Introduction

Two useful assays for quantifying the regulatory potential of genomic regions are DNase-seq, which indicates regions of open chromatin, and ChIP-seq, which can be used to determine the locations of histone modifications and transcription factor binding. Previous studies have suggested that individual variants can disrupt accessibility, histone modifications, and transcription factor binding [1, 2, 3], implying that there are sequence signatures underlying the regulatory potential of a region. Since the sequence features immediately disrupted by genetic variants partially determine a sequence's regulatory potential, we think that using these features can help us learn which variants affect regulation, even though the 150 bp centered on the eQTL may have a different chromatin state in the MPRA than in its native context due to differences in distal sequences.

Convolutional neural networks [4] are a powerful machine learning model that, when given a set of sequences labelled with a biological quantity of interest, can automatically determine a set of motifs associated with the quantity and compose them into a prediction of that quantity for other, unlabelled sequences. This approach is likely to be more useful than identifying known motif matches because many trancription factors' binding preferences have never been assayed either *in vivo* or *in vitro*. In fact, convolutional neural networks such as DeepBind [5], DeepSEA [6] and Basset [7] have achieved state-of-the-art accuracy at predicting transcription factor binding and chromatin accessibility.

Gradient boosting [8] is a machine learning method for learning how a target quantity depends on a set of features across many examples, where the features are only weakly correlated with the target. The core idea of boosting is to incrementally build a model of the target over multiple rounds, in each round focusing on the examples that were predicted poorly in previous rounds. We think that combining this approach with convolutional neural networks that are weakly predictive of the effect of a genetic variant on regulation is likely to reveal which variants modulate expression.

Methods

emVar probabilities were predicted by an ensemble of three gradient boosting classification models trained with the scikit-learn library [9] on the 3,044 eQTLs in the provided training set. Each model used one of these feature sets:

1. Average regulatory genomics signals across the 150bp centered on the eQTL:
    (a) The average ChIP-seq signal for each TF and histone mark with data available for GM12878 from the ENCODE [10] and Roadmap Epigenomics [11] projects.
    (b) The average DNase-seq signal from GM12878 [11].
2. The intragenomic replicate (IGR) scores [12] for ChIP-seq experiments from ENCODE.
    (a) Scores were calculated for both 7-mers and 8-mers, and both sets of scores were used as features when training the model.
    (b) Predictions were made for each base in the 150 bp window centered on the reference allele of the eQTL.
    (c) Individual scores were fold enrichment over input averaged per TF across all present experiments, using only regions that fell within the DNase track of the corresponding cell type. Thus, only TFs assayed in cell types with DNase were used, resulting in 173 TF tracks.
    (d) For the skew estimation task, IGR scores were calculated for both reference and alternate alleles and the max was taken for each kmer that overlapped the SNP/indel of interest, as described previously [12]. In addition, features were calculated from the baseline affinities, probability of shuffling the overlapping context window and observing the given affinity by chance, and the correlation between orientations along the 400bp window centered at each kmer.

3. DNase hypersensitivity predictions in 164 cell lines from Basset:
    (a) Scores for both the reference and alternate allele sequences surrounding the eQTL (600 bp centered on the eQTL).
    (b) The absolute difference between the reference and alternate scores.

Allelic skew was predicted by an ensemble of three gradient boosting regression models trained on the same feature sets.

