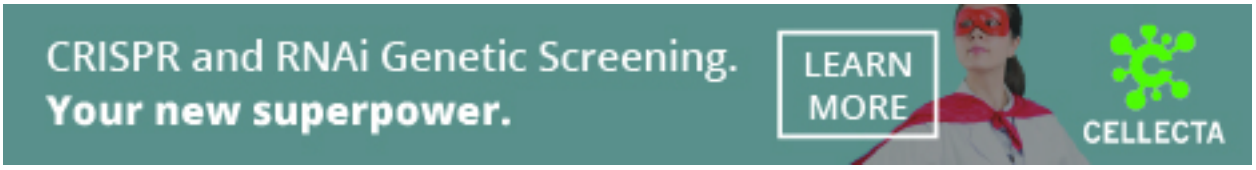
Comparative motif discovery combined with comparative transcriptomics yields accurate targetome and enhancer predictions

Marina Naval-Sánchez, Delphine Potier, Lotte Haagen, et al.

Genome Res. 2013 23: 74-88 originally published online October 15, 2012
Access the most recent version at doi:10.1101/gr.140426.112



Comparative motif discovery combined with comparative transcriptomics yields accurate targetome and enhancer predictions

Marina Naval-Sánchez,[1] Delphine Potier,[1] Lotte Haagen,[1] Máximo Sánchez,[2] Sebastian Munck,[3] Bram Van de Sande,[1] Fernando Casares,[2] Valerie Christiaens,[1] and Stein Aerts[1,4]

[1] *Laboratory of Computational Biology, Department of Human Genetics, University of Leuven, 3000 Leuven, Belgium;* [2] *Centro Andaluz de Biología del Desarrollo (CABD) CSIC-UPO-Junta de Andalucía, 41013 Sevilla, Spain;* [3] *LiMoNe, VIB Center for the Biology of Disease, 3000 Leuven, Belgium*

The identification of transcription factor binding sites, enhancers, and transcriptional target genes often relies on the integration of gene expression profiling and computational *cis*-regulatory sequence analysis. Methods for the prediction of *cis*-regulatory elements can take advantage of comparative genomics to increase signal-to-noise levels. However, gene expression data are usually derived from only one species. Here we investigate tissue-specific cross-species gene expression profiling by high-throughput sequencing, combined with cross-species motif discovery. First, we compared different methods for expression level quantification and cross-species integration using Tag-seq data. Using the optimal pipeline, we derived a set of genes with conserved expression during retinal determination across *Drosophila melanogaster*, *Drosophila yakuba*, and *Drosophila virilis*. These genes are enriched for binding sites of eye-related transcription factors including the zinc-finger Glass, a master regulator of photoreceptor differentiation. Validation of predicted Glass targets using RNA-seq in homozygous *glass* mutants confirms that the majority of our predictions are expressed downstream from Glass. Finally, we tested nine candidate enhancers by in vivo reporter assays and found eight of them to drive GFP in the eye disc, of which seven colocalize with the Glass protein, namely, *scrt*, *chp*, *dpr10*, *CG6329*, *retn*, *Lim3*, and *dmrt99B*. In conclusion, we show for the first time the combined use of cross-species expression profiling with cross-species motif discovery as a method to define a core developmental program, and we augment the candidate Glass targetome from a single known target gene, *lozenge*, to at least 62 conserved transcriptional targets.

[Supplemental material is available for this article.]

Developmental programs depend on complex transcriptional regulation to accomplish the correct and timely expression changes of thousands of genes during the course of patterning, cell specification, and differentiation. The genomic code that implements this intricate regulatory control is to a large extent contained within *cis*-regulatory modules (CRM), harboring binding sites for specific transcription factors (TF) or TF combinations. The annotation and characterization of CRMs is a key challenge in genome biology because a better understanding of *cis*-regulation can deliver mechanistic insight into developmental, evolutionary, and disease processes. For example, the characterization of the "even-skipped" stripe II enhancer, with binding sites of KR, GT, HB, and BCD, has revealed how a striped pattern of gene expression in the *Drosophila* embryo emerges from the combination of TF concentrations, on the one hand, and the genome sequence, on the other hand (Small et al. 1993; Davidson 2001; Carroll et al. 2009). A better knowledge of CRMs can also contribute to the understanding of disease processes, for example, by providing an interpretation of polymorphisms and mutations in the noncoding genome found to be whole-genome sequencing or GWAS studies (Worsley-Hunt et al. 2011). Finally, CRMs can provide insight into evolutionary processes because they account for a large fraction of morphological divergence in the animal kingdom (Wray 2007; Wittkopp and Kalay 2012).

Computational methods are indispensable in the quest for CRMs in the genome and are often used in combination with high-throughput experiments (for a recent review, see Aerts 2012). For example, ChIP-seq against a TF yields whole-genome binding locations for the TF, which can be further classified as directly bound regions versus indirectly bound regions using motif discovery (Gordân et al. 2009). A second example is the application of CRM prediction methods on whole-genome chromatin accessibility data, such as those obtained by DNase I-seq (Sabo et al. 2006). DNase I-seq or FAIRE-seq yield "open regions" that are strongly enriched for functional enhancers. By combining such data with motif discovery or CRM prediction, direct TF target CRMs can be identified (Won et al. 2010; Pique-Regi et al. 2011; Song et al. 2011). A third kind of strategy involves motif discovery or CRM prediction methods on sets of coexpressed genes, to identify shared motifs and CRMs and to predict the upstream regulators (Aerts et al. 2003; Frith et al. 2004; Ho Sui et al. 2007; Roider et al. 2009; McLeay and Bailey 2010). We and others have proposed ways to increase the performance of motif discovery using whole-genome CRM predictions across species, combined with GSEA-like enrichment analysis (Van Loo et al. 2008; Warner et al. 2008; Aerts et al. 2010; Potier et al. 2012). These extensions make the methods more complex but allow analyzing much larger sequence search

[4]Corresponding author
E-mail stein.aerts@med.kuleuven.be
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.140426.112.

23:74–88 © 2013, Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/13; www.genome.org

spaces around each gene in the genome, up to tens of kilobases upstream of and downstream from a gene's transcription start site. Nevertheless, the result of motif discovery depends largely on the input set of coexpressed genes, which can be noisy and usually contains direct and indirect target genes of many different TFs.

In this study, we show a new strategy for motif discovery on coexpressed gene sets, by determining sets of coexpressed genes across multiple species, to focus on the conserved core of a biological process and, consequently, to improve the accuracy of motif and CRM discovery. We apply this multispecies approach to three *Drosophila* species, namely, *Drosophila melanogaster*, *Drosophila yakuba*, and *Drosophila virilis*, with the aim to gain further knowledge of the transcriptional program underlying eye development, and to identify new eye enhancers. The master regulators of *Drosophila* eye development frame the conserved retinal determination gene network (RDGN), with five highly conserved TFs, namely, *eyeless* (*ey*; PAX6 homolog), *twin of eyeless* (*toy*; PAX6 homolog), *dachsund* (*dac*; DACH1-2 homolog), *sine oculis* (*so*; SIX1/2 homolog), and *eyes absent* (*eya*; EYA1-4 homolog) (Silver and Rebay 2005; Amore and Casares 2010; Kumar 2010). Downstream from the RDGN, cells become specified and further differentiate either as one of the eight types of photoreceptors (R1–R8) or as accessory cells. This system has been extensively used in forward genetic screens and has played an important role in deciphering signaling pathways including the Notch, EGFR, Smoothened, Dpp, Wnt, and Hippo signaling cascades. Although a considerable number of genes and genetic interactions are already known in retinal determination (e.g., nearly 500 genes are annotated with the GO term "eye development"), little is known about the regulatory interactions among these genes. Indeed, only a handful of Eyeless target genes are known (Ostrin et al. 2006), around 20 Atonal target genes have recently been described (Aerts et al. 2010), and for other TFs, few anecdotal targets are known (e.g., Pauli et al. 2005; Rogers et al. 2005; Jemc and Rebay 2007).

We generated whole-genome expression data in eye and wing imaginal discs in three *Drosophila* species by Tag-sequencing and derived conserved eye-enriched gene sets. By applying advanced motif discovery methods, including CRM conservation cues, we identify enriched motifs in these gene sets for multiple eye-related TFs, such as Glass, SoxNeuro, Scratch, Eyeless, and Suppressor of Hairless. We then validated the predicted conserved Glass targets in *D. melanogaster* using RNA-seq in mutant eye discs and put forward a set of 62 genes that are activated by Glass. Enhancer validations by in vivo reporter assays, both using cloned enhancers and using the Janelia Farm GAL4 lines (Pfeiffer et al. 2008), achieved a success rate of 77%, with seven out of nine tested enhancers showing GFP expression in the eye disc in third instar larvae and colocalizing with Glass expression. We conclude that cross-species expression profiling, combined with robust regulatory sequence analysis, provides a straightforward strategy for enhancer discovery and gene regulatory network mapping and is generally applicable to probe homologous developmental programs across species.

Results

Tag sequencing in two tissues and three species

Tag-sequencing libraries were generated for *D. melanogaster*, *D. yakuba*, and *D. virilis* eye-antennal and wing imaginal discs, yielding between 2.4 and 6.8 million expressed sequence tags (EST) of 21 bp (Supplemental Table S1). Gene expression levels were obtained by a custom data processing pipeline, involving low-count filtering and normalization (see Methods; Supplemental Figs. S1–S5). For *D. melanogaster*, 89% of the uniquely mapped reads could be assigned to an annotated gene, and expression levels were obtained for 6021 and 5825 genes in eye-antennal and wing imaginal discs, respectively. The log ratio of the normalized eye-versus-wing levels was validated by qRT-PCR for eight genes (Fig. 1A) and correlates well with previously obtained microarray data in eye and wing imaginal discs (Supplemental Fig. S6; Ostrin et al. 2006; Aerts et al. 2010). By ranking all genes according to log(eye/wing) values, we found this ranking to contain a highly significant "leading edge" for eye-related Gene Ontology functions and for eye-related gene expression (using FlyBase TermLink gene sets), providing additional confirmation for accurate gene expression measurements obtained by Tag-seq (Table 1; Supplemental Fig. S7; Supplemental Table S2). Similar results can be obtained by ranking all genes according to *P*-values for differential expression calculated by DESeq (Anders and Huber 2010), edgeR (Robinson et al. 2010), or NOISeq (Tarazona et al. 2011), although on this specific data set the log-ratio ranking is slightly more robust (Supplemental Fig. S6).

Next, we turned to the mapping results for *D. yakuba* and *D. virilis*. We expected the quality of the genome annotation for these species to be not as high as for *D. melanogaster*. Indeed, with a comparable fraction of reads aligning to the genome, now the percentage of reads that could be assigned to a gene is much lower (41% and 50%, respectively) (Fig. 1B,C, blue bars; see Supplemental Fig. S8 for an example). To solve this issue, we implemented two different procedures (Fig. 1D,E). In the first approach, we assigned the remaining free peaks to a gene when they are located downstream from an annotated transcript. This way, we were able to assign an expression value to an additional 2887 genes in *D. yakuba* and 3016 in *D. virilis* compared with the baseline, using a maximum distance to the annotated 3′ end of 5 kb (Supplemental Fig. S9). In the second approach, we used pairwise whole-genome alignments and assigned peaks in *D. yakuba* or *D. virilis* to an annotated *D. melanogaster* gene. This approach allows us to assign expression values to *D. melanogaster* genes for an additional 2572 and 1054 genes for *D. yakuba* and *D. virilis*, respectively, compared with the baseline (see Supplemental Fig. S10 for a comparison between both approaches). Note that for genes where both approaches assign a different peak and thus a different expression value to the same gene, we selected the value that is most similar to the orthologous *D. melanogaster* gene, working under the conservative assumption that the majority of genes are conserved in gene expression rather than divergent (Supplemental Fig. S10; Supplemental Table S3).

