Bachelor's Degree in Bioinformatics (UPF-UPC-UB)
Final Grade Project

Imputing eQTL from tissue specific and across tissues data

Laura Aviñó[1]

Scientific director: Daniel Zerbino[1]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK

Abstract

Motivation: Most interactions between regulatory regions such as enhancers and their target genes are still unknown. To address this issue, different methods have been developed to assess cis-regulatory interactions, using genetics, genomics, epigenomics and molecular biology. Currently, existing analysis methods to predict enhancer-promoter interactions from this data are tied either to linear data (e.g. gene expression and methylation) or to interaction data between a gene and its putative enhancers. Moreover, data sources are rarely being exploited across tissues.

Results: Here we present a method to jointly analyze all these pieces of evidence and infer cis-regulatory interactions between enhancers and promoters. In order to test the technique, we benchmarked against the GTEX eQTL dataset using eGTEx data. In this study, we demonstrate that combining assays increases the performance of the cis-regulatory interaction models and of the imputation of eQTLs .

Supplementary information: Supplementary data and software are available at https://github.com/LauAvinyo/inteql

1 Introduction

The human genome's estimated 19,000 protein-coding genes and numerous non-coding genes are expressed at different levels depending on cell type and cell environment, thanks to cis-regulatory elements called enhancers that modulate transcription at the promoters. They have been found to be involved in (1) cell differentiation and cell lineage, (2) phenotype differences including morphological differences between species[1] and (3) complex diseases[2–4]. Huge efforts are being made to annotate them, both computationally[5] and experimentally[6].

Detecting Promoter-Enhancer interactions is not a simple task[7], therefore most of them are still unknown. Irrespective of the linear genomic order, orientation and distance, an enhancer and its promoter target can be in physical proximity. Moreover, a single regulatory region can act on different genes depending on the cell type or tissue, as an activator or silencer, and differently in different cell types adding more complexity to the regulatory network.

To address this issue, different assays have been developed to identify interactions between enhancers and promoters. We can classify them into four different categories. First, the genetic approaches based on quantitative trait locus cohorts across individuals. These techniques find statistical correlations between markers and intermediate phenotypes such as gene expression (eQTLs) or protein levels (pQTL). Second, genomic approaches define the three-dimensional structure of the genome and infer regulatory interactions from physical contacts. Examples of these are Hi-C[8], Promoter Capture Hi-C[9] and ChiA-PET[10]. Third, epigenomic approaches measure the covariance of epigenomic mark intensity between genomic regions. Finally, molecular biology approaches such as Massively Parallel Enhancer Validations (MPRAs)[11] and high-throughput CRISPR variations directly test molecular variants and their effects.

Recently, data production projects based on these approaches have been carried out leading to large scale sources of data. For example: GTEx has collected and analysed data for eQTLs (gtexportal.org/); ENCODE[12] and Roadmap-epigenomics[7] for epigenomic chromatin data and 4DNucleosome[13] for Hi-C data. Moreover, different studies have published their data for MPRAs and CRISPR techniques[14,15].

To process the results of these new assays, new analysis methods have been developed. They can be classified depending on the data they use as input. We will call 1D the methods that use linear data such as gene expression, chromatin marks, or methylation The input data of all these methods are different, but, they can be explained as a matrix Locus x feature of that locus. In this group, we find methods like JEME[16], eQTL[17], TargetFinder[18], EpiTensor[19], RIPPLE[20], PresTIGE[21]. Alternatively, there is a second group that use interaction data, i.e. physical interactions captured with Hi-C[22], ChiA-PET[10] or derived methods such as the correlations between measures of different features of the enhancers and promoters. Some of them directly take physical contacts data, such as PSYCHIC[23] that infers regulatory interactions with a certain probability. Others use the same data to train models, such as EP2Vec[24] and SPEID[25]. Finally, some methods use correlations as input for their models. For instance, Im-PET[26] uses the co-evolution of the sequences, enhancer-promoter activity profile correlation and the correlation between the expression of the Transcription Factors that bind the enhancer with the expression of the gene. However, to the best of our knowledge, no method integrates both 1D and interaction data.

In an effort to develop a robust and widely applicable method to infer cis-regulatory interactions, we set out to evaluate whether an ensemble approach that aggregates as much experimental evidence as available would provide reliable results. Using finemapped eQTLs as an accurate measure of regulatory causality in a specific cell-type context, we explored various approaches to first model then impute these measurements across a greater number of tissues.

