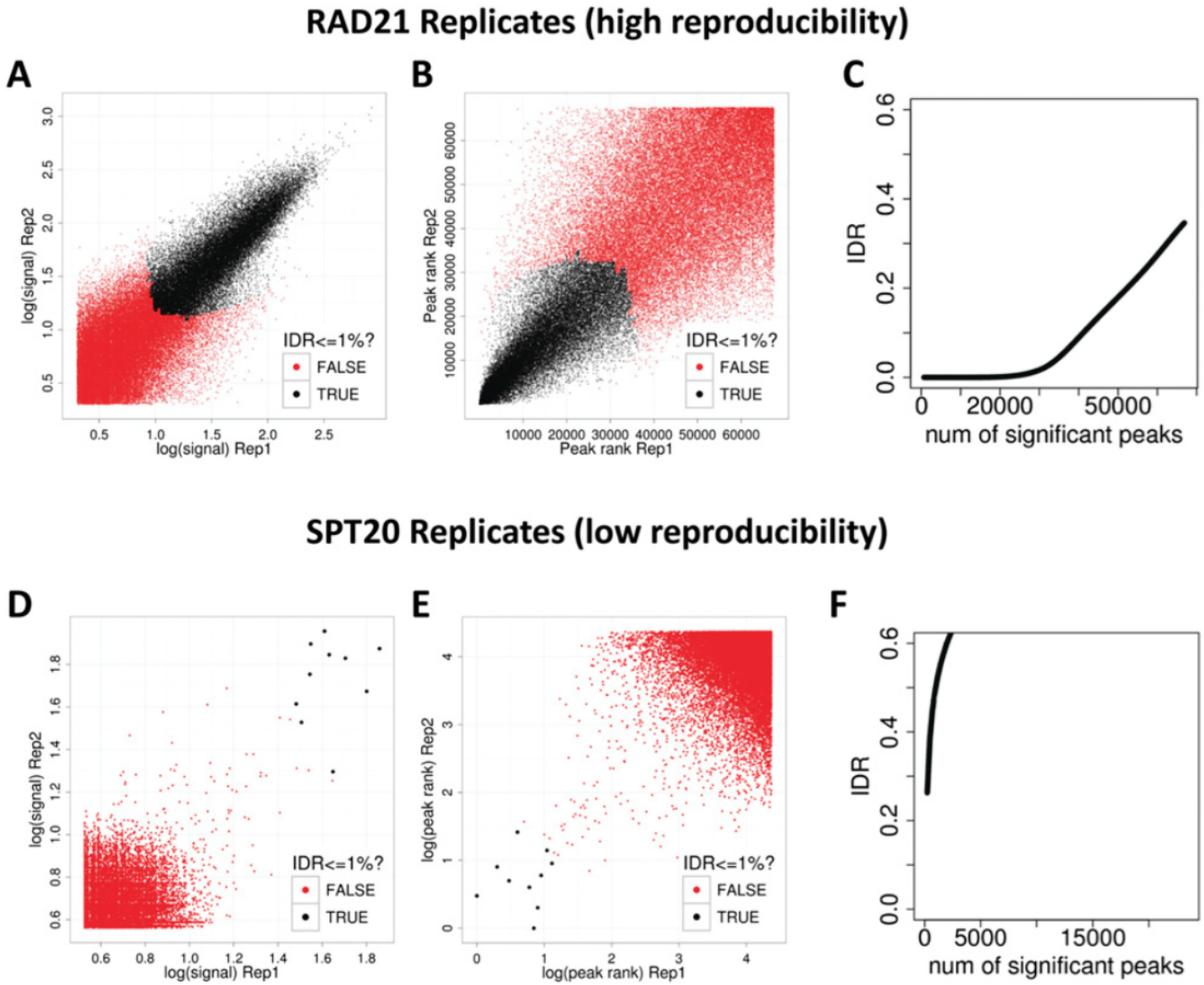
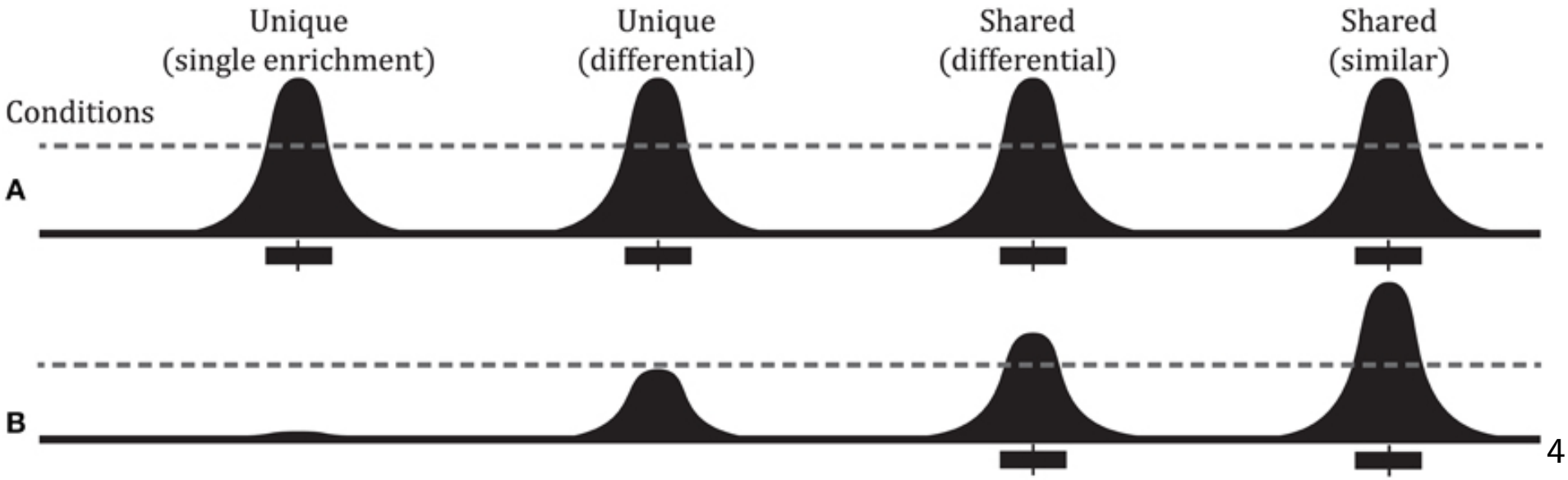


