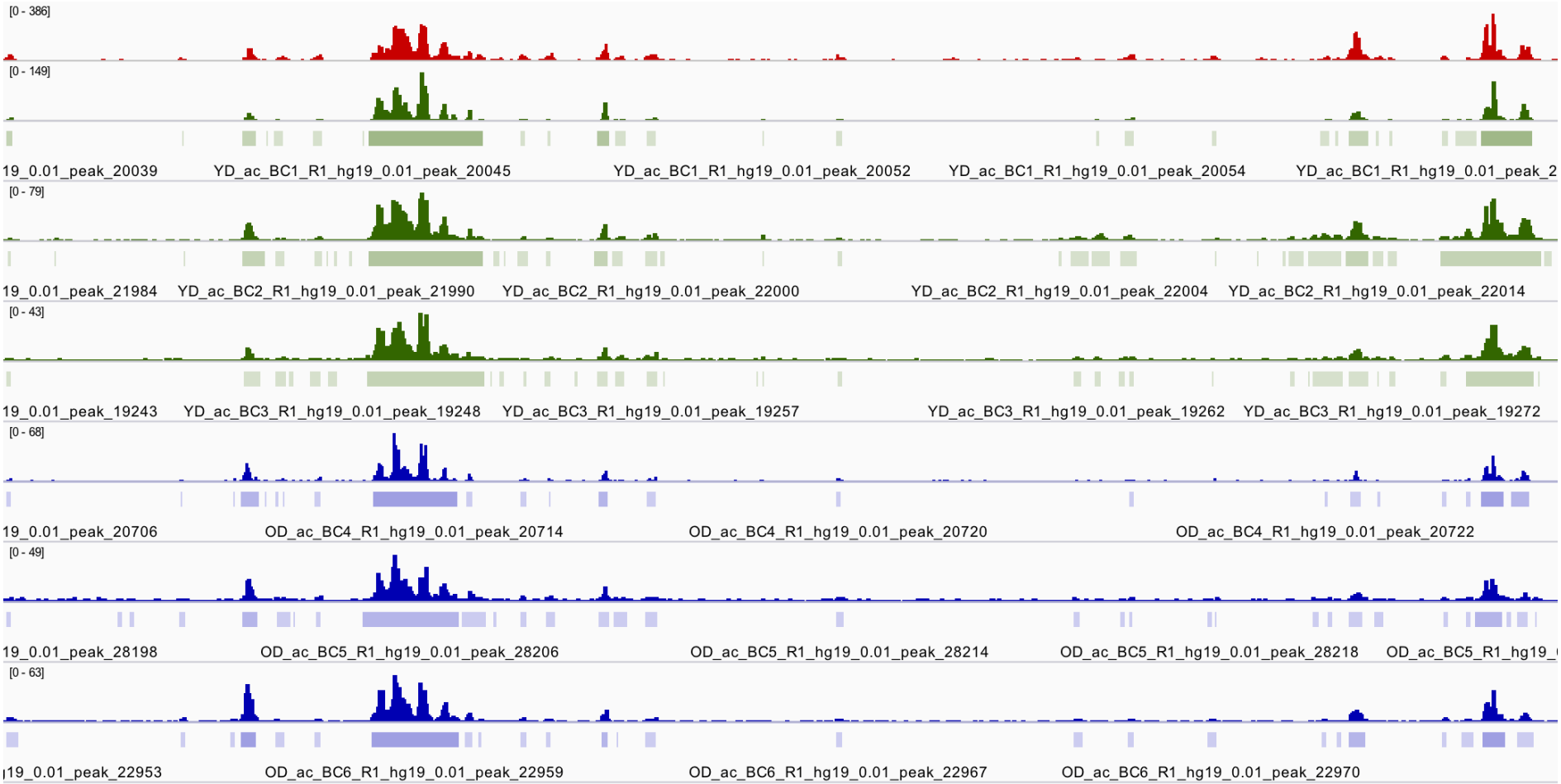
ChIP-Seq peak calling

Oleg Shpynov

JetBrains Biolabs

February 22, 2017

Peaks


[0 - 386]
[0 - 149]
19_0.01_peak_20039
YD_ac_BC1_R1_hg19_0.01_peak_20045
YD_ac_BC1_R1_hg19_0.01_peak_20052
YD_ac_BC1_R1_hg19_0.01_peak_20054
YD_ac_BC1_R1_hg19_0.01_peak_2
[0 - 79]
19_0.01_peak_21984
YD_ac_BC2_R1_hg19_0.01_peak_21990
YD_ac_BC2_R1_hg19_0.01_peak_22000
YD_ac_BC2_R1_hg19_0.01_peak_22004
YD_ac_BC2_R1_hg19_0.01_peak_22014
[0 - 43]
19_0.01_peak_19243
YD_ac_BC3_R1_hg19_0.01_peak_19248
YD_ac_BC3_R1_hg19_0.01_peak_19257
YD_ac_BC3_R1_hg19_0.01_peak_19262
YD_ac_BC3_R1_hg19_0.01_peak_19272
[0 - 68]
19_0.01_peak_20706
OD_ac_BC4_R1_hg19_0.01_peak_20714
OD_ac_BC4_R1_hg19_0.01_peak_20720
OD_ac_BC4_R1_hg19_0.01_peak_20722
[0 - 49]
19_0.01_peak_28198
OD_ac_BC5_R1_hg19_0.01_peak_28206
OD_ac_BC5_R1_hg19_0.01_peak_28214
OD_ac_BC5_R1_hg19_0.01_peak_28218
OD_ac_BC5_R1_hg19_
[0 - 63]
19_0.01_peak_22953
OD_ac_BC6_R1_hg19_0.01_peak_22959
OD_ac_BC6_R1_hg19_0.01_peak_22967
OD_ac_BC6_R1_hg19_0.01_peak_22970

Agenda

- Reads QC
- Alignment
- Peaks calling
- Peak caller comparison
- ENCODE DNANexus pipeline

Reads QC

NGS QC Toolkit: A Toolkit for Quality Control of Next Generation Sequencing Data

Ravi K. Patel, Mukesh Jain*

Functional Genomics and Bioinformatics Laboratory, National Institute of Plant Genome Research (NIPGR), New Delhi, India

Only a few standalone tools for QC of NGS data are publicly available other than commercial softwares supplied with the sequencing machines, which are not sufficiently optimal.

Table 1. Comparison of various features of NGS QC toolkit and other available QC tools.

Feature\Tools	NGS QC Toolkit v2.2	FastQC v0.10.0	PRINSEQ-lite v0.17[1]	TagDust	FASTX-Toolkit v0.0.13	SolexaQA v1.10	TagCleaner v0.12[1]	CANGS v1.1
Supported NGS platforms	Illumina, 454	FASTQ[2]	Illumina, 454	Illumina, 454	Illumina	Illumina	Illumina, 454	454
Parallelization	Yes	Yes	No	No	No	No	No	No
Detection of FASTQ variants	Yes	Yes	Yes	No	No	Yes	No	No
Primer/Adaptor removal	Yes	No[3]	No	Yes	Yes	No	Yes[4]	Yes
Homopolymer trimming (Roche 454 data)	Yes	No	No	No	No	No	No	Yes
Paired-end data integrity	Yes	No	No	No	No	No	No	No
QC of 454 paired-end reads	Yes	No	No	No	No	No	No	No
Sequence duplication filtering	No	No[5]	Yes	No	Yes	No	No	Yes
Low complexity filtering	No	No	Yes	No	Yes	No	No	No
N/X content filtering	No	No[6]	Yes	No	Yes	No	No	Yes
Compatability with compressed input data file	Yes	Yes	No	No	No	No	No	No
GC content calculation	Yes	Yes	Yes	No	No	No	No	No
File format conversion	Yes	No	No	No	No	No	No	No
Export HQ and/or filtered reads	Yes	No	Yes	Yes	Yes	No	Yes	Yes
Graphical output of QC statistics	Yes	Yes	No[7]	No	Yes	Yes	No[7]	No
Dependencies	Perl modules: Parallel::ForkManager, String::Approx, GD::Graph (optional)	-	-	-	Perl module: GD::Graph	R, matrix2png	-	BLAST, NCBI nr database

[1]Standalone version.
[2]Data of any platform in FASTQ file format.
[3]only detection.
[4]only one primer/adaptor sequence at a time.
[5]only reports duplication and that too is for only first 200,000 reads.
[6]only reports N/X content.
[7]yes, in case of online version.
doi:10.1371/journal.pone.0030619.t001

