Isoform discovery by targeted cloning, 'deep-well' pooling and parallel sequencing

Kourosh Salehi-Ashtiani[1,2,5], Xinping Yang[1,2,5], Adnan Derti[1,3,5], Weidong Tian[1,3,5], Tong Hao[1,2,5], Chenwei Lin[1,2], Kathryn Makowski[4], Lei Shen[4], Ryan R Murray[1,2], David Szeto[1,2], Nadeem Tusneem[4], Douglas R Smith[4], Michael E Cusick[1,2], David E Hill[1,2], Frederick P Roth[1,3] & Marc Vidal[1,2]

Describing the 'ORFeome' of an organism, including all major isoforms, is essential for a system-level understanding of any species; however, conventional cloning and sequencing approaches are prohibitively costly and labor-intensive. We describe a potentially genome-wide methodology for efficiently capturing new coding isoforms using reverse transcriptase (RT)-PCR recombinational cloning, 'deep-well' pooling and a next-generation sequencing platform. This ORFeome discovery pipeline will be applicable to any eukaryotic species with a sequenced genome.

Experimental definition of the complete set of protein-coding transcript sequences ('ORFeome') is fundamental for complete understanding of any organism, but this has not been achieved to date for any metazoan. Adding to the uncertainty, many eukaryotic genes exhibit alternative splicing, leading to a diversity of open reading frames (ORFs) encoded by a single gene. Currently, ~74% of human genes and ~13% of *Caenorhabditis elegans* genes are predicted to undergo alternative splicing[1,2]. Expansion of the 'isoform space' in more complex organisms may partly explain the paradoxical lack of correlation between organismal complexity and gene number, and underscores the need to efficiently and comprehensively capture the full ORFeome. Historically, determination of intron-exon boundaries in eukaryotes has been addressed mainly by large-scale sequencing of random cDNAs (expressed sequence tags; ESTs) followed by alignment to a reference genomic DNA sequence. Although EST collections are extremely helpful, the human isoform space remains underexplored. A targeted cloning and full-length sequencing strategy could provide the desired information but is impractically resource-intensive.

Next-generation parallel sequencing technologies, such as the Roche 454 FLX, offer the prospect of sequencing at a much faster pace and lower cost than conventional Sanger sequencing–based capillary platforms[3]. Most applications described so far have entailed resequencing of megabase-scale genomic DNA fragments[4–7] or of small sequence tags[8–11]. A disadvantage of the latter approach is that *cis* connectivity is lost between the reads; therefore, although the reads can be assembled into contigs, mRNAs cannot be assembled unambiguously when splice variants are involved. Sequencing of kilobase-scale DNA fragments from complex pools in which fragments have heterogeneous abundance has not yet been tested, nor has correct assembly of hundreds to thousands of full-length cDNAs in parallel from a complex mixture been proven feasible.

Previous and ongoing full-length cDNA isolation projects aim to discover one isoform per gene, without attempting to investigate the depth of 'isoform space'. Here we describe and demonstrate the

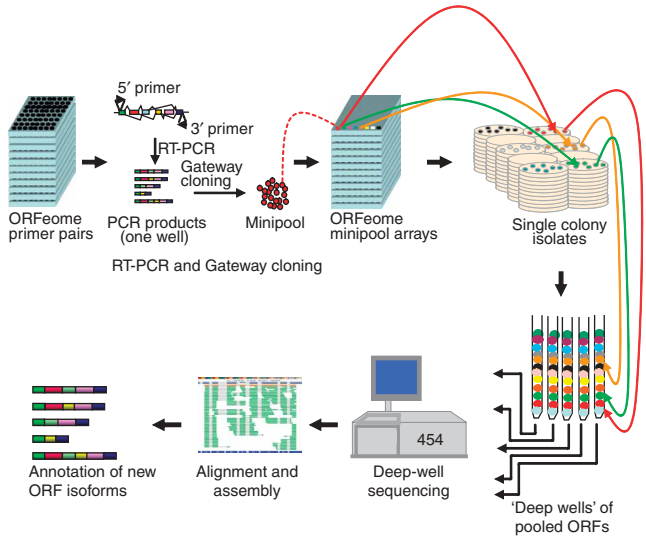

Figure 1 | The isoform discovery pipeline. First, ORFs are captured in RT-PCR experiments, cloned and transformed into *Escherichia coli*. Minipools of transformants for each gene may contain different isoforms. Second, deep-well pools are constructed by pooling the PCR-amplified ORF sequence from one transformant for each of many genes. This method of pooling ensures normalization of ORFs and avoids concurrent sequencing of multiple isoforms. Third, parallel sequencing is carried out separately on each deep well. The obtained reads are assembled using an SBA algorithm. Resulting ORF contigs are filtered for the presence of noncanonical splice acceptor/receptor sites and prior presence in sequence databases to identify unique 'novel' isoforms.

