Basal Contamination of Bulk Sequencing: Lessons from the GTEx dataset

Tim O. Nieuwenhuis[1,2], Stephanie Yang[2], Vamsee Pillalamarri[2], Dan E. Arking[2], Avi Z. Rosenberg[1], Matthew N. McCall[3], Marc K. Halushka[1]

[1] Department of Pathology, Johns Hopkins University SOM, Baltimore, MD, USA

[2]McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University SOM, Baltimore, MD, USA

[3]Department of Biostatistics and Computational Biology, University of Rochester Medical Center, Rochester, NY, USA

Email Addresses:

tnieuwe1@jhmi.edu

syang93@jhmi.edu

vpillal1@jhmi.edu

arking@jhmi.edu

arosen34@jhmi.edu

matthew_mccall@urmc.rochester.edu

* Correspondence and address for reprints to:
Marc K. Halushka, M.D., Ph.D.
Johns Hopkins University School of Medicine
Ross Bldg. Rm 632B
720 Rutland Avenue
Baltimore, MD 21205
410-614-8138 (ph)
410-502-5862 (fax)
mhalush1@jhmi.edu

Abstract

Background: One of the challenges of next generation sequencing (NGS) is contaminating reads from other samples. We used the Genotype-Tissue Expression (GTEx) project, a large, diverse, and robustly generated dataset, as a useful resource to understand the factors that contribute to contamination.

Results: We obtained 11,340 RNA-Seq samples, DNA variant call files (VCF) of 635 individuals, and technical metadata from GTEx as well as read count data from the Human Protein Atlas (HPA) and a pharmacogenetics study. We analyzed 48 tissues in GTEx. Of these, 24 had variant co-expression clusters of four known highly expressed and pancreas-enriched genes (*PRSS1*, *PNLIP*, *CLPS*, and *CELA3A*). Fifteen additional highly expressed genes from other tissues were also indicative of contamination (*KRT4*, *KRT13*, *PGC*, *CPA1*, *GP2, PRL, LIPF, CTRB2, FGA, HP, CKM, FGG, MYBPC1, MYH2, ZG16B*). Sample contamination by non-native genes was highly associated with a sample being sequenced on the same day as a tissue that natively has high levels of those genes. This was highly significant for both pancreas genes (p= 2.7E-75) and esophagus genes (p= 8.9E-154). We used genetic polymorphism differences between individuals as validation of the contamination. Specifically, 11 SNPs in five genes shown to contaminate non-native tissues demonstrated allelic differences between DNA-based genotypes and contaminated sample RNA-based genotypes. Low-level contamination affected 1,841 (15.8%) samples (defined as ≥500 *PRSS1* read counts). It also led to eQTL assignments in inappropriate tissues among these 19 genes. In support of this type of contamination occurring widely, pancreas gene contamination (*PRSS1*) was also observed in the HPA dataset, where pancreas samples were sequenced, but not in the pharmacogenomics dataset, where they were not.

Conclusions: Highly expressed, tissue-enriched genes basally contaminate the GTEx dataset impacting on some downstream GTEx data analyses. This type of contamination is not unique to GTEx, being shared with other datasets. Awareness of this process will reduce assigning variable, contaminating low-level gene expression to disease processes.

Key Words: GTEx, RNA-Seq, Contamination, eQTL, PEER factors

Introduction

The rise of next generation sequencing has allowed for unparalleled data generation for a variety of nucleic acid studies including RNA expression. As cost per basepair decreases, more large-scale transcriptome projects can be performed that will inform on tissue expression patterns in health and disease [1-4]. These data sources are generally publicly-available and have been used by hundreds of researchers for secondary analyses of high impact [5, 6].

Limitations exists for all –omics technologies, including bulk RNA sequencing (RNA-Seq). Issues of hybridization biases, library preparation biases, and computational biases such as positional fragment bias are known limitations of RNA-Seq experiments [7-9]. Another challenge of high throughput RNA-Seq is contamination, leading to the presence of sequence data within a dataset of one sample that originates from a separate sample. This contamination can come from many different aspects of the modern sequencing process, such as human error, machine or equipment contamination, intrinsic preparation and sequencing errors, and computational errors, including errors that can occur based on the multiplexing methods used [10-12]. Contamination has been better characterized for DNA sequencing projects [13-15].

The Genotype-Tissue Expression project (GTEx) aims to create a large publicly available database of tissue-specific expression quantitative trait loci (eQTL) from over 40 tissues [1]. It is an ongoing project with over 700 individuals and 11,000 tissue samples. GTEx combines genotyping from whole genome sequencing with gene expression levels from bulk RNA-Seq. GTEx has made their RNA-Seq, phenotype, genotype, and technical data available for public access with permission.

In an analysis of variation in the GTEx RNA-Seq data (V7), we detected unexpected sources of variation that we hypothesized were likely contaminating sequence reads found at low, but variable levels across different tissues. Herein we describe how we identified the source of contamination and establish basal rates of contamination in the GTEx RNA-Seq data.

Results

Patterns of extreme tissue variation identified usual gene signatures

We embarked on a project to expand our initial description of the causes of lung expression variation in GTEx to all tissue samples using DEseq2 variance stabilizing transformation to normalize read counts from 11,340 samples [16, 17]. We filtered genes in each tissue keeping those with a mean transformed count >5. The median number of genes above the expression threshold was 17,729 with the highest and lowest gene counts being 23,930 and 13,807 in the testis and whole blood respectively. As previously described, we correlated and hierarchically clustered variable genes (>4 variance across samples) for all tissues with >70 samples (N=48) in the GTEx dataset [16]. Our algorithm identified multiple gene clusters per tissue, based on their Kendall's tau correlations. It additionally reported non-clustering, highly variable genes. Most clusters were the result of biologic and phenotypic features related to the tissues. For example, a cluster of Y chromosome genes and *XIST* appeared in 42 of 43 non-sex specific tissues. However, there was one consistent pattern of 3-4 genes (*PNLIP,* *PRSS1*, *CELA3A,* and/or *CLPS)* identified in 24 of the 48 tissues, that failed to have an intuitive explanation as these genes are highly-expressed and specific to the pancreas. We then determined if there were other highly expressed tissue enriched genes appearing variably in other samples. To further understand this, we utilized a list of tissue enriched proteins generated by the Human Protein Atlas (HPA) and cross-referenced this to GTEx TPM data (Table 1) [18, 19]. From this list, we noted 19 genes from 7 tissues including two esophagus genes *KRT13* and *KRT4* that are highly expressed in their native tissue and identified as variable in three or more other unrelated tissues (Fig. 1a, Additional File 1: Fig. S1).

