Supporting Methods

Serial Analysis of Gene Expression (SAGE) Protocol. Mouse tissue samples were collected in either RNAlater (Ambion) or TRIzol reagent (Invitrogen) or they were snap-frozen by using liquid nitrogen. Mechanical homogenization of tissues was performed by using a PowerGen 125 rotor-stator homogenizer (Fisher) and disposable generators, except for visual cortex tissue, for which a handheld Polytron PT 1200CL homogenizer (Kinematica, through Brinkmann Instruments), equipped with a 7-mm easy care generator (PT-DA 1207/2EC), was used. Total RNA was extracted with TRIzol reagent (Invitrogen) and phase-lock gel tubes (Brinkmann), according to the manufacturer’s protocol. Lipid-rich or fibrous samples were extracted with RNEasy Lipid or Fibrous kits (Qiagen, Valencia, CA). RNA was assessed for quality and quantified by using an Agilent 2100 Bioanalyzer (Agilent Technologies) and an RNA 6000 Nano or Pico LabChip kit (Caliper Technologies).

LongSAGE (1) libraries were constructed with at least 5 μg of DNase I- (Invitrogen) or DNA-free- (Ambion) treated total RNA by using the Invitrogen I-SAGE Long kit and protocol. Scale-up PCR was performed for 23 to 27 cycles on varying dilutions (1/20 to 1/80) of template and two 96-well plates with 50-μl reactions per well. Concatemers were cloned into pZErO-1 vectors and transformations performed by using One Shot TOP10 Electrocompetent *Escherichia coli*. After screening of transformants by colony PCR, the concatemer-size fractions were chosen for sequencing. Colony-picking was performed by using a Q-Pix robot (Genetix) and inoculations made into 2× YT media with 50 μg/ml zeocin and 7.5% glycerol. After overnight culture, glycerol stocks were used to inoculate larger-volume cultures for plasmid preparation by using a standard alkaline lysis procedure adapted for high-throughput processing with microtiter plates. DNA sequencing was performed with BigDye v3.1 dye terminator cycle sequencing reactions run on Tetrad thermal cyclers (MJ Research). Sequencing reaction products were purified by ethanol precipitation and analyzed on model 3700 and 3730xl capillary DNA sequencers (Applied Biosystems). These template preparation and sequencing protocols were described by Yang *et al.* (2).

Sequence data were collected automatically by using a custom DNA-sequencing laboratory information management system and processed by trimming reads for sequence quality and removal of nonrecombinant clones and linker-derived tags. Sufficient clones were sequenced to yield ≈100,000 LongSAGE tags per library. On average, 34 LongSAGE tags resulted from each sequencing read.

Samples with limiting (submicrogram) amounts of total RNA were subject to an amplification step similar to the SAGELite method (3). The chemistry for these amplified libraries is based on the SMART cDNA synthesis strategy (Clontech) for the generation of full-length cDNA libraries. In SMART (Switching Mechanism At the 5' end of RNA Transcripts) cDNA synthesis, only full-length first-strand cDNAs are extended with a polyC tail by a terminal transferase property inherent to the reverse transcriptase. A synthetic oligonucleotide with a 3' polyG stretch hybridizes to the first-strand cDNA and serves as a template for further extension of the cDNA. Thus, each full-length first-strand cDNA has incorporated a synthetic 5' priming site and a 3' site (biotinylated OligodT) which allow the cDNAs to be amplified by a subsequent PCR step. After PCR amplification, the cDNA is processed by the standard LongSAGE protocol.

LongSAGE Processing Pipeline. After sequencing, flanking vector sequences were removed and the tags extracted from each sequence read. The SAGE protocols generated concatemers in which the tags were present in pairs (ditags). A sequence quality factor (*QF*) was derived for each tag by using the following formula:

$$QF = \prod_{S=firstBase}^{S=lastBase} \left(1 - 10^{-S/10}\right),$$

where S is the PHRED score (4) for a particular base and the value is calculated over all bases in the tag. The quality factor was used in the calculation of tag-sequence-probability values.

Probability Values. Please refer to www.bcgsc.ca/downloads/genex/mouse_atlas/bioinfo_methods.

Tag-Sequence Clustering. Please refer to www.bcgsc.ca/downloads/genex/mouse_atlas/bioinfo_methods.