[1]Center for Cancer Systems Biology, Department of Cancer Biology, Dana Farber Cancer Institute, 44 Binney Street, Boston, Massachusetts 02115, USA. [2]Department of Genetics, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, Massachusetts 02115, USA. [3]Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, 250 Longwood Avenue, Boston, Massachusetts 02115, USA. [4]Agencourt Bioscience Corporation, 500 Cummings Center, Beverly, Massachusetts 01915, USA. [5]These authors contributed equally to this work. Correspondence should be addressed to M.V. (marc_vidal@dfci.harvard.edu) or K.S.-A. (kourosh_salehi-ashtiani@dfci.harvard.edu) or F.P.R. (fritz_roth@hms.harvard.edu).

RECEIVED 19 FEBRUARY; ACCEPTED 21 MAY; PUBLISHED ONLINE 15 JUNE 2008; DOI:10.1038/NMETH.1224

Table 1 | Splice variants captured from pooled RT-PCRs of brain, testis, heart, liver and placenta

			Novel	
Entrez identifier	Gene symbol	Non-novel[a]	GY-AG[b]	Other[c]
10449	*ACAA2*	1	1	0
27237	*ARHGEF16*	1	1	0
408	*ARRB1*	2	2	0
7918	*BAT4*	1	1	2
54930	*C14orf94*	3	1	0
79411	*GLB1L*	1	1	0
80270	*HSD3B7*	1	2	0
28981	*IFT81*	1	1	0
9776	*KIAA0652*	0	3	1
8569	*MKNK1*	2	1	0
55471	*PRO1853*	1	0	1
51100	*SH3GLB1*	1	1	0
10629	*TAF6L*	1	1	0
95	*ACY1*	1	0	0
123	*ADFP*	1	0	0
57332	*CBX8*	1	0	0
1848	*DUSP6*	2	0	0
7157	*TP53*	1	0	0
84790	*TUBA6*	1	0	0
26100	*WIPI-2*	1	0	0
84287	*ZDHHC16*	2	0	0
10617	*STAMBP*	0	0	0
Total		26	16	4

[a]Captured isoforms represented in their entirety by individual transcripts in GenBank (including ESTs), RefSeq or MGC. [b]Novel isoforms with GT-AG (canonical) or GC-AG splice donor-acceptor signals (Y, pyrimidine). [c]Novel isoforms with at least one splice donor-acceptor pair other than GT-AG or GC-AG. Redundant sequences were not counted.

feasibility of a pipeline for large-scale discovery and cloning of coding isoforms. We tested each individual component of the pipeline and demonstrated overall effectiveness for isolation of new coding isoforms, successfully sequencing and assembling ~820 ORFs in parallel.

The 'deep-well' strategy has three elements (**Fig. 1**): (i) efficient capture and cloning of ORF isoforms; (ii) 'deep-well' pooling; and (iii) parallel sequencing and assembly of the obtained fragmentary ORF sequence tags (fOSTs) into full-length ORFs. The capture of coding isoforms starts with RT-PCR using primers annealing to annotated ORFs. Complex mixtures of RNAs from one or more tissues are reverse transcribed, PCR amplified and cloned using the Gateway recombination methodology[12]. As products of each PCR can contain mixtures of several splice variants, the obtained bacterial transformants represent 'minipools' that potentially contain different coding isoforms of the same ORF. Individual colonies are picked from minipools and arrayed across 96- or 384-well plates for archival storage and subsequent consolidation into equimolar normalized pools of single-colony isolates. In 'deep-well' pooling, aliquots from the same individual well from each plate of a set of arrayed plates are combined such that each pool contains one colony from each of the targeted gene loci (in other words, only one colony from any given minipool is included in each deep well). Deep-well pooling creates a library that is perfectly normalized across genes, unlike non-normalized cDNA libraries, which may be dominated by a few abundantly transcribed genes. Transcripts are 'segregated' in the sense that each deep-well pool contains just one coding variant from each gene locus in the target set. This segregation is critical to ensure that each assembled contig is composed of sequence fragments arising from one specific transcript for any given gene.

