Bioinformatics
Submission Date: 11 October 2016
Master Thesis

Bioinformatics

Enhancer – gene networks for the identification of cancer driver genes affected by enhancer mutations

Author: Sokratis Kariotis

Supervisors : Jeroen de Ridder, Sjoerd Huisman

Bioinformatics Lab, Delft University of Technology

Abstract

Motivation: Reversible epigenetic modifications that happen on the DNA's histones, namely histone modifications, play an important role in gene regulation by controlling the accessibility of different functional genomic regions. Such modifications have been measured and primarily studied on genes or their promoters, but a currently interesting and less studied category of functional elements are enhancers. Enhancer regions can be found virtually anywhere along our non-coding DNA and through looping towards the promoters of target genes they contribute to their normal expression patterns. Abnormal epigenetic signatures and somatic mutations in those regions can interfere with the enhancer's looping procedure and result in irregular gene expression patterns, a crucial contributor in cancer development.

Results: In this paper we propose a method which utilizes epigenetic information across multiple cell types, to form reliable enhancer – gene pairs. The formation of each pair is based on the multiple correlation between epigenetic mark enrichment of an enhancer set and a gene's expression. The pairing procedure deals with the increased computational requirements caused by the multi-dimensional nature of the data and constructs pairs which follow the literature notion of frequent linearly proximal enhancers. The distribution of distances of each pair's elements, showed different results in regulation proximity between our method's pairs and a number of randomized sets of pairs. These pairs are assembled into multiple enhancer-gene (EG) networks which include multiple connected subnetworks of different sizes. On each EG network we overlaid non-coding somatic mutations and found that enhancers of cancer census genes have a higher percentage of mutated enhancers as well as more regulating enhancers, than non-cancer genes. The mutation percentages and the within-pair distances showed different behaviour among chromosomes as well as epigenetic marks. Finally, we inserted these into network databases so that they can accept queries and be easily extendable.

Conclusion: The core of this paper deals with the subject of enhancer cis-regulation and we have performed analyses in order to link regulating enhancers to their target genes, a recent and increasingly important task since the availability of high quality data. The subsequent EG networks depict the circuitry of regulation within a cell's nucleus as well as the mutational landscape and are excellent candidates for further exploration of either the role of specific enhancers in gene misregulation during cancer or more generic characteristics of cancer related enhancers that affect cancer driver genes.

Contact: sokratiskariotis@gmail.com

1 Background

In recent years, studies have focused on the epigenetic modifications present on our DNA or histone proteins, highlighting their important contribution in chromatin structure and functionality [1-3]. A subgroup of those chemical modifications, called histone post-translational modifications, were repeatedly found to enrich several functional regulatory elements [4,5], special segments of our DNA which play an integral part in the process of gene expression. Such modifications, or alternatively referred to as "epigenetic marks", can perform numerous functions on different genomic regions. However, their primary contribution is to control the accessibility of said regions by allowing or blocking the binding of transcription factors (TFs). *Table 1* presents a list of the most studied epigenetic marks and the regulatory elements they define. Since several studies were able to discriminate

different elements quite efficiently using enrichment patterns of these marks [6,32], they are currently being used as one of the most effective methods of identifying important elements on the human genome, like promoters or enhancers [5]. It is worth noting that despite the decent amount of epigenetic information discovered so far, we still do not have a complete picture of each mark's role. This is especially evident (and critical) in disease cases, like cancer, where the lack of complete and quality epigenomes is pronounced. ***(a more detailed description of epigenetics can be found in the supplementary section "Epigenetics, an integral part of chromatin structure")***

Mark	Region	Role
H3K4me3	Promoter regions	Correlated to active transcription
H3K4me1	Enhancer regions	Constitutively marks enhancer elements
H3K36me3	Transcribed regions	Correlated to active transcription
H3K9me3	Heterochromatin regions	Transcription repression
H3K27me3	Polycomb repression regions	Transcription repression
H3K4me1 + H3K27me3	Enhancers	Indicate a repressed state

Table 1: *Studied characteristics of several methylation marks [6,7,8]*

Functional regulatory elements have been studied extensively and catalogued based on the way they affect expressing genes. Among those, the enhancer elements are of special importance. Specific activator proteins bind first to such regions and through DNA looping (*Fig. 1a*), to the transcription factors found on the promoter of a gene. This interaction increases or reduces the transcription rate of the targeted gene. Therefore, enhancers are able to regulate transcription of proximal or distal, upstream or downstream genes [9, 10]. Additionally, enhancers belong to the non-coding part of the genome (which constitutes 98.8% of the total human genome) and can be found both in intergenic and intragenic regions of unrelated genes. Since they can be located virtually anywhere on the genome, enhancer identification is a very tedious procedure but their direct role in regular gene expression makes this task essential. Towards that direction, genome-wide scale, protein binding (ChIP-seq) and DNA hypersensitivity (DNase-seq) data have been examined in respect to enhancer activity. In several studies, transcription factors (like the well-studied p300 protein, a co-activator protein which enhances gene transcription through chromatin remodelling and acetylation of the histone proteins) and epigenetic modifications (initially the presence of H3K4me1 or H3K27ac) were found to be correlated with gene function [11, 12]. Such discoveries lead to the redefinition of enhancers as the regions that have a chromatin profile that is correlated with the transcription of a gene. (***more details about enhancers can be found in the supplementary section "Enhancers, the useful non-coding DNA"***)

After the connection between enhancer epigenetics and gene expression was established, many cancer studies shifted their interest from examining gene and promoter regions to enhancers. Cancer is a disease that involves abnormal/uncontrollable cell growth caused by over-expression or under-expression of certain genes and initially it was thought that only genetic abnormalities (DNA mutations) contribute to cancer development. After researching newly available epigenetic data, it was found that also epigenetic abnormalities (changes in the chromatin profile) are involved in different kinds of cancer [13]. Therefore, combinations of genetic and epigenetic "mistakes" are now thought to be associated with cancer. Epigenetic modifications can influence cancer in two ways [14]. First, specific abnormal epigenetic signatures can cause oncogenes to express. These genes are involved in cell growth and their activation allows the cell to continue replicating rapidly when normally it should not. Secondly, epigenetic changes can cause the inactivation of tumour suppressors, genes that are responsible for stopping or decreasing proliferation of cancerous cells. An example of the latter is visualised in *Fig. 1b.* The task of identifying the epigenomic features related to cancer is immense as all these features have been found to be cell-type and tissue specific [15,16,17]. An informative step towards that direction was the discovery that somatic mutations on gene regions can alter the local epigenetic landscape and create an irregular expression regulatory circuit which allows cancer development [19]. Equally informative are deletion experiments, where the few known (cell-type specific) cancer-related enhancers were removed from a subject and a decrease of tumorigenesis or increase of growth inhibition was observed [20-25]. Finally, one of the most recent findings was that mutations on enhancers

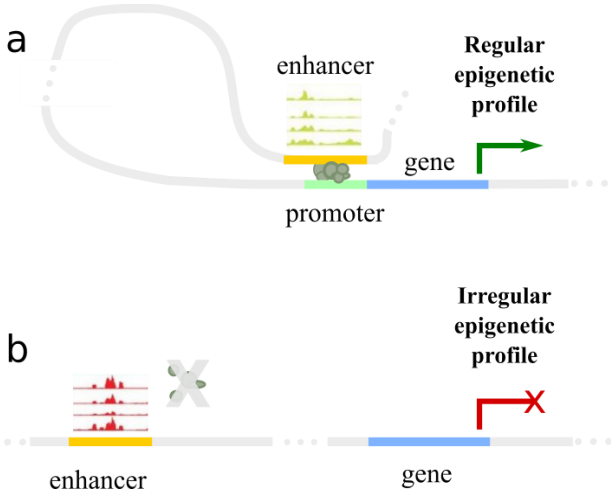

Figure 1 : *(a) The looping procedure through which an enhancer binds to the promoter of a gene and affects its expression, (b) A case where an abnormal epigenetic landscape of an enhancer region, blocks the up-regulation of a tumour suppressor, thus contributing in cancer development. Above every enhancer region, we can see a number of epigenetic signal tracks which are measuring several marks' enrichment. These marks are partly responsible for the functions of the enhancer. Green signals represent normal epigenetic signatures, while red represent abnormal signatures which can affect the expression pattern of a gene by not allowing transcription factors to bind on the enhancer's region, thus disabling the looping procedure.*

can contribute to cancer by causing them to gain new functions or lose those they already have, as extensively discussed in [26-29]. (***more details about the enhancer's epigenetic landscape and its role in cancer can be found in the supplementary section "Enhancer's epigenetic landscape involved in Cancer development"***)

