2009

The completion of the Mammalian Gene Collection (MGC)

Michael Brent
Washington University School of Medicine in St. Louis

Laura Langton
Washington University School of Medicine in St. Louis

Charles L.G. Comstock
Washington University School of Medicine in St. Louis

Michael Stevens
Washington University in St Louis

Chaochun Wei
Washington University School of Medicine in St. Louis

See next page for additional authors

The completion of the Mammalian Gene Collection (MGC)

The MGC Project Team, Gary Temple, Daniela S. Gerhard, et al.

Genome Res. 2009 19: 2324-2333 originally published online September 18, 2009
Access the most recent version at doi:10.1101/gr.095976.109

Resource

The completion of the Mammalian Gene Collection (MGC)

The MGC Project Team[1]

Since its start, the Mammalian Gene Collection (MGC) has sought to provide at least one full-protein-coding sequence cDNA clone for every human and mouse gene with a RefSeq transcript, and at least 6200 rat genes. The MGC cloning effort initially relied on random expressed sequence tag screening of cDNA libraries. Here, we summarize our recent progress using directed RT-PCR cloning and DNA synthesis. The MGC now contains clones with the entire protein-coding sequence for 92% of human and 89% of mouse genes with curated RefSeq (NM-accession) transcripts, and for 97% of human and 96% of mouse genes with curated RefSeq transcripts that have one or more PubMed publications, in addition to clones for more than 6300 rat genes. These high-quality MGC clones and their sequences are accessible without restriction to researchers worldwide.

[Supplemental material is available online at http://www.genome.org. The accession nos. and properties of all clones and sequences are listed in the Supplemental material and at ftp://ftp.ncbi.nih.gov/repository/MGC/MGC_project/.]

cDNA clones containing the entire protein-coding sequence of mRNA transcripts (full-CDS clones), together with corresponding high-quality sequences, are essential resources for annotating protein-coding genes on genomes and for expressing the protein products of those genes. The Mammalian Gene Collection (MGC) was established as a multi-institute effort at the National Institutes of Health (NIH) to provide the research community with unrestricted access to sequence-validated human and mouse full-CDS clones and their sequences (Strausberg et al. 1999).

The goal for MGC at the outset, in 2000, was to provide at least one sequence-validated, full-CDS clone for each known human and mouse gene. A similar cDNA cloning program was funded later for 6200 rat genes. The MGC high-throughput cloning, sequencing, and distribution infrastructure was also used for full-CDS cloning programs for *Danio rerio* (ZGC; http://zgc.nci.nih.gov/) and *Xenopus laevis* and *Xenopus tropicalis* (XGC; http://xgc.nci.nih.gov/).

MGC clones initially were obtained by randomly picking 5000–20,000 colonies from custom cDNA libraries and end-sequencing the plasmid inserts. Those representing genes absent from the collection were fully sequenced (Strausberg et al. 2002). Using this approach, by June 2004 the MGC had acquired full-CDS clones for 11,727 unique human genes, 10,171 unique mouse genes, and 828 unique rat genes (Fig. 1), isolated from 154 human, 131 mouse, and 33 rat libraries, derived from a wide variety of tissues and cell lines (Gerhard et al. 2004). Putative full-CDS clones were sequenced to high quality, that is, with no uncertain base calls and an average error rate of <1 error in 50,000 bp. Descriptions of these libraries and their tissue sources are available at http://mgc.nci.nih.gov/.

The progress of the MGC, XGC, and ZGC cloning programs, from their start to March 2009, is shown in Figure 1, which also gives the total numbers of genes represented and the total numbers of clones in each collection. This report focuses on the MGC efforts to complete the human, mouse, and rat collections since the last MGC publication, in 2004 (Gerhard et al. 2004).

As the number of human and mouse full-CDS clones approached approximately 10,000, random EST screening of cDNA libraries yielded cDNA clones for progressively fewer unique genes over time (Fig. 2), and for fewer genes per thousand ESTs analyzed (data not shown), significantly raising the cost of obtaining clones for the unrepresented genes. Pilot studies to improve yield with normalized cDNA libraries and protocols that favored cDNAs for rarer transcripts and longer inserts also introduced additional mutations, yielding an increased rate of nucleotide variation in the cDNAs compared to the genome (Gerhard et al. 2004). Therefore, two alternative approaches—directed RT-PCR cloning ("PCR rescue") and DNA synthesis—were implemented to obtain clones for the missing genes. These approaches and their results are described below.

Results

PCR rescue

We selected targets for PCR rescue from RefSeq transcripts (Pruitt et al. 2009b) for human and mouse genes not represented by full-CDS clones in MGC (see Methods). In cases of multiple transcript isoforms, we chose the isoform with the longest CDS supported by transcript and protein homology in other mammalian species. These targets were divided into two sets, with each assigned to one of two laboratories. The methods used by each laboratory have been described (Baross et al. 2004; Wu et al. 2004). Targets that failed to be isolated at one laboratory were exchanged with the other for a second attempt. A full list of the RefSeq transcripts and their corresponding genes assigned for PCR rescue (together with the CDS sizes and the outcome of each transcript's PCR rescue) is given in Supplemental Table A.

PCR rescue recovered 8862 full-CDS clones for 4088 human genes (Fig. 2A) and 4774 mouse genes (Fig. 2B). RT-PCR reactions frequently displayed additional bands on gel electrophoresis, which cloning and sequence analysis often revealed as alternative splice isoforms of the targeted transcript (M Hirst, T Zeng, K Tse, A Delaney, J Pang, J Wang, G Taylor, A Deng, M Moksa, K Fichter, et al., in prep.). Clones of isoforms with a CDS length at least 50% of the CDS length of the targeted transcript were accepted by MGC, as long as they met criteria consistent with a full CDS (see Methods).

[1]A complete list of authors and affiliations appears at the end of the paper, before the Acknowledgments section.
[2]Corresponding author.
E-mail gtemple@mail.nih.gov; fax (301) 480-2770.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.095976.109.

19:2324–2333 © 2009 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/09; www.genome.org

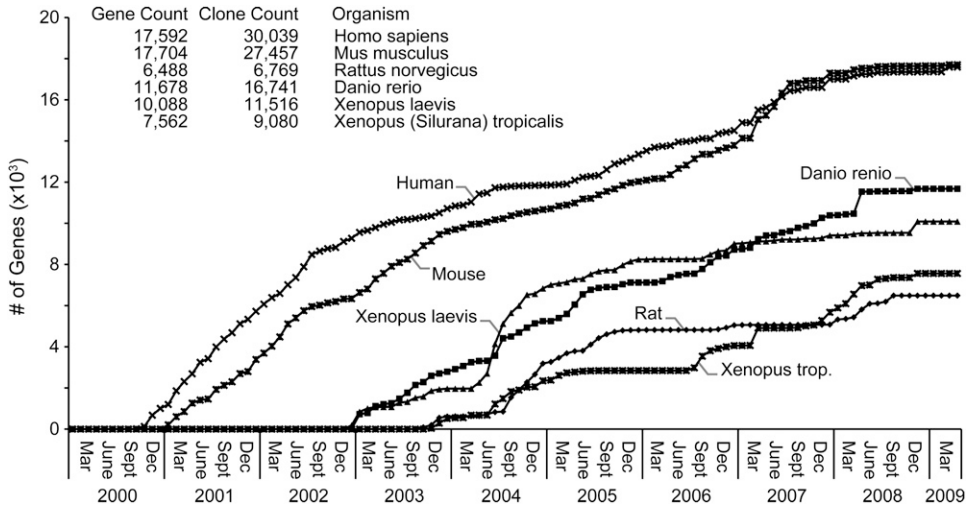

Figure 1. Cumulated gene counts for MGC, XGC, and ZGC. The progressive addition of clones, measured by genes represented in each collection, is shown for MGC, XGC, and ZGC from the beginning to conclusion of these programs. "Gene Count" is the total final number of RefSeq genes represented by each set of clones. This number includes some noncurated genes (XM accessions) that are not counted in Table 1. "Clone Count" includes all clones, including duplicate transcripts and isoforms. Isoforms constitute 2%–3% of the human, mouse, and rat collections.

