OLIVER: A Tool for Visual Data Analysis on Longitudinal Plant Phenomics Data

Oliver L Tessmer
Dept. of Energy Plant Research Lab
Michigan State University
East Lansing, USA
tessmero@msu.edu

David M Kramer*
Dept. of Energy Plant Research Lab
Michigan State University
East Lansing, USA
kramerd8@msu.edu

Jin Chen*
Inst. for Biomedical Informatics
University of Kentucky
Lexington, USA
chen.jin@uky.edu

***Abstract*—There is a critical unmet need for new tools to analyze and understand "big data" in the biological sciences where breakthroughs come from connecting massive genomics data with complex phenomics data. By integrating instant data visualization and statistical hypothesis testing, we have developed a new tool called OLIVER for phenomics visual data analysis with a unique function that any user adjustment will trigger real-time display updates for any affected elements in the workspace. By visualizing and analyzing omics data with OLIVER, biomedical researchers can quickly generate hypotheses and then test their thoughts within the same tool, leading to efficient knowledge discovery from complex, multi-dimensional biological data. The practice of OLIVER on multiple plant phenotyping experiments has shown that OLIVER can facilitate scientific discoveries. In the use case of OLIVER for large-scale plant phenotyping, a quick visualization identified emerging phenomena that are highly transient and heterogeneous. The unique circular heat map with false-color plant images also indicates that such emerging phenomena appear in different leaves under different conditions, suggesting that such previously unseen processes are critical for plant responses to dynamic environments.**

***Index Terms*—phenomics, instant data visualization, heat map**

I. Introduction

Modern genetic tools have revolutionized biology [27], [32]. It is now possible to rapidly and inexpensively sequence the genomes of any organism on the planet [28], [33]. However, this huge dataset does not, by itself, answer most major questions in biology, because we do not know how these genes control complex biological processes. The next revolution in biology will be to connect the genomes to the phenomes of organisms, i.e. to understand how the genome and its related metagenome control complex behaviors [4], [45]. To achieve this integration, we need to measure and interpret multiple phenotypes (i.e., the specific properties and behaviors of organisms) under a range of dynamic environmental conditions (or stimuli) in a large number of genetic variants [13], [22]. This is a hyper-complex problem requiring new approaches to measurements, data analyses and the generation of functional insights. Recent developments in plant photosynthesis phenotyping make it possible to probe plant growth, photosynthesis, and other properties under dynamic environmental conditions [7], [10], [17], [30], [43]. Similar approaches are impacting other fields, such as biochemistry, drug development and behavior studies [3], [5], [6], [39].

Despite these major phenotyping technological advances, biomedical scientists are facing the difficulty of analyzing longitudinal phenomics data, for the nonlinear temporal patterns in a high-dimensional space are difficult to detect. Particularly, plant photosynthetic phenomics experiments are often conducted under dynamic environmental conditions, in order to measure the influence of the environment (e.g. abrupt changes in light intensity or temperature) on photosynthesis (e.g. excessive losses in photosynthetic efficiency and/or increases in photodamage). In plants, photosynthesis must respond to changing environment to provide the optimal amount of energy to meet the needs of the organism, in the correct forms, without producing toxic byproducts, e.g. reactive oxygen species or glycolate [12], [23], [46], [49]. In this context, photosynthesis can be viewed as a set of integrated modules that form a self-regulating network that is sensitive to changes in both environmental parameters (e.g. light intensity) and metabolic or physiological factors [21]. Consequently, it is crucial to *simultaneously* model genotypes, phenotypes, and environmental factors and identifies complex relationships among multiple variables to arrive at a holistic characterization of plant performance [49].

In practice, for plant phenomics data analysis, a combination of visual and statistical data analytics approach emerges [7]. The concept of visual knowledge discovery, which has been extensively utilized in a wide variety of data mining and machine learning applications [9], [14]–[16], [41], adopts the unique human perceptual and cognitive abilities and combines it with modern machine learning techniques, leading to accelerated multi-dimensional pattern recognition [20].

A critical component of visual and statistical data analysis is visual data representation. Due to the limitation of human cognition and perception, the quantity of information a person can examine at a moment is limited, which significantly affects the volume of data a user can view and consequently the pattern a person can retrieve for any specific purposes [44], [47]. A *clean, simple and interactive phenomics data visualization interface* is required for understanding, comparing, and analyzing complicated relationships among individuals in

This project is supported by Department of Energy BES program (grant no. DE-FG02-91ER20021) and National Science Foundation ABI program (grant no. 1458556).

* to whom correspondence should be addressed.