The search space along an ORF is established by the choice of primer pairs. To focus on coding potential in the human genome, we directed our primer pairs solely to coding regions of the annotated human cDNAs available from the Mammalian Gene Collection (MGC)[13,14]. This strategy discovers new coding variants that share the primer sites with the original cDNA used to design the primer pairs for cloning ORFs[14], and has the additional advantage that the resulting clones are immediately useful for protein expression analysis.

To test whether coding regions of isoforms can be robustly amplified, we carried out a medium-scale RT-PCR experiment on 94 human ORFs (randomly chosen among genes with available primer pairs based on the Human ORFeome 3.1 collection[15]), using five normal human tissue RNA preparations as template (**Supplementary Fig. 1a** and **Supplementary Methods** online). Our PCR success rates were 75–88% for all five tissues. We then amplified by RT-PCR ~820 'disease' ORFs (those associated with one or more human disorders in the Online Mendelian Inheritance in Man (OMIM) database[16]), observing success rates of 67%, 66%, 78%, 34% and 53% from testis, brain, heart, liver and placenta RNA, respectively (**Supplementary Fig. 1b**).

We used subsets of these PCR products for subsequent Gateway cloning. To generate set 1 ('pooled tissue'), we pooled and cloned products of RT-PCRs for 22 ORFs (chosen from the 94 above) for each of five tissue RNAs (**Supplementary Fig. 1a**). Set 2 ('brain') and set 3 ('testis') each corresponded to RT-PCR amplifications of a different set of ORFs randomly selected from the OMIM set. We cloned these from brain and testes RNA, respectively (**Supplementary Fig. 1b**). As a control, we included *HSD3B7* (hydroxy-delta-5-steroid dehydrogenase, 3-beta- and steroid delta-isomerase 7) in set 1 as well as sets 2 and 3.

We summarize the cloning results in **Tables 1** and **2**, and present genomic alignments of three examples in **Figure 2** (**Supplementary Fig. 2** online shows the complete set of aligned sequences). We considered a sequence 'novel' unless it was recapitulated in its entirety by a single transcript from the MGC, RefSeq and GenBank resources, including dbEST. For set 1, in which we pooled the PCR products before cloning, we discovered 20 novel variants in the 22 genes tested. These included 16 novel variants in 12 genes with only GY-AG splice signals (that is, with canonical GT-AG or with GC-AG) and 4 novel variants with at least one unusual splice signal. Of the 22 genes examined in set 2 and set 3, we discovered 23 novel splice variants in 9 genes. These included 10 novel GY-AG variants in 6 genes for set 2, and 4 such variants in 3 genes for set 3. For *HSD37,* one novel GY-AG variant occurred in all three sets. In summary, we isolated an average of 18 clones per gene from a small number of tissues and discovered novel splice variants with canonical or typical alternative splice signals for almost half (19 out of 44) of the genes examined (see **Supplementary Note** online).

Next-generation sequencing technologies provide greatly reduced cost per raw base sequenced but have, to varying degrees, the serious drawback of shorter read lengths. To assess whether our deep-well strategy can be coupled to the 454 pyrosequencing technology, we tested the set of cloned OMIM disease ORFs (**Supplementary Fig. 1b**). To eliminate ambiguity in assembly of

Table 2 | Splice variants captured by RT-PCR from set 2 and set 3

			Novel, set 2 (brain)		Novel, set 3 (testis)	
Entrez identifier	Symbol	Non-novel[a]	GY-AG[b]	Other[c]	GY-AG[b]	Other[c]
445	*ASS1*	1	1	3	0	0
1497	*CTNS*	3	1	0	0	0
201163	*FLCN*[d]	1	3	1	2	0
3043	*HBB*	1	0	1	0	0
80270	*HSD3B7*[e]	1	0	0	0	0
4953	*ODC1*	1	0	1	0	0
80025	*PANK2*	1	0	0	1	0
5213	*PFKM*[d]	2	1	3	1	0
5660	*PSAP*	1	3	0	0	0
6102	*RP2*	1	1	0	0	0
8542	*APOL1*	2	0	0	0	0
1738	*DLD*	1	0	0	0	0
55670	*PEX26*	2	0	0	0	0
57104	*PNPLA2*	1	0	0	0	0
5538	*PPT1*	2	0	0	0	0
6403	*SELP*	3	0	0	0	0
219736	*STOX1*	1	0	0	0	0
1861	*TOR1A*	1	0	0	0	0
7391	*USF1*	1	0	0	0	0
7422	*VEGF*	3	0	0	0	0
8565	*YARS*	3	0	0	0	0
56652	*PEO1*	0	0	0	0	0
Total		33	10	9	4	0

