Human ORFeome Version 1.1: A Platform for Reverse Proteomics

Jean-François Rual,[1,2,9] Tomoko Hirozane-Kishikawa,[1,9] Tong Hao,[1,9] Nicolas Bertin,[1,3,9] Siming Li,[1,9] Amélie Dricot,[1] Ning Li,[1] Jennifer Rosenberg,[1] Philippe Lamesch,[1,2] Pierre-Olivier Vidalain,[1] Tracey R. Clingingsmith,[1] James L. Hartley,[4] Dominic Esposito,[4] David Cheo,[6] Troy Moore,[6] Blake Simmons,[6] Reynaldo Sequerra,[7] Stephanie Bosak,[7] Lynn Doucette-Stamm,[7] Christian Le Peuch,[3] Jean Vandenhaute,[2] Michael E. Cusick,[1] Joanna S. Albala,[1,8] David E. Hill,[1,10] and Marc Vidal[1,10]

[1]Center for Cancer Systems Biology and Department of Cancer Biology, Dana-Farber Cancer Institute and Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA; [2]Facultés Universitaires Notre-Dame de la Paix, 5000 Namur, Belgium; [3]Centre de Recherche en Biochimie Macromoleculaire, Centre National de la Recherche Scientifique FRE 2593, 34293 Montpellier Cedex 5, France; [4]Science Applications International Corporation/National Cancer Institute, Frederick, Maryland 21702, USA; [5]Atto Bioscience, Rockville, Maryland 20850, USA; [6]Open Biosystems, Inc., Huntsville, Alabama 35806, USA; [7]Agencourt Biosciences Corporation, Beverly, Massachusetts 01915, USA; [8]Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, University of California, Livermore, California 94551, USA

The advent of systems biology necessitates the cloning of nearly entire sets of protein-encoding open reading frames (ORFs), or ORFeomes, to allow functional studies of the corresponding proteomes. Here, we describe the generation of a first version of the human ORFeome using a newly improved Gateway recombinational cloning approach. Using the Mammalian Gene Collection (MGC) resource as a starting point, we report the successful cloning of 8076 human ORFs, representing at least 7263 human genes, as mini-pools of PCR-amplified products. These were assembled into the human ORFeome version 1.1 (hORFeome v1.1) collection. After assessing the overall quality of this version, we describe the use of hORFeome v1.1 for heterologous protein expression in two different expression systems at proteome scale. The hORFeome v1.1 represents a central resource for the cloning of large sets of human ORFs in various settings for functional proteomics of many types, and will serve as the foundation for subsequent improved versions of the human ORFeome.

[Supplemental material is available online at www.genome.org.]

The currently available annotations of the human genome sequence (Lander et al. 2001; Venter et al. 2001) and those of model organisms (Goffeau et al. 1996; The *C. elegans* Sequencing Consortium 1998; Adams et al. 2000; *Arabidopsis* Genome Initiative 2000; Waterston et al. 2002; Gibbs et al. 2004) provide a necessary framework, sometimes referred to as the "parts list", for the ongoing transition from molecular biology (detailed single-gene studies of protein function) to systems biology (local and global analyses of the molecular networks in which proteins function; Ideker et al. 2001; Vidal 2001; Kitano 2002). However, most gene products predicted from the currently available genome annotations remain functionally uncharacterized. One essential step in the development of global studies of molecular networks is the systematic mapping of macromolecular interactions and of biochemical reactions, using reverse proteomics approaches (Walhout and Vidal 2001). Reverse proteomics projects, in turn, require the cloning and manipulation of large numbers of protein-encoding sequences, or open reading frames (ORFs).

Reverse proteomics approaches such as high-throughput yeast two-hybrid (HT-Y2H) analyses (Walhout et al. 2000a; Li et al. 2004), pull-down of tagged protein complexes followed by mass spectrometry (Gavin et al. 2002; Ho et al. 2002), proteome chips (Zhu et al. 2001), and reverse transfection microarray strategies (Ziauddin and Sabatini 2001), all entail cloning of large numbers of protein-encoding ORFs into many different expression vectors, and subsequent heterologous expression of large numbers of proteins. Frequently, such proteins need to be expressed as covalent fusions to well-characterized protein tags, which act as experimental anchors or functional moieties. In sum, nearly complete sets of ORFs, or ORFeomes (Walhout et al. 2000b), cloned into flexible vectors, are a necessary antecedent to take full advantage of the information found in any genome sequence (Rual et al. 2004b).

A metazoan ORFeome resource was first generated for the nematode *Caenorhabditis elegans* using Gateway recombinational cloning (Reboul et al. 2003; Lamesch et al. 2004), and was used for reverse proteomic approaches such as HT-Y2H based protein–protein interaction mapping (Li et al. 2004), protein chips (Reboul et al. 2003), and large-scale phenotypic (phenome) mapping (Rual et al. 2004a). The *C. elegans* ORFeome project illustrates why ORFeome resources constitute a necessary bridge between whole-genome sequencing and downstream omics applications (Boone and Andrews 2003).

Here, we describe an improved version of the Gateway-cloning strategy and its use for the generation of a Gateway-cloned human ORFeome designated version 1.1 (hORFeome

[9]These authors contributed equally to this work.
[10]Corresponding authors.
E-MAIL marc_vidal@dfci.harvard.edu; FAX (617) 632-5739.
E-MAIL david_hill@dfci.harvard.edu; FAX (617) 632-5739.

Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.2973604.

14:2128–2135 ©2004 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/04; www.genome.org

v1.1), and we demonstrate that hORFeome v1.1 is amenable to proteome-scale protein expression in diverse expression systems.

RESULTS

Defining the First Version of the Human ORFeome

The ORFeome of an organism corresponds to its complete set of protein-encoding genes, cloned as full-length open reading frames (ORFs). An ORF consists of the entire coding sequence between the initiation and termination codons, excluding the 5′ and 3′ mRNA untranslated regions (UTRs; Walhout et al. 2000b). For any organism of interest, a cloned ORFeome resource should ideally include all variants of all genes expressed in all cells at all stages of development. In reality, the completeness and quality of an ORFeome resource depends greatly on the completeness and quality of the underlying genome annotation.

Significant challenges remain in identifying all human genes, particularly those for which limited experimental annotation exists (Collins et al. 2003). However, public collections of human cDNAs provide a useful starting point for the construction of a human ORFeome resource. We focused on cloning all unique human ORFs that are already available as full-length cDNAs in the Mammalian Gene Collection (MGC; Strausberg et al. 2002). The long-term goal of the MGC project is to identify and sequence most human cDNA clones that contain a full-length ORF (FL-ORF; Fig. 1A,B). Although useful for defining what constitutes an ORFeome, the pioneering MGC effort has limited application. The presence of 5′ and 3′ UTRs in the MGC clones precludes expression of the encoded proteins as N- or C-terminal fusions to protein tags. Second, the vectors used for cDNA cloning by MGC are usually not compatible with most expression systems, and are certainly not compatible with versatile and high-throughput recombinational cloning approaches (Brasch et al. 2004; Marsischky and LaBaer 2004). However, the MGC collection currently represents the best resource for the generation of a first version of the human ORFeome, both as a source of high-quality gene structure annotation and for use as template DNA for PCR amplification of ORFs (Fig. 1A,B,C).

