DE NOVO PREDICTION OF CIS-REGULATORY MODULES IN EUKARYOTIC ORGANISMS

by

Meng Niu

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Bioinformatics and Computational Biology

Charlotte

2014

Approved by:

_____
Dr. Zhengchang Su

_____
Dr. Anthony Fodor

_____
Dr. Daniel Janies

_____
Dr. Juntao Guo

_____
Dr. Bao-Hua Song

ABSTRACT

MENG NIU. *De novo* prediction of *cis*-regulatory modules in eukaryotic organisms. (Under the direction of DR. ZHENGCHANG SU)

Gene regulation networks (GRNs) are the bases for virtually all biological processes. To gain a global understanding of GRNs encoded in a genome, we first need to identify in all the *cis*-regulatory elements (CREs) recognized by transcription factors (TFs). In higher eukaryotes, CREs rarely work alone, instead, they regulate genes by forming combinatorial patterns called *cis*-regulatory modules (CRMs). Thus finding CREs as well as CRMs is the key to understanding GRNs in eukaryotes. However, identification of CREs and CRMs is a highly challenging task due to their short length and degeneracy while residing in long intergenic or intronic sequences. The recent wide adaptation of chromatin precipitation followed by DNA sequencing (ChIP-seq) techniques has churned out numerous datasets for locating CREs for TFs, providing an unprecedented opportunity to decipher CREs and CRMs in a genome. In this dissertation, we have developed a graph theory based algorithm DePCRM for genome-wide *de novo* predictions of CRMs and CREs by integrating a large number of ChIP datasets. Using this algorithm, we have predicted 1,108,018 and 5,186,520 CREs, and 115,932 and 807,365 CRMs in the *Drosophila melanogaster* and human genomes, respectively, using all the ChIP-seq datasets available to us in the two organism. We found that our predicted CRMs could recover more than 80% known CRMs, and that both the putative CREs and CRMs were more conserved than randomly selected sequences in both the genomes. Furthermore, trait-linked SNPs and DNaseI hypersensitive regions are highly enriched in our predicted CRMs in the human genome.

Thus, we have provided so far the most comprehensive maps of CREs and CRMs in the two genomes. Using the much larger number of human ChIP datasets, we also analyzed the saturation trends of predicted CRE motifs and their combinatory patterns using an increasing number of randomly selected datasets, datasets in different cell types and datasets for different TFs. We found that the saturation trends started to be notable with only a few datasets in each scenario. The results suggest ways to generate ChIP datasets more cost-effectively in the future. Finally, we analyzed the conservation and variation of the *cis*-regulatory systems between the two species. We found that although a large portion of CRMs are conserved in their motif composition in the two species, their target genes have been significantly changed. Thus, the majority of the GRNs have been rewired during the evolution from *D. melanogaster* to humans.

DEDICATION

To my family and teachers who introduced me into science.

ACKNOWLEDGEMENTS

First, I would like to thank my advisor Dr. Zhengchang Su for his guidance, support, encouragement and patience over the years. He's an excellent advisor of vision and dedication to research, and this dissertation would not have been possible without him. The journey with him is a priceless lesson to my life. My great appreciation to him is not only for the free academic environment and financial support he provided, but also the ways to think and implement as a researcher that he taught me.

I most sincerely thank my committee members – Dr. Anthony Fodor, Dr. Jun-tao Guo, Dr. Daniel Janies, and Dr. Bao-Hua (Alysa) Song for spending their precious time on meeting with me and providing invaluable suggestions and ideas for me to continuously make progress on my research.

I would also like to thanks the members in our group for their valuable suggestions and support, both past and current, and most especially, Ehsan Tabari. Ehsan has been a labmate and a dear friend of mine who was always willing to give suggestions and help unreservedly.

I would like to dedicate my deepest gratitude to my family for their love and support all these years. The support from my parents Mr. Tiejun Niu and Ms. Hongyan Meng made it possible for me to pursue the Ph.D. And the love and encouragement from my dear wife Liying Wang has always been my strength.

Finally, I thank the Department of Bioinformatics and Genomics for providing me with the opportunity to pursue Ph.D. in Bioinformatics and the Graduate Assistant Support Plan (GASP) for providing me with the financial support in the past years. My

appreciation also goes to all of the staff members in my department and the international students and scholars office for their friendly help.

INTRODUCTION

With the development of the powerful computational and experimental methods, we have gained a rather comprehensive understanding of genome-wide coding sequences of all the sequenced eukaryotic organisms [1, 2]. By contrast, functional non-coding sequences are only starting to be systematically studied due to the difficulties in their characterization. In particular, accumulative data have indicated that non-coding functional elements play a more important role in the adaptation of species to their environments during the course of evolution than originally thought [3]. For instance, comparative and functional genomics data now strongly support the long standing argument that the differences between humans and chimpanzees stem from the divergence in their regulatory elements rather than coding sequences [4-9]. Furthermore, variations in non-coding regulatory regions are more likely to account for phenotypic diversities among individuals of a given species than variations in coding sequences. For instance, recent genome wide association studies (GWAS) have found that vast majority of disease associated single nucleotide variations (SNVs) do not reside in coding sequences, instead they lie in the non-coding sequences overlapping with chromatin markers for non-coding regulatory regions [10-12].

In eukaryotic organisms, the transcriptional control elements including promoters, enhancers, silencers and insulators mostly reside in the non-coding regions of the genomes including intergenic sequences and introns; they along with epigenetic remodeling machineries determine which protein or RNA-specifying sequences should be transcribed for the cells under various physiological conditions [13]. All of these functional elements are bound by more than one transcriptional factor (TF) to their

individually recognized cis-regulatory elements (CREs) to perform their regulatory functions. In other word, TFs are barely working alone, they cooperate with each other to regulate the transcription process by binding to a group of CREs closely located within promoters, enhancers, silencers and insulators. These groups of CREs are called *cis*-regulatory modules (CRMs). The combinatorial use of a subset of the TFs results in a large number of CRMs of different constitutions that account for the unique regulation of each gene in the genome in different cell types, tissues, developmental stages and physiological conditions. Thus it is critical to identify all CREs and their combinations in the forms of CRMs, to elucidate dynamic GRNs in different cell types in the entire life of the organism. However, our general knowledge of the locations of these CRE and CRMs are still limited due to the difficulty to characterize them.

The difficulty in identifying CREs and CRMs by using the traditional computational or experimental methods is mainly due to the short and degenerate nature of their sequences as well as their usual locations within very long intergenic and intron sequences [14]. Fortunately, the process is largely facilitated by the development of next-generation sequencing (NGS)-base high throughput techniques including ChIP-seq [15-17] [18], DNase-seq [19-21], FAIRE-seq [20], Hi-C and RNA-seq [22, 23]. However, it remains a highly challenging computational problem to derive CREs and CRMs genome-wide in large eukaryotic genomes using large volumes of datasets produced by these techniques. For example, from a single ChIP-seq experiment for a TF in the human genome, typically thousands to tens of thousands of peaks with lengths of a few hundreds will be returned by the peak-calling tools such as PeakSeq or MACs [24, 25]. Hence, the actual location of CREs with length of 6~20 bs still need to be identified by a motif-

finding tool, which can be a nontrivial tasks. Researchers also have made efforts to develop algorithms for CRM prediction, such as SpaMo [26], CPModule [27] and [28], which are designed to identify CREs of cooperator TFs in a ChIP-seq dataset. However, these algorithms do not integrate multiple ChIP-seq datasets, and cannot predict novel motifs in CRMs, as they all depend on a library of known CREs such as TRANSFAC [29] or JASPAR [30] to scan for possible cooperative CREs in binding peaks.

In this dissertation, we attempted to tackle this problem by developing an algorithm for predicting CREs and CRMs by integrating a large number of ChIP datasets in a genome. Using this algorithm, we have predicted comprehensive maps of CREs and CRMs in the *Drosophila melanogaster* and human genomes with high accuracy judged by a few criteria. Furthermore, we also addressed the question of how far we are from obtaining a complete CRE and CRM map in the human genome and how to approach this goal cost-effectively.

Finally, *Drosophila melanogaster* as one of the most studied model animal, shows extensive conservation with humans at the levels of gene, pathway, organ and behavior [31]. This conservation has helped researchers succesfully develop models for a variety of human diseases [32-43]. However, these studies mostly focused on one specific tissue or disease, thus we still lack a global perspective of conservation and variation, especially, of the GRNs represented by TFs and their cognate CREs and CRMs and target genes. Our predictions of the CREs and CRMs in the two genomes allowed us to conduct a comprehensive comparison of the cis-regulatory systems in two species, and to provide evidences of the CRM evolution from *Drosophila melanogaster* genome to human genome.

TABLE OF CONTENTS

CHAPTER 1: DE NOVO PREDICTION OF CIS-REGULATORY ELEMENTS AND MODULES THROUGH INTEGRATIVE ANALYSIS OF A LARGE NUMBER OF CHIP DATASETS

1.1 Abstract

In eukaryotes, transcriptional regulation is mediated by interactions of multiple transcription factors (TFs) with their respective specific cis-regulatory elements (CREs) in the so-called cis-regulatory modules (CRMs) in DNA. Although the knowledge of CREs and CRMs in a genome is crucial to elucidate gene regulatory networks and understand many important biological phenomena, little is known about the CREs and CRMs in most eukaryotic genomes due to the difficulty to characterize them by either computational predictions or traditional experimental methods. However, the exponentially increasing number of TF binding location data produced by the recent wide adaptation of chromatin immunoprecipitation coupled with microarray hybridization (ChIP-chip) or high-throughput sequencing (ChIP-seq) technologies has provided an unprecedented opportunity to identify CRMs and CREs in genomes. Nonetheless, how to effectively mine the large volumes of ChIP data to identify CREs and CRMs is a challenging task.

We have developed a novel graph-theoretic based algorithm DePCRM for genome-wide de novo predictions of CRMs and CREs using a large number of ChIP datasets. DePCRM predicts CRMs by identifying overrepresented combinatorial CRE motif patterns in multiple ChIP datasets in an effective way. When applied to 168 ChIP datasets of 56 TFs from *D. melanogaster*, DePCRM identified 184 and 746

overrepresented CRE motifs and their combinatorial patterns, respectively, and predicted a total of 115,932 CRMs in the genome. The predictions recover 77.9% of known CRMs in the datasets, 89.3% of known CRMs containing at least one predicted CRE. We found that the putative CRMs and CREs as a whole in a CRM are more conserved than randomly selected sequences.

Our results suggest that the CRMs predicted by DePCRM are highly likely to be functional. Our algorithm is the first of its kind for de novo genome-wide prediction of CREs and CRMs using multiple TF ChIP datasets. The algorithm and predictions will hopefully facilitate the elucidation of gene regulatory networks in eukaryotes. All the predicted CREs, motifs, CRMs, and their target genes are available at http://bioinfo.uncc.edu/mniu/pcrms/www/

1.2 Background

Since the completion of sequencing the first metazoan genomes in 1998 [44], more than 311 important metazoan and plant genomes have been sequenced thus far [2], and enormous efforts have been made to understand how biological functions and diseases of these organisms including the humans can be explained by the genetic information stored in the genome sequences. Although significant progress has been made in the past 16 years, we are still far from the goal of understanding the biology of metazoans and plants solely from their genome sequences [45]. In fact, it turns out that interpreting a genome is more difficult and challenging than originally thought when a few eukaryotic genomes including the human genome were first released [45, 46]. With this recognition, the community has taken a more realistic approach by first identifying all the functional sequence elements in the genomes [47-49]. These functional elements

include transcribed sequences as well as transcriptional control elements, epigenetic features, and regulatory elements acting at the RNA level post-transcriptionally. In principle, while the transcribed sequences specify the potential part list in the cells in an organism, including proteins, various types of RNAs and metabolites, the transcriptional control elements including promoters, enhancers, silencers and insulators together with epigenetic remodeling machineries, determine which protein- or RNA-specifying sequences should be transcribed in each cell during development and under various physiological conditions, thereby specifying the cell's type during development and specific physiological functions, as it is the dynamic interactions of these components in a cell that determine the cell's type and specific physiological functions [13]. Once these functional elements are at least partially known, then we can move toward to the next step to identify dynamic interactions among the functional sequence elements and their products of proteins, RNAs and metabolites in different cell types in the entire life of the organism.

In the past we have gained a good understanding of transcribed sequences, particularly protein-coding sequences in numerous sequenced eukaryotic genomes thanks to the development of powerful computational and experimental methods for their characterization [50]. However, we have had only very limited understanding of transcriptional control elements, particularly promoters, enhancers and silencers in virtually all sequenced large eukaryotic genomes, even though these elements are as important as the transcribed sequences for the functions of an organism [51-53]. More specifically, promoters, enhancers and silencers are clusters of closely located cis-regulatory elements (CREs) that are recognized by specific transcription factors (TFs)

[54]. Thus, a CRE is also called a TF binding site (In this paper, we will refer to a set of similar CREs recognized by the same TF as a motif). These clusters of CREs are also called cis-regulatory modules (CRMs) [54]. The difficulty to identify CREs and CRMs either computationally or experimentally is due mainly to their short and degenerate nature while they mainly reside in very long intergenic or intronic background sequences [14]. To further confound the problem, they can be very far away from the target genes or even can be located on a different chromosome [55], making their characterization extremely difficult by computational methods such as comparative genomics approaches, although there are successful examples, in particular for developmental enhancers that tend to be more conserved [56, 57].

However, in the past a few years, the development of a plethora of next-generation sequencing (NGS)-based high throughput techniques has largely changed the way to characterize CREs or even CRMs genome-wide in large eukaryotic genomes. These techniques include ChIP-chip and ChIP-seq for locating CREs of a TF [15-17] and various chromatin modification marks [18], DNase-seq [19-21] and FAIRE-seq [20] for locating free nucleosome regions which tend to coincide with active CRMs, and Hi-C for measuring the physical proximity of linearly distal DNA segments [22, 23]. In particular, ChIP-seq techniques can potentially identify all possible (thousands to tens of thousands) binding regions of a TF in a cell type, tissue or developmental stage. However, these sequenced potential binding regions can be much longer than the CREs of the ChIP-ed TF. Thus, peak-calling algorithms and tools have been developed to identify the binding peaks in the potential binding regions. Even though the existing peak-calling algorithms can narrow down CREs of a ChIP-ed TF to a certain regions, typically from a few

hundred to a few thousand base pairs (bp) [58], they are still much longer than the typical lengths of CREs, which are typically 6~16bp long. Hence, the actual locations of CREs need to be identified by a motif-finding tool [59, 60]. Although a few new motif-finders have been developed to analyze large sequence sets from ChIP-seq experiments, such as seeder [61], Trawler [61, 62] , ChIPMunk [63], HMS [64], CMF [65], STEME [66], DREME [67], DECOD [68], RSAT [69], and POSMO [70], they are typically used to find the CREs of a ChIP-ed TF in a short region of sequences (~200bp) around the binding peak summits in order to reduce the searching space and increase prediction specificity in trading of sensitivity. Some of these tools [64, 70] use the locations of binding peaks to help find the CREs of a ChIP-ed TF. Thus, only CREs of the ChIP-ed TF are returned by these tools. However, CREs in higher eukaryotes rarely work alone. Instead, they cooperate with one another by forming CRMs for combinatorial regulations [54]. It has been shown that CREs of cooperative TFs of a ChIP-ed TF can be found in the neighborhoods of the binding peaks of the ChIP-ed TF [26, 27, 71-73]. In this sense, the information of CREs in a ChIP dataset is not fully explored by the majority of current studies that were mainly targeted to identify the CREs of a ChIP-ed TF.

With the continuous drop in costs of NGS technologies, TF ChIP-seq is becoming routine in numerous individual labs worldwide, and enormous ChIP-seq datasets are being produced in many important metazoans and plants, in addition to the large amount of ChIP data churned out by large consortiums such as the ENCODE [47, 74] and modENCODE [48] projects aimed at identifying all the functional sequence elements in the genomes of humans and the model organisms C. elegans [72] and D. melanogaster [71, 75]. It is highly expected that very soon, at least one ChIP-seq dataset will be

available in a certain cell type, tissue or developmental stage for the majority of TFs encoded in the genomes through both these efforts. Since certain combinations of TFs are often repeatedly used for regulating one or more groups (regulons) of genes in some cell types, tissues and developmental stages [51], the increasing number of ChIP-seq datasets contains a wealth of information about the combinatorial patterns of different TFs for transcriptional regulation [71, 72]. Thus, it is now possible to predict the CRMs and CREs genome-wide through integrating the information about co-occurrence of motifs in a large number of ChIP-seq datasets for different TFs from different cell types, tissues, developmental stages and physiological conditions. Although a few methods such as SpaMo [26], CPModule [27] and [28], have been made to identify CREs of cooperator TFs in a ChIP-seq dataset, they do not integrate multiple ChIP-seq datasets, and cannot predict novel motifs in CRMs, as they all depend on a library of known CREs such as TRANSFAC [29] or JASPAR [30] to scan for possible cooperative CREs in binding peaks. Consequently, simple and approximate methods were often used to find motifs in big ChIP datasets. For instance, in recent studies using the modENCODE [71] and ENCODE [76, 77] datasets, only the top 250 and 500 binding peaks with a length of 100bp and 200bp, respectively, in each dataset were used to find motifs. Hence, the wealthy information in the valuable ChIP datasets was not fully explored.

In this paper, we have developed a new algorithm DePCRM for genome-wide de novo prediction of CREs and CRMs by identifying overrepresented patterns of motif combinations in a large number of ChIP datasets in a sequenced eukaryotic organism. When applied to the *D. melanogaster* genome using a total of 168 ChIP-chip and ChIP-seq datasets for 56 TFs, DePCRM identified 184 CRE motifs and 115,932 CRMs,

recovering 77.9% of known CRMs located in the datasets and 89.3% of known CRMs containing at least one predicted CRE. Thus, the algorithm has achieved rather high prediction accuracy even using this limited number of datasets.

1.3 Results

1.3.1 Basic Idea of the Algorithm

As TFs in eukaryotes tend to work together by binding to their CREs in CRMs with a typical size of 500~3,000bp [78], we assume that although a ChIP experiment is mainly aimed to identify the binding locations of the ChIP-ed TF, if we extend shorter binding peaks toward the two ends to reach the typical size of CRMs (e.g., 3,000bp), then extended binding peaks are more likely to contain the CREs of different cooperative TFs in addition to the CREs of the ChIP-ed TF as illustrated in Figure 1.1. In other words, if two different TFs (e.g. the red circle and black circle TFs in Figure 1.1.) cooperatively regulate the same regulons in certain cell types by binding to their respective CREs in CRMs, then their extended ChIP binding peaks from these cell types should overlap with one another to some extent. Hence, if we have enough number of ChIP datasets for different TFs from the same and/or different cell types, then the datasets are likely to include overlapping binding peaks for cooperative TFs. Accordingly, our algorithm predicts CRMs through identifying overrepresented co-occurring putative motif patterns in a large number of ChIP datasets, ideally for different TFs in different cell types and developmental stages.

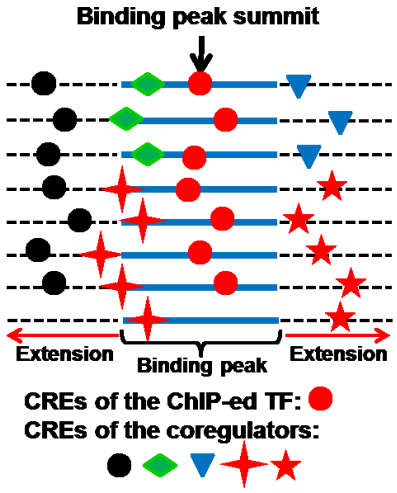

Figure 1.1. If the binding peak is shorter than 3,000bp, we equally extend from the two ends to have a length up to 3,000bp. We assume that in addition to the CREs of the ChIP-ed TF (red circle), CREs of different cooperative TFs (the other shapes) are also enriched in the neighborhoods of at least some subsets of the binding peak dataset. Each line represents an extended binding peak sequence.

More specifically, first, we identify all possible motifs in each of extended binding peak datasets (Figures 1.2.A and 1.2.B) using a fast motif finder. Second, we find overrepresented co-occurring motif pairs regardless of their distance in each of the datasets, and call them co-occurring pairs (CPs) (Figures 1.2B and 1.2.C). Third, we reason that if some highly similar CPs appear in multiple datasets, then all these similar CPs are likely to be subsets of the motifs of two certain TFs that cooperatively regulate regulons in different cell types or developmental stages, and therefore are likely to form CRMs by themselves or to be a part of larger CRMs. We identify such repeatedly occurring similar CPs in multiple datasets, and call them CP clusters (CPCs) (Figure 1.2.D). Presumably, each of the CPCs contains highly similar CPs for two certain TFs. Fourth, to predict CRMs containing more than two CREs, we cluster CPCs if they tend to co-occur in the same binding peaks (Figure 1.2.E). Each CPC cluster corresponds to a

possible combination of their motifs to form a part of or an entire CRM dependent on the sufficiency of the datasets, and thus we refer to them as CRM components (CRMCs) (Figure 1.2.F). Finally, we predict individual CRMs across the genome based on the motif pattern of the CRMCs and their close adjacency (Figure 1.2.G). Obviously, in order to accurately predict CRMs genome-wide, we need to have a sufficiently large number of diverse TF ChIP datasets, so that they likely include datasets for cooperative TFs in different cell types and developmental stages. We expect that the more diverse the datasets, the more accurate the predictions will be. The details of the algorithm are described in Methods.

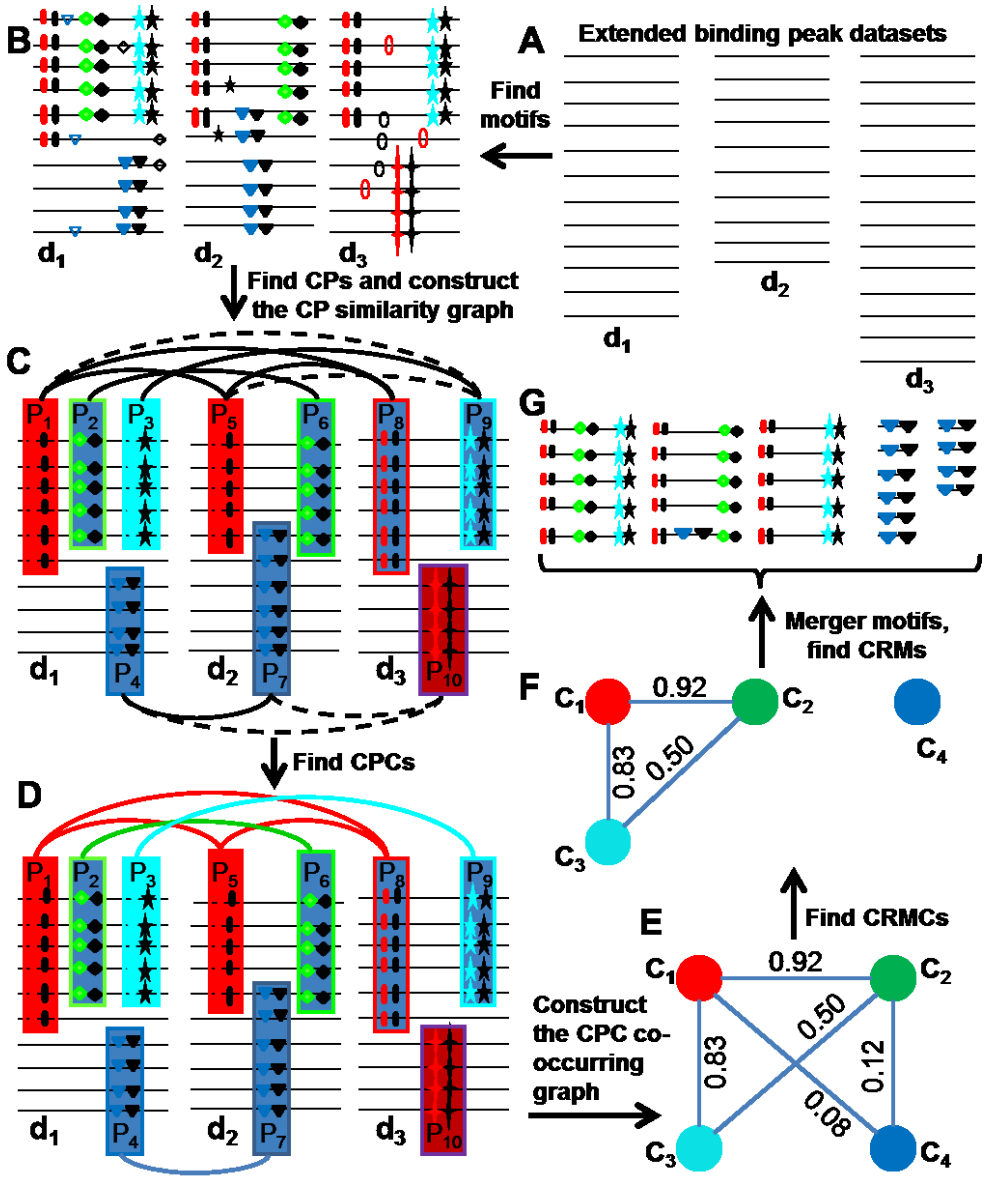

Figure 1.2. A schematic of the major steps of the DePCRM algorithm. A. Illustration of extended binding peaks from dataset d1, d2 and d3 respectively. B. Illustration of CREs found within each dataset, CREs of the same motif are shown in the same shape and color. C. Construction of CP similarity graph. [P1, P2, P3, P4], [P5, P6, P7] and [P8, P9, P10] are sets of CPs found in datasets d1, d2 and d3 respectively. For clarity, the CPs formed between motifs in P1 and motifs in P2 and so on in the datasets are not shown. Each CP (represented as a rectangle) is a node of the multi-partied similarity graph, and two nodes are linked by an edge if and only if their Ss $\geq\beta$, with Ss being the weight, which is not shown for clarity. D. By removing the dotted edges in panel C, MCL cuts the graph into five CP clusters (CPCs): C1=[P1, P5, P8]; C2=[P2, P6], C3=[P3, P9] , C4=[P4, P7] and C5=[P10]. CPs in a cluster are connected by edges in the same color. The singleton cluster C5=[P10] is discarded for its low density. E. For each pair Ci and Cj from the four CPCs, we find sets of CPs from the same dataset dk, and compute a co-occurring scores SCPC (Ci, Cj) for the two CPCs. F. Construction of the CPC co-

(Continued) occurring graph using the four CPCs. Cutting the graph using MCL results in two CRMCs, [C1,C2 ,C3] and [C4]. G. After merging motifs into Unique motifs (Umotifs), we project the CREs of CRMCs to the genome and predict the CRMs.

1.3.2 Overlap of the Extended Binding Peaks of Cooperative TFs in the Datasets

Since D. melanogaster has been long used to study gene transcriptional regulation in metazoans, a relatively large number of its CREs and CRMs have been experimentally characterized, and since a large number of ChIP-chip and ChIP-seq have been generated in the organism in the last few years, we evaluated our algorithm in this organism. To this end, we compiled a total of 168 ChIP-seq and ChIP-chip datasets for 56 distinct TFs, collected at different developmental stages (embryo, larva stage 1-3, pupa and adult female and male) and under different experimental conditions (heat shock and etc). More specifically, 42 ChIP-chip and 42 ChIP-seq datasets were from the ModENDCOE project [71, 75], 38 Chip-chip datasets were from the Berkeley Drosophila Transcription Network Project (BDTNP) [79], and 46 ChIP-chip datasets were from literature. Additional file 8: Table S1 summarizes the major features of the 168 datasets. As shown in Figure 1.3A, the majority of the binding peaks have a length around 1,000bp, and only 0.62% of them have a length longer than 5,000bp, which were not used in our study due to their low quality. Furthermore, if a binding peak is shorter than 3,000pb, we extended it up to 3,000pb (Methods) in order to include CREs of possible cooperative TFs (Figure 1.1). The final datasets contain a total of 445,252 sequences, each individual dataset containing 26 to 11,772 sequences (Figure 1.3B). These 445,252 sequences contain a total of 1,183,049,646bp, which are 7.0 times of the genome (168,736,537bp), but only cover 45.4% (76,555,033bp) of the genome (Table 1.1), indicating that some of these

sequences highly overlap with one another, thus confirming our aforementioned assumption. Of the 76,555,033bp genome sequence covered by the datasets, 64,033,300bp (86.3%) are in non-coding regions (NCRs, including introns and intergenic sequences), consisting of 47.7% of NCRs (134,207,178bp) in the genome (Figure 1.3C and Table 1.1). The remaining 12,521,733 (16.4%) sequences are in coding regions (CDRs), consisting of 36.3% of CDRs (34,529,359bp) in the genome (Figure 1.3C and Table 1.1). Thus we have included a considerable portion of CDRs in the datasets, because some binding peaks are located in CDRs. Currently, there are 1,830 known CRMs in D. melanogaster in the REDfly database [80], and 1,330 (72.7%) of which are located in the extended binding peaks, indicating that the available datasets are biased to the best-studied TFs in the organism. We will evaluate our algorithm for its ability to recover these 1,330 known CRMs in the extended binding peaks.

Table 1.1. Summary of the coverage of the datasets, predicted CRMs and CREs on the CDRs and NCRs of the genome

Categories	Size (bs)	% of genome	CDRs			NCRs		
			Size (bs)	% of category	% of genome	Size (bs)	% of category	% of genome
Genome	168,736,537	100.0	34,529,359	20.5	100.0	134,207,178	79.5	100.0
Datasets	76,555,033	45.4	12,521,733	16.4	36.3	64,033,300	83.6	47.7
CRMs	49,796,159	29.5	2,915,215	5.9	8.4	46,880,944	94.1	34.9
CREs	9,045,115	5.4	461,299	5.1	1.3	8,583,816	94.9	6.4

Table 1.2. Summary of the predictions of CRMs in the D. melanogaster genome at the major steps of the algorithm

Steps	Motifs		CPs		CPCs		CRMC	CRM	Known CRMs	
	Number	Percentage	Number	Percentage	Number	Percentage			Number	Percentage
Motif Finding	17890	NA	1,308,592	N/A	N/A	N/A	N/A	N/A	1,061	79.77%
CP Finding	1589	8.88%	4,891	0.37%	N/A	N/A	N/A	N/A	1,041	98.11%
CPC Finding	1376	86.60%	2,842	58.11%	951	N/A	N/A	N/A	1,036	99.52%
CRMC Finding	1316	95.64%	2,807	98.77%	937	98.53%	815	N/A	1,036	100.00%
CRM Finding	N/A	N/A	N/A	N/A	N/A	N/A	746	115,932	947	91.41%
Overall percentage		7.36%		0.21%		98.53%				77.89%

Each percentage is calculated based on the immediate previous step, except for the overall percentages which are based on the relevant initial step.

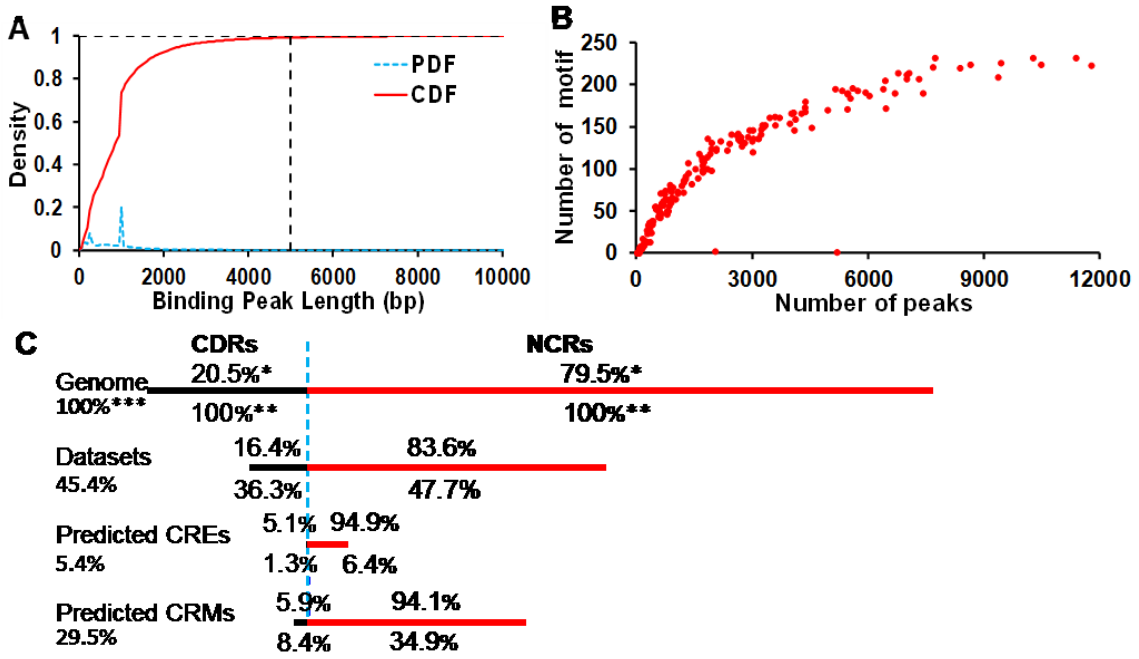

Figure 1.3. **A**. Distribution of the binding peak lengths in the 168 original datasets. Vast majority (99.38%) of the binding peaks are shorter than 5,000bp **B**. Number of motifs found in each of the 168 datasets as a function of the number of binding peaks the datasets. **C.** Coverage of the datasets, predicted CRMs and CREs on the CDRs and NCRs in the genome. The numbers above the lines are the proportions of the CDRs and NCRs in the corresponding sequence categories; the numbers below the lines are the proportions of CDRs and NCRs with respect to the entire CDRs and NCRs in the genome, respectively.