Figure 6. The irreproducible discovery rate (IDR) framework for assessing reproducibility of ChIP-seq data sets. (*A–C*) Reproducibility analysis for a pair of high-quality RAD21 ChIP-seq replicates. (*D,E*) The same analysis for a pair of low quality SPT20 ChIP-seq replicates. (*A,D*) Scatter plots of signal scores of peaks that overlap in each pair of replicates. (*B,E*) Scatter plots of ranks of peaks that overlap in each pair of replicates. Note that low ranks correspond to high signal and vice versa. (*C,F*) The estimated IDR as a function of different rank thresholds. (*A,B,D,E*) Black data points represent pairs of peaks that pass an IDR threshold of 1%, whereas the red data points represent pairs of peaks that do not pass the IDR threshold of 1%. The RAD21 replicates show high reproducibility with ~30,000 peaks passing an IDR threshold of 1%, whereas the SPT20 replicates show poor reproducibility with only six peaks passing the 1% IDR threshold.

Differential peak calling



[4]

[4] Identifying differential transcription factor binding in ChIP-seq

Differential peak calling

Two alternatives have been proposed.[5]

- qualitative - hypothesis testing on multiple overlapping sets of peaks[6].
- quantitative - proposes the analysis of differential binding between conditions based on the total counts of reads in peak regions or on the read densities, i.e., counts of reads overlapping at individual genomic positions[7]
- model based

[5] Practical Guidelines for the Comprehensive Analysis of ChIP-seq Data

[6] `http://bioinformatics.oxfordjournals.org/content/28/24/3318.short`

[7] `http://www.biomedcentral.com/content/pdf/gb-2010-11-10-r106.pdf`

Tools

A comprehensive comparison of tools for differential ChIP-seq analysis

Sebastian Steinhauser, Nils Kurzawa, Roland Eils and Carl Herrmann

Corresponding author. Carl Herrmann, IPMB Universitôt Heidelberg and Department of Theoretical Bioinformatics, DKFZ, Im Neuenheimer Feld 364, D-69120 Heidelberg, Tel.: (+49) 6221 423612. E-mail: carl.herrmann@uni-heidelberg.de

Problems

- Results obtained by different centers - systematic bias correction required before data integration can be achieved
- Reproducibility of the assays is often limited, especially in cases with additional constraints, for example low input material
- ENCODE project showed that the amount of noise is *geq* 90% (FRiP)
- IDR for the rescue!

