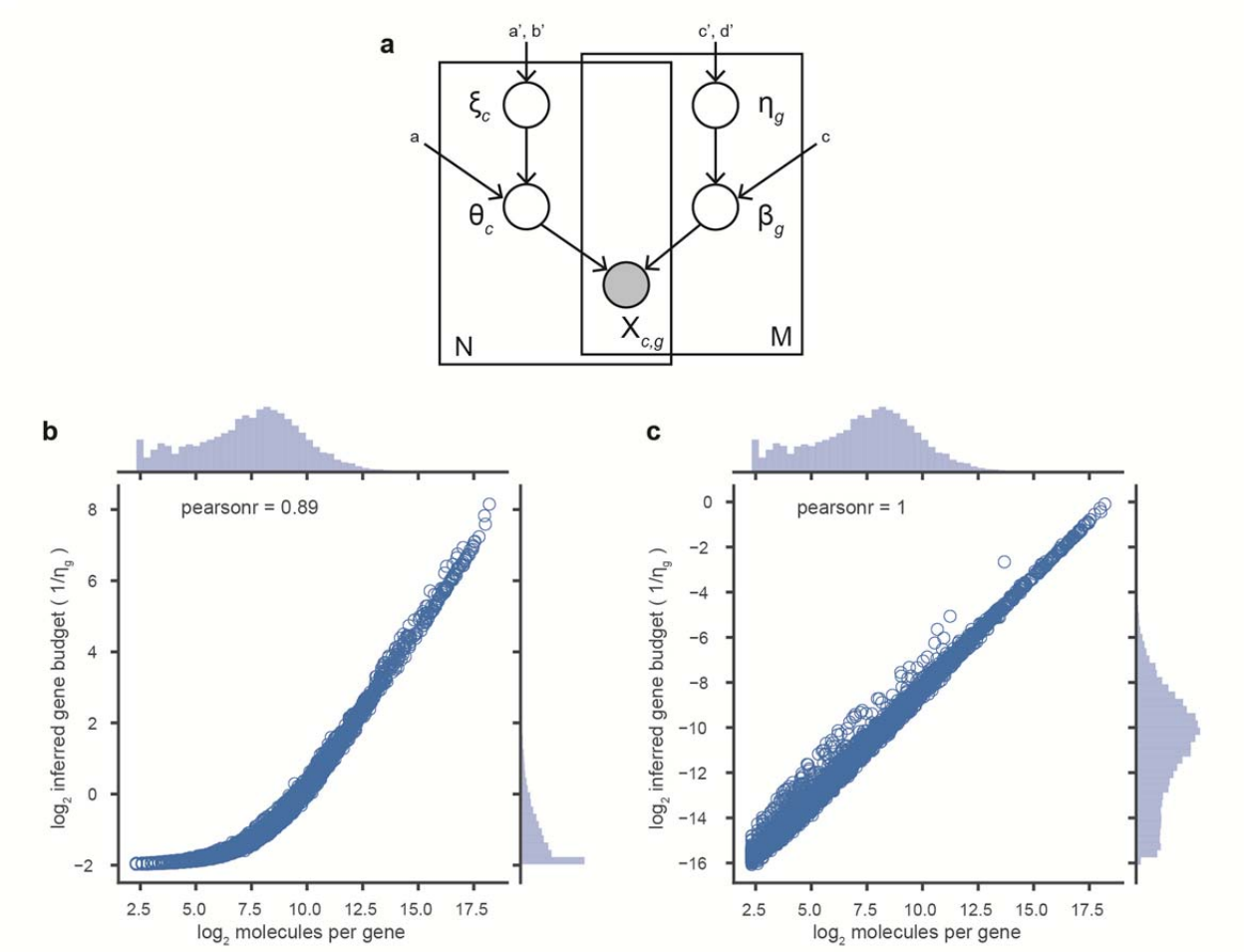
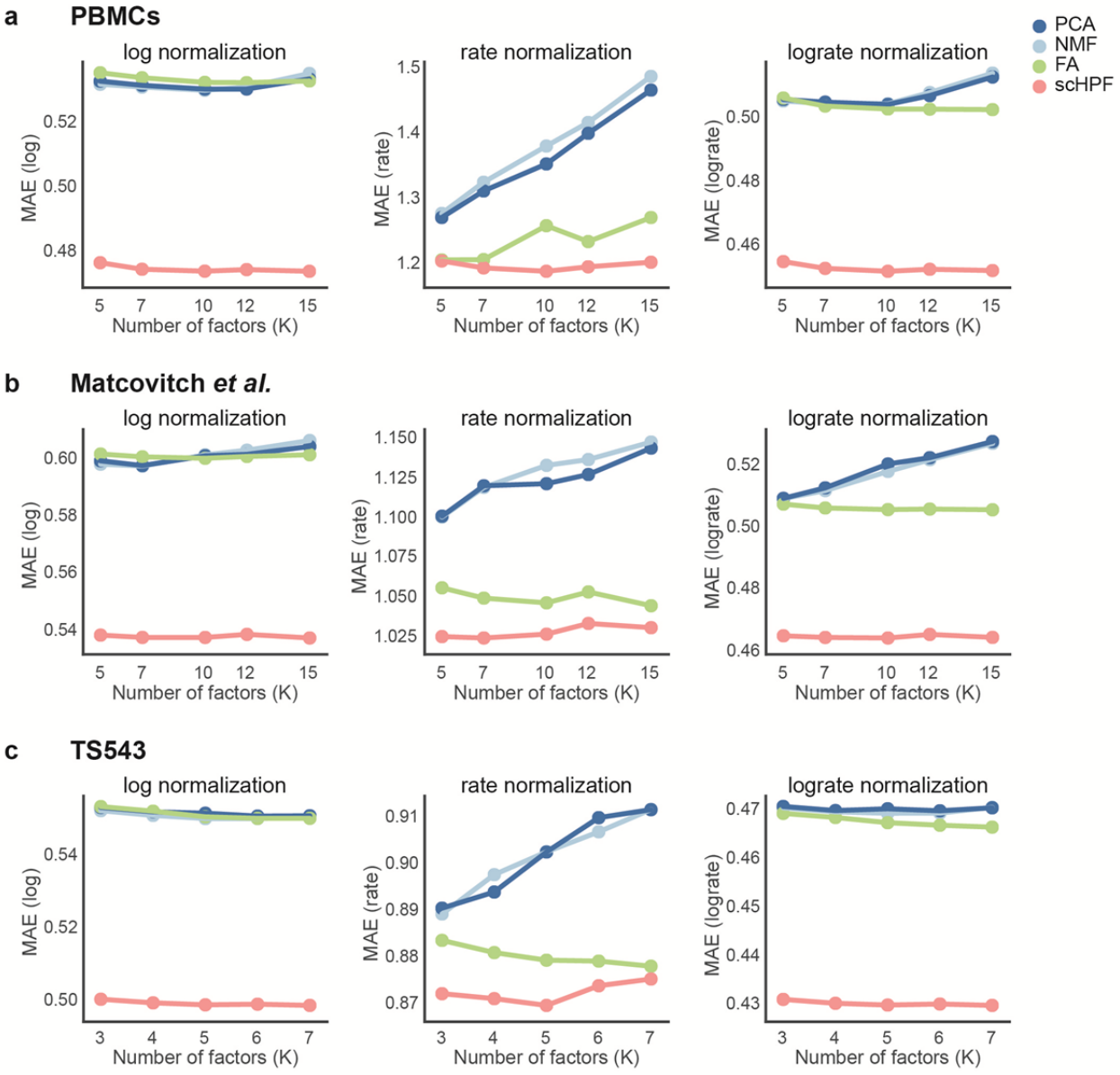
	PBMCs	Matcovitch et al.	TS543	HGG
scRNA-seq platform	10x Chromium	MARS-seq	Yuan & Sims	Yuan & Sims
Number of cells	4340	3456	9924	3109 core 3000 margin
Minimum number of cell expressing gene (for filtering)	5	5	10	10
Number of protein-coding genes after filtering	13030	8086	11807	14730
Sparsity of filtered data	90%	94%	92%	93%
