Genome Biology

Genome Biology
Biology for the post-genomic era

This Provisional PDF corresponds to the article as it appeared upon acceptance. Copyedited and fully formatted PDF and full text (HTML) versions will be made available soon.

ModuleMiner: improved computational detection of cis-regulatory modules. Different modes of gene regulation in embryonic development and adult tissues?

Genome Biology 2008, **9**:R66 doi:10.1186/gb-2008-9-4-r66

Peter Van Loo (Peter.VanLoo@med.kuleuven.be)
Stein Aerts (Stein.Aerts@med.kuleuven.be)
Bernard Thienpont (Bernard.Thienpont@med.kuleuven.be)
Bart De Moor (Bart.DeMoor@esat.kuleuven.be)
Yves Moreau (Yves.Moreau@esat.kuleuven.be)
Peter Marynen (Peter.Marynen@med.kuleuven.be)

ISSN 1465-6906

Article type Method

Submission date 30 December 2007

Acceptance date 7 April 2008

Publication date 7 April 2008

Article URL http://genomebiology.com/2008/9/4/R66

This peer-reviewed article was published immediately upon acceptance. It can be downloaded, printed and distributed freely for any purposes (see copyright notice below).

Articles in *Genome Biology* are listed in PubMed and archived at PubMed Central.

For information about publishing your research in *Genome Biology* go to

http://genomebiology.com/info/instructions/

© 2008 Van Loo *et al.*, licensee BioMed Central Ltd.
This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

MODULEMINER: improved computational detection of *cis*-regulatory modules. Different modes of gene regulation in embryonic development and adult tissues?

Peter Van Loo[1,2,3,§], Stein Aerts[1,2], Bernard Thienpont[2], Bart De Moor[3], Yves Moreau[3], Peter Marynen[1,2]

[1]Department of Molecular and Developmental Genetics, VIB, Herestraat 49, box 602, B-3000 Leuven, Belgium

[2]Department of Human Genetics, University of Leuven, Herestraat 49, box 602, B-3000 Leuven, Belgium

[3]Bioinformatics group, Department of Electrical Engineering (ESAT-SCD), University of Leuven, Kasteelpark Arenberg 10, B-3001 Heverlee, Belgium

[§]Corresponding author

Email addresses:

PVL: Peter.VanLoo@med.kuleuven.be

SA: Stein.Aerts@med.kuleuven.be

BT: Bernard.Thienpont@med.kuleuven.be

BDM: Bart.DeMoor@esat.kuleuven.be

YM: Yves.Moreau@esat.kuleuven.be

PM: Peter.Marynen@med.kuleuven.be

Abstract

We present MODULEMINER, a novel algorithm for computationally detecting *cis*-regulatory modules (CRMs) in a set of co-expressed genes. MODULEMINER outperforms other methods for CRM detection on benchmark data and successfully detects CRMs in tissue-specific microarray clusters and in embryonic development gene sets. Interestingly, CRM predictions for differentiated tissues show a strong enrichment close to the transcription start site, while CRM predictions for embryonic development gene sets are depleted in this region.

Background

The identification and functional annotation of transcriptional regulatory sequences in the human genome is lagging far behind the rapidly increasing knowledge of protein-coding genes. These transcriptional regulatory sequences are often build up in a modular fashion and exert their function in *cis* through the concerted binding of multiple transcription factors (and co-factors), resulting in the formation of protein complexes that interact with RNA polymerase II [1,2]. These sequences are called *cis*-regulatory modules (CRMs). In theory, these CRMs can be detected by the presence of multiple transcription factor binding sites. However, in practice, the reliable detection of functional transcription factor binding sites is difficult and results in many false positives, partly because these binding sites are too short and too degenerate [3]. Hence, the computational detection of functional regulatory sequences in the human genome remains a formidable challenge.

Multiple method have been developed that aim to computationally detect regulatory sequences [4-8]. Promising and validated results have been delivered mostly in model organisms with relatively compact genomes (e.g. *Drosophila melanogaster*) [9-11]. In the larger human genome, deep sequence conservation (e.g. up to zebrafish) or extreme sequence conservation (e.g. perfect conservation in mouse over 200 base pairs), irrespective of transcription factor binding site detection, remains the method of choice for approaches validating regulatory sequences *in vitro* or *in vivo* [12-14]. While these conservation approaches are quite successful in predicting which regions have a regulatory function, they provide no information on what expression pattern these regions produce and by which transcription factors they are targeted.

When several similar CRMs have been characterized, and the regulatory factors and binding sites have been elucidated, one can use this knowledge to find new examples of similar CRMs directing the transcription of other genes involved in the same process. A number of computational methods have been described that apply this approach [15-17]. These methods have been highly successful [10,11,18], but in practice, apart from in *Drosophila* embryonic development, the lack of available data often precludes the application of these approaches.

When this knowledge is not available, the detection of tissue- or process-specific CRMs can be tackled by looking for recurring combinations of transcription factor binding sites in putative regulatory regions of a set of co-expressed genes. A few methods applying this approach have been developed [19-22]. However, in part because this is a more complex problem, these methods have only been applied on a limited scale and did not report many successful predictions. To our knowledge, only our ModuleSearcher method [20] has provided results subjected to experimental validation [23].

Here, we develop MODULEMINER, a novel algorithm to detect similar CRMs in a set of co-expressed genes, focussed on the human genome. MODULEMINER does not require prior knowledge of regulating transcription factors or annotated binding sites, but uses only a library of position weight matrices (PWMs). Contrary to existing algorithms that require *a priori* unknown CRM properties (such as the length of the CRMs or the number of binding sites) as input parameters, MODULEMINER is parameterless. In addition, MODULEMINER differs from existing similar approaches in that it implements a whole-genome optimization strategy to specifically look for signals that discriminate the given co-expressed genes from all other genes in the genome. By leave-one-out cross-validation on benchmark data, we show that

MODULEMINER outperforms other methods that computationally detect CRMs. Finally, we demonstrate that MODULEMINER can successfully detect similar CRMs in microarray clusters with a tissue specific expression profile, as well as in custom-build gene sets related to specific embryonic developmental processes. In total, MODULEMINER predicted 257 CRMs near the genes studied, as well as an additional 1400 CRM predictions resulting from full genome scans for new target genes. We further analyze these CRM predictions to elucidate differences between CRMs directing transcription in differentiated tissues and CRMs directing transcription during embryonic development.

Results

MODULEMINER: detection of similar *cis*-regulatory modules in a set of co-expressed genes

We developed MODULEMINER, a novel algorithm to detect similar CRMs in a set of co-expressed genes. MODULEMINER models similar CRMs as a combination of motifs (represented by PWMs), as in [20]. These models are called "transcriptional regulatory models" (TRMs) [24]. We postulate that a good TRM is able to retrieve targets in the genome. Therefore, we express the fitness of a TRM in terms of its target gene recovery and we select the TRM that has maximum specificity for the given set of co-expressed genes, by a whole-genome optimization strategy. To determine the fitness of a TRM, each gene's search space is first scored with the TRM, where we define a gene's search space as the collection of all conserved non-coding sequences within 10 kb 5' of the transcription start site (see Materials and methods). These scores are then used to rank all genes in the genome. Finally, the ranks of the given co-expressed genes are determined, and the probability of observing this collection of ranks by chance is calculated using order statistics (see Materials and methods). If a large part of the co-expressed genes are ranked high, the order statistic is highly significant, and hence the TRM is considered to have a high fitness for modelling similar *cis*-regulatory modules regulating these genes. MODULEMINER searches the TRM with the most significant order statistic (i.e. the best fitness) using a genetic algorithm (detailed in Materials and methods).

We introduce MODULEMINER and its rigorous validation procedure by an example case study. We constructed a high quality set of 12 smooth muscle marker genes [25], and performed leave-one-out cross-validation (LOOCV). In each validation run, one

gene was left out and MODULEMINER constructed a TRM using the remaining 11 genes. This TRM was then used to rank all genes in the genome and the position of the left-out gene was determined. The set of 12 ranks obtained in this way was used to calculate sensitivity/specificity pairs, which were subsequently plotted on a receiver operator characteristic (ROC) curve. We used the area under this curve (AUC) as a measure of MODULEMINER performance on this set of co-expressed genes.

We repeated the LOOCV for three sets of candidate transcription factor binding sites (TFBSs): (i) predicted binding sites in human-mouse conserved non-coding sequences (CNSs), obtained by aligning 10 kb 5' of all human-mouse orthologs and selecting regions of at least 75 % identity over a minimum of 100 base pairs; (ii) binding sites from (i), retaining only the PWMs for which in both the human and mouse CNS an instance is predicted (we follow the nomenclature in [10] and call these sites *preserved* sites); (iii) as in (ii), but here the CNSs are obtained by aligning 10 kb 5' of all human genes to 110 kb 5' + 100 kb 3' of the transcription start site of their mouse orthologs (and hence correcting for possible differences in transcription start site annotation) (Table 1). The resulting ROC curves are shown in Figure 1A. In all three cases, the AUC values are significantly above 50 % (the theoretical value obtained if the left-out genes would be ranked randomly), indicating that the TRMs obtained are sensitive and specific in predicting *cis*-regulatory modules near the left-out genes.

We observed that similar TRMs have a similar fitness and a similar order statistic. The TRM that is selected by MODULEMINER (the one that has the lowest order statistic) is surrounded by similar TRMs with order statistics that are only slightly larger. The selection of one TRM out of these similar TRMs is inherently arbitrary and depends only marginally on the true regulatory signals. To make MODULEMINER more robust to this "noise", we cluster the top-scoring TRMs and select the most

prominent cluster, instead of the single optimal TRM. We call this cluster of TRMs a "transcriptional regulatory global model" (TRGM). The results of a LOOCV when using these TRGMs (Figure 1B) show that this indeed has a positive effect on MODULEMINER performance, as the AUCs increased by 6 % on average. Furthermore, these TRGMs provide additional information compared to singular TRMs, since they allow an estimate of the relative importance of each PWM involved, as discussed below.

When comparing the performance of MODULEMINER (using TRGMs) on the three sets of candidate binding sites, a large difference between selecting all detected binding sites (set 1, AUC value of 84.6 %) and restricting to preserved sites only (set 2, AUC value of 92.8 %) is apparent. Correcting for transcription start site (TSS) differences in human and mouse (set 3, AUC value of 92.5 %) did not increase this performance further. Thus, for this high quality set of co-expressed genes, the preservation of binding sites is highly beneficial for efficient detection of *cis*-regulatory modules. This strongly suggests that for this gene set, the trans-acting factors are conserved between human and mouse.

We next applied the MODULEMINER algorithm to the full set of 12 smooth muscle marker genes, using the site preservation measure (set 2). The resulting TRGM identifies SRF, SMAD4, SP1 and ATF3 as the main transcription factors involved in the co-regulation of these genes (detailed MODULEMINER output is reported at our website). Importantly, MODULEMINER implicates SRF as the most important smooth muscle regulator, and suggests that smooth muscle specific regulation often entails two or more SRF binding sites, in agreement with literature [26].

To verify the added value of the resulting combination of PWMs over SRF alone, we manually generated a TRGM containing only PWMs for SRF and compared this to

MODULEMINERs performance. When we applied this "SRF only" TRGM to rank the genome, we obtained an AUC of 79.9 %, significantly smaller than the 92.8 % AUC of MODULEMINER (obtained in an LOOCV setting).

Sensitivity to noise

To assess the performance of MODULEMINER as a function of the composition of the input set of co-expressed genes, we performed LOOCV on input sets that contain a varying percentage of genuinely co-regulated genes ("true positives"). As true positive genes, we selected the set of 10 smooth muscle markers that share similar *cis*-regulatory modules that can be identified by MODULEMINER (these 10 genes all are ranked within the top 7 % of the genome by a LOOCV, as shown in Figure 1B). We approximated negative genes (genes that do not contain the smooth muscle *cis*-regulatory module) by random genes.

In a first analysis, we kept the number of true positive genes constant at 10, and we added a varying number of negative genes. The decrease in performance as a function of an increasing number of negative genes was surprisingly small (Figure 2). Even when only 10 of 50 genes contained the smooth muscle *cis*-regulatory module, MODULEMINER was able to pick up this signal (the AUC was 85.2 %, and SRF and SP1 were still found as key factors).

In a second analysis, we kept the total number of genes constant at 10, and we varied the percentage of negative genes. We now observed a steep decrease in MODULEMINER performance as a function of an increasing percentage of negative genes (Figure 2).

We conclude from these experiments that MODULEMINER requires a critical mass of true positive genes for successful detection of similar *cis*-regulatory modules.

However, when this critical mass is present, MODULEMINER is highly robust to false positive genes.

Comparison to other CRM detection algorithms

We next compared MODULEMINER to other *in silico* approaches for CRM detection on benchmark data. From PAZAR [27], we selected all 'boutiques' containing annotated regulatory regions directing expression in a particular system: (i) M02, muscle; (ii) M03, liver; (iii) M08, ORegAnno Stat1 and (iv) M09, ORegAnno Erythroid. As a fifth benchmark set, we used the 12 smooth muscle genes described above. On each of these 5 sets, we compared the performance of MODULEMINER to that of 4 state-of-the-art publicly available algorithms to detect similar CRMs in co-expressed genes: ModuleSearcher [28], CREME [19], CisModule [22] and EMCMODULE [29]. We also included the Clover algorithm [30], which looks for individual overrepresented transcription factor binding sites in putative regulatory sequences of a set of co-expressed genes. We note that our analysis does not focus specifically on the known enhancers, but in contrast, we consider all CNSs in the entire 10 kb 5' of the TSS (which may or may not contain the known enhancer, as well as other sequences). This effectively mimics a real-life situation, where the exact location of the regulatory sequences is not known *a priori*.