The most prominent topics that bioinformatics in this field have to deal with are enhancer prediction and their matching with target genes. In the first case, three general strategies are being used to predict enhancers. Clustering of similar epigenetic profiles composed by several marks was able to identify numerous different classes of genomic elements [30], while TF binding motifs, coupled with individual epigenetic marks' location have identified genomic element markers [31]. In [32], a computational prediction algorithm compares the similarity of an element's epigenetic profile to a trained set of profiles, achieving a relatively high recognition rate. Moreover, learning based algorithms like SVMs [33,34] combine different classification models using features that derive from histone modification marks or sequence characteristics to discriminate enhancer from non-enhancer elements. But one of the most recent models of enhancer prediction uses a Hidden Markov Model (HMM) with integrated information from multiple histone marks to annotate every region of human DNA [35,36,37]. As mentioned, the second big challenge for bioinformatics is the discovery of promoter – enhancer interactions (PEIs). Several studies used the cell-type specific correlation between one epigenetic mark and gene expression, in fixed regions around genes [38] or DHS correlations [39], while a more recent method uses a Random Forest classifier based on four features like enhancer–promoter activity profile correlation and co-evolution of enhancers and promoters [40]. In a different approach, the framework developed in [18] predicts gene expression utilizing the combinatorial effect of different epigenetic marks through a deep convolutional neural network. Looking the problem from a different angle, a few studies attempted to explain the link between non-coding SNPs and non-cancer diseases using chromatin regulators and transcription factors [41,42] or Hi-C data [43].

Following these recent advances, we propose a method to infer regulating enhancer – gene pairs based on the correlation between the epigenetic signature of enhancer regions and a gene's expression, which is currently considered as one of the most effective methods to find such relationships [50,51]. Subsequently, we construct reliable enhancer – gene (EG) networks for different epigenetic marks, which attempt to draw a picture of the regulatory circuitry that governs gene regulation. EG-networks are then enriched with known cancer information, in the form of somatic mutations and cancer related genes from several cancer types, aiming to discriminate cancer related enhancers. These enriched networks serve as an easily extendable source of information since we can integrate a wide variety of data having as an end goal the discovery of currently unexplored characteristics of gene regulation through enhancer involvement.

Although a number of studies use epigenetic marks to relate enhancers to gene expression [52], their methods do not allow the parallel visualization of information stemming from multiple histone marks, whose role we only partially know. Also, despite the existence of a few enhancer predicting and gene pairing methods, there have been no attempts for creating networks that utilize all these pairs in a unified context. Combining information of multiple pairs at the same time can enable the discovery of broader relations between more than two participating genomic elements.

2 Approach

In this project we are concerned with the construction of enhancer – gene (EG) networks and their potential as a source of information about the regulatory landscape in a cancer context. Our approach can be described in three main steps, each presenting a different set of challenges and corresponding solutions. A complete overview of the approach can be found in *Fig. 2*.

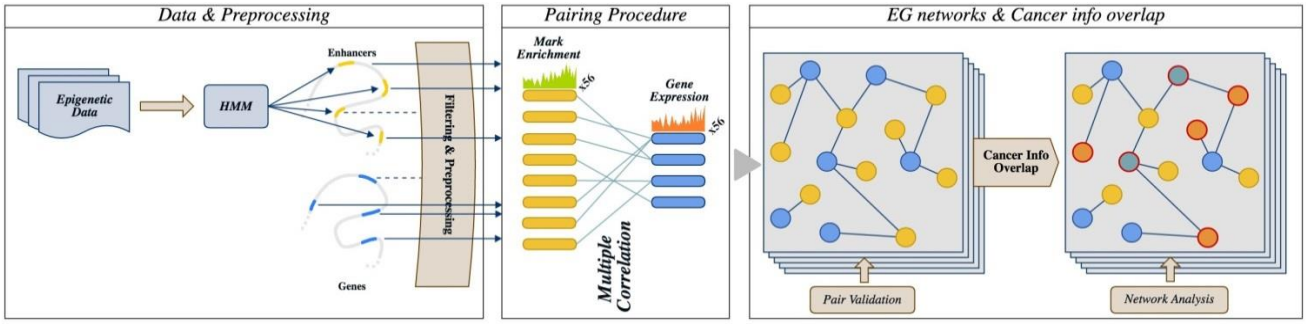

Figure 2: *The three steps of our approach. The first step concerns the acquisition, pre-processing and filtering of our epigenetic mark and enhancer position data to be used as input for the following steps. The second step describes the regression pipeline we use to pair our candidate enhancers with protein-coding genes. The pipeline contains two types of regressions (Lasso and stepwise) and utilizes multiple correlation (many predictors) between each enhancer's mark enrichment and a gene's expression, across 56 cell types. The final step includes assembling the pairs to EG-networks (per chromosome) and overlapping cancer information (somatic mutations and cancer census genes) in order to examine mutated enhancers.*

The first step (*Fig.2, left box*) begins with the acquisition of the genomic regions that will play the role of enhancers. Since we approach this problem from an epigenetic point of view, we require a prediction method that utilizes relevant information. As stated in the previous chapter, the HMM built in [35] and used by the Roadmap Epigenomics Consortium in [15] offers the most recent and widely applied epigenetic method to predict enhancer regions in 56 different cell types. Since different cell types have different patterns of epigenetic marks across their DNA sequence, any HMM model will produce different sets of candidate enhancers, which reflect the active enhancers in each cell type. We are interested in constitutive, global regulatory elements. Therefore we filter the cell-type specific enhancers by introducing a threshold (CT) of enhancer activity in a minimum number of cell types. To validate the candidate enhancers we cross-reference them with the small set of experimentally verified enhancers that is publicly available [44]. In addition to the enhancers' positioning we acquire the expression of nearly 20,000 genes residing on the human genome. To avoid cluttering our networks with uninteresting pairs later on, we filter the low variance genes. We perform preprocessing of the data concerning both sets of elements. The enrichment of the five epigenetic marks on enhancers requires normalization and averaging, while the gene expression also needs to be normalized since it heavily suffers from skewness, extreme values and large variance differences.

The core of this approach is the formation of enhancer – gene pairs, that is the identification of the regulating enhancers of every gene (*Fig.2, middle box*). We are taking advantage of the observation noted earlier on, that the enrichment of an epigenetic mark on an enhancer should correlate with the expression of the regulated genes. While this is the main premise, we also must take into account the co-operative nature of enhancer regulation. For that reason we created a model which computes the multiple correlation (across 56 cell types) between a gene's expression and an epigenetic mark's enrichment on a set of candidate enhancers. The enormous number of candidate enhancers for each gene leads to dimensionality issues, where the small number of data-points (56) are unable to explain the combined effect of around 150,000 enhancers. To counter that problem we use two methods, successively. First, we restrict the space around each gene within which we allow candidate enhancers, a decision that excludes rare distal regulation but facilitates the extremely more frequent proximal regulation. Secondly, we use a multivariate Lasso regression model which uses as input the remaining candidates and picks a small number of enhancers whose mark enrichment is most correlated

