BINDER: computationally inferring a gene regulatory network for *Mycobacterium abscessus*

Patrick M. Staunton[1], Aleksandra A. Miranda-CasoLuengo[2], Brendan J. Loftus[1], and Isobel Claire Gormley*[3]

[1]School of Medicine, Conway Institute, University College Dublin

[2]Moyne Institute of Preventive Medicine, Department of Microbiology, Trinity College Dublin

[3]School of Mathematics & Statistics, Insight Centre for Data Analytics, University College Dublin

Abstract

Background: Although many of the genic features in *Mycobacterium abscessus* have been fully validated, a comprehensive understanding of the regulatory elements remains lacking. Moreover, there is little understanding of how the organism regulates its transcriptomic profile, enabling cells to survive in hostile environments. Here, to computationally infer the gene regulatory network for *Mycobacterium abscessus* we propose a novel statistical computational modelling approach: BayesIan gene regulatory Networks inferreD via gene coExpression and compaRative genomics (BINDER). In tandem with derived experimental coexpression data, the property of genomic conservation is exploited to probabilistically infer a gene regulatory network in *Mycobacterium abscessus.*

Inference on regulatory interactions is conducted by combining 'primary' and 'auxiliary' data strata. The data forming the primary and auxiliary strata are derived from RNA-seq experiments and sequence information in the primary organism *Mycobacterium abscessus* as well as ChIP-seq data extracted from a related proxy organism *Mycobacterium tuberculosis.* The primary and auxiliary data are combined in a hierarchical Bayesian framework, informing the apposite bivariate likelihood function and prior distributions respectively. The inferred relationships provide insight to regulon groupings in *Mycobacterium abscessus.*

Results: We implement BINDER on data relating to a collection of 167,280 regulator-target pairs resulting in the identification of 54 regulator-target pairs, across 5 transcription factors, for which there is strong probability of regulatory interaction.

Conclusions: The inferred regulatory interactions provide insight to, and a valuable resource for further studies of, transcriptional control in *Mycobacterium abscessus*, and in the family of *Mycobacteriaceae* more generally. Further, the developed BINDER framework has broad applicability, useable in settings where computational inference of a gene regulatory network requires integration of data sources derived from both the primary organism of interest and from related proxy organisms.

Keywords— gene regulatory network, *Mycobacterium abscessus*, Bayesian inference, data integration

*claire.gormley@ucd.ie

1 Background

Mycobacterium abscessus is a rapidly growing mycobacteria capable of causing a variety of soft tissue infections, primarily affecting subjects with immuno-deficiencies. *Mycobacterium abscessus* (*M. abscessus*) is considered a major pathogen involved in broncho-pulmonary infection in patients with cystic fibrosis or chronic pulmonary disease (1). In addition, *M. abscessus* is responsible for several skin and soft tissue diseases, central nervous system infections, bacteremia, and ocular and other infections (2). Owing to a range of cellular mechanisms, one of the most salient aspects of pathogenesis resulting from *M. abscessus* infection is its multi-drug resistance. It is the most chemotherapy-resistant rapid-growing mycobacterium (3).

While many genic features in *M. abscessus* have been fully validated and characterised in terms of the expression landscape at the transcriptional, post-transcriptional and translational levels (4), a comprehensive understanding of regulatory elements is lacking. Without functional identification of the modes of regulation present, a complete understanding of how *M. abscessus* modulates its transcriptomic tendencies, enabling cells to survive and thrive in hostile environments such as in the presence of antibiotics or in the host sputum, remains out of reach.

Gene regulatory network (GRN) resources are typically split into two categories: generalist resources and specialist resources. The former category provides regulatory information (such as transcription factors, putative and confirmed target genes/operon structures, transcription factor binding sites (TFBS) motifs, upstream location coordinates) for a wide group of organisms. CollecTF (5) is one such resource that hosts a large collection of DNA binding sites for prokaryotic transcription factors. Although CollecTF comprises a small amount of regulatory information pertaining to mycobacteria, it currently does not contain any information on *M. abscessus*. Indeed most generalist resources tend not to comprise much content on regulatory information directly relevant to *M. abscessus*.

Specialist resources tend to provide regulatory information for a much narrower subgroup of organisms such as a single species or genus; RegulonDB (6) is one such resource which comprises information regarding transcriptional regulation in *Escherichia coli*. Most resources of both types provide curation based on techniques such as SELEX-based methods (7) as well as ChIP-seq (8). Currently, for *M. abscessus*, there is no such existing specialist resource.

Many approaches have been designed for *in silico* inference of prokaryotic GRNs. Two popular strategies for regulon mapping include (1) the use of conservation data arising from comparative genomics analyses and (2) expression data in the form of transcriptional abundance comparison. The conservation approach relies on the observation that TFBSs are often conserved between related species. This implies that regulatory resources from a given organism can be leveraged to elucidate on transcriptional control in closely related organisms (9). Further, if two organisms with a non-distant common ancestor share an orthologous gene that is understood to assist in achieving a certain biological process (such as transcriptional regulation) in one organism, it is likely to perform a similar role in the other organism (10). Phylogenetic footprinting provides a conservation-based approach for determining conserved noncoding sequences and associated TFBSs; such methods typically involve quantifying the rate of occurrence of noncoding DNA sequences in the upstream regions of orthologs of genes of interest in related species (11; 12).

Expression-based approaches tend to model the expression of a target gene candidate as a function of the expression or activation of a regulator gene. The GENIE3 (13) method frames the problem of deriving a regulatory network between p genes as p different regression tree-based ensemble models where the expression pattern of one gene is predicted by the expression pattern of all other genes in the collection. Other authors have noted the observed property that genes sharing a common network have a greater tendency to exhibit strong coexpression (14). Weighted correlation network analysis (WGCNA) (15) is a software package that implements a suite of correlation-based methods for describing the coexpression patterns among genes across experimental samples designed with a view to uncovering gene networks of several varieties.

The literature on prokaryotic gene regulation is replete with ChIP-seq experiments detailing the specifics of transcriptomic control (16; 17). ChIP-seq provides a means of isolating target DNA sequences and transcription factor bound protein complexes stimulated in response to induced transcription factor production. This process facilitates the ascertaining of relationships between specific transcription factors and target binding site DNA sequences (including their downstream genic and intergenic units). Such data are not presently available for *M. abscessus*, due to its status as an emerging pathogen (3). However, similar resources exist to varying degrees of completeness for closely related organisms, such as those in the family of Mycobacteriaceae (18; 19). Many efforts have focussed on the integration of ChIP-seq experimental data with RNA-based expression results to improve GRN inference (20).

In general, the concept of designing hybrid models that integrate existing regulatory information and expression abundance results has been the focus of much research. For example, iRafNet (21) implements a random forest approach to inferring GRNs while incorporating prior regulatory knowledge such that putative regulators used to build individual trees are sampled in accordance with the provided prior information. GRACE (22) integrates biological *a priori* data as well as heterogeneous data and makes use of Markov random fields to infer regulatory networks in eurkaryotic organisms. The RNEA (23) approach also combines prior knowledge from manual literature curation and experimental data with enrichment analysis to infer relevant subnetworks under experimental conditions. The multi-species cMonkey approach (24) includes gene expression data for multiple related organisms in addition to upstream sequence information and other network knowledge, iteratively building biclusters to detect putative co-regulated gene groupings.

Hierarchical Bayesian frameworks provide a natural choice for heterogenous data integration; Bayesian methods like COGRIM (25) and CRNET (26) have sought to exploit this quality. With a view to inferring GRNs, integrative Bayesian methods have focussed on directly modelling putative target gene expression data as a function of regulator activity in addition to binding strength and sequence information.

Herein, we introduce a novel statistical modelling approach to computationally inferring the GRN for *M. abscessus*: BayesIan gene regulatory Networks inferreD via gene coExpression and compaRative genomics (BINDER). BINDER is an integrative approach, hybridising coexpression data and comparative genomics profiles to infer prokaryotic regulons. BINDER requires two organisms: an organism of interest, here *M. abscessus*, and an annotated proxy organism, here *Mycobacterium tuberculosis* (*M. tuberculosis*). To computationally infer the GRN for *M. abscessus* we leverage existing resources: specifically we exploit several RNA-seq libraries elicited from *M. abscessus* generated across a range of experimental conditions, and the unique availability of a high-quality and comprehensively catalogued ChIP-seq-derived regulatory network in *M. tuberculosis* (27). BINDER utilises a primary data stratum and an auxiliary data stratum. Here, the data forming the primary and auxiliary strata are derived from RNA-seq experiments and sequence information from *M. abscessus* as well as ChIP-seq data extracted from the related *M. tuberculosis*. BINDER is a Bayesian hierarchical model that appositely models the type and structure of both this primary and auxiliary data to infer the probability of a regulatory interaction between a regulator-target pair. The auxiliary data inform the prior distributions and the posterior distributions are updated by accounting for the primary coexpression data in a novel, apposite bivariate likelihood function. BINDER's Bayesian framework facilitates the borrowing of information across the genome yielding estimates of the probability of regulation between regulator and target candidate genes, as well as quantification of the inherent uncertainty in a probabilistically principled manner.

In what follows, we explore the performance of BINDER under a range of challenging simulated data settings, as well as in two case studies using *Bacillus subtilis* (*B. subtilis*) and *Escherichia coli* (*E. coli*) as the primary organisms of interest, for which regulatory interactions have been well-established. We present the regulatory interactions inferred on *M. abscessus* by BINDER, and explore in detail the putative inferred regulon corresponding to the transcriptional regulator zur. We also include an exploration of prior sensitivity concerns and some discussion. The Methods section (Section 5) describes the data utilised and details the architecture of the BINDER approach.

The results of this effort provide insight to, and a valuable resource for further studies of, transcriptional control in *M. abscessus*, and in the family of *Mycobacteriaceae* more generally. Further, the developed BINDER framework has broad applicability, useable in settings where computational inference of a GRN requires integration of data sources derived from both the primary organism of interest and from a related proxy organism. A software implementation for BINDER is provided by its associated R package, which is freely available from github.com/ptrcksn/BINDER.

2 Results

2.1 Exploring *M. abscessus* and *M. tuberculosis* shared orthology

It has been established that there is high retention of gene regulation in prokaryotes between species (28). Moreover, it has been demonstrated that gene function is also retained across wide phylogenetic distances in prokaryotes (29). Given the availability of a large number of experimentally validated regulatory networks in *M. tuberculosis* (27), from the standpoint of inferring a GRN in *M. abscessus* using conservation phenomena, we quantifed the extent to which genes present in *M. tuberculosis* are conserved in *M. abscessus*. To do so, we employ the Ortholuge (64) procedure which facilitates bacterial and archaeal comparative genomic analysis and large-scale ortholog predictions. Through Ortholuge, we categorise orthologs as belonging to one of five tiers, ranging from more reliable to less reliable: supporting-species-divergence (SSD), borderline supporting-species-divergence (borderline SSD), reciprocal best blast (RBB), similar non-supporting-species-divergence (similar non-SSD) and non-supporting-species-divergence (non-SSD). We found 1,343 SSD putative orthologs, 116 borderline SSD putative orthologs, 845 genes that satisfied the RBB criteria but did not undergo any further analysis, 6 similar non-SSD putative orthologs and 85 non-SSD putative orthologs. In total, we found 2,395 predicted orthologs of all qualities, equating to $\approx 48\%$ of all annotated genes in *M. abscessus*.

In terms of regulatory interactions, for 34 orthologous regulators of interest and where possible, we performed a one-to-one mapping of all validated regulatory interactions in *M. tuberculosis* to their corresponding orthologs in *M. abscessus*. We found a mean regulon size in *M. tuberculosis* of 107.91 genes (sd: 128.78). Of these 34 regulons, the mean regulon proportion comprising orthologous interactions in *M. abscessus* is 0.61 (sd: 0.16) (Figure 1). These results are suggestive of conserved regulatory interactions between *M. tuberculosis* and *M. abscessus*.

2.2 BINDER Simulation Study

In order to evaluate the performance of BINDER (Section 5.2), we perform a simulation study across a number of settings. Our focus is on exploring the impact of BINDER's hierarchical Bayesian model structure and on the influence of the inclusion of the auxiliary data when inferring a GRN. Specifically we focus on the parameter $\theta_{r,t}$ representing the probability of an interaction in the (r,t)th regulator-target pair and consider two simplified versions of the BINDER model:

- *Deterministic model*: each $\theta_{r,t}$ is modelled deterministically as a linear function of the auxiliary data. Thus BINDER's prior on $\theta_{r,t}$ is replaced by:

$$\text{logit}(\theta_{r,t}) = \zeta_r + \tau_{\text{ME}_r}\text{ME}_{r,t} + \tau_{\text{PE}_r}\text{PE}_{r,t}$$

- *Non-auxiliary model*: no auxiliary data are used during inference on $\theta_{r,t}$, which are instead inferred based on the primary data only. In this case BINDER's prior on $\theta_{r,t}$ is instead replaced by the prior $\text{logit}(\theta_{r,t}) \sim \mathcal{U}(-\infty, \infty)$.

In addition, the impact on inference of noisy primary data and of large variability in the true underlying $\theta_{r,t}$ parameters is also of interest. Since the primary data CP and CM are assumed to be $\mathcal{N}_l\{\text{logit}(\theta_{r,t}), \psi_{k_r}\}$ for $k \in \{\text{CP, CM}\}$, larger values of ψ_{k_r} reflect noisier primary data.

