Next-generation sequencing to generate interactome datasets

Haiyuan Yu[1–3], Leah Tardivo[1,2,6], Stanley Tam[1,2,6], Evan Weiner[1,2,5], Fana Gebreab[1,2], Changyu Fan[1,2], Nenad Svrzikapa[1,2], Tomoko Hirozane-Kishikawa[1,2], Edward Rietman[1,2], Xinping Yang[1,2], Julie Sahalie[1,2], Kourosh Salehi-Ashtiani[1,2,5], Tong Hao[1,2], Michael E Cusick[1,2], David E Hill[1,2], Frederick P Roth[1,4,5], Pascal Braun[1,2] & Marc Vidal[1,2]

Next-generation sequencing has not been applied to protein-protein interactome network mapping so far because the association between the members of each interacting pair would not be maintained in *en masse* sequencing. We describe a massively parallel interactome-mapping pipeline, Stitch-seq, that combines PCR stitching with next-generation sequencing and used it to generate a new human interactome dataset. Stitch-seq is applicable to various interaction assays and should help expand interactome network mapping.

At any time hundreds of thousands of macromolecular interactions occur in a cell, mediating functions that maintain normal cellular activities. High-throughput approaches have been developed to determine interactions in many organisms at large scale. Current high-throughput protein-protein 'interactome' datasets are of high quality but have low coverage[1,2]. For humans, more than 95% of the interactome remains to be mapped[1].

A bottleneck for high-throughput interactome-mapping methods, such as yeast one-[3], two-[4] and three-hybrid[5] systems, is determining the identities of the interacting protein, DNA or RNA molecules. Implementation of next-generation DNA sequencing technologies[6–8], as opposed to Sanger technology, would substantially increase throughput and decrease cost. Although highly effective for genome and transcriptome 'shotgun' sequencing, next-generation DNA sequencing technologies are not readily applicable for identification of interacting pairs. The necessary pooling of PCR amplicons in the preparation of interacting sequence tags (ISTs) (**Fig. 1a**) would inevitably eliminate the association in each pair of DNA sequences coding for interacting molecules.

Here we describe a massively parallel interactome-mapping strategy that incorporates next-generation DNA sequencing (**Fig. 1a**) and test the strategy in a high-throughput yeast two-hybrid (Y2H) system. This general scheme can be readily extended to increase throughput and decrease cost for other interactome-mapping methods, particularly other binary protein-protein interaction assays[1], yeast one-hybrid[3] or genetic screens in which pairs of DNA molecules are selected and identified[9].

In current protocols of high-throughput Y2H screens, the open reading frames (ORFs) or cDNAs encoding selected pairs of interacting hybrid proteins (*X* fused to a DNA-binding domain (DB-*X*) and *Y* fused to an activation domain (AD-*Y*)) are amplified directly from yeast transformants and subsequently identified by Sanger DNA sequencing[4] (**Supplementary Fig. 1**). As *X* and *Y* originate from recorded positions in paired PCR plates, they can be computationally reassembled to form pairs of ISTs[10].

The first step of our methodology, termed Stitch-seq, is PCR stitching, which places a pair of sequences encoding interacting proteins on the same PCR amplicon[11]. PCR stitching consists of two rounds of PCR (**Fig. 1b**). In the first round, *X* and *Y* (present on the Y2H DB-*X* and AD-*Y* vectors) are amplified with DB- and AD-vector-specific upstream primers, respectively (**Supplementary**