Tools

Our criterion for tool selection was the availability of a working software that could be implemented without the need for extensive efforts for porting the code.[8]

[8] ChipDiff is not there - authors had problems installing it. Indeed.

Different aims - different approaches - different tools

Tools

- Replicated or not
- ChIPSeq or TF data
- Control is required or not

Test data

NO gold standard for differential enrichment in ChIP signal.

- Each category is tested separately, 2 biological replicates for each condition[9]
- TF: FoxA1 ChIP-seq for (E2)- and vehicle (Veh)-treated MCF7 (GSE59530) + expression data (GSE59531)
- Shar ChIP-Seq: H3K27ac for hESC-H1 and mesenchymal stem cells (GSE16256)
- Broad ChIP-Seq: H3K36me3 for myelome cells TKO vs NTKO (GSE57632)
- Simulated datasets to estimate Sensitivity/Specificity
- hg19, BWA, filter out duplicated/bad quality reads
- MACS2
  narrow: '-g hs -q 0.1 –call-summits'
  broad: '-g hs -q 0.1 –broad'

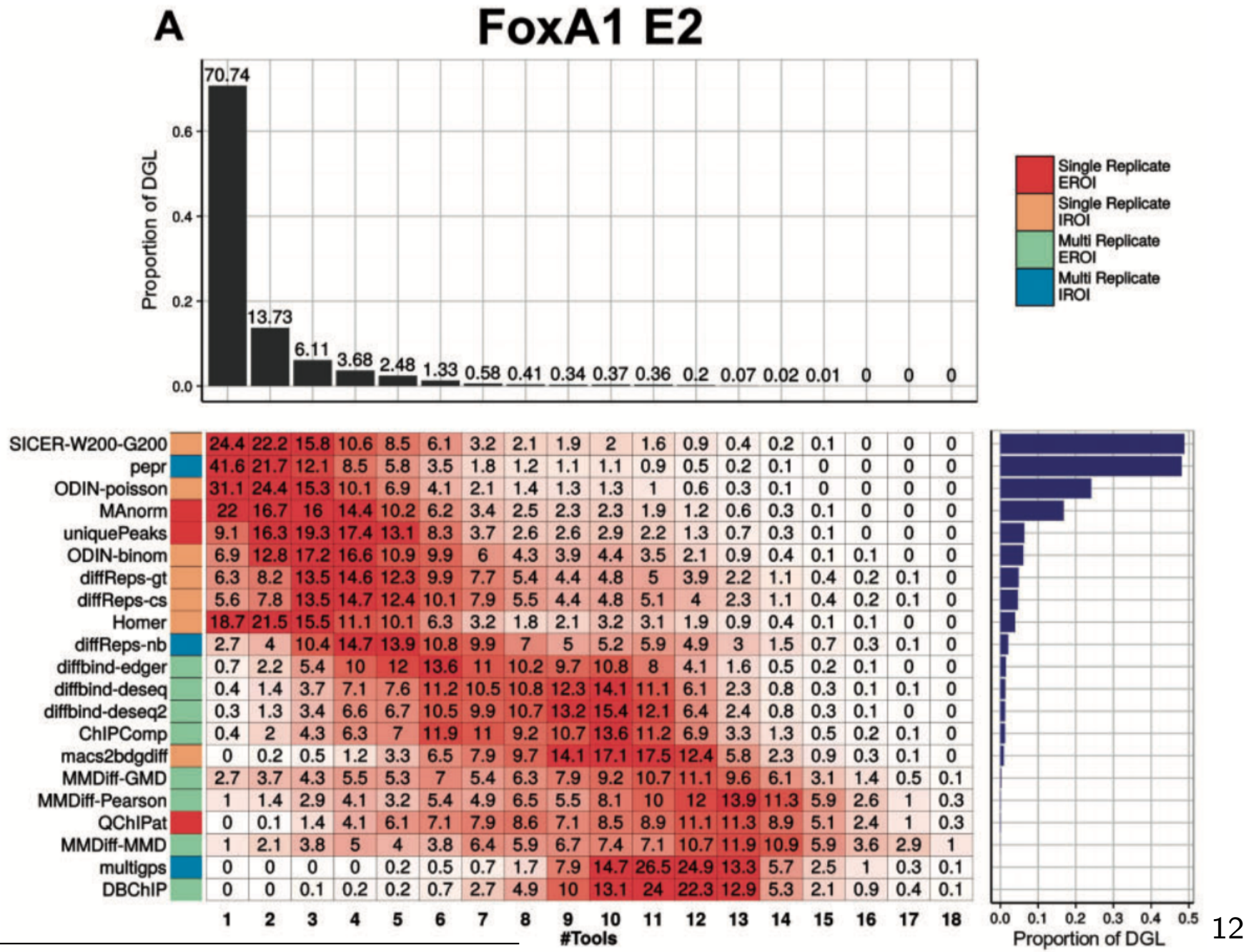
[9] Replicates were pooled for single source tools

