Motif signatures of transcribed enhancers

Dimitrios Kleftogiannis[1], Haitham Ashoor[2], Nikolaos Zarokanellos[3], and Vladimir B. Bajic[2,*]

[1] Centre for Evolution and Cancer, Division of Molecular Pathology, The Institute of Cancer Research (ICR), London, SW7 3R, United Kingdom

[2] Computational Bioscience Research Center (CBRC), Computer, Electrical and Mathematical Sciences and Engineering Division (CEMSE), King Abdullah University of Science and Technology (KAUST), Thuwal, 23955-6900, Saudi Arabia

[3] Red Sea Research Center (RSRC), Biological and Environmental Sciences & Engineering Division (BESE), King Abdullah University of Science and Technology (KAUST), Thuwal, 23955-6900, Saudi Arabia

* Corresponding author:
Vladimir B. Bajic Tel: +966 (12) 808-2386; Fax: +966 (12) 808-2386; Email: Vladimir.bajic@kaust.edu.sa

SUPPEMENTARY MATERIAL

TELS additional implementation details

To achieve more robust results and provide a more comprehensive view of the recognition performance, we repeat the learning process for every classification problem 300 times, and compute the average classification performance for every combination of top-ranked features. This process guarantees stable selection of combinations of features since it is based on the average classification performance of multiple runs that involve random splits of the input data. In summary, we generate the following classification problems:

1. All-facets vs. all-facets random controls: 112 cell types/tissues, and for each cell type/tissue we evaluate 346 combination of motifs using 300 random splits of the data.
2. Robust set vs. robust random controls: for this set we evaluate 346 combinations of motifs using 300 random splits of the data.
3. Exclusively transcribed vs. exclusively transcribed negative controls: 96 cell types/tissues, and for each cell type/tissue we evaluate 346 combination of motifs using 300 random splits of the data.

Classification performance metrics

To assess the classification performance and identify combinations of motifs that minimize classification error we consider the following performance metrics:

(1) $GM = \sqrt{Sensitivity * Specificity}$

with

$$Sensitivity = \frac{TP}{TP + FN} \text{ and } Specificity = \frac{TN}{TN + FP}$$

(2) $$PPV = \frac{TP}{TP + FP}$$

(3) $$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}},$$

where TP, FP, FN, TN, GM, PPV, and MCC denote True Positives, False Positives, False Negatives, True Negatives, Geometric mean of Sensitivity and Specificity, Positive Predictive Value, and Mathews Correlation Coefficient, respectively.

Dealing with the class imbalance problem

To test the impact of class-imbalance on the classification performance we focus on the 'robust set' of TrEn. We repeat the learning process using various ratios between positive and negative samples. We test ratios 1:1, 1:2, 1:3 up to 1:10, which means that we progressively increase the number of negatives. For every run we measure the classification performance and identify the most informative sets of motifs. Our experimentation shows that as we increase the number of negatives, the selection of features appears quite consistent with small discrepancies on the selected sets, but we observe a drop on the classification performance. Supplementary Figure 4 shows the levels of MCC, as the most indicative performance metric. Apparently for ratios 1:1 to 1:5 (i.e., lowly unbalanced sets) the MCC levels have standard deviation of 0.05 MCC, which is low. Unfortunately, the MCC levels drop much

more for the highly unbalanced cases (i.e., ratios 1:6 to 1:10). From all the above we conclude that the ratio 1:1 gives us the highest classification performance.

However, we would like to note that the objective of this study is not to predict TrEn in a genome-wide scale. In contrary, TELS addresses the problem of identifying motif signatures that allow effective characterization of TrEn, and to explore the degree to which different combinations of short nucleotide motifs operate in a context-specific manner. As importantly, TELS is neither a general feature selection method for imbalanced datasets nor a genome-wide de novo TrEn predictor.

In fact, to develop an effective de-novo enhancer predictor, imbalance-learning techniques are useful (e.g., SMOTE or ensemble learning) since the non-enhancer sequences outnumber real enhancers (as far as we know). A good practise to achieve high sensitivity and specificity in a genome wide scale is to learn models with some 'realistic' ratio between enhancers and non-enhancers. However, this is not the case for the discrimination problem we present in TELS.