To see the overlapping patterns of binding peaks upon which our algorithm is based, we computed pair-wise overlapping scores (formula (1.1) in Methods) of the extended binding peaks among the 168 datasets for the 56 TFs (Table 1.1), and clustered the datasets using the overlapping scores. As shown in Figure 1.4, consistent with the above analysis, there are significant overlaps among the binding peaks in even these limited 168 datasets for only 5.3% (56/1,052) of the 1,052 annotated TFs encoded in the genome (flytf.org). As expected there are overlaps among datasets of the same TFs collected at differently developmental stages and/or under different experimental

conditions, indicating that these TFs might function similarly under these circumstances. For example, the datasets 2625 and 2626 from the ModENCODE project were collected using the same TF Caudal (CAD) at the embryonic stages 0-4 hours and adult female, respectively, and they have an overlapping score of 0.5. On the other hand, there are also numerous overlaps among datasets of different TFs. Interestingly, the datasets of TFs that are known to work cooperatively form clusters. The two highlighted boxes in Figure 1.4 show two examples of such clusters. The upper cluster is formed by the binding peaks for TFs Medea (MED), Dichaete (D), Dorsal (DL), Twist (TWI) and Daughterless (DA). It has been reported that DL and TWI cooperatively regulate the expression of Snail (SNA) in the mesoderm of the embryo [81]. The lower cluster is formed by the binding peaks of the global regulator CREB-binding protein (CBP), gap regulators Kruppel (KR), Giant (GT), CAD and Hunch back (HB). It has been well documented that these TFs bind to CRMs (enhancers/silencers) of genes involved in the segmentation process of early embryogenesis of D. melanogaster [80]. To further evaluate the overlaps of the binding peaks of distinct TFs, we analyzed the 56 out of the 168 datasets, each being for a different TF (if there are multiple datasets of a TF, we selected the one with the largest size), and the same conclusion can be drawn about the overlaps of the binding peaks of different TFs. The similar results also were reported in the ENCODE datasets in D. melanogaster [71] and human [82] datasets. Thus these results validate our assumption of the overlaps of binding peaks, and indicate that the datasets might contain sufficient information to predict at least portion of CRMs in the genome.

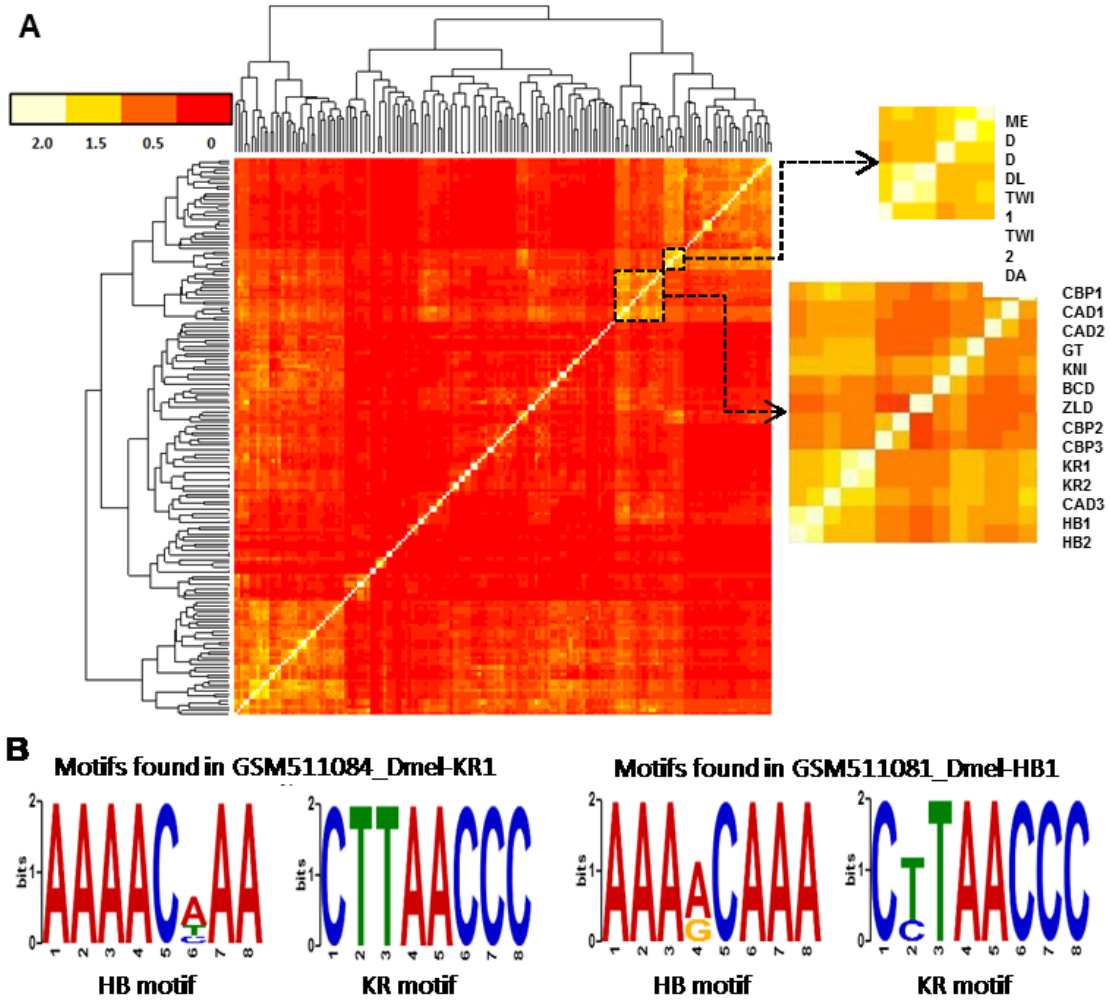

Figure 1.4. **A**. Hierarchical clustering of the 168 datasets for 56 TFs based on their pair-wise binding peak overlapping scores S_o. The blow-ups show two clusters for cooperative TFs (see Results). **B**. The motifs of TFs KR and HB are both found in the overlapping datasets GSM511084_Dmel-KR1 ChIP-ed by KR and GSM511081_Dmel-HB1 ChIP-ed by HB.

1.3.3 Identification of Motifs in the Extended Binding Peaks

Our goal now is to identify in each of the extended binding peak datasets all possible TF binding motifs of the ChIP-ed TFs as well as of its cooperative TFs (Figures 1.1, 1.2A and 1.2B). Because accurate motif-finding is still a notoriously difficult problem [14, 83-85], to achieve this goal we consider all overrepresented motifs returned

by DREME [67] in each extended binding peak datasets to maximally include possible true motifs. As shown in Figure 3B, depending on the size and quality of the datasets, a varying number (0~231) of motifs were found in each dataset. Particularly, in a total of six datasets that generally contain fewer binding peaks and are of low quality (26, 26, 28, 28, 70 and 5,188 sequences, Figure 3B), none or only a single motif could be identified. As no motif pairs can be formed in these datasets, they did not contribute to the final CRE and CRM predictions. In other words, they were filtered out by the motif-finder. On the other hand, putative CREs were found in the vast majority (99.98%) of the 439,886 extended binding peaks in the remaining 162 datasets, indicating that they were highly enriched with motifs. In this sense, the motif finding step servers as a quality control to filter out low quality datasets without the need of human involvement, conferring additional robustness to the algorithm. The returned motifs from the 162 datasets for 56 TFs (no TF was eliminated by discarding the six datasets) generally have high information contents (Figure 1.5A). Importantly, the known motifs of the ChIP-ed TFs were found by DREME for 99 of the 168 datasets, and were generally ranked high by the program, although they were usually not the top hit of DREME (Figure 1.5B), suggesting that it is necessary to consider a sufficient number of returned motifs to include the true ones. Moreover, when the datasets of different TFs have significant overlaps, we can identify all the motifs of the ChIP-ed TFs in all the overlapping datasets. For instance, the dataset GSM511084 for TF KR significantly overlaps with the dataset GSM511081 for TF HB, and motifs highly similar to the known binding sites of KR and HB were found in both the datasets (Figure 1.4B). Overall, we identified a total of 17,890 putative motifs corresponding to 35,359,819 putative CREs in the 168 datasets. These 35,359,819

putative CREs contain 275,857,398bp which are 1.6 times of the genome, but only cover 30.9% (52,078,901bp) of the genome, indicating that some of them still overlap with one another. At least one putative CRE was found in 1,061 (79.8%) of the 1,330 known CRMs in the sequences (Table 1.2). The failure to find CREs in the remaining 269 known CRMs in the datasets could be due to the fact that the CREs in these CRMs were not enriched in the datasets. Nonetheless, these results strongly suggest that in addition to the CREs of the ChIP-ed TFs, CREs of cooperative TFs, and thus at least partial CRMs are highly enriched in the extended binding peaks. This conclusion is in agreement with an early study based on 38 ChIP datasets in D. melanogaster [71] and also is supported by two recent studies using human datasets [82, 86].

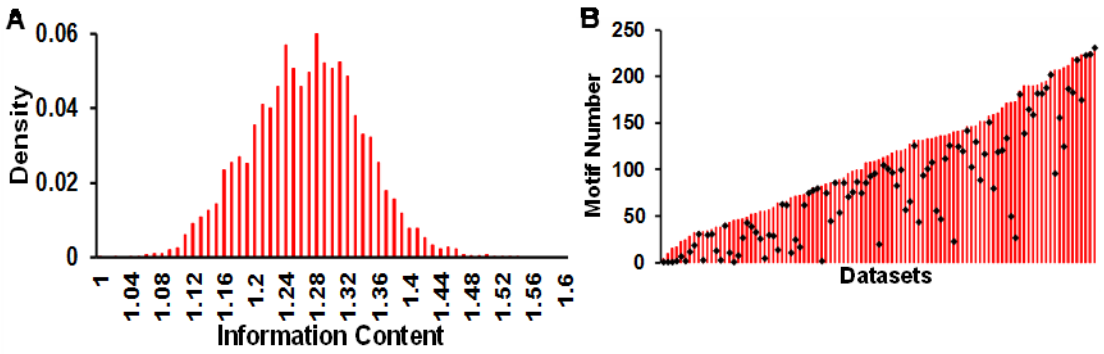

Figure 1.5. **A**. Distribution of the information content of the predicted motifs in the datasets. **B**. The rank of the ChIP-ed TF's motif among the predicted motifs in the 99 datasets in which the motifs of the ChIP-ed TFs can be identified. The diamond on the bar indicates the rank of the ChIP-ed TF's motif among the predicted motifs in the dataset. The higher the position of the diamond, the higher the rank of the target TF's motif.

1.3.4 Prediction of CRMs by Iteratively Enriching Repeatedly Used Motif Combinatorial Patterns

Clearly, as we used a rather loose stringency in motif finding to maximally include true motifs, there are inevitably a large number of spurious predictions in the 17,890 putative motifs identified in the datasets. Thus, our algorithm takes these 17,890 putative motifs as the input, and predicts CREs and CRMs by iteratively enriching repeatedly used motif combinatorial patterns though gradually filtering out spurious ones. Specifically, DePCRM first identifies highly co-occurring motif pairs (CPs) in each dataset by computing a co-occurring score (S_c) (formula (1.2)) for each pair of putative motifs found in each dataset (Figures 1.2C). As shown in Figure 1.6A, the distribution of S_c is strongly skewed toward right, indicating that there are multiple components of the S_c values. The left low-scoring component can be well fitted to a Gaussian distribution with a mean and standard deviation 0.19 and 0.0043, respectively. The motif pairs accounting for this component are more likely to co-occur by chance, and thus, they are likely spurious motif pairs. On the other hand, the right high-scoring portion of the distribution is more likely to attribute to true cooperative motif pairs. To find the S_c cutoff α by which a maximal number of motif pairs occurring by chance are filtered out while a maximal number of possible true motif pairs are kept, we plotted the proportion of the motif pairs with a $S_c \geq \alpha$ as a function of α. As shown in Figures 1.6A and 1.6B, when α =0.7, 1,303,701(1,303,701/1,308,592=99.6%) motif pairs and 16,301 motifs (16,301/17,890=91.1%) were filtered out, while putative CREs in only 20 (1.8%) the known 1,061 CRMs containing predicted CREs were completely left out. Thus we selected the motif pairs with $S_c \geq \alpha$ =0.7 as CPs for further analysis, thereby discarding the vast majority of presumably randomly occurring motif pairs (99.63%) and motifs

(91.12%). This results in 4,891(4,891/1,308,592=0.4%) CPs containing 1,589 (1,589/17,890=8.9%) motifs (Table 1.2) for further analysis, which are presumably enriched for true motif pairs and motifs.

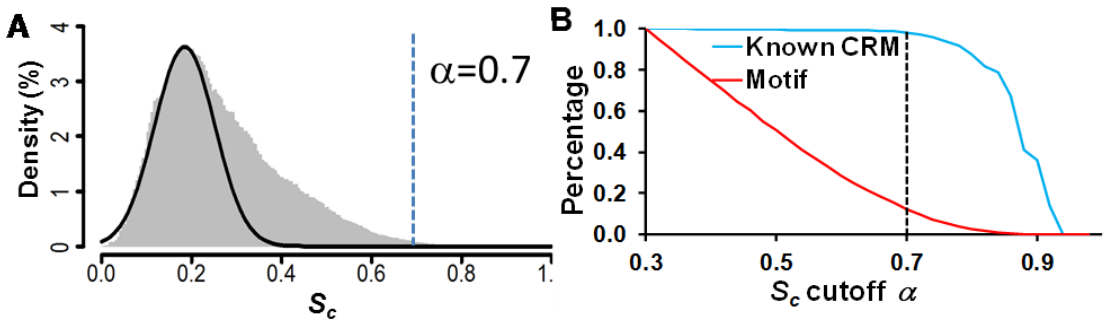

Figure 1.6 **A**. Distribution of co-occurring scores S_c of the motif pairs found in the 168 datasets. The curve is a fitting of the left portion of the distribution to a Gaussian distribution N($\mu = 0.19$, $\sigma = 0.067$). **B**. The remaining proportions of predicted motifs and known CRMs as functions of the S_c cutoff α. The vertical line indicates the position of the chosen cutoff $\alpha = 0.7$ for selecting co-occurring motif pairs (CPs).

To further enrich true motif pairs and motifs, the algorithm identifies repeatedly used CPs by clustering highly similar CPs in different datasets. To this end, we computed a similarity scores S_s (formula (1.3)) for each pair of CPs, each from two different datasets; and then constructed a CP similarity graph based on an S_s cutoff value β (Figure 1.2C). As shown in Figure 1.7A, with the increase in β, the density of the graph drops rapidly, but the dropping starts slowing down around β =1.36; meanwhile the number of nodes (CPs) in the graph starts decreasing rapidly around β =1.36 (Figure 1.7B). Thus, we set β =1.36 to construct the CP similarity graph (Methods). Applying the Markov chain clustering (MCL) algorithm [87] to the graph (Figure 1.2D) resulted in 951 CP clusters (CPCs) containing 2,842 (2,842/4,891=58.1%) CPs and 1,376 (1,376/1,589

=86.6%) motifs (Table 1.2). Thus we further filtered out 2,049 (2,049/4,891=41.9%) CPs and 213 (213/1,589=13.4%) putative motifs.

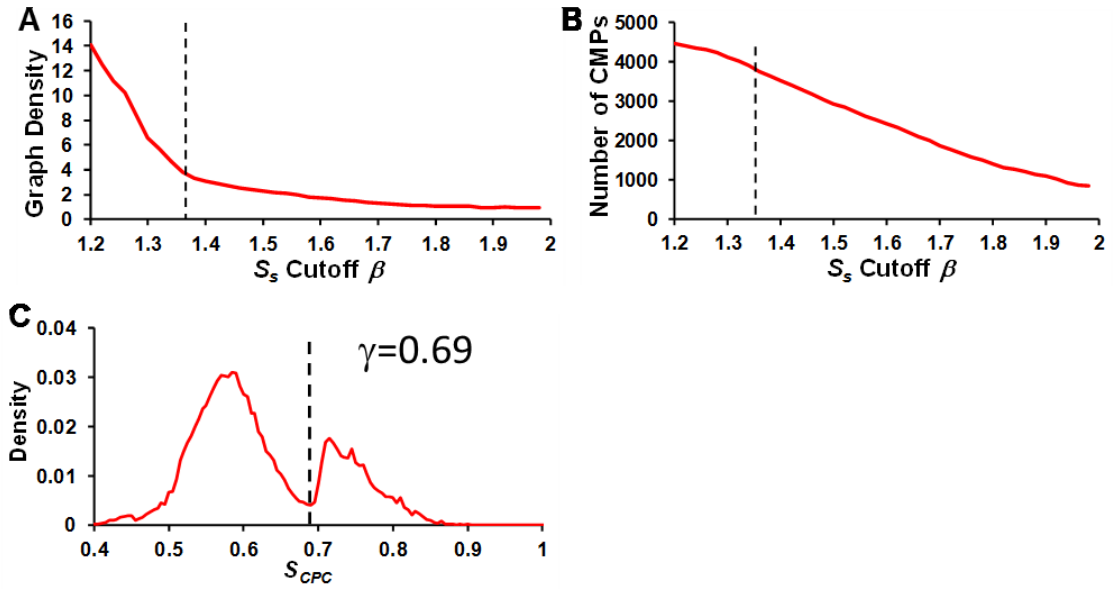

Figure 1.7. **A**. The density of the CP similarity graph drops rapidly with the increase in the S_s cutoff β, but the trend of decrease slows down around β =1.36. **B**. The number of CRM in the graph also starts to drop rapidly around β=1.36. Thus we set β = 1.36 for construing the final CP similarity graph. **C**. The distribution of CPC co-occurring scores Figure 1.7(Continued) S_{CPC} are well separated into a low-scoring component and a high-scoring component. The vertical line indicates the S_{CPC} cutoff γ = 0.69 at the deepest valley between the two peaks, for constructing the CPC co-occurring graph.

Next, to identify larger repeatedly used motif patterns, we computed a co-occurring score S_{CPC} (formula (1.5)) for each pair of CPCs across the datasets in which both the CPCs have motifs. Interestingly, as shown in Figure 1.7C, the S_{CPC} scores display a well-separated bimodal distribution, and the low-scoring peak is likely mainly due to random motif patterns, while the high-scoring one is more likely attributable to truly cooperative motifs, thus we considered CPC pairs with an $S_{CPC} \geq \gamma = 0.69$ (at the valley between the two peaks) for further analysis. Applying the MCL algorithm to the resulting CPC co-occurring graph (Figures 1.2D and 1.2E, Methods), gave rise to 815

CRM components (CRMCs) containing 937(937/951=98.5%) CPCs, 2,807(2,807/2,842=98.8%) CPs and 1,316 (1,316/1,376=95.6%) motifs (Table 1.2). The compositions and structures of these 815 CRMCs are shown in Figure 1.8, each containing 1~9 CPCs. Overall, 16,574 (92.6%) of the original 17,890 input motifs were filtered out by the algorithm (Table 1.2), suggesting that at least the vast majority (92.6%) of the putative motifs found in the datasets are spurious predictions.

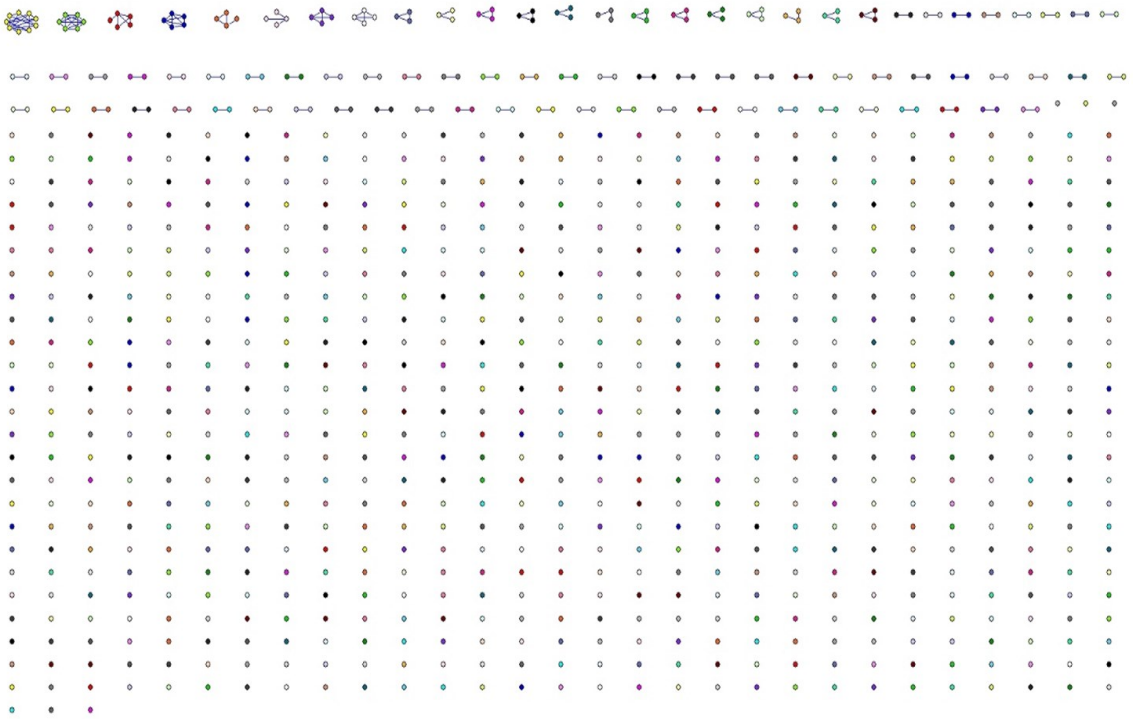

Figure 1.8. Structures of the 815 CRMCs. Each node in graphs is a CPC, and each connected graph represents a CRMC.

As expected, some of the resulting 1,316 motifs found in different datasets are highly similar and often overlap with one another as demonstrated by the examples shown in Figure 1.4B. They are likely recognized by the same TFs or closely related ones, thus need to be combined into non-redundant and unique ones. To this end, we

iteratively clustered the final 1,316 motifs based on their similarities (Method), resulting in 184 clusters. We consider each cluster as a unique motif and refer to it as a Umotif, each containing 1 or 2~108 highly similar motifs and 255~88,702 CREs (Figure 1.9, Table 1.3). When compared with the known motifs in multiple built-in databases including DMMPMM, iDMMPMM, flyreg and fly factor survey using TOMTOM [88-91], 111 (60.3%) of the Umotifs are highly similar to known motifs in D. melanogaster at $p<0.001$ (Supplementary file 1), strongly suggesting that they are likely to be true motifs. Examples of such Umotifs, their constituent motifs and the known motifs hit are shown in Figures 1.10A and 1.10B. The rest 73 Umotifs that does not resemble any known motif might be novel ones. Examples of such Umotifs, their constituent motifs are shown in Figures 1.10C and 1.10D. Furthermore, 106 (29.4%), 203 (56.2%) and 269 (74.5%) of 381 possibly redundant motifs found in the earlier study [71] were recovered by the Umotifs with a p-value cutoff of 0.001, 0.005 and 0.01, respectively. We replaced the motifs in the CRMCs with the Umotifs that they belong to, and each of the CRMCs is represented by their constituent Umotifs. Some CRMCs contain the same combination of Umotifs, thus we merged them in a unique one, resulting in 746 CRMCs.

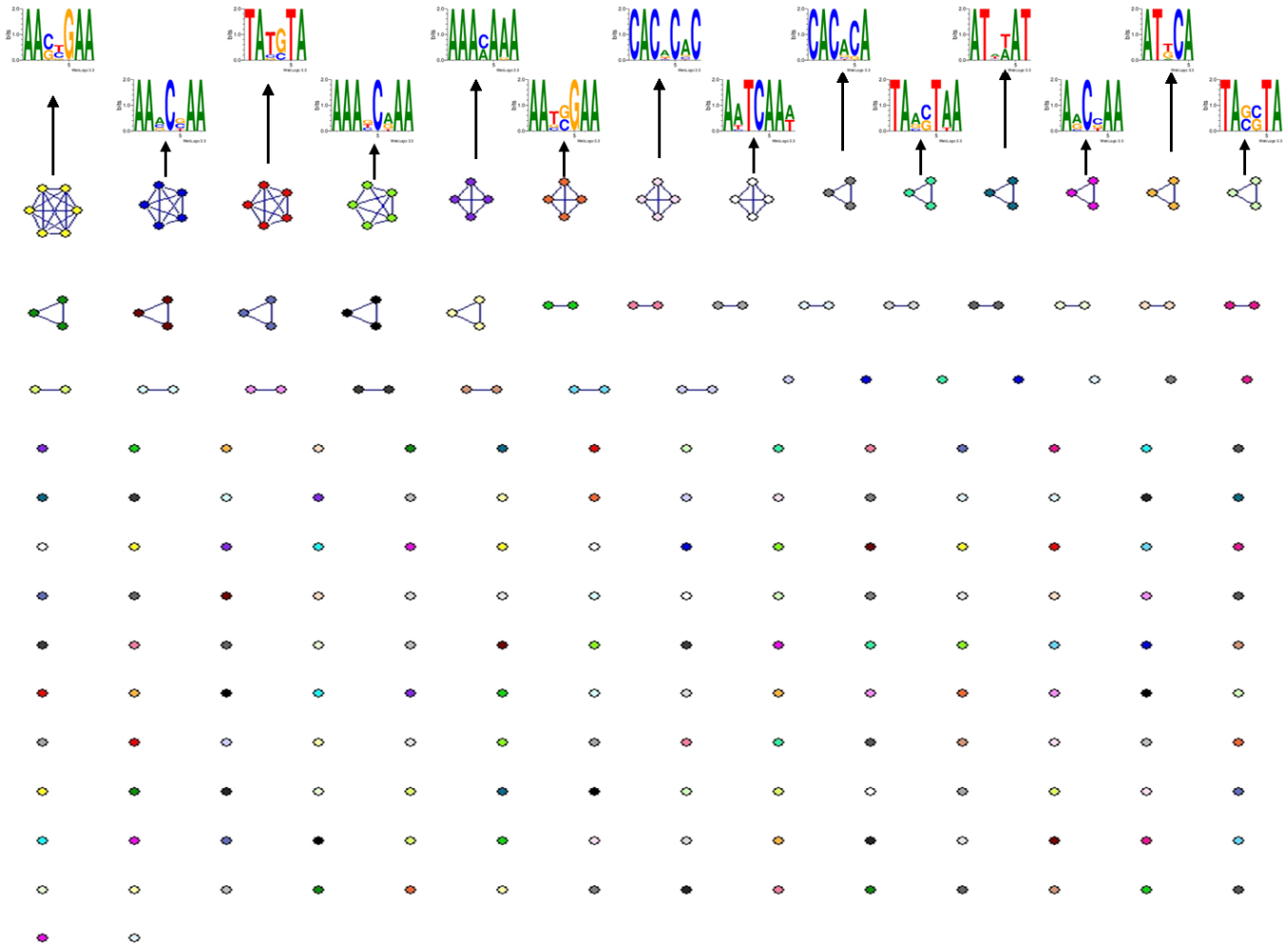

Figure 1.9. Structures of the 184 Umotifs. Each node in graphs is a putative motif, and each connected graph represents a Umotif. The logos are for the indicated Umotifs.

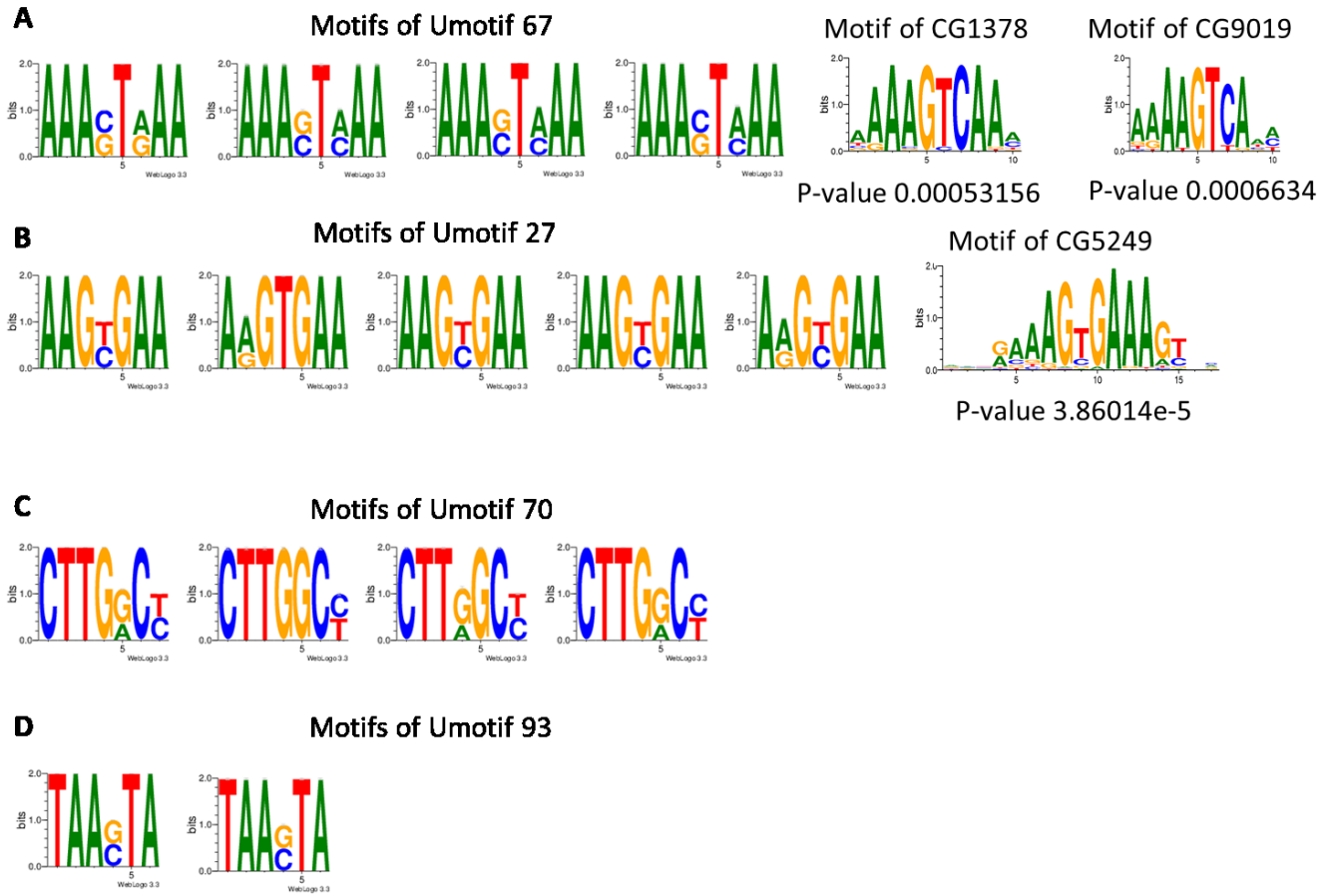

Figure 1.10. **A**. Umotif 67 and its four individual constituent motifs found in different datasets. Umotif 72 is similar to known motifs CG1378 and CG9019. **B**. Umotif 27 and its five individual constituent motifs. Umotif 27 is similar to known motif CG5249. **C**. Umotif 70 and its 4 individual constituent motifs found in different datasets. **D**. Umotif 93 and its 2 individual constituent motifs found in different datasets.

1.3.5 Genome-wide Predictions of CREs and CRMs in D. melanogaster

Projecting the CREs in these 746 CRMCs back to the D. melanogaster genome (Methods) resulted in a total of 1,108,018 non-overlapping CREs with an average of 8.2±2.8bp, with 53,785 (4.9%) of which being entirely located in CDRs. These 1,108,018 CREs cover 9,045,115bp (5.4%) genome sequence, of which 8,583,816bp (94.9%) are in NCRs, consisting of 6.4% of NCRs; the remaining 461,299bp (5.1%) are in CDRs, consisting of 1.3% of CDRs (Figure 1.3C and Table 1.1). By connecting these putative CREs (Methods), we predicted a total of 115,932 non-overlapping CRMs, 71,817 (61.9%) of which are entirely located in NCRs, and the remaining 44,115 (38.1%) contain CDRs. These 115,932 CRMs cover 49,796,159bp (29.5%) genome sequence, 46,880,944bp (94.1%) of which are in NCRs, consisting 34.9% of NCRs; the remaining 2,925,215bp (5.9%) are in CDRs, consisting of 8.4% of CDRs (Figure 1.3C and Table 1.1). These putative CRMs tend to have shorter lengths than those of the known CRMs (Figure 1.11A). Furthermore, the putative CRMs harbor 2 to 146 with a median of 7 CREs, and the distances between adjacent two putative CREs are largely similar to those in known CRMs, except that a small portion of the putative CRMs tend to have a short distance between adjacent two putative CREs (Figure 1.11B). These results suggest that we might have missed certain CREs in the predicted CRMs, particularly at the two ends, presumably due to insufficient information in the limited number of available ChIP datasets used in this study. In other words, some of our predictions might consist of only a part of real CRMs with possible missing CREs at the two ends of the CRM. Clearly, in order to make more accurate and complete predictions, more and highly diverse ChIP datasets are needed.

