Yeast one-hybrid assays for gene-centered human gene regulatory network mapping

John S Reece-Hoyes[1–3], A Rasim Barutcu[2,4], Rachel Patton McCord[1,2,4], Jun Seop Jeong[5–10], Lizhi Jiang[5–10], Andrew MacWilliams[11,12], Xinping Yang[11,12], Kourosh Salehi-Ashtiani[11,12], David E Hill[11,12], Seth Blackshaw[5–10], Heng Zhu[5–10], Job Dekker[1,2,4,11] & Albertha J M Walhout[1–3,11]

Gateway-compatible yeast one-hybrid (Y1H) assays provide a convenient gene-centered (DNA to protein) approach to identify transcription factors that can bind a DNA sequence of interest. We present Y1H resources, including clones for 988 of 1,434 (69%) predicted human transcription factors, that can be used to detect both known and new interactions between human DNA regions and transcription factors.

Interactions between regulatory genomic DNA and transcription factors provide the first level of gene control and, therefore, the backbone of gene regulatory networks. Two complementary types of approaches can be used to identify such interactions[1]. Transcription factor–centered or protein-to-DNA approaches (for example, chromatin immunoprecipitation (ChIP)) reveal genomic regions bound by a transcription factor. Gene-centered or DNA-to-protein approaches (for example, yeast one-hybrid (Y1H)), in contrast, define the repertoire of transcription factors that can bind a DNA fragment of interest. The advantages and disadvantages of these techniques have been discussed elsewhere[2–4]. In Y1H assays, a target DNA sequence ('DNA bait') is cloned upstream of two reporter genes (*HIS3* and *LacZ)* to generate two DNA bait–reporter constructs[5]. After integration of these constructs into the yeast genome to generate a 'DNA bait strain', interacting transcription factors ('protein preys') can be identified either by screening complex cDNA or transcription factor mini-libraries or by testing individual protein preys in a directed pairwise manner[5,6]. Activation of the *HIS3* reporter permits growth on medium lacking histidine and containing 3-amino-1,2,4-triazole (3AT), a competitive inhibitor of the His3 enzyme, whereas activation of the *LacZ* reporter is detected by a colorimetric assay in which beta-galactosidase turns 5-bromo-4-chloro-3-indolyl-β-D-galactoside (X-gal) into a blue compound.

We have previously combined Y1H assays with Gateway cloning to transfer multiple DNA baits in parallel into the two Y1H reporter vectors[5] and have applied these assays to delineate *Caenorhabditis elegans* gene regulatory networks[7–10] and to screen *Arabidopsis thaliana* gene promoters[11]. In an accompanying paper, we describe the development of a *C. elegans* enhanced Y1H (eY1H) pipeline[4]. For eY1H, a robotic setup is used with an arrayed collection of yeast strains expressing transcription factor preys that can be mated with a DNA bait strain.

Currently, to our knowledge no gene-centered assays are available to identify human DNA–transcription factor interactions in a high-throughput manner. Here we present a resource of human transcription factor–encoding open reading frames (ORFs) fused to the sequence encoding the Gal4 activation domain (Gal4-AD) and applied this collection to several Y1H configurations, including the high-throughput eY1H pipeline.

The human genome encodes 1,434 regulatory transcription factors, 1,116 of which are currently available in large clone collections[12,13] (Online Methods and **Supplementary Table 1**). We transferred these ORFs to the AD-2μ Y1H prey vector by Gateway cloning and, after sequence verification, obtained 988 full-length transcription factor prey clones (**Fig. 1a** and **Supplementary Table 1**). These clones can be transformed directly into DNA bait strains for haploid-based Y1H experiments[6]. We transformed these clones into the Y1H prey strain to generate a human transcription factor yeast array (**Supplementary Table 2**) that can be used in small-scale mating-based Y1H experiments as well as in eY1H assays[4]. We also added 236 clones for unconventional DNA-binding proteins[13] (**Supplementary Tables 2** and **3**).

