MACHINE LEARNING FOR THE STUDY OF GENE REGULATION AND COMPLEX TRAITS

By

Anne Sonnenschein

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Genetics - Doctor of Philosophy

2017

ABSTRACT

MACHINE LEARNING FOR THE STUDY OF GENE REGULATION AND COMPLEX TRAITS

By

Anne Sonnenschein

Functional elements are found in DNA outside of protein coding regions; an important class of these elements are 'enhancers', which govern when and where transcription occurs. Predicting the identity and function of potential enhancers based on DNA sequence remains a major goal of genomics. A number of features are associated with the enhancer state, but even combinations of these features in well-studied systems such as *Drosophila* have limited predictive accuracy. I have examined the current limits of computational enhancer prediction, and analyzed which features are most useful for this task, by applying machine-learning methods to an extensive set of genomic features.

Inferring the genetic underpinning of even well-characterized phenotypes is equally challenging, although similar analytical methods can be applied. Phenotypes are frequently defined based on a set of characteristic features; when images are used as specimens, these features are frequently based on morphometric landmarks, although computational pattern-recognition has been used as an alternative. I use *Drosophila* wing shape as a model for a complex phenotype, and use machine learning to predict underlying genotype using both traditional landmarks and features extracted using 'computer vision'.

This thesis is dedicated to my parents, Mark Sonnenschein and Gerry Franklin Sonnenschein, who made it possible through years of love, support, and many cups of coffee. It is also dedicated to Malvika, who didn't want me to do a PhD, but would be glad to know I'd finished it (you are missed).

ACKNOWLEDGEMENTS

This dissertation was made possible by the contributions, help and guidance of a number of mentors, friends, and committee members. First and foremost, my co-advisers Dr. David N. Arnosti and Dr. Ian Dworkin have been incredible, both as mentors and by acting as examples of exemplary scientists, who love what they do and share that passion for science with others. I have been lucky to learn from them. In the first few years of my degree Ian introduced me to programming, experimental design, and the fundamentals of genetics that I've since gone on to teach. In the last few years of my degree, David took over most of the tasks of day-to-day mentorship, and was an endless source of organization, patience career advice, and entertaining trivia. My committee members, including Dr. Barry Williams, Dr. Patricia Wittkopp, Dr. Cathy Ernst and Dr. Shin-Han Shiu have all provided a great deal of valuable guidance and encouragement both in and out of committee meetings. Dr. Terri McElhinny, although not on my committee, has been a mentor, motivator and friend in the realm of teaching focused careers.

In Lansing I have been lucky to be surrounded by co-workers and classmates who were also friends. Both the Dworkin lab and Arnosti labs (past and present) have been amazing communities in which to learn and do science. Over the last few years, Drs. Rewatee Gokhale, Sandhya Payankaulam and Yiliang Wei, as well Rima Mouawad have provided a great deal of feedback, advice and short-notice pet-sitting. I especially owe thanks to the last of the 'Michigan-Dworkinites', Dr. Will Pitchers and Amanda Charbonneau, for their friendship, support, and help with data analysis. Outside of the lab, Roshan Angoshtari, David Tack, Ann Ripberger, Laura Hurtig, Michael DeNieu and Sarah Marzac all spent a lot of after-hours time with me in coffee shops and all-night diners on weekends and evenings, with the sometimes successful goal of

getting work done. Acknowledgments are also due to my long-distance friend Dr. Conor D. Cox, for seven years of encouragement, conversation, and frequent emergency proof-reading.

Lastly, I owe thanks to my family-- my brother Matthew for his sympathy, conversation and links to funny comics, my sister Susie for putting up with my nonsense during my comprehensive exams, and my parents, Mark and Gerry Sonnenschein, for everything.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I: INTRODUCTION

Introduction

Advances in biotechnology and computing have led to an enormous expansion of available data in the life sciences[1,2]. These developments have contributed to a movement toward accessibility of results, public data sharing, and the development of new analytical tools for combining heterogeneous datasets[3,4]. Interpreting the available data has become an overarching challenge for modern biology[5]. However, a bottleneck exists at making biological inferences across disciplines; translating findings from molecular biology through to physiology and behavior in evolving populations, and vice versa. Drawing conclusions about phenotype from genotype (e.g. determining if a pathogen is resistant to a treatment based on its DNA sequence) and genotype from phenotype (e.g. which genetic variants are causing a complex disease in a patient) are equally challenging problems[6,7]. Machine learning is a common approach to addressing these questions.

Machine learning algorithms identify the features associated with a concept, and optimize or 'learn' with the addition of new data. They allow incorporation of very heterogeneous data into models, a feature that is well suited to addressing biological questions. Current tasks that profit from such an approach include combining DNA sequence and protein binding information to predict genomic features, or using different metabolic and environmental factors to identify a disease state. In the life sciences, machine learning approaches have been effectively applied across a wide range of disciplines and questions[8–10]. Machine learning can approach two related goals: identifying potentially informative trends from data by looking at the distributions of different states (sometimes referred to as generative models), or making classifications or predictions by finding the most effective way to separate subgroups within data (discriminative models)[9]. Depending on the goals of the experiment and the availability of pre-existing data,

machine learning algorithms may be 'supervised', that is, optimized on a training set, or 'un-supervised', in which they learn from the patterns that exist within the data[9,10]. Within these categories there are a variety of frequently used algorithms, best suited to different kinds of data and questions[8,11].

Since the 1980s, these methods have been extensively applied in the life sciences, in fields ranging from determining the conditions best suited for plants[12] diagnosing and predicting the development of cancer[13,14] and identifying the conditions that might be contributing to population changes in endangered species[15]. In molecular biology machine learning has been used for interpreting data from next generation sequencing[16] and extensively applied to the study of gene regulation. The development of large publicly available datasets focused on features associated with gene regulation, such as ENCODE and modENCODE, have stimulated studies that make inferences about genes that are co-regulated in gene regulatory networks[17–19]. These genome-wide data sets have also been used directly for genome annotation, identifying the location of open reading frames, splice sites, and regulatory regions such as enhancers[10]. The relatively constrained features of transcription units and exons facilitate their genomic annotation, but the more flexible nature of enhancers has made their characterization a challenging problem.

Identification and interpretation of cis-regulatory elements

In 1933, while discussing the transfer of biological information between generations, Thomas Hunt Morgan observed that it was irrelevant if genes were real or a 'purely fictitious...hypothetical unit'[20]. Either way the information was clearly there, localized in chromosomes, and passed from parent to offspring. Within thirty years, genes had been established as definitively real, embodied in nucleic acids, expressed through transcription into

messenger RNA and translation into protein. However, the sequencing of the human genome revealed that only ~2% of the information is protein coding[21]. Regulatory information, determining when, where and to what extent these protein-coding genes are expressed, is believed to occupy a much greater percentage of the genome (estimates vary from ~8% to ~80%)[22]. Changes in genomic regulatory regions are believed to be critically important to the differences between species, and evolutionary change over time. Countless studies have linked mutations in regulatory DNA sequences to disease, as well as to both small differences between species and major reorganizations in body-plan patterning between phylogenetic groups[23,24].

Many elements of gene regulation are common to all eukaryotes, and predate the existence of a common ancestor of plants and animals over a billion years ago. A number of gene families encoding transcription factors important to cell cycle regulation and multi-cellular body-planning, such as E2F and MADs-box genes, originated before this split, though they are deployed in very different ways across different phyla[25,26]. All eukaryotic genes are regulated by a core promoter coincident with the transcriptional initiation site, and generally require distally-acting enhancer elements for more than basal transcription. These enhancers are characterized by open chromatin, and physically interact with the promoter via the co-activator Mediator protein. Unlike the simpler regulatory landscape found in yeast, control of animal genes can feature a complexity of enhancer architecture, with multiple enhancers sometimes acting at great genomic distances to regulate spatial and temporal expression (reviewed in Shlyueva et al. 2014)[27].

Core promoters are comparatively well understood, representing compact regions of DNA flanking the transcription start site. Sequence elements in the core promoter permit interactions with the highly conserved general transcription factors necessary for basal transcription in eukaryotes. There are several common motifs that frequently occur at the core promoter (e.g.

TATA, Inr), though the composition and identities of these are variable between genes, and between species[28,29]. Promoters have been categorized based on a number of criteria, including the distribution of dinucleotides, CAGE peaks, histone tail modifications, general transcription factors, and the presence or absence of bi-directional transcription[29,30]. A common organization is differentiating 'dispersed' promoters (weaker transcription, multiple transcription start sites), and 'focused' promoters (stronger transcription, one or few transcription start sites). The former is functionally associated with housekeeping genes, and is more common in metazoans than in single-celled eukaryotes like yeast[30,31].

By comparison, enhancers are far more variable between and within species, and less clearly defined. Enhancers typically are stretches of DNA bound by transcription factors, and were originally differentiated from proximal-acting regulatory sequences by their ability to activate gene expression from distal locations, in either orientation. Virtually all genes from multi-cellular eukaryotes require enhancers for activation. They are incredibly diverse in terms of size[32], organization[33] and function[27,34]. Much like Morgan's 'hypothetical' genes, it is not clear what characteristics are distinct to enhancers, how their unique qualities contribute to their biological function, and what distinguishes them from background DNA.

In metazoans, enhancers are typically characterized by clusters of sequence motifs associated with the binding of specific transcription factors[21]. In the course of development, enhancers can be initially first bound by pioneer transcription factors, which recruit proteins to promote chromatin-remodeling, before recruitment of additional transcription factors and co-activators (Schulz 2015). The degree to which different transcription factors bound to the same enhancer influence each other and cooperate to regulate gene expression is highly variable. Although a common repertoire of regulatory mechanisms can be found across many pathways[35], how they

are specifically deployed within enhancers can be associated with specific regulatory pathways and functions[36].

Several models have been proposed as broad categories for enhancers based on the interactions of constituent transcription factors. The most highly structured enhancers are classified as 'enhanceosomes', with proteins that exhibit highly interdependent binding, in which single-base pair changes in the sequence lead to total loss of function. In contrast, some enhancers are more analogous to 'billboards', wherein transcription factors are more loosely organized, and largely direct gene expression independently[37]. Loss of a single transcription factor binding site might only partially alter the enhancer's function, or have no impact at all due to built-in redundancy. Such enhancers are not entirely unconstrained in their organization; they can still demonstrate cis-regulatory grammar, that is the constraints on regulatory DNA sequences as a consequence of the way that arrangements and affinities of transcription factor binding motifs generate differential outputs[37,38]. Interestingly, simulations have suggested that in some cases, such apparent constraints may be entirely an artifact of neutral sequence evolution leading to clustering of independent regulatory regions, that are adjacent but otherwise unrelated[39]. A third model proposed to describe a category of enhancer organization is the “transcription factor collective”, in which the binding of transcription factors and co-activators has a strong impact on the binding of additional proteins, which themselves may not rely on sequence motifs[40]. There are numerous biological examples that fit each of these models, but without extensive experimental manipulation it is difficult to categorize newly discovered enhancers, or make general statements about the relative frequency of enhanceosomes, billboards and collectives.

In-depth analysis of specific enhancers has also highlighted examples that do not clearly fit within previously defined models in which enhancers act as small independent modules. In

Drosophila melanogaster, a single wing gene was found to have clusters of transcription factor binding dispersed over a region spanning about 10kb; these clusters direct a single gene in unique but highly overlapping expression patterns[41]. Several of these clusters could be removed without causing any visible phenotypic changes[42]. With enhancer information distributed semi-redundantly over a large area, it is not clear whether each cluster should be viewed as a discrete enhancer, or as part of a larger regulatory 'billboard'. Similarly, the *insulin receptor* gene in *Drosophila* has complex and partially redundant regulatory information for various developmental stages extending across over forty kilobases that may represent cis-regulatory elements that necessarily act in concert rather than independently[43]. In mammalian genomes, large regulatory regions with very high levels of transcriptional output have been referred to as 'super-enhancers', and hypothesized to have functionality that is distinct from smaller regulatory elements, and a role in defining cell-type specificity. In some cases, these regions have been shown to act as a cooperative unit, in which subsections do not function independently[44]. Even for smaller enhancers, the exact borders of the functional region are not readily defined; sequences with no known protein binding have been shown to influence or modify enhancer function[45]. In some cases, large regions of AT-rich DNA may modify chromatin structure to act as a booster sequence for a single enhancer, without containing any transcription factor binding motifs[46].

Evolutionary conservation is of limited use for predicting the functional boundaries of enhancers, or categorizing different classes of enhancer types. There are a number of well-studied enhancers that have substantially diverged in sequence over evolutionary time, while the timing and location of transcriptional output are maintained[47,48]. However, there are also some cases where regulatory elements are conserved across very long phylogenetic distances[49,50]; it has been

speculated that these may represent enhancers that conform to 'enhanceosome' type architecture, where small changes in sequence are highly detrimental to function[51]. However, this type of distinction is difficult to make in *Drosophila*, which possesses a genome in which both intergenic and intronic regions exhibit a high degree of conservation[52]. Conservation of regulatory elements tends to be more common in organisms with greater biological complexity or larger genomes such as mammals, but even in humans the percentage of the genome that is predicted to have regulatory importance is much greater than the portion that is highly conserved[53]. Changes in sequence also do not consistently correspond to proportional differences in expression[54–56]. In some cases, transcription-factor binding sites for a single protein have been found to be under different levels of purifying selection throughout the genome; some sites are unchanged over long periods of time while others are subject to rapid turnover. It has been speculated that this is due to these transcription factors having different functions (with correspondingly different specificity requirements) for these sites[51].

Within populations, variation in regulatory elements is common in the form of single nucleotide polymorphisms, insertions and deletions[57]. Entire regulatory elements are sometimes present in subsets of a population[58]. Some of this variation is likely due to genetic drift; in humans, this variation likely plays a role in the inheritance of complex diseases[57]. In other cases, it may contribute to standing cryptic genetic variation, and only influence phenotype under certain conditions[59]. Many regulatory elements have a certain amount of internal redundancy, in the form of extra potential binding sites[60,61], or backup 'shadow' enhancers, which can contribute to buffering of transcriptional precision[62,63]. Even highly conserved regulatory elements that are clearly under strong purifying selection appear to have some level of redundancy; removal of classical 'enhanceosomes' does not always show a visible phenotypic effect[64].

Enhancer variation can also come from directional selection. Gene regulatory elements are a major source of adaptation, ranging from subtle shifts in phenotype[65] to rewiring of entire gene networks[66]. It is frequently suggested that changes in cis-regulatory elements are in fact the most important drivers of evolutionary change[23]. In *Drosophila*, the changes necessary to lead to pigmentation in the adult wing (which has a marked effect on *Drosophila* behavior and courtship) can be reached with modifications of a few nucleotides[24]. At the other end of the spectrum, extensive changes in regulatory interactions may be consistent with conserved function of genetic interactions. For instance, the *HOX* genes that pattern appendages include genes that have largely conserved function across all metazoans[23]. The human eye development gene *Pax6* can be substituted for its fruit fly ortholog, *eyeless*, and drive eye formation in *Drosophila*, despite the profound differences in these organs across this evolutionary distance[67–69]. While primary function is conserved, there have clearly been substantial changes in how these genes are deployed in their respective networks[68].

There are many models of regulatory evolution[70], and the relationships between the structure, function, and evolution of regulatory elements are not fully understood. These complexities make it difficult to accurately predict how changes in regulatory DNA leads to changes in function, or how to measure different types of selective pressure[71]. In several studies, the precise changes that led to phenotypic changes have been mapped out. In the case of the *Drosophila* enhancer *sparkling*, rapid sequence change over time was directly connected to cis-regulatory architecture. The enhancer produced the correct level of expression when it included several weak binding sites, as opposed to fewer strong binding sites. Weak binding sites require a less specific consensus sequence[72]. Also in *Drosophila*, an enhancer that drives expression of a gene in the optic lobe was compared between two very closely related species. Although there had

been a number of changes, the function was conserved; however, not all paths to sequence divergence were equally neutral. Many mutations in this regulatory element would have led to ectopic expression[55]. Presumably, purifying selection was playing a role despite the overall high levels of sequence turnover. It is plausible that similar mechanisms have been at work in the enhancer for the *even-skipped* gene in *Drosophila*. The enhancer that drives the 'stripe two' expression pattern has conserved function between different species of *Drosophila*, despite substantial sequence divergence. However, chimeric fusions of the two enhancers do not produce a normal expression pattern, suggesting that compensatory mutations have accompanied the turnover of binding sites[73,74].

Predicting the output of enhancers

Experimental results like those from the *even-skipped* enhancer have led to another approach for studying the rules governing enhancer function, that is, building it from first principles, either using quantitative models or synthetic biology approaches (reviewed in Ay and Arnosti 2011[75]). Modeling approaches can generally be grouped into several categories. These include boolean approaches, simple decision trees with binary outcomes, very analogous to the 'genetic switch' terminology, thermodynamic or fractional occupancy models that model that treat gene activity as a statistical scenario based on the probability of individual transcription factors associating with the available binding sites, and interacting with each other and key co-activators, and differential equations models that predict output of a gene as a function. Dynamic parameters representing time and transcription factor availability are estimated[75]. A limitation to these approaches is that we don't know all the essential inputs even for well-studied of regulatory regions[76]. Furthermore, the more biologically accurate methods are extremely computationally intensive, and require a very solid biological understanding of what proteins are involved in a

given regulon (a question which is complicated by extensive cross-binding between different regulatory networks, and the key role of co-activators that do not themselves bind to the DNA as illustrated by the 'transcription factor collective' described in Spitz and Furlong 2012[40]). It is also challenging to apply the rules governing expression for one gene to another gene or even another regulatory element, as the degree to which transcription factors interact with each other in different enhancers is highly variable[77,78]. It has been suggested that enhancers can be roughly grouped into three categories based on the biological function of the genes they regulate-- simple binary switches for housekeeping genes, multiple homotypic binding sites for genes regulated by a gradient, and complex heterotypic regulatory architecture for genes associated with cell differentiation, that are highly tissue or time specific[79]. However, experimental evidence has suggested that even ostensible housekeeping genes like that encoding the Insulin-like receptor protein have regulatory architecture that is highly complex, and not easy to categorize[43].

It would be easier to make inferences about which changes in DNA sequence will lead to changes in enhancer function if it were better understood how common different types of enhancers are throughout the genome. Are the majority of enhancers like *sparkling,* with a number of weak binding sites[72]? Are enhancers of a certain size more likely to be highly interdependent 'super-enhancers'[32] or 'enhanceosomes'[37]? Does cooperativity between different transcription factors within binding sites influence evolutionary rates within binding sites[80]? Can enhancers be described as a single group with some traits in common, or even broken down into sub-categories, or are enhancers highly individual? Regulatory elements identified through the same methods will share some features that may not be universal; some methods for experimentally identifying enhancers can lead to a false appearance of uniformity with respect to size, conservation, and chromatin state[33]. The availability of new molecular biology and

genomics tools has contributed to the protean understanding of what constitutes an enhancer[79].

Experimental identification of enhancers

Gene regulation was first worked out in bacterial and phage systems, with the lytic 'genetic switch' in phage lambda and the operons defined by Jacob and Monod in bacterial systems. Mutations in regulatory DNA produced characteristic phenotypes, such as constitutive expression of LacZ. There, the concept of regulatory effects operating in cis and trans were first worked out[21]. Regulatory mutations have also been the basis of classic genetic research in eukaryotes as well, including studies of homeotic transformations in *Drosophila* and maize kernel coloration[81]. Modern genome-wide approaches to identify regulatory alleles that impact expression in cis or trans have been widely deployed in model systems such as yeast and *Drosophila*, allowing estimation of the frequency of such regulatory alleles. At a finer scale, unique complementation has been taken advantage of in experiments designed to find the impact of cis-regulatory change on species divergence[82].

Molecular biological approaches were the first to identify enhancers as such. Reporter assays are generally considered the 'gold standard' of enhancer identification. In these, coding sequence for an easily identifiable transgenic protein (lacZ, firefly luciferase, or green fluorescent protein) along with a basal promoter is fused to a putative enhancer sequence. These are most commonly incorporated into plasmids for transfection into cultured cells or allowed to integrate into the host genome to generate transgenic organisms[83]. Sequences that enable the production of the transgenic gene or protein are likely enhancers, and the quantity or conditions under which they direct expression can provide insight into their in-vivo function. A disadvantage of this method is that it requires prior knowledge about where regulatory sequences are located, although recently high throughput methods have been employed to randomly test large fractions of the genome for

enhancer activity[84,85]. Additionally, many transgenic approaches involve the gene of interest being randomly integrated into the genome of the organism, which can produce variable outcomes depending on where it lands. The use of site-specific integrases to target specific docking sites in the target organism can avoid these position effects[86]. Another limitation is that molecular cloning is best suited to sequences under a specific size; it has been proposed that the general perception of enhancers as being between 500 and 2000 base pairs is an artifact of what sequences sizes can be amplified most efficiently using molecular cloning[33].

Genetics defines elements based on what is necessary and sufficient-- with reporters, this frequently focuses on 'sufficient'-- what minimal element can reconstitute an entire expression pattern for a gene. 'Necessary' can be tested with rescue experiments, where an organism that is null for an enhancer or for the gene itself has a copy of the enhancer or enhancer and gene re-introduced, either through standard molecular cloning or BAC recombineering[87]. A rescue construct that is both necessary and sufficient can be further dissected to see how individual binding sites or the order of binding sites contribute to the overall function, and how changes in sequence are associated with changes in phenotype[88]. More recently, in-vivo gene editing techniques like CRISPR-Cas9 have been used to alter sequences without the use of transgenics[89]. As genomic engineering becomes more efficient, it may provide a great deal of information on how very high resolution changes to regulatory elements can influence gene function.

While genetics methods target the regulatory elements for individual genes, much of enhancer identification has moved on to genomics methods, where regulatory elements are identified without knowledge of the genes they control, based on features associated with enhancer activity throughout the genome[17]. DNase hypersensitivity identifies regions of open chromatin, which is a prerequisite for active enhancers[90]. Chromatin-immunopreciptation has also been extensively

used for identifying regulatory elements, both by locating genomic regions bound by transcription factors or combinations of transcription factors[91], and by identifying histone modifications that are associated with open chromatin that has either enhancer and/or promoter activity[90]. Programs like ChromHMM[92] can synthesize data from multiple chromatin signatures to identify the genomic regions that are most likely to exhibit enhancer activity. These methods have been used by community efforts like the ENCODE and MODENCODE projects to broadly map all regulatory regions throughout the genome[18,19,93].

Although many features correlate with enhancer activity, none of them are universal or exclusive[94]. There are numerous active enhancers that are not characterized by the typical histone modifications, and chromatin immunoprecipitation shows that transcription factors bind widely throughout the genome, including in many regions that are not active enhancers[95]. Even among the 'true-positives' that can be identified through these methods, they neither provide certain insight into what their specific function may be (although the identities of bound activators and repressors can provide clues about gene regulatory networks, as in Zeitlinger et al. 2007[91]), nor indicate which specific genes the enhancers are activating[96]. A common approach is to assign putative enhancers to the nearest transcription start site, or if RNA-sequencing data is available, to the nearest active transcription start site[96]. There is some biological basis for this; active enhancers do tend to influence the activity of the nearest gene[85], although there are numerous cases where they have been found to regulate genes that are tens or hundreds of kilobases and many genes away[97,98]. Chromosome conformation capture methods (4C, 5C, HiC, Chia-Pet and others) can provide physical proof of interaction between a regulatory element and a specific promoter[99,100], although obtaining high-resolution data is not easily or economically obtained from genome-wide applications of this method.

Exhaustive testing of all genomic sequences in mass reporter assays, such as that used in STARR-seq, provides a more direct indication of enhancer locations, by using enhancers to drive their own expression. This approach has the advantage of being very direct, because the readout can provide both location and strength of the enhancer. It has the disadvantage of only working if the enhancer is modular, distance and orientation independent, and is tied to the specific core promoter used in the cloning vector. In addition, readout is limited to function in specific cultured cells, rather than the whole organism[84].

Enhancers are also predicted from DNA sequence alone. Deeply conserved DNA can indicate enhancer activity. Very rapid change can indicate directional selection, and enhancers that are associated with a phenotypic shift. A broadly applicable approach to identification of regulatory regions is the combination of phenotypic information and DNA sequence information[101,102]. In eQTL, genomic regions associated with changes in levels of gene expression can be used to pinpoint enhancers. In the related approach of GWAS, comparisons of populations can be used to identify variable regions that may be associated with variable gene expression, and may be in linkage disequilibrium with relevant enhancers (reviewed in Albert and Kruglyak 2015[103]).

Machine learning and enhancer identification

Machine learning is frequently used as a supplement to experimental methods for predicting enhancers. In addition to being a tool for predicting enhancers, it can potentially provide insight into what features best define enhancers, and potential limiting factors. This has been accomplished using both supervised methods[104], and unsupervised models (e.g. ChromHMM[92]). They have based these predictions on features ranging from motif presence to transcription factor occupancy and chromatin modifications, although approaches that incorporate multiple diverse datasets are generally the most successful[105]. In the model system *Drosophila melanogaster*,

computational enhancer identification has been studied extensively by the lab of Dr. Eileen Furlong, who have used both combinations of transcription factors, and chromatin marks as predictors for enhancers in embryonic mesoderm development[91,106]. Experimental validation of predicted enhancers demonstrated that predictions of enhancer activity were highly accurate, although determining the spatial expression of that activity was more challenging. From these methods they were able to make genome-wide predictions about the frequency and distribution of mesoderm enhancers, and the levels of enhancer redundancy[107].

In Kvon et al. 2014 the lab of Dr. Alexander Stark published a database of in-vivo reporters covering 15% of the *Drosophila* genome, and the individual spatial and temporal expression of those that showed enhancer activity. Since these were identified as reporters, with relatively consistent sizes, instead of based on histone marks or other features associated with enhancers, there are ad-hoc fewer assumptions about what defines an enhancer[85]. This data set is thus uniquely suited to serving as a training set for determining which characteristics are associated with enhancer activity. Moreover, since approximately half of the reporters they tested exhibited no expression, this study provides an unprecedented resource of 'negative results'. In Chapter 2, we use this database and combined it with a wide range of features associated with enhancer activity, to analyze how well different combinations of features could be used to train effective machine learning classifications.

A comparison of morphometrics and computer vision for interpreting phenotype from images

Enhancer prediction is arguably an effort to go from genotype to phenotype. Going from phenotype to genotype is an equally difficult problem. This is sometimes referred to as 'phenomics', as a complement to 'genomics': the study how different genotypes interact with the

environment to produce a distribution of phenotype[7]. Phenotype can be described as anything from how genes are translated into proteins, to how genes are translated into behavior. However, images are frequently used for recording phenotype across the entire spectrum. These images can then be quantitatively analyzed, and interpreted. For complex traits, the interpretation can rely on machine learning. This has been applied to a variety of situations, ranging estimating the prognosis of cancer based on images of the tumor, to automatically identifying a plant species based photographs of its leaves. Some success has even been achieved using 2D and 3D cranial-facial phenotypes to infer the genetic cause of disease[108–110], and to track the progress of disease[111].

Wing shape as a model system for morphometrics and computer vision

As with using machine learning for genome annotations, the features chosen for training a classifier have a substantial impact on the performance. Analysis of images is usually done using morphometrics methods, which identify corresponding landmarks across images, and use them for quantitative comparison. These landmarks, or analysis based on them, can then be used for classifications. Wing shape in *Drosophila* is a major model system for morophometric methods. It is typically measured using the intersections of veins as landmarks; more recent techniques generate splines between landmarks to automate quantification of both size and shape. These methods can extract information from wings to provide extremely in-depth insight into the underlying genotype and characteristics of the fly from which it came[112].

However, morphometric methods do have some inherent limitations. Even very sophisticated morphometrics reduce the phenotype to essentially one dimension (in wings) or two dimensions (if images include three dimensional shape information, as with skull shape). Moreover, morphometric landmarks are sometimes selected primarily because they are easily measured

rather than biologically relevant, especially in cases where the basis of a phenotype is not fully understood. It is very possible that there are other aspects of image data that are not visually obvious, but do provide relevant biological information. Computer vision, in which software extracts information about images without prior information about which features are relevant, has been proposed as an alternative or complement to morphometrics methods. It essentially takes feature selection from images out of the hands of the biologist. Features generated from computer vision technology have been extensively applied in facial recognition. Some preliminary studies have shown that it can also be used to classify *Drosophila* wing shape[113,114], but it has not been a subject of extensive research.

A wing shape database

Facial recognition technology has greatly improved over the last decade[115]. This is partly due to tremendous interest in the subject due to its use in technology developed for social media and national security. However, it has also been encouraged by the existence of databases providing training data, allowing researchers to determine how best to identify individuals from images across a range of conditions[116–118]. In Chapter 3, we discuss the development of a database of *Drosophila* wing images, with variation incorporated for sex and genotype. We also analyze some initial comparisons of how well existing morphometric features compare with features extracted through alternative methods can be used to identify sex and genotype.

REFERENCES

REFERENCES

1. Marx V. Biology: The big challenges of big data. *Nature*. 2013;498(7453):255-260.

2. Stephens ZD, Lee SY, Faghri F, et al. Big data: Astronomical or genomical? *PLoS Biol*. 2015;13(7):7.

3. Goecks J, Nekrutenko A, Taylor J, Team TG. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*. 2010;11.

4. Joseph H. The Open Access Movement Grows Up: Taking Stock of a Revolution. *PloS Biol*. 2013;11:10.

5. Schatz MC. Biological data sciences in genome research: Figure 1. *Genome Res*. 2015;25(10):1417-1422.

6. Dowell RD, Ryan O, Jansen A, et al. Genotype to Phenotype: A Complex Problem. *Science (80-)*. 2010;328(5977):469.

7. Houle D, Govindaraju DR, Omholt S. Phenomics: the next challenge. *Nat Rev Genet*. 2010;11(12):855-866.

8. Larrañaga P, Calvo B, Santana R, et al. Machine learning in bioinformatics. *Brief Bioinform*. 2006;7(1):86-112.

9. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. *Nat Rev Genet*. 2015;16(6):321-332.

