

# Imogene: identification of motifs and cis regulatory modules underlying gene co-regulation

Hervé Rouault<sup>1,2,+</sup>, Marc Santolini<sup>3,+</sup>, François Schweisguth<sup>1,2</sup> and Vincent Hakim<sup>3</sup>

<sup>1</sup> Institut Pasteur, Developmental Biology Department, 75015 Paris, France

<sup>2</sup> CNRS, URA2578, F-75015 Paris, France

<sup>3</sup> Laboratoire de Physique Statistique, CNRS, Université P. et M. Curie, Université Paris-Diderot, École Normale Supérieure, Paris, France.

<sup>+</sup> Have contributed equally

Email: Hervé Rouault - herve.rouault@pasteur.fr; Marc Santolini - santolin@lps.ens.fr; François Schweisguth\* - francois.schweisguth@pasteur.fr; Vincent Hakim\* - hakim@lps.ens.fr;

\*Corresponding author

April 17, 2013

## Abstract

Cis-regulatory modules (CRMs) and motifs play a central role in tissue and condition-specific gene expression. Their identification could be facilitated by the development of suitable bio-informatic tools. Here we present and test *Imogene* an algorithm that we have implemented in a publicly available software (<http://mobylye.pasteur.fr/cgi-bin/portal.py#forms::imogene>). Starting from a small training set of mammalian or fly CRMs that drive similar gene expression profiles, *Imogene* determines *de novo* cis-regulatory motifs that underlie this co-expression as well as to predict on a genome wide scale other CRMs with a regulatory potential similar to the training set. The algorithm makes central use of information provided by other sequenced genomes, based on previously developed statistical tools and explicit evolutionary models. We further show, using two mouse neural and limb specific CRM sets as well as CRMs active during fly mesoderm development, that CRMs related to different developmental programs can be distinguished on the basis of *Imogene de novo* generated motifs. We thus expect *Imogene* to be a useful tool to decipher transcriptional gene regulation in higher eukaryotes.

## Background

The identification and functional characterization of the non-coding sequences that direct the spatio-temporal specificity of gene expression in eukaryotes is of fundamental importance in developmental biology [1] and can find crucial applications in medicine [2]. These regulatory sequences are generally located distally from gene promoters and termed

enhancers or more generically cis-regulatory modules (CRMs) since they can either enhance or repress gene expression [3]. They usually are of the order of 500 nucleotides (nts) long and can be located as far as several mega base-pairs away from the transcription start sites (TSSs) of the genes that they regulate. CRMs are composed of transcription factor binding sites (TFBSs) which bring spatio-temporal

specificity to the expression of their target promoters [4]. Detailed studies in both flies and vertebrates [5] have shown that CRMs contains multiple binding sites for transcription factors that can be either identical (homotypic clustering) or different (heterotypic clustering). Homotypic clustering can provide cooperatively and sharp on-off gene expression whereas heterotypic clustering allows for combinatorial gene regulation. The extent to which the order and relative positioning of the different TFBSs in CRMs matter, remains however debated [6, 7].

With the advent of ChIP-seq techniques, genome-wide studies are providing large amount of data on the binding loci of tissue-specific transcription factors [8], as well as on other factors that regulate transcription e. g. by modifying chromatin structure (p300, CTCF, histone marks, etc) [9, 10]. This protein binding data has helped the identification of numerous CRMs specific of well-defined developmental processes and it has brought important information on CRM structure. However, genome wide studies suffer from limitations. A full characterization of regulatory mechanisms would require ChIP-seq analysis to be performed for every potential regulatory factor, on every tissue, at multiple developmental stages. The results would also have to be obtained for the often heterogeneous cells that constitute the tissue of interest instead of being averaged over them as it usually needs to be the case. Finally, and very importantly, binding cannot be equated to functional regulation.

Therefore, *in silico* identification of CRMs form a useful complement to genome-wide binding studies. By classic case-by-case studies or through larger scale analyses [11], as previously described, several CRMs have been identified as active players in the co-regulation of a subset of genes, in specific biological systems or in the formation of different organs at various stages of developement. Identifying the important binding sites on these known sequences would help to bypass some of the limitations of large scale studies by providing information on the factor involved, both known and new, as well as on the existence of a regulatory grammar. It should also help one to determine other CRMs providing specific expression patterns, a difficult task at present given the absence of close association [12] between CRMs and their target genes. These labor-intensive experimental tasks could be eased by bioinformatics. To this end, we have previously developed [13] statistical tools to determine cis-regulatory elements

*de novo*, in a set of input DNA sequences encoding a common transcriptional regulation. They allow the determination of regulatory elements from input DNA sequences without any prior information on the transcription factors acting in cis or on their binding sites. They make by central use of the phylogenetic information contained in the aligned DNA sequences of related species. The method was applied to the *D. melanogaster* gene expression program in sensory organ precursor cell (SOPs), a specific type of neural progenitor cells [13]. Predicted motifs included already characterized TFBS as well as new motifs and were successfully tested by mutational analysis. These motifs were used to rank intergenic DNA fragments genome wide for their regulatory potential in SOPs. Of the top 29 predicted CRMs, 38% were found by transgenic assays to direct transcription in SOP. A larger fraction (65%) drove more generally transcription in neural precursors.

This successful application to a *Drosophila* transcriptional program led us to wonder how the tools developed in ref. [13] would perform in the case of mammalian CRMs. The tool performance certainly needed to be assessed since the task of determining cis-regulatory elements is considerably more challenging for mammalian genomes, which are an order of magnitude richer in intergenic sequences than *Drosophila* ones. To tackle this challenge, we have developed *Imogene*, a computer algorithm and software that we present here and characterize. *Imogene* aims at:

1. predicting cis-regulatory sequences (of about 10 nt long) responsible for specific gene co-regulation within these CRMs, as well as to build a set of Probability Weight Matrices (PWM) or motifs [14, 15] characterizing the DNA-binding specificity of the associated putative factors.
2. predicting novel CRMs at the genomic scale with the same expression pattern as the starting set of CRMs, based on the set of build PWMs.

*Imogene* is based on the statistical tools introduced in ref. [13] and described in detail there. It extends this previous version by:

- i) allowing the use of a refined evolutionary model,
- ii) including the eutherian genomes and correspond-

ing alignment data, as described below,  
 iii) being accessible through the publicly available interface that we have developed and that we here describe.

In the following, the general methodology of *Imogene* is first presented. Then *Imogene* performance is tested on two sets of mammalian CRMs pertaining to neural tube and limb developmental programs during embryogenesis. We then consider the distinct but related task of discriminating CRMs with different specificities, rather than discriminating a set of specific CRM from background intergenic sequences. The discrimination of the two sets of mammalian neural tube and limb CRMs is first addressed. To further assess the performance of *Imogene*, it is applied to the discrimination of five sets of mesodermal fly CRMs, a task previously considered in ref. [16]. Finally, the developed publicly available *Imogene* interface is presented.

## Results and Discussion

### Description of Imogene

*Imogene* has two modes that can be used in succession, as sketched on Figure 1 and summarized here (see *Methods* for details of their implementation).

The first mode, *genmot*, aims at extracting statistically meaningful PWMs from a “training set” of functionally related CRMs on a reference genome (the mouse *M. musculus* genome for mammals; the *D. melanogaster* genome for flies). The size of the training set could in principle be unlimited, but in practice computer execution time requires it to stay below 100 Kbp. It should also be above a few Kbp to provide a sufficient amount of information (a training set of about 20 Kbp appears as a good compromise). Starting from a chosen training set, *Genmot* performs its task in two steps (I and II in Figure 1): I. *Genmot* first enlarges the training set with orthologous sequences in other related sequenced genomes, shown in Figure 2 (for the mouse, the 11 other aligned eutherian sequenced genomes with high coverage presently available on the Ensembl project [43]) the 11 other *Drosophilae* sequenced genomes [44] for the fly). This comparative genomics step results in the creation of the “enlarged training set” (step I in Figure 1).

II. In this second central step, *Genmot* build PWMs of given length  $\ell$  (10 nt is the default value) by scanning the training set, in an iterative manner

(step II in Figure 1). Each sequence of  $\ell$  nucleotides in the training set is used in turn to create an initial PWM using a Bayesian prior. This PWM is then refined by scanning the training set to find all the PWM binding sites in the training set, i.e. all  $\ell$  nucleotide long sequences in the training set that have a binding score above a generation threshold score  $S_g$ , chosen at the procedure onset ( $S_g = 13$  bits is the default value). These binding sites are filtered using conservation, that is only sites that have orthologues in distant species are kept. A shift in alignment between a binding site on the reference species and its orthologues in other species to correct for eventual alignment errors (20nt is the shift default value). The ensemble of conserved binding sites and their orthologues serve, using an evolutionary model, to build a refined PWM. The procedure is then iterated by finding the binding sites of the refined PWM and using them to build a further refined PWM until convergence to a stable set of binding sites.

The need of an evolutionary model to properly assemble binding sites [25,26,45] is simply explained. A binding site in the reference genome and its orthologues are all related through descent from their last common ancestor, and cannot therefore be considered as independent observations. In order to correctly quantify the amount of information provided by the observation of orthologous sites, one has to estimate their potential of change through mutation since their last common ancestor. To account for this, *Imogene* can, in its present implementation, make use of either one of two evolutionary models of TFBS evolution at the user choice. The first option, “*Felsenstein*”, is a simple and computationally fast model proposed in [45]. Mutations are generated at the same rate in a PWM binding site than in the background intergenic sequences. However, the mutated nucleotide in a binding site is drawn according to its frequency in the PWM at the mutated position. This is analogous to the simplest model of DNA evolution [46] but with nucleotides neutral relative abundances replaced by PWM nucleotide frequencies. This *Felsenstein* model is the simplest model that provides at evolutionary equilibrium, nucleotide frequencies that agree with those prescribed by the PWM at the different positions in the binding site. The second option, “*Halpern-Bruno*” [47] uses an evolutionary model that is more complex than the *Felsenstein* model and that has previously been used for TFBS evolution in [25]. It allows for the inclusion of different mutational prob-

abilities between different bases in the neutral background intergenic mutation model. Additionally, it includes a fitness-dependent fixation probability for a mutation in a TFBS, based on classical population genetics estimates for the fixation of a mutant allele appearing in an homogeneous population [48]. The relative fitnesses of different nucleotides are determined by the requirement that binding site convergence to evolutionary equilibrium leads to the PWM nucleotide frequencies (see Methods for details).

The described procedure produces a PWM for every  $\ell$  nucleotide long sequence in the training set. In a series of final steps, this long list is pruned and ranked based on comparing the PWM bindings sites on the training set to a “background” set of intergenic sequences in the reference genome (20 Mb of *M. Musculus* or *D. melanogaster* genomic DNA). The PWM corresponding to repeated sequences are first eliminated on the basis of the non-poissonian distribution on their binding sites in the background set. Then for each remaining PWM, the distribution of its conserved binding sequences on the training set is compared to the distribution of the PWM conserved binding sequences on a set of background intergenic sequences. The larger the statistical deviation between the two distributions, the larger its score and the more meaningful the PWM is deemed. In a final step, PWMs in the ranked list are compared and, among similar ones, only the highest scoring one is kept. Although the identity of the transcription factors corresponding to the different PWMs of interest is not directly assessable by the algorithm, the comparison between the produced PWMs and existing databases can provide relevant information on their identity, as will be shown in the following sections.

In its second mode, *scangen*, *Imogene* determines intergenic sequences in the reference genome that are considered as putative CRMs with the same functional specificity as the training set. This second mode (step III in Figure 1) is based on the inferred PWMs in the *genmot* mode. The algorithm scan the entire non-coding repeat-masked reference genome and find all the conserved binding sites above the scanning binding score  $S_s$  for the  $N$  first PWMs in the ranked list. The intergenic sequences of a given length (the default value is 1000 nt) are then scored according to their similarity to the training set in their content of PWM binding sites. The closest the similarity in its motif content with the training set, the most likely an intergenic sequences is deemed to

be functionally related to the training set.

