BMC Genomics

METHODOLOGY ARTICLE Open Access

De novo prediction of *cis*-regulatory elements and modules through integrative analysis of a large number of ChIP datasets

Meng Niu, Ehsan S Tabari and Zhengchang Su*

Abstract

Background: In eukaryotes, transcriptional regulation is usually 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 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 these large volumes of ChIP data to identify CREs and CRMs at nucleotide resolution is a highly challenging task.

Results: We have developed a novel graph-theoretic based algorithm DePCRM for genome-wide *de novo* predictions of CREs and CRMs using a large number of ChIP datasets. DePCRM predicts CREs and 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 and 89.3% of known CRMs containing at least one predicted CRE. We found that the putative CRMs as well as CREs as a whole in a CRM are more conserved than randomly selected sequences.

Conclusion: 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 larger number of transcription factor ChIP datasets. The algorithm and predictions will hopefully facilitate the elucidation of gene regulatory networks in eukaryotes. All the predicted CREs, CRMs, and their target genes are available at http://bioinfo.uncc.edu/mniu/pcrms/www/.

Keywords: *cis*-regulatory elements, *cis*-regulatory modules, ChIP-chip, ChIP-seq, *Drosophila melanogaster*

Background

Since the completion of sequencing the first metazoan genomes in 1998 [1], 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 [3]. 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 [3,4]. With this recognition, the community has taken a more realistic approach by first identifying all the functional sequence elements in the genomes [5-7]. These functional elements include transcribed sequences as well as transcriptional control elements, epigenetic features, and

* Correspondence: zcsu@uncc.edu
Department of Bioinformatics and Genomics, College of Computing and Informatics, The University of North Carolina at Charlotte, 9201 University City Blvd, Charlotte, NC 28223, USA

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 [8]. 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 [9]. 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 [10-12]. More specifically, promoters, enhancers and silencers are clusters of closely located *cis*-regulatory elements (CREs) that are recognized by specific transcription factors (TFs) [13]. 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) [13]. 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 [15], 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 [16,17].

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 [18-20] and various chromatin modification marks [21], DNase-seq [22-24] and FAIRE-seq [23] 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 [25,26]. While a single epigenetic dataset derived from DNase-seq [22-24], FAIRE-seq [23] or enhancer mark ChIP-seq potentially contains location information of all CRMs active in a cell or tissue type, CREs and CRMs for specific TFs cannot be easily identified in such a dataset, as it lumps all CREs and CRMs active in the cell or tissue type. In contrast, a TF ChIP-seq dataset is highly enriched for the CREs of the TF, thus they can be potentially identified at single nucleotide resolution in a cell or tissue or type. However, the sequenced potential binding regions in a TF ChIP-seq dataset can be still 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) [27], they are still much longer than the typical lengths of CREs, which are typically 6 ~ 16 bp long. Hence, the actual locations of CREs need to be identified by a motif-finding tool [28,29]. Although a few new motif-finders have been developed to analyze large sequence sets from ChIP-seq experiments, such as seeder [30], Trawler [30,31], ChIPMunk [32], HMS [33], CMF [34], STEME [35], DREME [36], DECOD [37], RSAT [38], and POSMO [39], they are typically used to find the CREs of a ChIP-ed TF in a short region of sequences (~200 bp) around the binding peak summits in order to reduce the searching space and increase prediction specificity in trading of sensitivity. Some of these tools [33,39] 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 [13]. 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 [40-44]. 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 [5,45] and modENCODE [6] projects aimed at identifying all the functional sequence elements in the genomes of humans and the model organisms *C. elegans* [43] and *D. melanogaster* [42,46]. 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 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 [10], the increasing number of ChIP-seq datasets contains a wealth of information about the combinatorial patterns of different TFs for transcriptional regulation [42,43]. Thus, it is now possible to predict the CREs and CRMs 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 [40], CPModule [41] and [47], 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 [48] or JASPAR [49] 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 [42] and ENCODE [50,51] datasets, only the top 250 and 500 binding peaks with a length of 100 bp and 200 bp, 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* [p]rediction of [CRMs] and CREs 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.

Results

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 ~ 3,000 bp [52], 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., 3,000 bp), then extended binding peaks are more likely to contain the CREs of different cooperative TFs (TFs that co-act in a CRM) in addition to the CREs of the ChIP-ed TF as illustrated in Figure 1. In other words, if two different TFs (e.g. the red circle and black circle TFs in Figure 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.

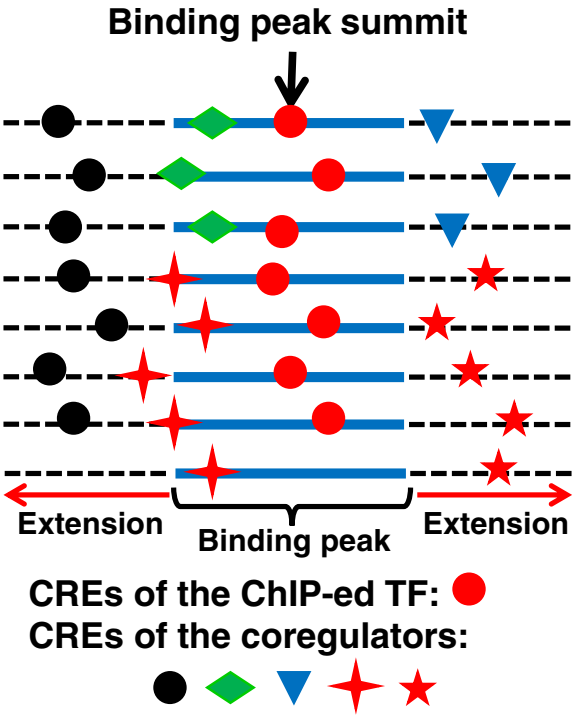

Figure 1 A schematic view of our hypothesis. If the binding peak is shorter than 3,000 bp, we equally extended from the two ends to have a length up to 3,000 bp. We assume that in addition to the CREs of the ChIP-ed TF (red circle), CREs of different cooperative TFs (the other shapes) are also enriched 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 (Figure 2A and 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) (Figure 2B and 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

