A comprehensive collection of chicken cDNAs

Citation for published version:
Boardman, PE, Sanz-Ezquerro, J, Overton, IM, Burt, DW, Bosch, E, Fong, WT, Tickle, C, Brown, WRA, Wilson, SA & Hubbard, SJ 2002, 'A comprehensive collection of chicken cDNAs', *Current biology : CB*, vol. 12, no. 22, pp. 1965-9. https://doi.org/10.1016/S0960-9822(02)01296-4

Digital Object Identifier (DOI):
10.1016/S0960-9822(02)01296-4

Published In:
Current biology : CB

Current Biology, Vol. 12, 1965–1969, November 19, 2002, ©2002 Elsevier Science Ltd. All rights reserved. PII S0960-9822(02)01296-4

A Comprehensive Collection of Chicken cDNAs

Paul E. Boardman,[1] Juan Sanz-Ezquerro,[2] Ian M. Overton,[1] David W. Burt,[3] Elizabeth Bosch,[4,7] Willy T. Fong,[4] Cheryll Tickle,[2] William R.A. Brown,[5] Stuart A. Wilson,[1] and Simon J. Hubbard[1,6]
[1]Department of Biomolecular Sciences
University of Manchester Institute of Science and Technology
P.O. Box 88
Manchester M60 1QD
United Kingdom
[2]The Wellcome Trust Biocentre
Medical Sciences Institute
University of Dundee
Dow Street
Dundee DD1 5EH
United Kingdom
[3]Department of Genomics and Bioinformatics
Roslin Institute
Roslin, Midlothian, EH25 9PS
United Kingdom
[4]Incyte Genomics
3160 Porter Drive
Palo Alto, California 94304
[5]Institute of Genetics
Nottingham University
Queen's Medical Centre
Nottingham, NG7 2UH
United Kingdom

Summary

Birds have played a central role in many biological disciplines, particularly ecology, evolution, and behavior. The chicken, as a model vertebrate, also represents an important experimental system for developmental biologists, immunologists, cell biologists, and geneticists. However, genomic resources for the chicken have lagged behind those for other model organisms, with only 1845 nonredundant full-length chicken cDNA sequences currently deposited in the EMBL databank. We describe a large-scale expressed-sequence-tag (EST) project aimed at gene discovery in chickens (http://www.chick.umist.ac.uk). In total, 339,314 ESTs have been sequenced from 64 cDNA libraries generated from 21 different embryonic and adult tissues. These were clustered and assembled into 85,486 contiguous sequences (contigs). We find that a minimum of 38% of the contigs have orthologs in other organisms and define an upper limit of 13,000 new chicken genes. The remaining contigs may include novel avian specific or rapidly evolving genes. Comparison of the contigs with known chicken genes and orthologs indicates that 30% include cDNAs that contain the start codon and 20% of the contigs represent full-length cDNA sequences. Using this dataset, we estimate that chickens have approximately 35,000 genes in total, suggesting that this number may be a characteristic feature of vertebrates.

[6]Correspondence: simon.hubbard@umist.ac.uk
[7]Present address: Five Prime Therapeutics, 951 Gateway Blvd., South San Francisco, California 94080.

Results and Discussion

The aim of the project was novel gene discovery, and this dictated the choices in sequencing strategy and tissues from which to construct libraries. For tissue selection, we took into account experimental research on chicken systems and the existing chicken EST data generated from other projects [1, 2]. The tissues from which clones were sequenced, numbers of unique reads for the whole project and each tissue, numbers of successful reads, and inter-library redundancy statistics are detailed in Table 1. For each tissue, we constructed one standard [3] and at least two normalized libraries in order to enhance gene discovery. The only exception to this was for the adult adipose tissue, for which normalization failed completely. The average length of the cDNA inserts was at 1200–1400 bp across all libraries; a sample size of 196 clones per library was used. We sequenced clones from the 5′ end and used poly A-trimmed libraries to enable future 3′ sequencing. Initially, 250,000 sequencing reads were assigned across the 21 tissues. Cross-tissue comparisons from these reads were used for determining which libraries provided the highest rate of gene discovery. When assigning the remaining 100,000 sequencing reads, we favored those libraries yielding the highest rate of gene discovery. For example, libraries from stage 20–21 embryos, ovaries, and chondrocytes were rich sources of novel genes and were extensively sequenced (Table 1). The rate of gene discovery was consistently highest in high-stringency normalized libraries, as expected. In some cases, technical difficulties in sequencing, picking clones, or obtaining transformants prevented some apparently high-quality libraries from being further investigated.