Identification of tissue-specific genes across species: Species as replicates versus rank aggregation

Having obtained normalized expression levels for as many genes as possible in each species individually, the next step of our analysis was to use the expression data in *D. yakuba*, *D. virilis*, and *D. melanogaster* to identify a core set of eye developmental genes in *Drosophila*. To this end, we propose a rank aggregation method based on order statistics (OS) (Aerts et al. 2006), integrating the three species-specific rankings into one "conserved" eye-versus-wing ranking. As species-specific rankings, we used the log-ratio-based ranking, although rankings generated above by DESeq, edgeR, or NOISeq can also be used (Supplemental Fig. S11). We compared this approach with the *direct* statistical comparison of

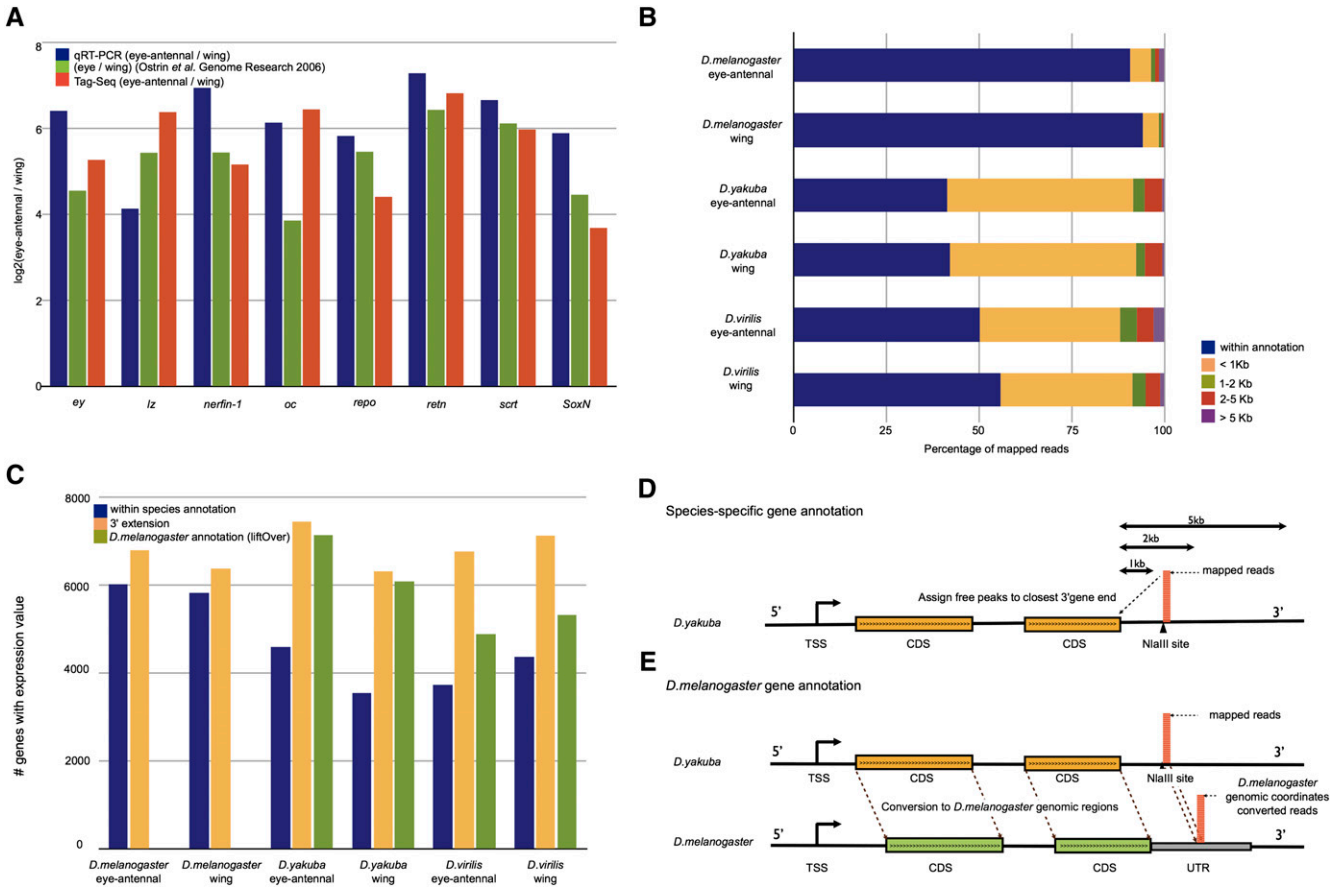


Figure 1. Tag-seq analysis *D. melanogaster* and other species. (*A*) Relative expression measures of eye-antennal imaginal discs versus wing imaginal discs, as fold-changes, of eight genes involved in eye development, namely, *ey*, *lz*, *nerfin-1*, *oc*, *repo*, *retn*, *scrt*, and *SoxN*. (Blue) Measures by qRT-PCR; (green) microarray; (orange) Tag-seq. (*B*) Percentage of mapped reads falling within the currently available species-specific gene annotation (blue), compared with reads falling 1 kb (yellow), 2 kb (green), 5 kb (red), or >5 kb (purple) downstream from an annotated gene. (*C*) Number of genes with more than 10 mapped reads in any of the two tissues, using different annotation procedures. (*D*,*E*) Overview of the two different methods to obtain gene expression levels in the other species, one using species-specific annotation with 3′ extension (*D*), the other by exploiting orthologous positions in *D. melanogaster* with *D. melanogaster* gene annotations (*E*).

eye and wing samples, using the three species as replicates, with edgeR (Robinson et al. 2010), DESeq (Anders and Huber 2010), or NOISeq (Tarazona et al. 2011), and found that the OS approach outperforms these other methods (Fig. 2A; Supplemental Fig. S12). Importantly, regardless of the method used, the integration of three species greatly improves the accuracy of detecting eye-specific genes, compared with *D. melanogaster* only. For example, the five core RDGN genes (*ey*, *toy*, *dac*, *eya*, and *so*) are all ranked within the top 170 genes in the OS cross-species ranking, compared with the top approximately 500 in each species individually (Fig. 2B). To prevent applying an arbitrary threshold on the OS-based ranking, we determined the optimal threshold using Gene Ontology enrichment with *GOrilla* (Eden et al. 2009) and compared the GO enrichment in the cross-species ranking with the individual rankings for each species (Table 1). For all GO terms related to eye, photoreceptor, or neuronal development, the enrichment in the cross-species rankings is higher than the enrichment in any of the individual species. Hence, the combination of three species to identify genes involved in a conserved process confers gene selection robustness. The best enrichment of eye-antennal related terms is for the term "compound eye photoreceptor cell development" (GO:0042051; *P*-value is 3.6×10^{-12}). This enrichment is found at the optimal threshold of 245 genes, and we further use this set of 245 genes as the conserved eye-enriched gene set (see Supplemental Fig. S13 for a heatmap with expression values). Interestingly, within these 245 genes, we observe a high percentage of "unknown genes." More precisely, 99 of the 245 genes (40.4%) are annotated only with a "CG number" and are largely uncharacterized. For these genes, we can now assign a role in eye development because they are highly enriched in the eye, across three *Drosophila* species. Another interesting finding is that a large proportion of these 245 genes, namely, 18.77% (28.08% of the known genes) are TFs, as annotated by flyTF (Supplemental Table S3; Pfreundt et al. 2010). We conclude that cross-species Tag-seq, integrated with order statistics, identifies conserved tissue-specific gene expression and thereby enables the association of many unknown genes to the core eye developmental program in *Drosophila*.

Motif discovery on conserved eye-specific genes

To identify novel regulatory interactions underlying eye photoreceptor development, we analyzed the set of 245 conserved eye-specific genes across species for shared motifs in their regulatory

Table 1. Gene Ontology enrichment on eye-versus-wing gene expression rankings

		Single species ranking			Cross-species ranking			
GO ID	GO Term	*D. melanogaster*	*D. yakuba*	*D. virilis*	edgeR	DESeq	NOISeq	OS
GO:0042051	Compound eye photoreceptor development	9.66×10^{-08} (596, 17)	7.06×10^{-8} (142, 10)	9.00×10^{-07} (559, 15)	1.87×10^{-07} (418, 14)	1.11×10^{-07} (52, 7)	2.50×10^{-10} (93, 10)	3.60×10^{-12} (245, 15)
GO:0042462	Eye photoreceptor cell development	2.76×10^{-07} (596, 17)	1.23×10^{-7} (142, 10)	2.37×10^{-06} (559, 15)	4.35×10^{-07} (418, 14)	1.70×10^{-07} (52, 7)	4.47×10^{-10} (93, 10)	9.12×10^{-12} (245, 15)
GO:0042461	Photoreceptor cell development	7.19×10^{-07} (596, 17)	1.96×10^{-7} (142, 10)	5.36×10^{-06} (559, 15)	6.41×10^{-07} (418, 14)	2.24×10^{-07} (52, 7)	7.35×10^{-10} (93, 10)	2.11×10^{-11} (245, 15)
GO:0001654	Eye development	1.48×10^{-04} (523, 22)	1.06×10^{-6} (351, 21)		2.57×10^{-05} (605, 25)	9.20×10^{-07} (25, 7)	2.88×10^{-10} (756, 35)	3.28×10^{-11} (545, 31)
GO:0030182	Neuron differentiation	4.56×10^{-04} (523, 12)	2.7×10^{-7} (499, 17)	1.48×10^{-06} (353, 14)	5.75×10^{-08} (418, 17)	1.55×10^{-06} (237, 12)	2.85×10^{-07} (108, 9)	5.01×10^{-11} (308, 17)
GO:0048749	Compound eye development	5.1×10^{-04} (523, 20)	1.08×10^{-5} (71, 9)	7.08×10^{-05} (9, 4)	6.35×10^{-05} (418, 19)	6.05×10^{-07} (25, 7)	9.83×10^{-10} (756, 33)	1.07×10^{-09} (545, 28)
GO:0048666	Neuron development	3.67×10^{-05} (807, 27)	6.75×10^{-7} (147, 13)	7.89×10^{-05} (339, 16)	1.22×10^{-06} (995, 31)	1.13×10^{-07} (52, 9)	1.49×10^{-09} (902, 34)	1.24×10^{-09} (170, 16)
GO:0001754	Eye photoreceptor cell differentiation		1.05×10^{-5} (405, 12)	2.48×10^{-05} (353, 11)	1.85×10^{-05} (418, 12)	1.33×10^{-05} (202, 9)	4.23×10^{-07} (108, 8)	1.20×10^{-08} (245, 12)
GO:0048663	Neuron fate commitment	1.39×10^{-04} (50, 5)	3.37×10^{-5} (101, 7)	1.16×10^{-05} (131, 8)	1.65×10^{-05} (239, 10)	7.86×10^{-06} (52, 6)	3.91×10^{-07} (93, 8)	1.75×10^{-08} (170, 11)
GO:0046530	Photoreceptor cell differentiation		2.05×10^{-5} (405, 12)	4.71×10^{-05} (353, 11)		2.23×10^{-05} (202, 9)	6.49×10^{-07} (108, 8)	2.41×10^{-08} (245, 12)

P-values are based on *GOrilla* (Eden et al. 2009). The optimal rank threshold and the number of genes annotated to the respective GO term within the threshold are given in parentheses. Shading indicates maximum enrichment.

