Analysis of Grapevine Gene Expression Data using Node-Based Resilience Clustering

Jeffrey Dale  
*Computer Science Dept.*  
Missouri State University  
Springfield, USA  
dale2755@live.missouristate.edu

John Matta  
*Computer Science Dept.*  
Southern Illinois University Edwardsville  
Edwardsville, USA  
jmatta@siue.edu

Susanne Howard  
*College of Agriculture*  
Missouri State University  
Mountain Grove, USA  
susannehoward@MissouriState.edu

Gunes Ercal  
*Computer Science Dept.*  
Southern Illinois University Edwardsville  
Edwardsville, USA  
gercal@siue.edu

Wenping Qiu  
*College of Agriculture*  
Missouri State University  
Mountain Grove, USA  
wenpingqiu@MissouriState.edu

Tayo Obafemi-Ajayi  
*Engineering Program*  
Missouri State University  
Springfield, USA  
tayoobafemiayaji@missouristate.edu

Abstract—Powdery mildew is the most economically important disease of cultivated grapevines worldwide. In the agricultural community, there is a great need for better understanding of the complex genetic basis of powdery mildew (PM) resistance by delineating possible gene biomarkers associated with the plants' defense mechanisms. Machine learning techniques can be applied to analysis of gene expression data to aid knowledge discovery of disease fighting genes. In this work, we apply a data-driven computational model, utilizing a graph-based clustering algorithm—Node-Based Resilience Clustering (NBR-Clust), to analyze grapevine gene expression data to identify possible gene biomarkers associated with powdery mildew disease defense mechanisms. We investigated two graph representations (geometric and kNN) on the mean differences of PM inoculated vs. mock inoculated gene expression values of Cabernet and Norton (PM disease resistant) species across 6 time points. By applying the contrarian approach, we hypothesized that smaller sized clusters will contain genes that do not follow general patterns, hence, could display distinct expression patterns of PM-induced transcripts across the time points that may insinuate biological relevance. We compared the smaller clusters obtained in Norton in contrast with the ones from Cabernet in terms of the genes that clustered in common between both (intersection of sets) as well as the differences of the sets. The results obtained demonstrated the usefulness of the geometric graphs for this domain application in contrast to the kNN graphs. Some genes that belong to biologically relevant pathways were identified that displayed differences in patterns across the time points between Norton and Cabernet species.

Index Terms—plant disease resistance, genes, clustering, graph theory, resilience measures.

I. INTRODUCTION

Machine learning (ML) is increasingly becoming a fundamental foundational approach for interdisciplinary research as intelligent data analytic systems span diverse applications beyond computer science and engineering. ML methodologies are useful in biological applications to characterize, process and integrate data to uncover useful information and actionable knowledge. ML encompasses different approaches of learning including supervised learning, unsupervised learning (also known as clustering) and reinforcement learning. Clustering techniques are particularly useful in domains where there is no ground truth or class information.

In this work, we investigate a data driven methodological framework using a novel graph-based method, node-based resilience clustering (NBR-Clust) [1], to infer meaningful biomarkers from heterogeneous noisy biological gene expression data. This study is a translational research that focuses on transforming concepts and theories into practical applications. Our domain application addresses disease resistance mechanisms of grapevine. The most economically important disease of cultivated grapevines worldwide is powdery mildew. In the agricultural community, there is a great need for better understanding of the complex genetic basis of powdery mildew resistance, and subsequently, genes associated with plants' defense mechanisms. Large-scale gene expression data coming from micro-array experiments provide new means to reveal fundamental cellular processes, investigate functions of genes, and understand relations and interactions among them. Our objective is to use this data to identify and characterize the set of genes responsible for powdery mildew resistance using novel graph-based clustering techniques.

Grapevine is the most cultivated fruit crop in world, having been cultivated for human consumption for over 7000 years [2], [3]. Powdery mildew (PM), the number one grape disease worldwide, is an economically important disease of grapevines that causes significant losses in yield and reduction in berry quality [3], [4]. To prevent this disease, fungicide is applied during production which translates to increased production costs. This also poses significant risk to the health of growers and pollution of the environment. There are certain grapevine species that have demonstrated a high degree of resistance to this disease but have a lower grape quality [4], [5]. In par-
particular, the Norton (Vitis aestivalis), the most widely planted grape in Missouri and its State Grape, is resistant. Cabernet Sauvignon (Vitis vinifera), a variety of superior quality and the most cultivated species, is very susceptible to the disease.

