Published online *Nucleic Acids Research, , Vol. , No.* ***1–27***
doi:10.1093/nar/gkn000

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

Hervé Rouault [1,2,+,*], **Marc Santolini** [3,+], **François Schweisguth** [1,2] **and Vincent Hakim** [3*]

[1] Institut Pasteur, Developmental Biology Department,75015 Paris, France, [2]CNRS, URA2578, F-75015 Paris, France, [3]Laboratoire de Physique Statistique, CNRS, École Normale Supérieure, Université P. et M. Curie, Université Paris-Diderot,
[+] Have contributed equally
[*] present address: Janelia Farm Research Campus, Howard Hughes Medical Institute, Ashburn, VA 20147, USA

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 *Imogene* an ensemble of statistical tools that we have developed and implemented in a publicly available software. 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. It can then predict on a genome-wide scale other CRMs with a regulatory potential similar to the training set. *Imogene* bypasses the need of large data sets for statistical analyses by making central use of the information provided by the sequenced genomes of multiple species, based on the developed statistical tools and explicit models for transcription factor binding site evolution. We test *Imogene* on characterized tissue-specific mouse developmental CRMs. Its ability to identify CRMs with the same specificity based on its *de novo* created motifs equals that of the previously evaluated best motif-blind methods. We further show, both in flies and in mammals, that *Imogene de novo* generated motifs are sufficient to discriminate CRMs related to different developmental programs. Notably, *Imogene* performs as well in this discrimination task purely based on sequence data, than a previously reported learning algorithm based on ChIP data for multiple transcription factors. We thus expect *Imogene* to be a useful tool to decipher transcriptional gene regulation in higher eukaryotes.

INTRODUCTION

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 contain multiple binding sites for transcription factors (TFs) that can be either identical (homotypic clustering) or different (heterotypic clustering). Homotypic clustering can provide cooperative TF binding 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 to 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 forms a useful complement to genome-wide binding studies. Classic case-by-case studies or large scale binding data (11), as previously

[*]To whom correspondence should be addressed. Email: vincent.hakim@ens.fr

described, often provide a moderate number (about ten to a few tens) of CRMs, active in the co-regulation of a subset of genes, in specific biological systems or in the formation of different organs at various stages of development. 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 (12). It should also help one to determine other CRMs providing specific expression patterns, a difficult task at present given the absence of close association (13) between CRMs and their target genes in higher eukaryotes. These labor-intensive experimental tasks could be eased by computational work. To this end, we have previously developed (14) 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 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 (14). 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 try and extend the method developed in ref. (14) to the case of mammalian CRMs. The task of determining cis-regulatory elements is even more difficult for mammalian genomes than for *Drosophila* ones since they are an order of magnitude richer in intergenic sequences (15, 16). To tackle this challenge, we have developed *Imogene*, a computer algorithm and software that we present here and characterize. *Imogene* predicts:

1. cis-regulatory sequences (of about 10 nt long) within a moderate set size of 10-30 CRMs, responsible for specific gene co-regulation, as well as a set of Probability Weight Matrices (PWM) or motifs (17, 18) characterizing the DNA-binding specificity of the associated putative factors.

2. novel CRMs at the genomic scale with the same expression pattern as the starting set of CRMs, based on the set of build PWMs.

Numerous algorithms have already been developed to try and map cis- underlying transcriptional regulation (see e.g. (3, 17, 19, 20, 21) for recent reviews). *Imogene* differs from previous methods in several respects. *Imogene* aim is most similar to the goal of the algorithms analyzed in (22). These algorithms have been specially designed to decode cis-regulatory regulation in a small set of CRMs, contrary to other algorithms which are aimed at the analysis of large datasets such as whole ChIP-seq peak regions (23). Both work *de novo* instead of using already characterized binding motifs (24, 25, 26, 27, 28, 29, 30, 31, 32). Faced to the weak statistical discriminative power offered by the starting set of characterized CRMs, the best algorithms of ref. (22) try and distinguish regulatory sequences by their entire content in short nucleotide sequences as also proposed in other works (33, 34, 35, 36, 37). On the contrary, *Imogene* insists on building cis-regulatory motifs since those are important for experimental work. It instead relies on conservation and the comparison of multiple sequenced genomes.

In the following, the general methodology of *Imogene* is first presented. Then, *Imogene* performance on mammalian CRMs is assessed. Imogene is trained on CRMs pertaining to neural tube and limb developmental programs during embryogenesis. It is shown to successfully classify other CRMs in the same class based on its *de novo* created list of best motifs which contained both new and already known motifs. We then consider the distinct but related task of discriminating CRMs with different specificities, rather than discriminating a set of specific CRMs from background intergenic sequences. *Imogene* is shown to accurately discriminate mammalian neural tube from limb CRMs on the basis of very few learned motifs. 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. (38). The CRM classification solely based on *Imogene de novo* generated motifs is found to be of similar quality as the results obtained in ref. (38) based on ChIP binding data for multiple transcription factors at several developmental time points. Finally, the developed publicly available *Imogene* interface is presented.

MATERIALS AND METHODS

Genome alignments

The alignments were downloaded from `ftp://ftp.ensembl.org/pub/release-63/emf/ensembl-compara/epo_12_eutherian` for mammals and from `http://www.biostat.wisc.edu/~cdewey/fly_CAF1/data` for Drosophilae. 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). These scripts are available in the *Imogene* distribution.

Annotations

The CDS coordinates were downloaded from `ftp://ftp.ensembl.org/pub/release-64/gtf/mus_musculus` for mammals (mm9 coordinates) and from `ftp://ftp.flybase.net/releases/FB2011_06/dmel_r5.38/gff` for Drosophilae (release 5 coordinates). In the case of mammals, the TSS coordinates were obtained separately from `http://hgdownload.cse.ucsc.edu/goldenPath/mm9/database`. Mammalian alignments were already masked for repeat sequences. Drosophilae alignments were masked using the coordinates indicated in the *gff* file.

Phylogenetic trees

The phylogenetic trees used within *Imogene* are displayed in Figure 2. For drosophilae, the distances are taken from Heger and Pontig (39). For mammals, they are obtained from the Ensembl (16) website (`www.ensembl.org`).

Background sequences

Imogene computes the statistical over-representation of the predicted motifs by comparing them to 20 Mb of background intergenic DNA (10^4 regions of 2 Kb). 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 used mammalian training sets (limb, neural tube) were obtained from `http://enhancer.lbl.gov`, based on the work of (11, 40). 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 to 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 ±20nt, our flexibility criterion 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 limb CRMs (minimal length 789 bp, maximal length 9052 bp, average length 3045 bp) and 29 neural tube CRMS (minimal length 585 bp, maximal length 3045 bp, average length 2419 bp).

The Drosophilae training sets were obtained from (38). Coordinate files are given as Supplementary Material.

Main program

The main program is written in C++ and adapted from the program used in a previous study (14). It is distributed under the GNU GPL license and available as a git repository at `http://github.com/hrouault/Imogene`. The user manual is available at `http://hrouault.github.io/Imogene/`. The program can be accessed through a web interface at `http://mobyle.pasteur.fr/cgi-bin/portal.py#forms::imogene`.

Binding site scores

A given motif is represented by a PWM with frequency $w_{i,b}$ for the base b at position i. The index i runs from 1 to l_m the size of the motif which is a parameter in the program which takes the same value for all considered motifs. The binding score of a sequence s_i for such a motif, is defined through the corresponding PWM as:

$$S = \sum_i \log_2 \left(\frac{w_{i,s_i}}{\pi_b} \right) \qquad \textbf{(1)}$$

where π_b is the mean frequency of the base b within intergenic regions ($\pi_{\mathrm{A,T}} = 0.30$ and $\pi_{\mathrm{C,G}} = 0.20$) as measured on the "background sequences" (see methods *Background sequences* subsection for their detailed description). A sequence is considered as a binding site in the reference species (*D. melanogaster* or *Mus musculus*) when its score S is larger than the score threshold (S_s or S_g) defined by the user of *Imogene*.

Conservation requirements for binding sites

Imogene iteratively builds PWM from binding sites that have conserved instances in different species. The conservation requirement 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) considered conserved by *Imogene*.

Evolutionary models

Imogene can use two different evolutionary models, which vary in complexity and computational time, to compare orthologous binding sites. In both models, the bases within a site evolve independently from each other.

Felsenstein model. The simplest models of TFBS nucleotides evolution are copied on models of neutral evolution for genomic nucleotides. This procedure has been proposed by Sinha *et al* (24, 41) with the Felsenstein model of neutral evolution (42). In this TFBS evolution model, the transition probability from nucleotide b to b' at position i in two sites at evolutionary distance d is defined as:

$$p^i_{b \to b'} = q\ \delta_{b,b'} + (1-q)\ w_{i,b'} \qquad \textbf{(2)}$$

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 π_b:

$$q = \exp\left(-\frac{d}{1/2 + 4\pi_{\mathrm{A,T}}\pi_{\mathrm{C,G}}} \right) \qquad \textbf{(3)}$$

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

Halpern-Bruno model. The Halpern-Bruno model (HB) (43) differs in two ways from the simplest *Felsenstein* model. It uses the more complex Hasegawa, Kishino and Yano model (HKY) (44) 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. Substitutions 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 κ which is the ratio of the transition rate over the transversion rate. It is measured to be $\kappa = 2$ in flies and $\kappa = 3.7$ in mammals (45).

Within a TFBS, the HB model extends the HKY model to take into account an additional purifying selection on the nucleotide identities (43). It is formulated by the following transition probabilities:

$$p_{b\to b'} = \exp(t\mathbf{H})_{b,b'} \tag{4}$$

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

$$H_{b,b'} = \begin{cases} \pi_b \, h_{b'\to b} & \text{if } b \neq b' \\ -\sum_{b'\neq b} H_{b,b'} & \text{if } b = b' \end{cases} \tag{5}$$

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

$$t = \frac{d}{1/2 + 4\kappa \pi_{\mathrm{A,T}} \pi_{\mathrm{C,G}}} \tag{6}$$

Finally, the transition rates are defined by:

$$h_{b\to b'} = \frac{w_{\mathrm{b'}}}{\pi_{\mathrm{b'}}} \frac{\log\left(\frac{\pi_{\mathrm{b}} w_{\mathrm{b'}}}{\pi_{\mathrm{b'}} w_{\mathrm{b}}}\right)}{w_{\mathrm{b'}}/\pi_{\mathrm{b'}} - w_{\mathrm{b}}/\pi_{\mathrm{b}}} \alpha_{b\to b'} \tag{7}$$

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

Inference

The algorithm infers in a Bayesian way the PWM $\mathbf{w}$ frequencies $w_{i,b}$ based on observations of binding sites, as previously described in (14). In a Bayesian framework, the posterior probability $\mathcal{P}(\mathbf{w}|\{\mathcal{A}\})$ that the matrix $\mathbf{w}$ represents the PWM binding to a set of aligned nucleotides $\{\mathcal{A}\}$ is proportional to the product of:

-the *a priori* probability $\mathcal{P}_{ap}(\mathbf{w})$, the '*prior*', that the matrix $\mathbf{w}$ represents a PWM

-the probability $\mathcal{P}(\{\mathcal{A}\}|\mathbf{w})$ of observing the set of aligned nucleotides given that they belong to binding sites for the PWM $\mathbf{w}$.