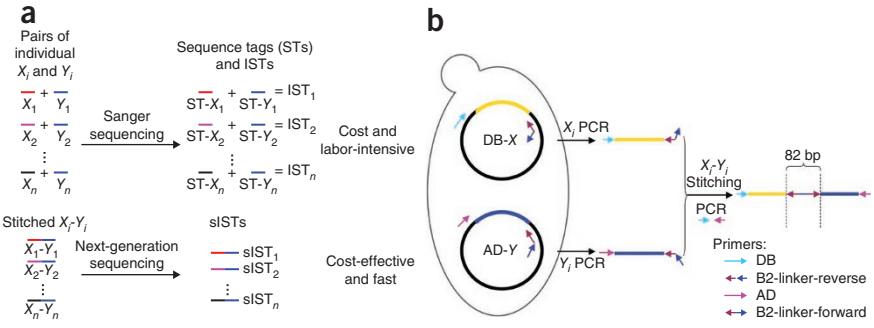

Figure 1 | Stitch-seq interactome mapping. (**a**) Outline of interactome mapping using different sequencing technologies. Each DNA fragment in each interacting pair is PCR-amplified individually and Sanger-sequenced; the association is tracked via position on the plate (top). Or each pair of DNA fragments is placed on the same PCR amplicon by PCR stitching; the amplicons are then collected and subjected to next-generation sequencing (bottom). (**b**) Outline of a PCR-stitching experiment.

[1]Center for Cancer Systems Biology (CCSB), Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA. [2]Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA. [3]Department of Biological Statistics and Computational Biology and Weill Institute for Cell and Molecular Biology, Cornell University, Ithaca, New York, USA. [4]Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, Massachusetts, USA. [5]Present addresses: Weill Cornell Graduate School of Medical Sciences, New York, New York, USA (E.W.), New York University Abu Dhabi, Abu Dhabi, United Arab Emirates and Center for Genomics and Systems Biology, Department of Biology, New York University, New York, New York, USA (K.S.-A.), and Donnelly Centre for Cellular and Biomolecular Research, University of Toronto and Samuel Lunenfeld Research Institute, Mt. Sinai Hospital, Toronto, Ontario, Canada (F.P.R.). [6]These authors contributed equally to this work. Correspondence should be addressed to P.B. (pascal_braun@dfci.harvard.edu) or M.V. (marc_vidal@dfci.harvard.edu).