K	10	10	5	14

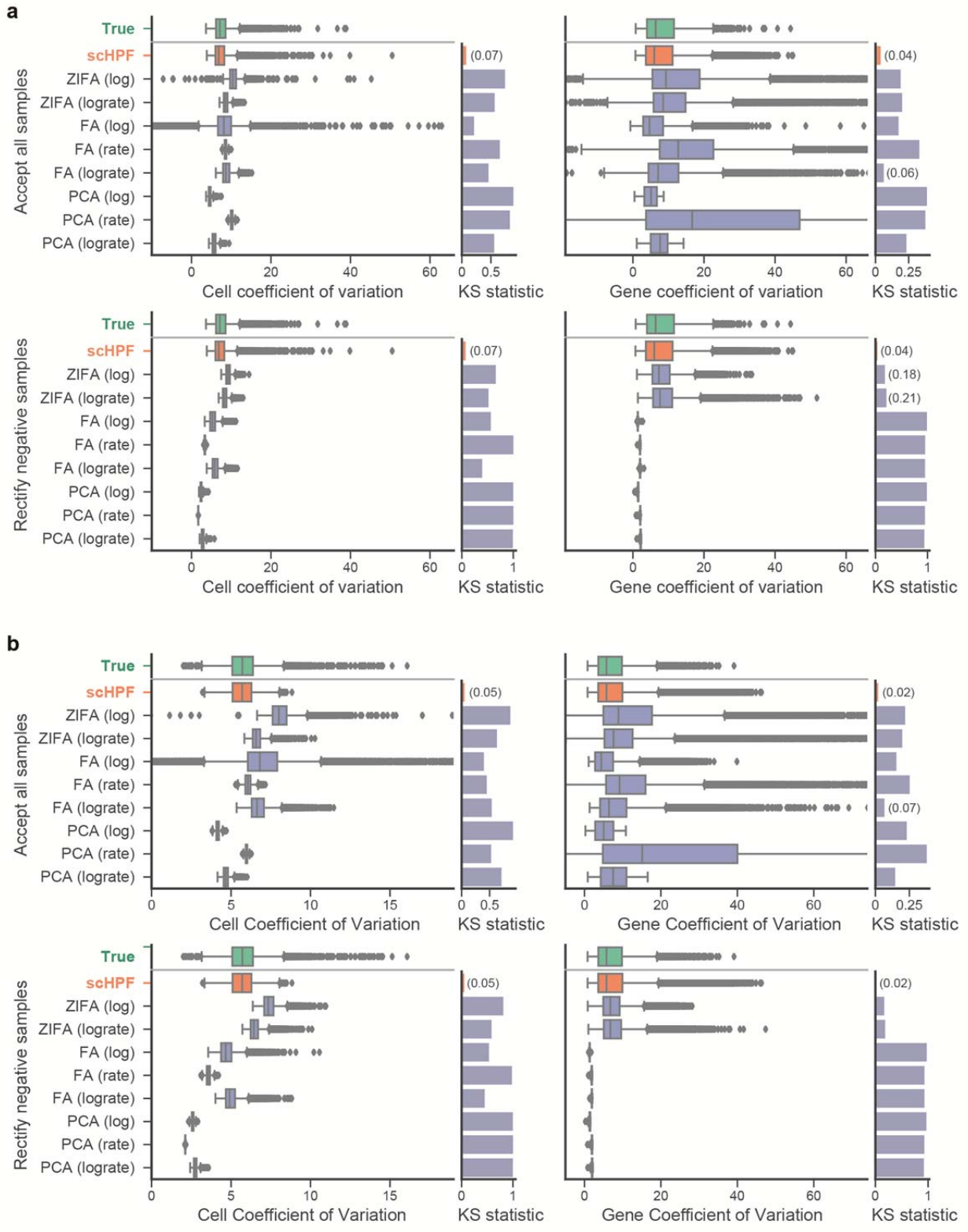
Supplementary Table 1: Datasets and parameters used.



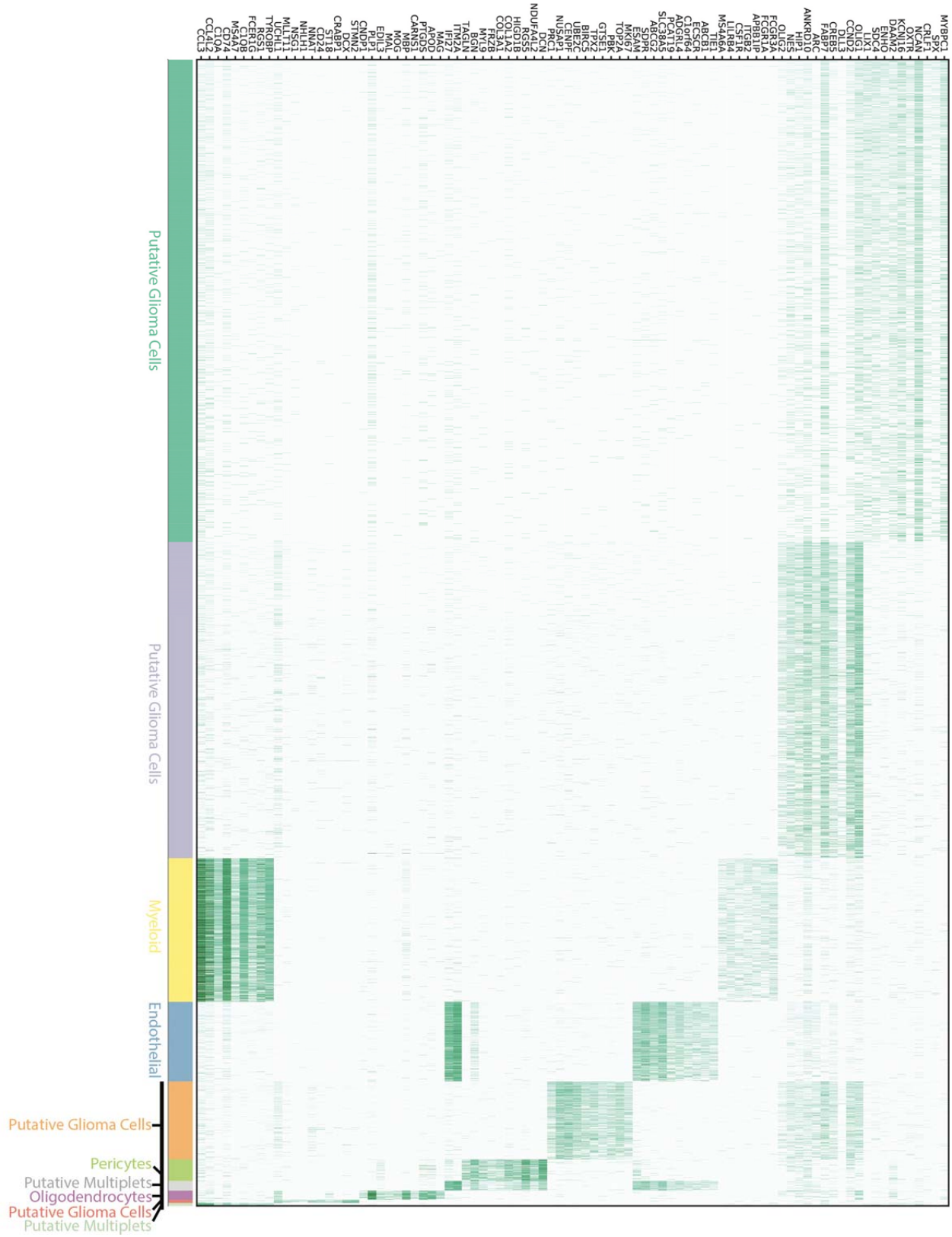
Supplementary Figure 1: (a) scHPF models the data matrix $X_{c,g}$ using a set of per-cell latent factors θ_c and per-gene latent factors β_g. scHPF places hierarchical priors over the latent factors through the latent variables ξ_c and η_g, which probabilistically determine the observed transcriptional output for the cell or gene. **(b & c)** Scatter plots of log2 molecules per gene (x-axes) versus the log2 inferred gene budgets (y-axes), with hyperparameters (b) a', b', c' and d' set to 1 or (c) determined empirically in a representative experiment on peripheral blood mononuclear cells. Histograms on top and right show the marginal probability distributions along each axis.



