Cell

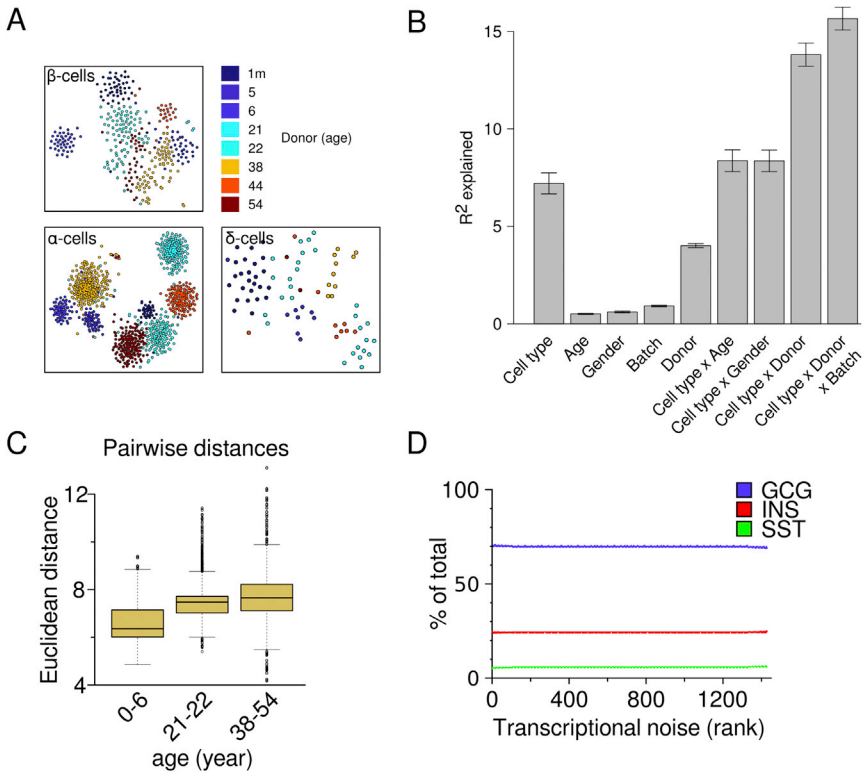

Figure S1. Single-Cell RNA-Seq of Human Pancreas, Related to Figure 1

(A) tSNE plot of cells from the major endocrine cell types. Colors are by donor (as specified by age, top right panel). Cells cluster by donor suggesting that our data could not find support for sub cell types that have a stronger cell identity than individual variation, but does not preclude the existence of more subtle sub-cell types.

(B) Relative contributions of cell type, age, gender, donor, and library preparation batch. Error bars are mean +-SEM.

(C) Boxplot of pairwise euclidean distances between 10000 random pairs of endocrine cells from each donor is plotted by age group. Whole-transcriptome cell-to-cell variability between β-cells from adult donors is higher than variability between cells from juvenile donors. Boxes indicate the middle quartiles, separated by median line. Whiskers indicate last values within 1.5 × the interquartile range for the box.

(D) Cell type composition is constant between endocrine pancreatic cells with low and high transcriptional noise. Lines are running mean (k = 200) of fractional cell type content, by rank of transcriptional noise (low to high).