FastQC

[ЦИТИРОВАНИЕ] **FastQC**: A quality control tool for high throughput sequence data
S Andrews - Reference Source, 2010
Цитируется: 408 Похожие статьи Цитировать Сохранить

FastQC Report

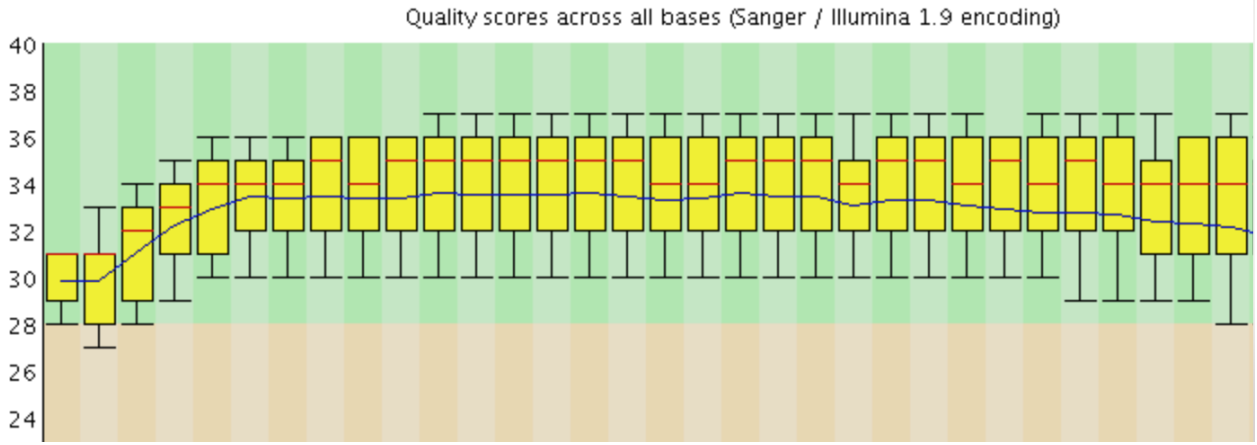
Wed 11 May 2016
SRR568364.fastq

Summary

Basic Statistics

Measure	Value
Filename	SRR568364.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	22999576
Sequences flagged as poor quality	0
Sequence length	36
%GC	48

Per base sequence quality



MultiQC

Aggregate results from bioinformatics analyses across many samples into a single report.[1]

[1] Aggregate results from bioinformatics analyses across many samples into a single report

General Statistics

Copy table | Configure Columns | Showing 8 rows.

Sample Name	% Assigned	M Assigned	% Aligned	M Aligned	% Trimmed	% Dups	% GC	Length
SRR3192396	67.5%	71.9	93.7%	97.8	4.0%	78.9%	51%	97
SRR3192397	66.6%	63.0	94.7%	87.1	3.5%	77.2%	49%	97
SRR3192398	50.9%	36.5	88.2%	58.7	5.0%	55.3%	47%	97
SRR3192399	52.3%	42.3	88.2%	65.6	5.0%	57.4%	47%	97
SRR3192400	70.3%	63.4	77.3%	73.4	7.2%	74.1%	45%	93
SRR3192401	71.2%	63.8	76.4%	72.8	6.3%	76.3%	45%	94
SRR3192657	73.1%	67.1	91.2%	85.0	3.1%	82.2%	51%	98
SRR3192658	71.2%	66.9	89.7%	87.1	3.4%	82.3%	52%	97

Alignment

Hatem *et al. BMC Bioinformatics* 2013, **14**:184
http://www.biomedcentral.com/1471-2105/14/184

RESEARCH ARTICLE Open Access

Benchmarking short sequence mapping tools

Ayat Hatem[1,2], Doruk Bozdağ[2], Amanda E Toland[3] and Ümit V Çatalyürek[1,2*]

Alignment

- SNPs: GSNAP's filtered output helped in detecting the largest number of SNPs.
- The evaluation of Bowtie, Bowtie2, BWA, mrsFAST, and Novoalign show their ability to correctly map the reads.
- Genome: further investigations are required to understand the different properties of the genomes and their effect on the different mapping techniques.
- Mismatches: Bowtie output mappings are more accurate than the other tools.
- Indels: Tools like Maq and Bowtie will not map reads if there is an insertion or deletion. I have used BWA to map 75bp Illumina reads at 20x coverage to a 30Mb fungal genome with good results.[2]
- In general, there is no **the-best** tool among all of the tools; each tool was **the-best** in certain conditions.

[2] `https://www.biostars.org/p/97197/`

Popular aligners

BWA

Fast and accurate short read alignment with Burrows-Wheeler transform.
www.ncbi.nlm.nih.gov/pubmed/19451168 ▾ Перевести эту страницу
автор: H Li - 2009 - Цитируется: 8015 - Похожие статьи
Bioinformatics. 2009 Jul 15;25(14):1754-60. doi: 10.1093/bioinformatics/btp324. Epub 2009 May 18.
Fast and accurate short read alignment with ...

Bowtie

[PDF] **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome**
B Langmead, C Trapnell, M Pop, SL Salzberg - Genome biol, 2009 - biomedcentral.com
Abstract Bowtie is an ultrafast, memory-efficient alignment program for aligning short DNA sequence reads to large genomes. For the human genome, Burrows-Wheeler indexing allows Bowtie to align more than 25 million reads per CPU hour with a memory footprint of ...
Цитируется: 7635 Похожие статьи Все версии статьи (55) Цитировать Сохранить Ещё

Less popular

MAQ

Mapping short DNA sequencing reads and calling variants using - NCBI
www.ncbi.nlm.nih.gov/pubmed/18714091 ▾ Перевести эту страницу
автор: H Li - 2008 - Цитируется: 2186 - Похожие статьи
19 авг. 2008 г. - **Mapping short DNA sequencing reads and calling variants using mapping quality scores**. Li H(1), Ruan J, Durbin R. Author information:

PASH

[HTML] Pash 3.0: A versatile software package for read mapping and integrative analysis of genomic and epigenomic variation using massively parallel DNA …
C Coarfa, F Yu, CA Miller, Z Chen… - BMC …, 2010 - bmcbioinformatics.biomedcentral. …
Background Massively parallel sequencing readouts of epigenomic assays are enabling integrative genome-wide analyses of genomic and epigenomic variation. Pash 3.0 performs sequence comparison and read mapping and can be employed as a module within ...
Цитируется: 36 Похожие статьи Все версии статьи (19) Цитировать Сохранить Ещё

Mega projects

- ENCODE: MAQ
  ChIP-seq reads were aligned to human genome build HG18 with MAQ using default parameters. All reads were truncated to 36 bases before alignment.[3]
- Roadmapepigenomics: PASH
  Sequenced data sets from the Release 9 of the Epigenome Atlas involved mapping a total of 150.21 billion sequencing reads onto hg19 assembly of the human genome using the PASH read mapper. These read mappings were used (except for RNA-seq data sets) for constructing the 111 consolidated epigenomes. Only uniquely mapping reads were retained and multiply-mapping reads were filtered out (subsampling to 30mln reads, ignoring blacklisted regions).[4]

