Title
Exploiting Regulatory Heterogeneity to Systematically Identify Enhancers with High Accuracy

Author
Arbel, Hamutal

Publication Date
2017

Peer reviewed|Thesis/dissertation

Exploiting Regulatory Heterogeneity to Systematically Identify Enhancers with High Accuracy

By

Hamutal Arbel

Dissertation submitted in partial satisfaction of the

requirements for degree of

Doctor of Philosophy

In

Engineering – Applied Science and Technology

In the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Peter J. Bickel, chair
Professor Haiyan Huang
Professor Mohammad R. K. Mofrad

Fall 2017

Abstract

Exploiting regulatory heterogeneity to systematically identify enhancers with high accuracy

by

Hamutal Arbel

Doctor of Philosophy in Applied Science and Technology

University of California, Berkeley

Professor Peter Bickel, Chair

Enhancer discovery through computational means has long been a goal of the genomics community. The tools developed for this purpose, however, tend to underperform when tested on completely held out test sets. Here we use the pregrastrula patterning network of *Drosophila melanogaster* to demonstrate that loss in accuracy in held out data results from heterogeneity of functional signatures in enhancer elements. We show that at least two classes of enhancer are active during early *Drosophila* embryogenesis and that by focusing on a single, relatively homogeneous class of elements, extremely high (>98%) prediction accuracy can be achieved in a balanced, held-out test set. The homogenous set is composed predominantly of enhancers driving multi-stage, large segmentation patterns in the early embryo, and hence we term them segmentation driving enhancers (SDE). Prediction is primarily driven by transcription factors DNA occupancy with almost no power derived from histone modifications, including H3K27ac, casting further doubt on the utility of histone modifications to demarcate enhancer elements. The transcription factors used in the prediction process constitute over half of the transcription factors identified in genetic screens as patterning the early embryo, and hence provide a remarkably expansive view of this process. Applying this method to a genome-wide scan, we predict 1,600 SDEs, 916 of which are novel, covering approximately 1.6% of the euchromatic genome. We verified these predictions by testing 41 novel SDEs using *in situ* whole embryo imaging of stably integrated reporter constructs. We confirmed 39 of these predictions, a 95% precision on a genome-wide scan with an estimated recall of 98%, indicating that our reported collection of SDEs may be close to comprehensive.

Acknowledgment

First and foremost, I would like to thank my “acting adviser”, James Bentley Brown, for all the support, guidance and patient explanations you have given me throughout this process, without which this work could never have been done. Through his mentorship, I have become a better scientist, and through his friendship I hope I became a better person.

I am forever grateful to Prof. Peter Bickel for taking a chance on accepting me as a foreign chemistry student seeking a degree in a new field, and helping me to grow academically into a scientist ready to graduate. He is one of the kindest brightest people I have been privileged to know, and his counsel and guidance meant more to me than I can say.

I would also like to thank Prof. Haiyan Huang for all her patient tutelage in and out of the classroom, particularly as I was taking my first (and belated) steps into the world of statistics.

To Sue Celniker, who graciously hosted me in her lab for 6 years, I wish to thank for her support and assistance, both personal and academic. It has been an honor and a delight to work beside her for all these years.

This work was only possible thanks to the combined efforts of the BDNTP consortium. I wish to acknowledge and thank Mark Biggin, who led the project, for all his insight and support. To Sue Celniker, who led the biological section of the work. To Bill Fisher and Ann Hammond for creating the drosophila embryos validating the result and to Soile Keranen who imaged them and tirelessly annotated expression patterns in our validation as well as in the original Stark lab samples.

I would also like to thank all my past and present lab mates Nathan, Marcus, Ke, Omid, Sumanta and Sarah, who made my time in the lab productive and enjoyable. I would especially like to thank Marcus Stber, the source of all awk knowledge, who endured my pestering questioning for six years with good grace and unfailing willingness to stop everything until we found an answer.

Thank you all for everything.

Table of contents

List of Figures

List of Figures (continued)

List of Tables

Commonly Used Terms and Abbreviations:

TF : Transcription factor

RF: Random Forests

Oob: out of bag, i.e. – held out test data

TP: True positive

FP: False positive

FN: False negative

TN: True negative

Recall: Fraction of positives captured by prediction, $\frac{TP}{TP+TN}$

PR: Precision, fraction of true positives in predicted positives, $\frac{TP}{TP+FP}$

FDR: False Discovery Rate, or the fraction of false positive in the prediction. 1 – PR

Accuracy: Fraction of correct predictions, $\frac{TP+TN}{TP+FP+FN+FP}$

Type I error: Incorrect rejection of a true null hypothesis. In this work, refers to incorrectly classifying a non-enhancer DNA segment as an enhancer.

Type II error: Failure to reject a false null hypothesis. In this work, refers to incorrectly classifying an enhancer as an inactive DNA segment.

Enhancers: DNA segments which acts to upregulate the expression of a gene under the correct conditions, either temporal, spatial or external signaling. For the purpose of this work, it refers to DNA segments inducing gene expression in stage 4-6 drosophila embryos.

Active enhancers: Enhancers acting to upregulate a gene in the present conditions.

Non-enhancers: Any DNA segment not acting to upregulate gene expression. For the Purpose of this work, this refers to DNA segments not inducing gene expression in stage 4-6 drosophila embryos.

Class I enhancers: The set of enhancers that can be correctly classified at least 75% of the time. AKA Segmentation driving enhancers or SDE's

Class II enhancers: The set of enhancers which cannot be reliably correctly classified at least 75% of the time

SDE: Segmentation driving enhancers

Prediction score: fraction of trees classifying a DNA segment as being an enhancer.

INTRODUCTION:

Genomic regulation and enhancers

The fundamental question of genetics has shifted in the last few decades from "which genes are present in the genome" to "which genes are expressed in the genome", marking a fundamental emphasis on regulation. While the genomic composition of an organism defines the potential of genomic variability, it is the temporally, localized and environmentally attuned manner of expression which allows organisms to thrive in a verity of conditions, and indeed for multicellular organisms to exist. Advances in sequencing techniques and the success of varying sequencing projects have provided us with ample knowledge of the genomic makeup of multiple species, from human to platypus, with more added to the list seemingly daily [1]. RNA sequencing (RNA-seq) allows us direct observations to the genomic expression of cells at different times and conditions, simultaneously revealing which genes are transcribed and challenging our definition of what a gene is by exposing how much of the genome is indeed transcribed [2]. But even if one learns what genes are present in the genome - and even what genes are expressed in a variety of conditions, the complexity of the regulatory mechanism still defies complete elucidating, making predictions of gene activation at different conditions a yet distant goal.

In eukaryotes, transcription is mostly governed by Polymerase II initiating the transcription process upon binding to the DNA proximal to the transcription start site (TSS), a binding aided by sequence specific transcription factors which binds the promotor. PolII association with the 6 general transcription factors TFIIA-H and other factors to form the preinitiation complex (PIC) helps in its positioning, stabilizes the binding, opens the DNA and even protects the exposed single-stranded DNA from damage by insulating it from the environment. The universal nature of this action allows for little in regulation control, which is instead provided by the sequence specific TF's binding the promotor, and through the massive mediator complex, which helps recruit and position the different PIC elements. The mediator is a massive and versatile complex, composed of over 20 different core proteins with varying possible compositions[3]. Mediator is essential to transcription initiation and elongation; its size and versatility allows it to interact with many proteins at once, and It can change the local chromatin structure and DNA three-dimensional structure - allowing many more elements of the genome to participate in regulating distant genes, vastly increasing the possibility of genomic regulation.

Genomic elements participating in gene regulation thus include trans-elements operating proximal to the gene being regulated, such as promotors, and more distance cis-regulatory elements or modules (CRM) operating at a distance (in the 2D), such as enhancers[4,5], silencers [6,7] and insulators[8]. True to their name, these elements act to enhance or repress expression, and to insure correct localized regulation by isolating enhancers from other proximal genes. The complexity of the regulatory architecture varies among species, growing with that of the organism. In Drosophila, a gene may be regulated by several enhancers, silencers as well as by transcription factors bound to the

promotor, and it is the combined action of all these factors which determines whether a gene is transcribed.

Enhancers are DNA sections rich in transcription factor binding sites which may interact with the promotor and polII directly or through cofactors, helping to form the PCI or mediator complex, and thus initiate, stabilize and localize transcription. Functionally, enhancers are genomic modules capable of driving expression when placed near a gene in a precise spatial and temporal manner, though enhancer location relative to the gene they regulate is extremely variable. Enhancer regions can be found a few hundred base pairs to millions of base pairs upstream or downstream to the gen being regulated, in either orientation; they may be placed within the intron of a gene being regulated or in the intron of a distant gen, or even on a different chromosome [9,10]. Their proximity to the gene is in the three-dimensional space, not in the 2-dimensional sequence, making enhancer discovery a challenging proposition.

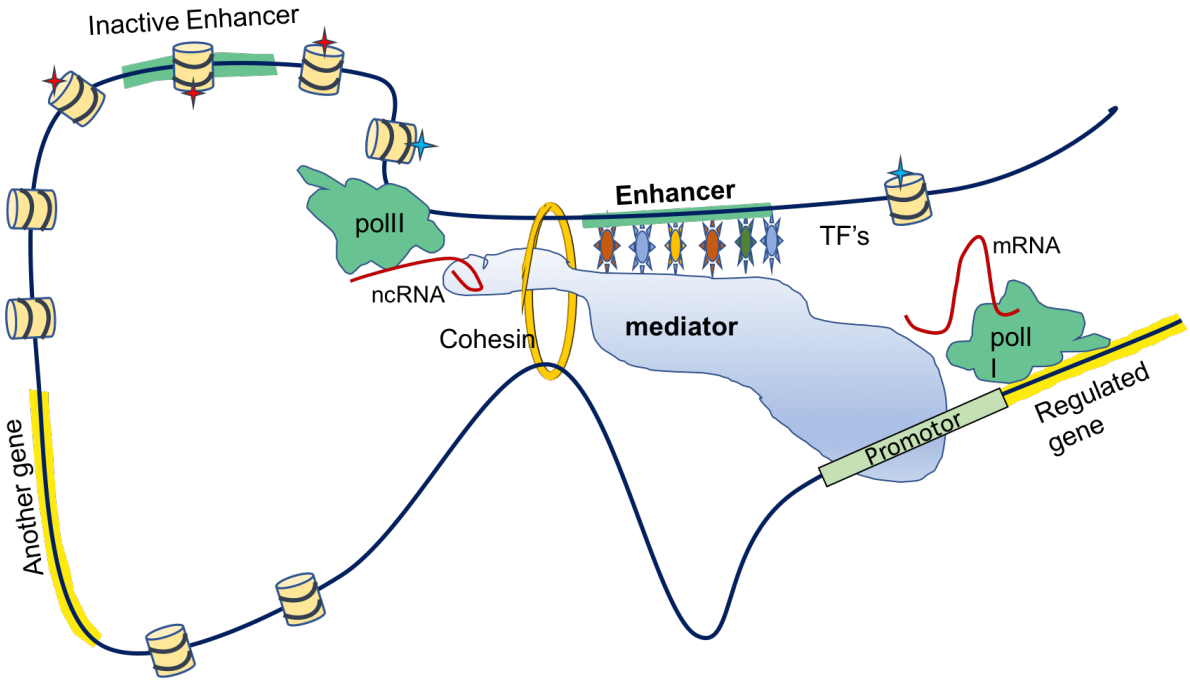

Figure 1: *Illustration of an active enhancer (in green) brought into proximity of the promotor of its targeted gene, past an irrelevant gene (in yellow). The proximity is held by the Cohesion complex, holding the DNA together. The different transcription factors on the enhancer interact with the mediator complex, help to assemble and stabilize it. Enhancers are also often associated with paused polII, transcribing non-coding (ncRNA) which is thought to aid in mediator stabilization. These are often bi-directional, as they are not guided by a promotor docking. The enhancer regions are unpacked and do not contain nucleosomes, and the nucleosomes around an active enhancer are often marked by molecular modifications (blue) distinct from those near an inactive enhancer (red).*

A key determinant of enhancer activity is the abundance of the varying of transcription factors which bind it. This is indeed the primary patterning and symmetry braking tool responsible for embryonic development and differentiation. But while we know that that order, orientation, competition and cooperation amongst transcription factors determines regulation, the details of these interactions are not yet fully resolved even in simple systems. Chromatin structure and modeling – i.e. DNA accessibility – is another key factor in determining activity[11]. It has also been shown that chemical modifications such as methylation and acetylation of the histone core proteins mark and help regulate cellular processes such as transcriptional regulation, and a great deal of work has gone in recent years to identify the epigenetic enhancer signature in the nucleosome data, leading to the classification of three distinct enhancer states: inactive, active and poised. The inactive enhancer is identified by a tightly packed chromatin structure, while the poised enhancer, while still inactive, is identified by an open chromatin state and a particular set of histone modifications, such as the tri-methylation of the lysine on protein H3 of the nucleosome core (H3K27me3). The tri-methylation is replaced by acetylation (H3K27ac) in the active enhancer by a p300, a sub unit of the mediator complex and a common target of transcription factor [12].

Another possible mark of enhancer activity may be the presence of polII or the proximity to a PolII peak, as it has been shown that PolII may pause when passing an enhancer [13]. In passing and localizing, PolII was also found to transcribe short, bidirectional RNA transcripts of the enhancer regions. Whether those play a functional role in transcription has not yet been resolved, but it has been suggested these may indicate enhancers [14]. It has also been suggested that as enhancers are functional units, there will be evolutionary pressure to ensure slower divergence in them then in other parts of the DNA, but though several attempts have been made to locate enhancers by utilizing conservation scores, results have been mixed.

Tools that measure predictive accuracy in terms of indirect evidence of enhancer activity, e.g. H3K27ac positive regions or p300 enriched regions, often display excellent accuracy based on these limited criteria [15-18]. When algorithms are benchmarked on held-out in vivo tests of enhancer activity, however, positive predictive power on genome-wide scans in metazoan systems have been lower than expected. By targeting transcription factors in a specific biological processes a precision of 56% was achieved in a randomly selected sample through transient transfection [19]. Higher precision has been reported when tests were confined to the top of the ranking list [20], but such numbers are unlikely to represent the precision of the prediction set as a whole. In general, precision in metazoan system rarely exceed 40% [15,16,21]. There are several possible explanations for this. For instance, transient in vivo enhancer assays often employed to test predictions may suffer a high false-negative rate due to the loss of local chromatin context. However, studies indicate that local features are the principle drivers of the competency of a genomic element in an enhancer assay [21] [22]. Alternatively, the data provided to the prediction algorithms might be insufficient: for example, while H3K27ac can partially distinguish between active and poised enhancers [12], it remains unclear whether any chromatin mark or combination of chromatin mark uniquely identifies enhancers among all sequences in a genome [19,23]. In mammals p300-mediated acetylation of H3K27 is sufficient to activate gene expression from promoters and known enhancers [24], but it is unclear whether p300 activity is sufficient to induce an arbitrary genomic element to act

as an enhancer for proximal genes. Further, enhancers that lack H3K27ac and admit patterns of hypermethylation are found in vertebrates and are known to be essential during early development [25]. Hence, there may be more than a single class of genomic element that drives patterned expression, or, more precisely, the term “enhancer” may encapsulate a mechanistically diverse class of functional elements. Transcription factor (TF) occupancy is a better predictor of enhancer activity than canonical chromatin marks (including H3K27ac, H3K4me1, and H3K4me3) in mouse and humans [19]. One could envision that groups of TFs may partition enhancers into mechanistic sub-types, and thus such heterogeneity could explain the difficulties encountered to date in the computationally identifying enhancer elements from genomic data.

Drosophila Melanogaster embryogenesis

The Drosophila Melanogaster fruit fly undergoes a complex set of developmental processes, with 3 distinct larval stages and pupation, taking about 2 weeks from egg to adult.

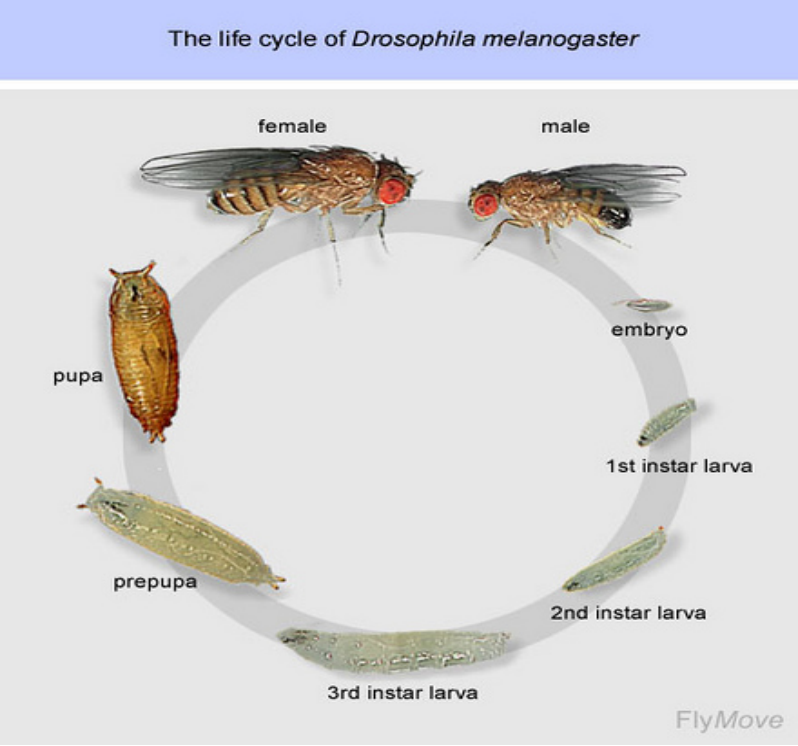

Figure 2: *the drosophila melanogaster life cycle, from embryo to adult fly. The embryonic stage last less then 24 hours, resulting in the first instar larva. multiple stages of maturation are accompanied by shedding of larval outer shell, and culminating in pupation. The process take ~3 weeks (depending on temperature), and the resulting adult begins procreation within hours of hatching.*

the development of the egg is commonly referred to as the embryonic stage in drosophila development, and occurs in less than a day (20 – 24 hours, depending on

temperature) through 17 distinctly characterized developmental stages. At stage 1, the fertilized nuclei begin rapid rounds of division, from 2 nuclei at the commencement of stage 2, 25 minutes (at 25^{oC}) post laying, to ~6000 cells at the close of stage 4 (80-90 minutes). This amazingly rapid multiplication represents the fastest cell division in any metazoan. At stage 4 the nuclei migrate periphery of the egg and broad differentiation patterns begin to appear, as embryonic protein production starts, followed by formation of cellular walls (stage 5, 130-170 minutes, 12-14 cellular division) and appearance of segmentation and narrowing of the broad pattern. At stages 4-6 (80 – 180 minutes) all of the embryo cells are located at the periphery, and it is at these stages that pattern, and segmentation and differentiation commences. A small cohort of ~30 spatially patterned transcription factors drive body patterning in concert with another 30 or so ubiquitously expressed sequence specific transcription factors [26-36]. The critical stage of symmetry breaking in the embryo is accomplished a-priori in the fly, with a transcription factor gradient established by nascent cells attached to one side of the embryo depositing a *Bicoid* transcription factor gradient along the anterior-posterior (a-p) axis (**Fig. 3**). Additional transcription factors, such as Caudal and Nanog, are also maternally deposited creating complementary gradients along the anterior-posterior (A-P) axis. A separate set of maternally deposited transcription factors, such as Dorsal and Twist, pattern the dorsal-ventral (D-V) axis.

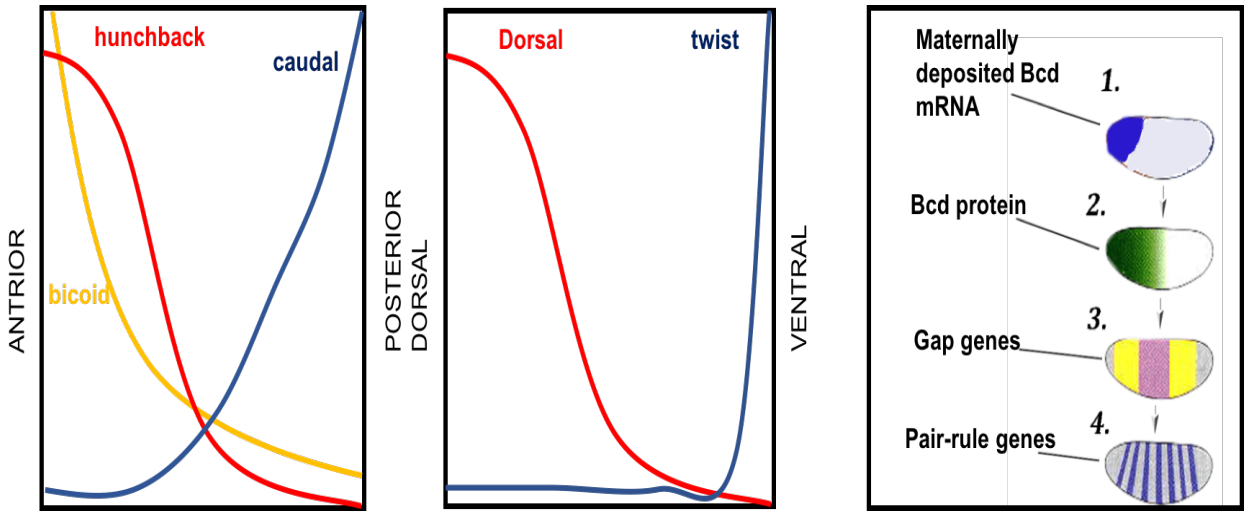

Figure 3: *Cartoon representation of a transcription factor gradients along the drosophila embryo, along the A-P* ***(a)*** *or the D-V* ***(b)*** *axis.* ***(c)*** *distribution and effect of Bicoid gradient along the embryo at different stages: (1) mRNA is maternally injected into the embryo by the attached nascent cells, gradient forms by diffusion (2) Bicoid protein is produced by the embryo as the nuclei migrate to the periphery, and the gradient in production follows the mRNA diffusion. (3) The varying levels of bicoid and other proteins form broad patterns of zygotic expression in stage 4, in particular with the regulation of transcription factors responsible for patterning. The combined action of the maternally deposited TF's and the zygotic transcribed ones works to narrow the broad patterns into finely delineated regions.*

Over a 90 minute period corresponding to developmental stages 4 and 5, these proteins act in concert with zygotically expressed A-P and D-V transcription factors to refine initially broad patterns of transcription into narrower striped patterns that define the basic segmental body plan of the fruit fly [37], known as the *gap genes* or *pair-rule* genes. The pregrastrula fly network is thus a particularly well defined model system for studying the relationship between transcription factor DNA binding and spatially patterned enhancer activity.

Random Forests

Random Forest is a supervised machine learning tool for classification or regression based on an ensemble of decision trees in which a randomization in data and features is introduced [38]. In each tree, a subset of data points is sampled with replacement, while the other data points serve as held out test set, commonly known as "out of bag" data points. At each node, a small fraction of the features is randomly selected and the data is split along the criteria which minimizes the variability in the resulting subdivided groups. The votes of all the trees in the forest are summed to give the overall estimation of the regression or classification of the data. More formally [39], for classification set, let: $X \in [0,1]^p$ be i.i.d observation , with unknown binary response vector $Y = \{0,1\}$ we wish to estimate. Given a training set $D_n = (X_1,Y_1)..(Xn,Yn)$, classification of point x by tree j in the forests is a function of x, D_n and a random variable Θ_j, i.e. $m_n(x,D_n,\Theta_j)$ where Θ is an i.i.d variable, deriving from the randomization of the tree. This randomization is achieved by sampling $a_n \in (1..n)$, a subset of data points with replacement from the data set and choosing mtry $\in \{1..\ p\}$ splitting criteria to be tried at each cell (leaf) A to be split.

At each cell A containing N(A) points, the split criteria along any feature $j \in \{1..p\}$ at value z creating 2 cells A_L and A_R. The classification criteria used by the forest is the Gini impurity measure:

$$L_{class,n}(j,z) = p_{0,n}(A)\ p_{1,n}(A) - (N_n(A_L)/\ N_n(A)\ )\ p_{0,n}(A_L)p_{1,n}(A_L) - (N_n(A_R)/\ N_n(A)\ )p_{0,n}(A_R)\ p_{1,n}(A_L)$$

Where $p_{0,n}$ and $p_{1,n}$ are the empirical probabilities of data of class 0,1 respectively falling into cell A. The optimal cut is that which minimizes the impurity at the resulting cells, such that the best split criteria satisfying is that which maximizes $L_{class,n}$:

$$(j_n^*, z_n^*) = \ argmax_{j \in mtry} L_{class,n}(j,z)$$

For classification, it has been recommended [40] that cut be continued until each node contains a single data point with mtry = $\sqrt{p}$.

The random tree estimator is thus given by the majority of points which falls in region A:

$$m_n(x, \Theta_n, D_n) = \begin{cases} 1 & if: \sum_{i=1}^{n} \mathbf{1}_{x_i \in A, y_i = 1} > \sum_{i=1}^{n} \mathbf{1}_{x_i \in A, y_i = 0} \\ 0 & otherwise \end{cases}$$

As each tree is a week classifier casting a vote, the Random Forest estimation of a forest of M trees is:

$$M_{m,n} = \begin{cases} 1 & if \ \frac{1}{M} \sum_{j=1}^{M} m_n(x, \Theta_j, D_n) > \frac{1}{2} \\ 0 & Otherwise \end{cases}$$

The algorithm of random forest can be summarized as:

1) Randomly select a_n data points from training set D_n
2) Select mtry $\in$ {1..p} features to seek optimal cut
3) For a cell A, find feature-value pair (j,z) which minimizes the Gini impurity measure L_n(j,z). return the 2 cells A_L and A_R.
4) Repeat until each cell A holds the desired *nodesize* number of data points
5) Calculate $m_n(x, \Theta_n, D_n)$ and classify the out of bag data points, i.e $D_n \cap a_n$
6) Calculate prediction of test set
7) Repeat M (large number of times)
8) Aggregate results to determine the forest estimator $M_{m,n}$ for the out of bag data points, calculate and return the forest error rate
9) Aggregate results to determine the forest estimator $M_{m,n}$ for the test set and return the predicted classification.

The out-of-bag error gives class-wise error rate estimation, or an overall error estimation by averaging over. It also allows tuning the random forest parameters.

