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/.

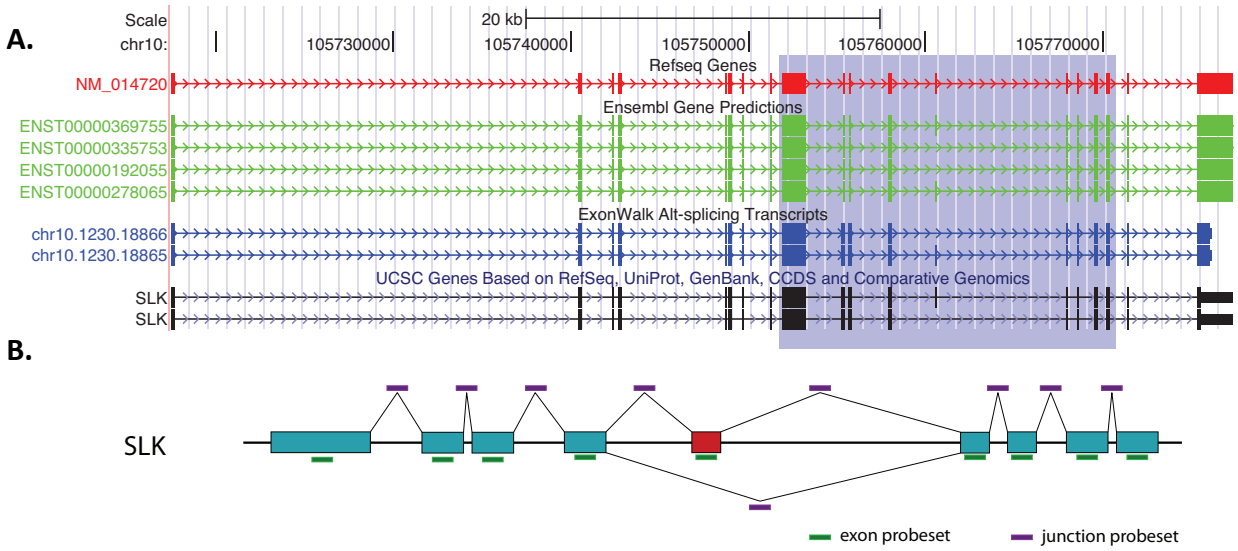

Fig. S1. An example of the GG-H array annotation of genes (transcript clusters). **A.** Annotations of transcripts, exons, and junctions of SLK gene from RefSeq, Ensembl, UCSC Known Genes, Exonwalk and other database sources. **B.** Design of exon and junction probesets targeting the gene SLK. Shown 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.

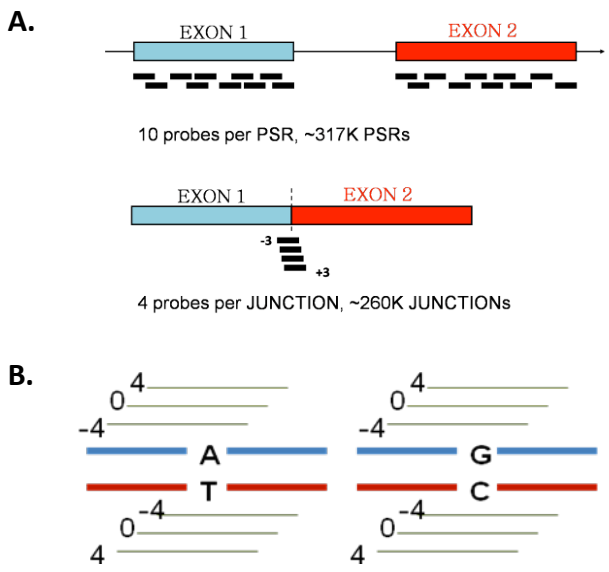

Fig. S2. Design of Exon and junction probes (A) and SNP (B)

