# Chipseq - peaks

02/05/2016

# ENCODE

Projects >

## (2014) ENCODE: Histone modification ChIP-seq uniform peak calls

**Defining functional DNA elements in the human genome.**
Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, Ward LD, Birney E, Crawford GE, Dekker J, Dunham I, Elnitski LL, Farnham PJ, Feingold EA, Gerstein M, Giddings MC, Gilbert DM, Gingeras TR, Green ED, Guigo R, Hubbard T, Kent J, Lieb JD, Myers RM, Pazin MJ, Ren B, Stamatoyannopoulos JA, Weng Z, White KP, Hardison RC.
Proc Natl Acad Sci U S A. 2014 Apr 23.

Histone modification ChIP-seq datasets were processed to identify regions of ChIP enrichment relative to corresponding sequenced input-DNA controls. Read alignment files were filtered to discard multi-mapping reads and duplicates.

We used the MACS2 peak caller (v 2.0.10.20130712) to identify regions of enrichment over a wide range of signal strength. Enriched regions were scored on individual replicates, pooled data (reads pooled across replicates) and on subsampled pseudoreplicates (obtained by pooling reads from all replicates and randomly subsampling, without replacement, two pseudoreplicates with half the total number of pooled reads).

We used MACS2 to identify three types of regions of enrichment: (i) narrow peaks of contiguous enrichment (narrowPeaks) that pass a Poisson *p*-value threshold of 0.01; (ii) broader regions of enrichment (**broadPeaks**) that pass a Poisson *p*-value threshold of 0.1 (using MACS2's broad peak mode); (iii) gapped/chained regions of enrichment (**gappedPeaks**) defined as broadPeaks that contain atleast one strong narrowPeak.

In order to obtain reliable regions of enrichment, we restricted to enriched regions identified using pooled data that were also independently identified in both pseudoreplicates. The coverage and conservation analysis only used histone modification datasets from the Broad Institute Production group. We used the gappedPeak representation for the histone marks with relatively compact enrichment patterns. These include H3K4me3, H3K4me2, H3K4me1, H3K9ac, H3K27ac and H2A.Z.

For the diffused histone marks, H3K36me3, H3K79me2, H3K27me3, H3K9me3 and H3K9me1, we used the broadPeak representation. These peak calls were not optimally thresholded by design so as to allow for analysis of genomic coverage over a wide range of signal enrichment.

The gappedPeak and broadPeak files can be downloaded from http://www.broadinstitute.org/~anshul/projects/encode/rawdata/peaks_histone/mar2012/broad/combrep_and_ppr/
The narrowPeak files (not used in any of the analyses) can be downloaded from
http://www.broadinstitute.org/~anshul/projects/encode/rawdata/peaks_histone/mar2012/narrow/combrep_and_ppr/
Negative $\log_{10}$ of Poisson *p*-values of enrichment present in Column 8 of the peak files were used as scores for the peaks in the coverage analysis.

# ENCODE

- We used MACS2 to identify three types of regions of enrichment:
    1. narrow peaks of contiguous enrichment (narrowPeaks) that pass a Poisson P value threshold of 0.01;
    2. broader regions of enrichment (broadPeaks) that pass a Poisson P value threshold of 0.1 (using MACS2's broad peak mode);
    3. gapped/chained regions of enrichment (gappedPeaks) defined as broadPeaks that contain at least one strong narrowPeak.
- We used the gappedPeak representation for the histone marks with relativelycompact enrichment patterns. These include H3K4me3, H3K4me2, H3K4me1, H3K9ac, H3K27ac, and H2A.Z
- For the diffused histone marks, H3K36me3, H3K79me2, H3K27me3, H3K9me3, and H3K9me1, we used the broadPeak representation
- The narrowPeak files were not used in any of the analyses
- The negative log10 of Poisson P values of enrichment present in column 8 of the peak files was used as scores for the peaks in the coverage analysis.

# MACS2

2. NAME_peaks.narrowPeak is BED6+4 format file which contains the peak locations together with peak summit, pvalue and qvalue. You can load it to UCSC genome browser. Definition of some specific columns are:

    - 5th: integer score for display
    - 7th: fold-change
    - 8th: -log10pvalue
    - 9th: -log10qvalue
    - 10th: relative summit position to peak start

    The file can be loaded directly to UCSC genome browser. Remove the beginning track line if you want to analyze it by other tools.

