

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,*}

¹Developmental and Stem Cell Biology Department, Institut Pasteur, F-75015 Paris, France, ²CNRS, URA2578, F-75015 Paris, France and ³Laboratoire de Physique Statistique, CNRS, École Normale Supérieure, Université P. et M. Curie, Université Paris-Diderot

Received November 26, 2013; Revised February 25, 2014; Accepted February 27, 2014

ABSTRACT

***Cis*-regulatory modules (CRMs) and motifs play a central role in tissue and condition-specific gene expression. Here we present *Imogene*, an ensemble of statistical tools that we have developed to facilitate their identification 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 datasets 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 is comparable to that of previously evaluated ‘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, purely relying on sequence data, *Imogene* performs as well in this discrimination task as a previously reported learning algorithm based on Chromatin Immunoprecipitation (ChIP) data for multiple transcription factors at multiple developmental stages.**

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) that 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 TFs (8), as well as on other factors that regulate transcription, e.g. by modifying chromatin structure (p300, CTCF, histone marks, etc.) (9,10). These protein binding data have 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

*To whom correspondence should be addressed. Telephone: 33144323768, Fax: 33144323433. Email: vincent.hakim@ens.fr

†Have contributed equally.

Present address:

Hervé Rouault, Janelia Farm Research Campus, Howard Hughes Medical Institute, Ashburn, VA 20147, USA.

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 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 TFs 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 *Drosophila melanogaster* gene expression program in sensory organ precursor cells (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 (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 (PWMs) 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 built PWMs.

Numerous algorithms have already been developed to try and map *cis*-underlying transcriptional regulation (see, e.g. (3,17,19–21) for recent reviews). *Imogene* differs from previ-

ous methods in several respects. *Imogene* aim is most similar to the goal of the ‘motif-blind’ algorithms analyzed in (22). These algorithms have been specially designed to characterize the specificity of a small set of CRMs, contrary to other algorithms that are aimed at the analysis of large datasets such as whole ChIP-seq peak regions (23). As *Imogene*, they work *de novo* instead of using already characterized binding motifs (24–32). Faced to the weak statistical discriminative power offered by the starting set of characterized CRMs, the algorithms of (22) try and distinguish regulatory sequences by their entire content in short nucleotide sequences as also proposed in other works (33–37). 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 that contained both new and already known motifs. *Imogene* is furthermore found to perform comparably to ‘motif-blind’ algorithms using the benchmark and methodology of (22). 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 (38). Remarkably, the CRM classification solely based on *Imogene de novo* generated motifs is found to be of similar quality as the results obtained in (38) based on ChIP binding data for multiple TFs 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

Drosophila (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. *Drosophila* alignments were masked using the coordinates indicated in the *gff* file.

Phylogenetic trees

The phylogenetic trees used within *Imogene* are displayed in Supplementary Figure S1. For *Drosophila*, the distances are taken from Heger and Ponting (39). For mammals, they are obtained from the Ensembl (15) 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 and neural tube) were obtained from <http://enhancer.lbl.gov> based on the work of Visel *et al.* (11). They were manually curated to produce a high-quality dataset, 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 ± 20 nt, 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 and average length 3045 bp) and 29 neural tube CRMs (minimal length 585 bp, maximal length 3045 bp and average length 2419 bp).

The *Drosophila* 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://moby.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 that 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:

$$S = \sum_i \log_2 \left(\frac{w_{i,s_i}}{\pi_{s_i}} \right), \quad (1)$$

where π_b is the mean frequency of the base b within intergenic regions ($\pi_{A,T} = 0.30$ and $\pi_{C,G} = 0.20$) as measured on the 'background sequences' (see the '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 three distant species, including the reference species. For mammals, the six groups of related species are composed of: *M. 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 five groups composed of: *D. melanogaster*, *D. sechellia*, *D. simulans*, *D. yakuba* and *D. erecta*; *D. ananassae*; *D. pseudoobscura* and *D. persimilis*; *D. willistoni*; *D. grimshawi*, *D. mojavensis* and *D. virilis*.

A site instance must be found in at least three of these groups (with an allowed shift of up to 20 nt with the reference species) to be 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 of each other.

Felsenstein model. The simplest models of TFBS nucleotide evolution are copied on models of neutral evolution for genomic nucleotides. This procedure has been proposed by Sinha *et al.* (29,41) 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

$$p_{b \rightarrow b'}^i = q \delta_{b,b'} + (1 - q) w_{i,b'}, \quad (2)$$

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 :

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

with $\pi_{A,T}$ (resp. $\pi_{C,G}$) 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:

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

where H is the rate matrix defined by

$$H_{b,b'} = \begin{cases} \pi_b h_{b' \rightarrow b} & \text{if } b \neq b' \\ -\sum_{b' \neq b} H_{b,b'} & \text{if } b = b'. \end{cases} \quad (5)$$

The evolutionary time t is expressed in terms of the evolutionary distance by

$$t = \frac{d}{1/2 + 4\kappa \pi_{A,T}\pi_{C,G}}. \quad (6)$$

Finally, the transition rates are defined by

$$h_{b \rightarrow b'} = \frac{w_{b'}}{\pi_{b'}} \frac{\log\left(\frac{\pi_b w_{b'}}{\pi_{b'} w_b}\right)}{w_{b'}/\pi_{b'} - w_b/\pi_b} \alpha_{b \rightarrow b'} \quad (7)$$

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

Inference

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

- the *a priori* probability $\mathcal{P}_{\text{ap}}(w)$, the ‘prior’, that the matrix w represents a PWM,
- the probability $\mathcal{P}(\{\mathcal{A}\}|w)$ of observing the set of aligned nucleotides given that they belong to binding sites for the PWM w .

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

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

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

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

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

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

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

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

$$\tilde{w}_{i,b} = \frac{N_b + \alpha_b}{\sum_b N_b + \alpha_b}. \quad (11)$$

Equation (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 Supplementary Figure S1, 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 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 parameter $\alpha_b + 1$ 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 Supplementary Figure S1, 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 Equation (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 Supplementary Figure S2 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 $d \rightarrow \infty$, the two instances of nucleotide X form two independent observations. Using the previous result (Equation (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

$$w_Y = \frac{N_Y + \alpha}{N_Y + N_X + 2\alpha}. \quad (12)$$

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

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

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

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

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

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

where \mathcal{A} stands for the simple alignment considered in Supplementary Figure S2 and we used $w_X = 1 - w_Y$. From

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

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

Note that for the maximum posterior estimate, $w_{Y,ma}$, the prior exponent $\alpha + 1$ has been used instead of α as explained above. So, the two estimates coincide at $q = 0$ and $q = 1$. Both estimates are plotted as function of the evolutionary distance d in Supplementary Figure S2 ($\alpha = 0.1$).

