

Human transcriptome array for high-throughput clinical studies

Weihong Xu^{a,1}, Junhee Seok^{a,1}, Michael N. Mindrinos^{a,1}, Anthony C. Schweitzer^b, Hui Jiang^c, Julie Wilhelmy^a, Tyson A. Clark^b, Karen Kapur^c, Yi Xing^{c,d}, Malek Faham^b, John D. Storey^e, Lyle L. Moldawer^f, Ronald V. Maier^g, Ronald G. Tompkins^h, Wing Hung Wong^c, Ronald W. Davis^{a,2}, Wenzhong Xiao^{h,2}, and the Inflammation and Host Response to Injury Large-Scale Collaborative Research Program³

^aStanford Genome Technology Center and Department of Biochemistry, Stanford University, Palo Alto, CA 94304; ^bAssay and Application Product Development, Affymetrix, Inc., Santa Clara, CA 95051; ^cDepartment of Statistics and Department of Health Research and Policy, Stanford University, Stanford, CA 94305; ^dDepartment of Internal Medicine and Department of Biomedical Engineering, University of Iowa, Iowa City, IA 52242; ^eLewis-Sigler Institute and Department of Molecular Biology, Princeton University, Princeton, NJ 08544; ^fDepartment of Surgery, University of Florida College of Medicine, Gainesville, FL 32610; ^gDepartment of Surgery, University of Washington, Seattle, WA 98104; and ^hMassachusetts 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.

gene isoform | next-generation sequencing | exon junction | blood leukocyte

Developments 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.

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 fine-needle biopsy, enriched cell populations, or laser capture microdissection, where the RNA yield is typically as low as 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-

Author contributions: M.N.M., J.D.S., R.V.M., R.G.T., W.H.W., R.W.D., W. Xiao, and Inflammation and Host Response to Injury Large-Scale Collaborative Research Program designed research; W. Xu, J.S., M.N.M., J.W., and W. Xiao performed research; A.C.S., H.J., T.A.C., K.K., Y.X., M.F., and W. Xiao contributed new reagents/analytic tools; W. Xu, J.S., H.J., W.H.W., and W. Xiao analyzed data; and W. Xu, J.S., L.L.M., W.H.W., R.W.D., and W. Xiao wrote the paper.

The authors declare no conflict of interest.

Freely 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](https://doi.org/10.1101/11319), [GSE26072](https://doi.org/10.1101/26072), and [GSE26109](https://doi.org/10.1101/26109)).

¹W. Xu, J.S., and M.N.M. contributed equally to this work.

²To whom correspondence may be addressed. E-mail: dbowe@stanford.edu or wxiao1@partners.org.

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

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1019753108/-DCSupplemental.

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. 1*A* 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 S1*A*) 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 S1*B*) 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 eval-

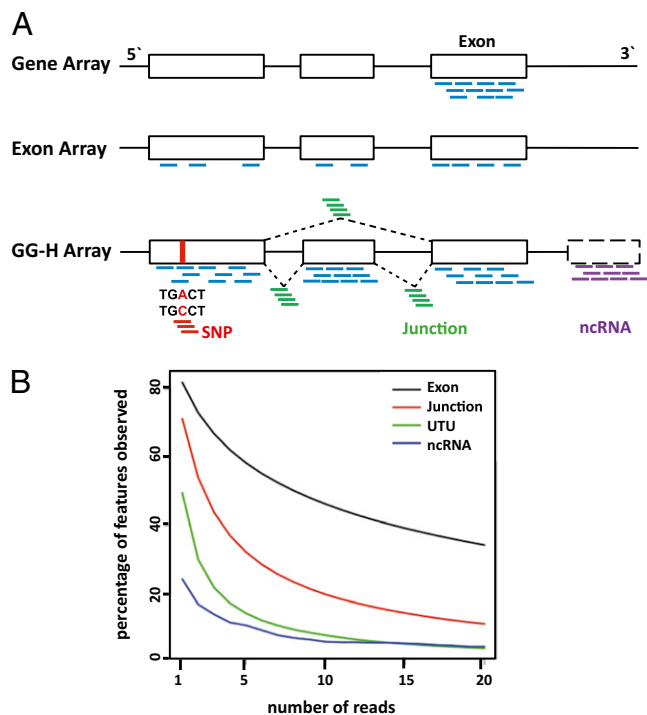


Fig. 1. Design of the GG-H human transcriptome array for comprehensive examination of gene and exon expression, alternative splicing, and additional contents of human transcriptome. (A) On a 3' gene array, such as the Affymetrix HU-133 Array, 11 probes were designed for the 3'-end exon(s) of each gene. On an exon array, such as the Affymetrix Human Exon 1.0 ST Array, there are two to four probes for each exon of the gene. In contrast, the GG-H array uses, on average, 10 probes for each exon/PSR and 4 probes for each exon–exon junction; in addition, 6 pairs of probes were designed for each coding SNP, and 10 probes were designed for each noncoding RNA transcript. (B) Comparison of the GG-H array contents with mRNA sequencing data on multiple tissues. The percentages of exons, junctions, UTUs, and ncRNAs (y axis) supported by at least a specified number of sequencing reads (x axis) are shown.

