Mapping and quantifying mammalian transcriptomes by RNA-Seq

Ali Mortazavi[1,2], Brian A Williams[1,2], Kenneth McCue[1], Lorian Schaeffer[1] & Barbara Wold[1]

We have mapped and quantified mouse transcriptomes by deeply sequencing them and recording how frequently each gene is represented in the sequence sample (RNA-Seq). This provides a digital measure of the presence and prevalence of transcripts from known and previously unknown genes. We report reference measurements composed of 41–52 million mapped 25-base-pair reads for poly(A)-selected RNA from adult mouse brain, liver and skeletal muscle tissues. We used RNA standards to quantify transcript prevalence and to test the linear range of transcript detection, which spanned five orders of magnitude. Although >90% of uniquely mapped reads fell within known exons, the remaining data suggest new and revised gene models, including changed or additional promoters, exons and 3′ untranscribed regions, as well as new candidate microRNA precursors. RNA splice events, which are not readily measured by standard gene expression microarray or serial analysis of gene expression methods, were detected directly by mapping splice-crossing sequence reads. We observed 1.45×10^5 distinct splices, and alternative splices were prominent, with 3,500 different genes expressing one or more alternate internal splices.

The mRNA population specifies a cell's identity and helps to govern its present and future activities. This has made transcriptome analysis a general phenotyping method, with expression microarrays of many kinds in routine use. Here we explore the possibility that transcriptome analysis, transcript discovery and transcript refinement can be done effectively in large and complex mammalian genomes by ultra-high-throughput sequencing.

Expression microarrays are currently the most widely used methodology for transcriptome analysis, although some limitations persist. These include hybridization and cross-hybridization artifacts[1–3], dye-based detection issues and design constraints that preclude or seriously limit the detection of RNA splice patterns and previously unmapped genes. These issues have made it difficult for standard array designs to provide full sequence comprehensiveness (coverage of all possible genes, including unknown ones, in large genomes) or transcriptome comprehensiveness (reliable detection of all RNAs of all prevalence classes, including the least abundant ones that are physiologically relevant). Other approaches to large-scale RNA analysis are serial analysis of gene expression (SAGE)[4,5] and related methods such as massively parallel signature sequencing (MPSS)[6], which use DNA sequencing of previously cloned tags 17–25 base pairs (bp) from terminal 3′ (or 5′) sequence tags. These sequence tags are then identified by informatic mapping to mRNA reference databases or, for longer tag lengths, to the source genome. A strength of SAGE and SAGE-like methods is that they produce digital counts of transcript abundance, in contrast to the analog-style signals obtained from fluorescent dye–based microarrays. However, SAGE-family assays provide no information about splice isoforms or new gene discovery, and fully comprehensive measurements of lower-abundance-class RNAs have not been achieved owing to cost and technology constraints. Expressed sequence tag (EST) sequencing of cloned cDNAs has long been the core method for reference transcript discovery[7–9]. It has both qualitative and quantitative limitations, imposed partly by historic sequencing capacity and cost issues, and more crucially by bacterial cloning constraints that affect which sequences are represented and how sequence-complete each clone is. Recently, dense whole-genome tiling microarrays have been developed and applied to transcriptomes for measuring expression and for transcript discovery[10–14]. In contrast to expression arrays, these tiling arrays can discover new genes and exons, but they require large amounts of input RNA and have other limitations that affect sensitivity, specificity and direct splice detection.

A simpler and potentially more comprehensive way to measure transcriptome composition and to discover new exons or genes is by direct ultra-high-throughput sequencing of cDNA (**Fig. 1a**). This RNA-Seq approach avoids the need for bacterial cloning of the cDNA input. The resulting sequence reads are individually mapped to the source genome and counted to obtain the number and density of reads corresponding to RNA from each known exon, splice event or new candidate gene. The presence and amount of each RNA can be calculated and subsequently compared with the amount in any other sequenced sample, now or in the future. If enough reads (>40 million) are collected from a sample, it should in theory be possible to detect and quantify RNAs from all biologically relevant abundance classes and to map RNA splice choices for transcripts of high and moderate abundance.

[1]Division of Biology, MC 156-29, California Institute of Technology, Pasadena, California 91125, USA. [2]These authors contributed equally to this work. Correspondence should be addressed to B.W. (woldb@caltech.edu).