Comparison of Gini-index to alternative FS methods

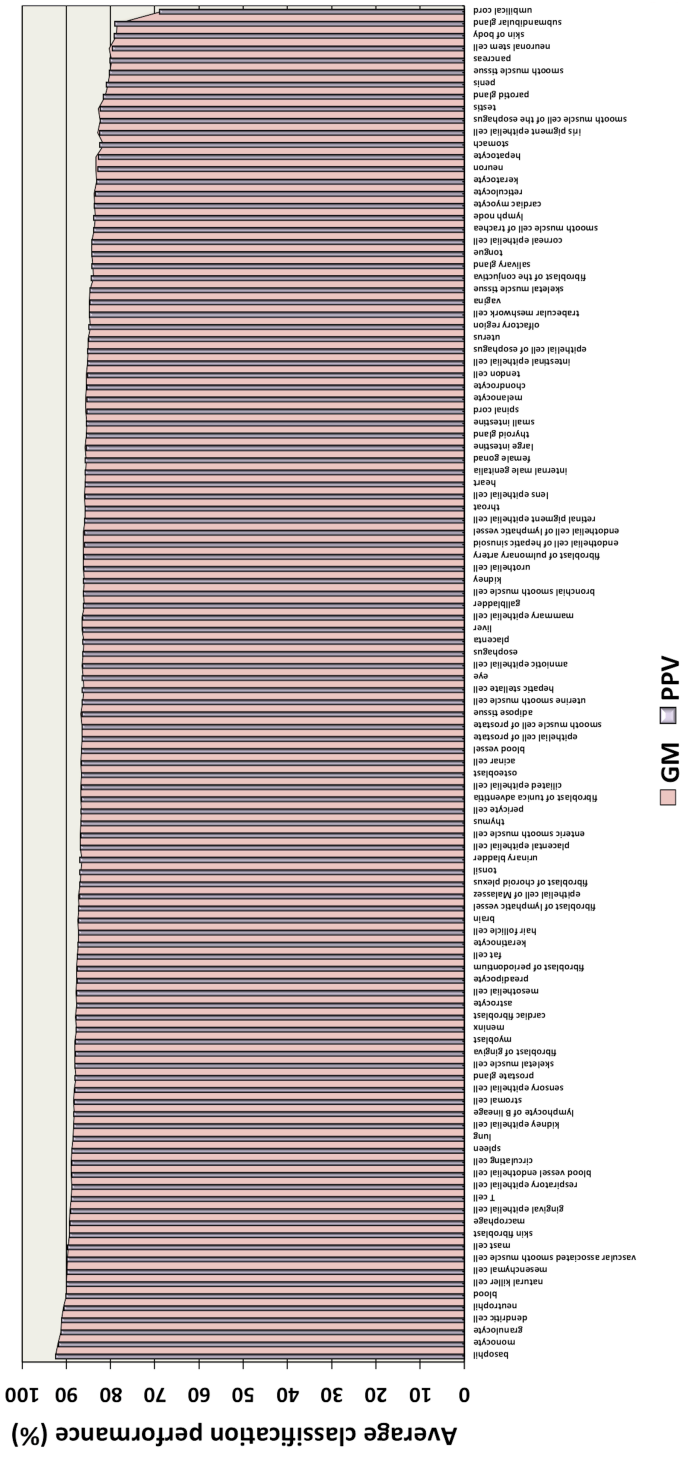
We compare the recognition performance of LR with Gini index using two other state-of-the-art algorithms for FS, namely minimum redundancy maximum relevance criterion (mRMR) and Fisher's test-based FS. To conduct a fair comparison, we follow the same protocols summarized in Supplementary Figure 3. For all competitor methods we repeat the feature ranking 300 times using a random subset of the data samples equal to 20%. From them, we select the most frequent ranking as the best. We point here, that the feature ranking is a completely independent process that does not involve any interaction with classifier or other algorithm.