The CREME algorithm was unable to identify similar CRMs in any of the 5 benchmark sets, most likely in part because of its focus on larger sets of more loosely co-expressed genes [19]. Using the remaining algorithms, we performed LOOCV on each of the 5 benchmark sets. For this LOOCV, we used each algorithm to train a TRM or TRGM using gene sets where one gene is left out (see Materials and methods for details). Hence, as training data, we used all CNSs in the 10 kb 5' of the TSS of the benchmark set, except for the left-out gene. For CisModule and EMCMODULE,

the inputs were the sequences of the CNSs; for Clover, the inputs where the sequences of the CNSs as well as all TRANSFAC and JASPAR vertebrate PWMs; for ModuleSearcher, the inputs were the predicted binding sites within those CNSs, using all TRANSFAC and JASPAR vertebrate PWMs. The combination of PWMs that each algorithm provided as output was used to build a TRM or TRGM. We subsequently used the ModuleScanner algorithm to rank all genes in the genome based on the predicted TRM/TRGM, and we used the results to construct ROC curves. We used the site preservation measure (candidate TFBS set 2) for the MODULEMINER runs (as this was the set where we obtained the best results for the smooth muscle genes). Since the other algorithms do not use site preservation in the discovery step, we used candidate TFBS set 1 (without preservation) also in their genome ranking step. We also constructed random ROC curves based on genome ranking using random TRMs (see Materials and methods for details). On the OregAnno Erythroid benchmark set neither MODULEMINER nor any of the other algorithms seem to perform better than random (Figure 3A). As this is the smallest set, containing only 6 genes with human-mouse CNSs, this is consistent with the results we obtained in the previous section, where we concluded that a critical number of co-regulated genes is required for CRM detection. In contrast, on each of the 4 other benchmark sets, MODULEMINER performs better than random TRMs, as do some of the other algorithms (Figure 3B-E). Comparing the performance of all CRM detection algorithms, MODULEMINER seems to show the best performance in all 4 cases. Interestingly, only MODULEMINER can compete with “simple” transcription factor binding site overrepresentation in this setup, emulating a real-life situation where the regulatory sequences are not known. Indeed, only MODULEMINER outperforms Clover on four of the five benchmark sets.

On the fifth benchmark set (muscle), Clover and MODULEMINER seem to be closely matched, with the Clover method showing a steeper start of the ROC curve.

The performance of the other CRM detection algorithms can be improved by using site preservation (TFBS set 2) in the genome ranking step (Figure 3F-I), although here as well, MODULEMINER outperforms all other CRM detection algorithms, suggesting that the TRMs predicted by MODULEMINER are more informative or more specific than those suggested by other methods. Candidate TFBS set 2 was not in all cases the optimal choice for MODULEMINER: on the muscle benchmark set, candidate TFBS set 3 performed better (Figure 3J).

We noticed the CRM predictions MODULEMINER made on the muscle, liver and ORegAnno Stat1 sets, correspond well with the known regulatory elements. The TRGMs MODULEMINER contructed contain PWMs for SRF, MEF2, Myf and MyoD (muscle), HNF1, HNF3, HNF4 and CEBP (liver) and STAT (ORegAnno Stat1), even though we used all CNSs in the 10 kb upstream region. In addition, the CRM predictions mostly overlap the true enhancer, when the real regulatory sequence was in our CNS collection. Indeed, for the muscle set, in 9 of the 11 cases where the known enhancer was in our CNS set, MODULEMINER was ably to identify this region. For the liver set, MODULEMINER identified 7 out of 8 regulatory elements (data not shown).

Detection of *cis*-regulatory modules in microarray clusters

Realizing that clustering of microarray data provides a rich source of large co-expressed gene sets, where robustness to genes that are not co-regulated ("false positive genes") is critical, our sensitivity to noise analysis above encouraged us to apply MODULEMINER to microarray clusters on a larger scale. The GNF SymAtlas [31] contains expression profiles of 140 human and mouse tissues. Nelander *et al.*

[32] obtained gene clusters by hierarchically clustering this dataset, followed by a Pearson's correlation coefficient cut-off. From this clustering, we selected all clusters with at least 25 genes in our dataset (i.e. genes with at least one CNS within 10 kb 5' of the TSS). This results in 10 clusters with sizes ranging from 26 to 214 genes. Large clusters were randomly divided in a training set of 50 genes, and a test set containing the remaining genes.

As it was our goal here to identify similar *cis*-regulatory modules within a subset of the genes in each microarray cluster, we used a two-step procedure, first detecting which subset of genes potentially share *cis*-regulatory modules, and next detecting the actual *cis*-regulatory modules in their upstream regions (Figure 4A). The first step consisted of a five-fold cross-validation, where in each validation run, we used MODULEMINER to train a TRGM on four-fifth of the genes in a cluster, and next we determined which of the other one-fifth left-out genes were targets of the TRGM. If the total number of true target genes among left-out genes would not be significantly higher than random, we concluded that MODULEMINER is not able to detect similar CRMs within this cluster. If on the other hand there is a significant enrichment of these true target genes, we concluded that MODULEMINER is able to detect similar CRMs, and we use these high scoring genes in the second step. In this second step, MODULEMINER was applied to this focussed subcluster, identifying similar *cis*-regulatory modules regulating these genes. As an extra validation, LOOCV was used to confirm the presence of similar *cis*-regulatory modules, as done previously on the smooth muscle and other benchmark sets.

Application of this procedure to the microarray clusters described above resulted in successful *cis*-regulatory module detection in 9 of the 10 clusters (Table 2, Figure 4B). In each case, this success was confirmed by a LOOCV on the selected subcluster

(all AUCs were significantly above 50 %, with an average AUC of 90.3 %, Figure 4C). For the TRGMs obtained for clusters containing over 50 genes, the number of targets in the independent test set was determined. This was significantly higher than random in three of the five cases (Table 2). In total, we predicted 209 CRMs. These MODULEMINER predictions can be viewed in detail at our website.

Detection of *cis*-regulatory modules in embryonic development gene sets

In the previous section, we detected CRMs in microarray clusters expressed in different adult tissues. Next, we aimed to predict CRMs involved in embryonic development processes.

We constructed 5 gene sets involved in specific embryonic development processes, based on literature (Table 3). Contrary to the previous section, where we aimed to detect similar CRMs in a subset of the genes in the microarray clusters (using a two-step approach), here we can assume that the embryonic development gene set are more focussed, and hence we can directly apply MODULEMINER to these sets (as in our high quality smooth muscle gene set). We performed LOOCV, confirming that MODULEMINER was able to successfully detect similar CRMs in all five gene sets (Table 3).

Characterization of the *cis*-regulatory modules

The transcriptional regulatory global models that were predicted by MODULEMINER in each of the 10 microarray clusters and each of the 5 embryonic development gene sets are summarized in Tables 4 and 5. Apart from this TRGM, MODULEMINER also provides additional information characterizing the *cis*-regulatory modules. We will discuss here the results we obtained on cluster 9, which contains genes related to cardiac muscle function.

First, MODULEMINER characterizes the given input genes, retrieving descriptions and commonly used identifiers (e.g. HGNC) from the Ensembl database. In addition, the Gene Ontology (GO) terms annotated to the input genes are retrieved, and the overrepresented GO terms are reported. For the cardiac muscle subcluster, "muscle contraction" (GO:0006936), "muscle development" (GO:0007517), "organogenesis" (GO:0009887), "contractile fibre" (GO:0043292) and "regulation of heart contraction rate" (GO:0008016) were among the overrepresented GO terms.

Next, MODULEMINER determines the weight of each PWM in the transcriptional regulatory global model (see Materials and methods). By grouping similar PWMs, the weight of each trans-factor involved is determined. The cardiac muscle TRGM contains PWMs for SRF, MEF2A, myogenin, SP3, a thyroid hormone response element (all with weights of approximately 1), and a muscle TATA box (with weight approximately 0.5). MODULEMINER also displays the *cis*-regulatory modules it identifies on the input genes. Figure 4D shows this for the heart muscle genes.

As our approach uses only human and mouse sequences to model *cis*-regulatory modules, sequenced genomes of other species can be used as validation data. MODULEMINER employs the rat and dog genomes for this purpose, by checking for *cis*-regulatory modules that fit the obtained TRGM in rat-dog conserved non-coding sequences. For the cardiac muscle genes, 11 orthologs were present in our rat-dog TFBS database, 7 of which were ranked within the top 10 % of the genome ($p = 2.28 \times 10^{-5}$).

Finally, MODULEMINER selects putative new target genes of the TRGM from the complete genome. We aim to minimize noise in these target gene predictions by using network level conservation [33], particularly through phylogenetic fusion of target gene rankings. To this end, first all genes in the human-mouse transcription factor

binding site database (excluding the input genes), and all (non-input) genes in the dog-rat transcription factor binding site database are ranked separately. MODULEMINER then fuses these two rankings into one global ranking using order statistics (similar to the approach used in [23] and [34]). Among the 100 top ranking new target genes of the cardiac muscle TRGM were MYL3 ("Cardiac myosin light chain 1"), MYOD1 ("Myoblast determination protein 1"), TNNI1 ("Troponin I") and MYH3 ("Myosin heavy chain, embryonic skeletal muscle").

The results we obtained on all sets of co-expressed genes discussed in this work, can be viewed at [35].

Where are the *cis*-regulatory module predictions located?

MODULEMINER successfully detected 9 sets of similar CRMs in the 10 microarray clusters and 5 sets of similar CRMs in the 5 embryonic development gene sets. In total, 257 CRMs were predicted. In addition to this, MODULEMINER predicted 100 new target genes of each TRGM. We next used this compendium of 1657 CRMs to examine their positions relative to the TSS of the genes they regulate.

Since a gene's search space was defined as all CNSs within 10 kb 5' of the TSS, we first examined the distributions of CNS locations, as these represent the background distribution to which the CRM locations will be compared. A first important observation is that the CNSs are highly overrepresented close to the TSS, as shown in Figures 5A and 5B. The type of gene set, namely adult tissue versus embryonic development, introduces a second CNS location bias (Figure 5C). Indeed, the adult tissue CNS set is enriched in sequences close to the TSS (< 200 base pairs) ($p = 7.6 \times 10^{-16}$ by a Wilcoxon rank sum test), while the embryonic development CNS set is depleted in sequences close to the TSS and enriched in sequences further from the TSS (2000 – 4000 base pairs) ($p = 5.6 \times 10^{-7}$). When evaluating each of the gene sets

separately (Figure 5F), 8 of the 9 adult tissue CNS sets are enriched in sequences less than 200 base pairs from the TSS (in 6 cases, this was statistically significant by a Chi-square test), while all 5 embryonic development CNS sets are depleted in sequences less then 200 base pairs from the TSS (in 3 cases, this was statistically significant).

Next, we examine the location distribution of the CRMs that were identified by MODULEMINER. For adult tissue genes, CRMs are strongly overrepresented close to the TSS (Figure 5D). Sixty-three percent of these CRMs are within 200 base pairs of the TSS. In contrast, the CRMs MODULEMINER identified near the embryonic development genes are depleted close to the TSS and enriched further away (1000 – 2000 base pairs). These conclusions remain valid even when controlling for both biases mentioned above: comparing Figure 5D to Figure 5C (the predicted CRMs in Figure 5D can be considered as a selection from the CNS sets in Figure 5C), the enrichment of predicted CRMs directing expression in adult tissues close to the TSS persisted: $p = 2.6 \times 10^{-27}$ (calculated as follows: the distances to the TSS of (i) the predicted CRMs and (ii) all CNSs of the genes in the microarray clusters were ranked and the Wilcoxon rank sum test was applied). For the CRMs directing expression in embryonic development, no statistically significant deviation from random selection from the embryonic development CNS sets could be observed ($p = 0.18$). When considering the gene sets separately, in 8 microarray clusters expressed in adult tissues, CRMs are enriched in sequences close to the TSS (Figure 5G) (this was statistically significant when controlling for bias in 6 cases). In contrast, in 4 embryonic development gene sets, CRMs are depleted close to the TSS (markedly, for three of these sets, no CRMs were predicted within 200 base pairs of the TSS).

A similar difference in TSS distance distribution was also seen for the new target genes (Figure 5E). Here as well, the distances to the TSS of the CRMs predicted to direct expression in adult tissues were clearly non-randomly distributed compared to all CNSs ($p = 3.6 \times 10^{-74}$ by a Wilcoxon rank sum test). For the CRMs predicted to direct expression in embryonic development, no statistically significant difference was observed (by a Wilcoxon rank sum test). However, these sequences seem to be (slightly) depleted within 200 base pairs of the TSS ($p = 1.5 \times 10^{-4}$ by a Chi-square test). Considering each of the gene sets separately (Figure 5H), in 7 adult tissue microarray clusters, CRMs were significantly enriched within 200 base pairs of the TSS, while for 2 embryonic development gene sets, CRMs were significantly depleted close to the TSS. Although in six cases this effect was highly significant ($p < 10^{-9}$), it was smaller than the effect within the clusters (compare Figures 5D and 5E). In summary, the *cis*-regulatory modules MODULEMINER detected were non-randomly positioned in the genome. CRMs predicted to direct expression in adult tissues were highly enriched very close to the transcription start site, while CRMs predicted to direct expression in embryonic development were depleted very close to the transcription start site.

Discussion

Although the sequence of the human genome has been available for a considerable time now, our ability to chart the regions controlling gene expression is still very limited. The situation seems to improve as a function of smaller genome size. Indeed, in the *Drosophila* early segmentation network, CRMs can be predicted based on known examples [10,11]. In the yeast *Saccharomyces cerevisiae*, with an even much smaller genome, it is possible to go one step further and predict the expression of genes based only on upstream sequences [36]. Here we focus on the computational detection of CRMs in the human genome, and hence this work is a contribution in bridging this gap.

MODULEMINER detects CRMs by taking as input a set of co-expressed genes, under the assumption that a subset of these are co-regulated, and looking for a recurrent pattern of (computationally predicted) transcription factor binding sites. The advantages of this approach are that it does not require known examples and that it allows prediction of a probable function for the detected CRMs.

MODULEMINER is similar in scope to ModuleSearcher [20,28] and CREME [19]. It differs from these previous approaches in that MODULEMINER maximizes specificity for the given set of co-expressed genes by performing a whole genome optimization. Indeed, MODULEMINER optimizes the combined rankings of the given gene set in a ranking of the complete genome. In addition, this approach allows comparison between TRMs with different parameters (e.g. maximum CRM length, number of PWMs in the TRM). Therefore, MODULEMINER is able to optimize over these parameters, and hence, our approach effectively eliminates the parameters required by the previous approaches.

Other algorithms have been developed that aim to detect similar CRMs in a set of co-expressed genes that (contrary to the approaches above) do not use a library of PWMs [21,22,29,37]. Instead, these algorithms optimize, besides the combination of motifs, also the motifs themselves. Hence, these methods attempt to solve a problem with considerably higher complexity, resulting in lower performance, as confirmed by our comparison on benchmark data. Given the extremely poor performance of motif detection methods in other organisms than yeast [38], we have opted to circumvent motif optimization by using experimentally determined PWMs. Note that this decision not necessarily limits the search to known PWMs, as libraries of computationally predicted PWMs are also available (e.g. the phylofacts PWM library [39]). In addition, we believe that with the emergence of the protein binding microarray technology [40], high quality PWMs will soon become available for a large fraction of the human transcription factor repertoire. Even though the currently available libraries of experimental PWMs show high redundancy and may contain low quality PWMs, our new approach of clustering similar TRMs is able to group redundant PWMs and our validations show that in many cases a combination of five experimental PWMs can capture enough information of a CRM to yield acceptable genome-wide specificity levels.