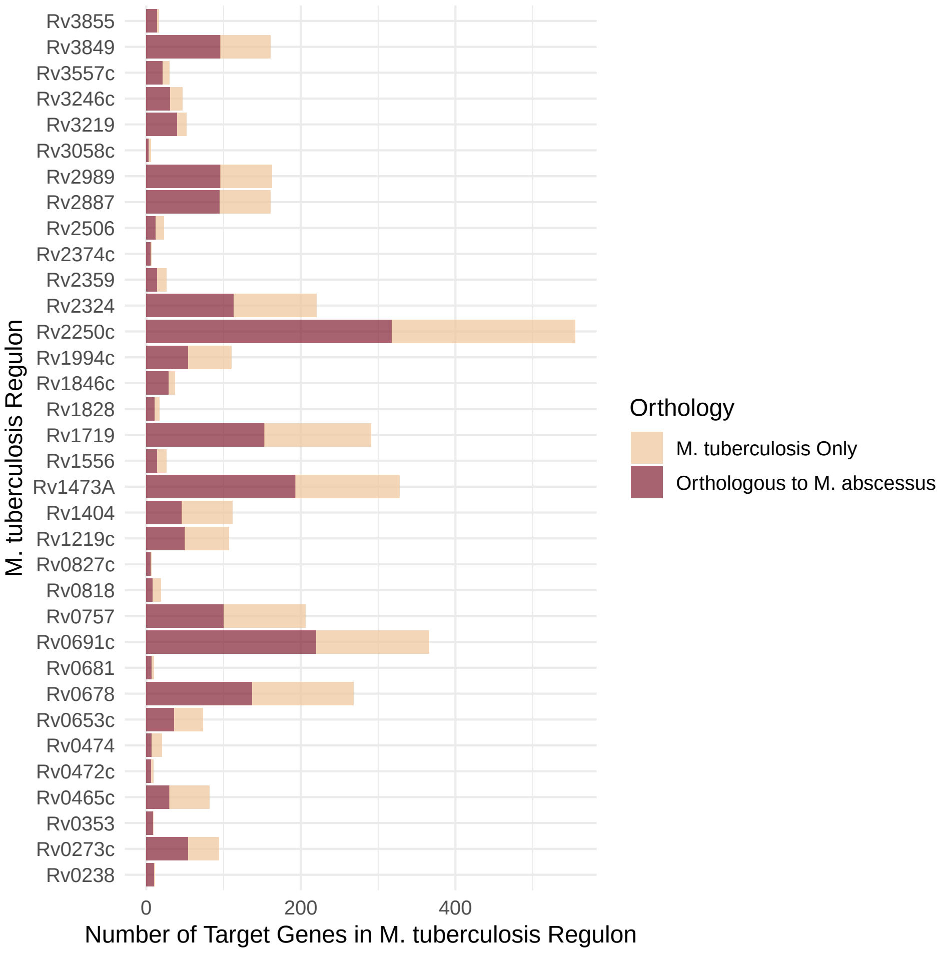

Figure 1: Number of target genes in the 34 orthologous *M. tuberculosis* regulons. Also illustrated is the the extent of orthology between *M. tuberculosis* and *M. abscessus*.

Similarly, $\text{logit}(\theta_{r,t}) \sim \mathcal{N}(\gamma_{r,t}, \phi_r)$, with larger values of ϕ_r reflecting larger variation in the underlying regulatory interaction probabilities. Hence, we compare the performance of BINDER, the deterministic model and the non-auxiliary model on 9 distinct dispersion parameterisations corresponding to the Cartesian product of $\boldsymbol{\psi_r} = \{\psi_{\text{CM}_r}, \psi_{\text{CP}_r}\} = \{\text{low} = 1, \text{mid} = 2, \text{high} = 3\}$ and $\phi_r = \{\text{low} = 1, \text{mid} = 2, \text{high} = 3\}$.

For each of the nine dispersion settings, we simulate three data sets, each with $N = 1,000$ regulator-target pairs. To challenge the BINDER model, we consider weakly informative auxiliary data: ME and PE are generated from a Bernoulli distribution with success parameter 0.1. We compute $\gamma_{r,t}$ according to (1) where $(\zeta_r, \tau_{\text{ME}_r}, \tau_{\text{PE}_r}) = (-3.5, 3.8, 2.9)$ and simulate $\text{logit}(\theta_{r,t}) \sim \mathcal{N}(\gamma_{r,t}, \phi_r)$. Finally, for the primary data, we simulate $\text{CM}_{r,t} \sim \mathcal{N}_l(\text{logit}(\theta_{r,t}), \psi_{\text{CP}_r})$ and $\text{CP}_{r,t} \sim \mathcal{N}(\text{logit}(\theta_{r,t}), \psi_{\text{CM}_r})$. Model performance across the 27 settings considered was assessed using the mean absolute deviation (MAD) (30) between each true simulated $\theta_{r,t}$ and its resulting posterior mean estimate.

We observed competitive performance of the BINDER approach over both the deterministic and non-auxiliary approaches for the majority of settings considered in terms of lower MAD (Figure 2). Specifically, the mean for the MAD statistics for the BINDER approach was 0.087 (sd: 0.034) as compared with 0.120 (sd: 0.050) and 0.120 (sd: 0.056) for the deterministic and non-auxiliary approaches respectively (standard deviations in parentheses). The deterministic approach has a tendency to perform worse in instances where the dispersion around each $\theta_{r,t}$ value is large (i.e. high values for ϕ_r). This is to be expected as the deterministic approach has insufficient flexibility to model $\theta_{r,t}$ values that lie distant from their mean value resulting in higher MAD statistics.

On the contrary, the deterministic approach does well in the setting of low ϕ_r. In contrast, the non-auxiliary approach tends to be less sensitive to changes in the dispersion around the mean of the distribution of $\theta_{r,t}$. However, given that the non-auxiliary approach only uses the primary data to infer $\theta_{r,t}$, when the level of dispersion around the mean of CP and CM is high (i.e. high values for $\boldsymbol{\psi_r}$) the primary data contain a weaker signal leading to poor estimation of the true $\theta_{r,t}$ and resulting in higher MAD statistics. As a compromise between the deterministic and non-auxiliary approaches, BINDER utilises the information contained in the auxiliary data whilst, simultaneously, providing the flexibility to accommodate observation-specific variation in the regulation interaction probabilities resulting in more accurate inference. BINDER outperforms the non-auxiliary model in all settings considered, and is only marginally outperformed in a minority of cases by the deterministic model in settings where ϕ_r is mid or low.

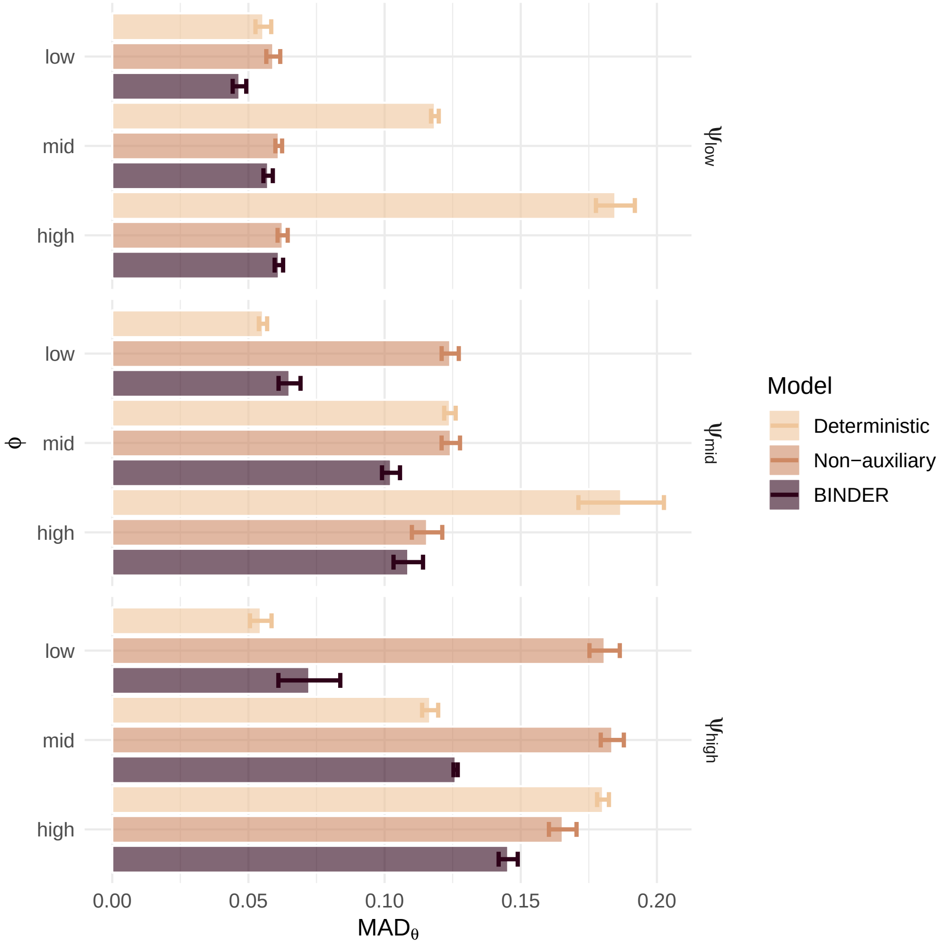

Figure 2: Simulation results illustrating the mean absolute deviation (MAD) between the true and estimated regulation interaction probabilities achieved by the deterministic, non-auxiliary and BINDER approaches across a range of dispersion parameter settings.

2.3 Application of BINDER to *Escherichia coli* and *Bacillus subtilis* data

As a benchmarking exercise to assess the performance of BINDER on a *bona fide* regulatory interaction data set, we investigated BINDER's ability to infer interaction plausibility for the fur and lexA regulons in *Escherichia coli* (31) and *Bacillus subtilis* (32). Where *E. coli* constitutes the organism of interest, *Pseudomonas aeruginosa* (*P. aeruginosa*) (33) constitutes the proxy organism and where *B. subtilis* is the organism of interest, *Listeria monocytogenes* (*L. monocytogenes*) (34) fulfils the role of the proxy organism. Considering two regulons across these well researched settings allows for intra-regulon and inter-regulon analysis as well as intra-organism and inter-organism analysis.

The ferric uptake regulator, or fur, is a transcriptional factor originally described as a repressive regulator of genes involved in iron import. Since then, aside from iron-homeostasis, fur has been shown to be associated with processes such as resistance to oxidative stress, pH homeostasis and quorum sensing as well as other cellular mechanisms (35). In bacteria, the SOS response provides the means for responding to DNA damage; the expression of genes comprising the SOS regulatory network is under the control of lexA (36). lexA is a global transcription factor that undergoes cleavage during stress permitting expression of DNA repair functions (37). lexA also regulates genes that are not comprised within the SOS response program (36).

Here we avail of well-established regulator-target interactions as detailed by RegulonDB (6) for *E. coli* and well-established regulator-target interactions as per SubtiWiki (38) for *B. subtilis.* To build the primary data, we used *E. coli* expression data from COLOMBOS (39) and *B. subtilis* expression data from SubtiWiki (40). For the auxiliary data, we use regulatory sequence motifs and orthologous target interactions from *P. aeruginosa* and *L. monocytogenes* curated by collecTF (5).

We consider the BINDER, deterministic and non-auxiliary approaches to infer the GRNs in *Escherichia coli* and in *Bacillus subtilis* from their primary and auxiliary data. Non-informative priors were employed with mean hyperparameters set to 0 and standard deviation hyperparameters set to 3, with the exception of the prior on ϕ_r which was set to $\phi_r \sim \mathcal{N}_{(0,\infty)}(1, 0.1)$ for regularisation purposes. Further, we also consider iRafNet (21) which employs an integrative prior-information-based approach to random forest inference of GRNs from expression data. For iRafNet, we applied the algorithm to each target candidate of interest individually using the fur and lexA regulator genes as predictors; further, in addition to the standardised expression matrix, for the iRafNet prior information matrix W, the element w_{ij}, corresponding to the ith regulator and jth target candidate, was configured such that $w_{ij} = \exp(1)$ if $\text{ME} = 1$ or $\text{PE} = 1$ and $w_{ij} = \exp(0)$ for $i \neq j$.

In total, of the 4,221 uniquely labelled genes present in RegulonDB with available expression data, 67 correspond to well-established regulatory interactions concerning fur and 23 correspond to well-established interactions concerning lexA in *E. coli.* For *B. subtilis*, of the 4,162 uniquely labelled genes with available expression data, 58 correspond to well-established regulatory interactions with fur and 57 to well-established regulatory interactions with lexA.

For the fur regulon in *E. coli*, BINDER achieved an area under curve (AUC) of 0.880. Notably however, in contrast to BINDER, iRafNet omits data recorded under conditions for which expression levels for all genes are not available. Thus, in order to fairly compare performance with iRafNet, we applied BINDER to a reduced expression matrix comprising fewer conditions such that no missing data were present. BINDER achieved an AUC of 0.787 as compared with 0.710, 0.654 and 0.725 for the non-auxiliary, deterministic and iRafNet approaches respectively (Figure 3, Table 1).

Interestingly, for BINDER applied to the reduced coexpression data, the mean posterior 50th percentile $\theta^{50\%}_{\text{fur},t} \forall t \in T$ corresponding to validated regulatory interactions was only 0.0050 as compared with 0.0016 for the mean $\theta^{50\%}_{\text{fur},t}$ corresponding to observations without evidenced regulatory interactions (Figure 4). That this BINDER implementation achieved a corresponding AUC of 0.787 suggests that the distribution of $\theta^{50\%}_{\text{fur},t}$ values is highly skewed to the right, and thus their relative magnitude is of importance when observing BINDER's output. Interestingly, we did not observe this effect when BINDER was applied to the complete expression data. Thus, we imposed a more informative prior $\phi_{\text{fur}} \sim \mathcal{N}_{(0,\infty)}(10, 0.01)$ and applied BINDER again resulting in a mean $\theta^{50\%}_{\text{fur},t}$ corresponding to validated regulatory interactions of 0.2427 as compared with 0.1833 for the mean $\theta^{50\%}_{\text{fur},t}$ corresponding to observations without evidenced regulatory interactions (Figure 4). However, with this informative prior the AUC dropped to 0.729. This is almost identical to the AUC for the non-auxiliary implementation which is intuitive because as ϕ_{fur} increases, the auxiliary stratum provides diminishing influence (Figure 3, Table 1).