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

Recognition Particle Database (http://psyche.uthct.edu/dbs/SRPDB/), Noncoding regulatory RNA database (http://biobases.ibch.poznan.pl/ncRNA/), NONCODE (http://www.bioinfo.org.cn/noncode/), RNAdb (http://jsm-research.imb.uq.edu.au/rnadb/), Rfam (http://www.sanger.ac.uk/Software/Rfam/), fRNAdb (http://www.ncrna.org/), H-Invitational database (http://www.h-invitational.jp/) as well as primary literature._730 curated ncRNA species (f-ncRNA) were identified with experimental evidence to support their biological functions (excluding rRNA and tRNA); these diverse functions include chromatin architecture/epigenetic memory, DNA replication, transcription, RNA splicing, editing, translation, protein transport and turnover, stress induced, drug resistance, and disease related (Table S1A). Ten probes were designed for each of these ncRNAs.

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

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

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

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

Min Length	0	25	100*
# bases	36Mb	36Mb	8.5Mb

*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

Table S2. Various control probes included in the GG-H array.

Content	Annotation	# Probes
Affymetrix polyA spike-ins	Affymetrix Eukaryotic PolyA RNA Control Kit - 4 species	1182
Affymetrix default controls	Various controls for manufacturing and scanning	16570
Affymetrix antigenomic 25 mers	GC bin background probes	16943
Affymetrix human non-transcribed sequences ("big GC")	Additional GC bin background probes	20282
Affymetrix "norm gene" sequences	3012 probes targeting exons and 10078 targeting introns	13090
ERCC probes	Targeting 140 species of ERCC	22358
Antigenomic 22 mers	GC bin background probes	8350
Mismatches to rRNA	0-4bp mismatches to rRNA	3182
Mismatches to PolyA spike-Ins	0-4bp mismatch and ins/del to Affymetrix polyA spike-ins	16314
Total		118271

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

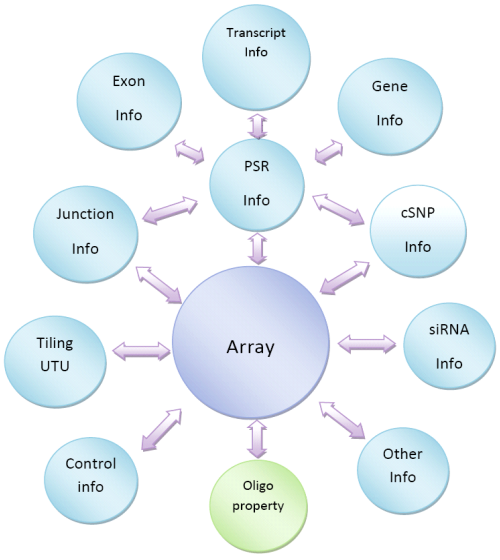

Fig. S3. Summary of the schema of the array annotation database.

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.

Tissue	PSR	JUNC	UTU	as-ncRNA	f-ncRNA	Total
Adipose	16.6M	1.4M	0.2M	77.0K	7.4K	18.2M
Brain	10.7M	0.6M	0.1M	29.4K	5.6K	11.5M
Breast	8.5M	0.8M	0.1M	46.9K	6.8K	9.4M
Colon	17.2M	1.3M	0.2M	72.4K	6.7K	18.8M
Heart	12.0M	0.7M	0.3M	30.4K	11K	13.0M
Liver	10.7M	0.9M	88.9K	42.6K	3.5K	11.8M
Lymph	13.0M	1.5M	0.2M	0.2M	6.3K	14.9M
Muscle	13.9M	1.3M	0.2M	51.8K	9.0K	15.4M
Testes	16.1M	1.6M	0.1M	85.2K	6.8K	18.0M
UHR (low cov.)	4.1M	0.5M	70.8K	29.1K	3.9K	4.6M
SUM	122.8M	10.7M	1.4M	0.7M	67.1K	136M

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 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.