2 Material and methods

2.1 Data

We focussed our analysis on the GM12878 cell line. This is a lymphoblastoid cell line produced by an Epstein-Barr Virus (EBV) transformation.

eQTL

We have used eQTL data from GTEx version 7. For modeling, we used the list of significant associations from the EBV-lymphoblastoid. This cell line contains 130 samples, 117 of them with donor genotype and 3845 eGenes (genes with at least one significant eQTL). For each variant-gene interaction, we matched the enhancer(s) for TargetFinder list[18] that contain that variant and the promoter of that gene. For the imputation task we used all variants that were significant eQTLs in at least one tissue. From GTEx we get the p-values of the association. To factor out artifactual correlations caused by Linkage Disequilibrium (LD) we performed finemapping using the FineMap[27] algorithm, thus producing Z-scores.

Chromatin Marks

The epigenomic data consisted of CAGE[28], DNase-seq[29], FAIREseq[30], Methylation[31] and Chip-seq[32] for different histone modifications and transcription factors data from the ENCODE project (See Supplementary Table 1 for datasets accessions). We used the processed bed files from the TargetFinder benchmark. We attached to each region (enhancer or promoter as annotated by ENCODE and Roadmap Epigenomics) all the peaks for each Transcription Factor, Methylation, Expression or Open Chromatin measures and presented it to the models as a unique feature.

Hi-C

We have used the in situ Hi-C Rao et al. data for GM1287. Concretely, we used the raw observed contacts matrix for the diploid GM12878 cell line at 5kb resolution. The data can be found in NCBI GEO acc=GSE63525. Raw Hi-C data was normalized different ways: Vanilla Coverage, Square Root of the previous normalization and KR normalization . Vanilla Coverage is a fast and robust way to normalize. This normalization method tends to over correct the data. One way to solve this is by taking the square root of it. This has been shown to perform similarly to much more complex algorithms. Random polymer interactions in close genomic regions can increase the number of contacts in such pairs. To take that into account Observed/Expected matrices should be constructed. The instructions normalizing and compute expected values can be found in Rao et. all supplementary information[22].

RNA Expression

We have used the gene expression levels reported in GTEx for each sample. We used the 5000 top expressed genes in the 11688 samples collected for the GTEx Version 7 release. We processed the data with log CPM normalization using the R function limma::voom().

2.2 Modeling

To determine the amount of cross-information between different types of experimental evidence, we first trained different models of the eQTL Z-score. We tested both Random Forest and Decision Trees with different subsets of input data: a) a set containing only linear chromatin data, b) a set containing only Hi-C, c) containing only eQTL data and finally d) the combination of all the above sets. After generating a data matrix with all the above data we used regression random forest and decision tree models to predict eQTL slopes/z-scores using epigenomic data, Hi-C contacts an association measures from other tissues' eQTL as predictors. All the models are imported from the Sklearn[33] python library (Version 0.20.2). Random forests have used 100 estimators. To split the data into validation and train datasets we have used the sklearn train_test_test. We used 30% of the data for validation. To quantify the importance of each feature in a Random Forest we used the feature_importances_ atribut of Sklearn ExtraTreeRegressor.

2.3 Imputation

One simple way to impute values is to compute the mean of each row and use this value as the eQTL score of our goal tissue. To do so, we have used mean function from the pandas library. This approach assumes that there is only minor variation between tissues, and does not allow for trends across tissues, let alone tissue-specific outliers. We assumed that that similar tissues will have similar eQTL network, therefore add similarity information may improve the imputation. One way to assess the distance between tissues is by using the gene expression. To perform a weighted mean imputation we have designed a four-step pipeline. a) Apply a dimensionality reduction technique to RNA data from all the tissues. b) Project the data into the two top factors space. c) Extract the Euclidean distance between tissues. d) Compute the weighted mean using as coefficient the inverse of the distance between genes. As a dimensionality reduction techniques, we have used PCA and tSNE. We compute them using R prcomp() and tsne::tsne() functions, respectively.

3 Results

Modeling

We used four different types of model: Random Forest with maximum depth of 5, default Random Forest, Decision Tree with maximum depth 5 and Decision Tree with default parameters. Each type model was trained with four different sets of data: epigenomics, Hi-C, Tissue Specific (epigenomics and Hi-C), eQTLs from other tissues and the combination of all. The Root Mean Squared Error of the prediction of the test data for each model is shown in table 1.

