JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 22, Number 4, 2015
© Mary Ann Liebert, Inc.
Pp. 253–265
DOI: 10.1089/cmb.2014.0290

Research Articles

Multi-Species Network Inference Improves Gene Regulatory Network Reconstruction for Early Embryonic Development in *Drosophila*

ANAGHA JOSHI[1] YVONNE BECK,[2,*] and TOM MICHOEL[2]

ABSTRACT

Gene regulatory network inference uses genome-wide transcriptome measurements in response to genetic, environmental, or dynamic perturbations to predict causal regulatory influences between genes. We hypothesized that evolution also acts as a suitable network perturbation and that integration of data from multiple closely related species can lead to improved reconstruction of gene regulatory networks. To test this hypothesis, we predicted networks from temporal gene expression data for 3,610 genes measured during early embryonic development in six *Drosophila* species and compared predicted networks to gold standard networks of ChIP-chip and ChIP-seq interactions for developmental transcription factors in five species. We found that (i) the performance of single-species networks was independent of the species where the gold standard was measured; (ii) differences between predicted networks reflected the known phylogeny and differences in biology between the species; (iii) an integrative consensus network that minimized the total number of edge gains and losses with respect to all single-species networks performed better than any individual network. Our results show that in an evolutionarily conserved system, integration of data from comparable experiments in multiple species improves the inference of gene regulatory networks. They provide a basis for future studies on the numerous multispecies gene expression datasets for other biological processes available in the literature.

Key words: functional genomics, gene networks, genomics, graphs and networks.

1. INTRODUCTION

In systems biology it is hypothesized that causal regulatory influences between transcription factors (TFs) and their target genes can be reconstructed by observing changes in gene expression levels during dynamic processes or in response to perturbing the cell by gene mutations or extracellular signals (Ideker et al., 2001; Kitano, 2002). As increasing amounts of gene expression data have become available, numerous computational and statistical methods have been developed to address the gene network inference problem

[1]Division of Developmental Biology and [2]Division of Genetics and Genomics, The Roslin Institute, The University of Edinburgh, Midlothian, Scotland, United Kingdom.
*Current address: Institute for Applied System Dynamics, Aalen University, Aalen, Germany.

[reviewed by Albert (2007); Bansal et al. (2007); Emmert-Streib et al. (2012); Friedman (2004); Gardner and Faith (2005); and Marbach et al. (2012a)]. Spurred by the observation that different methods applied to the same dataset can uncover complementary aspects of the underlying regulatory network (Marbach et al., 2010; Michoel et al., 2009), it is now firmly established that community-based methods that integrate predictions from multiple methods perform better than individual methods (Marbach et al., 2012a). A dimension that has remained unexplored in gene regulatory network inference is evolution: Does the integration of data from multiple related species lead to improved network inference performance? Numerous comparative analyses of gene expression data from multiple species have been performed (Bergmann et al., 2003; Brawand et al., 2011; Ihmels et al., 2005; Kalinka et al., 2010; Llinas et al., 2006; Lu et al., 2009; Miller et al., 2010; Movahedi et al., 2012; Mutwil et al., 2011; Rhind et al., 2011; Romero et al., 2012; Roy et al., 2013; Stuart et al., 2003; Thompson et al., 2013; Tirosh et al., 2006; Wang et al., 2009), but invariably these have studied conservation and divergence of individual gene expression profiles or coexpression modules. However, it is known that (co)expression can be conserved despite divergence of upstream *cis*-regulatory sequences, and although shuffling of TF-binding sites does not necessarily alter the topology of the TF–target network, cases have been documented where conserved coexpression modules are regulated by different TFs in different species (''TF switching'') (reviewed in Weirauch and Hughes, 2010). It is therefore not *a priori* obvious if and how multispecies expression data can be harnessed for gene regulatory network inference.

To address this question we decided to focus on a regulatory model system that is well characterized and conserved across multiple species. We were therefore particularly interested in a study where gene expression was measured at several time points during early embryonic development in six *Drosophila* species, including the model organism *D. melanogaster* (Kalinka et al., 2010). Early development of the animal body plan is a highly conserved process, controlled by gene regulatory network components resistant to evolutionary change (Davidson and Erwin, 2006). Furthermore, the binding sites of around half of all sequence-specific regulators controlling transcription in the blastoderm in *D. melanogaster* have been mapped on a genome-wide scale by ChIP-chip (MacArthur et al., 2009), and for several of these factors additional binding profiles mapped by ChIP-sequencing are available in other *Drosophila* species (Bradley et al., 2010; He et al., 2011; Paris et al., 2013). In this study, we took advantage of these unique gold standard networks of regulatory interactions across multiple species to predict and evaluate gene regulatory networks from gene expression data in six species, study their phylogeny and biology, and analyze how an integrated multispecies approach improves network inference performance.

2. RESULTS

2.1. Evolutionary and developmental dynamics have comparable effects on gene expression

We collected gene expression data for 3,610 genes in six *Drosophila* species measured at 9–13 time points during early embryonic development with 3–8 replicates per time point (200 samples in total) (Kalinka et al., 2010). To obtain a global view on the similarities and differences between samples, we performed multidimensional scaling using Sammon's nonlinear mapping criterion on the 3,610-dimensional sample vectors (cf., Methods Section and Fig. 1a). The first (horizontal) axis of variation corresponded to developmental time, with samples ordered along this dimension according to increasing developmental time points, while the second (vertical) axis of variation corresponded to evolutionary distance, with samples ordered along this dimension according to species. By expanding these two axes of variation into principal components, we found that the ''developmental'' dimension explained 34% of the total variation in the data, while the ''evolutionary'' dimension explained 11% (cf., Methods Section). This result confirms that variations in gene expression levels across *Drosophila* species at the same developmental time point are not greater than variations across time points within the same species. In this study, we were interested in whether this additional layer of interspecies expression variation can be harnessed in the reconstruction of gene regulatory networks.

