Human transcriptome array for high-throughput clinical studies

Weihong Xuab,1, Junhee Seokc,1, Michael N. Mindrinosb,1, Anthony C. Schweitzerb, Hui Jiangb, Julie Wilhelmya, Tyson A. Clarkb, Karen Kapurc, Yi Xingc,d, Malek Fahamb, John D. Storeye, Lyle L. Moldaweref, Ronald V. Maierg, Ronald G. Tompkinsb, Wing Hung Wongc, Ronald W. Davisa,2, Wenzhong Xiaoh,2, and the Inflammation and Host Response to Injury Large-Scale Collaborative Research Program3

Development, Affymetrix, Inc., Santa Clara, CA 95051; bDepartment of Statistics and Department of Health Research and Policy, Stanford University, Stanford, CA 94305; cDepartment of Internal Medicine and Department of Biomedical Engineering, University of Iowa, Iowa City, IA 52242; dLewis-Sigler Institute and Department of Molecular Biology, Princeton University, Princeton, NJ 08544; eDepartment of Surgery, University of Florida College of Medicine, Gainesville, FL 32610; fDepartment of Surgery, University of Washington, Seattle, WA 98110; and gMassachusetts General Hospital, Harvard Medical School, Shriners Hospital for Children, Boston, MA 02114

Contributed by Ronald W. Davis, January 10, 2011 (sent for review November 15, 2010)

A 6.9 million-feature oligonucleotide array of the human transcriptome [Glue Grant human transcriptome (GG-H array)] has been developed for high-throughput and cost-effective analyses in clinical studies. This array allows comprehensive examination of gene expression and genome-wide identification of alternative splicing as well as detection of coding SNPs and noncoding transcripts. The performance of the array was examined and compared with mRNA sequencing (RNA-Seq) results over multiple independent replicates of liver and muscle samples. Compared with RNA-Seq of 46 million uniquely mappable reads per replicate, the GG-H array is highly reproducible in estimating gene and exon abundance. Although both platforms detect similar expression changes at the gene level, the GG-H array is more sensitive at the exon level. Deeper sequencing is required to adequately cover low-abundance transcripts. The array has been implemented in a multicenter clinical program and has generated high-quality, reproducible data. Considering the clinical trial requirements of cost, sample availability, and throughput, the GG-H array has a wide range of applications. An emerging approach for large-scale clinical genomic studies is to first use RNA-Seq to the sufficient depth for the discovery of transcriptome elements relevant to the disease process followed by high-throughput and reliable screening of these elements on thousands of patient samples using custom-designed arrays.


The authors declare no conflict of interest.

Freedly available online through the PNAS open access option.

Data deposition: The data reported in this paper have been deposited in the Gene Expression Omnibus (GEO) database, www.ncbi.nlm.nih.gov/geo (accession nos. GPL11319, GSE26072, and GSE26109).

1. To whom correspondence may be addressed. E-mail: dbowe@stanford.edu or wxiao@partners.org.

2. See Supporting Information for a complete list of individuals in the Inflammation and Host Response to Injury Large-Scale Collaborative Research Program.

3. See Supporting Information for a complete list of individuals in the Inflammation and Host Response to Injury Large-Scale Collaborative Research Program.

To date, the investigation of exon-level expression and alternative splicing in diseases has been hampered by the availability of high-throughput, reliable, and cost-effective platforms as well as appropriate methods for processing limited amounts of RNA (nanograms) from clinical samples. The Affymetrix Human Exon 1.0 ST Array is the only commercially available microarray for expression profiling at both gene and exon levels; although it provides accurate assessments of gene expression (15, 16), only 20% of its probe sets are supported by high-confidence annotations, and each probe set contains only one to four probes. As a result, the array is susceptible to false positives for analysis of exons and alternative splicing. An alternative design is to use probes targeting exon–exon junctions in addition to exons (17, 18), which has been applied to several custom arrays (19–21). In one example, Castle et al. (21) developed a set of 17 oligonucleotide arrays to survey genome-wide alternative splicing in 48 human tissues using pooled RNA samples. Recently, mRNA sequencing using deep-sequencing technologies has provided a new platform for the discovery and quantitation of genes and isoforms. However, for large clinical studies, further developments will be required to improve the current limitations on cost and throughput and the requirement of micrograms of total RNA for mRNA isolation. A major challenge of clinical studies is the limited amount of available material, especially from single-digit nanograms. Such challenges have limited the widespread study of alternative splicing in a clinical setting. Thus, we sought to develop a single array platform for cost-effective and reproducible analysis of gene and exon expression from nanograms of total RNA.

In this article, we report on the design and performance of the Glue Grant human transcriptome (GG-H) array, a 6.9 million-

D evelopments in genome technologies have expanded the scope of molecular medicine. To investigate complex and common diseases, such as inflammation and cancer, a clinical study typically requires high-quality genomic data from hundreds or preferably, thousands of patients, often from small quantities of tissues or blood. High-throughput DNA microarray technologies (1, 2) have been applied in many clinical investigations to obtain gene expression profiles of patient samples for disease predisposition, diagnostics, prognostics, and individualized treatment regimens (3–7).

Although most microarrays for RNA applications measure the expression level of individual genes, the human transcriptome is undoubtedly more complex. Alternative splicing, the process by which individual exons of pre-mRNAs are spliced to produce different isoforms of mRNA transcripts from the same gene, is a major source of the diversity of proteins and their functions in humans and other higher organisms (8–10). Defects in mRNA splicing are an important cause of diseases such as Alzheimer’s disease, cystic fibrosis, and multiple cancers; furthermore, as many as 50% of disease mutations in exons may impact on splicing (reviewed in refs. 11–14). It is, therefore, of great importance to be able to effectively measure levels of exon expression and identify alternative isoforms in human health and disease.
feature oligonucleotide array for comprehensive human transcriptome analysis. The array uses high-density tiling of probes over a large collection of transcriptome contents well-supported by experimental evidence to assess gene- and exon-level expression, detect alternative splicing, and analyze coding SNP and noncoding transcripts. To be able to routinely process limited quantities of tissue from clinical studies, we adapted a sample preparation method that can be used to efficiently process <50 ng total RNA for whole transcriptome analysis.

We validated the GG-H array in a testing experiment of multiple independent sample preparations and measurements of high-quality RNA samples from two distinct human tissues, liver and muscle, and compared the results with those from mRNA sequencing analysis of the same samples. We show that the array is highly reproducible in estimating expression at both gene and exon levels and sensitive in detecting expression changes. In addition, the majority of alternative splicing events detected on the array can be verified by mRNA sequencing data. We implemented the use of this array in a multicenter clinical program and have obtained high-quality, reproducible data.

Results

Overall Design of the GG-H Array. We designed a 6.9 million-feature oligonucleotide array for comprehensive measurement of the expression of genes and exons and detection of alternative splicing as well as analysis of additional contents of the human transcriptome, such as coding SNPs and several collections of noncoding transcripts (Fig. Lfi and Table 1).