For the HB model, the analogous results have been computed numerically and are also shown for comparison in Supplementary Figure S2. The HB 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 $L = 2 \times 10^3$ nt. For a Poisson distribution, one would expect the number $N_m^{(p)}(j)$ of intergenic sequences containing j binding sites to be

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

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,

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

where Θ is the Heaviside function ($\Theta(x) = 0$ for $x < 0$, $\Theta(x) = 1$ for $x > 0$) that 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

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

where $\text{Prob}\{S(\mathbf{s}, \mathbf{w}) > S_{\text{th}}\}$ is the probability that a sequence \mathbf{s} drawn at random with the background frequencies π_b has a binding score $S(\mathbf{s}, \mathbf{w})$ (Equation (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 shifts in the motifs or in their orientation, $\text{Prox}(\mathbf{w}^{(1)}, \mathbf{w}^{(2)})$ 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 specificity). 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 0 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 $\lambda_m^{(\text{bg})}$ 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

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

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:

$$S(E) = \sum_i n(E, m_i) \log(\lambda_i^t / \lambda_i^b), \quad (20)$$

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:

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

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 (~20) defining the test set of ‘True Positives’. In addition, a set of background test regions was built using the 1 kb 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 toward 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. 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 motif 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 1 kb 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 Receiver Operating Characteristic (ROC) curve. The Area Under ROC Curve or AUC, a quantity that varies between 0 for absolute misclassification, 0.5 for random classification, and 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.

Comparison with Kantorovitz *et al.*

The data provided by Kantorovitz *et al.* (22) consisted of eight classes of human CRMs, each class containing between 10 and 67 CRMs, with an average of 30 CRMs per class. Coordinates of the human CRMs were obtained from <http://enhancer.lbl.gov> and were converted from the human hg19 to the mouse mm9 assembly using the UCSC LiftOver tool, yielding a loss of 5–10 unmapped CRMs per class. The extraction of the mammalian alignments with Imogene resulted in another loss of 1–2 CRMs per class for which no alignment could be retrieved. Overall, the ratio of finally retrieved over initially available number of CRMs per class was the following: dorsal root ganglion (7/10), eye (12/16), forebrain (50/67), heart (9/19), hindbrain rhombencephalon (25/32), limb (24/35), midbrain mesencephalon (36/42) and neural tube (20/23). The sensitivity of Imogene on these CRMs was then computed using a leave-one-out cross-validation scheme as described in (22). More precisely, we measured the ability of Imogene to retrieve a bona fide CRM of a given class (the test CRM) embedded in 10 kb of background intergenic DNA using the best motifs generated from the other CRMs of the class (the training set). The sequence containing the test CRM and the intergenic DNA was scanned using a window size equal to the average length of the CRMs in the given class. The window of highest score defined the predicted CRM. The prediction was considered valid when the test CRM and the predicted CRM overlapped over at least 100 bp. The sensitivity was finally computed as the proportion of test CRMs retrieved. This process was repeated using 10 different background regions. Imogene parameters were varied (evolutionary models: F and HB, S_g : 11 and 13, S_s : 5–13, number of motifs: 1–15), and the parameters giving the highest mean sensitivity over the 10 cross-validations were kept for each class. A p -value of 0.05 was computed empirically as described in (22). More precisely, the above process was repeated by drawing at random the best scoring fragment. The proportion of correctly predicted CRMs was computed. The process was repeated 100 000 times. The threshold sensitivity expected by chance with a p -value of 0.05 was obtained as the threshold proportion above which the 5000 higher computed proportions lied.

Two separate sets of tests were performed, either with or without making use of multiple sequenced genomes and conservation at the CRM retrieval step. Without conservation, tests were performed on human CRMs embedded in human intergenic background using the exact same composite sequences as in (22). For *Scangen* in its normal mode with conservation, the mouse is the reference species and alignments for the background sequences used in (22) needed to be retrieved. To do so, we followed the same protocol as for the training set sequences using the UCSC LiftOver tool for coordinate conversion and extracting alignments with Imogene. Because the length of the final background sequences could change during the process, we redefined 10 kb background regions around the centers of the sequences in the mouse genome. The training set alignments were then embedded in the center of the corresponding 10 kb background alignments and repeats were masked using repeatmasker (47). The resulting

sequences were finally used to conduct the leave-one-out cross-validation (LOOCV). We note that repeat masking did not significantly affect the results (data not shown).

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

$$S(E) = \sum_{j=1}^{N_c} (2\delta_{j,i} - 1) S_N^{C_j}(E), \quad (22)$$

where $S_N^{C_j}(E)$ 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 AUC for a false positive rate $FPR \leq 20\%$, which we notify as AUC20 and want to maximize.

In our case, we used a 2D parameter 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 HB 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 the TRANSFAC (48) and JASPAR (49) databases, as well as the list of motifs provided in (50) from HT-Selex experiments and the UniPROBE database (51) created from protein binding array data.

In order to avoid uninformative matches, we kept motifs that had an information content greater than 8 bits, a threshold approximately corresponding to four conserved nucleotides. This led us to keep 764 vertebrate motifs (934 total minus 170 below threshold) in the TRANSFAC database, 389 vertebrate motifs (476 total minus 87 below threshold) in the JASPAR database 575 HT-Selex motifs (580 total minus 5 below threshold) and 488 in the UniPROBE database (538 total minus 50 below threshold).

Each *de novo* motif was compared to all kept motifs in each database, using the PWM distance introduced in (14). During the comparison, motifs are shifted to find the best match with a minimal match of 5 nt. The shift is simply introduced by adding flanking nucleotides with background frequency on either side. The closest candidate was kept for identification.