10. Yip KY, Cheng C, Gerstein M. Machine learning and genome annotation: a match meant to be? *Genome Biol*. 2013;14(5):205.

11. Wolpert DH, MacCready WG. No Free Lunch Theorems for Optimization. *IEEE Trans Evol Comput*. 1997;1(1):67-82.

12. Ma C, Zhang HH, Wang X. Machine learning for Big Data analytics in plants. *Trends Plant Sci*. 2014;19(12):798-808.

13. Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. *Cancer Inform*. 2006;2:59-77.

14. Kourou K, Exarchos TP, Exarchos KP, Karamouzis M V., Fotiadis DI. Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J*. 2015;13:8-17.

15. Kampichler C, Wieland R, Calmé S, Weissenberger H, Arriaga-Weiss S. Classification in

conservation biology: A comparison of five machine-learning methods. *Ecol Inform*. 2010;5(6):441-450.

16. Kircher M, Stenzel U, Kelso J. Improved base calling for the Illumina Genome Analyzer using machine learning strategies. *Genome Biol*. 2009;10(8):R83.

17. Consortium TEP. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012;489(7414):57-74.

18. Marbach D, Roy S, Ay F, et al. Predictive regulatory models in *Drosophila melanogaster* by integrative inference of transcriptional networks. *Genome Res* . 2012;22(7):1334-1349.

19. The modENCODE Consortium, Roy S, Ernst J, et al. Identification of Functional Elements and Regulatory Circuits by Drosophila modENCODE. *Science (80-).* 2010;330(6012):1787 LP-1797.

20. Morgan TH. *The Nobel Prize in Physiology or Medicine 1933*. Nobelprize.org; 1933.

21. Pierce BA. *Genetics: A Conceptual Approach, 5th Edition*. NY: W. H. Freeman; 2014.

22. Chi KR. The dark side of the human genome. *Nature*. 2016;538:275-277.

23. Carroll SB. Evo-Devo and an Expanding Evolutionary Synthesis: A Genetic Theory of Morphological Evolution. *Cell*. 2008;134(1):25-36.

24. Prud'homme B, Gompel N, Rokas a, et al. Repeated morphological evolution through cis-regulatory changes in a pleiotropic gene. *Nature*. 2006;440(April):1050-1053.

25. Gramzow L, Ritz MS, Theißen G. On the origin of MADS-domain transcription factors. *Trends Genet*. 2010;26(4):149-153.

26. Nitta KR, Jolma A, Yin Y, et al. Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. Ren B, ed. *Elife*. 2015;4:e04837.

27. Shlyueva D, Stampfel G, Stark A. Transcriptional enhancers: from properties to genome-wide predictions. *Nat Rev Genet*. 2014;15(4):272-86.

28. Niemann UO, Heinrich. Identification and analysis of eukaryotic promotors: recent computational approaches. *Trends Genet*. 2001;17(2):56-60.

29. Danino YM, Even D, Ideses D, Juven-gershon T. The core promoter : At the heart of gene expression. *Biochim Biophys Acta*. 2015;1849(2014):1116-1131.

30. Rach EA, Winter DR, Benjamin AM, et al. Transcription initiation patterns indicate divergent strategies for gene regulation at the chromatin level. *PLoS Genet*. 2011;7(1):1.

31. Lenhard B, Sandelin A, Carninci P. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nat Rev Genet*. 2012;13(4):233-245.

32. Pott S, Lieb JD. What are super-enhancers? *Nat Genet*. 2015;47(1):8-12.

33. Barolo S. Shadow enhancers: Frequently asked questions about distributed cis-regulatory information and enhancer redundancy. *BioEssays*. 2012;34(2):135-141.

34. Wittkopp PJ, Kalay G. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nat Rev Genet*. 2012;13(1):59-69.

35. Barolo S, Posakony JW. Three habits of highly effective signaling pathways: principles of transcriptional control by developmental cell signaling. *Genes Dev*. 2002;16(10):1167-1181.

36. Erives A, Levine M. Coordinate enhancers share common organizational features in the Drosophila genome. *Proc Natl Acad Sci U S A*. 2004;101(11):3851-6.

37. Arnosti DN, Kulkarni MM. Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? *J Cell Biochem*. 2005;94(5):890-898.

38. Hewitt GF, Strunk BS, Margulies C, et al. Transcriptional repression by the Drosophila giant protein: cis element positioning provides an alternative means of interpreting an effector gradient. *Development*. 1999;126(6):1201-10.

39. Lusk RW, Eisen MB. Evolutionary mirages: Selection on binding site composition creates the illusion of conserved grammars in Drosophila enhancers. *PLoS Genet*. 2010;6(1):1.

40. Spitz F, Furlong EEM. Transcription factors : from enhancer binding to developmental control. *Nat Rev Genet*. 2012;13(9):613-626.

41. Yao L-C, Phin S, Cho J, Rushlow C, Arora K, Warrior R. Multiple modular promoter elements drive graded brinker expression in response to the Dpp morphogen gradient. *Development*. 2008;135(12):2183-2192.

42. Gafner L, Dalessi S, Escher E, Pyrowolakis G, Bergmann S, Basler K. Manipulating the Sensitivity of Signal-Induced Repression: Quantification and Consequences of Altered Brinker Gradients. *PLoS One*. 2013;8(8):8.

43. Wei Y, Gokhale RH, Sonnenschein A, Montgomery KM, Ingersoll A, Arnosti DN. Complex cis-regulatory landscape of the insulin receptor gene underlies the broad expression of a central signaling regulator. *Development*. 2016;143(19):3591 LP-3603.

44. Hnisz D, Shrinivas K, Young RA, Chakraborty AK, Sharp PA. A Phase Separation Model for Transcriptional Control. *Cell*. 2017;169(1):13-23.

45. Bickel RD, Kopp A, Nuzhdin S V. Composite effects of polymorphisms near multiple regulatory elements create a major-effect QTL. *PLoS Genet*. 2011;7(1):1.

46. Barrière A, Gordon KL, Ruvinsky I. Distinct Functional Constraints Partition Sequence Conservation in a *cis*-Regulatory Element. *PLoS Genet*. 2011;7(6):e1002095.

47. Hare EE, Peterson BK, Iyer VN, Meier R, Eisen MB. Sepsid even-skipped enhancers are functionally conserved in Drosophila despite lack of sequence conservation. *PLoS Genet*. 2008;4(6):6.

48. Kalay G, Wittkopp PJ. Nomadic enhancers: Tissue-specific cis-regulatory elements of yellow have divergent genomic positions among Drosophila species. *PLoS Genet*. 2010;6(11):441-450.

49. Siepel A, Bejerano G, Pedersen JS, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* . 2005;15(8):1034-1050.

50. Makunin I V., Shloma V V., Stephen SJ, Pheasant M, Belyakin SN. Comparison of ultra-conserved elements in drosophilids and vertebrates. *PLoS One*. 2013;8(12):1334-1349.

51. Rebeiz M, Castro B, Liu F, Yue F, Posakony JW. Ancestral and conserved cis-regulatory architectures in developmental control genes. *Dev Biol*. 2012;362(2):282-294.

52. Halligan DL, Eyre-Walker A, Andolfatto P, Keightley PD. Patterns of evolutionary constraints in intronic and intergenic DNA of Drosophila. *Genome Res*. 2004;14(2):273-279.

53. Bejerano G, Haussler D, Blanchette M. Into the heart of darkness: Large-scale clustering of human non-coding DNA. *Bioinformatics*. 2004;20(SUPPL. 1):i40-i48.

54. Bradley RK, Li XY, Trapnell C, et al. Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related drosophila species. *PLoS Biol*. 2010;8(3):3.

55. Glassford WJ, Rebeiz M. Assessing constraints on the path of regulatory sequence evolution. *Philos Trans R Soc Lond B Biol Sci*. 2013;368(1632):20130026.

56. Ramos AI, Barolo S. Low-affinity transcription factor binding sites shape morphogen responses and enhancer evolution. *Philos Trans R Soc Lond B Biol Sci*. 2013;368(20130):20130018.

57. Haraksingh RR, Snyder MP. Impacts of variation in the human genome on gene regulation. *J Mol Biol*. 2013;425(21):3970-3977.

58. Balhoff JP, Wray GA. Evolutionary analysis of the well characterized endo16 promoter reveals substantial variation within functional sites. *Proc Natl Acad Sci U S A*. 2005;102(24):8591-8596.

59. Gibson G, Dworkin I. Uncovering cryptic genetic variation. *Nat Rev Genet*. 2004;5(9):681-690.

60. Doniger SW, Fay JC. Frequent gain and loss of functional transcription factor binding sites. *PLoS Comput Biol 3 e99*. 2007;3(e99):5.

61. Gotea V, Visel A, Westlund JM, et al. Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. *Genome Res*. 2010;20:565-577.

62. Frankel N, Davis GK, Vargas D, Wang S, Payre F, Stern DL. Phenotypic robustness conferred by apparently redundant transcriptional enhancers. *Nature*. 2010;466(7305):490-493.

63. Perry MW, Boettiger AN, Bothma JP, Levine M. Shadow Enhancers Foster Robustness of Drosophila Gastrulation. *Curr Biol*. 2010;20(17):1562-1567.

64. Xiong N, Kang C, Raulet DH. Redundant and unique roles of two enhancer elements in the TCRgamma locus in gene regulation and gammadelta T cell development. *Immunity*. 2002;16(3):453-463.

65. Gompel N, Prud'homme B, Wittkopp PJ, Kassner VA, Carroll SB. Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in Drosophila. *Nature*. 2005;433(7025):481-487.

66. Erwin DH, Davidson EH. The evolution of hierarchical gene regulatory networks. *Nat Rev Genet*. 2009;10(2):141-148.

67. Halder G, Callaerts P, Gehring WJ. Induction of ectopic eyes by targeted expression of the eyeless gene in Drosophila. *Science*. 1995;267(5205):1788-1792.

68. Weasner B, Anderson J, Kumar JP. The Eye Specification Network in Drosophila. *Proc Indian Natl Sci Acad Part B, Biol Sci*. 2004;B70(5-6):517-530.

69. Weasner BM, Weasner B, DeYoung SM, Michaels SD, Kumar JP. Transcriptional activities of the Pax6 gene eyeless regulate tissue specificity of ectopic eye formation in Drosophila. *Dev Biol*. 2009;334(2):492-502.

70. Duque T, Samee MAH, Kazemian M, Pham HN, Brodsky MH, Sinha S. Simulations of Enhancer Evolution Provide Mechanistic Insights into Gene Regulation. *Mol Biol Evol*. 2014;31(1):184-200.

71. Yáñez-Cuna JO, Kvon EZ, Stark A. Deciphering the transcriptional cis-regulatory code. *Trends Genet*. 2013;29(1):11-22.

72. Swanson CI, Schwimmer DB, Barolo S. Rapid evolutionary rewiring of a structurally constrained eye enhancer. *Curr Biol*. 2011;21(14):1186-1196.

73. Ludwig MZ, Kreitman M. Evolutionary dynamics of the enhancer region of even-skipped in Drosophila. *Mol Biol Evol*. 1995;12(6):1002-11.

74. Ludwig MZ, Bergman C, Patel NH, Kreitman M. Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature*. 2000;403(6769):564-567.

75. Ay A, Arnosti DN. Mathematical modeling of gene expression: a guide for the perplexed biologist. *Crit Rev Biochem Mol Biol*. 2011;46(2):137-51.

76. Crocker J, Tsai A, Stern DL. A Fully Synthetic Transcriptional Platform for a Multicellular Eukaryote. *Cell Rep*. 2017;18(1):287-296.

77. Fakhouri W, Ay A, Sayal R. Deciphering a transcriptional regulatory code: modeling short-range repression in the Drosophila embryo. *Mol Syst …*. 2010;6(341):341.

78. Janssens H, Hou S, Jaeger J, et al. Quantitative and predictive model of transcriptional control of the Drosophila melanogaster even skipped gene. *Nat Genet*. 2006;38(10):1159-1165.

79. Levo M, Segal E. In pursuit of design principles of regulatory sequences. *Nat Rev Genet*. 2014;15(7):453-68.

80. Narasimhan K, Pillay S, Huang YH, et al. DNA-mediated cooperativity facilitates the co-selection of cryptic enhancer sequences by SOX2 and PAX6 transcription factors. *Nucleic Acids Res*. 2015;43(3):1513-1528.

81. Forms: CSBE. The Evolution of Gene Regulation and Morphological Diversity. *Cell*. 2000;101:577-580.

82. Wittkopp PJ, Haerum BK, Clark A. Genetic basis of regulatory variation within and between Drosophila species. *Nat Genet*. 2008;40:346-350.

83. Ipe J, Swart M, Burgess KS, Skaar TC. High-Throughput Assays to Assess the Functional Impact of Genetic Variants: A Road Towards Genomic-Driven Medicine. *Clin Transl Sci*. 2017;10(2):67-77.

84. Arnold CD, Gerlach D, Stelzer C, Boryn LM, Rath M, Stark A. Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq. *Science (80-).* 2013;339(6123):1074-1077.

85. Kvon EZ, Kazmar T, Stampfel G, et al. Genome-scale functional characterization of Drosophila developmental enhancers in vivo. *Nature*. 2014;512(7512):91-5.

86. Groth AC, Fish M, Nusse R, Calos MP. Construction of Transgenic Drosophila by Using the Site-Specific Integrase from Phage qC31. *Genetics*. 2004;166(4):1775-1782.

87. Venken KJT, He Y, Hoskins RA, Bellen HJ. Pacman: a BAC transgenic platform for targeted insertion of large DNA fragments in D. melanogaster. *Science (80-).* 2006;314(5806):1747-1751.

88. Sharan SK, Thomason LC, Kuznetsov SG, Court DL. Recombineering: a homologous recombination-based method of genetic engineering. *Nat Protoc*. 2009;4(2):206-223.

89. Fulco CP, Munschauer M, Anyoha R, et al. Systematic mapping of functional enhancer–

promoter connections with CRISPR interference. *Science (80-).* 2016;354(6313):1-8.

90. Heintzman ND, Stuart RK, Hon G, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet*. 2007;39(3):311-8.

91. Zeitlinger J, Zinzen RP, Stark A, et al. Whole-genome ChIP-chip analysis of Dorsal, Twist, and Snail suggests integration of diverse patterning processes in the Drosophila embryo. *Genes Dev*. 2007;21(4):385-390.

92. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods*. 2012;9(3):215-6.

93. Devailly G, Mantsoki A, Michoel T, Joshi A. Variable reproducibility in genome-scale public data: A case study using ENCODE ChIP sequencing resource. *FEBS Lett*. 2015;589(24):3866-3870.

94. Henikoff S, Shilatifard A. Histone modification: Cause or cog? *Trends Genet*. 2011;27(10):389-396.

95. Spivakov M. Spurious transcription factor binding: Non-functional or genetically redundant? *BioEssays*. 2014;36(8):798-806.

96. Whitaker JW, Nguyen TT, Zhu Y, Wildberg A, Wang W. Computational schemes for the prediction and annotation of enhancers from epigenomic assays. *Methods*. 2015;72(C):86-94.

97. Claussnitzer M, Dankel SN, Kim K-H, et al. FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. *N Engl J Med*. 2015;373(10):895-907.

98. Williamson I, Lettice LA, Hill RE, Bickmore WA. Shh and ZRS enhancer colocalisation is specific to the zone of polarising activity. *Development*. 2016;143(16):2994-3001.

99. Barutcu AR, Fritz AJ, Sayyed KZ, et al. C-ing the genome: A compendium of chromosome conformation capture methods to study higher-order chromatin organization. *J Cell Physiol*. 2015;231(1):1097-4652.

100. Yao L, Berman BP, Farnham PJ. Demystifying the secret mission of enhancers: linking distal regulatory elements to target genes. *Crit Rev Biochem Mol Biol*. 2015;50:550-573.

101. Gittelman RM, Hun E, Ay F, et al. Comprehensive identification and analysis of human accelerated regulatory DNA. *Genome Res*. 2015;25(9):1245-1255.

102. Visel A, Bristow J, Pennacchio LA. Enhancer identification through comparative genomics. *Semin Cell Dev Biol*. 2007;18(1):140-152.

103. Albert FW, Kruglyak L. The role of regulatory variation in complex traits and disease. *Nat Rev Genet*. 2015;16(4):197-212.

104. Narlikar L, Ovcharenko I. Identifying regulatory elements in eukaryotic genomes. *Briefings Funct Genomics Proteomics*. 2009;8(4):215-230.

105. Erwin GD, Oksenberg N, Truty RM, et al. Integrating Diverse Datasets Improves Developmental Enhancer Prediction. *PLoS Comput Biol*. 2014;10(6):6.

106. Wilczynski B, Liu YH, Yeo ZX, Furlong EEM. Predicting Spatial and Temporal Gene Expression Using an Integrative Model of Transcription Factor Occupancy and Chromatin State. *PLoS Comput Biol*. 2012;8:12.

107. Cannavo E, Khoueiry P, Garfield DA, Geeleher P, Zichner T, et al. Shadow enhancers are pervasive features of developmental regulatory networks. *Curr Biol*. 2016;26:1-14.

108. Allanson JE, Bohring A, Dorr HG, Dufke A, Gillessen-Kaesbach G, Horn D, Konig R, Kratz CP, Kutsche K, Pauli S, et al.. The Face of Noonan Syndrome: Does Phenotype Predict Genotype. *Am J Med Genet A*. 2010;152:1960-1966.

109. Boehringer S, van der Lijn F, Liu F, et al. Genetic determination of human facial morphology: links between cleft-lips and normal variation. *Eur J Hum Genet*. 2011;19(11):1192-1197.

110. Ferry Q, Steinberg J, Webber C, et al. Diagnostically relevant facial gestalt information from ordinary photos. *Elife*. 2014;3:e02020.

111. Zou L, Adegun OK, Willis A, Fortune F. Facial biometrics of peri-oral changes in Crohn's disease. *Lasers Med Sci*. 2014;29(3):869-874.

112. Zelditch M. *Geometric Morphometrics for Biologists : A Primer*. Elsevier Academic Press; 2004.

113. Ahmad F, Roy K, O'Connor B, Shelton J, Dozier G, Dworkin I. Fly Wing Biometrics Using Modified Local Binary Pattern, SVMs and Random Forest. *Int J Mach Learn Comput*. 2014;4(3):279-285.

114. Payne M, Turner J, Shelton J, et al. Fly wing biometrics. *IEEE Work Comput Intell Biometrics Identity Manag CIBIM*. 2013:42-46.

115. Sagonas C, Antonakos E, Tzimiropoulos G, Zafeiriou S, Pantic M. 300 Faces In-The-Wild Challenge: database and results. *Image Vis Comput*. 2016;47:3-18.

116. Steve IK, Daniel S, Dec C V. The MegaFace Benchmark : 1 Million Faces for Recognition at Scale. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. ; 2016:4873--4882.

117. Shelton J, Bryant K, Abrams S, et al. Genetic & Evolutionary Biometric Security: Disposable Feature Extractors for Mitigating Biometrics Replay Attacks. *Procedia Comput Sci*. 2012;8:351-360.

118. Schroff F, Kalenichenko D, Philbin J. FaceNet: A Unified Embedding for Face Recognition and Clustering. In: *The IEEE Conference on Computer Vision and Pattern Recognition*. ; 2015:815-823.

CHAPTER II: THE SHEEP AND THE GOATS: DISTINGUISHING TRANSCRIPTIONAL ENHANCERS IN A COMPLEX CHROMATIN LANDSCAPE

Abstract

Predicting regulatory function of non-coding DNA using genomic information remains a major goal in genomics, and an important step in interpreting cis-regulatory code. Regulatory capacity can be partially inferred from transcription factor occupancy, histone modifications, motif enrichment, and evolutionary conservation. However, even combinations of these features in well-studied systems such as *Drosophila* have limited predictive accuracy. Here we examine the current limits of computational enhancer prediction by applying machine-learning methods to an extensive set of genomic features, validating predictions with the Fly Enhancer Resource, which characterized the transcriptional activity of approximately 15 percent of the genome. Supervised machine learning trained on a range of genomic features identify active elements with a high degree of accuracy, but are less successful at distinguishing tissue-specific expression patterns. Consistent with previous observations of their widespread genomic interactions, many transcription factors were associated with enhancers not known to be direct functional targets. Interestingly, no single factor was necessary for enhancer identification, although binding by the 'pioneer' transcription factor Zelda was the most predictive feature for enhancer activity. Using an increasing number of predictive features improved classification with diminishing returns. Thus, additional single-timepoint ChIP data may have only marginal utility for discerning true regulatory regions. On the other hand, spatially- and temporally-differentiated genomic features may provide more power for this type of computational enhancer identification. Inclusion of new types of information distinct from current chromatin-immunoprecipitation data may enable more precise identification of enhancers, and further insight into the features that distinguish their biological functions.

Introduction

Enhancers, cis-regulatory elements that coordinate the input of transcriptional activators and repressors, are the primary determinants of eukaryotic gene expression. Enhancers can regulate genes from distances of hundreds to hundreds of thousands of base pairs, and genes are frequently regulated by multiple enhancers to provide complex spatial and temporal regulation[1–3]. Changes in enhancers play pivotal roles in the evolution of multi-cellular life [4–6], and disruption in enhancer function has been linked in many cases to disease[3,7–9].

Identifying the complete repertoire of enhancer(s) for a specific gene is difficult, although there are a number of features that can provide insight. Enhancers are frequently characterized by enrichment of binding motifs for various transcription factors[10–12]. This may rely on prior knowledge of binding affinities for specific factors and curated position weight matrices (PWMs)[13,14], or by searching for over-represented motif clusters in sequences adjacent to co-expressed genes[15,16]. However the presence of motifs does not indicate when an enhancer would be active in development, and motifs are not a perfect indicator of transcription factor binding[17]. Genome-wide studies have used features such as chromatin immunoprecipitation data for transcription factor occupancy and histone modifications to infer function of non-coding DNA; however, features associated with enhancers are also found in regions of open chromatin that are not necessarily active[18–20]. Many enhancers will have features of regulatory DNA even while inactive[21] and transcription factors and polymerases will frequently bind in preparation for future activity[22].

Going beyond simple predictions of overall activity, several studies have attempted to predict spatial and temporal aspects of enhancer function. These have had some success at classifying enhancers into broad categories[23–25], and making inferences about the frequency and distribution

of specific types of enhancers (e.g. 'shadow enhancers')[25]. However, results from thermodynamic models used to predict specific expression patterns based on transcription factor binding have shown that achieving that level of predictive specificity is a challenging problem[26,27].

After identifying a prospective enhancer, an additional challenge is determining which gene or genes it regulates[22]. Enhancers can be very distal to the target gene, which may not be the adjacent transcription unit [28–30] and a single enhancer can influence multiple genes[25,28,31]. Most computational approaches do rely on proximity, as a substantial percentage of enhancers control the nearest gene[32–34]. However, since many enhancers direct distal genes, this heuristic is far from universally applicable[35].

Chromatin conformation data can provide more direct information about which genes are regulated by enhancers[6,34,36]. Many of these experiments have been conducted in cell culture, but increasingly datasets from specific organ systems are also becoming available[37]. However, identifying tissue specific enhancers from conformation data is complicated by the large number of non-transcription interactions between genomic regions[38]. Combinatorial protein binding can also provide insight into enhancer locations; clusters of transcription factor binding can indicate enhancer activity. Genes known to be involved in a developmental network are likely to be associated with enhancers bound by common transcription factors regulating that network[39,40].

In *Drosophila melanogaster* enhancer identification and characterization has traditionally relied on genetic analysis[41] and in-vivo reporter assays[1]. However, even in organisms with relatively small genomes like *Drosophila*, this is a daunting task on a genome-wide scale[42]. High-throughput methods that assay the entire genome, such as STARR-seq, offer a more comprehensive perspective, but have thus far been confined to analysis of potential enhancers in cultured cells, and only provide insight on transcriptional activity, not function in control of

specific genes. Most of these assays are also limited to showing enhancer readout on specific basal promoters (which can have a substantial impact on enhancer responsiveness), and neglect the influence of genomic neighborhood[43–45]. Evolutionary conservation has been successfully used to identify *Drosophila* enhancers where functional regions of non-coding DNA are more highly conserved than background[46–48], although this method is complicated by the relatively uniform conservation across most upstream and intronic regions[49]. Furthermore, in many cases enhancers have conserved function despite sequences being highly diverged[50], and non-coding regions with functions unrelated to enhancer activity can also be conserved above background[47,51,52].

Ultimately, no one genomic feature is an exclusive or universal indicator of enhancer location or activity[42]. The increasing availability of genomics datasets for enhancer correlates, such as nucleosome occupancy and transcription factor binding[53], as well as spatial information from techniques like HiC[54] and ChIA-Pet[55], has opened up the door for compound approaches wherein multiple features are used to infer the activity of enhancers[56,57]. Machine learning approaches combining multiple genomics datasets have employed a broad range of supervised and unsupervised methods, as well as different combinations of enhancer features[58]. Some studies have also successfully assigned enhancers to genes using a probabilistic approach, where both proximity and biological information are considered[59,60]. Taken together, these studies suggest that compound approaches, using multiple features and diverse types of features as predictors are the most successful[60–62]. However, distinguishing between functional binding from background represents a major challenge to using genome-wide datasets to predict enhancer activity, and the accuracy of enhancer predictions is not entirely known. Many efforts at estimating the accuracy of enhancer predictions have been limited to testing a handful of representative cases[63], or

correlations in place of broad-scale verification. Enhancers are identified based on correlating features (e.g. motifs, transcription factor occupancy), and the accuracy of these calls are estimated from different correlates (e.g. DNase hypersensitivity[61,64], and eRNAs[65]). Until recently, more direct methods for estimating accuracy have not been available[62,66].

The recently published Fly Enhancer Resource is a database of information for nearly 8000 *Drosophila* genomic regions, showing the activity of in-vivo reporters with these regions at various developmental stages. It is also an ideal resource for determining both the predictive power of different features associated with enhancers. Importantly, the database provides detailed information about active enhancers through development, as well as many genomic regions that are not transcriptionally active. This data can be used as a training set for distinguishing active enhancers from inactive regions, and determining which features are most effective for distinguishing these groups. Here, we exploit the possibilities for validation provided by the extensive functional survey by Kvon et al. 2014[67], and use extant data on transcription factor occupancy, chromatin marks, and DNA sequence to push the limits in enhancer identification and classification, and explore what conditions could best be used to distinguish this state.

Materials and Methods

Datasets used as predictive features

Predictive features were obtained from a variety of publicly available datasets (see Table 2-1). For each dataset, information that overlapped with the coordinates of DNA elements were categorized as features of those DNA elements. These include conservation information between *D. melanogaster* and other species in the Drosophila genus, motif scores based on position weight matrices (PWMs) associated with transcription factors involved in dorsal-ventral embryonic patterning, and chromatin immunopreciptation (ChIP) data for a number of

transcription factors and histone modifications. ChIP data all came from stage 4-6 embryos; we chose to focus on early embryonic development, as the regulatory networks governing this stage of development are extremely well understood, and many genes are expressed in simple, easily categorized expression patterns. This stage also has the fewest active genes and the smallest number of active DNA elements (as indicated in Kvon et al. 2014[67]), which simplifies the task of assigning enhancers to genes.

Pairwise alignments were downloaded in axt file format from the University of California Santa-Cruz Genome Browser[68] (http://hgdownload.soe.ucsc.edu/downloads.html). Alignments used were *D. melanogaster* version dm3 to *D. grimshawi* droGri2, *D. willistoni* droWil1, *D. ananassae* droAna3, *D. pseudoobscura* dp4, *D. erecta* droEre2, *D. yakuba* droYak2, *D. sechellia* droSec1, and *D. simulans* droSim1. These were chosen to reflect a range of phylogenetic distances from *D. melanogaster.* Summary lines from each chromosome axt file were used to create a single summary file for each genome. A custom python script was used to determine the average BLASTZ score per 100 base pairs for a region, which was plotted by species on a log10 scale. Perfect conservation over 100 base pairs would yield a score near 10,000. Scripts and intermediate files are available on github at https://github.com/asonnens/crm_analysis. BLASTZ scores were used for pairwise comparisons[69].