A systematic examination of annotated transcripts yielded a comprehensive collection of unique transcripts from 35,123 transcript clusters (genes). Comparisons of these unique transcripts defined a set of 249,240 exon clusters and 315,137 probe selection regions (PSRs) for exon analysis. Furthermore, a set of 260,488 unique exon–exon junctions was defined based on the observed junctions between the adjacent exons on each transcript, with 32% constitutive and 68% alternatively spliced junctions. As an example, the SLK gene has, collectively, 9 unique transcripts, consisting of 19 exon clusters and 23 PSRs, as well as 19 junctions, including 16 constitutive and 3 alternatively spliced junctions (SI Appendix, Fig. S1). To achieve sensitive and robust measurement of exon-level quantitation and distinguish gene paralogs, 10 probes on average were designed for each PSR, and additional probes were selected for homologous regions and exons of long length, resulting in 119 unique probes on average for each gene. Four probes were designed for each exon–exon junction to aid the identification of alternative splicing.

SNPs in the coding regions or UTRs (86,954) were identified from National Center for Biotechnology Information (NCBI) dbSNP, including coding-synonymous (23,825), coding-nonsynonymous (21,950), and UTR SNPs (41,334). In addition, 2,828 DNA variations from 229 genes of drug-metabolizing enzymes and transporters (DMET) were also included. Six probes were designed for each allele of the 89,782 (86,954 + 2,828) variations, which allows for the study of allele-specific expression.

Noncoding RNAs (ncRNAs) are usually excluded during mRNA isolation. However, recent evidence suggests that the majority of the transcripts of human and other mammalian genomes is likely ncRNAs (8, 22, 23). From ncRNA databases, we identified 730 curated ncRNA species with known biological functions (f-ncRNA) (SI Appendix, Table S1A) and 50,783 antisense transcripts that overlap with RefSeq genes. Ten probes were designed for each of the content. In addition, we identified as targets 49,957 highly transcribed fragments of unknown functions (UTUs) (SI Appendix, Table S1B) in cytosol and nuclear regions of cells from Affymetrix tiling array data (24). Further details of the design of the array and software for array processing and data analysis are available in SI Appendix.

Array Contents Are Well Supported by RNA Sequencing of Multiple Tissues. Deep mRNA sequencing provides a comprehensive and objective survey of polyadenylated transcripts. To further evaluate the experimental evidence of the array targets, we compared the annotations of the array with publicly available sequencing data of mRNA from 10 diverse human tissue transcriptomes, where 5.3–20.2 million 32-bp uniquely mappable reads were acquired for each tissue (9). If a genomic region targeted by the array has uniquely mapped sequencing reads from one or more human tissues, it would support the potential existence of mRNA transcription from the targeted region.

The design of the GG-H array is well-supported by mRNA sequencing data from the multiple tissues (Fig. 1B). Overall, 94.5% of the uniquely mapped mRNA sequencing reads across the 10 tissues fell into the targeted regions of the array, including 85% on exons and 7% on junctions. Among the exons designed

Table 1. Summary of the contents of the GG-H array

<table>
<thead>
<tr>
<th>Array components</th>
<th>No. of targets</th>
<th>No. of probes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene exons</td>
<td>315,137</td>
<td>3,292,929</td>
</tr>
<tr>
<td>Exon–exon junctions</td>
<td>260,488</td>
<td>1,060,703</td>
</tr>
<tr>
<td>Coding SNPs and DMET variations</td>
<td>89,782</td>
<td>982,941</td>
</tr>
<tr>
<td>Noncoding functional RNA (f-ncRNA)</td>
<td>730</td>
<td>5,869</td>
</tr>
<tr>
<td>Noncoding antisense expression</td>
<td>50,783</td>
<td>563,097</td>
</tr>
<tr>
<td>(as-ncRNA)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unannotated transcribed units (UTUs)</td>
<td>49,957</td>
<td>488,581</td>
</tr>
<tr>
<td>Other probes including controls</td>
<td>496,840</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>6,892,960</td>
<td></td>
</tr>
</tbody>
</table>
on the array, 81% had at least one uniquely mapped read in sequencing, and 72% had at least two reads; additionally, 47% of the exons were supported by 10 or more reads. Similarly, 72% of the junctions targeted on the array were supported by at least one unique read; however, the percentage of supported junctions dropped quickly with the increased number of mapped reads, with 54% for at least 2 reads and only 18% for more than 10 reads. This is likely because junction regions are typically shorter than exons. Interestingly, 49% of the previously unannotated UTUs also had uniquely mapped reads (30% for at least two reads), although this coverage was lower than that of exons and junctions. As a comparison, functional noncoding RNAs are covered by roughly 20% because of their frequent lack of polyadenylation.

A fraction of the exon and junction targets designed on the array did not have reads in the RNA sequencing (RNA-Seq) data. This is expected, because a percentage of the exons and junctions can be expressed in other tissues. In addition, deeper sequencing is generally required for better coverage of low-abundance transcripts, as suggested by the observed decreasing coverage of exons and junctions with respect to the number of mapped reads in the dataset. Finally, although only uniquely mapped reads were included in the analysis, the sequencing errors of 1–2% can lead to mismappings; however, the false-positive rate is estimated to be quite low (no more than 0.1%) (9).

Evaluation of the Reproducibility of the Array. We designed experiments to test the performance of the array. Using reference RNA samples of liver and muscle tissues, we processed each tissue independently four times and hybridized to a total of eight arrays. To be able to routinely process very limited quantities of RNA in clinical studies, 50 ng total RNA were chosen to be the amount of material analyzed in each sample. In parallel, 2 μg same liver and muscle samples were used to perform four independent replicates of mRNA isolation, processing, and sequencing. mRNA sequencing reads were mapped over the genome as well as junction regions and generated 46 million uniquely mappable reads per sample or a total of 366 million, among which 94% fall within the target regions of the array, including 85% on exons and 9% on junctions. For each of the eight runs, on average, 39 million reads were uniquely mapped to exons, and 4 million reads were mapped to junctions, which were included in the further analysis. Overall, 75% of the genes and 62% of the exons detected in a tissue have more than five reads, with 85% and 40%, respectively, having more than 20 reads (SI Appendix, Fig. S4). The expression level of the genes and exons was calculated for each sequencing run as reads per kilobase per one million mapped reads (RPKM).

The GG-H array uses high-density tiling of probes over transcriptome targets for the robust estimation of expression levels. As an example, SI Appendix, Fig. S5A shows the reproducibility of the raw signal of probes for several adjacent exons of the gene SLK, which was measured by a total of 306 probes on the GG-H array, including 10 probes targeting each of its 23 exons and 4 probes targeting each of its 19 junctions. Compared with 11 probes per gene on the Affymetrix U133 array or 2–4 probes per exon on the Affymetrix exon Array, the GG-H array uses 10 probes targeting one exon to ensure the reproducibility for exon-level expression similar to the gene-level measurement on the Affymetrix U133 array, and it provides even more robustness for the expression estimation of genes with more than 100 probes. In addition, previous studies had revealed that exon 15 of SLK (chr10: 105,760,564–105,760,656) is alternatively spliced between liver and muscle (25). In our data (SI Appendix, Fig. S5B), hierarchical clustering of the probe signals from liver and muscle samples also classified the SLK probes into two groups inversely correlated with each other: a tightly clustered smaller group consisting of mainly probes targeting exon 15 and its adjacent junctions and a larger group of other probes targeting the remaining exons and junctions. Here, the larger cluster reflects changes of gene expression level between the two tissues, and the smaller cluster indicates alternative splicing. As expected, several probes were misclassified; these probes had low signal levels indistinguishable from the background noise and failed to reflect the expression of their targets. It is important to note that, although these probes performed poorly, the multiple additional probes from the high-density tiling ensure the capture of differential gene and exon expression as well as alternative splicing events.