Sequencing statistics for the entire project are shown in Table 2. These statistics refer to data involving bioinformatic quality control, which removes chicken rRNAs, mitochondrial DNA, cloning vectors, and bacterial contamination, together with sequences shorter than 50bp. In total 323,670 EST sequences were obtained, with an average length of 671 bp and several long reads in excess of 1000 bp, as shown in Figure 1. EST sequences were annotated with the top BLASTX [4] hit to SWISS-PROT/TrEMBL [5] (full BLAST reports are available on the web site). In addition, these searches were used for assigning gene ontology (GO) [6] terms to 123,249 EST sequences and 45,740 contigs.

After vector removal and contamination checks, the sequences were assembled. Sequences were first clustered into 64,760 gene bins based on pairwise similarity from BLASTN searches. Gene bins were assembled with PHRAP (http://www.phrap.org), for which quality scores generated with the PHRED base-calling program were

Table 1. Tissues Selected and Statistics for EST Reads

Tissue	Sequences Obtained	Percent Redundancy	Percent Unique Singletons	Percent Tissue-Specific Singletons
Stage 36 trunks	13,448	42.8	28.4	10.6
Stage 36 limbs	19,088	50.8	27.6	10.1
Stage 36 hearts	8,647	48.0	23.1	7.9
Stage 36 heads	15,379	46.0	29.1	10.8
Stage 22 limbs	15,748	49.1	28.8	9.7
Stage 22 heads	17,915	33.7	40.6	17.0
Stage 20–21 whole embryos	26,381	34.9	42.0	23.5
Stage 10 whole embryos	12,579	49.6	28.2	8.6
16-day embryo brain	16,714	44.5	36.6	13.9
Adult ovary	29,720	46.6	35.0	20.1
Adult muscle	9,333	28.1	41.8	14.8
Adult chondrocytes	31,635	51.4	32.1	18.5
Adult small intestine	17,800	49.9	28.1	10.5
Adult pancreas	7,965	79.9	9.6	4.9
Adult liver	13,131	44.2	30.7	12.0
Adult kidney + adrenal	19,290	49.5	28.4	11.3
Adult heart	9,300	48.8	29.5	8.6
Adult brain—other parts	15,408	42.1	36.0	13.0
Adult brain—cerebrum	14,005	44.4	33.0	12.2
Adult brain—cerebellum	14,230	41.7	37.0	11.9
Adult adipose	2,672	34.1	34.8	11.5
Total	330,388			
Average		54.3	31.4	12.4

Sequence obtained refers to reads for which greater than 100 bp of high-quality cDNA, which also passed vector, mitochondrial DNA, and ribosomal RNA contamination checks, was obtained. Percent redundancy = percent of ESTs that are found in tissue more than once (or, if you prefer, percent singletons = 100 − percent redundancy). Percent unique singletons = percent of ESTs that have no SWISS-PROT/TrEMBL hits and are also singletons (see above). A high number represents a gene-rich library. Percent tissue-specific singletons = percent of ESTs that don't have SWISS-PROT/TrEMBL hits, are singletons, and don't match any other ESTs in any other tissues. A sign of a particularly rich source of novel genes (it is hoped), which is why many reads from ovary, chondrocytes, and 20-to-21-stage whole embryos were attempted.

used [7, 8]. This resulted in 85,486 individual contigs, of which 38,812 (45%) contained more than one clone and 46,674 (55%) were singletons. The average number of clones per contig is 3.8, and the average length of the contigs is 874 bp (Figure 1). Strikingly, a number of significantly larger contigs have been assembled; for example, contig 354630.6 is over 7 Kb long, corresponding to 86% of the full chicken myosin open reading frame and including more than 1500 bases of 3′ UTR (Figure 2).

