Wisdom of crowds for robust gene network inference

Daniel Marbach[1,2,11], James C Costello[3–5,11], Robert Küffner[6,11], Nicole M Vega[3–5], Robert J Prill[7], Diogo M Camacho[3–5,10], Kyle R Allison[3–5], The DREAM5 Consortium[8], Manolis Kellis[1,2], James J Collins[3–5,9] & Gustavo Stolovitzky[7]

Reconstructing gene regulatory networks from high-throughput data is a long-standing challenge. Through the Dialogue on Reverse Engineering Assessment and Methods (DREAM) project, we performed a comprehensive blind assessment of over 30 network inference methods on *Escherichia coli*, *Staphylococcus aureus*, *Saccharomyces cerevisiae* and *in silico* microarray data. We characterize the performance, data requirements and inherent biases of different inference approaches, and we provide guidelines for algorithm application and development. We observed that no single inference method performs optimally across all data sets. In contrast, integration of predictions from multiple inference methods shows robust and high performance across diverse data sets. We thereby constructed high-confidence networks for *E. coli* and *S. aureus*, each comprising ~1,700 transcriptional interactions at a precision of ~50%. We experimentally tested 53 previously unobserved regulatory interactions in *E. coli*, of which 23 (43%) were supported. Our results establish community-based methods as a powerful and robust tool for the inference of transcriptional gene regulatory networks.

'The wisdom of crowds' refers to the phenomenon in which the collective knowledge of a community is greater than the knowledge of any individual[1]. Based on this concept, we developed a community approach to address one of the long-standing challenges in molecular and computational biology, which is to uncover and model gene regulatory networks. Genome-scale inference of transcriptional gene regulation has become possible with the advent of high-throughput technologies such as microarrays and RNA sequencing, as they provide snapshots of the transcriptome under many tested experimental conditions. From these data, the challenge is to computationally predict direct regulatory interactions between a transcription factor and its target genes; the aggregate of all predicted interactions comprises the gene regulatory network. A wide range of network inference methods have been developed to address this challenge, from those exclusive to gene-expression data[2,3] to methods that integrate multiple classes of data[4–7]. These approaches have been

successfully used to address many biological problems[8–11], yet when applied to the same data, they can generate disparate sets of predicted interactions[2,3].

Understanding the advantages and limitations of different network inference methods is critical for their effective application in a given biological context. The DREAM project is a framework to enable such an assessment through standardized performance metrics and common benchmarks[12] (http://www.the-dream-project.org/). DREAM is organized around annual challenges, whereby the community of network inference experts is solicited to run their algorithms on benchmark data sets, participating teams submit their solutions to the challenge and the submissions are evaluated[12–14].

Here we present the results for the transcriptional network inference challenge from DREAM5, the fifth annual set of DREAM systems biology challenges. The community of network inference experts was invited to infer genome-scale transcriptional regulatory networks from gene-expression microarray data sets for a prokaryotic model organism (*E. coli*), a eukaryotic model organism (*S. cerevisiae*), a human pathogen (*S. aureus*) and an *in silico* benchmark (**Fig. 1**).

The predictions made from this challenge enabled the comprehensive characterization of network inference methods across different species and data sets, providing insights into method performance, data requirements and inherent biases. We found that the performance of inference methods varies, with a different method performing best in each setting. Taking advantage of variation, we integrated predictions across inference methods and demonstrated that the resulting community-based consensus networks are robust across species and data sets, achieving the best overall performance by far. Finally, we constructed high-confidence consensus networks for *E. coli* and *S. aureus* and experimentally tested novel regulatory interactions in *E. coli*.

We make all benchmark data sets and team predictions, along with the integrated community predictions, available as a public resource (**Supplementary Data 1–5**). In addition, we provide a web interface through the GenePattern genomic-analysis platform[15]

[1]Computer Science and Artificial Intelligence Laboratory (CSAIL), Massachusetts Institute of Technology (MIT), Cambridge, Massachusetts, USA. [2]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [3]Howard Hughes Medical Institute, Boston, Massachusetts, USA. [4]Department of Biomedical Engineering, Boston University, Boston, Massachusetts, USA. [5]Center for BioDynamics, Boston University, Boston, Massachusetts, USA. [6]Department of Informatics, Ludwig-Maximilians University, Munich, Germany. [7]IBM T.J. Watson Research Center, Yorktown Heights, New York, USA. [8]Full lists of members and affiliations appears at the end of the paper. [9]Wyss Institute, Harvard University, Boston, Massachusetts, USA. [10]Present address: Computational Sciences Center of Emphasis, Pfizer Worldwide Research & Development, Cambridge, Massachusetts, USA. [11]These authors contributed equally to this work. Correspondence should be addressed to G.S. (gustavo@us.ibm.com).

Figure 1 | The DREAM5 network inference challenge. Assessment involved the following steps (from left to right). (1) Participants were challenged to infer the genome-wide transcriptional regulatory networks of *E. coli*, *S. cerevisiae* and *S. aureus* as well as an *in silico* (simulated) network. (2) Gene-expression data sets for a wide range of experimental conditions were compiled. Anonymized data sets were released to the community with the identities of the genes hidden. (3) Twenty-nine participating teams inferred gene regulatory networks. In

addition, we applied six off-the-shelf inference methods. (4) Network predictions from individual teams were integrated to form community networks. (5) Network predictions were assessed using experimentally supported interactions from *E. coli* and *S. cerevisiae* as well as the known *in silico* network.
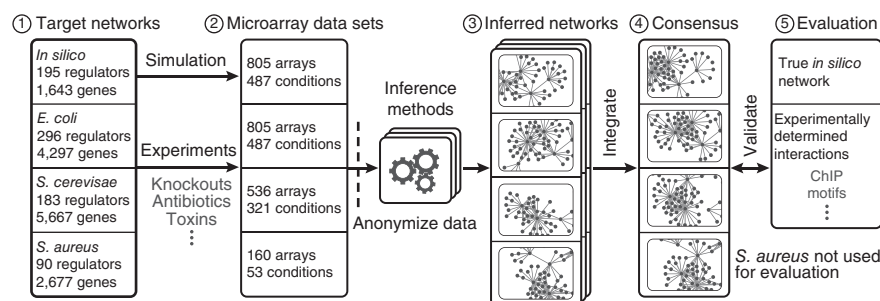
(GP-DREAM, http://dream.broadinstitute.org/), which allows researchers to apply top-performing inference methods and construct consensus networks.

RESULTS
Network inference methods
Proceeding from the DREAM5 challenge (**Supplementary Notes 1–3**), we compared 35 individual methods for inference of gene regulatory networks: 29 submitted by participants and an additional 6 commonly used 'off-the-shelf' tools (**Table 1**). Based on descriptions provided by participants, the methods were classified into six categories: regression, mutual information, correlation, Bayesian networks, meta (methods that combine several different approaches) and other (methods that do not belong to any of the previous categories) (**Table 1**).

Performance of network inference methods
We used three gold standards for performance evaluation: experimentally validated interactions from a curated database (RegulonDB[16]) for *E. coli*, a high-confidence set of interactions supported by genome-wide transcription-factor binding data[17] (ChIP-chip) and evolutionarily conserved binding motifs[18] for *S. cerevisiae*, and the known network for the *in silico* data set (Online Methods). We evaluated performance on *S. aureus* separately (see below), as a sufficiently large set of experimentally validated interactions currently does not exist.