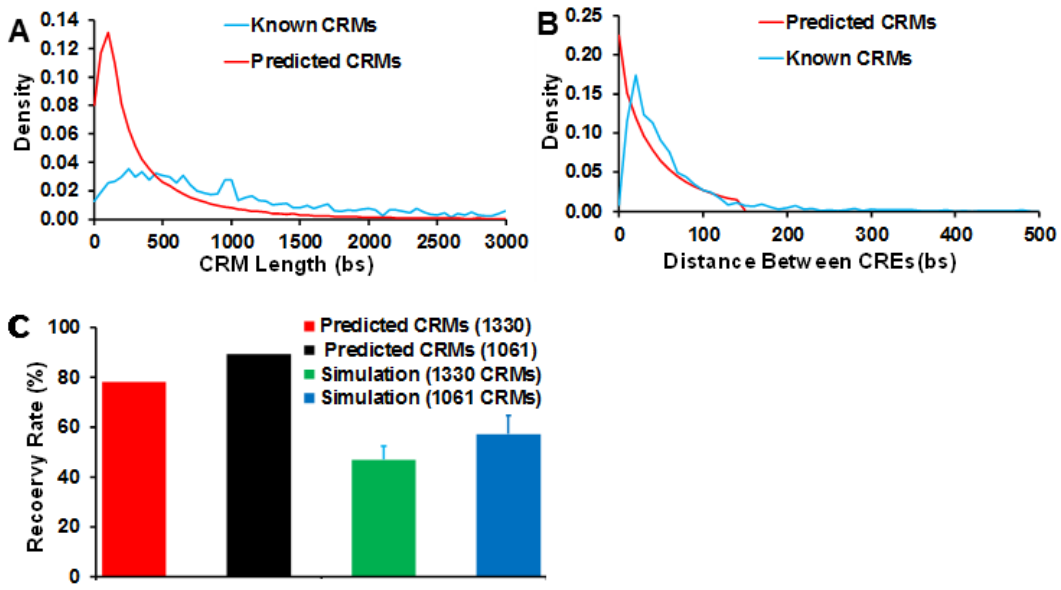

Figure 1.11. **A**. Distribution of the lengths of the known and predicted CRMs. **B**. Distribution of the distances (bp) between two adjacent CREs in the known and predicted CRMs. **C**. Recovery rates of the known CRMs in the datasets (1330) and the known CRMs containing a predicted CRE (1061) by the predicted CRMs and the corresponding same number and length sequences randomly selected from NCRs.

To evaluate the sensitivity of our predicted CRMs, we first computed the recovery rate by the predicted CRMs of the 1,330 known CRMs contained in the datasets. We consider a known CRM is recovered if it overlaps with a predicted CRM by at least half of its length. Remarkably, 1,036 (77.9%) of the 1,330 known CRMs were recovered by the 115,932 putative CRMs (Table 1.2). By contrast, when the same number and length sequences were randomly selected from the genome region covered by the datasets, only 46.9±5.5% (n=50) (Figure 1.11C) of the 1,330 known CRMs could be recovered. The recovery rate for the 1,061 known CRMs, in which at least a putative CRE was found, was even higher (947/1,061=89.3%). By contrast, when the same number and length sequences were randomly selected from the genome region covered by the dataset, only

57.2±7.6% (n=50) (Figure 1.11C) of the known CRMs were recovered. Hence, our algorithm has achieved rather a high recovery rate or sensitivity of CRM predictions, in particular when a putative CRE could be identified in them, even using just the limited 168 datasets for only 56 TFs. Importantly, some of the known CREs in these recovered CRMs overlap with our predicted CREs. For example, CRM(3R:21859748.. 21862775) containing Umotif 34 recovers a known CRM of gene e(spl); and a putative CRE of Umotif 34 overlaps with the known CRE of TF DA in the CRM, while Umotif 34 is highly similar to the known motif of DA (Figure 1.12.A). Furthermore, CRM (2L: 15731775..15732968) containing Umotifs 106 and 114 recovers the known CRM of gene cycE; moreover, Umotifs 106 and 114 are highly similar to the known motifs of HTH and KNI which also have CREs located in the recovered CRM, respectively (Figure 1.12.B and C). In addition, many of our novel predictions also have strong experimental data supports thus are likely to be authentic. For example, our predicted CRMs 3R:8896195..8898063 , 3R: 12636031..12636729 and 2R: 5984055..5984519 share Umotifs 3 and 14, and they recover the known CRMs of genes abd-A, jun-realted antigen (jra) and single-minded (sim). It has been shown that these three genes are involved in nervous system development [92-94], and thus are likely to be coregulated. Consistent with this, we identified CREs of Umotifs 3 and 14 in the regulatory regions of these genes. Interestingly, Umotifs 3 and 14 are highly similar to the known motifs of hormone receptor 51 (HR51) and ladybird early (LBE), receptively (Figure 1.12.D and E), and it has been reported that HR51 and LB regulate neurogenesis [95, 96]. Thus, HR5 and LB might carry out their functions by binding to the putative CREs of Umotifs 3 and 14. Furthermore, we have predicted a CRM 2R: 16831599..16832019 overlaps with the first

intron of gene actin57B (Figure 1.13) containing Umotif 27 and 23, which are highly similar to the known motifs of TFs myocyte enhancer factor 2 (MEF2) and chorion factor 2 (CF2), respectively (Figure 1.12.F and G). It has been shown that these two TFs cooperatively regulate Actin57B by binding to its promoter region [97]. Thus, MET2 and CF2 might also regulate actin57B through binding to the putative CRES of Umotifs 27 and 23 located in its first intron (Figure 1.13). Therefore, our predicted CREs and CRMs can help biologists identify potential enhancers for genes of interest.

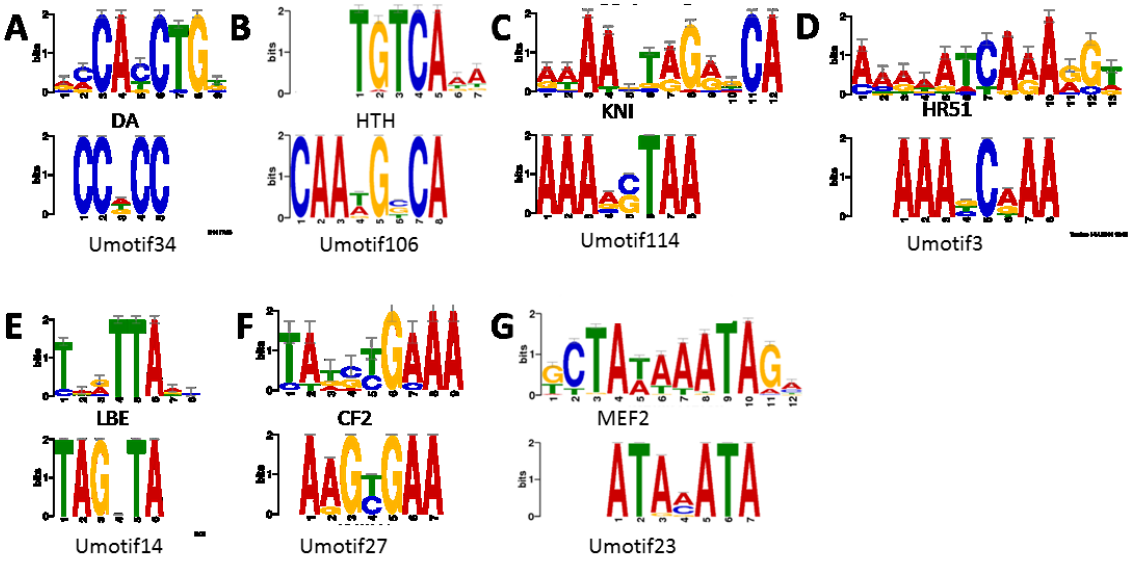

Figure 1.12. Examples of known CREs in the recovered known CRMs that overlap with our predicted CREs, their corresponding Umotifs are similar the known motifs.

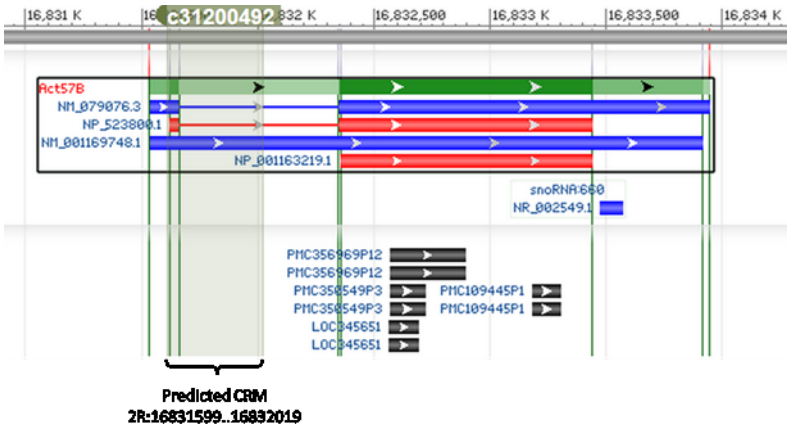

Figure 1.13. A putative CRM (shown in gray shadow) is located in the first intron of gene act57B.

1.3.6 The Predicted CRMs as well as CREs in a CRM as a Whole are More Conserved than Randomly Selected Sequences

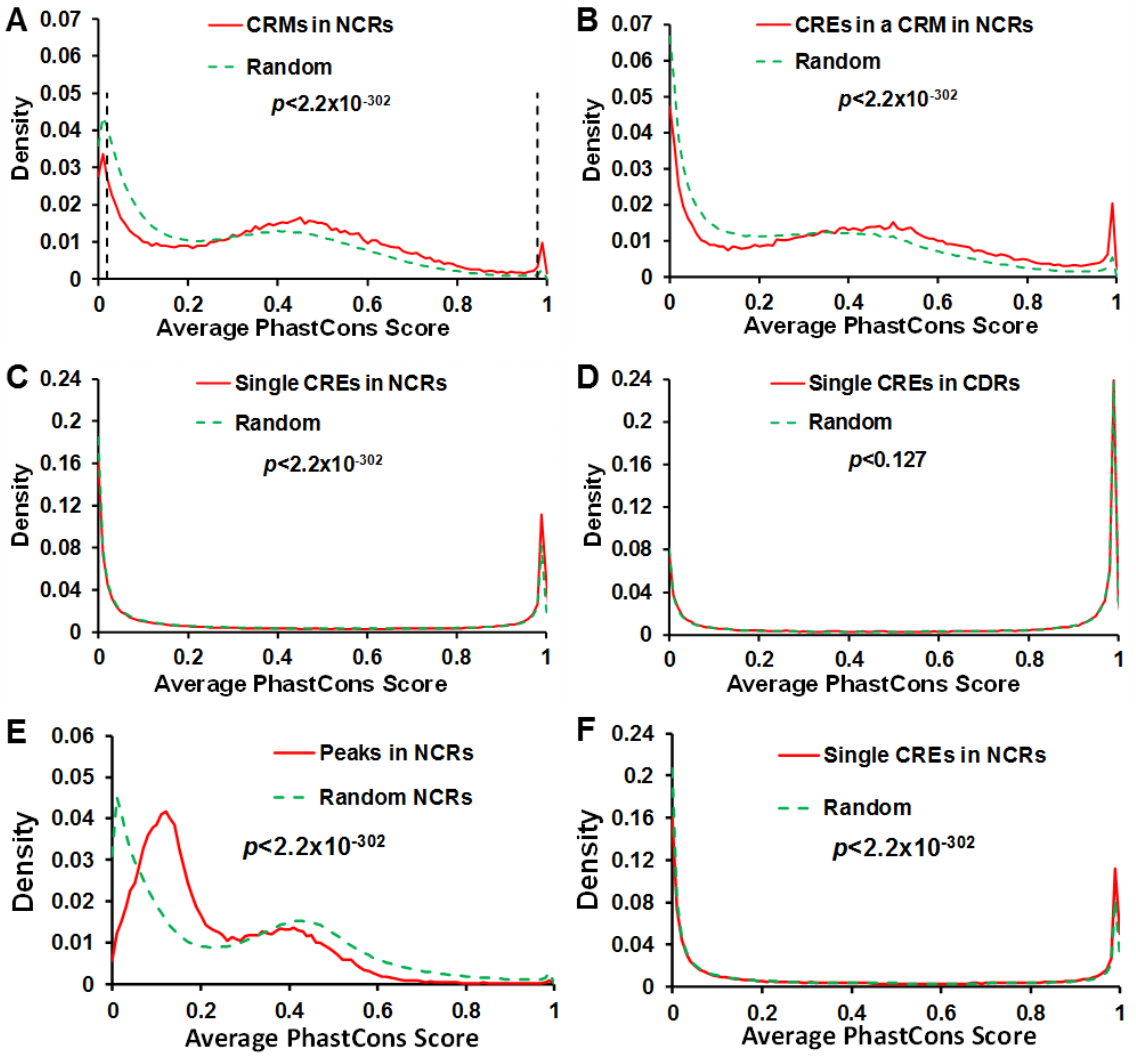

Figure 1.14. A. Distribution of average phastCons scores of the predicted CRMs in NCRs and of the same number and length sequences randomly selected from NCRs. The vertical dashed lines indicate the PhastCons score cutoffs for highly conserved (≥0.98) and non-conserved (≤0.02) CRMs. B. Distribution of average phastCons scores of all putative CREs in a predicted CRMs in NCRs and of the same number and length sequences randomly selected from NCRs. C. Distribution of average phastCons scores of single predicted CREs in NCRs and of the same number and length sequences randomly selected from NCRs. D. Distribution of average phastCons scores of single predicted CREs in CDRs and of the same number and length sequences randomly selected from CDRs. E. Distribution of average phastCons scores of the non-redundant original binding peaks in NCRs and of the same number and length sequences randomly selected from NCRs. F. Distribution of average phastCons scores of single predicted CREs in the

original binding peaks in NCRs and of the same number and length sequences randomly selected from the original binding peaks in NCRs.

As functional sequences tend to be more conserved than non-functional ones, to further evaluate our predicted CRMs and CREs, we first compared the average phastCons conservation scores [98] of the nucleotides in each of the putative 71,817 CRMs entirely located in NCRs with those of the same number and length sequence randomly selected from NCRs. The phastCons score is computed as the posterior probability for a nucleotide to be conserved given a multiple alignment of genomes and their phylogenetic tree [98]. As shown in Figure 1.14A, although the average phastCons scores of both the predicted CRMs in NCRs and the randomly selected sequences have tri-modal distributions, they are significantly different ($p<2.2x10\text{-}302$, Kolmogorov–Smirnov test). Specifically, the right peak with very low phastCons scores, which reflects highly mutated sequences is much larger for the former than for the latter, and the opposite is true for the left peak with very high phastCons score, which reflects highly conserved sequences [98]. Moreover, the middle peak with intermediate phastCons scores, which reflects neutral to moderately conserved sequences [98], shifts about 0.04 to right for the former relative to that for the latter. Thus, the nucleotides in the predicted CRMs in NCRs tend to be more conserved than those in the randomly selected sequences. As the spacing sequences between CREs in a CRM may not necessarily be functional and thus conserved, we next compared average phastCons scores of putative CREs in each of the 71,817 predicted CRMs in CDRs with those of the same number and length sequences randomly selected from NCRs. As shown in Figure 1.14B, average phastCons scores of CREs in a CRM and randomly selected sequences from NCRs also show tri-modal

distributions, however again, they are significantly different ($p < 2.2 \times 10^{-302}$, Kolmogorov–Smirnov test) in the similar way as for those of the full length putative CRMs and the corresponding randomly selected sequences (Figure 1.14A). However, there are subtle differences between the two cases: compared to the difference between the peaks for the putative CRMs and the randomly selected sequence (Figure 1.14A), the right peak for the putative CREs is much larger than that of the randomly selected sequences (Figure 1.14B), and the middle peak for the putative CREs shifts more (0.15 vs. 0.04 unit) to right relative to that of the randomly selected sequences (Figure 1.14B). Hence, putative CREs in a CRM as a whole are much more conserved than the randomly selected NCRs, and also more conserved than spacer sequences in the putative CRMs. We further compared average phastCons scores of nucleotides in the 646,143 putative single CREs in the 71,817 predicted CRMs in NCRs and in the 53,785 putative CREs in CDRs with the same number and length sequences randomly selected from NCRs and CDRs, respectively. As shown in Figure 1.14C and 1.14D, the average phastCons scores of single putative CREs in both NCRs and CDRs and those of the corresponding randomly selected short k-mer sequences all show well separated bi-modal distributions, with each peak located near the two extremes (0 and 1) of phastCons scores. This result indicates that nucleotides in single putative CREs in both NCRs and CDRs and their corresponding randomly selected short k-mers all tend to have either a very low (near zero) or a very high (near 1) average phastCons score, implying that the nucleotides in short sequences tend to be simultaneously highly conserved or non-conserved. This observation is consistent with the findings that the D. melanogaster genome is highly compact, and vast majority of its sequences are either negatively or positive selected, and

thus are likely to be functional [99-105]. However, interestingly, there are striking differences between the predicted CREs in NCRs (Figure 1.14C) and those in CDRs (Figure 1.14D). First, the distribution for single putative CREs in NCRs is significantly different from that for the corresponding randomly selected sequences ($p<2.2x10-302$, Kolmogorov–Smirnov test), as the right peak of the former is slightly larger than that of the latter (Figure 1.14C), indicating that a small fraction of single predicted CREs in NCRs are more conserved than the randomly selected short k-mers. By contrast, the distributions for single putative CREs in CDRs and the corresponding randomly selected short k-mers are not significantly different ($p<0.127$, Kolmogorov–Smirnov, Figure 1.14D), indicating that single putative CREs in NCRs are not more conserved than the randomly selected short k-mers. Second, the right peaks for single predicted CREs in NCRs and the randomly selected short k-mers are slightly smaller than their own left peaks (Figure 1.14C), indicating that there are slightly fewer conserved short sequences than non-conserved ones in NCRs. By contrast, the right peaks for single putative CREs in CDRs and the randomly selected short k-mers are much larger than their own left peaks (Figure 1.14D), indicating that there are much more conserved short sequences than non-conserved ones in CDRs, which is expected as most CDRs are highly conserved. Third, the right peaks for single putative CREs in NCRs and the corresponding randomly selected k-mers are much smaller than those of single putative CREs in CDRs and the corresponding randomly selected short k-mers, and the opposites are true for the left peaks (Figure 1.14C and 1.14D), indicating that short sequences in CDRs are more conserved than those in NCRs as expected. Finally, to see the extent to which the original binding peaks (without length extension) in the datasets were enriched

for CRMs and CREs, we computed average phastCons scores of the non-redundant original binding peaks and the CREs contained as well as of the same number and length sequences randomly selected from NCRs and NCRs in the binding peaks, respectively. As shown in Figure 1.14E, the distribution of average phastCons scores of non-redundant original binding peaks was quite different from that of putative CRMs. In particular, the peak at the score =1 in the latter distribution was almost missing in the former distribution. Moreover, the original binding peaks with an average phastCons score > 0.32 even tended to be less conserved than randomly selected NCRs, and the opposite was true for the putative CRMs, indicating that the predicted CRMs contains more conserved sequences than do the original binding peaks. Furthermore, the distribution difference between average phastCons scores of CREs predicted in the original binding peaks and those of randomly selected NCRs with the same lengths is similar to that between average phastCons scores of CREs and those of the randomly selected NCRs of the binding peaks (Figure 1.14F). Thus, our predicted CREs in extended binding peaks as a whole are of similar quality to the predicted CREs in the original binding peaks. In summary, although only a small fraction of the single predicted CREs in NCRs are more conserved than the randomly selected short k-mers, predicted CREs in a putative CRM as a whole and predicted CRMs are significantly more conserved than the corresponding randomly selected sequences, thus they are highly likely to be functional.

1.3.7 Highly Conserved and Non-conserved CRMs Regulate Distinct Classes of Genes

To further evaluate our predicted CRMs, we examined whether or not the highly conserved predicted CRMs (with an average phastCons score ≥ 0.98) and highly non-conserved predicted CRMs (with an average phastCons score ≤0.02) (Figure 1.14A) have

distinct regulatory functions. To this end, we assigned each of the predicted CRMs a target gene whose transcription start site has the shortest distance to the predicted CRM. Thus, a predicted CRM can only be assigned to a gene while a gene can have multiple assigned putative regulating CRMs. A total of 763 and 2,319 genes are predicted as targets of the highly conserved and highly non-conserved putative CRMs, of which 601 and 2,053 have gene ontology (GO) annotations, respectively. As shown in Supplementary file 2, 134 (22.3%) the putative target genes of the 601 highly conserved putative CRMs are clustered into 11 functional groups using the DAVID program [106] with an enrichment score ≥ 1.5 and $p<0.01$ (hyper-geometric test with Benjamini correction). Intriguingly, these genes are enriched for developmental functions (8 groups), neurological functions (1 group), motility (1 group) and transcriptional regulations (1 group). On the other hand, 481(23.4%) putative target genes of the 2,053 highly non-conserved putative CRMs are clustered into 10 functional groups with an enrichment score ≥ 1.5 and $p<0.01$, In contrast to the putative target genes of highly conserved putative CRMs, these genes are enriched for plasma membrane functions (6 groups), metabolism (2 groups), and chemical sensory perception (2 groups) (Supplementary file 3). Thus, the highly conserved putative CRMs and highly non-conserved putative CRMs do regulate distinct groups of genes. The results are in excellent agreement with the fact that highly conserved CRMs are mainly involved in embryonic development in both insects [107, 108] and vertebrates [109], while CRMs for genes with other functions in particular those related to environmental adaptations evolve extremely fast [110], strongly suggesting that both the highly conserved putative and non-conserved putative CRMs are likely to be functional. The predicted CREs, Umotifs,

CRMs, average phastCons scores and putative target genes are stored in a searchable relational database (http://bioinfo.uncc.edu/mniu/pcrms/www/) for public use. The query results and relevant knowledge are displayed using the NCBI graphical sequence viewer.

1.4 Discussion

ChIP-seq techniques have been proven a powerful means to locate CREs for specific TFs genome-wide in various cell types, tissues, developmental stages and physiological conditions. However, precise identification of CREs in the binding peaks from ChIP experiments is still a challenging computational problem [85]. Efforts have been made to narrow down the binding peaks through improving experimental procedures [111], thereby facilitating the identification of CREs. On the other hand, once the binding peak summits of a TF are identified, information about the CREs of its cooperative TFs around the summits can provide a good opportunity to identify the relevant CRMs. With the accumulation of a large number of ChIP datasets in many important metazoans and plants, it is tantalizing to predict CRMs around CREs of the ChIP-ed TFs by integrating information in a large number of ChIP datasets in an organism. In this study, we have explored this idea and developed a novel algorithm DePCRM for such a purpose. The algorithm is largely based on the fact that similar TF combinatorial patterns are often repeatedly used to regulate multiple similar or different regulons in different cell types, tissues, developmental stages or physiologically conditions. As the number of possible combinations of TFs is extremely large, DePCRM identifies possible real motif combinatorial patterns in a sufficiently large number of ChIP datasets through iteratively filtering out randomly occurring spurious motifs, thereby effectively reducing the searching space in each step (Table 1.2). Clearly, in

order for the algorithm to make reasonable predictions, the ChIP datasets have to be sufficiently large and diverse, so that they are likely to include datasets for cooperative TFs in different cell types, tissues, developmental stages and physiological conditions.

Using the currently available 168 ChIP datasets for 56 TFs in D. melanogaster, the algorithm was able to recover 77.9% of the known CRMs in the datasets and even 89.3% known CRMs in which a putative CRE could be identified (Table 1.2). Thus, our algorithm has achieved rather high prediction sensitivity even only using these 168 limited datasets, in particular when a putative CRE can be located in the CRMs by a motif finding tool. Although we cannot rigorously evaluate the prediction specificity of the algorithm due to the limited knowledge of CRMs in the genome, it should not be too low for the following reasons. First, the chance for such high recovery rate of known CRMs to happen by chance is virtually impossible as indicated by our simulation studies (Figure 1.11C). Second, our predicted CRMs as well as CREs in a CRM as a whole are more conserved than the corresponding randomly selected sequences (Figures 1.12A and 1.12B). Third, the highly conserved predicted CRMs tend to be located in the close neighborhoods of genes involved in embryonic development (Supplementary file 2), which is consistent with the existing knowledge [107-109]. Fourth, the highly non-conserved predicted CRMs tend to be located in the close neighborhoods of genes involved in neural transmission, chemical sensation and metabolism, which is also in excellent agreement with the observations that gene regulatory networks for genes involved in responses to environmental factors tend to evolve very rapidly through rewiring by degrading existing CREs (death), or gaining new CREs (birth), a process called CRE turnover [110, 112]. This form of genetic changes plays a more pivotal role

in functional evolution of organisms than previously thought [110, 113]. Therefore, both the conserved and non-conserved putative CRMs are highly likely to be functional.

As vast majority of known CRMs are located in NCRs, we did not attempt to predict CRMs that are entirely located in CDRs, thus, we only allow the extended binding peaks to include at most the adjacent exon (Methods). Nevertheless, 5.9% of our predicted CRMs at least partially include the first or last exon of genes. Although putative CREs in CDRs are more likely to be conserved than those in NCRs (Figures 1.12C and 1.12D), they are not more conserved than the randomly selected short k-mers in CDRs (Figure 1.14D). Therefore, putative CREs in CDRs are not necessarily under a higher selection pressure than are the randomly selected short k-mers in CDRs. On the other hand, the other 94.1% of our predicted CRMs are entirely located in NCRs (Figure 1.3C) and consist of 34.9% of all NCRs in the genome. Interestingly, it has been shown that there are more than three times as many functional NCRs as CDRs in the D. melanogaster genome, because these NCRs are under at least the same level of natural selection as CDRs [99, 101, 105]. In other words, more than 75% of NCRs in the genome are likely to be functional. In this regarding, we have predicted less than half of possible CRMs in the genome. Furthermore, our predicted CRMs are based on 746 combinatorial patterns (i.e., CRMCs) of 184 identified Umotifs. Since TFs of the same structural family tend to recognize highly similar motifs [114, 115], our predicted Umotif might correspond to multiple highly similar motifs of different TFs of the same structural family. Hence, we may have actually predicted more than 184 motifs for some of the 1,052 annotated TFs in the genomes, and many of them are likely novel motifs. However, our predicted motifs might be far away from covering all the annotated TFs as our

predicted CRMs only cover 34.9% of NCRs. Extrapolating our results based on these datasets, we predict that 13.42% and 73.17% of NCRs in the D. melanogaster genome might code for CREs and CRMs, respectively, which is in excellent good agreement with the earlier conclusion that at least 75% of NCRs are likely to have transcriptional regulatory functions [99, 101, 105].

Nonetheless, our results demonstrate that even these limited 168 datasets for just 56 TFs can result in highly meaningful predictions of CRMs and CREs genome-wide. In other words, these datasets contain sufficient information for repeatedly used motif patterns as indicated by the significant overlaps of their binding peaks (Figures 1.4). On the other hand, because these datasets were not generated by random efforts of the community, rather, they are strongly biased to well-studied cooperative TFs, and their CRMs are relatively well documented in the literature. Therefore, if the datasets were generated by random efforts and the known CRMs were characterized by uncorrelated efforts, then we might need a much larger number of datasets to achieve the similar prediction accuracy. Moreover, as indicated above, although we have achieved a rather high recovery rate (89.3%) of known CRMs with a putative CRE, more and diverse ChIP datasets are needed to further improve the predictions, in particular to predict all CRMs in the genome. Fortunately, with ChIP-seq techniques becoming routine and the progress of the ENCODE projects, more and more ChIP-seq datasets will be churned out for numerous and even all TFs encoded in the organisms. Thus, our algorithm could be very useful for elucidating CRMs encoded any genome once a sufficient number of diverse ChIP-seq datasets become available in the organism.

Clearly, the result of our algorithm is only a static map of CREs and CRMs encoded in the genome, and for many putative CREs in the predicted CRMs, we may not know their cognate TFs and functional state (active, poised or inactive) in specific cell types, tissues, developmental stages or physiological conditions. However, once such a global CRMs map is available for an organism, it is relatively straightforward to infer the functional states to CRMs if epigenetic data in a certain cell type, tissue, developmental stage or physiological condition are available, such as ChIP-seq data for histone modification markers (e.g., mono-, bi- and tri-methylation at lysine 4 of histone 3 or H3K4m1, H3K4m2, K3K4m3, etc.) at active promoters, enhancers and silencers [18, 49, 116-119], and DNAse-seq data for nucleosome free regions [19, 21, 111, 120, 121]. Thus, future development is to incorporate the epigenetic datasets, hereby predicting the functional states of all the predicted CRMs in a certain cell type, tissue, developmental stage or physiological condition [116-119].

1.5 Methods

1.5.1 Datasets

We attempted to collect all possible ChIP-seq and ChIP-chip datasets from D. melanogaster available to us from three sources: the modENDCOE project [122], the Berkeley drosophila transcription network project (BDTNP) [79]and literature. We used the binding peak summits in each dataset, provided in the original publications as the data owners might have a better understanding of their datasets for background subtraction and normalization. We removed binding peaks that overlap with high occupancy target (HOT) regions [71, 72]We used the binding peaks in the datasets identified by the original data owners. Because the typical lengths of known CRMs are 1,000-2,000bp

[80], we extended the binding peaks shorter than 3,000bp to up to 3,000bp by padding equal length of flanking genomic sequences to the two ends. If the extension on either end reaches an adjacent exon, we only included up to the full length sequence of the exon as majority of CRMs are located in NCRs. We discarded the binding peaks longer than 5,000bp as they generally have low quality score and consist of only a small portion in the datasets (Figure 3A). The remaining extended binding peaks in each dataset were used for motif finding. The known CREs and CRMs in D. melanogaster were downloaded from the REDfly database [80].

1.5.2 Measurement of the Overlap of Binding Peaks in Two Datasets

We quantify the overlapping level of binding peaks in two datasets d_i for TF F_i and d_j for TF F_j, defined as,

$$S_o(d_i, d_j) = o_i(d_i, d_j)/|d_i| + o_j(d_i, d_j)/|d_j| \quad (1.1)$$

where $|d_i|$ and $|d_j|$ are the number of binding peaks in d_i and d_j, respectively, and $o_i(d_i, d_j)$ the number of sequences in d_i that are overlapped by a sequence in d_j.

1.5.3 Finding Motifs in Binding Peak Datasets

Based on an initial evaluation of multiple motif-finding tools for large ChIP datasets, including seeder [61], Trawler [61, 62], ChIPMunk [63], HMS [64], CMF[65], STEME [66], DREME [67], DECOD [68], RSAT [69], and POSMO [70], we selected DREME to identify all possible motifs in each of the extended binding peak dataset for its computational efficiency and capability to return enough number of over-represented motifs in a dataset [67]. As DREME requires a negative dataset for more accurate predictions, we generated a random sequence set for each input dataset using a third order Markov chain model based on the transition probabilities of the sequences in the dataset.

In addition, since it is highly unlikely that one can find a large number of high quality motifs in such a random dataset or in a low quality ChIP dataset, we also used DREME as a quality control measure to filter out low quality datasets in which no or only a single motif could be identified.

1.5.4 The Algorithm

Our DePCRM algorithm predicts CRMs through the following steps using the putative motifs as the input found in the modified binding peaks from all ChIP-seq and/or ChIP-chip datasets.

1.5.4.1 Identify Co-occurring Motif Pairs (CPs) in Each Dataset

For each pair of motifs $M_d(i)$ and $M_d(j)$ found in the same dataset d regardless of their distance, we compute a motif co-occurring score S_c defined as,

$$S_c(M_d(i), M_d(j)) = o(M_d(i), M_d(j)) / \max\{|M_d(i)|, |M_d(j)|\}, \quad (1.2)$$

where $|M_d(i)|$ and $|M_d(j)|$ are the number of binding peaks containing CREs of motifs $M_d(i)$ and $M_d(j)$, respectively; and $o(M_d(i), M_d(j))$ the number of binding peaks containing CREs of both the motifs. We select motif pairs with a $S_c \geq \alpha$ as co-occurring motif pairs (CPs) for further analysis (Figures 1.2B and 1.2C). The cutoff α is chosen such that the predicted motifs in known CRMs are minimally excluded (Figures 1.4A and 1.4B). If there are not enough known CRMs in the genome, a default $\alpha = 0.7$ is used based on the data from REDfly (see Results).

1.5.4.2 Compute Similarity Scores among All Pairs of CPs in Different Datasets

For each pair of datasets a and b, we compute a similarity score S_S between each pair of CPs $P[M_a(i), M_a(j)]$ from a and $P[M_b(m), M_b(n)]$ from b, defined as,

$$S_s\{P[M_a(i),M_a(j)],P[M_b(m),M_b(m)]\}$$
$$=\max_{k\in\{i,j\},l\in\{m,n\}}\{Sim[M_a(k),M_b(l)]\}+Sim[M_a(r),M_b(s)],\quad r\in\{i,j\},r\neq k;s\in\{m,n\},s\neq l, \quad (1.3)$$

where Sim(M, N) is the similarity score between motifs M and N using a metric called SPIC that we proposed previously considering both the frequency matrixes and position specific weight matrixes (PSWMs) of both the motifs [123-125]. We have shown that SPIC outperforms the existing metrics for measuring motif similarities [123-125]. Note that to compute S_s we first select the highest similarity among all the four possible motif pairs, and then sum it with the similarity of the remaining pair.

