EMD MILLIPORE

PEAK CALLING FOR ChIP-SEQ

ChIP-SEQ AND PEAK CALLING

Chromatin-immunoprecipitation sequencing (ChIP-seq) is the most widely used technique for analyzing Protein:DNA interactions. Very briefly, cells are cross-linked, fragmented and immunoprecipitated with an antibody specific to the target protein; the resulting ChIP DNA fragments are used as the template in a next-generation sequencing library preparation and many millions of short sequence reads are generated for analysis. The computational analysis is heavily dependent on the detection of "peaks", regions of the genome where multiple reads align that are indicative of protein binding. This article focuses on peak calling, presents the major tools used today and lists additional tools for ChIP Seq.

IN THIS GUIDE WE WILL REVIEW THE FOLLOWING TOPICS:

- Introduction of ChIP-seq and peak calling
- What do peak callers do?
- Choosing a peak calling algorithm
- How do peak callers compare

WHAT DO PEAK CALLERS DO?

Peak calling programs help to define sites of Protein:DNA binding by identifying regions where sequence reads are enriched in the genome after mapping. The common assumption is that the ChIP-seq process is relatively unbiased so reads should accumulate at sites of protein binding faster than in background regions of the genome. The millions of sequencing reads generated in a ChIP-seq experiment are first aligned to a reference genome using tools such as BWA (1) and Bowtie (2). The choice of alignment algorithm and the parameters used can impact peak calling. The number of mismatches allowed can affect the percentage of sequences that can be successfully aligned and the use and placement of reads that map to multiple locations e.g. in repeat regions, can mask true binding events. It is important to understand if and how the aligner and peak caller will work together.

Peak calling requires several distinct analyses be carried out to generate the final peak list; read shifting, background estimation, identification of enriched peaks, significance analysis and removal of artifacts. A 2009 review by Pepke et al details each of these steps and discusses how peak finding tools approach the separate steps very differently (3). A follow up review by Wilbanks et al evaluated the performance of 11 ChIP-seq peak callers nearly all of which are still widely used today (4). Each step can have parameters that can be adjusted by the user, but changing these can significantly affect the final peak lists. Care must be taken that data sets are analysed using the same methods. The ENCODE consortium produced guidelines for analysis of the dispersed data sets to avoid issues created by analysis parameter differences (5). This project used MACS, PeakSeq and SPP.

Read shifting: The aligned reads are from fragments of 150-300bp in length and, as most ChIP-seq data is from single-end sequencing, only one end of a fragment is read. Reads therefore align to either the sense/antisense strands and the 3' or 5' extremes of the DNA fragments pulled down in the immunoprecipitation. The reads are shifted and the data from both strands combined to determine the most likely bases involved in protein binding. How much to "shift" is determined by the fragment size generated in the ChIP-seq library preparation, this can be determined empirically or estimated from the sequence data. Comparison of these two measurements can be an effective quality control, as can the ratio of reads from different strands, where one would expect the ratio to be close to 1.

Background estimation: Control ChIPs are processed in the same way to allow either a genomic background to be determined (input controls) or for regions enriched through the ChIP process with no antibody specificity to be identified (IgG controls). Some peak callers work without control data and assume an even background signal, others make use of blacklist tools, that mask regions of the genome e.g. RepeatMasker and the "Duke excluded regions" list that was developed for the ENCODE project.

Peak identification: A peak is called where either the number of reads exceeds a pre-determined threshold value or where there is a minimum enrichment compared to background signal, often in a sliding window across the genome. Some tools apply both methods. The parameters for identifying peaks can be adjusted, sometimes leading to very different numbers of peaks being called. The user must determine if fewer high-quality peaks are preferred over lower quality peaks.

Significance analysis: Many peak callers compute a P value for called peaks, others use the height of the peaks and/or enrichment over background to rank peaks, but these do not provide statistical significance values. The false discovery rate (FDR) is often used to provide a truer peak list and this can be computed from the P values provided. Some packages make use of the control data to determine an empirical FDR and generate a ratio of peaks in the sample vs. control.

Artifact removal: Two major classes of ChIP-seq artifacts are generally removed before the final peak list is used in downstream applications. First, peaks containing either a single read, or just a few reads are assumed to be PCR amplification artifacts and discarded. Then, peaks where there is a significant imbalance between the numbers of reads on each strand are removed. This second filtering is more difficult in complex regions where binding may be occurring at multiple co-located sties.

Unfortunately ChIP-seq does have biases, but these are gradually being understood. In experiments where deproteinized, sheared, non-cross-linked DNA was used as the template for ChIP-seq studies, it was possible to identify some of the factors affecting background noise. The authors of this study also developed a model-based approach called MOSAiCS to find peaks more reliably, although this has not yet been widely adopted.