MODULEMINER outputs the predicted CRMs, and a transcriptional regulatory global model (TRGM). This TRGM can be considered as a bag of PWMs (selected from TRANSFAC and JASPAR), with a weight associated to each PWM. Therefore, this TRGM not only predicts the transcription factors functioning in the process under study, but in addition also allows an assessment of the relative importance of each of these transcription factors.

TRGMs do not contain spatial relations between transcription factor binding sites (except for the total size of the CRMs and a Boolean parameter indicating whether different binding sites can overlap or not). Although certain spatial relations between transcription factors working in concert are known to exist (e.g. [41,42]), we did not find any reports indicating that this is the rule rather then the exception. Therefore, we reasoned that any such relationships should not be hard-coded in the TRGMs, but rather would become apparent by inspection of the predicted CRMs. Upon inspection of the predicted CRMs presented above, no such spatial relationships surfaced.

Our method for scoring a sequence using a TRM or TRGM (see Materials and methods) does not take homotypic clustering of transcription factor binding sites into account (like HMM based methods do [15,17,43]). However, this cooperative binding of one transcription factor can nevertheless be modelled in our framework by the construction of a TRM or TRGM that contains multiple instances of the same PWM. Therefore, if multiple instances of a specific transcription factor are important for the regulation of a set of co-regulated genes, this is represented accordingly in the optimal model. For example, when applying MODULEMINER to the tightly co-expressed set of smooth muscle markers, the transcription factor SRF occurs 2 or 3 times in each of the TRMs in the resulting TRGM, suggesting an extensive cooperation between SRF binding sites for smooth muscle specific transcription regulation. In contrast, the SMAD4, SP1 and ATF3 PWMs occur exactly once in 97.5 % of the TRMs (SMAD4 and SP1 occur twice in 1.5 % and 1 % of the TRMs respectively).

MODULEMINER takes the genomic background sequence into account in two ways. Firstly, a third order background model is used in the process of annotating putative transcription factor binding sites. Secondly, our optimization strategy selects the TRM (or TRGM) that optimally separates the given genes (sequences) from all other genes

in the genome. Hence, our system corrects both for local sequence properties (by the third order background model) as for more global sequence properties (by selecting against combinations of transcription factor binding sites that occur independently of the given sequences).

We included all CNSs up to 10 kb 5' of the transcription start site in our pipeline. Although this choice is inherently arbitrary, it is motivated by the following arguments: (i) sequences 3' of the transcription start site might harbour translational regulatory signals, which we do not want to model here; (ii) potential regulatory sequences far upstream can be difficult to assign to a target gene; (iii) selecting 10 kb 5' of the transcription start site has proven to be valuable in our previous study [20], and others have made similar choices as well [44]; (iv) in a previous study where CRMs were predicted in an unbiased way across the complete human genome [8], it was shown that CRMs are highly depleted between 10 kb and 30 kb 5' of the transcription start site.

The validation framework we use, combining genome-wide ranking with leave-one-out cross-validation, could also be useful in evaluating or comparing hypotheses regarding the working principles of transcription regulation, and in this regard can be considered similar in scope to CodeFinder [24]. In this work, two such tests are implicitly performed: (i) CRMs driving a tissue-specific expression pattern are compared to CRMs driving an embryonic development expression pattern and (ii) by the comparison of the three sets of putative transcription factor binding sites (e.g. Figure 1, Figure 3J, Figure 4B), the importance of binding site preservation is evaluated as well as the impact of a correction for differences in transcription start sites between human and mouse.

Construction of a high-quality set of co-regulated genes involved in a certain process under study is not always straightforward. In this regard, robustness to noise in a set of putative co-expressed genes is highly desirable in an algorithm to detect similar CRMs. We found MODULEMINER to be highly robust to the quality of this input gene set. Indeed, in our experiments with smooth muscle marker genes, we observed that even when only 10 of 50 given genes are really co-regulated, MODULEMINER was still able to pick up the correct signal (Figure 2). These properties of MODULEMINER prompted us to apply the algorithm to gene sets obtained from clustering microarray data. In 9 out of 10 microarray clusters, MODULEMINER succeeded in finding similar CRMs in a subset of the genes. Perhaps unsurprisingly, a critical mass of co-regulated genes is required for MODULEMINER to detect similar CRMs. However, this minimum required number of co-regulated genes is sufficiently small so as not to preclude application of the algorithm. This is illustrated both by our results obtained on the smooth muscle genes (Figure 2), and by the successful CRM detection in two small heart development gene sets (Table 3).

Application of MODULEMINER to the smooth muscle marker genes resulted in CRMs with multiple binding sites for SRF, and with single binding sites for SMAD4, SP1 and ATF3. Both SRF and SP1 have been shown to play a role in regulating smooth muscle specific expression [26]. Furthermore, SMADs are effectors of the TGF-β signalling pathway, and have been shown to work in concert with SRF to control smooth muscle cell differentiation [45]. MODULEMINER identified transcription factors known to play a key role in other co-expressed gene sets as well. Examples are GATA-factors, NFATs and HAND1 in heart development, HNF-1 and HNF-4 in liver-specific gene expression, PU.1 in lymphocyte specific gene expression, and

Myogenin, SRF, the thyroid hormone receptor, and MEF2 in heart specific gene expression.

Imposing trans-factor conservation by motif preservation between human and mouse sequences of a CNS significantly improved the performance of MODULEMINER on the set of smooth muscle marker genes. A similar approach has also been shown to improve CRM detection performance in the *Drosophila* early segmentation gene network [10]. When we applied MODULEMINER to the microarray clusters and the embryonic development gene sets, in some cases this trans-factor conservation also increased performance (microarray clusters 6, 7 and 9, and the neural crest cell gene set), but in other cases it did not.

Correcting for possible differences in transcription start site in human and mouse by a three-step alignment procedure (see Materials and methods), resulted in increased performance for most of the microarray clusters, but not for the development gene sets. This marked difference may be related to the different locations of the detected CRMs in these two different systems.

We observed a significant difference in the locations of the CRMs MODULEMINER predicted to direct expression in adult tissues and the CRMs MODULEMINER predicted to direct expression in embryonic development. CRMs driving tissue-specific expression are highly overrepresented within 200 base pairs of the TSS. In contrast, CRMs driving expression in embryonic development are more evenly distributed in the 10 kb sequences we considered, and seem to be underrepresented within 200 base pairs of the TSS. These results suggest that transcription regulation of tissue-specific expression is mainly exerted by proximal promoters, while transcription regulation of expression during embryonic development seems to be mainly exerted by more distal enhancers.

MODULEMINER can be applied to 3 conceptually different tasks: (i) prediction of transcription factors that play a role in regulating a set of co-regulated genes, (ii) prediction of regulatory regions and (iii) predictions of new target genes of a TRGM. It is important to realize that the accuracy of the predictions differs between those tasks. Although exact performance statistics can only be obtained through the careful experimental testing of our predictions, which is outside the scope of the present study, the results we obtained in this work can be used to provide rough estimates of the predictive accuracy. When we applied MODULEMINER to the two well-studied benchmark sets, we obtained HNF1, CEBP, HNF3, GATA1, PAX6 and HNF4 for the liver benchmark set, and MZF1, PPARγ, SRF, MEF2, the Epstein-Barr virus transcription factor R, MYF and MYOD for the muscle benchmark set. Comparing this to literature [4,46] and to the PWM libraries we use, we obtain a sensitivity of 70 % (7 out of 10 known PWMs are recovered), a specificity of 99.6 % (630 of 633 (liver) and 619 of 621 (muscle) likely incorrect PWMs are rejected) and a positive predictive power of 62 % (8 of 13 total predicted PWMs are correct). These values need to be regarded with some reservations when extrapolating to other cases, since both liver and muscle are well-studied systems with high quality PWMs available. Nevertheless, we can conclude that MODULEMINER is quite accurate in selecting PWMs/transcription factors that play a key role in the regulation of the genes under study. Regarding detection of regulatory sequences, MODULEMINER was able to detect 16 of 24 known muscle/liver enhancers, when a total of 24 predictions where made. This is a sensitivity of 67 % and a positive predictive power of 67 %, although we emphasize that this last value is an underestimate as some of our predictions may be yet unknown enhancers. Notwithstanding some reservations on extrapolating these data, we conclude that the predictive accuracy of MODULEMINER for detection of

regulatory regions (CRMs) near a set of co-regulated genes is quite high. Regarding the predictive accuracy of MODULEMINER for the detection of new target genes given a TRGM, the results of our LOOCV procedure can provide some estimates. From the resulting ROC curves, one can see that for a sensitivity of 50 %, the specificity is about 90 %, and for a sensitivity of 80 %, specificity is about 80 %, although the differences between different gene sets can be large. However, typically only a few dozen new target genes can be tested, and thus specificity may not be high enough to select the right targets from the complete genome. In our previous study [23], we confirmed that the predictive accuracy of new target genes is quite low, although we showed it to be detectably present. We note that in that study, we used our previous ModuleSearcher algorithm which was shown here to have a lower performance than MODULEMINER. In addition, MODULEMINER's use of network level conservation between human/mouse and rat/dog predictions of new target genes might increase performance. Finally, the results we obtained in the transcription start site distribution of the CRMs predicted near the new target genes are consistent with these performance predictions: Figures 5E and 5H show a similar trend as Figures 5D and 5G, but to a lesser extend, hence pointing to a substantial amount of noise, but also indicating that a signal can be picked up, even in a whole genome scan.

Conclusions

We present MODULEMINER, the first algorithm to detect similar *cis*-regulatory modules in the human genome that is based on whole-genome optimization. MODULEMINER is generally applicable, and outperforms other similar approaches to detect CRMs on benchmark data. In addition, MODULEMINER can detect similar CRMs in noisy sets of co-expressed genes, such as microarray clusters. We successfully applied the algorithm to sets of genes expressed in adult tissues and sets of genes expressed in embryonic development processes. We show that CRMs predicted to regulate genes expressed in adult tissues are highly overrepresented within 200 base pairs of the transcription start site, while CRMs predicted to regulate genes involved in embryonic development processes are depleted within this region. These findings suggest that expression in adult tissues is mainly directed by proximal promoters, while expression in embryonic development is more often regulated by distal enhancers.

Materials and methods

Construction of 3 sets of candidate transcription factor binding sites

We constructed three sets of genome-wide candidate transcription factor binding sites in human-mouse conserved non-coding sequences (CNSs). The first set contains all predicted binding sites in all CNSs. Sequences 10 kb 5' (+ 50 bp 3') of the transcription start site of all human genes and their mouse orthologs were obtained from Ensembl (version 36). When another gene was encountered, only the sequence up to that gene was included. Conserved non-coding sequences were selected by LAGAN alignments [47]. Thresholds were set to 75 % conservation over at least 100 base pairs. Transcription factor binding site predictions were performed using MotifScanner [48], with the prior set to 0.2. Both TRANSFAC [49] (version 9.4) and JASPAR [39] were used as PWM libraries.

The second set aims to restrict the candidate binding sites by enforcing that the regulatory factors should be conserved. This is achieved by selecting only binding sites in each human region for transcription factors for which we also detect binding sites in the orthologous mouse region (preserved sites). We note that this constraint does not impose that the binding sites should be conserved, nor that they should align.

In the construction of the third set we aimed to correct for differences in human and mouse transcription start sites (TSSs), and for possible annotation errors of TSSs. To this end, we extended the mouse sequences used in the alignments by 100 kb in both directions. Alignment errors were kept in check by applying a multi-step alignment procedure. The human 10 kb sequence was aligned to (A) the 10 kb mouse sequence, (B) the mouse sequence extended by 10 kb in both directions, and (C) the mouse sequence extended by 100 kb in both directions. If in alignment (A) CNSs were

predicted, we assumed that the correct orthologous region in the mouse is not off by more then 10 kb, and hence we used the CNSs from alignment (A), supplemented by all additional CNSs from alignment (B). CNSs that were truncated in alignment (A) because they extended over the sequence borders, were replaced by their counterpart from alignment (B). If in alignment (A) no CNSs were predicted, we reasoned that the correct orthologous region in the mouse might be off by more then 10 kb, and we used the CNSs from alignment (C). Here also, for each CNS (in human), we selected only preserved binding sites.

The same procedure was used with the dog and rat sequences to create sets of candidate transcription factor binding sites corresponding to the three human-mouse sets. As neither dog nor rat could serve as a reference species, we did not extend the sequences in the dog-rat candidate transcription factor binding site set that corresponds to human-mouse set 3.

Transcriptional regulatory models

We model similar *cis*-regulatory modules in a set of co-expressed genes by transcriptional regulatory models. These TRMs are parameterized as in [20]. A TRM is a combination of PWM instances (up to 6), supplemented by three parameters: (i) the maximum length of *cis*-regulatory modules, (ii) a Boolean parameter stating whether different binding sites can overlap or not, and (iii) a Boolean parameter that indicates whether incomplete modules will be penalized or not. Given a TRM and a sequence, a score S_{seq} can be calculated, as detailed in [20]. A TRM may contain multiple instances of one specific PWM: in the calculation of S_{seq}, each PWM in the TRM is matched to at most one binding site – thus if a PWM occurs twice, up to two binding sites for the corresponding transcription factor can be taken into account. We assign a score S_g to a gene by taking the maximum of S_{seq} for all CNSs of that gene.

The S_g scores for the given set of co-regulated genes are used to determine a 'fitness score' of a TRM. This fitness score of a TRM for a given set of co-expressed genes is determined by the positions of the co-expressed genes in a ranking of S_g for all genes in the genome. We use order statistics to assign a probability to the combination of ranks of the given co-expressed genes (using the numerical approach detailed in [23]). Hence, the resulting p-value represents how well that TRM models the given set of co-expressed genes, compared to all other genes in the genome. We use 1 minus that p-value as the fitness score for the TRM.

