Electronic Supplementary Material (ESI) for Analyst.

Supporting information for:

A microfluidic approach to study variations in *Chlamydomonas reinhardtii* alkaline phosphatase activity in response to phosphate availability

Methods

Cluster Analysis via HDBSCAN: Data from the single-cell response of algae cells was analyzed using an unsupervised machine learning approach to identify emerging patterns associated with the availability of P. Unsupervised Machine Learning (UL) comprises a set of methods in which a computational model is fitted not to predict an outcome, but to identify patterns in the data. For two N-dimensional data points x_1, $x\_2 \in R^N$ a similarity measure can be calculated from their features, that is, the entries of the N-dimensional vector describing each data point. This information can then be used to organize data points that are similarly close to each other and far apart from data points that are dissimilar. The result is a representation of the data with highly dense regions, called clusters, separated by regions of low-density. This representation eases the interpretation of the underlying patterns in the data. A number of computational approaches exist to obtain clusters from data. These clustering methods can be roughly classified as flat, if all clusters are created simultaneously, or hierarchical, if clusters are created sequentially. The former type of clustering methods requires to know or assume the number of clusters to be created, which limits its application. Clustering methods can also be classified as centroid-based, or density-based. Centroid-based clustering methods start by finding a candidate center for the cluster and then populate the cluster following a given distribution assumed for the data. Density-based clustering on the other hand does not assume any distribution *a priori* and is therefore more versatile than centroid-based clustering. To reduce the number of assumptions to be made about the data, the hierarchical density-based spatial clustering for applications with noise (HDBSCAN) method [1,2] is adopted in this work. Hyperparameters are user-defined variables that specify or constrain the way the clustering method searches for patterns in the data. For HDBSCAN, these hyperparameters are *min_cluster_size* which controls how small are the clusters allowed

to be, *cluster_selection_epsilon* which affects how the clusters are split up and counterbalances min_cluster_size, and *min_samples* that regulates how many data points are considered as noise. These hyperparameters are set by inspection as their effect is highly dependent on the nature of the data being analyzed. In the present work, HDBSCAN was implemented with min_cluster_size = 10, cluster_selection_epsilon = 0.0, min_samples = 18.

Since HDBSCAN is based on density, its performance could be hindered by high-dimensional data. To deal with this, dimensionality reduction (DR) is often implemented on the data before the clustering step. Generally, DR methods can be divided into matrix factorization methods, and neighbor graphs methods. A classic example of matrix factorization DR, the principal component analysis (PCA) method [3], seeks for the loading matrix W that, when multiplied by the original data matrix X, produces the reduced representation of the data T=XW. This procedure contains an implicit covariance analysis that limits the methods to only consider second order interactions in the data and its application is not suitable for non-linear data. A neighbor's graph method, called uniform manifold approximation and projection (UMAP) [4], starts by creating a neighbors graph in the original high-dimensional space and then finds the topological manifold that minimizes the error between the original data and the data reconstruction from the low-dimensional representation. An important characteristic of UMAP is that is also tries to preserve the relationship among the variables in the original space in the low dimensional space. Since the data cannot be assumed to be linear, and since the relationships in the original space are important, UMAP was adopted for the analysis here presented. Like clustering, DR methods require hyperparameters to be set for a particular implementation. For UMAP, the final number of dimensions (n_comp), the number of neighbors to be considered around each point (n_neighbors), and the minimum distance allowed for two points in the low dimensional space (min_distance). Together, these hyperparameters control whether the global or local structure of the high-dimensional space is prioritized and how the low-dimensional representation should look like. The hyperparameters selected for UMAP here were n_comp = 3, n_neighbors = 25, and min_dist = 0.0, for all cases. The DR and clustering algorithms, as well as their corresponding hyperparameters were selected to minimize the assumptions to be made about the data and its natural distribution. With

this, similarities and differences in cell response for different cohorts and spiked P exposures can be detected, leveraging single-cell response analysis data.

