# Final Report Storyline

## Methods

### Multiomics data analysis and statistical testing

For the omics data, preliminary data analysis and visualization were performed in R 4.4.1. T tests of the transcriptomics data were likely performed using the DESeq2 package in R, and p values were adjusted using the Benjamini-Hochberg method. For the proteomics data, median rather than mean log2 fold changes (log2FC) were used to be more robust against outliers in proteomics data. Then one sample T tests were performed with the null hypothesis log2FC equal to 1, implying no change between the groups. The p values were then adjusted using the Benjamini-Hochberg method to account for multiple testing.

### Functional annotation

Gene Ontology (GO) annotation was performed separately for up-regulated, down-regulated genes, and up-regulated proteins using ShinyGo 0.82 (add ref). Protein-protein interaction (PPI) network was constructed including all significant genes and proteins using the STRING database (add ref). Analysis and visualization of the network was performed in Cytoscape 3.10.3 (add ref).

## Results

### Transcriptomics and proteomics data distributions

Significantly more genes (3,470) than proteins (679) were detected in the samples, reflecting differences in the underlying detection methods. In transcriptomics, sequencers directly output nucleotide sequences that can be mapped to a reference genome, enabling more comprehensive detection. In contrast, proteomics relies on inferring protein identities through database matching of mass and retention time data, which results in lower detection sensitivity. While neither approach captures the full transcriptome or proteome, it is important to note that proteomic data represent a particularly limited subset of the bacterial proteome, which should be considered when interpreting significant findings.

For both transcriptomic and proteomic data, most absolute log2FC cluster around zero (i.e., raw fold change of 1), suggesting a maintained homeostatic state critical for bacterial survival (Figure X). Adaptations to phosphate and light levels reflected by larger fold changes may thus occur to maintain or minimally disrupt this homeostasis.

The standard error of transcriptomics and proteomics data both inversely correlate with absolute fold change (Figure X). The distributions show that smaller fold changes may be more attributed to technical variation rather than true biological variation.

In the transcriptomic data, fold changes show an inverse correlation with the

mean expression value across samples (*baseMean*) (Figure X). Genes with higher expression levels tend to exhibit smaller fold changes, likely because further up-regulation is energetically costly and may disrupt bacterial homeostasis more significantly.

### Transcriptome and proteome correlation

Overall, there is little correlation between transcriptomic and proteomic log2FC (Figure XA). Only a small number of genes exhibited proteomic fold changes that correspond to transcriptomic fold changes. In particular, after filtering for genes or proteins with absolute log2FC greater than 1, a subset of genes with either low transcriptomic log2FC/high proteomic log2FC or vice versa are identified, contracting expectations of the central dogma (Figure XB). These discrepancies suggest the presence of intermediate regulatory mechanisms that may be important for understanding bacterial responses to phosphate conditions and require further investigation.

### Duplicate translations

Four proteins detected are each translated from two different genes (Table X). The reasons here, highlighting the need for protein isoform specification in the database. Interestingly, genes for some of the same protein do not have close baseMean and log2FC, suggesting more nuance regulatory mechanisms for protein levels (Figure X).

| Protein | Gene | Function |
|---------|------|----------|
| Q6YRS9 | slr6016, slr6075 | Unknown |
| P16033 | psbA2, psbA3 | Photosystem II protein D1 |
| P09192 | psbD2, psbD | Photosystem II protein D2 |
| P74750 | DnaE-c, DnaE-n | DNA polymerase III subunit alpha |

### Investigation of NA ratios in proteomics data

The proteomic data include many proteins with missing (NA) log2FC values. This happens when a protein is not detected in one or both experimental groups. Since these proteins appear in the dataset, they were detected in at least one sample. If they are missing in all three pairs of replicates, there are 3 NA values, making it impossible to calculate a meaningful fold change between conditions. There are at least four plausible, non-mutually exclusive explanations for these missing values:

1. Spurious hit: a fragment from another protein may have been incorrectly matched to the protein in question. This can occur due to:
   1. Structural similarity between proteins, causing shared fragments with similar retention times.

2. Structural isomers, fragments with the same mass-to-charge ratio (m/z) but different structures, being indistinguishable by mass spectrometry, especially if they elute at similar times.
3. Low-resolution mass spectrometry that cannot distinguish between fragments with very close m/z values (e.g., 70.03 vs. 70.05), leading to misidentification.

