Supporting Information

Cell size contributes to single-cell proteome variation

Michael C. Lanz[*], Lucas Valenzuela[*], Joshua E. Elias, Jan M. Skotheim[+]
[*]Contributed equally
[+]Address correspondence to skothiem@stanford.edu

Table of Contents

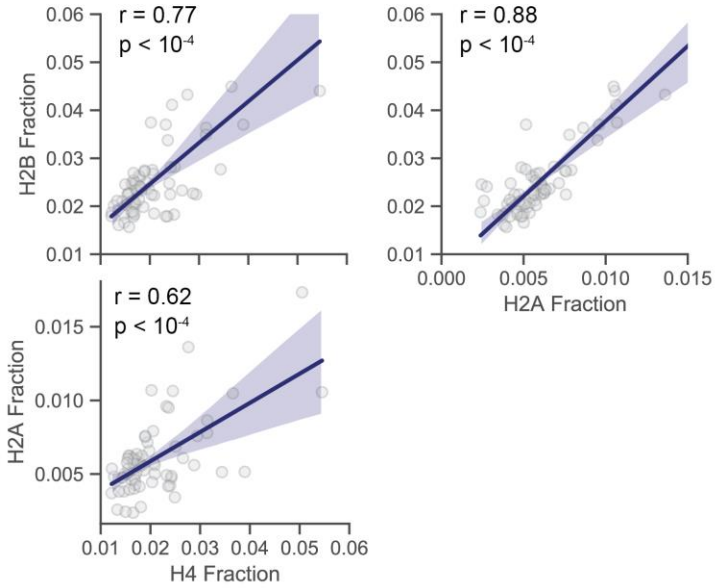

Figure S1: Multiple core histones can be used to estimate cell size.
The concentration of Histone H4 correlated with the concentration of other core histones in single cells. A regression line is plotted in dark blue with 95% confidence intervals. Pearson r value and its associated p-value are shown.

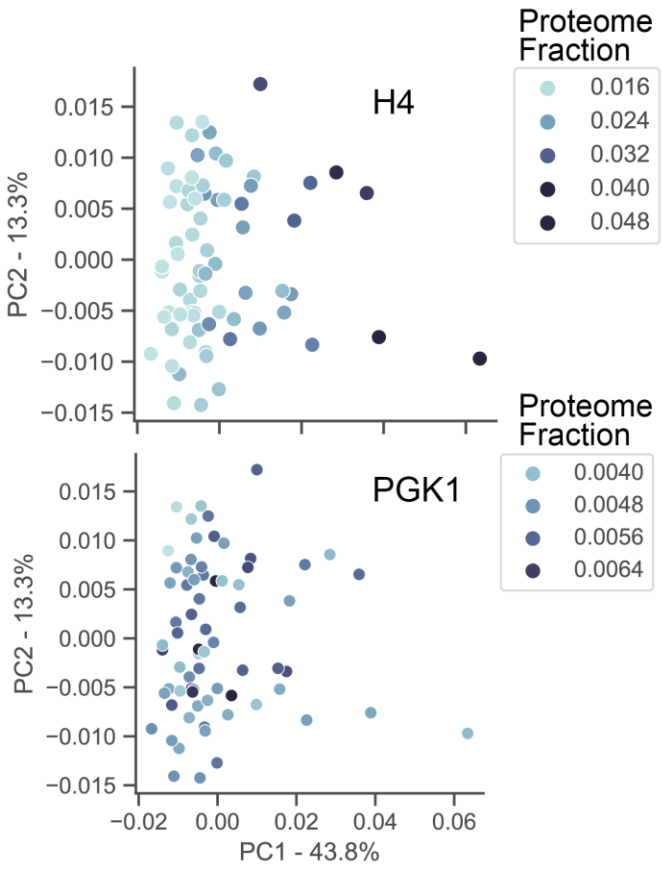

Figure S2: Correlation of principal components with the relative concentration of histone H4 and PGK1, a protein whose concentration is not expected to change with cell size.
Plot of the first two principal components from a PCA analysis of 70 single cell proteomes. Each dot represents the proteome of a G1 cell from Brunner *et al.* and its color indicates the fraction of the proteome represented by either H4 (top) or PGK1 (bottom).

