Statistical modeling of bacterial promoter sequences for regulatory motif discovery with the help of transcriptome data: application to *Listeria monocytogenes*

Ibrahim Sultan[1*], Vincent Fromion[1], Sophie Schbath[1], Pierre Nicolas[1*]

1 MaIAGE UR1404, INRA, Université Paris-Saclay, Jouy-en-Josas, France

* ibrahim.sultan@inra.fr (IS); pierre.nicolas@inra.fr (PN)

Abstract

Powerful algorithms have been developed for the discovery of bacterial transcription factors binding sites but automatic *de novo* identification of the main regulons of a bacterium from genome and transcriptome data remains a challenge. To address this task, we propose a statistical model of the DNA sequence that can use information on exact positions of the transcription start sites and condition-dependent expression profiles. Two main novelties of this model are to allow overlaps between motif occurrences and to incorporate covariates summarizing expression profiles into the probabilities of motif occurrences in promoter regions. Covariates may correspond to the coordinates of the genes on axes such as obtained by Principal Component Analysis or Independent Component Analysis, or to the positions of the genes in a tree such as obtained by hierarchical clustering. All parameters are estimated in a Bayesian framework using a dedicated trans-dimensional Markov chain Monte Carlo algorithm that adjusts, simultaneously, for many motifs and many expression covariates: the width and palindromic properties of the corresponding position-weight matrices, the number of parameters to describe position with respect to the transcription start site, and the choice of relevant covariates. A data-set of 1,512 transcription start sites and 165 expression profiles available for *Listeria monocytogenes* served to validate the approach, which proved able to predict regulons corresponding to many known transcription factors (SigB, CcpA, Rex, LiaR, Fur, LexA, VirR, Spx, BglR2, SigL, PrfA). The results provides thus a new global prediction on the transcription regulatory network of this model food-borne pathogen. For instance, a previously unreported motif was found upstream the promoters controlling the expression of many ribosomal proteins which suggests that an unknown transcription factor might play an important role in the regulation of growth.

Author summary

Statistical models of DNA sequences using position-weight matrix to represent motifs recognized by transcription factors are widely used for the reconstruction of transcriptional regulatory networks. However, due to motif degeneracy and limited number of motif occurrences, these approaches are usually successful only when applied to biologically relevant sets of sequences enriched for a particular motif. Such sets are most often defined based on experimental data (expression or protein-DNA interaction profiles). Hence, transcriptional regulatory network reconstruction tends to be an incremental process in which new components of the network are added one by one. In

this work, we developed a coherent statistical framework to tackle motif discovery at a system level by incorporating two types of "global" experimental data that are increasingly available for bacteria: maps of the transcription start sites and rich transcriptome data-sets. Rather than to search for DNA sequence features able to explain/predict these experimental data, we use them to improve the modeling of the DNA sequences. Thereby, the benefits of the use of simple well-established probabilistic sequence models are preserved and it remains possible to find motifs that are not connected to the transcriptome data.

Introduction

Motif discovery in DNA sequences is one of the oldest problem in the field of bioinformatics [1] but automatic *de novo* identification of the main regulons of an organism as "simple" as a bacterium remains a challenge. There have been hundreds of publications detailing motif discovery methods and numbers of surveys on these methods [2–4]. The representation chosen for the motifs, which can be position-weight matrix (PWM) or word-based, is at the heart of the methodology. Word-based methods usually represent motifs as consensus strings written in an alphabet allowing degenerate symbols (IUPAC system) and rely, for scoring, on an hypothesis testing framework to define a notion of unexpected words with respect to a null hypothesis. Examples of null hypotheses are that words are generated from a background Markov model [5] or that the frequency of occurrences are the same between sequences associated with genes in different co-expression clusters [6, 7]. PWM representation aim to offer more precise description of the motifs than word-based methods. A motif of length $\mathcal{W}$ is represented by a $4 \times \mathcal{W}$ matrix, that records the frequencies of the 4 DNA nucleotides (A,C,G,T) at each position within the motif. Beside this probabilistic representation of the motif, a full probabilistic model involves also a background model accounting for the nucleotides outside motif occurrences and a model for the position of the motif in the sequence set. Motif discovery is then cast as the problem of estimating PWMs. MEME [8] and Gibbs Motif Sampler [9] were the first algorithms implementing these ideas and remain among the most powerful and widely used tools to search for motifs based on the sole properties of nucleotide composition of the set of sequences under consideration.

Chromatin immunoprecipitation (ChIP) coupled with hybridation on microarray (ChIP-Chip) and later with DNA sequencing (ChIP-Seq) permits the genome-wide identification of protein-DNA interactions *in vivo*. It is probably the experimental technique that had the deepest impact on the field of motif discovery [4]. Nevertheless, ChIP has intrinsic limitations for *de novo* motif discovery at a system level because the approach requires *a priori* selection of the combinations of TFs and biological conditions that are relevant and dedicated ChIP experiments are needed for each of these combinations. Furthermore, ChIP analysis of a specific TF–DNA interaction is not completely straightforward: it requires either a specific antibody to recognize the TF or genetic engineering of a functionally active tagged TF that can be recognized by an already available antibody. In parallel to the development of ChIP, microarrays and RNA-Seq have been widely applied to compare gene expression between growth conditions ("*in vitro*" stress, stage of growth, infection condition, ...) and/or specific mutants. All types of data can be informative and have been early and widely used for bacterial transcriptional network reconstruction [10, 11]. When the goal is to unravel globally with limited resources the transcription regulatory network of a bacterium on which we have limited knowledge, the methods that rely on expression data do not present the aforementioned disadvantages of the ChIP-based method. For instance, expression profiles across many biologically relevant growth and stress conditions can be obtained without focusing on specific TFs and without genetic engineering [12, 13].

In a systems biology perspective, our work focuses on making use of transcriptome data. The development of specific methods for *de novo* motif discovery using expression profiles has probably attracted less attention than the use of ChIP data. It is also clearly a more difficult task because the experimental data is not directly relevant for a specific TF–DNA interaction. Nevertheless, a diversity of approaches has been proposed and implemented, these are based on different methodological concepts such as: mutual information in FIRE [14], enrichment test in GEMS [6], or regression of expression data y given sequence data x in REDUCE [15] and MatrixREDUCE [16]. The first two approaches involve transforming the expression data into one-dimensional categorical values, while the third approach can directly accommodate multidimensional continuous expression data. The algorithm implemented in RED2 [7] intended to bypass the need for clustering of the first two approaches by applying enrichment tests or computing mutual information for overlapping sets of genes that are close in the expression space (neighborhoods).

These approaches face the difficulty to find appropriate summary or probabilistic models for expression data and present the obvious limitation of being entirely focused on the discovery of sequence motifs that are linked to the available expression data. An alternative viewpoint, which is adopted in this work, consists in modeling the sequence data x given the expression data y. A key benefit of this modeling choice is to build upon the powerful sequence modeling approaches for *de novo* motif discovery based on PWMs and full probabilistic modeling of the sequences, such as MEME and the Gibbs Motif Sampler [8, 9]. This makes it possible to envision the simultaneous use of the expression data and of all the statistical properties of the sequence, establishing a continuum between discovery of motifs related and unrelated to the available expression data. In [12], we already proposed a statistical model for the discovery of Sigma factor binding sites based on modeling $x|y$. The generality of the model was however limited since it was specifically tailored for the discovery of sigma factor binding sites whose specificity is to delineate and partition the promoter space [17, 18]. For this reason, regulation by sigma factors tends to correspond to a preponderant and non-overlapping level of regulation that is particularly well captured by the structure of a hierarchical clustering tree and did not justify to model the occurrence of more than one motif per sequence [12].

Along with condition-dependent expression data, another type of transcriptome data which is now increasingly available for bacteria are genome-wide maps of TSSs at single base resolution that are obtained by specific RNA-Seq protocols. The general idea consists of ligating an adapter on the 5'-end of the RNA molecule during the library preparation and sequencing from this adapter, its implementation comes in different refinements [19–22]. These data are particularly relevant with respect to the objective of the present work because they provide a repertoire of promoter regions that can be aligned with respect to the TSS and thus will allow to focus the search for regulatory motifs. Incorporation of positional information in the sequence model for *de novo* motif discovery proved helpful to discover sigma factor binding sites [12, 23] whose positions are strongly constrained with respect to the TSS.

To address the task of automatic *de novo* identification of the main regulons (not restricted to sigma factors) of a bacterium from genome and transcriptome data, this work develops a coherent probabilistic model of the DNA sequence allowing to search simultaneously for multiple motifs and to integrate precise information on the position of the transcription start sites and of condition-dependent transcription profiles. Two main novelties of the proposed model are to allow overlaps between motif occurrences and to incorporate covariates summarizing transcription profiles into the probability of occurrence in a given promoter region. Covariates can correspond to the coordinate of the gene on an axis such as obtained by PCA (see [24] for an early application to

transcriptome data) or ICA [25, 26] but we also show how to use as covariates positions in a hierarchical clustering trees [10, 12, 27]. All the parameters are estimated in a Bayesian framework using a dedicated trans-dimensional MCMC algorithm. In order to validate the approach, we applied it to the food-borne pathogenic bacterium *L. monocytogenes*. Owing to its status as important human pathogen and model bacterium for the study of host-pathogen interaction and bacterial transcriptomics [28], a wealth of transcriptomics data has been collected on this bacterium. In particular, one could cite the landmark study using RNA-Seq and high-density tiling arrays [29], the early use of genome-wide TSS mapping [30], and the work done to aggregate the diversity of transcriptomics data in a single database [31].

Materials and methods

Probabilistic model of promoter sequences

Model overview

The sequence data that we consider consists of a set of $\mathcal{N}$ DNA sequences of same length $\mathcal{L}$. Let denote by $x = (x_n)_{n=1:\mathcal{N}}$ the set of DNA sequences and $x_{n,l} \in \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ the nucleotide in position l of the nth sequence. Our probabilistic representation for these sequences involves $\mathcal{M} + 1$ unobserved components that capture the heterogeneity of nucleotide composition. These components consist of $\mathcal{M}$ motif models (PWMs), denoted by $(\theta_1, \ldots, \theta_{\mathcal{M}})$, with respective widths $(w_1, ..., w_{\mathcal{M}})$ and a background Markov model of order v whose parameters are denoted by θ_0. All the $\mathcal{M}$ motifs are simultaneously searched for and the model makes use of expression data to define the probability of each motif in each sequence. For simplicity, the model assumes zero or one occurrence per sequence (in keeping with the simple ZOOPS model of MEME [8]) of each motif and the occurrences of the different motifs are modeled as mutually independent.

Statistical inference is carried out in a Bayesian framework and the parameters are thus treated as random variables drawn from prior distributions. The Directed Acyclic Graph (DAG) shown in Fig 1 represents the factorization of the joint probability distribution of the variables (parameters, hidden variables, observed data) that compose the model. We present below the most salient characteristics of our model whose complete and detailed mathematical presentation is found in S1 Appendix.

The observed sequence data x is at the center of this DAG. Connected to x, the hidden variables a and b respectively describe the position of occurrences of the motifs and disambiguate the contribution of the overlapping occurrences. Namely, $a = (a_{m,n})_{m=1:\mathcal{M},\, n=1:\mathcal{N}}$ is a layer of random variables that encode the position of the occurrences of the motifs, with $a_{m,n} \in \{0, 1, \ldots, \mathcal{L}\}$ being the position of motif m in sequence n; 0 encodes the absence of occurrence. For practical reasons pertaining to the implementation of the update of the motif width w_m in the MCMC algorithm we do not systematically rely on the first position within motif m to record its position. Instead, we use the reference position $r_m \in \{1, \ldots, w_m\}$ which allows motif occurrences to extend outside of sequences. A motif can extend up to $r_m - 1$ positions upstream of the sequence (case $a_{m,n} = 1$) and up to $w_m - r_m$ positions downstream of the sequence (case $a_{m,n} = \mathcal{L}$).

