Exposing and Characterizing Subpopulations of Distinctly Regulated Genes by K-Plane Regression [⋆]

Fabrizio Frasca[1][0000−0002−5165−1394],
Matteo Matteucci[1][0000−0002−8306−6739],
Marco J. Morelli[2,3][0000−0003−1862−667X], and
Marco Masseroli[1][0000−0003−2574−1174]

[1] Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano,
20133 Milan, Italy
{fabrizio.frasca@mail., matteo.matteucci@, marco.masseroli@}polimi.it
[2] Center for Genomic Science of IIT@SEMM, Istituto Italiano di Tecnologia (IIT),
20139 Milan, Italy
[3] Current address: Center for Translational Genomics and Bioinformatics, IRCCS
San Raffaele Scientific Institute, Via Olgettina 58, 20132 Milan, Italy
morelli.marco@hsr.it

Abstract. Understanding the roles and interplays of histone marks and transcription factors in the regulation of gene expression is of great interest in the development of non-invasive and personalized therapies. Computational studies at genome-wide scale represent a powerful explorative framework, allowing to draw general conclusions. However, a genome-wide approach only idientifies generic regulative motifs, and possible multi-functional or co-regulative interactions may remain concealed. In this work, we hypothesize the presence of a number of distinct subpopulations of transcriptional regulative patterns within the set of protein coding genes that explain the statistical redundancy observed at a genome-wide level. We propose the application of a K-Plane Regression algorithm to partition the set of protein coding genes into clusters with specific shared regulative mechanisms. Our approach is completely data-driven and computes clusters of genes significantly better fitted by specific linear models, in contrast to single regressions. These clusters are characterized by distinct and sharper histonic input patterns, and different mean expression values.

Keywords: Gene expression · Epigenetic transcriptional regulation · K-Plane Regression

1 Background

In both complex and simple organisms, the regulation of gene expression is crucial in allowing cellular differentiation and response to environmental stimuli.

[⋆] This work was supported by the Advanced ERC Grant "Data-Driven Genomic Computing (GeCo)" project (2016-2021), funded by the European Research Council.

Among the many layers of gene regulation, those occurring at the stage of the initiation of transcription are considered to be the most flexible and effective [10]. Transcriptional regulation typically act at the epigenetic level, i.e., without any modification of the underlying sequence, rather with a combination of binding of regulating molecules (transcription factors, or TF) to the DNA, and changing the structure of chromatin through the addition or removal of chemical residues on histone molecules (histone marks, HM).

Many studies have revealed the implications of epigenetic regulations: for example, their oncogenic role played in cancer etiology by gene expression alterations [12]. At the same time, epigenetic interactions can be targeted by non-invasive, promising therapeutic possibilities, leveraging on their intrinsic reversibility [2]. These premises have contributed to the birth of the field of targeted cancer therapy, where epigenetic approaches could be used to treat cancer in a personalized manner, by kick-starting specific immune responses, or bringing back gene expression to physiological levels. Understanding the fundamental mechanisms by which histone marks and transcription factors operate to regulate the expression of specific genes is then of paramount interest as it is the necessary prerequisite in order to design effective and precise "epigenetic" drugs.

Next-generation sequencing (NGS) technologies nowadays routinely allow genome-wide measurements of gene expression and epigenetic signals in the cell lines or tissues of interest [8]. Large datasets are publicly available in repositories generated by multinational projects, such as ENCODE [6] and Roadmap Epigenomics [7]. One can now leverage on these large collections of data to quantatively model the processes of interests and understand the specific roles, interplays and effects of epigenetic transcriptional regulators.

In particular, several statistical models have been conceived to study the association between gene-related epigenetic signals and messenger RNA (mRNA) abundance at a genome-wide scale [5]. Within this context, the problem is usually framed as a regression or classification task, where all the protein-coding genes are *samples*, signals from HMs and/or binding of TFs are *input features* and the aim is to predict the *response value*, i.e., either mRNA levels (regression), or activities of genes (binary classification).

