Automated protein-DNA interaction screening of *Drosophila* regulatory elements

Korneel Hens[1], Jean-Daniel Feuz[1], Alina Isakova[1], Antonina Iagovitina[1], Andreas Massouras[1], Julien Bryois[1], Patrick Callaerts[2,3], Susan E Celniker[4] & Bart Deplancke[1]

Drosophila melanogaster **has one of the best characterized metazoan genomes in terms of functionally annotated regulatory elements. To explore how these elements contribute to gene regulation, we need convenient tools to identify the proteins that bind to them. Here we describe the development and validation of a high-throughput yeast one-hybrid platform, which enables screening of DNA elements versus an array of full-length, sequence-verified clones containing over 85% of predicted *Drosophila* transcription factors. Using six well-characterized regulatory elements, we identified 33 transcription factor–DNA interactions of which 27 were previously unidentified. To simultaneously validate these interactions and locate the binding sites of involved transcription factors, we implemented a powerful microfluidics-based approach that enabled us to retrieve DNA-occupancy data for each transcription factor throughout the respective target DNA elements. Finally, we biologically validated several interactions and identified two new regulators of *sine oculis* gene expression and hence eye development.**

Since its adoption over 100 years ago, *Drosophila* has been a model organism used for studying the basic principles underlying many developmental and cellular processes, including transcriptional regulation. Specifically, the availability of a high-quality genome sequence[1], a large-scale enhancer trapping assay[2], chromatin immunoprecipitation (ChIP)–microarray or sequencing data revealing genome-wide *cis*-regulatory modules and specific chromatin states[3,4], a convenient transgenesis system to screen the activity of regulatory elements[5] as well as powerful comparative genomics methodologies[6] has led to the identification of many functional regulatory elements. To explore how these elements contribute to gene regulation and function in the context of gene regulatory networks, we need a technique to identify the transcription factors binding to these elements. Although several genome-wide techniques exist to determine which DNA elements are bound by a specific transcription factor (for example, ChIP, protein-binding microarrays and DNA adenine methyltransferase identification), techniques that identify the full complement of transcription factors binding to a specific DNA element often suffer from low throughput or high technical complexity[7]. Here we describe the development and validation of a high-throughput yeast one-hybrid (Y1H) system that enables interrogation of binding of transcription factors to selected DNA baits. By creating a nearly complete *Drosophila* transcription factor open reading frame (ORF) library, we optimized and validated this Y1H system for the fly, which allowed us to screen DNA baits versus the majority of predicted *Drosophila* transcription factors. As such, this technique may be instrumental to construct *Drosophila* gene regulatory networks.

RESULTS

A transcription factor ORF library

Building on previous efforts in *Caenorhabditis elegans*[8], we developed a gene-centered, Y1H-based platform that allows the high-throughput screening of DNA elements of interest versus the nearly complete *Drosophila* transcription factor repertoire. To obtain the latter, we determined, based on bioinformatic analyses[9] and manual curation, that the *Drosophila* genome contains 755 sequence-specific transcription factor–coding genes (**Supplementary Table 1**). Less than 15% of these have been characterized in terms of target genes[10]. Through incorporation of existing cDNA collections and *de novo* cloning, we generated 722 (96%) Gateway-compatible Entry clones (Invitrogen) containing the ORF of each transcription factor. We sequence-verified several Entry clones for each transcription factor using a recently developed high-throughput sequencing-based method[11], enabling us to confirm the identity of 692 transcription factors (92%) of which the majority is fully sequence-verified (588 or 78%) (**Fig. 1**). Cloned ORFs were distributed uniformly among all major transcription factor families (**Supplementary Fig. 1**).

The *Drosophila* high-throughput Y1H system

Most Y1H screens have so far been performed using direct transformation of the prey proteins in a haploid yeast strain in whose

[1]Laboratory of Systems Biology and Genetics, Institute of Bioengineering, School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland. [2]Laboratory of Developmental Genetics, Vlaams Instituut voor Biotechnologie, Leuven, Belgium. [3]Laboratory of Developmental Genetics, Department of Human Genetics, Catholic University of Leuven, Leuven, Belgium. [4]Department of Genome Dynamics, Berkeley *Drosophila* Genome Project, Lawrence Berkeley National Laboratory, Berkeley, California, USA. Correspondence should be addressed to B.D. (bart.deplancke@epfl.ch).

RECEIVED 24 JUNE; ACCEPTED 23 AUGUST; PUBLISHED ONLINE 30 OCTOBER 2011; DOI:10.1038/NMETH.1763

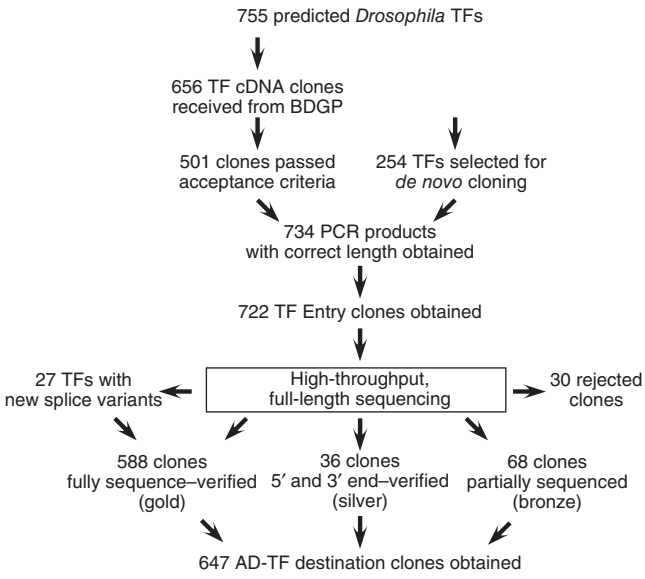

Figure 1 | Workflow underlying the generation of the *Drosophila* transcription factor (TF) ORF clone resource and the *Drosophila* Y1H AD transcription factor library. Of 755 predicted *Drosophila* transcription factors, 501 were available as cDNA clones from the Berkeley *Drosophila* Genome Project (BDGP). The remaining transcription factors were targeted for *de novo* cloning. Transcription factor ORFs were PCR-amplified and cloned into the pDONR221 Entry vector. The resulting Entry clones were sequence-verified by high-throughput sequencing and categorized according to the quality and the coverage of the sequencing into three classes: gold for fully sequence–verified clones, silver for 5′ and 3′ end-sequenced clones, and bronze for partially sequenced clones. All nonrejected clones were transferred into the Y1H-compatible AD vectors pAD-DEST-ARS/CEN and pAD-DEST-2μ by Gateway cloning.

genome the DNA bait is integrated. Recent efforts demonstrated that this haploid format allows a more comprehensive protein-DNA interaction coverage than mating-based assays, in which diploid strains are used to pair transcription factors with DNA baits[12]. However, haploid transformation is more laborious and expensive than mating-based assays, for which high-throughput platforms have recently been established[13,14], as hundreds of transcription factors must be manually transformed per screen. To pair optimal coverage with higher throughput and lower cost, we engineered a robotic platform that completely automates the haploid yeast transformation process (http://www.youtube.com/watch?v=PM8WWXgE1-A). In addition, we substantially decreased overall reagent consumption by scaling down the protocol to enable direct transformation in 384-well format (**Fig. 2**). Together, this allowed us to screen several DNA baits per day in fully automated fashion versus a *Drosophila* transcription factor array consisting of two 384-well plates currently containing 647 transcription factors and three negative controls (empty AD vector). We performed two independent screens per bait using selection reproducibility as the key criterion to filter out potential false positives. This procedure has been shown to be very effective in reducing false positives in yeast two-hybrid (Y2H) screens[15].