### Application to eutherian developmental programs

In order to assess *Imogene* performance on mammalian transcriptional regulation, we applied it to two sets of mammalian specific CRMs, that have previously been identified starting from p300 ChIP-seq data and functionally tested in a transient transgenic assay for activity in stage 10 mouse embryo [11, 49]. We chose CRMs active in neural tube and limb, as characterized in the VISTA website (<http://enhancer.lbl.gov>). For each developmental program, a subset of CRMs was visually selected for specificity and strength of expression in the tissue of interest, from the provided expression pattern. Among these selected sets, 2 limb CRMs and 4 neural tube CRMs contained no sequence that could possibly be used to learn motifs by *Imogene*, due to its conservation requirements, either because of repeat masking or because of low conservation (see *Methods*). Elimination of these uninformative sequences produced curated training sets of 29 neural and 39 limb CRMs.

A cross-validation scheme was then used to measure *Imogene* predictability power (see *Methods* for details). In brief, for each developmental program, the CRMs were divided into a training set composed of 15 CRMs chosen at random, and a test set composed of the other CRMs used as True Positives. These test CRMs were ranked against a ‘background test set’, a set of  $\sim 60$  regions of 1Kb taken from the flanking sequences of the initial set of CRMs (see *Methods*).

The training set was used for motifs generation using *Imogene genmot* mode. This procedure was conducted for both evolutionary models using different values of the generation parameter  $S_g$  and scanning threshold  $S_s$  to obtain the optimal values of these parameters for each model and each training set (see Figure 3 and Figure S1). For different parameter sets, the test CRMs as well as the intergenic sequences of the background set were scored. The proportion of retrieved test set CRMs above a given score (True Positive Rate or TPR) was plotted against the proportion of appearing test background regions above the same score (False Positive Rate or FPR) as this score decreased, to produce a so-called ROC curve [?]. The ROC curves corresponding to different parameters values were then compared using the Area Under ROC Curve (AUC), a quantity

that is maximal at best prediction.

Figure S1 shows the AUC as a function of the number of motifs  $N$  for different values of the scanning threshold  $S_s$ . One can see that the AUC increases quickly with the 5 first motifs generated, and has nearly converged to its maximum value when 10 motifs are kept. Therefore we restricted ourselves to  $N = 10$  motifs, and constrained the other parameters using AUC maximization. Figure 3 shows the ROC curves obtained for the optimal parameters which are seen to be similar for both models and both training sets. Figure S1 shows how changing  $N$  impacts these ROC curves, making it clear that  $N = 10$  is already nearly optimal.

For the limb CRMs, 60% of the test set CRMs are retrieved at 5% FPR whereas an even larger proportion of 70% is obtained for the neural tube CRMs. The *Halpern-Bruno* and the *Felsenstein* models are seen in Figure 3 to yield very similar results in both cases, with a slight superiority for the *Halpern-Bruno* one. It should be noted that the chosen measure really provides a lower estimate of *Imogene* success rate since we considered as ‘False Positives’ all flanking CRMs sequences, whereas, in reality, some could be *bona fide* positive CRMs.

In the cross-validation procedure, different ranked lists of motifs were created for each randomly drawn test set. In order to provide a list of motifs generated by the algorithm, we ran *Imogene* on the full set of CRMs for each class. The corresponding 10 best motifs are shown in Figure S2. The closest TRANSFAC PWM assigned to each motif by *Imogene* PWM distance is also shown in Figure S2. Previously characterized motifs belonging to the considered developmental programs appear in each class (e.g. Oct and NeuroD motif in the neural CRMs). The motif content of each CRM is also provided in Figures S3, S4. It is seen that the 10 best motifs appear on most CRMs of the training set.

### Discrimination of tissue-specific CRMs in the mouse

Given *Imogene* ability to distinguish specific CRMs from background sequences, we found it interesting to apply it to the related but distinct task of distinguishing different classes of CRMs. The question was previously considered for *D. melanogaster* CRMs based on ChIP-seq data at different developmental time points [16], as detailed in the next section. It consists in learning features that distin-

guishes the CRMs of a given class from the CRMs of other classes, in order to be able to predict the class of a newly observed CRM. The task differs from distinguishing CRMs from background intergenic sequences since learning motifs shared among different classes, for instance characterizing the binding of generic CRM factors, is of no use for discrimination purposes. As a test case, we considered the neural tube and limb sets of mammalian CRMs used in the previous section. Given the nature of the task, we selected in each set the CRMs with an expression that appeared mostly restricted to neural tube and limb. This yielded 12 neural and 15 limb CRMs.

As in ref. [16], we used a leave-one-out cross validation (LOOCV) scheme in which the learning set constituted all but one of the elements of a class, the remaining one being used as a test sequence. The process can be summarized as follow. We call the class of interest the positive class and the classes against which we wish to learn the negative classes. The LOOCV process begins with the exclusion of a (positive or negative) CRM which serves as an unobserved test CRM. Then, a set of  $N$  motifs is learnt on the remaining CRMs of each class, yielding positive and negative motifs. These motifs are used to build a weighted score giving positive (resp. negative) contributions to positive (resp. negative) motifs (see *Methods*). Finally, the rank of the test CRM among all CRMs is kept as an indicator of the classification. We expect positive CRMs to be on top of the list and have low ranks while negative CRMs should be attributed high ranks. After processing for all CRMs, we have a list of ranks for the positive and negative CRMs that can be represented by a ROC curve indicating the True Positives Rate and False Positive Rate for increasing rank. We optimized parameters (the threshold for motifs generation  $S_g$ , the threshold for sequences scanning  $S_s$ , and the number of motifs  $N$  used to score sequences) by maximizing the Area Under the ROC Curve for a  $FPR \leq 0.2$ .

The results are shown in Figure 4. We focus on the results obtained with the *Halpern-Bruno* evolutionary model, which does slightly better than the *Felsenstein* model, as for the previous CRMs ranking task. Results (motifs, thresholds) are nonetheless very comparable in the two cases. Motifs are shown on the right of the ROC plots and were generated on the positive classes with optimal parameters. The two classes were optimally discriminated using only 2 motifs in each class, with specificities  $S_g = 11$ ,  $S_s = 8$ , comparable to that found in the learning

task of the previous section. The best ranking motif of the neural CRMs was found to be unequivocally associated to the Transfac Oct1/Pou2f3 Transcription Factor, known to be involved in the neural tube formation [51].

### Discrimination of *Drosophila* tissue-specific CRMs

In order to further test the discriminating power of *Imogene de novo* generated motifs, we applied it to the CRM classification task reported in ref. [16]. In this work, previously characterized *D. melanogaster* CRM were divided in 5 classes corresponding to the different tissue types in which they were active: mesoderm (Meso), somatic muscle (SM), visceral muscle (VM), mesoderm and somatic muscle (Meso & SM) and visceral and somatic muscle (VM & SM). Ref. [16] made use of a collection of Chip-seq binding data for different factors and at different developmental time points to attribute to each CRM a total of 15 peak height values. It was then tested whether classical machine learning techniques could be used to discriminate the different CRM classes, on the basis of these extensive data. This was indeed found possible with a high success rate in a standard cross-validation scheme: CRMs predicted with probability higher than 95% to belong to a given class were indeed found to belong to that class with a high success rate of 80%.

This led us to wonder whether *Imogene* would succeed in classifying these different CRMs, without using any binding data, but rather on the basis of combinations of *de novo* motifs that it would itself generate. We used the set of well-characterized CRMs belonging to 5 different classes assembled in ref. [16]. We then proceeded similarly to the previous case with eutherian CRMs.

*Imogene* results are shown together with the machine learning results of ref [16] in Figure 5. For clarity, we here show results obtained with the *Felsenstein* model. Results obtained with the *Halpern-Bruno* model are comparable. Strikingly, without any binding data *Imogene* prediction rates are comparable to the machine learning ones, in the high specificity range ( $FPR \leq 5\%$ ) used for CRM prediction in [16]. Its performance is even better for the Meso and SM classes at high score. The latter case is of particular interest. The machine learning algorithm essentially used Mef2 ChIPseq peak heights to predict SM CRMs, resulting in an incorrect classification at high scores since this TF is required for

the differentiation of all muscle types. However, the use of the specific Mef2 motif obtained *de novo* from the SM training set allows one to restore a correct classification at high score (Figure 5C).

On the side of each ROC plot, the *de novo* motifs generated on the whole training set are displayed. The number of motifs shown is the optimal number used for CRM scoring in the leave-one-out cross-validation. Among the generated motifs, one can recognize 4/5 TFs for which ChIPseq data was used in [16], namely Twist (motif 2, Meso & SM), Mef2 (motif 1, SM), Bin and Tin (motifs 1 and 2, VM). The Bap motif was not found by the algorithm, and correspondingly it was not shown to be of importance in ref. [16].

In summary, our analysis indicates that *Imogen* can not only identify *de novo* functionally relevant binding sites within a set of CRMs but can also be used to identify the more subtle differences in binding sites that underlie functional differences between related sets of CRMs.

### Web interface

The described algorithm is available through a user-friendly web interface that provides motif and CRM predictions for the community. This interface is powered by the Pasteur Institute Internet server through the mobyle framework [52]. The input web page and an example output web page are shown in Figure 6 and 7.

The input form (see Figure 6) is divided into several sections. One of the two available algorithm modes should be chosen at start:

- **genmot:** given a list of coordinates of typically 15 enhancers of 1 kb (training set), generates *de novo* motifs ranked by p-value.
- **scangen:** given the previously generated motifs, produces a list of genome-wide predicted CRMs with conserved binding sites. The rank of a CRM is based on a poissonian score that takes into account the motif content, as explained in [13].

The group of species considered should also be specified. The algorithm can be used on *Drosophilae* (with reference species *D. melanogaster*) or mammals (with reference species *Mus musculus*). The different algorithm parameters such as the sought motif width, threshold specificity for binding sites or

allowed position shifts between different species (see *Methods* for a detailed description) are set by default to values that have been found to provide reasonable results. They can be modified by the user to optimize the results for the considered training set.

In mode *genmot*, the user should enter the training set CRM coordinates. The chosen evolutionary model for the TFBS should also be specified. The *Felsenstein* mode is computationally faster than the *Halpern-Bruno* one. The results of the two modes have been found to be comparable, with a slight superiority in the the performance of the *Halpern-Bruno* model (see Figure 3 and 4).

In mode *scangen*, the algorithm scores and ranks intergenic sequences in the reference species, using a list of motifs, as described in the first *Results* section and in *Methods*. The list of *de novo genmot* motifs is used as input. The user can set the length of the ranked sequences (1 Kb is the default value) and the number of scoring motifs (5 is the default value for computational speed but this can be changed to improve results, see Figure S1)

An example of *Imogene* output is displayed in Figure 7. The *genmot* mode creates from the provided training set a list of ranked motifs together with their significance and over-representations (see *Methods*). The positions of these motifs on the CRM of the training set and on their homologous sequences in other species are also provided, as illustrated in Figure 7A for 2 motifs. Figure 7B shows the output of the *scangen* mode for these two motifs. The ordered list of best-ranking intergenic sequences is given together with information on the closest TSSs.

## Conclusions

We have presented *Imogene*, a computer algorithm able to predict *de novo* relevant motifs in functionally related sets of CRMs and able to infer novel CRMs with a low false positive rate in both drosophilae and mammalian genomes. *Imogene* mode of inference internally makes use of quantitative models for binding site evolution. This allows it to systematically exploits the information available in multiple sequenced-genomes, and to work efficiently from a CRM set of modest size.

Numerous algorithms have already been developed to try and map cis-regulatory information underlying transcriptional regulation (see e.g. [3,14,17–

19] for recent reviews). *Imogene* differs from previous methods in several respects. It has been specially designed to decode cis-regulatory regulation in a small set of CRMs while other algorithms are aimed at the analysis of large datasets such as whole ChIP-seq peak regions [20]. It works *de novo* while many algorithms make use of well-characterized binding motifs [21–29]. Several that work *de novo* try and distinguish regulatory sequences by their entire content in short nucleotide sequences [30–35]. Phylogenetic conservation between multiple sequenced genomes has been shown to provide useful information on cis-regulatory motifs [36–38], but cannot *per se* address the question of specific spatio-temporal expression. In order to analyze a small set of CRMs with well-characterized expression, *Imogene* makes full use of several sequenced genomes instead of focusing a single genome [29] or simply comparing it with another one [39–41]. Moreover, it does not simply add orthologous sequences [42] but uses an evolutionary model to properly weigh this additional information.

