Identifying Emerging Phenomenon in Plant Long Temporal Phenotyping Experiments

Jiajie Peng, Junya Lu, Donghee Hoh, Ayesha S Dina,
Xuequn Shang, David M Kramer, Jin Chen

***Abstract*—The rapid improvement of phenotyping capability, accuracy, and throughput have greatly increased the volume and diversity of phenomics data. A remaining challenge is an efficient way to identify phenotypic patterns to improve our understanding of the quantitative variation of complex phenotypes, and to attribute gene functions. To address this challenge, we developed a new algorithm to identify emerging phenomena from large-scale temporal plant phenotyping experiments. An emerging phenomenon is defined as a group of genotypes who exhibit a coherent phenotype pattern during a relatively short time. Emerging phenomena are highly transient and diverse, and are dependent in complex ways on both environmental conditions and development. Identifying emerging phenomena may help biologists to examine potential relationships among phenotypes and genotypes in a genetically diverse population and to associate such relationships with the change of environments or development. We present an emerging phenomenon identification tool called Temporal Emerging Phenomenon Finder (TEP-Finder). Using large-scale longitudinal phenomics data as input, TEP-Finder first encodes the complicated phenotypic patterns into a dynamic phenotype network. Then, emerging phenomena in different temporal scales are identified from dynamic phenotype network using a maximal clique based approach. Meanwhile, a directed acyclic network of emerging phenomena is composed to model the relationships among the emerging phenomena. The experiment that compares TEP-Finder with two state-of-art algorithms shows that the emerging phenomena identified by TEP-Finder are more functionally specific, robust, and biologically significant. The source code, manual, and sample data of TEP-Finder are all available at: http://phenomics.uky.edu/TEP-Finder/.**

I. Introduction

Biomedical studies have been ushered into a new era by the rapid development of large-scale genotyping and phenotyping technologies [2], [8], [11], [13], [17]. Recent studies demonstrate that by integrating both phenomics and genomics, we can better understand organism behaviors and identify new genes that govern phenotypes and response to the varying environments [7], [9], [29]. More specifically, by analyzing large-scale plant photosynthetic phenotype data, researchers can identify complex aggregate phenotypic traits, and explore the processes or genetic components that control a trait and the essential conditions under which the trait emerge [15], [31].

The main computational challenge in omics data analysis arises from its unsupervised nature. It is generally believed that the **emerging phenomena** among multiple phenotypes measured across several genotypes (e.g., gene knockouts) reveals, to a great extent, the common regulatory roles of the knocked out genes in the biological system. An emerging phenomenon refers to a phenotypic pattern that multiple genotypes have correlated phenotype values during a serial of continuous time points [8]. A sample emerging phenomenon in plant photosynthesis phenotype data is shown in Figure 1. The experiment was done under the fluctuating light conditions (between 0 and $1000\mu molm^{-2}s^{-1}$). Five selected genotypes ($P_1...P_5$) were measured using three photosynthetic phenotypes, namely photosynthetic system II activity (Φ_{II}), photoprotection (q_{ESV}), and photoinhibition (q_I). The relative phenotype values were calculated by comparing each genotype with the reference (col-0) using logged fold change. The shadowed area indicates an emerging phenomenon of the five plants between 12:30 and 14:30, during which, all the five genotypes have similar phenotype values.

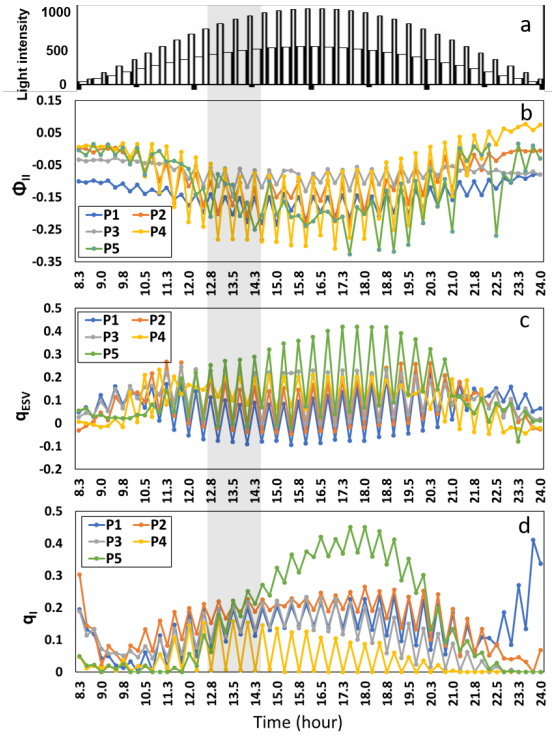

Figure 1. A sample emerging phenomenon (shadowed area) identified in a plant photosynthesis phenotyping experiment under fluctuating light conditions (a). In the experiment, five genotypes (chloroplast-targeted single mutant lines of *Arabidopsis thaliana*) were measured using three photosynthetic phenotypes (Φ_{II}, q_{ESV}, and qI). In the shadowed area (b,c,d), all the genotypes have similar phenotype values.

Emerging phenomena is universal in phenotyping experiments *esp.* under dynamic environmental conditions. They are highly transient and diverse, dependent in complex ways on both environmental conditions and development [8]. Re-

vealing emerging phenomena is vital towards the identification of meaningful differences in biological function among genotypes, which may help biologists to examine potential phenome-genome relationships in a genetically diverse population and to associate such relationships with the change of environments or development. It is, however, *unclear* what specific patterns biomedical researchers should look for given the complexity of the biological system and its responses to environmental perturbations. Besides, the large variance in phenotypes, due to the biodiversity and the variance in environmental distribution, adds more challenges to the already difficult task [8], [11], [33].

To address this challenge, we propose a new tool called **Temporal Emerging Phenomenon Finder** (TEP-Finder) as the first approach to capture emerging phenomena with various temporal scales and arbitrary phenotype variation shapes (see Figure 2). TEP-Finder automatically transforms large-scale phenomics data into emerging phenomenon patterns, thus facilitates the translation of information into knowledge. TEP-Finder has two phases. First, TEP-Finder encodes phenotype-based relationships into a dynamic network using nonparametric clustering and generates seeds. It then identifies all the emerging phenomena in different temporal scales and constructs a directed acyclic network of emerging phenomena. To demonstrate the effectiveness of TEP-Finder, we applied TEP-Finder on a large-scale plant photosynthesis phenotyping experiment, and the results show that TEP-Finder can reliably and accurately identify high quality emerging phenomena from data. Comparing with the existing models, TEP-Finder has the following advantages:

- TEP-Finder is the first approach to capture emerging phenomena with diverse scales systematically;
- TEP-Finder constructs a network of emerging phenomena to provides a graph-based representation of the complex hierarchy of emerging phenomena;
- TEP-Finder successfully discovers emerging phenomena in an Arabidopsis photosynthetic phenotyping experimental data with high biological significance.

II. Background

An emerging phenomenon is defined as a group of genotypes who has a pattern of correlated phenotypes in a serial of continuous time points [8]. In the literature, given a set of predefined patterns, the minimal genotype contributor set can be identified using existing data mining techniques such as association rule mining [16], [32] or subspace trajectory clustering [1], [26]. However, given the unsupervised nature, most emerging phenomena are not pre-definable. To our knowledge, there is no existing algorithm exactly designed for emerging phenomena identification. Tools, such as DHAC and NPM [12], [19], may be slightly modified to achieve the goal. Here we discuss two existing approaches with additional steps adopted for emerging phenomenon discovery.