[3] http://www.nature.com/nature/journal/v473/n7345/abs/nature09906.html

[4] http://www.nature.com/nature/journal/v518/n7539/full/nature14248.html

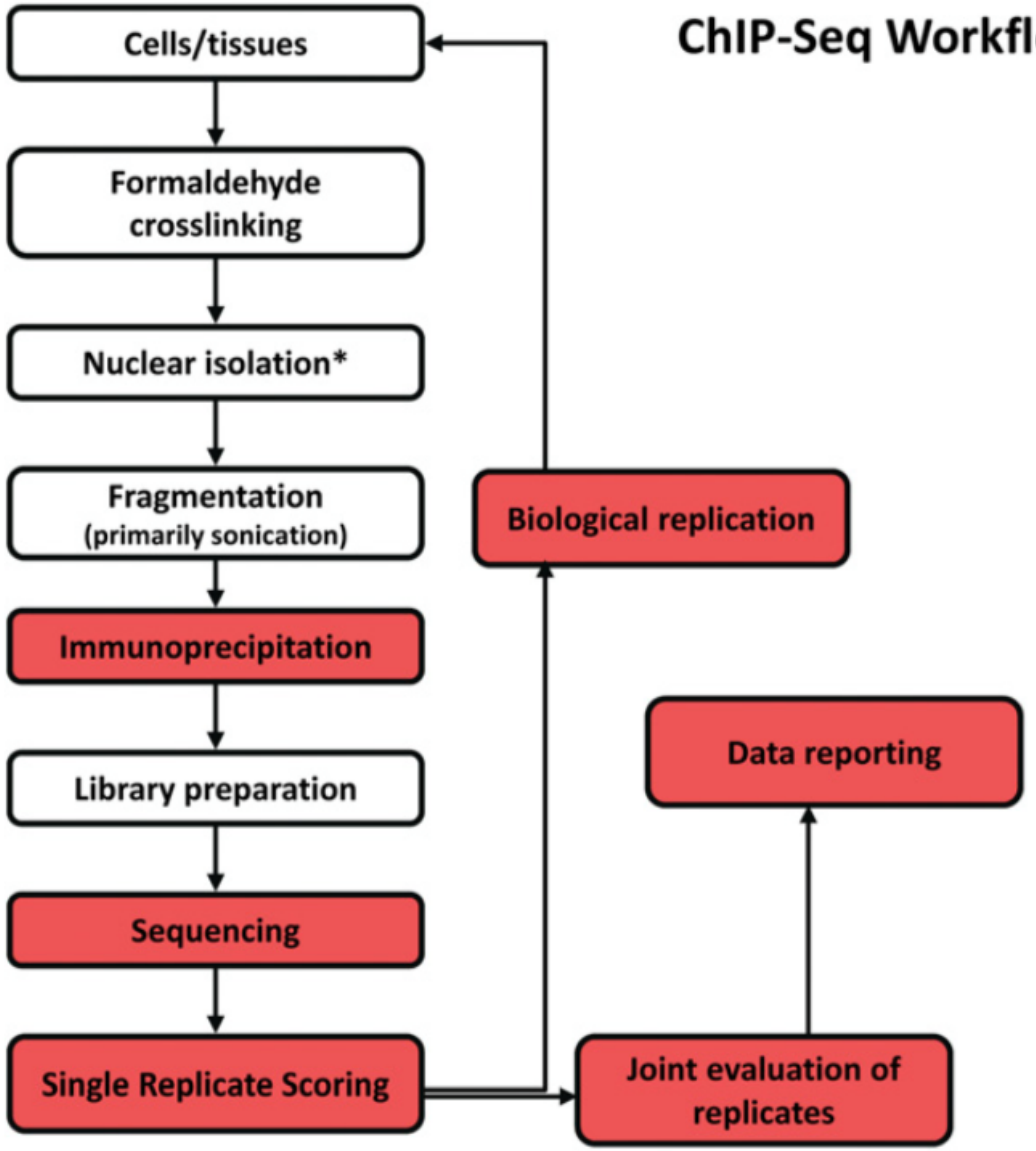
ENCODE guideline

ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia

Stephen G. Landt,[1,26] Georgi K. Marinov,[2,26] Anshul Kundaje,[3,26] Pouya Kheradpour,[4] Florencia Pauli,[5] Serafim Batzoglou,[3] Bradley E. Bernstein,[6] Peter Bickel,[7] James B. Brown,[7] Philip Cayting,[1] Yiwen Chen,[8] Gilberto DeSalvo,[2] Charles Epstein,[6] Katherine I. Fisher-Aylor,[2] Ghia Euskirchen,[1] Mark Gerstein,[9] Jason Gertz,[5] Alexander J. Hartemink,[10] Michael M. Hoffman,[11] Vishwanath R. Iyer,[12] Youngsook L. Jung,[13,14] Subhradip Karmakar,[15] Manolis Kellis,[4] Peter V. Kharchenko,[12] Qunhua Li,[16] Tao Liu,[8] X. Shirley Liu,[8] Lijia Ma,[15] Aleksandar Milosavljevic,[17] Richard M. Myers,[5] Peter J. Park,[13,14] Michael J. Pazin,[18] Marc D. Perry,[19] Debasish Raha,[20] Timothy E. Reddy,[5,27] Joel Rozowsky,[9] Noam Shoresh,[6] Arend Sidow,[1,21] Matthew Slattery,[15] John A. Stamatoyannopoulos,[11,22] Michael Y. Tolstorukov,[13,14] Kevin P. White,[15] Simon Xi,[23] Peggy J. Farnham,[24,28] Jason D. Lieb,[25,28] Barbara J. Wold,[2,28] and Michael Snyder[1,28]

[1–25]*[Author affiliations appear at the end of the paper.]*

The ENCODE and modENCODE consortia have performed more than a thousand individual ChIP-seq experiments for more than 140 different factors and histone modifications in more than 100 cell types in four different organisms.


ChIP-Seq Workflow
Cells/tissues
Formaldehyde crosslinking
Nuclear isolation*
Fragmentation (primarily sonication)
Immunoprecipitation
Library preparation
Sequencing
Single Replicate Scoring
Biological replication
Joint evaluation of replicates
Data reporting

Different sizes

- **Point-source** factors and certain chromatin modifications are localized at specific positions that generate highly localized ChIP-seq signals. This class includes most sequence-specific transcription factors, their cofactors, and, with some caveats, transcription start site or enhancer-associated histone marks. These comprise the *majority* of ENCODE and modENCODE determinations and are therefore the primary focus of this work.
- **Broad-source** factors are associated with large genomic domains. Examples include certain chromatin marks (H3K9me3, H3K36me3, etc.) and chromatin proteins associated with transcriptional elongation or repression.
- **Mixed-source** factors can bind in point-source fashion to some locations of the genome, but form broader domains of binding in others. RNA polymerase II, as well as some chromatin modifying proteins (e.g., SUZ12) behave in this way.

Wet lab

Problem: poor reactivity against the intended target and/or cross-reactivity with other DNA-associated proteins.
One of five criteria:

- factor "knockdown" by mutation or RNAi
- independent ChIP experiments using antibodies against more than one epitope on a protein or against different members of the same complex
- immunoprecipitation using epitope-tagged constructs
- affinity enrichment followed by mass spectrometry
- binding-site motif analysis

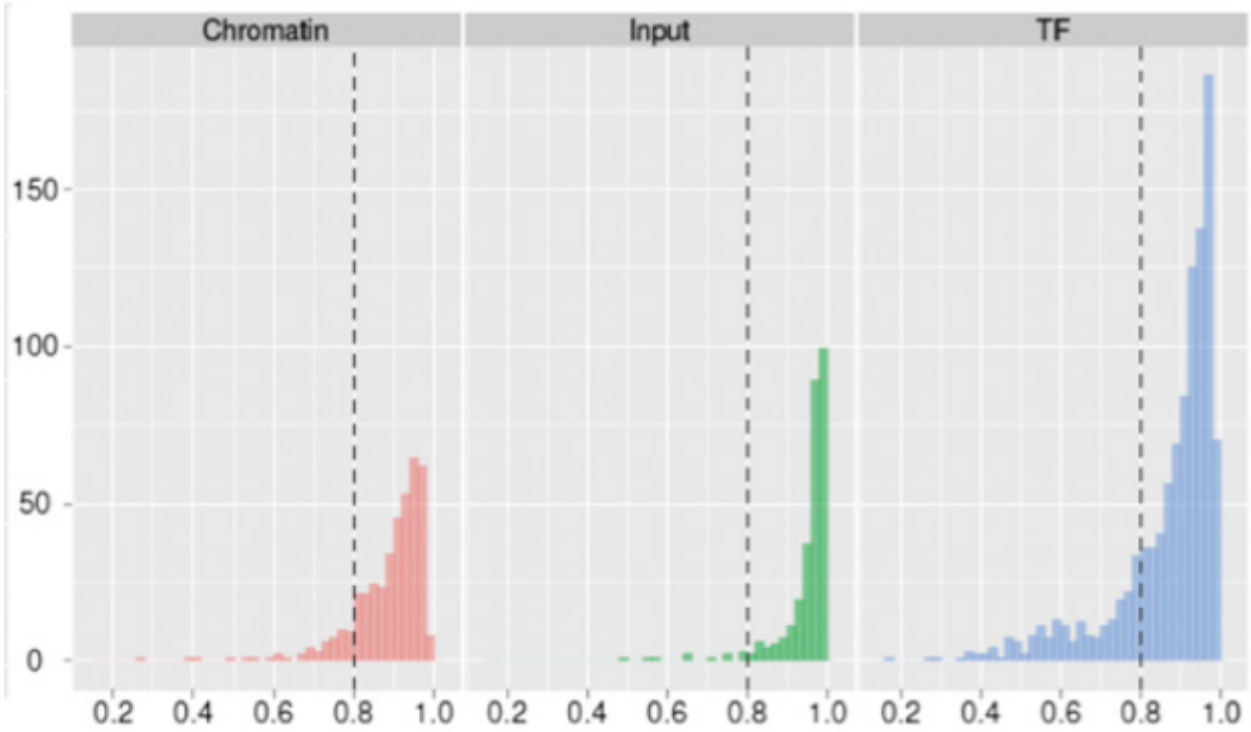
20% (44 of 227) of the tested commercially available antibodies against **transcription factors** meet these characterization guidelines and also function in ChIP-seq assays.