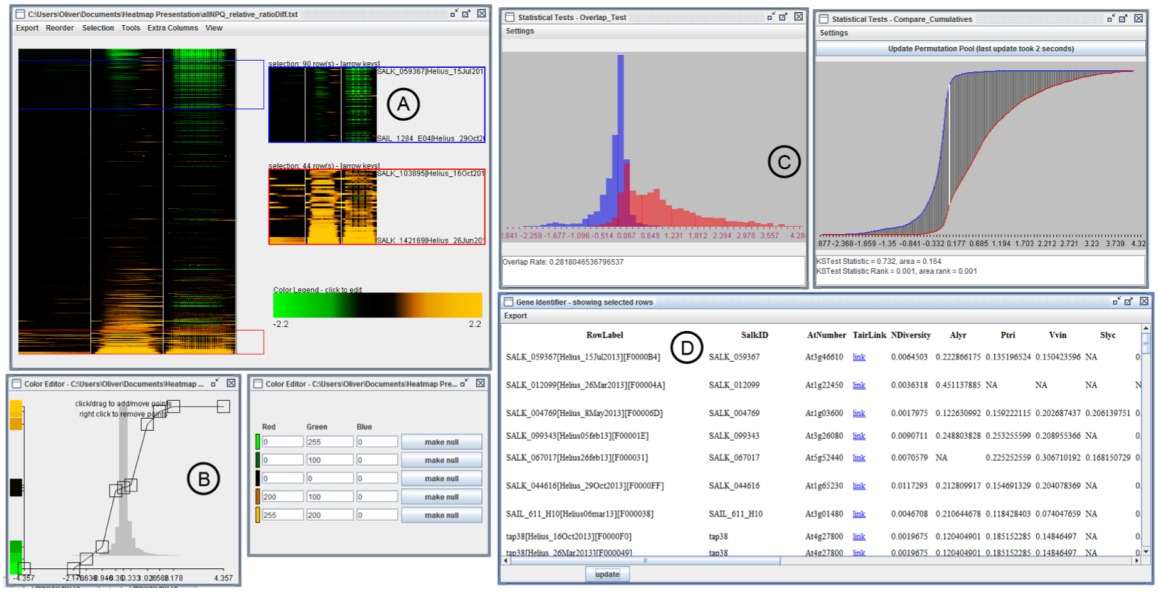

Fig. 1. OLIVER workspace with (a) a phenotype-over-time heat map loaded from a spreadsheet, with two subsets of rows selected, (b) two dialogs for editing the color curve of the heat map, one to edit the RGB channel intensities (right), and the other to change the way the base colors are interpolated (left), (c) statistics dialogs showing overlaid histograms representing the frequency distributions of values in the selected subsets (left), as well as a comparison of their cumulative curves and a p-value (right), (d) the Gene Identifier dialog, which presents external data corresponding to a row selection in the heat map.

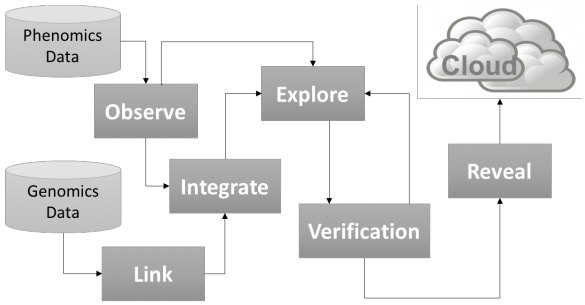

Fig. 2. The workflow of OLIVER. OLIVER is the acronym of six key processes in visual data analysis: Observe, Link, Integrate, Verify, Explore, and Reveal.

a multi-dimensional space [7], [19]. With the phenomics data visualization interface, users can effectively gain insights from big data to interpret complex behaviors, to determine which genes are responsible to certain phenotypes under specific environmental conditions, or to identify genes involved in certain biological processes and test mechanistic hypotheses.

To facilitate visual phenomics knowledge discovery, we present an interactive phenomics data visualization tool called **OLIVER**, which is the acronym of six key processes in visual data analysis, i.e. Observe, Link, Integrate, Verify, Explore, and Reveal (see the main interface in Figure 1 and the workflow in Figure 2.). OLIVER visually represents data using interactive animated heat maps. Comparing with the existing data visualization tools such as KaleidaGraph [42], Origin, and Microsoft Excel that are often used to generate heat maps, OLIVER is advantageous on its design principles, i.e. *any user adjustment will trigger real-time display updates for any affected elements in the workspace*. Note that not all the functions in OLIVER are conceptually new. In fact, functions like sorting or searching are available in almost every heat map tool, and functions like amplification and color curve adjustment are common in many photo-editing tools such as Adobe Photoshop and GIMP. But OLIVER is the first tool to bring all these into one platform for visual data analysis. Furthermore, to our knowledge, the key of the heat map visualization is not to appropriately display a heat map, it is instead to *instantly display a heat map when adjustments (such as sorting, clustering, merging, splitting, re-coloring) are applied*. This enables the trial-and-error (or "what if?") strategy in the data analysis workflow. It effectively increases the incidence of new discoveries (also called exploratory research or even serendipity) [35]. For example, the user may notice variations in a region that was assumed to be homogeneous as they click-and-drag the color scheme editor, leading to new discoveries on developmental regulation [7].

In summary, OLIVER is a visual data analysis tool specially designed for big longitudinal phenomics data with a unique feature that any user adjustment will trigger real-time display updates, allowing biologists to quickly integrate and compare phenomics data resulting from large-scale phenotyping experiments, and genomics information in public databases.

II. Background

In the last decade, with the rapid development of data-driven techniques, many data visualization tools have been developed to visually recognize patterns and trends in big data and to facilitate statistical analysis and data mining [15], [34], [41]. Among all the data visualization techniques, a particularly useful tool is the heat map, which is a graphical representation of data making full use of color to extend perception into multiple dimensions by allowing properties to be mapped both spatially and as color gradients [48]. Specifically, the 2D heat

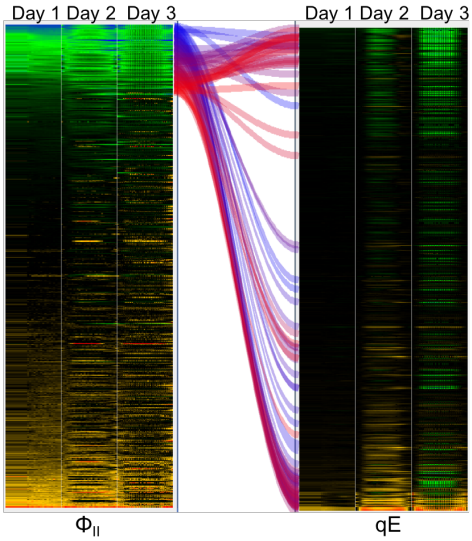

Fig. 3. For a 3-day plant phenotyping experiment, two heat maps of photosynthetic phenotype parameters (i.e. the photosynthetic rate of photosystems II Φ_{II} and the pH- or energy-dependent component of non-photochemical quenching qE) have their rows linked based on the matching row labels. A row-based selection is made based on the row order of the left heat map, and bands are automatically drawn between the selected rows and the corresponding rows in the other heat map.

map visualization allows researchers to quickly grasp the state and impact of a large number of variables at one time [36]. 2D heat map tools have been widely used in visualizing data in gridded formats, such as genome-wide gene expression or genetic mutations [29], [31].