The algorithm which lies at *Imogene* core was previously applied to gene co-regulation in *Drosophila* [13]. Motifs predicted to be important for Sensory-Organ-Precursors development were confirmed by site-directed mutagenesis. A significant fraction of top predicted new CRMs were also shown to direct expression in SOP or more generally in the peripheral nervous system. The ability of the algorithm to provide meaningful information on cis-regulatory elements in *Drosophila* was further confirmed in a subsequent application to epidermal morphogenesis and trichome development [53]. The algorithm provided an informative PWM for the master regulator Ovo/Shavenbaby and predicted as well a functionally important novel motif.

In spite of its successful application to gene co-regulation in *Drosophila*, it was not clear that *Imogene* would be able to decipher cis-regulatory information in the notoriously more difficult case of mammalian gene expression. We have here provided bio-informatic evidence that *Imogene* provide meaningful results in this case also. *Imogene* was shown to successfully recognize CRMs belonging to neural and limb development programs solely based on motifs that it has constructed *de novo* on other CRMs. Furthermore, the created PWMs appear to comprise both known and new motifs, in strong analogy with the previous studied cases in the fly.

There is currently numerous cases for which a

small number of CRMs belonging to the same program of gene expression has been characterized. Therefore, the possibility to use *Imogene* should provide helpful service to the community. We have further shown that *Imogene* can discriminate between classes of CRMs. In this task, it should play a complementary role to ChIP-seq data that are currently obtained for many developmental programs. Whereas, ChIP-seq provides information on the binding of already known factors, *Imogene* is able to propose new PWMs and help to identify new involved DNA-binding cofactors and their binding sites. We thus believe that *Imogene* is a useful complement to existing softwares [19]. A user-friendly version has been made it publicly available on the Pasteur Institute web platform <http://mobyle.pasteur.fr/cgi-bin/portal.py#forms::imogene>. The full computer code is also freely available at <http://github.com/hrouault/Imogene>.

## Acknowledgments

We wish to thank I Leroux, S Meilhac and B Robert who helped us to characterize the patterns of expression of the mammalian CRMs used in the present work. We acknowledge the Centre d’Informatique pour la Biologie at the Pasteur institute for its help in the design of a mobyle front-end to *Imogene*. This work was supported by core funding from Centre National de la Recherche Scientifique, Ecole Normale Supérieure and Institut Pasteur and by a specific grant from the Agence Nationale pour la Recherche (ANR-08-BLAN-0235).

## Methods

### Genome alignments

The alignments were downloaded from [ftp://ftp.ensembl.org/pub/release-63/emf/ensembl-compara/epo\\_12.eutherian](ftp://ftp.ensembl.org/pub/release-63/emf/ensembl-compara/epo_12.eutherian) for eutherians and from [http://www.biostat.wisc.edu/~cdewey/fly\\_CAF1/data](http://www.biostat.wisc.edu/~cdewey/fly_CAF1/data) for *Drosophila*. For the latter case, we have used the alignments engineered by A. Caspi with the help of the Mercator and MAVID programs. In both cases, the alignments were processed through a customized script to produce alignments in fasta format, mask for coding sequences (CDS) and simple repeats (see below).

### Annotations

The CDS coordinates were downloaded from [ftp://ftp.ensembl.org/pub/release-64/gtf/mus\\_musculus](ftp://ftp.ensembl.org/pub/release-64/gtf/mus_musculus) for eutherians (mm9 coordinates) and from [ftp://ftp.flybase.net/releases/FB2011\\_06/dmel\\_r5.38/gff](ftp://ftp.flybase.net/releases/FB2011_06/dmel_r5.38/gff) for *Drosophila* (release 5 coordinates). In the case of eutherians, the TSS coordinates were obtained separately from <http://hgdownload.cse.ucsc.edu/goldenPath/mm9/database>. Eutherian alignments repeat sequences were already masked for repeat sequences, appearing as small capital nucleotides in the emf files. *Drosophila* alignments were masked using the coordinates indicated in the gff file.

### Background sequences

*Imogene* computes the statistical overrepresentation of the predicted motifs by comparing them to 20 Mb of background intergenic region. The script that generates the random coordinates is included in the distribution of *Imogene* as well as the actual coordinates of the produced intergenic regions.

### Training sets

The two mammalian training sets (limb, neural tube) were obtained from <http://enhancer.lbl.gov>, based on the work of [11, 49]. They were manually curated to produce a high-quality data set, with respectively 41 CRMs for the limb, and 33 for the neural tube. We further pruned out uninformative CRMs for which no motifs could be generated, either because of repeat masking or because of lack of conservation. More precisely, the reference species sequence was scanned using a window size corresponding the motif size. If a sequence did not contain any masked nucleotide, we looked in the other species for any unmasked sequence in the surrounding neighborhood of  $\pm 20$ nt, our flexibility criterium when defining a conserved instance. If putative orthologous sequences were found in enough species to satisfy our conservation requirements (see below), the site was declared as a putative conserved site for a regulatory motif. This filtering step resulted in final sets of 39 CRMs for the limb and 29 for the neural tube. The *Drosophila* training sets were obtained from [16]. Coordinate files are given as Supplemental Material.



## Mammalian predictions

**Training and background test sets** For each class, the CRMs were divided into a learning set composed of 15 CRMs chosen at random, the other CRMs ( $\sim 20$ ) defining the test set of 'True positives'. In addition, a set of background test regions was built using the 1Kb flanking sequences of the full list of CRMs. Such an 'adapted' background test set was used to provide a more stringent and informative test of the algorithm by preventing discrimination on the training set from the background test set, based on other features than the sought high-information-content motifs, such as a local composition bias. Furthermore, in order to avoid biasing the results towards the true positives, uninformative sequences for *Imogene* (i.e. sequences where no binding site could possibly be found given *Imogene* conservation requirements) were also removed from this background test set. These regions were also filtered for uninformative elements. This yielded background test sets of 72 CRMs for the limb and 57 for the neural tube.

**Cross-validation protocol** The learning set was used to learn the motifs content. The 10 best motifs were then used to score test set CRMs and background regions. Because the length of the training set CRMs could vary, we decided to keep for each test sequence the best scoring 1kb fragment. This process was repeated 40 times, and both generation and scanning threshold were varied. The retrieval rate of test set CRMs (True Positives) among background elements (False Positives) as a function of the score was used to build a ROC curve. The Area Under ROC Curve or AUC, a quantity that varies between 0 for absolute misclassification, 0.5 for random classification, to 1 for perfect classification, was used to evaluate the quality of prediction. The parameter set yielding the highest AUC was chosen as the best set.

## Leave-one-out cross-validation for the CRM discrimination task.

Let us note  $\mathcal{C}_i$  the tissue class of interest. There are  $M_i$  corresponding CRMs. Let  $N_c$  denote the total number of classes. Our goal is to find the particular motif signature that distinguishes these  $M_i$  CRMs from the  $N_c - 1$  other classes of CRMs. This signature corresponds in our case to a number  $N$  of top ranked motifs with generation and scanning thresholds  $S_g$  and  $S_s$ . These are the three parameters we wish to constrain with a leave-one-out cross-

validation (LOOCV) procedure.

Let us detail this procedure in the case where we distinguish class  $\mathcal{C}_i$  from the other classes  $\mathcal{C}_j$ . The  $M_i$  CRMs of  $\mathcal{C}_i$  are termed 'positive' CRMs and the  $M_j$  CRMs of each of the other classes are termed 'negative' CRMs. Let us note  $M = \sum_i M_i$  the total number of CRMs. The LOOCV consists in withdrawing one 'test' CRM from these  $M$  CRMs, learn the motifs on the  $M - 1$  resulting CRMs, and use them to score the left alone test CRM. For the learning step, motifs are generated with threshold  $S_g$  on each class (one class being deprived of one CRM), yielding  $N_c$  sets of motifs: one set of positive motifs from class  $\mathcal{C}_i$  and  $N_c - 1$  sets of negative motifs from the other classes. The  $N$  top ranked motifs from each set are then used to scan the  $M$  CRMs for conserved instances with scanning threshold  $S_s$ . Each CRM  $E$  is scored with respect to these  $N_c$  sets of motifs by:

$$S(E) = \sum_{j=1}^{N_c} (2\delta_{j,i} - 1) S_N^{\mathcal{C}_j}(E) \quad (1)$$

where  $S_N^{\mathcal{C}_j}(E)$  is the CRM score for the  $N$  top motifs of class  $\mathcal{C}_j$  as defined later in the *Methods* section, and  $\delta_{j,i} = 1$  if  $j = i$ , and 0 otherwise. This score simply gives positive contributions if positive motifs are found on the CRM, and negative contributions if negative motifs are found. This scoring procedure allows to rank the test CRM among the other  $M - 1$  CRMs. Ties are resolved by using the mean rank among equally scored CRMs. Here, the rank is used rather than the raw score of the test CRM to avoid any artifact stemming from normalization problems. Indeed, the raw score is dependent on the generated motifs, which differ at each step of the LOOCV. This procedure is repeated over all  $M$  CRMs, yielding a corresponding list of  $M$  ranks. This list is finally used to build a ROC curve discriminating True Positives (CRMs from class  $\mathcal{C}_i$ ) from False Positives (the other CRMs). The discrimination is quantified by the area under the ROC curve for a False Positive Rate  $FPR \leq 20\%$ , which we note AUC20 and that we want to maximize.

In our case, we used a 2D parameters grid with  $S_g$  varying between 7 and 13 bits by steps of 1, and  $S_s$  varying between  $S_g - 5$  and  $S_g$  by steps of 1. Both *Felsenstein* and *Halpern-Bruno* models were used for motif generation. For each parameter set, the number of motifs used for scanning was increased from 1 to a maximum number of 10 (actually never at-

tained) until the addition of a new motif decreased the AUC20, yielding an optimal number of motifs  $N$ . Finally, for each class, the parameter set  $\{S_g, S_s, N\}$  yielding the highest AUC20 was selected as the best parameter set.

### Motifs identification

In order to identify the known TFs that might correspond to the *de novo* generated motifs, we used Transfac database [54]. In order to avoid uninformative matches, we kept Transfac motifs that had an information content greater than 8 bits, a threshold approximately corresponding to 4 conserved nucleotides. This gets rid of 170 vertebrate motifs and 32 insect motifs, yielding a total of respectively 765 and 37 motifs.

Each *de novo* motif was compared to all Transfac motifs from the corresponding clade (vertebrates or insects) using the PWM distance introduced in [13]. During the comparison, motifs are shifted to find the best match, with a minimal match of 5 nts. The shift is simply introduced by adding flanking nucleotides with background frequency on either side. the closest candidate was kept for identification.

### Main program

The main program is written in C++ and adapted from the program used in a previous study [13]. It is distributed under the GNU GPL licence and available as a git repository at <http://github.com/hrouault/Imogene>.

### Binding site and CRM scores

Binding sites as well as CRMs are scored in the same manner as in [13].

For a given PWM with the weight  $w_{i,b}$  for the base  $b$  at position  $i$ , the score of a sequence  $s_i$  is defined as:

$$S = \sum_i w_{i,s_i} \quad (2)$$

A sequence is considered as a binding site when  $S > S_{th}$ , where  $S_{th}$  is the score threshold defined by the user of *Imogene*.

An CRM  $E$  is scored with respect to a set of motifs  $m_i$  by:

$$S(E) = \sum_i n(E, m_i) \log(\lambda_i^t / \lambda_i^b) \quad (3)$$

where  $n(E, m_i)$  is the number of binding sites for the motif  $m_i$  on  $E$  and  $\lambda_i^t, \lambda_i^b$  and the average number of binding sites per base on the training set and background respectively. It is important to note that the previously found motif binding sites are masked when scanning with successive motifs. Thus motifs with lower ranks that resemble high-ranking motifs, but could not be fused properly, do not increase artificially the CRM weight by predicting the same binding sequences twice.