Given the significance and pressing need to identify the role of certain genes in disease resistance, some work [2], [4]–[7] has been done in genomic analysis of grapevine species. Qui et al. in [3] present a detailed review of current knowledge from multiple studies on the genetic basis of the powdery mildew disease. Genetic marker analysis has identified some genes that have shown differential expression levels between powdery mildew resistance in the wild Vitis species and the susceptible V. vinifera varieties [4]. For example, there is some evidence suggesting existence of a PEN1-mediated secretory pathway is an important component of pattern triggered immunity against powdery mildew in grapevine [3]. Hence, there is still a critical need for more targeted approaches to explore grapevine genes involved in powdery mildew susceptibility and pathways underlying this process.

We present a data driven methodological framework to investigate what set of gene features is responsible for the disease resistance mechanism of the Norton variety in contrast to the Cabernet Sauvignon species. The hypothesis is that there are certain genes associated with powdery mildew defense mechanisms in grapevines. Analysis of gene expression data from both Norton and Cabernet species using novel machine learning analytics methods should reveal and more clearly delineate disease resistance patterns and features that would increase our understanding of how plants fight against disease.

The remainder of this paper is organized as follows. In Section II, we present an overview of the NBR-Clust algorithm and graph types. In Section III, we describe the proposed framework for detection of significant set of genes that are could be associated with PM disease fighting mechanism. The results obtained are demonstrated and analyzed in Section IV. The conclusion and next steps are discussed in Section V.

II. BACKGROUND

To provide a context for the gene analysis methodology presented in this work, we briefly describe the underlying node-based resilience clustering (NBR-Clust) algorithm, the fundamentals of graph representations considered, and the grapevine gene expression dataset.

A. Node-Based Resilience Clustering

Node-based resilience clustering [1] is a graph-based clustering framework that utilizes the critical attack sets returned by graph theoretic resilience measures to cluster a graph. Every node-based resilience measure involves computation of a critical attack set of vertices S such that the removal of S results in a relatively significant disruption to the remaining network. Attack sets identify weaknesses or bottlenecks in graphs [8], and removal of the corresponding nodes breaks the graph into disjoint partitions. The NBR-Clust algorithm has been explored for five resilience measures [1]. Matta et al. in [1] discuss the different properties of the NBR-Clust algorithm based on each type of resilience measure. In this work, we apply NBR-Clust using the normalized integrity measure [9]. Normalized integrity is defined as:

$$I(G) = \min_{S \subseteq V} \left\{ \frac{|S| + C_{\max}(V - S)}{|V|} \right\},$$  (1)

where $G = (V, E)$ denotes a graph with a set of V vertices and E edges, S is the attack set which is a subset of V, and $C_{\max}(V - S)$ is the largest connected component in V − S.

One of the main advantages of NBR-Clust is the ability to cluster in one step (i.e. not requiring multiple recursive iterations) where the number of clusters is not known a priori. Results from clustering using a resilience measure like integrity have been shown to indicate a natural number of clusters for a graph [1]. The experimental results in [1] demonstrated the usefulness of integrity based clustering in finding the optimal number of clusters. Since the number of clusters sought in this work was unknown, NBR-Clust based on integrity was employed in this work. The NBR-Clust framework, as applied, consists of three key steps as follows:

<table>
<thead>
<tr>
<th>NBR-Clust Framework</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Convert data to a graph representation, $G$.</td>
</tr>
<tr>
<td>2 Compute integrity resilience measure ($I(G)$) and its corresponding attack set S.</td>
</tr>
<tr>
<td>3 Remove the attack set nodes $S$ from the graph and note the resulting clusters. (Attack set nodes are not reassigned in this work).</td>
</tr>
</tbody>
</table>

Given the computation hardness of computing integrity [10], a heuristic method known as GreedyBC was utilized to estimate integrity in step 2 of the NBR-Clust framework. GreedyBC relies on graph-theoretic measure called betweenness centrality. Using GreedyBC, the highest betweenness nodes are repeatedly removed from the graph, and with each removal the integrity resilience measure $I(S, G)$ is computed at that configuration. The configuration with the lowest integrity score is taken to be $I(G)$ and nodes $S$ removed to that point are considered to be the attack set. The GreedyBC algorithm based on integrity resilience measure $I(G)$ can be summarized as:

Greedy-BC Heuristic

1 $I_{min} = I(G), S_{min} = \emptyset$
2 repeat |V| times
3 $v = \arg\max_{v \in V} BC(v)$
4 $G = G \setminus \{v\}$ and $S = S \cup \{v\}$
5 if $I(S, G) < I_{min}$ then
6 $I_{min} = I(S, G)$ and $S_{min} = S$
6 return $S_{min}$