The relevance of genome-wide modeling resides in its omnicomprensive, explorative, as general conclusions can be drawn about the role and interplay of TFs and HMs. Interestingly, if at this level of resolution such features have been shown to be predictive for mRNA abundance [5], they have also been observed to exhibit certain *statistical redundancy* within themselves. In [4], the genome-wide resolution level itself has been addressed as the main cause for such observed redundancy in the regression task, as variations in the relative predictive power of TFs and HMs are observed at the finer resolution of groups of ontology-classified biological processes.

However, the work in [4] resorts to manually curated external sources of information; in contrast, in the case of investigative analyses it is useful to let conclusions directly arise from data, including the least possible prior knowledge. This is, for instance, the case of targeted cancer therapy and personalized medicine, where the main objects of research are the possibly unknown alterations in epigenetic patterns and anomalies in their regulative effects: here, it is essential for statistical modeling to be data-driven.

So far, only few attempts have tried overcome the observed statistical redundancy observed in [4] in a data-driven manner: in [11] a mixture of Bayesian linear elastic nets revealed to better fit transcriptional regulation w.r.t. a single regression model and assign a distinct predictive relevance to epigenetic features. With this modeling approach, the statistical redundancy addressed in [4] is further specified in terms of feature *multi-functionality*: the epigenetic-transcriptional association is better modeled with an ensemble of models where HMs and TFs may assume different roles. However, even though in [11] the models accounting to the mixture are distinctly defined, genes in the dataset are only softly clustered, as the expression for a gene is the weighted sum of the outputs of all models.

As the 'soft' approach hinders the interpretation of the results, in this work we hypothesize the presence of a number of distinct, heterogeneous subpopulations within the set of protein coding genes, with different transcriptional regulative behaviors. Accordingly, we perform a hard partitioning of the whole gene set in a data-driven manner, defining clusters where specific linear regression models are fitted to learn the regulative dynamics of each gene sub-group. In this setting, our aim is not just the enhancement of the fitting of a global model, but rather the definition of a completely data-driven and fully interpretable procedure to overcome the aforementioned statistical redundancy [4] and feature multi-functionality [11], yielding well-defined, hard clusters of protein coding genes sharing specific dynamics of epigenetic transcriptional regulation. With our approach, a one-to-one association between linear models and gene clusters follows, and interpretative analyses are supported at best: regulative patterns can be investigated both at a gene-specific level and, statistically, at a gene-cluster level, and the regulative behavior can be a posteriori matched with the most-represented biological processes within a group.

Our problem is therefore similar to a *piecewise linear affine model fitting*, usually approached with *hinging hyperplane* [3] or *bounded error* [1] methods. However, our main goal is not to fit a supposed non-linear dynamic with piecewise linear functions, but rather learning different linear models in a scenario where dynamics are likely to be overlapped, discontinuous, and partially lying on sub-dimensional manifolds.

A more suited approach is then represented by *K-Plane Regression* [9], firstly introduced in the general context of fitting possibly discontinuous functions with an ensemble of linear models. The method is based on a clustering approach: it finds a fixed number (K) of hyperplanes such that each point in the training set is close to one of the hyperplanes, and all points in a partition are as close as possible in the input feature space. Resulting hyperplanes are found by minimizing the following objective function with an Expectation-Maximization (EM) algorithm:

$$E(\Theta) = \sum_{k=0}^{K-1} \sum_{i \in \Theta(k)} (t_i - \tilde{\boldsymbol{w}}_k^T \tilde{\boldsymbol{x}}_i)^2 + \gamma \|\boldsymbol{x}_i - \boldsymbol{\mu}_k\|_2^2, \tag{1}$$

where K is the pre-defined number of clusters, Θ defines the partitioning over the dataset – with $\Theta(k)$ the set of samples in cluster k, $\tilde{\boldsymbol{x}}_i$ and t_i are, respectively, the input feature and the target value for sample $i \in \Theta(k)$, $\boldsymbol{w}_k$ is the weight vector of the least square solution for those points, $\boldsymbol{\mu}$ terms refer to centroids in the feature space and γ is a user-defined parameter deciding the relative weight of the two additive terms in the objective function. The 'tilde' notation is used to indicate the inclusion of the bias term in the regression.

Given the capability of K-Plane Regression to fit discontinuous functions and the increased flexibility offered by a clustering approach, we built upon this last piece of work to solve our problem of modeling epigenetic transcriptional regulation with a hard ensemble of linear regression models.

