
ChIP-seq-analysis

Snakemake pipelines

I developed a Snakemake based ChIP-seq pipeline: `pyflow-ChIPseq`. and ATACseq pipeline: `pyflow-ATACseq`

Resources for ChIP-seq

1. ENCODE: Encyclopedia of DNA Elements ENCODEExplorer: A compilation of metadata from ENCODE. A bioc package to access the meta data of ENCODE and download the raw files.
2. ENCODE Factorbook
3. ChromNet ChIP-seq interactions
paper: Learning the human chromatin network using all ENCODE ChIP-seq datasets
4. The International Human Epigenome Consortium (IHEC) epigenome data portal
5. GEO. Sequences are in .sra format, need to use `sratools` to dump into fastq.
6. European Nucleotide Archive. Sequences are available in fastq format.
7. Data bases and software from Sheirly Liu's lab at Harvard
8. Blueprint epigenome
9. A collection of tools and papers for nucelosome positioning and TF ChIP-seq
10. review paper:Deciphering ENCODE
11. EpiFactors is a database for epigenetic factors, corresponding genes and products.
12. biostar handbook. My ChIP-seq chapter is out April 2017!
13. ReMap 2018 An integrative ChIP-seq analysis of regulatory regions. The ReMap atlas consits of 80 million peaks from 485 transcription factors (TFs), transcription coactivators (TCAs) and chromatin-remodeling factors (CRFs) from public data sets. The atlas is available to browse or download either for a given TF or cell line, or for the entire dataset.

Papers on ChIP-seq

1. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia
2. Practical Guidelines for the Comprehensive Analysis of ChIP-seq Data
3. Systematic evaluation of factors influencing ChIP-seq fidelity
4. ChIP-seq: advantages and challenges of a maturing technology

-
5. ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions
 6. Beyond library size: a field guide to NGS normalization
 7. ENCODE paper portol
 8. Enhancer discovery and characterization
 9. 2016 review Recent advances in ChIP-seq analysis: from quality management to whole-genome annotation
 10. bioinformatics paper: Features that define the best ChIP-seq peak calling algorithms compares different peak callers for TFs and histones.
 11. Systematic comparison of monoclonal versus polyclonal antibodies for mapping histone modifications by ChIP-seq The binding patterns for H3K27ac differed substantially between polyclonal and monoclonal antibodies. However, this was most likely due to the distinct immunogen used rather than the clonality of the antibody. Altogether, we found that monoclonal antibodies as a class perform as well as polyclonal antibodies. Accordingly, we recommend the use of monoclonal antibodies in ChIP-seq experiments.
 12. A nice small review: Unraveling the 3D genome: genomics tools for multiscale exploration
 13. Three very interesting papers, Developmental biology: Panoramic views of the early epigenome
 14. ChIP off the old block: Beyond chromatin immunoprecipitation. A nice review of the past and future of ChIPseq.
 15. Histone Modifications: Insights into Their Influence on Gene Expression **Protocols**
 16. A computational pipeline for comparative ChIP-seq analyses
 17. Identifying ChIP-seq enrichment using MACS
 18. Spatial clustering for identification of ChIP-enriched regions (SICER) to map regions of histone methylation patterns in embryonic stem cells
 19. ENCODE tutorials
 20. A User's Guide to the Encyclopedia of DNA Elements (ENCODE)
 21. A toolbox of immunoprecipitation-grade monoclonal antibodies to human transcription factors
The data portal <https://proteincapture.org/>

Quality Control

Data downloaded from GEO usually are raw fastq files. One needs to do quality control (QC) on them.

-
- fastqc
 - multiqc Aggregate results from bioinformatics analyses across many samples into a single report. Could be very useful to summarize the QC report.

Peak calling

Be careful with the peaks you get:

Active promoters give rise to false positive 'Phantom Peaks' in ChIP-seq experiments

It is good to have controls for your ChIP-seq experiments. A DNA input control (no antibody is applied) is preferred. The IgG control is also fine, but because so little DNA is there, you might get many duplicated reads due to PCR artifact.

For cancer cells, an input control can be used to correct for copy-number bias.

- tools used by IHEC consortium

A quote from Tao Liu: who developed MACS1/2

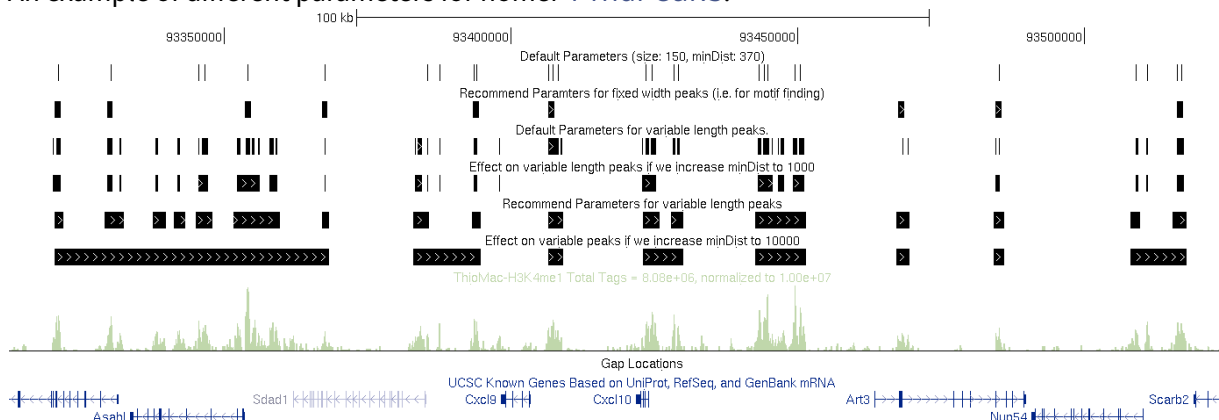
I remember in a PloS One paper last year by Elizabeth G. Wilbanks et al., authors pointed out the best way to sort results in MACS is by $-10 \cdot \log_{10}(\text{pvalue})$ then fold enrichment. I agree with them. You don't have to worry about FDR too much if your input data are far more than ChIP data. MACS1.4 calculates FDR by swapping samples, so if your input signal has some strong bias somewhere in the genome, your FDR result would be bad. Bad FDR may mean something but it's just secondary.