We assessed method performance for the *E. coli*, *S. cerevisiae* and *in silico* data sets using the area under the precision-recall (AUPR) and receiver operating characteristic (AUROC) curves[14] as well as an overall score that summarizes the performance across the three networks (Online Methods and **Supplementary Note 4**). **Figure 2a** shows the overall score and the performance on each network for all applied inference methods. On average, regulatory interactions were recovered more reliably for the *in silico* and *E. coli* data sets than for *S. cerevisiae*.

Notably, well-established off-the-shelf inference methods, such as CLR (context likelihood of relatedness)[11] and ARACNE (algorithm for reconstruction of accurate cellular networks)[9] (categorized as mutual information methods 1 and 3), were substantially outperformed by several teams. The two teams with the best overall score used novel inference approaches based on Random Forests[19] and ANOVA[20] (other 1 and 2), respectively (**Table 1**). However, when their performance on individual networks was considered, the Random Forest and ANOVA-based methods were the best scorers for *E. coli* only. Two regression

methods achieved the best AUPR for the *in silico* benchmark (regression 1 and 2), and two meta predictors did so for *S. cerevisiae* (meta 1 and 5).

Performance also varied within each category of inference methods (**Fig. 2a**). For example, the overall scores obtained by regression methods range from the third best of the challenge down to the fourth lowest. A similar spread in performance can be observed for other categories. We conclude that there is no category of inference methods that is inherently superior and that performance depends largely on the specific implementation of each individual method. For example, several inference methods used the same sparse linear-regression approach (Lasso[21]), but they exhibited large variation in performance because they implemented different data resampling strategies (**Table 1** and **Fig. 2a**).

Complementarity of different inference methods
To examine the observed variation in performance, we analyzed complementary advantages and limitations of the different methods. As a first step, we explored the predicted interactions of all the assessed methods by principal-component analysis (PCA; Online Methods). The top principal components reveal four clusters of inference methods, which coincide with the major categories of inference approaches (**Fig. 2b**). Even though the prediction accuracy of methods from the same category varied (**Fig. 2a**), PCA revealed that they have an intrinsic bias toward predicting similar interactions.

We next analyzed how method-specific biases influenced the recovery of different connectivity patterns (network motifs), and we observed characteristic trends for different method categories (**Fig. 2c**). For example, feed-forward loops were recovered most reliably by mutual-information and correlation-based methods, whereas sparse-regression and Bayesian-network methods performed worse at this task. The reason for this is that the latter approaches preferentially select regulators that independently contribute to the expression of target genes. However, the assumption of independence is violated for genes regulated by mutually dependent transcription factors, as in the case of feed-forward loops. Indeed, linear cascades were more accurately predicted by regression and Bayesian-network methods. This shows that current methods experience a trade-off between performance on cascades and performance on feed-forward loops.

For a subset of the transcription factors contained in the gold standards, knockout or overexpression experiment data were supplied to DREAM5 participants, and several inference

Table 1 | Network inference methods

ID	Synopsis	Reference
	Regression: transcription factors are selected by target gene–specific (i) sparse linear-regression and (ii) data-resampling approaches.	
1	Trustful Inference of Gene REgulation using Stability Selection (TIGRESS): (i) Lasso; (ii) the regularization parameter selects five transcription factors per target gene in each bootstrap sample.	33[a]
2	(i) Steady-state and time-series data are combined by group Lasso; (ii) bootstrapping.	34[a]
3	Combination of Lasso and Bayesian linear regression models learned using reversible-jump Markov chain Monte Carlo simulations.	35[a]
4	(i) Lasso; (i) bootstrapping.	36
5	(i) Lasso; (ii) area under the stability selection curve.	36
6	Application of the Lasso toolbox GENLAB using standard parameters.	37
7	Lasso models are combined by the maximum regularization parameter selecting a given edge for the first time.	36[a]
8	Linear regression determines the contribution of transcription factors to the expression of target genes.	—[a,b]
	Mutual information: edges are (i) ranked based on variants of mutual information and (ii) filtered for causal relationships.	
1	Context likelihood of relatedness (CLR): (i) spline estimation of mutual information; (ii) the likelihood of each mutual information score is computed based on its local network context.	11[a,b]
2	(i) Mutual information is computed from discretized expression values.	38[a,b]
3	Algorithm for the reconstruction of accurate cellular networks (ARACNE): (i) kernel estimation of mutual information; (ii) the data processing inequality is used to identify direct interactions.	9[a,b]
4	(i) Fast kernel-based estimation of mutual information; (ii) Bayesian local causal discovery (BLCD) and Markov blanket (HITON-PC) algorithm to identify direct interactions.	39[a]
5	(i) Mutual information and Pearson's correlation are combined; (ii) BLCD and HITON-PC algorithm.	39[a]
	Correlation: edges are ranked based on variants of correlation.	
1	Absolute value of Pearson's correlation coefficient.	38
2	Signed value of Pearson's correlation coefficient.	38[a,b]
3	Signed value of Spearman's correlation coefficient.	38[a,b]
	Bayesian networks: optimize posterior probabilities by different heuristic searches.	
1	Simulated annealing (catnet R package, http://cran.r-project.org/web/packages/catnet/), aggregation of three runs.	—
2	Simulated annealing (catnet R package, hyperlink above).	—
3	Max-min parent and children algorithm (MMPC), bootstrapped data sets.	40
4	Markov blanket algorithm (HITON-PC), bootstrapped data sets.	41
5	Markov boundary induction algorithm (TIE*), bootstrapped data sets.	42
6	Models transcription factor perturbation data and time series using dynamic Bayesian networks (Infer.NET toolbox, http://research.microsoft.com/infernet/).	—[a]
	Other approaches: network inference by heterogeneous and novel methods.	
1	GENIE3: a Random Forest is trained to predict target gene expression. Putative transcription factors are selected as tree nodes if they consistently reduce the variance of the target.	19[a]
2	Codependencies between transcription factors and target genes are detected by the nonlinear correlation coefficient η^2 (two-way ANOVA). Transcription-factor perturbation data are up-weighted.	20[a]
3	Transcription factors are selected by maximizing the conditional entropy for target genes, which are represented as Boolean vectors with probabilities to avoid discretization.	43[a]
4	Transcription factors are preselected from transcription-factor perturbation data or by Pearson's correlation and then tested by iterative Bayesian model averaging (BMA).	44
5	A Gaussian noise model is used to estimate whether the expression of a target gene changes in transcription-factor perturbation measurements.	45
6	After scaling, target genes are clustered by Pearson's correlation. A neural network is trained (genetic algorithm) and parameterized (back-propagation).	46[a]
7	Data is discretized by Gaussian mixture models and clustering; interactions are detected by generalized logical network modeling (χ^2 test).	47[a]
8	The χ^2 test is applied to evaluate the probability of a shift in transcription-factor and target-gene expression in transcription-factor perturbation experiments.	47[a]
	Meta predictors: (i) apply multiple inference approaches and (ii) compute aggregate scores.	
1	(i) z scores for target genes in transcription-factor knockout data, time-lagged CLR for time series, and linear ordinary differential-equation models constrained by Lasso (Inferelator); (ii) resampling approach.	48[a]
2	(i) Pearson's correlation, mutual information and CLR; (ii) rank average.	—
3	(i) Calculates target-gene responses in transcription-factor knockout data, applies full-order, partial correlation and transcription factor-target codeviation analysis; (ii) weighted average with weights trained on simulated data.	—[a]
4	(i) CLR filtered by negative Pearson's correlation, least-angle regression (LARS) of time series, and transcription factor perturbation data; (2) combination by z scores.	49
5	(i) Pearson's correlation, differential expression (limma), and time-series analysis (maSigPro); (ii) naive Bayes.	—[a]

Methods have been manually categorized based on participant-supplied descriptions. Within each class, methods are sorted by overall performance (see **Fig. 2a**). Note that generic references have been used if more specific ones were not available.
[a]Detailed method description included in **Supplementary Note 10**; [b]Off-the-shelf algorithm applied by challenge organizers.

