The ORFeome Collaboration: a genome-scale human ORF-clone resource

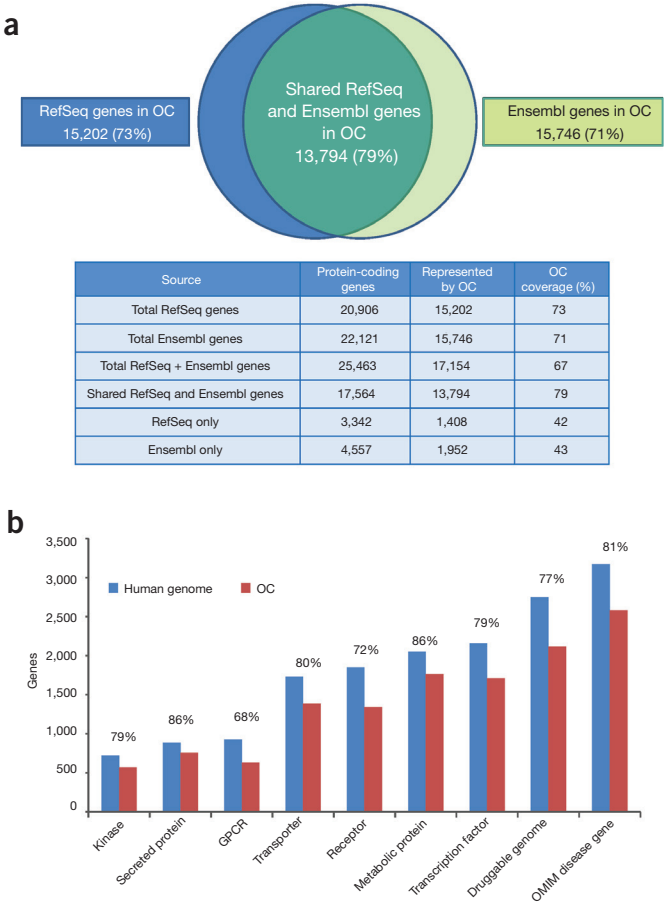
To the Editor: Here we describe the ORFeome Collaboration (OC) open reading frame (ORF) clone collection, created by the OC (http://www.orfeomecollaboration.org/), an international collaboration of academic and commercial groups committed to providing genome-scale clone resources for human genes via worldwide commercial and academic clone distributors.

Proteins are the predominant functional modules determining the fate of cells, tissues and organisms. An encyclopedic understanding of cellular physiology requires protein expression for protein-protein interaction screening, cellular functional screening, validation of knockout and knockdown phenotypes, and numerous other approaches. Performing such studies on individual proteins or at the proteome scale requires a comprehensive collection of human protein expression clones.

Our collection comprises ORF clones (**Supplementary Note**) and covers 17,154 RefSeq and Ensembl genes, nearly 73% of human RefSeq genes (http://www.ncbi.nlm.nih.gov/refseq/rsg/) and 79% of the highly curated Consensus Coding DNA Sequence Project (CCDS) human genes (http://www.ncbi.nlm.nih.gov/CCDS/CcdsBrowse.cgi) (**Fig. 1a** and **Supplementary Data**). The collection includes clones of transcript variants for 6,304 (37%) of those genes. All major functional categories of human genes are substantially represented (**Fig. 1b**).

All clones are provided in the Gateway vector format (Life Technologies), permitting high-throughput, precise and directional transfer of ORFs to a large variety of vectors for protein expression in biological systems such as *Escherichia coli*, yeast and mammals or using cell-free protein expression[1] (**Supplementary Note**). OC clones were generated primarily by PCR amplification of the ORF from full-length, sequence-verified human cDNA clones of the Mammalian Gene Collection[2] or the German cDNA Consortium[3]; ORFs were also prepared by directed RT-PCR cloning[4] or DNA synthesis[2]. All 5′ and 3′ untranslated regions were excluded, permitting direct expression of ORFs as fusions to amino- or carboxy-terminal polypeptides, or as native protein, after transfer to a Gateway-expression vector[1]. The clones are designed to maintain the correct reading frame for both amino- and carboxy-fusion proteins. Among all genes represented in the OC collection, 64% of clones are without stop codons, 5% have stop codons, and 31% are present in both versions.