1. The most popular peak caller by Tao Liu: MACS2. Now `--broad` flag supports broad peaks calling as well.
2. TF ChIP-seq peak calling using the Irreproducibility Discovery Rate (IDR) framework and many Software Tools Used to Create the ENCODE Resource
3. SICER for broad histone modification ChIP-seq
4. HOMER can also be used to call Transcription factor ChIP-seq peaks and histone modification ChIP-seq peaks.
5. MUSIC
6. permseq R package for mapping protein-DNA interactions in highly repetitive regions of the genomes with prior-enhanced read mapping. Paper on PLoS Comp.
7. Ritornello: High fidelity control-free chip-seq peak calling. No input is required!

8. Tumor samples are heterogeneous containing different cell types. MixChIP: a probabilistic method for cell type specific protein-DNA binding analysis
9. Detecting broad domains and narrow peaks in ChIP-seq data with hiddenDomains tool
10. BroadPeak: a novel algorithm for identifying broad peaks in diffuse ChIP-seq datasets
11. epic: diffuse domain ChIP-Seq caller based on SICER. It is a re-written of SICER for faster processing using more CPUs. (Will try it for broad peak for sure). epic2 paper is out <https://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/btz232/54215>
12. Cistrome: The best place for wet lab scientist to check the binding sites. Developed by Shierly Liu lab in Harvard.
13. ChIP-Atlas is an integrative and comprehensive database for visualizing and making use of public ChIP-seq data. ChIP-Atlas covers almost all public ChIP-seq data submitted to the SRA (Sequence Read Archives) in NCBI, DDBJ, or ENA, and is based on over 78,000 experiments.
14. A map of direct TF-DNA interactions in the human genome UniBind is a comprehensive map of direct interactions between transcription factor (TFs) and DNA. High confidence TF binding site predictions were obtained from uniform processing of thousands of ChIP-seq data sets using the ChIP-eat software.
15. Accounting for GC-content bias reduces systematic errors and batch effects in ChIP-Seq peak callers tool in github
16. SUPERmerge:ChIP-seq coverage island analysis algorithm for broad histone marks
17. PeakRanger heard that it is good for broad peaks of H3K9me3 and H3K27me3.

Different parameters using the same program can produce drastic different sets of peaks especially for histone modifications with variable enrichment length and gaps between peaks. One needs to make a valid argument for parameters he uses

An example of different parameters for homer `findPeaks`:



Tutorial

- tutorial by Simon van Heeringen at bioinfosummer

Binding does not infer functionality

- A significant proportion of transcription-factor binding sites may be nonfunctional A post from Judge Starling
- Several papers have shown that changes of adjacent TF binding poorly correlates with gene expression change: Extensive Divergence of Transcription Factor Binding in Drosophila Embryos with Highly Conserved Gene Expression
Transcription Factors Bind Thousands of Active and Inactive Regions in the Drosophila Blastoderm

The Functional Consequences of Variation in Transcription Factor Binding

>” On average, 14.7% of genes bound by a factor were differentially expressed following the knock-down of that factor, suggesting that most interactions between TF and chromatin do not result in measurable changes in gene expression levels of putative target genes. ”

- paper A large portion of the ChIP-seq signal does not correspond to true binding
- BIDCHIPS: Bias-Decomposition of ChIP-seq Signals
mappability, GC-content and chromatin accessibility affect ChIP-seq read counts.
- ChIP bias as a function of cross-linking time

We analyzed the dependence of the ChIP signal on the duration of formaldehyde cross-linking time for two proteins: DNA topoisomerase 1 (Top1) that is functionally associated with the double helix in vivo, especially with active chromatin, and green fluorescent protein (GFP) that has no known bona fide interactions with DNA. With short time of formaldehyde fixation, only Top1 immunoprecipitation efficiently recovered DNA from active promoters, whereas prolonged fixation augmented non-specific recovery of GFP dramatizing the need to optimize ChIP protocols to minimize the time of cross-linking, especially for abundant nuclear proteins. Thus, ChIP is a powerful approach to study the localization of protein on the genome when care is taken to manage potential artifacts.

Gene set enrichment analysis for ChIP-seq peaks

The Gene Ontology Handbook Read it for basics for GO.

-
1. Broad Enrich
 2. ChIP Enrich
 3. GREAT predicts functions of cis-regulatory regions.
 4. ENCODE ChIP-seq significance tool. Given a list of genes, co-regulating TFs will be identified.
 5. cscan similar to the ENCODE significance tool.
 6. CompGO: an R package for comparing and visualizing Gene Ontology enrichment differences between DNA binding experiments
 7. interactive and collaborative HTML5 gene list enrichment analysis tool
 8. GeNets from Broad. Looks very promising.
 9. Bioconductor EnrichmentBrowser
 10. clusterProfiler by Guangchuan Yu, the author of [ChIPseeker](#).
 11. fgsea bioconductor package Fast Gene Set Enrichment Analysis.
 12. paper: A Comparison of Gene Set Analysis Methods in Terms of Sensitivity, Prioritization and Specificity
 13. UniBind Enrichment Analysis predicts which sets of TFBSs from the UniBind database are enriched in a set of given genomic regions. Enrichment computations are performed using the LOLA tool.
 14. BEHST from Hoffman group: genomic set enrichment analysis enhanced through integration of chromatin long-range interactions
 15. ChEA3: transcription factor enrichment analysis by orthogonal omics integration

Chromatin state Segmentation

1. ChromHMM from Manolis Kellis in MIT. >In ChromHMM the raw reads are assigned to non-overlapping bins of 200 bps and a sample-specific threshold is used to transform the count data to binary values
2. Segway from Hoffman lab. Base pair resolution. Takes longer time to run.
3. epicseg published 2015 in genome biology. Similar speed with ChromHMM.
4. Spectacle: fast chromatin state annotation using spectral learning. Also published 2015 in

genome biology.