For the lexA regulon in *E. coli*, BINDER achieves an AUC of 0.888. Once again, in order to compare performance with iRafNet, we re-applied BINDER to a reduced expression matrix comprising fewer conditions such that no missing data were present. For the reduced expression data BINDER achieved an AUC of 0.857 as compared with 0.768, 0.778 and 0.829 for the non-auxiliary, deterministic and iRafNet approaches respectively(Figure 3, Table 1).

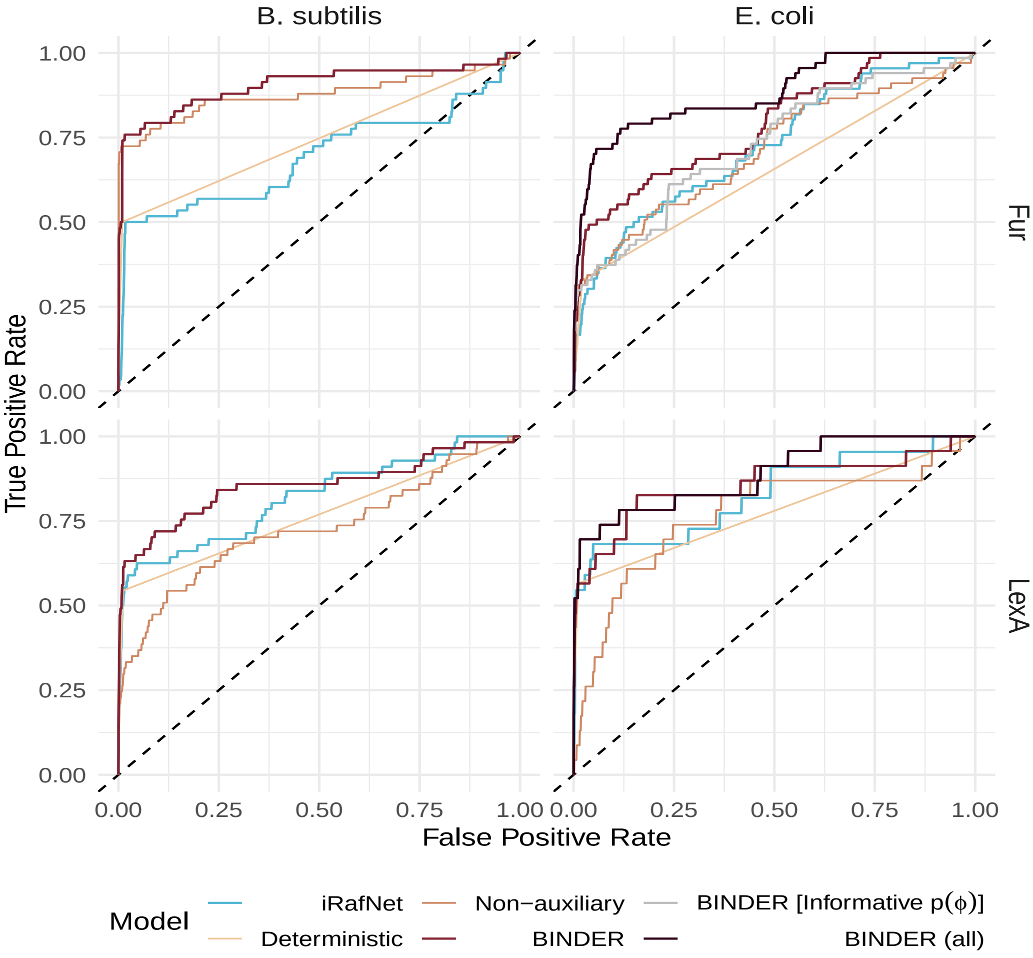

Figure 3: ROC analysis for $\theta_{r,t}^{50\%}$ posterior estimates for the BINDER, deterministic and non-auxiliary approaches and gene importance estimates for iRafNet for the r = fur and r = lexA regulons in *E. coli* and *B. subtilis*. BINDER (all) denotes results from analysis of BINDER applied to the complete coexpression data; BINDER relates to its application to the reduced data set.

Performance was similar for the *B. subtilis* organism (Figure 3, Table 1). For the fur regulon, BINDER achieved an AUC of 0.905 as compared with 0.878, 0.746 and 0.694 for the non-auxiliary, deterministic and iRafNet approaches respectively. For the lexA regulon, BINDER achieves an AUC of 0.855 as compared with 0.728, 0.767 and 0.819 for the non-auxiliary, deterministic and iRafNet approaches respectively.

Not only does BINDER out perform all other considered approaches in terms of AUC, but, considering false positive rates in the neighbourhood of 0, BINDER tends to achieve higher true positive rates than any of the other approaches. This is particularly important because, owing to sparse regulatory connectivity across a given genome, regulon mapping is typically a minority class problem i.e. the vast majority of target candidates will constitute negatives for most regulators. This implies that a low false positive rate can still translate to a large *number* of false positives.

The ability of BINDER to integrate and borrow information across primary and auxiliary data

Table 1: AUC scores achieved by each modelling approach for each regulon in each organism.

Model	fur (*E. coli*)	lexA (*E. coli*)	fur (*B. subtilis*)	lexA (*B. subtilis*)
iRafNet	0.725	0.829	0.694	0.819
Deterministic	0.654	0.778	0.746	0.767
Non-auxiliary	0.710	0.768	0.878	0.728
BINDER	0.787	0.857	0.905	0.855
BINDER (all)	0.880	0.888	-	-
BINDER (informative $p(\phi)$)	0.729	-	-	-

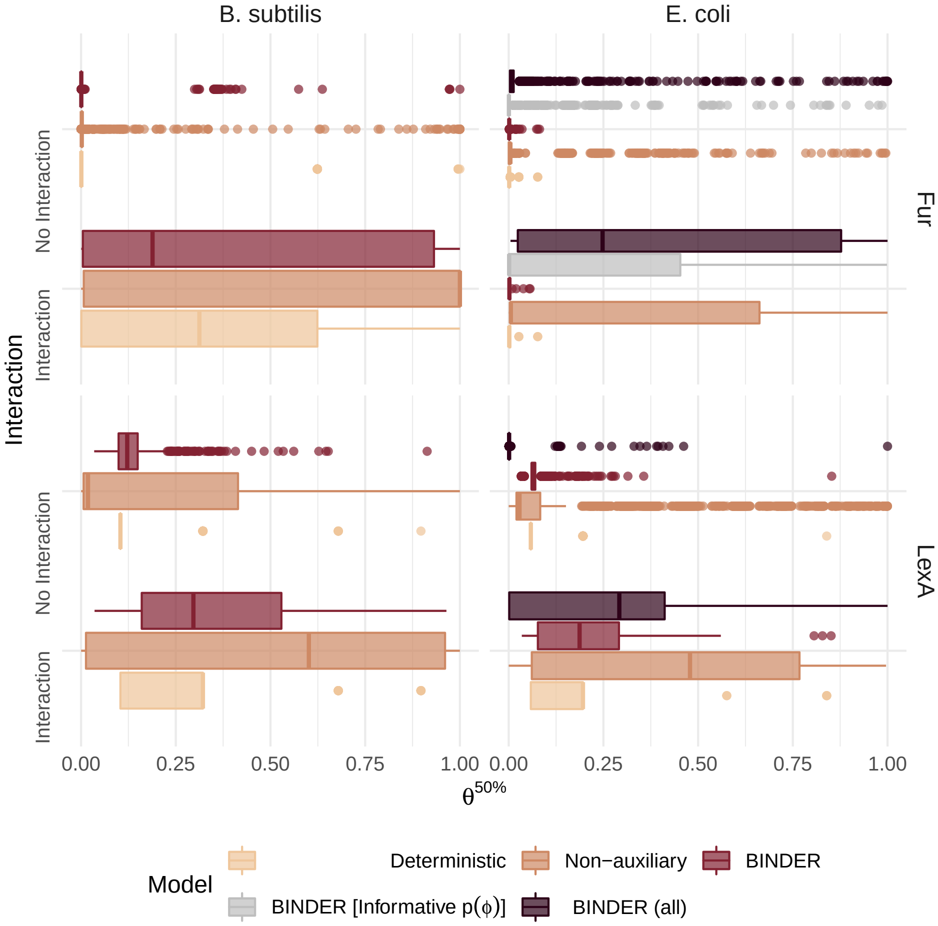

Figure 4: Posterior estimates of $\theta^{50\%}_{r,t}$ for the BINDER, deterministic and non-auxiliary approaches for $r =$ fur and $r =$ lexA regulons in *E. coli* and *B. subtilis*, factored by established interaction status.

when inferring a GRN is demonstrated in Figure 5 for the particular case of the lexA regulator in *B. subtilis* when there is no auxiliary evidence. Only the full BINDER implementation is capable of tempering estimates when there is disagreement between interaction status and auxiliary evidence; when there is an interaction but no auxiliary evidence BINDER is capable of exploiting the individual primary data values, CM and CP, to provide higher estimates to the regulator-target candidate; however, the deterministic approach lacks the flexibility to provide any high $\theta^{50\%}_{\text{lexA},t}$ estimates in the absence of auxiliary evidence. Similarly, owing to the lack of auxiliary evidence, BINDER is capable of tempering its estimates for $\theta^{50\%}_{\text{lexA},t}$ when there is no interaction and no auxiliary evidence; in contrast, the non-auxiliary approach results in high $\theta^{50\%}_{\text{lexA},t}$ estimates for all observations with high primary data values CM and CP. BINDER's hierarchical modelling structure and ability to borrow local *and* global information from both the primary and auxiliary data sources results in more realistic estimates: higher $\theta^{50\%}_{\text{lexA},t}$ estimates for putative interactions and lower $\theta^{50\%}_{\text{lexA},t}$ estimates for putative non-interactions in general. Synoptically, BINDER's ability to integrate the information on whether a given regulator-target pair has an affinity for the predicted motif and/or an orthologous regulatory interaction in the proxy organism with the information provided in the primary data stratum provides greater flexibility.

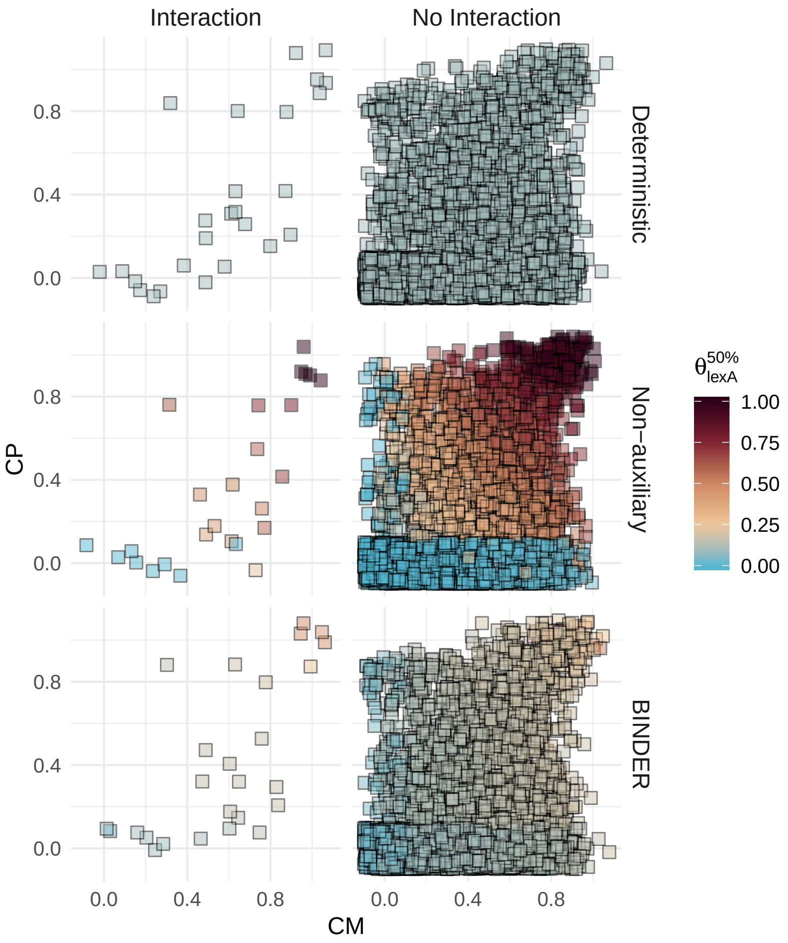

Figure 5: For the lexA regulon in *B. subtilis* and for targets where the auxiliary data ME = 0 and PE = 0, estimates of $\theta^{50\%}_{\text{lexA},t}$ for the BINDER, deterministic and non-auxiliary approaches, factored by known interaction status. The primary data values are CM and CP; points are jittered slightly for visibility.

2.4 Application of BINDER to *M. abscessus* data

With a view to producing a model of regulation in *M. abscessus*, we leveraged data from across 34 orthologous ChIP-seq validated interactions in *M. tuberculosis* and from 32 RNA-seq libraries from across 16 distinct experimental conditions in *M. abscessus*. We considered $R = 34$ orthologous regulators in *M. tuberculosis*, and $T = 4920$ target candidates in the *M. abscessus* genome, yielding $N = 167,280$ regulator-target pairs. For computational efficiency, given the likelihood function can be factored by regulator, we run BINDER on the $R = 34$ orthologous regulators' data in parallel. To computationally infer the gene regulatory network for *M. abscessus* the posterior distribution

$p(\theta_{r,t}|\ldots)$ is of key interest, for $r \in R$ and $t \in T$ with $\ldots$ denoting all auxiliary and primary data and other model parameters.

