Multi-species network inference improves gene regulatory network reconstruction for early embryonic development in Drosophila

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,*]

[1]Division of Developmental Biology and [2]Division of Genetics and Genomics, The Roslin Institute, The University of Edinburgh, Midlothian EH25 9RG, Scotland, United Kingdom

[♯]Current address: Institute for Applied System Dynamics, Aalen University, Beethovenstrasse 1, 73430 Aalen, Germany

[*]Corresponding author, E-mail: tom.michoel@roslin.ed.ac.uk

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 which 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 multi-species gene expression datasets for other biological processes available in the literature.

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 extra-cellular signals [1, 2]. 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 (reviewed in [3–8]). Spurred by the observation that different methods applied to the same dataset can uncover complementary aspects of the underlying regulatory network [9, 10], it is now firmly established that community-based methods which integrate predictions from multiple methods perform better than individual methods [8]. 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 [11–26], but invariably these have studied conservation and divergence of individual gene expression profiles or co-expression 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 co-expression modules are regulated by different TFs in different species ("TF switching") (reviewed in [27]). It is therefore not a priori obvious if and how multi-species 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* [18]. Early development of the animal body plan is a highly conserved process, controlled by gene regulatory network components resistant to evolutionary change [28]. 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 [29] and for several of these factors additional binding profiles mapped by ChIP-sequencing are available in other *Drosophila* species [30–32]. 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 multi-species 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) [18]. To obtain a global view on the similarities and differences between

samples, we performed multi-dimensional scaling using Sammon's nonlinear mapping criterion on the 3,610-dimensional sample vectors (cf. Methods and Figure 1). 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). 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 whether this additional layer of inter-species 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 [33] with Pearson correlation as a similarity measure to predict regulatory interactions in each species separately from the 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 [29]. 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). 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 [18] and that the 14 TFs comprise one-third of all sequence-specific regulators controlling transcription in the *D. melanogaster* blastoderm embryo [29].

We compared the predicted regulatory networks in all six species to the *D. melanogaster* gold standard network using standard recall and precision measurements [34]. 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 Figure 2a). Any differences in performance between species were found to be small (nearly identical areas under the curve (AUC), Figure 2a). 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 (Figure 2a). The levels of accuracy in network prediction obtained here have previously only been observed for bacteria [8, 9] and demonstrate the importance of using a gold standard network measured in an appropriate experimental condition. Indeed, when we used the more heterogeneous modENCODE [35] or Flynet [36] *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 (Figure 2a). 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*) [32] and one TF (TWI) in four species (*melanogaster*, *simulans*, *ananassae* and *pseudoobscura*) [31], and created ChIP-seq gold standard networks for five species (cf. Methods). The recall-precision curves generated from the *D. melanogaster* ChIP-seq gold standard network (Figure 2b) 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 (Figure 2c–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 analysed 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. Figure 3a shows that the predicted networks were significantly enriched for interactions ubiquitous to all species (Z-score$= 37.7$) and depleted for species-specific interactions (Z-score$= -39.5$).

We then calculated if individual TFs were biased towards 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 [37] 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 (Figure 2) could be explained if only ubiquitous interactions (predicted in all species) were true positives. Although interactions predicted only in one species have a lower precision compared to interactions predicted in four or more species (Figure 3b), about a third of all true positives come from species-specific interactions. 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* [30], less than 30% of those binding sites were also conserved in the more distant *D. pseudoobscura* [32]. 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 (Figure 3c). 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%, Figure 3c). This is consistent with a higher degree of binding site conservation for this factor with up to 60% conserved binding sites across six species [31].

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

Since also the species-specific predicted interactions contain known gold standard interactions, we hypothesized that the differences between these networks are not solely due to random variations in the expression data. To analyse 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 minimises 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 [38] and to reconstruct and predict transition states of developmental lineage trees based on gene expression data [39]. 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 which split the species in three groups - *melanogaster* (top), *obscura* (middle), *virilis* (bottom) (cf. Methods and Figure 3d). This tree is in full agreement with the tree reconstructed based on gene content [40]. To ensure the robustness of the tree, we applied a standard bootstrap procedure which predicted 100% bootstrap confidence on all branches of the tree (Figure 3d). 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 towards 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 (Figure 3a).

We further explored whether the nine branch points (numbered 1–9 in Figure 3d) reflect the biology behind the evolution of the *Drosophila* species. We created gene lists at each branch point containing target genes which 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, non-sweet substrates throughout the year [41]. Gene ontology analysis of all target sets revealed that many gene sets were enriched for transcription regulation (Supplementary Table S1), i.e. 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 TableS2). 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 Multi-species 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 [9, 10] and this has formed the basis for integrative approaches which combine the predictions from multiple algorithms [8]. 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 multi-species analysis which 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 a consensus ranking of predicted interactions. More precisely, sorting interactions from high to low confidence, we defined a consensus rank as (i) the maximum rank over all species ("intersection" method, a prediction has low consensus rank (i.e. high confidence) if it has low rank in all species), (ii) the minimum rank over all species ("union" method, a prediction has low consensus rank if it has low rank in at least one species) and (iii) the average rank over all species ("average" method, also used in [8]). 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). Figure 4 shows the recall vs. precision curves for the five consensus networks against the six ChIP-chip and ChIP-seq gold standard networks.