RECEIVED 13 DECEMBER 2010; ACCEPTED 3 MARCH 2011; PUBLISHED ONLINE 24 APRIL 2011; DOI:10.1038/NMETH.1597

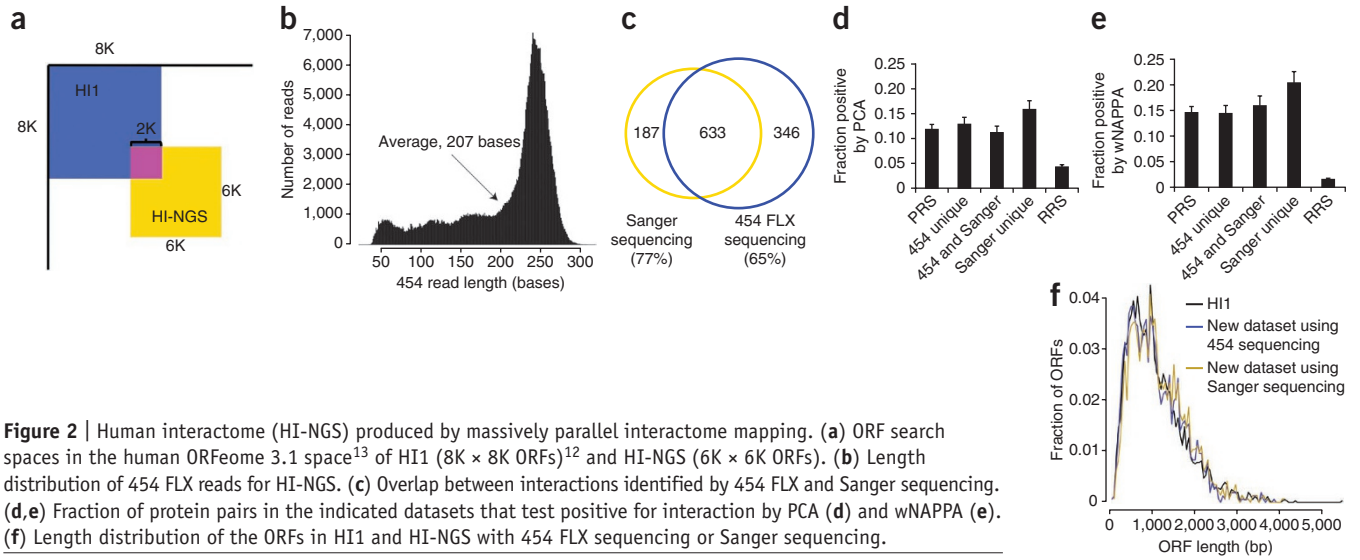

Figure 2 | Human interactome (HI-NGS) produced by massively parallel interactome mapping. (**a**) ORF search spaces in the human ORFeome 3.1 space[13] of HI1 (8K × 8K ORFs)[12] and HI-NGS (6K × 6K ORFs). (**b**) Length distribution of 454 FLX reads for HI-NGS. (**c**) Overlap between interactions identified by 454 FLX and Sanger sequencing. (**d**,**e**) Fraction of protein pairs in the indicated datasets that test positive for interaction by PCA (**d**) and wNAPPA (**e**). (**f**) Length distribution of the ORFs in HI1 and HI-NGS with 454 FLX sequencing or Sanger sequencing.

Table 1a). A common sequence of the downstream primers is complementary to the Gateway-specific *attB2* site immediately after the *X* and *Y* ORFs. We tested the PCR-stitching concept for Y2H experiments using Gateway clones, though the approach can be generalized to other interactome-mapping assays with different vectors. In the second round of PCR (**Supplementary Table 1b**), we used *X* and *Y* amplicons from the first round as templates to produce a concatenated PCR product composed of *X* and *Y* ORFs connected by an 82-bp linker sequence (**Fig. 1b**). Then we pooled all PCR products and sequenced them by next-generation DNA sequencing to produce stitched ISTs (sISTs).

Concatenated PCR products should, on average, be twice the length of single ORFs (**Fig. 1b**). To test the length limit of PCR stitching, we chose four DB-*X* and four AD-*Y* constructs of various ORF lengths: 500 bp, 1 kb, 2 kb and 3 kb (**Supplementary Fig. 2a**). As expected, in the first-step colony PCRs all eight ORFs were amplified (**Supplementary Fig. 2b**). In second-step PCRs, we tested all 16 possible combinations, with the longest combination (A4–D4) being greater than 6 kb. Concatenated ORF pairs up to 6 kb in total length were generated efficiently and accurately (**Supplementary Fig. 2c**).

We next applied PCR stitching to pairs of ORFs identified from a Y2H screen aimed at expanding the human interactome map[12]. After Y2H screening of a 6,000 (6K) by 6K ORF search space of human ORFs in the ORFeome 3.1 set[13] (**Fig. 2a**) with two rounds of phenotype testing, we selected ~5,200 (interaction-positive) colonies. PCR stitching applied to these colonies produced ~5,000 stitched PCR amplicons. We sequenced stitched amplicons with the 454 FLX platform[7], producing ~400,000 reads (**Table 1**). The average read length was 207 bases (**Fig. 2b**), which is 125 bases longer than the 82-bp linker sequence, so that many reads could unambiguously identify pairs of unique *X* and *Y* ORFs, thereby generating sISTs. To identify ORFs encoding pairs of interacting proteins, we selected reads that contained the linker sequence (~10%) and also covered at least 15 bases of ORF-specific sequences on both ends of the linker. After matching these sequences to human ORFeome v3.1 (ref. 13) we identified 2,089 unique sISTs.

We experimentally retested by pairwise Y2H all sISTs starting from fresh yeast transformants stored in our collection and confirmed 1,318 pairs of ORFs as demonstrably encoding Y2H interacting proteins (**Table 1**). Because the collection contains multiple ORFs for some genes (for example, splice isoforms), the final tally was 979 interactions among proteins encoded by 997 genes (**Table 1**). This confirmation rate is almost identical to that previously described for Y2H screens using Sanger sequencing[12]. Furthermore, the confirmation rate did not vary between sISTs discovered uniquely and sISTs discovered multiple times (**Supplementary Note 1** and **Supplementary Fig. 3**).

For comparison we also sequenced all of the ~5,200 interaction-positive colonies individually by Sanger sequencing and identified 820 interactions among proteins encoded by 914 genes (**Fig. 2c** and **Table 1**). Of these, we also identified 633 interactions by 454 FLX sequencing. This overlap is higher than the expected overlap of ~70% (**Supplementary Fig. 4**), even taking into account a ~5% failure rate of PCR and Sanger sequencing reactions. We detected 19% more interactions using our Stitch-seq strategy than using Sanger sequencing, but that was probably because of the higher coverage of the 454 FLX sequencing and the inherent failure rate of Sanger sequencing.