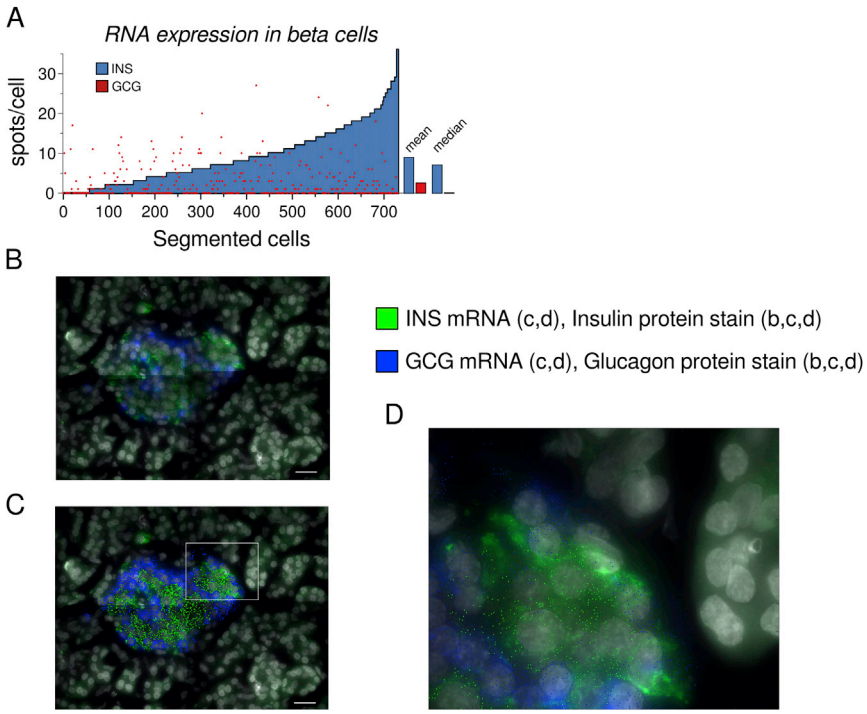

Figure S2. Quantification of Cell-Atypical Hormone Expression In Situ, Related to Figure 2

(A) Cells were ranked by number of *INS* spots per cell (blue bars), with the number of *GCG* spots in the same cell shown in red. There was no significant dependency between *INS* expression and *GCG* expression ($p = 0.859$, linear regression, $n = 730$).

(B–D) Parallel protein and RNA staining in situ. A representative image at 63x magnification of a pancreatic islet containing cells with atypical hormone expression. Scale bar is 20 μm. (B), protein stain only (green: insulin, blue: glucagon); (C), in situ RNA-staining (dots) + protein stain (green dots: *INS* gene specific, blue dots: GCG gene specific); (D) magnified version of (B).