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 Z-scores by comparison to AUCs and PREC10s of networks generated by randomly assigning ranks to all possible edges in the corresponding gold standard network. While the AUC assesses the overall performance of a predicted network, PREC10 measures the quality of the top-ranked predictions, a property that may be of greater practical relevance. As the overall score, we considered the average Z-scores over the six ChIP-chip and ChIP-seq gold standard networks (similar to the procedure of [8]). As expected, this analysis showed that no predicted network performs best for either measure across all gold standards (Table 2). The single-species *virilis* networks performed best for 5 out of 12 AUC-Z and PREC10-Z scores, albeit not for the ChIP-seq network measured in its own species, and achieves highest average PREC10-Z score of all single-species networks. 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 method performed best for 5 out of 12 AUC-Z and PREC10-Z scores. Overall, the union, average and centroid methods all had higher average AUC-Z score than the best single-species network, but only the centroid method had higher average PREC10-Z score than the best single-species net-

work. The centroid method had the highest average AUC-Z and PREC10-Z scores of all single- and multi-species methods and should be considered the optimal network integration method, at least on this dataset.

2.7 The centroid regulatory network is clustered according to temporal expression

Finally we investigated the topology and biology of the centroid network at the 50% precision level (corresponding to 12% recall) with respect to the *D. melanogaster* ChIP-chip network. This centroid network consists of two disconnected components with regulators in each component organized according to their temporal expression profile (Figure 5). The first (left) component predominantly consists of shared targets of the TFs D, KR, HKB, PRD, RUN, SLP1 and TWI whose expression profiles have a characteristic peak at the second time point. This component is enriched for the functional categories *developmental protein* ($P = 4.3 \times 10^{-32}$) and *cell fate determination* ($P = 2.5 \times 10^{-13}$). The second (right) component mainly consists of targets of the TF DL and is enriched for *mitochondrion* ($P = 3.6 \times 10^{-17}$) and *cofactor metabolic process* ($P = 8.9 \times 10^{-7}$). The functional enrichment of the first component is not surprising since the genes selected for the gene expression study were known to be expressed during embryonic development [18]. The functional enrichment of the second component however, together with the fact that many targets of DL (Dorsal) are validated by ChIP-chip data (Figure 5) suggests a new biological hypothesis that DL might regulate cellular energy metabolism processes.

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 multi-species 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 which minimizes the total sum of edge gains and losses with respect to all individual networks. The union, rank average and centroid method, but not the intersection method, showed better overall performance than the single-species networks, consistent with the observation that correct predictions are not restricted to interactions predicted in all species. The centroid method performed best overall and had the additional advantage of having a higher rate of true positives among the top-ranked predictions.

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. Future challenges will be to investigate if these results 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 multi-species gene expression datasets.

4 Methods

4.1 Gene expression data

Embryonic developmental time-course expression data in 6 *Drosophila* species (*D. melanogaster* ("amel"), *D. ananassae* ("ana"), *D. persimilis* ("per"), *D. pseudoobscura* ("pse"), *D. simulans* ("sim") and *D. virilis* ("vir")) was obtained from [18] (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 for each of 3610 genes on the array absolute expression levels over all reporters for that gene followed by taking the $\log_2$ transform.

4.2 Multi-dimensional scaling and variance explained

We used two-dimensional scaling using the Eucledian 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, i.e. $\sum_{i=1}^{200} \sigma_i^2 = 1$. 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 [29]. 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 [32]. 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 [31]. 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 [33] 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 [42] 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 [43]. 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 [37] and enrichment was calculated using a hypergeometric test.

References

[1] Ideker T, Galitski T and Hood L. A new approach to decoding life: systems biology. *Annu Rev Genomics Hum Genet* **2**:343–372 (2001).

[2] Kitano H. Systems biology: a brief overview. *Science* **295**:1662–1664 (2002).

[3] Friedman N. Inferring cellular networks using probabilistic graphical models. *Science* **308**:799–805 (2004).

[4] Gardner TS and Faith JJ. Reverse-engineering transcription control networks. *Phys Life Rev* **2**:65–88 (2005).

[5] Bansal M *et al.* How to infer gene networks from expression profiles. *Mol Syst Biol* **3**:78 (2007).

[6] Albert R. Network inference, analysis, and modeling in systems biology. *Plant Cell* **19**:3327–3338 (2007).

[7] Emmert-Streib F *et al.* Statistical inference and reverse engineering of gene regulatory networks from observational expression data. *Front Genet* **3**:8 (2012).

[8] Marbach D *et al.* Wisdom of crowds for robust gene network inference. *Nat Methods* **9**:796–804 (2012).

[9] Michoel T *et al.* Comparative analysis of module-based versus direct methods for reverse-engineering transcriptional regulatory networks. *BMC Syst Biol* **3**:49 (2009).