Tag-Sequence Mapping. Tag sequences were mapped to the genome sequence, MGC genes (ftp://ftp.ncbi.nih.gov/repository/MGC/MGC.sequences), RefSeq genes (ftp://ftp.ncbi.nih.gov/refseq/daily/), and Ensembl genes (Ensembl v20). All mappings were transformed to genomic coordinates (chromosome, position, and strand) on the mouse sequence (assembly 32) (5), with the aid of the Ensembl PERL API (application programming interface) (6). The mapping of RefSeq genes to genome contigs used data from Ensembl. The mapping of Mammalian Gene Collection (MGC) genes to genome contigs used data from the University of California, Santa Cruz (UCSC) Genome Browser site (7). If the mapping could not be transformed to genomic coordinates (mostly the result of inconsistencies between different databases and the failure of transformation routines to convert contig coordinates to chromosomal coordinates), the original mapping information was retained. For this article, tag sequences mapping to multiple positions on the genome were used only to determine the percentage of tags mapped.

We counted gene identifiers to calculate the number of gene loci represented by the data. To avoid double-counting different identifiers used to name the same gene in different databases, identifiers found at the same genomic location were assumed to represent the same gene.

“Known” Ensembl genes are those confirmed by full-length sequences deposited in public sequence databases. “Novel” Ensembl genes are those predicted by computational methods and confirmed by ESTs.

RT-PCR Validation. An RT-PCR method was used to confirm the presence of transcripts corresponding to singleton longSAGE tags that hit unannotated genomic sequence. The singleton tags were filtered by removing those that matched against RefSeq sequences (standard, X, and GS), MGC sequences, UniGene sequences, Ensembl EST genes, and Ensembl mappings of ESTs onto the genome. PCR primers were designed by using genomic sequence, Primer3 (8), and custom scripts to generate amplicons with an average length of 120 bp. Primers were designed that flank the tag sequence such that the tag would be included in the amplicon. The amplicons were each amplified from RNA representing the developmental stage and tissue in which the singleton tag was observed. DNase-treated RNAs (2 μg) remaining from the construction of the LongSAGE libraries was used as template to produce cDNA by using an Oligo(dT)$_{20}$ primer and the SuperScript III First-Strand Synthesis system (Invitrogen) following the manufacture's recommended protocol. The cDNA was amplified by using the Phusion High-Fidelity PCR kit (MJ Research, Cambridge, MA) following the manufacture's recommended protocol, with the addition of DMSO to a final concentration of 3%. The cycling conditions consisted of an initial denaturation at 98°C for 30 sec, followed by 10 touchdown PCR cycles starting with 98°C for 10 sec, 72°C (decreased by 1°C in each subsequent cycle) for 15 sec, 72°C for 30 sec, and 29 cycles of 98°C for 10 sec, 62°C for 15 sec, and 72°C for 30 sec, followed by an extension at 72°C for 10 min. Two microliters of the PCR reaction for each sample was loaded on a 3% MetaPhor Agarose gel (Cambrex, Walkersville, MD) and resolved for 3.5 h at 110 mA in 1XTBE cooled to 4°C. The gel was stained with SYBR green (Mandel) and visualized by using a Typhoon 9400 Variable Mode Imager (Amersham Pharmacia). Control experiments (data not shown) demonstrated that amplicons were RNA-dependent: RNase-A-treated RNA samples failed to produce amplicons, indicating that amplicons were derived from RNA and not from genomic DNA potentially contaminating the RNA.

Representation of Gene Families. Genes in each category were identified by their gene ontology (GO) classification (9), except for those reported in Messina *et al.* (10), which were taken directly from that paper. The GO classification of human genes was used and the mouse orthologue determined by using the Ensembl database.

1. Saha, S., Sparks, A. B., Rago, C., Akmaev, V., Wang, C. J., Vogelstein, B., Kinzler, K. W. & Velculescu, V. E. (2002) *Nat. Biotechnol.* **20,** 508–512.

2. Yang, G. S., Stott, J. M., Smailus, D., Barber, S. A., Balasundaram, M., Marra, M. A. & Holt, R. A. (2005) *BMC Genomics* **6,** 2.

3. Peters, D. G., Kassam, A. B., Yonas, H., O'Hare, E. H., Ferrell, R. E. & Brufsky, A. M. (1999) *Nucleic Acids Res.* **27,** e39.

4. Ewing, B. & Green, P. (1998) *Genome Res* **8,** 186–194.

5. Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., *et al.* (2002) *Nature* **420,** 520–562.

6. Stabenau, A., McVicker, G., Melsopp, C., Proctor, G., Clamp, M. & Birney, E. (2004) *Genome Res.* **14,** 929–933.

7. Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M. & Haussler, D. (2002) *Genome Res.* **12,** 996–1006.

8. Rozen, S. & Skaletsky, H. (2000) *Methods Mol. Biol.* **132,** 365–386.

9. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., *et al.* (2000) *Nat. Genet.* **25,** 25–29.

10. Messina, D. N., Glasscock, J., Gish, W. & Lovett, M. (2004) *Genome Res.* **14,** 2041–2047.