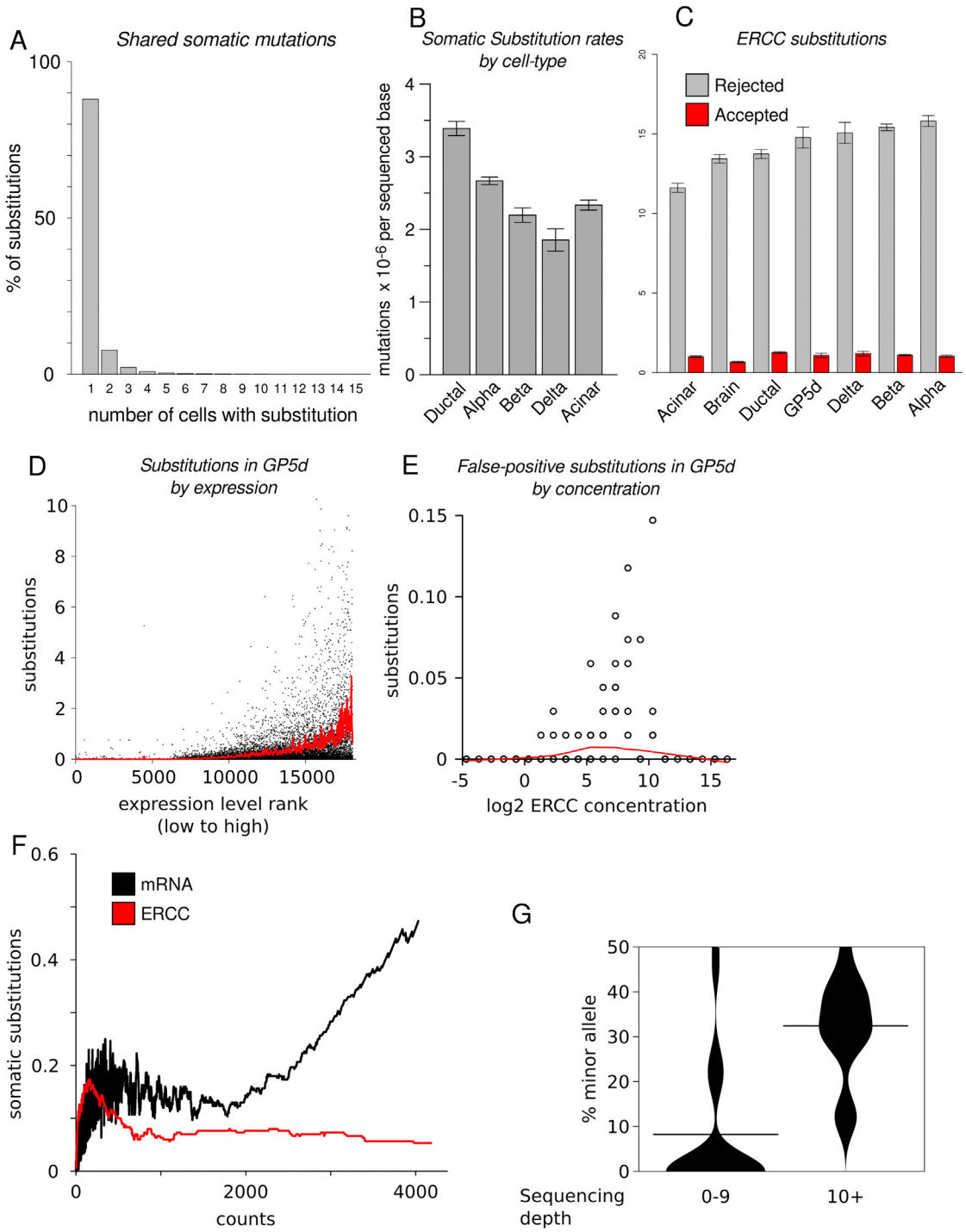

Figure S3. Characteristics of Somatic Substitutions in Single-Cell RNA-Seq Data, Related to Figure 3

(A) The distribution of the number of occurrences of distinct somatic (non-germline) substitutions. As expected, somatic mutations that are shared between more than one cell are rare.

(B) Somatic substitution rates vary between cell types in the same organ (bars are mean +- SEM).

(C) Numbers of substitution calls in ERCC control are similar between cell types. Shown are mean numbers (± SEM) of putative substitution calls in ERCC controls, that were rejected (gray bars) or accepted (red bars) by our variation calling method. Red bars constitute false-positive calls.

(D) Substitutions/cell in genes in GP5d cells, ordered by mean expression. Only genes that were expressed in at least one cell are shown. Both clonal somatic and non-clonal substitutions are counted. Red line is running mean (k = 100).

(E) Substitutions in ERCC controls by concentration of each spike-in RNA. Red line is a local regression (loess) fit.

(legend continued on next page)

(F) Somatic substitutions in individual mRNA or ERCC control transcripts in a cell as a function of the number of reads mapped to the transcript/cell. Substitutions in highly expressed genes are more likely to be detected, whereas PCR errors are less likely to pass QC thresholds. Lines are running mean (k = 300).

(G) Allelic imbalance is negatively correlated with the depth of sequencing used to call the substitution.