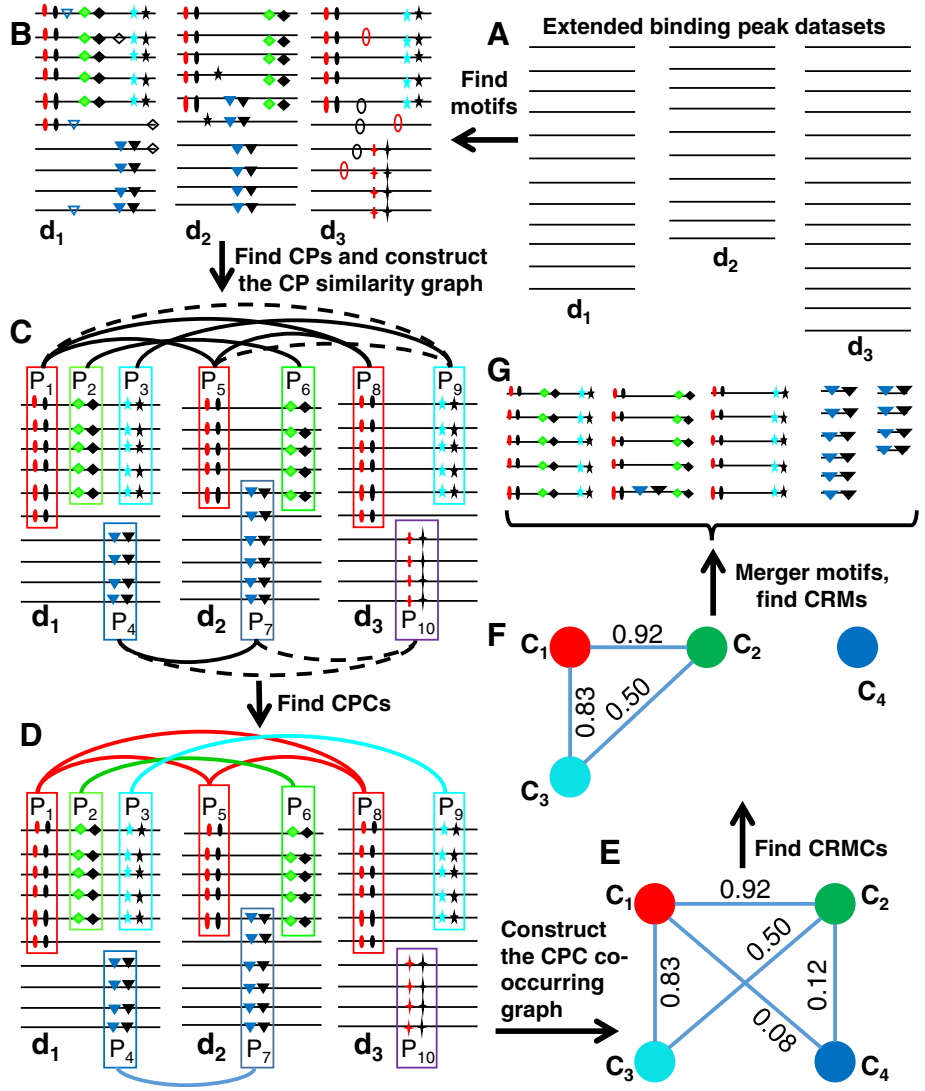

Figure 2 A schematic of the major steps of the DePCRM algorithm. A. Illustration of extended binding peaks from dataset d_1, d_2 and d_3 respectively. **B**. Illustration of CREs found within each dataset, CREs of the same motif are shown in the same shape and color. **C**. Construction of CP similarity graph. $\{P_1, P_2, P_3, P_4\}$, $\{P_5, P_6, P_7\}$ and $\{P_8, P_9, P_{10}\}$ are sets of CPs found in datasets d_1, d_2 and d_3 respectively. For clarity, the CPs formed between motifs in P_1 and motifs in P_2 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 $S_s \geq \beta$, with S_s 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): $C_1 = \{P_1, P_5, P_8\}$; $C_2 = \{P_2, P_6\}$, $C_3 = \{P_3, P_9\}$, $C_4 = \{P_4, P_7\}$ and $C_5 = \{P_{10}\}$. CPs in a cluster are connected by edges in the same color. The singleton cluster $C_5 = \{P_{10}\}$ is discarded for its low density. **E**. For each pair C_i and C_j from the four CPCs, we find sets of CPs from the same dataset d_k, and compute a co-occurring scores S_{CPC} (C_i, C_j) for the two CPCs. **F**. Construction of the CPC co-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.

CPs in multiple datasets, and call them *CP clusters* (CPCs) (Figure 2D). 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 2E). 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 2F). Finally, we predict individual CRMs across the genome based on the motif pattern of the CRMCs and their close adjacency (Figure 2G). 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 complete the datasets, the more accurate the predictions will be. The details of the algorithm are described in Methods.

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 modENCODE project [42,46], 38 Chip-chip datasets were from the Berkeley Drosophila Transcription Network Project (BDTNP) [53], and 46 ChIP-chip datasets were from literature. Additional file 1: Table S1 summarizes the major features of the 168 datasets. As shown in Figure 3A, the majority of the binding peaks have a length around 1,000 bp, and only 0.62% of them have a length longer than 5,000 bp, 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). The datasets contain a total of 445,252 sequences, each individual dataset containing 26 to 11,772 sequences (Additional file 2: Figure S8). These 445,252 sequences contain a total of 1,183,049,646 bp, which are 7.0 times of the genome (168,736,537 bp), but only cover 45.4% (76,555,033 bp) of the genome (Additional file 3: Table S2), indicating that some of these sequences highly overlap with one another, thus confirming our aforementioned assumption. Of the 76,555,033 bp genome sequence covered by the datasets, 64,033,300 bp (86.3%) are in non-coding regions (NCRs, including introns and intergenic sequences), consisting of 47.7% of NCRs (134,207,178 bp) in the genome (Figure 4 and Additional file 3: Table S2). The remaining 12,521,733 (16.4%) sequences are in coding regions (CDRs), consisting of 36.3% of CDRs (34,529,359 bp) in the genome (Additional file 3: Table S2 and Figure 4). 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 [54], and 1,330 (72.7%) of which are located in the extended binding peaks. We will evaluate our algorithm for its ability to recover these 1,330 known CRMs in the extended binding peaks.

To see the overlapping patterns of binding peaks upon which our algorithm is based, we computed pair-wise overlapping scores (formula (1) in Methods) of the extended binding peaks among the 168 datasets for the 56 TFs (Additional file 1: Table S1), and clustered the datasets using the overlapping scores. As shown in Figure 5A, consistent with the above analysis, there are significant

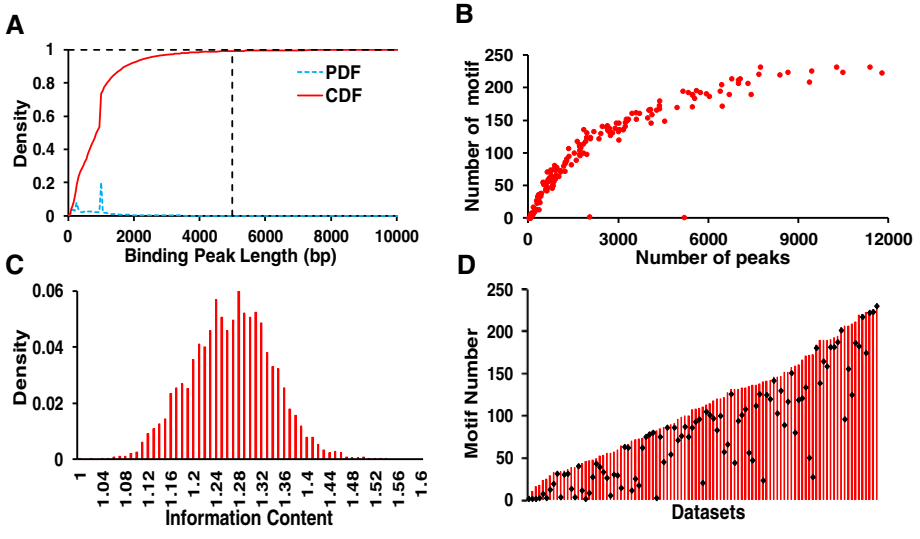

Figure 3 Analysis of the original datasets and motif finding results. A. Distribution of the binding peak lengths in the 168 original datasets. Vast majority (99.38%) of the binding peaks are shorter than 5,000 bp. **B**. Number of motifs found in each of the 168 datasets as a function of the number of binding peaks the datasets. **C**. Distribution of the information content of the predicted motifs in the datasets. **D**. 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.

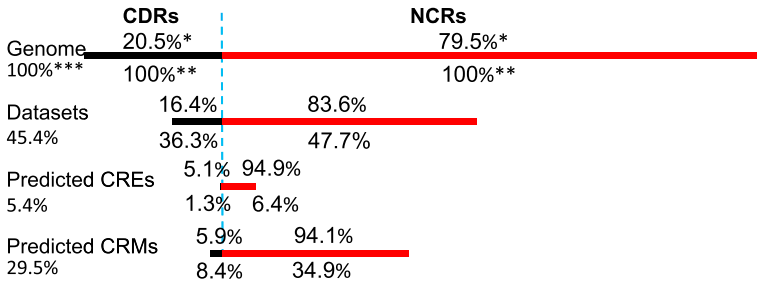

Figure 4 Coverage of the datasets, predicted CREs and CRMs on the genome and its CDRs and NCRs. *, the numbers above a line (sequence category) are the percentages of the CDRs and NCRs in the category. **, the numbers below a line are the percentages of CDRs and NCRs of the category with respect to the entire CDRs and NCRs in the genome. ***, the number on the right of a line is the percentage of the category with respect to the genome.