2.2. Single-species network reconstruction recovers known transcriptional regulatory interactions in early Drosophila development

We used the context-likelihood of relatedness (CLR) algorithm (Faith et al., 2007) with Pearson correlation as a similarity measure to predict regulatory interactions in each species separately from the

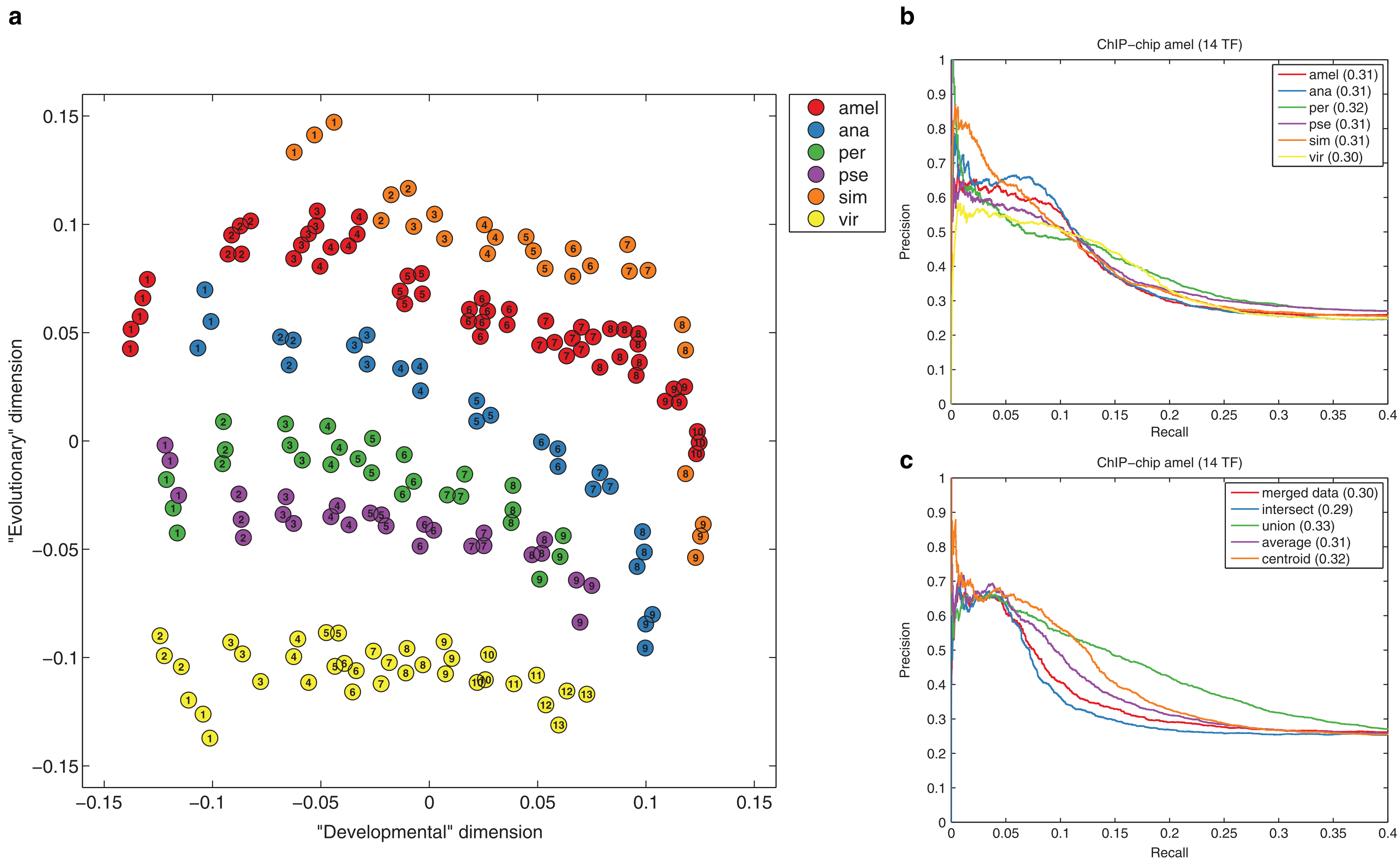


FIG. 1. **(a)** Two-dimensional scaling plot of the gene expression data using Sammon's nonlinear mapping criterion. Each dot represents one sample (200 samples total) positioned such that the two-dimensional distances reflect the Euclidean distances between the 3610-dimensional data vectors. Samples are colored by species and the number in each dot is the developmental time point of the sample. **(b)** Recall vs. precision curves for predicted regulatory networks in six *Drosophila* species using a gold-standard network of ChIP-chip interactions for 14 transcription factors (TFs) in *D. melanogaster*.

developmental gene expression data. As candidate regulators we used a set of 14 sequence-specific transcription factors (TFs) present on the expression array whose binding sites have been mapped by ChIP-chip in *D. melanogaster* at developmental time points relevant for the present study (MacArthur et al., 2009). A gold standard network of known transcriptional regulatory interactions in *D. melanogaster* development was constructed by assigning binding sites of these TFs to their closest gene (cf., Methods Section). The gold standard network was dense (25% of all possible edges were present), consistent with the fact that genes on the expression array were selected from genes known to be expressed during embryonic development (Kalinka et al., 2010) and that the 14 TFs comprise one-third of all sequence-specific regulators controlling transcription in the *D. melanogaster* blastoderm embryo (MacArthur et al., 2009).

We compared the predicted regulatory networks in all six species to the *D. melanogaster* gold standard network using standard recall and precision measurements (Stolovitzky et al., 2009). Without exception all six predicted networks showed percentages of true positives close to or in excess of 50% at a recall level of 10%, corresponding to networks with 1,300–1,400 predicted interactions (Table 1 and Fig. 1b). Any differences in performance between species were found to be small (nearly identical areas under the curve [AUC], Fig. 1b). The recall cut-off of 10% in Table 1 was chosen because it was closest for most species to the inflection point where precision starts to drop more rapidly with increasing recall. The levels of accuracy in network prediction obtained here have previously only been observed for bacteria (Marbach et al., 2012a; Michoel et al., 2009) and demonstrate the importance of using a gold standard network measured in an appropriate experimental condition. Indeed, when we used the more heterogeneous modENCODE (Roy et al., 2010) or Flynet (Marbach et al., 2012b) *D. melanogaster* reference networks, performance dropped dramatically (data not shown).

2.3. Chip-sequencing data confirms similar network reconstruction performance independent of species

Although the gold standard network reconstructed from ChIP-chip data was in *D. melanogaster*, perhaps surprisingly the *D. melanogaster* predicted network did not perform better overall than the networks predicted for the other species (Fig. 1b). To get confidence in this observation, we downloaded ChIP-sequencing data for three TFs (BCD, KR, HB) in three *Drosophila* species (*melanogaster, pseudoobscura*, and *virilis*) (Paris et al., 2013) and one TF (TWI) in four species (*melanogaster, simulans, ananassae*, and *pseudoobscura*) (He et al., 2011), and created ChIP-seq gold standard networks for five species (cf.,

Table 1. Transcription Factors and Their Number of Target Genes

TF	*ChIP*	*Amel*	*Ana*	*Per*	*Pse*	*Sim*	*Vir*
zD	1166	158 (129)	145 (122)	171 (137)	154 (124)	163 (132)	132 (102)
kr	518	125 (86)	128 (86)	196 (125)	176 (109)	127 (80)	207 (143)
mad	40	11 (0)	0 (0)	1 (0)	4 (0)	0 (0)	0 (0)
bcd	157	13 (0)	4 (0)	0 (0)	0 (0)	0 (0)	0 (0)
cad	274	8 (0)	0 (0)	40 (7)	0 (0)	133 (7)	85 (13)
da	795	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
dl	1503	216 (163)	234 (183)	67 (52)	137 (110)	289 (216)	111 (83)
hb	358	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
hkb	206	131 (49)	181 (61)	167 (45)	172 (48)	135 (43)	122 (34)
prd	313	44 (21)	38 (15)	65 (28)	58 (27)	41 (10)	55 (22)
run	158	134 (52)	117 (49)	186 (56)	154 (56)	127 (47)	167 (62)
slp1	212	178 (57)	155 (45)	221 (62)	192 (57)	154 (47)	192 (54)
sna	291	170 (78)	169 (73)	207 (83)	191 (76)	174 (72)	197 (81)
twi	1163	98 (80)	96 (81)	177 (120)	153 (108)	74 (61)	149 (121)
Total	7154	1286	1267	1498	1391	1417	1417
Precision		0.56	0.56	0.48	0.51	0.50	0.50