The prior is taken to be a Dirichlet distribution with parameters α_β at each PWM position

$$\mathcal{P}_{ap}(w_i) \propto \prod_{b\in\{A,T,C,G\}} w_{i,b}^{\alpha_b - 1} \tag{8}$$

The nucleotides at different positions are assumed to be independent and the prior for the full site is taken to be the product of the $\mathcal{P}_{ap}(w_i)$ over the different positions. The parameters α_b are taken to be equal for Watson-Crick complementary nucleotides since a sequence and its reverse complement are not distinguished in the description of binding sites (i.e. we assume that binding is not biased toward a particular DNA strand). The two values of α_b are fully determined by assuming that i) TFBS *a priori* have the same nucleotide frequencies as the background and ii) that a PWM mean *a priori* information content is equal to the input threshold score S_g.

The probability $\mathcal{P}(\{\mathcal{A}\}|\mathbf{w})$ of observing the set of aligned nucleotides given the PWM $\mathbf{w}$ is computed in a standard way (42) by recursion for a given PWM $\mathbf{w}$ and a given evolutionary model.

The posterior distribution of the nucleotides frequencies at position i is thus obtained under the form,

$$\mathcal{P}(w_i|\{\mathcal{A}\}) \propto \prod_{a\in\{\mathcal{A}\}} \mathcal{P}(a|w_{i,b}) \prod_{b\in\{A,T,C,G\}} w_{i,b}^{\alpha_b - 1} \tag{9}$$

where we omit the normalization factor.

In the idealistic case where the aligned nucleotides represent independent observations (infinitely distant species), the likelihood reduces to a multinomial distribution and the posterior is given by:

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

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 mean posterior estimate is expressed as:

$$\tilde{w}_{i,b} = \frac{N_b + \alpha_b}{\sum_b N_b + \alpha_b} \tag{11}$$

Eq. (**11**) coincides with the maximum likelihood estimate for a Dirichlet 'prior' with parameters $\alpha_b + 1$.

In the case of a non-trivial evolutionary tree (like those of Fig. 2), the orthologous sites are correlated by their evolution from common ancestors. The probability $\mathcal{P}(a|w_{i,b})$ is a polynomial function of the $w_{i,b}$'s. However, it generally lacks a simple analytical expression and the mean posterior estimate should be computed numerically.

Mean Posterior Estimation

The mean posterior estimate was initially computed using a Markov chain Monte Carlo (MCMC) procedure (46). This turned out to be a time-consuming step in the algorithm. To speed it up, we observed, as noted above, that the mean posterior estimate for a prior with Dirichlet parameters α_b coincided with the maximum likelihood estimate for a prior Dirichlet parameters α_b+1 in the case of uncorrelated observations as well as fully correlated ones (i.e. reducing to a single observation). We thus reasoned that maximization with this modified Dirichlet prior could give a quick satisfying approximation for the phylogenetic trees of Fig. 2, which was checked on different examples. This procedure is thus adopted in the present version of *Imogene* and for the results shown here. The posterior distribution obtained with the modified prior is maximized by using the Nelder-Mead simplex algorithm, as implemented in the GNU GSL. The initial value for the estimation is taken to be the mean estimator in the independent species regime given in Eq. (**10**). This allows one to start close to the quadratic region and ensures fast convergence.

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 consider in Figure S5 a dinucleotidic genome with bases X and Y and a simple phylogenetic tree 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 $d\to\infty$, the two instances of nucleotide X form two independent observations. Using the previous result (Eq. (**11**)) 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:

$$w_Y=\frac{N_Y+\alpha}{N_Y+N_X+2\alpha} \qquad (\mathbf{12})$$

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

$$w_Y=\frac{\alpha}{1+2\alpha} \qquad (\mathbf{13})$$

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

$$w_Y=\frac{\alpha}{2+2\alpha} \qquad (\mathbf{14})$$

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:

$$\begin{aligned}\mathcal{P}(\mathcal{A}|w)&=w_X\left[q+(1-q)w_X\right]^2+w_Y(1-q)^2w_X^2\\&=q^2w_X+(1-q^2)w_X^2\end{aligned} \qquad (\mathbf{15})$$

where $\mathcal{A}$ stands for the simple alignment considered in Figure 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 ($d\to\infty$, $q=0$) 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 posterior, $w_{Y,ma}$ analytic estimates for finite d read

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

Note that for the maximum posterior estimate, $w_{Y,ma}$, the prior exponent $\alpha+1$ has been used instead of α 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 be comparable to the difference between the evolutionary models.

Filtering of motifs coming from simple repeats

Imogene pre-processes the training set by masking repeated sequences with repeat masker (47) but this is not sufficient to eliminate the production of motifs corresponding to repeated sequences. These motifs have a non-poissonian distribution of binding sites on intergenic sequences: one binding site has a high probability to be followed by another one after a multiple of the repeat period. This anomalous distribution of binding sites biases motif ranking and diminishes the algorithm CRM predicting power (14). Motifs corresponding to repeated sequences are thus filtered out using the non-poissonian characteristics of their binding site distribution. The binding sites of each motif m are determined on the above-described set of $N_{bg}=10^4$ background sequences of length $L=2\times10^3$ nt. For a Poisson distribution, one would expect the number $N_m^{(p)}(j)$ of intergenic sequences containing j binding sites to be

$$N_m^{(p)}(j)=N_{bg}\frac{(\lambda_m^{(bg)}L)^j}{j!}\exp(-\lambda_m^{(bg)}L) \qquad (\mathbf{16})$$

where $\lambda_m^{(bg)}$ is the computed density of binding sites of the motif m in the set of background sequences. The deviation from this theoretical Poisson distribution is quantitatively assessed by computing the χ^2-like value,

$$\chi^2(m)=\sum_j \frac{[N_m(j)-N_m^{(p)}(j)]^2}{N_m^{(p)}(j)}\Theta(N_m(j)) \qquad \textbf{(17)}$$

where Θ is the Heaviside function ($\Theta(x)=0$ for $x<0,\Theta(x)=1$ for $x>0$) which restricts the sum to non-zero values of $N_m(j)$. Only the 75 % motifs with the lowest $\chi^2(m)$-value are retained for subsequent computations.

Distance between motifs

The similarity between two motifs is quantitatively assessed based on the overlap between the sets of their binding sites. The 'strict proximity' between motifs represented by two PWMs $\mathbf{w_1}$ and $\mathbf{w_2}$, is defined by

$$\mathrm{Prox}(\mathbf{w}^{(1)},\mathbf{w}^{(2)})=$$
$$2\frac{\mathrm{Prob}\left\{\left[S(\mathbf{s},\mathbf{w}^{(1)})>S_{th}\right] \text{ and } \left[S\left(\mathbf{s},\mathbf{w}^{(2)}\right)>S_{th}\right]\right\}}{\mathrm{Prob}\{S(\mathbf{s},\mathbf{w}^{(1)})>S_{th}\}+\mathrm{Prob}\{S(\mathbf{s},\mathbf{w}^{(2)})>S_{th}\}} \qquad \textbf{(18)}$$

where $\mathrm{Prob}\{S(\mathbf{s},\mathbf{w})>S_{th}\}$ is the probability that a sequence $\mathbf{s}$ drawn at random with the background frequencies π_b has a binding score $S(\mathbf{s},\mathbf{w})$ (Eq. (**1**)) above the threshold S_{th} for the frequency matrix $\mathbf{w}$. The strict proximity is computed analytically as explained in (14), where it was defined. To take into account potential shifs in the motifs or in their orientation, $\mathrm{Prox}(\mathbf{w}^{(1)},\mathbf{w}^{(2)})$is computed for all possible alignments of the two matrices (with a maximum shift of $l_m/2$ where l_m is motif size) in the two possible orientations. When shifted matrices are compared, they are completed by additional columns with the background frequencies (i. e. with no specifity). The proximity between two motifs is obtained simply by taking the maximum over the obtained strict proximities. It goes from 1 for two identical motifs to zero for motifs that do not share any binding site above the threshold. *Imogene* distance between two motifs is defined as minus the logarithm of their proximity.

Ranking motifs

The previous filtering step provides for each considered motif m, the density $\lambda_m^{(bg)}$ of its binding sites on the background sequences and ensures that these sites are approximately distributed in a poissonian way. The deviation from this baseline distribution on the CRM of the training set (t.s.) is used to score each motif. This is quantified by the poissonian log-likelihood of the training set

$$Pl(m)=\sum_{t\in\{\text{t.s.}\}}\log\left(\frac{\left(L_t\lambda_m^{(bg)}\right)^{k_t}\exp(-L_t\lambda_m^{(bg)})}{k_t!}\right) \qquad \textbf{(19)}$$

where k_t is the number of instances of m on the training set sequence t of length L_t. Larger deviations from the baseline poissonian distribution are supposed to reflect motif specificity for the training set and correspond to more negative/better scores.

Scoring intergenic sequences

Given a list of motifs m_i, a CRM E is scored as follows:

$$S(E)=\sum_i n(E,m_i)\log(\lambda_i^t/\lambda_i^b) \qquad \textbf{(20)}$$

where $n(E,m_i)$ is the number of binding sites for the motif m_i on E and λ_i^t, λ_i^b are 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, do not increase artificially the CRM weight by predicting the same binding sequences twice.

Selection of optimal intergenic sequences

When ranking genome-wide intergenic sequences, with a list of N motifs, the best intergenic sequence at a given position is determined as follows. The list of motifs is used to scan the genome for conserved binding sites above a given threshold. Binding sites are then 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:

$$E_i=\frac{X_1+X_N+l_m-1}{2} \qquad \textbf{(21)}$$

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

Mammalian predictions

Learning sets, test sets 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. It prevents 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 let 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:

$$S(E) = \sum_{j=1}^{N_c} (2\delta_{j,i} - 1) S_N^{\mathcal{C}_j}(E) \qquad \textbf{(22)}$$

where $S_N^{\mathcal{C}_j}(E)$ is the CRM score for the N top motifs of class $\mathcal{C}_j$ as defined below in the 'Main program' description, 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 attributing their mean rank to equally scored CRMs. The rank of the test CRM is used rather than its raw score to avoid potential bias stemming from score normalization. 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 attained) 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 (48). 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 (14). 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.

Statistical analysis

All statistical analyses were performed using R (49).