### Evolutionary models

*Imogene* can use two different evolutionary models, which vary in complexity and computational time. To simplify the computation, we suppose in both models that the bases within a site evolve independently from each other.

**Felsenstein** The simplest models of nucleotides evolution are copied on model of neutral selection. This procedure has been proposed by Sinha *et al* [26, 45] with the Felsenstein model of neutral evolution [46]. In this TFBS evolution model, the transition probability from nucleotide  $b$  to  $b'$  at position  $i$  in two sites at evolutionary distance  $d$  writes:

$$p_{b \rightarrow b'}^i = q \delta_{b,b'} + (1 - q) w_{i,b'} \quad (4)$$

where  $\delta_{b,b'}$  is the Kronecker symbol,  $w_{i,b'}$  is the mean frequency of base  $b'$  at position  $i$  of the site (as given by the PWM model), and  $q$  is the probability of conservation for an evolutionary distance  $d$  under neutral selection (see below).

When two species are close to one another,  $q \sim 1$  and the probability that the observed bases are identical is high. On the contrary, when the two considered species are distant ( $q \sim 0$ ), the observed bases are uncorrelated and reflect the PWM probabilities  $w_{i,b}$ .

The probability of conservation  $q$  can then be computed within this model by setting the PWM probabilities  $w_{i,b}$  to the mean genomic frequencies  $\pi_b$ :

$$q = \exp\left(-\frac{d}{1/2 + 4\pi_{A,T}\pi_{C,G}}\right) \quad (5)$$

with  $\pi_{A,T}$  (resp.  $\pi_{C,G}$ ) the common genomic frequency of A and T (resp. G and C).

**Halpern-Bruno** The Halpern-Bruno model (HB) differs in two ways from the simplest *Felsenstein* model. It uses the more complex Hasegawa, Kishino and Yano model (HKY) [55] for the neutral evolution of nucleotides and adds a fixation probability based on fitness differences for the evolution of nucleotides within the TFBS.

The HKY model improves on the Felsenstein model by taking into account the observed dependence of the mutation rate on the chemical nature of the bases. Mutations between bases of the same chemical nature (purine or pyrimidine), also called transitions, are generally more frequent than the other type of mutations, called transversions. This is encapsulated in the HKY model by the parameter  $\kappa$  which is the ratio of transition rate to the transversion rate. It is measured to be  $\kappa = 2$  in flies and  $\kappa = 3.7$  in mammals.

Within a TFBS, the HB model extends the HKY model to take into account an additional purifying selection on the nucleotide identities. It is formulated by the following transition probabilities:

$$p_{b \rightarrow b'} = \exp(t\mathbf{H})_{b,b'} \quad (6)$$

where  $\mathbf{H}$  is the rate matrix defined by:

$$H_{b,b'} = \begin{cases} \pi_b h_{b' \rightarrow b} & \text{if } b \neq b' \\ -\sum_{b' \neq b} H_{b,b'} & \text{if } b = b' \end{cases} \quad (7)$$

The evolutionary time  $t$  is expressed in term of the evolutionary distance by:

$$t = \frac{d}{1/2 + 4\kappa \pi_{A,T} \pi_{C,G}} \quad (8)$$

Finally, the transition rates are defined by:

$$h_{b \rightarrow b'} = \frac{w_{b'}}{\pi_{B'}} \frac{\log\left(\frac{\pi_B w_{b'}}{\pi_{B'} w_b}\right)}{w_{b'}/\pi_{B'} - w_b/\pi_B} \alpha_{b \rightarrow b'} \quad (9)$$

with  $\alpha_{b \rightarrow b'} = \kappa$  for a transition and  $\alpha_{b \rightarrow b'} = 1$  for a transversion.

## Inference

The algorithm performs Bayesian inference in order to infer the frequencies  $w_{i,b}$  based on observations of binding sites, as previously described in [13]. In the Bayesian framework, one can write the Posterior

distribution of  $w_i$  given the observation of a set of aligned nucleotides  $\{\mathcal{A}\}$  as

$$\mathcal{P}(w_i|\{\mathcal{A}\}) \propto \prod_{a \in \{\mathcal{A}\}} \mathcal{P}(a|w_{i,b}) \prod_{b \in \{A,T,C,G\}} w_{i,b}^{\alpha_b - 1} \quad (10)$$

where we omit the normalization factor. The first product is the likelihood function and the second one is the prior, taken to be a Dirichlet distribution with pseudo-counts  $\alpha_\beta$  computed as in [13]. In the idealistic case where the aligned nucleotides would represent independent observations (infinitely distant species), the likelihood reduces to a multinomial distribution and the posterior writes:

$$\mathcal{P}(w_i|\{\mathcal{A}\}) \propto \prod_{b \in \{A,T,C,G\}} w_{i,b}^{N_b + \alpha_b - 1} \quad (11)$$

where  $N_b$  is the number of times the base  $b$  is observed in  $\{\mathcal{A}\}$ . This formula allows simple analytic formulations for the estimator of mean and maximum posterior probability. The estimator of the mean posterior distribution is expressed as:

$$\tilde{w}_{i,b} = \frac{N_b + \alpha_b}{\sum_b N_b + \alpha_b} \quad (12)$$

The estimator of maximum probability has the same value if one uses the “transformed” posterior where  $\alpha_b \rightarrow \alpha_b + 1$ .

In our case, we take into account the evolutionary history that correlates the orthologous sites to the one on the reference species. In that case,  $\mathcal{P}(a|w_{i,b})$  is a polynomial function of the  $w_{i,b}$ ’s and generally lacks a simple analytical formulation.

## Mean Posterior Estimation

The transformed posterior distribution is maximized by using a simplex algorithm implemented by the GNU GSL to fit the mean estimator. The initial value for the estimation is taken to be the mean estimator in the independent species regime given in Eq. (11). This allows to start close to the quadratic region and ensures fast convergence. The estimation was consistently retrieved when using an independent MCMC approach to compute the mean estimator.

## A simple example of nucleotide inference using the two evolutionary models

To illustrate the inference of ancestral nucleotides and the main features of the two models, we con-

sider a dinucleotidic genome with bases  $X$  and  $Y$  and the simple alignment shown in Figure S5 with an ancestral species at equal evolutionary distance from the reference species and a daughter species. We suppose that the observed nucleotide at position  $i$  of an observed binding site is  $X$  both in the reference and the orthologous species.

Our goal is to infer the frequencies  $w_Y$  and  $w_X = 1 - w_Y$ . First, there are two simple cases. For  $d = 0$ , the observations of the same nucleotide in the two evolutionary branches really constitute only one observation of  $X$ . On the contrary, for very long evolutionary branches  $d \rightarrow \infty$ , the two instances of nucleotide  $X$  form two independent observations. Using the previous result (Eq. (12)) with  $\alpha_X = \alpha_Y = \alpha$ , the estimator of the maximum transformed posterior distribution for  $N_X$  and  $N_Y$  independent instances of  $X$  and  $Y$  is:

$$w_Y = \frac{N_Y + \alpha}{N_Y + N_X + 2\alpha} \quad (13)$$

Thus, for  $d = 0$ , the inferred frequency is:

$$w_Y = \frac{\alpha}{1 + 2\alpha} \quad (14)$$

while for  $d \rightarrow \infty$ , it tends toward:

$$w_Y = \frac{\alpha}{2 + 2\alpha} \quad (15)$$

Between these two extreme cases, an evolutionary model has to be used to estimate  $w_Y$ , for finite evolutionary branches of length  $d$ .

For the Felsenstein model, the likelihood function writes:

$$\begin{aligned} \mathcal{P}(\mathcal{A}|w) &= w_X [q + (1 - q)w_X]^2 + w_Y (1 - q)^2 w_X^2 \\ &= q^2 w_X + (1 - q^2) w_X^2 \end{aligned} \quad (16)$$

where  $\mathcal{A}$  stands for the simple alignment drawn on fig. S5 and we used  $w_X = 1 - w_Y$ . From this expression it can clearly be seen that the evolutionary model simply interpolates between the independent species case ( $d \rightarrow \infty$ ,  $q = 0$ ) where there are two observations of base  $X$ :  $\mathcal{P}(w|\mathcal{A}) = w_X^2$ , and the fully correlated case ( $d = 0$ ,  $q = 1$ ) where the two species merge and we have only one observation:  $\mathcal{P}(w|\mathcal{A}) = w_X$ . The corresponding mean,  $w_{Y,me}$  and maximum likelihood,  $w_{Y,ma}$  analytic estimates for fi-

nite  $d$  read

$$\begin{aligned} w_{Y,me} &= \frac{\alpha}{2} \frac{1 + q^2}{\alpha + 1 + \alpha q^2} \\ w_{Y,ma} &= \frac{1}{4(\alpha + 1)(1 - q^2)} \left[ 3\alpha + 2 - (\alpha + 1)q^2 \right. \\ &\quad \left. - \sqrt{[\alpha + 2 - 3(\alpha + 1)q^2]^2 + 8q^2(1 - q^2)(\alpha + 1)^2} \right] \end{aligned}$$

Note that for the maximum likelihood estimate,  $w_{Y,ma}$ , the prior exponent  $\alpha + 1$  has been used instead of  $\alpha$  as explained above. So, the two estimates coincides at  $q = 0$  and  $q = 1$ . Both estimates are plotted as of function of the evolutionary distance  $d$  in Figure S5 ( $\alpha = 0.1$ ).

For the Halpern-Bruno model, the analogous results have been computed numerically and are also shown for comparison in Figure S5. The Halpern-Bruno model results are seen to be closer to the large distance limit than the Felsenstein model ones. Moreover, the difference between the nature of the estimates is seen to dominate the difference between the evolutionary models.

## Phylogenetic trees

The phylogenetic trees used within *Imogene* are displayed in Figure 2. For drosophilae, the distances are taken from Heger and Pontig [56]. For eutherian, they are extracted from the Ensembl website ([www.ensembl.org](http://www.ensembl.org)).

## Conservation requirements

*Imogene* builds PWM from binding sites that have conserved instances in different species. The conservation requirements is that orthologous instances are found in at least 3 distant species, including the reference species. For mammals, the 5 following groups of related species are composed of: *Mus musculus* and *Rattus norvegicus*; *Callithrix jacchus*, *Macaca mulatta*, *Pongo abelii*, *Gorilla gorilla*, *Homo sapiens* and *Pan troglodytes*; *Bos taurus*; *Sus scrofa*; *Canis familiaris*; *Equus caballus*. Similarly for flies, there are 5 groups composed of: *Drosophilae melanogaster*, *sechellia*, *simulans*, *yakuba* and *erecta*; *Drosophila ananassae*; *Drosophilae pseudoobscura* and *persimilis*; *Drosophila willistoni*; *Drosophilae grimshawi*, *mojavensis* and *virilis*.

A site instance must be found in at least 3 of these 5 groups (with an allowed shift of up to 20

nt with the reference species) to be considered conserved by *Imogene*.

## Selection of CRMs

A number  $N$  of motifs are used to scan the genome for conserved instances above a given threshold. Instances are ranked according to their genomic position and grouped in successive CRMs of size  $L$  such as to maximize clustering. The position  $E_i$  of the center of the enhancer  $i$  is chosen to be the center of the motifs cluster:

$$E_i = \frac{X_1 + X_N + w - 1}{2} \quad (17)$$

where  $X_1$  and  $X_N$  are the starting positions of the first and last TFBSs in the cluster and  $w$  is the width of the motif.

## Statistical analysis

All statistical analyses were performed using R [57].

## Authors contributions

H. R., M. S., F. S., V. H. designed research. H. R., M. S. performed research and wrote the software. H. R., M. S., F. S., V. H. wrote the paper.

