Basal Contamination of Sequencing: Lessons from the GTEx dataset

Tim O. Nieuwenhuis[1,2], Stephanie Yang[2], Rohan X. Verma[1], 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, Department 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

rverma6@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

One of the challenges of next generation sequencing (NGS) is read contamination. We used the Genotype-Tissue Expression (GTEx) project, a large, diverse, and robustly generated dataset, to understand the factors that contribute to contamination. We obtained GTEx datasets and technical metadata and validating RNA-Seq from other studies. Of 48 analyzed tissues in GTEx, 26 had variant co-expression clusters of four known highly expressed and pancreas-enriched genes (*PRSS1*, *PNLIP*, *CLPS*, and/or *CELA3A*). Fourteen additional highly expressed genes from other tissues also indicated contamination. Sample contamination by non-native genes was associated with a sample being sequenced on the same day as a tissue that natively expressed those genes. This was highly significant for pancreas and esophagus genes (linear model, p=9.5e-237 and p=5e-260 respectively). Nine SNPs in four genes shown to contaminate non-native tissues demonstrated allelic differences between DNA-based genotypes and contaminated sample RNA-based genotypes, validating the contamination. Low-level contamination affected 4,497 (39.6%) samples (defined as ≥10 *PRSS1* TPM). It also led to eQTL assignments in inappropriate tissues among these 18 genes. We note this type of contamination occurs widely, impacting bulk and single cell data set analysis. In conclusion, highly expressed, tissue-enriched genes basally contaminate GTEx and other datasets impacting analyses. Awareness of this process is necessary to avoid assigning inaccurate importance to low-level gene expression in inappropriate tissues and cells.

Key Words: GTEx, RNA-Seq, Contamination, scRNA-Seq, 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 exist for all –omics technologies, including 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]. For single cell RNA-Seq (scRNA-Seq), doublets or multiplets (2+ cells partitioned together) can cause cross-contamination resulting in expression hybrids[13]. Compared to RNA-Seq, contamination has been better characterized for DNA sequencing projects[14-16].

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 identifies eQTLs by associating genotypes called from whole genome

sequencing with gene expression levels obtained 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 bulk RNA-Seq data. We further demonstrate the universality of highly-expressed genes contaminating other samples.

Results

Patterns of extreme tissue variation identified usual and unusual gene signatures

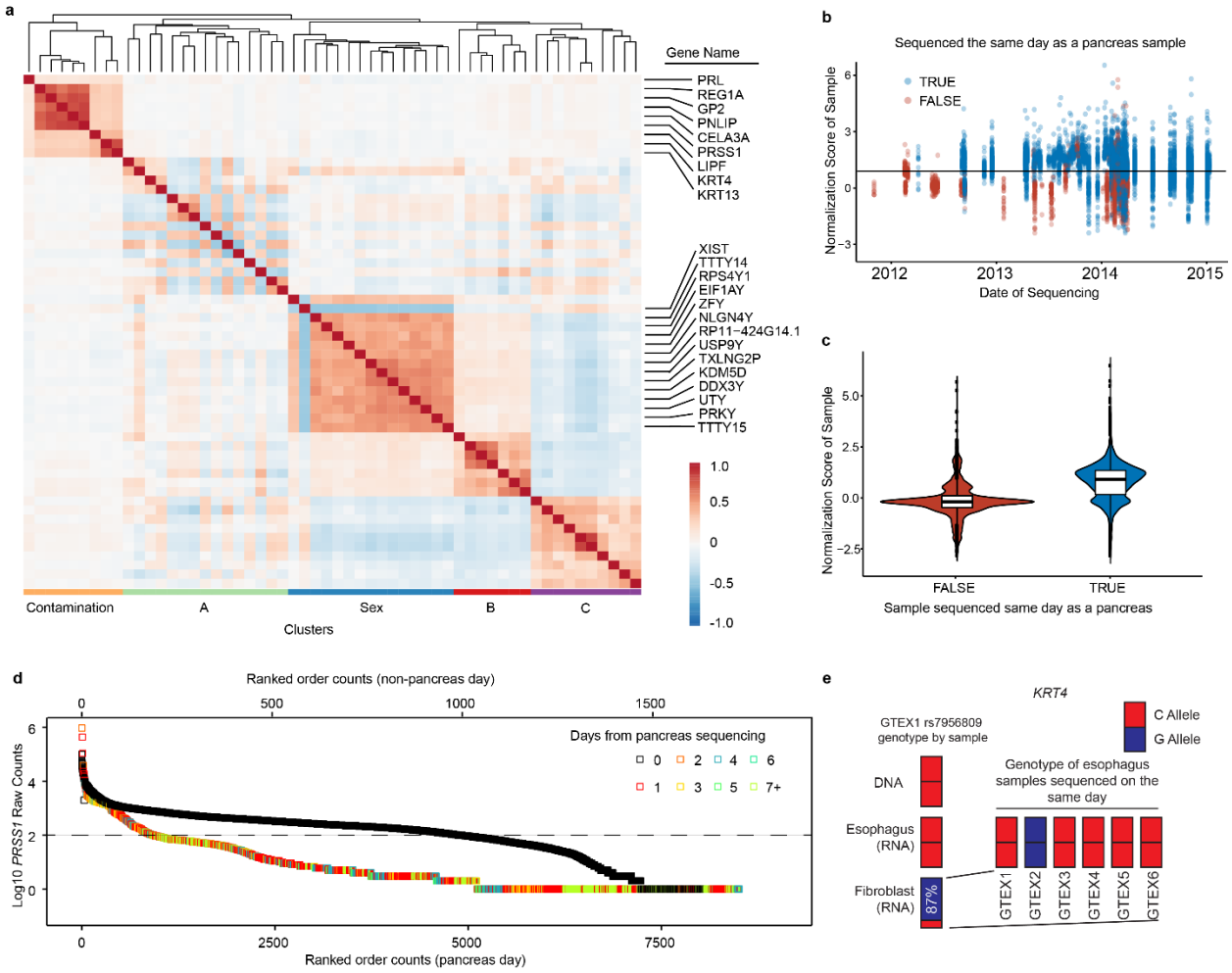
We embarked on a project to expand our initial description of the causes of lung expression variation in GTEx to all tissue samples [17]. We used DESeq2 variance stabilizing transformation to normalize read counts from 11,340 samples[17,18]. Then we filtered genes in each tissue keeping those with a 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 V7 [17]. Our algorithm identified multiple gene clusters per tissue, based on their Kendall's tau correlations (Fig. 1a). 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 cluster of 3-4 genes (*PNLIP, PRSS1*, *CELA3A,* and/or *CLPS)* identified in 26 of the 48 tissues that did not have an intuitive biological explanation. These genes are highly-expressed and pancreas acinar-cell specific[19]. To identify other highly expressed tissue-enriched genes appearing variably in other samples, we cross-referenced a list of tissue-enriched proteins generated by the Human Protein Atlas (HPA) to the GTEx transcripts per million (TPM) data (Table 1)[20,21]. We noted 18 genes from seven tissues including two esophagus genes *KRT13* and *KRT4* that are highly expressed in their native tissue and identified as variable in five or more other unrelated tissues (Fig. 1a, Supplementary Fig. 1).

Table 1 Eighteen highly expressed genes that appear to contaminate five or more GTEx tissues

Gene	Times identified as variable in other tissues	Highest expressed GTEx tissue	GTEx TPM	2nd Highest expressed GTEx Tissue		
				Tissue	GTEx TPM	Independent Study TPM
PRSS1	43	Pancreas	99,100	Stomach	28.36	0.92
PNLIP	34	Pancreas	33,660	Ovary	3.67	0
CPA1	31	Pancreas	54,500	Testis	8.16	0
GP2	29	Pancreas	14,280	Prostate	18.34	13.5
CELA3A	24	Pancreas	27,130	Stomach	14.1	0.03
KRT13	20	Esophagus	33,960	Vagina	13,140	3,961
PGC	19	Stomach	36,720	Lung	83.84	12.42
KRT4	18	Esophagus	22,290	Vagina	1,375	15,069
PRL	17	Pituitary	54,500	Testis	5.15	0
LIPF	15	Stomach	29,380	Testis	4.33	0
CLPS	14	Pancreas	51,640	Stomach	5.35	0
CTRB2	9	Pancreas	20,760	Ovary	3.21	0
MYBPC1	6	Skeletal muscle	3,587	Prostate	54.5	21.3

MYH2	6	Skeletal muscle	1,064	Colon	0.52	0.02
ZG16B	6	Salivary gland	17,540	Prostate	48.1	187.1
FGA	5	Liver	5,717	Stomach	12.43	6.02
HP	5	Liver	12,710	Adipose	140.1	2.7
CKM	5	Skeletal muscle	11,138	Heart	2,987	2,339
Independent study samples were taken from RNA-seq experiments of the 2nd highest GTEx tissue and were not co-sequenced with the highest expressed GTEx tissue.						