To test the use of Y1H assays for the identification of human DNA-transcription factor interactions, we first generated a small positive reference set (PRS) via literature curation (**Supplementary Table 4**). We predominantly focused on the well-studied beta-globin locus, but also included a few other regulatory regions and gene promoters (**Supplementary Table 5**). We tested each of the PRS

[1]Program in Systems Biology, University of Massachusetts Medical School, Worcester, Massachusetts, USA. [2]Program in Gene Function and Expression, University of Massachusetts Medical School, Worcester, Massachusetts, USA. [3]Program in Molecular Medicine, University of Massachusetts Medical School, Worcester, Massachusetts, USA. [4]Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, Worcester, Massachusetts, USA. [5]Department of Pharmacology and Molecular Sciences, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA. [6]Solomon H. Snyder Department of Neuroscience, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA. [7]Department of Neurology, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA. [8]Department of Ophthalmology, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA. [9]Center for High-Throughput Biology, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA. [10]Institute for Cell Engineering, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA. [11]Center for Cancer Systems Biology and Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA. [12]Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA. Correspondence should be addressed to A.J.M.W. (marian.walhout@umassmed.edu) or J.S.R.-H. (john.reece-hoyes@umassmed.edu).

RECEIVED 20 JUNE; ACCEPTED 22 AUGUST; PUBLISHED ONLINE 30 OCTOBER 2011; DOI:10.1038/NMETH.1764

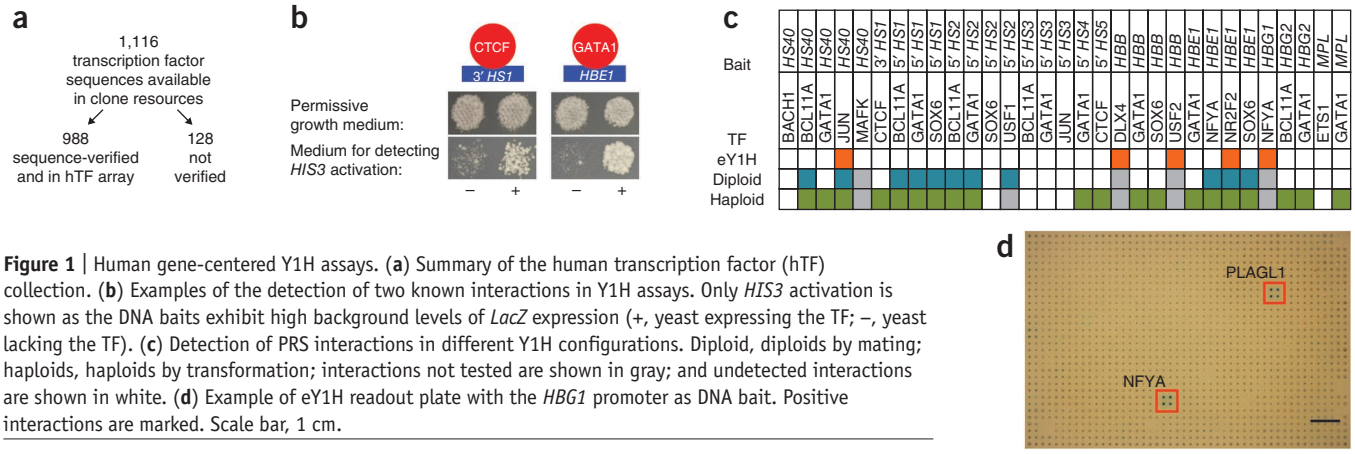

Figure 1 | Human gene-centered Y1H assays. (**a**) Summary of the human transcription factor (hTF) collection. (**b**) Examples of the detection of two known interactions in Y1H assays. Only *HIS3* activation is shown as the DNA baits exhibit high background levels of *LacZ* expression (+, yeast expressing the TF; –, yeast lacking the TF). (**c**) Detection of PRS interactions in different Y1H configurations. Diploid, diploids by mating; haploids, haploids by transformation; interactions not tested are shown in gray; and undetected interactions are shown in white. (**d**) Example of eY1H readout plate with the *HBG1* promoter as DNA bait. Positive interactions are marked. Scale bar, 1 cm.