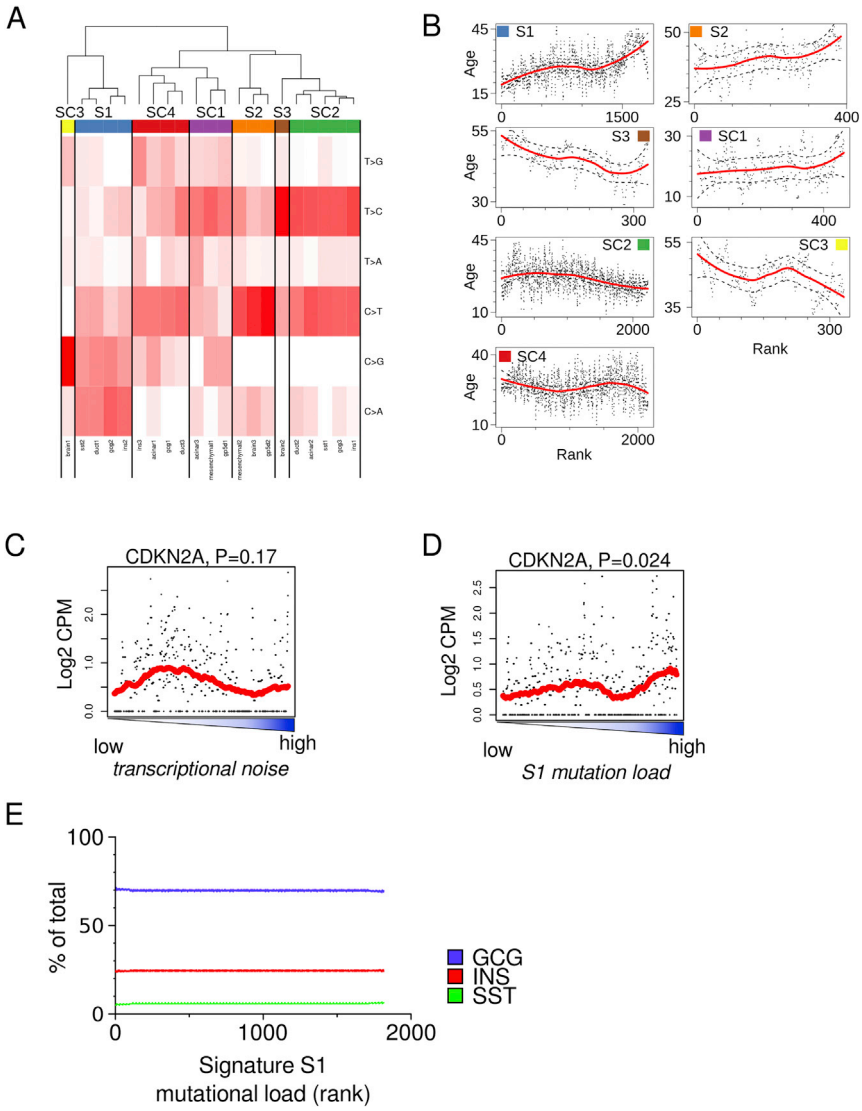

Figure S4. Mutational Signatures, Related to Figure 4

(A) Heatmap showing raw signatures from non-negative matrix factorization. Dendrogram (top) indicates hierarchical clustering, and clusters at the 6th branch point shown as colored bar between dendrogram and heatmap. The spatial median of each cluster is shown in Figure 4A.

(B) Association of signatures S1-3,SC4-7 to age. Cells were ordered according to the fraction of mutations attributed to the indicated signature. Dots are running mean of age, k = 10. Line is loess fit, dotted lines indicate +-.999 confidence interval.

(C and D) CDKN2A expression in cells ordered according to their level of transcriptional noise (C) or fraction of mutations attributed to signature S1 (D). Transcriptional noise is not associated with *CDKN2A* expression, while S1 mutational load is weakly associated to it.

(E) Cell type composition is constant between cells with low and high signature S1 mutational load. Lines are running mean (k = 200) of fractional cell type content, by rank of signature S1 specific mutational load (low to high).