1.5.4.3 Construct the CP Similarity Graph

We then construct a CP similarity graph using the CPs as the nodes, and connecting two CPs with an edge with their score S_s being the weight if and only if S_s is above a cutoff β. As edges are only allowed among CPs from different datasets, thus the resulting similarity graph is a multi-partied graph (Figure 1.2C). The value of β is chosen based on the relationship between the graph density as well as the number of nodes in the graph and different β values. The graph density is defined as:

$$D=|E|/|CP|, \quad (1.4)$$

where |CP| and |E| are the numbers of CPs and edges in the graph, respectively. We choose a β value such that the resulting graph is as spars as possible and has as many nodes/CPs as possible (Figures 1.7A and 1.7B).

1.5.4.4 Cut the CP Similarity Graph into Dense Sub-graphs, CP Clusters (CPCs)

We use the Markov Chain Clustering algorithm (MCL) [87]to cut the graph into dense sub-graphs, each corresponding to a cluster of repetitively occurring CPs across multiple datasets (Figure 1.2D). MCL iteratively computes random walks determined by

a Markov chain by alternately executing two operations (expansion and inflation) on a stochastic matrix [87]. It ranks the identified dense sub-graphs according to their sizes in a descending order. It has been shown that MCL works very well in finding dense sub-graphs in very large weighted sparse graphs [87, 123, 124, 126-130]. We discard the clusters containing fewer than τ CPs (τ =2 in this study. So we only discarded singleton CPs) (Figure 1.2D). Presumably, the remaining clusters contain highly similar CPs for certain two TFs. For example, cluster C1 (P1, P5, P8) in Figure 1.2D contains highly similar motifs (red and black ova) for two distinct TFs. For this reason, we call these clusters CP clusters (CPCs) (Figure 1.2D).

1.5.4.5 Compute a Co-occurring Score for Each Pair of CPCs

Let C_i and C_j be two CPCs, and $\Omega_{d_k}(C_i, C_j)$ be the set of the CPs in C_i and C_j from the same dataset d_k. We define a co-occurring score between C_i and C_j as,

$$S_{CPC}(C_i, C_j) = \frac{1}{D}\sum_{k=1}^{D} \frac{1}{N(\Omega_{d_k}(C_i, C_j))} \sum_{(P_s \in C_i, P_t \in C_j) \subset \Omega_{d_k}(C_i, C_j)} [o(P_s, P_t)/|P_s| + o(P_s, P_t)/|P_t|], \quad (1.5)$$

where D is the number of datasets in which CPs of both C_i and C_j occur, P_s and P_t two CPs from C_i and C_j, respectively, $o(P_s,P_t)$ the number of binding peaks where P_s and P_t co-occur, $|P|$ the size of P, and $N(\Omega_{dk}(C_i,C_j))$ the number of unique comparisons among the CPs in $\Omega_{dk}(C_i, C_j)$.

1.5.4.6 Construct the CPC Co-occurring Graph

We construct a CPC co-occurring graph using each CPC as a node, and connecting two CPCs C_i and C_j by an edge with $S_{CPC}(C_i, C_j)$ being the weight if and only if $S_{CPC}(C_i, C_j) \geq = \gamma$ (Figure 1.2E). The cutoff γ is chosen based on the bimodal distribution of the S_{CPC} sores (Figure 1.7C).

1.5.4.7 Cut the CPC co-occurring Graph into Dense Subgraphs

We apply MCL to cut the CPC co-occurring graph into dense sub-graphs (Figure 1.2F). Each of these sub-graphs is assumed to correspond to a possible combination of their motifs to form a CRM based on the datasets used. For this reason, we refer to these CPC clusters as CRM components (CRMCs) (Figure 1.2E).

1.5.4.8 Combine Highly Similar Motifs in Unique Ones

Some motifs in the CRMCs may have overlapping CREs, and can be very similar to one another. It is highly likely that they consist of the same or similar CREs of the same TF or closely related ones. Thus we need to combine such highly similar and possibly redundant motifs into unique ones. To this end, we calculate the pairwise motif similarity of all the motifs in the CRMCs using the SPIC motif similarity metric [123-125]. We construct a motif similarity graph using the motifs as nodes, and connecting two nodes by an edge with the similarity being the weight if and only if the similarity of the corresponding motifs is greater than 0.7. We identify high density subgraphs in the graph using MCL. For each subgraph, we extend each CRE of each associated motif by padding 5 bp original genomic sequence at each of its two ends. We then identify the common motif in each set of the extended CREs using DREME. For the resulting motifs with more than 50% CRE overlapping and a similarity score more than 0.4, we repeat the above procedure until no two motifs meet the criteria. Each resulting motif has a similarity smaller than 0.4 and an overlapping rate lower than 0.5 with any other motifs. Thus we call each of them a unique motif or Umotif. Each motif in the identified CRMCs is then represented by its Umotif.

1.5.4.9 Predict CRMs in the Genome

We project CREs of all the CRMCs back to their locations in the genome. If the projected CREs overlap with one another, we merge them in a non-overlapping one. We then connect any two adjacent CREs if their distance is shorter than a preset value δ (δ =150bp in this study) according to the distribution of the distances between the CREs in known CRMs (Figure 1.11B) and the connection cannot span over an exon unless it contain a binding site. We predict a CRM as a segment of sequence connected by CREs of Umotifs in one or multiple CRMCs.

1.5.5 Comparison of Our Algorithm with a Naïve Algorithm

Since CRMs are likely to be enriched in our extended peaks, a naïve method that randomly selects sequences from the extended peaks can recover true CRMs. To compare our algorithm with such a naïve method, we concatenated all the genome sequences that are covered by the extended binding peaks according to the order of the sequences on the chromosomes X, Y, 2, 3 and 4, and we connected the two ends of the concatenated sequence to form a circular DNA. For each of CRM predicted by our algorithm, we randomly selected a segment of sequence with the same length as the predicted CRM from the circular DNA. We repeated the process 50 times and compared their averaged results to our predictions.

1.6 Conclusion

The exponentially increasing number of TF binding location data produced by the recent wide adaptation of chromatin immunoprecipitation coupled with microarray hybridization (ChIP-chip) or high-throughput sequencing (ChIP-seq) technologies has provided an unprecedented opportunity to identify CRMs and CREs in genomes.

However, how to effectively mine the large volumes of ChIP data to identify CREs and CRMs is a challenging task. We have developed a novel graph-theoretic based algorithm DePCRM for genome-wide de novo predictions of CRMs and CREs using a large number of ChIP datasets. DePCRM predicts CRMs by identifying overrepresented combinatorial motif patterns in multiple ChIP datasets in an effective way. When applied to 168 ChIP datasets of 56 TFs from *D. Melanogaster*, DePCRM identified 184 and 746 overrepresented motifs and their combinatorial patterns, respectively, and predicted a total of 115,932 CRMs in the genome. The predictions recover 77.9% of known CRMs in the datasets, 89.3% of known CRMs containing at least one predicted CRE. These putative CRMs and CREs as a whole in a CRM are more conserved than randomly selected sequences, thus, they are highly likely to be functional. Thus, the algorithm can be used to predict CRMs and CREs in other eukaryotic genomes from which a sufficient number of diverse ChIP datasets are available. All the predicted CREs, motifs, CRMs, and their target genes are available at http://bioinfo.uncc.edu/mniu/pcrms/www/.

CHAPTER 2: PREDICTION OF CIS-REGULATORY MODULES IN THE HUMAN GENOME

2.1. Abstract

It has been shown that about 70% of the conserved human genome are NCRs, suggesting that they might be involved in gene transcriptional regulation. However, due to the computational and experimental difficulties, we are still far from a comprehensive understanding of the human regulatory genome. In this chapter, we applied DePCRM to 359 human ChIP-seq datasets for 148 TFs in 68 different cell or tissue types, and identified 636 overrepresented motifs, 1,991 CRMCs, and 807,365 CRMs in the genome. The predictions recovered 95.55% of known enhances in the datasets, 48.84% of trait-linked SNPs from dbGAP. Furthermore, 50.96% of our predicted CRMs overlaps with DNase I hypersensitive sites (DHSs). Thus, our predictions reached a rather high sensitivity. We also found that our predicted CREs and CRMs tend to be more conserved than randomly selected sequences, suggesting they are more likely to be functional. Furthermore, we analyzed the saturation trends of predicted CRMCs and motifs using increasing number of datasets in three scenarios. We found that with a practical number of datasets, a complete list of CRMCs and Umotifs are reachable in specific cell types, for specific TFs (with its co-working TFs) and in the whole genome. We predicted the proportion of the complete list of Umotifs and CRMCs one can reach with the number of datasets available.

2.2 Introduction

The human genome encodes around 21,000 protein-coding genes that consist of only 1.5% of the whole genome. Comparative genomic studies suggest that 5% of the human genome are conserved [131], implying that 70% of conserved NCRs might be involved in gene transcriptional regulation. Furthermore, the human proteome is different from the chimpanzee proteome by only 1.23% [132], while their noncoding regulatory sequences differ much more [133]. Therefore, the observable differences between chimpanzee and humans may be due more to differences in the regulation of gene expression than to differences in protein-coding genes. Moreover, current genome-wide association studies suggest that noncoding genetic variation among individuals plays a major role in the variation in human phenotypes and disease susceptibility [88], therefore, a better understanding of CRMs in the human genome will lead to an understanding of not only human gene regulation and evolution, but also various diseases. Enormous efforts have been made to identify CREs and CRMs in the human genome, through large consortiums such as the ENCODE project, as well as studies by individuals around the world using a variety of NGS-based technologies.

However, compared with the *D. melanogaster* genome (139.5Mbp), the human genome (3.2 Bbp), is 22.9 times larger, and encodes more genes (21,000 vs 13,600) and more TFs (2,886 vs 1,030), indicting more complex combinatorial usages of TFs, and thus CREs and CRMs in the human genome. Furthermore, there are far more cell and tissue types in human that need to be explored to reveal all the encoded CREs and CRMs than in D. melanogaster. Thus, the characterization of CREs and CRMs in the human genome can be a more challenging task than in the *D. melanogaster* genome. For

instance, although, the ever increasing number of TF ChIP-seq datasets hold promise for the characterization of CREs and CRMs in the human genome, ChIP-seq datasets obtained from human tissues or cells are on average 2-3 times larger than from *D. melanogaster* tissues or cells, and may contain far more number of motif combinatory patterns, making their analysis and integration more challenging. Can the DePCRM algorithm we developed earlier work on much bigger human ChIP-seq datasets, and predict the CREs and CRMs in the genome with high accuracy? Moreover, given the great efforts that have been made word-wide to generate a large number of ChIP-seq datasets from various human tissues and cell types, , what is the status of these efforts, how far will we need to go, and how should we proceed to achieve the goals faster and more cost-effectively?

To address these questions, we first speeded up the DePCRM algorithm by splitting datasets with a large number for binding peaks into multiple smaller ones for motif-finding, as motif-finding is the rate-limiting step of the DePCRM algorithm, in particular for very large datasets. We then applied the algorithm to a total of 359 ChIP-seq datasets for 148 TFs in 68 different types. We identified 636, 1,991 and 807,365 Umotif, CRMCs and CRMs in the genome, respectively. The predictions recovered 95.55% of known enhances in the datasets, 48.84% of trait-linked SNPs from dbGAP. Furthermore, 47.11% of our predicted CRMs overlaps with DNase I hypersensitive sites (DHSs). We also found that the putative CRMs and CREs as a whole in a CRM are more conserved than randomly selected sequences. Thus, they are likely to be functional. Furthermore, using these datasets, we analyzed the saturation trend of CRM predictions in three different scenarios: 1) How does the number of predicted CRMs change using an

increasing number of datasets for different TFs from the same type of cells? 2) How does the number of predicted CRMs change using an increasing number of datasets for the same TF in different cell types? And 3) how does the number of predicted CRMs change using an increasing number randomly selected datasets?

2.3 Materials and Methods

2.3.1 Datasets

A total of 359 ChIP-seq datasets for 148 TFs in 68 different cell or tissue types were downloaded from the Encyclopedia Of DNA Elements (ENCODE) consortium website (http://ftp.ebi.ac.uk/pub/databases/ensembl/encode/integration_data_jan2011/byDataType/peaks/jan2011/peakSeq/optimal/hub/). The binding peaks were identified by the peak-calling and refinery procedure designed by Kundaje and colleague [134]. A total of 850 experimentally verified the sequences containing the enhancers in the human genome (version hg19) were downloaded from the Vista Ehancer Browser database [135]. These human enhancer fragments have an average length of 1,925 bps. The coordinates of a total of 30,572 trait-linked SNPs were downloaded from the database of Genotypes and Phenotypes (dbGaP) [136]. Majority of these SNPs were identified by genome-wide association studies (GWAS), medical sequencing, molecular diagnostic assays, as well as association between genotype and non-clinical traits. Additionally, the coordinates of a total of 1,281,988 non-overlapping DHSs in 125 cell types produced by ENCODE were downloaded from the UCSC Genome Browser database (http://genome.ucsc.edu/cgi-bin/hgTrackUi?db=hg19&g= wgEncodeRegDnaseClusteredV2).

2.3.2 Measurement of the Overlap of Binding Peaks in Two Datasets

We extended the binding peaks shorter than 3kbp to up to 3kbp by padding equal length of flanking genomic sequences to the two ends. If the extension of either side of the peak reaches an exon, we then only include the first exon since majority of the CREs locate within NCR. We quantify the overlapping level of extended binding peaks in two datasets d_i for TF F_i and d_j for TF F_j, defined as,

$$S_o(d_i, d_j) = o_i(d_i, d_j)/|d_i| + o_j(d_i, d_j)/|d_j|, \quad (2.1)$$

where $|d_i|$ and $|d_j|$ are the number of binding peaks in d_i and d_j, respectively, and $o_i(d_i, d_j)$ the number of sequences in d_i that are overlapped by a sequence in d_j.

2.3.3 Finding Motifs in Binding Peak Datasets

We used DREME to identify all possible motifs in each of the extended binding peak dataset for its computational efficiency and capability to return enough number of over-represented motifs in a dataset [67]. As DREME requires a negative dataset for more accurate predictions, we generated a random sequence set for each input dataset using a third order Markov chain model based on the transition probabilities of the sequences in the dataset. As the size of a dataset becomes large, even a fast algorithm such as DREME cannot run in a practical time, thus we split a dataset with a size over 10,000 peaks into multiple sub-datasets with similar number of peaks smaller than 10,000, i.e., the size sub-datasets is equal to s/ (mod(s/10,000) +1), where s > 10,000 is the size of the original dataset.

2.3.4 The Algorithm

The motifs found in each dataset were fed to the DePCRM detailed in Chapter 1, to predict CRE and CRMs in the genome. Briefly, for each pair of motifs $M_d(i)$ and $M_d(j)$,

found in the same dataset d, regardless of their distance, we compute a motif co-occurring score Sc. We select motif pairs with a Sc ≥ α as co-occurring motif pairs (CPs) for further analysis. The cutoff α is chosen such that the predicted motifs in known CRMs are minimally excluded (Figures 2.1). If there are not enough known CRMs in the genome, a default α=0.6 is used based on the data from VISTA. For each pair of datasets a and b, we compute a similarity score S_s between each pair of CPs $P[M_a(i), M_a(j)]$ from a and $P[M_b(m), M_b(n)]$ from b. We then construct a CP similarity graph using the CPs as the nodes, and connecting two CPs with an edge with their score S_s being the weight if and only if S_s is above a cutoff β. We choose a β value such that the resulting graph is as sparse as possible and has as many nodes/CPs as possible. We use the Markov Chain Clustering algorithm (MCL)[137] to cut the graph into dense sub-graphs, each corresponding to a cluster of repetitively occurring CPs across multiple datasets. We discard the clusters containing fewer than τ CPs (τ=2 in this study. Thus we only discarded singleton CPs). Presumably, the remaining clusters contain highly similar CPs for certain two TFs. We call these clusters CP clusters (CPCs). For each pair of CPCs, C_i and C_j, we calculate a co-occurring score S_{cpc} for their concurrence over all the datasets, and construct a CPC co-occurring graph using each CPC as a node, and connecting two CPCs C_i and C_j by an edge with being the weight if and only if $S_{CPC}(C_i, C_j) \geq = \gamma$. The cutoff γ is chosen based on the bimodal distribution of the S_{cpc} sores. We apply MCL to cut the CPC co-occurring graph into dense sub-graphs. Each of these sub-graphs is assumed to correspond to a possible combination of their motifs to form a CRM based on the datasets used. For this reason, we refer to these CPC clusters as CRM components (CRMCs). Some motifs in

the CRMCs may have overlapping CREs, and can be very similar to one another. It is highly likely that they consist of the same or similar CREs of the same TF or closely related ones. Thus, we combine such highly similar and possibly redundant motifs into unique ones. We call each of them a unique motif or Umotif. Each motif in the identified CRMCs is then represented by its Umotif. We project CREs of all the CRMCs back to their locations in the genome. If the projected CREs overlap with one another, we merge them in a non-overlapping one. We then connect any two adjacent CREs if their distance is shorter than a preset value δ (δ=150bp in this study) according to the distribution of the distances between the CREs in known CRMs and the connection cannot span over an exon unless it contain a binding site. We predict as a CRM each segment of the sequence connected by CREs of Umotifs in one or multiple CRMCs.

2.3.5 Prediction Saturation Analysis

We analyzed the saturation trends of predicted Umotifs and CRMCs in the following three scenarios: 1) changes in the number of Umotifs and CRMCs with increasing number of datasets for different TFs from the same cell type or tissue; 2) changes in the number of Umotifs and CRMCs with increasing number of datasets in different cell types or tissues for the same TF; and 3) changes in the number of Umotifs and CRMCs with increasing number of randomly selected datasets. Specifically, for the first two scenarios, we used the Umotifs and CRMCs predicted using the 359 dataset as the standard sets, and count the number of the Umotifs and CRMCs predicted using the selected datasets, which are recovered by the standard sets. For the third scenario, we randomly selected different numbers (n=100, 200, 250 and 300) of datasets from the 359 dataset, and applied the algorithm to each of the randomly selected datasets with the same

parameter setting. For all the three scenarios, we repeated the process 50 times for each dataset size, and present the averaged results to minimize the effect caused by the combination and the order of datasets used. We fitted the results to a sigmoid function,

$$f(n) = \delta + \frac{\alpha - \delta}{1 + e^{[\beta \times log(n/\gamma)]}}, \quad (2.2)$$

where α, β γ and δ are constant, and n is the number of datasets used for the predictions.

2.4 Results

2.4.1 Overlap of the Extended Peaks.

As shown in Figure 2.1A, each of the 359 datasets contains 19~58,505 binding peaks, and vast majority (99.5%) of the peaks have a length shorter than 5,000bp (Figure 2.1B). We extended the binding peaks shorter than 3,000bp to up to 3,000bp. The reason we chose 3,000bs as the extension length is that most of the known enhancer segments from VISTA are shorter than 3,000bs.

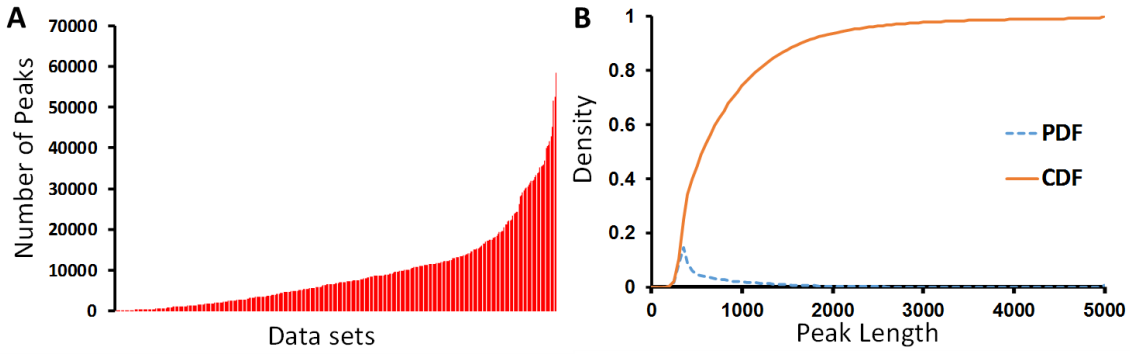

Figure 2.1. **A**. The number of binding peaks in the 359 original datasets sorted by their sizes in the ascending order. **B**. Distribution of binding peak lengths in the 359 datasets, vast of them are shorter than 5,000bs.

The 359 datasets contain a total of 3,423,090 sequences with 131 contain more than 10,000 sequences, after the length extension they contain a total of

10,269,627,767bp, which are 3.27 times of the genome (3,137,161,264bs), but only cover 34.8% (1,091,718,950bs) of the genome (Table 2.1 and Figure 2.2), indicating that some of these sequences highly overlap with one another. Of the 1,091,718,950bs genome sequence covered by the datasets, 1,048,072,608bs (96%) are in non-coding regions (NCRs, including introns and intergenic sequences), consisting of 34.3% of NCRs (3,059,588,382bs) in the genome (Table 2.1 and Figure 2.2). The remaining 43,646,342 (4%) sequences are in coding regions (CDRs), consisting of 56.3% of CDRs (77,572,882bs) in the genome (Table 2.1 and Figure 2.2). Thus, we have included a considerable portion of CDRs in the datasets because some binding peaks are located in CDRs.

Table 2.1. Summary of the coverage of the datasets, predicted CRMs and CREs on the CDRs and NCRs of the genome.

Categories	Size (bs)	% of genome	CDRs			NCRs		
			Size(bs)	% of category	% of genome	Size (bs)	% of category	% of genome
Genome	3,137,161,264	100.0	77,572,882	2.5	100.0	3,059,588,382	97.5	100.0
Covered by dataset	1,091,718,950	34.8	43,646,342	4.0	56.3	1,048,072,608	96.0	34.3
Covered by putative CRMs	216,940,383	6.9	4,017,661	1.9	5.2	212,922,722	98.1	7.0
Covered by putative CREs	47,268,468	1.5	489,386	1.0	0.6	46,779,082	99.0	1.5

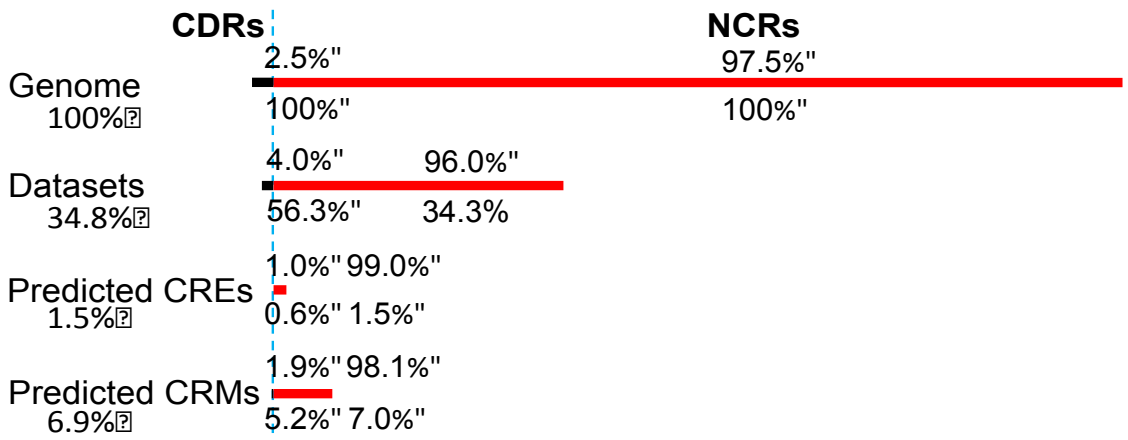

Figure 2.2. Coverage of the datasets, predicted CREs and CRMs on the CDRs and NCRs in the genome. The numbers above the lines are the proportions of the CDRs and NCRs in the corresponding sequence categories; the numbers below the lines are the proportions of CDRs and NCRs with respect to the entire CDRs and NCRs in the genome, respectively.

To see the patterns of the overlapping in the datasets, we computed the pair-wise overlapping score among the 359 datasets using the formula 1.1, and clustered the datasets based on the score. As shown in Figure 2.3A, there are numerous clear clusters formed by datasets of TFs. Interestingly, many TFs whose datasets form cluster are known to work cooperatively in regulating genes. For example, the cluster of datasets highlighted in Figure 2.3B involves TFs RAD21, CTCF and SMC3. RAD21 and SMC3 are the members of the cohesin complex, and it has been reported that cohesin co-localizes with CTCF at more than 80% of sites genome-wide [138]. Another example is the cluster formed by the datasets of TFs ZNF274, KAP1 and SETBD1 (Figure 2.3C). It has been shown that knockdown of ZNF274 with siRNAs reduced the levels of KAP1 and SETDB1 recruitment to the

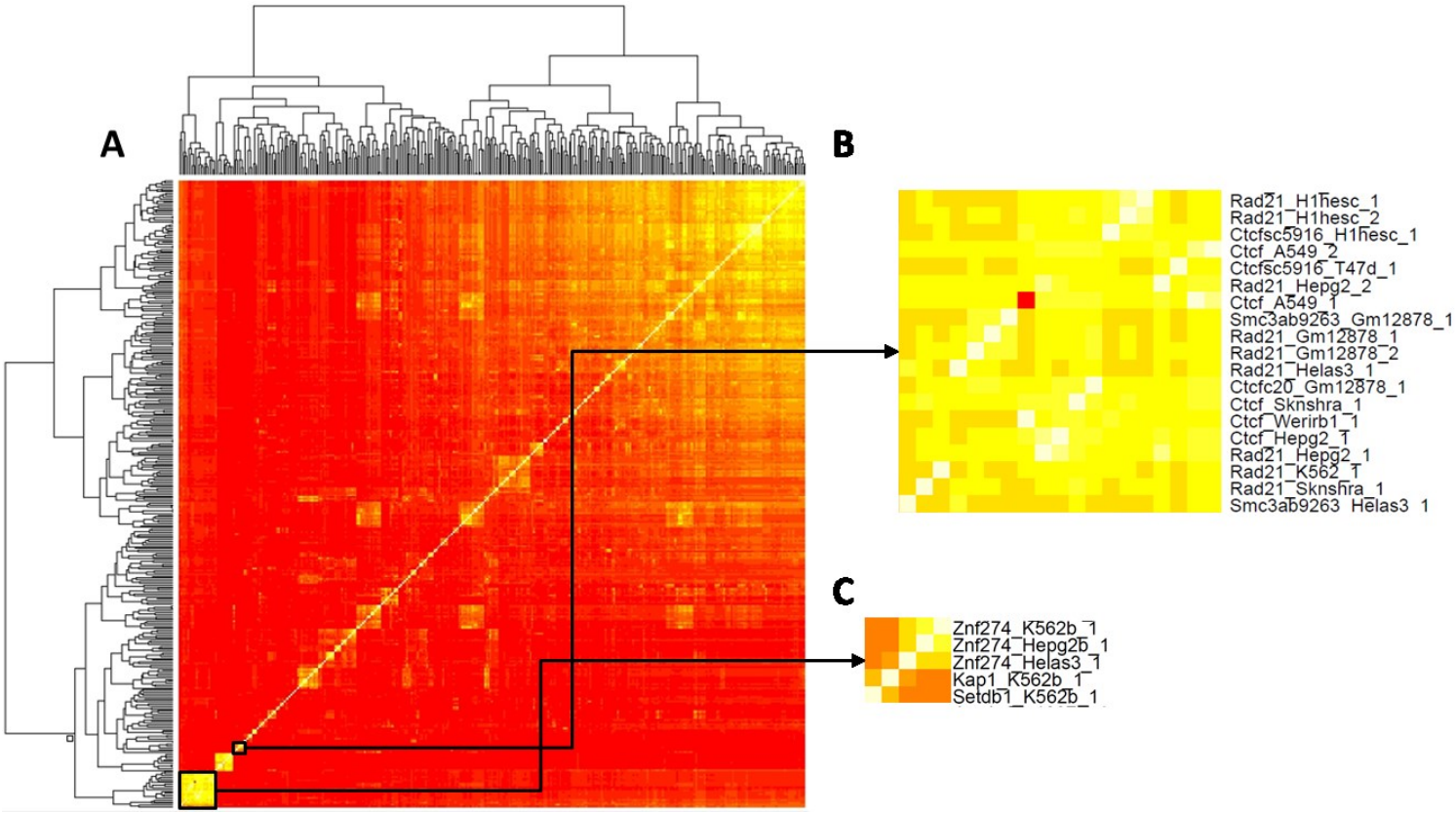

Figure 2.3 **A.** Hierarchical clustering of the 168 datasets for 56 TFs based on their pair-wise binding peak overlapping scores S_o. **B.** The blow-up of a cluster of datasets in which target TFs are RAD21, CTCF, SMC3 and ZNF143. **C.** The blow-up of a cluster of datasets in which target TFs are ZNF274, Kap1 and Setdb1.

ZNF274 binding regions, suggesting that ZNF274 is involved in the recruitment of the KAP1 and SETDB1 to specific regions of the human genome [139]. Therefore, these results indicate that the datasets might contain sufficient information to predict at least a portion of CREs and CRMs in the genome.

2.4.2 Identification of the Motifs

Our goal is to find all possible TF binding motifs of the ChIP-ed TFs and their cooperative TFs in each dataset. To facilitate motif finding in the 131 large datasets containing more than 10,000 binding peaks, we split them into a total 342 dataset, thus we ended up with a total of 570 datasets each contain fewer than 10,000 binding peaks. As shown in Figure 2.4, we found 0~339 motif in each of the datasets depending on the quality and size of the dataset (Figure 2.4A). On the one hand, in the dataset with 19 peaks, DREME was not able to find any motif, so the dataset is filtered out at this step. Thus, the motif-finding step serves as a quality control step that involve minimum human input. On the other hand, the vast majority (3,423,064) of the 3,423,210 binding peaks contain members of the identified motifs, indicating that the peaks are enriched with motifs. To see the effects of splitting a large dataset in smaller ones on the motif finding results, we randomly split three datasets with 22,314, 30,924, and 40,670 peaks in three, four and five sub-datasets, respectively, so each sub-dataset contains fewer than 10,000 peaks, and find motifs in each of the sub-dataset. We repeated this process by 10 times. As shown in Figure 2.4C, the number of motifs identified for each splitting are quite similar, and are also similar to the number of motifs identified by the way of splitting used in the algorithm. Therefore, the way to split a large dataset does not significantly affect at least the number of motifs identified in the split sub-datasets and thus the

dataset. The returned motifs generally have high information content (Figure 2.4B). Overall, we identified 130,812 putative motifs corresponding to 465,045,660 putative CREs containing 2,650,760,262bs, which is 2.42 times of the genome covered by our datasets, indicating that many motifs were identified from overlapping sequences in different datasets as DREME does not return overlapping motifs from in same datasets.

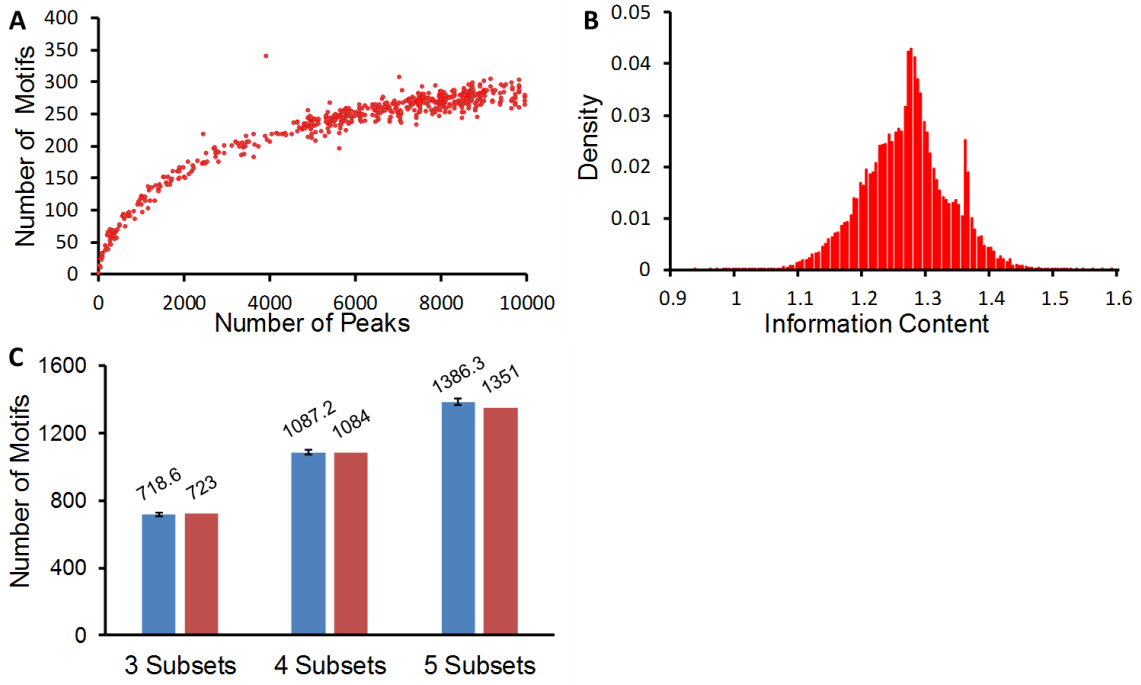

Figure 2.4 **A.** Number of motifs found as a function of the size of the 570 datasets. **B.** Distribution of the motif information content. **C.** The average (n=10) total numbers of motifs found in the randomly split three, four and five sub-datasets of the corresponding large datasets (Blue), and the total number of motifs found by the way of splitting used in this study.