Table 1 GTEx and HPA highly expressed, tissue-enriched genes present in other tissues through contamination

Gene	Times identified as variable in other tissues	Highest expressed GTEx/HPA tissue	GTEx TPM	HPA TPM	Second highest expressed HPA tissue	HPA TPM in second tissue
PRSS1	41	Pancreas	99,100	81,683	Ovary	257
PNLIP	33	Pancreas	33,660	93,703	Ovary	288
CPA1	30	Pancreas	54,500	48,857	Ovary	133
GP2	29	Pancreas	14,280	7,530	Duodenum	36
CELA3A	23	Pancreas	27,130	56,988	Ovary	162
KRT13	20	Esophagus	33,960	35,139	Tonsil	1,728
PGC	19	Stomach	36,720	22,276	Duodenum	1,302
KRT4	18	Esophagus	22,290	14,862	tonsil	599
PRL	17	Pituitary	54,500	--	--	--
LIPF	14	Stomach	29,380	22,415	Duodenum	259
CLPS	13	Pancreas	51,640	56,632	Ovary	214
CTRB2	8	Pancreas	20,760	29,060	Ovary	74
FGA	6	Liver	5,717	9,265	Stomach	39
HP	6	Liver	12,710	28,407	Bone marrow	155.8
CKM	5	Skeletal muscle	11,138	23,799	Heart	1,419
FGG	5	Liver	6,623	8,699	Lung	75

MYBPC1	5	Skeletal muscle	3,587	3,918	Prostate	125
MYH2	5	Skeletal muscle	1,064	4,306	Esophagus	44
ZG16B	5	Salivary gland	17,540	19,471	Prostate	87

As both abundant and tissue-enriched genes were unlikely to be randomly and lowly expressed in a range of other tissues, we performed analyses to determine the source of the contamination.

Nucleic acid isolation is a minor source of contamination

We first questioned if the contamination occurred during tissue harvesting, hypothesizing that occasionally small fragments of a tissue could contaminate a separate sample from shared dissection tools or surfaces. For that to be true, we reasoned that organs near the pancreas/esophagus, or temporally collected relative to the pancreas/esophagus would be most affected. However, a pancreas gene contamination cluster was found in transformed fibroblasts which were grown over multiple passages and would not retain other cell types over that time period, excluding this possibility (Additional File 1: Figure S1). Using the available technical metadata, we found a modest association between nucleic acid isolation date and the presence of contamination (p= 0.003, linear regression model). Thus, date of nucleic acid isolation may represent a small aspect of the contamination.

Identification of sequencing date as a correlate to contamination

We then ascertained if the contamination was occurring at the time of sequencing. A linear regression model estimated that contamination was 0.85 standard deviations higher when a sample was sequenced on the same day as a pancreas sample (p= 2.66e-75). (Fig. 1b,c). When the model included both nucleic acid isolation date and sequencing date, the association with nucleic acid isolation was not significant (p= 0.31), whereas the sequencing date remained strongly associated with contamination (p= 1.436e-73), suggesting that the sequencing date was the primary cause of contamination. A comparison of the aforementioned models using a one way anova test indicated nucleic acid isolation date did not significantly increase the variance explained in normalized contamination scores (p= 0.31). A similar association between sequencing data and contamination was observed with esophageal gene contamination, which in the same model, had a strong association with nucleic acid isolation date (p= 4.59e-16) but a stronger association with sequencing date (p= 8.95e-154). In the samples, contamination by esophagus-enriched genes had a negative association with having nucleic acid isolation on the same day as an esophagus (-0.306 Z-Score, p= 4.59e-16), discounting nucleic acid isolation date as the main point of contamination. Despite this strong correlation with sequencing, some high Z-scores came from samples that were not sequenced on the same days as pancreata. Further analysis showed that essentially all of these samples were sequenced within a few days of a pancreas (Fig. 1d). This additionally implicated the library preparation process (for which date information is lacking in GTEx) which is temporally related to sequencing, rather than the sequencing itself.

Genetic polymorphisms confirm contamination is derived from other samples

To prove that pancreas/esophagus transcripts were contaminating from other (non-self) samples we investigated for incongruencies between a person's genotype (from DNA data) and the genotype in matching loci in the pancreas/esophagus contaminated RNA-Seq samples. We required both the individuals' DNA genotype and their contamination source RNA-Seq as we are aware of both RNA editing and preferential allele expression. Based on sample requirements and limited by available raw sequencing files, we identified 11 contaminated tissues to evaluate. For each, we obtained and processed their raw RAN-Seq FASTQ sequences to identify variants in both their contaminated tissues and their matched pancreas or esophagus tissue (depending on the gene source of contamination). Additionally, we used the GTEx filtered VCF file from their sequenced DNA to further establish their SNP allele patterns. Across all tissues, 533 SNPs, rare variants, and private variants, were investigated in pancreas associated gene coding sequences (*PNLIP*, *CLPS*, and *CELA3A*) and 190 in esophagus associated gene coding sequences (*KRT13*, *KRT4*). As a comparison group, 287 variants were investigated in two control gene coding sequences (*GAPDH*, and *RAB7A*) that have near ubiquitous expression across all tissues. Of 1,010 variants obtained from the combined VCF files, 11 had some degree of allelic heterogeneity (Table 2). No incongruencies were found in the 287 variants of the two control genes.