5. chromstaR: Tracking combinatorial chromatin state dynamics in space and time
6. epilogos visualization and analysis of chromatin state model data.
7. Accurate promoter and enhancer identification in 127 ENCODE and Roadmap Epigenomics cell types and tissues by GenoSTAN
8. StatePaintR StateHub-StatePaintR: rules-based chromatin state annotations.
9. [IDEAS(<https://github.com/yuzhang123/IDEAS/>): an integrative and discriminative epigenome annotation system <http://sites.stat.psu.edu/~yzz2/IDEAS/>

deep learning in ChIP-seq

- Coda uses convolutional neural networks to learn a mapping from noisy to high-quality ChIP-seq data. These trained networks can then be used to remove noise and improve the quality of new ChIP-seq data. From Ashul lab.
- DeepChrome is a unified CNN framework that automatically learns combinatorial interactions among histone modification marks to predict the gene expression. (Is it really better than a simple linear model?)
- deep learning in biology

Peak annotation

1. Homer [annotatePeak](#)
2. Bioconductor package ChIPseeker by Guangchuan Yu
See an important post by him on 0 or 1 based coordinates.

Most of the software for ChIP annotation doesn't consider this issue when annotating peak (0-based) to transcript (1-based). To my knowledge, only HOMER consider this issue. After I figure this out, I have updated ChIPseeker (version $\geq 1.4.3$) to fix the issue.

3. Bioconductor package ChIPpeakAnno.
4. annotatr Annotation of Genomic Regions to Genomic Annotations.
5. geneXtender computes optimal gene extensions tailored to the broadness of the specific epigenetic mark (e.g., H3K9me1, H3K27me3), as determined by a user-supplied ChIP-seq peak input file. As such, geneXtender maximizes the signal-to-noise ratio of locating genes closest to and directly under peaks

- DNASHapeR predicts DNA shape features in an ultra-fast, high-throughput manner from genomic sequencing data

Differential peak detection

Look at a post and here describing different tools. A review paper A comprehensive comparison of tools for differential ChIP-seq analysis

12 | Steinhauser et al.

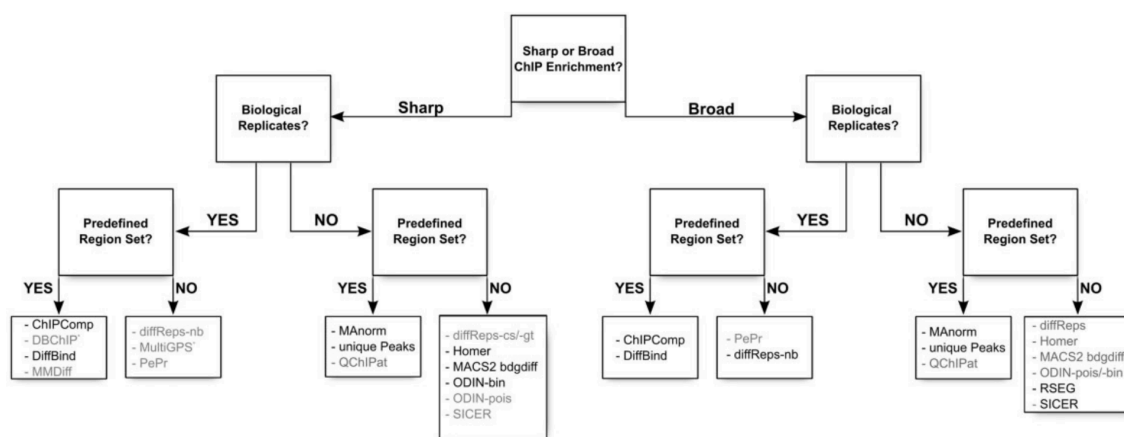


Figure 7. Decision tree indicating the proper choice of tool depending on the data set: shape of the signal (sharp peaks or broad enrichments), presence of replicates and presence of an external set of regions of interest. We have indicated in dark the name of the tools that give good results using default 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.

- ATAC-seq normalization method can significantly affect differential accessibility analysis and interpretation
- Comparison of differential accessibility analysis strategies for ATAC-seq data <https://github.com/Zhanglab/BeCorrect> to correct batch effect from the bedgraph files.

1. MultiGPS
2. PePr. It can also call peaks.
3. histoneHMM
4. diffreps for histone. developed by Shen Li's lab in Mount Sinai who also developed ngs.plot.
5. diffbind bioconductor package. Internally uses RNA-seq tools: EdgR or DESeq. Most likely, I will use this tool.
6. ChIPComp. Very little tutorial. Now it is on bioconductor.

-
7. csaw bioconductor package. Tutorial [here](#)
 8. chromDiff. Also from from Manolis Kellis in MIT. Similar with ChromHMM, documentation is not that detailed. Will have a try on this.
 9. MACS2 can detect differential peaks as well
 10. paper Identifying differential transcription factor binding in ChIP-seq

Motif enrichment

1. HOMER. It has really detailed documentation. It can also be used to call peaks.

For TF ChIP-seq, one can usually find the summit of the peak (macs14 will report the summit), and extend the summit to both sides to 100bp-500bp. One can then use those 100bp-500 bp small regions to do motif analysis. Usually, oen should find the motif for the ChIPed TF in the ChIP-seq experiment if it is a DNA binding protein.

It is trickier to do motif analysis using histone modification ChIP-seq. For example, the average peak size of H3K27ac is 2~3 kb. If one wants to find TF binding motifs from H3K27ac ChIP-seq data, it is good to narrow down the region a bit. MEME and many other motif finding tools require that the DNA sequence length to be small (~500bp). One way is to use [findPeaks](#) in homer turning on `-nfr` (nucleosome free region) flag, and then do motif analysis in those regions.

suggestions for finding motifs from histone modification ChIP-seq data from HOMER page: >Since you are looking at a region, you do not necessarily want to center the peak on the specific position with the highest tag density, which may be at the edge of the region. Besides, in the case of histone modifications at enhancers, the highest signal will usually be found on nucleosomes surrounding the center of the enhancer, which is where the functional sequences and transcription factor binding sites reside. Consider H3K4me marks surrounding distal PU.1 transcription factor peaks. Typically, adding the `-center` >option moves peaks further away from the functional sequence in these scenarios.