Position probability matrices for the transcription factors Dorsal, Snail, Twist and Zelda were obtained from the database Fly Factor Survey[70]. These values were formatted using a custom python script to enable them to be input into the MEME Suite program MAST[71]. Motif matches throughout the genome with a p-value of less than 0.0001 were obtained using MAST, and the release 5.37 version of the *Drosophila melanogaster* genome. Output files, containing coordinates of qualifying matches and their scores, were converted to bedgraph forma

Table 2-1 Features for enhancer classification

ID	Short ID	Features	Reference
Dorsal 2015 ChIP-seq	Dl 15 s	1748	Sun et al. 2015
Dorsal 2009 ChIP-chip	Dl 09 c	9357	Macarthur et al. 2009
Snail 2014 ChIP-chip	Sna 14 c	7735	Rembold et al. 2014
Snail 2009 ChIP-chip	Sna 09 c	596, 2800	Macarthur et al. 2009
Twist 2014 ChIP-chip	Twi 14 c	8629	Rembold et al. 2014
Twist 2011 ChIP-seq	Twi 11 s	8797	He et al. 2011
Twist 2009 ChIP-chip	Twi 09 c	6684, 7414	Macarthur et al. 2009
Bicoid 2013 ChIP-seq	Bcd 13 s	2061	Paris et al. 2013
Bicoid 2009 ChIP-chip	Bcd 09 c	619, 702	Macarthur et al. 2009
Caudal 2010 ChIP-seq	Cad 10 s	4045	Bradley et al. 2010
Caudal 2009 ChIP-chip	Cad 09 c	1590	Macarthur et al. 2009
Hunchback 2013 ChIP-seq	Hb 13 s	4986	Paris et al. 2013
Hunchback 2009 ChIP-chip	Hb 09 c	1831, 1717	Macarthur et al. 2009
Giant 2013 ChIP-seq	Gt 13 s	4194	Paris et al. 2013
Giant 2009 ChIP-chip	Gt 09 c	1069	Macarthur et al. 2009
Hairy 2009 ChIP-chip	Hry 09 c	1704, 2729	Macarthur et al. 2009
Knirps 2010 ChIP-seq	Kni 10 s	505	Bradley et al. 2010
Kruppel 2013 ChIP-seq	Kr 13 s	5309	Paris et al. 2013
Zelda 2011 ChIP-seq	Zld 11 s	9432	Harrison et al. 2011
H3K27ac 2015 ChIP-seq	H3K27ac 15	3055	Kok et al. 2015
H3K27ac 2010 ChIP-seq	H3K27ac 10	3658	Roy 2010
H3K4me1 2015 ChIP-seq	H3K4me1 15	1934	Kok et al. 2015
p300 2010 ChIP-seq	p300 10	3296	Roy et al. 2010
Zelda motif sanger	Zld m1	17179	Fly Factor Survey
Zelda motif solexa	Zld m2	19271	Fly Factor Survey
Dorsal motif FlyReg	Dl m1	71087	Fly Factor Survey
Dorsal motif NBT	Dl m2	26159	Fly Factor Survey
Snail motif FlyReg	Sna m1	38221	Fly Factor Survey
Snail motif sanger	Sna m2	31059	Fly Factor Survey
Snail motif solexa	Sna m3	42341	Fly Factor Survey
Twist motif da	Twi m1	29656	Fly Factor Survey
Twist motif FlyReg	Twi m2	61676	Fly Factor Survey
UCSC BlastZ scores	NA	NA	Rosenbloom et al. 2015

To obtain insight into factors driving enhancer activity, we used ChIP-chip and ChIP-seq data. Peak scores were obtained from public databases including ModEncode, the Berkeley *Drosophila* Genome Project, and other publications with relevant ChIP data (see Table 2-1 in Methods). In each case terminal data files were converted to bedgraphs. When necessary,

coordinates were converted to the release 5 genome using the Flybase coordinate conversion tool. As these datasets were analyzed using different pipelines, they are somewhat variable regarding distributions of scores, and thresholds for what was considered a significant peak. Intermediate files are available on github at https://github.com/asonnens/crm_analysis. The overlap (of at least one nucleotide) between peaks called by different datasets, and correlation between scores in overlapping peaks, was analyzed using a custom python script, and visualized using gplots (version 3.0.1) in R (version 3.3.2).

Organization and curation of data from Fly Enhancer Resource

DNA elements from the Stark Lab Fly Enhancer Resource were downloaded from the Kvon et al. 2014 supplementary tables. For comparisons made in stage 4-6 embryos, we used regions that were active in stage 4-6 embryos, filtered to include only those with activity scores of 3 or higher (as annotated by Kvon et al. 2014), and confirmed (by Kvon et al. 2014) inactive regions in stage 4-6. For examination of expression patterns, the highly active stage 4-6 DNA elements were re-annotated based on photographs of the slides, to identify those that exclusively fit characteristic Anterior/Posterior and Dorsal/Ventral expression patterns. 114 DNA elements were found to have strong expression distinctly in the Anterior region of the embryo, while 78 were only expressed in the Central or Posterior regions. 22 DNA elements were identified that had stereotypical mesodermal or neurogenic ectoderm expression patterns (as in enhancers for *snail* or *short gastrulation* respectively), and 192 had strong expression patterns that did not overlap with these two regions. All active and inactive DNA elements that Kvon et al. 2014 identified as 'verified' were used for classifications based on temporal specificity (e.g. ubiquitously active, or stage specific). DNA elements were grouped by whether they were active in all embryonic stages, only in stages 4-8, stages 9-12, stages 13-16 (Figure 2-13).

Intersection of features with Fly Enhancer Resource

To determine if these features could be used to distinguish between spurious binding and functional binding, we compared the distribution of peak scores for features that overlapped (by at least one base pair) with genomic regions that drive strong expression in stage 4-6 embryos (as reported by the Fly Enhancer Resource) and regions that do not drive expression in stage 4-6 embryos. Every DNA element tested by the Fly Enhancer Resource was assigned a score for each feature included. Scores from bedgraph items (peak scores in the case of ChIP datasets, MAST motif scores in the case of motif datasets, and BLASTZ scores for conservation information) that overlapped with a DNA element were assigned to form that DNA element's scores (multiple peaks within an element were combined; scores were only combined within a single dataset, thus normalization was not necessary between datasets). If a DNA element did not overlap with any features for a given dataset, its score for that feature was set to a numerical value one standard-deviation lower than the lowest score that indicated a hit for that dataset, based on the distribution of values for that feature. We calculated fractional overlap of these genomic features, both to determine whether individual features were reproducibly identified in different studies, and to see if there was overlap of features known to be functionally related. Each dataset was compared pair-wise with every other dataset; the degree of feature overlap was calculated as a percentage (the percentage of the features in the smaller dataset that overlap with one or more features in the larger dataset). In Fig. 2-11, the Fly Enhancer Resource DNA elements were filtered to only include those that fit a minimal threshold for a potential active enhancer, based on binding of transcription factors from at least one of the datasets described in Table 2-1. Subsets of this set of potential enhancers were created to allow comparisons between active DNA elements and inactive regions that resemble active DNA elements with regards to

transcription factor occupancy. This was done by restricting the data to include only elements bound by two or more transcription factors, or only DNA elements occupied by three or more transcription factors (see Results.) All sets of DNA elements were under-sampled from the larger category to balance the number of active and inactive DNA elements included in these prospective training sets.

The DNA elements with features assigned were scaled and visualized using a Principal Components Analysis using the stats package (version 3.2.3) in R (version 3.3.2).The percentage of variation captured is displayed in a scree plot (Figure 2-b).

Discrimination between classes of enhancers with Random Forest

To determine if it was possible to distinguish between regions that are transcriptionally active and regions that are soon-to-be active, we also classified the Fly Enhancer Resource DNA elements into several categories of activity. This included ubiquitous activity (active in every developmental stage measured Kvon et al.), total inactivity (never active in any measured stage), and stage specific regions that are active only in early, middle, or later embryonic stages. To facilitate comparisons of different sorts of data, values for ChIP, conservation, and motif scores were normalized using the 'scale' function in R. In Fig. 2-9, highly active (defined by a score of 3 or higher in stage 4-6) and inactive DNA elements were classified using a random forest algorithm (randomForest package, version 4.6-12) with 500 trees in R (version 3.2.3), using two thirds of DNA elements for training and one third for testing. Receiver operating characteristic (ROC), Precision Recall and the area under these curves was analyzed using ROCR package in R (version 1.0-7, Sing et al. 2005). In this and later figures, the DNA elements used for training were randomly re-sampled ten times. In Figure 2-11, 2-12 and Figure 2-16 where balanced datasets were used equal numbers of active and inactive elements were randomly selected using

the package dplyr (version 0.5.0). For subsets filtered for DNA elements containing the indicated ChIP binding protein or chromatin mark, the number of "active" elements ranged from ~100-300, and for balanced datasets, the same number of inactive elements was used. The same classification was attempted with datasets that were reduced to only include inactive regions that at least superficially resembled active enhancers (regions occupied by two or more transcription factors, or regions occupied by a combination of factors associated with development).

Feature importance analysis

Feature importance was measured using the mean decrease in Gini index within the randomForest function, which estimates variable importance during the training of the random forest. To determine if redundancy or correlation between features influenced importance scores, features were iteratively left out of analyses, so after each set of predictions, the most important or least important feature was dropped. In Fig. 12 where successive features are omitted from the analysis, the most and least important features were re-calculated after each run with subsets of the features. After exclusion of each feature, the order of feature importance frequently changed, as the removal of partially correlated data increased the relative importance of remaining features.

Identifying potential enhancers

In Fig. 10, to comprehensively assess regions that contain bound transcription factors and chromatin signatures associated with enhancers, we generated a list of potential enhancers *de-novo* based on clustering of transcription factors adjacent to specific genes of interest, which were selected based on annotated anterior/posterior, mesodermal, or neurogenic ectodermal expression patterns. Prospective enhancers were defined by clusters of overlapping features occurring within 50kb windows centered on the +1 position of genes of interest, using a custom

python script available on github at https://github.com/asonnens/crm_analysis. Genes of interest *brinker (brk), ventral nervous system defective (vnd), short gastrulation (sog), snail (sna), even-skipped (eve), Kruppel (Kr), hunchback (hb),* and *knirps (kni),* were selected based on their well studied regulation in early stage embryos. To locate clusters, we defined regions containing at minimum Zelda and one other ChIP signal (p300, H3K4me, H3K27ac, or any TF), or H3K4me1 and one other ChIP signal (as indicated in Fig. 10). The final enhancer was defined as a region around the center of the cluster or clusters of features, and set to a fixed size of 500-2000 bp. These regions were then classified as 'putative enhancers' or 'background binding' by random forest algorithms trained on regions from the Fly Enhancer Resource.

Validation of predictions

Annotated enhancers were defined based on verified regulatory regions in the RedFly database[72]. Each random forest trained on the Fly Enhancer Resource was used to classify predicted enhancers around genes of interest ten times, using different samples of the training set. A custom python script was used to determine the frequency that each model classified a cluster that overlapped with an annotated enhancer as active (defined as percentage overlap), and the frequency that each model classified non-overlapping segments as active (likely off target). The results of this were graphed with ggplot2.

Data Availability

Scripts and final files are available on github at https://github.com/asonnens/crm_analysis.

Results

Enhancers are not highly conserved

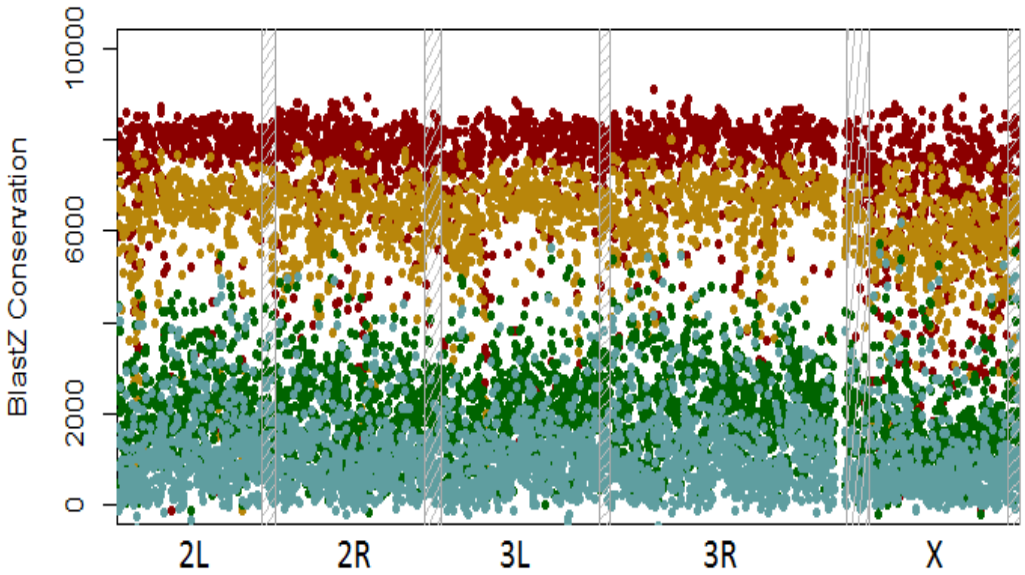
Sequence similarity in related species has been extensively used to identify functional non-coding DNA. Highly conserved regions may indicate purifying selection on regulatory elements[69], and conversely highly diverged sequences can be a signature of enhancers associated

with phenotype differences between species[73]. We looked at sequence conservation between *Drosophila melanogaster* and a range of other species in the Drosophila genus for DNA elements that were either active in every stage of embryonic development or inactive across all stages, to see if there were any visible trends. DNA elements do not show increased sequence conservation over inactive genomic regions as measured by BlastZ scores (Figure 2-1). This was true for all species comparisons.

Correlations of distinct chromatin immunoprecipitation data

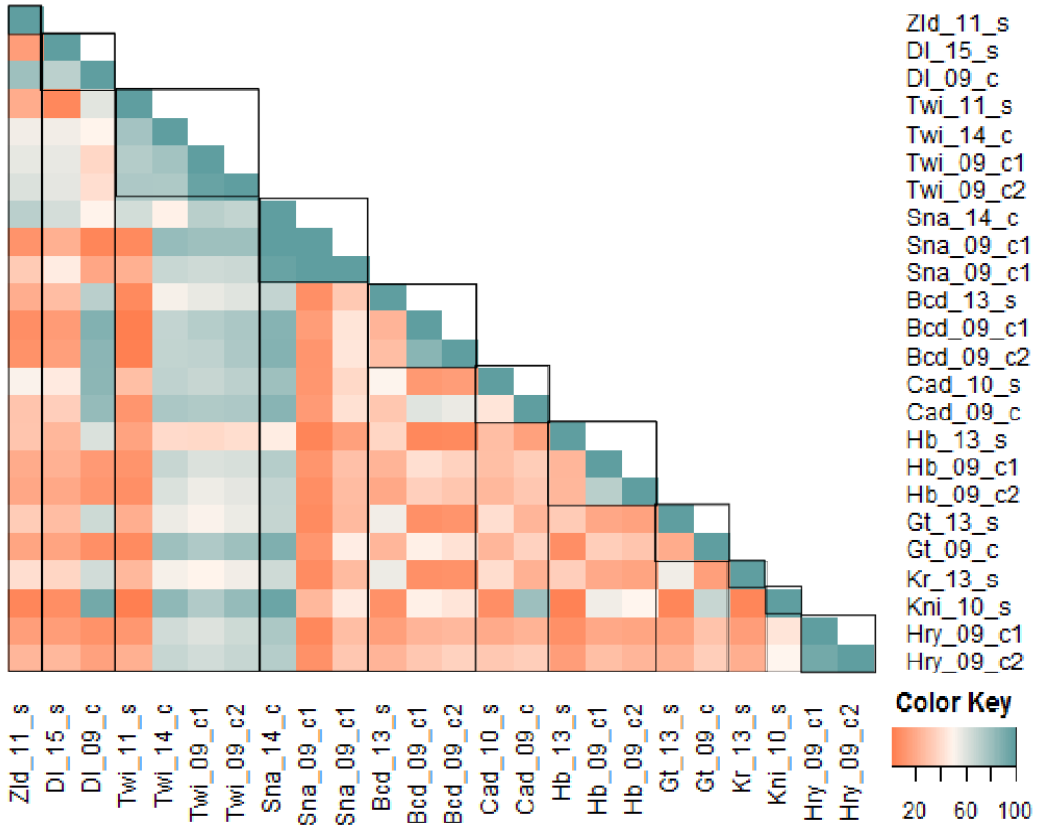
We collected a range of ChIP datasets to use for studying marks of enhancers. These datasets were produced using different methods (ChIP-chip vs. ChIP-seq) and a range of experimental conditions. Accordingly, the number of features contained within different datasets was highly variable (Table 2-1). We assessed the consistency and overlap of each dataset to determine how to most effectively combine them. Different datasets measuring binding of the same transcription factor typically do not completely overlap, but do to a greater degree than is seen in comparisons between different transcription factors (Figure 2-2, Figure 2-3).

Figure 2-1 Sequence conservation in active and inactive elements



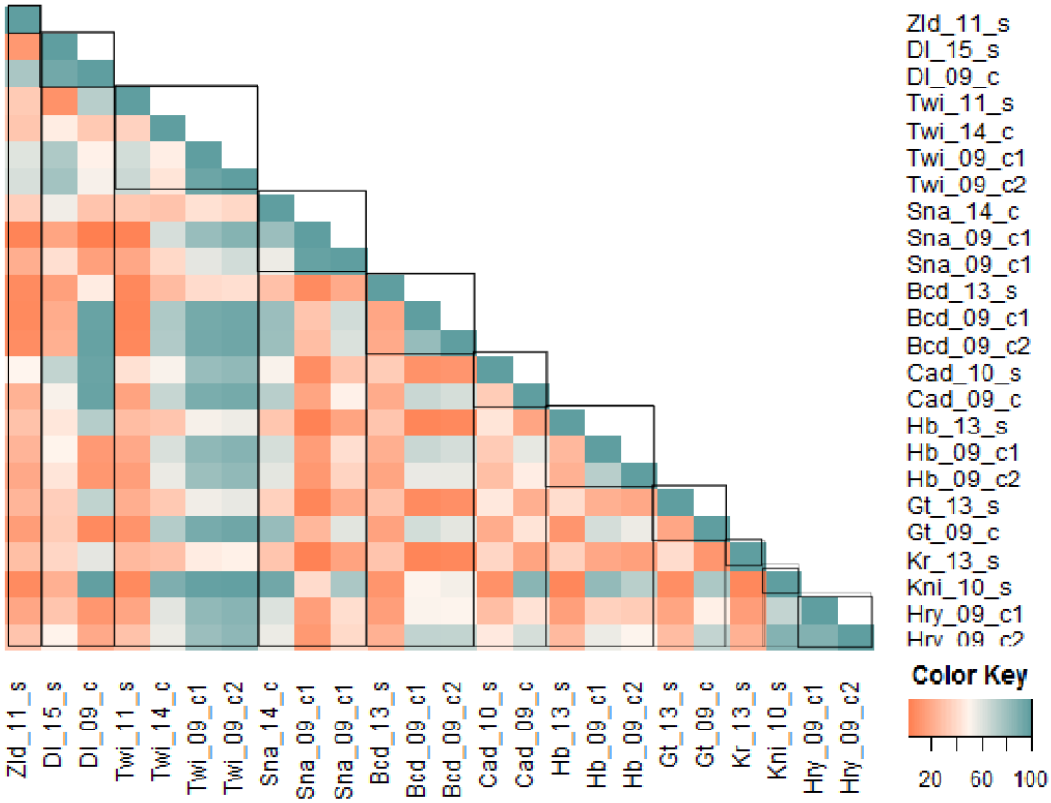
We compared the sequence conservation for reporters that drove expression in every stage of embryonic development relative to reporters that showed no activity at any stage. Levels of sequence conservation for these ~3,800 tested reporter elements from *D. melanogaster* versus *D.simulans* (red), *D.yakuba* (yellow), *D.pseudoobscura* (green), or *D.grimshawi* (blue), are averaged per 100bp, across the 2kb reporter elements.The average conservation scores closely match evolutionary distances-- sequences are consistently more similar to *Drosophila melanogaster* in more closely related species regardless of activity; but reporters with strong activity (shown in hatched areas) are not distinguishable from reporters that are inactive throughout embryonic development (rest of points, aligned by chromosomal location). BlastZ score of ~10,000 would indicate perfect sequence identity over 100 base pairs.

Figure 2-2 Correlations between genomic regulatory features



Similarities between ChIP datasets were assessed by determining percentage overlap of ChIP peaks between different datasets, ranging from full overlap (blue) to no overlap (red). Boxes on the diagonal highlight independent measurements of specific transcription factors; Twist occupancy shows a high degree of correlation, whereas Hunchback is lower. Three Twist datasets and one Snail dataset showed greater degrees of cross-correlation to other factors; lower levels of correlation were noted for complete ChIP peak datasets. In the reported bound regions with peak scores one standard deviation above average, overlap averages at 57 percent in datasets measuring the same transcription factor as reported by different labs, and 38 percent between different transcription factors.

Figure 2-3 Overlap of features between different datasets

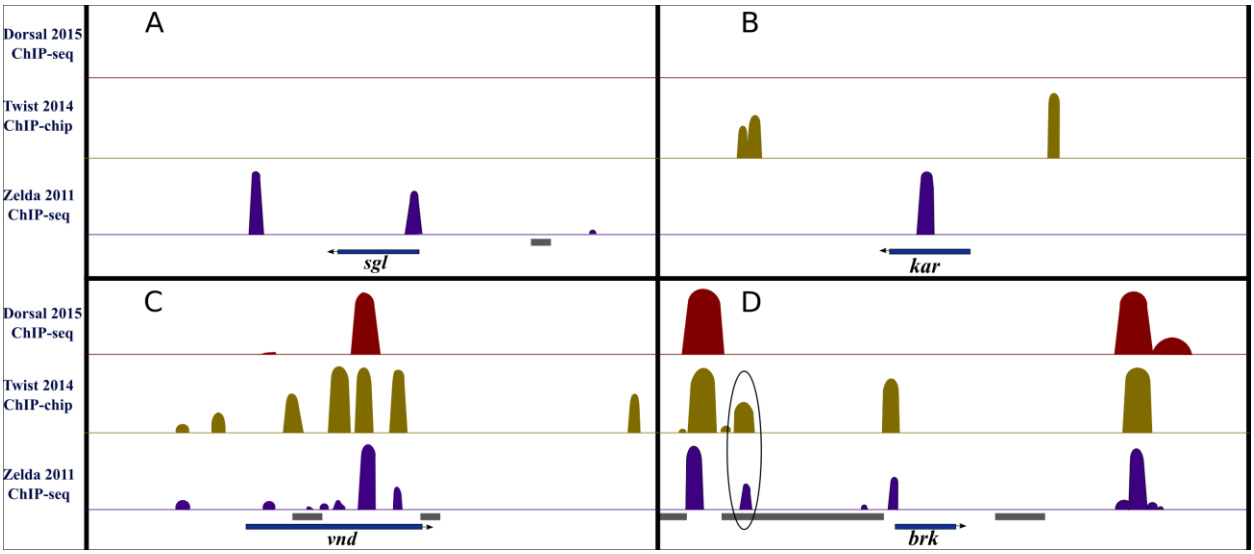


In many cases less than half of the peaks reported for a given transcription factor are reproducible between labs. This is more evident when examining all reported peaks, instead of the highest-scoring peaks (Figure 2-2). When all reported peaks are included, measurements of the same transcription factor between labs overlap on average 52 percent, compared with 49 percent overlap between different transcription factors.

Separate datasets produced by the same lab (either at different times, or using different antibodies for the same protein- e.g. Twist, Bicoid, and Hunchback) overlapped more than datasets produced by different research groups, although excluding lower scoring peaks increased this trend (Fig. 2-2 vs. Fig. 2-3). MacArthur et al. (2009)[74] reported a 91% overlap in ChIP results using different antibodies for the same transcription factor. In comparison, the average peak overlap for ChIP results of specific transcription factors from different laboratories ranges from 52% for all peaks to 57% for highest scoring peaks (Fig. 2-2 and Fig. 2-3). By contrast, overlap between different transcription factors ranged from 49 percent (high scoring peaks) to 38 percent (all peaks). Variability between ChIP datasets, even for duplicate measurements of the same factor, has been noted in previous studies, but overall trends indicate measurable differences in measured genomic occupancy of these factors.

Although ChIP datasets provide partially inconsistent information about occupancy, in combination they are an important guide to the regulatory potential of specific DNA elements. In some cases, loci that are inactive can be readily distinguished from loci around active genes based on ChIP occupancy alone. Sparse occupancy around weakly transcribed or inactive genes (Figures 2-4 A, 2-4 B) contrasts with abundant binding of factors near loci adjacent to robustly expressed genes (Figure 2-4 C, 2-4 D). However, around active genes, some bound regions may not be associated with known enhancers. We found many instances of transcription factor occupancy of tested, inactive regions in the Fly Enhancer Resource (e.g. circled region in 2-4 D). These regions are difficult to visually distinguish from active enhancers.

Figure 2-4 ChIP peaks correlate with active elements

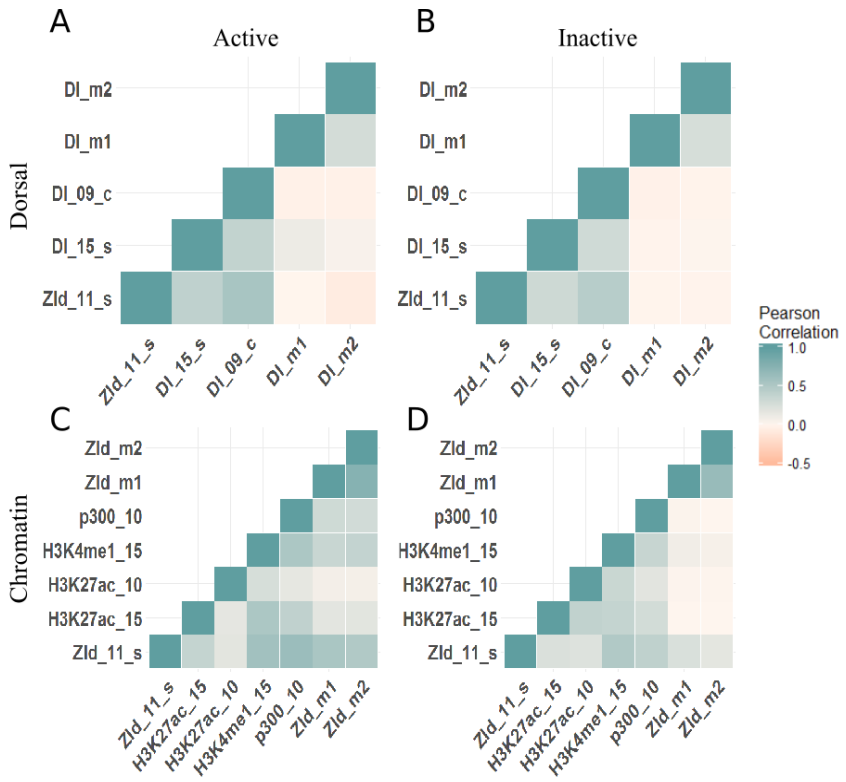


Low levels of binding by Dorsal, Twist and Zelda at A weakly expressed *sgl* locus and B inactive *kar* locus. C Numerous bound regions associated with actively transcribed *vnd* and D *brk* loci. However, some regions found to be inactive in Kvon et al. 2014 are bound by multiple transcription factors (circled region near *brk*). Gray bars indicate portions of genome assessed by Kvon et al. 2014. Blue bars indicate transcription units for genes.

Reproducibility between measurements of features was limited, but higher for stronger signals. To determine if these reproducible strong signals were more likely to be associated with active elements, we used the pairwise comparisons of features from Figure 2-2 to measure the correlation in scores for overlapping features in both DNA elements that were strongly active and inactive in stage 4-6 embryos. In the case of Zelda and Dorsal (transcription factors associated with a large number of ChIP-peaks, Table 2-1) and histone marks associated with enhancer status, occupancy scores for different datasets measuring the same transcription factors within a given DNA element generally correlate with each other, as do datasets measuring different but functionally related features, although the difference in this correlation between active and inactive regions is minimal (Figure 2-5). The same trend is largely true for other transcription factors (Fig. 2-6). This suggests that the intensity of peak scores are fairly reproducible between datasets, but that functional binding is not more reproducible or consistent than background binding. Motif scores generally do not correlate at all with occupancy scores for transcription factors that overlap with the motifs, except in the case of Zelda (Figure 2-5, Figure 2-6). This correlation seems to be much stronger in active DNA elements than in inactive elements (Figure 2-4 C and 2-4 D).

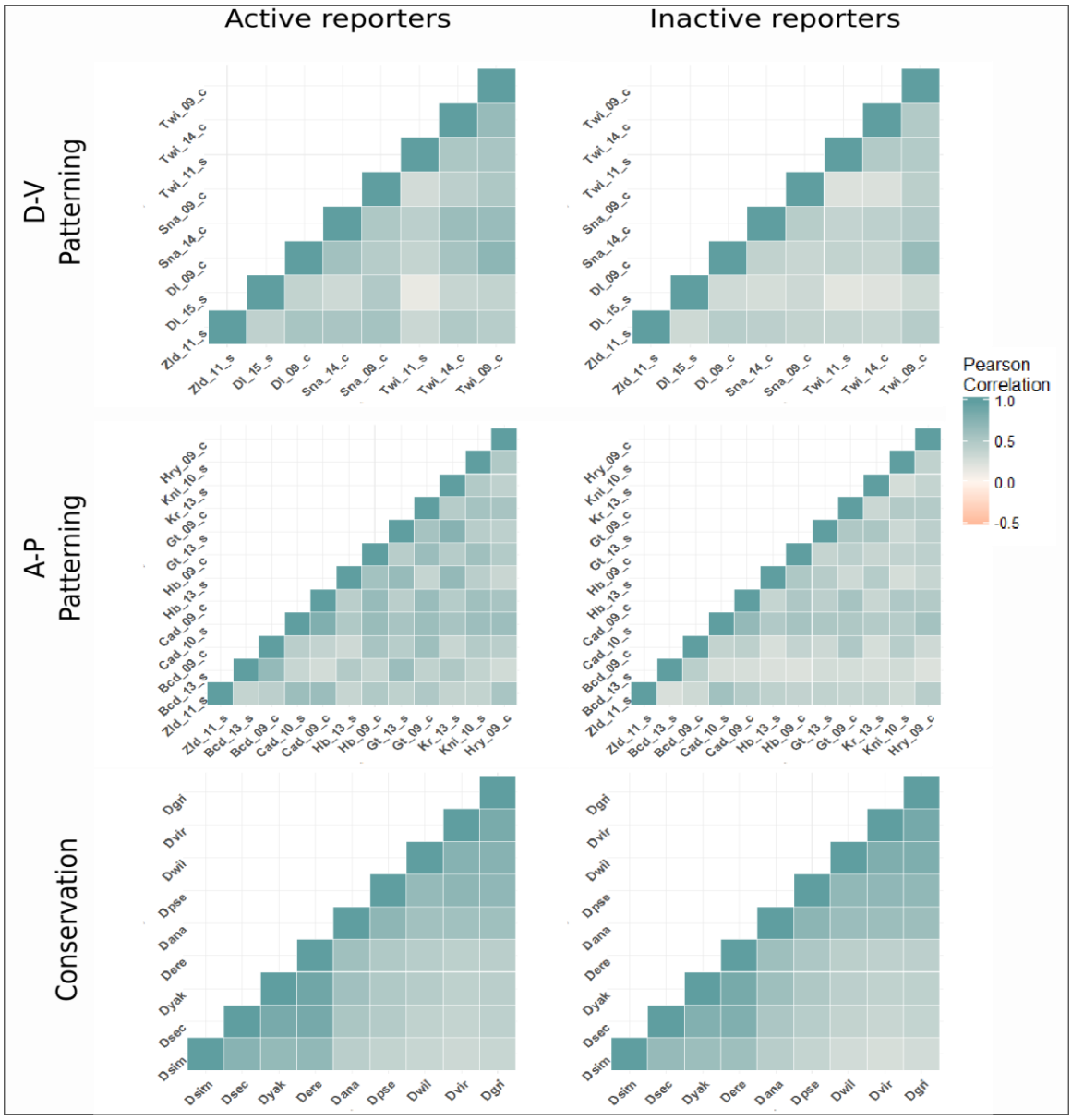
The trends shown in Figure 2-4 suggested that there may be qualitative differences between binding or chromatin modification that could be used to distinguish them. Therefore, we plotted the distribution of peak or motif scores for the same set of active or inactive DNA elements for individual ChIP datasets and motifs that were used in Fig. 2-5 (Figure 2-7). For some ChIP datasets, the higher-scoring peaks were associated with active DNA elements (e.g. Snail 2014 chip). This trend was not consistent for a given transcription factor, suggesting that different thresholds or levels of background binding in these datasets may obscure relevant signals.