RECEIVED 2 MAY; ACCEPTED 27 MAY; PUBLISHED ONLINE 30 MAY 2008; DOI:10.1038/NMETH.1226

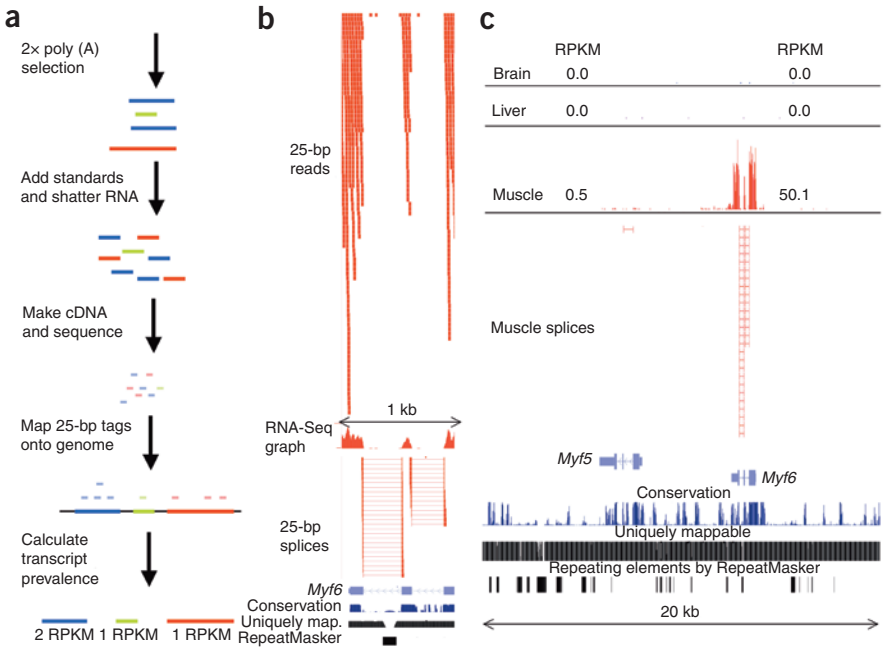

Figure 1 | Outline of RNA-Seq procedure. (**a**) After two rounds of poly(A) selection, RNA is fragmented to an average length of 200 nt by magnesium-catalyzed hydrolysis and then converted into cDNA by random priming. The cDNA is then converted into a molecular library for Illumina/Solexa 1G sequencing, and the resulting 25-bp reads are mapped onto the genome. Normalized transcript prevalence is calculated with an algorithm from the ERANGE package. (**b**) Primary data from mouse muscle RNAs that map uniquely in the genome to a 1-kb region of the *Myf6* locus, including reads that span introns. The RNA-Seq graph above the gene model summarizes the quantity of reads, so that each point represents the number of reads covering each nucleotide, per million mapped reads (normalized scale of 0–5.5 reads). (**c**) Detection and quantification of differential expression. Mouse poly(A)-selected RNAs from brain, liver and skeletal muscle for a 20-kb region of chromosome 10 containing *Myf6* and its paralog *Myf5*, which are muscle specific. In muscle, *Myf6* is highly expressed in mature muscle, whereas *Myf5* is expressed at very low levels from a small number of cells. The specificity of RNA-Seq is high: *Myf6* expression is known to be highly muscle specific, and only 4 reads out of 71 million total liver and brain mapped reads were assigned to the *Myf6* gene model.

Some challenges to performing RNA-Seq are expected to affect transcriptomes from all organisms similarly, such as how uniformly sequences from an entire transcript will be represented, how much sequence is required to reliably detect and measure the concentration of RNAs of lower abundance classes, how the data will be quantified and how relative quantification will be converted to absolute RNA concentrations. In addition, transcriptomes of organisms with large genomes, containing genes with more complicated structure, present some special challenges. Yeast and *Arabidopsis thaliana* transcriptomes have been profiled by RNA-Seq approaches concurrently with this study[15–17], but the mouse and human genomes are much larger and more complex than these. This increases the computational resources needed to simply map reads onto the genome, and it also requires the mapping of splice-crossing reads that span very large introns. These demands mean that it is not possible to use some tools that might be tenable in a small genome with few introns, such as BLAST[18]. Large genomes are also typically rich in families of paralogous genes, which presents the challenge of mapping reads that could map equally well to multiple sites in the genome. Here we have begun to address issues pertaining to the acquisition, standardization and analysis of large and complex transcriptomes by RNA-Seq. As a test mammalian case, we performed RNA-Seq, using Illumina/Solexa sequencing technology, on poly(A)-selected RNA from adult C57BL mouse brain, liver and skeletal muscle tissues.

RESULTS

RNA-Seq sample preparation and sequencing

Uniformity of sequence coverage across transcripts will affect sensitivity of detection, accuracy of quantification and completeness of splice and exon maps. In preliminary experiments, we found that controlled hydrolysis of RNA samples before cDNA synthesis steps significantly improved the uniformity of sequence coverage across transcripts, although coverage uniformity did not achieve the theoretical limit (**Supplementary Fig. 1a**,**b** online). The rationale for using hydrolysis of RNA before random priming rather than fragmentation of cDNA at the next step was twofold. First, cDNA priming at putatively random sites, if fully successful, will over-represent 5′ ends of transcripts, and this uneven representation will have differing impact on RNAs of different sizes. Second, preliminary data strongly suggested that there are some strongly favored and disfavored sites of random priming, and we observed this in samples that were primed without hydrolysis. It did not, however, correspond to simple GC content bias (**Supplementary Fig. 1b**). We reasoned that, at room temperature, some RNA secondary structure may shield parts of transcripts from priming while favoring other sites. By fragmenting the RNA, we expected to reduce the amount of such secondary structure, though not completely eliminate it. RNA fragmentation before copying would also be expected to greatly reduce 5′ bias. This protocol gave better overall uniformity than protocols without RNA fragmentation (**Supplementary Fig. 1**), although some residual and reproducible nonuniformity clearly persists for randomly primed substrates that was not observed in other kinds of Illumina sequencing substrates handled simultaneously, such as chromatin immunoprecipitation sequencing (ChIPSeq) samples (for example, **Supplementary Fig. 1c**).

For each transcriptome measurement (mouse liver, skeletal muscle and total brain), we made randomly primed cDNA from 100 ng poly(A)$^+$ RNA hydrolyzed to 200–300 nucleotides (nt), and constructed a Solexa molecular library (**Fig. 1**). We obtained 10–30 million 25-bp reads mapping to unique sites in the mouse genome from each library, with two independent libraries assayed for each source tissue. Additional reads were later mapped to RNA splices and to some regions not included in the 'uniquely mappable' genome. All primary sequence read data for both replicates of the three tissue RNAs have been submitted to the National Center for Biotechnology Information (NCBI) short-read archive (accession number SRA001030). RNA-Seq Summary data are in **Supplementary Figure 2** and **Supplementary Table 1** online.

High read number is relevant for RNA-Seq because our ability to reliably detect and measure rare, yet physiologically relevant, RNA species (those with abundances of 1–10 RNAs per cell) depends on the number of independent pieces of evidence (sequence reads) obtained for transcripts from each gene. This constraint influenced our sequencing strategy, choice of instrument and choice of the 25-bp read length.

The sensitivity of RNA-Seq will be a function of both molar concentration and transcript length. We therefore quantified transcript levels in reads per kilobase of exon model per million mapped reads (RPKM) (**Fig. 1a,c**). The RPKM measure of read density reflects the molar concentration of a transcript in the starting sample by normalizing for RNA length and for the total read number in the measurement. This facilitates transparent comparison of transcript levels both within and between samples.