We have estimated the number of cDNA clones that extend to the start codon for known chicken genes in our dataset by comparing the contigs with a set of 1845 nonredundant, full-length *Gallus gallus* cDNA sequences taken from the EMBL [9] databank. In total, 3181 of the 85,486 contigs had stringent BLASTN matches (greater than 98% identity over 100 bases), of which 996 (31%) extend completely to the start codon. In addition, 631 (20%) of these sequence matches extend from the start codon to the stop codon, suggesting that the complete coding sequence has been obtained for these genes. Extrapolating these figures to the complete dataset indicates that a large number of contigs contain component clones that encompass the start codon and may represent full-length cDNAs. Figure 3 illustrates the lengths of chicken gene coding sequences that the contigs cover completely; these sequences include many that extend up to and greater than 2 Kb in length. We obtained similar statistics by searching the contigs against a set of 1,676 nonredundant, full-length chicken sequences extracted from SWISS-PROT/TrEMBL, where 916 (37%) of the 2,442 matching contigs extend to within 5 amino acids of the N terminus of the matching chicken protein (96% identity over 30 amino acids).

In total, 31,005 (36%) of the contigs have a BLASTX match in the nonredundant SWISS-PROT/TrEMBL databank (with expectation values better than 10^{-6}), which corresponds to 14,801 different protein sequences. Given that more than 1,813 already have stringent matches to known chicken genes, this suggests that the database contains in excess of 13,000 sequences representing novel chicken genes that have orthologs in the known databanks. Of the remaining 54,481 contigs, a BLASTX search with the human peptidome dataset from ENSEMBL [10] revealed an additional 752 with homology to predicted human genes. The remaining 63% of

Table 2. Global Project Statistics for EST Sequencing

Individual Sequence Statistics	
Total number of reads attempted	350,980
Number of successful reads	339,314
Number of reads entering assembly	323,670
Percentage of reverse insert clones	2.3%
Average read length	671 bp
Contig Statistics	
Number of contigs	85,486
Singleton contigs	46,674 (55%)
Multi-component contigs	38,812 (45%)
Average number of reads/contig	3.8
Average contig read length	874 bp (1158 bp for multi-component contigs)

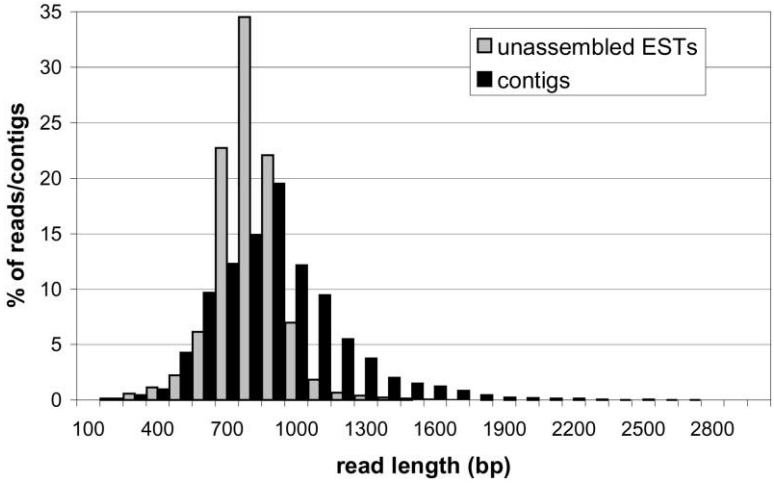

Figure 1. Sequence Length Statistics

Histogram of read lengths for unassembled ESTs and contig sequences.