Figure 2-5 Correlation between features for active and inactive elements



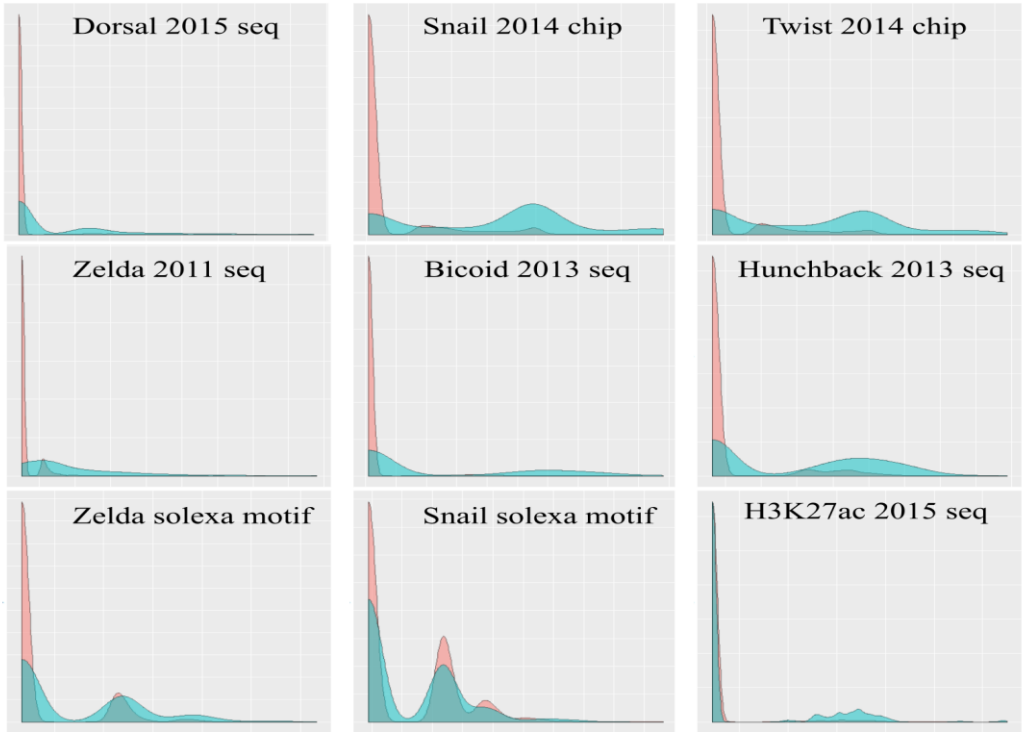
A,B: Levels of Dorsal and Zelda were assessed on active and inactive regions. Zelda and Dorsal peaks were correlated on both classes of elements, but little correlation is noted between protein occupancy and the respective binding motifs. C,D: For chromatin marks and pioneer factor Zelda protein and motifs, higher correlations were noted on active vs. inactive elements. Results from 392 highly active elements (left) and 6,858 stage 4-6 inactive elements (right). Comparisons for additional factors shown in Figure 2-6.

Figure 2-6 Correlations of feature scores for overlapping or replicate features



Most functionally related features have scores that partially correlate. The correlation is not noticeably weaker in inactive regions, which might indicate that the correlation is not due to the functional relationship.

Figure 2-7 Differential ChIP peak or binding motif enrichment values

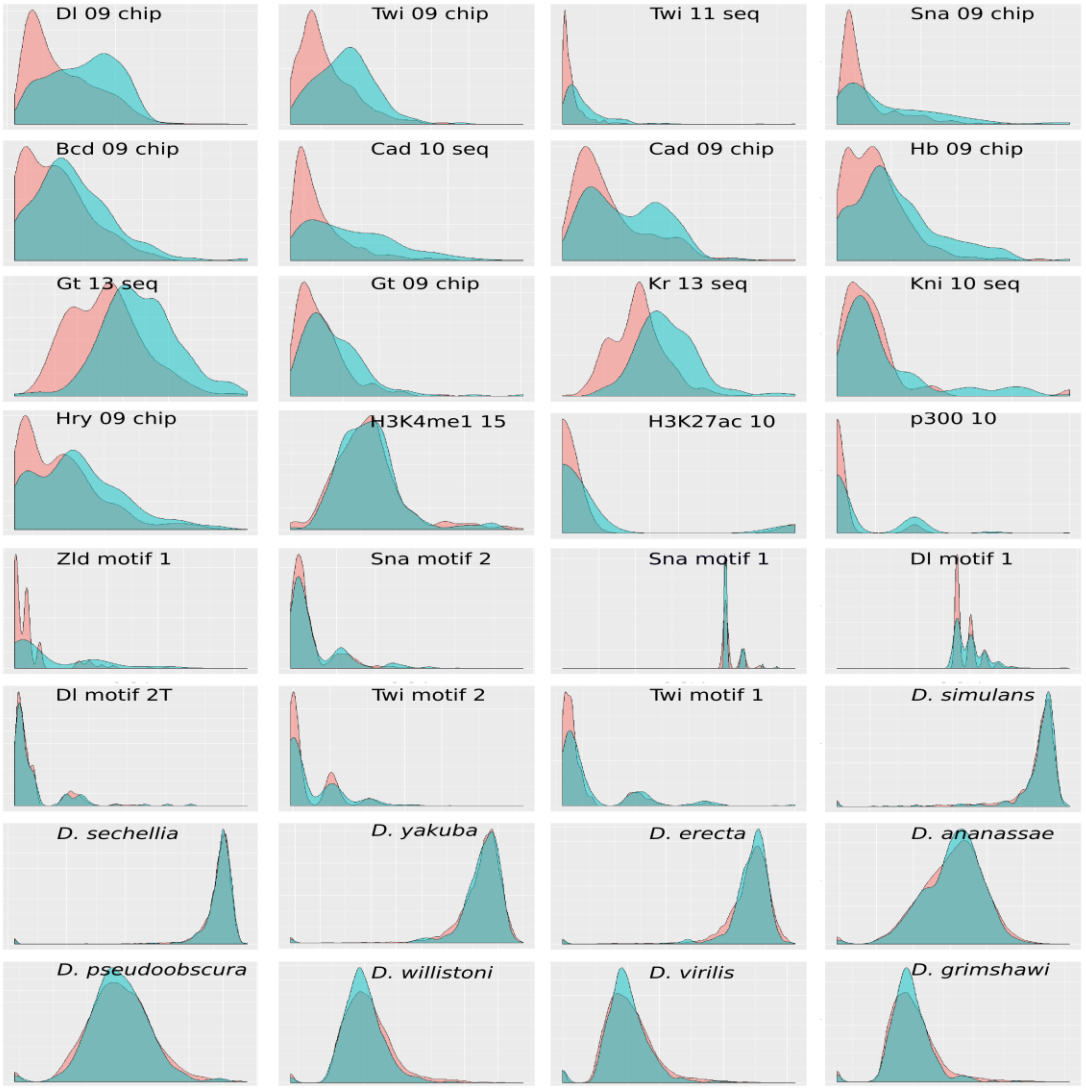


Histograms showing distributions of ChIP peak or binding motifs were plotted for DNA elements active (blue) and inactive (red) in stage 4-6 embryos. DNA elements from both categories that did not overlap with the signal were excluded from the comparison, and the histograms from both categories were scaled to integrate to 1. In most cases, there were a larger number of inactive regions bound by transcription factors than active regions, as there are 6,858 inactive regions vs. 392 active regions (Table 2-2). Some datasets showed substantial differences in the distributions between active and inactive elements e.g. Snail 2014 chip and H3K27ac 2015 seq. No large differences in enrichment of specific regulatory motifs were observed between active or inactive elements.

Table 2-2 Overlap of active and inactive reporters with features

Feature	Overlap Active	Overlap Inactive
All reporters	392	6858
Dorsal 2015 ChIP-seq	118	222
Dorsal 2009 ChIP-chip	326	2081
Snail 2014 ChIP-chip	279	1083
Snail 2009 ChIP-chip	168	413
Twist 2014 ChIP-chip	257	1115
Twist 2011 ChIP-seq	240	885
Twist 2009 ChIP-chip	320	1528
Bicoid 2013 ChIP-seq	118	253
Bicoid 2009 ChIP-chip	138	130
Caudal 2010 ChIP-seq	250	780
Caudal 2009 ChIP-chip	175	275
Hunchback 2013 ChIP-seq	192	777
Hunchback 2009 ChIP-chip	201	424
Giant 2013 ChIP-seq	190	661
Giant 2009 ChIP-chip	169	193
Hairy 2009 ChIP-chip	210	525
Knirps 2010 ChIP-seq	111	96
Kruppel 2013 ChIP-seq	204	842
Zelda 2011 ChIP-seq	292	940
H3K27ac 2015 ChIP-seq	98	335
H3K27ac 2010 ChIP-seq	123	537
H3K4me1 2015 ChIP	167	421
p300 2010 ChIP-seq	168	291
Zelda motif sanger	165	1073
Zelda motif solexa	148	1208
Dorsal motif FlyReg	249	3983
Dorsal motif NBT	149	1959
Snail motif FlyReg	138	2632
Snail motif sanger	119	2013
Snail motif solexa	156	2720
Twist motif da	134	1952
Twist motif FlyReg	178	3154

Figure 2-8 Similarity of distribution of feature scores for active vs. inactive elements



ChIP datasets show greater differences, while conservation scores are indistinguishable on active and inactive elements. Datasets from within the same lab, using different antibodies for the same transcription factor, were merged.

Transcriptional activity and stage specificity can be classified based on genomic features

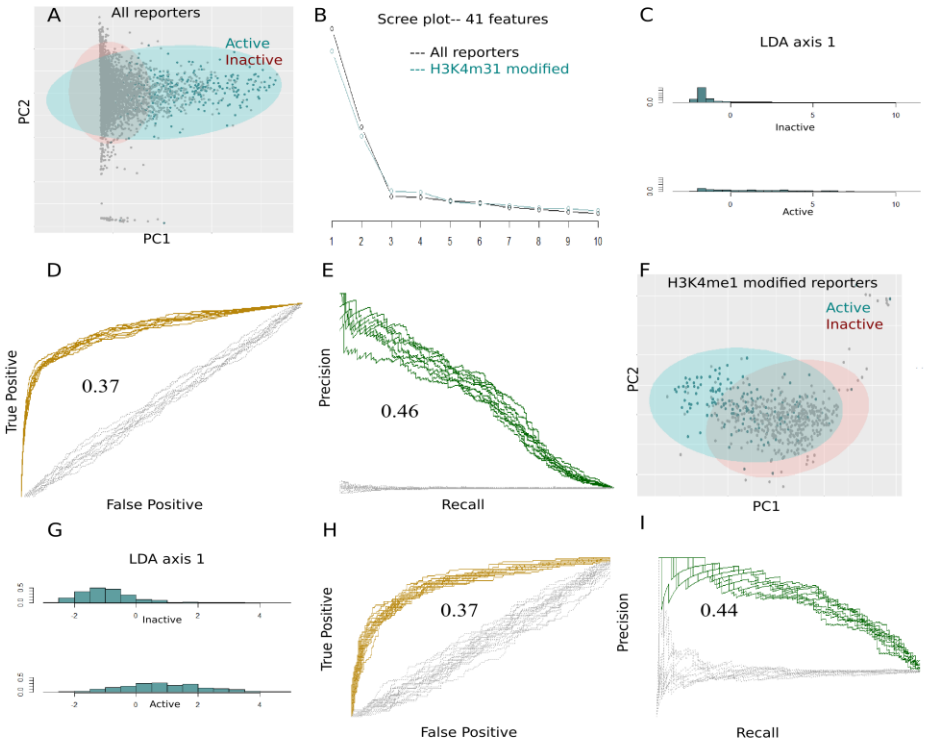
Our analysis indicated that no single feature conclusively correlates with regulatory activity, therefore, we asked if combinations of these features might distinguish active from inactive DNA elements. To judge the feasibility of this approach, we implemented a Principal Components Analysis (PCA) of the forty-one datasets. PCA of all features for all DNA elements (392 active and 6858 inactive) showed some differences in their distribution. (Figure 2-9 A). The largest contributors to PC1 (the greatest sources of variation in the dataset) are mostly occupancy information for transcription factors associated with Anterior-Posterior patterning, like Caudal and Hunchback (Data not shown). Interestingly, the largest contributors to PC2 are primarily the degree of evolutionary conservation in each DNA element, based on various species comparisons. Almost all variation in the dataset was captured in the first ten principal components (Figure 2-9 B).

Although active and inactive regions overlap, the visible partial separation based on the first principal component suggested it may be possible to classify activity using machine learning. Linear discriminant analysis of all DNA elements showed a degree of separation, although there was overlap in the distribution of where they fell on LD axis 1 (Figure 2-9 C, cross validation in Figure 2-10). Next, we used a random forest (chosen for its amenability to biological interpretation) to classify all active and inactive DNA elements based on the same 41 features used for PCA and LDA. A random forest trained on two thirds of the active and inactive DNA elements was able to classify the remaining one third with 96 percent accuracy (measured by total percentage of true positives and true negatives) over 10 trials. Receiver Operator Characteristic Curve (ROC) and Precision recall curves are shown in Figure 2-9 D,E. Classifications are much more successful as measured by ROC than by Precision-recall, as

indicated by their respective measurements of 'area under curve' (AUC). Zelda and Bicoid datasets were found to be the most important features (Data not shown).

Many of the inactive regions used in the PCA and classifications in Figure 2-9 C-E are relatively sparse in terms of transcription factor occupancy, like the regions shown in Figures 2-4 A, B. These unoccupied regions can be trivially distinguished from potential enhancers. A more challenging question is how to separate active regions from inactive regions that superficially resemble them, containing many of the same marks, as shown in Figures 2-4 D. Chromatin signatures are frequently used as a proxy for active elements, therefore we selected loci that are associated with H3K4 mono-methylation for training and testing, reducing the datasets to 167 highly active regions and 421 inactive regions. These are much less clearly separated by Principal Components (Figure 2-9 F) and LDA (Fig. 2-9 G). Turning to random forests, the AUC for ROC is nearly as high as before (compare Fig. 2-9 D, H), and Precision Recall is dramatically improved (compare Fig. 2-9 E,I). However, the reduced dataset is much closer to having a balanced number of samples for active and inactive elements, so the probability of correct classification due to random chance was greatly increased. Neither ROC nor Precision recall showed substantial improvements in AUC relative to random chance. Restricting the analysis to loci that were bound by one or more transcription factors similarly improved overall Precision Recall, but not relative to background (discussed in Fig. 2-11).

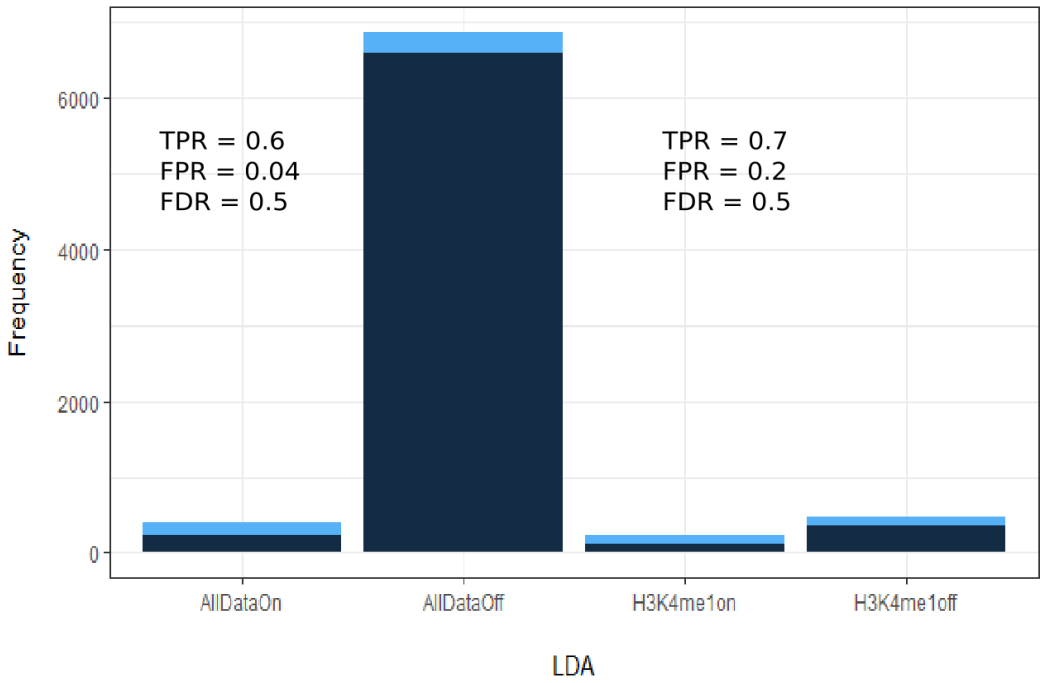
Figure 2-9 Active DNA elements can be separated from inactive regions



A: DNA elements with and without regulatory activity can be partially visually distinguished based on Principal Component 1; additional principal components showed little separation. There is a sharp boundary on the PC1 axis; this is probably due to PC1 variation primarily coming from ChIP occupancy scores, and the majority of DNA elements having no occupancy, and thus share an identical 'low score' numerical stand-in. The separation is far more visible using supervised machine learning. B: Scree plot for PCA in (A: black) and (F: blue). C: Separation between active and inactive regions with linear discriminant analysis, using all DNA elements (cross validation in Figure 2-10). D: Random forest trained on 2/3 of both unrestricted and restricted datasets (using all 41 features) was able to achieve 96% accuracy of total calls, and a high ratio of true positives relative to false positives, as shown in ROC plots (yellow), and a

difference of 0.37 in the area under the curve relative to classifications made on random data with the same sample size (whose ROC are shown in gray). However, precision recall (E, green plots) was much lower, although even higher relative to background (gray plots), with a relative AUC difference of 0.46. F: Using only active and inactive DNA elements that are distinguished by H3K4 mono-methylation reduces the sample size from 392 highly active regions and 6858 inactive regions, to 167 highly active and 421 inactive regions. Separation was substantially less visible when DNA elements were restricted to only include those marked by H3K4me1. This also lowered the degree of visible separation based on LDA (G) and decreased overall accuracy to 83% and area under the ROC curve (H) It led to substantially higher precision recall (I). However, it led to equal increases in the AUC of predictions made on random data of the same sample size, shown in gray.

Figure 2-10 LDA Cross validation

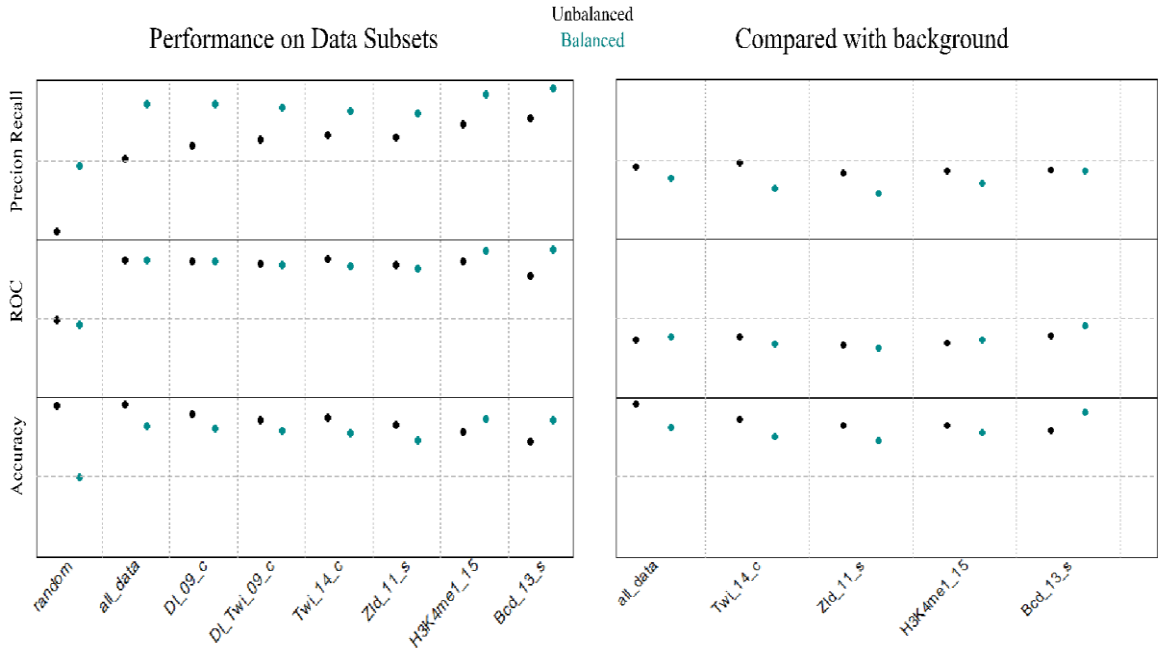


Classifications based on Linear Discriminant Analysis cross-validation, corresponding to analysis in Figure 2-9. Dark blue indicates active elements, light blue indicate active elements, columns indicate how these elements were classified. LDA was performed on training sets that included all active and inactive DNA elements from stage 4-6 in the Fly Enhancer resource, and also only those that were marked by H3K4me1. The numbers are provided that correspond to the True Positive Rate (TPR), False Positive Rate (FPR), and False Discovery Rate (FDR) for both classifications.

Impact of number of features on predictive accuracy

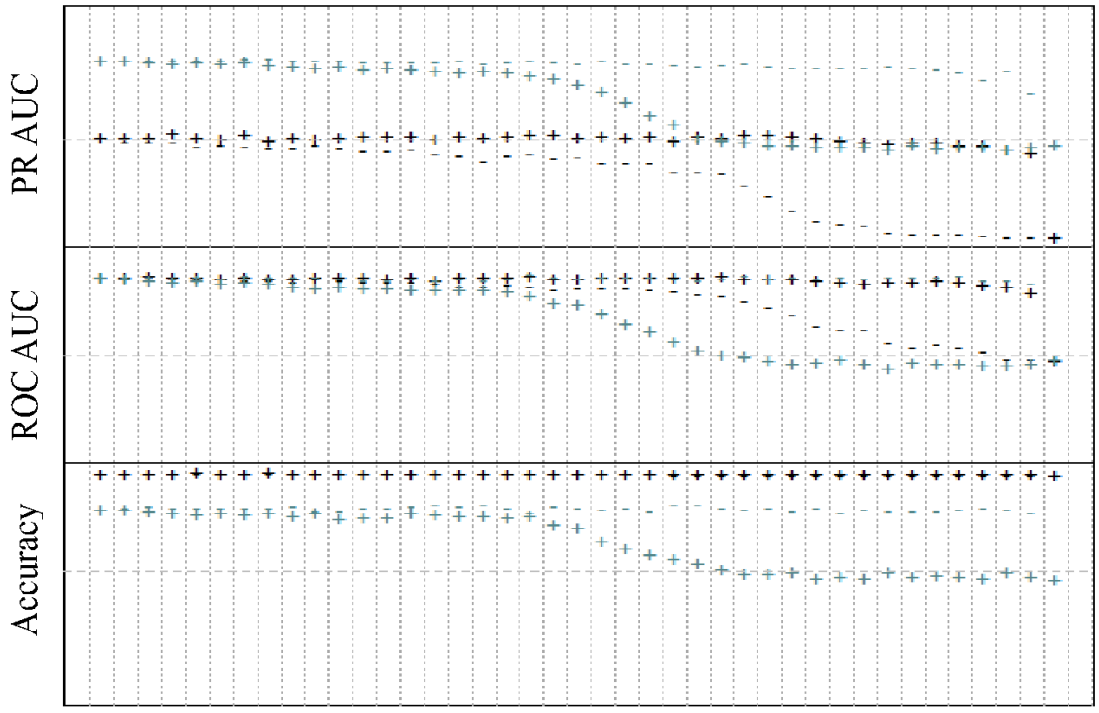
To determine if a specific combination of features would be the most effective for predictions (e.g. only using the most predictive features, or excluding the least predictive features), we re-performed random forest classification using the same training data, with different combinations of the 41 predictors. We iteratively left out individual features (Figure 2-12). After each run of ten random forests, the features were ranked using Gini Impurity Index. The next run then excluded either the most or least important feature based on the previous run, and performance was compared between each. We found that dropping the most important feature, or even top three features had little impact on precision recall; loss of additional features had some negative impact on classifications. After dropping the eighteen most important features (largely ChIP data), precision recall and ROC curves are depressed. However, dropping the least important features had no noticeable impact, until over thirty features are dropped. Analyses using only the top twelve most useful features were equivalent to employing all features. These results suggest that the features must contain a significant amount of redundancy, but the redundancy is not interfering with accurate predictions.

Figure 2-11 LDA Exclusion of unbound genomic regions and use of a balanced training set



Accuracy/Precision scores for random forest models trained on different subsets of the Fly Enhancer Resource DNA elements. Top: Area under precision recall curves. Middle: Area under ROC curves. Bottom: Accuracy, calculated as total percentage of correct calls. Unbalanced datasets are indicated in black, blue indicates models trained on datasets balanced through undersampling the majority category (inactive reporters). "Random" represents classifications with elements for which all feature scores were randomly generated, with values 0-1000; for other marks indicated, DNA elements were selected that contained that mark, or those two marks, and at least one other.