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 cumulated 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 aligned orthologous sequences in other related sequenced genomes (see *genome alignments* in *Methods*), as shown in Figure 2 (for the mouse, the 11 other aligned mammalian sequenced genomes with high coverage presently available on the Ensembl project (16), the 11 other *Drosophilae* sequenced genomes (15) 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 ℓ (10 nt is the default value) by scanning the training set, in an iterative manner (step II in Figure 1). Each sequence of ℓ 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 ℓ 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 further considered (see *Conservation requirements for binding sites* in *Methods*). A shift in alignment between a binding site on the reference species and its orthologues in other species is allowed for the correction of 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 (24, 25, 41) 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 model”*, is a simple and computationally fast model proposed in (41). 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 (42) 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 model”* (43) uses an evolutionary model that is more complex than the Felsenstein model but that is also more clearly grounded on theoretical population genetics ideas. It has previously been used for TFBS evolution in (25). It allows for the inclusion of different mutational probabilities 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 (50). 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 each ℓ nucleotide long sequences in the training set. In a series of final steps (see *Methods* for a mathematically detailed description), 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). *Imogene* pre-processes the training set by masking repeated sequences with repeat masker (47) but, as noted in ref. (14), this is not sufficient to eliminate some PWMs corresponding to repeated sequences from the produced list of PWMs. These PWMs have statistically anomalous distributions of binding sites that bias their subsequent ranking. Therefore, in a filtering first step, PWMs corresponding to repeated sequences are discarded on the basis of their anomalous distribution of their binding sites in the background set (see *Filtering of motifs coming from simple repeats* in *Methods*). 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 the 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 (see *Ranking motifs* in *Methods*). In a final step, PWMs in the ranked list are compared (see *Distance between motifs* in *Methods*)) 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 scans 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 (see *Scoring intergenic sequences* in *Methods*) 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 mammalian 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, 40). 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 (see *Training sets* in *Methods*).

A cross-validation scheme was then used to measure *Imogene* predictability power (see *Methods* for details). In brief, for each developmental program, the CRMs of the training set were divided into a learning set composed of 15 CRMs chosen at random, and a test set composed of the other CRMs used as True Positives.

The learning 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 learning set (see Figure 3 and Figure S1).

The test CRMs of the training set were then ranked, using motifs generated on the learning set, against a 'background test set', a set of ~ 60 regions of 1Kb taken from the flanking sequences of the initial set of CRMs (see *Methods*).

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 (51). 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. For the neural tube CRMs, 30% of the test set CRMs are retrieved at 1% FPR whereas an even larger proportion of 40% is obtained for the limb CRMs. The *Halpern-Bruno* and the *Felsenstein* models are seen in Figure 3 to yield very similar results in both cases. It should be noted that the test really provides only a lower estimate of *Imogene* success rate. Sequences of the background test set counted as 'False Positive' could, in reality, be *bona fide* positive CRMs.

The performance of *Imogene* is found to be comparable to the best motif-blind methods (22). Using a cross-validation protocol similar to the one used here, in which the CRMs to be tested were compared to flanking sequences, the 'HexMCD' was found to be top-scoring method for the set of limb CRMs. It recovered 60% of the training set for a $5-10\%$ FPR. For neural tube CRMs, the two best methods, 'PAC-rc' and 'D2z-cond-weights' recovered 80% and 74% of the test set for a $5-10\%$ FPR (see Figure S5 in ref. (22)).

One interesting feature of *Imogene* lies in its production of specific motifs. In our 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. Oct1/Pou2f3 family 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 the ability of *Imogene* 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 (38), as detailed in the next section. It consists in learning features that distinguishes 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. (38), 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 simple linear classifier based on a weighted score giving positive (resp. negative) contributions to positive (resp. negative) motifs (see *Methods*). Finally, the test CRM is ranked among all CRMs by the build classifier and this rank is registered. A successful classification would rank positive CRMs on top of the list and attribute worse ranks to negative CRMs. Therefore, after processing all CRMs, the list of ranks for the positive and negative CRMs is represented as a ROC curve indicating the True Positives Rate and False Positive Rate for increasing rank. This serves to optimize the different 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 ≤ 0.2.

The results are shown in Figure 4. We focus on the results obtained with the *Halpern-Bruno* evolutionary model. Results (motifs, thresholds) are 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 (52).

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. (38). 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. (38) 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. (38). We then proceeded as in the previous case of mammalian CRMs.

Imogene results are shown together with the machine learning results of ref (38) 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 specificity range (FPR $\leq 5\%$) used for CRM prediction in (38). 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 (38), 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. (38).

In summary, our analysis indicates that *Imogene* not only determines *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 ensemble of developed statistical tools and the allied computer codes are freely available at `http://github.com/hrouault/Imogene`. In addition, they can be used through a user-friendly web interface (`http://mobyle.pasteur.fr/cgi-bin/portal.py#forms::imogene`) that provides motif and CRM predictions for the community. This interface is powered by the Pasteur Institute Internet server through the mobyle framework (53). 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 their score ($Pl(m)$ in *Methods*).
- 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 described in *Methods*)

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 other training sets.

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 (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 can be 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). The default values have been chosen for computational efficiency but changes can 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.

DISCUSSION AND CONCLUSION

We have presented *Imogene*, a set of statistical tools and a computer software able to predict *de novo* relevant motifs in a moderate size set of functionally related 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. It leads it to achieve a performance comparable to the best motif-blind algorithms (22).

Phylogenetic conservation between multiple sequenced genomes has previously been shown to provide useful information on cis-regulatory motifs (54, 55, 56) but cannot *per se* address the question of specific spatio-temporal expression. The necessary information is provided to *Imogene* by the training set of CRMs with well-characterized expression. *Imogene* aim is to extract it optimally by making full use of several sequenced genomes, instead of focusing on a single genome (26) analysis, simply comparing the reference genome with another one (57, 58, 59) or simply adding orthologous sequences (60). Similarly to the *Monkey* algorithm of ref. (25), *Imogene* uses a model for the evolution of motif binding sites, to properly weigh this additional information. The two algorithms are however complementary since *Imogene* creates *de novo* motifs from the training set while *Monkey* tests already well-characterized binding motifs.

The algorithm which lies at *Imogene* core was previously applied to gene co-regulation in *Drosophila* (14). 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 (61). 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 the method could be successfully extended to decipher cis-regulatory information in the notoriously more difficult case of mammalian gene expression. We have here provided bioinformatics evidence that our developed algorithm indeed provides 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* from the analysis of 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. At the same time a large number of PWMs remain to be found. This is even more the case for CRMs. Therefore, the use of *Imogene* with its *de novo* motif building ability and allied CRM identification, should provide helpful service to the community.

We have further shown that *Imogene* can discriminate between classes of CRMs. In this task, it should usefully complement 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 motifs and help to identify new involved DNA-binding cofactors and their binding sites. We thus believe that *Imogene* is a useful addition to existing algorithms and softwares (26). We hope that it will serve as a helpful and timely tool in the difficult deciphering of gene regulation in higher eukaryotes.

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 and S. Eddy for his critical reading of the manuscript. 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).

Conflict of interest statement. None declared.