2. Post-translational modifications or isotope enrichment: these changes can alter the mass of protein fragments, preventing correct matching in the database.
3. Low abundance: the protein may be present at very low levels, detectable only in some samples when it happens to exceed the detection threshold.
4. True biological absence: the protein may be genuinely absent in one group and expressed in the other, making it a potentially important finding.

If case 4 is true, these NA values could reflect meaningful biological differences. Visual inspection of the transcriptomic fold changes and p values for these proteins showed no consistent pattern, suggesting they are not detected likely due to random variability and not associated with experimental bias (Figure X). However, without details about the instrumentation, analysis pipeline, and raw data, no further judgement on the likelihood of the first 3 cases can be made. The significance of these proteins across conditions could be further explored by examining their connections to the identified and quantified differentially expressed genes and proteins in functional annotations and PPI networks. However, given the unclear reasons for their non-detection and the generally weak correlation between transcriptomic and proteomic log2FCs, any relationships discovered would provide only limited support for more robust findings and new conclusions drawn from these connections would have limited reliability.

**Transcriptome and proteome T test**

Using a threshold of $|\text{log2FC}| > 1$ and adjusted p value $< 0.05$, 153 significantly up-regulated and 120 down-regulated genes were identified (Figure XA). Several adjusted p values became 0 due to being smaller than the lower limit of floating point representation and were treated as significant. Using a threshold of $|\text{log2FC}| > 1$ and adjusted p value $< 0.1$, 11 significantly up-regulated genes were identified (Figure XB). Overall the p values of proteins were higher than those of genes because the standard error of the proteomics data is much higher, leading to lower confidence in the fold changes observed (Figure X).

**GO functional annotation**

GO analysis of the up-regulated genes revealed significant enrichment for phosphate transporter-related terms, indicating increased phosphate transporter activity in the phosphate-limited group (Figure XA). In contrast, down-regulated genes were primarily enriched for ribosome and translation-related GO terms, suggesting a reduction in overall translational activity under phosphate limitation (Figure XB). GO analysis of the up-regulated proteins also identified enrich-

ment in phosphate transporter-related terms, confirming increased phosphate transporter activity at both the transcriptomic and proteomic levels (Figure XC). However, only 62% of up-regulated genes, 80% of down-regulated genes, and 82% of up-regulated proteins were annotated in the GO database. This incomplete coverage suggests that some relevant functional annotations may be missing from the analysis.

**Protein-protein interaction network analysis**

**Clusters identified in network** Three major clusters are evident in the network (Figure X). Two correspond to the phosphate transporter and ribosomal genes identified through GO analysis. The third cluster primarily includes genes and proteins involved in cell-cell communication, environmental sensing, and intracellular signaling, most of which are up-regulated. Further investigation of the genes linking these clusters may uncover key regulatory relationships and shed light on how these functional modules interact under phosphate-limited conditions.

**Connection between phosphate transporter gene and cell division gene**
The reduced cell count observed in the phosphate-limited group may result from decreased cell division. A network pathway linking the phosphate transporter cluster to a cell division gene supports this connection (Figure X). Functional annotations suggest that cellular anabolic processes mediate the relationship between phosphate availability and cell division (Table X). This aligns with the earlier hypothesis that bacteria adjust to environmental conditions in ways that preserve internal homeostasis.

| Gene | Functional annotation | Compartment | Gene state in data |
|------|----------------------|-------------|--------------------|
| pstA | phosphate transporter | Membrane | Up-regulated |
| sphX | phosphate binding | Membrane | Up-regulated |
| sll0540 | ATP-binding cassette transporter | Membrane | Up-regulated |
| sll0185 | Cellular anabolic processes | Intracellular | Up-regulated |
| slr0373 | Cellular anabolic processes regulation | Intracellular | Down-regulated |
| slr0374 | Cell division, cellular process | Intracellular | Down-regulated |

**Independent phospholipid synthesis gene** Because cell growth depends on phospholipid synthesis, we examined the list of significant genes and proteins with related functional annotations. We identified *slr0496*, an up-regulated gene involved in phospholipid synthesis, which is not connected to other nodes in the PPI network. Although the corresponding protein was not detected, the gene's up-regulation is consistent with the observed increase in cell volume under phosphate limitation. Its isolation from the network may imply the presence of yet unidentified regulatory pathways.