The MODULEMINER algorithm

MODULEMINER uses a genetic algorithm to find the TRM with the optimal fitness score. At the onset, a starting population of TRMs is obtained by running our ModuleSearcher algorithm [28] using many different combinations of parameters. This initial step is not absolutely required (one can start from a population of randomly generated CRMs), but it provides a speed advantage. These TRMs obtained by ModuleSearcher are assigned a fitness score, and the 200 best scoring TRMs are retained as starting population for the ModuleMiner genetic algorithm. During each 'generation' of the algorithm, 200 new individuals (TRMs) are generated (based on the TRM population at that time) and added to the population. This population of 400 TRM is then required to compete (by fitness score), and the 200 best scoring TRMs are retained. This procedure is repeated until the stop criterion is reached (at least 300 generations and at most 1000 generations). Generation of new individuals (TRMs) is done using 2 'parent' TRMs randomly selected from the population. Each of the TRM parameters (number of PWMs, length, overlap and penalization) is determined by random selection from both parents, allowing a small probability of mutation (i.e. each parameter is set to a random value with a probability of 0.1). Subsequently,

PWMs are selected at random from both parents. Here as well, each PWM can be 'mutated' (replaced by a PWM randomly selected from TRANSFAC and JASPAR) with a probability of 0.1. As stop criterion, we use homogeneity of the population: if more than 80 % of the TRMs can be grouped into one TRGM (see below) and at least 300 generations have passed, the algorithm is stopped. If this stop criterion is not reached, the algorithm is stopped after 1000 generations. The parameters of the ModuleMiner genetic algorithm (e.g. population size, mutation probability, …) were selected by optimizing for speed. The convergence of the algorithm is highly insensitive to these parameters over a wide range, and sensitivity of speed to these parameter settings is also limited (data not shown).

Transcriptional regulatory global models

Aiming to minimize the sensitivity of our models of similar *cis*-regulatory modules to noise in transcription factor binding site predictions, we constructed composite models (TRGMs) from multiple high-scoring TRMs. To this end, similar TRMs are clustered, and the largest cluster is returned as resulting TRGM. TRMs were clustered when the *cis*-regulatory modules they predict near the high scoring genes (out of the given set of co-expressed genes) occur in the same CNS. As a cut-off for determining which genes are among the "high scoring genes", we used the top 2.5 % in a ranking of the complete genome.

Scoring a sequence with a TRGM is performed by scoring this sequence for each TRM within the TRGM, subsequently normalizing this score (maximum CNS score = 1), and finally adding the normalized TRM scores.

As a TRGM is a collection of TRMs and TRMs each contain a collection of PWM instances, TRGMs are also collections of PWMs. In addition, a weight can be assigned to each PWM in the TRGM, quantifying the significance of the PWM for the

process under study. This weight of a PWM is calculated as follows: for each TRM in the TRGM, the number of instances of that PWM is counted, and this number is averaged over all the TRMs in the TRGM.

Performance comparison on benchmark data

Four benchmark data sets containing annotated regulatory regions directing expression in a particular system were selected from PAZAR [27]. We selected all human genes (or human orthologs) from each of these 'boutiques'. The regulatory sequence search space was defined as all CNSs within 10 kb 5' of the TSS (as throughout our study). We used this search space for all algorithms, except CREME [19], where only the online version was available that by default uses one CNS within 1.5 kb of the TSS. As the other CRM detection algorithms had multiple parameters (absent in MODULEMINER), these parameters were set to default options. For the ModuleSearcher algorithm [28], we used the same parameters as in the cell cycle case study reported [20]. For CisModule [22] and EMCMODULE [29] we used the default parameter settings. We used Clover [30] as follows: for each PWM found overrepresented, we constructed a TRM (with parameters: no overlap between binding sites, no penalization and a maximum distance of 1000 bp), and this way we constructed TRGMs containing enriched PWMs reported by Clover. We also generated 100 random TRMs (combinations of 3-6 PWMs with randomly generated parameters) and we used these to rank the genes of each benchmark set, as a proxy for a method unable to detect similar CRMs.

Availability

MODULEMINER can be accessed at our website [35]. A stand-alone version is available upon request.

List of abbreviations

AUC: area under the curve

CNS: conserved non-coding sequence

CRM: *cis*-regulatory module

GO: Gene Ontology

LOOCV: leave-one-out cross-validation

PWM: position weight matrix

ROC: receiver operator characteristic

TFBS: transcription factor binding site

TRGM: transcriptional regulatory global model

TRM: transcriptional regulatory model

TSS: transcription start site

Authors' contributions

PVL, SA and PM conceived and designed the experiments; PVL performed the experiments; PVL and SA analyzed the data; PVL, SA, BT, BDM and YM contributed reagents/materials/analysis tools; PVL wrote the paper; all authors read and approved the final manuscript.

Acknowledgements

PVL is supported by a PhD fellowship and SA by a postdoc fellowship of the Research Foundation – Flanders (FWO). BT is supported by a PhD fellowship of the IWT. Research performed by YM and BDM is supported by: (i) Research Council KUL: GOA AMBioRICS, CoE EF/05/007 SymBioSys, several PhD/postdoc & fellow grants; (ii) Flemish Government: FWO: PhD/postdoc grants, projects G.0241.04 (Functional Genomics), G.0499.04 (Statistics), G.0232.05 (Cardiovascular), G.0318.05 (subfunctionalization), G.0553.06 (VitamineD), G.0302.07 (SVM/Kernel), research communities (ICCoS, ANMMM, MLDM); IWT: PhD Grants, GBOU-McKnow-E (Knowledge management algorithms), GBOU-ANA (biosensors), TAD-BioScope-IT, Silicos; SBO-BioFrame; Belgian Federal Science Policy Office: IUAP P6/25 (BioMaGNet, Bioinformatics and Modeling: from Genomes to Networks, 2007-2011); EU-RTD: ERNSI: European Research Network on System Identification; FP6-NoE Biopattern; FP6-IP e-Tumours, FP6-MC-EST Bioptrain, FP6-STREP Strokemap.

This research was conducted utilizing high performance computational resources provided by the University of Leuven.

We are grateful to Irina Balikova, Boyan Dimitrov and Koen Devriendt for help with the construction of the development gene sets. The authors declare no competing financial interests.

References

1. Davidson EH: *Genomic Regulatory Systems: Development and Evolution*. San Diego, USA: Academic Press; 2001.

2. Balmer JE, Blomhoff R: **Anecdotes, data and regulatory modules.** *Biol Lett* 2006, **2:** 431-434.

3. Wasserman WW, Sandelin A: **Applied bioinformatics for the identification of regulatory elements.** *Nat Rev Genet* 2004, **5:** 276-287.

4. Wasserman WW, Fickett JW: **Identification of regulatory regions which confer muscle-specific gene expression.** *J Mol Biol* 1998, **278:** 167-181.

5. Elnitski L, Hardison RC, Li J, Yang S, Kolbe D, Eswara P, O'Connor MJ, Schwartz S, Miller W, Chiaromonte F: **Distinguishing regulatory DNA from neutral sites.** *Genome Res* 2003, **13:** 64-72.

6. 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:** 1034-1050.

7. Hallikas O, Palin K, Sinjushina N, Rautiainen R, Partanen J, Ukkonen E, Taipale J: **Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity.** *Cell* 2006, **124:** 47-59.

8. Blanchette M, Bataille AR, Chen X, Poitras C, Laganiere J, Lefebvre C, Deblois G, Giguere V, Ferretti V, Bergeron D, Coulombe B, Robert F: **Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression.** *Genome Res* 2006, **16:** 656-668.

9. Yuh CH, Brown CT, Livi CB, Rowen L, Clarke PJ, Davidson EH: **Patchy interspecific sequence similarities efficiently identify positive cis-regulatory elements in the sea urchin.** *Dev Biol* 2002, **246:** 148-161.

10. Berman BP, Pfeiffer BD, Laverty TR, Salzberg SL, Rubin GM, Eisen MB, Celniker SE: **Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in Drosophila melanogaster and Drosophila pseudoobscura.** *Genome Biol* 2004, **5:** R61.

11. Schroeder MD, Pearce M, Fak J, Fan H, Unnerstall U, Emberly E, Rajewsky N, Siggia ED, Gaul U: **Transcriptional control in the segmentation gene network of Drosophila.** *PLoS Biol* 2004, **2:** E271.

12. Nobrega MA, Ovcharenko I, Afzal V, Rubin EM: **Scanning human gene deserts for long-range enhancers.** *Science* 2003, **302:** 413.

13. 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:** e7.

14. Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD, Plajzer-Frick I, Akiyama J, De VS, Afzal V, Black BL, Couronne O, Eisen MB, Visel A, Rubin EM: **In vivo enhancer analysis of human conserved non-coding sequences.** *Nature* 2006, **444:** 499-502.

15. Rajewsky N, Vergassola M, Gaul U, Siggia ED: **Computational detection of genomic cis-regulatory modules applied to body patterning in the early Drosophila embryo.** *BMC Bioinformatics* 2002, **3:** 30.

16. Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, Levine M, Rubin GM, Eisen MB: **Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome.** *Proc Natl Acad Sci U S A* 2002, **99:** 757-762.

17. Sinha S, van NE, Siggia ED: **A probabilistic method to detect regulatory modules.** *Bioinformatics* 2003, **19**(Suppl 1)**:** i292-i301.

18. Halfon MS, Grad Y, Church GM, Michelson AM: **Computation-based discovery of related transcriptional regulatory modules and motifs using an experimentally validated combinatorial model.** *Genome Res* 2002, **12:** 1019-1028.

19. Sharan R, Ovcharenko I, Ben-Hur A, Karp RM: **CREME: a framework for identifying cis-regulatory modules in human-mouse conserved segments.** *Bioinformatics* 2003, **19**(Suppl 1)**:** i283-i291.

20. Aerts S, Van Loo P, Thijs G, Moreau Y, De Moor B: **Computational detection of *cis*-regulatory modules.** *Bioinformatics* 2003, **19**(Suppl 2)**:** II5-II14.

21. Thompson W, Palumbo MJ, Wasserman WW, Liu JS, Lawrence CE: **Decoding human regulatory circuits.** *Genome Res* 2004, **14:** 1967-1974.

22. Zhou Q, Wong WH: **CisModule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling.** *Proc Natl Acad Sci U S A* 2004, **101:** 12114-12119.

23. Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, De Smet F, Tranchevent LC, De Moor B, Marynen P, Hassan B, Carmeliet P, Moreau Y: **Gene prioritization through genomic data fusion.** *Nat Biotechnol* 2006, **24:** 537-544.

24. Philippakis AA, Busser BW, Gisselbrecht SS, He FS, Estrada B, Michelson AM, Bulyk ML: **Expression-guided in silico evaluation of candidate cis regulatory codes for Drosophila muscle founder cells.** *PLoS Comput Biol* 2006, **2:** e53.

25. Nelander S, Mostad P, Lindahl P: **Prediction of cell type-specific gene modules: identification and initial characterization of a core set of smooth muscle-specific genes.** *Genome Res* 2003, **13:** 1838-1854.

26. Kumar MS, Owens GK: **Combinatorial control of smooth muscle-specific gene expression.** *Arterioscler Thromb Vasc Biol* 2003, **23:** 737-747.

27. Portales-Casamar E, Kirov S, Lim J, Lithwick S, Swanson MI, Ticoll A, Snoddy J, Wasserman WW: **PAZAR: a framework for collection and dissemination of cis-regulatory sequence annotation.** *Genome Biol* 2007, **8:** R207.

28. Aerts S, Van Loo P, Moreau Y, De Moor B: **A genetic algorithm for the detection of new cis-regulatory modules in sets of coregulated genes.** *Bioinformatics* 2004, **20:** 1974-1976.

29. Gupta M, Liu JS: **De novo cis-regulatory module elicitation for eukaryotic genomes.** *Proc Natl Acad Sci U S A* 2005, **102:** 7079-7084.

30. Frith MC, Fu Y, Yu L, Chen JF, Hansen U, Weng Z: **Detection of functional DNA motifs via statistical over-representation.** *Nucleic Acids Res* 2004, **32:** 1372-1381.

31. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB: **A gene atlas of the mouse and human protein-encoding transcriptomes.** *Proc Natl Acad Sci U S A* 2004, **101:** 6062-6067.

32. Nelander S, Larsson E, Kristiansson E, Mansson R, Nerman O, Sigvardsson M, Mostad P, Lindahl P: **Predictive screening for regulators of conserved functional gene modules (gene batteries) in mammals.** *BMC Genomics* 2005, **6:** 68.

33. Pritsker M, Liu YC, Beer MA, Tavazoie S: **Whole-genome discovery of transcription factor binding sites by network-level conservation.** *Genome Res* 2004, **14:** 99-108.

34. Aerts S, van HJ, Sand O, Hassan BA: **Fine-Tuning Enhancer Models to Predict Transcriptional Targets across Multiple Genomes.** *PLoS ONE* 2007, **2:** e1115.

35. **ModuleMiner: computational detection of *cis*-regulatory modules** [http://www.esat.kuleuven.be/moduleminer]

36. Beer MA, Tavazoie S: **Predicting gene expression from sequence.** *Cell* 2004, **117:** 185-198.

37. Grad YH, Roth FP, Halfon MS, Church GM: **Prediction of similarly acting cis-regulatory modules by subsequence profiling and comparative genomics in Drosophila melanogaster and D.pseudoobscura.** *Bioinformatics* 2004, **20:** 2738-2750.

38. Tompa M, Li N, Bailey TL, Church GM, De MB, 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 HJ, 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:** 137-144.

39. Vlieghe D, Sandelin A, De Bleser PJ, Vleminckx K, Wasserman WW, van RF, Lenhard B: **A new generation of JASPAR, the open-access repository for transcription factor binding site profiles.** *Nucleic Acids Res* 2006, **34:** D95-D97.

40. Mukherjee S, Berger MF, Jona G, Wang XS, Muzzey D, Snyder M, Young RA, Bulyk ML: **Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays.** *Nat Genet* 2004, **36:** 1331-1339.

41. Fickett JW: **Coordinate positioning of MEF2 and myogenin binding sites.** *Gene* 1996, **172:** GC19-GC32.

42. Papatsenko D, Levine M: **A rationale for the enhanceosome and other evolutionarily constrained enhancers.** *Curr Biol* 2007, **17:** R955-R957.

43. Frith MC, Li MC, Weng Z: **Cluster-Buster: Finding dense clusters of motifs in DNA sequences.** *Nucleic Acids Res* 2003, **31:** 3666-3668.

44. Chang LW, Nagarajan R, Magee JA, Milbrandt J, Stormo GD: **A systematic model to predict transcriptional regulatory mechanisms based on**

overrepresentation of transcription factor binding profiles. *Genome Res* 2006, **16:** 405-413.

45. Nishimura G, Manabe I, Tsushima K, Fujiu K, Oishi Y, Imai Y, Maemura K, Miyagishi M, Higashi Y, Kondoh H, Nagai R: **DeltaEF1 mediates TGF-beta signaling in vascular smooth muscle cell differentiation.** *Dev Cell* 2006, **11:** 93-104.