DHAC models how a network change with time [19]. Assuming that the edges in a network are conditionally independent given group membership, DHAC uses a probabilistic model to translate a hierarchical stochastic block to

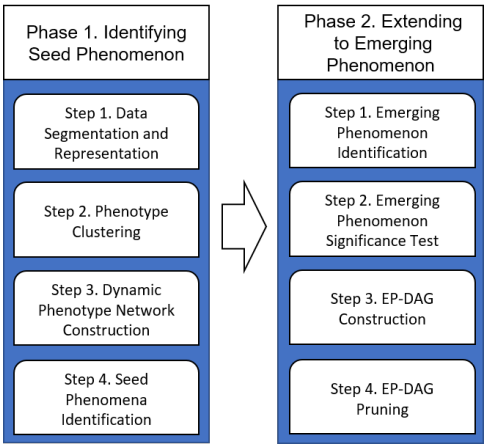

Figure 2. The workflow of TEP-Finder. Given the temporal phenomics data, it identifies significant seed phenomena in every time frame; by expanding each seed to longer time frames, it discovers emerging phenomena that appear and disappear subject to the change of environments or development; the relationships among all the emerging phenomena are modeled by a directed acyclic network called EP-DAG.

the dynamic domain, thus clustering a time-evolving network based on the observations at several specific time points. The rationale is that any node in a network cluster at a specific time point should be influenced by clusters at nearby time points. DHAC can be employed to group genotypes by matching clusters across multiple time points with additional steps that transform longitudinal phenomics data into a dynamic network (called DHAC+). However, to facilitate dynamic network clustering, DHAC considers global features on all temporal points rather than local features. Subsequently, the DHAC-based method cannot identify emerging phenomenon at different temporal scales.

NPM is a non-parametric clustering method that can simultaneously cluster subjects with arbitrary cluster shapes [12]. NPM represents the phenotypes of each genotype in a serial of continuous time points as a cloud of points. Each point of the cloud corresponds to a vector in the sequential phenotype measurements taken for the genotype. Two similar shapes of clouds represent that two genotypes have a coherent phenotype pattern in a given time frame. Note that NPM is more advantageous than the Pearson correlation on the identification of a set of genotypes with coherent phenomics data. It is because Pearson correlation requires all the variables to follow a normal distribution, which is not always held for the phenomics data, while NPM does not make any assumption about the underlying data distribution and thus is particularly suitable for phenomics data analysis. NPM can be employed to identify emerging phenomena by applying it repeatedly on every time frame of a longitudinal phenomics dataset (called NPM+). However, it is difficult to pre-define the time scale of emerging phenomena or to identify the relationships between overlapped emerging phenomena. Furthermore, NPM is not a deterministic method so that the results are dependent on the initialization and the selection of anchor points.

The unmet needs to effectively identify high-quality emerging phenomena necessitates the development of tools that can automatically transform large-scale phenomics data into

Individuals		T_1	T_2	T_3
A	B	0.9	0.9	0.2
A	C	0.9	0.9	0.9
A	D	0.9	0.2	0.2
A	E	0.2	0.9	0.2
A	F	0.2	0.2	0.2
A	G	0.2	0.2	0.9
B	C	1	1	0.2
B	D	0.9	0.2	0.2
B	E	0.2	0.9	0.2
B	F	0.2	0.2	0.2
B	G	0.2	0.2	0.2
C	D	0.9	0.2	0.2
C	E	0.2	0.9	0.9
C	F	0.2	0.9	0.9
C	G	0.2	0.2	0.9
D	E	0.9	0.2	0.2
D	F	0.9	0.2	0.2
D	G	0.9	0.2	0.2
E	F	0.9	0.9	0.9
E	G	0.2	0.2	0.9
F	G	0.9	0.9	0.9

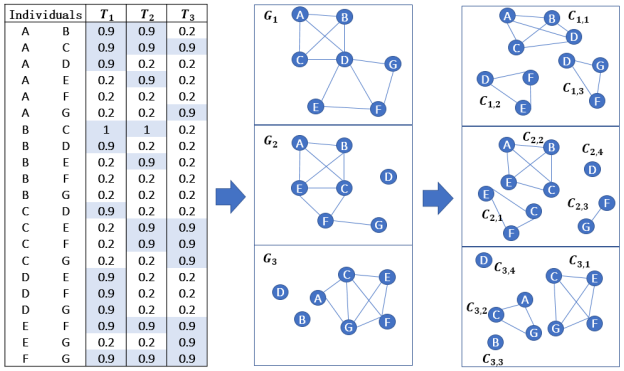

Figure 3. Illustrative example of the dynamic phenotype network construction. Values in the table represent the co-occurrence frequencies of any two genotypes being in the same cluster in three different time frames $\{T_1, T_2, T_3\}$. We add edge $\langle P_j, P_h \rangle$ to a slice of the dynamic phenotype network if the corresponding co-occurrence frequency of genotypes P_j and P_h is greater than a threshold (shaded blocks). The middle panel shows the dynamic network, in each slice of which, the maximal cliques are displayed in the right panel.

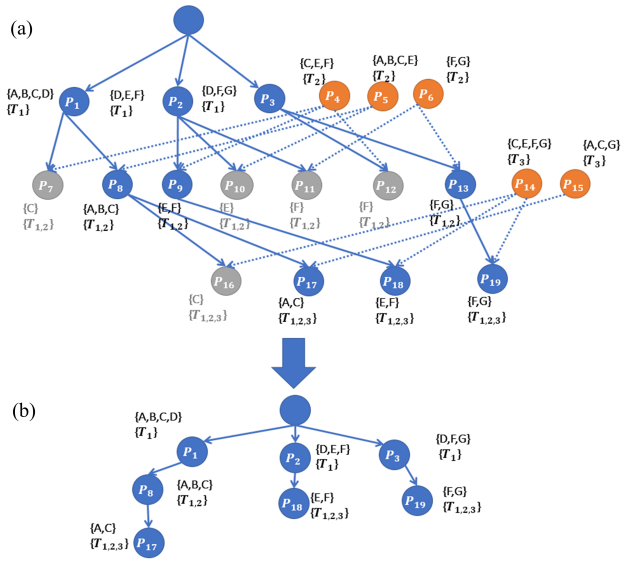

Figure 4. Illustrative example on extending a seed phenomenon to a longer time frame starting from the same time point. (a) The maximal cliques of G_1 are at the first level. Then they are joined with the maximal cliques of G_2 and G_3 to generate longer emerging phenomena. (b) The result is pruned using the procedure introduced in Section IV-B1.

emerging phenomenon patterns, thus facilitate the translation of information into knowledge. To precisely identify emerging phenomena with different temporal scales, we propose TEP-Finder. In our experiment, TEP-Finder has been compared with NPM+ and DHAC+. The results demonstrate that TEP-Finder is better for capturing emerging phenomena and relationships among them.