B. Graph Representations

The study of applicable graph representations is fundamental to the success of graph-based clustering approaches. The first step of NBR-Clust is to convert the input data into a graph representation. For the same dataset, there could exist some graph representations that result in a large number of
components even from the removal of very small $S$. In this work, we investigate the effectiveness of two different graph types: geometric graph and k-nearest neighbors (kNN) graph. Both are parametrized graph types, with radius threshold $r$ required for the geometric graph and number of nearest neighbors chosen, $k$, required for the kNN graph. In previous works [1], [11] it was noted that, depending on parameters used, the kNN graphs were generally more efficient to cluster and yielded better results. This was probably due to the significantly smaller number of edges which made it easier to partition the graph with a smaller attack set. In the current work, we apply a dimensionality space reduction tool (as described in details in Section III) and have thus obtained better results with geometric graphs.

1) Construction of Geometric Graphs: The Weighted Gene Correlation Network Analysis (WGCNA) R package [12] was employed to convert the input gene expression data to desired geometric graphs. Each node in a geometric graph is connected to all nodes within a certain radius, or distance. WGCNA returns an adjacency matrix whose entries are the correlations between the gene indicated by the row of the input data matrix and the gene indicated by the column of the input data matrix. All of the entries of the adjacency matrix are in $[-1, 1]$. WGCNA computes correlation using either Pearson or Biweight Mid-Correlation measure (bicor) [13]. We selected bicor as it has been demonstrated in literature to be more robust [14].

WGCNA requires specification of the thresholding power parameter $t_p$ to determine the minimum level of correlation between genes required for inclusion in the graph. Let $B$ denote the final adjacency matrix with the less-correlated genes filtered from the original adjacency matrix $A$. In general, for an entry $a_{ij}$ of $A$ and a soft thresholding power $t_p$, $b_{ij}$ is given by

$$b_{ij} = \begin{cases} a_{ij}, & \text{if } a_{ij}^{t_p} \geq 0.5 \\ 0, & \text{otherwise} \end{cases}$$

A soft thresholding power $t_p = 7$ (pow7) implies that genes with correlation less than approximately 0.906 will have their corresponding entry of $B$ set to zero. Higher powers tend to result in a more sparse matrix of strongly correlated genes. By varying the $t_p$ value for the same input data, we obtain graphs with different sizes and properties as illustrated in Table I. The analysis reported in this paper are based on $t_p = 7$ (pow7).

WGCNA also requires specifying the graph network type: signed, unsigned, signed hybrid, or distance. The network type, signed hybrid, was applied as recommended by the WGCNA R package documentation. The signed hybrid network is a hybrid of a weighted and unweighted network. The similarity is set to the correlation value if positive, otherwise it is equal to zero. The distance between two genes is inversely proportional to the correlation between their expression levels across all conditions. This means that two highly correlated gene will be geometrically close to each other, while uncorrelated genes will have a larger distance between them.

### Table I: Properties of Geometric vs. kNN Graphs Generated

<table>
<thead>
<tr>
<th>Species</th>
<th>Threshold Power/radius</th>
<th>Nodes</th>
<th>Edges</th>
<th>Average Degree</th>
<th>Modularity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Norton</td>
<td>pow7</td>
<td>9,113</td>
<td>16,876</td>
<td>4</td>
<td>0.91</td>
</tr>
<tr>
<td>Norton</td>
<td>kNN-30</td>
<td>9,113</td>
<td>113,297</td>
<td>25</td>
<td>0.73</td>
</tr>
<tr>
<td>Norton</td>
<td>pow5</td>
<td>9,113</td>
<td>534,786</td>
<td>117</td>
<td>0.66</td>
</tr>
<tr>
<td>Norton</td>
<td>pow6</td>
<td>9,112</td>
<td>388,486</td>
<td>85</td>
<td>0.68</td>
</tr>
<tr>
<td>Norton</td>
<td>kNN-5</td>
<td>9,113</td>
<td>294,993</td>
<td>65</td>
<td>0.70</td>
</tr>
<tr>
<td>Norton</td>
<td>kNN-50</td>
<td>9,113</td>
<td>16,876</td>
<td>4</td>
<td>0.91</td>
</tr>
<tr>
<td>Norton</td>
<td>pow5</td>
<td>9,113</td>
<td>534,786</td>
<td>117</td>
<td>0.66</td>
</tr>
<tr>
<td>Norton</td>
<td>pow6</td>
<td>9,112</td>
<td>388,486</td>
<td>85</td>
<td>0.68</td>
</tr>
<tr>
<td>Norton</td>
<td>kNN-5</td>
<td>9,113</td>
<td>294,993</td>
<td>65</td>
<td>0.70</td>
</tr>
<tr>
<td>Norton</td>
<td>kNN-50</td>
<td>9,113</td>
<td>16,876</td>
<td>4</td>
<td>0.91</td>
</tr>
</tbody>
</table>