Run	Liver	Muscle
1	47.7M	47.4M
2	40.0M	47.6M
3	34.5M	39.7M
4	55.5M	53.6M
Total	177.7M	188.4M

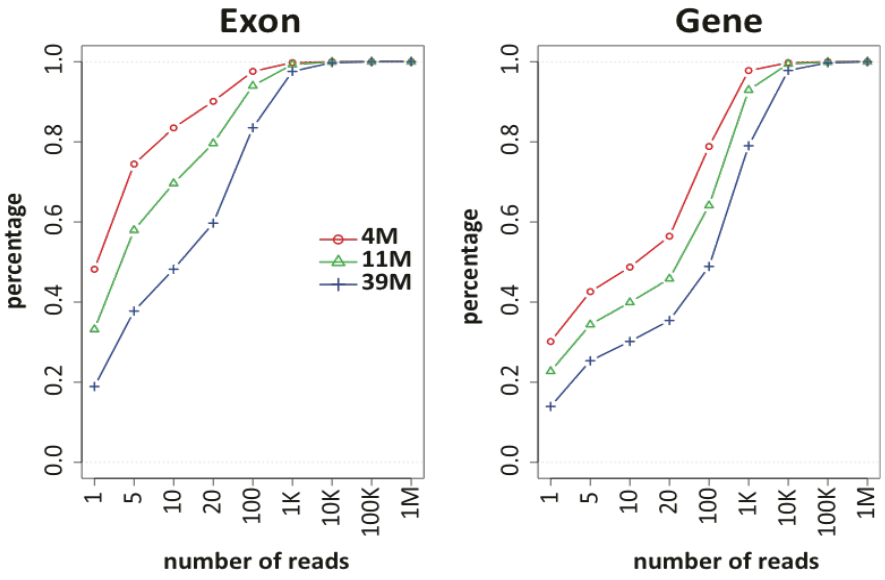

Fig. S4. The cumulative distribution of the coverage of sequencing reads on exons (left) and genes (right) for 4M (red), 11M (green) and 39M (blue) uniquely mapped reads to exons averaged over the four replicates. 4M and 11M reads are two sub-samplings of 39M. Y-axis represents the percentages of exons 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