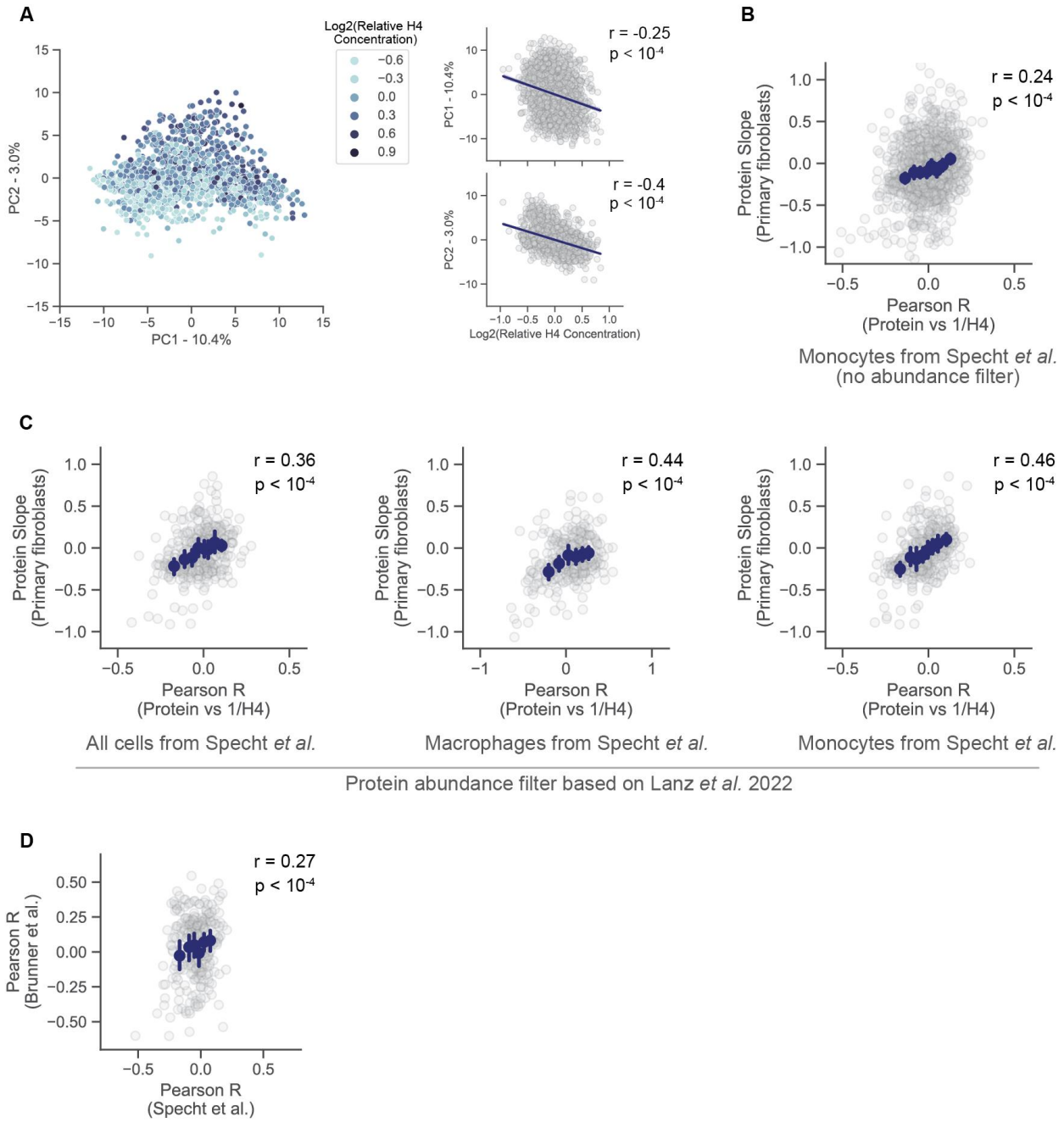

Figure S3: Reanalysis of single cell proteomics data from Specht *et al.*

A) PCA analysis of 1490 single cell proteomes from Specht *et al*. Proteins were quantified using tandem mass tags, so cell size was estimated using the relative concentration of MS2-level reporter ions for Histone H4. Each dot represents a proteome and its color indicates the relative H4 concentration. Correlation between the relative histone concentrations and PC1 and PC2. A regression line is plotted in dark blue with 95% confidence intervals. Pearson r value and its associated p-value are shown.