interactions in different types of Y1H assays, in haploids and diploids, at different readout times and under different Y1H conditions, because it has been demonstrated in other yeast-based assays that varying assay format and conditions results in a more comprehensive dataset[14]. We detected 24 of 31 known interactions (77%; **Fig. 1b**,**c** and **Supplementary Table 4**), with five of these interactions detected in eY1H assays (16%) (**Fig. 1d**). Although this PRS detection rate for eY1H assays was comparable to that observed for high-throughput yeast two-hybrid screens[15], this result contrasts to that observed with *C. elegans* bait strains for which eY1H is at least as sensitive as the other Y1H methodologies[4]. This disparity is likely due to the fact the DNA baits used in this study had higher background compared to most *C. elegans* baits[4]. We note that human DNA baits generated for other studies do not show this high background (data not shown), suggesting that this is not a systematic problem of Y1H assays with human DNA. All of the interactions detected by Y1H approaches other than eY1H were observed with only the *HIS3* reporter gene because of high background *LacZ* expression (data not shown). Essentially, the performance of various Y1H approaches is intrinsically linked to bait-strain behavior, and for optimal results the user should modulate the experimental settings accordingly (**Supplementary Fig. 1**).

In human eY1H assays, in which we screened 14 baits against the entire collection, we detected 175 DNA-protein interactions involving 13 DNA baits and 100 proteins (**Supplementary Table 6**). The proteins detected include 95 human transcription factors (~10% of the 988 tested) and five unconventional DNA-binding proteins (~2% of the 236 tested)[13]. The eY1H interactions did not exhibit a major bias for or against a particular type of DNA-binding domain (**Fig. 2a**), complementing our observations in *C. elegans* experiments[7]. We found a larger proportion of nuclear hormone receptors, however, with most of these exclusively interacting with the *CSF1* promoter (**Supplementary Table 6**), suggesting this enrichment is likely due to the small sample size of DNA baits.

Validating DNA–transcription factor interactions in complex systems is challenging, and a 'true negative' is nearly impossible to demonstrate[3]. Whereas several of the interactions detected with eY1H were part of the PRS and thus are known to have *in vivo* relevance, we wanted to assess the overall quality of the eY1H dataset. To this end, we evaluated the relationship between transcription