Figure 2 | Evaluation of network inference methods. Inference methods are indexed according to **Table 1**. (**a**) The plots depict the performance for the individual networks (area under precision-recall curve, AUPR) and the overall score summarizing the performance across networks (Online Methods). R, random predictions; C, integrated community predictions. (**b**) Methods are grouped according to the similarity of their predictions via principal-component analysis. The second versus third principal components are shown; the first principal component accounts mainly for the overall performance (**Supplementary Note 4**). (**c**) The heat map depicts method-specific biases in predicting network motifs. Rows represent individual methods and columns represent different types of regulatory motifs. Red and blue show interactions that are easier and harder to detect, respectively.
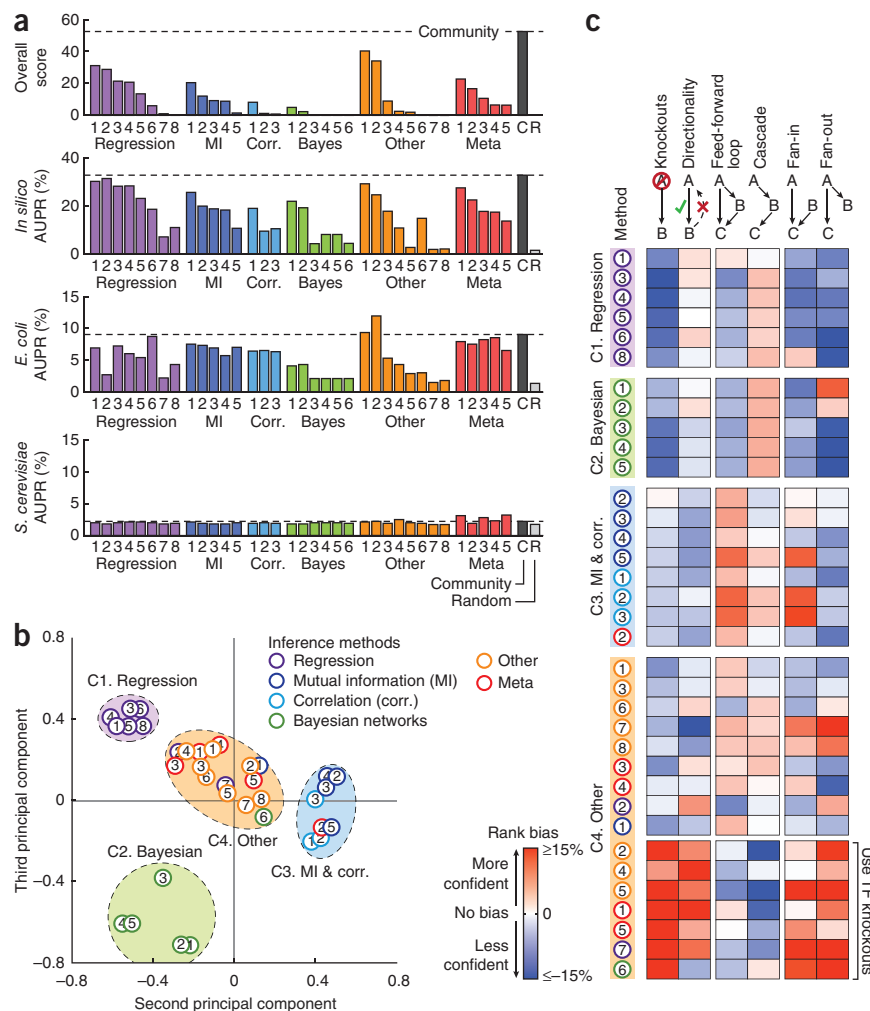
methods explicitly used this information. Consequently, these methods recovered target genes of deleted transcription factors more reliably than the inference methods that did not leverage this information (**Fig. 2c**). Explicit use of such knockouts also helped methods to draw the direction of edges between transcription factors more reliably. These observations suggest that measurements of transcription-factor knockouts can be informative for network reconstruction. In particular, this is the case for the *E. coli* data set, which contained the largest number of such experiments (Online Methods). To further explore the information content of different experiments, we employed a machine learning framework[22] to systematically analyze the information gain from microarrays grouped according to the type of experimental perturbation (knockouts, drug perturbations, environmental perturbations and time series; **Supplementary Note 5**). We found that experimental conditions independent of transcription factor knockout and overexpression also provide information, though at a reduced level.

Community networks outperform individual inference methods
Network inference methods have complementary advantages and limitations under different contexts, which suggests that combining the results of multiple inference methods could be a good strategy for improving predictions. We therefore integrated the predictions of all participating teams to construct community networks by rescoring interactions according to their average rank across all methods (**Supplementary Note 6**). The integrated community network ranks first for *in silico*, third for *E. coli* and sixth for *S. cerevisiae* out of the 35 applied inference methods, which shows that the community network is consistently as good or better than the top individual methods (**Fig. 2a**). Thus it has by far the best performance reflected in the overall score. We stress that, even though top-performing methods for a given network are competitive with the integrated community method, the performance of individual methods does not generalize across networks.

Given the biological variation among organisms and the experimental variation among gene-expression data sets, it is difficult to determine beforehand which methods will perform optimally for reconstructing an unknown regulatory network. In contrast, the community approach performs robustly across diverse data sets.

We next analyzed how the number of integrated methods affects the performance of community predictions by examining randomly sampled combinations of individual methods. On average, community methods perform better than individual inference methods even when integrating small sets of individual predictions: for example, just five teams (**Fig. 3a**). Performance increases further with the number of integrated methods. For instance, given 20 inference methods, their integration ranks first or second in 98% of the cases (**Fig. 3b**). We also found that the performance of the community network can be improved by increasing the diversity of the underlying inference methods. Consensus predictions from teams using similar methodologies were outperformed by consensus predictions from diverse methodologies (**Fig. 3c**).

A key feature in taking a community network approach is robustness to the inclusion of a limited subset (up to ~20%) of poorly performing inference methods (**Fig. 3d**). Poor predictors essentially contributed noise, but this did not affect the performance of the community approach as a whole. This finding is crucial because

Figure 3 | Analysis of community networks compared to individual inference methods. (**a**) The plot shows the overall score, which summarizes performance across the *E. coli*, *S. cerevisiae* and *in silico* networks, for individual inference methods or various combinations of integrated methods. The first box plot depicts the performance distribution of individual inference methods ($K = 1$). Subsequent box plots show the performance when $K > 1$ randomly sampled methods were integrated. The red bar shows the performance when all methods ($K = 29$) were integrated. Box plots depict performance distributions with respect to the minimum, the maximum and the three quartiles. (**b**) The probability that the community network ranks among the top $x\%$ of the K individual methods used to construct the community network. The diagonal shows the expected performance when an individual method was chosen ($K = 1$). (**c**) The integration of complementary methods is particularly beneficial. The first box plot shows the performance of individual methods from clusters 1–3 (as defined in **Fig. 2b**). The second and third box plots show performance of community networks, which were obtained by integrating three randomly selected inference methods (i) from the same cluster or (ii) from different clusters. (**d**) The plots show the overall score for an initial community network formed by integrating all individual methods except for the best five or worst five. The worst five (left) and best (right) five methods were added one by one to form additional community networks.


the performance of individual methods when inferring regulatory networks for poorly studied organisms is not known a priori and is hard to evaluate empirically: even top performers on a benchmark network (such as *E. coli*) have varied performance when inferring a new, unknown network (such as *S. aureus*). On the other hand, adding good performers substantially increased the performance of the community approach (**Fig. 3d**), which highlights the importance of developing high-quality individual inference methods.