B) A Pearson correlation coefficient was calculated by regressing the relative concentration of each individual protein against a proxy for each cell's size (histone H4 concentration). The r value for each protein in the Specht *et al.* dataset is plotted against the Protein Slope value (Lanz *et al.*, 2022). Histone H4 was excluded from the plot. Error bars for the binned data represent the 99% confidence interval of the mean. In Figure 1H, only the most abundant proteins are depicted. All proteins identified by Specht *et al.* are included.

C) The same analysis as in (B) but looking at all the different cell populations measured by Specht *et al.* The left most panel is Figure 1H. Unlike panel (B), a threshold for protein abundance is used to filter out noisy protein measurements. Error bars for the binned data represent the 99% confidence interval of the mean.

D) Correlation of Pearson r values (1/Histone) generated from the Specht *et al.* and Brunner *et al.* datasets.

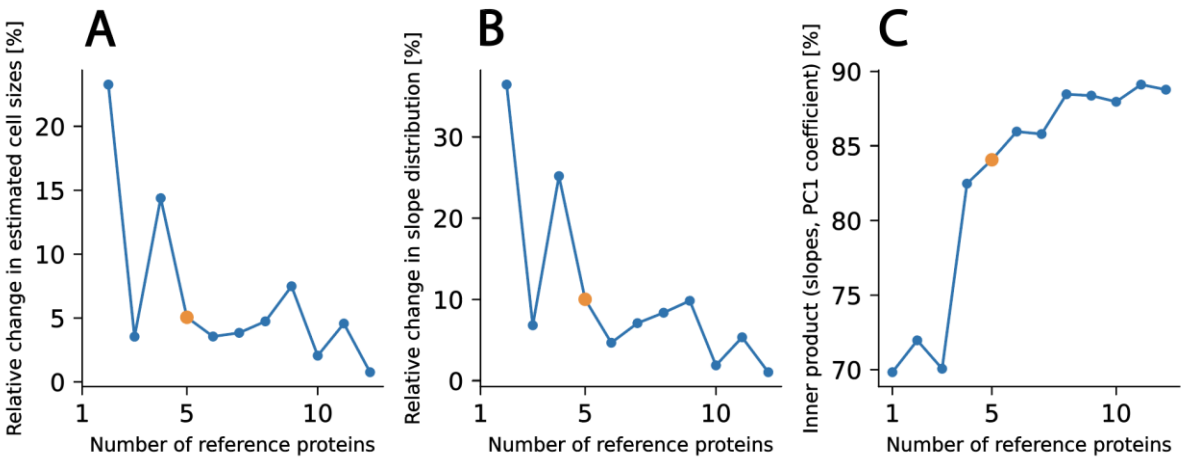

Figure S4 – Metrics used to determine the number of reference proteins.

The orange dot represents the number of reference proteins that is presented in the main text. Relative change denotes the value for n reference proteins minus the value for n-1 reference proteins divided by value for n reference proteins.

A) Relative change in inferred estimated cell size distribution as more reference proteins are added. **B)** Relative change in estimated slope distribution as more reference proteins are added. **C)** Normalized inner product between the slopes and the PC1 coefficient as a function of the number of reference proteins . Reference proteins were discarded from the dataset before computing this metric. We chose 5 reference proteins because beyond this number changes produced only marginal differences.