Random forest has been widely used in a variety of application. One of the most commonly used classification methods, it has been used extensively in biological computation[39,41], including enhancement discovery[18]. A benchmark comparison of 179 classifiers in 17 families on 121 data sets recently found it the best family of classifiers [42]. Yet the theoretical underpinnings and properties of Random Forests have proven difficult to elucidate. While no true understanding of its properties exists to date, several theoretical results have emerged. In his initial report in 2001, Breiman presented an upper bound of Random forests generalization error, which decreases as the forest grows. The classification error also depends the strength of the trees, and on their correlation: the higher the correlation the higher the error. A relationship between Random Forests and adaptive nearest neighbor has been has been shown[43], and this and other frameworks have been perused to establish the methods general properties[44]. A link between the error of the infinite and finite forest has also been established in 2015 [45]. Biau, Devroye and Lugosi established in 2008 that averaging classifiers are consistent, and that certain types of Random Forest (including classic Random Forests discussed in this work) are

consistent[46]. Wager has further showed that the Random Forest prediction are asymptotically normal, and provided a method to consistently estimate the errors[47].

One reason for the popularity and power of Random Forests is their ability to produce consistent classification in with high number of features, and even when the number of features exceed the training set size. Biau[48] explored the consistency of Random Forests in the simplified centered tree model proposed by Breiman[49]. Biau showed the model convergence rate depends only on the number of strong predictors S, regardless of how many noisy additional features are present. While this result applies to a very simplified version of the model, Jean-Philippe Vert also suggests an explanation to Random Forest's additivity to sparsity in the context of additive regression models[50]

While the decisions path behind the forest voting is not readily transparent, Random Forests do provide two measures of importance[40,51]. *Mean decrease impurity* (also known as *mean decrease Gini* or *Gini importance*) measures the total decrease in node impurity stemming from splitting a node on a feature, averaged over all trees. Note that as the method requires repeated attempts to split the node on a feature, features for which more attempts can be made are more likely to be selected by chance. This makes the mean decrease Gini biased towards categorical data with more classes, or numerical data with higher values. Assuming infinite sampling, it has been shown that this importance measure is unaffected by the presence of irrelevant features in the set [52]. Mean decrease accuracy measures the difference in out-of-bag error in all trees upon permutation of feature values. While it is not prone to the bias towards larger categorical features or larger scale variable mean decrees Gini is prone to, Mean decrease accuracy measure performance is more sensitive to correlated features. [53], as correlated features may mask each other's importance upon permutation. While not as sensitive, mean decrees Gini is also prone to underestimate importance of correlated data [54,55]

Another measure related to mean decrease accuracy is the Random Forest local importance, defined as the decrease of accuracy calculated for each member of the training set $X_i \in D_n$ separately, rather than for the whole set. This measure can be used to understand the importance of each feature in the classification of each class.

RESULTS:

Data, Feature and feature selection

Whole-embryo imaging data from three sources were combined in our data set. In all cases, DNA segments constructs to be tested were inserted into a genomic construct near a reporter gene whose expression determine their enhancer activity. Kvon et al. [56] conducted a semi-automated in situ hybridization and imaging of 7705 genomic regions (http://enhancers.starklab.org/) specifying their activity at developmental stages throughout early embryogenesis. While the high throughput nature of the work allowed for an unprecedented number of genomic areas to be tested, the small number of embryos per collection plate leads to increased misclassifications in the data. The activity of an additional 282 genomic segments was manually tested by the BDGP group [57] (unpublished data) (**Table 1**). Altogether, 7987 genomic regions were examined and 731 were experimentally found to be enhancers in Drosophila embryonic stages 4-6[58]. As the BDGP genomic segments were tested by carful practiced imaging of dozens of embryos at each stage, those were taken as ground truth. By manually comparing the labeling of overlapping genomic regions in the BDGP database with the larger Stark lab data we estimate a 10% false negative rate in the latter.

As this work focuses on stages 4-6 embryogenesis, all 7256 regions which were found not to induce expression at those stages were considered non-enhancers, though 4031 of them were found to be enhancers at later stages. This was done both because it is desirable to be able to separate enhancers using all data available if at all possible, and since most feature data (described below) is stage-specific, so enhancer prediction is expected to be independent (this assumption is explicitly tested later in this work).

Table 1: all features used in prediction

Segments #	Segment type	Description
7705	Vienna tiles	High-throughput automated scan; high volume but error prone
282	BDGP	small number of very high quality manual annotation
7987	**Training data set**	**Full training set used in this analysis**
Training set composition:		
731	Enhancers	induce test gene expression at stages 4-6
7256	Non-enhancers	Don't induce test gene expression at stages 4-6
Non-enhancer set is composed of:		
4031	Late-stage enhancers	while not inducing test gene expression at stages 4-6, do induce it at late stages of embryogenesis.
3225	absolute non-enhancers	Segments not inducing test gene throughout embryogenesis

The different DNA segments cover all the drosophila chromosomes, spanning euchromatic and heterochromatic region. While the size of the most genomic segments was ~2-2.5Kb, the sizes ranged from 0.1-4.5Kb for the entire training set. The size of the enhancer containing segments ranged from 0.5-4.5Kb, with the size of the contained enhancer unknown (**Fig 4**).

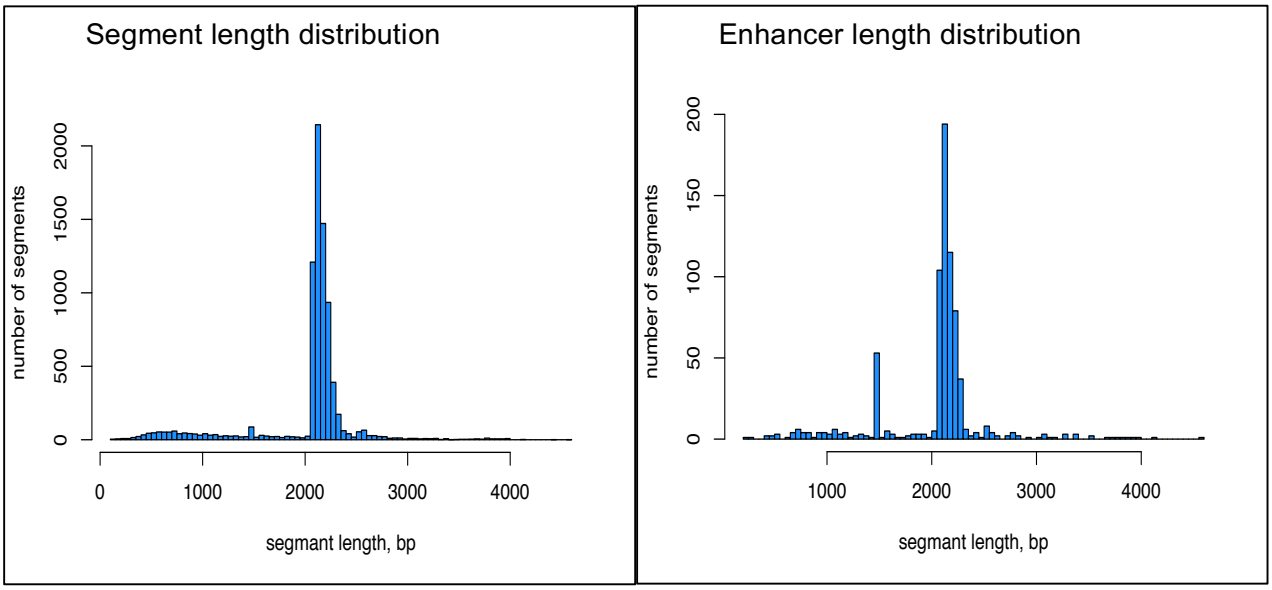

Fig. 4: *Size distribution of* ***(a)*** *the training set or* ***(b)*** *the enhancers in the training set.*

Features used in the initial model included ChIP-chip data for 20 of the ~30 A-P and D-V transcription factors shown to be important for transcriptional patterning in Drosophila 4-6 embryo [59-61]; ChIP-seq data for the ubiquitous transcription factors ZLD and z as well as 45 chromatin proteins and histone modifications in cycles 8-14 [62] gave us additional feature data, as did DNase accessibility data [11,63,64] and evolutionary conservation scores [65-67]. Also considered were the presence of bidirectional RNA transcripts, exon and intron coverage, distance to RNA Polymerase II ChIP-chip binding peaks, and distance to transcription start sites. A summarized list of features is presented in **Table 2**. For a full list and description please refer to **methods.**

Table 2: all features used in prediction

	Features included
Histone and Histone modifications (cell cycle indicated by _c#)	H3_c12 H3_c14a H3_c14c H3_c8 H3K18ac_c12 H3K18ac_c14a H3K18ac_c14c H3K18ac_c8 H3K27ac_c12 H3K27ac_c14a H3K27ac_c14c H3K27ac_c8 H3K27me3_c12 H3K27me3_c14a H3K27me3_c14c H3K36me3_c12 H3K36me3_c14a H3K36me3_c14c H3K4me1_c12 H3K4me1_c14a H3K4me1_c14c H3K4me1_c8 H3K4me3_c12 H3K4me3_c14a H3K4me3_c14c H3K4me3_c8 H3K9ac_c12 H3K9ac_c14a H3K9ac_c14c H3K9ac_c8 H4K5ac_c12 H4K5ac_c14a H4K5ac_c14c H4K5ac_c8 H4K8ac_c12 H4K8ac_c14a H4K8ac_c14c H4K8ac_c8 wt_H3 wt_H3K18ac wt_H3K4me1
ChiP-seq input files (sequencing control, used here as importance control)	input_c12 input_c14a input_c14c input_c8
A-P Transcription Factor data (duplicates indicated by number)	bcd1 bcd2 cad1 D1 ftz3 gt2 h1 h2 hb1 hb2 hkb1 hkb2 hkb3 kni1 kni2 kr1 kr2 prdBQ prdFQ run1 run2 slp1 tll1
D-V Transcription Factor data (duplicates indicated by number)	da2 dl3 mad2 med2 shn2 shn3 sna1 sna2 twi1 twi2
Ubiquitous Transcription Factor data	z2 zld
Transcription factor combinatorics	Sum of all TF, sum of all duplicates for: bcd twi sna shn run kr kni hkb prd hb h
Conservation scores	Mean, Max sliding window of:200, 500 and 1000, longest continues stretch
Zld ChiP-seq measurements	Mean, Max sliding window of:200, 500 and 1000, longest continues stretch
DNA accessibility	dnase, dnase2
Bi-directional RNA binding	Distance, absolute distance, maximal signal
Exon/intron data	Coding Exons Coverage, All Exons Coverage, Introns Coverage, binary indicators for weather segments contain exons, coding exons or introns
Transcriptional data	Distance to PolII binding peak, distance to closest transcription start site

.

For the purpose of analysis, a single value or small set of values was needed to summarize the base-wise information available. This task was complicated by the uncertainty in the enhancer boundaries inside the fragments – averaging on a long segment containing a short enhancer will dilute the signal relative to an identical enhancer tested with a shorter construct. Likewise, normalizing for segment length is unhelpful as the relevant normalization is the unknown contained enhancer. Therefore, maximal values were chosen as a more robust measure. While it is not optimal as it is more prone to outliers and experimental errors, yet in the absence of any knowledge of the underlying size of the enhancers and given the vast heterogeneity in the genomic segments tested (**Fig. 4**) this was found to be the least biased solution, when a signal value was needed to identify ChIP-chip and ChIP-seq signal strength.

By the same token, identifying the distance to and prevalence of genic constitution such as number of genes, number of introns and exons, distance to PolII binding or determining the nearest gene and so forth are also impeded by our ignorance as to the true genomic boundaries of the enhancers. It is impossible to determine which of the several genes overlapping the tested region, if any, also overlap the embedded enhancer or enhancers, or which overlapping gene is closer to it, particularly as the orientation of the putative enhancer is also unknown. In the absence of a data driven solution to this problem, we chose an inclusive rather than an exclusive approach; In the absence of a mechanism to choose between genes overlapping the segment, all overlapping genes were considered. If no gene overlapped the tested genomic segment, the nearest gene on either side of the fragment was included in the analysis. For the purpose of counting introns/exons and calculating genomic distances, the entire segment was treated as a single enhancer with those boundaries.

With this data, we trained and tested Random Forest classifier, a supervised machine learning approach based on an ensemble of decision trees [38-40] . To reduce parameter number and prevent overfitting, a preliminary culling of the feature sets was performed based on the error rates in their presence and absence (**see methods**). We found transcription factors and histone modification data were sufficient to minimize the error rate in the absence of all other feature sets in both held-out training or out-of-bag (oob) data and in test data (**Fig. 5a-b**). We note that while DNase accessibility did not contribute to Random Forest predictive power in the presence of TF binding data, it significantly improved predictive power in its absence, suggesting multi-collinearity in the data. Conversely, conservation scores did not contribute to the predictive power in any fitted model, and error rate utilizing solely conservation scores was neared 50%, suggesting conservation is not a feature of enhancers in the drosophila embryo. As data sets beyond TF binding data and histone modification added little predictive power, they were abandoned and not included in subsequent analyses.

To increase prediction accuracy in a data set highly enriched in non-enhancers, a forests-voting sampling scheme was developed, were multiple forests built with equal number enhancers and non-enhancers were used in the prediction (**Fig. 5c**) (**see methods**).

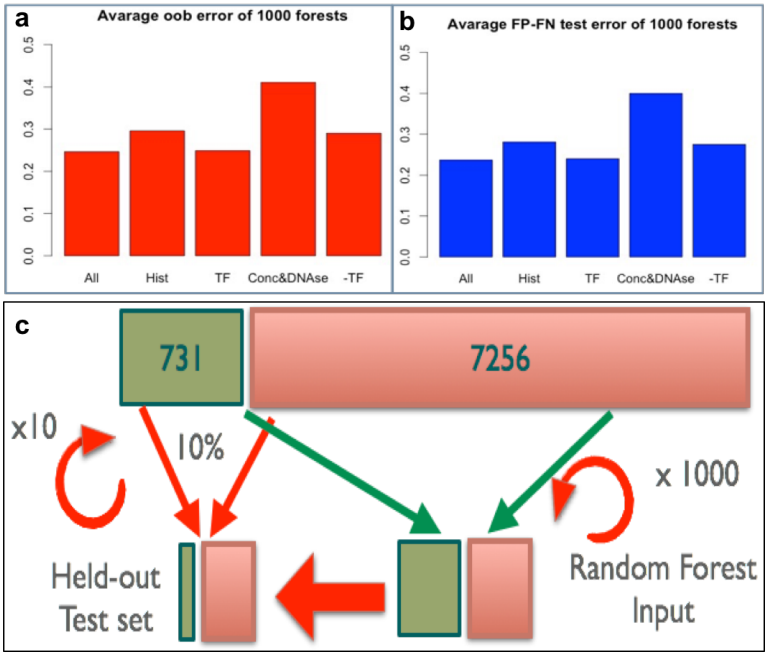

Figure 5: *average error rate of the out of bag data* ***(a)*** *or in the test set* ***(b)*** *of 1000 forests of 500 trees trained on the full data set (all), histone and histone modification only (Hits), transcription factors data set only, (TF), with conservation and DNase accessibility only (Conc & DNase), and with all the above except transcription factor data (-TF). It is clear most of the signal is located in transcription factor data, seconded by signal held in the histone modifications. There is no difference in performance between a set containing only histones and that containing all data sets besides TF data, emphasizing the redundancy of other data sets. (c) The Random Forests sampling scheme. 10% of the data serves as a held-out set, with 1000 balanced (equal number enhancers and non-enhancers) samples randomly selected from the remaining data used as training set of 1000 forests, with the held-out data serving as a test set. The scheme is thus repeated with the next 10% of the data held out, until each segment in the data set is predicted by 50,000 trees from 1000 different forests for which it was a completely held out test set.*

Heterogeneity among enhancer elements

With our optimal feature set our error rate in a single forest was nearly 30%, performance of the Forest voting probabilities indicates a likely success rate comparable to others in literature, indicated by the area under the ROC curve, AUC=0.82 (**Fig.6A**). However, while the overall predictive power falls short of that required for predicting enhancers genome wide, some enhancers were consistently correctly classified, while others were consistently misclassified. Hypothesizing that the model's poor performance may be due to heterogeneity in the enhancer set, enhancers were separated into two classes. **Class I Enhancers** contained the 358 enhancer segments that were correctly classified 75% of the time and **class II Enhancers** containing the 373 which were not. When excluding class II enhancers from test sample, the single forest error rate drops to ~3%, and the area under the ROC curve is ~0.99 (**Fig.6A**). When excluding Class I

enhancers, errors of a single forest are ~40%, and the roc curve indicated performance only marginally better than random guessing (**Fig. 6A**). To establish that the enhancer heterogeneity is data-driven and not an artifact of our choice of method, logistic regression and naïve bays models of the data were also constructed. In both cases the removal of the class II enhancer set significantly improves the model's predictive power (**Fig.6B**). Interestingly, the effect of retaining and removing class I and class II enhancers appears to have almost identical effect regardless of the method, and indeed the ROC curves are nearly overlapping (**Fig.6B**). This is particularly noteworthy as the underlying assumption of both models - primarily, feature additivity and independence – are unlikely to be present in the data, yet both perform as well as Random Forests, which does not require such assumptions. This may indicate that the problem of enhancer discovery becomes relatively trivial once heterogeneity is accounted for. Precision-recall curves do show Random Forests is the better classifier, as will be discussed in detail later in the text (**Fig. 27-29**).

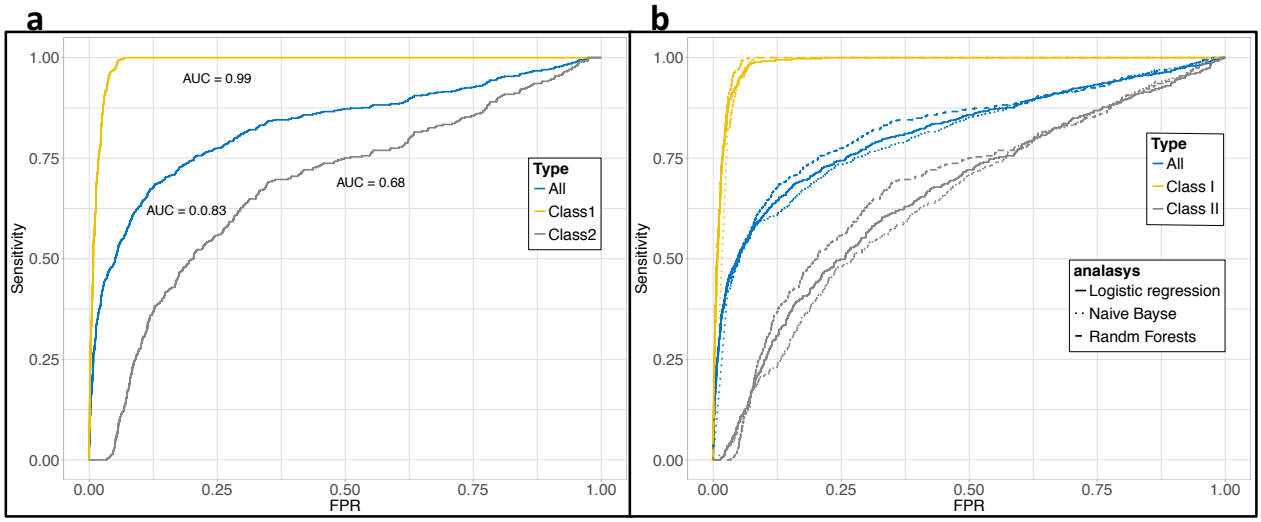

Figure 6: (a) *Random Forest ROC curve for the data set (blue) shows mediocre performance, with area under the curve (AUC) of 0.83. Predicting class I enhancers, recall rises sharply, while prediction of class II enhancers is close to random guess.* ***(b)*** *ROC curves for Random Forests, logistic regression and Naïve Bayes classifier are nearly overlapping, with AUC =0.99 in all three methods when predicting the class I homogenous enhancer set.*

This seperation by the model of our enhancer class into 2 subclasses can be understood through examining the Principal Component Analysis (PCA) of the data (**Fig.7a**), where it is clear Class II enhancers collocate with non-enhancers while class I enhancers are separated from both along both PCA primary axes. Examination of feature space statistics of the 3 groups shows Class II enhancers are indistinguishable from non-enhancers along our entire feature space – including TF binding, histone marking, conservation and DNase accessibility - while class I enhancers segregated from both on multiple features (**Fig.7b, Fig.8-9**). A distinct difference in the distribution of Class I enhancers is particularly notable for TF binding (**Fig. 8a**) though a few histone marks (**Fig. 8b**) - particularly H3K4me1 - also demonstrate a difference between Class I enhancers

and the other DNA segments. The separation in transcription factor binding profiles may indicate a possible reason and mechanism for the separation of the 2 classes – it may be Class II enhancers are regulated by TF's for whom data is not currently present. However, differences in distribution are also present for many other features (**Fig. 9**), suggesting a more profound difference between the sets. The large difference in DNase accessibility (**Fig 9b**) may indicate Class II enhancers are active in a small number of cells, with the enhancer tightly packed in the rest. In that case, Class II enhancers operate just as type I enhancers but the signal is too weak and cannot be detected in whole-embryo sequencing tests as those employed here. This supposition is further confirmed by the difference in expression patterns induced by the two classes, which will be described in detail later in the text. It is It is clear that there is no separation in any of the conservation scores between enhancers and non-enhancers (**Fig 9a**), explaining why this measure was found uninformative. Notably, in no feature do class II enhancers separated from non-enhancers. The separation of Class I and Class II enhancers in feature space demonstrates Random Forests can be readily used to separate heterogeneous enhancer sets.

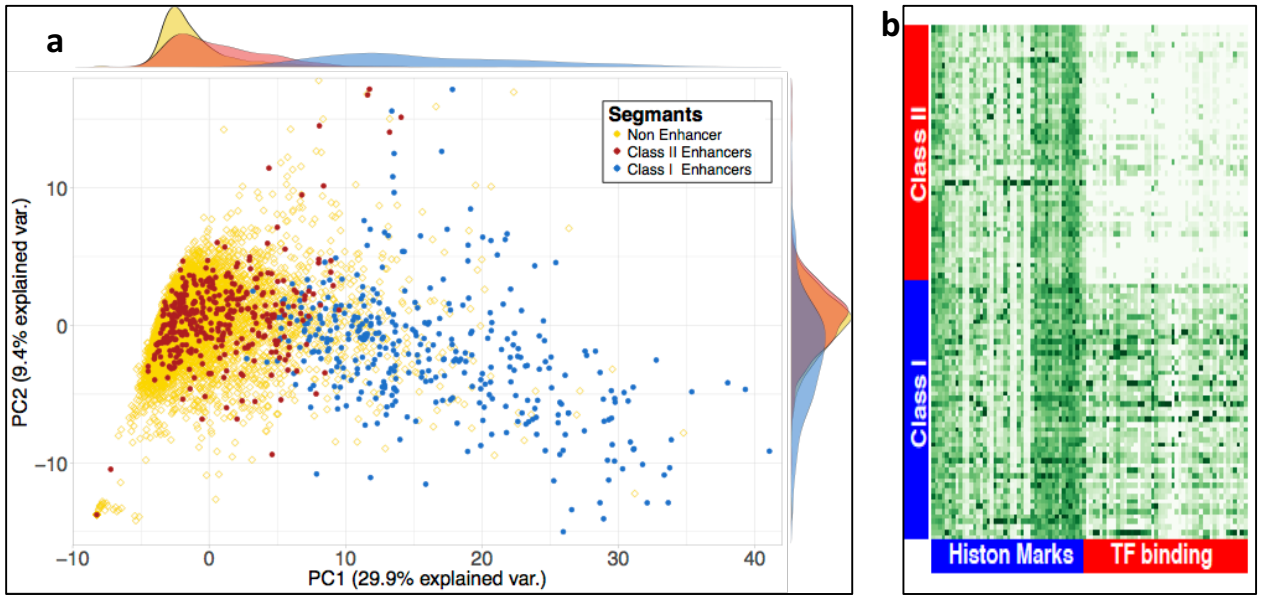

Figure 7: *The separation into enhancer classes can be explained by the co-localization of class II enhancers and non-enhancers in the PCA projection in* ***(a)****. The separation is mainly driven by the transcription factors as exemplified by the normalized ChiP strength across features of 200 randomly selected class I and class II enhancers* ***(b)***

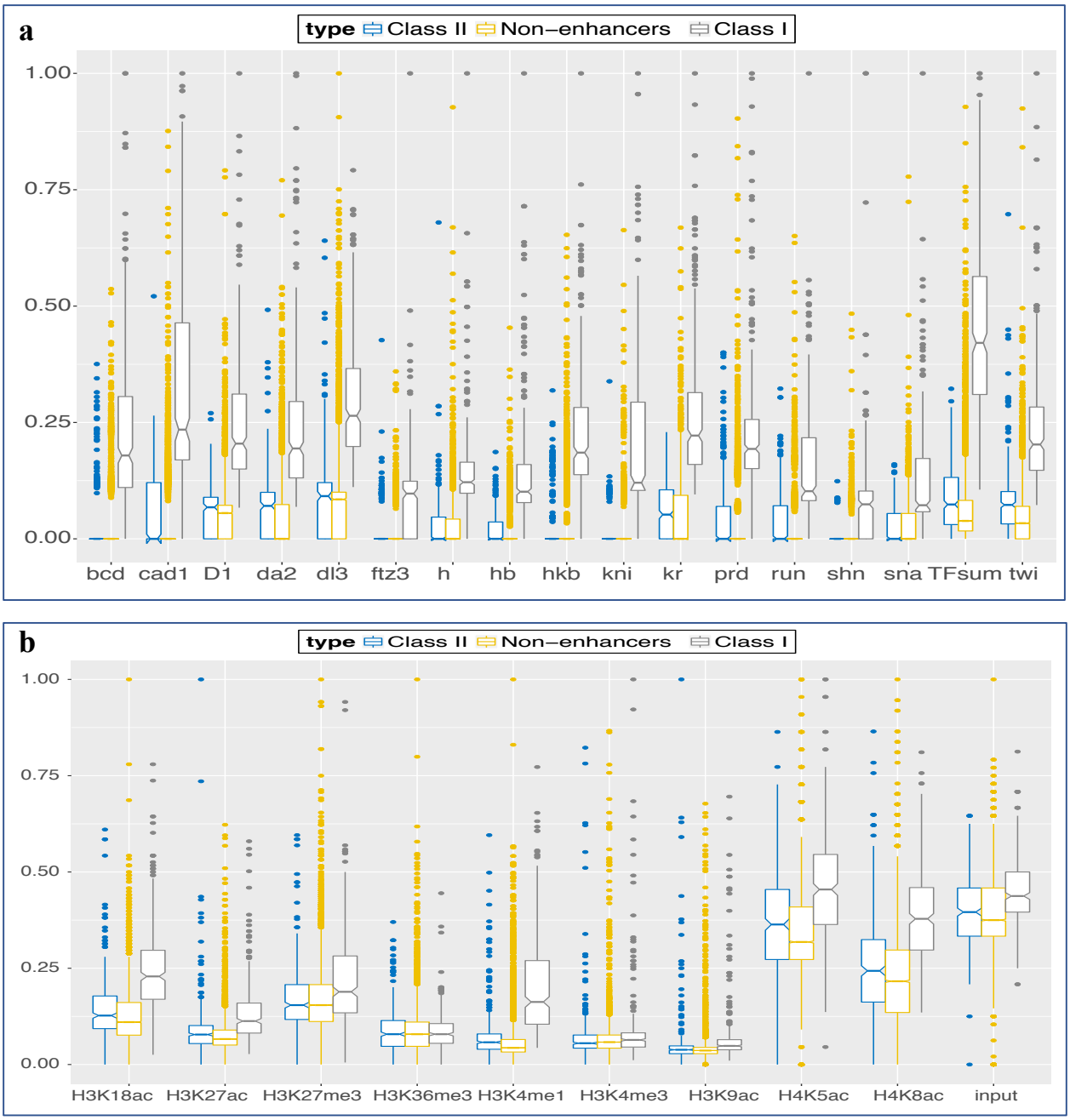

Figure 8: *box plots showing the distribution data by summarizing quantiles: 0.25-0.75 quantiles represented as a box, the notch indicates the median, whiskers extend an additional 1.5 interquartile range in each direction, and outlies shown as points. Shown here are the distribution of* ***(a)*** *selected Transcription factor or* ***(b)*** *selected histone modifications, for non-enhancers (yellow), class I enhancers (gray) and class II enhancers (yellow).*