2.4.1 Prior Sensitivity Analysis

In order to assess the sensitivity of inference to the prior distribution specifications, we constructed three different prior parameterisation settings and compared the resulting inferences. The three settings considered were labelled as 'non-informative', 'informative' and 'precise' (Table 2). In particular, the informative settings reflect *a priori* beliefs that: (1) the auxiliary data PE and ME would encode a reliable positive indication as to whether a given regulatory interaction exists and (2) a negative intercept would be required to correctly model interaction plausibility. The precise setting reflects more extreme versions of the informative setting (in terms of smaller auxiliary data scale hyperparameters).

Inference was relatively insensitive to prior specification in terms of MAD scores for $\theta_{r,t}^{50\%}$ (uninformative versus informative: 0.0040, sd: 0.0094; uninformative versus precise: 0.0183, sd: 0.0466; informative versus precise: 0.0168, sd: 0.0437, Figure 6). Using a classification criterion such that regulator-target pairs with a posterior 50th percentile $\theta_{r,t}^{50\%} > 0.9$ are classified as positive regulation cases, comparing uninformative to informative positive regulation cases yielded an adjusted Rand index (41) of 0.9247, versus 0.5203 and 0.5553 for uninformative versus precise and informative versus precise respectively (an adjusted Rand index of 1 indicates perfect agreement). Thus, for the remainder of this work, with a view to allowing the data to determine the parameter estimates without imposing strong beliefs, we focus on the uninformative parameterisation.

Table 2: Prior parameterisation settings considered for sensitivity analysis of BINDER.

Hyperparameter	**Uninformative**	**Informative**	**Precise**
μ_{ζ_r}	0	-3	-3
σ_{ζ_r}	3	1	0.1
$\mu_{\tau_{\text{ME}_r}}$	0	3	3
$\sigma_{\tau_{\text{ME}_r}}$	3	1	0.1
$\mu_{\tau_{\text{PE}_r}}$	0	3	3
$\sigma_{\tau_{\text{PE}_r}}$	3	1	0.1
μ_{ϕ_r}	0	0	0
σ_{ϕ_r}	1	0.5	0.1
$\mu_{\psi_{\text{CP}_r}}$	0	0	0
$\sigma_{\psi_{\text{CP}_r}}$	3	1.5	0.5
$\mu_{\psi_{\text{CM}_r}}$	0	0	0
$\sigma_{\psi_{\text{CM}_r}}$	3	1.5	0.5

2.4.2 Inferred regulatory interactions in *M. abscessus*

Of the $N = 167,280$ regulator-target pairs considered in *M. abscessus*, under the uninformative parameterisation, BINDER identified 54 pairs across 5 transcription factors with a posterior 50th percentile $\theta_{r,t}^{50\%} > 0.9$ (Table 3). Of these 54 interactions, 24 are known to have validated orthologous regulatory interactions in *M. tuberculosis* as per ChIP-seq data (Figure 7); the number of interaction pairs almost doubles by reducing the threshold by 0.1 (102 pairs with 31 known orthologous interactions satisfying $\theta_{r,t}^{50\%} > 0.8$). In comparison, under the informative parameterisation, a similar effect was observed with 54 pairs with 21 known orthologous interactions satisfying $\theta_{r,t}^{50\%} > 0.9$. A more conservative effect was observed for the precise settings: 33 pairs across 28 transcription factors with a posterior 50th percentile $\theta_{r,t}^{50\%} > 0.9$. As expected, for all

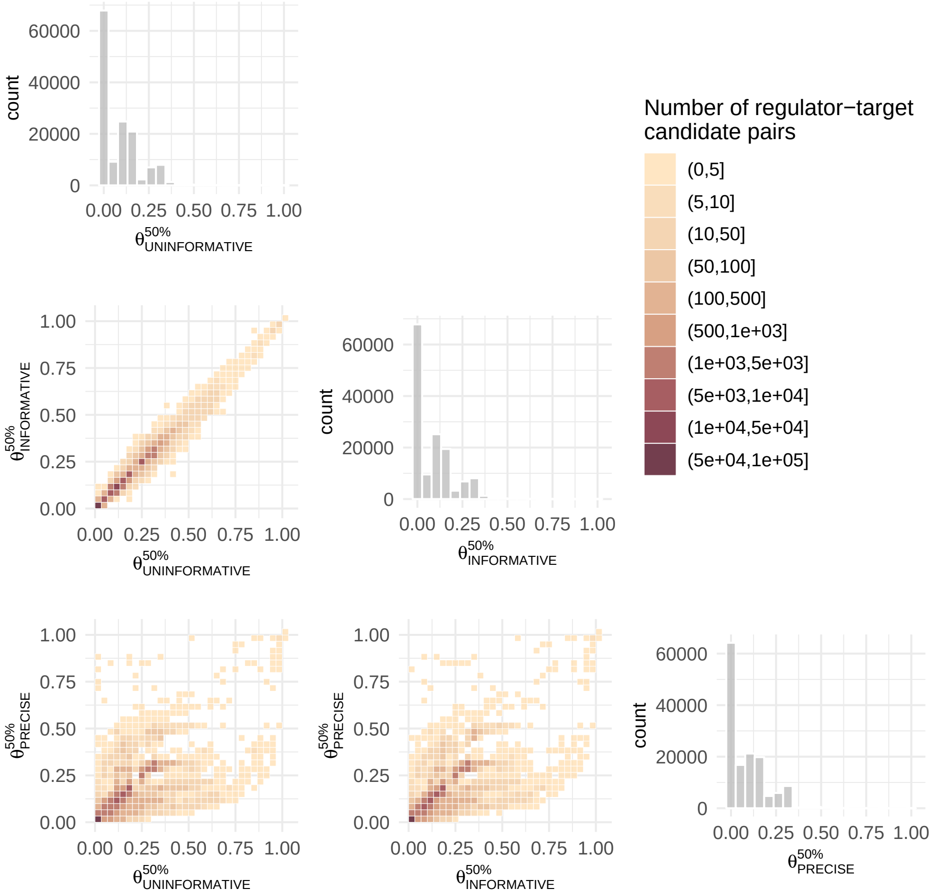

Figure 6: Heat map illustrating the similarity between mean predicted $\theta_{r,t}^{50\%}$ values achieved by BINDER under three distinct prior distribution parameterisations (uninformative, informative, precise) on the set of $N = 167,280$ regulator-target pairs.

parameterisations, the vast majority of posterior distributions of $\boldsymbol{\theta}$ were centred at low values, suggesting low levels of regulatory connectivity across the *M. abscessus* interactome; the mean 50th percentile for all of $\boldsymbol{\theta}$ was 0.085 (sd: 0.106) for the uninformative parameterisation and 0.087 (sd: 0.105) and 0.0885 (sd: 0.0995) for the informative and precise parameterisations respectively. It should be noted that in the benchmarking exercise (Section 2.3) we observed that the nominal value of a regulator-target pair's $\theta_{r,t}^{50\%}$ is not always as informative as its relative magnitude to $\{\theta_{r,1}, \ldots, \theta_{r,N}\}$. In general, whilst there were many instances of plausible conserved interactions, the results suggest evidence for many non-conserved interactions that may be unique to *M. abscessus*. Further, it can be observed that for a given regulator, many of the regulated genes appear to be spatially clustered along the genome (Figure 7). This observation lends support to the concept of gene colocalization arising as a means to affect efficient transcription (42; 43).

The parameter ζ_r in the auxiliary component influences the inferred probability of regulator-target interaction before any further regulator-target pair information is taken into account, with larger values of ζ_r meaning higher interaction probabilities. In this sense, each ζ_r is related to the ubiquity of regulation by regulator r across the genome. Under the uninformative parameterisation, we observed an average posterior mean of -6.63 across all regulator models (sd: 4.07). Hence, intuitively, conditional on the auxiliary data ME and PE being zero, the probability of a regulatory interaction is low.

The parameter τ_{ME_r} captures the influence the auxiliary ME data has on the prior mean of the inferred probability of a regulatory interaction between regulator r and target t, given all other covariates. Across all regulators, under the uninformative parameterisation, we observed

Table 3: Regulator-target pairs achieving a posterior $\theta_{r,t}^{50\%} > 0.9$ in *M. abscessus* by regulator under the uninformative parameterisation.

Regulator Locus Tag	Regulator Gene Name	Total Interactions	Conserved Interactions	Unconserved Interactions
MAB_0599	-	8	4	4
MAB_1678c	zur	15	8	7
MAB_4086	-	8	1	7
MAB_4270c	hspR	7	4	3
MAB_4449c	-	16	7	9

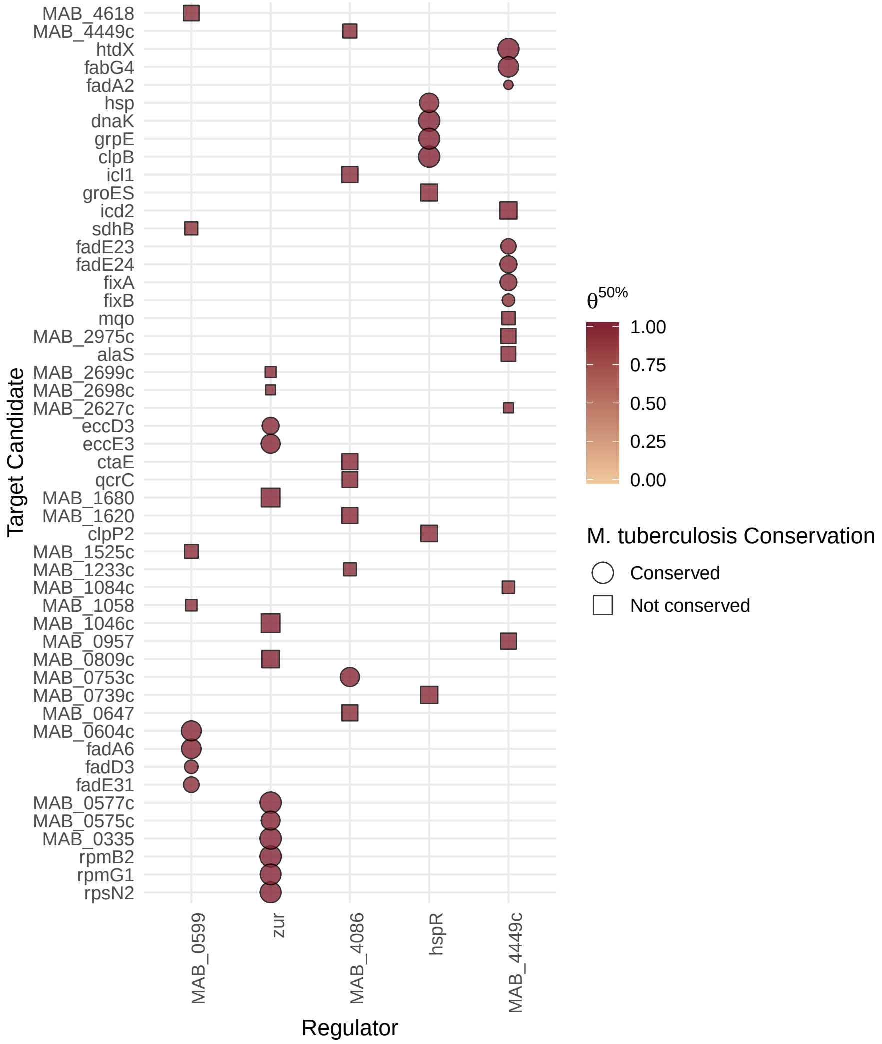

Figure 7: Abacus plot illustrating interaction candidates achieving $\theta_{r,t}^{50\%} > 0.9$ for the uninformative parameterisation; larger points are suggestive of less uncertainty; circles correspond to validated regulatory interactions in *M. tuberculosis*; shading corresponds to the posterior $\theta_{r,t}^{50\%}$ estimate. Regulators and targets are arranged by genomic position.

an average posterior mean for τ_{ME_r} of 1.43 (sd: 0.9982) (Figure 8). The parameter τ_{PE_r} has a similar interpretation for the auxiliary data PE. Across all regulators, under the uninformative parameterisation, we observed an average posterior mean for τ_{PE_r} of 1.95 (sd: 1.8981) (Figure 8). These results suggest that, on average, both ME and PE are positively correlated with the primary data in the likelihood. Given the phenomenon of genomic conservation, this is as we would expect and lends credence to the BINDER approach. Furthermore, although the mean posterior means for τ_{ME_r} and τ_{PE_r} are quite similar, the latter has larger variation suggesting higher volatility in the influence of PE than in the influence of ME.

