TF binding models

QBiC predictions are based on quantitative TF binding models trained on universal protein binding microarray (PBM) data from the following databases:

* UniPROBE: http://the\_brain.bwh.harvard.edu/uniprobe/
* CisBP: http://cisbp.ccbr.utoronto.ca/
* PBM data generated in the Gordan laboratory (reported in Shen et al. Cell Systems 2018, or included with the release of QBiC)

We downloaded the entire database from CisBP (v1.02) consisting of 3,342 universal PBM data sets. Next, we added 245 PBM data sets from the UniPROBE database, which were not included in CisBP, and 22 PBM data sets generated in our lab, resulting in a total of 3,609 PBM data sets. Then we annotated the quality of each uPBM data based on: 1) the number of 8-mers with binding enrichment score (E-scores) exceeding 0.45; and 2) the performance of our models trained on the data, which enabled us to filter out low quality PBM data sets. After the filtering process, we utilized The Human Transcription Factors Project[cite1], as well as manual curation, to map PBM data sets to human TFs. By keeping only the best quality or the highest amino-acid identity PBM data sets for each TF, we finally arrived at a total of 666 data sets characterizing the binding of 577 human TFs. The full mapping of human TFs to available PBM data sets is available here.

Our TF binding models are linear regression models trained on universal PBM data[cite2]. We use the counts of 6-mers or gapped 6-mers (with a gap of 1 or 2 bases) as features, and the binding intensity signal as the outcome variable. Then, we use ordinary least squares (OLS) to estimate the coefficients for all 6-mer features, as well as the distributions of these estimates. The effect of each mutation can be written as a linear combination of all the coefficients of 6-mers covering it and can therefore be estimated as well as tested for statistical significance, as described in detail in our previous work[cite2]. Thus, for each variant and each PBM data set for a TF protein, our model predicts:

* **Difference** = the difference in TF binding (i.e. in PBM signal) between the two alleles
* **z-score** = the normalized difference in TF binding, computed based on the estimated parameter distributions; the z-scores are comparable among different TFs and data sets
* **p-value** = the significance of the predicted difference, computed based on the estimated parameter distributions; the p-values depend on the magnitude of difference, as well as the quality of the model and the training data.

[[[the citations below can simply be links to the paper websites. No need to put the actual citations in the about page.]]

[cite1]Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, Chen X, Taipale J, Hughes TR, Weirauch MT.(2018) [The Human Transcription Factors](https://www.sciencedirect.com/science/article/pii/S0092867418301065?via%3Dihub). Cell. 172(4):650-665. doi: 10.1016/j.cell.2018.01.029. Review.

[cite2] Zhao J, Li D, Seo J, Allen AS, Gordân R. Quantifying the Impact of Non-coding Variants on Transcription Factor-DNA Binding. *Res Comput Mol Biol*. 10229:336-352