# BaseSpace ChIP-Seq output conversion to bioset

We currently do not support the curation of public ChIP-Seq analyses as a service for Correlation Engine. This document outlines the general process by which we would transform the output of the BaseSpace ChIP-Seq application into a format suitable for upload as a bioset. There are two data files generated by the BaseSpace Sequence Hub ChIP-Seq application that can be used for bioset creation for upload into Correlation Engine.



If the experiment is characterizing the genome occupancy by modified histones, the data ChIP-Seq\_peaks.xls file is used. If the study was of transcription factor binding or RNA polymerase binding, the data from the ChIP-Seq\_peaks.narrowPeak file is as well as settings information provided in the ChIP-Seq\_peaks.xls file.

## Histone binding

The ChIP-Seq\_peaks.xls contains a header area offset with hashes (#) with settings values and other metadata around processing and data section with the following column headers:

|  |  |
| --- | --- |
| Original header | Renamed header |
| chr | chr |
| start | Start |
| end | End |
| length |  |
| abs\_summit |  |
| pileup |  |
| =-LOG10(pvalue) | p-value |
| fold\_enrichment | Enrichment fold |
| =-LOG10(qvalue) | q-value |
| name |  |

The renamed header column indicates which data should be used and how the header names should appear in the bioset before upload. The values in the p-value and q-value columns should be multiplied by -1 and unlogged to generate the original p- and q-values.



The data columns should be pulled out, transformed appropriately and added to tab-delimited text file with a header describing the settings used in processing. For example:

|  |
| --- |
| Bioset summary = Mammary epithelial cell genome occupancy by MYC ChIP-Seq |
|  |  |  |  |  |  |  |
| Bioset description = Mammary Epithelial cells expressing HA-tagged MYC + MYC ChIP \_vs\_ input ChIP-Seq |
|  |  |  |  |  |  |  |
| Organism = Homo sapiens |  |  |  |  |
|  |  |  |  |  |  |  |
| Data pre-processing = FASTQ files were downloaded from Sequence Read Archive (SRA). No other preprocessing was performed. |
|  |  |  |  |  |  |  |
| Analysis summary = Identification of chromosome regions significantly enriched by chromatin immunoprecipitation (ChIP). Sequence reads from each IP experiment are analyzed using the BaseSpace Sequence Hub ChIPSeq application version 1.0.2 https://basespace.illumina.com/apps/2561559/ChIPSeq?preferredversion. Sequence reads are aligned to the Homo sapiens reference genome (NCBI37/hg19) using BWA and processed using MACS2 2.1.0.. Parameter settings used = [Fragment size = 200, Q-value cutoff = 0.01, Fold enrichment = 10,30, Sample normalization = Down Sample]. Regions, defined by the chromosome start and end positions, are ranked by fold change as determined by comparison to input DNA or IgG ChIP, as provided. For transcription factor binding studies, narrow peak calls are reported. For histone modifications, broad peak calls are reported. |
|  |  |  |  |  |  |  |
| Test group = Immortalized\_mammary\_epithelial\_cell\_DNA\_expressing\_HA-tagged\_MYC\_+\_MYC\_chIP |
| (total replicates = 1) |  |  |  |  |  |
|  |  |  |  |  |  |  |
| Control group = Immortalized\_mammary\_epithelial\_cell\_DNA\_expressing\_HA-tagged\_MYC\_-\_input (total replicates = 1) |
|  |  |  |  |  |  |  |
| Platform(s) = Illumina iGenome UCSC, hg19, March 6, 2013 RefSeq |
|  |  |  |  |  |  |  |
| Internal ID - GSE66252\_1 |  |  |  |  |
|  |  |  |  |  |  |  |
| chr | start | end | Enrichment fold | q-value | p-value |  |
| chr7 | 4681591 | 4681891 | 34.34069 | 4.92E-97 | 3.97E-105 |  |
| chr17 | 61851121 | 61851429 | 28.33336 | 2.10E-96 | 1.90E-104 |  |
| chr17 | 75084699 | 75084869 | 26.33382 | 2.71E-85 | 7.89E-93 |  |
| chr1 | 1.59E+08 | 1.59E+08 | 21.00002 | 4.24E-81 | 1.86E-88 |  |
| chr2 | 2.17E+08 | 2.17E+08 | 36.57338 | 5.76E-81 | 2.63E-88 |  |

Q-value cut-offs are determined in the processing set-up, usually with a default of q-value = 0.01. No additional cut-offs need be applied before upload.