2.4.3 Prediction of CRM Clusters by Mining the Combinatorial Patterns of Motifs

Both spurious and true motifs are included in the 130,821 putative motifs returned by DREME clearly. However, it is necessary to return this relatively large number of the putative motifs in order to include the majority, if not all, of the true motifs in the dataset. Our algorithm DePCRM is designed to filter out spurious motifs as many as possible

while keeping the true ones. Thus, we use these 130,821 motifs as the input of the algorithm. DePCRM first identifies highly co-occurring motif pairs (CPs) in each dataset by computing a co-occurring score (S_c) for each pair of putative motifs found in each dataset using formula 1.2. As shown in Figure 2.5A, the distribution of S_c is strongly skewed toward right, indicating that the low-scoring Gaussian-like component is likely due to the motif pairs that occurred by chance, thus are spurious. To find a proper cutoff for S_c such that most of motif-pairs occurred by chance are filtered out while most of the true motifs and motif pairs are kept, we plot the motif number and VISTA enhancer coverage as a function of the S_c cutoff α. As shown in Figure 2.5B, when α=0.6, 123,233 (94.2%) of the 130821 input motifs are filtered out; and 16,142,379 (99.9%) of the 16,161,265 CPs are filtered out; meanwhile only 19 enhancer segments from VISTA which overlaps with our predicted CREs are lost. Thus, we chose S_c>α=0.6 as the cutoff. This results in 18,886 (18,886/16,161,265=0.12%) CPs containing 7,603 (7,603/130,821=5.8%) motifs for further analysis.

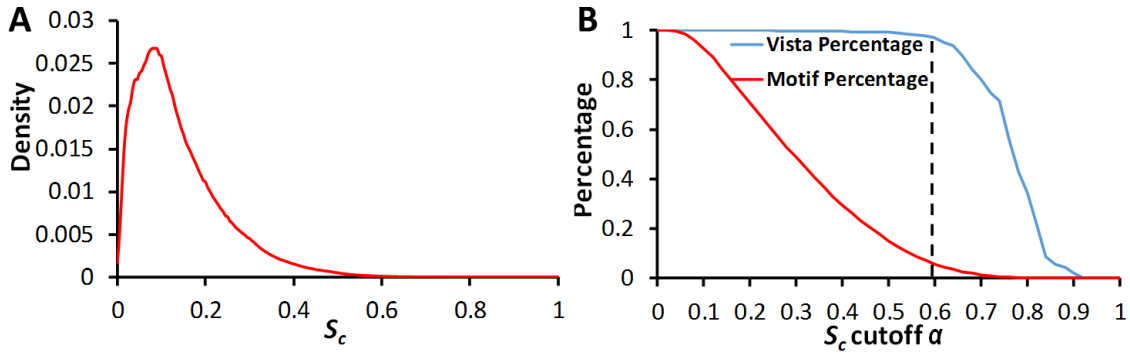

Figure 2.5. **A.** Distribution of S_c scores. **B.** The remaining proportion (red line) of the motifs found in the datasets as a function of S_c score cutoff. Recovery rate of the VISTA enhancer segments (blue line) as a function of S_c score cutoff.

To further enrich true motif pairs and motifs, the algorithm identifies repeatedly used CPs by clustering highly similar CPs in different datasets. Thus, for each pair of CPs from different datasets, we calculate the S_s score using formula 1.3. Then we construct a CP similarity graph using the CPs as the nodes, and the S_s scores as the weights on the edges (Methods). As shown in Figure 2.6A, with the increase in β, the density of the graph drops rapidly, but the dropping starts slowing down around β =1.38; meanwhile the number of nodes (CPs) in the graph starts decreasing rapidly around β =1.368 (Figure 7B). Thus, we set β =1.38 to construct the CP similarity graph (Methods). Applying the Markov chain clustering (MCL) algorithm [137] to the graph resulted in 2,444 CP clusters (CPCs) containing 7,851 (7,851/18,886=41.6%) CPs and 5,665 (5,665/7,603 =74.5%) motifs.

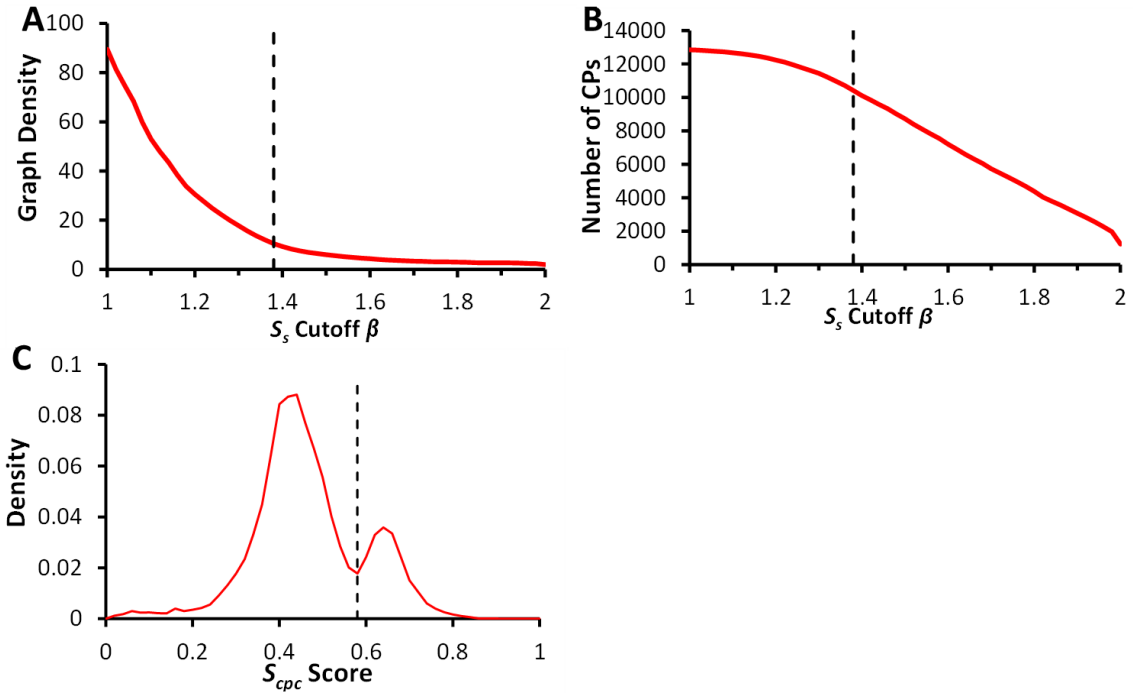

Figure 2.6. **A**. The density of the CP similarity graph drops rapidly with the increase in the S_s cutoff β, but the trend of decrease slows down around β =1.38. **B**. The number of CRM in the graph also starts to drop rapidly around β=1.38. Thus, we set β = 1.38 for

construing the final CP similarity graph. **C**. The distribution of CPC co-occurring scores S_{CPC} is well separated into a low-scoring component and a high-scoring component. The vertical line indicates the S_{CPC} cutoff $\gamma = 0.58$ at the deepest valley between the two peaks, for constructing the CPC co-occurring graph.

In order to identify larger combinatorial motif patterns, we calculate S_{cpc} for each pair of CPCs using formula 1.5. As shown in Figure 2.6, the distribution of S_{cpc} displays a well-separated bimodal distribution, and the low-scoring peak is likely mainly due to random motif patterns, while the high-scoring one is more likely attributable to truly co-acting motifs, thus we considered CPC pairs with an $S_{cpc} \geq \gamma = 0.58$ (at the valley between the two peaks) for further analysis. This result is consistent with our CRM prediction in *D. melanogaster* genome. We apply the MCL algorithm to the resulting CPC co-occurring graph, and found a total of 2,032 CRMCs (Figure 2.7). Overall, 125,156 (95.7%) motifs are filtered out from the original 130,821 input motifs by the algorithm, suggesting that majority of the motifs found in the dataset are spurious predictions.

Figure 2.7. Structures of the 2,032 CRMCs. Each node in graphs is a CPC, and each connected graph represents a CRMC.

Given the fact that extensive overlaps exist among datasets and motifs found in them, the resulting 5,665 motifs may contain duplicates. The duplicated motifs are likely to be recognized by the same TF or closely related ones. Thus, they need to be combined into non-redundant and unique ones. We iteratively clustered the final 5,665 motifs based on their similarities (Methods), resulting in 636 clusters (Figure 2.8). We consider each cluster as a unique motif and refer to it as a Umotif, each containing 1~259 highly similar

motifs and 8~394,519 CREs (Supplementary file 4 Table S6). When compared with the known motifs from Jolma et. al [115] and JASPAR CORE vertebrate [140], 317 (49.8%) Umotifs are highly similar to known motifs in Human at $p<0.001$, suggesting that they are highly likely to be true motifs. For example, as shown in Figure 2.9, Umotif 13 and Umotif 14 are very similar to the binding profile of MA0162.2 and MA0131.1, respectively, and the member of both Umotifs are very similar to one another,. We replaced the motifs in the CRMCs with the Umotifs that they belonged to, and each of the CRMCs is represented by their constituent Umotifs. Some CRMCs contain the same combination of Umotifs. Thus, we merged them in a unique one, resulting in 1,991 CRMCs.

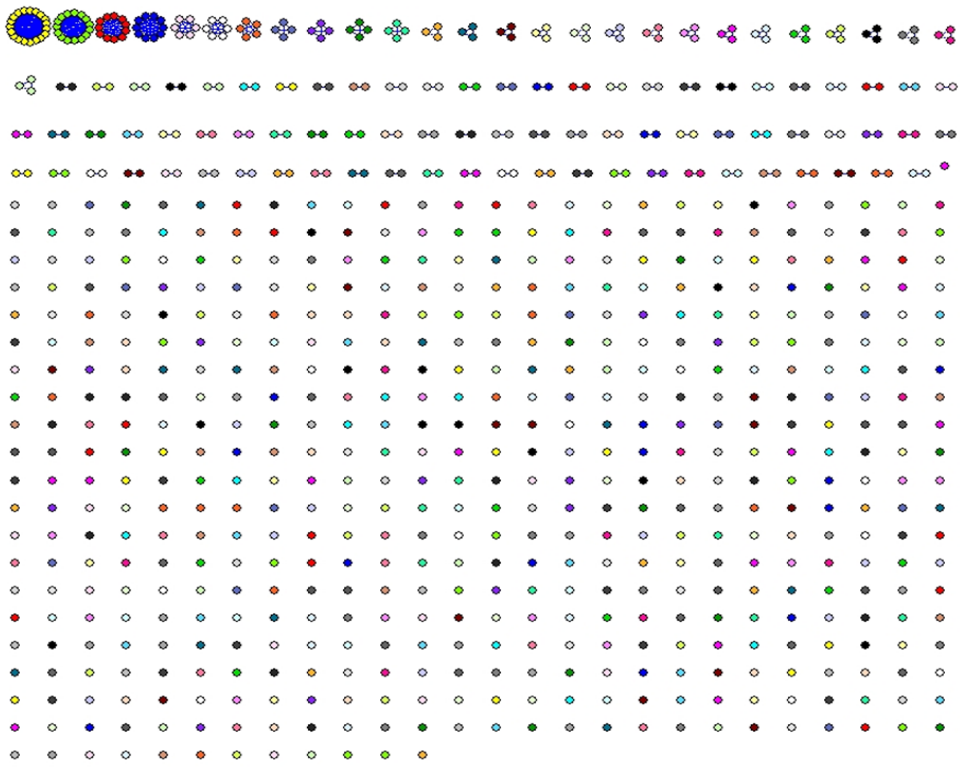

Figure 2.8. Structures of the 636 Umotifs and their member motifs. Each node in graphs is a member motif, and each connected graph represents a Umotif.

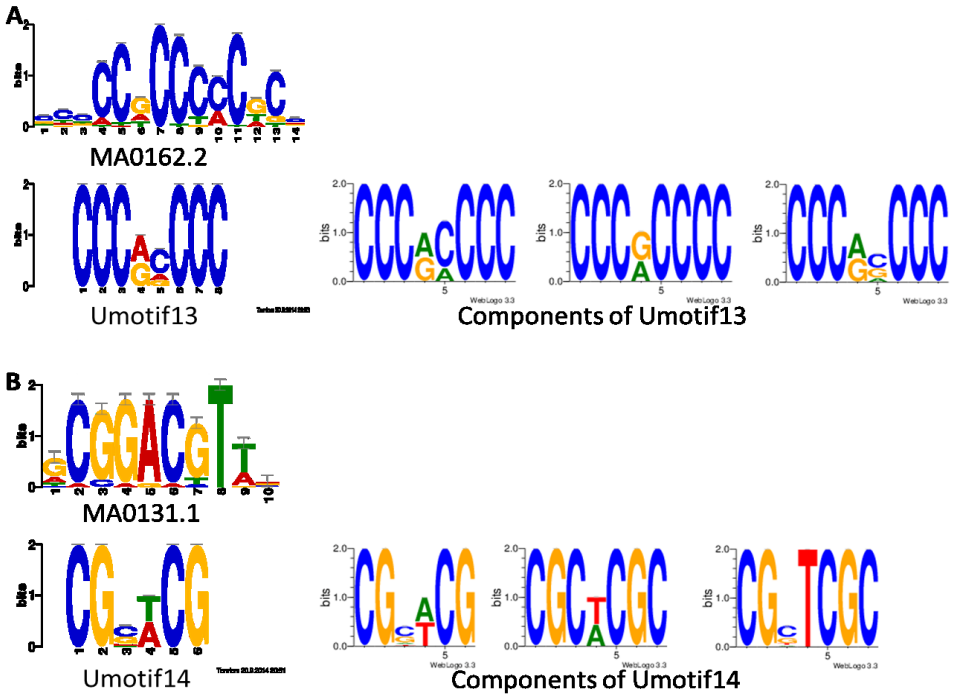

Figure 2.9. **A.** Umotif 13 and its known motif hit MA0162.1, and its three member motifs. **B.** Umotif 14 and its known motif hit MA0131.1, and its three member motifs.

2.4.4 Genome-Wide Prediction of CREs and CRMs in the Human Genome

Projecting the CREs in these 1,991 CRMCs back to the human genome (Methods) resulted in a total of 5,186,520 non-overlapping CREs, with 78,640 (1.5%) of which being entirely located in CDRs. These 5,186,520 CREs cover 47,268,468bp (5.4%) genome sequence, of which 46,779,082bp (94.9%) are in NCRs, consisting of 1.5% of NCRs; the remaining 489,386bp (1.0%) are in CDRs, consisting of 0.6% of CDRs (Figure 2.2 and Table 2.1). By connecting these putative CREs (Methods), we predicted a total of 807,365 non-overlapping CRMs, 726,954 (90.0%) of which are entirely located in NCRs, and the remaining 80,411 (10.0%) contain CDRs. These 807,365 CRMs cover 216,940,383bp (6.9%) of genome sequence, 212,922,722bp (98.1%) of which are in NCRs, consisting 7.0% of NCRs; the remaining 4,017,661bp (1.9%) are in CDRs,

consisting of 5.2% of CDRs (Figure 2.2 and Table 2.1). These putative CRMs tend to have shorter lengths than those of the known CRMs (Figure 2.10A). Furthermore, the putative CRMs harbor 2 to 2,199 CREs with a median of 6, only a small portion of the putative CRMs tends to have a short distance between adjacent two putative CREs (Figure 2.10B). These results suggest that we might have missed certain CREs in the predicted CRMs, particularly at the two ends, presumably due to insufficient information in the limited number of available ChIP datasets used in this study. In other words, some of our predictions might consist of only a part of real CRMs with possible missing CREs at the two ends of the CRM. Clearly, in order to make more accurate and complete predictions, more and highly diverse ChIP datasets are needed.

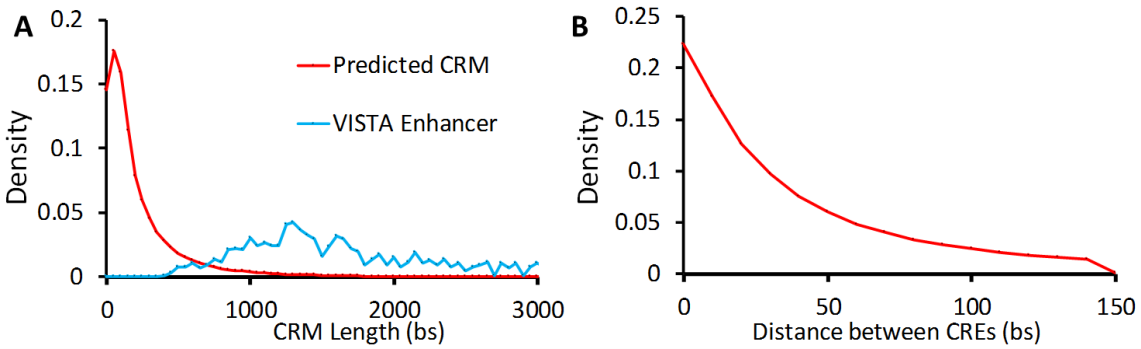

Figure 2.10. **A**. Distribution of the lengths of the known and predicted CRMs. **B**. Distribution of the distances (bs) between two adjacent CREs.

2.4.5 Predicted CRMs and CREs Are More Conserve Than Randomly Selected Sequences.

It is widely recognized that functional elements such as protein-coding exons, non-coding RNAs, and regulatory sequences usually are under more selective constrains, thus are more likely to be conserved than non-functional sequences that are selectively neutral. Therefore, we first evaluate our predicted CRMs and CRE by comparing the

conversation levels of the nucleotides in the putative CRMs and CREs with those of the same number and length sequences randomly selected from relevant background sequences. We quantified the conservation levels of each nucleotide using the GERP++ scores [141]. GERP++ (gerpcol) estimates the substitution rate on each position of multiple alignments of genomic sequences from human and 33 other mammalian species, and provides a rejected substitution (RS) score for the nucleotides at the position relative to selectively neutral sequences. Thus a positive RS score indicates that the position has been under purifying selection, thus is conserved; a negative RS score means that the position has been under positive selection; and a RS score around 0 suggest that the position is selectively neutral or nearly so. As shown in Figure 2.11A, although the average conservation score of both putative CRMs and randomly selected sequences from NCR (repeated by 50 times) have a bell-shape distribution, they are significantly different ($p<2.2x10-302$, Kolmogorov–Smirnov test). Specifically, the randomly selected NCR sequences have a left-skewed and much narrower peak distribution centered at 0 than do the putative CRMs (Figure 2.11A), indicating the randomly selected NCR sequences are largely selectively neutral or not conserved. By contrast, the predicted CRMs have a lightly right-skewed and much broader distribution (Figure 2.11A) than do the randomly selected NCR sequences, indicating that our predicted CRMs are more likely to be under purifying selection and thus conserved. As the spacing sequences between CREs in a CRM may not necessarily be functional and conserved, we next compared average RS scores of putative CREs in each of the 726,954 predicted CRMs in NCRs with those of the same number and length sequences randomly selected from NCRs. As expected, although the distribution of average RS scores for the randomly

selected sequences for CREs in a CRM (Figure 2.11B) are very similar to that for the randomly selected sequences for CRMs (Figure 2.11A), the distribution for CREs in a CRM (Figure 2.11B) is much more right-skewed than that for CRMs (Figure 2.11A). Consequently, the difference between the two new distributions (Figure 2.11B) is more contrast than between the two earlier distributions (Figure 2.11A), in particular in the high RS score (> 0.2) range. Specifically, in Figure 2.11A, 17.09% of the predicted CRMs have a RS score greater than 0.2 comparing to 9.73% of the randomly selected sequences, the difference is 17.9%-9.73%=8.17%. In figure 2.11B, 23.87% of the predicted CRMs have a RS score greater than 0.2 comparing to 13.57% of the randomly selected sequences, the difference is 23.87%-13.57%=10.3%. Hence, putative CREs in a CRM as a whole are even much more conserved than the randomly selected NCRs, and also more conserved than spacer sequences in the putative CRMs.

We further compared average RS scores of nucleotides in single CREs in the 5,083,613 predicted CRMs in NCRs and in the 78,640 predicted CRMs in CDRs with those of the same number and length sequences randomly selected from NCRs and CDRs, respectively. Interestingly, as shown in Figure 2.11C, the predicted individual CREs in NCRs are either more likely to be highly conserved with a RS score greater than 0.2 (35.28%), or more likely to be moderately positively selected with a RS score within [-0.5, -2] (30.8%) compared with the randomly selected sequences. Therefore, the putative individual CREs in the NCRs are likely to be authentic as functional elements can either under negative or positive selections. Intriguingly, as shown in Figure 2.11D, even though both individual CREs in CDRs and randomly selected CDR sequences display a bimodal distribution with one peaking at the RS score around 0 and the other

peaking at a large positive RS value, they shows significant differences. First, the peak around RS=0 for CREs in CDRs are very small, indicating that CREs in CDRs are largely either negatively or positively selected, very few are selectively neutral. By contrast, a large peak around RS=0 for randomly selected CDR sequences, indicate a considerate portion of randomly selected CDRs is selectively neutral. Second, the high-scoring peak for randomly selected CDRs is more right-skewed than that for CREs in CDRs, while the low-scoring peak for CREs in CDRs is more left-skewed than that for randomly selected CDRs. These results suggest that randomly selected CDRs are more likely to be strongly negatively selected than CREs in CDRs, and that CREs in CDRs are more likely to be positively selected. These findings might help to solve the long-standing question whether or not CREs in CDRs have dual functions: serving both as coding sequences and regulatory sequences. The lower RS scores of the some CREs in CDRs than those of randomly selected CDRs might indicate that these CREs is under less coding constrain, allowing them to be versatile enough to gain the regulatory functions while still possibly coding for a spacer in the protein sequences. On the other hand, individual CREs in CDRs are more conserved than those in NCRs (Figures 2.11C and 2.11D), strongly suggesting that CREs in CDRs are under more evolutionary constraints that the CREs in NCRs, thus they may also have coding functions in addition to regulatory functions.

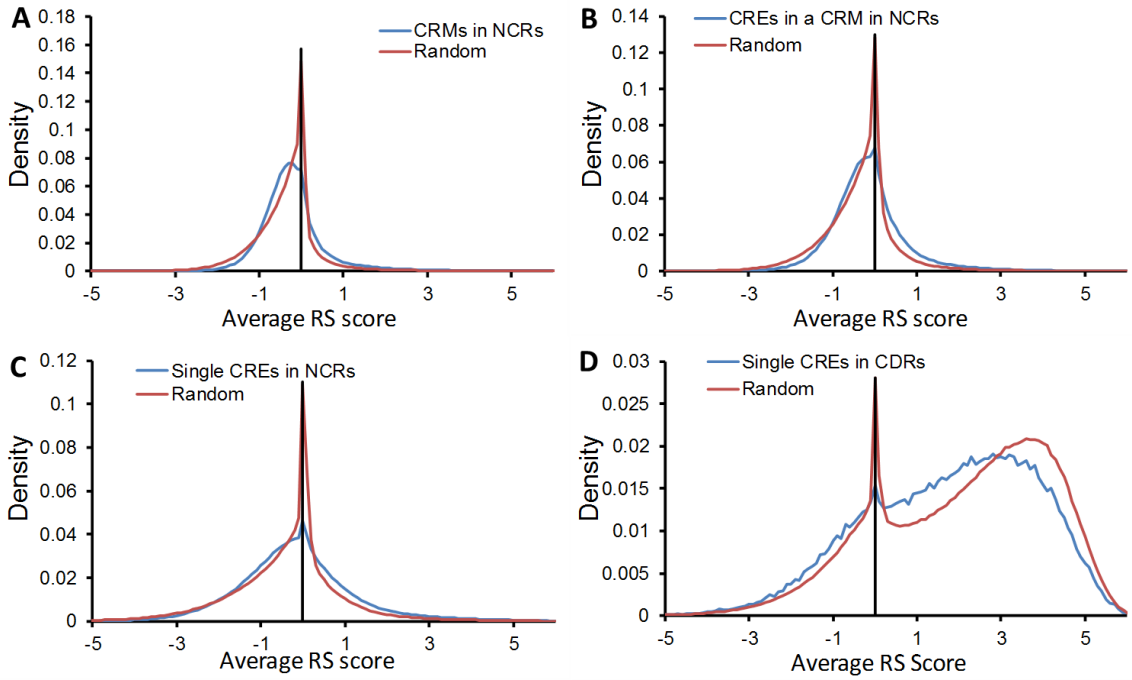

Figure 2.11. **A.** Distribution of average RS scores of the predicted CRMs in NCRs and of the same number and length sequences randomly selected from NCRs. **B**. Distribution of average RS scores of all putative CREs in a predicted CRMs in NCRs and of the same number and length sequences randomly selected from NCRs. **C.** Distribution of average RS scores of single predicted CREs in NCRs and of the same number and length sequences randomly selected from NCRs. D. Distribution of average RS scores of single predicted CREs in CDRs and of the same number and length sequences randomly selected from CDRs.

2.4.6 Functional Elements Revealed by Independent Studies Are Highly Enriched in Predicted CRMs

To further evaluate our predicted CREs and CRMs, we analyzed the enrichment of trait-linked SNPs in our predictions. Of the 30,572 SNPs documented in the dbGAP database, 14,344 (46.9%) are located in our extended biding peaks in the datasets that cover 34.8% of the genome, indicating that these SNPs are already highly enriched in the datasets. This result is not surprising as it has been shown that about 95% of these trait-linked SNPs are located in the NCRs, and most likely are in the regulatory sequences [10]. We then wanted to know how well these 14,344 SNPs that were already enriched in

our input datasets could be further enriched in our predicted CRMs by the algorithm. To this end, we consider a SNP is recovered by our predicted CRM if it is within a window of 200bs to our predicted CRMs, considering the possibility that a SNP itself may not necessarily be functional but in disequilibrium with nearby functional sequences. Remarkably, 7,001 (48.81%) of the 14,344 SNPs were recovered by the predicted CRMs. By contrast, when the same number and length sequences were randomly selected from the genome regions that are covered by the datasets, only 34.22% (repeated 50 times) (Figure 2.12A) of the 14,344 SNPs were recovered. Thus, the SNPs were further enriched in our predicted CRMs by 14.59% compared with a naïve way that randomly selects sequences from ChIP datasets that are already highly enriched for the SNPs.

Furthermore, DHSs are the regions in the genome that have less condense structure, and are open (sensitive) to cleavage by DNase I enzyme in certain cell types and tissues. They are also likely bound by TFs in these cell types and tissues, and work as CRMs. A large number of DHSs in 125 human cell or tissue types have been recently determined by the ENCODE consortium, thus, we used them as additional lines of independent evidence to further validate our predicted CRMs. We consider a DHS is recovered by a predicted CRMs if the DHS overlaps with at least one putative CRE in the predicted CRM because our predicted CRMs always start and end with a putative CRE. Of the 1,281,988 non-overlapping DHSs provided by the ENCODE consortium, 815,790 (63.63%) are located in our input datasets, indicating that they are also highly enriched in the datasets. As shown in Figure 2.12B, 437,397 (50.96%) of the DHSs in the datasets are recovered by our predicted CRMs. By contrast, the same number and length sequenced randomly selected from the dataset can only recover 252,630 (29.43%) of the DHSs.

Thus, the DHSs were further enriched in our predicted CRMs by 36.37% compared with a naïve way that randomly selects sequences from ChIP datasets that are already highly enriched for the DHSs.

Additionally, we validated our predicted CRMs with enhancer segments from VISTA [135]. Of the 850 experimentally verified enhancer segments from VISTA, 625 (73.53%) are located in our input datasets. Although we used them to help set the S_c cutoff in the very early step of the algorithm, if our algorithm does not work, they can still be lost during filtering process of the algorithm as vast majority of input motif and CPs are dropped out by the algorithm. However, 558 (89.28%) of the 625 enhancers were recovered by the algorithm. By contrast, the same number and length sequence as the predicted CRMs. randomly selected from the datasets could only recover an average of 403 (64.47%) of enhancer segments (Figure 2.12C). Taken together, all these three lines of independent evidence indicate that our algorithm has achieved a high recovery rate of possible CRMs.

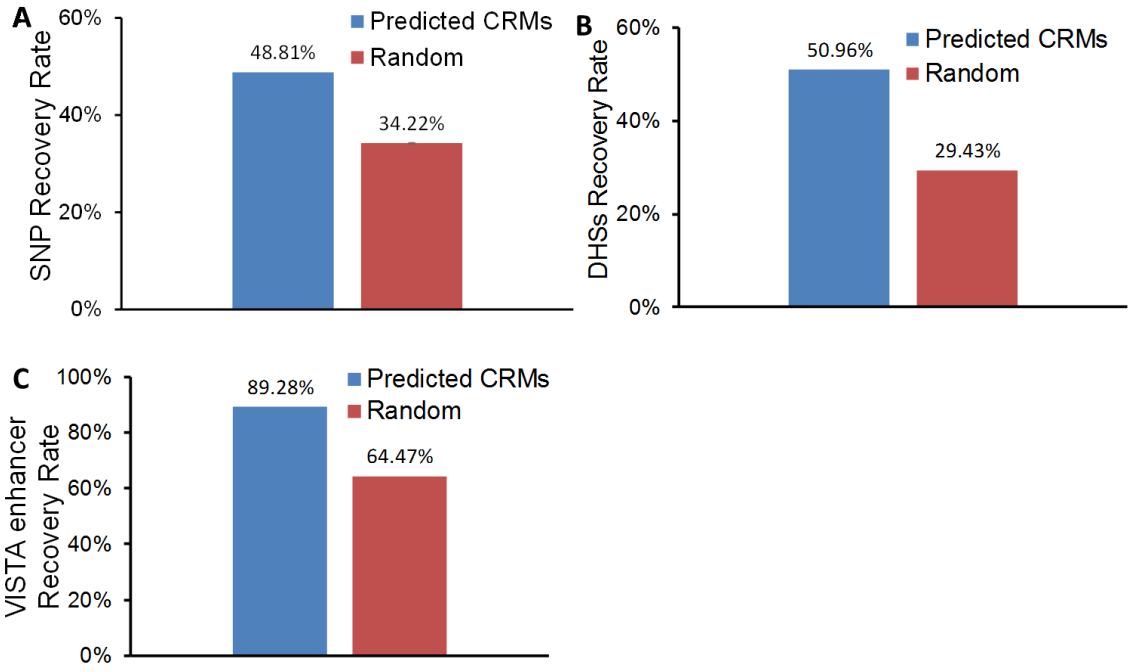

Figure 2.12. **A.** Enrichment of trait-linked SNPs in the predicted CRMs. **B.** Enrichment of DNase I hypersensitive sites in the predicted CRMs. C. Enrichment of VISTA enhancer segments in the predicted CRMs.

2.4.7 Prediction Saturation Analysis

Next, we analyzed the trends of changes in the numbers of CRMCs and Umotifs predicted using an increasing number of datasets in three scenarios: 1) an increasing number of the datasets for different TFs in the same cell type or tissue; 2) an increasing number of datasets in different cell types or tissues for the same TF; and 3) an increasing number of datasets randomly selected from all the datasets.

For the first scenario, we used cell lines in which enough number of datasets for different TFs are available, including K562 (the first human immortalised myelogenous leukemia cell line), GM12878 (a lymphoblastoid cell line) and Helas3 (a sub-clone of the HeLa cell line), in which 62, 65 and 39 datasets are available, respectively. The results based on the datasets in K562 are shown in Figures 2.13A and 2.13B, in which we

plotted the first to fourth quarter trend lines. The decreasing slopes of these trend lines clearly indicate a saturation trend for both the predicted Umotifs (Figure 2.13A) and CRMCs (Figure 2.13B) with the increasing number of datasets used. Surprisingly, the trends of saturation begin to be notable when as few as 5~9 datasets were used for the predictions. The saturation trends for both cases can be well fitted to a sigmoid function (formula 2.2). Extrapolation of the fitting functions suggests that up to 959 Umotifs and 5,489 of their combinatorial patterns in the form of CRMCs using enough datasets in the K562 cells, and that the current 62 datasets in K562 covers 31.55% (300) of all possible Umotifs (959) and 16.18% (883) of all possible CRMCs (5,489) that potentially function in the K562 cells (Table 2.2, Table 2.3). Additionally, as shown in the Figure 2.14, the fitting curve also shows that the contribution by increasing number of datasets decreases as more datasets are used. Thus, it is would be beneficial to understand the optimal number of datasets to be produced in each cell or tissue type to predict a desired proportion of Umotifs or CRMCs for the highest cost efficiency. We attempted to answer this question using the fitting function, and the results are shown in Table 2.2 and 2.3 for Umotifs and CRMCs, respectively. For example, in cell line K562, if one wants to identify 70% of the Umotifs that possible works as CRMC, he/she might need 733 datasets for different TFs; and if one wants to increase the percentage to 80%, then he/she might need 1,071 datasets (338 more) for different TFs, assuming the same level of diversity of the TFs used to generate the datasets as those used in the analysis. Similar results were obtained using datasets in the GM12878 and HeLaS3 cell lines as shown Figures 15 to Figure 18.

Table 2.2. Number of datasets needed for predicting desired portions of Umotifs based on the fitting function.

% of Umotifs	Cell line			TF				
	Gm12878	K562	HelaS3	CTCF	NFKB	NRSF	TAF1	Random
50%	97	173	104	97	35	201	9	251
55%	125	225	130	128	45	263	11	333
60%	162	294	165	168	57	345	13	443
65%	215	390	212	224	74	459	15	596
70%	282	526	275	305	98	622	19	821
75%	386	733	368	428	132	868	23	1165
80%	554	1071	513	632	187	1274	31	1742
85%	858	1695	767	1011	285	2021	42	2816
90%	1534	3114	1310	1874	498	3737	63	5403

Table 2.3 Number of datasets needed for predicting desired portions of CRMCs based on the fitting function.