46. Krivan W, Wasserman WW: **A predictive model for regulatory sequences directing liver-specific transcription.** *Genome Res* 2001, **11:** 1559-1566.

47. Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, Green ED, Sidow A, Batzoglou S: **LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA.** *Genome Res* 2003, **13:** 721-731.

48. Aerts S, Thijs G, Coessens B, Staes M, Moreau Y, De Moor B: **Toucan: deciphering the cis-regulatory logic of coregulated genes.** *Nucleic Acids Res* 2003, **31:** 1753-1764.

49. Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, Kloos DU, Land S, Lewicki-Potapov B, Michael H, Munch R, Reuter I, Rotert S, Saxel H, Scheer M, Thiele S, Wingender E: **TRANSFAC: transcriptional regulation, from patterns to profiles.** *Nucleic Acids Res* 2003, **31:** 374-378.

50. Buckingham M, Meilhac S, Zaffran S: **Building the mammalian heart from two sources of myocardial cells.** *Nat Rev Genet* 2005, **6:** 826-835.

51. Stoller JZ, Epstein JA: **Cardiac neural crest.** *Semin Cell Dev Biol* 2005, **16:** 704-715.

52. Graw J: **The genetic and molecular basis of congenital eye defects.** *Nat Rev Genet* 2003, **4:** 876-888.

53. Epstein CJ, Erickson BP, Wynshaw-Boris A: *Inborn Errors of Development: The Molecular Basis of Clinical Disorders of Morphogenesis.* New York, USA: Oxford University Press; 2004.

Figure legends

Figure 1 - Performance of MODULEMINER on a set of smooth muscle marker genes, using the three different sets of candidate transcription factor binding sites

ROC curves are shown, representing results for leave-one-out cross-validations on the set of smooth muscle markers, (A) using singular transcriptional regulatory models (TRMs) and (B) using transcriptional regulatory global models (TRGMs).

Figure 2 - Sensitivity of MODULEMINER's performance to the quality of the input genes

The ratio of true positive genes (containing the smooth muscle CRM) to negative genes (approximated by random genes) was varied. Each time, a leave-one-out cross-validation was performed, an ROC curve was constructed, and the area under the curve was calculated. These AUCs were plotted as a function of the ratio negative genes/positive genes. As an AUC of 50 % signifies random ordering of the left-out genes (and hence indicates that no CRMs can be detected), this value was taken as the origin on the Y-axis. Blue: the number of positive genes was kept constant at 10, and the number of negative genes was varied. Red: the total number of genes was kept constant at 10, and the ratio negative genes/positive genes was varied.

Figure 3 - Comparison to other CRM detection algorithms

(A-E) ROC curves for the LOOCV using MODULEMINER, ModuleSearcher, CisModule, EMCMODULE, Clover and random TRMs for each of the 5 benchmark sets: (A) ORegAnno Erythroid, (B) liver, (C) muscle, (D) ORegAnno Stat1 and (E) smooth muscle. (F-I) ROC curves when using TFBS preservation (TFBS set 2) in the genome ranking step for all algorithms, on the 4 benchmark sets that performed above

random: (F) liver, (G) muscle, (H) ORegAnno Stat1 and (I) smooth muscle. (J) MODULEMINER performance for the three TFBS sets on the muscle benchmark data.

Figure 4 - Application of MODULEMINER to microarray clusters

(A) The 2-step procedure used to detect similar *cis*-regulatory modules in a subset of genes within a given microarray cluster. In the first step, a five-fold cross-validation is performed, and the number of left-out genes considered as target genes is counted. If this number is significantly more than expected under a random distribution of the ranks, these genes are transferred to the second step. In this second step, MODULEMINER is used to model the similar *cis*-regulatory modules regulating the genes in this focused subcluster. (B) Results of the first step of the procedure in (A) for the 10 microarray clusters and the three different sets of candidate transcription factor binding sites. Significantly higher numbers of target genes among the left-out genes than randomly expected are depicted by an asterisk. Clusters 7 and 10 only contained sufficient genes (≥ 25) in transcription factor binding site set 3 and therefore are omitted for the other two sets. (C) Leave-one-out cross-validation results on the subclusters with a significant enrichment of target genes from (B). Each left-out gene was ranked using the TRGM obtained on the remaining genes. Next, sensitivity/specificity pairs where calculated for different detection thresholds, and these were used to construct ROC curves. The area under these ROC curves (AUC) was calculated and is depicted here. Colors: as in (B). (D) Example of a set of similar *cis*-regulatory modules identified by MODULEMINER. These results were obtained on the cardiac muscle genes by the procedure depicted in (A). Each horizontal line represents a human-mouse conserved non-coding sequence upstream of a gene within the cluster. The different colored boxes represent binding sites of different

transcription factors. Detailed results, including descriptions of the genes shown, and the exact positions of the CNSs are available at [35].

Figure 5 - Distribution of distance to transcription start site for CNSs and predicted *cis*-regulatory modules

(A) All human-mouse CNSs in TFBS sets 1 and 2 (both are based on the same set of CNSs), and in TFBS set 3. (B) The distribution from (A), when divided in 6 unequal bins. (C) Distribution of all conserved non-coding sequences upstream of genes within the microarray clusters (of genes expressed in different adult tissues) and the embryonic development gene sets, where CRMs could successfully be detected (Tables 2 and 3), divided in the same 6 bins as under (B). (D) Distribution of the distance to transcription start for the *cis*-regulatory modules MODULEMINER identified near the genes from (C). (E) Distribution of distance to transcription start for the *cis*-regulatory modules MODULEMINER identified in a whole genome scan (genes in (D) where removed, such that only new target genes where represented here). Note that (B), (C), (D) and (E) are drawn to the same scale. (F) Portion of CNSs near the genes in the different microarray clusters and embryonic development sets that is located within 200 bp of the transcription start site. (G) Portion of predicted CRMs near the genes in the different microarray clusters and embryonic development sets that is located within 200 bp of the transcription start site. (H) Portion of CRMs, predicted in a whole genome scan for the TRGM built for the different gene sets that is located within 200 bp of the transcription start site. The blue line in (F), (G) and (H) indicates the portion of all CNSs (within 10 kb 5' of all human genes) that is less then 200 bp of the transcription start site.

Tables

Table 1 - Genome-wide databases of candidate transcription factor binding sites

Nr	Database properties	Nr genes	Nr regions	Nr binding sites
1	human-mouse conserved regions, 10 kb 5' of TSS	8759	22582	1858800
2	(1) + limited to binding sites occurring both in the human and mouse CNS	8759	22582	878338
3	(2) + correct for possible mouse TSS differences (add 100 kb of mouse sequence 5' and 3')	11653	35021	1316927

Table 2 - Summary of MODULEMINER's results for the 10 microarray clusters

Cluster	Annotation	TFBS set	Nr target genes after cross-validation (p-val)	AUC on target genes	Nr target genes in independent test set (p-val)	Total nr of CRMs
1	Protein synthesis	1	10 / 50 (p = 0.025)	0.96	14 / 123 (p = 0.35)	30
2	Oocyte / fertilized egg	3	10 / 50 (p = 0.025)	0.98	30 / 164 ($p = 8.6 \times 10^{-4}$)	43
3	Neural tissues	3	10 / 50 (p = 0.025)	0.84	15 / 122 (p = 0.24)	29
4	Lymphocytes	3	10 / 50 (p = 0.025)	0.87	23 / 85 ($p = 7.0 \times 10^{-6}$)	36
5	Testis / spermatogenesis	-	-	-	-	-
6	Liver	3	14 / 50 ($p = 2.9 \times 10^{-4}$)	0.93	7 / 29 (p = 0.022)	23
7	Mitochondrion	3	9 / 31 (p = 0.0026)	0.87	-	12
8	Extracellular matrix	2	7 / 32 (p = 0.036)	0.92	-	10
9	Cardiac muscle	3	17 / 32 ($p = 6.6 \times 10^{-10}$)	0.95	-	16
10	Energy metabolism	3	7 / 26 (p = 0.012)	0.82	-	10

TFBS: transcription factor binding site. TFBS sets: set 1: human-mouse CNSs, 10 kb 5' of TSS; set 2: set 1 + binding site preservation; set 3: set 2 + correction for TSS differences. For clusters where multiple TFBS sets resulted in successful CRM detection, only the result showing the best cross-validation performance is shown. Genes (in the cluster) that by cross-validation where ranked within the top 10 % of the genome where considered target genes of the TRGM. The total number of CRMs constitutes all successful CRM predictions near genes in the cluster. CRM predictions were considered successful if the TRGM score was sufficient to rank the target gene within the top 10 % of the genome. In some cases, multiple CRMs are found controlling the same target gene.

Table 3 - Summary of MODULEMINER's results for the 5 embryonic development gene sets

Embryonic development process	TFBS set	Nr target genes after leave-one-out cross-validation (p-val)	AUC
Primary heart field [50]	1	6 / 7 ($p = 6.4 \times 10^{-6}$)	0.92
Secondary heart field [50]	1	6 / 9 ($p = 6.4 \times 10^{-5}$)	0.79
Neural crest cells [51]	2	6 / 10 ($p = 1.5 \times 10^{-4}$)	0.86
Eye development [52]	1	10 / 15 ($p = 1.9 \times 10^{-7}$)	0.79
Limb development [53]	1	10 / 24 ($p = 5.2 \times 10^{-5}$)	0.77

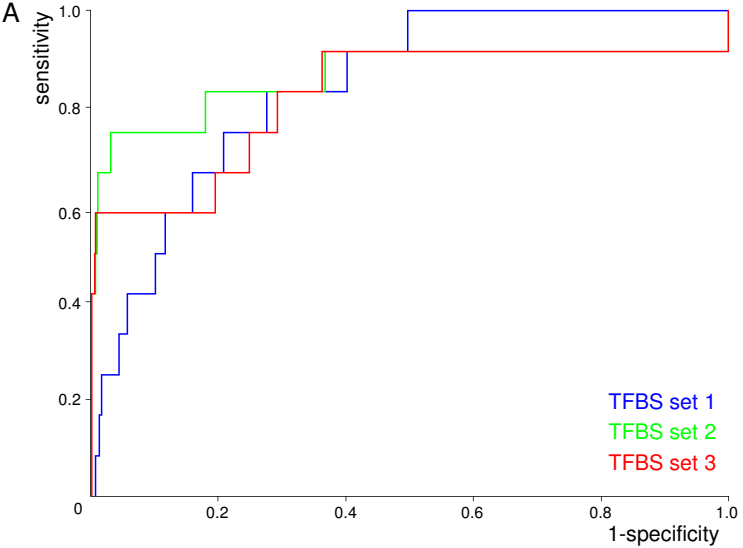
A key review or book used as a basis for construction of the development gene set is given in the first column. The genes in each set, as well as the detailed results can be viewed at our website [35]. TFBS: transcription factor binding site. TFBS sets: set 1: human-mouse CNSs, 10 kb 5' of TSS; set 2: set 1 + binding site preservation; set 3: set 2 + correction for TSS differences. For clusters where multiple TFBS sets resulted in successful CRM detection, only the result showing the best cross-validation performance is shown. Genes (in the cluster) that by cross-validation where ranked within the top 10 % of the genome where considered target genes of the TRGM.

Table 4 - Transcriptional regulatory global models constructed for the 10 microarray clusters

Cluster	Key transcription factors and binding sites in TRGM (weight)
Protein synthesis	NF-Y (1.59), DEC (1.13), HIC1 (1.09), general initiator sequence (0.47), CCAAT box (0.44), TCF-4 (0.32)
Oocyte / fertilized egg	T3R (1.00), NF-Y (1.00), ETS/PEA3 (0.99), MAZ (0.92), AP2α (0.78), SP1 (0.30)
Neural tissues	UF1-H3β (1.13), CRE-BP/CJUN/ATF-1 (1.00), AP-2 (0.87), ETF (0.55), AP-1/NF-E2 (0.33)
Lymphocytes	STAT6 (1.00), PU.1 (0.99), ETS (0.96), STAT5/STAT (0.95), SP1 (0.89)
Testis / sper-matogenesis	-
Liver	TCF1/HNF-1 (1.00), NF-1 (1.00), C/EBP (0.99), HNF-4/COUP (0.99), PPAR/HNF-4/COUP/RAR (0.66), MYC-MAX (0.58), PPAR (0.33)
Mitochondrion	c-ETS (1.35), VDR (1.00), GATA-1/GATA-2 (1.00), ZID (0.82), AR (0.43), ROAZ (0.34)
Extracellular matrix	AP-1/NF-E2/BACH1 (2.00), FOXD1 (1.00), BLIMP1 (1.00), SRF (0.70), MEF-2/RSRFC4 (0.51), STAT5/STAT6 (0.35)
Cardiac muscle	SP-3 (1.00), Myogenin (1.00), MEF2A (1.00), SRF (1.00), Tyroid hormone receptor/RAR/RXR (0.91), Muscle TATA box (0.48)
Energy metabolism	CREB/ATF/HLF (1.01), WHN (1.00), SPIB (0.71), PPARγ/RXRα (0.65), general initiator sequence (0.51), RFX (0.31)

Table 5 - Transcriptional regulatory global models constructed for the 5 embryonic development sets

Development process	Key transcription factors and binding sites in TRGM (weight)
Primary heart field	D type LTRs (1.12), HAND1/TCF3 (1.01), STAT3 (0.92), STAT5A (0.89), GATA1/GATA2 (0.63), ELK1 (0.32)
Secondary heart field	HNF3α (1.56), STAT5A/STAT5B (1.00), GATA2 (0.56), NFAT (0.56), GATA/GATA3 (0.48), WHN (0.35)
Neural crest cells	FREAC-7 (1.00), Poly A (1.00), TBX5 (1.00), HSF (0.89), FREAC-2 (0.30)
Eye development	RREB1 (1.00), IRF (0.96), POU3F2 (0.92), ZF5 (0.80), GATA/GATA1 (0.46), LMO2 (0.39), NKX6-1 (0.32)
Limb development	TEF (1.00), PLZF (1.00), PAX4 (0.96), EGR (0.87), AP-2 (0.65), PBX (0.63), Ikaros 1 (0.37)


A
sensitivity
1.0
0.8
0.6
0.4
0.2
0
0.2
0.4
0.6
0.8
1.0
1-specificity
TFBS set 1
TFBS set 2
TFBS set 3