III. Definition of Emerging Phenomenon

In a long temporal phenotype dataset $\mathcal{M}(\mathcal{P}, \mathcal{T})$, T_i is a time frame associated with experimental environments, treatments, and outcomes ($T_i \in \mathcal{T}$), and P_j is an genotype to study ($P_j \in \mathcal{P}$), e.g., a gene knockout or a inbred line. The phenotype values of genotype P_j in time frame T_i are represented by a set of data points $D_{i,j} = \{x_1^{i,j}, \ldots, x_m^{i,j}\}$. In the plant photosynthetic phenotyping experiment using DEPI [8], the phenotypes are mainly photosynthetic system II activity (Φ_{II}), photoprotection (q_{ESV}), and photoinhibition (q_I). An emerging phenomenon e_i is defined as follows (see example in Figure 1).

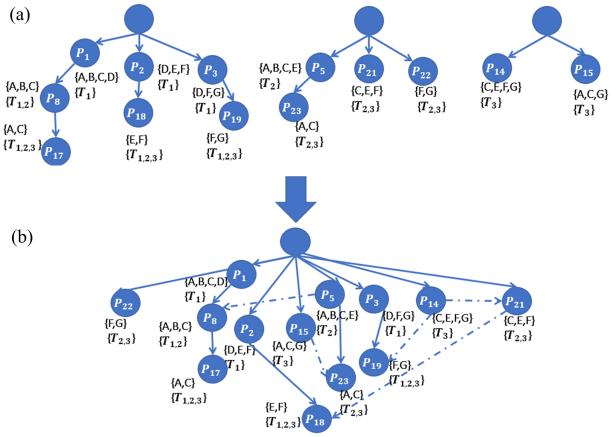

Figure 5. Illustrative example on EP-DAG construction. (a) Three sub-DAGs are built based on different starting time points. (b) All of them are merged into one DAG using the procedure introduced in Section IV-B3.

Definition III.1. Emerging Phenomenon. Given the temporal phenotype data $\mathcal{M}(\mathcal{P}, \mathcal{T})$, an emerging phenomenon $C(P_\lambda, T_\lambda)$ is a group of genotypes P_λ ($P_\lambda \subseteq \mathcal{P}$) that exhibit coherent phenomena during continuous temporal range T_λ ($T_\lambda \subseteq \mathcal{T}$), where $|P_\lambda| \geq K_1$, $|T_\lambda| \geq K_2$, and the percentage of significant phenotype values in e is greater than or equal to K_3. K_1, K_2, and K_3 are user specified thresholds.

Note that in an emerging phenomenon $C(P_\lambda, T_\lambda)$, certain percentage of phenotype values of should be significantly different from the reference. The definition does not require all the phenotype values of $C(P_\lambda, T_\lambda)$ to be significant because, when the environmental conditions vary dynamically, phenotype values often periodically switch between significance and insignificance (see Figure 1). Hence, it is more reasonable to require a certain portion but not all of the phenotype values to be significantly different from that of the reference. Since *A Priori* does not apply, new algorithms are needed for the identification of emerging phenomenon.

For a large-scale phenotyping experiment, the total number of identified emerging phenomena could be large. To better manage and use them, we construct an EP-DAG $\mathcal{G}$ defined as:

Definition III.2. Emerging Phenomenon DAG. An emerging phenomenon DAG (EP-DAG) $\mathcal{G}$ is a directed acyclic network (DAG), where each node in $\mathcal{G}$ represents represents an emerging phenomenon $C(P_\lambda, T_\lambda)$, and node $C(P_i, T_j)$ is a descendent of node $C(P_h, T_k)$ if and only if $P_i \supset P_h$ and $T_j \subset T_k$.

The outputted EP-DAG is available in the OBO format. It, once generated from data, can be visualized with Cytoscape [25] or OntoVisT [28]. It automatically supports emerging phenomenon search, phenotype relationship identi-

fication, and multiple phenotyping experiments comparison, leading to improved computational efficiency and succinct representation. To our knowledge, there is no existing work focused on the construction of EP-DAGs.

IV. Methods

To systematically identify emerging phenomena in long-term phenotyping experiments and to examine the interactions between emerging phenomena and dynamic environments in a genetically diverse population, we introduce a new algorithm called **TEP-Finder**. TEP-Finder has two phases. First, it identifies seed phenomena in every time frame of a longitudinal phenomics dataset, where a time frame is a predefined minimal temporal range of any emerging phenomena. Second, by expanding each identified seed phenomena to longer time frames, TEP-Finder discovers emerging phenomena that appear and disappear subject to the change of environments or time. Multiple emerging phenomena are then merged, pruned, and connected to form a phenomenon network to facilitate phenotype search, comparison, and functional analysis. The diagram of the whole process is shown in Figure 2.

A. TEP-Finder Phase 1. Identifying Seed Phenomenon

An emerging phenomenon e is considered as the phenotypes of multiple genotypes that have similar variation trends in a continuous time period. Biologically, such time period can be transient or can last for a relatively long time. To identify e with varying length, we consider a seed-based approach. Namely, we segment the whole experiment duration into multiple time frames, each being the minimal temporal range of any emerging phenomena. Then, at every time frame, we seek seed phenomena that are potentially extendable to a longer time period. The seed identification phase can be divided into four steps.

1) Data Segmentation and Data Representation: Given the temporal phenotype data $\mathcal{M}(\mathcal{P}, \mathcal{T})$, we segment $\mathcal{T}$ into separated time frames with a fixed length m using the sliding window approach. Here, the window width is the smallest temporal period of any emerging phenomenon (e.g., 30 minutes) that users can specify.

We adopt a meta-clustering approach to identify the relationships among all the tested genotypes in each time frame T_i [5], [18]. In the meta-clustering process, we repeatedly cluster the phenotype values of all the genotypes $\mathcal{P}$ in T_i using non-parametric clustering with random anchor points [12]. The center of non-parametric clustering is a cloud-of-points representation. Since all the phenotype values are collected in a relatively short time, we examine the dependence among different phenotypes while ignoring the sequential order among the values and simply characterizing the phenotypes of genotype P_j in time frame T_i by the set of data points $\mathcal{D}_{i,j} = \{x_1^{i,j}, \ldots, x_m^{i,j}\}$, which we refer to as *cloud-of-points* representation.

2) Phenotype Clustering: Following the standard framework of mixture models, we assume that there are K different underlying distributions in time frame T_i, where each distribution is introduced to capture a different "shape" of the cloud-of-points representation, and all the phenotype values observed in the cloud-of-points representation are drawn independently from one of the K distributions [10]. More specifically, let $f_1(\cdot), \ldots, f_K(\cdot)$ be the density functions for the K underlying distributions, and let $p_1, \ldots, p_K$ be the prior probabilities for choosing each distribution. Then, for genotype P_j in time frame T_i, the likelihood of observing the cloud-of-points representation $\mathcal{D}_{i,j}$ is then given by

$$\Pr(\mathcal{D}_{i,j}) = \sum_{k=1}^{K} p_k \Pr(\mathcal{D}_{i,j}|f_{i,j}) = \sum_{k=1}^{K} \left(p_k \prod_{h=1}^{m} f_{i,j}(\mathbf{x}_h^{i,j}) \right) \quad (1)$$

Following the framework of maximum likelihood estimation, we find the optimal density functions $\{f_j(\cdot)\}_{j=1}^{K}$ by solving the optimization problem

$$\max_{f_1,\ldots,f_K,\mathcal{P}} \sum_{j=1}^{n} \log \Pr(\mathcal{D}_{ij}) \quad (2)$$

where $\Pr(\mathcal{D}_{ij})$ is given in Equation 1, and n is the total number of genotypes in $\mathcal{M}$.