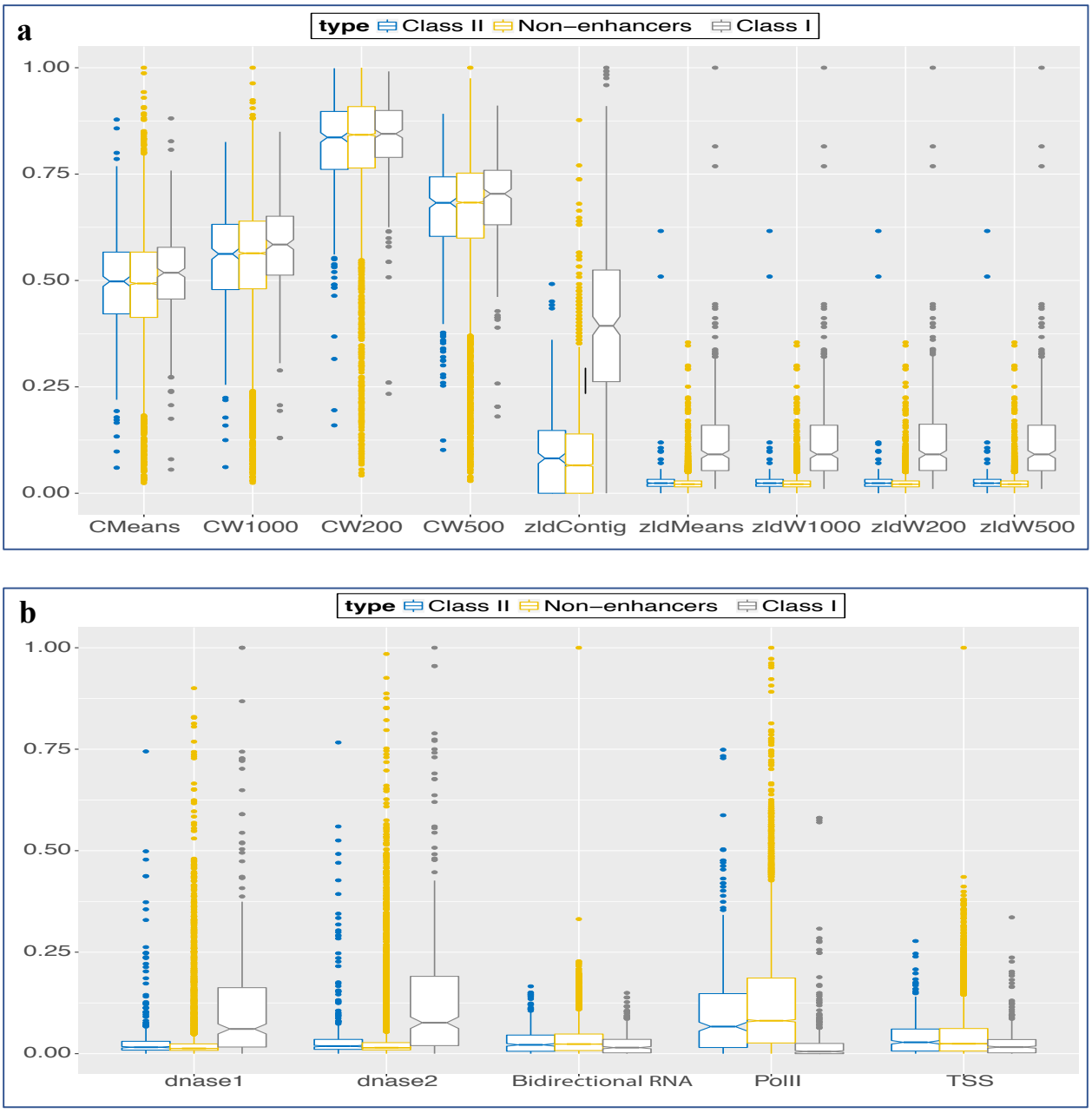

Figure 9: *box plots showing the distribution data by summarizing quantiles: 0.25-0.75 quantiles represented as a box, the notch indicates the median, whiskers extend an additional 1.5 interquartile range in each direction, and outlies shown as points. Shown here are the distribution of* ***(a)*** *various conservation and ZLD binding scores: average (CMeans, zldMeans), highest values along sliding windows (CW1000/200/500, zld1000/200/500) or longest contiguous starch (zldContig)* ***(b)*** *DNase accessibility, distance to bidirectional RNA, distance to PolII 2 or distance to transcription start site.*

Allowing for enhancer heterogeneity by excluding Class II enhancers from the sampled training set gives us unprecedented prediction accuracy. On a balanced held out test set, where half genomic segments were enhancers and half were determined to be not functional, more than 98% of Class I enhancers are discovered with better than 95% precision. This level of accuracy far surpasses any previous report in any metazoan system, and is unique to the Class I enhancer prediction. Using the same methods on the full heterogeneous set does not allow for accuracy and prediction above 85%, and with Class II enhancers predictive accuracy is close to that of random guess. This high predictive power in the model is mostly attributable to a small number of transcription factors (Kr, Med, Twi, Dl, D), as will discussed later in this work. It is possible Class II enhancers are difficult to predict as they are controlled by TF's not included in this work, and that additional transcription factors will allow for this accuracy to be extended to Class II enhancers as well.

In a true scan of the complete genome, we expect the accuracy to be lower. As one moves away from a balanced set, the frequency of false positives rate increases as a small fraction mislabeled non-enhancers can overwhelm the much smaller true positive set. To demonstrate the point, Random Forests were trained on a balanced set and tested on increasingly imbalanced test set, at various degrees of stringency (**Fig.10A**). It is interesting to note the sharp rise in false discovery rate in both model accuracy and non-enhancer fold increase. This can also be seen in the marginal (**Fig.10B-C**): unless the sample is very close to balanced, the rise in false discovery rate in the test set is extremely sharp. Conversely, in genomic scans where the non-enhancers are likely to be at least a hundred-fold more prevalent, a precision considerably better then 95% is needed to retain predictive power with better then 75% false discovery rate.

If all segments not in the balanced test set are considered in the test set, the test set contains 20 times more non-enhancers than enhancers. Under those conditions 90% of the enhancers were discovered with 60% precision. The prediction accuracy is likely considerably higher than this analysis implies, however, due to false negatives in the data published by Kvon el al. Reassessment of their gene expression image data for the top 100 genomic regions that our method predicted to be enhancers but which were reported as non-enhancers by Kvon et al. revealed that only 15 were true non-enhancers, 47 were clearly enhancers, and the remainder could not be conclusively classified due to inadequate or insufficient data (**Fig. 11C**).

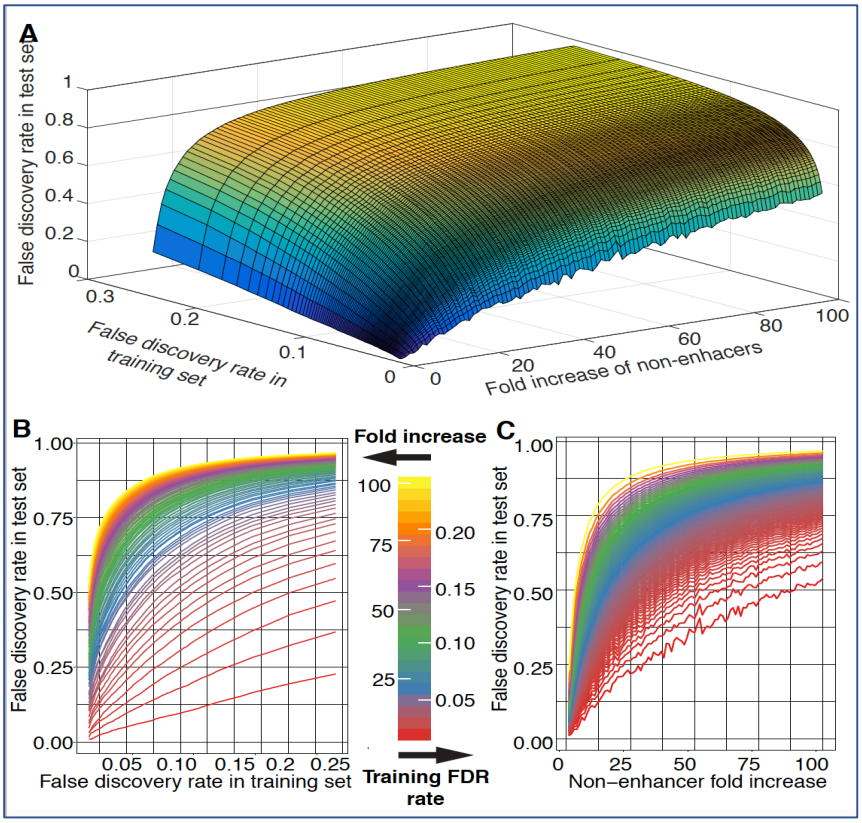

Figure 10: *False positive rate is a function of method accuracy and imbalance in the test data* ***(A)*** *3-dimensional surface plot, showing a sharp increase in the test's false positive rate in both axes. In genomic settings, where the imbalance cannot be controlled, a very high degree of accuracy is required.* ***(B-C)*** *Marginal of the 3D image above, demonstrating the sharp rise in test inaccuracy with regards to both false positive rate in the training set or dilution of enhancer class in the test set.*

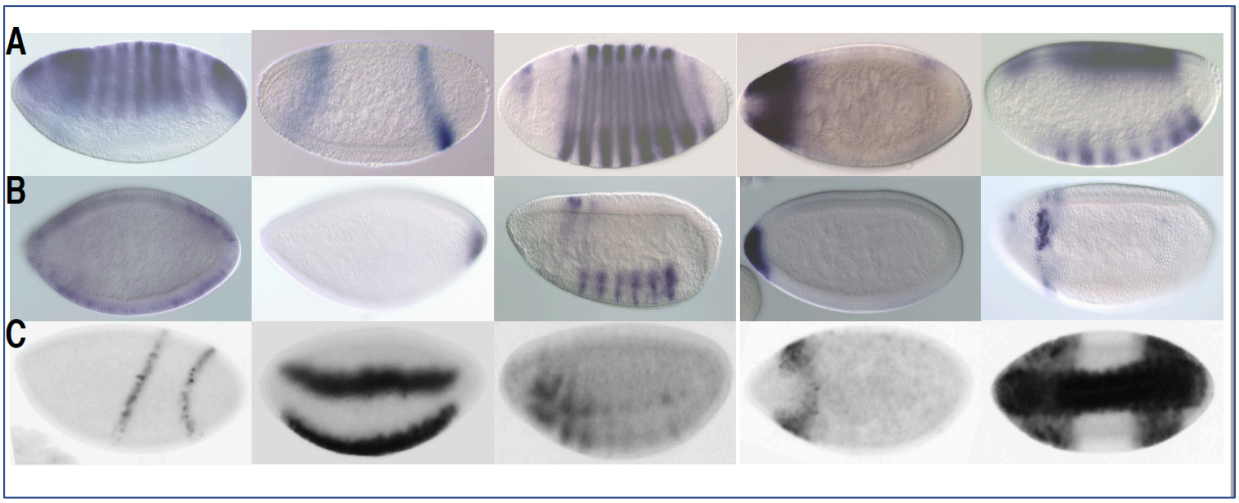

Figure 11: *Examples of expression patterns in* ***(A)*** *Class I enhancers* ***(B)*** *Class II enhancers* ***(C)*** *Embryos misclassified as non-enhancers in stages 4-6 in Kvon et al.*

Segmentation driving enhancers (SDE)

We next sought to understand if the separation of the enhancers in our feature space is related to their biology. To understand possible differences in function, we first examined possible differences in the expression of the reporter gene mRNA induced by the class I and Class II enhancers. A detailed reexamination of images for 76 randomly selected class I and 66 randomly selected class II enhancers was performed. We found Class II enhancers tend to be expressed in only a small percent of cells. 82% of class II enhancers had a tight expression pattern (expressed in <= 15% of cells) vs. 45% of class I. While separation is not complete, it is unlikely that these differences in expression are due to chance (P-value < 10^{-5}). The separation between the two classes in prediction scores are very sharp, demonstrating two distinct underlying distributions (**Fig.12a).** In addition, there is a difference in the time scale of activation. We find that class I enhancers are more likely to remain active throughout embryogenesis, while class II enhancers tend to work intermittently or for shorter periods (**Fig.12 b, c)**

Additionally, we looked at expression patterns of class I and class II enhancers, as annotated by Kvon et al. [56]. Analysis of the annotation terms show a significant (P value < 10^{-4}) enrichment for the expression in A-P stripes, posterior or gap gene like patterns (**table 3)**. As this implies class I terms may be A-P patterning while class II are D-V patterning, we wished to test the assumption by collecting all terms relating to one of these primary patterns either at stage 4-6 or a progenitor to a patterned organ later in development. However, we found no difference in axis patterning as a whole between the classes, as the distribution between them was remarkably even (**Table 3)**. GO-term analysis of the genes proximal to class I enhancers also showed a highly significant enrichment of terms related to segmentation (**Fig.13**), while those of class II enhancers showed much lower enrichment for any GO terms and no significant enrichment for any particular pathway (**Fig.13**). We therefore hypothesize that class I enhancers are likely to drive patterns of expression needed for establishing the segmented body plan (segmentation driving enhancers (SDE)), and term class I and class II enhancers SDE and non-SDE enhancers respectively. We note that while the differences are significant, there is not a clear separation in function as a minority of non-SDE enhancers direct patterns of expression resembling those of SDEs (**Fig.11A,3B**).

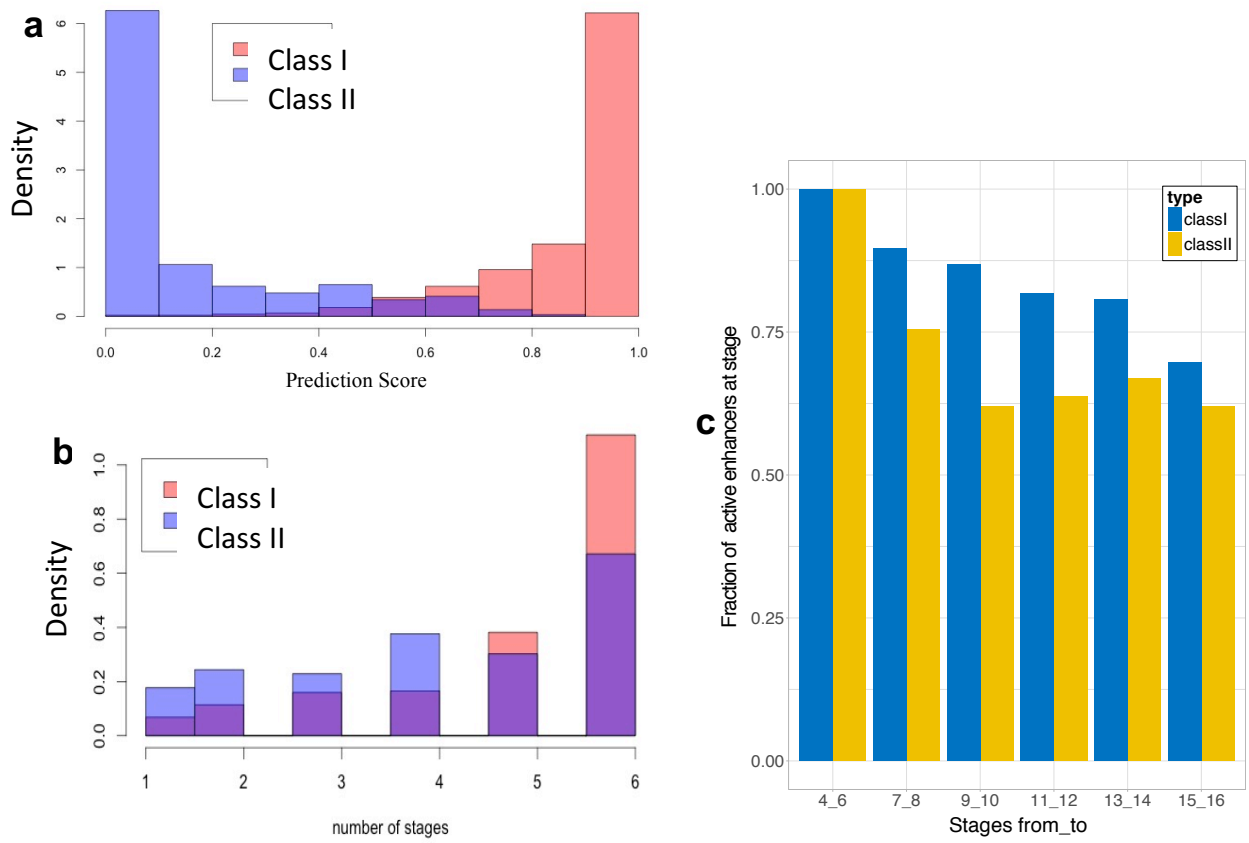

Figure 12: (a) *Histogram of the prediction score (fraction of trees classifying a DNA segment as enhancer) for Class I and Class II enhancers. The separation between the 2 classes distribution indicates fundamental differences in classification between the two classes.* ***(b)*** *Proportion of enhancers at each class vs. the number of annotated stages at which they were found to be active* ***(c)*** *Proportion of enhancers at each of the annotated stages. As for the purposes of this work "Enhancer" is defined as a DNA segment inducing gene expression at stages 4-6, all enhancers of both classes are active at that stage. Class I enhancers are more likely to remain active in later stages.*

Table 3: Annotated expression terms in stages 4-6				
Annotated expression	**Class I**	**Class II**	**all**	**P-value**
posterior	32	4	36	6.80E-06
gap	30	5	35	4.98E-05
AP_stripes	46	14	60	6.28E-05
ventral_ectoderm_AISN	33	7	40	7.72E-05
head_mesoderm_AISN	16	36	52	0.008418044
posterior_endoderm_AISN	25	47	72	0.013328329
dorsal_ectoderm_AISN_broad	17	6	23	0.037056219
trunk_mesoderm_AISN_broad	12	25	37	0.048519738
anterior	21	10	31	0.072486085
trunk_mesoderm_AISN_subset	7	16	23	0.095292838
yolk	10	3	13	0.096092329
anterior_endoderm_AISN	51	70	121	0.101763505
middle	11	4	15	0.12133525
segment_polarity	3	0	3	0.248213079
AP_stripe	14	8	22	0.286422023
amnioserosa_AISN_subset	2	6	8	0.288844366
dorsal_ectoderm_AISN_subset	21	29	50	0.322198806
pair_rule	11	6	17	0.331975467
procephalic_ectoderm_AISN	90	77	167	0.353102706
mesectoderm_anlage	4	1	5	0.37109337
apically_cleared	2	0	2	0.479500122
AP_semistripes_ventral	4	7	11	0.546493595
ubiquitous	22	27	49	0.567709166
hindgut_AISN	18	14	32	0.595883091
amnioserosa_AISN	19	15	34	0.606905427

table 3: Annotated expression terms in stages 4-6 (continued)				
Annotated expression	**Class I**	**Class II**	**all**	**P-value**
AP_semistripes_dorsal	6	4	10	0.751829634
brain_anlage_AISN	1	0	1	1
mesectoderm_AISN	1	0	1	1
mesoderm_AISN	15	15	30	1
AP_semistripe_dorsal	**3**	**3**	**6**	**1**
All patterned expression	**280**	**342**	**622**	**0.014450022**
A-P terms	**301**	**240**	**541**	**0.009891439**
D-V terms	**244**	**244**	**488**	**1**

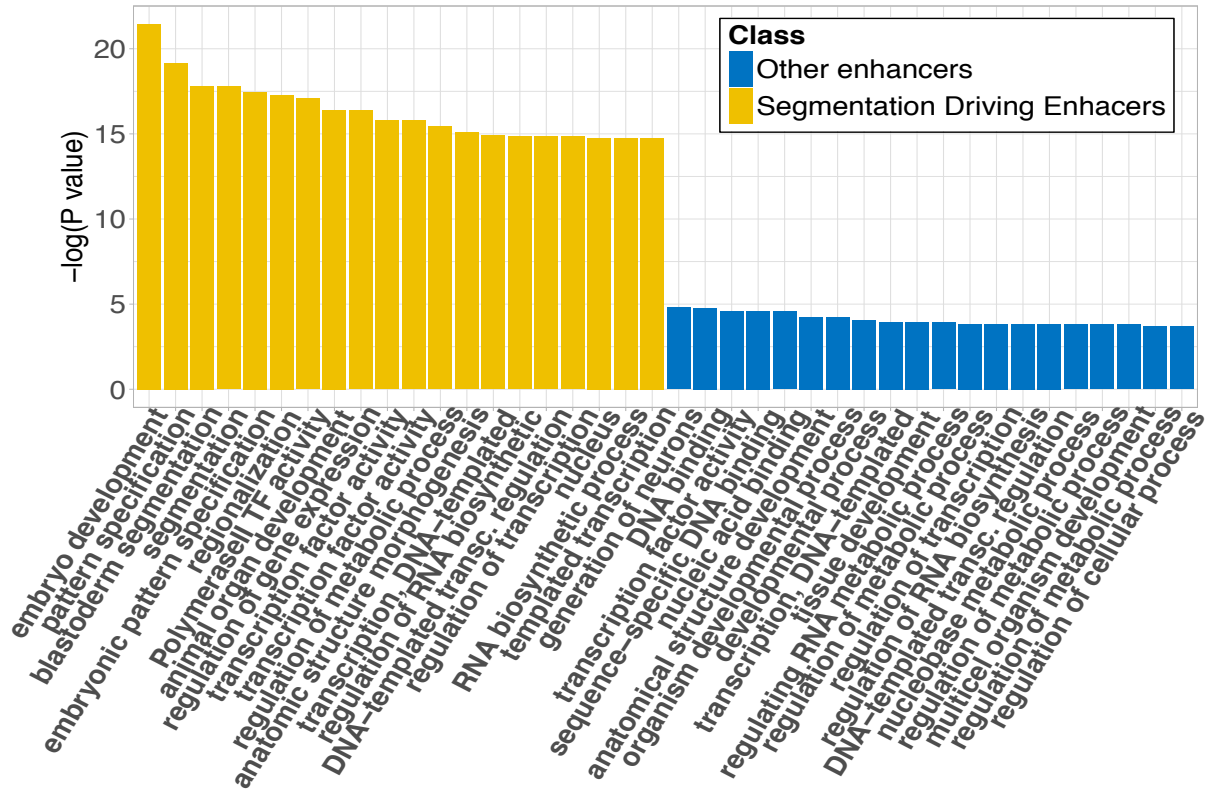

Figure 13: *Enrichment of the GO-terms proximal to enhancers as compared to Go-terms proximal to the entire DNA segment training set. Negative log of the P-value is shown for the 20 most enriched GO-terms foe class I and class I enhancers. As class I enhancers are highly enriched in terms related to segmentation we rename these* ***segmentation driving enhancers*** *or* ***SDE****.*

Feature importance is dominated by transcription factors

The random forest importance measures “mean decrease accuracy” and “mean decrease Gini” [38] varied widely from sample to sample but in all cases a small set of the transcription factors were found near the top of the importance ranking list. This can be seen by the spread of the bootstrap confidence interval of the two importance variables calculated in 50,000 trees (**Fig.15**). The sum of transcription factor binding along with a small number of transcription factors (Kr, Med, Twi, Dl, D) were most important by both measures, and were also the most often used by the model **(Fig.14**). Other transcription factors such as Bcd and Ftz [37], were found to be uninformative despite their importance in embryo segmentation. This can be at least partially explained by low coverage in the ChIP-chip data. Here, we define coverage as the fraction of DNA segments in the data set which have a non-zero binding value anywhere along its length. Low coverage may indicate a TF has few binding sites and low non-specific binding to DNA; It may also indicate low quality in the test. Both explanations may influence a TF importance to the prediction, one due to smaller number of nodes affected by feature and the other by dilution of the signal, and both are artifacts which hide the true importance of a feature. Ideally, coverage and feature importance are independent, however we find there is a clear correlation (r = 0.7) between coverage and importance measure mean decrease accuracy. The correlation vanishes (r = -0.1) when very low coverage data such as Bcd and Ftz are excluded. (**Fig.16 a**). Similar correlation is found for mean decrease Gini importance ad coverage.