Transcription factors and their number of target genes in the *D. melanogaster* ChIP-chip gold standard network and in the predicted networks for six *Drosophila* species at the 10% recall level (in brackets for each TF the number of true positive predictions). The bottom two rows are the total number of interactions in each network and the overall precision (percentage of true positives) of the predicted networks. TF, transcription factor, amel, *D. melanogaster;* ana, *D. ananassae*; per, *D. persimilis*; pse, *D. pseudoobscura*; sim, *D. simulans;* vir, *D. virilis*.

Methods Section). The recall-precision curves generated from the *D. melanogaster* ChIP-seq gold standard network (Supplementary Fig. S1b, available online at www.liebertpub.com/cmb) were in good agreement with the ChIP-chip data, demonstrating again that the *D. melanogaster* predicted network performed no better than other *Drosophila* species. We also calculated recall-precision curves using the *D. ananassae, D. pseudobscura, D. simulans*, and *D. virilis* ChIP-seq gold standard networks. Again, the regulatory network in that species did not perform better compared to the other species (Supplementary Fig. S1c–f).

2.4. Reconstructed regulatory networks are enriched for ubiquitous interactions

The result that network reconstruction performance is similar across species regardless of the species origin of the gold standard network suggests that each species-specific dataset represents a different perturbation of an underlying conserved regulatory network. To better understand how the predicted networks in each species relate to each other, we analyzed the reconstructed regulatory networks at the 10% recall level (Table 1) in greater detail. Taken together, these networks contained 3,329 regulatory interactions between 14 TFs and 1098 genes. About 10% of these interactions (382) were predicted in all species. To systematically evaluate if this overlap can occur by chance, we randomized independently each interaction network keeping its in- and out-degree distribution constant and calculated the frequencies of having one to six edges overlap in 100 randomized networks. The predicted networks were significantly enriched for interactions ubiquitous to all species (Z-score = 37.7) and depleted for species-specific interactions (Z-score = -39.5) (Fig. 2a).

We then calculated if individual TFs were biased toward species-specific or ubiquitous interactions. Zygotic factors such as SNA ($P = 9.8 \times 10^{-60}$) shared statistically significant predicted targets among all six species, whereas maternal factors such as CAD did not share a single target across the six species. This, together with the observation that early zygotic genes at sequence level evolved much slower (Mensch et al., 2013), leads to the hypothesis that not only the sequences of early zygotic lineage genes but also the transcriptional program controlling their expression has evolved slower. The early zygotic genes are indeed overrepresented in the targets with conserved interactions across all species ($P = 1.2 \times 10^{-5}$).

The observation that prediction performance is independent of species (Fig. 1b) could be explained if only ubiquitous interactions (predicted in all species) were true positives. Although more true positives are found among interactions shared by three or more species than expected based on the total distribution of predicted interactions (Fig. 2b), and with precision increasing by the number of species (Fig. 2c), ubiquitous interactions account for only 18% of all true positives. Another possible explanation for the species-independent performance could be that binding events are highly conserved across species. Although it has been noted that more than 90% of TF binding sites overlapped between *D. melanogaster* and the closely related *D. yakuba* (Bradley et al., 2010), less than 30% of those binding sites were also conserved in the more distant *D. pseudoobscura* (Paris et al., 2013). Furthermore, it is also not true that conserved gold standard interactions for these TFs (BCD, HB, and KR) are more likely to be inferred. Indeed, the recall for species-specific gold standard interactions or those conserved in two or three species for these factors in the 10% recall networks did not differ from the overall recall value (Fig. 2d). In contrast, for the factor TWI, gold standard interactions conserved in three or four species were more likely to be included in the 10% recall networks (recall values resp. 19% and 36%; Fig. 2d). This is consistent with a higher degree of binding site conservation for this factor, with up to 60% conserved binding sites across six species (He et al., 2011).

2.5. Differences between predicted transcriptional regulatory networks reflect known phylogeny and biology

Since conservation of predicted or known gold standard interactions across species does not fully explain the observed species-independent network reconstruction performance, we hypothesized that the differences between these networks are not solely due to random variations in the expression data. To analyze these differences, we constructed a phylogenetic tree between the species based on the gain or loss of predicted interactions using the principle of maximum parsimony. This method minimizes the number of state changes in all transitions in a tree and has been used previously to reconstruct the evolutionary history of species based on gene content (Martens et al., 2008) and to reconstruct and predict transition states of developmental lineage trees based on gene expression data (Joshi and Göttgens, 2011). Using a binary matrix representing the presence or absence of all 3,329 predicted TF-target interactions in each of the 10% recall networks, a rooted tree was reconstructed that split the species in three groups: *melanogaster* (top), *obscura* (middle), and *virilis* (bottom) (cf., Methods Section and Fig. 2d). This tree is in full agreement with the tree reconstructed based on gene content (Stark et al., 2007). To ensure the robustness of the tree, we applied a standard bootstrap procedure that predicted 100% bootstrap confidence on all branches of the tree (Fig. 2d). The parsimony tree, moreover, predicts the network state transitions at each branch in terms of interactions gained or lost at a given transition. The transitions show a bias toward gain of interactions at most branch points over the loss. This is probably due to the presence of a large number of species-specific interactions (Fig. 2a).

We further explored whether the nine branch points (numbered 1–9 in Fig. 2d) reflect the biology behind the evolution of the *Drosophila* species. We created gene lists at each branch point containing target genes that gained or lost transcriptional interactions at that branch point. The maximum number of genes (361) gained interactions from branch point ‘A’ to *D. virilis* and were enriched for neuron differentiation ($P=1.2 \times 10^{-6}$) and embryonic morphogenesis ($P=3.1 \times 10^{-8}$). Genes gaining interactions from branch point ‘D’ to *D. simulans* were enriched for response to organic substances ($P=3.4 \times 10^{-2}$), in line with the fact that *D. simulans*, unlike *D. melanogaster*, lives on diverse rotting, nonsweet substrates throughout the year (David et al., 2007). Gene ontology analysis of all target sets revealed that many gene sets were enriched for transcription regulation (Supplementary Table S1), that is, transcriptional regulators were more likely to gain or lose interactions in the network rewiring. At each branch point, we found TFs losing or gaining interactions more than expected by chance (Supplementary Table S2). For instance, SLP1 is predicted to lose its interactions with genes involved in wing disc formation only in *D. ananassae* while, dorsal (DL) is predicted to regulate mitochondrial genes only in the *melanogaster* subgroup. Taken together, a biologically relevant evolutionary network history can be reconstructed using the individual predicted regulatory networks in six *Drosophila* species.

