Supporting Information

Han et al. 10.1073/pnas.0902417106

SI Methods

Sequencing Quality, Trimming, and Mapping. Sequencing quality was illustrated by building base score distributions on Galaxy (http://galaxy.bx.psu.edu/) (Fig. S1). First, go to Galaxy website and upload the sequencing quality file. Raw scores from sequencing results were used. Then click "Short Read Analysis" and go to "Build distribution of base quality". According to the median value in the distribution, we set the cut-off length to be 27 bp for maintaining a balance between read length and its quality.

Mouse genome sequence from the latest assembly (mm9, July 2007) was downloaded from UCSC genome browser (http://genome.ucsc.edu/). In the analysis of single-end data, all chromosome sequences were concatenated to a long pseudosequence, because Rmap, the mapping program we used, only tolerated 1 sequence as target. We then used Rmap (version 0.41, mismatch 3, read length 27, seed length 11) for aligning the high-quality portion of reads to the mouse genome. In the analysis of paired-end data, we followed the same process and mapped each end of the paired-end reads respectively to the target sequence. Later, the pairs were restored using the read ids. There were 2 additional criteria: 1) both ends must be mapped to the same gene; 2) 1 end must be mapped at the positive and the other on the negative strand. We did not use the distance between a pair of ends as a threshold because of the unknown intron lengths. Only uniquely mapped reads were selected for further analyses. For paired-end data, we required both ends were uniquely mapped.

Rmap generated 1 file of mapping position of uniquely mapped reads. This file follows BED format defined by the UCSC Genome Browser. The mapping positions in the second and third column from Rmap output were coordinates in the pseudosequence and they were then transformed back into coordinates of the July 2007 mouse genome assembly.

Assigning Mapped Reads to Known or Predicted Genes (UCSC Clusters). We downloaded coordinates of exons UCSC genome browser through Galaxy using the following procedure. First, go to the Galaxy website, click "get data" and go to "UCSC Main table browser". Next, choose the "mouse" genome, the "July 2007" assembly, the "Genes and Gene Prediction Tracks" group, the "UCSC genes" track, the "KnownGene" table and then click "get output". Choose "Exons Plus" and click "send query to Galaxy". Similarly, coordinates of exons are provided in BED format.

Several Perl scripts were developed to compare the mapping position of each read to the coordinates of every exon in the corresponding chromosome. All scripts mentioned in this paper would be available upon request.

Because exons can potentially overlap or are on different strands of the same genomic region, and because the UCSC database treats the same exon in alternatively spliced transcripts as different entries, the read in our data could be located in >1 exon entry in the UCSC database. Thus, to determine the reads number per gene, when the mapping position of a reads are amid multiple exons and these exons are from the same gene, this read should be counted for our purpose; but when a read is located in exons from 2 or more genes, it should not be counted because of the ambiguousness. Therefore, we needed to have a gene with which each exon was associated. For this purpose, we downloaded the "knownIsoforms" table showing the relationship between transcripts and the UCSC genes/clusters. Several scripts were developed to parse the last column in the comparison result based on "knownIsoforms" table. If all of the exons in the last column belong to the same UCSC gene, the read was recorded as from this gene. If not, this read was not used in the following analysis. Finally the number of reads each gene had was calculated. In paired-end analysis, both ends were only counted once.

Differential Expression Analysis. We had samples from 2 developmental stages, E18 and P7. The number of reads each gene had in either stage was calculated, respectively. To access the significance of differences in read counts corresponding to a particular gene between the 2 developmental stages, we performed Fisher's exact test on reads count with "sagenhaft" library in Bioconductor. Each gene was given a P value to measure the significance of difference in reads number between the 2 stages. FDR was also calculated. Since FDR was much smaller than P value in our data, we just used P value to tell the difference.

The test was carried out in 2 different ways as reported in the paper. To balance the false positive rate and the false negative rate, we summed reads count from 2 paired-end data and then performed the test. To be more stringent, we did the test on either paired-end data, respectively. Then those differentially expressed ones supported by both datasets were extracted. We required the p value is equal to or smaller than 0.01 in both datasets and the gene has the same direction of change in reads count.

The results of reads counting and differential expression analysis were summarized in Table S1. The last column was gene annotation taken from "kgXref" table in UCSC. The "kgXref" table provided associated gene name and functional annotation to every transcript of the same gene. Because these transcripts were collected from different sources by UCSC, each one might have different annotation or associated gene name. The way we generated annotation for Table S1 is to take the annotation of every transcript and then trim to get the nonredundant set.