After training the models we check that they make sense from a biological point of view. In all cases the models gave importance to features known to be important for each region. Figure 1 shows the importance of each feature (tissue) of a Random Forest trained with only eQTL data across tissues. Whole Blood, the reasonable closest tissue to the Lymphoblast shows higher level of importance. Tissues with high level of immune cells also have higher importance in the model, such as Whole Blood, Lung or Mucosa. In the same fashion, Skin exposed to sun has more immune cells then Skin not exposed to sun, the model may show this fact assigning more importance to the Skin exposed to sun. Following the same hypothesis we may explain why adipose underskin has a higher effect than visceral fat. Another example is shown in Supplementary Figure 1 where a Decision Tree trained with tissue specific data is represented. The tree uses Histone Marks for enhancers and promoters already shown in literature to be predominantly shown in that type region. Another important point is that the branches organization follows a logic. For example, after checking the presence at the promoter of H3K4me3, a histone mark found in inactive promoters, the decision tree checks whether the promoter is bound by CUX1, a transcription factor thought to be a repressor.

Table 1. Root Mean Squared Error of the different models.

Data	RF5	RF	DT5	DT
Epigenomics	0.3798	0.2414	0.4307	0.2608
Hi-C	0.4626	0.3225	0.4726	0.3303
Tissue Specific	0.3798	0.2414	0.430	0.279
eQTL	0.3005	0.2528	**0.3165**	0.2693
All	**0.2819**	**0.0976**	0.3245	**0.1449**

Each column is one type of model: Random Forest with max depth 5, default Random Forest, Decision Tree with max depth 5 and a default Decision Tree. Each row is one data subset: Epigenomics, Hi-C, Tissue specific (Epigenomics and Hi-C), eQTL from other tissues and the combination of the previous.