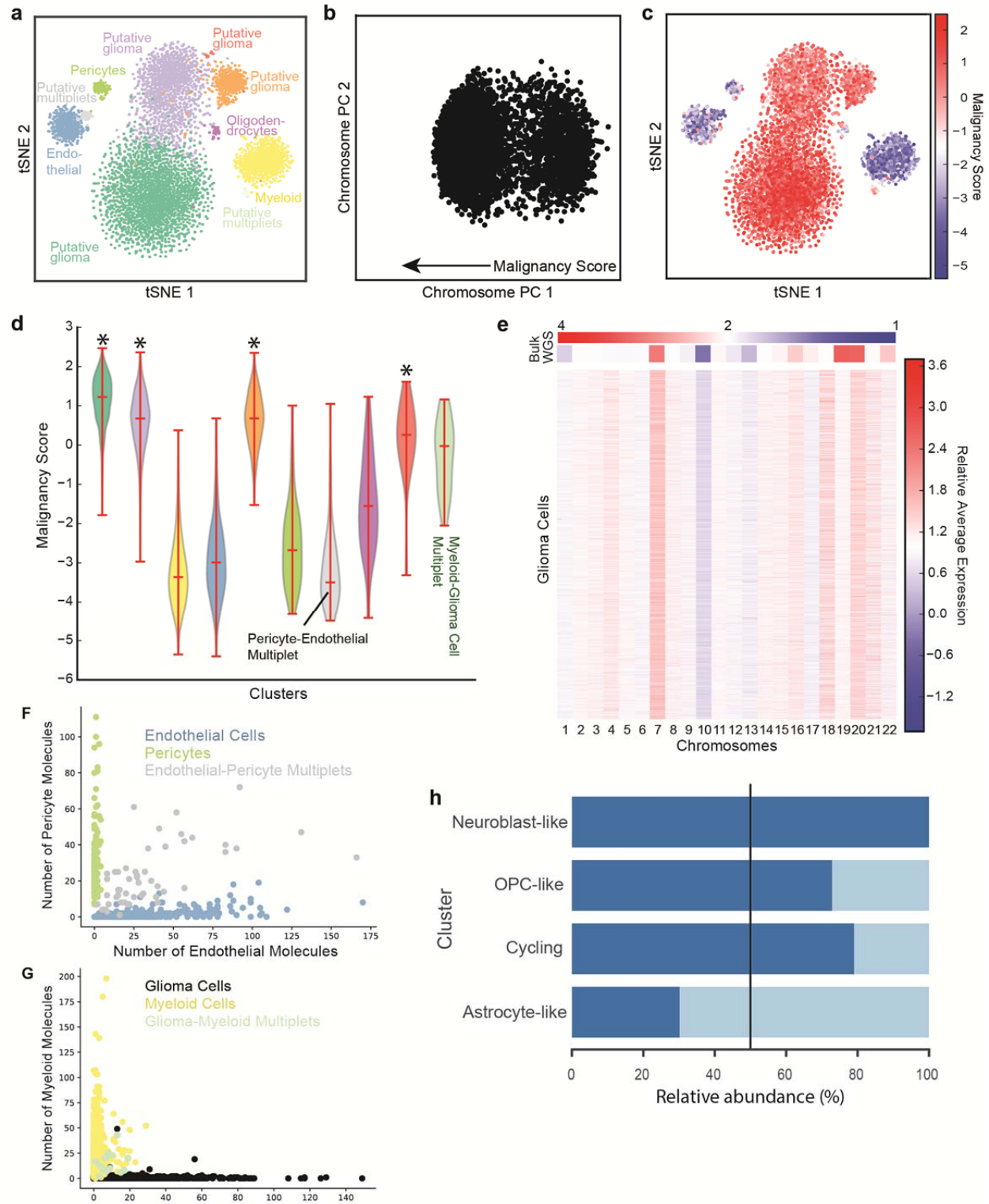
Supplementary Figure 2: Different method and normalization combinations' mean absolute error (MAE) on a withheld partition of the **(a)** PBMC, **(b)** Matcovitch *et al.*, and **(c)** TS543 datasets as compared to scHPF for several different numbers of factors. scHPF's predictions were normalized before calculating error.