Spicing Variant Analysis. In our data, there were still many corresponding reads only in 1 exon for genes having many spliced transcripts, suggesting these reads were from transcript specific region. Thus, these reads provided evidence that certain transcripts were transcribed. Moreover, in paired-end analysis, if the 2 ends mapped to unique exon-exon combination, this read pair was also used as evidence for expression of the corresponding transcript.

First, we sorted out reads satisfying the following 2 standards:

1, mapped in only 1 transcript

2, from genes having alternative splicing. The "knownIsoforms" table was used to tell whether a gene have multiple transcripts.

Then several scripts were developed to summarize these reads to genes.

Finally the splicing events in E18 and P7 were compared. Genes which expressed different transcripts in 2 stages were recorded.

Analysis of Unknown Transcripts. The downloaded exon coordinates were used to define exon boundaries. In single-end reads analysis, 20 bp or full length (if the exon was 20 bp) beside the exon boundaries were extracted and concatenated to create the pseudosequence (Fig. S3). Overlapping exons were considered only if their boundaries were different and single exon genes

were excluded. All possible combinations of pairs of exons without the violation of the transcription direction were constructed. All possible junctions were distinguished based on annotation information from UCSC Known Gene. Rmap was used for mapping reads to exon junctions, allowing at most 3 mismatches, read length 27, and seed length 11 (the same process as mapping reads to the whole genome). Reads that were already mapped in the mouse genome were excluded. Then reads mapped to junctions were selected as evidence for transcripts.

In paired-end analysis, the process was similar but modified a little to fully exploit the advantage of paired ends. The pseudosequence was constructed by concatenating all possible junctions and all full-length exons so that both ends can be mapped. Mapping was performed separately for either end. Pair relations were restored using the read IDs and only unique hits were selected for the analyses. Reads that were mapped in the mouse genome were excluded. Coordinates in the pseudosequence were then transformed to coordinates in the real genome. How 2 ends mapped to junctions and exons were shown in Fig. S3. In addition to detection of transcripts by looking for reads mapped on junctions, reads in which 2 ends mapped to exon-exon combinations were also selected as evidence for unreported transcripts. The complete list of unreported splicing form variants was shown in Table S3.