Table 2 Allelic incongruencies found in contaminated samples

			Enriched Tissue			Contaminated		
Individual	Gene	SNP	Major/ Minor	Reads	Major Allele %	Tissue Type	Reads	Major Allele %
GTEX-1	*KRT13*	rs903	C/A	101,908	0%	Fibroblast Cells	252	50%
GTEX-1	*KRT4*	rs7959052	T/C	74,468	100%	Fibroblast Cells	203	12%
GTEX-1	*KRT4*	rs7956809	C/G	85,803	100%	Fibroblast Cells	204	13%
GTEX-1	*KRT4*	rs2035879	T/C	72,978	51%	Fibroblast Cells	164	7%

GTEX-1	*KRT4*	rs17119475	G/A	71,592	49%	Fibroblast Cells	226	98%
GTEX-9	*CELA3A*	rs3820285	C/G	98,896	1%	Adipose	5,178	48%
GTEX-9	*CELA3A*	rs9187	C/T	105,462	75%	Adipose	6,082	97%
GTEX-9	*CELA3A*	rs12908	G/A	108,681	75%	Adipose	6,313	98%
GTEX-8	*CELA3A*	rs9187	C/T	162,318	73%	Tibial Nerve	1,155	100%
GTEX-8	*CELA3A*	rs12908	G/A	169,394	74%	Tibial Nerve	1,215	100%
GTEX-10	*CLPS*	rs3748050	T/C	80,019	47% C	Artery	1,117	99%

One SNP site (rs7956809), was particularly informative. SNP rs7956809 (C/G), located in *KRT4*, had a relatively low allelic variation, with only 5 individuals in the entire GTEx cohort homozygous for the alternative allele (G). One sample (arbitrarily GTEX1) was homozygous C at rs7956809 in both its DNA (VCF file) and matched esophagus (RNA-Seq FASTQ data) (Fig. 1e). However, the rs7956809 SNP in the GTEX1 fibroblast sample was 87% G and 13% C. Six esophagus samples were sequenced on the same day as the GTEX1 fibroblast sample. No other esophagus samples were sequenced within 4 days. One of those six samples, GTEX2, was homozygous G at rs7956809. The five other samples were homozygous C. This strongly implicates the GTEX2 esophagus sample as the dominant contaminant of the GTEX1 fibroblast sample.

We further investigated the relationship between the GTEX1 fibroblast sample and the GTEX2 esophagus sample finding no clear connection. The two samples were sequenced on different machines and in different flow cells. Of some interest, the sequencing sample adapters (molecular indexes) were similar (Additional File 2: Table S1).

The extent of highly expressed, tissue-enriched gene contamination in GTEx

After establishing that contamination exists in GTEx, identifying a temporal association and polymorphism validation, we then attempted to address the extent of contamination in the GTEx dataset. To characterize this we investigated the various levels of pancreas enhanced gene expression in non-pancreatic tissue (Table 1). In the 10,298 non-pancreas samples investigated, <0.5% had >10,000 read counts of PRSS1, the most abundant pancreas gene (Table 3). However, at a threshold of >100 read counts, over half of samples contained some *PRSS1*.

Table 3 Extent of contamination of 11,092 non-pancreas samples by pancreas genes.

Gene	Read Count > 10,000	Read Count > 1,000	Read Count > 100
PRSS1	49 (0.44%)	782 (7.1%)	5802 (52.3%)
PNLIP	30 (0.27%)	278 (2.5%)	4511 (40.6%)
CELA3A	24 (0.22%)	253 (2.3%)	4102 (37.1%)
CLPS	13 (0.12%)	122 (1.1%)	2587 (23.3%)

Numbers indicate the amount of affected samples and their percentage

PEER factor normalization does not fully correct for contamination

The GTEx analysis pipeline uses probabilistic estimation of expression residuals (PEER) factor to correct for possible confounders [20, 21]. This method identifies hidden factors that explain much of the expression variability and can be used to normalize RNA expression data. We focused on just one tissue, lung, and followed the GTEx analysis pipeline to determine the extent to which PEER factor normalization can identify and correct for this contamination. Sixty PEER factors were identified with the top two identifying a difference between “in hospital” (short postmortem interval) and “outside of hospital” (longer postmortem interval) deaths (Fig. 2a). This relationship is consistent with our prior report of variation in lung [16]. Similar to the global findings of Fig. 1, *PNLIP* expression was increased in lung samples sequenced on the same day as a pancreas. Despite correcting for 35 or even 60 PEER factors, this difference was not fully accounted for (Fig. 2b). Indeed, of five genes evaluated, only one gene (*KRT4*) was fully corrected for by PEER factors (Table 4). We then explored if this lack of full correction impacted eQTL analysis in the GTEx program.

Table 4. Significance of same-day sequencing of lung with contaminating tissues on gene expression.

Gene	P. value before PEER correction	P. value after correcting for 35 PEER factors	P. value after correcting for 60 PEER factors	Beta estimate after correction
PNLIP	4.34e-14	1.38e-11	3.03e-06	0.54
PRSS1	6.29e-14	8.07e-11	5.18e-06	0.52
CELA3A	5.91e-14	8.78e-11	4.86e-06	0.52
KRT4	0.0034	0.055	0.22	0.15
KRT13	8.29e-17	3.70e-08	0.0050	0.36

P. values are shown before and after PEER correction.

Contamination affects GTEx eQTL reporting

Using the GTEx eQTL browser, we identified 75 tissues reported as having significant eQTLs for the 19 genes listed in Table 1. Eight tissues matched the known dominant expression patterns of the genes. An additional 25 tissues were deemed possible based on expression patterns noted by RNA and protein immunohistochemistry in which expression (in TPM) was above the basal level of all tissues. However, 42 inappropriate tissues were identified as harboring eQTLs even though these genes are not natively expressed in these tissues, appearing only as a result of contamination (Table 5).

Table 5 Distribution of GTEx eQTLs by tissue type in contaminating genes

Genes	Appropriate Tissues	Possible tissues	Inappropriate tissues
PRSS1	--	Small intestine	Liver, coronary, skin, lung
PNLIP	--	--	--
CPA1	--	--	Coronary
GP2	--	--	Brain
CELA3A	Pancreas	Stomach	Liver
KRT13	Vagina	--	Lung
PGC	--	Lung, pancreas	Tibial artery
KRT4	Esophagus	Skin, lung	Colon, brain, thyroid
PRL	--	--	Gastroesophageal junction, skin, tibial artery
LIPF	Stomach	--	--
CLPS	Pancreas	--	--
CTRB2	Pancreas	--	Aorta, brain, lung, thyroid
FGA	Liver	Stomach	--
HP	--	Whole blood, adipose (2), artery (3), lung, tibial nerve	brain, esophagus mucosa, heart,
CKM	--	--	Aorta, whole blood
FGG	Liver	Lung, adrenal	--
MYBPC1	--	Heart, prostate, brain (2)	Esophagus (2), colon, lung, thyroid
MYH2	--	--	Colon, lung
ZG16B	--	Skin (2), stomach, prostate, colon	Adipose, adrenal, esophagus, fibroblasts, lung, pituitary, spleen, testis, thyroid, whole blood

Non-GTEx data sets confirm contamination

To determine if highly-expressed tissue-enriched contamination is a feature of sequencing in general, we searched for RNA-Seq datasets that had similar protocols to GTEx, that both included or did not include pancreas samples. We identified an HPA sequencing study which included pancreas [22] and a pharmacogenetics study which did not include pancreas [23]. Both studies were sequenced on Illumina 2000 or 2500 sequencers. The HPA study multiplexed their samples, 15 per lane, but the pharmacogenetics study did not report multiplexing. These data sets demonstrate *PRSS1* contamination of the HPA data (N=19), with essentially no *PRSS1* contamination in the pharmacogenetics study (N=74) (Fig. 2c).