Each OC clone was isolated from a single colony and is fully sequenced. Individual clone sequences have been deposited in the GenBank, EMBL and DDBJ databases. The OC website provides a searchable database with annotation of all OC clones, their respective genes, and clone confidence levels based on CCDS and RefSeq annotations (**Supplementary Note**) along with links to the UCSC and RIKEN browsers (http://genome.ucsc.edu/cgi-bin/hgGateway and http://fantom.gsc.riken.jp/zenbu/gLyphs/#config), which provide graphical representations of the gene structures and transcripts. OC clones are distributed via a good faith agreement, giving unrestricted clone access to all scientists worldwide. The OC website lists OC clone distributors.



Source	Protein-coding genes	Represented by OC	OC coverage (%)
Total RefSeq genes	20,906	15,202	73
Total Ensembl genes	22,121	15,746	71
Total RefSeq + Ensembl genes	25,463	17,154	67
Shared RefSeq and Ensembl genes	17,564	13,794	79
RefSeq only	3,342	1,408	42
Ensembl only	4,557	1,952	43

Figure 1 | RefSeq and Ensembl genes and functional gene categories represented in the OC. (**a**) Numbers of protein-coding genes represented in the OC collection from RefSeq (blue) and Ensembl (green) gene catalogs. The table summarizes these numbers, together with OC coverage for RefSeq-only and Ensembl-only genes. (**b**) Numbers of human RefSeq genes represented in the OC collection versus in the human genome, compared in nine functional categories; percentages of genes in the OC are presented above the bars. The methods used to calculate the gene numbers in each category are explained in the **Supplementary Note** and contrasted to the standard Gene Ontology categories. An expanded list of the top Gene Ontology categories is also provided in the **Supplementary Note.** The data underlying the graphs are provided as **Supplementary Data**.

The value of the OC resource has been demonstrated in numerous studies covering a broad range of applications. These include large-scale binary protein-protein interaction mapping[5], production of recombinant human proteins[6], mapping of co-complex associations, fluorescent protein tagging for human protein localization in mammalian cells and microscopy-based functional screening of proteins, development of disease-specific protein interaction

networks, coexpression to rescue RNA interference– or CRISPR-CAS9–induced reduction of endogenous transcripts, and expression of ORFs carrying a mutation of interest to allow measurement of the mutation effect in the absence of the wild-type background.

High-level gene coverage, combined with the versatility of Gateway cloning, and full access to OC clones make this collection a unique and valuable resource for the scientific community that should aid in the functional characterization of new protein targets and testing of disease-relevant mutations on a large scale. The OC resource will continue to be expanded in the future to increase human gene coverage, provide additional isoforms where available, provide clones with medically relevant mutations and add additional species, including ORFs from *Xenopus* and *Drosophila*.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper (http://dx.doi.org/10.1038/nmeth.3776).