The remainder of this paper is organized as follows. In Section 2, we discuss the data used and the techniques applied: we describe the epigenetic features, data sources and pre-processing we considered, and the specific versions of K-Plane Regression algorithm we employ in our pipelines. In section 3 we present the results, specifically focusing on the models obtained: we illustrate their predictive accuracy and we describe the regulative patterns. The conclusions are addressed in Section 4.

2 Materials and Methods

This work aims at modeling epigenetic transcriptional regulation with a hard ensemble of linear regression models, each one explaining mRNA levels as a function of epigenetic signals for a specific gene sub-group, i.e., a cluster of genes.

2.1 Biological Setting and Data

All considered measurements are over the chronic myeloid leukemia K562 immortalized cell line (human blood tissue), and only involve protein coding genes. GENCODE v10 reference annotation for the hg19 assembly was used to retrieve their transcription start sites (TSSs). The Roadmap Epigenomics Mapping Consortium's (REMC) [7] repository was chosen as the only data source in this work.

Genes are epigenetically characterized by data in the form of processed ChIP-Seq peaks for all the $m = 12$ histone modifications measured in K562 cell line within the REMC proejct (no TF was accounted for). Peaks for the considered HMs were retrieved in the formats of 'bed narrowPeak' or 'bed broadPeak', according to the nature of the considered mark, i.e., TF-like (sharp) marks or broad marks. These choices are summarized in Table 1. The epigenetic status of the generic gene g is represented as an m-dimensional input vector $\boldsymbol{x}_g$, whose

Table 1: HMs and their chosen data format.

Marks	Format
H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K9me1, H3K27ac, H2A.Z	narrowPeak
H3K9me3, H3K27me3, H3K36me3, H3K79me2, H4K20me1	broadPeak

elements contain the maximum peak enrichment value attained within a symmetric window region of 10 kbases centered on g's TSS, thus summarizing the g-related status of a specific monitored HM. In accordance to [5], signals closer to the genes' TSSs (roughly, within promoters) are the most informative to predict gene expression. These vectors, considered together for all our $n = 19,794$ genes, form an input matrix X, with dimensions $n \times m$.

For the transcriptional characterization of genes, we consider mRNA quantifications, measured with RNA-sequencing. The transcriptional status of gene g is encoded by $t_g = \sqrt{\log(1 + \tau_g)}$, where τ_g is the original mRNA quantification, log represents the natural logarithm and the application of two sub-linear, monotonically increasing functions aims at reducing the heteroskedasticity in regression residuals. Finally, consider t_g as the $(g+1)$-th element in n-dimensional target vector T collecting the transcriptional statuses of all the genes.

Together, X and T form our dataset $D = \langle X, T \rangle$, which is going to be partitioned by K-Plane Regression algorithm.

2.2 K-Plane Regression

The K-plane algorithm we used in our work is designed to minimize the following objective function:

$$E(\Theta) = \sum_{k=0}^{K-1} \sum_{i \in \Theta(k)} (t_i - \tilde{\boldsymbol{w}}_k^T \tilde{\boldsymbol{x}}_i)^2, \tag{2}$$

where we dropped the second additive term $\gamma \|\boldsymbol{x}_i - \boldsymbol{\mu}_k\|_2^2$ originally present in Equation 1, which enforces points belonging to the same partition to be close to each other. Such a 'closeness' term was explicitly introduced in [9] to avoid EM finding sub-optimal solutions, and to force the obtained partitions not to contain points from disjoint regions of the input feature space. In our application, where the objective is the disambiguation of overlapped expression dynamics from heterogeneous gene subpopulations, not only the l2-distance measure might not be suitable over histonic inputs, but also partitions spanning disjoint feature space regions are not necessarily to be avoided. Concerning the issue of possible sub-optimality, we just resorted to a simple multiple re-initialization technique.

The K-Plane Regression procedure is run R times; each time it starts by a random partition and optimizes Equation 2 as described in [9], i.e., by iteratively alternating a Maximization step – hyperplanes to clusters fitting – and an Expectation step – gene-cluster reassignments. A tolerance parameter τ is used to verify numerical convergence of the optimization procedure. Initializations are designed to construct a completely random partitioning made up of equally-sized clusters. In the end, among the R solutions obtained, the one attaining the best objective value is returned.