Incorporating expression data as covariates using the probit model

Information from expression data is incorporated into the probability of occurrence of a motif via a probit regression framework. Implementation of this model relies on a data augmentation scheme [32] that introduces a Gaussian latent variable, denoted here by

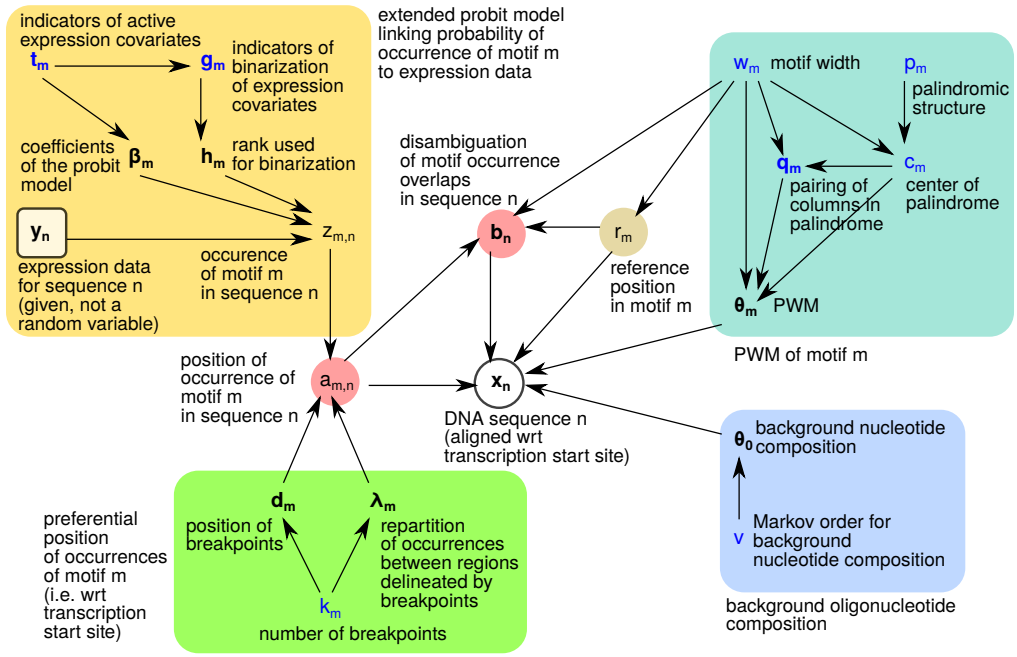

Fig 1. Hierarchical model of promoter sequences. Directed Acyclic Graph of the model. Vertices represent the variables (parameters, hidden variables, observed data) and edges show the factorization of the joint probability distribution. Colored areas highlight groups of variables that contribute to a same ingredient of the model. The upper-left part of the DAG is dedicated to the incorporation of expression data in the sequence model. In this representation we show the variables for only one motif m and one sequence n. Multidimensional variables (vectors or matrices) are boldfaced; variables whose values change the dimension of the model are in blue; continuous random variables are denoted by Greek letters.

$z_{m,n}$, with the event $\{z_{m,n} > 0\}$ corresponding to the occurrence of motif m in sequence n, which corresponds also to $\{a_{m,n} > 0\}$.

Let's consider the availability of a number $\mathcal{C}$ of vectors of continuous variables denoted by $y = (y_{n,c})_{n=1:\mathcal{N},c=1:\mathcal{C}}$ that can be used as covariates. The standard probit model would then write the probability of occurrence of motif m in sequence n as

$$\pi(a_{m,n} > 0|y, \beta) \quad = \quad \Phi(\beta_{m,0} + \sum_{c=1:\mathcal{C}} \beta_{m,c} y_{n,c}) \,, \tag{1}$$

where β contains the regression coefficients ($\beta_{m,c} \in \mathbb{R}$) and Φ is the cumulative distribution function of the standard normal distribution which serves to map any value of $\beta_{m,0} + \sum_{c=1:\mathcal{C}} \beta_{m,c} y_{n,c}$ into a probability between 0 and 1.

We developed an extension for this model that can be interpreted as to allow binarization of expression covariates according to automatically adjusted breakpoints. This extension is very appealing in our modeling context because it allows to model sharp switches in the probability of occurrences as a function of the position of the sequence in the expression space, like if we were also entering as covariates all the co-expression clusters that could be defined on the basis of gene coordinates on axis c. Importantly, this does not implies to model the probability of occurrence as jumping between 0 and 1 as when modeling sharp switches with the standard probit.

To allow automatic selection of the covariates relevant to predict occurrence of motif m and the aforementioned binarization (both choices impacting on model dimension), our model involves the following variables,

- $t_m = (t_{m,1}, \ldots, t_{m,\mathcal{C}})$ where $t_{m,c} \in \{0,1\}$ indicates whether covariate c should be taken into account;
- $\beta_m = (\beta_{m,0}, \beta_{m,1}, \ldots, \beta_{m,\mathcal{C}})$ where $\beta_{m,c} \in \mathbb{R}$ represents, when $t_{m,c} = 1$, a coefficient specific to covariate c, $\beta_{m,0}$ is the intercept parameter;
- $g_m = (g_{m,1}, \ldots, g_{m,\mathcal{C}})$ where $g_{m,c} \in \{0,1\}$ indicates, when $t_{m,c} = 1$, whether the values in the vector $y_{\cdot,c}$ is binarized ($g_{m,c} = 1$) or not ($g_{m,c} = 0$);
- $h_m = (h_{m,1}, \ldots, h_{m,\mathcal{C}})$ where $h_{m,c} \in \{1, \ldots, \mathcal{N}-1\}$ indicates, when $t_{m,c} = 1$ and $g_{m,c} = 1$, the rank in the vector $y_{\cdot,c}$ of the value used for binarization, $y_{(h_{m,c}),c}$ denotes the corresponding cut-off.

In keeping with the probit regression framework, the probability of occurrence of motif m in sequence n writes then as

$$\pi(a_{m,n} > 0 | y, t, \beta, b, h) \quad = \quad \Phi\big(\beta_{m,0} + \sum_c \beta_{m,c} \mathbb{I}\{t_{m,c} = 1\} \tilde{y}_{m,n,c}\big)\,, \tag{2}$$

where $\mathbb{I}$ is the indicator function ($\mathbb{I}\{z\} = 1$ if z is true, 0 otherwise) and $\tilde{y}_{m,n,c}$ corresponds to $y_{n,c}$ after an motif-specific eventual binarization and

$$\begin{aligned} \tilde{y}_{m,n,c} \quad = \quad & \mathbb{I}\{g_{m,c} = 0\} y_{n,c} \\ & + \mathbb{I}\{g_{m,c} = 1\} \left[\Big(\frac{h_{m,c}}{\mathcal{N}} - 1 \Big) \mathbb{I}\{y_{n,c} \leq y_{(h_{m,c}),c}\} + \frac{h_{m,c}}{\mathcal{N}} \mathbb{I}\{y_{n,c} > y_{(h_{m,c}),c}\} \right]. \end{aligned} \tag{3}$$

When binarization is active (case $g_{m,c} = 1$), Eq (3) maps $y_{n,c} > y_{(h_{m,c}),c}$ to $h_{m,c}/N$ and $y_{n,c} \leq y_{(h_{m,c}),c}$ to $h_{m,c}/N - 1$. While mapping to other values would in principle be equivalent provided that the β's are adjusted accordingly, this mapping to $h_{m,c}/N$ and $h_{m,c}/N - 1$ presents the two-fold advantage to ensure the centering to 0 of $\tilde{y}_{m,n,c}$ whatever the value of $h_{m,c}$ when $g_{m,c} = 1$ and to ensure a difference of 1 between the two possible values of the binarized covariates (hence preserving the interpretation of β across the possible values for $h_{m,c}$). The centering is important for the mixing of the MCMC algorithm since the $\mathcal{C}$ blocks of variables $(t_{m,c}, g_{m,c}, \beta_{m,c}, h_{m,c})$ are updated separately for each (m, c).

Fig 2A and 2B illustrate the relationship between expression covariates and probability of motif occurrence allowed by the simple probit model and by our extension involving binarization.

Handling trees as covariates

Further extending the model described by Eq (2) and Eq (3), we consider that covariates can come not only in the form of vectors of continuous variables but also in the form of trees. Indeed, the binary representation of the covariates makes it also possible to incorporate whole tree structures in the regression model. In this case, the binarization breakpoint is located along the branches of the tree instead of along a simple axis.

In our context, the trees are obtained by usual techniques of hierarchical clustering. They are thus rooted binary trees in which all the leaves are at a same distance from the root. Such a tree covariate c, contains $\mathcal{N}-1$ internal nodes and can be conveniently encoded (topology and branch lengths) as a vector of internal node heights $h_c = (h_{c,i})_{i=1:\mathcal{N}-1}$ ordered such as $h_{c,i} \leq h_{c,i+1}$ and a $(\mathcal{N}-1) \times 2$ matrix $s_c = (s_{c,i,1}, s_{c,i,2})_{i=1:\mathcal{N}-1}$ where $(s_{c,i,1}, s_{c,i,2})$ identifies the pair of subtrees merged at internal node i. In practice, each subtree is identified by its top node, itself identified by a value in $\{-\mathcal{N}, \ldots, -1\} \cup \{1, \ldots, \mathcal{N}-1\}$, with negative values corresponding to

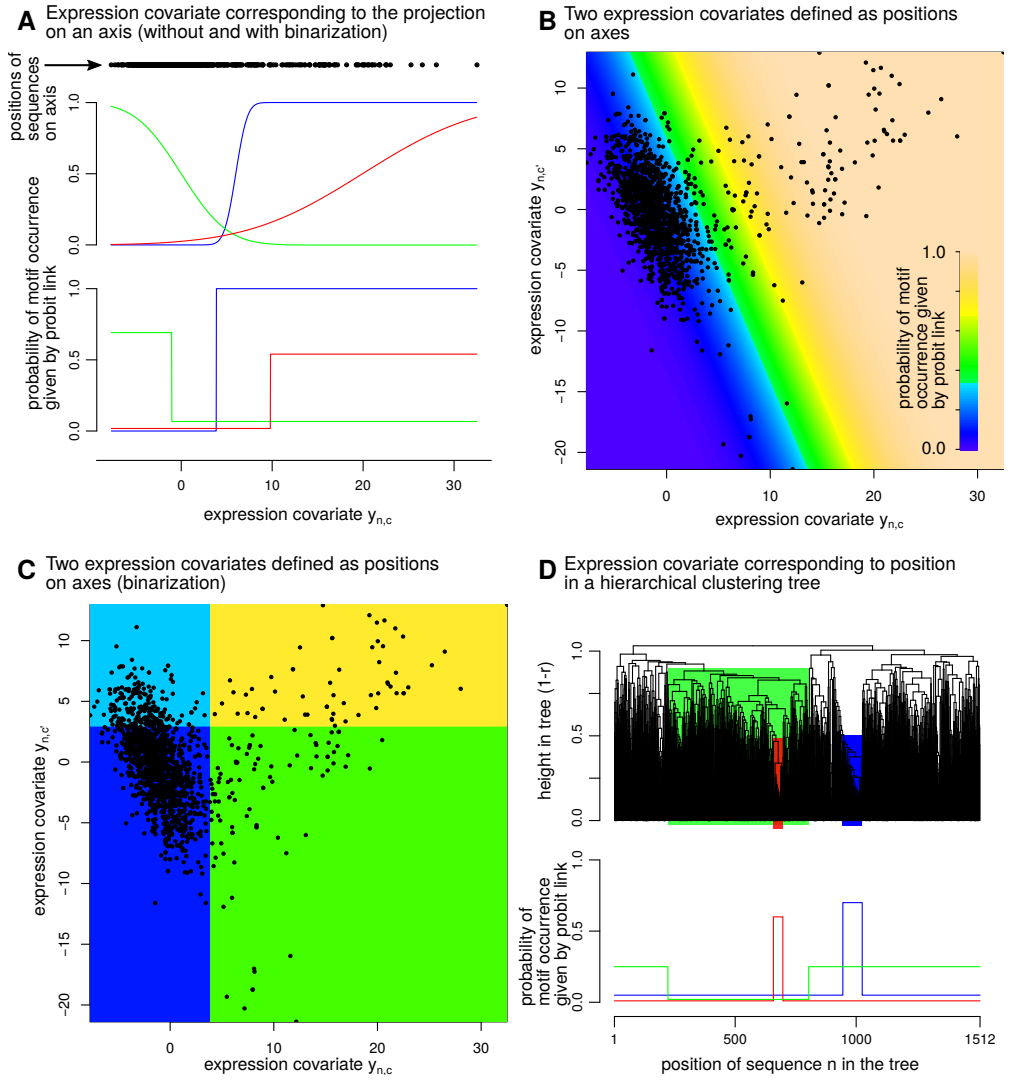

Fig 2. Extended probit model. **A** Graphical illustration of the use of the probit model to link probability of motif occurrence to one expression covariate. Upper plot: classical probit model. Lower plot: probit model with covariate binarization. **B** Graphical illustration of the use of the probit model of probability of motif occurrence to link probability of motif occurrence to two expression covariates. **C** Graphical illustration of the use of the probit model of probability of motif occurrence to link probability of motif occurrence to two expression covariates with binarization. **D** Graphical illustration of the use of the probit model of probability of motif occurrence to link probability of motif occurrence to the position in a tree.

terminal nodes (in our case the indexes of the sequences) and positive values corresponding to internal nodes ordered as in vector h_c.

The vector h_c and the two-columns matrix s_c, taken together, replace the vector y_c when the covariate c is a tree. Furthermore, $g_{m,c} = 1$ since trees can be accommodated by our model only in the form of binarized covariates. Compared to a covariate of type "vector", the interpretation of the $h_{m,c}$ is slightly changed such as it identifies a subtree (i.e. a row in matrix s_c). The mapping of the covariate to $\tilde{y}_{m,n,c}$, is also adapted. We used, $\tilde{y}_{m,n,c} = |h_{m,c}|/\mathcal{N} - 1$ if sequence n belongs to the subtree $h_{m,c}$ and $|h_{m,c}|/\mathcal{N}$

otherwise, where we use the notation $|h_{m,c}|$ for the number of sequences in the subtree $h_{m,c}$. Figure 2C illustrates how the selection of a subtree allows to associate different probability of motif occurrence to different sequences.

