Errors in RNA-seq transcript abundance estimates

Sonali Arora[1], Siobhan S. Pattwell[1], Eric C. Holland[1]*, Hamid Bolouri[1]*

[1] Division of Human Biology, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA

* Correspondence: eholland@fhcrc.org (E. C. Holland), hbolouri@fhcrc.org (H. Bolouri)

RNA-sequencing data is widely used to identify disease biomarkers and therapeutic targets. Here, using data from five RNA-seq processing pipelines applied to 6,690 human tumor and normal tissues, we show that for >12% of protein-coding genes, in at least 1% of samples, current RNA-seq processing pipelines differ in their abundance estimates by more than four-fold using the same samples and the same set of RNA-seq reads, raising clinical concern.

To explore the effects of RNA-seq data processing differences on gene expression estimates, we downloaded pan-tissue uniformly-processed RNA-seq abundance values from five different processing pipelines (details in **Supplementary Table 1)** for 4,800 tumor samples from The Cancer Genome Atlas (TCGA), and calls from four pipelines for 1,890 normal-tissue samples from the Genotype-Tissue Expression (GTEx) project. We also included an additional dataset for the same samples with batch effect correction between TCGA and GTEx. To ensure fair comparisons, we limited all our analyses to protein coding genes that appear across all data sets (16,109 for TCGA; 16,518 for GTEx).

Four of the above data sources are only available in units of reads/fragments per kilobase of transcript per million mapped reads (RPKM/FPKM)[1]. RPKM/FPKM counts do not take transcript lengths into account and are thus sensitive to differences in reference transcriptome annotations (**Supplementary Table 1**), as illustrated in **Fig. 1a,b**. To overcome this issue, we converted all data sets to units of transcripts per million (TPM) in an identical manner[1] (see **Online Methods**).

As expected, after the above unit conversion, the different versions of TCGA and GTEx data appear highly similar at the whole-genome level (**Fig. 1c,d and Supplementary Fig. 1,2**), and inter-tissue differences become more pronounced than data-source effects (**Fig. 1e,f)**. Moreover, inter-batch differences do not appear to confound comparisons within either TCGA or GTEx data sets, irrespective of data source (**Supplementary Fig. 3,4**). Nonetheless, we note that gene expression differences between TCGA and GTEx samples may still be subject to batch effects[2], as illustrated in the before-and-after batch-correction plots in **Supplementary Fig. 5**.

To quantify inter-pipeline differences not related to batch-effects, we analyzed data from the TCGA and GTEx projects separately, and compared only pipelines that do not include batch-effect correction (five sources for TCGA; four sources for GTEx). We searched for genes whose expression in a given sample is >32 TPM according to one pipeline (suggesting high expression), but <8 TPM (i.e. >4-fold lower) in the same sample according to at least one other pipeline. To ensure that such large differences between abundance estimates are not due to a small number of outlier samples, we further required that the expression of a gene should be discordant (i.e. meet the above criteria) in at least 1% of all samples (48 samples for TCGA, 19 samples for GTEx). We refer to genes that meet all the preceding criteria as "discordantly quantified" genes.

We found 1,637 discordantly quantified genes (~10% of all genes considered) in TCGA data and 1,214 discordantly quantified genes (~7%) in GTEx data (**Fig. 2a,b**). Across all TCGA and GTEx data sets, of 16,738 genes analyzed, 2,068 genes (12.36%) are discordantly quantified (**Supplementary Table 2**). Of note, many discordantly quantified genes have divergent expression values in large numbers of samples (>500, **Fig. 2c,d** and **Supplementary Table 3**). Notably, the observed discrepancies are not attributable to a particular subset of processing pipelines or a particular subset of samples. Even the two pipelines with the greatest level of agreement (MSKCC and Xena/TOIL), still include multiple genes that are discordantly quantified in large numbers of samples.

It is important to note that we imposed very stringent criteria in our identification of genes with discordant expression estimates across pipelines. In particular, three and two-fold expression differences between pipelines affect many more genes (**Supplementary Tables 2b,c)**. For most discordantly quantified genes, differences between pipelines are quantitative and threshold dependent. As an example, **Fig. 2e** shows the expression pattern of an example discordantly quantified gene (the splicing regulator U2AF1, which is frequently mutated in Myelodyplastic Syndrome[3]) in five versions of TCGA data (see **Supplementary Fig. 6a,b** for additional examples). We note that differences in abundance estimates for U2AF1 have diverse characteristics. In some cases, estimates differ by a scaling factor. In other cases, a gene is essentially not expressed according to one pipeline, but highly expressed according to another pipeline. In yet other cases, abundance estimates from two pipelines are simply poorly correlated. These large uncertainties in abundance estimates pose a significant challenge to biomedical research.