uate 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. 1*B*). 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

Array components	No. of targets	No. of probes
Gene exons	315,137	3,292,929
Exon–exon junctions	260,488	1,060,703
Coding SNPs and DMET variations	89,782	982,941
Noncoding functional RNA (f-ncRNA)	730	5,869
Noncoding antisense expression (as-ncRNA)	50,783	563,097
Unannotated transcribed units (UTUs)	49,957	488,581
Other probes including controls		498,840
Total		6,892,960

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 μ g same liver and muscle samples were used to perform four independent repeats 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, whereas 65% and 40%, respectively, have 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 1 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 HU133 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

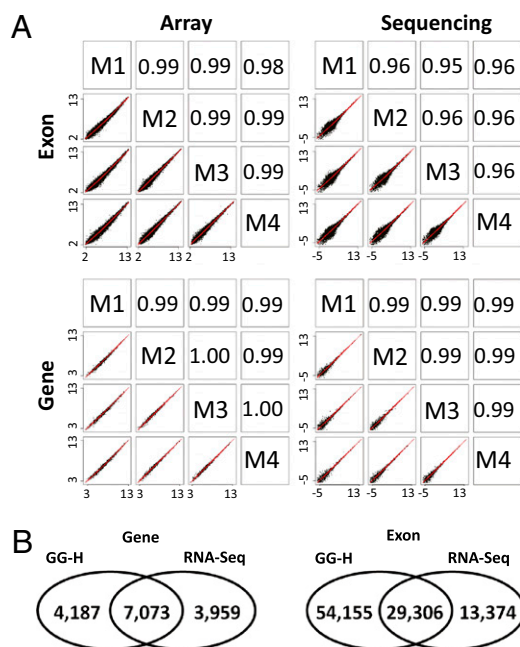


Fig. 2. Comparison of the performances of the GG-H array and mRNA sequencing. (A) Reproducibility of exon and gene expression measured by array and sequencing. In each panel, scatter plots of logged expression values between four independent replicates of a reference muscle sample are shown in the bottom 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.

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 junction reveal the alternative splicing to the isoform that skips exon 15.

By combining such information on exons and corresponding junctions, more than 6,600 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

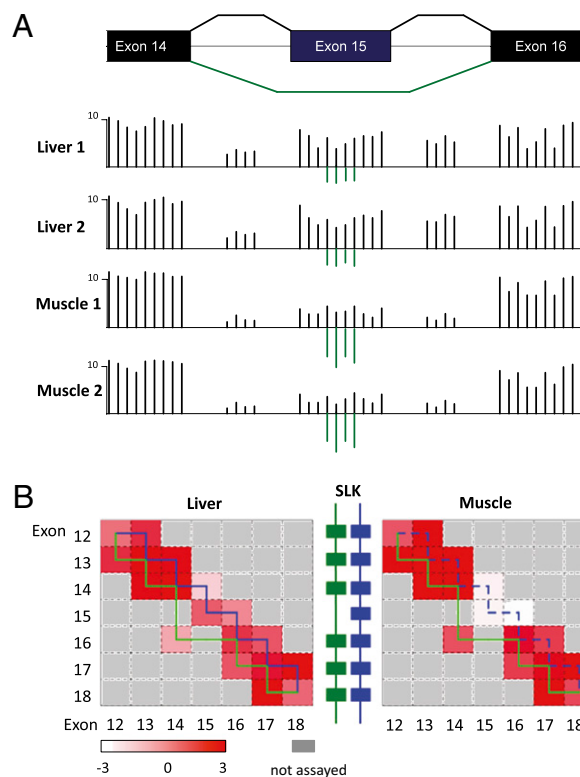


Fig. 3. Detection of alternative splicing events using exon and junction probes on the array. Two isoforms of *SLK* are alternatively spliced between liver and muscle; the green lines represent an isoform skipping exon 15 (ENST00000335753), and the blue lines represent another isoform including exon 15 (ENST00000369755). (A) Changes of the signal of junction and exon probes. (B) Changes of the calculated expression of exons and junctions. Exon expression is shown at the diagonal, and junction expression is shown off the diagonal of two connecting exons.