A draft of our K-Plane Regression pipeline is reported below.

```
procedure ClusterHyperplanes($\mathcal{D}, K, \tau, R$)
   $\mathcal{S} \leftarrow$ empty solution dictionary
   $\mathcal{O} \leftarrow$ array of $R$ elements
   for $r = 0, 1, \ldots, R-1$                                    ▷ perform $R$ runs
      $\Theta_0 \leftarrow$ RandomInit($|\mathcal{D}|, K$)               ▷ randomly initialize
      $s, o \leftarrow$ K-PlaneRegression($\mathcal{D}, K, \Theta_0, \tau$)          ▷ run K-Plane
      $\mathcal{S}[r] \leftarrow s$
      $\mathcal{O}_r \leftarrow o$
   $best \leftarrow \arg\max_r \mathcal{O}_r$                                  ▷ choose best
   return $\mathcal{S}[best]$
```

```
procedure K-PlaneRegression($\mathcal{D}, K, \Theta, \tau$)
   $\mathcal{X}, \mathcal{T} \leftarrow$ input matrix and target vector from $\mathcal{D}$
   $o^{prev} \leftarrow \infty$
   $o^{curr} \leftarrow E(\Theta)$
   while ($o^{prev} - o^{curr} > \tau$)                ▷ loop EM until convergence
      $\mathcal{M} \leftarrow$ empty model dictionary
      for ($k = 0, 1, \ldots, K-1$)                          ▷ Maximization
         $\mathcal{M}[k] \leftarrow$ OLS solution over $\langle \mathcal{X}[\Theta(k)], \mathcal{T}[\Theta(k)] \rangle$
      $\Theta \leftarrow$ empty cluster assignment
      for ($i = 0, 1, \ldots, |\mathcal{D}|-1\}$)                       ▷ Expectation
         $\Theta[i] \leftarrow \arg\min_k |\mathcal{M}[k](\mathcal{X}_i) - \mathcal{T}_i|$
      $o^{prev} \leftarrow o^{curr}$
      $o^{curr} \leftarrow E(\Theta)$
   return $\Theta$, $o^{curr}$
```

```
procedure RandomInit($n, K$)
   $\sigma \leftarrow$ random permutation of $[0, 1, \ldots, n-1]$
   $\Theta \leftarrow$ empty cluster assignment
   for ($k = 0, 1, \ldots, K-1$)
      $\Theta[k] \leftarrow$ k$^{th}$ subpart of $\sigma$
   return $\Theta$
```

3 Results

In our experimentsm,the procedure ClusterHyperplanes has been run with the following parameters: $\mathcal{D} = D, K \in [2 \dots 6], \tau = 0.1, R = 30$. The parameter K ranges in $[2 \dots 6]$ as the best value is hardly guessable a priori and might depend on the nature of the specific problem. Better solutions, in terms of cost functions, have been observed for larger values of K: Figure 1 depicts the trend of the convergence objective value as a function of this parameter. In the following, results for the value $K = 4$ are discussed: this choice is less prone to overfit spurious correlations, still yielding a good value of the objective function. In other words, it represents a convenient trade-off between goodness of fit and biological interpretability, given the current knowledge about HM (co-)activity, .

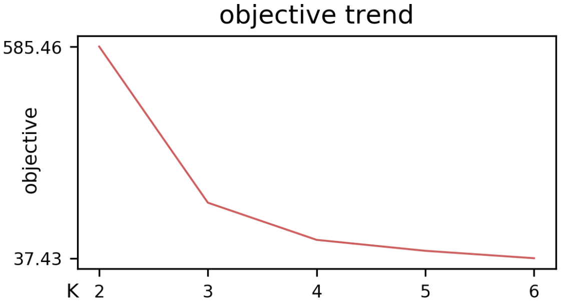

Fig. 1: Values of best solutions (objective) from re-initialized K-Plane Regression as a function of number of clusters (K).