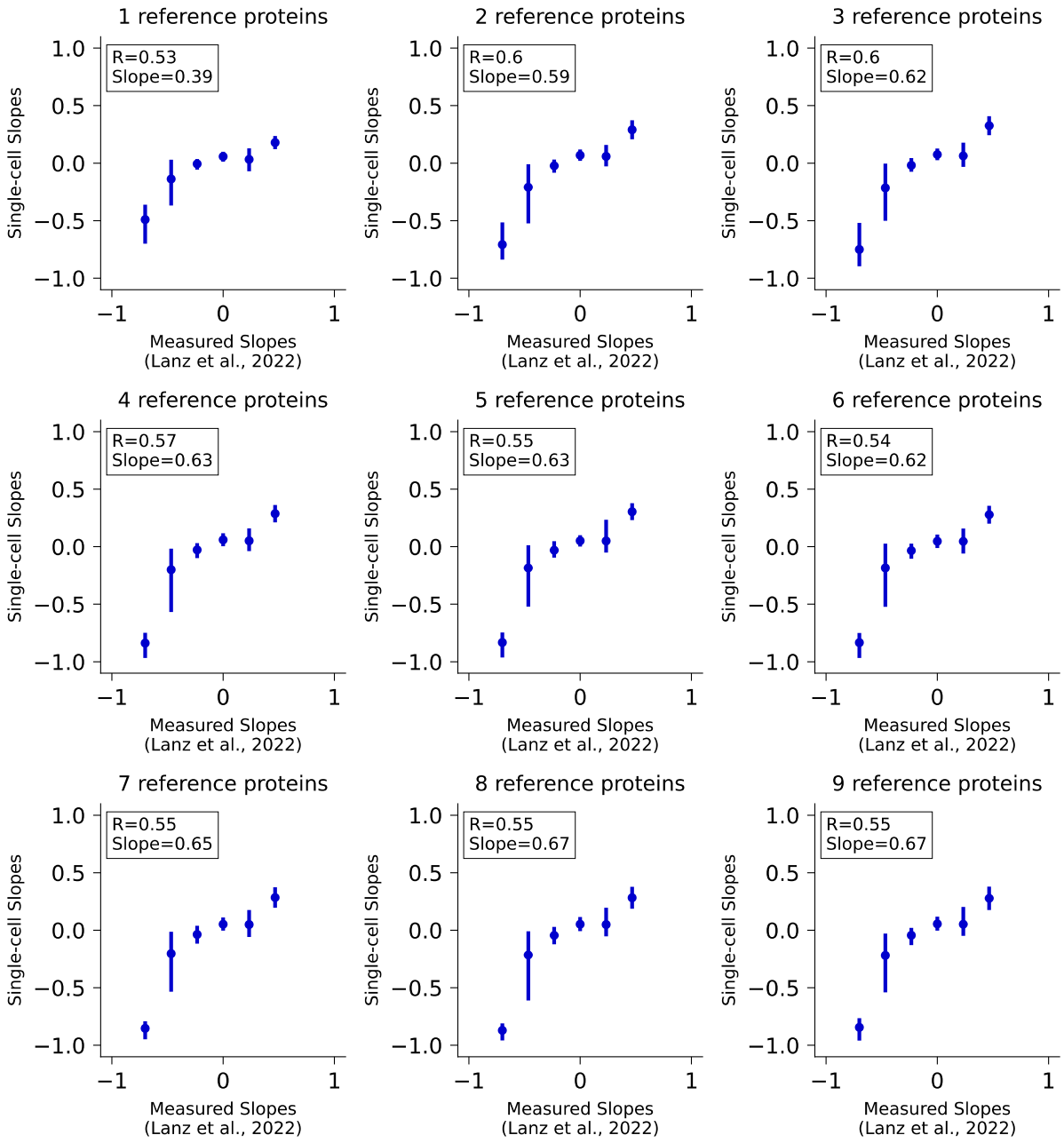

Figure S5 – Impact of increasing number of reference proteins on single cell slopes Each panel reports the slopes estimated with a given number of reference proteins. Orange dots denote the proteins used as reference and blue dots denote binned data and associated confidence intervals.

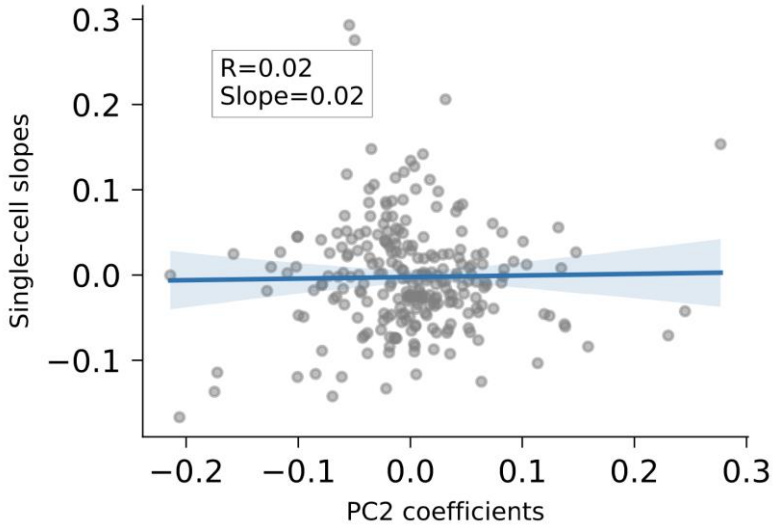

Figure S6 – Correlation of estimated single-cell protein slopes with the coefficients from PC2 from a PCA analysis of *Brunner et al.* as shown for PC1 in Fig. 2E.

