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

Follow this and additional works at: http://digitalcommons.wustl.edu/open_access_pubs

Recommended Citation
Brent, Michael; Langton, Laura; Comstock, Charles L.G.; Stevens, Michael; Wei, Chaochun; van Baren, Marijke J.; and et al., “The completion of the Mammalian Gene Collection (MGC).” Genome Research.19, 2324-2333. (2009).
http://digitalcommons.wustl.edu/open_access_pubs/1941

This Open Access Publication is brought to you for free and open access by Digital Commons@Becker. It has been accepted for inclusion in Open Access Publications by an authorized administrator of Digital Commons@Becker. For more information, please contact engeszer@wustl.edu.
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

**Supplemental Material**
http://genome.cshlp.org/content/suppl/2009/09/22/gr.095976.109.DC1.html

**References**
This article cites 35 articles, 21 of which can be accessed free at:
http://genome.cshlp.org/content/19/12/2324.full.html#ref-list-1

**Creative Commons License**
This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see http://genome.cshlp.org/site/misc/terms.xhtml). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 3.0 Unported License), as described at http://creativecommons.org/licenses/by-nc/3.0/.

**Email Alerting Service**
Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or click here.

To subscribe to *Genome Research* go to:
http://genome.cshlp.org/subscriptions

Copyright © 2009 by Cold Spring Harbor Laboratory Press
The completion of the Mammalian Gene Collection (MGC)

The MGC Project Team

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/]
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 DNA synthesis, 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.

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.

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 DNA synthesis, 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.

Multiple-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...
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
Supplemental Table S2 presents the rates of sequence discrepancy, based on total sequence 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 $9.1 \times 10^{-4}$, 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 $\times 10^{-4}$) 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, 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.

Figure 3. PCR rescue success versus target size. (Black bars) The number of assigned targets in each size range; (white bars) 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 suppressed, or withdrawn 233 in total)

Figure 4. Synthesis success versus target size. (Black bars) The number of assigned targets in each size range; (white bars) 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 (233 in total)
Table 1. MGC achievement

<table>
<thead>
<tr>
<th>Gene classes(^a)</th>
<th>Protein-coding genes</th>
<th>Protein-coding genes in MGC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Human</td>
<td>Mouse</td>
</tr>
<tr>
<td>A. All genes with curated RefSeq transcripts(^b)</td>
<td>18,877</td>
<td>19,357</td>
</tr>
<tr>
<td>B. Genes with ≥1 PubMed articles and curated RefSeq transcripts(^b)</td>
<td>14,614</td>
<td>12,434</td>
</tr>
<tr>
<td>C. Genes with known disease phenotype(^c)</td>
<td>2306</td>
<td>2208</td>
</tr>
<tr>
<td>D. Genes with CCDS transcripts(^d)</td>
<td>13,884</td>
<td>15,263</td>
</tr>
</tbody>
</table>

\(^a\)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://mge.ncbi.nlm.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. Non-synonymous (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 (≤0.2 × 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/QCoverall.shtml.

Revised genome annotation

While the MGC PCR rescue program was under way, concurrent progress in human and mouse genome annotations forced MGC
to retire and assign ~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', exte nding 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; Gutman 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
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.

\*11% are in the CDS (0.79% total discrepancies).

\*NS Discrepancies in CDS 20,205 (0.72) 27,188 0.89 0.64 54.6%

\*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).

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) do not indicate that the CDS is partial; (2) the reading frame is consistent with RefSeq and supported by ORFeome Collaboration clones aligned with individual human genes. Links lead to additional information on the clone, associated protein, and to "Order cDNA Clone."
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 (TAG) 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 one version, with a stop codon (TAG). Subsequent assignments to both companies requested only in two versions, with a stop codon (TAA) and without a stop codon (TAC). Subsequent assignments to both companies requested only in two versions, with a stop codon (TAA) and without a stop codon (TAC). Subsequent assignments to both companies requested only in two versions, with a stop codon (TAA) and without a stop codon (TAC).

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 RTBank 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 \( P = 0.25N^m \) 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. Collins3

Genome Research www.genome.org 2331

Downloaded from genome.cshlp.org on December 18, 2013 - Published by Cold Spring Harbor Laboratory Press
The MGC Project Team

Primary bioinformatics team
Wonhee Jang, Lukas Wagner, and Carolyn M. Shennan

Additional bioinformatics and MGC website

Human gene predictions
Adam Siepel, Brona Brejova, Mark Diekhans, Rachel Harte, Robert Baertsch, Jim Kent, David Haussler, Michael Brent, Laura Langton, Charles L.G. Comstock, Michael Stevens, Chaochun Wei, Marijke J. van Baren, Kourosh Salehi-Ashtiani, Ryan R. Murray, Lila Ghamsari, Elizabeth Mullikin, Eric D. Green, Mark C. Dickson, Alex C. Howard, Jacob Harte, Robert Baertsch, Jim Kent, David Haussler, Michael Holt, Marco A. Marra, Sun Lu, and Shuwei Yang