Figures and Tables

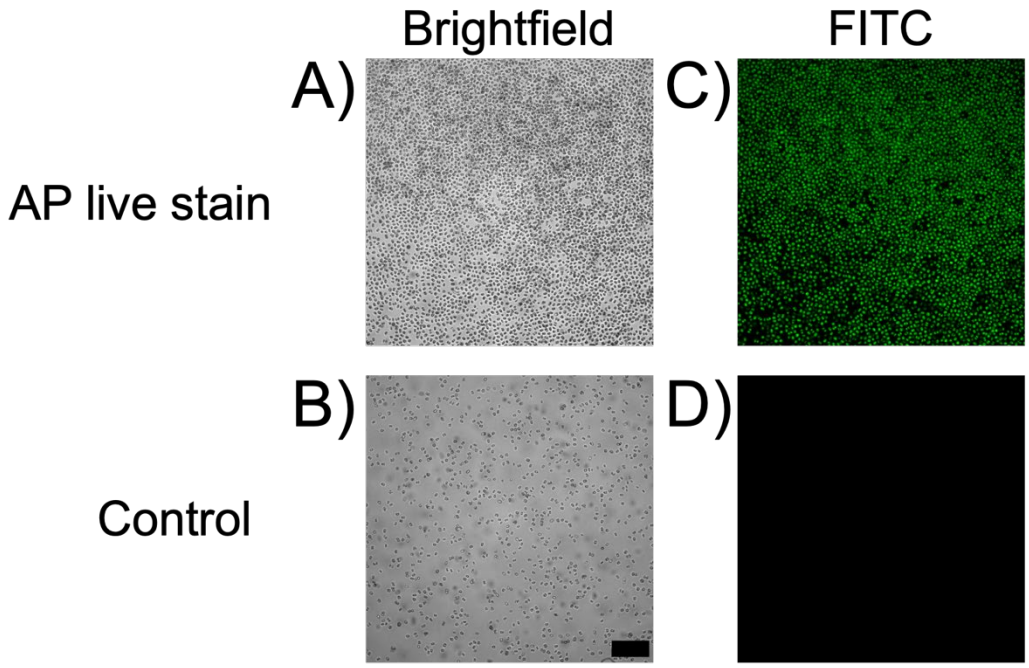

Figure S1 – Fluorescence microscopy images of APA in *C. reinhardtii* cells using the AP live stain outside of the microfluidic device. Algal cells were cultured, starved, and spiked in NTAP with 1 mM P as described in the methods. 1 mL of the spiked culture was centrifuged, re-suspended in spiked NTAP supplemented with AP live stain, and imaged in a 6-well plate. Representative images of cells treated with either AP live stain (top row) or vehicle control (bottom row) were obtained using both brightfield (A-B) and the FITC filter set (C-D). Scale bar is 100 μm.

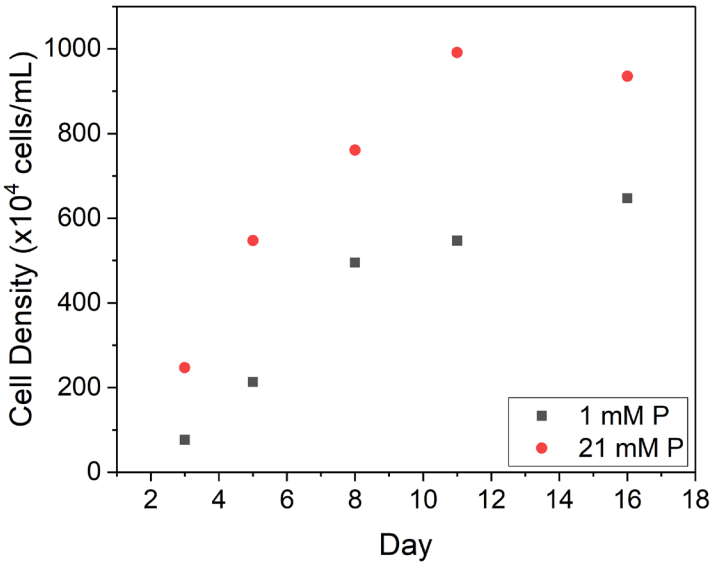

Figure S2 – Observed changes in algal growth rate under different basal P concentrations. *C. reinhardtii* were seeded in 250 mL flasks as described in the methods in either 1- or 21-mM P. ~50 µL from each flask was collected on days 3, 5, 8, 11, and 16 of culture and counted using a hemocytometer. The data is plotted as cell density vs. day number (from the beginning of the culture period).