When we began this effort, the MGC collection consisted of 12,710 available cDNA clones, arrayed in 133 96-well plates (Fig. 1B). From this collection, 10,196 distinct full-length ORFs were identified. Removed from the initial set of 12,710 were: (1) 1068 clones for which the corresponding ORFs were reported as "partial coding sequence" at NCBI (no 5′ ATG detected); (2) multiple copies of identical cDNA clones (276 clones); and (3) clones that contain an identical coding sequence, but differ in their 5′ and/or 3′ UTRs (1170 clones; Fig. 1B). An additional 42 clones were discarded that have ORF length smaller than 100 nucleotides, a threshold three times smaller than the convention (300 nucleotides or 100 amino acids) initiated during the annotation of the yeast genome (Goffeau et al. 1996). The remaining 10,154 MGC clones were arrayed by increasing ORF length into 109 96-well plates, and constituted our starting point (Fig. 1C).

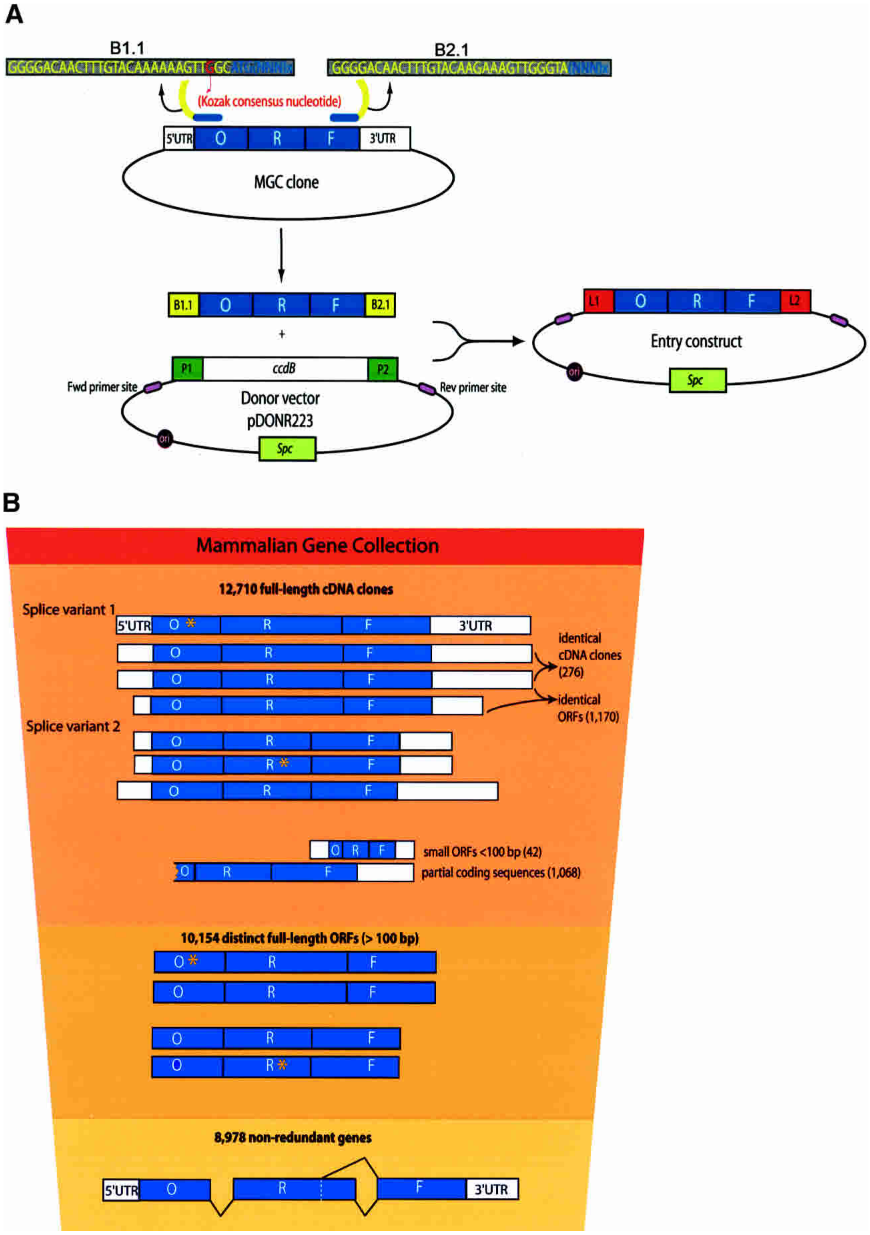
A
B1.1
GGGGACAACTTTGTACAAAAAAGTTGGCATGNNNNN
B2.1
GGGGACAACTTTGTACAAGAAAGTTGGGTANNNNN
(Kozak consensus nucleotide)
5'UTR
O
R
F
3'UTR
MGC clone
B1.1
B2.1
+
P1
ccdB
P2
Fwd primer site
Rev primer site
Donor vector
pDONR223
ori
Spc
L1
L2
Entry construct
B
Mammalian Gene Collection
12,710 full-length cDNA clones
Splice variant 1
Splice variant 2
identical cDNA clones (276)
identical ORFs (1,170)
small ORFs <100 bp (42)
partial coding sequences (1,068)
10,154 distinct full-length ORFs (> 100 bp)
8,978 non-redundant genes

Figure 1 (Continued on next page)

The sequences of the 10,154 unique ORFs in the starting collection were aligned to the human genome using Golden Path release hg16 (Karolchik et al. 2003). From this, we identified 602 splice variants, 523 polymorphic ORFs, and 51 ORFs that could not be aligned. These unique ORFs thus represent at least 8978 nonredundant genes in the starting MGC collection.