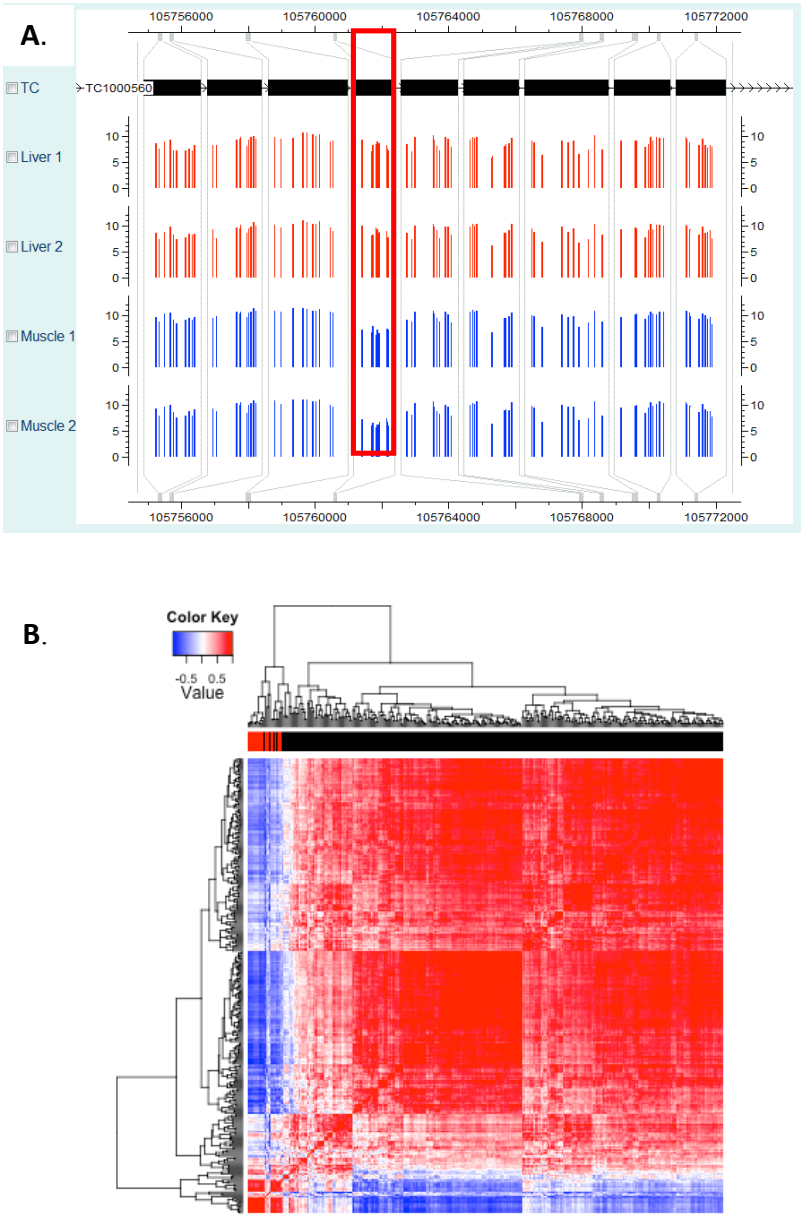

Fig. S5. Reproducibility of array measurements at probe level. **A**. log2 transformed raw signal of probes targeting several adjacent exons of SLK gene (TC1000560) as visualized using cis-genome browser. The red box indicates an exon (exon 15, chr10: 105,760,564 - 105,760,656) known to be alternatively 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