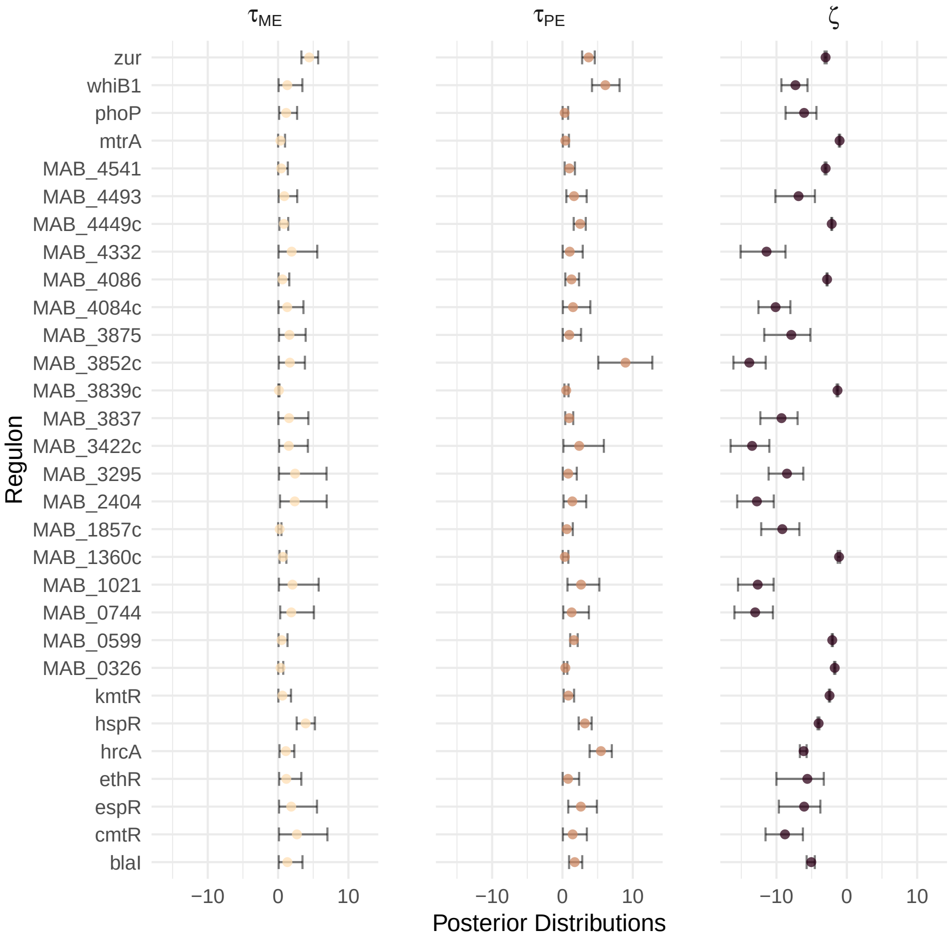

Figure 8: Central 95% of mass of the posterior distributions for τ_{ME_r}, τ_{PE_r} and ζ_r under the uninformative parameterisation with posterior means indicated by dots for each of the $R = 34$ regulators.

In terms of scale parameters, under the uninformative parameterisation, ϕ tended to have the lowest posterior mean values (average posterior mean of 1.12 with standard deviation 1.0067) (Figure 9). Both ψ_{CM_r} and ψ_{CP_r} yielded larger posterior mean estimates. In particular, under the uninformative parameterisation, ψ_{CM_r} yielded an average posterior mean of 4.23 (sd: 1.7713) and ψ_{CP_r} yielded an average posterior mean of 3.63 (sd: 1.4499), suggesting that the primary CM data tend to lie further from $\mathrm{logit}(\theta_{r,t})$ than CP (Figure 9). Also, the larger average posterior mean associated with ψ_{CM_r} compared with that of ψ_{CP_r} is intuitive, given the extra uncertainty associated with motif inference (comprised within CM) compared with validated orthologous interactions comprised within CP.

2.4.3 Interpretation of results: composition of the zur regulon

As an example of a putative discovery facilitated by BINDER, we examine the inferred regulon corresponding to the transcriptional regulator zur (MAB_1678c). The zur regulator present in *M. tuberculosis* and *M. abscessus* is a zinc-responsive transcription factor. Zinc is an essential element for life in many organisms (44). In addition to its role as a structural scaffold for many

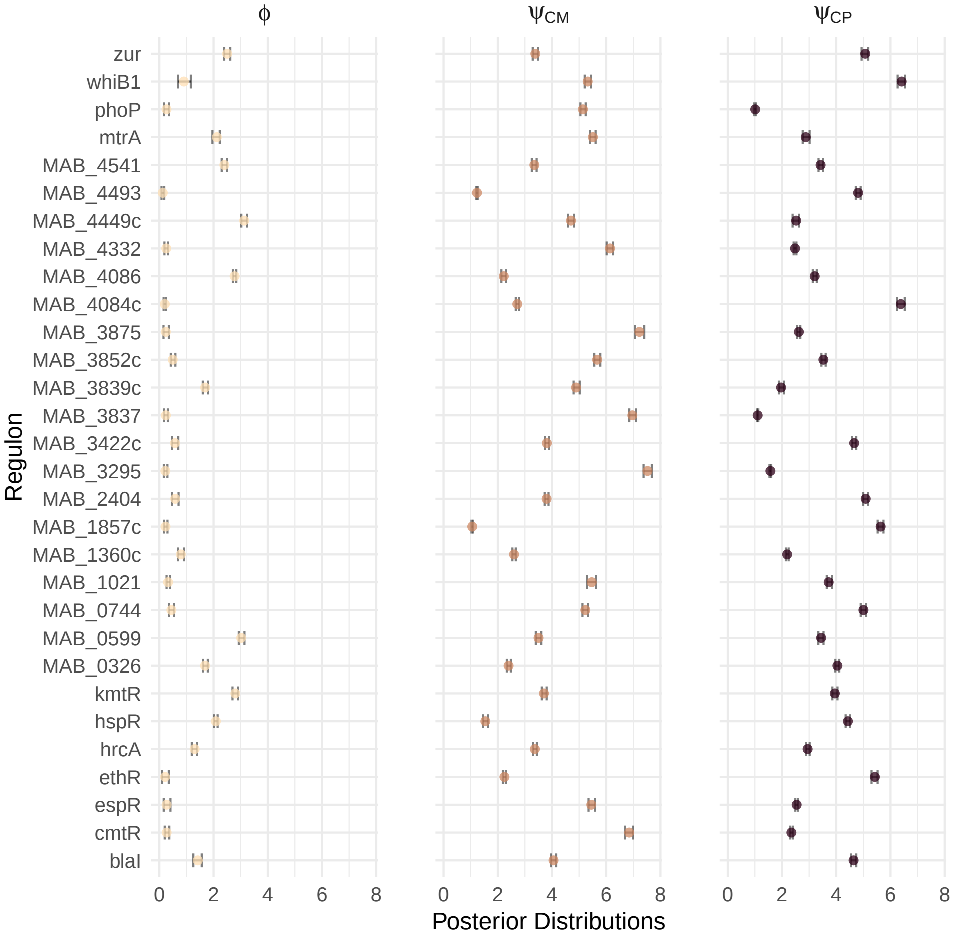

Figure 9: Central 95% of mass of posterior distributions for ϕ_r, ψ_{CM_r} and ψ_{CP_r} under the uninformative parameterisation with posterior mean values denoted by dots for each of the $R = 34$ regulators.

proteins, it fulfils a critical function as a frequent enzyme and DNA-binding protein cofactor (45). However, zinc can be toxic at high concentrations (46). For prokaryotes, efficient zinc acquisition, concentration and tolerance are critical processes for survival and pathogenicity (47). Zinc homeostasis in prokaryotes is achieved via cellular import and export, zinc binding, and zinc-sensing (47). Cellular zinc levels are maintained by importer and exporter proteins which are then regulated at the transcriptional level by several zinc-responsive transcription factors (48), including the zur regulator.

As per ChIP-seq results, the original regulon pertaining to zur in *M. tuberculosis* (Rv2359/furB) comprised 26 target genes (12 directly regulated targets); under the uninformative parameterisation, of these targets, 14 (53.8%) contained orthologs in *M. abscessus*. Using the cutoff criterion $\theta^{50\%}_{\text{zur},t} > 0.9$, BINDER suggested 15 target candidate genes in *M. abscessus* be considered valid targets of zur, 8 of which correspond to evidenced interactions in *M. tuberculosis*. Gene ontological analysis carried out on the putative targets provided intuitive insight, revealing up-regulated biological processes ($p \leq 0.05$) corresponding to metal ion transport.

BINDER also identified a number of interesting non-conserved putative targets for zur. For example, MAB_1046c, is annotated as a cobalamin synthesis protein. This is interesting as MAB_0335, one of the identified conserved targets, is also annotated as a cobalamin synthesis protein. This is perhaps owing to the role of cobalamin as a cofactor for cobalamin dependent methionine synthase in prokaryotes. Cobalamin dependent methionine synthase is involved in zinc ion binding (49). Further, MAB_2698c and its immediately adjacent neighbour MAB_2699c also yield high $\theta^{50\%}_{\text{zur},t}$ posterior estimates; gene ontology suggests that MAB_2699c, another unconserved putative target, is involved in pseudouridine synthesis/pseudouridine synthase activity; pseudouridine synthases catalyse the isomerisation of uridine to pseudouridine in RNA molecules and are thought to act as RNA chaperones. Intriguingly, pseudouridine synthase I (TruA) (50), one of the four distinct families of pseudouridine synthases, contains one atom of zinc essential for its native conformation and tRNA recognition (51). Another unconserved target is the PPE-like gene MAB_0809c; PPE genes are widely considered to play a key role in pathogenesis. Interestingly, phagosomes containing PPE genes found to disrupt lysosome-phagosome fusion have been shown to display differences in zinc levels relative to corresponding phagosomes containing PPE-knockout mutants (52). Another highly-probable unconserved interaction, MAB_1680, is annotated as a putative transmembrane protein. Given its association with zur, MAB_1680 is perhaps involved with zinc uptake in *M. abscessus*.

3 Discussion

In this work we have inferred the GRN in *M. abscessus* using the BINDER approach, the primary purpose of which is to infer the probability of pairwise interactions in a collection of regulator-target pairs. BINDER exploits experimental coexpression data in tandem with the property of genomic conservation to probabilistically infer a GRN in *M. abscessus*. To infer a GRN, BINDER proceeds by binding information from data in primary and auxiliary strata.

BINDER facilitates information sharing horizontally (by sharing parameters in the same layer of the model hierarchy) and vertically (by sharing of parameters in distinct strata of the hierarchy). The likelihood function assumes independence of the assumed logit-normal distributed primary data variables, conditional on the shared parameter of interest $\theta_{r,t}$, representing the probability of an interaction in the $(r,t)^{th}$ regulator-target pair. Further, the mean of this interaction probability's logit-normal distribution is informed by a linear function of the auxiliary data, serving as a proxy for genomic conservation information. Thus inference is strengthened through the borrowing of information across variables and strata.

With the exception of PE, the construction of all variables considered (i.e. ME, CM and CP) involves the choice of thresholds and/or decisions. For example, from the outset we have formed a TFBS-based module binary membership structure and an orthologous target binary membership structure, recorded in the auxiliary binary variables ME and PE respectively, on which the primary variables CM and CP rely. However, in order to circumvent potential loss of information associated with such hard membership, a "soft" approach using scale free topology or clustering coefficients may be worth exploring. Under these scenarios, the idea of membership has a continuous representation (15). Further, the auxiliary variable ME is derived from thresholding a p-value and as such is sensitive to the cutoff point ϵ selected. The BINDER approach also implements a further two threshold points δ_{CM} and δ_{CP}; clearly it is of paramount importance to choose these thresholds in an informed and careful manner. We have employed a hypergeometric framework for CM and CP, but any mapping to $[0,1]$ is possible. Again, topological overlap mapping or clustering coefficent mapping (15) are alternative approaches. With a view to foregoing the need to choose a threshold at all, simply mapping a regulator-target pair to the mean of its coexpression with members of the ME and PE modules is possible because the mean of a group of unsigned coexpressions will also lie in $[0,1]$; validation studies suggests that this approach, although convenient, does not perform quite as well as the hypergeometric framework.

It should be noted that, for our purposes, we had a relatively small-scale expression compendium with which to form our coexpression networks. Both the volume and diversity of RNA-seq conditions used to construct the coexpression networks may not be fully sufficient to computationally infer the entire GRN in *M. abscessus*. Small coexpression data sets are more likely to comprise noisy correlation results and similar experimental conditions have the effect of duplicating expression information leading to low numbers in terms of effective sample sizes. Similarly, for some regulators, we observed a lack of specificity in binding sites (owing to very long binding regions and small numbers of binding interactions); this has the effect of negatively impacting motif inference (i.e. false discovery of erroneous motifs). Naturally, more reliable data are preferable, however where data are less reliable, it is possible to account for this uncertainty through specification of the hyperparameters in the priors on the variable-specific parameters. Regardless, as the signal deteriorates (e.g. erroneous consensus motifs, inaccurate binding interactions), inference will suffer and thus it is important to ensure that all data sources are as accurate as possible. For the above reasons, it may be worthwhile to examine the more conservative BINDER parameterisations (i.e. the precise parameterisations) detailed above. This parameterisation implements a less diffuse prior distribution such that candidates lacking auxiliary support are less likely to achieve high $\theta_{r,t}$ estimates.

Through the course of this analysis, with a view to focusing on inferred highly probable regulator-target interactions, we have examined pairs for which the posterior median $\theta_{r,t}^{50\%} > 0.9$. However, the intention behind this model is not to define interaction probability on the basis of a single point estimate, but rather to provide a posterior distribution of $\theta_{r,t}$. This allows for a more nuanced analysis on interaction probability estimates than is typically provided by a simple binary classifier. Instead, we recommend that estimates are received in the context of the scientific question posed; varying the the number and severity of thresholds and tolerances will allow for differing results. Similarly, as noted in the fur regulon inference for *E. coli* explored in the benchmarking results, under certain scenarios BINDER estimates low values for all interaction candidates (both positive and negative cases); this is either due to influential hyperparameter settings and/or poor agreement between the auxiliary and primary data. However, even under these scenarios, BINDER can still estimate higher estimates for positive interaction cases. In such cases, as is good statistical practise, prior sensitivity analyses should be conducted or it may be worthwhile to consider regulator results individually.