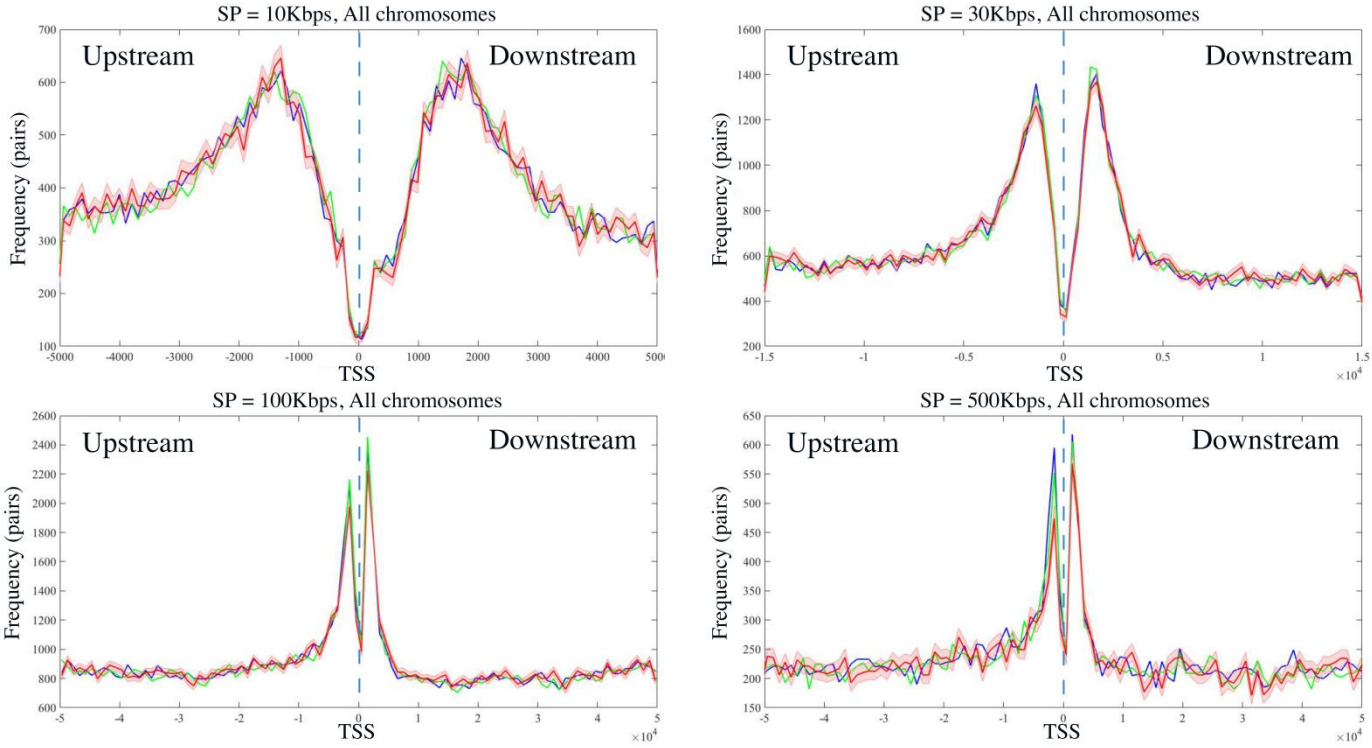

***Figure 3**: The four distributions of pair distances using different SPs. The blue line represents the pairs that resulted from our regression pipeline. The red line represents the pairs that resulted from the randomization experiment where we randomly picked enhancers that are strictly located within the SPs and are equal in numbers with those produced by the pipeline. The green line represents the pairs the resulted from the randomization experiment where we permuted the gene expression vector of every gene and then re-run our pipeline. All the subplots in this figure were produced by the correlation between gene expression and H3K4me1 enrichment.*

with the expression of a gene. The resulting enhancer set is the input of the last part of this approach step, where we apply an additional stepwise regression model. This type of regression is very powerful but quite computationally costly, however the small number of predictors allows for viable runtimes.

For each of our epigenetic marks the above procedure provides a set of enhancer – gene pairs. In order to assemble them to EG networks (*Fig.2, right box*) we first have to evaluate their reliability. For that purpose and due to the lack of reliable pre-existing pairs from other approaches, we focus on the linear distance between the pair elements. Research about the linear distance between enhancers and regulated genes has repeatedly shown that this type of cis-regulation majorly consists of short distance (< 50Kbps) pairs while more distal pairs are quite infrequent [48]. Therefore a set of such pairs must follow this distance distribution. Since we do not have enough data to afford an independent test-set, we measure the future performance of our resulting regression models by cross-validation. We validate our fitting procedure by producing mainly positive leave-one-out R^2 values. Next, we overlap our networks with a layer of cancer information in the form of confirmed somatic cancer mutations and consensus cancer genes. This overlap enables us to label cancer related enhancers based on the literature statement that somatic mutations on enhancer regions affect the target's expression [26-29]. This paper is concluded after the exploration of a number of characteristics about these augmented EG networks.

3 Results & Discussion

3.1 Constitutive candidate enhancers

The HMM provided 56 large sets of candidate enhancers (CEs) spread across the genome, one for each cell type we considered in our research. The first filtering step, aims to determine the number of cell types our enhancers must be active in. Consequently, we set up a threshold, termed CT, which indicates the constitutive function of an enhancer. We compare multiple candidate enhancer sets based on different CTs and validate the final choice of our threshold based on the retrieval rate of the only existing set of experimentally verified enhancers found on the human genome and assessed in transgenic mice [44]. Several performance metrics as well as the strict cell-type specificity of enhancers, indicated that low thresholds and more specifically a CT = 10% performed best and was therefore chosen. ***(additional information on the CT experiments can be found in the supplementary section "Candidate enhancer constitutive threshold")***

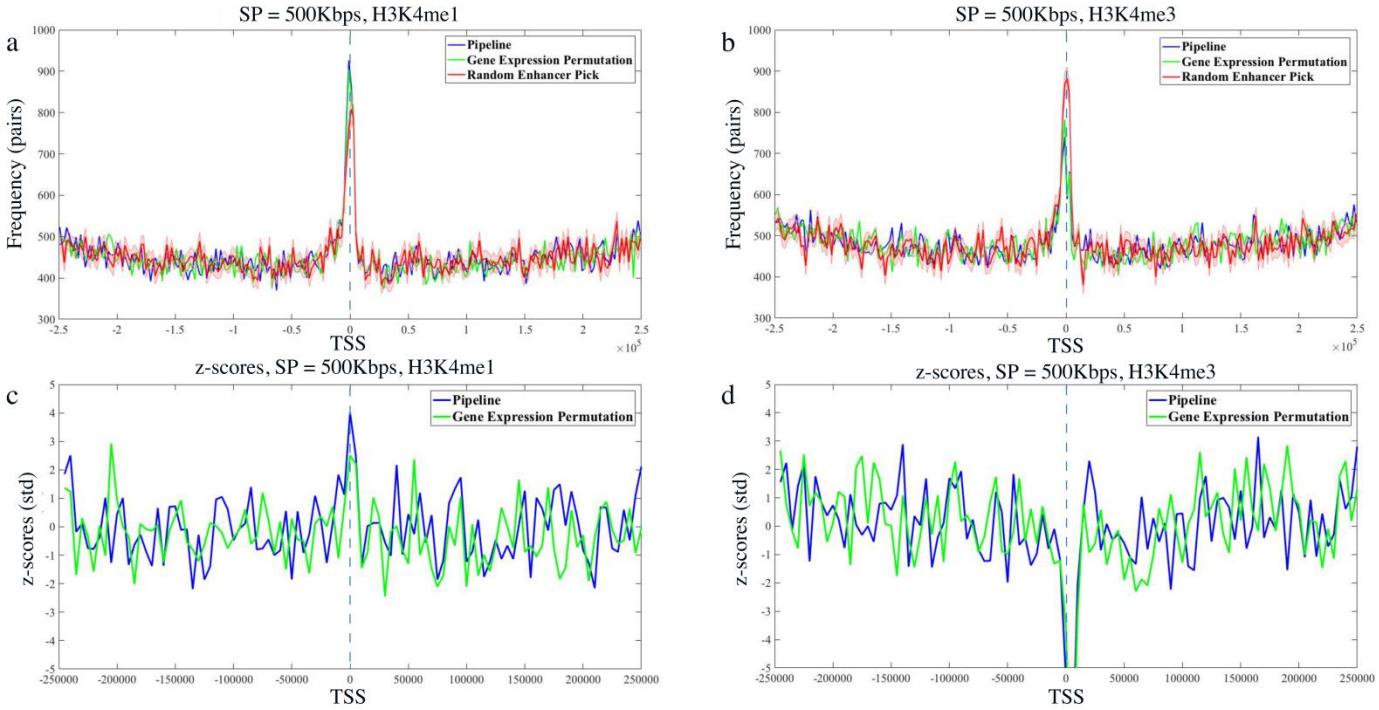

Figure 4*: Subplots (a, b) depict the distribution of distances for SP = 500Kbps and two differently performing epigenetic marks, the best performing mark H3K4me1 (a) and the worst performing mark H3K4me3 (b). Subplots (c, d) depict the distributions of z-scores for every distance around the TSS. The z-scores measure how many standard deviations away from the red randomization's (random enhancer pick from within the search space) mean is each line. The blue line represents the pairs from our pipeline, while the green line represents the pairs from the randomization where the pipeline was re-run with permuted gene expression vectors. The red line (random enhancer pick) is not shown as in every case it is the horizontal line y = 0.*

