doi:10.1038/nature11247

Contents

Introduction

We have structured the Supplement for this paper in a logical manner, trying to isolate the different analyses that underlie each different component of the main paper. In each section we state the main analysts responsible for the work, the figure or text section in the main paper this refers to, a text description of the method and finally a code bundle (see below). In cases where the majority of the method is encompassed by a specific companion paper, the Supplement is appropriately short and refers the reader to that companion paper.

The "Code Bundles" are a new aspect of supplementary information that we believe is helpful for these large, data rich papers. Every analysis presented in the paper has a series of source data files, which are then transformed into usually a single output file for that analysis from which the final Figure and statements in the paper are made. Where possible raw data files are assigned tracking IDs from the ENCODE DCC (as well as often accessions numbers for the rawest data forms at SRA), and each Supplementary section or Code Bundle explicitly lists these input files. The supplement methods then give a textual description of the methodology, and the code bundle provide the actual scripts and manipulations that correspond to that methodology. When we use specific analytical programs (e.g., peak callers, HMM methods) we reference appropriate papers (or websites) describing those programs.

We have emphasized transparency in this process (at the cost of consistency), meaning that we have exposed the large diversity of scripting languages and software components used by the various analysts in the project. For example, if a critical step was in fact executed by a complex piped UNIX command line, this command line is provided explicitly. This diversity in analysis methods should not be a surprise to any scientist working in large-scale genomics, but might be confusing or frustrating with people with less large-scale data handling experience. We apologize in advance for this diversity, but it is important to realize that our goal here is not to provide easy-to-use programs, or robust engineering solutions (there are separately funded projects to create such things), but rather to provide scientific transparency of our analytical results. By having the input data sets, a text description of the method, the actual code implementing the method and finally the output, along with a well-defined section with named analysts, we hope to provide a highly transparent view of the analysis we have performed.

In addition we have established a virtual machine instance of the software, using the code bundles, where each analysis program has been tested and run. Where possible the VM enables complete reproduction of the analysis as it was performed to generate the figures, tables or other information. However in some cases the analysis involves highly parallelised processing within a specialised multiprocessor environment. In these cases, a partial example has been implemented leaving it to the reader to decide whether and how to scale to a full analysis. We hope that this structure gives readers the opportunity to run the same analyses in the wild. During implementation of the code bundles to establish the VM, there have necessarily been tweaks to the code and installation of packages that had been omitted from the code bundle through oversight. Therefore we recommend use of the VM as a first step. Instructions for obtaining and running the VM can be found at http://encodeproject.org/ENCODE/integrativeAnalysis/VM.

For inquiries about the content of this supplementary information and specifically the content of the code, please email first the joint author email "encode_authors@ebi.ac.uk" , stating the inquiry and section; please do not email the analyst directly without making contact through this address so we can ensure that commonly asked questions are not a burden to analysts. Although the main analytical programs can be run on many different datasets in different environments, please do not consider this collection of programs and scripts to be a portable analysis system.

Introduction to Data Sources.

ENCODE data is submitted to the ENCODE Data Coordinating Centre (DCC) at the University of California, Santa Cruz (UCSC) (see http://genome.ucsc.edu/ENCODE/). Data is quality reviewed and released for display as tracks in the UCSC Genome Browser (http://genome.ucsc.edu/cgi-bin/hgTracks?db=hg19) and for download at http://genome.ucsc.edu/ENCODE/downloads.html. There are a number of useful tools including track search and file search (see http://genome.ucsc.edu/ENCODE/search.html) to assist with location of data. The ENCODE portal (http://encodeproject.org) provides organized access to the data, along with instructional material and additional resources shedding light on the data provided.

The analysis process was a distributed effort between many groups. Individual analysts will have downloaded and processed files from the ENCODE download site, and we have created intermediate and final analysis products in various forms. We have attempted to organise this process for our own sanity by establishing centralised descriptions of files and analysis processes, particularly through the private ENCODE wiki site. However inevitably data ends up in many places and for good reason. For instance we have established repositories of data close to the large compute resources such as the EBI to allow us to process many files simultaneously. We have also been faced with the problems of moving large files over the internet which we have solved by a number of methods including disc transfer and using the Aspera transfer client (http://www.asperasoft.com/downloads/connect). Now we have brought the analysis to completion we are making the analysis data available for viewing either through UCSC tracks, or through a UCSC datahub (Link required). In addition we provide links to the ftp server at ebi.ac.uk which contains an organised file structure with the ENCODE data. Analysis datasets are located in ftp.ebi.ac.uk:pub/databases/ensembl/encode/supplementary/integration_data_jan2011 in the directories in byDataType. We provide code bundles, in particular at ftp.ebi.ac.uk:pub/databases/ensembl/encode/supplementary/. In the code bundles we try to provide the data files used, where this makes sense or links to the correct file location. Finally in some cases the archival site for an analysis or file may be another remote URL. In the case of file downloads at UCSC, the references below list the public site (hgdownload), although some files may not yet have completed review and must be accessed from the staging site (hgdownload-test). Such references below are prefixed with asterisks to alert the reader.

The ENCODE information is also available through other genomics portals, including Ensembl (www.ensembl.org) and the NCBI Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/info/ENCODE.html) and the raw data are deposited in the sequence read archives (SRA: http://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?, http://www.ebi.ac.uk/ena/).

Note for reviewers: The Main ENCODE portal (www.encodeproject.org) and other publically accessible routes for ENCODE data include all the ENCODE data used in the paper but also data submitted post the cut-off for data processing into this paper. To provide clarity of the precise datasets used for the analysis, we have grouped the main element-based datasets described in this paper at http://genome.ucsc.edu/cgi-bin/hgHubConnect, and this can be accessed by loading the URL http://www.ebi.ac.uk/~anshul/public/encodeRawData/dcc/hub/awgHub/hub.txt in the My Hubs section.

We would appreciate your opinion about whether such a frozen dataset associated with these papers would be valuable to the community over the long term, or whether this is just relevant for review.

Section A: Summary of ENCODE Data

Main Analysts:

Ian Dunham

Principally Related to:

Deprecated figure removed from the final version. However the code is still on the VM and can be run there, as well as the code bundle still being available. Internal references to fig1 have been left as they are inside the code bundle. Think of it as an easter egg.

Methods:

A. Circos image of ENCODE data. ENCODE data files for GM12878 were downloaded from the analysis ftp site (ftp.ebi.ac.uk:pub/databases/ensembl/encode/supplementary/integration_data_jan2011), grouped into appropriate data types and processed into 100kb windows using slc2density.pl (in CircosFig_code_bundle/bin). For TFs, single linkage clusters (slc) were produced for each of the TFSS, TFNS and remainder (other) divisions of the TFs using saturate (in CircosFig_code_bundle/bin) from the IDR processed spp peak calls (see filelists in CircosFig_code_bundle/circos_files/TFs/).

Circos (Krzywinski, M. et al. Circos: an Information Aesthetic for Comparative Genomics. Genome Res (2009) 19:1639-1645, PMID: 19541911) was used to generate the image using the config file and ancillary files in the code bundle. The Circos software (http://circos.ca/) can be downloaded from http://circos.ca/software/download/. The order and configuration of tracks is specified in CircosFig_code_bundle/fig1.conf . After placing the files in appropriate directories and registering the locations in CircosFig_code_bundle/fig1.conf, run "circos -conf fig1.conf".

B. Panel B was an example figure that was removed in review. The files remain in the code bundle. ENCODE data was viewed in the UCSC Genome browser (http://genome.ucsc.edu/) using the session file in CircosFig_code_bundle/browser. The browser image was exported as pdf and then edited in Adobe Illustrator for clarity and aesthetics.

Supplementary Table 1, section A gives details of the individual methods and their acronyms.

Location of Code Bundle:

ftp.ebi.ac.uk:pub/databases/ensembl/encode/supplementary/CircosFig_code_bundle.tar.gz

Datasets used:

GM12878 SPP uniform calls
GM12878 UW DNase hotspots
GM12878 FAIRE peaks
Gencode version 7 annotation
GM12878 RNA elements for whole cell polyA plus and polyA minus RNA-seq
GM12878 RRBS methylation data
GM12878 Combined Segmentation

All available in the CircosFig_code_bundle/circos_files directory downloadable from

ftp://ftp.ebi.ac.uk/pub/databases/ensembl/encode/supplementary/CircosFig_code_bundle.tar.gz

Section B: Conservation among humans and mammals.

Main Analysts:

Luke Ward, Javier Herrero

Principally Related to:

Figure 1.

Methods:

Gencode v7 annotations were parsed as follows: all features annotated as "CDS" were selected as CDS and all protein-coding genes annotated as "UTR" were selected as UTRs. Protein-coding and non-coding genes were selected as genes, and the set difference between bases annotated as being in a gene and bases annotated as being exonic were selected as intronic.

The genome was masked as follows: Autosomes from the hg19 were selected, and the following regions were excluded: RepeatMasker and SimpleRepeat regions, from the UCSC table browser; ENCODE blacklist regions (both the Duke and empirically-defined regions); all CpG islands from the UCSC table browser, and any dinucleotide that is "CG" in either the reference genome or when mutated to a 1000 Genomes SNP observed in the YRI population; any regions not falling into either an EPO alignment block or a 1000 Genomes callable region.

Certain features were split as follows: Novel intronic RNA contigs were selected from each experiment by selecting contigs that entirely overlap an intron and have no overlap with an exon or any base within 2kb of a TSS, and novel intergenic RNA contics were selected from each experiment by selecting contigs that are entirely annotated as intergenic and have no overlap with any base within 2kb of a TSS. A subset of 49 PWM families were selected from the literature, for which at least at least one ChIP-seq experiment for a corresponding protein displayed a significant enrichment. Instances of motifs matching these PWMs were merged within each family. Then, these motif instances were split into those that were either bound in at least one ChIP-seq experiment by a matching factor, or never bound by a matching factor in any experiment.

Over all features, mean GERP score, heterozygosity, and derived allele frequency (DAF) were calculated. Heterozygosity was calculated basewise as 2pq, where p and q are allele frequencies estimated from the pilot sample of the 1000 Genomes YRI population; the heterozygosity for any bases without detected variants segregating was estimated as zero. For SNPs with a defined ancestral allele as called by the 1000 Genomes Project, the frequency of the derived allele was used to calculate a mean DAF across all SNPs overlapping a feature.

In the bottom-right panel (conservation of bound GR motifs), all positive GERP conservation scores on occupied motifs are considered. The mean GERP score is compared to the information content of the motif using a Pearson correlation. The analysis is expanded 10 base pairs each side to give some context although these positions are not used in the correlation test.

Location of Code Bundle:

ftp://ftp.ebi.ac.uk/pub/databases/ensembl/encode/supplementary/Fig1_code_bundle.tar.gz

Has the source of scripts to perform the analysis and make the figures, as well as the data table underlying the figures.

Datasets used:

Basewise conservation scores by GERP
1000 Genomes pilot SNPs from the YRI population
Gencode v7 annotations and TSS clusters (CAGE and non-CAGE all included)
All SPP narrowPeak Chip-Seq files
(ftp.ebi.ac.uk:pub/databases/ensembl/encode/supplementary/integration_data_jan2011/byDataType/peaks/jan2011/spp/optimal/)
All FAIRE and DNAse data (will look up DCC accession)
All long and short RNA contigs (will look up DCC accession)
All human literature motifs from Pouya Kheradpour and his manual annotation of them into families cross-referenced with ChIP proteins (http://www.broadinstitute.org/~pouyak/encode-motif-disc/ also at
ftp.ebi.ac.uk:pub/databases/ensembl/encode/supplementary/integration_data_jan2011/byDataType/motifs/jan2011/)

Section C: Derived allele frequency at ENCODE elements

Main Analysts:

Luke Ward

Principally Related to:

Main Figure 1, histogram panel.

Methods:

The genome was annotated, 1000 Genomes pilot data from the YRI population was accessed, and the genome was masked as described for Figure 1.

The derived allele spectrum was inspected both genomewide and within primate-specific elements. Primate-specific elements were defined as follows: from the Ensembl 57 11-way mammal EPO alignments, regions were selected that were at least 200 base pairs and contained only primate sequences, but were part of a longer alignment block containing at least one non-primate species.

Location of Code Bundle:

ftp://ftp.ebi.ac.uk/pub/databases/ensembl/encode/supplementary/Figure1.R

R Script parses the same results.txt that is included in the Figure 1 bundle above (ftp://ftp.ebi.ac.uk/pub/databases/ensembl/encode/supplementary/Fig1_code_bundle.tar.gz).

Datasets used:

1000 Genomes pilot SNPs from the YRI population

Gencode v7 annotations and TSS clusters (CAGE and non-CAGE all included).

All SPP narrowPeak Chip-Seq files (ftp://ftp.ebi.ac.uk:pub/databases/ensembl/encode/supplementary/integration_data_jan2011/byDataType/peaks/jan2011/spp/optimal/)

All FAIRE and DNAse data (ftp://ftp.ebi.ac.uk:pub/databases/ensembl/encode/supplementary/integration_data_jan2011/byDataType/openchrom/jan2011/)

All long and short RNA contigs (ftp://ftp.ebi.ac.uk:pub/databases/ensembl/encode/supplementary/integration_data_jan2011/byDataType/rna_elements/jan2011/LongRnaSeq/ and ftp://ftp.ebi.ac.uk:pub/databases/ensembl/encode/supplementary/integration_data_jan2011/byDataType/rna_elements/jan2011/ShortRnaSeq/)

All human literature motifs from Pouya Kheradpour and his manual annotation of them into families cross-referenced with ChIP proteins (http://www.broadinstitute.org/~pouyak/encode-motif-disc/

also at
ftp://ftp.ebi.ac.uk:pub/databases/ensembl/encode/supplementary/integration_data_jan2011/byDataType/motifs/jan2011/)

Section D: Integration of ENCODE data around known genomic features: Promoter Anchored Integration.

Main Analysts:

Xianjun Dong, Chao Cheng

Principally Related to:

Figure 2. Panels showing aggregated signals of histone modification (HM) around the TSS and TTS of all HCP and LCP promoters and the aggregated signals of transcription factor (TF) binding sites around the TSS of all HCP and LCP promoters were removed to supplementary information (Supplementary Figures 1 and 2, section Z) after review. Figure 2A shows the consistency of the predictive expression with experimental measurement of PolyA+ RNA by CAGE in K562 cell, as well as the relative importance of 14 in the classification and the regression models. Figure 2B shows the consistency of the predictive expression with experimental measurement by CAGE, as well as the relative importance of 40 TFs in the classification and the regression models.

Methods:

In Supplementary Figure 1A, section Z: (1) The DNA region around each Gencode transcript was divided into 80 bins -- 40 bins for [-2kb, +2kb] around TSS and 40 bins for [-2kb, +2kb] around TTS, each of 100bp in length; (2) The average binding signal of each HM in each bin was calculated for all transcripts; (3) The aggregated signal of each histone modification (HM) was calculated for all high CpG content (HCP) and low CpG content (LCP) promoters, respectively. Supplementary Figure 1A, section Z shows the binding of four selected HMs in the 80 bins in K562, each representing a functional category (e.g. H3k9ac for activation, H3k27me3 for repression, H3k4me1 for enhancer and H3k36me3 for elongation) with the red and green curves showing the binding signal for HCP and LCP, respectively.

In Supplementary Figure 1B, section Z: (1) The DNA region around each Gencode TSS [-2kb, 2kb] was divided into 40 bins, each of 100bp in length; (2) The average binding signal of each TF in each bin was calculated for all TSSs; (3) The aggregated signal of each TF was calculated for all expressed high CpG content (HCP) and low CpG content (LCP) promoters, respectively. Supplementary Figure 1B, section Z shows the binding of those four selected TFs in the 40 bins in K562. The red and green curves show the binding signal for HCP and LCP, respectively. Only promoters expressed in Poly A+ extracted from K562 whole cells (cpkw) were used in the calculation.