Replication, sequencing depth, library complexity

- RNA polymerase II ChIP-seq experiments showed that more than two replicates did not significantly improve site discovery. Each ChIP-Seq should performed on two independent biological replicates
- One cannot a priori set a specific target threshold for ChIP peak number or ChIP signal strength that will assure inclusion of all functional sites
- For point-source factors in mammalian cells, a minimum of 10 million uniquely mapped reads are used by ENCODE for each biological replicate.
- Input DNA or IgV control required - the same as the number of PCR amplification cycles, read length, etc.

NFR

Nonredundant fraction or NRF = the ratio between the number of positions in the genome that uniquely mappable reads map to and the total number of uniquely mappable reads.
Target is NRF $\geq$ 0.8 for 10 million uniquely mapped reads.
Distribution over ENCODE datasets:



Peak calling

- Punctate-source: several peak calling algorithms and corresponding software packages, including SPP, PeakSeq, and MACS. Too relaxed thresholds lead to a high proportion of false positives for each replicate, but subsequent replicates based analysis can strip false positives from a final joint peak determination.
- Broad-source factors or Mixed-source: more challenging. Methods to identify such regions are emerging: ZINBA, Scripture[5] and MACS2.[6]

[5] Originally developed for lincRNAs

[6] SICER not mentioned as of 2012

Peak calling Metrics

- Successful experiments generally identify thousands to tens of thousands of peaks for most TFs. A true signal is expected to show a clear asymmetrical distribution of reads mapping to the forward and reverse strands around the midpoint (peak) of accumulated reads.
- Visual control: inspection of mapped sequence reads using a genome browser
- FRiP - fraction of reads in peaks. ENCODE data sets have a FRiP enrichment of 1% - quality threshold.[7]
- Strand cross-correlation: Pearson linear correlation between the Crick strand and the Watson strand, after shifting Watson by k base pairs, where k is fragment size(fragment-length cross correlation).
- NSC - normalized strand coefficient. The normalized ratio between the fragment-length cross-correlation peak and the background cross-correlation (normalized strand coefficient) and the ratio between the fragment- length peak and the read-length peak (relative strand correlation, RSC).[8]

[7] ZNF274 and human RNA polymerase III have very few true binding sites and a FRiP of ¡1% is obtained

[8] TODO: not clear enough

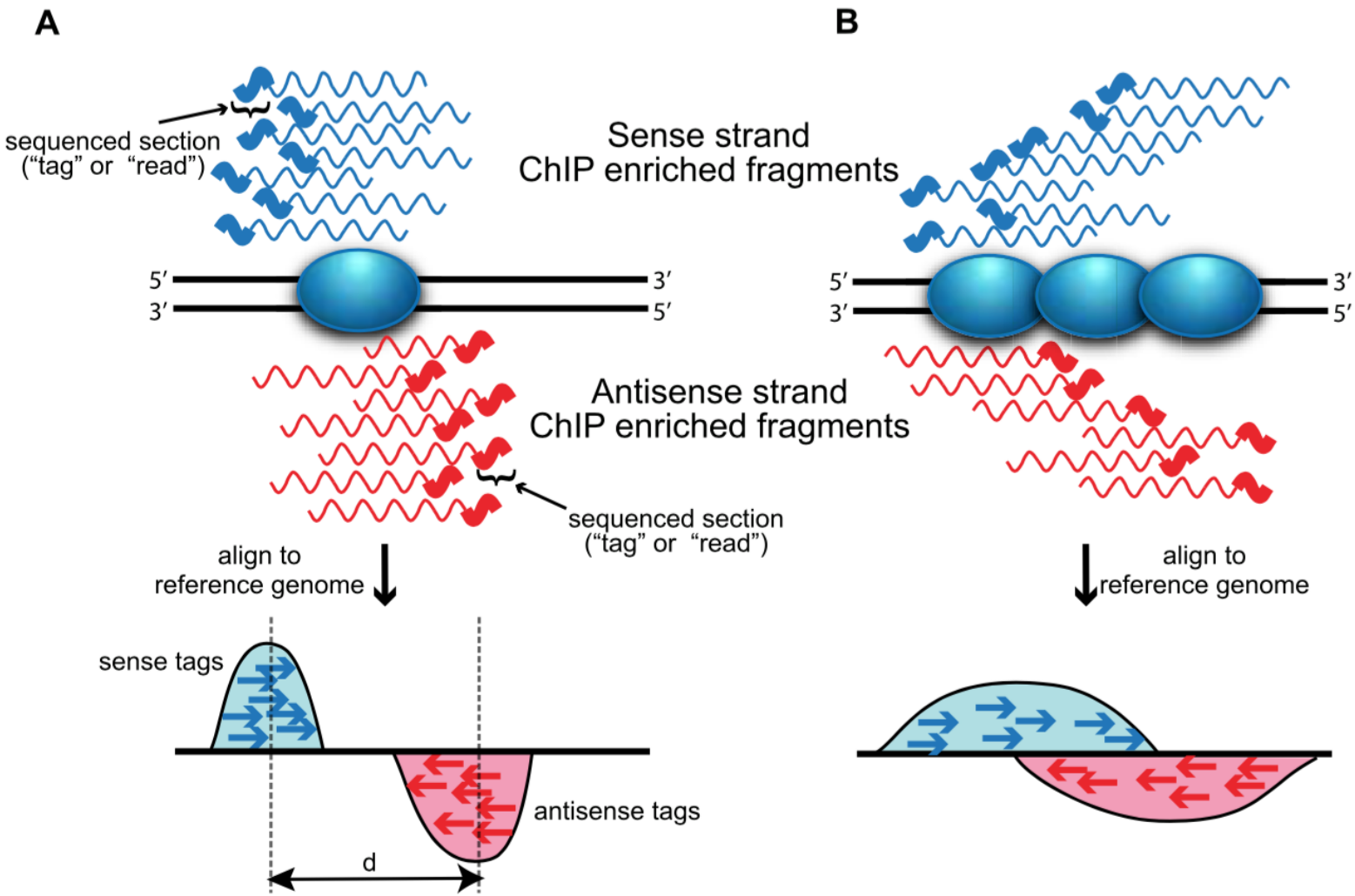


Figure 1. Strand-dependent bimodality in tag density. The 5′ to 3′ sequencing requirement and short read length produce stranded bias in tag distribution. The shaded blue oval represents the protein of interest bound to DNA (solid black lines). Wavy lines represent either sense (blue) or antisense (red) DNA fragments from ChIP enrichment. The thicker portion of the line indicates regions sequenced by short read sequencing technologies. Sequenced tags are aligned to a reference genome and projected onto a chromosomal coordinate (red and blue arrows). (**A**) Sequence-specific binding events (*e.g.* transcription factors) are characterized by "punctuate enrichment" [11] and defined strand-dependent bimodality, where the separation between peaks (d) corresponds to the average sequenced fragment length. Panel A was inspired by Jothi *et al.* [32]. (**B**) Distributed binding events (*e.g.* histones or RNA polymerase) produce a broader pattern of tag enrichment that results in a less defined bimodal pattern. doi:10.1371/journal.pone.0011471.g001 [9]

[9] http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0011471

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).[10]