One obvious limitation of any model that exploits conservation phenomena to perform inference in scarcely annotated organisms is that such a model can only make inference based on existing conservation data; indeed BINDER cannot infer interaction that may exist in *M. abscessus* on regulators not considered here. There are modelling approaches for “de novo” network inference that are based exclusively on coexpression analysis or other non-conservation based predictors, but such approaches can contain many false positives (53). Instead BINDER aims to overcome such issues by allowing coexpression-based data have partial influence on model inference. Moreover, while BINDER requires a consensus sequence motif and a collection of orthologous regulator-target interactions to perform inference, it is possible to run BINDER with a consensus sequence motif *or* a collection of orthologous interactions only. In this case, BINDER comprises one variable in the auxiliary stratum and one variable in the primary stratum.

One mechanism used by cells to refine and maintain transcription factor levels is autoregulation. It has been argued that the occurrence of autoregulation positively correlates with the developmental or physiological importance of the transcription factor (54). Given that any gene will have a perfect coexpression with itself, most expression-based approaches (such as GENIE3 and iRafNet) to GRN inference are unable to detect transcription factor autoregulation. For a given regulator, BINDER uses the coexpression profiles of a target gene with genes under the control of the regulator to inform the probability of a regulator-target interaction. BINDER does not examine the coexpression of the target candidate with regulator directly. As a result, BINDER is able to detect autoregulation.

For each regulator considered here, we applied the BINDER approach to all 4,920 annotated protein-coding genes in *M. abscessus*. However, in theory, BINDER could be applied to any desired subset of genes. With a view to accurately describing whole-population behaviour we recommend including all available data, albeit acknowledging the associated additional computational cost.

Pearson’s correlation was employed here as a measure of coexpression. Although there are other options, with a view to remaining conservative and reducing false positives, Pearson’s correlation gives high values when expression values are strongly linearly related. Common alternatives include the more flexible Spearman’s method, but often with increased flexibility comes an increase in less biologically significant relationships. Although use of Pearson's correlation can come at the cost of increased false negatives, studies have suggested that many coexpression relationships are linear and monotonic so this issue may be overstated (55).

Recent studies have suggested that implementing an ensemble approach to motif identification can improve detection results (56). BINDER could be extended to augment the number of motif search tools used in the analysis. Similarly, another suggestion might be to augment the number of proxy organisms from a single proxy organism to k proxy organisms, similar in vein to (24). A spike-and-slab prior distribution (57) for the associated model parameters would provide insight on the information contained in the individual proxy organisms. Furthermore, it is possible to extend the dimensionality of the primary stratum. In general, data that are binary or lie in $[0, 1]$ can be appended to the primary stratum: for example, the direct coexpression between a given regulator-target pair could be used to form a trivariate primary stratum. Although we have used exclusively binary variables in the auxiliary stratum, there is no restriction on the form of auxiliary data that can be modelled by BINDER.

It may be worthwhile to investigate the effect of incorporating more sophisticated levels of dependency in the BINDER model. Such dependencies could be based on operon comembership, on regulator family membership (e.g. the whiB-like family (58)), on target reoccurrence or on gene function using GO (59) or COG (60), for example. Here, we only consider the gene immediately downstream of a confirmed or putative TFBS to be under the regulation of the associated regulator. Recent studies suggest that operon organisation is dynamic and, hence, operon structures are capable of changing across conditions (61). However, given that BINDER considers not only the existence of a precedent interaction and/or motif match for a given candidate, but also the coexpression of that candidate with other candidates that do comprise a precedent interaction and/or motif match, BINDER is capable of detecting adjacent gene coregulation. Members of operon structures that are cotranscribed across all conditions considered will exhibit greater coexpression than those that are only cotranscribed under a fraction of conditions considered; as a result, BINDER is able to reflect that behaviour through the $\theta_{r,t}$ posteriors. Furthermore, it is possible to construct prior distribution parameterisations such that BINDER will tend to estimate higher $\theta_{r,t}$ median values for genes in cotranscribed structures if they comprise a precedent interaction and/or motif match; this may facilitate the determination of gene importance in cotranscribed structures. Owing to the lack of assumptions made by BINDER with respect to transcription start sites and operon co-membership, we expect that the results generated by BINDER will sufficiently aid in the generation of dynamic regulatory networks, as well as the understanding of transcriptional unit plasticity.

4 Conclusions

We have sought to determine the evidence for gene regulation in *M. abscessus* using a range of expression data from *M. abscessus* and experimentally validated regulatory network data from *M. tuberculosis*. We have demonstrated the extent to which there is a correlation between gene regulation in *M. tuberculosis* and transcriptome coexpression in *M. abscessus*. Our results imply not only strong genic conservation between *M. abscessus* and *M. tuberculosis* but also evidence of conservation with respect to the modes of transcriptomic control between these two organisms.

We have implemented a Bayesian modelling approach to quantifying the probability of an interaction across a collection of 167,280 regulatory-target pairs. Of these, 54 regulator-target pairs across 5 transcription factors, were inferred to have a posterior 50th percentile for $\theta_{r,t} > 0.9$ in *M. abscessus*.

The interactions identified in this study will form a valuable resource for further studies of transcriptional control in *M. abscessus* and in the family of *Mycobacteriaceae* more generally. Further, the BINDER framework is applicable across a wider range of organisms for which similar data are available.

5 Methods

5.1 Data

Given the paucity of data available from the primary organism *M. abscessus* (MAB), BINDER integrates data from a proxy organism *M. tuberculosis* (MTB) into the inferential procedure. Specifically, we leverage data from across orthologous ChIP-seq validated interactions in *M. tuberculosis* as proxy data and extract the primary data from 32 RNA-seq libraries across 16 distinct experimental conditions in *M. abscessus*. Thus we consider the set of all possible regulator-target interaction candidate pairs, arising from the set $R = 34$ orthologous regulators in *M. tuberculosis*, and $T = 4920$ target genes in the *M. abscessus* genome yielding $N = 167,280$ regulator-target pairs of interest.

5.1.1 Auxiliary data: motif evidence (ME) and precedent evidence (PE)

Motif Evidence: With respect to a given regulator r, the TFBS status of a target t is encoded through a binary variable termed motif evidence (ME). Specifically, for a regulator-target pair, ME takes the value 1 if the corresponding target contains a putative TFBS for the regulator's motif in its upstream region and a value of 0 otherwise. Here, the binding motif is assumed to be identical to the binding motif in the proxy organism.

With a view to determining regulator motifs, we extracted binding sequences using the NCBI *M. tuberculosis* (Accession: AL123456) complete chromosome sequence and annotation, S_{MTB}. The evidenced binding region coordinates were provided by ChIP-seq data sets ranging across several induced transcription factor experiments in *M. tuberculosis*. We subsequently categorised these binding sequences by regulator with a view to discovering binding sequence consensus motifs. The MEME motif discovery tool (62) was used to infer a single consensus binding motif M_r for each regulator $r \in R$: in particular, using a DNA alphabet, we searched on both strands seeking zero or one occurrence per binding sequence of a single consensus motif between 10 and 30 nucleotides long.

To find putative TFBSs for the derived motifs in the *M. abscessus* genome, we defined a sequence region U_t corresponding to the region -300nt to +50nt of the start of each target of interest $t \in T$. This interval size was chosen in light of the distribution of intergenic region lengths in the *M. abscessus* genome. In order to find putative TFBSs for each M_r, we searched in each U_t using the complete chromosome sequence and annotation S_{MAB} provided by NCBI for *M. abscessus* (Accession: NC010397). In the scenario that the most upstream coordinate of an immediately adjacent upstream gene was annotated to occur within 300nt of an upstream region of interest, the upstream region of interest was truncated to the most upstream coordinate of the upstream gene. To perform this search, we used the FIMO tool (63) to find the high-scoring upstream sequences with a q-value $\leq \epsilon = 0.1$. We provided a background file encoding 0-order nucleobase probabilities based on all upstream sequences of interest.

In summary, for each regulator-target pair (r, t) for $r = 1, \ldots, R$ and $t = 1, \ldots, T$ the motif evidence $\mathrm{ME}_{r,t}$ is computed where:

$$\mathrm{ME}_{r,t} = \begin{cases} 1 & \text{if for } M_r \text{ the FIMO q-value for } U_t \leq \epsilon \\ 0 & \text{otherwise.} \end{cases}$$

For a given regulator r, we refer to the set of all genes where $\mathrm{ME}_{r,t} = 1$ as the 'ME_r module'.

Precedent Evidence: The presence of an annotated orthologous regulator-target interaction in the proxy organism is encoded in the binary variable termed precedent evidence (PE). For a regulator-target pair, PE takes the value of 1 if such an orthologous interaction exists and takes the value of 0 otherwise.

Specifically, given both the proxy genome G_{MTB} and the primary genome of interest G_{MAB}, Orthologue (64) derived one-to-one orthologs were used to map orthologous regulator-target interactions from G_{MTB} to G_{MAB}. ChIP-seq data sets drawn from 34 induced transcription factor experiments in G_{MTB} were scanned for orthologous regulator-target interactions with respect to G_{MAB}; orthologous regulator-target pairs were subsequently grouped by regulator to derive a rudimentary orthology of regulons in G_{MAB}.

Thus, given the rudimentary orthology, for a given regulator r and target t:

$$\mathrm{PE}_{r,t} = \begin{cases} 1 & \text{if orthologous evidence of } r \text{ regulating } t \text{ in } G_{\mathrm{MTB}} \\ 0 & \text{otherwise.} \end{cases}$$

As in the ME case, for a given regulator r, we refer to the set of all genes where $\mathrm{PE}_{r,t} = 1$ as the 'PE_r module'.

5.1.2 Primary data: coexpression of motif and precedent evidence

Coexpression of Motif Evidence: Exploiting the property that genes sharing a common regulator exhibit strong coexpression (14), we computed a measure termed coexpression of motif evidence (CM). For a given regulator, using the motif derived from the proxy organism, CM quantifies the extent to which a target gene coexpresses with genes that have strong affinity for the putative regulator motif in the primary organism.

Specifically, for a regulator binding sequence motif M_r inferred from G_{MTB}, we define $\mathrm{CM}_{r,t}$ for a given gene regulator-target pair (r, t) in G_{MAB}. We define the reduced primary genome $G_{\mathrm{MAB},-O_t} = G_{\mathrm{MAB}} \setminus O_t$, where O_t is a t-inclusive set of genes in G_{MAB} that should not be used in the calculation of $\mathrm{CM}_{r,t}$. This set will naturally include t, but can contain any other genes that are not desired for calculation of $\mathrm{CM}_{r,t}$. The variable $\mathrm{CM}_{r,t}$ lies in $[0, 1]$, where values closer to 1 represent stronger correlation between expression levels of the target t with genes in $G_{\mathrm{MAB},-O_t}$ producing strong matches to the inferred sequence motif M_r. Specifically, for a regulator-target pair

$$\mathrm{CM}_{r,t} = \begin{cases} \text{hypergeometric}(a|b, c, d) & \text{for } a, b, d \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

where hypergeometric$(a|b, c, d)$ represents the cumulative distribution function of a hypergeometric random variable a with parameters b, c and d where, for some threshold δ_{CM},

- a is the number of genes in $G_{\mathrm{MAB},-O_t}$ that belong to the ME_r module and have an absolute expression correlation with gene $t > \delta_{\mathrm{CM}}$
- b is the number of genes in $G_{\mathrm{MAB},-O_t}$ exhibiting an absolute expression correlation with gene $t > \delta_{\mathrm{CM}}$
- c is the number of genes in $G_{\mathrm{MAB},-O_t}$ exhibiting an absolute expression correlation with gene $t \leq \delta_{\mathrm{CM}}$
- d is the number of genes in $G_{\mathrm{MAB},-O_t}$ that belong to the ME_r module.

A Benjamini and Hochberg adjustment (65) is applied to these probabilities to relax the observed polarisation of probabilities around 0 and 1; for a given regulator r, the adjustment is relative to all targets $t \in T$. We expect genes under the control of regulator r to coexpress strongly with members of the ME_r module. For our purposes, we vary the threshold such that each δ_{CM} is specific to each target. For a given target t, assuming $\mathrm{CX}_{i,j}$ represents the coexpression between genes i and j, we choose δ_{CM} to be equal to the 95th percentile of all values in the set $\{\mathrm{CX}_{t,g} \text{ for } g \in G_{\mathrm{MAB},-O_t}\}$.

Coexpression of Precedent Evidence: Analogous to CM, we develop a score of coexpression of precedent evidence, CP. For a given regulator, CP quantifies the extent to which a target gene coexpresses with orthologs of genes comprising regulator-target interactions in the proxy organism.

Specifically, for regulator r, we define the regulon P_r as the collection of orthologous interactions annotated in G_{MTB}. For a given gene regulator-target pair (r, t) in G_{MAB} the variable $\mathrm{CP}_{r,t}$ is defined on the interval $[0,1]$, where values closer to 1 represent stronger expression correlation of gene t with orthologs of genes from P_r in $G_{\mathrm{MAB},-O_t}$. That is,

$$\mathrm{CP}_{r,t} = \begin{cases} \text{hypergeometric}(a|b,c,d) & \text{for } a,b,d \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

where, for a threshold δ_{CP}

- a is the number of genes in $G_{\mathrm{MAB},-O_t}$ that belong to the PE_r module and have an absolute expression correlation with gene $t > \delta_{\mathrm{CP}}$
- b is the number of genes in $G_{\mathrm{MAB},-O_t}$ containing an ortholog in G_{MTB} and exhibit an absolute expression correlation with gene $t > \delta_{\mathrm{CP}}$
- c is the number of genes in $G_{\mathrm{MAB},-O_t}$ containing an ortholog in G_{MTB} and exhibit an absolute expression correlation with gene $t \leq \delta_{\mathrm{CP}}$
- d is the number of genes in $G_{\mathrm{MAB},-O_t}$ that belong to the PE_r module.

