Overview of the supplementary website of the Roadmap Epigenomics project

Wouter Meuleman
http://compbio.mit.edu/roadmap
• Website with supplementary information for the Roadmap Epigenomics Project

• Data (raw and uniformly processed)

• Annotations

• Methods

http://compbio.mit.edu/roadmap
GRID VISUALIZATION

Select Initialize Grid Visualization to obtain a grid of uniformly processed data sets (columns) across all consolidated and/or unconsolidated epigenomes (rows).
Select data views (signal tracks, peak calls, read alignments)
Select grid-cells
Visualize in the epigenome browser

| Uniform Signal Coverage Tracks (-log10(p-value)) (Recommended) | 30 |
| Uniform Signal Coverage Tracks (fold enrichment) | 30 |
| Imputed Signal Coverage Track (predicted -log10(p-value)) | 30 |
| Uniform Peak Calls (data-specific default) |
| Narrow Peaks for histone Chip-Seq and DNase-seq (Recommended) |
| Gapped Peaks for histone Chip-Seq (Recommended) |
| Reprocessed Filtered Alignments (36bp uniform mappability) |
| Compressed view of core chromatin state model (all 127 consolidated epigenomes) |
| Epilo of core chromatin state model (all 127 consolidated epigenomes) |
| Differentially Methylated Regions |
| Reduced Representation Bisulfite Sequencing (RRBS) methylation calls |
| Whole Genome Bisulfite Sequencing (WGBS) methylation calls |
| MeDIP/MRE methylation calls |

Options

- Ignore unconsolidated epigenome data
- Open in a new page (deactivate pop-up blockers)

[Visualize] [Reset all] [Expand all] [Collapse all]

<table>
<thead>
<tr>
<th>TSS-STATE MODEL</th>
<th>TSS-STATE MODEL</th>
<th>Enhancer</th>
<th>Promoter</th>
<th>Gene Body</th>
<th>Histone</th>
<th>DNAse</th>
</tr>
</thead>
<tbody>
<tr>
<td>16-State Model</td>
<td>16-State Model</td>
<td>Enhancer</td>
<td>Promoter</td>
<td>Gene Body</td>
<td>Histone</td>
<td>DNAse</td>
</tr>
<tr>
<td>36-State Model</td>
<td>36-State Model</td>
<td>Enhancer</td>
<td>Promoter</td>
<td>Gene Body</td>
<td>Histone</td>
<td>DNAse</td>
</tr>
</tbody>
</table>

- Adipose (1)
  - Adipose Nuclei
    - Adipose (1)

- Blood & T-cell (14)
- Brain (10)
- Digestive (12)
- ENCODE (16)
- Epithelial (8)
- ES-deriv (6)
- ESC (8)
- Heart (5)
- HSC & B-cell (9)
- iPS (5)
- Mesench (4)
- Muscle (5)
- Myocardial (1)
- Neuroepith (2)
- Other (11)
- Sm. Muscle (4)
- Thymus (2)

Showing 1 to 127 of 127 entries (filtered from 300 total entries)
Example: all available data for Adipose Nuclei
DNaseI tracks across epigenomes

Colors correspond to Roadmap sample (group) colors (Figure 2 in paper)
15-state chromatin state model (ChromHMM)
Sample naming/IDs, sample grouping, data quality, raw data filenames, metadata

The spreadsheet contains 3 sheets (See bottom of sheet)

- Consolidated_EpigeneomId_summary_Table: Main metadata table for 127 consolidated epigenomes
- Consolidated_EpigeneomId_QC: QC measures for DNase-seq and Histone ChIP-seq datasets from all 127 consolidated epigenomes
- Unconsolidated_Release9_QC: QC measures for DNase-seq and Histone ChIP-seq datasets from all unconsolidated epigenomes