115

116

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. The meaning of cluster A is unknown. Cluster B may relate to the percentage of smooth muscle cells and cluster C includes acute phase reactants **b** Contamination normalized score values for non-pancreas tissue samples (N=11,366) colored relative to being sequenced on the same day as a pancreas tissue. **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* read counts 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** *KRT4* contaminating reads in GTEX1's fibroblast RNA-Seq appear to have originated from GTEX2 esophagus mucosa tissue. 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.

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.

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. 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, challenging this as a source of contamination (Supplementary Fig. 1).

Linear mixed models reveal sequencing date as the source of contamination

We then queried if other technical sources of contamination may be related. Other GTEx metadata available included nucleic acid isolation date and sequencing date, which were both interrogated for associations with this contamination. The normalized score of four pancreas-enriched genes (*PRSS1*, *CLPS*, *PNLIP*, and *CELA3A*) was significantly higher in non-pancreas samples if they were sequenced on the same day as a pancreas sample (Wilcox Rank Sum Test, $p < 5e\text{-}324$, Fig 1b,c). We then performed linear mixed model analyses across non-contaminating tissues to understand expression levels of genes known to be highly-enriched and highly-expressed in the pancreas or esophagus, but were found lowly and variably in other tissues (ex. coronary artery, liver, tibial nerve). Because the majority of samples were sequenced on the same day as a pancreas or esophagus – mucosa sample, for model robustness, we limited our models to only tissues with >40 samples not sequenced on the same data as a respective contaminating tissue (Supplementary Data 1 and 2).

After adjusting for tissue type, the pancreas contamination score (average of normalized expression values for *PRSS1*, *CLPS*, *PNLIP*, and *CELA3A*) was on average 0.175 higher if the tissue was isolated on the same day a pancreas sample ($p=3.9e\text{-}12$), but on average 0.863 higher if the tissue was sequenced on the same day as a pancreas sample ($p=9.5e\text{-}237$) (Supplementary Data 3).

A similar linear mixed model was used to evaluate the esophagus contamination score (average of normalized expression values for two highly-expressed esophagus-enriched genes *KRT4* and *KRT13*). The same model as above showed only sequencing a tissue on the same day as a esophagus sample to be significant (p = 5e-260) (Supplementary Data 4).

Despite the significance of with sequencing date, some high scores came from samples that were not sequenced on the same days as pancreata. Focusing on just one these genes, *PRSS1*, it was clear that 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) as a potential source for contamination, as it is temporally related, but not identical, to sequencing date.

With this understanding of the temporal importance of sequencing date to gene expression, we then revisited the four genes making up the pancreas normalized score gene cluster. For all tissues, median normalized scores were higher on pancreas sequencing dates, and the division between the 26 tissues with clustering vs. the 21 without was a result of the high threshold for TPM we set at the VST step of our pipeline (Supplementary Figure 2).

Genetic polymorphisms confirm contamination is derived from other samples

To prove that pancreas/esophagus transcripts were contaminating from other (non-self) samples we identified 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 RNA-Seq files

from contaminating tissues, in order to account for both RNA editing and preferential allele expression. Based on these sample requirements and limited by available raw sequencing files, we identified 11 contaminated tissues to evaluate. For each, we obtained and processed their raw RNA-Seq FASTQ sequences to identify nucleotide 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 the individual's 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 inconsistencies 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-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-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-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 individual (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 (Supplementary Data 5).

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

After establishing that contamination exists in GTEx by identification of a temporal association and validation through polymorphisms, 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 (Fig. 2). In the 11,366 non-pancreas samples investigated, roughly 25% had 0 reads for each of 4 pancreas-enriched genes and another ~50% of samples had TPM <10.

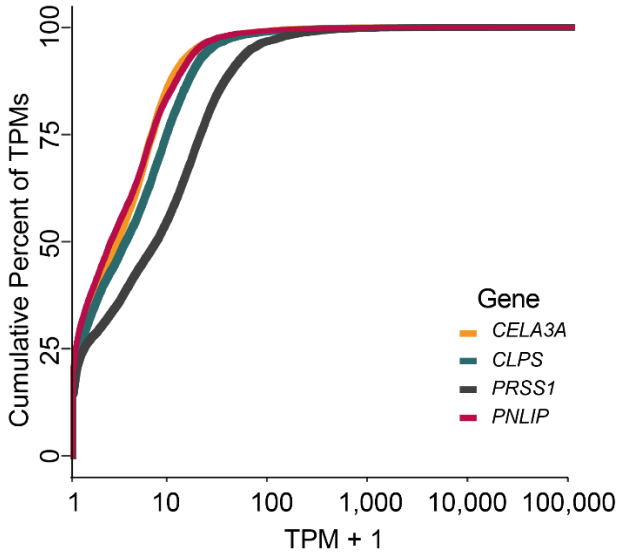


Fig. 2 A density plot of 11,366 non-pancreas RNA-seq samples and their cumulative TPM expression of four pancreas genes.

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[22,23]. 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 capturing a difference between “in hospital” (short postmortem interval) and “outside of hospital” (longer postmortem interval) deaths (Fig. 3a). This relationship is consistent with our prior report of variation in lung[17]. 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. 3b). Indeed, of five genes evaluated, only one gene (*KRT4*) was fully corrected for by PEER factors (Table 3). We then explored if this lack of full correction impacted eQTL analysis in the GTEx program.

Table 3. 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	1.39e-07	1.94e-05	0.49
PRSS1	6.29e-14	5.83e-07	2.63e-05	0.48
CELA3A	5.91e-14	3.36e-07	2.41e-05	0.49
KRT4	0.0034	0.17	0.15	0.18
KRT13	8.18e-17	0.00048	0.0036	0.37

P. values are shown before and after PEER correction.

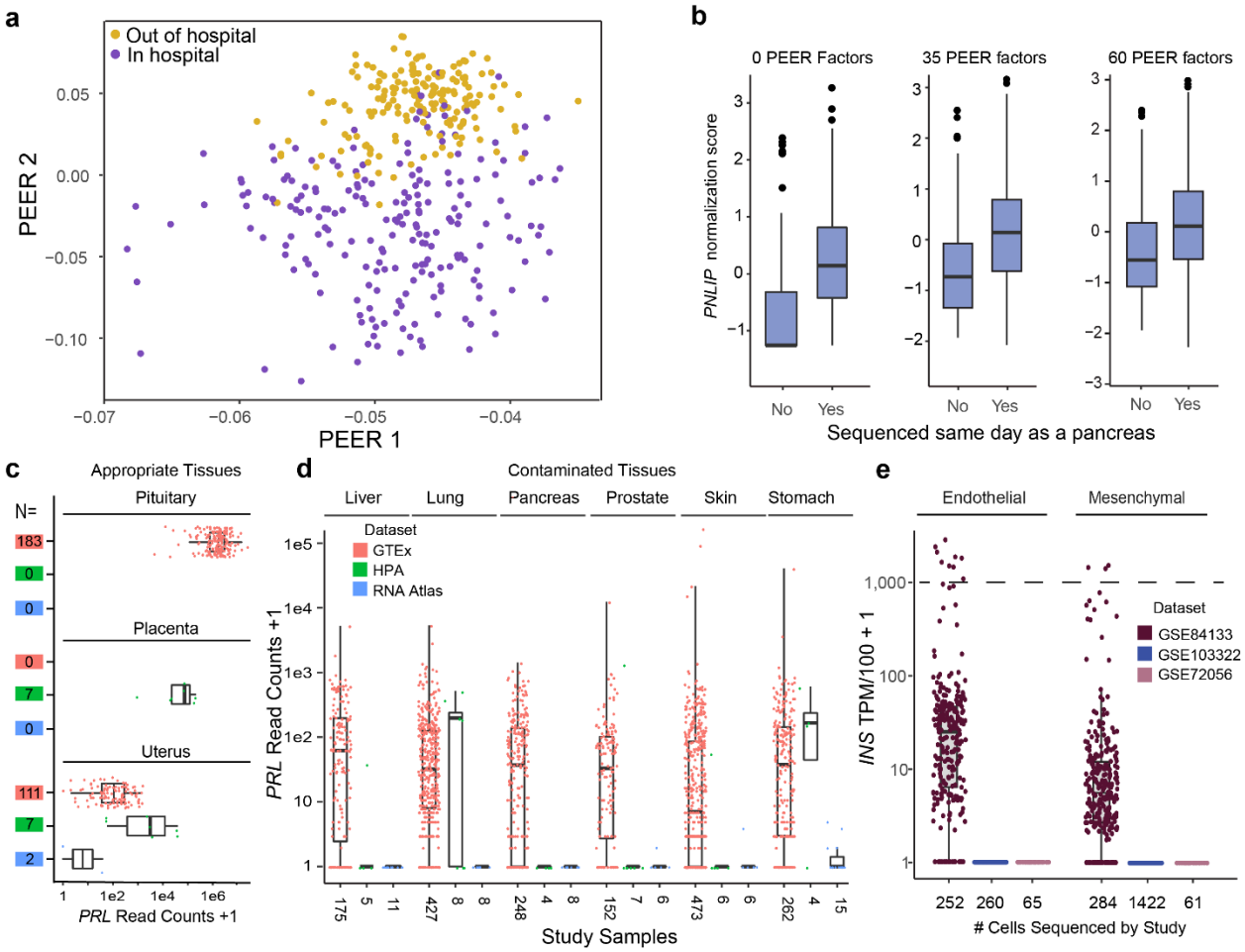


Fig. 3 Impact of PEER factors on contamination. **a** The top two PEER factors separated in hospital from out of hospital deaths (N=427). **b** With no PEER factor correction there is a significant increase in *PNLIP* expression normalized scores in lung samples if they were sequenced on the same day as a pancreas (No = 96, Yes = 331; linear regression, p= 4.34e-14). After 35 (p= 1.38e-11) or 60 (p= 3.03e-06) PEER factor corrections, the difference remained. **c** *PRL* read counts in pituitary (high), placenta (medium), and uterus (low), where *PRL* is known to be expressed across GTEx, HPA, and the RNA Atlas. The numbers in colored boxes indicate sample sizes and the color