In Figure 2A: We used a two-step model to predict the expression levels of Gencode genes. In this figure, we show the results of the model in K562 for predicting expression in Poly A+ whole cell RNAs. Only the promoters that are expressed in at least one cell line were selected. The density signals of DNase I plus 12 HMs in the bin that are correlated best with expression levels were used as the predictors. (1) We constructed a random forest classification model to predict whether a promoter is expressed or not. (2) We constructed a linear regression model to predict the expression levels of a promoter. (3) The two models were combined by setting the predicted values $\hat{y}_i = I(C_i = 1)R_i$, where C_i is the results from the classification model (C_i =1 if promoter i is predicted to be expressed, and 0 otherwise); R_i is the predicted value for promoter i by the regression model.

In Figure 2B: We used a two-step model to predict the expression levels of Gencode promoters based on the Random Forest (RF) method. In this Figure, we show the results of the model in K562 for predicting expression in Poly A+ whole cell RNAs. Only the promoters that are expressed whole cell Poly A+ RNAs from at least one cell line were selected. The binding signals of 40 TFSSs in the 100bp bin right at the TSS were used as the predictors (if there are multiple ChIP-Seq datasets for a TF in K562, we choose the one best correlated with expression levels).

(1) We constructed a RF classification model to predict whether a promoter is expressed. (2) We constructed a RF regression model to predict the expression levels of a promoter. (3) The two models were combined by setting the predicted values $\hat{y}_i = I(C_i = 1)R_i$, where C_i is the results from the classification model (C_i =1 if promoter *i* is predicted to be expressed, and 0 otherwise); R_i is the predicted value for promoter *i* by the regression model.

The performance of the classification model, the regression model and the combined two-step model were evaluated based on cross-validation. Namely, the data was divided into a training data and a testing data. A model was trained using the training data and then applied to the testing data to make predictions. We used AUC to represent the accuracy of the classification model, which measured the area under the ROC curve (sensitivity versus 1-specificity of a classification model). For the regression model, the predictive accuracy was measured by *r* (the Pearson correlation coefficient between the predicted value and the experiment value), R^2 (the fraction of variance of gene expression explained by the model), and RMSE (rooted mean squared error).

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y}_i)^2}$$

$$RMSE = \sqrt{\sum_i (y_i - \hat{y}_i)^2 / n}$$

In Figure 2A, it shows the predictive accuracy of the model in K562 Poly A+ whole cell RNA sample. Only the promoters that are expressed in at least one cell line were selected.

The relative importance of a HM was calculated based on its Gini index in the classification model, and R^2 decomposition for linear regression model. The relative importance of a TF was also calculated based on its Gini index in the classification model, and as increase in node purity for regression model. All the calculation was implemented in the R package "`randomForest`" and "`relaimpo`".

Location of Code Bundle:

The R package "`randomForest`" was used to implement the RF model and basic lm() for linear regression model.

A code bundle is provided at
ftp.ebi.ac.uk:pub/databases/ensembl/encode/supplementary/Fig2_code_bundle.tar.gz

Datasets used:
* Histone modification data
The normalized bigWig files for all HMs from the Broad Institute:
Version: hg19 / v7
URL:
ftp://ftp.ebi.ac.uk:pub/databases/ensembl/encode/supplementary/integration_data_jan2011/byDataType/signal/jan2011/bigwig

* TF binding data
The "bedGraph" or "bigWig" files for all TF binding tracks.
Version: hg19 / v7
URL: http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/

* Gene annotation (v7)
Version: hg19 / v7
URL: ftp://ftp.sanger.ac.uk/pub/gencode/release_7/gencode.v7.annotation.gtf.gz

* RNA transcription quantifications -- TSS-based

Version: hg19 / v7
URL: ftp://genome.crg.es/pub/Encode/data_analysis/TSS/Gencodev7_TSS_July2011.gff.gz

Section E: Patterns of chromatin marks around transcription factor binding sites.

Main Analysts:

Anshul Kundaje, Sofia Kyriazopoulou-Panagiotopoulou

Principally Related to:

Figure 3: Patterns of histone modifications and nucleosome positioning at transcription factor binding sites

Methods:

Figure 3 A: Clustered Aggregation Plots (Supplementary Figures 5 and 6, section E)

We used clustered aggregation plots (CAGT) to explore the diversity of magnitude and shapes of signal profiles of specific histone modifications and nucleosome occupancy (target marks) in the vicinity of binding sites of each transcription factor (target TF) in a different cell-line (target cell-line)[1].

We define the binding sites of a target TF in a target cell-line as the set of confident and reproducible peak summits identified from ChIP-seq data. We then extract normalized signal profiles (Hoffman *et al*., manuscript in preparation) of the target mark in a 1000 bp (+/- 500 bp) window centered at each binding site. Since ChIP-seq peaks do not contain explicit strand information, the binding sites are always defined on the '+' strand and each target mark profile is oriented in the 5' to 3' on the '+' strand.

First, we compute a global aggregation profile (the subplot labeled "all" followed by the total number of target TF peaks). The global aggregate profile shows the position-wise mean (black line), lower 10th and upper 90th percentile (grey borders) computed using signal profiles centered at all binding sites of the factor. However, such a global average is only informative if one makes the strong assumption that all the profiles are sampled from a homogeneous population. Results show that this assumption is almost always violated.

In order to dissect the effect of differences in overall magnitude of the target marks signal, we then split target mark profiles into a low signal and high signal category. A profile belongs to the high signal bin only if its robust maximum value (0.99 quantile) over all positions is greater than 5. We compute the summary statistics (median, upper and lower quantiles) for all profiles in each bin. These subplots are labeled as "high" and "low" in the figures with the corresponding fraction of profiles that fall into each category. The high signal fraction indicates the fraction of binding sites of the target TF that are actually enriched for the target mark. We drop the low signal component from further analysis as these profiles simply represent noise.

We then focus on clustering the high signal profiles into a natural set of groups based on shapes/patterns of the profiles. We first standardize each profile (i.e subtract mean and divide by the standard deviation across the profile) in order to remove any magnitude effect and simply retain shape information. We then use a combination of k-means clustering followed by hierarchical clustering to get a robust and stable set of pattern groups (See ref [1] for more details). A key feature of this step is that it also accounts for hidden directionality in the shapes of the patterns. Functional elements of various types (such the binding site of another TF or a gene transcription start site) on either side of the binding site can affect the overall pattern of the target mark. Since we do not know a-priori, the identity and relative orientation of these hidden variables and how they affect pattern shape and orientation; our clustering procedure automatically flips (reverses) and merges pattern groups that are mirror images of each other

and also keeps track of which profiles get flipped. Thus, we simultaneously obtain a concise summary of the distinct pattern shapes (irrespective of orientation) and also uncover hidden directionality for patterns that are asymmetric about the target TF binding sites. The subplots labeled "pattern X" in the figures represent the learned pattern groups with the fraction of total number of binding sites that belong to each group.

We typically obtain 4-10 distinct pattern groups and there is a predominant tendency for patterns groups to have highly asymmetric shapes

Figure 3B: Asymmetry ratio

In order to quantify the degree of shape asymmetry of profiles of a particular target mark over all TFs, we compute an asymmetry ratio for each target TF dataset with respect to the target mark and then plot the empirical cumulative distribution of asymmetry ratios over all unique TFs. The asymmetry ratio of a target mark for a target TF dataset is the fraction of all high signal profiles of the target mark that show asymmetric patterns around the binding sites of the target TF (See ref [1] for detailed parameters). Some TFs are represented more often than others due to multiple experiments in different cell-lines by different production groups. In order to avoid this bias, the asymmetry ratio for each unique TF is computed as the weighted average of asymmetry ratios for all datasets corresponding to the TF. The weights are proportional to the number of peaks in each dataset.

Motif proximity

For GATA1 and CTCF, we wanted to test whether the asymmetry of patterns is related to the relative orientation of the DNA binding sequence motif for these TFs. Since we cluster asymmetric patterns that are mirror images of each other into the same group, each asymmetric pattern group almost always contains profiles that are flipped. Thus, we can test whether the flipping of a profile in a particular pattern group correlates with a corresponding strand flip of the sequence motif of the TF within the peak region (Fisher test). We restrict the analysis to TF peaks that have a motif. If a peak has multiple motifs we use the motif that is nearest to the peak summit. Motif hits within peaks were obtained (Kheradpour and Kellis, manuscript in preparation). Below, we show the contingency tables for (i) pattern groups of H3k27me3 at CTCF sites in H1hesc (ii) pattern groups of H3k27ac at GATA1 sites in K562 (Supplementary Table 1 and 2, section E). In both cases, none of the pattern groups show significant p-values indicating no significant relationship between pattern asymmetry and motif orientation.

Relationship to nucleosome occupancy

Supplementary Figure 1, section E: nucleosomes@CTCF in K562

We ran the CAGT pipeline using nucleosome sequencing data (MNase-seq) as the target mark around CTCF sites in K562. We see that 99.6% of the profiles belong to the high signal category indicating that almost all CTCF peaks have significant nucleosome occupancy and well-positioned nucleosomes flanking them. However, we also observe distinct diverse shapes of nucleosome positioning, the largest cluster (pattern 1) being largely symmetric and the remaining clusters showing various asymmetric patterns.

Supplementary Figure 2, section E: H3K9ac@CTCF and corresponding nucleosomes occupancy

We used CAGT to analyze H3k9ac profiles at CTCF peaks in K562. We observe that only 27.45% of all CTCF peaks are enriched for H3k9ac. We then wanted to analyze the relationship of shape patters of H3k9ac to corresponding nucleosome occupancy profiles. Rather than averaging profiles that belong to each shape cluster using the original scale, we standardize each H3k9ac

profile and then compute the median of the standardized profiles that belong to each shape cluster (shown as black lines in the pattern subplots). We then extract nucleosome occupancy profiles for all peaks that belong to each H3k9ac pattern cluster; standardize them; flip/reverse profiles around peaks whose corresponding H3k9ac profiles were flipped in the CAGT analysis and compute the median nucleosome occupancy profile. We observe that while the H3k9ac profiles are largely asymmetric, the nucleosome occupancy profiles show strong symmetry with well-positioned nucleosomes on either side of the CTCF. This indicates that histones on either side of the CTCF site tend to be differentially marked with H3k9ac.

Supplementary Figure 3, section E: H3K79me2@CTCF and corresponding nucleosome occupancy

We performed a similar analysis of H3k79me2 at CTCF peaks in K562. Once again only 18% of the CTCF peaks are enriched for H3k79me2. We observe asymmetric shapes of H3k79me2 but symmetric positioning and occupancy of nucleosomes. Only ~50% of these peaks are within 5Kbp of TSS and > 80% are within GENCODEv7 gene boundaries (which is in concordance with the enrichment of H3k79me2 in actively transcribed domains). Hence, proximity to TSSs does not entirely explain the pattern asymmetry of H3k79me2.

Supplementary Figure 4, section E: TAF1@H3k4me3 with nucleosome

For TSS-proximal TFs such as TAF1 and for the histone mark H3k4me3 which is typically enriched at active promoters, we observe that the asymmetry of H3k4me3 is strongly correlated with asymmetry of corresponding nucleosome occupancy patterns.

Location of Code Bundle:

Link to CAGT code: http://code.google.com/p/cagt/

Code to generate normalized signal files: http://code.google.com/p/align2rawsignal/

Code to extract signal: http://code.google.com/p/extractsignal/

Code to produce plots: http://code.google.com/p/cagt/

A code bundle is provided at ftp.ebi.ac.uk:pub/databases/ensembl/encode/supplementary/Fig3-SupplementaryData.tar.gz

Datasets used:

Target TF Peak calling datasets (SPP peak caller)

All peak call datasets:
ftp://ftp.ebi.ac.uk:pub/databases/ensembl/encode/supplementary/integration_data_jan2011/byDataType/peaks/jan2011/spp/optimal/

- TAF1 in Gm12878:
spp.optimal.wgEncodeHaibTfbsGm12878Taf1Pcr1xAlnRep0_VS_wgEncodeHaibTfbsGm12878ControlPcr1xAlnRep0.narrowPeak.gz

- GATA1 in K562:
spp.optimal.wgEncodeSydhTfbsK562bGata1UcdAlnRep0_VS_wgEncodeSydhTfbsK562bInputUcdAlnRep1.narrowPeak.gz

- CTCF in H1hesc:
spp.optimal.wgEncodeBroadHistoneH1hescCtcfStdAlnRep0_VS_wgEncodeBroadHistoneH1hescControlStdAlnRep0.narrowPeak.gz

- CTCF in K562:
spp.optimal.wgEncodeBroadHistoneK562CtcfStdAlnRep0_VS_wgEncodeBroadHistoneK562ControlStdAlnRep1.narrowPeak.gz

Target mark normalized signal datasets

All normalized signal datasets:
ftp://ftp.ebi.ac.uk:pub/databases/ensembl/encode/supplementary/integration_data_jan2011/byDataType/signal/jan2011/bedgraph

- H3k4me3 in Gm12878:
wgEncodeUwHistoneGm12878H3k4me3StdAln_2Reps.norm5.rawsignal.bedgraph.gz

- H3k27ac in K562:
wgEncodeBroadHistoneK562H3k27acStdAln_2Reps.norm5.rawsignal.bedgraph.gz

- H3k9ac in K562:

wgEncodeBroadHistoneK562H3k9acStdAln_2Reps.norm5.rawsignal.bedgraph.gz

- H3k79me2 in K562:

wgEncodeBroadHistoneK562H3k79me2StdAln_2Reps.norm5.rawsignal.bedgraph.gz

- nucleosome occupancy in K562:
wgEncodeSydhK562NucleosomeRep0.bw30.norm5.rawsignal.bedgraph.gz

Extracted signal files of all target marks around all TF peaks
http://www.ebi.ac.uk/~anshul/public/encodeRawData/extractSignal/jan2011/cagt_1000/

CAGT result tables that were used for plots

https://sites.google.com/site/anshulkundaje/projects/cagt

Motif data

http://www.broadinstitute.org/~pouyak/encode-motif-disc/

CTCF known motif: CTCF_known1

GATA1 known motif: GATA_known1

Section F: Co-association of transcription factor binding sites

Main Analysts:

Kevin Y. Yip, Nitin Bhardwaj, Ben Brown, Anshul Kundaje, Manoj Hariharan, Nathan Boley, Joel Rozowsky, Peter Bickel, Mike Snyder, Mark Gerstein

Principally Related to:

Figure 4

Methods:

The binding peaks from all transcription factor (TF) ChIP-seq datasets were collected from the ENCODE uniform peak calling pipeline. We used the set of TF binding peaks called by PeakSeq[2]. Peak calling thresholds were determined based on consistency and reproducibility using the Irreproducible-discovery rate (IDR) framework (Kundaje *et al.,* manuscript in preparation). Problematic regions, such as repeats, were removed. For every dataset x and every dataset y from the same cell line, a base overlap ratio between the binding peaks in the two datasets was computed as $|X \cap Y|/|X|$, where X and Y are the base positions covered by the binding peaks from x and y respectively, $X \cap Y$ is their intersection, and $|A|$ denotes the number of base positions in any set A. The base overlap ratio measure is asymmetric.

To evaluate whether a base overlap ratio is statistically significant, we applied a sampling procedure that adopts genome structure correction (GSC) [3]. Briefly, we first segmented the genome into three types of regions based on DNase I hypersensitivity signals, namely DNase I peaks, DNase I non-peak hotspots, and DNase I insensitive regions. We then performed segmented block sampling [3]to get a background distribution of base overlap ratios for random regions given the distribution of binding peaks in the two datasets. Block sampling was applied instead of sampling individual base positions in order to capture the dependency of adjacent positions. The segmentation component of the sampling procedure was to take into account the fact that TF binding sites are not uniformly distributed in the whole genome, but rather highly correlated with DNase signals.

It was proved that if the segmented genome is piecewise stationary, the sampled base overlap ratios will be normally-distributed[3]. A z-score can then be derived as a measure of significance of the observed base overlap ratio. We noticed that in some of our cases the sampled base overlap ratios were non-Gaussian. To quantify the normality of the sampled values, we computed the skewness (defined as the third moment about the mean divided by the cube of the standard deviation) and robust kurtosis of each distribution. Centrality is indicated by small absolute values of the two measures.