RESULTS

Description of Imogene

Imogene has two modes that can be used in succession, as shown in Figure 1 and summarized here (see the ‘Materials and Methods’ section 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 kb. It should also be above a few kb to provide a sufficient amount of information (a training set of about 20 kb 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 the Materials and Methods section), as shown in Supplementary Figure S1 (for the mouse, the 11 other aligned mammalian sequenced genomes with high coverage presently available on the Ensembl project (15), the 11 other *Drosophilae* sequenced genomes (16) 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* builds PWMs of a 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

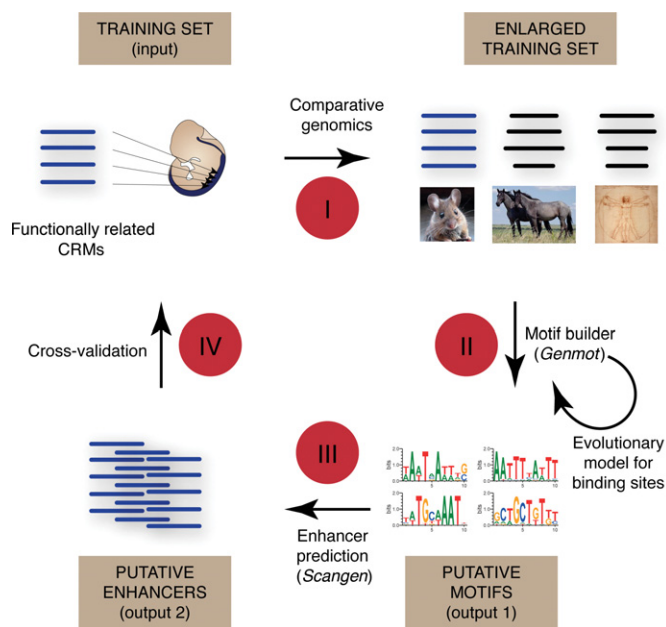


Figure 1. Imogene workflow. The algorithm takes as input a list of functionally related CRMs. Homologous sequences from closely related species are automatically retrieved (I) and scanned in order to generate a list of putative transcription factor motifs (II). These motifs fuel the last step consisting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

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, i.e. only sites that have orthologs in distant species are further considered (see ‘Conservation requirements for binding sites’ in the Materials and Methods section). A shift in alignment between a binding site on the reference species and its orthologs in other species is allowed for the correction of eventual alignment errors (20 nt is the shift default value). The ensemble of conserved binding sites and their orthologs 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 (28,29,41) is simply explained. A binding site in the reference genome and its orthologs 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 nucleotide neutral relative abundances replaced by PWM nucleotide frequencies. The 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 (43), ‘HB model’, uses an evolutionary model that is more complex than the Felsenstein model but is also more clearly grounded on theoretical population genetics ideas. It has previously been used for TFBS evolution in (28). 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 a homogeneous population (52). 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 the Materials and Methods section 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 the Materials and Methods section for a mathematically detailed description), this long list is pruned and ranked based on comparison of 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 (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 the Materials and Methods section). 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 the PWM score and the more meaningful the PWM is deemed (see ‘Ranking motifs’ in the Materials and Methods section). In a final step, PWMs in the ranked list are compared (see ‘Distance between motifs’ in the Materials and Methods section) and, among similar ones, only the highest scoring one is kept. Although the identity of the TFs 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 PWMs inferred in the *Genmot* mode. The algorithm scans the entire non-coding repeat-masked reference genome and finds 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 the Materials and Methods section). The closest the similarity in its motif content with the training set, the most likely an intergenic sequence 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). 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, two limb CRMs and four 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 the Materials and Methods section). Elimination of these uninformative sequences produced curated training sets of 29 neural and 39 limb CRMs (see ‘Training sets’ in the Materials and Methods section).

A cross-validation scheme was then used to measure *Imogene* predictability power (see ‘Materials and 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 motif generation by *Imogene* in its *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 2 and Supplementary Figure S3).

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 1 kb taken from the flanking sequences of the initial set of CRMs (see the Materials and Methods section).

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 (FPR) as this score decreased to produce a so-called ROC curve (53). The ROC curves corresponding to different parameter values were then compared using the AUC, a quantity that is maximal at best prediction. Supplemen-