An Improved Gateway Recombinational Cloning Strategy

Cloning thousands of ORFs in dozens of different vectors would be virtually impossible using conventional restriction enzymes and DNA ligase technology. A robust, standardized, and flexible

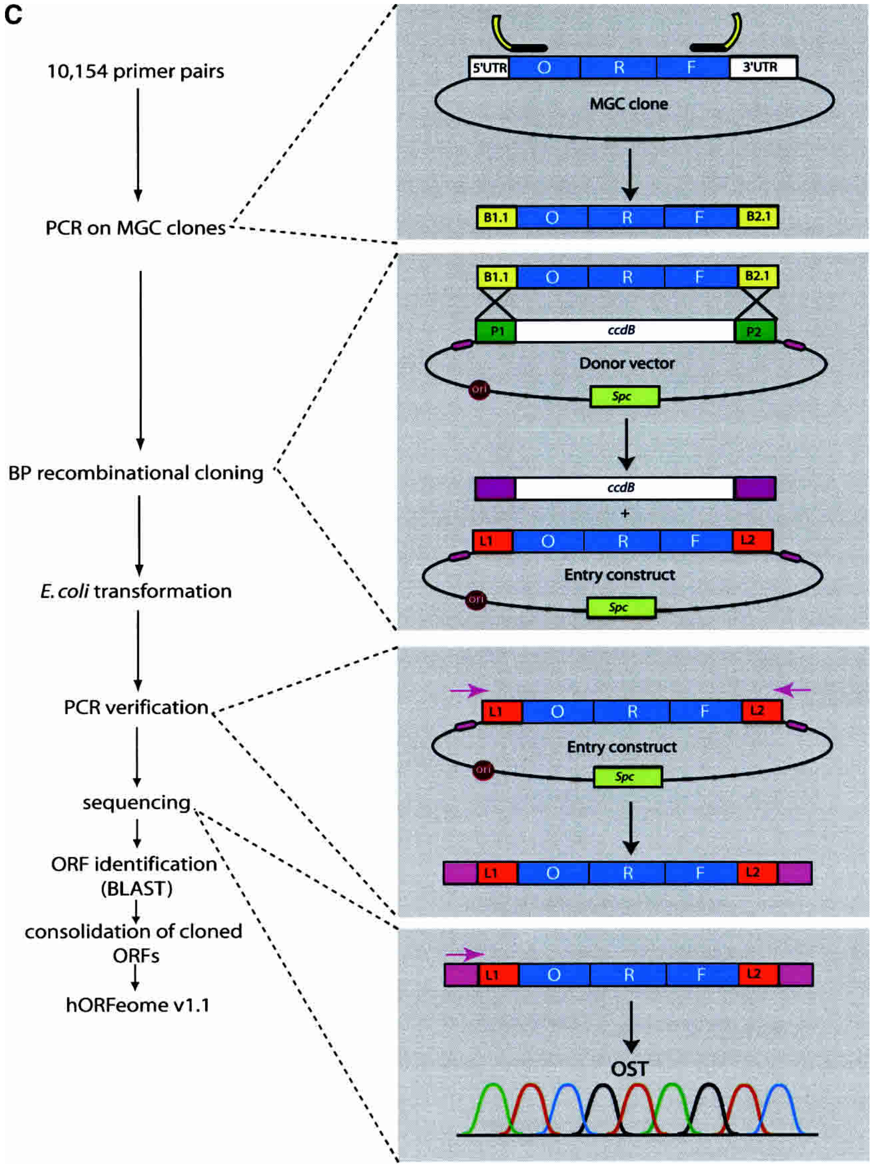

Figure 1 (*A*) Outline of the Gateway recombination reaction used for generation of hORFeome v1.1. PCR amplification of human ORFs (blue boxes) was performed on isolated MGC cDNA clones. In the depiction of the gene-specific primers, yellow nucleotides represent the altered *attB* recombination sites (*attB1.1* and *attB2.1*), and blue nucleotides represent the coding sequence in the ORF. PCR-amplified ORFs are cloned by a unidirectional recombinational cloning reaction via their flanking *attB1.1* and *attB2.1* recombination sites into the pDONR223 Gateway Donor vector. On the Donor vector, the universal Fwd and Rev sequencing primers, the origin of replication (ORI), and the spectinomycin (*Spc*) selectable marker are indicated. (*B*) Defining a first version of the human ORFeome starting from the MGC (Mammalian Gene Collection) Resource. The MGC contained 12,710 available full-length human cDNA clones at the time this project was begun. We designed pairs of primers for the PCR amplification of 10,154 distinct full-length ORFs. Asterisks indicate sequence polymorphisms. (*C*) Scheme for the generation of the hORFeome v1.1 resource. A total of 10,154 pairs of ORF-specific primers were designed to PCR amplify all nonredundant ORFs present in the MGC collection as of March 2004. Amplified ORFs were cloned into the pDONR223 Donor vector by BP recombinational cloning. PCR amplification of cloned inserts for sequence verification was then done using the Fwd and Rev primers that flank the cloning site. Sequencing was performed on the 5′ end with the Fwd primer to generate ORF Sequence Tags (OSTs), and ORFs were then identified by BLAST analysis against the complete MGC set of 10,154 nonredundant clones. Successfully cloned ORFs were consolidated onto new plates, generating the hORFeome v1.1 resource.

methodology is required for ORFeome-scale cloning, a need satisfied by the Gateway recombinational cloning (RC) system (Hartley et al. 2000; Simpson et al. 2000; Walhout et al. 2000b). Gateway mimics the site-specific recombination events of bacteriophage λ integration into and excision from the *Escherichia coli* chromosome. This RC system allows the automation of unidirectional ORF cloning into an initial resource vector (Donor vector) for the generation of Entry constructs (Fig. 1A,C). Gateway-cloned ORFs can subsequently be transferred from Entry constructs into expression vectors (Destination vectors). An appealing advantage of recombinational cloning is the generation of a resource of cloned ORFs available to the scientific community, which allows the use of an identical set of ORF constructs for different functional experiments performed by multiple investigators.

Compared with our previous efforts cloning the *C. elegans* ORFeome (Reboul et al. 2001, 2003; Lamesch et al. 2004), we improved the Gateway-cloning protocol (Table 1). We performed PCR amplification of ORFs using individually purified plasmids containing MGC cDNA clones as template DNA, as opposed to cDNA libraries. By using KOD Hot Start DNA polymerase (Novagen), an extremely high-fidelity thermostable DNA polymerase (Takagi et al. 1997), coupled with reducing the number of PCR cycles to 20, we greatly minimized the rate of PCR-induced mutations. For comparison, the *C. elegans* ORFeome project (Reboul et al. 2003; Lamesch et al. 2004) used a cDNA library as template, requiring 35 PCR cycles, and used Platinum *Taq* DNA Polymerase High Fidelity (Invitrogen). Although using purified individual plasmids rather than cDNA libraries as templates ensures that each template is adequately represented, a reduction of PCR cycles reduces the amount of DNA amplified, and hence might decrease the overall success rate of the Gateway reactions. To circumvent this potential difficulty, we also improved the Gateway cloning system (Fig. 1A,C).