Modeling occurrence position with respect to the TSS

We consider DNA sequences that are aligned with respect to experimentally determined TSSs such that the position of motif occurrence encoded in $a_{m,n}$ corresponds to a precise position relative to the TSS.

Given the presence of an occurrence of motif m in sequence n (event $\{a_{m,n} > 0\}$ modeled by the extended probit), the probability distribution of its exact position is defined by a piecewise constant function with k_m breakpoints located at $d_m = (d_{m,k})_{k=1:k_m}$, with $1 < d_{m,k} < d_{m,k+1} \leq \mathcal{L}$. The probability of occurrence of the motif at a particular position $a_{m,n}$ in region k (i.e. $d_{m,k-1} \leq a_{m,n} < d_{m,k}$) is $\lambda_{m,k}/(d_{m,k} - d_{m,k-1})$ where $\lambda_{m,k} \geq 0$ is the probability to find this occurrence in region k ($\sum_{k=1}^{k_m+1} \lambda_{m,k} = 1$).

Allowing motif occurrences to overlap

The model allows motif occurrences to overlap. This feature underpins the simplifying assumption of mutual independence between the occurrences of the $\mathcal{M}$ motifs made when modeling the precise position of the motif with respect to the TSS in conjunction with the relationship between expression data and probability of motif occurrence. Another key benefit of allowing overlap between motif occurrences is to allow the update motif width (w_m) without having to avoid motif collisions.

When $o_{n,l}$ motifs overlap position l in sequence n, $x_{n,l}$ is modeled as drawn from a distribution that is the arithmetic mean of the PWM columns associated with the different motifs which overlap. Namely,

$$\pi(x_{n,l}|a, r, \theta, o_{n,l} \geq 1) \quad = \quad \frac{1}{o_{n,l}} \sum_{m \in O_{n,l}} \theta_{m,l-a_{m,n}+r_m,x_{n,l}} \,, \tag{4}$$

where $O_{n,l}$ denotes the set of motifs that overlap position l, and $\theta_{m,w,u}$ is the probability of nucleotide u at position w of motif m. This modeling choice implies that resulting density can be seen as the marginal of an equal-weight mixture model. In keeping with the usual data augmentation scheme for Bayesian inference of mixture models [33], a layer of latent random variables is introduced to encode the disambiguation of the overlaps between motifs: $b = (b_{n,l})_{n=1:\mathcal{N},\, l=1:\mathcal{L}}$, with $b_{n,l} \in \{0, 1, \ldots, \mathcal{M}\}$ denoting the model component (background or one of the $\mathcal{M}$ motifs) responsible for the probability distribution function of the nucleotide $x_{n,l}$, i.e.

$$\begin{aligned} \pi(x_{n,l}|a, b, r, \theta) \quad = \quad & \mathbb{I}\{b_{n,l} = 0\}\theta_{0,x_{n,l-v}\ldots x_{n,l-1},x_{n,l}} \\ & + \mathbb{I}\{b_{n,l} \geq 1\}\theta_{b_{n,l},l-a_{m,n}+r_m,x_{n,l}} \,, \end{aligned} \tag{5}$$

where we use the notation $\theta_{0,u_1\ldots u_v,u}$ for the probability of nucleotide u after the word $u_1 \ldots u_v$ of length v in the background.

Palindromic constraints in PWMs

Dimeric nature and symmetry properties of many transcription factors explain that many of the known binding motifs are palindromic in the sense that pairs of PWM columns at symmetric positions with respect to the center of the motif appear as mirrored according to Watson-Crick base pairing rule (`A` : `T` and `C` : `G`). Modeling these palindromic constraints on PWMs can be useful because they reduce the dimension of

the model and thereby increase the amount of data available to estimate each parameter and simultaneously decrease the size of the search space for the parameter values.

Instead of imposing a strict palindromic constraint on all or a subset of the motifs, we developed a more flexible model that allows, for each motif m, intermediate states between non-palindromic and fully-palindromic structures. This approach relies on the introduction of a set of variables that define the active constraints on $\theta_m = (\theta_{m,w,u})_{w=1:w_m, u\in\{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}}$:

- p_m, a binary variable indicating if the motif m is allowed to have a (possibly partial) palindromic structure (case $p_m = 1$),
- $c_m \in \{1.5, 2, 2.5, \ldots, w_m - 0.5\}$, a discrete variable recording the position of the center of symmetry of the palindromic structure; the range for c_m allows to types of palindromic structure ("even" and "odd" types, where the odd type contains a central unpaired column at $w = c_m$);
- $q_{m,w}$, a binary variable to indicate whether columns w and $2c_m - w$ are paired, i.e. $\theta_{m,w,u} = \theta_{m,2c_m-w,\bar{u}}$ when $(u, \bar{u})$ is a Watson-Crick pair.

There are several important motivations for allowing intermediate levels of constraints between the non-palindromic and the fully-palindromic structures. First, the overarching goal to set up an algorithm that could identify as many motifs as possible simultaneously is incompatible with the idea of having all (or none of) the motifs palindromic. Second, the variables (p_m, c_m, q_m) allow smooth transitions between non-palindromic and palindromic motif representation which makes it possible to design algorithms that gradually increase or decrease the number of free parameters in θ_m. Finally, while some level of palindromic structure is often obvious it may be unclear to which extent a biological motif is fully palindromic or partially palindromic.

Bayesian inference

The priors that we used intend to be non-informative. Their choices takes also into account considerations on the tractability of the MCMC updates and, as usual in Bayesian inference, most of these priors are mutually independent conjugate priors [34].

A Markov Chain Monte Carlo (MCMC) algorithm was built with the joint posterior

$$a, b, v, \theta_0, w, p, c, q, \theta, r, k, d, \lambda, t, \beta, \delta, h \mid x, y \tag{6}$$

as target distribution. Since the dimension of the target is high, this MCMC algorithm is composed of many different steps whose purposes are to update separate subsets of variables [33, 35], a type of algorithm that has been referred to as block MCMC sampler [36]. A sweep consists of different steps such as to allow the update of all the variables. Each step is designed to preserve the target distribution by sampling from the conditional distribution of a block of variables given all the other variables. The block MCMC sampler that we implemented involves 14 types of steps that can be conveniently divided into three categories: Metropolis-Hastings (MH) steps that consist of drawing from a proposal distribution and applying an accept/reject procedure; Gibbs steps that consist of sampling directly from the target (conditional distribution of the variable or block of variables); Reversible-Jump MH steps that are generalization of the Metropolis-Hastings steps when the attempted moves change the dimension [37]. These dimension changing steps are needed to update the Markov order of the background (variable v), the active covariates and their possible binarization (block of variables $t_{m,c}$, $g_{m,c}$, $h_{m,c}$), and the variables encoding the palindromic structure of the motif (block of variables p_m, c_m, q_m). Of note, Gibbs steps can be seen as a special case of Metropolis-Hastings steps where the proposal coincides with the target conditional

distribution. Similarly, under circumstances where the probability distribution of the variables whose dimension is modified can be integrated-out, the Reversible-Jump MH steps can be done in a Gibbs manner, i.e. direct sampling from the conditional distribution of the variable that defines the dimension. We proceeded this way for the joint update of $(t_{m,c}, g_{m,c}, h_{m,c})$ and for the joint update of (p_m, c_m, q_m). Correctness of the MCMC algorithm was carefully checked by implementing a successive-conditional simulator to reveal analytical and coding errors [38].

This MCMC algorithm, whose all mathematical details are provided in S1 Appendix, was implemented in C++ program called Multiple available at `https://github.com/eisultan/Multiple` and distributed under the GNU General Public License.

Data set for application to *Listeria monocytogenes*

Transcription start sites

To define promoter sequences and to align them with respect to TSSs, we used the repertoire of 2,299 TSSs mapped at 1 bp resolution on *L. monocytogenes* EGDe genome sequence [30]. Promoter sequences were defined as the 121 bp spanning from position -100 to +20 relative to each TSS, in keeping with the size of the regions that we previously found to be enriched for the presence of known transcription factor binding sites in *Staphylococcus aureus* [13]. The set of promoter sequences was further trimmed to avoid overlaps on a same strand which would be misinterpreted by the model as shared sequence features and thereby could cause false positive motif predictions. Since each TSS is accompanied by a read count reflecting its level of experimental support and transcriptional activity [30], the trimming procedure kept the TSS with the highest read count when choices were made to avoid overlaps of promoter sequences. In practice, a simple greedy procedure incorporated non-overlapping promoter regions one-by-one in the order of decreasing read-count which led to a set of 1,545 non-overlapping promoter regions (67% of the initial list of TSSs).

Expression data set

For the expression data, we relied on the compendium data-set established by aggregation of data sets published in many different studies to build the listeriomics website [31]. As downloaded, the data had dimensions 3,159×254, where each row corresponds to one gene and each column corresponds to the log2 of an expression ratio (log2 fold-change) comparing two samples (mutants, growth conditions, strains . . .) from a same study. All these studies used the genome and annotation of *L.monocytogenes* EGDe as reference. Some columns (pairs of conditions) and rows (genes) contained many missing values due the heterogeneity of technologies (one- and two-color arrays, RNA-Seq) used by the different studies. The number of columns was reduced from 254 to 165 after computing the median of the number of missing values per column and discarding the columns with a number of missing values higher than 1.5 times this median. In parallel, the number of rows was reduced from 3,159 to 2,825 based on the same criterion. Finally, the gene name associated with each TSS [30] permitted to match the 1,512 out of 1,545 non overlapping promoter sequences to one of these 2,825 genes with expression data. This resulted in an expression data-set of 1,512 promoter sequences with 165 expression values. The 165 expression columns represented 31 published studies, and included 17 RNASeq profiles and 25 tiling array profiles.

Building covariates for motif discovery

Both hierarchical clustering and projection methods were used as dimensionality reduction techniques to summarize the information of the expression data set (dimension 1,512×165). Two types of hierarchical clustering were applied corresponding to different options of hclust function provided by R stats package: Ward clustering and average-link based on Pearson distance ($1 - r$, where r is Pearson correlation coefficient). The distances needed to build the trees were computed after centering and scaling of the rows (genes) of symmetrized expression matrix obtained by duplicating each column (pairs of conditions) with a negative sign to remove the effect of the arbitrary orientation of the log2 fold-changes.

As projection methods, we applied PCA and ICA implemented in functions prcomp (package stats) and fastICA [39] of R without scaling the expression data. For PCA, we kept as covariates for motif discovery the 20 first components of the PCA (accounting for 68.8% of the total variance) after examining the rate of decrease of the residual variance. When applying ICA, the target dimension (number of components) is fixed before numerical optimization of the decomposition and the algorithm can converge to different projections that share only a subset of their components. In keeping with the idea developed by [26], we choose the target dimension K after examining the stability of the components and only stable components were used. Here, we used average-link clustering based on absolute Pearson correlation coefficient (r) between columns of the source matrix (dimension 1,512×K) and a cut-off $|r| \geq 0.8$ to compare components between runs. The target dimension was $K = 40$, the algorithm was run 100 times which lead to 26 components found in at least 80% of the runs that were kept as expression covariates for motif discovery.

The final set of 50 covariates that we used in our motif discovery analysis consisted of the 20 first PCA components (covariates numbered 1 to 20), the 26 stable ICA components (covariates numbered 21 to 46), and two hierarchical clustering trees obtained by Ward and average-link methods (covariates numbered 47 and 48). The two trees were further duplicated (covariates 49 and 50) to make it possible for the model to use two branches of the same tree. Expression covariates and sequence data are made available with the program.

Results

Exploration of the posterior landscape

Examination of the dynamics of evolution of the parameters describing motif components during MCMC runs confirmed that the algorithm was able to adjust simultaneously the characteristics of many PWMs. To illustrate the algorithm behavior, Fig 3A depicts the parallel evolution in a same MCMC of two PWMs in terms of width, nucleotide composition, and palindromic structure.

For *de novo* motif discovery, it was important to show that stability of the parameters of a motif component across tens of thousands of MCMC sweeps as seen in Fig 3A were not caused by slow-mixing but truly reflected attraction to peaks of the posterior density in the parameter space and should therefore be treated as relevant motif predictions. Since reaching similar motif components independently from different starting points is indicative of the second scenario, we performed 10 independent parallel runs of the MCMC algorithm. Each run consisted of 50,000 MCMC sweeps from a random starting point and the number of PWM components $\mathcal{M}$ was fixed to 75 ($\approx$ 4 weeks of CPU time). The two trajectories shown Fig 3A correspond to the first of these runs. The first 40,000 sweeps served as a burn-in period. The last 10,000 sweeps

were used in our analysis to characterize the posterior distributions after recording the values of the variables every 100 sweeps.