While there are many available heat map programs for genomics data visualization [1], [2], [8], [11], [18], [38], to our knowledge, tools focused on large-scale phenomics data visualization are still absent. Phenomics data, in contrast to genomics data, are collected from numerous sources, represent vastly different phenomena, and are time and condition dependent [22]. In general, the rows and columns of a phenomics dataset are the individuals to test and the time points the individuals were phenotyped, respectively. Phenotype values change dynamically with environments or stimuli whereas the genome is constant, making phenotyping data highly complex and condition-dependent [7]. The relationships among testing objects (such as mutant lines), environmental conditions (such as light intensity and temperature), and the behavior of organisms are mutually interacting so that effects may be seen only under certain sets of conditions and genetic backgrounds [26].

In this article, we introduce OLIVER, which is designed to visualize and analyze complex phenotypic datasets in ways that reveal emerging phenomena and trends and thus to generate and test biological hypotheses and communicate results visually. OLIVER was developed with the general purpose of extracting knowledge from large-scale phenomics data. By developing the software in intimate collaboration between experts in biology, bioinformatics, and computer science, we have demonstrated that OLIVER is able to identify the insights of plant photosynthetic phenomics data [7].

III. Architecture

By integrating instant data visualization and hypothesis testing, we have developed a phenomics visual data analysis tool called **OLIVER**. OLIVER is the acronym of "Observe, Link, Integrate, Verify, Explore, and Reveal", which are the six critical processes of visual knowledge discovery. The workflow of OLIVER is shown in Figure 2.

A. OBSERVE Emerging Phenomena in Phenomics Data

The 2D heat map, the main view of OLIVER, is designed to instantly visualize large-scale time-resolved phenomics data and to assistant users to optimize data visualization. The core display functions are aimed towards presenting one or more entire phenomics data sets in the same workspace, providing essential operations for both broad and focused analysis. A typical longitudinal phenomics dataset can be represented as a 2D matrix, where each row is a plant to phenotype and each column is a time point at which every plant is phenotyped.

In very long or large-scale phenomics experiments, such as to measure the growth rates of plants in their whole life span or to phenotype thousands of plants at the same time, the number of columns/rows could be even higher than the number of pixels in the width/height of the heat map display. To render large heat maps at the real-time, loss-less heat map snapshots are periodically generated in the memory. Computer graphics card(s) are used for rapid scaling/stretching these snapshots to any size to facilitate heat map visualization. When adjustments on the colors or the order of rows and columns are made by the users, real-time previews can be generated by sampling only the heat map cells that are visible on the screen.

In the main view, rows of a heat map can be sorted or clustered according to the columns that users manually or automatically select. In the sorting dialog, the unsorted and sorted heat maps are displayed side-by-side for comparison, and the sorted heat map will update in real-time as users click-and-drag to select the columns to sort (see Figure 6). A heat map can be instantly re-colored using an interactive color curve interface (similar to the curve dialog in Photoshop). In a heat map, appropriate coloring allows users to better sort or cluster the rows and the columns and to maximally differentiate phenotype values. With OLIVER's interactive value-to-color mapping interface (see Figure 5), new color schemes can be rapidly developed and saved for future use. The purpose of re-coloring is to allow users to precisely control the value-to-color mapping, which is particularly useful to capture extreme values in non-normal distributions. In the re-coloring interface, color adjustments can be triggered using intuitive click-and-drag gestures. As one of the central features of OLIVER, any adjustment will trigger real-time display updates for any affected elements in the workspace.

Faithfully visualizing rows and columns in a heat map is imperative to gaining appropriate insights from sophisticated phenomics data. A unique property of phenotyping experiments is that to obtain the data with the best quality, the phenotyping sampling rate often varies within the same experiment. Accordingly, the width of each column in the

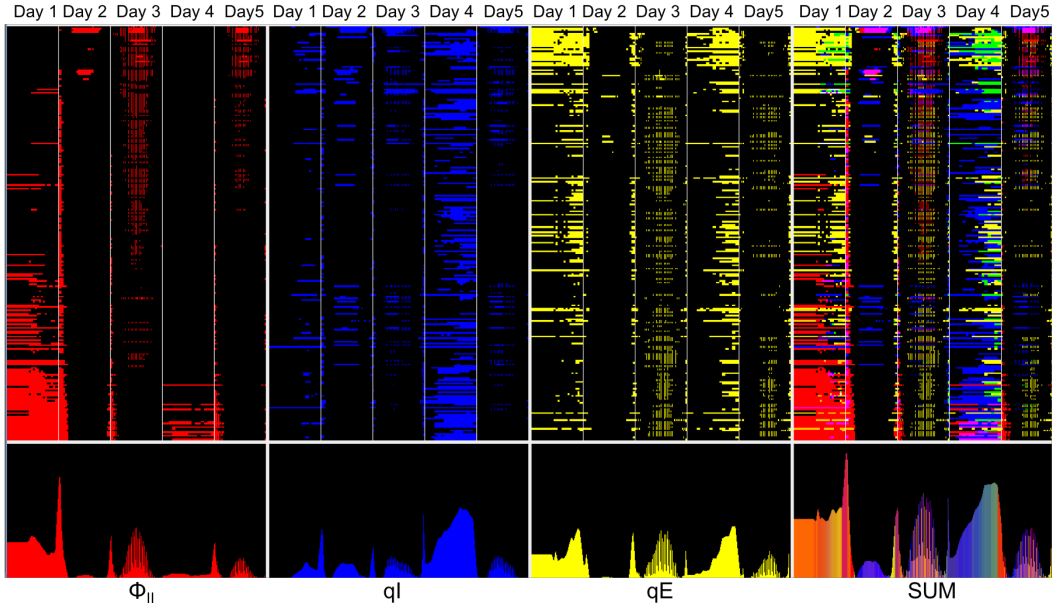

Fig. 4. For a 5-day plant phenotyping experiment, the binarized heat maps on the left represent three different photosynthetic phenotype parameters (Φ_{II}, qI, and qE). They were created by applying thresholds to the original heat maps. The heat map on the right is a combined visualization created by assigning logical colors to every possible combination of the three corresponding values on the left. On the bottom row, totals have been accumulated for each column to give a summarized view of expression-over-time for each phenotype parameter, as well as for the combined result. Users can adjust the binarization thresholds and instantly view the changes in this display.