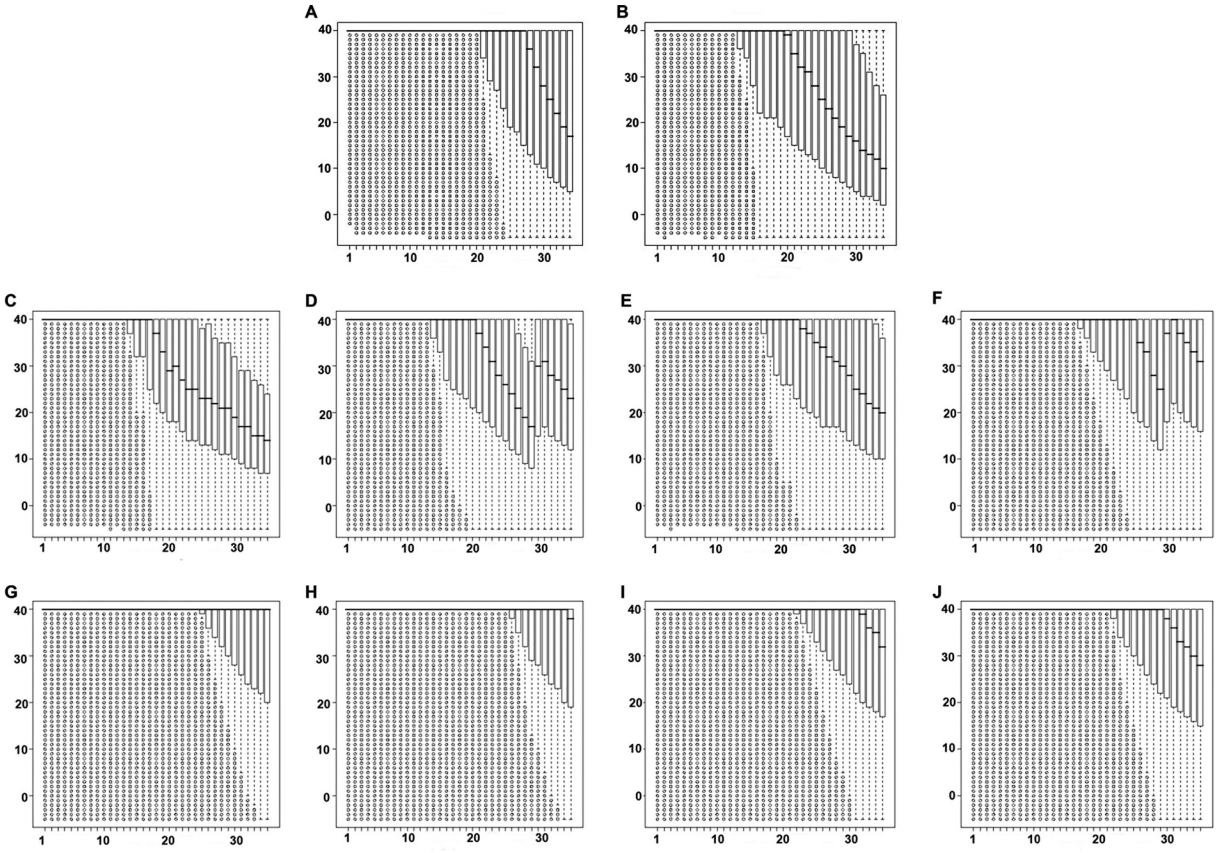

Fig. S1. The box plot of sequencing quality score at each base. The *y* axis was the quality score (Phred-equivalent metric). The *x* axis showed each base position in a read. The short bold horizontal line was the median of quality score at that base and the medians >40 of the first several bases were not shown. At each base the box plot of quality score was shown. Small circles stand for outliers. (*A*) Single-end reads from E18. (*B*) Single-end reads from P7. (*C*) The first end of the first paired-end reads from E18. (*D*) The second end of the first paired-end reads from E18. (*E*) The first end of the first paired-end reads from P7. (*F*) The second end of the first paired-end reads from P7. (*G*) The first end of the second paired-end reads from E18. (*H*) The second end of the second paired-end reads from E18. (*I*) The first end of the second paired-end reads from P7. (*J*) The second end of the second paired-end reads from P7. In all of the 10 plots, the first 27 bases have quality scores >20 (<1% error rate).