First, increased efficiency of cloning PCR products via *attB*–*attP* (BP) recombination was achieved by altering the sequences of the Integrase (Int) binding sites within the *attB* sites (Table 1), so that they more closely resemble the consensus sequence, caacttnnt, in bacteriophage λ (Weisberg et al. 1983). The *attP1* and *attP2* sites remained unchanged. BP cloning of PCR products that contain these new sites was about fourfold more efficient than the cloning of PCR products containing the original *attB1* and *attB2* sites.

ORF-specific primers used for PCR contained at their 5′ end the Gateway consensus sequence for B1.1 (forward primer) or B2.1 (reverse primer), followed by 18–31 nucleotides of ORF-specific sequence (Table 1). The forward ORF-specific primer also incorporates a Kozak consensus nucleotide, G, at position −3 relative to the ATG, thereby allowing

Table 1. **Comparison of the Gateway Cloning System for the Current Human ORFeome Project to the Previous *C. elegans* ORFeome Project (Reboul et al. 2003)**

	***C. elegans* ORFeome**	**Human ORFeome**
PCR DNA Template	cDNA library	isloated cDNA clone
Number of PCR Cycles	35	20
Taq Polymerase	Platinum *Taq* DNA Polymerase High Fidelity	KOD Hot Start DNA polymerase
Mutation Rate	1/1500	1/35,000
AttB1 Site Sequence	B1:acaaGtttgtacaaaaaagCAggct	B1.1: acaaCtttgtacaaaaaagTTg
AttB2 Site Sequence	B2: acCactttgtacaagaaagCtgggt	B2.1: acAactttgtacaagaaagTtg
pDONR Vector	pDONR201	pDONR223
5′ ORF Specific Primer	5′-ggggacaagtttgtacaaaaaagcaggcttg(nnn)x	5′-ggggacaactttgtacaaaaaagttggcatg(nnn)x
3′ ORF Specific Primer	5′-ggggaccactttgtacaagaaagctgggta(nnn)x	5′-ggggacaactttgtacaagaaagttggg(t or c)a(nnn)x

efficient translational initiation from the initiation codon of the ORF.

Second, the efficiency of the BP reaction was improved by modifying the composition of the BP buffer BP3, which works best with the *attB1.1*/*attP1* and *attB2.1*/*attP2* recombination sites. Pilot experiments showed that cloning is about twofold more efficient with the new BP3 buffer than with the standard buffer.

Third, cloning was performed in a new Gateway Donor vector, pDONR223, which contains several improvements over the original Donor vector pDONR201 (Fig. 1A; Supplemental Fig. 1). In pDONR223, the universal Forward (Fwd) and Reverse (Rev) sequencing primer sites flank the *attP1* and *attP2* recombination sites, the replication origin has been improved to increase copy number, and the selectable marker is spectinomycin resistance (instead of kanamycin resistance in pDONR201), which permits the use of either ampicillin or kanamycin resistance markers in Destination vectors. The sequences of the *attP1* and *attP2* sites in pDONR223 are identical to those in the original pDONR201 Entry vector.

After BP recombinational cloning of PCR amplified ORFs into pDONR223 and transformation into *E. coli*, pools of 10–500 (average ~100) primary transformants were collected, and the resulting Entry constructs were subjected to ORF Sequence Tag (OST) analysis (Reboul et al. 2001) to determine insert identity (Fig. 1C). OSTs are long enough (300–500 nucleotides in a single read) to permit unique identification of the cognate ORF, while sparing the considerable expense of full insert sequencing.

OST Analysis

Our Gateway ORFeome cloning strategy was applied to the 10,154 unique ORFs present in the starting collection from MGC. An ORF was counted as successfully cloned if the OST obtained matched the sequence of the expected ORF. As observed previously with high-throughput (HT) Gateway cloning (Reboul et al. 2003), cloning success was dependent on ORF length, with shorter ORFs cloned more frequently than longer ones (Supplemental Fig. 2). The ORFs that were successfully cloned and sequenced were then rearrayed in order of increasing length, thereby generating version 1.1 of the human ORFeome (hORFeome v1.1). In total, hORFeome v1.1 contains 8076 cloned ORFs, a resource large enough to constitute a platform for reverse proteomics. These 8076 cloned ORFs include 413 splice variants, 362 polymorphisms, and 38 ORFs that could not be aligned to the human genome sequence (Golden Path release hg16), and thus comprise at least 7263 nonredundant human genes.

hORFdb, the Human ORFeome Project Database

The data pertaining to hORFeome v1.1, as well as future versions of our human ORFeome cloning project, are stored and integrated in the human ORF database (hORFdb; http://horfdb.dfci.harvard.edu). hORFdb will serve as a central data repository for the scientific community regarding availability and quality of cloned human ORFs. The hORFdb database, modeled on WorfDB (Worm ORFeome DataBase) for integration and dissemination of information on the *C. elegans* ORFeome (Vaglio et al. 2003), contains information for each of the 10,154 cloning attempts, including the sequences of the PCR primers, OSTs for each ORF, accession numbers, and links to various NCBI databases. All OSTs are deposited in GenBank.

Quality Assessment of hORFeome v1.1

In hORFeome v1.1, each Entry construct represents a mini-pool of PCR products amplified from a unique MGC cDNA, rather than a single-colony isolate. This strategy facilitates subsequent genome-scale reverse proteomics approaches, without the expense of having individually isolated clones (Brasch et al. 2004). Each mini-pool contains exact copies of the original full-length coding sequence for each cDNA, but it is possible that byproducts generated during the cloning may be present in the mini-pool. First, although PCR conditions were optimized, mutations can still occur during the PCR amplification. Second, inactivating mutations in the *ccdB* toxic selectable marker can give rise to viable transformants in the absence of any transfer of PCR inserts into the Donor vector. Finally, as in any HT approach, well-to-well cross contaminations can occur during processing.

To assess the overall quality of hORFeome v1.1, we analyzed representative subsets of mini-pools of cloned ORFs. A first test assessed the misincorporation rate by performing sequence analysis on 88 isolated colonies (11 different randomly selected ORFs, eight isolated colonies per ORF, lengths ranging from 777 to 1260 bp). Only good-quality sequences (PHRED score for each nucleotide >20 over more than 100 nucleotides) underwent consideration. The sequence analysis had two parts as follows: (1) examination of the ORF between the primer sequences; (2) examination of only the original PCR primer sequences.

For the first part, of ~70,000 nucleotides sequenced, the misincorporation rate after 20 cycles of PCR was one nucleotide substitution every 35,000 nucleotides using the KOD polymerase versus one substitution every 2000 nucleotides with Platinum Taq High Fidelity polymerase. The mutation rate observed for clones during the *C. elegans* ORFeome project was one mutation every 1500 nucleotides (Table 1; Reboul et al. 2003).