We initially also evaluated 'interactions' based on the expression of a second reporter, *LacZ*, but found that it was less sensitive than the *HIS3* reporter. For example, we found six interactions with the *LacZ* reporter versus 11 interactions identified with the *HIS3* reporter for one of the tested elements (*so10*). Additionally, we found the majority of interactions from the *lacZ* screen (5 of 6) in both independent *HIS3* screens (**Supplementary Figs. 2** and **3** and **Supplementary Table 2**), consistent with results obtained previously, that positives from one screen are typically positive for both reporters[16]. However, the *LacZ* screen can still be performed if additional stringency in selection of interactions is required.

An automated protein-DNA interaction detection tool

The identification of interactions by eye is often confounded by a varying background across the same yeast plate. To allow a more objective detection of interactions, we generated a Matlab-based image-analysis program, transcription factor-DNA interaction detection in yeast (TIDY). This program semiautomatically calls interactions by convoluting the image with the pattern of four bright spots on a dark background, which has the advantage of ignoring the noisy background of the image and only detecting the yeast colony array. TIDY also takes the uniformity of the quadrant colonies into account to filter out high-intensity values derived from only one or two contributing colonies.

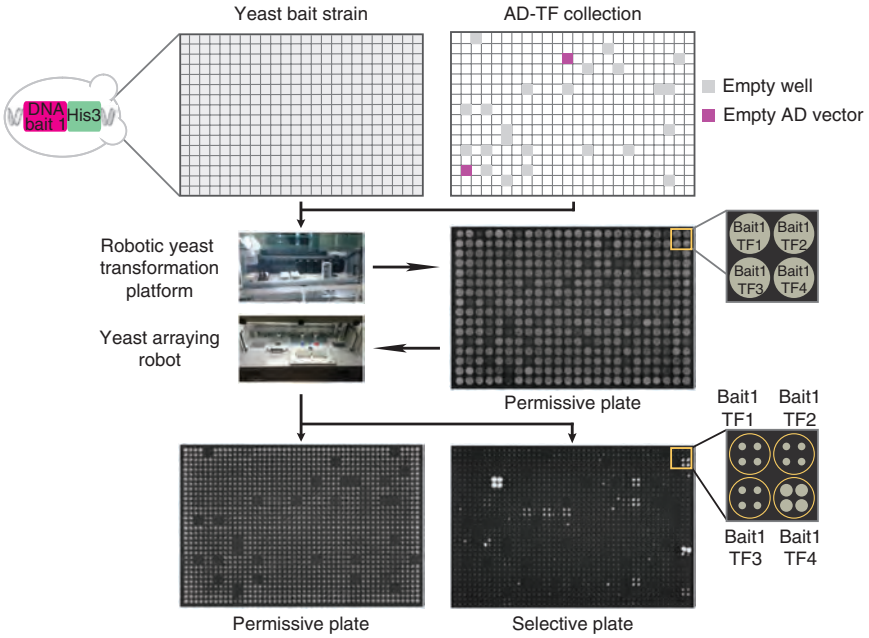

Figure 2 | *Drosophila* high-throughput Y1H platform. A yeast DNA-bait strain was distributed over a 384-well plate. Each well of this plate was then transformed with a different AD transcription factor clone from the *Drosophila* Y1H AD transcription factor library by a robotic yeast transformation platform, which additionally spotted the 384 individually transformed yeast strains on a permissive agar plate. A colony-pinning robot then transferred the yeast colonies onto a permissive and a selective plate, quadruplicating each colony in a square pattern in the process. Transcription factor–DNA bait interactions were identified based on growth on a selective, 3-amino-1,2,4-triazole–containing yeast plate.

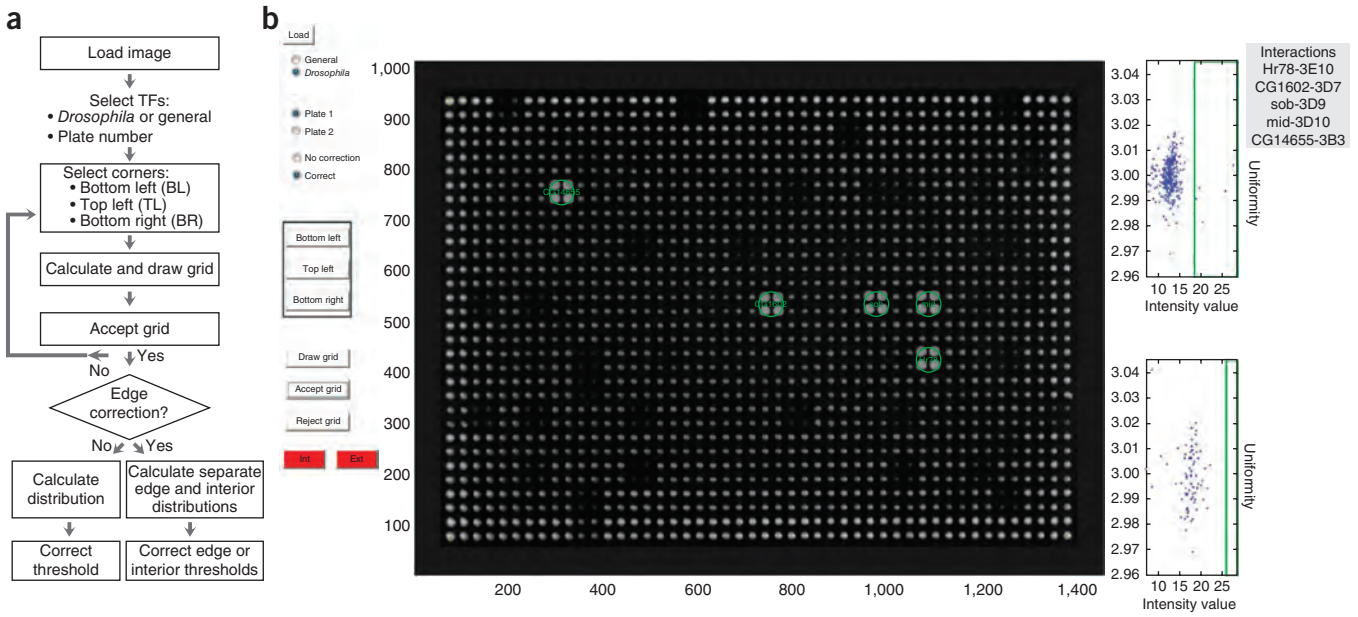

Figure 3 | Overview of the TIDY program. (**a**) Flowchart of TIDY program steps. (**b**) Screenshot of the TIDY output upon image analysis of a selective plate from a Y1H screen. In this example, five interactions were observed (green circles). A different threshold was used for plate-interior and plate-exterior yeast colonies.

Thus, uniform yeast quadrants whose resulting intensity values score above the threshold are identified as positives and labeled in green (**Fig. 3**). TIDY has the option to perform a separate background normalization for exterior versus interior yeast colonies as we often observed that colonies on the border of the plate grew faster than those in the middle (**Supplementary Fig. 4a**,**b**). Finally, TIDY allows the user to manually change the default threshold. In some cases, slightly lowering the threshold resulted in the inclusion of additional interactions that clearly still score above the highest background intensity value. We considered such interactions 'weak' and labeled them in magenta to indicate their distinct status (**Supplementary Fig. 4c**).