This optimization problem can be effectively solved by employing NPM, a non-parametric clustering method for phenomics data analysis [12]. Based on the Nadaraya-Watson method for kernel density estimation [20], [24], [27] and following the framework of maximum likelihood estimation, NPM uses optimal density functions and applies a non-parametric clustering technique to group genotypes into the same cluster if their clouds-of-points share similar arbitrary shapes. The non-parametric approach avoids the parametric assumption of the underlying distribution so that NPM is suitable to model the nonlinear interactions among multiple phenotypes [12]. Since the clustering process is dependent on the initialization and the selection of anchor points, we repeat NPM multiple times to obtain all the meta-clustering results.

3) Dynamic Phenotype Network Construction: In this step, we construct a dynamic phenotype network $G(\mathcal{P}, \mathcal{E}, \mathcal{T})$, where $\mathcal{P}$ is the set of genotypes, $\mathcal{T}$ is the set of time frames, and $\mathcal{E} = \{E_1, E_2, ..., \mathcal{E}_k\}$ represents edges in different time frames. In each time frame T_i, we check whether any two genotypes P_j and P_h are frequently grouped into the same cluster in meta-clustering. If the co-occurrence is greater than a predefined threshold K_4, we add edge $\langle P_j, P_h \rangle$ to E_i. In the dynamic network G, while the nodes are identical, the edges vary over time, indicating emerging phenomena emerge and disappear with the change of time.

A running example is shown in Figure 3. In the example, $\mathcal{T} = \{T_1, T_2, T_3\}$ and $\mathcal{P} = \{A, B, C, D, E, F, G\}$. Given the frequency of concurrence of every two genotypes in T_1, T_2 and T_3 (the table on the left), and let K_4 be 0.8, we identify all the edges (shaded blocks) and construct the dynamic network in the middle panel of Figure 3.

4) Seed Phenomena Identification: We identify the seed phenomena by repeatedly applying a maximal clique based approach on every time frame of the dynamic network $G(\mathcal{P}, \mathcal{E}, \mathcal{T})$. Clique is a special structure such that any two nodes in it are adjacent, implying a close relationship among all the nodes that belong to the same clique. A maximal clique is a clique that cannot be extended by including one more adjacent node, meaning it is not a subset of a larger clique.

More specifically, we adopt the Bron-Kerbosch algorithm to identify all the maximal cliques [4]. The basic form of the Bron-Kerbosch algorithm is the recursive backtracking that searches for all maximal cliques in a given network. Its performance has been further improved by defining a pivot vertex set, allowing it to backtrack more quickly in branches of the search that contain no maximal cliques [6], [30].

Let $\mathbb{C}$ be the set of all the maximal cliques in the dynamic network $G(\mathcal{P}, \mathcal{E}, \mathcal{T})$, maximal clique $C(P_j, T_i) \in \mathbb{C}$ defines a seed phenomenon with its genotype set being P_j and its time frame being T_i. A running example is shown in the right panel of Figure 3. The maximal cliques in T_1 are $C_{1,1} = \{A, B, C, D\}$, $C_{1,2} = \{D, E, F\}$, $C_{1,3} = \{D, F, G\}$.

B. TEP-Finder Phase 2. Extending from Seeds to Emerging Phenomenon

After identifying all the seed phenomena in the minimal time frames, we extend them to longer time frames. This phase has four steps.

1) Emerging Phenomenon Identification: To identify emerging phenomena, the general idea is to combine seed phenomena in adjacent time frames. More specifically, for $C(P_j, T_i)$, which is the jth seed phenomenon in time frame T_i, we join it with every seed in time frame T_{i+1} $C(P_k, T_{i+1})$, resulting in $C(P_{j,k}, T_{i,i+1})$, where $P_{j,k}$ represents the intersection of P_j and P_k. Then, we determine whether $C(P_{j,k}, T_{i,i+1})$ is a new emerging phenomenon using the following rules developed based on the definition of the emerging phenomenon (see Definition III.2). The combination process will continue on the followed time frames (i.e. $T_{i+2}, \ldots, T_n$), until all the seed phenomena are examined.

- Discard $C(P_{j,k}, T_{i,i+1})$ if $|P_{j,k}| < K_1$, is not an emerging phenomenon;
- Replace $C(P_j, T_i)$ with $C(P_{j,k}, T_{i,i+1})$ if $P_{j,k} = P_j \geq K_1$.
- Replace $C(P_k, T_{i+1})$ with $C(P_{j,k}, T_{i,i+1})$ if $P_{j,k} = P_k \geq K_1$.
- Accept $C(P_{j,k}, T_{i,i+1})$ as a new emerging phenomenon if $S_{j,k} \neq S_j$ and $S_{j,k} \neq S_k$ and $|S_{j,k}| \geq K_1$

Following the example of the dynamic phenotype network and maximal cliques in Figure 3, the emerging phenomenon identification procedure starting from T_1 is shown in Figure 4. In Figure 4(a), one of the seed phenomena that start from T_1 or T_2 is $P_2 = C(\{D, E, F\}, \{T_1\})$ and $P_4 = C(\{C, E, F\}, \{T_2\})$ respectively. The join of P_2 and P_4 is $P_9 = C(\{E, F\}, \{T_{1,2}\})$, which, according to Definition III.1, is saved as an emerging phenomenon in time frame $T_{1,2}$. Similarly, for the other seeds in T_1 and T_2, we join them pairwisely and save all the qualified emerging phenomena (see the blue colored notes in Figure 4(a)). Next, we join all the emerging phenomena in time frame $T_{1,2}$ with the seeds in T_3, resulting in the emerging phenomena in time frame $T_{1,2,3}$. For example, $P_{18} = C(\{E, F\}, \{T_{1,2,3}\})$ is the result by joining $P_9 = C(\{E, F\}, \{T_{1,2}\})$ and $P_{14} = C(\{C, E, F, G\}, \{T_3\})$. Note that P_{18} replaces P_9 since they have the same genotypes and the time frame of P_{18} contains that of P_9. Those who do not qualify the definition of emerging phenomenon are discarded (all the gray notes in Figure 4(a)).

2) Significance Test: Given the temporal phenotype data $\mathcal{M}(\mathcal{P}, \mathcal{T})$, we compare the phenotype values of every genotype P_i with the reference using logged fold change, resulting in the relative phenotype values. The reference could be the wild-type in mutant experiments, the parental lines in recombinant inbred line experiments, or the average of all the genotypes in population experiments. Without losing generality, all the significant phenomena can be identified using a user given logged fold change threshold or with the computation of the false discovery rate. Other significance tests can also be applied for the same purpose. If the percentage of significant phenotype values of an emerging phenomenon is less than a user given threshold K_3, the emerging phenomenon is discarded.