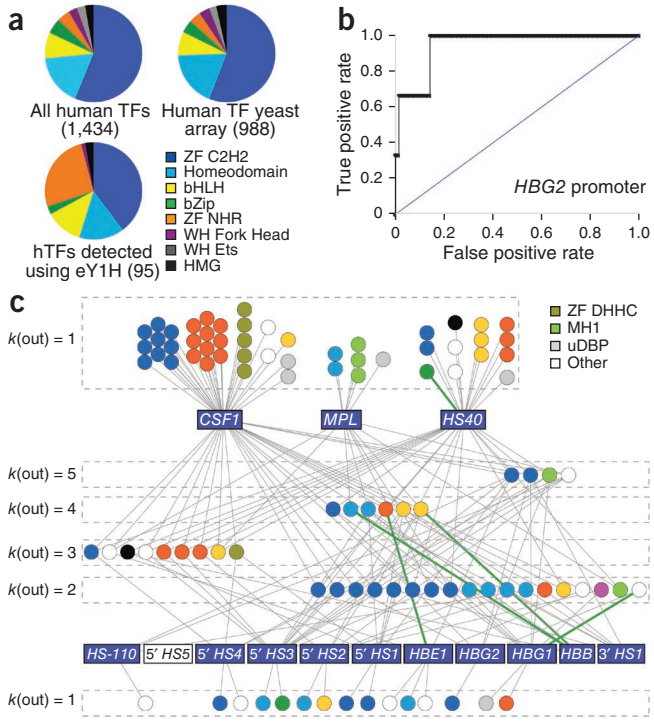

Figure 2 | eY1H data analysis. (**a**) DNA binding domain analysis of the transcription factor (TF) compendium, the transcription factor prey yeast array, and the transcription factors detected in Y1H interactions. ZF, zinc finger; C2H2, Cys-Cys-His-His; bHLH, basic region helix-loop-helix; bZip, basic leucine zipper; NHR, nuclear hormone receptor; WH, winged helix; HMG, high mobility group. (**b**) Receiver operating characteristic (ROC) curve of DNA binding site analysis for the *HBG2* promoter. Binding sites in the DNA bait sequence were ranked from best to worst (in terms of match to position weight matrix) along the *x* axis; only the 'best' binding site match for each transcription factor was used. As the curve progresses along the *x* axis, it steps up only for binding sites of transcription factors detected by eY1H. If the binding sites provided no information regarding eY1H interactions (that is, no match between interactions predicted and detected), the curve would be largely below the diagonal. (**c**) Transcription factor–DNA interactions detected by eY1H depicted in a gene regulatory network. The DNA bait 5′ *HS5* is depicted as a clear box because it had no interactions. Unless otherwise noted, colors indicate transcription factor families as in **a**. MH1, MAD homology 1 domain; DHHC, Asp-His-His-Cys; and uDBP, unconventional DNA-binding protein. The outgoing degree *k*(out) (number of DNA baits bound per transcription factor) is indicated. Green edges indicate detected PRS interactions.

factor interactions observed in eY1H assays and their reported DNA binding sites. We first compiled DNA-binding specificity information for human transcription factors or their orthologs[16] (Online Methods). Based on the potential transcription factor binding sites in each DNA bait, we predicted which factors are expected to bind. Then we compared our experimentally detected factors to these predictions. We generated a receiver operating characteristic curve for each DNA-bait sequence (**Fig. 2b** and **Supplementary Fig. 2**) and calculated the area under each curve (AUC). We found significant enrichment for transcription factor binding sites in five of 11 DNA baits (AUC > 0.5, $P < 0.05$; **Supplementary Table 7**). This compares favorably to a similar analysis for ChIP-sequencing (ChIP-seq) data (**Supplementary Tables 8** and **9**), which suggests that eY1H assays may be more likely to capture direct physical interactions between DNA and transcription factors than ChIP. Two DNA baits (the *MPL* and *HBB* promoters) did not exhibit a correlation between the predicted transcription factor binding sites they contain and the transcription factors retrieved in eY1H assays. This could be because these factors require interactions with cofactors and so are missed in eY1H or because of differences in binding sites between human transcription factors and their orthologs. A more likely explanation is the high background reporter gene expression we observed for both these DNA baits, which makes them difficult to assay.

We visualized all eY1H interactions using Cytoscape[17], generating to our knowledge the first gene-centered human gene regulatory network (**Fig. 2c** and **Supplementary Fig. 3**). Although it is small in size, we can already observe both specific as well as more promiscuous transcription factors. For instance, four factors each bind five DNA baits, whereas the majority of factors only interact with a single DNA bait. We also find several instances in which multiple members of a transcription factor family interacted with the same DNA sequence. For instance, all four nuclear factor 1 transcription factors (NFIA, NFIB, NFIC and NFIX) interacted with the *MPL* promoter. This observation could reflect that these transcription factors have similar DNA binding specificities and may be relevant in different cells or tissues, or under varying physiological conditions.

In summary, we developed a collection of human transcription factor prey clones and a human transcription factor yeast array, and combined these resources with our newly developed eY1H platform[4], facilitating the mapping of human gene-centered regulatory networks. The human eY1H pipeline will be a powerful complement to transcription factor–centered methods, by enabling large-scale characterization of the DNA-binding activity of transcription factors that may be expressed or active only under restricted conditions or in a few cells. However, these resources can also easily be used for mating or direct DNA transformations of one or a few human DNA baits in small-scale studies.