2) Construction of kNN Graphs: The kNN graphs were derived from the data using the Class Cover Catch Digraphs (CCCD) R package\(^1\). It creates a kNN graph such that an edge is added between vertices $u$ and $v$ if $u$ is one of the $k$ nearest neighbor to $v$, and $v$ is one of the $k$ nearest neighbor to $u$. Due to this symmetry requirement, each vertex will have a degree of at most $k$. Depending on the parameters used, kNN graphs tend to have fewer edges than geometric graphs created from the same data, as illustrated in Table I for $k = 5$ and $k = 30$.

Table I also includes the modularity value for the various graphs constructed and considered. Modularity is a measure that quantifies the strength of modules (analogous to clusters) created when clustering a graph. A graph with high modularity has more than expected edges internal to its modules, and fewer than expected edges between modules. We applied modularity as a quick way to evaluate the “clusterability” of a graph. As can be observed from Table I, the kNN graphs had the higher modularity values with kNN-5 ranking the highest. However, from preliminary experiments, we observed that the lower $k$ values resulted in graphs that were very sparse (not well-connected). Hence, we performed subsequent analysis on the kNN-30 graphs even though they were denser.

### C. Grapevine Gene Expression Dataset

Gene expression data consists of expression levels of a large set of genes measured across a set of conditions or samples organized as a data matrix with rows corresponding to genes and columns corresponding to samples or conditions. To identify differentially expressed genes in the two grapevine genotypes, *Vitis vinifera* (Cabernet sauvignon) and *Vitis aestivalis* (Norton), in response to the PM fungus, gene expression data [4], [6] was obtained by inoculating PM onto detached leaves of both Norton and Cabernet species. One leaf was harvested from each plant and ten leaves were pooled as one sample at each time point. Powdery mildew conidia-inoculated and mock-inoculated samples were ground separately in liquid nitrogen. Samples of three replicates were processed for analysis. The gene expression data was generated on Affymetric GeneChip *Vitis vinifera* genome array and extracted using the GeneChip operating software version 1.2.

\(^1\)Available at https://CRAN.R-project.org/package=cccd
(The Affymetrix data is publicly available on the NCBI Gene Expression Omnibus website, accession number is GSE6404.)

The input gene expression data utilized in this work consisted of 36 samples for each Norton and Cabernet species i.e. 72 in total. For each species, there are 3 biological replicates of inoculated and non-inoculated (“mock”) specimens at 6 different time points: initially (0 hour), and after 4, 8, 12, 24 and 48 hours. If any of the 72 samples for a particular feature (gene) was deemed absent, that gene was removed from further consideration. A total of 9113 genes were verified to be 100% present or marginal across all conditions for both species.

III. METHODOLOGY

The objective of this work was to apply a ML framework to identify features (genes) that significantly change in response to the PM inoculation and differ in pattern between Norton and Cabernet species. The overall data-driven approach, illustrated in Figure 1, consists of 5 key phases which we describe in detail below.

A. Data Preprocessing

To investigate a suited data-driven ML framework, it is important to understand key characteristics of the problem at hand ‘and how it translates to a computational intelligence framework. The preprocessing phase takes care of all necessary modifications that need to be addressed to translate the data effectively. We are interested in observing the differences or similarity in the expression patterns of Cabernet vs. Norton across biological relevant genes in response to the PM inoculation across the different time points. Hence, we compute the mean differences in expression values between the inoculation and mock treatments over the six time points. The three biological replicates at each time point were aggregated using geometric mean ($d = \sqrt[3]{d_1d_2d_3}$, where $d_x$ represents the data point from replica $x$), prior to computing the mean differences. We also separated the input matrix into two separate data files since we were interested in comparing Cabernet against Norton. Thus, for subsequent analysis, we had two data matrices (Cabernet and Norton) with 6 conditions (representing each time point), a significantly reduced matrix from an input matrix of 72 conditions.