[a]Captured isoforms represented in their entirety by individual transcripts in GenBank (including ESTs), RefSeq, or MGC. [b]Novel isoforms with GT-AG (canonical) or GC-AG splice donor-acceptor signals (Y, pyrimidine). [c]Novel isoforms with at least one splice donor-acceptor pair other than GT-AG or GC-AG. Redundant sequences were not counted. [d]A novel GY-AG variant was detected in both sets but is reported only for Set 2 (see **Supplementary Fig. 2**). [e]A novel GY-AG variant was detected in both sets but is already reported in **Table 1**.

short reads, we chose a set such that no two ORFs shared any 50-bp subsequence with >90% identity. The resulting ORF set, taken from the human ORFeome 3.1 collection[15], encompassed a broad size range (~0.15–5.1 kb), and corresponded to ~96% of the OMIM ORFs. PCR amplification using vector primers flanking the ORFs successfully amplified ~820 ORFs (**Supplementary Fig. 1b**). We pooled these PCR products and sequenced them by the 454 platform. The sequencing run produced 145,318 reads with an average read length of 240 bases (approximately 35 million bases total) and ~25-fold coverage of each base in the set (**Supplementary Fig. 3** online).

We assembled contigs from fOSTs using the human genome sequence as the template for assembly. Knowing the genomic location of ORF-targeted primers allows us to limit template sequences to genomic regions between targeting primers. In aligning fOSTs to the genome, bridging of consecutive exons by BLAT requires that a 'bridging' read has a sufficient length of exonic sequence on each side of the intervening intron. Although these length requirements may be relaxed when strong hypotheses about the locations of exon ends are available, such information is not currently used by BLAT. After initially using a pipeline that combines existing software packages, we developed a more advanced pipeline that better reveals intron-exon structure ('smart bridging assembly' or SBA; **Supplementary Methods** and **Supplementary Fig. 4** online). SBA introduces two features not present in our initial conventional assembly method. First, adjacent contigs with inner termini that correspond approximately to exon ends may be 'bridged' by examination of unaligned sequence segments that may have been too short for BLAT to justify the introduction of an intron-sized gap. Second, where two contigs aligned to the genome are separated by a gap that is too small to contain an intron, these contigs are 'bridged' by filling the gap with the known genomic sequence.

We applied both conventional and SBA assembly methods to all the 454 FLX reads and computed the percentage of ORFs with 100% correctly assembled gene structure (**Fig. 3a**). We also simulated the effects of reduced coverage by assembling a randomly chosen subset of 20%, 40%, 60% and 80% of the 454 FLX reads corresponding to 5-, 10-, 15- and 20-fold coverage. We repeated this procedure 10 times to compute the average percentage of correctly assembled ORFs. The SBA method outperformed the

