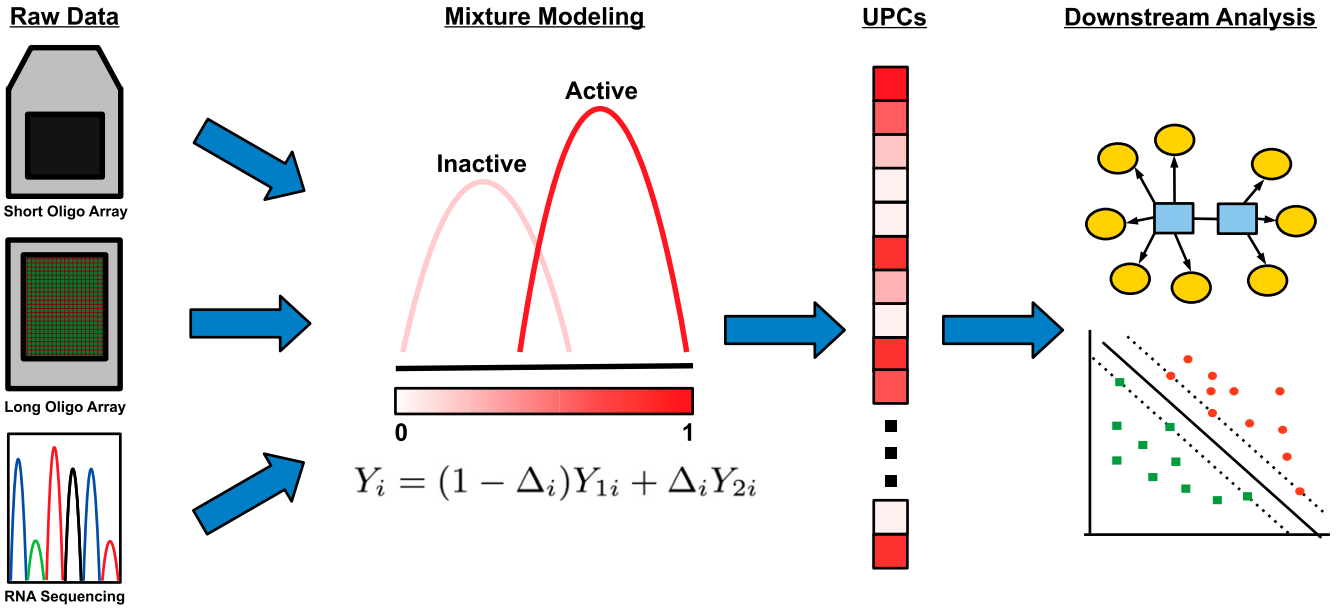
Supporting Information

Piccolo et al. 10.1073/pnas.1305823110



A. Adjust for platform-specific effects due to probe/gene composition, dye effects, transcript length.

B. Use mixture modeling to estimate probability of expression above background.

C. Generate platform-agnostic, probabilistic barcode values.

D. Develop downstream analysis tools that can be applied in a platform-agnostic manner.

Fig. S1. Universal exPression Codes (UPCs) characterize transcriptional activation for microarray and RNA-sequencing (RNA-Seq) platforms. This diagram depicts the UPC method's ability to process genomic data acquired through various high-throughput profiling techniques. Raw values are adjusted for genomic base composition, dye effects, and/or transcript length, and then values representing transcriptional activation are produced for each input platform. These values can then be used in downstream analytical tools uniformly, irrespective of the input platform used.