Examination of a well-characterized locus

Data from a 21-million-read transcriptome measurement of adult mouse skeletal muscle (**Fig. 1b,c**) illustrate some key characteristics of our results. *Myf6* (also known as *Mrf4*) is a much-studied myogenic transcription factor gene that is expressed specifically and modestly in muscle, as expected, but silent in liver and brain. Evidence for *Myf6* expression in skeletal muscle (**Fig. 1b**) consisted of 1,295 sequence reads 25 bp in length that map uniquely to *Myf6* exons, and 30 reads that cross splice junctions; another four reads fell within the introns. Brain and liver measurements of similar total read number had 1 and 0 reads on *Myf6* exons, illustrating favorable signal-to-noise characteristics, absolute signal and specificity (**Fig. 1c**).

RNA-Seq global data properties

Technical replicate determinations of transcript abundance were reproducible ($R^2 = 0.96$, **Fig. 2a**). Summing the replicates over an entire transcriptome (**Fig. 2b**, liver; **Supplementary Table 2** online) showed that the vast majority of reads (93%) mapped to known and predicted exons, even though the exons comprise <2% of the entire genome; 4% of reads were within introns; and only 3% fell in the large intergenic territory. We expected to observe some intronic reads in total poly(A)$^+$ RNA because such preparations are known to include partially processed nuclear RNAs and because some genes might have internal exons that have not yet been added to the gene models. The 3% intergenic fraction places a rough upper bound on possible noise reads.

To assess the dynamic range of RNA-Seq and to test for possible effects of starting transcript length on the observed transcript abundance, we introduced into each experimental sample a set of known RNA standards transcribed *in vitro* from *Arabidopsis* and phage lambda templates (**Fig. 2c**). These standards comprised long (~10,000 nt), intermediate (~1,500 nt) and short (~300 nt) transcripts, and they were designed to span the range of abundance (~0.5–50,000 transcripts per cell) typically observed in natural transcriptomes. RNA-Seq data for the standards were linear across a dynamic range of five orders of magnitude in RNA concentration. Sequence coverage over test transcripts was highly reproducible and quite uniform (**Supplementary Fig. 1c**). At current practical sequencing capacity and cost (~40 M mapped reads), transcript detection was robust at 1.0 RPKM and above for a typical 2-kilobase (kb) mRNA (~80 individual sequence reads resulting in a P value $<10^{-16}$). Beyond simple detection confidence, we analyzed the impact of different amounts of sequencing on our ability to measure the concentration of a given transcript class (defined on the basis of RPKM) within ±5% (**Fig. 2d**). When these RNA standards are used in conjunction with information on cellular RNA content, absolute transcript levels per cell can also be calculated. For example, on the basis of literature values for the mRNA content of a liver cell[19] and the RNA standards, we estimated that 3 RPKM corresponds to about one transcript per liver cell. For C2C12 tissue culture cells, for

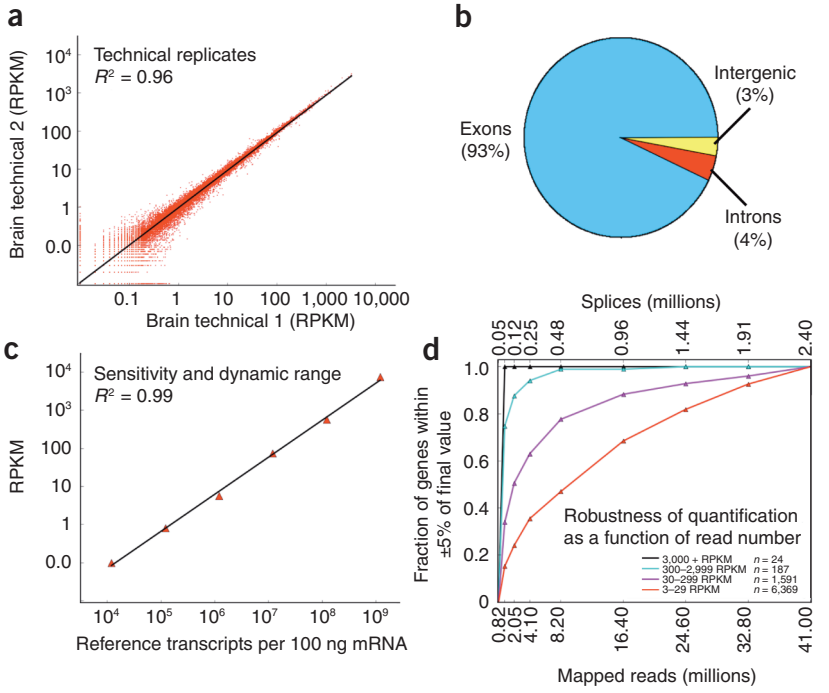

Figure 2 | Reproducibility, linearity and sensitivity. (**a**) Comparison of two brain technical replicate RNA-Seq determinations for all mouse gene models (from the UCSC genome database), measured in reads per kilobase of exon per million mapped sequence reads (RPKM), which is a normalized measure of exonic read density; $R^2 = 0.96$. (**b**) Distribution of uniquely mappable reads onto gene parts in the liver sample. Although 93% of the reads fall onto exons or the RNAFAR-enriched regions (see **Fig. 3** and text), another 4% of the reads falls onto introns and 3% in intergenic regions. (**c**) Six *in vitro*–synthesized reference transcripts of lengths 0.3–10 kb were added to the liver RNA sample (1.2×10^4 to 1.2×10^9 transcripts per sample; $R^2 > 0.99$). (**d**) Robustness of RPKM measurement as a function of RPKM expression level and depth of sequencing. Subsets of the entire liver dataset (with 41 million mapped unique + splice + multireads) were used to calculate the expression level of genes in four different expression classes to their final expression level. Although the measured expression level of the 211 most highly expressed genes (black and cyan) was effectively unchanged after 8 million mappable reads, the measured expression levels of the other two classes (purple and red) converged more slowly. The fraction of genes for which the measured expression level was within ±5% of the final value is reported. 3 RPKM corresponds to approximately one transcript per cell in liver. The corresponding number of spliced reads in each subset is shown on the top x axis.

which we know the starting cell number and RNA preparation yields needed to make the calculation, a transcript of 1 RPKM corresponds to approximately one transcript per cell.