[10] Marbach D *et al.* Revealing strengths and weaknesses of methods for gene network inference. *PNAS* **107**:6286–6291 (2010).

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

[12] Stuart JM *et al.* A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302**:249–255 (2003).

[13] Ihmels J *et al.* Comparative gene expression analysis by a differential clustering approach: application to the Candida albicans transcription program. *PLoS Genet* **1**:e39 (2005).

[14] Llinas M *et al.* Comparative whole genome transcriptome analysis of three Plasmodium falciparum strains. *Nucl Acids Res* **34**:1166–1173 (2006).

[15] Tirosh I *et al.* A genetic signature of interspecies variations in gene expression. *Nat Genetics* **38**:830–834 (2006).

[16] Wang K *et al.* Meta-analysis of inter-species liver co-expression networks elucidates traits associated with common human diseases. *PLoS Comp Biol* **5**:e1000616 (2009).

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

[18] Kalinka AT *et al.* Gene expression divergence recapitulates the developmental hourglass model. *Nature* **468**:811–814 (2010).

[19] Miller JA, Horvath S and Geschwind DH. Divergence of human and mouse brain transcriptome highlights alzheimer disease pathways. *PNAS* **107**:12698–12703 (2010).

[20] Rhind N *et al.* Comparative functional genomics of the fission yeasts. *Science* **332**:930–936 (2011).

[21] Brawand D *et al.* The evolution of gene expression levels in mammalian organs. *Nature* **478**:343–348 (2011).

[22] Mutwil M *et al.* Planet: combined sequence and expression comparisons across plant networks derived from seven species. *Plant Cell* **23**:895–910 (2011).

[23] Romero IG, Ruvinsky I and Gilad Y. Comparative studies of gene expression and the evolution of gene regulation. *Nat Revi Genet* **13**:505–516 (2012).

[24] Movahedi S *et al.* Comparative co-expression analysis in plant biology. *Plant, Cell & Environment* **35**:1787–1798 (2012).

[25] Roy S *et al.* Arboretum: reconstruction and analysis of the evolutionary history of condition-specific transcriptional modules. *Genome Res* **23**:1039–1050 (2013).

[26] Thompson DA *et al.* Evolutionary principles of modular gene regulation in yeasts. *Elife* **2** (2013).

[27] Weirauch MT and Hughes TR. Conserved expression without conserved regulatory sequence: the more things change, the more they stay the same. *Trends Genet* **26**:66–74 (2010).

[28] Davidson EH and Erwin DH. Gene regulatory networks and the evolution of animal body plans. *Science* **311**:796–800 (2006).

[29] MacArthur S *et al.* 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 (2009).

[30] Bradley RK *et al.* Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related Drosophila species. *PLoS Biol* **8**:e1000343 (2010).

[31] He Q *et al.* High conservation of transcription factor binding and evidence for combinatorial regulation across six drosophila species. *Nat Genet* **43**:414–420 (2011).

[32] Paris M *et al.* Extensive divergence of transcription factor binding in *Drosophila* embryos with highly conserved gene expression. *PLoS Genet* **9**:e1003748 (2013).

[33] Faith JJ *et al.* Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. *PLoS Biol* **5**:e8 (2007).

[34] Stolovitzky G, Prill RJ and Califano A. Lessons from the DREAM2 challenges. *Ann New York Acad Sciences* **1158**:159–195 (2009).

[35] Roy S *et al.* Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science* **330**:1787–1797 (2010).

[36] Marbach D *et al.* Predictive regulatory models in drosophila melanogaster by integrative inference of transcriptional networks. *Genome Res* **22**:1334–1349 (2012).

[37] Mensch J *et al.* Positive selection in nucleoporins challenges constraints on early expressed genes in drosophila development. *Genome Biol Evol* **5**:2231–2241 (2013).

[38] Martens C, Vandepoele K and Van de Peer Y. Whole-genome analysis reveals molecular innovations and evolutionary transitions in chromalveolate species. *PNAS* **105**:3427–3432 (2008).

[39] Joshi A and Göttgens B. Maximum parsimony analysis of gene expression profiles permits the reconstruction of developmental cell lineage trees. *Devel Biol* **353**:440–447 (2011).

[40] Stark A *et al.* Discovery of functional elements in 12 drosophila genomes using evolutionary signatures. *Nature* **450**:219–232 (2007).

[41] David JR *et al.* The historical discovery of the nine species in the drosophila melanogaster species subgroup. *Genetics* **177**:1969–1973 (2007).

[42] Felsenstein J. Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Meth Enzymology* **266**:418–427 (1996).

[43] Huang DW *et al.* Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**:44–57 (2008).