Supplementary Figure 3: Same as Figure 2b-c, but for **(a)** Matcovitch *et al*. and **(b)** TS543. X-axes limits for boxplots are set to include all coefficients of variation from the true distribution and scHPF, and as many coefficients of variation from other methods as possible.

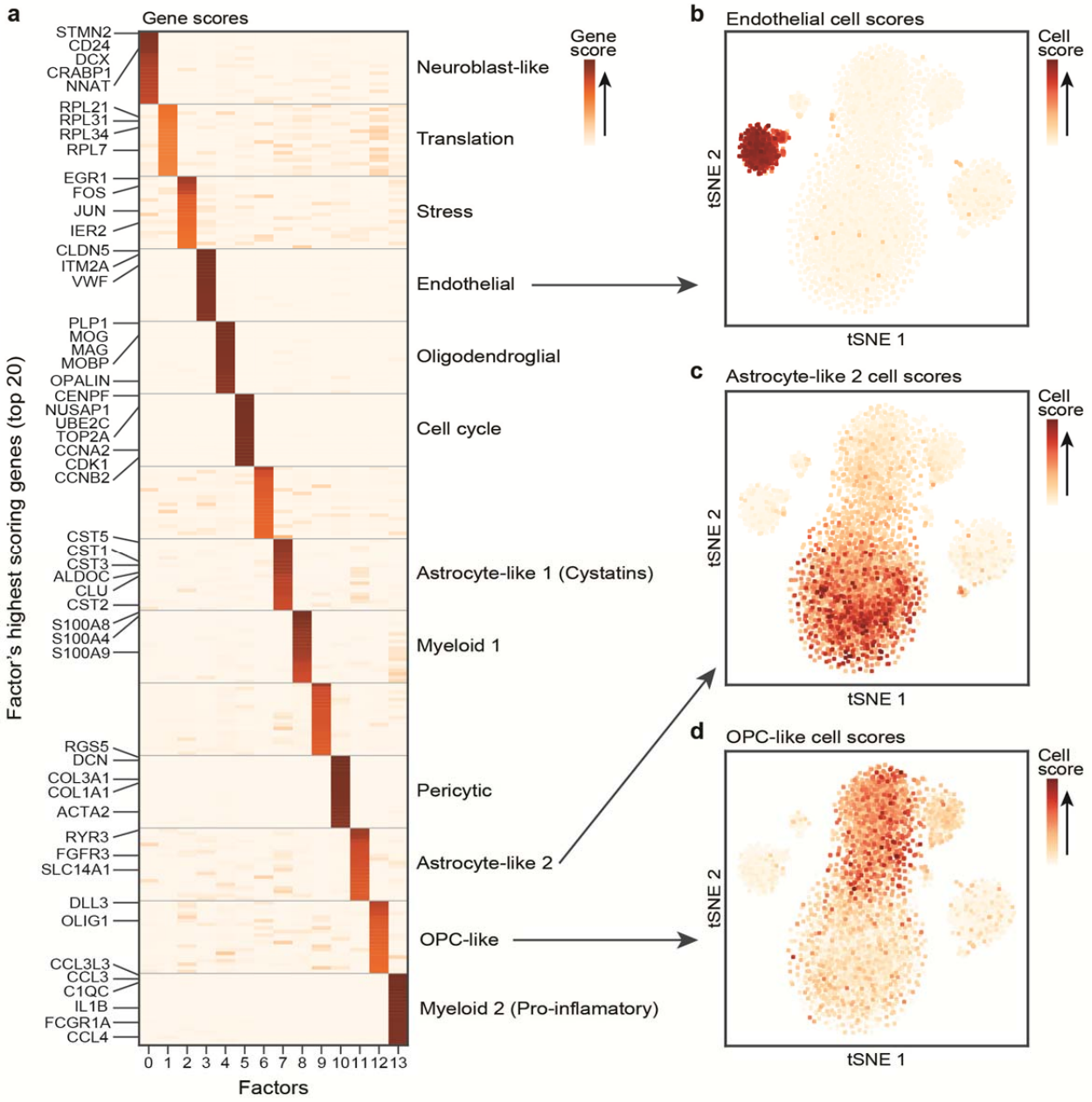


Supplementary Figure 4: Heatmap gene expression in a high-grade glioma with cells (columns) ordered by Louvain cluster (Methods) and genes (rows) selected as the top ten most specific genes in each cluster. Bottom color bar shows clusters and putative labels based on expression of canonical marker genes and aneuploidy analysis (see Figure S5).