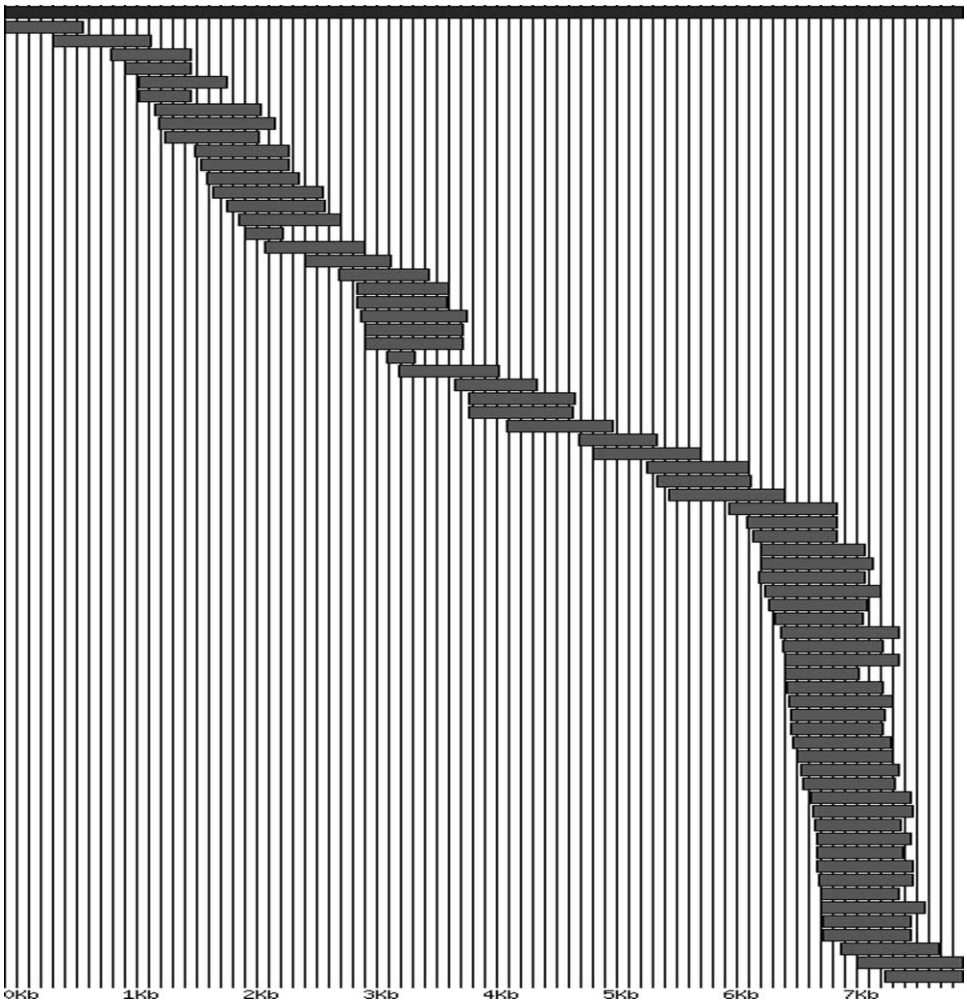

Figure 2. Contig Sequence Representing Chicken Nonmuscle Myosin Heavy Chain

This is a graphical representation of the composition of contig sequence 354630.6. The long bar at the top represents the contig sequence. Smaller bars represent alignment positions of the component EST sequences. The spacing of vertical lines is equivalent to 100 bp in the gapped alignment. (Chicken myosin accession number, Q02015).

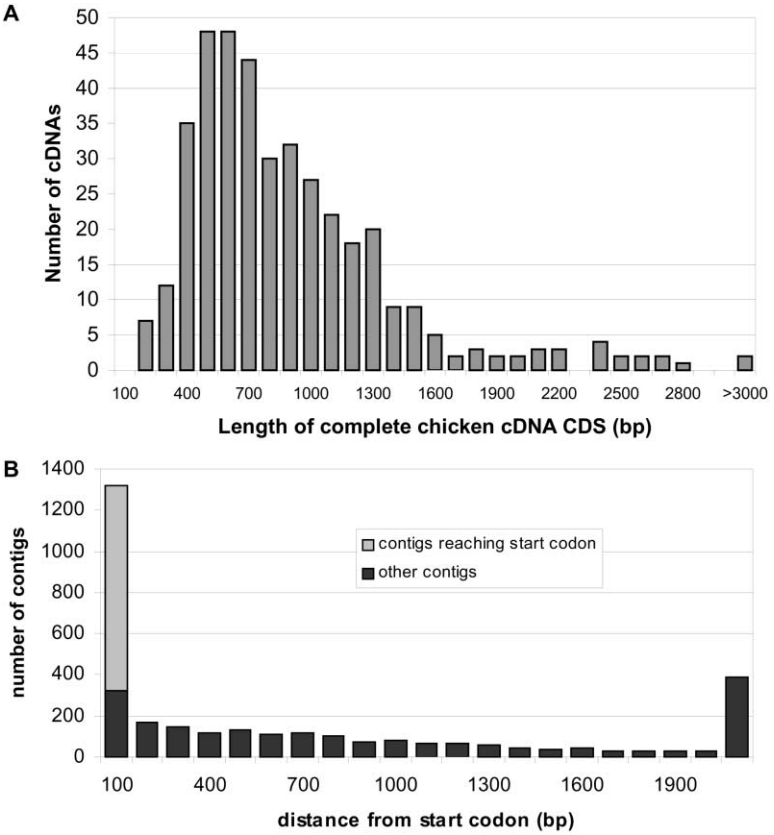

Figure 3. Comparison of Contig Sequences with Full-Length Chicken cDNAs

(A) Distribution of full-length chicken cDNA coding regions completely matched by individual contigs.

