Detecting Divergent Subpopulations in Phenomics Data using Interesting Flares

Methun Kamruzzaman
Washington State University
Pullman, Washington
md.kamruzzaman@wsu.edu

Ananth Kalyanaraman
Washington State University
Pullman, Washington
ananth@wsu.edu

Bala Krishnamoorthy
Washington State University
Vancouver, Washington
kbala@wsu.edu

ABSTRACT

One of the grand challenges of modern biology is to understand how genotypes (G) and environments (E) interact to affect phenotypes (P), i.e., $G \times E \rightarrow P$. Phenomics is the emerging field that aims to study large and complex data sets encompassing combinations of {genotypes, environments, phenotypes} readings. A phenomenon of crucial interest in this context is that of *divergent subpopulations*, i.e., how certain subgroups of the population show differential behavior under different types of environmental conditions. We consider the fundamental task of identifying such "interesting" subpopulation-level behavior by analyzing high-dimensional phenomics data sets from a large and diverse population. However, delineation of such subpopulations is a challenging task due to the large size, high dimensionality, and complexity of phenomics data. We present a new framework to extract such subpopulation-level information from phenomics data. Our approach is based on principles from algebraic topology, a branch of mathematics that studies shapes and structure of data in a robust manner. In particular, our framework identifies and quantifies "flares", which are structural branching features in data that characterize divergent behavior of subpopulations, in an unsupervised manner. We present algorithms to detect and rank flares, and demonstrate the utility of the proposed framework on two real-world plant phenomics data sets.

CCS CONCEPTS

• **Applied computing → Bioinformatics**; **Systems biology**; • **Mathematics of computing → Paths and connectivity problems**;

KEYWORDS

Phenomics; topological data analysis; Mapper; interesting flares

ACM Reference Format:
Methun Kamruzzaman, Ananth Kalyanaraman, and Bala Krishnamoorthy. 2018. Detecting Divergent Subpopulations in Phenomics Data using Interesting Flares. In *ACM-BCB'18: 9th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, August 29-September 1, 2018, Washington, DC, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3233547.3233593

ACM-BCB'18, August 29-September 1, 2018, Washington, DC, USA

ACM ISBN 978-1-4503-5794-4/18/08…$15.00
https://doi.org/10.1145/3233547.3233593

1 INTRODUCTION

Advances in agricultural and biomedical sciences for the next decade are poised to be driven by the increasing availability of data generated by high-throughput technologies. In precision agriculture, in addition to genotypic data generated using DNA sequencing technologies, massive volumes of data are also being gathered from various field sensing technologies that measure crop phenotypes (e.g., plant height, pollen shed) alongside environmental variables (e.g., temperature, humidity, soil pH, etc.). In the field of medicine, a diverse range of data is being collected on patient genotypes along with a multitude of disease- and treatment-related phenotypes (e.g., drug efficacy, molecular biomarkers) as well as hospital or healthcare-related environmental variables (e.g., length of stay, hand hygiene practices).

Consequently, the new branch of *phenomics* [2, 6] has emerged, whose goal is to study data sets that contain combinations of {genotype, environment, phenotype} observations. Phenomics data are inherently high-dimensional, rich in variety, and also typically spatio-temporal (resulting from longitudinal applications of technologies). Figure 1 shows a simplified schematic view of phenomics data. Compartmentalized tools developed for genomic application silos or traditional genome-wide association (GWAS) tools that map genotypic alleles to phenotypic traits are no longer adequate to analyze modern phenomics data sets. Instead, tools that are able to process complex phenomics data and enable extraction of "interesting" features and questions are needed.

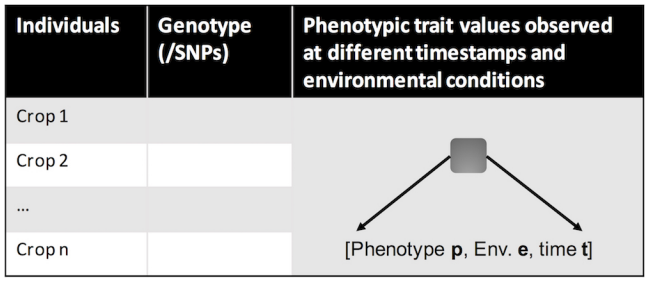

Figure 1: Schematic (table) view of a multi-dimensional plant phenomics data set.