The reproducibility of GG-H array at the gene and exon levels was examined and compared with mRNA sequencing results over the four independent replicates of liver and muscle samples (Fig. 2A). Pearson correlation coefficients were calculated over logged expression values. For both array and mRNA sequencing, the calculation of correlation included the same set of 140,000 exons and 17,000 genes that have nonzero reads in sequencing. Although quantile normalization across multiple chips was often used in microarray analysis to improve reproducibility and reduce variation (26), here, we chose instead to scale probe intensities of each array by their median value, similar to the approach of RPKM in mRNA sequencing that normalizes the expression values with the total number of reads. As shown in Fig. 2A, the level of exon and gene expression measured by GG-H array is highly reproducible. Between independent repeats, the correlation coefficients were >0.99 for genes and 0.99 for exons. The mRNA sequencing showed gene correlation of 0.99, similar to the array, whereas the correlation of exon expression was 0.96, lower than that of the array. The observed lower reproducibility of exons in sequencing is most obvious with low-abundance exons in which a difference of one or two reads between repeats can induce a large variance in the estimation when there is only one to a few reads for the exon. This effect is less problematic when estimating gene expression, because on average, a gene includes 10 exons, thereby producing ten times the amount of mapped reads.

Similar observations can be made when comparing the coefficient of variation (CoV) between the array and sequencing.

Fig. 2. Comparison of the performances of the GG-H array and mRNA sequencing. (A) Reproducibility of exon and gene expression measured by array and sequencing. In each panel, scatter plots of logged expression values between four independent replicates of a reference muscle sample are shown in the bottom left corner and in the top right corner, the corresponding Pearson correlation coefficients are shown. (B) Comparison of differentially expressed exons and genes identified by GG-H array and mRNA sequencing.

Xu et al.
data. As shown in SI Appendix, Fig. S6A, the gene and exon expression measured by GG-H array has low variance over a wide range, and the median CoV of gene and exon expression is 0.06 and 0.11, respectively. In comparison, mRNA sequencing data showed median CoV of 0.12 and 0.31 for genes and exons, respectively, and large variances are seen for those genes and exons covered with a small number of reads. In this experiment, 80% genes and 85% exons show lower CoVs in the array estimation than mRNA-Seq, with 46 million uniquely mapped reads. Both sample preparation and sequencing step contribute to the observed variance in the sequencing data, and deeper sequencing is required to improve its reproducibility (SI Appendix, Fig. S7).

**Array Has High Sensitivity and Specificity in Detecting Differential Expression.** We identified genes and exons significantly differentially expressed between liver and muscle tissues using the four independent array repeats of each tissue and compared the results with the four sequencing repeats for each of the same tissues. To avoid potential bias from different parametric models, we chose to analyze the data using a nonparametric permutation test that assumes no underlying signal distributions. The method described by Tusher et al. (27) was applied to the analysis of both array and sequencing data, and differentially expressed genes and exons were detected with a false discovery rate (FDR) of <0.005.

A significant percentage of the exons and genes identified by the GG-H arrays was also identified in the sequencing analysis (Fig. 2B) (P value < 10−16). Between liver and muscle tissues, ~14,000 genes and ~114,000 exons were detected as differentially expressed by the arrays. About 80% of these genes and exons were covered by more than five sequencing reads in at least one tissue, among which more than 90% showed the same direction of expression change in the sequencing data. In addition, as shown in SI Appendix, Fig. S8, whereas array and sequencing identified a similar number of significant expression changes among abundant genes, sequencing identified many fewer genes among those detected by less than 20 reads. Deeper sequencing is required to detect the expression of these genes and exons (~35% of genes and ~60% of exons). Overall, these results show that the GG-H array has high sensitivity and specificity in measuring differential expression of exons and genes.

**Improved Identification of Alternative Splicing by the Array.** Data from multiple exon and junction probe sets on the array were used to detect alternative splicing events. Junction probes can significantly improve the percentage of true-positive detections in alternative splicing analysis when combined with exon probes. For example, the two isoforms of SLK distinguish each other by either skipping or including exon 15. As shown in Fig. 3, changes of proportions between these two isoforms can be observed from the changes of signals of the corresponding exon as well as its neighbor junctions. In liver, the abundant expression of exon 15, its two connecting junctions, and the bridging junction between exon 14 and 16 imply that both isoforms are present, whereas in muscle, the reduced expression of exon 15 and its connecting junctions accompanied by the increased expression of the bridging junctions reveal the alternative splicing to the isoform that skips exon 15.

By combining such information on exons and corresponding junctions, more than 6,000 alternative spliced events were identified from the array data between liver and muscle. Briefly, for each exon, its expression level was normalized to the expression of its corresponding gene, and candidate exons of alternative splicing were identified based on the significant changes of their normalized expression levels between the conditions. Similarly, junction probe sets were prescreened to remove those not significantly above the background across all of the conditions, and candidate junctions were identified based on the significant changes of their expression levels normalized to gene expression. To reduce false positives, each alternative spliced event identified required a candidate exon for alternative splicing supported by at least one of its connecting junctions. Compared with data from sequencing, 62% of the exons identified by array were covered by, on average, more than five sequencing reads in at least one tissue, among which 56% showed a change of the same direction by more than twofold, indicating that the array is a reliable platform to evaluate alternative splicing.

**Measuring Noncoding RNA Using the Array.** The GG-H array also measures the expression of less well-studied contents of the human transcriptome, such as UTUs and ncRNA. In addition, transcripts lacking polyadenylation are usually excluded during the mRNA isolation. Our array data revealed that substantial numbers of these transcripts lacking polyA were well-detected in liver and muscle (SI Appendix, Fig. S9A). For example, many ncRNAs of diverse functions were highly expressed, including signal recognition particle RNA 7SLs, small cytoplasmic Alu RNAs (scAlu), small nuclear RNA 7sk, spliceosomal RNA U2, and several small nuclear RNAs (snRNAs) and small Cajal body-specific RNAs (scaRNAs). As expected, none were well-detected in our mRNA sequencing data. Interestingly, the two functional families of snoRNAs, the C/D box family that is associated with RNA methylation and the H/ACA box family associated with pseudouridylation, were strongly differentially expressed between the tissues: C/D box snoRNAs were overexpressed in muscle, whereas H/ACA box snoRNAs were dominant in liver (SI Appendix, Fig. S9B). Further studies are necessary to understand the biological implications of these findings.
Discussion

High-throughput genomic approaches can greatly accelerate our understanding of the molecular mechanisms of diseases. However, a clinically useful genomic technique has several important requirements. First, a disease study typically involves acquiring genomic data from hundreds to thousands of patient samples at multiple time points, which demands rapid turnaround and cost effectiveness. Second, clinical studies are restricted in the quantity of biological materials that can be sampled from a patient. Third, patient enrollment is demanding and often costly; therefore, high sensitivity and specificity are required, and the power of a study is dependent on the reproducibility of the analytical measurement.