(B) Distribution of distances from the start codon for the contigs that match to known full-length chicken cDNAs. In total, 996 contigs extend to the start codon. Full-length cDNAs were extracted from the EMBL sequence database.

the contigs are likely to include an unknown number of avian-specific or rapidly evolving genes.

Comparison of the EST dataset with other databases confirms the extensive nature of this resource. For example, the EST resource has matches for 72% of the known full-length chicken cDNAs, and the resource also matches 84% of *Gallus gallus* ESTs in EMBL (BLASTN hits better than $E = 10^{-30}$). A comparison against the human proteome set indicated matches (BLASTX hits better than $E = 10^{-6}$) to 36% of the 27,000 ENSEMBL confirmed human protein set (available from http://www.ensembl.org), and strong matches (TBLASTN at $E = 10^{-20}$) to 90% of the genes in the morbid map of the OMIM human disease database (http://www.ncbi.nlm.nih.gov/omim/) were found.

The large number of cDNA sequences in this dataset allows the estimation of the total number of genes in the chicken genome. The method of Ewing and Green [11] was applied to the contigs containing at least two component ESTs and two reference sets; 1,845 clustered, full-length chicken cDNAs (taken from EMBL), and 1,676 nonredundant complete protein sequences (taken from SWISS-PROT/TrEMBL). The estimated total number of genes predicted via this approach were 33,228 and 35,682 for the nonredundant cDNA and protein reference sets, respectively. This is in reasonable agreement with gene number estimates for the human [11, 12] and *Fugu rubripes* genome [13], suggesting a common baseline of around 35,000 genes for vertebrates.

In partnership with the Sanger Institute, we are currently working to determine the full cDNA insert sequence for the entire nonredundant clone set. This dataset will be the most extensive coverage of the chicken transcriptome to date and will provide an invaluable tool for exon definition when the chicken genome sequence is determined. The second-phase cDNA data will be made available through our web site (http://www.chick.umist.ac.uk), which currently allows users to conduct BLAST and keyword searches and to download the whole dataset. The EST sequences have been submitted to Genbank.

Experimental Procedures

Library Construction and Sequencing

Library construction and subsequent sequencing was carried out by Incyte Genomics in a public-private partnership. Standard methods were used for preparing cDNA, which were poly A-trimmed [3] prior to being cloned between the NotI and EcoRI sites in pBluescript II KS+. For each tissue, a standard cDNA library was constructed together with two normalized libraries of varying stringencies. EST sequences were obtained from the 5′ end of clones by standard methods. Clones are available from our distributors, the MRC Geneservice (http://www.hgmp.mrc.ac.uk/geneservice/) and ARK genomics at the Roslin Institute (http://www.ark-genomics.org/).

Computational Analysis

Successful reads were passed through a bioinformatic quality-control process. Vector clipping involved a Smith-Waterman search [14] against the sequence 5′ and 3′ of the NotI and EcoRI restriction sites in pBluescript II KS+. We performed further checks against the complete vector, the *E. coli* genome, and a complete set of chicken rRNAs and the chicken mitochondrial genome in order to remove other contaminating sequences prior to assembly. We then assembled sequences by first clustering ESTs into gene bins by

using BLASTN to group sequences (98% identity over 50 nucleotides), then assembling sequences in gene bins with PHRAP (http://www.phrap.org) to produce the final assembled contig sequences.

Bioinformatic processing and annotation were performed with a variety of software tools, including BLASTX, BLASTN, SSEARCH3 [15], and InterProScan [16]. Standard searches were performed against a nonredundant SWISS-PROT/TrEMBL database containing 725,000 sequences.