Other strategy similar to `-nfr` was developed in this paper: Dissecting neural differentiation regulatory networks through epigenetic footprinting. In the method part of the paper, the authors computed a depletion score within the peaks, and use the footprinted regions to do motif analysis. (Thanks kadir for pointing out the paper)

<http://homer.ucsd.edu/homer/ngs/peakMotifs.html>

Region Size (“-size <#>”, “-size <#>,<#>”, “-size given”, default: 200) The size of the region used for motif finding is important. If analyzing ChIP-Seq peaks from a transcription factor, Chuck would recommend 50 bp for establishing the primary motif bound by a given transcription factor and 200 bp for finding both primary and “co-enriched” motifs for a transcription factor. When looking at histone marked regions, **500-1000 bp is probably a good idea (i.e. H3K4me or H3/H4 acetylated regions)**. In theory, HOMER can work with very large regions (i.e. 10kb), but with the larger the regions comes more sequence and longer execution time. These regions will be based off the center of the peaks. If you prefer an offset, you can specify “-size -300,100” to search a region of size 400 that is centered 100 bp upstream of the peak center (useful if doing motif finding on putative TSS regions). If you have variable length regions, use the option “-size given” and HOMER will use the exact regions that were used as input.

I just found PARE. PARE is a computational method to Predict Active Regulatory Elements, specifically enhancers and promoters. H3K27ac and H3K4me can be used to define active enhancers.

2. MEME suite. It is probably the most popular motif finding tool in the papers. protocol:Motif-based analysis of large nucleotide data sets using MEME-ChIP
3. MEME R package
4. JASPAR database
5. pScan-ChIP
6. MotifMap
7. RAST Regulatory Sequence Analysis Tools.
8. ENCODE TF motif database
9. oPOSSUM is a web-based system for the detection of over-represented conserved transcription factor binding sites and binding site combinations in sets of genes or sequences.
10. my post how to get a genome-wide motif bed file
11. Many other tools here
12. A review of ensemble methods for de novo motif discovery in ChIP-Seq data
13. melina2. If you only have one sequence and want to know what TFs might bind there, this is a very useful tool.
14. STEME. A python library for motif analysis. STEME started life as an approximation to the

Expectation-Maximisation algorithm for the type of model used in motif finders such as MEME. **STEME's EM approximation runs an order of magnitude more quickly than the MEME implementation for typical parameter settings.** STEME has now developed into a fully-fledged motif finder in its own right.

15. CENTIPEDE: Transcription factor footprinting and binding site prediction. Tutorial
16. msCentipede: Modeling Heterogeneity across Genomic Sites and Replicates Improves Accuracy in the Inference of Transcription Factor Binding
17. DiffLogo: A comparative visualisation of sequence motifs
18. Weeder (version: 2.0)
19. MCAST: scanning for cis-regulatory motif clusters Part of MEME suite.
20. Sequence-based Discovery of Regulons iRegulon detects the TF, the targets and the motifs/-tracks from a set of genes.
21. Regulatory genomic toolbox
22. Parse TF motifs from public databases, read into R, and scan using 'rtfbs'
23. Romulus: Robust multi-state identification of transcription factor binding sites from DNase-seq data: Romulus is a computational method to accurately identify individual transcription factor binding sites from genome sequence information and cell-type-specific experimental data, such as DNase-seq. It combines the strengths of its predecessors, CENTIPEDE and Wellington, while keeping the number of free parameters in the model robustly low. The method is unique in allowing for multiple binding states for a single transcription factor, differing in their cut profile and overall number of DNase I cuts.
24. moca: Tool for motif conservation analysis.
25. gimmotifs Suite of motif tools, including a motif prediction pipeline for ChIP-seq experiments. looks very useful, will take a look!
26. YAMDA: thousandfold speedup of EM-based motif discovery using deep learning libraries and GPU
27. motif clustering
28. RSAT matrix-clustering: dynamic exploration and redundancy reduction of transcription factor binding motif collections

Super-enhancer identification

The fancy “super-enhancer” term was first introduced by Richard Young in Whitehead Institute. Basically, super-enhancers are enhancers that span large genomic regions (~12.5kb). The concept of super-enhancer is not new. One of the most famous example is the Locus Control Region (LCR) that controls

the globin gene expression, and this has been known for decades.

A review in Nature Genetics What are super-enhancers?

paper: Genetic dissection of the α -globin super-enhancer in vivo

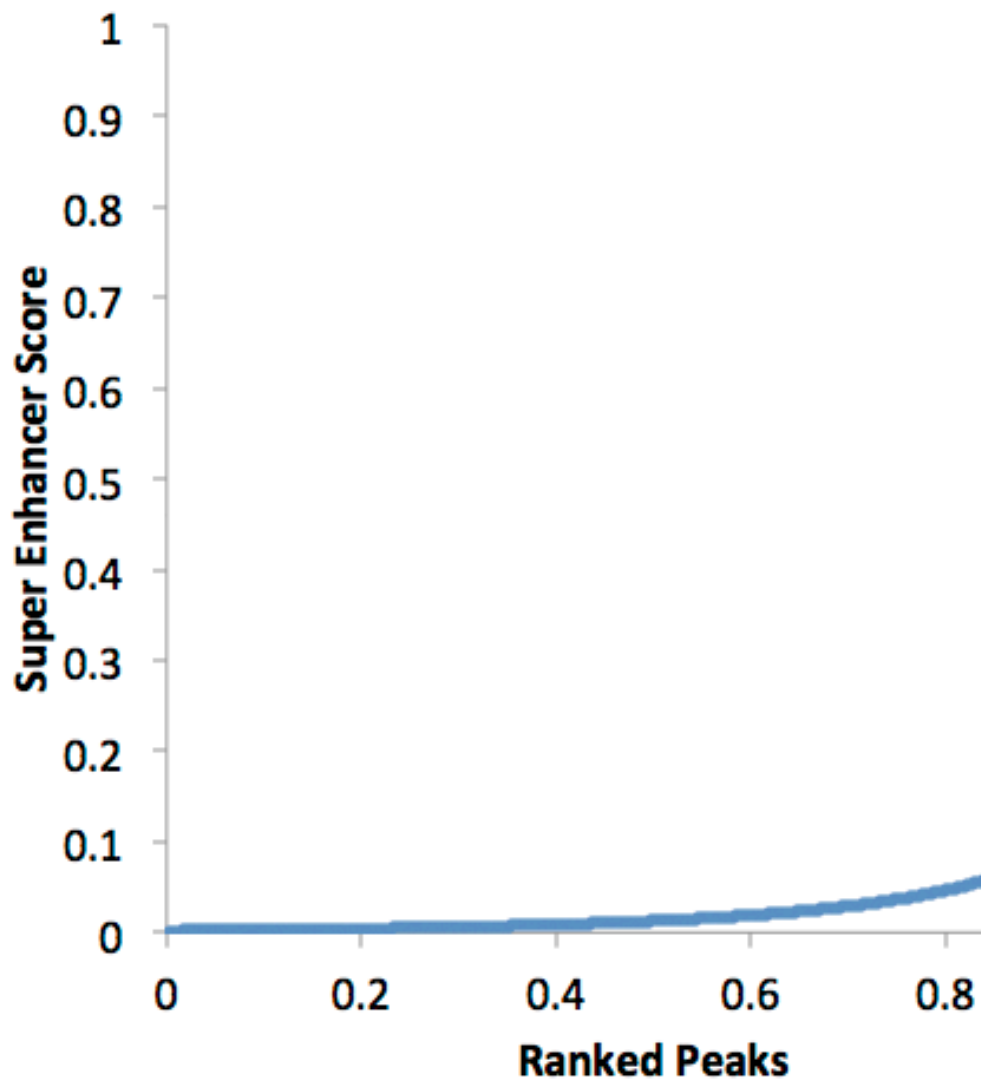
By generating a series of mouse models, deleting each of the five regulatory elements of the α -globin super-enhancer individually and in informative combinations, we demonstrate that each constituent enhancer seems to act independently and in an additive fashion with respect to hematological phenotype, gene expression, chromatin structure and chromosome conformation, without clear evidence of synergistic or higher-order effects.