## References

- Davidson EH: *The regulatory genome: gene regulatory networks in development and evolution*. Burlington, MA: Academic 2006, [http://www.loc.gov/catdir/enhancements/fy0668/2006445256-d.html].
- Dorer DE, Nettelbeck DM: **Targeting cancer by transcriptional control in cancer gene therapy and viral oncolysis**. *Adv Drug Deliv Rev* 2009, **61**(7-8):554-71.
- Hardison RC, Taylor J: **Genomic approaches towards finding cis-regulatory modules in animals**. *Nat. Rev. Genet.* 2012, **13**(7):469-483.
- Lelli KM, Slattery M, Mann RS: **Disentangling the many layers of eukaryotic transcriptional regulation**. *Annu. Rev. Genet.* 2012, **46**:43-68.
- Levine M: **Transcriptional enhancers in animal development and evolution**. *Curr. Biol.* 2010, **20**(17):R754-763.
- Arnosti DN, Kulkarni MM: **Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards?** *J Cell Biochem* 2005, **94**(5):890-8.
- Swanson CI, Evans NC, Barolo S: **Structural rules and complex regulatory circuitry constrain expression of a Notch- and EGFR-regulated eye enhancer**. *Dev Cell* 2010, **18**(3):359-70.
- Johnson DS, Mortazavi A, Myers RM, Wold B: **Genome-wide mapping of in vivo protein-DNA interactions**. *Science* 2007, **316**(5830):1497-502.
- 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):823-37.
- Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP, Lee W, Mendenhall E, O'Donovan A, Presser A, Russ C, Xie X, Meissner A, Wernig M, Jaenisch R, Nusbaum C, Lander ES, Bernstein BE: **Genome-wide maps of chromatin state in pluripotent and lineage-committed cells**. *Nature* 2007, **448**(7153):553-60.
- Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, Afzal V, Ren B, Rubin EM, Pennacchio LA: **ChIP-seq accurately predicts tissue-specific activity of enhancers**. *Nature* 2009, **457**(7231):854-8.
- Amano T, Sagai T, Tanabe H, Mizushima Y, Nakazawa H, Shiroishi T: **Chromosomal dynamics at the Shh locus: limb bud-specific differential regulation of competence and active transcription**. *Dev. Cell* 2009, **16**:47-57.
- Rouault H, Mazouni K, Couturier L, Hakim V, Schweisguth F: **Genome-wide identification of cis-regulatory motifs and modules underlying gene coregulation using statistics and phylogeny**. *Proc Natl Acad Sci U S A* 2010, **107**(33):14615-20.
- Wasserman WW, Sandelin A: **Applied bioinformatics for the identification of regulatory elements**. *Nat. Rev. Genet.* 2004, **5**(4):276-287.
- Stormo G, Fields D: **Specificity, free energy and information content in protein-DNA interactions**. *Trends in biochemical sciences* 1998, **23**(3):109-113.
- Zinzen RP, Girardot C, Gagneur J, Braun M, Furlong EE: **Combinatorial binding predicts spatio-temporal cis-regulatory activity**. *Nature* 2009, **462**:65-70.
- Su J, Teichmann SA, Down TA: **Assessing computational methods of cis-regulatory module prediction**. *PLoS Comput. Biol.* 2010, **6**(12):e1001020.
- Elnitski L, Jin VX, Farnham PJ, Jones SJ: **Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques**. *Genome Res.* 2006, **16**(12):1455-1464.
- Aerts S: **Computational strategies for the genome-wide identification of cis-regulatory elements and transcriptional targets**. *Curr. Top. Dev. Biol.* 2012, **98**:121-145.
- Machanick P, Bailey TL: **MEME-ChIP: motif analysis of large DNA datasets**. *Bioinformatics* 2011, **27**(12):1696-1697.

21. Berman B, Nibu Y, Pfeiffer B, Tomancak P, Celniker S, Levine M, Rubin G, Eisen M: **Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the *Drosophila* genome.** *Proceedings of the National Academy of Sciences* 2002, **99**(2):757.
22. Halfon M, Grad Y, Church G, Michelson A: **Computation-based discovery of related transcriptional regulatory modules and motifs using an experimentally validated combinatorial model.** *Genome research* 2002, **12**(7):1019.
23. Rebeiz M, Reeves N, Posakony J: **SCORE: A computational approach to the identification of cis-regulatory modules and target genes in whole-genome sequence data.** *Proceedings of the National Academy of Sciences* 2002, **99**(15):9888.
24. Schroeder M, Pearce M, Fak J, Fan H, Unnerstall U, Emberly E, Rajewsky N, Siggia E, Gaul U: **Transcriptional control in the segmentation gene network of *Drosophila*.** *PLoS biology* 2004, **2**:1396–1410.
25. Moses AM, Chiang DY, Pollard DA, Iyer VN, Eisen MB: **MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model.** *Genome Biol* 2004, **5**(12):R98.
26. Siddharthan R, Siggia E, van Nimwegen E: **PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny.** *PLoS Comput Biol* 2005, **1**(7):e67.
27. 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.
28. Pierstorff N, Bergman C, Wiehe T: **Identifying cis-regulatory modules by combining comparative and compositional analysis of DNA.** *Bioinformatics* 2006, **22**(23):2858.
29. Herrmann C, Van de Sande B, Potier D, Aerts S: **i-cisTarget: an integrative genomics method for the prediction of regulatory features and cis-regulatory modules.** *Nucleic Acids Res* 2012.
30. Nazina A, Papatsenko D: **Statistical extraction of *Drosophila* cis-regulatory modules using exhaustive assessment of local word frequency.** *BMC bioinformatics* 2003, **4**:65.
31. Abnizova I, te Boekhorst R, Walter K, Gilks W: **Some statistical properties of regulatory DNA sequences, and their use in predicting regulatory regions in the *Drosophila* genome: the fluffy-tail test.** *BMC bioinformatics* 2005, **6**:109.
32. Chan B, Kibler D: **Using hexamers to predict cis-regulatory motifs in *Drosophila*.** *BMC bioinformatics* 2005, **6**:262.
33. Leung G, Eisen M, Provart N: **Identifying Cis-Regulatory Sequences by Word Profile Similarity.** *PLoS ONE* 2009, **4**(9):e6901.
34. Kantorovitz M, Kazemian M, Kinston S, Miranda-Saavedra D, Zhu Q, Robinson G, G
- ottgens B, Halfon M, Sinha S: **Motif-Blind, Genome-Wide Discovery of cis-Regulatory Modules in *Drosophila* and Mouse.** *Developmental Cell* 2009, **17**(4):568–579.
35. Brody T, Yavatkar AS, Kuzin A, Kundu M, Tyson LJ, Ross J, Lin TY, Lee CH, Awasaki T, Lee T, Odenwald WF: **Use of a *Drosophila* genome-wide conserved sequence database to identify functionally related cis-regulatory enhancers.** *Dev Dyn* 2012, **241**:169–89.
36. Xie X, Lu J, Kulbokas E, Golub T, Mootha V, Lindblad-Toh K, Lander E, Kellis M: **Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals.** *Nature* 2005, **434**(7031):338–345.
37. Ettwiller L, Paten B, Souren M, Loosli F, Wittbrodt J, Birney E: **The discovery, positioning and verification of a set of transcription-associated motifs in vertebrates.** *Genome Biology* 2005, **6**(12):R104.
38. Stark A, Lin M, Kheradpour P, Pedersen J, Parts L, Carlson J, Crosby M, Rasmussen M, Roy S, Deoras A, et al.: **Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures.** *Nature* 2007, **450**(7167):219.
39. Wang T, Stormo G: **Combining phylogenetic data with co-regulated genes to identify regulatory motifs.** *Bioinformatics* 2003, **19**(18):2369.
40. Grad Y, Roth F, Halfon M, Church G: **Prediction of similarly acting cis-regulatory modules by subsequence profiling and comparative genomics in *Drosophila melanogaster* and *D. pseudoobscura*.** *Bioinformatics* 2004, **20**(16):2738.
41. Zhao G, Schrieffer L, Stormo G: **Identification of muscle-specific regulatory modules in *Caenorhabditis elegans*.** *Genome research* 2007, **17**(3):348.
42. Busser BW, Taher L, Kim Y, Tansey T, Bloom MJ, Ovcharenko I, Michelson AM: **A machine learning approach for identifying novel cell type-specific transcriptional regulators of myogenesis.** *PLoS Genet* 2012, **8**(3):e1002531.
43. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S, Gil L, Gordon L, Hendrix M, Hourlier T, Johnson N, Kahari AK, Keefe D, Keenan S, Kinsella R, Komorowska M, Koscielny G, Kulesha E, Larsson P, Longden I, McLaren W, Muffato M, Overduin B, Pignatelli M, Pritchard B, Riat HS, Ritchie GR, Ruffier M, Schuster M, Sobral D, Tang YA, Taylor K, Trevanion S, Vandrovicova J, White S, Wilson M, Wilder SP, Aken BL, Birney E, Cunningham F, Dunham I, Durbin R, Fernandez-Suarez XM, Harrow J, Herrero J, Hubbard TJ, Parker A, Proctor G, Spudich G, Vogel J, Yates A, Zadiisa A, Searle SM: **Ensembl 2012.** *Nucleic Acids Res.* 2012, **40**(Database issue):84–90.
44. Clark A, Eisen M, Smith D, Bergman C, Oliver B, Markow T, Kaufman T, Kellis M, Gelbart W, Iyer V, et al.: **Evolution of genes and genomes on the *Drosophila* phylogeny.** *Nature* 2007, **450**(7167):203–218.
45. Sinha S, van Nimwegen E, Siggia ED: **A probabilistic method to detect regulatory modules.** *Bioinformatics* 2003, **19 Suppl 1**:i292–301.

46. Felsenstein J: **Evolutionary trees from DNA sequences: a maximum likelihood approach.** *J Mol Evol* 1981, **17**(6):368–76.
47. Halpern AL, Bruno WJ: **Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies.** *Mol Biol Evol* 1998, **15**(7):910–7.
48. Kimura M: **On the probability of fixation of mutant genes in a population.** *Genetics* 1962, **47**:713–9.
49. May D, Blow MJ, Kaplan T, McCulley DJ, Jensen BC, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Afzal V, Simpson PC, Rubin EM, Black BL, Bristow J, Pennacchio LA, Visel A: **Large-scale discovery of enhancers from human heart tissue.** *Nat. Genet.* 2011, **44**:89–93.
50. Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, Kuan S, Wagner U, Dixon J, Lee L, Lobanenkov VV, Ren B: **A map of the cis-regulatory sequences in the mouse genome.** *Nature* 2012.
51. Kiyota T, Kato A, Altmann CR, Kato Y: **The POU homeobox protein Oct-1 regulates radial glia formation downstream of Notch signaling.** *Dev Biol* 2008, **315**(2):579–92.
52. Neron B, Menager H, Maufrais C, Joly N, Maupetit J, Letort S, Carrere S, Tuffery P, Letondal C: **Mobyle: a new full web bioinformatics framework.** *Bioinformatics* 2009, **25**(22):3005–3011.
53. Menoret D, Santolini M, . . . , Payre F, Plaza S: **Decoding the transcriptional program of epidermal cell morphogenesis.** (*Submitted*) 2012.
54. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, Voss N, Stegmaier P, Lewicki-Potapov B, Saxel H, Kel AE, Wingender E: **TRANSFAC and its module TRANSCOMP: transcriptional gene regulation in eukaryotes.** *Nucleic Acids Res* 2006, **34**(Database issue):D108–10.
55. Hasegawa M, Kishino H, Yano T: **Dating of the human-ape splitting by a molecular clock of mitochondrial DNA.** *J Mol Evol* 1985, **22**(2):160–74.
56. Heger A, Ponting CP: **Variable strength of translational selection among 12 Drosophila species.** *Genetics* 2007, **177**:1337–1348.
57. R Development Core Team: *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria 2011, [http://www.R-project.org/]. [ISBN 3-900051-07-0].