The 10 runs produced information on 750 (10×75) motif components that were named from M1.0 (random seed 1, motif 0) to M10.74 (random seed 10, motif 74). We analyzed and compared these motif components to extract distinct well supported motifs that were not only stable across the last 10,000 sweeps but also found in at least two runs. To declare that two motif components correspond in fact the same motif, we compared the set of positions in the sequences that were predicted to be covered by the occurrences of each motif, which summarizes the information of the different ingredients of the motif model (nucleotide composition in PWM, preferential position with respect to TSS, link with expression data). Namely, we considered the positions with estimated posterior probability of coverage greater than 0.5 and the pairwise distance between two motif components i and j was computed as

$$d(i,j) \quad = \quad 1 - \frac{2O(i,j)}{N(i)+N(j)}\,, \tag{7}$$

where $N(i)$ and $N(j)$ denote the total number of positions covered by each motif and $O(i,j)$ is the number of positions covered by both motifs. With this formula, $d(i,j) = 0$ if the positions covered are exactly the same and $d(i,j) = 1$ if they do not overlap. A special case of no overlap corresponds to $N(i) = 0$ which concerned 190 of the 750 motifs and typically reflected instability during the last 10,000 sweeps. The minimal distance that we observed between two motif components within the same run was 0.59 and 92.7% of the motif components where at distance ≥ 0.9 of any other motif components from the same run. This shows that allowing motif occurrences to overlap in the model is an effective strategy to enforce the convergence of motif components to distinct motifs.

A comprehensive set of distinct well supported motifs were identified by average-link hierarchical clustering of the 750 motifs based on the distance defined in Eq (7). This tree whose a section is shown in Fig 3B was cut at two different heights: 0.75 (i.e. 25% of sequence positions covered by occurrences are shared between the two motifs) to define high-level clusters that separate well-distinct motifs; 0.25 to define low-level cluster containing very similar motifs (i.e. 75% of the sequence positions are shared). High-level clusters are represented here by different colors. One representative motif was selected for each high-level cluster that contained at least two motifs at distance less 0.25 (i.e. at least one low-level cluster). This representative was selected as the member of the low-level clusters with the highest $N(i)$ (see Eq (7); when no low-level cluster existed the high-level cluster was not represented in the final list of motifs. This procedure defined the set of 40 representative motifs reported in Table 1.

Fig 4 shows several examples of estimated links between expression covariates and motif occurrences, including covariates that were binarized or not and trees. The average number of expression covariates used (i.e. with $t_{m,c} = 1$) to describe the patterns of presence/absence of each of the 40 motifs in the set of sequences ranged from 1.02 (for M61.5) to 6.45 (for M71.2). Distinguishing the two types of covariates, the range was 0.28 to 4.94 for the covariates of type "vector" and 0.60 to 2.65 for the covariates of type "tree". Fig 4ABC show three examples (M33.8, M20.1, and M68.1) of strong associations defined as those where a covariate has an estimated posterior probability of being used greater than 0.9. As expected given that we did not try to avoid redundancy between the covariates, the strong association with a specific covariate did not cover all the cases of clear association between the presence/absence of a motif and the expression covariates. For instance, none of the covariates fulfilled this condition for M29.8 (SigB -10 box, see below) whereas the average number of expression covariates used was as high as 4.38 and a clear association with expression data is

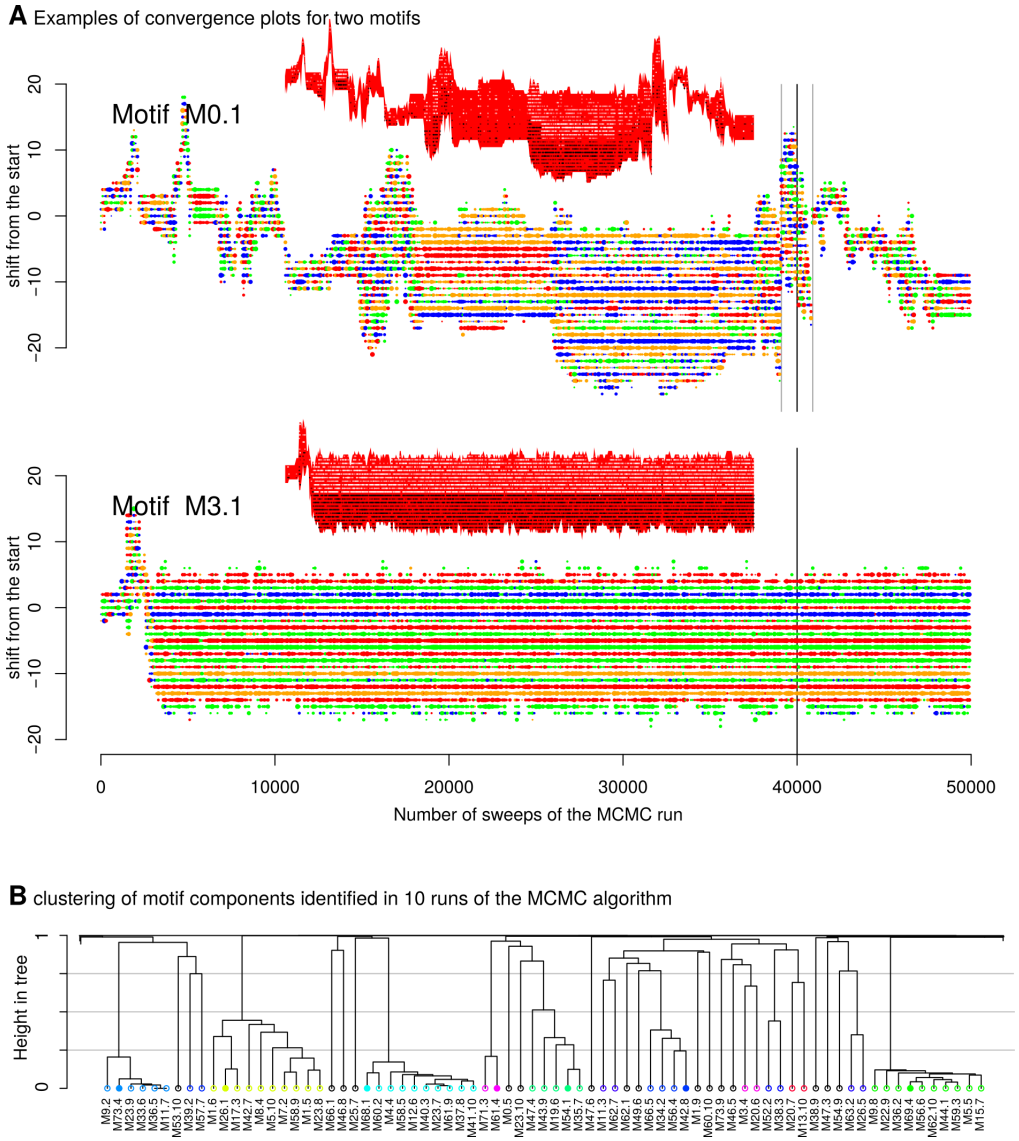

Fig 3. Exploration of the posterior landscape. A Convergence plots for two motifs across the 50,000 sweeps of the algorithm. Colored dot points correspond to PWM columns, whose color reflects the most likely nucleotides (green for `A`, blue for `C`, orange for `G`, red for `T`) and diameter the information content of the nucleotide frequency. Increase or decrease of the width of the PWM on the 5'- and 3'-sides are represented by change in the number of colored dots (on lower and upper side of the colored area, respectively). A black vertical line is drawn at sweep number 40,000 that delimited the end of our burn-in period. Grey vertical lines indicate sweeps at which the y-coordinate was re-centered for the purpose of the representation. Evolution of the palindromic structure is represented in insert plots (at lower scale) where the evolution of PWM width is represented by a red area and paired columns are indicated by black and white dots. **B** Hierarchical clustering of motif components to extract stable motifs. Only a fraction of the tree is represented here (75 motifs out of 750). Motifs belonging to the final list of 40 selected representative stable motifs are indicated by closed circles (open circles are used for non selected motifs).

captured by the extended probit model (see Fig 4DEF). A total of 9 motifs were nevertheless strongly associated with one covariate of type vector and remarkably all these strong associations concerned covariates defined by ICA (none defined by PCA).

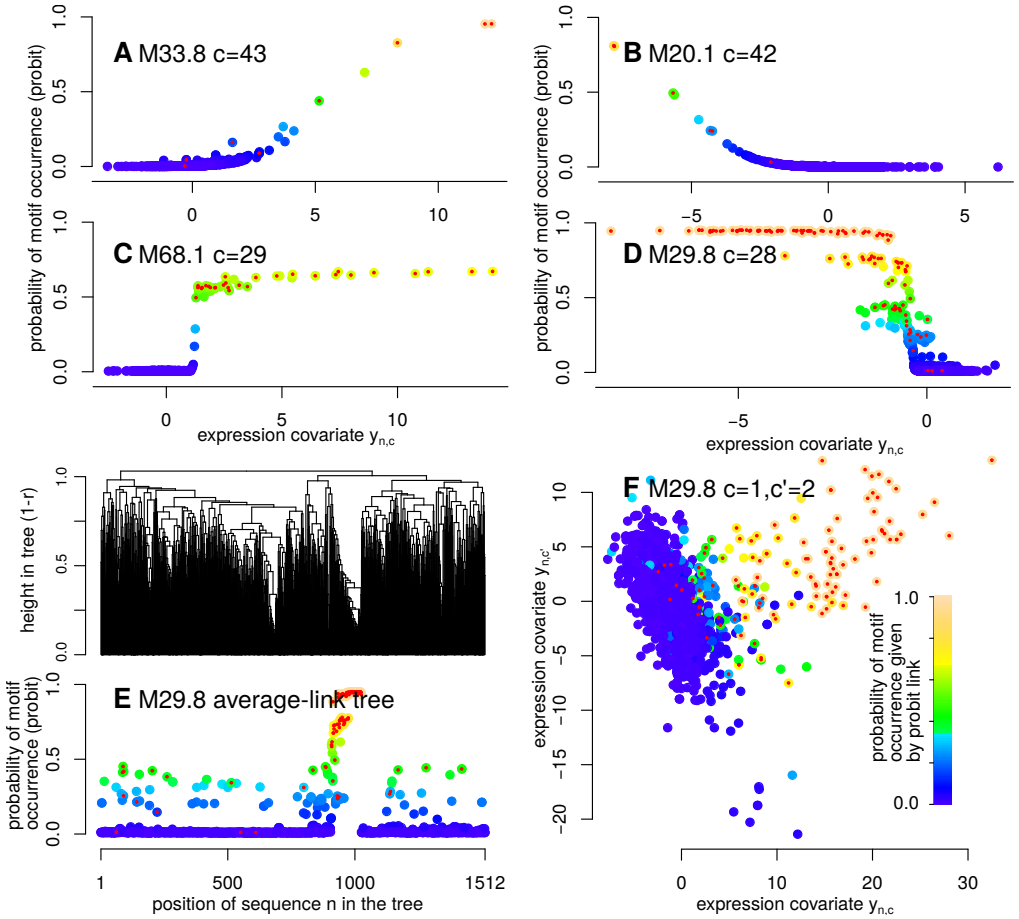

Fig 4. Estimated links between expression covariates and motif occurrences. Examples are shown for four different motifs in subplots **A**, **B**, **C**, and **DEF**, respectively. Color gradient from blue to yellow (see color scale in subplot **F**) reflects the estimated probability of motif occurrence given by the extended probit model (i.e. summarizing the information from the expression data). Red dots indicate sequences in which the motif was predicted to occur after taking into account the nucleotide sequence information (estimated posterior probability of motif occurrence was ≥ 0.5). **A** and **B** correspond to links established without binarizing the covariate ($g_{m,c} = 0$) while **C** and **D** correspond to cases where the covariate was automatically binarized ($g_{m,c} = 1$). The juxtaposition of **D**, **E** and **F** illustrates a degree of redundancy between the covariates: for this motif (M29.8), several covariates are able to separate the sequences for which the probit model gives a high probability of motif occurrence (yellow points).

From validation by comparison with known motifs and regulons to new insights on *L. monocytogenes* transcription network

The 40 distinct well supported motifs exhibit considerable diversity in terms of abundance, preferred position with respect to the TSS, link with expression data, and PWM characteristics (width, information content, palindromic vs. non-palindromic

structure). Table 1 summarizes the main characteristics of each motif. Fig 5 provides a graphical representation for two of these motifs that includes a graphical representation of the PWM as a sequence logo, the estimated probability density function for the position of the motif occurrences relative to the TSS, and the average log2 fold-change across expression conditions. Similar figures are available for all motifs in S1 Fig and the numerical values of the PWM parameters are found in S1 File. The list of genes associated with each motif and the exact position of occurrences are reported in S2 File and S3 File.