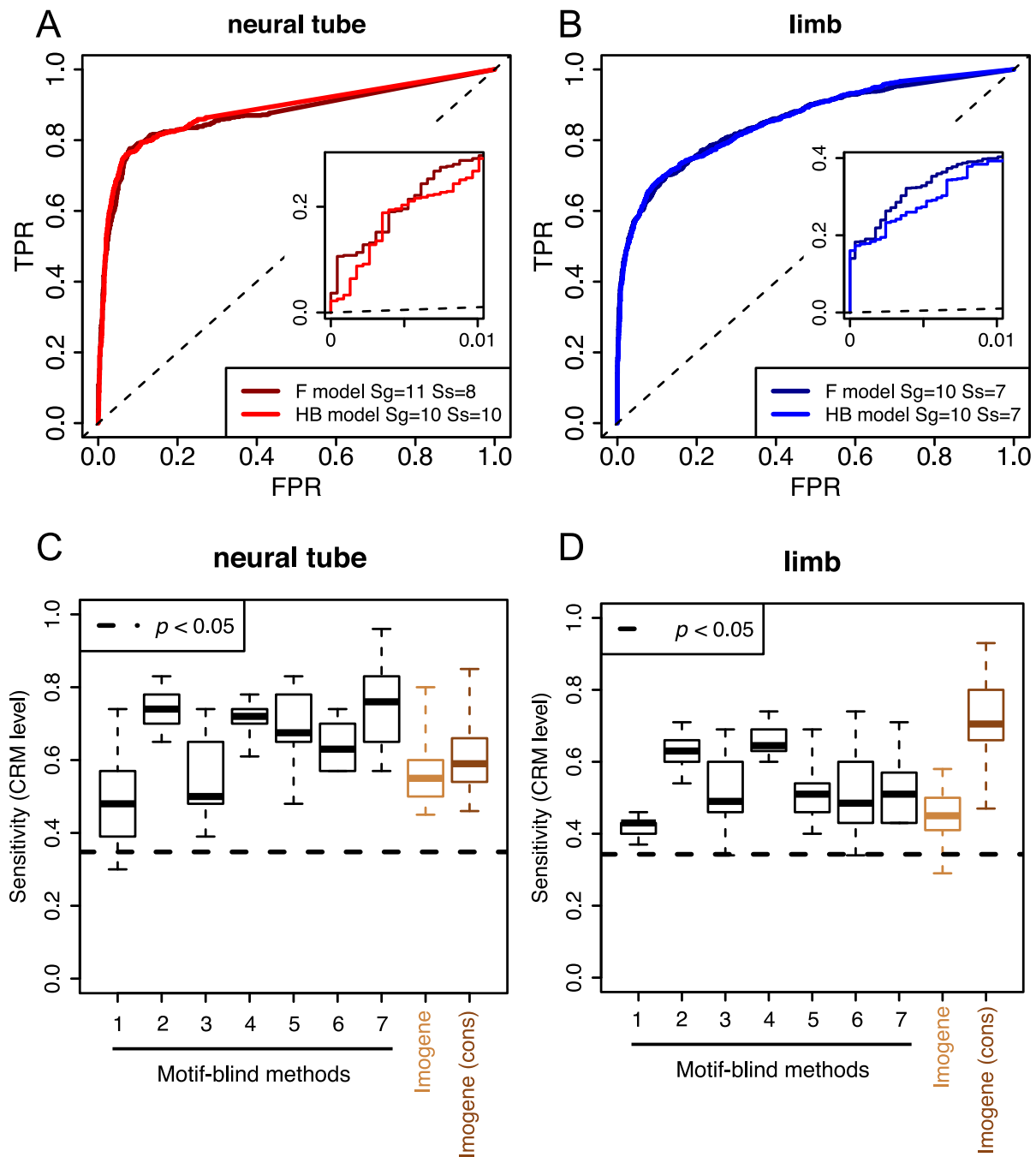


Figure 2. Analysis of well characterized developmental processes. We tested the algorithm on mammalian CRMs driving expression at E11.5 in neural tube (A) and limb (B). For each class, CRMs were divided into a training set and a test set. Motifs were learned on the training set and used to score CRMs from the test set along with background regions consisting of the CRMs' 1 kb flanking sequences (see the Materials and Methods section). The displayed ROC curves show the proportion of test set CRMs recovered above a given score (True Positive Rate denoted by TPR) versus 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 ≤ 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 motif generation and sequence scanning are given in the figures. Black dashed lines show random discrimination. (C and D) Imogene ability to predict neural tube (C) and limb (D) CRMs is compared to that of the 'motif-blind' methods assessed in (22). A leave-one-out cross-validation scheme is used to assess the efficiency of the different methods to predict a bona fide CRM embedded in a 10 kb intergenic region, as described in (22) and in the Materials and Methods section. The box plot summarizes the distributions of sensitivities obtained using 10 different intergenic regions. The dashed line indicates the sensitivity value above which results have a probability smaller than 5% to be generated by chance. The results obtained with *Imogene* are shown when only sequences from a single species (human) is considered at the prediction stage (light brown) or when conservation between mammalian genomes is also used (dark brown). The different methods examined in (22) are denoted by numbers in the figure with the following correspondence with the terminology of (22): 1 = HexDiff.rc, 2 = PAC.rc, 3 = HexYMF.s200.rc, 4 = HexMCD, 5 = D2z.cond.s100, 6 = D2z.cond.m01.weights.rc and 7 = D2z.cond.weights. The HexDiff and HexYMF methods consider the words most associated in a statistical sense with the training set, and then use a weighted sum of word counts to score a given sequence. The PAC method (for Poisson Additive Conditional) also considers the words that are most associated with the training set, and then examines how over-represented each of these words is in the test sequence, relative to the assumed background, by computing a Poissonian p -value. The HexMCD method trains separate fifth-order Markov chains on training modules and background sequences, and quantifies which model matches the test sequence better using a score proposed in (65). The three different variants of the D2z method compute dot products between the k -mer frequency distributions of training and test set sequences, and use their statistical significance (z score) as a scoring scheme. A fuller description of each method is provided in (22).

tary Figure S3 shows the AUC as a function of the number of motifs N for different values of the scanning threshold S_g . One can see that the AUC increases quickly with the first five 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 2 shows the ROC curves obtained for the optimal parameters. They 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 HB and the Felsenstein models are seen in Figure 2 to yield very similar results in both cases. This standard procedure provides a test of the two modes of *Imogene*. Its success indicates that meaningful motifs were generated at the *Genmot* stage and that they were properly used at the *Scangen* stage to recognize the test CRM from background sequences.

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.

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 S4. Figure S4 also shows the closest PWM to each motif in the TRANSFAC (48), JASPAR (49), HT-Selex (50) and UniPROBE (51) list of motifs, as computed by *Imogene* PWM distance. Previously characterized motifs belonging to the considered developmental programs appear in each class (e.g. Oct/Pou TF family and NeuroD motif in the neural CRMs). The motif content of each CRM is also provided in Supplementary Figures S5 and S6. It is seen that the 10 best motifs appear on most CRMs of the training set.

Among the existing algorithms, Kantorovitz *et al.* (22) concluded that 'motif-blind' methods were the most successful for characterizing the specificity of a small set of CRMs. In order to further assess *Imogene*, we thus chose to compare its performance to that of these algorithms. Kantorovitz *et al.* (22) benchmarked the prediction of the algorithms they examined on eight mammalian CRM datasets promoting expression in diverse tissues. We quantified the ability of *Imogene* to characterize these different sets of CRMs using the cross-validation protocol of (22) in which the CRMs to be tested were compared to intergenic sequences with similar GC content (see the Materials and Methods section). Conservation and phylogeny were used for the generation of motifs (i.e. in *Imogene Genmot* mode). The CRM scoring was performed both using conservation, the normal *Scangen* operating mode, and without it in order to provide a fair comparison between *Imogene* and the 'motif-blind' methods that do not take advantage of conservation. The results are displayed in Figure 2 for the neural tube and limb CRM datasets and in Supplementary Figure S7 for the other CRM datasets. In all cases, *Imogene* is found to perform comparably to the 'motif-blind' methods (22). Without conservation, it appears nonetheless less efficient

than the best 'motif-blind' methods, such as 'HexMCD'. The use of conservation significantly enhances *Imogene* predicting power and actually makes it the top predictor for several datasets (5/8). The prediction of specific motifs is, of course, the interesting complementary feature of *Imogene* in both cases.