## Figures

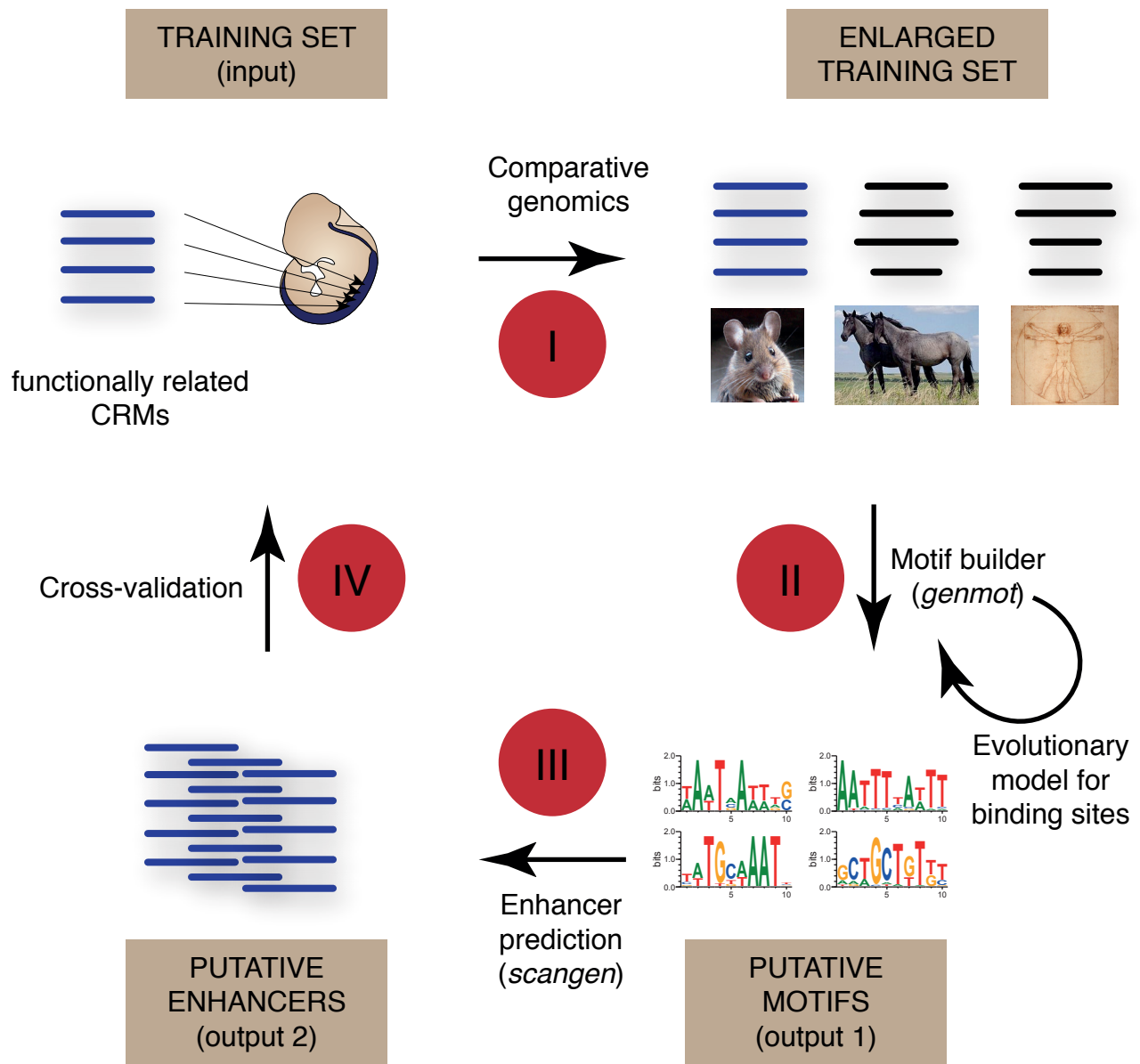


Figure 1: **Imogene workflow**. The algorithm takes as input a list of functionally related CRMs. Homologous sequences from closely related species are automatically retrieved (I) and scanned in order to generate a list of putative transcription factor motifs (II). These motifs fuel the last step consisting in the inference of related novel CRMs (III). These predicted CRMs can finally be compared to a set of test CRMs to evaluate the predictability power (IV).



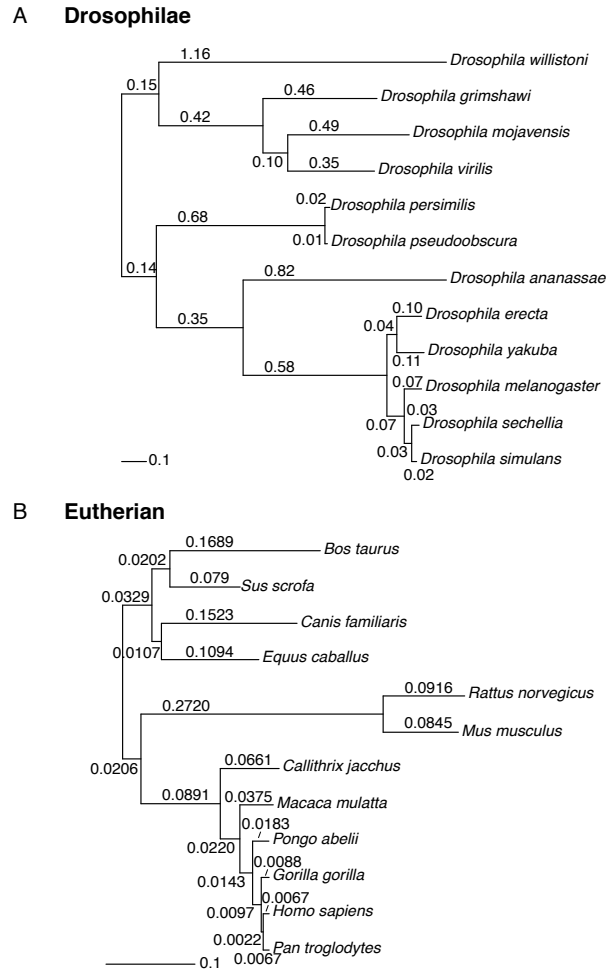


Figure 2: **Phylogenetic trees and phylogenetic distances used by *Imogene*.** The branch lengths represent the evolutionary distances  $d$  used by the evolutionary models at the motif construction stage.

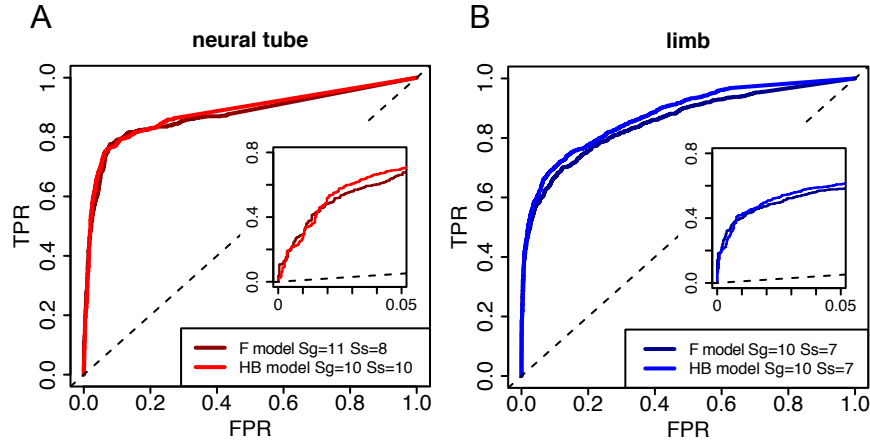


Figure 3: **Analysis of well characterized developmental processes.** We tested the algorithm on mammal CRMs expressed at E11.5 in neural tube (A) and limb (B). For each class, CRMs were divided into a training set and a test set. Motifs were learned on the training set and used to score CRMs from the test set along with background regions consisting of the CRMs 1kb flanking sequences. Finally, True Positive Rates or TPR (resp. False Positive Rates or FPR) were defined as the proportion of test set CRMs (resp. background sequences) recovered above a given score. ROC plots summarize the results averaged over 40 trials. Insets emphasize the  $\text{FPR} \leq 5\%$  region. Evolutionary models, along with thresholds  $S_g$  and  $S_s$  used for motifs generation and sequences scanning are indicated. F and HB models respectively stand for Felsenstein and Halpern-Bruno models. Black dashed lines show random discrimination.

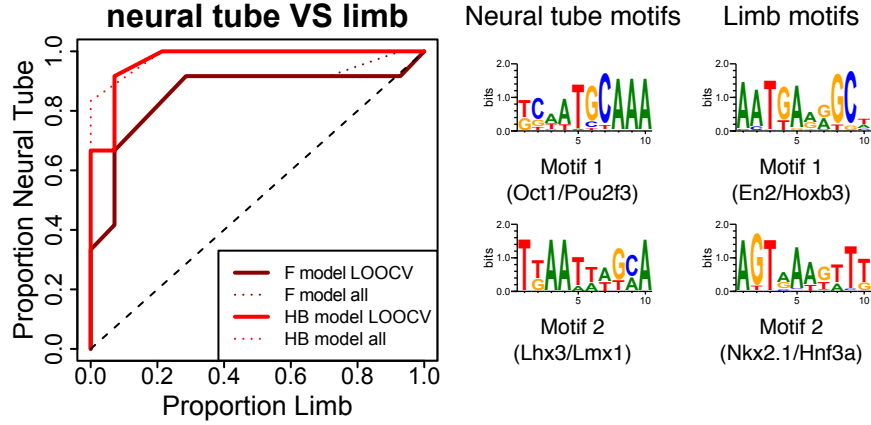


Figure 4: **Pattern recognition (mammals).** ROC plots showing the discrimination between limb and neural CRMs. Neural and limb classes are compared to each other. Thick lines correspond to a leave-one-out cross-validation (LOOCV) scheme with a score function based on the *de novo* generated motifs from *Imogene*, while colored dashed lines represent the discrimination based on the whole training set, systematically showing overfitting compare to LOOCV. Two evolutionary models are used: Felsenstein (solid dark red line,  $S_g = 11$ ,  $S_s = 9$ ) and Halpern-Bruno (solid light red line,  $S_g = 11$ ,  $S_s = 8$ ). Black dashed line show random discrimination.