The only histone mark to have an importance above random noise was H3k4 mono-methylation (H3K4me1), a histone mark previously reported as an enhancer indicator [68]. All 40-other histone and histone modifications, including the H3K27 acetylation (H3K27ac) that has been widely regarded as a key indicator of enhancer regions [12,69] were found uninformative by the model in the presence of the transcription factor data, and had importance measure comparable to the RNA-seq input data, which are sequencing bias-control files and are not expected to contain any enhancer-relevant data. Features with importance measures comparable to the various input files may be assumed to be not important.

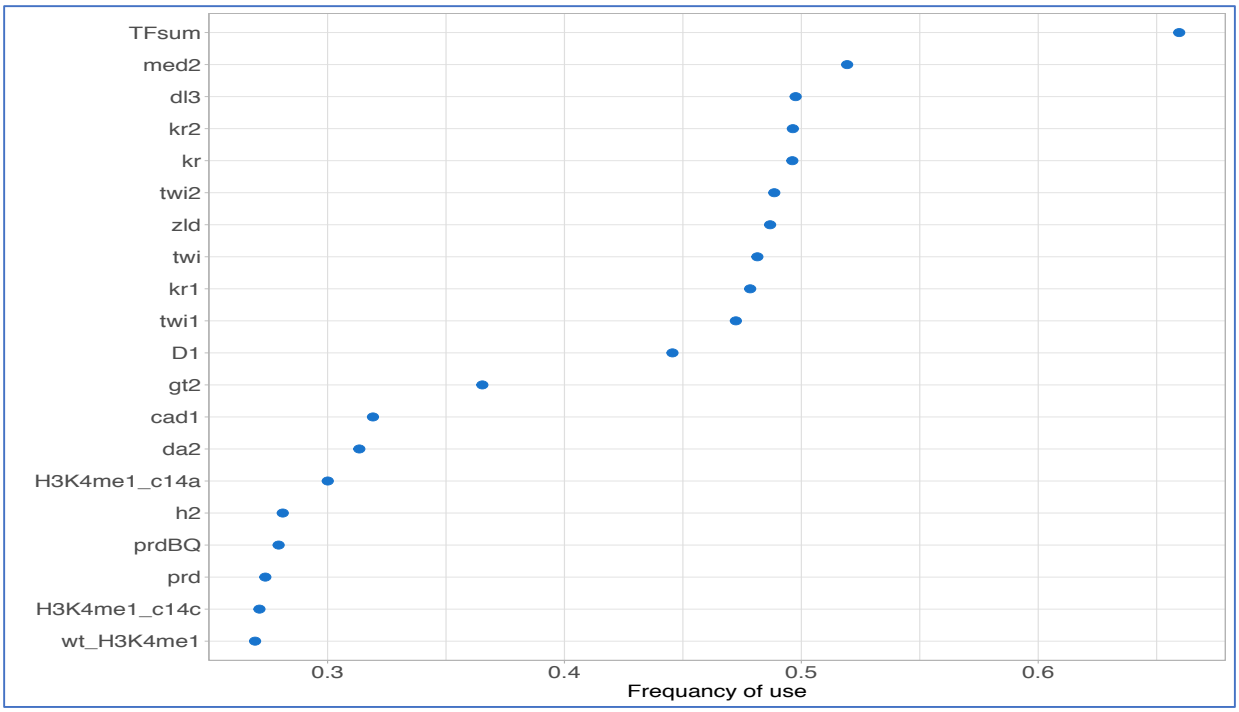

Figure 14: *How frequently each of the top 25 features was used by Random Forest in predicting SDE.*

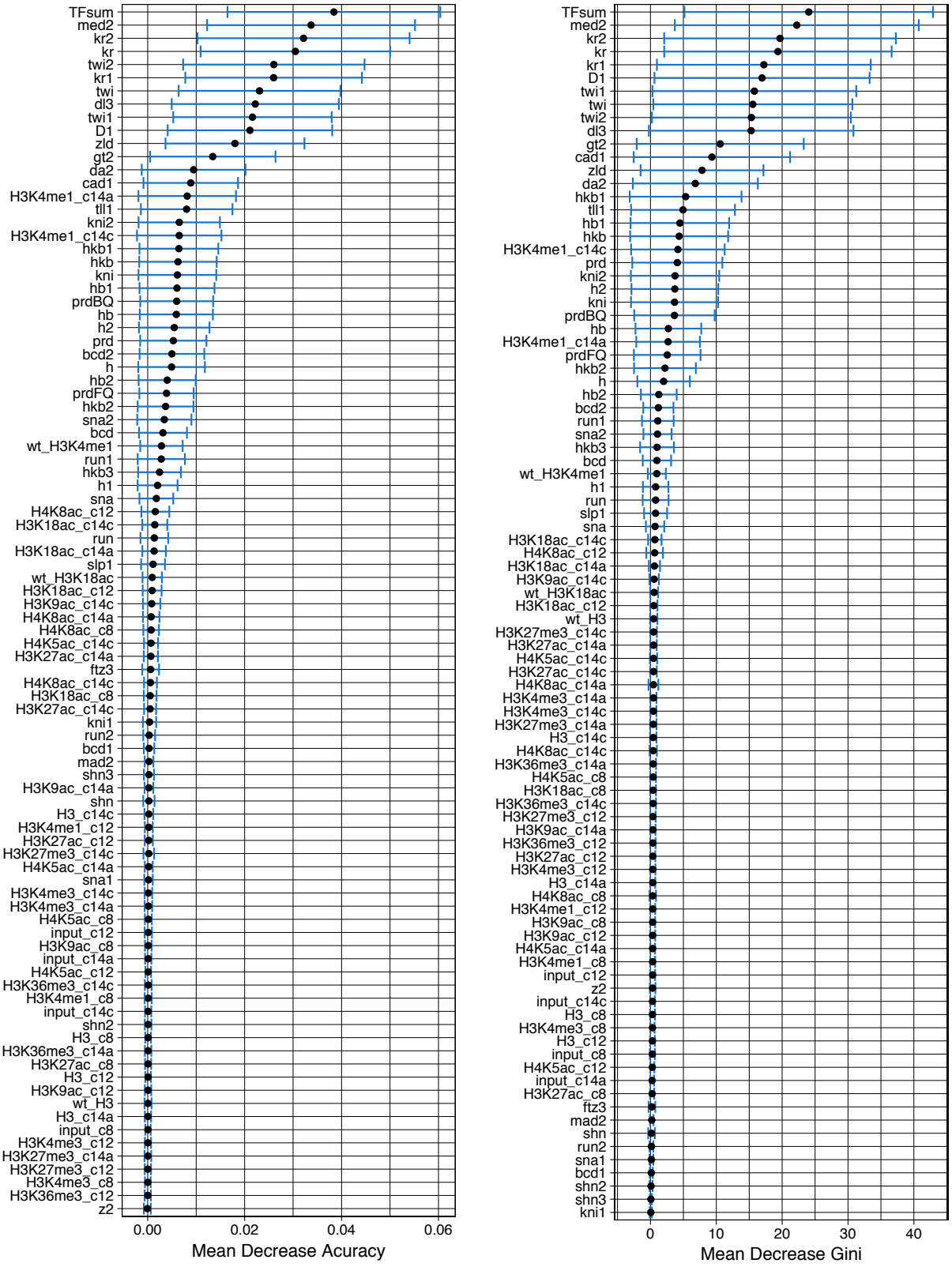

Figure 15: *Mean decrease Gini and mean decrees accuracy measure averaged over 50,000 forest, trained on SDE and non-enhancers.*

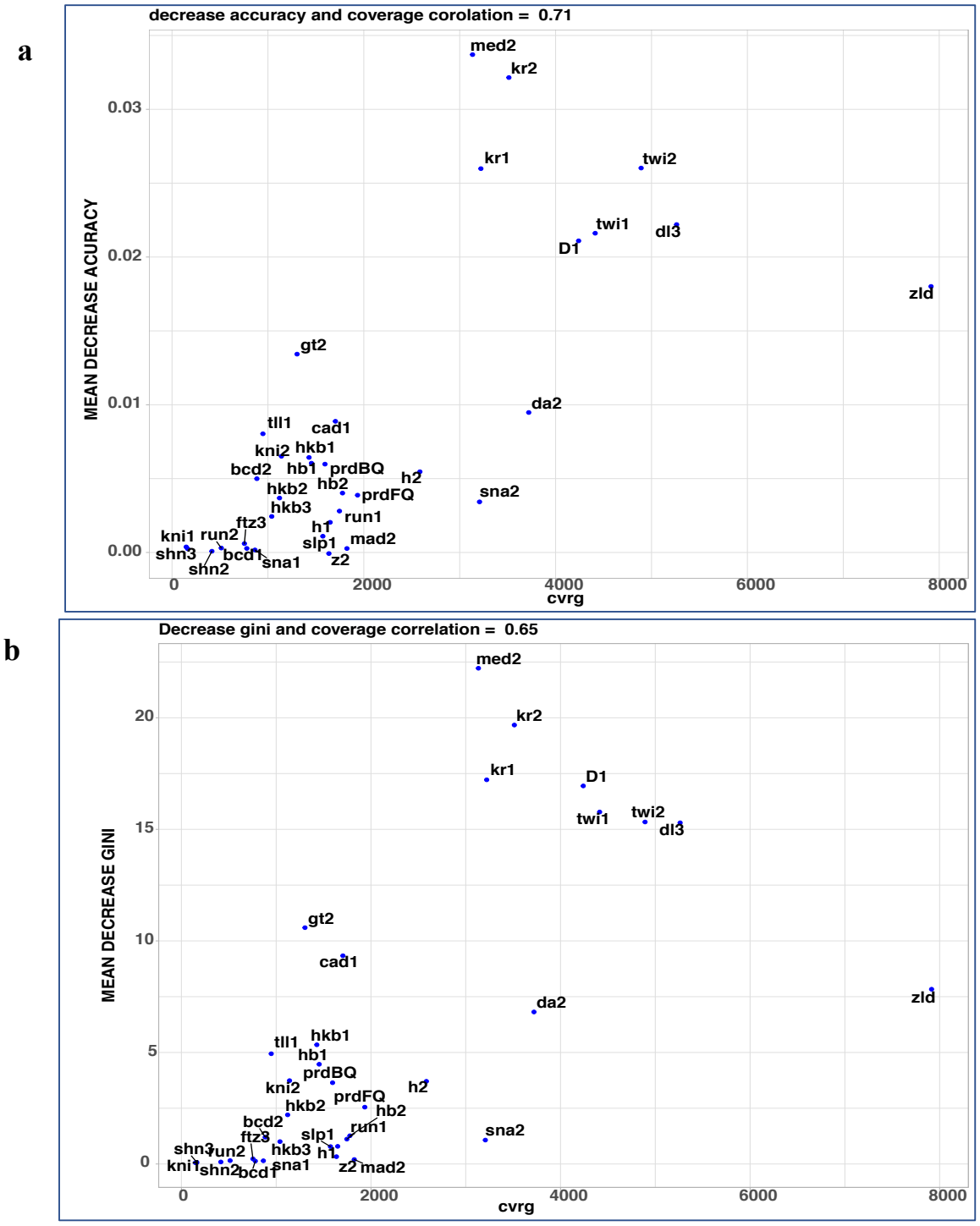

Figure 16: *Correlation between feature importance of transcription factors as measured by mean decrease accuracy* ***(a)*** *or mean decrease Gini* ***(b)*** *and the number of DNA segments in the training set which contain peaks above the 25% FDR of these transcription factors. It seems there are two cluster of transcription factors: low coverage low importance and high coverage high importance. Though there appears no correlation inside the clusters, there is an overall correlation of r ~ 0.7 between importance and coverage.*

The Random Forest power to distinguish non-SDE enhancers is limited, but not non-existent, as demonstrated by the ROC curves in (**Fig.6)**. The error rate for a single forest is ~35%, vs. 3% when seeking SDE enhancers, but while this error rate is high it does indicate some capacity of Random Forests to correctly classify these enhancers. Interestingly, it appears an entirely different set of features are used by the random forest to classify non-SDE enhancers (**Fig.17).** While several transcription factors are of high importance in SDE classification, with a great deal of redundancy and variation in importance ranking amongst them, Twist (Twi) is considerably and consistently the most prominent feature in classifying non-SDE enhancers, with most of the variation issuing from the redundancy of the two biological replicas and their summation. Given its high prominence, it is interesting to note that the overall error rate of a single forest is unaffected by the exclusion of Twist, indicating additional redundancies. Twist is a prominent dorsal-ventral transcription factor essential for the embryonic D-V differentiation, which originally led us to hypothesize class I and class II enhancers may be A-P vs. D-V pattern forming, however this was later disproved by comparison of the expression patterns (**Table 2)**. Furthermore, Twist and other dorsal-ventral transcription factors such as Dorsal are highly ranked in in SDE classification importance feature list as well (**Fig.15**). Twist is also one of the most prevalent transcription factors in the drosophila embryo, with numerous binding sites, as can demonstrated by its high coverage (**Fig.16**), raising the possibility it may be a serving as a surrogate for DNase accessibility, which was trimmed from the feature set at an early stage. However, the ubiquitously expressed Zelda has the greatest coverage in our training set (**Fig.16**), and is yet of only modest importance. Furthermore, when all features are reintroduced, twist continues to top the importance rank lists, with DNase accessibility biological replicates appearing in places 32 and 56 of the mean decrease Gini and mean decrease accuracy, respectively (**Fig.18**).

The ascendancy of transcription factors as predictors is also reduced in the non-SDE set, with histone modification becoming more prominent (**Fig.17**), though surprisingly H3K27ac still retain a relatively low rank in importance space, with other histone markers not normally associated with enhancer discovery, such as such as H3K18ac having higher importance ranking. With the reintroduction of the full feature set, many other features (**Fig.18**) are now found to be more important, including distance to polymerase peaks (polII, BedPolII), distance to bidirectional mRNA transcripts (biDistance), distance to transcription start site (TS) and others. There is a much greater heterogeneity in the features used, while the reduction in importance value along the ranked features is more gradual, so that even features at the bottom of the rank list are occasionally used by the forest. Despite that, there is no improvement in predictive power with the reintroduction of the previously removed feature sets, and the error rate remains at ~35%.

Feature importance when attempting to classify all enhancers as a heterogeneous set resembles that of non-SDE classification, with twist consistently the most used feature, and several histone modifications in the top 25 ranked (**Fig.19**). In contrast, when using random forest to separate SDE and non-SDE enhancers the parameter list as well as the error rates are the same as those used to separate SDE from non-enhancers (**Fig.20**).

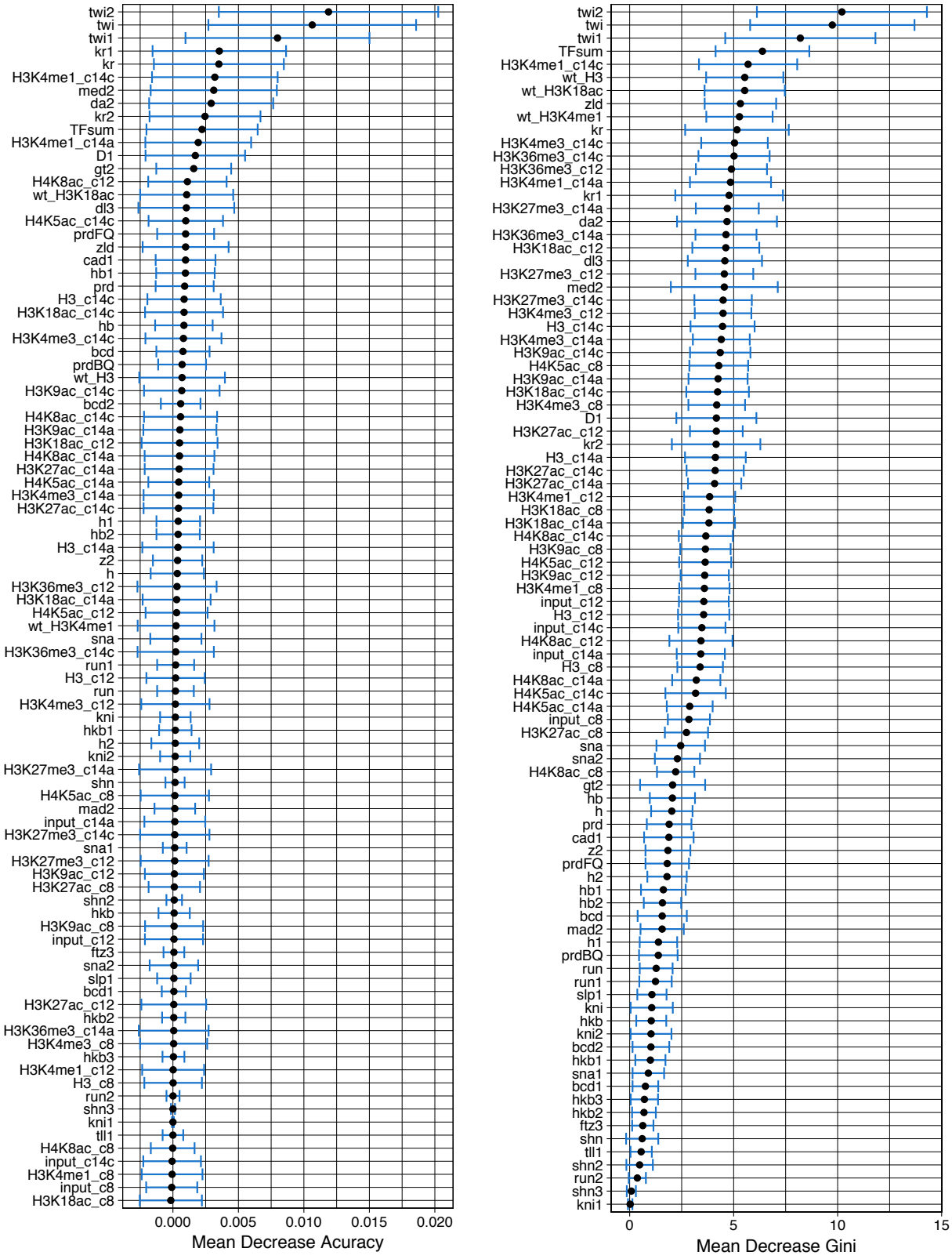

Figure 17: *Mean decrease Gini and mean decrees accuracy measure averaged over 50,000 forest, trained on non-SDE Vs. non-enhancers.*

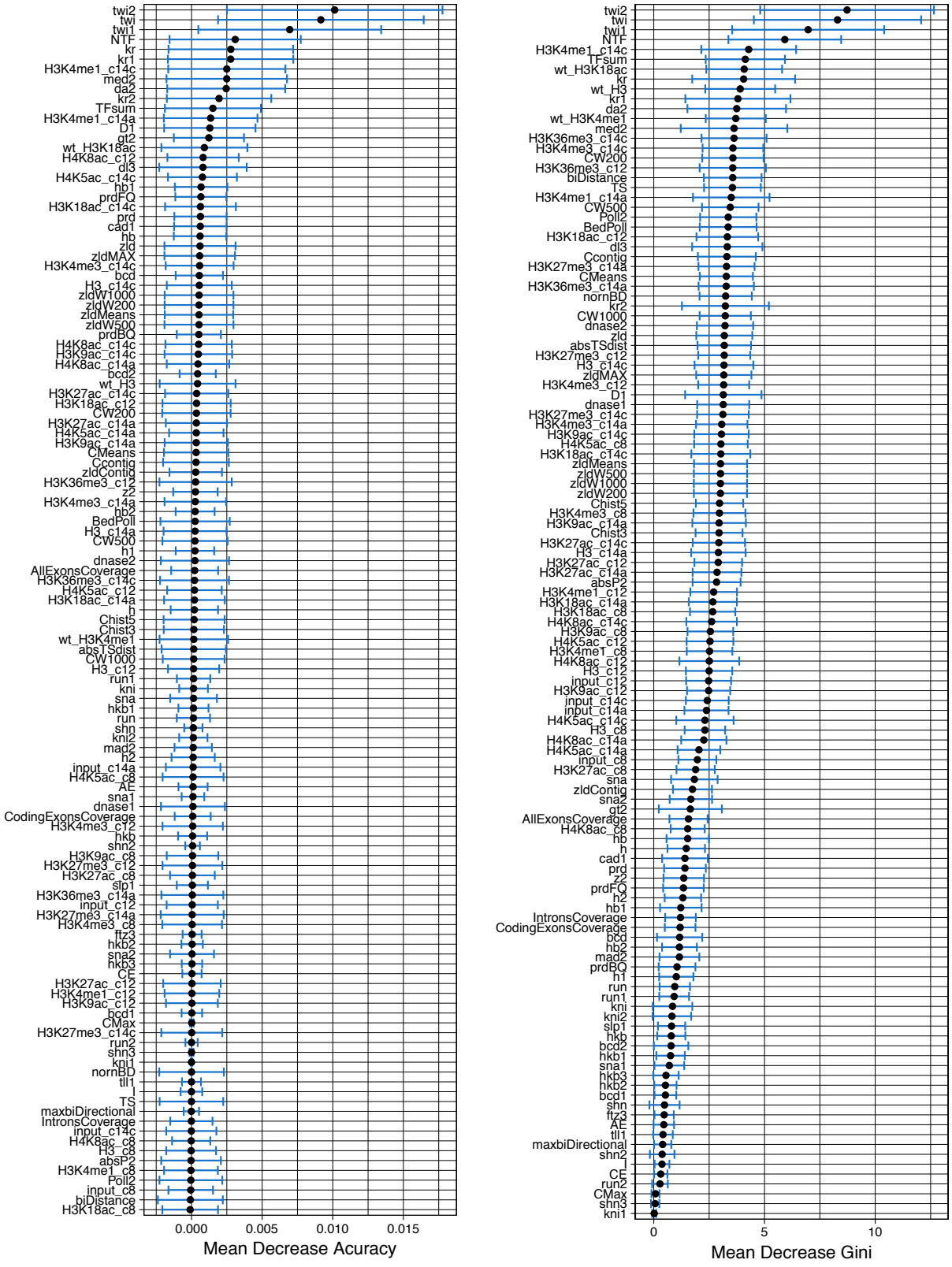

Figure 18: *Mean decrease Gini and mean decrees accuracy measure averaged over 50,000 forest, trained on non-SDE Vs. non-enhancers using all 124 features.*

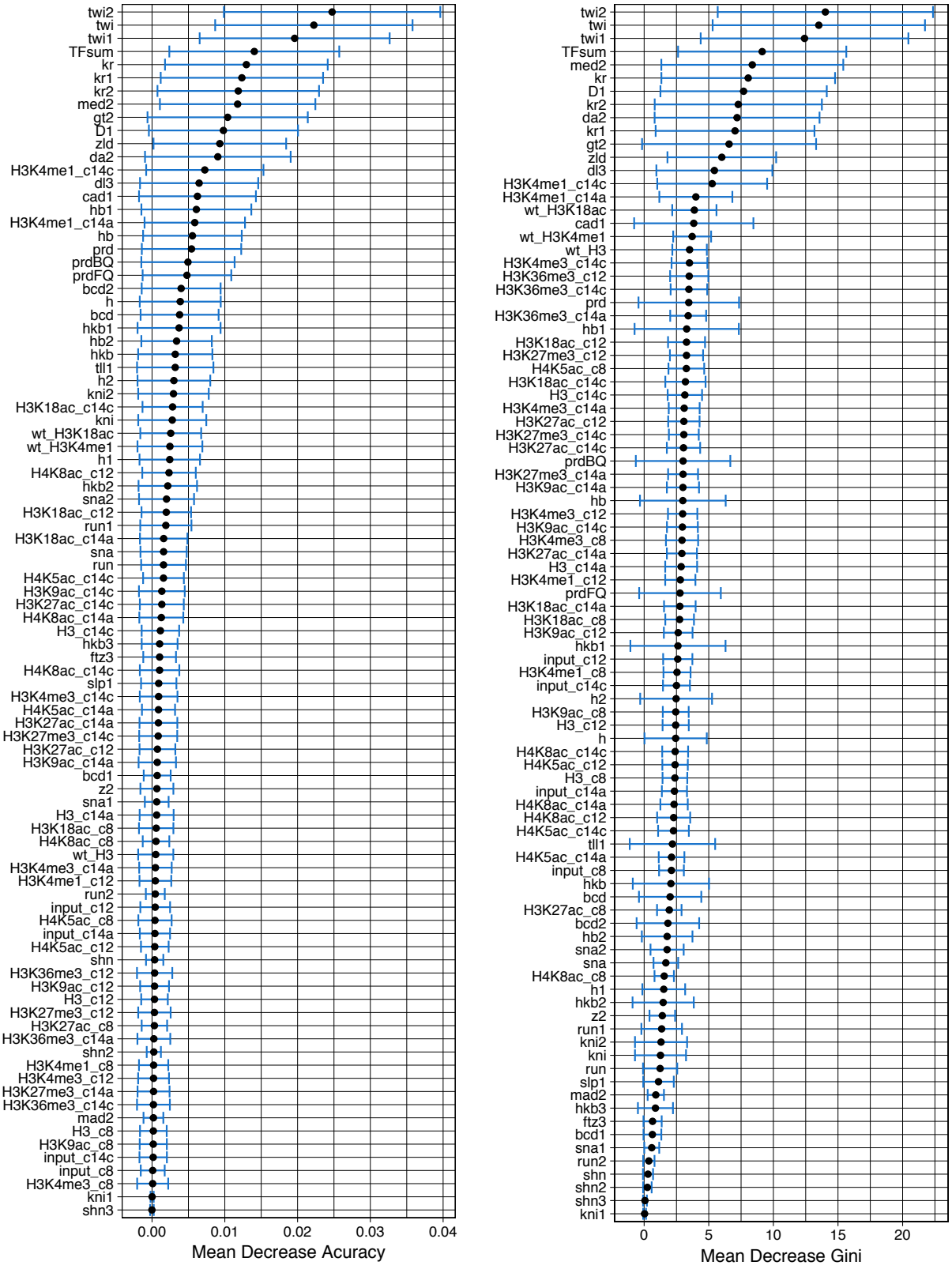

Figure 19: *Mean decrease Gini and mean decrees accuracy measure averaged over 50,000 forest, trained on all enhancers (SDE and non-SDE), Vs. non-enhancers using all 124 features.*