rendered heat map of OLIVER is set to be proportional to the phenotyping sampling rate. The OLIVER workspace can also display multiple heat maps simultaneously, allowing them to be independently rendered, ordered, and related to each other via common features. OLIVER also allows users to visualize heat maps at multiple levels of data granularity ranging from the averaged phenotypes of every cluster to the sample groups measured using different devices, to the view of individuals without losing any details, thus facilitating investigations at different biological aspects (see Figure 1a, red and blue boxes).

B. LINK Metadata from External Sources

OLIVER can connect the input phenomics data with a range of online databases via external links to facilitate a better understanding of the phenomics data. As shown in Figure 1d fourth column, external links to the corresponding genes/proteins in the TAIR database [24] provide the most recently updated gene structure of Arabidopsis thaliana along with gene product, gene expression, DNA and seed stocks, genome maps, genetic and physical markers, and publications, which is particularly useful for Arabidopsis phenotyping experiments. Besides TAIR, the OLIVER API allows for integration with a variety of existing databases.

The other columns in Figure 1d show results generated based on the lookup-tables customized by users. Inspired by the modern object-relational database model [40], all the look-up tables in OLIVER are designed to be accessed in a chained fashion, i.e. the result of the look-up table loaded first becomes the input of the second one and so on.

The external data (such as sub-cellular localization in Gene Ontology), once being linked, can be represented visually as extra columns to a heat map (see Figure 8). Extra columns are typically rendered independent to the original heat map and are given a unique color scheme to represent linked results that are either qualitative (e.g. the experimenter who generated the data for each row), or quantitative (e.g. the evolutionary rate of a protein for each row). Extra data can be saved in the same data file containing the input phenomics data for easier data management.

C. INTEGRATE Results for Hypothesis Generation

The power of phenomics data analysis comes from the integration of multiple kinds of phenotype parameters (such as the photosynthetic rate of photosystems II Φ_{II} and the pH- or energy-dependent component of non-photochemical quenching qE, see Figure 3). This is because while a single phenotyping parameter describes a general effect, checking multiple phenotype parameters will lead to the identification of specific pathways and genes that are key to phenotype regulation. Thus phenomics data is inherently many-dimensional and difficult for humans to perceive.

OLIVER allows users to statistically test interactions among multiple heat maps, revealing trends in multi-dimensional phenomics datasets. In OLIVER, the data for each phenotype parameter can be visualized as a binary heat map (see Figure 4, left three heat maps). Then, the multiple binary heat maps can be combined into a single one, which is colored by assigning logical colors to every possible combination of the corresponding values in binary heat maps, as shown in Figure 4, the right heat map. In the integration process, any operation (e.g. changing the percentages to mitigate the rate of loss-of-detail caused by binarization) triggers real-time display updates, allowing for rapid exploration of hypotheses considering multiple phenotype parameter combinations.

The integration function enables the discovery of new functional relationships between phenotype parameters. Figure 3

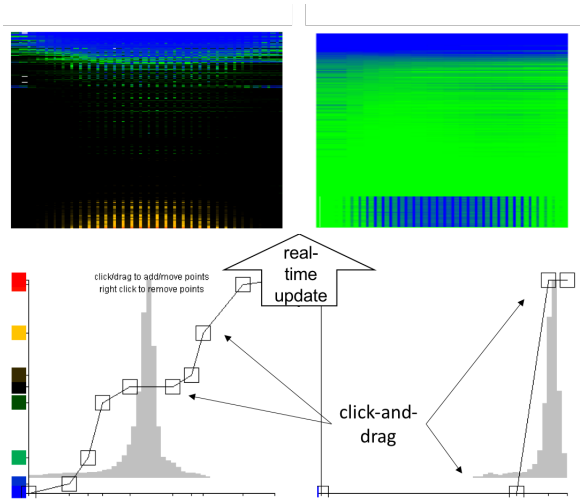

Fig. 5. Color curve interface (bottom row) and the associated heat maps (top row). The right side features data with a single-tailed distribution, while the left side contains two tails. This interface features both manual and automated options for assigning an appropriate color scheme to a heat map.

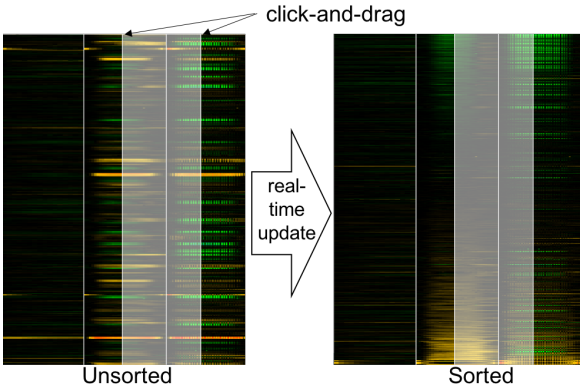

Fig. 6. Sorting Dialog. Both the unsorted (left) and sorted (right) heat maps are displayed, where the heat map on the right updates in real-time as users click-and-drag to select the time-range to sort (highlighted in white).

demonstrates one such case using an intuitive visualization. In the figure, a set of individuals with a single extreme phenotype value under one phenotype parameter (left) are linked to a second phenotype parameter (right), where the corresponding individuals are clustered on two different extremes.