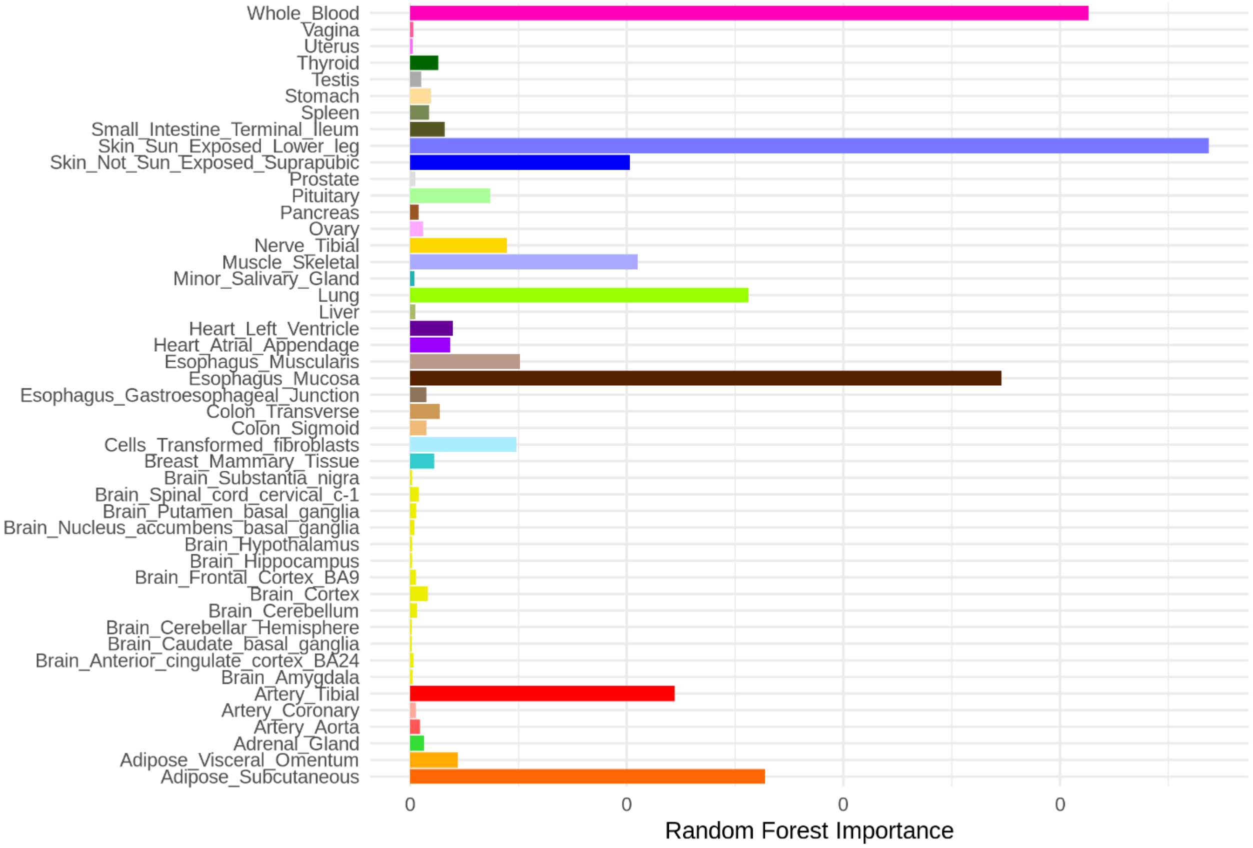

Fig.1. Random Forest Tissue importance. Shows the feature importance of each tissue for a default Extra Trees Regressor trained with eQTL across tissues data.

Next, we examined the effect of the different subsets of features in the prediction. As expected, epigenomics is able to predict eQTLs[18]. Hi-C alone also contains information to predict Enhancer-Promoter interactions and Random Forest and Decision Trees use them in combination of linear data or eQTLS. However, is the subset of data that performs worst. Moreover, adding Hi-C to the other subsets does not improve predictions or it does so in a very modest way. Another remarkable result that using across tissue data performs quite good: eQTLs from other tissues are the best predictors. It is reasonable to think that is a good practice to use all the types of evidence in the same model. Nevertheless, we are aware that joining data for different tissues with fundamentally different experiments used to generate it may be a challenge.

Table 2. R value of the regression of the predicted and real values for the validation data set.

Data	GTEx Effect size	FineMap Z-score
Epigenomics	0.8094	0.8251
Hi-C	0.7725	0.7733
Tissue Specific	0.8149	0.8259
eQTL	0.9592	0.9609
All	0.9681	0.9715

Each row is the subset of features used to train the model. The first columns shows the result when using the GTEx reported effect size and the second columns shows the result when using the FineMap Z-scores. All models are default Random Forest.

One drawback of eQTLs is Linkage Disequilibrium: some SNPs in enhancers may not be causal variants but are genetically linked to the variant responsible of the eQTL. To factor out this confounding factor we ran FineMap on chromosome 3. Table 2 shows the R-value of the regression predicted eQTL scores and the real values for the test data set. These scores are effect size of the eQTL reported by GTEx and the Z-score computed by FineMap. Using Z-scores improve the predictions. This is especially clear with chromatin marks, which have comparatively high resolution in genomic coordinates and define tight boundaries around regulatory elements. eQTLS from other tissues may better predict the effect size of the eQTLs in our tissue of interest since they may be subject to the same linkage effect.

Imputation

We then imputed one tissue's eQTLs using other tissues' eQTLs. Here, we did not restrict to Enhancer-Promoter interactions. Thus, we use all the eQTLs that are significant in at least one tissue.

We first tested the performance of a basic imputation using the mean. We held back the eQTL values for Cells EVB-transformed lymphoblast. Then, for each eQTL we compute the mean of the eQTL effect scores in those tissues where it is significant. Figure 2 shows the real values versus the imputed values for these tissue. The Root Mean Squared Error of these imputation is 0.3387. Despite being insensitive to tissue identity, this method performed very well, as compared to the different models described above.

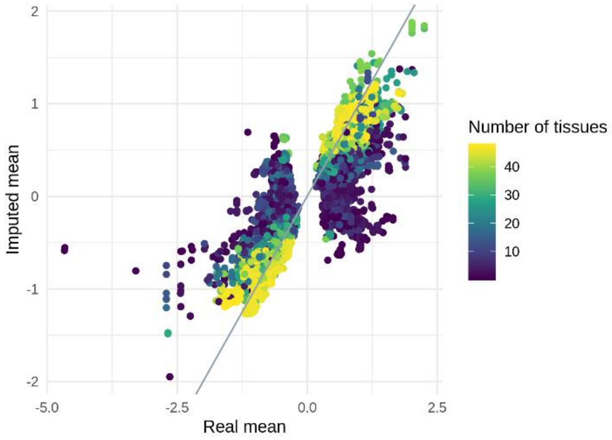

Fig. 2. Real eQTL values for the Cells Transformed Limfoblast and their mean imputed values. The imputation is done only with the other tissues in which that eQTL is significant. In grey is shown a one to one line.