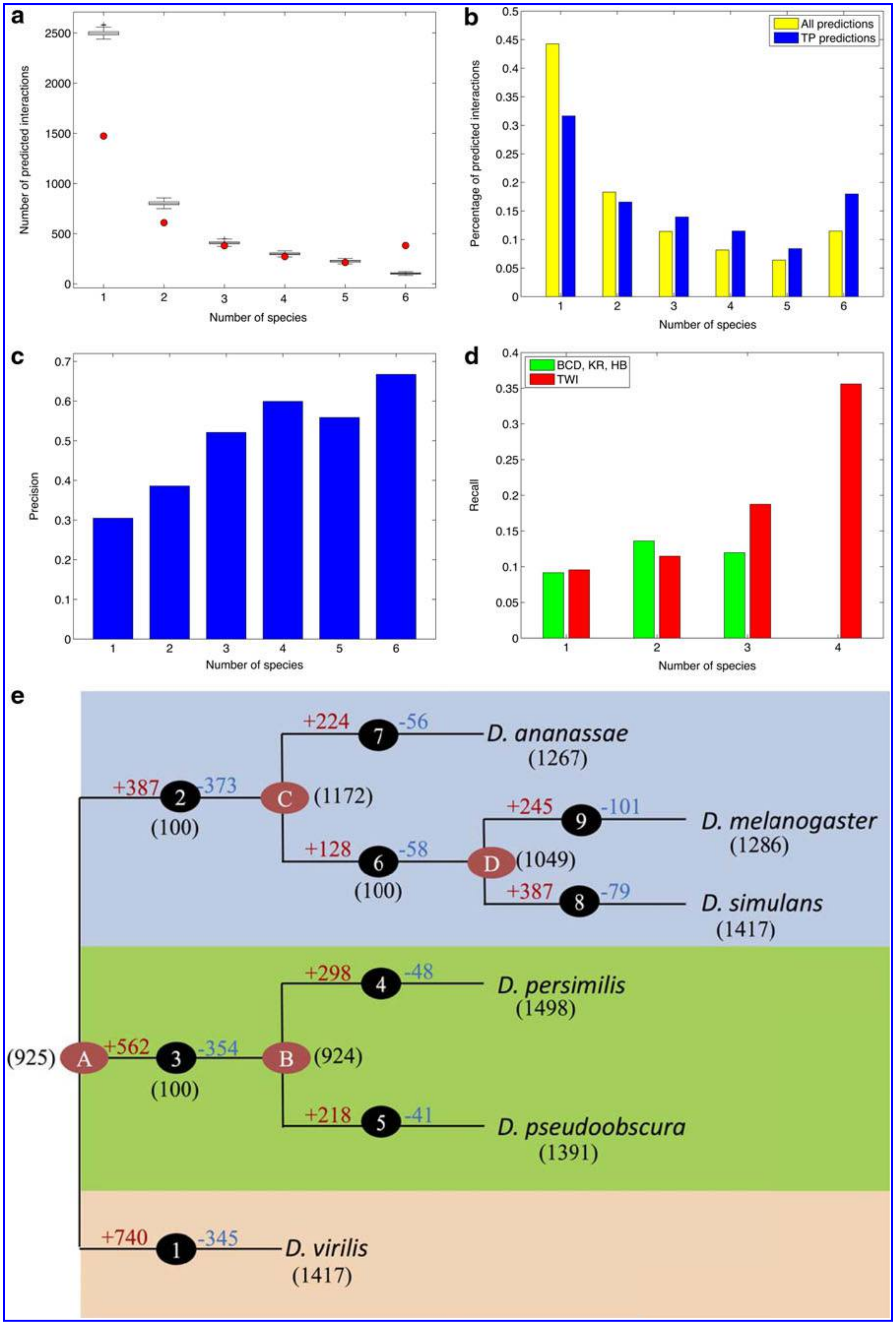
2.6. Multispecies analysis improves network reconstruction

It has been observed that different network inference algorithms applied to the same data uncover complementary aspects of the true underlying regulatory network (Marbach et al., 2010; Michoel et al., 2009) and this has formed the basis for integrative approaches that combine the predictions from multiple algorithms (Marbach et al., 2012a). In our case, since the networks predicted from different species equally well recover known transcriptional interactions while their differences reflect known phylogeny and biology, we reasoned that a multispecies analysis that combines predictions across species should also lead to a better network reconstruction. To test this hypothesis we considered several integrative approaches. Firstly, we combined the expression data from all species into one dataset to which we again applied the CLR algorithm (‘‘merged data’’ method). Secondly, we kept CLR scores from the individual species and applied rank-aggregation methods to derive an ‘‘intersection,’’ ‘‘union,’’ and ‘‘average’’ consensus ranking of predicted interactions (cf., Methods Section). Finally, motivated by the phylogenetic tree reconstruction, we also constructed a consensus ranking as the centroid of the six species-specific rankings for the cityblock distance, which for discrete networks corresponds to counting total number of edge gains and losses between two networks (‘‘centroid’’ method, see Methods Section for details).

To quantitatively compare different methods across different gold standard networks we considered the area under the recall–precision curve (AUC) and the precision at 10% recall (PREC10) as performance measures and converted them to *P*-values by comparison to AUCs and PREC10s of networks generated by randomly assigning ranks to all possible edges in the corresponding gold standard network (cf., Methods Section and Supplementary Fig. S2 for the recall vs. precision curves). While the AUC assesses the overall

→

FIG. 2. **(a)** Number of interactions found in 1–6 species in the inferred gene regulatory networks at 10% recall level (*red dots*) and in 100 randomized networks with the same in- and out-degree distribution as the inferred networks (*boxplots*). **(b)** Percentage of all predicted interactions (*yellow*) and of all true-positive predictions (*blue*) in 1–6 species. **(c)** Precision of interactions found in 1–6 species. **(d)** Recall of ChIP-seq gold-standard interactions conserved in 1–3 species (*green*; data for BCD, KR, and HB) and 1–4 species (*red*; data for TWI). **(e)** Phylogenetic tree between 6 *Drosophila* species reconstructed from the inferred interactions at 10% recall level, with the total number of interactions in each species shown *in brackets*. The tree correctly splits the species in three groups: *melanogaster* (top), *obscura* (middle), and *virilis* (bottom). Each branch (*numbered 1–9*) represents a inferred network state transition. At each network state transition, the number of interactions inferred to be gained (*red*) or lost (*blue*) as well as the bootstrap value for each branch (*in brackets*) is indicated.