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 5A show two examples of such clusters. The upper cluster is

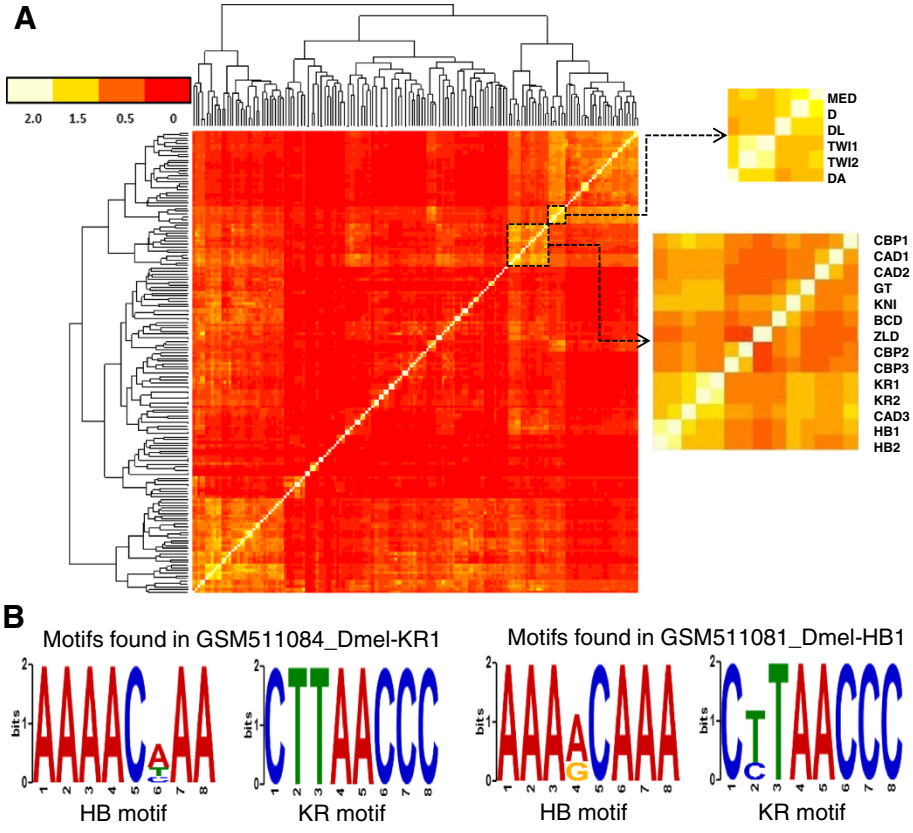

Figure 5 Overlapping analysis of the datasets. A. Hierarchical clustering of the 168 datasets for 56 TFs based on their pair-wise binding peak overlapping scores S_o. The blow-ups show two clusters for cooperative TFs (see Results). **B**. The motifs of TFs KR and HB are both found in the overlapping datasets GSM511084_Dmel-KR1 ChIP-ed by KR and GSM511081_Dmel-HB1 ChIP-ed by HB.

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 [55]. 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* [54]. An example of such CRMs is shown in Additional file 4: Figure S1. 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 (Additional file 5: Figure S2). The similar results also were reported in *D. melanogaster* [42] and human [56] 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.

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, 2A and B). Because accurate motif-finding is still a notoriously difficult problem [14,57-59], to achieve this goal we consider all overrepresented motifs returned by DREME [36] 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 ($0 \sim 231$) 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 3C). Importantly, the known motifs of the ChIP-ed TFs were found by DREME for 99 of the 162 datasets, and were generally ranked high by the program, although they were usually not the top hit of DREME (Figure 3D), 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 is significantly overlap 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 5B). Overall, we identified a total of 17,890 putative motifs corresponding to 35,359,819 putative CREs in the 162 datasets. These 35,359,819 putative CREs contain 275,857,398 bp which are 1.6 times of the genome, but only cover 30.9% (52,078,901 bp) 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). 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

Table 1 Summary of the predictions of CREs and 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 findings	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.

is in agreement with an early study based on 38 ChIP datasets in *D. melanogaster* [42], and also is supported by two recent studies using human datasets [56,60].

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 (2)) for each pair of putative motifs found in each dataset (Figures 2C). As shown in Figure 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 5A and B, 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) for further analysis, which are presumably enriched for true motif pairs and motifs.

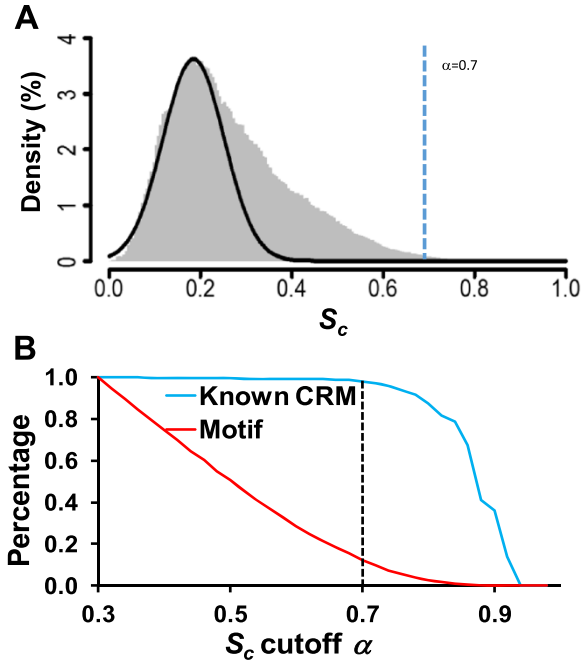

Figure 6 Setting S_C cutoff α. A. Distribution of co-occurring scores S_c of the motif pairs found in the 16 datasets. The curve is a fitting of the left portion of the distribution to a Gaussian distribution $N(\mu = 0.19, \sigma = 0.067)$. **B**. The remaining proportions of predicted motifs and known CRMs as functions of the S_c cutoff α. The vertical line indicates the position of the chosen cutoff α = 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 (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 2C). As shown in Figure 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 7B). Thus, we set β =1.36 to construct the CP similarity graph (Methods). Applying the Markov chain clustering (MCL) algorithm [61] to the graph (Figure 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). Thus we further filtered out 2,049 (2,049/4,891 = 41.9%) CPs and 213 (213/1,589 = 13.4%) putative motifs.

Next, to identify larger repeatedly used motif patterns, we computed a co-occurring score S_{CPC} (formula (5)) for each pair of CPCs across the datasets in which both the CPCs have motifs. Interestingly, as shown in Figure 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 (Figure 2D and E, 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). The compositions and structures of these 815 CRMCs are shown in Additional file 6: Figure S3, each containing 1 ~ 9 CPCs. Overall, 16,574 (92.6%) of

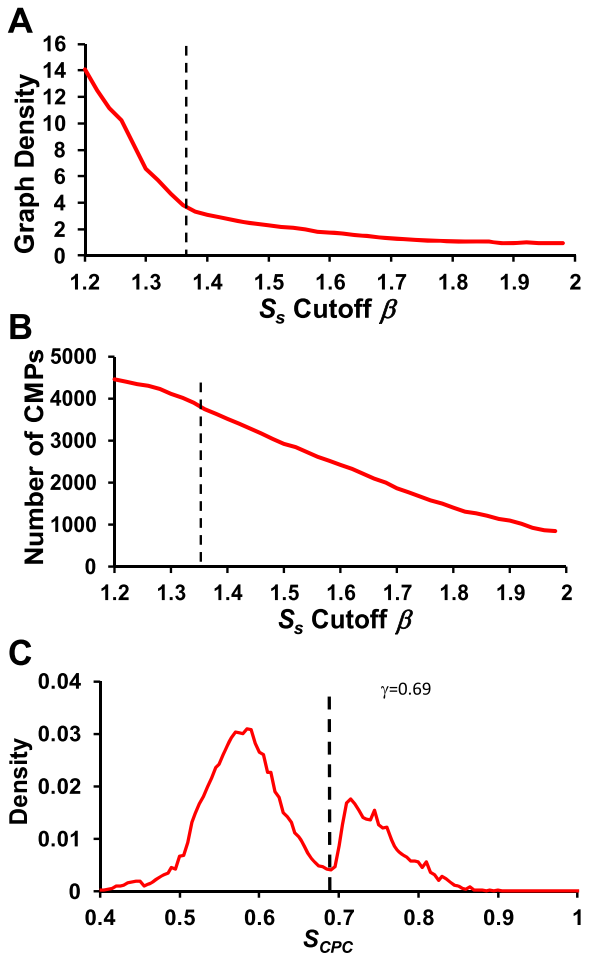

Figure 7 Setting S_S cutoff β and S_{CPC} cutoff γ. A. The density of the CP similarity graph drops rapidly with the increase in the S_s cutoff β, but the trend of decrease slows down around β =1.36. **B**. The number of CRM in the graph also starts to drop rapidly around β = 1.36. Thus we set β =1.36 for construing the final CP similarity graph. **C**. The distribution of CPC co-occurring scores 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.

the original 17,890 input motifs were filtered out by the algorithm (Table 1), suggesting that at least the vast majority (92.6%) of the putative motifs found in the datasets are spurious predictions.