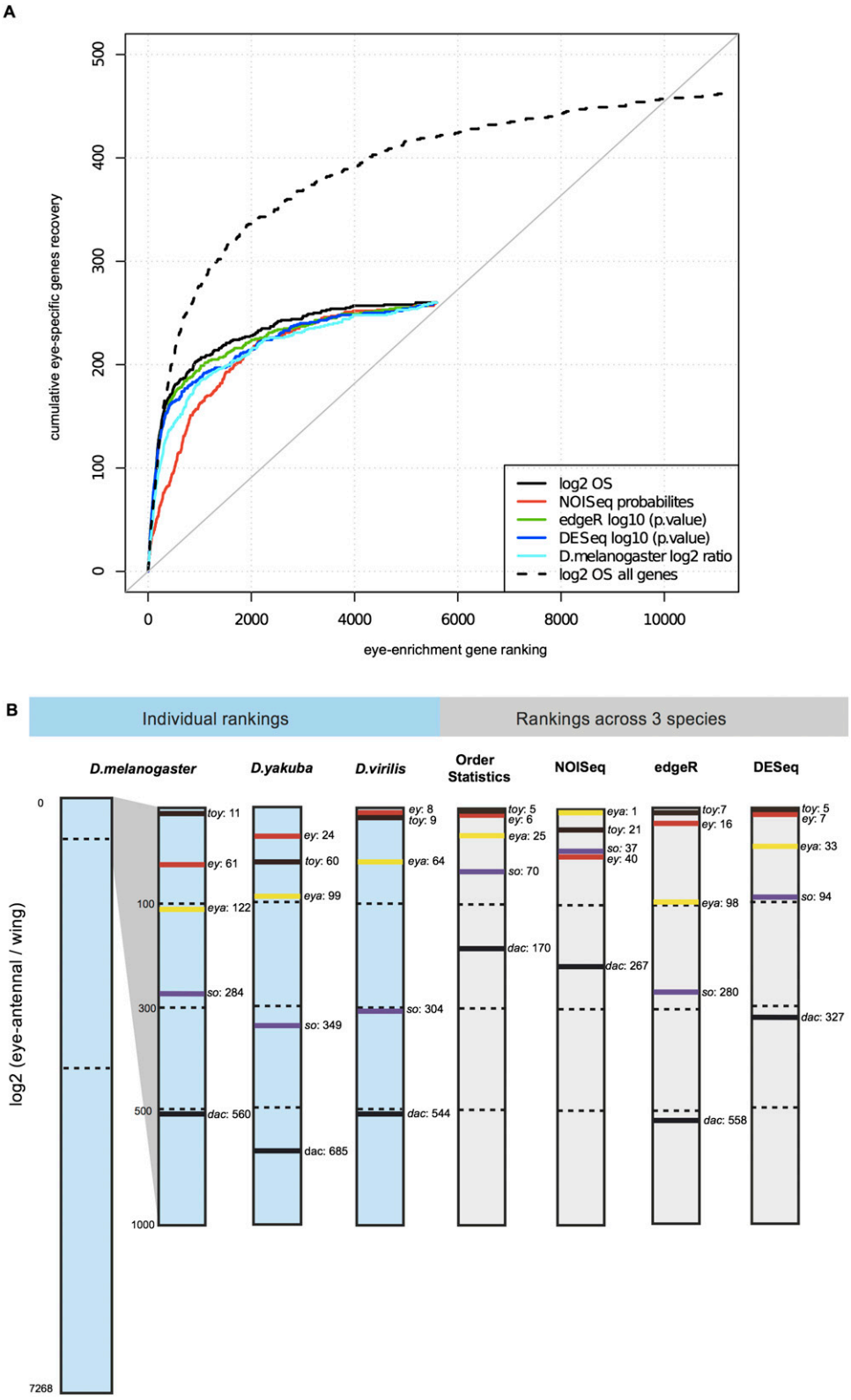


Figure 2. Cross-species analysis more robustly identifies eye-specific genes. (*A*) Comparison of rank aggregation using Order Statistics (OS; black curves) to integrate expression levels across species, with differential expression analysis using the species as replicates (blue, green, red); and with a single-species log-ratio ranking (cyan). (Black dashed curve) The recovery using all genes; (solid curves) using only genes with expression values in the three species (no missing values). The true-positive set is 507 eye-enriched genes from *D. melanogaster* obtained from microarray data (Ostrin et al. 2006). (*B*) Schematic visualization of individual and cross-species eye-versus-wing rankings, indicating the rank position of the RDGN genes *so*, *toy*, *ey*, *eya*, and *dac,* showing an increasing rank for all RDGN members in the cross-species ranking, in particular the OS ranking.

sequences. We used the method cisTargetX (Aerts et al. 2010; Potier et al. 2012), which combines whole-genome scorings of clustered binding sites for a library of position weight matrices (PWMs), across all sequenced *Drosophila* species, with enrichment analysis. Out of 3731 PWMs, our set of genes showed an over-representation of 47 motifs with a normalized enrichment score (NES) >2.5. Due to motif redundancy, the 47 motifs could be clustered in 16 significantly distinct motifs (Mahony and Benos 2007). The highest enriched motif is for Glass (GL; rank = 1, NES = 4.7523), a TF previously known to be involved in photoreceptor differentiation (Moses et al. 1989). Thus far, and to our knowledge, only one target gene is known for Glass at the same developmental time point, namely, *lozenge* (*lz*) (Yan et al. 2003). (A second known Glass target, *ninaE*, is activated during pupal development and is not expressed in the tissue under study [Moses and Rubin 1991].) From the 245 genes as input, cisTargetX predicts 96 direct target genes of Glass, including *lz*, and also predicts *gl* itself as an auto-regulatory target gene (Table 2; Supplemental Table S4).

Among the remaining enriched motifs, several more are related to TFs that play a role in eye development, such as a Sox-related motif for SoxNeuro (SoxN); the motif for SU(H); and the motif for the Zinc Finger TF Scratch (SCRT) (Fig. 3; Zhu et al. 2011). SOXN and SCRT are themselves among the list of 245 conserved eye-specific genes, which justifies the assignment of these TFs to their candidate motifs. Although the SCRT motif is similar to an E-box motif (CANNTG), such as the one bound by Atonal, the SCRT motif identifies a significantly different set of target genes compared with ATO. Indeed, by comparing the SCRT targetome found here, with ATO target predictions from our earlier work (Aerts et al. 2010), we find 81 specific SCRT targets, 71 specific ATO targets, and 17 genes in common. Therefore, we believe that the SCRT motif is indeed related to SCRT target genes, rather than target genes of a basic helix–loop–helix factor such as Atonal. From the analysis of the top 245 conserved eye-specific genes, we did not find motifs for the RDGN factors, such as Eyeless. However, when we increased the stringency of the cutoff, using, for example, only the top 75 or top 100 of conserved eye-specific genes, we found the Eyeless motif (PAX6 position weight matrix) significantly enriched (Fig. 3). When the motif discovery results on conserved eye-enriched genes are compared with the motif discovery results on single-species gene expression data, only one of these five meaningful motifs could be identified, illustrating the robustness of motif discovery after cross-species integration of tissue-specific gene expression (Supplemental Figs. S14, S15). Putting all the regulator-target interactions together yields a gene regulatory network (Fig. 3B), which shows a high degree of regulatory cross talk, with many genes regulated by more than one of these TFs. The highly interconnected network contains 96 potential Glass targets, 39 SU(H) targets, 17 SOXN targets, 99 SCRT targets, and 18 EY targets (Fig. 3B; Supplemental Table S4).

Validation of predicted Glass target genes by RNA-seq in *D. melanogaster*

To validate the target gene predictions in the above gene regulatory network, we focus on one of the five TFs, for which the targetome was largely unknown before—Glass. For *glass*, homozygous viable mutant lines are available, and *glass* phenotypes have been described as having disrupted ommatidial patterning and a lack of photoreceptors (PR) (Moses et al. 1989; Ellis et al. 1993). To validate our Glass target gene predictions, we performed RNA-seq on *D. melanogaster glass* mutant eye-antennal imaginal discs and wild-type discs (see Methods). We found that the set of 96 predicted Glass targets is significantly enriched at the top of the wild-type-versus-mutant gene ranking, by a Gene Set Enrichment Analysis (GSEA FDR <0.001) (Fig. 4A, inset), regardless of the method used to assess differential expression between wild-type and mutant samples (DESeq, edgeR, NOISeq, or the log ratio) (see Supplemental Fig. S16). This result globally validates our direct target gene predictions based on cisTargetX. To assess the significance further, we compared the fold changes of the 96 predicted Glass targets in mutant versus wild-type discs with several other gene sets and found that the predicted Glass targets are significantly more down-regulated than any control set ($P < 0.005$ by Wilcoxon test) (Fig. 4A). Analysis of differential expression using DESeq identifies a subset of 62 validated direct Glass targets (FDR <0.05), which we define as the "Glass targetome" (boxed network in Fig. 4B). Note that we choose DESeq because of its slightly better GSEA results compared with the other methods (Supplemental Fig. S16).

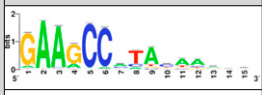
Based on existing knowledge and tools available for some of the predicted targets, we could confirm that using "*down-regulation in the glass mutant*" as a filter is useful to distinguish true-positive from false-positive predictions. Among the set of 34 *in*validated targets, we found several genes that are unlikely to be Glass targets, either because they are known to act upstream of *glass* in the eye gene regulatory network (Amore and Casares 2010; Aerts et al. 2010; Kumar 2010), such as *eyes absent*, *Optix*, and *atonal*, or because they are not expressed in differentiating photoreceptors,

Table 2. Selection of predicted Glass target genes

Gene	*D. melanogaster* gene FBgn	*D. melanogaster* wild type[a]	*D. yakuba* wild type[a]	*D. virilis* wild type[a]	gl[60j][b]	cisTargetX predicted Glass binding region
gl	FBgn0004618	8.34	5.03	8.10	−1.27	chr3R:14199286–14201326
chp	FBgn0000313	3.54	4.51	6.85	−4.44	chr3R:27035441–27035982
dpr10	FBgn0052057	2.49	3.60	1.43	−2.67	chr3L:10166262–10167555
Lim3	FBgn0002023	3.73	2.97	3.67	−2.07	chr2L:19085176–19086688
amon	FBgn0023179	4.89	1.72	4.59	−1.89	chr3R:22530095–22531144
retn	FBgn0004795	6.82	5.18	3.83	−1.81	chr2R:19523270–19524801
lz	FBgn0002576	6.38	3.99	4.50	−1.75	chrX:9180815–9181775
dmrt99B	FBgn0039683	2.71	3.51	6.56	−1.72	chr3R:25514163–25515780
scrt	FBgn0004880	5.97	5.04	7.05	−1.68	chr3L:3981801–3982640
CG6329	FBgn0033872	4.68	0.73	3.78	−1.66	chr2R:9715391–9716094
Nrt	FBgn0004108	2.72	2.38	2.45	−0.46	chr3L:16754263–16755474

[a]Log2(eye/wing) of Tag-seq-derived expression values.
[b]Log2(eye gl[60j]/eye wild type) of RNA-seq-derived expression values.