a
Number of predicted interactions
0
500
1000
1500
2000
2500
Number of species
1
2
3
4
5
6
b
Percentage of predicted interactions
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
All predictions
TP predictions
Number of species
1
2
3
4
5
6
c
Precision
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Number of species
1
2
3
4
5
6
d
Recall
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
BCD, KR, HB
TWI
Number of species
1
2
3
4
e
+224
7
-56
D. ananassae
(1267)
+387
2
-373
(100)
C
(1172)
+245
9
-101
D. melanogaster
(1286)
+128
6
-58
(100)
D
(1049)
+387
8
-79
D. simulans
(1417)
+298
4
-48
D. persimilis
(1498)
(925)
A
+562
3
-354
(100)
B
(924)
+218
5
-41
D. pseudoobscura
(1391)
+740
1
-345
D. virilis
(1417)

performance of a predicted network, PREC10 measures the quality of the top-ranked predictions, a property that may be of greater practical relevance. This analysis showed that no predicted network performs best for either measure across all gold standards (Fig. 3a–f). The single-species *virilis* networks performed best for 5 out of 12 AUC and PREC10 scores, albeit not for the ChIP-seq network measured in its own species. This overall good performance is consistent with *virilis* having the highest number of measured time points in the data (Supplementary Table S3). *D. melanogaster* also had more data points available than the other four species, but its time series were less complete (Supplementary Table S3). Among the integrative methods, the centroid and union methods both performed best for 5 out of 12 AUC and PREC10 scores (Fig. 3a–f). Both also had higher average AUC scores than the best single-species network, but only the centroid method had a higher average PREC10 score than the best single-species network (Fig. 3g). The most important result however is the fact that the single-species network for the species where the gold standard network was measured never has the highest single-species AUC and only twice has the highest PREC10. In contrast, the centroid method always performs as good, and in most cases better, than the single-species network for the reference species (Fig. 3a–f). We conclude that the centroid method is the most robust network integration method achieving consistently high AUC and PREC10 scores, at least on this dataset.

3. DISCUSSION

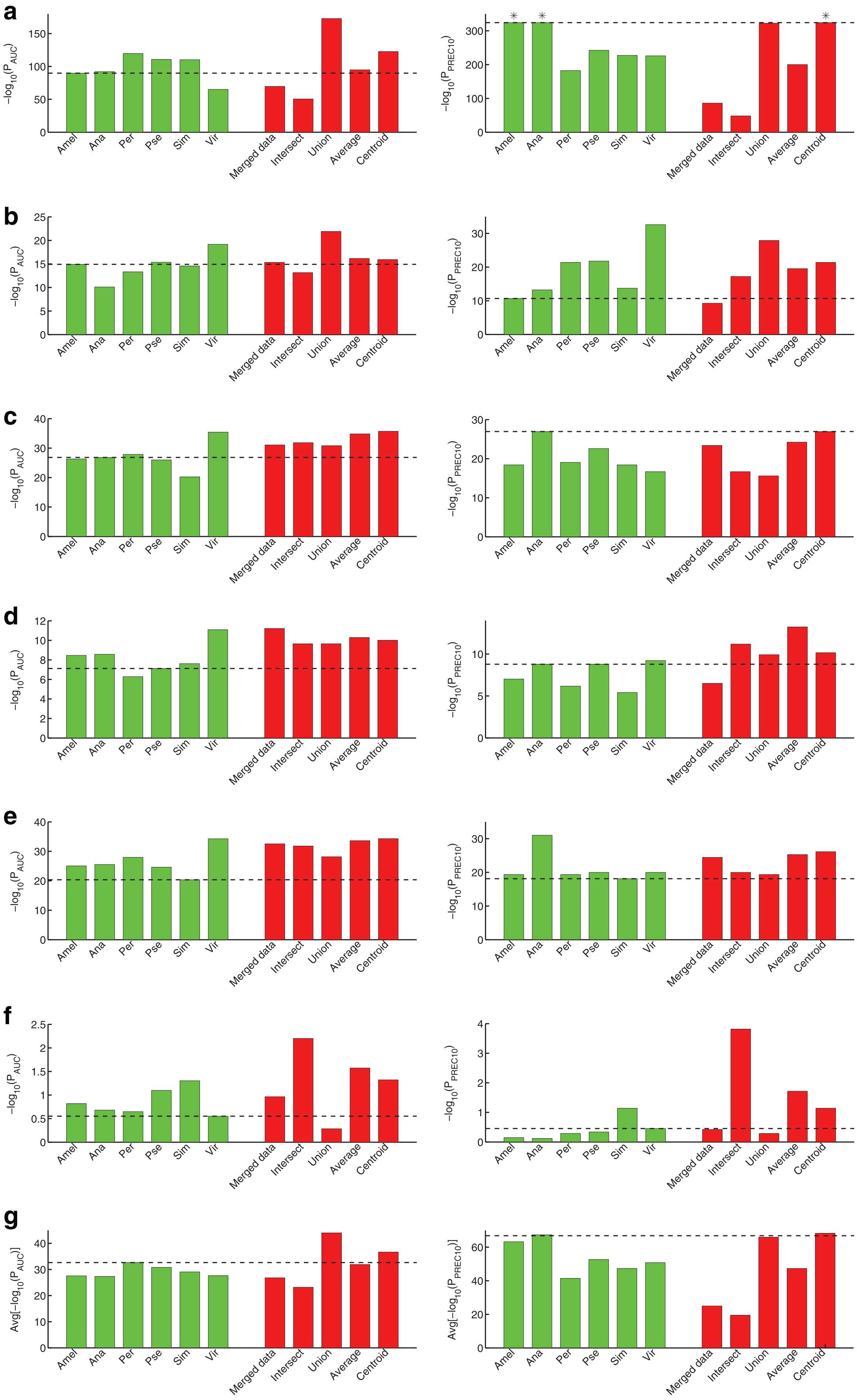
Here we predicted and evaluated developmental gene regulatory networks from temporal gene expression data in six *Drosophila* species, studied their phylogeny and biology, and analyzed how an integrated multispecies analysis improved network inference performance using gold standard networks of regulatory interactions measured by ChIP-chip and ChIP-seq in five species.

We unexpectedly found that network prediction performance of the single-species networks was independent of the species where the gold standard was measured. With precision values around or greater than 50% at a recall level of 10% for all predicted networks, this result was clearly not due to poor overall prediction performance. Although there was a trend that interactions predicted in all species had higher precision than interactions predicted in only one species and that conserved interactions in the gold standard networks for at least one of the TFs had higher chance to be correctly predicted, neither trend was sufficiently strong to account for the observed performance similarities.