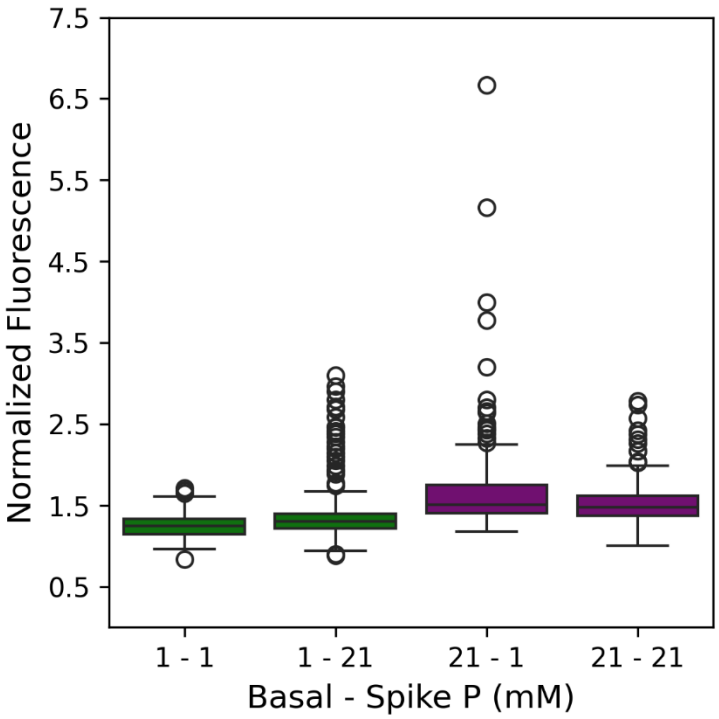

Figure S3 – Side by side comparison of the single cell APA response from four cohorts of cells with spiked concentrations equivalent to those of their basal levels of 1- and 21 mM P. APA was quantified in cells in four separate cohorts following the week-long culture in NTAP supplemented with either 1- or 21-mM P. Cells from the 1 mM culture (green) were treated with either basal (1 mM) or spiked (21 mM) P concentrations while cells from the 21 mM culture (purple) were treated with either spiked (1 mM) or basal (21 mM) P concentrations. APA was quantified for each cohort (minimum 183 cells per cohort).

Table S1. Statistical metrics for all seven cohorts under 1 mM basal P levels. Metrics including number of cells analyzed from each cohort, mean normalized fluorescence (corresponding to APA), and standard deviation (cohort heterogeneity), corresponding to the data demonstrated in Figures 2 and 3. The values for mean and standard deviations are gradient colored from lowest (dimmest) to highest (strongest), to facilitate comparison of the numerical values.

Basal [P] (mM)	Spike [P] (mM)	Cell Count	Mean	Standard Deviation
1	0.1	200	1.464	0.209
	0.5	302	1.531	0.274
	1	255	1.245	0.144
	11	286	1.372	0.292
	21	183	1.415	0.404
	31	245	1.376	0.277
	41	346	1.646	0.390

Table S2. Statistical analysis of single-cell APA from seven cohorts of different spike P levels acclimated to a 1 mM basal P level. One-way ANOVA and Fisher's test of means, both with a p-value of 0.05, were performed on all cohorts. *MeanDiff*: difference between mean values, *Prob*: probability, *Sig*: significance. All two-cohort combinations (first two columns) are compared with each other. Rows that are boxed in red indicate no statistically significant difference between the respective cohorts (significance = 0). Minimum of 183 cells per each cohort with an average number of 260 cells quantified for each cohort.

Spiked P (mM)		*MeanDiff*	*t Value*	*Prob*	*Sig*
0.5	0.1	0.067	2.445	0.015	1
1	0.1	-0.219	-7.744	0.000	1
1	0.5	-0.285	-11.221	0.000	1
11	0.1	-0.092	-3.349	0.001	1
11	0.5	-0.159	-6.443	0.000	1
11	1	0.126	4.908	0.000	1
21	0.1	-0.049	-1.614	0.107	0
21	0.5	-0.116	-4.142	0.000	1
21	1	0.169	5.845	0.000	1
21	11	0.043	1.516	0.130	0
31	0.1	-0.088	-3.074	0.002	1
31	0.5	-0.154	-5.999	0.000	1
31	1	0.131	4.902	0.000	1
31	11	0.005	0.181	0.856	0
31	21	-0.038	-1.308	0.191	0
41	0.1	0.182	6.854	0.000	1
41	0.5	0.115	4.901	0.000	1
41	1	0.401	16.239	0.000	1
41	11	0.274	11.481	0.000	1
41	21	0.231	8.468	0.000	1
41	31	0.270	10.800	0.000	1

Table S3. Statistical metrics for all seven cohorts under 21 mM basal P levels. Metrics including number of cells analyzed from each cohort, mean normalized fluorescence (corresponding to APA), and standard deviation (cohort heterogeneity), corresponding to the data demonstrated in Figure 3. The values for mean and standard deviations are gradient colored from lowest (dimmest) to highest (strongest), to facilitate comparison of the numerical values.