Discussion

The GTEx dataset represents an ideal resource to study sequence contamination. Its 11,000+ samples from 700+ individuals from a diverse set of tissues with all library preparation and sequencing performed at one center is unique. During our initial variation analysis of 46 tissues spanning 10,294 samples, we detected a variable signal of pancreas genes in 24 of those tissues. From there we noticed genes that were highly expressed in esophagus, stomach, pituitary and other tissues also appearing in shared clusters across unrelated tissues. These highly expressed, tissue-enriched genes were found at low, variable levels in other organs and represented some of the most frequent causes of variation between samples of the same tissue type.

We found that contamination is best linked to the date of sequencing (linear regression model, p = 2.66e-75). However, both due to contamination being noted in some samples that are sequenced a few days apart from a possible contaminating source and the SNP-based evidence, we suspect the majority of the contamination occurred during library preparation rather than the sequencing itself. Library preparation dates were not documented (personal communication, GTEx Help Desk).

A variety of contamination causes have been reported, all of which could have had some role in our findings. Contamination during the collection of samples from individuals is possible, especially if non-disposable tools such as forceps are not cleaned properly in between collections [24]. During tissue manipulation, a “floater” or tiny piece of tissue could end up in the fixation kit (PAXgene) [24]. Although we did not see either type of contamination, it would be the hardest to prove due to the shared genotype.

While the nucleic acid isolation date was only modestly associated with contamination, physical contamination can easily occur at this stage. GTEx RNA isolation was manually done in batches of 12 tissues, purposefully with a mix of donors and tissues to minimize batch effects. Samples were individually cut and placed into cryovials for homogenization, followed by further manipulations [25].

At the stage of library preparation or sequencing where our data indicates most of the contamination occurred, there are multiple steps that could be implicated. The library preparation was completed automatically in 96 well plates with a mix of tissues and individuals to prevent batch effects [25]. Fluidic carryover could have occurred here. At the sequencing level, a major concern is index contamination where index oligonucleotides used for multiplexing can ligate to other sample transcripts, thus contaminating the data after demultiplexing. Index based contamination is machine and lane specific and can even occur at the creation of the indexes when multiple indexes are purified on the same high-performance liquid chromatography column [26].

Additionally, if steps to clean libraries of free adapters/primers are not properly executed, the remaining indexes can contaminate clusters in the flow cells [11]. Molecular recombination of indexes during sequencing can also lead to read misassignment as multiplex clusters can become contaminated by other samples that acquire the indices of the native sample (index hopping). GTEx's use of dual indices reduces the amount of index hopping that can occur [25, 26].

Using other sequencing datasets with similar sequencing methods, HPA and the pharmacogenetics study, we validated that it is contamination, not low-level transcription, which causes these unusual expression findings. This also shows the generalizability of this type of contamination regardless of the labs in which they take place.

So how big is the contamination problem? It depends on how the data is to be used. Fortunately, in the GTEx data, the levels are overall low with only 0.46% of samples having relatively high levels of *PRSS1*. Thus, for many uses of GTEx data, this level is irrelevant. However, for groups that are investigating differential expression in the GTEx dataset, these genes will repeatedly appear due to their variable levels of contamination. As well, we note that the GTEx standard normalization pipeline using PEER factors did not entirely eliminate this source of variation and an abundance of eQTLs that were identified for the 19 genes described herein were located in incorrect tissues (84%).

Many publications have reported rare, but variable gene expression in their samples claiming their importance or disease-related behaviors [27]. Our findings call these reports into question. The extent of cross-contamination, where one laboratories' samples get prepped and sequenced at the same time as a different laboratories' unrelated samples through a university core sequencing facility or sequencing company is unknown, but likely frequent [28, 29]. The xenomiR story, that rice miRNAs are found in human blood through dietary means [30], was shown to result from library preparation contamination [31, 32]. Also, our work supports that work flows must be considered carefully in very-low DNA mutation detection analysis in clinical cancer samples as samples with higher tumor burdens may contaminate samples with lower tumor burdens and falsely suggest treatment approaches [33, 34]. Specific to GTEx, their data is available in many outlets including the UCSC Genome Browser and variable, low-level expression of *PRSS1*, *CELA3A* and others may falsely intrigue researchers, particularly within the reported eQTLs.

Conclusion

We described low-level, variable expression contamination in the GTEx RNA-Seq dataset. The contamination was most noticeable for 19 highly-expressed, tissue-enriched genes. This contamination strongly correlates with the library preparation and sequencing of the samples. Similar contamination was observed in the HPA dataset, suggesting a universality to this type of contamination. Evaluating low-level variable gene expression in RNA-Sequencing data sets must be performed with precaution and awareness of potential sample contamination.

Methods

Retrieval of GTEx RNA-Seq dataset, FASTQ files, and sample Data

The gene read counts of the RNA-Seq GTEx version 7 dataset (GTEx_Analysis_2016-01-15_v7_RNASeQCv1.1.8_gene_reads.gct.gz) were downloaded from the GTEx Portal (https://gtexportal.org/home/datasets), along with the de-identified sample annotations (GTEx_v7_Annotations_SampleAttributesDS.txt). From dbGaP with the required permissions, the FASTQ files of the tissue samples and the variant call file (VCF) files of appropriate individuals were downloaded.

Retrieval of Human Protein Atlas tissue enriched gene list

We obtained the HPA tissue enriched genes by downloading a CSV file from this filtered site (https://www.proteinatlas.org/search/tissue_specificity_rna:any;Tissue%20enriched+AND+sort_by:tissue+specific+score, visited on 6/21/18).