Let $\Theta = \{\vartheta_0, \dots, \vartheta_{K-1}\}$ be our obtained solution, with $K = 4$ and ϑ_k representing the $(k+1)$-th cluster of genes computed by the algorithm. In correspondence with this partitioning, an ensemble of cluster-wise linear models can be considered as $M = \{\mu_0, \dots, \mu_{K-1}\}$, where μ_k represents the hyperplane being the least square solution over genes in ϑ_k. Our solution Θ is compared with $\Theta_{gw} = \{\vartheta_{gw}\}, \vartheta_{gw} = \{0, 1, \dots, n-1\}$, the degenerate partitioning made of a single cluster indexing the whole dataset D. This solution corresponds to setting $K = 1$, i.e., to the use of a single linear model fitted over the entire dataset D. In the following, we call this model the genome-wide model, labeled as μ_{gw}.

3.1 Enhanced (Cluster-wise) Fitting

Our K-Plane Regression managed to cluster genes with common regulative behaviors, as the obtained model ensemble effectively enhanced data fitting.

Not only the objective value associated with Θ_{gw} is much larger than that associated with our solution Θ (3892.08 vs. 339.83), but also fitting is better at the level of all the computed clusters. The regression scores computed specifically

Table 2: Cluster specific figures of merit. For cluster k, RSS_{cw} and RSS_{gw} are the residual sum of squares of μ_k, μ_{gw} over ϑ_k, while R^2_{cw} and R^2_{gw} refer to the coefficients of determination of μ_k, μ_{gw} over ϑ_k.

cluster (cardinality)	RSS_{cw}	RSS_{gw}	R^2_{cw}	R^2_{gw}
$0(2,717)$	79.22	1393.00	0.80	−2.50
$1(7,547)$	82.05	1514.44	0.54	−7.57
$2(5,045)$	85.64	714.70	0.84	−0.37
$3(4,485)$	92.90	269.92	0.92	0.76

over clusters in Θ, for both cluster-wise and genome-wide models, are reported in Table 2, in terms of residual sum of squares (RSS) and coefficients of determination (R^2). The i-th row of the table contains scores for models μ_i and μ_{gw} over Cluster ϑ_i – subscripts '$_{cw}$' and '$_{gw}$', respectively.

In Table 2, the effectiveness of our approach is confirmed by the fact that clusters are always better fitted by cluster-wise models than by μ_{gw}. Moreover, the specific linear models are very good in explaining the epigenetic transcriptional regulation of a large part of the genes (see to column "R^2_{cw}"). In three clusters out of four, the R^2 scores from μ_{gw} are negative, implying that the fitting over the genes of the clusters is worse than the constant mean model.

The intuition that μ_{gw} is likely to only capture the regulative mechanisms of genes with "intermediate" regulative behaviour, such as those in Cluster 3, is supported by what observed in Figure 2, where cluster-specific residuals ($\boldsymbol{y}$-axis) from cluster-wise and genome-wide models are plotted against target values ($\boldsymbol{x}$-axis).

Residuals from the genome-wide model are generally more disperse and heteroskedastic, except for Cluster 3 – the only one where μ_{gw} attains positive R^2 – where they are similar to those from the cluster-wise model μ_3, fitting genes very well. The overall R^2 of 0.66 attained by μ_{gw} on the whole dataset D is, consequently, an intermediate value resulting from putting together mildly-modeled genes (Cluster 3) with the remaining ones, where the genome-wide model seems to be rather inadequate.

Hard hyperplanes clustering has revealed the criticality of single genome-wide regression by exposing subsets of genes under-fitted by μ_{gw}. Clearly, in a realistic setting, such as targeted cancer therapy, it would be completely unacceptable to fit only 23% of protein coding genes (Cluster 3), as conceptually wrong conclusions might be drawn about the epigenetic regulative behaviors of the remaining genes.

3.2 Cluster Characterization

The effectiveness of hyperplanes clustering also emerges by observing how the obtained clusters are distinct in terms of the input patterns and mean expression value for the genes they contain. In this sub-section we leverage on the