To analyze the results, for each cell line we first selected one representative dataset for each TF based on the normality measures. The z-scores of all pairs of selected datasets were put into a matrix. We then performed a two-way average-link hierarchical clustering using the correlation of two rows/columns as similarity measure.

We repeated the whole process described above three times, respectively for binding peaks in the whole genome, binding peaks in promoter regions only (defined as regions within 2000bp of an annotated transcription start site in Gencode version 7 level 1 and level 2 annotations), and binding peaks in intergenic regions only (defined as regions at least 10000bp from any annotated gene in Gencode 7).

The final heatmap visualization was based on the clustering for the whole-genome case. A full list of co-associations is provided in Supplementary Table 1, section F. Images for the final figure and

heatmaps for the other cell lines analysed are provided in the file figures/TFCoassociation_matrices_v6.pdf within the code bundle.

Detailed steps:

1. TF binding peaks, blacklist regions, DNase peaks and hotspots, and Gencode annotation files were downloaded from the web sites listed above.

2. For each TF binding experiment, the binding peaks within blacklist regions were filtered. The remaining binding peaks were put into three lists: all of them, only those within 2,000bp from a promoter, and only those at least 10,000bp away from genes.

3. For each pair of TF binding experiments, the base-overlap ratio between their binding peaks were calculated for the three lists separately.

4. At the same time, for each cell line, the whole human genome not within blacklist regions was divided into three types of regions: regions with DNase peaks, regions with non-peak DNase hotspts, and all other regions. The regions of each type were collected to form three artificial chromosomes.

5. For each TF binding experiments, the binding peaks were re-numbered according to the three artificial chromosomes. Steps 2-5 were all performed using the Java program CoassociationAnalyzer.java.

6. To evaluate the statistical significance of the base overlap ratios, we then called the program block_bootstrap.py for each pair of TF experiments from the same cell line three times: once for the TF binding peaks in the whole genome, once for those in the promoter-proximal regions, and once for those in the gene-distal regions. The scripts gsc_bm_r0.1_n10000_all_dnasesegment_jobs, gsc_bm_r0.1_n10000_promoter_dnasesegment_jobs and gsc_bm_r0.1_n10000_distal_dnasesegment_jobs were used to call the block bootstrap program for all these cases, based on the renumbered TF binding peaks generated in step 5.

7. The block bootstrap program outputted the base overlap ratios of the sampled regions and the final z-score and p-values. We collected all these values and computed the skewness and robust kurtosis values for each pair of TF binding experiments using the program GSCCoassociationAnalyzer.java.

8. The program outputted three matrices for each of the three types of regions: GSC z-scores, skewness, and robust kurtosis values. We put them all in an Excel file and performed the following: a) added annotation information of the experiments, b) sorted the data, c) separated data for different cell lines, d) selected datasets that passed quality and reproducibility tests, e) for each cell line, selected one dataset for each TF.

9. We then performed two-way hierarchical clustering of the z-score matrices using the HCE software (www.cs.umd.edu/hcil/hce/). No additional normalization was performed, and the clusters were produced by average-linkage (UPGMA-type) clustering based on Euclidean distance. Nodes were ordered by keeping right child small.

10. The resulting order of experiments were then imported back to the Excel file, and the z-score, skewness and robust kurtosis matrices were all ordered accordingly.

11. Finally, heatmaps were generated based on these matrices using the VBA scripts in the Excel file.

Location of Code Bundle:

The software for performing segmented block sampling can be downloaded at http://www.encodestatistics.org/releases/block_bootstrap-0.8.1.zip

A code bundle is provided at
ftp.ebi.ac.uk:pub/databases/ensembl/encode/supplementary/Fig4_code_bundle.tar.gz

Datasets used:

Blacklist regions:
http://www.ebi.ac.uk/~anshul/public/encodeRawData/blacklists/wgEncodeHg19ConsensusSignalArtifactRegions.bed.gz

TF binding peaks:
ftp://ftp.ebi.ac.uk:pub/databases/ensembl/encode/supplementary/integration_data_jan2011/byDataType/peaks/jan2011/peakSeq/optimal/

Gencode version 7: ftp://ftp.sanger.ac.uk/pub/gencode/release_7/

DNase peaks and hotspots:
**http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwDnase/

Section G: ENCODE Genome Segmentations

Main Analysts:

Ian Dunham, Steve Wilder, Michael Hoffman, Jason Ernst, Bob Harris.

Principally Related to:

Figure 5

Methods:

A. UCSC Browser shot of three ENCODE genome segmentations for data from cell line Gm12878, Segway 25 state, ChromHHM 25 state and consensus 7 state, along with the uniform processed signal data that was the input to the segmentations. BigWig or wig tracks were uploaded for display in the UCSC browser of hg19. Required tracks were loaded into the UCSC Genome Browser (Preview version http://genome-preview.ucsc.edu/) using the tracklist at http://www.ebi.ac.uk/~dunham/ENCODE/Fig5A.tracklist.txt i.e. use URL

http://genome-preview.ucsc.edu/cgi-bin/hgTracks?org=human&position=chr22&hgt.customText=http://www.ebi.ac.uk/~dunham/ENCODE/Fig5A.tracklist.txt

The Figure was made using the preview browser at genome-preview.ucsc.edu.
The tracklist file is also in the Fig5_code_bundle/browser directory of the code bundle. The configuration of the browser was also saved to the session file Fig5_browser in the code bundle Fig5_code_bundle/browser directory and can also be used to upload the session. PDF was exported from the browser and edited in Adobe Illustrator for clarity and aesthetics.

B. The combined segmentation in bed format was compared for overlap with either long and short RNA seq IDR filtered elements or SPP IDR filtered peak calls for the appropriate cell line using segway_compare.pl (in Fig5_code_bundle/bin).

e.g.
run_segway_compare.pl -rna -segfile Gm12878.segmentation.bed GM12878_*.bed

or
run_segway_compare.pl -segfile chromhmm.segway.gm12878.comb11.concord4.mne.bed spp.optimal.wgEncode*Gm12878*narrowPeak

A directory of csv format output files is processed to amalgamate results and calculate observed/expected overlap for either element counts or genome coverage based on the genomic sizes of the respective segments using element_csv2matrix.pl (in Fig5_code_bundle/bin) N.B. this requires from Encode_modules.tar.gz.

e.g.
element_csv2matrix.pl csv_overlapdirectory/*.csv > matrix.R

This outputs an R data matrix that can be processed to the heatmap figure using either RNA_heatmap.cmds.R or TF_heatmap.cmds.R in the bin directory. Redundant tracks were removed for CTCF and POLR2A.

C. For each cell line combined segmentation, the segmentation bed file was separated out into individual states types according to Fig5_code_bundle/bin/panelC.cmds in the code bundle. For each state type, states from each of the 6 cell lines were overlapped using the saturate program found in Fig5_code_bundle/bin/saturate, to give a bit string of states at each genomic region

(commands in Fig5_code_bundle/bin/panelC.cmds). Note that saturate assumes elements are on the same strand in this version. Bit strings are then separated on a per cell basis using segment_variability.pl (in Fig5_code_bundle/bin) to give files of state overlaps for each cell line for each state. R commands in Fig5_code_bundle/bin/panelC.R will then plot the number cells each state location is found in the same state per cell line and mean counts across all the cells for each state. Intermediate data files are in Fig5_code_bundle/data/panelC and figure output in Fig5_code_bundle/figures/.

D. The combined segmentation in bed format was compared to the RRBS methylation data for overlap using segway_methylation.pl found in the bin directory of the code bundle. segway_methylation.pl requires the co-occur.pl script also in the Fig5_code_bundle/bin directory. A copy of the RRBS data for the Tier 1 and 2 cell lines is found in the Fig5_code_bundle/data/panelD/ directory. The output is an R data matrix object of the percent methylation score associated with each segmentation state across the whole genome. The R matrix objects are also given in the code bundle in Fig5_code_bundle/data/panelD/. Violin plots are used to display the distribution of methylation scores for each state, using the cmds in Fig5_code_bundle/bin/panelD.R. Output for the figure is given in Fig5_code_bundle/figures/.

Location of Code Bundle:

ftp.ebi.ac.uk:pub/databases/ensembl/encode/supplementary/Fig5_code_bundle.tar.gz
ftp.ebi.ac.uk:pub/databases/ensembl/encode/supplementary/Encode_modules.tar.gz

Datasets used:

Panel A. Use http://genome-preview.ucsc.edu/cgi-bin/hgTracks?org=human&position=chr22&hgt.customText=http://www.ebi.ac.uk/~dunham/ENCODE/Fig5A.tracklist.txt or the tracklist file in Fig5_code_bundle/browser

Panel B. Segmentation files can be found at ftp://ftp.ebi.ac.uk:pub/databases/ensembl/encode/supplementary/integration_data_jan2011/ in the directories in byDataType/segmentations/jan2011/.

IDR filtered RNA elements are located in ftp://ftp.ebi.ac.uk:pub/databases/ensembl/encode/supplementary/integration_data_jan2011/ byDataType/rna_elements/jan2011/LongRnaSeq/idrFilt and byDataType/rna_elements/jan2011/ShortRnaSeq/idrFilt

SPP peak calls are found in ftp://ftp.ebi.ac.uk:pub/databases/ensembl/encode/supplementary/integration_data_jan2011/ byDataType/ byDataType/peaks/jan2011/spp/optimal

Panel C. Segmentation files can be found at ftp://ftp.ebi.ac.uk:pub/databases/ensembl/encode/supplementary/integration_data_jan2011/ in the directories in byDataType/segmentations/jan2011/.

Panel D. Segmentation files can be found at ftp://ftp.ebi.ac.uk:pub/databases/ensembl/encode/supplementary/integration_data_jan2011/ in the directories in byDataType/segmentations/jan2011/. A summary RRBS data file in the correct format for the analysis is included in the data directories as above.

Section H: Fine-grained integrative analysis of large-scale segmentation using Self-Organizing Maps.

Main Analysts:

Ali Mortazavi @ UC Irvine, Shirley Pepke @ Caltech

Principally Related to:

Figure 7.

Methods:

See Mortazavi *et al.* Integrating ChIP-seq and DNAse-seq ENCODE data from multiple cell types using Self-Organizing Maps (manuscript in preparation).

Location of Code Bundle:
The SOM code and all of the files generated in this analysis can be found here:
http://woldlab.caltech.edu/~alim/ENCODESOM/ENCODESOM.bundle.current.tgz

This URL will move to a more permanent URL by the end of 2011. The file is 967 Mb.

Also available at
ftp.ebi.ac.uk:pub/databases/ensembl/encode/supplementary/ENCODESOM.bundle.current.tgz

Datasets used:

The segmentation used is the round8 chromHMM "stacked" segmentation on ENCODE Tier 1 and Tier 2 data found at

http://www.broadinstitute.org/~jernst/ROUND8_ChromHMM/fourcol_ChromHMM_stacked_25.bed.gz

Signal densities were calculated over those segments using the same processed bedgraphs that were used as for the large-scale segmentation, the ENCODE gencode+CAGE TSS file, and the NIH GWA catalog.

Section I: Experimental testing of putative enhancers.

Main Groups:

Jay Gertz, Tim Reddy, Rick Myers

Len Pennachio and Laboratory, LBL

Jochen Wittbrodt and Laboratory, University of Heidelberg

Felix Schlesinger and Tom Gingeras, Cold Spring Harbor Laboratory

Principally Related to:

Figure 6

Methods:

Candidate Region Selection

The proposed fragments were picked in the following manner:

- They had a particular reason to be chosen (see below)
- They were repeat free (RepeatMasker)
- They did not overlap Exons (GenCode version 3c) or within 2Kb of a Gencode TSS. Intronic picks are allowed
- The centre of the pick was extended by 500bp
- The flanking 100bp was also required to be repeat free in which primer design could occur

The first criteria was variable as follows

- A Naive set, where a match to a TransFac motif hit, as defined by TransFac track on the genome browser.
- A segmentation set, where the overlap enhancer predictions from ChromHMM and Segway (corresponding to the E set of enhancers in the consolidated set) were. This input list can be found as part of the code bundle.
- A discriminative prediction set, where the overlapping discriminative methods from Stanford and Yale were intersected. This input list can be found as part of the code bundle.

We expected that the cloning pipelines in both the Mouse and Fish transgenic experiments would have a proportion of drop outs; we placed enough clones into the pipelines to ensure a reasonable number of tested constructs.

Due to the higher cost of mouse transgenic experiments, and the long experience of the mouse transgenic assay in terms of a low false positive rate on negative DNA, it was decided to only focus on the two test predictions in this system, *i.e.*, the Segmentation set and the Discriminative methods set. As expected, a proportion of cloning attempts failed, leaving 58 tested constructs, with around 10 embryos per construct. From previous studies the background rate in this assay is very low, at least below 5%.

For the Fish enhancers assay all the above sets were attempted for cloning. As expected, a proportion of cloning attempts failed, leaving 37 tested constructs, with around 100 injections per construct.

Mouse Transgenic Assays

Transgenic mouse embryos were generated by pronuclear injection. Embryos were collected at e11.5 and stained for β-galactosidase activity with 5-bromo-4-chloro-3-indolyl β-D-galactopyranoside (X-Gal) as previously described [4]. Only patterns that were observed in at least three different embryos resulting from independent transgenic integration events of the same construct were considered reproducible [5]. Expression patterns were classified according to X-Gal staining in broadly defined anatomical regions[5]. Detailed annotations and photos of embryos for the tested elements are available at http://enhancer.lbl.gov. All mouse work was performed in accordance with protocols reviewed and approved by the Lawrence Berkeley National Laboratory Animal Welfare and Research Committee.

Fish Transgenic Assays

Reporter construct:

We used an Hsp70 basal promoter driving eGFP in a cassette flanked by ISceI restriction sites for efficient integration at early stages of embryonic development. To detect enhancer activity, the Vector was opened by a XmnI digest and PCR fragments containing putative regulatory elements were TA cloned upstream of the HSP70 promoter. Resulting constructs were tested by restriction digest and injection grade DNA was prepared using Qiagen Midi preps.

Microinjection:

Injections were performed following the meganuclease approach as described previously[6,7]. The efficiency of the meganuclease mediated transgenesis results in uniform expression patterns and a low degree of mosaicism. This facilitates the effective detection of enhancer activity even in few cells. In brief, medaka embryos were microinjected at the one cell stage. The concentration of the reporter construct was at 10 ng/µl. DNA was diluted in 1x ISceI buffer, containing ISceI enzyme (NEB) at a concentration of 1U/µl. DNA/enzyme mix was kept on ice prior to microinjection. For each construct at least 120 embryos were injected. Constructs were reinjected if more than 50% of the embryos died before the end of gastrulation. Embryos were maintained in their injection tray and grown at 23 °C until hatching. Individual embryos were scored twice at day 4 and 5 after the injection.

The expression patterns of the injected embryos were monitored under an Olympus MVX10 Stereomicroscope with a Leica DC500 digital camera (Leica).

Scoring:

The reporter system drives basal expression in the lens. All embryos showing lens expression from stage 28 onwards (day 4 at 23°C) were scored as positively injected and are the basis for further analysis. An injection was considered successful if at least 20 lens positive embryos were obtained at day 4. Embryos were analyzed individually on two consecutive days and expression domains were noted and counted for each expression construct.

Significance Tests

For the fish scenario, we had tested a matched set of normals and so could do either a categorical comparisons (fisher's exact) or a continuous test of means (t-test) on the ratio of specific patterns found.