We have developed a human transcriptome array platform that meets these requirements. The array uses a high-density tiling approach for the measurement of gene and exon expression and genome-wide identification of alternative splicing as well as analysis for coding SNP detection and noncoding transcripts. Care was taken in probe selection to optimize probe performance and avoid cross-hybridization. The content designed on the array is well-supported by RNA sequencing data of multiple tissues. We validated the array using multiple independent preparations of 50 ng total RNA as the starting material and compared the results with RNA sequencing analysis at 2 μg total RNA. Our results showed that the GG-H array is highly reproducible in estimating mRNA abundance at the gene and exon level and is sensitive in detecting changes. In addition, the exon–exon junction probes of this array were shown to improve the detection of alternative splicing events.

Recent developments in mRNA sequencing provide a comprehensive and objective survey of the polyadenylated transcriptome. In addition, the dynamic range of signal of deep sequencing can be greater than that of microarrays (SI Appendix, Fig. S6B). Because the sequencing platform counts a transcript in an RNA sample according to its abundance, the most-abundant species can be sampled hundreds of thousands or millions of times before the less-abundant species are sampled; therefore, the coverage and reproducibility of low-abundance exons and genes are critically dependent on the total number of reads. In our data of 46 M uniquely mapped reads per replicate, 35% of the genes and 60% of the exons detected in a tissue are covered by fewer than 20 reads, where large variation is observed and fewer significant changes are identified by sequencing. Sequencing with fewer reads would lead to higher percentages of genes and exons inadequately covered (SI Appendix, Fig. S4). In the Glue Grant clinical studies, expression of many immune mediators, such as IFNs, ILs, and chemokines, is observed at the lower 35% of the transcriptomes of T cells, monocytes, and neutrophils. Deeper sequencing, therefore, is required to adequately cover these genes, especially when considering the dollar cost of patient enrollment and sample collection (tens of thousands dollars per subject in the Glue Grant). Besides, because mRNA transcripts usually count for only a small percentage of total cellular RNA, purification of polyA RNA becomes necessary for mRNA sequencing, which at the present time, requires micrograms of starting material. In contrast, microarrays use predesigned hybridization probes to allow the detection of transcriptome targets in the presence of other RNA species and without prior polyA RNA enrichment. To meet the requirements for large-scale clinical studies, reduction in both the quantity of starting materials and cost of each assay as well as sharp reduction in run time would be required.

For example, current throughput is considerably different with the two approaches. With similar initial instrument cost, an average analytical core laboratory can be equipped with one Affymetrix 7G scanner and four hybridization stations for arrays (~$400,000), one Illumina Genome Analyzer II (~$600,000), or one HiSeq 2000 analyzer (~$750,000). A typical large-scale clinical study, such as our clinical trial examining the genomic response to severe injury, performs transcriptome analysis on about 5,000 clinical samples. Table 2 compares the current throughput using microarrays and sequencing with the Genome Analyzer II or anticipated throughput of the HiSeq 2000 as well as the time required for processing these samples using each of the three platforms. With current technologies, completing the analysis of a clinical trial of this size would be challenging with present mRNA sequencing technologies alone.

However, the information generated by RNA sequencing can benefit from array design and verification tremendously. In this study, although 94% of the sequencing reads fell into targeted regions of the array, the remaining were mapped to other genomic regions, which provide valuable information to improve the array design. As an example, we performed de novo identification of exon–exon junctions from the RNA sequencing data (28) (SI Appendix, Fig. S10). Although overall GG-H junctions cover the highly expressed junctions identified by the de novo method very well, a total of 6,581 additional de novo junctions supported by more than four reads (~1 RPKM) were discovered and will be included in the update of array design. In addition, by analyzing the array signals across a large number of samples over multiple tissues and comparing array results with those of sequencing, expression estimation of the array can be calibrated, and nonperforming probes can be detected and removed.

All in all, an approach has emerged to apply RNA sequencing and arrays in combination to large-scale clinical studies, where RNA-Seq with sufficient depth is first conducted to discover all of the transcriptome elements potentially relevant to the disease of study followed by high-throughput and reliable screening of these elements on thousands of patient samples using custom-designed arrays.

We have implemented this array platform within an ongoing multicenter clinical study to examine the temporal genomic re-

Table 2. Comparison of the throughput of transcriptome analysis

<table>
<thead>
<tr>
<th>Gene</th>
<th>Throughput (samples per week)</th>
<th>Processing time for 5,000 samples (wk)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GG-H array</td>
<td>200</td>
<td>25</td>
</tr>
<tr>
<td>RNA-Seq (GA II)</td>
<td>2</td>
<td>5,000</td>
</tr>
<tr>
<td>RNA-Seq (HiSeq 2000)</td>
<td>30</td>
<td>167</td>
</tr>
</tbody>
</table>

The number of reads required for RNA sequencing depends on the objectives of the study. To achieve the same level of reproducibility as the GG-H array, ~150 million reads for genes and 200 million reads for exons are likely required (SI Appendix, Fig. S7). To detect differential expression, the array detects about the same number of genes and about two times the number of exons as RNA-Seq of 46 million uniquely mappable reads, and deeper sequencing is required for low-abundance genes (SI Appendix, Fig. S8). Here, for the throughput comparison, we assumed 46 and 92 million uniquely mappable reads for differential expression analysis of genes and exons, respectively. The throughput of HiSeq 2000 for RNA sequencing was estimated based on 700 million uniquely mappable reads per run and two runs per week.
spontaneous response to severe injury in blood leukocytes, enriched monococytes, T cells, and neutrophils (29, 30). Using macroscale and micro-fluidic approaches to isolate these cell populations from <14 mL blood, we have successfully analyzed more than 1,800 samples of 25–200 ng total RNA obtained from more than 100 patients. With current technology and a throughput of hundreds of samples per week in an average core facility at a current cost of approximately $400/sample, we anticipate that this array platform will have a wide range of applications in high-throughput clinical studies.

Materials and Methods

Design of the Human Transcriptome Array. Details of the design of the array and its availability as software for array processing and data analysis are available in SI Appendix.

Design of exon and junction probes. On average, 10 probes were designed for each PSR, and 119 unique probes, on average, were designed for each gene; care was taken in probe selection to optimize probe performance and avoid cross-hybridization. To design probes for the array, we considered three important factors: (i) probe performance by thermodynamics calculation, (ii) sequence uniqueness against the transcribed regions and whenever possible, whole genome (the sequence of a desirable probe is unique without any 17mer or more off-target perfect matches or up to three base off-target mismatches, including insertions/deletions), and (iii) spreadness of the selected probes across the PSR. Four probes were designed for each exon–exon junction at positions −3, −1, +1, and +3.

Various control probes. Quality control is important for microarray studies of clinical samples (31). In addition to the standard control probes on Affymetrix GeneChips, several additional probe sets were included for the purpose of quality control of the assay, background modeling, estimation of cross-hybridization, and monitoring of the ribosomal RNA signal in the amplified material.