E. coli and *S. aureus* community networks

To gain insights into transcriptional gene regulation for two bacteria, *E. coli* and *S. aureus*, we constructed networks for both organisms by integrating the predictions of all teams using the average-rank method. **Figure 4** shows the community networks for both organisms at a cutoff of 1,688 edges, which corresponds to an estimated precision of 50% for the *E. coli* network based on the gold standard of experimentally validated interactions from RegulonDB (Online Methods). At this cutoff, 50% of the *de novo* predicted regulatory edges were recovered known interactions; the remaining 50% may be false positives or newly discovered true interactions.

The precision of the *S. aureus* network cannot be measured accurately because there are comparatively few experimentally supported interactions available. Nevertheless, we confirmed the robustness of the consensus predictions by evaluating the network using the largely computationally derived interactions from the RegPrecise database[23] (**Supplementary Note 7**).

We found that the *E. coli* and *S. aureus* networks both have a modular structure[24]: that is, they comprise clusters of genes that are more densely connected amongst themselves than with other parts of the network. After identifying these modules[24], we tested them for enrichment of Gene Ontology terms (**Supplementary Note 7**). Network modules are strongly enriched for very specific biological processes. This allowed us to assign unique functions to most of the identified modules in both networks (**Fig. 4** and **Supplementary Data 6**). As a specific example of an enriched module, 27 genes in *S. aureus* are highly enriched for pathogenic genes (**Fig. 4b**).
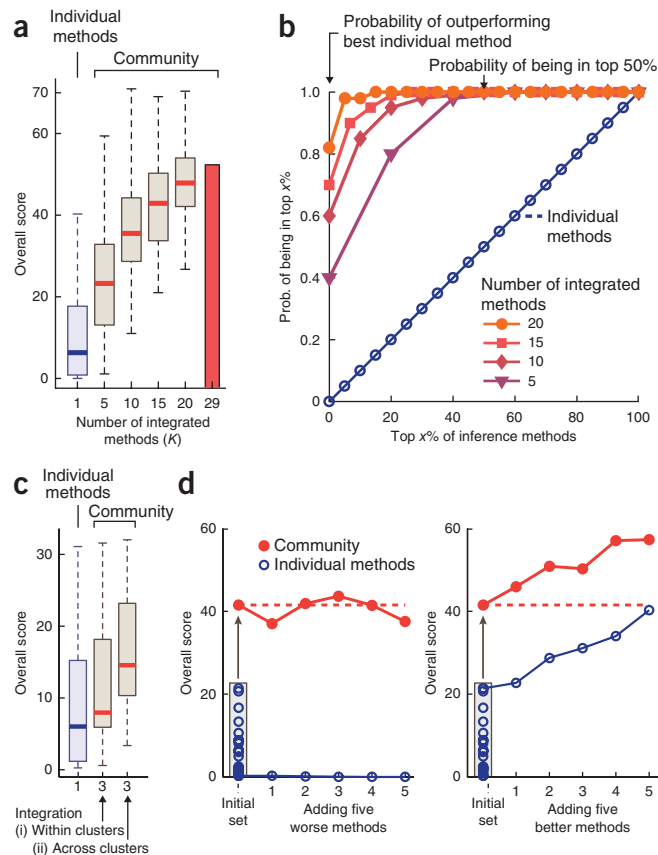
These include genes encoding exotoxins (*set7*, *set8*, *set11*, *set14*), genes responsible for biofilm formation (*tcaR*) and antibiotic metabolism (*tetR*), as well as one encoding a cell surface protein (*fnb*). The remaining 20 genes of this module are uncharacterized, but the predicted connections suggest their role in pathogenesis. This example illustrates how the inferred networks generate specific hypotheses regarding both the regulation and function of uncharacterized genes, enabling targeted validation efforts.

Experimental support of novel interactions

In addition to validation against known interactions from the RegulonDB gold standard, we experimentally tested a subset of novel predictions from the *E. coli* community network described above. We selected five transcription factors (*rhaR*, *cueR*, *purR*, *mprA* and *gadE*), and then we tested each of the 53 corresponding target gene predictions individually (**Supplementary Note 8**). Using qPCR, we measured the expression of each predicted target gene in the absence or presence of a chemical inducer known to activate the corresponding transcription factor (rhamnose for *rhaR*, copper sulfate for *cueR*, adenine for *purR*, carbonyl cyanide *m*-chlorophenylhydrazone for *mprA* and hydrogen chloride for *gadE*). We also measured target gene expression in transcription-factor deletion strains, again in the absence or presence of the chemical inducer. Putative targets were considered confirmed if they showed (i) strong response to the inducer of the respective transcription factor in the wild type and (ii) no response to the inducer in the transcription-factor deletion strain. We observed a clear difference between the two responses (>1.8 fold) for 23 novel targets out of 53 tested (**Fig. 4c**); this corresponds to a precision of