For the second part, the primer analysis, eight of the 176 primer sequences examined did not generate readable sequence. Of the remainder, 23 (14%) were mutated, arising from errors in primer synthesis. The mutations observed for these 23 primers are (1) 14 frameshifts due to nucleotide deletion; (2) six missense mutations; and (3) three silent mutations. Similar rates of primer mutation were obtained previously for both the *C. elegans* ORFeome project (9.8%; Reboul et al. 2003) and the *C. elegans* promoterome project (20.9%; Dupuy et al. 2004). On average, there were 6.1 exact matches to the cDNA template of the eight isolates in each mini-pool examined.

A second test assessed the accuracy of cloning by measuring insert length on three sets of 94 individual ORF mini-pools (282 total), spanning lengths of from 789 to 894 (short), 1590 to 2004 (medium), and 2487 to 3177 (long) nucleotides (Supplemental Fig. 2; Supplemental Table 1: plates 11006 [small], 11023 [medium], 11025 [long]). By comparison, the average length of coding sequences among the initial 10,154 unique ORFs is 1135 nucleotides. Twelve single colonies were isolated for each of the selected 282 ORF mini-pools, 3384 clones in total. For each clone, PCR amplification was performed using the universal Fwd and Rev primers.

These PCR products were analyzed by electrophoresis, and the length of the observed products was compared with their expected length. Of the 3384 ORFs, 19 did not give rise to a PCR product, whereas four products appeared as multiple bands, likely arising from cross-contamination during colony isolation. For the remaining 3361 PCR products, 3265 (97%) had the expected length, whereas 96 (3%) of them showed a length different than expected (a detectable difference is ~10% of the total length of the PCR product). There was at least one clone of the correct size band obtained for each of the 282 ORF mini-pools. Of the products of incorrect size, 50% belonged to the long set of ORFs.

We next sequenced two PCR products for each ORF of the correct length (564 PCR products altogether) as well as all of the 96 incorrectly sized PCR products. Of the 523 PCR products that produced readable sequence from the set of 564 products of correct length, 99% matched the sequence of the expected ORF. Six (~1%) matched the sequence of other ORFs, likely arising from cross-well contamination. Turning to the 96 incorrectly sized PCR products, 14 did not give readable sequence. All of the remaining 82 represented abnormal PCR products arising from aberrant primer annealing within the target ORF. These aberrant ORFs were truncated at the 5′ end (40), the 3′ end (39), or both ends (3).

Overall, these quality assessments indicate that the hORFeome v1.1 mini-pools are of high quality, making them amenable to reverse proteomics approaches.

hORFeome v1.1 as a Platform for Proteome-Wide Protein Expression

Having the hORFeome v1.1 resource in hand permits rapid straightforward cloning of thousands of ORFs into multiple protein expression vectors. We tested the use of hORFeome v1.1 for HT protein expression. The same 282 ORF mini-pools used above for quality assessment were transferred by automated methods from the Entry constructs into Destination vectors for two different protein expression systems as follows: (1) in *E. coli* as amino-terminal His6-tag fusion (hexa-histidine tag), and (2) in mammalian cells as a carboxyl-terminal GFP (green fluorescent protein) fusion where translation begins at the authentic ATG initiation codon of the ORF.

We examined bacterial protein expression by Western blot analysis using an antibody against His6, observing a band of the appropriate size 55% of the time (Table 2; Fig. 2A; Supplemental Table 1). The rate of successful protein expression decreased with increasing ORF length, as is commonly observed with recombinant expression from bacteria, with the set of short lengths showing 79% successful expression of proteins in the 30–35 kDa range, the medium set showing 57% successful expression of proteins in the 55–69 kDa range, and the long set showing 30% successful expression of proteins in the 90–114 kDa range (Fig. 2A). These results compare favorably with previous proteome-scale expression of human proteins in bacteria (Braun et al. 2002).

Table 2. Summary of the Proteome-Scale Protein Expression Data

Short ORFs	
GFP-Mammalian Expression	His-Bacterial Expression
67/94	74/94
71%	79%
Medium ORFs	
GFP-Mammalian Expression	His-Bacterial Expression
74/94	54/94
79%	57%
Long ORFs	
GFP-Mammalian Expression	His-Bacterial Expression
51/94	28/94
54%	30%
Total	
GFP-Mammalian Expression	His-Bacterial Expression
192/282	156/282
68%	55%

The same sets of proteins expressed in 293T mammalian cells produced a detectable signal 68% of the time (71%, 79%, and 54% success for short, medium, and long ORF sets respectively; Table 2; Fig. 2B; Supplemental Table 1). Most of the failures of expression in mammalian cells likely arise from poor transfection of DNA into cells.

Overall, the high success rate obtained for protein expression suggests that the hORFeome v1.1 is a useful resource for the expression of proteins at proteome scale in multiple protein expression systems.

DISCUSSION

The generation of flexible resources of cloned human ORFs represents a necessary antecedent for human systems biology (Vidal 2001; Rual et al. 2004a). The production here of the hORFeome v1.1 provides the starting resource for this objective.

The human ORFeome is a dynamic and flexible resource that represents a long-term ongoing effort. The hORFeome v1.1 presented here is merely the first step of the human ORFeome cloning project, and many challenges remain before its completion. Similarly to the worm ORFeome project (Lamesch et al. 2004), new ORFs will be added to the human ORFeome as improved human genome annotation becomes available, such as the recently released H-invDB platform (Imanishi et al. 2004), and as additional cDNA resources are generated. Further attempts at cloning the remaining ORFs not yet captured will also rely on improved algorithms for computational gene prediction (Reboul et al. 2003; Lamesch et al. 2004). Ultimately, a complete ORFeome should contain nearly all alternatively spliced forms as well as most relevant polymorphic ORF sequences.