A

	Glass			SCRT			SU(H)			SOXN			EY		
Motifs															
	flyfactorsurvey-gl_SOLEXA_5_FBgn0004618			flyfactorsurvey-scrt_2.5_FBgn0004880			jaspar4-MA0085.1-Su_H			transfac10.1-M00410-V-SOX9_B1			transfac10.1-M00097-V-PAX6_01		
Input	Motif rank	NES	Number of Target Genes	Motif rank	NES	Number of Target Genes	Motif rank	NES	Number of Target Genes	Motif rank	NES	Number of Target Genes	Motif rank	NES	Number of Target Genes
75	1	5.3455	36	3	4.0350	33	11	3.2393	25	6	3.7252	16	31	2.6491	8
100	1	4.5654	41	2	4.3835	41	8	3.3746	11	13	3.0670	17	74	2.3316	18
245	1	4.7523	96	5	3.5069	82	15	3.1040	38	65	2.3562	17			
545	8	3.8510	177	10	3.4508	172	19	3.0350	100	249	1.5604	(--)			

B

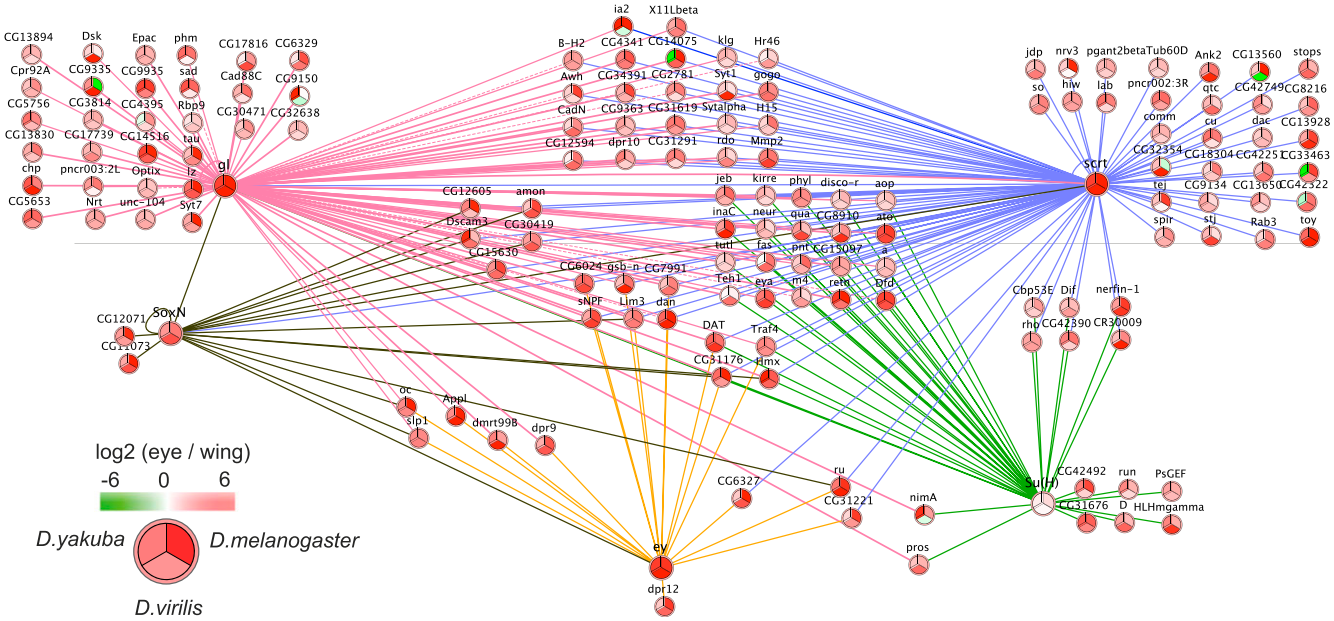


Figure 3. Motif discovery results from cisTargetX. (*A*) Motifs and candidate TFs found by cisTargetX, using different gene set sizes as input (top 75, 100, 245, and 545 genes from the OS-based integrative ranking). (NES) Normalized enrichment score from cisTargetX; (Motif Rank) rank of the motif from a collection of 3731 motifs used by cisTargetX. (*B*) Gene regulatory network showing the predicted regulatory interactions of Glass, SCRT, SU(H), SOXN, and EY, obtained from the cisTargetX analyses on the top100 and the top245 eye-specific genes. (Dashed arrows) Glass target gene predictions that are found to be up-regulated in the *glass* mutant by RNA-seq.

such as *Awh* (Curtiss and Heilig 1997; Roignant et al. 2010) and *Dfd* (Diederich et al. 1991). For several of these genes, we could verify by immunohistochemistry that their expression is indeed independent of Glass and also that Glass is not repressing these genes (see Supplemental Fig. S17).

Next, we compared the 62 validated targets with the 34 invalidated genes to search for characteristic features of true-positive predictions (Fig. 4C). An additional cisTargetX analysis on the 62 validated targets finds the Glass motif more strongly enriched than in the original set of 245 genes, while a cisTargetX analysis of the 34 invalidated genes does not find the Glass motif over-represented. This suggests that there indeed exist differences between both gene sets at the motif level (Supplemental Fig. S18). First, we found that the genomic rank given by cisTargetX of the validated targets is better than the genomic rank of the nonvalidated targets (a median rank of 900 vs. 1602, respectively). Interestingly, using the top 500 as genomic rank cutoff instead of the automatically calculated cutoff at 3327 yields 27 predicted targets, of which 24 are validated in the RNA-seq experiment (Supplemental Fig. S19), representing an increase of the positive predictive value (PPV) from 64.5% (62/97) to 88.9%. Therefore, the cisTargetX ranking can be used as a valid filter. Second, the validated targets are enriched for Glass-only target

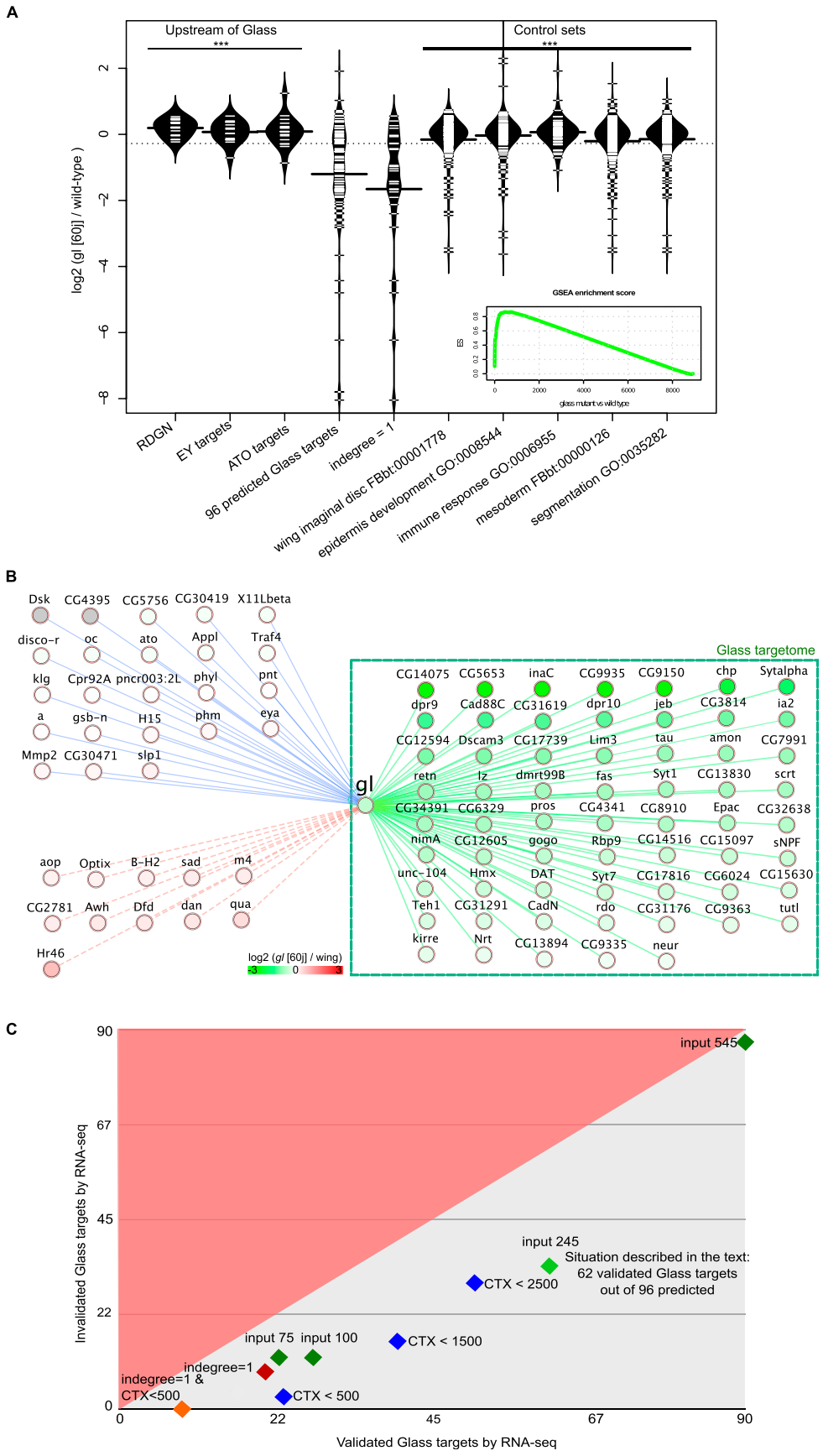


Figure 4. (Legend on next page)

predictions, while the 34 invalidated targets are enriched for targets of multiple TFs (e.g., Glass + Scratch). The Glass-only targets are much more strongly down-regulated than the targets with multiple inputs (see bean plot in Fig. 4A). This means that the node in degree can be used as a post-filter, retaining 30 genes from the initial set of 96 genes, of which now 21 can be validated. Several other features we investigated were not significantly different between these sets. For example, we found no obvious difference between the maximal scoring PWM instances in both sets, indicating that the differences in binding site clustering and conservation at the CRM-level (jointly captured by the cisTargetX ranking) are more indicative for being a true target gene than differences in nucleotide preference in one single binding site (Supplemental Fig. S20). Also, using smaller sizes of input genes (75 or 100 genes) did not result in a lower false-positive rate (although using a larger size, namely, 545 genes, resulted in a higher false-positive rate) (Supplemental Fig. S21). Based on these results, we combined the two best filters (top 500 for cisTargetX and Glass-only targets) into a high-stringency filter. Although this filter retains only eight predicted Glass target genes (including *chp* and the positive control *lozenge*), it corresponds to a PPV of 100%, because all eight targets are validated by RNA-seq (orange diamond in Fig. 4C).