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 members of the Walhout lab for discussions and critical reading of the manuscript. Research in the Walhout lab is supported by US National Institutes of Health (NIH) grants DK068429 and GM082971. This work was supported by NIH grant HG003143 and a W.M. Keck Foundation Distinguished Young scholar award to J.D., an Ellison Foundation grant and Dana Farber Cancer Institute Sponsored Research funds awarded to Center for Cancer Systems Biology. Research in the Blackshaw lab is funded by a W.M. Keck Foundation Distinguished Young scholar award and a grant from the Ruth and Milton Steinbach Fund. H.Z. is supported by NIH grant GM076102.

AUTHOR CONTRIBUTIONS

A.J.M.W. conceived the project; J.D., J.S.R.-H. and A.J.M.W. designed the project; J.S.R.-H. and A.R.B. performed the experiments; J.S.R.-H. and A.J.M.W. analyzed the data; J.S.J., L.J. and A.M. picked transcription factor ORF clones; J.S.J. and L.J. assisted A.R.B. and J.S.R.-H. with Gateway cloning. R.P.M. performed the binding-site analysis; H.Z., S.B., K.S.-A., X.Y., A.M. and D.E.H. provided transcription factor ORFeome clones; J.S.R.-H. and A.J.M.W. wrote 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. Walhout, A.J.M. *Genome Res.* **16**, 1445–1454 (2006).
2. Arda, H.E. & Walhout, A.J.M. *Brief. Funct. Genomics* **9**, 4–12 (2009).
3. Walhout, A.J.M. *Genome Biol.* **12**, 109 (2011).
4. Reece-Hoyes, J.S. *et al. Nat. Methods* doi:10.1038/nmeth.1748 (this issue).
5. Deplancke, B., Dupuy, D., Vidal, M. & Walhout, A.J.M.A. *Genome Res.* **14**, 2093–2101 (2004).
6. Vermeirssen, V. *et al. Nat. Methods* **4**, 659–664 (2007).
7. Deplancke, B. *et al. Cell* **125**, 1193–1205 (2006).
8. Vermeirssen, V. *et al. Genome Res.* **17**, 1061–1071 (2007).
9. Martinez, N.J. *et al. Genes Dev.* **22**, 2535–2549 (2008).
10. Arda, H.E. *et al. Mol. Syst. Biol.* **6**, 367 (2010).
11. Brady, S. *et al. Mol. Syst. Biol.* **7**, 459 (2011).
12. Lamesch, P. *et al. Genomics* **89**, 307–315 (2007).
13. Hu, S. *et al. Cell* **139**, 610–622 (2009).
14. Chen, Y.C., Rajagopala, S.V., Stellberger, T. & Uetz, P. *Nat. Methods* **7**, 667–668 (2010).
15. Braun, P. *et al. Nat. Methods* **6**, 91–97 (2009).
16. Badis, G. *et al. Science* **324**, 1720–1723 (2009).
17. Shannon, P. *et al. Genome Res.* **13**, 2498–2504 (2003).

ONLINE METHODS

Human transcription factor–encoding genes and ORF clones. All sequence-verified clones are available upon request. We define regulatory human transcription factors as proteins that have a predicted sequence-specific DNA-binding domain[18]. We considered two primary curated catalogs of transcription factor annotations[19,20] to generate a nonredundant compendium of 1,405 transcription factors. Interpro DNA-binding domain (DBD) identifiers were incorporated into the list using the InterPro DBD database. These were supplemented with 29 homologs of *C. elegans* transcription factors[18] that have types of DNA-binding domains that were missing in the first two collections. These include transcription factors with the following domains or annotations: PUR (3 factors), RPEL (5 factors), WT1 (5 factors), YL1 (1 factor), ZF-A20 (7 factors) and ZF-DHHC (8 factors). Redundant genes and annotated pseudogenes were removed from the list. The final list of 1,434 transcription factor genes is available in **Supplementary Table 1**. For each transcription factor, HUGO Gene Nomenclature Committee (HGNC) numbers, gene name aliases, Ensembl gene identifiers and Entrez gene identifiers incorporated using the HGNC webpage as of 17 August 2010 are included.