REFERENCES

1. Davidson, E. H. (2006) The regulatory genome: gene regulatory networks in development and evolution, Academic, Burlington, MA.
2. Dorer, D. E. and Nettelbeck, D. M. (Jul, 2009) Targeting cancer by transcriptional control in cancer gene therapy and viral oncolysis. *Adv Drug Deliv Rev,* **61**(7-8), 554–71.
3. Hardison, R. C. and Taylor, J. (2012) Genomic approaches towards finding cis-regulatory modules in animals. *Nat. Rev. Genet.,* **13**(7), 469–483.
4. Lelli, K. M., Slattery, M., and Mann, R. S. (2012) Disentangling the many layers of eukaryotic transcriptional regulation. *Annu. Rev. Genet.,* **46**, 43–68.
5. Levine, M. (Sep, 2010) Transcriptional enhancers in animal development and evolution. *Curr. Biol.,* **20**(17), R754–763.
6. Arnosti, D. N. and Kulkarni, M. M. (Apr, 2005) Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards?. *J Cell Biochem,* **94**(5), 890–8.
7. Swanson, C. I., Evans, N. C., and Barolo, S. (Mar, 2010) Structural rules and complex regulatory circuitry constrain expression of a Notch- and EGFR-regulated eye enhancer. *Dev Cell,* **18**(3), 359–70.
8. Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B. (Jun, 2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science,* **316**(5830), 1497–502.
9. Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (May, 2007) High-resolution profiling of histone methylations in the human genome. *Cell,* **129**(4), 823–37.
10. Mikkelsen, T. S., Ku, M., Jaffe, D. B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.-K., Koche, R. P., Lee, W., Mendenhall, E., O'Donovan, A., Presser, A., Russ, C., Xie, X., Meissner, A., Wernig, M., Jaenisch, R., Nusbaum, C., Lander, E. S., and Bernstein, B. E. (Aug, 2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature,* **448**(7153), 553–60.
11. Visel, A., Blow, M. J., Li, Z., Zhang, T., Akiyama, J. A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., Afzal, V., Ren, B., Rubin, E. M., and Pennacchio, L. A. (Feb, 2009) ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature,* **457**(7231), 854–8.
12. Arnosti, D. N. and Kulkarni, M. M. (Apr, 2005) Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards?. *J. Cell. Biochem.,* **94**(5), 890–898.
13. Amano, T., Sagai, T., Tanabe, H., Mizushina, Y., Nakazawa, H., and Shiroishi, T. (Jan, 2009) Chromosomal dynamics at the Shh locus: limb bud-specific differential regulation of competence and active transcription. *Dev. Cell,* **16**(1), 47–57.
14. Rouault, H., Mazouni, K., Couturier, L., Hakim, V., and Schweisguth, F. (Aug, 2010) Genome-wide identification of cis-regulatory motifs and modules underlying gene coregulation using statistics and phylogeny. *Proc Natl Acad Sci U S A,* **107**(33), 14615–20.
15. Clark, A., Eisen, M., Smith, D., Bergman, C., Oliver, B., Markow, T., Kaufman, T., Kellis, M., Gelbart, W., Iyer, V., et al. (2007) Evolution of genes and genomes on the Drosophila phylogeny. *Nature,* **450**(7167), 203–218.
16. Flicek, P., Amode, M. R., 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., Kähäri, A. K., 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, H. S., Ritchie, G. R. S., Ruffier, M., Schuster, M., Sobral, D., Tang, Y. A., Taylor, K., Trevanion, S., Vandrovcova, J., White, S., Wilson, M., Wilder, S. P., Aken, B. L., Birney, E., Cunningham, F., Dunham, I., Durbin, R., Fernández-Suarez, X. M., Harrow, J., Herrero, J., Hubbard, T. J. P., Parker, A., Proctor, G., Spudich, G., Vogel, J., Yates, A., Zadissa, A., and Searle, S. M. J. (Jan, 2012) Ensembl 2012. *Nucleic Acids Res,* **40**, D84–90.
17. Wasserman, W. W. and Sandelin, A. (Apr, 2004) Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.,* **5**(4), 276–287.
18. Stormo, G. and Fields, D. (1998) Specificity, free energy and information content in protein–DNA interactions. *Trends in biochemical sciences,* **23**(3), 109–113.
19. Su, J., Teichmann, S. A., and Down, T. A. (2010) Assessing computational methods of cis-regulatory module prediction. *PLoS Comput. Biol.,* **6**(12), e1001020.
20. Elnitski, L., Jin, V. X., Farnham, P. J., and Jones, S. J. (Dec, 2006) Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques. *Genome Res.,* **16**(12), 1455–1464.
21. Aerts, S. (2012) Computational strategies for the genome-wide identification of cis-regulatory elements and transcriptional targets. *Curr. Top. Dev. Biol.,* **98**, 121–145.
22. Kantorovitz, M., Kazemian, M., Kinston, S., Miranda-Saavedra, D., Zhu, Q., Robinson, G., Göttgens, B., Halfon, M., and Sinha, S. (2009) Motif-Blind, Genome-Wide Discovery of cis-Regulatory Modules in Drosophila and Mouse. *Developmental Cell,* **17**(4), 568–579.
23. Machanick, P. and Bailey, T. L. (Jun, 2011) MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics,* **27**(12), 1696–1697.
24. Siddharthan, R., Siggia, E., and van Nimwegen, E. (2005) PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput Biol,* **1**(7), e67.
25. Moses, A. M., Chiang, D. Y., Pollard, D. A., Iyer, V. N., and Eisen, M. B. (2004) MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. *Genome Biol,* **5**(12), R98.
26. Herrmann, C., Van de Sande, B., Potier, D., and Aerts, S. (Aug, 2012) i-cisTarget: an integrative genomics method for the prediction of regulatory features and cis-regulatory modules. *Nucleic Acids Res,* **40**(15), e114.
27. Berman, B., Nibu, Y., Pfeiffer, B., Tomancak, P., Celniker, S., Levine, M., Rubin, G., and Eisen, M. (2002) 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,* **99**(2), 757.
28. Halfon, M., Grad, Y., Church, G., and Michelson, A. (2002) Computation-based discovery of related transcriptional regulatory modules and motifs using an experimentally validated combinatorial model. *Genome research,* **12**(7), 1019.
29. Rebeiz, M., Reeves, N., and Posakony, J. (2002) 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,* **99**(15), 9888.
30. Schroeder, M., Pearce, M., Fak, J., Fan, H., Unnerstall, U., Emberly, E., Rajewsky, N., Siggia, E., and Gaul, U. (2004) Transcriptional control in the segmentation gene network of Drosophila. *PLoS biology,* **2**, 1396–1410.
31. Hallikas, O., Palin, K., Sinjushina, N., Rautiainen, R., Partanen, J., Ukkonen, E., and Taipale, J. (2006) Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. *Cell,* **124**(1), 47–59.
32. Pierstorff, N., Bergman, C., and Wiehe, T. (2006) Identifying cis-regulatory modules by combining comparative and compositional analysis of DNA. *Bioinformatics,* **22**(23), 2858.
33. Nazina, A. and Papatsenko, D. (2003) Statistical extraction of Drosophila cis-regulatory modules using exhaustive assessment of local word frequency. *BMC bioinformatics,* **4**(1), 65.
34. Abnizova, I., te Boekhorst, R., Walter, K., and Gilks, W. (2005) Some statistical properties of regulatory DNA sequences, and their use in predicting regulatory regions in the Drosophila genome: the fluffy-tail test. *BMC bioinformatics,* **6**(1), 109.
35. Chan, B. and Kibler, D. (2005) Using hexamers to predict cis-regulatory motifs in Drosophila. *BMC bioinformatics,* **6**(1), 262.
36. Leung, G., Eisen, M., and Provart, N. (2009) Identifying Cis-Regulatory Sequences by Word Profile Similarity. *PLoS ONE,* **4**(9), e6901.
37. Brody, T., Yavatkar, A. S., Kuzin, A., Kundu, M., Tyson, L. J., Ross, J., Lin, T.-Y., Lee, C.-H., Awasaki, T., Lee, T., and Odenwald, W. F. (Jan, 2012) Use of a Drosophila genome-wide conserved sequence database to identify functionally related cis-regulatory enhancers. *Dev Dyn,* **241**(1), 169–89.
38. Zinzen, R. P., Girardot, C., Gagneur, J., Braun, M., and Furlong, E. E. (Nov, 2009) Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature,* **462**, 65–70.
39. Heger, A. and Ponting, C. P. (Nov, 2007) Variable strength of translational selection among 12 Drosophila species. *Genetics,* **177**, 1337–1348.
40. May, D., Blow, M. J., Kaplan, T., McCulley, D. J., Jensen, B. C., Akiyama, J. A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Afzal, V., Simpson, P. C., Rubin, E. M., Black, B. L., Bristow, J., Pennacchio, L. A., and Visel, A. (2011) Large-scale discovery of enhancers from human heart tissue. *Nat. Genet.,* **44**, 89–93.
41. Sinha, S., van Nimwegen, E., and Siggia, E. D. (2003) A probabilistic

method to detect regulatory modules. *Bioinformatics,* **19 Suppl 1**, i292–301.

42. Felsenstein, J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol,* **17**(6), 368–76.
43. Halpern, A. L. and Bruno, W. J. (Jul, 1998) Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Mol Biol Evol,* **15**(7), 910–7.
44. Hasegawa, M., Kishino, H., and Yano, T. (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol,* **22**(2), 160–74.
45. Seplyarskiy, V. B., Kharchenko, P., Kondrashov, A. S., and Bazykin, G. A. (Aug, 2012) Heterogeneity of the transition/transversion ratio in Drosophila and Hominidae genomes. *Mol. Biol. Evol.,* **29**(8), 1943–1955.
46. Bishop, C. et al. (2006) Pattern recognition and machine learning, Springer New York, .
47. Bao, Z. and Eddy, S. (2002) Automated de novo identification of repeat sequence families in sequenced genomes. *Genome research,* **12**(8), 1269–1276.
48. Matys, V., Kel-Margoulis, O. V., 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, A. E., and Wingender, E. (Jan, 2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res,* **34**(Database issue), D108–10.
49. R Development Core Team R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing Vienna, Austria (2011) ISBN 3-900051-07-0.
50. Kimura, M. (Jun, 1962) On the probability of fixation of mutant genes in a population. *Genetics,* **47**, 713–9.
51. Hastie, T., Tibshirani, R., Friedman, J., et al. (2001) The elements of statistical learning: data mining, inference, and prediction, Springer New York, .
52. Kiyota, T., Kato, A., Altmann, C. R., and Kato, Y. (Mar, 2008) The POU homeobox protein Oct-1 regulates radial glia formation downstream of Notch signaling. *Dev Biol,* **315**(2), 579–92.
53. Neron, B., Menager, H., Maufrais, C., Joly, N., Maupetit, J., Letort, S., Carrere, S., Tuffery, P., and Letondal, C. (Nov, 2009) Mobyle: a new full web bioinformatics framework. *Bioinformatics,* **25**(22), 3005–3011.
54. Xie, X., Lu, J., Kulbokas, E., Golub, T., Mootha, V., Lindblad-Toh, K., Lander, E., and Kellis, M. (2005) Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature,* **434**(7031), 338–345.
55. Ettwiller, L., Paten, B., Souren, M., Loosli, F., Wittbrodt, J., and Birney, E. (2005) The discovery, positioning and verification of a set of transcription-associated motifs in vertebrates. *Genome Biology,* **6**(12), R104.
56. Stark, A., Lin, M., Kheradpour, P., Pedersen, J., Parts, L., Carlson, J., Crosby, M., Rasmussen, M., Roy, S., Deoras, A., et al. (2007) Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures.. *Nature,* **450**(7167), 219.
57. Wang, T. and Stormo, G. (2003) Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics,* **19**(18), 2369.
58. Grad, Y., Roth, F., Halfon, M., and Church, G. (2004) Prediction of similarly acting cis-regulatory modules by subsequence profiling and comparative genomics in Drosophila melanogaster and D. pseudoobscura. *Bioinformatics,* **20**(16), 2738.
59. Zhao, G., Schriefer, L., and Stormo, G. (2007) Identification of muscle-specific regulatory modules in Caenorhabditis elegans. *Genome research,* **17**(3), 348.
60. Busser, B. W., Taher, L., Kim, Y., Tansey, T., Bloom, M. J., Ovcharenko, I., and Michelson, A. M. (Mar, 2012) A machine learning approach for identifying novel cell type-specific transcriptional regulators of myogenesis. *PLoS Genet,* **8**(3), e1002531.
61. Menoret, D., Santolini, M., Fernandes, I., Spokony, R., Zanet, J., Gonzalez, I., Latapie, Y., Ferrer, P., Rouault., H., White, K., Besse, P., Hakim, V., Aerts, S., Payre, F., and Plaza, S. (2012) Genome-wide analyses of Shavenbaby target genes reveals distinct features of enhancer organization. *(Submitted),*.