We next quantitatively evaluated the quality of this interactome dataset based on orthogonal interaction assays[1,14]. We selected 94 protein pairs at random from all verified interactions that we identified by (i) only 454 FLX sequencing ('454 unique'); (ii) only Sanger sequencing ('Sanger unique'); or (iii) both ('454 and Sanger') (**Fig. 2c**). We combined these 282 interactions with positive reference set (PRS) and random reference set (RRS) interactions consisting of 92 interactions each[2,14], to benchmark assay performance[14]. We tested

Table 1 | Interactome mapping and IST-sIST identification

	Sanger	454 FLX
Search space (pairs of ORFs)	1.8×10^7	
Colonies	~5,200	
PCRs	~10,400	~15,600
Reads	~8,840	395,873
Reads with linker	NA	39,211
ISTs or sISTs	~8,840	18,853
Candidate interactions (pairs of ORFs)	1,602	2,089
Verified interactions (pairs of ORFs)	1,032	1,318
Verified interactions (pairs of genes)	820	979
Total verified interactions (pairs of genes)	1,166	

NA, not applicable.

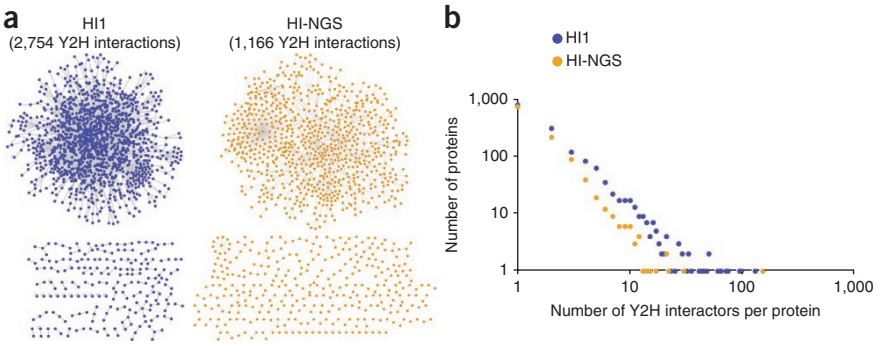

Figure 3 | HI-NGS network. (**a**) Network view (main connected component at the top and unconnected components below) of HI-NGS produced with PCR stitching compared to that in HI1. (**b**) Distribution of HI-NGS compared to HI1.

the 466 pairs by two assays orthogonal to Y2H: a protein complementation assay (PCA)[14] and a nucleic acid programmable protein array in wells (wNAPPA)[14]. In all three groups, the detection rate of new interactions was statistically indistinguishable from the PRS detection rate of both PCA and wNAPPA (all 'P' > 0.2; see Online Methods for calculation) and significantly higher than that of the RRS pairs (all $P < 0.001$) (**Fig. 2d**,**e**). We recovered PRS interactions in the search space at the expected rate and found no RRS pairs (**Supplementary Note 2**). Because shorter products can amplify more efficiently than longer ones in a PCR, our stitching scheme might have favored the identification of shorter ORFs, but the size distributions of ORFs, as determined by both 454 FLX and Sanger sequencing, were identical to those of the ORFs in the previous human interactome version 1 (HI1)[12] (**Fig. 2f**). Thus, large numbers of high-quality sISTs can be identified in a single next-generation sequencing reaction.

Combining 454 FLX and Sanger sequencing results produced a high-quality human interactome dataset, 'Human Interactome produced with Next-Generation Sequencing' (HI-NGS) containing 1,166 interactions among proteins encoded by 1,147 human genes (**Fig. 3a** and **Supplementary Table 2**; Molecular interaction database IM-15361). This is a 42% (1,149 new interactions) increase over HI1 data (ref. 12). The overlap of 127 interactions between the two datasets matched the expected overlap of 138 pairs[1] (**Supplementary Note 3**). The distribution of numbers of interactors per protein in HI-NGS was similar to that of previous datasets (**Fig. 3b** and **Supplementary Fig. 5**).