Analysis strategy and software

To analyze these data, we developed Enhanced Read Analysis of Gene Expression (ERANGE), which is outlined in **Figure 3a** and is available as **Supplementary Software** online and at http://woldlab.caltech.edu/RNA-Seq. The functions of ERANGE are to (i) assign reads that map uniquely in the genome to their site of origin and, for reads that match equally well to several sites ('multireads'), assign them to their most likely site(s) of origin; (ii) detect splice-crossing reads and assign them to their gene of origin; (iii) organize reads that cluster together, but do not map to an already known exon, into candidate exons or parts of exons; and (iv) calculate the prevalence of transcripts from each known or newly proposed RNA, based on normalized counts of unique reads, spliced reads and multireads. The new candidate RNA regions produced can be thought of as ESTs, and, like ESTs, some are provisionally appended to existing gene models if they meet several additional criteria. Remaining unassigned candidate transcribed regions (labeled RNAFAR features) can then be used in conjunction with other confirming data to develop new or revised gene models. Final RPKM values for each dataset, together with intermediate values calculated at earlier steps in ERANGE, are in **Supplementary Datasets 1, 2,** and **3** online.

Although RNA-Seq is not affected by background from cross-hybridization, as microarrays are, it is not free of ambiguities

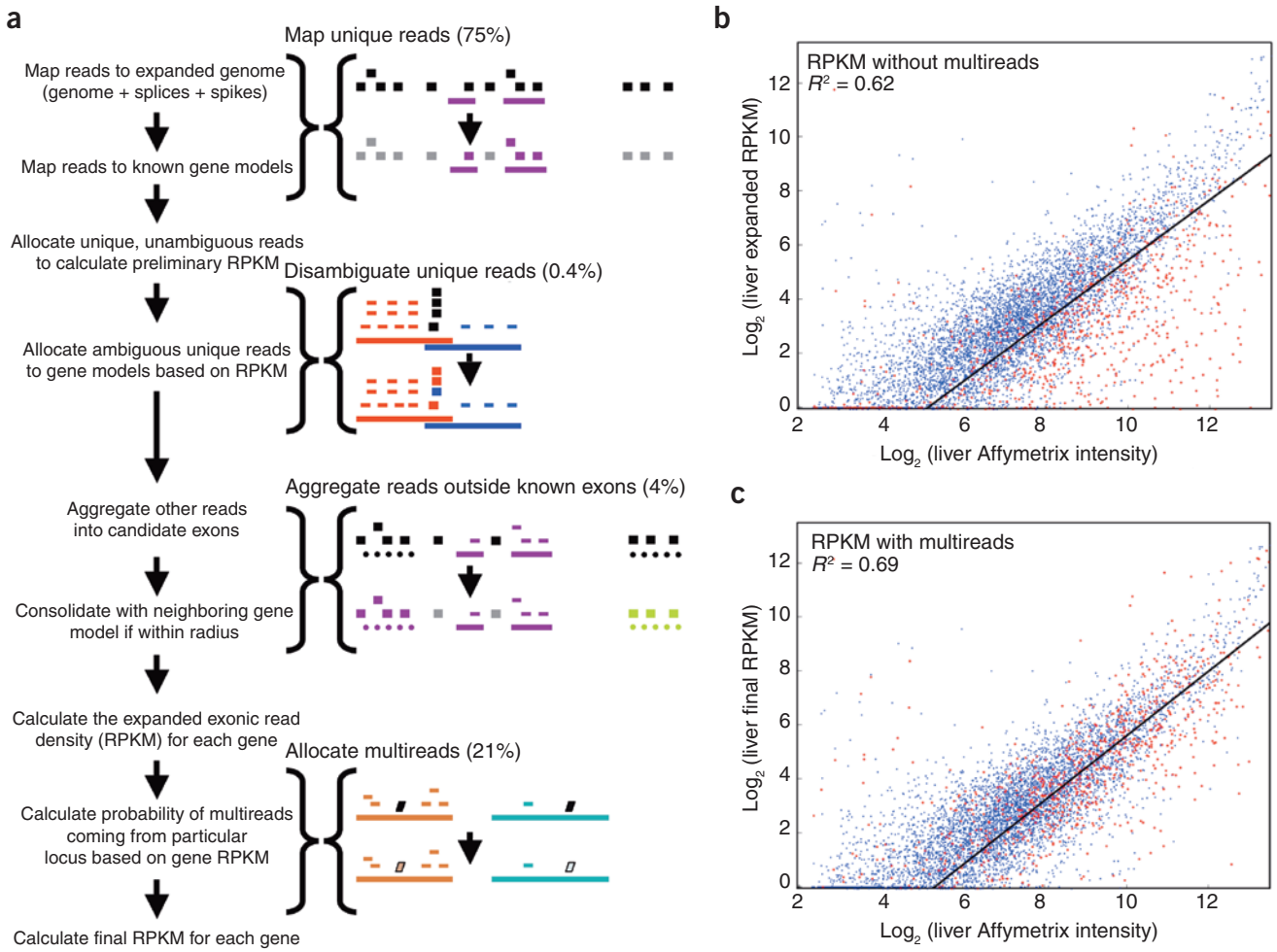

Figure 3 | Enhanced Read Analysis of Gene Expression (ERANGE) and the allocation of multireads. (**a**) The main steps in the computational pipeline are outlined at left, with different aspects of read assignment and weighting diagrammed at right and the corresponding number of gene model reads treated in muscle shown in parentheses. In each step, the sequence read or reads being assigned by the algorithm are shown as a black rectangle, and their assignment to one or more gene models is indicated in color. Sequence reads falling outside known or predicted regions are shown in gray. RNAFAR regions (clusters of reads that do not belong to any gene model in our reference set) are shown as dotted lines. They can either be assigned to neighboring gene models, if they are within a specified threshold radius (purple), or assigned their own predicted transcript model (green). Multireads (shown as parallelograms) are assigned fractionally to their different possible locations based on the expression levels of their respective gene models as described in the text. (**b**) Comparison of mouse liver expanded RPKM values to publicly available Affymetrix microarray intensities from GEO (GSE6850) for genes called as present by Rosetta Resolver. Expanded RPKMs include unique reads, spliced reads and RNAFAR candidate exon aggregation, but not multireads. Genes with >30% contribution of multireads to their final RPKM (**Supplementary Fig. 4**) are marked in red. (**c**) Comparison of Affymetrix intensity values with final RPKMs, which includes multireads. Note that the multiread-affected genes that are below the regression line in **b** straddle the regression line in **c**.