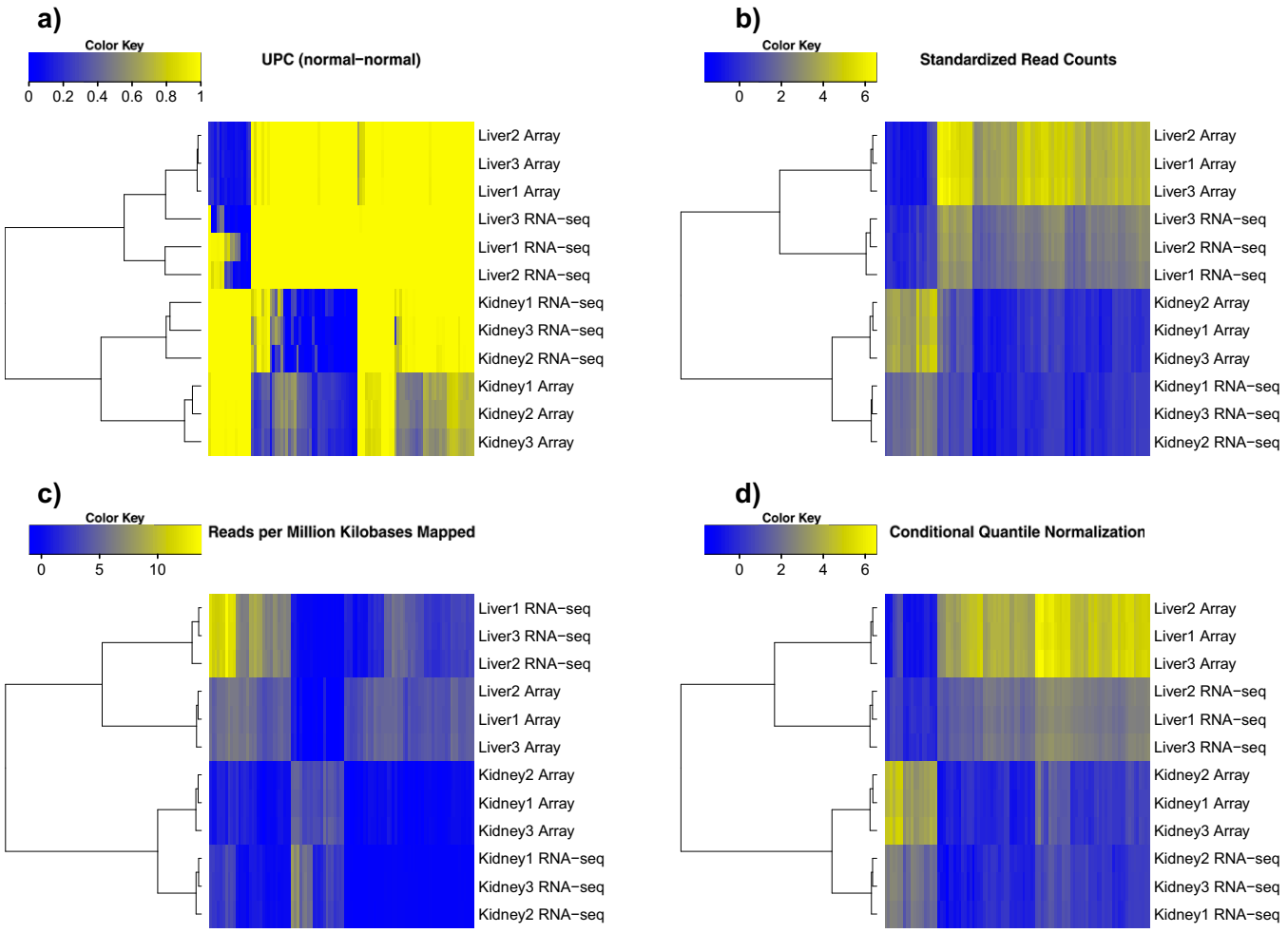


Fig. S2. Heat maps of differentially expressed genes for kidney and liver data (aligned using TopHat). For each normalization method, values were calculated for microarrays and RNA-Seq, and the 100 genes that exhibited the largest absolute fold change between tissue types in the microarray data were plotted. In *A*, UPC values were used for both platforms. In *B*, single-channel array normalization (SCAN) values were used for microarrays, and log-transformed, *z*-score standardized read counts were used for RNA-Seq. In *C* and *D*, SCAN values were used for microarrays, and reads per million kilobases mapped (RPKM) or conditional quantile normalization (CQN) values were used, respectively, for RNA-Seq; the latter were also log-transformed (before normalization) and *z*-score standardized (after normalization). For each normalization method, the samples cluster correctly by tissue type. However, even after *z*-score standardization, value ranges for the competing methods varied substantially by technology type.