Drosophila Y1H validation

We selected 10 well-characterized *cis*-regulatory modules of 82–1,007 bp from the Regulatory Element Database for *Drosophila* (REDfly)[10] and the literature, based on the criterion of covering as many distinct transcription factors as possible. Four baits exhibited high self-activation (data not shown), and we did not consider them further as initial tests revealed that the interaction reproducibility dropped sharply with increasing self-activation. The six remaining elements together contributed 22 reported interactions (**Supplementary Table 3**). For 19 of these 22 interactions, the interacting transcription factor was present in our library. In total, we detected 33 transcription factor-DNA interactions that overlapped between two independent screens, involving 25 unique transcription factors belonging to 9 of the 11 main transcription factor families defined in **Supplementary Figure 1**. Representative TIDY-processed images are shown in **Supplementary Figures 2** and **5–9**, and the detected transcription factor-DNA interactions are listed in **Supplementary Table 2** and REDfly. We reproducibly detected five of 19 (26%) reported interactions, each involving a distinct transcription factor. This percentage falls in the range of Y1H and Y2H screen detection rates[8,17].

To evaluate whether some interactions were missed owing to the high-throughput nature of the screen, we retested the 19 reported interactions by manual transformation. Of the 14 interactions that were not detected previously, we recovered only two by manual transformation, showing the robustness of the automated Y1H system (**Supplementary Fig. 10** and **Supplementary Table 4**). The 10 remaining interactions may be missed because some transcription factors may require other proteins or post-translational modifications to bind DNA (that is, they may not be detectable in the Y1H system at all). However, half of the tested reported interactions so far have been observed using only one method, most of which were *in vitro* techniques such as electrophoretic mobility shift assays and DNase I footprinting and therefore involve naked DNA. Y1H DNA baits are integrated in the yeast genome and are thus 'chromatinized', which may result in more biologically relevant DNA binding behavior. Furthermore, some of the positive controls acted as repressors in *Drosophila* (for example Giant and Krüppel binding to the *eve-stripe2* element). It is possible that the repressive function of some of these transcription factors can overcome the activating function of the GAL4 activation domain, thereby preventing the *trans*-activation of the reporter gene, consistent with what was previously observed for the repressor TRA-1 in *C. elegans*[18]. We do not believe that this finding can be generalized to all repressors because we have reproducibly detected binding of repressing transcription factors (for example, Snail binding to *eve-stripe2* and Goosecoid binding to *so10*). Additionally, the requirement for interactions to test positive in two independent screens may sometimes be too stringent. Indeed, we found an additional reported interaction (*dpp813* and EXD) in one replicate.

To investigate putative factors influencing the positive detection rate, we retested the *dpp813* element using the *LacZ* reporter. Consistent with results obtained with the *so10* element, we identified fewer interactions using the *LacZ* reporter than in both *HIS3*

reporter screens and the *lacZ* screen recovered no reported interactions that were missed in the *HIS3* screens (**Supplementary Fig. 11** and **Supplementary Table 2**). Finally, we tested the influence of bait size or orientation by dividing the *dpp813* element into three overlapping elements or inverting the full-length *dpp813* element (**Supplementary Table 3**). Overall, reducing the size or inverting the element did not have a clear impact on overall coverage (**Supplementary Figs. 12–15** and **Supplementary Table 5**) but we found both reported interactions at least once more in the additional screens. Therefore we propose, for elements showing limited overlap between two independent screens, to perform additional repeats of the screen and use the number of times an interaction is observed as a confidence level to distinguish between spurious and likely true interactions.

Microfluidics-based validation and binding site mapping

We next estimated the proportion of interactions found by the Y1H screen that could be recapitulated with an alternative protein-DNA interaction detection technique. To this end, we used a microfluidic method based on mechanically induced trapping of molecular interactions (MITOMI)[19] for the analysis of regulatory elements (MARE) that we initially developed to validate mouse transcription factor–DNA interactions (C. Gubelmann, A. Isakova, A. Iagovitina, K.H., S.M. Waszak, J.-D.F. *et al.*; unpublished data).

First, we analyzed the *sine oculis* enhancer *so10*, as DNase I footprinting data for the well-known interactors Eyeless (EY) and Twin of eyeless (TOY) have previously been published[20], and can thus be used to benchmark the technique for *Drosophila*. We divided *so10* into 50 fragments of 36 base pairs (bp) with each fragment overlapping the previous one by 24 bp. We tested each fragment on-chip for recognition by Y1H-identified transcription factors and plotted DNA occupancy data for each 12-bp stretch (**Fig. 4** and **Supplementary Fig. 16**). We detected site-specific binding for eight of the 11 Y1H-selected transcription factors. EY and TOY reproducibly showed strong and similar binding patterns, consistent with the fact that they are both homologs of vertebrate PAX6 and have been shown to exhibit similar DNA binding properties[21]. The site yielding highest DNA occupancy overlapped with known EY and TOY binding sites[20], validating the MARE technology. Furthermore, five of six transcription factors for which positional weight matrix (PWM) data are available had a predicted binding site in *so10* that overlapped with a DNA occupancy peak detected by MARE.

We similarly tested the *yp1-1* element with the Y1H-detected transcription factors Doublesex (DSX) and Traffic jam (TJ). Using MARE, we detected binding sites for both transcription factors in the *yp1-1* element in specific locations. Additionally, PWM-based binding site prediction and DNase I footprinting[22] for DSX showed two binding sites in the highest DNA occupancy peak found using MARE (**Fig. 4** and **Supplementary Fig. 17**). Overall, we observed site-specific DNA binding for ten of 13 (77%) tested transcription factors, and the remaining three transcription factors produced mostly nonreproducible background signal, likely reflective of nonspecific binding.

In vivo relevance of detected Y1H interactions

We chose the *so10* element to estimate the proportion of interactions found by Y1H that could be relevant *in vivo*. Modulation of *so* expression results in readily observable eye phenotypes[23,24]; knockdown of transcription factors that regulate *so* expression *in vivo* should therefore also result in such phenotypes. We knocked down transcription factors identified to interact with so10 by crossing distinct upstream activating site–siRNA (UAS-RNAi) fly lines obtained from the Transgenic RNAi Project (TRiP) and Vienna *Drosophila* RNAi Center (VDRC) collections with *so10-GAL4* and *ok107* driver lines. As a first approach, we evaluated the effect on eye development by visual inspection of the adult eye. RNAi-mediated knockdown resulted in observable eye phenotypes for EY, Tramtrack (TTK) and CG9797 (**Fig. 5a–d**, **Supplementary Fig. 18** and **Supplementary Table 6**). Knockdown