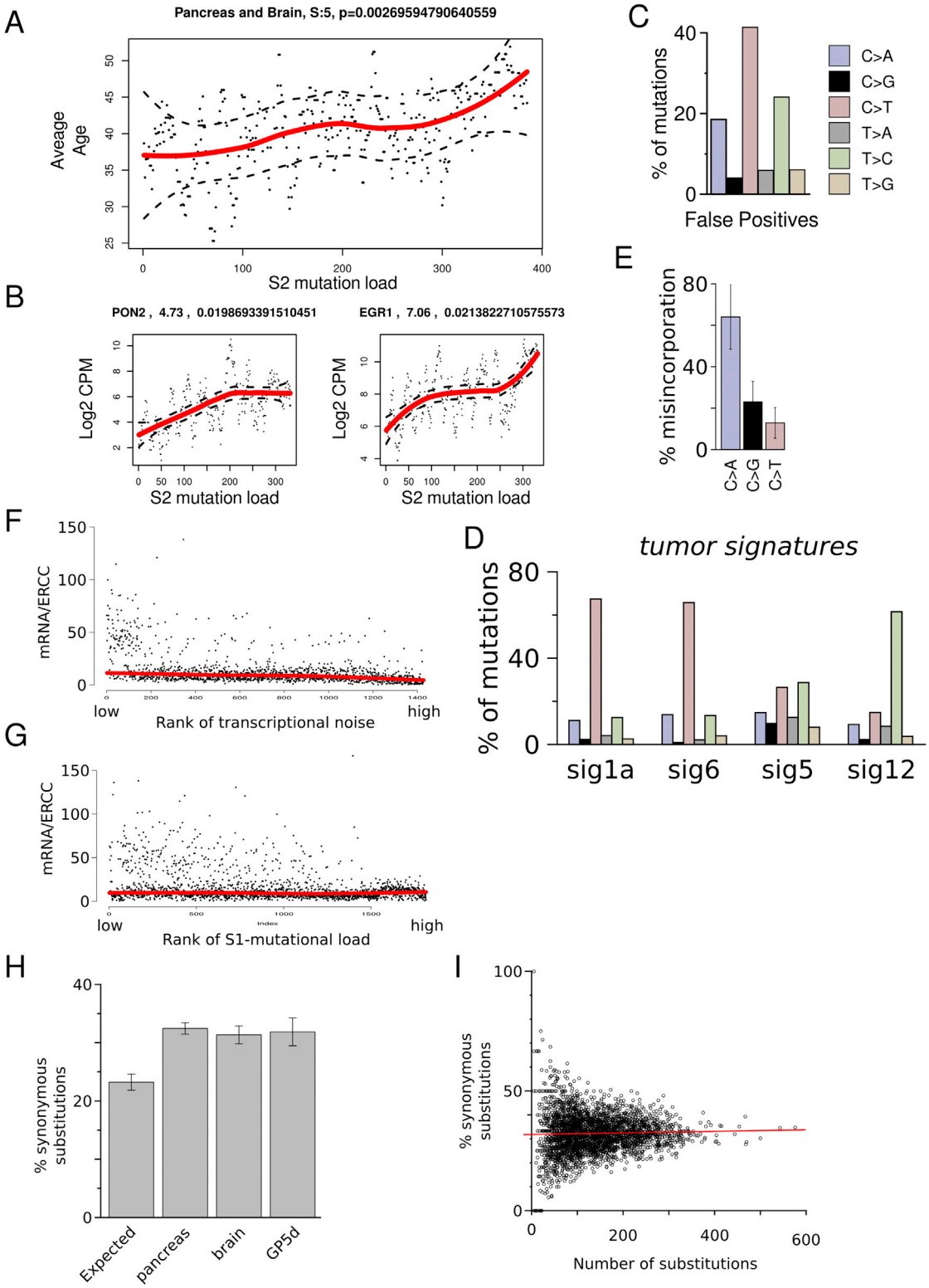

Figure S5. Transcriptional Correlates of Mutational Signatures, Related to Figure 6

Brain cells were ordered according to the fraction of mutations attributed to Signature S2.

(A) Average age is higher in cells with high signature S2 load (p = 2.7E-3, n = 398. linear rank regression). Line is loess fit +-.999 confidence interval. Dots are running mean, k = 10.

(legend continued on next page)

(B) Each gene was tested for association with signature S2 (linear rank regression), shown are the top genes by coefficient, with p < 5E-2 (FDR corrected). Line is loess fit +-.999 confidence interval. Dots are individual observations.
(C) Signature of raw substitution rates in ERCC spike-in RNA constitutes a false-positive signature.
(D) Tumor signatures from Alexandrov et al. (2013b) collapsed into substitution types without 3′/5′ context by addition.
(E) Empirical misincorporation rates caused by 8-Hydroxyguanosine in vitro. Bars are mean ± SEM. Data from from Kamiya et al. (Kamiya et al., 2009).
(F and G) Ratio of human mRNA to spike in control in cells, ordered by rank of transcriptional noise (F) or rank of signature S1 mutational load (G).
(H) Synonymous substitutions generating an identical codon as the reference sequence are enriched in somatic variation from all tissues.
(I) The fraction of synonymous substitutions is not positively correlated with overall mutation load.