Supplementary Information Section 21

Sequence Strategy

Depending in part upon the length, and in part upon the evolution of technologies during the course of this project, individual cDNAs were sequenced using a variety of strategies, including one-pass sequencing from both ends of the clone, shotgun sequencing after PCR of the insert and concatenation of 48 inserts of similar size, transposon shotgun sequencing, and primer walking. The cDNA sequences were assembled using the PHRED/PHRAP/CONSED [1-3] system, taking advantage of the previously collected 5' and 3' end sequences. Each assembly was checked for cloning site sequences and cloning direction to eliminate potential artefacts. In total, we generated 60,770 high-quality, full-insert cDNA sequences.

Clustering with ClusTrans [Supplementary Information 19]

Pairs of cDNAs sharing 98% or greater identity over 500 bp or more were placed into clusters. cDNA pairs with greater than 98 % identity but match lengths between 250 bp and 500 bp were clustered if the shortest of the non matching regions was less than 100 bp. Finally, clones that did not meet the above criteria but matching another clone at greater than 98% over more than 95 % of its length were clustered with their best match. Using these criteria, the initial 60,770 were placed into 35,367 candidate clusters. ClusTrans has also the ability to select a representative cDNA from each cluster. Representatives were selected using a simple comparison of sum of pairwise alignment scores. For example, when sequences A, B, and C are in a cluster, sequence A is selected as the representative if Alignment score(A:B)+ Alignment score (A:C)> Alignment score (B:C)+Alignment score(B:A)> Alignment score (C:A)+ Alignment score (C:B). In case where only a pair of sequence is in the cluster, the sequence with the longest predicted ORF was selected. For clustering of the FANTOM 2 dataset with

known genes from the public databases, a modified version of ClusTrans was used with SSEARCH replaced with BLAT [4] as BLAT produced identical clusters with a great reduction in the time required for pairwise searches (unpublished).

Creating a Representative Transcript and Protein Set (RTPS) [Supplementary Information 9]

To derive a comprehensive yet unique Representative Transcript and Protein Set (RTPS), it was necessary to address the inherent redundancy in FANTOM2 data and in the sources of known mouse transcript-based sequence data. Our assumption in this analysis is that when two nucleotide sequences of cDNAs have a long identical region, these sequences should be placed into a single cluster (Supplementary Information 19). The 60,770 FANTOM 2 cDNA sequences and 28,330 mouse transcript sequences compiled from the LocusLink, Mouse Genome Informatics (non-EST) and GenBank (non-EST) databases were clustered initially by our cDNA clustering method ClusTrans to reduce redundancies within the FANTOM2 cDNA set.

This clustering algorithm also selected a single sequence from each FANTOM2 cluster as the best (or Representative) sequence for the cluster, based on the quality and length of clustered sequences and their CDS regions. For a subset of the FANTOM2 clusters that showed significant sequence similarity to genes in Mouse Genome Informatics database (MGI), MATRICS curators evaluated cluster integrity, taking advantage of three independent cluster builds of FANTOM2 data (a FANTOM2-only build by the RIKEN group, and two builds of FANTOM2 + all mouse cDNAs and ESTs by UniGene and TIGR Gene Index, respectively) and the extensive cluster viewing tools available in the FANTOM2 web interface. The validity of computational sequence clustering was addressed by examining the biological context of sequence alignments, including clone orientation, repetitive sequence, transcript processing, close

paralogues, and other factors. Where FANTOM2 clusters were found to be identical to known genes in MGI, relationships were established. Clusters were split if sequences were found to be inappropriately grouped. Curators then selected a representative FANTOM 2 sequence for each resolved cluster, based on CDS integrity, sequence quality and overall sequence length.

A multi-layered approach was used to integrate the 35,637 cluster-resolved Representative FANTOM2 sequences with the set of known mouse transcripts compiled from public databases. To produce a high quality representative transcript set for further analysis, we avoided mouse sequences in the EST division of GenBank during construction of the RTPS. A table was used to track transcription unit redundancy and the information needed to select Representative Transcripts and Proteins for each unique TU. We started with a curated set of 19,401 representative transcript sequences from mouse LocusLink records as a core of known mouse transcripts, followed by additional RNA-derived sequences from all records in MGI that were not associated with the starting LocusLink sequences. Then, all remaining mouse GenBank RNA-derived (non-EST) sequences that were not contained in MGI or LocusLink were added, along with all 35,637 cluster-resolved Representative FANTOM2 sequences. In all, 63,967 mouse transcript sequences were considered, including 28,330 non-FANTOM2 cDNAs from the public domain.