An alternative or additional explanation could be that the ''true'' gene expression and binding profiles are highly conserved between these six species, but the observed profiles show species-dependent variation due to the inherent noisyness of high-throughput data. Because such random fluctuations in gene expression and binding data would be unrelated, one would then indeed expect similar performance independent of species. This explanation, however, conflicts with the published findings that binding divergence for these TFs increases with evolutionary distance and our observation that the differences between the predicted regulatory networks are consistent with the known phylogeny and differences in biology between these six *Drosophila* species. Future work in other species will have to elucidate if the observed species-independent performance is an artefact of this particular dataset, a consequence of the highly conserved nature of the underlying biological process, or a more general feature of this type of analysis. Motivated by the result that all species-specific networks showed good inference performance and that their differences reflected true phylogenetic relations, we pursued integrative approaches whereby predicted networks from all species were combined into consensus networks. In addition to established aggregation methods such as taking the intersection, union, or rank average of individual predictions, we also considered a novel centroid method that minimizes the total sum of edge gains and losses with respect to all individual networks. Multispecies methods showed better overall performance than the single-species networks, consistent with the

→

FIG. 3. Performance scores with respect to the gold-standard ChIP-chip network for 14 TFs in *D. melanogaster* **(a)** and the ChIP-seq networks for *D. melanogaster* (**b**, 4 TFs), *D. ananassae* (**c**, 1 TF), *D. pseudoobscura* (**d**, 4 TFs), *D. simulans* (**e**, 1 TF), *D. virilis* (**f**, 4 TFs), and their averages over all gold-standard networks **(g)**. In each panel, the left and resp. right figures show $-\log 10(P_{\mathrm{AUC}})$ and resp. $-\log 10(P_{\mathrm{PREC10}})$ for the six single-species predicted networks (*green*) and the five prediction aggregation methods (*red*). The *dashed lines* indicate the performance level of the single-species network for the gold-standard species **(a–f)** or of the best performing single-species network **(g)**. Values with an *asterisk* (*) in **(a)** indicate numerical underflow values truncated to the smallest nonzero *p*-value (10^{-324}).


a
$-\log_{10}(P_{AUC})$
0
50
100
150
Amel
Ana
Per
Pse
Sim
Vir
Merged data
Intersect
Union
Average
Centroid
$-\log_{10}(P_{PREC10})$
*
*
*
0
100
200
300
Amel
Ana
Per
Pse
Sim
Vir
Merged data
Intersect
Union
Average
Centroid
b
$-\log_{10}(P_{AUC})$
0
5
10
15
20
25
Amel
Ana
Per
Pse
Sim
Vir
Merged data
Intersect
Union
Average
Centroid
$-\log_{10}(P_{PREC10})$
0
10
20
30
Amel
Ana
Per
Pse
Sim
Vir
Merged data
Intersect
Union
Average
Centroid
c
$-\log_{10}(P_{AUC})$
0
10
20
30
40
Amel
Ana
Per
Pse
Sim
Vir
Merged data
Intersect
Union
Average
Centroid
$-\log_{10}(P_{PREC10})$
0
10
20
30
Amel
Ana
Per
Pse
Sim
Vir
Merged data
Intersect
Union
Average
Centroid
d
$-\log_{10}(P_{AUC})$
0
2
4
6
8
10
12
Amel
Ana
Per
Pse
Sim
Vir
Merged data
Intersect
Union
Average
Centroid
$-\log_{10}(P_{PREC10})$
0
5
10
Amel
Ana
Per
Pse
Sim
Vir
Merged data
Intersect
Union
Average
Centroid
e
$-\log_{10}(P_{AUC})$
0
10
20
30
40
Amel
Ana
Per
Pse
Sim
Vir
Merged data
Intersect
Union
Average
Centroid
$-\log_{10}(P_{PREC10})$
0
10
20
30
Amel
Ana
Per
Pse
Sim
Vir
Merged data
Intersect
Union
Average
Centroid
f
$-\log_{10}(P_{AUC})$
0
0.5
1
1.5
2
2.5
Amel
Ana
Per
Pse
Sim
Vir
Merged data
Intersect
Union
Average
Centroid
$-\log_{10}(P_{PREC10})$
0
1
2
3
4
Amel
Ana
Per
Pse
Sim
Vir
Merged data
Intersect
Union
Average
Centroid
g
Avg[$-\log_{10}(P_{AUC})$]
0
10
20
30
40
Amel
Ana
Per
Pse
Sim
Vir
Merged data
Intersect
Union
Average
Centroid
Avg[$-\log_{10}(P_{PREC10})$]
0
20
40
60
Amel
Ana
Per
Pse
Sim
Vir
Merged data
Intersect
Union
Average
Centroid

observation that correct predictions are not restricted to interactions predicted in all species. Of note, the single-species network matching the gold standard species was almost never the best performing single-species network. Because in real-world applications the aim of network inference is usually to reconstruct a TF-target network for a species of interest in the absence of gold standard ChIP-seq/chip data, our results suggest that by combining predicted networks from multiple closely related species, a better network will be inferred than by using data for the species of interest only, and that the combined network is likely to perform better, or at least as good as, the best single-species network. A novel multispecies network integration method that reconstructs an ''ancestral'' network minimizing the number of edge gains and losses to each single-species network appeared to be particularly promising in this regard.

Our work has shown that in an evolutionarily conserved system such as early embryonic development, integration of data from comparable experiments in multiple species improves the inference of gene regulatory networks. Although the data for the present study came from a well-controlled experiment in a model organism, with matching time-course data adjusted for differences in developmental time between species, our approach is based solely on comparing expression profiles of different genes within the same species, and expression levels in different species were never directly compared. We therefore expect that our results should also hold for other biological processes, when more heterogeneous data are used or when data from more distantly related species are combined, in order to cover the entire spectrum of available multispecies gene expression datasets.

4. METHODS

4.1. Gene expression data

Embryonic developmental time-course expression data in six *Drosophila* species [*D. melanogaster* (''amel''), *D. ananassae* (''ana''), *D. persimilis* (''per''), *D. pseudoobscura* (''pse''), *D. simulans* (''sim'') and *D. virilis* (''vir'')] was obtained from (Kalinka et al., 2010) (ArrayExpress accession code E-MTAB-404). The data consists of 10 (amel), 13 (vir), or 9 (ana, per, pse, sim) developmental time points with several replicates per time point, resulting in a total of 56 (amel), 36 (vir), or 27 (ana, per, pse, sim) arrays per species (Supplementary Table S3). The downloaded data was processed by averaging absolute expression levels over all reporters for a gene followed by taking the $\log_2$ transform.

4.2. Multidimensional scaling and variance explained

We used two-dimensional scaling using the Euclidean distance and Sammon's nonlinear mapping criterion on the 3,610-dimensional sample vectors using the built-in ''mdscale'' function of Matlab. To estimate the variance explained by each of the two dimensions, we first calculated the principal components of the data matrix. These are a set of 200 mutually orthogonal (200×1)–dimensional vectors, each explaining a proportion σ_i^2 of the total variance, that is, $\sum_{i=1}^{200} \sigma_i^2 = 1$. Each dimension in Figure 1 also corresponds to a (200×1) vector Y, and the proportion of variance explained by Y is found by expansion into principal components, $\sigma_Y^2 = \sum_{i=1}^{200} \sigma_i^2 (Y^T V_i)^2$, where it is assumed that Y and all V_i have unit norm. To correct for systematic biases in the data, genes were standardized to have mean zero and standard deviation one over all 200 samples.