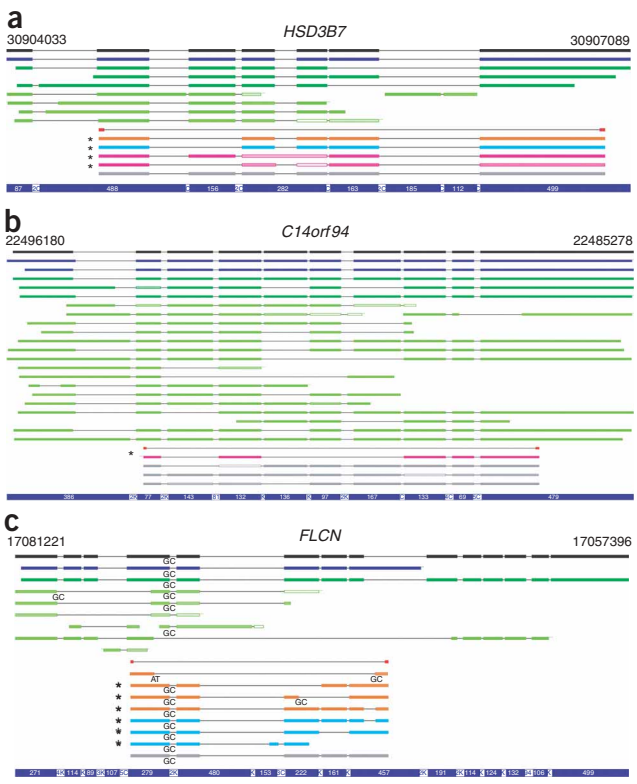

Figure 2 | Examples of identified ORFs. (**a**–**c**) Genomic alignments of representative genes from sets 1–3 compared with RefSeq (black), MGC (blue), GenBank (dark green) and dbEST (light green) after removal of redundant alignments. Results are shown for 3 of 44 genes from which ORFs were cloned (the complete set is in **Supplementary Fig. 2**). Transcripts with exon/intron structures that were exactly recapitulated over the entire length by individual MGC, Refseq or GenBank transcripts, including ESTs, are shown in gray, and novel ones are shown for the pooled tissue (purple), brain (orange) and testis (cyan) cloning experiments. The positions of primers used for RT-PCR are shown in red. Color saturation indicates percent identity, ranging from light (≤90% identity) to dark (≥99% identity). Splice signals other than the canonical GT donor and AG acceptor are shown for all sequences. Novel isoforms with only canonical or GC-AG signals are indicated by an asterisk. For simplicity, ESTs with unusual splice signals are not shown, but they were included in the assessment of novelty. Chromosomal coordinates are indicated at the top of each panel. The blue bar at the bottom of each panel indicates the lengths of exonic (white on blue) and intronic (reversed) segments, in base pairs (C = 100; K = 1,000); introns are compressed to highlight exons.

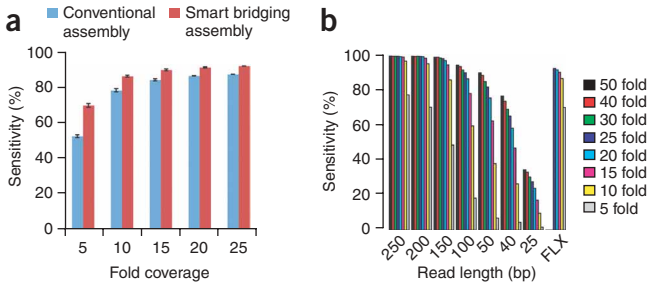

Figure 3 | Sequence assembly results and simulation. (**a**) Success rates of assembly using conventional and smart bridging assembly methods at indicated fold coverage. The percentage of ORFs with 100% correctly assembled gene structure (exon-intron) was computed ($n = 10$ repeats). Error bars represent s.d. (**b**) The set of ORF sequences used in the 454 FLX run were randomly fragmented *in silico* with average fragment size of 550 base pairs and range of 300–800 bp. Different sequence read lengths and fold coverages were simulated. For each ORF, we assembled contigs based on all available sequence reads that have a corresponding best match in the genomic region of the ORF. The graphs illustrate sensitivity by gene, that is, the percentage of ORFs whose gene structure (all exons) is 100% correctly assembled.

conventional method, particularly when coverage was low (**Fig. 3a**). For example, at fivefold coverage, the SBA method assembles 70% of ORFs correctly, as compared with 52% for the conventional method.

Other next-generation sequencing platforms offer substantially shorter reads than the 454 FLX system, with reduced per-base cost of sequencing. To evaluate compatibility of such platforms with the deep-well strategy, we tested *in silico* the rates of successful assembly at different read lengths and different depths of coverage (**Fig. 3b**). Not surprisingly, the fraction of genes with completely correct assembly of all exons was strongly affected by read length. As a reference we used sequence reads from the actual 454 FLX reactions, finding that at least 15-fold coverage was needed to achieve 90% per-gene sensitivity. For short read lengths, the same extent of assembly could be achieved by increased overall coverage. In the simulation, to achieve a similar assembly success rate for read lengths of 200, 150 and 100 bases, fold coverages of 10, 15 and 25 were needed, respectively (**Fig. 3b**). Even at 50-fold coverage, sequence reads of 25 bases achieved a per-gene sensitivity of only 34%; however, 50-fold coverage allowed 76% and nearly 90% per-gene sensitivity, respectively, for 40- and 50-base read lengths. Taken together, our experiments suggest that transcripts can be assembled accurately with read lengths smaller than those produced by the 454 FLX technology, but substantial increases in coverage would be needed for read lengths shorter than 40 bp (see also **Supplementary Figs. 5–10** and **Supplementary Note** online for details).