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 distinguish 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 motifs shared among different classes that, for instance, characterize the binding of generic CRM factors, are 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 (38), we used a 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 follows. 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 that 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 the Materials and Methods section). 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 TPR and FPR for increasing rank. This serves to optimize the different parameters (the threshold for motif generation S_g , the threshold for sequences scanning S_s , and the number of motifs N used to score sequences) by maximizing the AUC for a FPR ≤ 0.2 .

The results are shown in Figure 3. We focus on the results obtained with the HB evolutionary model. Results (motifs and thresholds) are very comparable with the Felsenstein model. 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 two 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 closest motifs in the TRANSFAC (48)

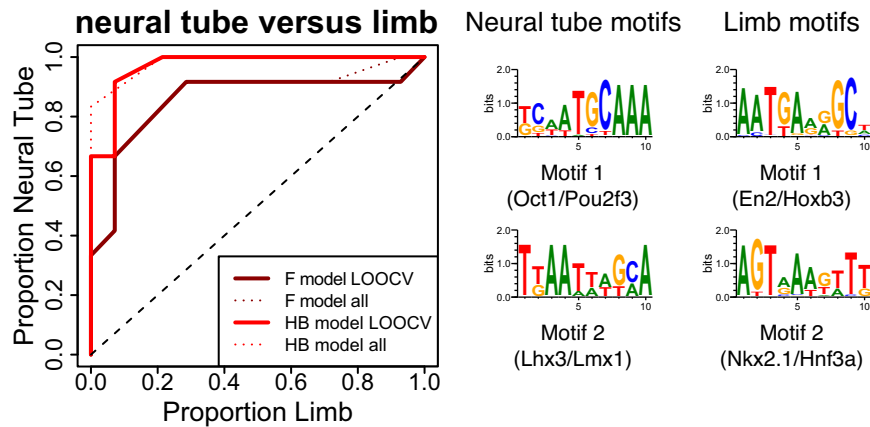


Figure 3. Pattern recognition (mammals). ROC plots showing the discrimination between limb and neural CRMs using a simple linear classifier. Neural and limb classes are compared to each other. Thick lines correspond to a leave-one-out cross-validation (LOOCV) scheme with a score function based on the *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 $S_g = 11$, $S_s = 9$; 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 is still comparable to that computed by the LOOCV, 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 shows random discrimination.

and JASPAR databases (49), as well as in the HT-Selex experiment list (50) and UniPROBE database (51), are shown in Supplementary Figure S8. The best ranking motif of the neural CRMs is found to be unequivocally associated with the TRANSFAC Oct/Pou Transcription Factor known to be involved in the neural tube formation (54).

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 (38). In this work, previously characterized *D. melanogaster* CRMs were divided into five 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). Zinzen *et al.* (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 to belong to a given class with a probability higher than 95% 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 five different classes assembled in (38). We then proceeded as in the previous case of mammalian CRMs.

Imogene results are shown together with the machine learning results of (38) in Figure 4. For clarity, we here show

results obtained with the Felsenstein model. Results obtained with the HB 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 ChIP-seq 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 4C).

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 ChIP-seq data were 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 (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://mobyte.pasteur.fr/cgi-bin/portal.py#forms::imogene>) that provides motif and CRM predictions for the community. This interface is pow-