RNA editing analysis
Praveen Sethupathy

mRNA preparation
Michael Brownstein, Narayan Bhat, Joseph Lazar, and Howard Jacob

cDNA library preparation
Chris E. Gruber and Mark R. Smith

DNA cloning (PCR rescue) and sequencing
John McPherson, Angela M. Garcia, Preethi H. Gunaratne, Jiaqian Wu, Donna Munzy, Richard A. Gibbs, Alice C. Young, Gerard G. Bouffard, Robert W. Blakesley, Jim Mullikin, Eric D. Green, Mark C. Dickson, Alex C. Rodriguez, Jane Grimwood, Jeremy Schmutz, Richard M. Myers, Martin Hirst, Thomas Zeng, Kane Tse, Michelle Moksa, Merinda Deng, Kevin Ma, Diana Mah, Johnson Pang, Greg Taylor, Eric Chuah, Athena Deng, Keith Fichter, Anne Go, Stephanie Lee, Jing Wang, Malachi Griffith, Ryan Morin, Michael Mayo, Sarah Munro, Susan Wagner, Steven J.M. Jones, Robert A. Holt, Marco A. Marra, Sun Lu, and Shuwei Yang

EST sequencing
James Hartigan

DNA synthesis
Marcus Graf, Ralf Wagner, Stanley Letovksy, Jacqueline C. Pulido, and Keith Robison

Synthetic clone functional analysis
Dominic Esposito, James Hartley, Vanessa E. Wall, Ralph F. Hopkins, Osamu Ohara, and Stefan Wiemann

1National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA.
2National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA.
3National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland 20892, USA.
4SAIC-Frederic, Inc., National Cancer Institute at Frederick, Frederick, Maryland 21702, USA.
5National Center for Biotechnology Information, National Library of Medicine, Bethesda, Maryland 20894, USA.
6National Cancer Institute, Center for Bioinformatics, Rockville, Maryland 20852, USA.
7National Institute of Mental Health, National Institutes of Health, Bethesda, Maryland 20892, USA.
8Biological Statistics and Computational Biology, Cornell University, Ithaca, New York 14853, USA.
9Department of Computer Science, Comenius University, 842 48 Bratislava, Slovakia.
10Department of Molecular Molecular, Celluar and Developmental Biology, Yale University, New Haven, Connecticut 06620, USA.
11NIH Intramural Sequencing Center, National Human Genome Institute, National Institutes of Health, Bethesda, Maryland 20892, USA.
12Center for Cancer Systems Biology and Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, Massachusetts 02115, USA.
13Present address: Pediatrics Department, Stanford University School of Medicine, Stanford, California 94305, USA.
14The I.M.A.G.E. Consortium, Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, California 94550, USA.
15Present address: Lawrence Berkeley National Laboratory, Walnut Creek, California 94598, USA.
16Present address: School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240 and Shanghai Center for Bioinformation Technology, Shanghai 200235, China.
17Present address: Center for Cancer Systems Biology and Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, Massachusetts 02115, USA.
18Present address: Department of Pediatrics, Stanford University School of Medicine, Stanford, California 94305, USA.
19Present address: Department of Computer Science, Washington University, St. Louis, Missouri 63130, USA.
20Primary address: Department of Computer Science, Washington University, St. Louis, Missouri 63130, USA.
21Present address: School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240 and Shanghai Center for Bioinformation Technology, Shanghai 200235, China.
22Present address: Lawrence Livermore National Laboratory, Livermore, California 94550, USA.
23Present address: Lawrence Berkeley National Laboratory, Walnut Creek, California 94598, USA.
24Present address: National Ignition Facility, Lawrence Livermore National Laboratory, Livermore, California 94550, USA.
25Present address: Computing Applications and Research Department, Lawrence Livermore National Laboratory, Livermore, California 94550, USA.
26Present address: Open Biosystems, now a part of ThermoFisher Scientific, Huntsville, Alabama 35806, USA.
27Present address: HudsonAlpha Institute for Biotechnology, Huntsville, Alabama 35806, USA.
29Department of Dermatology, Physiology, Medical College of Wisconsin, Milwaukee, Wisconsin 53226, USA.
30Genexpress Genomics, Inc., Frederick, Maryland 21701, USA.
31Baylor College of Medicine Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas 77030, USA.
32University of Houston, Houston, Texas 77004, USA.
33Present address: Department of Molecular, Celluar and Developmental Biology, Yale University, New Haven, Connecticut 06620, USA.
34Present address: National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA.
35Present address: National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA.
36Present address: Department of Genetics, Stanford University School of Medicine, Stanford, California 94305, USA.
37Present address: Cardioids, Inc., Palo Alto, California 94303, USA.
38Present address: Baxter International, Inc., Deerfield, Illinois 60015, USA.
39Present address: Helicos Biosciences Corporation, Cambridge, Massachusetts 02139, USA.
40Protein Expression Laboratory, NCI/SAIC-Frederick, Frederick, Maryland 21702, USA.
41GeneCopoeia Inc., Rockville, Maryland 20850, USA.
42Beckman Coulter Genomics, Beverly, Massachusetts 01915, USA.
43Present address: Codon Devices, Inc., Cambridge, Massachusetts 02139, USA.
44Protein Expression Laboratory, NCI/SAIC-Frederick, Frederick, Maryland 21702, USA.
45Kazusa DNA Research Institute, Kisarazu, Chiba 292-0818, Japan.
46German Cancer Research Center, D-69120 Heidelberg, Germany.

www.genome.org
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 Schulier and Karl Sirotkin provided valuable bioinformatics advice to MGC. We thank the Vitek 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 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
Bass BL. 2002. RNA editing by adenosine deaminases that act on RNA.

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