Target size most strongly influenced the outcome, with 64%–70% success for 0.1–3-kb targets falling progressively to zero for targets of 9 kb and larger (Fig. 3). Success also correlated inversely with the level of mRNA expression (Supplemental Fig. S1). Overall, one or more full-CDS clones were obtained for 65% of targeted genes.

DNA synthesis of full-CDS clones

After two attempts at PCR rescue, MGC still lacked full-CDS clones for about 2200 human genes and about 1800 mouse genes with curated RefSeq transcripts (accession prefix NM_). Compared to the expense for further attempts at PCR rescue, DNA synthesis provided a cost-effective alternative for obtaining clones for transcripts of most outstanding genes in MGC.

DNA synthesis also made it practical to synthesize the CDS precisely, without additional 5′- or 3′-untranslated region (UTR) sequences, facilitating the subsequent use of these clones to produce proteins with N-terminal and C-terminal fusion tags using the Gateway cloning system. The MGC full-CDS clones generated by DNA synthesis were prepared in a Gateway Entry vector, permitting the subsequent transfer of inserts into a wide range of expression vectors by site-specific recombination (Hartley et al. 2000), a transfer method with a very low risk of introducing mutations into the transferred inserts (JL Hartley, unpubl.).

The protein-coding sequences of 3647 RefSeq accessions supported by known transcripts and protein orthologs were assigned for DNA synthesis to two companies (Methods). The numbers and sizes of human and mouse transcripts assigned for synthesis and the rate of success for each size category are given in Figure 4. The contributions of DNA synthesis to the total MGC human and mouse full-CDS clone collections are displayed in Figure 2, A and B. DNA synthesis provided MGC with full-CDS clones for 86% of the 3414 outstanding genes assigned. Synthesis succeeded for 94% of targets with a CDS of 4 kb or less, but success fell dramatically for larger targets (11% of 46 targets with CDS >10 kb).

Finally, 318 cDNA clones for 126 high-priority genes that had failed one or more attempts at synthesis and stable cloning of a full-CDS were accepted by MGC with the CDS cloned in two or more fragments. For 92 of these genes, the partial-CDS clones together compose the entire CDS (Supplemental Table B).

Predictions of new human genes

Multi–exon gene predictions

Starting in 2005, MGC sought to predict human multi-exon genes absent from the RefSeq and other major gene catalogs. We used algorithms that relied primarily on comparative sequence data, with or without existing EST or cDNA evidence: N-SCAN (Gross and Brent 2006), N-SCAN_EST (Wei and Brent 2006), Exoniphy (Siepel and Haussler 2004), and TransMap (Zhu et al. 2007). Results were confirmed by sequencing RT-PCR products of two or more spliced exons in the predicted transcripts from each postulated gene locus. This effort identified 734 novel gene fragments (NGFs) containing 2188 exons with little or no prior cDNA support, corresponding to an estimated 563 distinct genes. At the time of this analysis, 327 of these genes were completely absent from the cDNA-based RefSeq and Vega gene catalogs (Wilming et al. 2008; Pruitt et al. 2009b), and 178 were also absent from the Ensembl collection (Hubbard et al. 2009). Many other gene fragments were identified that represented extensions of known genes. These novel fragments contributed transcript evidence for 480 RefSeq accessions later assigned for PCR rescue. For seven of these accessions, the NGFs provided the only direct transcript support. Details of the methods and results of this program were published in 2007 (Siepel et al. 2007). Subsequent to our analysis, 42 genes overlapping the novel gene fragments have been added to RefSeq.

Single–exon gene predictions

To minimize the inclusion of pseudogene transcripts and other non-protein-coding sequences in the MGC, our random-EST cloning and PCR rescue efforts intentionally excluded transcripts of single-exon genes (SEGs) and transcripts potentially encoding proteins of fewer than 100 amino acids (Strausberg et al. 2002). These criteria excluded the isolation of transcripts of authentic single-exon genes and some multi-exon genes encoding short protein-coding transcripts, such as for some human olfactory

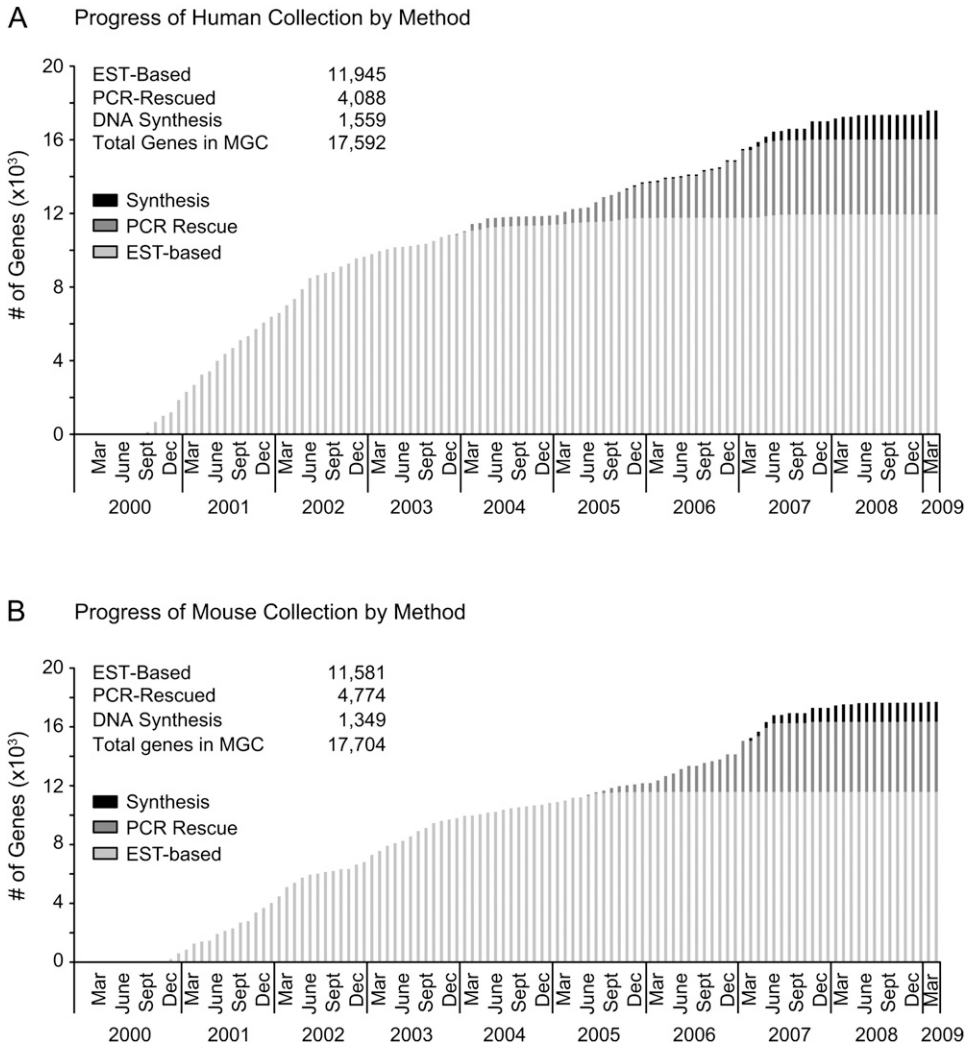

Figure 2. MGC progress represented over time by method. (*A*) Human; (*B*) mouse. The absolute contribution (by genes represented) of each cloning method is shown for EST-based cloning, PCR-Rescue, and DNA synthesis, over time.

receptors (Carninci et al. 2005; Glusman et al. 2006; The ENCODE Project Consortium 2007).