Figure 2-12 Random forest performance with least and most informative features

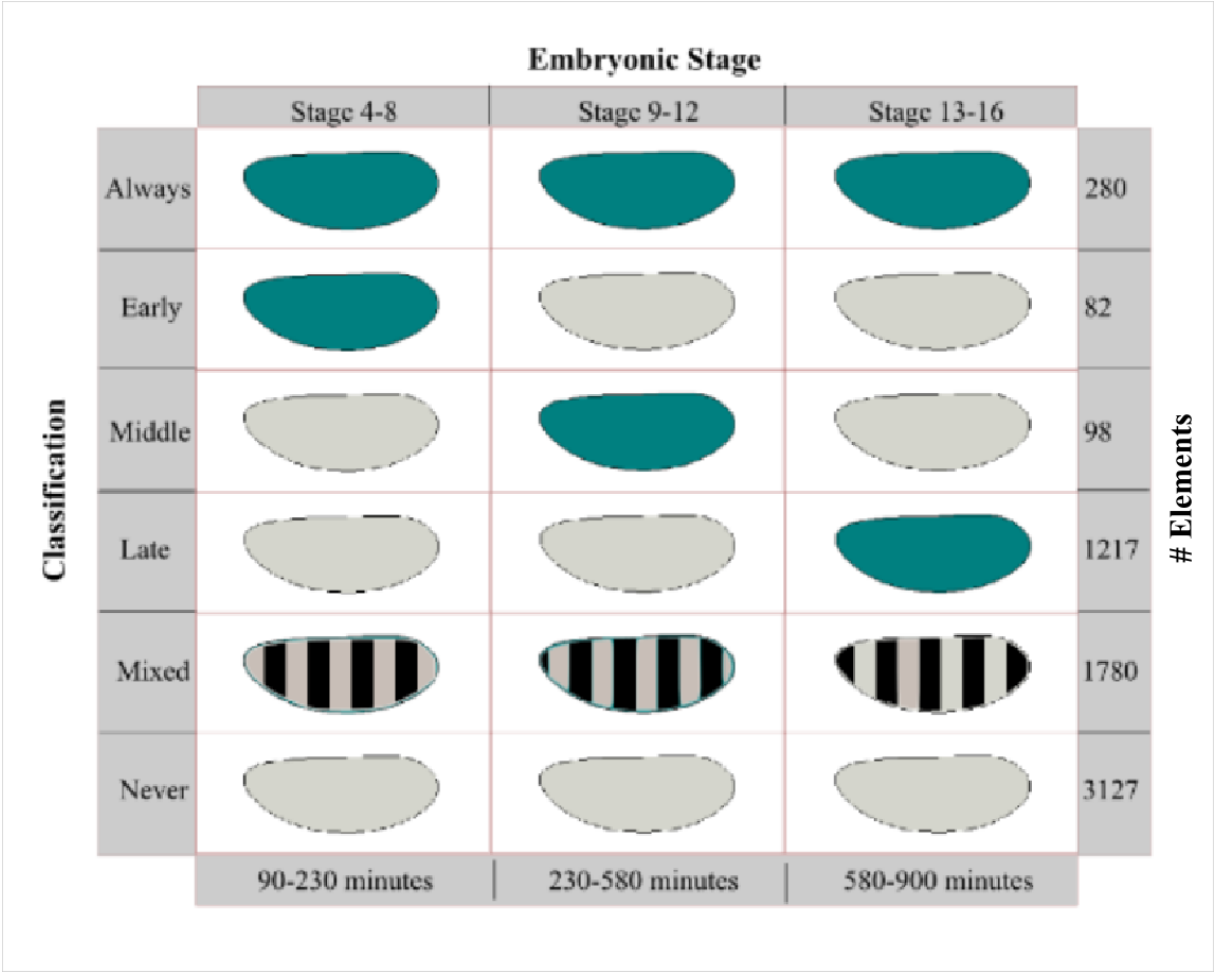


Predictive success of models with differing numbers of features, using unbalanced (black) datasets using all active and inactive elements, and balanced datasets (blue), in which 200 active and inactive elements are used for training, and 100 for testing. Predictions were made by iteratively dropping the most informative feature to least informative feature (+), or least informative to most informative (-). The scores running from left to right represent predictions made using 41 features, ranging to predictions made using only two features. The furthermost right scores are predictions made based on 41 randomly generated features. Note that balanced datasets have substantially higher 'background', or scores based on random datasets of the same size (Fig. 2-9)

Classifications by stage specificity and expression pattern

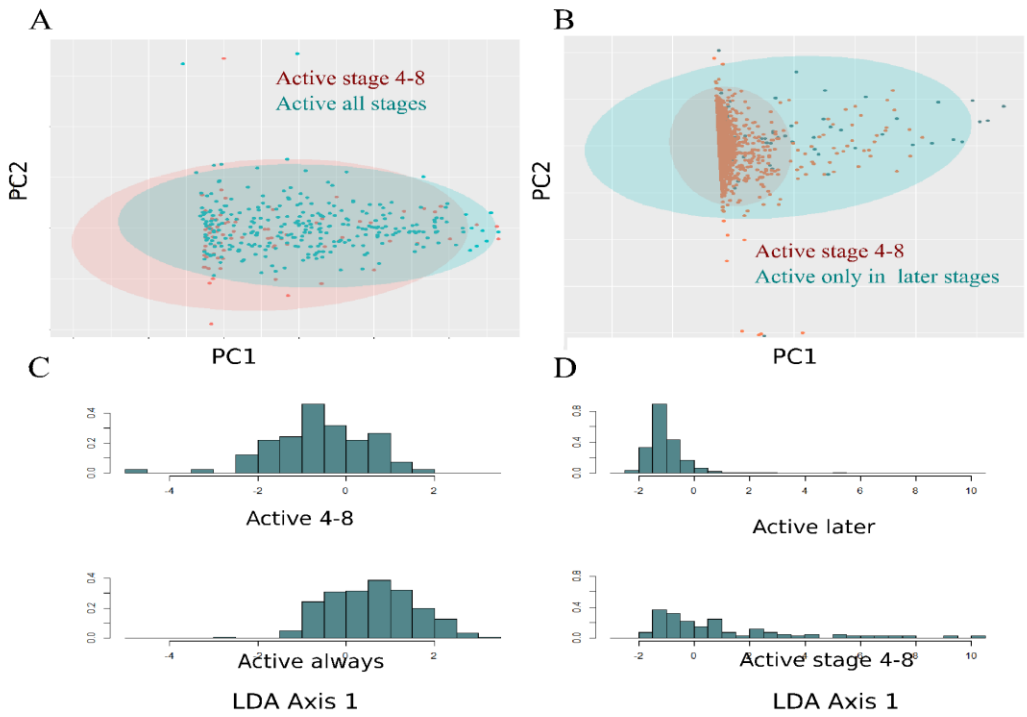
Some studies have shown that genes that are active throughout development have features that make them distinct from genes and regulatory regions that are tissue or stage specific[75], or that separate poised enhancers from those that are actively expressed[76]. However, a comparison of DNA elements that are active in all stages with those that are specific to early embryonic development do not suggest that they can be as easily distinguished as active and inactive elements (Figure 2-14 A,B). A possible mitigating factor is that regulatory regions often gain binding of transcription factors and chromatin marks in a successive fashion during development, prior to becoming active. Thus, some of the elements considered above as "inactive" may actually be false negatives. Here, some DNA elements are exclusively active in embryonic stages 4-8, whereas are not active in this earlier period, but become active in stages 9-12 or 13-16 (Figure 2-13). We tested if the 41 features would distinguish such elements that become active only in later stages based on their features measured at an earlier time point. Compared to the results with active vs. constitutively inactive, PCA and LDA analysis did not show much separation (Figures 2-14 B,D). Consistent with these results, random forest classification was only slightly more successful than would be expected by chance (data not shown). The low accuracy may stem in part from the relatively smaller training set, however, it is also possible that the considered genomic features are simply not informative about early vs. late active DNA elements. Moreover, most of the features used for this classification were from datasets collected at the early embryonic stage of development; additional information about chromatin conformation or later transcription factor occupancy would likely improve predictions for later enhancers.

Figure 2-13 Distinct classes of elements tested for function in embryos



To assess predictions of different classes of active elements, DNA elements (Kvon et al. 2014) were assigned to one of six classifications. "Mixed" refers to elements that exhibit activity in two of the three indicated stages.

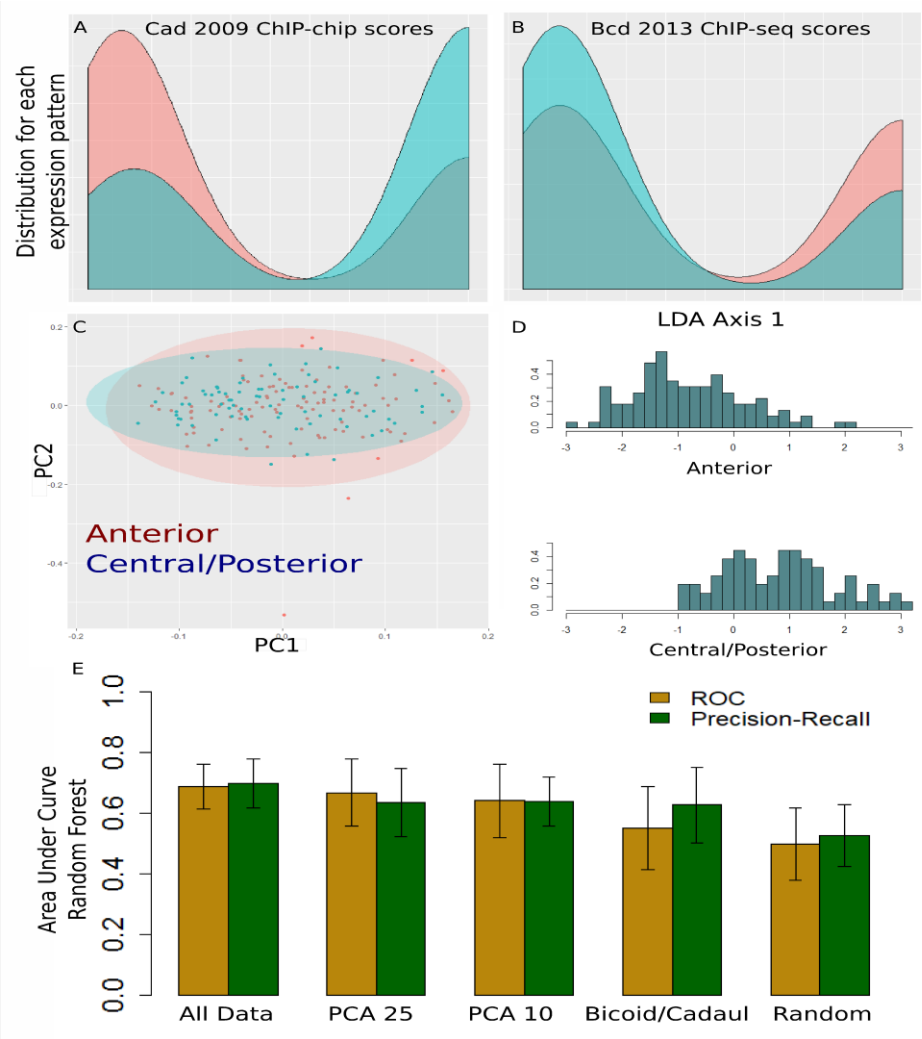
Figure 2-14 Principal component analysis, Random Forest classification, of subsets



Principal components analysis shows very little separation based on the first two principal components of DNA elements that are active or inactive at specific developmental times. (A) This is true when they are separated based on continuous activation (identified as 'always' in Fig 2-13) vs. stage-specific activation (identified as 'Early' in Fig 2-13), or (B) early stage specific activity from activity specific to later stages (later stages combining groups 'Middle' and 'Late' from Fig 2-13). Linear discriminant analysis also shows very little separation along LD axis 1 for either comparison (C-D)

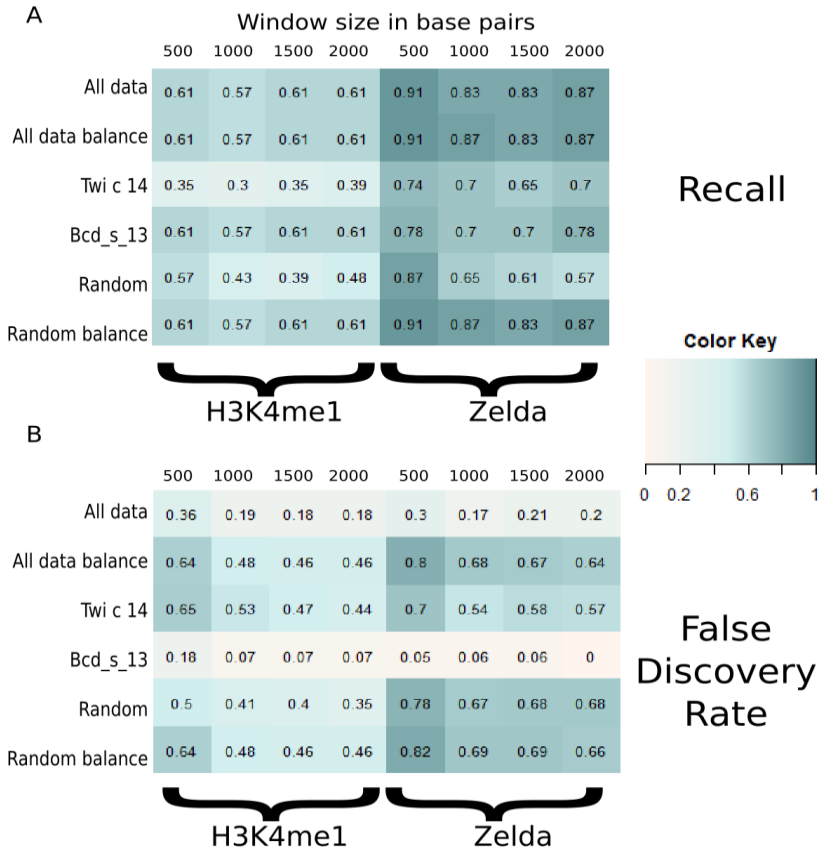
Predicting expression patterns is a more difficult problem than predicting activity. Although Kvon et al. 2014 annotated expression patterns for all DNA elements at each stage of embryonic development, their annotations were in many cases non-exclusive, permitting a DNA element with distinct A/P stripe activity to be annotated "mesoderm" because of its partial expression in that region. We re-annotated expression patterns based on the images available at the Fly Enhancer Resource, grouping patterns into exclusive and non-overlapping categories. We categorized 114 elements as "anterior", and 78 as "central or posterior", and 22 with "mesodermal or neurogenic ectoderm" expression patterns. Interestingly, a substantial fraction of Anterior/Posterior and Dorsal/Ventral enhancers were ubiquitously active throughout embryonic development. Occupancy of certain transcription factors associated with Anterior vs. Central or Posterior expression patterns correlated well with the corresponding categories of DNA elements: Bicoid peak scores were higher on Anterior elements, and Caudal peak scores were higher on Central/Posterior elements (Figure 2-15 A and 2-15 B). Considering all 41 features, however, PCA and LDA did not show much separation between these two expression patterns (Figure 2-15 C, D). A random forest trained on all features performed better than predictions using randomized data, but not as well as previously observed for differentiation of active vs. inactive DNA elements (Fig. 2-15). Excluding Bicoid and Caudal related data had a dramatic effect on predictions. Small training sets such as those used here are generally not compatible with large numbers of predictors, therefore, we tested dimensionally reduced datasets, to see if this would improve precision. Reducing the number of predictors using principal components did not improve performance (Figure 2-15). Because Bicoid and Caudal play such key roles in these predictions, Anterior vs. Central/Posterior expression may be predicted well if a larger data set of specifically Anterior/Central-Posterior elements were available for training.

Figure 2-15 Features correlate with tissue specific expression patterns



Although some features clearly correlate with Anterior vs. Central/Posterior expression patterns (A,B), and these categories are largely indistinguishable when viewed by PCA (C) and show only slight separation based on linear discriminant analysis D: Random forests can classify these active elements by expression pattern with greater success than when using random data, but not with sufficiently high precision or accuracy to make useful predictions. Reducing the numbers of features does little to improve accuracy of classifications (E).

Figure 2-16 Recall relative to false discovery rate when identifying enhancers.



Clusters of transcription factors around Zelda binding OR around H3K4me1 marks, that involved two or more overlapping peaks, were used to define putative enhancers. These were defined as 500, 1000, 1500, or 2000 base pairs centered around the clustered binding. Random Forests were trained on subsets of the Fly Enhancer Resource, using all strongly active and inactive DNA elements, or only active and inactive elements that overlap with Bicoid ChIP-seq peaks, or only active and inactive elements that overlap with Twist ChIP-chip peaks (two conditions that performed very well on the Fly Enhancer Resource in Fig 2-11). Random forests were also trained on randomized DNA elements, where no features distinguish active and inactive elements.

Testing against validated enhancers

Enhancer classification described above depends on the extensive data from the Fly Enhancer Resource, which randomly samples ~15% of the entire genome, and has limited information about which genes the putative enhancers regulate. Several intensively studied developmental genes in stage 4-6 embryos have more detailed information about which regions are important for expression, as found in the RedFly database[72]. To use these loci as independent guides for enhancer classification efforts, we tested eight different models derived from random forest analysis of balanced or unbalanced data. These models were used to predict active DNA elements within a 50kb window encompassing the eight target genes (expressed in Anterior/Posterior, Mesoderm, and Neuroectoderm patterns). Our assumption is that previously annotated regulatory regions around these genes are effective measures of true-positives, and that other regions lack enhancer activity. Thus, any additional enhancers identified within this window that do not overlap with curated enhancers are likely to be false positives.

To ensure that our random forest models evaluate all relevant regions around these genes, we needed to include additional DNA elements, as the Fly Enhancer Resource does not cover the entire genome. Therefore, we identified all putative enhancers by presence of bound transcription factors. In Figure 2-16, putative enhancers are defined as any locus that is binding of Zelda, as well as one or more additional transcription factors, or alternatively any locus binding H3K4me1 as well as one or more transcription factors. The 'borders' of each putative enhancer were set to 2000 base pairs (approximately the same as the window size used in the Fly Enhancer Resources). There are 23 curated enhancers for these genes total. However, there are an additional 31 clusters of enhancer-like binding around the protein Zelda. Almost every model trained on the Fly Enhancer Resource had high recall, and was able to identify the majority of the

23 'true positives' (2-16 A). However, many conditions led to an almost equal number of 'false positives', or a high false discovery rate (FDR) (2-16 B). Random forests trained on balanced datasets had high false discovery rates, identifying many presumably non-functional clusters as active enhancers. Models trained on unbalanced data were far more successful at rejecting the non-functional binding, although the balanced datasets did performed better at training and testing within the Fly Enhancer Resource. Unbalanced datasets may more accurately reflect natural conditions; the frequency of enhancer-like binding in stage 4-6 embryos appears to be far more common than actual enhancers. We performed the same test again, setting the putative enhancers to window sizes of 500, 1000 and 1500 base pairs, and the same trends were observed. There are exactly 23 clusters of binding in 2kb regions around H3K4me1 marks, which do not encompass every known enhancer. As a result, both Recall (percentage of true positives identified) and FDR were much lower.

Discussion

Many genomic databases include information about regulatory regions. However, many primarily collate 'true positives' that show activity, or simply display chromatin features that are positively correlated with activity. From these resources, it is difficult to know how many real regulatory are missed, or the frequency of false positives. The Fly Enhancer Resource is the first publicly available database that includes a large set of reporters that are inactive at various developmental stages. We used features associated with enhancer activity and cross-listed them with the coordinates defining active and inactive regions in the Fly Enhancer Resource. From this we were able to draw conclusions about which features are the most indicative of enhancer activity, and how combinations influence classification.

Features correlating with enhancer activity are not guarantees of enhancer function

Many features have been associated with enhancer activity, including sequence conservation, transcription factor occupancy, chromatin marks, and motif enrichment. Sequence conservation has been observed in some *Drosophila* regulatory elements, but using it as a predictor is complicated by the generally rapid rate of divergence in regulatory sequences[50], and the compact and largely conserved nature of the *Drosophila* genome[49]. Although ChIP-seq data correlates with specific enhancer functions, it is challenging to use it as a predictor without a priori knowledge of the genes being controlled. For example, previous studies have observed that the transcription factors responsible for Dorsal-Ventral patterning (Dorsal, Snail and Twist) are also observed at DNA elements for genes controlling Anterior-Posterior patterning[40], although not generally considered canonical players in A-P development[77].Some features (e.g. Zelda ChIP-seq) have many low scores that overlap with inactive regions, probably indicating a bioinformatics threshold (where scores below a certain minimum were excluded as probable background binding, and the reported peaks only represent a fraction of the original distribution). It is possible that using an even higher threshold for ChIP peak scores in these cases could remove a large amount of potentially spurious features. However, this would almost certainly leave out many true positives. Motif enrichment is largely indistinguishable between active and inactive DNA elements. Consistently across all classes of features, there are strong signals that overlap with regions that do not drive gene expression, possibly due to regions like those observed in Figure 4-C and 4-D, which share many of the characteristics of active DNA elements. However, most genomic regions that did not drive expression overlapped with far fewer features associated with enhancers (data not shown). This likely reflects regions like those shown in Figures 4-A and 4-B, where inactive or weakly active genes are in loci with little

protein occupancy.

We also found that datasets measuring the same feature in some cases were highly variable. This is not surprising; differences in resolution between ChIP-chip and ChIP-seq are well documented[78], as are the many factors that can lead to higher reproducibility within labs than between labs[79]. However, not all results were inconsistent. In three out of four datasets measuring Twist binding, the transcriptional activator visibly correlates with all transcription factors, suggesting that Twist can play a role in diverse transcriptional networks, or that it binds easily to all regions of open chromatin (Figure 2-2). It is plausible that further increasing the number of features used would improve results; overlapping features would be more likely to be non-spurious, and consequently to have high peak scores (in the case of ChIP data) or motif scores (in the case of transcription factor PWMs), and greater reproducibility. However, the degree to which classification accuracy plateaued suggests inclusion of more data may have marginal returns.

Zelda a strong but not definitive indicator of early enhancer activity

Zelda occupancy is by far the most informative feature (Data not shown), which may indicate that Zelda, more than other factors, plays a determinate role in establishing active biological enhancers during early embryonic development. Interestingly, conservation scores, especially in the more distantly related species, are among of the more important features for this classification indicated by measures of feature importance from random forests using all available data (data not shown). This is surprising given the lack of visible correlation in Figure 2-2. It is possible that the importance of conservation information measured by Gini Impurity Index is inflated by the high level of variation contributed by this data, as shown in PCA. It is also possible that other features are providing redundant or correlated information, elevating the

importance scores for conservation. Leave-one-out analysis (Figure 2-12) suggest although some features are substantially more informative than others, no one feature is essential for predicting active DNA elements, and many combinations of features can produce equally accurate predictions of activity. Even classifications made using the least informative features still outperform classifications based on random data. The largest drops in classification accuracy were after dropping most features based on ChIP-seq and ChIP-chip of transcription factor proteins. ChIP-seq of histone modifications was less informative than transcription factor occupancy, but still more valuable for classifications than motif information or evolutionary conservation.

Potential limitations and stage specificity

A possible complication for predictive accuracy is that many regulatory regions assemble much of the transcriptional machinery well before the developmental time in which they are active[21]. We used a random forest to classify DNA elements as stage-specific or ubiquitous, using the same 41 features from previous classifications. Separating stage-specific vs. ubiquitously active DNA elements has substantially lower precision relative to recall, which is unsurprising given the smaller size of the dataset (only 82 of the DNA elements from the Fly Enhancer Resource are exclusively specific to early development), that both stage specific and ubiquitous DNA elements are active in stage 4-6, and many of the predictive features are specific to that time point. Interestingly, most tissue specific specific expression patterns (e.g. DV stripes at the Mesoderm, or AP pair-rule type stripes) were not associated with either category-- approximately half of these DNA elements were stage specific, and the other half were active driving other expression patterns later in development (data not shown). Inclusion of predictive features that are taken from later developmental time-points may allow more effective pinpointing of which (if any)

features can be used to identify enhancers that are active throughout development.

Characteristics and Classification of DNA elements with Varying Expression Patterns

Predicting expression patterns is a more difficult problem than predicting activity. We used the Fly Enhancer Resource for this as well. Although Kvon et al. 2014 annotated expression patterns for all elements at each stage of embryonic development, their annotations were in many cases overlapping-- any expression in developing germ layers, or in the anterior or posterior halves of the embryo, were associated with dorsal-ventral and anterior-posterior patterns. We manually re-annotated expression patterns based on the photos of embryos available at the Fly Enhancer Resource, to group patterns into exclusive and non-overlapping categories. A majority of the expression patterns are not easily categorized; they either have multiple expression patterns evident, high variability between different embryos imaged, or very amorphous or faint distribution of protein. We categorized 114 elements as driving expression only in the Anterior part of the embryo, and 78 that were exclusive to the Central or Posterior ends of the embryo. We identified 22 regions that displayed distinct Mesodermal or Neuroectoderm expression patterns, as characterized by those for the Dorsal-Ventral patterning genes *snail*, *brinker,* and similar. This expression pattern is strongly associated with a well understood biological function, and may correspond to a unique combination of features. However, the very small sample size did not make it ideal for machine learning. for comparison. Interestingly, neither Anterior/Posterior nor Dorsal/Ventral DNA elements displayed a tendency toward ubiquitous activity throughout development or toward stage specificity. Approximately half of each category was ubiquitous--elements that drive expression at every stage of embryonic development, while an equal number showed some level of stage specificity.

Transcription factors known to be associated with Anterior vs. Central or Posterior expression

patterns did show expected correlations (e.g. Figure 2-15 A,B-- Bicoid peak scores were higher DNA elements driving Anterior activity, and Caudal peak scores were higher on Central/Posterior elements), although PCA based on all 41 features did not show separation between these two expression patterns (Figure 2-15 C). A random forest trained on all features was able to perform better than predictions based on randomized data, but not as well as classifications of general activity. Although random forest algorithms are generally robust to co-linearity between features these predictions were trained on a much smaller sample size, and small training sets are not as compatible with large numbers of predictors. We created secondary datasets wherein the first ten or first twenty-five Principal Components were used instead of scores for individual features. Reducing the number of predictors using principal components did not hurt the accuracy of classification, but also did not engender substantial improvement (Figure 2-15 E). However, dropping individual features known to be important for defining Anterior expression patterns (Bicoid and Caudal binding) brought ROC and Precision Recall closer to what would be expected based on random chance (10e). Given the elevated importance of Bicoid and Caudal for these classifications, it seems likely that anterior expression could be predicted with reasonably high accuracy given a larger data set of specifically Anterior/Central-Posterior DNA elements for training.

Identification of enhancers based on genomic features

For testing on experimentally identified enhancers from other sources than the Fly Enhancer Resource, we assembled a list of putative enhancers by clustering of transcription factors, and classified them with random forests trained on the Fly Enhancer Resource. Random forests were trained on all active DNA elements from the Fly Enhancer Resource (as shown in Figure 2-9), and also on all regions that had been balanced to equal numbers of active and inactive elements

through under-sampling. We used several different criteria for identifying clusters to use as putative enhancers: Zelda binding that overlapped with at least one other transcription factor or chromatin mark, or H3K4 mono-methlyation that overlapped with at least one transcription factor or chromatin mark. There are far more Zelda ChIP-seq peaks than H3K4me1 peaks (Table 2-1), so the former criterion identified far more putative enhancers. Surprisingly, using either criteria, random forests trained on a balanced dataset identified 100 percent of clusters as active, which in some cases led to a very high false discovery rate (Figure 2-16). As the primary goal is to identify active enhancers, low precision and high false-discovery was a significant concern. A potential factor contributing to low precision of classification is the large fraction of inactive regions in stage 4-6 embryos with very little transcription factor occupancy. If enhancers are predicted based on the presence of clustered transcription factors, these unbound regions are less useful for distinguishing between cases where background binding is also clustered, and resembles truly active regions. Balancing by random under-sampling runs the risk of excluding most if not all of the ambiguous, occupied-but-inactive DNA elements. We used several other forms of under-sampling, only training on DNA elements that were bound by Bicoid, or bound by Twist (two of the more successful approaches in Fig. 2-11). When H3K4me1 marks were used as the standard for identifying putative enhancers, recall was capped at the number of H3K4me1 marks that overlapped with curated enhancers (about 60%), whereas when the more promiscuously binding Zelda was used, recall could reach over 90% (Figure 2-16 A). Correspondingly, false discovery rate was much lower across the board when H3K4me1 modification was used as the minimum standard for putative enhancers (2-16 B). The most overall successful combinations (high recall, low FDR) seemed to be those that combined a permissive setting with a restrictive setting, with regards to defining putative enhancers and what

data was used for training.

A potential source of inaccuracy with regards to classifying clustered binding as active or inactive is that the Fly Enhancer Resource is used as the sole source of training data. Although it covers a substantial portion of the genome, the included reporters may bisect biological enhancers, and not represent in-vivo expression patterns due to loss of essential activators or repressors. The reporters also cannot account for position specific effects, or the potential influence of promoter specificity. This necessarily means that some of the reporters will be incorrectly classified as active or inactive for specific expression patterns, and introduces some minimal levels of error. When predicting enhancers around well-studied genes, it is possible that some clusters of transcription factor binding that strongly resemble enhancers might be regulating other genes. The assumption that all putative enhancers within a window around an active gene are acting on that gene introduces another source of potential mis-classification.

It is also possible that elements currently classified as false positives are actually regulatory elements that are active at an earlier or a later time in development. However, based on these features there is no combination that appears to separate DNA elements active later in development from ubiquitously inactive regions. Although some regions that are classified as false positives may be active at later stages, many of the regions active later in development do not share these characteristics. Presumably there are additional features that are not currently known that allow biological systems to make this distinction.

Conclusions

Genomic features can be used to predict overall enhancer activity with a high degree of accuracy. The goals of enhancer prediction usually entail identifying as many true positives as possible, while also minimizing false-positives. Using a conservative approach when establishing the

training set is key to high-precision results. A wide range of features can also lead to the best accuracy, but ChIP-seq and ChIP-chip data for a range of transcription factors and chromatin modifications seem to be far and away the most essential information. However, no single feature is a guarantee of enhancer function, and datasets measuring the same feature frequently have limited reproducibility. This has been previously observed with genome-wide studies, but it is not clear how this influences inference of function. Moreover, many features associated with a specific function (e.g. Bicoid occupancy with Anterior-Posterior patterning) can be found binding at genes associated with other regulatory networks. This may indicate functional cross-binding and integration of regulatory networks, or it could suggest promiscuous binding at regions of open chromatin. Zelda, the pioneer transcription factor, is consistently the single most important predictor of enhancer activity. However, Zelda binding is not necessary for accurate predictions. Excluding Zelda has minimal effect on overall prediction accuracy; overall accuracy appears to asymptotically level off after either a small number of highly informative features are used, or a large number of less informative features. It seems unlikely that inclusion of further ChIP-chip and ChIP-seq data would lead to meaningful improvements over the current level of performance. It also is not clear to what extent all ChIP data is serving as a proxy for DNA accessibility in these predictions, as opposed to each protein providing unique information. Although in this study motif enrichment and evolutionary conservation were not the most informative features, it is possible that more in-depth analysis in these directions, or other features that are equally orthogonal to Chromatin Immunoprecipitation data, would be necessary to reach higher levels of precision. High accuracy at predicting activity also did not translate to equally accurate identification of expression patterns. The majority of the expression patterns driven by DNA elements from the Fly Enhancer Resource were not easily categorized into clear-

cut categories; the paradigm of enhancers being specific to a single gene regulatory network, and therefore to a distinct spatial expression pattern, may be biologically inaccurate.

REFERENCES

REFERENCES

1. Small S, Arnosti DN, Levine M. Spacing ensures autonomous expression of different stripe enhancers in the even-skipped promoter. *Development*. 1993;119(3):762-72.

2. Yao L-C, Phin S, Cho J, Rushlow C, Arora K, Warrior R. Multiple modular promoter elements drive graded brinker expression in response to the Dpp morphogen gradient. *Development*. 2008;135(12):2183-2192.

3. Pennacchio LA, Bickmore W, Dean A, Nobrega MA, Bejerano G. Enhancers: five essential questions. *Nat Rev Genet*. 2013;14(4):288-295.

4. Carroll SB. Evo-Devo and an Expanding Evolutionary Synthesis: A Genetic Theory of Morphological Evolution. *Cell*. 2008;134(1):25-36.

5. Wittkopp PJ, Kalay G. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nat Rev Genet*. 2012;13(1):59-69.

6. Shlyueva D, Stampfel G, Stark A. Transcriptional enhancers: from properties to genome-wide predictions. *Nat Rev Genet*. 2014;15(4):272-86.

7. Emilsson V, Thorleifsson G, Zhang B, et al. Genetics of gene expression and its effect on disease. *Nature*. 2008;452(7186):423-428.

8. Epstein DJ. Cis-regulatory mutations in human disease. *Briefings Funct Genomics Proteomics*. 2009;8(4):310-316.

9. Corradin O, Saiakhova A, Akhtar-Zaidi B, et al. Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. *Genome Res* . 2014;24(1):1-13.