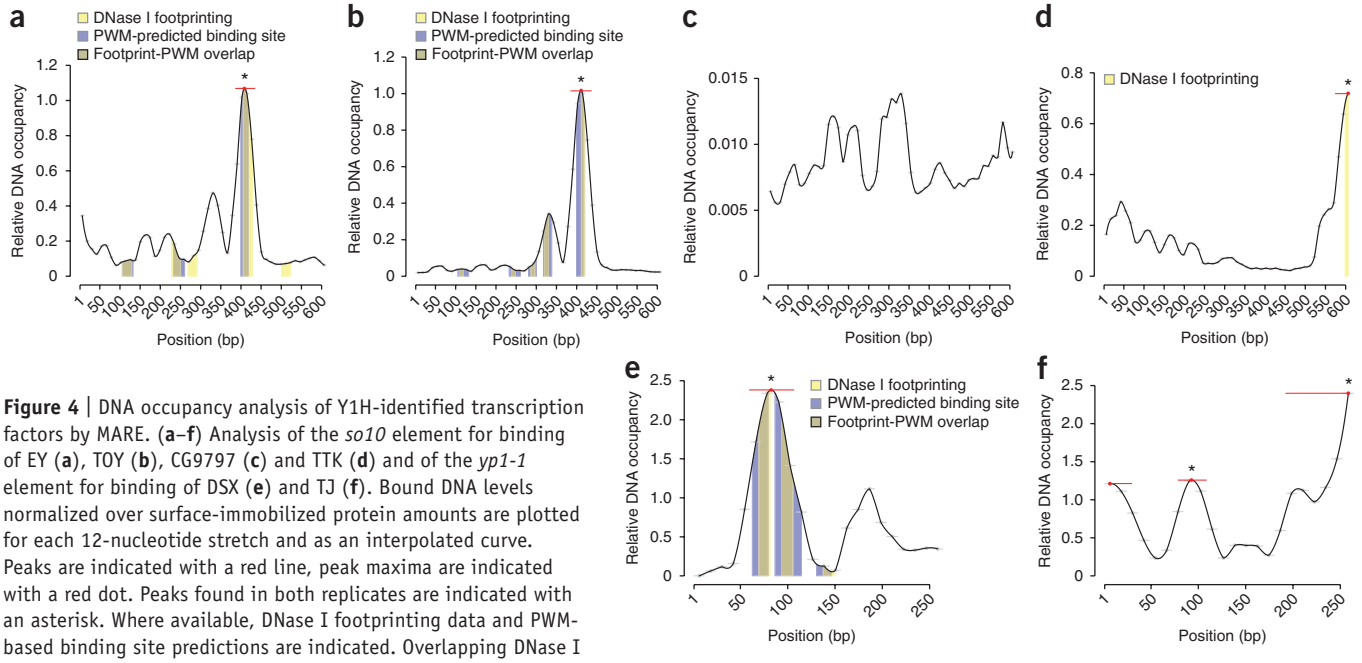

Figure 4 | DNA occupancy analysis of Y1H-identified transcription factors by MARE. (**a**–**f**) Analysis of the *so10* element for binding of EY (**a**), TOY (**b**), CG9797 (**c**) and TTK (**d**) and of the *yp1-1* element for binding of DSX (**e**) and TJ (**f**). Bound DNA levels normalized over surface-immobilized protein amounts are plotted for each 12-nucleotide stretch and as an interpolated curve. Peaks are indicated with a red line, peak maxima are indicated with a red dot. Peaks found in both replicates are indicated with an asterisk. Where available, DNase I footprinting data and PWM-based binding site predictions are indicated. Overlapping DNase I footprinting data and PWM-based binding site predictions are indicated. Note that as DNA occupancy is plotted as a relative signal normalized for the protein level in the microfluidics chamber, the scale of the *y* axis may vary between replicates.

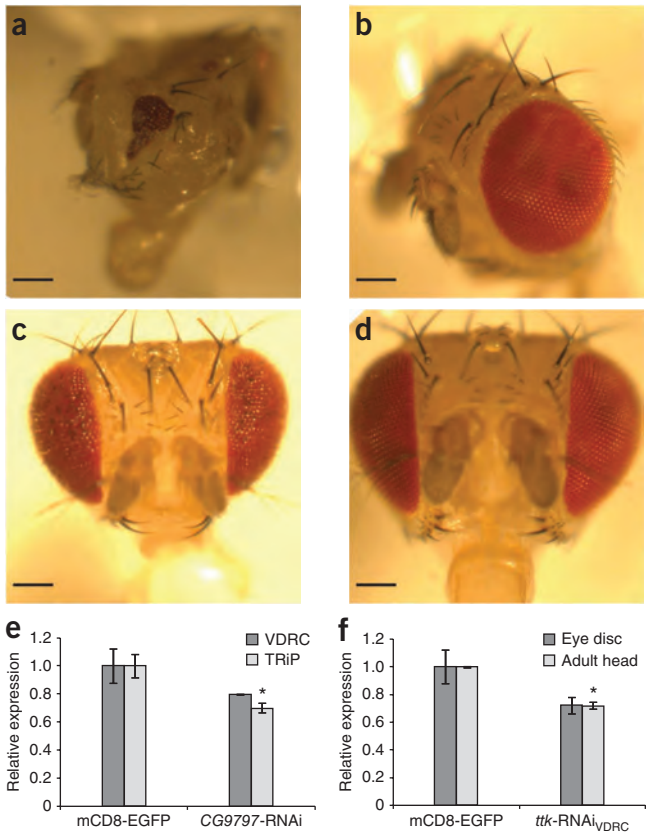

Figure 5 | *In vivo* effects of RNAi-mediated knockdown of Y1H-identified transcription factors binding the *so10* element. (**a**,**b**) Bright-field microscopy images of adult eyes, lateral view of *OK107>CG9797*-RNAi$_{TRiP}$ (**a**) and *OK107>UAS-mCD8-GFP* (**b**) flies. (**c**,**d**) Bright-field microscopy images of adult eyes, frontal view of *OK107>ttk*-RNAi$_{VDRC}$ (**c**) and *OK107>UAS-mCD8-GFP* (**d**) flies. Scale bars, 100 µm. (**e**,**f**) Quantitative real-time PCR analysis of *so* expression in third instar eye-antennal discs of *OK107>CG9797*-RNAi$_{VDRC}$ and *OK107>CG9797*-RNAi$_{TRiP}$ flies (**e**) and in the indicated tissues of *OK107>ttk*-RNAi$_{VDRC}$ flies (**f**). Values are relative to the corresponding controls. Error bars, s.e.m. ($n = 3$). *$P < 0.05$.

of *ey* and *CG9797* resulted in variable but similar eye phenotypes, ranging from completely absent to near-wild-type eyes, similar to the phenotype described for hypomorphic *ey* alleles[25]. *so10>CG9797*-RNAi$_{VDRC}$ flies (flies in which the expression of the VDRC *CG9797* RNAi construct is driven by the *so10* element) had a somewhat distinct phenotype, resulting in a protrusion of the eye coupled to a reduction of the eye perimeter. *OK107>ttk*-RNAi$_{VDRC}$ and *so10>ttk*-RNAi$_{VDRC}$ phenotypes were largely similar and had ommatidial degeneration consistent with the reported role of TTK in promoting photoreceptor cell differentiation at the late stages of eye development[26]. Both *so* and *ttk* are expressed in photoreceptor cells, and both mutants of *so* and *ttk* display defects in adult photoreceptor rhabdomeres[23,26], strengthening the hypothesis that TTK acts through SO in regulating photoreceptor cell differentiation.