4.3. ChIP-chip data

ChIP-chip data for 21 sequence-specific *Drosophila* transcription factors (TFs) measured in *D. melanogaster* embryos was obtained from MacArthur et al., (2009). We considered the 1% FDR bound regions and defined target genes for each TF by assigning to each bound region its closest gene, if the distance between the region and the gene was less than 5,000 base pairs. For TFs with repeat measurements, target lists were defined by taking the union over replicates. Fourteen of the TFs were present on the array and used to construct a gold standard regulatory network.

4.4. ChIP-sequencing data

The peaks for three transcription factors present on the array (BCD, HB, and KR) for three species (*D. melanogaster, D. pseudoobscura*, and *D. virilis*) were obtained from Paris et al., (2013). Genes with normalized peak height greater than 0 were selected as the gold standard targets of a given transcription factor. The peaks for one factor (TWI) for four species (*D. melanogaster, D. ananassae, D. pseudoobscura*, and *D. simulans*) were obtained from He et al., (2011). Peaks were mapped to the nearest transcription start site of genes by using the gene annotation from FlyBase (FB2013_03). Genes with peak height greater than 10 were selected as the gold standard targets for each species.

4.5. Transcriptional regulatory network reconstruction

We used the CLR (Context Likelihood of Relatedness) algorithm (Faith et al., 2007) using Pearson correlation as a similarity measure to predict transcriptional regulatory networks in each species using the aforementioned 14 TFs as candidate regulators. Because the CLR algorithm only considers the right-hand tail of similarity values for every TF–gene combination, in theory the absolute values of the Pearson correlations should be provided to the CLR software. However, we observed improved performance with respect to all gold standard networks when the Pearson correlations were *not* transformed to absolute values before calling the CLR algorithm (effectively ignoring negative correlations) and therefore used this approach for all reported results. Pearson correlation followed by CLR also performed better than the default mutual information similarity measure followed by CLR as well as using Pearson correlation or mutual information without CLR (data not shown).

4.6. Phylogenetic tree construction

We created a binary matrix of 3,329 rows and 6 columns representing predicted TF–target interactions in each species at a CLR *Z*-score cutoff corresponding to 10% recall with respect to the *D. melanogaster* ChIP-chip network. In this matrix, the $(i, j)^{\text{th}}$ element denotes whether the interaction i is present in the species j or not. Network states and state changes were mapped onto the branches of inferred phylogenetic trees using the PARS program from the PHYLIP package (Felsenstein, 1996) by defining *D. virilis* as the root of the tree. Bootstrapping was performed using the SEQBOOT program from the PHYLIP package where 100 datasets were generated by randomly replacing a given six-species network matrix. A consensus tree with a bootstrap confidence on each branch of the tree was reconstructed using the CONSENSE program from the PHYLIP package.

4.7. Enrichment analyses

Gene set enrichment for each phylogenetic tree state change was calculated using the DAVID suite of programs (Huang et al., 2008). For each transcription factor, enrichment of overlap of the candidate target gene set with each transition state gene set was calculated using a hypergeometric test. Early zygotic, late zygotic, maternal, and adult gene lists were downloaded from Mensch et al., (2013), and enrichment was calculated using a hypergeometric test.

4.8. Prediction aggregation methods

To combine predicted networks from multiple species, we considered five prediction aggregation methods. The first method combined expression data from all species into one dataset, to which we again applied the CLR (''merged data'') method. For the four other methods, predictions from each species were first ranked by their respective CLR-scores, such that the highest score received the highest rank value and tied values were given their average rank value, using Matlab's ''tiedrank'' function. Three methods used standard functions to combine the edge ranks of the six single-species predicted networks, namely the *minimum* (''intersection'' method), *maximum* (''union'' method), and *average* (''average'' method) rank value. The intersection and union methods are named such because if a threshold is used to convert a fixed number of top-ranked predictions to a binary graph, these would result precisely in the intersection and union of the binary graphs over all species. Conversely, on binary graphs, the phylogenetic tree construction infers ancestral networks by minimizing the number of edge gains and losses between single-species networks. This corresponds to minimizing their ''cityblock'' distance, defined for two n-dimensional vectors x and y as $d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$. As the final prediction aggregation method, we therefore considered the centroid network for the six single-species networks for the cityblock distance, defined by the edge weights C_{ij}, which minimize

$$\sum_{k=1}^{6} d(C,\ R^{(k)}) = \sum_{k=1}^{6} \sum_{ij} |C_{ij} - R_{ij}^{(k)}|$$

where $R^{(k)}$ is the matrix of edge rank values for species k. The matrix C is easily computed using Matlab's ''kmeans'' function by specifying the cityblock distance and grouping species into one cluster.

4.9. Network reconstruction performance

To compare the network reconstruction performance of several predicted networks across multiple gold standard networks, we used the area under the precision-recall curve (AUC) and the precision at 10% recall (PREC10) as scoring measures. Absolute scores were converted to *P*-values following established protocols of the DREAM project (Marbach et al., 2012a). Briefly, 100,000 random predictions were generated for each gold standard network by assigning a random rank to each possible TF-target interaction. Next, an asymmetric stretched exponential function of the form

$$f(x) = \begin{cases} he^{-b+(x-x_{\max})^{c+}} & x \geq x_{\max} \\ he^{-b-(x_{\max}-x)^{c-}} & x < x_{\max} \end{cases}$$

was fitted to each histogram (1000 bins) of random AUCs and random PREC10s, using the ''fit'' function in Matlab's curve fitting toolbox. Finally, *P*-values for real AUCs and PREC10s were calculated by integrating the right tail of the corresponding fitted and normalized stretched exponential distribution function using Matlab's ''trapz'' function.

AUTHOR DISCLOSURE STATEMENT

The authors declare that no competing financial interests exist.

REFERENCES

Albert, R. 2007. Network inference, analysis, and modeling in systems biology. *Plant Cell* 19, 3327–3338.

Bansal, M., Belcastro, V., Ambesi-Impiombato, A., et al. 2007. How to infer gene networks from expression profiles. *Mol. Syst. Biol.* 3, 78.

Bergmann, S., Ihmels, J., and Barkai, N. 2003. Similarities and differences in genome-wide expression data of six organisms. *PLoS Biol.* 2, e9.

Bradley, R.K., Li, X.-Y., Trapnell, C., et al. 2010. Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related *Drosophila* species. *PLoS Biol.* 8, e1000343.

Brawand, D., Soumillon, M., Necsulea, A., et al. 2011. The evolution of gene expression levels in mammalian organs. *Nature* 478, 343–348.