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 7SKs, spliceosomal RNA U2, and several small nucleolar RNAs (snoRNAs) 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 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 from 2 μ g 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 thou-

sands 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 remainders 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

	GG-H array	RNA-Seq (GA II)	RNA-Seq (HiSeq 2000)
Gene			
Throughput (samples per week)	200	2	30
Processing time for 5,000 samples (wk)	25	2,500	167
Exon			
Throughput (samples per week)	200	1	15
Processing time for 5,000 samples (wk)	25	5,000	333

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.

sponse to severe injury in blood leukocytes, enriched monocytes, T cells, and neutrophils (29, 30). Using macroscale and microfluidic 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 and 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 F5450 and GeneChip Scanner 3000 7G.

Sample preparation for Illumina sequencing. Library preparation and sequencing were performed as described (32). Briefly, starting with 2 μ g total human liver and muscle RNA (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 v2 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 Multiarray 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 RNAs, 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 abundances 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 Balis for assistance with the sequencing experiments. This work was supported in part by National Institutes of Health Grants U54GM062119, P01HG000205, and R01HG004634.

- Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270:467–470.
- Chee M, et al. (1996) Accessing genetic information with high-density DNA arrays. *Science* 274:610–614.
- Heller RA, et al. (1997) Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. *Proc Natl Acad Sci USA* 94:2150–2155.
- Golub TR, et al. (1999) Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* 286:531–537.
- Alizadeh AA, et al. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 403:503–511.
- Bittner M, et al. (2000) Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* 406:536–540.
- Perou CM, et al. (2000) Molecular portraits of human breast tumours. *Nature* 406:747–752.
- ENCODE Project Consortium, et al. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447:799–816.
- Wang ET, et al. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature* 456:470–476.
- Pan Q, et al. (2004) Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform. *Mol Cell* 16:929–941.
- Faustino NA, Cooper TA (2003) Pre-mRNA splicing and human disease. *Genes Dev* 17:419–437.
- García-Blanco MA, Baraniak AP, Lasda EL (2004) Alternative splicing in disease and therapy. *Nat Biotechnol* 22:535–546.
- Kim E, Goren A, Ast G (2008) Insights into the connection between cancer and alternative splicing. *Trends Genet* 24:7–10.
- Blencowe BJ (2006) Alternative splicing: New insights from global analyses. *Cell* 126:37–47.
- Kapur K, Xing Y, Ouyang Z, Wong WH (2007) Exon arrays provide accurate assessments of gene expression. *Genome Biol* 8:R82.
- Robinson MD, Speed TP (2007) A comparison of Affymetrix gene expression arrays. *BMC Bioinformatics* 8:449.
- Clark TA, Sugnet CW, Ares M, Jr. (2002) Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays. *Science* 296:907–910.
- Black DL (2000) Protein diversity from alternative splicing: A challenge for bioinformatics and post-genome biology. *Cell* 103:367–370.
- Johnson JM, et al. (2003) Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science* 302:2141–2144.
- Ule J, et al. (2005) Nova regulates brain-specific splicing to shape the synapse. *Nat Genet* 37:844–852.
- Castle JC, et al. (2008) Expression of 24,426 human alternative splicing events and predicted cis regulation in 48 tissues and cell lines. *Nat Genet* 40:1416–1425.
- Gingeras TR (2007) Origin of phenotypes: Genes and transcripts. *Genome Res* 17:682–690.
- Kapranov P, Willingham AT, Gingeras TR (2007) Genome-wide transcription and the implications for genomic organization. *Nat Rev Genet* 8:413–423.
- Kapranov P, et al. (2007) RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* 316:1484–1488.
- Clark TA, et al. (2007) Discovery of tissue-specific exons using comprehensive human exon microarrays. *Genome Biol* 8:R64.
- Irizarry RA, et al. (2003) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res* 31:e15.
- Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA* 98:5116–5121.
- Au KF, Jiang H, Lin L, Xing Y, Wong WH (2010) Detection of splice junctions from paired-end RNA-seq data by SpliceMap. *Nucleic Acids Res* 38:4570–4578.
- Kotz KT, et al. (2010) Clinical microfluidics for neutrophil genomics and proteomics. *Nat Med* 16:1042–1047.
- Laudanski K, et al. (2006) Cell-specific expression and pathway analyses reveal alterations in trauma-related human T cell and monocyte pathways. *Proc Natl Acad Sci USA* 103:15564–15569.
- Tumor Analysis Best Practices Working Group (2004) Expression profiling—best practices for data generation and interpretation in clinical trials. *Nat Rev Genet* 5:229–237.
- Illumina I (2009) mRNA Sequencing: Sample Preparation Guide (Illumina, San Diego), Part 1004898, Rev D.
- Jiang H, Wong WH (2008) SeqMap: Mapping massive amounts of oligonucleotides to the genome. *Bioinformatics* 24:2395–2396.
- Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 5:621–628.