To verify that the phenotypes of *ttk* and *CG9797* knockdown were caused by the misregulation of *so* expression, we quantified *so* mRNA levels in third instar eye-antennal discs of *OK107>CG9797*-RNAi$_{TRiP}$ and *OK107>ttk*-RNAi$_{VDRC}$ flies. As the *ttk* knockdown phenotype resembled ommatidial degeneration in the adult stage, we also evaluated *so* expression in adult heads of *so10>ttk*-RNAi$_{VDRC}$ flies. We observed a 20% and 30% reduction of *so* mRNA levels in third instar eye-antennal discs of *OK107>CG9797*-RNAi$_{VDRC}$ and *OK107>CG9797*-RNAi$_{TRiP}$ flies, respectively, but only the latter was significant ($P < 0.05$, $n = 3$). Knockdown of *ttk* resulted in a 30% reduction of *so* levels in both eye-antennal discs and adult heads, with the difference in adult heads being significant ($P < 0.05$, $n = 3$) (**Fig. 5e**,**f**). These results provide evidence that the observed phenotypes after RNAi-mediated knockdown of the transcription factors were likely caused by a reduction in *so* expression. Taken together with the interaction data, these results suggest that at least four out of 11 *so10* interactors identified by our Y1H system may be involved in the regulation of *so* expression *in vivo*.

DISCUSSION

The presented library is, to our knowledge, one of the most comprehensive, full-length, sequence-verified transcription factor ORF clone collections for a metazoan organism. The ORFs were cloned open-ended (without a stop codon) in the versatile Gateway system. Using this resource we developed an automated, yeast-based protein-DNA interaction detection system providing a powerful tool to deorphanize in a high-throughput manner the many functional *Drosophila* promoters and *cis*-regulatory modules for which the interacting transcription factors are still unknown. We benchmarked our system using previously characterized *cis*-regulatory modules, and identified 26% of control interactions. Although this detection rate is in the range of previously reported Y1H and Y2H data[8,17,27], we believe that this number is a conservative estimate given the absence of a high-confidence protein-DNA interaction collection comparable to the one available to validate protein-protein interaction assays[17].

We confirmed binding of the transcription factors found in the Y1H *in vitro* using MARE, which enables refinement of the identified interactions to the level of individual binding sites (C. Gubelmann, A. Isakova, A. Iagovitina, K.H., S.M. Waszak, J.-D.F. *et al.*; unpublished data). Coupled to the high-throughput Y1H system, this pipeline uniquely enabled us to identify transcription factors binding to an uncharacterized *cis*-regulatory module, and subsequently locate the specific binding site for each of these transcription factors in this element. Although we obtained a high validation rate of Y1H-detected interactions using MARE, not all detected positives showed *in vitro* site-specific binding. For example, both Y1H data and *in vivo* validation suggested that CG9797 can directly interact with the *so10* DNA bait, yet we did not recover it using MARE. This may indicate that CG9797 binding to *so10* is chromatin-dependent, showing the complementarity of both techniques.

In addition to our direct *in vitro* and *in vivo* data providing support for Y1H-detected interactions, we have indirect evidence that at least two other Y1H-observed interactions may also have biological imporance. For example, we detected binding of the homeobox transcription factor extradenticle (*exd*) to the stripe 2 enhancer of the even skipped (*eve*) gene. Although this interaction is previously unidentified for *Drosophila*, in the cricket *Gryllus bimaculatus* it has been shown that RNAi-mediated knockdown of *exd* leads to reduced *eve* expression[28], suggesting that the network regulating *eve* expression may at least be partly conserved in these insect species. A second example involves the binding of the

bZIP transcription factor slow border cells (*slbo*) to the fat body enhancer of the Yolk protein 1 (*Yp1*) gene. Although this interaction has been found by DNase I footprinting[29], it is unlikely that *slbo* regulates yolk expression *in vivo* because *slbo* is not expressed in the fat body of adult flies and yolk haemolymph levels are unchanged in *slbo* mutant flies[29]. Our Y1H screen picked up a different bZIP transcription factor, namely TJ. This transcription factor is involved in female gonad development[30] and is therefore a putative candidate to regulate *Yp1* expression *in vivo*. Together, our results indicate that the high-throughput Y1H technique described here is a useful method to uncover previously unknown interactions with putative biological importance.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturemethods/.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

We thank the members of the Lausanne genomic technologies facility for performing the Illumina sequencing, K.H. Wan for managing cDNA sequencing and transcription factor cDNA clone production, J. Reece-Hoyes and M. Walhout (University of Massachusetts Medical School, Worcester) for discussions of this work and for providing the Y1H-aS2 strain, N. Gheldof for making figures, N.W. Kelley (Biozentrum, University of Basel) for providing PWMs, S. Waszak for MARE data analysis, S. Plaza (Centre de Biologie du Développement, Université de Toulouse) for providing *so10*-GAL4 flies, and members of the TRiP at Harvard Medical School (US National Institutes of Health National Institute of General Medical Sciences R01-GM084947) and the Vienna *Drosophila* RNAi Center for providing transgenic RNAi fly stocks used in this study. This work was supported by funds from the Swiss National Science Foundation and SystemsX.ch, by a Marie Curie International Reintegration grant (BD) from the Seventh Research Framework Programme, by the Frontiers in Genetics National Centres of Competence in Research Program and by Institutional support from the Ecole Polytechnique Fédérale de Lausanne.