3.2 Enhancer – gene pairs

3.2.1 Enhancer – gene proximity

Due to the immense number of candidate enhancers in each chromosome we have to apply constraints on the regions we will consider in our pairing pipeline. In order to set an effective search space (SP) we executed our regression pipeline using different sets of enhancers that come from different SPs. We also created two sets of randomized pairs, with different levels of randomness, to compare with our original pairs. The hypothesis that governs this experiment is that proximal pairs are much more frequent than distal pairs [48] and this should also be reflected in our selected set of pairs. *Fig.3* showcases three lines that represent: our pipeline's pairs (red line), pairs that were also created by the pipeline but after we permuted each gene's expression vector (green line) and pairs for which we randomly chose the enhancer but from within the chosen search space (red line). Although the high frequency of proximal pairs was shown to hold for all three cases to some extent, we can see the gradual distinction of the blue line (generated from the pipeline pairs) as the SP increases in *Fig. 3*. According to the distances found in the literature [48], we consider SPs up to ten times larger than the range within which we usually observe enhancers, therefore we do not enforce the picking of proximal enhancers. Despite the freedom we allow, the regions on either side of –and proximal to- the TSS show a prevalence of the blue line (pipeline pairs) for larger SPs. There is a clear signal of proximal pair preference manifested specifically with an SP of 500Kbps. To follow the changes of each line and their relation, *Fig. 3* demonstrates this signal using the blue (pipeline pairs), red and green (randomized) lines for four increasing SPs (10, 30,100 and 500 Kbps). Also visible, is the sudden drop immediately around the transcription start site (TSS). This drop happens because this area is populated by either promoters or the first exons of a gene.

The z-scores of the blue and green distributions were calculated against the random enhancer randomization (red). Our pipeline's z-score at the proximal space was found to be ~5 standard deviations away from the red random case and ~2 standard deviations from the less-random green distribution. During the data pre-processing step we observed that many genes only have small differences in gene expression across cell types (even after the variance filtering). As a result, the gene expression vector permutation will not change the vector enough to correlate with different (than the pipeline's) enhancers and we will end up with a number of same (possibly proximal) pairs. The z-score distributions confirmed that the SP of 500Kbps holds the strongest signal as it most deviates from y = 0, near the TSS ***(more about the distance distributions and the z-scores in the supplementary section "Search space determination").***

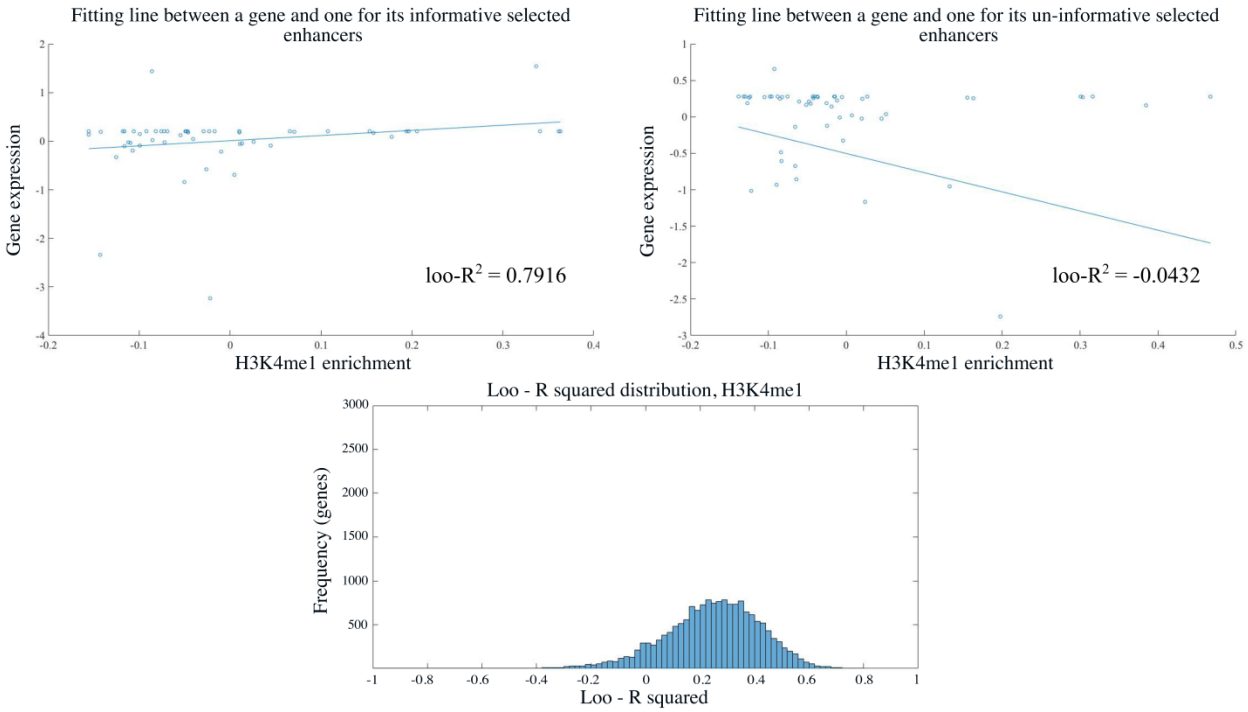

Figure 5*: Each of the top plots describes the fitting line between the elements of a pair. These lines were calculated using all active enhancers of the specific gene, but in the plots we present the same fitting line only in respect to one of the regulating enhancers. In the left plot we can see an enhancer that positively contributes in the fitting of this gene's expression, while the opposite holds for the case on the right. Notice how much the right fitting line is affected by outliers, in contrast to the left one. At the bottom we can see the mainly positive distribution of loo-R^2 values for all genes while using the H3K4me1 mark.*

We also examined the consistency of the proximal pairs across epigenetic marks and whether our findings coincide with the current knowledge on our epigenetic marks. According to the experiments, the five marks demonstrated a very different behaviour. H3K4me1 revealed the clearest signal as we expected, since according to *Table 1* (see *Background* section) it is found mainly on enhancer regions and is implicated in its functions. The remaining marks provide less signal, with H3K4me3 forming considerably less proximal pairs than the complete random case (*Fig 4b,d*). The H3K4me3 mark is the only one that shows this kind of negative performance in comparison to the random case. This indicates that while using this mark we preferentially create non-proximal pairs. One possible explanation is that the presence of H3K4me3 on enhancers is correlated with the deposition of Poll II in the same regions, which marks active distal enhancers [53]. Therefore, the frequency by which we select proximal enhancers using H3K4me3 correlation, greatly drops. The rest of the marks appear not to be able to correlate with gene expression to provide realistic pairs and thus their role as regulation intermediates on enhancers, is doubtful. We also noted a difference on the number of enhancers picked by our pipeline when using different marks, with H3K4me1 supplying the most correlated regulators than any other signal.

3.2.2 Model fit

Since the knowledge of enhancer elements is relatively incomplete, the means to validate a regulatory pair are limited. But these results provide an indication of the biological validity of our pairs based on the current research concerning the within-pair distance and the relevance of the particular epigenetic mark to the function of an enhancer. To validate the predictive performance of our models we used a cross-validation test in the form of a leave-one-out R^2. Since our regression is trying to explain gene expression based on up to 10 enhancers/predictors,