Figures

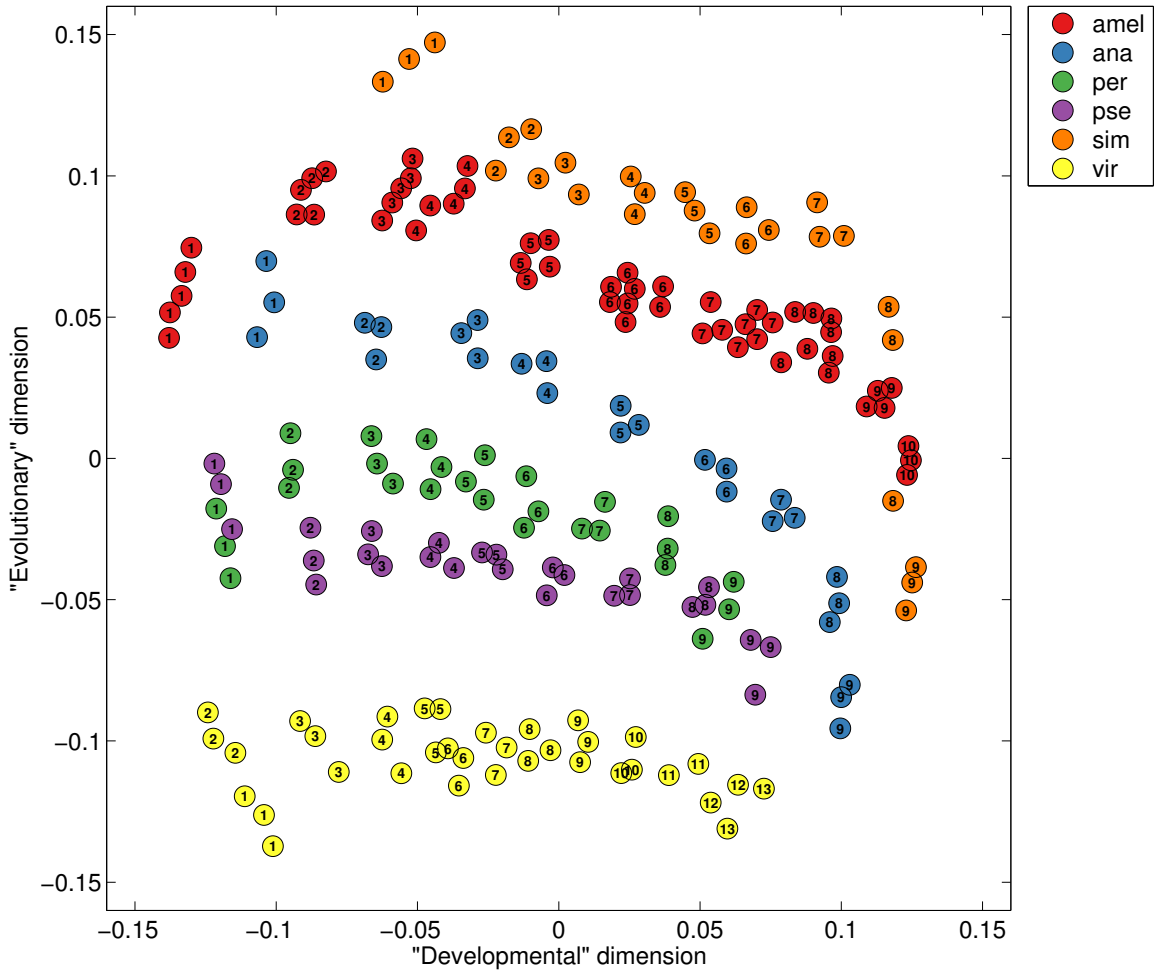

Figure 1: Two-dimensional scaling plot of the gene expression data using Sammon's non-linear mapping criterion. Each dot represents one sample (200 samples total) positioned such that the two-dimensional distances reflect the Euclidean distances between the 3610-dimensional data vectors. Samples are colored by species and the number in each dot is the developmental time point of the sample.