Params

- NO silver bullet: NO common significant threshold.
  When possible: FDR ≤ 0.01 or P-value ≤ 0.05.
- Gene centric approach: ranked differential peaks by significance - select top 1000 closest genes[10]
- Functional annotation of regions $\pm$ 1.5kb around TSS[11].
- Jaccard index used as measure of differential peaks correspondence
- Analyze single tool DR fraction in union of all DR from all tools

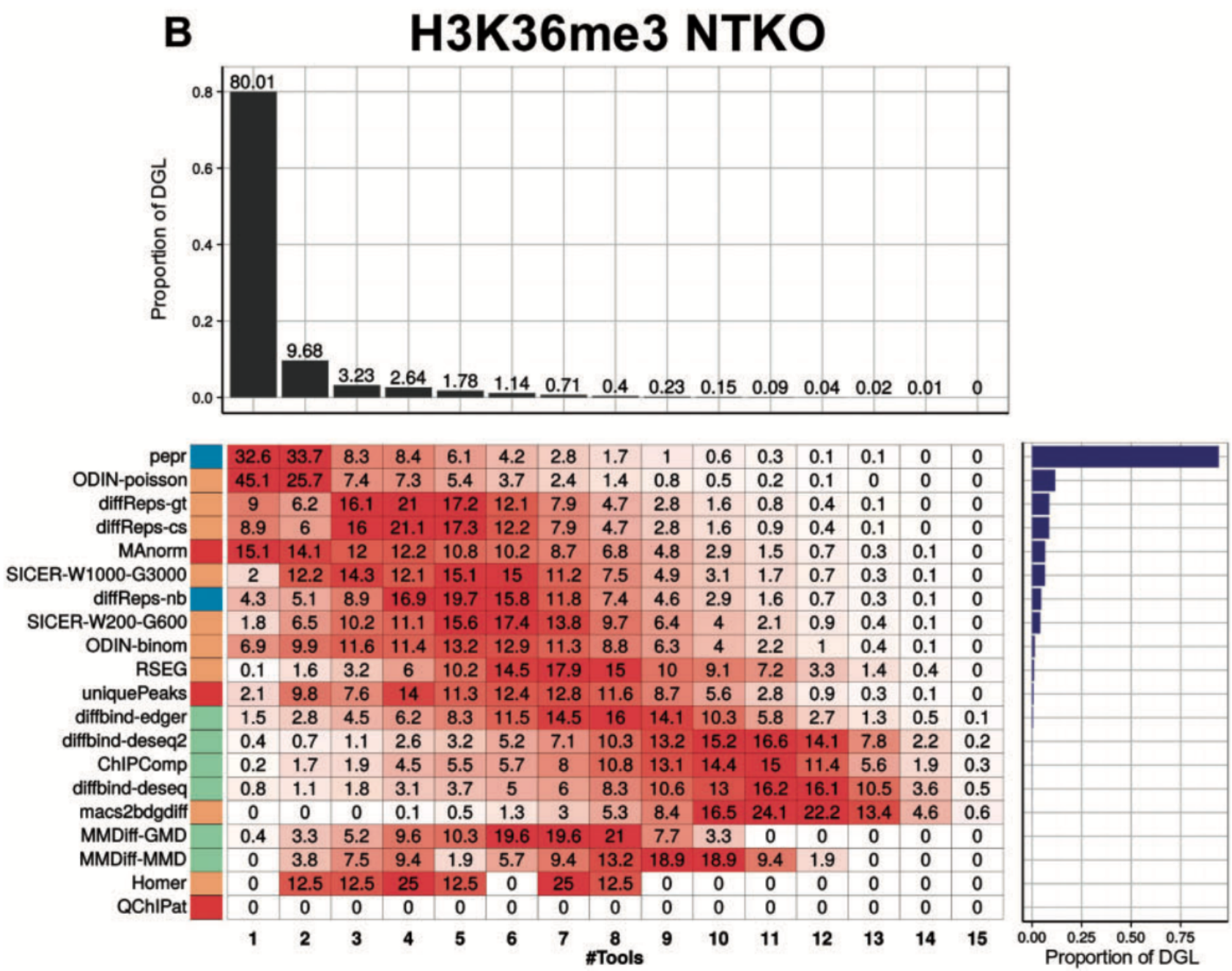
[10] ChIPSeeker

[11] GREAT: single nearest gene



[12]

[12] EROI/IROI-external regions required
Tools - DR is identified by number of tools


B
H3K36me3 NTKO
Proportion of DGL
0.0
0.2
0.4
0.6
0.8
80.01
9.68
3.23
2.64
1.78
1.14
0.71
0.4
0.23
0.15
0.09
0.04
0.02
0.01
0
pepr 32.6 33.7 8.3 8.4 6.1 4.2 2.8 1.7 1 0.6 0.3 0.1 0.1 0 0
ODIN-poisson 45.1 25.7 7.4 7.3 5.4 3.7 2.4 1.4 0.8 0.5 0.2 0.1 0 0 0
diffReps-gt 9 6.2 16.1 21 17.2 12.1 7.9 4.7 2.8 1.6 0.8 0.4 0.1 0 0
diffReps-cs 8.9 6 16 21.1 17.3 12.2 7.9 4.7 2.8 1.6 0.9 0.4 0.1 0 0
MAnorm 15.1 14.1 12 12.2 10.8 10.2 8.7 6.8 4.8 2.9 1.5 0.7 0.3 0.1 0