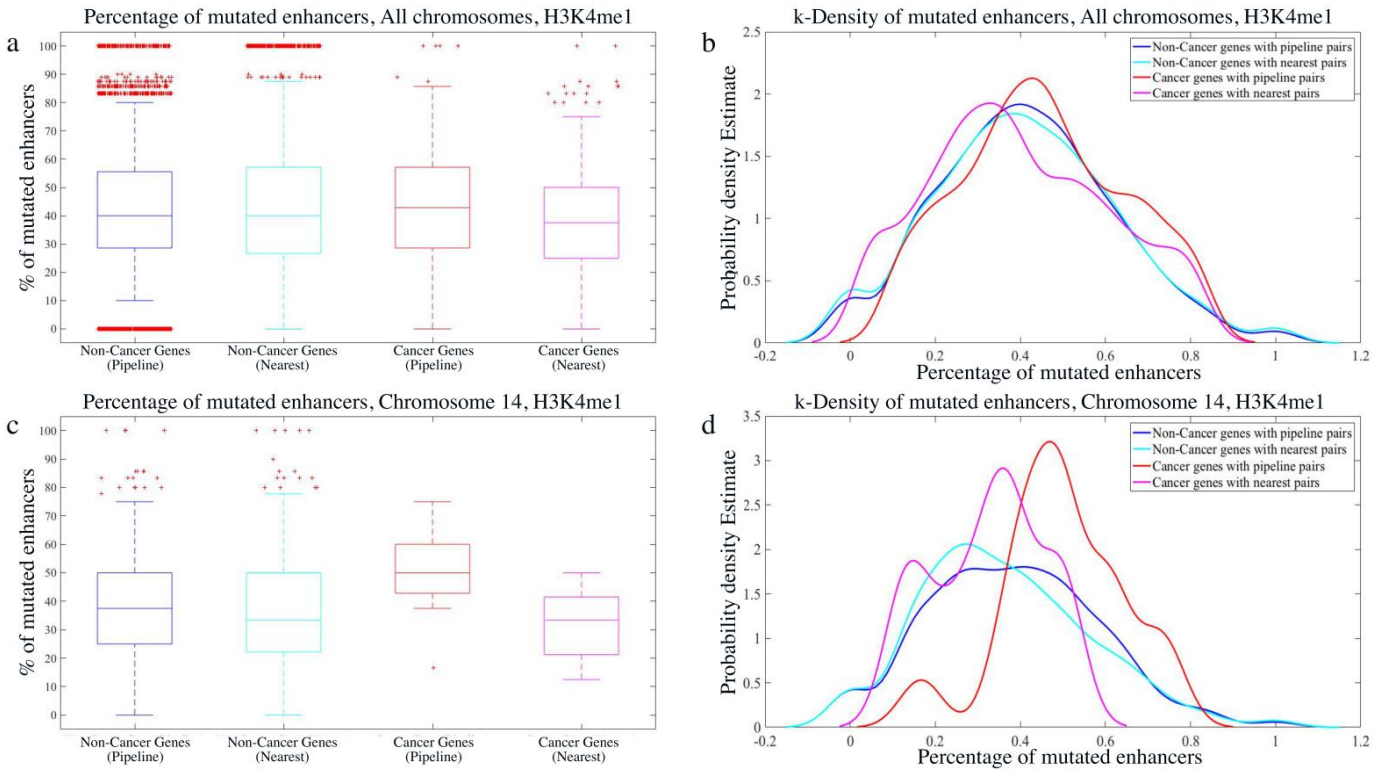

Figure 6*: The boxplots on the left side (a,c) show the percentage of the regulating enhancers that got mutated at least once during the somatic mutation overlap. For each sub plot we have four types of experiments with different combinations of genes and enhancers. The dark blue experiment concerns the non-cancer genes (do not belong to the group of cancer census genes) whose enhancers were determined by our pipeline. The cyan type concerns the same type of genes but we paired them with their nearest enhancers (equal number of enhancers as the previous experiment). The red experiment includes only the cancer census genes and the enhancers outputted by our pipeline. The magenta experiment includes the same type of genes but with the nearest enhancers assigned to them. Boxplot (a) describes the percentage over all chromosomes while (c) only for chromosome 14. Density plots (b,d) use the same colours and show the different distributions (converted in probability density estimates) for the four experiments. Plot (b) concerns all the chromosomes, while plot (d) only chromosome 14. For the upper figures 1507 non-cancer genes and 473 cancer genes we used along with the best performing epigenetic mark H3K4me1. The bottom figure refers to chromosome 14 that includes 15 cancer and 543 non-cancer genes. According to the COSMIC database [46] a gene has on average 18 cancer genes.*

the loo-R^2 statistic is also measured with respect to multiple enhancers. In the top-left section of *Fig.5* we can observe a case were a model fits the gene's expression and one enhancer's H3K4me1 enrichment in a satisfying degree. In contrast, on the right side of the same figure we can see a case where the model fails to describe the scattered data-points. The distribution of all our loo-R^2 scores (for the best performing mark H3K4me1) is shown at the bottom of *Fig.5* and includes a few negative scores (models where the enhancers could not explain a specific gene). Although the distribution is not describing very high R^2 values, we should consider that we could not have a much higher result, since the expression of a gene depends on many more variables than just the correlation with an enhancer's epigenetic signature. These scores are additionally important because they show that despite the large-scale predictor removal we still can explain a considerable part of the gene expression. ***(more loo-R^2 scores can be found in the supplementary section "Leave one our R squared").***

3.3 Mutation overlap results

By enriching the EG-networks with cancer-information, we can compare cancer to non-cancer genes. We also introduce a new set of enhancers for each gene type. These new sets include the most proximal enhancers, to each gene, therefore we now have four categories of genes: non-cancer and cancer genes paired with our pipeline's enhancers, and non-cancer and cancer genes paired with an equal number of the nearest, to each gene, enhancers. These are ideal for comparison purposes, since in many studies the most proximal enhancers are assumed to be the regulators as it is easier for them to loop to nearby genes [11]. As seen in *Fig. 6a*, our pipeline's enhancers that belong to cancer genes are more frequently mutated than in any other combination of genes and enhancers. More specifically, for each cancer gene a slightly higher percentage of our enhancers are mutated, compared to the non-cancer genes using the same type of enhancers. Concerning all pairs from all chromosomes, the t-test between the aforementioned distributions has a p-value of 0.1910. A more clear distinction between the two distributions (red and blue, respectively) is visible if we observe the corresponding density plot (*Fig. 6b*), where cancer genes show higher probability density estimates for higher mutation percentages. When the required mutations (to label an enhancer as "mutated") are increased to two, we once again see the same picture but the p-value increased to 0.4372. However, these plots concern every gene in every chromosome and the p-values are not supporting the significance of this difference overall. Specific chromosomes provide lower p-values. As an example *Fig. 6c,d* presents the case of chromosome 14 where the difference is easier observable and the p-value is 0.0758. Similar differences

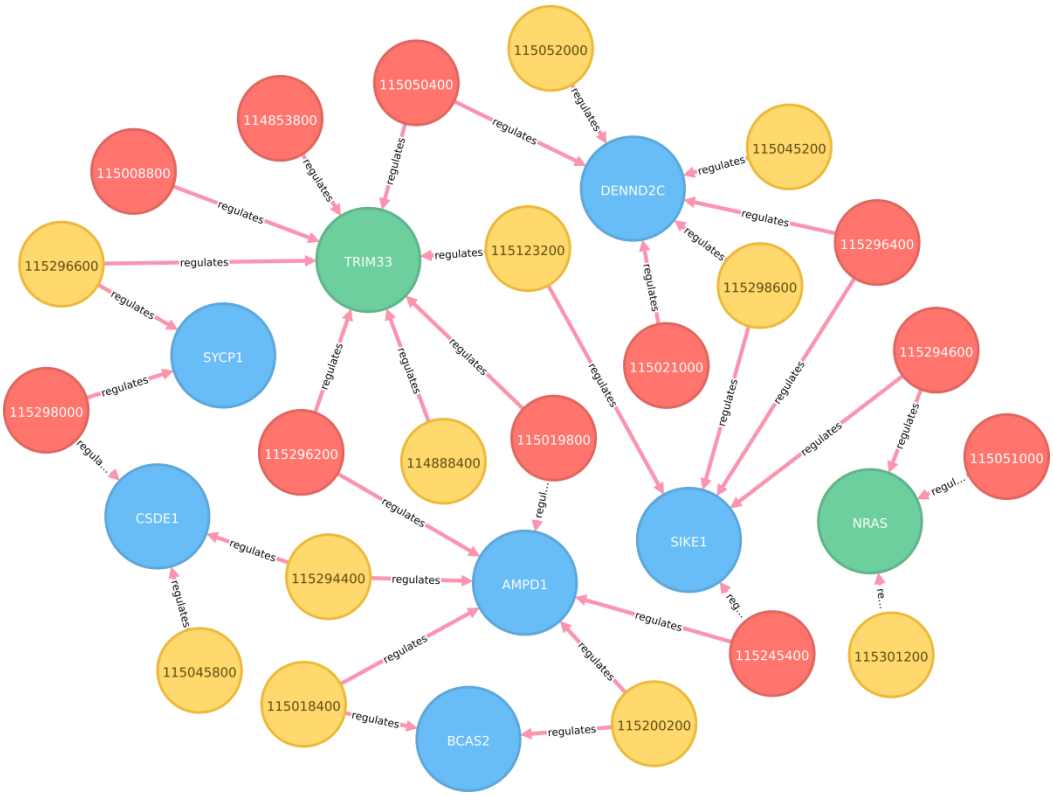

Figure 7: *An example subnetwork that includes two genes related to thyroid cancer (green nodes). Blue nodes represent non-cancer genes, yellow nodes represent non-mutated enhancers and red nodes represent the enhancers that got mutated at least once during the mutation overlap.*

(in p-values) were observed between our pipeline's and the nearest enhancers. Both the cancer and non-cancer genes' enhancers got mutated in a higher percentage than their nearest counterparts.