indicates respective study. **d** *PRL* contamination reads across 6 tissues from 3 studies that correlate with levels of likely contamination based on the other sequenced organs. **e** *INS* contamination across three scRNA-Seq data sets. Only in the pancreas dataset (GSE84133), where beta cells were also sequenced, does *INS* appear to be lowly expressed in endothelial and mesenchymal cells. Cells with expression above the dotted line at 1,000 TPM/100 are likely doublets or multiplets.

Contamination affects GTEx eQTL reporting

Using the GTEx eQTL browser, we identified 72 tissues reported as having significant eQTLs for the 18 genes listed in Table 1. Seven tissues matched the known dominant expression patterns of the genes. An additional 34 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, 31 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 4).

Table 4 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, heart	Brain, esophagus mucosa
CKM	--	--	Aorta, whole blood
MYBPC1	--	Heart, testis, colon, prostate, brain (2)	Esophagus (2), lung, thyroid
MYH2	--	--	Colon, lung
ZG16B	--	Skin (2), adrenal, fibroblasts, lung, stomach, prostate, spleen, colon, testis, whole blood	Adipose, esophagus, pituitary

Non-GTEx bulk sequencing data sets confirm contamination

To determine if highly-expressed tissue-enriched contamination is a feature of sequencing in general, we identified two additional RNA-Seq datasets containing multiple organs/tissues[24,25,26]. Neither study sequenced the pituitary, which is the organ with the highest levels of prolactin (*PRL*) expression (Fig. 3c). Both studies performed multiplexed sequencing on Illumina 2000 or 3000 sequencers. These data sets demonstrate the extent of *PRL* expression contamination across six tissues is dependent on the amount of *PRL* expressed in the appropriate tissues (GTEx pituitary median 54,500 TPM and uterus median 4.01 TPM (Fig. 3d). To additionally characterize the breadth of contamination in the literature, we identified ten studies with two or more organs/cells sequenced together. All ten had cross-sample contamination of the type described herein (Supplementary Data 6). We also analyzed the METSIM study, containing bulk sequencing of adipose tissue in 434 subjects.[27] The data from that study demonstrates how, in the absence of sequencing at the same time as a contaminating tissue (ex. stomach, pancreas), contaminating reads are minimal do not correlate. This was opposite that observed in GTEx adipose tissue (Supplementary Fig. 3).

scRNA-Seq may be similarly impacted by this type of contamination.

We then investigated this type of highly-expressed gene contamination in scRNA-Seq. Although doublets are well-characterized, and barcode swapping has been reported, other forms of contamination are not well-described[13,28]. Also, due to the low depth of sequencing in current scRNA-seq compared to bulk sequencing, the type of contamination concerned herein would not be expected to generate large numbers of contaminating reads. However, investigations across two cell types, endothelial cells and mesenchymal cells (fibroblasts, activated stellate cells, cancer-associated fibroblasts), suggested otherwise.

Three different scRNA-Seq data sets (GSE84133, GSE103322, GSE72056) from normal pancreas, metastatic melanoma, and head and neck squamous cell carcinoma were used[29-31]. From the pancreas sample, insulin (*INS*) represented a highly-expressed, cell type-enriched gene that is exclusive to beta cells (TPM value of 167,144; transcripts per 10,000 cells = 1,671). *INS* expression was modestly elevated in endothelial cells and mesenchymal cells from the same data set (GSE84133), but is absent in endothelial and mesenchymal cells from the other data sets in which beta cells were not sequenced (Fig. 3e). Contamination of pancreatic endothelial cells is further supported by a recent snATAC-seq study that demonstrated closed chromatin at the *INS* locus in pancreatic endothelial cells[32].

scRNA-seq contamination drives aberrant cell clustering

Tabula Muris is a single cell transcriptome atlas of 20 mouse tissues[33]. We gathered endothelial cell data from 9 tissues and generated tSNE plots of this data based on adjustment by either 6 or 10 principal components (PCs) (Supplementary Fig. 4a, b). With 10 PCs, a subset of pancreatic endothelial cells incorrectly clustered separately as a result of contaminating pancreatic acinar genes that drive PCs 8 and 10. HPA staining validated the localization of these proteins to acinar cells (Supplementary Fig. 4c).

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 48 tissues spanning 10,294 samples, we detected a variable signal of pancreas genes in 26 of those tissues. From there we noticed genes that were highly expressed in esophagus, stomach, pituitary and other tissues also appeared 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 for both pancreas genes and esophagus genes (linear model, p=1.3e-14 and p=6.1e-49 respectively 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).

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 [34].

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 [34]. 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 [35]. 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 faulty read assignment as multiplex clusters can become contaminated by other samples that acquire the indices of the native sample (index hopping). This has been demonstrated in scRNA-Seq, and seems to occur more often on a HiSeq 4000 platform[28].

Specific to this study, GTEx samples were run on either HiSeq 2000 or 2500 machines in high output mode using cbot clustering. The flowcells were not patterned, and therefore not prone to index-swapping (personal communication, GTEx Help Desk). GTEx's use of custom i7 (later Illumina kit-based) dual indices also reduced the amount of index hopping that can occur[34,35].

Contamination of scRNA-Seq, may or may not be occurring by yet an additional mechanism. Ambient RNA, from disrupted ("broken") cells may be a major source of contamination, as recently described[36]. If true, this would occur before the library preparation and sequencing steps, would cause only a modest expression of the gene, and would not be detectable by current doublet correcting mechanisms[13,37]. This method of contamination might have implications for Human Cell Atlas studies, as we demonstrated it clearly affects cross-tissue comparisons of common cell types in Tabula Muris[33,38].

Using other bulk RNA-Seq data sets with similar sequencing methods and scRNA-Seq data sets, with different preparation methods, we validated that it is predominately contamination, not low-level transcription, which results in non-zero expression values in inappropriate tissues/cells (Fig. 3d,e, Supplementary Fig. 3, Supplementary Fig. 4, Supplementary Data 6). These extra data sets also uncover the generalizability of this highly-expressed transcript contamination regardless of the labs in which they take place or the methods employed.

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 2.85% of samples having relatively high levels of *PRSS1* (TPM > 1,00). Thus, for many uses of

GTEx data, this level is irrelevant. However, for experiments that involve differential expression profiling, these genes will repeatedly appear due to their variable levels of contamination. We additionally note that the GTEx standard normalization pipeline using PEER factors did not entirely eliminate this source of variation and numerous eQTLs that were identified for the 18 genes described herein were located in incorrect tissues (>40%). We caution that our results do not suggest that one should design a study in which sample type (e.g. tissue type or disease state) is perfectly confounded with library preparation or sequencing date. In fact, it was only because the GTEx study sequenced samples from different individuals on the same date that we were able to definitively show that contamination between samples occurred.

In other scenarios, this basal contamination may be more worrisome. Many publications have reported rare, but variable gene expression in their samples claiming their importance or disease-related behaviors[39]. Our findings call these reports into question. The extent of cross-contamination, where one laboratories' samples are 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[40,41]. The xenomiR story, that rice miRNAs are found in human blood through dietary means[42], was shown to result from library preparation contamination[43,44]. For scRNA-Seq, this contamination falsely implies some gene expression is ubiquitous across cells, which influences computational heterogeneity analyses of cell types, beyond the known challenges of other biological and technical artifacts[33,36,45]. Also, our work highlights the fact 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[46,47]. In particular, GTEx data is available in many outlets, including the UCSC Genome Browser. Variable, low-level expression of *PRSS1*, *CELA3A* and others may falsely intrigue researchers, particularly within the reported eQTLs.

We described low-level, variable expression contamination in the GTEx RNA-Seq dataset. This variation was most noticeable for 18 highly-expressed, tissue-enriched genes and strongly correlates with the library preparation and sequencing of the samples. Similar contamination was observed in other bulk and scRNA-Seq data sets 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

Ethics Statement

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

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). The FASTQ files of the

tissue samples and the variant call file (VCF) files of appropriate individuals were downloaded from dbGaP (phs000424.v7.p2) with the required permissions.