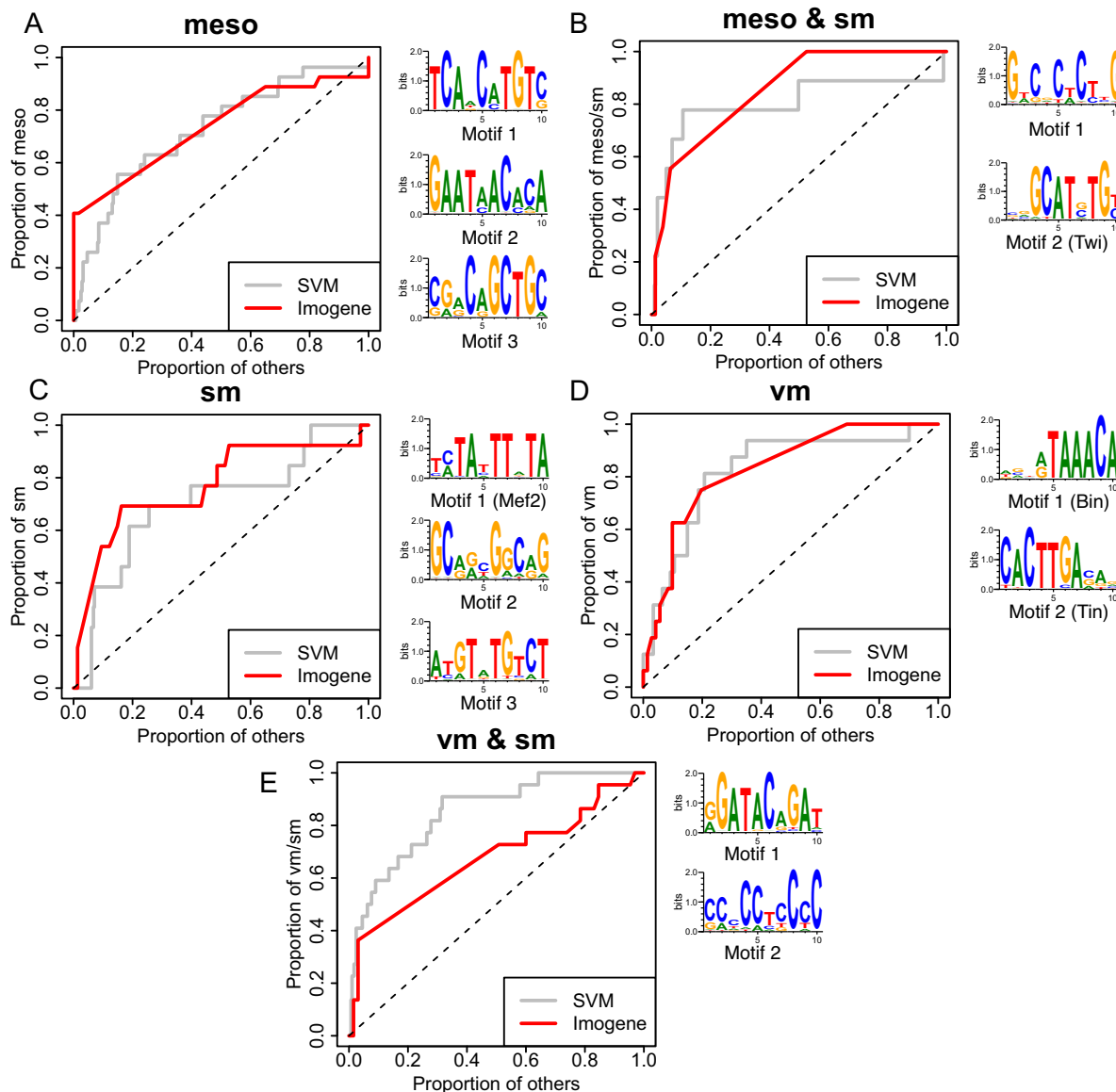


Figure 4. Pattern recognition (*Drosophila*). Recognition of classes of CRMs expressed in five tissue types: mesoderm (meso), somatic muscle (sm), visceral muscle (vm), mesoderm and somatic muscle (meso & sm) and visceral and somatic muscle (vm & sm). ROC plots are obtained using a leave-one-out cross-validation scheme. Two classifiers are compared: a Support Vector Machine using 15 ChIP-on-chip peak heights (gray, replotted using the data and the program provided in (38)), and *Imogene* using the *de novo* generated motifs with Felsenstein evolutionary model (red) and a simple linear classifier (see the Materials and Methods section). 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 ($S_g = 10$, $S_s = 10$) and vm & sm ($S_g = 11$, $S_s = 8$).

ered by the Pasteur Institute Internet server through the mogle framework (55). The input web page and an example output web page are shown in Figures 5 and 6 respectively.

The input form (see Figure 5) 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 the Materials and Methods section).
- Scangen: given the previously generated motifs, produces a list of genome-wide predicted CRMs with conserved bind-

ing sites. The rank of a CRM is based on a Poissonian score that takes into account the CRM content in motifs (as described in the Materials and Methods section)

The group of species considered should also be specified. The algorithm can be used on *Drosophila* (with reference species *D. melanogaster*) or mammals (with reference species *M. musculus*). The different algorithm parameters such as the sought motif width, threshold specificity for binding sites or allowed position shifts between different species (see the Materials and Methods section 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.

* Execution mode ?

General options

* Family of species to consider ?

* Width of the motifs ?

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

Genmot options

* Evolutionary model used for motif generation ?

* Threshold used for motif generation ?

* Threshold used to scan training set sequences for display ?

* 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 ?

* Width of selected enhancers ?

* Number of motifs to consider at maximum ?

* File containing a list of motif definitions ?

Enter your data below:

Figure 5. Web-based interface: input web page. A copy input web page for *Imogene* powered by the mobile bioinformatics framework is shown.

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 HB one. The results of the two modes have been found to be comparable (see Figures 2 and 3).

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 ‘Materials and 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 Supplementary Figure S3).

An example of *Imogene* output is displayed in Figure 6. The *Genmot* mode creates from the provided training set a list of ranked motifs together with their significance and over-representations (see the Materials and Methods section). 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 6A for two motifs. Figure 6B 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

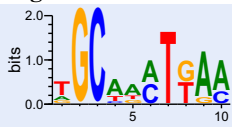
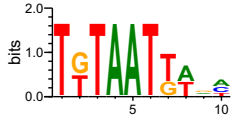
We have presented *Imogene*, a set of statistical tools and a computer software able to predict *de novo* relevant motifs in a moderate size set of functionally related CRMs and able to infer novel CRMs with a low FPR in both *Drosophila* and mammalian genomes. In contrast to methods dedicated to the general discovery of CRMs (see 2156 for recent reviews and assessment), the task requires to extract specific information from a training set that by itself offers only weak statistically discriminative power. This challenge was previously tackled by ‘motif-blind’ methods (22) that aimed to characterize the statistical distribution of short sequences in the training without providing information on specific TFBS. *Imogene* makes use of a different strategy to work efficiently from a CRM set of modest size. It systematically exploits the information available in multiple sequenced genomes with a mode of motif inference that makes intrinsic use of quantitative models for binding site evolution. This leads it to achieve a performance for CRM identification comparable to ‘motif-blind’ algorithms (22) or even superior when motif conservation between different species is used as the CRM identification stage. But, in contrast to ‘motif-blind’ methods, *Imogene* also provides specific information on TFBS, an element of particular biological interest as well as a crucial ingredient for further biological tests of bioinformatic predictions.

Imogene relies on conservation between different species both as filtering step and to enlarge its training set. Phylogenetic conservation between multiple sequenced genomes has previously been shown to provide useful information on *cis*-regulatory motifs (57–59). Although many binding sites are not conserved (60), methods that use conservation among multiple genomes were found superior to single genome methods in a recent assessment of methods devoted to general CRM prediction (56). A simple peak phast-Cons score (61) was in fact found to be surprisingly efficient (56). In addition, ultraconservation has been found to reliably point out functional CRMs (62) in transgenic assays although subsequent deletion of these CRMs did not result in a marked phenotype (63) perhaps because of redundancy or too crude experimental assays.

Phylogenetic conservation, however, 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 (32) analysis, simply comparing the reference genome with another (64–66) or simply adding orthologous sequences (67). Similarly to the *Monkey* algorithm of (28) *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 that lies at *Imogene* core was previously applied to gene co-regulation in *Drosophila* (14). Motifs predicted to be important for sensory-organ-precursor development were confirmed by site-directed mutagenesis. A significant fraction of top predicted new CRMs based on

A Motifs

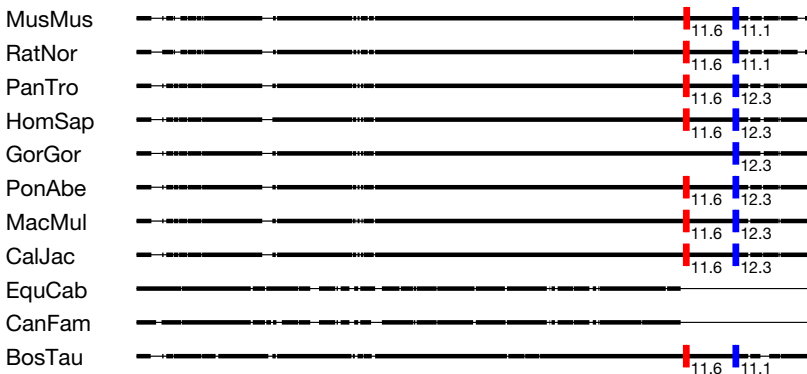
Color	Rank	Logo	P-value (log10)	Over-representation
1	1		-79.4772	55.122
2	2		-76.2578	38.1757