Human genes often encode multiple splice variants. For the first human Y1H resource, we only considered a single variant per transcription factor because ORFeome collections usually only contain a single variant and because different variants often have the same DNA-binding domain[12,13]. We picked transcription factor–encoding ORFs from two ORFeome collections[12,13] and transferred the ORFs to a Gateway Destination AD-2μ vector (Life Technologies) by a Gateway LR reaction as described previously[21]. We verified 988 AD-2μ clones by sequencing and transformed them into the Yα1867 Y1H prey strain[4] to create the human transcription factor yeast array (**Supplementary Table 1**). Another 29 clones were included in the array although they could not be confirmed for various reasons: some were incorrect, and for others a clear sequence could not be obtained (**Supplementary Table 1**, labeled as "maybe" in column 1). We will continue our attempts to obtain sequence-verified clones for all 1,116 available transcription factors and will consider *ab initio* cloning additional ORFs that are not yet available.

Generation of DNA baits. DNA baits were obtained by PCR amplification using either genomic DNA from K562 cells, or the bacterial artificial chromosome (BAC) CTD-264317, as a template. PCR amplicons were Gateway BP–cloned into pDONR-P4-P1R as described previously[22]. Primer sequences are listed in **Supplementary Table 5**. DNA baits were subsequently transferred to the pMW#2 and pMW#3 Y1H reporter Destination vectors that carry the *HIS3* and *LacZ* reporter genes, respectively[7,22]. DNA bait–*HIS3* and DNA bait–*LacZ* constructs were linearized and simultaneously integrated into the genome of the Y1H-aS2 yeast strain[4]. Up to 12 independent integrants were examined for autoactivation of the two reporter genes[22]. Integrants with lowest auto-activity were selected for Y1H assays.

Yeast one-hybrid assays. In eY1H assays, each DNA–transcription factor interaction is tested in quadruplicate in a 1,536-colony format using a robotic platform by plating diploid yeast on medium containing both X-gal and 3AT; only yeast in which both *HIS3* and *LacZ* are activated by transcription factor binding to the DNA bait will turn blue[4]. The human transcription factor prey yeast array consists of five plates of 1,536 colonies, each containing up to 380 transcription factors in quadruplicate (there were empty spaces available in the array; **Supplementary Table 2**). We performed eY1H assays using the 14 DNA bait strains (**Supplementary Table 5**). eY1H assays were performed as described in the accompanying paper[4] with the exception that the human eY1H assay plates were evaluated for interacting transcription factors on a daily basis on readout plates containing 5 mM, 10 mM or 20 mM 3AT rather than at a fixed 1-week time point with only 5 mM 3AT. Traditional Y1H assays were performed as described previously[6] with diploid or transformed haploid yeast plated on 1 mM, 3 mM, 5 mM, 10 mM, 20 mM and 40 mM 3AT and evaluated twice daily for up to 7 d. Each DNA bait was screened twice, and for 12 of 14 DNA baits two independent integrant yeast strains were tested.

Positive reference set of literature-curated human transcription factor-DNA interactions. The PRS list is available in **Supplementary Table 4**. Interactions were only considered when they were detected *in vivo* and when the interaction was presented in the figures of the original publication. The majority of interactions in the PRS are from experiments that targeted a DNA region of interest (double-stranded DNA oligo shifted using extract from cultured human cells then supershifted with transcription factor–specific antibody or ChIP from cultured human cells). However, some were from publications that performed genome-wide ChIP-microarray anlaysis (ChIP-chip) studies that then presented interactions for a DNA bait(s) in this study. Interactions detected by ChIP-seq efforts of the Encyclopedia of DNA elements (ENCODE) Consortium were not considered, as we used these interactions in the DNA binding site analysis.