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).

Second highest tissue expression of tissue enriched genes and expression in single tissue studies

Using the GTEx portal (v7) we noted the TPM for the second highest expressing tissue with greater than 60 samples. We then used GEO, SRA Run Selector, and Recount2 (https://jhubiostatistics.shinyapps.io/recount/) to find datasets where only one tissue was sequenced, with the exception of the ovary sample which was obtained from the Illumina BodyMap (16 tissue types, absent of pancreas)[48]. Due to our inability to find salivary gland data we used prostate tissue, the third highest expressing tissue for *ZG16B*. Heart (ERP009437), adipose tissue (SRP053101), lung (SRP032833), prostate (SRP003611), stomach (ERP010889), and colon (SRP029880) TPMs were acquired through Recount2 and were selected for normal samples. The testis data was downloaded from GEO Omnibus (GSE103905). Vagina and ovary FASTQ data was downloaded from SRA run selector (GSE68229 and GSE30611). FASTQ files were mapped to the Genome Reference Consortium Human Build 38 (GRCh38) using the software HISAT2 version 2.1.0 [49]. The output SAM files were turned into BAM files and

indexed using samtools version 1.9 [50,51]. Assembly was completed using StringTie v1.3.4d and StringTie TPMs were used from the output [52,53].

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 3.6.1 [18]. 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 3.6.1. In each tissue, a threshold of a >4 variance of variance stabilizing transformed 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 normalized expression scores

Normalized expression scores allow one to summarize the expression of a gene cluster in a sample by the average normalized score of the genes in that cluster. These normalized 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. The equation is as follows, where x is the VST normalized expression of gene j in sample i, $t(i)$ is the tissue type for sample i, and J is the number of genes in a given cluster:

$$S_i = \frac{1}{J} \sum_{j=1}^{J} \frac{x_{ij} - \bar{x}_j}{\underset{m:t(m)=t(i)}{Median} \left| x_{mj} - \bar{x}_j \right|}$$

A “pancreas” contamination score was defined as the average normalized score using the genes *PRSS1, CELA3A, PNLIP,* and *CLPS*, across all tissues. An “esophagus” normalized score was defined as the average normalized score using genes *KRT4* and *KRT13* in a given tissue. All normalized scores were calculated independently for each tissue, to account for tissue-specific between-sample variation in the expression of these genes.

Linear mixed model analysis for pancreas gene contamination

Linear models were used to find associations between nucleic acid isolation or sequencing on the same day as a potential contaminating tissue (source of gene expression). These analyses were completed in R version 3.6.1 using the lmer function in package lme4 (v1.1-21). In the linear mixed models we used all available tissues with >40 samples not sequenced on the same day as a pancreas sample (15 tissues, N=6,258). For the model, tissue was used as a covariate along with “pancreas sequencing day” and “pancreas nucleic acid isolation day” were coded binarily as True/False based on whether a given non-pancreas sample was sequenced or

underwent RNA isolation on the same day as a pancreas sample. Our model identified an association between the pancreas contamination score with pancreas sequencing day (T/F; 1/0), pancreas nucleic acid isolation day (T/F; 1/0), and tissue as covariates, with the subject IDs included as a random effect. The equation for our model is below where $t(i)$ gives the tissue type of sample i, $s(i)$ is defined as the sequencing day of the sample, and $d(i)$ is the isolation date of sample i. Lastly T is the number of tissue types, $u_{n(i)}$ is the random effect of subject ID for sample i, and ε_i is the error term for said sample:

$$Y_i = \beta_0 + \beta_1 I\left(s(i) \in \left\{ s\left(m: t(m) = t_{pancreas}\right)\right\}\right)$$
$$+ \beta_2 I\left(d(i) \in \left\{ d\left(m: t(m) = t_{pancreas}\right)\right\}\right)$$
$$+ \sum_{k=3}^{T+1} \beta_k I(t(i) = k - 1) + u_{n(i)} + \varepsilon_i$$

Linear mixed model analysis for esophagus gene contamination

We repeated the above analysis with the esophagus mucosa, coding shared sequencing day and shared nucleic acid isolation day similarly. In the linear mixed model we used all tissues with a sample size >40 samples not sequenced on the same day as the esophagus mucosa (8 tissues, N=3,917). The model used esophagus nucleic acid isolation day (T/F; 1/0), esophagus sequencing day (T/F; 1/0), and tissue as covariates along with subject IDs as a random effect on the esophagus contamination score.

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 [49]. The output SAM files were turned into BAM files and indexed using samtools version 1.9 [50,51]. Preliminary analysis and development of figures were generated using the Integrative Genome Viewer version 2.4.13 [54,55]). 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 [22]. 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[23]. Normalized scores for each gene are based on TMM-normalization, inverse-normal transformation, and scaling/centering at zero. A linear regression was performed, with either raw inverse-normalized expression values or expression normalized scores (after correcting for PEER factors) as the outcomes and same-day pancreas (/esophagus) sequencing status as the predictor. Beta estimates represent how many standard deviations greater the mean

expression of a gene is when samples are sequenced on the same day as a contaminating tissue, even when accounting for variance explained by 60 PEER factors.

Cross-referencing eQTLs with contamination findings

We obtained and tallied eQTL reports for the 18 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 HPA[56].

Acquiring Human Protein Atlas, RNA Atlas RNA-Seq and METSIM Data

Using the R package recount version 1.8.2, we downloaded HPA RNA-Seq data, accession ERP003613 [24]. The RNA Atlas was acquired by downloading their raw RNA-seq counts from Gene Expression Omnibus (GSE120795)[25]. The HPA RNA-Seq was performed across 27 tissues and the RNA Atlas was across 20 tissues. We filtered samples down to the shared tissues of liver, lung, pancreas, prostate, skin, and stomach as well as the potential sources of *PRL* contamination of pituitary, placenta and uterus. Only sun exposed skin was used for GTEx analysis. TPM RNA-seq data from the METSIM study was downloaded through the GEO Omnibus (GSE135134)[27].

Acquiring and Normalizing scRNA-Seq Data

We identified three publications with human scRNA-Seq data sets that all contained endothelial and mesenchymal cells[29-31]. All three sets had median read counts/cell >100,000. The processed read counts from the three studies were obtained from the Gene Expression Omnibus (GEO) (GSE84133; GSE103322; GSE72056). For each study we used the supplied cell type information to identify endothelial and mesenchymal-type cells. Mesenchymal cells were labeled as “activated stellate” (GSE84133), “fibroblast” (GSE103322), and “cancer-associated fibroblast” (GSE72056). The GSE103322 and GSE72056 read data were generated using Smart-Seq2 on a NextSeq 500 instrument[57]. Raw FASTQ data sets were normalized using RSEM to calculate $E_{i,j} = log_2(TPM_{i,j}/10+1)$, where $TPM_{i,j}$ refers to transcript-per-million for gene i in sample j [58]. The $TPM_{i,j}$ was then divided by 10, to be on the order of 100,000 transcripts to account for single cell read depth and the effect of 0s in downstream analysis[29]. GSE84133 was generated using the inDrop method and sequenced on a HiSeq 2500 instrument[59]. Human islets were isolated using a modified Ricordi method which includes collagenase, mechanical agitation, pumping and centrifugation[60] . To make single cells, the islets were then centrifuged twice at 250 rpm and treated with TrypLE Express before mechanical dispersion with a P100 pipette, followed by centrifugation at 500 rpm. The two cancer sample libraries (GSE103322, GSE72056) were generated using Smart-Seq2 and sequenced on a NextSeq 500 instrument[57]. Read counts were scaled using TPM/100 to equalize expression level counts to the other two studies. The calculation of the percent contamination of endothelial cells by *INS* was the ratio of average normalized *INS* read counts per endothelial cell divided by the normalized average read count in beta cells and excluding the 9 cells with >1000 TPM/100.

We obtained the Tabula Muris read data from https://figshare.com/articles/Robject_files_for_tissues_processed_by_Seurat/5821263 (visited on 8/8/19)[33]. We used the R package Seurat (v3.1.1) and selected endothelial cells from nine organs/tissues (brain, fat, heart, kidney, limb muscle, liver, lung, pancreas, trachea) based on the annotation of the seurat object in the metadata slot cell_ontology_class labeled 'endothelial cell'[61]. tSNE plots were generated based on the top 6 or top 10 PCs. The average number of PCs used by Tabula Muris per tissue/organ to create tSNE maps was 12 and the median was 10. Images of protein expression of contaminating pancreas genes were obtained from the HPA[56].

Data availability

No new sequencing data was created for this study. All data used in this study is available through dbGap (https://www.ncbi.nlm.nih.gov/gap/), GEO (https://www.ncbi.nlm.nih.gov/gds), recount2 (https://lcolladotor.github.io/project/recount2/) or Tabula Muris (https://figshare.com/articles/Robject_files_for_tissues_processed_by_Seurat/5821263).