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

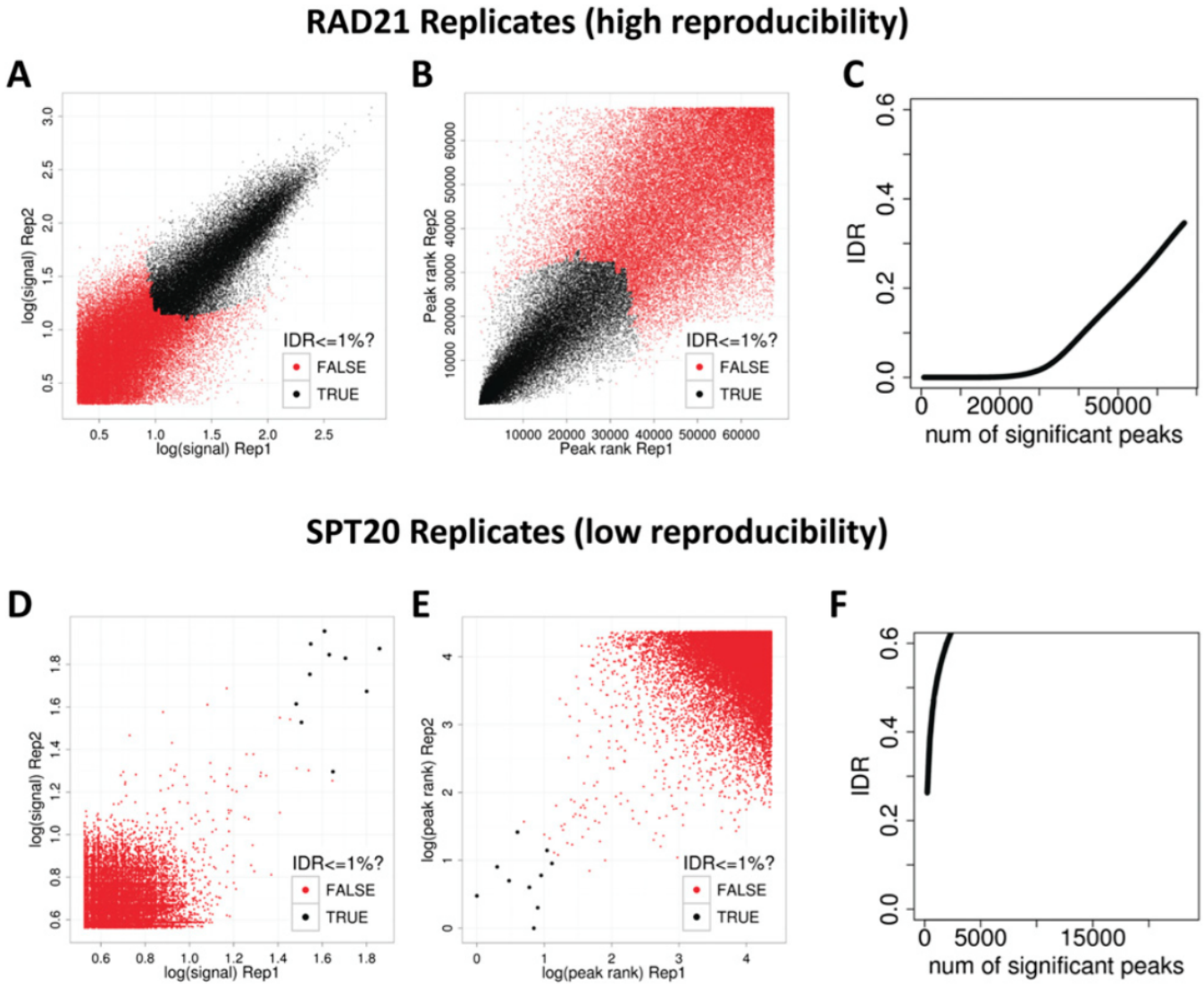


Figure 6. The irreproducible discovery rate (IDR) framework for assessing reproducibility of ChIP-seq data sets. (*A–C*) Reproducibility analysis for a pair of high-quality RAD21 ChIP-seq replicates. (*D,E*) The same analysis for a pair of low quality SPT20 ChIP-seq replicates. (*A,D*) Scatter plots 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.

Summary[12]

- Read depth $\geq$ 10 mln reads
- 2 biological replicates
- NFR[11]
- FRiP
- Strand cross-correlation
- NSC for sharp histone modifications
- IDR

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

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

OPEN ACCESS Freely available online

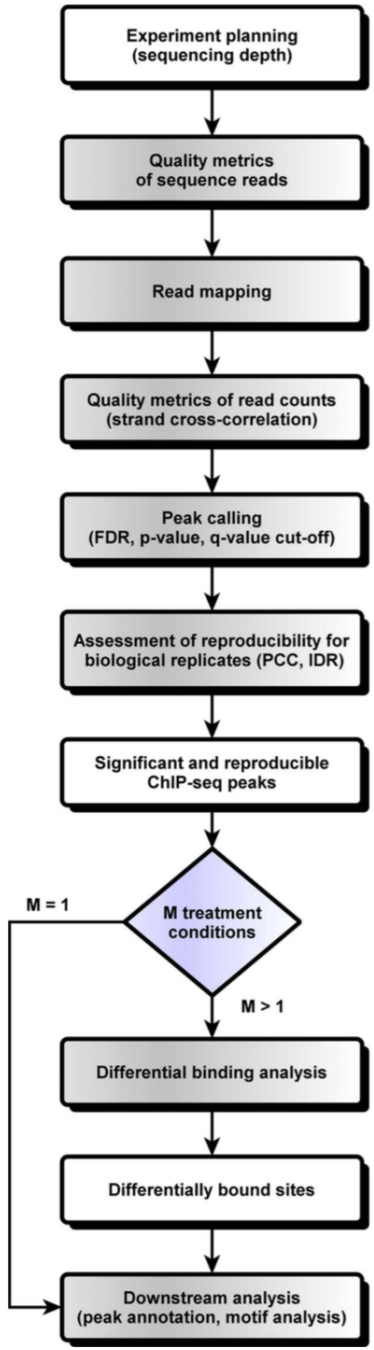
PLOS COMPUTATIONAL BIOLOGY

Education

Practical Guidelines for the Comprehensive Analysis of ChIP-seq Data

Timothy Bailey[1¶*], Pawel Krajewski[2¶], Istvan Ladunga[3¶], Celine Lefebvre[4¶], Qunhua Li[5¶], Tao Liu[6¶], Pedro Madrigal[2¶*], Cenny Taslim[7¶], Jie Zhang[7¶]

1 Institute for Molecular Bioscience, The University of Queensland, Brisbane, Australia, **2** Department of Biometry and Bioinformatics, Institute of Plant Genetics, Polish Academy of Sciences, Poznań, Poland, **3** Department of Statistics, Beadle Center, University of Nebraska-Lincoln, Lincoln, Nebraska, United States of America, **4** Inserm U981, Cancer Institute Gustave Roussy, Villejuif, France, **5** Department of Statistics, Penn State University, University Park, Pennsylvania, United States of America, **6** Department of Biochemistry, University at Buffalo, Buffalo, New York, United States of America, **7** Department of Biomedical Informatics, The Ohio State University, Columbus, Ohio, United States of America


Experiment planning
(sequencing depth)
Quality metrics
of sequence reads
Read mapping
Quality metrics of read counts
(strand cross-correlation)
Peak calling
(FDR, p-value, q-value cut-off)
Assessment of reproducibility for
biological replicates (PCC, IDR)
Significant and reproducible
ChIP-seq peaks
M treatment
conditions
M = 1
M > 1
Differential binding analysis
Differentially bound sites
Downstream analysis
(peak annotation, motif analysis)

Library complexity

- Sufficient read depth? - saturation analysis. The peaks called should be consistent performed on increasing numbers of reads chosen at random from the actual reads (built in SPP peak caller).
- Library complexity is a common quality measure for ChIP-seq libraries (preseq or PCR bottleneck coefficient from ENCODE tools[13]).
- NSC and RSC: very successful ChIP experiments generally have NSC ≥ 1.05 and RSC ≥ 0.8.
- IP strength: the software CHANCE assesses IP strength by estimating and comparing the IP reads pulled down by the antibody and the background, using a method called signal extraction scaling.
- Duplicate (identical) reads - a challenge. One can remove a certain number of duplicates to call confident peaks, and then put duplicates back to refine properties of these peaks such as peak height and boundaries.

[13] `https://code.google.com/p/phantompeakqualtools/`

Alignment

- The percentage varies between organisms, and for human, mouse, etc, above 70% uniquely mapped reads is normal, whereas less than 50% may be cause for concern.
- SNR - signal-to-noise ratio. SNR can be estimated by strand cross-correlation or IP enrichment estimation using the software package CHANCE. Strand cross-correlation analysis is built into some peak callers (e.g., SPP or MACS2).
- The reproducibility of the reads can be measured by computing the Pearson correlation coefficient (PCC) of the (mapped) read counts at each genomic position[14]. High quality experiments: PCC ≥ 0.9. (0.3 for unrelated samples).