B. Graph Representation

During this phase, the preprocessed data is converted to a graph form. As mentioned in Section II-B, we explored two different graph representations: geometric and kNN graphs using the mean difference between inoculated and mock values, for both Norton and Cabernet species. In this study, we selected the geometric graph with $t_p = 7$ (pow7) given that it was the least dense of the geometric graphs along with the kNN-30 graphs. The denser the graph, the greater the computational complexity. From a biological standpoint, the lower the power, the more likely noisy the resulting data would be. Thus, selecting a graph with higher thresholding power probably implies more relevant results.

C. Cluster Analysis

We conducted cluster analysis on the graphs using the NBR-Clust clustering framework with integrity resilience measure, as described in section II-A. The cluster analysis returns sub-partitions (clusters) of graph that represent genes that follow similar patterns of behavior i.e., the genes react similarly in terms of the mean difference between the inoculation and mock values across the six different time points (conditions). The question then is: how do we merge the clustering results obtained from the Cabernet graphs with the Norton graphs? How do we determine which are the biologically relevant clusters? We address these questions in the next phase.

D. Contrarian Approach

When clustering a very large dataset such as gene expression data or genotype data, it is known that the relevant sets of genes are usually very small in comparison to the initial data size. Most genes behave similarly, in accordance with basic functions of the cell. Usually, the interesting subset of genes is much smaller. The contrarian approach, as discussed in [15], applies clustering as a tool to filter out a majority of the non-essential (or non-interesting) genes and focuses on the outliers or the smaller clusters. It is assumed that the large clusters group the non-essential and basic functional genes together. As denoted by the term “contrarian approach” meaning “contrary”, we are interested in the small clusters of genes that act “contrary” to the generally observed patterns. Our hypothesis is that these clusters will contain genes that do not follow general patterns, hence, could display distinct expression patterns of PM-induced transcripts across the time points that may insinuate biological relevance. We compared the smaller clusters obtained in Norton in contrast with the ones from Cabernet in terms of the genes that clustered in common between both (intersection of sets) as well as the differences of the sets.

E. Biological Analysis

To determine the biological relevance of the results obtained, we employed the Database for Annotation, Visualization and Integrated Discovery (DAVID) online tool [16], [17]. Given a list of genes, the DAVID tool returns sets of genes that belong to known enriched functional-related gene groups and pathways among the input genes. Possible relevant pathways of genes that confer resistance to powdery mildew in the Norton species include catabolic process, biosynthesis, kinase, plant pathogen, and membrane. It is important to note that the membrane pathway could be also regarded as basic plant functional pathway. However, depending on the significance of the differences in reaction between both species, it could be interesting. We compared the means and standard deviations across time periods determine if the differences in the clusters of patterns observed are significance. For further analysis, we plan to incorporate more rigorous statistical tests to compare difference in means across all time points between both species.
We computed a set difference of the smaller sized clusters in both Cabernet and Norton graphs. The emphasis is on genes that are in the Norton cluster but not in the Cabernet cluster, given that the prior work had observed differences in the Cabernet but not the Norton species. These genes were analyzed by the DAVID tool. The results revealed several genes that matched important keywords related to possible functional pathways that could influence plant disease resistance mechanism. Figures 2, 3 and 4 illustrate a graph network visualization of the set difference between two interesting clusters, one of Norton and one of Cabernet. The important genes related to key pathways are denoted by different colors for ease of visualizations. In Figures 2, 3 and 4, the red nodes represent genes identified with catabolic process, orange nodes: biosynthesis, yellow nodes: plant pathogen, green nodes: kinase, and cyan nodes: membrane respectively. Nodes that are not matched with a pathway of interest are colored in gray.

IV. RESULTS AND ANALYSIS

Cluster analysis on the geometric (pow7) graphs vs. the kNN-30 graphs yielded different results as shown in Table II. It is interesting to note that for both types of graphs, the size of the critical attack set is very large. The geometric graphs produced one smaller size cluster (size 604) on the Norton dataset and 3 other large clusters (> 1700). The Cabernet graph seemed to partition readily into more clusters: about 10 with 6 of them less than 300 genes. The kNN graphs resulted in a higher number of clusters for both the Cabernet and Norton graphs with a close range of size distribution. The geometric graphs also seemed to result in more biologically relevant results as demonstrated in Tables III, IV and V.