Ideally, a comprehensive ORFeome would contain sequence-verified single colony isolates for all ORFs. To achieve the throughput needed to generate an ORFeome collection useful for subsequent reverse proteomics such as HT-Y2H, we chose to bypass single colony isolation of individual ORF clones. Each Entry construct in hORFeome v1.1 actually is a mini-pool corresponding to PCR-amplified inserts of the expected ORFs, in addition to any by-products (e.g., PCR-induced mutations) produced during cloning. Although individual clones can be isolated for detailed single-gene studies (Reboul et al. 2003; Lamesch et al. 2004), the generation of wild-type isolated clones for each and every ORF (version 2.1) is cumbersome, costly, and time consuming (Brasch et al. 2004). Our in-depth analyses of representative subsets of

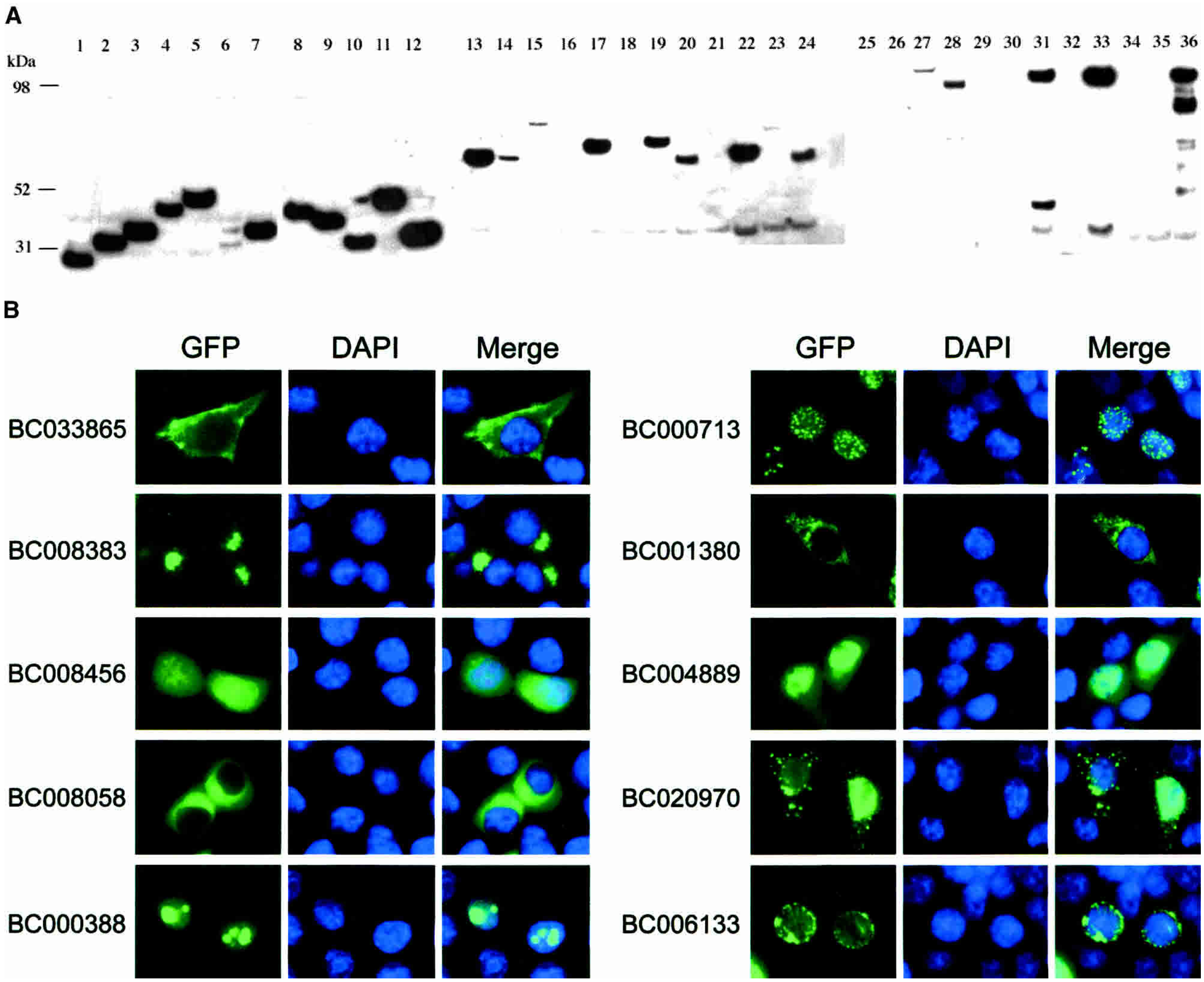

Figure 2 (*A*) Recombinant protein expression in *E. coli*. The 282 Entry constructs were transferred into a His6 N-terminal fusion vector (pDEST17). Representative samples from the 282 small, medium, and long ORFs are shown. (Lanes *1–12*) Plate 11006, Row F01–12, small ORFs (*1*) BC007838; (*2*) BC007355; (*3*) BC027953; (*4*) BC027899; (*5*) BC021719; (*6*) BC038108; (*7*) BC002826; (*8*) BC030277; (*9*) BC005991; (*10*) BC036723; (*11*) BC025403; (*12*) BC025760; (lanes *13–24*) Plate 11023, Row A1–12, medium ORFS (*13*) BC001061; (*14*) BC000453; (*15*) BC000017; (*16*) BC001167; (*17*) BC001150; (*18*) BC000480; (*19*) BC001221; (*20*) BC001142; (*21*) BC002472; (*22*) BC000723; (*23*) BC000770; (*24*) BC001665; (lanes *25–36*) Plate 11025, Row C1–12, large ORFS (*25*) BC037313; (*26*) BC035818; (*27*) BC007670; (*28*) BC007897; (*29*) BC012177; (*30*) BC037491; (*31*) BC005033; (*32*) BC032597; (*33*) BC036216; (*34*) BC001571; (*35*) BC012064; (*36*) BC034237. The positions of molecular weight markers (31–98 kDa) are indicated. All visible proteins migrate at the expected size. (*B*) Recombinant protein expression in mammalian cells. The 282 Entry constructs were transferred into a GFP C-terminal fusion vector (pcDNA-DEST47). Expression of GFP fusion proteins were assessed in transiently transfected 293T cells. Shown are 10 representative images of GFP fusion protein expression, with GFP images (*left*) showing the distribution of the fusion proteins, DAPI images (*middle*) indicating nuclei, and the GFP/DAPI merged images (*right*). MGC clone numbers are indicated to the *left* of each panel.

the Entry construct collection, plus our pilot protein-expression experiments, demonstrate that hORFeome v1.1 is a resource suitable for reverse proteomics studies. Numerous studies done with the *C. elegans* ORFeome v1.1 strongly support the validity of a mini-pool ORFeome cloning strategy (Brasch et al. 2004; Li et al. 2004).

The hORFeome v1.1 can also be used for the generation of an ORFeome library, as was done with the *C. elegans* ORFeome v1.1 (Reboul et al. 2003). To generate an ORFeome library, all or specific subsets of ORF Entry constructs are transferred into a suitable Destination vector; then, all Destination constructs are pooled together. ORFeome libraries are normalized compared with conventional cDNA libraries, making them a valuable resource for reverse proteomics applications such as protein–protein interaction mapping projects (Li et al. 2004).

All information pertaining to hORFeome v1.1 is available in the hORFdb database (http://horfdb.dfci.harvard.edu), and the physical resource of ORF Entry minipools is available from Open Biosystems Inc. (http://www.openbiosystems.com). The distribution of the hORFeome v1.1 resource among the biological community will meet the needs of systems biologists, and should satisfy conventional molecular biologists as well. The hORFeome v1.1 will provide the foundation for future high-throughput protein analyses, working toward a greater understanding of complex biological systems.