Sample Processing and Analysis. Sample preparation for GG-H array. Reference RNA samples of liver and muscle tissues were obtained from Ambion. The detailed protocol is described in SI Appendix. The array was washed, stained, and scanned using Affymetrix Fluidics Station FS450 and GeneChip Scanner 3000 7G.

Sample preparation for Illumina sequencing. Library preparation and sequencing were performed as described (32). Briefly, starting with 2 μg total human liver and muscle mRNA (Ambion), mRNA extraction was performed using the Illumina mRNA-Seq Sample Prep Kit. Refer to the user guide for the full protocol through final library generation. The Single Read Cluster Generation Kit and Sequencing Kit v3 by Illumina were used to sequence the libraries. Three or four lanes were used for each sample.

GG-H array data processing. Raw signals of the arrays were processed using Affymetrix Power Tools, including Robust Multisarray Analysis (RMA) for background correction, median scaling for normalization, and median polish for summarization (26). Gene and exon expressions were calculated according to annotations of the GG-H array. Sense and antisense probe set signals were combined for UTU expression. Expression of junctions, non-coding exons, and other features was calculated as expression of the corresponding probe sets. For alternative splicing analysis, briefly, Junction and Exon array Toolkits for Transcriptome Analysis (JETTA) software was used with the cutoff criteria of Microarray Data Analysis System (MIDAS) (25) P value <0.01 and fold change > 2 for the alternatively spliced exon and at least one of its adjacent junctions. Details of the computational algorithm, software, and visualization tools are described in SI Appendix. The data have been deposited in Gene Expression Omnibus (GEO) under the accession number GSE26072.

mRNA-Seq data processing. mRNA-Seq reads were mapped over the genome as well as junction regions defined by the GG-H array annotation using 36 bp. Junction regions were defined to span 27 bp over each side of two connected exons. SeqMap was used for the mapping, with allowance of 2-bp mismatches (33). About 67% of the total reads from each run were mapped to one single location of the genome, resulting in, on average, 13 million uniquely mappable reads per lane to the genome sequence, which was comparable with other published studies (9, 34). Expression indices of the abundance of genes, exons, and other features were calculated as RPKM (9). The abundance of genes calculated in liver and muscle tissues also agree well with other published datasets; for example, compared with data on the human muscle tissue from Wang et al. (9), the Pearson correlation coefficient was over 0.9. De novo identification of exon–exon junctions was performed using SpliceMap from mRNA on 120 million 58-bp RNA-Seq data (28). The data have been deposited to the NCBI short-read archive under the accession number GSE26109.

ACKNOWLEDGMENTS. We thank Dr. Weng-Onn Lui for assistance and comments during the development of the GG-H array and the groups of Drs. Michael Snyder, Steven Kunkel, and Ulysses Ball for assistance with the sequencing experiments. This work was supported in part by National Institutes of Health Grants U54GM062619, P01HG000205, and R01HG006434.

Supporting Information

1. Design of the GG-H Array

Annotation of gene structure

To systematically examine the ‘annotated’ and experimentally ‘observed’ transcripts in terms of exons and exon-exon junctions, annotated contents were collected from RefSeq, Ensembl, and UCSC Known Genes based on human genome assembly of HG18, and complemented by information of EBI's AEDB exons (literature-confirmed) and UCLA ASAP2 cassette exons. ExonWalk program was then used to merge EST evidence and annotated contents together to predict full-length isoforms, including alternative transcripts. This yielded a comprehensive collection of 335,663 unique transcripts, consisting of a total of 370,295 unique exons. These unique transcripts formed 35,123 transcript clusters (genes), and the set of unique exons defined 249,240 exon clusters and 315,137 probe selection regions (PSRs). The set of 260,488 exon-exon junctions was defined based on observed junctions between unique exons, with ~32% constitutive and ~68% alternatively spliced junctions.

As an example, SLK gene has collectively nine input transcripts in RefSeq, Ensembl, UCSC KG and Exonwalk as one transcript cluster. This transcript cluster includes 19 exon clusters and 23 PSRs, as well as totally 19 junctions including 16 constitutive and 3 alternatively spliced (Fig. S1). A database of the annotations of gene structures is available on our supporting website at http://gluegrant1.stanford.edu/~DIC/GGHarray/.
**Fig. S1.** An example of the GG-H array annotation of genes (transcript clusters). A. Annotations of transcripts, exons, and junctions of SLK gene from RefSeq, Ensembl, UCSC Known Genes, Exonwalk and other database sources. B. Design of exon and junction probesets targeting the gene SLK. Shown are probesets for nine selected exons and associated junctions.

**Design of exon and junction probes**

On average, ten probes were designed for each PSR and additional probes selected for homologous regions and exons longer than 2KB, resulting in 119 unique probes on average for each gene; four probes were designed for each exon-exon junction at positions -3, -1, +1, +3 relative to the splicing site (Fig. S2A).

To design probes for the exons, we considered three important factors: (1) probe performance by thermodynamics calculation; (2) sequence uniqueness against the transcribed regions and whenever possible, against the whole genome - a desirable probe is unique without any off-target 17mer perfect matches and no up to three base off-target mismatches including insertions/deletions; and (3) spreadness of the selected probes across the probe selection region.
The performance of a candidate probe was examined by a thermodynamic model proposed in Mei et al. 2003 (1). The sequence uniqueness of a candidate probe was analyzed against both the transcribed region only and the whole genome in three steps. First, the number of 17-mer off-target perfect matches to a candidate probe of 25 nucleotides is used as a measure of potential cross-hybridization, which should be minimized. Second, the number of mismatch targets up to three bases, including insertions and deletions, should be minimized. An ideal probe would have no mismatch targets at all. Third, if there were mismatches, the location of its mismatch targets on the genome should not match the sequence of other selected probes for the array. In addition, when multiple probes are to be designed for a probe selection region (PSR), the distribution of probes should be evaluated to avoid the possibility of picking best probes clustered in a small region.

For the GG-H array, one-half of the ten probes for each PSR were selected to maximize probe thermodynamics and sequence uniqueness over the PSR region, and the other half were selected to maximize the spreadness of the probes by picking the probe with best performance and uniqueness for each 1/6 window of the PSR. Further, additional probes for PSR longer than 2kb were added at a pace of one probe every 100bp over 2KB. For homologous PSRs, additional
probes were also added specifically targeting unique regions to help distinguish the expression of homologous PSRs.

**Probes for SNPs in coding and untranslated (UTR) regions**

From the SNP126 database, we selected SNPs in exons or 3’, 5’ UTR regions of genes meeting the following criteria: (1) the SNP falls within any exon or UTR region defined as above; (2) the SNP is validated; (3) the SNP sequence is mapped to a single location on the genome. This resulted in 86,954 coding SNPs (cSNPs), including coding-synonymous SNPs – 23,825, coding-nonsynonymous SNPs – 21,950, and UTR SNPs – 41,334. In addition, 2,828 DNA variations in 229 genes of drug metabolizing enzymes and transporters (DMET) were also identified.