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 5B. 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 (Methods), 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 (Additional file 7: Figure S4, Additional file 8: Table S3). When compared with the known motifs in multiple built-in databases including DMMPMM, iDMMPMM, flyreg and fly factor survey using TOMTOM [62-65], 111 (60.3%) of the Umotifs are highly similar to known motifs in *D. melanogaster* at $p < 0.001$ (Additional file 8: Table S3), strongly suggesting that they are likely to be true motifs. As shown in Additional file 9: Figure S9, a p-value cutoff of 0.001 is sufficient to identify highly similar motifs. More examples of such Umotifs, their constituent motifs and the known motifs hit are shown in Additional file 10: Figures S5A and 5B. 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 Additional file 10: Figures S5C and 5D. Furthermore, 106 (29.4%), 203 (56.2%) and 269 (74.5%) of 381 possibly redundant motifs found in the earlier study [42] 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.

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 ± 2.8 bp, with 53,785 (4.9%) of which being entirely located in CDRs. These 1,108,018 CREs cover 9,045,115 bp (5.4%) genome sequence, of which 8,583,816 bp (94.9%) are in NCRs, consisting of 6.4% of NCRs; the remaining 461,299 bp (5.1%) are in CDRs, consisting of 1.3% of CDRs (Figure 4 and Additional file 3: Table S2). 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 44115 (38.1%) contain CDRs. These 115,932 CRMs cover 49,796,159 bp (29.5%) genome sequence, 46,880,944 bp (94.1%) of which are in NCRs, consisting 34.9% of NCRs; the remaining 2,925,215 bp (5.9%) are in CDRs, consisting of 8.4% of CDRs (Figure 4 and Additional file 3: Table S2). These putative CRMs tend to have shorter lengths than those of the known CRMs (Figure 8A). 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 tends to have a short distance between adjacent two putative CREs (Figure 8B). 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

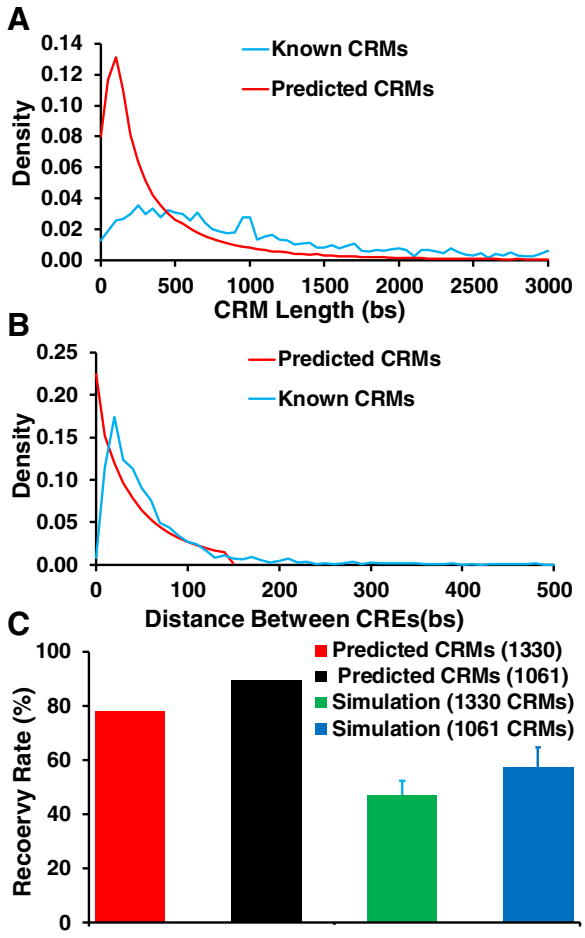

Figure 8 Summary of the predicted CRMs. A. Distribution of the lengths of the known and predicted CRMs. **B**. Distribution of the distances (bp) between two adjacent CREs in the known and predicted CRMs. **C**. Recovery rates of the known CRMs in the datasets (1330) and the known CRMs containing a predicted CRE (1061) by the predicted CRMs and the corresponding same number and length sequences randomly selected from NCRs.

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.

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). 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 8C) 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 8C) 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 number of datasets for only 56 TFs. Importantly, some of 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 (Additional file 11: Figure S6A). Furthermore, CRM (2 L: 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 (Additional file 11: Figure S6B and S6C). 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 [66-68], 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 (Additional file 11: Figures S6D and S6E), and it has been reported that HR51 and LB regulate neurogenesis [69,70]. 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* (Additional file 12: Figure S7) 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 (Additional file 11: Figures S6F and S6G). It has been shown that these two TFs cooperatively regulate Actin57B by binding to its promoter region [71]. Thus MET2 and CF2 might also regulate *actin57B* through binding to the putative CRES of Umotifs 27 and 23 located in its first intron (Additional file 12: Figure S7). Therefore, our predicted CREs and CRMs can help biologists identify potential enhancers for genes of interest.

The predicted CRMs as well as CREs in a CRM as a whole are more conserved than randomly selected sequences

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 [72] 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 [72]. As shown in Figure 9A, 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 ($p < 2.2\times10^{-302}$, Kolmogorov–Smirnov test). Specifically, the right peak with very low phastCons scores, which reflects highly conserved sequences [72] 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 non-conserved sequences [72]. Moreover, the middle peak with intermediate phastCons scores, which reflects neutral to moderately conserved sequences [72], 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 9B, average phastCons scores of CREs in a CRM

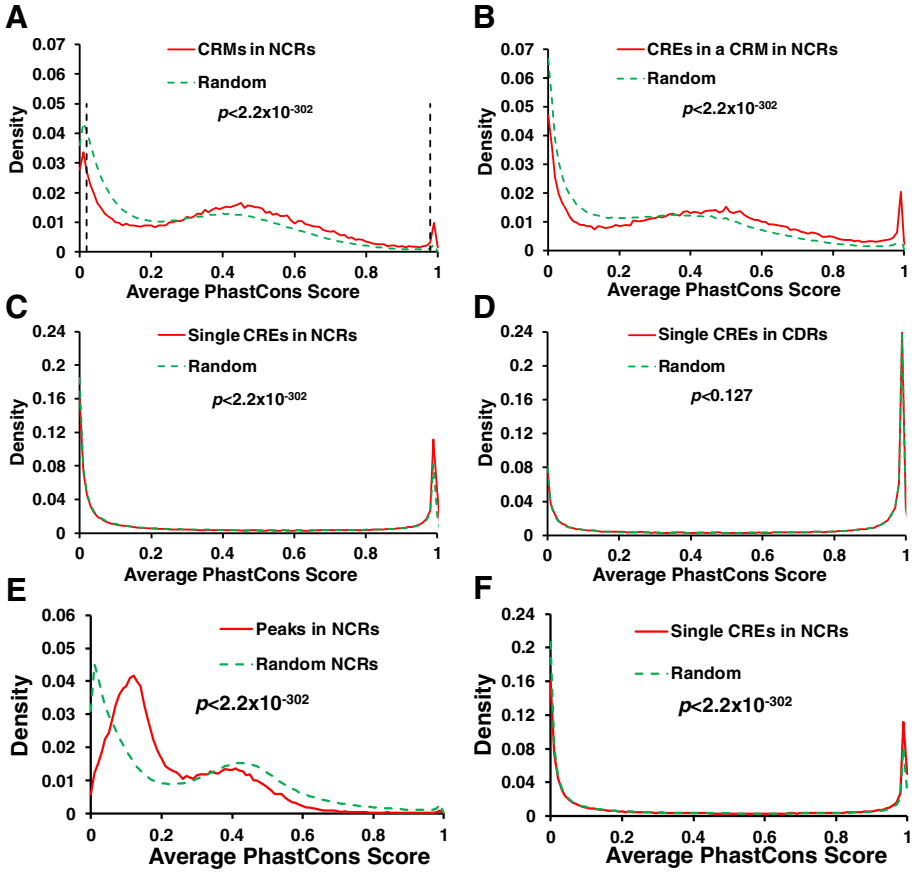

Figure 9 Conservation analysis of the CRMs. A. Distribution of average phastCons scores of the predicted CRMs in NCRs and of the same number and length sequences randomly selected from NCRs. The vertical dashed lines indicate the PhastCons score cutoffs for highly conserved (≥0.98) and non-conserved (≤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.

and randomly selected sequences from NCRs also show tri-modal distributions, however again, they are significantly different (p <2.2×10^{-302}, Kolmogorov–Smirnov test) in the similar way as for those of the full length putative CRMs and the corresponding randomly selected sequences (Figure 9A). 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 9A), the right peak for the putative CREs is much larger than that of the randomly selected sequences (Figure 9B), and the middle peak for the putative CREs shift more (0.15 vs. 0.04 unit) to right relative to that of the randomly selected sequences (Figure 9B). 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 single CREs in the 71,817 predicted CRMs in CDRs and in the 53,785 predicted CRMs in CDRs with those of the same number and length sequences randomly selected from NCRs and CDRs, respectively. As shown in Figure 9C and D, 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 [73-79]. However, interestingly, there are striking differences between the predicted CREs in NCRs (Figure 9C) and those in CDRs (Figure 9D). First, the distribution for single putative CREs in NCRs is significantly different from that for the corresponding randomly selected sequences ($p < 2.2\times10^{-302}$, Kolmogorov–Smirnov test), as the right peak of the former is slightly larger than that of the latter (Figure 9C), 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 9D), 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 9C), 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 9D), 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 9C and D), 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 9E, 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 (Figure 9F). 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.