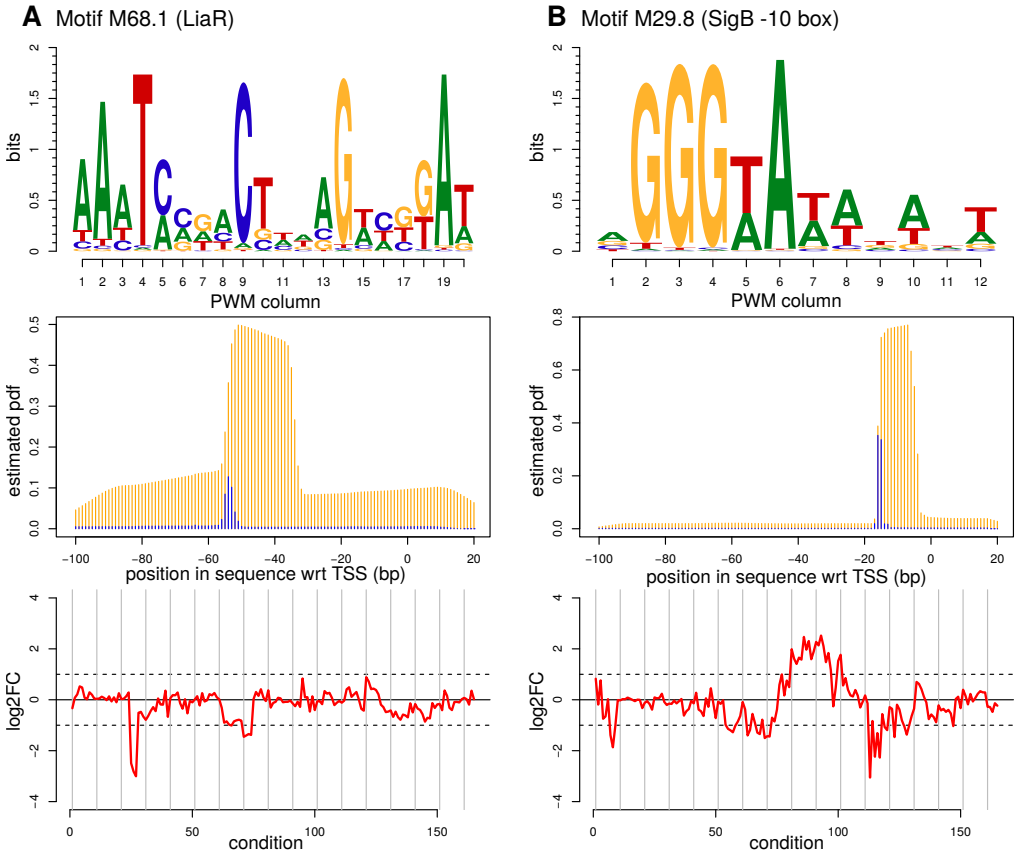

Fig 5. Two examples motifs for known transcription factors rediscovered by our algorithm. These motifs were identified as corresponding to the binding sites of LiaR (M68.1 subplot **A**) and as the -10 box of the SigB binding sites (M29.8 subplot **B**). First row: sequence logo where the relative heights of the letters in one column represent the nucleotide composition and the total height represent the information content (in bits). Second row: estimated probability distribution function for the 5'-end of the occurrence (in blue) and probability of having the position covered by the occurrence (in orange). These probabilities are conditional on the presence of the motif in the promoter sequences; position is reported as the distance to the TSS (negative when upstream of TSS). Third row: average log2 fold-change of expression level for downstream genes across the 165 pairs of conditions that were used to define the expression covariates.

Three complementary approaches were used to identify links between the 40 motifs discovered by our *de novo* approach and known regulons. The first was the comparison with lists of genes collected from tables published in several expression studies and positions of transcription factor binding sites recorded in the RegPrecise database [40];

the second approach was the systematic comparison with 188 reference PWMs derived from sequence alignments extracted from the propagated RegPrecise database accessed in July, 2018 for different taxonomic groups in the Firmicutes phylum: *Listeriaceae* (25 PWMs) and *Staphylococcaceae* (39 PWMs) and *Bacillales* (124 PWMs). The third approach consisted in dedicated literature searches associated with a careful manual examination of (i) the set of genes downstream the promoters in which the motif was predicted to occur (ii) the comparisons of conditions in which the log2 fold-change deviated the most from 0 (iii) the characteristics of the PWM and the preferred position of motif occurrences with respect to the TSS. For many of the identified transcription factors, these lines of observation provided convergent clues.

Links to known regulons identified with this combination of approaches are reported in the rightmost column of Table 1. Most of the motifs with high number of occurrences were found to describe general characteristics of promoter regions (variations on the themes of SigA -10 and -35 boxes, nucleotide composition around TSS, Ribosome Binding Site). Systematic comparison with RegPrecise PWMs was particularly informative for the identification of BglR2, CcpA, CcpB, Fur, LexA, LiaR, and Rex. Among the other identified transcription factors, SigB and SigL were identified based on position with respect to the TSS, comparison with sigma factor consensus defined for *B. subtilis* [12], as well as overlap with previous experimental data on SigL (also known as RpoN or sigma54) and SigB regulons in *L. monocytogenes* [41–44]. S2 Fig provides a graphical representation of the prediction of the SigB regulon from the joint occurrence of boxes -10 and -35 and of the comparison with previous studies. In brief, 89 promoter regions are predicted to contain both the -10 and -35 boxes, 88 out of them were previously reported as probable members of the SigB regulon in either [43] or [44]. Spx was identified based on (i) its position with respect to the TSS just upstream of the SigA -35 box, (ii) sequence properties reported for Spx in *B. subtilis* described as an AGCA element at position -44 [45, 46], and (iii) literature data on the Spx-regulon of *L. monocytogenes* [47]. PWMs found for PrfA and VirR, two key transcription regulator involved in *L. monocytogenes* virulence, were in line with previously described sequence properties [48–50]. Taken together, correspondences with known motifs validate our approach for *de novo* discovery and suggest that other motifs listed in Table 1 may also correspond to biologically relevant motifs.

Among the predicted regulons that we have not been able to connect to literature data, the most spectacular by its size and the functional homogeneity of the regulated genes (genes encoding the translation apparatus, including ribosomal proteins), is associated with motif M58.7. Importantly, even when motifs could be linked to previously known transcription factors, being able to identify them in a *de novo* and automated manner can shed a new light on the corresponding regulons. For instance, our results suggest that LiaR regulon may be about twice larger than previously identified by differential expression analysis of the *liaS*-deletion mutant vs. wild-type [51]. Ordered by the number of TSSs as in Table 1, the predicted regulons for which we have identified a transcription factor are: SigB, CcpA, Rex, LiaR, Fur, LexA, VirR, Spx, BglR2, SigL, PrfA. The discussion section points, for some of these transcription factors, to new biological observations made from these predictions.

Detailed information on the predicted regulons associated to each of the 40 motifs can be found in S2 File in which transcriptional units defined by [30] were used to propagate information of motif occurrence to downstream genes. Pearson correlation computed on log2 expression ratios aggregated by [31] and operons delineated by [29] are also reported to assess the level of support for propagated information. S3 File reports the most likely positions of occurrences for the 40 motifs in the 1,512 promoter regions (121 bp around the selected TSSs). We noted for instance from the positions of occurrences reported in this table that occurrences of M20.1 and M63.1 occur

simultaneously in a longer highly conserved sequence element. Position information is provided for 17,065 possible occurrences corresponding to those for which the posterior probability of occurrence in the promoter sequence was estimated to be above 0.1.

Contributions of different sources of information to motif discovery

To examine the contributions of different ingredients of our model to the results obtained on the *L. monocytogenes* data set, we compared the results obtained with the full model with those obtained with 5 other models that differed with respect to three ingredients that they incorporate (position with respect to TSS, expression profiles, and palindromic structures). Specifically, we tested 5 out of the 7 possible combinations that discards one or more of these three model ingredients: we considered the effect of removing a single ingredient (3 models), and also considered the 2 models that discard both the information of preferential position with respect to the TSS and expression profiles. Of note, even the simplest model still allow motif overlap and automatic adjustment of motif width which are key ingredients of the approach developed here. As nomenclature, we used $\mathtt{t}^1\mathtt{e}^1\mathtt{p}^1$ for the full model and $\mathtt{t}^0\mathtt{e}^0\mathtt{p}^0$ for the simplest model where $\mathtt{t}$ stands for TSS, $\mathtt{e}$ for expression, and $\mathtt{p}$ for palindromic.

The results of the different runs were post-processed independently to identify the set of stable motifs in each setting. A total of 208 stable motifs were identified (from 40 for $\mathtt{t}^1\mathtt{e}^1\mathtt{p}^1$ to 28 for $\mathtt{t}^0\mathtt{e}^0\mathtt{p}^0$). We used Eq (7) to build a distance matrix between the motifs of the different settings. Table 2 provides a global view of the 40 motifs obtained with the full setting, used as reference, and reports their closest matches that were obtained with each of the other settings. Using a cut-off on the distance at 0.25, the number of matches for the 40 motifs in the different settings varied from 18 for $\mathtt{t}^1\mathtt{e}^1\mathtt{p}^0$ to 6 for $\mathtt{t}^0\mathtt{e}^0\mathtt{p}^1$ and $\mathtt{t}^0\mathtt{e}^0\mathtt{p}^0$. These numbers show the strong impact of the different model ingredients on the definition of the motifs and prediction of the targets. It is also interesting to note that among the three ingredients tested, taking into account the palindromic structure had the least overall impact on the results.

However, as expected given the diversity of the motifs, all the motifs were not sensitive to the same model ingredients. For instance, only 1 motif had a match at distance ≤ 0.25 in the 5 model settings tested (M62.3, Fur) but 25 motifs had at least 1 match in 1 of these 5 settings. Somewhat unsurprisingly, motifs sensitive to modeling the exact position with respect to the TSS tended to have a more constrained position of occurrence in the sequence (see column "Pos" in Table 2; e.g. SigB, BglR2, Spx), motifs sensitive to taking into expression data tended to be associated with marked variations of expression across conditions (see column "FC"; e.g. Sig B, LiaR, SigL, PrfA), and motifs whose detection was sensitive to the modeling of the palindromic structure were palindromic motifs (see column "Pal"; e.g. Rex, CcpB). Importantly, the discovery or prediction of the motif occurrences of many known (and therefore biologically relevant) motifs were sensitive to several ingredient of the model which attests of the interest of an integrative statistical model.

Comparison with other methods

We also compared our results to those of different algorithms able to handle the data-set of $1,512$ sequences and $1,512 \times 165$ expression matrix for *de novo* motif discovery. These algorithms are based on PWM estimation from the sole sequence data-set (MEME, [8]) or search specifically for motifs that are connected to the expression data using motif representations that can be either consensus string written in IUPAC degenerated symbols (FIRE and RED2, [7, 14]) or PWMs (MatrixREDUCE, [16]). Input data, parameter settings, and results are detailed in S2 Appendix.

Motif discovery without auxiliary data using MEME (see S2 Appendix) returned up to 8 motifs with E-value less than at 1. These motifs included motifs describing general properties of the *L. monocytogenes* promoter regions, such as SigA -10 box, the presence of a RBS and of T-rich elements (our M59.9), as well as more specific motifs that are relatively abundant and/or of high information content such as M70.6, CcpA, and Fur. A variant of the sequence element described by M20.1 and M61.3 is also found (high information content, see column IC1 in Table 1).

The three algorithms (MatrixREDUCE, FIRE, and RED2) searching for motifs related to expression data returned only short motifs (lengths up to 9) due to the use of k-mers as seeds in the initial steps of the searches (see S2 Appendix). With some parameter settings, they were all able to retrieve SigB -10 box (M29.8) which presents the particularity of being abundant and to exhibit a strong link to expression profile. It is also short and contains a stretch of adjacent high information content PWM columns which makes it ideally suited for approaches based on k-mers (Fig 5B). MatrixREDUCE also discovered a short motif that corresponds to the central part of PrfA. This motif presents the particularity of exhibiting, by far, the strongest link with expression data among all the motifs that we discovered (see column FC in Table 1). Thus, comparison of our results with those of these four algorithms illustrates the interest of the statistical model described in this work that allows to combine the benefits of the purely sequence-based approaches, like MEME, and the benefits of algorithms that can only find motifs linked to expression data like MatrixREDUCE, FIRE and RED2.

Discussion

The model and associated trans-dimensional MCMC algorithm provides an integrated framework regulatory motif discovery in bacteria with the help of transcriptome data (exact positions of TSSs and expression profiles). The algirthm adjusts, simultaneously for many motifs and many transcription covariates: the width and palindromic properties of the corresponding position weight matrices, the number of parameters to describe position with respect to the transcription start site, and the selection of the expression covariates that are relevant. Application to *L. monocytogenes* permitted to validate the approach and detailed predictions are made available for this important pathogen. The program Multiple is freely distributed under a GPL license which gives full access to the source code and make it possible to analyze other data-sets from other bacteria.