ACKNOWLEDGMENTS

The authors acknowledge valuable encouragement for initiating the OC from F. Collins (US National Institutes of Health) and E. Harlow (Harvard Medical School). Some of the cDNAs used as PCR templates for ORF cloning and other support were received from the German cDNA Consortium: K. Köhrer (University Düsseldorf, Germany), W. Ansorge (EMBL Heidelberg, Germany), H. Blöcker (Helmholtz Center Braunschweig, Germany), W. Mewes, C. Amid (Helmholtz Center Munich, Germany), J. Lauber, A. Bahr (Qiagen, Hilden, Germany), D. Heubner, R. Wambutt (Agowa, Berlin, Germany), B. Ottenwälder, B. Obermaier (Medigenomix, Ebersberg, Germany), H. Blum, H. Domdey (University Munich, Germany), I. Schupp, S. Bechtel and A. Poustka (DKFZ, Heidelberg, Germany). The German cDNA Consortium was funded by the Federal Ministry of Education and Research (BMBF) in the frame of the German Genome Project (DHGP) and the German National Genome Research Network (NGFN) programs (to S.W.). This work was supported by the Ellison Foundation (grant to M.V. and D.E.H.); the DFCI Institute (Sponsored Research funds to M.V. and D.E.H.); the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan (research grants to Y.H. at the RIKEN Omics Science Center and to P.C. at the RIKEN Center for Life Science Technologies); and the MEXT Genome Network Project (grant to Y.H.).

AUTHOR CONTRIBUTIONS

M.V., D.S.G., J.L., G.T. and D.E.H. founded the OC and conceived of the project. S.W., Y. Hu, G.B., P.C., I.D., Y. Hayashizaki, J.K., O.O., A.R., K.S.-A., J.S., R.W., M.V. and D.E.H. generated and contributed entry clones. S.W., Y. Hu, G.B., P.C., I.D., Y. Hayashizaki, S.H., B.K., J.K., C.K., S.L., A.R., K.S.-A., J.S., R.W., S.Y., M.V., J.L. and D.E.H. generated sequence-verified ORF clones. S.W., Y. Hu, G.B., I.D., T.H., O.H., S.H., A.H.-W., W.J., A.J., C.K., A.L., S.L., J.P., B.S., L.W. and S.Y. performed bioinformatics analysis and clone annotation. C.P., A.A., C.K., A.L., S.L., A.v.B.S. and S.Y. re-arrayed libraries and carried out quality control of ORF clones. C.P., P.H., M.H., A.A., M.B., J.W.H., B.K., C.L., S.L., J.M., T.M., O.O., J.P., A.R., C.S., B.S., A.v.B.S., T.W. and S.Y. performed archiving and distribution of the ORF clone resource to the public. S.W., C.P., P.H., M.H., M.D., O.H., S.L., J.P., C.S. and S.Y. carried out website and database development. S.W., M.H., M.V., D.S.G., J.L., G.T. and D.E.H. led, oversaw and steered development of the consortium and wrote the paper.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper (http://dx.doi.org/10.1038/nmeth.3776).

The ORFeome Collaboration: Stefan Wiemann[1,2], Christa Pennacchio[3], Yanhui Hu[4], Preston Hunter[5], Matthias Harbers[6,7], Alexandra Amiet[8], Graeme Bethel[9], Melanie Busse[10], Piero Carninci[7], Mark Diekhans[11], Ian Dunham[9], Tong Hao[12–14], J Wade Harper[15], Yoshihide Hayashizaki[16], Oliver Heil[2], Steffen Hennig[17], Agnes Hotz-Wagenblatt[2], Wonhee Jang[18], Anika Jöcker[1], Jun Kawai[16], Christoph Koenig[17], Bernhard Korn[19], Cristen Lambert[20], Anita LeBeau[21], Sun Lu[22,23], Johannes Maurer[17], Troy Moore[24], Osamu Ohara[25], Jin Park[5], Andreas Rolfs[26], Kourosh Salehi-Ashtiani[12–14], Catherine Seiler[5], Blake Simmons[21,24], Anja van Brabant Smith[8], Jason Steel[5], Lukas Wagner[18], Tom Weaver[10], Ruth Wellenreuther[1], Shuwei Yang[22], Marc Vidal[12–14], Daniela S Gerhard[27], Joshua LaBaer[5,28], Gary Temple[20] & David E Hill[12–14]