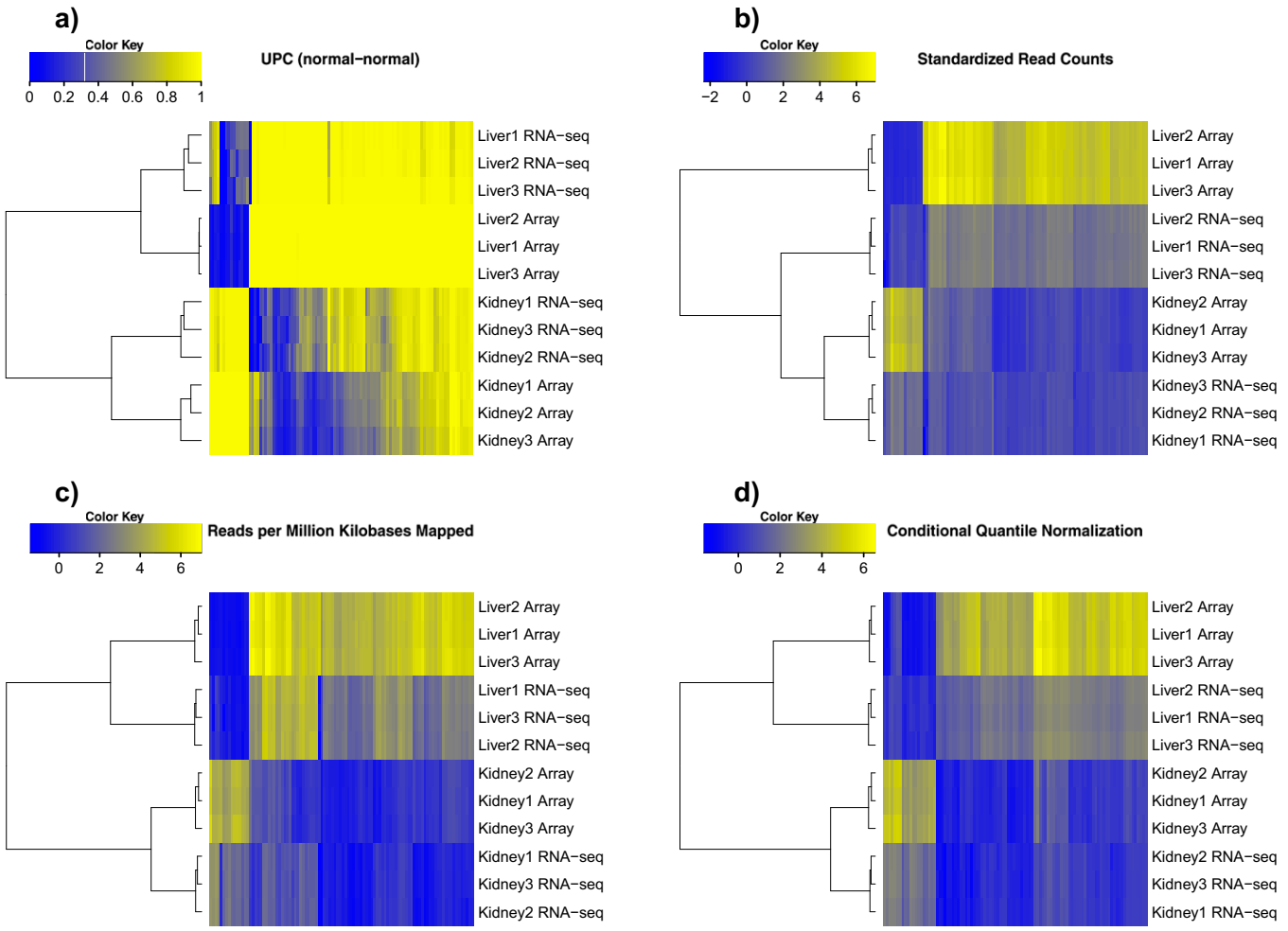


Fig. S3. Heat maps of differentially expressed genes for kidney and liver data, aligned using Genomic Next-generation Universal MAPper (GNUMAP). For each normalization method, values were calculated for microarrays and RNA-Seq, and the 100 genes that exhibited the largest absolute fold change between tissue types in the microarray data were plotted. In *A*, UPC values were used for both platforms. In *B*, single-channel array normalization (SCAN) values were used for microarrays, and log-transformed, z-score standardized read counts were used for RNA-Seq. In *C* and *D*, SCAN values were used for microarrays, and reads per million kilobases mapped (RPKM) or conditional quantile normalization (CQN) values were used, respectively, for RNA-Seq; the latter were also log-transformed (before normalization) and z-score standardized (after normalization).