10. Halfon MS, Grad Y, Church GM, Michelson AM. Computation-Based Discovery of Related Transcriptional Regulatory Modules and Motifs Using an Experimentally Validated Combinatorial Model. *Genome Res* . 2002;12(7):1019-1028.

11. Stathopoulos A, Drenth M, Erives A, Markstein M, Levine M. Whole-genome analysis of dorsal-ventral patterning in the Drosophila embryo. *Cell*. 2002;111.

12. Narlikar L, Ovcharenko I. Identifying regulatory elements in eukaryotic genomes. *Briefings Funct Genomics Proteomics*. 2009;8(4):215-230.

13. Bailey TL, Elkan C. Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Mach Learn*. 1995;21(1):51-80.

14. Stormo GD. DNA binding sites: representation and discovery . *Bioinformatics*. 2000;16(1):16-23.

15. Sinha S, Blanchette M, Tompa M. PhyME: A probabilistic algorithm for finding motifs in sets of orthologous sequences. *BMC Bioinformatics*. 2004;5(1):170.

16. Elemento O, Slonim N, Tavazoie S. A Universal Framework for Regulatory Element Discovery across All Genomes and Data Types. *Mol Cell*. 2017;28(2):337-350.

17. Junion G, Spivakov M, Girardot C, et al. A Transcription Factor Collective Defines Cardiac Cell Fate and Reflects Lineage History. *Cell*. 2017;148(3):473-486.

18. Fisher WW, Li JJ, Hammonds AS, et al. DNA regions bound at low occupancy by transcription factors do not drive patterned reporter gene expression in Drosophila. *Proc Natl Acad Sci* . 2012;109(52):21330-21335.

19. Graur D, Zheng Y, Price N, Azevedo RBR, Zufall RA, Elhaik E. On the Immortality of Television Sets: "Function" in the Human Genome According to the Evolution-Free Gospel of ENCODE. *Genome Biol Evol*. 2013;5(3):578-590.

20. Spivakov M. Spurious transcription factor binding: Non-functional or genetically redundant? *BioEssays*. 2014;36(8):798-806.

21. Ghavi-Helm Y, Klein FA, Pakozdi T, et al. Enhancer loops appear stable during development and are associated with paused polymerase. *Nature*. 2014;512(7512):96-100.

22. Kleftogiannis D, Kalnis P, Bajic VB. Progress and challenges in bioinformatics approaches for enhancer identification. *Brief Bioinform*. 2016;17(6):967-979.

23. Zinzen RP, Girardot C, Gagneur J, Braun M, Furlong EEM. Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature*. 2009;462(7269):65-70.

24. Wilczynski B, Liu YH., Yeo ZX, Furlong EEM. Predicting Spatial and Temporal Gene Expression Using an Integrative Model of Transcription Factor Occupancy and Chromatin State. *PLoS Comput Biol*. 2012;8:12.

25. Cannavo E, Khoueiry P, Garfield DA, Geeleher P, Zichner T, et al. Shadow enhancers are pervasive features of developmental regulatory networks. *Curr Biol*. 2016;26:1-14.

26. Dresch JM, Arnosti DN. The Wisdom of Crowds: Can Mathematical Models Crack the cis Regulatory Code? *Cell Syst*. 2017;1(6):379-380.

27. Sayal R, Dresch JM, Pushel I, Taylor BR, Arnosti DN. Quantitative perturbation-based analysis of gene expression predicts enhancer activity in early Drosophila embryo. Barkai N, ed. *Elife*. 2016;5:e08445.

28. Frankel N, Davis GK, Vargas D, Wang S, Payre F, Stern DL. Phenotypic robustness conferred by apparently redundant transcriptional enhancers. *Nature*. 2010;466(7305):490-493.

29. Montavon T, Soshnikova N, Mascrez B, et al. A Regulatory Archipelago Controls Hox Genes Transcription in Digits. *Cell*. 2017;147(5):1132-1145.

30. Claussnitzer M, Dankel SN, Kim K-H, et al. FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. *N Engl J Med*. 2015;373(10):895-907.

31. Perry MW, Boettiger AN, Bothma JP, Levine M. Shadow Enhancers Foster Robustness of Drosophila Gastrulation. *Curr Biol*. 2010;20(17):1562-1567.

32. Sanyal A, Lajoie BR, Jain G, Dekker J. The long-range interaction landscape of gene promoters. *Nature*. 2012;489(7414):109-113.

33. Shen Y, Yue F, McCleary DF, et al. A map of the cis-regulatory sequences in the mouse genome. *Nature*. 2012;488(7409):116-120.

34. Whitaker JW, Nguyen TT, Zhu Y, Wildberg A, Wang W. Computational schemes for the prediction and annotation of enhancers from epigenomic assays. *Methods*. 2015;72(C):86-94.

35. Whalen S, Truty RM, Pollard KS. Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nat Genet*. 2016;48(5):488-496.

36. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods*. 2012;9(3):215-6.

37. Williamson I, Berlivet S, Eskeland R, et al. Spatial genome organization: contrasting views from chromosome conformation capture and fluorescence in situ hybridization. *Genes Dev* . 2014;28(24):2778-2791.

38. Mora A, Sandve GK, Gabrielsen OS, Eskeland R. In the loop: promoter–enhancer interactions and bioinformatics. *Brief Bioinform*. 2016;17(6):980-995.

39. Markstein M, Zinzen R, Markstein P, Yee KP, Erives A, Stathopoulos A. A regulatory code for neurogenic gene expression in the Drosophila embryo. *Development*. 2004;131.

40. Zeitlinger J, Zinzen RP, Stark A, et al. Whole-genome ChIP-chip analysis of Dorsal, Twist, and Snail suggests integration of diverse patterning processes in the Drosophila embryo. *Genes Dev*. 2007;21(4):385-390.

41. Fu W, Duan H, Frei E, Noll M. shaven and sparkling are mutations in separate enhancers of the Drosophila Pax2 homolog. *Development*. 1998;125(15):2943-50.

42. Wang C, Zhang MQ, Zhang Z. Computational Identification of Active Enhancers in Model Organisms. *Genomics Proteomics Bioinformatics*. 2013;11(3):142-150.

43. Arnold CD, Gerlach D, Stelzer C, Boryn LM, Rath M, Stark A. Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq. *Science (80-)*. 2013;339(6123):1074-1077.

44. Arnold CD, Zabidi MA, Pagani M, et al. Genome-wide assessment of sequence-intrinsic enhancer responsiveness at single-base-pair resolution. *Nat Biotech*. 2017;35(2):136-144.

45. Zabidi MA, Arnold CD, Schernhuber K, et al. Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. *Nature*. 2015;518(7540):556-559.

46. Emberly E, Rajewsky N, Siggia ED. Conservation of regulatory elements between two species of Drosophila. *BMC Bioinformatics*. 2003;4(1):57.

47. Glazov EA, Pheasant M, McGraw EA, Bejerano G, Mattick JS. Ultraconserved elements in insect genomes: A highly conserved intronic sequence implicated in the control of homothorax mRNA splicing. *Genome Res* . 2005;15(6):800-808.

48. Marcovitz A, Jia R, Bejerano G. "Reverse Genomics" Predicts Function of Human Conserved Noncoding Elements. *Mol Biol Evol*. 2016;33(5):1358-1369.

49. Halligan DL, Keightley PD. Ubiquitous selective constraints in the Drosophila genome revealed by a genome-wide interspecies comparison. *Genome Res*. 2006;16(7):875-884.

50. Hare EE, Peterson BK, Iyer VN, Meier R, Eisen MB. Sepsid even-skipped enhancers are functionally conserved in Drosophila despite lack of sequence conservation. *PLoS Genet*. 2008;4(6):6.

51. Zhang B, Pan X, Cannon CH, Cobb GP, Anderson TA. Conservation and divergence of plant microRNA genes. *Plant J*. 2006;46(2):243-259.

52. Ahituv N, Zhu Y, Visel A, et al. Deletion of Ultraconserved Elements Yields Viable Mice. *PLOS Biol*. 2007;5(9):e234.

53. The modENCODE Consortium, Roy S, Ernst J, et al. Identification of Functional Elements and Regulatory Circuits by Drosophila modENCODE. *Science (80-).* 2010;330(6012):1787 LP-1797.

54. Lieberman-Aiden E, van Berkum NL, Williams L, et al. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science (80-).* 2009;326(5950):289 LP-293.

55. Li G, Fullwood MJ, Xu H, et al. ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. *Genome Biol*. 2010;11(2):R22.

56. Erwin GD, Oksenberg N, Truty RM, et al. Integrating Diverse Datasets Improves Developmental Enhancer Prediction. *PLoS Comput Biol*. 2014;10(6):6.

57. Capra JA. Extrapolating histone marks across developmental stages, tissues, and species: an enhancer prediction case study. *BMC Genomics*. 2015;16(1):104.

58. Li Y, Shi W, Wasserman WW. Genome-Wide Prediction of cis-Regulatory Regions Using Supervised Deep Learning Methods. *bioRxiv*. February 2016.

59. He B, Chen C, Teng L, Tan K. Global view of enhancer–promoter interactome in human cells. *Proc Natl Acad Sci* . 2014;111(21):E2191-E2199.

60. Kleftogiannis D, Kalnis P, Bajic VB. DEEP: a general computational framework for predicting enhancers. *Nucleic Acids Res*. 2015;43(1):e6-e6.

61. Griffon A, Barbier Q, Dalino J, van Helden J, Spicuglia S, Ballester B. Integrative analysis of public ChIP-seq experiments reveals a complex multi-cell regulatory landscape. *Nucleic Acids Res*. 2015;43(4):e27-e27.

62. van Duijvenboden K, de Boer BA, Capon N, Ruijter JM, Christoffels VM. EMERGE: a flexible modelling framework to predict genomic regulatory elements from genomic signatures. *Nucleic Acids Res*. 2016;44(5):e42-e42.

63. Gurdziel K, Vogt KR, Schneider G, Richards N, Gumucio DL. Computational prediction and experimental validation of novel Hedgehog-responsive enhancers linked to genes of the Hedgehog pathway. *BMC Dev Biol*. 2016;16(1):4.

64. Fletez-Brant C, Lee D, McCallion AS, Beer MA. kmer-SVM: a web server for identifying predictive regulatory sequence features in genomic data sets. *Nucleic Acids Res*. 2013;41.

65. Chae M, Danko CG, Kraus WL. groHMM: a computational tool for identifying unannotated and cell type-specific transcription units from global run-on sequencing data. *BMC Bioinformatics*. 2015;16(1):222.

66. Coppola CJ, C. Ramaker R, Mendenhall EM. Identification and function of enhancers in the human genome. *Hum Mol Genet*. 2016;25(R2):R190-R197.

67. Kvon EZ, Kazmar T, Stampfel G, et al. Genome-scale functional characterization of Drosophila developmental enhancers in vivo. *Nature*. 2014;512(7512):91-5.

68. Rosenbloom KR, Armstrong J, Barber GP, et al. The UCSC Genome Browser database: 2015 update. *Nucleic Acids Res*. 2015;43(D1):D670-D681.

69. Visel A, Bristow J, Pennacchio LA. Enhancer identification through comparative genomics. *Semin Cell Dev Biol*. 2007;18(1):140-152.

70. Zhu LJ, Christensen RG, Kazemian M, Hull CJ, Enuameh MS, Basciotta MD. FlyFactorSurvey: a database of Drosophila transcription factor binding specificities determined using the bacterial one-hybrid system. *Nucleic Acids Res*. 2011;39.

71. Bailey TL, Gribskov M. Combining evidence using p-values: application to sequence homology searches. *Bioinformatics*. 1998;14(1):48-54.

72. Gallo SM, Gerrard DT, Miner D, et al. REDfly v3.0: toward a comprehensive database of transcriptional regulatory elements in Drosophila. *Nucleic Acids Res*. 2011;39(suppl_1):D118-D123.

73. Gittelman RM, Hun E, Ay F, et al. Comprehensive identification and analysis of human accelerated regulatory DNA. *Genome Res*. 2015;25(9):1245-1255.

74. MacArthur S, Li XY, Li J, Brown JB, Chu HC, Zeng L. 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*. 2009;10.

75. Liu Y, Pelham-Webb B, Di Giammartino DC, et al. Widespread Mitotic Bookmarking by Histone Marks and Transcription Factors in Pluripotent Stem Cells. *Cell Rep*. 2017;19(7):1283-1293.

76. Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A*. 2010;107.

77. St Johnston D, Nusslein-Volhard C. The origin of pattern and polarity in the Drosophila embryo. *Cell*. 1992;68.

78. Park PJ. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet*. 2009;10.

79. Devailly G, Mantsoki A, Michoel T, Joshi A. Variable reproducibility in genome-scale public data: A case study using ENCODE ChIP sequencing resource. *FEBS Lett*. 2015;589(24):3866-3870.

CHAPTER III: AN IMAGE DATABASE OF *DROSOPHILA MELANOGASTER* WINGS FOR PHENOMIC AND BIOMETRIC ANALYSIS

The work described in this chapter was published in the following manuscript: Sonnenschein, A., VanderZee, D., Pitchers, W.R., Chari, S., and Dworkin, I. (2015). An image database of *Drosophila melanogaster* wings for phenomic and biometric analysis. *GigaScience*, *4*(25).

Abstract

Extracting important descriptors and features from images of biological specimens is an ongoing challenge. Features are often defined using landmarks and semi-landmarks that are determined *a priori* based on criteria such as homology or some other measure of biological significance. An alternative, widely used strategy uses computational pattern recognition, in which features are acquired from the image *de novo*. Subsets of these features are then selected based on objective criteria. Computational pattern recognition has been extensively developed primarily for the classification of samples into groups, whereas landmark methods have been broadly applied to biological inference. To compare these approaches and to provide a general community resource, we have constructed an image database of *Drosophila melanogaster* wings - individually identifiable and organized by sex, genotype and replicate imaging system - for the development and testing of measurement and classification tools for biological images. We have used this database to evaluate the relative performance of current classification strategies. Several supervised parametric and nonparametric machine learning algorithms were used on principal components extracted from geometric morphometric shape data (landmarks and semi-landmarks). For comparison, we also classified phenotypes based on *de novo* features extracted from wing images using several computer vision and pattern recognition methods as implemented in the Bioimage Classification and Annotation Tool (BioCAT). Because we were able to thoroughly evaluate these strategies using the publicly available *Drosophila* wing database, we believe that this resource will facilitate the development and testing of new tools for the measurement and classification of complex biological phenotypes.

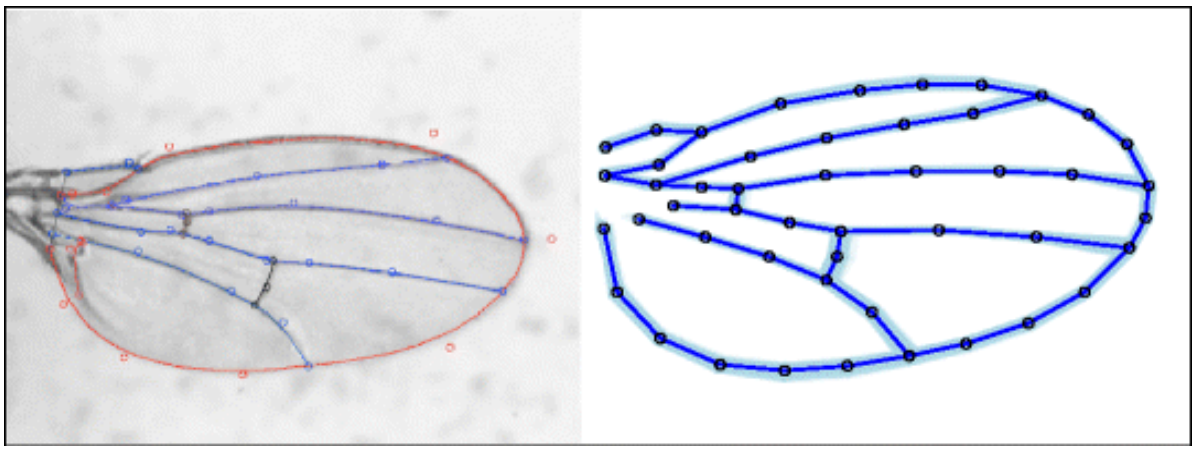
Introduction

Understanding the causes and consequences of phenotypic variation is a unifying goal across

many biological disciplines. One aim of phenomics is to comprehensively measure this variation. However, biological traits are complex and multidimensional and this presents challenges for both measurement and analysis[1]. The complete 'phenome' of an individual includes more phenotypes than can realistically be measured and the most informative subset of measurable features is not necessarily known, even for specific traits[2]. Manually selected features benefit from prior knowledge of the biological system, whereas computationally selected image properties are generally optimized for discrimination between groups. However, it is not clear how these strategies compare in their classification of images into groups (sex, genotype, species) or in their potential to derive broader biological inferences.

Geometric morphometrics and computational pattern recognition represent very different strategies for extracting and quantifying phenotypes from image data. Geometric morphometrics measures shape by using homologous landmarks (or curves) across specimens as features[3,4]. Methodologically, these landmarks are determined *a priori*based on biological considerations of both homology and potential informativeness. Information about the shape of the specimen is extracted from the configuration by removing variation in size, location and orientation of the specimen, resulting in an explicit geometric representation of shape (Fig. 3-1)[5-7].

Figure 3-1 Wing landmarks and semi-landmarks



Example wing image from *D. melanogaster* that has been splined using WINGMACHINE. After landmark and semi-landmark data is extracted, data is translated (centered to origin), scaled by centroid size and superimposed (Procrustes superimposition for landmarks) data all lies in a common subspace. Image represents 50 individual configurations from specimens to demonstrate some of the variation among individuals.

Computational pattern recognition represents a school of alternative approaches, in which features are extracted from image data with computer vision tools[8,9]. Pattern recognition uses features such as the statistical distribution of pixels, or descriptions of texture or edges. A subset of informative features is generally selected based on an objective function, such as the classification of samples into groups, often using machine learning techniques[9]. The degree of informativeness of these features is usually assessed by cross-validation. Whereas geometric morphometrics requires a comprehensive understanding of the biological relevance and evolutionary history of the feature, computational pattern recognition can be applied without prior knowledge and can also detect informative patterns that are not visually perceptible[9].

Both geometric morphometrics and computational pattern recognition have practical applications in biological research. There have been varying levels of success using two-dimensional and three-dimensional cranial-facial morphometric phenotypes to infer the genetic causes of disease[10-,12] and to track disease progression[13]. Morphometrics and computational pattern recognition have also been successfully used with machine learning algorithms to classify complex morphological phenotypes by species[14-,16]. Similarly, computer vision and pattern recognition have been crucial in the development of tools for the related field of biometrics[17], which uses phenotypes to distinguish individuals. Biometrics tools may be useful for interpreting phenomics data, thereby extending the amount of informative variation that can be extracted from biological images.

A potential biological application for biometrics is the interpretation of *Drosophila* wing shape. Wing shape is an established model system for phenomics[1,18], the genetic basis of shape[19- 21] and for phenotypic evolution[22,23]. Although *Drosophila* wings can be evaluated qualitatively[24] or by metrics such as length and surface area[25], they are often measured within a geometric

morphometric framework[14,19,26,27]. Landmarks are based on vein intersections[26,27] with semi-landmarks defining curves (Fig. 1)[14,23]. Biometric facial recognition tools have had some success at classifying images of *Drosophila* wings into biological categories[28,29]. The 'eigenface' method, which is a classic technique for facial recognition, has been modified into 'eigenwings' using features extracted from *Drosophila* wings to classify individuals by their sex[29]. Another facial recognition method that uses a genetic algorithm to select texture features[30] has also been used with similar goals[28-30], with up to a 94 % successful classification rate[29].

The success of facial recognition programs that rely on texture features instead of vein positioning raises the questions of what other features might also be useful for classifying *Drosophila* wings and how tools that are already used in biometrics may be applied to phenomics datasets. However, our ability to evaluate different approaches − whether for classification or biometric identification, as well as for long-term goals of further biological inference − remains limited by the lack of open databases of images for comparison.

In this article, we describe the creation and implementation of a database of wing images from *Drosophila melanogaster* for the development and testing of such methods. The database was designed to include multiple levels of replication encompassing both biological and technical variation. It allows the assessment of variation and classification by genotype, sex and individual identity (right and left wings from the same fly). To introduce sources of technical noise common to biological images, it includes several images of each wing, captured on various microscopes and at multiple magnifications.

Using landmark and semi-landmark measurements extracted from images in this database, we have analyzed the relative success of a number of machine learning algorithms at classifying *Drosophila* genotype and sex. We compare the success of these methods with the

performance of the same classifier algorithms using features extracted by the Bioimage Classification and Annotation Tool (BioCAT), a pattern-recognition program designed for image analysis[31]. The database of images, landmark data and all source code have been made publicly available to serve as a resource for the testing and development of biometrics tools.

Methods

Fly genetics and sample preparation

Fly stocks were obtained from the Bloomington Stock Center. These lines include wing mutations in the genes *Epidermal growth factor receptor* (*Egfr*), *mastermind* (*mam*), *thickveins* (*tkv*) and *Star* (*S*; see Table 3-1 for allele and stock information). All four mutations are caused by insertions of P-element transposable elements, each marked with a mini-w^+resulting in partial rescue of wild-type red eyes. The wild-type strain used as a background was an isogenic Samarkand (SAM), marked with a w^- mutation to enable identification of the mutant alleles[38].

Each P-element-bearing strain was initially introgressed into SAM by repeatedly backcrossing into the SAM background genotype (as described in[27]). These have since been maintained heterozygous balanced over a CyO (also in the SAM background) with the exception of the *tkv* mutant, which was maintained as a homozygote. Before initiating the experiment, these flies were maintained for one generation in an incubator (Percival Model : I41VLC8 set to 24 °C, 65 % relative humidity and a 12-hour light/dark cycle) to acclimatize them to the environment. Under these same growth conditions, the lines carrying mutant alleles were then backcrossed for two additional generations into the SAM wild-type background prior to rearing flies for data collection for the database. Because of the extensive back-crossing, each mutant-bearing strain is close to co-isogenic to the SAM wild type, with the exception of the focal allele and a small

genomic fragment in linkage disequilibrium to that allele.

For each mutant strain, populations were expanded in five replicate bottles. Each bottle contained 10 mutant males (red-eyed) crossed to 20 SAM virgin females (white-eyed). This experimental design was also applied to the SAM control, with 10 SAM males and 20 SAM virgin females. The flies were allowed to lay eggs for 4 days, after which the adults were discarded. After 7 days, paper towel was added to the bottles to soak up excess moisture and provide additional substrate for pupation. From days 14-18, emerging flies were phenotyped (based on the w^+ marker) and sexed. They were stored separately by sex and genotype in microtubes containing 70 % ethanol at room temperature for wing dissections. Fly wings were dissected in phosphate buffered saline (PBS) and mounted on slides in a solution of 70 % glycerol and 30 % PBS. All wings dissections were performed by the same person (AS). If one wing was torn or damaged, both wings from that fly were discarded.

Imaging

Each wing was imaged at 20× and 40× magnification on both an Olympus BX51 and Leica M125 microscope, using the DP controller (V.3,1,1208) and Leica App Suite (V.3) imaging software respectively. Two individuals imaged wings, one using the Olympus microscope (AS) and the other the Leica (DV) microscope. All images have unique names, using the format 'genotype_sex_side_microscope_magnification_fly-number'. If images contained tears or folds in the wing or indicated errors in dissection or mounting, all images from that fly were discarded.

Geometric morphometric data acquisition and preparation

Images were first converted to grayscale and cropped with the Gnu Image Manipulation Program (GIMP, version 2.8[39]) in batches using the David's Batch Processor plugin (version 1.1.8[40]). Two starting landmarks were manually labeled at the humeral break and alula notch, using tpsDig

(version 2.17[41]). For more details on re-sizing and cropping, see Additional file 1: Supplementary Methods. WINGMACHINE (Wings version 3.7.2[14]) was used to generate wing splines, which were manually reviewed and adjusted as necessary. CPR (version 1.01r[42]) was used to scale wings by centroid size, perform a Procrustes superimposition and extract landmark and semi-landmark coordinates. Further details on processing of this data are available in Appendix: Supplementary Methods. All further statistical analysis was done in R (version 3.1.0[35]) on images of wings taken at 40× magnification on the Olympus BX51 microscope. Scripts can be found at the Dworkin Lab github page[43] and together with the data at GigaDB[44].

Procrustes coordinate values and centroid for left and right wings from the same fly were averaged using the R *plyr* package (V.1.8.1). In total, this included 12 two-dimensional landmarks and 36 semi-landmarks. However, because of image registration, scaling and Procrustes superimposition, four dimensions do not contain any information. Furthermore, the semi-landmarks are constrained to slide along a curve and therefore have approximately one degree of freedom. This results in approximately 58 dimensions of potential data. Thus, the first 58 principal components contributing to shape (excluding centroid) were extracted and used for all further analyses.

Morphometric analysis

Two-thirds of the samples from each genotype were defined as the training set and one-third as a testing set. These were used to train and test 'lda' and 'qda' functions from the MASS package (V. 7.3-33), 'mda' and 'fda' functions from the *mda* package (V. 0.4-4), the 'bagging' function from the *adabag* package (V.3.2), random forest from the *randomForest*package (using 500 trees, version 4.6-7), the 'svm' function from the *e1071* package (version 1.6-3) and a neural network from the *nnet* package (V. 7.3-8). K-nearest neighbors (KNN) from the *class* package

(V. 7.3-10) was also tested, using k values from 1 to 100. Confidence intervals were approximated by re-sampling the training and testing sets over 1,000 repetitions. All functions were used with default arguments, with the exception of 'svm', 'knn' and random forest. 'svm' was optimized for kernel function shape and 'knn' for the value of k. Random forests were tested over a range of 10-1,000 trees.

SVM and random forest (using 10 trees and 1,000 trees) were repeated on several subsets of the original dataset, using only left wings (prior to averaging) and left-female wings, to facilitate comparisons with the BioCAT results. The same analysis was also used to classify wings by sex, using only wings from the SAM (wild-type) genotype.

BioCAT analysis

For the BioCAT analysis[14], we used a Fisher feature selector to identify 50 eigenvalues of the Hessian matrix[37] from 144 left male and female wings (77 each) from the SAM genotype. BioCAT applied these features to train an SVM classifier and two random forest classifiers: one with 10 trees and one with 1,000 trees. These models were used to annotate 30 male and 30 female left SAM wings that were not included in the training set. Annotation accuracy was determined by counting the number of correct and incorrect classifications. Although BioCAT allows for cross-validation during combined training/testing, the quantity and size of our data made re-sampling for confidence intervals infeasible. This process was repeated for classification by genotype using 70 left-female wings from each of the five genotypes as a training set for classification by genotype. The genotype classification models were tested on 30 left-female wings from each genotype. Images used as training and testing sets have been organized with their respective models at the database[44].

Results

The Drosophila wing database comprises a large number of high-quality wing images and contains both biological and technical variation. Sources of biological variation include genotype (there are four mutant genotypes (listed in Table 3-1) in the wild-type background of Samarkand (SAM), as well as the SAM wild-type background itself). In addition, sex and within-individual (left and right wings) variation is included. There are 100-130 individual samples for each combination of biological variables (Table 3-2). The mutant genotypes included in the database are heterozygous loss-of-function mutations for the genes that encode the Epidermal growth factor receptor (*Egfr*), mastermind (*mam*), Star (*S*) and thickveins (*tkv*) (see Table 3-1 for allele information, Fig. 3-2 for the relative impact of each mutation on phenotype and Methods for additional details). As heterozygotes, these mutations all have quantitative effects on shape, although they are qualitatively indistinguishable from the wild-type background (Fig. 3-3). The mutations represent perturbations of multiple signaling pathways: for example, *tkv* is a receptor kinase in the Transforming growth factor-β (TGF-β) pathway[32] and *mam* is a transcription factor in the Notch signaling pathway[33]. *Egfr* and *Star* genetically interact as *Star* modulates signaling through the Egfr pathway[32, 34]. These specific mutations were selected because previous studies have shown that when heterozygous, they have a range of quantitative effects on wing shape[27].