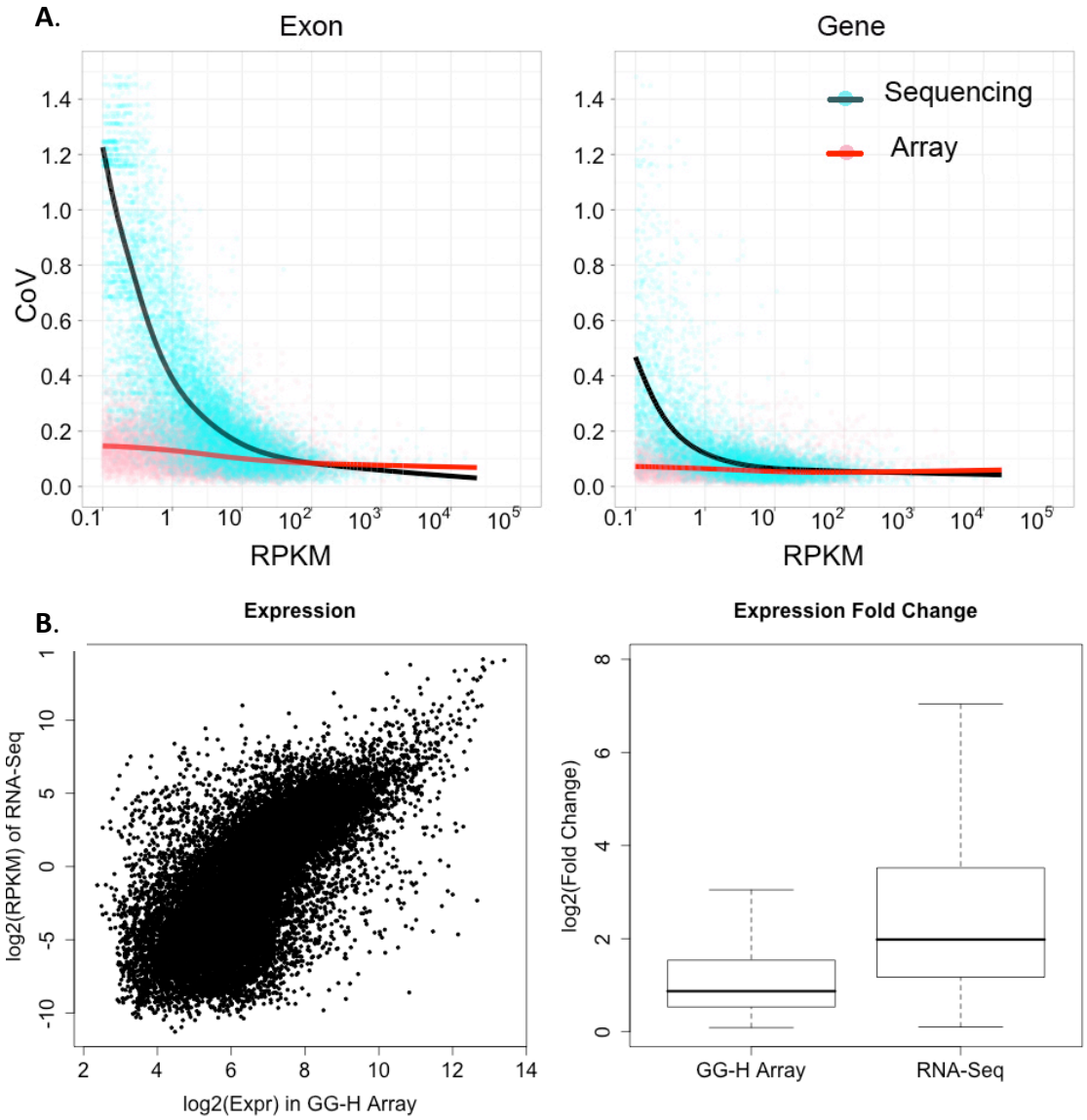

Fig. S6. Comparison of GG-H and RNA-Seq. **A.** Coefficient of variation (CoV) of measured expression levels of exons and genes between four replicates of the same muscle sample. Exons and genes with zero reads in sequencing were excluded in the calculation. The trend of change in CoV versus RPKM in 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 (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).