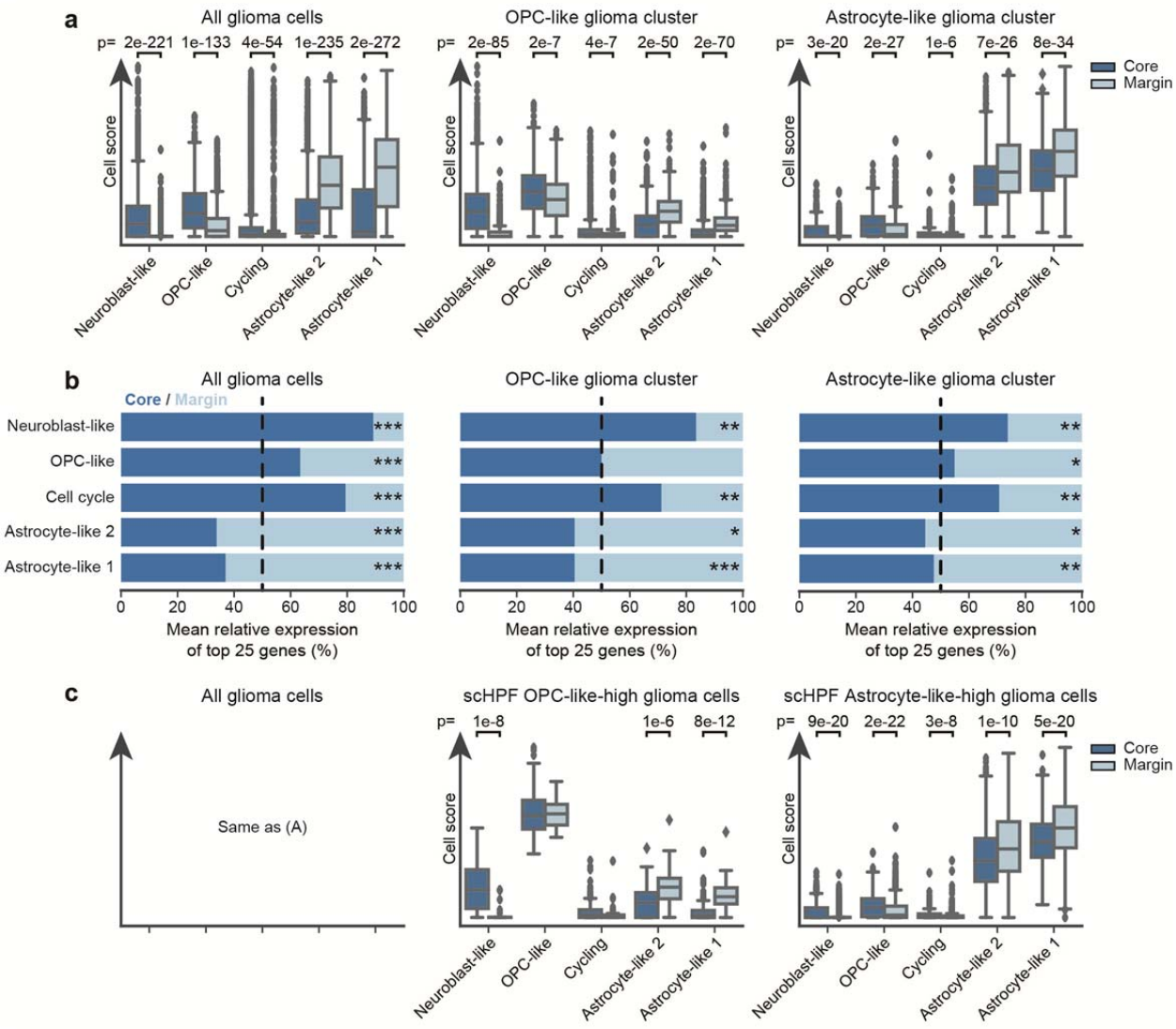


Supplementary Figure 5: (a) t-distributed Stochastic Neighbor Embedding (tSNE) [1] plot of tumor cells, labeled by cluster (also see figure S4). **(b)** PCA of whole-chromosome expression for each cell. The first principle component (PC1), which we call a malignancy score, separates putative glioma from non-malignant cells. **(c)** tSNE plot of all cells, colored by malignancy score. **(d)** Violin plots of malignancy scores for each cluster. Putative glioma clusters are starred. **(e)** Main heatmap shows putative glioma cells' (rows) relative average expression of

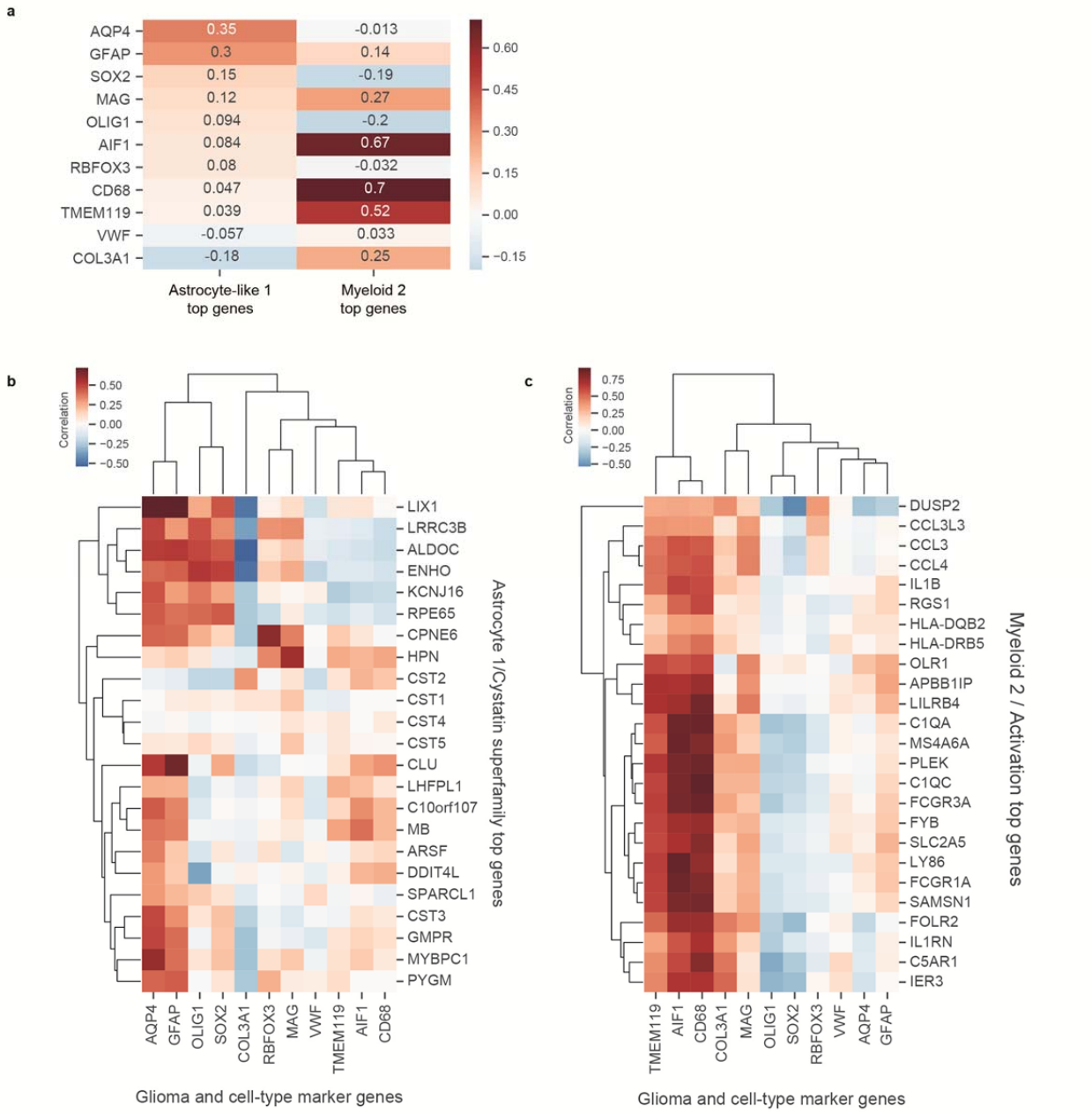
each chromosome (columns). Values generally agree with bulk whole genome sequencing (WGS) of the tumor (top heatmap). **(f)** Barnyard plot of cells in the endothelial (blue), pericyte (green) or endothelial-pericyte multiplet (gray) clusters. Total number of molecules for the ten most endothelial-specific genes by a binomial test are on the x-axis, and total number of molecules for the top ten most pericyte-specific genes are on the y-axis. **(g)** Barnyard plot of all putative glioma cells (black), cells in the myeloid cluster (yellow), and cells in the putative myeloid-glioma multiplet cluster (green). Total number of molecules of the ten most glioma-specific genes by a binomial test are on the x-axis, and total number of molecules of the ten most myeloid-specific genes are on the y-axis. **(h)** Relative abundance of glioma subpopulations in the core (navy) and margin (light blue).