Fisher's Exact:

Fisher's Exact Test for Count Data

data: table(fish$CLASS, fish$Call)

p-value = 0.03884

alternative hypothesis: two.sided

T-test:

> t.test(fish[fish$CLASS == 'HMM',]$positive_pattern_ratio,fish[fish$CLASS == 'NAIVE',]$positive_pattern_ratio)

Welch Two Sample t-test

data: fish[fish$CLASS == "HMM",]$positive_pattern_ratio and fish[fish$CLASS == "NAIVE",]$positive_pattern_ratio

t = 1.9342, df = 18.739, p-value = 0.06833

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-0.02041418 0.51152926

sample estimates:

mean of x mean of y

0.5398547 0.2942971

> t.test(fish[fish$CLASS == 'DISCR',]$positive_pattern_ratio,fish[fish$CLASS == 'NAIVE',]$positive_pattern_ratio)

Welch Two Sample t-test

data: fish[fish$CLASS == "DISCR",]$positive_pattern_ratio and fish[fish$CLASS == "NAIVE",]$positive_pattern_ratio

t = 1.1038, df = 21.406, p-value = 0.2819

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-0.1007159 0.3291502

sample estimates:

mean of x mean of y

0.4085143 0.2942971

For mouse we used a binomial test of the observed numbers vs the upper bound of background rate of human sequence of 0.05:

For HMMs: Pvalue = < 2e-16

For Discriminative case: Pvalue = 0.0003

The fisher test for whether Discriminative picks are separate from HMM was close to significant, but not quite:

> fisher.test(table(mouse$CLASS,mouse$Call))

Fisher's Exact Test for Count Data

data: table(mouse$CLASS, mouse$Call)

p-value = 0.1643

alternative hypothesis: true odds ratio is not equal to 1

95 percent confidence interval:

0.726581 9.934512

sample estimates:

odds ratio

2.550075

RNA Enrichment over Enhancers (panel omitted after review).

Details of the methods can be found in the Readme.txt file in the eRNA.supplement. In brief the scripts in the supplement look for RNA elements (RNA-seq contigs and CAGE Tag Clusters) nearby a set of enhancer predictions. They then compute the aggregate pattern of the positioning of these elements

relative to the center of the enhancer prediction. This is done separately for intergenic and genic enhancers and both strands. As a negative control the enhancer positions are randomly shuffled in the non-repeat portion of the genome.

Contact

Chenghai Xue (xuec@cshl.edu)

Felix Schlesinger (schlesin@cshl.edu)

Gingeras Lab

(gingeras@cshl.edu)

1) Gencode V7 annotation

2) Segmentation Enhancer set

awk '$4 == "E"' chromhmm.segway.gm12878.comb11.concord4.bed > chromhmm.segway.gm12878.comb11.concord4.onlyE.bed

---

We are using RNA pooled from Whole Cell, Cytoplasmic and Nuclear Fractions.

---

cat wgEncodeCshlLongRnaSeqGm12878* > Gm12878.Pooled.contigs.bed

Location of Code/Data Bundle:

Enhancer assay code bundle is at:

ftp.ebi.ac.uk: pub/databases/ensembl/encode/supplementary/ Figure6_code_bundle.tar.gz

eRNA code bundle is located at

ftp.ebi.ac.uk: pub/databases/ensembl/encode/supplementary/eRNA.supplement.zip

Datasets used:

Gencode version 7: ftp://ftp.sanger.ac.uk/pub/gencode/release_7/

Segmentation (version 7):

Discriminative Enhancer Picks

TransFac track from UCSC

RepeatMasker track of UCSC

Section J: Allele Specific Information.

Main Analysts:

Robert Altshuler, (Tim Reddy)

Principally Related to:
Allele Specific Information and Figure 8.

Methods:
Reads from ENCODE ChIP-Seq assays on the GM12878 cell line were aligned using bowtie (http://bowtie-bio.sourceforge.net/index.shtml). Reads were first aligned to the EBV genome to filter out reads that aligned to EBV. The remaining reads were aligned to the maternal chromosomes, and separately to the paternal chromosomes of a personalized genome for GM12878 (http://sv.gersteinlab.org/NA12878_diploid/). Alignments were post-processed to identify reads aligning to only one of the parental haplotypes. Reads aligning to SNPs were required to have an exact match to the SNP sequence. The counts of maternal and paternal reads at each variant were reported in files in the SNPCov format.

In order to improve sensitivity the maternal and paternal read counts were aggregated across replicates, and then aggregated over regions of Gencode v7 gene bodies and ChromHMM segments.

For each gene body chromHMM segment the allele-specific bias ratio defined as (#paternal_reads + #maternal_reads) / #total_reads was calculated. Pairwise correlations between assays were evaluated by fitting linear models to the allele-specific bias ratio data using weighted least squares. Weights were calculated as the product of the total number of reads from each assay for a particular region. Regions with fewer than seven total reads in either assay were excluded.

Location of Code Bundle:

ftp.ebi.ac.uk:pub/databases/ensembl/encode/supplementary/Fig8_code_bundle.tar.gz

(copy from http://www.broadinstitute.org/~rca/ENCODE_2012/index.html)

Datasets used:

Gencode version 7: ftp://ftp.sanger.ac.uk/pub/gencode/release_7/

The following ENCODE Experiments were used (to locate data sets search with the accession at http://genome.ucsc.edu/cgi-bin/hgFileSearch?db=hg19):

```
wgEncodeEH000028 ChipSeq Broad GM12878 antibody=H3K4me3
wgEncodeEH000029 ChipSeq Broad GM12878 antibody=CTCF
wgEncodeEH000030 ChipSeq Broad GM12878 antibody=H3K27ac
wgEncodeEH000031 ChipSeq Broad GM12878 antibody=H3K27me3
wgEncodeEH000032 ChipSeq Broad GM12878 antibody=H3K36me3
wgEncodeEH000033 ChipSeq Broad GM12878 antibody=H3K4me1
wgEncodeEH000034 ChipSeq Broad GM12878 antibody=H3K4me2
wgEncodeEH000035 ChipSeq Broad GM12878 antibody=H3K9ac
wgEncodeEH000036 ChipSeq Broad GM12878 antibody=H4K20me1
wgEncodeEH000037 ChipSeq Broad GM12878 antibody=Input
wgEncodeEH000147 RnaSeq CSHL GM12878 localization=cytosol rnaExtract=longPolyA
wgEncodeEH000170 RnaSeq CSHL GM12878 localization=nucleus rnaExtract=longPolyA
wgEncodeEH000187 RnaSeq CSHL GM12878 localization=nucleus rnaExtract=longNonPolyA
wgEncodeEH000394 ChipSeq UW GM12878 antibody=CTCF
wgEncodeEH000395 ChipSeq UW GM12878 antibody=H3K4me3
wgEncodeEH000428 ChipSeq UW GM12878 antibody=H3K27me3
wgEncodeEH000445 ChipSeq UW GM12878 antibody=H3K36me3
wgEncodeEH000463 ChipSeq UW GM12878 antibody=Input
```

wgEncodeEH000492 DnaseSeq UW GM12878
wgEncodeEH000528 ChipSeq UT-A GM12878 antibody=Input
wgEncodeEH000532 ChipSeq UT-A GM12878 antibody=CTCF
wgEncodeEH000534 DnaseSeq Duke GM12878
wgEncodeEH000592 ChipSeq UT-A GM12878 antibody=Pol2
wgEncodeEH000625 ChipSeq Yale GM12878 antibody=Input control=std
wgEncodeEH000626 ChipSeq Yale GM12878 antibody=Pol2 control=std
wgEncodeEH000690 ChipSeq Stanford GM12878 antibody=NFKB control=IgG-rab treatment=TNFa
wgEncodeEH000706 ChipSeq Stanford GM12878 antibody=Input control=IgG-mus
wgEncodeEH000707 ChipSeq Stanford GM12878 antibody=NFKB control=IgG-rab
wgEncodeEH000708 ChipSeq Stanford GM12878 antibody=Pol2 control=IgG-mus
wgEncodeEH000749 ChipSeq Stanford GM12878 antibody=Rad21 control=IgG-rab
wgEncodeEH000771 ChipSeq Stanford GM12878 antibody=Input control=IgG-rab
wgEncodeEH001033 ChipSeq Broad GM12878 antibody=H2A.Z
wgEncodeEH001034 ChipSeq Broad GM12878 antibody=H3K79me2
wgEncodeEH001035 ChipSeq Broad GM12878 antibody=H3K9me3
wgEncodeEH001462 ChipSeq HudsonAlpha GM12878 antibody=GABP protocol=PCR2x
wgEncodeEH001463 ChipSeq HudsonAlpha GM12878 antibody=Pol2 protocol=PCR2x
wgEncodeEH001464 ChipSeq HudsonAlpha GM12878 antibody=SRF protocol=PCR2x
wgEncodeEH001465 ChipSeq HudsonAlpha GM12878 antibody=NRSF protocol=PCR2x
wgEncodeEH001467 ChipSeq HudsonAlpha GM12878 antibody=RevXlinkChromatin protocol=PCR1x
wgEncodeEH001468 ChipSeq HudsonAlpha GM12878 antibody=USF-1 protocol=PCR2x
wgEncodeEH001469 ChipSeq HudsonAlpha GM12878 antibody=RevXlinkChromatin protocol=PCR2x
wgEncodeEH001475 ChipSeq HudsonAlpha GM12878 antibody=POU2F2 protocol=PCR1x
wgEncodeEH001476 ChipSeq HudsonAlpha GM12878 antibody=PU.1 protocol=PCR1x
wgEncodeEH001477 ChipSeq HudsonAlpha GM12878 antibody=Pbx3 protocol=PCR1x
wgEncodeEH001478 ChipSeq HudsonAlpha GM12878 antibody=TAF1 protocol=PCR1x
wgEncodeEH001479 ChipSeq HudsonAlpha GM12878 antibody=BATF protocol=PCR1x
wgEncodeEH001480 ChipSeq HudsonAlpha GM12878 antibody=EBF1_(SC-137065) protocol=PCR1x
wgEncodeEH001484 ChipSeq HudsonAlpha GM12878 antibody=IRF4_(SC-6059) protocol=PCR1x
wgEncodeEH001485 ChipSeq HudsonAlpha GM12878 antibody=TCF12 protocol=PCR1x
wgEncodeEH001486 ChipSeq HudsonAlpha GM12878 antibody=BCL11A protocol=PCR1x
wgEncodeEH001487 ChipSeq HudsonAlpha GM12878 antibody=p300 protocol=PCR1x
wgEncodeEH001488 ChipSeq HudsonAlpha GM12878 antibody=ZBTB33 protocol=PCR1x
wgEncodeEH001489 ChipSeq HudsonAlpha GM12878 antibody=PAX5-C20 protocol=PCR1x
wgEncodeEH001495 ChipSeq HudsonAlpha GM12878 antibody=PAX5-N19 protocol=PCR1x
wgEncodeEH001496 ChipSeq HudsonAlpha GM12878 antibody=SP1 protocol=PCR1x
wgEncodeEH001517 ChipSeq HudsonAlpha GM12878 antibody=Pol2-4H8 protocol=PCR1x
wgEncodeEH001541 ChipSeq HudsonAlpha GM12878 antibody=RXRA protocol=PCR1x
wgEncodeEH001542 ChipSeq HudsonAlpha GM12878 antibody=SIX5 protocol=PCR1x
wgEncodeEH001562 ChipSeq HudsonAlpha GM12878 antibody=ATF3 protocol=PCR1x
wgEncodeEH001563 ChipSeq HudsonAlpha GM12878 antibody=BCLAF1_(SC-101388)
protocol=v041610.1
wgEncodeEH001564 ChipSeq HudsonAlpha GM12878 antibody=ETS1 protocol=PCR1x
wgEncodeEH001565 ChipSeq HudsonAlpha GM12878 antibody=MEF2A protocol=PCR1x
wgEncodeEH001617 ChipSeq HudsonAlpha GM12878 antibody=ELF1_(SC-631) protocol=v041610.1
wgEncodeEH001624 ChipSeq HudsonAlpha GM12878 antibody=SRF protocol=v041610.1
wgEncodeEH001632 ChipSeq HudsonAlpha GM12878 antibody=Egr-1 protocol=v041610.1
wgEncodeEH001634 ChipSeq HudsonAlpha GM12878 antibody=RevXlinkChromatin
protocol=v041610.1
wgEncodeEH001640 ChipSeq HudsonAlpha GM12878 antibody=Rad21 protocol=v041610.1
wgEncodeEH001645 ChipSeq HudsonAlpha GM12878 antibody=ZEB1_(SC-25388)
protocol=v041610.2
wgEncodeEH001647 ChipSeq HudsonAlpha GM12878 antibody=RevXlinkChromatin
protocol=v041610.2
wgEncodeEH001648 ChipSeq HudsonAlpha GM12878 antibody=MEF2C_(SC-13268)
protocol=v041610.1
wgEncodeEH001657 ChipSeq HudsonAlpha GM12878 antibody=YY1_(SC-281) protocol=PCR1x
wgEncodeEH001658 ChipSeq HudsonAlpha GM12878 antibody=BCL3 protocol=v041610.1
wgEncodeEH001756 ChipSeq USC GM12878 antibody=ZNF274 control=std
wgEncodeEH001787 ChipSeq Stanford GM12878 antibody=WHIP control=IgG-mus
wgEncodeEH001798 ChipSeq Stanford GM12878 antibody=TBP control=IgG-mus
wgEncodeEH001810 ChipSeq Stanford GM12878 antibody=RFX5_(200-401-194) control=IgG-mus
wgEncodeEH001812 ChipSeq Stanford GM12878 antibody=USF2 control=IgG-mus
wgEncodeEH001831 ChipSeq Stanford GM12878 antibody=CHD2_(AB68301) control=IgG-mus
wgEncodeEH001832 ChipSeq Stanford GM12878 antibody=EBF1_(SC-137065) control=std
wgEncodeEH001833 ChipSeq Stanford GM12878 antibody=SMC3_(ab9263) control=IgG-mus
wgEncodeEH001846 ChipSeq Stanford GM12878 antibody=Nrf1 control=IgG-mus
wgEncodeEH001851 ChipSeq Stanford GM12878 antibody=CTCF_(SC-15914) control=std
wgEncodeEH001853 ChipSeq Stanford GM12878 antibody=Znf143_(16618-1-AP) control=std
wgEncodeEH001858 ChipSeq Stanford GM12878 antibody=Pol2(phosphoS2) control=IgG-mus
wgEncodeEH002025 ChipSeq Stanford GM12878 antibody=BHLHE40_(NB100-1800) control=IgG-
mus
wgEncodeEH002026 ChipSeq Stanford GM12878 antibody=Mxi1_(AF4185) control=IgG-mus
wgEncodeEH002034 ChipSeq Stanford GM12878 antibody=Input control=IgG-rab
treatment=TNFa
wgEncodeEH002037 ChipSeq Stanford GM12878 antibody=p300_(SC-584) control=std

Section K: Examining ENCODE Elements on a per individual basis in the NA12878/GM12878 Genome.

Main Analysts:

Joel Rozowsky

Principally Related to:

Figure 9A & 9B

Methods:

In order to investigate the effect of using the true personal NA12878 diploid genome sequence as a reference for peak calling for the GM12878 ChIP-Seq datasets as apposed to using the GRCh37(hg19) reference sequence we performed the following analysis. Reads for GM12878 datasets were independently mapped using Bowtie [8] against the maternal and paternal haplotype sequences for NA12878[9] (http://alleleseq.gersteinlab.org/). The maternal and paternal haplotypes sequences for NA12878 were constructed using phased variants (SNPs, indels and deletions) from the pilot phase of the 1000 Genomes Project[10]. Only reads that map uniquely to one location on either allele were used for the following peak calling procedure (the same procedure that was used for reads mapped to the GRCh37 reference genome sequence). Reads mapping to each haplotype were then scored using PeakSeq[2] using the default parameters. The same IDR threshold that was applied to the peaks called using the reads mapped to the GRCh37 reference sequence was also applied to the haplotype specific ranked peak lists.

We then compared the peaks that were called using the maternal and paternal haplotypes as well as the peaks called the GRCh37 reference sequence. In Supplementary Figure 1, section K we plot the percentage of maternal or paternal specific peaks (i.e. peaks that are present using either only the maternal or paternal haplotypes but not when using the GRCh37 reference genome sequence) for all GM12878 transcription factors. We see on average about ~2% of peaks called are either maternal or paternal specific. An example of one of these is shown in Figure 9A. An additional example shown in Supplementary Figure 2, section K is of a peak that is present on the paternal allele (but not the maternal allele) for both POU2F2 and EBF. An interesting observation is in this example the region of the peak on chromosome 16 does not contain any sequence variant that differs between the maternal and paternal alleles for NA12878. The reason for the difference in peak calling is due to a maternal specific insertion on chromosome 1 (i.e. a maternal specific duplication of this region) which causes the reads on the maternal allele not to map uniquely on the maternal haplotype, causing the difference in peak calling. It is ambiguous whether this is due to a difference in binding at this location between the maternal and paternal alleles, but it does demonstrate some of the technical issues associated with peak-calling.