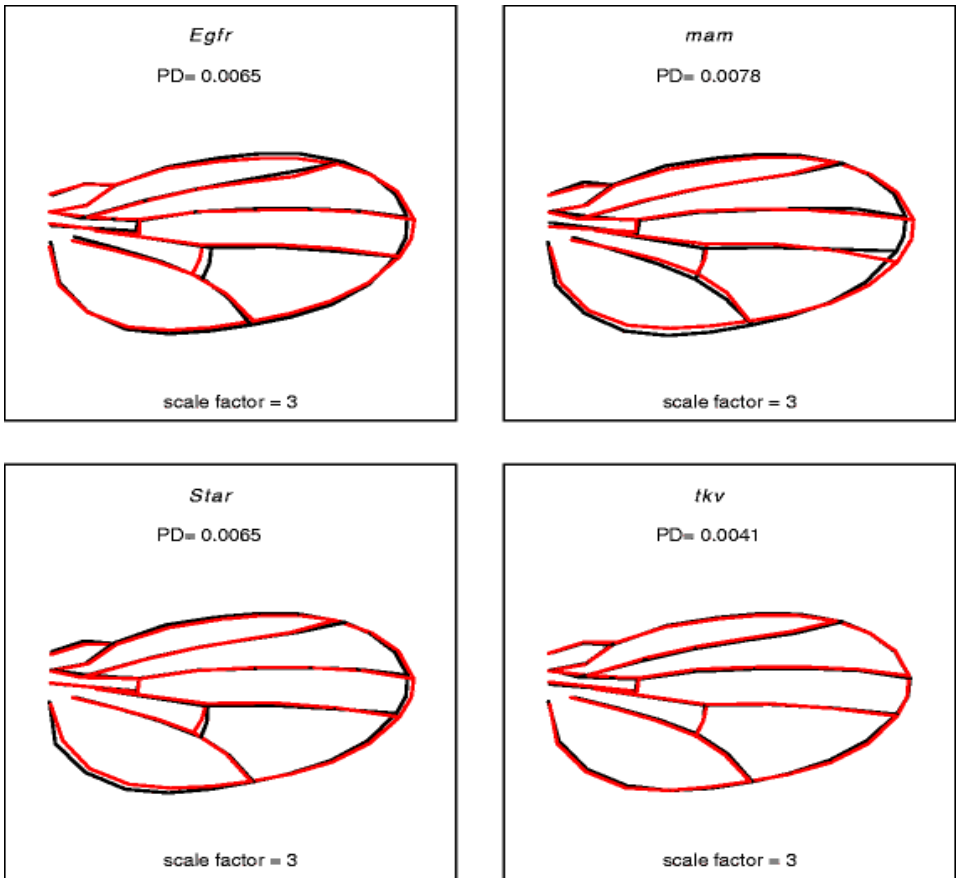
Table 3-1 *Drosophila* allele information

Bloomington #	Gene name	Gene symbol	Allele name
10385	*Epidermal growth factor receptor*	*Egfr*	P{lacW}*Egfr*^{k05115}
14189	*mastermind*	*mam*	P{SUPor-P}*mam*^{kG02641}
10418	*Star*	*S*	P{lacW}*S*^{k09530}
14403	*thickveins*	*tkv*	P{SUPor-P}*tkv*^{KG01923}

Table 3-2 *Drosophila* wings dissected by sex and genotype

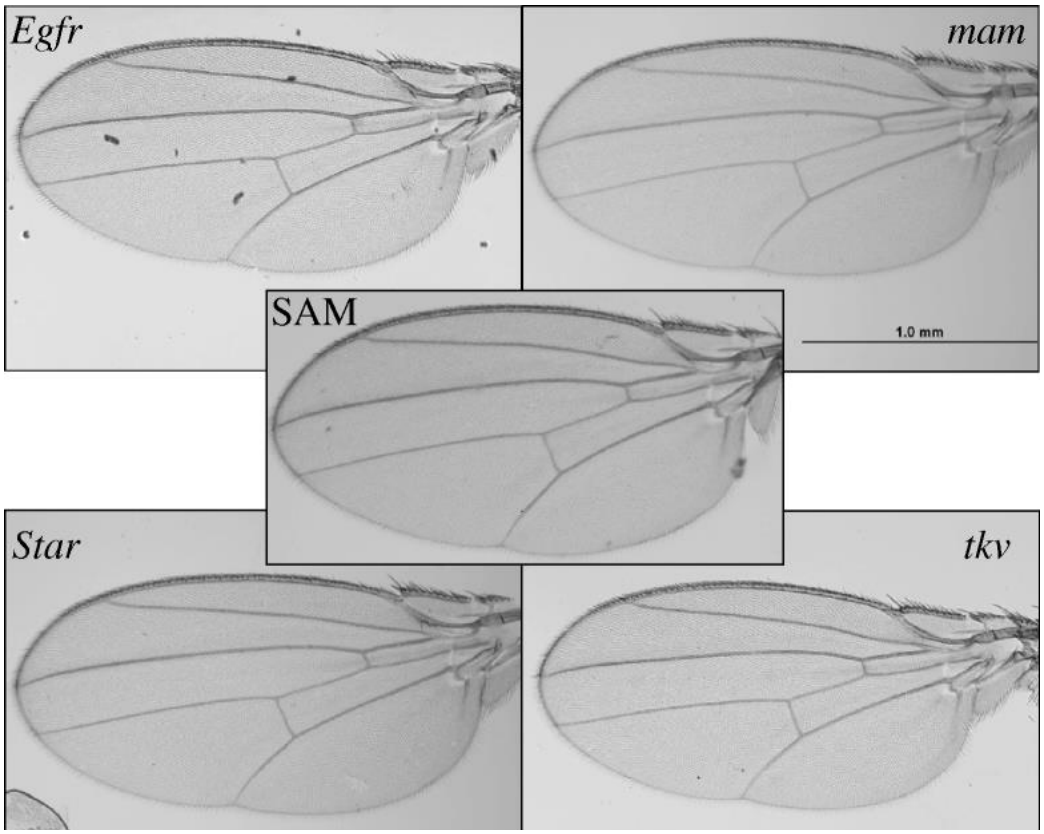
Genotype	Female	Male
Egfr	116	118
mam	106	130
Samarkand (SAM)	107	100
Star	115	111
tkv	116	116

Figure 3-2 Mutation effects



Magnitude and direction of the effect of each mutation (*red*) relative to Samarkand wild type (*black*). Magnitudes are in units of Procrustes distance (PD), which for this (tangent approximation) is equivalent to the Euclidean distance between the mean vector of each mutant and the Samarkand (SAM) wild type. The vectors of shape differences are magnified three-fold.

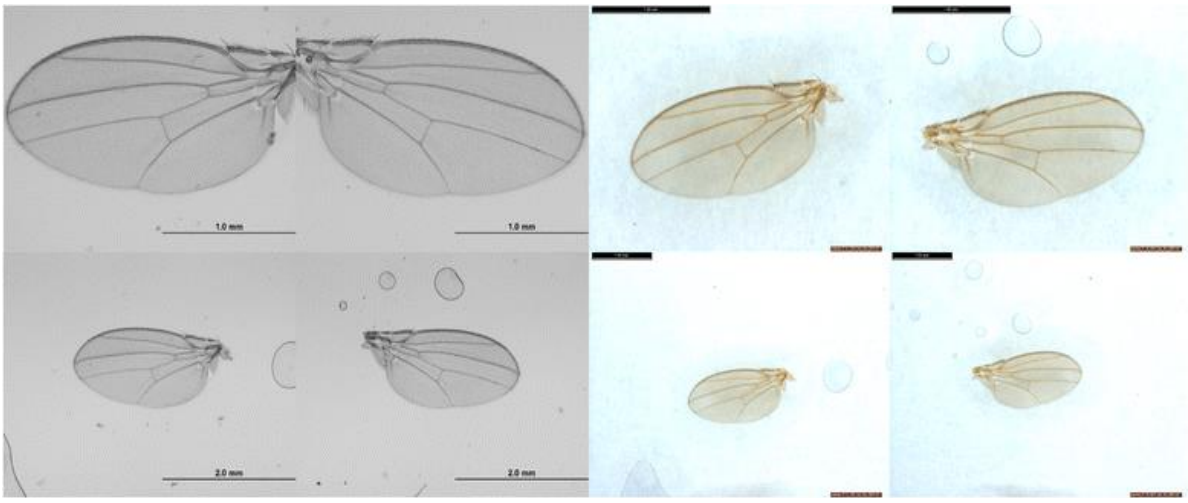
Figure 3-3 Heterozygote mutant wings



Representative images from the database. From right top corner counter-clockwise: *mastermind*, *Epidermal growth factor receptor*, *Star* and *thickveins. mastermind*, *Egfr* and *Star*mutations are all homozygous lethal and *thickveins* has a qualitative defect as a homozygote. As heterozygotes, they are qualitatively indistinguishable from the Samarkand (SAM) wild type (center)

To allow researchers to compare various classification algorithms across a range of technical conditions, the measurements were subject to technical variation including the microscope and software used to capture images and the magnification setting of the microscope. Each wing in the database was imaged on two different microscope models, at both 40× and 20× magnification, so each wing in the database was imaged a total of four times (see Fig.3-4: left and right wings from the same fly imaged under all four technical variation conditions). Also included in the database is landmark and semi-landmark coordinate information extracted from all images in the database using WINGMACHINE software[14], so information extracted from these images can be compared with existing standards for wing analysis. We also repeated the morphometrics analysis (landmarking, fitting curves (splining), editing splines and superimposition) for a small subset of the wing images (50 left and right female wings from two genotypes), to provide information on the technical variation in this process.

Figure 3-4 Technical replicates conditions



Left and right wings from the same female (SAM) fly, imaged four times. Top left are images taken on Olympus BX51 microscope at 40× magnification, top right are taken on Leica M125 at 40× magnification. Bottom left and right are images taken at 20× magnification

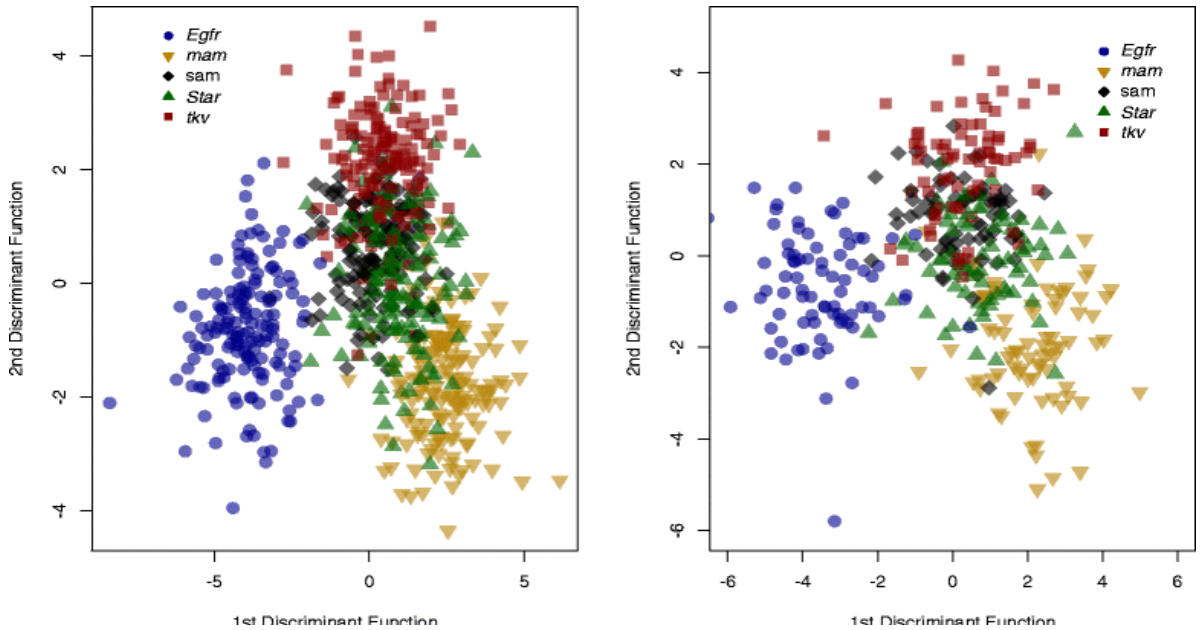
Classification based on geometric morphometric data shows a high degree of accuracy

Although the primary goal of this project was to develop the image database, we also wished to provide future users with some baseline data to evaluate classifiers. We used a wide range of supervised machine learning algorithms in R (version 3.1.0[35]) for classification based on the landmark and semi-landmark data extracted from the images. All data for this analysis was from images taken on an Olympus BX51 microscope at 40× magnification. The data analyzed included the 58 principal components of shape generated from the landmark and semi-landmark coordinates (representing all non-zero eigenvalues). We chose algorithms to represent a wide range of models; standard errors were estimated by re-sampling training and testing sets. When classifying wings within a common genotype (SAM) by sex, all algorithms except for quadratic discriminant analysis (QDA) were able to predict the sex of a test set with more than 95 % accuracy (Table 3). When classifying wings by both genotype and sex, linear discriminant analysis (LDA), flexible discriminant analysis (FDA) and mixture discriminant analysis (MDA) were all able to correctly categorize test data with 85 % accuracy or higher. Support vector machines (SVM) and neural networks were also accurate with over 80 % of wings (Table 3-3). The high accuracy of most methods, especially of LDA (Fig. 3-5), suggests that classifications using this data, based on both sex and genotype, are robust to assumptions of linearity and common covariance matrices between factors (genotype and sex)[36].

Table 3-3 Classification accuracy of machine learning algorithms using landmarks

Algorithm	Sex (± Standard error)	Genotype (± Standard error)
LDA	98.2 % (±1.6)	86.1 % (±1.5)
QDA	81.5 % (±6.4)	68.7 % (±2.2)
FDA	98.2 % (±1.6)	86.0 % (±1.5)
MDA	98.1 % (±1.6)	84.8 % (±1.6)
Bagging	93.3 % (±2.9)	57.6 % (±2.9)
Random forest	94.6 % (±2.7) 100 trees	74.9 % (±2.1) 1,000 trees
SVM	96.8 % (±2.1) sigmoid	83.8 % (±1.6) radial
Neural network (size 10)	98.3 % (±1.6)	81.2 % (±2.2)
KNN	98.3 % (±1.5) k = 4	59.3 % (±2.1) k = 32

Figure 3-5 Separation of specimens using landmark data using LDA



Separation of specimens for each of the five genotype by linear discriminant analysis (LDA) in training set (left panel) and testing set (right panel), plotting the first discriminant function by the second discriminant function. This includes data for both males and females, but averaged (left and right wings) per specimen

Computational feature detection and sub-setting for classification using BioCAT

We tested several methods of classification using the image analysis software BioCAT[31], which allows combinations of feature selectors, extractors and classifiers. Using the Fisher feature selection criterion, we tested several combinations of features and classification algorithms (Appendix: Supplementary Methods). After training a random forest classifier with 50 FeatureJ Hessians[37] extracted from a training set of wing images, we were able to classify individuals by sex (in a common genotype) in a test set of wing images with 85 % accuracy (Table 3-4). Classification of wing images by genotype (within sex) had an accuracy of only up to 52 %, although this is higher than the 20 % success rate that would be expected for random classification.

Comparisons between BioCAT and geometric morphometric descriptors for classification

BioCAT feature selectors act on raw images and therefore had access to both shape and size information for wings, whereas morphometric analyses were performed after scaling by centroid. When the parameter for centroid size was included with landmark and semi-landmark coordinates in morphometric analysis, the relative effectiveness of different algorithms was largely the same (Table 3-4), although classification accuracy generally increased for both sex and genotype.

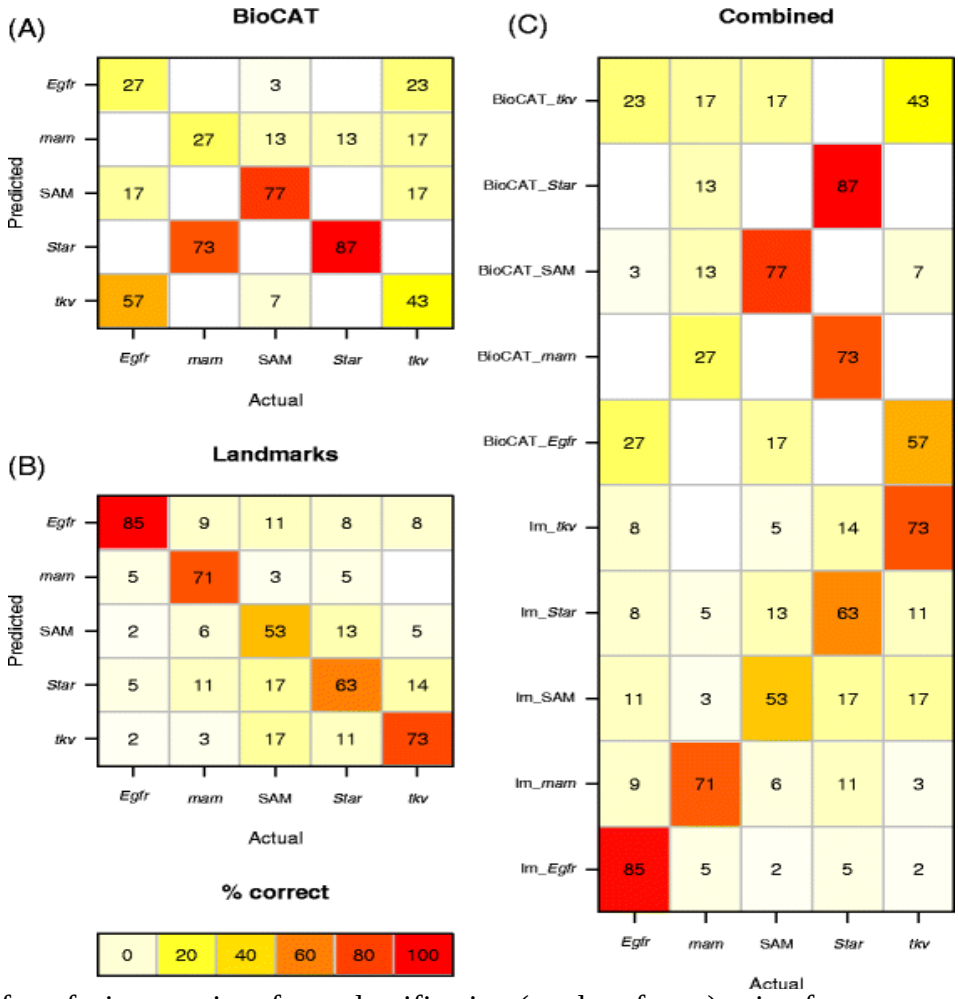
Classification based on features extracted by BioCAT and those using landmarks and semi-landmark coordinates differed in the distribution of classification errors between genotypes. BioCAT classified some genotypes far more consistently than others and errors frequently skewed towards a particular genotype (Fig. 3-6). Notably, *mastermind* was misclassified as *Star* 90 % of the time (27/30 mis-identifications in the test set). There is no similar trend of *Star* and *mam* phenotype mis-identification in classifications based on landmarks (Fig. 3-6).

Table 3-4 Classification accuracy of machine learning algorithms compared with BioCAT

Classification	Algorithm	Hessian	Shape	Shape + Size
Sex	Random forest (10)	85.0 %	92.3 % (±3.7)	94.7 % (±2.6)
	Random forest (1,000)	85.0 %	96.1 % (±2.2)	95.9 % (±2.1)
	SVM	81.7 %	99.0 % (±1.2)	99.0 % (±1.2)
Genotype	Random forest (10)	52.0 %	43.3 % (±3.5)	44.7 % (±3.7)
	Random forest (1,000)	46.7 %	69.1 % (±3.4)	70.2 % (±2.8)
	SVM	43.3 %	75.1 % (±2.8)	75.8 % (±2.7)

Hessian column represents accuracy of classifications based on Hessian features extracted with BioCAT. Shape column represents classification accuracy based on landmarks and semi-landmarks, not including centroid. Shape + size represents classification accuracy based on landmarks and semi-landmarks, including centroid

Figure 3-6 Confusion matrices



Heatmap of confusion matrices from classification (random forest) using features extracted using BioCAT **(a)** compared with landmark and semi-landmark data **(b)**. The data in **(a)** and **(b)** is shown together in **(c)** to facilitate comparison. Numbers represent percentage of correct classifications. lm_* represent the landmark/semi-landmark data. BioCAT features were mis-classified more consistently as some genotypes, e.g. mis-classification of *mam* mutants as *Star* **(a)**. This pattern is not evident in the classification using the landmark data **(b)**. Scale represents frequency of classification

Comparisons between BioCAT and geometric morphometric descriptors for classification

Both geometric morphometric methods and BioCAT were able to classify images by sex across sources of technical variation (i.e. images taken on different microscopes). Geometric morphometric methods showed very little loss in accuracy when classifying wings at the same magnification across microscopes (where an LDA was trained on images taken on the Olympus at 40× magnification and tested on images from the Leica microscope at the same magnification). However, accuracy dropped substantially from 98.2 % to 81.2 % when the LDA was trained on images taken on the Olympus at 40× magnification and tested on images taken on the same microscope at 20× magnification (Table 3-5). Images from each microscope and magnification were superimposed separately and simultaneous superimposition might substantially increase the accuracy of classification across datasets.

Machine learning algorithms using landmark and semi-landmark features, compared with Hessian features extracted by BioCAT, trained and tested across microscopes and magnifications was not able to make accurate classifications across datasets using unedited images. When trained on images taken on the Olympus at 40×, it uniformly classified wings in images taken on the Olympus at 20× magnification as males (using Hessian features and a random forest classifier).

Table 3-5 Classification accuracy of machine learning algorithms compared with BioCAT

Method	Training images	Testing images	Sex (± SE)
BioCAT	Olympus 40×	Olympus 40×	85.0 %
	Olympus 40×	Olympus 20×	50.0 %
	Olympus 40× cropped	Olympus 20× cropped	50.0 %
	Leica 40× cropped	Leica 40× cropped	93.0 %
	Olympus 40× cropped	Leica 40× cropped	73.7 %
	Olympus & Leica 40× cropped	Olympus 40× cropped	73.3 %
	Olympus & Leica 40× cropped	Leica 40× cropped	86.0 %
Landmarks	Olympus 40× landmarks	Olympus 40× landmarks	98.2 % (±1.6)
	Olympus 40× landmarks	Olympus 20× landmarks	81.2 % (±1.4)
	Leica 40× landmarks	Leica 40× landmarks	97.8 % (±0.69)
	Olympus 40× landmarks	Leica 40× landmarks	79.1 % (±1.3)

Using images cropped to the same dimensions as used for measuring splines (cropping images was also necessary for geometric morphometric analysis), BioCAT still had difficulty classifying across magnifications, but was able to correctly identify sex from wing images taken at the same magnification on the Leica microscope with 73.7 % accuracy (Table 5). Interestingly, BioCAT performed better on images taken on the Leica - when both trained and tested on images taken on the Leica at a common magnification, it classified images by sex with 93 % accuracy, relative to 85 % accuracy when classifying images from the Olympus microscope (Table 3-5).

Discussion

Although this database is primarily intended to serve as a resource for the development and testing of measurement tools, we also investigated whether the image collection could provide insights into the comparative effectiveness of existing pattern recognition and morphometrics methods. In particular, we compared *a priori* biologically informed landmark data as features analyzed within a geometric morphometrics framework with *de novo* feature extraction,

identification and optimization. Using both types of features, we evaluated the classification of wings by genotype and by sex and the accuracy of various statistical learning methods. The performance of the classifiers based on landmark data was generally superior with respect to classifying test data. For a number of reasons, this success must be considered within the context of the methods examined in this study. The availability of the database now provides a test bed for further refinement.

Computational pattern recognition and morphometrics software are likely to extract different features. In addition to considering how well geometric morphometric approaches compare to 'computer vision' *de novo* feature extraction (see below), it is also worth comparing the efficiency with which the feature data can be obtained. The WINGMACHINE pipeline was designed with a single goal and has been optimized for extracting landmark and semi-landmark data from *Drosophila* wings. By contrast, BioCAT was designed (and therefore chosen for this study) for its accessibility and flexibility, which allow it to be immediately applied to raw wing images. The FeatureJ features extracted with BioCAT describe image texture[31], whereas a geometric morphometric approach uses biologically defined, homologous landmarks and curves as features, defined by vein intersections and outlines[14]. Extracting the large number of landmarks and semi-landmarks used in this study is laborious for most biological systems. Even using the WINGMACHINE pipeline requires multiple stages of image processing, some manual landmark acquisition and manual correction of splines after automated fitting. By contrast, both feature extraction and classification using BioCAT were performed without *a priori* annotation or editing. Thus, despite the overall success in classification using the landmark and semi-landmark data, the efficiency of acquiring the data must also be considered for other studies.

Perhaps unsurprising given the different nature of the features extracted, the machine learning

algorithms that were most able to classify wings (for sex and genotype) using BioCAT's Hessian features differed from those that were most able to classify wings using landmark and semi-landmark data. Whereas SVMs consistently performed better than random forests for classification using morphometric data, the reverse was true using features extracted with BioCAT (Table 3-4). Classifications based on landmarks and semi-landmarks were also substantially improved by increasing the number of trees in the random forest from 10 (the BioCAT default) to 1,000 trees. BioCAT classification success was unaffected or slightly lowered by an increase in the number of trees.

The *Drosophila* wing database contains large numbers of wing images representing multiple genotypes. It also includes several built-in controls for technical variation that should make it amenable to the development of biometric classification tools. Using the landmark and semi-landmark data extracted with WINGMACHINE, wings can be classified by sex and genotype with high levels of accuracy. We were also able to classify wings by sex and genotype with relatively good accuracy using texture features extracted by the computer vision software BioCAT. We hope that this database will serve as a resource for research into the sources of variation contributing to wing shape and for the development and testing of measurement tools for image-based phenomics.

APPENDIX

IMAGE PROCESSING AND CLASSIFICATION USING BIOCAT

Supplementary Methods

1. Image processing for WINGMACHINE and CPR

1.1 Cropping wing images

Wing images taken at 2x on the Olympus BX51 microscope and at both 2x and 4x on the Leica M125 microscope were cropped (using the David's Batch Processor plugin in GIMP[1]) for analysis with WINGMACHINE. All images taken on the Olympus were initially 1360 x 1024 pixels (px). 2x images were cropped to 750 x 555px. All images on the Leica were initially 1392 x 1040px. Leica 4x images were cropped to 960 x 718px, except for a small subset in which the orientation of the wing required larger dimensions and these were cropped to 1200 x 898px. Leica 2x images were primarily cropped to 500 x 374px, except for a small subset that were cropped to 600 x 449px. The dimensions that were used for each image are indicated in the files 'Leica_2X_coords.tsv' and 'Leica_2X_coords.tsv' in the *GigaScience* GigaDB repository[2].

1.2 Landmarking wing images with tpsUtil and tpsDig

Wing images were resized for landmarking to the dimensions 632 x 480 pixels. A.tps file was created using tpsUtil (version 1.58[3]), and landmarked in tpsDig2[3]. The.tps file was converted to an.asc file using a custom python script.

1.3 Splining wing images

Images for splining were resized to the dimensions 316 x 240px. All WINGMACHINE splines were manually checked and edited.

1.4 Superimposition in CPR

WINGMACHINE output was superimposed in CPR. After superimposition and sliding of semi-landmarks, the most proximal semi-landmark on vein 4 was removed as described in[4]. Data for

the first five principal components were visually checked as a scatterplot and outliers caused by incorrect spline alignment were manually corrected.

1.5 Identifying scale in cropped images

Scale mm/px values were manually calibrated in ImageJ version 1.48[5].

2. Supplementary methods for BioCAT

This section includes a subset of informative results from testing various BioCAT settings with different subsets of wings. These subsets of the wing images, and the classification models, are available at the *GigasScience* GigaDB repository[2]. For these trials, training was done using BioCAT's 'Training only' option, followed by annotation using the 'test set'. In all cases, only left wings were used from the Olympus 40x dataset. Unless otherwise indicated, default settings were used. Data that was entered into the confusion matrices for Figure 6 are indicated (see code in[2]).

2.1 Classifying by sex within Samarkand genotype, using approximately equal numbers of each class for training and testing. Overall, the best results were with Hessian features combined with Random forest classifiers. F, female; M, male; RF, random forest; SAM, Samarkand; SVM, support vector machine.

Table 3-6 Results with BioCAT for sex

Features	Classifier	Training set	Testing set	Success rate
Structure 30	RF 10 trees	68 M SAM wings 71 F SAM wings (Sex training set 1)	15 M SAM wings 15 F SAM wings (Sex test set 1)	14/15 F correct 11/15 M correct
Structure 50	RF 10 trees	68 M SAM wings 71 F SAM wings (Sex training set 1)	15 M SAM wings 15 F SAM wings (Sex test set 1)	14/15 F correct 11/15 M correct
Laplacian 30	RF 10 trees	68 M SAM wings 71 F SAM wings (Sex training set 1)	15 M SAM wings 15 F SAM wings (Sex test set 1)	12/15 F correct 8/15 M correct
Laplacian 30	SVM linear	68 M SAM wings 71 F SAM wings (Sex training set 1)	15 M SAM wings 15 F SAM wings (Sex test set 1)	Very low
Stats 30	RF 10 trees	68 M SAM wings 71 F SAM wings (Sex training set 1)	15 M SAM wings 15 F SAM wings (Sex test set 1)	13/15 F correct 12/15 M correct
Derivatives 30	SVM linear	68 M SAM wings 71 F SAM wings (Sex training set 1)	15 M SAM wings 15 F SAM wings (Sex test set 1)	11/15 F correct 14/15 M correct
Hessian 30	SVM linear	68 M SAM wings 71 F SAM wings (Sex training set 1)	15 M SAM wings 15 F SAM wings (Sex test set 1 and Sex test set 2)	Test 1: 13/15 F correct 14/15 M correct Test 2: 8/15 F correct 15/15 M correct
Structure 50	RF 10 trees	68 M SAM wings 71 F SAM wings (Sex training set 1)	15 M SAM wings 15 F SAM wings (Sex test set 1 and Sex test set 2)	Test 1: 12/15 F correct 11/15 M correct Test 2: 10/15 F correct 15/15 M correct
Structure 50	SVM linear	68 M SAM wings 71 F SAM wings (Sex training set 1)	15 M SAM wings 15 F SAM wings (Sex test set 1 and Sex test set 2)	Test 1: 11/15 F correct 13/15 M correct Test 2: 8/15 F correct 15/15 M correct

Table 3-6 (cont'd)

Hessian 50	RF 10 trees	68 M SAM wings 71 F SAM wings (Sex training set 1)	15 M SAM wings 15 F SAM wings (Sex test set 1 and Sex test set 2)	Test 1: 13/15 F correct 13/15 M correct Test 2: 10/15 F correct 15/15 M correct
Hessian 50	SVM linear	68 M SAM wings 71 F SAM wings (Sex training set 1)	15 M SAM wings 15 F SAM wings (Sex test set 1 and Sex test set 2)	Test 1: 11/15 F correct 14/15 M correct Test 2: 9/15 F correct 15/15 M correct
Hessian 50	RF 1,000 trees	68 M SAM wings 71 F SAM wings (Sex training set 1)	15 M SAM wings 15 F SAM wings (Sex test set 1 and Sex test set 2)	Test 1: 13/15 F correct 13/15 M correct Test 2: 10/15 F correct 15/15 M correct

2.2 Classifying by genotype - Samarkand vs. mutant phenotypes (Egfr, mam, S, tkv) with varying numbers of trees. The best results were again achieved with Hessian combined with Random forest. SVM linear performed much better than SVM with alternate kernel shapes. RF, random forest; SAM, Samarkand; SVM, support vector machine.