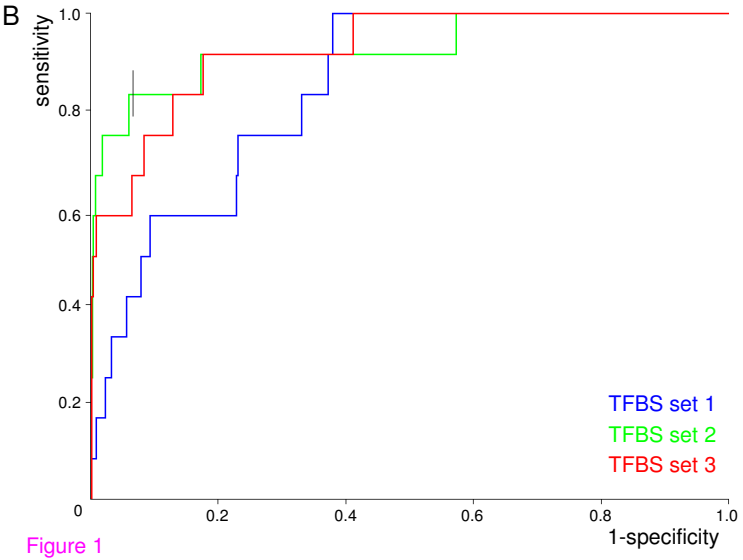
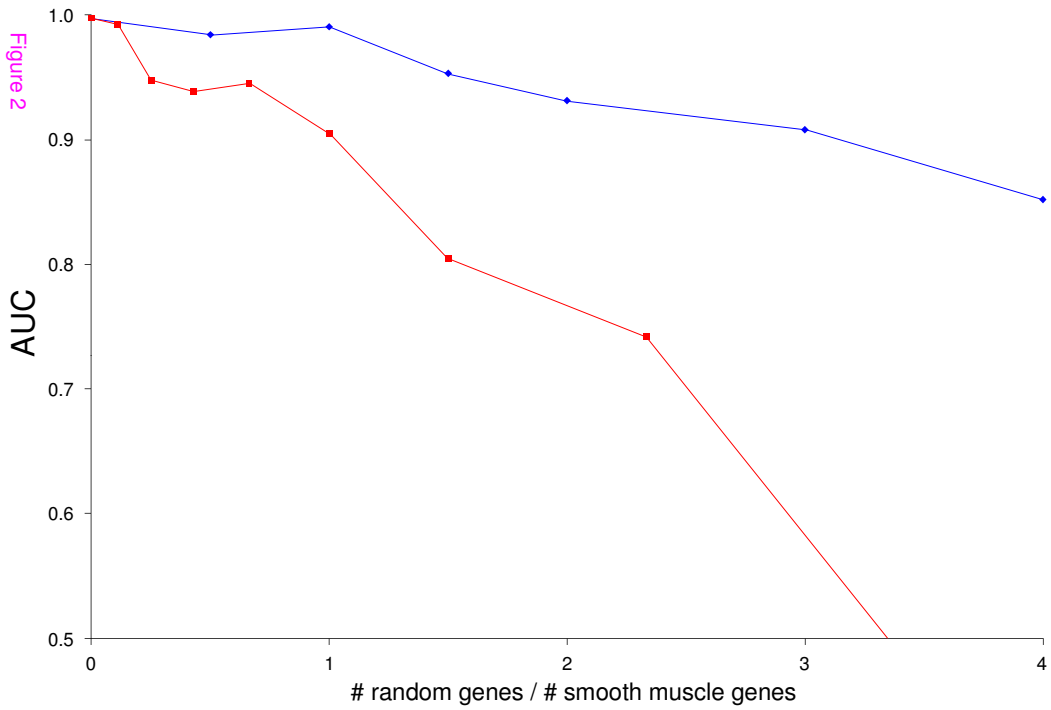
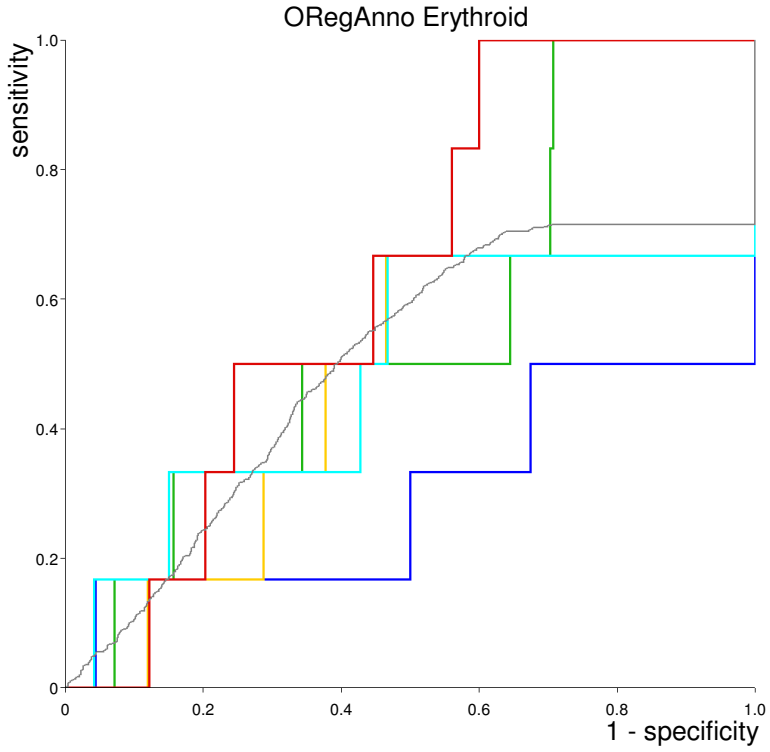

B
sensitivity
1.0
0.8
0.6
0.4
0.2
0
0.2
0.4
0.6
0.8
1.0
1-specificity
TFBS set 1
TFBS set 2
TFBS set 3