D. Rapid Statistical VERIFICATION of Hypotheses

By conducting on-the-fly statistical tests on phenomics data sets, OLIVER allows users to verify whether a hypothesis is statistically sound or not directly on the visualized results. To provide statistic tests, OLIVER takes advantage of R and its libraries in statistics. For example, OLIVER performs the Kolmogorov-Smirnov test [25] to check whether two input data have significantly different distributions. The result shown in Figure 7 (bottom-right) was generated by calling R function $ks.test$ in the background [37]. In OLIVER, any modification that would affect the statistical tests (e.g. by changing the row or column of selection) will result in an instant update in the statistical test.

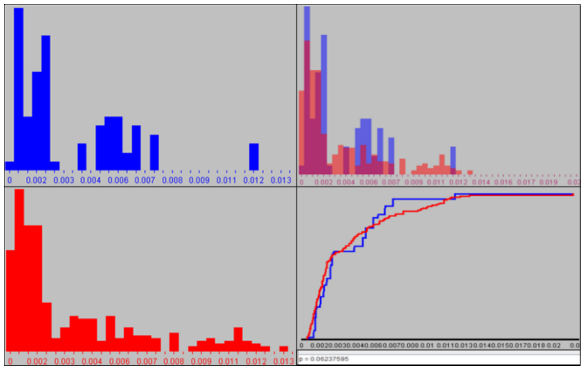

Fig. 7. Color-coded frequency distributions of two heat maps (top and bottom-left), which are overlaid (top-right) and are compared using Kolmogorov-Smirnov test (bottom-right).

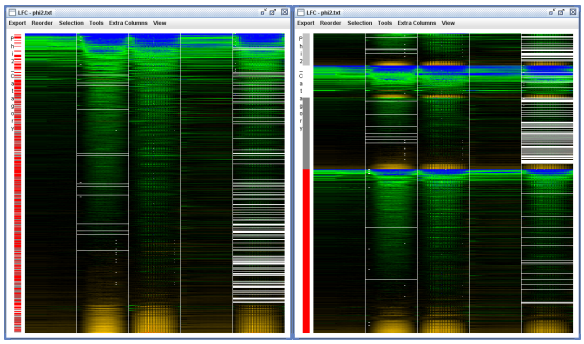

Fig. 8. A heat map is shown on the left with rows ordered according to the averaged values, and one additional column containing categorical information pertaining to the individual row entries. On the right, the same heat map has been reordered such that the extra column categories take precedence in sorting. This allows users to explore the relationships between phenotype measurements and external information.

E. Visually EXPLORE Multiple Hypotheses

The seamless integration of all the modules in OLIVER allows users to test and compare multiple "what if" hypotheses. For instance, users can explore phenotype relationships by adjusting the color scheme of a heat map (Figure 1b) to highlight only values far from the control, and then reorder the rows (Figure 6) based on each row's variance in a given temporal range. The resulting heat map may identify mutant lines with diverse responses over the course of specific stimuli, leading to insights about genes that affect the given phenotypes. Furthermore, by taking advantage of heuristic options (such as preset color maps, automatic sorting, and scripting), this data processing workflow can be completed within a few minutes. This greatly increases the rate at which researchers can test their hypotheses.

OLIVER provides enough flexibility towards the design of workflow so that users are free to explore data at their own pace. OLIVER has a variety of functions allowing users to gain new insights by exploring the results of every step, or it can run all the steps and show the final results. Fast-paced users will reach end-goal visualizations as quickly as possible, while exploitative users will be provided with fully interactive in-depth visualizations for each step in an open-ended workflow.

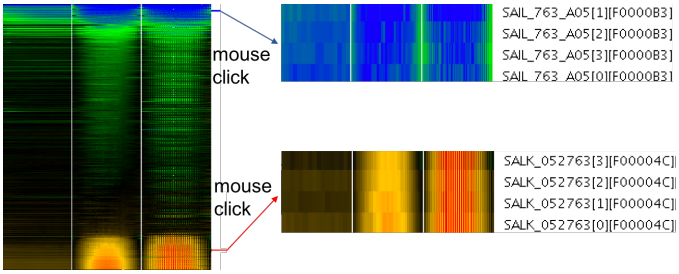

Fig. 9. A phenotype-over-time heat map, where the values in the heat map are the averaged values of multiple biological replicates of each genotype. The right heat maps show the raw data of a genotype before averaging.

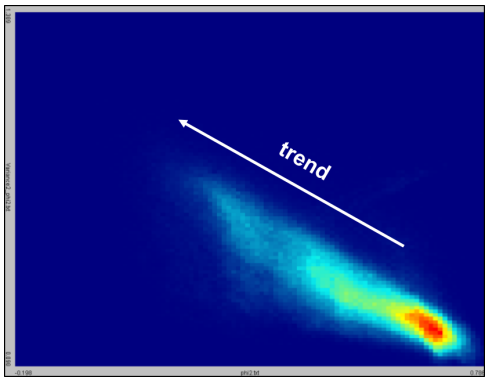

Fig. 10. Two phenotype parameters visualized in a 2D histogram. The 2D histogram is a heat map where each cell is colored based on the number of data points that fall into the corresponding bin. An animation is built showing how the relationship between the two phenotype parameters changes over time.

F. REVEAL New Knowledge by Sharing Complex Results

OLIVER provides a web-based module powered by Java applet to distribute its workspace in case-specific configurations so that multiple users can collaboratively view and explore the same phenomics data. For instance, a user can create and share a workspace of OLIVER with a group of viewers. Then the viewers would be presented with a workspace populated with the same data using the same settings. In addition, the results of OLIVER can be saved in graphical forms that are readily interpretable by users, as well as files in the form of publishable graphics, spreadsheets, or the internal binary file types that OLIVER can reload. All of these files can be easily shared among users.

G. Software Implementation

OLIVER was developed using Netbeans IDE 8.0, and compiled using JDK 1.8. Java Swing API was used exclusively for all display functions. OLIVER uses R and its statistics libraries [37] to provide a wide variety of statistical analysis. For plant phenotyping, OLIVER includes look-up tables that can link Arabidopsis mutant lines to the corresponding genes in the TAIR web site [24]. Here we introduce the implementation details of three key functions in OLIVER.