Figure 4 | *E. coli* and *S. aureus* community networks. (**a**,**b**) At a cutoff of 1,688 edges, the *E. coli* community network (**a**) connects 1,505 genes (including 204 transcription factors, shown as diamonds), and the *S. aureus* network (**b**) connects 1,084 genes (85 transcription factors). Network modules were identified and tested for Gene Ontology–term enrichment, as indicated (gray genes do not show enrichment). A network module enriched for Gene Ontology terms related to pathogenesis is highlighted in the *S. aureus* network. (**c**) The schematics depict newly predicted *E. coli* regulatory interactions that were experimentally tested. The pie chart depicts the breakdown of strongly and weakly supported targets (Online Methods). The positive controls were six known interactions from RegulonDB.
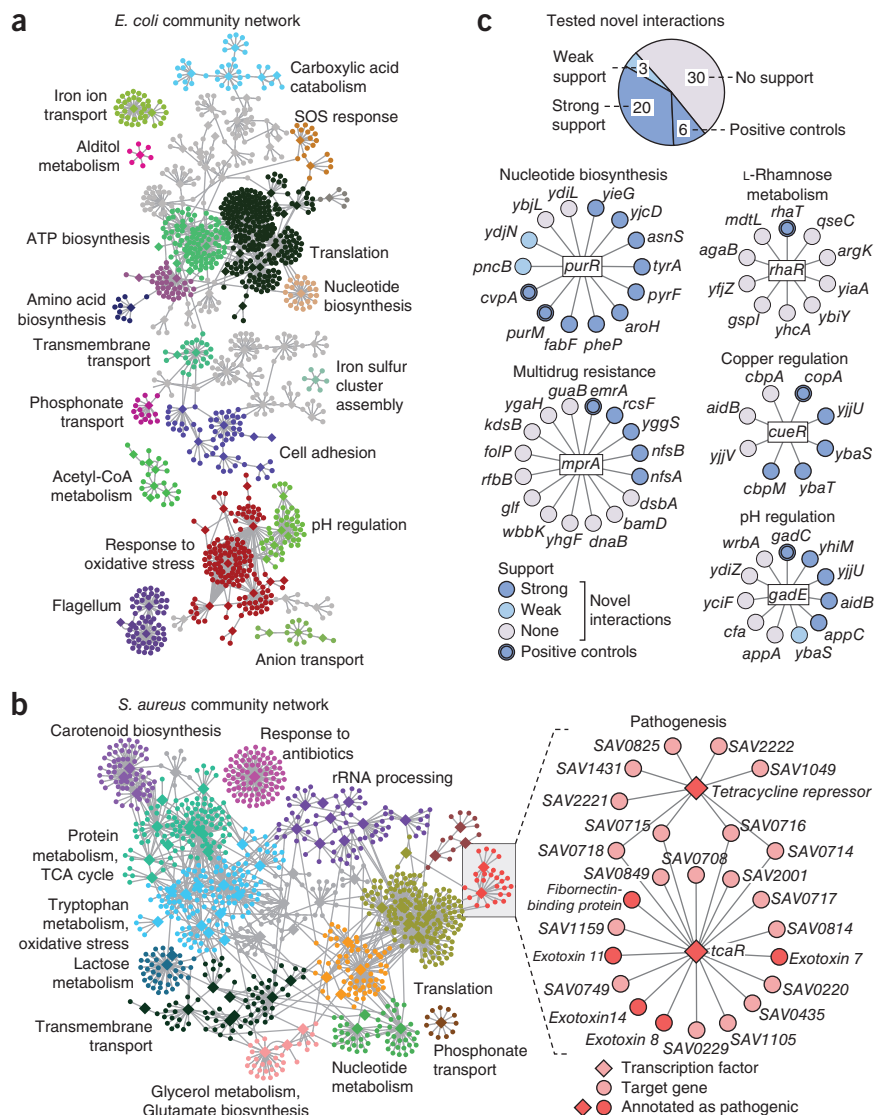
~40% for novel interactions, which is in line with our estimate of ~50% precision based on known interactions from RegulonDB. We note that these data support a direct regulatory effect of the tested transcription factor on the target gene, but chromatin immunoprecipitation experiments would be required to determine physical binding.

We observe a large variation in experimental validation among individual transcription factors (**Fig. 4c**). For *purR*, a key regulator in purine nucleotide metabolism, 10 of the 12 predicted target genes were experimentally supported. Nucleotide metabolism is a fundamental biological process that is affected across multiple conditions, and thus *purR* regulation is well sampled across the *E. coli* data set. However, in the case of *rhaR*, a key regulator in L-rhamnose deg-

radation, none of the novel target-gene predictions showed signs of regulation. L-Rhamnose degradation is a specialized process that is only activated in the presence of L-rhamnose, and there were no conditions in the *E. coli* data set in which L-rhamnose degradation was explicitly tested. In the instance of *cueR*, a transcriptional regulator activated in the presence of copper, four out of seven novel target-gene predictions were confirmed. As with *rhaR*, there were no conditions in the data set that explicitly tested copper regulation, yet unlike with *rhaR*, network inference methods were able to identify true positive *cueR* regulatory interactions. These results suggest that although the overall precision for the network is high, the reliability of predictions for individual transcription factors can vary. When constructing a compendium of microarrays for global network inference, one should thus avoid any bias toward oversampling a narrow set of experimental conditions.

DISCUSSION
The DREAM project provides a unique framework where network inference methods from a community of experts are collected and impartially assessed on benchmark data sets.

The collection of 35 inference methods assessed here constitutes a unique resource, as it spans all commonly used approaches in the field. In addition, the collection includes novel approaches (including the two best individual team performers of the challenge), representing a snapshot of the latest developments in the field.

Our analyses revealed specific advantages and limitations of different inference approaches (see **Supplementary Note 9** and the full description of approaches in **Supplementary Note 10**). Sparse linear-regression methods performed well, but only when data resampling strategies such as bootstrapping were used (the best-performing regression methods all used data resampling, whereas the worst-performing methods did not). Sparsity constraints employed by these methods effectively increased performance for cascade motifs at the cost of missing interactions in feed-forward loops, fan-in motifs and fan-out motifs. Bayesian-network methods exhibited below-average performance in this challenge, likely because they use heuristic searches, which are often too costly for systematic data resampling and may be better suited for smaller networks. Information theoretic methods performed better than correlation-based methods, but the two

approaches had similar biases in predicting regulatory relationships. They also performed better than regression and Bayesian-network methods on feed-forward loops, fan-ins and fan-outs (the more densely connected parts of the network), but they had an increased rate of false positives for cascades. Meta predictors performed more robustly across data sets than other categories of methods; however, they could not match the robustness and performance of the community predictions, presumably because they combine methods that do not provide sufficient diversity. Among all categories, methods that made explicit use of direct transcription-factor perturbations (knockout or overexpression) greatly improved prediction accuracy for downstream targets (albeit at an increased false-positive rate for cascades). For improving individual inference approaches, we suggest the following: (i) optimally exploit direct transcription-factor perturbations; (ii) employ strategies to avoid overfitting, such as data resampling; and (iii) develop more effective approaches to distinguish direct from indirect regulation (feed-forward loops versus cascades).

Overall, methods performed well for the *in silico* and prokaryotic (*E. coli*) data sets. However, inferring gene regulatory networks from the eukaryotic (*S. cerevisiae*) data set proved to be a greater challenge. A fundamental assumption of network inference algorithms is that mRNA levels of transcription factors and their targets tend to be correlated; we found that this is true for *E. coli*, but not for *S. cerevisiae* (**Supplementary Note 5**). Although the lower coverage of *S. cerevisiae* gold standards may also play a role (*E. coli* has the best-known regulatory network of any free-living organism[16]), the poor correlation at the mRNA level in *S. cerevisiae* is likely due to the increased regulatory complexity and prevalence of post-transcriptional regulation in eukaryotes, which would suggest that accurate inference of eukaryotic regulatory networks requires additional inputs, such as promoter sequences and data sets for transcription-factor binding and chromatin modification[7].

Individual studies that introduce a novel inference method naturally tend to focus on its advantages in a particular application, which can paint an overoptimistic picture of performance[13]. Whereas previous studies have explored strengths and weaknesses of inference approaches[2,3], the present assessment shows that method performance is not robust across species and varies greatly even within the same category of inference methods (**Table 1**). This implies that performance is related more to the details of implementation than the choice of underlying methodology.

In network inference, variation in performance presents a problem, but at the same time offers a solution. By integrating the predictions from individual methods into community networks, we show that advantages of different methods complement each other and limitations tend to be canceled out. Instead of relying on a single inference method with uncertain performance on a previously unseen network, integrating predictions across inference methods becomes the best strategy. We note that not all of the 29 methods are required for enhanced performance. By considering complementary methods, we have shown that performance can be substantially improved with as few as three methods (**Fig. 3c**).

Ensemble-based methods have a storied past, with applications ranging from economics[1] to machine learning[25]. In systems biology, robust models are often constructed from ensembles of instances (for example, different parameterizations or model

structures) that are derived from experimental data via a single approach[26–30], such as Monte Carlo sampling. In contrast, we formed consensus predictions from a large array of heterogeneous inference approaches. These 'meta predictors' have been successful in other machine learning competitions[31,32]. We have observed from previous DREAM challenges anecdotal evidence that community predictions can rank among the top performers[13], but we did not previously attempt a systematic study of prediction integration for network inference. Here we established, through rigorous assessments and experimentally derived data sets, the performance robustness of prediction integration for transcriptional gene network inference.