In conclusion, the RNA-seq data validate a significant number of the 96 initially predicted Glass targets and show that Glass mainly acts as an activator (also see Discussion). The rate of true positives can be increased from 64.5% to 100% using additional filters, providing an interesting strategy when additional transcriptomics data in a TF perturbation are not available. In this study, having RNA-seq data available, we continue using the set of 62 validated target genes in the remainder of the text. A selection of these genes is shown in Table 2 (namely, *lozenge*, *glass*, and nine genes for which we test the enhancer in the next section), together with their expression levels across the three species and the change in expression in the *glass* mutant (for all genes, see Supplemental Table S5).

Validation of predicted Glass binding sites by in vivo enhancer-reporter assays

Each of the 62 predicted Glass target genes is based on a high-scoring *cis*-regulatory module (CRM) harboring a cluster of one or more Glass binding sites (Table 2). To test some of these CRM predictions, we made use of the collection of GAL4 lines made available from Janelia Farm (Pfeiffer et al. 2008). By overlapping our CRM predictions with the currently available GAL4 lines in the Bloomington stock center (3029 lines on November 28, 2011), we found nine GAL4 lines that cover a predicted CRM. We crossed these lines to UAS-GFP and assayed GFP expression, in combination with the expression of Glass and Elav (Fig. 5; Supplemental Fig. S22). We found eight of the nine lines to drive GFP in the eye disc, of which seven are active in photoreceptor cells and overlap with Glass and ELAV (Fig. 5). Only the CRM prediction for the a*mon* gene showed no GFP in the eye-antennal disc, and the CRM prediction for *Nrt* showed GFP in the eye disc, but in glial cells rather than in photoreceptor cells (Supplemental Fig. S23).

Some of the GAL4 lines are constructed with relatively large genomic regions. We assessed for three positive CRMs whether the actual CRM prediction recapitulates the GFP expression pattern observed for the encompassing GAL4 line. We chose the CRM located near *chp* because it shows a delay in expression compared with *glass*; and the CRMs near *scrt* and *retn*, because these are interesting TFs themselves with phenotypes manifest in the eye or photoreceptors (FlyBase phenotypes). Transgenic flies carrying the *scrt*, *chp*, and *retn* CRMs, directly linked with GFP, generate the same expression pattern as their corresponding GAL4 lines in the Rubin GAL4 collection (Fig. 5). Finally, we verified for all three CRMs whether the CRM activity is affected in the *glass* mutant, by crossing the enhancer-GFP reporter into the homozygous *glass* mutant background. Indeed, for all three enhancers, the activity is entirely gone (Fig. 5; Supplemental Fig. S24). We could furthermore confirm that the Chaoptic and Lozenge proteins, for which antibodies are available, are either gone (Chaoptic) or severely affected (Lozenge) in the glass mutant (Supplemental Fig. S17), in agreement with previous reports (Treisman and Rubin 1996; Firth and Baker 2007). Overall, we have achieved a high success rate of Glass target gene predictions, both in terms of their perturbation in the mutant eye, and in terms of PR-specific enhancer-reporters.

Discussion

Sequencing-based expression profiling using RNA-seq or Tag-seq provides the opportunity to obtain genome-wide quantitative gene expression levels in any tissue and in any species (McManus et al. 2010; Hong et al. 2011). Here we use comparative transcriptomics in combination with comparative motif discovery. The comparative transcriptomics was performed during retinal determination in three *Drosophila* species (*D. melanogaster*, *D. yakuba*, and *D. virilis*) using Tag-sequencing (Tag-seq, sometimes also called EDGE) (Saha et al. 2002; Hong et al. 2011). By using Tag-seq, the expression levels are based on one expressed sequence tag (EST) per transcript, corresponding to the 21 bp downstream from the most 3′-located NlaIII restriction site. After mapping the sequence reads, assigning ESTs to genes, and optimizing the filtering and normalization steps, we found this technique to deliver accurate gene expression levels in *D. melanogaster*. The nonmodel species *D. yakuba* and *D. virilis* have a lower-quality genome annotation and therefore require annotation amendments toward the 3′ end. Taking these annotation imperfections into account, either by extending the 3′ end of a gene, or by comparing the EST location to

Figure 4. Validation of predicted Glass target genes by RNA-seq. (*A*) Bean plots representing the $\log_2$ (gl[60j]/wild-type) for the 96 predicted Glass target genes, compared with the same values for control gene sets. As control sets we used three sets of genes expressed upstream of *glass*, namely, all genes from the Retinal Determination Gene Network (RDGN) from Amore and Casares (2010); Eyeless target genes from Ostrin et al. (2006); and Atonal target genes from Aerts et al. (2010). We also used five other negative control sets, unrelated to eye development. (***) Significant difference (Wilcoxon FDR <0.05), compared with the Glass target sets. (*Inset*) Gene Set Enrichment Analysis (GSEA) showing a significant (FDR <0.001) enrichment of the 96 genes at the top of the ranking. (*B*) The predicted Glass targetome with 96 genes, showing overall down-regulation (green). (Dashed red edges) Up-regulation; (blue edges) no expression changes in the mutant. (*C*) Comparison of the amount of validated, or true-positive Glass target predictions (*x*-axis; genes that are down-regulated in the glass mutant) and invalidated, or false-positive predictions (*y*-axis; genes that are not down-regulated in the mutant). All situations show a higher number of true positives than false positives. (Blue diamonds) Different filters based on the cisTargetX genomic ranking; (green diamond) the situation in the text (62/96 validated targets); (orange diamond) a very stringent filtering with 100% positive predictive value, although retaining only eight Glass target genes.

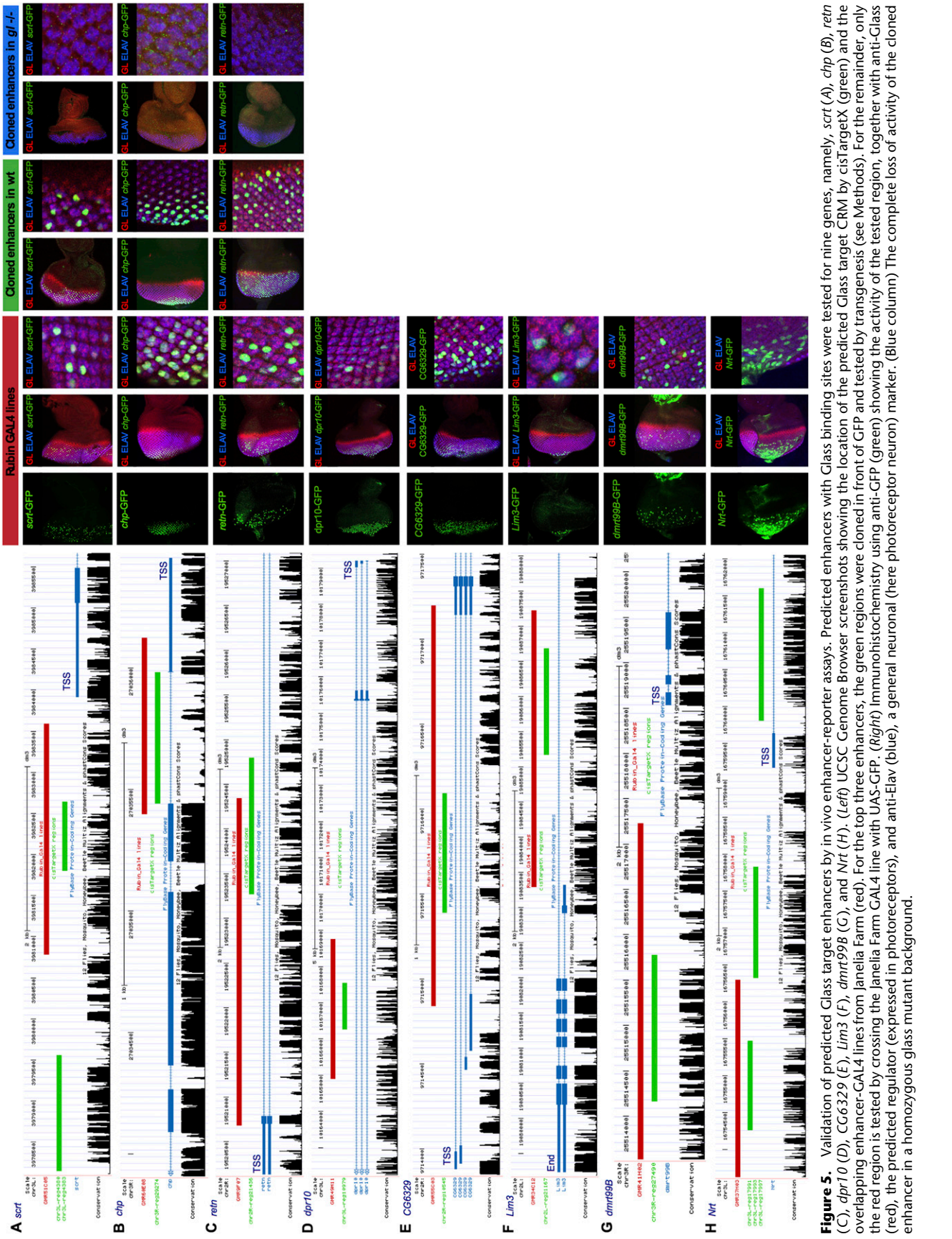


Figure 5. Validation of predicted Glass target enhancers by in vivo enhancer-reporter assays. Predicted enhancers with Glass binding sites were tested for nine genes, namely, *scrt* (*A*), *chp* (*B*), *retn* (*C*), *dpr10* (*D*), *CG6329* (*E*), *Lim3* (*F*), *dmrt99B* (*G*), and *Nrt* (*H*). (*Left*) UCSC Genome Browser screenshots showing the location of the predicted Glass target CRM by cisTargetX (green) and the overlapping enhancer-GAL4 lines from Janelia Farm (red). For the top three enhancers, the green regions were cloned in front of GFP and tested by transgenesis (see Methods). For the remainder, only the red region is tested by crossing the Janelia Farm GAL4 line with UAS-GFP. (*Right*) Immunohistochemistry using anti-GFP (green) showing the activity of the tested region, together with anti-Glass (red), the predicted regulator (expressed in photoreceptors), and anti-Elav (blue), a general neuronal (here photoreceptor neuron) marker. (Blue column) The complete loss of activity of the cloned enhancer in a homozygous glass mutant background.

orthologous positions in *D. melanogaster*, we were able to obtain accurate expression measures for *D. yakuba* and *D. virilis*.