After applying individual feature ranking, we use greedy selection method and estimate the classification performance using all top-N ranked feature subsets where N equals 1,2,3,…, 346. We repeat the learning process 300 times, and we select the top-ranked subset that achieves the maximum MCC. All implementations are made in Matlab R2014b using the FEAST library for FS. Results obtained by mRMR and the Fisher's test with LR are summarized in Supplementary Figures 5 and 6 for all cell type/tissue specific enhancers included in the 'all-facets' dataset. FS using a combination of LR with Gini index achieves higher recognition performance in almost all of the tested

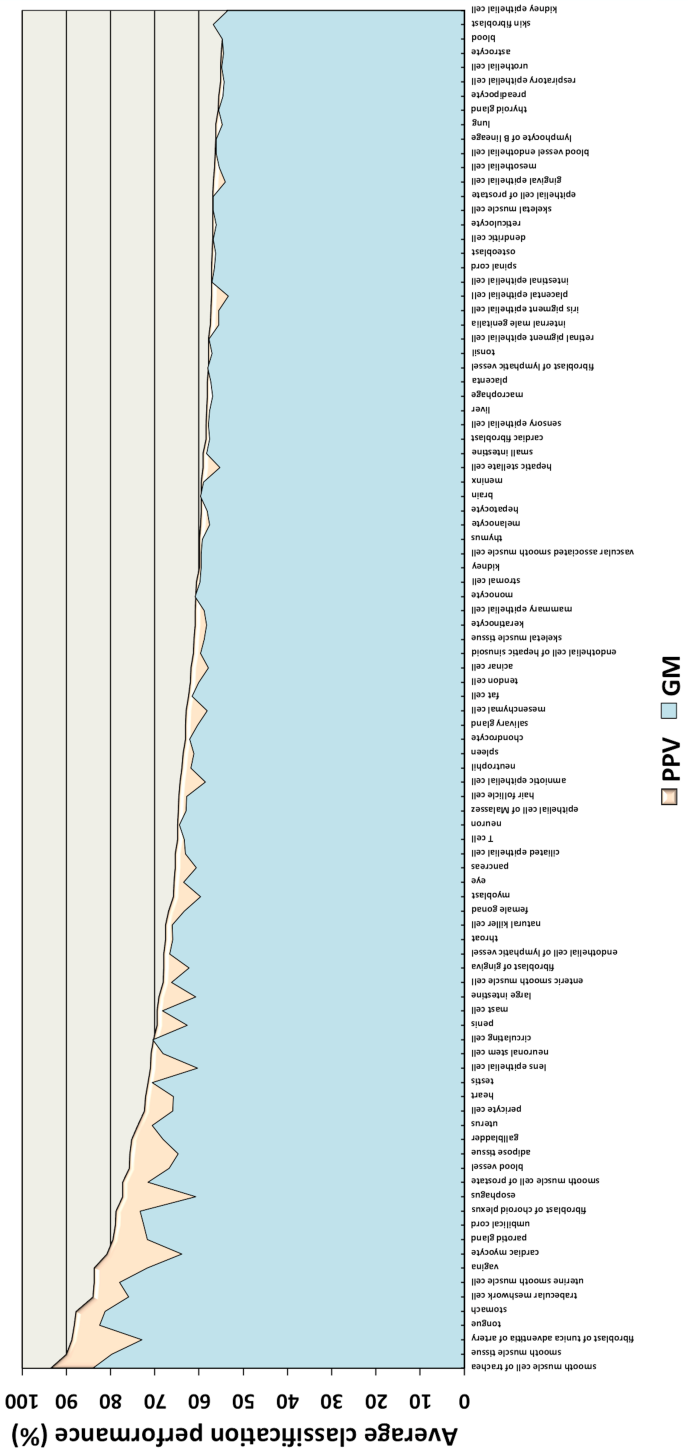
cases. In particular, FS based on Gini index achieves an average PPV of 85.94% and MCC of 0.72 across all studied tissues and cell-types. On the other hand, FS by mRMR achieves average PPV of 84.05% and MCC of 0.67, whereas FS by Fisher's test achieves average PPV of 85.13% and MCC of 0.71, respectively. This clearly suggests that FS that combines LR with Gini index under the greedy forward selection is the best choice.