DNA binding site analysis. We evaluated whether transcription factors that interact with a DNA bait either in eY1H (**Supplementary Table 6**) or ChIP (**Supplementary Table 8**) assays are enriched among those transcription factors that are *ab initio* predicted to bind according to the putative binding sites in the DNA bait.

We compiled previously published position weight matrices (PWMs) for human transcription factors or their orthologs (in **Supplementary Tables 6** and **8** indicated by a "Y" in the column "PWM available"). These PWMs convey sequence specificity information and were either derived from protein binding microarray (PBM) data[23] or obtained from the 'core' set of transcription factors in the curated Jaspar database[24]. Of the 100 transcription factors detected by eY1H screens, 39 had PWMs available. Of the 79 transcription factors assayed by ChIP in the ENCODE project[25], 47 had peaks that overlapped with one or more of the DNA baits used in our study, and of these, 18 had PWMs available.

To predict the likelihood of a transcription factor binding to a given DNA bait sequence, we used the generalizable occupancy model of expression regulation (GOMER) scoring framework, a physically principled approach that is used to calculate the binding probability of a transcription factor over the entire length of a sequence according to that transcription factor's PWM[26]. PWMs were downloaded from Jaspar or

Uniprobe and trimmed from both the 5′ and 3′ ends until a position was reached that exceeded an information content threshold of 0.3, a trimming approach previously found to be effective when scoring sequences by PWMs from PBM data[25]. After GOMER scoring of each bait with each PWM, the same PWM was used to score 800 additional human promoter or DNase I–hypersensitive site sequences that were length and G+C content matched for each bait. The GOMER scores for each bait-transcription factor pair were then normalized by dividing by the median GOMER score for this transcription factor across the 800 G+C content– and length-matched sequences. This allowed Gomer scores derived from PWMs with varying properties to be compared fairly. A receiver operating characteristic (ROC) curve was then generated for each bait sequence by comparing the Gomer ranks of transcription factors that interacted with the bait (foreground set) to the ranks of the non-interacting transcription factors (background set). The area under the curve (AUC) was then calculated for each ROC curve. Randomly permuting the assignments of transcription factors to foreground (interacting with a given bait) and background (not interacting with a given bait) sets 1000 times allowed the calculation of a *P* value for each AUC. Interacting transcription factors were considered significantly enriched among predicted interactors for each bait if the AUC was above 0.5 with $P < 0.05$. In **Supplementary Tables 7** and **9** we show the results of this analysis using interactions detected by eY1H and ChIP, respectively. For the eY1H assay analysis, the enrichment results were strongest when we scored only the 500 base pairs of the bait sequence that was closest to the reporter gene, though most significant enrichments (AUC > 0.5, $P < 0.05$) were also present when the full-length sequence was considered (data not shown).

18. Reece-Hoyes, J.S. *et al. Genome Biol.* **6**, R110 (2005).
19. Vaquerizas, J.M., Kummerfeld, S.K., Teichmann, S.A. & Luscombe, N.M. *Nat. Rev. Genet.* **10**, 252–263 (2009).
20. Fulton, D.L. *et al. Genome Biol.* **10**, R29 (2009).
21. Walhout, A.J.M. *et al. Methods Enzymol* **328**, 575–592 (2000).
22. Deplancke, B., Vermeirssen, V., Arda, H.E., Martinez, N.J. & Walhout, A.J.M. *CSH Protocols* doi:10.1101/pdb.prot4590 (2006).
23. Newburger, D.E. & Bulyk, M.L. *Nucleic Acids Res.* **37**, D77–D82 (2009).
24. Bryne, J.C. *et al. Nucleic Acids Res.* **36**, D102–D106 (2008).
25. ENCODE Project consortium. *Nature* **447**, 799–816 (2007).
26. Granek, J.A. & Clarke, N.D. *Genome Biol.* **6**, R87 (2005).

doi:10.1038/nmeth.1764