Again, the probabilities are subject to Benjamini and Hochberg adjustment relative to all target candidates $t \in T$. We expect genes under the control of regulator r to coexpress strongly with members of the PE_r module. Thus again we choose δ_{CP} to be equal to the 95th percentile of all values in the set $\{\mathrm{CX}_{t,g} \text{ for } g \in G_{\mathrm{MAB},-O_t}\}$.

With a view to quantifying coexpression in G_{MAB}, the expression profiles (using RPKM (66)) of all genes constituting the NCBI GenBank annotation for the G_{MAB} genome were computed across 32 RNA-seq libraries (comprising 16 distinct experimental conditions) elicited from a range of astringent response and control experiments. In order to compute the corresponding coexpression profiles, we generated the unsigned Pearson correlation coefficient of all possible pairwise annotated gene-pair combinations. All read files were aligned using Bowtie (version 1.2.2) (67) and totalled using Samtools (version 1.7) (68). RNA-seq libraries can be found on NCBI's Gene Expression Omnibus (Accession: GSE78787).

5.2 The BINDER model for inferring a GRN

Borrowing strength across the primary and auxiliary data sets, we computationally infer the GRN for *M. abscessus* through a novel statistical modelling approach: BayesIan gene regulatory Networks inferreD via gene coExpression and compaRative genomics (BINDER). BINDER is a Bayesian hierarchical model that appositely models the type and structure of both the primary and auxiliary data to infer the probability of a regulatory interaction between a regulator-target pair candidate. Each of $N = |R| \times |T|$ observations is a regulator and target candidate pair (r,t) from the set of regulators R and the set of target candidates T in the *M. abscessus* genome. Interest lies in the probability $\theta_{r,t}$ of there being an interaction between regulator r and target t. Thus, inferring $\theta_{r,t}$ facilitates inference of the *M. abscessus* GRN.

As stated, BINDER integrates primary data from *M. abscessus* with data from the proxy organism *M. tuberculosis*. Specifically, the variables CM and CP (Section 5.1.2) constitute the primary data stratum whilst ME and PE (Section 5.1.1) constitute the auxiliary stratum. As BINDER is a Bayesian hierarchical model, the auxiliary data inform the prior distribution for each $\theta_{r,t}$; the posterior distribution for each $\theta_{r,t}$ is then updated by accounting for the primary data.

To define the likelihood function of the BINDER model we appositely model the primary data type and assume logit-normal distributions for CM and CP. As such, in the case where $\mathrm{CM}_{r,t}$ or $\mathrm{CP}_{r,t}$ were 0 or 1, they were increased or decreased respectively by a small factor (10^{-4}). Further we assume, given $\theta_{r,t}$, the regulator-target pairs and primary variables are conditionally independent:

$$\mathcal{L}(\boldsymbol{\theta}, \psi_{\mathrm{CM}}, \psi_{\mathrm{CP}}|\mathrm{CM},\mathrm{CP}) \;=\; \prod_{\substack{r\in R\\ t\in T}} \mathcal{N}_l\{\mathrm{CM}_{r,t}|\mathrm{logit}(\theta_{r,t}), \psi_{\mathrm{CM}_r}\}\mathcal{N}_l\{\mathrm{CP}_{r,t}|\mathrm{logit}(\theta_{r,t}), \psi_{\mathrm{CP}_r}\}$$

Here $\mathcal{N}_l(x|a,b)$ denotes the logit-normal distribution of x with location and standard deviation parameters a and b respectively. The location parameter is common across the distributions for CM and CP. This shared parameter enables the borrowing of information across variables, in addition to facilitating tractability through the conditional independence assumption. The conditional independence assumption is widely employed in other settings, such as latent class analysis (69; 70).

As with any Bayesian hierarchical model, prior distributions are specified on the BINDER model parameters. For each $\theta_{r,t}$ we posit a logistic normal prior such that $\text{logit}(\theta_{r,t}) \sim \mathcal{N}(\gamma_{r,t}, \phi)$ where ϕ is the standard deviation parameter controlling the level of dispersion around the mean. The mean $\gamma_{r,t}$ is informed by the auxiliary data ME and PE on the regulator-target pair (r,t) through a linear model. Specifically:

$$\gamma_{r,t} \quad = \quad \zeta_r + \tau_{\text{ME}_r}\text{ME}_{r,t} + \tau_{\text{PE}_r}\text{PE}_{r,t} \tag{1}$$

Independent priors are then posited on the parameters in (1) such that the intercept $\zeta_r \sim \mathcal{N}(\mu_\zeta, \sigma_\zeta)$ and a truncated normal prior is assumed on the slope parameters: $\tau_{k_r} \sim \mathcal{N}_{(0,\infty)}(\mu_{\tau_k}, \sigma_{\tau_k})$ for $k \in \{\text{ME}, \text{PE}\}$. This truncated normal prior with mass on the positive real line reflects the assumption that the presence of regulation in regulator-target pair (r,t) in the proxy organism is suggestive of the presence of such regulation in *M. abscessus.* To complete the model setup, prior distributions are placed on the scale parameters such that $\psi_{l_r} \sim \mathcal{N}_{(0,\infty)}(\mu_{\psi_l}, \sigma_{\psi_l})$ for $l \in \{\text{CP}, \text{CM}\}$. The hyperparameters of all the specified prior distributions must be set by the practitioner and their values are potentially influential; sensitivity of inference to their choice is explored in Section 2.4.1.

In order to infer the GRN for *M. abscessus*, the set of parameters $\{\theta_{r,t} : r \in R, t \in T\}$ are of primary interest. Thus the required posterior distribution is

$$p(\boldsymbol{\theta}|\text{CM}, \text{CP}, \text{ME}, \text{PE}, \boldsymbol{\mu}, \boldsymbol{\sigma}) \;=\; \int_{\boldsymbol{\tau}} \dots \int_{\boldsymbol{\psi}} p(\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\phi}, \boldsymbol{\tau}, \boldsymbol{\zeta}|\text{CM}, \text{CP}, \text{ME}, \text{PE}, \boldsymbol{\mu}, \boldsymbol{\sigma}) d\boldsymbol{\psi} d\boldsymbol{\phi} d\boldsymbol{\zeta} d\boldsymbol{\tau}$$

This posterior distribution is explored using Stan (71), a state-of-the-art platform for statistical modelling and computation for large data sets that employs Hamiltonian Monte Carlo methods (72) to draw samples from the posterior distribution of interest. An illustration of the BINDER model is provided in Figure 10.

Abbreviations

AUC: Area under curve; *B. subtilis*: *Bacillus subtilis*; BINDER: BayesIan gene regulatory Networks inferreD via gene coExpression and compaRative genomics; ChIP-Seq: Chromatin immunoprecipitation followed by sequencing; CM: Coexpression of motif evidence; CP: Coexpression of precedent evidence; DNA: Deoxyribonucleic acid; *E. coli*: *Escherichia coli*; GRN: Gene regulatory network; *L. monocytogenes*: *Listeria monocytogenes*; *M. abscessus*: *Mycobacterium abscessus*; *M. tuberculosis*: *Mycobacterium tuberculosis*; MAB: *Mycobacterium abscessus*; MAD: Mean absolute deviation; ME: Motif evidence; MTB: *Mycobacterium tuberculosis*; NCBI: National Center for Biotechnology Information; PE: Precedent evidence; PPE: proline-proline-glutamate; RPKM: Reads per kilobase per million; *P. aeruginosa*: *Pseudomonas aeruginosa*; RBB: Reciprocal-best-BLAST; RNA: Ribonucleic acid; RNA-seq: RNA sequencing; ROC: receiver operating characteristic; SSD: Supporting-species-divergence; TFBS: Transcription factor binding site; tRNA: Transfer ribonucleic acid;

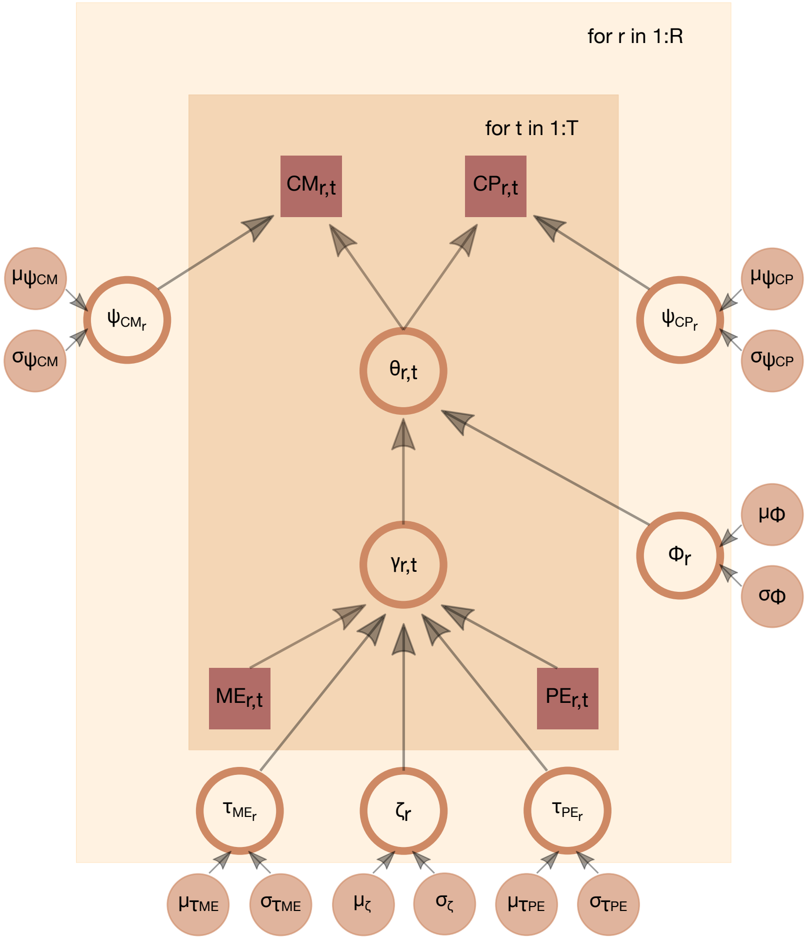

Figure 10: Graphical representation of the hierarchical BINDER model; squares correspond to observed data, large discs correspond to random parameters and small discs correspond to fixed hyperparameters; the surrounding boxes denote observation-specific parameters and data.

Declarations

Availability of data and materials

An implementation of the BINDER approach is available as an R package at github.com/ptrcksn/BINDER. The datasets generated and analysed in the current study are available at github.com/ptrcksn/BINDER_paper_analysis.

Funding

This research is supported by The Wellcome Trust (grant no. 099817/Z/12/Z) and by Science Foundation Ireland (grant no. 12/RC/2289_P2).

References

[1] Nessar R, Cambau E, Reyrat JM, Murray A, Gicquel B. Mycobacterium abscessus: a new antibiotic nightmare. J. Antimicrob. Chemother. 2012; doi:10.1093/jac/dkr578. 2

[2] Lee MR, Sheng WH, Hung CC, Yu CJ, Lee LN, Hsueh PR. Mycobacterium abscessus Complex Infections in Humans. Emerging Infect. Dis. 2015; doi:10.3201/2109.141634. 2

[3] Baranyai Z, Krtk M, Vinov J, Szab N, Senoner Z, Horvti K, Stola?kov J, Dvid S, B?sze S. Combating highly resistant emerging pathogen Mycobacterium abscessus and Mycobacterium tuberculosis with novel salicylanilide esters and carbamates. Eur J Med Chem. 2015; doi:10.1016/j.ejmech.2015.07.001. 2, 3

[4] Miranda-CasoLuengo AA, Staunton PM, Dinan AM, Lohan AJ, Loftus BJ. Functional characterization of the Mycobacterium abscessus genome coupled with condition precise transcriptomics reveals conserved molecular strategies for host adaptation and persistence. BMC Genomics. 2016; doi:10.1186/s12864-016-2868-y. 2

[5] Kili S, White ER, Sagitova DM, Cornish JP, Erill I. CollecTF: a database of experimentally validated transcription factor-binding sites in Bacteria. Nucleic Acids Res. 2014; doi:10.1093/nar/gkt1123 2, 7

[6] Santos-Zavaleta A, Salgado H, Gama-Castro S, Snchez-Prez M, Gmez-Romero L, Ledezma-Tejeida D, Garca-Sotelo JS, Alquicira-Hernndez K, Muiz-Rascado LJ, Pea-Loredo P, Ishida-Gutirrez C, Velzquez-Ramrez DA, Del Moral-Chvez V, Bonavides-Martnez C, Mndez-Cruz CF, Galagan J, Collado-Vides J. RegulonDB v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in E. coli K-12. Nucleic Acids Res. 2018; doi:10.1093/nar/gky1077. 2, 7

[7] Darmostuk M, Rimpelova S, Gbelcova H, Ruml T. Current approaches in SELEX: An update to aptamer selection technology. Biotechnol Adv. 2015; doi:10.1016/j.biotechadv.2015.02.008. 2

[8] Mundade R, Ozer HG, Wei H, Prabhu L, Lu T. Role of ChIP-seq in the discovery of transcription factor binding sites, differential gene regulation mechanism, epigenetic marks and beyond. Cell Cycle. 2014; doi:10.4161/15384101.2014.949201. 2

[9] Doniger SW, Huh J, Fay JC. Identification of functional transcription factor binding sites using closely related Saccharomyces species. Genome Res. 2005;15(5):701-9. 2

[10] Koonin EV. Orthologs, paralogs, and evolutionary genomics. Annu. Rev. Genet. 2005;39:309-38. 2

[11] Van de Velde J, Van Bel M, Vaneechoutte D, Vandepoele K. A Collection of Conserved Noncoding Sequences to Study Gene Regulation in Flowering Plants. Plant Physiol. 2016; doi:10.1104/pp.16.00821. 2