[14] Only position with reads are taken into account

Peak calling

- Punctate-source: SPP and MACS2 use cross-correlation to find the lag between reads mapped to the minus and the plus strand as the size of actual protein-DNA interacting regions. After smoothing, background models are then used to remove noise either directly from the control sample or from features of the genome sequence such as GC content or mappability.
- Broad enriched: Several peak callers are specifically designed for predicting broad regions from ChIP-seq data, including SICER, CCAT, ZINBA, and RSEG.
- Mixed: Some tools have options for both narrow and broad peak calling, such as SPP, MACS, ZINBA, and PeakRanger.

Summary[15]

- Unique alignment rate
- Saturation analysis
- IP strength
- PCC

[15] Additional to previous

Peak calling

Guidelines above are better, these are old and don't cover some tools:

2010

OPEN ACCESS Freely available online

PLoS one

Evaluation of Algorithm Performance in ChIP-Seq Peak Detection

Elizabeth G. Wilbanks[1,3], Marc T. Facciotti[1,2,3]*

1 Graduate Group in Microbiology, University of California Davis, Davis, California, United States of America, **2** Department of Biomedical Engineering, University of California Davis, Davis, California, United States of America, **3** Genome Center, University of California Davis, Davis, California, United States of America

2014

OPEN ACCESS Freely available online

PLOS | ONE

A Comparison of Peak Callers Used for DNase-Seq Data

Hashem Koohy[1,2]*, Thomas A. Down[1], Mikhail Spivakov[2], Tim Hubbard[1]*

1 The Babraham Institute, Babraham Research Campus, Cambridge, United Kingdom, **2** Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, United Kingdom

2016

Bioinformatics. 2016 Oct 24. pii: btw672. doi: 10.1093/bioinformatics/btw672. [Epub ahead of print]

Optimizing ChIP-seq peak detectors using visual labels and supervised machine learning.

Hocking TD[1], Goerner-Potvin P[1], Morin A[1], Shao X[1], Pastinen T[1], Bourque G[1].

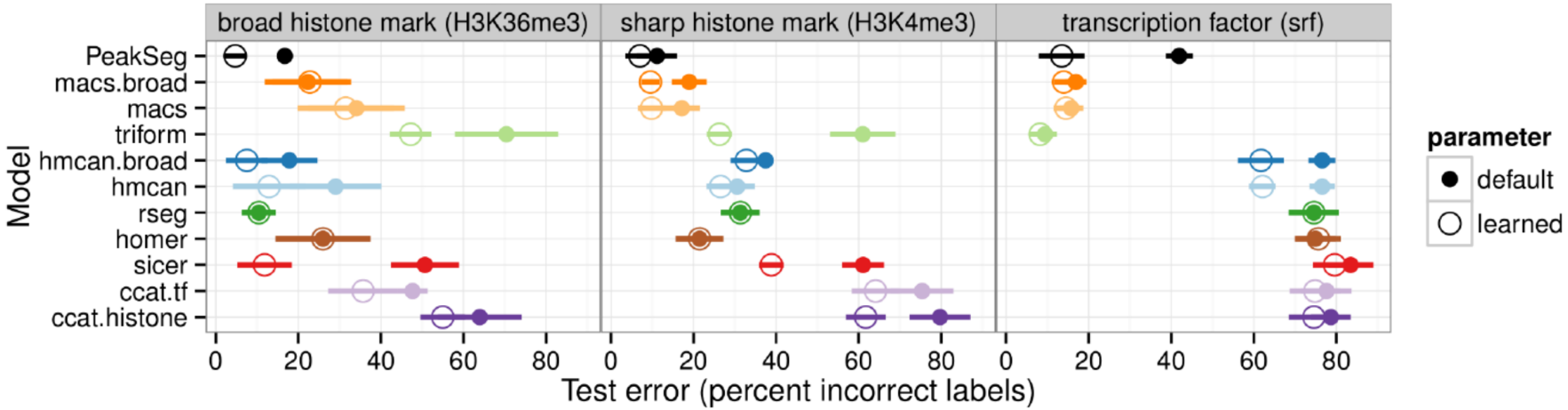


Fig. 5. Model parameters which are learned using labels provide more accurate peak predictions than default model parameters. Four-fold cross-validation was used to estimate test error rates (mean ± standard deviation over four test folds). Closed circles represent default parameters (labels not used for model training), and open circles represent learned parameters (with minimal incorrect labels in each training data set). It is clear that some algorithms are accurate in several data types and others only work in one data type.

Popular

MACS2[16]

[HTML] **Model-based** analysis of ChIP-Seq (**MACS**)

Y Zhang, T Liu, CA Meyer, J Eeckhoute… - Genome …, 2008 - biomedcentral.com

Abstract We present **Model-based** Analysis of ChIP-Seq data, **MACS**, which analyzes data generated by short read sequencers such as Solexa's Genome Analyzer. **MACS** empirically models the shift size of ChIP-Seq tags, and uses it to improve the spatial resolution of **...**

Цитируется: 2590 Похожие статьи Все версии статьи (26) Цитировать Сохранить Ещё

Scipture

Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs

M Guttman, M Garber, JZ Levin, J Donaghey… - Nature …, 2010 - nature.com

Massively parallel cDNA sequencing (RNA-Seq) provides an unbiased way to study a transcriptome, including both coding and noncoding genes. Until now, most RNA-Seq studies have depended crucially on existing annotations and thus focused on expression ...

Цитируется: 753 Похожие статьи Все версии статьи (17) Цитировать Сохранить Ещё

PeakSeq

PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls

J Rozowsky, G Euskirchen, RK Auerbach… - Nature …, 2009 - nature.com

Abstract Chromatin immunoprecipitation (ChIP) followed by tag sequencing (ChIP-seq) using high-throughput next-generation instrumentation is fast, replacing chromatin immunoprecipitation followed by genome tiling array analysis (ChIP-chip) as the preferred **...**

Цитируется: 467 Похожие статьи Все версии статьи (17) Цитировать Сохранить Ещё

[16] Roadmapepigenomics: Integrative analysis of 111 reference human epigenomes

Less popular

SPP

Design and analysis of ChIP-seq experiments for DNA-binding proteins

PV Kharchenko, MY Tolstorukov, PJ Park - Nature biotechnology, 2008 - nature.com

Abstract Recent progress in massively parallel sequencing platforms has enabled genome-wide characterization of DNA-associated proteins using the combination of chromatin immunoprecipitation and sequencing (ChIP-seq). Although a variety of methods exist for ...

Цитируется: 431 Похожие статьи Все версии статьи (15) Цитировать Сохранить Ещё

SICER

A clustering approach for identification of enriched domains from histone modification ChIP-Seq data

C Zang, DE Schones, C Zeng, K Cui, K Zhao… - …, 2009 - Oxford Univ Press

... Using a scaling approach for evaluation of false positives that is based on the digitized nature of **ChIP-Seq** data, and two datasets with experimental validation, we demonstrated that **SICER** outperforms other **ChIP-Seq** methods in dealing with histone modification data. **...**

Цитируется: 400 Похожие статьи Все версии статьи (17) Цитировать Сохранить Ещё

CCAT

[HTML] A signal–noise model for significance analysis of ChIP-seq with negative control

H Xu, L Handoko, X Wei, C Ye, J Sheng, CL Wei… - …, 2010 - Oxford Univ Press

Abstract Motivation: ChIP-seq is becoming the main approach to the genome-wide study of protein–DNA interactions and histone modifications. Existing informatics tools perform well to extract strong ChIP-enriched sites. However, two questions remain to be answered:(i) to ...

Цитируется: 107 Похожие статьи Все версии статьи (12) Цитировать Сохранить Ещё

Even less popular

ZINBA[17]

[PDF] ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions

NU Rashid, PG Giresi, JG Ibrahim, W Sun... - Genome ..., 2011 - biomedcentral.com

Abstract ZINBA (Zero-Inflated Negative Binomial Algorithm) identifies genomic regions enriched in a variety of ChIP-seq and related next-generation sequencing experiments (DNA-seq), calling both broad and narrow modes of enrichment across a range of signal- ...

Цитируется: 88 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

PeakSeg

[PDF] **PeakSeg**: constrained optimal segmentation and supervised penalty learning for peak detection in count data.