Allowing motifs to overlap is the cornerstone of the model, since it makes it possible to model occurrence of the different motifs in the same sequence independently from each other. This simplifies not only the model but also MCMC updates and it facilitates the mixing of the algorithm. Importantly, we show here that this modeling prevents different PWM components of the model to converge to the same motif. In this aspect, explicit modeling of motif occurrences overlap appears as a proper statistical framework to implement what [52] named "repulsion" and for which they proposed a heuristic that consists in incorporating *ad-hoc* repulsive forces between parallel MCMC runs. Binding sites do overlap in bacterial genomes [53] and our choice of modeling the contribution of different motifs that overlap as an average of the nucleotide emission probability density functions (PWM columns) is probably not the most biologically realistic. However, it has the advantage of being the simplest choice. Indeed, we also considered a more complicated model in which PWM columns corresponding to motifs that overlap contribute as a function of their information content. This more complex model satisfies the intuition that if two motifs overlap and one motif has a strong preference for a nucleotide at a particular position whereas the other has no or little preference, the motif with the strong preference tends to "impose" its choice. In S1 Appendix, we refer

to this model as the θ-dependent weight mixture model for motif overlaps and describe how it can be implemented. An important drawback of this more sophisticated model is to complicate the dependence structure described in the DAG (Fig 1) by adding an arrow from the PWM parameters θ to the disambiguation variable b. This dependence makes that θ can no longer be marginalized out or updated by direct drawing from its conditional distribution (i.e. in Gibbs manner). By turn, the efficient update that we implemented for the palindromic structure does not work. Thus, our code does not accommodate coexistence of the palindromic structure model and θ-dependent mixture model for motif overlap and we have decided here to focus only on the results obtained with the simplest model of motif overlap. Other extensions of the model that could be envisioned in the future include accounting for multiple occurrences of the same motif in each promoter region and for overlapping promoter regions.

Our results show that the choice of incorporating the expression data y as a covariate that helps to model the sequence data x preserves the benefits of the pure sequence based approach grounded on well-established statistical models of DNA sequences, such as implemented in MEME [8]. We also showed that it effectively allows to make use of expression data by automatically selecting the features in expression space that are relevant for each motif on the basis that they improve the probabilistic representation of the sequence (likelihood). This use of external data can be connected to earlier works that have incorporated external data in sequence models for motif discovery via the definition of informative priors in order to favor motif whose occurrences are found in regions of the genome that are more likely to contain binding sites, for instance because they are conserved between species or are depleted in nucleosomes [54–57]. In our work, the relevant covariates that describe how the probability of occurrence of a motif differs between promoters are automatically selected toghether with the coefficients that specifies their contributions and this is achieved in a simultaneous manner for many motifs. We also show how we can account for covariates with complex structures such as the position in a tree as covariate. The idea of using a tree whose topology and branch lengths reflect similarities between activity profiles is to provide an alternative to the use of a predefined set of co-expression clusters. It is also found in the previous model and algorithm that we specifically developed for *de novo* discovery of sigma factor binding sites [12]. However, the probabilistic models are completely different. In [12], the occurrences of the different possible motifs are modeled as resulting from an "evolution" process along the branches of the tree which makes that the sequences that are close in the tree (hence in the expression space) are more likely to harbor binding sites for the same sigma factor. The motivation for introducing here the use of a probit model for modeling $x|y$ was to handle more complex representations of the expression space than a single tree. With our extended probit model, several concurrent description of the expression spaces can coexist (trees, continuous vectors) and the algorithm selects the most relevant, with the possibility to binarize continuous vectors according to automatically adjusted cut-off values.

A limitation of our approach is that each run takes several weeks on a data-set like the one studied in this study. Most of the time is spent in the update of a (the positions of the motif occurrences) and thus the time complexity of a MCMC sweep is approximately proportional to the cumulated length of the sequences times the cumulated width of the PWMs plus a small term roughly in the square of the cumulated width for taking into account motif occurrence overlaps. Furthermore, the analysis that we conducted is based on several runs that serve to unambiguously identify peaks in the posterior landscape. These peaks are identified by the convergence of several runs to the same neighborhood in a space of very large dimension. Here, we limited our analysis to 10 runs that were conducted in parallel. These 10 runs allowed to identify 40 stable motifs but it would not be very surprising to find several other motifs by adding more

runs, since some of the biologically known motifs have only be found by 2 or 3 runs (Rex, Spx, CcpB). Future studies on other data-sets may decide to include more runs.

As briefly mentioned in the introduction, *L. monocytogenes* was well-suited for a proof-of-principle application of the method since it combines good experimental data sets (rich expression data [31], repertoire of TSSs at single base resolution [30]) with a body of established biological knowledge available in the literature on its regulatory network that contains features shared with the related Gram-positive model bacterium *B. subtilis* and features that are specific, such as those involved in pathogenicity. Comparison of our list of 40 motifs with literature data validated the method by proving its ability to re-discover, in a pure *de novo* manner, regulons of many known transcription factors. In essence, the method is based on the search for "over-represented" motifs whose modeling significantly improves the probabilistic representation of the DNA sequence (as measured in the likelihood). It is thus only adapted to identify the regulons of transcription factors playing the coordinating roles of regulating the expression of several to many transcription units. These so-called global transcription factors are opposed to local transcription factors that account the vast majority of transcription factors in bacteria but regulate only one or a very few target in a specific biological pathway [58, 59]. Global transcription factors have very different roles. Many have been subjected to dedicated transcriptomics experiments in *L. monocytogenes*, with roles ranging from control of large set of genes involved in the general stress response by SigB [43, 44] to response to specific stresses such as: the so-called SOS response to DNA damage mediated by LexA [60], response to iron limitation stress mediated by Fur [61], and response to cell envelope stress mediated by LiaR [51]. In the case of a pathogen like *L. monocytogenes*, a special attention has also been paid to regulators involved in the coordination of the virulence function such as PrfA and VirR [49, 50, 62].

Our results provide *de novo* predictions made in a unified framework based on the presence of motif occurrences, taking into account experimentally determined TSS position and expression data for the regulons of these transcription factors that have been subjected to analyses by transcriptomics. This is interesting since contributions of direct and indirect regulations have not always been fully disentangled in literature by thorough identification of transcription factors binding sites. The degree of novelty of our predictions vary depending of the concordance with previous knowledge, from very high concordance for SigB or PrfA, were virtually no additional targets are identified, to a significant increase in the predicted number of targets for Fur (14 occurrences of Fur-box, listed in [61]) or LiaR (17 transcription units whose expression is modulated by LiaSR two-component systems, listed in [51]). It should, however, be noted that we cannot formally exclude the possibility that some of our predicted regulons may aggregate several transcription factors with related binding sites. For instance, we note a consensus binding site close to Fur for NanR which regulates sialic acid catabolism in *S. aureus* [63]. New predicted targets whose annotated functions correlate well with known functions of the regulators include: for Fur regulon, *lmo0484*, annotated as the gene of heme-degrading monooxygenase IsdG; and for LexA regulon, *lmo0233*, annotated as encoding DNA repair protein RadA, *lmo1514*, annotated as recombination factor protein RarA, and *lmo0476*, annotated as encoding oxetanocin A resistance protein oxrB. We also noted the presence of *lmo1948* and *lmo1947*, encoding the two-component system ResDE, in our prediction of SigL regulon. This prediction adds a piece of information to why the inactivation of ResDE, known in other bacteria such as *B. subtilis* to be involved in the regulation of respiration, was reported to reduce the expression of mptA, a gene involved in the uptake of mannose, and to lead to impaired growth on selected carbohydrates [64]. Indeed, these effects are fully consistent with the functions of other genes in *L. monocytogenes* SigL regulon [41, 43].

CodY is a previously studied transcription factor which plays an important role in the regulation of the metabolism of Gram-positive bacteria and was found to also co-regulate motility and virulence pathways in *L. monocytogenes* [65]. We were initially puzzled to have missed CodY in our predictions but this is not surprising since CodY binds most often downstream the promoter regions analyzed here (ending 20 bp after the TSS) [66].

Our results also contain the first published global *de novo* predictions for the regulons of several transcriptions factors whose importance is suggested by knowledge in other bacteria such as *B. subtilis*, but which have not yet been experimentally studied in *L. monocytogenes*. We find in this category CcpA, Rex, and Spx. CcpA, the Carbon catabolite control protein A, has been deeply studied in *B. subtilis*, where it stands as a major regulator controlling about 100 operons (see SubtiWiki for details, [67]). CcpA is also present in *Listeria*, but has not been really studied, probably because it was early reported to play no role in carbon source regulation of virulence genes [68]. With 87 promoter regions identified, CcpA is second after SigB in terms of number of binding sites identified (Table 1). The number of targets of CcpA in *L. monocytogenes* is thus close to *B. subtilis* and, even if CcpA is not directly involved in virulence, it plays certainly an important role, possibly during the life of the bacterium in the soil. Rex has not been studied yet in *L. monocytogenes* but was suspected to be, like in *B. subtilis*, a direct sensor of NAD+/NADH imbalance caused by Redox stresses that regulates carbohydrate and energy metabolism [69]. The size of the regulon predicted here for Rex (39 sites, third regulon in terms of number of binding sites after SigB and CcpA) is striking compared to the 7 targets known in *B. subtilis* (see SubtiWiki for details) although, again, we cannot formally exclude aggregation with targets of other transcription factors (for instance, Fnr of *B. subtilis* and ArcR of *S. aureus* have binding sites very close to those of Rex). Among the predicted targets genes, we find for instance *lmo1586*, encoding an inorganic polyphosphate-ATP-NAD kinase, which is regulated by the non coding RNA RoxS in *B. subtilis* [70]. Finally, Spx is another key regulator of the response to Redox stresses that detects disulfide stress caused by the oxidation of cysteine residues whose regulon has not yet been studied globally in *L. monocytogenes* [47,69]. With 19 predicted targets, the size of the predicted regulon in *L. monocytogenes* is similar to *B. subtilis*. Among the predicted targets not identified in *B. subtilis* we find *lmo2072* encoding Rex suggesting an intricated Redox stress response regulation. For our method, we considered the detection of the Spx binding sites as an important achievement since its sequence signature consists of a very short motif with a constrained position of occurrence directly upstream of SigA -35 box that has remained elusive in *B. subtilis* until the use of dedicated ChIP experiments and regression analyses [45].

One or the motivation of our work was to open a door for the discovery of new important regulons. A very interesting result with respect to this prospect, is the identification of a partially palindromic motif whose central PWM columns corresponds to palindromic consensus `ACGTAYYCGT` (M58.7 in Table 1 and S1 Fig). We predicted for this motif at least 27 occurrences that are very consistently found upstream of genes encoding the translation apparatus. These targets include 23 promoter regions upstream of genes encoding ribosomal proteins but are also found upstream *lmo2610*, *lmo1785*, and *lmo2654*, encoding respectively translation initiation factors IF-1 and IF-2, and elongation factor EF-G (see S2 File). In the Gram-positive model bacterium *Bacillus subtilis*, as well as in *Escherichia coli*, growth rate directly reflects the availability of resources in the environment [71]. In these cells, a tight coupling between growth and available resources is ensured by a sophisticated regulatory network whose main actors are a GTP pyrophosphokinase (RelA/SpoT) and two key metabolites (ppGpp and GTP) that jointly implement a mechanism known as the stringent

response [71, 72]. In both bacteria, the goal of this regulatory mechanism is apparently to decrease (or even stop) the production of the translation apparatus components when the resources in the medium become too scarce. At molecular level, modification of the transcription of the translation apparatus is mediated by the level of ppGpp in *E. coli* [73] or GTP in *B. subtilis* [74] and probably also *L. monocytogenes* which also possesses RelA [75, 76]. Univocal coupling between available resources and growth rate is not necessarily the most appropriate in all circumstances. For instance, it is possible that staying calm is a better strategy against host defense at certain stages of infection, even if the conditions inside the host provides enough nutrients for the growth. Existence of dedicated regulatory mechanisms allowing to keep control of the growth rate, even in the presence of nutrients, may thus not seem unexpected. We hypothesize that the new regulon that we detected in *L. monocytogenes* may be an instance of such a mechanism whose biological role and regulatory molecules remain to be identified.

Supporting information

S1 Fig. Summary plots for the 40 stable motifs.

S2 Fig. Prediction of SigB regulon and comparison with previous studies.

S1 File. PWMs for the 40 stable motifs. Table in XLSX format.

S2 File. Detailed information on the predicted regulons associated to each of the 40 stable motifs. Table in XLSX format.

S3 File. Most likely positions for the occurrences for the 40 motifs in the 1,512 promoter regions (121 bp around the selected TSSs). Table in XLSX format.

S1 Appendix. Detailed description of the statistical model and the associated MCMC algorithm.

S2 Appendix. Comparison with results of some other algorithms for *de novo* motif discovery.

Acknowledgments

We wish to thank: Birgitte H. Kallipolitis, Pascal Piveteau and the List_Maps consortium for helpful discussions; the INRA MIGALE bioinformatics platform (http://migale.jouy.inra.fr) for providing computational resources.

Funding

Ibrahim Sultan's PhD thesis was funded by the European Union's Horizon 2020 research and innovation program under Marie Skłodowska-Curie, grant agreement number 641984 (project "List_Maps").