Six probes were designed for each allele at -4, 0, and +4 positions relative to the SNP, except that when the SNP position is less than 17 bases to one end of an exon, the probes were shifted to the other side (Fig. S2B). Similar probe design strategy was utilized in commercial and research Affymetrix GeneChips™ for SNP genotyping. These SNP probes will also allow us to study allele specific expression in human samples.

In addition, DNA TAG probes (2) were included on the array allowing the flexibility to analyze up to 70-80 thousand selected SNPs simultaneously using MIPS technique.

**Survey of noncoding transcripts**

We manually reviewed entries in several functional and regulatory ncRNA databases, including SILVA rRNA database (http://www.arb-silva.de/), genomic tRNA database (http://lowelab.ucsc.edu/GtRNAdb/), snoRNABase (http://www-snorna.biotoul.fr/), Signal

Table S1. Survey of noncoding RNA species on the GG-H Array.

(A) Known functional and regulatory noncoding RNA selected for the array.

<table>
<thead>
<tr>
<th>Family</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>C/D box snoRNA</td>
<td>255</td>
</tr>
<tr>
<td>snRNA</td>
<td>111</td>
</tr>
<tr>
<td>Disease related</td>
<td>104</td>
</tr>
<tr>
<td>H/ACA box snoRNA</td>
<td>94</td>
</tr>
<tr>
<td>Y RNA</td>
<td>50</td>
</tr>
<tr>
<td>SRP_7SL RNA</td>
<td>35</td>
</tr>
<tr>
<td>Cajal body-specific ScaRNA</td>
<td>25</td>
</tr>
<tr>
<td>7SK RNA</td>
<td>13</td>
</tr>
<tr>
<td>scAlu RNA</td>
<td>13</td>
</tr>
<tr>
<td>Imprinted</td>
<td>11</td>
</tr>
<tr>
<td>RNase MRP RNA</td>
<td>9</td>
</tr>
<tr>
<td>Vault RNA</td>
<td>5</td>
</tr>
<tr>
<td>Telomerase RNA</td>
<td>4</td>
</tr>
<tr>
<td>RNase P RNA</td>
<td>1</td>
</tr>
</tbody>
</table>

(B) Unannotated transcribed units (UTUs) identified in all the 8 cytosol and 2 nuclear conditions from (3).

<table>
<thead>
<tr>
<th>Min Length</th>
<th>0</th>
<th>25</th>
<th>100*</th>
</tr>
</thead>
<tbody>
<tr>
<td># bases</td>
<td>36Mb</td>
<td>36Mb</td>
<td>8.5Mb</td>
</tr>
</tbody>
</table>

*Selected for GG-H array

From NATsDB (http://natsdb.cbi.pku.edu.cn/), we identified 6,025 antisense transcripts that overlap with RefSeq genes, and ten probes were designed for each transcript. In addition, to survey the potential antisense transcripts overlapping with the UTR region, we targeted the antisense strands of 44,758 3’ and 5’ UTR regions of the RefSeq genes, and designed probes at the density of one probe per 50 bp of UTR and with a minimum of six probes per region.
We also analyzed the transcribed fragments of unknown functions (UTUs) from Affymetrix tiling array data of polyA RNA isolations from eight cytosol and two nuclear conditions by Kapranov et al., 2007 (3). Since there can be differences in transcribed RNAs between the nuclear region and the cytosolic region, we identified those expressed in either all of the cytosol or nuclear conditions. About 50 thousand of such transcripts having a minimum length of 100 bases, were included as targets for the array (Table S1B). Since the tiling array experiment did not distinguish the strand of the transcript, ten probes were designed for both strands of each UTU (five probes per strand x two).

Control probes.

Quality control is important for microarray studies of clinical samples (4); therefore we included in the array design several sets of control probes (Table S2). First, the complete sets of controls probes for the Affymetrix GeneChips™ are used including quality controls (polyA spike-ins, default controls, and additional controls) and background modeling (antigenome 25 mers, big GC and norm genes). Second, we included as controls probes targeting 140 external RNA control
species (ERC) developed by the External RNA Control Consortium (ERCC) (5). These ERCs are synthetic RNA to be added to a sample for the purpose of quality control of the assay, and have been extensively prototyped and optimized for performance on microarray platforms. Third, we also included both 22 mer probes of antigenomic sequences that are not homologous to human genome, for background modeling of small RNA probes. Fourth, to better understand cross hybridization of oligonucleotides on the array, additional probes were designed with 0 (perfect match) to 4 base mismatches as well as insertion(s)/deletion(s) to the sequences of Affymetrix poly-A spike-in controls. Finally, to monitor the ribosomal RNA signal in the amplified material, we included probes with 0-4 base mismatches to rRNA, as the rRNA signal can be overwhelming.

**Database of Array Annotation**

The array annotation database is designed with the concept of entity-relation model such that probe information (oligo property and array layout), target information (transcript, gene, SNP, etc) and probe-to-target matching information are stored in different tables. A summary of the schema is shown in Fig. S3, where entities and relations are illustrated as circles and double-headed arrows respectively. Since the contents of the multiple assays on the array are stored independently, our design allows each table to be updated individually without affecting the rest, making it possible for each table to be synchronized to its public databases regularly. Besides, in each target information table, new relations can be

---

*Fig. S3.* Summary of the schema of the array annotation database.
built to include new information from other databases. Further details of the design of the array and the annotation database are available at http://gluegrant1.stanford.edu/~DIC/GGHarray/.

2. Software for array processing and data analysis

We have developed a pipeline for array processing and data analysis, which includes: (1) Quality control, low-level analysis and expression analysis using Affymetrix Power Tools (APT); (2) High-level exploratory analysis using dChip; (3) Alternative splicing analysis using Junction and Exon array Toolkits for Transcriptome Analysis (JETTA). The details of array processing and data analysis, including supporting library files and annotation files, can be found at our supporting website at http://gluegrant1.stanford.edu/~DIC/GGHarray/.

3. Protocol of sample processing for GG-H array

Sample processing protocol was developed to work efficiently with small amount of starting material. Briefly, fifty nanogram aliquots of total cellular RNA were converted to double stranded cDNA using custom-designed random primers containing the T7 polymerase promoter region and conventional enzymatic steps. Subsequently, T7 RNA polymerase was used to produce and amplify antisense cRNA, which was used as starting material to produce double stranded labeled cDNA for hybridization. At this step, random primers were annealed to the cRNA and the subsequent first and second strand synthesis reactions were performed using dNTP’s with both thymine and uracil at a ratio of 4:1, utilizing conventional enzymatic steps.
The double stranded cDNA was then fragmented by uracil DNA glycosylase and the digested fragments were labeled with deoxynucleotidyl transferase (rTdT) and the biotin-conjugated nucleotide analogue, DLR—1a. After the labeling reaction, the sample was hybridized overnight. The array was washed, stained, and scanned using Affymetrix Fluidics Station FS450 and GeneChipScanner3000 7G. The detailed protocol is described in the protocol section of our supporting website at http://gluegrant1.stanford.edu/~DIC/GGHarray/.