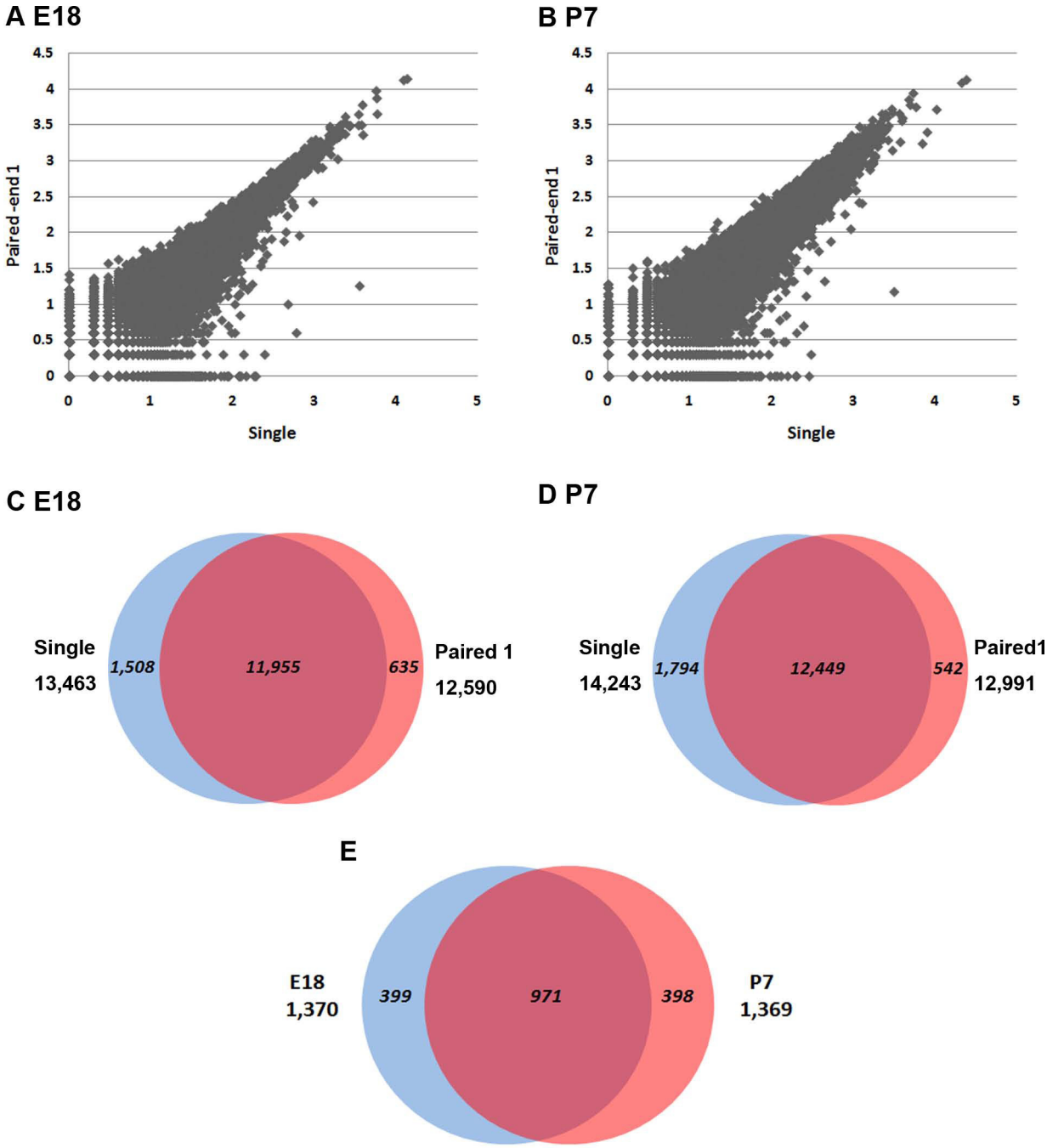

Fig. S2. Single-end analysis results and comparison with the first paired-end data (from the same cDNA sample). (*A* and *B*) The comparison of reads per gene between single-end analysis and the first paired-end analysis. Because the read number per gene ranges from 0 to >10,000, the read numbers adding 1 were transformed by $\log_{10}$. (A) E18 R = 0.96; (B) P7 R = 0.95. (*C* and *D*) The comparison of the number of expressed genes between single-end analysis and the first paired-end analysis. Genes having 2 or more reads were regarded as expressing at certain stage. (*C*) E18; (*D*) P7. (*E*) The number of highly expressed genes in single-end analysis. The top 5% genes were regarded as highly expressed ones.