1) Sorting and Clustering: Sorting and clustering are basic functions in OLIVER that provide users with accessible options for re-ordering and grouping rows of a heat map. Users can sort the heat map based on phenotype value sums, standard deviations, absolute values, *etc*. Users can also group

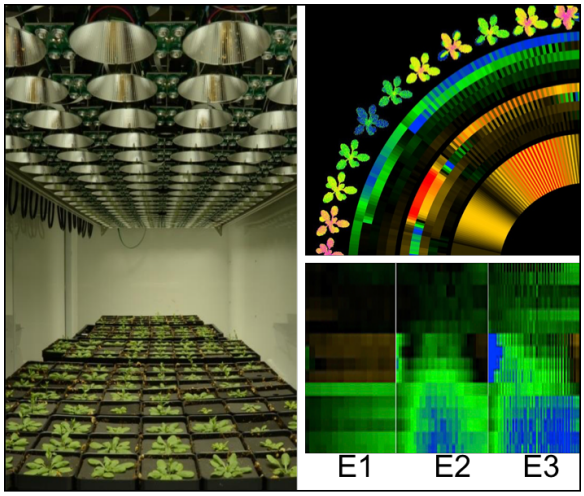

Fig. 11. A use case of OLIVER for large-scale plant photosynthetic performance phenotyping [7]. (Left panel) the plant photosynthetic performance of *Arabidopsis thaliana* mutant lines was monitored in a DEPI chamber under three kinds of ramped environmental perturbations (i.e. $E1$, $E2$, and $E3$). (Bottom-right panel) a quick visualization using OLIVER indicates that the emerging phenomena in the Arabidopsis mutants are highly transient and heterogeneous, depending in complex ways on both environmental conditions and plant developmental age. (Top-right panel) a circular heat map with false-color plant images generated using OLIVER, indicating that such emerging phenomena appear in different leaves under different conditions.

the rows using hierarchical or k-means clustering. If a look-up table is provided, sorting/clustering can be performed on the extra information in the look-up table in addition to the phenotype values in the heat map. A unique feature is that users can sort/cluster multiple linked heat maps synchronously, and conduct statistical tests on them as if they work on a single heat map. This allows users to explore the relationships among multiple phenotype parameters. Users can also propagate the row ordering of a heat map to other heat maps. Another unique feature is that users can preview sorting/clustering results before actually running the function in the entire data set (see Figure 6). A preview is generated and updated in real-time as the column/row selection is tweaked and the sorting/clustering method is chosen. Users can instantly see the individuals being sorted or grouped and the results of statistical tests when they specify sorting/clustering/testing parameters.

2) Zoomed View: While an input phenomics data set can be displayed on a global level, the nature of phenomics data analysis demands to visualize multiple levels of granularity simultaneously. To this end, an interactive "zoomed view" function has been implemented. Consider a phenomics data set containing multiple biological replicates for each genotype. The input data can be consolidated to an *averaged heat map*, where each row contains the averaged phenotype values for each genotype in the input. A row selection can be made conveniently by clicking-and-dragging the mouse over the heat map. The selection will instantly result in a zoomed view of the values in the selected area at a finer level of granularity. With the more detailed information provided by OLIVER, users can visually check sample variance, experiment consistency, and distribution of raw data (see Figure 9). Phenotype data are

often extracted from images, and a measured phenotype value is often the average of many pixels. In this case, OLIVER supports yet another level of granularity, where a detailed image is presented for each cell in the heat map. For grey scale images, the same value-to-color mapping is adopted to create a false-color image (see Figure 11). This level of detail is helpful in understanding unexpected results or identifying errors in data collection.

3) Animation: Gaining insights from phenomics data typically requires visualizing and operating upon multiple heat maps. OLIVER can generate a 2D-histogram (see Figure 10) to visualize the temporal correlations between two heat maps using animation. The 2D-histogram itself is a heat map where each cell is a bin, where the coordinates of a bin represent the values in the first and second heat maps, and the color of a bin is proportional to the number of phenotype values in the bin. The animation of temporal correlation can be adjusted in real-time and the results can be exported using the graphics interchange format (GIF) showing how the relationship between two phenotypes changes over time. The entire workspace can also be captured in animation, in which a sliding-window selection of heat map rows is gradually shifted between frames. Any related frames visible in the workspace (e.g. other heat maps, plots, and histograms) are automatically updated for each frame and are included in the animation.

IV. Results and Discussion

Use cases in plant phenomics were collected to demonstrate the functionality of OLIVER. In this article, we report a typical use case with large-scale phenomics data and present more use cases in the OLIVER website.

Figure 11 shows a use case of OLIVER for large-scale plant photosynthetic performance phenotyping [7]. The left panel shows that the plant photosynthetic performance of more than 100 *Arabidopsis thaliana* mutant lines was monitored in a DEPI chamber under ramped light intensity perturbations. A quick visualization of the phenomics data using OLIVER (shown in the bottom-right panel) indicates that the emerging phenomena in the Arabidopsis mutant lines are highly transient and heterogeneous, depending in complex ways on both environmental conditions and plant developmental age. Furthermore, the top-right panel shows a circular heat map with false-color plant images generated using OLIVER, indicating that such emerging phenomena appear in different leaves under different conditions. These emerging phenomena appear to be caused by a range of phenomena, suggesting that such previously unseen processes are critical for plant responses to dynamic environments [7].

V. Conclusion

High-throughput phenotyping has revolutionized biological sciences. The power of new phenotyping techniques lies in the fact that they enable the large-scale high-resolution observation of phenotypes, which is key to understanding the molecular mechanisms underlying gene functions in phenotype-related pathways. As a consequence, the quantity of phenomics data has been expanding exponentially in the last few years.