The deep-well isoform discovery methodology described and validated here can now be used for genome-wide isoform discovery projects. One potential source of ambiguity in the deep-well strategy is the presence of paralogs with high nucleotide sequence similarity. This difficulty is easily addressed by separating clones from paralogs into distinct deep-well pools. Deep-well pools can be easily assembled with available robotic liquid handling systems such that 'tailored' sets of pools of any size and composition can be generated. For genome-scale implementation of the deep well strategy for humans using 454 FLX technology, the set of targeted genes (currently estimated to be between ~20,500 (ref. 17) to ~34,000 (ref. 18) might be best separated into deep-well pools of ~4,000 genes each. Each pool would contain ~4 Mb of unique sequence, and 454 FLX sequencing could produce 10× coverage at current capacity. The optimal number of clones in each deep well depends on both the raw quantity of sequence that a sequencing run can generate as well as the read length (as this determines the required sequence coverage). If we were to limit our search space to the current RefSeq transcripts, we can expect as few as 18×4 deep wells (corresponding to 18 colonies sequenced for 19,000 RefSeq ORFs or the equivalent of ~342,000 sequencing reactions) to yield novel isoforms for about half of the RefSeq genes, relative to GenBank, including 7.8 million EST entries. Given the efficiency of this approach, conjoined with rapidly increasing capacity and read length of emerging sequencing methods, whole-genome experiments can be carried out with increasing speed and cost-effectiveness.

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

ACKNOWLEDGMENTS

This work was funded in part by a grant from the Ellison Foundation (awarded to M.V.) and in part by the Dana Farber Cancer Institute Strategic Initiative in support of Center for Cancer Systems Biology. F.P.R. acknowledges support from US National Institutes of Health grants NS054052, HG003224, HL081341. W.T. was supported in part by National Institutes of Health grant DK070078. We thank the West Quad Computing Group at Harvard Medical School as well as Research Computing at Massachusetts General Hospital for assistance with computational resources. We thank G. Temple for helpful comments on the manuscript.

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturemethods/.

Published online at http://www.nature.com/naturemethods/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Johnson, J.M. *et al. Science* **302**, 2141–2144 (2003).
2. Zahler, A.M. *WormBook* **2005**, 1–13 (2005).
3. Schuster, S.C. *Nat. Methods* **5**, 16–18 (2008).
4. Margulies, M. *et al. Nature* **437**, 376–380 (2005).
5. Shendure, J. *et al. Science* **309**, 1728–1732 (2005).
6. Moore, M.J. *et al. BMC Plant Biol.* **6**, 17 (2006).
7. Oh, J.D. *et al. Proc. Natl. Acad. Sci. USA* **103**, 9999–10004 (2006).
8. Torres, T.T., Metta, M., Ottenwalder, B. & Schlotterer, C. *Genome Res.* **18**, 172–177 (2008).
9. Porreca, G.J. *et al. Nat. Methods* **4**, 931–936 (2007).
10. Emrich, S.J., Barbazuk, W.B., Li, L. & Schnable, P.S. *Genome Res.* **17**, 69–73 (2007).
11. Wicker, T. *et al. BMC Genomics* **7**, 275 (2006).
12. Walhout, A.J. *et al. Methods Enzymol.* **328**, 575–592 (2000).
13. The MGC Project Team. *Genome Res.* **14**, 2121–2127 (2004).
14. Rual, J.F. *et al. Genome Res.* **14**, 2128–2135 (2004).
15. Lamesch, P. *et al. Genomics* **89**, 307–315 (2007).
16. Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A. & McKusick, V.A. *Nucleic Acids Res.* **33**, D514–D517 (2005).
17. Clamp, M. *et al. Proc. Natl. Acad. Sci. USA* **104**, 19428–19433 (2007).
18. Yamasaki, C. *et al. Nucleic Acids Res.* **36**, D793–D799 (2008).