AUTHOR CONTRIBUTIONS

B.D. supervised the study. K.H. and B.D. designed the study. K.H. and J.B. built the transcription factor clone collection. K.H. and J.-D.F. performed Y1H screens. K.H. performed *in vivo* validations. A. Iagovitina developed image analysis software. A. Isakova performed MARE analyses. A.M. analyzed high-throughput sequencing data. P.C. provided cDNA clones and financial support. S.E.C. identified transcription factors with sequence-specific DNA-binding domains used in this study and provided transcription factor cDNA clones. K.H. and B.D. provided the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturemethods/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Adams, M.D. *et al.* The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185–2195 (2000).
2. O'Kane, C.J. & Gehring, W.J. Detection *in situ* of genomic regulatory elements in *Drosophila*. *Proc. Natl. Acad. Sci. USA* **84**, 9123–9127 (1987).
3. Zinzen, R.P., Girardot, C., Gagneur, J., Braun, M. & Furlong, E.E.M. Combinatorial binding predicts spatio-temporal *cis*-regulatory activity. *Nature* **462**, 65–70 (2009).
4. Filion, G.J. *et al.* Systematic protein location mapping reveals five principal chromatin types in *Drosophila* cells. *Cell* **143**, 212–224 (2010).
5. Bischof, J., Maeda, R.K., Hediger, M., Karch, F. & Basler, K. An optimized transgenesis system for *Drosophila* using germ-line-specific phic31 integrases. *Proc. Natl. Acad. Sci. USA* **104**, 3312–3317 (2007).
6. Stark, A. *et al.* Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. *Nature* **450**, 219–232 (2007).
7. Simicevic, J. & Deplancke, B. DNA-centered approaches to characterize regulatory protein-DNA interaction complexes. *Mol. Biosyst.* **6**, 462–468 (2010).
8. Deplancke, B. *et al.* A gene-centered *C. elegans* protein-DNA interaction network. *Cell* **125**, 1193–1205 (2006).
9. Adryan, B. & Teichmann, S.A. Flytf: A systematic review of site-specific transcription factors in the fruit fly *Drosophila melanogaster*. *Bioinformatics* **22**, 1532–1533 (2006).
10. Gallo, S.M. *et al.* Redfly v3.0: Toward a comprehensive database of transcriptional regulatory elements in *Drosophila*. *Nucleic Acids Res.* **39**, D118–D123 (2011).
11. Massouras, A., Decouttere, F., Hens, K. & Deplancke, B. Webprinses: Automated full-length clone sequence identification and verification using high-throughput sequencing data. *Nucleic Acids Res.* **38** (suppl.), W378–W384 (2010).
12. Vermeirssen, V. *et al.* Matrix and steiner-triple-system smart pooling assays for high-performance transcription regulatory network mapping. *Nat. Methods* **4**, 659–664 (2007).
13. Reece-Hoyes, J.S. *et al.* Yeast one-hybrid assays for gene-centered human gene regulatory network mapping. *Nat. Methods* doi.10.1038/nmeth.1764 (30 October 2011).
14. Reece-Hoyes, J.S. *et al.* Enhanced yeast one-hybrid assays for high-throughput gene-centered regulatory network mapping. *Nat. Methods* doi:10.1038/nmeth.1748 (30 October 2011).
15. Koegl, M. & Uetz, P. Improving yeast two-hybrid screening systems. *Brief. Funct. Genomics Proteomics* **6**, 302–312 (2007).
16. Vermeirssen, V. *et al.* Matrix and steiner-triple-system smart pooling assays for high-performance transcription regulatory network mapping. *Nat. Methods* **4**, 659–664 (2007).
17. Braun, P. *et al.* An experimentally derived confidence score for binary protein-protein interactions. *Nat. Methods* **6**, 91–97 (2009).
18. Deplancke, B., Dupuy, D., Vidal, M. & Walhout, A.J. A gateway-compatible yeast one-hybrid system. *Genome Res.* **14** 10B, 2093–2101 (2004).
19. Maerkl, S.J. & Quake, S.R. A systems approach to measuring the binding energy landscapes of transcription factors. *Science* **315**, 233–237 (2007).
20. Punzo, C., Seimiya, M., Flister, S., Gehring, W.J. & Plaza, S. Differential interactions of eyeless and twin of eyeless with the sine oculis enhancer. *Development* **129**, 625–634 (2002).
21. Czerny, T. *et al.* Twin of eyeless, a second pax-6 gene of *Drosophila*, acts upstream of eyeless in the control of eye development. *Mol. Cell* **3**, 297–307 (1999).
22. Burtis, K.C., Coschigano, K.T., Baker, B.S. & Wensink, P.C. The doublesex proteins of *Drosophila melanogaster* bind directly to a sex-specific yolk protein gene enhancer. *EMBO J.* **10**, 2577–2582 (1991).
23. Serikaku, M.A. & Otousa, J.E. Sine oculis is a homeobox gene required for *Drosophila* visual-system development. *Genetics* **138**, 1137–1150 (1994).
24. Cheyette, B.N.R. *et al.* The *Drosophila* sine oculis locus encodes a homeodomain-containing protein required for the development of the entire visual-system. *Neuron* **12**, 977–996 (1994).
25. Callaerts, P. *et al.* *Drosophila* pax-6/eyeless is essential for normal adult brain structure and function. *J. Neurobiol.* **46**, 73–88 (2001).
26. Lai, Z.C. & Li, Y. Tramtrack69 is positively and autonomously required for *Drosophila* photoreceptor development. *Genetics* **152**, 299–305 (1999).
27. Chen, Y.C., Rajagopala, S.V., Stellberger, T. & Uetz, P. Exhaustive benchmarking of the yeast two-hybrid system. *Nat. Methods* **7**, 667–668 (2010).
28. Mito, T. *et al.* Divergent and conserved roles of extradenticle in body segmentation and appendage formation, respectively, in the cricket *Gryllus bimaculatus*. *Dev. Biol.* **313**, 67–79 (2008).
29. Hutson, S.F. & Bownes, M. The regulation of yp3 expression in the *Drosophila melanogaster* fat body. *Dev. Genes Evol.* **213**, 1–8 (2003).
30. Li, M.A., Alls, J.D., Avancini, R.M., Koo, K. & Godt, D. The large maf factor traffic jam controls gonad morphogenesis in *Drosophila*. *Nat. Cell Biol.* **5**, 994–1000 (2003).

ONLINE METHODS

Clone and software availability. The *Drosophila* transcription factor ORF collection is available upon request and will eventually be distributed via Addgene (http://www.addgene.org/bart_deplancke/). TIDY is freely available for academic use and can be downloaded from http://updepla1srv1.epfl.ch/software/.

Gateway cloning of *Drosophila* transcription factors. Transcription factor ORFs were PCR-amplified using primers containing the *attB1* and *attB2* Gateway tails at the 5′ end of the forward and reverse primer, respectively (primer sequences are available in **Supplementary Table 1**). The gene-specific part of the primer was designed to have a melting temperature of ~60 °C and a G+C content close to 50%, although these parameters often had to be relaxed to find an appropriate primer. We omitted the stop codon, generating open-ended clones. The PCR was performed using iProof High-Fidelity DNA Polymerase (Bio-Rad) according to manufacturer's specifications. In a first approach we used BDGP clones as DNA template. We first compared the cDNA clone sequence with the reference sequence for each transcription factor. Clones were rejected if they contained partial ORFs, nonsense mutations, missense mutations in a known functional protein domain or more than five missense mutations in total compared to the reference sequence. Applying these criteria reduced the number of acceptable cDNA clones from 656 to 501. When no acceptable cDNA clone was retrieved, a reverse transcription–PCR strategy was adopted by extracting total RNA from whole *Drosophila* embryos, larvae or adult flies using Tri Reagent (Sigma) followed by a clean-up step using the RNeasy mini kit (Qiagen). Five micrograms of this RNA was used as an input to generate cDNA using the SuperScript III First-Strand Synthesis kit (Invitrogen). The resulting cDNA was subsequently used as a template for PCR amplification. PCR-amplified transcription factor ORFs were cloned into the pDONR221 vector using Gateway cloning by mixing 100 µg of the pDONR221 vector, 2 µl of the PCR product and 0.5 µl of BP clonase II enzyme mix (Invitrogen). After incubating for 18 h at 25 °C, this mix was transformed into competent DH5α cells and single colonies, typically four per transcription factor, were analyzed by colony PCR with M13F and M13R primers using standard protocols. The transcription factors that were successfully cloned in pDONR221 (below called transcription factor Entry clones) were then analyzed by high-throughput sequencing.

High-throughput sequencing of transcription factor clone ORFs. The transcription factor Entry clones were pooled equimolarly and subsequently fragmented using a Covaris S2 Adaptive Focused Acoustics instrument (Covaris) using the settings: duty cycle, 20%; intensity, 5; cycles per burst, 200; and time, 90 s. Five micrograms of the fragmented plasmid pool was then used for sequencing library preparation using the Illumina DNA Sample Prep kit (Illumina) according to the protocol supplied with the reagents. The sequencing library was loaded into one lane of a flow cell, sequencing clusters were generated using the Illumina Single-Read Cluster Generation Kit v2 and the flowcell sequenced on the Illumina Genome Analyzer IIx using Illumina Cycle Sequencing Kit v3 reagents according to the protocol provided by the supplier, producing 76-bp reads. The output data were processed using the Genome Analyzer Pipeline Software v1.4. The resulting file containing the short reads was submitted to the WebPrInSeS server[11] together with a file containing the reference sequences for automated assembly of the reads and evaluation of the resulting ORFs in comparison with the respective reference sequences. The transcription factor Entry clones were evaluated for sequencing coverage and quality of the assembled sequence. Clones that are fully covered by sequencing and that meet the criteria used for the evaluation of the cDNA clones described above following the BDGP convention were labeled 'gold' (588 clones or 78%). Clones of which the 5′ and 3′ were covered by sequencing (that is, standard ORFeome quality), and for which quality criteria were met, were labeled 'silver' (36 or 5%). Clones which were only partially covered by sequencing, but for which the resulting assembled sequence met the quality criteria, were labeled 'bronze' after pooling all clones that were available for a specific transcription factor (typically four) to maximize the chance of having a functional clone in this mix (68 clone mixes or 9%).