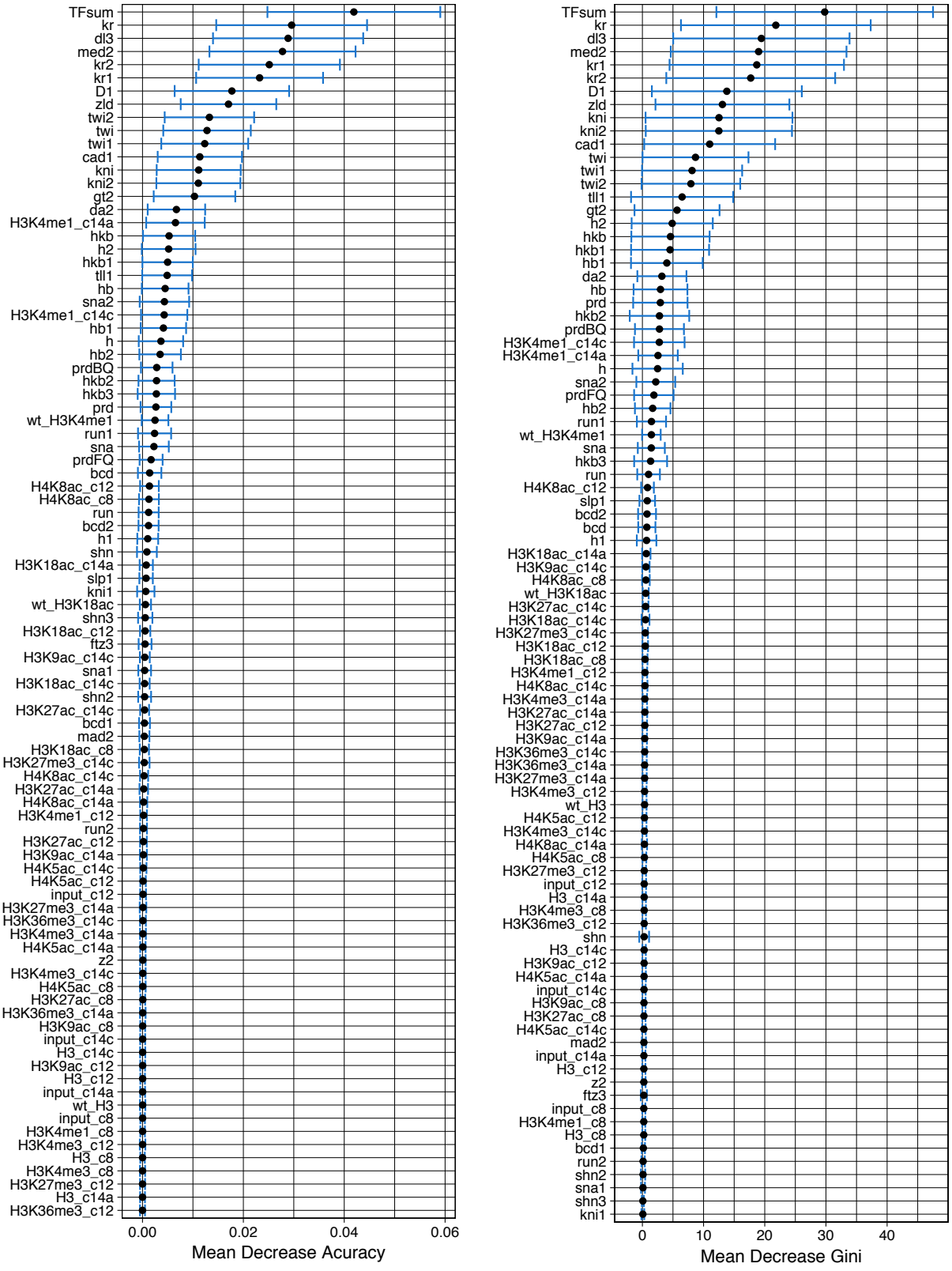

Figure 20: *Mean decrease Gini and mean decrees accuracy measure averaged over 50,000 forest, trained on SDE Vs. non-SDE.*

Attempts to determine which features were used by the Naïve Bayes classifier and logistic regression were unsuccessful, as the measures failed to converge, and were greatly affected by the order in which the parameters were presented to the model. As both these methods are considerably slower than Random Forest, only a few thousand runs could be made in each within a reasonable time frame, and the results were inconsistent. In an attempt to get a more stable estimation, 500 stepwise logistic regression were performed, and the number of times each feature appeared was calculated. However, almost without exception all features were used between 200-300 times (**table 4**). It should be noted that the features used most often are some of the transcription factors topping the Random forest importance list: Zld, TFsum, Prd and Twi. The exception in the list is *harry* (h), which was only used 5 times in the final step-wise regression. While harry is not an exceptionally critical transcription factor, its relevant scarcity has probably more to do with its location in the data matrix as the last feature introduced, coupled with the high correlative value between the transcription factors. In stepwise regressions were the feature order was randomly permuted, h was not found to have significantly less importance than other features.

Table 4: Number of times features were present in final model of 500 stepwise logistic regressions

features	#Used	features	#Used	features	#Used
H3_c12	217	**H3K4me3_c14a**	280	**D1**	301
H3_c14a	235	**H3K4me3_c14c**	246	**da2**	291
H3_c14c	293	**H3K4me3_c8**	209	**dl3**	252
H3_c8	249	**H3K9ac_c12**	246	**ftz3**	260
H3K18ac_c12	237	**H3K9ac_c14a**	262	**gt2**	272
H3K18ac_c14a	288	**H3K9ac_c14c**	248	**mad2**	232
H3K18ac_c14c	264	**H3K9ac_c8**	210	**med2**	325
H3K18ac_c8	264	**H4K5ac_c12**	285	**slp1**	298
H3K27ac_c12	227	**H4K5ac_c14a**	256	**tll1**	270
H3K27ac_c14a	239	**H4K5ac_c14c**	265	**z2**	259
H3K27ac_c14c	239	**H4K5ac_c8**	224	**zld**	386
H3K27ac_c8	265	**H4K8ac_c12**	239	**TFsum**	386
H3K27me3_c12	224	**H4K8ac_c14a**	260	**bcd**	331
H3K27me3_c14a	225	**H4K8ac_c14c**	301	**twi**	386
H3K27me3_c14c	255	**H4K8ac_c8**	244	**sna**	279
H3K36me3_c12	251	**input_c12**	240	**shn**	258
H3K36me3_c14a	251	**input_c14a**	238	**run**	265
H3K36me3_c14c	245	**input_c14c**	229	**kr**	281
H3K4me1_c12	269	**input_c8**	230	**kni**	241
H3K4me1_c14a	276	**wt_H3**	248	**hkb**	325
H3K4me1_c14c	248	**wt_H3K18ac**	214	**prd**	364
H3K4me1_c8	252	**wt_H3K4me1**	251	**hb**	306
H3K4me3_c12	227	**cad1**	254	**h**	5

Local feature-importance measures and clustering

Random Forests local importance provides a detailed determination of the importance of each feature in classifying each segment, allowing a more direct understanding on the Random Forests decision making process. Random Forest local importance measures were calculated for forests attempting to classify SDE and non-enhancers, non-SDE and non-enhancers, and SDE and non-SDE enhancers (**Fig.21-22).** It is clear the same small set of features are used to distinguish SDE and non-enhancers (**Fig.21A**) as are used to distinguish SDE from non-SDE enhancers (**Fig.21B**), while the attempted separation of non-SDE and non-enhancers (**Fig.21C**) shows no variable which can consistently used in separation while many more parameters are used. The increase in used featues and the blurring of decision-criteria is also seen when non-SDE are presented to random forest as enhancers (**Fig.22A**) rather then non-enhancers (**Fig.22B**). This is result is in acordance with the results of the feature importance presented in the last section.

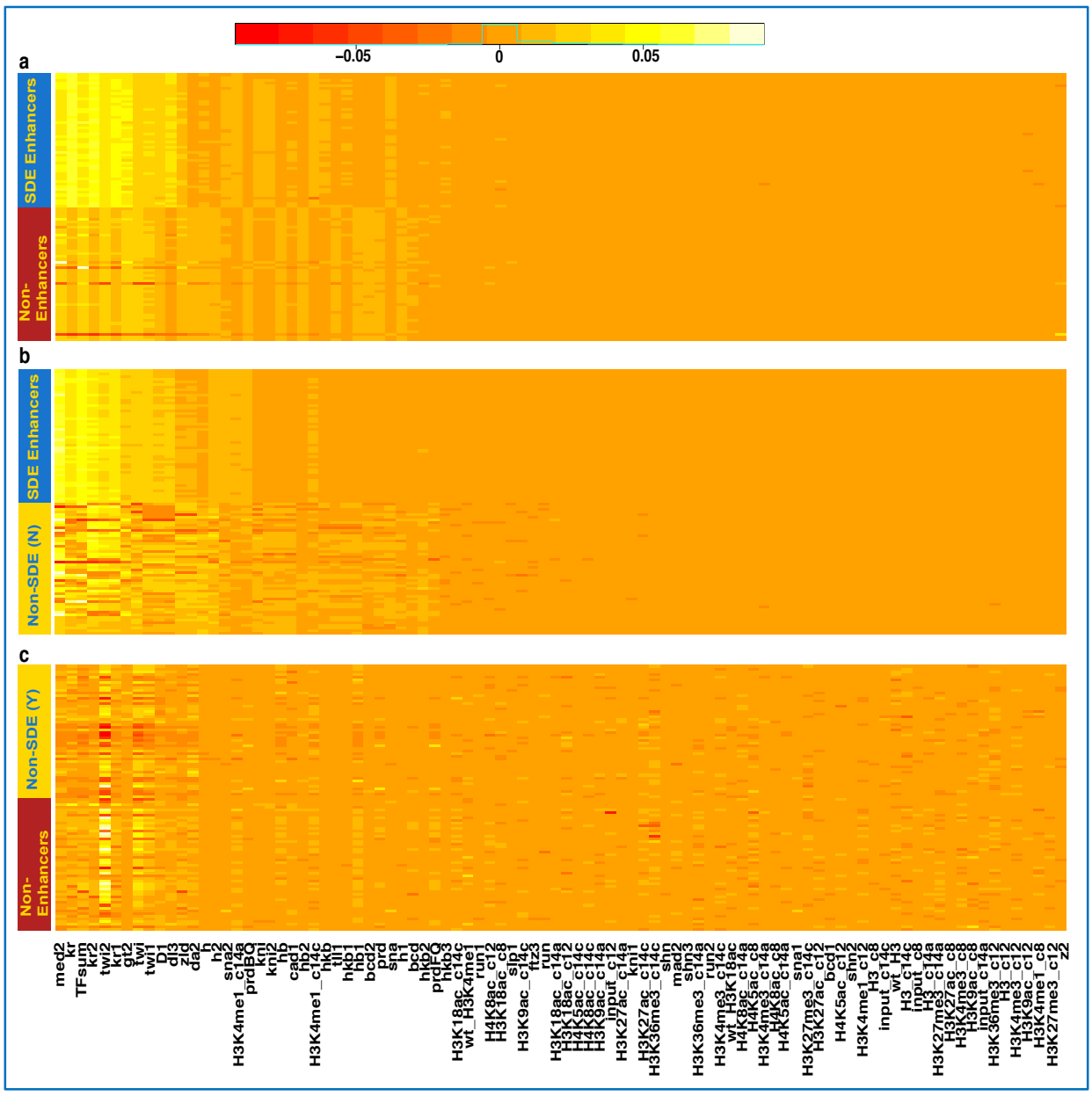

Figure 21: Local importance measurements *of randomly selected segments indicting how important each feature was in the segment classification, when forest was trained on* ***(a)*** *SDE vs. non-enhancers* ***(b)*** *SDE vs. non-SDE enhancers or* ***(c)*** *Non-SDE vs. non-enhancers.*

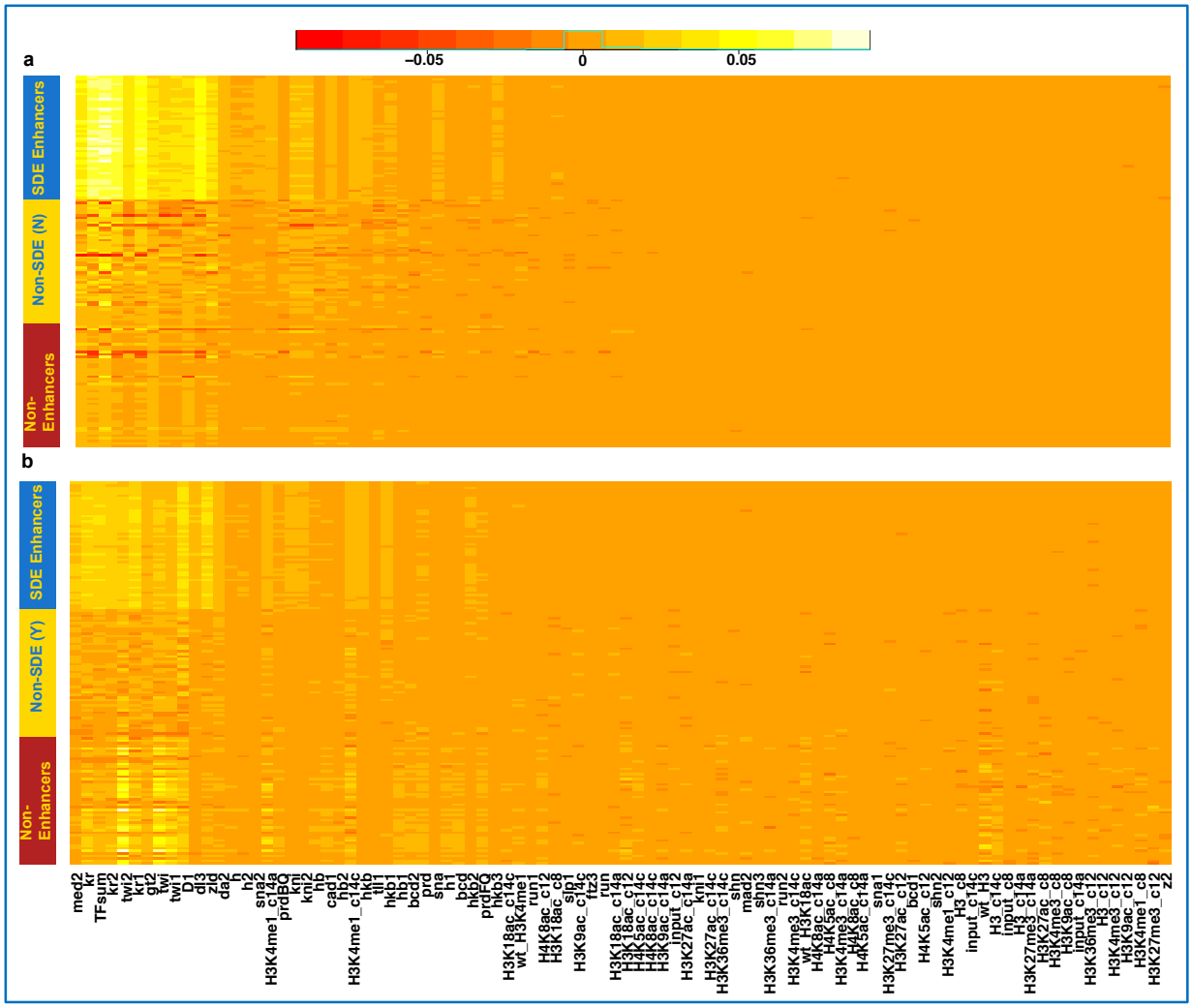

Figure 22: Local importance measurements *of randomly selected segments indicting how important each feature was in the segment classification, with non-SDW enhancers included in the analysis with a label* ***(a)*** *non-enhancers* ***(b)*** *Enhancers. The difference in the Random forest use of features is marked.*

Spectral clustering is a technique that relies on the eigenvector of the similarity or afinity matrix projection of the data, usually followed by k-nearest neighbors or k-means clustering [70]. It is an efficiant way of dimension reduction, and the number of clusters in the data can often be infered by the eigenvalues. Applying kknn spectral clustering [71] to our data fails to separate enhancers, and the eigenvalues of the affinity matrix indicates a single cluster (**Fig.23A**). Applying spectral clustering to the local importance matrix yields a good separation of the data (**Fig.23B**), with a sharp jump after the second eigenvalue (**Fig.24)** not indicating further fine structure in the data. The same results were obtained by using Kern spectral clustering (**Fig.25A).** K-means clustering of the data directly with k = 2 clustered all genomic segments together, with only 17 in the second cluster. The same trend continues as the number of clusters (K) increases. Applying K-clustering to the local importance does show good separation of the data, showing dimension reduction is not necessary for correct classification of the data (**Fig.25B**)

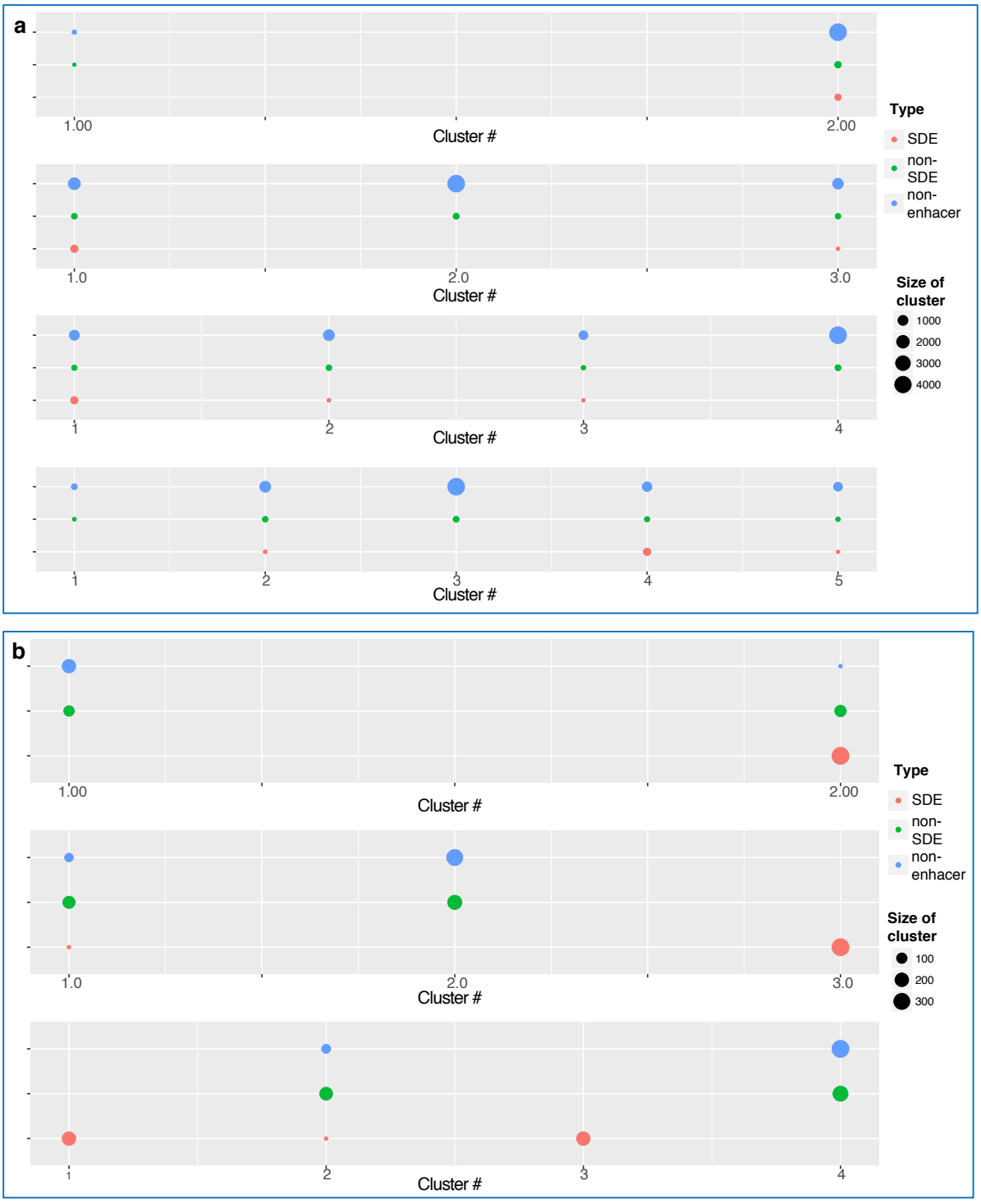

Figure 23: *Number of DNA segments of each type present in each, cluster size indicated by size of the dot, type by color and hight.* ***(a):*** *KKNN spectral clustering of the training data with 2,3,4 and 5 clusters. All fail to correctly separate enhancers of any kind from non-enhancers.* ***(b)*** *KKNN spectral clustering of local importance requiring 2,3, and 4 cluster correctly separate SDE enhancers from non-enhancers. Additional clusters splits SDE enhancers rather than separate non-SDE enhancers.*

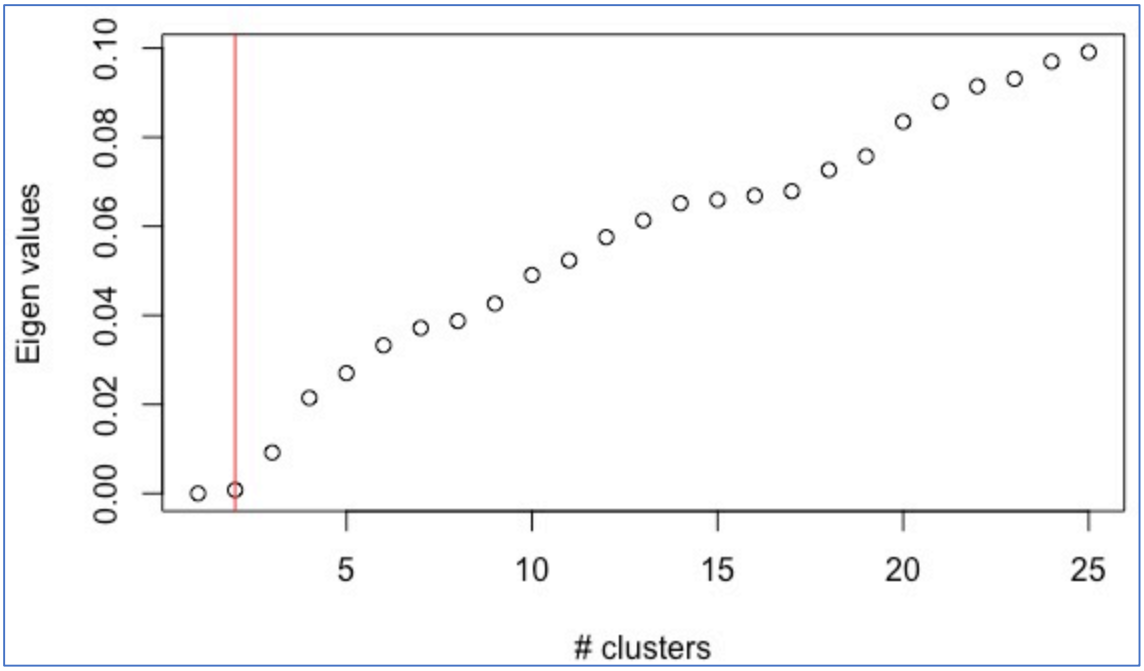

Figure 24: *The ordered eigenvalues of the affinity matrix (7 nearest neighbors of Euclidian distance based similarity matrix) of the Random Forests local importance matrix. A jump of 3 orders of magnitude occurs between the second and third eigenvalue, indicating a 2-cluster structure.*

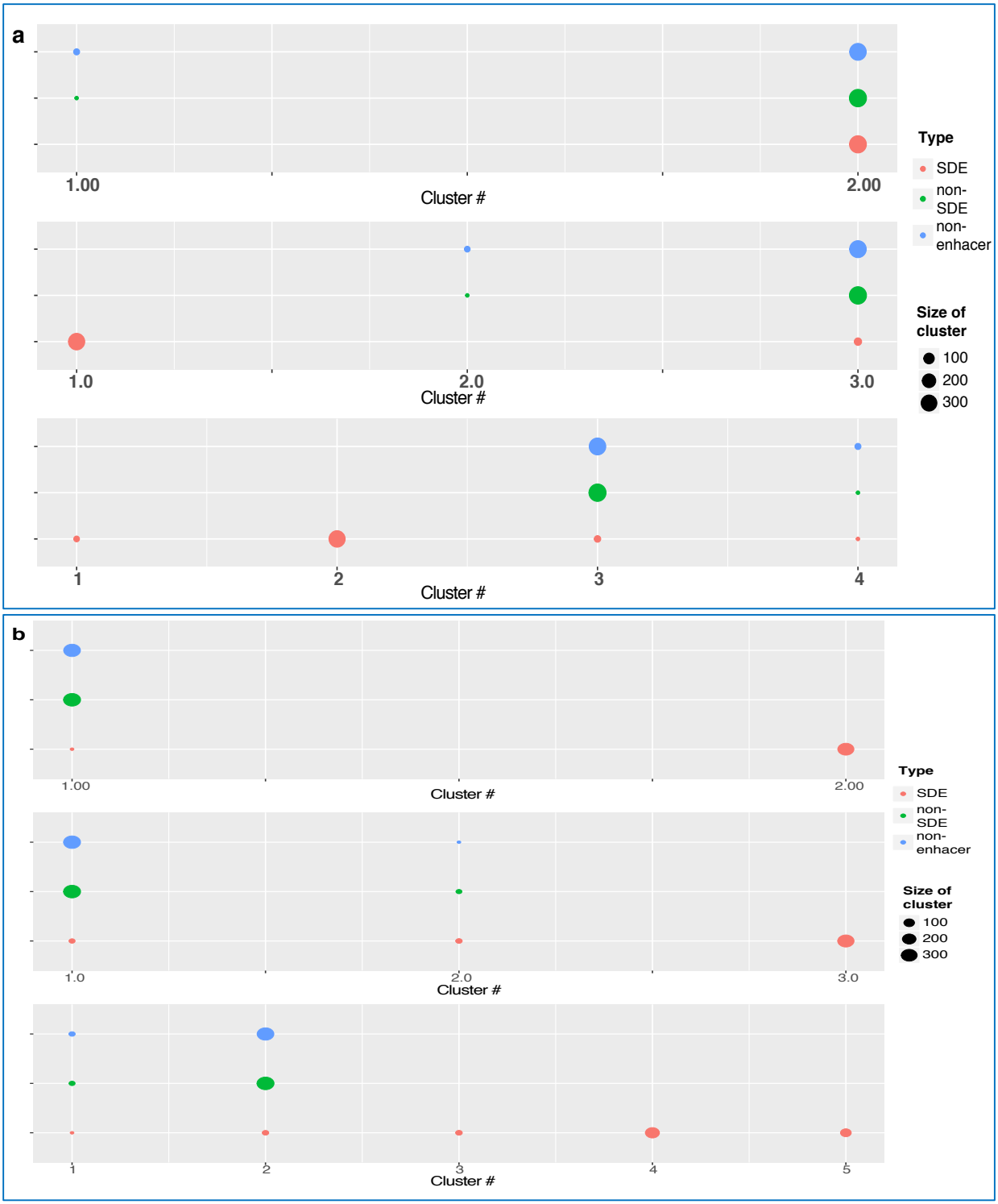

Figure 25: *Number of DNA segments of each type present in each, cluster size indicated by size of the dot, type by color and height* ***(a)*** *Kern-lab spectral clustering of local importance 2,3, and 4 correctly separate SDE enhancers from non-enhancers, as does K-means clustering of local importance* ***(b)*** *K-means clustering applied to the local importance matrix directly, successfully clustering the enhancers. without further dimensionality reduction*

To explore further the concept of substructures, two further clustering attempts were made. Hierarchical Clustering such as Agnes agglomerative clustering [72] can occasionally perform well in locating cluster structures or reveling underlying clusters, however in our data AGNES failed to distinguish either enhancers both in the data set and in local importance (**Fig.26**). A more direct method to discover underlying structures is through hypothesis testing for automated community detection in networks [73]. This method was deliberately developed to uncover hidden communities and sub-clusters in the data, and as it attempts to discover hidden structures, a prior knowledge of the number of clusters in the data is not required. The underlying assumption of the model is that the network in question can be treated as a stochastic block-model, which is not always true, and the bipartitioning hypothesis is measured against an idealized case, which can lead to over-partitioning.