Bulk sequencing processing

The acquired raw read counts were segmented into separate tissue subsets (48 tissues with ≥70 samples each) and their read counts were normalized using the Variance Stabilizing Transformation feature in DESeq2 version 1.22.1 in R version 3.5.1 [17]. This method incorporates estimated size factors based on the median-ratio method, and transformed by the dispersion-mean relationship. We then filtered the 56,202 genes based on their mean expression (mean transformed count > 5) to reduce noise and lessen the inflated effect of low expressing genes on correlations.

Identification of highly variable genes and clusters

All analyses were completed in R version 3.5.1 (2018/07/02). In each tissue a threshold of a >4 variance of normalized read counts was used as our cut off for highly variable transcripts. These genes were then clustered using hierarchal clustering on a distance generated by 1 - Kendall's rank-correlation coefficient. A tau critical value was calculated based on the number of samples and genes expressed. The correlation-based dendrogram was cut to produce gene clusters with average within cluster correlation of at least the tau critical value.

Calculation of average gene expression Z-Scores

Approximate z-scores were calculated by subtracting the mean expression and dividing by the median absolute deviation of the expression values for each gene across all samples within a given tissue. These Z-scores provide a standard measure of expression for all genes and allow one to summarize the expression of a gene cluster in a sample by the average Z-score of the genes in that cluster.

Base pair incongruency analysis

Base pair incongruency analysis required a contaminated tissue expression FASTQ, a native tissue expression FASTQ, and the individual’s VCF file. FASTQ files were mapped to the Genome Reference Consortium Human Build 37 (hg19) using the software HISAT2 version 2.1.0 [35]. The output SAM files were turned into BAM files and indexed using samtools version 1.9 [36, 37]. Preliminary analysis and development of figures were generated using the Integrative Genome Viewer version 2.4.13 [38, 39]). Protein coding SNPs, rare variants, and personal variants (collectively referred to as variants in this paper), were manually selected using IGV as a reference. Using the tool bam-readcount version 0.8.0 in combination with a Python 3.6.2 script, a list of RNA-Seq and genomic incongruencies were generated for the acquired sample BAM files.

PEER factor analysis

We obtained the GTEx RNA-Seq dataset from lung (N=427). The data underwent trimmed mean of m-values (TMM) normalization and filtering out of lowly expressed genes (< 0.1 TPM for 80% or more of the samples) before running PEER to identify potential confounders [20]. Following GTEx's pipeline (https://gtexportal.org/home/documentationPage#staticTextAnalysisMethods visited), we then performed an inverse normal transformation (INT) on the expression values for each gene in order to reduce the effect of outliers [21]. Z-scores for each gene are based on TMM-normalization, inverse-normal transformation, and scaling/centering at zero.

Cross-referencing eQTLs with contamination findings

We obtained and tallied eQTL reports for the 19 genes in Table 1 from the GTEx eQTL browser (https://gtexportal.org visited on March 26, 2019). eQTLs were identified by tissue association and conservatively placed in one of three categories: appropriate expression, possible expression, and inappropriate expression. The appropriateness of expression in any tissue was based on the evaluation of TPM levels in the tissue and immunohistochemistry staining patterns as noted in the Human Protein Atlas [40].

Acquiring Human Protein Atlas and Pharmacogenetic Study Variation RNA-Seq Data

Using the R package recount version 1.8.2, we downloaded HPA RNA-Seq data, accession ERP003613 [22], and the RNA-Seq data of a pharmacogenetic transcriptomic study, accession SRP060355 [23]. The HPA RNA-Seq was performed across 27 tissues including the pancreas and the pharmacogenetic RNA-Seq was across 4 tissues not including pancreas. We filtered samples down to only the shared tissues of liver, heart, and adipose.

Additional files

Additional File 1: Figure S1 A correlation heatmap of the highly variable gene clusters in 343 transformed fibroblast samples. Red shows a positive correlation. Genes within the contamination cluster are given. A, B and C represent other groups of co-variable genes.

Additional File 2: Table S1: A technical comparison of the GTEX1 fibroblast sample and its main contaminating GTEX2 esophagus sample.

Figure Legends –

Fig. 1 Identification and explanation of sequencing contamination **a** A correlation heatmap of highly variable subcutaneous adipose tissue genes across 442 subjects. Red shows a positive correlation. The genes within the contamination cluster and the sex cluster are given. Clusters A, B, and C represent other groups of co-variable genes. **b** Z-score values of non-pancreas tissue sample *PRSS1* reads coded by relationship to being sequenced on the same day as a pancreas tissue. (p<1.21e-67, linear model) over ~3 years. **c** Violin plot of the same data showing a strong, but not complete correlation of sequencing on a pancreas day. **d** Ranked order of all samples either sequenced on the same day as a pancreas sample (black) or on a non-pancreas sequencing day (colors) for *PRSS1* in log10. Among samples not sequenced on a pancreas day, 91% of samples with >100 reads were sequenced within 4 days of a

known sequenced pancreas. The dashed line represents 100 reads. **e** Contamination of GTEX1's fibroblast RNA-Seq predominately came from GTEX2. By DNA and RNA of the appropriate tissue source of *KRT4*, sample GTEX1 is homozygous for the C allele at rs7956809. The fibroblast sample is 87% G reads, primarily matching sample GTEX2. The read count depth at the SNP in the GTEX1 esophagus was 85,803 and 204 for the GTEX1 fibroblast.

Fig. 2 Impact of PEER factors on contamination. **a** The top two PEER factors separated in hospital from out of hospital deaths. **b** With no PEER factor correction there is a significant increase in *PNLIP* expression Z-scores in lung samples if sequenced on the same day as a pancreas (No = 96, Yes = 331; p= 4.34e-14). After 35 (p= 1.38e-11) or 60 (p= 3.03e-06) PEER factor corrections, the difference remained. **c** *PRSS1* contamination across three data sets. Only in data sets where pancreas was collected and sequenced (GTEx and HPA) are there notable contaminating PRSS*1* reads. Key: Pharma = Pharmacogenomics data set.

Declarations.

Acknowledgements – The authors thank the GTEx Help Desk for valuable assistance in understanding the technical data sets.