Fung et al. had identified a set of 626 genes that were differentially expressed between the PM-inoculated and the mock-inoculated in Cabernet [4]. Of these 626 genes, 491 were a subset of the 9113 genes explored in this work. We were also interested in observing which clusters contained a high percentage of these already known genes. In Table II, the numbers in parenthesis indicate the number of genes in the corresponding cluster that overlap with known subset of 491 genes. As can be observed from Table II, for the kNN-30 graphs, the already known genes appear spread out through all the clusters, unlike the pow7 geometric graphs that seem to have them less evenly distributed. Both graphs had approximately the same percentage of those known genes in the critical attack set across both species. We plan to explore the biological relevance of the critical attack set in future work.
that the biosynthesis nodes tend to cluster together, tend to have higher betweenness centrality, and in some cases groups of biosynthesis nodes are matched with a corresponding membrane node. In contrast, many membrane nodes are distributed throughout the graph, and many have a low importance in the network, as expected. The catabolic process genes in red also tend to be important, but there are fewer of them, and the pattern is harder to distinguish. The corresponding network visualization results for the kNN-30 graphs, displayed in Figure 5, do not seem to yield any significant results unlike the geometric graphs.

The results from the DAVID tool comparisons for the geometric graphs are summarized for a few of the cluster comparisons in Tables III, IV, and V. The top row for each pathway denotes the mean and standard deviation values for the Norton species while the second row denotes the results for the Cabernet species. As can be observed, across all the tables the Norton species generally appeared to have a lower variance around the mean. The difference in mean analysis for the geometric graph suggests significant different in patterns observed across the Norton species in contrast to the Cabernet species. The results obtained by applying the DAVID tool on the kNN graphs (Tables VI and VII) did not also yield interesting results. Figure IV illustrates the trends between the Norton and Cabernet species for certain cultivars belonging to the plant pathogen pathway found in cluster 0 of Norton (N0) but not in cluster 5 of Cabernet (C5). We can observe varied differences in the Cabernet species at this cultivar, as they end up in different clusters.

V. Conclusion

This paper investigated a data-driven computational model, utilizing a graph-based clustering algorithm – Node-Based Resilience Clustering (NBR-Clust), to analyze grapevine gene expression data to identify possible gene biomarkers associated with powdery mildew disease defense mechanisms. We explored two graph representations (geometric and kNN) on the mean differences of PM inoculated vs. mock inoculated gene expression values of Cabernet and Norton (PM disease resistant) species across 6 time points. The objective was to analyze the gene profiles obtained from mock inoculated vs. PM inoculated to identify sets of gene features responsible for the disease resistance mechanism of the Norton variety in contrast to the Cabernet Sauvignon species. By applying the contrarian approach, we hypothesized that smaller sized clusters will contain genes that do not follow general patterns, hence, could display distinct expression patterns of PM-induced transcripts across the time points that may insinuate biological relevance. Smaller clusters obtained in Norton were compared in contrast with the ones from Cabernet in terms of the genes that clustered together in one but not the other i.e. differences of the sets. The results obtained demonstrated the usefulness of the geometric graphs for this domain application in contrast to the kNN graphs. We also identified some genes that differ in patterns between Norton and Cabernet species that would be further explored to validate usefulness as potential biomarkers.
### TABLE IV
MEAN AND STANDARD DEVIATION COMPARISONS FOR THE SET DIFFERENCES OF GEOMETRIC GRAPH (Norton 0 (upper row) vs Cabernet 6 (lower row)) ACROSS CERTAIN BIOLOGICAL PATHWAYS.

<table>
<thead>
<tr>
<th>Pathway</th>
<th>0 hours</th>
<th>4 hours</th>
<th>8 hours</th>
<th>12 hours</th>
<th>24 hours</th>
<th>48 hours</th>
</tr>
</thead>
<tbody>
<tr>
<td>bioynthesis</td>
<td>10.92 ± 69.1</td>
<td>-4.49 ± 61.14</td>
<td>36.32 ± 72.12</td>
<td>-80.42 ± 92.27</td>
<td>-138.84 ± 136.35</td>
<td>174.41 ± 194.17</td>
</tr>
<tr>
<td>catabolic process</td>
<td>46.65 ± 32.78</td>
<td>40.88 ± 55.35</td>
<td>115.6 ± 115.54</td>
<td>-71.69 ± 59.97</td>
<td>-87.56 ± 97.1</td>
<td>175.38 ± 141.31</td>
</tr>
<tr>
<td>kinase</td>
<td>58.57 ± 73.43</td>
<td>-80.76 ± 93.51</td>
<td>108.68 ± 99.05</td>
<td>36.13 ± 103.82</td>
<td>95.52 ± 92.15</td>
<td>25.5 ± 130.82</td>
</tr>
<tr>
<td>pathogen</td>
<td>-5.46 ± 50.74</td>
<td>-14.28 ± 53.27</td>
<td>22.62 ± 51.92</td>
<td>-44.8 ± 78.34</td>
<td>-76.07 ± 131.59</td>
<td>100.08 ± 191.04</td>
</tr>
<tr>
<td>0.95 ± 226.64</td>
<td>-141.78 ± 285.07</td>
<td>-8.3 ± 115.65</td>
<td>30.42 ± 115.28</td>
<td>-95.51 ± 226.32</td>
<td>-135.89 ± 253.03</td>
<td>163.12 ± 170.11</td>
</tr>
<tr>
<td>1.06 ± 226.97</td>
<td>-199.1 ± 324.15</td>
<td>25.54 ± 128.89</td>
<td>-21.62 ± 159.01</td>
<td>-170.32 ± 287.52</td>
<td>-169.47 ± 284.2</td>
<td></td>
</tr>
</tbody>
</table>