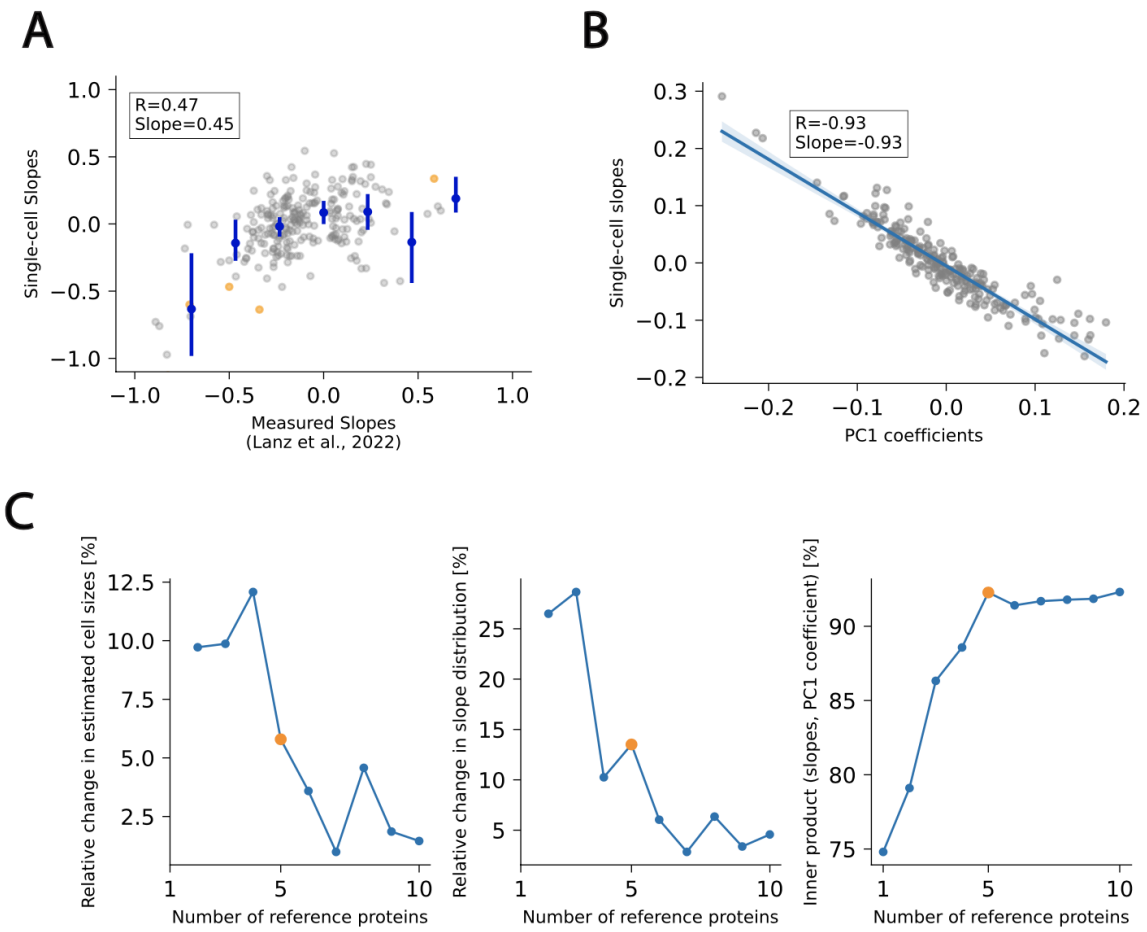

Figure S7 – Analysis of the *Brunner et al.* dataset using HLF protein slopes from *Lanz et al.* rather than RPE1 protein slopes.

A) Comparison of protein slopes estimated via the approach described in this paper and those measured in *Lanz et al*. Orange dots denote the proteins used as reference and blue dots denote binned data and associated confidence intervals.

B) Relationship between estimated slopes and the coefficients of the first principal component.

C) Relevant metrics with increasing number of reference proteins as in **Figure S4**.