To assess how many SEGs are not annotated in current gene catalogs, we considered all open reading frames (ORFs) longer than 200 bp in the human genome. We used logistic regression analysis to select 351 ORFs most likely to encode unannotated SEGs, based on features such as cross-species conservation, protein homology, and genome-wide expression data (Methods). These candidate SEGs were tested for expression by RT-PCR, with no-RT controls to detect results due to genomic contamination. Expression was confirmed in 198 out of 351 candidates (57%) (Supplemental Table S1). Additional RT-PCR experiments, using RNA from several tissues and variable numbers of PCR cycles, suggested that these SEG candidates are expressed at low levels and in a tissue-specific manner, especially in the testes and cerebellum. However, a large fraction of negative reference loci (selected from annotated pseudogenes and regions annotated as intronic or intergenic by the ENCODE pilot project) also showed evidence of expression by RT-PCR, consistent with previous reports (Carninci et al. 2005; Glusman et al. 2006; The ENCODE Project Consortium 2007). Attempts to confirm expression at the protein level were inconclusive, with only nine of 198 positive candidates and six of the 138 negative reference loci matching peptide mass spectrometry (MS) spectra (http://bioinfo2.ucsd.edu/MSGeneAnnotation/index.html), perhaps in part owing to low levels of protein expression and incomplete databases of peptide MS spectra. Thus, whether these 198 candidate SEGs are true protein-coding genes remains an open question.

These ambiguous results underscore the challenge of obtaining a fully comprehensive set of human protein-coding genes, given pervasive genomic transcription, expressed pseudogenes, and true genes that are expressed transiently and at low levels. Although our genome-wide search for candidate SEGs turned up relatively few instances with, at best, questionable evidence for protein-coding function, our methods could have overlooked some fast-evolving, very short, lineage-specific, or recently duplicated genes.

Final numbers of genes represented by MGC clones

Table 1 gives the final numbers of human, mouse, and rat genes represented by one or more full-CDS clones in the MGC, compared to the totals for four classes of protein-coding genes. The MGC now

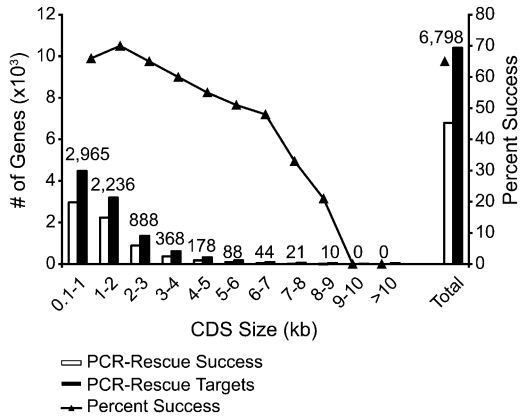

Figure 3. PCR rescue success versus target size. (Black bars) The number of assigned targets in each size range; (white bars) the number of assigned targets that were obtained as full-CDS clones, with the number of clones recovered shown *above* the bars. The triangles and trendline show the percentage recovered for each size group. Excluded from these calculations are RefSeq targets where the assigned CDS later was changed, suppressed, or withdrawn over the course of the PCR rescue program. Among 8764 human and mouse targets with changed annotation, we obtained a full-CDS clone for 3197 (36%), including one 10.8-kb clone (BC150731).

contains clones for 92% of human genes, 89% of mouse genes, and 41% of rat genes with RefSeq curated (NM accession) transcripts, regardless of publication status (line A); and contains full-CDS clones for 97% of human genes, 96% of mouse genes, and 44% of rat genes with one or more PubMed publication (line B). Table 1 also shows that MGC includes full-CDS clones for 93% of human genes linked to a disease phenotype (line C), and for 95% of human and 93% of mouse transcripts listed as highly curated Consensus CDS (CCDS) transcripts (line D) (Pruitt et al. 2009a). The RefSeq transcripts for the genes represented in Table 1 and the corresponding accessions of MGC clones are given in Supplemental Table C; and genes lacking full-CDS clones in MGC are listed in Supplemental Table D. The size distributions of the final MGC human and mouse clone collections, compared to the longest RefSeq transcript isoforms, are shown in Supplemental Table S8. The relationships between the MGC, RefSeq, and Ensembl human and mouse gene sets are shown in Figure 5.

Sequence variation in MGC clones

MGC clone variation versus dbSNP

The full-CDS sequences for all clones submitted to MGC were compared to their corresponding reference genome (human clones were also compared to the chimpanzee reference genome). Discrepancies between a cDNA and its reference genome are annotated in the GenBank records, with links to polymorphisms recorded in dbSNP.

Table 2 shows the sequence discrepancies (single-nucleotide mismatches and indels) found between MGC clones and the reference human (version 36.3) and mouse (version 37.1) genomes, expressed as the number of differences observed per clone. Among human clones, 57% contain no mismatch in the CDS and 72% no nonsynonymous (NS) mismatches. Similarly, 66% of mouse clones contain no mismatches in the CDS and 79% no NS mismatches. Thus, the majority of clones are free of any differences in the CDS; and 72% of human and 79% of mouse clones are free of NS changes.

Supplemental Table S2 presents the rates of sequence discrepancy, based on total sequences of human and mouse clones, together with the percentages of discrepancies that correspond to validated polymorphisms in dbSNP. Because the mouse reference genome sequence was derived from a single mouse strain, C57BL/6, the variation in MGC mouse clones was divided into three categories, based on the strains that provided the RNA: C57BL/6 and C57BL/6J; other known strains, including crosses to C57BL/6J; and undocumented strains.

The variation per nucleotide observed in human MGC clone coding and noncoding human sequences compared to the human reference genome (Supplemental Table S2) is 9.1×10^{-4}, 44.6% of which is validated polymorphism in dbSNP (defined in the footnote to Table 2). For MGC mouse clones, the variation frequency and percentage of variation in dbSNP vary with the strain of mouse RNA used to prepare the clones. As expected, both the variation (3.8×10^{-4}) and the percent variation documented in dbSNP (4.6%) are lowest for clones derived from C57BL/6.

Sequence variation due to RNA editing

Sequence discrepancies in MGC clones can also reflect post-transcriptional editing of mRNA, which in mammalian cells is due almost exclusively to A-to-I editing, mediated by the adenosine deaminases acting on RNA (ADAR) family of enzymes (Bass 2002; Gommans et al. 2008). The resulting inosine in the edited RNA is read as guanosine by the in vivo cellular machinery, as well as by the enzymes used in cDNA cloning and sequencing. To date, only about 70 human mRNAs have been reported to contain A-to-I editing sites in the CDS (Supplemental Table S3), whereas several thousand examples of A-to-I editing in noncoding sequences of the 5′ and 3′ UTRs and within introns of human pre-mRNA sequences have been reported (Athanasiadis et al. 2004; Kim et al. 2004; Levanon et al. 2004; Li et al. 2009).

We sought to identify candidate A-to-I editing sites in MGC clones. Because MGC has produced only a single full-CDS clone for most genes, we could not use the occurrence of coincident edits within multiple clones to identify loci of selective RNA editing. Therefore, we used two different tests to focus on identifying clones statistically enriched for clusters of A-to-G changes compared to the genome sequence (Supplemental Text S3). These two tests detected 118 MGC clones with potential editing sites, of which 87 were identified by both tests (Supplemental Tables S4, S5,

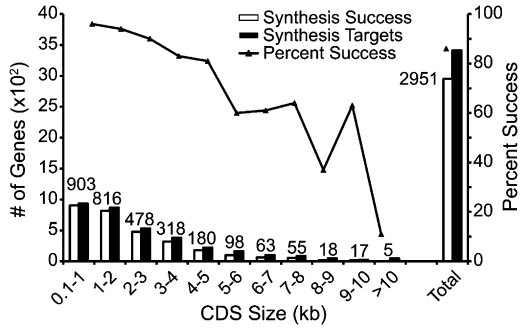

Figure 4. Synthesis success versus target size. (Black bars) The number of assigned targets in each size range; (white bars) the number of assigned targets that were obtained as full-CDS clones, with the number of clones recovered shown *above* the bars. The triangles and trendline show the percentage recovered for each size group. RefSeq targets where the assigned CDS later was changed, suppressed, or withdrawn (233 in total) were excluded from these calculations.