We have also computed various overlap statistics between maternal and paternal peaks, as well as intersections with the different variants types (SNPs, indels and deletions), which are included in the attached supplementary files. The code to perform the haplotype specific peak calling and related analysis are included as part of this supplement.

We also investigated the overlap of variants within the NA12878 genome versus annotations determined by ENCODE in order to see to what extent we can "annotate" the variants with respect to ENCODE annotation. We started with all the sequence variants (SNPs, indels and deletions) for NA12878 from the 1000 Genomes Project. We were then able to subdivide the variants into those that are rare (and correspondingly the complement which are common) using the low coverage sequencing of 179 individuals also done by the 1000 Genomes Project. Rare variants in NA12878 are defined as those not present in any of the 179 individuals. We computed overlap statistics for these variants (also partitioned by variants types and its homozygous vs

heterozygous classification) against the various GENCODE annotations (GENCODE v7): protein coding genes, pseudogenes and non-coding RNAs as well as for all the transcription factor binding sites determined using ChIP-Seq (we used the PeakSeq scored peaks but found equivalent results for the SPP scored peaks). In addition to performing these straightforward overlaps we also counted the number of variants that are likely to have a functional effect on the annotation it overlaps. For the cases of protein-coding genes likely functional variants are defined as those that would either cause a loss-of-function due to a premature stop, frame-shift or disruption of a splice site or those that would cause of a non-synonymous substitution. For binding sites functional variants are defined as those that overlap an identified motif within a transcription factor binding site.

We summarize the results of these overlap statistics in Supplementary Table 1, section K (which is encapsulated in Figure 9B – the published Figure 9B is a refined version that has been added to the code bundle as fig9b_v3.pdf). Supplementary files contain the overlap counts of variants with binding sites and motifs for each individual transcription factor. We also performed an equivalent element-centric version of these overlaps where instead of counting variants overlapping annotations, we count ENCODE annotations that overlap NA12878 sequence variants. The results of these element-centric variant overlap analyses are summarized in Supplementary Table 2, section K.

The overlap statistics of variants with respect to ENCODE annotations were computed using BEDTools[11]. We include as part of this supplement the code used to perform these overlaps.

Location of Code Bundle:

ftp.ebi.ac.uk:pub/databases/ensembl/encode/supplementary/
Supp_Material_Fig9_Data_Files.tar.gz

(from
http://homes.gersteinlab.org/people/rozowsky/ENCODE/Supp_Material_Fig11_Data_Files.tar.gz
)

Datasets used:

(http://encodewiki.ucsc.edu/EncodeDCC/index.php/Locations_of_ENCODE_Data#Post_Jan_2011_Freeze_data)

All peakcalls for TF ChIP-Seq datasets.

Gencode version 7: ftp://ftp.sanger.ac.uk/pub/gencode/release_7/

1000 Genomes Pilot Phase Variant Calls for NA12878 (attached in bundle).

Section L: Cancer mutations

Main Analysts:

Stephen C. J. Parker, Elliott H. Margulies, Ewan Birney, Ian Dunham

Principally Related to:

Somatic variants section; Figure 9C.

Methods:
We used DNaseI hypersensitive site (DHS) data sets for the following 34 different cell types: 8988t, A549, AosmcSerumfree, Chorion, Cll, Fibrobl, Gliobla, Gm12878, H1hesc, H9es, Helas3, Hepatocytes, Hepg2, Hmec, Hpde6e6e7, Hsmm, Htr8, Huh7, Huvec, Ips, K562, Lncap, Mcf7, Medullo, Melano, Myometr, Nhek, Osteobl, Panislets, Phte, Progfib, Stellate, T47d, Urotsa, which are listed in Supplementary Table 1, section L.

Single-linkage clustering was performed on all DHSs across the 34 cell lines to determine regions that are active in single, multiple, and all cell types. The DHS signature tree (Supplementary Figure 1, section L) was constructed by first creating a binary vector for each cell type that classifies a region as either present (1) or absent (0). Then, Euclidean distance was used as a metric to hierarchically cluster the binary vectors. Using these results, we identified all DHS regions that are cell-type specific and those that are ubiquitous (active in all 34 cell types).

We next divided the cell-type specific and ubiquitous DHSs into mutually exclusive categories based, in order, on the following genic landmark overlaps: coding regions, 5' UTR's, 3' UTR's, introns, intergenic transcription start site (TSS)-proximal (within 5,000 bp of a TSS), and intergenic TSS-distal (greater than 5,000 bp from a TSS). All genic landmarks are based on the GENCODE V7 annotation in hg19 and can be downloaded from the UCSC Genome Browser [genome.ucsc.edu].

We used the Genome Structure Correction (GSC) method[3] to calculate enrichment statistics for somatic single nucleotide variants (SSNVs) relative to different DHS sets. Supplementary Figure 2, section L shows the results of this analysis for each cancer type and each set of non-genic TSS-distal cell-specific and ubiquitous DHSs. We focused our analyses on non-genic TSS-distal DHS sets to minimize confounding with transcription-coupled repair. Notably, somatic variants accumulate significantly less than expected in ubiquitous non-genic TSS-distal DHSs for three of the four cancer sets. Melanoma SSNVs are significantly depleted in melanocyte-specific non-genic TSS-distal DHSs. Similarly, for one pancreatic cancer set, SSNVs are less likely to occur in pancreatic islet-specific non-genic TSS-distal DHSs, though this did not reach statistical significance for alpha = 0.05.

By orienting intronic mutations relative to the transcribed strand, we can observe asymmetry in the mutation repair process (Supplementary Figure 3, section L). This bias is different for ENCODE-annotated DHSs relative to bulk introns (compare left and right panels in Supplementary Figure 2, section L), suggesting a change in the underlying repair mechanism.

Location of Code Bundle:

ftp.ebi.ac.uk:pub/databases/ensembl/encode/supplementary/encode-cancer-analysis-code-bundle.tar.gz

(from http://zoo.nhgri.nih.gov/parker/encode/supplement/encode-cancer-analysis-code-bundle.tar.gz)

Datasets used:

There are two principle sets of data used for this analysis:

1. Whole genome single nucleotide somatic cancer mutations from ICGC [http://www.icgc.org/] for the following cancer types:
    a. Melanoma[12]
    b. Pancreatic cancer (two different samples):
        i. ftp://data.dcc.icgc.org/version_6/Pancreatic_Cancer-QCMG-AU/
        ii. ftp://data.dcc.icgc.org/version_6/Pancreatic_Cancer-OICR-CA/
    c. Small cell lung cancer[13]
2. DHSs from the DCC accessions listed in Supplementary Table 1, section L.

Gencode version 7: ftp://ftp.sanger.ac.uk/pub/gencode/release_7/

Section M: Common variation associated with human disease and phenotypes.

Main Analysts:

Belinda Giardine, Weisheng Wu, Robert S. Harris, Ross C. Hardison

Principally Related to:

Figure 10 panels a, b, c in the main text.

Supplementary Figures 1and 2, section M

Supplementary Tables 1 and 2, section M: GWAS SNPs overlapping (a) TF occupied segments and (b) DNase peaks, respectively.

Methods, Location of code bundle, and Datasets used

This information is provided for each of the analyses comprising this section of the paper.

A. GWAS catalog data file

Methods:

1. Download file from http://www.genome.gov/gwastudies/. This has over 6000 SNP-phenotype associations.

2. Parse file to keep unique positions in chromosomes 1-22 and X, sort by chromosome and start position.

3. Collapse redundant SNP-phenotype associations. These redundancies had different p-values initially and came from different studies, but p-values were removed, allowing the collapsing.

4. From that file, we grouped together SNPs with highly related phenotypes. This introduced some additional redundancies, which were then collapsed. This generates a file with 4492 SNPs involved in 4860 SNP phenotype associations. The file name is gwascatalog.june2011.phenoSimple.colapsedPheno2.filteredChr.bed

Code:

```
cat gwascatalog.june_16_2011.join.txt | perl -ne 'chomp; split(/\t/); if ($_[0] !~ /Y|M|G|rand|hap|Un/) { $_[0] =~ s/chrX/chr99/; print "$_[0]\t$_[1]\t$_[2]\n";} ' | sort -k1.4,1.5n -k2,2n | uniq | perl -ne 'chomp; s/chr99/chrX/; print "$_\n"; ' > ~/encode/cache/gwas_catalog.june_2011.bed
```

Datasets used:

GWAS data gwascatalog.june_16_2011.txt originally from http://www.genome.gov/admin/gwascatalog.txt

Can be downloaded from ftp://ftp.ebi.ac.uk:pub/databases/ensembl/encode/supplementary/integration_data_jan2011/ in directory byDatatype/GWAS/jan2011/

B. Sets of genotyping SNPs matched to GWAS SNPs ("Stam null set")

Methods:

1. SNPs from the Illumina 1M array were matched with the GWAS SNPs based on the CEU frequency, the distance from the nearest TSS, and the genomic location (exon|intron|utr|intergenic). This information was compiled for each of the GWAS SNPs and each of the SNPs in the Illumina 1M array using Galaxy tools. The exact tools and steps can be seen in the published history, ENCODE GWAS null set.

2. From the matched sets for all the GWAS SNPs, 1000 random samples were chosen. (3 SNPs had no matches, plus some of the GWAS SNPs had no frequency information, these were dropped.)

Files in Code Bundle:

Code bundle is at
ftp.ebi.ac.uk:pub/databases/ensembl/encode/supplementary/Fig10abcCode.tar.gz

Galaxy (published history, ENCODE GWAS null set)

Also used Galaxy (used to prep input files)

Datasets used:

Gwascatalog.june2011.positions.bed.gz

Can be downloaded from
ftp://ftp.ebi.ac.uk:pub/databases/ensembl/encode/supplementary/integration_data_jan2011/
in directory byDatatype/GWAS/jan2011/

snpArrayIllumnina1M.sortChr.merged.bed.gz

available at
ftp://ftp.ebi.ac.uk:pub/databases/ensembl/encode/supplementary/integration_data_jan2011/
in directory byDatatype/GWAS/jan2011/

UCSC: HapMap SNPs track

Gencode Genes V7: ftp://ftp.sanger.ac.uk/pub/gencode/release_7/

C. Figure 10a (intersections)

Methods:

Basic intersections based on position are done just keeping the counts of the number of intersections and total size of each SNP set (bed file).

Files in Code Bundle:

bed_intersect.py (bx.python module, https://bitbucket.org/james_taylor/bx-python/wiki/Home)

intersections.pl in Code bundle at
ftp.ebi.ac.uk:pub/databases/ensembl/encode/supplementary/Fig10abcCode.tar.gz

Datasets used:

Gwascatalog.june_16_2011.txt

snpArrayIllumnina1M.sortChr.merged.bed.gz

pgSnpsCombined24.hg19.noRand.noY.sortChr.bed.gz

low_coverage.2010_07.hg19.sorted.bed.gz

gwasNullSet.tar

all available at
ftp://ftp.ebi.ac.uk:pub/databases/ensembl/encode/supplementary/integration_data_jan2011/
in directory byDatatype/GWAS/jan2011/

D. Supplementary Figure 2, section M (enrichments)

Methods:

1. We start with the results of the combined segmentations from Steve Wilder and Ian Dunham, which are a combination of chromHMM and Segway results, which themselves were trained primarily on histone modification data. Four basic features (Stats) were computed for each of the segments: segment name, number of segments, number of bases in segments, and the average number of bases per segment.

2. Then overlap is computed for each of the SNP sets with each of the segments.

3. The counts of the overlaps and the stats of the segments are used to compute the enrichment. This is % of the SNP set overlapping divided by the % of the total segments this segment class makes up.

4. Then the log2 is taken of the result and graphed. For the matched samples a box plot showing the median, 95% range and outliers are drawn.

5. An enrichment (or depletion) is considered significant if the values for < 50 samples fall outside of GWAS value; this corresponds to an empirical p-value threshold of 0.05.

Files in Code Bundle:

files:drive_NFSvS_E in for_this_directory.r, feature_to_title, intervals_to_intersecting_bases.py, makeStatsFile.pl, accumSamples.pl, run.sh, runSample2.sh, runSamples.sh, stamNullSet.pl

in Code bundle at
ftp.ebi.ac.uk:pub/databases/ensembl/encode/supplementary/Fig10abcCode.tar.gz

Datasets used:

gwascatalog.june_16_2011.txt

snpArrayIllumnina1M.sortChr.merged.bed.gz

pgSnpsCombined24.hg19.noRand.noY.sortChr.bed.gz

low_coverage.2010_07.hg19.sorted.bed.gz

gwasNullSet.tar.gz

All available from
ftp://ftp.ebi.ac.uk:pub/databases/ensembl/encode/supplementary/integration_data_jan2011/
in directory byDatatype/GWAS/jan2011/

ENCODE combined segmentations files can be found at
ftp://ftp.ebi.ac.uk:pub/databases/ensembl/encode/supplementary/integration_data_jan2011/
in the directories in byDataType/segmentations/jan2011/

E. Figure 10b (generating tables)

Methods:

1. As presented in section A, the GWAS SNPs were filtered to retain only those on chrs 1-22 and X, and then were collapsed to get a set unique on position and phenotype. Then some similar phenotypes were collapsed, further reducing the number of rows.

2. The TF files were processed from the single linkage clustering data to have 1 file per factor.

3. The GWAS SNPs were then intersected with the singleLinkage TF and the DNase peaks (from Bob Thurman, containing the DNase peaks from both the UW and the Open Chromatin Group). This made 2 tables.

4. We combined the SNPs associated with identical phenotypes, and calculated the numbers of SNPs in each phenotype (shown in the second column), the numbers of SNPs in a phenotype that overlap with at least one TF occupied segment from any ChIP-seq measurement (shown in the third column), the numbers of SNPs in all phenotypes that overlap with the TF occupied segments from individual ChIP-seq measurements (shown in the second row), and the numbers of SNPs in each phenotype that overlap with the TF occupied segments from individual ChIP-seq measurements (shown in the remaining matrix). A subset of the data was chosen to contain only the phenotypes with at least five SNPs that overlap with TF occupied segments, and only the ChIP-seq measurements that overlap with at least 20 SNPs in all phenotypes. Hierarchical clustering was performed on both rows and columns, and the table was reordered based on the clustering. The phenotype-TF interaction cells were colored green if the empirical p-value of the enrichment is less than 0.01 and the counts are more than two. The upper left corner of this table is shown in the left matrix of Fig. 10B. The same counts and coloring were performed for the phenotype-DHS interactions, and some relevant DHS were selected to be shown in the right matrix of Fig. 10B.

Relevant files in Code Bundle:

files: minPheno.pl, gwasForClustering.pl, bed_intersect.py, splitTfFromSlc.pl

in Code bundle at
ftp.ebi.ac.uk:pub/databases/ensembl/encode/supplementary/Fig10abcCode.tar.gz

Datasets used:

gwascatalog.june_16_2011.txt available from
ftp://ftp.ebi.ac.uk:pub/databases/ensembl/encode/supplementary/integration_data_jan2011/
in directory byDatatype/GWAS/jan2011/

Gwascatalog.june2011.phenotypes.bed available from
ftp://ftp.ebi.ac.uk:pub/databases/ensembl/encode/supplementary/integration_data_jan2011/
in directory byDatatype/GWAS/jan2011/

Single Linkage Cluster (slc files) are downloadable from
ftp://ftp.ebi.ac.uk:pub/databases/ensembl/encode/supplementary/integration_data_jan2011/
in the directories in byDataType/slc/jan2011

Combined UW/Duke FDR 1% peaks from
http://www.uwencode.org/public/rthurman/encode_integrated_open_chrom/data/UW_hot.tgz

F. Partitioned graph (Supplementary Figure 1, section M)

Methods:

1. The genome was divided to non-overlapping 500kb windows.

2. Windows with more than 90% Ns were discarded.

3. The coverage of TFs, DHS hotspots, Gencode exons was computed on these windows.

4. For each coverage the windows were binned in 5 groups of equal numbers of windows from the lowest to the highest coverage.

