>12% of protein coding genes have >4-fold mRNA abundance difference depending on bioinformatic processing pipeline

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

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

FRED HUTCH CURES START HERE™

Abstract

Motivation: RNA-sequencing data is widely used to identify disease biomarkers and therapeutic targets. How reliable is it?

Approach: We compared data for 16730 genes from five "best-in-class" RNA-seq processing pipelines applied to 4,800 tumor and 1,890 normal human tissues from TCGA and GTEx.

Results: We show that for >12 % of protein-coding genes, RNA-seq expression estimates by different pipelines differ by > 4-fold in at least 10% of samples *using the same sequencing reads*.

Conclusions: A total of 2071 genes are discordantly quantified by current "best-in-class" RNA-seq processing pipelines.

Overview & Motivation

Our Pipeline

Download processed TCGA and GTEx data

↓

Subset for shared genes and samples/ tissues

Normalize to Transcripts Per Million (TPM)

↓

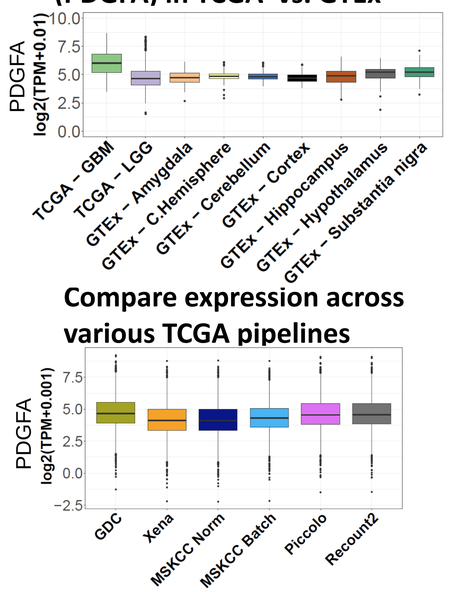
Compare RNASeq expression across TCGA and GTEx

Different Number of genes & samples from each source

TCGA Sources	Source	No. of genes	No. of samples
GDC	GDC website	19,560	11,037
XENA-Toil	UCSC	19,797	10,535
Recount2	John Hopkins	19,950	11,229
MSKCC-Normalized	cBio	20,242	6,142
MSKCC– Batch Corrected	cBio	18,154	6,142
Piccolo-Lab	Bringham Young Univ.	20,871	7,706

GTEx Sources	Source	No. of genes	No. of samples
GTEx	GTEX-v6	19,644	8,555
XENA-Toil	UCSC	19,797	7,791
Recount2	John Hopkins	19,950	9,662
MSKCC-Normalized	cBio	19,610	2,445
MSKCC– Batch Corrected	cBio	17,614	2,322

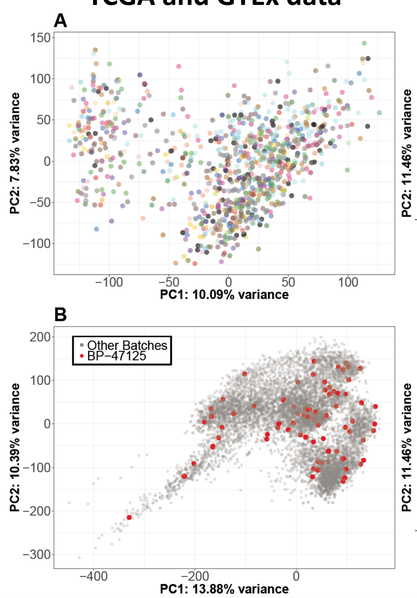
Compare expression of a gene (PDGFA) in TCGA vs. GTEx



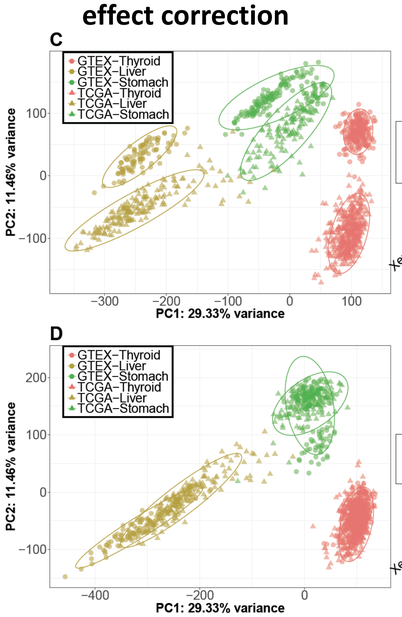
Compare expression across various TCGA pipelines

Fig 2: Expression estimates within individual pipelines are consistent, but differ across pipelines.

No Batch effects in TCGA and GTEx data



Inter TCGA/GTEx batch-effect correction



Discordant samples in TCGA and GTEx sources

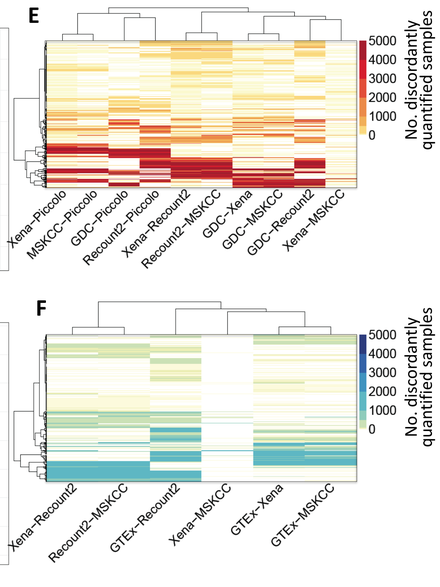
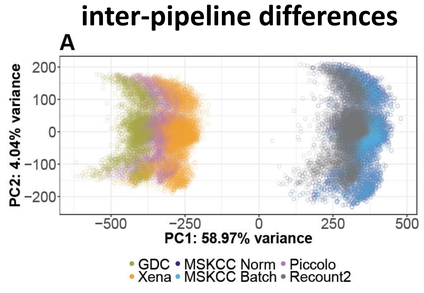
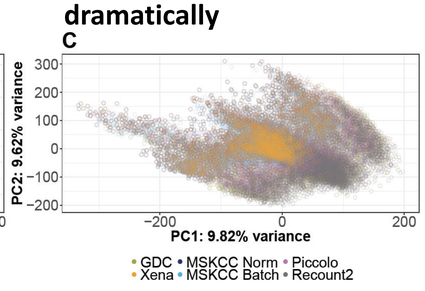


