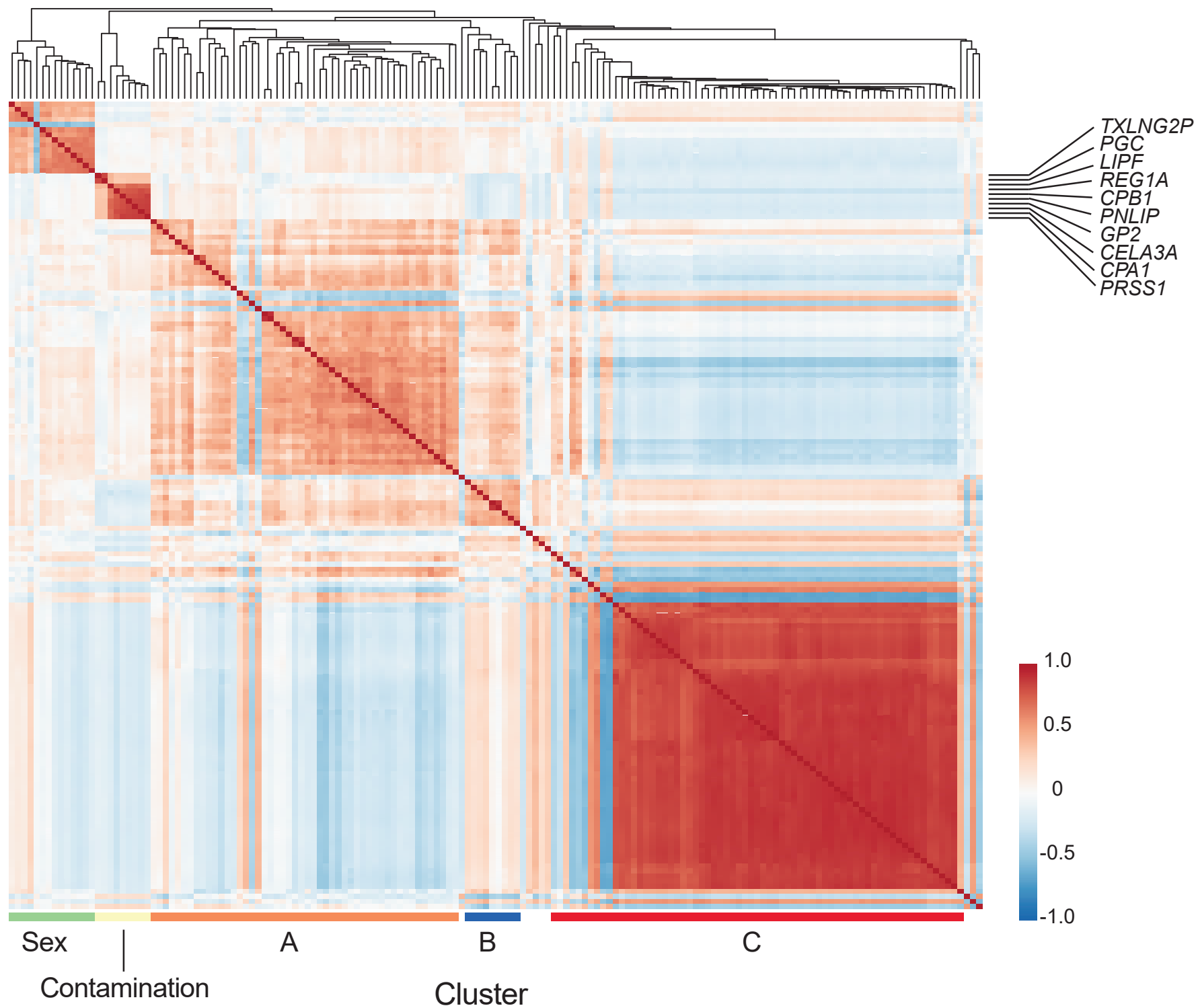


Consistent RNA Sequencing Contamination in GTEx and Other Datasets

Tim O. Nieuwenhuis et al.