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 9A) 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 Additional file 13: Table S4, 134 (22.3%) the putative target genes of the 601 highly conserved putative CRMs are clustered into 11 functional groups using the DAVID program [80] with an enrichment score ≥1.5 and $p < 0.01$ (hypergeometric 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) (Additional file 14: Table S5). 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 [81,82] and vertebrates [83], while CRMs for genes with other functions in particular those related to environmental adaptations evolve extremely fast [84], 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 PCRMs (http://bioinfo.uncc.edu/mniu/pcrms/www/) for public use. The query results and relevant knowledge are displayed using the NCBI graphical sequence viewer. Currently, the algorithm was implemented in Perl, and the scripts are available on the PCRMs website.

Discussion

The recent development of various ChIP-seq, DNase-seq and FAIRE-seq techniques for locating bind regions of specific TFs, chromatin marks, free nucleosome regions, has provided an unprecedented opportunity for deciphering all *cis*-regulatory sequences in eukaryotic genomes. These techniques and resulting datasets reveal similar or quite different aspects of *cis*-regulatory sequences, and have their pros and cons. On one hand, a single epigenomic dataset resulted from DNase-seq, FAIRE-seq, or enhancer mark such as H3K27ac ChIP-seq provide information of the locations of all functional CRMs in a cell or tissue type, thus these techniques could be less expensive. However, it is very difficult to predict novel CREs of specific TFs from such an epigenetic dataset, since it lumps the potential CREs for all TFs active in the cell or tissue type, and TF information is usually unavailable. Due to the lack of CRE locations, it is also difficult to predict CRMs at single nucleotide resolution using epigenetic datasets. On the other hand, a TF ChIP-seq dataset is highly enriched for the CREs of the ChIP-ed TF and of its co-operators, thus all these CREs can be potentially identified at single nucleotide resolution in a cell or tissue type. However, as a TF ChIP-seq dataset only contains location information of CREs of the ChIP-ed TF, a certain number of TFs that are potentially active in the cell or tissue type need to be analyzed to identify all the CREs and CRMs. Nevertheless, to fully understand the *cis*-regulatory genome and also for a wide spectrum of applications, it is necessary to real the exact locations of all CREs and CRMs in the genome. ChIP-seq datasets for various TF in different cell and tissue types can be the key to the goal.

However, precise identification of CREs in the binding peaks from ChIP experiments is still a challenging computational problem [59]. Efforts have been made to narrow down the binding peaks through improving experimental procedures [85], 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). 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). Thus our algorithm has achieved rather high prediction sensitivity even only using these limited 168 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 8C). Second, our predicted CRMs as well as CREs in a CRM as a whole are more conserved than the corresponding randomly selected sequences (Figure 9A and B). Third, the highly conserved predicted CRMs tend to be located in the close neighborhoods of genes involved in embryonic development (Additional file 13: Table S4), which is consistent with the existing knowledge [81-83]. 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 [84,86]. This form of genetic changes plays a more pivotal role in functional evolution of organisms than previously thought [84,87]. 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 (Figure 9C and D), they are not more conserved than the randomly selected short *k*-mers in CDRs (Figure 9D). 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 4), 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 [73,75,79]. 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 [88,89], 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, of which at least 75% are likely to have transcriptional regulatory functions [73,75,79].

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 (Figure 4 and Additional file 5: Figure S2). On the other hand, because these datasets were not generated by random efforts of the community, rather, they are likely 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, our predicted result 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 states (active, poised or inactive) in specific cell types and tissues. 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 [7,21,90-93], and DNAse-seq data for nucleosome free regions [22,24,85,94,95]. In this sense, various epigenetic datasets in different cell and tissue types can complement with TF ChIP datasets and speed up the process of deciphering the entire cis-regulatory genome of an organism. Thus, a 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 [90-93]. Then, it is also possible to predict the molecular, cellular and organismal phenotypes based on the functional states of the CRMs and their variations among individuals and species, given the recent indication of the importance of CRMs in determining the phenotypes of organism [96-104].

Conclusions

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/.

Methods

Datasets

We attempted to collect all possible ChIP-seq and ChIP-chip datasets from *D. melanogaster* available to us from three sources: the modENCODE project [46], the Berkeley drosophila transcription network project (BDTNP) [53] 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 [42,43]. Because the typical lengths of known CRMs are 1,000-2,000 bp [54], we extended the binding peaks shorter than 3,000 bp to up to 3,000 bp by padding equal length of flanking genomic sequences to the two ends. If the extension on either end reaches to 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 5,000 bp 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 [54].

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,

$$S_o(d_i, d_j) = o(d_i, d_j)/|d_i| + o(d_i, d_j)/|d_j| \quad (1)$$

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 that have at least one pb overlap between the sequences in the two datasets.

Finding motifs in binding peak datasets

Based on an initial evaluation of multiple motif-finding tools for large ChIP datasets, including seeder [30], Trawler [30,31], ChIPMunk [32], HMS [33], CMF [34], STEME [35], DREME [36], DECOD [37], RSAT [38], and POSMO [39], 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 [36]. 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.

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.

Step 1 identify co-occurring motif pairs (CPs) in each dataset

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

$$S_c(M_d(i), M_d(j)) = o(M_d(i), M_d(j)) / \max\{|M_d(i)|, |M_d(j)|\}, \quad (2)$$

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 (Figure 2B and C). The cutoff α is chosen such that the predicted motifs in known CRMs are minimally excluded (Figure 5A and B). 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).

Step 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,

$$S_s\{P[M_a(i), M_a(j)], P[M_b(m), M_b(m)]\} = \max_{k\in\{i,j\},\ l\in\{m,n\}} \{Sim[M_a(k), M_b(l)]\} + Sim[M_a(r), M_b(s)],\ r\in\{i,j\}, r\neq k; s\in\{m,n\}, s\neq l, \quad (3)$$

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 [105-107]. We have shown that SPIC outperforms the existing metrics for measuring motif similarities [105-107]. 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.

Step 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 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:

$$D = |E|/|CP|, \quad (4)$$

where $|CP|$ and $|E|$ are the numbers of CPs and edges in the graph, respectively. We choose an β value such that the resulting graph is as spars as possible and has as many nodes/CPs as possible (Figure 7A and B).

Step 4 cut the CP similarity graph into dense sub-graphs, CP clusters (CPCs)

We use the Markov Chain Clustering algorithm (MCL) [61] to cut the graph into dense sub-graphs, each corresponding to a cluster of repetitively occurring CPs across multiple datasets (Figure 2D). MCL iteratively computes random walks determined by a Markov chain by alternately executing two operations (expansion and inflation) on a stochastic matrix [61]. 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 [61,105,106,108-112]. We discard the clusters containing fewer than τ CPs ($\tau = 2$ in this study, τthus we only discarded singleton CPs) (Figure 2D). Presumably, the remaining clusters contain highly similar CPs for certain two TFs. For example, cluster C1 (P1, P5, P8) in Figure 2D contains highly similar motifs (red and black ova) for two distinct TFs. For this reason we call these clusters CP clusters (CPCs) (Figure 2D).

Step 5 compute a co-occurring score for each pair of CPCs

Let C_i and C_j be two CPCs, and $\Omega_{dk}(C_i, C_j)$ 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,

$$S_{CPC}(C_i, C_j) = \frac{1}{D}\sum_{k=1}^{D} \frac{1}{N(\Omega_{dk}(C_i, C_j))} \sum_{(P_s\in C_i, P_t\in C_j)\subset\Omega_{dk}(C_iC_j)} [o(P_s, P_t)/|P_s| + o(P_s, P_t)/|P_t|], \quad (5)$$

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 $\Omega_{dk}(C_i,\ C_j)$.