The shortcomings of individual methods revealed in our assessment present many opportunities for improving these methods. We also expect further improvements in performance from advanced community approaches that: (i) actively leverage the method-specific advantages with regard to the data sets and networks of interest; (ii) optimize diversity in the ensemble—for example, by weighting methods so as to balance the contribution of different method categories or PCA clusters; and (iii) employ more sophisticated voting schemes to negotiate consensus networks. To help spur developments in these areas, we provide the GP-DREAM web platform for the community to develop and apply network inference and consensus methods (http://dream.broadinstitute.org/). We will continue to expand this free toolkit with top-performing methods from the DREAM challenges as well as other methods contributed by the community.

METHODS
Methods and any associated references are available in the online version of the paper.

Note: Supplementary information is available in the online version of the paper.

R.K., N.M.V., R.J.P., K.R.A., M.K., J.J.C. and G.S. analyzed results; D.M., J.C.C., R.K., M.K., J.J.C. and G.S. wrote the paper; and challenge participants performed network inference and provided method descriptions.

1. Surowiecki, J. *The Wisdom of Crowds: Why the Many are Smarter than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations* (Doubleday, 2004).
2. De Smet, R. & Marchal, K. Advantages and limitations of current network inference methods. *Nat. Rev. Microbiol.* **8**, 717–729 (2010).
3. Marbach, D. *et al.* Revealing strengths and weaknesses of methods for gene network inference. *Proc. Natl. Acad. Sci. USA* **107**, 6286–6291 (2010).
4. Bar-Joseph, Z. *et al.* Computational discovery of gene modules and regulatory networks. *Nat. Biotechnol.* **21**, 1337–1342 (2003).
5. Reiss, D.J., Baliga, N.S. & Bonneau, R. Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks. *BMC Bioinformatics* **7**, 280 (2006).
6. Lemmens, K. *et al.* DISTILLER: a data integration framework to reveal condition dependency of complex regulons in *Escherichia coli*. *Genome Biol.* **10**, R27 (2009).
7. Marbach, D. *et al.* Predictive regulatory models in *Drosophila melanogaster* by integrative inference of transcriptional networks. *Genome Res.* published online (28 March 2012).
8. Friedman, N., Linial, M., Nachman, I. & Pe'er, D. Using Bayesian networks to analyze expression data. *J. Comput. Biol.* **7**, 601–620 (2000).
9. Margolin, A.A. *et al.* ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* **7** (suppl. 1), S7 (2006).
10. di Bernardo, D. *et al.* Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. *Nat. Biotechnol.* **23**, 377–383 (2005).
11. Faith, J.J. *et al.* Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. *PLoS Biol.* **5**, e8 (2007).
12. Stolovitzky, G., Monroe, D. & Califano, A. Dialogue on reverse-engineering assessment and methods: the DREAM of high-throughput pathway inference. *Ann. NY Acad. Sci.* **1115**, 1–22 (2007).
13. Stolovitzky, G., Prill, R.J. & Califano, A. Lessons from the DREAM2 Challenges. *Ann. NY Acad. Sci.* **1158**, 159–195 (2009).
14. Prill, R.J. *et al.* Towards a rigorous assessment of systems biology models: the DREAM3 challenges. *PLoS ONE* **5**, e9202 (2010).
15. Reich, M. *et al.* GenePattern 2.0. *Nat. Genet.* **38**, 500–501 (2006).
16. Gama-Castro, S. *et al.* RegulonDB version 7.0: transcriptional regulation of *Escherichia coli* K-12 integrated within genetic sensory response units (Gensor Units). *Nucleic Acids Res.* **39**, D98–D105 (2011).
17. Harbison, C.T. *et al.* Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**, 99–104 (2004).
18. MacIsaac, K.D. *et al.* An improved map of conserved regulatory sites for *Saccharomyces cerevisiae*. *BMC Bioinformatics* **7**, 113 (2006).
19. Huynh-Thu, V.A., Irrthum, A., Wehenkel, L. & Geurts, P. Inferring regulatory networks from expression data using tree-based methods. *PLoS ONE* **5**, e12776 (2010).
20. Küffner, R., Petri, T., Tavakkolkhah, P., Windhager, L. & Zimmer, R. Inferring Gene Regulatory Networks by ANOVA. *Bioinformatics* **28**, 1376–1382 (2012).
21. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Series B Stat. Methodol.* **58**, 267–288 (1996).
22. Mordelet, F. & Vert, J.-P. SIRENE: supervised inference of regulatory networks. *Bioinformatics* **24**, i76–i82 (2008).
23. Ravcheev, D.A. *et al.* Inference of the transcriptional regulatory network in *Staphylococcus aureus* by integration of experimental and genomics-based evidence. *J. Bacteriol.* **193**, 3228–3240 (2011).
24. Newman, M.E.J. Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA* **103**, 8577–8582 (2006).
25. Dietterich, T.G. Ensemble methods in machine learning. *Multiple Classifier Systems, First International Workshop* (eds. Kittler, J. & Roli, F.) **1857**, 1–15 (Springer, 2000).
26. Prinz, A.A., Bucher, D. & Marder, E. Similar network activity from disparate circuit parameters. *Nat. Neurosci.* **7**, 1345–1352 (2004).
27. Kuepfer, L., Peter, M., Sauer, U. & Stelling, J. Ensemble modeling for analysis of cell signaling dynamics. *Nat. Biotechnol.* **25**, 1001–1006 (2007).
28. Kaltenbach, H.-M., Dimopoulos, S. & Stelling, J. Systems analysis of cellular networks under uncertainty. *FEBS Lett.* **583**, 3923–3930 (2009).
29. Marbach, D., Mattiussi, C. & Floreano, D. Combining multiple results of a reverse-engineering algorithm: application to the DREAM five-gene network challenge. *Ann. NY Acad. Sci.* **1158**, 102–113 (2009).
30. Marder, E. & Taylor, A.L. Multiple models to capture the variability in biological neurons and networks. *Nat. Neurosci.* **14**, 133–138 (2011).
31. Moult, J. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Curr. Opin. Struct. Biol.* **15**, 285–289 (2005).
32. Bell, R.M. & Koren, Y. Lessons from the Netflix Prize Challenge. *SIGKDD Explor.* **9**, 75–79 (2007).
33. Haury, A.-C., Mordelet, F., Vera-Licona, P. & Vert, J.-P. TIGRESS: trustful inference of gene regulation using stability selection. Preprint at <http://arxiv.org/abs/1205.1181> (2012).
34. Yuan, M. & Lin, Y. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Series B Stat. Methodol.* **68**, 49–67 (2006).
35. Lèbre, S., Becq, J., Devaux, F., Stumpf, M.P.H. & Lelandais, G. Statistical inference of the time-varying structure of gene-regulation networks. *BMC Syst. Biol.* **4**, 130 (2010).
36. Meinshausen, N. & Bühlmann, P. Stability selection. *J. R. Stat. Soc. Series B Stat. Methodol.* **72**, 417–473 (2010).
37. van Someren, E.P. *et al.* Least absolute regression network analysis of the murine osteoblast differentiation network. *Bioinformatics* **22**, 477–484 (2006).
38. Butte, A.J. & Kohane, I.S. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac. Symp. Biocomput.* **2000**, 418–429 (2000).
39. Mani, S. & Cooper, G.F. A Bayesian local causal discovery algorithm. in *Proceedings of the World Congress on Medical Informatics, MedInfo 2004* (eds. Fieschi, M. *et al.*) 731–735 (IOS, 2004).
40. Tsamardinos, I., Aliferis, C.F. & Statnikov, A. Time and sample efficient discovery of Markov blankets and direct causal relations. in *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 673–678 (ACM, 2003).
41. Aliferis, C.F., Statnikov, A., Tsamardinos, I., Mani, S. & Koutsoukos, X.D. Local causal and Markov blanket induction for causal discovery and feature selection for classification part I: algorithm and empirical evaluation. *J. Mach. Learn. Res.* **11**, 171–234 (2010).
42. Statnikov, A. & Aliferis, C.F. Analysis and computational dissection of molecular signature multiplicity. *PLoS Comput. Biol.* **6**, e1000790 (2010).
43. Karlebach, G. & Shamir, R. Constructing logical models of gene regulatory networks by integrating transcription factor-DNA interactions with expression data: an entropy-based approach. *J. Comput. Biol.* **19**, 30–41 (2012).
44. Yeung, K.Y., Bumgarner, R.E. & Raftery, A.E. Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data. *Bioinformatics* **21**, 2394–2402 (2005).
45. Yip, K.Y., Alexander, R.P., Yan, K.-K. & Gerstein, M. Improved reconstruction of *in silico* gene regulatory networks by integrating knockout and perturbation data. *PLoS ONE* **5**, e8121 (2010).
46. Sîrbu, A., Ruskin, H.J. & Crane, M. Stages of gene regulatory network inference: the evolutionary algorithm role. in *Evolutionary Algorithms* (ed. Kita, E.) Ch. 27, 521–546 (Intech, 2011).
47. Song, M.J. *et al.* Reconstructing generalized logical networks of transcriptional regulation in mouse brain from temporal gene expression data. *EURASIP J. Bioinform. Syst. Biol.* **2009**, 545176 (2009).
48. Greenfield, A., Madar, A., Ostrer, H. & Bonneau, R. DREAM4: Combining genetic and dynamic information to identify biological networks and dynamical models. *PLoS ONE* **5**, e13397 (2010).
49. Watkinson, J., Liang, K.-C., Wang, X., Zheng, T. & Anastassiou, D. Inference of regulatory gene interactions from expression data using three-way mutual information. *Ann. NY Acad. Sci.* **1158**, 302–313 (2009).