caused by gene sequences that are closely related to each other. We surveyed the mouse genome informatically and found that 76% of it is in 25-bp sequence segments that are unique, whereas 6% is composed of 25-mers that occur 2–10 times in the genome (multireads). The remainder is composed of 25-mer reads that occur >10 times (**Supplementary Fig. 3a**,**b** online), and these are excluded from further analysis in this work. Of the 25-bp mappable reads from each mouse RNA-Seq sample, 13–24% fell into the multiread class, matching equally well at 2–10 different locations on the mouse genome rather than mapping best to a single site (**Supplementary Fig. 3**).

Most of the multireads in our datasets are attributable to known duplicated genes and segmental duplications. *Myf6* is a well-studied example of gene duplication, lying immediately adjacent to its paralog, *Myf5*, on mouse chromosome 10 (**Fig. 1c**). For *Myf6*, only seven additional reads (0.5%) were from the multiread class, and these mapped to the conserved basic helix-loop-helix (bHLH) DNA-binding domain that defines genes as members of this paralogous family. In contrast, members of other multigene families, such as the ubiquitin B family (**Supplementary Fig. 3b**), are dominated by multireads (42,642 = 97%) as compared with uniquely mappable reads (1,135). This is expected for paralogs that are very similar to each other and for internally repeated domains within some genes. If all multireads are simply discarded, as default settings in current Solexa software do, the end result will be to undercount greatly or even entirely fail to report expression for genes that have closely related paralogs, such as those of the ubiquitin family. To avoid this, ERANGE distributes multireads in proportion to the number of unique and splice reads recorded at similar loci. The inclusion and proportionate distribution of multireads will naturally have variable impact on RNA quantification, with smaller effects on paralogs that are more divergent and larger effects on those that are more similar to each other (**Supplementary Fig. 4** online). In some rare instances for which there are no unique reads across an entire gene model, usually reflecting very recent gene duplication, ERANGE will distribute multireads evenly among candidate paralogs in the genome (**Supplementary Dataset 4** online). The impact of identifying and allocating multireads in this manner, summed over the three transcriptomes, was to change RNA quantification for 28% of genes scoring above a 5-RPKM threshold in the muscle transcriptome by more than 30%.

The overall impact of allocating multireads in this manner was assessed by following where the affected transcript RPKM values fell before and after multiread allocation as a function of the correlation between our liver RNA-Seq data and an independent, publicly available Affymetrix mouse liver transcriptome measurement (**Fig. 3b,c**; multiread-affected genes shown in red). Multiread-affected transcript measurements generally moved from being systematically under-represented relative to their value in the array data (**Fig. 3b**) to a position straddling the mean after multiread allocation (**Fig. 3c**). This suggests that the overall impact of our computational allocation of multireads is beneficial, as it improves the correlation with the microarray results. The overall picture of the transcriptome obtained by the two methods is similar ($R^2 = 0.69$). However, unlike the rest of the distribution, the bottom quartile of the Affymetrix 'present' calls showed no correlation with the RNA-Seq data ($R^2 = 0.03$), suggesting that many of the putatively 'expressed' RNAs identified by the microarray analysis might be false positives.

Widespread alternate splice isoforms

Splice-crossing reads, such as are shown for *Myf6* (**Fig. 1b**), were identified by mapping otherwise unassigned sequence reads to a library of all known splice events in all University of California Santa Cruz genome database (UCSC) Mouse July 2007 (mm9) gene model splices. When we summed over the entire dataset, including all otherwise unmappable reads, splice-spanning reads comprised ~3% (**Supplementary Table 1**), which is consistent with splice frequency in gene models across the genome. To assess the efficiency of splice detection, we computationally predicted all reads expected to cross known splices in a transcriptome, by considering all UCSC gene models and their respective levels of expression based on exon reads. The observed instances of splice-crossing reads were in good agreement with predictions (**Fig. 4c**). Based on sequence coverage of control transcripts, we further calculated that a splice will be detected with 95% confidence if a transcript is represented at >11 RPKM in a 40-million-read transcriptome.

We next assessed the extent of alternative splice usage. Alternate splices were extensive. We observed more than one alternate splice form for 3,462 genes in all three tissues (**Supplementary Table 3** online). The vast majority (>93%) were multiple splice forms detected within at least one tissue (**Fig. 4d**), rather than being distinct splices restricted to one tissue as compared to another. However, as illustrated below for *Mef2d* transcripts, one splice form can be strongly quantitatively preferred over the other in each tissue, even though both are detectable. This initial splicing analysis provides minimum measures of the extent of alternative splicing, because splice detection is a function of transcript prevalence and because we only looked for known alternative splice events. We conclude that cDNA sequencing with 30–40 million read measurements readily detects major splice isoforms for abundant and moderately abundant transcripts, whereas splice detection for the lowest-abundance RNA classes and isoforms is sporadic.

Transcript annotation and novel transcript discovery

In addition to known alternate splice forms, other sequence features that are not in the NCBI and RefSeq annotations can also be found in the data. Summing over all three transcriptomes, we identified 16,923 regions that are not in the NCBI Gene annotation and have less than 10% overlap with repeat-masked features. The RNAFAR algorithm (part of ERANGE, **Fig. 3**) clustered reads that were not associated with gene models and assigned 92% of these candidate exons to neighboring gene models when they were within a specifiable distance of the model (here we used a permissive 20-kb distance parameter to help identify candidate additions to untranscribed regions (UTRs), including new external exons, and to discriminate these from RNAFAR features that are the best candidates to be previously unknown genes). An illustrative example is *Mef2d* (**Fig. 4**), which has a prominent muscle-preferred protein coding exon (**Fig. 4b**) and also has a much longer 3′ UTR than the one described in RefSeq (**Fig. 4a** and **Supplementary Fig. 5** online). We have tested some regions predicted by RNAFAR using RT-PCR, including the extended *Mef2d* 3′ UTR (**Supplementary Fig. 5**), and have noted partial or complete support for others in GenBank.