Furthermore, we examined the number of candidate enhancers (as provided by the HMM) that reside around cancer and non-cancer genes, prior to any kind of filtering. We tested a small region (20Kbps) and a larger one (500Kbps) to get a broader view of the distribution of enhancers around each gene. Cancer genes have significantly more neighbours than non-cancer genes (p-values of 0.0033 and 2.7151e-07, respectively). Since we are using multiple correlation to pick enhancers for each gene, more available enhancers means higher chances for a larger number of correlated enhancers to be picked. Our pipeline is not affected by this considerable difference, as it selects an almost identical number (~7) of enhancers for cancer and non-cancer genes. ***(additional information on plots and p-values in the supplementary section "Mutation overlap").***

3.4 Enhancer – gene network visualisation & tool

For each epigenetic mark and chromosome, we assembled the pairs (formed by our regression pipeline and enriched by the mutational overlap) into one interconnected EG network for which we implemented a tool (using graph databases) for visualisation and query purposes. These graphs include multiple types of nodes: "CancerEnhancers" (mutated at least once during the overlap), "Enhancers", "CancerGenes" (cancer census), and "Genes", while each edge represent the relation "regulates". *Fig. 7* demonstrates both the utility and the informational content of such a graph. In this example we wish to examine the regulation landscape around two cancer census genes involved in thyroid cancer, *TRIM33* and *NRAS*. Thus we queried our H3K4me1 graph for a subnetwork of a small size (30) that includes the two genes. From such a subnetwork we can extract information about mutation and/or enhancer clusters or other topological relations between enhancers and genes. The tool allows for the enquiry of more complex searches, like shortest paths between genes and/or enhancers.

4 Conclusions

4.1 Summary of results

In this paper we used a new approach to create pairs of genes and their regulating enhancer elements, based on the multiple correlation between epigenetic mark enrichment and gene expression. This approach resulted in pairs that significantly follow the linear within-pair distance described by literature and have a positive prediction performance despite the increased complexity caused by the plurality of possible parameters. Fine-tuning these parameters poses as a considerable task but it can lead to significant increase in the network's accuracy. The aforementioned pairs were assembled into EG-networks on which we overlaid an additional layer of cancer related information. In their current state, our networks can be queried in order to explore the regulatory landscape of each chromosome and provide its visualization. The enriched networks show promising capabilities on detecting enhancers of known cancer genes as well as exploring the differences between epigenetic marks, despite the limitations imposed on our models by the (important) requirement of constitutive enhancers. Finally, a very important characteristic of our networks is that they can be effortlessly combined, extended and integrate information from different sources (like eQTLs or 3D conformation). However, at the moment they are limited to single chromosomes and thus they cannot capture inter-chromosomal interactions or deliver the regulatory circuitry of genes that reside in different chromosomes (also a result of the search space definition).

4.2 Future work

A direct improvement in our pipeline's performance would be the acquisition of additional epigenomic information. The NIH (National Institutes of Health) is currently mapping additional epigenomes [49], which translate to more data-points, giving further explanatory power to our enhancer vectors and room for more predictors to our models. Also, a meaningful grouping of our cell-types before executing the pairing pipeline may unveil more specific, but still constitutive, enhancers with higher precision and consequently increase the pairing accuracy. Furthermore, higher resolution candidate enhancers or a more intricate mark representation of an enhancer's region (instead of averaging) can further improve the information quality an EG network can provide. However, this would increase the complexity of the procedure. Since the role of enhancers in cancer development was established, pairing enhancers to genes and exploring how their irregular behavior can contribute to tumorigenesis became a central topic in cancer research. As a step towards this direction, we are proposing this new type of network than aims to shed light to the relatively understudied landscape of enhancer cis-regulation and work towards the identification of enhancer markers to be used in cancer related prognosis and diagnosis.

5 Methods

5.1 Data

The data for this paper were recently (Jan. 2015) made available by the Roadmap Epigenomics Consortium [15]. The histone modification patterns, DNA accessibility and methylation along with RNA expression of 127 curated

epigenomes were integrated and analysed. The resulting high resolution maps, produced by a variety of assays followed by massively parallel sequencing, provide a global view of the epigenomic landscape in a wide variety of human cell types and tissues. However, we only used the 56 epigenomes that had complete, genome-wide data in the form of enrichment signal tracks for the five epigenetic marks H3K4me1, H3K4me3, H3K9me3, H3K27me3 and H3K36me3 as well as mRNA-seq gene expression data. We acquired our initial set of candidate enhancers from [35], an HMM also used by the Roadmap Epigenomics Consortium and made available through the Human Epigenome Atlas [45]. In this approach each epigenome is segmented and each segment is assigned a chromatin state which attaches regulatory characteristics to each successive 200bp region individually. These assignments are decided by a multivariate HMM based on DNA methylation and accessibility, regulator binding and evolutionary conservation. Additional data used include cancer somatic mutations and known cancer genes, downloaded from the COSMIC database [46]. The experimentally verified enhancers were download from the VISTA enhancer browser [44] and were filtered to only include enhancers found on the human genome. To determine the directionality of gene transcription, needed at the analysis of the enhancer – gene pairs and the location of promoters, we utilized the start and end codon gene information included in the evidence-based annotation of the human genome (GRCh37), version 10 (Ensembl 65) from GENCODE [47]. ***(additional description of the data and the HMM can be found in the supplementary sections "Data" and "Hidden Markov Model (HMM)", respectively)***

5.2 Pre-processing

The expression of protein coding genes is required in this project to associate the activity of an enhancer to a gene. We are using the expression of all protein-coding genes in the human genome, as provided by the Human Epigenome Atlas [45]. Due to the inherent problems of RPKM values we apply a standardized log2 normalization to achieve a more normal distribution of gene expression. These standardized RPKMs compose the two-dimensional Gene Expression Matrix (*GEM*) per chromosome, where the first dimension contains all the genes of a chromosome and the second the 56 cell types. As described in the *Data* section, the epigenetic mark enrichment of enhancer regions is measured in fold-enrichment and each base pair has one such value per epigenetic mark and cell type. Since the enhancer regions are all 200bps long, we average the 200 enrichment values (concerning only one epigenetic mark) of each enhancer region and produced one value that represents the mark enrichment of an enhancer in one cell type. For each chromosome and mark, we concatenate these 56 vectors and generated the Mark Enrichment Matrix (*MEM*) where the first dimension describes the enhancers and the second the 56 cell types.

5.3 Candidate enhancer constitutive threshold

In this section we examine the performance of different CTs concerning the retrieval rates of experimentally verified enhancers. We then conclude about the optimal CT value based on the metrics related to the ROC and PR curves as well as additional performance metrics. To define the TPs, TNs, FPs and FNs we first segment each chromosome in 200bps regions (same length as the HMM's enhancers). The candidate enhancers (*CEs*) will be the positive predictions and every other 200bp region

Definition	Description
True Positive (TP)	Number of CEs that overlap VRs
True Negative (TN)	Number of non-CEs that do not overlap VRs
False Positive (FP)	Number of CEs that do not overlap VRs
False Negative (FN)	Number of non-CEs that overlap VRs

Table 2: *The description of the each definition used in the contingency table*

(*non-CEs*) will be a negative prediction. As for the ground truth, 200bp regions that are overlapped (even partially) by a VISTA enhancer will be true enhancers (*VRs*) while the rest will be truly non-enhancer regions (*non-VRs*). Of course, the latter set of regions in reality contains true enhancers but for our comparative purposes we can ignore that. Also, this method of labelling regions produces a very large number of TNs and FNs, but by using PR curves in addition to ROC curves we adjust for that problem. For clarity purposes the above definitions can be found in *Table 2*. A number of performance metrics were used to capture different aspects of this experiment. ***(more on the different metrics can be found in the supplementary section "Candidate enhancer constitutive threshold")***

5.4 Determining search space

5.4.1 Search space definition

To find the most suitable search space around each gene we first have to define which regions a search space includes. For this purpose, three different methods were considered: *Intronic*, *Non-Intronic* and *TSS*. The latter method (*Fig. 8*) has more relevant to our task characteristics and was therefore chosen ***(additional information and a comparison between all three methods can be***

found in the supplementary section "Search space (SP) definition").