paper: Hierarchy within the mammary STAT5-driven Wap super-enhancer

paper: Enhancers and super-enhancers have an equivalent regulatory role in embryonic stem cells through regulation of single or multiple genes

From the HOMER page **How finding super enhancers works:**

Super enhancer discovery in HOMER emulates the original strategy used by the Young lab. First, peaks are found just like any other ChIP-Seq data set. Then, peaks found within a given distance are 'stitched' together into larger regions (by default this is set at 12.5 kb). The super enhancer signal of each of these regions is then determined by the total normalized number reads minus the number of normalized reads in the input. These regions are then sorted by their score, normalized to the highest score and the number of putative enhancer regions, and then super enhancers are identified as regions past the point where the slope is greater than 1.



Example of a super enhancer plot:

In the plot above, all of the peaks past 0.95 or so would be considered “super enhancers”, while the one’s below would be “typical” enhancers. If the slope threshold of 1 seems arbitrary to you, well... it is! This part is probably the ‘weakest link’ in the super enhancer definition. However, the concept is still very useful. Please keep in mind that most enhancers probably fall on a continuum between typical and super enhancer status, so don’t bother fighting over the precise number of super enhancers in a given sample and instead look for useful trends in the data.

Using ROSE from Young lab

ROSE: RANK ORDERING OF SUPER-ENHANCERS

**imPROSE - Integrated Methods for Prediction of Super-Enhancers

CREAM (Clustering of Functional Regions Analysis Method) is a new method for identification

of clusters of functional regions (COREs) within chromosomes. published in Genome Research by Mathieu Lupien group. paper: Identifying clusters of cis-regulatory elements underpinning TAD structures and lineage-specific regulatory networks.

Bedgraph, bigwig manipulation tools

WiggleTools

bigwig tool

bigwig-python

samtools

bedtools my all-time favorite tool from Araon Quinlan' lab. Great documentation! pyBedGraph: a Python package for fast operations on 1-dimensional genomic signal tracks. pyBigwig Hosting bigWig for UCSC visualization

My first play with GRO-seq data, from sam to bedgraph for visualization

convert bam file to bigwig file and visualize in UCSC genome browser in a Box (GBiB). megadept is pretty fast, can access bigWig files from the web, works on macOS, Linux & Windows, plus is also available via @Bioconductor <http://www.bioconductor.org/packages/release/bioc/html/megadept.html> which makes easy to use it in #rstats. For example, for quantifying expression of custom regions from recount3 data

Peaks overlapping significance test

The genomic association tester (GAT)

poverlap from Brent Pedersen. Now he is working with Aaron Quinlan at university of Utah.

Genometric Correlation (GenometriCorr): an R package for spatial correlation of genome-wide interval datasets

Location overlap analysis for enrichment of genomic ranges bioconductor package.

regioner Association analysis of genomic regions based on permutation tests similaRpeak: Metrics to estimate a level of similarity between two ChIP-Seq profiles