In our case, this tendency for over-partitioning was particularly evident, as the algorithm partitioned the training data and the local importance matrix into 32 and 28 clusters, respectively (**Table 5**). While the large numbers of clusters undermine the utility of this method to partition the data set, it is interesting to note that this method alone was more successful in separating SDE enhancers from the data set rather than the transformed local importance measures. While there is no enrichment of enhancers in any of the clusters resulting from local importance, ~90% of all SDE enhancers are found in a single cluster of the untransformed feature data. Though this technique still has high false-positive rate, as only 1/3 of the cluster members are enhancers, it is particularly noteworthy that this method - which clearly over partitioned the rest of the data - nevertheless agglomerated the SDE enhancers into a single cluster. This may be taken as a strong indication no secondary structures are present in the data. It should also be noted that no cluster, in either of the data sets, contained a significant enrichment of non-SDE enhancers, strengthening our proposition these are indistinguishable from non-enhancer in our feature-space.

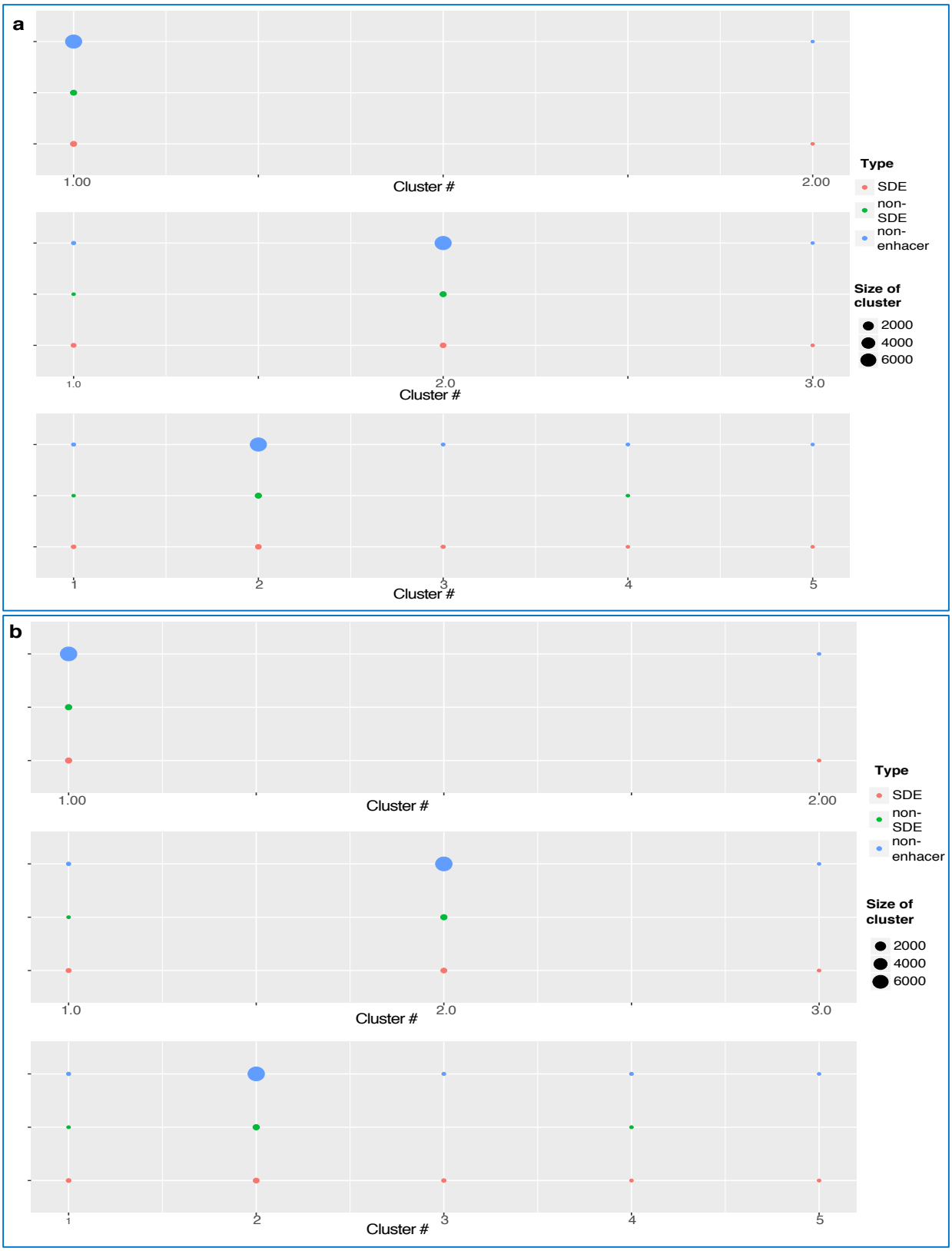

Figure 26: ***(a)*** *AGNES hierarchical clustering of the data* ***(a)*** *or of the local importance* ***(b)*** *failed to separate enhancers and non-enhancers when attempting 2,3 and 4 clusters.*

Table 5: Clusters created through automated hypothesis testing

All Data			Local importance data		
SDE	Non-SDE	Non-enhancer	SDE	Non-SDE	Non-enhancer
0	14	376	2	2	51
0	1	109	2	2	36
0	7	366	4	6	96
0	14	429	5	10	129
0	11	128	3	2	45
0	5	173	2	0	26
0	14	226	0	1	20
0	15	421	0	1	20
0	6	201	2	1	30
0	31	467	0	3	35
0	22	450	1	0	20
0	3	54	1	1	44
0	3	42	3	3	32
316	**49**	**708**	1	1	28
9	18	85	3	3	34
1	7	20	3	3	29
16	41	336	2	1	24
12	18	150	1	3	20
4	21	206	12	5	23
0	1	73	7	1	18
0	0	45	2	3	16
0	20	408	4	4	23
0	3	135	7	4	22
0	3	222	24	11	49
0	8	286	14	1	20
0	7	376	8	0	14
0	20	386			
0	9	339			
0	1	20			
0	1	19			

Enhancer activity in later stages impacts prediction accuracy

Of the 7256 genomic segments experimentally determined to be non-enhancers in stages 4-6, 4031 induce gene expression in later stages, while only 3225 were not found to act as enhancers at any stage. It should be noted though, that little to no data is present on enhancer activity at other points in the post-embryonic drosophila life, so it is impossible to say conclusively these are true non-enhancers. Thus far we considered the entire cohort of non-stage-5 enhancers as non-enhancers. This was done for several reasons; First, it is desirable to be able to separate enhancers in a specific stage from all other segments, without need for further partitioning of the data, if possible. Also, the feature data used in the analysis (transcription factors and histone modification ChIP-chip and ChIP-seq data) are stage specific. It was therefore presumed that late stage enhancers would be indistinguishable in our data set from non-enhancers, and indeed the high-accuracy of our results seemed to supports that theory. Another reason is that in the absence of data on later developmental stages, true non-enhancer data is an unknown and privileges embryonic stages. Yet it is plausible to hypotheses enhancer activity in later stages influence on prediction accuracy will be inversely proportional to the time difference between the stages. Thus, we sought to test the scope by which late-stage enhancement affects prediction accuracy and test our assumption of independence.

Random forest analysis was conducted again, utilizing the same balanced schema described in **methods** and depicted in (**Fig. 5c**), on the same DNA segments training set described at the start of this work, but with later-stage enhancers excluded from the test-set. This was repeated with the full data set, SDE enhancers only, or non-SDE enhancers. While ROC curves show improvement in predictive power (as demonstrated by the area under the ROC curves) when later-stage enhancers are excluded from the analysis, this improvement is consistently small, whether we consider the entire enhancer set or focus on SDE or non-SDE enhancers (**Fig.27a-c**). In contrast, the precision-recall curves show a marked improvement in all three cases (**Fig.27d-f**), particularly in the case of SDE enhancers, where the area under the curve rises from 0.75 to 0.95, an exceptionally high value not previously reported in this kind of studies. As precision measures the probability a genomic region to be an enhancer given we predict it to be an enhancer regardless of priors and sample-imbalance, it is arguably the more relevant measure in this case where we seek to minimalize the type-I errors (false positives) of a relatively rare occurrence (i.e. – enhancers). While some studies reported roc curves with areas above 0.9, albeit when examining small samples from the top-ranks of the enhancer prediction, the PR curve represents unprecedented precision far exceeding previous reports.

An additional motivation for this analysis has been the hope that it will aid in predicting non-SDE enhancers, but while a significant improvement in precision is present in analyzing the all enhancers and non-SDE enhancers (**Fig.27d-f**) the areas under the ROC and PR curve values are still indicative of performance not much better than random guessing, and clearly insufficient to allow us to make predictions. It also appears that the effect of inclusion of later-stage enhancers in the test set is independent of their presence in the training set (compare **Fig. 27a-f** to **Fig.27g-l**).

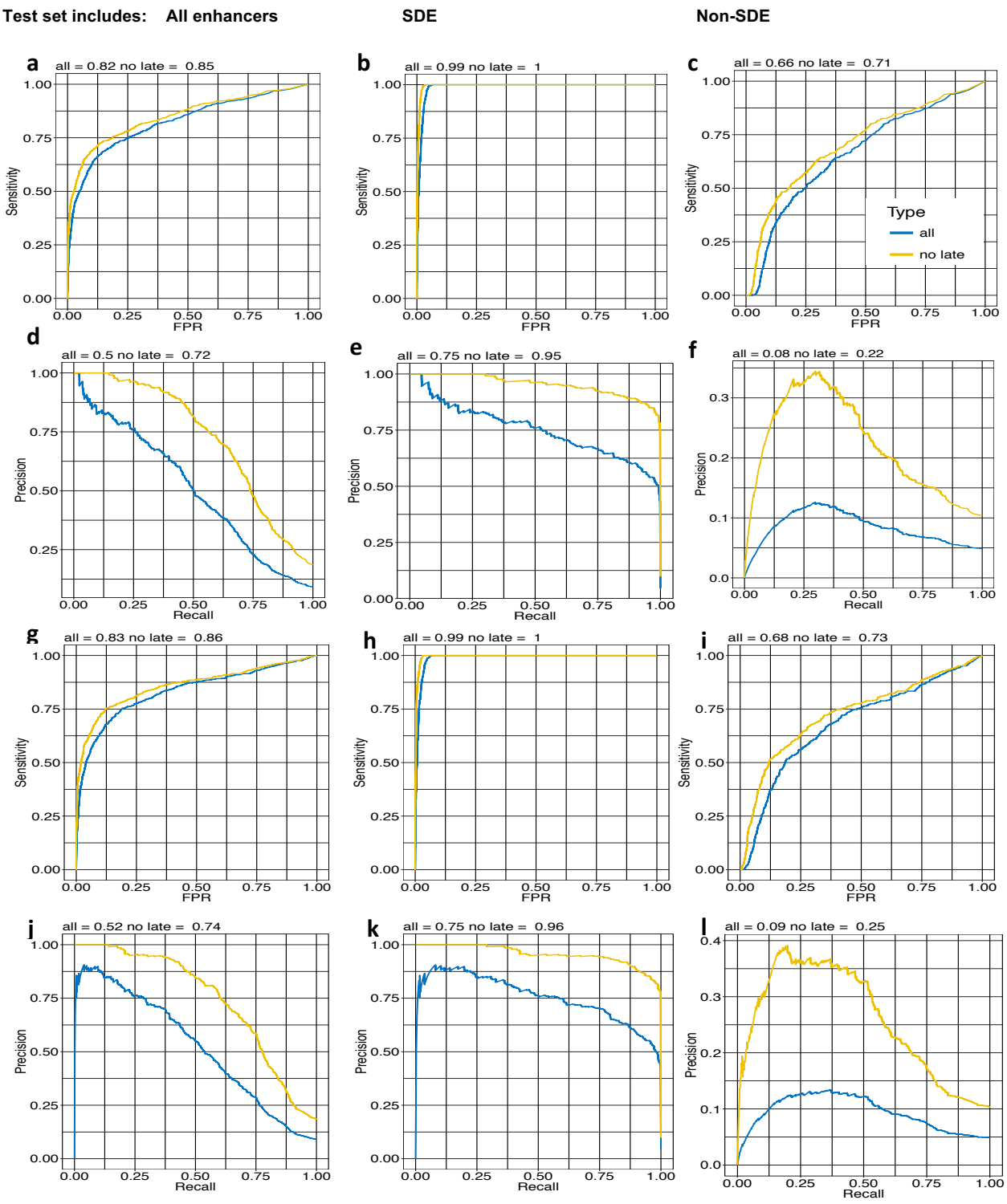

Figure 27: Random forests *ROC and PR curves with all non-enhancers (all, blue) and with only those which remain non-enhancers in later stages (no-late, yellow) present in the test set. Respective areas are indicated above the figures. First column shows all enhancers, second SDE enhancers and third non-SDE enhancers.* ***g-I*** *shows the same experiments, where late-stage enhancers are excluded from both training and test sets, with little to no effect.*

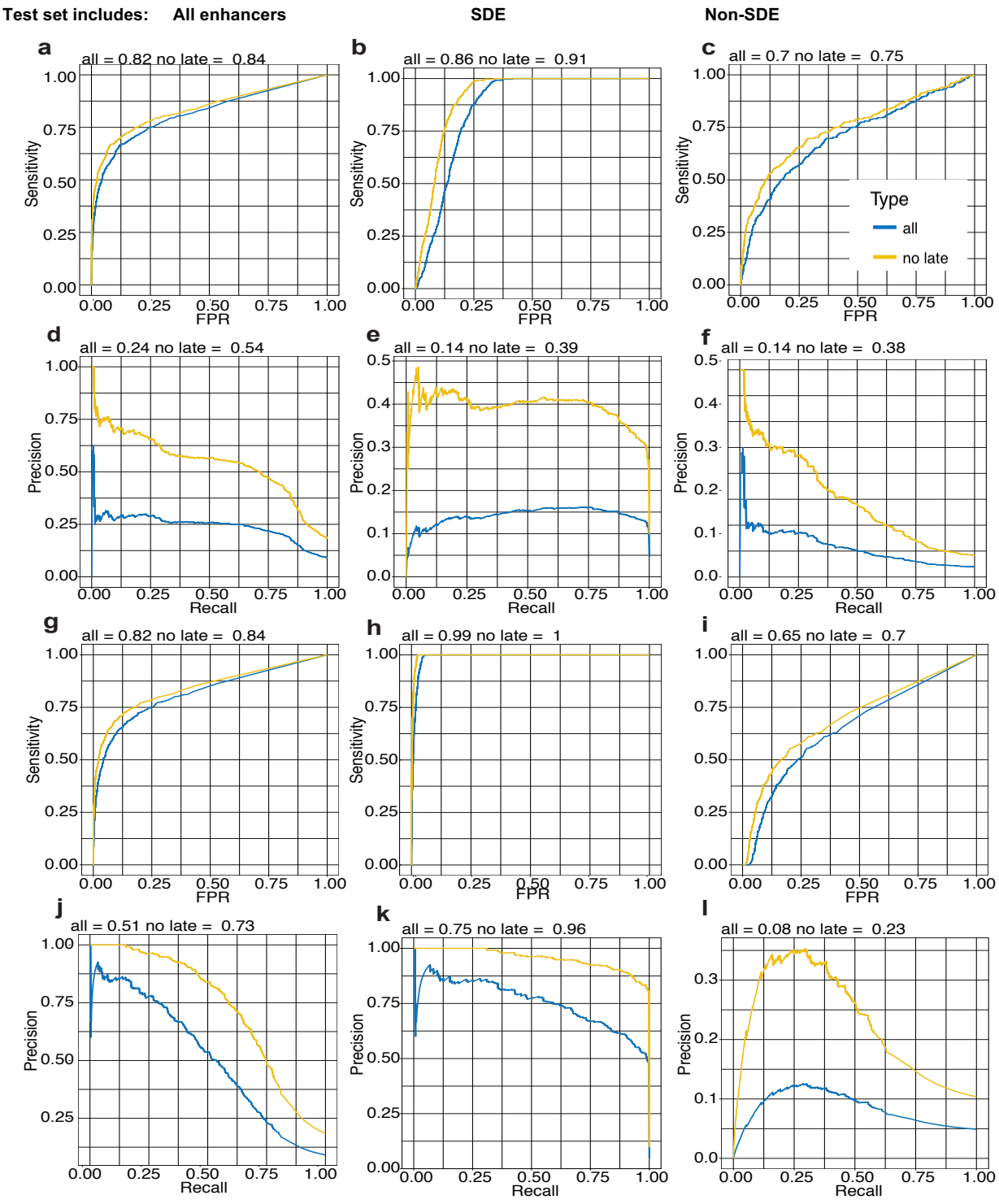

Figure 28: *Same as figure 27, except training set is limited to* ***non-SDE*** *enhancers and no late stage enhancers in control group (**a-f**), or to* ***SDE*** *enhancers and no late stage enhancers in control group (**g-l**). Note the change of scale on last figure in each set. Training on non-SDE enhancers drastically reduces SDE prediction precision with no gain in non-SDE prediction, while training on SDE enhancers only has little to no effect.*

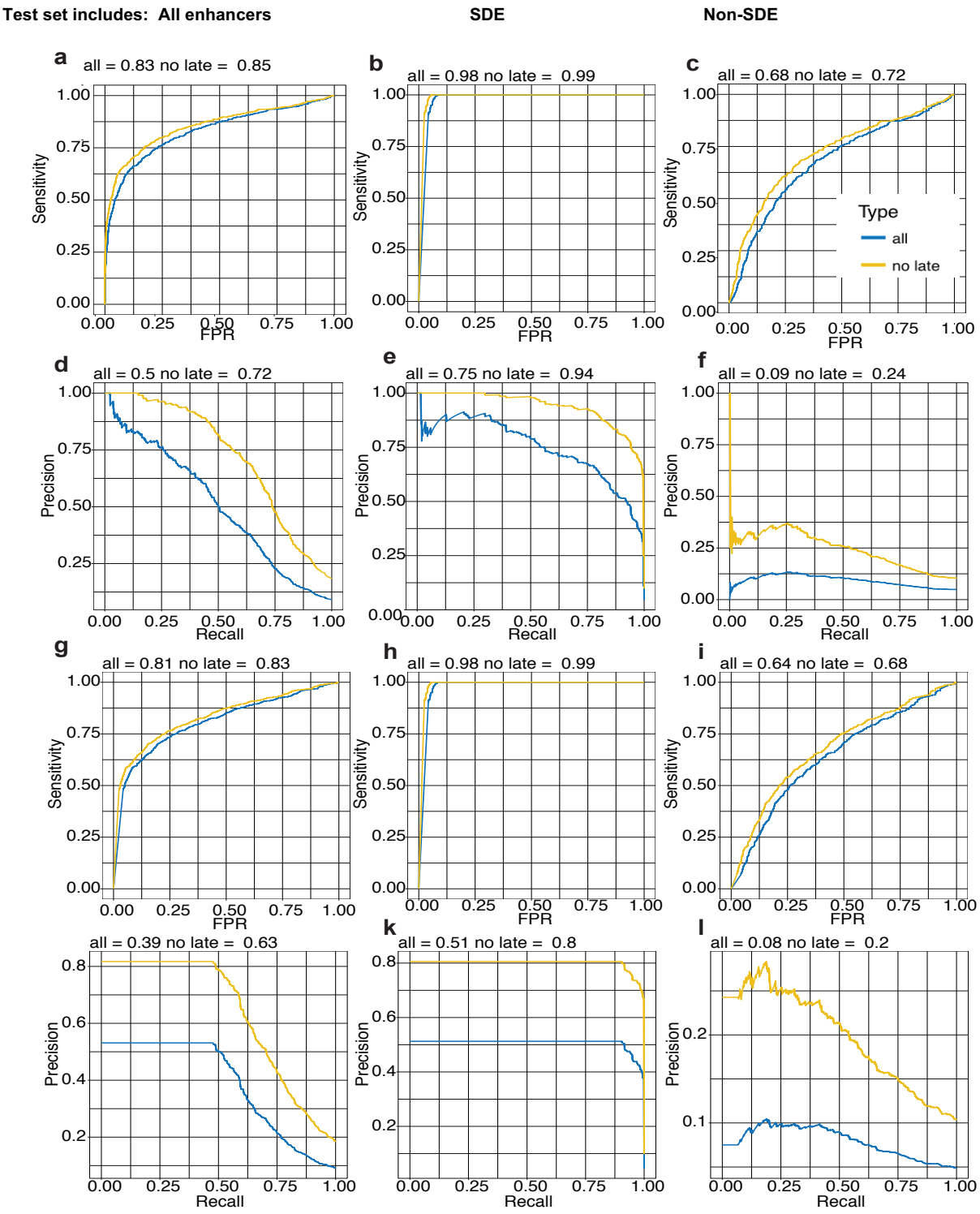

***Figure 29: Logistic regression** (a-f) and **naïve Bayes** (g-l) ROC and PR curves with all non-enhancers (all, blue) and with only those which remain non-enhancers in later stages (no-late, yellow) present in the test set. Respective areas are indicated above the figures. First column shows all enhancers, second SDE enhancers and third non-SDE enhancers. Note the change of scale on last figure of each set.*

To understand the relationship between data inclusion in the test and training data, the experiments were repeated with all combinations of All-enhancer/SDE only/Non-SDE-only and all-stages/No-late-enhancers present in the training and testing set. For example, (**Fig. 28 a-f**) presents ROC and Precision-Recall curves for random forests analyses in which non-SDE enhancers only were included in the training set as enhancers, and only those non-enhancers which do not drive later stage expression in the embryo are included as non-enhancers. They show training sets which include either all enhancers (**a, d**), only SDE enhancers (**b, e**) or only non-SDE enhancers (**c, f**), and either included (blue) or excluded (gold) later-stage enhancers. While the ROC curves remain relatively high despite the exclusion of SDE enhancers from the training set, a study of the PR curves quickly reveals that their absence removed most of the signal from the analysis, with the highest area under the PR curve ~0.55. It is interesting to note that even though SDE enhancers were excluded from the analysis, prediction of SDE was still more accurate (**Fig. 28 b, d vs. c, f**). This is probably due to the arbitrary cutoff by which SDE and non-SDE enhancers were separated, and suggest that most of the remaining signal in the non-SDE pool belongs to wrongly classified SDE enhancers. The exclusion of non-SDE enhancers from the training set had no effect on either ROC or PR curves (**Fig.28 g-l**), including prediction of non-SDE enhancers, furthering that supposition.

In all cases, removal of late-stage enhancers from the testing set improved precision, while its inclusion or exclusion from the training set had no impact (**data not shown**). It is likely that the signal in the SDE and true non-enhancers is strong enough to overcome the ambiguous signal by those non-SDE or later-stage enhancers to correctly classify them. The fact that the inclusion of non-SDE enhancers or late-stage enhancers is important, as there is no way to know a-priori which type of enhancer each data point will prove to be, or whether a non-enhancer remains a non-enhancer in later (non-embryonic) stages. It indicates that including all known enhancers is sufficient to the task of predicting SDE enhancers without the need of prior separation, and that while information of enhancer activity is useful for evaluating the result, it is not needed for accurate predictions.

To compare the veracity of this argument and in the hopes of finding a method to accurately predict all or some of the non-SDE enhancers, logistic regression and naïve Bayes analyses were conducted again, with and without later-stage enhancers in the test set (**Fig.29**). As with the Random Forests analysis, the exclusion of later-stage enhancers marginally enhanced recall, and significantly enhanced precision, but did not significantly enhance our ability to detect non-SDE enhancers. It should also be noted that while the ROC curves of all three methods indicate equally high performance, PR curves indicate random forests is the best performing method, while naïve Bayes performs quite poorly in comparison. The analyses were repeated again with all combinations of all-enhancers/SDE-enhancers/non-SDE enhancers, and all non-enhancers/no-late enhancers included in the training set, and evaluated with the inclusion or exclusion of those groups (**data not shown**). The results for both naïve Bayes and logistic regression were similar to those reported for Random Forests: exclusion of later-stage enhancers from the non-enhancer set makes a small difference in recall and large difference in precision, but does not grant sufficient power to predict non-SDE enhancers.

Genome-wide scan to identify all segmentation-driving enhancers in the early embryo

Given the high accuracy of the model on our training and held-out data set, a genome wide search for SDE enhancers was feasible. Random Forest was used to predict enhancer score on a computationally segmented genome (see methods). More than 0.82% of all segments had less than 0.01 score of being enhancer, and more then 93% had less than 0.1 predicted score **(Fig. 30a**). While it is hard to see at first glance, as the histogram is dominated by the 0-0.01 score bar, the histogram is in fact bimodal (**Fig. 30b, inset**), with a secondary peak around p = 0.95. It should be noted that a similar bimodality is also present in the score distribution of the training data, indicating this is a feature of the genome – or at least, of applying random forest analysis to the genome (**Fig. 30c**). The bimodality is not present if the split criteria between SDE and non-SDE enhancers is set at a significantly lower threshold, which serves as a validation to the segregation criteria between the two classes (**Fig. 30d**).