3) EP-DAG Construction: To model the complex relationships among all the emerging phenomena, we construct an EP-DAG $\mathcal{G}$. $\mathcal{G}$ is a DAG with a virtual root node P_{root}. We first connect all the emerging phenomena found in any individual time frame directly to P_{root} (see example in Figure 4(b)). Next, we add an edge pointing from every emerging phenomenon to another one if the latter is generated by joining the former with other ones and both of them start from the same time frame. For example, in Figure 4(b), an edge is pointing from $P_8 = C(\{A, B, C\}, \{T_{1,2}\})$ to $P_{17} = C(\{A, C\}, \{T_{1,2,3}\})$. Finally, we add an edge pointing from one emerging phenomenon $C(P_j, T_i)$ to another one $C(P_h, T_k)$, if $P_j \subset P_h$, $T_i \supset T_k$, and $C(P_h, T_k)$ is not a descendent of $C(P_j, T_i)$. For example, we add edges pointing from P_5 to P_8, P_{15} to P_{23}, and P_{14} to P_{21} (see the dotted edges in Figure 5(b)).

4) EP-DAG Pruning: Finally, to reduce the redundancy of the emerging phenomenon, we merge the highly overlapped emerging phenomena and remove emerging phenomena with insignificant phenotype values. Note that if an emerging phenomenon is discarded because it does not satisfy the user given thresholds (e.g., the percentage of significant values less than K_3), its children will be redirect to its patent emerging phenomena. See examples in Figure 4(a,b). Mathematically, given two emerging phenomena $C(P_j, T_i)$ and $C(P_h, T_i)$ in the same time frame, if $|P_j - P_h| \leq 1$ and $|P_h - P_j| \leq 1$, we remove the two emerging phenomena and compose a new one called $C(P_j \cup P_h, T_i)$. Meanwhile, the edges connecting to $C(P_j, T_i)$ and $C(P_h, T_i)$ are redirected to $C(P_j \cup P_h, T_i)$.

V. Results

A. Data Description

For performance evaluation, we used the long temporal plant photosynthesis phenomics data in Gao et al [12]. The phenotyping experiment was carried out by testing 182 chloroplast-targeted single mutant lines (each with at least four biological replicates) of *A. thaliana* under dynamic environmental conditions using DEPI [8]. Three kinds of phenotypes, i.e., photosynthetic system II activity (Φ_{II}), photoprotection (q_{ESV}), and photoinhibition (q_I) were collected at 112 time points. See experiment details in [8].

TEP-Finder was implemented with Python 2.7. The following parameters for TEP-Finder were used in the experiment:

number of genotypes $K_1 = 5$, number of time points $K_2 = 10$, percentage of significant phenomena $K_3 = 0.5$, number of time points per time frame 10, overlap rate between two adjacent time frames 90%; number of runs of clustering per time frame 100. The final results consist of 4,318 emerging phenomena and an EP-DAG with 7,789 edges.

B. Methods to Compare

We compared TEP-Finder with NPM+ and DHAC+. The latter two are the methods modified from NPM and DHAC respectively (see Section II). The major difference in these methods locates on the process of seed phenomenon identification. More specifically, NPM+ consists of the following two steps. First, given the phenotype data $\mathcal{M}(\mathcal{P}, \mathcal{T})$, we call NPM once at every time frame to obtain the clustering results. Second, the clusters are used as the inputs to TEP-Finder phase two (see Section IV-B). In DHAC+, we first preprocess the phenomics data $\mathcal{M}(\mathcal{P}, \mathcal{T})$ using the phase one of TEP-Finder, resulting in the dynamic phenotype network G. Second, DHAC is adopted to identify seed phenomena in G instead of searching for the maximal cliques. Finally, the seed phenomena are used as the inputs to TEP-Finder phase two (see Section IV-B). Note that the only difference between NPM+, DHAC+, and TEP-Finder is how the seed phenomena are identified. Comparing NPM+ and DHAC+ with TEP-Finder is critical because it can test whether our meta-clustering followed with maximal clique approach is appropriate to generate seeds, which form the basis for the identification of emerging phenomena.

C. Performance Evaluation using GO Enrichment

An emerging phenomenon consists of a list of chloroplast-targeted single mutant lines that exhibit coherent and significant phenomena in a continuous time frame. It is expected that the knockout genes would be involved in the same biological process or have a similar molecular function. Therefore, we tested whether the knockout genes in the same emerging phenomenon are also enriched in Gene Ontology (GO). GO includes three categories: biological process, molecular function, and cellular component. Given a set of genes and their GO annotations, GO enrichment analysis identifies the over-represented GO terms. In our experiment, data were downloaded from the GO website in March 2017, and clusterProfiler [34] was used for the enrichment test.

Figure 6(a) shows that the percentage of emerging phenomena at each level of the EP-DAG using GO biological process. Clearly, TEP-Finder is constantly better than DHAC+, *esp.* at deep levels of the EP-DAG. The high performance on deep levels is important because emerging phenomena at deep levels often represent abnormal photosynthetic behaviors in a relatively more extended time period. TEP-Finder and NPM+ have a similar trend, but TEP-Finder is still better than NPM+ on most of the cases. Specifically, the averaged percentage of the enriched emerging phenomena of TEP-Finder is 0.80, which is 0.70 for NMP+. Similar results are found on the GO enrichment test on the molecular function category. In general, the performance of TEP-Finder is higher than NPM+

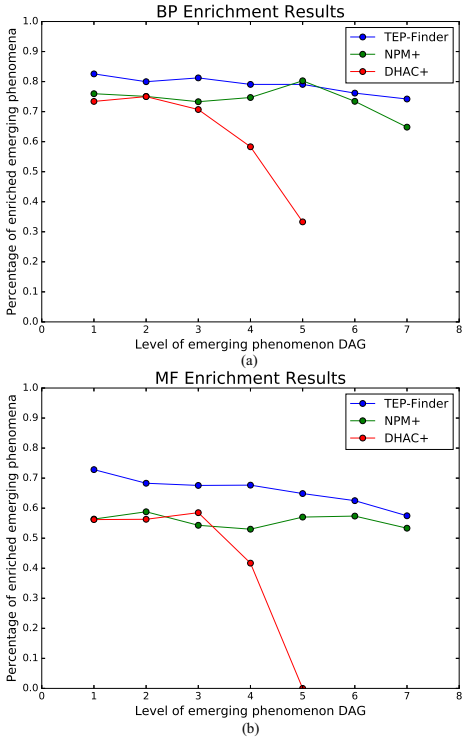

Figure 6. Evaluation of emerging phenomena using Gene Ontology enrichment on biological process (a) and molecular function (b). The x-axis represents the level of the EP-DAG. The y-axis represents the percentage of the emerging phenomena enriched in at least one GO term. Blue, green, and red represent the results of TEP-Finder, NPM+, and DHAC+.

and DHAC+ at each level of EP-DAG (Figure 6(b)). The averaged percentage of the enriched emerging phenomena of TEP-Finder is 0.66, while the values of NPM+ and DHAC+ are 0.56 and 0.43 respectively.

While the first experiment shows that TEP-Finder has more enriched emerging phenomena than the other two, it is not clear whether the enriched GO terms are at a shallow or deep level of the GO. Therefore, in the second experiment, we compared the distribution of the enriched GO terms among the three methods. Since the first level of the EP-DAG is the virtual root node, the comparison was carried out at the second, third and fourth level. We only tested the first three valid levels of EP-DAG because, in the results of DHAC+, the number of emerging phenomena after three levels are too few to compare. Figure 7 shows the cumulative distribution of the GO biological process terms and the molecular function terms at the first three valid EP-DAG levels. It is constant that there are more deep-level enriched GO terms in TEP-Finder than the other two.