Analyzing longitudinal phenomics data requires approaches and tools that are distinct from those used to analyze more static genomic data since phenotyping platforms can aggregate results from a wide range of different sensors, each with its own set of constraints, time resolution, and sensitivities. OLIVER is a visual phenomics data analysis tool that integrates data visualization, statistical tests, and real-time data operation, providing researchers an interactive tool, aimed at establishing a comprehensive integration of thoughts, knowledge, and data emerge from the interactive activities.

The two key differences between OLIVER and the existing heat map tools are that 1) any user adjustment will trigger real-time display updates for any affected elements in the workspace, and 2) multiple heat maps can be synchronously adjusted, operated and connected. These allow scientists to instantly compare multiple phenomics data sets resulting from the same large-scale experiments, or correlate them with genomics information in public databases, in an interactive visualization environment through heat map views. By creating unique images and diagrams to effectively convey messages to scientists, OLIVER will assist them to quickly integrate, visualize, and compare large phenomics data, thus enabling new biological discoveries. Also, OLIVER can be extended to visualize data from other domains with similar data types. To this end, the OLIVER API has been implemented allowing for data integration with a variety of domain-specific databases.

VI. Software and Data availability

The OLIVER program, sample data and the user's manual are available under the GNU General Public License at the OLIVER website https://caapp-msu.bitbucket.io/projects/oliver/. The latest version of Java should be installed before running OLIVER.

References

[1] Ethan Cerami, Jianjiong Gao, Ugur Dogrusoz, Benjamin E Gross, Selcuk Onur Sumer, Bülent Arman Aksoy, Anders Jacobsen, Caitlin J Byrne, Michael L Heuer, Erik Larsson, et al. The cbio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov*, 2(5):401–404, 2012.

[2] Hao Chen and Xiangfeng Wang. Crusview: A java-based visualization platform for comparative genomics analyses in brassicaceae species. *Plant Physiol*, 163(1):354–362, 2013.

[3] T Andrew Clayton, John C Lindon, Olivier Cloarec, Henrik Antti, Claude Charuel, Gilles Hanton, Jean-Pierre Provost, Jean-Loïc Le Net, David Baker, Rosalind J Walley, et al. Pharmaco-metabonomic phenotyping and personalized drug treatment. *Nature*, 440(7087):1073–1077, 2006.

[4] 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. *Theor Appl Genet*, 126(4):867–887, 2013.

[5] Jacqueline N Crawley. Behavioral phenotyping of transgenic and knockout mice: experimental design and evaluation of general health, sensory functions, motor abilities, and specific behavioral tests. *Brain Res*, 835(1):18–26, 1999.

[6] Jacqueline N Crawley. *What's wrong with my mouse: behavioral phenotyping of transgenic and knockout mice*. John Wiley & Sons, 2007.

[7] 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.

[8] Marc Fiume, Eric JM Smith, Andrew Brook, Dario Strbenac, Brian Turner, Aziz M Mezlini, Mark D Robinson, Shoshana J Wodak, and Michael Brudno. Savant genome browser 2: visualization and analysis for population-scale genomics. *Nucleic Acids Res*, 40(W1):W615–W621, 2012.

[9] William J Frawley, Gregory Piatetsky-Shapiro, and Christopher J Matheus. Knowledge discovery in databases: An overview. *AI Mag*, 13(3):57, 1992.

[10] Robert T Furbank and Mark Tester. Phenomics–technologies to relieve the phenotyping bottleneck. *Trends Plant Sci*, 16(12):635–644, 2011.

[11] David M Goodstein, Shengqiang Shu, Russell Howson, Rochak Neupane, Richard D Hayes, Joni Fazo, Therese Mitros, William Dirks, Uffe Hellsten, Nicholas Putnam, et al. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res*, 40(D1):D1178–D1186, 2012.

[12] D Großkinsky, J Svensgaard, S Christensen, and T Roitsch. Plant phenomics and the need for physiological phenotyping across scales to narrow the genotype-to-phenotype knowledge gap. *J Exp Bot*, 66(18):5429–40, 2015.

[13] Angela M Hancock, Benjamin Brachi, Nathalie Faure, Matthew W Horton, Lucien B Jarymowycz, F Gianluca Sperone, Chris Toomajian, Fabrice Roux, and Joy Bergelson. Adaptation to climate across the arabidopsis thaliana genome. *Science*, 334(6052):83–86, 2011.

[14] Geoffrey Holmes, Andrew Donkin, and Ian H Witten. Weka: A machine learning workbench. In *Intelligent Information Systems, 1994. Proceedings of the 1994 Second Australian and New Zealand Conference on*, pages 357–361. IEEE, 1994.

[15] Andreas Holzinger. Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Informatics*, 3(2):119–131, 2016.

[16] Andreas Holzinger and Igor Jurisica. Knowledge discovery and data mining in biomedical informatics: The future is in integrative, interactive machine learning solutions. In *Interactive knowledge discovery and data mining in biomedical informatics*, pages 1–18. Springer, 2014.

[17] Jan F Humplík, Dušan Lazár, Alexandra Husičková, and Lukáš Spíchal. Automated phenotyping of plant shoots using imaging methods for analysis of plant stress responses–a review. *Plant Methods*, 11(1):29, 2015.

[18] Günter Jäger, Alexander Peltzer, and Kay Nieselt. inphap: Interactive visualization of genotype and phased haplotype data. *BMC Bioinformatics*, 15(1):200, 2014.

[19] Daniel A Keim, Florian Mansmann, Jörn Schneidewind, and Hartmut Ziegler. Challenges in visual data analysis. In *Information Visualization. IV 2006. Tenth International Conference on*, pages 9–16. IEEE, 2006.

[20] Boris Kovalerchuk. *Visual knowledge discovery and machine learning*. Springer, 2018.

[21] D Kramer and J Evans. The importance of energy balance in improving photosynthetic productivity. *Plant Physiol*, 155(1):70–8, 2011.