METHODS

Gateway Cloning of the Human ORFeome v1.1

At the start of the human ORFeome project, the MGC collection contained 12,710 human cDNA clones. The nucleotide sequences of the available cDNA clones were collected from the NCBI Web site, and their coding sequences were compared with each other, resulting in the identification of 10,154 unique full-length nonredundant ORFs (length longer than 100 nucleotides). The number of nonredundant genes was also determined by aligning the sequences of the ORFs on the human genome. Alignment of ORFs to the genome sequence was done using data retrieved from the UCSC Web site http://genome.ucsc.edu.

For PCR amplification, both forward and reverse ORF-specific primers for each MGC clone were designed automatically using the OSP program (Hillier and Green 1991). The forward primer starts from A of the ATG initiation codon, whereas the reverse primer starts from the second nucleotide of the termination codon. The reverse *attB2.1* primers do not contain the last nucleotide of the termination codon, so as to allow generation of C-terminal fusion proteins. Primers arbitrarily 20 nucleotides long were chosen for 769 clones that have exceptionally high GC content. Cloning was successful for 542/769 (75%) of the high GC content clones, versus 7534/9385 (81%) of other clones. PCR was performed in 25 µL reactions containing 1 unit of KOD Hot Start DNA polymerase according to the manufacturer (Novagen).

The hORFeome v1.1 was generated essentially as described (Reboul et al. 2003) with minor changes to accommodate improved PCR conditions and an improved Donor vector. The sequence and the map of the new pDONR223 vector are available in Supplemental Figure 1. A BP recombination reaction contains 2 µL of 5× BP3 buffer; 2 µL of BP clonase; 2 µL of pDONR223 (75 ng/µL); 2 µL of PCR product (2–200 ng/µL); 2 µL H_2O. The 5× BP3 buffer consists of 100 mM Tris-Cl (pH 7.5); 20 mM EDTA; 30 mM spermidine-HCl; 25% glycerol; 225 mM NaCl. LR reactions were performed as described previously with minor changes (Reboul et al. 2003; Rual et al. 2004b). An LR recombination reaction consists of 1 µL of 5× LR buffer, 0.5 µL of LR clonase, 1 µL of Destination vector (75 ng/µL), 1.5 µL of Entry construct mini-prep DNA (~40 µg/ml), and 1 µL TE (Tris-EDTA) buffer.

Transformations of BP and LR products were done in liquid cultures, with antibiotic selection of spectinomycin at 50 µg/mL (BP) or ampicillin at 100 µg/mL (LR). To estimate the overall efficiency of both the Gateway reaction and bacterial transformations, an aliquot from one well of each 96-well plate was also transferred to solid medium, and the number of antibiotic resistant colonies determined. Cloned ORFs were PCR amplified using the universal Fwd and Rev primers and the resulting PCR product sequenced at the 5′ end with the Fwd primer, generating an ORF Sequence Tag (OST). OSTs were accepted only if their average PHRED score was at least 20 over a minimum of 200 nucleotides. An ORF is counted as cloned if the OST obtained matches the expected sequence of the ORF for the corresponding template cDNA from a specific plate and well position.

Protein Expression in *E. coli*

The 282 randomly selected Entry constructs were cloned via an LR reaction into pDEST17 vector (Invitrogen) for expression in *E. coli* (Braun et al. 2002). The resulting products were transformed into *E. coli* BL21 Star (DE)pLysS strain (Invitrogen), and heterologous protein expression was induced with 1 mM IPTG. Cells were lysed in SDS–polyacrylamide sample dye, followed by polyacrylamide gel electrophoresis. Proteins were transferred to PVDF membranes and standard Western blotting was performed. Recombinant proteins expressed from the pDEST17 vector incorporate an N-terminal His6 tag, which was detected by immunoblotting first with the anti-polyhistidine His-1 mouse monoclonal primary antibody (Sigma) at 1:2000 dilution, and then with a goat anti-mouse HRP-conjugated secondary antibody (Calbiochem) at 1:1000 dilution. Visualization was performed using enhanced chemiluminescence (Amersham).

Protein Expression in Mammalian Cells

The 282 randomly selected Entry constructs were cloned via an LR reaction into the pcDNA-DEST47 vector, which contains GFP as a C-terminal fusion tag with ORF expression under the control of the cytomegalovirus (CMV) promoter. The 293T cells were transfected in 96-well format using Lipofectamine 2000 according to the manufacturer (Invitrogen). Two days later, cells were fixed with 3.7% formalin for 20 min, followed by staining of nuclei with 4′,6-diamidino-2-phenylindole (DAPI, Sigma). Expression of GFP fusion proteins was imaged using a Nikon ECLIPSE (TE300) microscope.

ACKNOWLEDGMENTS

We thank Ed Benz, Stan Korsmeyer, David Livingston, Priya McCue, Jane Song, and the DFCI Strategic Planning Initiative for support, the NIH Mammalian Gene Collection program for making the MGC collection available, Ian Dunham and David Bentley for suggesting the use of the KOD Hot Start DNA polymerase, members of the Vidal Lab and the participants of the ORFeome meeting for discussions, and Carlene Fraughton for technical support. This work was supported by the High-Tech Fund of the Dana-Farber Cancer Institute (S. Korsmeyer), by an Ellison Foundation grant awarded to M.V., by NCI U01 CA105423-01, and by EMBO fellowship awarded to P.O.V.

REFERENCES

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., et al. 2000. The genome sequence of *Drosophila melanogaster. Science* **287:** 2185–2195.

Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana. Nature* **408:** 796–815.

Boone, C. and Andrews, B. 2003. ORFeomics: Correcting the wiggle in worm genes. *Nat. Genet.* **34:** 8–9.

Brasch, M.A., Hartley, J.L., and Vidal, M. 2004. ORFeome cloning and systems biology: Standardized mass production of the parts from the parts-list. *Genome Res.* (this issue).

Braun, P., Hu, Y., Shen, B., Halleck, A., Koundinya, M., Harlow, E., and LaBaer, J. 2002. Proteome-scale purification of human proteins from bacteria. *Proc. Natl. Acad. Sci.* **99:** 2654–2659.

The *C. elegans* Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science* **282:** 2012–2018.

Collins, F.S., Green, E.D., Guttmacher, A.E., and Guyer, M.S. 2003. A vision for the future of genomics research. *Nature* **422:** 835–847.

Dupuy, D., Li, Q.-R., Deplancke, B., Boxem, M., Hao, T., Lamesch, P., Sequerra, R., Bosak, S., Doucette-Stamm, L., Hope, I.A., et al. 2004. A first version of the *Caenorhabditis elegans* promoterome. *Genome Res.* (this issue).