D. Performance Evaluation using Gene Association

Given an EP-DAG, gene-to-gene similarity can be calculated based on the topological structure of the DAG. We thereby test the correlation between EP-DAG based gene similarities with the GO molecular function based gene similarities. To calculate the gene-to-gene similarities based on the EP-DAG, we adopted the widely used Resnik method [23]. Specifically, given any two genes g_i and g_j, we identify

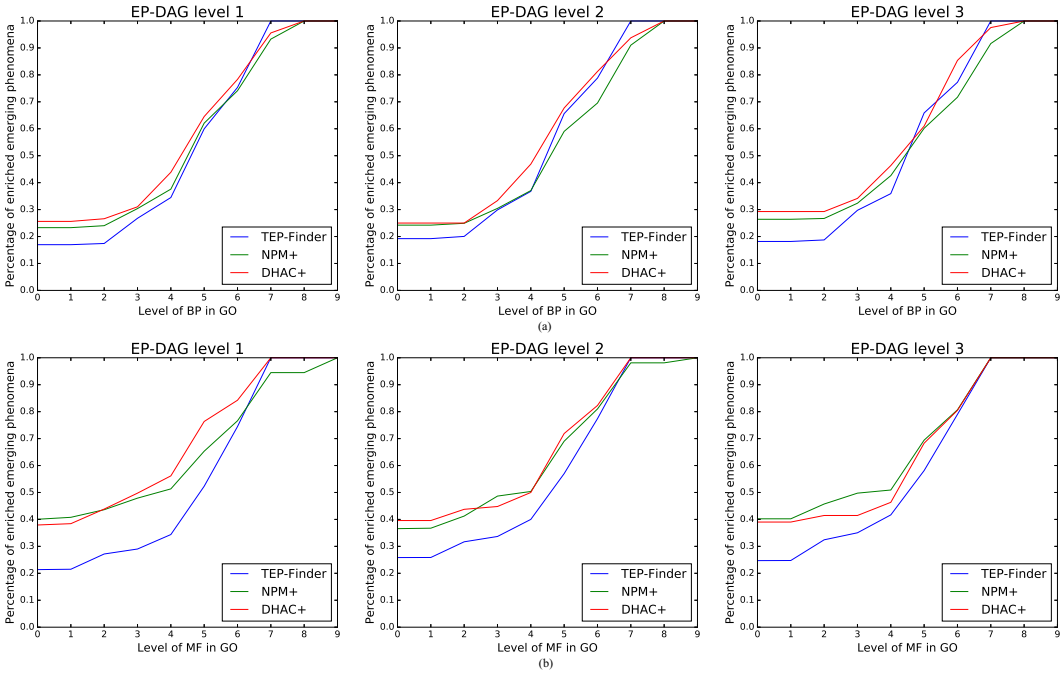

Figure 7. Cumulative distributions of identified emerging phenomena at different levels, which are enriched in Gene Ontology (GO) biological process category (a) and molecular function category (b). The x-axis represents the level of GO. The y-axis represents the percentage of emerging phenomena enriched at each GO level. The blue, green, and red line represent the result of TEP-Finder, NPM+, and DHAC+.

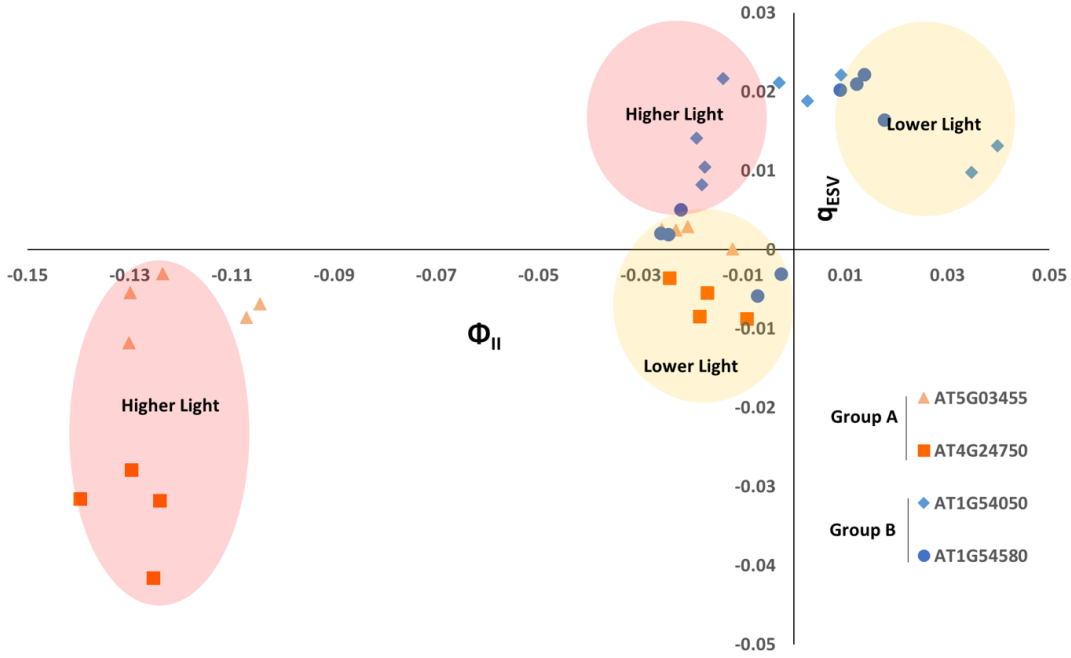

Figure 8. Two emerging phenomena found under strong fluctuating light conditions (between approximately $500\mu molm^{-2}s^{-1}$ (lower light) and $1000\mu molm^{-2}s^{-1}$ (higher light) four times repeated) have distinctively different photosynthetic phenotypes. Only two selected genotypes are shown for each group. In the first emerging phenomenon (group A, orange), plants have constantly low photoprotection yet the PS II activity decreased with the increased lights, indicating they are under stress. In the second one (group B, blue), less decrease of PS II activity and with high photoprotection as light increases, indicating they are well accommodated with the rapid changes of light.

their least common ancestor term and calculate the gene-gene similarity using $sim(g_i, g_j) = \log \frac{N}{N_i}$, where N is the total number of genes in the ontology and N_i is the number of genes annotated to the lowest common ancestor. The GO-based gene similarities were calculated using a web service named InteGO2 [21]. All the similarities were normalized to the range of $[0, 1]$.

Figure 9 shows the experimental results on the three net-

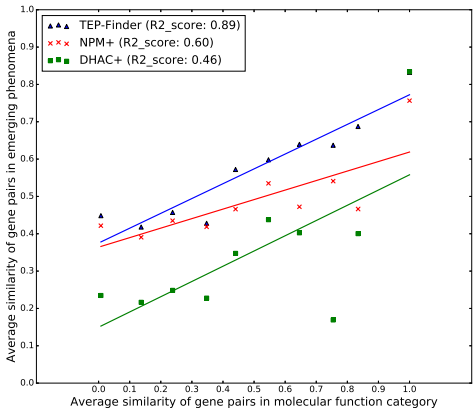

Figure 9. Comparing GO-based similarity with the EP-DAG based similarity. Gene pairs were clustered into 10 groups based on their GO-based similarities (x-axis), and for each group of gene pairs, we calculated the averaged EP-DAG based similarity (y-axis). Blue, green, and red represent the results of TEP-Finder, NPM+, and DHAC+.

works constructed using TEP-Finder, NPM+, and DHAC+, perspectively. In general, there is a strong correlation between the gene-gene similarities based on TEP-Finder and based on the GO. Specifically, the R^2 score of TEP-Finder is 0.89, significantly higher than that of the other two methods (i.e. 0.60 for NPM+ and 0.46 for DHAC+). This experiment suggests that the EP-DAG built by TEP-Finder is well organized.