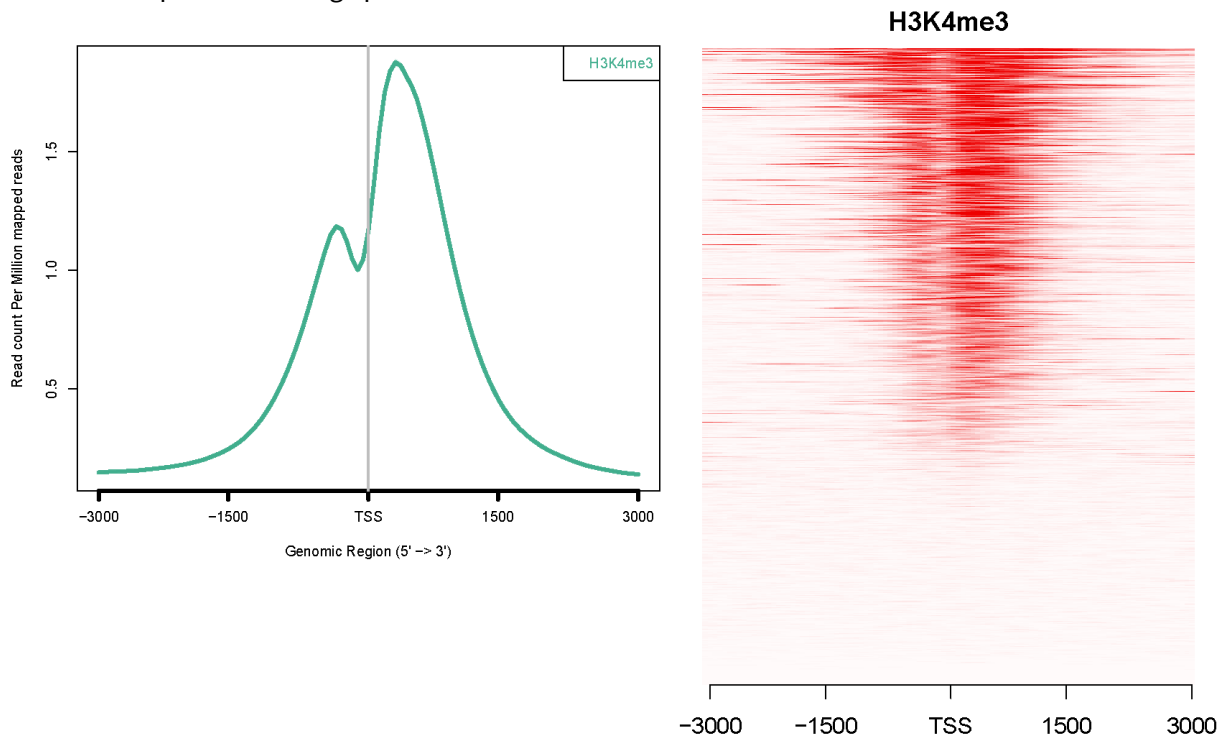
RNA-seq data integration

Beta from Shirley Liu's lab in Harvard. Tao Liu's previous lab.

Heatmap, meta-plot

Many papers draw meta-plot and heatmap on certain genomic regions (2kb around TSS, genebody etc) using ChIP-seq data.

See an example from the ngs.plot:



Tools

1. deeptools. It can do many others and have good documentation. It can also generate the heatmaps, but I personally use ngs.plot which is easy to use. (developed in Mount Sinai).
2. you can also draw heatmaps using R. just count (using either Homer or bedtools) the ChIP-seq reads in each bin and draw with heatmap.2 function. here and here. Those are my pretty old blog posts, I now have a much better idea on how to make those graphs from scratch.
3. You can also use bioconductor Genomation. It is very versatile.
4. ChAsE
5. Metaseq
6. EnrichedHeatmaps from Zuguang Gu based on his own package [ComplexHeatmaps](#). This is now my default go-to because of the flexibility of the package and the great user support. Thx!
7. A biostar post discussing the tools: Visualizations of ChIP-Seq data using Heatmaps

-
8. A bioconductor package to produce metagene plots
 9. Fluff is a Python package that contains several scripts to produce pretty, publication-quality figures for next-generation sequencing experiments I just found it 09/01/2016. looks promising especially for identifying the dynamic change.

One caveat is that the meta-plot (on the left) is an average view of ChIP-seq tag enrichment and may not reflect the real biological meaning for individual cases.

See a post from Lior Pachter How to average genome-wide data

I replied the post: >for ChIP-seq, in addition to the average plot, a heatmap that with each region in each row should make it more clear to compare (although not quantitatively). a box-plot (or a histogram) is better in this case . I am really uncomfortable averaging the signal, as a single value (mean) is not a good description of the distribution.

By Meromit Singer:

>thanks for the paper ref! Indeed, an additional important issue with averaging is that one could be looking at the aggregation of several (possibly very distinct) clusters. Another thing we should all keep in mind if we choose to make such plots..

A paper from Genome Research Ubiquitous heterogeneity and asymmetry of the chromatin environment at regulatory elements

Enhancer databases

- FANTOM project CAGE for promoters and enhancers.
- DENDb: database of integrated human enhancers
- VISTA enhancer browser
- Super-enhancer database
- Genome-wide identification and characterization of HOT regions in the human genome
- EnhancerAtlas: a resource for enhancer annotation and analysis in 105 human cell/tissue types
- review: Computational Tools for Stem Cell Biology
- Integrative analysis of 10,000 epigenomic maps across 800 samples for regulatory genomics and disease dissection from Manolis Kellis group.
- Index and biological spectrum of accessible DNA elements in the human genome DHS sites from John A Stamatoyannopoulos group.

Interesting Enhancer papers

- Multiplex enhancer-reporter assays uncover unsophisticated TP53 enhancer logic
-

Enhancer target prediction

- DoRothEA: collection of human and mouse regulons DoRothEA is a gene regulatory network containing signed transcription factor (TF) - target gene interactions. DoRothEA regulons, the collection of a TF and its transcriptional targets, were curated and collected from different types of evidence for both human and mouse. A confidence level was assigned to each TF-target interaction based on the number of supporting evidence.
- Assessing Computational Methods for Transcription Factor Target Gene Identification Based on ChIP-seq Data
- Protein binding and methylation on looping chromatin accurately predict distal regulatory interactions
- i-cisTarget
- protocol iRegulon and i-cisTarget: Reconstructing Regulatory Networks Using Motif and Track Enrichment
- Model-based Analysis of Regulation of Gene Expression: MARGE from Shirley Liu's lab. MARGE is a robust methodology that leverages a comprehensive library of genome-wide H3K27ac ChIP-seq profiles to predict key regulated genes and cis-regulatory regions in human or mouse.
- PrESSto: Promoter Enhancer Slider Selector Tool
- TargetFinder. paper: Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin
- C3D Cross Cell-type Correlation in DNaseI hypersensitivity. calculates correlations between open regions of chromatin based on DNase I hypersensitivity signals. Regions with high correlations are candidates for 3D interactions. It also performs association tests on each candidate and adjusts p-values.
- ABC Activity-by-Contact model of enhancer specificity from thousands of CRISPR perturbations. Blog post <https://jesseengreitz.wordpress.com/2019/02/10/preprint-activity-by-contact-model/>
- A curated benchmark of enhancer-gene interactions for evaluating enhancer-target gene prediction methods "We use BENGI to test several published computational methods for linking enhancers with genes, including signal correlation and the TargetFinder and PEP supervised learning methods. We find that while TargetFinder is the best-performing method, it is only modestly better than a baseline distance method for most benchmark datasets when trained

and tested with the same cell type and that TargetFinder often does not outperform the distance method when applied across cell types.”

Allele-specific analysis

- WASP: allele-specific software for robust molecular quantitative trait locus discovery
- ABC – (Allele-specific Binding from ChIP-Seq)
- SNPsplit: Allele-specific splitting of alignments between genomes with known SNP genotypes
- BaalChIP: Bayesian analysis of allele-specific transcription factor binding in cancer genomes. bioconductor package, seems to be very useful.