Combining all three transcriptomes, we consolidated the remaining 1,320 unaffiliated RNAFAR regions into 596 candidate transcript

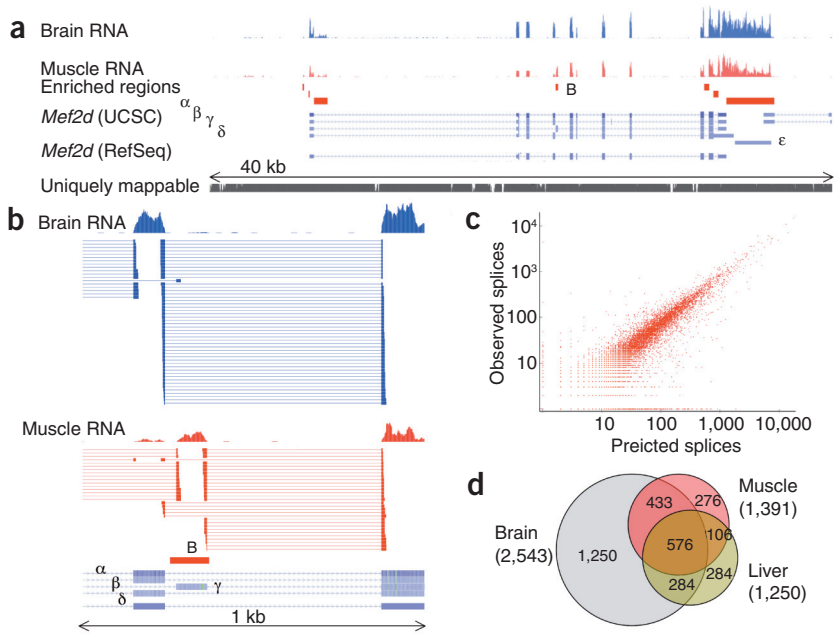

Figure 4 | Candidate new and revised exons identified by the RNAFAR algorithm. (**a**) A 40-kb region encompassing the *Mef2d* gene, which is expressed in adult muscle (28 RPKM in muscle and 45 RPKM in brain), and a neighboring gene that is expressed at a much lower level in brain. RefSeq has only a single annotation for *Mef2d*, but UCSC has five (labeled α–ε). The α form corresponds to the RefSeq model, and γ is a muscle-specific isoform[23]. The RNAFAR algorithm identified seven regions (red) enriched with reads that fell outside the NCBI gene annotations and were assigned by the algorithm to the *Mef2d* locus. (**b**) A 1.5-kb close-up of muscle-specific alternative splicing at the RNAFAR region labeled 'B' in panel **a**. The prevalence of splicing switches from the canonical exon in the brain sample to the RNAFAR exon in the muscle sample, as seen both in the ratio of spliced reads and in the number of reads falling on the two diagnostic exons. (**c**) The number of expected spliced reads for each gene model was predicted computationally, based on the number of introns and the exonic read density. The predicted number is then plotted against the number of splices observed ($R^2 = 0.90$). (**d**) The tissue distribution of genes with two splice isoforms in the same tissue.

models. Some are expressed in all three tissues, whereas others are strongly tissue specific, such as the new candidate precursor of the neuronal microRNA mir-124-1 (**Fig. 5** and **Supplementary Fig. 6** online). We did not attempt to assess whether other new RNAFAR models are likely to code for proteins.

DISCUSSION

This dataset of 140 million mapped sequence reads provides rich starting information for improving gene models across the mouse genome, although data from additional tissues and cell types, and from a variation on the current RNA-Seq methodology, will all be needed to drive a comprehensive reannotation. However, with the basic RNA-Seq used here and data from only three tissues, we identified ~17,000 features (RNAFAR clusters of reads) that are candidate new parts of existing genes (these are the majority of features that are not in the gene models), plus 596 new candidate transcripts. High-density tiling arrays have also been used to discover previously unknown RNAs[10–12], but our data are not directly comparable to the data from those studies because of differences in the species, tissue and type of RNA sample studied. Our data included evidence for (i) ~3,000 extended or newly identified 3′ UTRs, which are relevant because of their possible roles in microRNA-mediated control of translation and post-transcriptional RNA metabolism, and (ii) ~3,000 new or extended 5′ exons that imply different positions for promoters and alternate promoters, which should be useful for interpreting data on regulatory factor interactions with each gene. The precise number of candidate features identified as possible 3′ or 5′ extensions or new internal exons depends on investigator choices in regard to analysis parameters (see **Supplementary Table 4** online). At one extreme, one can exclude any new features adjacent to current exon models (by using a radius parameter of zero for RNAFAR in ERANGE) and regard all extensions as new features. Alternatively, one can set the parameter more permissively, which is useful when trying to find candidate unknown genes.

For relatively abundant candidate transcripts (we estimate >30 RPKM), detailed validation experiments can be readily designed. However, sequence coverage from experiments of this size is not sufficient for definitive mapping of moderate- and lower-abundance newly identified exons and genes, unless the experiment is adjusted to include prevalence normalization of input RNAs and use of longer-read sequencing, such as that provided by the current 454 sequencing platform.

Although it is possible to construct custom microarrays to detect splice events directly[20–22], RNA splicing has not generally been accessible for routine microarray methods, and it is not accessible by SAGE. Here, the sheer number of reads produced made it possible to identify splice events very effectively for high- and moderate-abundance RNAs (>15–25 RPKM). In only three tissues, we found evidence for 1.45×10^5 different splices, from a library of ~2×10^5 possible splices. As expected, splices were detected sporadically in rarer transcripts. It would clearly be desirable to build a fully comprehensive map of splice isoforms, and this should be possible through extension of the current RNA-Seq in two ways. First, detecting all splices for RNAs of all prevalence classes, and for rare alternative splice isoforms from any prevalence class, calls for prevalence-normalized input RNA or cDNA. This approach would distribute the sequence sampling power more evenly across all transcript species, and implementing it should require no new technology. Second, long-range contiguity of splice choices cannot be extracted from our data, and for this reason we explicitly did not attempt to quantify splice forms. We therefore reported all RPKM prevalence information on a per-locus basis.