Funding - M.K.H. was supported by grants 1R01HL137811, R01GM130564, and P30CA006973 from the National Institutes of Health and 17GRNT33670405 from the American Heart Association. T.O.N. was supported by grants R01GM130564 and

T32GM07814. D.E.A was supported by 1R01HL131573 and 1R01HL137811. M.N.M. was supported by R01HL137811 and the University of Rochester CTSA award number UL1TR002001. A.Z.R was supported by R01GM130564.

Availability of data and material – All data used in this study is available through dbGap or recount2.

Authors contributions – M.K.H., M.N.M and A.Z.R conceived of the experiments and assisted with the manuscripts. T.O.N. performed the experiments, analyzed the data and wrote the manuscript. S.Y., V.P. and D.E.A. performed experiments and assisted on the manuscript.

Ethics Approval – All human data was publicly available or used with approval of the GTEx consortium. Consent was obtained by those studies.

Competing interests – The authors declare no competing interests.

References

1. Consortium GT: **The Genotype-Tissue Expression (GTEx) project**. *Nat Genet* 2013, **45**(6):580-585.
2. Tomczak K, Czerwinska P, Wiznerowicz M: **The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge**. *Contemporary oncology* 2015, **19**(1A):A68-77.
3. Zeisel A, Munoz-Manchado AB, Codeluppi S, Lonnerberg P, La Manno G, Jureus A, Marques S, Munguba H, He L, Betsholtz C *et al*: **Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq**. *Science* 2015, **347**(6226):1138-1142.
4. Kumasaka N, Knights AJ, Gaffney DJ: **Fine-mapping cellular QTLs with RASQUAL and ATAC-seq**. *Nat Genet* 2016, **48**(2):206-213.
5. Gutman DA, Cooper LA, Hwang SN, Holder CA, Gao J, Aurora TD, Dunn WD, Jr., Scarpace L, Mikkelsen T, Jain R *et al*: **MR imaging predictors of molecular profile and survival: multi-institutional study of the TCGA glioblastoma data set**. *Radiology* 2013, **267**(2):560-569.

6. Ewing AD, Houlahan KE, Hu Y, Ellrott K, Caloian C, Yamaguchi TN, Bare JC, P'ng C, Waggott D, Sabelnykova VY *et al*: **Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection**. *Nature methods* 2015, **12**(7):623-630.

7. Okoniewski MJ, Miller CJ: **Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations**. *BMC Bioinformatics* 2006, **7**:276.

8. van Dijk EL, Jaszczyszyn Y, Thermes C: **Library preparation methods for next-generation sequencing: tone down the bias**. *Experimental cell research* 2014, **322**(1):12-20.

9. Tuerk A, Wiktorin G, Guler S: **Mixture models reveal multiple positional bias types in RNA-Seq data and lead to accurate transcript concentration estimates**. *PLoS computational biology* 2017, **13**(5):e1005515.

10. Lusk RW: **Diverse and widespread contamination evident in the unmapped depths of high throughput sequencing data**. *PLoS One* 2014, **9**(10):e110808.

11. **Effects of Index Misassignment on Multiplexing and Downstream Analysis** [https://www.illumina.com/content/dam/illumina-marketing/documents/products/whitepapers/index-hopping-white-paper-770-2017-004.pdf]

12. Rosenberg AZ, Wright C, Fox-Talbot K, Rajpurohit A, Williams C, Porter C, Kovbasnjuk O, McCall MN, Shin JH, Halushka MK: **xMD-miRNA-seq to generate near in vivo miRNA expression estimates in colon epithelial cells**. *Scientific reports* 2018, **8**(1):9783.

13. Merchant S, Wood DE, Salzberg SL: **Unexpected cross-species contamination in genome sequencing projects**. *PeerJ* 2014, **2**:e675.

14. Cibulskis K, McKenna A, Fennell T, Banks E, DePristo M, Getz G: **ContEst: estimating cross-contamination of human samples in next-generation sequencing data**. *Bioinformatics* 2011, **27**(18):2601-2602.

15. Ma X, Shao Y, Tian L, Flasch DA, Mulder HL, Edmonson MN, Liu Y, Chen X, Newman S, Nakitandwe J *et al*: **Analysis of error profiles in deep next-generation sequencing data**. *Genome Biol* 2019, **20**(1):50.

16. McCall MN, Illei PB, Halushka MK: **Complex Sources of Variation in Tissue Expression Data: Analysis of the GTEx Lung Transcriptome**. *American journal of human genetics* 2016, **99**(3):624-635.

17. Love MI, Huber W, Anders S: **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2**. *Genome Biol* 2014, **15**(12):550.

18. Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson A, Kampf C, Sjostedt E, Asplund A *et al*: **Proteomics. Tissue-based map of the human proteome**. *Science* 2015, **347**(6220):1260419.

19. **The human tissue specific proteome** [https://www.proteinatlas.org/humanproteome/tissue/tissue+specific]

20. Stegle O, Parts L, Piipari M, Winn J, Durbin R: **Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses**. *Nature protocols* 2012, **7**(3):500-507.

21. Consortium GT, Laboratory DA, Coordinating Center -Analysis Working G, Statistical Methods groups-Analysis Working G, Enhancing Gg, Fund NIHC, Nih/Nci, Nih/Nhgri, Nih/Nimh, Nih/Nida *et al*: **Genetic effects on gene expression across human tissues**. *Nature* 2017, **550**(7675):204-213.

22. Fagerberg L, Hallstrom BM, Oksvold P, Kampf C, Djureinovic D, Odeberg J, Habuka M, Tahmasebpoor S, Danielsson A, Edlund K *et al*: **Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics**. *Molecular & cellular proteomics : MCP* 2014, **13**(2):397-406.

23. Chhibber A, French CE, Yee SW, Gamazon ER, Theusch E, Qin X, Webb A, Papp AC, Wang A, Simmons CQ *et al*: **Transcriptomic variation of pharmacogenes in multiple human tissues and lymphoblastoid cell lines**. *The pharmacogenomics journal* 2017, **17**(2):137-145.

24. Sehn JK, Spencer DH, Pfeifer JD, Bredemeyer AJ, Cottrell CE, Abel HJ, Duncavage EJ: **Occult Specimen Contamination in Routine Clinical Next-Generation Sequencing Testing**. *American journal of clinical pathology* 2015, **144**(4):667-674.

25. Consortium GT: **Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans**. *Science* 2015, **348**(6235):648-660.

26. Kircher M, Sawyer S, Meyer M: **Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform**. *Nucleic Acids Res* 2012, **40**(1):e3.