To compare gene expression between species, we first normalized the gene expression levels in each species separately by dividing the expression in the eye disc by the expression in the wing imaginal disc. The wing imaginal disc is a good control because it also consists of epithelium and is taken at the same developmental time. Through a post-analysis, we confirmed that eye-enriched genes compared with the wing disc are also eye-enriched compared with the entire larva (Supplemental Fig. S25). We then focused on conserved eye-specific genes across the three species by ranking all genes according to the eye-versus-wing differential expression and integrating these rankings by order statistics (OS). We compared this procedure with a direct statistical assessment of eye-specific gene expression, using the three species as replicates (three eye samples vs. three wing samples). None of the available packages (DESeq, NOISeq, edgeR) outperformed the OS integrated ranking. Interestingly, the OS-based integration is robust to missing values, and genes with strong eye-enrichment for two species, but a missing value in the third species can still be ranked very high (for example, no orthologous gene exists in one species; or the transcript has no NlaIII site; or the coverage is too low). On the other hand, the statistical methods used for differential expression across the species are very sensitive to missing values, and we even had to restrict our comparison to 1-1-1 orthologs. This finding is important when more species are included in the analysis, which would result in too stringent filtering if expression levels are required for all species analyzed.

For the comparative motif discovery, we used a cross-species approach called cisTargetX. We have shown before that this approach delivers accurate motif discovery and CRM prediction results (Aerts et al. 2010) . This method takes a set of coexpressed genes as input and identifies enriched motifs (present in the 5 kb upstream and first intron regions) using whole-genome scoring for homotypic motif clusters across the 12 sequenced *Drosophila* species. As an input set, we selected the top 245 genes from the OS-based cross-species gene ranking. This cutoff was determined based on the finding that this "leading edge" contains the strongest enrichment of genes involved in photoreceptor differentiation, as determined by a ranking-based GO analysis (Eden et al. 2009). This finding is not unexpected because at this stage of development, the eye-antennal disc is strongly enriched for photoreceptor cells. We analyzed this set of 245 eye-specific genes for enriched motifs using cisTargetX and identified motifs for several eye-related TFs, including Glass, SU(H), SOXN, and Scratch. For each of the enriched motifs, cisTargetX predicts the optimal subset of direct target genes, which is then considered as the candidate "targetome" of the corresponding transcription factor. Putting the targets for all factors together in a network allowed us to connect already 149 of the 245 input genes. Although additional regulators and TF–target interactions can be found when alternative input sets are used (e.g., the top 545 genes, or filtered gene sets using other data sets) (see Supplemental Table S6), it remains a future challenge to predict a regulator for all genes in a signature.

We focused on the targetome of the highest ranked motif, namely, Glass, which consists of 96 predicted Glass target genes (out of the 245 genes used as input). Thus far, only one direct target gene of Glass has been reported in the tissue under study, namely, *lozenge*, which is an activating regulatory interaction. Together with the fact that we start from highly eye-enriched genes, often also highly expressed genes, we expected to find additional target genes that are activated by Glass, rather than repressed. In agreement with this hypothesis, we find a significant amount of predicted Glass targets to be down-regulated in the *glass* mutant. This results in 62 direct and activated candidate Glass target genes, including *lozenge*. If we include more genes as input to cisTargetX, additional Glass target genes are found (for example, *Pph13*, a known TF involved in photoreceptor morphogenesis) (Zelhof et al. 2003; Mishra et al. 2010), but we decided to present the core set of 62 targets in this work (the larger list of targets is available from our website, through the cisTargetX results). To test these candidate targets further, we examined the predicted CRMs with Glass binding sites using in vivo reporter assays. To this end, we have made use of a recently available collection of GAL4 lines from Janelia Farm (Pfeiffer et al. 2008) that contains overlapping genomic regions around genes involved in the development and functioning of the central nervous system. Because we are studying eye development and photoreceptor neuron differentiation, many of the eye-enriched genes also play a role in the CNS, and therefore we found several GAL4 lines, nine in particular, for genomic regions that overlap our CRM predictions (based on the current availability in the Bloomington stock center). Remarkably, eight of the nine tested lines (88.8%) show expression in the eye disc, downstream from Glass, and for seven of these eight GFP colocalizes with Glass in photoreceptor cells. This brings the number of in vivo–validated eye enhancers activated by Glass from one (the *lz* eye enhancer) (Yan et al. 2003) to eight. Note that these enhancers are found ab initio, without starting from a set of "training enhancers." Indeed, many computational methods for CRM prediction rely on a training set (for review, see Aerts 2012), but in our case, as in many other circumstances, such data are not available. Here we show that, purely based on gene expression measurements in wild-type tissue, enhancers with a specific function, and activated by a specific TF, can be identified on a genome-wide scale.

Following the identification of many true-positive targets, we also examined the set of invalidated genes that are not significantly down-regulated in the *glass* mutant, such as *eya* (fold-change = 1.14 up), *ato* (fold-change = 0.90 down), and *phyl* (fold change = 1.00 up). Among these 34 are also 11 genes that are significantly up-regulated in the *glass* mutant, although we note that the fold-change in expression is markedly smaller than the fold-changes of the down-regulated genes, with examples such as *Optix* (1.32-fold up) and *Dfd* (1.61-fold up). Using antibodies against some of the encoded proteins (*eya*, *Optix*, *Dfd*, *Ato*) (see Supplemental Fig. S17) or crossing candidate CRMs into a *glass* mutant background (*phyl*) (see Supplemental Fig. S24), we found no obvious changes in expression in the *glass* mutant, suggesting that they are, indeed, most likely false-positive predictions. Importantly, we find no evidence that Glass could have a repressive role, because (1) there is overall very little up-regulation in the glass mutant, compared with the very strong down-regulation of validated targets (up to 1000-fold); (2) the entire RDGN is, independently of Glass, slightly up-regulated because of noncell-autonomous effects (see bean plot in Fig. 4A); and (3) genes that are not expressed in Glass-expressing cells in wild-type discs are not activated in these cells in the *glass* mutant, as we have shown for *Optix* and *Dfd*. Altogether, we conclude that Glass is mainly an activator, and that the RNA-seq-based filter for down-regulation in the mutant allows separating true- from false-positive predictions. Interestingly, we identified particular parameter settings and filters that can shift the ratio of false-positive predictions, one of which leads to eight predicted Glass targets with a 100% positive predictive value.

Finally, to give nuances to the above separation of true and false positives, we note that such separation depends on the res-

olution of the assay. One can imagine that the global mRNA expression level of a gene does not change significantly in the mutant versus wild-type tissue. For example, it is likely that for many targets, Glass is not the only regulator, so removing Glass may not entirely abolish the activation of the target, but only cause a subtle change. Also, we find that Glass directly activates several repressors, such as Lozenge and Scratch, which, in turn, may also bind to Glass target CRMs (many genes in the network of Fig. 3B are predicted as targets of Glass *and* Scratch) yielding incoherent feed-forward loops. Removing Glass could result in a decrease of the repressor and a derepression (hence up-regulation) of the shared targets. One intriguing CRM we predicted where Glass could act as a cofactor is Eya. The predicted Glass binding sites are located right inside the known eye enhancer located just upstream of the *eya* transcription start site, where Eyeless is known to bind (Bui et al. 2000). Although the RNA-seq data invalidated the Glass → *eya* interaction, we investigated this candidate further because of the very strong CRM score (ranked 188th in the entire genome) and because *eya* is expressed in two domains, namely, anterior to the morphogenetic furrow (in the RDGN), and posterior to the furrow in all differentiating photoreceptors. Particularly, we assessed the levels of EYA protein quantitatively inside ELAV-positive versus ELAV-negative cells, before and after the morphogenetic furrow, in wild-type and *glass* mutant discs, and could detect a small change of EYA expression inside ELAV-positive cells posterior to the furrow (Supplemental Fig. S25). We believe that in future work, quantitative expression analysis will play an important role to identify small quantitative effects and to begin modeling the networks we began to map here (Jaeger et al. 2004).

In this study, we show that for systems that are not easily amenable to ChIP, cross-species transcriptomics followed by computational motif discovery allows accurate predictions of targets and CRMs. Subtle interactions remain uncertain at the resolution of our assays, such as the Glass → *eya* interaction. Such interactions would benefit from complementary ChIP-seq data against Glass and cofactors. Currently, to our knowledge, no ChIP-seq has been performed yet against sequence-specific TFs in specific cell types within eye imaginal discs. This may become possible in the future because technological advances, including recombineering-mediated tagging of transcription factors (Venken et al. 2008) and miniaturization of ChIP protocols (Adli and Bernstein 2011), are paving the way to overcome the challenge of ChIP on low input material and the need for ChIP-grade antibodies for every TF. Because such approaches will remain costly and technically challenging, we believe our strategy provides a straightforward alternative to map gene regulatory networks.

In conclusion, by integrating gene expression across three *Drosophila* species, we obtain a high-quality set of conserved tissue-specific genes, representing the core of the developmental process under study, in our case, *Drosophila* retinal determination. This core set of eye-specific genes shows stronger functional enrichment than eye-specific genes obtained from a single species only. The motif discovery results on the conserved set are more accurate, both in terms of specificity and sensitivity, which indicates that the genes with conserved expression are more tightly coregulated than genes derived from one species. This strategy is generally applicable to conserved organs and allows us to probe wild-type tissues without the requirement of genetic perturbations of transcription factors, or other enrichment procedures (e.g., cell-type-specific expression profiling using cell sorting, or chromatin immunoprecipitation). Massively parallel sequencing technologies thus allow using "species as replicates" to discover the conserved patterns in a developmental program.

Methods

Fly stocks and antibodies

The fly strains used were *D. melanogaster* Canton-S and *yw*. For *D. yakuba* and *D. virilis*, we used the sequenced strains, obtained from the San Diego Stock Center (stock number 14021-0261.01 and 15010-1051.87, respectively). All flies were raised at 25°C on standard fly food. For immunohistochemistry, imaginal discs of wandering third instar larva were dissected and processed as described (Wang et al. 2002). The anti-Optix antibody was a kind gift of F. Pignoni, and anti-Dfd of T. Kaufman. The antibodies against CHP raised by S. Benzer, LZ raised by U. Banerjee, ELAV and Glass raised by G.M. Rubin, and REPO raised by C. Goodman were obtained from the Developmental Studies Hybridoma Bank (DSHB), developed under the auspices of the NICHD, and maintained by The University of Iowa, Department of Biology (Iowa City, IA).

Imaginal disc dissections, RNA extraction, and qRT-PCR

Imaginal discs of wandering third instar larvae were dissected in PBS, and RNA for Tag-seq, RNA-seq, or qRT-PCR was extracted with the Mini RNA Isolation Kit (ZymoResearch). For qRT-PCR, we applied relative quantification with the comparative ddCT method (SDS User bulletin 2; Applied Biosystems) with the Roche Lightcycler 480 SYBR Green Master Mix 2 (Roche Diagnostics) on the Roche Lightcycler 480 instrument. Total RNA of eye-antennal imaginal discs was converted to cDNA using the QuantiTect Reverse Transcription Kit (QIAGEN). Primers were designed with a Roche Lightcycler 480 probe design and are available upon request. As housekeeping gene, we used *rpl32*. RNA of eye-antennal imaginal discs of Canton-S wild type was used as the control sample. After an initial denaturation step for 10 min at 95°C, thermal cycling conditions were 15 sec at 95°C and 1 min at 60°C for 40 cycles.