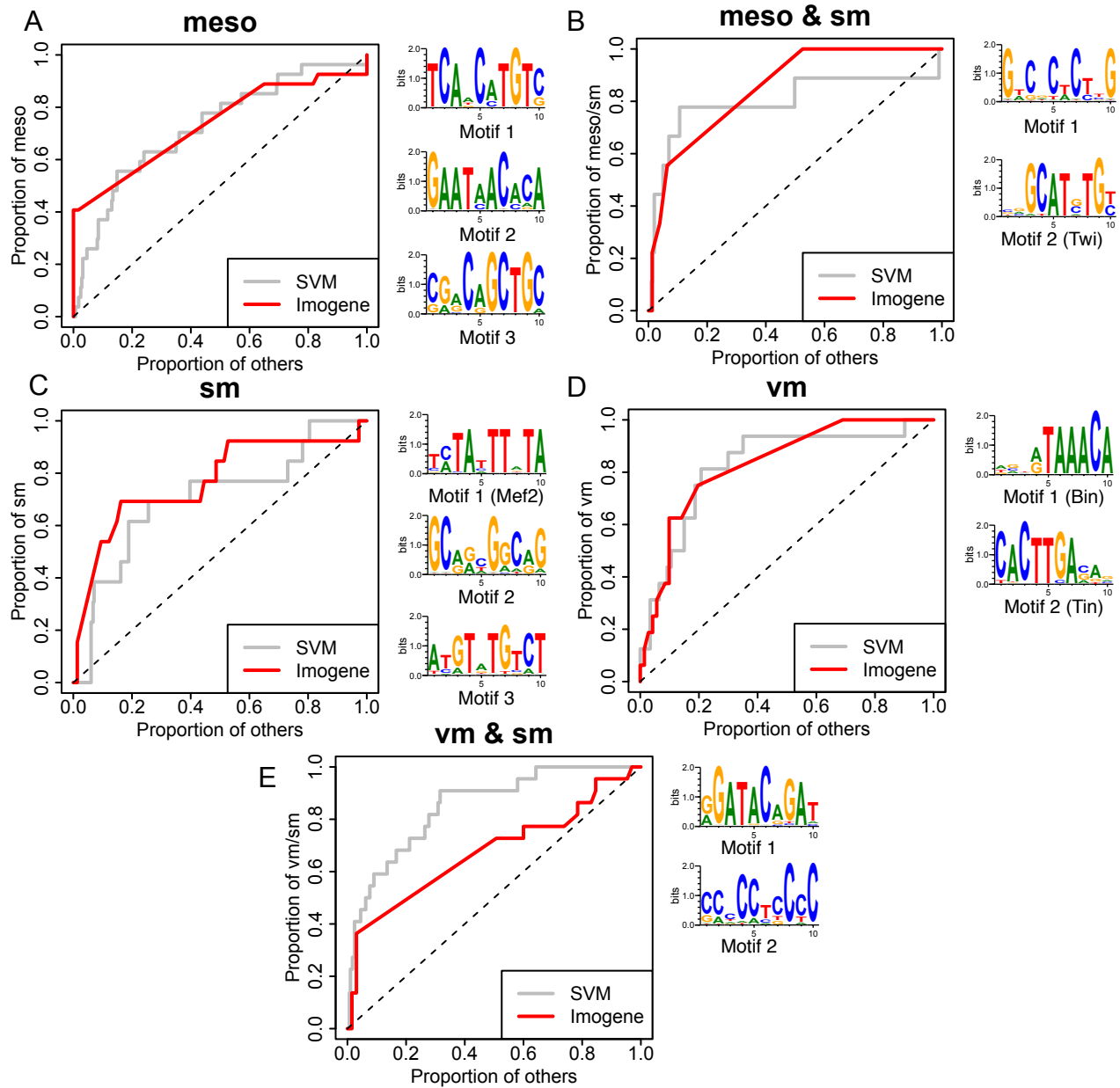


Figure 5: **Pattern recognition (Drosophila).** Recognition of classes of CRMs expressed in 5 tissue types: mesoderm (meso), somatic muscle (sm), visceral muscle (vm), mesoderm and somatic muscle (meso & sm) and visceral and somatic muscle (vm & sm). ROC plots are obtained using a leave-one-out cross-validation scheme. Two classifiers are compared: a Support Vector Machine using 15 ChIPseq peak heights (grey), and *Imogene* using the *de novo* generated motifs with Felsenstein evolutionary model (red). The following thresholds were used: meso ( $S_g = 12$ ,  $S_s = 12$ ), meso & sm ( $S_g = 10$ ,  $S_s = 10$ ), sm ( $S_g = 9$ ,  $S_s = 4$ ), vm ( $S_g = 10$ ,  $S_s = 10$ ), vm & sm ( $S_g = 11$ ,  $S_s = 8$ ).

\* Execution mode ? genmot: Generate motifs from a training set

#### General options

\* Family of species to consider ? Eutherians

\* Width of the motifs ? 10

\* Allowed shift of a binding site position in orthologous species ?

20

#### Genmot options

\* Evolutionary model used for motif generation ? Felsenstein model

\* Threshold used for motif generation ? 11.0

\* Threshold used to scan training set sequences for display ? 8.0

\* Training set sequences coordinates ?

paste

upload

EDIT

CLEAR

Enter your data below:

```
chr8 91462919 91464123 CYLD-SALL1
chr4 99040833 99042291 APG4C-FOXD3
chr14 118834760 118836087 SOX21-ABCC4
chr18 69658816 69660452 TCF4(intragenic)
chr6 138199417 138201368 MGST1-LMO3
chr12 51291542 51292872 FOXG1B-PRKD1
```

#### Scangen options

\* Threshold used to scan the genome ? 8.0

\* Width of selected enhancers ? 1000

\* Number of motifs to consider at maximum ? 5

\* File containing a list of motif definitions ?

paste

upload

EDIT

CLEAR

Enter your data below:

Figure 6: **Web based interface : input web page.** A copy input web page for *Imogene* powered by the mobile bioinformatics framework is shown.

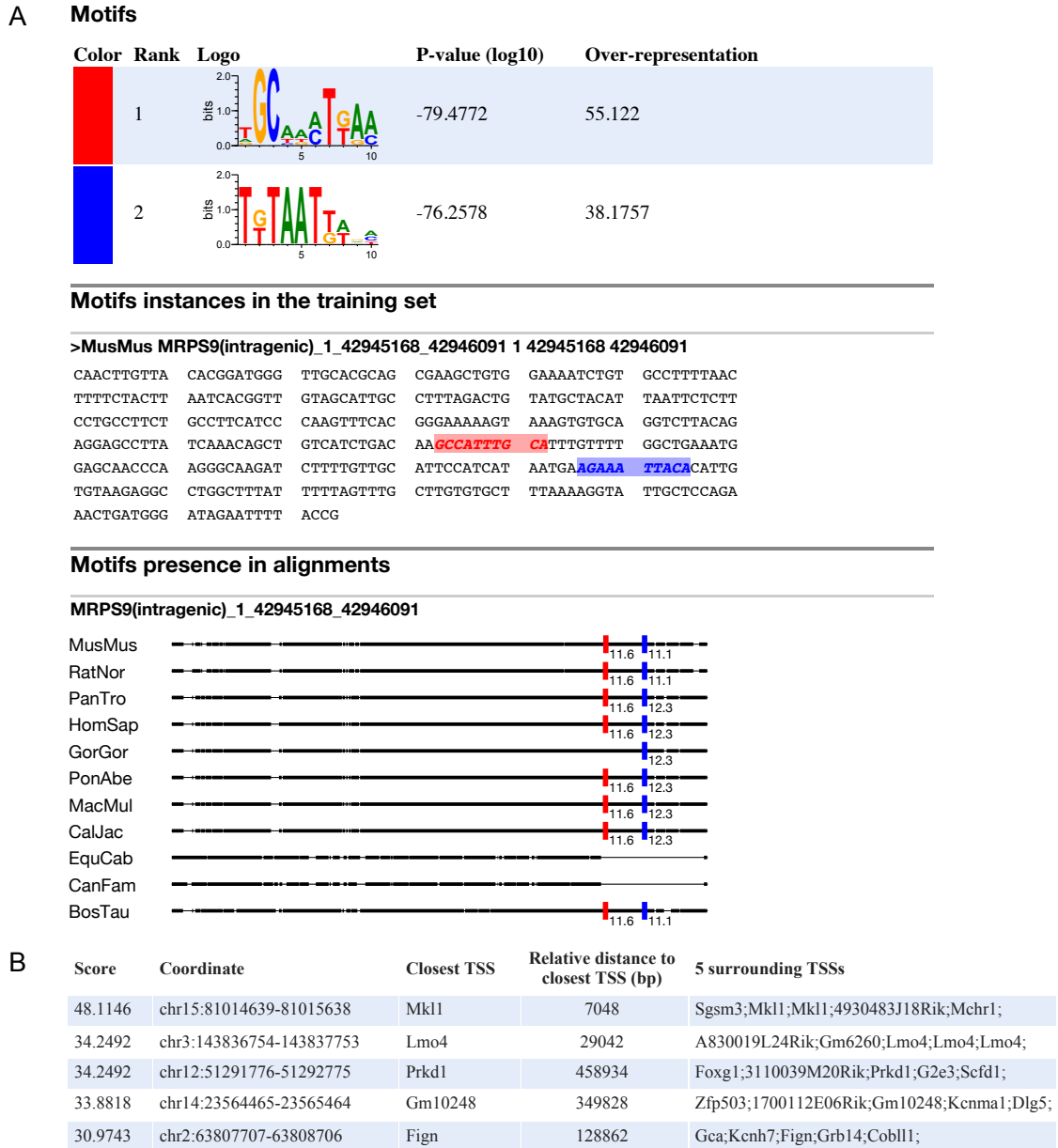


Figure 7: **Web based interface : output web page.** Example of an output web page for *Imogene* powered by the mobyle bioinformatics framework. A. Result page for the genmot mode. Two motifs were generated from the neural tube full training set (default is 5), using the same parameters as in Figure 3. Results are shown for the training set sequence MRPS9(intragenic). For display purposes, the beginning of the sequence, which contains no instances for the motifs, was cut in the middle panel. **In the alignments, thick lines correspond to sequences and thin lines to gaps.** B. Result page for the scangen mode. The two generated motifs were used to score putative regulatory sequences of 1kb in the mouse genome at optimal threshold  $S_s = 10$ . The 5 best ranking sequences are shown (default is 200).

## Supplementary Figures

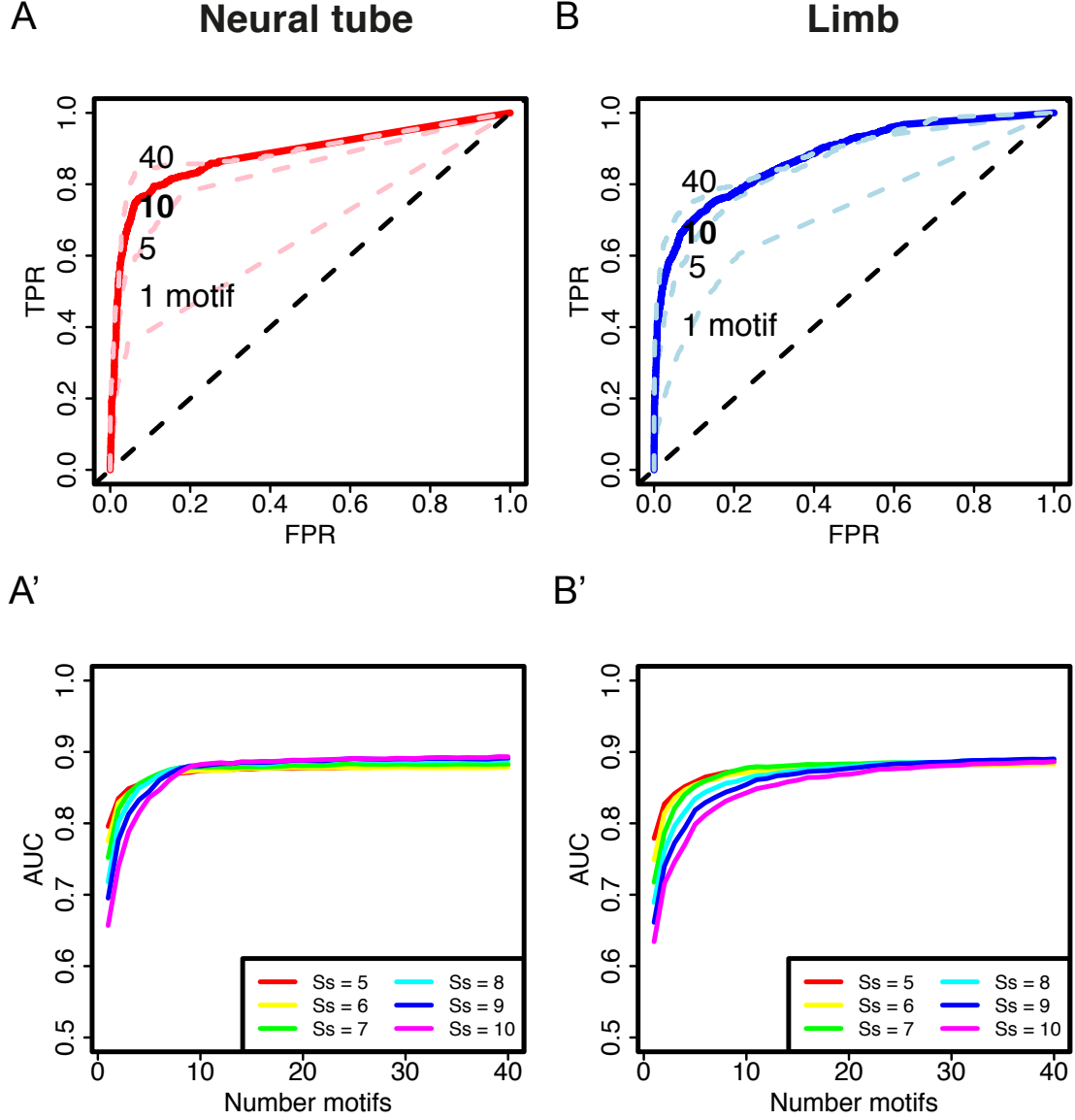


Figure S1: **Dependence of the predictions on the number of scoring motifs** ROC plots obtained at optimal scanning threshold using the Halpern-Bruno evolutionary model are shown for the neural tube (A) and limb (B) cases. Different curves are shown corresponding to sequences scored with different number of motifs: 1, 5 and 40 (light-color dashed lines), 10 (thick line). The ROC curves obtained for 10 motifs correspond to the ones shown in Fig. 3. To assess the degree of convergence, we computed the Area Under ROC Curve as a function of the number of motifs used (A',B',C'). We show the curves corresponding to the choice of different scanning thresholds  $S_s$ . In all cases, 10 motifs were sufficient for the AUC to converge. The optimal  $S_s$  was chosen as the one maximizing the AUC for 10 motifs.