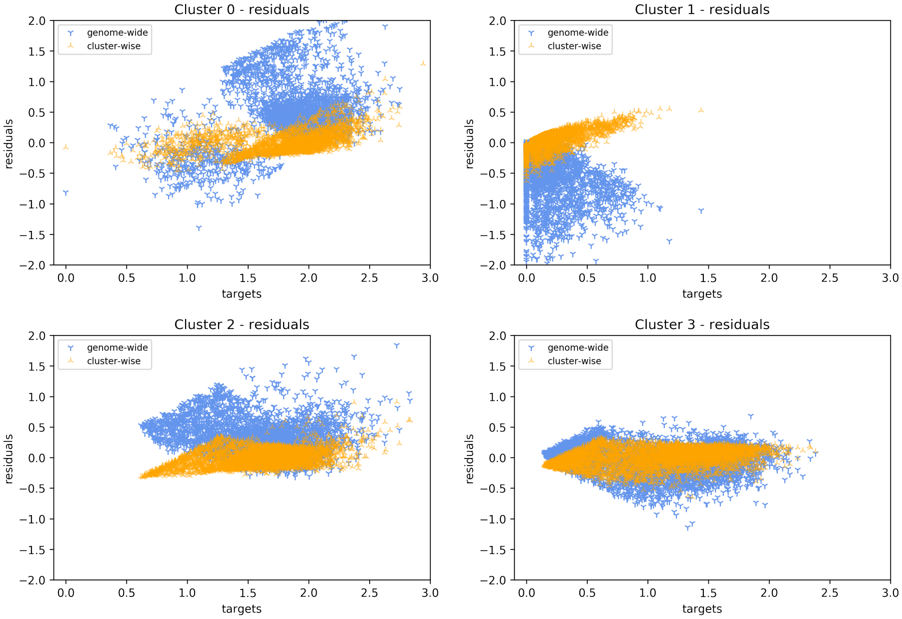

Fig. 2: Cluster specific residuals from cluster-wise (orange) and genome-wide (blue) models.

enhanced interpretability coming from a hard gene partitioning to characterize the computed clusters.

For the generic cluster-wise model μ_i, let $\tilde{\boldsymbol{w}}_i$ be its weight vector, made of the learnt intercept and regression coefficients ($m+1$ elements). For gene g in ϑ_i, let $\boldsymbol{\psi}_g$ be its weighted input vector, obtained by an element-wise multiplication between its input vector $\tilde{\boldsymbol{x}}_g$ and $\tilde{\boldsymbol{w}}_i$ – the input vector is 1-edged to account for bias. The weighted input vectors are an effective means to quickly assess the importance of single features in determining the predicted response value, as $y_g = \sum_{j=0}^{m}(\boldsymbol{\psi}_{g_j})$. The weighted input vectors for all the genes in a cluster generate feature-wise boxplots which illustrate the frequency distributions of cluster-specific histone contributions and their associated dispersion.

Figure 3 depicts the patterns obtained by considering the feature-wise medians of the weighted input vectors, specifically for each cluster, along with the 25^{th} and 75^{th} percentiles of their distributions. In the patterns, intercepts are in orange, whilst HMs are green if associated with positive regression weight (computed *activators*) and red otherwise (computed *repressors*), with semi-transparent rendering for weights not passing a statistical F-test with significance $\alpha = 0.01$. In this way, fictitious correlations are pruned, resulting in simpler and more robust patterns.

Cluster 1 is the most populated one and comprises genes with a negligible histonic activity and usually null expression: these genes are likely to never be