Despite the PCR-stitching protocol involving one additional PCR for each ORF pair compared to the traditional Y2H method, our strategy reduces the overall cost by at least ~40% and should therefore allow increased throughput (**Supplementary Fig. 6** and **Supplementary Note 4**). With continued improvement of next-generation DNA sequencing technologies, the cost of sequencing should keep decreasing[15]. Because 454 FLX sequencing can accommodate lower-capacity runs and because samples can be combined with other sequencing samples, there is no lower size limit for screens to which this method can be applied. The 82-bp linker has no identical sequence in all of GenBank, so sISTs can, in principle, be sequenced in combination with other samples (**Supplementary Note 5**). The approach should be equally effective with cDNA library screens as it was here for an ORFeome library screen.

The linker length of 82 bp requires that the average read length be >100 bp for reliable identification of sISTs. Among existing next-generation DNA sequencing technologies, the 454 technology is to our knowledge the only one that reliably produces reads of more than 100 bp on average[7]. The application of paired-end sequencing[16] to stitched PCR products would extend the approach to next-generation DNA sequencing platforms that have average read lengths less than 100 bp (**Supplementary Note 6**).

The Stitch-seq strategy implemented here for Y2H can be readily applied to other types of interaction assays, leading to improved capacity and expanded scope of interactome-network mapping.

METHODS

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

Accession codes. Molecular interaction database: IM-15361.

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

ACKNOWLEDGMENTS

We thank members of the Dana-Farber Cancer Institute Center for Cancer Systems Biology and the Roth Laboratory for helpful discussions, and members of the University of Pennsylvania DNA Sequencing Facility in the Department of Genetics for the 454 FLX sequencing. This work was funded in part by grants R01 HG001715 (to D.E.H., F.P.R. and M.V.) and R21 HG004756 (to F.P.R.) from the US National Human Genome Research Institute of the National Institutes of Health, a grant from the Ellison Foundation (to M.V.), a Canadian Institute for Advanced Research Fellowship and Canada Excellence Research Chair (to F.P.R.) and in part by Institute Sponsored Research funds from the Dana-Farber Cancer Institute Strategic Initiative in support of Center for Cancer Systems Biology. M.V. is supported as a Chercheur Qualifié Honoraire from the Fonds de la Recherche Scientifique (French Community of Belgium).

AUTHOR CONTRIBUTIONS

M.V. conceived of the project and the PCR-stitching methodology for next-generation sequencing and oversaw all aspects, including writing and editing of the manuscript; H.Y. developed the PCR-stitching process, did validations, analyzed data and co-wrote the manuscript; P.B. oversaw human interactome screening and co-wrote the manuscript; F.P.R. oversaw aspects of computational work, helped develop the next-generation sequencing strategy and contributed to editing the manuscript; D.E.H. helped with developing experimental protocols and editing the manuscript; T.H. handled database implementation and analysis of raw sequence data; L.T., S.T., E.W., F.G., N.S., T.H.-K. and J.S. carried out experimental processes for interactome mapping, PCR-stitching and validation tests; X.Y. carried out processing of PCR products and subsequent 454 FLX sequencing; C.F. and E.R. did computational and network graphing tasks; K.S.-A. contributed to troubleshooting the PCR-stitching methods; and M.E.C. contributed to writing and editing of 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. Venkatesan, K. *et al. Nat. Methods* **6**, 83–90 (2009).
2. Yu, H. *et al. Science* **322**, 104–110 (2008).
3. Deplancke, B., Dupuy, D., Vidal, M. & Walhout, A.J. *Genome Res.* **14**, 2093–2101 (2004).
4. Walhout, A.J. & Vidal, M. *Methods* **24**, 297–306 (2001).
5. Hook, B., Bernstein, D., Zhang, B. & Wickens, M. *RNA* **11**, 227–233 (2005).
6. Bennett, S.T. *et al. Pharmacogenomics* **6**, 373–382 (2005).
7. Margulies, M. *et al. Nature* **437**, 376–380 (2005).
8. Shendure, J. *et al. Science* **309**, 1728–1732 (2005).
9. Tong, A.H. *et al. Science* **294**, 2364–2368 (2001).
10. Walhout, A.J. *et al. Science* **287**, 116–122 (2000).
11. Nirantar, S.R. & Ghadessy, F.J. *Proteomics* **11**, 1335–1339 (2011).
12. Rual, J.F. *et al. Nature* **437**, 1173–1178 (2005).
13. Lamesch, P. *et al. Genomics* **89**, 307–315 (2007).
14. Braun, P. *et al. Nat. Methods* **6**, 91–97 (2009).
15. Coombs, A. *Nat. Biotechnol.* **26**, 1109–1112 (2008).
16. Maher, C.A. *et al. Proc. Natl. Acad. Sci. USA* **106**, 12353–12358 (2009).