It is important to note that the discordant expression abundance estimates reported here arise purely from data processing steps performed on exactly the same set of RNA-seq reads, and after correcting for differences in annotated transcript lengths.

Many differences among RNA-seq processing pipelines can contribute to variability in abundance estimates. For example, more complex genome annotations can increase the numbers of unmapped and multi-mapped reads[4]. Accordingly, 240 of our 2,068 discordantly quantified genes (11.61%) are known to be frequently affected by multi-mapping reads (**Supplementary Table 2**)[5]. Additional factors affecting RNA-seq

abundance estimates include the numbers of annotated variants, alignment-based versus transcript-based abundance estimation, and how reads that map partially or to multiple loci are handled.

Our list of 2,068 discordantly quantified genes includes 784 disease-associated genes (**Supplementary Table 4**), such as CEBPA, HIF1A, KRAS, NPM1, and PDGFRA, with important clinical implications. In addition, for approximately half of the TCGA genes with available protein abundance data, mRNA and protein levels show remarkably low levels of correlation (**Supplementary Fig. 7, Supplementary Tables 5a,b,c**) raising further concerns regarding the use of RNA-seq data for biomarker and target discovery.

The discordant gene expression estimates reported here do not imply any technical errors. Rather they highlight a degree of uncertainty inherent in processing noisy and complex data. Nonetheless, the end result is that for the discordantly quantified genes reported here (**Supplementary Table 2a**), we can have little confidence in the abundance estimates produced by any RNA-seq processing pipeline. A concerted, community-wide effort will be needed to develop a gold-standard for estimating the mRNA abundance of these genes.

Acknowledgements: We thank …. **Funding:** National Institutes of Health U54 CA193461 (E.C.H.), T32 CA9657-25 (S.S.P.), U54 DK106829 (S.S.P.), R21 CA223531 (S.S.P.); Jacobs Foundation Research Fellowship (S.S.P.). **Author Contributions:** All analysis was performed by SA and HB and discussed with SSP and ECH. All authors contributed to the final manuscript. **Competing Interests:** Authors declare no competing interests. **Data and materials availability:** https://github.com/sonali-bioc/UncertainityRNA.

Online Methods

All analyses were performed in R (https://www.r-project.org/) using Bioconductor (https://www.bioconductor.org/) packages. To maximize transparency and reproducible research, we have deposited all scripts, associated data, and a large number of additional plots in a Github repository: https://github.com/sonali-bioc/UncertainityRNA.

Figure Captions

Fig. 1. TCGA and GTEx data processed using diverse pipelines exhibit more variation by tissue source than by pipeline. (**a,b**) Principle Component Analysis (PCA) plots of TCGA and GTEx data using RPKM/FPKM values show very large inter-pipeline differences. (**c,d**) After transformation to TPM values, differences among pipelines are dramatically reduced. (**e, f**) Using TPM, variability between tissue types far exceeds inter-pipeline variability.

Fig. 2. Large-scale inter-pipeline variability in specific genes. (**a, b**) Number of discordant genes per data source for TCGA (C) and GTEx (D). Each panel shows both UpSet plots[6] and Venn diagrams. (**c, d**) Heatmaps showing the number of discordant samples per gene in TCGA and GTEx data. Each row represents one gene, each column represents a comparison between two pipelines (as labeled). (**e**) Example pairwise comparisons of expression abundance estimates for E2AF1 showing diverse modes of disagreement among different pipelines.

References

1. Conesa, A.*, et al.* A survey of best practices for RNA-seq data analysis. *Genome Biol* **17**, 13 (2016).
2. Wang, Q.*, et al.* Unifying cancer and normal RNA sequencing data from different sources. *Sci Data* **5**, 180061 (2018).
3. Ilagan, J.O.*, et al.* U2AF1 mutations alter splice site recognition in hematological malignancies. *Genome Res* **25**, 14-26 (2015).
4. Wu, P.Y., Phan, J.H. & Wang, M.D. Assessing the impact of human genome annotation choice on RNA-seq expression estimates. *BMC Bioinformatics* **14 Suppl 11**, S8 (2013).
5. Robert, C. & Watson, M. Errors in RNA-Seq quantification affect genes of relevance to human disease. *Genome Biol* **16**, 177 (2015).
6. Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R. & Pfister, H. UpSet: Visualization of Intersecting Sets. *IEEE Trans Vis Comput Graph* **20**, 1983-1992 (2014).