Comments in the column headers describe each column (scroll over column header if you open the spreadsheet in a new window or see bottom of sheet if you open it using the Visualize button)
ChIP-seq and DNase-seq uniform reprocessing for consolidated epigenomes

a. Read mapping

b. Mappability filtering, pooling and subsampling

c. Peak Calling

d. Genome-wide signal coverage tracks

e. Quality Control

RNA-seq uniform processing and quantification for consolidated epigenomes

Methylation data cross-assay standardization and uniform processing for consolidated epigenomes

Differentially Methylated Regions (DMRs) and DNA methylation variation

DMR calls across reference epigenomes

DMRs in hESC differentiation (Fig. 4h)

Additional DMR calls

For studying breast epithelia differentiation, DMRs were called from WGBS, requiring at least 5 aligned reads to a cluster (Gascard et al. (2015)). For studying tissue environment vs. developmental origin, DMRs were called from MeDIP-seq.
Imputed signal tracks

Peak calls on imputed data

Chromatin state model based on imputed data (25 state, 127-reference genome).

DATA SOURCE
- 12-mark/127-reference epigenome/25-state Imputation Based Chromatin State Model
- Open in a new page (deactivate pop-up blockers)
- Summarized visualization of all 127 epigenomes using epilogs
Chromatin state learning

DATA SOURCE

- Download URL:
  http://egg2.wustl.edu/roadmap/data/byFileType/chrmmSegmentations/

- Summarized visualization of all 127 epigenomes using epilogs
- Emission, transition probabilities and enrichment of states relative to variants
- MNCMOMICS BED FILES ((Epigenome_id)15_coreMarks_mnemonics.bed): Tab delimited 4 columns (chromosome, start (0-based), stop (1-based), state_label_mnemonic for mnemonics)
- ARCHIVE of all mnemonics.bed files
- BROWSER FRIENDLY FILES ((Epigenome_id)15_coreMarks_dense.bed.gz): The dense BigBED files will allow you to view each epigenome as a single UCSC Genome Browser or IGV
- ARCHIVE of all the dense BigBED files
- [Epigenome_id]15_coreMarks_dense.bed.gz (Same as above except in expanded BED format)
- ARCHIVE of all dense BED files
- [Epigenome_id]15_coreMarks_expanded.bed.gz: The expanded file contains all mnemonics and representative colors
- ARCHIVE of expanded BED files
- STATES FOR EACH 200bp BIN:
  http://egg2.wustl.edu/roadmap/data/byFileType/chromhmSegmentations/ChmmModels/coreMarks/jointModel/final/PCSTERIOR/Max. posterior state label for each 200 bp bin in each chromosome for all chromosomes: bins with the same state label are merged and a label is assigned to the cluster
- ARCHIVE of state-by-line files
- POSTERIOR PROBABILITY FOR EACH 200bp BIN:
  http://egg2.wustl.edu/roadmap/data/byFileType/chromhmSegmentations/ChmmModels/coreMarks/jointModel/final/PCSTERIOR/Posterior probabilities of each state in each 200 bp bin for all chromosomes in all epigenomes