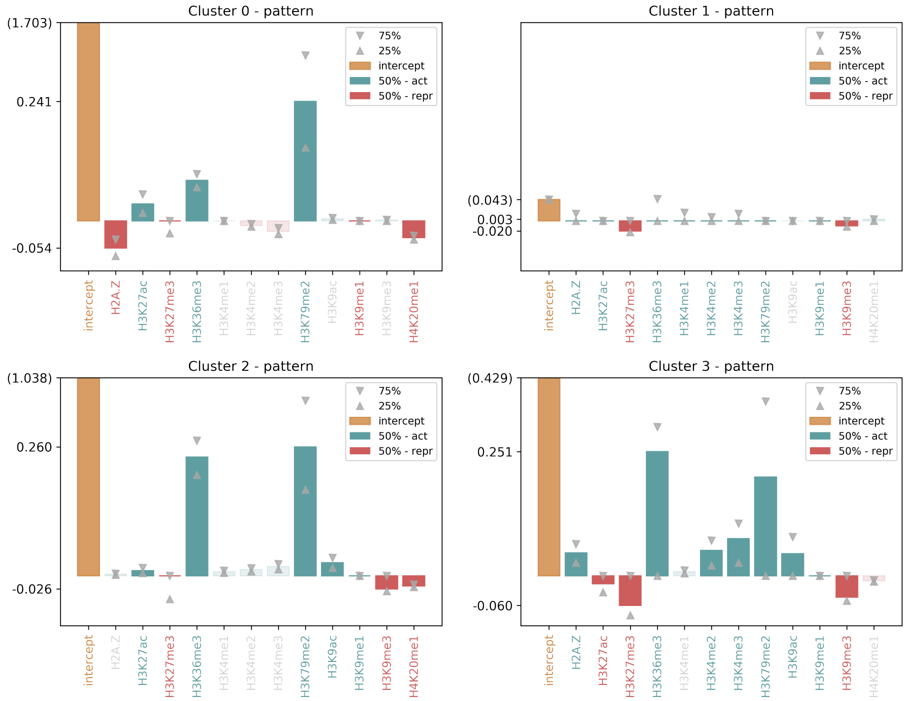

Fig. 3: Cluster specific input patterns; in green *computed activators* (positive weight), in red *computed repressors* (negative weight).

activated, being instead repressed at a chromatin level, like for example developmental genes. The flat input and the low intercept are consistent with the related expression distribution (0.0 RPKM median value). In such a scenario, a lower signal-to-noise-ratio is the probable cause of the mild attained R^2 score in this cluster (see Table 2).

Clusters 2 and 0 are made, respectively, of low and high expressed genes (RPKM medians 12.73 and 32.23). This characterization is confirmed by the intercepts and the predominant roles assumed by activator H3K79me2, and H3K36me3 specifically in Cluster 2, with their large variations explaining higher expression levels. The two clusters show different relative regulative relevance from repressors H2A.Z and H3K9me3, and activators H3K27ac and H3K9ac.

Cluster 3 embraces null to low transcriptional activity (RPKM median 2.32) and is characterized by a more complex input pattern: more relevant than in other clusters are H3K27me3, H3K4me2 and H3K4me3. The simultaneous presence of activating and repressive marks (H2A.Z and H3K27ac, respectively) recapitulate the characteristic of bivalent promoters, whose genes could be poised for fast activation when needed. Moreover, single HMs might counterbalance one another and/or co-work to induce particular effects.

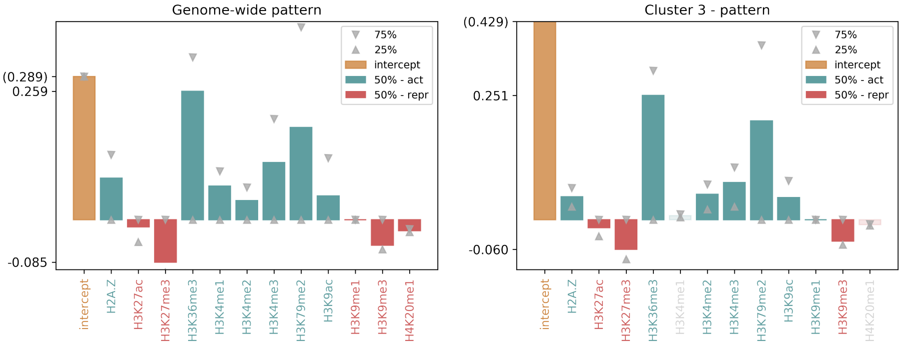

Fig. 4: Genome-wide input pattern (left) vs. Cluster 3 input pattern (right).

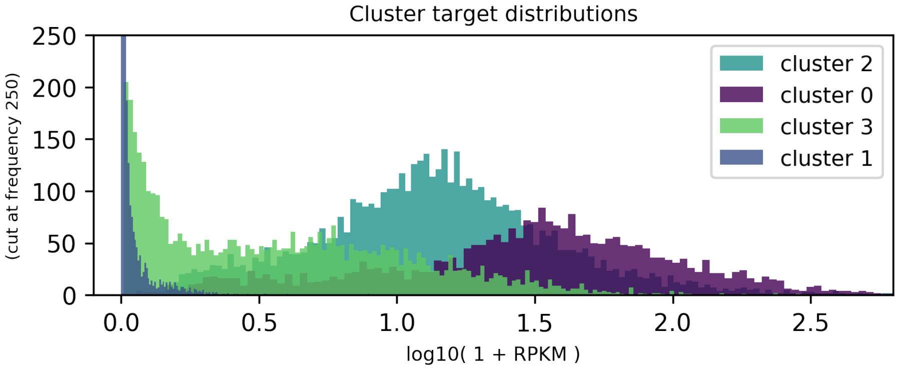

Fig. 5: Cluster target distributions (base-10 log of RPKMs).

