Unravelling higher order genome organisation [working title]

Methods section

Benjamin L. Moore

June 10, 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.

Table 1: Public Hi-C data used in this work.

Cell line	Total reads	Accession	Citation
Gm12878	31×10^6	SRX030113	5
H1 hESC	331×10^6	GSE35156	6
K562	36×10^6	GSE18199	7
Cortex	373×10^6	GSE35156	6
mESC	476×10^6	GSE35156	6
IMR90	355×10^6	GSE35156	6

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[8,9] 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[10] to produce fold-change relative to input chromatin. GC content was also calculated and used in the featureset.

1.1.3 Clustering input features

To quantify collinearity of input features, correlation matrices built from genome-wide vectors of input feature measures were build and hierarchicaly clustered. The ”significance” of observed clustering was assessed using sub- and super-sampled bootstrapping, with stable clusters deemed significant. The `pvclust` R package

1.2 MODELLING

1.2.1 Random Forest

Random Forest (RF) regression,[11] was used as implemented in the R package `randomForest`.[12] The RF algorithm makes use of a collective of regres-

sion trees (size *ntrees*), each built from a bootstrapped sample of the training set. In growing each tree, a small number of variables (*mtry*) is tested at each bifurcation node, and that which minimises the variance in child node subsets is selected at a specific threshold. Having trained a group of trees, these can then be used as predictive tools by inputting a vector of features to each tree and averaging the output leaf node value across the forest. RF regression was used as it is known to be one of the most powerful regression methods developed to date,[13,14] typically providing low bias and low variance predictions without the need for variable selection.[15] Additionally the RF method represents an example of "algorithmic modelling"[16] in that it makes no assumptions about the underlying data model. Parameters of $mtry = \frac{n}{3}$ (where n is the number of input features) and $ntrees = 200$ were assumed as they are known to be largely insensitive;[15,17] this was verified with the dataset used in this work (Fig. **??**).

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).[14,15]

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 eigenvector, 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.

1.2.3 Other modelling approaches

Linear regression was used as a baseline for comparison with more complicated approaches such as Random Forest. If the same modelling accuracy could be achieved with simple multiple linear regression, this would be a faster and more interpretable modelling framework.

Partial least squares regression was also used to model compartment profiles. This method is well-suited to highly correlated inputs.

1.2.4 Graphical lasso

Regularised models made use of the Graphical LASSO[18] (least absolute shrinkage and selection operator) as a method of L_1-norm based regularisation, implemented via the `glasso` R package. The graphical lasso provides tuneable regularisation which is capable of feature selection via minimising regression parameters to 0. It was chosen in this case due to the multicollinearity of the featureset, the algorithm's fast speed of execution and the intuitiveness a graphical model presents.[18]

More specifically, the graphical lasso regulates the number of os in the inverse covariance matrix, $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$, also known as the precision matrix. Then if element $\theta_{ij} = 0$, the variables X_i and X_j can be said to be conditionally independent, given the remaining variables.[19] The algorithm minimises a negative log-likelihood (Eqn. 1 [19]) given the tuning parameter λ, which was tuned in this case to leave a small number of variables (< 10) directly dependent on the eigenvector data.

$$\underset{\mathbf{\Theta} \prec \mathbf{0}}{\text{minimise}}\ f(\mathbf{\Theta}) := -\log\det(\mathbf{\Theta}) + \operatorname{tr}(\mathbf{S\Theta}) + \lambda\|\mathbf{\Theta}\|_1 \tag{1}$$

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.[20] 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.

This was repeated for other chromatin states.

1.4 BOUNDARIES

1.4.1 TADs

TAD boundaries were called using the software provided in Dixon *et al.*[6] using their recommended parameters. For the generation of boundary profiles, the same parameters were used: input features were averaged into 40 kb bins spanning ±500 kb from the boundary centre.

To align boundaries between cells ...

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 ±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.

1.4.3 MetaTADs

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.[21]

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, ...

1.6 NUCLEAR POSITIONING

Previously published data on chromosome positioning preference within the nucleus was used to label each chromosome as "inner", "middle" or "outer".[22] Chromosomes whose DAPI hybridisation signals were significantly enriched ($p \leq 2 \times 10^{-2}$) in the inner nuclear shell, as defined by Boyle *et al.*[22], made up the "inner" group and included chromosomes 1 and 16. Similarly the "outer" group had enriched signals ($p \leq 5 \times 10^{-3}$) in the outer shell relative to the inner nuclear shell and included chromosomes 2, 3, 11-13 and 18. The remaining chromosomes in our filtered dataset, 6, 14 and 15, were assigned to the "middle" group and showed no significant to either inner or outer nuclear shells ($p \geq 0.1$).[22] The significance of the difference in distribution of eigenvectors in the inner versus outer shell was determined by a one-sided Kolmogorov-Smirnov (K-S) test, with the alternative hypothesis that the empirical cumulative density function of the inner chromosome eigenvectors F_{inner} is greater-than or equal-to F_{outer}. This chromosomal positioning data was measured in lymphoblastoid cells though nuclear architecture is though to be largely conserved between cell types[23,24] and even higher primates.[25]

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] Kalhor R, Tjong H, Jayathilaka N, Alber F, Chen L (2012) Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nature biotechnology*, **30**(1): 90–8.

[6] 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.

[7] Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**(5950): 289–93.

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

[9] 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.

[10] 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.

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

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

[13] Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP (2003) Random forest: a classification and regression tool for compound classification and QSAR modeling. *Journal of chemical information and computer sciences*, **43**(6): 1947–58.

[14] 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.

[15] Dasgupta A, Sun YV, König IR, Bailey-Wilson JE, Malley JD (2011) Brief review of regression-based and machine learning methods in genetic epidemiology: the Genetic Analysis Workshop 17 experience. *Genetic epidemiology*, **35 Suppl 1**(Suppl 1): S5–11.

[16] Breiman L (2001) Statistical Modeling: The Two Cultures. *Statistical Science*, **16**(3): 199–231.

[17] 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.

[18] Friedman J, Hastie T, Tibshirani R (2008) Sparse inverse covariance estimation with the graphical lasso. *Biostatistics (Oxford, England)*, **9**(3): 432–41.

[19] Mazumder R, Hastie T (2012) The graphical lasso: New insights and alternatives. *Electronic Journal of Statistics*, pp. 1–21.

[20] 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.

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

[22] Boyle S, Gilchrist S, Bridger JM, Mahy NL, Ellis Ja, Bickmore Wa (2001) The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells. *Human molecular genetics*, **10**(3): 211–9.

[23] Chambers EV, Bickmore Wa, Semple CA (2013) Divergence of mammalian higher order chromatin structure is associated with developmental loci. *PLoS computational biology*, **9**(4): e1003017.

[24] de Wit E, Bouwman BaM, Zhu Y, Klous P, Splinter E, Verstegen MJaM, Krijger PHL, Festuccia N, Nora EP, *et al.* (2013) The pluripotent genome in three dimensions is shaped around pluripotency factors. *Nature*, pp. 1–7.

[25] Tanabe H, Müller S, Neusser M, von Hase J, Calcagno E, Cremer M, Solovei I, Cremer C, Cremer T (2002) Evolutionary conservation of chromosome territory arrangements in cell nuclei from higher primates. *Proceedings of the National Academy of Sciences of the United States of America*, **99**(7): 4424–9.