## Transcription factor binding

Transcription factor protein-DNA binding experiments should report calls provided in the ChIP-Seq\_peaks.narrowPeak file, which looks like the following:



Note the lack of column headers or any processing related information. The content of the columns is as follows:

1. Chromosome number
2. Start
3. End
4. Name
5. Integer score for display
6. Strand
7. Fold change
8. –log10pvalue
9. –log10qvalue
10. Relative summit position to peak start

Columns 1, 2, 3, 7, 8, and 9 should be brought together in a new table with the following headers applied:

1. chr
2. start
3. end
4. Enrichment fold
5. p-value
6. q-value

The p-values and q-values should be multiplied by -1 and unlogged.

Header information should be added as for histone biosets, drawing the settings information from the ChIP-Seq\_peaks.xls file. The genome build can be read from sample level statistics files if needed, located at <ChIP-Seq>/Statistics/<SRR#######\_S#.summary.csv>.

Correlation Engine annotates genomes upon upload using one particular genome build per supported species. Lift-over tools are available that allow the user to convert a subset of other builds to the Correlation Engine standard. The following table outlines what lift-over tools are available and the target build.

|  |  |  |
| --- | --- | --- |
| Species | Build | Lift-over to |
| Human |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|

|  |
| --- |
| hg19, Genome Reference Consortium GRCh37 |
| hg18, NCBI Build 36.1 |
| hg17, NCBI Build 35 |
| hg16, NCBI Build 34 |
| hg15, NCBI Build 33 |

 |

 | Hg18 |
| Mouse |

|  |
| --- |
| mm9, NCBI Build 37 |
| mm8, NCBI Build 36 |
| mm7, NCBI Build 35 |
| mm6, NCBI Build 34 |
| mm5, NCBI Build 33 |

 | Mm9 |
| Rat | Rn4, NCBI RGSC v3.4 | Rn4 |
| Worm |

|  |
| --- |
| ce6, NCBI WS190 |
| ce4, NCBI WS170 |
| ce2, NCBI WS120 |

 | Ce6 |
| Fly |

|  |
| --- |
| dm3, NCBI v5.2 |
| dm2, NCBI v4 |
| dm1 NCBI v3 |

 | Dm3 |
| Saccharomyces cerevisiae |

|  |
| --- |
| sacCer2 |
| sacCer1 |

 | sacCer2 |
| Monkey | rheMac2 | rheMac2 |
| Chicken | galGal3 | galGal3 |
| Chimp | panTro2 | panTro2 |
| Cow | bosTau4 | bosTau4 |
| Fish | danRer6 | danRer6 |
| Dog |

|  |
| --- |
| canFam2 |

 | canFam2 |

Upon upload vie the Correlation Engine user interface, select the appropriate data type and species and whether the experiment type was such that multiple samples were grouped together as aggregates for analysis, or whether just single samples were used, i.e., 1 IP and 1 input sample, for the comparison.

After choosing the biosets files and all the required information is recognized in the next page, select the Enrichment fold column for ranking, and Rank, descending for how the bioset should be ranked. Lastly on this page select the original genome build that was used for generating the data. For example, if hg19 was used for alignments, select hg19 to lift-over coordinates from that build to hg18. The user has an opportunity to examine any unrecognized rows or rows dropped by the build conversion.

The last step before commencing upload is to apply tags to the biosets. The standard types of tags are BIOSOURCE, TISSUE, the GENE that is the antibody target, the GENEMODE “ChIP antibody target” and an appropriate biodesign of Normal vs. normal, Disease vs. disease, Treatment vs. treatment. We do not normally ingest protein-DNA studies of designs other than ChIP \_vs\_ input or ChIP \_vs\_ Ig.