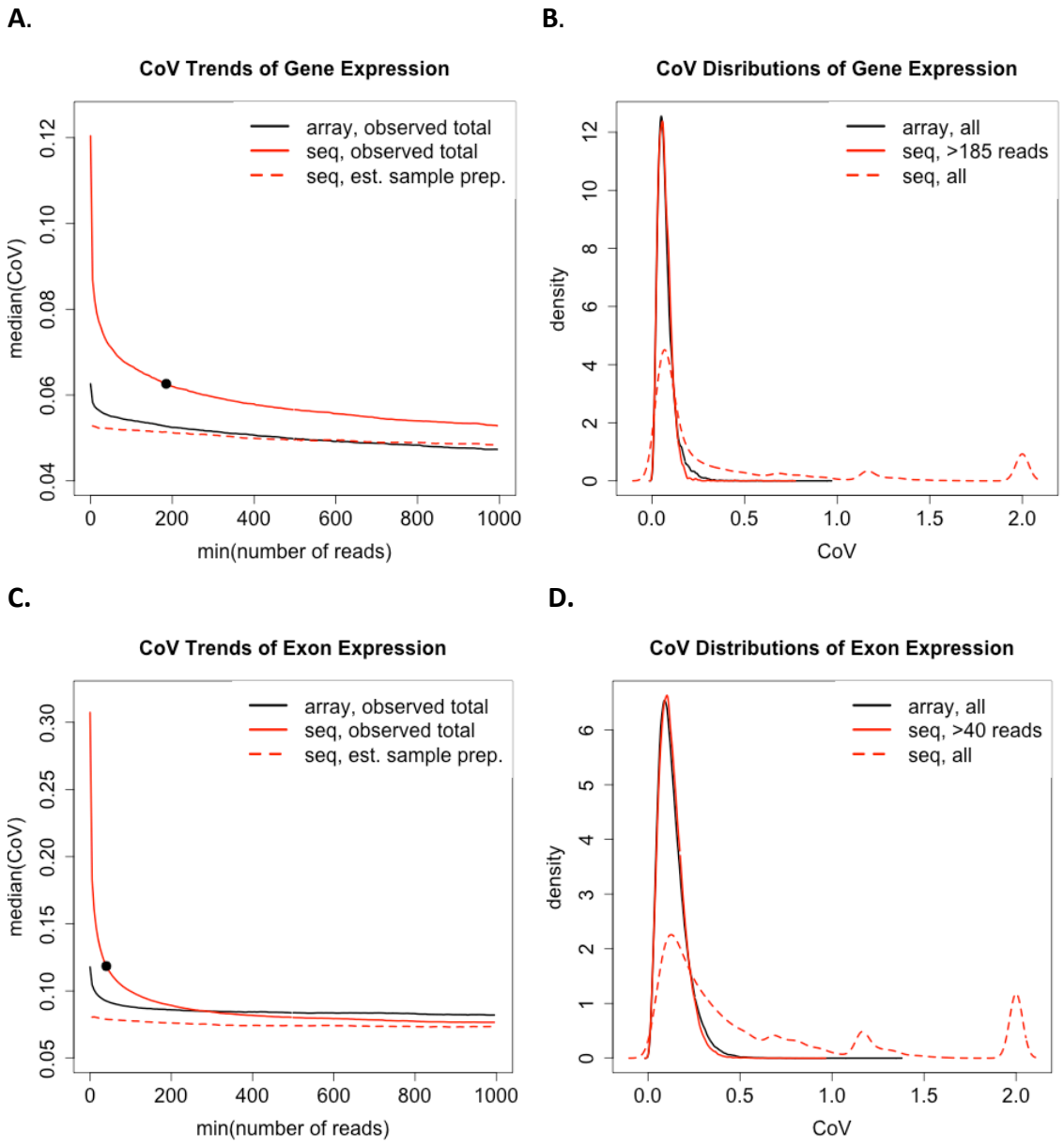

Fig. S7. Coefficients of variation (CoVs) in GG-H and RNA-Seq. **A**. Median CoVs of the subset of genes that has more than a minimum number of reads in sequencing. Black solid curve: the median observed CoV (y-axis) in array of the subset of genes above a required minimum number of reads (x-axis), red 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.

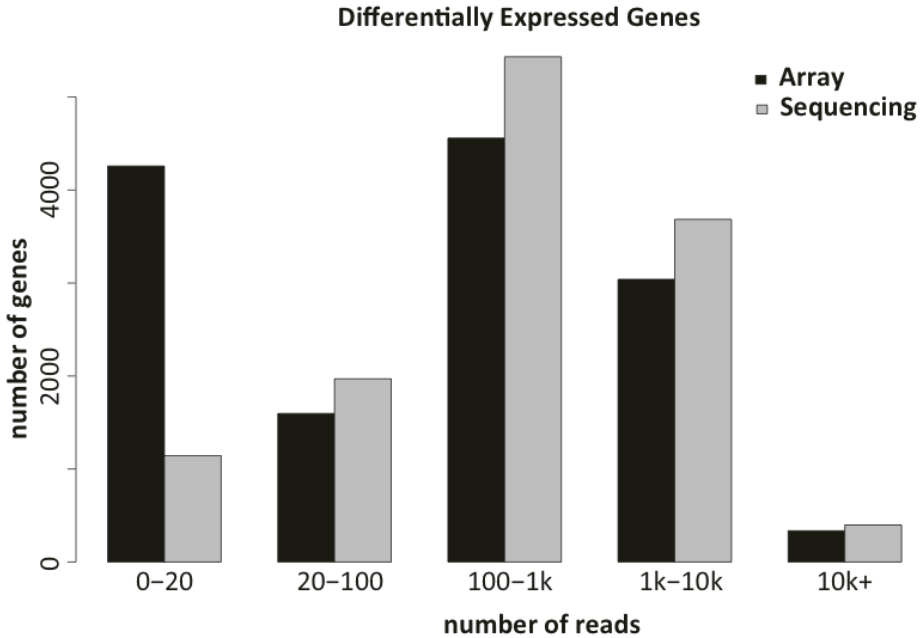

Fig. S8. RNA-Seq detection of differential expressed genes. Comparison of the number of differentially expressed genes (FDR<0.005) detected by GG-H array (black) and RNA-Seq (gray). While sequencing and array detects similar number of genes when their coverage is above 20 reads, sequencing detects much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

