Unravelling higher order genome organisation [working title]

Introduction

Benjamin L. Moore

May 5, 2015

igmm
INSTITUTE OF GENETICS
& MOLECULAR MEDICINE

1 METHODS

1.1 INPUT DATA

Some more stuff here.

1.1.1 Hi-C data

Raw Hi-C reads were downloaded from three published datasets through GEO[1] or the SRA[2] with identifiers: GSE35156 (H1 hESC), GSE18199 (K562) and SRX030113 (GM12878). These paired reads were mapped independently to theand mapped to the genome (hg19/GRCh37). Mapping was performed using the `hiclib` software package[3] and `bowtie2`[4] with the `--very-sensitive` flag. Mapped reads were then binned into contact maps and iteratively corrected[3]. The `hiclib` software was also used for eigenvector expansion of each intrachromosomal contact map, performed independently on each chromosome arm.

1.1.2 Locus-level features

Genome-wide ChIP-seq datasets for: 22 DNA binding proteins and 10 histone marks were made available by the ENCODE consortium[5,6] along with DNase I hypersensitivity and H2A.z occupancy, for each of the Tier 1 ENCODE cell lines used in this work: H1 hESC, K562 and GM12878. These data were pre-processed using MACSv2[7] to produce fold-change relative to input chromatin. GC content was also calculated and used in the featureset.

1.2 MODELLING

1.2.1 Random Forest

Random Forest regression[8] was used as implemented in the R package `randomForest`.[9] Parameters of $mtry = \frac{n}{3} = 12$ and $ntrees = 200$ were assumed as they approximate the defaults and are known to be largely insensitive.[10]

Variable importance within Random Forest regression models was measured using mean decrease in accuracy in the out-of-bag (OOB) sample. This represents the average difference (over the forest) between the accuracy of a tree with permuted and unpermuted versions of a given variable, in units of mean squared error (MSE).[11]

1.2.2 Model performance

The effectiveness of the modelling approach was measured by four different metrics. Prediction accuracy was assessed by the Pearson correlation coefficient between the predicted and observed eigenvectors (determined by 10-fold cross-validation), and the root mean-squared error (RMSE) of the same data. Classification error, when predictions where thresholded into $A \geq 0; B < 0$, was also calculated using accuracy (% correct classifications or True Positives) and area under the receiver operating characteristic (AUROC) curve. Together these give a comprehensive overview of the model performance, both in terms of regression accuracy of the continuous eigen-

vector, and in how that same model could be used to label discrete chromatin compartments.

For cross-application of cell type specific models, a single Random Forest regression model was learned from all 1 Mb bins for a given cell type. This was then used to predict all bins from each of the other two cell types.

1.3 VARIABLE REGIONS

1.3.1 Stratification by variability

Median absolute deviation (MAD) was chosen as a robust measure of the variability in a given 1 Mb block between the three primary cell types used in this work: H1, K562 and GM12878. Blocks were ranked by this measure and split into thirds that represented "low" variability (the third of blocks with the lowest MAD), "mid" and "high" variability. Each subgroup was then independently modelled using the previously-described Random Forest approach.

"Flipped" regions are those whose compartment state differs in one cell type relative to the other two. For example, if a 1 Mb bin was classified as "open" in H1 hESC and "closed" in both K562 and GM12878, this is said to be a "flipped" compartment (to open).

1.3.2 Enhancer enrichment

Enhancer annotations were collected from the ChromHMM / SegWay combined annotations in each cell type.[12] Enhancers were considered "shared" if there was an overlapping enhancer annotation in either of the two other cell types, and labelled as "tissue-specific" otherwise.

1.4 BOUNDARIES

1.4.1 TADs

TAD boundaries were called using the software provided in Dixon *et al.*[13] using their recommended parameters. For the generation of boundary profiles, the same parameters were used: input features were averaged into 40 kb bins spanning $\pm$500 kb from the boundary centre.

1.4.2 Compartments

Compartment boundaries were called by first training a two-state hidden Markov model (HMM) on the compartment eigenvector and then using the Viterbi algorithm to predict the most likely state sequence that produced the observed values. The point at which transitions occurred between states was taken as a boundary which was then extended $\pm$1.5 Mb to give a 3 Mb window in which a boundary was though to occur.

To test for the enrichment or depletion of a chromatin feature over a given boundary, a two tailed Mann-Whitney test was used to compare the boundary bin with the ten outermost bins of the window (5 from either side). The significance level at $\alpha = 0.01$ was then Bonferonni-adjusted for multiple testing correction, and results with p-values exceeding this threshold were deemed significantly enriched or depleted at a given boundary.

1.5 GIEMSA BAND COMPARISON

Cytogenic band data and Giemsa stain results were downloaded from the UCSC genome browser (table `cytoBandIdeo`). The genomic co-ordinates are an approximation of cytogenic band data inferred from a large number of FISH experiments.[14]

To compare G-band boundaries with our compartment data, we allowed for a $\pm$500 kb inaccuracy in G-band boundary. For each G-band boundary, the minimum absolute distance to any compartment or TAD boundary was calculated for each cell type. To generate a null model, ...

REFERENCES

[1] Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall Ka, Phillippy KH, Sherman PM, *et al.* (2013) NCBI GEO: archive for functional genomics data sets–update. *Nucleic acids research*, **41**(Database issue): D991–5.

[2] Leinonen R, Sugawara H, Shumway M (2011) The sequence read archive. *Nucleic acids research*, **39**(Database issue): D19–21.

[3] Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, Dekker J, Mirny LA (2012) Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature methods*, **9**(10): 999–1003.

[4] Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nature methods*, **9**(4): 357–9.

[5] ENCODE (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**(7414): 57–74.

[6] Boyle AP, Araya CL, Brdlik C, Cayting P, Cheng C, Cheng Y, Gardner K, Hillier LW, Janette J, *et al.* (2014) Comparative analysis of regulatory information and circuits across distant species. *Nature*, **512**(7515): 453–456.

[7] Zhang Y, Liu T, Meyer Ca, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome biology*, **9**(9): R137.

[8] Breiman L (2001) Random Forests. *Machine learning*, **45**(1): 5–32.

[9] Liaw A, Wiener M (2002) Classification and Regression by randomForest. *R News*, **2**(December): 18–22.

[10] Hastie T, Tibshirani R, Friedman J (2009) *Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2 edition. ISBN 978-0-387-84858-7.

[11] Cutler DR, Edwards TC, Beard KH, Cutler A, Hess KT, Gibson J, Lawler JJ (2007) Random forests for classification in ecology. *Ecology*, **88**(11): 2783–92.

[12] Hoffman MM, Ernst J, Wilder SP, Kundaje A, Harris RS, Libbrecht M, Giardine B, Ellenbogen PM, Bilmes Ja, *et al.* (2013) Integrative annotation of chromatin elements from ENCODE data. *Nucleic acids research*, **41**(2): 827–41.

[13] Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**(7398): 376–80.

[14] Furey TS (2003) Integration of the cytogenetic map with the draft human genome sequence. *Human Molecular Genetics*, **12**(9): 1037–1044.