Illumina Tag-sequencing

Around 1–3 μg of total RNA was used per sample (70–80 larvae), and NlaIII-Digital Gene Expression libraries were generated following the Illumina guidelines. In brief, the total RNA per sample was bound to oligo(dT) beads. Double-stranded cDNA was synthesized and digested using an NlaIII restriction enzyme. Next, adapters were added to the 5′ end of the fragments. A second digestion with MmeI cuts 17 bp downstream from the NlaIII site and is followed by 3′-adapter ligation and tag enrichment by PCR. Finally, sequencing was performed on the Illumina Genome Analyzer (GAII). Illumina's Pipeline's FireCrest was used to convert sequencing cycle images to signal intensities, and the Bustard algorithm (Bentley et al. 2008) was run to perform base and calculate quality scores for every base. Quality assessment analysis of Phred scores per sequencing cycle and per lane was performed using the ShortRead Bioconductor package (Morgan et al. 2009).

Tag-seq data analysis

Sequencing reads corresponding to 17-bp tags were converted to 21-bp tags by adding the 5′ NlaIII restriction site (CATG). The 21-bp tags were aligned to the corresponding FlyBase genome assemblies: *Drosophila melanogaster* release 5, *Drosophila yakuba* and *Drosophila virilis* release 1. *D. yakuba* and *D. virilis* samples were also

mapped to University of California, Santa Cruz (UCSC) genome assemblies (WUGSC 7.1/droYak2) and (droVir3), respectively. The mappings were performed using bowtie (Langmead et al. 2009) with maximally two mismatches per read. Only tags mapping uniquely to the reference genome were processed further. For each gene, we considered the position with the maximal number of tags to determine the gene expression level. Applied normalization methods were total count normalization, upper-quartile normalization (Bullard et al. 2010) and trimmed mean of M-values (TMM) (Robinson and Oshlack 2010). Total count and upper-quartile normalization consisted of dividing each expression value for the total sum of counts assigned to genes (library size), or by the upper quartile of gene counts and multiplied by the average total-count or upper-quartile gene counts between samples. TMM normalization was performed using the calcNormFactors function from the edgeR package and multiplying with the scaling factor for each library size. Differential expression (DE) analysis was performed by edgeR (version 2.05) (Robinson et al. 2010), DESeq (version 1.2.1) (Anders and Huber 2010), and NOISeq (Tarazona et al. 2011) packages. Zero values were adjusted by adding 1, to avoid missing values (infinity) in the log ratios.

Cross-species integration, coordinate orthology, and gene orthology

Gene orthology was obtained from EnsemblCompara GeneTrees (Vilella et al. 2009). Coordinate orthology was obtained from whole-genome alignment (*.chain*) files from the UCSC Genome Browser and the liftOver tool (Fujita et al. 2010). Order statistics was used as described (Aerts et al. 2006).

Comparison to publicly available data sets

Data set GSE4008 (Ostrin et al. 2006) was analyzed from the CEL files using BioConductor, RMA for normalization, and Limma for differential expression, yielding 507 *D. melanogaster* eye-enriched genes with $\log_2$(eye/wing) >2 and FDR <0.05. ROC curves were performed using the ROCR (version 1.0-4) BioConductor package.

Motif discovery

Motif discovery was performed with cisTargetX (http://med.kuleuven.be/lcb/cisTargetX) as described before (Aerts et al. 2010; Herrmann et al. 2012) using version 1 of the motif collection (3731 position weight matrices), and using the 5 kb upstream and first intron as search space. For each motif, this search space is scored for clusters of PWM matches using a Hidden Markov Model, and orthologous regions of 11 other *Drosophila* species are scored in parallel. This results in 12 whole-genome rankings per motif, and these are combined by rank aggregation into one whole-genome ranking for each motif. For an input set of genes, the motifs for which the gene ranking is significantly enriched for input genes at the highly ranked genes are identified, and for each significant motif, the optimal threshold is determined through a Receiver Operator Characteristic curve. For details, we refer to the original cisTargetX and i-cisTarget publications. Full analysis results are available from the cisTargetX website.

Gene Regulatory Network visualization

Gene Regulatory Network visualization was performed using Cytoscape 2.8.1 (Smoot et al. 2011). Expression levels for the three species were represented as $\log_2$(eye-antennal/wing) values in node colors using the MultiColoredNodes cytoscape plugin (version 2.4.12) (Warsow et al. 2010).

RNA-seq

Fly stocks from *D. melanogaster* wild type (Canton-S and strain *RAL-208* from the inbred collection of T. Mackay) (Jordan et al. 2007; Ayroles et al. 2009) and the *glass* mutant line (*gl[60j]*, stock 507 from the Bloomington Stock Center) were maintained at room temperature. Eye-antennal and wing imaginal discs were dissected, followed by RNA extraction, yielding ~3 μg of total RNA per sample, to be processed to libraries according to the Illumina TruSeq protocol with appropriate indices, pooled, and sequenced on the Illumina HISeq 2000.

RNA-seq data analysis

Reads containing residuals of adapter sequences were discarded (FastX clipper version 0.0.13 with option -M15). Quality control assessment on raw sequenced reads was performed using the software FastQC (version 0.9), checking for PHRED quality >20 and different primer contaminations. Reads passing the filtering were mapped against the *D. melanogaster* FlyBase genome release 5 with TopHat v.2.0 (default parameters) (Trapnell et al. 2009). Gene expression measures were computed by HT-Seq (Anders and Huber 2010) (option -str=no) using *D. melanogaster* gene annotation release 5.30. Differential expression between *[gl60j]* and wild type (Canton-S and strain RAL-208 from the inbred collection of T. Mackay) (Jordan et al. 2007; Ayroles et al. 2009) was calculated with DESeq (v.1.2.1) using FDR <0.05, where only genes with more than 1 read per million in two samples were assessed for differential gene expression.

Enhancer-reporter assays

Out of 3029 GAL4 lines made available from Janelia Farm (Pfeiffer et al. 2008) in Bloomington stock center (28 November 2011), nine GAL4 lines covered the cisTargetX-predicted glass binding CRM. We crossed these lines to UAS-GFP lines, to assess whether GFP expression was observed in *D. melanogaster* eye-antennal third instar wandering larvae. Enhancer regions containing the predicted *glass*-binding motif were PCR-amplified from genomic DNA of *D. melanogaster* or of *D. virilis* and cloned into the phiC31 and Gateway compatible reporter vector pH-attB-Dest (Aerts et al. 2010), injected into VK37 (Venken et al. 2006) by Genetivision, and crossed together to generate homozygous stocks.

Quantitative immunohistochemical analysis

To quantify changes of EYA expression in the *glass* mutant, confocal stacks of immunostained samples, including DAPI as nuclear marker, were used. Single nucleus resolution samples were obtained by thresholding the DAPI staining signal and fitting the spots to geometric spheroids by scanning different major semiaxis lengths within an interval that is characteristic of nuclear size in the eye imaginal disc. These data were spatially transformed along the anterior–posterior axis and registered relative to the morphogenetic furrow, so that the negative semiaxis corresponds to precursor cells anterior to the furrow, and the positive semiaxis corresponds to the differentiated photoreceptors and accessory cells, posterior to it. A full account of the imaging and computational method will be described elsewhere.

Data access

Tag-seq data (two tissues, three species) and RNA-seq data (two wild-type *D. melanogaster* strains, and the *glass* mutant) are available from GEO (accession number GSE39784). CisTargetX and

the results from our cisTargetX analyses are available from the cisTargetX website at http://med.kuleuven.be/lcb/cisTargetX.

Acknowledgments

We are grateful to T. Kaufman and F. Pignoni for providing antibodies. We thank Rekin's Janky and Katina Spanier for insightful comments on the manuscript. This work is funded by research grants from Research Foundation Flanders (FWO, grant G.0704.11N), University of Leuven (CREA/10/014 and PF/10/016), and Human Frontiers Science Program (RGY0070/2011). M.N.S. is funded by a PhD fellowship from FWO.