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 TAATTCCTT
CCTGCCTTCT GCCTTCATCC CAAGTTTCAC GGGAAAAAGT AAAGTGTGCA GGTCTTACAG
AGGAGCCTTA TCAAACAGCT GTCATCTGAC AAGCCATTG CATTGTTT GGCTGAAATG
GAGCAACCCA AGGCAAGAT CTTTGTGTC ATTCCATCAT AATGAAGAAA TTACACATFG
TGTAAGAGGC CTGGCTTTAT TTTTAGTTTG CTTGTGTGCT TAAAAGGTA 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 6. 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 five) using the same parameters as in Figure 2. 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 1 kb in the mouse genome at optimal threshold $S_3 = 10$. The five best ranking sequences are shown (default is 200).

this predicted motifs were also shown to direct expression in SOP or more generally in the peripheral nervous system. The interest of predicting motifs *de novo* was further illustrated by a subsequent application of the algorithm to epidermal morphogenesis and trichome development in *Drosophila* (68). The algorithm provided a refined 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 provided here 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 are 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, a capability that is clearly distinct from general CRM prediction (56). In this task, *Imogene* 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 helps to identify new involved DNA-binding cofactors and their binding sites. We anticipate that *Imogene* CRM discriminative ability is likely to be important for future studies of transcription regulation specificity in closely related cell types (e.g. different neuronal cell types) since even large-scale studies will probably not provide more than a few tens of differentially activated CRMs, the training set size targeted by *Imogene*.

We thus believe that *Imogene* is a useful addition to existing algorithms and softwares (32). 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 are also grateful to S. Sinha for sending us the precise benchmark used in Kantorovitz *et al.* (22). We acknowledge the Centre d'Informatique pour la Biologie at the Pasteur Institute for its help in the design of a mobile front-end to *Imogene*.

FUNDING

Centre National de la Recherche Scientifique; Ecole Normale Supérieure; Institut Pasteur; Agence Nationale pour la Recherche [ANR-08-BLAN-0235]. Source of open access funding: institutional funds.

Conflict of interest statement. None declared.

REFERENCES

- Davidson, E.H. (2006) *The Regulatory Genome: Gene Regulatory Networks in Development and Evolution*. Academic, Burlington, MA.
- Dorer, D.E. and Nettelbeck, D.M. (2009) Targeting cancer by transcriptional control in cancer gene therapy and viral oncolysis. *Adv. Drug. Deliv. Rev.*, **61**, 554–571.
- Hardison, R.C. and Taylor, J. (2012) Genomic approaches towards finding *cis*-regulatory modules in animals. *Nat. Rev. Genet.*, **13**, 469–483.
- Lelli, K.M., Slattery, M. and Mann, R.S. (2012) Disentangling the many layers of eukaryotic transcriptional regulation. *Annu. Rev. Genet.*, **46**, 43–68.
- Levine, M. (2010) Transcriptional enhancers in animal development and evolution. *Curr. Biol.*, **20**, R754–R763.
- Arnosti, D.N. and Kulkarni, M.M. (2005) Transcriptional enhancers: intelligent enhanceosomes or flexible billboards? *J. Cell. Biochem.*, **94**, 890–898.
- Swanson, C.I., Evans, N.C. and Barolo, S. (2010) Structural rules and complex regulatory circuitry constrain expression of a Notch- and EGFR-regulated eye enhancer. *Dev. Cell.*, **18**, 359–370.
- Johnson, D.S., Mortazavi, A., Myers, R.M. and Wold, B. (2007) Genome-wide mapping of *in vivo* protein-DNA interactions. *Science*, **316**, 1497–1502.
- Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I. and Zhao, K. (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.
- Mikkelsen, T.S., Ku, M., Jaffe, D.B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.-K., Koche, R.P. *et al.* (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, **448**, 553–560.
- Visel, A., Blow, M.J., Li, Z., Zhang, T., Akiyama, J.A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F. *et al.* (2009) ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, **457**, 854–858.
- Arnosti, D.N. and Kulkarni, M.M. (2005) Transcriptional enhancers: intelligent enhanceosomes or flexible billboards? *J. Cell. Biochem.*, **94**, 890–898.
- Amano, T., Sagai, T., Tanabe, H., Mizushima, Y., Nakazawa, H. and Shiroishi, T. (2009) Chromosomal dynamics at the *Shh* locus: limb bud-specific differential regulation of competence and active transcription. *Dev. Cell*, **16**, 47–57.
- Rouault, H., Mazouni, K., Couturier, L., Hakim, V. and Schweisguth, F. (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**, 14615–14620.
- Flicek, P., Amode, M.R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S. *et al.* (2012) Ensembl 2012. *Nucleic Acids Res.*, **40**, D84–D90.
- 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**, 203–218.
- Wasserman, W.W. and Sandelin, A. (2004) Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.*, **5**, 276–287.
- Stormo, G. and Fields, D. (1998) Specificity, free energy and information content in protein-DNA interactions. *Trends Biochem. Sci.*, **23**, 109–113.
- Su, J., Teichmann, S.A. and Down, T.A. (2010) Assessing computational methods of *cis*-regulatory module prediction. *PLoS Comput. Biol.*, **6**, e1001020.
- Elnitski, L., Jin, V.X., Farnham, P.J. and Jones, S.J. (2006) Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques. *Genome Res.*, **16**(12) 1455–1464.