It is interesting to notice the similarity between the pattern of Cluster 3 with that of the genome-wide one, as shown side-by-side in Figure 4. This is a further confirmation the algorithm has managed to expose the sub-group of genes possessing the largest leverage in bending one single regression hyperplane. As emerging from the cluster characterization above, the remaining population has instead been set apart in a well differentiated manner: genes lacking the single genome-wide fit have been naturally stratified according to their expression value in groups with distinct characteristic input patterns.

Such discussed stratification for transcriptional activity is finally recapitulated in Figure 5, which reports, for each cluster, the distribution of the expression response value in base-10 logarithm of RPKM.

4 Conclusion

We proposed the application of a randomly re-initialized version of K-Plane Regression to expose subpopulations of protein coding genes commonly regulated at an epigenetic-histonic level. The proposed approach has revealed how a single regression model only captures the fit of a sub-group of genes with null to low expression and how poor scores from μ_{gw} on the remaining genes are due to unfitting rather than linear under-fitting. The hard gene partitioning produced by our method allowed instead a statistical characterization of the computed clusters in terms of input contribution patterns, revealing how clusters stratify for higher and higher expression levels, with histone marks assuming specific roles of different relevance.

Future developments will involve the biological characterization of the found gene clusters: grouped genes will be analyzed to find possibly enriched biological processes they are involved into. Also, further investigations will target the optimal choice of the pre-determined number of clusters, i.e., K: we will analyze cluster-specific target distributions and patterns as a function of parameter K to understand the partitioning dynamic behavior of the algorithm.

Bibliography

[1] Amaldi, E., Mattavelli, M.: The MIN PFS problem and piecewise linear model estimation. Discrete Applied Mathematics **118**, 115–143 (2002)

[2] Bannister, A.J., Kouzarides, T.: Regulation of chromatin by histone modifications. Cell Research **21**(3), 381–395 (2011)

[3] Breiman, L.: Hinging hyperplanes for regression, classification, and function approximation. IEEE Transaction on Information Theory **39**, 999–1013 (1993)

[4] Budden, D., Hurley, D., Cursons, J., Markham, J., Davis, M., Crampin, E.: Predicting expression: the complementary power of histone modification and transcription factor binding data. Epigenetics & Chromatin **7**, 36 (2014)

[5] Cheng, C., Yan, K.K., Yip, K., Rozowsky, J., Alexander, R., Shou, C., Gerstein, M.: A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets. Genome Biology **12**, R15 (2011)

[6] ENCODE Project Consortium: An integrated encyclopedia of DNA elements in the human genome. Nature **489**(7414), 57–74 (2012)

[7] Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M., Amin, V., Whitaker, J., Schultz, M., Ward, L., Sarkar, A., Quon, G., Sandstrom, R., Eaton, M., Wu, Y.C., Lin, Y.: Integrative analysis of 111 reference human epigenomes. Nature **518**, 317–330 (2015)

[8] Levy, S.E., Myers, R.M.: Advancements in next-generation sequencing. Annual Review of Genomics and Human Genetics **17**, 95–115 (2016)

[9] Manwani, N., Sastry, P.: K-plane regression. Information Sciences **292**, 39–56 (2015)

[10] Maston, G., Evans, S., Green, M.R.: Transcriptional regulatory elements in the human genome. Annual Review of Genomics and Human Genetics **7**, 29–59 (2006)

[11] do Rego, T.G., Roider, H.G., de Carvalho, F.A.T., Costa, I.G.: Inferring epigenetic and transcriptional regulation during blood cell development with a mixture of sparse linear models. Bioinformatics **28**(18), 2297–2303 (2012)

[12] Vaquerizas, J., Kummerfeld, S., Teichmann, S., Luscombe, N.: A census of human transcription factors: function, expression and evolution. Nature Reviews. Genetics **10**, 252–263 (2009)