E. Biological Significance

Although we now have deep knowledge of the core processes of photosynthesis, the "ancillary components" essential for function in living cells under dynamic conditions are largely unexplored [35]. Intriguingly, these ancillary components probably evolved as plug-in functional modules to adapt the core processes to different conditions. Understanding their functions may allow us to combine these modules in different organisms, to achieve rapid improvements in the photosynthetic efficiency. The identification of the emerging phenomena of chloroplast-targeted single knockouts may enable systematic analysis of genotype-phenotype connections and provides a clue on the characterization of specific "ancillary processes" that support efficient photosynthesis. Note that many of the ancillary photosynthetic processes down-regulate the capture of light energy, preventing photodamage but at the cost of light-capture efficiency. From an evolutionary perspective, these processes can be viewed as balancing needs for energy and the avoidance of deleterious effects from photosynthesis.

We first analyzed the identified emerging phenomena from the gene evolution perspective. Since essential genes are often slow evolving compared with genes with nonlethal mutant phenotypes, the genes identified only in the emerging phenomena under fluctuating light varying conditions may evolve faster than those in the emerging phenomena under smooth light conditions. The ratio Ka/Ks, which measures the relative rates of synonymous and nonsynonymous substitutions at a particular site, is often used for the estimation of evolutionary rates [22]. In our experiment, the averaged Ka/Ks ratio of the 50 genes appeared only in the emerging phenomena under strong and smooth light conditions is 0.164, while the averaged Ka/Ks ratio of the 45 genes identified uniquely under fluctuating and strong light conditions is 0.192, significantly higher than the former (permutation test, p-value=0.013).

We then analyzed the emerging phenomena from the perspective of photosynthetic functionality. Two emerging phenomena (A and B) were categorized under the same strong fluctuating light conditions in the middle of the day (between $500\mu molm^{-2}s^{-1}$ and $1000\mu molm^{-2}s^{-1}$ four times repeated) due to distinctively different photosynthetic phenotypes (Figure 8 and S2, A, orange; B, blue). The emerging phenomenon A consists of mutant lines AT1G12250, AT1G80030, AT4G24750, and AT5G03455. They are sensitive to fluctuating light, showing large extent of decreases in PS II activity and decreases in q_{ESV} (photoprotection) under high light intensity compared to the low light. Mutant lines in emerging phenomenon B (AT1G14590, AT1G54580, AT2G40400, AT3G10470, AT4G31560, AT5G03455, AT5G39830) have less extent of decreases in PS II activity with higher q_{ESV} indicating less sensitivity to the fluctuating light. As the important genes responsive to dynamic light conditions, sensitivity of mutant would be increased. Thus, for mutant lines that are shown a sensitive phenotype under the conditions, it indicates that the mutated genes are responsible for maintaining robust photosynthesis under the stress conditions. Hence, we hypothesize that the genes in A may contribute to photoprotection in response to natural light dynamics (see the selected samples in Figure 8). According to the GO, these genes are involved in arsenate reductase activity and the photosynthesis-related biological processes, including arsenate reductase activity and oxidation-reduction process. Most of them are related to cellular redox balance, which are important for regulation photosynthesis, yet mode of function of found genes in this study is still partly remained elusive [3], [14]. This analysis may provide new insights and open the new possibility to understand how plants are adapted dynamic conditions. Mutant lines in B stay high q_E and minor decrease in PS II activity in the high light indicating those mutants are less sensitive to dynamic light conditions. It shows that the mutate genes in B are less likely responsible to adapt fluctuating light conditions. A functional analysis based on GO shows that these genes are involved in cell cycle, cell division and protein complex oligomerization, and protein folding fatty acid biosynthetic process cytochrome b_6f complex assembly. Also, the averaged Ka/Ks ratio of A and B is 0.24 and 0.22 perspectively, which is significantly higher than that of randomly selected chloroplast-targeted genes (permutation test, p-value=0.024 and 0.013).

The biological analysis demonstrates that accurately identifying emerging phenomena from plant phenotyping data may be valuable towards the characterization of specific ancillary processes that support efficient photosynthesis.

VI. Discussion and Conclusion

Comprehensive analysis of emerging phenomena is required to improve our understanding of the quantitative variation of complex phenotypes and to attribute gene functions [11].

However, unlike frequent patterns, emerging phenomena may re-occur frequently or may appear only once during an experimental period, depending on the experimental design. TEP-Finder is the first tool towards capturing the emerging phenomena in large-scale longitudinal phenotyping experiments, leading to the identification of the minimum set of distinct actors needed to produce an undefined, complex aggregate phenotypic trait. Particularly, TEP-Finder can identify emerging phenomena in different temporal scales from the data and also can construct a directed acyclic network (EP-DAG) for better data management. The Gene Ontology and gene-gene association based performance evaluation show that TEP-Finder is better than the existing tools regarding biological significance.

An important component of TEP-Finder is the meta-clustering that repeatedly calls NPM with random anchor points for kernel density estimation. We tested whether the meta-clustering approach can lead to more robust results. Specifically, given the same input, we ran TEP-Finder and NPM+ three times and calculated the differences between the results of the three runs. Figure S1 shows the average number of genes per emerging phenomenon (indicated by circle area) at each level of the EP-DAG. In Figure S1(a), the three runs of TEP-Finder are similar to each other, indicated by the highly overlapped circles, whereas the three runs of NPM+, as shown in Figure S1(b), are distinctively different. In summary, the adoption of the meta-clustering approach ensures TEP-Finder to be robust enough for emerging phenomenon mining.

A key parameter in capturing emerging phenomena is K_3, the percentage of significant phenotype values. Unlike K_1 and K_2 that define the dimension of an emerging phenomenon, which is common in pattern recognition, K_3 is difficult to specify. Here we fixed K_1 and K_2 and varied K_3 to explore rules for choosing K_3. Table S1 indicates that with the increase of K_3, the EP-DAG becomes more concise (more shallow and has less amount of nodes), and the majority of the removed nodes are intermediate nodes. It suggests that to choose an optimal K_3, we can start with a high value and then gradually reduce it. At the same time, we should check whether the leaf nodes (which has the longest time frames) captures long-term patterns. As a future work, we will develop new algorithms to automatically optimize the parameters of TEP-Finder.

References

[1] Rakesh Agrawal, Johannes Gehrke, Dimitrios Gunopulos, and Prabhakar Raghavan. *Automatic subspace clustering of high dimensional data for data mining applications*, volume 27. ACM, 1998.

[2] Anna Alemany, Maria Florescu, Chloé S Baron, Josi Peterson-Maduro, and Alexander van Oudenaarden. Whole-organism clone tracing using single-cell sequencing. *Nature*, 556(7699):108, 2018.