4. Comparison of array annotations with RNA sequencing data from multiple tissues

Table S3. Mapping of RNA sequencing data from multiple tissues to the target genomic regions of the array. Sequencing data of the 10 tissues were from Wang et al. 2008. Overall, 94.5% of the uniquely mapped RNA-Seq reads across the 10 tissues fall in the target regions of GG-H, including 85% on exons and 7% on junctions.

<table>
<thead>
<tr>
<th>Tissue</th>
<th>PSR</th>
<th>JUNC</th>
<th>UTU</th>
<th>as-ncRNA</th>
<th>f-ncRNA</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adipose</td>
<td>16.6M</td>
<td>1.4M</td>
<td>0.2M</td>
<td>77.0K</td>
<td>7.4K</td>
<td>18.2M</td>
</tr>
<tr>
<td>Brain</td>
<td>10.7M</td>
<td>0.6M</td>
<td>0.1M</td>
<td>29.4K</td>
<td>5.6K</td>
<td>11.5M</td>
</tr>
<tr>
<td>Breast</td>
<td>8.5M</td>
<td>0.8M</td>
<td>0.1M</td>
<td>46.9K</td>
<td>6.8K</td>
<td>9.4M</td>
</tr>
<tr>
<td>Colon</td>
<td>17.2M</td>
<td>1.3M</td>
<td>0.2M</td>
<td>72.4K</td>
<td>6.7K</td>
<td>18.8M</td>
</tr>
<tr>
<td>Heart</td>
<td>12.0M</td>
<td>0.7M</td>
<td>0.3M</td>
<td>30.4K</td>
<td>11K</td>
<td>13.0M</td>
</tr>
<tr>
<td>Liver</td>
<td>10.7M</td>
<td>0.9M</td>
<td>88.9K</td>
<td>42.6K</td>
<td>3.5K</td>
<td>11.8M</td>
</tr>
<tr>
<td>Lymph</td>
<td>13.0M</td>
<td>1.5M</td>
<td>0.2M</td>
<td>0.2M</td>
<td>6.3K</td>
<td>14.9M</td>
</tr>
<tr>
<td>Muscle</td>
<td>13.9M</td>
<td>1.3M</td>
<td>0.2M</td>
<td>51.8K</td>
<td>9.0K</td>
<td>15.4M</td>
</tr>
<tr>
<td>Testes</td>
<td>16.1M</td>
<td>1.6M</td>
<td>0.1M</td>
<td>85.2K</td>
<td>6.8K</td>
<td>18.0M</td>
</tr>
<tr>
<td>UHR (low cov.)</td>
<td>4.1M</td>
<td>0.5M</td>
<td>70.8K</td>
<td>29.1K</td>
<td>3.9K</td>
<td>4.6M</td>
</tr>
<tr>
<td><strong>SUM</strong></td>
<td>122.8M</td>
<td>10.7M</td>
<td>1.4M</td>
<td>0.7M</td>
<td>67.1K</td>
<td>136M</td>
</tr>
</tbody>
</table>
5. RNA-Seq analysis
Two micrograms of the same liver and muscle samples were used to perform four independent repeats of mRNA processing and sequencing analysis of each tissue utilizing the Illumina Genome Analyzer II. For each of the eight runs, on average, 39 million reads were uniquely mapped to exons and 3.9 million to junctions, which were included in the further analysis.

<table>
<thead>
<tr>
<th>Run</th>
<th>Liver</th>
<th>Muscle</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>47.7M</td>
<td>47.4M</td>
</tr>
<tr>
<td>2</td>
<td>40.0M</td>
<td>47.6M</td>
</tr>
<tr>
<td>3</td>
<td>34.5M</td>
<td>39.7M</td>
</tr>
<tr>
<td>4</td>
<td>55.5M</td>
<td>53.6M</td>
</tr>
<tr>
<td>Total</td>
<td>177.7M</td>
<td>188.4M</td>
</tr>
</tbody>
</table>

Table S4. Number of reads from the four independent repeats of RNA-Seq for human liver and muscle tissues that were uniquely mapped to the genome or junction regions.

Fig. S4. The cumulative distribution of the coverage of sequencing reads on exons (left) and genes (right) for 4M (red), 11M (green) and 39M (blue) uniquely mapped reads to exons averaged over the four replicates. 4M and 11M reads are two sub-samplings of 39M. Y-axis represents the percentages of exons or genes detected that have less or equal number of reads than specified by X-axis. The percentages of exons and genes covered by no more than 20 reads are 90% and 56% respectively at 4M total reads, 80% and 46% at 11M total reads, and 60% and 35% at 39M total reads.
6. Evaluation of the array performance

Fig. S5. Reproducibility of array measurements at probe level. A. log2 transformed raw signal of probes targeting several adjacent exons of SLK gene (TC1000560) as visualized using cis-genome browser. The red box indicates an exon (exon 15, chr10: 105,760,564 - 105,760,656) known to be alternatively spliced between liver and muscle tissues. B. Clustering of the raw probe signal shows the negative correlations between probes targeting the alternative exon and most of the probes targeting other exons of the gene. Red bars on the top of the columns indicate probes targeting known splicing exons and junctions, and black bars are the other probes.
7. Reproducibility and dynamic range between the array and RNA-Seq

**Fig. S6.** Comparison of GG-H and RNA-Seq. **A.** Coefficient of variation (CoV) of measured expression levels of exons and genes between four replicates of the same muscle sample. Exons and genes with zero reads in sequencing were excluded in the calculation. The trend of change in CoV versus RPKM in sequencing was estimated by locally weighted scatterplot smoothing (LOWESS). Shown are trend lines of CoV of array (red line) and mRNA sequencing (black line) versus the abundance estimated by RPKM of mRNA sequencing. In 39 million uniquely mapped reads to exons, large variation is observed in sequencing for genes lower than 0.5 RPKM and exons lower than 5 RPKM. **B.** The larger dynamic range of mRNA-Seq comparing with array. Left, average gene expression indices measured by RNA-Seq and GG-H array. Right, fold changes of gene expression between liver and muscle samples measured by RNA-Seq and GG-H array.
8. Analysis of the sources of variations in RNA-Seq

Variance across replicates was analyzed for GG-H and RNA-Seq. Genes or exons with zero reads across all the four repeats in a tissue were excluded as these genes are potentially not expressed in the tissue. Fig. S7A and C shows the median value of coefficients of variation (CoVs) of genes and exons in RNA-Seq that have at least a certain number of reads (solid red curves) and the median value of CoVs in GG-H array with the same set of genes (solid black curves).