Figure 1

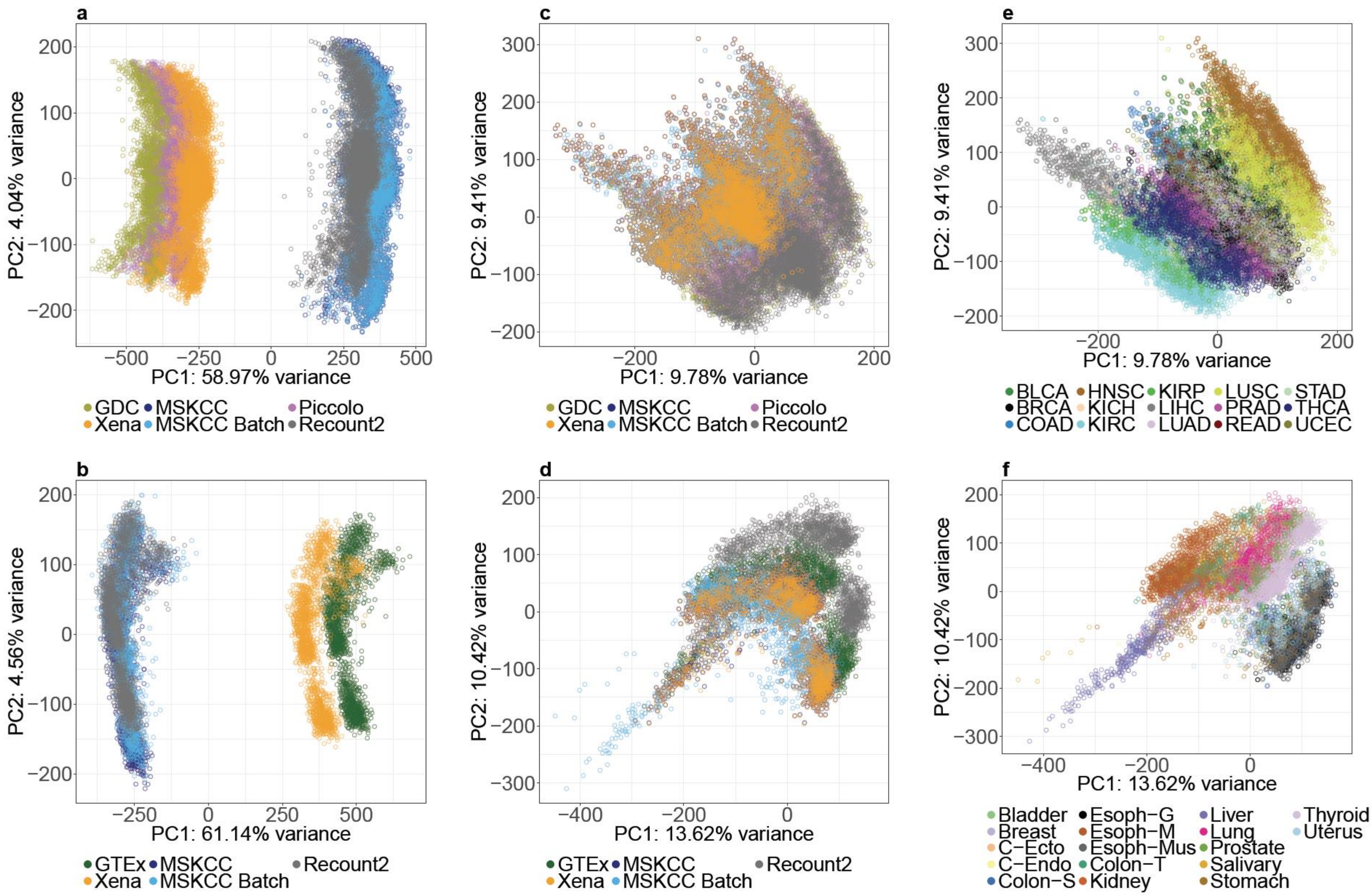
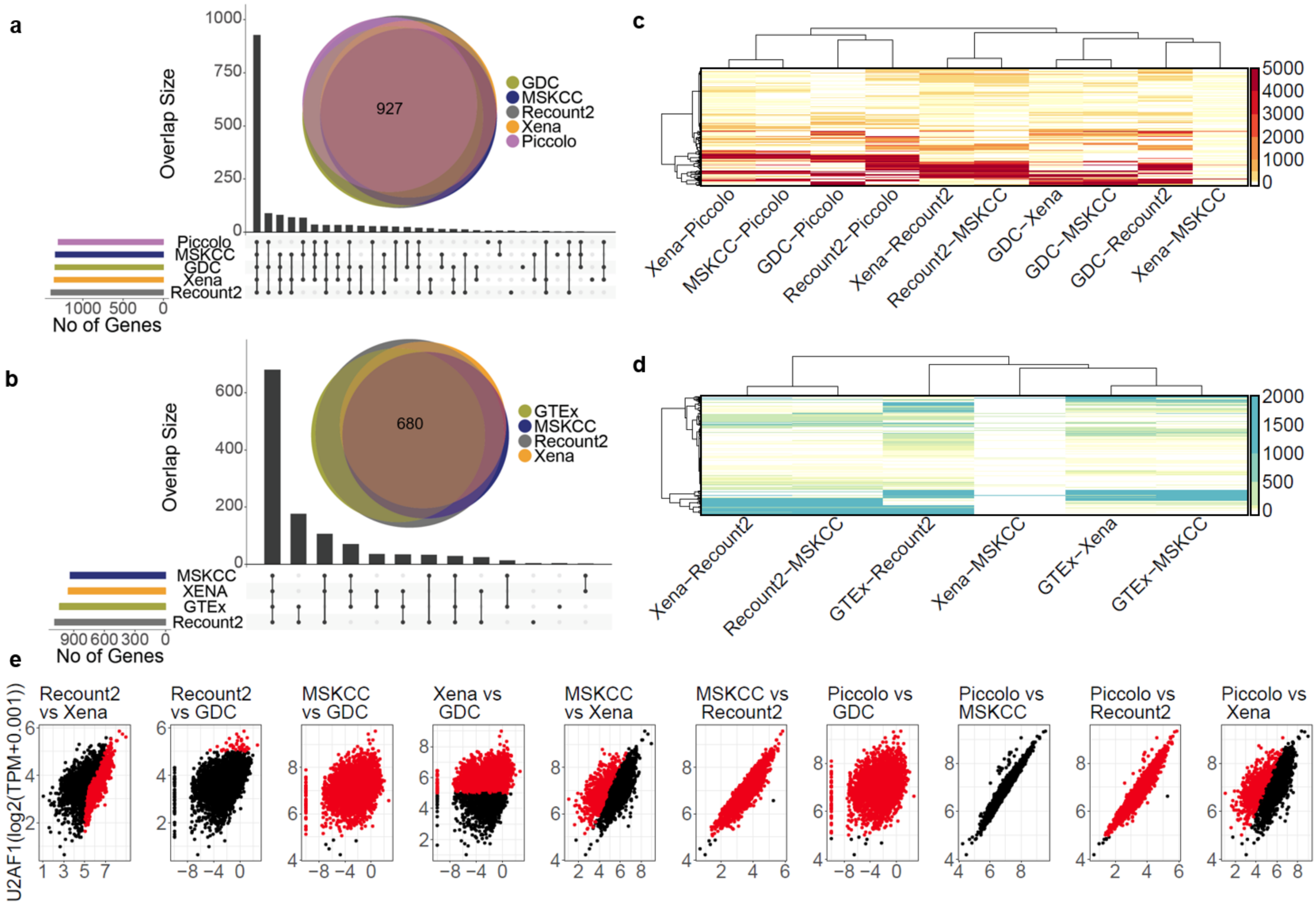