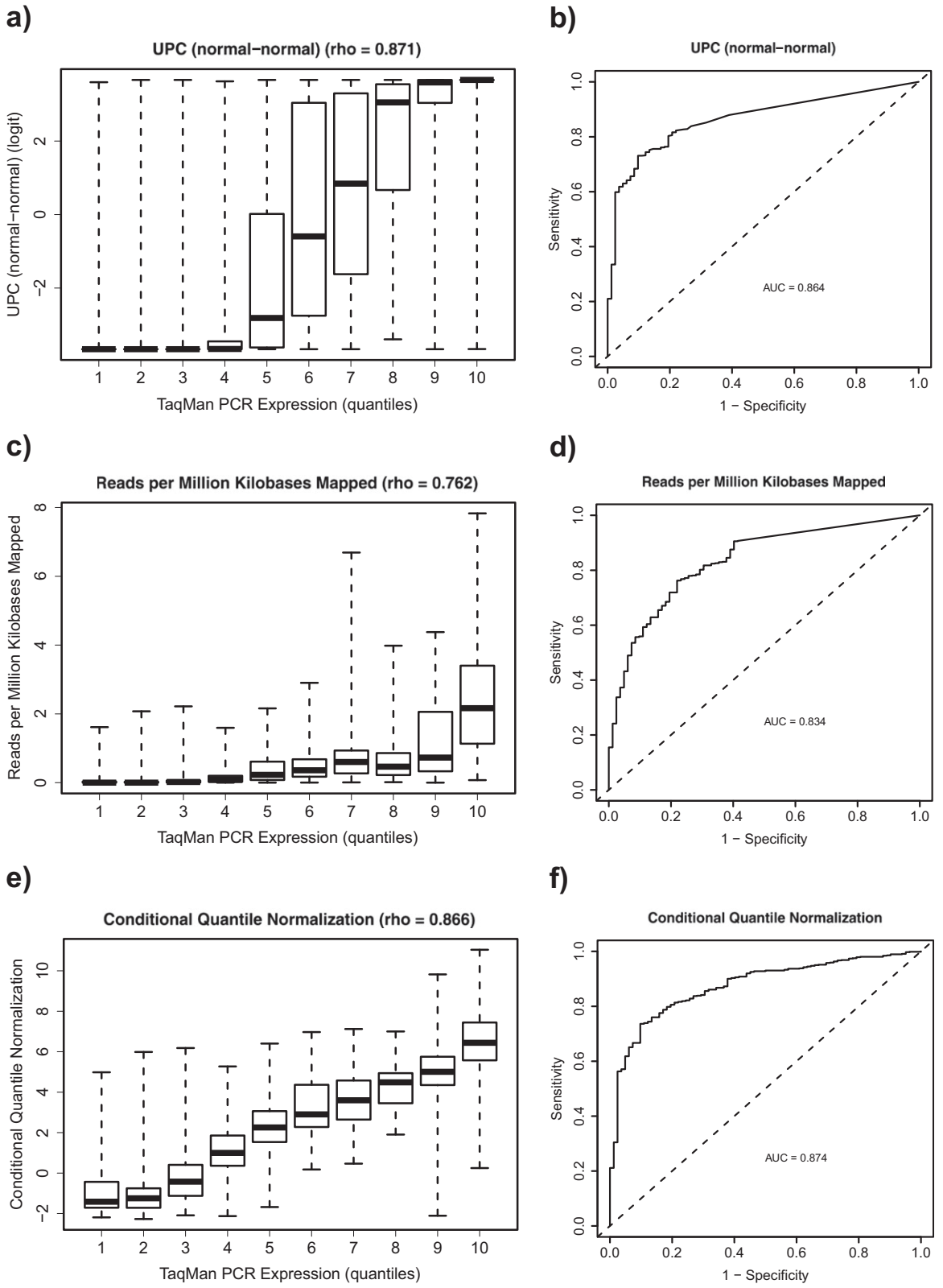


Fig. S4. Comparison of UPC, RPKN, and CQN (RNA-Seq) values against quantitative PCR (qPCR) data. Data came from the Microarray Quality Control (MAQC) project (1). qPCR values were compared against normalized RNA-Seq values using Spearman's rank correlation coefficient. Normalized RNA-Seq values were also compared against absent/present calls to assess whether genes called as present tended to have higher UPC, RPKN, and CQN values than genes called as absent. *A*, *C*, and *E* show that UPC- and CQN-normalized values correlated more strongly with qPCR values than RPKM values. *B*, *D*, and *F* show that UPC and CQN values tended to differentiate better between absent and present calls than RPKM.