[1]Division of Molecular Genome Analysis, German Cancer Research Center (DKFZ), Heidelberg, Germany. [2]Genomics & Proteomics Core Facility, German Cancer Research Center (DKFZ), Heidelberg, Germany. [3]IMAGE Consortium, Lawrence Livermore National Laboratories, Livermore, California, USA. [4]Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA. [5]Virginia G. Piper Center for Personalized Diagnostics (VGPCPD), Biodesign Institute, Arizona State University, Tempe, Arizona, USA. [6]DNAFORM Inc., Tsurumi-ku, Yokohama City, Kanagawa, Japan. [7]Division of Genomic Technologies, RIKEN Center for Life Science Technologies, RIKEN Yokohama Institute, Tsurumi-ku, Yokohama, Kanagawa, Japan. [8]Dharmacon, GE Healthcare, Lafayette, Colorado, USA. [9]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, UK. [10]Source BioScience, Nottingham, UK. [11]UC Santa Cruz Genomics Institute, University of California, Santa Cruz, California, USA. [12]Center for Cancer Systems Biology (CCSB), Dana-Farber Cancer Institute, Boston, Massachusetts, USA. [13]Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA. [14]Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA. [15]Dana-Farber–Harvard Cancer Center (DFHCC) DNA Resource Core and Department of Cell Biology, Harvard Medical School, Boston, Massachusetts, USA. [16]RIKEN Preventive Medicine & Diagnosis Innovation Program, RIKEN Yokohama Institute, Wako, Saitama, Japan. [17]imaGenes GmbH, Berlin, Germany. [18]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland, USA. [19]Ressourcenzentrum für Genomforschung gGmbH, Berlin, Germany. [20]National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland, USA. [21]HudsonAlpha Institute of Biotechnology, Huntsville, Alabama, USA. [22]GeneCopoeia, Inc., Rockville, Maryland, USA. [23]Guangzhou FulenGen, Ltd., Guangdong, China. [24]Open Biosystems, Inc., Huntsville, Alabama, USA. [25]Kasusa DNA Research Institute, Kisarazu, Chiba, Japan. [26]Department of Biological Chemistry & Molecular Pharmacology, Harvard Institute of Proteomics, Harvard Medical School, Boston, Massachusetts, USA. [27]Office of Cancer Genomics, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, USA. [28]Department of Chemistry and Biochemistry, Arizona State University, Tempe, Arizona, USA. Correspondence should be addressed to D.E.H. (david_hill@dfci.harvard.edu), S.W. (s.wiemann@dkfz.de), G.T. (gftemple@gmail.com), M.H. (matthias.harbers@riken.jp), M.V. (marc_vidal@dfci.harvard.edu) or J.L. (joshua.labaer@asu.edu).

1. Walhout, A.J. *et al. Methods Enzymol.* **328**, 575–592 (2000).
2. MGC Project Team *et al. Genome Res.* **19**, 2324–2333 (2009).
3. Wiemann, S. *et al. Genome Res.* **11**, 422–435 (2001).
4. Yang, X. *et al. Nat. Methods* **8**, 659–661 (2011).
5. Rolland, T. *et al. Cell* **159**, 1212–1226 (2014).
6. Yu, X. & LaBaer, J. *Nat. Protoc.* **10**, 756–767 (2015).

TeraFly: real-time three-dimensional visualization and annotation of terabytes of multidimensional volumetric images

To the Editor: New sample preparation and high-throughput light-sheet microscopy techniques[1] are increasingly capable of generating multidimensional (3D and higher) images easily exceeding the terabyte size. This has posed a significant challenge for scalable interactive visualization and quantitative annotation of such big image data. A common practice is to design a data-streaming and visualization tool to supply and display small parts of an image volume when needed[2,3]. However, existing tools allow only 2D slice-based rendering of 3D image stacks. Such 2D approaches not only are time consuming and low throughput but also bring bias to the understanding of intrinsic 3D properties of bioimage data[4]. A free, open-source and cross-platform software tool for true 3D visualization and 3D annotation of very large multidimensional volumes is highly desired (**Supplementary Note 1**).