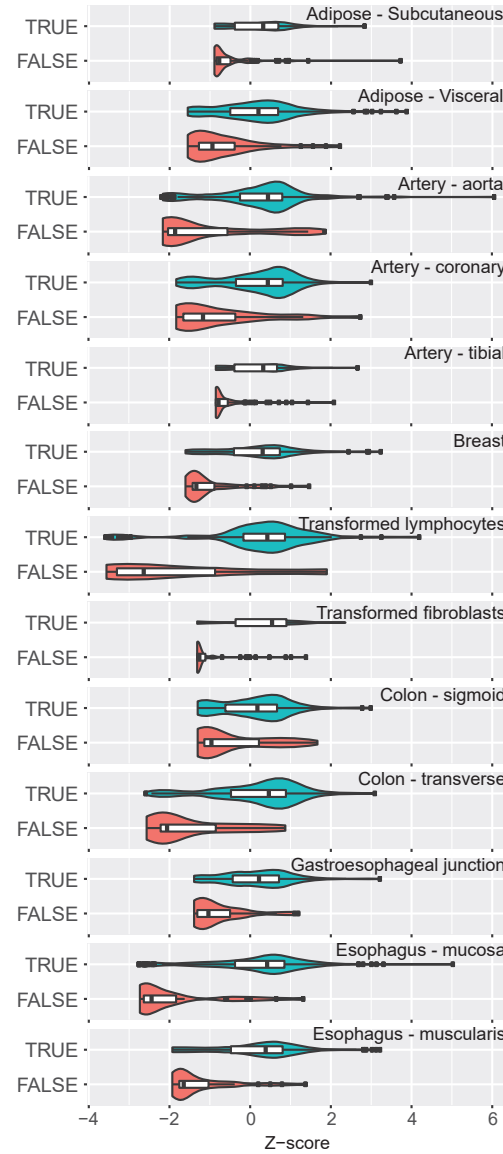
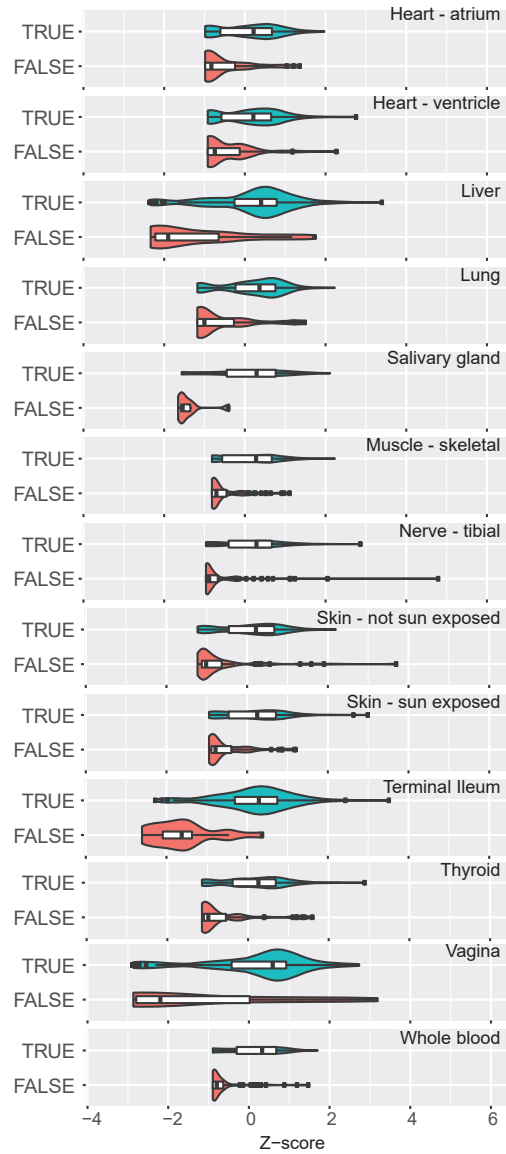
Supplementary Information