Figure 1

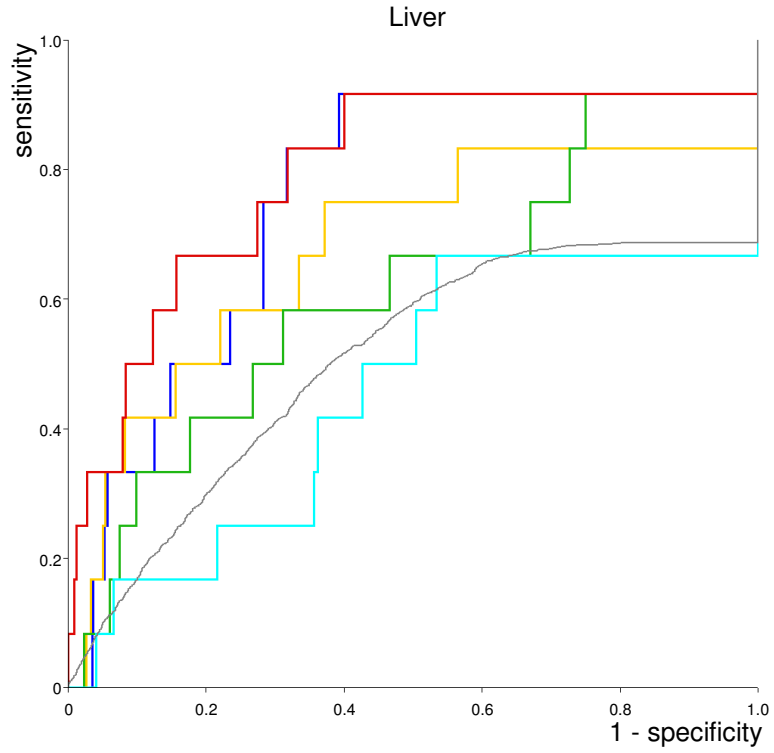
Figure 2



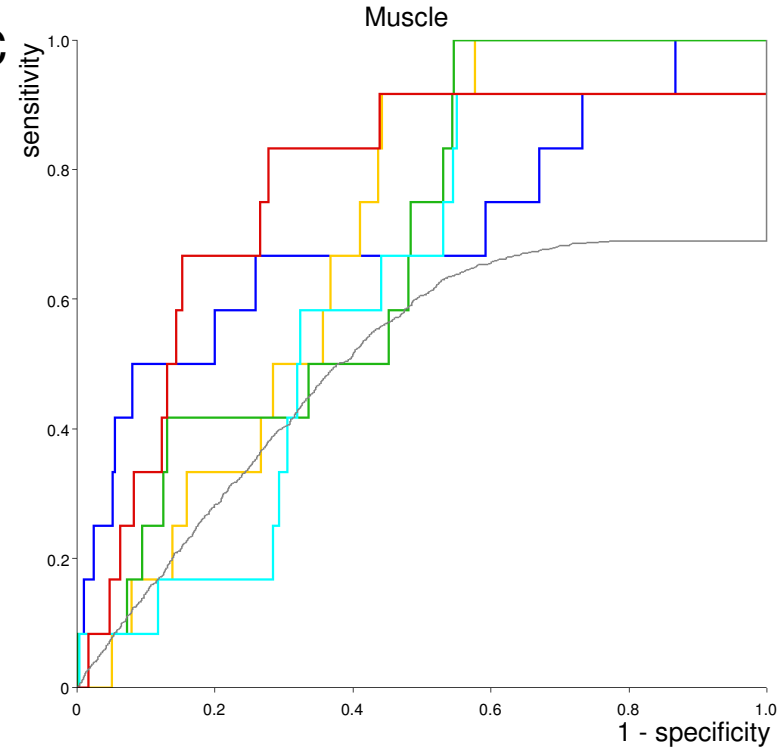
A



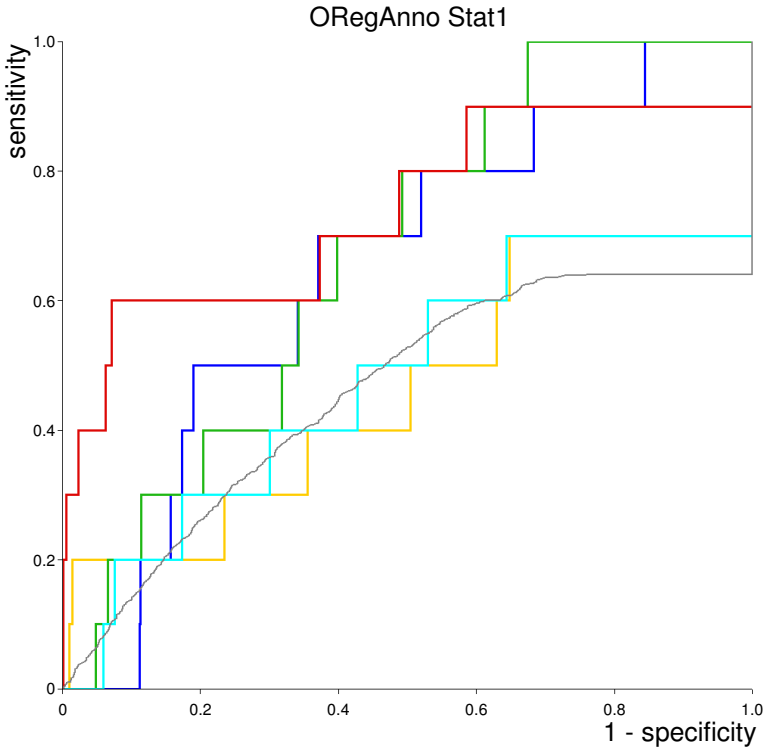
B



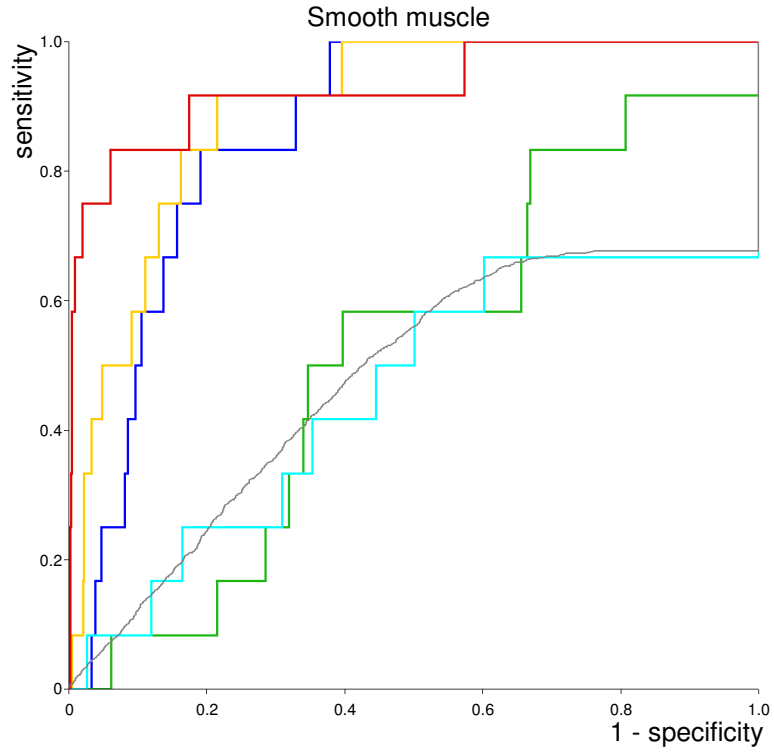
C



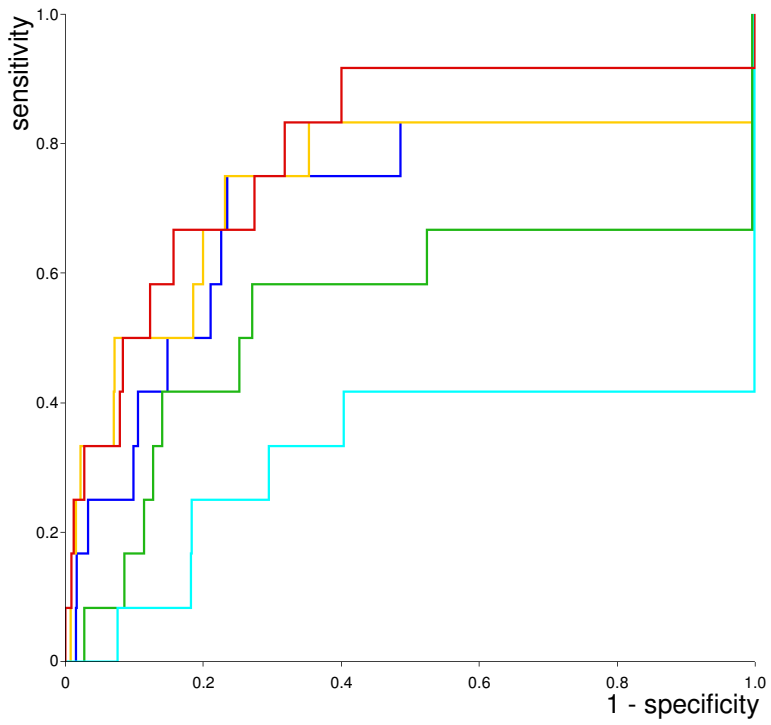
D



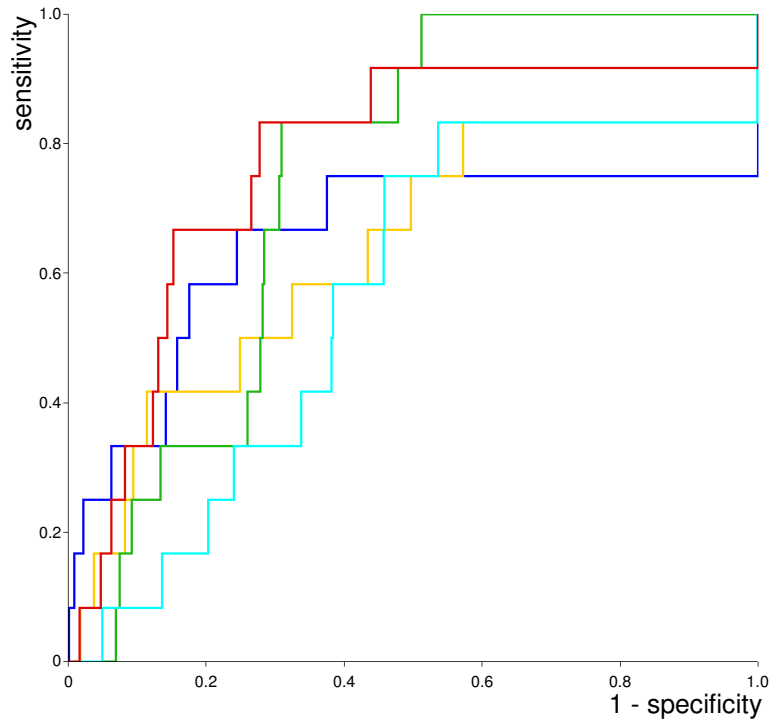
E



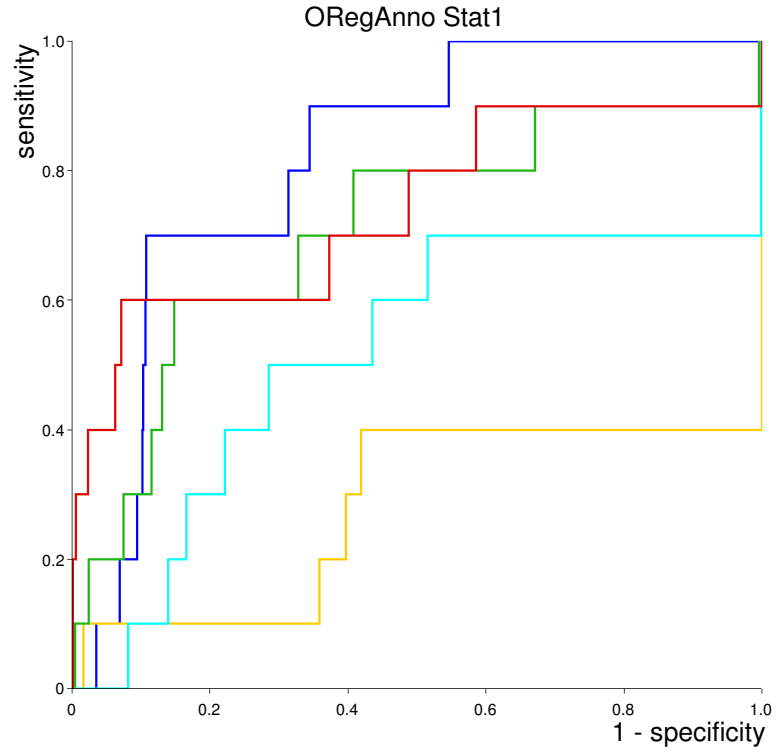
F



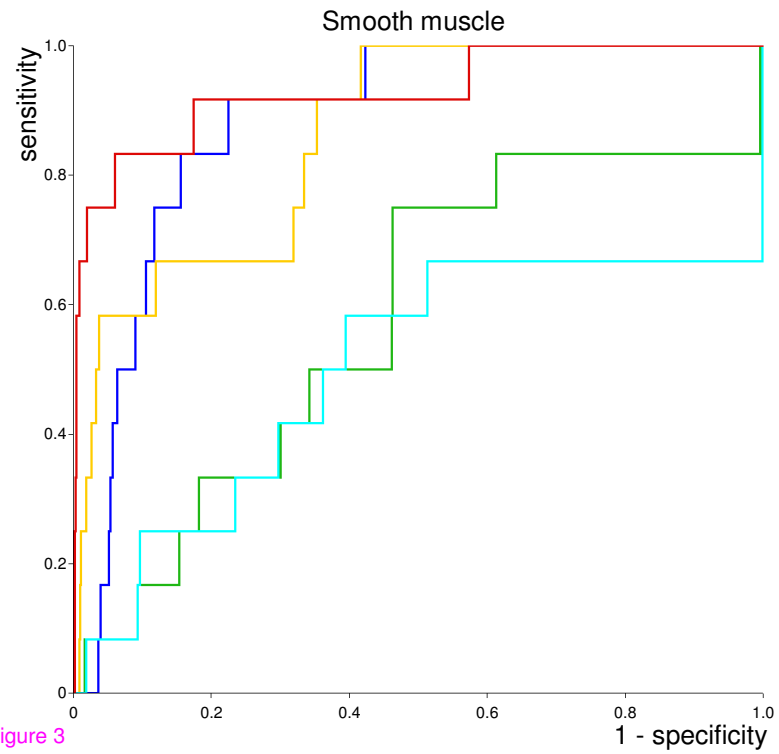
G



H



I



Legend A-I
ModuleMiner
ModuleSearcher
CisModule
EMCMODULE
Clover
Random TRMs

J

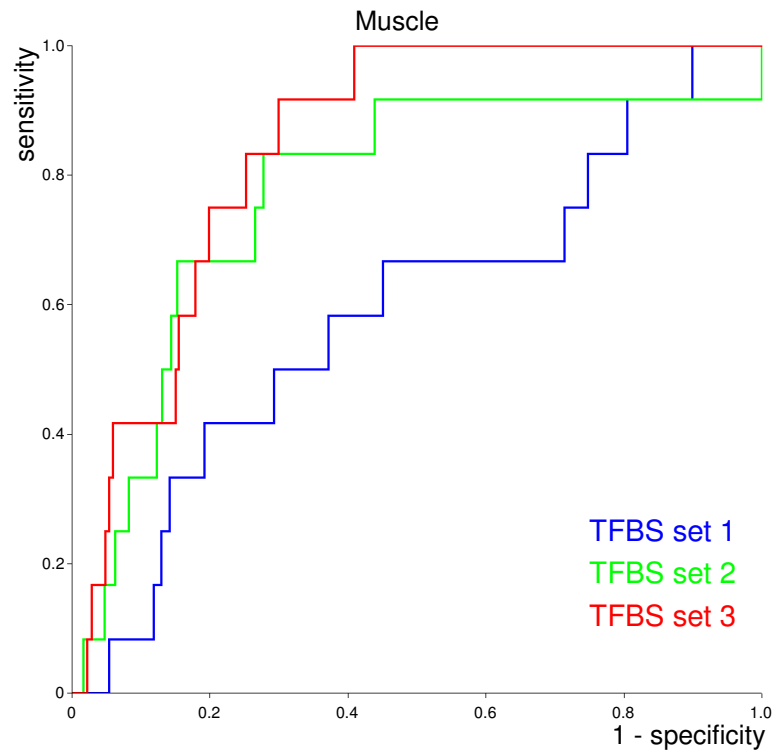


Figure 3

A

Selection of genes containing similar CRMs

Set of co-regulated genes

Leave out 1/5 of genes

ModuleMiner: Train a CRM model on the remaining 4/5 of genes

p,v
191 bp

p
211 bp

Score the full genome and consider positions of left-out genes

Are co-regulated genes overrepresented in the top 1000?

$p < 0.05$?

NO

No similar CRMs found

YES

Top-scoring genes

CRM detection

ModuleMiner: train CRM model

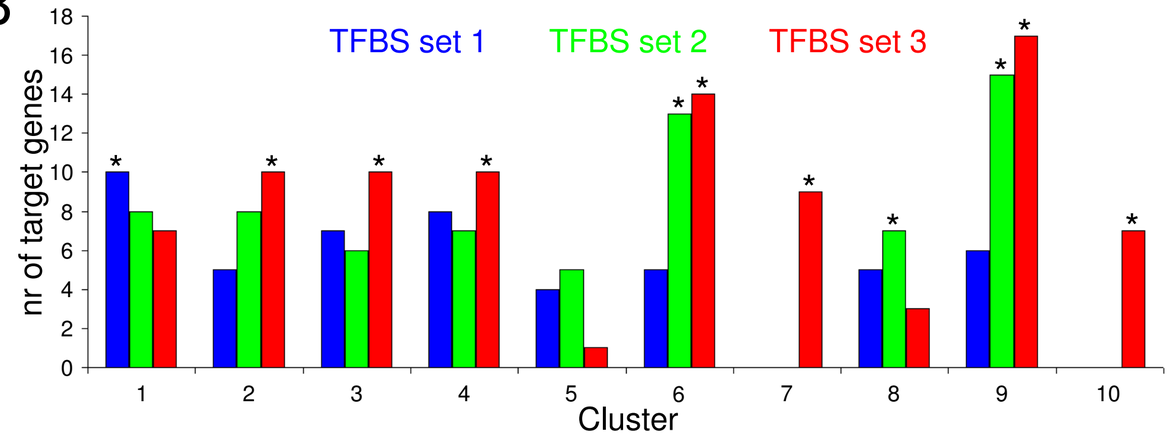
Validation: leave-one-out cross-validation