We anticipate that use of rapidly improving 'paired-end' variations of ultra-high-throughput sequencing will soon provide additional information needed to assemble full splice isoforms ('paired' sequences are determined for both ends of single segment of DNA, and starting DNA can be of a known length class) and sequence newly identified transcripts in a genome-wide fashion. Coupled with appropriate bioinformatics tools, this should provide a way to map long-range splice contiguity, which is not possible with the present

method. Combining prevalence normalization of input RNA with paired-end sequencing would presumably give the most complete splicing pattern map, and this analysis could also benefit from considerably longer read-length data, such as those from the 454 sequencing platform. Another unknown in our data is RNA strandedness. This has been successfully addressed in transcriptome studies of *Arabidopsis*[17]. The RNA-Seq sample preparation developed in those studies has the particular advantage of reporting strand specificity.

Our data were very reproducible and sensitive, and quantification was reliable over a broad range of RNA concentrations. However, limitations of current costs and sequencing capacity mean that low-prevalence RNAs from minority cells of naturally complex mixed tissues such as the brain will not be accessible under the current protocols. As has been true for microarrays, one path for improvement will be to push sample input requirements down, aiming for the single-cell level. For arrays, that path involves many rounds of amplification, and it is expected that similar approaches will be adapted for RNA-Seq. However, the actual number of molecules that can be sequenced on a single flow cell is similar to the number of 200-nt RNA pieces derived from a single animal cell. This means that it is theoretically possible to sequence the contents of a single cell with minimal prior amplification, though the technical challenges to implementation are considerable.

The strength of evidence for detecting any given rare transcript with RNA-Seq, especially if it has garnered multiple unique sequence reads, may be considerably greater than that provided by microarrays, because array fluorescence signals from a low-abundance true positive can be very difficult to distinguish, numerically and statistically, from background array signals due to cross-hybridization and dye binding. A 40-million-read transcriptome measurement provides reliable measurement of a single transcript per cell (between 1 and 3 RPKM for C2C12 or liver cells, as discussed in the Results). At 40 million reads, 1× sequence coverage of the transcriptome has been achieved (40 reads per kilobase of RNA).

This contrasts with an average read density of 0.03 RPKM in the sum of all regions that fall outside of exons and RNAFAR clusters, and constitutes an upper bound on what could be general 'background'. Although it is likely that this desirable specificity prevails for a majority of RNAs, a short sequence read that contains one or more errors (wrong base calls) can coincidentally match—in its mutated form—to another existing sequence in the genome. This has the potential to create a specific kind of false-positive background in RNA-Seq that is of interest because it will preferentially affect gene families. It would thus cause greatest mischief if one family member were very highly expressed (therefore generating occasional mutant reads) and the other gene to which it might map were not, in reality, expressed at all. Candidate instances of such miscalls can be culled for further evaluation if desired, but the sequencing errors themselves are being reduced by improvements in sequencing machines, use of longer high-quality sequence reads (30–40 bp) and improved base-calling algorithms. Increased read length and sequence quality will also improve splice mapping and elevate (though modestly) the fraction of the genome that will fall into the unique-read class. Finally, it will be important to develop more sophisticated probabilistic models for each transcriptome to further improve the certainty with which rare RNAs are called correctly and RNAs from related genes are quantified.

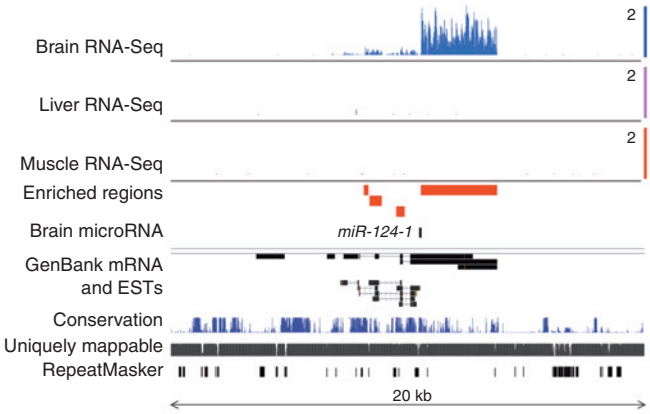

Figure 5 | Candidate microRNA precursor. A 20-kb region of chromosome 14 surrounding *mir-124-1*, which encodes a microRNA that is expressed in neurons and is embedded in a transcript that is not in the NCBI or RefSeq gene models. Our algorithm flagged the enriched regions that do not overlap repeat-masked regions as candidate exons into a candidate transcription unit, which correspond well to the mRNA and EST evidence available from GenBank.

METHODS

RNA preparation. Total RNA from pooled, adult C57BL6 mouse tissues was obtained from Stratagene (MVP Total RNA). Oligo(dT) selection was performed twice by using Dynal magnetic beads (Invitrogen) according to the manufacturer's protocol. After selection, a single 100-ng aliquot of mRNA was reserved for evaluation on the BioAnalyzer.

cDNA preparation. Total RNA (75 μg) was subjected to two rounds of hybridization to oligo(dT) beads (Dynal). 100 ng of the resulting mRNA was then used as template for cDNA synthesis. The mRNA was first fragmented by addition of 5× fragmentation buffer (200 mM Tris acetate, pH 8.2, 500 mM potassium acetate and 150 mM magnesium acetate) and heating at 94 °C for 2 min 30 s in a thermocycler and was then transferred to ice and run over a Sephadex-G50 column (USA Scientific) to remove the fragmentation ions. The reason for using random priming rather than dT priming is that the latter typically produces a bias in the product that favors 3′ end representation. The extension issues do not affect all transcripts identically, and the cDNA population that results for each RNA species is a complex function of its specific properties as a substrate for reverse transcription and the overall transcript length. We used 3 μg random hexamers, added to prime first-strand reverse transcription according to the manufacturer's protocol (Invitrogen cDNA synthesis kit). After the first strand was synthesized, a custom second-strand synthesis buffer (Illumina) was added, and dNTPs, RNase H and *Escherichia coli* polymerase I were added to nick translate the second-strand synthesis for 2.5 h at 16 °C. The reaction was then cleaned up on a QiaQuick PCR column (Qiagen) and eluted in 30 μl EB buffer (Qiagen).