Table 1. MGC achievement

Gene classes[a]	Protein-coding genes			Protein-coding genes in MGC		
	Human	Mouse	Rat	Human	Mouse	Rat
A. All genes with curated RefSeq transcripts[b]	18,877	19,357	15,389	17,421 (92%)	17,285 (89%)	6363 (41%)
B. Genes with ≥1 PubMed articles and curated RefSeq transcripts[b]	14,614	12,434	6236	14,102 (97%)	11,902 (96%)	2724 (44%)
C. Genes with known disease phenotype[c]	2306	2208	2075	2152 (93%)	2047 (93%)	782 (38%)
D. Genes with CCDS transcripts[d]	13,884	15,263	NA	13,131 (95%)	14,124 (93%)	NA

[a]Genes counted in Classes B, C, and D are subsets of Class A and not mutually exclusive.
[b]Curated RefSeq transcripts (NM-accession transcripts) are a subset of RefSeq transcripts that have been validated based on protein and DNA evidence.
[c]Human genes in this category were identified by searching OMIM for records with "phenotype description, molecular basis known" and "gene with known sequence and phenotype" and then retrieving Gene Links that are not in the phenotype-only category. Mouse and rat genes in this category were identified using NCBI HomoloGene links for the above-mentioned human genes.
[d]Consensus CDS (CCDS) includes a subset of transcripts with agreement on the full CDS by annotation specialists at NCBI, European Bioinformatics Institute, University of California at Santa Cruz, and the Wellcome Trust-Sanger Institute (Pruitt et al. 2009a); because the numbers are based on RefSeq mRNAs in the CCDS set that are current as of March 23, 2009, they are less than the total CCDS gene number. (NA) Not applicable; CCDS genes have not been defined for rat.

and S6), with an apparent false-positive rate of 2%. Eighty-nine percent of the clusters of A-to-G changes lie wholly or partially within *Alu* repeat sequences, and 88% are within UTRs, consistent with previous reports (Athanasiadis et al. 2004; Blow et al. 2004; Kim et al. 2004; Levanon et al. 2004). Eleven clones within this set of 87 show evidence of CDS editing, including clones for seven genes that to our knowledge have not been reported previously to have edits in the CDS (Supplemental Table S7).

Accessing MGC clones

From the years 2000 to 2007, MGC clones were archived at the IMAGE Consortium (Lawrence Livermore National Laboratories), which provided MGC clones to the scientific community through five commercial distributors: Open Biosystems, Life Technologies (formerly Invitrogen), and ATCC, in the United States; Gene Services Ltd, in the United Kingdom; and imaGenes, in Germany. In January 2008, all MGC, XGC, and ZGC clones were relocated to a permanent new archive, at the HudsonAlpha Institute for Biotechnology (HAIB), in Huntsville, Alabama. The HAIB website (http://image.hudsonalpha.org/) now lists all MGC clones. Scientists wishing to obtain MGC clones can order them, as before, from the same five commercial distributors.

Table 3 lists the URLs of websites that provide useful information and search tools for users seeking information on and access to MGC cDNA clones. Searches for MGC clones can begin at the MGC website (http://mgc.nci.nih.gov/), at the NCBI portal (http://www.ncbi.nlm.nih.gov/), or at the UCSC Genome Browser (http://genome.ucsc.edu/). Also listed are tutorials on how to locate MGC clones and details related to vectors, libraries, and tissue sources.

Discussion

The MGC now provides the scientific community with unrestricted access to high-quality, full-CDS clones and sequences for 92% of human genes and 89% of mouse genes with curated RefSeq (NM-accession) transcripts, and for 97% of human and 96% of mouse genes with curated RefSeq transcripts and at least one publication. The MGC also includes 6363 rat clones, representing 41% of rat genes encoding curated RefSeq protein-coding transcripts. A complete list of MGC full-CDS clones, including isoforms, is provided in Supplemental Table C.

MGC clone quality

The high sequencing standards used by MGC means that errors in MGC clone DNA sequence analysis are well below 1 in 50,000 bp. The protein-coding sequence in the majority of human and mouse clones perfectly matches its reference genome. Nonsynonymous (NS) changes are absent in 72% of human and 79% of mouse clones, and 45.7% of NS changes in human clones are documented as polymorphisms. For mouse clones, the percentage of NS changes documented as polymorphism varies depending on the strain used as the source of RNA. All differences from the reference genome are noted in the GenBank record for each MGC clone.

Assuming that cloning procedures for the human and mouse clones introduced mutations at roughly similar frequencies and that the 6.5% of C57BL/6 variation matching dbSNP (Supplemental Table S2) largely represents variation within different colonies of C57BL/6, the remaining 93.5% of the rate of sequence discrepancy in the CDS (2.7×10^{-4}) suggests an upper limit of 2.5×10^{-4} for the combined frequency of CDS mutations arising from the preparation of the clones, sequencing errors ($\leq 0.2 \times 10^{-4}$), and RNA editing in both mouse and human clones.

We identified a small percentage of MGC clones with changes suggesting A-to-I editing of pre-mRNA. As reported previously by others, most of these putative edited sites lie within UTR sequences and overlap *Alu* repeat sequences. We also identified new evidence of A-to-I editing in the CDS of MGC clones for seven human genes, detected by both of the tests we used (Supplemental Table S7).

Maintaining high clone quality also depends on researchers receiving the correct clone for the accessions they have ordered. To detect and correct well-to-well contamination and errors in the clone rearraying process, all clones on master plates at LLNL and HAIB are end-sequenced to confirm their identity, prior to sending replica plates to MGC commercial distributors. Incorrect clones are replaced, if a suitable replacement is available, or removed. Results of this QC process are posted at http://image.hudsonalpha.org/qc/html/QCoverall.shtml.

Revised genome annotation

While the MGC PCR rescue program was under way, concurrent progress in human and mouse genome annotations forced MGC

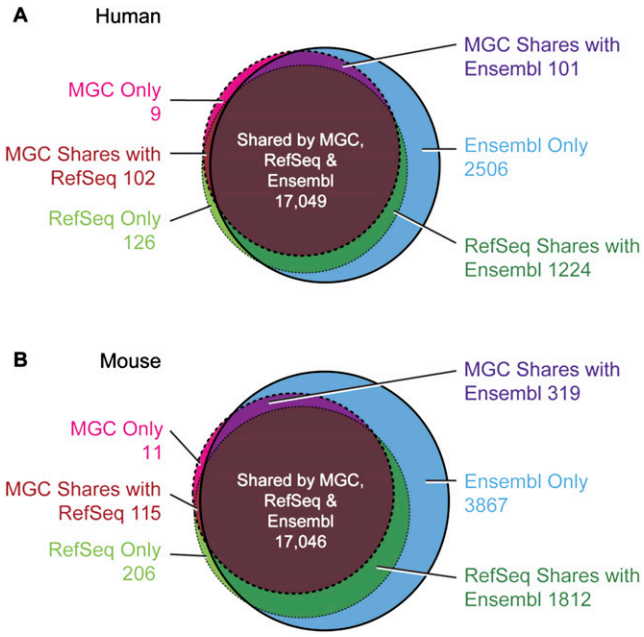

Figure 5. Venn diagram comparing the number of loci containing protein-coding genes from MGC, RefSeq, and Ensembl. (*A*) Human; (*B*) mouse. The loci were computed by clustering transcripts from all three gene sets based on the overlap of the genomic location of the CDS portion of the exons. When a transcript is not uniquely mapped to the genome, the clusters for all mappings of that transcript were combined and counted as one locus. For human, this resulted in 17,239 loci containing MGC clones, 18,494 loci with RefSeq mRNAs (Pruitt et al. 2009b), and 20,856 Ensembl gene loci (Hubbard et al. 2002). Mouse had 17,455 loci with MGC clones, 19,064 loci with RefSeq mRNAs, and 23,087 Ensembl gene loci. Genes counted as shared between any two gene sets exclude genes in the third set. BLAT (Kent 2002) alignments of MGC clones and RefSeq mRNAs (NM accessions) obtained from the UCSC Genome Browser database (Karolchik et al. 2008) for human genome assembly 36.1 and mouse assembly 37, and Ensembl Release 52 were used in the analysis. Genomic loci serve as an estimate of the number of genes in these data sets. The counts vary from those seen in Table 1, owing to the different method of computation.

to retire and reassign ~45% of the target sequences assigned between 2004 and 2007. Updated CDS annotation can, for example, reposition the annotated ATG start codon of a CDS further 5′, extending the CDS, or excise a length of CDS sequence deemed to be a retained intron, or retire a transcript from a likely pseudogene.