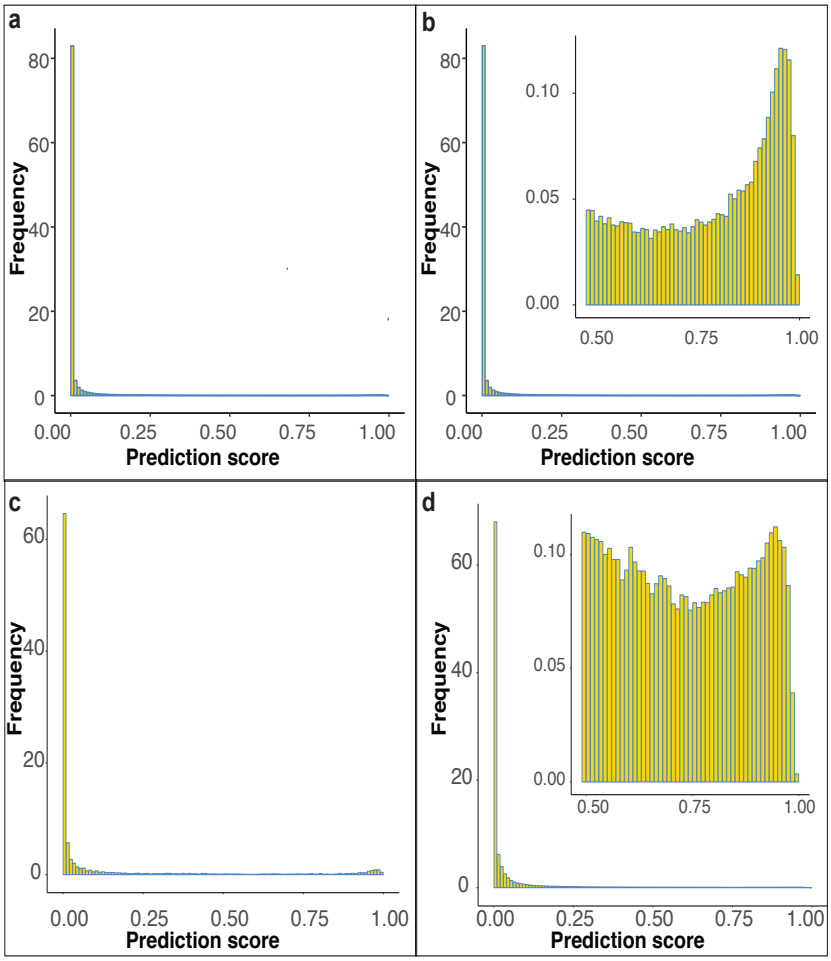

Figure 30: Predicted score distribution of the whole genome is bimodal. ***(a)*** *Histogram of the prediction score for every 1000bp sliding window along the drosophila genome is dominated by >80% of the genome having a score < 0.01* ***(b)*** *a closer examination of the second half of the histogram reveals the distribution is bimodal (inset)* ***(c)*** *This bimodality was also observed in the predicted score histogram of the training data* ***(d)*** *when repeating the process with a less-stringent separation of between SDE and non-SDE (treating 80 more enhancers as SDE enhancers), bimodality is not observed, or is at least much more ambiguous.*

In order to call enhancers, a threshold of p> 0.75 was established, covering ~1.6% of the genome and rediscovering *de novo* 98% of the training set. The threshold of 0.75 is a natural one for this sample, as it was also the threshold used to designate SDE enhancers. The resulting overlapping segments were combined into continues segments giving rise to 1374 predicted enhancers of varying lengths (**Fig. 31a**). While most enhancers were relatively short, with a peak at ~1700bp, some enhancers were several Kb long (one longer then 12Kb), which likely indicates an enhancer cluster rather than a single enhancer. Enhancers longer then 1.5Kb were further separated based on their transcription factor biding profile where possible (**see methods**). While this was not always possible, either due to a true single very long enhancer or peaks too close to separate, we were able to separate over 200 peaks, with none of the remaining longer then 4KB.

Random Forest predictions were conducted again with the resulting predicted enhancers as test case in order to evaluate the confidence in prediction of the new boundaries. For the most part, the score histogram shifted right as the flanking regions containing only partial signals were no longer clouding the predicted score landscape (**Fig. 31b-c**). In a few cases, the new boundary enhancers yielded predictions with scores below our threshold of 0.75 (**Fig. 31c**), likely due to the stochastic process of Random Forests. We chose to keep theses enhancers in the prediction pool, with the new scores assigned.

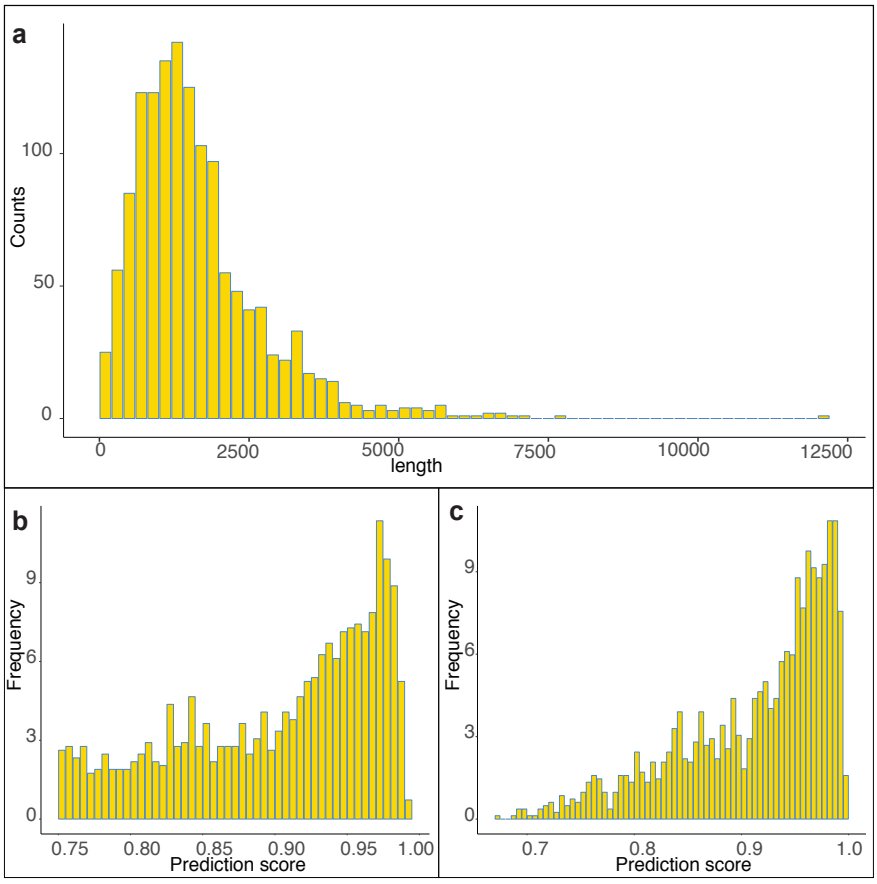

Figure 31: ***(a)*** *length distribution of predicted enhancers prior to peak splitting. While most are shorter then 2Kb, some are extremely long (>4Kb), indicating likely an enhancer cluster.* ***(b)*** *Histogram of the predicted score of the new boundaries enhancers, taken as the maximal score of the composing segments prior to unification.* ***(c)*** *Histogram of predicted score of being an enhancer based on a new Random forest prediction with the new boundaries.*

All together, we predict 1,640 SDE enhancers along the genome, 1174 of which do not contain overlap with training data. 364 overlap known CRM's identified in the Redfly database [74-76], while 916 are novel. Examples for how the model performed in the comprehensively investigated enhancers clusters Eve and Ftz can be viewed in **(Fig 32-33)**.

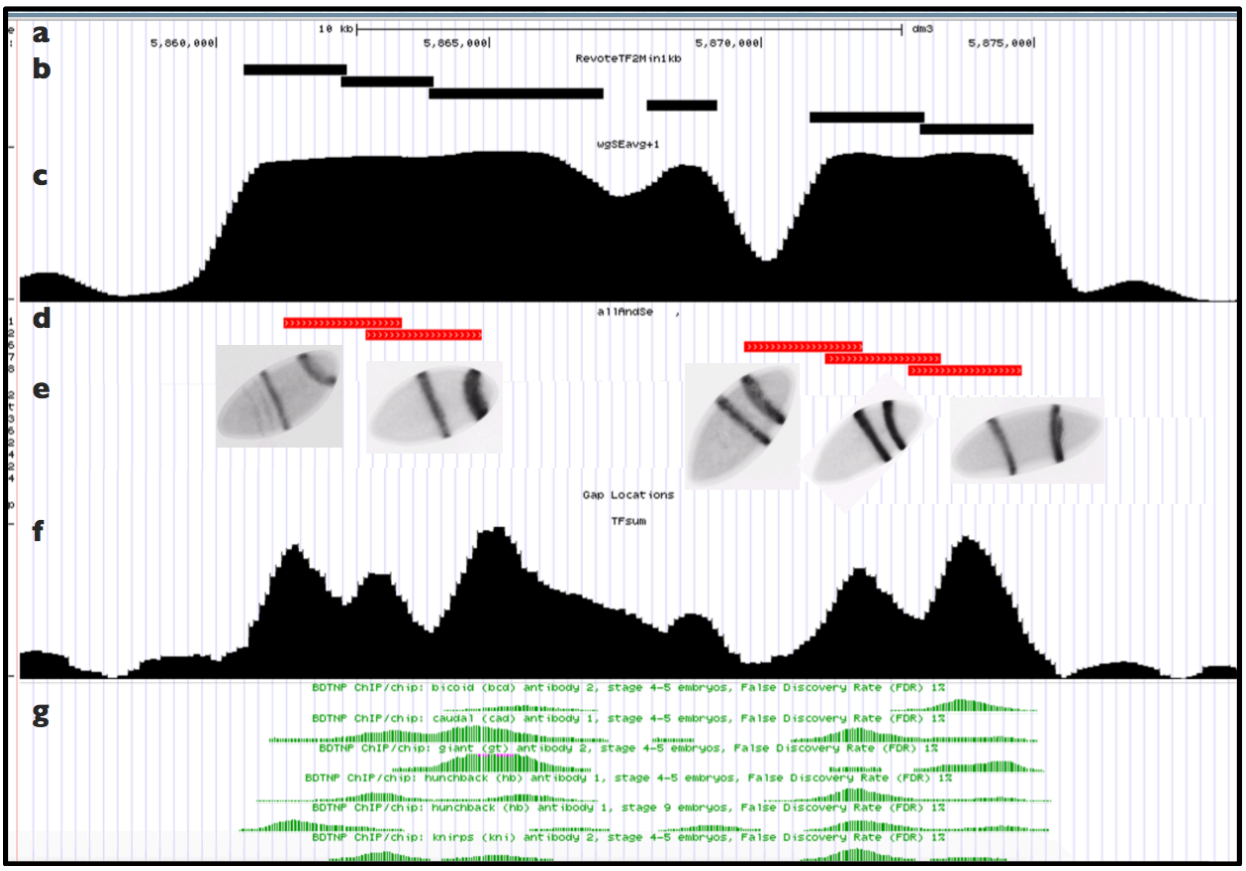

Figure 32: *A UCSC genome brewers of the enhancer cluster regulating the Eve gene expression, containing (from top to bottom):* ***(a)*** *Genomic location and scale* ***(b)*** *Predicted enhancers* ***(c)*** *Row data from which peaks were called and subsequently separate* ***(d)*** *location of training-set enhancers (red indicating they were found to induce expression in stage 4-6 embryos)* ***(e)*** *Figures of the embryos in the training set, upon which their enhancer ability was determined.* ***(f)*** *Transcription factor binding profile of the region, being the sum of all known individual transcription factors bindings, some examples of which are presented in* ***(g).*** *The model correctly predicts the location of all seven known enhancers, and correctly separates 6 of them into distinct enhancer regions.*

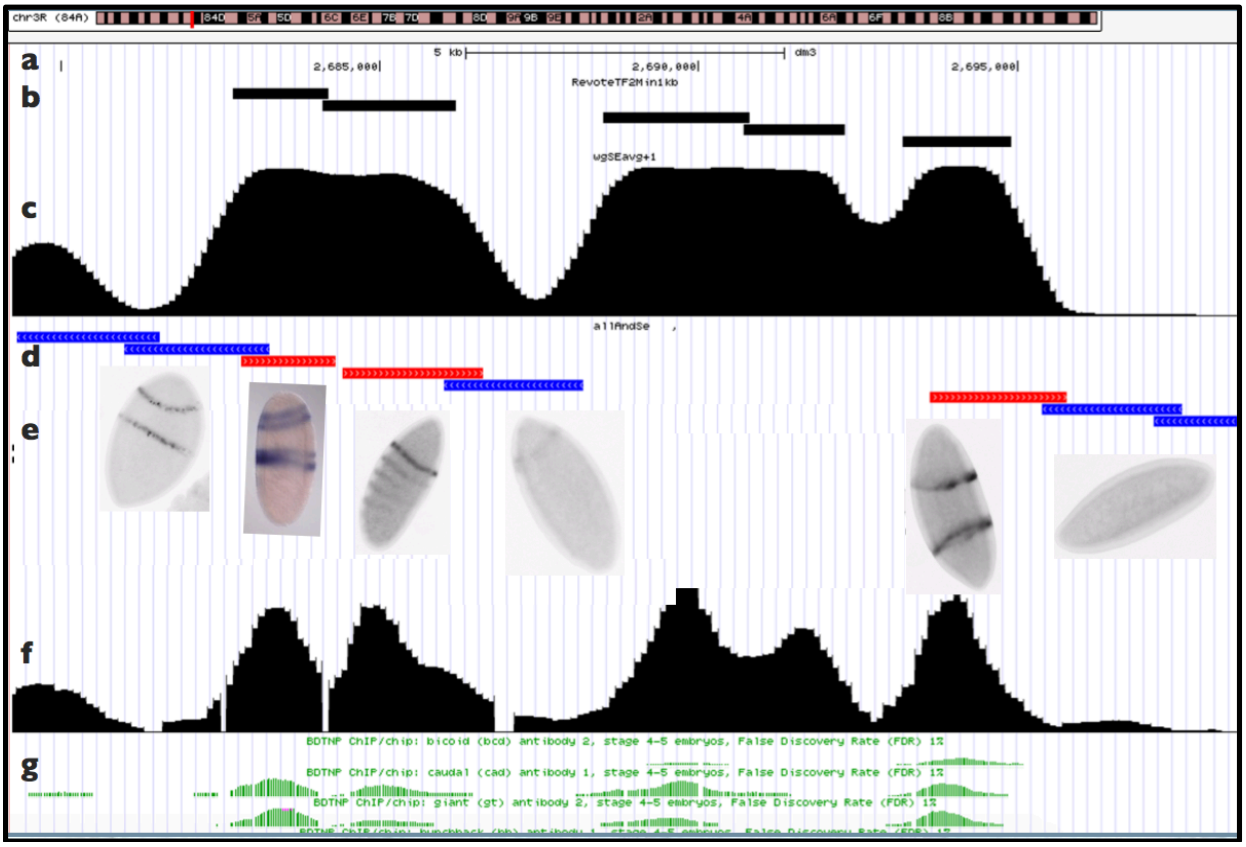

Figure 33: *A UCSC genome brewers of the enhancer cluster regulating the Eve gene expression, containing (from top to bottom):* ***(a)*** *Genomic location and scale* ***(b)*** *Predicted enhancers* ***(c)*** *Row data from which peaks were called and subsequently separate* ***(d)*** *location of training-set enhancers. Red indicates genomic segments found to be stage 4-6 enhancers enhancer, while blue indicates those which were not* ***(e)*** *Figures of the embryos in the training set, upon which their enhancer ability was determined.* ***(f)*** *Transcription factor binding profile of the region, being the sum of all known individual transcription factors binding, examples of which are presented in* ***(g).***
The model correctly predicts both the presence and absence of enhancer activity in regions overlapping the training set. In one case (fifth tested enhancer from left) the tested enhancer overlapped a predicted enhancer, and yet was determined to be a non-enhancer. Examination of the embryo below revels this is in fact a weak enhancer, an example of experimental false negative.

To validate our precision, an in-vivo expression-driving test were conducted. 6, 17 and 18 genomic regions were selected with predicted scores corresponding to expected false discovery rates of 4%, 25% and 50% respectively. Test regions were cloned into the pBPGUW expression vector then injected into flies using the attP integration system [77]. All but two of the enhancers, including all but 1 of those predicted to be in the 50% FDR region were found to be enhancers (**Fig. 34a-f**). We thus needed to adjust our FDR estimation, and assuming a Poisson distribution we placed a confidence bound of 0.15-

32.77% FDR on our predicted 25% FDR, and 0.14-30.95% FDR for the region we predicted to have 50% FDR, with most likely values (MLE)

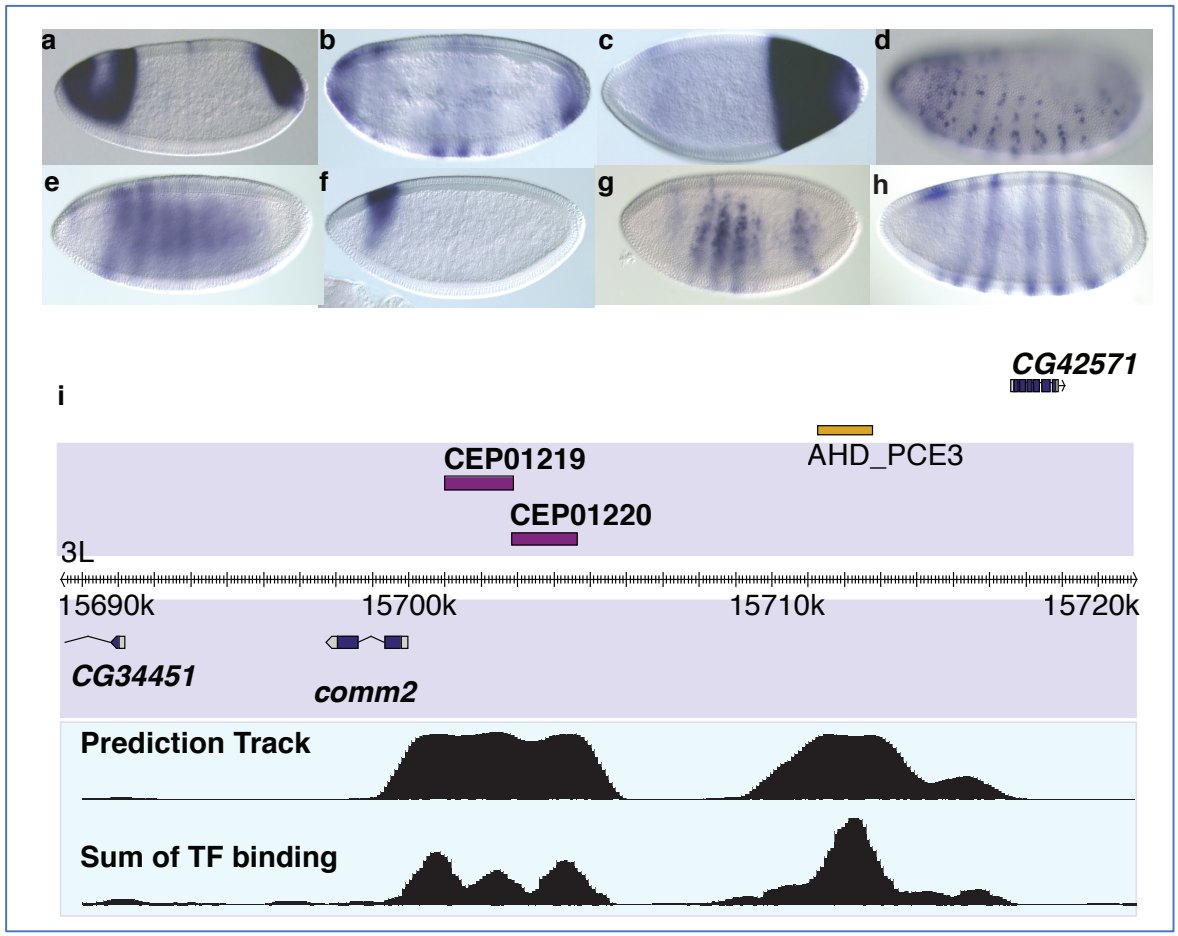

Figure 34: **(a-e)** As validation, predicted enhancers were inserted into drosophila genome, and were found to drive expression. **(f-g)** two enhancers are predicted proximal the to the comm2 gene. Each of their patterns is a different portion fractions of the comm2 gene expression pattern **(h). (i)** The genomic region of the two predicted enhancers is shown, along with the raw prediction track showing the predicted score of enhancer activity with 100bp resolution, and the sum of transcription factor binding at the same resolution.

of 5.88% and 5.56% respectively. Thus, the predicted FDR to the 1121predicted enhancers above the tested threshold is 5.73% and better then 31.9%. By fitting a polynomial model through the data and extrapolating for the missing value, we estimate an overall FDR of 6.28%, and no more than 33.27% (**Fig. 35**). This large range may be further reduced with further experimentations done, particularly on the lower ranked enhancers.

An interesting example and validation for the use of transcription factors to separate proximal enhancers (**methods)** can be seen in two predicted segments proximal to Commissureless (comm2) (**Fig. 34i)**, an important protein in axons guidance across embryo's midline[78,79]. The two predicted enhancers combined expression pattern (**Fig. 33f-g)** matches the comm2 more complicated expression pattern (**33h).**

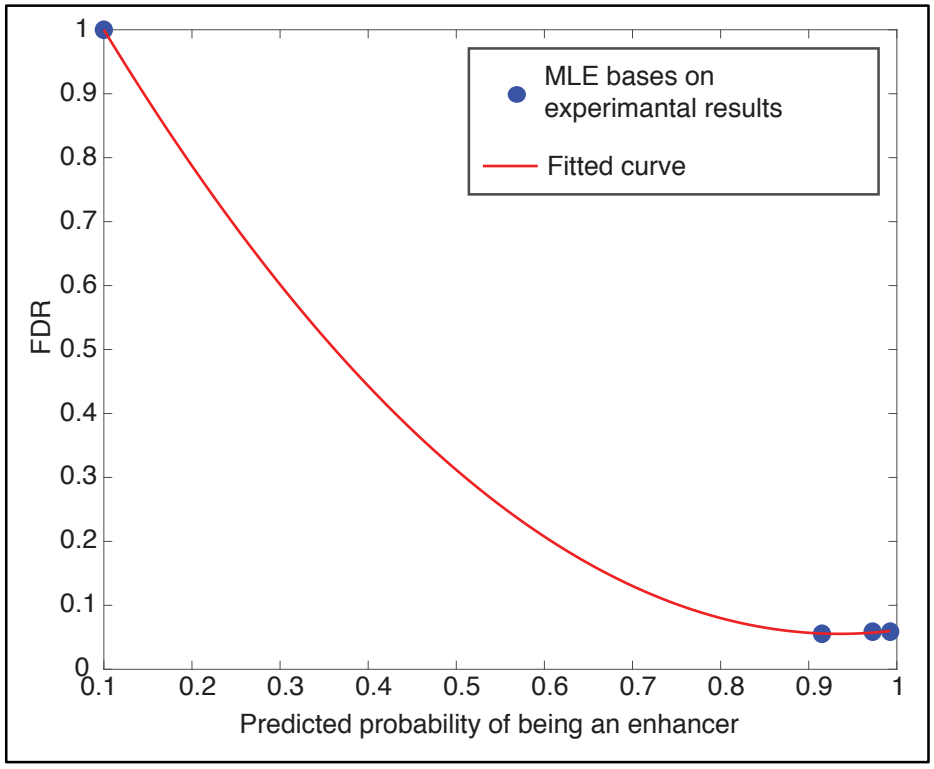

Figure 35: *To calculate an overall FDR for the entire cohort of predicted enhancers, the MLE of the three tested regions was calculated (assuming an underlying Poisson process). In addition, all DNA segments with scores <0.01 were assumed to be non-enhancers. These 4 data points were used to fit a polynomial function, which gives a predicted FDR for each predicted score, allowing us to find the most likely FDR of cohort. This was repeated using the two CI values for each region, to give the confidence interval for the FDR (not shown).*

DISCUSSION:

The identification of enhancer elements from genomics data has remained a challenging problem – in part due to the relative scarcity of enhancers in genome sequences. As we illustrated in (**Fig. 10**), even an extremely incisive enhancer prediction algorithm fitted on balanced training data (training sets with nearly equal numbers of positive and negative elements) is likely to generate high false discovery rates when applied to a genome-wide scan. Hence, to accurately discover enhancer elements using *in silico* techniques, extremely high fidelity models are needed.

Though high precision predictions were reported previously, validation method and measure varied greatly in literature. Many papers compare their success to the colocation of known epigenetic marks such as p3000 and H3K27ac, but it is yet to be established that these are exclusive to enhancers or that all enhancers possess them. Indeed, we report here a H3K27ac-free enhancer class. Others report results tested at the top of the rank list, which forms a biased estimation of the overall prediction accuracy. We suggest reporting of precision measured by testing throughout the prediction rank list is required to establish a uniform unbiased measure of success.

By far the most important factor in the prediction of SDE-enhancers is transcription factor binding profile, to the extent that utilizing transcription factors alone yields results similar to utilizing the full array features, and in combination with histone modification no other feature was required. Yet most other data sets are not irrelevant, merely redundant, as the error rate of all features without transcription factors is not much lower than that with it. DNase accessibility especially plays a much greater role in prediction when transcription factors are excluded, while conservation scores appear to have little to no ability to aid in correct classification (see **Fig. 5**). This redundancy of data indicates many different approaches may prove rewarding to the problem of enhancer discovery.

We found that the prediction of enhancer elements *en masse* was made difficult by heterogeneity among enhancer elements: for about half of validated enhancer elements, strong TF binding signal for multiple factors is indicative of enhancer activity in our system; for the other half of elements, enhancer activity and TF binding signal are only weakly associated. That is, a prediction engine that works well on one class tends to fail on the other. We posit that this challenge, heterogeneity in element classes, is a widespread and foundational challenge in genomics. For example, the emerging literature on “chromatin priming elements” [80] demonstrates the existence of “enhancer-like” functional elements that, while they share chromatin structure and similar patterns of transcription factor occupancies with enhancers, do not themselves drive patterned expression – rather they establish chromatin context that subsequently gives rise to enhancer activity for proximal elements. It may be that the class of elements we presently denote “enhancers” is in fact diverse, admitting elements that exert regulatory effects through a variety of underlying molecular mechanisms. Indeed, it remains unclear what fraction of enhancers require eRNAs for their activity [80], or whether priming elements are transcribed like enhancers.