Code availability

Code for all analyses is deposited at GitHub (https://github.com/mhalushka/gtex_contamination_code). Computational analyses were done using public R packages except when specifically noted otherwise.

Acknowledgements

The authors thank the GTEx Help Desk, Kristen Ardlie, Sheila Dodge, and NDRI for valuable assistance in understanding the technical data sets and islet isolation methodology. 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.

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., R.X.V., V.P. and D.E.A. performed experiments and assisted on the manuscript.

Conflicts of interest

The authors declare no conflicts of interest.

References

605 1 Consortium, G. T. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580-585,
606 doi:10.1038/ng.2653 (2013).
607 2 Tomczak, K., Czerwinska, P. & Wiznerowicz, M. The Cancer Genome Atlas (TCGA): an
608 immeasurable source of knowledge. *Contemporary oncology* **19**, A68-77,
609 doi:10.5114/wo.2014.47136 (2015).
610 3 Zeisel, A. *et al.* Brain structure. Cell types in the mouse cortex and hippocampus revealed by
611 single-cell RNA-seq. *Science* **347**, 1138-1142, doi:10.1126/science.aaa1934 (2015).
612 4 Kumasaka, N., Knights, A. J. & Gaffney, D. J. Fine-mapping cellular QTLs with RASQUAL and
613 ATAC-seq. *Nat Genet* **48**, 206-213, doi:10.1038/ng.3467 (2016).
614 5 Gutman, D. A. *et al.* MR imaging predictors of molecular profile and survival: multi-institutional
615 study of the TCGA glioblastoma data set. *Radiology* **267**, 560-569, doi:10.1148/radiol.13120118
616 (2013).
617 6 Ewing, A. D. *et al.* Combining tumor genome simulation with crowdsourcing to benchmark
618 somatic single-nucleotide-variant detection. *Nature methods* **12**, 623-630,
619 doi:10.1038/nmeth.3407 (2015).
620 7 Okoniewski, M. J. & Miller, C. J. Hybridization interactions between probesets in short oligo
621 microarrays lead to spurious correlations. *BMC Bioinformatics* **7**, 276, doi:10.1186/1471-2105-7-
622 276 (2006).
623 8 van Dijk, E. L., Jaszczyszyn, Y. & Thermes, C. Library preparation methods for next-generation
624 sequencing: tone down the bias. *Experimental cell research* **322**, 12-20,
625 doi:10.1016/j.yexcr.2014.01.008 (2014).
626 9 Tuerk, A., Wiktorin, G. & Guler, S. Mixture models reveal multiple positional bias types in RNA-
627 Seq data and lead to accurate transcript concentration estimates. *PLoS computational biology*
628 **13**, e1005515, doi:10.1371/journal.pcbi.1005515 (2017).
629 10 Lusk, R. W. Diverse and widespread contamination evident in the unmapped depths of high
630 throughput sequencing data. *PLoS One* **9**, e110808, doi:10.1371/journal.pone.0110808 (2014).
631 11 Illumina. *Effects of Index Misassignment on Multiplexing and Downstream Analysis*,
632 <https://www.illumina.com/content/dam/illumina-
633 marketing/documents/products/whitepapers/index-hopping-white-paper-770-2017-004.pdf> (
634 12 Rosenberg, A. Z. *et al.* xMD-miRNA-seq to generate near in vivo miRNA expression estimates in
635 colon epithelial cells. *Scientific reports* **8**, 9783, doi:10.1038/s41598-018-28198-z (2018).
636 13 McGinnis, C. S., Murrow, L. M. & Gartner, Z. J. DoubletFinder: Doublet Detection in Single-Cell
637 RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Syst* **8**, 329-337 e324,
638 doi:10.1016/j.cels.2019.03.003 (2019).
639 14 Merchant, S., Wood, D. E. & Salzberg, S. L. Unexpected cross-species contamination in genome
640 sequencing projects. *PeerJ* **2**, e675, doi:10.7717/peerj.675 (2014).
641 15 Cibulskis, K. *et al.* ContEst: estimating cross-contamination of human samples in next-generation
642 sequencing data. *Bioinformatics* **27**, 2601-2602, doi:10.1093/bioinformatics/btr446 (2011).
643 16 Ma, X. *et al.* Analysis of error profiles in deep next-generation sequencing data. *Genome Biol* **20**,
644 50, doi:10.1186/s13059-019-1659-6 (2019).
645 17 McCall, M. N., Illei, P. B. & Halushka, M. K. Complex Sources of Variation in Tissue Expression
646 Data: Analysis of the GTEx Lung Transcriptome. *American journal of human genetics* **99**, 624-
647 635, doi:10.1016/j.ajhg.2016.07.007 (2016).
648 18 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for
649 RNA-seq data with DESeq2. *Genome Biol* **15**, 550, doi:10.1186/s13059-014-0550-8 (2014).
650 19 Muraro, M. J. *et al.* A Single-Cell Transcriptome Atlas of the Human Pancreas. *Cell Syst* **3**, 385-
651 394 e383, doi:10.1016/j.cels.2016.09.002 (2016).

652 20 Uhlen, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419,
653 doi:10.1126/science.1260419 (2015).
654 21 Atlas, T. H. P. *The human tissue specific proteome*,
655 <https://www.proteinatlas.org/humanproteome/tissue/tissue+specific> (
656 22 Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression
657 residuals (PEER) to obtain increased power and interpretability of gene expression analyses.
658 *Nature protocols* **7**, 500-507, doi:10.1038/nprot.2011.457 (2012).
659 23 Consortium, G. T. *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**,
660 204-213, doi:10.1038/nature24277 (2017).
661 24 Fagerberg, L. *et al.* Analysis of the human tissue-specific expression by genome-wide integration
662 of transcriptomics and antibody-based proteomics. *Molecular & cellular proteomics : MCP* **13**,
663 397-406, doi:10.1074/mcp.M113.035600 (2014).
664 25 Suntsova, M. *et al.* Atlas of RNA sequencing profiles for normal human tissues. *Sci Data* **6**, 36,
665 doi:10.1038/s41597-019-0043-4 (2019).
666 26 Chhibber, A. *et al.* Transcriptomic variation of pharmacogenes in multiple human tissues and
667 lymphoblastoid cell lines. *The pharmacogenomics journal* **17**, 137-145, doi:10.1038/tpj.2015.93
668 (2017).
669 27 Raulerson, C. K. *et al.* Adipose Tissue Gene Expression Associations Reveal Hundreds of
670 Candidate Genes for Cardiometabolic Traits. *American journal of human genetics* **105**, 773-787,
671 doi:10.1016/j.ajhg.2019.09.001 (2019).
672 28 Griffiths, J. A., Richard, A. C., Bach, K., Lun, A. T. L. & Marioni, J. C. Detection and removal of
673 barcode swapping in single-cell RNA-seq data. *Nature communications* **9**, 2667,
674 doi:10.1038/s41467-018-05083-x (2018).
675 29 Puram, S. V. *et al.* Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor
676 Ecosystems in Head and Neck Cancer. *Cell* **171**, 1611-1624 e1624, doi:10.1016/j.cell.2017.10.044
677 (2017).
678 30 Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell
679 RNA-seq. *Science* **352**, 189-196, doi:10.1126/science.aad0501 (2016).
680 31 Baron, M. *et al.* A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals
681 Inter- and Intra-cell Population Structure. *Cell Syst* **3**, 346-360 e344,
682 doi:10.1016/j.cels.2016.08.011 (2016).
683 32 Chiou, J. *et al.* Single cell chromatin accessibility reveals pancreatic islet cell type- and state-
684 specific regulatory programs of diabetes risk. *BioRxiv*, doi:https://doi.org/10.1101/693671
685 (2019).
686 33 Tabula Muris, C. *et al.* Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris.
687 *Nature* **562**, 367-372, doi:10.1038/s41586-018-0590-4 (2018).
688 34 Consortium, G. T. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis:
689 multitissue gene regulation in humans. *Science* **348**, 648-660, doi:10.1126/science.1262110
690 (2015).
691 35 Kircher, M., Sawyer, S. & Meyer, M. Double indexing overcomes inaccuracies in multiplex
692 sequencing on the Illumina platform. *Nucleic Acids Res* **40**, e3, doi:10.1093/nar/gkr771 (2012).
693 36 Young, M. D. & Behjati, S. SoupX removes ambient RNA contamination from droplet based
694 single cell RNA sequencing data. *BioRxiv*, doi:https://doi.org/10.1101/303727 (2018).
695 37 DePasquale, E. A. K. *et al. DoubletDecon: Cell-State Aware Removal of Single-Cell RNA-Seq*
696 *Doublets* (https://www.biorxiv.org/content/10.1101/364810v2, 2019).
697 38 Regev, A. *et al.* The Human Cell Atlas. *eLife* **6**, doi:10.7554/eLife.27041 (2017).