The number of chicken genes was estimated after a search of the 38,812 multicomponent contigs with BLASTN against a set of 1845 clustered, nonredundant, full-length *Gallus gallus* cDNAs taken from the EMBL nucleotide databank and with BLASTX against a set of 1,676 nonredundant, full-length chicken protein sequences in SWISS-PROT/TrEMBL. A match was assigned to hits possessing 98% or better sequence identity over 100 bases or more for the cDNA set. A less-stringent cutoff of 96% identity and an overlap of 30 amino acids were used for the protein set because of the greater impact that minor sequencing and frameshift errors have on the translated nucleotide BLAST statistics. The total gene number estimate was then calculated as 38,812 x $n1/m1$, where $n1$ is the number of genes/proteins in the reference set, and $m1$ is the number of matching sequences from the contig set, as per the method of Ewing and Green [11].

Acknowledgments

We thank John Young and Yvonne Boyd (Institute for Animal Health), who provided some of the tissues for library construction, and Nick Cole, Mikiko Tanaka, and Megan Davey from Dundee for help with dissecting the embryos. We thank Alf Game for his pivotal role in the early stages of the project. We thank Aimal Pashtoonmal, Sarah Dyer and Christopher Musther (University of Manchester Institute of Science and Technology), who carried out preliminary bioinformatic analyses. We thank Kristina Kaufman, Phil-Eric Jiao, Irene Ni, Amber Pham, and Silvia Ramirez (Incyte Genomics) for their contribution to the project. We are grateful to Alf Game, Yvonne Boyd, and John Young for invaluable discussions regarding tissue choices for sequencing. This work was supported by a grant awarded by the Biotechnology and Biological Sciences Research Council to W.R.A.B., C.T., and S.A.W. A Medical Research Council Bioinformatics Studentship supported P.B. C.T. acknowledges support from the Medical Research Council and a Royal Society Professorship.

Received: August 22, 2002
Revised: September 10, 2002
Accepted: September 10, 2002
Published: November 19, 2002

References

1. Tirunaguru, V.G., Sofer, L., Cui, J., and Burnside, J. (2000). An expressed sequence tag database of T-cell enriched activated chicken splenocytes: sequence analysis of 5251 clones. Genomics *66*, 144–151.
2. Abdrakhmanov, I., Lodygin, D., Geroth, P., Arakawa, H., Law, A., Plachy, J., Korn, B., and Buerstedde, J.M. (2000). A large database of chicken bursal ESTs as a resource for the analysis of vertebrate gene function. Genome Res. *10*, 2062–2069.
3. Fu, G.K., Starnes, S. and Stuve, L. (2002). Construction of unidirectionally cloned cDNA libraries from messenger RNA for improved 3′ end DNA sequencing. United States Patent. Number 6,387,624.
4. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402.
5. Bairoch, A., and Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res. *28*, 45–48.
6. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene ontology: tool for the unification of biology. Nat. Genet. *25*, 25–29.
7. Ewing, B., Hillier, L., Wendl, M.C., and Green, P. (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. *8*, 175–185.
8. Ewing, B., and Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. *8*, 186–194.
9. Stoesser, G., Baker, W., van den Broek, A., Camon, E., Garcia-Pastor, M., Kanz, C., Kulikova, T., Leinonen, R., Lin, Q., Lombard, V., et al. (2002). The EMBL Nucleotide Sequence Database. Nucleic Acids Res. *30*, 21–26.
10. Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., et al. ((2002). The Ensembl genome database project. Nucleic Acids Res. *30*, 38–41.
11. Ewing, B., and Green, P. (2000). Analysis of expressed sequence tags indicates 35,000 human genes. Nat. Genet. *25*, 232–234.
12. Hogenesch, J.B., Ching, K.A., Batalov, S., Su, A.I., Walker, J.R., Zhou, Y., Kay, S.A., Schultz, P.G., and Cooke, M.P. (2001). A comparison of the Celera and Ensembl predicted gene sets reveals little overlap in novel genes. Cell *106*, 413–415.
13. Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J.M., Dehal, P., Christoffels, A., Rash, S., Hoon, S., Smit, A., et al. (2002). Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. Science *297*, 1301–1310.
14. Smith, T.F., and Waterman, M.S. (1981). Identification of common molecular subsequences. J. Mol. Biol. *147*, 195–197.
15. Pearson, W.R. (2000). Flexible sequence similarity searching with the FASTA2 program package. Methods Mol. Biol. *137*, 185–219.
16. Zdobnov, E.M., and Apweiler, R. (2001). InterProScan – an integration platform for the signature-recognition methods in InterPro. Bioinformatics *17*, 847–848.