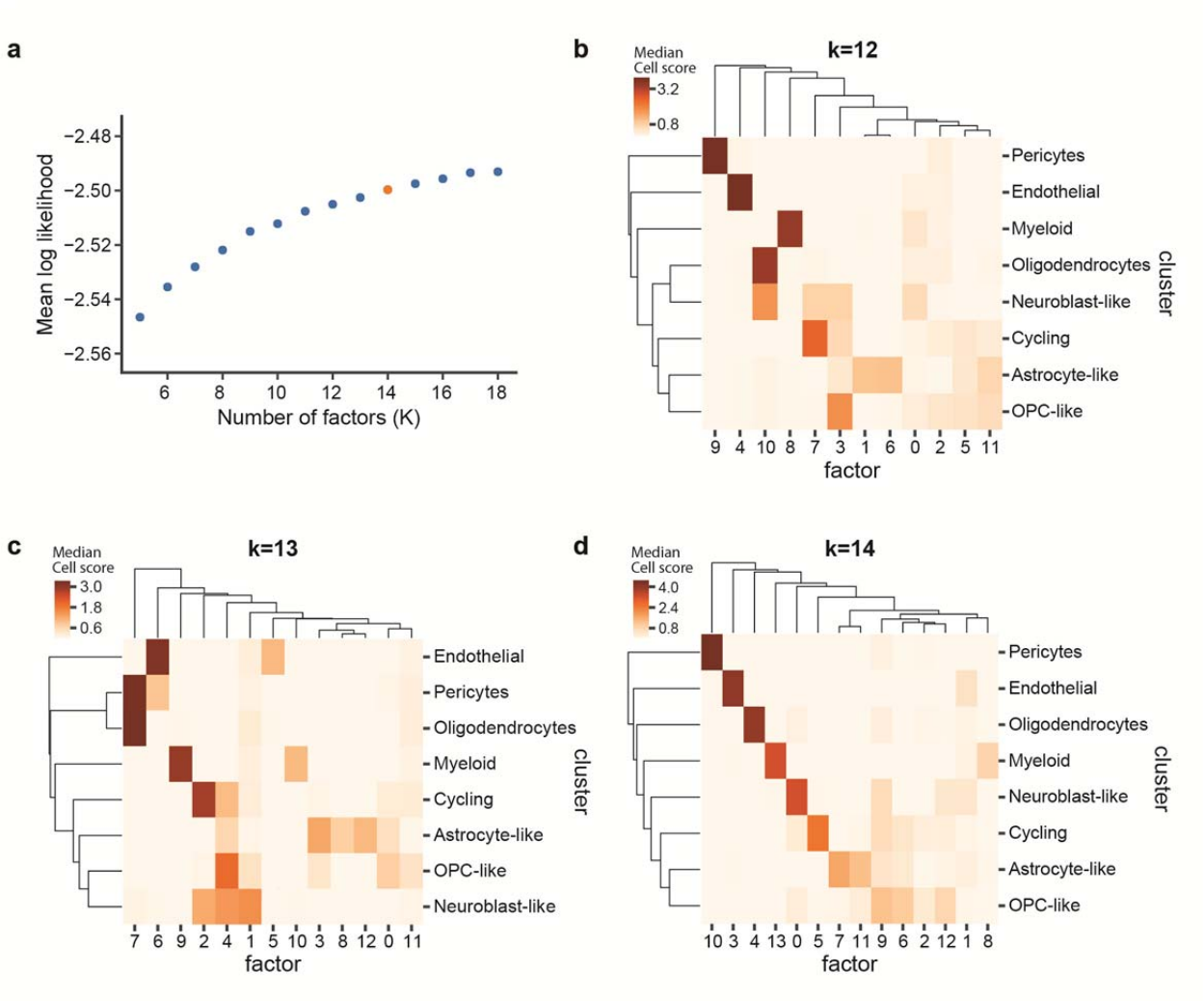
Supplementary Figure 6: (a) Heatmap of scHPF gene scores for each factor (columns) and the top twenty genes per factor (rows). Canonical marker genes and genes from a protein superfamily are highlighted. **(b-d)** tSNE of all cells colored by their scHPF cell scores for a factor that marks a discrete population of endothelial cells (b), one of two glioma-associated factors that highly ranks astrocyte marker genes (c), and a glioma-associated factor that highly ranks OPC maker genes.



Supplementary Figure 7: (a) Boxplots of scHPF cell scores for all glioma cells (left), OPC-like glioma cells (center), and astrocyte-like glioma cells (right) show strong regional bias towards the core (navy) or margin (light blue). Bracketed values show Bonferroni-corrected p-values from the Mann-Whitney U-test for the difference between two distributions. **(b)** Program scores, derived as the mean relative expression of the top 25 genes in each factor, recapitulate cell scores' regional biases. *** = $p < 10^{-50}$, ** = $p < 10^{-10}$, * = $p < 10^{-2}$. All p-values are Bonferroni corrected. Expression values were converted to counts per median and log10 scaled before averaging. **(c)** Same as (a), but with OPC-like and astrocyte-like glioma subpopulations defined as cells with maximal scHPF cell scores in the OPC-like factor or one of the two astrocyte-like factors, respectively.



Supplementary Figure 8: (a) Median correlation of the top 25 genes from two scHPF factors with glioma and cell type marker genes in TCGA GBM RNA-seq. scHPF Astrocyte-like 1 is best correlated the GBM-specific marker, SOX2, and astrocyte markers. In contrast, scHPF Myeloid 2 is best correlated with microglial/macrophage markers. **(b & c)** Hierarchically clustered correlation of marker genes with the top 25 genes from scHPF Astrocyte-like 1 (b) and scHPF Myeloid 2 (c).



Supplementary Figure 9: (a) Mean log likelihood for scHPF of a high-grade glioma at different values of K (higher is better). **(b-d)** Median factor score in each cluster at 12, 13, and 14 factors. With 12 factors (b), oligodendrocytes and neuroblast-like cells are both most closely associated with the same factor. Similarly, with K=13 (c), oligodendrocytes and pericytes are both most closely associated with the same factor. At K=14 (d), all clusters are most closely associated with at least one unique factor.

1. Maaten, L.v.d. and G. Hinton, *Visualizing data using t-SNE.* Journal of machine learning research, 2008. **9**(Nov): p. 2579-2605.