T Hocking, G Rigaill, G Bourque - ICML, 2015 - jmlr.org

Abstract Peak detection is a central problem in genomic data analysis, and current algorithms for this task are unsupervised and mostly effective for a single data type and pattern (eg H3K4me3 data with a sharp peak pattern). We propose **PeakSeg**, a new

Цитируется: 2 Похожие статьи Все версии статьи (6) Цитировать Сохранить Ещё [18]

[17] too slow `http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0096303`

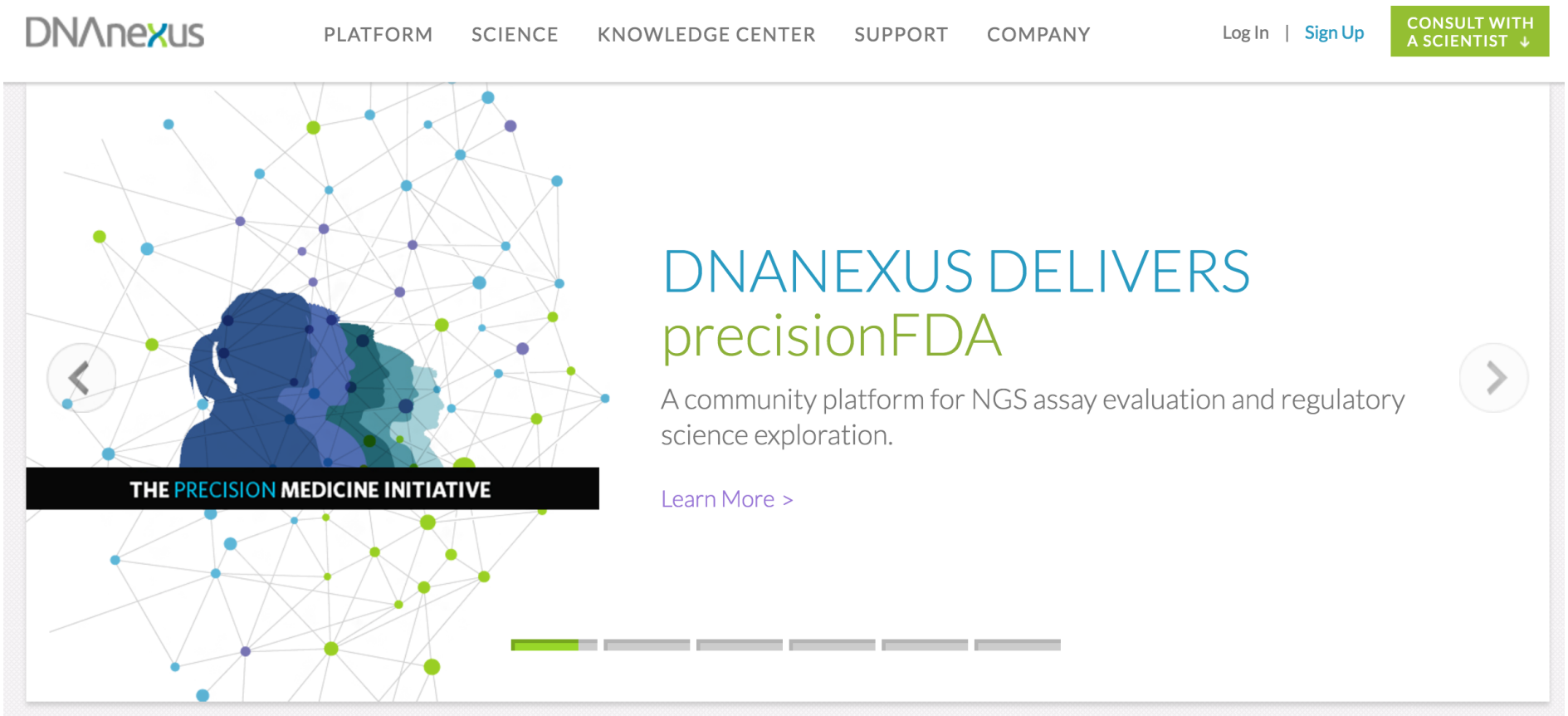
[18] supervised method

Peak calling summary

Consensus of 3 comparison papers:

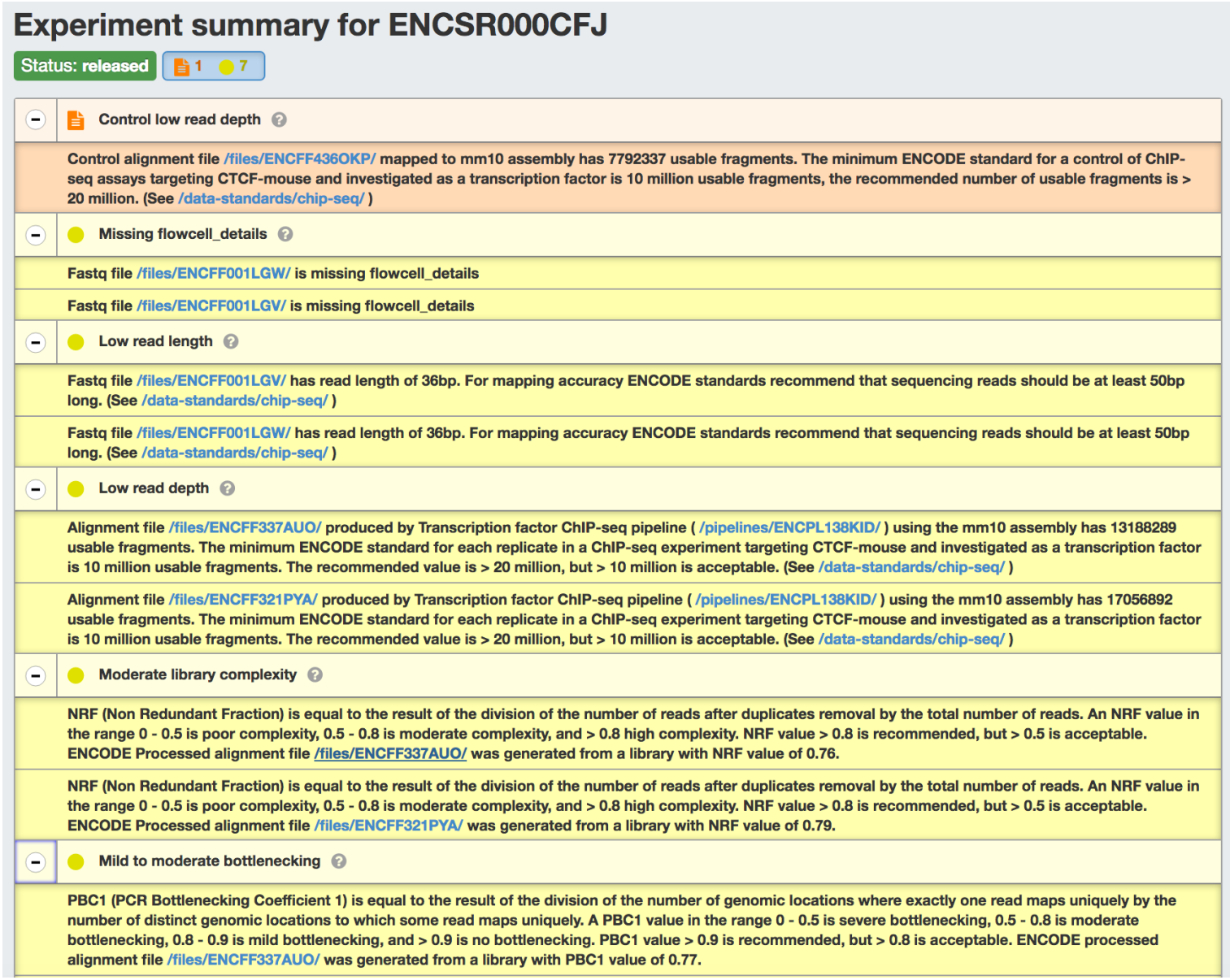
- Narrow peaks: MACS2 or SPP or PeakSeg
- Broad and mixed: SICER or ZINBA or PeakSeg or RSEG

DNANexus[19]



[19] https://www.dnanexus.com

ENCODE pipeline warnings[20]