% of CRMCs	Cell line			TF			
	Gm12878	K562	HelaS3	CTCF	NFKB	NRSF	Random
50%	226	355	226	36	43	100	192
55%	279	439	276	44	52	121	209
60%	345	545	339	55	63	149	227
65%	431	683	420	69	76	185	248
70%	546	870	529	88	94	232	272
75%	709	1135	681	116	119	298	301
80%	957	1539	909	158	154	396	339
85%	1376	2219	1291	229	213	560	391
90%	2214	3597	2056	377	326	888	472

Table 2.4 Number of datasets used in this project and their portions of Umotifs and CRMCs based on the fitting function.

	Cell line			TF				
	Gm12878	K562	HelaS3	CTCF	NFKB	NRSF	TAF1	Random
Datasets used	65	62	39	9	11	10	10	359
% of Umotif	42.11%	31.55%	30.12%	14.69%	27.72%	9.50%	53.80%	56.37%
% of CRMC	23.24%	16.18%	14.91%	21.95%	18.58%	9.07%	n/a	82.07%

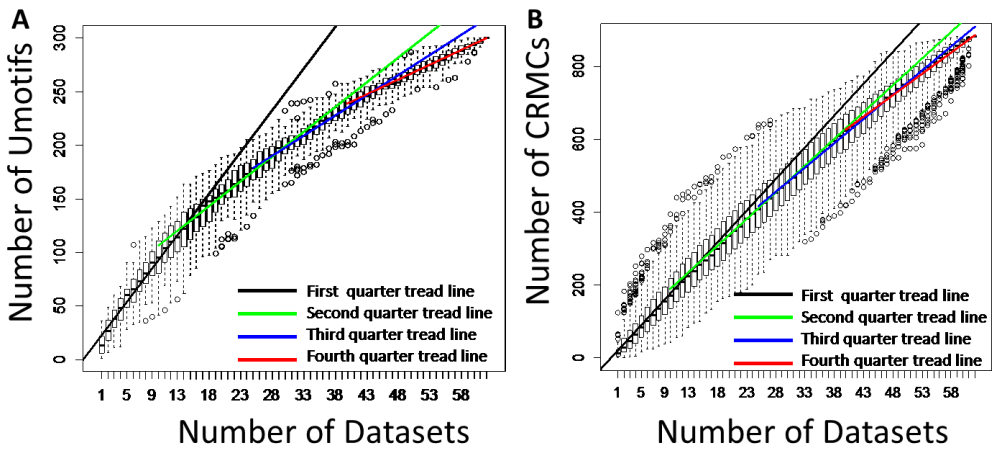

Figure 2.13. The number of recovered Umotifs (A) and CRMCs (B) in the K562 cells show a trend of saturation with the increase in the number of datasets used. The data point is presented using box-plot based on 50 repeats.

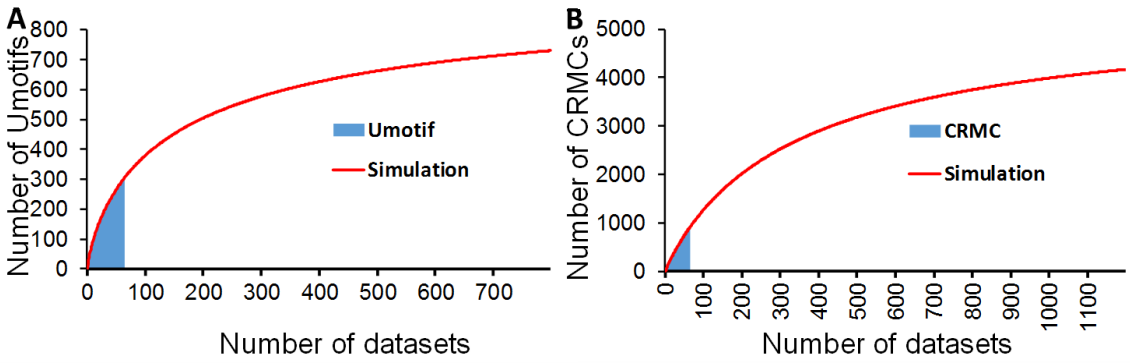

Figure 2.14 Extrapolation of the saturation trends of the number of recovered Umotifs (A) and CRMSs (B) based on the datasets in the K562 cells.

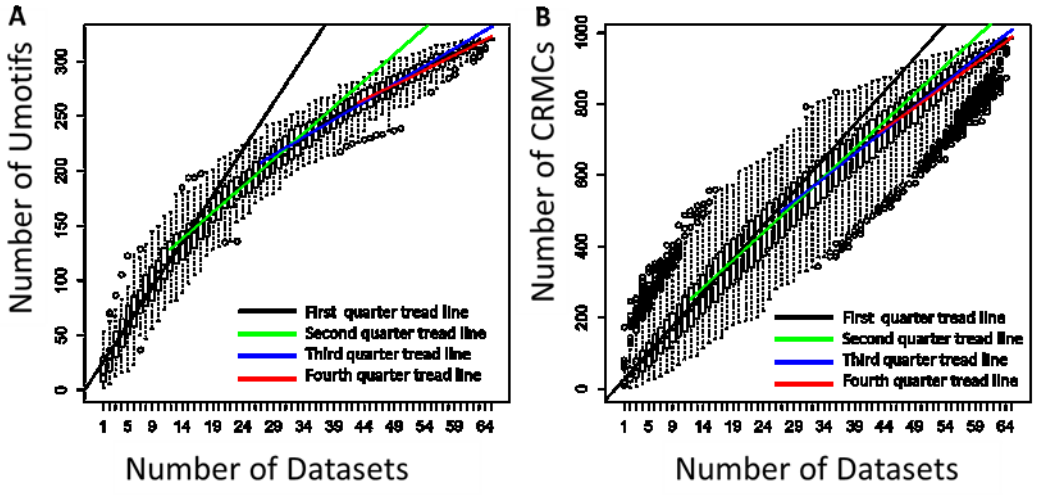

Figure 2.15. The number of recovered Umotifs (A) and CRMCs (B) in the GM12878 cells show a trend of saturation with the increase in the number of datasets used. The data point is presented using box-plot based on 50 repeats.

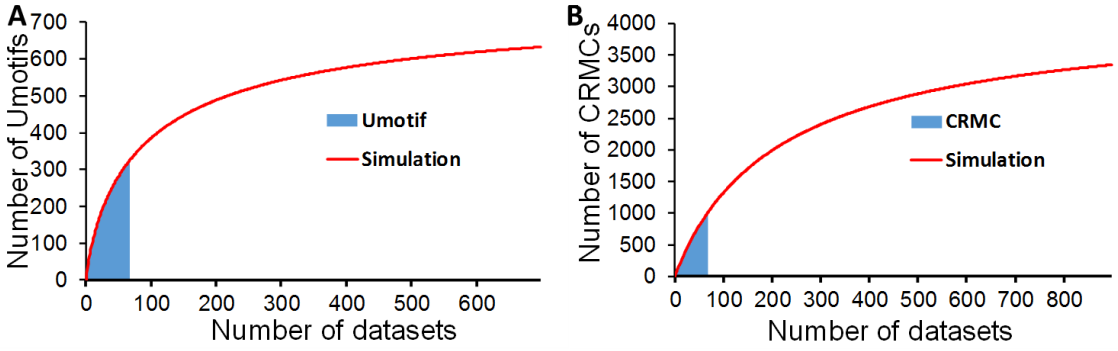

Figure 2.16 Extrapolation of the saturation trends of the number of recovered Umotifs (A) and CRMSs (B) based on the datasets in the GM12878 cells.

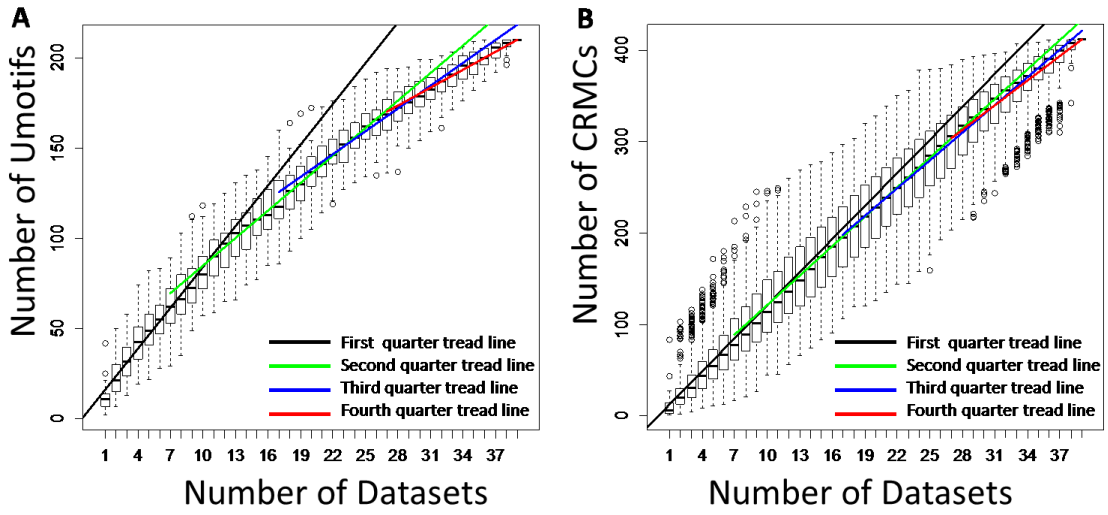

Figure 2.17. The number of recovered Umotifs (A) and CRMCs (B) in the HelaS3 cells show a trend of saturation with the increase in the number of datasets used. The data point is presented using box-plot based on 50 repeats.

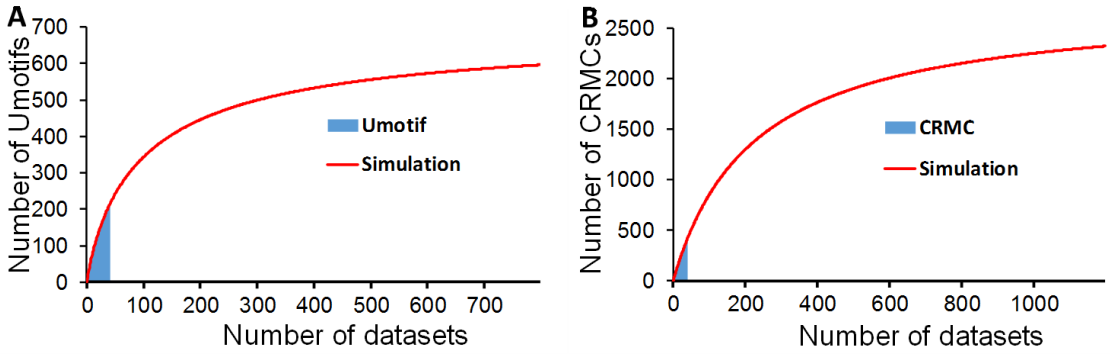

Figure 2.18 Extrapolation of the saturation trends of the number of recovered Umotifs (A) and CRMSs (B) based on the datasets in the HelaS3 cells.

For the second scenario, We used well studied TFs for which a relatively large number of datasets are available in different cell or tissue types, including CTCF, involved in insulator activity[142], V(D)J recombination [143], and regulation of chromatin architecture[144], for which nine datasets are available in different cell lines; NF-kB, involved in the immune and inflammatory responses, developmental processes,

cellular growth, and apoptosis, for which 11 datasets are available; NRSF, involved in the repression of neural genes in non-neuronal cells[145], for which 10 datasets are available; and TAF1, binding to the core promoter to position the polymerase properly and coordinating other related proteins for the initiation of transcription, for which 10 datasets are available. The results based on the nine datasets for NF-kB are shown in Figures 2.19A and Figure 2.19B, in which we only plotted, the first and second half saturation trend lines instead of four-quarters due to the lack of enough number of datasets. Interestingly, similar to the finding in the first scenario, the trends of saturation become notable when as few as 5 and 7 datasets were used for the predictions of Umotifs and CRMCs (Figure 2.19A, 2.19B), respectively. The results for both the Umotifs and CRMCs recovery numbers can be well fitted to a sigmoid function (Figure 2.20). Similar analyses were obtained for the TFs CTCF, NRSF and TAF1 as shown in Figures 21~26, except the fitting of the CRMC recovery number for TAF1 to the function cannot converge, for which more data might be needed for a successful fitting.

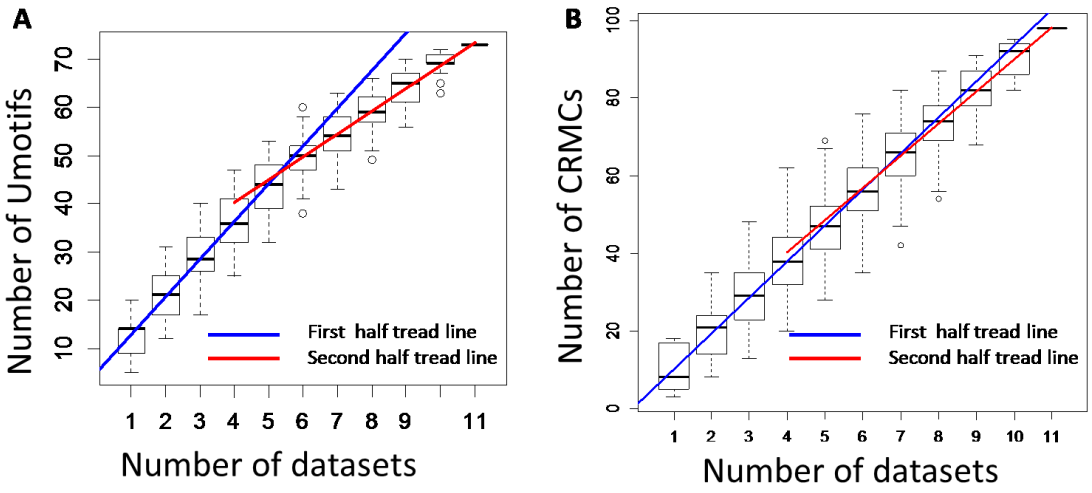

Figure 2.19. The number of recovered Umotifs (A) and CRMCs (B) for TF NF-kB show a trend of saturation with the increase in the number of datasets used from different cell types and tissues. The data point is presented using box-plot based on 50 repeats.

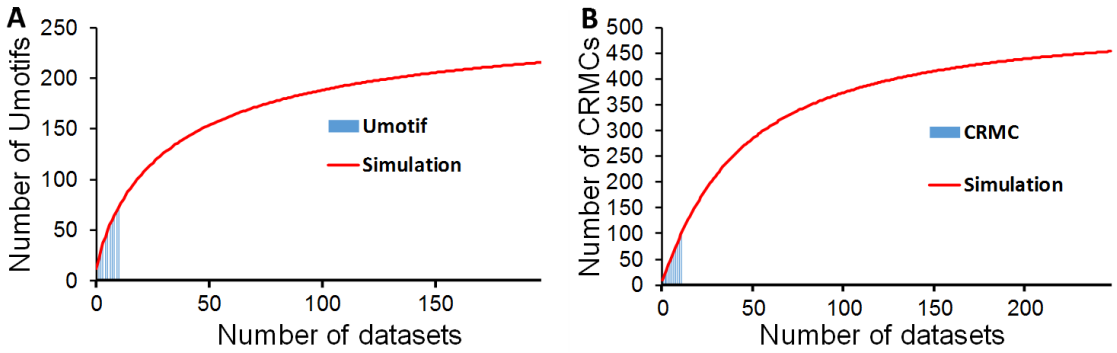

Figure 2.20 **A.** Simulation of the saturation trend based on the Umotif recovery number in TF NF-kB. **B.** Simulation of the saturation trend based on the CRMC recovery number in TF NF-kB.

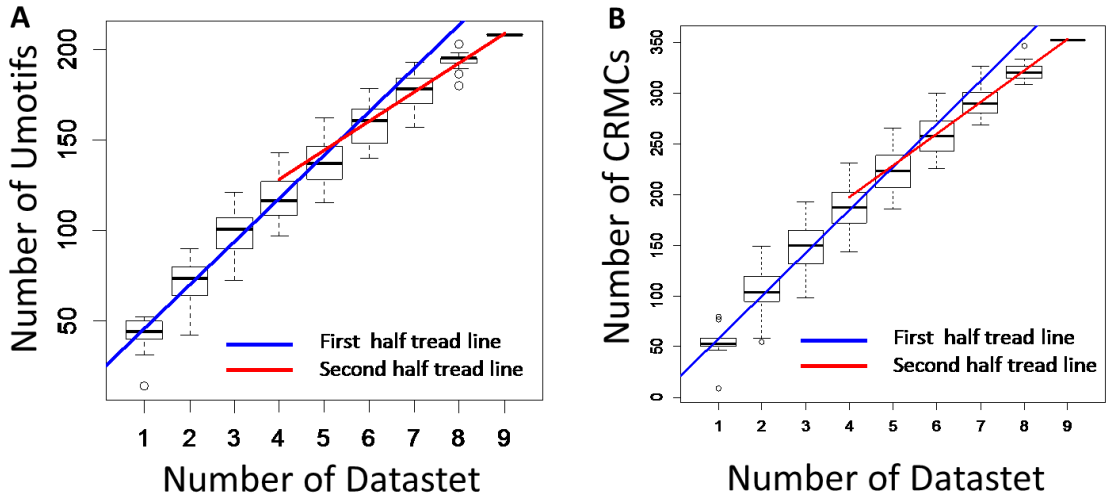

Figure 2.21. The number of recovered Umotifs (A) and CRMCs (B) for TF CTCF show a trend of saturation with the increase in the number of datasets used from different cell types and tissues. The data point is presented using box-plot based on 50 repeats.

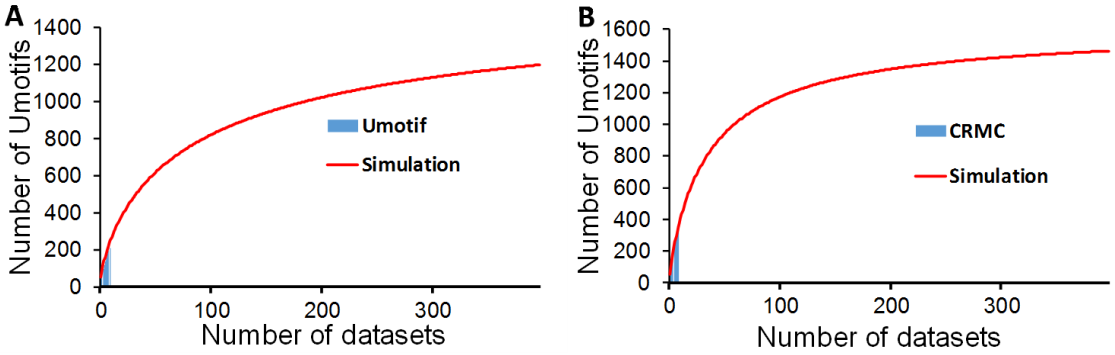

Figure 2.22 Extrapolation of the saturation trends of the number of recovered Umotifs (A) and CRMSs (B) based on the datasets using TF CTCF.

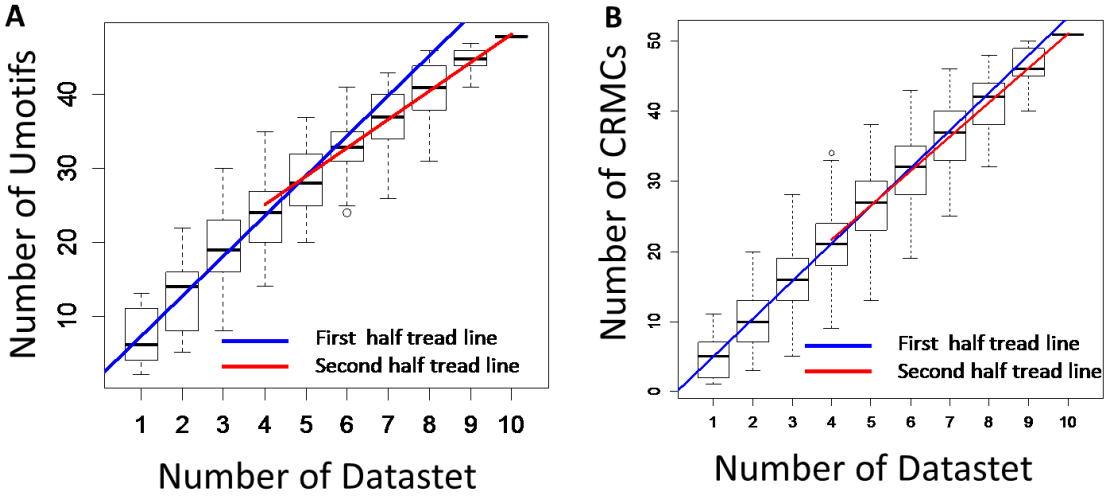

Figure 2.23. The number of recovered Umotifs (A) and CRMCs (B) for TF NRSF show a trend of saturation with the increase in the number of datasets used from different cell types and tissues. The data point is presented using box-plot based on 50 repeats.

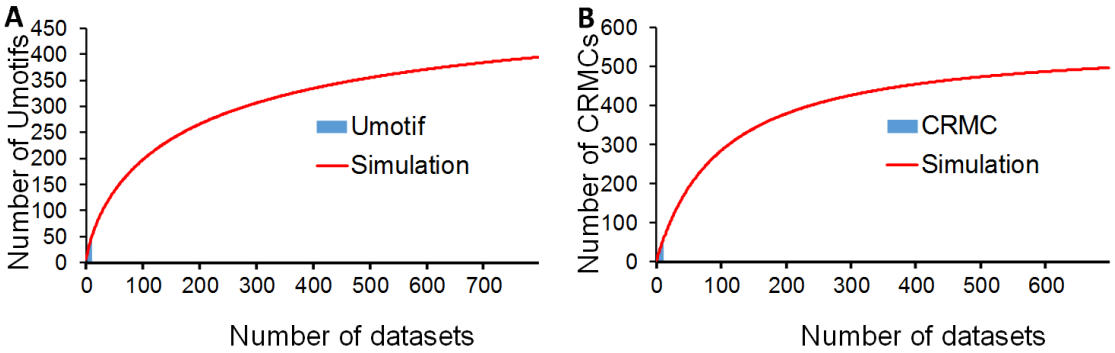

Figure 2.24 Extrapolation of the saturation trends of the number of recovered Umotifs (A) and CRMSs (B) based on the datasets using TF NRSF.

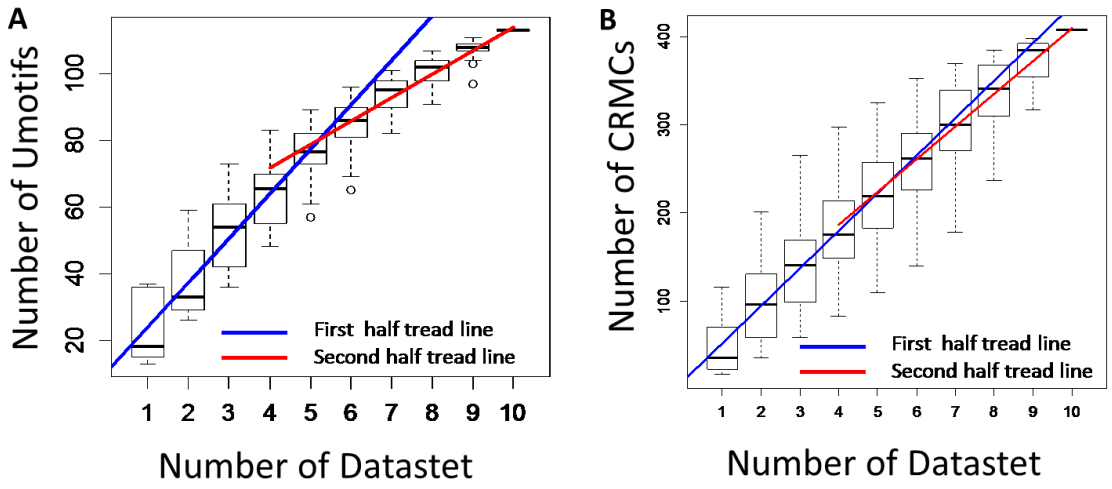

Figure 2.25. The number of recovered Umotifs (A) and CRMCs (B) for TF TAF1 show a trend of saturation with the increase in the number of datasets used from different cell types and tissues. The data point is presented using box-plot based on 50 repeats.

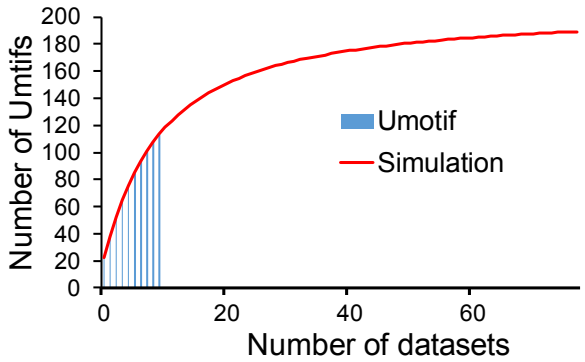

Figure 2.26 Extrapolation of the saturation trends of the number of recovered Umotifs based on the datasets using TF TAF1.

For the third scenario, we used randomly selected datasets with size of 100, 200, 250 and 300 to calculate the number of Umotifs and CRMCs we could predict. As shown in Figure 2.27A and 2.27B, both the numbers of putative Umotifs and CRMCs increased rapidly with the increase in the number of datasets used, but they entered a saturation phase when 200, and 250 datasets were used. Both the numbers can be well fitted the situation functions, with quite different parameters. Interestingly, the fitting of the

CRMCs data indicates that one can barely find the CRMCs using only a few randomly selected datasets, which makes sense because the chance for the relevant TFs to work together is very slim. However, when the number of datasets increased from 8 to 200, the number of CRMCs increased dramatically Extrapolation of the fitting function suggest that we could predict 1,128 Umotifs and 2,425 CMCs using a sufficiently large number of datasets with similar levels of diversity as the 359 datasets used this study, and that we have predicted about 636 (56.37%) of the 1,128 Umotifs and 1,991 (82.07%) of the 2,425 CRMCs using the available datasets. However, the rapid saturation of both the predicted Umotifs and CRMCs after 300 datasets were used, suggesting that it will not be cost effectively to generate more datasets with similar level of the diversity of the 148 TFs in the 68 different cell and tissue types.

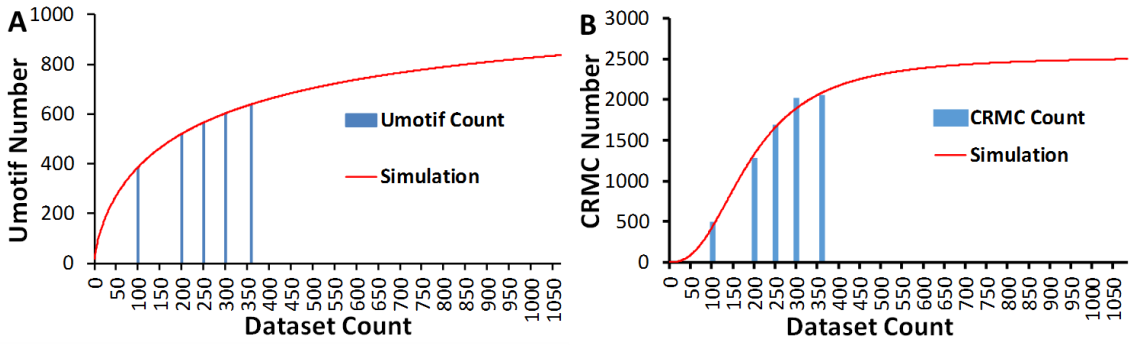

Figure 2.27 **A.** Simulation of saturation trend based on Umotifs recovered by different size of randomly selected datasets. **B.** Simulation of saturation trend based on CRMCs recovered by different size of randomly selected datasets.

2.5. Discussion

We designed the DePCRM algorithm largely based on the fact that similar TF combinatorial patterns are often repeatedly used to regulate multiple similar or different regulons in different cell types, tissues, developmental stages or physiologically

conditions. As the number of possible combinations of TFs is extremely large, DePCRM identifies possible real motif combinatorial patterns in a sufficiently large number of ChIP datasets through iteratively filtering out randomly occurring spurious motifs, thereby effectively reducing the searching space in each step. In order for the algorithm to make reasonable predictions, the ChIP datasets have to be sufficiently large and diverse, so that they are likely to include datasets for cooperative TFs in different cell types, tissues, developmental stages and physiological conditions. Having successfully demonstrated that DePCRM works on the *D. Melanogaster* genome, in this Chapter, we applied the algorithm to the much larger human genome with more and bigger ChIP datasets. In order to make it work more efficiently on large human datasets we modified the algorithm by splitting the large dataset into smaller ones. Such splitting has little effect on the motif-finding results, due probably to the information redundancy in large ChIP datasets. The modified DePCRM successfully worked on the human genome and rendered comparable results to those on the *D. Melanogaster* genome.

The results of conservation analysis on our predicted CREs and CRMs for the human genome is consistent with those for the *D. melanogaster* genome though different conversation measures were used: our predicted CREs and CRMs are more conserved than randomly selected sequences of the same length from NCRs. In particular, individual CREs in CRMs are more like to have gone either strongly negative selection, or moderately positive selection (Figure 2.11C), indicating that they are highly likely to be functional. This observation is in excellent agreement with the consensus that regulatory sequences tend to be more conserved due to negative selection, or to undergo rapid turnover by degrading existing CREs (death), or gaining new CREs (birth) due to

positive selection, a process called CRE turnover [112]. CRE turnover plays a more pivotal role in functional evolution of organisms than previously thought [110, 113], including the evolution for human-specific functions including intelligence.

Our predictions of numerous CREs in CDRs provide us an opportunity to address the long-standing question whether or not CREs in CDRs can serve both as coding and regulatory functions. On one hand, CREs in CDRs are generally more conserved than those in NCRs (Figures 2.11C vs 2.11D), suggesting that CREs in CDRs are under stronger negative selection than the CREs in NCRs, thus may also have coding functions. On the other hand, interestingly, CREs in CDRs are less likely to be selectively neutral, rather, they tend to be either positively or negatively selected. In the scenario of negative selection of CREs in CDRs, the selection pressure tends to weaker than that for randomly selected CDRs (Figure 2.11D). Thus CREs in CDRs might be more flexible to gain the regulatory functions while still possibly coding for a spacer in the protein sequences. In particular, a considerable proportion of randomly selected CDRs are selectively neutral (Figure 2.11D), they might be candidates to adapt novel functions including turning into CREs. However, the generality of these observations needs to be further verified in other larger mammal genomes as the same results were not seen in the *D. melanogaster* genome in which predicted CREs in CDRs are as conserved as the randomly selected CDRs (Figure 1.14D).

By the design of the DePCRM algorithm, theoretically, we need a sufficient number of ChIP-seq datasets to make a meaningful CRE and CRM prediction, if the datasets are produced by randomly selected TFs and cell or tissue types. Moreover, in order to predict all the CREs and CRMs in a genome, we need an even more sufficient

number of datasets that are also diverse enough to include information of all possible combinatory regulations of TFs in the genome. So what is the status of the current datasets in humans to reach the ultimate goal of predicting complete maps of CREs and CRMs in the genome, and are these datasets produced cost effectively for the purpose?

We addressed these questions by analyzing the saturation trends of the numbers of predicted Umotifs and CRMCs under three scenarios based on the 359 datasets collected for 148 TFs in 68 different cell and tissue types in humans. In all the three scenarios, the trends of saturation appeared when only a few datasets were used, suggesting that the datasets are biased to the well-studied cooperative TFs in the cell or tissue types, and this strategy is highly effective for revealing the functional CREs and CRMs in the cells lines and combinatorial utilization of specific TFs. Moreover, the saturation numbers of Umotifs predicted using datasets in different cell or tissue types for specific TFs are in good agreement with the known functions of the TFs. For instance, the saturation number of Umotif for TF CTCF is the highest (1618, Figure 2.22) among all the four TFs evaluated (536, 267 and 205 for Nrsf, Nf-KB and TAF, respectively). This result is consistent with the fact that CTCF generally binds insulators in all cell types tested [146]. Interestingly, except for CTCF, the saturation numbers of Umotifs predicted for specific TFs tested are smaller than those using datasets in specific cells for different TFs (960, 765 and 698 in cell lines K562, GM12878 and HelaS3, respectively) (Figure 2.14, 2.16, 2.18, 2.20, and 2.24). This discrepancy might suggest that these TFs only function in a limited number of cell lines tested and that there is a larger number of tested TFs working together in these cell lines. Furthermore, our saturation analysis suggests that we can potentially predict up to 1,128 Umotifs and 2,425 CRMCs using more than 1,000 datasets

with similar level of diversity as the 359 datasets used this study, and that our predictions using these datasets have recovered 56.37% and 82.07% of the saturation numbers of Umotifs and CRMCs, respectively (Table 2.4). Thus including more datasets with similar level of diversity to the existing datasets is no longer cost effective, rather more diverse TFs and cell lines or tissues should be used to generate more diverse datasets to recover more Umotifs and CRMCs in the human genome. Specifically, as the majority of cell or tissue types used in current studies are immortal cell lines, and all the TFs used are well studied ones, thus more primary tissues and less studied TFs should be included the generation of ChIP-seq datasets in the future. If TFs and cell types are appropriately combined to generate new datasets as the existing 359 datasets, we estimate that we need another 720 datasets in 296 cell or tissue types for 136 TFs to predict a complete map of CREs and CRMs in the human genome.