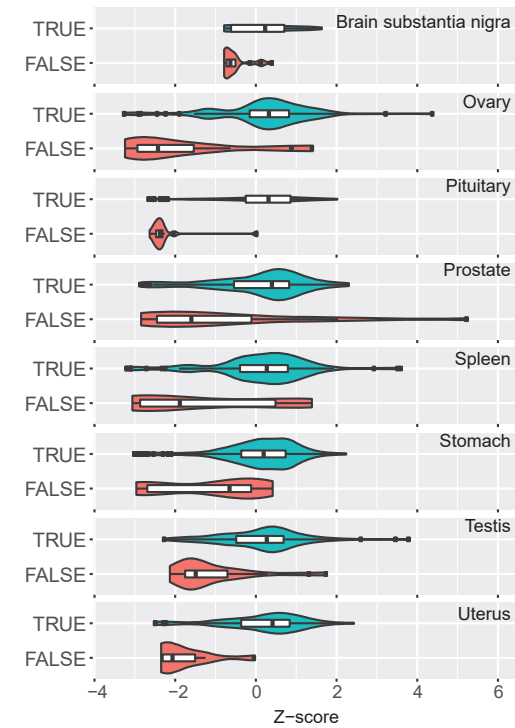
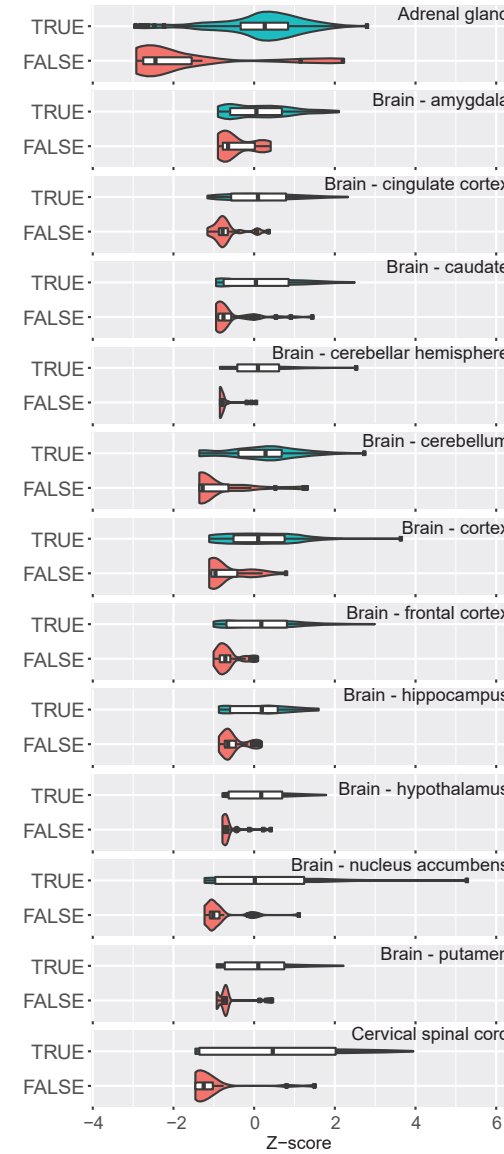
Supplementary Figure 1. A correlation heatmap of the highly variable gene clusters in 343 transformed fibroblast samples. Blue to red scale shows Kendall's tau correlation from -1 to 1. Genes within the contamination cluster are given. The etiology of clusters A and B are unknown. Cluster C represents co-variable genes related to mitosis.