Shuttling the transcription factor ORF to Gateway compatible AD vectors. To make the transcription factor (TF) clone resource Y1H compatible, we simultaneously subcloned each accepted TF in the same Gateway reaction to both a high- and low-copy Gal4 activation domain (AD)-containing vector (pAD-Dest-2µ and AD-Dest-ARS/CEN), resulting in an equimolar mix of both AD-ORF plasmids. The former allows higher TF expression than the latter, likely resulting in increased sensitivity. We kept the low-copy plasmid, which was used previously[8], as it may allow the detection of interactions involving TFs that are toxic to the yeast when expressed at high levels. The transcription factor ORFs were subcloned by mixing 2 µl of the transcription factor entry clone, 100 ng of the pAD-Dest mix and 0.5 µl of LR clonase II enzyme mix (Invitrogen). After incubating for 18 h at 25 °C, this mix was transformed into competent DH5α cells and single colonies were analyzed by colony PCR with the AD primer and a transcription factor-specific reverse primer using standard protocols. Plasmids were isolated for all subcloned transcription factors (647 clones) and diluted to a final concentration of 100 ng μl^{-1}. The plasmid preps were checked again by PCR to verify that no arraying errors were made during preparation.

The AD transcription factor clones are ordered in a similar way as the transcription factor ORF clone collection, but in a 384-well format. For example, for the transcription factor ORF clones in row A of 96-well plates 1, 2, 3 and 4, the corresponding AD transcription factor clone would reside in respectively the uneven wells of row A, the even wells of row A, the uneven wells of row B, and the even wells of row B of the 384-well AD transcription factor plate (**Supplementary Table 1**). Some of the empty wells in the 384-well AD-transcription factor plates were filled with the original pAD-DEST vectors as negative controls or with duplicates of some transcription factor clones of specific interest, as indicated in **Supplementary Table 1**. Interactions detected twice with a specific transcription factor are reported only once in **Supplementary Table 2**.

Cloning of *cis*-regulatory modules. *Cis*-regulatory modules (CRMs) were PCR-amplified using primers containing restriction enzyme recognition sites at the 5′ end of the forward and reverse primer, respectively, and cloned in the pENTRY-5′ vector using standard restriction-ligation techniques. The CRMs were then subcloned

in the Y1H-compatible pMW2 (*HIS3*) and pMW3 (*lacZ*) vectors by Gateway LR cloning as described above. Single colonies were selected and verified by Sanger sequencing. A double integration was performed with the resulting CRM destination clones (both pMW2-CRM and pMW3-CRM in a single yeast strain) in Y1H-aS2 (with the exception of element *so10* which was integrated in the YM4271 yeast strain) using lithium acetate (LiAc)–polyethylene glycol (PEG) transformation followed by selection on a synthetic complete medium (SC) lacking histidine and uracil (–His, –Ura).

High-throughput yeast transformation. The high-throughput yeast transformation protocol is based on the regular LiAc-PEG yeast transformation protocol but volumes were decreased to allow screening in 384-well format. Briefly, 2 µl of 100 ng μl^{-1} prey plasmid, 5 µl of competent yeast and 25 µl of TE-LiAc-PEG solution were added in a well of a 384 microwell plate and resuspended by pipetting. The yeast suspension was incubated for 30 min at 30 °C and subsequently heat-shocked for exactly 20 min at 42 °C in a hot-air incubator. The yeast was pelleted by centrifugation and the supernatant was removed. The cells were resuspended in 5 µl of sterile water and 1 µl of this suspension was spotted on a SC –His, –Ura, –Trp plate. We engineered and programmed a customized robotic system (Tecan Evo) equipped with a 384-pipetting-head, incubators and a centrifuge unit to perform the complete transformation and spotting process autonomously. After growing the yeast for 3 d at 30 °C, the colonies were transferred to selective SC –His, –Ura, –Trp plates containing varying 3-amino-1,2,4-triazole (3AT) (Sigma-Aldrich) concentrations. To evaluate activation of the *lacZ* reporter, positive colonies were picked, respotted four times in 384-well format onto permissive yeast plates covered by a nitrocellulose filter to perform a *lacZ* filter assay as described[16].

As a negative control, we also subcloned the multiple cloning site (MCS) of the pEntry5′ vector into the pMW2 vector and integrated it into the yeast genome. We then transformed this DNA bait yeast strain with all *Drosophila* transcription factors as described above. We detected a single, uncharacterized ZF-C2H2 transcription factor, CG14655, which interacted with the control vector (data not shown). This may be due to binding of this transcription factor to the minimal promoter of the *HIS3* gene or other vector parts like the Gateway sites or the MCS present in this vector. Consequently, interactions involving this transcription factor with other DNA baits (for example, the strongest growing quadrant in the upper left corner of the selective plate in **Fig. 2**) were considered as false positives and omitted from **Supplementary Table 2**.

Semiautomated detection of positive interactions. Despite the fact that the transformed yeast colonies were arrayed as quadruplicates to facilitate visual detection, manual inspection can still be inconsistent and subjective. To have more objective calls, we developed an image analysis software that allows semiautomatic processing of JPEG images of the Y1H selection plates. This custom-designed tool was written in Matlab (R2008b, Mathworks) and requires an image in grayscale as input. The user then has to define the three corner colonies (bottom left, top left and bottom right) by clicking on the image. This allows normalizing and reorienting of the image according to the array of yeast colonies. A uniform grid is created to define the position of each yeast colony quadrant. If the grid positioning is not precise, the user can reject the grid and redefine the corners of the image.

The quadruplicated yeast colony pattern was detected by convoluting the image with a pattern of four bright spots on a dark background. The intensity value of the convoluted image in the center of each quadrant is used as a measure for the size of the quadrant colonies with a greater value indicating a stronger interaction. TIDY then groups the intensity values in ten clusters. We achieved the most robust detection of strong positives when we considered (i) the highest intensity value in the largest of these clusters, representing most and thus likely negative interaction yeast quadrants, as the background threshold and (ii) quadrants scoring at least 20% above this background threshold as positives. Positives that fulfilled this criterion had intensity values that typically were at least 2 s.d. above the mean or median intensity value of the plate.

To avoid detection of interactions where only one or two out of four colonies show strong growth, we also measure the intensity of individual colonies. This is done by dividing the image in 1,536 squares, each defining the limit of a single colony, and integrating the intensity over each of these squares. A uniformity coefficient is computed for each colony by subtracting half of the maximal and minimal values from the sum of four intensities and dividing this number by the mean of the four values. Therefore a number close to 3 would indicate little variation in intensity between the four colonies whereas a number greater or lower than 3 would indicate respectively lower or higher growth of one of the quadrant colonies. A second threshold based on this value is empirically set at 2.96 as we specifically wanted to eliminate quadrants whose intensity values were derived from only one or two large colonies reflecting spotty yeast growth.