FIGURES

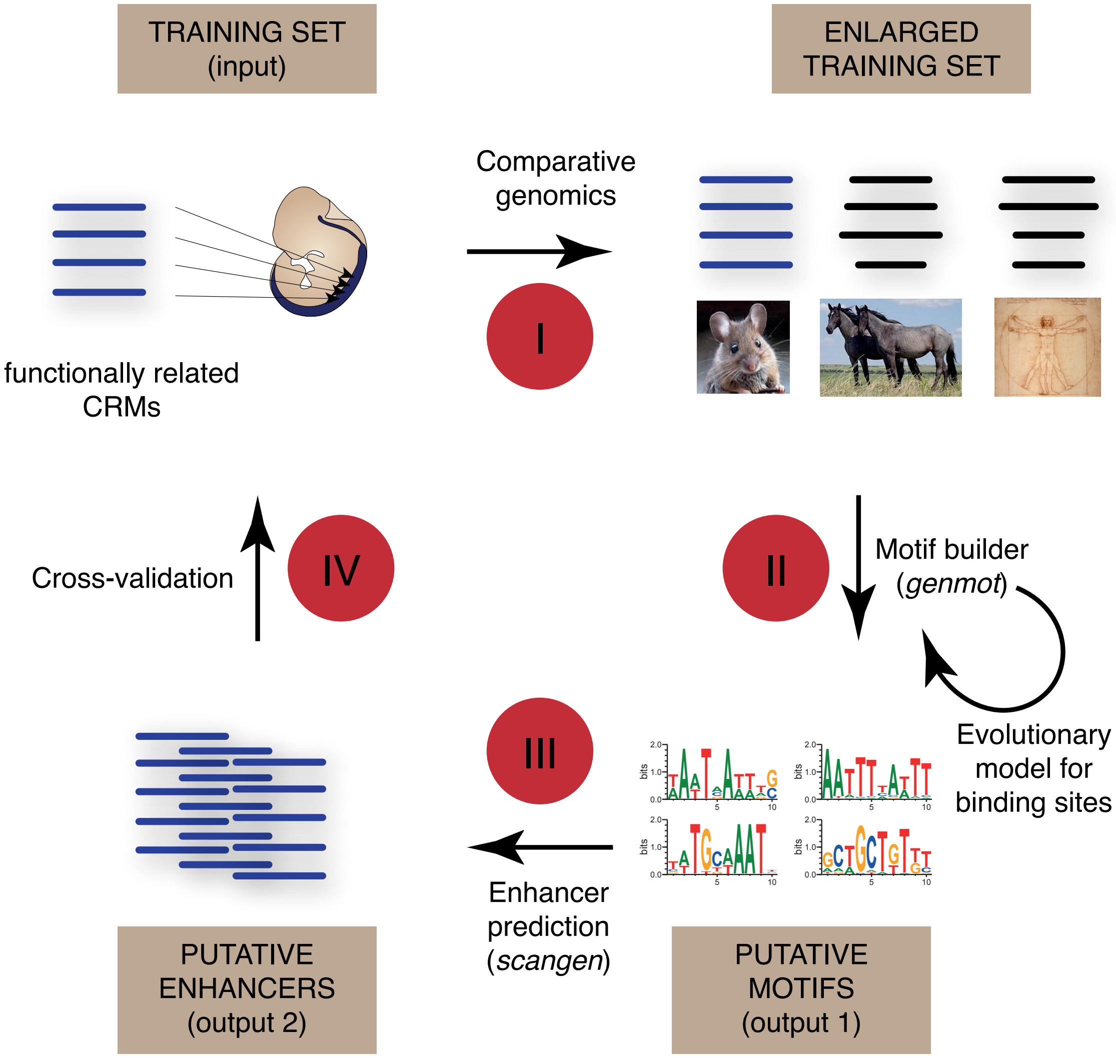


Figure 1. ***Imogene* workflow.** The algorithm takes as input a list of functionally related CRMs. Homologous sequences from closely related species are automatically retrieved (I) and scanned in order to generate a list of putative transcription factor motifs (II). These motifs fuel the last step consisting 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 of the whole procedure (IV).

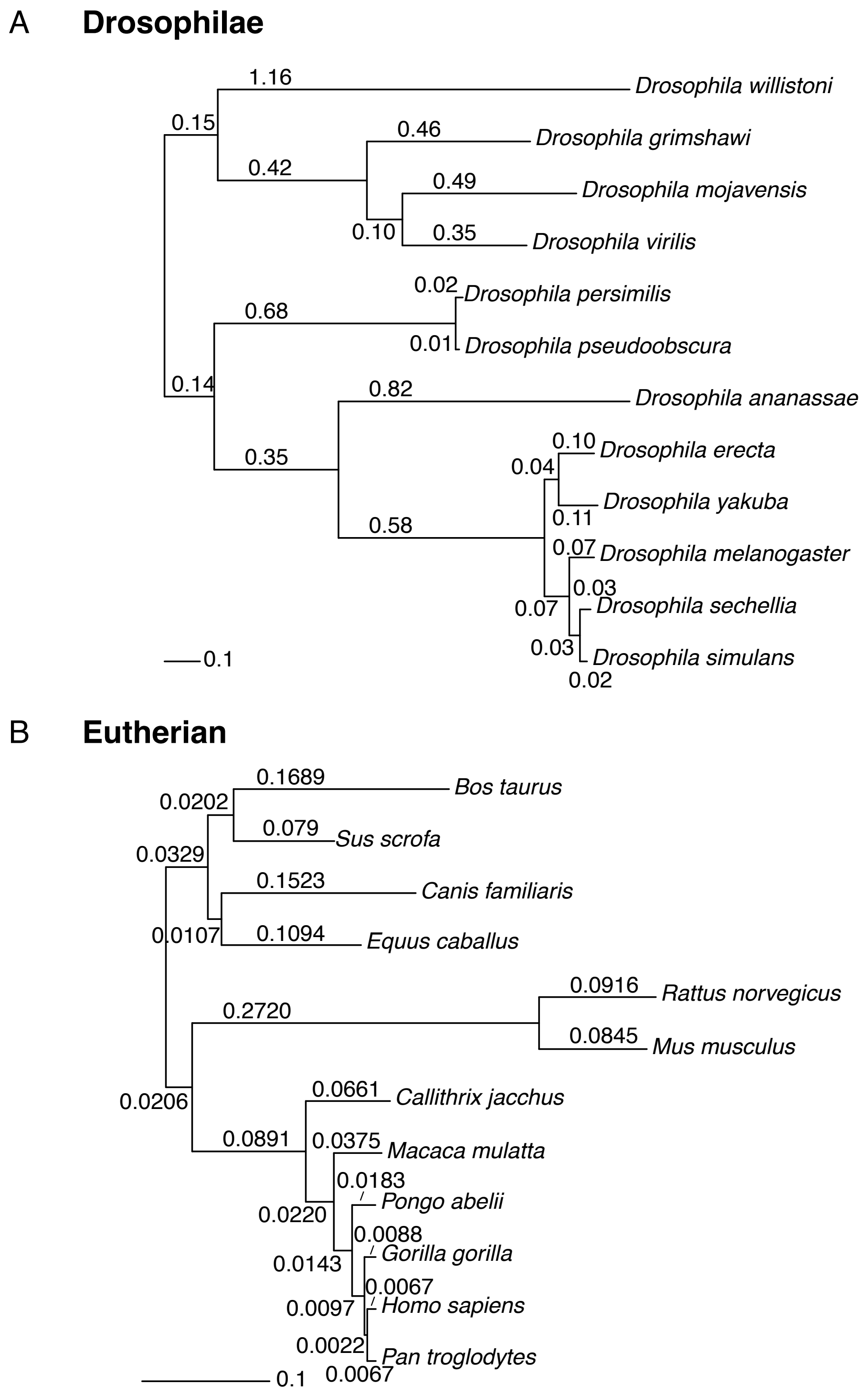


Figure 2. **Phylogenetic trees and phylogenetic distances used by *Imogene*.** The branch lengths represent the evolutionary distances d used by the evolutionary models at the motif construction stage.

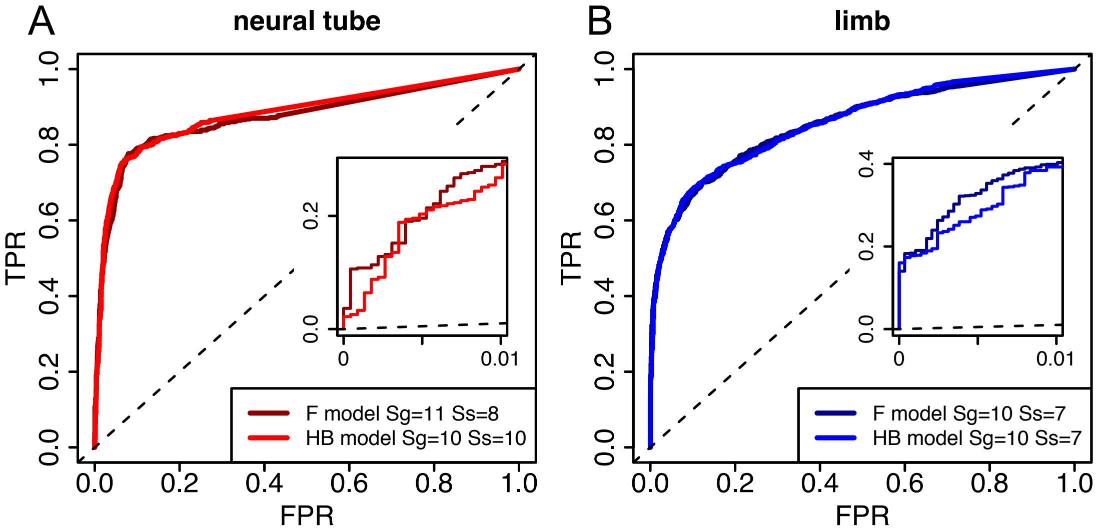


Figure 3. Analysis of well characterized developmental processes. We tested the algorithm on mammal CRMs driving expression at E11.5 in neural tube (A) and limb (B). For each class, CRMs were divided into a training set and a test set. Motifs were learned on the training set and used to score CRMs from the test set along with background regions consisting of the CRMs 1kb flanking sequences (see *Methods*). The displayed ROC curves show the proportion of test set CRMs recovered above a given score (True Positive Rate denoted by TPR) *vs.* the proportion of recovered background sequence at the same score for the Felsenstein (F) and Halpern-Bruno (HB) models. The shown ROC plots are the results of 40 trials. The FPR $\leq 1\%$ region of each curve is replotted in the insets for better visibility. For each test set and each evolutionary model, the thresholds S_g and S_s used for motifs generation and sequences scanning are given in the figures. Black dashed lines show random discrimination.

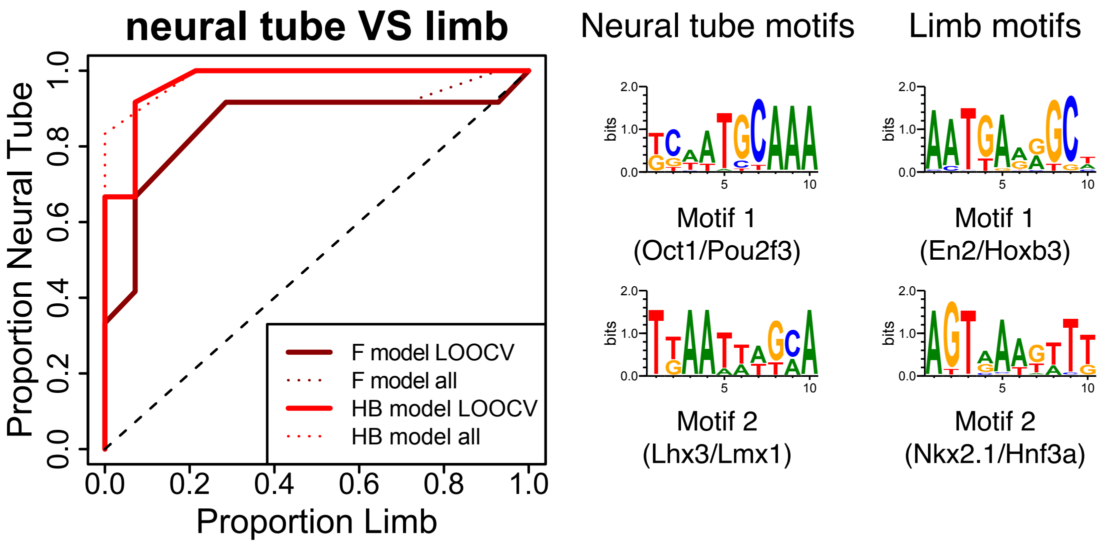


Figure 4. **Pattern recognition (mammals).** ROC plots showing the discrimination between limb and neural CRMs using a simple linear classifier. Neural and limb classes are compared to each other. Thick lines correspond to a leave-one-out cross-validation (LOOCV) scheme with a score function based on the *de novo* generated motifs from *Imogene*. The results obtained with the two evolutionary models are shown (Felsenstein model (F) solid dark red line, with threshold parameters $S_g = 11$, $S_s = 9$, and Halpern-Bruno (HB) model, solid light red line, with threshold parameters $S_g = 11$, $S_s = 8$). The analogous discrimination curves based on learning motifs on the whole training set (with the same threshold parameters) are shown for comparison (colored dashed lines). With this latter procedure, the discrimination is improved but still comparable to that computed by the LOOV, indicative of no strong overfitting of the training set. The corresponding discriminative motifs are shown for the whole training set learning with HB model (similar motifs are obtained with the F model). Black dashed line show random discrimination.

