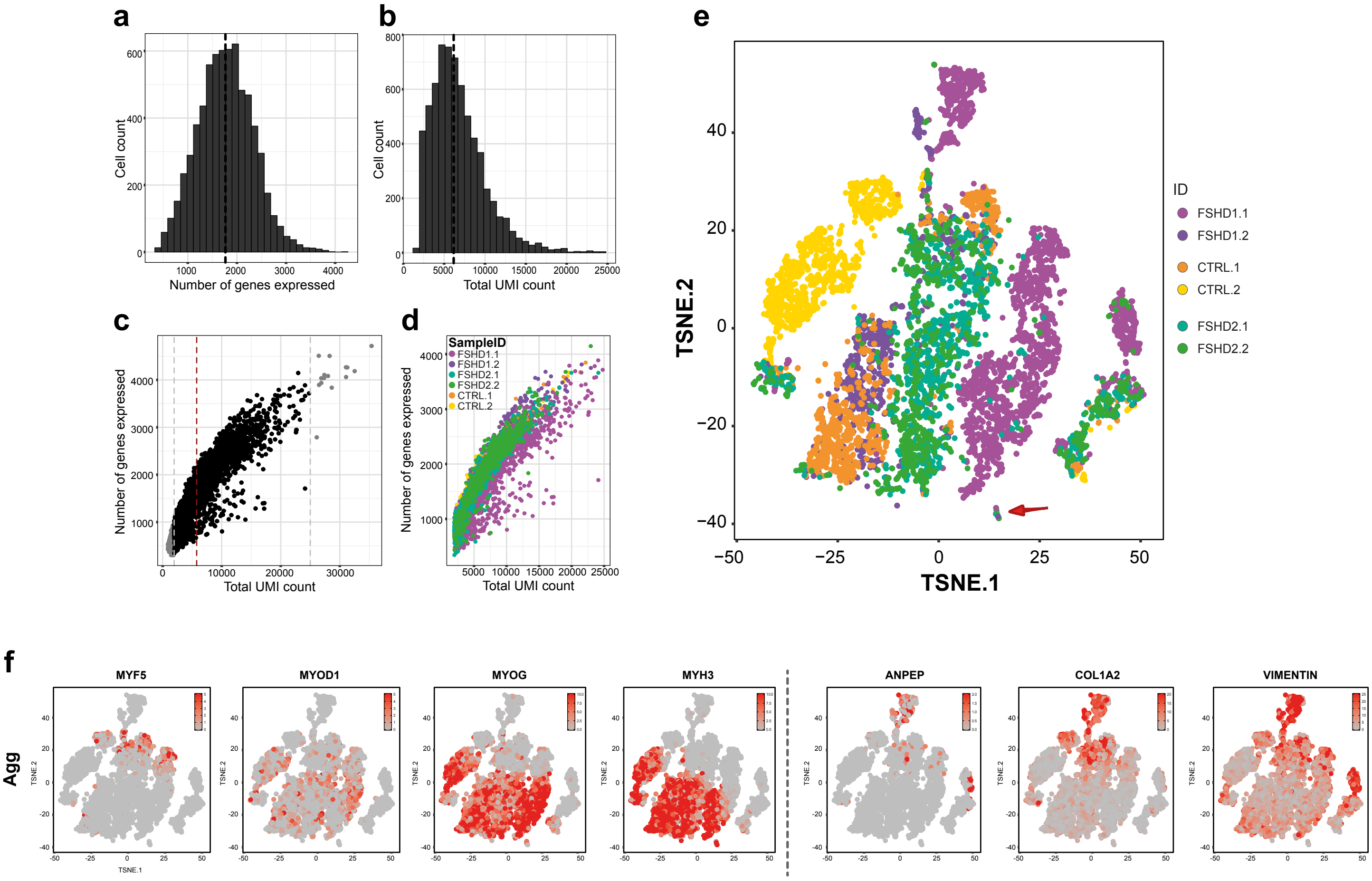

Figure S5. Detailed quality validation of the Agg dataset.

a-d) Agg sample sequence data metrics, with **a)** a histogram depicting the distribution of the number of expressed genes per cell (dashed line; median); **b)** a histogram depicting the distribution of the total UMI counts per cell (dashed line; median); **c)** a scatterplot depicting the relation between the total UMI counts versus total gene counts per cell (red dashed line; mean read counts per cell, grey dashed lines; selected threshold for filtering low quality cells per sample, grey dots; filtered low quality cells). **d)** The scatterplot as in (C), colored for sample identity (SampleID), showing that all samples are similarly represented in the Agg dataset. **e)** t-SNE projection (Cell Ranger) of the Agg dataset (as in Figure 2C) colored by sample identity (SampleID), with the FSHD-specific cell cluster highlighted by the red arrow. **f)** t-SNE projection as in (E), colored based on the expression levels of four myogenic markers; MYF5 (early marker), MYOD1 (intermediate marker), MYOG (intermediate-late marker) and MYH3 (late marker), and three fibroblast markers; ANPEP, COL1A2 and Vimentin. Color-scales depict the normalized expression levels for each specified gene.