Step 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 2E). The cutoff γ is chosen

based on the bimodal distribution of the S_{CPC} sores (Figure 7C).

Step 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 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 2E).

Step 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 [105-107]. 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.

Step 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 merged them in a non-overlapping one. We then connect any two adjacent CREs if their distance is shorter than a preset value δ (δ = 150 bp in this study) according to the distribution of the distances between the CREs in known CRMs (Figure 8B) and the connection cannot span over an exon unless it contain a binding site. We predict as a CRM each segment of sequence connected by CREs of Umotifs in one or multiple CRMCs.

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.

Additional files

Additional file 1: Table S1. Summary of the 168 ChIP datasets we collected.

Additional file 2: Figure S8. Number of binding peaks in the 168 ChIP datasets we collected. Datasets are sorted in ascending order according to their sizes.

Additional file 3: Table S2. Summary of the coverage of the datasets, predicted CRMs and CREs on the CDRs, NCRs and genome.

Additional file 4: Figure S1. An example of CRMs bound by TFs BCD, HB and KR from the REDfly database. The graph was shown using Gbrowser.

Additional file 5: Figure S2. Hierarchical clustering of the 56 datasets for distinct TFs based on their pair-wise binding peak overlapping scores S_o. The blow-up shows a cluster for cooperative TFs (see Results in the main text).

Additional file 6: Figure S3. Structures of the 815 CRMCs. Each node in the graphs is a CPC, and each connected graph represents a CRMC.

Additional file 7: Figure S4. Structures of the 184 Umotifs containing more than two motifs. Each node in the graphs is a putative motif, and each connected graph represents a Umotif. The logos are for the indicated Umotifs.

Additional file 8: Table S3. Summary of the 184 Umotifs.

Additional file 9: Figure S9. Examples of Umotifs and their matched known motifs with a p-value around 0.001 using TOMTOM.

Additional file 10: Figure S5. A. Umotif 72 and its four individual constituent motifs found in different datasets. Umotif 72 is similar to known motifs CG12287 and CG34395. B. Umotif 27 and its five individual constituent motifs. Umotif 27 is similar to known motif CG5249. C. Umotif 70 and its four individual constituent motifs found in different datasets. D. Umotif 93 and its two individual constituent motifs found in different datasets.

Additional file 11: Figure S6. Examples of known CREs in the recovered known CRMs that overlap with our predicted CREs, their corresponding Umotifs are similar the known motifs. See main text for the details.

Additional file 12: Figure S7. A putative CRM (shown in gray shadow) is located in the first intron of gene *act*57B.

Additional file 13: Table S4. Enriched GO terms for the putative target genes of highly conserved putative CRMs.

Additional file 14: Table S5. Enriched GO terms for the putative target genes of highly non-conserved putative CRMs.

Abbreviations

CRE: *cis*-regulatory element; CRM: *cis*-regulatory module; bp: Base pairs; CDRs: Coding regions; ChIP: Chromatin immunepreciptation; CPs: Co-occurring pairs; CPCs: Co-occurring pair clusters; CRMCs: CRM components; MCL: Markov chain clustering; NCRs: Non-coding regions; NGS: Next-generation sequencing; PSWMs: Position specific weight matrixes; TF: Transcription factor.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
Conceived and designed the algorithm and experiments: ZCS and MN. Analyzed the data: MN, ET and ZCS. Designed and implemented the database: ET and MN. Wrote the paper: ZCS and MN. All authors read and approved the final manuscript.

Acknowledgements
This work was supported by the National Science Foundation (EF0849615 and CCF1048261), and the UNC Charlotte (Faculty Research Grants). Funding for open access charge is from CCF1048261. We would like to thank members of the Su lab for discussions. We are thankful to Drs. Steve Gallo and Marc Halfon for providing the detailed CRM data in REDfly, James Johnson for providing a Perl script for matching DREME results to sequences, and Dr. Jason Lieb for his suggestions.

Received: 26 June 2014 Accepted: 19 November 2014
Published: 2 December 2014

References

1. Consortium CeS: **Genome sequence of the nematode C. elegans: a platform for investigating biology.** *Science* 1998, **282**(5396):2012–2018.
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–D579.
3. 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.
4. Collins F: **Has the revolution arrived?** *Nature* 2010, **464**(7289):674–675.
5. Consortium TEP: **The ENCODE (ENCyclopedia Of DNA Elements) Project.** *Science* 2004, **306**(5696):636–640.
6. Celniker SE, Dillon LA, Gerstein MB, Gunsalus KC, Henikoff S, Karpen GH, Kellis M, Lai EC, Lieb JD, MacAlpine DM, Micklem G, Piano F, Snyder M, Stein L, White KP, Waterston RH: **Unlocking the secrets of the genome.** *Nature* 2009, **459**(7249):927–930.
7. Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, Kellis M, Marra MA, Beaudet AL, Ecker JR, Farnham PJ, Hirst M, Lander ES, Mikkelsen TS, Thomson JA: **The NIH roadmap epigenomics mapping consortium.** *Nat Biotechnol* 2010, **28**(10):1045–1048.
8. Hartwell LH, Hopfield JJ, Leibler S, Murray AW: **From molecular to modular cell biology.** *Nature* 1999, **402**(6761 Suppl):C47–C52.
9. Temple G, Gerhard DS, Rasooly R, Feingold EA, Good PJ, Robinson C, Mandich A, Derge JG, Lewis J, Shoaf D, Collins FS, Jang W, Wagner L, Shenmen CM, Misquitta L, Schaefer CF, Buetow KH, Bonner TI, Yankie L, Ward M, Phan L, Astashyn A, Brown G, Farrell C, Hart J, Landrum M, Maidak BL, Murphy M, Murphy T, Rajput B, *et al*: **The completion of the Mammalian Gene Collection (MGC).** *Genome Res* 2009, **19**(12):2324–2333.
10. Maston GA, Evans SK, Green MR: **Transcriptional regulatory elements in the human genome.** *Annu Rev Genomics Hum Genet* 2006, **7**:29–59.
11. Narlikar L, Ovcharenko I: **Identifying regulatory elements in eukaryotic genomes.** *Brief Funct Genomic Proteomic* 2009, **8**(4):215–230.
12. 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.
13. Davidson EH: *The Regulatory Genome: Gene Regulatory Networks In Development and Evolution.* Waltham, Massachusetts: Academic Press; 2006.
14. Tompa M, Li N, Bailey TL, Church GM, De Moor B, Eskin E, Favorov AV, Frith MC, Fu Y, Kent WJ, Makeev VJ, Mironov AA, Noble WS, Pavesi G, Pesole G, Regnier M, Simonis N, Sinha S, Thijs G, van Helden J, Vandenbogaert M, Weng Z, Workman C, Ye C, Zhu Z: **Assessing computational tools for the discovery of transcription factor binding sites.** *Nat Biotechnol* 2005, **23**(1):137–144.
15. Heintzman ND, Ren B: **Finding distal regulatory elements in the human genome.** *Curr Opin Genet Dev* 2009, **19**(6):541–549.
16. Hardison RC, Taylor J: **Genomic approaches towards finding cis-regulatory modules in animals.** *Nat Rev Genet* 2012, **13**(7):469–483.
17. 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.
18. Johnson DS, Mortazavi A, Myers RM, Wold B: **Genome-wide mapping of in vivo protein-DNA interactions.** *Science* 2007, **316**(5830):1497–1502.
19. Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A, Thiessen N, Griffith OL, He A, Marra M, Snyder M, Jones S: **Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing.** *Nat Methods* 2007, **4**(8):651–657.
20. Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, Wong E, Orlov YL, Zhang W, Jiang J, Loh YH, Yeo HC, Yeo ZX, Narang V, Govindarajan KR, Leong B, Shahab A, Ruan Y, Bourque G, Sung WK, Clarke ND, Wei CL, Ng HH: **Integration of external signaling pathways with the core transcriptional network in embryonic stem cells.** *Cell* 2008, **133**(6):1106–1117.
21. 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.
22. 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.
23. Song L, Zhang Z, Grasfeder LL, Boyle AP, Giresi PG, Lee BK, Sheffield NC, Graf S, Huss M, Keefe D, Liu Z, London D, McDaniell RM, Shibata Y, Showers KA, Simon JM, Vales T, Wang T, Winter D, Clarke ND, Birney E, Iyer VR, Crawford GE, Lieb JD, Furey TS: **Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity.** *Genome Res* 2011, **21**(10):1757–1767.
24. Crawford GE, Holt IE, Whittle J, Webb BD, Tai D, Davis S, Margulies EH, Chen Y, Bernat JA, Ginsburg D, Zhou D, Luo S, Vasicek TJ, Daly MJ, Wolfsberg TG, Collins FS: **Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS).** *Genome Res* 2006, **16**(1):123–131.
25. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J: **Comprehensive mapping of long-range interactions reveals folding principles of the human genome.** *Science* 2009, **326**(5950):289–293.
26. 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, **58**(3):268–276.
27. 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.
28. Park PJ: **ChIP-seq: advantages and challenges of a maturing technology.** *Nat Rev Genet* 2009, **10**(10):669–680.
29. Pepke S, Wold B, Mortazavi A: **Computation for ChIP-seq and RNA-seq studies.** *Nat Methods* 2009, **6**(11 Suppl):S22–S32.
30. Fauteux F, Blanchette M, Stromvik MV: **Seeder: discriminative seeding DNA motif discovery.** *Bioinformatics* 2008, **24**(20):2303–2307.
31. Ettwiller L, Paten B, Ramialison M, Birney E, Wittbrodt J: **Trawler: de novo regulatory motif discovery pipeline for chromatin immunoprecipitation.** *Nat Methods* 2007, **4**(7):563–565.
32. 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.
33. 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.
34. Mason MJ, Plath K, Zhou Q: **Identification of context-dependent motifs by contrasting ChIP binding data.** *Bioinformatics* 2010, **26**(22):2826–2832.
35. Reid JE, Wernisch L: **STEME: efficient EM to find motifs in large data sets.** *Nucleic Acids Res* 2011, **39**(18):e126.
36. Bailey TL: **DREME: motif discovery in transcription factor ChIP-seq data.** *Bioinformatics* 2011, **27**(12):1653–1659.
37. 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.
38. 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.
39. 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.