[22] D.M. Kramer, C. Abernathy, J. Carpenter, J. Cruz, R. Zegarac, and B. Lucker. Photobioreactor/sensor array (pbsa) for high throughput screening of algal strains and growth conditions. *Patent*, 2011.

[23] M Kutsukake, X Meng, N Katayama, et al. An insect-induced novel plant phenotype for sustaining social life in a closed system. *Nat Comm*, 3:1187, 2012.

[24] Philippe Lamesch, Tanya Z Berardini, Donghui Li, David Swarbreck, Christopher Wilks, Rajkumar Sasidharan, Robert Muller, Kate Dreher, Debbie L Alexander, Margarita Garcia-Hernandez, et al. The arabidopsis information resource (tair): improved gene annotation and new tools. *Nucleic Acids Res*, 40(D1):D1202–D1210, 2011.

[25] Hubert W Lilliefors. On the kolmogorov-smirnov test for normality with mean and variance unknown. *J Am Stat Assoc*, 62(318):399–402, 1967.

[26] Stephen B Manuck and Jeanne M McCaffery. Gene-environment interaction. *Annua Rev Psychol*, 65:41–70, 2014.

[27] Elaine R Mardis. The impact of next-generation sequencing technology on genetics. *Trends Genet*, 24(3):133–141, 2008.

[28] Elaine R Mardis. A decade/'s perspective on dna sequencing technology. *Nature*, 470(7333):198–203, 2011.

[29] Laetitia Marisa, Aurélien de Reyniès, Alex Duval, Janick Selves, Marie Pierre Gaub, Laure Vescovo, Marie-Christine Etienne-Grimaldi, Renaud Schiappa, Dominique Guenot, Mira Ayadi, et al. Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. *PLoS Med*, 10(5):e1001453, 2013.

[30] Chandra H McAllister, Perrin H Beatty, and Allen G Good. Engineering nitrogen use efficient crop plants: the current status. *Plant Biotechnol J*, 10(9):1011–1025, 2012.

[31] Silvia Monticone, Namita G Hattangady, Koshiro Nishimoto, Franco Mantero, Beatrice Rubin, Maria Verena Cicala, Raffaele Pezzani, Richard J Auchus, Hans K Ghayee, Hirotaka Shibata, et al. Effect of kcnj5 mutations on gene expression in aldosterone-producing adenomas and adrenocortical cells. *J Clin Endocr Metab*, 97(8):E1567–E1572, 2012.

[32] Olena Morozova and Marco A Marra. Applications of next-generation sequencing technologies in functional genomics. *Genomics*, 92(5):255–264, 2008.

[33] Chandra Shekhar Pareek, Rafal Smoczynski, and Andrzej Tretyn. Sequencing technologies and genome sequencing. *Journal Appl Genet*, 52(4):413–435, 2011.

[34] Christian Perez-Llamas and Nuria Lopez-Bigas. Gitools: analysis and visualisation of genomic data using interactive heat-maps. *PLoS One*, 6(5):e19541, 2011.

[35] Eric F Pettersen, Thomas D Goddard, Conrad C Huang, Gregory S Couch, Daniel M Greenblatt, Elaine C Meng, and Thomas E Ferrin. Ucsf chimeraa visualization system for exploratory research and analysis. *J Comp Chem*, 25(13):1605–1612, 2004.

[36] Joachim D Pleil, Matthew A Stiegel, Michael C Madden, and Jon R Sobus. Heat map visualization of complex environmental and biomarker measurements. *Chemosphere*, 84(5):716–723, 2011.

[37] R Development Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2008. ISBN 3-900051-07-0.

[38] James T Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S Lander, Gad Getz, and Jill P Mesirov. Integrative genomics viewer. *Nat Biotech*, 29(1):24–26, 2011.

[39] U Roessner, L Willmitzer, and A Fernie. Metabolic profiling and biochemical phenotyping of plant systems. *Plant Cell Rep*, 21(3):189–196, 2002.

[40] James Rumbaugh, Michael Blaha, William Premerlani, Frederick Eddy, William E. Lorensen, et al. *Object-oriented modeling and design*, volume 199. Prentice-hall Englewood Cliffs, 1991.

[41] Jannis Stoppe, Oliver Keszocze, Maximilian Luenert, Robert Wille, and Rolf Drechsler. Bioviz: An interactive visualization engine for the design of digital microfluidic biochips. In *VLSI, 2017 IEEE Computer Society Annual Symposium on*, pages 170–175. IEEE, 2017.

[42] Joel Tellinghuisen. Nonlinear least-squares using microcomputer data analysis programs: Kaleidagraph in the physical chemistry teaching laboratory. *J Chem Edu*, 77(9):1233, 2000.

[43] Oliver L Tessmer, Yuhua Jiao, Jeffrey A Cruz, David M Kramer, and Jin Chen. Functional approach to high-throughput plant growth analysis. *BMC Sys Bio*, 7(Suppl 6):S17, 2013.

[44] Melanie Tory and Torsten Moller. Human factors in visualization research. *IEEE T Vis Comput Gr, volume=10, number=1, pages=72–84, year=2004, publisher=IEEE*.

[45] Günter P Wagner and Jianzhi Zhang. The pleiotropic structure of the genotype–phenotype map: the evolvability of complex organisms. *Nat Rev Genet*, 12(3):204–213, 2011.

[46] A Walter, F Liebisch, and A Hund. Plant phenotyping: from bean weighing to image analysis. *Plant Methods*, 11(1):14, 2015.

[47] Colin Ware. *Information visualization: perception for design*. Elsevier, 2012.

[48] Leland Wilkinson and Michael Friendly. The history of the cluster heat map. *Am Stat*, 63(2):179–184, 2009.

[49] M Wong, C Dong, V Andreev, M Arcos-Burgos, and J Licinio. Prediction of susceptibility to major depression by a model of interactions of multiple functional genetic variants and environmental factors. *Mol Psychiatr*, 17(6):624–33, 2012.