With the conclusion of the MGC project, the GenBank records of MGC clone sequences have been frozen, with no further updates. What constitutes a full-length coding region for some of the genes and transcripts for which MGC has clones is likely to change in the future; therefore, users planning to order MGC clones will need to monitor for these changes. Users can employ genome browsers and gene-specific databases, such as NCBI's Evidence Viewer, Entrez Gene, and the UCSC Genome Browser, to view relevant regions of the genome (browsers) or gene-related information (Entrez Gene). MGC has added a guide (see Table 3) to its website to help users evaluate MGC clone sequences in light of current genome annotation.

Future collections

Since its inception, the MGC approaches to cloning cDNAs for additional genes evolved by exploiting concurrent technical advances: dramatically cheaper DNA sequencing; improved bioinformatics methods for gene prediction and gene annotation; and cheaper, more accurate DNA synthesis for building cDNA clones. These advances made it feasible for MGC to achieve full-CDS clones for nearly 90% or more of a well-defined set of RefSeq transcript targets, transcribed from <2% of the human and mouse genomes (Lander et al. 2001; Venter et al. 2001; Waterston et al. 2002; Carninci et al. 2005).

Recently our view of the eukaryotic transcriptome has expanded dramatically in size and complexity to include multiple splice isoforms for 90% or more of multi-exon genes (Kuhn et al. 2009), and a vast network of sense and antisense non-protein-coding RNAs, some of which are well studied (Carthew and Sontheimer 2009), with many others still largely uncharacterized and of uncertain biological relevance (Kapranov et al. 2007; Pheasant and Mattick 2007; Guttman et al. 2009).

These major developments have implications for how one would build another collection of clones for RNAs of contemporary interest, such as for splice isoforms or non-protein-coding RNAs. Given the speed and cost efficiency of DNA synthesis, when the need arises for a particular transcript, a laboratory now can order most cDNAs to be synthesized. Indeed, this approach may suffice for many laboratories, given the MGC experience that only a handful of laboratories ordered entire collections of human or mouse cDNA clones for large-scale studies, while the overwhelming majority of customers ordered clones for <10% of the collection (C Pennacchio, unpubl.).

Yet high-throughput programs to study protein–protein interactions, protein structure, and protein function clearly profit from access to centralized collections of large numbers of clones. Such collections offer the scientific community benefits of scale, by providing clones of lower cost and more uniform quality; by reducing the waste of duplicated clone preparation within the community; and by relieving individual laboratories of the burden of clone quality control and distribution.

Less formal centralized approaches also can provide some of the same benefits. For example, the ORFeome Collaboration (OC; http://www.orfeomecollaboration.org/) is an informal network of laboratories, consisting of 10 contributing academic, commercial, and government groups (including the MGC), that are cooperating—largely without dedicated funding—to build a public collection of human cDNA clones in an expression-convenient format.

The growing emphasis on defining cellular networks, with myriad interactions of RNA, DNA, and protein, may result in an increased demand for such centralized collections in the future.

Methods

Target selection

From all protein-coding genes with RefSeq transcripts annotated on the human and mouse genomes (Pruitt et al. 2009b), we selected targets for genes outstanding from MGC based on two properties: their potential research and medical importance, and the level of supporting evidence that the transcript represents a CDS-complete product, as previously described (Strausberg et al. 2002).

For assigning PCR rescue targets, transcripts for human and mouse genes were ranked by the number of peer-reviewed publications associated with the genes. For genes lacking publications, orthologs and the number of gene-specific NCBI web queries were used for ranking. In the initial PCR rescue efforts, some potential

Table 2. Sequence variation in MGC clones versus RefSeq genomes of human and mouse

No. of discrepancies per clone	Clones with no discrepancies (Fx total)	No. of clones examined	Average no. of discrepancies per clone	Frequency (per kb)	Percent in dbSNP[c]
Human[a]					
Discrepancies in CDS + UTR	9713 (0.36)	27,188	1.87	0.91	44.6%
Discrepancies in CDS	15,551 (0.57)	27,188	0.89	0.64	54.6%
NS Discrepancies in CDS	19,636 (0.72)	27,188	0.47	0.33	45.7%
Mouse[b]					
Discrepancies in CDS + UTR	12,062 (0.47)	25,679	2.69	1.28	37.5%
Discrepancies in CDS	16,839 (0.66)	25,679	1.35	0.97	46.6%
NS Discrepancies in CDS	20,205 (0.79)	25,679	0.55	0.40	31.5%

Sequence discrepancies are accepted in MGC clones only if they do not change the phase of reading frame, alter the start or stop codons, or result in a CDS that is <50% of the length of the CDS of the longest isoform.
[a]89.2% of human discrepancies are single-nucleotide mismatches, and 10.8% are indels, of which 11% are in the CDS (1.2% total discrepancies).
[b]89.3% of mouse discrepancies are single-nucleotide mismatches, and 10.7% are indels, of which 8.8% are in the CDS (0.94% total discrepancies).
[c]Percent in dbSNP is based on dbSNP build 129 and represents validated SNPs identified as (1) SNPs with allele frequency data; (2) RefSNPs with at least two submitted SNPs, where at least one submitted SNP is by noncomputational method and is not a cDNA; (3) SNPs validated by submitter confirmation; or (4) SNPs validated by DoubleHit criteria.
(NS) Nonsynonymous. PCR rescue and DNA synthesis clones have less than all or none of the 5′- and 3′-UTR sequences represented (see Methods).

transcript targets uncharacterized in the literature were based on a single high-throughput cDNA cloning report and lacked protein orthologs; in these cases, the annotated transcript was also ranked based on the likelihood it contained a full-CDS. The full list of the transcripts targeted by PCR rescue is provided in Supplemental Table A.

For DNA synthesis, only RefSeq accessions (Pruitt et al. 2009b) confirmed to be current and well supported by known transcripts and protein orthologs (accession prefix NM_) were assigned. RefSeq transcripts containing predicted sequence (accession prefix XM_) were excluded. Candidates for DNA synthesis also underwent a final assessment by the NCBI RefSeq staff, to ensure that the CDS to be synthesized aligns to the genome and does not code for any nonsynonymous changes. Targets for outstanding genes were ranked for research and medical importance using the same criteria as used for PCR rescue, with additional weight given to disease genes listed in OMIM.

PCR rescue

Transcript targets for directed RT-PCR cloning were assigned to two research groups, at the Baylor College of Medicine and the British Columbia Cancer Agency Genome Sciences Center. The longest isoform generally was assigned for PCR rescue. Full descriptions of the PCR rescue protocols used by each center have been published (Baross et al. 2004; Wu et al. 2004). Both groups designed PCR primers flanking the target CDS, including varying amounts of UTR sequence, and RT-PCR was performed on RNA pooled from multiple tissues. Three to 12 clones from each RT-PCR reaction were isolated. Following EST sequencing and gel analysis, clones with the correct insert size and 5′- and 3′-end sequences became candidates for full-insert sequencing. After two RT-PCR cloning attempts, if one group failed to isolate a suitable cDNA clone for an assigned RefSeq transcript, that target was reassigned to the other group for another round of PCR rescue.

Clones containing CDS inserts shorter than the CDS of assigned transcripts were accepted into the MGC collection if they met the following criteria: (1) the protein alignment and hexamer analysis (Strausberg et al. 2002) does not indicate that the CDS is partial; (2) the reading frame is consistent with existing RefSeq records of the gene; and (3) the CDS length is equal to or greater than one-half of the longest RefSeq CDS of the gene.