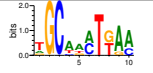
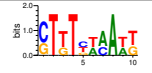
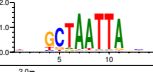
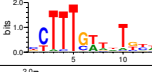
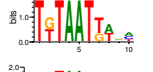
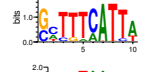
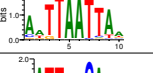
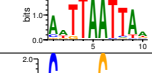
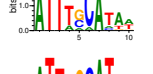
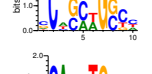
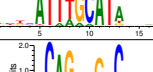
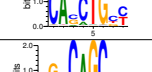
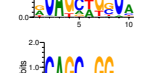
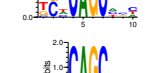
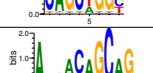
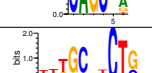
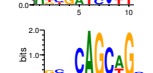
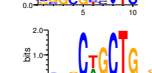
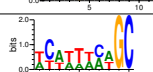
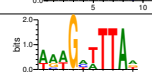
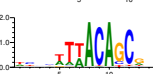
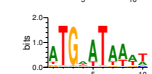
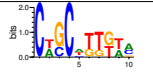
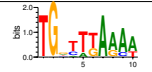
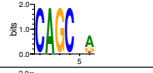
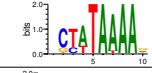
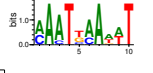
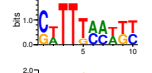
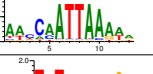
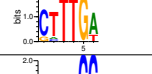
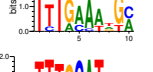
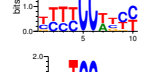
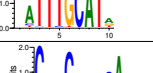
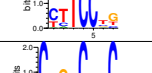
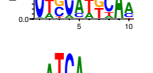
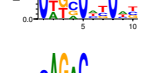
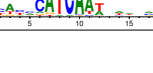
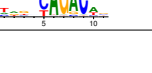
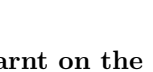

NEURAL		LIMB	
	Motif 1		Motif 1
	V\$CHX10_01		V\$TCF3_01
	Motif 2		Motif 2
	V\$LHX3_01		V\$LHX3_01
	Motif 3		Motif 3
	V\$OCT2_01		V\$MYOGENIN_Q6
	Motif 4		Motif 4
	V\$HEB_Q6		V\$CBF1_QX
	Motif 5		Motif 5
	V\$NEUROD_02		V\$NEUROD_02
	Motif 6		Motif 6
	V\$RHGX11_01		V\$POU1F1_Q6
	Motif 7		Motif 7
	V\$CBF1_QX		V\$TATA_C
	Motif 8		Motif 8
	V\$NKX61_01		V\$LEF1_Q2
	Motif 9		Motif 9
	V\$OCT1_B		V\$ETS2_Q6
	Motif 10		Motif 10
	V\$PBX1_04		V\$SMAD_Q6_01

Figure S2: **Motifs learnt on the full training sets.** The 10 best ranking motifs generated on the three CRMs training sets are shown together with the closest Transfac motifs (see *Methods* for a description of motif distance computation)

	Mot1	Mot2	Mot3	Mot4	Mot5	Mot6	Mot7	Mot8	Mot9	Mot10
ZIC4-ZIC1_9_91261697_91263041	2	0	1	0	0	0	2	1	1	0
TCF4(intragenic)_18_69658816_69660452	0	0	0	1	1	2	0	0	0	0
CEI-IRX1_13_72435297_72436784	3	4	2	2	0	3	0	1	3	2
NBEA(intragenic)_3_55768657_55770664	0	1	1	2	1	0	0	0	3	1
AKT3(intragenic)_1_179080168_179081586	2	1	1	3	2	1	2	0	2	0
FOXG1B-PRKD1_12_51291542_51292872	4	2	1	0	2	0	0	3	1	0
DACH1(intragenic)_14_98553917_98556433	5	0	2	4	0	2	3	1	3	2
FAM44A-CPEB2_5_42914188_42915270	1	0	1	3	1	1	2	0	1	0
IRX4-IRX2_13_73170587_73173631	0	0	0	2	0	0	0	0	4	0
EBF1(intragenic)_11_44469978_44471372	2	3	1	1	0	3	4	1	0	0
ATG4C-FOXO3_4_99240573_99241457	0	0	0	0	0	0	0	0	0	0
CYLD-SALL1_8_91462919_91464123	0	0	1	1	1	0	0	0	1	0
POU2F1(intragenic)_1_167864366_167866439	5	0	4	1	0	0	0	3	0	3
APG4C-FOXO3_4_99040833_99042291	0	0	0	0	0	0	0	0	0	0
MGC14798-HH114_2_115363420_115365044	2	2	2	0	1	0	0	2	0	0
MGST1-LMO3_6_138199417_138201368	5	1	1	1	1	3	0	3	1	1
APG4C-FOXO3_4_98961102_98962673	2	2	2	2	3	1	4	0	0	0
FLJ46321-RAEF4_4_73149468_73150526	0	0	2	4	0	1	1	1	0	1
TCF12(intragenic)_9_71823775_71824538	1	0	0	1	2	0	1	0	0	1
BMPER(intragenic)_9_23182371_23184296	2	1	1	1	0	2	1	0	1	0
SOX21-ABCC4_14_118834760_118836087	1	6	2	1	3	0	1	3	3	2
FANCL-BCL11A_11_25256346_25257683	0	2	1	0	3	0	2	0	0	0
DERA(intragenic)_6_137772070_137773298	1	5	0	1	1	1	2	0	0	0
MRPS9(intragenic)_1_42945168_42946091	1	1	0	2	1	1	1	0	0	0
YTHDF3-BHLHB5_3_16776170_16778776	2	2	0	1	0	0	0	0	1	0
STXBP6-NOVA1_12_47121350_47122759	1	4	2	3	0	2	2	0	0	0
IDH3B-CPXM1_2_130177541_130178125	0	0	0	0	0	0	0	0	0	0
LOC347487-SOX3_X_57972482_57973750	3	0	1	2	1	1	2	2	1	3

Figure S3: **Neural CRMs and motifs.** List of the neural CRMs used in this study. The number of motifs of different types on each CRM is given for the 10 best-ranking neural motifs shown in Figure S2

	Mot1	Mot2	Mot3	Mot4	Mot5	Mot6	Mot7	Mot8	Mot9	Mot10
hs1435_7_106105018_106107143	1	1	2	2	0	3	3	0	3	0
hs126_14_97485454_97486724	5	1	2	0	0	2	1	3	1	0
hs1477_2_59400401_59401189	2	0	1	1	2	1	1	1	1	0
hs521_1_91610325_91611486	0	1	2	4	0	1	0	0	8	0
mm422_2_4477190_4478921	0	0	1	1	0	0	0	0	0	0
hs1432_13_91326599_91329775	0	0	0	1	0	0	0	0	0	0
hs1433_3_30003454_30008202	8	4	8	5	5	5	4	5	6	1
hs208_9_100171947_100173392	2	2	3	3	5	1	1	1	4	2
hs1507_1_75765578_75770167	1	0	5	4	3	0	1	0	7	1
hs774_3_5329674_5330756	4	2	1	0	0	2	0	2	0	0
hs919_15_50496379_50498196	3	1	1	0	2	1	1	1	2	3
hs326_19_45568075_45569359	1	0	4	1	2	3	3	0	0	1
hs72_8_91978407_91979282	1	1	2	3	2	2	0	0	2	1
hs1484_4_97888231_97891318	0	1	0	0	2	0	0	1	1	0
mm423_2_4508631_4509808	0	0	1	0	0	0	0	0	0	0
mm428_5_38308981_38309833	0	2	1	0	0	0	1	0	4	0
hs741_3_66874217_66875516	4	2	1	0	1	2	0	2	1	0
hs1148_12_119941220_119942766	0	0	1	0	0	0	0	0	0	0
hs1109_13_79503055_79504129	2	1	1	1	0	1	1	0	1	0
hs2041_9_96280544_96283360	2	0	0	0	0	0	1	0	0	0
hs1473_13_56260379_56262548	1	1	7	1	8	0	0	0	1	0
hs1434_14_23833434_23842485	1	1	7	5	3	0	4	2	4	3
hs1465_6_51144711_51148222	0	2	6	3	1	1	1	0	3	0
mm94_6_122342623_122346341	0	0	2	1	1	0	0	0	3	0
hs1452_10_45612931_45614502	0	0	0	0	0	0	2	0	1	2
hs1468_10_125358093_125366026	0	0	1	0	0	1	0	0	0	0
hs1586_13_15640807_15642666	0	1	1	1	1	2	0	0	3	0
hs1273_12_9344323_9346407	2	2	2	1	3	4	1	4	3	1
hs1278_2_137073444_137074711	1	1	5	3	0	0	0	1	0	1
hs1500_14_22281464_22282917	0	0	4	1	2	0	0	0	2	0
mm458_15_63025492_63026343	2	0	1	0	0	1	0	0	4	0
hs388_12_26576441_26577229	4	4	2	2	1	0	0	0	0	1
hs1491_14_25804749_25806653	1	0	6	0	6	0	0	0	3	3
hs1428_3_99469238_99471067	0	2	4	2	2	0	1	0	3	0
hs1430_6_52917020_52919645	5	1	4	1	2	1	1	0	2	0
hs1475_16_72685882_72688547	0	0	1	0	1	0	1	4	0	1
hs1448_2_171555881_171562133	1	3	5	1	0	0	1	0	1	0
hs644_12_34884495_34885741	0	5	4	1	0	1	1	0	2	0

Figure S4: **Limb CRMs and motifs.** List of the limb CRMs used in this study. The number of motifs of different types on each CRM is given for the 10 best-ranking limb motifs shown in Figure S2

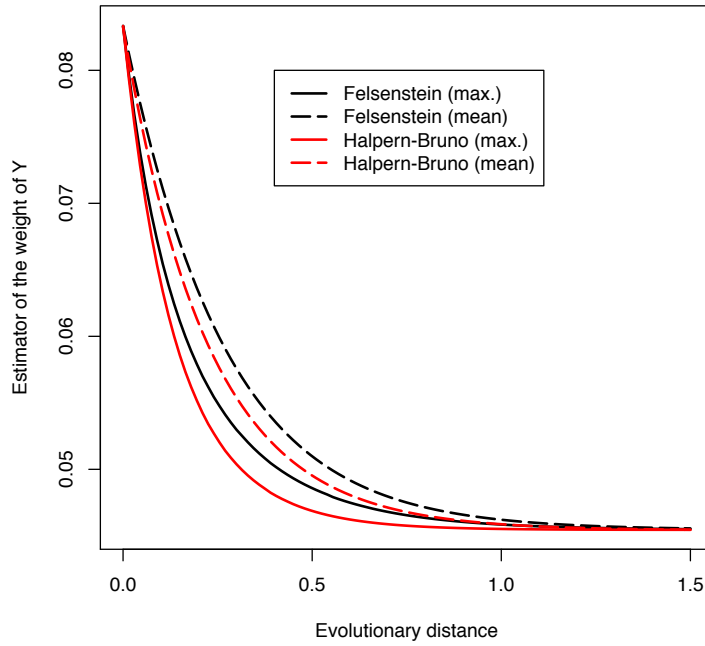


Figure S5: **Simple example of motif inference with Felsenstein and Halpern-Bruno evolutionary models** The inference of an ancestral base is compared in the simple case of two species at a phylogenetic distance  $d$  from their common ancestor, for a two nucleotide alphabet,  $X$  and  $Y$ . The mean and maximum likelihood estimate of observing  $Y$  in the common ancestor given that the two species share an  $X$  is shown as a function of evolutionary distance  $d$ , for the Felsenstein or Halpern-Bruno evolutionary models. The likelihood is always smaller with the Halpern-Bruno model, reflecting the model greater evolutionary rate.