In addition, our results might allow us to estimate the size or proportion of the human genome that are involved in transcriptional regulation. Within the 359 datasets covering about one third (34.8%) the genome, we can potentially predict ~1,100 Umotifs and ~2,500 CRMCs. If these results are extendable, then we predict that the human genome would encode at three times these numbers of Umotifs and CRMCs. In other words, we estimate that there are about 3,300 Umotifs and 7,500 CRMCs encoded in the human genome. The number of total Umotifs is consistent with the number of TFs in the human genome, which is 2,000~3,000 as estimated by [147] and 2,886 according to the DBD database [148]. Additionally, using these 359 datasets, our predicted CREs and CRMs covers 1.5% and 6.9% of the human genome, and 99% and 98.1% of which are in NCRs, respectively, (Table 2.1 and Figure 2.2). Assuming that these results are

extendable to the other two thirds of the genome that are not covered by our datasets, then we estimate that 4.5% and 20.70% of the genome, or 4.46% and 20.31% of the NCRs, might code for CREs and CRMs (constituent CREs plus spacer sequences), respectively. Furthermore as 32.2% of putative CRMs in NCRs are conserved with an RS score >0, 6.5% of genomes covered by predicted CRMs would be conserved. As 1.23% of the genome are conserved CDRs [132], we estimate that 7.73% of human genome are conserved, which is in agreement with the recent estimation that that approximately 7% of the human genome are conserved [141].

CHAPTER 3: EVOLUTION OF CRMS FROM DROSOPHILA MELANOGASTER TO HUMANS

3.1 Abstract

Despite the large evolution distance between human and *D. melanogaster*, a considerable portion of their genes, as well as expression patterns, are highly conserved [149-151]. Such conservation allows researchers to develop models for a variety of human diseases. The conservation of DNA binding domains in TFs may imply the conservation of their binding motifs. However, a genome scale comparison of the GRNs is still absent due to the lack of detailed information for such a comparison. In this study, we used our predicted CREs and CRMs in the two gnomes, to elucidate the conservation and variation of the GRNs from the perspective of motif composition of CRMCs and orthologous gene groups that are potentially regulated by the relevant CRMs. We found that 62 pairs of CRMCs were conserved both in motif composition and target genes, 1,865 pairs of the CRMCs were conserved in motif composition but not target genes; and 428 pairs of functionally related gene groups were conserved, but regulated by CRMs with different motif composition. Thus, although a large portion of CRMs are conserved in their motif composition, their target genes have been largely changed, meaning that the majority of the GRNs have been rewired during the evolution from *D. melanogaster* to humans.

3.2 Introduction

D. Melanogaster, one of the most studied model animal, shows extensive conservation with humans at the levels of gene, pathway, organ and behavior [31]. This conservation helps researchers successfully develop models for a variety of human diseases: nervous system-related diseases [32]; various muscular dystrophies [33-35]; responses to infection by human pathogens (e.g. [36, 37]); multi-symptom inherited disorders[38, 39]; heart disease [40, 41]; and cancer [42, 43]. Nearly 75% of human disease-causing genes have a functional homolog in *D. melanogaster* [152, 153]. Overall identity at the nucleotide level between *D. melanogaster* and mammals is approximately 40% between homologous genes, and in conserved functional domains, it can be 80% to 90% or higher [154]. Studies also show that some enhancers are conserved between the two species, and are highly likely to perform the similar functions [155-157]. Efforts have been made to construct a regulatory map of specific TFs in the two species, such as the RNA regulatory map between TFs Passilla(PS) in humans and their othologs NOVA1/2 in *D. melanogaster* [158]. However, since *cis*-regulatory systems tend to evolve much faster than the coding sequences [131], it remains to be seen the extent to which the cis-regulatory systems are conserved in the two organisms through a systematic comparison of their cis-regulatory systems. Our accurate predictions of the CREs and CRMs in the two genomes allow us to conduct a comprehensive comparison of the cis-regulatory systems in the two species. We found that although a large portion of CRMs is conserved in their motif composition in the two species, their target genes have been largely changed. Thus, the majority of the GRNs have been rewired during the evolution from *D. melanogaster* to humans.

3.3 Materials and Methods

3.3.1 Materials

We used the results from Chapters 1 and 2, including the 184 Umotifs, 746 CRMCs, 115,932 CRMs, and 14,647 target genes predicted in the *D. melanogaster* genome, and the 636 Umotifs, 1,991 CRMCs, 807,365 CRMs, and 17,561 target genes predicted in the human genome. Refseq mRNA and ncRNA IDs were downloaded from the NCBI website [159].

3.3.2 Mapping CRMCs between the Two Genomes by Target Gene Groups

We used DRSC Integrative Ortholog Prediction Tool (DIOPT) to predict orthologs between the *D. melanogaster* and human genomes. DIOPT integrates the results of 10 widely used orthologs prediction tools [31]. To make the analysis under a more stringent condition, we define that a CRM belongs to a CRMC only if the CRM contains at least one CRE of every Umotifs of the CRMC. For each such CRM, we assign it a potential target gene by the following rules: if the CRM is located in a NCRs, we assign it the gene whose transcription starting site (TSS) is the closest to the CRM; if the CRM is in an exon of a gene, we assign it the adjacent gene with the closest TSS to the CRM. Thus, for each CRMC, we have a target gene group that are potentially regulated by the CRMs of the CRMC. For target gene groups G_m and G_h of CRMCs m and h, each being from *D. melanogaster* and humans, respectively, we compute a score $S_{orth}(m, h)$:

$$S_{orth}(m,h) = \frac{O}{N_m + N_h - O} \quad (3.1)$$

where O is the number of genes that have orthologous relationships in G_m and G_h, N_n the number of genes in G_n. Thus, S_{orth} (m, h) measures the level of conservation between the target genes of CRMCs m and h as illustrated in Figure 3.1. For each CRMC in humans we find its corresponding CRMC in *D. melanogaster*, which has the highest $S_{orth}(m, h)$, and do the same for each CRMC in *D. melanogaster*. We call each of these CRMC pairs a best S_{orth} hit from humans to D. melanogaster and vice versa.

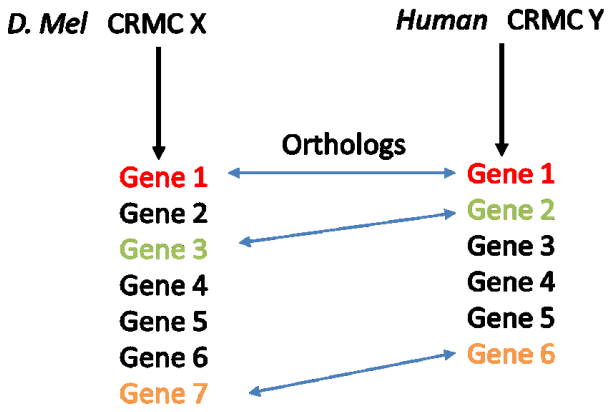

Figure 3.1. Illustration of the orthologous relationship between the target genes of two CRMCs from D. melanogaster and humans.

3.3.3 Mapping CRMCs between the Two Genomes by Motif Composition.

We calculate the motif similarities between motifs in the human and *D. melanogaster* genomes using SPIC [125]. For each pair of the CRMCs m and h from human and *D. melanogaster*, respectively, we compute a score for their similarity,

$$S_{sim}(m,h) = Max_{(i,j)}\Sigma_{(i,j)}\left(SPIC\left(M_m(i), M_h(j)\right)\right)/Min(|M_m|, |M_h|),$$

$$SPIC(M_m, M_h) \geq 0.5, \qquad (3.2)$$

where, $M_m(i)$ and $M_h(j)$ are motifs i and j from the CRMCs m and h, respectively. In computing S_{sim}, we consider all possible pairing between the motifs in m and h, and

sum up the similarity scores of all the pairs. However, we ignore motif pairs that have a similarity score (SPIC) lower than 0.5, as these pairs may increase noise and thus difficulty in identifying similar CRMCs. Then, we compare the total similarities scores of all the possible motif pairs, and divide the highest one by the minimum number of motifs in *m* or *h*. The result is the similarity score between the two CRMCs *m* and *h* from *D. melanogaster* and human. Figure 3.2 shows an example of computing S_{sim}. For each CRMC in human, we find its corresponding *D. melanogaster* CRMC with the highest $S_{sim}(m, h)$, and do the same for each CRMC in *D. melanogaster*. We call each of these CRMC pairs a best S_{sim} hit from humans to D. melanogaster and vice versa.

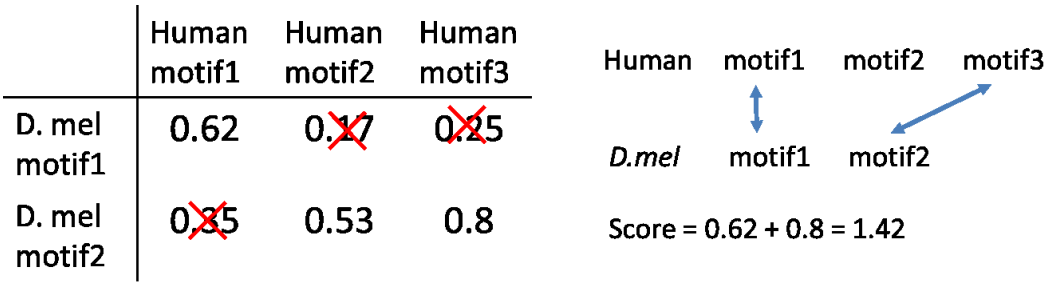

Figure 3.2 Illustration of calculating S_{sim} between a CRMC in D. melanogaster containing two motifs, and a CRMC in human, containing three motifs.

3.4 Results and Discussion

3.4.1 There are Extensive Orthologs between Humans and D. melanogaster

DIOPT integrates the results of 10 ortholog prediction tools that are based on phylogeny-algorithms (e.g. TreeFam, Phylome, Ensembl Compara), sequence similarity (e.g. InParanoid, orthoMCL and OMA) and protein-protein interaction (PPI) networks (e.g. NetworkBLAST, IsoBase). DIOPT predicts orthologous relationships between genes in two organisms using a weighed score of the individual tools. The higher the score, the more consent all the tools reaches. On the other hand, the higher the score, the

fewer ortholog relationships can be found due to the higher stringency. As shown in Figure 3.3A, at the DIOPT score cutoff 2, the decrease in the number of genes that may have orthologs between humans and *D. Melanogaster*, and in the number of orthologous pairs starts to slow down, thus, we chose 2 as the DIOPT score cutoff for the subsequent analysis, which renders 12,218 orthologous pairs between 6,998 *D. Melanogaster genes* and 10,903 Human genes (Figure 3.3B), because multiple mapping is allowed by DIOPT.

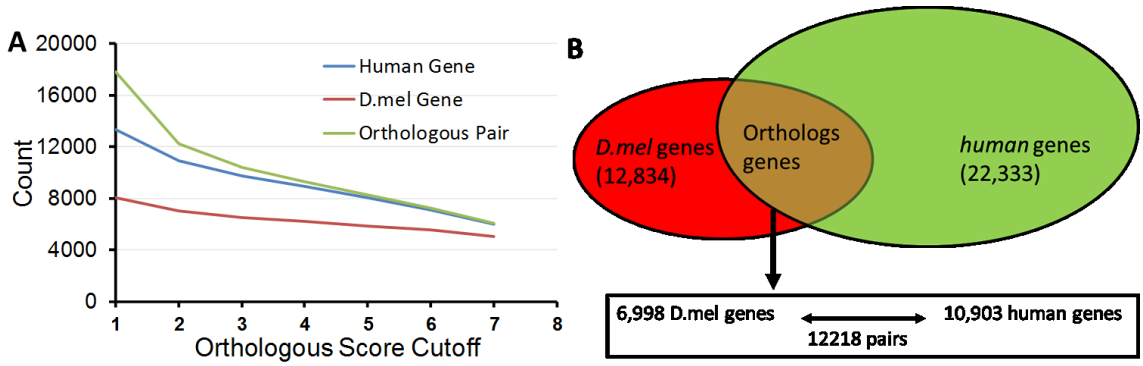

Figure 3.3. **A.** Number of genes that may have an ortholog and number of orthologous pairs, as a function of the DIOPT score cutoff **B.** Van diagram shows the orthologous relationsips between the human and D. melanogaster genes at a DIOPT score cutoff of 2. Many-to-many mapping is allowed by DIOPT.

3.4.2 Diverse Groups of Genes are Regulated by Conserved CRMCs in D. melanogaster and Humans.

We first asked how the CRMCs are conserved in their putative target genes of between D. melanogaster and humans. The distributions of the S_{orth} scores for the best hits from Human CRMCs to *D. melanogaster* CRMCs and vice versa show that almost all the CRMC pairs for both directions of comparison have S_{orth} score less than 0.35 (Figure 3.4). This means that almost all CRMCs have less than ~35% of their target genes conserved between the two species. Moreover, even with a S_{orth} cutoff of 0.04,

only 17 pairs of CRMCs are bidirectional best hit (BDBH) pairs for the S_{orth} score, i.e., each of the CRMCs in the pair is the best hit of the other in the two directions of comparison. These results suggest that the CRMCs have largely changed their target genes since the separation of the two species from their last common ancestor. Interestingly, the distribution for best hit S_{orth} scores from *D. melanogaster* to human shifts to right relative to that for the other way of comparison, indicating that co-regulated genes in *D. melanogaster* are more likely to have co-regulated orthologs in humans than are co-regulated human genes in *D. melanogaster*. This might be due to the fact that humans have evolved more additional ways and more complex cis-regulation mechanisms than does *D. melanogaster*, as also endorsed by our results that we have identified more CRMCs in human (1991) than in D. melanogaster (746).

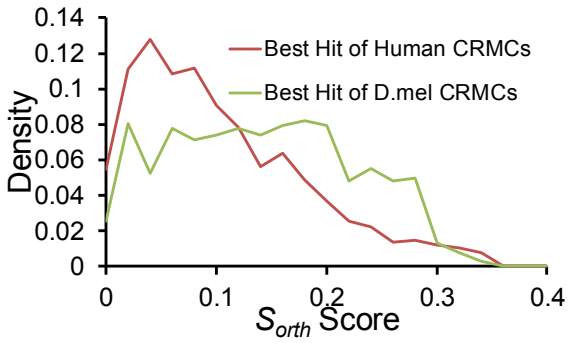

Figure 3.4. Distribution of S_{orth} scores for best hit CRMC pairs from human and *D. melanogaster* and vice versa.

3.4.3 CRMCs Tend to be Conserved in Their Motif Components between Humans and *D. melanogaster*

We then asked how the CRMCs are conserved in motif compositions between D. melanogaster and humans. To this end, we first find the similar motifs between the two

species. As shown in Figure 3.5, the pairs from D. melanogaster to humans tend to have a higher similarity score than the pairs from human to D. melanogaster, probably due to the same aforementioned reasons for the conservation of target gene groups of the CRMCs. In this context, we identified more Umotifs in the human (634) than in D. melanogaster (184). In both directions of comparison, a large portion of best hit motif pairs have a score >0.5. As we have shown in Chapter 2 that two motifs with a SPIC score > 0.5 are highly similar to each other, thus a large portion of TF binding motifs are likely conserved between the two species. Furthermore, 85 motif pairs are BDBHs, meaning that the two motifs in a pair are best hits of each other in the two directions of comparison. These BDBH pairs tend to have a higher similarity scores than those that are only best hit in one direction of comparison (Figure 3.5), suggesting that they are more likely to have orthologous relationships. These results are consistent with, and can be at least partially explained by the fact that the DNA binding domains of many TFs in the two organisms are largely conserved.

Using TOMTOM [88] we were able to predict the cognate TFs of some of these BDBH motif pairs in both organisms. The putative cognate TFs for a BDBH motif pair are often evolutionarily related. The top 5 BDBH motif pairs, their similarity scores and putative cognate TFs are shown in Figure 3.6. The cognate TFs for the third pair of Umotifs (Umotifs 605 in human and 131 in *D. melanogaster*) are POU5F1 and PDM2 in human and D. melanogaster, respectively, and POU5F1 and PDM2 are orthologs according to DIOPT at a score of 2. Moreover, for the fourth pair of motifs, the cognate TF NFIIL3 for human motif 216 is the ortholog of the cognate TF VRI for *D.*

melanogaster motif 46. However, to make a more stringent analysis, we will only consider the motif pairs with a SPIC similarity score >0.5 in the subsequent analysis.

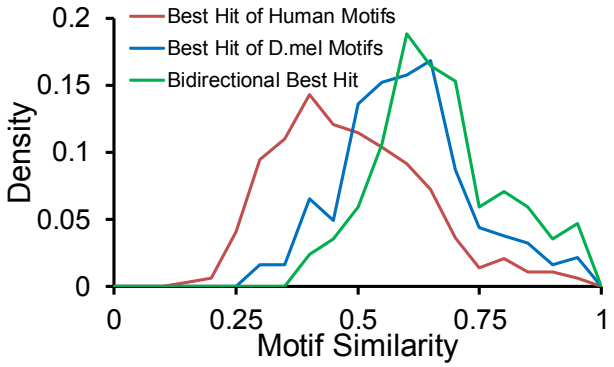

Figure 3.5 Distribution of motif similarity scores of best hits from human to *D. melanogaster*, from *D. melanogaster* to human, and the bidirectional best hits.

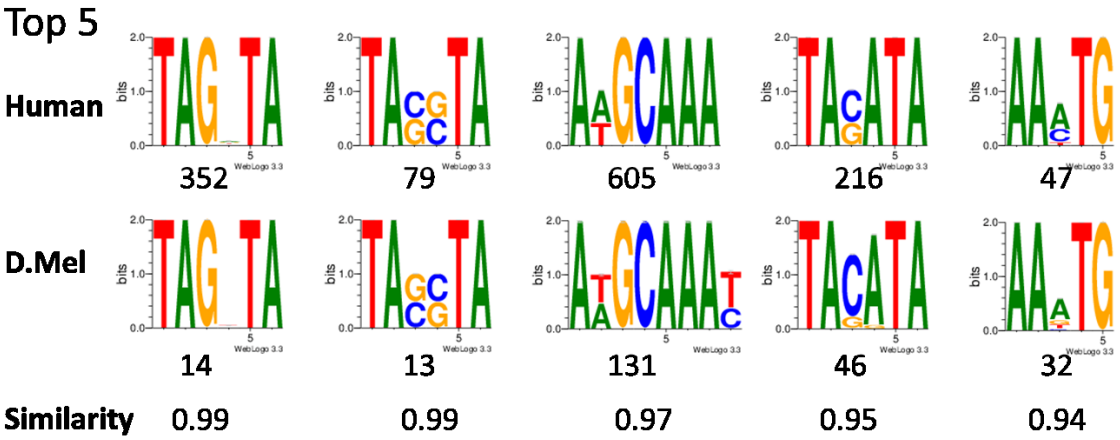

Figure 3.6. Examples of the top 5 bidirectional best-hit motif pairs between *D. melanogaster* and human; the similarity scores are shown under the logos.

Based on these identified best hit motif pairs, we analyzed the conservation of CRMCs for their motif compositions in the two organisms. As shown in Figure 3.7, the distribution of best hit S_{sim} scores shows a similar trend to that of best hit motif similarity scores (Figure 3.5) in that the S_{sim} scores from D. melanogaster to humans tend to be

higher than those from human to D. melanogaster. Moreover, 85 CRMC pairs are BDBHs for the S_{sim} scores, which tend to have the highest scores. These result suggests that not only motifs, but also a large number of motif combinatory usages are conserved between D. melanogaster and human. However, the result that only a few (17) BDBH CRMCs can be identified using S_{orth} score, suggests that the motifs composition of CRMs are more conserved than their target genes. In other words, the same or similar CRMs are used to regulate different set of genes, i.e., the GRNs have been largely rewired during the evolution from D. melanogaster to humans.

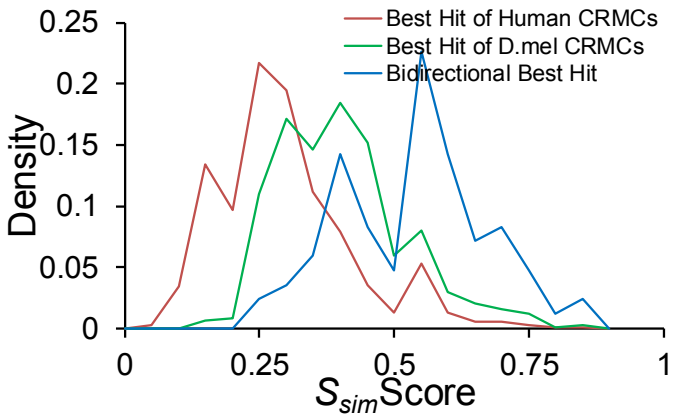

Figure 3.7. A. Distribution of score $\boldsymbol{S_{sim}}$ of best hits from human to *D. melanogaster* and from *D. melanogaster* to human, and bidirectional best hits, respectively.

3.4.4 Three Scenarios of the Evolution of CRMCs

We noted that there is no case where a CRMC pair is a best hit for both the S_{orth} and S_{sim} scores for either from *D. melanogaster* to humans, or from human to *D. melanogaster* comparisons, needless to say the possibility of BDBH CRMC pairs for both the S_{sim} and S_{orth} scores between the two organisms. These results again strongly suggest that the GRNs have been largely rewired during the speciation of the two organisms. To further investigate the relationships between a CRMCs best hit S_{sim} and the

corresponding S_{orth} scores and vice versa, we plotted the two scores for each CRMC in both the genomes for both comparisons. Overall, there is little correlation between a CRMC's best-hit S_{sim} core and the corresponding S_{orth} score for both comparisons (Figures 3.8A and 3.8B), and the same is true for a CRMC's best-hit S_{orth} core and the corresponding S_{sim} score for both comparisons (Figure 3.8C and 3.8D). However, interestingly, the distributions of CRMCs in these two plots for each comparison are almost exclusive, and there are clearly a few patterns worth noting.

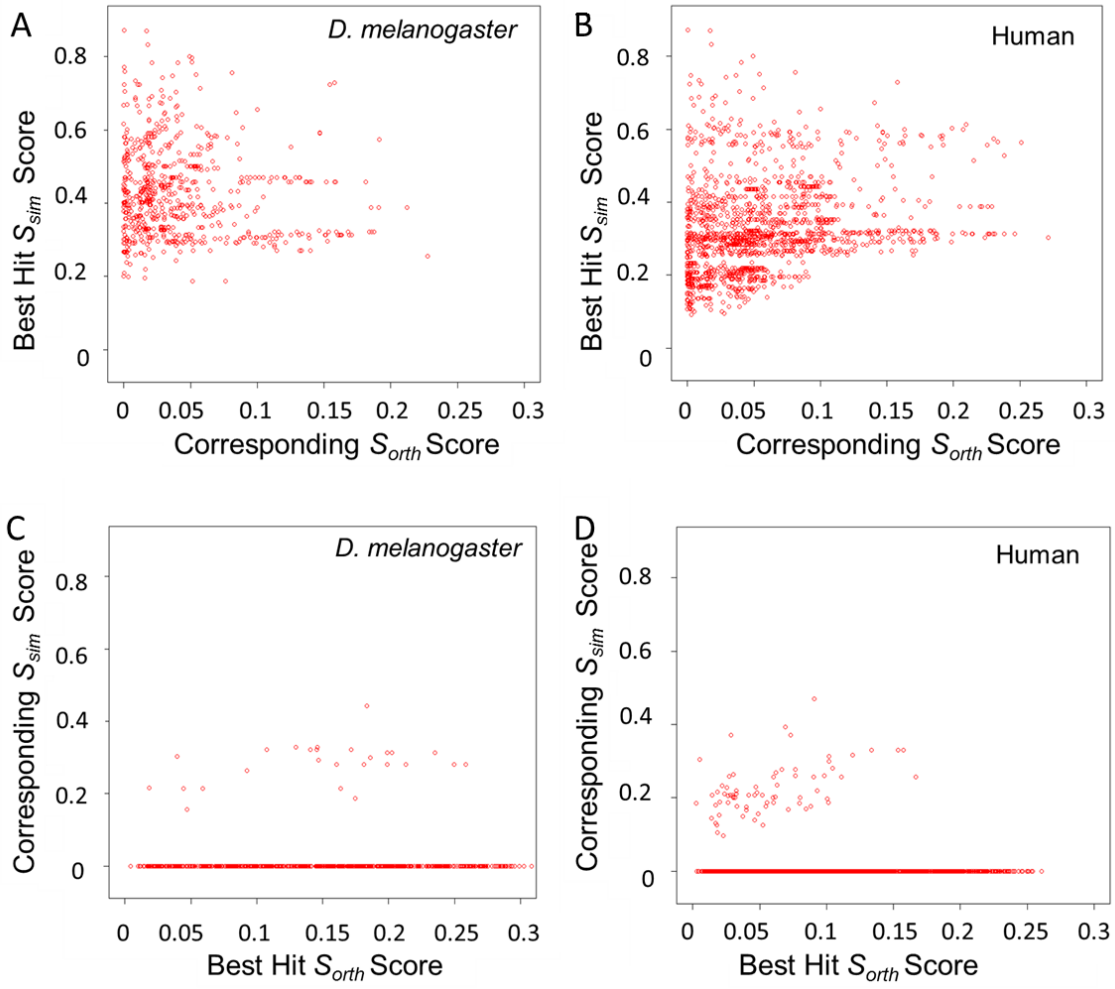

Figure 3.8. A. Relationship between the best hit S_{sim} score of a *D. melanogaster* CRMC with the corresponding S_{orth} score for *D. melanogaster* to human comparison. B.

(Continued) Relationship between the best hit S_{sim} score of a human CRMC with the corresponding S_{orth} score for human to *D. melanogaster* comparison. C. Relationship between of the best hit S_{orth} score of a *D. melanogaster* CRMC with the corresponding S_{sim} score for *D. melanogaster* to human comparison. D. Relationship between the best hit S_{orth} score of a human CRMC with the corresponding S_{sim} score for human to *D. melanogaster* comparison.

First, there are very few CRMCs with a relatively low best hit S_{orth} score but a high corresponding S_{sim} score (Figures 3.8C and 3.8D), suggesting that there are very few CRMCs that are conversed in motif compositions when their target genes are not conserved. By contrast, there are a large number of CRMCs with a high best-hit S_{sim} score but a relatively low corresponding S_{orth} score (Figures 3.8A and 3.8B), indicating that a large number of CRMCs are conserved in their motif composition, but might have adapted new functions during the course of evolution. For example, as shown in Figure 3.9, CRMC 1801 in human consisting of three motifs, Umotif 95, 473 and 535, and CRMC 676 in *D. melanogaster* consisting of three motifs, motif 10, 12 and motif 23, have a best hit S_{sim} score of 0.783, thus they are very similar to each other, respectively. By contrast, their corresponding S_{orth} score is as low as 0.04. Of the 832 and 2,965 putative target genes of CRMCs 1801 and 676, 777 and 2,395 have GO-term assigned, respectively. These putative target genes of CRMC 676 in *D. melanogaster* are enriched for imaginal disc development, cell motion and transcription activity, while those of the CRMC 1801 in human are enriched with for quite different functions, including regulation of metabolic process, cell adhesion and neuron development. Thus, these two CRMCs regulate groups of genes with quite different functions using highly similar motifs and thus likely similar TFs.

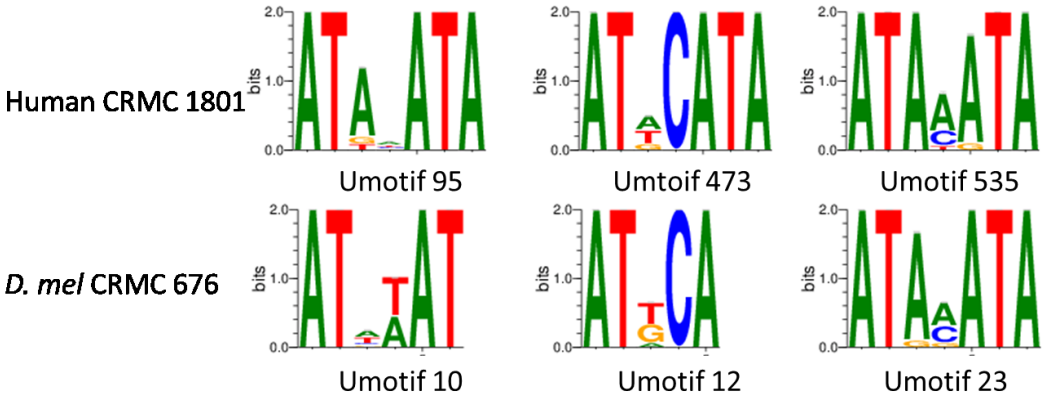

Figure 3.9 Umotifs of human CRMC 1801 and *D. melanogaster* CRMC 676.

Second, there is virtually no case where a CRMC has a very low best hit S_{sim} score but a high corresponding S_{orth} score (Figures 3.8A and 3.8B), indicating that when a CRMCs is not conserved in its motif composition, it also change its target genes. By contrast, there are a considerable number of CRMCs with a high best-hit S_{orth} score but almost zero corresponding S_{sim} score (Figures 3.8C and 3.8D), indicating that functionally similar groups of genes are can be regulated by different combination of motifs. These genes might have changed their regulators during the course of evolution. For example, human CRMC 1053 consisting of Umoits 99 and 103, and *D. melanogaster* CRMC 71 consisting of Umotifs 13, 14, and 45, have an S_{sim} Score = 0, meaning that the similarity score between any of the possible motif pairs is less than 0.5, thus they are not similar to each other (Figure 3.10). However, the S_{orth} score between the two CRMCs is as high as 0.18. Both of their target gene groups are enriched for transcriptional regulation, cell morphogenesis, neurogenesis regulation, neuron development, and phosphorus metabolic process.

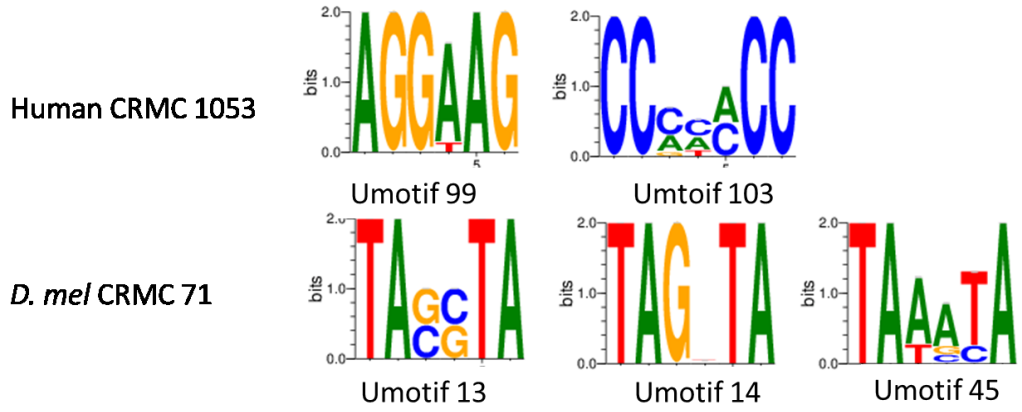

Figure 3.10. Umotifs of human CRMC 1053 and *D. melanogaster* CRMC 71.

Third, although there are very few CRMCs with both a high best hit S_{orth} score and a high corresponding S_{sim} score (Figures 3.8C and 3.8D), there are some CRMCs with both a high best-hit S_{sim} score and a high corresponding S_{orth} (Figures 3.8A and 3.8B). Thus, these latter CRMCs are conserved in both motif compositions and target genes. For example, human CRMC 1645 consisting of two Umotifs 128 and 323, and *D. melanogater* CRMC 364 consisting of two Umotifs 4 and 35, have both a high best-hit S_{sim} (0.5737) and a high corresponding S_{orth} (0.1912) score. These motifs are highly similar to each other, respectively (Figure 3.11). Moreover, of the 2,028 and 1614 targets genes assigned to the CRMCs 1645 and 364, 1938 and 1491 have GO-term assignments, and they are both enriched for transcription regulation, neuron development, and cytoskeleton organization.

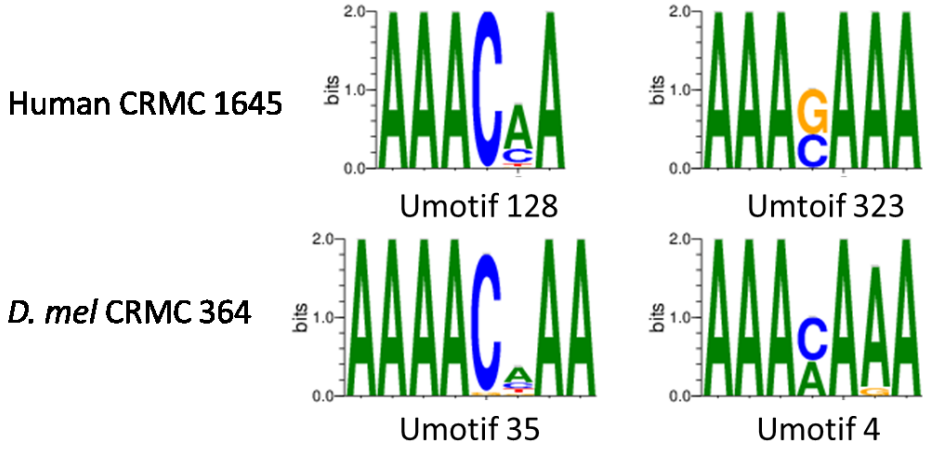

Figure 3.11 Umotifs of human CRMC 1645 and *D. melanogaster* CRMC 364.