[12] Van de Velde J, Heyndrickx KS, Vandepoele K. Inference of transcriptional networks in Arabidopsis through conserved noncoding sequence analysis. Plant Cell. 2014; doi:10.1105/tpc.114.127001. 2

[13] Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P. Inferring regulatory networks from expression data using tree-based methods. PLoS One. 2010; doi: 10.1371/journal.pone.0012776. 2

[14] Wang YX, Huang H. Review on statistical methods for gene network reconstruction using expression data. J Theor Biol. 2014; doi:10.1016/j.jtbi.2014.03.040. 2, 20

[15] Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008; doi: 10.1186/1471-2105-9-559. 2, 16

[16] Li J, Overall CC, Johnson RC, Jones MB, McDermott JE, Heffron F, Adkins JN, Cambronne ED. ChIP-Seq Analysis of the σE Regulon of Salmonella enterica Serovar Typhimurium Reveals New Genes Implicated in Heat Shock and Oxidative Stress Response. PLoS ONE. 2015; doi:10.1371/journal.pone.0138466. 3

[17] Peano C, Wolf J, Demol J, Rossi E, Petiti L, De Bellis G, Geiselmann J, Egli T, Lacour S, Landini P. Characterization of the Escherichia coli σ(S) core regulon by Chromatin Immunoprecipitation-sequencing (ChIP-seq) analysis. Sci Rep. 2015; doi:10.1038/srep10469. 3

[18] Jaini S, Lyubetskaya A, Gomes A, Peterson M, Park ST, Raman S, Schoolnik G, Galagan J. Transcription Factor Binding Site Mapping Using ChIP-Seq. Microbiol Spectr. 2014; doi:10.1128/microbiolspec.MGM2-0035-2013. 3

[19] Landick R, Krek A, Glickman MS, Socci ND, Stallings CL. Genome-Wide Mapping of the Distribution of CarD, RNAP σ^A, and RNAP β on the Mycobacterium smegmatis Chromosome using Chromatin Immunoprecipitation Sequencing. Genom Data. 2014;2:110-113. 3

[20] Angelini C, Costa V. Understanding gene regulatory mechanisms by integrating ChIP-seq and RNA-seq data: statistical solutions to biological problems. Front Cell Dev Biol. 2014; doi:10.3389/fcell.2014.00051. 3

[21] Petralia F, Wang P, Yang J, Tu Z. Integrative random forest for gene regulatory network inference. Bioinformatics. 2015; doi:10.1093/bioinformatics/btv268. 3, 7

[22] Banf M, Rhee SY. Enhancing gene regulatory network inference through data integration with markov random fields. Sci Rep. 2017; doi: 10.1038/srep41174. 3

[23] Chouvardas P, Kollias G, Nikolaou C. Inferring active regulatory networks from gene expression data using a combination of prior knowledge and enrichment analysis. BMC Bioinformatics. 2016; doi: 10.1186/s12859-016-1040-7. 3

[24] Waltman P, Kacmarczyk T, Bate AR, Kearns DB, Reiss DJ, Eichenberger P, Bonneau R. Multi-species integrative biclustering. Genome Biol. 2010; doi:10.1186/gb-2010-11-9-r96. 3, 18

[25] Chen G, Jensen ST, Stoeckert CJ Jr.. Clustering of genes into regulons using integrated modeling-COGRIM. Genome Biol. 2007; doi:10.1186/gb-2007-8-1-r4. 3

[26] Chen X, Gu J, Wang X, Jung JG, Wang TL, Hilakivi-Clarke L, Clarke R, Xuan J. CRNET: an efficient sampling approach to infer functional regulatory networks by integrating large-scale ChIP-seq and time-course RNA-seq data. Bioinformatics. 2018; doi:10.1093/bioinformatics/btx827. 3

[27] Galagan JE, Minch K, Peterson M, Lyubetskaya A, Azizi E, Sweet L, Gomes A, Rustad T, Dolganov G, Glotova I, Abeel T, Mahwinney C, Kennedy AD, Allard R, Brabant W, Krueger A, Jaini S, Honda B, Yu WH, Hickey MJ, Zucker J, Garay C, Weiner B, Sisk P, Stolte C, Winkler JK, Van de Peer Y, Iazzetti P, Camacho D, Dreyfuss J, Liu Y, Dorhoi A, Mollenkopf HJ, Drogaris P, Lamontagne J, Zhou Y, Piquenot J, Park ST, Raman S, Kaufmann SH, Mohney RP, Chelsky D, Moody DB, Sherman DR, Schoolnik GK. The Mycobacterium tuberculosis regulatory network and hypoxia. Nature. 2013; doi:10.1038/nature12337. 3, 4

[28] Snel B, van Noort V, Huynen MA. Gene co-regulation is highly conserved in the evolution of eukaryotes and prokaryotes. Nucleic Acids Res. 2004; 32(16):4725-31. 4

[29] Okuda S, Kawashima S, Goto S, Kanehisa M. Conservation of gene co-regulation between two prokaryotes: Bacillus subtilis and Escherichia coli. Genome Inform. 2005; 16(1):116-24. 4

[30] Nyamundanda G, Gormley IC, Brennan L A dynamic probabilistic principal components model for the analysis of longitudinal metabolomics data J. Royal Stat. Soc. 2014; doi:10.1111/rssc.12060. 5

[31] Jang J, Hur HG, Sadowsky MJ, Byappanahalli MN, Yan T, Ishii S. Environmental Escherichia coli: ecology and public health implications-a review. J Appl Microbiol. 2017; doi:10.1111/jam.13468. 6

[32] Earl AM, Losick R, Kolter R. Ecology and genomics of Bacillus subtilis. Trends Microbiol. 2008; doi:10.1016/j.tim.2008.03.004. 6

[33] de Lorenzo V. Pseudomonas aeruginosa: the making of a pathogen. Environ Microbiol. 2015; doi:10.1111/1462-2920.12620. 6

[34] Listeria monocytogenes, a food-borne pathogen. Farber JM, Peterkin PI. Microbiol Rev. 1991; 55(3): 476?511. 6

[35] Harrison A, Santana EA, Szelestey BR, Newsom DE, White P, Mason KM. Ferric Uptake Regulator and Its Role in the Pathogenesis of Nontypeable Haemophilus influenzae Infect Immun. 2013; doi:10.1128/IAI.01227-12. 7

[36] Fornelos N, Browning DF, Butala M. The Use and Abuse of lexA by Mobile Genetic Elements. Trends Microbiol. 2016; doi:10.1016/j.tim.2016.02.009. 7

[37] Butala M, Zgur-Bertok D, Busby SJ. The bacterial lexA transcriptional repressor. Cell Mol Life Sci. 2009; doi:10.1007/s00018-008-8378-6. 7

[38] Zhu B, Stlke J. SubtiWiki in 2018: from genes and proteins to functional network annotation of the model organism Bacillus subtilis Nucleic Acids Res. 2017; doi:10.1093/nar/gkx908. 7

[39] Meysman P, Sonego P, Bianco L, Fu Q, Ledezma-Tejeida D, Gama-Castro S, Liebens V, Michiels J, Laukens K, Marchal K, Collado-Vides J, Engelen K. COLOMBOS v2.0: an ever expanding collection of bacterial expression compendia. Nucleic Acids Res. 2014; doi:10.1093/nar/gkt1086. 7

[40] Condition-dependent transcriptome reveals high-level regulatory architecture in Bacillus subtilis. Nicolas P, Mder U, Dervyn E, Rochat T, Leduc A, Pigeonneau N, Bidnenko E, Marchadier E, Hoebeke M, Aymerich S, Becher D, Bisicchia P, Botella E, Delumeau O, Doherty G, Denham EL, Fogg MJ, Fromion V, Goelzer A, Hansen A, Hrtig E, Harwood CR, Homuth G, Jarmer H, Jules M, Klipp E, Le Chat L, Lecointe F, Lewis P, Liebermeister W, March A, Mars RA, Nannapaneni P, Noone D, Pohl S, Rinn B, Rgheimer F, Sappa PK, Samson F, Schaffer M, Schwikowski B, Steil L, Stlke J, Wiegert T, Devine KM, Wilkinson AJ, van Dijl JM, Hecker M, Vlker U, Bessires P, Noirot P. Science. 2012; doi:10.1126/science.1206848. 7

[41] Hubert L, Arabie P. Comparing Partitions. Journal of the Classification. 1985; 2:193-218. 11

[42] Michalak P. Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes. Genomics. 2008;91(3):243-8. 12

[43] Pannier L, Merino E, Marchal K, Collado-Vides J. Effect of genomic distance on coexpression of coregulated genes in E. coli. PLoS One. 2017; doi:10.1371/journal.pone.0174887. 12

[44] Mikhaylina A, Ksibe AZ, Scanlan DJ, Blindauer CA. Bacterial zinc uptake regulator proteins and their regulons. Biochem Soc Trans. 2018; doi:10.1042/BST20170228. 14

[45] Vallee BL, Falchuk KH. The biochemical basis of zinc physiology. Physiol Rev. 1993; doi:10.1152/physrev.1993.73.1.79. 15

[46] Blencowe DK, Morby AP. Zn(II) metabolism in prokaryotes. FEMS Microbiol Rev. 2003; doi:10.1016/S0168-6445(03)00041-X. 15

[47] Capdevila DA, Wang J, Giedroc DP. Bacterial Strategies to Maintain Zinc Metallostasis at the Host-Pathogen Interface. J Biol Chem. 2016; doi:10.1074/jbc.R116.742023. 15

[48] Shin JH, Helmann JD. Molecular logic of the zur-regulated zinc deprivation response in Bacillus subtilis. Nat Commun. 2016; doi:10.1038/ncomms12612. 15

[49] Pejchal R, Ludwig ML. Cobalamin-independent methionine synthase (MetE): a face-to-face double barrel that evolved by gene duplication. PLoS Biol. 2005; doi:10.1371/journal.pbio.0030031. 15

[50] Ramamurthy V, Swann SL, Spedaliere CJ, Mueller EG. Role of cysteine residues in pseudouridine synthases of different families. Biochemistry. 1999; 38(40):13106-11. 16

[51] Arluison V, Hountondji C, Robert B, Grosjean H. Transfer RNA-pseudouridine synthetase Pus1 of Saccharomyces cerevisiae contains one atom of zinc essential for its native conformation and tRNA recognition. Biochemistry. 1998; 37(20):7268-76. 16

[52] Samradhni S Jha, Lia Danelishvili, Dirk Wagner, Jrg Maser, Yong-jun Li, Ivana Moric, Steven Vogt, Yoshitaka Yamazaki, Barry Lai, Luiz E Bermudez. Virulence-related Mycobacterium avium subsp hominissuis MAV_2928 gene is associated with vacuole remodeling in macrophages BMC Microbiol. 2010; doi:10.1186/1471-2180-10-100 16

[53] Song WM, Zhang B. Multiscale Embedded Gene Co-expression Network Analysis. PLoS Comput Biol. 2015; doi:10.1371/journal.pcbi.1004574. 17

[54] Crews ST, Pearson JC. Transcriptional autoregulation in development. Curr Biol. 2009; doi:10.1016/j.cub.2009.01.015 17

[55] Song L, Langfelder P, Horvath S. Comparison of co-expression measures: mutual information, correlation, and model based indices. BMC Bioinformatics. 2012; doi:10.1186/1471-2105-13-328. 18

[56] Lihu A, Holban S. A review of ensemble methods for de novo motif discovery in ChIP-Seq data. Brief Bioinform. 2015; doi:10.1093/bib/bbv022. 18

[57] Ishwaran H, Rao SJ. Spike and slab variable selection: Frequentist and Bayesian strategies. Ann. Statist. 2005; doi:10.1214/009053604000001147. 18

[58] Alam MS, Garg SK, Agrawal P. Studies on structural and functional divergence among seven WhiB proteins of Mycobacterium tuberculosis H37Rv. FEBS J. 2009; doi:10.1111/j.1742-4658.2008.06755.x. 18

[59] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25(1):25-9. 18

[60] Tatusov RL, Galperin MY, Natale DA, Koonin EV. The COG database: a tool for genome-scale analysis of protein functions and evolution Nucleic Acids Res. 2000;28(1): 33-6. 18

[61] Fortino V, Tagliaferri R, Greco D. CONDOP: an R package for CONdition-Dependent Operon Predictions. Bioinformatics. 2016;32(20):3199-3200. 18

[62] Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol. 1994;2:28-36. 19

[63] Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics. 2011; doi:10.1093/bioinformatics/btr064. 19

[64] Whiteside MD, Winsor GL, Laird MR, Brinkman FS. OrthologeDB: a bacterial and archaeal orthology resource for improved comparative genomic analysis. Nucleic Acids Res. 2013; doi:10.1093/nar/gks1241. 4, 19

[65] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Royal Stat Soc. Series B, 1995; 57:289?300. 20

[66] Li P, Piao Y, Shon HS, Ryu KH. Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-Seq data. BMC Bioinformatics. 2015; doi:10.1186/s12859-015-0778-7. 21

[67] Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009; doi:10.1186/gb-2009-10-3-r25. 21

[68] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; doi:10.1093/bioinformatics/btp352. 21

[69] Linzer, D. A., Lewis, J. B. poLCA: An R package for polytomous variable latent class analysis. Journal of statistical software. 2011; 42(10):1–29 22

[70] White, A. and Murphy, T. B. BayesLCA: An R package for Bayesian latent class analysis. Journal of Statiscal Software. 2014; 61(13):1–28. 22

[71] Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, Brubaker M, Guo J, Li P, and Riddell A. Stan: A probabilistic programming language. J. Stat. Softw. 2017; doi:10.18637/jss.v076.i01. 22

[72] Duane S, Kennedy AD, Pendleton BJ, Roweth D. Hybrid Monte Carlo. Phys. Lett. 1987; doi:10.1016/0370-2693(87)91197-X. 22