ONLINE METHODS

Yeast two-hybrid assay. High-throughput Y2H screens were carried out according to published protocols[17]. Briefly, ~6,000 Entry clones contained in the human ORFeome version 3.1 (ref. 13) were transferred into pDEST-AD and pDEST-DB vectors (**Supplementary Note 7**) by Gateway LR reactions[18] encoding the activation domain (AD) and DNA-binding domain (DB), respectively. The vectors are available upon request. These LR recombination products were used directly to transform *Escherichia coli* (DH5α-T1^R). Transformed cells were selected on LB medium containing ampicillin, and plasmid DNA was extracted and purified. All AD-*Y* and DB-*X* plasmids were transformed into Y2H strains *MAT***a** Y8800 and *MAT*α Y8930 (genotype: *leu2-3,112 trp1-901 his3Δ200 ura3-52 gal4Δ gal80Δ GAL2::ADE2 GAL1::HIS3@LYS2 GAL7::lacZ@MET2 cyh2^R*), respectively. To identify autoactivators[19] all DB-*X* constructs were screened for growth on synthetic complete medium lacking leucine and histidine (SC −Leu, −His) supplemented with 1 mM of 3-amino-1,2,4-triazole (3-AT). All autoactivators were removed.

The AD-*Y*–containing yeast cells were combined into minipools of 188 AD-*Y* strains. To create these, first 5 µl from glycerol stocks of 94 individual AD-*Y* yeast strains were each inoculated into 500 µl of liquid SC without tryptophan (SC −Trp) medium in 96-well deep-well plates and grown for 4 d on a shaker at 30 °C. Settled yeast were resuspended by thoroughly vortexing the culture plates, and the OD_{600} of every well was measured to verify homogenous yeast growth and hence equal representation of each AD-*Y* yeast strain in the minipool. The contents of two 96-well culture plates (188 different AD-*Y* strains) were transferred into a sterile trough and mixed thoroughly to ensure equal representation of all AD-*Y* yeast strains in the pool. Archival glycerol stocks for storage at −80 °C were then prepared by combining 80 µl of the pooled yeast cultures with 80 µl of 40% (wt/vol) autoclaved glycerol in round-bottom 96-well microtiter plates.

Proteome-wide Y2H screens were carried out as described previsously[17], including inoculation of AD-minipool and DB-*X* yeast cultures, mating onto rich yeast extract peptone dextrose (YEPD) medium[17], and replica-plating onto selective SC −Leu, −Trp, −His with 1 mM 3-AT (SC −His) and SC −Leu, −His with 1 mM 3-AT plates containing 1 mg l^{-1} cycloheximide (SC −His + CHX). The latter control plates select for cells that do not have the AD plasmid owing to plasmid shuffling. Growth on selective medium thus identifies spontaneous autoactivators[19]. Only pairs that activated at least one reporter gene and were CHX-sensitive on both control plates were included in the final map. All pairs that exhibited CHX resistance on selective plates or did not pass the retest were excluded. Six Y2H controls with known phenotypes were included on all Y2H screening plates[17]. The plates were incubated overnight at 30 °C and 'replica-cleaned' the next day by placing each plate on a piece of velvet stretched over a replica-plating block and pushing evenly on the plate to remove excess yeast cells. Plates were then incubated for another 3 d, after which colonies were picked and used to inoculate liquid cultures in SC −Leu, −Trp medium. After overnight growth at 30 °C, a 5-µl aliquot was spotted onto each of the four plates for secondary phenotype confirmation (phenotyping II) (SC −His; SC −His + CHX; SC −Leu, −Trp, −adenine; SC −Leu, −adenine + CHX) to test for CHX-sensitive expression of the *LYS2::GAL1-HIS3* and *GAL2-ADE2* reporter genes. All plates were replica-cleaned the next day and scored after an additional 3 d to identify colonies that grew on SC −His or on SC −adenine but not on SC −His + CHX or on SC −adenine + CHX.