39 Witwer, K. W. & Halushka, M. K. Toward the promise of microRNAs - Enhancing reproducibility and rigor in microRNA research. *RNA biology* **13**, 1103-1116, doi:10.1080/15476286.2016.1236172 (2016).
40 Kryukov, K. & Imanishi, T. Human Contamination in Public Genome Assemblies. *PLoS One* **11**, e0162424, doi:10.1371/journal.pone.0162424 (2016).
41 Longo, M. S., O'Neill, M. J. & O'Neill, R. J. Abundant human DNA contamination identified in non-primate genome databases. *PLoS One* **6**, e16410, doi:10.1371/journal.pone.0016410 (2011).
42 Zhang, L. *et al.* Exogenous plant MIR168a specifically targets mammalian LDLRAP1: evidence of cross-kingdom regulation by microRNA. *Cell Res* **22**, 107-126, doi:10.1038/cr.2011.158 (2012).
43 Tosar, J. P., 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* **20**, 754-757, doi:10.1261/rna.044263.114 (2014).
44 Zhang, Y. *et al.* Analysis of plant-derived miRNAs in animal small RNA datasets. *BMC Genomics* **13**, 381, doi:10.1186/1471-2164-13-381 (2012).
45 Li, H. *et al.* Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nat Genet* **49**, 708-718, doi:10.1038/ng.3818 (2017).
46 Tian, S. K. *et al.* 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*, doi:10.5858/arpa.2016-0108-RA (2016).
47 Van Allen, E. M. *et al.* Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. *Nature medicine* **20**, 682-688, doi:10.1038/nm.3559 (2014).
48 Collado-Torres, L., Nellore, A. & Jaffe, A. E. recount workflow: Accessing over 70,000 human RNA-seq samples with Bioconductor. *F1000Research* **6**, 1558, doi:10.12688/f1000research.12223.1 (2017).
49 Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nature methods* **12**, 357-360, doi:10.1038/nmeth.3317 (2015).
50 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).
51 Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987-2993, doi:10.1093/bioinformatics/btr509 (2011).
52 Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature biotechnology* **33**, 290-295, doi:10.1038/nbt.3122 (2015).
53 Pertea, M., Kim, D., Pertea, G. M., Leek, J. T. & Salzberg, S. L. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nature protocols* **11**, 1650-1667, doi:10.1038/nprot.2016.095 (2016).
54 Thorvaldsdottir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics* **14**, 178-192, doi:10.1093/bib/bbs017 (2013).
55 Robinson, J. T. *et al.* Integrative genomics viewer. *Nature biotechnology* **29**, 24-26, doi:10.1038/nbt.1754 (2011).
56 Uhlen, M. *et al.* Towards a knowledge-based Human Protein Atlas. *Nature biotechnology* **28**, 1248-1250, doi:10.1038/nbt1210-1248 (2010).
57 Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature methods* **10**, 1096-1098, doi:10.1038/nmeth.2639 (2013).

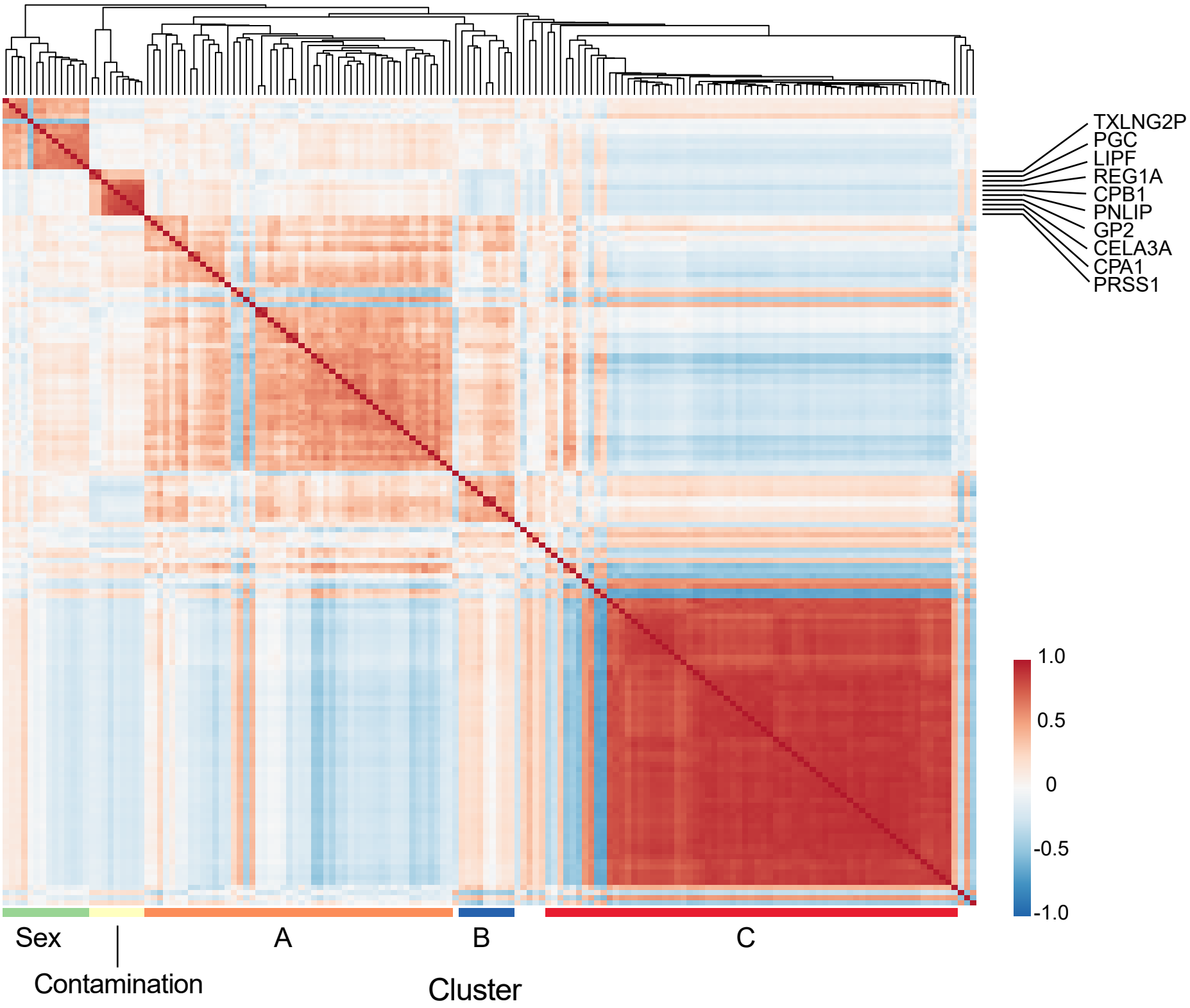
744 58 Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or
745 without a reference genome. *BMC Bioinformatics* **12**, 323, doi:10.1186/1471-2105-12-323
746 (2011).
747 59 Klein, A. M. *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic stem
748 cells. *Cell* **161**, 1187-1201, doi:10.1016/j.cell.2015.04.044 (2015).
749 60 Ricordi, C., Lacy, P. E., Finke, E. H., Olack, B. J. & Scharp, D. W. Automated method for isolation
750 of human pancreatic islets. *Diabetes* **37**, 413-420 (1988).
751 61 Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell
752 gene expression data. *Nature biotechnology* **33**, 495-502, doi:10.1038/nbt.3192 (2015).