SICER-W1000-G3000 2 12.2 14.3 12.1 15.1 15 11.2 7.5 4.9 3.1 1.7 0.7 0.3 0.1 0
diffReps-nb 4.3 5.1 8.9 16.9 19.7 15.8 11.8 7.4 4.6 2.9 1.6 0.7 0.3 0.1 0
SICER-W200-G600 1.8 6.5 10.2 11.1 15.6 17.4 13.8 9.7 6.4 4 2.1 0.9 0.4 0.1 0
ODIN-binom 6.9 9.9 11.6 11.4 13.2 12.9 11.3 8.8 6.3 4 2.2 1 0.4 0.1 0
RSEG 0.1 1.6 3.2 6 10.2 14.5 17.9 15 10 9.1 7.2 3.3 1.4 0.4 0
uniquePeaks 2.1 9.8 7.6 14 11.3 12.4 12.8 11.6 8.7 5.6 2.8 0.9 0.3 0.1 0
diffbind-edger 1.5 2.8 4.5 6.2 8.3 11.5 14.5 16 14.1 10.3 5.8 2.7 1.3 0.5 0.1
diffbind-deseq2 0.4 0.7 1.1 2.6 3.2 5.2 7.1 10.3 13.2 15.2 16.6 14.1 7.8 2.2 0.2
ChIPComp 0.2 1.7 1.9 4.5 5.5 5.7 8 10.8 13.1 14.4 15 11.4 5.6 1.9 0.3
diffbind-deseq 0.8 1.1 1.8 3.1 3.7 5 6 8.3 10.6 13 16.2 16.1 10.5 3.6 0.5
macs2bdgdiff 0 0 0 0.1 0.5 1.3 3 5.3 8.4 16.5 24.1 22.2 13.4 4.6 0.6
MMDiff-GMD 0.4 3.3 5.2 9.6 10.3 19.6 19.6 21 7.7 3.3 0 0 0 0 0
MMDiff-MMD 0 3.8 7.5 9.4 1.9 5.7 9.4 13.2 18.9 18.9 9.4 1.9 0 0 0
Homer 0 12.5 12.5 25 12.5 0 25 12.5 0 0 0 0 0 0 0
QChIPat 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
#Tools
0.00 0.25 0.50 0.75
Proportion of DGL