5. These bins were then intersected with the SNP sets.

6. The percent of total SNPs in each bin was then graphed.

7. In another graph only the SNPs that intersected TF sites were intersected with the bins, giving the percent of SNPs in the bin that intersect a binding site.

Code Bundle:

Galaxy (Published workflows: Partition genome into 5 bins based on coverage, Intersect annotation with 5 partitions(bins))

Datasets used:

gwascatalog.june_16_2011.txt

snpArrayIllumnina1M.sortChr.merged.bed.gz

pgSnpsCombined24.hg19.noRand.noY.sortChr.bed.gz

low_coverage.2010_07.hg19.sorted.bed.gz

gwasNullSet.tar

all from
ftp://ftp.ebi.ac.uk:pub/databases/ensembl/encode/supplementary/integration_data_jan2011/
in directory byDatatype/GWAS/jan2011/

Combined UW/Duke FDR 1% peaks from
http://www.uwencode.org/public/rthurman/encode_integrated_open_chrom/data/UW_Duke_fdr0.01_pks.tgz

500kb Windows

Gencode v7 exons (UCSC browser track)

G. Empirical p-values for association of GWAS SNPs with TF occupied segments and DNase peaks

Methods:

G1. Using the non-GWAS genotyping SNPs matched to the GWAS SNPS

As described in Section B, we generated a set of genotyping SNPs not associated with phenotypes but matched to the GWAS SNPS for allele frequency, distance from the transcription start site and genic location (or intergenic). For 1000 random samplings from this set, each SNP was assigned to a phenotype such that the resulting dataset had the same number of SNPs pseudo-associated with a phenotype as found in the observed GWAS SNP set. Then the SNPs in each of the 1000 pseudo-GWAS SNP-phenotype sets were examined for overlap with TF occupancy and DNase peak files (as in Section E). The results were 1000 tables of number of overlaps between pseudo-SNPs for each phenotype and each TF-cell combination or the DNase peaks in each cell type. Thus a given number of overlaps seen between GWAS SNPs for a phenotype could be compared with the frequency of the number of overlaps between pseudo-SNPs for each phenotype and each TF-cell combination, or overlaps of pseudo-SNPs for each phenotype with DNase peaks.

For the 105 SNPs associated with Crohn's disease, we found 5 and 3 that overlapped DNA segments occupied by GATA2 and cFOS, respectively, in Huvec cells. In the 1000 rounds of analysis of pseudo-SNPs associated with Crohn's disease, we found the following distributions of overlap frequencies:

number of overlaps	number of times that number of overlaps was seen	
	GATA2	cFOS
0	567	454
1	323	351
2	90	147
3	19	38
4	1	8
5	0	2

Thus, the empirical p-value is <0.001 for overlaps with GATA2 and 0.048 (i.e. (38+8+2)/1000) for overlaps with cFOS.

The same approach was used to estimate a p-value for the overall level of overlap between GWAS SNPs and TF occupied DNA segments. For all the GWAS SNPs, we found 600 overlaps with TF occupied DNA segments. For the 1000 samplings of pseudo-GWAS SNPs, the number of overlaps ranged between 352 and 547. Since none of the pseudo-GWAS SNP sets had as many overlaps as for the observed GWAS SNPs, we empirically estimate the p-value as <0.001.

G2. Estimating p-values by permutation

Since the real number of SNPs that are linked to Crohn's disease is 105, we did a simple permutation by randomly selecting 105 SNPs from the entire 4,860 GWAS SNPs and treated them as pseudo Crohn's disease-associated SNPs. From each permutation, we calculated how many SNPs out of the chosen 105 overlap with HuvecGata2 or HuvecCfos. In 1000 tests of the GWAS SNPs with phenotype labels randomly permuted, we observed 3 sets with at least 5 SNPs overlapping HuvecGata2 (the level seen for the real Crohn's disease SNPs. Thus, for this approach, the empirical p-value is 3/1000 or 0.003. A similar test for the overlaps of GWAS SNPs (phenotype label permuted) found at least 3 SNPs overlapping HuvecCfos to occur 23 out of 1000 samplings, for an empirical p-value of 0.023.

Location of Code Bundle:

ftp.ebi.ac.uk:pub/databases/ensembl/encode/supplementary/Fig10abcCode.tar.gz

Datasets used:

GWAS_DHSpeaks_all.xlsx, GWAS_TFandDHSpeaks_all.xlsx, GWAS_TFandDHSpeaks_all_subset.xlsx
all downloadable from
ftp://ftp.ebi.ac.uk:pub/databases/ensembl/encode/supplementary/integration_data_jan2011/
in directory byDatatype/GWAS/jan2011/

Section N: Classification of ENCODE TFs (Table 1 and Supplementary Table 1, section N)

Main Analysts:

Ian Dunham

Principally Related to:

Table 1

Supplementary Table 1, section N

Methods:

TFs were classified by manual annotation from the literature primarily by Peggy Farnham and Mike Pazin and the classification stored in the Factorbook (www.factorbook.org) metadata table at
https://spreadsheets.google.com/spreadsheet/ccc?key=0AiUFPJFcN7XldF9iTTk5T1JWbVBoTS10Y3VxR3Q5QVE&authkey=CLOqnDY&hl=en_US#gid=1

The set of TF used from the ENCODE January 2011 data freeze after removing poor quality datasets is documented at
https://docs.google.com/spreadsheet/ccc?key=0Am6FxqAtrFDwdFQ4YTh0TmtCOTBvVDk4dHlkSWlnLXc#gid=0

tsv dumps of both tables are included in the code bundle and can be combined together with paper_TF_stats.pl to give the output tables which are also present in the code bundle.

Location of Code Bundle:

ftp.ebi.ac.uk: pub/databases/ensembl/encode/supplementary/Table1_code_bundle.tar.gz

Note that the paper_TF_stats.pl script requires modules from

ftp.ebi.ac.uk: pub/databases/ensembl/encode/supplementary/Encode_modules.tar.gz

Datasets used:

ENCODE TF ChIP-seq data.

Section O: No O because it looks like 0.

Section P: Summary of ENCODE Data production.

Main Analysts:

Ian Dunham

Principally Related to:

Supplementary Table 1, section P

Methods:

This table aims to summarise the numbers of experiments and files generated by the ENCODE project and available to download. Files are available to download from either the "public" site at http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC, or from the "test" site at http://hgdownload-test.cse.ucsc.edu/goldenPath/hg19/encodeDCC. Public site files have been through the DCC quality control process in particular for consistency of metadata, etc. However there is more data on the test data. One can also look at NCBI36 (hg18) genome build data, although since July 2011 almost all data has been migrated to GRCh37 (hg19).

On these download sites, the various ENCODE data files are organised into directories according to the data submitter and the type of data. Within each directory is a file (files.txt) that lists the files, and the metadata tags that belong with the file. For more details of metadata tags see http://genome.ucsc.edu/ENCODE/otherTerms.html. Metatags give the cell, data type, file type, antibody and so on, and can be interpreted via the DCC controlled vocabulary.

Within ENCODE there is the concept of an experiment which comprises several datasets that are linked. For instance in a ChIP-seq experiment there will be two or more experimental replicate datasets and at least one input control datasets. Experiments are grouped under a single accession number, the DCC Accession. Each experiment can have multiple files available for download e.g. the sequence reads, their mappings, peak calls and so on. Experiments may subsequently be analysed in an integrated fashion along with other experiments.

ENCODE has had a series of data freezes. The freeze of interest for this paper is the January 2011 Freeze. Again the datasets have a metatag associated with them that records the data freeze.

This table was generated by downloading the file lists from each ENCODE data download directory (in this case from the test site on hg19), and then parsing the metatags to interpret the types of data, and to group data files by experiment (DCC Accession). The output is a count of files and experiments split by various metatag categories including data freeze, method, etc. including csv format data that can be converted directly into Table 2, section S in excel, or equivalent. At the same time a csv file is generated that can be used as a lookup table for all experiment files (encode_data.csv). Various parsings are done within the code to consolidate terms, and standardise output (for instance normalising antibody names to the HGNC gene name for the protein detected by the antibody). Data output includes data for the Jan 2011 data freeze and the current state of the site at the date of running. Supplementary Table 1, section P is the result of running on the 21 September 2011 on the test site (hg19).

The script encode_data.pl in the bin directory of the code bundle does the download and parsing. You will need to install various modules from CPAN plus the ENCODE modules as described below. If no options are given it will run on http://hgdownload-test.cse.ucsc.edu/goldenPath/hg19/encodeDCC by default. To run on released datasets, use the command 'encode_data.pl –site public'.

Location of Code Bundle:

ftp.ebi.ac.uk:pub/databases/ensembl/encode/supplementary/TableP1_code_bundle.tar.gz

You will also need the modules from Encode_modules.tar.gz in the same location:

ftp.ebi.ac.uk:pub/databases/ensembl/encode/supplementary/Encode_modules.tar.gz

Datasets used:

All ENCODE data from http://hgdownload-test.cse.ucsc.edu/goldenPath/hg19/encodeDCC as of 21 September 2011.

Section Q: ENCODE element counts and element lengths for major data types.

Main Analysts:

Ian Dunham

Principally Related to:

Supplementary Figure 1, section Q

Methods:

For each of the major datasets the overall length and count of elements per cell line is determined from the appropriate Bed file. The files are the uniformly processed and IDR's elements, or FDR filtered, and are located in the data structure at ftp://ftp.ebi.ac.uk:pub/databases/ensembl/encode/supplementary/integration_data_jan2011/ in the directories in byDataType/. Files to be processed are specified by directory in the bin/directories file in the code bundle. Element count is achieved by counting the number of lines in the bed file. Element length is determined from the bed file coordinates. The assumption is made that elements are in half-open format and non-overlapping, other than for RNA types where the elements are stranded and can overlap on opposite strands. Coverage for RNA types should thus be considered on a genome of two strands. No attempt has been made to cluster RNA types in this analysis, although that is done in other analyses.

Run element_matrix.pl –dir directories

to output element_matrix.tab and length_matrix.tab tab separated files. These were imported into excel to create Supplementary Table 1, section Q. The tsv format files can also be converted to R matrices for dispal using bin/matrix.pl.

Location of Code Bundle:

ftp.ebi.ac.uk:pub/databases/ensembl/encode/supplementary/TableQ1_code_bundle.tar.gz

element_matrix.pl requires perl modules in ftp.ebi.ac.uk:pub/databases/ensembl/encode/supplementary/Encode_modules.tar.gz

and bedlengths.pm in the bin directory of the code bundle.

Datasets used:

ftp://ftp.ebi.ac.uk:pub/databases/ensembl/encode/supplementary/integration_data_jan2011/

in the directories

byDataType/rna_elements/jan2011/ShortRnaSeq/idrFilt # short rna seq elements idr

byDataType/rna_elements/jan2011/LongRnaSeq/idrFilt #long rna seq elements idr

byDataType/peaks/jan2011/spp/optimal # Chip-seq elements idr spp

byDataType/peaks/jan2011/histone_macs/conservative/ # histone mod peak calls idr (conservative)

byDataType/openchrom/jan2011/labPeaks # open chrom FDR$ filtered peaks (from R Thurman) includes FAIRE

Section R: Saturation Analysis

Main Analysts:

Steven Wilder, Ian Dunham

Principally Related to:

"Summary of ENCODE elements"

Methods:

Saturation code was written in C++. The unions of experimental replicates (from multiple groups, for instance) peak calls are first sorted, then overlapping regions are joined to form elements of maximum size 5000 bp, and the cell type coverage of the element was compared to 1,000 (CTCF) or 20,000 (DNase1) precalculated randomly generated cell type combinations for each coverage value. Hence the distribution of number of unique elements for any number of cell types is approximated. This distribution was modelled using a Weibull distribution, and hence interpolated. The fit is robust to different thresholds of element calling (and restriction to only primary tissue samples.

Supplementary Figure 1, section R CTCF with Weibull Fit

Supplementary Figure 2, section R DNase with Weibull fit

Location of Code Bundle:

ftp.ebi.ac.uk:pub/databases/ensembl/encode/supplementary/Saturation.tar.gz

Datasets used:
Lists of the files used are at
encode-box-01@fasp.encode.ebi.ac.uk:byDataType/slc/jan2011/ or
ftp://ftp.ebi.ac.uk:pub/databases/ensembl/encode/supplementary/integration_data_jan2011/byDataType/slc/jan2011

Dnase/combined_peaks.list
Ctcf/filelist

Dnase files from
encode-box-01@fasp.encode.ebi.ac.uk:byFreeze/jan2011/openchrom/combined_peaks

SPP "Optimal" IDR CTCF Peaks from
encode-box-01@fasp.encode.ebi.ac.uk:byDataType/peaks/jan2011/spp/optimal

Section S: Characterization of transcription factor ChIP-peaks peaks with only low affinity recognition sequences

Main Analysts:

Ben Brown, Pouya Kheradpour

Principally Related to:

Text in section "*Regions bound by transcription factors, transcriptional machinery, and other proteins*"

Methods:

Identification of ChIP peaks with only low affinity motif instances
For each factor, we considered both known and discovered motifs. For the details our motif instance identification and scoring pipeline, including the motif discovery pipeline, see Kheradpour and Kellis, manuscript in preparation.

We defined "moderate to high affinity recognition sequences" as those with scores > 0.25, which corresponds to a frequency of random occurrence in the genome less than, on average, once per kilobase of sequence. All peaks without moderate to high affinity recognition sequences we defined as "peaks with only low affinity recognition sequences". However, because we compare only general trends between the bottom of the peak affinity rank list and the remaining majority of peaks, we found that the particular threshold (i.e. 0.25) did not materially effect our results.

Overlap statistics and significance estimation

Wilcoxon rank sum statistics were computed with MATLAB v7.10.0 function "ranksum.m".

Genome Structure Correction statistics were computed using the GSC statistical package available at encodestatistics.org. Because the current implementation of the code does not explicitly support two-sample tests, we computed the standard deviation of the overlap statistics under the null using the command line arguments: -r 0.1 -s 0.1 -t rm -n 10000. The -r and -s command line options were chosen based on the stability criterion [3]. The two sample z-score was then formulated in the usual way.

All overlaps between sets of peaks and/or bed files were computed using bedtools version 2.13.0[11], which can be downloaded at: http://code.google.com/p/bedtools/.

Datasets used:

ENCODE's GENCODE and CAGE merged promoter set:

ftp://genome.crg.es/pub/Encode/data_analysis/TSS/Gencodev7_CAGE_TSS_clusters_June2011.gff.gz

ENCODE's predicted and known motif instances in ChIP peaks:

ftp://encodeftp.cse.ucsc.edu/users/benbrown/

Section T: Element and coverage calculation

Main Analysts:

Steven Wilder, Ian Dunham

Principally Related to:

"Summary of ENCODE elements"

Methods:

Using the saturate program, all the relevant sets of input files were sorted and non-redundant overlaps were calculated.

The countdistsgaps.pl program was subsequently used to calculate the distribution of distances to the nearest biochemical event, adjusting for genome gaps.

Location of Code Bundle:

ftp.ebi.ac.uk:pub/databases/ensembl/encode/supplementary/saturatedists.tar.gz

Datasets used:
ftp://ftp.ebi.ac.uk:pub/databases/ensembl/encode/supplementary/integration_data_jan2011/

in the directories

byDataType/rna_elements/jan2011/ShortRnaSeq/idrFilt # short rna seq elements idr

byDataType/rna_elements/jan2011/LongRnaSeq/contigsWithAtLeast5reads #long rna seq elements

byDataType/peaks/jan2011/spp/optimal # Chip-seq elements idr spp

byDataType/peaks/jan2011/histone_macs/conservative/ # histone mod peak calls idr (conservative)

byDataType/openchrom/jan2011/labPeaks # open chrom FDR$ filtered peaks (from R Thurman) includes FAIRE

byDataType/gencode/jan2011/gencode.v7.exons.bed #GENCODE version7 exons

Section U: Gencode

See Supplementary Table 1, section U for additional Gencode annotation and refs [14,15].