40. Whitington T, Frith MC, Johnson J, Bailey TL: **Inferring transcription factor complexes from ChIP-seq data.** *Nucleic Acids Res* 2011, **39**(15):e98.
41. 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.
42. Negre N, Brown CD, Ma L, Bristow CA, Miller SW, Wagner U, Kheradpour P, Eaton ML, Loriaux P, Sealfon R, Li Z, Ishii H, Spokony RF, Chen J, Hwang L, Cheng C, Auburn RP, Davis MB, Domanus M, Shah PK, Morrison CA, Zieba J, Suchy S, Senderowicz L, Victorsen A, Bild NA, Grundstad AJ, Hanley D, MacAlpine DM, Mannervik M, *et al*: **A cis-regulatory map of the Drosophila genome.** *Nature* 2011, **471**(7339):527–531.
43. Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, Liu T, Yip KY, Robilotto R, Rechtsteiner A, Ikegami K, Alves P, Chateigner A, Perry M, Morris M, Auerbach RK, Feng X, Leng J, Vielle A, Niu W, Rhrissorrakrai K, Agarwal A, Alexander RP, Barber G, Brdlik CM, Brennan J, Brouillet JJ, Carr A, Cheung MS, Clawson H, Contrino S, *et al*: **Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project.** *Science* 2010, **330**(6012):1775–1787.
44. Zhang Z, Chang CW, Goh WL, Sung WK, Cheung E, Web Server issue: **CENTDIST: discovery of co-associated factors by motif distribution.** *Nucleic Acids Res* 2011, **39:**W391–W399.
45. ENCODE: **A user's guide to the encyclopedia of DNA elements (ENCODE).** *PLoS Biol* 2011, **9**(4):e1001046.
46. Roy S, Ernst J, Kharchenko PV, Kheradpour P, Negre N, Eaton ML, Landolin JM, Bristow CA, Ma L, Lin MF, Washietl S, Arshinoff BI, Ay F, Meyer PE, Robine N, Washington NL, Di Stefano L, Berezikov E, Brown CD, Candeias R, Carlson JW, Carr A, Jungreis I, Marbach D, Sealfon R, Tolstorukov MY, Will S, Alekseyenko AA, Artieri C, Booth BW, *et al*: **Identification of functional elements and regulatory circuits by Drosophila modENCODE.** *Science* 2010, **330**(6012):1787–1797.
47. Chen G, Zhou Q: **Searching ChIP-seq genomic islands for combinatorial regulatory codes in mouse embryonic stem cells.** *BMC Genomics* 2011, **12:**515.
48. Wingender E, Chen X, Fricke E, Geffers R, Hehl R, Liebich I, Krull M, Matys V, Michael H, Ohnhauser R, Pruss M, Schacherer F, Thiele S, Urbach S: **The TRANSFAC system on gene expression regulation.** *Nucleic Acids Res* 2001, **29:**281–283.
49. 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–D97.
50. Wang J, Zhuang J, Iyer S, Lin X, Whitfield TW, Greven MC, Pierce BG, Dong X, Kundaje A, Cheng Y, Rando OJ, Birney E, Myers RM, Noble WS, Snyder M, Weng Z: **Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors.** *Genome Res* 2012, **22**(9):1798–1812.
51. Wang J, Zhuang J, Iyer S, Lin XY, Greven MC, Kim BH, Moore J, Pierce BG, Dong X, Virgil D, Birney E, Hung JH, Weng Z: **Factorbook.org: a Wiki-based database for transcription factor-binding data generated by the ENCODE consortium.** *Nucleic Acids Res* 2013, **41**(Database issue):D171–D176.
52. Davidson EH, Rast JP, Oliveri P, Ransick A, Calestani C, Yuh CH, Minokawa T, Amore G, Hinman V, Arenas-Mena C, Otim O, Brown CT, Livi CB, Lee PY, Revilla R, Rust AG, Pan Z, Schilstra MJ, Clarke PJ, Arnone MI, Rowen L, Cameron RA, McClay DR, Hood L, Bolouri H: **A genomic regulatory network for development.** *Science* 2002, **295**(5560):1669–1678.
53. Li XY, MacArthur S, Bourgon R, Nix D, Pollard DA, Iyer VN, Hechmer A, Simirenko L, Stapleton M, Luengo Hendriks CL, Chu HC, Ogawa N, Inwood W, Sementchenko V, Beaton A, Weiszmann R, Celniker SE, Knowles DW, Gingeras T, Speed TP, Eisen MB, Biggin MD: **Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm.** *PLoS Biol* 2008, **6**(2):e27.
54. Gallo SM, Gerrard DT, Miner D, Simich M, Des Soye B, Bergman CM, Halfon MS MS: **REDfly v3.0: toward a comprehensive database of transcriptional regulatory elements in Drosophila.** *Nucleic Acids Res* 2011, **39**(Database issue):D118–D123.
55. 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.
56. Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan KK, Cheng C, Mu XJ, Khurana E, Rozowsky J, Alexander R, Min R, Alves P, Abyzov A, Addleman N, Bhardwaj N, Boyle AP, Cayting P, Charos A, Chen DZ, Cheng Y, Clarke D, Eastman C, Euskirchen G, Frietze S, Fu Y, Gertz J, Grubert F, Harmanci A, Jain P, Kasowski M, *et al*: **Architecture of the human regulatory network derived from ENCODE data.** *Nature* 2012, **489**(7414):91–100.
57. Machanick P, Bailey TL: **MEME-ChIP: motif analysis of large DNA datasets.** *Bioinformatics* 2011, **27**(12):1696–1697.
58. Mathelier A, Wasserman WW: **The next generation of transcription factor binding site prediction.** *PLoS Comput Biol* 2013, **9**(9):e1003214.
59. Tran NT, Huang CH: **A survey of motif finding Web tools for detecting binding site motifs in ChIP-Seq data.** *Biol Direct* 2014, **9**(1):4.
60. Bolouri H, Ruzzo WL: **Integration of 198 ChIP-seq datasets reveals human cis-regulatory regions.** *J Comput Biol* 2012, **19**(9):989–997.
61. van Dongen S: *A cluster Algorithm for Graphs.* Amsterdam: National Research Institute for Mathematics and Computer Science in the Netherlands; 2000.
62. Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS: **Quantifying similarity between motifs.** *Genome Biol* 2007, **8**(2):R24.
63. 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.
64. Zhu LJ, Christensen RG, Kazemian M, Hull CJ, Enuameh MS, Basciotta MD, Brasefield JA, Zhu C, Asriyan Y, Lapointe DS, Sinha S, Wolfe SA, Brodsky MH: **FlyFactorSurvey: a database of Drosophila transcription factor binding specificities determined using the bacterial one-hybrid system.** *Nucleic Acids Res* 2011, **39**(Database issue):D111–D117.
65. Kulakovskiy IV, Makeev VJ: **Discovery of DNA motifs recognized by transcription factors through integration of different experimental sources.** *Biophysics* 2009, **54**(6):667–674.
66. Brand AH, van Roessel PJ: **Region-specific apoptosis limits neural stem cell proliferation.** *Neuron* 2003, **37**(2):185–187.
67. 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.
68. Sanyal S, Narayanan R, Consoulas C, Ramaswami M: **Evidence for cell autonomous AP1 function in regulation of Drosophila motor-neuron plasticity.** *BMC Neurosci* 2003, **4:**20.
69. 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.
70. 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.
71. Tanaka KK, Bryantsev AL, Cripps RM: **Myocyte enhancer factor 2 and chorion factor 2 collaborate in activation of the myogenic program in Drosophila.** *Mol Cell Biol* 2008, **28**(5):1616–1629.
72. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D: **Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes.** *Genome Res* 2005, **15**(8):1034–1050.
73. 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.
74. 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.
75. Andolfatto P: **Adaptive evolution of non-coding DNA in Drosophila.** *Nature* 2005, **437**(7062):1149–1152.
76. Casillas S, Barbadilla A, Bergman CM: **Purifying selection maintains highly conserved noncoding sequences in Drosophila.** *Mol Biol Evol* 2007, **24**(10):2222–2234.
77. 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.
78. 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.
79. Kondrashov AS: **Evolutionary biology: fruitfly genome is not junk.** *Nature* 2005, **437**(7062):1106.
80. 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):3.