Table 3-7 Results with BioCAT for genotype

Features	Classifier	Training set	Testing set	Success rate	Notes
Hessian 30	RF 10 trees	Genotype training 1 327 mutant females 77 SAM females	Genotype test 1 15 of each mutant 15 SAM	Identifies all wings as Mutant	
Hessian 30	SVM linear	Genotype training 1 327 mutant females 77 SAM females	Genotype test 1 15 of each mutant 15 SAM	9/15 SAM wrong 6/60 mutants wrong	
Hessian 30	SVM radial	Genotype training 1 327 mutant females 77 SAM females	Genotype test 1 15 of each mutant 15 SAM	Identifies all wings as Mutant	
Hessian 30	SVM sigmoid	Genotype training 1 327 mutant females 77 SAM females	Genotype test 1 15 of each mutant 15 SAM	Very low accuracy	
Structure 30	RF 10 trees	Genotype training 1 327 mutant females 77 SAM females	Genotype test 1 15 of each mutant 15 SAM	Very low accuracy	
Structure 30	SVM linear	Genotype training 1 327 mutant females 77 SAM females	Genotype test 1 15 of each mutant 15 SAM	2 mutants wrong 9/15 SAM wrong	
Structure 30	SVM radial	Genotype training 1 327 mutant females 77 SAM females	Genotype test 1 15 of each mutant 15 SAM	Very low accuracy	
Hessian 50	SVM linear	Genotype training 2 77 mutant females 77 SAM females	Genotype test 2 15 mutants total (3-4 of each) 15 SAM	Test 1: 6/15 SAM correct 13/15 mutants correct Test 2: 2/15 SAM correct 15/15 mutants correct	Equal representation in train/test sets greatly improved accuracy

Table 3-7 (cont'd)

Structure 50	SVM linear	Genotype training 2 77 mutant females 77 SAM females	Genotype test 2 15 mutants total 15 SAM	Test 1: 12/15 SAM correct 7/15 mutant correct Test 2: 13/15 SAM correct 10/15 mutant correct	
Structure 50	RF 10 trees	Genotype training 2 77 mutant females 77 SAM females	Genotype test 2 15 mutants total 15 SAM	Test 1: 10/15 SAM correct 11/15 mutants correct Test 2: 9/15 SAM correct 12/15 mutants correct	
Hessian 50	RF 10 trees	Genotype training 2 77 mutant females 77 SAM females	Genotype test 2 15 mutants total 15 SAM	Test1: 11/15 SAM correct 13/15 mutants correct Test 2: 11/15 SAM correct 12/15 mutants correct	

2.3 Classifying by genotype - all genotypes (females). Egfr, epidermal growth factor receptor; mam, mastermind; RF, random forest; SAM, Samarkand; SVM, support vector machine; tkv, thickveins.

Table 3-8 Results with BioCAT across genotype

Features	Classifier	Training set	Testing set	Success rate	Notes
Hessian 50	RF 10 trees	Genotype training 3 70 wings of each genotype, females	Genotype test 3 15 wings of each genotype, females x2 test sets	EgfrT1: 20%; 3 correct; 2 SAM; 10 tkv EgfrT2: 33.3%; 5 correct; 3 SAM; 7 tkv MamT1: 20%; 3 correct; 12 star MamT2: 33.3%; 5 correct; 10 star SAMT1: 86.7%; 1 mam; 1 Egfr; 13 correct SAMT2: 66.7%; 3 mam; 10 correct; 2 tkv StarT1: 100%; 15 correct Star T2: 73.3%; 11 correct; 4 mam TkvT1: 26.7%; 4 correct; 4 mam; 5 Egfr; 2 SAM TkvT2: 60%; 9 correct; 2 Egfr; 1 mam; 3 SAM	52% combined Table 4 Figure 6

Table 3-8 (cont'd)

Hessian 50	SVM linear	Genotype training 3 70 wings of each genotype, females	Genotype test 3 15 wings of each genotype, females	EgfrT1: 1 correct; 1 SAM; 13 tkv EgfrT2: 1 correct; 2 mam; 1 SAM; 11 tkv MamT1: 1 correct; 14 Star MamT2: 2 correct; 13 Star SAMT1: 11 correct; 1 Egfr; 3 mam SAMT2: 4 correct; 6 Egfr; 4 mam; 1 tkv StarT1: 14 correct; 1 mam StarT2: 13 correct; 1 mam; 1 tkv TkvT1: 6 correct; 3 Egfr; 6 SAM TkvT2: 13 correct; 2 mam	44% overall Table 4

Table 3-8 (cont'd)

Structure 50	SVM linear	Genotype training 3 70 wings of each genotype, females	Genotype test 3 15 wings of each genotype, females	EgfrT1: 3 correct; 2 SAM; 10 tkv EgfrT2: 1 correct; 1 mam; 5 SAM; 8 tkv MamT1: 15 correct MamT2: 1 correct; 14 Star SAMT1: 11 correct; 1 Egfr; 3 mam SAMT2: 9 correct; 5 Egfr; 1 mam StarT1: 14 correct; 1 mam StarT2: 14 correct; 1 mam TkvT1: 6 correct; 2 Egfr; 7 SAM TkvT2: 9 correct; 1 Egfr; 5 SAM	45.3 % overall

Table 3-8 (cont'd)

Structure 50	RF 10 trees	Genotype training 3 70 wings of each genotype, females	Genotype test 3 15 wings of each genotype, females	EgfrT1: 5 correct; 3 SAM; 7 tkv EgfrT2: 3 correct; 1 mam; 3 SAM; 8 tkv MamT1: 3 correct; 12 Star MamT2: 3 correct; 12 Star SAMT1: 13 correct; 1 Egfr; 1 mam SAMT2: 10 correct; 3 Egfr; 2 tkv StarT1: 14 correct; 1 mam StarT2: 13 correct; 2 mam TkvT1: 1 correct; 6 Egfr; 4 mam; 1 SAM TkvT2: 6 correct; 5 Egfr; 4 SAM	47.3 % overall

Table 3-8 (cont'd)

Hessian 50	RF 1,000 trees	Genotype training 3 70 wings of each genotype, females	Genotype test 3 15 wings of each genotype, females	EgfrT1: 0 correct; 2 SAM; 13 tkv EgfrT2: 2 correct; 4 SAM; 9 tkv MamT1: 1 correct; 14 Star MamT2: 2 correct; 13 Star SAMT1: 12 correct; 1 Egfr; 1 mam SAMT2: 10 correct; 1 Egfr; 2 mam; 2 tkv StarT1: 15 correct StarT2: 13 correct; 2 mam TkvT1: 6 correct; 2 Egfr; 5 mam; 2 SAM TkvT2: 9 correct; 1 Egfr; 1 mam; 4 SAM	47% overall Table 4

2.4 Classifying by sex across microscopes/magnifications. Egfr, epidermal growth factor receptor; F, female; M, male; mam, mastermind; RF, random forest; SAM, Samarkand; SVM, support vector machine; tkv, thickveins.

Table 3-9 Results with BioCAT for sex across technical conditions

Features	Classifier	Training set	Testing set	Success rate	Notes
Hessian 50	RF 10 trees	68 M SAM wings 71 F SAM wings Olympus, 40x mag uncropped	Test 1 15 M SAM wings 15 F SAM wings Test 2 15 M SAM wings 15 F SAM wings Olympus 20x uncropped	Identifies 100% as male 50% overall	Table 5
Hessian 50	RF 10 trees	68 M SAM wings 71 F SAM wings Olympus, 20x mag uncropped	Test 1 15 M SAM wings 15 F SAM wings Test 2 15 M SAM wings 15 F SAM wings Olympus 20X uncropped	Test 1: 13/15 F correct 10/15 M correct Test 2: 12/15 F correct 15/15 M correct 83.3% overall	
Hessian 50	RF 10 trees	68 M SAM wings 71 F SAM wings Olympus, 40x mag cropped for splining	Test 1 15M SAM wings 15 F SAM wings Test 2 15M SAM wings 15 F SAM wings Olympus 40x cropped for splining	Test 1 12/15 F correct 13/15 M correct Test 2 10/15 F correct 15/15 M correct 83.3% overall	
Hessian 50	RF 10 trees	68 M SAM wings 71 F SAM wings Olympus, 40x mag cropped for splining	Test 1 15M SAM wings 15 F SAM wings Test 2 15M SAM wings 15 F SAM wings Olympus 20x cropped for splining	Test 1 0/15 F correct 14/15 M correct Test 2 1/15 F correct 15/15 M correct 50% overall	Table 5

Table 3-9 (cont'd)

Hessian 50	RF 10 trees	68 M SAM wings 71 F SAM wings Olympus, 40x mag cropped for splining	Test 1 15 M SAM wings 15 F SAM wings Test 2 15M SAM wings 12 F SAM wings Leica 40x cropped for splining	Test 1 7/15 F correct 14/15 M correct Test 2 7/12 F correct 14/15 M correct 73.7% overall	Table 5
Hessian 50	RF 10 trees	68 M SAM wings 71 F SAM wings Leica, 40x mag cropped for splining	Test 1 15 M SAM wings 15 F SAM wings Test 2 15 M SAM wings 12 F SAM wings Leica 40x cropped for splining	Test 1 15/15 F correct 15/15 M correct Test 2 12/12 F correct 11/15 M correct 93.0% overall	Table 5
Hessian 50	RF 10 trees	68 M SAM wings 71 F SAM wings Leica, 40x mag cropped for splining	Test 1 15M SAM wings 15 F SAM wings Test 2 15M SAM wings 15 F SAM wings Olympus 40x cropped for splining	Test 1 14/15 F correct 0/15 M correct Test 2 14/15 F correct 0/15 M correct	
Hessian 50	RF 10 trees	Mixed microscopes 15 males Oly 4x 15 males Lei 4x 15 females Oly 4x 15 females Lei 4x	Test 1 15M SAM wings 15 F SAM wings Test 2 15M SAM wings 15 F SAM wings Olympus 40x	Test 1 13/15 F correct 8/15 M correct Test 2 15/15 F correct 8/15 M correct	Table 5 73.3%
Hessian 50	RF 10 trees	Mixed microscopes 15 males Oly 4x 15 males Lei 4x 15 females Oly 4x 15 females Lei 4x	Test 1 15M SAM wings 15 F SAM wings Test 2 15M SAM wings 15 F SAM wings Leica 40x	Test 1 13/15 F correct 14/15 M correct Test 2 9/12 F correct 13/15 M correct	Table 5 86.0%

2.5 Classifying by genotype using Leica 4x images. Egfr, epidermal growth factor; mam,

mastermind; RF, Random forest; SAM, Samarkand; tkv thickveins.

Table 3-10 Results with BioCAT for genotype across technical conditions

Features	Classifier	Training set	Testing set	Success rate	Notes
Hessian 50	RF 10 trees	Genotype train Leica 64 wings from each genotype Leica 4x Cropped prior to landmark, splining 960 x 718 pixels	Genotype test Leica All from Leica 4x Test set 1 15 wings from each genotype Test set 2 15 wings from each genotype	EgfrT1: 9/15 correct 6/15 SAM EgfrT2: 12/15 correct 3/15 SAM MamT1: 1/15 correct 10/15 Star 4/15 tkv MamT2: 2/15 correct 2/15 Egfr 3/15 SAM 4/15 Star 4/15 tkv SAMwT1: 14/15 correct 1/15 mam SAMwT2: 8/15 correct 7/15 mam StarT1: 14/15 correct 1/15 mam StarT2: 15/15 correct TkvT1: 2/15 correct 4/15 mam 2/15 SAMw 7/15 Star TkvT2: 3/15 correct 3/15 mam 9/15 Star	53.3 % accurate overall. This isn't substantially higher than the accuracy using the Olympus images (52% overall), unlike sex predictions.

REFERENCES

REFERENCES

1. Houle D, Govindaraju DR, Omholt S. Phenomics: the next challenge. Nat Rev Genet. 2010;11:855–66.

2. Gerlai R. Phenomics: fiction or the future? Trends Neurosci. 2002;25:506–9.

3. Adams DC, Rohlf FJ, Slice DE. Geometric morphometrics: Ten years of progress after the ‘revolution’. Ital J Zool. 2004;71:5–16.

4. Slice DE. Geometric Morphometrics. Ann Rev Anthropol. 2007;36:261–81.

5. Klingenberg CP. MorphoJ: an integrated software package for geometric morphometrics. Mol Ecol Resour. 2011;11:353–7.

6. Mitteroecker P, Gunz P. Advances in geometric morphometrics. Evol Biol. 2009;36:235–41.

7. Zelditch ML, Swiderski DL, Sheets HD. Geometric Morphometrics for Biologists: A Primer. 2nd ed. San Diego: Elsevier; 2012.

8. Danuser G. Computer Vision in Cell Biology Cell. 2011;147:973–8.

9. PubMedShamir L, Delaney JD, Orlov N, Eckley DM, Goldberg IG. Pattern Recognition Software and Techniques for Biological Image Analysis. PLoS Comput Biol. 2010;6:e1000974.

10. Allanson JE, Bohring A, Dorr H, Dufke A, Gillessen-Kaesbach G, Horn D, et al. The face of Noonan syndrome: Does phenotype predict genotype. Am J Med Genet A. 2010;152A:1960–6.

11. Boehringer S, Guenther M, Wurtz RP, Horsthemke B, Wieczorek D. Automated syndrome detection in a set of clinical facial photographs. Am J Med Genet A. 2011;155A:2161–9.

12. Ferry Q, Steinberg J, FitzPatrick DR, Ponting CP, Zisserman A, Nellaker C. Diagnostically relevant facial gestalt information from ordinary photos. Elife. 2014;3:e02020.

13. Zou L, Adegun OK, Willis A, Fortune F. Facial biometrics of peri-oral changes in Crohn’s disease. Lasers Med Sci. 2014;29:869–74.

14. Houle D, Mezey J, Galpern P, Carter A. Automated measurement of Drosophila wings. BMC Evol Biol. 2003;3:25.

15. Kumar N, Belhumeur PN, Biswas A, Jacobs DW, Kress WJ, Lopez IC, et al. Leafsnap: A

Computer Vision System for Automatic Plant Species Identification. In: Fitzgibbon A, Lazebnik S, Perona P, Sato Y, Schmid C, editors. Computer Vision - ECCV 2012. Berlin: Springer; 2012. p. 502–16.

16. Perrard A, Baylac M, Carpenter JM. Evolution of wing shape in hornets: why is the wing venation efficient for species identification? J Evol Biol. 2014;27:2665–75.

17. Unar JA, Seng WC, Abbasi A. A review of biometric technology along with trends and prospects. Pattern Recogn. 2014;47:2673–88.

18. Houle D. Numbering the hairs on our heads: the shared challenge and promise of phenomics. Proc Natl Acad Sci U S A. 2010;107:1793–9.

19. Debat V, Debelle A, Dworkin I. Plasticity, canalization and developmental stability of the Drosophila wing: joint effects of mutations and developmental temperature. Evolution. 2009;63:2864–76.

20. Klingenberg CP. Morphometric integration and modularity in configurations of landmarks: tools for evaluating a priori hypotheses. Evol Dev. 2009;11:405–21.

21. Paloniswamy S, Thacker NA, Klingenberg CP. Automatic identification of landmarks in digital images. IET Comput Vis. 2010;4:247–60.

22. Gidaszewski NA, Baylac M, Klingenberg CP. Evolution of sexual dimorphism of wing shape in the Drosophila melanogaster subgroup. BMC Evol Biol. 2009;9:110.

23. Pitchers W, Pool JE, Dworkin I. Altitudinal clinal variation in wing size and shape in African Drosophila melanogaster: one cline or many? Evolution. 2013;67:38–42.

24. Miles WO, Korenjak M, Griffiths LM, Dyer MA, Provero P, Dyson NJ. Post-transcriptional gene expression control by NANOS is up-regulated and functionally important in pRb-deficient cells. EMBO J. 2014;33:2201–15.

25. Gafner L, Dalessi S, Escher E, Pyrowolakis G, Bergmann S, Basler K. Manipulating the sensitivity of signal-induced repression: quantification and consequences of altered brinker gradients. PLoS One. 2013;8:e71224.

26. Birdsall K, Zimmerman E, Teeter K, Gibson G. Genetic variation for the positioning of wing veins in Drosophila melanogaster. Evol Dev. 2000;2:16–24.

27. Dworkin I, Gibson G. Epidermal Growth Factor Receptor and Transforming Growth Factor-β signaling contributes to variation for wing shape in Drosophila melanogaster. Genetics. 2006;173:1417–31.

28. Ahmad F, Roy K, O'Connor B, Shelton J, Dozier G, Dworkin I. Fly Wing Biometrics Using Modified Local Binary Pattern SVMs and Random Forest. International Journal of Machine Learning and Computing. 2014;4:279–85.

29. Payne M, Turner J, Shelton J, Adams J, Carter J, Williams H, et al. Fly wing biometrics. IEEE Symposium Series on Computational Intelligence and Biometrics. 2013;42-6. doi:10.1109/CIBIM.2013.6607912.

30. Shelton J, Bryant K, Abrams S, Small L, Adams J, Leflore D, et al. Genetic & Evolutionary Biometric Security: Disposable Feature Extractors for Mitigating Biometrics Replay Attacks. Procedia Comput Sci. 2012;8:351–60.

31. Zhou J, Lamichhane S, Sterne G, Ye B, Peng H. BIOCAT: a pattern recognition platform for customizable biological image classification and annotation. BMC Bioinformatics. 2013;14:291.

32. Blair SS. Wing vein patterning in Drosophila and the analysis of intercellular signaling. Annu Rev Cell Dev Biol. 2007;23:293–319.

33. Bray SJ. Notch signaling: a simple pathway becomes complex. Nat Rev Mol Cell Biol. 2006;7:678–89.

34. Guichard A, Biehs B, Sturtevant MA, Wickline L, Chacko J, Howard K, et al. Rhomboid and Star interact synergistically to promote EGFR/MAPK signaling during Drosophila wing vein development. Development. 1999;126:2663–76.

35. PubMedR Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: the R Foundation for Statistical Computing. 2011. http://www.R-project.org. Accessed 21 April 2015.

36. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning: with applications in R. New York: Springer; 2013.

37. Meijering E, Jacob M, Sarria JCF, Steiner P, Hirling HL, Unser M. Design and validation of a tool for neurite tracking and analysis in fluorescence microscopy images. Cytom Part A. 2004;58:167–76.

38. Chandler C, Chari S, Tack D, Dworkin I. Causes and consequences of genetic background effects illuminated by integrative genomic analysis. Genetics. 2014;196:1321–36.

39. Gnu Image Manipulation Program, version 2.8. 2015. http://www.gimp.org. Accessed 21 April 2015.

40. David's Batch Processor version 1.1.8. 2015. http://members.ozemail.com.au/~hodsond/dbp.html. Accessed 21 April 2015.

41. Rohlf, FJ. tpsDig version 2.17. 2015. http://life.bio.sunysb.edu/morph/soft-dataacq.html. Accessed 21 April 2015.

42. CPR. 2015. Houle Lab: Software. http://bio.fsu.edu/~dhoule/wings.html. Accessed 29 April 2015.

43. DworkinLab Github account: Wing_Biometrics_2015. https://github.com/DworkinLab/Wing_Biometrics_2015. Accessed 29 April 2015.

44. David's Batch Processor version 1.1.8. 2015. http://members.ozemail.com.au/~hodsond/dbp.html. Accessed 24 April 2015.

45. tpsUtil. 2015. http://life.bio.sunysb.edu/morph/soft-utility.html. Accessed 24 April 2015

46. Pitchers W, Pool JE, Dworkin I. Altitudinal clinical variation in wing size and shape in African Drosophila melanogaster: one cline or many? Evolution. 2013;67:38-42

47. Image J. 2015. http://imagej.nih.gov. Accessed 24 April 2015.

CHAPTER IV: CONCLUSIONS AND FUTURE PERSPECTIVE

Introduction

A major goal of modem biology is to use newly available diverse sets of quantitative ‘omic data to make useful predictions relevant to health, agriculture, the environment, and basic biological research[1]. This objective applies in particular to 'precision medicine'. The same genetic or clinical properties in different people can mean very different things depending on individual factors and environmental context-- being able to take the relevant features and interpret their meaning without direct experimental testing would be invaluable in a medical or scientific setting[2]. However, predictions tend to be most accurate when they have been produced from an individually tailored dataset, which is not feasible for all the circumstances where predictions might be applied[3]. Experimentally, testing all treatments on all possible combination of inputs is not feasible, so despite the above limitations of prediction, computational methods using these new and growing data sets must be developed. Technological breakthroughs have provided modern biologists with an ever-increasing treasure trove of publicly available biologically relevant data necessary for this task; the appropriate matched computational resources are still being developed.

Experimental approaches are changing more quickly than the analytical methods for interpreting them. DNA sequencing technology, for example, has been increasing in speed and decreasing in cost at an exponential rate[4]. In just the last ten years, new techniques like STARR-Seq, RNA-seq, ChIP-seq and chromatin conformation capture (3C) analysis have replaced older methods like RT-PCR and microarrays[5–9]. In the last five years, CRISPR has gone from a niche subject to the “Swiss army knife” of molecular biology[10]. Despite this data explosion, exhaustively measuring how cells, organisms or ecosystems with complex genetic architectures would generate differential responses to diverse environmental circumstances, is still impossible. Thus,

computational methods are the only way forward. However, the machine learning methods that are used to make predictions and classifications from this data are largely unchanged in the last twenty years[11,12].

From genotype to gene expression

In an effort to address this deficiency in one important area of genomic research, in Chapter 2, I describe my efforts to use random forests to distinguish active cis-regulatory DNA from background, using the Fly Enhancer Resource developed by Alexander Stark's lab as a training set[13]. The existence of such a large database of in-vivo reporters allowed an in-depth analysis of how effective these predictions were, and how well they generalize. I found random forests made good predictions when tested on reporters set aside for testing from the Fly Enhancer Resource, but produced substantially higher false-discovery rate when applied to putative enhancers identified around well-characterized genes. I also found that attempting to classify by more specific categories (expression pattern, or temporal activity) was much less accurate. However, these are both dynamic categories that depend on the developmental window being considered, thus trying to classify elements into discrete groups may be a fundamentally flawed approach. Probabilistic models, like what is used in some software like Manolis Kellis ChromHMM, may be more computationally effective and biologically accurate, as they characterize distributions rather than classify based on discrete groups[14,15].

It seems likely that the currently existing datasets should be able to yield features that show differences between poised and currently active enhancers (as experimentally shown for H3K27 acetylation in Koenecke et al. 2016[16]) , even if we are not currently able to concretely classify them. Based on the degree to which classification accuracy of my random forest classification levels off asymptotically as more datasets are incorporated, it seems likely that including more

information from chromatin immunoprecipitation of transcription factors would have marginal returns. Future efforts along these lines may have more success using data along different lines, such as RNA-seq information for the relevant locus, chromatin conformation data, histone marks that are not explicitly associated with enhancer function (which may help parse out non-enhancers that share some similar features, like promoters) and information about proximity to prospective transcription start sites. More information about the binding of cofactors may also be relevant for regulatory elements that fit into the category of 'transcription factor collectives' and are less reliant on the underlying DNA sequence, or longer enhancer regions that do not necessarily fall into a neat 2kb window[17,18].

A more nuanced use of sequence conservation data than applied in my study would be helpful; information about variation between species and within populations is a rich resource that is abundantly available for *Drosophila*, but measuring average sequence change over large windows is not likely to yield a result[19,20]. Using cross-species ChIP data to look for regions that appear to be under purifying or directional selection may be more fruitful, as would using programs like EMMA from Saurabh Sinha's laboratory that are explicitly intended for making comparisons of enhancers between species by modeling turnover of transcription factor binding sites[21,22].

From wing shape to genotype

In Chapter 3, I describe a database we developed for the testing of alternative methods for extracting and analyzing phenotype features from images of *Drosophila* wings. I also discuss some preliminary analysis of how existing methods for classifying *Drosophila* into sex and genotype using standard morphometric features compares with classification based on features extracted by Biocat, an open source computer vision tool[23]. We found that morphometric

methods (which have been in development for decades) were substantially more successful, but features extracted from images using Biocat were quite successful at classifying wings by the sex of their origin fly, and by genotype[24].

In the future, it would be interesting to see how well computer vision methods could be applied to more complex classifications, like looking for the interactions between genes, recognizing a given genotype that has been influenced by different environmental conditions, or identifying a common mutation in different species. It would also be useful to see whether combination approaches that take advantage of both morphometric features and features extracted using computer vision methods can outperform either or both methods alone.

New methods of measuring phenotype and correlating it with underlying genetic state may lead to a more nuanced view of how subtle genetic perturbations influence phenotype. For example, the phenotypic effects seen in the removal of 'shadow' enhancers may be evident under a range of conditions, given sufficiently sophisticated measurement[25,26].

Conclusions

Although there are still limitations on using computational predictions in place of experimental work, it is a promising future direction. Experimental technology is constantly advancing, but the relevant features of the phenome that biologists are interested in exploring appears to be too large to ever exhaustively explore it through experiments-- it will always be necessary to draw conclusions from unique combinations of inputs. We used a large number of publicly available features to determine how accurately enhancers could be distinguished from background. Our results suggest that based on the currently existing data, there are limits to the success of these predictions-- surpassing these limits will either require a new method for analysis, a better understanding of the enhancer state, or a new way of interpreting the existing features. In image

analysis, computer vision with machine learning has been used as an alternative to traditional human noticeable features. Automatic extraction of features allows identifying important elements that may not appear to have relevance at first glance. Based on our initial analysis, however, automatic analysis alone is not comparable to the well-developed traditional methods. We have created a database of wing images in the hopes that it will be used for the creation of novel tools, which may find new features to use as the basis of predictions. When our ability to analyze data catches up with our current rate of accumulation, we may end up with a much stronger understanding of how genotype actually relates to phenotype.

REFERENCES

REFERENCES

1. Stephens ZD, Lee SY, Faghri F, et al. Big data: Astronomical or genomical? *PLoS Biol*. 2015;13(7):7.

2. Mirnezami R, Nicholson J, Darzi A. Preparing for Precision Medicine. *N Engl J Med*. 2012;366(6):489-491.

3. Xue Y, Lameijer E-W, Ye K, et al. Precision Medicine: What Challenges Are We Facing? *Genomics Proteomics Bioinformatics*. 2016;14(5):253-261.

4. Mardis ER. A decade's perspective on DNA sequencing technology. *Nature*. 2011;470(7333):198-203.

5. Arnold CD, Gerlach D, Stelzer C, Boryn LM, Rath M, Stark A. Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq. *Science (80-).* 2013;339(6123):1074-1077. doi:10.1126/science.1232542.

6. Bumgarner R. DNA microarrays: Types, Applications and their future. *Curr Protoc Mol Biol*. 2013;0 22:Unit-22.1.

7. Denker A, de Laat W. The second decade of 3C technologies: detailed insights into nuclear organization. *Genes Dev* . 2016;30(12):1357-1382.

8. Ho JWK, Bishop E, Karchenko P V, Nègre N, White KP, Park PJ. ChIP-chip versus ChIP-seq: Lessons for experimental design and data analysis. *BMC Genomics*. 2011;12(1):134.

9. Tachibana C. Transcriptomics today: Microarrays, RNA-seq, and more. *Science (80-).* July 2015.

10. Sternberg SH, Doudna JA. Expanding the Biologist's Toolkit with CRISPR-Cas9. *Mol Cell*. 2015;58(4):568-574.

11. Kourou K, Exarchos TP, Exarchos KP, Karamouzis M V., Fotiadis DI. Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J*. 2015;13:8-17.

12. Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. *Cancer Inform*. 2006;2:59-77.

13. Kvon EZ, Kazmar T, Stampfel G, et al. Genome-scale functional characterization of Drosophila developmental enhancers in vivo. *Nature*. 2014;512(7512):91-95.

14. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods*. 2012;9(3):215-216.

15. Ernst J, Vainas O, Harbison CT, Simon I, Bar-Joseph Z. Reconstructing dynamic regulatory maps. *Mol Syst Biol*. 2007;3(1).

16. Koenecke N, Johnston J, Gaertner B, Natarajan M, Zeitlinger J. Genome-wide identification of Drosophila dorso-ventral enhancers by differential histone acetylation analysis. *Genome Biol*. 2016;17(1):196.

17. Junion G, Spivakov M, Girardot C, et al. A Transcription Factor Collective Defines Cardiac Cell Fate and Reflects Lineage History. *Cell*. 2017;148(3):473-486.

18. Pott S, Lieb JD. What are super-enhancers? *Nat Genet*. 2015;47(1):8-12.

19. Lack JB, Lange JD, Tang AD, Corbett-Detig RB, Pool JE. A Thousand Fly Genomes: An Expanded Drosophila Genome Nexus. *Mol Biol Evol*. 2016;33(12):3308-3313.

20. Halligan DL, Keightley PD. Ubiquitous selective constraints in the Drosophila genome revealed by a genome-wide interspecies comparison. *Genome Res*. 2006;16(7):875-884.

21. He Q, Bardet AF, Patton B, Purvis J, Johnston J, Paulson A. High conservation of transcription factor binding and evidence for combinatorial regulation across six Drosophila species. *Nat Genet*. 2011;43..

22. He X, Ling X, Sinha S. Alignment and Prediction of cis-Regulatory Modules Based on a Probabilistic Model of Evolution. *PLOS Comput Biol*. 2009;5(3):e1000299.

23. Zhou J, Lamichhane S, Sterne G, Ye B, H. P. Biocat: a pattern recognition platform for customizable biological image classification and annotation. *BMC Bioinformatics*. 2013;14:291.

24. Zelditch M. *Geometric Morphometrics for Biologists : A Primer*. Elsevier Academic Press; 2004.

25. Cannavo E, Khoueiry P, Garfield DA, Geeleher P, Zichner T, et al. Shadow enhancers are pervasive features of developmental regulatory networks. *Curr Biol*. 2016;26:1-14.

26. Gafner L, Dalessi S, Escher E, Pyrowolakis G, Bergmann S, Basler K. Manipulating the Sensitivity of Signal-Induced Repression: Quantification and Consequences of Altered Brinker Gradients. *PLoS One*. 2013;8(8):8.