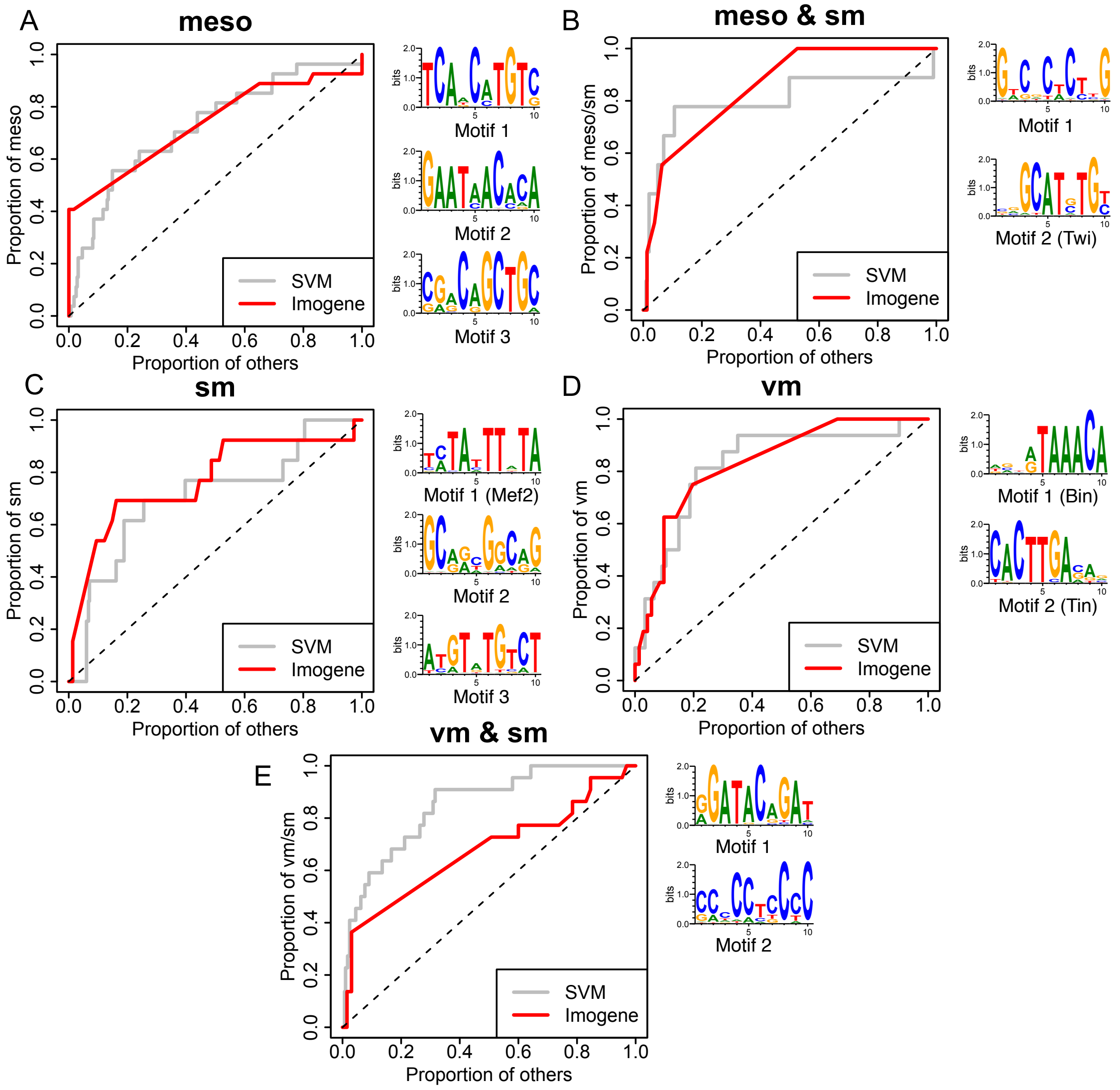
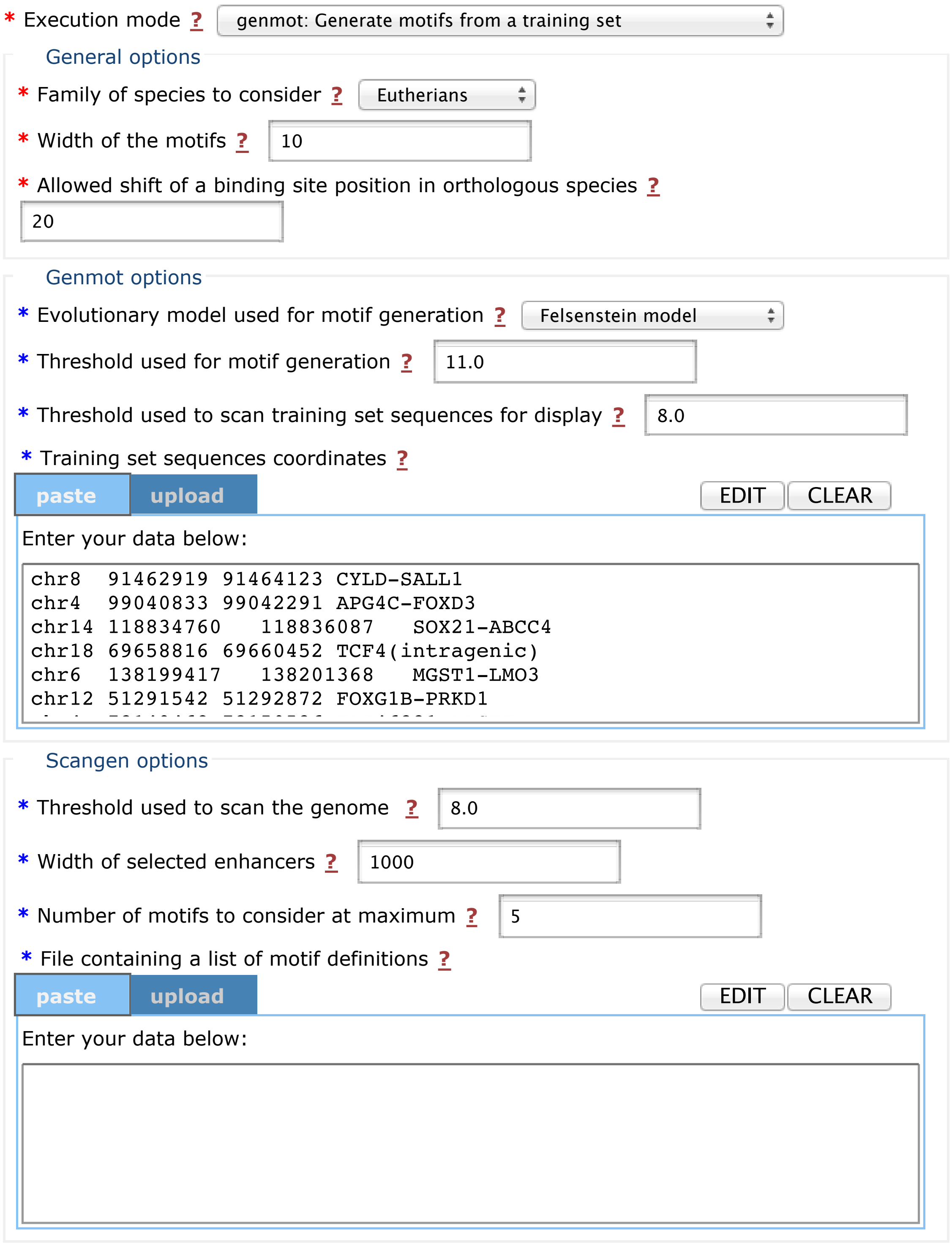


Figure 5. **Pattern recognition (Drosophilae).** Recognition of classes of CRMs expressed in 5 tissue types: mesoderm (meso), somatic muscle (sm), visceral muscle (vm) , mesoderm and somatic muscle (meso & sm) and visceral and somatic muscle (vm & sm). ROC plots are obtained using a leave-one-out cross-validation scheme. Two classifiers are compared: a Support Vector Machine using 15 ChIPseq peak heights (grey, replotted using the data and the program provided in ref. (38)), and *Imogene* using the *de novo* generated motifs with Felsenstein evolutionary model (red) and a simple linear classifier (see *Methods*). 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 ?