CHAPTER 4: CONCLUSION

In this dissertation, we aimed to develop and novel algorithm to predict CREs and CRMs in large eukaryotic genomes by effectively integrating a large number of big ChIP-seq datasets in the organisms, and predict a comprehensive map of CREs and CRMs in major model organisms and humans. Our results presented in Chapters 1~3 indicate that we have largely achieved the goals.

In chapter 1, we developed the DePCRM algorithm that predicts CREs and CRMs in a genome by identifying over-represented combinatorial motifs patterns in a large number of ChIP datasets. We evaluated the algorithm using 168 ChIP datasets for 56 TFs in *D. melanogaster*. We identified 184 over-represented CRE motifs and their 746 combinatorial patterns, and predicted a total of 115,932 CRMs in the genome. We validated our results using the known CRMs in the REDfly database [80], and recovered 77.9% of known CRMs in the datasets, and 89.3% of known CRMs containing at least one predicted CRE. Our predicted CRE and CRMs in the NCRs are more conserved than randomly selected NCRs, thus our predictions are highly likely to be functional. By extrapolating the our predictions based these currently available ChIP-seq datasets, we predict that 11.9% and 65% of the *D. melanogaster* genome code for CREs and CRMs, respectively, which is consistent with the results from comparative genomics analysis.

In Chapter 2, we applied the DePCRM algorithm to the human genome using 359 ChIP-seq datasets for 148 TFs in 68 different types. Although the datasets from human

cells or tissues are much larger than those form C. melanogaster tissues, the algorithm works as efficiently on the human ChIP-seq datasets as on the *D. melanogaster* datasets as we split the very big datasets into small ones without compromising motif finding. We identified 636 overrepresented motifs, 1,991 of their combinatorial patterns and 807,365, CRMs in the genome. As in the case for *D. melanogaster* genome, the predicted CREs and CRMs in NCRs tend to be more conserved than randomly selected NCRs. Furthermore, the predicted CRMs can recover 18.28% of known CRMs in the VISTA database, and are highly enriched for both DHSs, and trait-linked SNPs from dbGAP. We also analyzed the trend of saturation of the predicted CRMCs and Umotifs using increasing numbers of datasets in specific cell types, for specific TFs and randomly selected datasets. By extrapolating the our predictions using the currently available ChIP-seq datasets, we predict that 4.3% and 19.8% of the human genome might code for CRE and CRMs, respectively, which as in the case of D. melanogaster, is consistent with the results from comparative genomics analysis. We found that the trends of saturation of both Umotifs and CRMCs become notable when a few datasets are used in all the three scenarios. Based the trends of such saturation, we are able to predict the number of datasets needed to cover a certain the proportion of CRMCs and Umotifs in the cell or tissue types, for the TFs and in the whole genome, which can be used to guide experiment design for more cost efficiency.

In Chapter 3, we capitalized on the large number of high quality CREs and CRMs in the *D. melanogaster* and human genomes predicted in the previous Chapters, and analyzed the conservation and variation of CRMs and their target genes between the two genomes. We found 85 BDBH Umotifs pairs, and 12,218 orthologs gene pairs between

the two genomes, suggesting that a larger number of motifs and genes are conserved in the two species. Moreover, a large number of CRMCs are conserved in their motif composition, but not their target genes, while a considerable number CRMCs are conserved in their target genes but not their motif composition. Only a small number of CRMCs are conserved in both their motif composition and target genes. Thus, the GRNs have been largely rewired sine the separation of the two species from their last common ancestor.

To summarize, we have developed a novel, efficient and accurate algorithm for de novo prediction of CREs and CRMs in eukaryotic genomes by integrating a large number of ChIP-seq datasets. The algorithm is robust to work in both relatively small compact genomes such as the *D. melanogaster* genome and very large sparse genomes such as the human genome. Using this algorithm, we have predicted so-far the most comprehensive CRE and CRMs maps the *D. melanogaster* and human genomes. With more ChIP-seq datasets available in the future, this algorithm will be very useful for deciphering the cis-regulatory codes in the genomes.

REFERENCES

1. Maher B: ENCODE: The human encyclopaedia. *Nature* 2012, 489(7414):46-48.

2. Pagani I, Liolios K, Jansson J, Chen IM, Smirnova T, Nosrat B, Markowitz VM, Kyrpides NC: The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res* 2012, 40(Database issue):D571-579.

3. Rubinstein M, de Souza FS: Evolution of transcriptional enhancers and animal diversity. *Philosophical transactions of the Royal Society of London Series B, Biological sciences* 2013, 368(1632):20130017.

4. Gazave E, Marques-Bonet T, Fernando O, Charlesworth B, Navarro A: Patterns and rates of intron divergence between humans and chimpanzees. *Genome Biol* 2007, 8(2):R21.

5. Pai AA, Bell JT, Marioni JC, Pritchard JK, Gilad Y: A genome-wide study of DNA methylation patterns and gene expression levels in multiple human and chimpanzee tissues. *PLoS Genet* 2011, 7(2):e1001316.

6. Arbiza L, Gronau I, Aksoy BA, Hubisz MJ, Gulko B, Keinan A, Siepel A: Genome-wide inference of natural selection on human transcription factor binding sites. *Nat Genet* 2013, 45(7):723-729.

7. Domene S, Bumaschny VF, de Souza FS, Franchini LF, Nasif S, Low MJ, Rubinstein M: Enhancer turnover and conserved regulatory function in vertebrate evolution. *Philosophical transactions of the Royal Society of London Series B, Biological sciences* 2013, 368(1632):20130027.

8. Ye K, Lu J, Raj SM, Gu Z: Human expression QTLs are enriched in signals of environmental adaptation. *Genome biology and evolution* 2013, 5(9):1689-1701.

9. Lappalainen T, Dermitzakis ET: Evolutionary history of regulatory variation in human populations. *Human molecular genetics* 2010, 19(R2):R197-203.

10. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J *et al*: Systematic localization of common disease-associated variation in regulatory DNA. *Science* 2012, 337(6099):1190-1195.

11. Ramos EM, Hoffman D, Junkins HA, Maglott D, Phan L, Sherry ST, Feolo M, Hindorff LA: Phenotype-Genotype Integrator (PheGenI): synthesizing genome-wide association study (GWAS) data with existing genomic resources. *European journal of human genetics : EJHG* 2014, 22(1):144-147.

12. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA: Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* 2009, 106(23):9362-9367.

13. Hartwell LH, Hopfield JJ, Leibler S, Murray AW: From molecular to modular cell biology. *Nature* 1999, 402(6761 Suppl):C47-52.

14. Tompa M, Li N, Bailey TL, Church GM, De Moor B, Eskin E, Favorov AV, Frith MC, Fu Y, Kent WJ *et al*: Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol* 2005, 23(1):137-144.

15. Johnson DS, Mortazavi A, Myers RM, Wold B: Genome-wide mapping of in vivo protein-DNA interactions. *Science* 2007, 316(5830):1497-1502.

16. Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A *et al*: Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nature methods* 2007, 4(8):651-657.

17. Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, Wong E, Orlov YL, Zhang W, Jiang J *et al*: Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* 2008, 133(6):1106-1117.

18. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K: High-resolution profiling of histone methylations in the human genome. *Cell* 2007, 129(4):823-837.

19. Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, Furey TS, Crawford GE: High-resolution mapping and characterization of open chromatin across the genome. *Cell* 2008, 132(2):311-322.

20. Song L, Zhang Z, Grasfeder LL, Boyle AP, Giresi PG, Lee BK, Sheffield NC, Graf S, Huss M, Keefe D *et al*: Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome Res* 2011, 21(10):1757-1767.

21. Crawford GE, Holt IE, Whittle J, Webb BD, Tai D, Davis S, Margulies EH, Chen Y, Bernat JA, Ginsburg D *et al*: Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Res* 2006, 16(1):123-131.

22. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO *et al*: Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 2009, 326(5950):289-293.

23. Belton JM, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J: Hi-C: A comprehensive technique to capture the conformation of genomes. *Methods* 2012.

24. Rozowsky J, Euskirchen G, Auerbach RK, Zhang ZD, Gibson T, Bjornson R, Carriero N, Snyder M, Gerstein MB: PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat Biotechnol* 2009, 27(1):66-75.

25. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W *et al*: Model-based analysis of ChIP-Seq (MACS). *Genome Biol* 2008, 9(9):R137.

26. Whitington T, Frith MC, Johnson J, Bailey TL: Inferring transcription factor complexes from ChIP-seq data. *Nucleic Acids Res* 2011, 39(15):e98.

27. Sun H, Guns T, Fierro AC, Thorrez L, Nijssen S, Marchal K: Unveiling combinatorial regulation through the combination of ChIP information and in silico cis-regulatory module detection. *Nucleic Acids Res* 2012, 40(12):e90.

28. Chen G, Zhou Q: Searching ChIP-seq genomic islands for combinatorial regulatory codes in mouse embryonic stem cells. *BMC genomics* 2011, 12:515.

29. Wingender E, Chen X, Fricke E, Geffers R, Hehl R, Liebich I, Krull M, Matys V, Michael H, Ohnhauser R *et al*: The TRANSFAC system on gene expression regulation. *Nucleic Acids Res* 2001, 29:281-283.

30. Vlieghe D, Sandelin A, De Bleser PJ, Vleminckx K, Wasserman WW, van Roy F, Lenhard B: A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. *Nucleic Acids Res* 2006, 34(Database issue):D95-97.

31. Hu Y, Flockhart I, Vinayagam A, Bergwitz C, Berger B, Perrimon N, Mohr SE: An integrative approach to ortholog prediction for disease-focused and other functional studies. *BMC Bioinformatics* 2011, 12:357.

32. Jeibmann A, Paulus W: Drosophila melanogaster as a model organism of brain diseases. *International journal of molecular sciences* 2009, 10(2):407-440.

33. Chartier A, Benoit B, Simonelig M: A Drosophila model of oculopharyngeal muscular dystrophy reveals intrinsic toxicity of PABPN1. *EMBO J* 2006, 25(10):2253-2262.

34. Mosqueira M, Willmann G, Ruohola-Baker H, Khurana TS: Chronic hypoxia impairs muscle function in the Drosophila model of Duchenne's muscular dystrophy (DMD). *PLoS One* 2010, 5(10):e13450.

35. Ueyama M, Akimoto Y, Ichimiya T, Ueda R, Kawakami H, Aigaki T, Nishihara S: Increased apoptosis of myoblasts in Drosophila model for the Walker-Warburg syndrome. *PLoS One* 2010, 5(7):e11557.

36. Buchon N, Broderick NA, Poidevin M, Pradervand S, Lemaitre B: Drosophila intestinal response to bacterial infection: activation of host defense and stem cell proliferation. *Cell host & microbe* 2009, 5(2):200-211.

37. Castonguay-Vanier J, Vial L, Tremblay J, Deziel E: Drosophila melanogaster as a model host for the Burkholderia cepacia complex. *PLoS One* 2010, 5(7):e11467.

38. Grant J, Saldanha JW, Gould AP: A Drosophila model for primary coenzyme Q deficiency and dietary rescue in the developing nervous system. *Disease models & mechanisms* 2010, 3(11-12):799-806.

39. Melicharek DJ, Ramirez LC, Singh S, Thompson R, Marenda DR: Kismet/CHD7 regulates axon morphology, memory and locomotion in a Drosophila model of CHARGE syndrome. *Human molecular genetics* 2010, 19(21):4253-4264.

40. Yu W, Clyne M, Dolan SM, Yesupriya A, Wulf A, Liu T, Khoury MJ, Gwinn M: GAPscreener: an automatic tool for screening human genetic association literature in PubMed using the support vector machine technique. *BMC Bioinformatics* 2008, 9:205.

41. Nishimura M, Ocorr K, Bodmer R, Cartry J: Drosophila as a model to study cardiac aging. *Experimental gerontology* 2011, 46(5):326-330.

42. Yedvobnick B, Moberg K: Linking model systems to cancer therapeutics: the case of Mastermind. *Disease models & mechanisms* 2010, 3(9-10):540-544.

43. Das T, Cagan R: Drosophila as a novel therapeutic discovery tool for thyroid cancer. *Thyroid : official journal of the American Thyroid Association* 2010, 20(7):689-695.

44. Fraissinet-Tachet L, Baltz R, Chong J, Kauffmann S, Fritig B, Saindrenan P: Two tobacco genes induced by infection, elicitor and salicylic acid encode glucosyltransferases acting on phenylpropanoids and benzoic acid derivatives, including salicylic acid. *FEBS letters* 1998, 437(3):319-323.

45. Heard E, Tishkoff S, Todd JA, Vidal M, Wagner GP, Wang J, Weigel D, Young R: Ten years of genetics and genomics: what have we achieved and where are we heading? *Nat Rev Genet* 2010, 11(10):723-733.

46. Collins F: Has the revolution arrived? *Nature* 2010, 464(7289):674-675.

47. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* 2004, 306(5696):636-640.

48. Celniker SE, Dillon LA, Gerstein MB, Gunsalus KC, Henikoff S, Karpen GH, Kellis M, Lai EC, Lieb JD, MacAlpine DM *et al*: Unlocking the secrets of the genome. *Nature* 2009, 459(7249):927-930.

49. Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, Kellis M, Marra MA, Beaudet AL, Ecker JR *et al*: The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol* 2010, 28(10):1045-1048.

50. Temple G, Gerhard DS, Rasooly R, Feingold EA, Good PJ, Robinson C, Mandich A, Derge JG, Lewis J, Shoaf D *et al*: The completion of the Mammalian Gene Collection (MGC). *Genome Res* 2009, 19(12):2324-2333.

51. Maston GA, Evans SK, Green MR: Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet* 2006, 7:29-59.

52. Narlikar L, Ovcharenko I: Identifying regulatory elements in eukaryotic genomes. *Brief Funct Genomic Proteomic* 2009, 8(4):215-230.

53. Alexander RP, Fang G, Rozowsky J, Snyder M, Gerstein MB: Annotating non-coding regions of the genome. *Nat Rev Genet* 2010, 11(8):559-571.

54. Davidson EH: The Regulatory Genome: Gene Regulatory Networks In Development And Evolution: Academic Press; 2006.

55. Heintzman ND, Ren B: Finding distal regulatory elements in the human genome. *Current opinion in genetics & development* 2009, 19(6):541-549.

56. Hardison RC, Taylor J: Genomic approaches towards finding cis-regulatory modules in animals. *Nat Rev Genet* 2012, 13(7):469-483.

57. Taher L, McGaughey DM, Maragh S, Aneas I, Bessling SL, Miller W, Nobrega MA, McCallion AS, Ovcharenko I: Genome-wide identification of conserved regulatory function in diverged sequences. *Genome Res* 2011, 21(7):1139-1149.

58. Laajala TD, Raghav S, Tuomela S, Lahesmaa R, Aittokallio T, Elo LL: A practical comparison of methods for detecting transcription factor binding sites in ChIP-seq experiments. *BMC genomics* 2009, 10:618.

59. Park PJ: ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet* 2009, 10(10):669-680.

60. Pepke S, Wold B, Mortazavi A: Computation for ChIP-seq and RNA-seq studies. *Nature methods* 2009, 6(11 Suppl):S22-32.

61. Fauteux F, Blanchette M, Stromvik MV: Seeder: discriminative seeding DNA motif discovery. *Bioinformatics* 2008, 24(20):2303-2307.

62. Ettwiller L, Paten B, Ramialison M, Birney E, Wittbrodt J: Trawler: de novo regulatory motif discovery pipeline for chromatin immunoprecipitation. *Nature methods* 2007, 4(7):563-565.

63. Kulakovskiy IV, Boeva VA, Favorov AV, Makeev VJ: Deep and wide digging for binding motifs in ChIP-Seq data. *Bioinformatics* 2010, 26(20):2622-2623.

64. Hu M, Yu J, Taylor JM, Chinnaiyan AM, Qin ZS: On the detection and refinement of transcription factor binding sites using ChIP-Seq data. *Nucleic Acids Res* 2010, 38(7):2154-2167.

65. Mason MJ, Plath K, Zhou Q: Identification of context-dependent motifs by contrasting ChIP binding data. *Bioinformatics* 2010, 26(22):2826-2832.

66. Reid JE, Wernisch L: STEME: efficient EM to find motifs in large data sets. *Nucleic Acids Res* 2011, 39(18):e126.

67. Bailey TL: DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics* 2011, 27(12):1653-1659.

68. Huggins P, Zhong S, Shiff I, Beckerman R, Laptenko O, Prives C, Schulz MH, Simon I, Bar-Joseph Z: DECOD: fast and accurate discriminative DNA motif finding. *Bioinformatics* 2011, 27(17):2361-2367.

69. Thomas-Chollier M, Herrmann C, Defrance M, Sand O, Thieffry D, van Helden J: RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. *Nucleic Acids Res* 2012, 40(4):e31.

70. Ma X, Kulkarni A, Zhang Z, Xuan Z, Serfling R, Zhang MQ: A highly efficient and effective motif discovery method for ChIP-seq/ChIP-chip data using positional information. *Nucleic Acids Res* 2012, 40(7):e50.

71. Negre N, Brown CD, Ma L, Bristow CA, Miller SW, Wagner U, Kheradpour P, Eaton ML, Loriaux P, Sealfon R *et al*: A cis-regulatory map of the Drosophila genome. *Nature* 2011, 471(7339):527-531.

72. Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, Liu T, Yip KY, Robilotto R, Rechtsteiner A, Ikegami K *et al*: Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science* 2010, 330(6012):1775-1787.

73. Zhang Z, Chang CW, Goh WL, Sung WK, Cheung E: CENTDIST: discovery of co-associated factors by motif distribution. *Nucleic Acids Res* 2011, 39(Web Server issue):W391-399.

74. ENCODE: A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol* 2011, 9(4):e1001046.

75. Roy S, Ernst J, Kharchenko PV, Kheradpour P, Negre N, Eaton ML, Landolin JM, Bristow CA, Ma L, Lin MF *et al*: Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Journal Name: Science* 2010:Medium: ED.

76. Wang J, Zhuang J, Iyer S, Lin X, Whitfield TW, Greven MC, Pierce BG, Dong X, Kundaje A, Cheng Y *et al*: Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res* 2012, 22(9):1798-1812.

77. Wang J, Zhuang J, Iyer S, Lin XY, Greven MC, Kim BH, Moore J, Pierce BG, Dong X, Virgil D *et al*: Factorbook.org: a Wiki-based database for transcription factor-binding data generated by the ENCODE consortium. *Nucleic Acids Res* 2013, 41(Database issue):D171-176.

78. Davidson EH, Rast JP, Oliveri P, Ransick A, Calestani C, Yuh CH, Minokawa T, Amore G, Hinman V, Arenas-Mena C *et al*: A genomic regulatory network for development. *Science* 2002, 295(5560):1669-1678.

79. Li XY, MacArthur S, Bourgon R, Nix D, Pollard DA, Iyer VN, Hechmer A, Simirenko L, Stapleton M, Luengo Hendriks CL *et al*: Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm. *PLoS Biol* 2008, 6(2):e27.

80. Gallo SM, Gerrard DT, Miner D, Simich M, Des Soye B, Bergman CM, Halfon MS: REDfly v3.0: toward a comprehensive database of transcriptional regulatory elements in Drosophila. *Nucleic Acids Res* 2011, 39(Database issue):D118-123.

81. Ip YT, Park RE, Kosman D, Yazdanbakhsh K, Levine M: dorsal-twist interactions establish snail expression in the presumptive mesoderm of the Drosophila embryo. *Genes Dev* 1992, 6(8):1518-1530.

82. Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan KK, Cheng C, Mu XJ, Khurana E, Rozowsky J, Alexander R *et al*: Architecture of the human regulatory network derived from ENCODE data. *Nature* 2012, 489(7414):91-100.

83. Machanick P, Bailey TL: MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics* 2011, 27(12):1696-1697.

84. Mathelier A, Wasserman WW: The next generation of transcription factor binding site prediction. *PLoS Comput Biol* 2013, 9(9):e1003214.

85. Tran NT, Huang CH: A survey of motif finding Web tools for detecting binding site motifs in ChIP-Seq data. *Biology direct* 2014, 9(1):4.

86. Bolouri H, Ruzzo WL: Integration of 198 ChIP-seq datasets reveals human cis-regulatory regions. *J Comput Biol* 2012, 19(9):989-997.

87. van Dongen S: A cluster algorithm for graphs. Amsterdam: National Research Institute for Mathematics and Computer Science in the Netherlands; 2000.

88. Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS: Quantifying similarity between motifs. *Genome Biol* 2007, 8(2):R24.

89. Bergman CM, Carlson JW, Celniker SE: Drosophila DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, Drosophila melanogaster. *Bioinformatics* 2005, 21(8):1747-1749.

90. Zhu LJ, Christensen RG, Kazemian M, Hull CJ, Enuameh MS, Basciotta MD, Brasefield JA, Zhu C, Asriyan Y, Lapointe DS *et al*: FlyFactorSurvey: a database of Drosophila transcription factor binding specificities determined using the bacterial one-hybrid system. *Nucleic Acids Res* 2011, 39(Database issue):D111-117.

91. Kulakovskiy IV, Makeev VJ: Discovery of DNA motifs recognized by transcription factors through integration of different experimental sources. *BIOPHYSICS* 2009, 54(6):667-674.

92. Brand AH, van Roessel PJ: Region-specific apoptosis limits neural stem cell proliferation. *Neuron* 2003, 37(2):185-187.

93. Thomas JB, Crews ST, Goodman CS: Molecular genetics of the single-minded locus: a gene involved in the development of the Drosophila nervous system. *Cell* 1988, 52(1):133-141.

94. Sanyal S, Narayanan R, Consoulas C, Ramaswami M: Evidence for cell autonomous AP1 function in regulation of Drosophila motor-neuron plasticity. *BMC neuroscience* 2003, 4:20.

95. De Graeve F, Jagla T, Daponte JP, Rickert C, Dastugue B, Urban J, Jagla K: The ladybird homeobox genes are essential for the specification of a subpopulation of neural cells. *Dev Biol* 2004, 270(1):122-134.

96. Bates KE, Sung CS, Robinow S: The unfulfilled gene is required for the development of mushroom body neuropil in Drosophila. *Neural Dev* 2010, 5:4.

97. Tanaka KK, Bryantsev AL, Cripps RM: Myocyte enhancer factor 2 and chorion factor 2 collaborate in activation of the myogenic program in Drosophila. *Molecular and cellular biology* 2008, 28(5):1616-1629.

98. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S *et al*: Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 2005, 15(8):1034-1050.

99. 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.

100. 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.

101. Andolfatto P: Adaptive evolution of non-coding DNA in Drosophila. *Nature* 2005, 437(7062):1149-1152.

102. Casillas S, Barbadilla A, Bergman CM: Purifying selection maintains highly conserved noncoding sequences in Drosophila. *Mol Biol Evol* 2007, 24(10):2222-2234.

103. Bergman CM, Kreitman M: Analysis of conserved noncoding DNA in Drosophila reveals similar constraints in intergenic and intronic sequences. *Genome Res* 2001, 11(8):1335-1345.

104. Singh ND, Arndt PF, Clark AG, Aquadro CF: Strong evidence for lineage and sequence specificity of substitution rates and patterns in Drosophila. *Mol Biol Evol* 2009, 26(7):1591-1605.

105. Kondrashov AS: Evolutionary biology: fruitfly genome is not junk. *Nature* 2005, 437(7062):1106.

106. Dennis G, Jr., Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA: DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol* 2003, 4(5):P3.

107. Ciglar L, Furlong EE: Conservation and divergence in developmental networks: a view from Drosophila myogenesis. *Current opinion in cell biology* 2009, 21(6):754-760.

108. Zeitlinger J, Stark A: Developmental gene regulation in the era of genomics. *Dev Biol* 2010, 339(2):230-239.

109. Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T, Smith SF, North P, Callaway H, Kelly K *et al*: Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol* 2005, 3(1):e7.

110. Wray GA: The evolutionary significance of cis-regulatory mutations. *Nat Rev Genet* 2007, 8(3):206-216.

111. Zhang Z, Pugh BF: High-resolution genome-wide mapping of the primary structure of chromatin. *Cell* 2011, 144(2):175-186.

112. Moses AM, Pollard DA, Nix DA, Iyer VN, Li XY, Biggin MD, Eisen MB: Large-scale turnover of functional transcription factor binding sites in Drosophila. *PLoS Comput Biol* 2006, 2(10):e130.

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

114. Sandelin A, Wasserman WW: Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. *Journal of molecular biology* 2004, 338(2):207-215.

115. Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M, Taipale M, Wei G *et al*: DNA-binding specificities of human transcription factors. *Cell* 2013, 152(1-2):327-339.

116. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M *et al*: Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* 2011, 473(7345):43-49.

117. Ram O, Goren A, Amit I, Shoresh N, Yosef N, Ernst J, Kellis M, Gymrek M, Issner R, Coyne M *et al*: Combinatorial patterning of chromatin regulators uncovered by genome-wide location analysis in human cells. *Cell* 2011, 147(7):1628-1639.

118. Zhou VW, Goren A, Bernstein BE: Charting histone modifications and the functional organization of mammalian genomes. *Nat Rev Genet* 2011, 12(1):7-18.

119. Zhu J, Adli M, Zou JY, Verstappen G, Coyne M, Zhang X, Durham T, Miri M, Deshpande V, De Jager PL *et al*: Genome-wide chromatin state transitions associated with developmental and environmental cues. *Cell* 2013, 152(3):642-654.

120. Jiang C, Pugh BF: Nucleosome positioning and gene regulation: advances through genomics. *Nat Rev Genet* 2009, 10(3):161-172.

121. Ioshikhes I, Hosid S, Pugh BF: Variety of genomic DNA patterns for nucleosome positioning. *Genome Res* 2011, 21(11):1863-1871.

122. Roy S, Ernst J, Kharchenko PV, Kheradpour P, Negre N, Eaton ML, Landolin JM, Bristow CA, Ma L, Lin MF *et al*: Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science* 2010, 330(6012):1787-1797.

123. Zhang S, Xu M, Li S, Su Z: Genome-wide de novo prediction of cis-regulatory binding sites in prokaryotes. *Nucleic Acids Res* 2009, 37(10):e72.

124. Zhang S, Li S, Pham PT, Su Z: Simultaneous prediction of transcription factor binding sites in a group of prokaryotic genomes. *BMC Bioinformatics* 2010, 11:397.

125. Zhang S, Jiang L, Du C, Su Z: A novel information content-based similarity metric for comparing transcription factor binding site motifs. *IEEE 6th International Conference on Systems Biology (ISB)* 2012:32-36.

126. van Dongen S, Abreu-Goodger C: Using MCL to extract clusters from networks. *Methods Mol Biol* 2012, 804:281-295.

127. Vlasblom J, Wodak SJ: Markov clustering versus affinity propagation for the partitioning of protein interaction graphs. *BMC Bioinformatics* 2009, 10:99.

128. Brohee S, van Helden J: Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics* 2006, 7:488.

129. Samuel Lattimore B, van Dongen S, Crabbe MJ: GeneMCL in microarray analysis. *Comput Biol Chem* 2005, 29(5):354-359.

130. Enright AJ, Van Dongen S, Ouzounis CA: An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* 2002, 30(7):1575-1584.

131. Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE *et al*: Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 2007, 447(7146):799-816.

132. Demuth JP, De Bie T, Stajich JE, Cristianini N, Hahn MW: The evolution of mammalian gene families. *PLoS One* 2006, 1:e85.

133. Washietl S, Kellis M, Garber M: Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals. *Genome Res* 2014, 24(4):616-628.

134. Rosenbloom KR, Sloan CA, Malladi VS, Dreszer TR, Learned K, Kirkup VM, Wong MC, Maddren M, Fang R, Heitner SG *et al*: ENCODE data in the UCSC Genome Browser: year 5 update. *Nucleic Acids Res* 2013, 41(Database issue):D56-63.

135. Visel A, Minovitsky S, Dubchak I, Pennacchio LA: VISTA Enhancer Browser--a database of tissue-specific human enhancers. *Nucleic Acids Res* 2007, 35(Database issue):D88-92.

136. Tryka KA, Hao L, Sturcke A, Jin Y, Wang ZY, Ziyabari L, Lee M, Popova N, Sharopova N, Kimura M *et al*: NCBI's Database of Genotypes and Phenotypes: dbGaP. *Nucleic Acids Res* 2014, 42(Database issue):D975-979.

137. Dongen Sv: Graph Clustering by Flow Simulation. *PhD thesis, University of Utrecht* 2000.

138. Hou C, Dale R, Dean A: Cell type specificity of chromatin organization mediated by CTCF and cohesin. *Proc Natl Acad Sci U S A* 2010, 107(8):3651-3656.

139. Frietze S, O'Geen H, Blahnik KR, Jin VX, Farnham PJ: ZNF274 recruits the histone methyltransferase SETDB1 to the 3' ends of ZNF genes. *PLoS One* 2010, 5(12):e15082.

140. Mathelier A, Zhao X, Zhang AW, Parcy F, Worsley-Hunt R, Arenillas DJ, Buchman S, Chen CY, Chou A, Ienasescu H *et al*: JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res* 2014, 42(Database issue):D142-147.

141. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S: Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol* 2010, 6(12):e1001025.

142. Martin D, Pantoja C, Fernandez Minan A, Valdes-Quezada C, Molto E, Matesanz F, Bogdanovic O, de la Calle-Mustienes E, Dominguez O, Taher L *et al*: Genome-wide CTCF distribution in vertebrates defines equivalent sites that aid the identification of disease-associated genes. *Nature structural & molecular biology* 2011, 18(6):708-714.

143. Chaumeil J, Skok JA: The role of CTCF in regulating V(D)J recombination. *Current opinion in immunology* 2012, 24(2):153-159.

144. Phillips JE, Corces VG: CTCF: master weaver of the genome. *Cell* 2009, 137(7):1194-1211.

145. Chong JA, Tapia-Ramirez J, Kim S, Toledo-Aral JJ, Zheng Y, Boutros MC, Altshuller YM, Frohman MA, Kraner SD, Mandel G: REST: a mammalian silencer protein that restricts sodium channel gene expression to neurons. *Cell* 1995, 80(6):949-957.

146. Wang H, Maurano MT, Qu H, Varley KE, Gertz J, Pauli F, Lee K, Canfield T, Weaver M, Sandstrom R *et al*: Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome Res* 2012, 22(9):1680-1688.

147. Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM: A census of human transcription factors: function, expression and evolution. *Nat Rev Genet* 2009, 10(4):252-263.

148. Wilson D, Charoensawan V, Kummerfeld SK, Teichmann SA: DBD--taxonomically broad transcription factor predictions: new content and functionality. *Nucleic Acids Res* 2008, 36(Database issue):D88-92.

149. Shubin N, Tabin C, Carroll S: Deep homology and the origins of evolutionary novelty. *Nature* 2009, 457(7231):818-823.

150. Fukushige T, Brodigan TM, Schriefer LA, Waterston RH, Krause M: Defining the transcriptional redundancy of early bodywall muscle development in C. elegans: evidence for a unified theory of animal muscle development. *Genes Dev* 2006, 20(24):3395-3406.

151. Avidor-Reiss T, Maer AM, Koundakjian E, Polyanovsky A, Keil T, Subramaniam S, Zuker CS: Decoding cilia function: defining specialized genes required for compartmentalized cilia biogenesis. *Cell* 2004, 117(4):527-539.

152. Brooks AN, Yang L, Duff MO, Hansen KD, Park JW, Dudoit S, Brenner SE, Graveley BR: Conservation of an RNA regulatory map between Drosophila and mammals. *Genome Res* 2011, 21(2):193-202.

153. Reiter LT, Potocki L, Chien S, Gribskov M, Bier E: A systematic analysis of human disease-associated gene sequences in Drosophila melanogaster. *Genome Res* 2001, 11(6):1114-1125.

154. Lloyd TE, Taylor JP: Flightless flies: Drosophila models of neuromuscular disease. *Annals of the New York Academy of Sciences* 2010, 1184:e1-20.

155. Pandey UB, Nichols CD: Human disease models in Drosophila melanogaster and the role of the fly in therapeutic drug discovery. *Pharmacological reviews* 2011, 63(2):411-436.

156. Chen H, Rossier C, Antonarakis SE: Cloning of a human homolog of the Drosophila enhancer of zeste gene (EZH2) that maps to chromosome 21q22.2. *Genomics* 1996, 38(1):30-37.

157. Bargiela A, Llamusi B, Cerro-Herreros E, Artero R: Two enhancers control transcription of Drosophila muscleblind in the embryonic somatic musculature and in the central nervous system. *PLoS One* 2014, 9(3):e93125.

158. Stifani S, Blaumueller CM, Redhead NJ, Hill RE, Artavanis-Tsakonas S: Human homologs of a Drosophila Enhancer of split gene product define a novel family of nuclear proteins. *Nat Genet* 1992, 2(4):343.

159. Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O, Farrell CM, Hart J, Landrum MJ, McGarvey KM *et al*: RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res* 2014, 42(Database issue):D756-763.

APPENDIX A: LINK OF SUPPLEMENTARY DATA FILES

The supplementary files can be downloaded from
http://bioinfo.uncc.edu/mniu/dissertation_supplementary_files/