The output of the program plots in green the abbreviated names of the transcription factors corresponding to the interactions scoring 20% above the background threshold. In addition, the transcription factor names are shown in a text box next to the image plot and are returned in the Matlab command line from where they can be easily copied. A plot visualizing the intensity value distribution also appeared beside the image with the intensity values on the horizontal axis and the uniformity coefficient on the vertical axis. The user can modify the area set by the default thresholds by directly clicking on this plot to evaluate the detection stringency. In some cases, this allows the inclusion of weaker interactions that clearly score above background, but below the conservative 20% threshold. The user-defined threshold is drawn in red on the plot and the newly detected interactions appear in magenta indicating their distinct status.

Finally, on some yeast plates, exterior colonies exhibit higher growth than interior ones, potentially biasing the detection threshold. We therefore included an option in TIDY that allows the user to correct for this artefact. In the case where the correction option is selected, we separate the exterior colonies from the interior ones and treat them as two separate distributions. The clustering and definition of the thresholds is done in the same way as explained earlier except that the number of clusters for the exterior distribution is set at six because of the lower number of involved quadrants.

MITOMI-based analysis of regulatory elements. MARE analysis was performed essentially as will be described (C. Gubelmann, A. Isakova, A. Iagovitina, K.H., S.M. Waszak, J.-D.F. *et al.*; unpublished data). In brief, a library of 36 bp sequences was

doi:10.1038/NMETH.1763

designed to cover the whole DNA bait so that each sequence has a 24 bp overlap with the next one in the library and each 12 bp region is covered by three different fragments. Note that the first and last region is only covered by one fragment, and the second and penultimate by two fragments. Each sequence was purchased as a single-strand oligonucleotide (Invitrogen) which served as a template for generating labeled double-stranded oligonucleotides as described[19]. transcription factors were subcloned from the Entry clones into the pMARE vector by standard Gateway cloning, fusing the eGFP coding sequence to the 3′ end of the transcription factor ORF. Subsequently, linear expression templates containing 5′ end 3′ UTR sequences and the transcription factor-eGFP fusion were generated by PCR using standard techniques. Linear expression templates were printed on top of the DNA baits on an epoxy-coated glass slide using a Qarray (Genetix) microarrayer. Microfluidics device design, fabrication, alignment and surface chemistry was as described (C. Gubelmann, A. Isakova, A. Iagovitina, K.H., S.M. Waszak, J.-D.F. *et al.*; unpublished data). Transcription factor protein was synthesized by loading TNT SP6 High-Yield wheat germ extract mixture (Promega) onto the device. MITOMI was performed and the device was imaged as described[19]. MARE data analysis was performed as will be described (C. Gubelmann, A. Isakova, A. Iagovitina, K.H., S.M. Waszak, J.-D.F. *et al.*; unpublished data). In brief, for each 12 bp region, the average signal S of the 3 fragments in which it is represented was calculated. For each 12 bp region we defined the mid position as the representative binding event position. Signal values at positions other than representative binding event positions were estimated by cubic interpolation (interp1 function, signal package, R). Specific transcription factor protein-DNA interactions were identified by clustering the signal of each position into two distinct classes, that is, specific binding positions (SBPs) and nonspecific binding positions (NSBPs), using the *k*-means clustering algorithm (function kmeans, R; settings: centers = 2, algorithm = Hartigan-Wong, nstart = 1,000). The center of the NSBP class was defined as the DNA bait-specific mean background signal (MBS). For each SBP, we defined the relative enrichment over non-specific binding as E(SBP) = S(SBP)/MBS and filtered out SBPs that have an E < 2. Specific binding regions (SBRs) were defined by joining consecutive SBPs and SBPs with the largest enrichment within a SBR were defined as the SBR maxima. Each MARE experiment was performed two times. Note that, as DNA occupancy is plotted as a relative signal normalized for the protein level in the microfluidics chamber, the scale of the *y* axis may vary between replicates. Therefore the overall trend of the DNA occupancy signal was compared between replicates. A peak was considered present in both replicates if the SBR maximum of the first replicate overlapped with the SBR of the second replicate and vice versa.

Transcription factor binding site analysis. We used the online matrix-scan tool of the regulatory sequence analysis tools (RSAT) package[31]. PWMs were from the transcription factor binding site databases Jaspar and Transfac[32,33] (**Supplementary Data**). The upper detection threshold was set at $P < 10^{-3}$.

Fly stocks. Flies were maintained at 25 °C on standard agar-cornmeal medium. *UAS*-RNAi lines were from the VDRC[34] and TRiP[35] collections and are listed in **Supplementary Table 6**. Additional fly stocks used were *OK107*, UAS-mCD8-GFP (available from the Bloomington stock center), *y,w[1118];P{attP,y[+],w[3′]}* (available from the VDRC stock center) and *so10*-GAL4.

Analysis of phenotypes. Virgin females of the UAS-RNAi lines were crossed with males of the *OK107* and *so10*-GAL4 driver lines. Adult eyes were examined using bright-field microscopy by comparing the size, overall shape and roughness of each knock-down eye to the eye of a control fly (*OK107>attP, OK107>mCD8-GFP, so10>attP* and *so10>mCD8-GFP*). Bright-field microscopy images were obtained on a Leica MZ 16 1FA stereomicroscope equipped with a DFC 480 color camera.

RNA extraction, cDNA synthesis and quantitative real-time PCR. Total RNA was isolated from 30 eye-antennal discs per genotype from wandering third-instar larvae or ten adult heads from adults overexpressing the appropriate transgene for RNAi using the Nucleospin RNA XS kit (Macherey-Nagel) according to manufacturer's specifications. First-strand cDNA synthesis was performed using the Superscript VILO cDNA synthesis kit (Invitrogen) starting from 500 ng total RNA. Primer sets were pulled from the GETPrime primer database[36] (*RpL32*, 5′-TAAGCTGTCGCACAAATGG-3′ and 5′-GGGCATCAGATACTGTCCC-3′; *so*, 5′-CTGTGTTT GCGAGGTTCTC-3′ and 5′-TTATCACATTGTGGCAGCG-3′). Quantitative real-time PCR (qRT-PCR) was performed in 384-well plates with three technical replicates on the ABI-7900HT Real-Time PCR System (Applied Biosystems) using Power SYBR Green Master Mix (Applied Biosystems) using standard procedures. *RpL32* expression levels were used as endogenous control and relative expression ratios were calculated using the ΔΔCt method with the expression levels in *OK107>mCD8-GFP* and *so10>mCD8-GFP* flies as calibrators. qRT-PCR data were derived from three independent biological replicates and *P* values were derived using a *t*-test.

31. Turatsinze, J.V., Thomas-Chollier, M., Defrance, M. & van Helden, J. Using rsat to scan genome sequences for transcription factor binding sites and cis-regulatory modules. *Nat. Protoc.* **3**, 1578–1588 (2008).
32. Bryne, J.C. *et al.* Jaspar, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res.* **36**, D102–D106 (2008).
33. Matys, V. *et al.* Transfac and its module transcompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* **34**, D108–D110 (2006).
34. Dietzl, G. *et al.* A genome-wide transgenic rnai library for conditional gene inactivation in *Drosophila*. *Nature* **448**, 151–156 (2007).
35. Ni, J.Q. *et al.* A *Drosophila* resource of transgenic rnai lines for neurogenetics. *Genetics* **182**, 1089–1100 (2009).
36. Gubelmann, C. *et al.* Getprime: A gene- or transcript-specific primer database for quantitative real-time PCR. *Database* doi:10.1093/database/bar040 (2011).