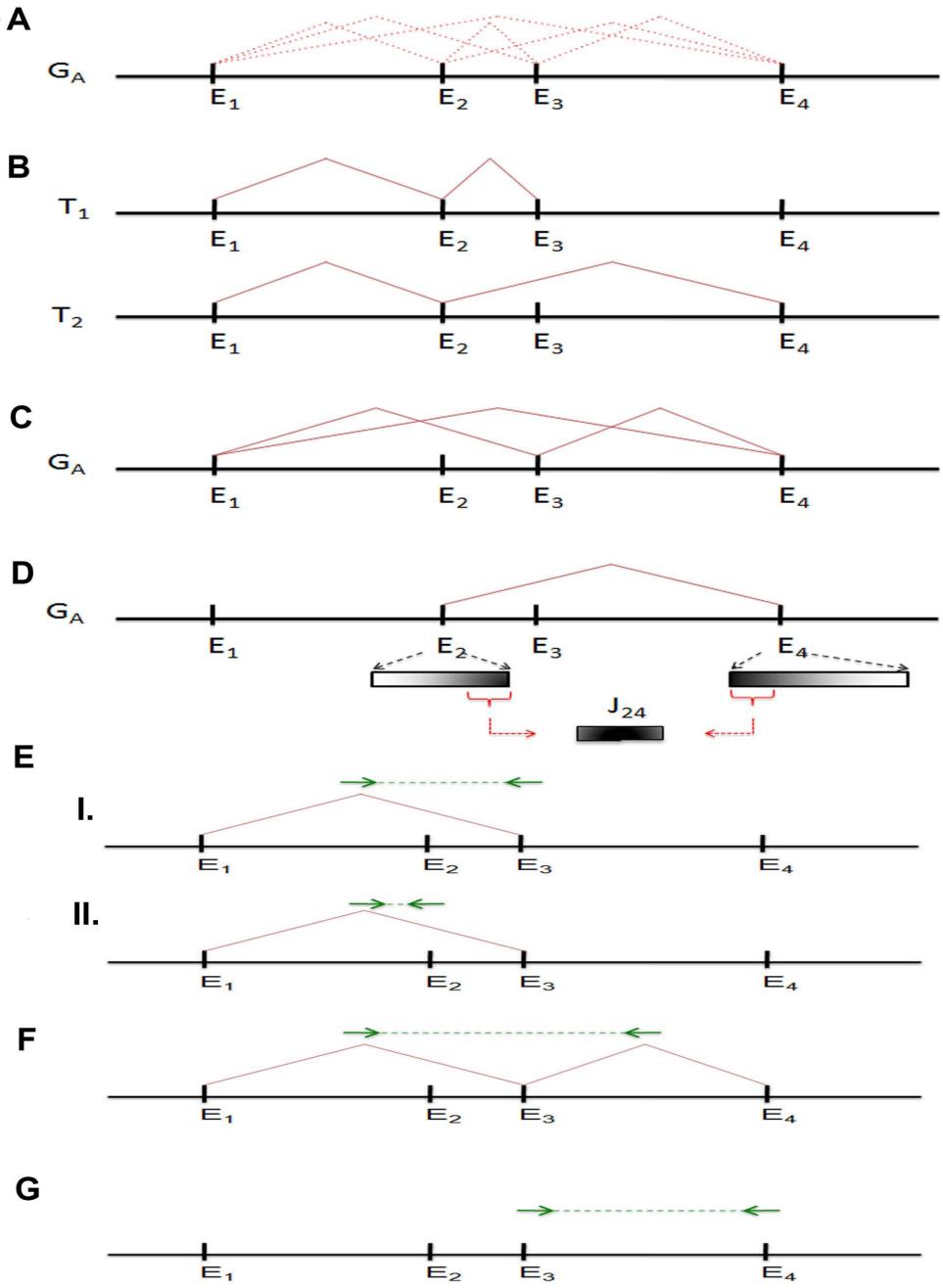

Fig. S3. A hypothetical example of the strategy to detect distinct transcripts. (*A*) Gene A had 4 exons, E1, E2, E3, and E4. (*B*) Two transcripts, T1 and T2, were alternatively splicing variants from gene A. T1 had E1, E2, and E3. T2 had E1, E2, and E4. (*C*) For gene A, junctions between E1 and E3, E1 and E4, and E3 and E4 were unknown junctions, which were not present in known transcripts. (*D*) Twenty base pairs on either side of every possible junction were taken and attached to form junction sequences computationally. For gene A, there were 6 possible junctions in total. In single-end analysis, all junction sequences from UCSC Known Gene database were concatenated into 1 long pseudosequence (because Rmap only tolerated 1 sequence as target) and reads were mapped to the concatenated sequence with Rmap. Reads mapping to unknown junctions were selected and considered as evidence for alternative transcripts. In paired-end analysis, all junction sequences and exon sequences were concatenated together to accommodate both ends. One or both ends mapping to unknown junctions were selected and considered as evidence for alternative transcripts. In addition, paired-end reads mapped to distinct exon-exon combinations were selected and regarded as evidence for unknown transcripts as well. (*E*–*G*) Detailed illustration of paired-end reads. Paired-end reads not only can detect alternative transcripts by junctions (*E* and *F*), but also through exon-exon combinations (*G*). (*E*) One end mapped on a junction; the other end mapped in an exon. (i) Valid mapping: 1 end of a paired-end read mapped on a junction (J13); the other end mapped in E3. This paired-end read indicates a splicing form of E1 and E3. (ii) Invalid mapping: 1 end of a paired-end read mapped at J13; the other end mapped at E2. It is not reasonable to have a splicing form of this kind. (*F*) Both ends were mapped on junctions. One end of a paired-end read mapped on a junction (J13); the other end also mapped at a junction (J34). This indicated a distinct splicing form of E1, E3, and E4. (*G*) Both ends were mapped in exons. One end mapped to E3 and the other end mapped to E4. An alternative transcript with E3 and E4 was indicated.

Other Supporting Information Files

Table S1 (PDF)
Table S2 (PDF)
Table S3 (PDF)
Table S4 (PDF)
Table S5 (PDF)