For colonies that scored positive, the identities of DB-*X* and AD-*Y* were determined using PCR stitching followed by massively parallel 454 FLX sequencing. Independently all pairs were identified by Sanger sequencing as described previously[17]. Before PCR, yeast cells from positive colonies were lysed in 15 µl of lysis buffer (2.5 mg ml^{-1} zymolase 20T (21,100 U g^{-1}; Seikagaku) dissolved in 0.1 M sodium phosphate buffer (pH 7.4)) in each well of a 96-well PCR plate. A small amount of yeast cells (not more than what fit on the end of a standard 200-µl tip) were picked and resuspended in lysis buffer in soft-shell, V-bottom 96-well microtiter plate (hereafter called PCR plate). PCR plates were put on a thermocycler to run the following lysis program: 37 °C for 15 min, 95 °C for 5 min and hold at 10 °C.

Afterward 100 µl of filter-sterilized water were added to each well. The PCR plates were centrifuged for 10 min at 800*g* and stored at −20 °C. Conditions and primers for the two rounds of PCRs in PCR stitching are available in **Supplementary Table 1**. From each PCR, 5 µl were combined together. A 1-ml aliquot of the pooled stitched PCR products was purified using QIAquick PCR Purification kit (Qiagen). A 200-µl aliquot of the purified stitched PCR products was sent to the University of Pennsylvania DNA Sequencing Facility for 454 FLX sequencing.

At the sequencing facility PCR products were processed using Roche kits (GS Standard DNA Library Preparation kit; GS FLX Standard emPCR kit (Shotgun); GS FLX PicoTiterPlate kit (70 × 75); and GS FLX Standard LR70 Sequencing kit) according to manufacturer's instructions. Briefly, 3–5 µg of the pooled PCR products were fragmented by nebulization for 1 min under nitrogen gas pressure of 30 p.s.i. (2.1 bar), the DNA fragments were size-selected and subjected to end-polishing and adaptor ligation. The library was then immobilized onto streptavidin-coated beads followed by the fill-in reaction to repair the gaps generated by the ligation of nonphosphorylated adaptors to the fragments. A single-stranded library was created by melting off the nonbiotinylated strand of bead-bound fragments. Subsequent quality assessment and quantitation were done by 96-well plate fluorometry and analysis on a Bioanalyzer with Agilent RNA Pico 6000 LabChip kit. The amount of library DNA needed for optimal results in the emulsion-based clonal amplification (emulsion PCR) procedure was determined by emulsion titration assay according to manufacturer's instructions. The library of DNA fragments was amplified from a single bead-bound copy to millions of copies per bead using water-in-oil emulsion PCR. Subsequently, emulsions were broken and the beads carrying the amplified library were recovered using biotinylated amplification primers and streptavidin-coated magnetic beads with manufacturer-provided protocols. Beads were counted, the enrichment ratio was calculated and the recommended amount of sequencing primer was added to bead-bound amplified fragments. After annealing and mixing of DNA-loaded beads with packing beads, the wells of a GS FLX Standard PicoTiterPlate were loaded according to manufacturer's protocols, that is, subsequent layers of enzyme beads, the mix of DNA and packing beads, another layer of enzyme beads followed by a layer of apyrase beads. The loaded PicoTiterPlate was inserted into the 454 FLX instrument and run using the standard protocol.

From the 454 FLX sequencing data, we first identified all usable sequencing reads containing the 82-bp linker using 'cross_match'[20]. Then sISTs were identified by mapping both ends of the usable reads to the screened ~6,000 ORFs in human ORFeome 3.1, using BLASTN (mismatches allowed) with an *E*-value cutoff of 10^{-3}.