Sequencing and read mapping on the genome and across splices. Libraries were sequenced as 32-mers using the standard Solexa pipeline (version 0.2.6). Raw reads were then truncated as 25-mers and remapped with version 0.3 of the Efficient Local Alignment of Nucleotide Data (ELAND; A.J. Cox, personal communication)

software using the – –multi option against an expanded genome consisting of the standard mouse mm9 genome build and 42-mers representing the last 21 bp of the upstream exon and the first 21 bp of the corresponding downstream exon of each mRNA splice documented in the knownGene table for mm9 plus our spike sequences. We called a locus alternatively spliced whenever two or more different splice reads either started on different exons and terminated on the same exon, or vice versa.

Normalized gene locus expression level analysis and multiread probability assignment. We calculated normalized gene locus expression levels using the ERANGE package, which we developed for this purpose (see **Supplementary Software** and **Supplementary Methods** online).

Additional methods. Details of adaptor ligation, size selection and amplification, spike control derivation and validation, sequencing and read mapping on the genome, algorithm, and conversion of RPKM into absolute transcript numbers are available in the **Supplementary Methods.** Additional information, including more raw data and the software code (see **Supplementary Software**), are also available at http://woldlab.caltech.edu/~alim/RNA-seq/. Subsequent versions of the code will be posted on that website as they are developed.

Accession numbers. NCBI Short Read Archive SRA001030 (short tag data).

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

This work was supported by The Beckman Foundation, The Simons Foundation and US National Institutes of Health (NIH) grant U54 HG004576 to B.W. and R. Myers. A.M. was supported by an NIH training grant. The authors especially thank D. Trout and B. King for professional data handling and G. Schroth, I. Khrebtukova and S. Luo, of Illumina, for exchanges of preliminary data and protocols under development. M. Liu and J.L. Riechmann, along with others from the laboratories of B. Wold, R. Myers, J. Allman and P. Sternberg, are gratefully acknowledged for many helpful discussions, as are R. Myers and S. Mango for manuscript assistance.

1. Casneuf, T., Van de Peer, Y. & Huber, W. *In situ* analysis of cross-hybridisation on microarrays and inference of expression correlation. *BMC Bioinformatics* **8**, 461 (2007).
2. Eklund, A.C. *et al.* Replacing cRNA targets with cDNA reduces microarray cross-hybridization. *Nat. Biotechnol.* **24**, 1071–1073 (2006).
3. Okoniewski, M.J. & Miller, C.J. Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. *BMC Bioinformatics* **7**, 276 (2006).
4. Velculescu, V.E., Zhang, L., Vogelstein, B. & Kinzler, K. Serial analysis of gene expression. *Science* **270**, 484–487 (1995).
5. Harbers, M. & Carninci, P. Tag-based approaches for transcriptome research and genome annotation. *Nat. Methods* **2**, 495–502 (2005).
6. Brenner, S. *et al.* Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.* **18**, 630–634 (2000).
7. Boguski, M.S. & Toltoshev, C.M. Gene discovery in dbEST. *Science* **265**, 1993–1994 (1994).
8. Gerhard, D.S. *et al.* The status, quality, and expansion of the NIH full-length cDNA project: The Mammalian Gene Collection (MGC). *Genome Res.* **14**, 2121–2127 (2004).
9. Dias Neto, E.D. *et al.* Shotgun sequencing of the human transcriptome with ORF expressed sequence tags. *Proc. Natl. Acad. Sci. USA* **97**, 3491–3496 (2000).
10. Bertone, P. *et al.* Global identification of human transcribed sequences with genome tiling arrays. *Science* **306**, 2242–2246 (2004).
11. Cheng, J. *et al.* Transcription maps of 10 human chromosomes at 5-nucleotide resolution. *Science* **308**, 1149–1154 (2005).
12. Kapranov, P. *et al.* RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* **316**, 1484–1488 (2007).
13. Royce, T.E. *et al.* Issues in the analysis of oligonucleotide tiling microarrays for transcript mapping. *Trends Genet.* **21**, 466–475 (2005).
14. Kapranov, P., Willingham, A.T. & Gingeras, T.R. Genome-wide transcription and the implications for genomic organization. *Nat. Rev. Genet.* **8**, 413–423 (2007).
15. Nagalakshmi, U. *et al.* The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* published online, doi:10.1126/science.1158441 (1 May 2008).
16. Wilhelm, B.T. *et al.* Dynamic repertoire of a eukaryotic transcriptome surveyed at single nucleotide resolution. *Nature* advance online publication, doi:10.1038/nature07002 (2008).
17. Lister, R. *et al.* Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* **133**, 523–536 (2008).
18. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
19. Galau, G.A., Klein, W.H., Britten, R.J. & Davidson, E.H. Significance of rare mRNA sequences in liver. *Arch. Biochem. Biophys.* **179**, 584–599 (1977).
20. Kapur, K., Xing, Y., Ouyang, Z. & Wong, W.H. Exon arrays provide accurate assessments of gene expression. *Genome Biol.* **8**, R82 (2007).
21. Lee, C. & Roy, M. Analysis of alternative splicing with microarrays: successes and challenges. *Genome Biol.* **5**, 231 (2004).
22. Johnson, J.M. *et al.* Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science* **302**, 2141–2144 (2003).
23. Martin, J.F. *et al.* A Mef2 gene that generates a muscle-specific isoform via alternative mRNA splicing. *Mol. Cell. Biol.* **14**, 1647–1656 (1994).