Regulatory hit probabilities for the reference and alternate alleles were predicted using an ensemble of three gradient boosting classification models trained on the above features (using both the sample dataset and the eQTLs from part 2 of the challenge, which are all known to be regulatory hits), along with the following five sequence-based models trained on external data:

4. A deep convolutional neural network based on the Basset architecture and trained with the Keras neural network library [13]. Our inputs to the model were the reference genome sequences underlying DNase-seq peaks that contain variants in dbSNP build 146 [14]. Our positive examples were sequences underlying DNase-seq peaks in GM12878, and our negative examples were sequences underlying peaks in other cell types. We considered only the 5% most significant uniformly reprocessed peaks in each cell type from the Roadmap Epigenomics project [11]. The sole input to the model was the 1000 bp sequence centered on each variant; the reverse complement of each sequence was provided as an additional training example. We thought that this classifier would be able to identify LCL-specific accessibility sequence signatures. To mitigate the imbalance between the number of positive and negative examples, the learning rate for each class was weighted in inverse proportion to the number of examples in the class. The model was trained for one epoch with the Adam optimizer [15] and early stopping, with validation taking place after each 128,000 training examples on a held-out 1% of the dataset. Minor modifications to the Basset architecture included the use of parametric rectified linear units [16] instead of rectified linear units and the placement of batch normalization [17] after rather than before rectification.

5. The same model as above, using H3K27ac ChIP-seq peaks instead of DNase-seq peaks.

6. The same model as above, using H3K9ac ChIP-seq peaks.

7. The same model as above, using H2A.Z ChIP-seq peaks.

8. The same model as above, using the union of DNase-seq peaks and all ChIP-seq peaks mentioned above.

An eQTL's regulatory hit probability was defined as the maximum of the probabilities for its reference and alternate alleles. Log expression for the reference and alternative alleles was approximated by regulatory hit probability. Standard deviations were estimated by the standard deviation of the predictions across classifiers in the ensemble.

In addition to submitting predictions for the full ensemble of eight models described above, our six submissions include ensembles of the following subsets of the eight models:

Submission	Part 1 models	Part 2 models
1	1	1
2	2	2
3	3	3
4	1, 2, 3	1, 2, 3
5	1, 2, 3, 4	1, 2, 3
6	All 8	1, 2, 3

References

[1] Jacob F Degner et al. "DNase [thinsp] I sensitivity QTLs are a major determinant of human expression variation". In: *Nature* 482.7385 (2012), pp. 390–394.

[2] Zhihao Ding et al. "Quantitative genetics of CTCF binding reveal local sequence effects and different modes of X-chromosome association". In: (2014).

[3] Fabian Grubert et al. "Genetic control of chromatin states in humans involves local and distal chromosomal interactions". In: *Cell* 162.5 (2015), pp. 1051–1065.

[4] Yann LeCun et al. "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.

[5] Babak Alipanahi et al. "Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning". In: *Nature Biotechnology* (2015).

[6] Jian Zhou and Olga G Troyanskaya. "Predicting effects of noncoding variants with deep learning-based sequence model". In: *Nature Methods* 12.10 (2015), pp. 931–934.

[7] David R Kelley, Jasper Snoek, and John Rinn. "Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks." In: *bioRxiv* (2015), p. 028399.

[8] Jerome H Friedman. "Greedy function approximation: a gradient boosting machine". In: *Annals of Statistics* (2001), pp. 1189–1232.

[9] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[10] ENCODE Project Consortium et al. "The ENCODE (ENCyclopedia of DNA elements) project". In: *Science* 306.5696 (2004), pp. 636–640.

[11] Anshul Kundaje et al. "Integrative analysis of 111 reference human epigenomes". In: *Nature* 518.7539 (2015), pp. 317–330.

[12] Richard Cowper-Sal·lari et al. "Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression". In: *Nature Genetics* 44.11 (2012), pp. 1191–1198.

[13] Francois Chollet. *Keras*. `https://github.com/fchollet/keras`. 2015.

[14] Stephen T Sherry et al. "dbSNP: the NCBI database of genetic variation". In: *Nucleic Acids Research* 29.1 (2001), pp. 308–311.

[15] Diederik Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[16] Kaiming He et al. "Delving deep into rectifiers: Surpassing human-level performance on Imagenet classification". In: *arXiv preprint arXiv:1502.01852* (2015).

[17] Sergey Ioffe and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift". In: *arXiv preprint arXiv:1502.03167* (2015).