Basal [P] (mM)	Spike [P] (mM)	Cell Count	Mean	Standard Deviation
21	0.1	308	1.524	0.375
	0.5	325	1.616	0.406
	1	273	1.659	0.519
	11	452	1.417	0.306
	21	257	1.502	0.288
	31	272	1.504	0.237
	41	135	1.265	0.228

Table S4. Statistical analysis of single-cell APA from seven cohorts of different spiked levels acclimated to a 21 mM basal P level. One-way ANOVA and Fisher's test of means, both with a p-value of 0.05, were performed on all cohorts. *MeanDiff*: difference between mean values, *Prob*: probability, *Sig*: significance. All two-cohort combinations (first two columns) are compared with each other. Rows that are boxed in red indicate no statistically significant difference between the respective cohorts (significance = 0). Minimum of 183 cells per each cohort with an average number of 260 cells quantified for each cohort.

Spiked P (mM)		*MeanDiff*	*t Value*	*Prob*	*Sig*
0.5	0.1	0.092	3.254	0.001	1
1	0.1	0.134	4.540	0.000	1
1	0.5	0.042	1.445	0.149	0
11	0.1	-0.107	-4.059	0.000	1
11	0.5	-0.199	-7.682	0.000	1
11	1	-0.241	-8.836	0.000	1
21	0.1	-0.022	-0.730	0.466	0
21	0.5	-0.114	-3.839	0.000	1
21	1	-0.156	-5.051	0.000	1
21	11	0.085	3.050	0.002	1
31	0.1	-0.020	-0.690	0.490	0
31	0.5	-0.113	-3.847	0.000	1
31	1	-0.155	-5.075	0.000	1
31	11	0.086	3.160	0.002	1
31	21	0.002	0.049	0.961	0
41	0.1	-0.259	-7.049	0.000	1
41	0.5	-0.351	-9.633	0.000	1
41	1	-0.393	-10.502	0.000	1
41	11	-0.152	-4.360	0.000	1
41	21	-0.237	-6.265	0.000	1
41	31	-0.239	-6.366	0.000	1

Table S5. Fisher's test of means for cohorts of data presented in Figure S3. Analysis was performed using a p-value of 0.05. *MeanDiff*: difference between mean values, *Prob*: probability, *Sig*: significance. Results indicate a statistically significant difference (Sig=1) across all pairwise combinations of cohorts.

Cohort		*MeanDiff*	*t Value*	*Prob*	*Sig*
1-21	1-1	0.169	4.765	<0.0001	1
21-1	1-1	0.413	12.940	<0.0001	1
21-1	1-21	0.244	6.963	<0.0001	1
21-21	1-1	0.257	7.925	<0.0001	1
21-21	1-21	0.088	2.470	0.014	1
21-21	21-1	-0.156	-4.906	<0.0001	1

Table S6. Two-way analysis of variance (ANOVA) on the combined data sets for all seven cohorts under different basal P levels. The analysis was performed to investigate the statistical significance of both factors studied: basal and spiked P levels. Results indicate that the effect of each factor individually (one-way ANOVA) and the interaction between the two factors (two-way) were both significant.

Factor	*DF*	*Sum of Squares*	*Mean Square*	*F Value*	*P Value*
Spiked P	6	11.705	1.951	17.887	0
Basal P	1	3.476	3.476	31.871	1.77E-08
Interaction	6	35.813	5.969	54.726	0
Model	13	54.781	4.214	38.636	0
Error	3825	417.181	0.109	–	–
Corrected Total	3838	471.962	–	–	–

References

(1) McInnes, L.; Healy, J.; Astels, S. Hdbscan: Hierarchical Density Based Clustering. *Journal of Open Source Software* **2017**, *2* (11), 205. https://doi.org/10.21105/joss.00205.

(2) Campello, R. J. G. B.; Moulavi, D.; Sander, J. Density-Based Clustering Based on Hierarchical Density Estimates. In *Advances in Knowledge Discovery and Data Mining*; Pei, J., Tseng, V. S., Cao, L., Motoda, H., Xu, G., Eds.; Lecture Notes in Computer Science; Springer: Berlin, Heidelberg, 2013; pp 160–172. https://doi.org/10.1007/978-3-642-37456-2_14.

(3) Jolliffe, I. Principal Component Analysis. In *Wiley StatsRef: Statistics Reference Online*; John Wiley & Sons, Ltd, 2014. https://doi.org/10.1002/9781118445112.stat06472.

(4) McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv September 17, 2020. https://doi.org/10.48550/arXiv.1802.03426.