27. Witwer KW, Halushka MK: **Toward the promise of microRNAs - Enhancing reproducibility and rigor in microRNA research**. *RNA biology* 2016, **13**(11):1103-1116.

28. Kryukov K, Imanishi T: **Human Contamination in Public Genome Assemblies**. *PLoS One* 2016, **11**(9):e0162424.

29. Longo MS, O'Neill MJ, O'Neill RJ: **Abundant human DNA contamination identified in non-primate genome databases**. *PLoS One* 2011, **6**(2):e16410.

30. Zhang L, Hou D, Chen X, Li D, Zhu L, Zhang Y, Li J, Bian Z, Liang X, Cai X *et al*: **Exogenous plant MIR168a specifically targets mammalian LDLRAP1: evidence of cross-kingdom regulation by microRNA**. *Cell Res* 2012, **22**(1):107-126.

31. Tosar JP, Rovira C, Naya H, Cayota A: **Mining of public sequencing databases supports a non-dietary origin for putative foreign miRNAs: underestimated effects of contamination in NGS**. *RNA* 2014, **20**(6):754-757.

32. Zhang Y, Wiggins BE, Lawrence C, Petrick J, Ivashuta S, Heck G: **Analysis of plant-derived miRNAs in animal small RNA datasets**. *BMC Genomics* 2012, **13**:381.

33. Tian SK, Killian JK, Rekhtman N, Benayed R, Middha S, Ladanyi M, Lin O, Arcila ME: **Optimizing Workflows and Processing of Cytologic Samples for Comprehensive Analysis by Next-Generation Sequencing: Memorial Sloan Kettering Cancer Center Experience**. *Archives of pathology & laboratory medicine* 2016.

34. Van Allen EM, Wagle N, Stojanov P, Perrin DL, Cibulskis K, Marlow S, Jane-Valbuena J, Friedrich DC, Kryukov G, Carter SL *et al*: **Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine**. *Nature medicine* 2014, **20**(6):682-688.

35. Kim D, Langmead B, Salzberg SL: **HISAT: a fast spliced aligner with low memory requirements**. *Nature methods* 2015, **12**(4):357-360.

36. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S: **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics* 2009, **25**(16):2078-2079.

37. Li H: **A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data**. *Bioinformatics* 2011, **27**(21):2987-2993.

38. Thorvaldsdottir H, Robinson JT, Mesirov JP: **Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration**. *Briefings in bioinformatics* 2013, **14**(2):178-192.

39. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP: **Integrative genomics viewer**. *Nature biotechnology* 2011, **29**(1):24-26.

40. Uhlen M, Oksvold P, Fagerberg L, Lundberg E, Jonasson K, Forsberg M, Zwahlen M, Kampf C, Wester K, Hober S *et al*: **Towards a knowledge-based Human Protein Atlas**. *Nature biotechnology* 2010, **28**(12):1248-1250.