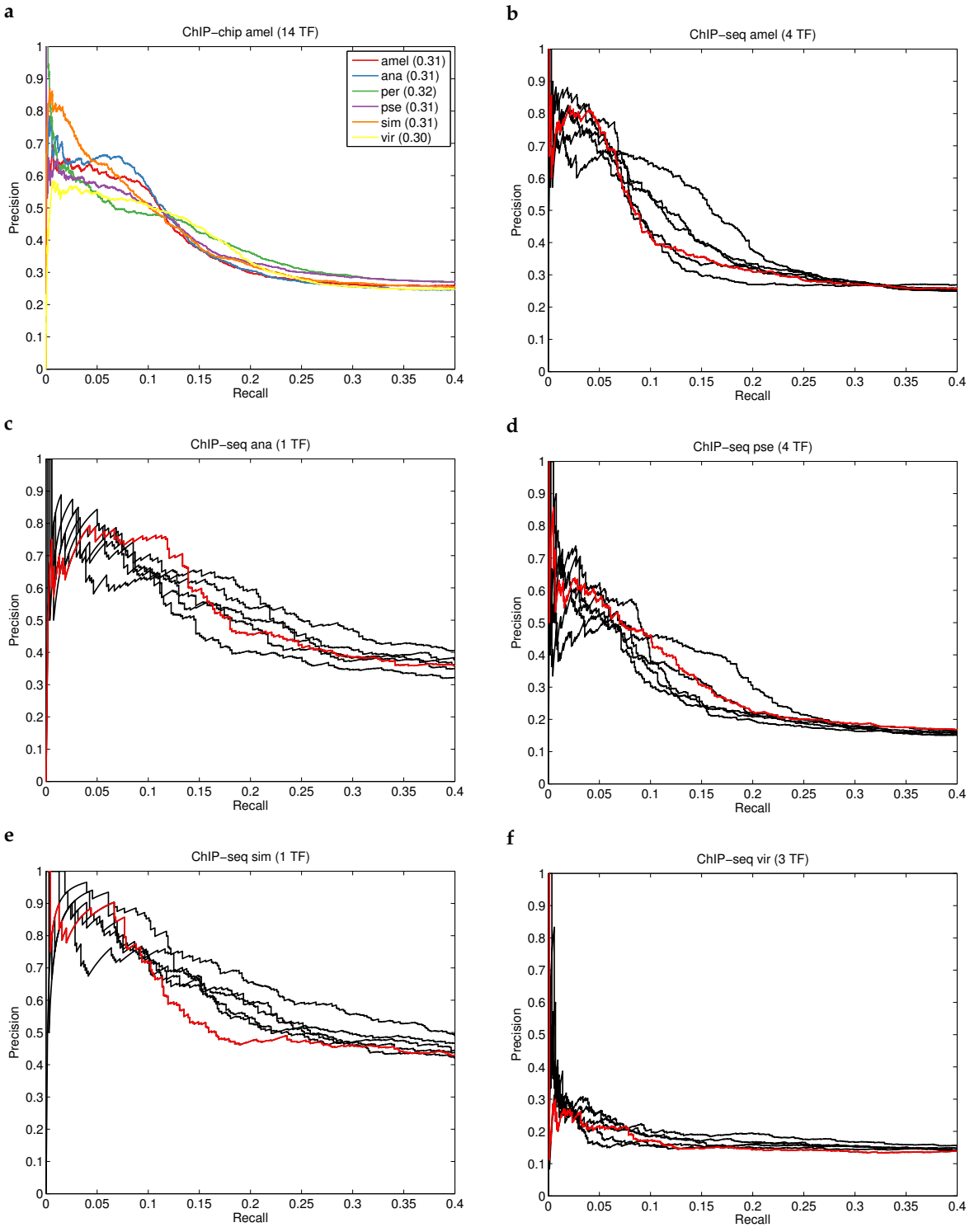

Figure 2: Recall vs. precision curves for predicted regulatory networks in six *Drosophila* species. The gold standard networks were the ChIP-chip network for 14 TFs in *D. melanogaster* (**a**) and the ChIP-seq networks for *D. melanogaster* (**b**, 4 TFs), *D. ananassae* (**c**, 1 TF), *D. pseudoobscura* (**d**, 4 TFs), *D. simulans* (**e**, 1 TF) and *D. virilis* (**f**, 4 TFs). In panel **a**, the numbers in the legend are the area under the curve for each species. In panel **b–f**, the curve for the reference species is in red while the other species are in black.