SUPPLEMENTARY FIGURES

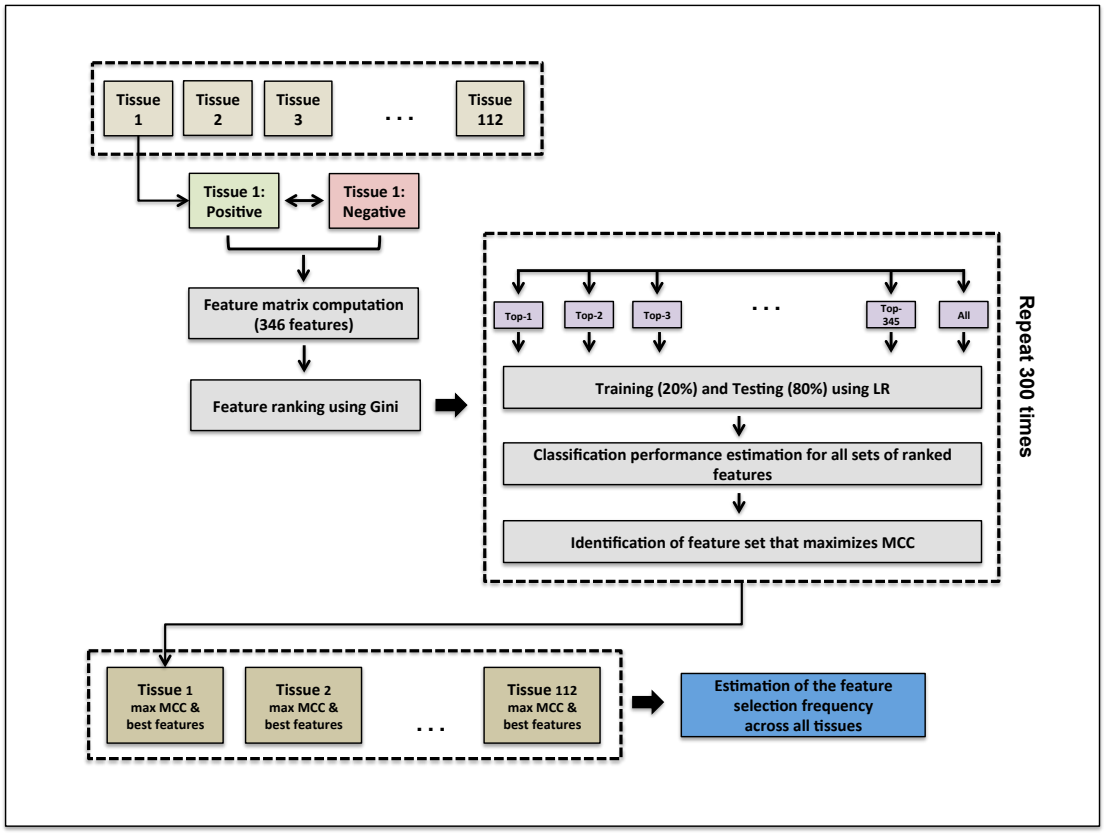
Supplementary Figure 1: Discrimination of TrEn from 'all-facets' dataset versus non-enhancers: Classification performance indicated by GM(%) and PPV(%) using the most informative combinations of motifs for 112 cell types/tissues.