Figure 1

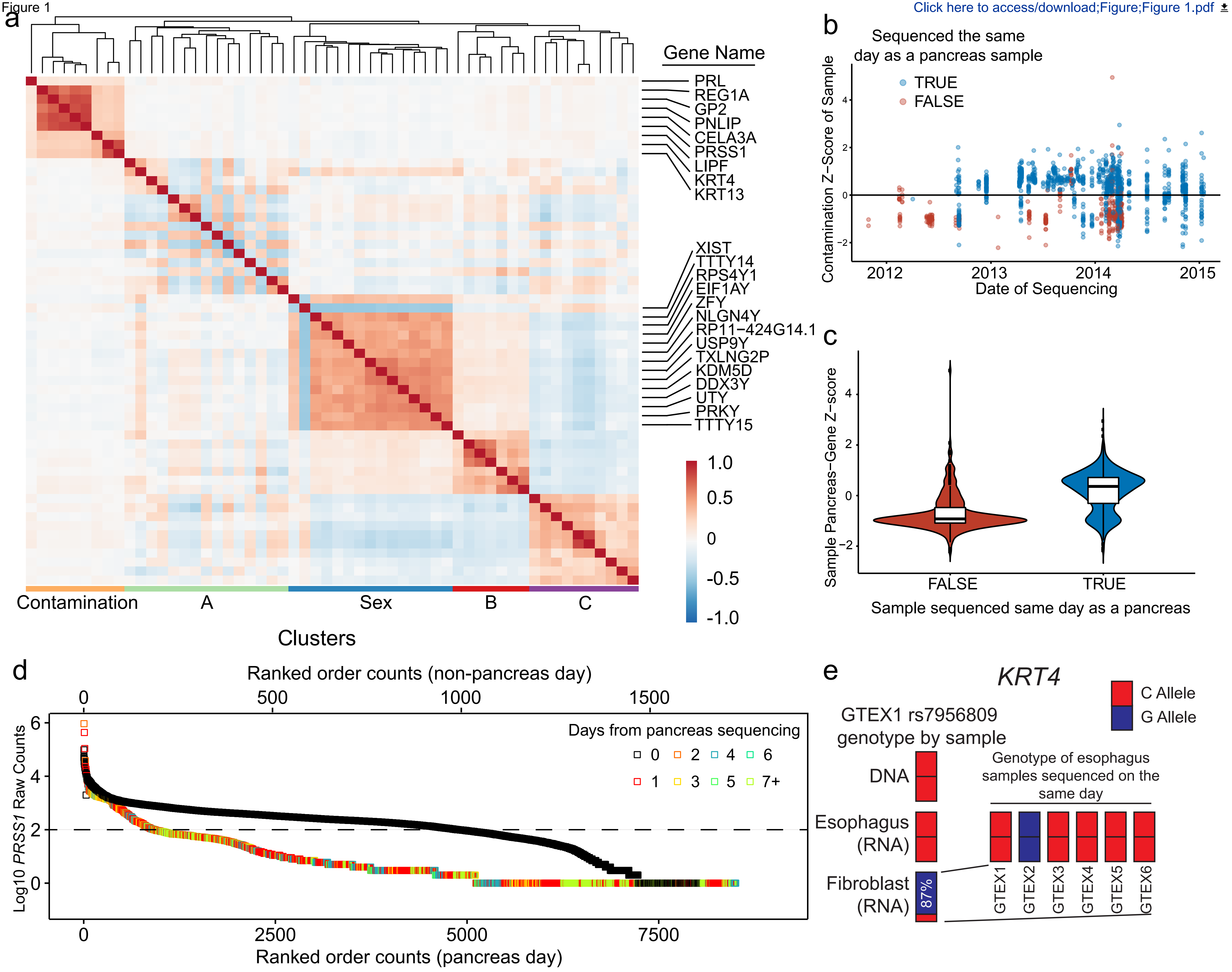
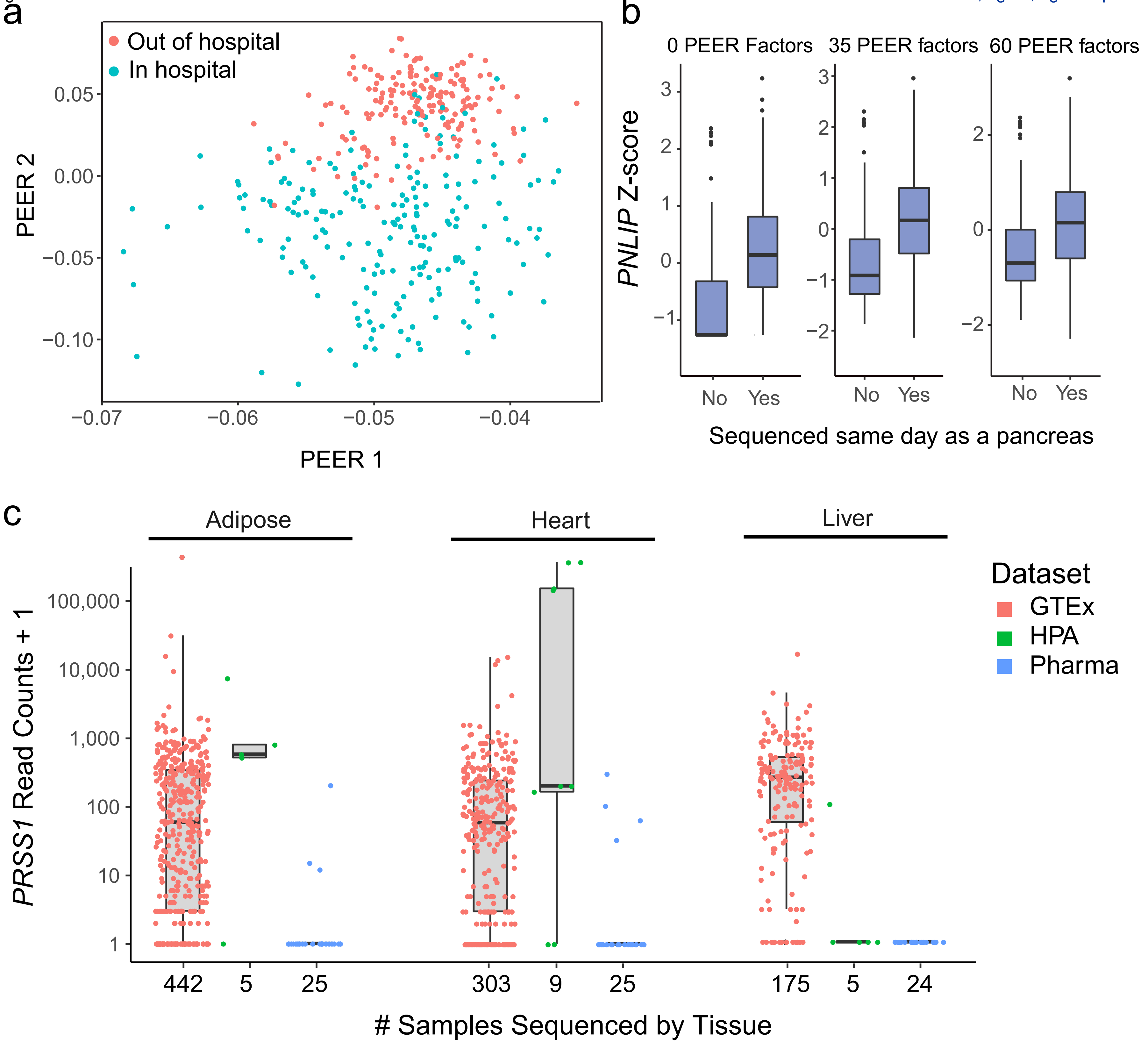

a
Gene Name
PRL
REG1A
GP2
PNLIP
CELA3A
PRSS1
LIPF
KRT4
KRT13
XIST
TTTY14
RPS4Y1
EIF1AY
ZFY
NLGN4Y
RP11−424G14.1
USP9Y
TXLNG2P
KDM5D
DDX3Y
UTY
PRKY
TTTY15
1.0
0.5
0
-0.5
-1.0
Contamination
A
Sex
B
C
Clusters
b
Sequenced the same
day as a pancreas sample
TRUE
FALSE
Contamination Z−Score of Sample
4
2
0
−2
2012
2013
2014
2015
Date of Sequencing
c
Sample Pancreas−Gene Z−score
4
2
0
−2
FALSE
TRUE
Sample sequenced same day as a pancreas
d
Ranked order counts (non-pancreas day)
0
500
1000
1500
Log10 PRSS1 Raw Counts
6
4
2
0
Days from pancreas sequencing
0
2
4
6
1
3
5
7+
0
2500
5000
7500
Ranked order counts (pancreas day)
e
KRT4
C Allele
G Allele
GTEX1 rs7956809
genotype by sample
DNA
Esophagus
(RNA)
Fibroblast
(RNA)
87%
Genotype of esophagus
samples sequenced on the
same day
GTEX1
GTEX2
GTEX3
GTEX4
GTEX5
GTEX6

Figure 2


a
Out of hospital
In hospital
0.05
0.00
−0.05
−0.10
PEER 2
−0.07
−0.06
−0.05
−0.04
PEER 1
b
0 PEER Factors
35 PEER factors
60 PEER factors
PNLIP Z-score
3
2
1
0
−1
−2
No
Yes
Sequenced same day as a pancreas
c
Adipose
Heart
Liver
PRSS1 Read Counts + 1
100,000
10,000
1,000
100
10
1
442
5
25
303
9
25
175
5
24
# Samples Sequenced by Tissue
Dataset
GTEx
HPA
Pharma