Fifth Annual Dialogue on Reverse Engineering Assessment and Methods (DREAM5) Consortium:

Andrej Aderhold[12,13], Kyle R Allison[3–5], Richard Bonneau[14–17], Diogo M Camacho[3–5], Yukun Chen[18], James J Collins[3–5,9], Francesca Cordero[19,20], James C Costello[3–5], Martin Crane[21], Frank Dondelinger[12,22], Mathias Drton[23], Roberto Esposito[19], Rina Foygel[23], Alberto de la Fuente[24], Jan Gertheiss[25], Pierre Geurts[26,27], Alex Greenfield[16], Marco Grzegorczyk[28], Anne-Claire Haury[29–31], Benjamin Holmes[1,2], Torsten Hothorn[25], Dirk Husmeier[12], Vân Anh Huynh-Thu[26,27], Alexandre Irrthum[26,27], Manolis Kellis[1,2], Guy Karlebach[32], Robert Küffner[6], Sophie Lèbre[33], Vincenzo De Leo[24,34], Aviv Madar[14,15], Subramani Mani[18], Daniel Marbach[1,2], Fantine Mordelet[29–31,35], Harry Ostrer[36], Zhengyu Ouyang[37], Ravi Pandya[38], Tobias Petri[6], Andrea Pinna[24], Christopher S Poultney[14,15], Robert J Prill[7], Serena Rezny[23], Heather J Ruskin[21], Yvan Saeys[39,40], Ron Shamir[32], Alina Sîrbu[21], Mingzhou Song[37], Nicola Soranzo[24], Alexander Statnikov[41], Gustavo Stolovitzky[7], Nicci Vega[3–5], Paola Vera-Licona[29–31], Jean-Philippe Vert[29–31], Alessia Visconti[19], Haizhou Wang[37], Louis Wehenkel[26,27], Lukas Windhager[6], Yang Zhang[37] & Ralf Zimmer[6]

[12]Biomathematics and Statistics Scotland, Edinburgh & Aberdeen, UK. [13]School of Biology, University of St. Andrews, St. Andrews, UK. [14]Department of Biology, New York University, New York, NY, USA. [15]Center for Genomics & Systems Biology, New York University, New York, NY, USA. [16]Computational Biology Program, New York University Sackler School of Medicine, New York, New York, USA. [17]Computer Science Department, Courant Institute of Mathematical Sciences, New York University, New York, New York, USA. [18]Department of Biomedical Informatics, Vanderbilt University, Nashville, Tennessee, USA. [19]Department of Computer Science, University of Turin, Torino, Italy. [20]Department of Clinical and Biological Sciences, University of Turin, Orbassano, Italy. [21]Centre for Scientific Computing and Complex Systems Modelling, School of Computing, Dublin City University, Dublin, Ireland. [22]School of Informatics, University of Edinburgh, Edinburgh, UK. [23]Department of Statistics, University of Chicago, Chicago, Illinois, USA. [24]CRS4 Bioinformatica, Parco Tecnologico della Sardegna, Pula, Italy. [25]Department of Statistics, Ludwig-Maximilians University, Munich, Germany. [26]Department of Electrical Engineering and Computer Science, Systems and Modeling, University of Liège, Liège Belgium. [27]GIGA-Research, Bioinformatics and Modeling, University of Liège, Liège, Belgium. [28]Department of Statistics, TU Dortmund University, Germany. [29]Mines ParisTech, Center for Computational Biology (CBIO), Fontainebleau, France. [30]Institut Curie, Centre de Recherche, Paris, France. [31]Institut National de la Santé et de la Recherche Médicale (INSERM), U900, Paris, France. [32]The Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv, Israel. [33]Computer Sciences and Remote Sensing Laboratory (LSIIT), Mixed Research Unit (UMR) UdS-CNRS 7005, Université de Strasbourg, Strasbourg, France. [34]Linkalab, Complex Systems Computational Laboratory, Cagliari, Italy. [35]Centre de Recherche en Économie et Statistique (CREST), National Institute of Statistics and Economic Studies (INSEE), Malakoff, France. [36]Human Genetics Program, Department of Pediatrics, New York University Langone Medical Center, New York, New York, USA. [37]Department of Computer Science, New Mexico State University, Las Cruces, New Mexico, USA. [38]Microsoft Research, Redmond, Washington, USA. [39]Department of Plant Systems Biology, Flanders Institute for Biotechnology (VIB), Gent, Belgium. [40]Department of Plant Biotechnology and Bioinformatics, Ghent University, Gent, Belgium. [41]Center for Health Informatics and Bioinformatics, New York University, New York, New York, USA.