We then explored whether we can improve these results by performing weighted means, such that similar tissues to our target tissue are given more importance when imputing the value. Since we do not know how close our target tissue is from others in the eQTL space, we used GTEx RNA expression data as a proxy for tissue similarity. We used all the sample expression for the 5000 most expressed genes. After log CPM normalization we performed PCA and tSNE in order to reduce the dimensionality of the data. We then computed the tissue centroids in the two top dimensions of each technique.

To support our assumption, we plotted in Figure 3 the Euclidean distance of the centroid of each tissue in RNA two top PC space and the Euclidean distance between tissues in the eQTL two top PC space. There is a weak (Pearson's r: 0.53) but significant (p-value < 0.05) correlation between those two measures. Moreover, most of the outliers are present in the bottom right corner of the plot. This is expected as the number of tissues in which a gene is active is higher than the number of tissues a non-housekeeping eQTLs is significant.

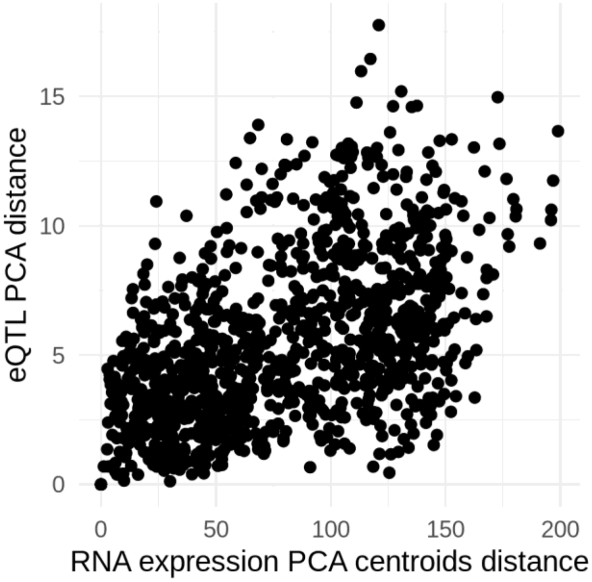

Fig. 3. Tissue pairwase distance PCA RNA vs PCA eQTL. X axis shows the Euclidean distance between the centroid of each tissue in the space formed by the two top components of the RNA expression PCA. Y axis shows the distance between tissues in the space formed by the top components of the eQTL PCA.

We then tested weighted mean imputation, using the inverse of the Euclidean distance as coefficient. Even though the distance in a tSNE plot is not necessarily a good proxy of closeness between tissues we tested its efficiency as well. Figure 4 shows the Root Mean Square Error and the R-value of the regression between real and imputed values for each tissue in GTEx. For both PCA and tSNE the weighted means approach brought a small but consistent improvement. PCA weights performed better than tSNE weights as we expected due to the nature of these techniques. We added to the plot the same values for a Random Forest model. Is not surprising that when there was data for that tissue (as in the case of models) the predictions were much better. It is also notable that tissues with higher number of eQTLs and samples overall are easier to impute[34].

4 Discussion

Hi-C data informs us about physical interaction, and is often assumed to imply regulatory interaction[18,25]. We therefore used eQTL data as a measurement of interaction. This approach has some caveats: First, the amount of tissues and cell types covered by eQTL datasets is smaller than Hi-C. This sparsity of data is limiting to complex models, like Deep Learning techniques that require a big amount of training data. Second eQTL may contain a lot of false negatives because of the low frequency of some causal variants.

Imputation may help to expand the coverage of eQTLs to tissues for which no eQTL data is available but expression data is available. We could also improve predictions using new features or preprocessing them in different ways. For instance, we tried to add the state (Ie. Active, inactive, repressed, poised) of promoters and enhancers, so that the models do not need to use information that we already know, eg. the histone marks associated with either region. This did not increase the predictive power of the model, however it changed the weight of the features in the models. For instance, decision trees used Hi-c as the root. Factoring out already know information may be useful for the models, but this approach still needs refinement.

One limitation of imputation is the loss tissue specific eQTLs One possible solution is using first epigenomic and Hi-C data to predict the variant-gene pairs with significant interaction and then use imputation across tissues to impute a score value. Selecting the subset of significant eQTLs is a task that remains to be done.