Table 3. Websites providing information on the Mammalian Gene Collection

URL	Description
http://mgc.nci.nih.gov/[a]	MGC website provides search engines, lists of genes and libraries, and information on library construction, vectors, tissue sources, and resources for human, mouse, and rat. This site includes "A Guide to Finding MGC Clones and Evaluating Their Sequence."
http://xgc.nci.nih.gov/	As above, but for *Xenopus laevis* and *Xenopus tropicalis* clones and information
http://zgc.nci.nih.gov/	As above, but for *Danio rerio* clones and information
http://www.ncbi.nlm.nih.gov/genome/clone/finding_cdna.shtml	"Tips for Finding cDNA Clones" is an NCBI page with extensive details on locating MGC clones.
http://www.ncbi.nlm.nih.gov/gene	Entrez Gene supports query by MGC clone designation and directs users to the cDNA clone order page.
http://www.ncbi.nlm.nih.gov/unigene	UniGene provides links to MGC/XGC/ZGC, supports retrieval of clusters with MGC clones, and directs users to the cDNA clone order page.
http://genome.ucsc.edu/	UCSC Genome Browser provides tracks that can be activated to display MGC and ORFeome Collaboration clones aligned with individual human genes. Links lead to additional information on the clone, associated protein, and to "Order cDNA Clone.
http://genome.ucsc.edu/goldenPath/help/ucscGeneFishing.pdf	This tutorial, "Fishing for Genes in the UCSC Browser," includes a section on accessing information on MGC clones.

[a]The MGC web page allows users to search for available clones by keyword or gene symbol. Entering "p53" as a keyword or "TP53" as a gene symbol yields the result that one MGC clone exists, named "BC003596"; the library name and IMAGE ID are also given. Clicking on the library name shows that this library was derived from a renal cell carcinoma and gives information on the library construction, vector, and bacterial host strain used. Following the link to BC003596 provides full details on the clone nucleotide sequence and encoded protein sequence. This page provides an "Order cDNA Clone" link (upper right corner) that lists the IMAGE distributors that offer this clone for sale, together with direct links to each distributor.

DNA synthesis

Following a successful Pilot Study (Supplemental Text S2), MGC assigned native protein-coding sequences (CDS) of RefSeq NM-accessions for synthesis to GeneArt (2564) and Codon Devices (1177). A net total of 3647 targets (minus duplicates) were assigned. GeneArt synthesized the first ~90% of its assigned targets in two versions, with a stop codon (TAA) and without a stop codon (TAC). Subsequent assignments to both companies requested only one version, with a stop codon (TAG). The largest CDS assigned was 20,721 nt. Assigned sequences were designed with uniform 5′- and 3′-flanking sequences that include Gateway recombination sites; and the full-CDS sequences plus flanking sequences were cloned into a Gateway Entry vector, pENTR223.1, as described (http://mgc.nci.nih.gov/Vectors). Final clones provided to MGC were sequenced to MGC standards, as described below. The delivered CDS sequences were required to match exactly the assigned RefSeq transcript.

In a small fraction of cases, the inserted sequence could not be stably propagated in pENTR223.1, and those sequences were provided to MGC in alternative vectors (indicated in the GenBank record). MGC also required that its clones be delivered in phage-resistant strains of *Escherichia coli* (*tonA-, tonB-*). Rarely, full-CDS clones proved unstable in one or more phage-resistant *E. coli* strains and were provided in non-phage-resistant strains. For 144 high-priority genes where the full-CDS insert proved unstable in multiple vectors and host strains, MGC accepted stable clones containing the CDS in multiple subfragments (listed in Supplemental Table B). A list of all the assigned transcripts, with their size and synthesis outcome, is included in Supplemental Table A.

Single-exon gene (SEG) predictions

Computational methods were used to screen the NCBI human genome sequence (Build 36.1) for all open reading frames (ORFs) of length at least 200 bp and to select the 351 most promising SEG candidates. To distinguish likely SEGs from pseudogenes, we used syntenic alignments between the human and other mammalian genomes; conservation of ORFs in multiple alignments with mouse, rat, and dog; homology with known proteins and domain profiles; whole-genome gene expression data; and other properties of each ORF. These features were integrated by logistic regression, after training the algorithms with both positive examples (known SEGs) and negative (known and predicted pseudogenes).

Expression was confirmed for 198 out of 351 selected candidates by RT-PCR, with RT-controls to detect results due to genomic contamination. Many of the weakest candidates appear to be fragments of pseudogenes. Indeed, we obtained an even higher percentage of expressed ORFs (67% vs. 56%) among a negative reference set of loci selected from annotated pseudogenes and regions annotated as intronic or intergenic by the ENCODE pilot project. Positive results for predicted SEGs and negative reference loci were confirmed by DNA sequence analysis of "mini-pools" of cloned RT-PCR products. A list of the SEG candidate and negative reference loci is provided in Supplemental Table S1. To find possible matches against human proteins, Vineet Bafna's group (University of California, San Diego) screened our set of predicted SEGs against an existing database of MS/MS spectra from human kidney cell lines (http://bioinfo2.ucsd.edu/MSGeneAnnotation/index.html; Tanner et al. 2007), verifying protein products of nine of the 198 putative expressed SEGs (V Bafna, pers. comm.).

DNA sequence submissions

DNA sequencing was performed by standard capillary-based methods, as described (Strausberg et al. 2002). All cDNA sequences were submitted to GenBank together with *phred* quality scores, and trace data were submitted to the NCBI Trace Archive. Clones obtained from RT-PCR were required to meet the same stringent sequencing quality that had been applied to clones from MGC cDNA libraries (Strausberg et al. 2002; Gerhard et al. 2004): less than one error per 50,000 bp, no uncertain base calls, and a *phred* score of 30 or higher at each base pair. Synonymous and non-synonymous changes were permitted within the protein-coding sequences of PCR rescue clones, but changes that altered the phase of CDS reading frame or introduced premature stop codons were not permitted. Clones with 5′ UTRs longer than 500 nt were manually curated. Clones with a stop codon more than 55 nt 5′ to a splice junction and with a CDS at least 50% of the longest isoform CDS were accepted into MGC, but were annotated in the GenBank record as likely NMD candidates. All sequence differences between the cDNA sequences and their genome are annotated in the GenBank entry (misc_feature).

RNA editing analysis

Two tests were used to identify clones with putative A-to-G edits. For test 1, we followed Kim et al. (2004) to identify clones with at least one window of 100 nt that has: (1) more than five A-to-G changes and (2) more than half of the total number of differences with the genomic DNA as A-to-G changes. The 113 clones that meet these criteria are given in Supplemental Table S4. Putative edits reported as validated single-nucleotide polymorphisms (multiple observed polymorphisms or genotyped polymorphisms with minor allele frequency exceeding 2%) were discarded. No clones with equivalent windows of G-to-A changes were identified, although two clones with equivalent T-to-C changes and three clones with C-to-T changes were identified, suggesting a false-positive rate of ~2% in this list of A-to-G candidate edits.

Our second test identified clones that harbor at least one 100-nt window of sequence with enough changes of a single type that the probability of observing this window by chance is 10^{-8}. We defined the probability of observing a window with m changes as $P = 0.25Nr^m$ where r is the observed mismatch rate per clone and N is the number of genomic instances of the original nucleotide in the sequence window (number of As for A-to-I editing). (Since transitions are more common than transversions, 0.25 is a slight underestimate of the number of changes expected for any single type of transition.) We set P to 10^{-8}, which means that for each 100-nt window, we assessed whether there are m changes where $m = \log(10^{-8}/0.25N)/\log(r)$. We identified 118 clones with at least one such window of m A-to-G changes and two clones with G-to-A changes (Supplemental Table S5). These 118 clones include 87 clones identified by test 1 (Supplemental Table S6). Additional methods and results are described in Supplemental Text S3.