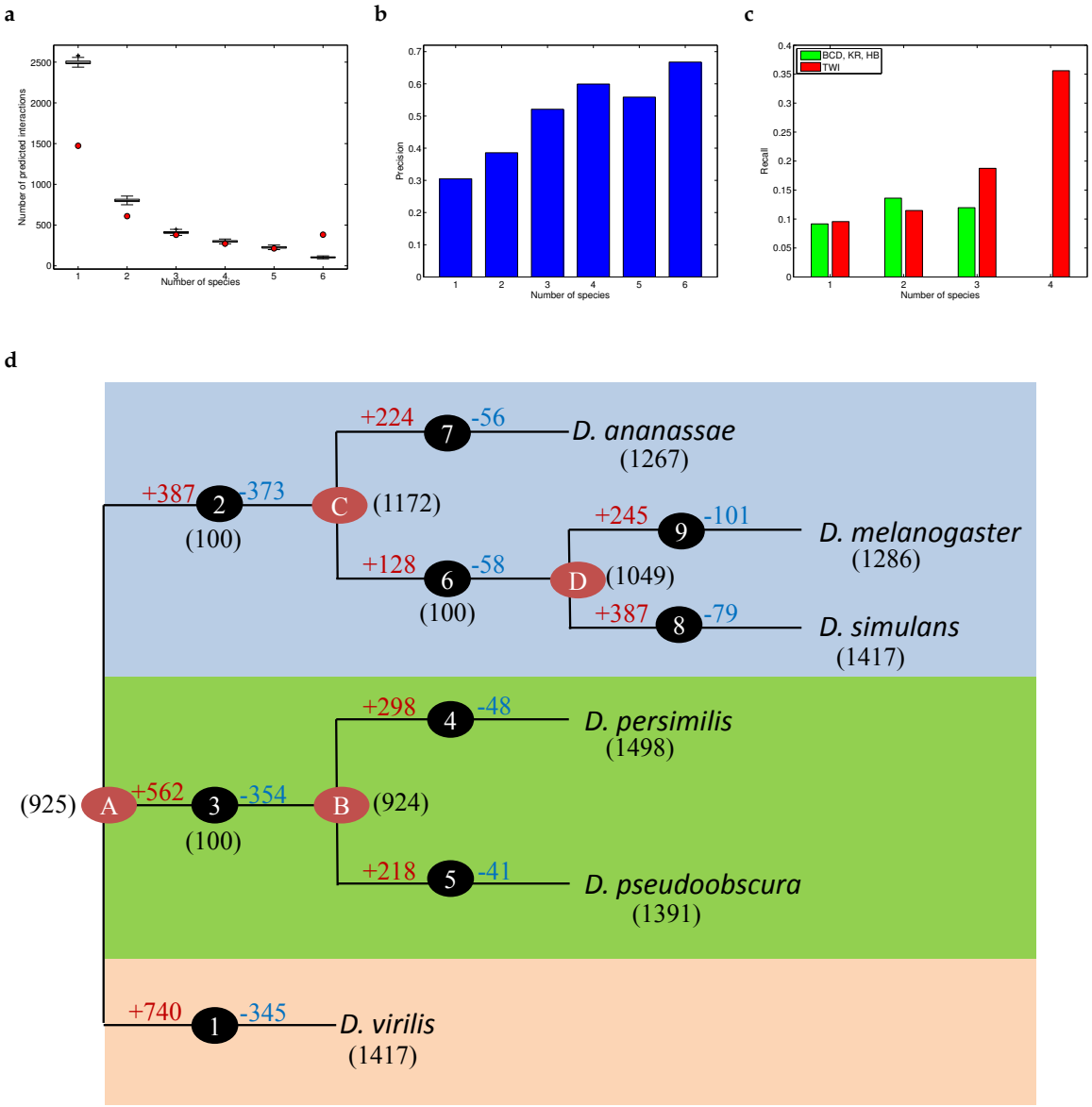

Figure 3: **a.** Number of interactions found in one up to six species in the inferred gene regulatory networks at 10% recall level (red dots) and in 100 randomized networks with the same in- and out-degree distribution as the inferred networks (boxplots). **b.** Precision (percentage of true positives) of interactions found in one up to six species in the inferred networks at 10% recall level. **c.** Recall (percentage of known interactions inferred) of ChIP-seq gold standard interactions conserved in one to three species (green; data for BCD, KR, HB) and one to four species (red; data for TWI) in the inferred networks at 10% recall level. **d.** Phylogenetic tree between six *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 3 groups – *melanogaster* (top), *obscura* (middle), *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.