Results

- Tools for differential ChIP-seq analysis show important differences in the number and size of detected differential regions (DR).
- Methods taking into account replicates appear to be more robust than those handling single replicate data sets.
- Inconsistent sets of DR will affect results based on sequence analysis, like detection of enriched transcription factor binding sites.
- Analysis of functional enrichments based on neighboring genes appears to be more robust.
- Some tools give good results with default parameters, like ChIPComp or diffBind when replicates are available, or MAnorm, Homer, macs2bdgdiff and RSEG with single replicates. The other tools would require more extensive fine-tuning of parameters to achieve satisfactory results.

How do I choose?

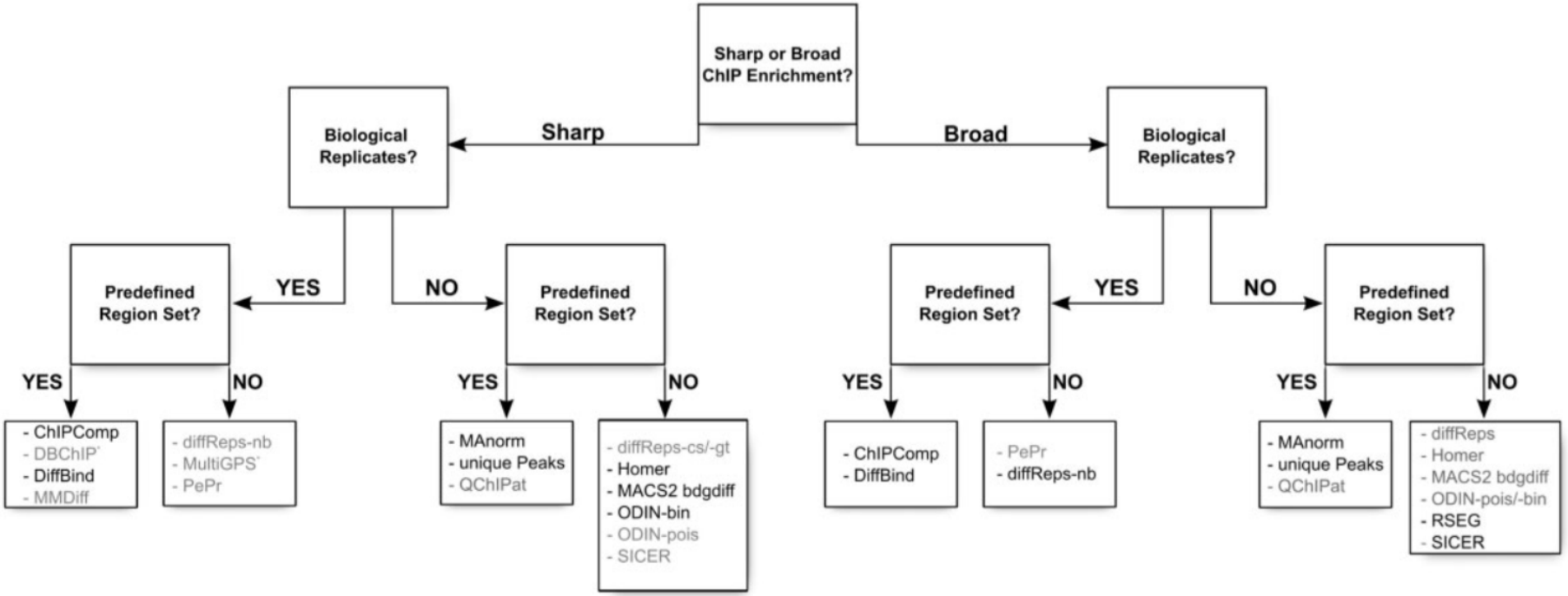


Figure 7. Decision tree indicating the proper choice of tool depending on the data set: shape of the signal (sharp peaks or broad enrichments), presence of replicates and presence of an external set of regions of interest. We have indicated in dark the name of the tools that give good results using default settings, and in gray the tools that would require parameter tuning to achieve optimal results: some tools suffer from an excessive number of DR (PePr, ODIN-pois), an insufficient number of DR (QChIPat, MMDiff, DBChIP) or from an imprecise definition of the DR for sharp signal (SICER, diffReps-nb). *MultiGPS has been explicitly developed for transcription factor ChIP-seq.