References

1. Korn LJ, Queen CL, Wegman MN. Computer analysis of nucleic acid regulatory sequences. Proc Natl Acad Sci U S A. 1977;74(10):4401–4405.
2. Das MK, Dai HK; BioMed Central. A survey of DNA motif finding algorithms. BMC Bioinformatics. 2007;8(7):S21.
3. Sandve GK, Drabløs F. A survey of motif discovery methods in an integrated framework. Biology Direct. 2006;1(1):11.
4. Zambelli F, Pesole G, Pavesi G. Motif discovery and transcription factor binding sites before and after the next-generation sequencing era. Briefings in Bioinformatics. 2012;14(2):225–237.
5. Schbath S. An overview on the distribution of word counts in Markov chains. Journal of Computational Biology. 2000;7(1-2):193–201.
6. Young JA, Johnson JR, Benner C, Yan SF, Chen K, Le Roch KG, et al. In silico discovery of transcription regulatory elements in *Plasmodium falciparum*. BMC Genomics. 2008;9(1):70.
7. Lajoie M, Gascuel O, Lefort V, Bréhélin L. Computational discovery of regulatory elements in a continuous expression space. Genome Biology. 2012;13(11):R109.
8. Bailey TL, Elkan C. Unsupervised learning of multiple motifs in biopolymers using expectation maximization. Machine Learning. 1995;21(1-2):51–80.
9. Neuwald AF, Liu JS, Lawrence CE. Gibbs motif sampling: detection of bacterial outer membrane protein repeats. Protein Science. 1995;4(8):1618–1632.
10. Mao L, Mackenzie C, Roh JH, Eraso JM, Kaplan S, Resat H. Combining microarray and genomic data to predict DNA binding motifs. Microbiology. 2005;151:3197–3213.
11. Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, et al. Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. PLoS Biol. 2007;5(1):e8.
12. Nicolas P, Mäder U, Dervyn E, Rochat T, Leduc A, Pigeonneau N, et al. Condition-dependent transcriptome reveals high-level regulatory architecture in *Bacillus subtilis*. Science. 2012;335(6072):1103–1106.
13. Mäder U, Nicolas P, Depke M, Pané-Farré J, Debarbouille M, van der Kooi-Pol MM, et al. *Staphylococcus aureus* Transcriptome Architecture: From Laboratory to Infection-Mimicking Conditions. PLoS Genet. 2016;12(4):e1005962.
14. Elemento O, Slonim N, Tavazoie S. A universal framework for regulatory element discovery across all genomes and data types. Molecular Cell. 2007;28(2):337–350.
15. Bussemaker HJ, Li H, Siggia ED. Regulatory element detection using correlation with expression. Nature Genetics. 2001;27(2):167.
16. Foat BC, Morozov AV, Bussemaker HJ. Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. Bioinformatics. 2006;22(14):e141–e149.
17. Gruber TM, Gross CA. Multiple sigma subunits and the partitioning of bacterial transcription space. Annu Rev Microbiol. 2003;57:441–466.

18. Paget MSB, Helmann JD. The sigma70 family of sigma factors. Genome Biol. 2003;4(1):203.

19. Wurtzel O, Sapra R, Chen F, Zhu Y, Simmons BA, Sorek R. A single-base resolution map of an archaeal transcriptome. Genome Research. 2010;20(1):133–141.

20. Irnov I, Sharma CM, Vogel J, Winkler WC. Identification of regulatory RNAs in *Bacillus subtilis*. Nucleic Acids Research. 2010;38(19):6637–6651.

21. Ettwiller L, Buswell J, Yigit E, Schildkraut I. A novel enrichment strategy reveals unprecedented number of novel transcription start sites at single base resolution in a model prokaryote and the gut microbiome. BMC Genomics. 2016;17(1):199.

22. Yan B, Boitano M, Clark TA, Ettwiller L. SMRT-Cappable-seq reveals complex operon variants in bacteria. Nature Communications. 2018;9:3676.

23. Nicolas P, Tocquet AS, Miele V, Muri F. A reversible jump Markov chain Monte Carlo algorithm for bacterial promoter motifs discovery. J Comput Biol. 2006;13(3):651–667.

24. Alter O, Brown PO, Botstein D. Singular value decomposition for genome-wide expression data processing and modeling. Proc Natl Acad Sci U S A. 2000;97(18):10101–10106.

25. Liebermeister W. Linear modes of gene expression determined by independent component analysis. Bioinformatics. 2002;18(1):51–60.

26. Kairov U, Cantini L, Greco A, Molkenov A, Czerwinska U, Barillot E, et al. Determining the optimal number of independent components for reproducible transcriptomic data analysis. BMC Genomics. 2017;18(1):712.

27. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A. 1998;95(25):14863–14868.

28. Sorek R, Cossart P. Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity. Nature Reviews Genetics. 2010;11(1):9.

29. Toledo-Arana A, Dussurget O, Nikitas G, Sesto N, Guet-Revillet H, Balestrino D, et al. The *Listeria* transcriptional landscape from saprophytism to virulence. Nature. 2009;459(7249):950.

30. Wurtzel O, Sesto N, Mellin JR, Karunker I, Edelheit S, Bécavin C, et al. Comparative transcriptomics of pathogenic and non-pathogenic *Listeria* species. Molecular Systems Biology. 2012;8(1):583.

31. Bécavin C, Koutero M, Tchitchek N, Cerutti F, Lechat P, Maillet N, et al. Listeriomics: an interactive web platform for systems biology of *Listeria*. MSystems. 2017;2(2):e00186–16.

32. Albert JH, Chib S. Bayesian analysis of binary and polychotomous response data. Journal of the American Statistical Association. 1993;88(422):669–679.

33. Robert CP, Casella G. Monte Carlo statistical methods; 2nd ed. Springer texts in statistics. Berlin, Heidelberg: Springer-Verlag; 2004.

34. Robert CP. The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation; 2nd ed. Springer Texts in Statistics. Springer New York; 2007.

35. Liu JS. Monte Carlo Strategies in Scientific Computing. Springer series in statistics. Berlin, Heidelberg: Springer-Verlag; 2004.

36. Andrieu C, De Freitas N, Doucet A, Jordan MI. An introduction to MCMC for machine learning. Machine Learning. 2003;50(1-2):5–43.

37. Green PJ. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika. 1995;82(4):711–732.

38. Geweke J. Getting It Right: Joint Distribution Tests of Posterior Simulators. Journal of the American Statistical Association. 2004;99(467):799–804.

39. Hyvarinen A. Fast and robust fixed-point algorithms for independent component analysis. IEEE Transactions on Neural Networks. 1999;10(3):626–634.

40. Novichkov PS, Kazakov AE, Ravcheev DA, Leyn SA, Kovaleva GY, Sutormin RA, et al. RegPrecise 3.0–a resource for genome-scale exploration of transcriptional regulation in bacteria. BMC Genomics. 2013;14(1):745.

41. Arous S, Buchrieser C, Folio P, Glaser P, Namane A, Hebraud M, et al. Global analysis of gene expression in an rpoN mutant of *Listeria monocytogenes*. Microbiology. 2004;150(5):1581–1590.

42. Raengpradub S, Wiedmann M, Boor KJ. Comparative analysis of the sigma B-dependent stress responses in *Listeria monocytogenes* and *Listeria innocua* strains exposed to selected stress conditions. Appl Environ Microbiol. 2008;74(1):158–171.

43. Chaturongakul S, Raengpradub S, Palmer ME, Bergholz TM, Orsi RH, Hu Y, et al. Transcriptomic and phenotypic analyses identify coregulated, overlapping regulons among PrfA, CtsR, HrcA, and the alternative sigma factors σB, σC, σH, and σL in *Listeria monocytogenes*. Applied Environ Microbiol. 2011;77(1):187–200.

44. Palmer ME, Chaturongakul S, Wiedmann M, Boor KJ. The *Listeria monocytogenes* σB regulon and its virulence-associated functions are inhibited by a small molecule. MBio. 2011;2(6):e00241–11.

45. Rochat T, Nicolas P, Delumeau O, Rabatinová A, Korelusova J, Leduc A, et al. Genome-wide identification of genes directly regulated by the pleiotropic transcription factor Spx in *Bacillus subtilis*. Nucleic Acids Research. 2012;40(19):9571–9583.

46. Lin AA, Walthers D, Zuber P. Residue substitutions near the redox center of *Bacillus subtilis* Spx affect RNA polymerase interaction, redox control and Spx-DNA contact at a conserved cis-acting element. Journal of Bacteriology. 2013; p. JB–00645.

47. Whiteley AT, Ruhland BR, Edrozo MB, Reniere ML. A redox-responsive transcription factor is critical for pathogenesis and aerobic growth of *Listeria monocytogenes*. Infection and immunity. 2017; p. IAI–00978.

48. Mandin P, Fsihi H, Dussurget O, Vergassola M, Milohanic E, Toledo-Arana A, et al. VirR, a response regulator critical for *Listeria monocytogenes* virulence. Molecular Microbiology. 2005;57(5):1367–1380.

49. Scortti M, Monzó HJ, Lacharme-Lora L, Lewis DA, Vázquez-Boland JA. The PrfA virulence regulon. Microbes and Infection. 2007;9(10):1196–1207.

50. de las Heras A, Cain RJ, Bielecka MK, Vázquez-Boland JA. Regulation of *Listeria* virulence: PrfA master and commander. Current Opinion in Microbiology. 2011;14:118–127.

51. Fritsch F, Mauder N, Williams T, Weiser J, Oberle M, Beier D. The cell envelope stress response mediated by the LiaFSRLm three-component system of *Listeria monocytogenes* is controlled via the phosphatase activity of the bifunctional histidine kinase LiaSLm. Microbiology. 2011;157(2):373–386.

52. Ikebata H, Yoshida R. Repulsive parallel MCMC algorithm for discovering diverse motifs from large sequence sets. Bioinformatics. 2015;31(10):1561–1568.

53. Hermsen R, Tans S, ten Wolde PR. Transcriptional regulation by competing transcription factor modules. PLoS Comput Biol. 2006;2(12):e164.

54. Bailey TL, Bodén M, Whitington T, Machanick P. The value of position-specific priors in motif discovery using MEME. BMC Bioinformatics. 2010;11:179.

55. Narlikar L, Gordân R, Hartemink AJ. A nucleosome-guided map of transcription factor binding sites in yeast. PLoS Computational Biology. 2007;3:e215.

56. Gordân R, Narlikar L, Hartemink AJ. Finding regulatory DNA motifs using alignment-free evolutionary conservation information. Nucleic Acids Research. 2010;38:e90.

57. Klepper K, Drabløs F. PriorsEditor: a tool for the creation and use of positional priors in motif discovery. Bioinformatics. 2010;26:2195–2197.

58. Ma HW, Kumar B, Ditges U, Gunzer F, Buer J, Zeng AP. An extended transcriptional regulatory network of *Escherichia coli* and analysis of its hierarchical structure and network motifs . Nucleic Acids Research. 2004;32(22):6643–6649.

59. Goelzer A, Bekkal Brikci F, Martin-Verstraete I, Noirot P, Bessières P, Aymerich S, et al. Reconstruction and analysis of the genetic and metabolic regulatory networks of the central metabolism of *Bacillus subtilis*. BMC Systems Biology. 2008;2:20.

60. van der Veen S, van Schalkwijk S, Molenaar D, de Vos WM, Abee T, Wells-Bennik MHJ. The SOS response of *Listeria monocytogenes* is involved in stress resistance and mutagenesis. Microbiology. 2010;156:374–384.

61. Ledala N, Sengupta M, Muthaiyan A, Wilkinson BJ, Jayaswal RK. Transcriptomic response of *Listeria monocytogenes* to iron limitation and Fur mutation. Applied Environ Microbiol. 2010;76:406–416.

62. Camejo A, Buchrieser C, Couvé E, Carvalho F, Reis O, Ferreira P, et al. In vivo transcriptional profiling of *Listeria monocytogenes* and mutagenesis identify new virulence factors involved in infection. PLoS Pathog. 2009;5(5):e1000449.

63. Olson ME, King JM, Yahr TL, Horswill AR. Sialic acid catabolism in *Staphylococcus aureus*. Journal of Bacteriology. 2013;195:1779–1788.

64. Larsen MH, Kallipolitis BH, Christiansen JK, Olsen JE, Ingmer H. The response regulator ResD modulates virulence gene expression in response to carbohydrates in *Listeria monocytogenes*. Molecular Microbiology. 2006;61:1622–1635.

65. Lobel L, Herskovits AA. Systems Level Analyses Reveal Multiple Regulatory Activities of CodY Controlling Metabolism, Motility and Virulence in *Listeria monocytogenes*. PLoS Genet. 2016;12(2):e1005870.

66. Lobel L, Sigal N, Borovok I, Belitsky BR, Sonenshein AL, Herskovits AA. The metabolic regulator CodY links *Listeria monocytogenes* metabolism to virulence by directly activating the virulence regulatory gene prfA. Molecular Microbiology. 2015;95:624–644.

67. Zhu B, Stülke J. SubtiWiki in 2018: from genes and proteins to functional network annotation of the model organism *Bacillus subtilis*. Nucleic Acids Research. 2018;46:D743–D748.

68. Behari J, Youngman P. A Homolog of CcpA Mediates Catabolite Control in *Listeria monocytogenes* but Not Carbon Source Regulation of Virulence Genes. Journal of Bacteriology. 1998;180(23):6316–6324.

69. Ruhland BR, Reniere ML. Sense and sensor ability: redox-responsive regulators in *Listeria monocytogenes*. Current Opinion in Microbiology. 2019;47:20–25.

70. Durand S, Braun F, Lioliou E, Romilly C, Helfer AC, Kuhn L, et al. A nitric oxide regulated small RNA controls expression of genes involved in redox homeostasis in *Bacillus subtilis*. PLoS Genetics. 2015;11:e1004957.