It may also be that the non-SDE or “class II” enhancers we study here are simply regulated by cohorts of TFs we have yet to survey, or that activity in a smaller fraction of the embryo tends to make these elements less amenable to interrogation through whole-

animal ChIP-seq. However, the differences between the two enhancer classes are statistical, not categorical – while there is a significant enrichment in segmentation GO-terms in SDE compared to non-SDE enhancers, some non-SDE enhancers also display segmentation-driving expression patterns, and many (20%) are active in as much or more of the embryo than the median for SDEs.

The separation of SDE from non-SDE enhancers in all parameter space allows us to utilize Random Forest directly to separate the classes. Studying the feature importance of this analysis shows the same features are utilized in this as are used in the separation of SDE's and non-enhancers. The same can also be seen in observing the local importance heat maps, where it is clear the same features can be used to separate SDE and non SDE enhancers. Local importance also shows us how the presence of non-SDE's in the training set changes the way features are utilized by the Random Forests to predict enhancers; the importance of the top transcription factors drops and more features are used though with less consistency (**Fig. 21**). This is a surprising result as ROC and PR curves seem to be unchanged by the presence or absence of non-SDE's in the training set (**Fig. 26-27**). It may be that while the Forest use the critical features less in the presence of SDE enhancers, it is still enough to correctly classify SDE enhancers in the test set if the signal is strong enough.

It is noteworthy in that respect that while little separation of non-SDE enhancers from non-enhancers was possible with our method and data, that separation relied heavily on single transcription factor – Twist, a crucial Dorsal-Ventral patterning [37], which is also found to be consistently most important in SDE classification. As it fails to correctly classify non-SDE enhancers most of the time, it may be Twist is part of a transcription network with other members not included in this work, leading to its limited power. Alternatively, it may be that Twist is weakly correlated with other factors related to non-SDE activity though not directly involved on their regulation, and so it has some limited predictive ability on that cohort. Yet another possibility is that the weak discriminatory power Random Forest classifier has on the classification of non-SDE enhancers stems from SDE enhancers included in the non-SDE cohort by the arbitrary threshold imposed to separate the classes. As error rate in the classification does not drop when Twist is removed from the analysis, this last seems the most probable conclusion, but It will be necessary to conduct further study to determine definitively which of these hypotheses is correct.

When comparing the performance of Random Forests to logistic regression and naïve Bayes, it was surprising how closely ROC curves resembled – indeed, overlaid – each other, given the far more stringent underlining assumptions in both these second models (**Fig. 6b**). However, as the ultimate goal of this analysis is to allow researchers to accurately predict enhancers, it is the precision and not recall which is the more relevant measure. When studying the PR curves of the different methods (**Fig, 26-28**) it is clear naïve Bayes is the worst performing of the three, unsurprising as it is manifestly untrue that the features used in the prediction are independent. While logistic regression performs better, Random Forests is clearly the best performing method of the three.

In contrast, clustering methods consistently fail to separate the data, indicating this problem is not tractable by unsupervised learning methods. The best performing method was Clustering through automated hypothesis testing, which successfully classified most

SDE enhancers together, yet less then 1/3 of this cluster members were SDE enhancers (this despite over segmentation of the method on the whole). Classifications conducted on the feature importance space were consistently successful, and spectral clustering performed on it suggests 2 classes are present in the data (**Fig. 23**). Together these may be taken as a strong indication as to the homogeneity of the SDE enhancer set.

Precision can be further increased when non-enhancer data is limited to those regions which do not induce expression in the embryo, not merely in the early blastoderm. This indicates that the separation in time between active and inactive enhancers is not complete with regards to transcription factor binding and histone modification, and that poised enhancers may be confused with active enhancers in feature space. This is not surprising, as their varying regulatory functions such as inhibitors which rely on the removal or addition of a single factor to initiate transcription. In such cases, all the enhancer signals mast be already present beforehand for this regulatory scheme to work. When late stage enhancers are removed from the test set rather than considered non-enhancers, both ROC and PR curve for predicting SDE enhancers has an AUC > 0.99, (**Fig. 27b**) indicating almost perfect classification. It is therefore reasonable to hypothesize that some of our false negative predictions may turn out to be expressed in later stages. Further experiments are needed to validate this supposition.

While the presence of late stage enhancers or non-SDE enhancers in the test set greatly effects our precision, their presence in the training set has no effect on either recall or precision (**Fig. 27-29**). This result seems to suggests that all known data may be given to a Random Forest classifier without the preliminary separation into classes as was done here. And yet, when it is sufficient to slightly relax the separation criterion between SDE and non-SDE enhancer to alter the whole-genome prediction so that the distribution is no longer bimodal, and indeed Random Forests clarifications previously reported were unable to achieve such accuracy [18]. It may be that the difference lies simply in the difference between cross validation used to construct the PR and ROC curves and true held out test set, such as is represented in the whole genome analysis. More probably, it lies with the high number of enhancers belonging to the same class (SDE) present in out training set. This inclusion was not intentional, but the lucky result of extensive study into segmentation which provided us with a trove a single process specific data.

Our validation assays revealed that cross validation has led to significant overestimation of the false discovery rate for SDEs. We attribute this to an abundance of false negatives in our training set – perhaps 10% of negatives are erroneously labeled, which, if true, would double the number of positives, and explain the significant gap in our anticipated versus validated false discovery rates. Another possible explanation to the discrepancy is selection bias in our training set, as genomic regions to be tested were selected as likely candidates for enhancer activity and not at random. Thus, it is possible that the genomic data as a whole gives clearer separation of features. Overall, we recover 98% of the training set SDEs with an estimated false discovery rate of less than 7%, indicating that our genome-wide predicted catalog of these elements may be close to comprehensive. Further experiments, particularly concentrated at high false discovery rates, are needed to better assess the boundary between functional and non-functional elements. At this time, it appears that at least 1600 elements, composing more than 1.5%

of the *Drosophila* genome, are involved in establishing early body patterning in the blastoderm.

MATERIALS AND MATHODS:

Data acquisition and processing:

25% FDR Transcription factor ChIP-chip data was taken from the drosophila TF network project (bdntp, http://bdtnp.lbl.gov/Fly-Net/), containing data for 22 transcription factors: bcd, cad, D, da, dl, ftz, gt, h, hb, hkb, kni, kr, mad, med, prd, run, shn, slp, sna, tll, twi, z, some with biological duplicates to give 34 tracks. 1% FDR ChiP-chip data for polII binding was also taken from BDTNP [59-61]. Histone, histone modification and zld chip-seq data were retrieved from UCSC genome browser track provided by Li el al.[62]. Histone modifications data collected in Zld mutant strain were not used, all other tracks were included in the analysis. DNase accessibility data and 12-fly conservation phastCons scores were obtained from UCSC genome browser[81-83], as was FlyBase gene data for exon, coding exons and intron location data[84]. Bidirectional RNA transcript data was obtained from lab unpublished data. Transcription start site was taken from FlyBase's mRNA data.

Though 80% of the DNA segments in the training set were between 2-2.5Kb long, segment sizes varied from 100bp to 4.5KB in the set, and the percent of enhancer region contained by each segment is unknown, making averages a biased estimator. Thus, the maximum of ChiP data was calculated over every segment in the training set and the segmented genome using bedtools and UCSC genome browser utilities for TF data, histones, conservation score and DNase accessibility. In addition, the sum of TF biological replicas and the sum of all TF tracks was also calculated and included as features in the model. In addition to the maximum score, for Zld higher-resolution ChiP-seq data and for the conservation phastCons conservation scores we also calculated the average over the segment, maximal score over a sliding window of 200,500 and 1000bp, and the longest continues stretch of scores above the 0.85 quantile. For the gene data, bedtools coverage was used to calculate percent of segment covered by exons, coding exons or introns and 3 binary tracks indicate the presence or absence of intron and exons. Bedtools closest to calculate distance to the closest tss and to polII binding peak.

Modeling:

Random Forests were modeled in R [85] using RandomForest[40] . Initial feature set culling was done through error rates average of 1000 forests of 500 trees when excluding/adding one feature at a time. Our training data is highly unbalanced, with only 10% of segments being enhancers. To improve Random Forest performance balanced samples were used as training set. To improve stability of the prediction, and counteract the sampling process employed by balancing the training set, we relied on forest voting. 1000 Forest of 50 trees each were trained on a randomly selected sets of 300 enhancer and 300 non-enhancers with 10% of the data held out of the samples and used as test set. The fraction of trees in all forests voting for each segment serving as the predicted

score of being an enhancer. this was repeated until such score was computed for each segment in the set. The same sampling and testing scheme was employed for logistic regression, and naïve Bayes [86].

Importance measures varied from sample to sample and averages required 10,000 Forest of 50 trees to converge. To increase stability of the importance measure, the average of 50,000 Random Forests mean decrease in accuracy and mean decrease in Gini index were used to find the importance Random forests confidence intervals. For local importance calculations, we used a single forest of 50,000 trees produced using all enhancers and a balanced non-enhancers subsample.

Analyses:

ROC curve areas were calculated with R package PRROC [87]. PCA was done using prcomp [85]. Go term analysis used bedtools [88] to find FlyBase genes located inside training enhancer regions, or to identify the closest genes if none are overlapping. David bioinformatics resource [89,90] was used to find and quantify Go term and Go-term enrichment, with the full set of ~8000 genomic regions as the genomic background. To find Affinity matrix of the data we converted Euclidian distance into a similarity matrix, and calculated 7 nearest neighbors for each segment. Spectral clustering and eigenvalue extraction was done using kknn [71] with default settings. We used a masked strategy to assess expression size and pattern on an unannotated randomly ordered set of both enhancer classes.

Genome wide prediction:

A sliding window of 1000bp with 100bp distance was used to create segments of the entire drosophila genome, and 1000 trained on SDE and non-enhancers only with our usual sampling scheme was used to predict enhancers genome wide, with the %trees taken as score. For each 100bp segment the average of the overlapping segments was calculated, and those above the 0.75 threshold were kept. adjacent segments were merged. Segments longer then 1.5Kb were separated based on peaks in the sum of transcription factors data when possible: The normalized sum of transcription factor binding was calculated for each 100bp window, second derivative used to detect peaks, and peaks closer then 200bp merged. If more than one peak remained, the minima between adjacent peaks was used to separate the longer predicted enhancers. Once boundaries were established, the genomic prediction scheme was repeated to establish score of the entire enhancer.

REFERANCES:

1 Warren, W. C. *et al.* Genome analysis of the platypus reveals unique signatures of evolution. *Nature* **453**, 175-183, doi:10.1038/nature06936 (2008).

2 Derrien, T. *et al.* The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Res* **22**, 1775-1789, doi:10.1101/gr.132159.111 (2012).

3 Allen, B. L. & Taatjes, D. J. The Mediator complex: a central integrator of transcription. *Nat Rev Mol Cell Biol* **16**, 155-166, doi:10.1038/nrm3951 (2015).

4 Bulger, M. & Groudine, M. Functional and Mechanistic Diversity of Distal Transcription Enhancers. *Cell* **144**, 327-339, doi:10.1016/j.cell.2011.01.024 (2011).

5 Levine, M. Transcriptional enhancers in animal development and evolution. *Curr Biol* **20**, R754-763, doi:10.1016/j.cub.2010.06.070 (2010).

6 Vokes, S. A., Ji, H., Wong, W. H. & McMahon, A. P. A genome-scale analysis of the cis-regulatory circuitry underlying sonic hedgehog-mediated patterning of the mammalian limb. *Genes Dev* **22**, 2651-2663, doi:10.1101/gad.1693008 (2008).

7 Petrykowska, H. M., Vockley, C. M. & Elnitski, L. Detection and characterization of silencers and enhancer-blockers in the greater CFTR locus. *Genome Res* **18**, 1238-1246, doi:10.1101/gr.073817.107 (2008).

8 Gaszner, M. & Felsenfeld, G. Insulators: exploiting transcriptional and epigenetic mechanisms. *Nat Rev Genet* **7**, 703-713, doi:10.1038/nrg1925 (2006).

9 Lomvardas, S. *et al.* Interchromosomal interactions and olfactory receptor choice. *Cell* **126**, 403-413, doi:10.1016/j.cell.2006.06.035 (2006).

10 Geyer, P. K., Green, M. M. & Gorces, V. G. Tissue-Specific Transcriptional Enhancers May Act in Trans on the Gene Located in the Homologous Chromosome - the Molecular-Basis of Transvection in Drosophila. *Embo Journal* **9**, 2247-2256 (1990).

11 Li, X. Y. *et al.* The role of chromatin accessibility in directing the widespread, overlapping patterns of Drosophila transcription factor binding. *Genome Biology* **12**, doi:ARTN R34 10.1186/gb-2011-12-4-r34 (2011).

12 Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A* **107**, 21931-21936, doi:10.1073/pnas.1016071107 (2010).

13 Ghavi-Helm, Y. *et al.* Enhancer loops appear stable during development and are associated with paused polymerase. *Nature* **512**, 96-100, doi:10.1038/nature13417 (2014).

14 Nechaev, S. *et al.* Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in Drosophila. *Science* **327**, 335-338, doi:10.1126/science.1181421 (2010).

15 Hoffman, M. M. *et al.* Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res* **41**, 827-841, doi:10.1093/nar/gks1284 (2013).

16 Firpi, H. A., Ucar, D. & Tan, K. Discover regulatory DNA elements using chromatin signatures and artificial neural network. *Bioinformatics* **26**, 1579-1586, doi:10.1093/bioinformatics/btq248 (2010).
17 Fernandez, M. & Miranda-Saavedra, D. Genome-wide enhancer prediction from epigenetic signatures using genetic algorithm-optimized support vector machines. *Nucleic Acids Res* **40**, doi:ARTN e77
10.1093/nar/gks149 (2012).
18 Rajagopal, N. *et al.* RFECS: a random-forest based algorithm for enhancer identification from chromatin state. *PLoS Comput Biol* **9**, e1002968, doi:10.1371/journal.pcbi.1002968 (2013).
19 Dogan, N. *et al.* Occupancy by key transcription factors is a more accurate predictor of enhancer activity than histone modifications or chromatin accessibility. *Epigenetics & Chromatin* **8**, 16, doi:10.1186/s13072-015-0009-5 (2015).
20 Zinzen, R. P., Girardot, C., Gagneur, J., Braun, M. & Furlong, E. E. Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature* **462**, 65-70, doi:10.1038/nature08531 (2009).
21 Kwasnieski, J. C., Fiore, C., Chaudhari, H. G. & Cohen, B. A. High-throughput functional testing of ENCODE segmentation predictions. *Genome Res* **24**, 1595-1602, doi:10.1101/gr.173518.114 (2014).
22 White, M. A., Myers, C. A., Corbo, J. C. & Cohen, B. A. Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks. *Proceedings of the National Academy of Sciences* **110**, 11952-11957, doi:10.1073/pnas.1307449110 (2013).
23 Kellis, M. *et al.* Defining functional DNA elements in the human genome. *Proc Natl Acad Sci U S A* **111**, 6131-6138, doi:10.1073/pnas.1318948111 (2014).
24 Hilton, I. B. *et al.* Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. *Nature Biotechnology* **33**, 510-U225, doi:10.1038/nbt.3199 (2015).
25 Kaaij, L. J. *et al.* Enhancers reside in a unique epigenetic environment during early zebrafish development. *Genome Biol* **17**, 146, doi:10.1186/s13059-016-1013-1 (2016).
26 Sandmann, T. *et al.* A core transcriptional network for early mesoderm development in Drosophila melanogaster. *Genes Dev* **21**, 436-449, doi:10.1101/gad.1509007 (2007).
27 Jimenez, G., Paroush, Z. & IshHorowicz, D. Groucho acts as a corepressor for a subset of negative regulators, including Hairy and Engrailed. *Gene Dev* **11**, 3072-3082, doi:DOI 10.1101/gad.11.22.3072 (1997).
28 Liaw, G. J. *et al.* The torso response element binds GAGA and NTF-1/Elf-1, and regulates tailless by relief of repression. *Genes Dev* **9**, 3163-3176 (1995).
29 Guichet, A. *et al.* The nuclear receptor homologue Ftz-F1 and the homeodomain protein Ftz are mutually dependent cofactors. *Nature* **385**, 548-552, doi:10.1038/385548a0 (1997).
30 Pritchard, D. K. & Schubiger, G. Activation of transcription in Drosophila embryos is a gradual process mediated by the nucleocytoplasmic ratio. *Gene Dev* **10**, 1131-1142, doi:DOI 10.1101/gad.10.9.1131 (1996).

31 Harrison, M. M., Botchan, M. R. & Cline, T. W. Grainyhead and Zelda compete for binding to the promoters of the earliest-expressed Drosophila genes. *Dev Biol* **345**, 248-255, doi:10.1016/j.ydbio.2010.06.026 (2010).
32 Simpson, P. Maternal-Zygotic Gene Interactions during Formation of the Dorsoventral Pattern in Drosophila Embryos. *Genetics* **105**, 615-632 (1983).
33 Lawrence, P. A. *The making of a fly : the genetics of animal design*. (Blackwell Scientific, 1992).
34 Parkhurst, S. M. & Meneely, P. M. Sex determination and dosage compensation: lessons from flies and worms. *Science* **264**, 924-932 (1994).
35 Nusslein-Volhard, C. & Wieschaus, E. Mutations affecting segment number and polarity in Drosophila. *Nature* **287**, 795-801 (1980).
36 Mahowald, A. P. & Hardy, P. A. Genetics of Drosophila embryogenesis. *Annu Rev Genet* **19**, 149-177, doi:10.1146/annurev.ge.19.120185.001053 (1985).
37 Stjohnston, D. & Nussleinvolhard, C. The Origin of Pattern and Polarity in the Drosophila Embryo. *Cell* **68**, 201-219 (1992).
38 Breiman, L. Random forests. *Mach Learn* **45**, 5-32, doi:Doi 10.1023/A:1010933404324 (2001).
39 Biau, G. & Scornet, E. A random forest guided tour. *Test-Spain* **25**, 197-227, doi:10.1007/s11749-016-0481-7 (2016).
40 Andy Liaw, M. W. Classification and Regression by randomForest. *R News* **2**, 18-22 (2002).
41 Boulesteix, A.-L., Janitza, S., Kruppa, J. & König, I. R. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **2**, 493-507, doi:10.1002/widm.1072 (2012).
42 Fernández-Delgado, M., Cernadas, E., Barro, S. & Amorim, D. Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? *Journal of Machine Learning Research* **15**, 3133-3181, doi:citeulike-article-id:13458170 (2014).
43 Lin, Y. & Jeon, Y. Random Forests and Adaptive Nearest Neighbors. *Journal of the American Statistical Association* **101**, 578-590, doi:10.1198/016214505000001230 (2006).
44 Biau, G. & Devroye, L. On the layered nearest neighbour estimate, the bagged nearest neighbour estimate and the random forest method in regression and classification. *Journal of Multivariate Analysis* **101**, 2499-2518, doi:https://doi.org/10.1016/j.jmva.2010.06.019 (2010).
45 Scornet, E. Random Forests and Kernel Methods. *IEEE Transactions on Information Theory* **62**, 1485-1500, doi:10.1109/TIT.2016.2514489 (2016).
46 Lugosi, G. B. L. D. G. Consistency of Random Forests and Other Averaging Classifiers. *Journal of Machine Learning Research* **9**, 2015-2033 (2008).
47 Wager, S. *Asymptotic Theory for Random Forests*. (2014).
48 Biau, G. Analysis of a random forests model. *J. Mach. Learn. Res.* **13**, 1063-1095 (2012).
49 Breiman, L. CONSISTENCY FOR A SIMPLE MODEL OF RANDOM FORESTS. *Technical Report 670* (2004).

50 Scornet, E., Biau, G. & Vert, J.-P. Consistency of random forests. *Ann. Statist.* **43**, 1716-1741, doi:10.1214/15-AOS1321 (2015).
51 Louppe, G., Wehenkel, L., Sutera, A. & Geurts, P. in *Advances in neural information processing systems.* 431-439.
52 Louppe, G., Wehenkel, L., Sutera, A. & Geurts, P. in *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1* 431-439 (Curran Associates Inc., Lake Tahoe, Nevada, 2013).
53 Genuer, R., Poggi, J.-M. & Tuleau-Malot, C. Variable selection using random forests. *Pattern Recognition Letters* **31**, 2225-2236, doi:https://doi.org/10.1016/j.patrec.2010.03.014 (2010).
54 Archer, K. J. & Kimes, R. V. Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis* **52**, 2249-2260, doi:https://doi.org/10.1016/j.csda.2007.08.015 (2008).
55 Auret, L. & Aldrich, C. Empirical comparison of tree ensemble variable importance measures. *Chemometrics and Intelligent Laboratory Systems* **105**, 157-170, doi:https://doi.org/10.1016/j.chemolab.2010.12.004 (2011).
56 Kvon, E. Z. *et al.* Genome-scale functional characterization of Drosophila developmental enhancers in vivo. *Nature* **512**, 91-95, doi:10.1038/nature13395 (2014).
57 Fisher, W. W. *et al.* DNA regions bound at low occupancy by transcription factors do not drive patterned reporter gene expression in Drosophila. *Proc Natl Acad Sci U S A* **109**, 21330-21335, doi:10.1073/pnas.1209589110 (2012).
58 Campos-Ortega, J. A. & Hartenstein, V. *The Embryonic Development of Drosophila melanogaster*. (Springer Berlin Heidelberg, 2013).
59 Moses, A. M. *et al.* Large-scale turnover of functional transcription factor binding sites in Drosophila. *PLoS Comput Biol* **2**, e130, doi:10.1371/journal.pcbi.0020130 (2006).
60 MacArthur, S. *et al.* Developmental roles of 21 Drosophila transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions. *Genome Biol* **10**, R80, doi:10.1186/gb-2009-10-7-r80 (2009).
61 Li, X. Y. *et al.* Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm. *PLoS Biol* **6**, e27, doi:10.1371/journal.pbio.0060027 (2008).
62 Li, X. Y., Harrison, M. M., Villalta, J. E., Kaplan, T. & Eisen, M. B. Establishment of regions of genomic activity during the Drosophila maternal to zygotic transition. *Elife* **3**, doi:10.7554/eLife.03737 (2014).
63 Thomas, S. *et al.* Dynamic reprogramming of chromatin accessibility during Drosophila embryo development. *Genome Biol* **12**, R43, doi:10.1186/gb-2011-12-5-r43 (2011).
64 Kaplan, T. *et al.* Quantitative Models of the Mechanisms That Control Genome-Wide Patterns of Transcription Factor Binding during Early Drosophila Development. *Plos Genet* **7**, doi:ARTN e1001290
10.1371/journal.pgen.1001290 (2011).
65 Yang, Z. H. A Space-Time Process Model for the Evolution of DNA-Sequences. *Genetics* **139**, 993-1005 (1995).
66 Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**, 1034-1050, doi:DOI 10.1101/gr.3715005 (2005).

67 Felsenstein, J. & Churchill, G. A. A hidden Markov Model approach to variation among sites in rate of evolution. *Mol Biol Evol* **13**, 93-104 (1996).
68 Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* **39**, 311-318, doi:10.1038/ng1966 (2007).
69 Rada-Iglesias, A. *et al.* A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**, 279-283, doi:10.1038/nature09692 (2011).
70 Ng, A. Y., Jordan, M. I. & Weiss, Y. On spectral clustering: Analysis and an algorithm. *Adv Neur In* **14**, 849-856 (2002).
71 kknn: Weighted k-Nearest Neighbors (2016).
72 cluster: Cluster Analysis Basics and Extensions (2016).
73 Bickel, P. J. & Sarkar, P. Hypothesis testing for automated community detection in networks. *J R Stat Soc B* **78**, 253-273, doi:10.1111/rssb.12117 (2016).
74 Halfon, M. S., Gallo, S. M. & Bergman, C. M. REDfly 2.0: an integrated database of cis-regulatory modules and transcription factor binding sites in Drosophila. *Nucleic Acids Res* **36**, D594-598, doi:10.1093/nar/gkm876 (2008).
75 Gallo, S. M., Li, L., Hu, Z. & Halfon, M. S. REDfly: a Regulatory Element Database for Drosophila. *Bioinformatics* **22**, 381-383, doi:10.1093/bioinformatics/bti794 (2006).
76 Gallo, S. M. *et al.* REDfly v3.0: toward a comprehensive database of transcriptional regulatory elements in Drosophila. *Nucleic Acids Res* **39**, D118-123, doi:10.1093/nar/gkq999 (2011).
77 Pfeiffer, B. D. *et al.* Tools for neuroanatomy and neurogenetics in Drosophila. *Proc Natl Acad Sci U S A* **105**, 9715-9720, doi:10.1073/pnas.0803697105 (2008).
78 van den Brink, D. M., Banerji, O. & Tear, G. Commissureless Regulation of Axon Outgrowth across the Midline Is Independent of Rab Function. *Plos One* **8**, doi:ARTN e64427
10.1371/journal.pone.0064427 (2013).
79 Keleman, K. *et al.* Comm sorts Robo to control axon guidance at the Drosophila midline. *Cell* **110**, 415-427, doi:Doi 10.1016/S0092-8674(02)00901-7 (2002).
80 Bevington, S. L. *et al.* Inducible chromatin priming is associated with the establishment of immunological memory in T cells. *Embo Journal* **35**, 515-535 (2016).
81 Karolchik, D. *et al.* The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* **32**, D493-D496, doi:10.1093/nar/gkh103 (2004).
82 Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res* **12**, 996-1006, doi:10.1101/gr.229102 (2002).
83 *UCSC genome browser*, <http://genome.ucsc.edu/> (
84 Crosby, M. A. *et al.* FlyBase: genomes by the dozen. *Nucleic Acids Res* **35**, D486-491, doi:10.1093/nar/gkl827 (2007).
85 R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing, Vienna, Austria).
86 naiveBayes (e1071).
87 Keilwagen, J., Grosse, I. & Grau, J. Area under precision-recall curves for weighted and unweighted data. *PLoS One* **9**, e92209, doi:10.1371/journal.pone.0092209 (2014).

88 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842, doi:10.1093/bioinformatics/btq033 (2010).
89 Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* **37**, 1-13, doi:10.1093/nar/gkn923 (2009).
90 Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44-57, doi:10.1038/nprot.2008.211 (2009).