10. Detection of other contents of human transcriptome

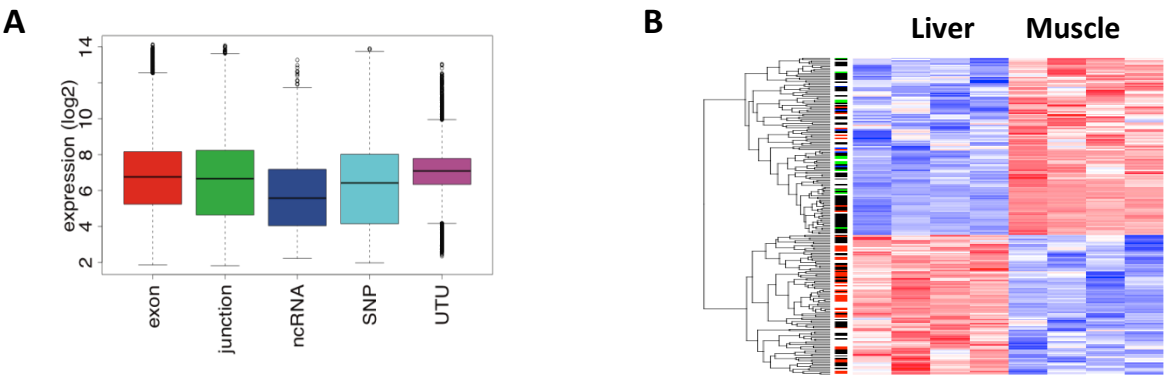

Fig. S9. Detection of other contents of human transcriptome. **A**. Boxplots of the expression levels of exons, junctions, ncRNAs, SNPs and UTUs. **B**. Hierarchical clustering of functional ncRNAs differentially expressed between liver and muscle tissues. C/D box snoRNA family (each indicated by a black 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.

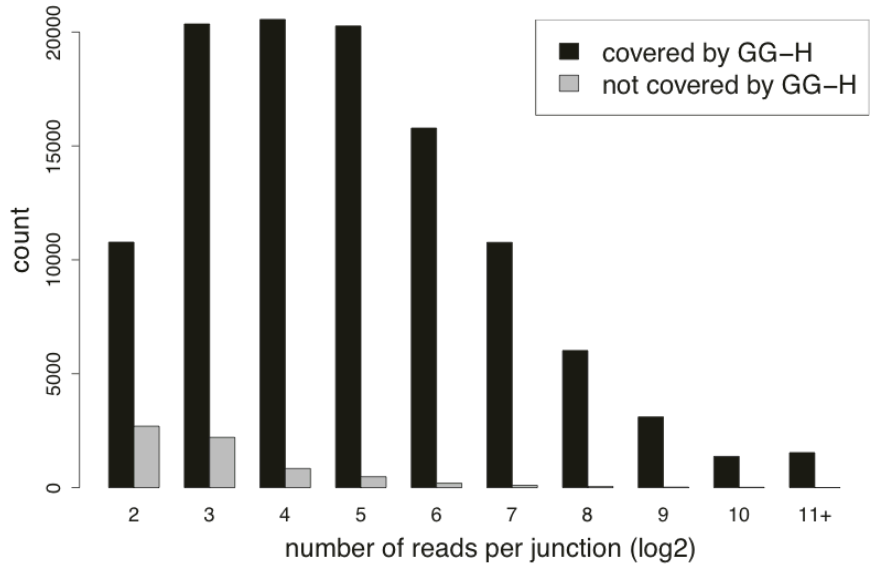

Fig. S10. *De novo* identification of junctions from mRNA sequencing data compared with the array design. The x-axis is the log2 number of reads supporting a junction and the y-axis shows the number of junctions supported. The black bars indicate the number of junctions in the array design and the light 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

1. Mei R*, et al.* (2003) Probe selection for high-density oligonucleotide arrays. *Proc Natl Acad Sci U S A* 100(20):11237-11242.
2. Shoemaker DD, Lashkari DA, Morris D, Mittmann M, & Davis RW (1996) Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy. *Nat Genet* 14(4):450-456.
3. Kapranov P*, et al.* (2007) RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* 316(5830):1484-1488.
4. Tumor Analysis Best Practices Working Group (2004) Expression profiling--best practices for data generation and interpretation in clinical trials. *Nat Rev Genet* 5(3):229-237.
5. Baker SC*, et al.* (2005) The External RNA Controls Consortium: a progress report. *Nat Methods* 2(10):731-734.
6. Mortazavi A, Williams BA, McCue K, Schaeffer L, & Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 5(7):621-628.