EN.601.448/648 Computational genomics: Final Project - Midterm report

Nae-Chyun Chen (cnaechy1)
Department of Computer Science

Arun Das (adas21)
Department of Computer Science

Xinyu Feng (xfeng17)
Department of Biology

April 12, 2019

Due: 4/12/2019

1 Problem

Prediction of Pioneer Transcription Factors Using Chromatin Accessibility and Gene Expression

The eukaryotic genome is packaged into chromatin. Concerted efforts from the ENCODE Consortium have revealed that different cell types vary in their accessible chromatin landscape. Open chromatin regions play an important role in cell identity, as they mark the gene regulatory regions which are accessible to cellular environment. However, the principles behind how open chromatin landscape is established and regulated are poorly understood. Here, we aim to identify transcription factors (TFs) that regulate chromatin accessibility, or pioneer transcription factors. Our preliminary results show that for each cell type, there are a unique set of TFs that predict chromatin accessibility. The cell type variability may be due to differential expression of pioneer transcription factors, which regulate different downstream targets. In this project, our goal is to incorporate TF expression level into the model and reveal how expression level is related to the "pioneer potential of transcription factors.

2 Data

- ENCODE Project [6]: DNase-seq, poly-A RNA-seq and total RNA-seq data
- JASPAR database [5]: TF motifs position weight matrix (PWM)

3 Methods

The positive training data is obtained from ENCODE DNase-seq data and an equal-size false training data set is randomly sampled with the same length as the positive set. The sequences are scanned against 537 human TF motifs position weight matrices (PWMs) for significant matches using FIMO [8]. This data pre-processing step generates a 2Nx537 matrix, where N is the number of DNase peaks.

Our preliminary results have shown that few universal candidate pioneer TFs are predicted across different cell lines. To explain this, a hypothesis is chromatin accessibility is associated with gene expression. If a TF gene has low expression level, it's unlikely to act as the pioneer. Therefore, expression levels obtained from poly-A are considered. Python package mygene is used to convert a TF gene into the ensembl gene format. To generate expression level for each transcription factor, we consider all the possible transcripts in each gene by summing the TPM values for each transcript.

Logistic regression classifier is used to model whether a motif is predictive of open or closed chromatin. The data matrix is randomly shuffled and split into for training and for testing. The training features include motif counts for each of the 537 TFs and their expression-weighted motif counts, which are calculated by multiplying motif counts by their corresponding expression level. Training is performed using sklearn.linear_model.LogisticRegressionCV, where the regularization parameter is optimized using 5-fold cross-validation. To evaluate the performance, we use the model learned from the training data to predict the test data.

After the midterm report, we plan to switch the training package into statsmodels to report the p-values of different features. Thereby we have the level of significance to test if weighted numbers of matches are effective features. Also, currently we label summed TPM for both a missing gene and a non-expressed gene as zero. This makes the model unaware of missing data, since some of the TF genes might not be recorded by the ENCODE dataset we use. We'll remove the weighted motif count from the features when a gene is missing. To explore other sources of expression data, total RNA-seq data can also be included as features. Once the pipeline is set up, we will compare the most effective features across different cell lines to see if they can be used to identify TFs with pioneer factor properties.

4 Evaluation metrics

To evaluate our performance, we compared the findings of our models to previous results using motif counts only without expression data. We used traditional accuracy metrics on our model, including reserving one third of the data for testing, using cross-validation to optimize training parameters and using ROC curves to evaluate the model's accuracy.

5 Results (at least 2 figures)

Figure 1 is a heatmap of transcription factor expression levels from two ENCODE Tier 1 cell lines. For K562 cells, we explored transcript quantification files from two isogenic replicates (ENCFF297CNO and ENCFF879WBJ) and a gene quantification file (ENCFF285HUZ); for GM12878, we used transcript quantification files from two isogenic replicates (ENCFF853TRI and ENCFF305QBE).