CHOOSING A PEAK CALLING ALGORITHM

There does not appear to be a clear winner among the thirty or more peak calling algorithms available today. Ask ten bioinformaticians which is best and you will likely get ten different answers. The answer very much depends on the type of experiment being considered, some peak callers are better for transcription factors whilst others produce more reliable data for long range interactions such as polymerase binding, e.g. MACS vs SICER. However much depends on user experience. Many peak callers have multiple parameters that can affect the number of peaks called, understanding these parameters takes time and once comfortable with a particular setup many are unlikely to change. This is perhaps one of the reasons MACs is still dominant. Although one of the oldest peak callers it compares well to newer tools and many people have experience with it.

HOW DO PEAK CALLERS COMPARE

Papers that compare the various peak calling algorithms are typically out of date the moment they are written, let alone published but they point out important areas for consideration. The comparison methods used in different papers could be usefully updated and presented in a non-static electronic format. The winners, as far as the number of citations the primary publication received to-date are; E-Range (from the Wold group at Caltech), ChIP-seq peak finder (from the Genome Institute of Singapore) and MACS (from the Liu lab at Dana Faber), with over 4000 citations between them. However these are also three of the oldest packages released in the early days of ChIP-seq analysis.

At least one group has tried to produce benchmark datasets that can be used for comparison of peak callers (6). One of their aims was to provide datasets that were independent of those used to develop analysis tools, making an unbiased comparison easier. Their analysis of five programs showed that control data was essential for reduction of false-positive peaks, but that even without this a manual visual inspection allowed 80% of false-positives to be removed suggesting that the shape of the peaks could be used to improve analysis methods. They suggested a meta-approach that used features from four of the programs tested which gave improved results for the benchmark dataset. Other groups have also suggested a multi-tool approach using several peak callers to generate consensus peak lists.

ChIP-SEQ PEAK CALLERS

Rather than giving a detailed description of all peak-finding packages, here we have picked four: MACS, which is one of the most popular tools, and three others that offer something different over the majority of programs. A more comprehensive list of current packages can be found at the end of this guide.

MACS: MACS is (for TF peaks) one of the most popular peak callers, it is also one of the oldest and this probably contributes to its success. It is a good method, good enough for many experimental conditions and requires very little justification if cited as the tool used in a publication. MACS performs removal of redundant reads, read-shifting to account for the offset in forward or reverse strand reads. It uses control samples and local statistics to minimize bias and calculates an empirical FDR.

SCICER: Not all ChIP-seq users are interested in the "peaky" data as seen with transcription factors. However nearly all peak callers were developed for exactly this kind of data. SCICER was developed for more diffuse chromatin modifications that can span kilobases or megabases of the genome. Their method scans the genome in windows and identifies clusters of spatial signals that are unlikely to appear by chance. These clusters or "islands" are used rather than fixed length windows, gaps in the islands are allowed to overcome technical issues (under-saturated experiments, repeat regions, etc). And this gap size can be adjusted for different types of chromatin modification. The program makes use of control data or a random background model (7).

T-PIC: This package uses the shape of putative peaks to identify true peaks from the background noise. They compared their approach to MACS and PeakSeq and demonstrated improved results. The package first extends short reads to the estimated fragment length, it then divides the genome into regions for which it constructs "trees" for shape analysis and uses the tree shape statistic to identify true peaks (8).

Genome wide event finding and motif discovery (GEM): This is one of the latest tools published in mid-2012. Its unique feature is the combination of peak finding and motif analysis to improve the resolution of the final peaks called. The paper presents an analysis of 63 transcription factors in 214 ENCODE experiments and improves the spatial resolution and motif discovery when compared to previous tools. The tool also allows discovery of spatially-constrained binding events which was demonstrated using the well understood Sox2-Oct4 transcription factor complex. This paper presents almost 400 spatially-constrained transcription factor binding events. This tool appears to be an exciting development for ChIP-seq studies.

SUMMARY

The abundance of algorithms for peak calling is a testament to the evolving and diverse needs in the research community. Researchers are using ChIP-seq in a diverse range of biological and technical scenarios. Although, there probably won't be one perfect solution, we hope that this introduction to this dynamic area of bioinformatics provides a useful starting point for your ChIP-seq data analysis.