Redundancy detection and tracking followed a stepwise process where sequences determined to be from the same TUs were assigned the same TU identifiers. Relationships between sequences in curated databases were considered first, followed by the curated relationships between FANTOM2 sequences and MGI genes established during cluster curation. To integrate the remaining FANTOM2 Representative sequences and non-RIKEN sequences from public databases with these TUs, the sequences were clustered using the same algorithm that clustered the complete

FANTOM2 set earlier (Adachi et al. manuscript in preparation). For this step, sequences with no TU assignments to this point that clustered together were assigned new TU IDs, and those that clustered with existing TUs were grouped with these TUs. This step did not merge existing TUs defined by the original core transcript set, but resolved the input transcripts to 49,062 TUs. Curators then inspected this set and removed TUs defined by partial, repetitive or other sequences considered redundant in the context of our TU definition. To detect additional redundancy and reveal instances of over clustering, all sequences that contributed to RTPS construction were mapped to the mouse genome assembly. For this step, a preliminary representative sequence was selected for each TU following the rules stated below. If two or more representative or unassigned sequences overlapped by at least one base pair in an exon, then these were merged into a single TU. For TUs representing multiple sequences in the table, if any sequence associated with the TU mapped to a different chromosome than the representative sequence, then that TU was split into separate TUs for each different chromosome. The net effect of genome mapping was condensation of the TU number to 36,089 TUs. Finally, to maintain consistency with our TU definition, which considers opposite strand transcripts as separate TUs, we looked for sequences from the same TU with opposite strand orientation, and established new TUs for these sequences. This step resulted in an increase in the number of TUs to 37,086.

After completing detection of TU sequence redundancy, a single transcript sequence was chosen to represent each TU following rules of a selection hierarchy, in the following order, best cluster sequence from curated clusters > LocusLink core transcript sequence > clustering algorithm best choice > clustering algorithm 2nd best choice > sequence with longest CDS > longest transcript sequence. For protein coding TUs, a single polypeptide sequence was chosen to represent each TU with the following selection rules, longest SWISS-PROT record > longest RefSeq protein (NP_) record >

longest CDS sequence. A total of 18,768 TUs contain a representative protein sequence.

A Variant-based Proteome Set (VPS) was prepared by identifying alternative transcription products within FANTOM2 cDNA clusters that have different CDS sequences compared to the CDS sequences of cluster representatives (J. Adachi *et al*. manuscript in preparation). Each FANTOM2 clone with a unique CDS due to alternative transcript production, contributed to the VPS. The VPS contains all 18,768 protein sequences in the RTPS plus and 15,513 coding sequences.

Proteins domain/motif analysis and functional assignment

The constructed mouse proteome sequence set was analyzed using a number of protein domain and motif databases including InterPro [5], Superfamily [6] and MDS [7]. The sequences were pre-processed before the domain/motif scanning to identify identical matches to entries in public protein sequences databases. We used IPI (International Protein Index, http://www.ebi.ac.uk/IPI), which contains a nonredundant set of sequences derived from SPTR, EnsEMBL, and RefSeq.

1. Ewing, B., Hillier, L., Wendl, M. C. & Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* **8**, 175-85. (1998).

2. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* **8**, 186-94. (1998).

3. Gordon, D., Abajian, C. & Green, P. Consed: a graphical tool for sequence finishing. *Genome Res* **8**, 195-202. (1998).

4. Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome Res* **12**, 656-64. (2002).

5. Apweiler, R. et al. InterPro--an integrated documentation resource for protein families, domains and functional sites. *Bioinformatics* **16**, 1145-50. (2000).

6. Gough, J., Karplus, K., Hughey, R. & Chothia, C. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J Mol Biol* **313**, 903-19. (2001).

7. Kawaji, H. et al. Exploration of novel motifs derived from mouse cDNA sequences. *Genome Res* **12**, 367-78. 2002 [doi] (2002).