1. Shi L, et al. (2006) The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol* 24(9):1151–1161.

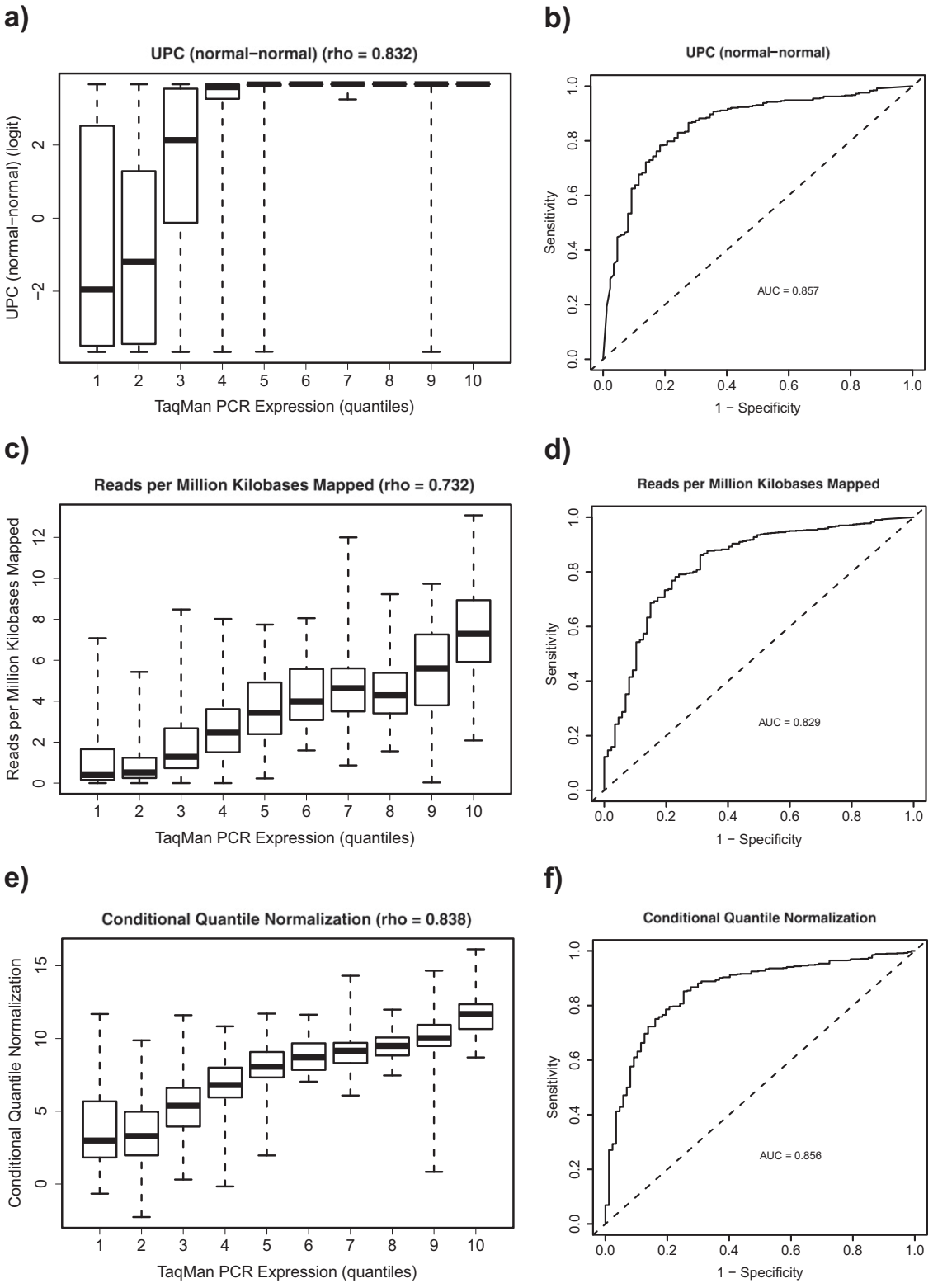


Fig. S5. Comparison of UPC, RPKN, and CQN (RNA-Seq) values against qPCR data using the GNUMAP read aligner. *A*, *C*, and *E* show that UPC- and CQN-normalized values correlated more strongly with qPCR values than RPKM values. *B*, *D*, and *F* show that UPC and CQN values tended to differentiate better between absent and present calls than RPKM.