21. Aerts, S. (2006) Computational strategies for the genome-wide identification of *cis*-regulatory elements and transcriptional targets. *Curr. Top. Dev. Biol.*, **98**, 43–68.
22. Kantorovitz, M., Kazemian, M., Kinston, S., Miranda-Saavedra, D., Zhu, Q., Robinson, G., Götting, B., Halfon, M. and Sinha, S. (2006) Motif-blind, genome-wide discovery of *cis*-regulatory modules in *Drosophila* and mouse. *Dev. Cell*, **17**, 568–579.
23. Machanick, P. and Bailey, T.L. (2011) MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics*, **27**, 1696–1697.
24. 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. *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 757–762.
25. 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 Res.*, **12**, 1019–1028.
26. 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. *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 9888–9893.
27. 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 Biol.*, **2**, 43–68.
28. Moses, A.M., Chiang, D.Y., Pollard, D.A., Iyer, V.N. and Eisen, M.B. (2006) MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. *Genome Biol.*, **5**, R98.
29. Siddharthan, R., Siggia, E. and van Nimwegen, E. (2005) PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput. Biol.*, **1**, e67.
30. 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**, 47–59.
31. Pierstorff, N., Bergman, C. and Wiehe, T. (2006) Identifying *cis*-regulatory modules by combining comparative and compositional analysis of DNA. *Bioinformatics*, **22**, 2858–2864.
32. Herrmann, C., Van de Sande, B., Potier, D. and Aerts, S. (2012) i-cisTarget: an integrative genomics method for the prediction of regulatory features and *cis*-regulatory modules. *Nucleic Acids Res.*, **40**, e114.
33. Nazina, A. and Papatsenko, D. (2003) Statistical extraction of *Drosophilacis*-regulatory modules using exhaustive assessment of local word frequency. *BMC Bioinformatics*, **4**, 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**, 109.
35. Chan, B. and Kibler, D. (2005) Using hexamers to predict *cis*-regulatory motifs in *Drosophila*. *BMC Bioinformatics*, **6**, 262.
36. Leung, G., Eisen, M. and Provart, N. (2009) Identifying *cis*-regulatory sequences by word profile similarity. *PLoS ONE*, **4**, 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. *et al.* (2012) Use of a *Drosophila* genome-wide conserved sequence database to identify functionally related *cis*-regulatory enhancers. *Dev. Dyn.*, **241**, 169–189.
38. Zinzen, R.P., Girardot, C., Gagneur, J., Braun, M. and Furlong, E.E. (2009) Combinatorial binding predicts spatio-temporal *cis*-regulatory activity. *Nature*, **462**, 43–68.
39. Heger, A. and Ponting, C.P. (2007) Variable strength of translational selection among 12 *Drosophila* species. *Genetics*, **177**, 43–68.
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**, 368–376.
43. Halpern, A.L. and Bruno, W.J. (1998) Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Mol. Biol. Evol.*, **15**, 910–917.
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**, 160–174.
45. Seplyarskiy, V.B., Kharchenko, P., Kondrashov, A.S. and Bazykin, G.A. (2012) Heterogeneity of the transition/transversion ratio in *Drosophila* and *Hominidae* genomes. *Mol. Biol. Evol.*, **29**, 1943–1955.
46. Bishop, C. (2006) *Pattern Recognition and Machine Learning*. Springer, New York, NY.
47. Bao, Z. and Eddy, S. (2002) Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res.*, **12**, 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. *et al.* (2006) TRANSFAC and its module TRANSCOMP: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, **34**(Database issue), D108–D110.
49. Portales-Casamar, E., Thongjuea, S., Kwon, A.T., Arenillas, D., Zhao, X., Valen, E., Yusuf, D., Lenhard, B., Wasserman, W.W. and Sandelin, A. (2010) JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **38**(Database issue), D105–D110.
50. Jolma, A., Yan, J., Whittington, T., Toivonen, J., Nitta, K.R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G. *et al.* (2013) DNA-binding specificities of human transcription factors. *Cell*, **152**, 327–339.
51. Newburger, D.E. and Bulyk, M.L. (2009) UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic Acids Res.*, **37**(Database issue), D77–D82.
52. Kimura, M. (1962) On the probability of fixation of mutant genes in a population. *Genetics*, **47**, 43–68.
53. Hastie, T., Tibshirani, R. and Friedman, J. (2001) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, New York, NY.
54. Kiyota, T., Kato, A., Altmann, C.R. and Kato, Y. (2008) The POU homeobox protein Oct-1 regulates radial glia formation downstream of Notch signaling. *Dev. Biol.*, **315**, 579–592.
55. Neron, B., Menager, H., Maufrais, C., Joly, N., Maupetit, J., Letort, S., Carrere, S., Tuffery, P. and Letondal, C. (2009) Mobylye: a new full web bioinformatics framework. *Bioinformatics*, **25**, 3005–3011.
56. Su, J., Teichmann, S.A. and Down, T.A. (2010) Assessing computational methods of *cis*-regulatory module prediction. *PLoS Comput. Biol.*, **6**, e1001020.
57. 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**, 338–345.
58. 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**, R104.
59. 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**, 219–232.
60. Schmidt, D., Wilson, M.D., Ballester, B., Schwalie, P.C., Brown, G.D., Marshall, A., Kutter, C., Watt, S., Martinez-Jimenez, C.P., Mackay, S. *et al.* (2010) Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science*, **328**, 1036–1040.
61. Siepel, A. and Haussler, D. (2004) Combining phylogenetic and hidden Markov models in biosequence analysis. *J. Comput. Biol.*, **11**, 413–428.
62. Pennacchio, L.A., Ahituv, N., Moses, A.M., Prabhakar, S., Nobrega, M.A., Shoukry, M., Minovitsky, S., Dubchak, I., Holt, A., Lewis, K.D. *et al.* (2006) *In vivo* enhancer analysis of human conserved non-coding sequences. *Nature*, **444**, 499–502.
63. Ahituv, N., Zhu, Y., Visel, A., Holt, A., Afzal, V., Pennacchio, L.A. and Rubin, E.M. (2007) Deletion of ultraconserved elements yields viable mice. *PLoS Biol.*, **5**, e234.
64. Wang, T. and Stormo, G. (2003) Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics*, **19**, 2369–2380.
65. 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**, 2738–2750 .
66. Zhao,G., Schriefer,L. and Stormo,G. (2007) Identification of muscle-specific regulatory modules in *Caenorhabditis elegans*. *Genome Res.*, **17**, 348–357 .
67. Busser,B.W., Taher,L., Kim,Y., Tansey,T., Bloom,M.J., Ovcharenko,I. and Michelson,A.M. (2012) A machine learning approach for identifying novel cell type-specific transcriptional regulators of myogenesis. *PLoS Genet.*, **8**, e1002531.
68. Menoret,D., Santolini,M., Fernandes,I., Spokony,R., Zanet,J., Gonzalez,I., Latapie,Y., Ferrer,P., Rouault,H., White,K.P. *et al.* (2013) Genome-wide analyses of Shavenbaby target genes reveals distinct features of enhancer organization. *Genome Biol.*, **14**, R86.