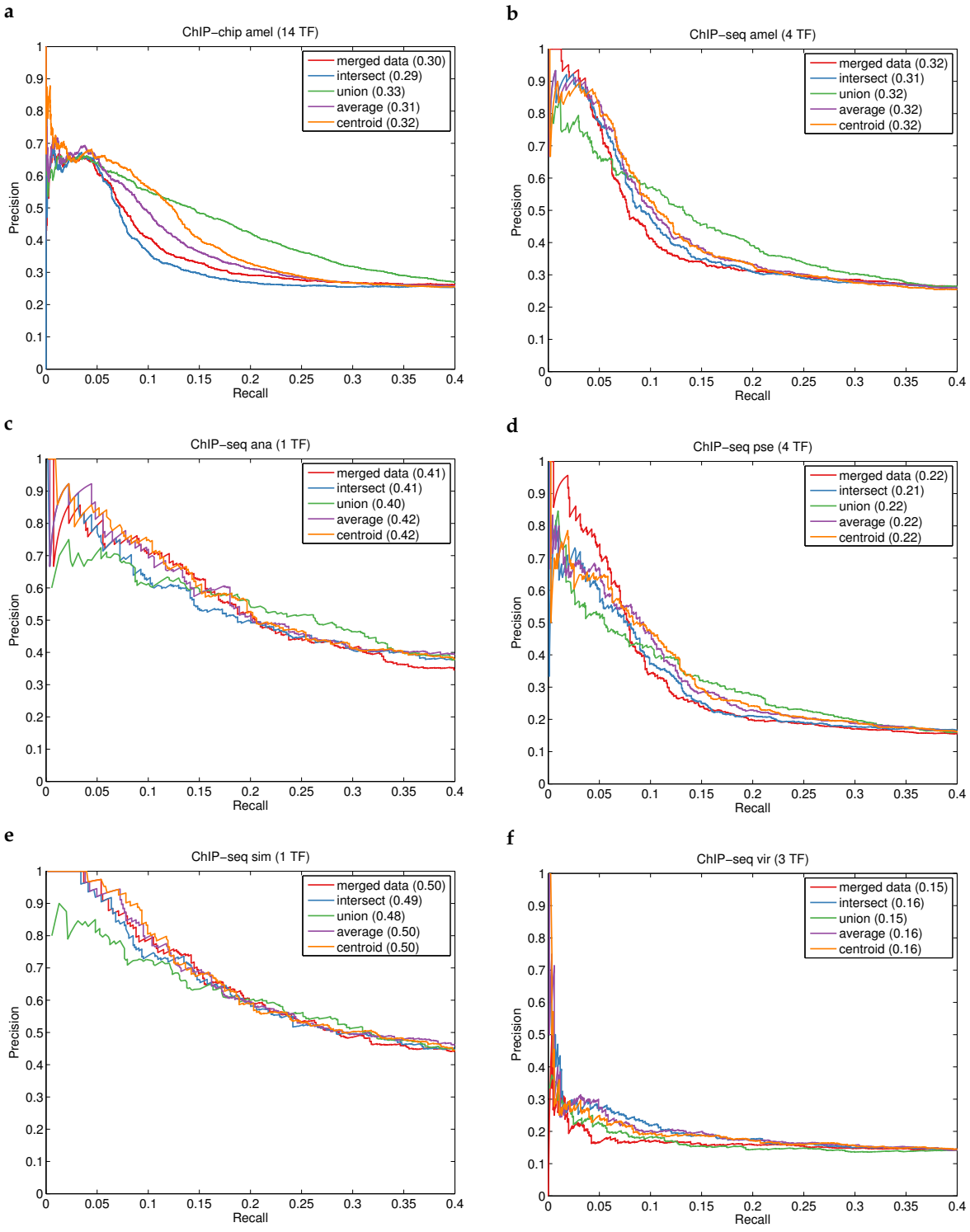

Figure 4: Recall vs. precision curves for predicted regulatory networks for five multi-species meta-analysis methods. The gold standard networks were the ChIP-chip network for 14 TFs in *D. melanogaster* (**a**) and the ChIP-seq networks for *D. melanogaster* (**b**, 4 TFs), *D. ananassae* (**c**, 1 TF), *D. pseudoobscura* (**d**, 4 TFs), *D. simulans* (**e**, 1 TF) and *D. virilis* (**f**, 4 TFs). The numbers in each legend are the area under the curve for each method.

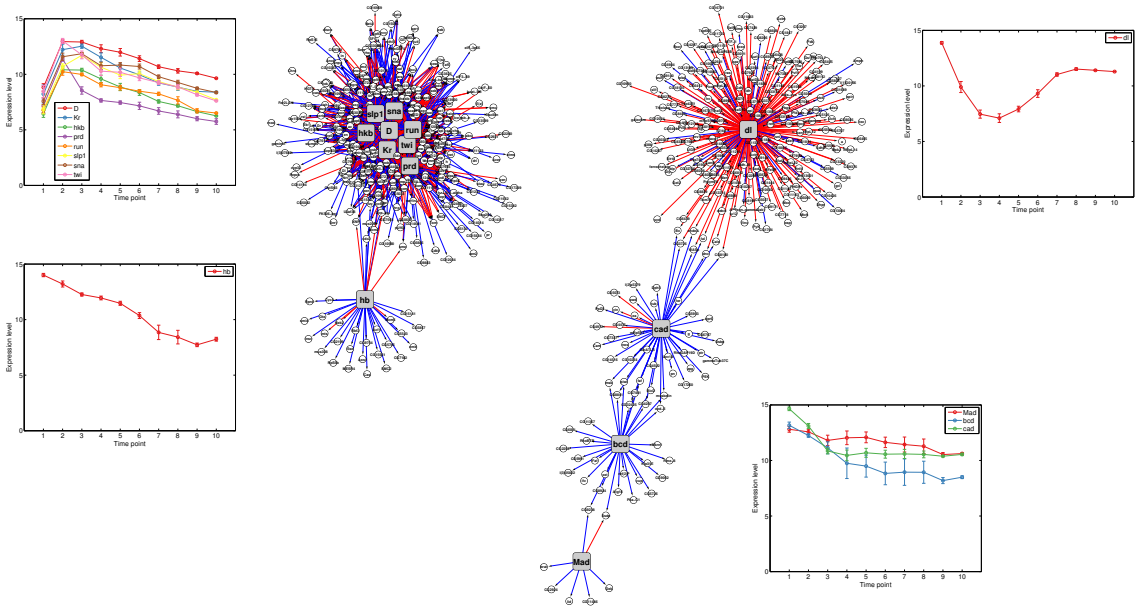

Figure 5: Centroid regulatory network inferred by combining predicted networks across six species. Red edges indicate interactions confirmed by ChIP-chip data in *D. melanogaster*. The insets show the temporal expression profiles in *D. melanogaster* for the regulator(s) in the corresponding subnetworks, clockwise from top left: D, KR, HKB, PRD, RUN, SLP1, TWI; DL; MAD, BCD, CAD; HB.

Tables

TF	ChIP	Amel		Ana		Per		Pse		Sim		Vir	
D	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	

Table 1: 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.