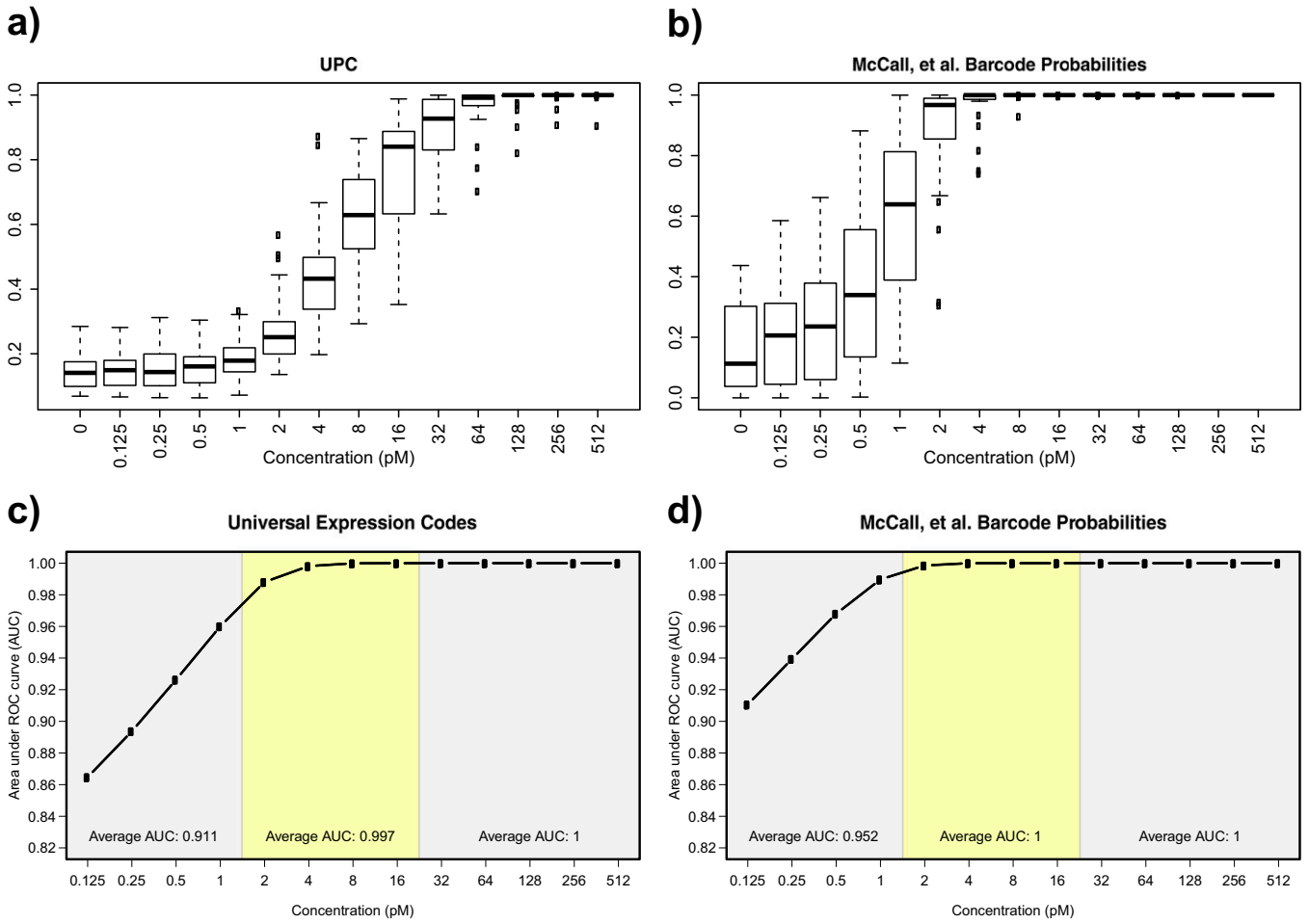


Fig. S6. Comparisons between UPC and prior barcode approach on microarray spike-in data. (*A* and *B*) For the Affymetrix U133A Latin Square data, transcriptional activation was estimated using the UPC method and the prior barcode method (1, 2) for the various spike-in concentrations. As spike-in levels increase, so do UPC and barcode values. In *C* and *D*, we evaluated the methods' sensitivity for detecting any level of transcription. We tested whether expression values for spike-in concentrations greater than zero were higher than values derived from the zero-concentration samples. For each concentration threshold, we considered only nonzero concentrations that were greater than the given threshold.

1. Zilliox MJ, Irizarry RA (2007) A gene expression bar code for microarray data. *Nat Methods* 4(11):911–913.
2. McCall MN, Uppal K, Jaffee HA, Zilliox MJ, Irizarry RA (2011) The Gene Expression Barcode: Leveraging public data repositories to begin cataloging the human and murine transcriptomes. *Nucleic Acids Res* 39(Database issue):D1011–D1015.