The states are as follows

<table>
<thead>
<tr>
<th>STATE NO.</th>
<th>MNEMONIC</th>
<th>DESCRIPTION</th>
<th>COLOR NAME</th>
<th>COLOR CODE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>TssA</td>
<td>Active TSS</td>
<td>Red</td>
<td>255,0,0</td>
</tr>
<tr>
<td>2</td>
<td>TssAFlnk</td>
<td>Flanking Active TSS</td>
<td>Orange Red</td>
<td>255,69,0</td>
</tr>
<tr>
<td>3</td>
<td>TxFlnk</td>
<td>Transcrip. at gene 5' and 3'</td>
<td>LimeGreen</td>
<td>50,205,50</td>
</tr>
<tr>
<td>4</td>
<td>Tx</td>
<td>Strong transcription</td>
<td>Green</td>
<td>0,128,0</td>
</tr>
<tr>
<td>5</td>
<td>TxWk</td>
<td>Weak transcription</td>
<td>DarkGreen</td>
<td>0,100,0</td>
</tr>
<tr>
<td>6</td>
<td>EnhG</td>
<td>Genic enhancers</td>
<td>GreenYellow</td>
<td>194,225,5</td>
</tr>
<tr>
<td>7</td>
<td>Enh</td>
<td>Enhancers</td>
<td>Yellow</td>
<td>255,255,0</td>
</tr>
<tr>
<td>8</td>
<td>ZNF/Rpts</td>
<td>ZNF genes &amp; repeats</td>
<td>Medium Aquamarine</td>
<td>102,205,170</td>
</tr>
<tr>
<td>9</td>
<td>Het</td>
<td>Heterochromatin</td>
<td>PaleTurquoise</td>
<td>138,145,208</td>
</tr>
<tr>
<td>10</td>
<td>TssBiv</td>
<td>Bivalent/Poised TSS</td>
<td>IndianRed</td>
<td>205,92,92</td>
</tr>
<tr>
<td>11</td>
<td>BivFlnk</td>
<td>Flanking Bivalent TSS/Enh</td>
<td>DarkSalmon</td>
<td>233,150,122</td>
</tr>
<tr>
<td>12</td>
<td>EnhBiv</td>
<td>Bivalent Enhancer</td>
<td>DarkKhaki</td>
<td>189,183,107</td>
</tr>
<tr>
<td>13</td>
<td>ReprPC</td>
<td>Repressed PolyComb</td>
<td>Silver</td>
<td>128,128,128</td>
</tr>
<tr>
<td>14</td>
<td>ReprPCwK</td>
<td>Weak Repressed PolyComb</td>
<td>Gainsboro</td>
<td>192,192,192</td>
</tr>
<tr>
<td>15</td>
<td>QuiesK</td>
<td>Quiescent/Low</td>
<td>White</td>
<td>255,255,255</td>
</tr>
</tbody>
</table>
Clustering of epigenomes reveals common lineages, common properties

DATA SOURCE

Correlation matrices (Data format: RData)
- Download URL: http://egg2.wustl.edu/roadmap/data/byDataType/celltype_clustering/correlation_matrices/

Newick formatted optimally ordered hierarchical trees, annotated with bootstrap scores
- Download URL: http://egg2.wustl.edu/roadmap/data/byDataType/celltype_clustering/bootstrap_results/

For each analyzed mark, we calculated Pearson correlation values between all pairwise combinations of reference epigenomes using the marks signal confidence scores (-log10(Poisson p-value)) within 200bp of the genomic regions deemed relevant for that mark. Relevance of regions is determined by whether a region was called in a particular (mark-matched) chromatin state with posterior probability of > 0.95 in any of the reference epigenomes. For H3K4me1, H3K27ac and H3K9ac we used state Enh, for H3K4me3 state TssA, for H3K27me3 state ReprPC, for H3K36me3 state Tx and for H3K9me3 state Hei, unless otherwise noted (all based on the 15-state core model).

The resulting correlation matrices were used as the basis for a distance matrix for complete-linkage hierarchical clustering, followed by optimal leaf ordering (Bar-Joseph et al. (2001)). Bootstrap support values are derived from 1,000 random samplings with replacement from all regions considered for a particular mark and a bootstrap tree was estimated for each resampling. The bootstrap support for a branch corresponds to the fraction of bootstrapped trees that support the bipartition induced by the branch.

In parallel to this, all correlation matrices mentioned above were used to perform Multi-Dimensional Scaling (MDS) analyses using R. Some concept code that can be used with the provided correlation matrices to perform hierarchical clustering and MDS analyses:

```r
mark <- "H3K4me1"; state <- 7; # example mark/state combination: H3K4me1 in the Enh state
load(paste("cor_", mark, ",", state, ",.RData", sep=""));
d <- as.dist(1-markcor);
hclust_res <- hclust(d);
MDS_res <- cmdscale(d, eig=TRUE, k=nrow(markcor)-1);
```
Delineation of DNasel-accessible regulatory regions

Clustering of DNasel-accessible regulatory regions to identify modules of coordinated activity