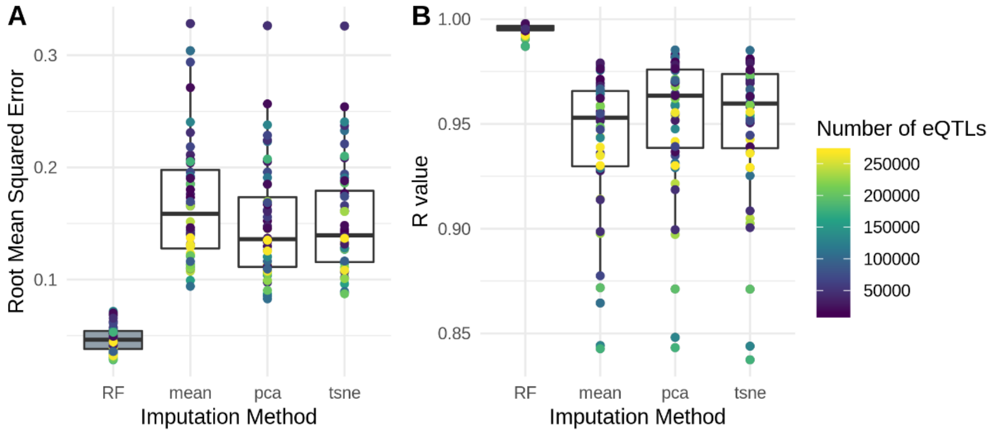

Fig. 4. Imputation Figures of Merit. A shows the Root Mean Square Error of the eQTL imputation for each tissue. Each white box is one imputation method. In grey the RMSE of the predictions of a default Random Forest trained with 70% of the eQTLs. **B** shows the R value of the regression line between the real values and the imputed values for each tissue and method. The grey box shows the R Value distribution of the regression lines of predicted and real values for the Random Forest models. The color scale shows the amount of significant eQTLs each tissue has.

We have observed that there is a relation between eQTL and RNA similarity, and this could be expanded on. One approach would be to use Tensor Decomposition methods such as MOFA[35] on a combination of transcriptomic, epigenomic and cis-regulatory data. Those methods decompose together different types of data into a small number of latent variable signatures The inverse mathematical transformation is then used to impute missing data.

5 Conclusions

We have shown that models can use tissue-specific and cross-tissue data to predict eQTL in a tissue of interest. Moreover, such models are biologically meaningful. Despite being frequently used as proxy for cis-regulatory interactions, Hi-C is a modest predictor. Thus, physical contacts should not be assumed to be in direct relation with regulatory interaction. eQTL data from other tissues are however good predictors. Finally, the combination of all features trained models produced the best models.

We explored the feasibility of imputing eQTL for a target tissue assuming no data for that tissue at all. We show that the mean across tissues with the same eQTL has already reasonable performance, although it does not account for tissue-specificity at all. We refined this imputation using tissue-specific weighted means. In those means, tissues closer to the target tissue in the RNA expression space have more importance. Finally, we verified that indeed, the two top components of eQTL PCA and RNA expression PCA for GTEx data are related.

Acknowledgements

I would like to express my special thanks to Daniel for the support I have received these months. Also, to my colleague and friend Rachel for the invaluable help she offered me expecting nothing in return. Finally, this project would have been impossible without the staff from the Bachelor's who not only teach me materials but how to be a scientific thinker.

Funding

LA was funded from core EMBL funds

Conflict of Interest: none declared.

References

1. Carroll, S. B. Evolution at Two Levels: On Genes and Form. *PLoS Biol.* **3**, e245 (2005).
2. Farh, K. K.-H. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337–343 (2014).
3. Adam, R. C. *et al.* Pioneer factors govern super-enhancer dynamics in stem cell plasticity and lineage choice. *Nature* **521**, 366–370 (2015).
4. Spitz, F. & Furlong, E. E. M. Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.* **13**, 613–626 (2012).
5. Zerbino, D. R., Wilder, S. P., Johnson, N., Juettemann, T. & Flicek, P. R. The Ensembl Regulatory Build. *Genome Biol.* **16**, 56 (2015).

6. Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L. A. VISTA Enhancer Browser–a database of tissue-specific human enhancers. *Nucleic Acids Res.* **35**, D88–D92 (2007).
7. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
8. Belaghzal, H., Dekker, J. & Gibcus, J. H. Hi-C 2.0: An optimized Hi-C procedure for high-resolution genome-wide mapping of chromosome conformation. *Methods* **123**, 56–65 (2017).
9. Schoenfelder, S., Javierre, B.-M., Furlan-Magaril, M., Wingett, S. W. & Fraser, P. Promoter Capture Hi-C: High-resolution, Genome-wide Profiling of Promoter Interactions. *J. Vis. Exp.* (2018). doi:10.3791/57320
10. Goh, Y. *et al.* Chromatin Interaction Analysis with Paired-End Tag Sequencing (ChIA-PET) for Mapping Chromatin Interactions and Understanding Transcription Regulation. *J. Vis. Exp.* (2012). doi:10.3791/3770
11. Inoue, F. & Ahituv, N. Decoding enhancers using massively parallel reporter assays. *Genomics* **106**, 159–164 (2015).
12. consortium, T. E. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
13. Dekker, J. *et al.* The 4D nucleome project. *Nature* **549**, 219–226 (2017).
14. Smith, R. P. *et al.* Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nat. Genet.* **45**, 1021–1028 (2013).
15. Xie, S., Duan, J., Li, B., Zhou, P. & Hon, G. C. Multiplexed Engineering and Analysis of Combinatorial Enhancer Activity in Single Cells. *Mol. Cell* **66**, 285–299.e5 (2017).
16. Cao, Q. *et al.* Reconstruction of enhancer–target networks in 935 samples of human primary cells, tissues and cell lines. *Nat. Genet.* **49**, 1428–1436 (2017).
17. Aguet, F. *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
18. Whalen, S., Truty, R. M. & Pollard, K. S. Enhancer–promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nat. Genet.* **48**, 488–496 (2016).
19. Zhu, Y. *et al.* Constructing 3D interaction maps from 1D epigenomes. *Nat. Commun.* **7**, (2016).
20. Roy, S. *et al.* A predictive modeling approach for cell line-specific long-range regulatory interactions. *Nucleic Acids Res.* **43**, 8694–8712 (2015).
21. Corradin, O. *et al.* Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. *Genome Res.* **24**, 1–13 (2013).
22. Rao, S. S. P. *et al.* A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* **159**, 1665–1680 (2014).
23. Ron, G., Globerson, Y., Moran, D. & Kaplan, T. Promoter-enhancer interactions identified from Hi-C data using probabilistic models and hierarchical topological domains. *Nat. Commun.* **8**, (2017).
24. Zeng, W., Wu, M. & Jiang, R. Prediction of enhancer-promoter interactions via natural language processing. *BMC Genomics* **19**, (2018).
25. Singh, S., Yang, Y., Poczos, B. & Ma, J. Predicting Enhancer-Promoter Interaction from Genomic Sequence with Deep Neural Networks. (2016). doi:10.1101/085241
26. He, B., Chen, C., Teng, L. & Tan, K. Global view of enhancer-promoter interactome in human cells. *Proc. Natl. Acad. Sci.* **111**, E2191–E2199 (2014).
27. Benner, C. *et al.* FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493–1501 (2016).
28. Takahashi, H., Kato, S., Murata, M. & Carninci, P. CAGE (Cap Analysis of Gene Expression): A Protocol for the Detection of Promoter and Transcriptional Networks. in *Methods in Molecular Biology* 181–200 (Humana Press, 2011). doi:10.1007/978-1-61779-292-2_11
29. Song, L. & Crawford, G. E. DNase-seq: A High-Resolution Technique for Mapping Active Gene Regulatory Elements across the Genome from Mammalian Cells. *Cold Spring Harb. Protoc.* **2010**, pdb.prot5384–pdb.prot5384 (2010).
30. Giresi, P. G., Kim, J., McDaniell, R. M., Iyer, V. R. & Lieb, J. D. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res.* **17**, 877–885 (2007).
31. Garrett-Bakelman, F. E. *et al.* Enhanced Reduced Representation Bisulfite Sequencing for Assessment of DNA Methylation at Base Pair Resolution. *J. Vis. Exp.* (2015). doi:10.3791/52246
32. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-Wide Mapping of in Vivo Protein-DNA Interactions. *Science* **316**, 1497–1502 (2007).
33. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
34. GTExConsortium. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
35. Argelaguet, R. *et al.* Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Mol. Syst. Biol.* **14**, e8124 (2018).