References

Adli M, Bernstein BE. 2011. Whole-genome chromatin profiling from limited numbers of cells using nano-ChIP-seq. *Nat Protoc* **6:** 1656–1668.
Aerts S. 2012. Computational strategies for the genome-wide identification of *cis*-regulatory elements and transcriptional targets. *Curr Top Dev Biol* **98:** 121–145.
Aerts S, Thijs G, Coessens B, Staes M, Moreau Y, De Moor B. 2003. Toucan: Deciphering the *cis*-regulatory logic of coregulated genes. *Nucleic Acids Res* **31:** 1753–1764.
Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, De Smet F, Tranchevent L-C, De Moor B, Marynen P, Hassan B, et al. 2006. Gene prioritization through genomic data fusion. *Nat Biotechnol* **24:** 537–544.
Aerts S, Quan X-J, Claeys A, Naval Sanchez M, Tate P, Yan J, Hassan BA. 2010. Robust target gene discovery through transcriptome perturbations and genome-wide enhancer predictions in *Drosophila* uncovers a regulatory basis for sensory specification. *PLoS Biol* **8:** e1000435. doi: 10.1371/journal.pbio.1000435.
Amore G, Casares F. 2010. Size matters: The contribution of cell proliferation to the progression of the specification *Drosophila* eye gene regulatory network. *Dev Biol* **344:** 569–577.
Anders S, Huber W. 2010. Differential expression analysis for sequence count data. *Genome Biol* **11:** R106. doi: 10.1186/gb-2010-11-10-r106.
Ayroles JF, Carbone MA, Stone EA, Jordan KW, Lyman RF, Magwire MM, Rollmann SM, Duncan LH, Lawrence F, Anholt RRH, et al. 2009. Systems genetics of complex traits in *Drosophila melanogaster*. *Nat Genet* **41:** 299–307.
Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456:** 53–59.
Bui QT, Zimmerman JE, Liu H, Gray-Board GL, Bonini NM. 2000. Functional analysis of an eye enhancer of the *Drosophila eyes absent* gene: Differential regulation by eye specification genes. *Dev Biol* **221:** 355–364.
Bullard JH, Purdom E, Hansen KD, Dudoit S. 2010. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* **11:** 94. doi: 10.1186/1471-2105-11-94.
Carroll SB, Grenier J, Weatherbee S. 2009. *From DNA to diversity: Molecular genetics and the evolution of animal design*, 2nd ed. Wiley, New York. http://books.google.be/books?id=tayrCszYKdkC.
Curtiss J, Heilig JS. 1997. Arrowhead encodes a LIM homeodomain protein that distinguishes subsets of *Drosophila* imaginal cells. *Dev Biol* **190:** 129–141.
Davidson EH. 2001. *Genomic regulatory systems*. Academic Press, San Diego.
Diederich RJ, Pattatucci AM, Kaufman TC. 1991. Developmental and evolutionary implications of labial, deformed and engrailed expression in the *Drosophila* head. *Development* **113:** 273–281.
Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. 2009. *GOrilla*: A tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* **10:** 48. doi: 10.1186/1471-2105-10-48.
Ellis MC, O'Neill EM, Rubin GM. 1993. Expression of *Drosophila glass* protein and evidence for negative regulation of its activity in non-neuronal cells by another DNA-binding protein. *Development* **119:** 855–865.
Firth LC, Baker NE. 2007. Spitz from the retina regulates genes transcribed in the second mitotic wave, peripodial epithelium, glia and plasmatocytes of the *Drosophila* eye imaginal disc. *Dev Biol* **307:** 521–538.
Frith MC, Fu Y, Yu L, Chen J-F, Hansen U, Weng Z. 2004. Detection of functional DNA motifs via statistical over-representation. *Nucleic Acids Res* **32:** 1372–1381.
Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, Cline MS, Goldman M, Barber GP, Clawson H, Coelho A, et al. 2010. The UCSC Genome Browser database: Update 2011. *Nucleic Acids Res* **39:** D876–D882.
Gordân R, Hartemink AJ, Bulyk ML. 2009. Distinguishing direct versus indirect transcription factor–DNA interactions. *Genome Res* **19:** 2090–2100.
Herrmann C, Van de Sande B, Potier D, Aerts S. 2012. i-cisTarget: An integrative genomics method for the prediction of regulatory features and cis-regulatory modules. *Nucleic Acids Res* **40:** e114. doi: 10.1093/nar/gks543.
Hong LZ, Li J, Schmidt-Kuntzel A, Warren WC, Barsh GS. 2011. Digital gene expression for non-model organisms. *Genome Res* **21:** 1905–1915.
Ho Sui SJ, Fulton DL, Arenillas DJ, Kwon AT, Wasserman WW. 2007. oPOSSUM: Integrated tools for analysis of regulatory motif over-representation. *Nucleic Acids Res* **35:** W245–W252.
Jaeger J, Surkova S, Blagov M, Janssens H, Kosman D, Kozlov KN, Manu , Myasnikova E, Vanario-Alonso CE, Samsonova M, et al. 2004. Dynamic control of positional information in the early *Drosophila* embryo. *Nature* **430:** 368–371.
Jemc J, Rebay I. 2007. Identification of transcriptional targets of the dual function transcription factor/phosphatase eyes absent. *Dev Biol* **310:** 416–429.
Jordan KW, Carbone MA, Yamamoto A, Morgan TJ, Mackay TFC. 2007. Quantitative genomics of locomotor behavior in *Drosophila melanogaster*. *Genome Biol* **8:** R172. doi: 10.1186/gb-2007-8-8-r172.
Kumar JP. 2010. Retinal determination: The beginning of eye development. *Curr Top Dev Biol* **93:** 1–28.
Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10:** R25. doi: 10.1186/gb-2009-10-3-r25.
Mahony S, Benos PV. 2007. STAMP: A web tool for exploring DNA-binding motif similarities. *Nucleic Acids Res* **35:** W253–W258.
McLeay RC, Bailey TL. 2010. Motif Enrichment Analysis: A unified framework and an evaluation on ChIP data. *BMC Bioinformatics* **11:** 165. doi: 10.1186/1471-2105-11-165.
McManus CJ, Coolon JD, Duff MO, Eipper-Mains J, Graveley BR, Wittkopp PJ. 2010. Regulatory divergence in *Drosophila* revealed by mRNA-seq. *Genome Res* **20:** 816–825.
Mishra M, Oke A, Lebel C, McDonald EC, Plummer Z, Cook TA, Zelhof AC. 2010. Pph13 and orthodenticle define a dual regulatory pathway for photoreceptor cell morphogenesis and function. *Development* **137:** 2895–2904.
Morgan M, Anders S, Lawrence M, Aboyoun P, Pagès H, Gentleman R. 2009. ShortRead: A bioconductor package for input, quality assessment and exploration of high-throughput sequence data. *Bioinformatics* **25:** 2607–2608.
Moses K, Rubin GM. 1991. Glass encodes a site-specific DNA-binding protein that is regulated in response to positional signals in the developing *Drosophila* eye. *Genes Dev* **5:** 583–593.
Moses K, Ellis MC, Rubin GM. 1989. The *glass* gene encodes a zinc-finger protein required by *Drosophila* photoreceptor cells. *Nature* **340:** 531–536.
Ostrin EJ, Li Y, Hoffman K, Liu J, Wang K, Zhang L, Mardon G, Chen R. 2006. Genome-wide identification of direct targets of the *Drosophila* retinal determination protein Eyeless. *Genome Res* **16:** 466–476.
Pauli T, Seimiya M, Blanco J, Gehring WJ. 2005. Identification of functional *sine oculis* motifs in the autoregulatory element of its own gene, in the eyeless enhancer and in the signalling gene *hedgehog*. *Development* **132:** 2771–2782.
Pfeiffer BD, Jenett A, Hammonds AS, Ngo T-TB, Misra S, Murphy C, Scully A, Carlson JW, Wan KH, Laverty TR, et al. 2008. Tools for neuroanatomy and neurogenetics in *Drosophila*. *Proc Natl Acad Sci* **105:** 9715–9720.
Pfreundt U, James DP, Tweedie S, Wilson D, Teichmann SA, Adryan B. 2010. FlyTF: Improved annotation and enhanced functionality of the *Drosophila* transcription factor database. *Nucleic Acids Res* **38:** D443–D447.
Pique-Regi R, Degner JF, Pai AA, Gaffney DJ, Gilad Y, Pritchard JK. 2011. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res* **21:** 447–455.
Potier D, Atak ZK, Sanchez MN, Herrmann C, Aerts S. 2012. Using cisTargetX to predict transcriptional targets and networks in *Drosophila*. *Methods Mol Biol* **786:** 291–314.
Robinson MD, Oshlack A. 2010. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* **11:** R25. doi: 10.1186/gb-2010-11-3-r25.
Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26:** 139–140.
Rogers EM, Brennan CA, Mortimer NT, Cook S, Morris AR, Moses K. 2005. Pointed regulates an eye-specific transcriptional enhancer in the *Drosophila hedgehog* gene, which is required for the movement of the morphogenetic furrow. *Development* **132:** 4833–4843.
Roider HG, Manke T, O'Keeffe S, Vingron M, Haas SA. 2009. PASTAA: Identifying transcription factors associated with sets of co-regulated genes. *Bioinformatics* **25:** 435–442.

Roignant J-Y, Legent K, Janody F, Treisman Jessica E. 2010. The transcriptional co-factor Chip acts with LIM-homeodomain proteins to set the boundary of the eye field in *Drosophila*. *Development* **137:** 273–281.
Sabo PJ, Kuehn MS, Thurman R, Johnson BE, Johnson EM, Cao H, Yu M, Rosenzweig E, Goldy J, Haydock A, et al. 2006. Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays. *Nat Methods* **3:** 511–518.
Saha S, Sparks AB, Rago C, Akmaev V, Wang CJ, Vogelstein B, Kinzler KW, Velculescu VE. 2002. Using the transcriptome to annotate the genome. *Nat Biotechnol* **20:** 508–512.
Silver SJ, Rebay I. 2005. Signaling circuitries in development: Insights from the retinal determination gene network. *Development* **132:** 3–13.
Small S, Arnosti DN, Levine M. 1993. Spacing ensures autonomous expression of different stripe enhancers in the even-skipped promoter. *Development* **119:** 767–772.
Smoot ME, Ono K, Ruscheinski J, Wang P-L, Ideker T. 2011. Cytoscape 2.8: New features for data integration and network visualization. *Bioinformatics* **27:** 431–432.
Song L, Zhang Z, Grasfeder LL, Boyle AP, Giresi PG, Lee BK, Sheffield NC, Gräf S, Huss M, Keefe D, et al. 2011. Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome Res* **21:** 1757–1767.
Tarazona S, García-Alcalde F, Dopazo J, Ferrer A, Conesa A. 2011. Differential expression in RNA-seq: A matter of depth. *Genome Res* **21:** 2213–2223.
Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics* **25:** 1105–1111.
Treisman JE, Rubin GM. 1996. Targets of *glass* regulation in the *Drosophila* eye disc. *Mech Dev* **56:** 17–24.
Van Loo P, Aerts S, Thienpont B, De Moor B, Moreau Y, Marynen P. 2008. ModuleMiner—improved computational detection of *cis*-regulatory modules: Are there different modes of gene regulation in embryonic development and adult tissues? *Genome Biol* **9:** R66. doi: 10.1186/gb-2008-9-4-r66.
Venken KJT, He Y, Hoskins RA, Bellen HJ. 2006. P[acman]: A BAC transgenic platform for targeted insertion of large DNA fragments in *D. melanogaster*. *Science* **314:** 1747–1751.
Venken KJT, Kasprowicz J, Kuenen S, Yan J, Hassan BA, Verstreken P. 2008. Recombineering-mediated tagging of *Drosophila* genomic constructs for in vivo localization and acute protein inactivation. *Nucleic Acids Res* **36:** e114. doi: 10.1093/nar/gkn486.
Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. 2009. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res* **19:** 327–335.
Wang VY, Hassan BA, Bellen HJ, Zoghbi HY. 2002. *Drosophila* atonal fully rescues the phenotype of Math1 null mice: New functions evolve in new cellular contexts. *Curr Biol* **12:** 1611–1616.
Warner JB, Philippakis AA, Jaeger SA, He FS, Lin J, Bulyk ML. 2008. Systematic identification of mammalian regulatory motifs' target genes and functions. *Nat Methods* **5:** 347–353.
Warsow G, Greber B, Falk SSI, Harder C, Siatkowski M, Schordan S, Som A, Endlich N, Schöler H, Repsilber D, et al. 2010. *ExprEssence*—revealing the essence of differential experimental data in the context of an interaction/regulation net-work. *BMC Syst Biol* **4:** 164. doi: 10.1186/1752-0509-4-164.
Wittkopp PJ, Kalay G. 2012. *Cis*-regulatory elements: Molecular mechanisms and evolutionary processes underlying divergence. *Nat Rev Genet* **13:** 59–69.
Won K-J, Ren B, Wang W. 2010. Genome-wide prediction of transcription factor binding sites using an integrated model. *Genome Biol* **11:** R7. doi: 10.1186/gb-2010-11-1-r7.
Worsley-Hunt R, Bernard V, Wasserman WW. 2011. Identification of *cis*-regulatory sequence variations in individual genome sequences. *Genome Med* **3:** 65. doi: 10.1186/gm281.
Wray GA. 2007. The evolutionary significance of *cis*-regulatory mutations. *Nat Rev Genet* **8:** 206–216.
Yan H, Canon J, Banerjee U. 2003. A transcriptional chain linking eye specification to terminal determination of cone cells in the *Drosophila* eye. *Dev Biol* **263:** 323–329.
Zelhof AC, Koundakjian E, Scully AL, Hardy RW, Pounds L. 2003. Mutation of the photoreceptor specific homeodomain gene Pph13 results in defects in phototransduction and rhabdomere morphogenesis. *Development* **130:** 4383–4392.
Zhu LJ, Christensen RG, Kazemian M, Hull CJ, Enuameh MS, Basciotta MD, Brasefield JA, Zhu C, Asriyan Y, Lapointe DS, et al. 2011. FlyFactorSurvey: A database of *Drosophila* transcription factor binding specificities determined using the bacterial one-hybrid system. *Nucleic Acids Res* **39:** D111–D117.

Received March 14, 2012; accepted in revised form September 24, 2012.