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

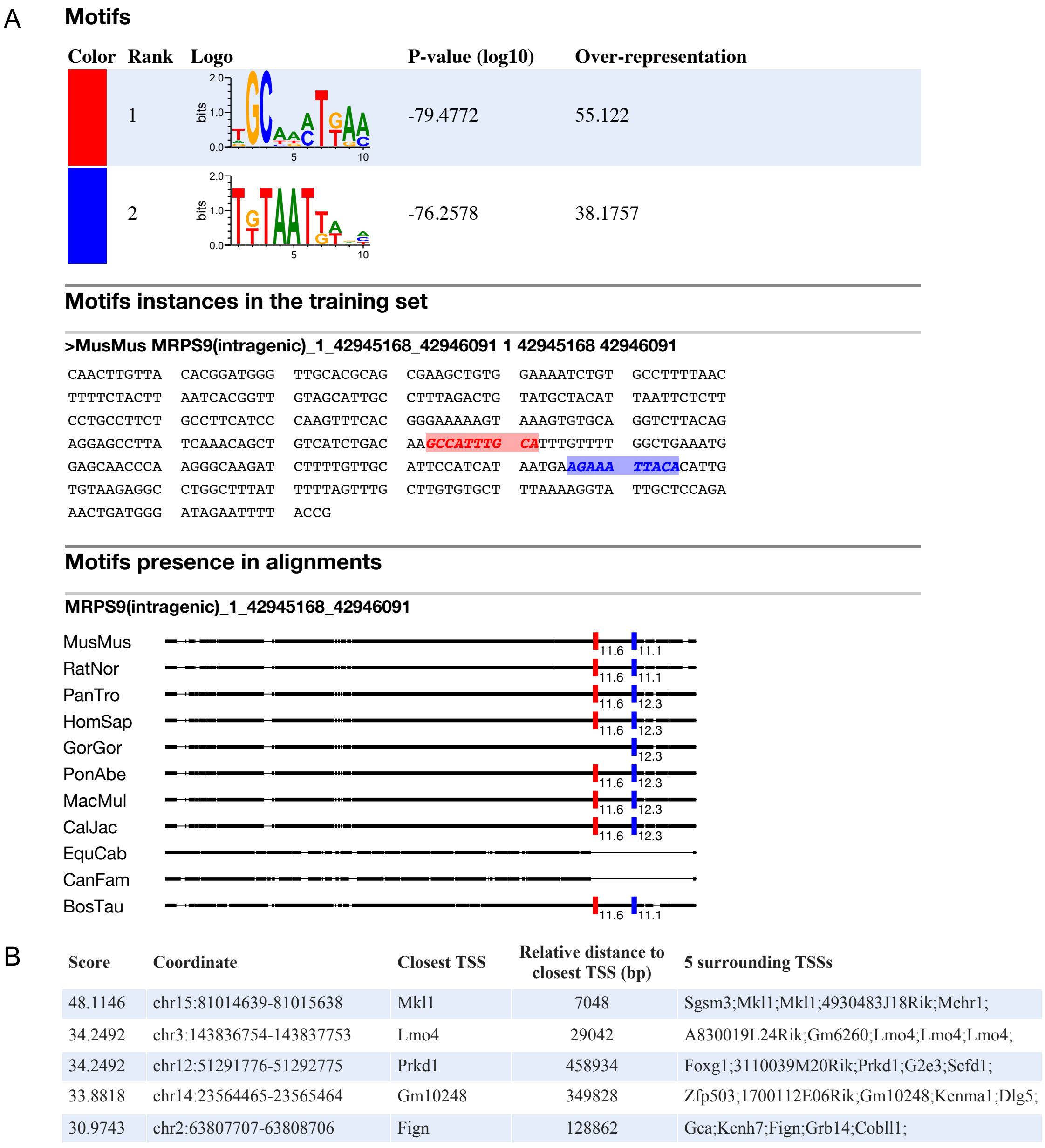
* Width of selected enhancers ? 1000

* Number of motifs to consider at maximum ? 5

* File containing a list of motif definitions ?

Enter your data below:

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



A

Motifs

Color	Rank	Logo	P-value (log10)	Over-representation
🟥	1		-79.4772	55.122
🟦	2		-76.2578	38.1757

Motifs instances in the training set

>MusMus MRPS9(intragenic)_1_42945168_42946091 1 42945168 42946091

```
CAACTTGTTA   CACGGATGGG   TTGCACGCAG   CGAAGCTGTG   GAAAATCTGT   GCCTTTTAAC
TTTTCTACTT   AATCACGGTT   GTAGCATTGC   CTTTAGACTG   TATGCTACAT   TAATTCTCTT
CCTGCCTTCT   GCCTTCATCC   CAAGTTTCAC   GGGAAAAAGT   AAAGTGTGCA   GGTCTTACAG
AGGAGCCTTA   TCAAACAGCT   GTCATCTGAC   AAGCCATTTG   CATTTGTTTT   GGCTGAAATG
GAGCAACCCA   AGGGCAAGAT   CTTTTGTTGC   ATTCCATCAT   AATGAAGAAA   TTACACATTG
TGTAAGAGGC   CTGGCTTTAT   TTTTAGTTTG   CTTGTGTGCT   TTAAAAGGTA   TTGCTCCAGA
AACTGATGGG   ATAGAATTTT   ACCG
```

Motifs presence in alignments

MRPS9(intragenic)_1_42945168_42946091

B

Score	Coordinate	Closest TSS	Relative distance to closest TSS (bp)	5 surrounding TSSs
48.1146	chr15:81014639-81015638	Mkl1	7048	Sgsm3;Mkl1;Mkl1;4930483J18Rik;Mchr1;
34.2492	chr3:143836754-143837753	Lmo4	29042	A830019L24Rik;Gm6260;Lmo4;Lmo4;Lmo4;
34.2492	chr12:51291776-51292775	Prkd1	458934	Foxg1;3110039M20Rik;Prkd1;G2e3;Scfd1;
33.8818	chr14:23564465-23565464	Gm10248	349828	Zfp503;1700112E06Rik;Gm10248;Kcnma1;Dlg5;
30.9743	chr2:63807707-63808706	Fign	128862	Gca;Kcnh7;Fign;Grb14;Cobll1;

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).

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

Hervé Rouault, Marc Santolini , François Schweisguth, Vincent Hakim

July 15, 2013

Supplementary Figures

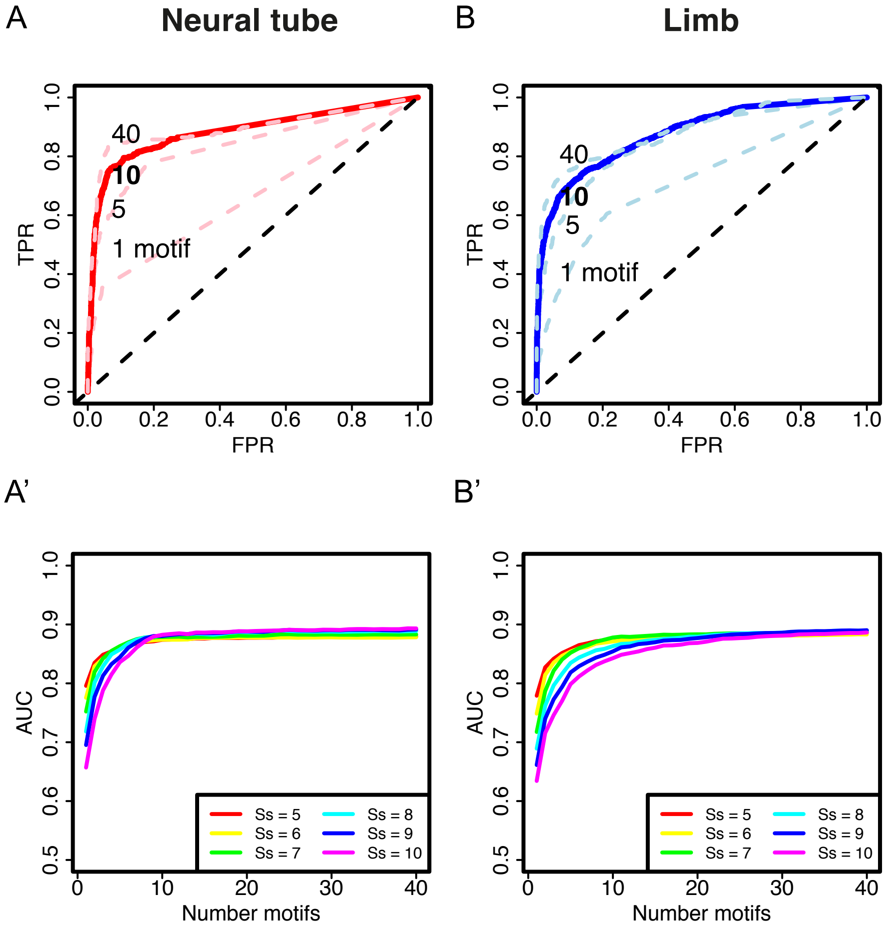


Figure S1. Dependence of the predictions on the number of scoring motifs ROC plots obtained at optimal scanning threshold using the Halpern-Bruno evolutionary model are shown for the neural tube (A) and limb (B) cases. Different curves are shown corresponding to sequences scored with different number 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 reach convergence. The optimal S_s was chosen as the one maximizing the AUC for 10 motifs.