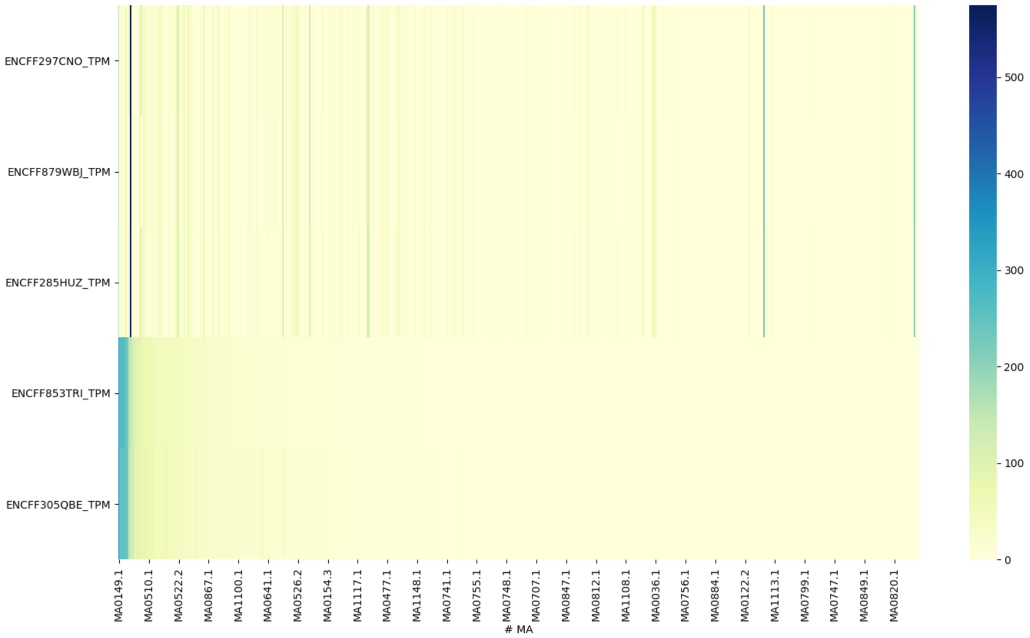

Figure 1: Heatmap of TF expression levels from K562 and GM12878 cell lines. TFs are named by their motif IDs on JASPAR database and are sorted by their expression levels in GM12878.

Going forward, we exclude the results from ENCFF285HUZ, as it is a gene quantification file, while the others are transcript quantifications. The correlation of replicates are very high within each cell line, for K562 the Pearson correlation coefficient is 0.9852 and for GM12878 it's 0.9958. The correlation coefficient between K562 and GM12878 is around 0.31.

After incorporating expression-weighted features to our logistic regression model, we observe that the performance is similar to without expression data. Using motif counts alone, the model predicted GM12878 chromatin accessibility with 83% accuracy; after adding expression-weighted motif counts, the accuracy is 80%. For K562 cell line, motif counts alone accurately predicted 77% of the test samples, while combined with expression-weighted features predicted 75%. The performance is repeatable across RNA-seq replicates (Fig 2, blue and green curves), thus in the future we plan to average the

RNA-seq expression data to train one model for each cell line.

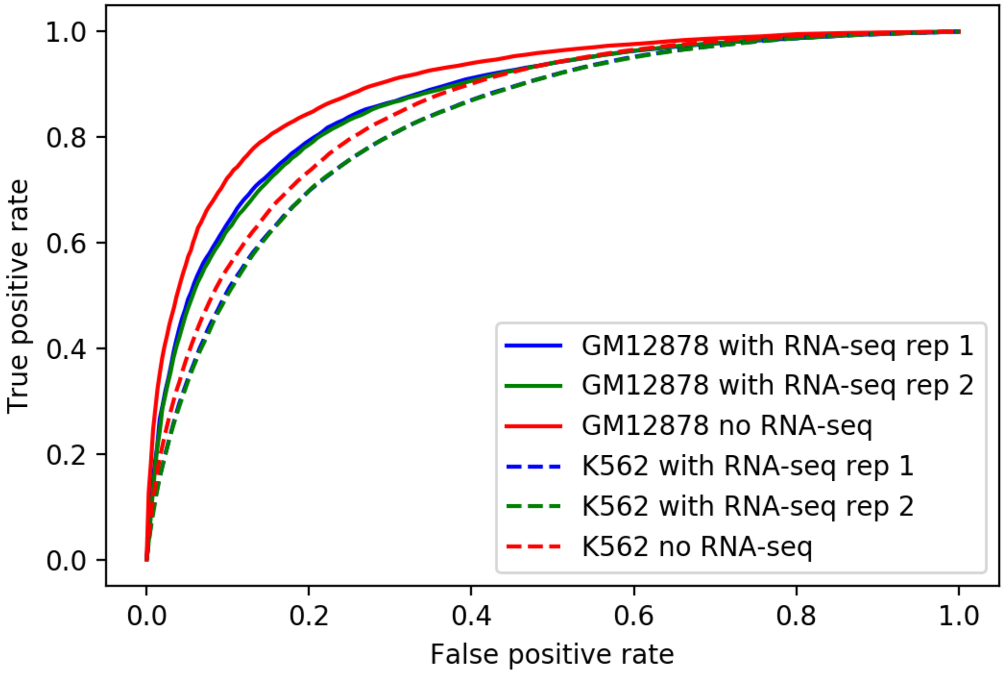

Figure 2: Performance of logistic regression models after incorporating RNA-seq data. For GM12878 and K562 cell lines, RNA-seq replicates show the same performance, with an area under the curve (AUC) of 0.87 for GM12878 and 0.83 for K562. Previous model that did not consider RNA-seq data gave an AUC of 0.90 for GM12878 and 0.85 for K562.