David, J.R., Lemeunier, F., Tsacas, L., et al. 2007. The historical discovery of the nine species in the drosophila melanogaster species subgroup. *Genetics* 177, 1969–1973.

Davidson, E.H., and Erwin, D.H. 2006. Gene regulatory networks and the evolution of animal body plans. *Science* 311, 796–800.

Emmert-Streib, F., Glazko, G., De Matos Simoes, R., et al. 2012. Statistical inference and reverse engineering of gene regulatory networks from observational expression data. *Front. Genet.* 3, 8.

Faith, J.J., Hayete, B., Thaden, J.T., et al. 2007. Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. *PLoS Biol.* 5, e8.

Felsenstein, J. 1996. Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Methods Enzymol.* 266, 418–427.

Friedman, N. 2004. Inferring cellular networks using probabilistic graphicalmodels. *Science* 308, 799–805.

Gardner, T.S., and Faith, J.J. 2005. Reverse-engineering transcription control networks. *Phys. Life Rev.* 2, 65–88.

He, Q., Bardet, A.F., Patton, B., et al. 2011. High conservation of transcription factor binding and evidence for combinatorial regulation across six drosophila species. *Nat. Genet.* 43, 414–420.

Huang, D.W., Sherman, B.T., Lempicki, R.A., et al. 2008. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4, 44–57.

Ideker, T., Galitski, T., and Hood, L. 2001. A new approach to decoding life: Systems biology. *Annu. Rev. Genomics Hum. Genet.* 2, 343–372.

Ihmels, J., Bergmann, S., Berman, J., et al. 2005. Comparative gene expression analysis by a differential clustering approach: Application to the Candida albicans transcription program. *PLoS Genet.* 1, e39.

Joshi, A., and Göttgens, B. 2011. Maximum parsimony analysis of gene expression profiles permits the reconstruction of developmental cell lineage trees. *Dev. Biol.* 353, 440–447.

Kalinka, A.T., Varga, K.M., Gerrard, D.T., et al. 2010. Gene expression divergence recapitulates the developmental hourglass model. *Nature* 468, 811–814.

Kitano, H. 2002. Systems biology: A brief overview. *Science* 295, 1662–1664.

Llinas, M., Bozdech, Z., Wong, E.D., et al. 2006. Comparative whole genome transcriptome analysis of three *Plasmodium falciparum* strains. *Nucleic Acids Res.* 34, 1166–1173.

Lu, Y., Huggins, P., and Bar-Joseph, Z. 2009. Cross species analysis of microarray expression data. *Bioinformatics* 25, 1476–1483.

MacArthur, S., Li, X.-Y., Li, J., et al. 2009. Developmental roles of 21 drosophila transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions. *Genome Biol.* 10, R80.

Marbach, D., Costello, J.C., Küffner, R., et al. 2012a. Wisdom of crowds for robust gene network inference. *Nat. Methods* 9, 796–804.

Marbach, D., Prill, R.J., Schaffter, T., et al. 2010. Revealing strengths and weaknesses of methods for gene network inference. *Proc. Natl. Acad. Sci. USA* 107, 6286–6291.

Marbach, D., Roy, S., Ay, F., et al. 2012b. Predictive regulatory models in drosophila melanogaster by integrative inference of transcriptional networks. *Genome Res.* 22, 1334–1349.

Martens, C., Vandepoele, K., and Van de Peer, Y. 2008. Whole-genome analysis reveals molecular innovations and evolutionary transitions in chromalveolate species. *Proc. Natl. Acad. Sci. USA* 105, 3427–3432.

Mensch, J., Serra, F., Lavagnino, N.J., et al. 2013. Positive selection in nucleoporins challenges constraints on early expressed genes in drosophila development. *Genome Biol. Evol.* 5, 2231–2241.

Michoel, T., De Smet, R., Joshi, A., et al. 2009. Comparative analysis of module-based versus direct methods for reverse-engineering transcriptional regulatory networks. *BMC Syst. Biol.* 3, 49.

Miller, J.A., Horvath, S., and Geschwind, D.H. 2010. Divergence of human and mouse brain transcriptome highlights alzheimer disease pathways. *Proc. Natl. Acad. Sci. USA* 107, 12698–12703.

Movahedi, S., Van Bel, M., Heyndrickx, K.S., et al. 2012. Comparative coexpression analysis in plant biology. *Plant Cell Environ.* 35, 1787–1798.

Mutwil, M., Klie, S., Tohge, T., et al. 2011. Planet: Combined sequence and expression comparisons across plant networks derived from seven species. *Plant Cell* 23, 895–910.

Paris, M., Kaplan, T., Li, X., et al. 2013. Extensive divergence of transcription factor binding in *Drosophila* embryos with highly conserved gene expression. *PLoS Genet.* 9, e1003748.

Rhind, N., Chen, Z., Yassour, M., et al. 2011. Comparative functional genomics of the fission yeasts. *Science* 332, 930–936.

Romero, I.G., Ruvinsky, I., and Gilad, Y. 2012. Comparative studies of gene expression and the evolution of gene regulation. *Nat. Rev. Genet.* 13, 505–516.

Roy, S., Ernst, J., Kharchenko, P., et al. 2010. Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science* 330, 1787–1797.

Roy, S., Wapinski, I., Pfiffner, J., et al. 2013. Arboretum: reconstruction and analysis of the evolutionary history of condition-specific transcriptional modules. *Genome Res.* 23, 1039–1050.

Stark, A., Lin, M.F., Kheradpour, P., et al. 2007. Discovery of functional elements in 12 drosophila genomes using evolutionary signatures. *Nature* 450, 219–232.

Stolovitzky, G., Prill, R.J., and Califano, A. 2009. Lessons from the DREAM2 challenges. *Ann. NY Acad. Sci.* 1158, 159–195.

Stuart, J.M., Segal, E., Koller, D., et al. 2003. A gene-coexpression network for global discovery of conserved genetic modules. *Science* 302, 249–255.

Thompson, D.A., Roy, S., Chan, M., et al. 2013. Evolutionary principles of modular gene regulation in yeasts. *Elife* 2, e01114.

Tirosh, I., Weinberger, A., Carmi, M., et al. 2006. A genetic signature of interspecies variations in gene expression. *Nat. Genet.* 38, 830–834.

Wang, K., Narayanan, M., Zhong, H., et al. 2009. Metaanalysis of inter-species liver co-expression networks elucidates traits associated with common human diseases. *PLoS Comput. Biol.* 5, e1000616.

Weirauch, M.T., and Hughes, T.R. 2010. Conserved expression without conserved regulatory sequence: the more things change, the more they stay the same. *Trends Genet.* 26, 66–74.

Address correspondence to:
Dr. Tom Michoel
The Roslin Institute
The University of Edinburgh
The Roslin Institute Building
Easter Bush EH25 9RG
United Kingdom

E-mail: tom.michoel@roslin.ed.ac.uk