Normalization score of pancreas contamination

Pancreas Gene Cluster Identified in These Tissues



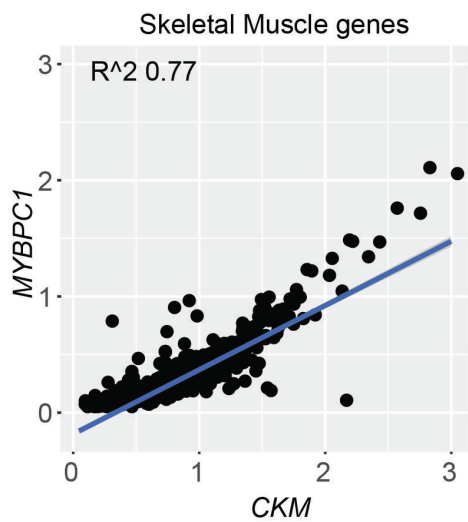
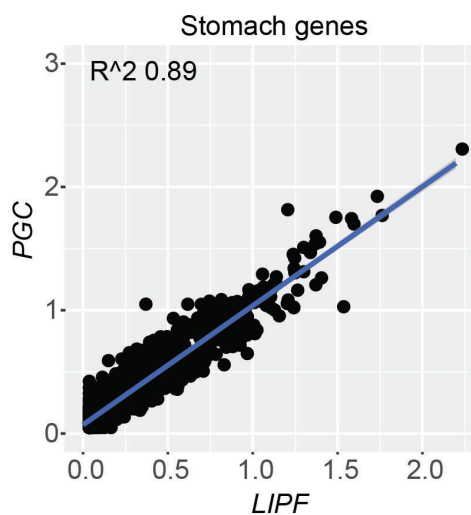
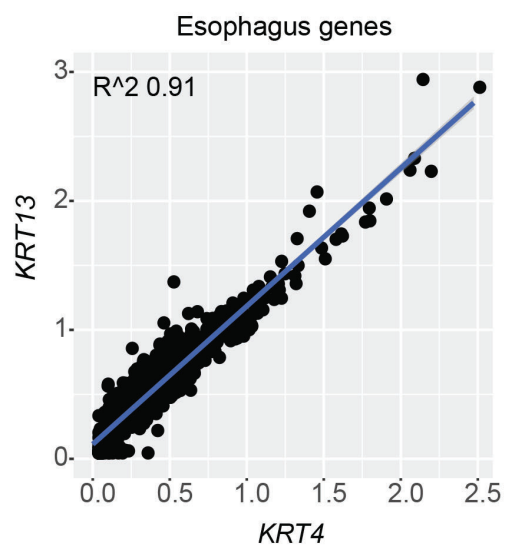
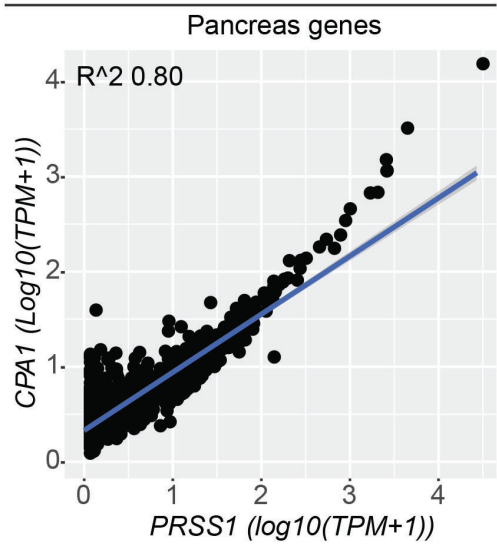
Pancreas Gene Cluster Not Identified in These Tissues