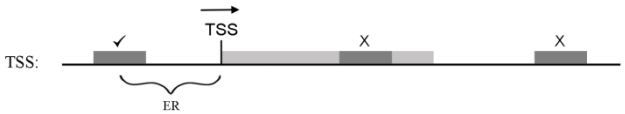

Figure 8: *The chosen definition of a search space. Only checked enhancers are included in the search space. Important note for this example: if the SP region was large enough it would include the rightmost enhancer that is not within the gene*

5.4.2 Simple distribution of distances for different SPs

Next, we apply a simple multivariate Lasso regression model, which predicts the regulating enhancer for each gene, based on the correlation between the enrichment of an epigenetic mark on the enhancer region and the expression of a gene. To reduce the complexity of the experiment we allow the regression model to pick only the single most correlated enhancer (using the LARS algorithm). For this experiment we exclude the genes that had less than two enhancers for a particular SP. Additionally, we generate 100 equally sized randomized sets of enhancer – gene pairs, where the enhancers satisfy the SP restriction. For the real and random sets we measure the distances (in base pairs) between the enhancer and the gene. Then we compute the frequency in which every distance occurs for every type of pair. The frequencies from the 100 randomizations are averaged to form one distribution of frequencies.

5.5 Regression pipeline

5.5.1 Multivariate Lasso regression

The first step of our regression pipeline consists of a multivariate regression model, which aims to considerably reduce the number of available enhancers/predictors for a given gene. To achieve the elimination of predictors, we use the Lasso analysis method as presented in [54] using as input the MEMs and GEMs described in the Data section. This regression disqualifies the predictors that are least correlated with the dependable variable based on the regularization parameter λ. Higher values of λ increase the penalization of the Lasso and more predictors are excluded. Instead of tuning this parameter, we use the LARS (Least Angle Regression) algorithm (MatLab package [55]), which allows us to directly choose the number of predictors we desire. In addition to the pre-processing done so far, LARS required the centering of the gene expression vectors and the normalization of the mark enrichment vectors. The number of predictors chosen for the LARS method is set to 10 for two reasons. Firstly, we need an adequate number of candidate enhancers per gene, since literature dictates that a gene is rarely regulated by a single enhancer but rather from a joint effect of multiple enhancers. Secondly, our dataset is composed by 56 data-points (cell types) for every enhancer and every gene and in order to extract their information (or explain them) we can use up to 10 variables (about 1/5 of our data-point count), otherwise we would contribute to the curse of dimensionality. The genes that do not surpass a variance threshold of 10% across the 56 cell types are excluded from the fitting procedure as uninteresting (housekeeping genes).The end goal of this step is to provide a correlation-filtered small subset of candidate enhancers to the stepwise regression that follows. ***(additional information on the Lasso regression can be found in the supplementary section "Multivariate Lasso Regression Model")***

5.5.2 Stepwise regression

The second step of the pipeline consists of a stepwise linear regression using the small set of candidate enhancers that resulted from the Lasso fitting. Each step of this regression adds the best available term(predictor) when the p-value for an F-test of the change in the sum of squared error is lower than 0.05. If no terms can be added, the regression examines the terms currently in the model, and removes the worst one if an F-test for removing it has a p-value > 0.10. This procedure can connect from 1 to 10 enhancers to any given gene and these pairs compose our EG-networks.

5.6 Cross-validation R squared

The Lasso fitting provides the stepwise regression with a set of candidate enhancers that were picked based on their correlation with the gene expression. Therefore, a standard R^2 measure (even the adjusted version) would be very optimistic about the fitting performance. To avoid this faulty coefficient of determination, we are using cross-validation which indicates the performance of our regression models on unseen data. More specifically, to produce the figures found in the supplementary material, we used the leave-one-out R^2 (loo-R^2), where we leave one of the 56 data-point out of every gene expression and mark enrichment vector and calculate the estimated gene expression based on 55 data-points each time. ***(additional description of the loo-R^2 can be found in the supplementary section "Leave one out R squared")***

5.7 Distribution of distances

In every distribution of distances figure, the x-axis presents the distances (in base pairs) between the TSS of a gene and the closest (to the gene) end of the paired enhancer. The whole axis is segmented in equal sized dis-

tance bins and each bin holds the frequency each type of pairs (blue, red or green) is observed to be apart. The blue line represents the distances that concern the pairs which were the product of our regression pipeline. The red line concerns our first type of randomization where, for each gene, we randomly pick candidate enhancers that are located within the specific search space (SP) we have set. The number of enhancers chosen this way is equal to the number of enhancers our pipeline selected for the same gene. This randomization is performed 1000 times and the frequencies for each bin are averaged. The green line depicts a second type of randomization where we permute the 56-long gene expression vector of every gene and then we re-run our pipeline with the new vectors. This randomization was performed twice and the final green line results from the averaging of the frequencies per bin.

5.8 Mutation overlap

The mutations used in our overlap experiment are drawn from the COSMIC catalogue [46] and are categorized under "Non coding variants". From the 10,937,718 non-coding mutation we filtered out those that were not confirmed somatic variants. Therefore, all the mutations we overlapped were confirmed to be somatic in an experiment by sequencing both the tumour and the matched normal from the same patient. After the overlap of somatic mutations the genes were labelled as cancer or non-cancer based on the current collection of cancer census genes.