4. NAME_peaks.broadPeak is in BED6+3 format which is similar to narrowPeak file, except for missing the 10th column for annotating peak summits.

5. NAME_peaks.gappedPeak is in BED12+3 format which contains both the broad region and narrow peaks. The 5th column is 10*-log10qvalue, to be more compatible to show grey levels on UCSC browser. Tht 7th is the start of the first narrow peak in the region, and the 8th column is the end. The 9th column should be RGB color key, however, we keep 0 here to use the default color, so change it if you want. The 10th column tells how many blocks including the starting 1bp and ending 1bp of broad regions. The 11th column shows the length of each blocks, and 12th for the starts of each blocks. 13th: fold-change, 14th: -log10pvalue, 15th: -log10qvalue. The file can be loaded directly to UCSC genome browser.

# Number of narrow peaks

| labExpId | old-input | new-input | mark |
|---|---|---|---|
| H018H3K27acX1 | 54083 | 159435 | H3K27ac |
| H000H3K27me3X1 | 56614 | 200581 | H3K27me3 |
| H003H3K27me3X1 | 47276 | 46199 | H3K27me3 |
| H003H3K27me3X2 | 77913 | 87551 | H3K27me3 |
| H006H3K27me3X1 | 84237 | 98512 | H3K27me3 |
| H009H3K27me3X1 | 84671 | 96076 | H3K27me3 |
| H012H3K27me3X1 | 58202 | 63639 | H3K27me3 |
| H018H3K27me3X1 | 67815 | 102694 | H3K27me3 |
| H024H3K27me3X1 | 71657 | 84988 | H3K27me3 |
| H036H3K27me3X1 | 84923 | 103310 | H3K27me3 |
| H048H3K27me3X1 | 77351 | 100562 | H3K27me3 |
| H072H3K27me3X1 | 75238 | 88966 | H3K27me3 |
| H120H3K27me3X1 | 64592 | 121408 | H3K27me3 |
| H168H3K27me3X1 | 79019 | 131513 | H3K27me3 |
| H168H3K36m3X1 | 151225 | 286487 | H3K36m3 |
| H000H3K4me1X1 | 153698 | 222949 | H3K4me1 |
| H003H3K4me1X1 | 124037 | 160196 | H3K4me1 |
| H006H3K4me1X1 | 158751 | 220182 | H3K4me1 |
| H012H3K4me1X1 | 153126 | 222981 | H3K4me1 |
| H024H3K4me1X1 | 144159 | 195934 | H3K4me1 |

# Number of narrow peaks

| labExpId | old-input | new-input | mark |
|---|---|---|---|
| H048H3K4me2X1 | 139299 | 197425 | H3K4me2 |
| H072H3K4me2X1 | 116145 | 163088 | H3K4me2 |
| H120H3K4me2X1 | 132674 | 177044 | H3K4me2 |
| H168H3K4me2X1 | 187581 | 311084 | H3K4me2 |
| H000H3K4me3X1 | 23006 | 50113 | H3K4me3 |
| H003H3K4me3X1 | 19569 | 30460 | H3K4me3 |
| H006H3K4me3X1 | 24876 | 53460 | H3K4me3 |
| H009H3K4me3X1 | 25244 | 66788 | H3K4me3 |
| H012H3K4me3X1 | 23201 | 63965 | H3K4me3 |
| H018H3K4me3X1 | 21647 | 57397 | H3K4me3 |
| H024H3K4me3X1 | 22832 | 56318 | H3K4me3 |
| H036H3K4me3X1 | 27996 | 64668 | H3K4me3 |
| H048H3K4me3X1 | 27572 | 62818 | H3K4me3 |
| H072H3K4me3X1 | 24718 | 52390 | H3K4me3 |
| H120H3K4me3X1 | 24326 | 45975 | H3K4me3 |
| H168H3K4me3X1 | 25320 | 52604 | H3K4me3 |
| H018H3K9acX1 | 25902 | 72364 | H3K9ac |
| H000H3K9me3X1 | 96137 | 240741 | H3K9me3 |
| H003H3K9me3X1 | 94989 | 193950 | H3K9me3 |
| H018H3K9me3X1 | 79027 | 188173 | H3K9me3 |
| H024H3K9me3X1 | 104084 | 178029 | H3K9me3 |
| H120H3K9me3X1 | 127583 | 249261 | H3K9me3 |