Experiment summary for ENCSR000CFJ

Status: released | 1 | 7

Control low read depth

Control alignment file /files/ENCFF436OKP/ mapped to mm10 assembly has 7792337 usable fragments. The minimum ENCODE standard for a control of ChIP-seq assays targeting CTCF-mouse and investigated as a transcription factor is 10 million usable fragments, the recommended number of usable fragments is > 20 million. (See /data-standards/chip-seq/)

Missing flowcell_details

Fastq file /files/ENCFF001LGW/ is missing flowcell_details

Fastq file /files/ENCFF001LGV/ is missing flowcell_details

Low read length

Fastq file /files/ENCFF001LGV/ has read length of 36bp. For mapping accuracy ENCODE standards recommend that sequencing reads should be at least 50bp long. (See /data-standards/chip-seq/)

Fastq file /files/ENCFF001LGW/ has read length of 36bp. For mapping accuracy ENCODE standards recommend that sequencing reads should be at least 50bp long. (See /data-standards/chip-seq/)

Low read depth

Alignment file /files/ENCFF337AUO/ produced by Transcription factor ChIP-seq pipeline (/pipelines/ENCPL138KID/) using the mm10 assembly has 13188289 usable fragments. The minimum ENCODE standard for each replicate in a ChIP-seq experiment targeting CTCF-mouse and investigated as a transcription factor is 10 million usable fragments. The recommended value is > 20 million, but > 10 million is acceptable. (See /data-standards/chip-seq/)

Alignment file /files/ENCFF321PYA/ produced by Transcription factor ChIP-seq pipeline (/pipelines/ENCPL138KID/) using the mm10 assembly has 17056892 usable fragments. The minimum ENCODE standard for each replicate in a ChIP-seq experiment targeting CTCF-mouse and investigated as a transcription factor is 10 million usable fragments. The recommended value is > 20 million, but > 10 million is acceptable. (See /data-standards/chip-seq/)

Moderate library complexity

NRF (Non Redundant Fraction) is equal to the result of the division of the number of reads after duplicates removal by the total number of reads. An NRF value in the range 0 - 0.5 is poor complexity, 0.5 - 0.8 is moderate complexity, and > 0.8 high complexity. NRF value > 0.8 is recommended, but > 0.5 is acceptable. ENCODE Processed alignment file /files/ENCFF337AUO/ was generated from a library with NRF value of 0.76.

NRF (Non Redundant Fraction) is equal to the result of the division of the number of reads after duplicates removal by the total number of reads. An NRF value in the range 0 - 0.5 is poor complexity, 0.5 - 0.8 is moderate complexity, and > 0.8 high complexity. NRF value > 0.8 is recommended, but > 0.5 is acceptable. ENCODE Processed alignment file /files/ENCFF321PYA/ was generated from a library with NRF value of 0.79.

Mild to moderate bottlenecking

PBC1 (PCR Bottlenecking Coefficient 1) is equal to the result of the division of the number of genomic locations where exactly one read maps uniquely by the number of distinct genomic locations to which some read maps uniquely. A PBC1 value in the range 0 - 0.5 is severe bottlenecking, 0.5 - 0.8 is moderate bottlenecking, 0.8 - 0.9 is mild bottlenecking, and > 0.9 is no bottlenecking. PBC1 value > 0.9 is recommended, but > 0.8 is acceptable. ENCODE processed alignment file /files/ENCFF337AUO/ was generated from a library with PBC1 value of 0.77.

[20] https://www.encodeproject.org/experiments/ENCSR000CFJ/

ENCODE ChIP-Seq standards[21]

ENCODE3 Standards

- Experiments should have two or more biological replicates, isogenic or anisogenic. Assays performed using EN-TEx samples may be exempted due to limited availability of experimental material.
- Antibodies must be characterized according to standards set by the ENCODE Consortium. Please see the linked documents for transcription factor standards (May 2016), histone modification and chromatin-associated protein standards (October 2016), and RNA binding protein standards (November 2016).
- Each ChIP-seq experiment should have a corresponding input control experiment with matching run type, read length, and replicate structure.
- Library complexity is measured using the Non-Redundant Fraction (NRF) and PCR Bottlenecking Coefficients 1 and 2, or PBC1 and PBC2. Preferred values are as follows: NRF>0.9, PBC1>0.9, and PBC2>10.
- The experiment must pass routine metadata audits in order to be released.

Uniform Processing Pipeline Restrictions

- The read length should be a minimum of 50 base pairs, though longer read lengths are encouraged.
- The sequencing platform used should be indicated.
- Replicates should match in terms of read length and run type.
- Pipeline files are mapped to either the GRCh38 or mm10 sequences.

Target-specific Standards

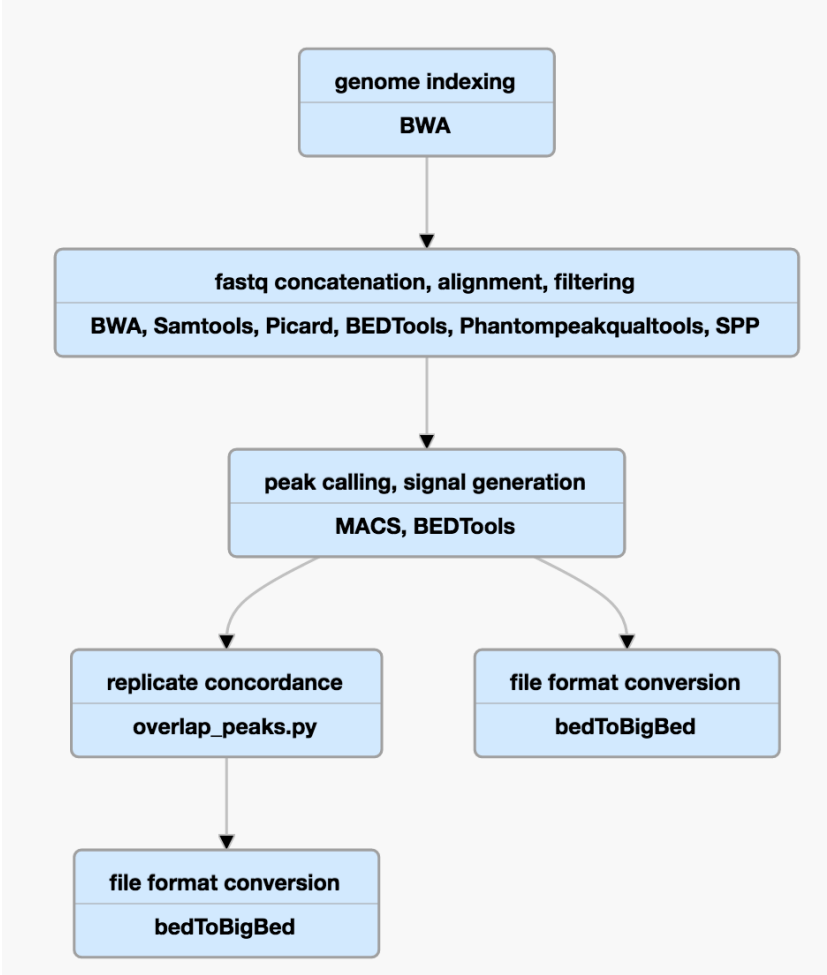
- For narrow-peak histone experiments, each replicate should have 20 million usable fragments.
- For broad-peak histone experiments, each replicate should have 45 million usable fragments.
- For transcription factor experiments, each replicate should have 20 million usable fragments.
- H3K9me3 is an exception as it is enriched in repetitive regions of the genome. Compared to other broad marks, there are few H3K9me3 peaks in non-repetitive regions of the genome in tissues and primary cells. This results in many ChIP-seq reads that map to a non-unique position in the genome. Tissues and primary cells should have 45 million total mapped reads per replicate.

Broad Marks	H3F3A	H3K27me3	H3K36me3	H3K4me1	H3K79me2	H3K79me3	H3K9me1	H3K9me2	H4K20me1
Narrow Marks	H2AFZ	H3ac	H3K27ac	H3K4me2	H3K4me3	H3K9ac			
Exceptions	H3K9me3								

- For transcription factor experiments, replicate concordance is measured by calculating IDR values (Irreproducible Discovery Rate). The experiment passes if both rescue and self consistency ratios are less than 2.

[21] https://www.encodeproject.org/data-standards/chip-seq/

ENCODE DNANexus pipeline[22]



Source: `https://github.com/ENCODE-DCC/chip-seq-pipeline`

[22] `https://www.encodeproject.org/chip-seq/histone/`

Thank you!

oleg.shpynov@jetbrains.com
https://research.jetbrains.org/groups/biolabs