Gencode Annotations can be downloaded from

ftp://ftp.sanger.ac.uk/pub/gencode/

Gencode version 7 is the default for the ENCODE intergrated analysis and can be downloaded from: ftp://ftp.sanger.ac.uk/pub/gencode/release_7/ or from the UCSC browser at http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeGencodeV7/. Ac opy of the same files exists at ftp://ftp.ebi.ac.uk:pub/databases/ensembl/encode/supplementary/integration_data_jan2011/ in the directories byDataType/gencode/jan2011.

Section V: Pseudogenes

Main Analysts:

Baikang Pei, Mark Gerstein.

Principally Related to:

Pseudogene section

Methods:

Details of GENCODE pseudogene annotation and their related genomic features are discussed in ref [16].

GENCODE pseudogene files can be found at http://Pseudogene.org/

Section W: Agreement in cell type similarities across assays

Main Analysts:

Steven Wilder, Ian Dunham, Ewan Birney

Principally Related to:

"Summary of ENCODE elements"

Methods:

Ten cell types were identified as having the most complete and comparable data sets for CTCF ChIP-seq, DNase-seq, PolyA- whole cell RNA-seq and PolyA+ whole cell RNA-seq; the cell types were:
GM12878, K562, H1-hESC, HeLa-S3, HepG2, HUVEC, AG04450, BJ, NHEK, and SK-N-SH_RA.

The overlaps of the peaks and RNA elements from the cell types were calculated using the overlap program (see Supplementary_Info-Saturation), to form elements of up to 5000 bp.

The Jaccard distance between two cell types for an assay was calculated using presence/absence in these elements. Within each assay, the pairwise cell type distances were ranked, and Kendall's coefficient of concordance was used to calculate the significance of the agreement in ranks across assays ($W = 0.614$, $p = 2.7\times10^{-7}$).

Location of Code Bundle:

ftp.ebi.ac.uk:pub/databases/ensembl/encode/supplementary/supplementary/Assay_similarity.Rhistory.gz

Datasets used:
ftp://ftp.ebi.ac.uk:pub/databases/ensembl/encode/supplementary/integration_data_jan2011/

in the directory

byDataType/slc/jan2011/Assay_similarity

The original data is stored in the directories

byDataType/peaks/jan2011/spp/optimal # Chip-seq elements idr spp

byDataType/openchrom/jan2011/labPeaks # open chrom FDR$ filtered peaks (from R Thurman)

byDataType/rna_elements/jan2011/LongRnaSeq/contigsWithAtLeast5reads # long rna seq elements

Section X: Establishment of Uniform Processing Pipeline for TF ChIP-seq data

Main Analysts:

Anshul Kundaje.

Principally Related to:

General peak calling

Methods

Peak Caller Comparison

There are several peak callers that are widely used. For practical considerations, the AWG decided to converge on two peak callers to be used in parallel for all analysis. We used the irreproducible discovery rate (IDR) framework to evaluate the reproducibility characteristics a set of popular peak callers namely

Peakseq (http://www.gersteinlab.org/proj/PeakSeq/),

SPP (http://compbio.med.harvard.edu/Supplements/ChIP-seq/),

MACS (http://liulab.dfci.harvard.edu/MACS/),

Fseq (http://fureylab.web.unc.edu/software/fseq/),

Hotspot (http://www.uwencode.org/proj/hotspot-ptih/),

Erange (http://woldlab.caltech.edu/rnaseq),

cisgenome (http://www.biostat.jhsph.edu/~hji/cisgenome/),

QuEST (http://mendel.stanford.edu/SidowLab/downloads/quest/index.html)

and SISSRS (http://sissrs.rajajothi.com/).

We decided to select two peak callers that showed the best IDR-based rank consistency between independent peak call lists for replicate ChIP-seq datasets for CTCF and Pol2 as well as replicate DNase-seq datasets.

We observed that PeakSeq[2], SPP[17] and MACS[18] outperformed most other peak callers on the basis of both the number of reproducible peaks called at a defined IDR threshold and the proportion of identified peaks that overlapped with predicted binding regions based on mapping of the appropriate TF position weight matrix. Although the motif hits data does not serve as a ground truth, it serves as an independent source for a crude evaluation of the accuracy of peak callers. It was decided to select PeakSeq and SPP for all further analysis. NOTE: QuEST[19], erange [20]and cisgenome[21] results are not as bad as they seem. They use extremely stringent default peak calling thresholds so the peak lists never really reach the inconsistent component and makes it impossible for the IDR model to fit the data effectively. It is hence, important to use relaxed peak calling thresholds. The results are better when more relaxed thresholds are used for these peak callers. However, SPP, PeakSeq and MACS continue to outperform them.

Uniform peak calling pipeline

The SPP and PeakSeq peak caller were established in two independent processing pipelines. All TF ChIP-seq datasets were paired with their corresponding 'Control' datasets based on UCSC DCC metadata. Details on metadata matching and code for matching datasets to control can be found https://docs.google.com/document/edit?id=1aJNyHvdqPrm_Uoz2uIXdX_hCp0aClHJ_DX2-Dm76Pgg&hl=en. Peaks were called on all TF ChIP-seq replicate datasets using a relaxed FDR threshold of 0.7. Control replicates were pooled together but TF ChIP-seq replicates are NOT pooled at this step. We refer to these as ReplicatePeakCall. Aligned reads from all replicates for a particular TF ChIP-seq experiment are now pooled. Peaks were called on the pooled ChIP-seq data wrt. pooled Control data. Once again an FDR threshold of 0.7 was used. We refer to these as PooledPeakCall. Then for each unique TF ChIP-seq dataset, we have one PooledPeakCall file and two or more ReplicatePeakCall files (depending on the number of replicates).

Given a set of peak calls for a pair of replicate datasets, we can rank each set of peaks based on some criterion of significance, such as the p-value, q-value or ChIP to input enrichment or read coverage for each peak. We expect true peaks to have high significance, be reproducible i.e. exist in both replicates and be rank consistent i.e. be placed similarly in the two ranked lists. The IDR statistic formally quantifies this notion. The IDR analysis considers all pairs of matched peaks to be sampled from one of two populations – a consistent population where rank consistency is maintained and an inconsistent population where rank consistency breaks down. The method goes down the pair of ranked lists to identify the rank at which the consistency begins to break down. Since it is based on a probabilistic model, each peak (pair) can be assigned a probability that it belongs to the inconsistent population. This is the local IDR. We would like to select peaks with low IDR values (analogous to peaks with low p-values or q-values). We can then set a reasonable IDR threshold that can be used to select the number of peaks that are statistically consistent. We can then call peaks on data pooled over all replicates, rank the peaks and use the IDR selected threshold to select a final set of confident, consistent peaks.

A valuable advantage of the IDR statistic is that one can select a single confidence threshold over datasets of varying quality. False discovery rate (FDR) thresholds that are typically used for peak calling need significant fine tuning to extract the optimal amount of signal from datasets of varying quality. FDR thresholds can also be quite unstable i.e. small changes in the FDR threshold can result in large changes in the number of peaks selected. The IDR thresholds on the other hand tend to show smoother behavior and so selection of an optimal threshold does not require as much fine tuning. Also, if ChIP-datasets are compared against inherently correlated (aggressive) input datasets, a significant FDR threshold will tend to call fewer peaks. The IDR threshold automatically adjusts to data quality and doesn't simply restrict to peaks with strong enrichment over input. It can select weaker peaks that are highly reproducible and rank consistent. More details on the IDR analysis can be found elsewhere[22].

We perform IDR/consistency analysis on all pairs of ReplicatePeakCall files that correspond to each PooledPeakCall file. Consistency is evaluated in terms of the rank of the peaks and their reproducibility. A copula mixture model is fitted to pairs of replicates. For each pairwise comparison of ReplicatePeakCall files, we obtain the number of peaks that pass an IDR threshold 0.02. We refer to this as PairwiseNumPeakCutoff. For each PooledPeakCall file, we obtain the largest of all the corresponding PairwiseNumPeakCutoff thresholds (i.e. the threshold based on the most consistent pair of replicates). We refer to this as MaxPairwiseNumPeakCutoff. We use this threshold to trim the PooledPeakCall file i.e. we keep the top N peaks where N = MaxPairwiseNumPeakCutoff.

If MaxPairwiseNumPeakCutoff < 100 for a particular PooledPeakCall file, it generally means that the dataset has very low signal to noise enrichment. This happens is a few cases (~6 datasets). In such cases, the IDR based threshold can be too conservative. In order to squeeze the most signal out of these datasets, we opt to select a threshold on signal enrichment. We only keep peaks where the signal score > 25. This is based on the observation that for a large fraction of 'good' datasets, the IDR based threshold tends to be equivalent to a signal score of ~25.

For datasets with no replicate data, we use the signal score cutoff of 25 to trim the peak call file. There are NN such datasets.

Blacklist filtering

After IDR thresholding, we further filter peaks that lie in signal artefact regions. A set of blacklists for Grch37/hg19 is provided here
ftp://encodeftp.cse.ucsc.edu/users/akundaje/rawdata/blacklists/hg19/

Or

**http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeMapability/wgEncodeDacMapabilityConsensusExcludable.bed.gz

Caveats of uniform peak calling pipeline

There are a few subtleties when working with the IDR statistic.

1. Selection of ranking measure: The IDR model works better in the absence of too many ties in ranks. So the measure of significance used to rank peaks should be as continuous as possible. Some peak callers like SPP provide 'blocky' q-value estimates. Signal enrichment is a better measure for ranking. One should try different ranking measures if there is no obvious choice. However, if a set of peaks are genuine ties (not an artefact of the scoring measure used), it is better to break the ties randomly than to use some ad-hoc procedure to deterministically break the ties. This has sound theoretical justification as well. Breaking 'real' ties randomly can be proved to maintain concordance between ranks.

2. A dataset that shows poor rank consistency could be due to a sparsely binding transcription factor or poor data quality. This is difficult to infer directly from the IDR analysis. One can defer to strand cross-correlation analysis and other prior biological knowledge to dig deeper into the causes for the poor consistency. Furthermore, the IDR analysis is unable to identify which replicate within an inconsistent pair is responsible for the poor results. Strand cross-correlation and other quality measures can be used to identify the bad replicates

3. The current IDR model is unable to directly leverage information from > 2 replicates. All replicates are analyzed in pairs and the most consistent set is used to learn thresholds that are then applied to all peak calls on pooled data from all replicates.

4. It is VERY important to use a relaxed threshold when calling peaks on the individual replicates. The IDR model assumes the existence of two populations – a consistent and an inconsistent one. Hence, the peak calls must contain a reasonable fraction of false peaks for the model to learn parameters appropriately. The exact peak caller significance threshold is immaterial. The IDR analysis is immune to the initial thresholds used for peak calling provided they are relaxed. Typically try to use ~150k to 300k peaks as input into the IDR pipeline.

Pooling data from multiple replicates typically increases the confidence of peak calling and hence the discovery power. Since the cutoffs are learned based on pairwise analysis of replicates, it is optimal to use a slightly relaxed IDR threshold. An IDR threshold of 0.02-0.03 works well in practice as an optimal tradeoff between sensitivity and specificity on pooled data.

Useful resources for Uniform Peak Calling pipeline

Matching datasets to controls:
ftp://encodeftp.cse.ucsc.edu/users/akundaje/fuzzyMatchMetadata/

Original SPP package (Park Lab Harvard): http://compbio.med.harvard.edu/Supplements/ChIP-seq/

Modified SPP code: ftp://encodeftp.cse.ucsc.edu/users/akundaje/phantomPeakQuality/

PeakSeq:

http://www.gersteinlab.org/proj/PeakSeq/

IDR Software: Kundaje *et al.* manuscript in preparation.

Section Y: ChIA-PET

ChIA-PET libraries were constructed as previously described (Li et al., manuscript in review; Fullwood et al., Nature, 2009). For RNAPII, Monoclonal antibody 8WG16 (Covance, MMS-126R) was used. For CTCF, 07-729 (Millipore) was used. RNAPII libraries were constructed in MCF7, HCT116, HeLa, K562 and NB4. CTCF libraries were constructed in K562, GM12878 and MCF7 cell lines. Libraries were analyzed as previously described[23]{Li, 2012 #343. The libraries are available at the GIS ChIA-PET web site (http://chiapet.gis.a-star.edu.sg; username: "encode", password "human"), and also the ENCODE data center at UCSC web site (**http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeGisChiaPet/). See also Supplementary Figure 1, section Y.

Section Z: Discriminative Training Methods

We developed two methods that involve discriminative learning for identifying potential enhancers in the human genome {Yip, 2012 #411}. Briefly, both methods learned models that primarily used chromatin features (histone marks and open-chromatin) to discriminate between a positive set of potential enhancer sites consisting of ChIP-seq peak locations of a limited set of likely enhancer binding TFs and a negative set consisting of different types of control genomic locations. The trained enhancer models were then used to scan and score sliding windows across the entire genome to obtain high-confidence enhancer predictions. Specifically, in the first method a machine-learning procedure was used to identify binding active regions (BARs) and potential promoters in the K562 cell line. BARs that are close to an annotated transcription start site (TSS), overlap a coding exon or have a high promoter score were discarded. Among the remaining regions, those with a binding motif of a transcription factor (TF) expressed in K562 were selected as the first list of candidate enhancer regions. The second method involved a two-stage machine-learning procedure that uses chromatin features alongside sequence conservation and proximity to annotated TSSs. In the first stage, broad high-scoring regions were predicted in K562 using mainly signal magnitude features of chromatin marks. In the second stage, a model was trained to discriminate TF binding peaks from flanking regions using shapes of chromatin mark signals. The second-stage model was used to refine the precision of predictions obtained from the first stage. Finally, only predictions that involved the use of H3K4me1 or H3K4me3 features were retained. All regions that did not overlap coding exons and were distant from annotated TSSs were selected as the second list of candidate enhancer regions. The two lists of candidates were then intersected and size-adjusted to satisfy experimental requirements. Since our models were not trained on bona fide enhancers, our predictions may capture enhancers as well as other types of regulatory elements. So to obtain a final set of high confidence enhancer predictions we filtered the intersected list to only retain predictions with strong H3K4me1 or moderate H3K4me3 signals. This gave us a final list of 13,539 potential enhancer regions for testing.

Section AA: References for Papers Using ENCODE Consortium Data.

Collected publications that have been identified as utilising ENCODE data, from outside the Consortium, as of the beginning of October 2011. References from within the consortium are collected at http://encodeproject.org/ENCODE/pubs.html. Since tracking of data sets in the wild is problematic this is almost certainly an incomplete list.

1. Adrianto I, Wen F, Templeton A, Wiley G, King JB, Lessard CJ, Bates JS, Hu Y, Kelly JA, Kaufman KM *et al*: **Association of a functional variant downstream of TNFAIP3 with systemic lupus erythematosus**. *Nat Genet* 2011, **43**(3):253-258.

2. Bau D, Sanyal A, Lajoie BR, Capriotti E, Byron M, Lawrence JB, Dekker J, Marti-Renom MA: **The three-dimensional folding of the alpha-globin gene domain reveals formation of chromatin globules**. *Nat Struct Mol Biol* 2011, **18**(1):107-114.

3. Bell JT, Pai AA, Pickrell JK, Gaffney DJ, Pique-Regi R, Degner JF, Gilad Y, Pritchard JK: **DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines**. *Genome biology* 2011, **12**(1):R10.

4. Bhattacharyya S, Tian J, Bouhassira EE, Locker J: **Systematic targeted integration to study Albumin gene control elements**. *PLoS One* 2011, **6**(8):e23234.

5. Boeva V, Surdez D, Guillon N, Tirode F, Fejes AP, Delattre O, Barillot E: **De novo motif identification improves the accuracy of predicting transcription factor binding sites in ChIP-Seq data analysis**. *Nucleic acids research* 2010, **38**(11):e126.

6. Bohlig L, Friedrich M, Engeland K: **p53 activates the PANK1/miRNA-107 gene leading to downregulation of CDK6 and p130 cell cycle proteins**. *Nucleic acids research* 2011, **39**(2):440-453.