One advantage of logistic regression is that the coefficients assigned to each feature gives us an idea of whether a feature is predictive of open or closed chromatin. A positive coefficient indicates the feature can be predictive of open chromatin, whereas a negative coefficient implies that the feature predicts closed chromatin. In our previous work using only TF motifs to predict open chromatin, CTCF motif usually has one of the most positive coefficients across different cell types. This agrees with the current understanding that CTCF binds in open chromatin regions, especially open regions that are ubiquitous across different cell types [9]. Now, after incorporating TF expression data into the model, we still see CTCF motif among the top features predictive of open chromatin. In addition, expression-weighted CTCF motif is also among the top features indicative of open chromatin (data not shown). In general, we observe that the sign of motif count features and the corresponding expression-weighted features stay the same, which is expected. Further investigation is needed to compare whether expression-weighted motif counts are more robust for predicting open chromatin than motif counts alone.

6 Issues (optional)

The only issue we encountered so far was the difference in between gene and transcript quantification files for different cell lines. In the gene quantification files, the total expression across all transcripts is given, whereas the transcript quantification file provides transcript-by-transcript expression levels. As a result, if there are transcripts present in the gene quantification file that are not present in our list of transcripts, we are unable to use the value provided in the gene quantification file. Future analysis will be done on the transcript quantifications only.

Other than this, we've had no issues with the project, and have made no modifications on the original proposal.

7 Reference

1. Iwafuchi-Doi,M., Zaret, K. (2014) Pioneer transcription factors in cell reprogramming, *Genes & Dev.*, **28**, 2679-2692.

2. Lamparter, D., Marbach D., Rueedi, R., Bergmann, S., Kutalik, Z. (2014) Genome-wide association between transcription factor expression and chromatin accessibility reveals regulators of chromatin accessibility, *PLOS Comput. Biol.*, doi: 10.1371/journal.pcbi.1005311.

3. Sherwood, R.I., Hashimoto, T., O'Donnell, C.W., Lewis, S., Barkal, A.A., van Hoff, J.P., Karun, V., Jaakkola, T., Gifford, D.K. (2014) Discovery of directional and nondirectional pioneer transcription fac-

tors by modeling DNase profile magnitude and shape. *Nat. Biotechnol.*, **32**, 171-178.

4. Voong, L.N., Xi, L., Sebeson, A.C., Xiong, B., Wang, J.P., Wang, X. (2016) Insights into nucleosome organization in mouse embryonic stem cells through chemical mapping. *Cell*, **167** (6), 1555-1570.

5. Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J.A., van der Lee, R., Bessy A., Chneby, J., Kulkarni, S.R., Tan, G., Baranasic, D., Arenillas, D.J., Sandelin, A., Vandepoele, K., Lenhard, B., Ballester, B., Wasserman W.W., Parcy, F., Mathelier, A. (2018) JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.*, **46** (D1), D260-D266.

6. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, Garg K, John S, Sandstrom R, Bates D, Boatman L, Canfield TK, Diegel M, Dunn D, Ebersol AK, Frum T, Giste E, Johnson AK, Johnson EM, Kutyavin T, Lajoie B, Lee BK, Lee K, London D, Lotakis D, Neph S, Neri F, Nguyen ED, Qu H, Reynolds AP, Roach V, Safi A, Sanchez ME, Sanyal A, Shafer A, Simon JM, Song L, Vong S, Weaver M, Yan Y, Zhang Z, Zhang Z, Lenhard B, Tewari M, Dorschner MO, Hansen RS, Navas PA, Stamatoyannopoulos G, Iyer VR, Lieb JD, Sunyaev SR, Akey JM, Sabo PJ, Kaul R, Furey TS, Dekker J, Crawford GE, Stamatoyannopoulos JA. (2012) The accessible chromatin landscape of the human genome. *Nature*, **489**, 75-82.

7. Swinstead EE, Miranda TB, Paakinaho V, Baek S, Goldstein I, Hawkins M, Karpova TS, Ball D, Mazza D, Lavis LD, Grimm JB, Morisaki T, Grntved L, Presman DM, Hager GL. (2016) Steroid Receptors Reprogram FoxA1 Occupancy through Dynamic Chromatin Transitions. *Cell*, **165**, 593-605.

8. Song L, Zhang Z, Grasfeder LL, Boyle AP, Giresi PG, Lee BK, Sheffield NC, Grf S, Huss M, Keefe D, Liu Z, London D, McDaniell RM, Shibata Y, Showers KA, Simon JM, Vales T, Wang T, Winter D, Zhang Z, Clarke ND, Birney E, Iyer VR, Crawford GE, Lieb JD, Furey TS. (2011) Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome Res*, **21**, 1757-1767.