Gavin, A.C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.M., Cruciat, C.M., et al. 2002. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415:** 141–147.

Gibbs, R.A., Weinstock, G.M., Metzker, M.L., Muzny, D.M., Sodergren, E.J., Scherer, S., Scott, G., Steffen, D., Worley, K.C., Burch, P.E., et al. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428:** 493–521.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., et al. 1996. Life with 6000 genes. *Science* **274:** 546, 563–567.

Hartley, J.L., Temple, G.F., and Brasch, M.A. 2000. DNA cloning using in vitro site-specific recombination. *Genome Res.* **10:** 1788–1795.

Hillier, L. and Green, P. 1991. OSP: A computer program for choosing PCR and DNA sequencing primers. *PCR Methods Appl.* **1:** 124–128.

Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S.L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., et al. 2002. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415:** 180–183.

Ideker, T., Galitski, T., and Hood, L. 2001. A new approach to decoding life: Systems biology. *Annu. Rev. Genomics Hum. Genet.* **2:** 343–372.

Imanishi, T., Itoh, T., Suzuki, Y., O'Donovan, C., Fukuchi, S., Koyanagi, K.O., Barrero, R.A., Tamura, T., Yamaguchi-Kabata, Y., Tanino, M., et al. 2004. Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *PLoS Biol.* **2:** E162.

Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S., Hinrichs, A., Lu, Y.T., Roskin, K.M., Schwartz, M., Sugnet, C.W., Thomas, D.J., et al.

2003. The UCSC Genome Browser Database. *Nucleic Acids Res.* **31:** 51–54.
Kitano, H. 2002. Systems biology: A brief overview. *Science* **295:** 1662–1664.
Lamesch, P., Milstein, S., Hao, T., Rosenberg, J., Li, N., Sequerra, R., Bosak, S., Doucette-Stamm, L., Vandenhaute, J., Hill, D., et al. 2004. *C. elegans* ORFeome version 3.1: Increasing the coverage of ORFeome resources with improved gene predictions. *Genome Res.* (this issue).
Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.
Li, S., Armstrong, C.M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P.O., Han, J.D., Chesneau, A., Hao, T., et al. 2004. A map of the interactome network of the metazoan *C. elegans*. *Science* **303:** 540–543.
Marsischky, G. and LaBaer, J. 2004. Many paths to many clones: A look at high-throughput cloning methods. *Genome Res.* (this issue).
Reboul, J., Vaglio, P., Tzellas, N., Thierry-Mieg, N., Moore, T., Jackson, C., Shin-i, T., Kohara, Y., Thierry-Mieg, D., Thierry-Mieg, J., et al. 2001. Open-reading-frame sequence tags (OSTs) support the existence of at least 17,300 genes in *C. elegans*. *Nat. Genet.* **27:** 332–336.
Reboul, J., Vaglio, P., Rual, J.F., Lamesch, P., Martinez, M., Armstrong, C.M., Li, S., Jacotot, L., Bertin, N., Janky, R., et al. 2003. *C. elegans* ORFeome version 1.1: Experimental verification of the genome annotation and resource for proteome-scale protein expression. *Nat. Genet.* **34:** 35–41.
Rual, J.F., Ceron, J., Koreth, J., Hao, T., Nicot, A.S., Hirozane-Kishikawa, T., Vandenhaute, J., Orkin, S., Hill, D.E., van den Heuvel, S., et al. 2004a. Towards improving *Caenorhabditis elegans* phenome mapping with an ORFeome-based RNAi library. *Genome Res.* (this issue).
Rual, J.F., Hill, D.E., and Vidal, M. 2004b. ORFeome projects: Gateway between genomics and omics. *Curr. Opin. Chem. Biol.* **8:** 20–25.
Simpson, J.C., Wellenreuther, R., Poustka, A., Pepperkok, R., and Wiemann, S. 2000. Systematic subcellular localization of novel proteins identified by large-scale cDNA sequencing. *EMBO Rep.* **1:** 287–292.
Strausberg, R.L., Feingold, E.A., Grouse, L.H., Derge, J.G., Klausner, R.D., Collins, F.S., Wagner, L., Shenmen, C.M., Schuler, G.D., Altschul, S.F., et al. 2002. Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. *Proc. Natl. Acad. Sci.* **99:** 16899–16903.
Takagi, M., Nishioka, M., Kakihara, H., Kitabayashi, M., Inoue, H., Kawakami, B., Oka, M., and Imanaka, T. 1997. Characterization of DNA polymerase from *Pyrococcus sp.* strain KOD1 and its application to PCR. *Appl. Environ. Microbiol.* **63:** 4504–4510.
Vaglio, P., Lamesch, P., Reboul, J., Rual, J.F., Martinez, M., Hill, D., and Vidal, M. 2003. WorfDB: The *Caenorhabditis elegans* ORFeome Database. *Nucleic Acids Res.* **31:** 237–240.
Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. 2001. The sequence of the human genome. *Science* **291:** 1304–1351.
Vidal, M. 2001. A biological atlas of functional maps. *Cell* **104:** 333–339.
Walhout, A.J. and Vidal, M. 2001. Protein interaction maps for model organisms. *Nat. Rev. Mol. Cell. Biol.* **2:** 55–62.
Walhout, A.J., Boulton, S.J., and Vidal, M. 2000a. Yeast two-hybrid systems and protein interaction mapping projects for yeast and worm. *Yeast* **17:** 88–94.
Walhout, A.J., Temple, G.F., Brasch, M.A., Hartley, J.L., Lorson, M.A., van den Heuvel, S., and Vidal, M. 2000b. GATEWAY recombinational cloning: Application to the cloning of large numbers of open reading frames or ORFeomes. *Methods Enzymol.* **328:** 575–592.
Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.
Weisberg, R.A., Enquist, L.W., Foeller, C., and Landy, A. 1983. Role for DNA homology in site-specific recombination. The isolation and characterization of a site affinity mutant of coliphage λ. *J. Mol. Biol.* **170:** 319–342.
Zhu, H., Bilgin, M., Bangham, R., Hall, D., Casamayor, A., Bertone, P., Lan, N., Jansen, R., Bidlingmaier, S., Houfek, T., et al. 2001. Global analysis of protein activities using proteome chips. *Science* **293:** 2101–2105.
Ziauddin, J. and Sabatini, D.M. 2001. Microarrays of cells expressing defined cDNAs. *Nature* **411:** 107–110.

WEB SITE REFERENCES

http://genome.ucsc.edu; UCSC Genome Browser Web site.
http://horfdb.dfci.harvard.edu; Source for human ORFeome data.
http://www.openbiosystems.com; Source for request of hORFeome v1.1 Entry constructs.

Received July 7, 2004; accepted in revised form August 19, 2004.