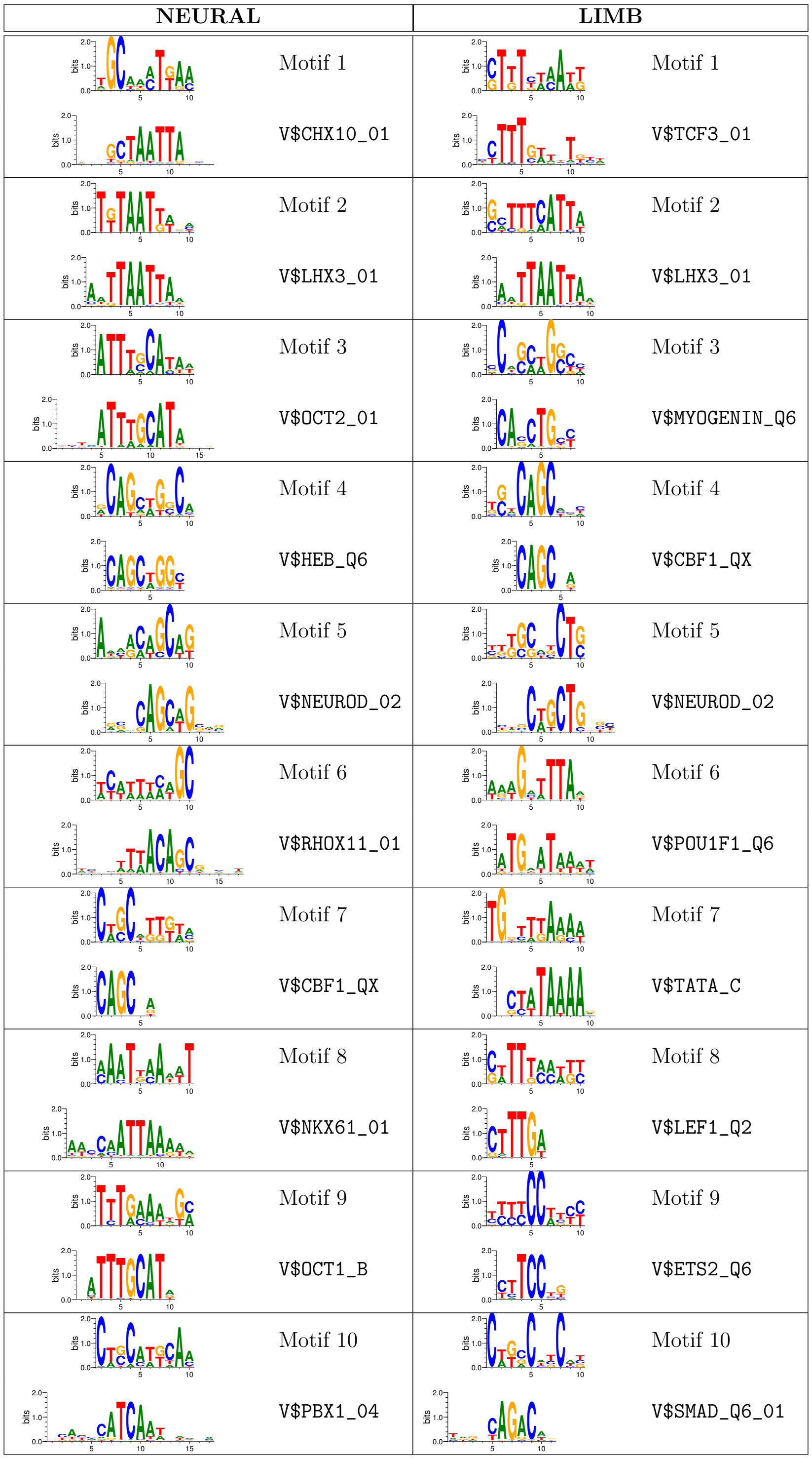
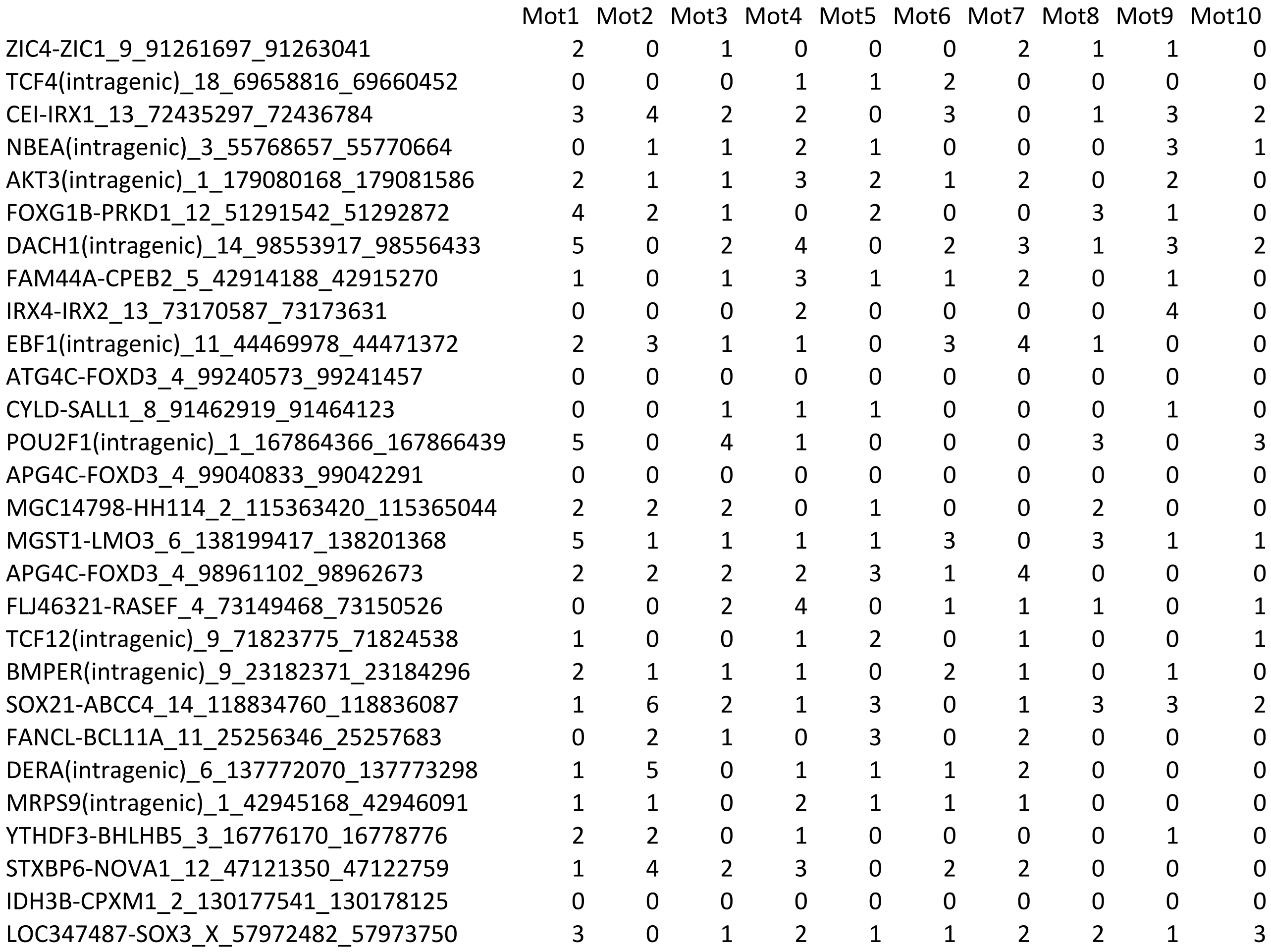
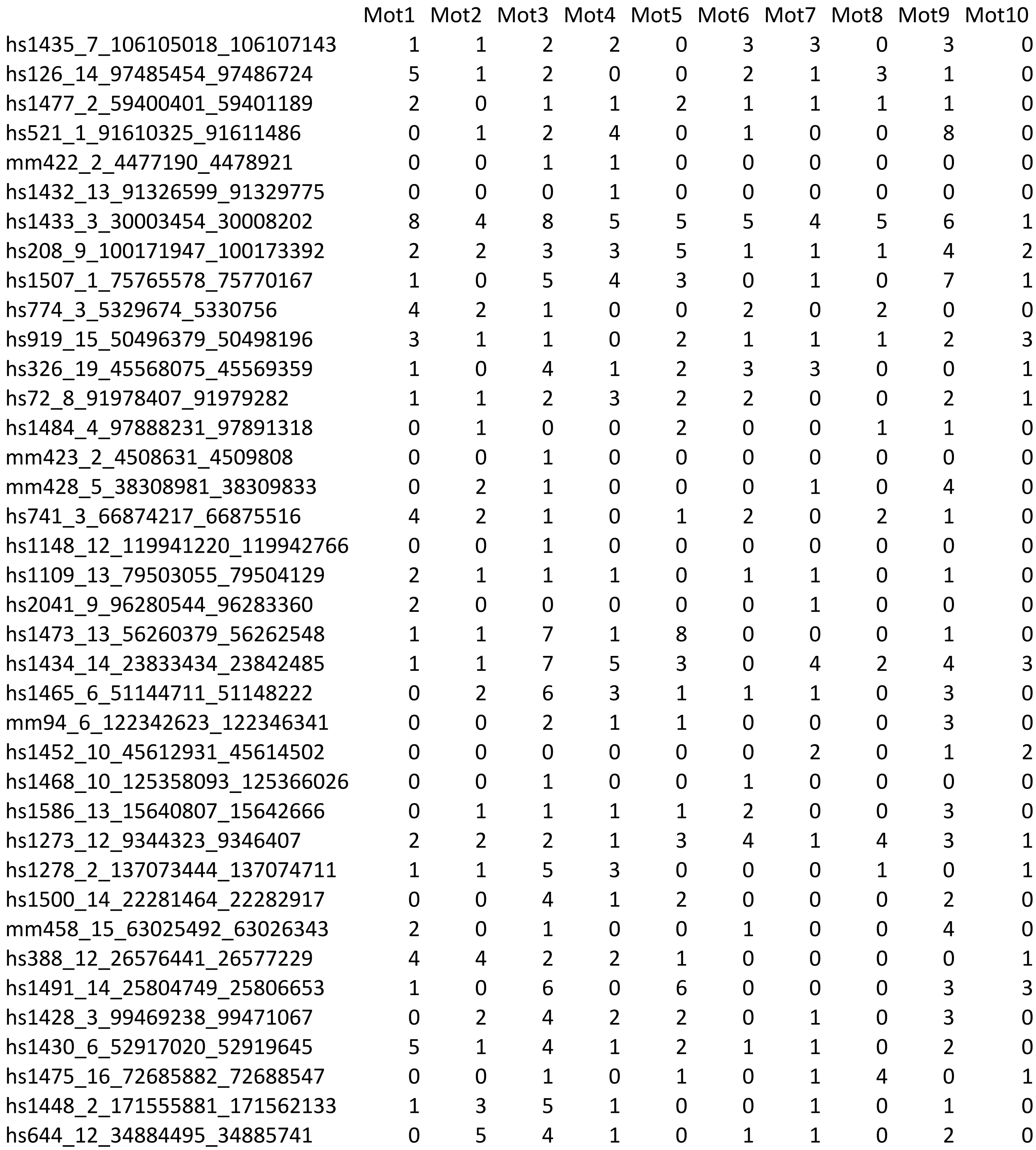


Figure S2. Motifs learnt on the full training sets. The 10 best ranking motifs generated on the CRMs training sets are shown together with the closest Transfac motifs (see *Distance between motifs* in *Methods* for details of motif distance computation).



	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-FOXD3_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-FOXD3_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-FOXD3_4_98961102_98962673	2	2	2	2	3	1	4	0	0	0
FLJ46321-RASEF_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.

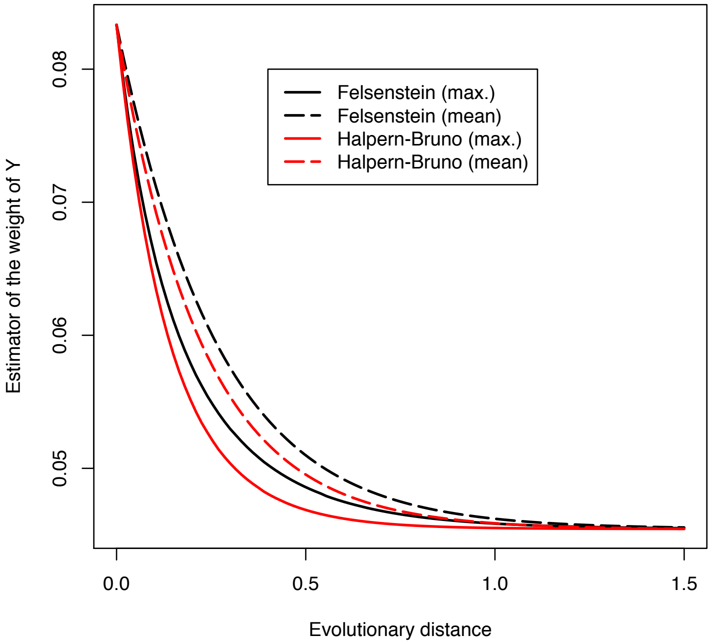


Figure S5. Simple example of motif inference with Felsenstein and Halpern-Bruno evolutionary models The inference of an ancestral base is compared in the simple case of two species at a phylogenic distance d from their common ancestor, for a two nucleotide alphabet,X and Y. The mean and maximum likelihood 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.