a
PC2: 4.04% variance
PC1: 58.97% variance
GDC MSKCC Piccolo
Xena MSKCC Batch Recount2
c
PC2: 9.41% variance
PC1: 9.78% variance
GDC MSKCC Piccolo
Xena MSKCC Batch Recount2
e
PC2: 9.41% variance
PC1: 9.78% variance
BLCA HNSC KIRP LUSC STAD
BRCA KICH LIHC PRAD THCA
COAD KIRC LUAD READ UCEC
b
PC2: 4.56% variance
PC1: 61.14% variance
GTEx MSKCC Recount2
Xena MSKCC Batch
d
PC2: 10.42% variance
PC1: 13.62% variance
GTEx MSKCC Recount2
Xena MSKCC Batch
f
PC2: 10.42% variance
PC1: 13.62% variance
Bladder Esoph−G Liver Thyroid
Breast Esoph−M Lung Uterus
C−Ecto Esoph−Mus Prostate
C−Endo Colon−T Salivary
Colon−S Kidney Stomach

Figure 2


a
Overlap Size
1000
750
500
250
0
927
GDC
MSKCC
Recount2
Xena
Piccolo
Piccolo
MSKCC
GDC
Xena
Recount2
1000 500 0
No of Genes
b
Overlap Size
600
400
200
0
680
GTEx
MSKCC
Recount2
Xena
MSKCC
XENA
GTEx
Recount2
900 600 300 0
No of Genes
c
5000
4000
3000
2000
1000
0
Xena−Piccolo
MSKCC−Piccolo
GDC−Piccolo
Recount2−Piccolo
Xena−Recount2
Recount2−MSKCC
GDC−Xena
GDC−MSKCC
GDC−Recount2
Xena−MSKCC
d
2000
1500
1000
500
0
Xena−Recount2
Recount2−MSKCC
GTEx−Recount2
Xena−MSKCC
GTEx−Xena
GTEx−MSKCC
e
U2AF1(log2(TPM+0.001))
Recount2 vs Xena
Recount2 vs GDC
MSKCC vs GDC
Xena vs GDC
MSKCC vs Xena
MSKCC vs Recount2
Piccolo vs GDC
Piccolo vs MSKCC
Piccolo vs Recount2
Piccolo vs Xena