ONLINE METHODS

Expression data and gold standards. The design of the DREAM5 network inference challenge is outlined in **Figure 1** (full description in **Supplementary Note 1**). Affymetrix gene expression data sets were compiled for *E. coli*, *S. aureus* and *S. cerevisiae* from the Gene Expression Omnibus (GEO) database[50]. Microarray data sets were uniformly normalized using Robust Multichip Averaging (RMA)[51]. Each data set queries the underlying regulatory network in hundreds of different conditions ranging from time courses to gene, drug and environmental perturbations. Note that the number of measurements of transcription factor–specific perturbations varies among the data sets (*S. aureus*: 0/161, *E. coli*: 67/806 and yeast: 3/537). The fourth data set is an *in silico* counterpart to the *E. coli* data set, generated using GeneNetWeaver[52,53] (version 4.0). The structure of the *in silico* network corresponds to the *E. coli* transcriptional regulatory network from RegulonDB[16] (10% random edges were added, resulting in 3,940 interactions). In addition to the gene-expression data, we provide a list of putative transcription factors for each data set and several descriptive features for each microarray experiment (for example, the target of a gene deletion, or the time point of a time-series experiment). It is important to note that the identity of the organisms from which the data was generated was unknown to the participants. This was achieved by encrypting certain aspects of the data and by anonymizing gene names.

Participants were presented the challenge of inferring direct regulatory interactions between transcription factors and target genes from the given gene-expression data sets. The submission format was a ranked list of predicted regulatory relationships for each network[3].

The gold standard set of known transcriptional interactions for *E. coli* was obtained from RegulonDB[16]. We only included well-established interactions annotated with 'strong evidence' according to RegulonDB evidence classification (2,066 interactions). For *S. cerevisiae*, we considered several alternative gold standards derived from orthogonal data sets, namely ChIP binding data and evolutionary conserved transcription-factor binding motifs[18] as well as systematic transcription-factor deletions[54] (**Supplementary Note 3**). For the results reported in the main text, we used the most stringent gold standard, which includes only interactions that have strong evidence of both binding and conservation[18].

All data and scripts are available in **Supplementary Data 1** and at the DREAM website (http://wiki.c2b2.columbia.edu/dream/index.php/D5c4). The original microarray data sets are also publicly available at the Many Microbe Microarrays Database[55] (M3D, http://m3d.bu.edu/dream/).

Performance metrics. A detailed description of all performance metrics is given in **Supplementary Note 4**. Briefly, transcription factor–target predictions were evaluated as a binary classification task. The gold-standard networks represent the true positive interactions; the remaining pairs are considered negatives. Only the top 100,000 edge predictions were accepted. Pairs of nodes not part of the submitted list were considered to appear randomly ordered at the end of the list. Performance was assessed using the area under the ROC curve (AUROC) and the area under the precision vs. recall curve (AUPR)[14]. Note that predictions

for genes that are not part of the gold standard, i.e., for which no experimentally supported interactions exist, were ignored in this evaluation.

AUROC and AUPR were separately transformed into *p* values by simulating a null distribution for 25,000 random networks. Random edge lists were constructed by sampling edges from the submitted edge lists of the participants and assigning these edges random ranks between 1 and 100,000. The histogram of randomly obtained AUROC and AUPR values was fit using stretched exponentials to extrapolate the distribution to values beyond the immediate range of the histogram[14]. To compute an overall score that summarizes the performance over the three networks with available gold standards (*E. coli*, *S. cerevisiae* and *in silico*), we used the same metric as in the previous two editions of the challenge[3,14], which is defined as the mean of the (log-transformed) network-specific *p* values

$$\text{score}_{\text{ROC}} = \frac{1}{3} \sum_{i=1}^{3} -\log_{10} p_{\text{ROC},i}$$

$$\text{score}_{\text{PR}} = \frac{1}{3} \sum_{i=1}^{3} -\log_{10} p_{\text{PR},i}$$

$$\text{score} = \frac{\text{score}_{\text{ROC}} + \text{score}_{\text{PR}}}{2}$$

Clustering of inference approaches by principal-component analysis. We constructed a prediction matrix P in which rows correspond to edges (transcription factor–target pairs) and columns correspond to inference methods. The element $p_{i,j}$ of this matrix is thus the rank assigned to edge i by inference method j. We only considered edges that figured in the top 100,000 predicted edges of at least three inference methods, which yielded 1,175,525 interactions across the four data sets. Note that knowledge of a gold-standard network is not required for the PCA; thus the *S. aureus* predictions were included in this analysis. The dimensionality of the combined prediction matrix (including the predictions for all four data sets) was reduced by PCA using SVDLIBC with standard parameters (http://tedlab.mit.edu/~dr/SVDLIBC/). Results are consistent when performing PCA for each of the four data sets separately (**Supplementary Note 4**).

Network motif analysis. The goal of the network motif analysis is to evaluate, for a given network inference method, whether some types of edges of motifs are systematically predicted less (or more) reliably than expected[3]. We considered the six motif types illustrated in **Figure 2**. For each type of motif m, we identified all instances in the gold-standard network and determined the average rank r_m assigned to its edges by the inference method. We further determined the average rank $r_{\overline{m}}$ assigned to all edges that are *not* part of this motif type. The prediction bias is given by the difference $r_m - r_{\overline{m}}$. See **Supplementary Note 4** for details.

Experimental materials and design. Novel predictions were selected from the *E. coli* community network with greater than 50% predicted precision. Transcription factors with at least eight novel predictions were selected, including *rhaR*, *cueR*, *purR*, *mprA* and *gadE* (note that the data set supplied to the DREAM5 participants did not contain any knockout measurement for these transcription factors). Primers were designed for all novel target

gene predictions after accounting for operon structure, and at least one known target of the transcription factor was included as a positive control. A total of 53 predictions and 6 positive controls were tested (**Supplementary Data 7**).

For each transcription factor, a knockout strain was generated from the background *E. coli* strain BW25113. Each transcription factor was induced by a different stimulus: rhamnose for *rhaR*, copper sulfate for *cueR*, adenine for *purR*, carbonyl cyanide m-chlorophenylhydrazone for *mprA* and HCl for *gadE*. Four experimental conditions were used for each transcription factor: background strain without inducer (WT(−)), background strain with inducer (WT(+)), deletion strain without inducer (Δ(−)) and deletion strain with inducer (Δ(+)). Three biological replicates were generated for all experimental conditions. Cultures were grown in LB media or minimal media (**Supplementary Note 8**), and incubation was performed in darkened shakers (300 r.p.m.) at 37 °C. PCR primers were designed for all target genes. Target genes were quantified through qPCR using LightCycler 480 SYBR Green I Master Kit (Roche Applied Science). True positive interactions were expected to meet two criteria: (i) a strong response to the TF inducer in wild type and (ii) no or weak response to the TF inducer in the TF-deletion strain. Target gene interactions were considered to have 'strong support' if the ratio of criteria 1 to criteria 2, (WT(+)/WT(−)) / (Δ(+)/Δ(−)), was greater than 2 and 'weak support' if the ratio was between 1.8 and 2 (**Supplementary Data 7**).

50. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets–10 years on. *Nucleic Acids Res.* **39**, D1005–D1010 (2011).
51. Bolstad, B.M., Irizarry, R.A., Astrand, M. & Speed, T.P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185–193 (2003).
52. Marbach, D., Schaffter, T., Mattiussi, C. & Floreano, D. Generating realistic *in silico* gene networks for performance assessment of reverse engineering methods. *J. Comput. Biol.* **16**, 229–239 (2009).
53. Schaffter, T., Marbach, D. & Floreano, D. GeneNetWeaver: *in silico* benchmark generation and performance profiling of network inference methods. *Bioinformatics* **27**, 2263–2270 (2011).
54. Hu, Z., Killion, P.J. & Iyer, V.R. Genetic reconstruction of a functional transcriptional regulatory network. *Nat. Genet.* **39**, 683–687 (2007).
55. Faith, J.J. *et al.* Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental metadata. *Nucleic Acids Res.* **36**, D866–D870 (2008).