SNPs affect on TF binding

- RegulomeDB Use RegulomeDB to identify DNA features and regulatory elements in non-coding regions of the human genome by entering dbSNP id, chromosome regions or single Nucleotides.
- motifbreakR A Package For Predicting The Disruptiveness Of Single Nucleotide Polymorphisms On Transcription Factor Binding Sites.
- GERV: A Statistical Method for Generative Evaluation of Regulatory Variants for Transcription Factor Binding From the same group as above.
- PRIME: Predicted Regulatory Impact of a Mutation in an Enhancer
- paper: Which Genetics Variants in DNase-Seq Footprints Are More Likely to Alter Binding? website
- paper: Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo
- A Science paper: Survey of variation in human transcription factors reveals prevalent DNA binding changes
- paper Estimating the functional impact of INDELs in transcription factor binding sites: a genome-wide landscape
- paper: Mutational Biases Drive Elevated Rates of Substitution at Regulatory Sites across Cancer Types
- paper: Recurrent promoter mutations in melanoma are defined by an extended context-specific mutational signature
- sasquatch Predicting the impact of regulatory SNPs from cell and tissue specific DNase-footprints

co-occurring TFs

- In-silico Search for co-occurring transcription factors: INSECT
- INSECT 2
- CO-factors associated with Uniquely-bound GENomic Regions: COUGER
-

Conservation of the peak underlying DNA sequences

- bioconductor annotation package phastCons100way.UCSC.hg19 see this post how to use it.
-

Integration of different data sets

methyPipe and compEpiTools: a suite of R packages for the integrative analysis of epigenomics data

Copy number information from targeted sequencing using off-target reads bioconductor CopywriteR package.

3CPET: Finding Co-factor Complexes in Chia-PET experiment using a Hierarchical Dirichlet Process

New single/few cell epigenomics

- GeF-seq: A Simple Procedure for Base Pair Resolution ChIP-seq
- Ultra-low input CUT&RUN (ulicUT&RUN) enables interrogation of TF binding from low cell numbers
- We describe Cleavage Under Targets and Release Using Nuclease (CUT&RUN), a chromatin profiling strategy in which antibody-targeted controlled cleavage by micrococcal nuclease releases specific protein-DNA complexes into the supernatant for paired-end DNA sequencing another cut&run method. maybe useful for scChIP-seq?
- Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state
- Calling Cards enable multiplexed identification of the genomic targets of DNA-binding proteins this is potentially can work with single cells.
- Ultra-parallel ChIP-seq by barcoding of intact nuclei as low as 1000 cells.

-
- single-cell chromatin overall omic-scale landscape sequencing (scCOOL-seq) to generate a genome-wide map of DNA methylation and chromatin accessibility at single-cell resolution
 - High-Throughput ChIPmentation: freely scalable, single day ChIPseq data generation from very low cell-numbers
 - CUT&Tag for efficient epigenomic profiling of small samples and single cells
 - CUT&Tag Data Processing and Analysis Tutorial protocols.io link <https://www.protocols.io/view/cut-amp-tag-data-processing-and-analysis-tutorial-bjk2kkye>
 - Simultaneous quantification of protein-DNA contacts and transcriptomes in single cells scDamID&T.
 - Self-reporting transposons enable simultaneous readout of gene expression and transcription factor binding in single cells piggyBac transposase.
 - Mapping Histone Modifications in Low Cell Number and Single Cells Using Antibody-guided Chromatin Tagmentation (ACT-seq) by Keji Zhao group.
 - Single-cell chromatin immunocleavage sequencing (scChIC-seq) to profile histone modification by Keji Zhao group.
 - CoBATCH for high-throughput single-cell epigenomic profiling Protein A in fusion to Tn5 transposase is enriched through specific antibodies to genomic regions and Tn5 generates indexed chromatin fragments ready for the library preparation and sequencing.
 - High-throughput single-cell ChIP-seq identifies heterogeneity of chromatin states in breast cancer

ChIP-exo

- Characterizing protein-DNA binding event subtypes in ChIP-exo data
- paper: Simplified ChIP-exo assays

ATAC-seq

Some may notice that the peaks produced look both like peaks produced from the TF ChIP-seq pipeline as well as the histone ChIP-seq pipeline. This is intentional, as ATAC-seq data looks both like TF data (narrow peaks of signal) as well as histone data (broader regions of openness).

- paper From reads to insight: a hitchhiker's guide to ATAC-seq data analysis
- ATACseqQC a bioconductor package for quality control of ATAC-seq data.

-
- RASQUAL (Robust Allele Specific QUantification and quality control) maps QTLs for sequenced based cellular traits by combining population and allele-specific signals. paper: Fine-mapping cellular QTLs with RASQUAL and ATAC-seq
 - ATAC-seq Forum
 - Single-cell ATAC-Seq
 - A rapid and robust method for single cell chromatin accessibility profiling
 - Global Prediction of Chromatin Accessibility Using RNA-seq from Small Number of Cells from RNA-seq to DNA accessibility. tool on github
 - NucleoATAC: Python package for calling nucleosomes using ATAC-Seq data
 - chromVAR: Inferring transcription factor variation from single-cell epigenomic data scATAC-seq
 - ENCODE ATAC-seq guidelines
 - Brockman is a suite of command line tools and R functions to convert genomics data into DNA k-mer words representing the regions associated with a chromatin mark, and then analyzing these k-mer sets to see how samples differ from each other. This approach is primarily intended for single cell genomics data, and was tested most extensively on single cell ATAC-seq data
 - Reproducible inference of transcription factor footprints in ATAC-seq and DNase-seq datasets via protocol-specific bias modeling
 - msCentipede is an algorithm for accurately inferring transcription factor binding sites using chromatin accessibility data (Dnase-seq, ATAC-seq) and is written in Python2.x and Cython.
 - The Differential ATAC-seq Toolkit (DASTk) is a set of scripts to aid analyzing differential ATAC-Seq data.
 - Identification of Transcription Factor Binding Sites using ATAC-seq We propose HINT-ATAC, a footprinting method that addresses ATAC- seq specific protocol artifacts
 - HMMRATAC splits a single ATAC-seq dataset into nucleosome-free and nucleosome-enriched signals, learns the unique chromatin structure around accessible regions, and then predicts accessible regions across the entire genome. We show that HMMRATAC outperforms the popular peak-calling algorithms on published human and mouse ATAC-seq datasets.