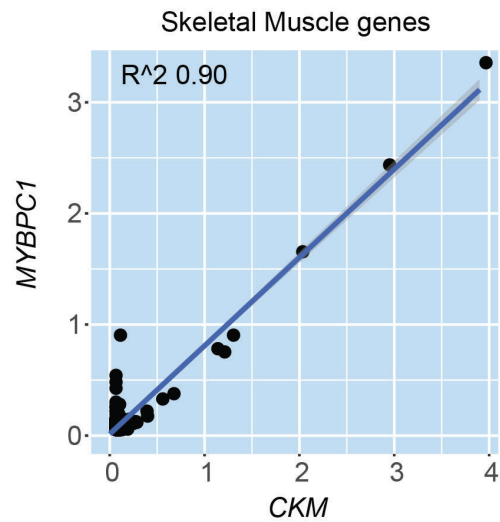
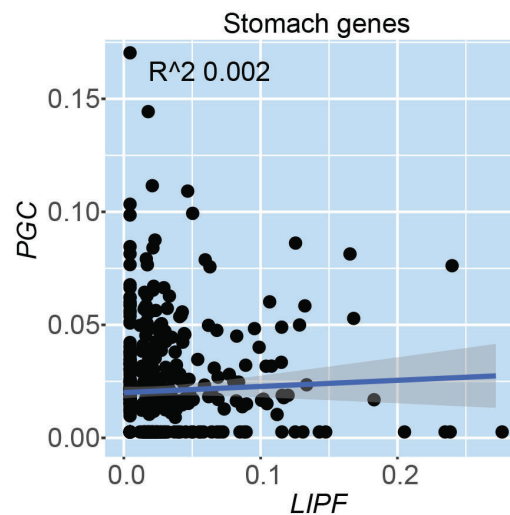
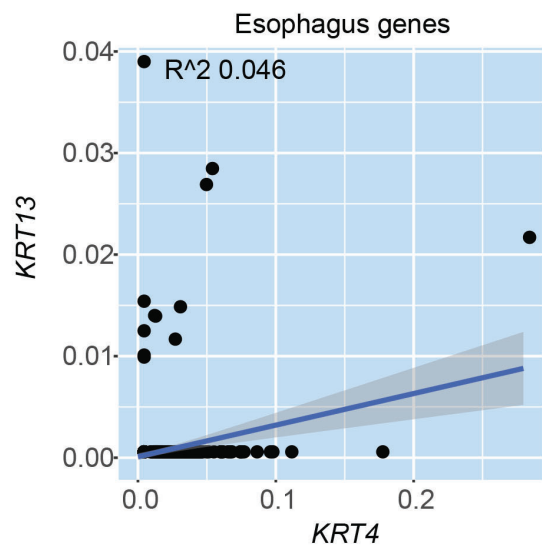
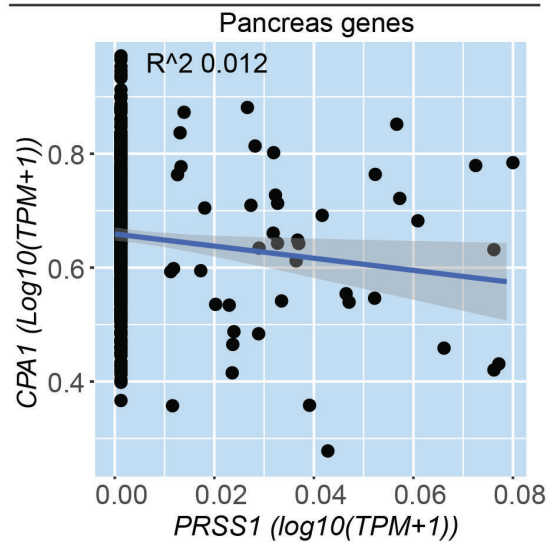
	Pancreas Cluster?	
	Yes	No
	TPM	TPM
<i>CELA3A</i>	9.06	5.43
<i>CLPS</i>	14.05	6.03
<i>PRSS1</i>	31.93	16.43
<i>PNLIP</i>	10.31	4.34

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). The solid line in all boxplots represents the median of the data, while the lower and upper hinges correspond to the 25th percentile and 75th percentile respectively. The whiskers represent the interquartile range $\times 1.5$, and any outliers beyond the whiskers are represented as dots