753

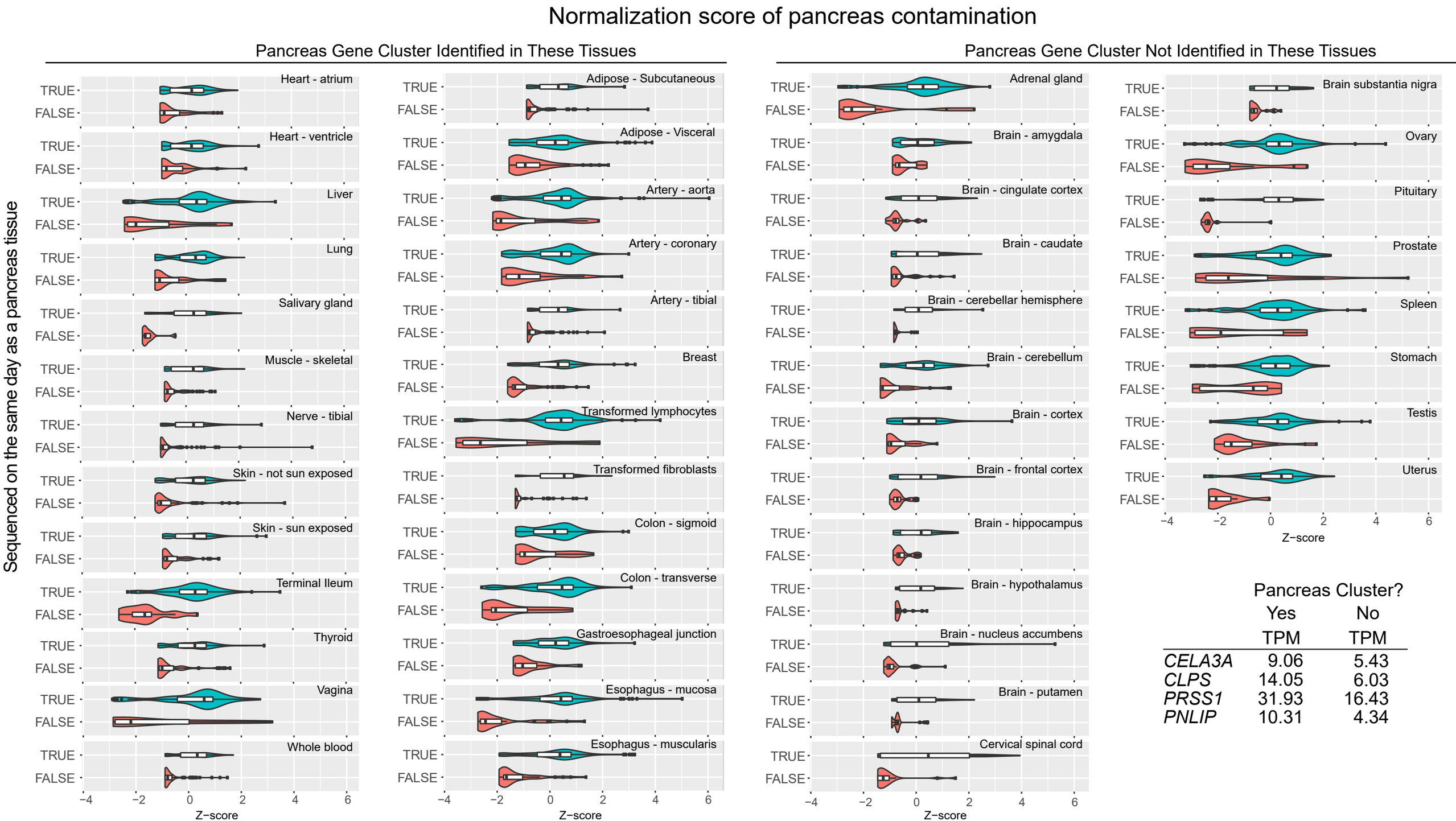
Basal Contamination of Sequencing: Lessons from the GTEx dataset

Tim O. Nieuwenhuis et al.

Supplementary Information

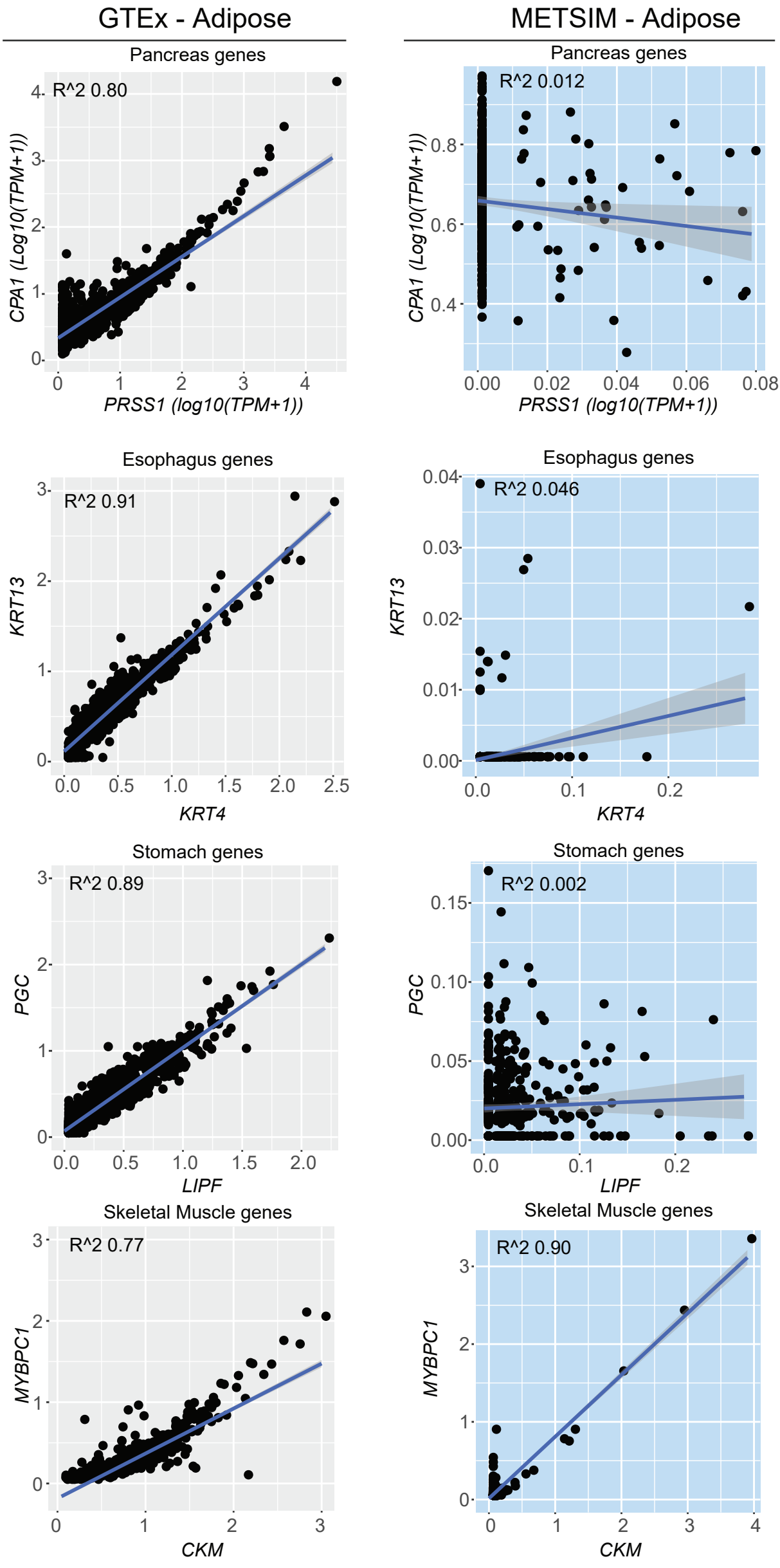

TXLNG2P
PGC
LIPF
REG1A
CPB1
PNLIP
GP2
CELA3A
CPA1
PRSS1
1.0
0.5
0
-0.5
-1.0
Sex
Contamination
A
B
C
Cluster

Supplementary Figure 1. 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. The etiology of clusters A and B are unknown. Cluster C represents co-variable genes related to mitosis.