	ChIP-chip amel (14)		**ChIP-seq amel (5)**		**ChIP-seq ana (1)**		**ChIP-seq pse (5)**		**ChIP-seq sim (1)**		**ChIP-seq vir (4)**		**Average**	
Amel	26.10	52.20	20.01	18.02	20.55	13.24	13.42	16.85	10.76	11.43	2.53	0.75	15.56	18.75
Ana	26.79	**53.61**	18.82	17.50	20.77	**16.54**	14.56	20.12	10.86	**15.04**	2.13	0.68	15.66	20.58
Per	30.65	38.90	21.35	29.52	21.21	13.50	14.37	18.49	11.44	11.43	3.78	2.43	17.13	19.04
Pse	29.40	45.10	23.45	30.40	20.40	14.92	16.35	**25.41**	10.64	11.65	**5.58**	**3.54**	**17.64**	21.84
Sim	**28.93**	43.51	20.05	19.20	17.72	13.24	12.21	13.71	9.55	11.00	4.30	2.31	15.46	17.16
Vir	22.37	43.51	**25.87**	**39.69**	**24.23**	12.48	**17.64**	24.83	**12.82**	11.65	2.12	1.67	17.51	**22.31**
Merged data	23.05	26.92	22.32	17.16	22.16	15.23	16.75	15.93	12.27	13.09	2.53	1.38	16.51	14.95
Intersect	18.64	19.62	22.12	23.40	22.39	12.48	15.05	18.30	12.09	11.65	**4.90**	**4.31**	15.86	14.96
Union	**36.64**	51.19	**24.96**	**32.37**	21.81	12.01	16.08	22.60	11.09	11.21	2.55	2.23	18.86	21.93
Average	26.97	40.38	24.00	26.39	23.99	15.55	17.10	26.01	12.66	13.35	4.81	3.29	18.25	20.83
Centroid	30.86	**53.21**	23.98	28.39	**24.22**	**16.54**	**17.38**	**26.47**	**12.78**	**13.62**	4.44	2.90	**18.94**	**23.52**

Table 2: Area under the curve (1st column) and precision at 10% recall (2nd column) Z-scores for single-species and multi-species predicted networks with respect to six gold standard networks and their averages across these six networks. The number in brackets is the number of TFs in the gold standard network and bold values indicate the highest value in each column for both the single-species and aggregated methods.

A Supplementary tables

Transition	Functional category	*P*-value
A → B loss	post-embyonic organ development	5.8×10^{-5}
	regulation of transcription	1.2×10^{-4}
A → B gain	cell fate commitment	1.0×10^{-9}
	regulation of transcription	4.4×10^{-8}
A → C loss	cell–cell adhesion	2.0×10^{-4}
	exocrine system development	5.9×10^{-4}
A → C gain	cell fate commitment	1.1×10^{-12}
	regulation of transcription	2.2×10^{-7}
A → vir loss	regulation of transcription	2.1×10^{-5}
	ectoderm development	4.2×10^{-5}
A → vir gain	neuron differentiation	1.2×10^{-6}
B → per loss	positive regulation of apoptosis	6.6×10^{-3}
B → per gain	translation factor activity	2.3×10^{-5}
	regulation of transcription	2.4×10^{-4}
B → pse loss	sensory organ development	4.6×10^{-3}
	transcription factor activity	8.2×10^{-3}
B → pse gain	intracellular organelle lumen	1.3×10^{-3}
C → ana loss	appendage development	4.0×10^{-6}
C → ana gain	regulation of transcription	7.2×10^{-6}
C → D loss	gastrulation	9.7×10^{-7}
C → D gain	mitochondrion	9.5×10^{-5}
D → amel loss	positive regulation of apoptosis	1.1×10^{-2}
D → amel gain	tissue morphogenesis	7.2×10^{-3}
D → sim loss	rRNA processing	1.2×10^{-4}
	response to organic substances	3.4×10^{-2}
D → sim gain	regulation of transcription	4.5×10^{-6}

Table S1: Functional enrichment for the gene sets gaining or losing interactions at each transition state in the phylogenetic tree in Figure 3d.

TF	Transition	Functional category	*P*-value
BCD	A → B loss		
BCD	A → vir loss		
BCD, HKB	C → ana gain		
BCD, MAD	D → amel gain		
DL	B → per loss	oxidation reduction	4.9×10^{-2}
DL	C → D gain	mitochondrion	1.2×10^{-8}
MAD	B → pse gain		
SLP1	C → ana loss	wing disc development	3.5×10^{-5}
		appendages development	1.8×10^{-5}
		leg disc pattern formation	2.8×10^{-4}
TWI	B → pse loss		
TWI	C → D loss	gastrulation	5.8×10^{-5}
		gland development	6.5×10^{-4}
		tube development	5.2×10^{-3}

Table S2: Transcription factors significantly enriched ($P < 0.05$) for targets in gene sets gaining or losing interactions at transition states in the phylogenetic tree in Figure 3d and the functional enrichment of these target sets.

Species	Time points	Series	Samples	Completeness
Amel	10	8	56	0.7
Ana	9	3	27	1
Per	9	3	27	1
Pse	9	3	27	1
Sim	9	3	27	1
Vir	13	3	39	0.92

Table S3: Expression data summary, listing for each species the number of time points, the number of replicate series, the total number of samples, and the completeness of the data (number of samples divided by number of time points times number of series).