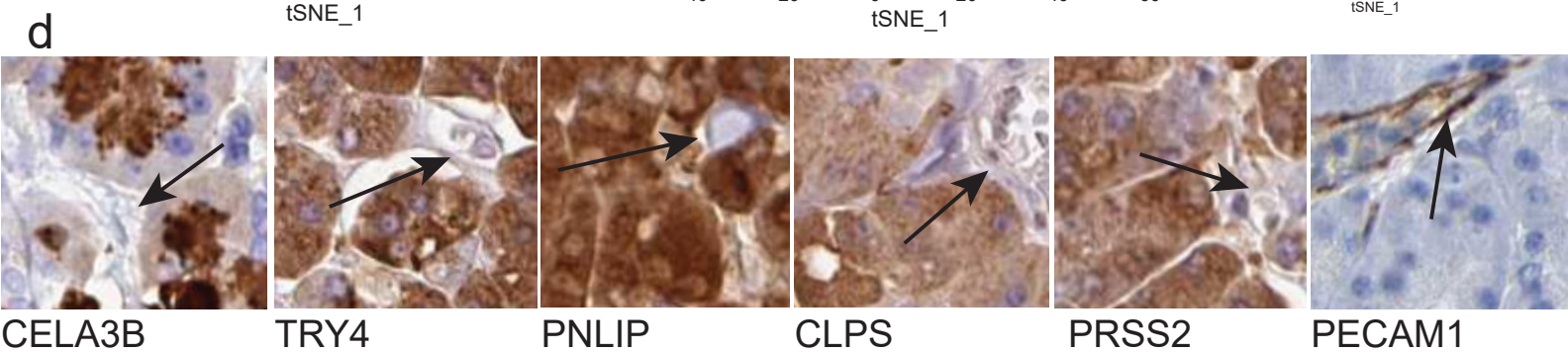
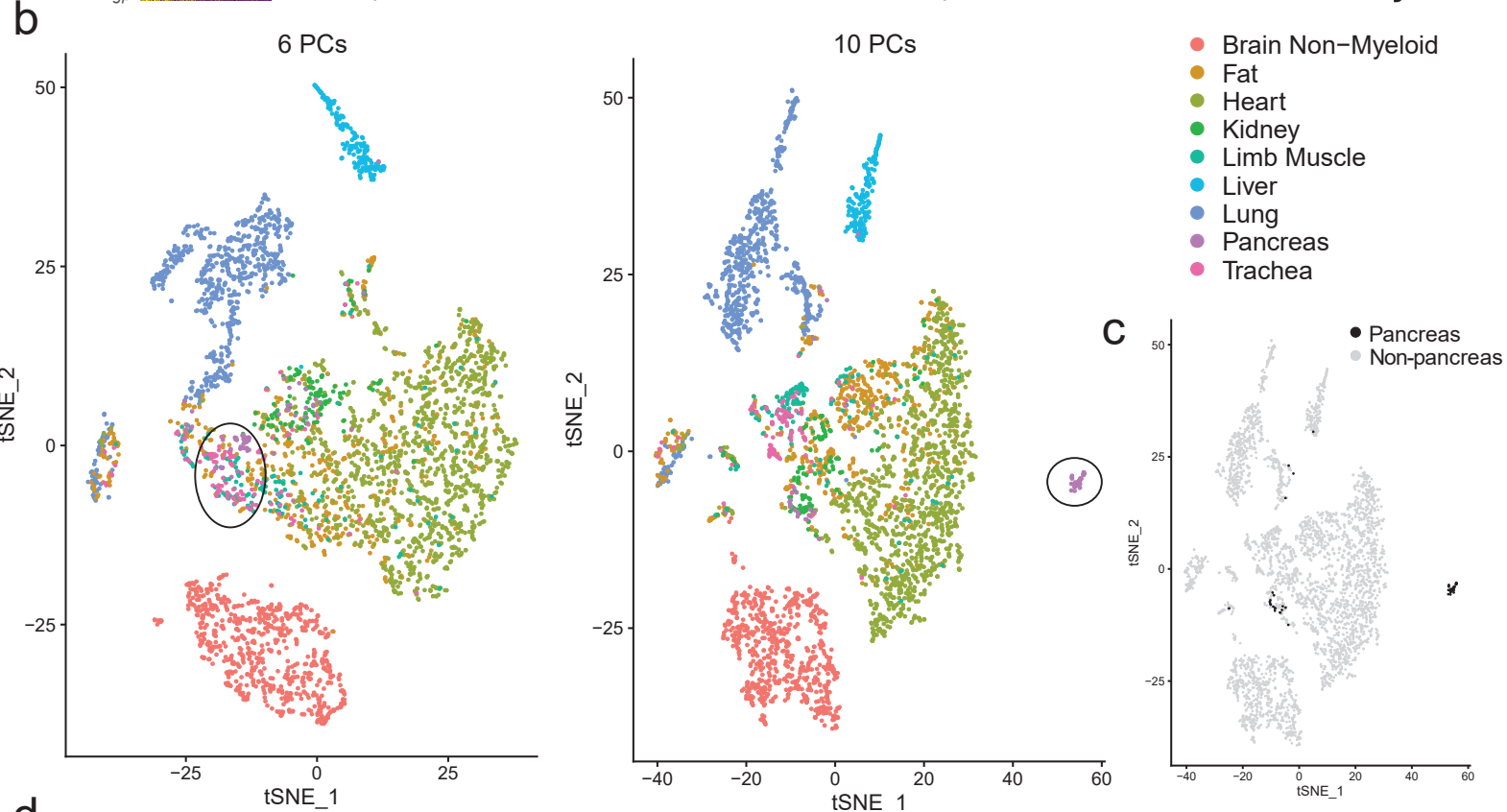
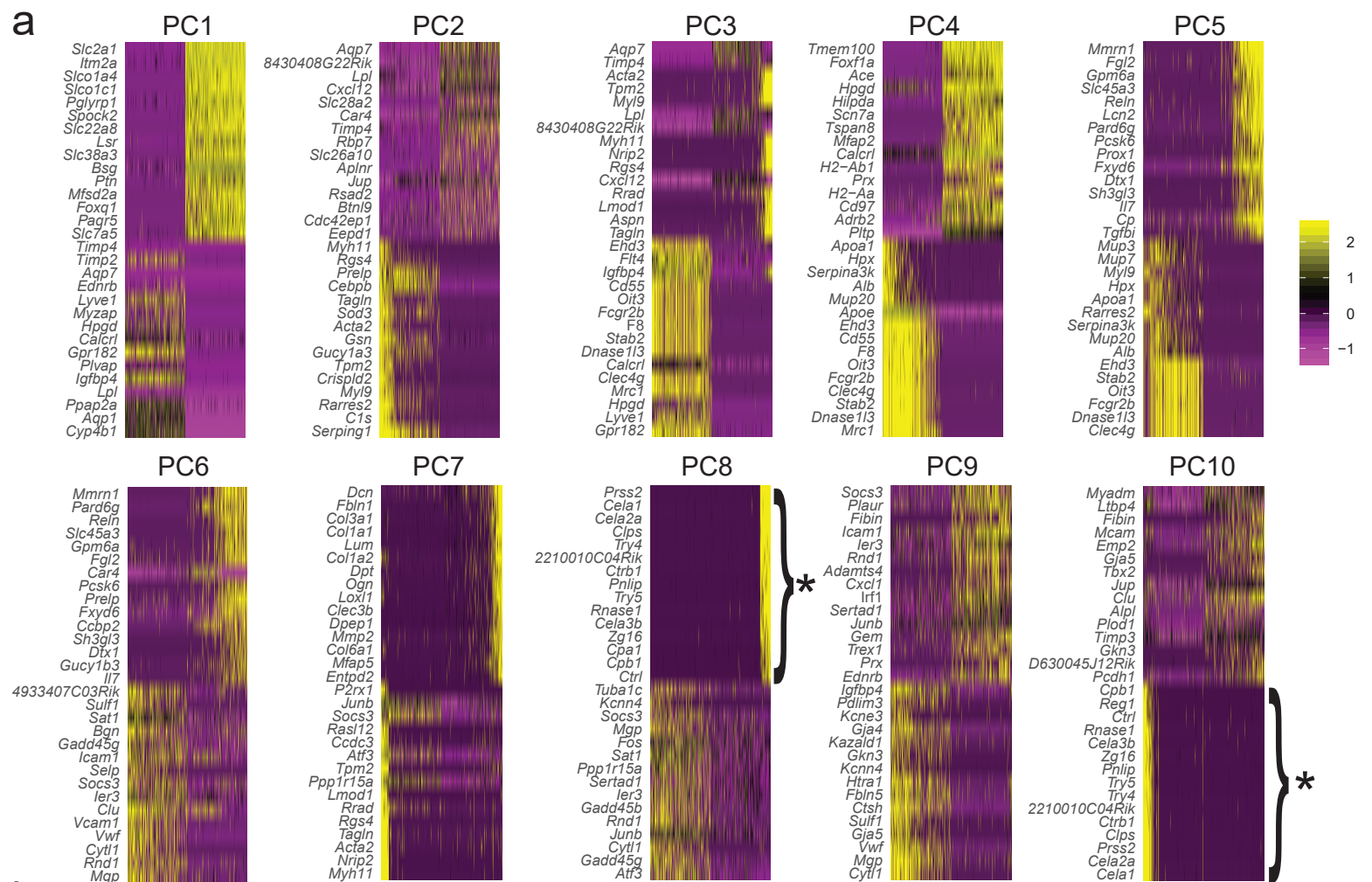
GTEx - Adipose



METSIM - Adipose



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.



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.