### TABLE V
MEAN AND STANDARD DEVIATION COMPARISONS FOR THE SET DIFFERENCES OF GEOMETRIC GRAPH (Norton 0 (upper row) vs Cabernet 8 (lower row)) ACROSS CERTAIN BIOLOGICAL PATHWAYS.

<table>
<thead>
<tr>
<th>Pathway</th>
<th>0 hours</th>
<th>4 hours</th>
<th>8 hours</th>
<th>12 hours</th>
<th>24 hours</th>
<th>48 hours</th>
</tr>
</thead>
<tbody>
<tr>
<td>bioynthesis</td>
<td>10.45 ± 68.68</td>
<td>-4.39 ± 60.67</td>
<td>35.83 ± 71.68</td>
<td>-79.43 ± 91.92</td>
<td>-137.14 ± 136.01</td>
<td>171.98 ± 193.69</td>
</tr>
<tr>
<td>catabolic process</td>
<td>46.65 ± 32.78</td>
<td>40.88 ± 55.35</td>
<td>115.6 ± 115.54</td>
<td>-71.69 ± 59.97</td>
<td>-87.56 ± 97.1</td>
<td>175.38 ± 141.31</td>
</tr>
<tr>
<td>kinase</td>
<td>58.57 ± 73.43</td>
<td>-80.76 ± 93.51</td>
<td>108.68 ± 99.05</td>
<td>36.13 ± 103.82</td>
<td>95.52 ± 92.15</td>
<td>25.5 ± 130.82</td>
</tr>
<tr>
<td>pathogen</td>
<td>-5.55 ± 53.33</td>
<td>-14.98 ± 55.95</td>
<td>27.14 ± 53.05</td>
<td>-50.09 ± 81.07</td>
<td>-84.73 ± 137.11</td>
<td>119.58 ± 197.07</td>
</tr>
<tr>
<td>0.63 ± 227.77</td>
<td>-159.97 ± 297.1</td>
<td>-7.96 ± 121.41</td>
<td>-33.31 ± 120.71</td>
<td>-101.8 ± 236.42</td>
<td>-157.35 ± 260.84</td>
<td>163.12 ± 170.11</td>
</tr>
<tr>
<td>1.06 ± 226.97</td>
<td>-199.1 ± 324.15</td>
<td>25.54 ± 128.89</td>
<td>-21.62 ± 159.01</td>
<td>-170.32 ± 287.52</td>
<td>-169.47 ± 284.2</td>
<td></td>
</tr>
</tbody>
</table>

### TABLE VI
MEAN AND STANDARD DEVIATION COMPARISONS FOR THE SET DIFFERENCES OF kNN GRAPH (Norton 10 (upper row) vs Cabernet 10 (lower row)) ACROSS CERTAIN BIOLOGICAL PATHWAYS.

