Supplemental Information

N. Belliveau, G. Chure, J. Theriot, R. Phillips

March 18, 2020

1 Summary of Proteome Datasets.

Here we briefly summarize the datasets that were considered for the work of the main text. The goal of this section is to give an overview of each dataset considered, including the main experimental details, and to provide a more detailed look at how well each compares.

Table 1 provides an overview of the proteomic datasets that we found in the literatrue. These are predominately mass spectrometry-based, with the exception of the work from Li *et al.* (2014) which used ribosomal profiling, and the fluorescence-based counting done in Taniguchi *et al.* (2010). The general strategy taken in these works is to quantify fractional abundance of each protein and then to convert these to absolute abundance by multiplying these fractions by the bulk measusured total cellular protein abundance. Note that the work of Peebo *et al.* (2014) did not perform any measurement of cell count or volume, and thus were only able to report cellular protein concentration.

Exceptions to this are found in Schmidt *et al.* and Taniguchi *et al.*. A key distinction in the work of Schmidt *et al.* is that in addition to determining relative abundance by mass spectrometry, they also selected 41 enzyme that cover over four orders of magnitude in cellular abundance to use in absolute protein quantification. Specifically, synthetic peptides were generated for each of these 41 enzymes and used to provide a calibration between measured mass spectrometry intensities and absolute protein abundances (using stable isotope dilution (SID) and selected reaction monitoring (SRM), though the details of this are beyond the scope of this section). In the work of Taniguchi *et al.*, the authored tagged each protein with a yellow fluorescent protein (YFP) and used fluoresence as readout of cellular expression.

Author Reported Quantity	Method fractional coverage (by count)	strain fractional coverage (by mass)
Taniguchi *et al.* (2010) fg/copies per cell		YFP-fusion, cell fluorescence
Valgepea *et al.* (2012) fg/copies per cell		Mass spectrometry
Peebo *et al.* (2014) fg/copies per fL		Mass spectrometry
Li *et al.* (2014) protein synthesis rate		Ribosomal profiling
Soufi *et al.* (2015) fg/copies per cell		Mass spectrometry
Schmidt *et al.* (2016) fg/copies per cell		Mass spectrometry
Caglar *et al.* (2017) relative abundance		Mass spectrometry

Figure **??** shows the distribution in reported protein abundance for a subset of the data.

An important consideration is whether the reported abundance per cell are correlated. while we expect some variability in expression of each protein due to growth rate, the reported values are nonetheless expected to be correlated. Figure **??** compares each dataset to the copy numbers from Schmidt *et al.*, grown in M9 minimal media supplemented with glucose.

2 Adjustments to Copy Number Data.

The datasets encompass a range of bacterial growth conditions, different *e. coli* strains, and for those that report quantities on a cell basis, different methods to normalize by cell count and volume. It was therefore important to consider if certain discrepencies exist across the data and whether these might be reasonably dealt with to make the compilated dataset internally consistent. - give reference to what was done in example of yeast proteome data corrections. However, given the work of [cite] and others, there are well-documented expectations about how characteristics such as total protein mass per cell and cell volume should scale with growth rate. We were therefore inclined to only renormalize data in a way that took into account such expectations. Figure **??** shows the total protein mass reported as a function of growth rate for

each experiment. Indeed, with the exception of the work of Peebo et al., the total mass per cell is generally consistent as a fuction of growth rate, and provide some confidence in such an approach.

In the remainder of this section we describe the rescaling that was done to each dataset, with a particular focus on correcting for discrepencies in cellular protein concentration, which may reflect differences in protein extraction efficiency. It is important to note that with the exception of the work from Peebo *et al.* (which is discussed more below), any rescaling is only performed within the data of individual authors and not performed globally. We felt this was important in order not to bias any individuals' work since we lack any true standard of protein abundance.

2.1 Corrections to Enforce a Consistent Cellular Protein Concentration

One parameter that we do not expect to change substantially across growth conditions is cellular protein concentration. As a general rule of thumb, we expect an *e. coli* cell to have about 30% dry mass, with about 55% of this expected from protein. With a density of about 1.1 g/ml, we find that the protein concentration in a cell should be approximately 180 fg/fL. The cellular density and dry mass are essentially fixed, with the fraction of cellular protein varying from [X-Y; refs??]. Hence, this parameter provides a useful reference point that datasets should agree on. Indeed, out of concern over differences in protein extraction efficiency in growth phases like stationary phase, Schmidt *et al.* applied a correction to their measured protein abundances to ensure cellular protein concentrations were internally consistent.

From the work of Schmidt *et al.* they reported an ability to consistently get high protein yield from cells grown in M9 minimal media supplemented with glucose. In order to account any protein loss during extraction, they use their measured protein concentration from this sample as a reference for which total protein concentration in all other growth conditions should match. This is shown in Figure **??**A. One challenge in performing this calculation is that cell volume must be known; the authors use volumes that were measured by flow cytometry in previous work [cite]. These volumes are shown in Figure **??**B. While it is difficult to assess the accuracy of these numbers, we find them to be quite inconsistent with the expected scaling that is reported by Taheri-Araghi *et al.* (2015), carefully measured as a function of growth rate [and other work?].

In addition, since cell volume was not determined in all studies, and to be consistent throughout, we instead use the predicted cell volumes from Taheri-Araghi *et al.*. Dealing with each

dataset seperately, we apply correction factors to correct for discrepencies in protein concentration across the different growth conditions considered [NB: I wonder if in these other datasets, the more appropriate thing to do is match to the average measured protein concentration]. Specifically, the scaling factor ϕ is given by,

$$\phi = \frac{P_i}{V_i} \cdot [P]_r \tag{1}$$

where P_i is the total protein mass in conditino i, V_i is the estimated cell volume, and $[P]_r$ is the reference protein concentration (i.e. growth in glucose for the Schmidt data).

2.2 Peebo *et al.*: Conversion from copies/ fL to copies per cell

In the work of Peebo *et al.*, the authors only report protein concentration. In order to determine protein per cell, we multiple these concentrations by expected cell volumes using the predictions from Taheri-Araghi *et al.* This is shown in Figure **??**A, where we see that reported mass is substantially lower than the other work considered here; as well as work from others [Sinauer, 1990].

Indeed, both Schmidt *et al.* and Li *et al.* reported a total protein mass of about 250 fg per cell at a growth rate of about $\lambda \approx 0.5 hr^{-1}$ (M9 minimal media with glucose and MOPS minimal media, respectively). Given this descrepency, in addition to requiring that cellular protein concentration be internally consistent across the growth conditions they reported on, we also required that total cellular mass be consistent with the work Schmidt *et al.* and Li *et al.* This amounted to performing a linear regression between total protein mass and growth rate, and using this to scale the Peebo *et al.* dataset according to this trend.