Popular tools

ChiPDiff

An HMM approach to genome-wide identification of differential histone modification sites from ChIP-seq data

H Xu, CL Wei, F Lin, WK Sung - Bioinformatics, 2008 - Oxford Univ Press

... Results: Aiming at identifying DHMSs, we propose an approach called **ChIPDiff** for the genome-wide comparison of histone modification sites identified by ChIP-seq. ... Here, we propose a HMM-based approach called **ChIPDiff** to solve the problem. ...

Цитируется: 111 Похожие статьи Все версии статьи (11) Цитировать Сохранить Ещё

RSEG

[HTML] Identifying dispersed epigenomic domains from ChIP-Seq data

Q Song, AD Smith - Bioinformatics, 2011 - Oxford Univ Press

Abstract Motivation: Post-translational modifications to histones have several well known associations with regulation of gene expression. While some modifications appear concentrated narrowly, covering promoters or enhancers, others are dispersed as ...

Цитируется: 85 Похожие статьи Все версии статьи (15) Цитировать Сохранить Ещё

MAnorm

[HTML] **MAnorm**: a robust model for quantitative comparison of **ChIP-Seq** data sets

Z Shao, Y Zhang, GC Yuan, SH Orkin, DJ Waxman - Genome biology, 2012 - Springer

... following **MAnorm** are robust to different peak cutoffs; integrating multiple replicates in **ChIP-seq** data set comparison; derivation of the P-value that quantifies the significance of **differential** binding at peak regions; using **MAnorm** to compare H3K36me3 **ChIP-seq** data; assessing ...

Цитируется: 70 Похожие статьи Все версии статьи (10) Цитировать Сохранить Ещё

Less popular

diffReps

[HTML] **diffReps**: detecting differential chromatin modification sites from ChIP-seq data with biological replicates

L Shen, NY Shao, X Liu, I Maze, J Feng, EJ Nestler - PloS one, 2013 - journals.plos.org

Abstract ChIP-seq is increasingly being used for genome-wide profiling of histone modification marks. It is of particular importance to compare ChIP-seq data of two different conditions, such as disease vs. control, and identify regions that show differences in ChIP

Цитируется: 59 Похожие статьи Все версии статьи (14) Цитировать Сохранить Ещё

DiffBind

[ЦИТИРОВАНИЕ] **DiffBind**: differential binding analysis of ChIP-Seq peak data

R Stark, G Brown - R package version, 2011

Цитируется: 24 Похожие статьи Все версии статьи (18) Цитировать Сохранить

ODIN

[HTML] Detecting **differential** peaks in **ChIP-seq** signals with ODIN

M Allhoff, K Seré, H Chauvistré, Q Lin, M Zenke… - …, 2014 - Oxford Univ Press

... Here, we propose an One-stage **DIffereNtial** peak caller (ODIN), an HMM-based approach to detect and analyze DPs **...** This allows us to perform a comparative analysis with all competing methods (DBChIP, **MAnorm**, DESeq, MACS2 and ChIPDiff) for **ChIP-seq** data from **...**

Цитируется: 4 Похожие статьи Все версии статьи (9) Цитировать Сохранить Ещё

Less popular

ChipComp

[PDF] **ChIPComp**: A novel statistical method for quantitative comparison of multiple ChIP-seq datasets

L Chen, C Wang, Z Qin, H Wu - 2015 - pdfs.semanticscholar.org

This vignette introduces the use of the Bioconductor package **ChIPComp**, which is designed for differential binding sites analyses based on high-throughput sequencing data. The core of **ChIPComp** is a new procedure to incorporate the control sequencing data in a linear **...**

Цитировать Сохранить Ещё

Thank you!

oleg.shpynov@jetbrains.com

https://research.jetbrains.org/groups/biolabs