The total variance observed from four replicates of sequencing can be approximated as the sum of sample preparation variance, and sampling variance of sequencing. The latter can be approximately estimated as $1/\sqrt{n}$ (average number of reads per replicate) under the assumption of Poisson sampling, which is 0.10 for 100 reads, 0.22 for 20 reads, 0.45 for 5 reads, and 1 for 1 read. Therefore the sample preparation variance can be estimated as the difference between total observed variance and the Poisson sampling variance (dashed red line in Fig. S7A and C). Significantly, on average more than half of the observed variance is estimated to come from sample preparation for genes and exons with more than 4 reads, or roughly a minimum abundance of 0.1 RPKM and 1 RPKM respectively. In addition, while the Poisson sampling variation decreases when the total number of reads increases in an experiment, the variations introduced by the sample preparation steps are unlikely to change.
We set out to estimate the total number of reads required in RNA-Seq to achieve the same overall median CoV as that of the array. For gene expression analysis, the overall median CoVs of the array is 0.062, and in sequencing the subset of ~8,600 genes with a minimum of 185 reads (median of 927 reads) achieves the same median CoV (Fig. S7A and B). To bring the median CoV of all the ~20,000 genes with at least one read (median of 114 reads) to the same level of CoV while keeping the overall expression distribution requires ~290M (8.1 x 36M) reads. A more conservative estimation, which only takes consideration of genes that have a minimum estimated abundance of 0.1 RPKM, or roughly 4 reads per replicate for a gene with a length of 1KB, resulted in ~150M reads. Similarly, 390M reads are estimated for exons (Fig. S7C and D), and conservatively 200M reads for exons with minimum abundance of 1RPKM or roughly 4 reads per replicate for an exon with a length of 100bp).
Fig. S7. Coefficients of variation (CoVs) in GG-H and RNA-Seq. A. Median CoVs of the subset of genes that has more than a minimum number of reads in sequencing. Black solid curve: the median observed CoV (y-axis) in array of the subset of genes above a required minimum number of reads (x-axis), red solid curve: the median observed CoV in sequencing, and red dashed curve: estimated sample preparation CoV in sequencing. A black dot on the red solid curve indicates that a minimum number of 210 reads is necessary in sequencing to achieve the same median CoV as in the array. B. CoV distribution for all genes in array (black solid curve), subset of genes with more than 185 reads in sequencing (red solid curve), and all genes in sequencing (red dashed curve). C. Median CoVs for exons, and a minimum number of 35 reads is required to achieve the same median CoV as in the array. D. CoV distribution for exon expression.
9. Detection of differential expression in RNA-Seq

Since the sequencing platform ‘counts’ a transcript in an RNA sample according to its abundance, the highest abundant species can be sampled hundreds of thousands or millions of times before the low abundant species are sampled. Therefore, to be able to detect the expression of exons and genes that are less abundant, it is critical to have sufficient number of reads in sequencing. Consistent with previously published results (6), from 39M uniquely mapped reads to exons, large variation is observed for genes lower than 0.5 RPKM and exons lower than 5 RPKM (Fig. S6A), or fewer than roughly 20 reads for genes and exons. This is expected, as the Poisson sampling in sequencing alone introduces CoV of 0.22 at 20 reads. Similarly, from 11M and 4M uniquely mapped reads, large variation occurs for genes lower than 1.5 RPKM and 10 RPKM respectively.

Fig. S4 shows the cumulative distribution of the coverage of sequencing reads on genes and exons for 39M uniquely mapped reads to exons averaged over the four replicates, as well as sub-samplings at 4M and 11M reads. Genes or exons with zero reads across all the four repeats were excluded. At 39M, 35% of the genes and 60% of the exons detected in a tissue are covered by fewer than 20 reads. Sequencing with fewer reads will further reduce the coverage, especially on low abundant species. For examples, sequencing of 11M uniquely mapped reads or approximately one lane leads to 46% of the genes and 80% of the exons covered by <= 20 reads, and further reducing the sequencing reads to 4M will lead to 56% and 90% for <=20 reads, and 43% and 74% for <=5 reads.
Similar observations can be made when comparing the detection of differentially expressed genes by array and sequencing with 39M uniquely mapped reads to exons. As shown in Fig. S8, while sequencing and array identify similar number of differentially expressed genes when the gene coverage is high, for 35% of the genes and 60% of the exons detected by less than 20 reads in tissues (roughly 0.5 RPKM for genes and 5 RPKM for exons), sequencing identifies much fewer as statistically significant. Deeper sequencing is required to detect the expression of these genes and exons. Further, at least 46% of the genes and more than 80% of the exons detected in a sample cannot be measured adequately with 11M or fewer uniquely mapped reads.

![Differentially Expressed Genes](image)

**Fig. S8. RNA-Seq detection of differential expressed genes.** Comparison of the number of differentially expressed genes (FDR<0.005) detected by GG-H array (black) and RNA-Seq (gray). While sequencing and array detects similar number of genes when their coverage is above 20 reads, sequencing detects much fewer significant genes below 20 reads.
10. Detection of other contents of human transcriptome

**Fig. S9.** Detection of other contents of human transcriptome. **A.** Boxplots of the expression levels of exons, junctions, ncRNAs, SNPs and UTUs. **B.** Hierarchical clustering of functional ncRNAs differentially expressed between liver and muscle tissues. C/D box snoRNA family (each indicated by a black bar next to the dendrogram) is over-expressed in muscle, and H/ACA box snoRNA (red bar) is over expressed in liver.

11. De novo identification of junctions from mRNA sequencing data and comparison with the array design.

**Fig. S10.** De novo identification of junctions from mRNA sequencing data compared with the array design. The x-axis is the log2 number of reads supporting a junction and the y-axis shows the number of junctions supported. The black bars indicate the number of junctions in the array design and the light gray bars indicate the number of new junctions. While GG-H junctions cover well highly-expressed junctions identified by the de novo method, a total of 6,581 additional de novo junctions supported by more than four reads (~1RPKM) were discovered and will be included in the next revision of the array.
Acknowledgements

The participating investigators of the Glue Grant Program include Lily Altstein, Ph.D., Henry V. Baker, Ph.D., Ulysses G.J. Balis, M.D., Paul E. Bankey, M.D., Ph.D., Timothy R. Billiar, M.D., Bernard H. Brownstein, Ph.D., Steven E. Calvano, Ph.D., David G. Camp II, Ph.D., J. Perren Cobb, M.D., Joseph Cuschieri, M.D., Asit K. De, Ph.D., Celeste C. Finnerty, Ph.D., Richard L. Gamelli, M.D., Hong Gao, Ph.D., Nicole S. Gibran, M.D., Brian G. Harbrecht, M.D., Douglas L. Hayden, M.A., Laura Hennessy, R.N., David N. Herndon, M.D., Shari E. Honari, R.N., Marc G. Jeschke, M.D., Ph.D., Jeffrey L. Johnson, M.D., Matthew B. Klein, M.D., Stephen F. Lowry, M.D., Philip H. Mason, Ph.D., Grace P. McDonald-Smith, M.Ed., Bruce A. McKinley, Ph.D., Carol L. Miller-Graziano, Ph.D., Joseph P. Minei, M.D., Ernest E. Moore, M.D., Frederick A. Moore, M.D., Avery B. Nathens, M.D., Ph.D., M.P.H., Grant E. O'Keefe, M.D., M.P.H., Laurence G. Rahme, Ph.D., Daniel G. Remick, M.D., David A. Schoenfeld, Ph.D., Michael B. Shapiro, M.D., Richard D. Smith, Ph.D., Jason Sperry, M.D., Robert Tibshirani, Ph.D., Mehmet Toner, Ph.D., H. Shaw Warren, M.D., Michael A. West, M.D., PhD., Bram Wispelwey, M.S.

The GG-H array can be ordered from Affymetrix as a custom array. For more information, please contact the authors.

References