Fig 1: Normalization removes unwanted variability across Pipelines.

TCGA and GTEx FPKM values show very large inter-pipeline differences



TPM normalization reduces inter-pipeline differenmces dramatically



Using TPM, tissue type variability exceeds inter-pipeline variability

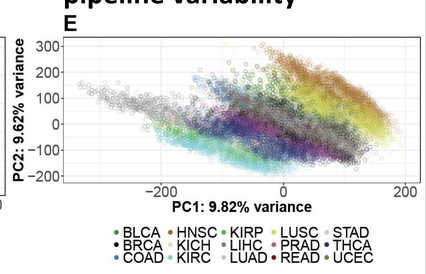
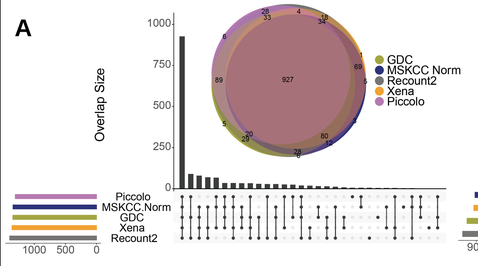
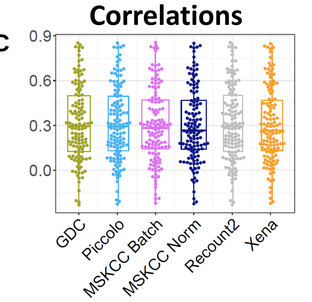


Fig 3: 2071 (12.4%) genes are discordantly quantified across pipelines.

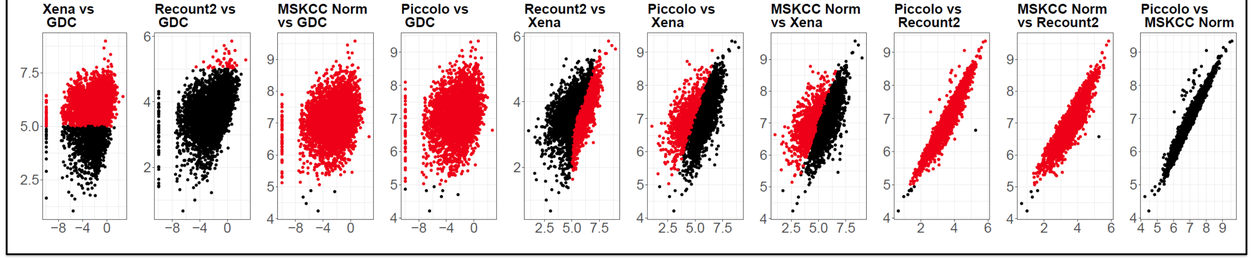
Different genes are discordantly quantified by different pipelines



Low protein-RNA Correlations



D Example pairwise expression comparisons using the U2AF1 gene and TCGA data



References

1. Grossman, Robert L., Heath, Allison Pet al. (2016) Toward a Shared Vision for Cancer Genomic Data. New England Journal of Medicine
2. Vivian J, Rao AA, Nothaft FA, et al. (2017) Toil enables reproducible, open source, big biomedical data analyses. Nature biotechnology.
3. Collado-Torres L, Nellore A, et al (2017) Reproducible RNA-seq analysis using recount2. Nature biotechnology.
4. Rahman M, et al. (2015) Alternative preprocessing of RNA-Sequencing data in TCGA leads to improved analysis results. Bioinformatics.
5. Q. Wang, J Armenia, et al. (2018) Unifying cancer and normal RNA sequencing data from different sources. Scientific Data
6. The GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. (2013) Nature genetics.

Example Impacted Cancer genes: ABCB6, ACY1, AKR1C1, AMACR, ANXA2, ARID4B, ARL2BP, ASPSCR1, ATF7IP, BMI1, BOP1, BRAF, BRD2, CD163L1, CEBPA, CNPY2, CSNK1A1, CSNK2B, DDR1, DLEC1, EIF4E, FAM168A, FCGBP, FHIT, GAGE1, GINS2, GNG11, GPR27, GSTM2, H3F3A, H3F3B, HIST1H1E, HIST1H3B, HOXD9, HRH2, KLRK1, KRT6B, KRT7, KYNU, LGALS7, LHFPL5, MIA, MORF4L1, MRPS18B, MUC6, MYCBP, P2RY8, PGAM1, PHLDB1, PKMYT1, PPIA, PSORS1C2, QKI, RHEB, RPL13, RPL14, RPL18, RPS3, RXRB, SLC19A1, SOD2, SPRR1A, STK19, STRADA, TATDN1, TNF, TPM3, TRIM27, TUBA1C, TXNDC17, UGT2B17, VDAC2, VPREB1, WDR46, WFDC1, ZNF593, ZNF668