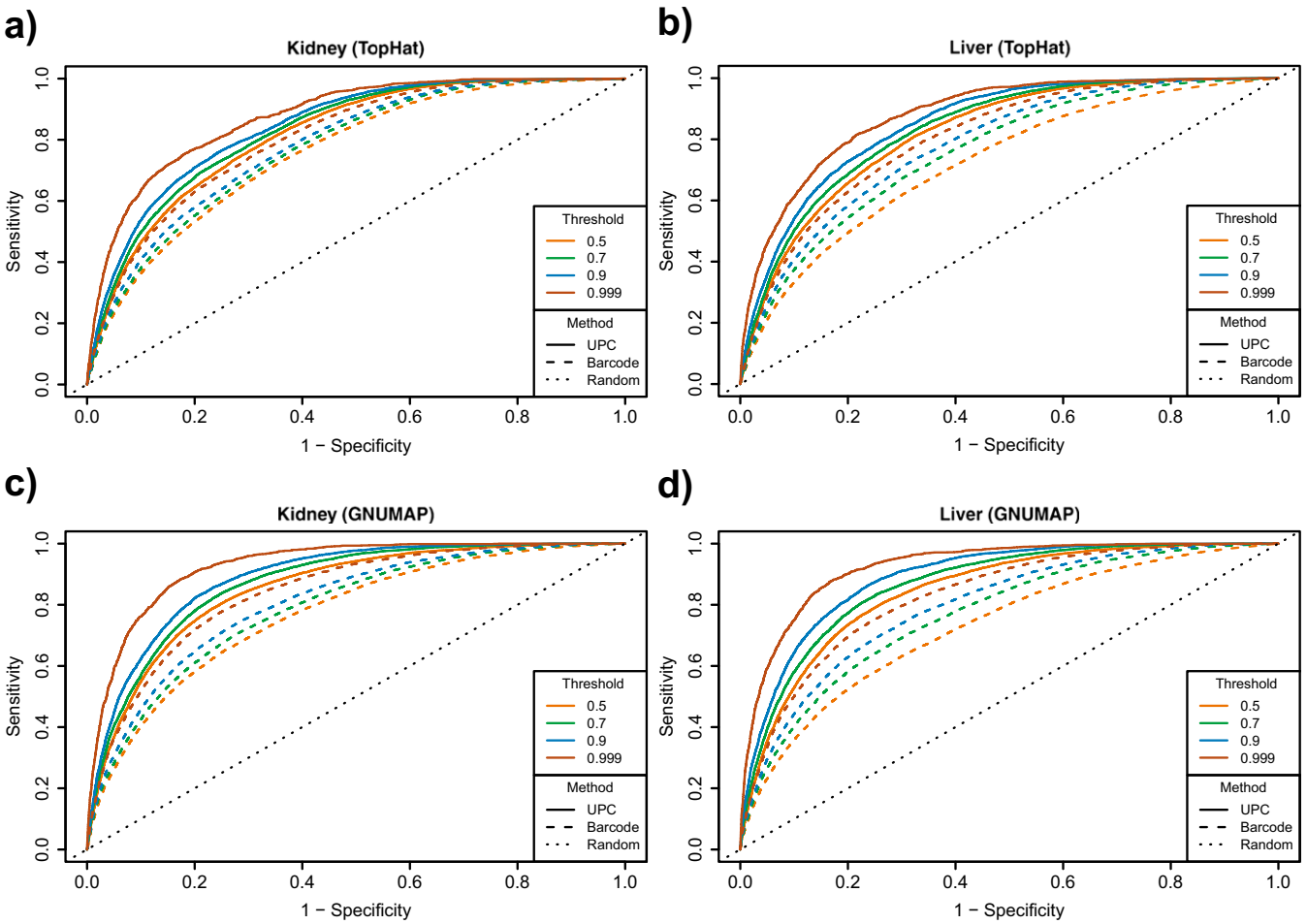


Fig. S7. Comparisons between UPC and barcode approaches using RNA-Seq read counts as a reference standard. These ROC curves illustrate each method's ability to detect active transcripts using RNA-Seq read counts in kidney and liver tissue (1) as a reference standard. Microarray values were transformed to active/inactive calls using thresholds ranging between 0.5 and 0.999. In *A* and *B*, the TopHat read aligner was used. In *C* and *D*, the GNUMAP read aligner was used. For both read aligners, UPCs performed better at each threshold.

1. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y (2008) RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* 18(9): 1509–1517.

Table S1. Summary of comparisons performed across RNA-Seq normalization methods (alternative read aligners)

Dataset	Comparison	UPC (nn)	UPC (ln)	UPC (nb)	RPKM	CQN
Marioni et al. (1)*	Correlation between microarray and RNA-Seq†	0.652	0.360	0.304	0.599	0.669
	Genes designated as active in microarray or inactive in RNA-Seq	77.1%	38.0%	99.8%	N/A	N/A
	Concordance for microarray active/RNA-Seq inactive genes	96.5%	99.4%	99.8%	N/A	N/A
MAQC*	Correlation between RNA-Seq and qPCR†	0.832	0.715	0.625	0.732	0.838
	AUC for present/absent calls	0.857	0.831	0.820	0.829	0.856
Body Map 2.0‡	Accuracy in predicting Wang et al. (2) tissue types (10 genes)§	0.844	0.894	0.672	0.728	0.733
	Accuracy in predicting Wang et al. tissue types (50 genes)§	0.950	0.839	0.528	1.000	1.000
	Accuracy in predicting Wang et al. tissue types (100 genes)§	0.950	0.944	0.744	1.000	1.000
	Accuracy in predicting Wang et al. tissue types (500 genes)§	0.950	0.944	0.900	1.000	1.000

ln, log-normal; nb, negative-binomial; nn, normal-normal.

*Aligned to the reference genome using GNUMAP. The log-normal and negative-binomial models exhibited less stability when GNUMAP was used, perhaps because the read counts exhibited less bimodality.

†Correlation coefficients were calculated using Spearman's rank-based method.

‡Aligned to the reference genome using BWA.

§Prediction accuracy is represented by the area under the receiver operating characteristic curve.

1. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y (2008) RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* 18(9):1509–1517.
2. Wang ET, et al. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature* 456(7221):470–476.