71. Goelzer A, Fromion V. Bacterial growth rate reflects a bottleneck in resource allocation. Biochimica et Biophysica Acta (BBA) - General Subjects. 2011;1810(10):978 – 988.

72. Magnusson LU, Farewell A, Nyström T. ppGpp: a global regulator in *Escherichia coli*. Trends in Microbiology. 2005;13(5):236 – 242.

73. Paul BJ, Ross W, Gaal T, Gourse RL. rRNA Transcription in *Escherichia coli*. Annual Review of Genetics. 2004;38(1):749–770.

74. Krásný L, Gourse RL. An alternative strategy for bacterial ribosome synthesis: *Bacillus subtilis* rRNA transcription regulation. The EMBO Journal. 2004;23(22):4473–4483.

75. Bennett HJ, Pearce DM, Glenn S, Taylor CM, Kuhn M, Sonenshein AL, et al. Characterization of relA and codY mutants of Listeria monocytogenes: identification of the CodY regulon and its role in virulence. Mol Microbiol. 2007;63(5):1453–1467.

76. Whiteley AT, Pollock AJ, Portnoy DA. The PAMP c-di-AMP Is Essential for Listeria monocytogenes Growth in Rich but Not Minimal Media due to a Toxic Increase in (p)ppGpp. [corrected]. Cell host & microbe. 2015;17:788–798.

Table 1. Summary of the 40 stable motifs identified in *L. monocytogenes* EGD-e

motif[a]	#[b]	#(0.8-0.2)[c]	W[d]	IC1[e]	Pal[f]	Pos[g]	FC[h]	Runs[i]	Comment[j]
M71.2	1,325	1,174-1,435	9	3	1.3	-9 [1]	0.44	7:10	SigA -10
M71.8	1,308	848-1,506	6	2	1.6	1 [1]	0.46	13:16	TSS (A)
M36.4	1,033	552-1,382	6	3	0	-32 [2]	0.48	10:10	SigA -35 (TTG)
M55.7	836	117-1,486	5	0	1.7	-25 [40]	0.58	3:7	SigA -10 extension (CT)
M8.3	792	331-1,273	5	1	0	4 [1]	0.53	10:15	TSS (G)
M9.10	735	447-1,165	8	4	0.6	-13 [1]	0.50	7:8	SigA (extended -10, TG)
M61.5	561	171-1,153	5	5	0	14 [10]	0.41	6:8	RBS (GGAGG)
M59.9	297	96-844	21	7	0	-60 [42]	0.36	4:10	T-rich element
M26.1	252	107-640	13	5	2.1	-79 [23]	0.73	6:10	SigA -10 on reverse strand
M27.6	240	153-590	6	5	0.4	-33 [2]	0.48	6:6	SigA -35 (TTGAC)
M29.8	122	91-150	12	4	2.2	-13 [2]	3.06	10:10	SigB -10
M9.1	116	43-368	22	2	5.6	-75 [37]	1.18	2:8	loose
M18.2	108	47-355	6	6	0.3	9 [12]	0.52	2:8	RBS (GAGGTG)
M12.4	97	71-109	7	4	0.6	-32 [3]	3.87	10:10	SigB -35
M69.4	87	59-128	15	8	**10.9**	-30 [53]	1.75	10:10	CcpA (CRE-box)
M42.8	62	14-295	17	0	0.2	-86 [21]	0.61	2:4	loose
M31.4	39	16-81	23	8	**18.8**	-40 [61]	1.88	2:2	Rex
M68.1	31	20-47	20	5	**15**	-46 [25]	3.00	10:10	LiaR
M58.7	27	14-40	19	6	**9.9**	-47 [56]	3.23	9:10	-
M62.3	26	18-34	20	13	**16.2**	-16 [67]	1.89	10:10	Fur
M38.10	22	18-38	15	8	**12.1**	-19 [28]	2.36	9:10	LexA
M53.9	20	8-50	20	8	**15.1**	-47 [60]	1.52	10:10	VirR
M2.6	19	12-53	9	4	1.1	-49 [2]	1.12	3:5	Spx
M13.7	19	7-48	21	5	1.4	-58 [69]	0.89	2:3	-
M61.6	14	9-16	25	14	**16.3**	-29 [3]	0.81	9:10	BglR2
M54.1	14	6-40	23	8	1	-56 [61]	1.04	2:5	-
M61.4	13	7-25	25	6	0.9	-37 [62]	1.20	2:2	-
M50.10	13	7-43	21	5	**9.6**	-77 [28]	0.93	2:2	-
M70.6	11	7-19	24	22	**19.2**	-44 [63]	1.28	10:10	-
M3.1	9	4-20	21	9	**15.5**	-43 [62]	1.52	10:10	-
M33.8	8	7-13	22	11	4.2	-24 [23]	2.78	9:10	SigL
M49.4	7	6-11	23	12	**16.8**	-35 [19]	10.55	10:10	PrfA
M20.1	7	5-7	25	16	5.4	-55 [3]	6.64	10:10	-
M17.4	7	5-14	24	9	**14.8**	-39 [61]	1.14	3:4	CcpB
M73.4	6	4-7	20	5	4.2	-44 [8]	5.03	6:6	-
M61.3	6	5-10	25	22	6.4	-33 [4]	3.14	10:10	-
M2.1	6	4-8	22	9	**13**	-33 [56]	1.05	2:5	-
M18.6	4	3-8	25	9	**17.2**	-51 [55]	1.25	6:7	-
M29.1	3	2-6	25	1	3	-39 [62]	1.64	9:10	loose
M59.2	2	0-13	23	2	5.1	-52 [63]	2.43	4:5	loose

[a] unique motif identifier build as Mxx.yy where yy identifies the run and xx the motif in the run; [b] number of promoter regions where the motif is predicted to occur (estimated posterior probability ≥ 0.5), used to order the motifs; [c] number of TSSs when the posterior probability cut-off is set to 0.8 (very likely) or 0.2 (possible); [d] motif width corresponding to the number of columns included in the PWM with posterior probability ≥ 0.5; [e] number of columns in the PWM with high information content, i.e. $2 + \sum_{u \in \{\mathtt{A,C,G,T}\}} \theta_{m,w,u} \log_2(\theta_{m,w,u}) \geq 1$; [f] estimated number of paired columns in the PWM reflecting the degree of palindromness (in boldface when high); [g] median position of occurrence for the middle of the motif with respect to the TSS (inter-quartile range reported between brackets), both numbers are derived from the estimated probability density function for the position described by the variables K_m, $\lambda_{m,\cdot}$ and $d_{m,\cdot}$; [h] maximum across the 165 pairs of conditions for the median of the expression values (log2 fold-change) associated with the TSSs counted in the second column; [i] number of parallel runs (out of 10) in which this motif was found as obtained by clustering based on amount of overlaps between occurrences, written in the format xx:yy where xx and yy are the numbers obtained with cut-offs 75% of overlap and 25% of overlap, respectively; [j] link to known transcription factors if identified or other observations.

Table 2. Contribution of model ingredients to identification of the 40 stable motifs.

	Motif information[a]							Model (dist. to closest motif)[b]				
Motif	**#**	**IC1**	**Pal**	**Pos**	**FC**	**Runs**	**Note**	$t^0e^1p^1$	$t^1e^0p^1$	$t^1e^1p^0$	$t^0e^0p^1$	$t^0e^0p^0$
M71.2	1325	3	1.3	-9 [1]	0.44	7:10	SigA -10	0.522	0.354	**0.129**	0.546	0.500
M71.8	1308	2	1.6	1 [1]	0.46	13:16	TSS	0.857	0.448	**0.057**	0.770	0.719
M36.4	1033	3	0	-32 [2]	0.48	10:10	SigA -35	0.974	**0.108**	**0.081**	0.978	0.983
M55.7	836	0	1.7	-25 [40]	0.58	3:7	SigA -10	0.976	0.991	**0.195**	0.986	0.970
M8.3	792	1	0	4 [1]	0.53	10:15	TSS	0.939	0.687	0.428	0.871	0.844
M9.10	735	4	0.6	-13 [1]	0.5	7:8	SigA -10	0.630	0.418	0.295	0.738	0.644
M61.5	561	5	0	14 [10]	0.41	6:8	RBS	0.625	0.332	0.422	0.654	0.651
M59.9	297	7	0	-60 [42]	0.36	4:10	-	0.463	**0.230**	0.401	0.512	0.577
M26.1	252	5	2.1	-79 [23]	0.73	6:10	SigA -10	0.488	0.545	0.325	0.560	0.496
M27.6	240	5	0.4	-33 [2]	0.48	6:6	SigA -35	0.931	**0.081**	**0.239**	0.974	0.980
M29.8	122	4	2.2	-13 [2]	3.06	10:10	SigB -10	0.326	0.349	**0.055**	0.969	0.885
M9.1	116	2	5.6	-75 [37]	1.18	2:8	loose	0.504	0.943	0.946	0.471	0.508
M18.2	108	6	0.3	9 [12]	0.52	2:8	RBS	0.327	0.442	0.273	0.970	0.325
M12.4	97	4	0.6	-32 [3]	3.87	10:10	SigB -35	0.988	0.946	**0.035**	0.982	0.887
M69.4	87	8	10.9	-30 [53]	1.75	10:10	CcpA	**0.226**	0.379	**0.083**	0.349	0.387
M42.8	62	0	0.2	-86 [21]	0.61	2:4	-	0.987	0.355	0.976	0.968	0.996
M31.4	39	8	18.8	-40 [61]	1.88	2:2	Rex	0.331	0.987	0.990	0.988	0.989
M68.1	31	5	15	-46 [25]	3	10:10	LiaR	**0.200**	0.914	**0.228**	0.994	0.996
M58.7	27	6	9.9	-47 [56]	3.23	9:10	-	0.425	0.979	0.400	0.966	0.973
M62.3	26	13	16.2	-16 [67]	1.89	10:10	Fur	**0.066**	**0.085**	**0.065**	**0.112**	**0.091**
M38.10	22	8	12.1	-19 [28]	2.36	9:10	LexA	0.371	0.394	**0.119**	0.409	0.382
M53.9	20	8	15.1	-47 [60]	1.52	10:10	VirR	0.294	0.297	0.308	0.454	0.356
M2.6	19	4	1.1	-49 [2]	1.12	3:5	Spx	0.982	0.992	0.371	0.993	0.993
M13.7	19	5	1.4	-58 [69]	0.89	2:3	-	0.956	0.434	0.986	0.991	0.956
M61.6	14	14	16.3	-29 [3]	0.81	9:10	BglR2	0.560	**0.043**	**0.120**	0.572	0.469
M54.1	14	8	1	-56 [61]	1.04	2:5	-	0.890	**0.158**	0.301	**0.160**	0.993
M61.4	13	6	0.9	-37 [62]	1.2	2:2	-	1.000	0.923	0.937	0.963	0.973
M50.10	13	5	9.6	-77 [28]	0.93	2:2	-	0.992	0.863	0.984	1.000	0.972
M70.6	11	22	19.2	-44 [63]	1.28	10:10	-	0.282	**0.119**	**0.181**	0.277	**0.208**
M3.1	9	9	15.5	-43 [62]	1.52	10:10	-	**0.122**	**0.112**	0.585	**0.194**	0.520
M33.8	8	11	4.2	-24 [23]	2.78	9:10	SigL	**0.224**	0.633	0.332	0.413	0.381
M49.4	7	12	16.8	-35 [19]	10.55	10:10	PrfA	**0.169**	0.992	**0.176**	1.000	0.999
M20.1	7	16	5.4	-55 [3]	6.64	10:10	-	**0.057**	0.524	**0.042**	0.554	0.589
M17.4	7	9	14.8	-39 [61]	1.14	3:4	CcpB	0.946	**0.081**	0.944	0.948	0.946
M73.4	6	5	4.2	-44 [8]	5.03	6:6	-	0.998	0.994	0.378	0.998	0.958
M61.3	6	22	6.4	-33 [4]	3.14	10:10	-	**0.167**	0.251	**0.036**	0.254	**0.194**
M2.1	6	9	13	-33 [56]	1.05	2:5	-	**0.110**	0.734	0.706	**0.152**	**0.206**
M18.6	4	9	17.2	-51 [55]	1.25	6:7	-	**0.010**	**0.005**	0.915	**0.025**	0.920
M29.1	3	1	3	-39 [62]	1.64	9:10	-	0.401	**0.153**	**0.159**	**0.000**	**0.006**
M59.2	2	2	5.1	-52 [63]	2.43	4:5	-	**0.000**	**0.246**	**0.186**	0.329	**0.156**
Number of motifs with matches at distance ≤ 0.25								11	12	18	6	6
at distance ≤ 0.75								26	29	32	21	21

[a] Columns as in Table 1. [b] Distance to the closest stable motif found with different model settings, model ingredients are encoded in a three-term format: t^1/t^0 for with or without distance to TSS, e^1/e^0 for with or without expression covariates; p^1/p^0 for with our without enabling palindromic structures. Distances below 0.25 are highlighted (boldface).