Complete list of authors

MGC program management team

Gary Temple,[2,3] Daniela S. Gerhard,[4] Rebekah Rasooly,[5] Elise A. Feingold,[3] Peter J. Good,[3] Cristen Robinson,[3] Allison Mandich,[3] Jeffrey G. Derge,[6] Jeanne Lewis,[6] Debonny Shoaf,[6] and Francis S. Collins[3]

Primary bioinformatics team

Wonhee Jang,[7] Lukas Wagner,[7] and Carolyn M. Shenmen[7]

Additional bioinformatics and MGC website

Leonie Misquitta,[8] Carl F. Schaefer,[8] Kenneth H. Buetow,[8] Tom I. Bonner,[9] Linda Yankie,[7] Ming Ward,[7] Lon Phan,[7], Alex Astashyn,[7] Garth Brown,[7] Catherine Farrell,[7] Jennifer Hart,[7] Melissa Landrum,[7] Bonnie L. Maidak,[7] Michael Murphy,[7] Terence Murphy,[7] Bhanu Rajput,[7] Lillian Riddick,[7] David Webb,[7] Janet Weber,[7] Wendy Wu,[7] Kim D. Pruitt,[7] and Donna Maglott[7]

Human gene predictions

Adam Siepel,[10] Brona Brejova,[10,11] Mark Diekhans,[12] Rachel Harte,[12] Robert Baertsch,[12] Jim Kent,[12] David Haussler,[12] Michael Brent,[13,14] Laura Langton,[13] Charles L.G. Comstock,[13] Michael Stevens,[14] Chaochun Wei,[13,15] Marijke J. van Baren,[13] Kourosh Salehi-Ashtiani,[16] Ryan R. Murray,[16] Lila Ghamsari,[16] Elizabeth Mello,[16] and Chenwei Lin[16,17]

cDNA clone management

Christa Pennacchio,[18,19] Kirsten Schreiber,[18,20] Nicole Shapiro,[18,21] Amber Marsh,[18,22] Elizabeth Pardes,[18,23] Troy Moore,[24] Anita Lebeau,[25] Mike Muratet,[25] Blake Simmons,[24] David Kloske,[24] Stephanie Sieja,[24] and James Hudson[25]

RNA editing analysis

Praveen Sethupathy[3]

mRNA preparation

Michael Brownstein,[9] Narayan Bhat,[7,26] Joseph Lazar,[27] and Howard Jacob[27]

cDNA library preparation

Chris E. Gruber[28] and Mark R. Smith[28]

cDNA full-insert cloning (PCR rescue) and sequencing

John McPherson,[29] Angela M. Garcia,[29] Preethi H. Gunaratne,[29,30] Jiaqian Wu,[29,31] Donna Muzny,[29] Richard A. Gibbs,[29] Alice C. Young,[32] Gerard G. Bouffard,[32,33] Robert W. Blakesley,[32,33] Jim Mullikin,[32,33] Eric D. Green,[32,33] Mark C. Dickson,[34,35] Alex C. Rodriguez,[34,36] Jane Grimwood,[34,37] Jeremy Schmutz,[34,37] Richard M. Myers,[34,37] Martin Hirst,[38] Thomas Zeng,[38] Kane Tse,[38] Michelle Moksa,[38] Merinda Deng,[38] Kevin Ma,[38] Diana Mah,[38] Johnson Pang,[38] Greg Taylor,[38] Eric Chuah,[38] Athena Deng,[38] Keith Fichter,[38] Anne Go,[38] Stephanie Lee,[38] Jing Wang,[38] Malachi Griffith,[38] Ryan Morin,[38] Richard A. Moore,[38] Michael Mayo,[38] Sarah Munro,[38] Susan Wagner,[38] Steven J.M. Jones,[38] Robert A. Holt,[38] Marco A. Marra,[38] Sun Lu,[39] and Shuwei Yang[39]

EST sequencing

James Hartigan[40]

DNA synthesis

Marcus Graf,[41] Ralf Wagner,[41] Stanley Letovksy,[42,43] Jacqueline C. Pulido,[42] and Keith Robison[42]

Synthetic clone functional analysis

Dominic Esposito,[44] James Hartley,[44] Vanessa E. Wall,[44] Ralph F. Hopkins,[44] Osamu Ohara,[45] and Stefan Wiemann[46]

[3]National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA.
[4]National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA.
[5]National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland 20892, USA.
[6]SAIC-Frederic, Inc., National Cancer Institute at Frederick, Frederick, Maryland 21702, USA.
[7]National Center for Biotechnology Information, National Library of Medicine, Bethesda, Maryland 20894, USA.
[8]National Cancer Institute, Center for Bioinformatics, Rockville, Maryland 20852, USA.
[9]National Institute of Mental Health, National Institutes of Health, Bethesda, Maryland 20892, USA.
[10]Biological Statistics and Computational Biology, Cornell University, Ithaca, New York 14853, USA.
[11]Present address: Department of Computer Science, Comenius University, 842 48 Bratislava, Slovakia.
[12]Center for Biomolecular Science & Engineering, University of California, Santa Cruz, California 95064, USA.
[13]Center for Genome Sciences, Washington University, St. Louis, Missouri 63130, USA.
[14]Department of Computer Science, Washington University, St. Louis, Missouri 63130, USA.
[15]Present address: School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240 and Shanghai Center for Bioinformation Technology, Shanghai 200235, China.
[16]Center for Cancer Systems Biology and Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, Massachusetts 02115, USA.
[17]Present address: Pediatrics Department, Stanford University School of Medicine, Stanford, California 94305, USA.
[18]The I.M.A.G.E. Consortium, Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, California 94550, USA.
[19]Present address: Lawrence Berkeley National Laboratory, Walnut Creek, California 94598, USA.
[20]Present address: Norgren Systems, Ronceverte, West Virginia 24970, USA.
[21]Present address: Department of Energy Joint Genome Institute, Walnut Creek, California 94598, USA.
[22]Present address: National Ignition Facility, Lawrence Livermore National Laboratory, Livermore, California 94550, USA.
[23]Present address: Computing Applications and Research Department, Lawrence Livermore National Laboratory, Livermore, California 94550, USA.
[24]Open Biosystems, now a part of ThermoFisher Scientific, Huntsville, Alabama 35806, USA.
[25]HudsonAlpha Institute for Biotechnology, Huntsville, Alabama 35806, USA.
[26]Present address: United States Patent and Trademark Office, Alexandria, Virginia 22314, USA.
[27]Department of Dermatology, Physiology, Medical College of Wisconsin, Milwaukee, Wisconsin 53226, USA.
[28]Express Genomics, Inc., Frederick, Maryland 21701, USA.
[29]Baylor College of Medicine Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas 77030, USA.
[30]University of Houston, Houston, Texas 77004, USA.
[31]Present address: Department of Molecular, Cellular and Developmental Biology, Yale University, New Haven, Connecticut 06620, USA.
[32]NIH Intramural Sequencing Center, National Human Genome Institute, National Institutes of Health, Bethesda, Maryland 20892, USA.
[33]Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA.
[34]Stanford Human Genome Center, Department of Genetics, Stanford University School of Medicine, Stanford, California 94305, USA.
[35]Present address: Cardiodx, Inc., Palo Alto, California 94303, USA.
[36]Present address: Baxter International, Inc., Deerfield, Illinois 60015, USA.
[37]Present address: HudsonAlpha Institute for Biotechnology, Huntsville, Alabama 35806, USA.
[38]Genome Sciences Centre, BC Cancer Agency, Vancouver BC, V5Z 4S6 Canada.
[39]GeneCopoeia Inc., Rockville, Maryland 20850, USA.
[40]Beckman Coulter Genomics, Beverly, Massachusetts 01915, USA.
[41]Geneart AG, Regensburg, Germany 93053.
[42]Codon Devices, Inc., Cambridge, Massachusetts 02139, USA.
[43]Present address: Helicos BioSciences Corporation, Cambridge, Massachusetts 02139, USA.
[44]Protein Expression Laboratory, NCI/SAIC-Frederick, Frederick, Maryland 21702, USA.
[45]Kazusa DNA Research Institute, Kisarazu, Chiba 292-0818, Japan.
[46]German Cancer Research Center, D-69120 Heidelberg, Germany.