6 References

1. Putiri EL, Robertson KD: **Epigenetic mechanisms and genome stability.** *Clin. Epigenetics* 2011, 2:299–314.
2. Dixon JR, Jung I, Selvaraj S, Shen Y, Antosiewicz-Bourget JE, Lee AY, Ye Z, Kim A, Rajagopal N, Xie W, Diao Y, Liang J, Zhao H, Lobanenkov VV, Ecker JR, Thomson JA, Ren B: **Chromatin architecture reorganization during stem cell differentiation.** *Nature* 2015, 518:331–336.
3. Tremblay J, Hamet P: **Impact of genetic and epigenetic factors from early life to later disease.** *Metabolism* 2008, 57 Suppl 2:S27–31.
4. Morse RH: **Epigenetic marks identify functional elements**. *Nat. Genet.* 2010, **42**:282–284.
5. He HH, Meyer CA, Shin H, Bailey ST, Wei G, Wang Q, Zhang Y, Xu K, Ni M, Lupien M, Mieczkowski P, Lieb JD, Zhao K, Brown M, Liu XS: **Nucleosome dynamics define transcriptional enhancers**. *Nat. Genet.* 2010, **42**:343–347.
6. Bonn S, Zinzen RP, Girardot C, Gustafson EH, Perez-Gonzalez A, Delhomme N, Ghavi-Helm Y, Wilczyński B, Riddell A, Furlong EEM: **Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development**. *Nat. Genet.* 2012, **44**:148–156.
7. Black JC, Van Rechem C, Whetstine JR: **Histone lysine methylation dynamics: establishment, regulation, and biological impact**. *Mol. Cell* 2012, **48**:491–507.
8. Wagner EJ, Carpenter PB: **Understanding the language of Lys36 methylation at histone H3**. *Nat. Rev. Mol. Cell Biol.* 2012, **13**:115–126.
9. White RJ: **Transcription by RNA polymerase III: more complex than we thought**. *Nat. Rev. Genet.* 2011, **12**:459–463.
10. Maston GA, Evans SK, Green MR: **Transcriptional regulatory elements in the human genome**. *Annu. Rev. Genomics Hum. Genet.* 2006, **7**:29–59.
11. Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, Afzal V, Ren B, Rubin EM, Pennacchio LA: **ChIP-seq accurately predicts tissue-specific activity of enhancers**. *Nature* 2009, **457**:854–858.
12. May D, Blow MJ, Kaplan T, McCulley DJ, Jensen BC, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Afzal V, Simpson PC, Rubin EM, Black BL, Bristow J, Pennacchio LA, Visel A: **Large-scale discovery of enhancers from human heart tissue**. *Nat. Genet.* 2012, **44**:89–93.
13. Novak K: **Epigenetics changes in cancer cells**. *MedGenMed* 2004, **6**:17.
14. Ting AH, McGarvey KM, Baylin SB: **The cancer epigenome—components and functional correlates**. *Genes Dev.* 2006.
15. Roadmap Epigenomics Consortium, Kundaje A, et al., **Integrative analysis of 111 reference human epigenomes**. *Nature* 2015, **518**:317–330.
16. Polak P, Karlić R, Koren A, Thurman R, Sandstrom R, Lawrence MS, Reynolds A, Rynes E, Vlahoviček K, Stamatoyannopoulos JA, Sunyaev SR: **Cell-of-origin chromatin organization shapes the mutational landscape of cancer**. *Nature* 2015, **518**:360–364.
17. Schultz MD, He Y, Whitaker JW, Hariharan M, Mukamel EA, Leung D, Rajagopal N, Nery JR, Urich MA, Chen H, Lin S, Lin Y, Jung I, Schmitt AD, Selvaraj S, Ren B, Sejnowski TJ, Wang W, Ecker JR: **Human body epigenome maps reveal noncanonical DNA methylation variation**. *Nature* 2015, **523**:212–216.
18. Singh R, Lanchantin J, Robins G, Qi Y: **DeepChrome: deep-learning for predicting gene expression from histone modifications**. *Bioinformatics* 2016, **32**:i639–i648.
19. Huether R, et al: **The landscape of somatic mutations in epigenetic regulators across 1,000 paediatric cancer genomes**. *Nat. Commun.* 2014, **5**:3630.
20. Tak YG, Hung Y, Yao L, Grimmer MR, Do A, Bhakta MS, O'Geen H, Segal DJ, Farnham PJ: **Effects on the transcriptome upon deletion of a distal element cannot be predicted by the size of the H3K27Ac peak in human cells**. *Nucleic Acids Res.* 2016, **44**:4123–4133.
21. Webster DE, Barajas B, Bussat RT, Yan KJ, Neela PH, Flockhart RJ, Kovalski J, Zehnder A, Khavari PA: **Enhancer-targeted genome editing selectively blocks innate resistance to oncokinase inhibition**. *Genome Res.* 2014, **24**:751–760.
22. Huether R et al.: **The landscape of somatic mutations in epigenetic regulators across 1,000 paediatric cancer genomes**. *Nat. Commun.* 2014, **5**:3630.
23. Mansour MR, Abraham BJ, Anders L, Berezovskaya A, Gutierrez A, Durbin AD, Etchin J, Lawton L, Sallan SE, Silverman LB, Loh ML, Hunger SP, Sanda T, Young RA, Look AT: **Oncogene regulation. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element**. *Science* 2014, **346**:1373–1377.
24. Herranz D, Ambesi-Impiombato A, Palomero T, Schnell SA, Belver L, Wendorff AA, Xu L, Castillo-Martin M, Llobet-Navás D, Cordon-Cardo C, Clappier E, Soulier J, Ferrando AA: **A NOTCH1-driven MYC enhancer promotes T cell development, transformation and acute lymphoblastic leukemia**. *Nat. Med.* 2014, **20**:1130–1137.
25. Zhang X, Choi PS, Francis JM, Imielinski M, Watanabe H, Cherniack AD, Meyerson M: **Identification of focally amplified lineage-specific super-enhancers in human epithelial cancers**. *Nat. Genet.* 2016, **48**:176–182.
26. Brazel AJ, Vernimmen D: **The complexity of epigenetic diseases**. *J. Pathol.* 2016, **238**:333–344.
27. Cavalli G, Misteli T. **Functional implications of genome topology.** Nat Struct Mol Biol 2013;20(3):290–9. doi: 10.1038/ nsmb.2474.

28. Horan M, Ballard J. **Application of Chromosome Conformation Capture (3C) to the Study of Human Genetic Disease.** eLS. Chichester: John Wiley & Sons, Ltd.; 2012. 112.
29. Schierding W, Cutfield WS, O'Sullivan JM. **The missing story behind Genome Wide Association Studies: single nucleotide polymorphisms in gene deserts have a story to tell.** Front Genet 2014;5:39. doi: 10.3389/fgene.2014.00039.
30. Hon G, Ren B, Wang W. **ChromaSig: a probabilistic approach to finding common chromatin signatures in the human genome.** PLoS Comput Biol 2008;4:e1000201.
31. Rye M, Saetrom P, Handstad T, et al. **Clustered ChIP-Seqdefined transcription factor binding sites and histone modi- fications map distinct classes of regulatory elements.** BMC Biol 2011;9:80.
32. Heintzman ND, Stuart RK, Hon G, et al. **Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome.** Nat Genet 2007;39:311–8.
33. Kleftogiannis D, Kalnis P, Bajic VB. **DEEP: a general computational framework for predicting enhancers.** Nucleic Acids Res 2015;43:e6.
34. Fletez-Brant C, Lee D, McCallion AS, et al. **kmer-SVM: a web server for identifying predictive regulatory sequence features in genomic data sets.** Nucleic Acids Res 2013; 41:W544–56.
35. Ernst J, Kellis M. **ChromHMM: automating chromatin-state discovery and characterization.** Nat Methods 2012;9:215–6.
36. Won KJ, Zhang X, Wang T, et al. **Comparative annotation of functional regions in the human genome using epigenomic data.** Nucleic Acids Res 2013;41:4423–32.
37. Won KJ, Chepelev I, Ren B, et al. **Prediction of regulatory elements in mammalian genomes using chromatin signatures.** BMC Bioinformatics 2008;9:547.
38. Andersson R, Gebhard C, Miguel-Escalada I, et al. **An atlas of active enhancers across human cell types and tissues.** Nature 2014;507(7493):455–61. doi: 10.1038/nature12787
39. Thurman RE, Rynes E, Humbert R, et al. **The accessible chromatin landscape of the human genome.** Nature 2012;489(7414):75–82. doi: 10.1038/nature11232.
40. He B, Chen C, Teng L, et al. **Global view of enhancer-promoter interactome in human cells.** Proc Natl Acad Sci USA 2014;111(21):E2191–9. doi: 10.1073/pnas.1320308111.
41. Hnisz D, Abraham BJ, Lee TI, et al. **Super-enhancers in the control of cell identity and disease**. Cell 2013;155(4):934–47. doi: 10.1016/j.cell.2013.09.053.
42. Davison LJ, Wallace C, Cooper JD, et al. **Long-range DNA looping and gene expression analyses identify DEXI as an autoimmune disease candidate gene.** Hum Mol Genet 2012;21(2): 322–33. doi: 10.1093/hmg/ddr468.
43. Dixon JR, Selvaraj S, Yue F, et al. **Topological domains in mammalian genomes identified by analysis of chromatin interactions.** Nature 2012;485(7398):376–80. doi: 10.1038/nature11082.
44. **VISTA Enhancer Browser, whole genome enhancer browser** [http://enhancer.lbl.gov/frnt_page_n.shtml].
45. **Human Epigenome Atlas** [https://www.genboree.org/epigenomeatlas/index.rhtml].
46. **Cosmic Catalogue of somatic mutations in cancer**, Available at: https://cancer.sanger.ac.uk/cosmic/ (2015).
47. **GENCODE - GENCODE Release Files** [https://www.gencodegenes.org/releases/10.html].
48. Chepelev I. et al., **Characterization of genome-wide enhancer-promoter intereactions reveals c0-expression of interacting genes and modes of higher order chromatin organization**. Cell 22, 490-503 (2012).
49. **NIH – National Institutes of Health,** [https://www.nih.gov/news-events/news-releases/nih-supported-researchers-map-epigenome-more-100-tissue-cell-types]
50. Ong C-T, Corces VG: **Enhancer function: new insights into the regulation of tissue-specific gene expression.** Nat. Rev. Genet. 2011, 12:283–293.
51. Karlić R, Chung H-R, Lasserre J, Vlahovicek K, Vingron M: **Histone modification levels are predictive for gene expression.** Proc. Natl. Acad. Sci. U. S. A. 2010, 107:2926–2931.
52. Li G et al.: **Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation.** Cell 2012, 148:84–98.
53. Pekowska A, Benoukraf T, Zacarias-Cabeza J, Belhocine M, Koch F, Holota H, Imbert J, Andrau J-C, Ferrier P, Spicuglia S: **H3K4 tri-methylation provides an epigenetic signature of active enhancers.** EMBO J. 2011, 30:4198–4210.
54. Tibshirani R., **Regression Shrinkage and Selection vai the Lasso.** (1996), 58: 267-288.
55. MatLab **SpaSM package**, DTU Informatik [http://www2.imm.dtu.dk/projects/spasm/]