7. Calva D, Dahdaleh FS, Woodfield G, Weigel RJ, Carr JC, Chinnathambi S, Howe JR: **Discovery of SMAD4 promoters, transcription factor binding sites and deletions in juvenile polyposis patients**. *Nucleic acids research* 2011, **39**(13):5369-5378.

8. Carstensen L, Sandelin A, Winther O, Hansen NR: **Multivariate Hawkes process models of the occurrence of regulatory elements**. *BMC Bioinformatics* 2010, **11**:456.

9. Chang HW, Cheng YH, Chuang LY, Yang CH: **SNP-RFLPing 2: an updated and integrated PCR-RFLP tool for SNP genotyping**. *BMC Bioinformatics* 2010, **11**:173.

10. Dong X, Navratilova P, Fredman D, Drivenes O, Becker TS, Lenhard B: **Exonic remnants of whole-genome duplication reveal cis-regulatory function of coding exons**. *Nucleic acids research* 2010, **38**(4):1071-1085.

11. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M *et al*: **Mapping and analysis of chromatin state dynamics in nine human cell types**. *Nature* 2011, **473**(7345):43-49.

12. Evans KJ: **Most transcription factor binding sites are in a few mosaic classes of the human genome**. *BMC Genomics* 2010, **11**:286.

13. Farrell JJ, Sherva RM, Chen ZY, Luo HY, Chu BF, Ha SY, Li CK, Lee AC, Li RC, Yuen HL *et al*: **A 3-bp deletion in the HBS1L-MYB intergenic region on chromosome 6q23 is associated with HbF expression**. *Blood* 2011, **117**(18):4935-4945.

14. Ferraiuolo MA, Rousseau M, Miyamoto C, Shenker S, Wang XQ, Nadler M, Blanchette M, Dostie J: **The three-dimensional architecture of Hox cluster silencing**. *Nucleic acids research* 2010, **38**(21):7472-7484.

15. Firpi HA, Ucar D, Tan K: **Discover regulatory DNA elements using chromatin signatures and artificial neural network**. *Bioinformatics* 2010, **26**(13):1579-1586.

16. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Chen Y, Clapham P, Coates G, Fairley S, Fitzgerald S *et al*: **Ensembl 2011**. *Nucleic acids research* 2011, **39**(Database issue):D800-806.

17. Gaulton KJ, Nammo T, Pasquali L, Simon JM, Giresi PG, Fogarty MP, Panhuis TM, Mieczkowski P, Secchi A, Bosco D *et al*: **A map of open chromatin in human pancreatic islets**. *Nat Genet* 2010, **42**(3):255-259.

18. Ghedira K, Hornischer K, Konovalova T, Jenhani AZ, Benkahla A, Kel A: **Identification of key mechanisms controlling gene expression in Leishmania infected macrophages using genome-wide promoter analysis**. *Infect Genet Evol* 2011, **11**(4):769-777.

19. Gheldof N, Smith EM, Tabuchi TM, Koch CM, Dunham I, Stamatoyannopoulos JA, Dekker J: **Cell-type-specific long-range looping interactions identify distant regulatory elements of the CFTR gene**. *Nucleic acids research* 2010, **38**(13):4325-4336.

20. Gotea V, Visel A, Westlund JM, Nobrega MA, Pennacchio LA, Ovcharenko I: **Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers**. *Genome Res* 2010, **20**(5):565-577.

21. Guillou E, Ibarra A, Coulon V, Casado-Vela J, Rico D, Casal I, Schwob E, Losada A, Mendez J: **Cohesin organizes chromatin loops at DNA replication factories**. *Genes Dev* 2010, **24**(24):2812-2822.

22. Handstad T, Rye MB, Drablos F, Saetrom P: **A ChIP-Seq benchmark shows that sequence conservation mainly improves detection of strong transcription factor binding sites**. *PLoS One* 2011, **6**(4):e18430.

23. Harismendy O, Notani D, Song X, Rahim NG, Tanasa B, Heintzman N, Ren B, Fu XD, Topol EJ, Rosenfeld MG *et al*: **9p21 DNA variants associated with coronary artery disease impair interferon-gamma signalling response**. *Nature* 2011, **470**(7333):264-268.

24. Higareda-Almaraz JC, Enriquez-Gasca Mdel R, Hernandez-Ortiz M, Resendis-Antonio O, Encarnacion-Guevara S: **Proteomic patterns of cervical cancer cell lines, a network perspective**. *BMC Syst Biol* 2011, **5**:96.

25. Ho ES, Gunderson SI: **Long conserved fragments upstream of Mammalian polyadenylation sites**. *Genome Biol Evol* 2011, **3**:654-666.

26. Hubisz MJ, Lin MF, Kellis M, Siepel A: **Error and error mitigation in low-coverage genome assemblies**. *PLoS One* 2011, **6**(2):e17034.

27. Karnani N, Taylor CM, Malhotra A, Dutta A: **Genomic study of replication initiation in human chromosomes reveals the influence of transcription regulation and chromatin structure on origin selection**. *Mol Biol Cell* 2010, **21**(3):393-404.

28. Kehayova P, Monahan K, Chen W, Maniatis T: **Regulatory elements required for the activation and repression of the protocadherin-{alpha} gene cluster**. *Proc Natl Acad Sci U S A* 2011, **108**(41):17195-17200.

29. Kratz A, Arner E, Saito R, Kubosaki A, Kawai J, Suzuki H, Carninci P, Arakawa T, Tomita M, Hayashizaki Y *et al*: **Core promoter structure and genomic context reflect histone 3 lysine 9 acetylation patterns**. *BMC Genomics* 2010, **11**:257.

30. Lamina C, Coassin S, Illig T, Kronenberg F: **Look beyond one's own nose: Combination of information from publicly available sources reveals an association of GATA4 polymorphisms with plasma triglycerides**. *Atherosclerosis* 2011.

31. Lessard CJ, Adrianto I, Kelly JA, Kaufman KM, Grundahl KM, Adler A, Williams AH, Gallant CJ, Anaya JM, Bae SC *et al*: **Identification of a systemic lupus erythematosus susceptibility locus at 11p13 between PDHX and CD44 in a multiethnic study**. *Am J Hum Genet* 2011, **88**(1):83-91.

32. Martin MM, Ryan M, Kim R, Zakas AL, Fu H, Lin CM, Reinhold WC, Davis SR, Bilke S, Liu H *et al*: **Genome-wide depletion of replication initiation events in highly transcribed regions**. *Genome Res* 2011.

33. Martin P, Barton A, Eyre S: **ASSIMILATOR: a new tool to inform selection of associated genetic variants for functional studies**. *Bioinformatics* 2011, **27**(1):144-146.

34. McLeay RC, Leat CJ, Bailey TL: **Tissue-specific prediction of directly regulated genes**. *Bioinformatics* 2011, **27**(17):2354-2360.

35. Mikula M, Gaj P, Dzwonek K, Rubel T, Karczmarski J, Paziewska A, Dzwonek A, Bragoszewski P, Dadlez M, Ostrowski J: **Comprehensive analysis of the palindromic motif TCTCGCGAGA: a regulatory element of the HNRNPK promoter**. *DNA Res* 2010, **17**(4):245-260.

36. Mokry M, Hatzis P, Schuijers J, Lansu N, Ruzius FP, Clevers H, Cuppen E: **Integrated genome-wide analysis of transcription factor occupancy, RNA polymerase II binding and steady-state RNA levels identify differentially regulated functional gene classes**. *Nucleic acids research* 2011.

37. Mudge JM, Frankish A, Fernandez-Banet J, Alioto T, Derrien T, Howald C, Reymond A, Guigo R, Hubbard T, Harrow J: **The origins, evolution, and functional potential of alternative splicing in vertebrates**. *Mol Biol Evol* 2011, **28**(10):2949-2959.

38. Myers RM, Stamatoyannopoulos J, Snyder M, Dunham I, Hardison RC, Bernstein BE, Gingeras TR, Kent WJ, Birney E, Wold B *et al*: **A user's guide to the encyclopedia of DNA elements (ENCODE)**. *PLoS Biol* 2011, **9**(4):e1001046.

39. Naumova N, Dekker J: **Integrating one-dimensional and three-dimensional maps of genomes**. *J Cell Sci* 2010, **123**(Pt 12):1979-1988.

40. Pickrell JK, Gaffney DJ, Gilad Y, Pritchard JK: **False positive peaks in ChIP-seq and other sequencing-based functional assays caused by unannotated high copy number regions**. *Bioinformatics* 2011, **27**(15):2144-2146.

41. Pique-Regi R, Degner JF, Pai AA, Gaffney DJ, Gilad Y, Pritchard JK: **Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data**. *Genome Res* 2011, **21**(3):447-455.

42. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A: **Detection of nonneutral substitution rates on mammalian phylogenies**. *Genome Res* 2010, **20**(1):110-121.

43. Raj B, O'Hanlon D, Vessey JP, Pan Q, Ray D, Buckley NJ, Miller FD, Blencowe BJ: **Cross-regulation between an alternative splicing activator and a transcription repressor controls neurogenesis**. *Mol Cell* 2011, **43**(5):843-850.

44. Ramagopalan SV, Heger A, Berlanga AJ, Maugeri NJ, Lincoln MR, Burrell A, Handunnetthi L, Handel AE, Disanto G, Orton SM *et al*: **A ChIP-seq defined genome-wide map of vitamin D receptor binding: associations with disease and evolution**. *Genome Res* 2010, **20**(10):1352-1360.

45. Rushefski M, Aplenc R, Meyer N, Li M, Feng R, Lanken PN, Gallop R, Bellamy S, Localio AR, Feinstein SI *et al*: **Novel variants in the PRDX6 Gene and the risk of Acute Lung Injury following major trauma**. *BMC Med Genet* 2011, **12**:77.

46. Sadri J, Diallo AB, Blanchette M: **Predicting site-specific human selective pressure using evolutionary signatures**. *Bioinformatics* 2011, **27**(13):i266-274.

47. Sankaran VG, Menne TF, Scepanovic D, Vergilio JA, Ji P, Kim J, Thiru P, Orkin SH, Lander ES, Lodish HF: **MicroRNA-15a and -16-1 act via MYB to elevate fetal hemoglobin expression in human trisomy 13**. *Proc Natl Acad Sci U S A* 2011, **108**(4):1519-1524.

48. Su J, Teichmann SA, Down TA: **Assessing computational methods of cis-regulatory module prediction**. *PLoS Comput Biol* 2010, **6**(12):e1001020.

49. Taft RJ, Hawkins PG, Mattick JS, Morris KV: **The relationship between transcription initiation RNAs and CCCTC-binding factor (CTCF) localization**. *Epigenetics Chromatin* 2011, **4**:13.

50. Teng L, Firpi HA, Tan K: **Enhancers in embryonic stem cells are enriched for transposable elements and genetic variations associated with cancers**. *Nucleic acids research* 2011, **39**(17):7371-7379.

51. Terrenoire E, McRonald F, Halsall JA, Page P, Illingworth RS, Taylor AM, Davison V, O'Neill LP, Turner BM: **Immunostaining of modified histones defines high-level features of the human metaphase epigenome**. *Genome biology* 2010, **11**(11):R110.

52. Valen E, Sandelin A, Winther O, Krogh A: **Discovery of regulatory elements is improved by a discriminatory approach**. *PLoS Comput Biol* 2009, **5**(11):e1000562.

53. Veiga DF, Deus HF, Akdemir C, Vasconcelos AT, Almeida JS: **DASMiner: discovering and integrating data from DAS sources**. *BMC Syst Biol* 2009, **3**:109.

54. Whitington T, Frith MC, Johnson J, Bailey TL: **Inferring transcription factor complexes from ChIP-seq data**. *Nucleic acids research* 2011, **39**(15):e98.

55. Yasukochi Y, Maruyama O, Mahajan MC, Padden C, Euskirchen GM, Schulz V, Hirakawa H, Kuhara S, Pan XH, Newburger PE *et al*: **X chromosome-wide analyses of genomic DNA methylation states and gene expression in male and female neutrophils**. *Proc Natl Acad Sci U S A* 2010, **107**(8):3704-3709.

56. Zentner GE, Saiakhova A, Manaenkov P, Adams MD, Scacheri PC: **Integrative genomic analysis of human ribosomal DNA**. *Nucleic Acids Res* 2011, **39**(12):4949-4960.

57. Zhang Z, Zhang MQ: **Histone modification profiles are predictive for tissue/cell-type specific expression of both protein-coding and microRNA genes**. *BMC Bioinformatics* 2011, **12**:155.

Section AB: Supplementary Information References

1 Kundaje, A. *et al.* Ubiquitous heterogeneity and asymmetry of the chromatin environment at regulatory elements. *Genome research* **22**, doi: 10.1101/gr.136366.111 (2012).

2 Rozowsky, J. *et al.* PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nature biotechnology* **27**, 66-75, doi:10.1038/nbt.1518 (2009).

3 Bickel, P. J., Boley, N., Brown, J. B., Huang, H. Y. & Zhang, N. R. Subsampling Methods for Genomic Inference. *Annals of Applied Statistics* **4**, 1660-1697, doi:Doi 10.1214/10-Aoas363 (2010).

4 Pennacchio, L. A. *et al.* In vivo enhancer analysis of human conserved non-coding sequences. *Nature* **444**, 499-502, doi:10.1038/nature05295 (2006).

5 Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L. A. VISTA Enhancer Browser--a database of tissue-specific human enhancers. *Nucleic acids research* **35**, D88-92, doi:10.1093/nar/gkl822 (2007).

6 Rembold, M., Lahiri, K., Foulkes, N. S. & Wittbrodt, J. Transgenesis in fish: efficient selection of transgenic fish by co-injection with a fluorescent reporter construct. *Nature protocols* **1**, 1133-1139, doi:10.1038/nprot.2006.165 (2006).

7 Thermes, V. *et al.* I-SceI meganuclease mediates highly efficient transgenesis in fish. *Mechanisms of development* **118**, 91-98 (2002).

8 Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology* **10**, R25, doi:10.1186/gb-2009-10-3-r25 (2009).

9 Rozowsky, J. *et al.* AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Molecular systems biology* **7**, 522, doi:10.1038/msb.2011.54 (2011).

10 1000_Genomes_Project_Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073, doi:nature09534 [pii]10.1038/nature09534 (2010).

11 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842, doi:10.1093/bioinformatics/btq033 (2010).

12 Pleasance, E. D. *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**, 191-196, doi:10.1038/nature08658 (2010).

13 Pleasance, E. D. *et al.* A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* **463**, 184-190, doi:10.1038/nature08629 (2010).

14 Harrow, J. *et al.* GENCODE: The reference human genome annotation for the ENCODE project. *Genome research* **22(9)** doi: 10.1101/gr.135350.111 (2012).

15 Derrien, T. *et al.* The GENCODE v7 catalogue of human long non-coding RNAs: Analysis of their gene structure, evolution and expression. *Genome research* **22(9)** doi: 10.1101/gr.132159.111 (2012).

16 Pei, B. *et al.* The GENCODE Pseudogene Resource: Integration of Functional Genomics Evidence Allows Comprehensive Annotation of Partial Activity. *Genome biology* **Manuscript under review.** (2012).
17 Kharchenko, P. V., Tolstorukov, M. Y. & Park, P. J. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nature biotechnology* **26**, 1351-1359, doi:10.1038/nbt.1508 (2008).
18 Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome biology* **9**, R137, doi:10.1186/gb-2008-9-9-r137 (2008).
19 Valouev, A. *et al.* Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nature methods* **5**, 829-834, doi:10.1038/nmeth.1246 (2008).
20 Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods* **5**, 621-628, doi:10.1038/nmeth.1226 (2008).
21 Ji, H. *et al.* An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nature biotechnology* **26**, 1293-1300, doi:10.1038/nbt.1505 (2008).
22 Li, Q. H., Brown, J. B., Huang, H. Y. & Bickel, P. J. Measuring Reproducibility of High-Throughput Experiments. *Annals of Applied Statistics* **5**, 1752-1779, doi:Doi 10.1214/11-Aoas466 (2011).
23 Li, X. Y. *et al.* The role of chromatin accessibility in directing the widespread, overlapping patterns of Drosophila transcription factor binding. *Genome biology* **12**, R34, doi:10.1186/gb-2011-12-4-r34 (2011).