DNase-seq

- pyDNase - a library for analyzing DNase-seq data. paper: Wellington-bootstrap: differential DNase-seq footprinting identifies cell-type determining transcription factors
- paper: Analysis of computational footprinting methods for DNase sequencing experiments tool
- paper: A practical guide for DNase-seq data analysis: from data management to common applications

-
- Two nature prime: Genome-wide footprinting: ready for prime time? Genomic footprinting
 - PING biocondcutor package: Probabilistic inference for Nucleosome Positioning with MNase-based or Sonicated Short-read Data
 - Basset Convolutional neural network analysis for predicting DNA sequence activity]
 - Analysis of optimized DNase-seq reveals intrinsic bias in transcription factor footprint identification

Chromatin Interaction data (ChIA-PET, Hi-C)

- ChIA-PET2 a versatile and flexible pipeline for analysing different variants of ChIA-PET data
- TopDom : An efficient and Deterministic Method for identifying Topological Domains in Genomes
- DBPnet: Inferring cooperation of DNA binding proteins in 3D genome
- Systematic identification of cooperation between DNA binding proteins in 3D space
- DiffHiC package maintained by Aaron Lun, who is the author of csaw and InteractionSet as well.
- protocol: Practical Analysis of Genome Contact Interaction Experiments
- 4D genome: a general repository for chromatin interaction data
- CCSI: a database providing chromatin–chromatin spatial interaction information. only hg38 for human and mm10 for mouse.
- LOGIQA is a database hosting local and global quality scores assessed over long-range interaction assays (e.g. Hi-C). Based on the concept applied by the NGS-QC Generator over ChIP-seq and related datasets, LOGIQA infers quality indicators by the comparison of multiple sequence reads random sampling assays.
- Computational Identification of Genomic Features That Influence 3D Chromatin Domain Formation
- Feng Yue's lab in PSU developed tools for Hi-C, 4C
- QuIN: A Web Server for Querying and Visualizing Chromatin Interaction Networks
- paper: Three-dimensional disorganization of the cancer genome occurs coincident with long-range genetic and epigenetic alterations
- Exploring long-range genome interactions using the WashU Epigenome Browse
- MAPPING OF LONG-RANGE CHROMATIN INTERACTIONS BY PROXIMITY LIGATION ASSISTED CHIP-SEQ
- HiChIP: Efficient and sensitive analysis of protein-directed genome architecture HiChIP improves the yield of conformation-informative reads by over 10-fold and lowers input requirement over 100-fold relative to ChIA-PE
- A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome paper from Bing Ren's group. 21 tissue-specific TADs.

Caleb's take on HiChIP analysis

From Caleb, the author of hichipper <https://twitter.com/CalebLareau/status/1098312702651523077> thx!

In HiChIP data analyses, there are two primary problems that we are trying to solve. A) Which anchors (i.e. genomic loci) should be used as a feature set and B) which loops (i.e. interactions between pairs of loci) are important in the data. 2/n

Depending on what you are hoping to use your data for, there are a variety of ways to think about anchors and loops. Two uses of HiChIP that come to mind are “which gene is this enhancer talking to” and “which loops are differential between my celltype/condition of interest” 3/n

When Martin and I wrote hichipper, we envisioned the second question being more used (i.e. building out a framework for differential loop calling), so we wanted a pre-processing pipeline that was as inclusive of potential loops as possible that could be subsetted downstream 4/n

To these ends, we reported an improved version of anchor detection from HiChIP data by modeling the restriction enzyme cut bias explicitly, which helped identify high-quality anchors from the data itself 5/n

(we achieve this by re-parametrizing MACS2 peak calling by essentially fitting a loess curve to the data in the previous picture) 6/n

Unfortunately, based on user feedback, this modified background winds up with a very, very conservative peak calling if the library preparations are sub-par. Thus, the safest way to approach HiChIP data analyses is often to use a pre-defined anchor set 7/n

These can be from either a complementary ATAC-seq or ChIP-seq dataset for the conditions that you are interested in. From what I've seen, you can supply a bed file to hichipper or other tools directly. Hichipper does some other modifications by default to this bed file FYI 8/n

In terms of the second problem of identifying loops, hichipper didn't make any revolutionary progress. We recommend some level of CPM-based filtering + mango FDR calculation (implemented in hichipper) for identifying single-library significant loops. 9/n

Where I've personally done the most is getting multiple libraries from multiple conditions and using some sort of between-replicate logic to filter to a reasonable (~10,000-20,000) number of loops (see e.g. <https://github.com/caleblareau/k562-hichip> ...) 10/n

Other tools (that I admittedly have not tried) use a variety of statistical techniques to (probably more intelligently from what I can tell) merge anchors or filter loops for analyses. A brief run down of those that I'm aware of (not exhaustive)- 11/n

MAPS (<https://www.biorxiv.org/content/biorxiv/early/2018/09/08/411835.full.pdf> ...) uses a measure

of reproducibility with ChIP-seq to define a normalization and significance basis for loop calling. Given HiChIP-specific restriction enzyme bias, this seems sensible 12/n

FitHiChIP (<https://www.biorxiv.org/content/early/2018/10/29/376194.full.pdf> ...) provides automatic merging of nearby anchors to solve the “hairball” problem, which is clearly shown Fig. 1. When I compared hichipper to FitHiC, the bias regression seemed to perform well, but I ran into memory issues which high... 13/n

resolution (i.e. ~2.5kb) HiChIP data, which the authors have apparently solved in FitHiChIP. 14/n

Additionally, there is CID, which uses a density-based method to further collapse anchors to solve the “hairball” problem. 15/n

There are certainly other tools out there, but from my experience, any of these four (hichipper, MAPS, FitHiChIP, and CID) will probably give you something sensible (again acknowledging that I myself haven’t actually run these other 3 tools) 16/n

And if you’re still reading this, I’ll be a bit more specific about how I view hichipper pros/cons from both my own use and others in the community: hichipper provides the most “vanilla” functionality to given sensible yet exhaustive anchors and loops. 17/n

I prefer it this way because I find that for each data set, I have to apply variable downstream threshold and cutoffs because the assay is so variable depending on which experimentalist performs the protocol and the biological question often varies so much 18/n

This may be a negative for individuals new to bioinformatics or HiChIP data but seemingly a positive for someone more experienced in working with related data. It’s not obvious to me which other tools may be more applicable to a novice 19/n

Hope this helps paint a picture– do let me know what you find if you compare tools! I think that it would be useful for the community. 20/20