THE REST OF THE ChIP-SEQ PEAK CALLERS

- AREM
- BayesPeak
- CEAS
- ChIP-Peak
- CisGenome
- CSDeconv
- E-RANGE: E-RANGE is a dual-use package for RNA-seq and ChIP-seq, it is based on the ChIPSeq mini peak finder published by the Wold group in 2007.
- EpiChip
- F-Seq
- FindPeaks: Is part of the Vancouver Short Read Analysis Package.
- HPeak
- MOSAiCS
- PeakSeq: Corrects for mappability and GC content biases to generate more accurate peak calls
- QuEST
- SIPeS: Uses paired-end data.
- SISSRS: Is a directional tool that iidentifies where reads "strand-shift" and can generate precise calls for sharp peaks. It is not so useful if you are interested in broader ChIP signals.
- Sole-Search
- SPP: Accounts for the read offset and read-shifts to improve results. The package makes use of background or control data, and estimate read saturation allowing the user to determine if more reads are required or not.
- SWEMBL
- Useq

Useful ChIP Seq Antibodies and Reagents

Product Name	Catalog Number
Kits	
Magna ChIP-Seq™ Chromatin Immunoprecipitation and Next Generation Sequencing Library Preparation Kit	17-1010
Magna ChIP™ A/G Chromatin Immunoprecipitation Kit	17-10085
EZ Magna ChIP A/G Chromatin Immunoprecipitation Kit	17-10086
Magna ChIP™ Protein A+G Magnetic Beads	16-663
Magna ChIP™ Protein A Magnetic Beads	16-661
Magna ChIP™ Protein G Magnetic Beads	16-662
ChIP-seq Qualified Antibodies	
ChIPAb+ Trimethyl-Histone H3 (Lys9)	17-625
ChIPAb+ Trimethyl-Histone H3 (Lys4)	17-614
Anti-acetyl-Histone H4 (Lys12) Antibody	07-595
Anti-trimethyl-Histone H3 (Lys4) Antibody	07-473
Anti-acetyl-Histone H4 (Lys8) Antibody	07-328
Anti-trimethyl-Histone H3 (Lys4) Antibody, clone MC315	04-745
Anti-phospho-Histone H3 (Ser10) Antibody, clone MC463	04-817
Anti-acetyl-Histone H3 (Lys14) Antibody	07-353
Anti-dimethyl-Histone H3 (Lys4) Antibody	07-030
Anti-acetyl-Histone H4 (Lys16) Antibody	07-329
Anti-phospho (Ser10)-acetyl (Lys14)-Histone H3 Antibody	07-081
Anti-acetyl-Histone H3 Antibody	06-599
ChIPAb+ EZH2, clone AC22	17-662
Anti-acetyl-Histone H3 (Lys 4) Antibody	07-539
Anti-trimethyl-Histone H3 (Lys9) Antibody, clone 6F12-H4	05-1242
Anti-Myc Tag Antibody, clone 4A6	05-724
Anti-Histone H4 Antibody, pan, clone 62-141-13	05-858
Anti-acetyl-Histone H3 (Lys14) Antibody, clone 13HH3-1A5	MABE351
Anti-CTCF Antibody	07-729
Anti-acetyl-Histone H4 (Lys5) Antibody	07-327
Anti-RNA polymerase II Antibody, clone CTD4H8	05-623
Anti-monomethyl-Histone H3 (Lys27) Antibody	07-448
Anti-E2F-4 Antibody, clone GG22-2A6	05-312
Anti-trimethyl-Histone H3 (Lys27) Antibody	07-449
Anti-dimethyl-Histone H3 (Lys27) Antibody	07-452
Anti-EZH2 Antibody	07-689
Anti-Androgen Receptor Antibody, PG-21	06-680
Anti-Methylcytosine dioxygenase TET1 Antibody	09-872
Anti-phospho-H2A.X (Ser139) Antibody	07-164
Anti-trimethyl-Histone H3 (Lys9) Antibody	07-442
Anti-monomethyl-Histone H3 (Lys4) Antibody	07-436

References

1. Li H. and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics, 25:1754-60.
2. (B.Langmead, C.Trapnell & S.L.Salzberg. Ultrafast and memory- efficient alignment of short DNA sequences to the human genome. Genome 2009;10(3):R25).
3. S.Pepke, B.Wold & A.Mortazavi, Computation for ChIP-seq and RNA-seq studies. Nature Methods: 2009 Nov;6(11 Suppl):S22-32)
4. (E.G.Wilbanks, M.T.Facciotti, Evaluation of Algorithm Performance in ChIP-Seq Peak Detection. PLoS One: 2010 Jul;5,7)
5. (S.G.Landt, et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res. 2012 Sep;22(9):1813-31)
6. Morten Beck Rye, Pal Sætrom and Finn Drabløs. A manually curated ChIP-seq benchmark demonstrates room for improvement in current peak-finder programs Nucleic Acids Research 2011
7. (C.Zang, D.E.Schones, C.Zeng, K.Cui, K.Zhao & W.Peng. A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. Bioinformatics (2009) 25 (15): 1952-1958).
8. (V.Hower, S.N.Evans & L.Pachter, Shape-based peak identification for ChIP-Seq. BMC Bioinformatics: 2011, 12:15)

www.millipore.com/epigenetics

This item is intended for research use only. Not for use in diagnostic procedures. Information, descriptions, and specifications in this publication are subject to change without notice.