**DATA SOURCE**

BED files with coordinates for regions in each module:

- **Promoter**: 81,232 promoter regions (1.43958% of genome)
  Download URL: http://egg2.wustl.edu/roadmap/data/byDataType/regulatory_modules/modules_prom/BED_files/
  ![Visualize](no_image)
  ![Open in a new page](no_button)
- **Enhancer**: 2,328,936 putative enhancer regions (12.6385% of genome)
  Download URL: http://egg2.wustl.edu/roadmap/data/byDataType/regulatory_modules/modules_enh/BED_files/
  ![Visualize](no_image)
  ![Open in a new page](no_button)
- **Dyadic**: 129,960 dyadic promoter/enhancer regions (0.985296% of genome)
  Download URL: http://egg2.wustl.edu/roadmap/data/byDataType/regulatory_modules/modules_dyadic/BED_files/
  ![Visualize](no_image)
  ![Open in a new page](no_button)

High resolution figures of module heatmaps for each module:

- **Promoter**: PDF, PNG
- **Enhancer**: PDF, PNG
- **Dyadic**: PDF, PNG

The order in which modules are plotted in the heatmaps:

- **Promoter**: TXT
- **Enhancer**: TXT
- **Dyadic**: TXT
Predicting motifs and active regulators in each cell-type/tissue/lineage

DATA SOURCE

- Primary motif resource:
  http://compbio.mit.edu/encode-motifs/
- Data underlying Figures Extended Data 8a, S13a (clustered), S13b,c (unclustered), S13d,e (unique):
  http://egg2.wustl.edu/roadmap/data/byDataType/motifanalysis/pouyak/
- Motif enrichments (enhancer cluster centric):
  http://egg2.wustl.edu/roadmap/data/byDataType/motifanalysis/pouyak/viewByCluster/bycluster.html

DNA Motif Positional Bias in Digital Genomic Footprinting Sites

DATA SOURCE

- Primary motif resource:
  http://compbio.mit.edu/encode-motifs/
- DNase/DGF Footprint calls:
  http://egg2.wustl.edu/roadmap/data/byDataType/dgfootprints/
  **Format:** 5 column BED files. 4th column is footprint ID. 5th column is FOS score. See below.
- Data underlying Figures 8, Extended Data 9b, 9c:
  http://egg2.wustl.edu/roadmap/data/byDataType/motifanalysis/zhizhou/
Tissue-specific activity of disease-associated regions

We tested the enrichment of SNPs from individual Genome-wide Association Studies (GWAS) for the gapped peak call sets for histone marks H3K4me1, H3K4me3, H3K36me3, H3K9me3, H3K27me3, and H3K27ac as well as the DNase peak call set based on MACS2 in each reference epigenome where available. The SNPs used were curated into the NHGRI GWAS catalog (Welter et al. (2014)) and obtained through the UCSC Table Browser (Karolchik et al. (2004)) on September 12, 2014. We restricted the enrichment analysis to chr1-22 and chrX. We defined a study to be a unique combination of annotated trait and PubMedID. To reduce dependencies between pairs of SNPs assigned to the same study, we pruned SNPs such that no two SNPs were within 1MB of each other on the same chromosome. The pruning procedure considered each SNP in ranked order of p-value with the the most significant coming first, and we retained a SNP if there was no already retained SNP on the same chromosome within 1MB. We computed hypergeometric p-values for the enrichment of each pruned set of SNPs overlapping peak calls against the pruned GWAS catalog as the background. We estimated separately for each mark a mapping from a p-value to a false discovery rate across tests for all study and reference epigenome combinations by generating 100 randomized versions of the pruned GWAS catalogs shuffling which SNPs were assigned to which study and computing the average fraction of reference epigenomes study combinations that reached that level of significance (in a continuous mapping of p-values to FDR) using randomized catalogs divided by the number based on the actual GWAS catalog.

Further data available in Table S6 online
Summarized visualization of all 127 epigenomes using epilogs

epilogs