<table>
<thead>
<tr>
<th>Pathway</th>
<th>0 hours</th>
<th>4 hours</th>
<th>8 hours</th>
<th>12 hours</th>
<th>24 hours</th>
<th>48 hours</th>
</tr>
</thead>
<tbody>
<tr>
<td>bioynthesis</td>
<td>10.34 ± 46.37</td>
<td>-7.8 ± 38.88</td>
<td>4.11 ± 26.19</td>
<td>11.46 ± 47.16</td>
<td>-29.29 ± 42.93</td>
<td>-0.23 ± 50.95</td>
</tr>
<tr>
<td>kinase</td>
<td>21.75 ± 35.01</td>
<td>-22.48 ± 18.21</td>
<td>3.27 ± 10.72</td>
<td>17.65 ± 23.84</td>
<td>-10.13 ± 8.42</td>
<td>-6.8 ± 11.48</td>
</tr>
<tr>
<td>membrane</td>
<td>11.48 ± 60.53</td>
<td>-14.26 ± 27.92</td>
<td>16.02 ± 48.57</td>
<td>25.66 ± 81.45</td>
<td>24.96 ± 40.1</td>
<td>4.48 ± 31.5</td>
</tr>
<tr>
<td>37.16 ± 106.37</td>
<td>-15.29 ± 21.79</td>
<td>1.46 ± 16.46</td>
<td>10.69 ± 28.49</td>
<td>-20.68 ± 45.06</td>
<td>-17.09 ± 34.6</td>
<td>45.02 ± 109.81</td>
</tr>
</tbody>
</table>

### TABLE VII
MEAN AND STANDARD DEVIATION COMPARISONS FOR THE SET DIFFERENCES OF kNN GRAPH (Norton 12 (upper row) vs Cabernet 9 (lower row)) ACROSS CERTAIN BIOLOGICAL PATHWAYS.

<table>
<thead>
<tr>
<th>Pathway</th>
<th>0 hours</th>
<th>4 hours</th>
<th>8 hours</th>
<th>12 hours</th>
<th>24 hours</th>
<th>48 hours</th>
</tr>
</thead>
<tbody>
<tr>
<td>bioynthesis</td>
<td>-9.17 ± 13.82</td>
<td>15.45 ± 35.73</td>
<td>-5.66 ± 17.28</td>
<td>-1.34 ± 32.35</td>
<td>-17.01 ± 36.96</td>
<td>6.9 ± 46.6</td>
</tr>
<tr>
<td>kinase</td>
<td>-27.24 ± 46.63</td>
<td>-1.51 ± 41.11</td>
<td>-18.48 ± 40.58</td>
<td>-34.53 ± 56.53</td>
<td>-55.5 ± 118.71</td>
<td>5.32 ± 25.7</td>
</tr>
<tr>
<td>membrane</td>
<td>-8.92 ± 17.35</td>
<td>0.24 ± 34.32</td>
<td>2.84 ± 16.43</td>
<td>25.19 ± 25.9</td>
<td>7.39 ± 33.94</td>
<td>13.49 ± 42.22</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pathway</th>
<th>0 hours</th>
<th>4 hours</th>
<th>8 hours</th>
<th>12 hours</th>
<th>24 hours</th>
<th>48 hours</th>
</tr>
</thead>
<tbody>
<tr>
<td>bioynthesis</td>
<td>-5.1 ± 36.68</td>
<td>18.6 ± 20.13</td>
<td>1.72 ± 9.96</td>
<td>-16.98 ± 19.0</td>
<td>18.85 ± 7.75</td>
<td>5.04 ± 19.44</td>
</tr>
<tr>
<td>kinase</td>
<td>-8.92 ± 17.35</td>
<td>0.24 ± 34.32</td>
<td>2.84 ± 16.43</td>
<td>25.19 ± 25.9</td>
<td>7.39 ± 33.94</td>
<td>13.49 ± 42.22</td>
</tr>
<tr>
<td>membrane</td>
<td>-3.65 ± 10.95</td>
<td>4.0 ± 13.31</td>
<td>-5.05 ± 11.96</td>
<td>-8.92 ± 14.93</td>
<td>5.63 ± 21.16</td>
<td>-4.57 ± 18.1</td>
</tr>
<tr>
<td>-9.39 ± 18.13</td>
<td>5.97 ± 14.93</td>
<td>0.08 ± 11.58</td>
<td>0.14 ± 24.47</td>
<td>8.65 ± 46.0</td>
<td>21.33 ± 56.84</td>
<td></td>
</tr>
</tbody>
</table>
Fig. 5. Three cluster comparisons of set differences between Norton and Cabernet based on the kNN-30 graph.

Fig. 6. Trend in difference of expression levels across the 6 time points (0, 4, 8, 12, 24, 48 hrs) for selected cultivar belonging to Norton cluster 0 (N0) but not in Cabernet cluster 5 (C5). The blue dashed lines represent Norton while the orange solid lines - Cabernet. Below each chart, we specify the cluster each gene is found in the corresponding Cabernet graph. C-S denotes critical attack set of Cabernet

REFERENCES