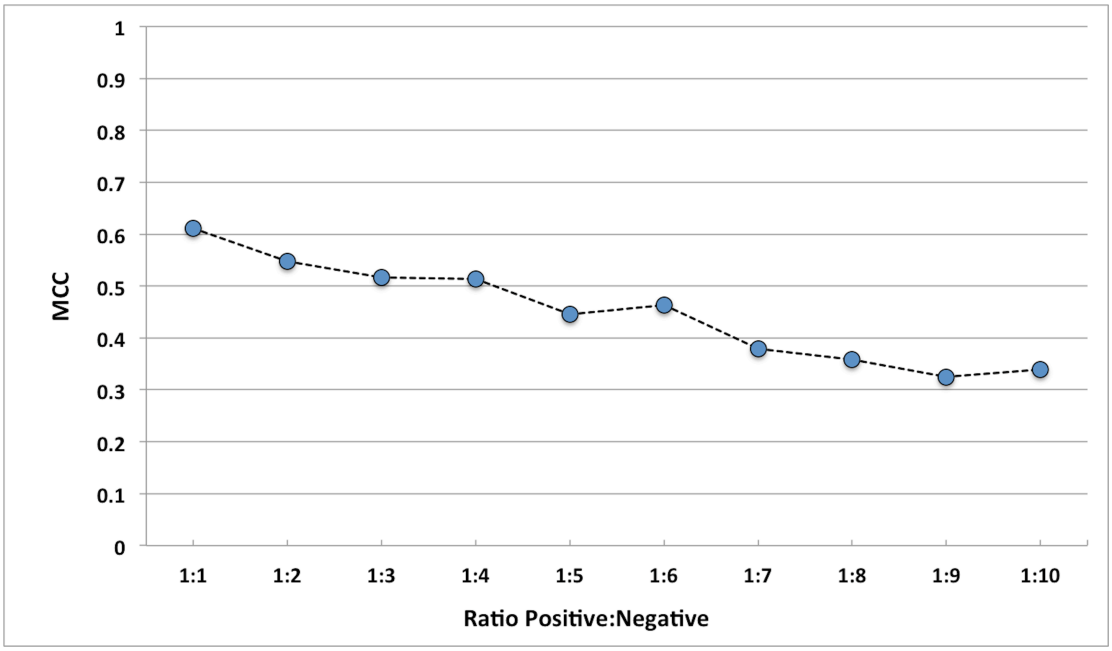
Supplementary Figure 2: Discrimination of TrEn expressed in only one cell-type or tissue from all other exclusively transcribed TrEn: Classification performance indicated by GM(%) and PPV(%) using the most informative combinations of motifs for 96 cell types/tissues.



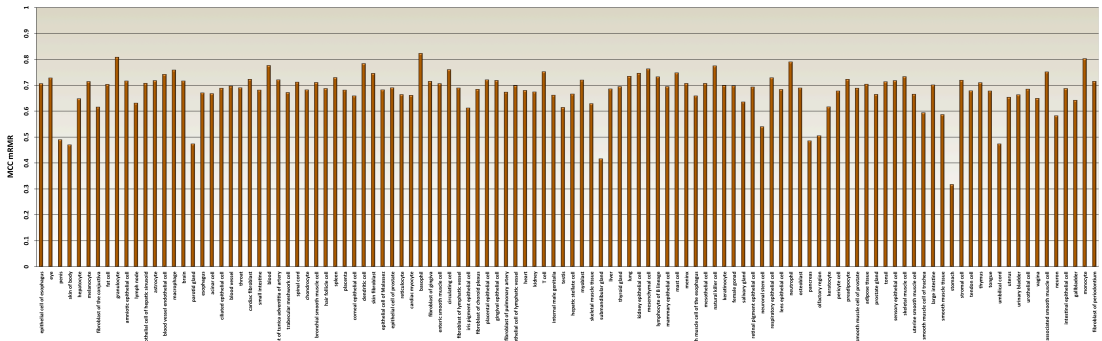
Supplementary Figure 3: Flowchart of TELS.



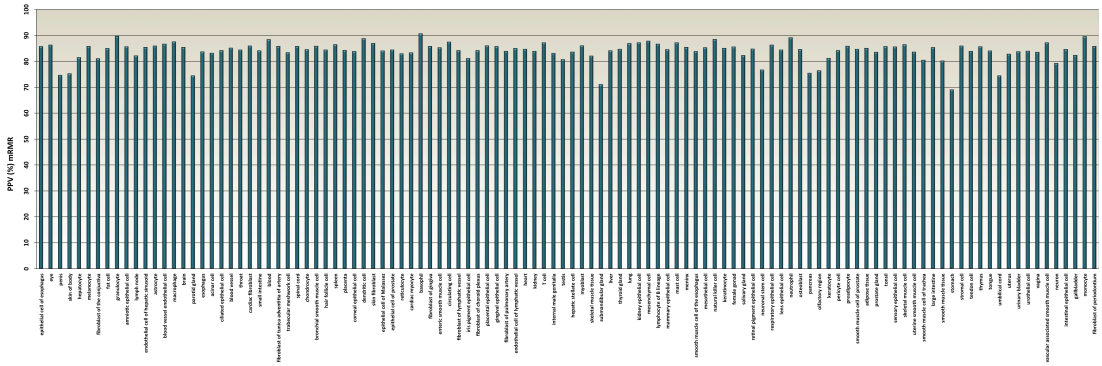
Supplementary Figure 4: The effect of class imbalance between positive and negative data on the classification performance.