**Hypotheses for observed changes**

In the phosphate-limited group, we observed the following changes:

1. Up-regulation of phosphate transporter expression
2. Down-regulation of ribosomal genes and reduced translational activity
3. Decreased cell number
4. Reduced ploidy
5. Increased cell volume

To explain these observations, we propose three hypotheses, illustrated in Supplementary Figure X. The following assumptions (A) are applied either universally or selectively across the hypotheses:

1. For all of the hypotheses the underlying assumption is that under resource-limited conditions, bacteria prioritize individual survival over reproduction to remain competitive.
2. In a phosphate limited environment, growth is favored over reproduction due to the phosphate requirement for DNA replication.
3. More protein needs to be translated during division than growth. Hence there is a energy gain when division is stopped.
4. There is a physical limit to the number of phosphate transporters that can occupy the membrane due to constraints like transporter size, lipid packing, and interactions with other membrane components.
5. DNA is used as phosphate storage in cyanobacteria.

**Hypothesis 1: Division halt supports transporter expression**

Under low-phosphate conditions, bacteria increase phosphate transporter expression (Observation 1). To support this, they halt cell division (Assumption 2) and redirect the energy saved from reduced protein synthesis (Assumption 3) toward transporter production. The cessation of division decreases the need for duplicating DNA and proteins, explaining Observations 2, 3, and 4. Continued growth without division leads to increased cell volume (Observation 5).

**Hypothesis 2: Growth-driven transporter up-regulation**

Phosphate scarcity triggers a shift from division to growth (Assumption 2), resulting in larger cell volume (Observation 5). As the cell expands, new membrane synthesis naturally includes more phosphate transporters (Observation 1). The energy saved from suppressed division (Assumption 3) supports this biosynthesis, which also leads to Observations 2, 3, and 4, as in Hypothesis 1.

**Hypothesis 3: Membrane demand drives growth and DNA degradation**

In response to phosphate limitation, bacteria increase phosphate transporter expression (Observation 1). To accommodate more transporters, additional membrane area is required (Assumption 4), driving cell growth and thus Observation 5. To meet the phosphate demand for new membrane synthesis, DNA is degraded to release phosphate (Assumption 5), explaining the reduced ploidy

(Observation 4). Ongoing DNA degradation prevents cell division (Observation 3) and reduces the need for translation of division-related proteins (Observation 2).

## Follow-up experiments

### Hypothesis 3

In order to validate hypothesis 3, the assumption that there is an upper limit to phosphate transporter density on the membrane needs to be first validated. A modified strain with fluorescently labeled phosphate transporters can be grown in low and normal phosphate conditions with 4 replicates each like in this study. In addition, an unlabeled control strain should be grown in the same set-up to confirm fluorescent labeling does not disrupt normal transporter activity and cell function. Using fluorescence microscopy combined with cell number and cell size calculated from the CASY counter, a normalized fluorescence intensity reflecting phosphate transporter density can be calculated for each group. If the intensity is approximately equal, it demonstrates that in low and normal phosphate conditions, the bacteria maintains a similar phosphate transporter density.

Hypothesis 3 further proposes that the need for more membrane space promotes cell size increase. Therefore, if the upper limit for phosphate transporter density can be removed, allowing more phosphate transporters to be allocated while keeping the membrane area the same, the cell size increase observed should decrease or diminish. The transporter density could be altered with two approaches. First, genes that regulate phosphate transporter density could be identified and knocked out. However, since it likely involves a complex set of genes, a second approach that screens or engineers a cyanobacteria strain with naturally higher maximum phosphate transporter density can be applied. One possible strain engineering approach is using adaptive laboratory evolution (add ref, https://microbialcellfactories.biomedcentral.com/articles/10.1186/1475-2859-12-64). Cyanobacteria strains would be grown under a low phosphate and low carbon condition to promote higher transporter expression while limiting growth. The phosphate transporter density can be dynamically tracked using fluorescence microscopy. The strain with the highest phosphate transporter density can then be selected.

Then two groups, control (unmodified) and KO (gene knockout, screened, or engineered strain), could each be grown in phosphate limited and normal phosphate conditions with 4 replicates. Their cell number, cell size, and phosphate transporter density can be quantified and compared to confirm if the bacteria size growth is the result of needing more space for phosphate transporters.

## Discussion

- need detailed LC/GC-MS protocol to understand NA ratios

- proteomics data have high standard error, need more sample size and tighter control of experimental variability
- need more comprehensive proteomics workflow to capture more proteins