p,v
187 bp

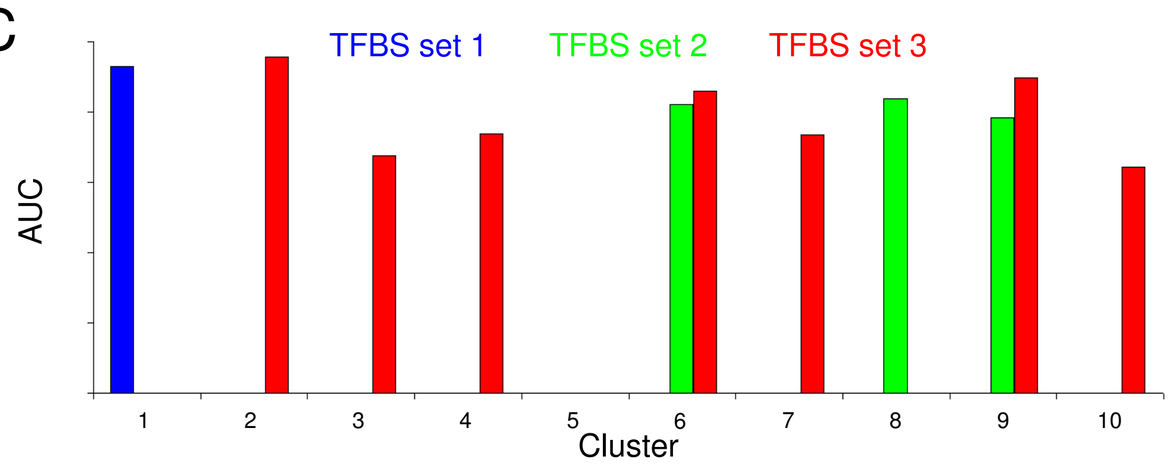
Target genes

Search for other similar CRMs in non-target genes

B



C



D

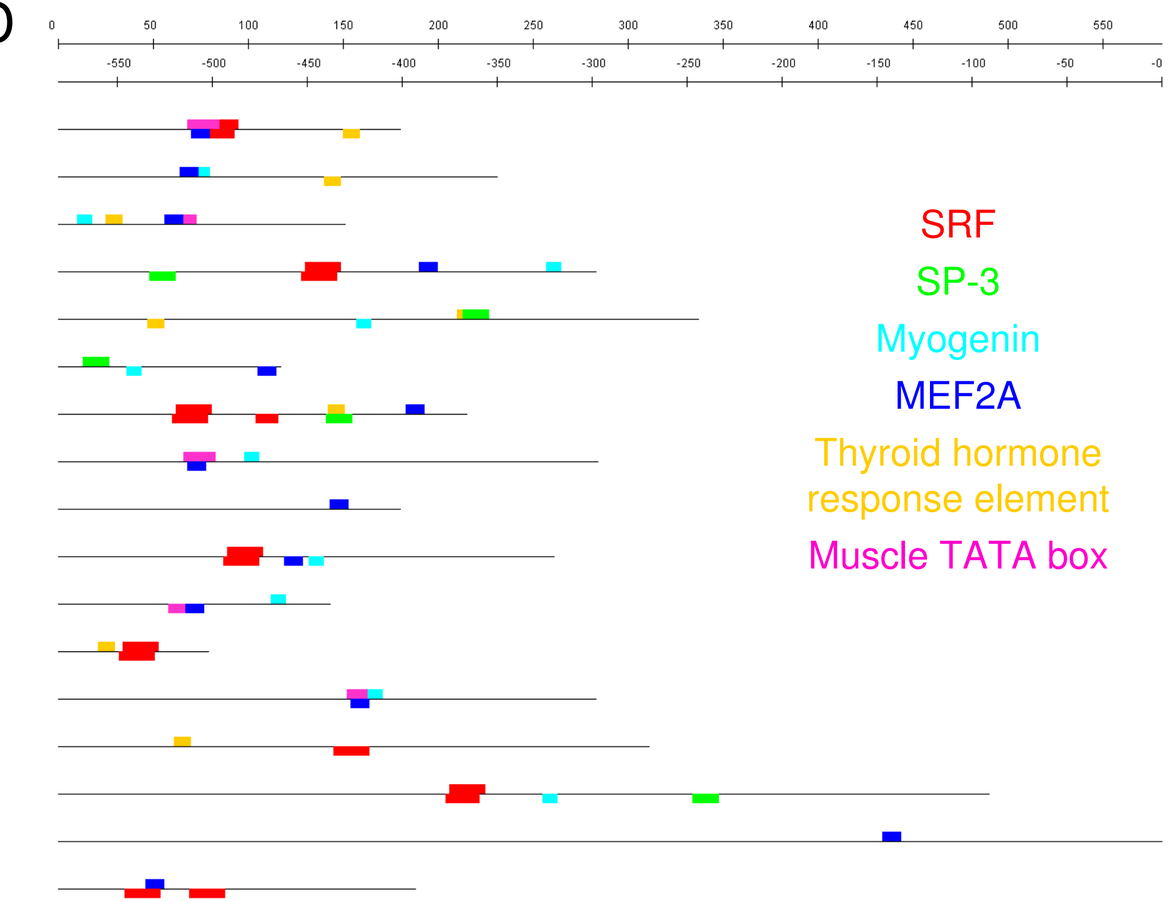
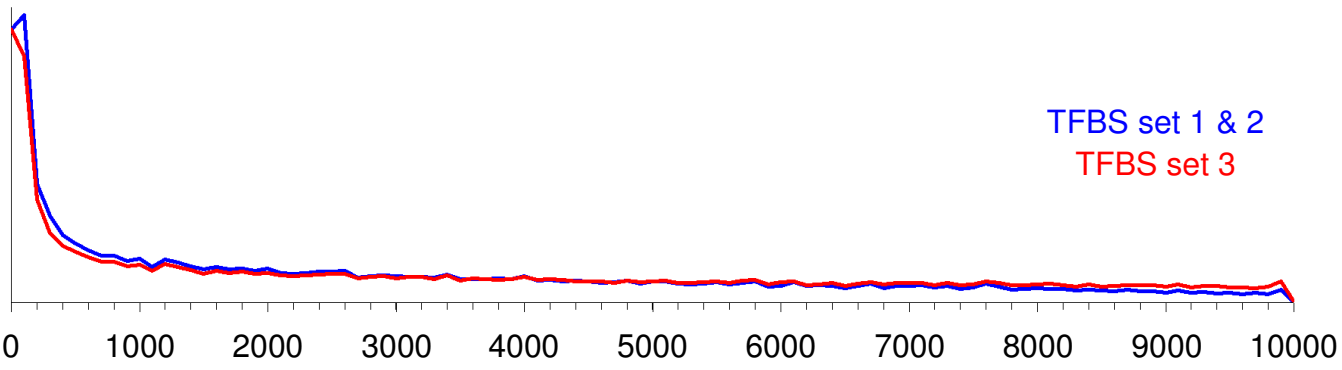
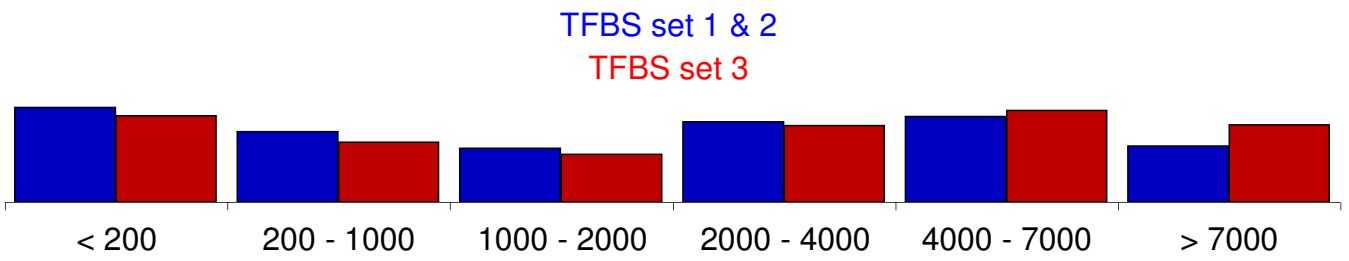


Figure 4

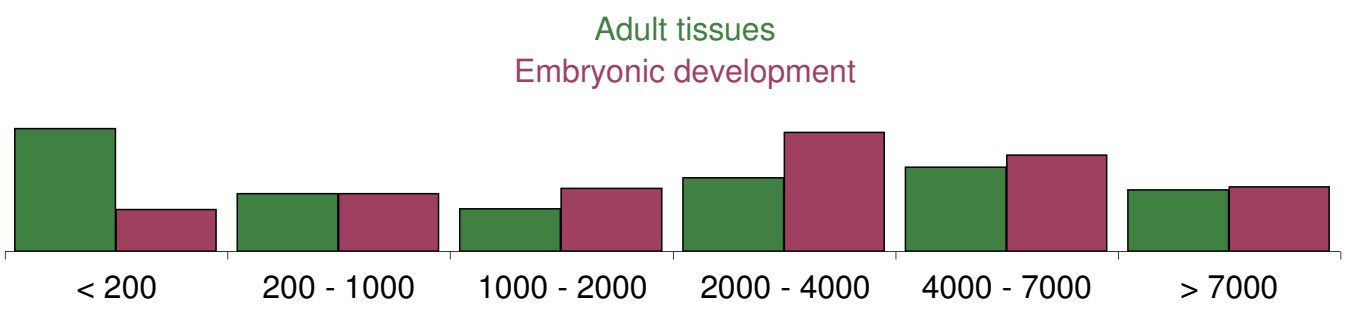
A



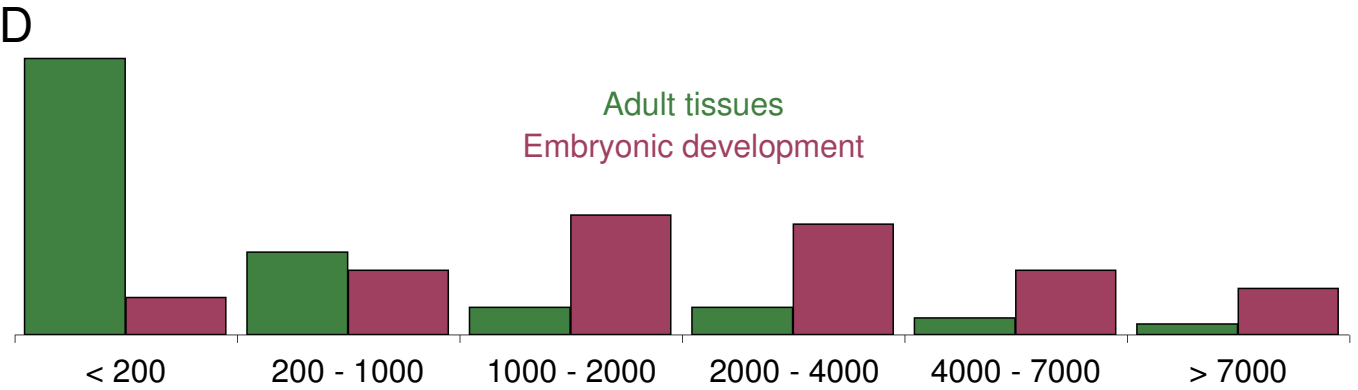
B



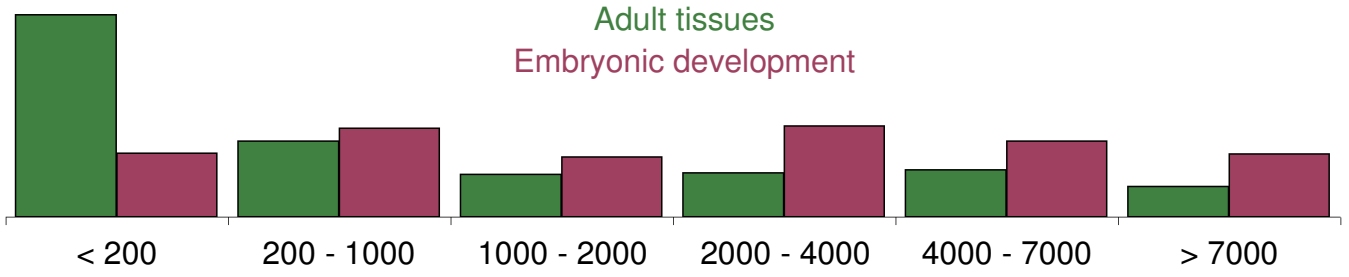
C



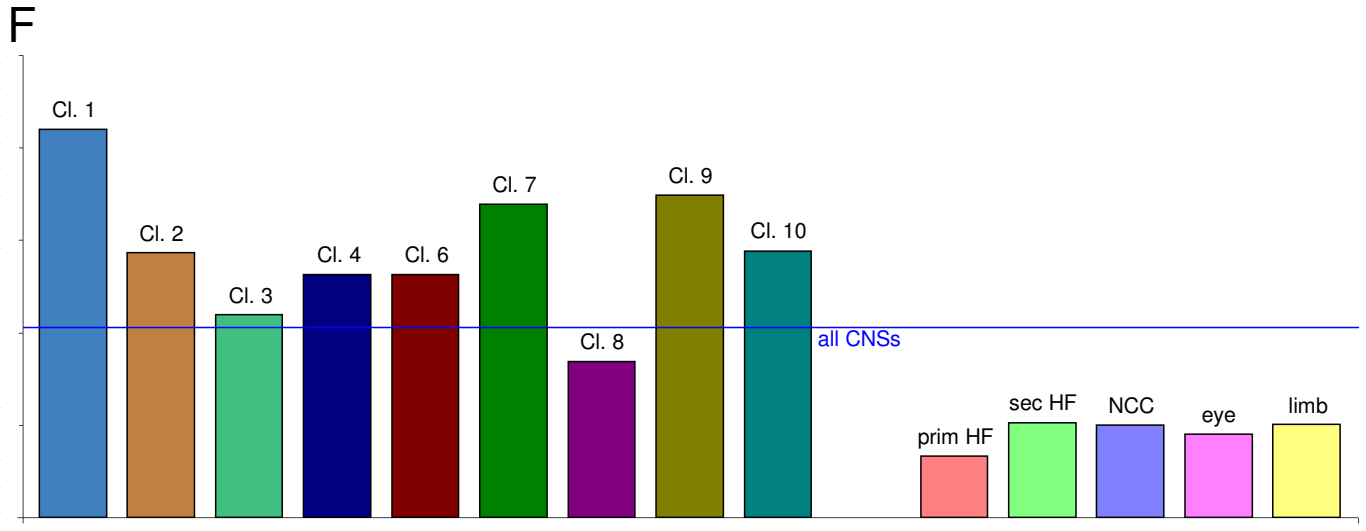
D



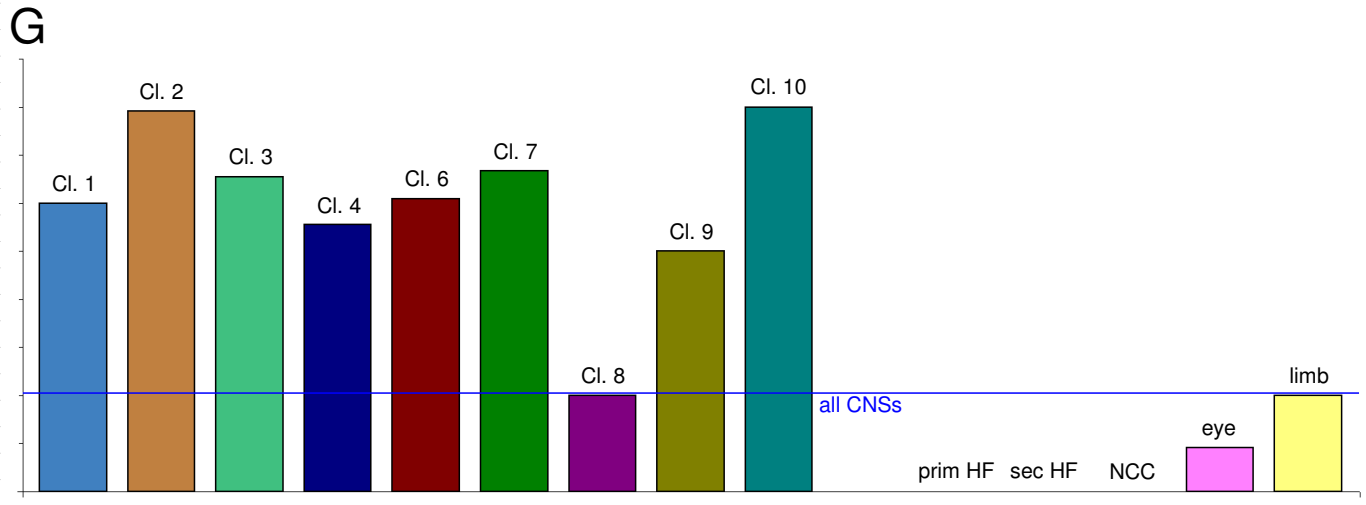
E



F



G



H

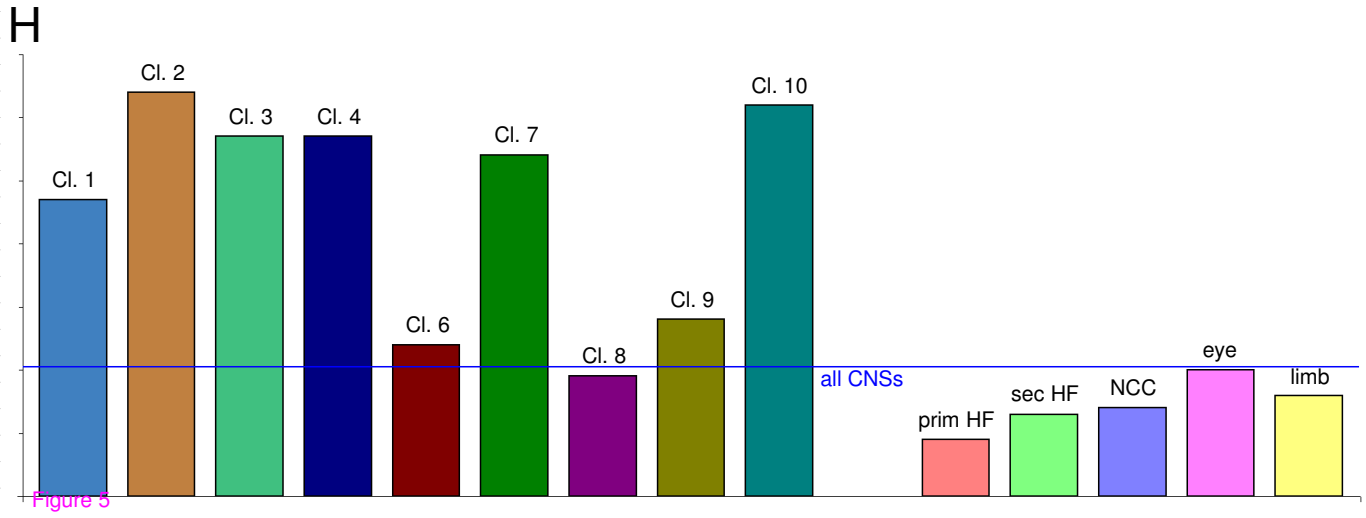


Figure 5