81. Ciglar L, Furlong EE: **Conservation and divergence in developmental networks: a view from Drosophila myogenesis.** *Curr Opin Cell Biol* 2009, **21**(6):754–760.
82. Zeitlinger J, Stark A: **Developmental gene regulation in the era of genomics.** *Dev Biol* 2010, **339**(2):230–239.
83. Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T, Smith SF, North P, Callaway H, Kelly K, Walter K, Abnizova I, Gilks W, Edwards YJ, Cooke JE, Elgar G: **Highly conserved non-coding sequences are associated with vertebrate development.** *PLoS Biol* 2005, **3**(1):e7.
84. Wray GA: **The evolutionary significance of cis-regulatory mutations.** *Nat Rev Genet* 2007, **8**(3):206–216.
85. Zhang Z, Pugh BF: **High-resolution genome-wide mapping of the primary structure of chromatin.** *Cell* 2011, **144**(2):175–186.
86. 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.
87. Wittkopp PJ, Kalay G: **Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence.** *Nat Rev Genet* 2012, **13**(1):59–69.
88. Sandelin A, Wasserman WW: **Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics.** *J Mol Biol* 2004, **338**(2):207–215.
89. Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M, Taipale M, Wei G, Palin K, Vaquerizas JM, Vincentelli R, Luscombe NM, Hughes TR, Lemaire P, Ukkonen E, Kivioja T, Taipale J: **DNA-binding specificities of human transcription factors.** *Cell* 2013, **152**(1–2):327–339.
90. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, Ku M, Durham T, Kellis M, Bernstein BE: **Mapping and analysis of chromatin state dynamics in nine human cell types.** *Nature* 2011, **473**(7345):43–49.
91. Ram O, Goren A, Amit I, Shoresh N, Yosef N, Ernst J, Kellis M, Gymrek M, Issner R, Coyne M, Durham T, Zhang X, Donaghey J, Epstein CB, Regev A, Bernstein BE: **Combinatorial patterning of chromatin regulators uncovered by genome-wide location analysis in human cells.** *Cell* 2011, **147**(7):1628–1639.
92. Zhou VW, Goren A, Bernstein BE: **Charting histone modifications and the functional organization of mammalian genomes.** *Nat Rev Genet* 2011, **12**(1):7–18.
93. Zhu J, Adli M, Zou JY, Verstappen G, Coyne M, Zhang X, Durham T, Miri M, Deshpande V, De Jager PL, Bennett DA, Houmard JA, Muoio DM, Onder TT, Camahort R, Cowan CA, Meissner A, Epstein CB, Shoresh N, Bernstein BE: **Genome-wide chromatin state transitions associated with developmental and environmental cues.** *Cell* 2013, **152**(3):642–654.
94. Jiang C, Pugh BF: **Nucleosome positioning and gene regulation: advances through genomics.** *Nat Rev Genet* 2009, **10**(3):161–172.
95. Ioshikhes I, Hosid S, Pugh BF: **Variety of genomic DNA patterns for nucleosome positioning.** *Genome Res* 2011, **21**(11):1863–1871.
96. Fraser HB: **Gene expression drives local adaptation in humans.** *Genome Res* 2013, **23**(7):1089–1096.
97. Ye K, Lu J, Raj SM, Gu Z: **Human expression QTLs are enriched in signals of environmental adaptation.** *Genome Biol Evol* 2013, **5**(9):1689–1701.
98. Babak T, Garrett-Engele P, Armour CD, Raymond CK, Keller MP, Chen R, Rohl CA, Johnson JM, Attie AD, Fraser HB, Schadt EE: **Genetic validation of whole-transcriptome sequencing for mapping expression affected by cis-regulatory variation.** *BMC Genomics* 2010, **11**:473.
99. Stranger BE, Montgomery SB, Dimas AS, Parts L, Stegle O, Ingle CE, Sekowska M, Smith GD, Evans D, Gutierrez-Arcelus M, Price A, Raj T, Nisbett J, Nica AC, Beazley C, Durbin R, Deloukas P, Dermitzakis ET: **Patterns of cis regulatory variation in diverse human populations.** *PLoS Genet* 2012, **8**(4):e1002639.
100. Vernot B, Stergachis AB, Maurano MT, Vierstra J, Neph S, Thurman RE, Stamatoyannopoulos JA, Akey JM: **Personal and population genomics of human regulatory variation.** *Genome Res* 2012, **22**(9):1689–1697.
101. Zheng W, Zhao H, Mancera E, Steinmetz LM, Snyder M: **Genetic analysis of variation in transcription factor binding in yeast.** *Nature* 2010, **464**(7292):1187–1191.
102. Kasowski M, Kyriazopoulou-Panagiotopoulou S, Grubert F, Zaugg JB, Kundaje A, Liu Y, Boyle AP, Zhang QC, Zakharia F, Spacek DV, Li J, Xie D, Olarerin-George A, Steinmetz LM, Hogenesch JB, Kellis M, Batzoglou S, Snyder M: **Extensive variation in chromatin states across humans.** *Science* 2013, **342**(6159):750–752.
103. Kasowski M, Grubert F, Heffelfinger C, Hariharan M, Asabere A, Waszak SM, Habegger L, Rozowsky J, Shi M, Urban AE, Hong MY, Karczewski KJ, Huber W, Weissman SM, Gerstein MB, Korbel JO, Snyder M: **Variation in transcription factor binding among humans.** *Science* 2010, **328**(5975):232–235.
104. Haraksingh RR, Snyder MP: **Impacts of variation in the human genome on gene regulation.** *J Mol Biol* 2013, **425**(21):3970–3977.
105. 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.
106. 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.
107. 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.
108. van Dongen S, Abreu-Goodger C: **Using MCL to extract clusters from networks.** *Methods Mol Biol* 2012, **804**:281–295.
109. Vlasblom J, Wodak SJ: **Markov clustering versus affinity propagation for the partitioning of protein interaction graphs.** *BMC Bioinformatics* 2009, **10**:99.
110. Brohee S, van Helden J: **Evaluation of clustering algorithms for protein-protein interaction networks.** *BMC Bioinformatics* 2006, **7**:488.
111. Samuel Lattimore B, van Dongen S, Crabbe MJ: **GeneMCL in microarray analysis.** *Comput Biol Chem* 2005, **29**(5):354–359.
112. 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.

doi:10.1186/1471-2164-15-1047
Cite this article as: Niu *et al.*: De novo prediction of *cis*-regulatory elements and modules through integrative analysis of a large number of ChIP datasets. *BMC Genomics* 2014 **15**:1047.