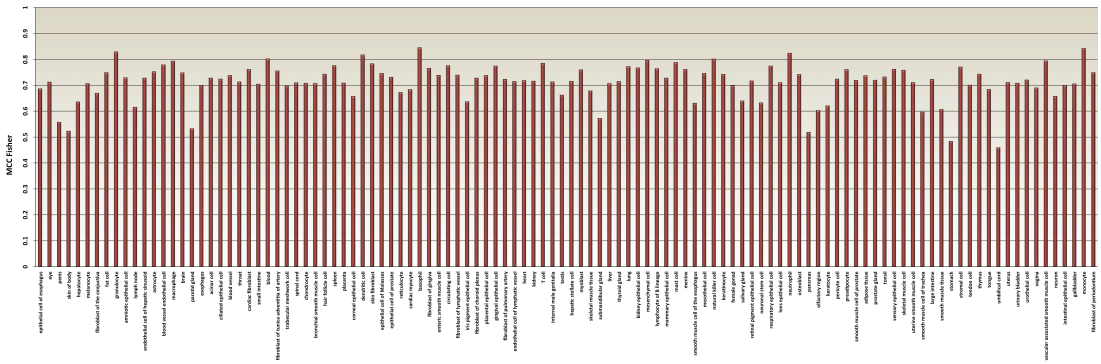
Supplementary Figure 5a: Classification performance indicated by MCC using optimized sets of nucleotide markers selected by mRMR for all tissues and cell-types from FANTOM5 (all-facets dataset).



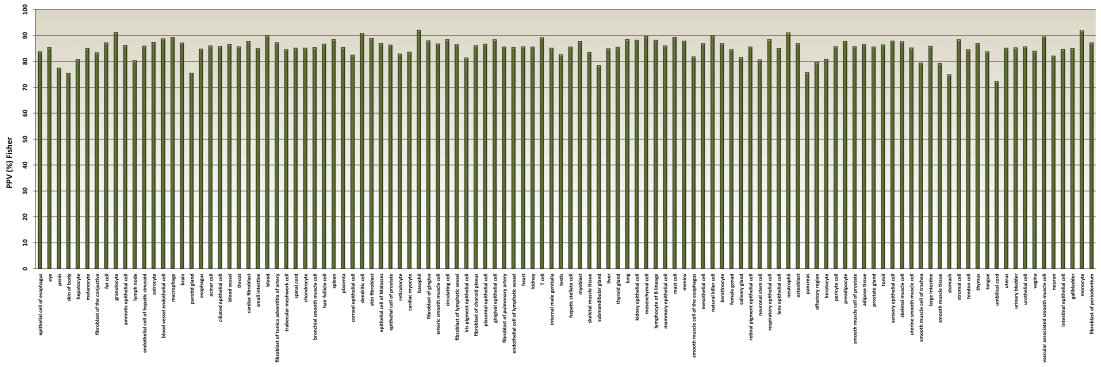
Supplementary Figure 5b: Classification performance indicated by PPV (%) optimized sets of nucleotide markers selected by mRMR for all tissues and cell-types from FANTOM5 (all-facets dataset).



Supplementary Figure 6a: Classification performance indicated by MCC optimized sets of nucleotide markers selected by Fisher for all tissues and cell-types from FANTOM5 (all-facets dataset).



Supplementary Figure 6b: Classification performance indicated by PPV (%) optimized sets of nucleotide markers selected by Fisher for all tissues and cell-types from FANTOM5 (all-facets dataset).



Supplementary Figure 7: Discrimination of TrEn from 'all-facets' dataset versus random controls: (a) Dendrogram of the hierarchical cluster tree constructed from the motif set similarity matrix; (b) Dendrogram of the hierarchical cluster tree constructed from input sequence set similarity matrix.