From the set of successfully sequenced DB-*X* and AD-*Y* pairs, all interacting protein pairs were verified in a single-pass pairwise retesting to ensure the robustness of the His^+ or Ade^+ phenotypes and to exclude the possibility that physiologic and genetic changes that occurred during the experiment gave rise to experimental artifacts[17]. For retesting, liquid cultures of individual yeast strains with corresponding AD and DB constructs were inoculated from archival glycerol stocks, grown overnight and arrayed for pairwise Y2H analysis using the procedure outlined above. Briefly, from the YEPD mating plates, yeast colonies were replica-plated onto SC −Leu, −Trp plates to select diploid yeast cells containing both AD and DB constructs. Diploid yeast cells were subsequently replica-plated onto four Y2H assay plates identical to the ones used for phenotyping II. All interactions so verified have been deposited with the International Molecular Exchange (IMEx) consortium.

Protein complementation assay. For PCA[17] human ORFs available in Gateway Entry vectors were transferred by Gateway LR reactions into vectors encoding the two fragments of a yellow fluorescent protein (Citrine variant) fused to the N terminus of the tested proteins. Baits were fused to the F1 fragment (amino acids 1–158 of Citrine) and preys to the F2 fragment (amino acids 159–239 of Citrine). After bacterial transformation, minipreps were prepared on a Qiagen BioRobot, and DNA concentrations were determined by PicoGreen assay (Invitrogen) in 96-well format according to the manufacturer's protocols. A 30 ng aliquot of each vector encoding the two proteins was added to 140 ng of a *CFP* control plasmid for transfection into CHO-K1 cells in 96-well plates, using Lipofectamine 2000 (Invitrogen) according to the manufacturer's instructions. At ~18 h after transfection, cells were washed twice with PBS, trypsinized with 15 µl cell culture grade trypsin, suspended in 130 µl PBS (pH 7.4) and analyzed by fluorescence-activated cell sorting on a Canto II FACS (Becton Dickinson) equipped with a 96-well autosampler. Viable cyan-fluorescent cells, that is, transfected cells, were selected and analyzed for yellow fluorescent protein signal. A pair was considered interacting if at least 30% of CFP-expressing cells had a yellow fluorescent protein signal, and the yellow/cyan signal ratio was at least twice as high as the ratio of the average yellow fluorescence signal across the entire plate over the average cyan fluorescence ratio on that plate.

Nucleic acid programmable protein array in wells. For the wNAPPA assay[17] ORFs encoding the interacting proteins were cloned into Gateway-compatible pCITE-HA and pCITE-GST vectors by the LR reaction. After transformation, growth, DNA minipreps and determination of DNA concentration, ~0.5 µg of each plasmid were added to Promega TnT coupled transcription-translation mix and incubated for 1.5 h to express proteins. During this time GST antibody–coated 96-well plates (Amersham 96-well GST detection module) were blocked at room temperature (25 °C) with PBS containing 5% dry milk powder. After protein expression, the expression mix was diluted in 100 µl blocking solution and added to the emptied preblocked 96-well plates. Binding was done for 2 h at 16 °C with agitation. After capture, plates were washed three times and developed by incubation with primary and secondary antibody. Signal was visualized using enhanced chemiluminescence (Pierce PicoWest ECL reagent) with a Bio-Rad ChemiDoc. Signal was manually assigned a score between 0 and 5 (0 corresponding to background in empty controls, and 5 being a saturated signal). Wells that scored ≥ 2 in either configuration were deemed to contain positively interacting pairs.

***P* value calculations.** All *P* values were calculated by the following equation:

$$z = \frac{p_1 - p_2}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where x_1 is the number of positives in dataset 1 detected by a given assay; n_1 is the total number of pairs in dataset 1; x_2 is the number of positives in dataset 2 detected by the assay; n_2 is the total number of pairs in dataset 2; $p_1 = x_1 / n_1$; $p_2 = x_2 / n_2$ and

$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

$P = 1 - \text{probability}(-z < Z < z)$.

17. Dreze, M. *et al. Methods Enzymol.* **470**, 281–315 (2010).
18. Walhout, A.J. *et al. Methods Enzymol.* **328**, 575–592 (2000).
19. Walhout, A.J. & Vidal, M. *Genome Res.* **9**, 1128–1134 (1999).
20. Ewing, B., Hillier, L., Wendl, M.C. & Green, P. *Genome Res.* **8**, 175–185 (1998).

doi:10.1038/nmeth.1597