Acknowledgments

We thank Robert L. Strausberg and Richard D. Klausner, who initiated the MGC project and helped to manage it during its first four years. The MGC Program received excellent guidance from members of the External Scientific Committee: Barbara Wold, Philip Sharp, Geoffrey Duyk, Connie Cepko, Stewart Scherer, Lincoln Stein, Ronald Davis, Howard Jacob, and Edward Harlow. The Mammalian Gene Collection Program was an NIH interinstitute effort that received financial and scientific support from 19 institutes within the NIH. A complete list of these institutes is provided on the MGC website. Greg Schuler and Karl Sirotkin provided valuable bioinformatics advice to MGC. We thank the Vineet Bafna laboratory for screening our predicted SEGs against its MS/MS spectra database. We also appreciate the valuable support and advice of the NIH MGC Inter-Institute Coordinating Committee: Andrea Beckel-Mitchener, Anthony Carter, Hemin Chin, Jennifer Couch, Pete Dudley, Nancy Freeman, Maria Giovanni, Q. Max Guo, John Harding, Steven Klein, Cheryl Kraft, Gabrielle Leblanc, Anna McCormick, Susan Old, Raymond O'Neill, Jane Peterson, Jonathan Pollock, Zhaoxia Ren, John Satterlee, William Sharrock, Rochelle Small, Phillip Smith, Susan Sparks, Danilo Tagle, and Jose Velazquez. Assya Abdallah provided excellent assistance with the final manuscript preparation. The XGC Program received excellent guidance from Aaron Zorn, Ken Cho, Bruce Blumberg, Richard Harland, Robert Grainger, and Jane Rogers. The ZGC Program received excellent guidance from the members of its advisory committee: Bruce Birren, Will Talbot, Monte Westerfield, and Len Zon. Four rat cDNA libraries were contributed by the Cancer Genome Anatomy Project, NCI. This project was funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, under contract no. N01-C0-12400. The portion of this work carried out at the Center for Cancer Systems Biology, at the Dana-Farber Cancer Institute, was funded in part by a grant from the Ellison Foundation.

References

Athanasiadis A, Rich A, Maas S. 2004. Widespread A-to-I RNA editing of *Alu*-containing mRNAs in the human transcriptome. *PLoS Biol* **2:** e391. doi: 10.1371/journal.pbio.0020391.

Baross A, Butterfield YS, Coughlin SM, Zeng T, Griffith M, Griffith OL, Petrescu AS, Smailus DE, Khattra J, McDonald HL, et al. 2004. Systematic recovery and analysis of full-ORF human cDNA clones. *Genome Res* **14:** 2083–2092.

Bass BL. 2002. RNA editing by adenosine deaminases that act on RNA. *Annu Rev Biochem* **71:** 817–846.

Blow M, Futreal PA, Wooster R, Stratton MR. 2004. A survey of RNA editing in human brain. *Genome Res* **14:** 2379–2387.

Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, et al. 2005. The transcriptional landscape of the mammalian genome. *Science* **309:** 1559–1563.

Carthew RW, Sontheimer EJ. 2009. Origins and mechanisms of miRNAs and siRNAs. *Cell* **136:** 642–655.

The ENCODE Project Consortium. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447:** 799–816.

Gerhard DS, Wagner L, Feingold EA, Shenmen CM, Grouse LH, Schuler G, Klein SL, Old S, Rasooly R, Good P, et al. 2004. The status, quality, and expansion of the NIH full-length cDNA project: The Mammalian Gene Collection (MGC). *Genome Res* **14:** 2121–2127.

Glusman G, Qin S, El-Gewely MR, Siegel AF, Roach JC, Hood L, Smit AF. 2006. A third approach to gene prediction suggests thousands of additional human transcribed regions. *PLoS Comput Biol* **2:** e18. doi: 10.1371/journal.pcbi.0020018.

Gommans WM, Dupuis DE, McCane JE, Tatalias NE, Maas S. 2008. Diversifying exon code through A-to-I RNA editing. In *RNA and DNA editing: Molecular mechanisms and their integration into biological systems* (ed. HC Smith), pp. 3–30. Wiley, New York.

Gross SS, Brent MR. 2006. Using multiple alignments to improve gene prediction. *J Comput Biol* **13:** 379–393.

Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP, et al. 2009. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458:** 223–227.

Hartley JL, Temple GF, Brasch MA. 2000. DNA cloning using in vitro site-specific recombination. *Genome Res* **10:** 1788–1795.

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.

Hubbard TJ, Aken BL, Ayling S, Ballester B, Beal K, Bragin E, Brent S, Chen Y, Clapham P, Clarke L, et al. 2009. Ensembl 2009. *Nucleic Acids Res* **37:** D690–D697.

Kapranov P, Willingham AT, Gingeras TR. 2007. Genome-wide transcription and the implications for genomic organization. *Nat Rev Genet* **8:** 413–423.

Karolchik D, Kuhn R, Baertsch R, Barber G, Clawson H, Diekhans M, Giardine B, Harte R, Hinrichs A, Hsu F, et al. 2008. The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res* **36:** D773–D779.

Kent WJ. 2002. BLAT—the BLAST-like alignment tool. *Genome Res* **12:** 656–664.

Kim DD, Kim TT, Walsh T, Kobayashi Y, Matise TC, Buyske S, Gabriel A. 2004. Widespread RNA editing of embedded *Alu* elements in the human transcriptome. *Genome Res* **14:** 1719–1725.

Kuhn RM, Karolchik D, Zweig AS, Wang T, Smith KE, Rosenbloom KR, Rhead B, Raney BJ, Pohl A, Pheasant M, et al. 2009. The UCSC Genome Browser Database: Update 2009. *Nucleic Acids Res* **37:** D755–D761.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, 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.

Levanon EY, Eisenberg E, Yelin R, Nemzer S, Hallegger M, Shemesh R, Fligelman ZY, Shoshan A, Pollock SR, Sztybel D, et al. 2004. Systematic identification of abundant A-to-I editing sites in the human transcriptome. *Nat Biotechnol* **22:** 1001–1005.

Li JB, Levanon EY, Yoon JK, Aach J, Xie B, Leproust E, Zhang K, Gao Y, Church GM. 2009. Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. *Science* **324:** 1210–1213.

Pheasant M, Mattick JS. 2007. Raising the estimate of functional human sequences. *Genome Res* **17:** 1245–1253.

Pruitt KD, Harrow J, Harte RA, Wallin C, Diekhans M, Maglott DR, Searle S, Farrell CM, Loveland JE, Ruef BJ, et al. 2009a. The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res* **19:** 1316–1323.

Pruitt KD, Tatusova T, Klimke W, Maglott DR. 2009b. NCBI Reference Sequences: Current status, policy and new initiatives. *Nucleic Acids Res* **37:** D32–D36.

Siepel A, Haussler D. 2004. Computational identification of evolutionarily conserved exons. In *Proceedings of the Eighth International Conference on Research in Computational Molecular Biology*, pp. 177–186. ACM Press, New York.

Siepel A, Diekhans M, Brejova B, Langton L, Stevens M, Comstock CL, Davis C, Ewing B, Oommen S, Lau C, et al. 2007. Targeted discovery of novel human exons by comparative genomics. *Genome Res* **17:** 1763–1773.

Strausberg RL, Feingold EA, Klausner RD, Collins FS. 1999. The mammalian gene collection. *Science* **286:** 455–457.

Strausberg RL, Feingold EA, Grouse LH, Derge JG, Klausner RD, Collins FS, Wagner L, Shenmen CM, Schuler GD, Altschul SF, 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.

Tanner S, Shen Z, Ng J, Florea L, Guigo R, Briggs SP, Bafna V. 2007. Improving gene annotation using peptide mass spectrometry. *Genome Res* **17:** 231–239.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, et al. 2001. The sequence of the human genome. *Science* **291:** 1304–1351.

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, 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.

Wei C, Brent MR. 2006. Using ESTs to improve the accuracy of de novo gene prediction. *BMC Bioinformatics* **7:** 327. doi: 10.1186/1471-2105-7-327.

Wilming LG, Gilbert JG, Howe K, Trevanion S, Hubbard T, Harrow JL. 2008. The Vertebrate Genome Annotation (Vega) database. *Nucleic Acids Res* **36:** D753–D760.

Wu JQ, Garcia AM, Hulyk S, Sneed A, Kowis C, Yuan Y, Steffen D, McPherson JD, Gunaratne PH, Gibbs RA. 2004. Large-scale RT-PCR recovery of full-length cDNA clones. *Biotechniques* **36:** 690–700.

Zhu J, Sanborn JZ, Diekhans M, Lowe CB, Pringle TH, Haussler D. 2007. Comparative genomics search for losses of long-established genes on the human lineage. *PLoS Comput Biol* **3:** e247. doi: 10.1371/journal.pcbi.0030247.

Received May 12, 2009; accepted in revised form September 8, 2009.