[3] Michael Bauer and Jutta Papenbrock. Identification and characterization of single-domain thiosulfate sulfurtransferases from arabidopsis thaliana. *FEBS letters*, 532(3):427–431, 2002.

[4] C Coen Bron, Jagm Joep Kerbosch, and Hj Henk Schell. Finding cliques in an undirected graph. *Tech.univ.ndhoven*, 1972.

[5] Rich Caruana, Mohamed Elhawary, Nam Nguyen, and Casey Smith. Meta clustering. In *Data Mining, 2006. ICDM'06. Sixth International Conference on*, pp. 107–118. IEEE, 2006.

[6] F. Cazals and C. Karande. A note on the problem of reporting maximal cliques. *Theoretical Computer Science*, 407(1):564–568, 2008.

[7] Joshua N Cobb, Genevieve DeClerck, Anthony Greenberg, Randy Clark, and Susan McCouch. Next-generation phenotyping: requirements and strategies for enhancing our understanding of genotype–phenotype relationships and its relevance to crop improvement. *Theoretical and Applied Genetics*, 126(4):867–887, 2013.

[8] Jeffrey A Cruz, Linda J Savage, Robert Zegarac, Christopher C Hall, Mio Satoh-Cruz, Geoffry A Davis, William Kent Kovac, Jin Chen, and David M Kramer. Dynamic environmental photosynthetic imaging reveals emergent phenotypes. *Cell Systems*, 2(6):365–377, 2016.

[9] George Emanuel, Jeffrey R Moffitt, and Xiaowei Zhuang. High-throughput, image-based screening of pooled genetic-variant libraries. *nature methods*, 14(12):1159, 2017.

[10] Mario A. T. Figueiredo and Anil K. Jain. Unsupervised learning of finite mixture models. *IEEE Transactions on pattern analysis and machine intelligence*, 24(3):381–396, 2002.

[11] Pádraic J Flood, Willem Kruijer, Sabine K Schnabel, Rob Schoor, Henk Jalink, Jan FH Snel, Jeremy Harbinson, and Mark GM Aarts. Phenomics for photosynthesis, growth and reflectance in arabidopsis thaliana reveals circadian and long-term fluctuations in heritability. *Plant Methods*, 12(1):14, 2016.

[12] Qiaozi Gao, Elisabeth Ostendorf, Jeffrey A Cruz, Rong Jin, David M Kramer, and Jin Chen. Inter-functional analysis of high-throughput phenotype data by non-parametric clustering and its application to photosynthesis. *Bioinformatics*, 32(1):67–76, 2015.

[13] Daniel F Gudbjartsson, Hannes Helgason, Sigurjon A Gudjonsson, Florian Zink, Asmundur Oddson, Arnaldur Gylfason, Soren Besenbacher, Gisli Magnusson, Bjarni V Halldorsson, Eirikur Hjartarson, et al. Large-scale whole-genome sequencing of the icelandic population. *Nature genetics*, 47(5):435, 2015.

[14] Michael Hall, Alejandro Mata-Cabana, Hans-Erik Åkerlund, Francisco J Florencio, Wolfgang P Schröder, Marika Lindahl, and Thomas Kieselbach. Thioredoxin targets of the plant chloroplast lumen and their implications for plastid function. *Proteomics*, 10(5):987–1001, 2010.

[15] Iris M Heid and Thomas W Winkler. A multitrait gwas sheds light on insulin resistance. *Nature genetics*, 49(1):7, 2017.

[16] Jochen Hipp, Ulrich Güntzer, and Gholamreza Nakhaeizadeh. Algorithms for association rule mining—a general survey and comparison. *ACM sigkdd explorations newsletter*, 2(1):58–64, 2000.

[17] Sebastian Kuhlgert, Greg Austic, Robert Zegarac, Isaac Osei-Bonsu, Donghee Hoh, Martin I Chilvers, Mitchell G Roth, Kevin Bi, Dan TerAvest, Prabode Weebadde, et al. Multispeq beta: a tool for large-scale plant phenotyping connected to the open photosynq network. *Royal Society open science*, 3(10):160592, 2016.

[18] Pawan Lingras, Farhana Haider, and Matt Triff. Granular meta-clustering based on hierarchical, network, and temporal connections. *Granular Computing*, 1(1):71–92, 2016.

[19] Yongjin Park and Joel S. Bader. How networks change with time. *Bioinformatics*, 28(12):40–8, 2012.

[20] Emanuel Parzen. On estimation of a probability density function and mode. *Ann.math.statis*, 33(3):1065–1076, 1962.

[21] Jiajie Peng, Hongxiang Li, Yongzhuang Liu, Liran Juan, Qinghua Jiang, Yadong Wang, and Jin Chen. Intego2: a web tool for measuring and visualizing gene semantic similarities using gene ontology. *BMC genomics*, 17(5):553, 2016.

[22] G. I. Peterson and J Masel. Quantitative prediction of molecular clock and ka/ks at short timescales. *Molecular Biology & Evolution*, 26(11):2595–603, 2009.

[23] Resnik and Philip. Using information content to evaluate semantic similarity in a taxonomy. pp. 448–453, 1995.

[24] Murray Rosenblatt. Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, 27(3):832–837, 1956.

[25] Michael E Smoot, Keiichiro Ono, Johannes Ruscheinski, Peng-Liang Wang, and Trey Ideker. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, 27(3):431–432, 2010.

[26] Mahdi Soltanolkotabi, Ehsan Elhamifar, Emmanuel J Candes, et al. Robust subspace clustering. *The Annals of Statistics*, 42(2):669–699, 2014.

[27] P Sprent. Introduction to nonparametric estimation. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(4):944–945, 2009.

[28] Alok Kumar Srivastava and Narinder Singh Sahni. Ontovist: A general purpose ontological visualization tool. *Bioinformation*, 6(7):288–90, 2011.

[29] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. Uk biobank: an open access resource for identifying the causes of

a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3):e1001779, 2015.

[30] Etsuji Tomita, Akira Tanaka, and Haruhisa Takahashi. The worst-case time complexity for generating all maximal cliques. In *International Computing and Combinatorics Conference*, pp. 161–170, 2004.

[31] Peter M Visscher, Naomi R Wray, Qian Zhang, Pamela Sklar, Mark I McCarthy, Matthew A Brown, and Jian Yang. 10 years of gwas discovery: biology, function, and translation. *The American Journal of Human Genetics*, 101(1):5–22, 2017.

[32] Christian H Weiß. Association rule mining. *Wiley StatsRef: Statistics Reference Online*, pp. 1–6, 2014.

[33] Yifan Yang, Lei Xu, Zheyun Feng, Jeffrey A Cruz, Linda J Savage, David M Kramer, and Jin Chen. Phenocurve: capturing dynamic phenotype-environment relationships using phenomics data. *Bioinformatics*, 33(9):1370–1378, 2017.

[34] Guangchuang Yu, Ligen Wang, Yanyan Han, and Qingyu He. clusterprofiler: an r package for comparing biological themes among gene clusters. *Omics A Journal of Integrative Biology*, 16(5):284–287, 2012.

[35] Xin-Guang Zhu, Stephen P Long, and Donald R Ort. What is the maximum efficiency with which photosynthesis can convert solar energy into biomass? *Current opinion in biotechnology*, 19(2):153–159, 2008.