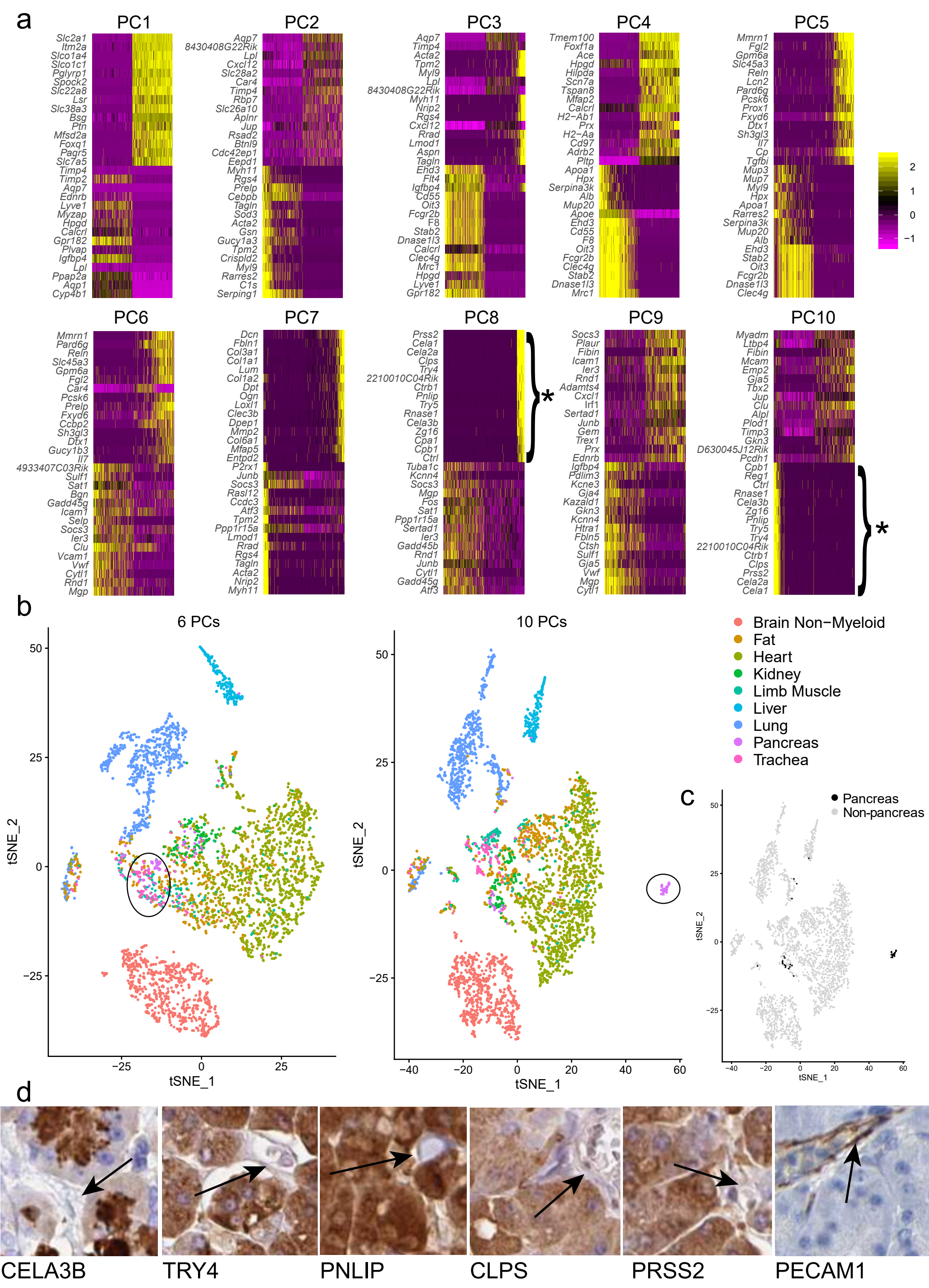


	Pancreas Cluster?	
	Yes	No
	TPM	TPM
CELA3A	9.06	5.43
CLPS	14.05	6.03
PRSS1	31.93	16.43
PNLIP	10.31	4.34

Supplementary Figure 2. Violin plots representing the pancreatic gene (*CELA3A*, *CLPS*, *PRSS1*, *PNLIP*) normalization scores of samples sequenced the same day as a pancreas sample and those that were not. The bottom right corner shows the median TPM of pancreas genes in non-pancreas tissues. These were divided base on the tissue having all four genes in their contamination cluster (N=26) or not (N=21).


GTEx - Adipose
METSIM - Adipose
Pancreas genes
R^2 0.80
CPA1 (Log10(TPM+1))
PRSS1 (log10(TPM+1))
Pancreas genes
R^2 0.012
CPA1 (Log10(TPM+1))
PRSS1 (log10(TPM+1))
Esophagus genes
R^2 0.91
KRT13
KRT4
Esophagus genes
R^2 0.046
KRT13
KRT4
Stomach genes
R^2 0.89
PGC
LIPF
Stomach genes
R^2 0.002
PGC
LIPF
Skeletal Muscle genes
R^2 0.77
MYBPC1
CKM
Skeletal Muscle genes
R^2 0.90
MYBPC1
CKM

Supplementary Figure 3. A scatterplot comparison of different contaminating genes between the GTEx adipose tissues (n = 797) and the adipose tissues harvested in the METSIM study (n = 434). We compared *CPA1* vs *PRSS1*, *KRT13* vs *KRT4*, *PGC* vs *LIPF*, and *MYBPC1* vs *CKM* as each of those gene pairs are highly expressed together in a contaminating tissue type. In GTEx the expression is consistently higher and has a stronger positive correlation compared to the METSIM study for three of those pairs Because skeletal muscle can be captured in adipose harvesting, *MYBPC1* and *CKM* also showed strong correlation in METSIM as a result of a direct contamination.


a
PC1
Slc2a1
Itm2a
Slco1a4
Slco1c1
Pglyrp1
Spock2
Slc22a8
Lsr
Slc38a3
Bsg
Ptn
Mfsd2a
Foxq1
Paqr5
Slc7a5
Timp4
Timp2
Aqp7
Ednrb
Lyve1
Myzap
Hpgd
Calcrl
Gpr182
Plvap
Igfbp4
Lpl
Ppap2a
Aqp1
Cyp4b1
PC2
Aqp7
8430408G22Rik
Lpl
Cxcl12
Slc28a2
Car4
Timp4
Rbp7
Slc26a10
Aplnr
Jup
Rsad2
Btnl9
Cdc42ep1
Eepd1
Myh11
Rgs4
Prelp
Cebpb
Tagln
Sod3
Acta2
Gsn
Gucy1a3
Tpm2
Crispld2
Myl9
Rarres2
C1s
Serping1
PC3
Aqp7
Timp4
Acta2
Tpm2
Myl9
Lpl
8430408G22Rik
Myh11
Nrip2
Rgs4
Cxcl12
Rrad
Lmod1
Aspn
Tagln
Ehd3
Flt4
Igfbp4
Cd55
Oit3
Fcgr2b
F8
Stab2
Dnase1l3
Calcrl
Clec4g
Mrc1
Hpgd
Lyve1
Gpr182
PC4
Tmem100
Foxf1a
Ace
Hpgd
Hilpda
Scn7a
Tspan8
Mfap2
Calcrl
H2−Ab1
Prx
H2−Aa
Cd97
Adrb2
Pltp
Apoa1
Hpx
Serpina3k
Alb
Mup20
Apoe
Ehd3
Cd55
F8
Oit3
Fcgr2b
Clec4g
Stab2
Dnase1l3
Mrc1
PC5
Mmrn1
Fgl2
Gpm6a
Slc45a3
Reln
Lcn2
Pard6g
Pcsk6
Prox1
Fxyd6
Dtx1
Sh3gl3
Il7
Cp
Tgfbi
Mup3
Mup7
Myl9
Hpx
Apoa1
Rarres2
Serpina3k
Mup20
Alb
Ehd3
Stab2
Oit3
Fcgr2b
Dnase1l3
Clec4g
2
1
0
−1
PC6
Mmrn1
Pard6g
Reln
Slc45a3
Gpm6a
Fgl2
Car4
Pcsk6
Prelp
Fxyd6
Ccbp2
Sh3gl3
Dtx1
Gucy1b3
Il7
4933407C03Rik
Sulf1
Sat1
Bgn
Gadd45g
Icam1
Selp
Socs3
Ier3
Clu
Vcam1
Vwf
Cytl1
Rnd1
Mgp
PC7
Dcn
Fbln1
Col3a1
Col1a1
Lum
Col1a2
Dpt
Ogn
Loxl1
Clec3b
Dpep1
Mmp2
Col6a1
Mfap5
Entpd2
P2rx1
Junb
Socs3
Rasl12
Ccdc3
Atf3
Tpm2
Ppp1r15a
Lmod1
Rrad
Rgs4
Tagln
Acta2
Nrip2
Myh11
PC8
Prss2
Cela1
Cela2a
Clps
Try4
2210010C04Rik
Ctrb1
Pnlip
Try5
Rnase1
Cela3b
Zg16
Cpa1
Cpb1
Ctrl
Tuba1c
Kcnn4
Socs3
Mgp
Fos
Sat1
Ppp1r15a
Sertad1
Ier3
Gadd45b
Rnd1
Junb
Cytl1
Gadd45g
Atf3
*
PC9
Socs3
Plaur
Fibin
Icam1
Ier3
Rnd1
Adamts4
Cxcl1
Irf1
Sertad1
Junb
Gem
Trex1
Prx
Ednrb
Igfbp4
Pdlim3
Kcne3
Gja4
Kazald1
Gkn3
Kcnn4
Htra1
Fbln5
Ctsh
Sulf1
Gja5
Vwf
Mgp
Cytl1
PC10
Myadm
Ltbp4
Fibin
Mcam
Emp2
Gja5
Tbx2
Jup
Clu
Alpl
Plod1
Timp3
Gkn3
D630045J12Rik
Pcdh1
Cpb1
Reg1
Ctrl
Rnase1
Cela3b
Zg16
Pnlip
Try5
Try4
2210010C04Rik
Ctrb1
Clps
Prss2
Cela2a
Cela1
*
b
6 PCs
10 PCs
tSNE_1
tSNE_2
50
25
0
−25
−40
−20
0
20
40
60
Brain Non−Myeloid
Fat
Heart
Kidney
Limb Muscle
Liver
Lung
Pancreas
Trachea
c
Pancreas
Non-pancreas
d
CELA3B
TRY4
PNLIP
CLPS
PRSS2
PECAM1

Supplementary Figure 4. Pancreatic acinar cell contamination drives endothelial cell diversity in Tabula Muris. **a** The major genes of the top 10 principal components (PCs) for nine groups of endothelial cells. * represents contaminating acinar cell expression. **b** tSNE plots based on 6 or 10 PCs demonstrate a large deviation of pancreas endothelial cells from a shared cluster to a unique cluster based on additional PCs. **c** In this black and white representation, a subset of pancreas cells are noted to move. **d** Representative images (HPA) of several highly-expressed acinar proteins that do not mark endothelial cells (black arrows) and *PECAM1*, a known endothelial cell marker.