One such fundamental question revolves around *subpopulations*. Given data from a large population, how do subpopulations (defined simply as arbitrary subsets of the original population) differ in the way they perform (or behave) in response to environmental variations? For instance, certain drug treatments or therapies tend to have a more pronounced effect on certain groups of individual

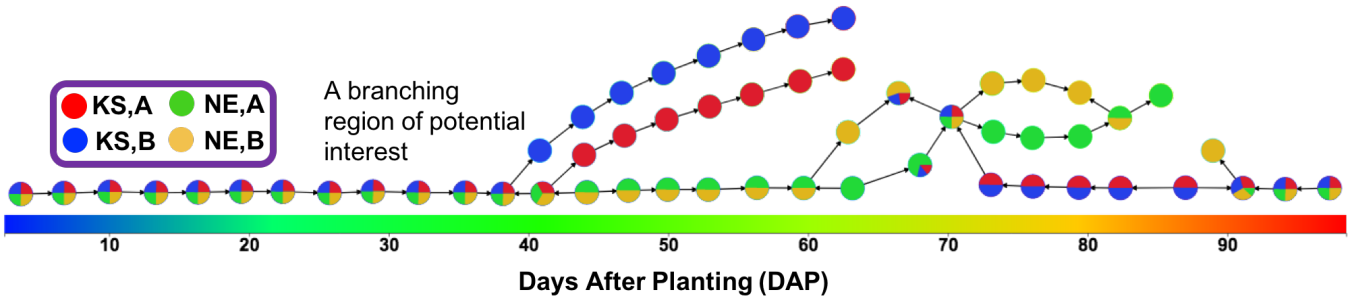

Figure 2: A sample output produced by our software suite Hyppo-X, showing a branching region of potential interest. The figure shows the topological object generated by our method from a real-world 400-point maize phenomics data set, with two different genotypes (A and B) from two different locations (KS: Kansas and NE: Nebraska). Various phenotypic traits and environmental conditions were measured over the course of the growth period which is about 100 days after planting (DAP). The horizontal color bar indicates the gradient of DAP (value increases from left to right). The topological object is rendered as a directed graph, where each node is a cluster of points distributed over time (DAP); and an edge from one node u to another node v implies that a) the two corresponding clusters share an intersection of plant individuals; and b) the mean phenotype (growth rate) increases from u to v. Further, each node is rendered as a pie-chart showing the distribution of individuals from the different <location,genotype> combinations. One of the interesting features is the branchings (divergence) in subpopulations visible between days 39 and 42—it is these types of features that we aim to detect as "flares".

patients than others, depending on the circumstances under which the treatments are administered. Similarly, certain subsets of crop genotypes show better resilience to harsher climactic conditions or environmental shifts than other groups. Identifying such subpopulations is potentially interesting to the domain scientist as it could directly aid in their ability to design targeted experiments for testing plausible hypotheses. However, standard statistical approaches that perform correlation tests between variables are typically not suited to highlight such subpopulation level variations.

In this paper, we propose a new approach for identifying and quantifying divergent subpopulations based on algebraic topology, a branch of mathematics that studies shapes and structures of data. More precisely, we build on our recent work and develop a new capability for detecting "flares", which are structural branching features that detect differentially behaving subpopulations from the data in an entirely unsupervised manner.

There are two major steps in our approach to detect flares:

- Given a large high-dimensional phenomics data set, we first use the *Mapper* framework [14] to build a compact topological (and visual) representation of the data. The Mapper framework has recently found increasing number of applications in the analysis of high dimensional data (see Section 2 for a summary). Section 3.1 presents an overview of the Mapper algorithm.
- Next, we define *flares* as branching features in the topological object, with certain properties (as defined in Section 3.2), and provide efficient algorithms to detect the flares and score them so that they can be ranked in non-increasing order of interest (see Section 3.2).

To the best of our knowledge, this is the first formal treatment given to the detection and quantification in terms of interestingness of flares.

All our implementations are available as part of our Hyppo-X software suite for analyzing phenomics data [9]. Figure 2 shows a sample output generated by our Hyppo-X suite from a real-world maize data set containing two genotypes from two different U.S. locations. The figure gives a glimpse of the branching regions that look potentially interesting from the point of view of delineating subpopulation-level variations (and are therefore the target of our flare detection procedure). Note that our method operates in an unsupervised manner, and the information about the various data sources (KS/NE, A/B) was applied to the topological representation only *after* the output was generated (for use in the interpretation step). Consequently, our method's automatic ability to separate the different subpopulations (KS from NE and A from B) at various stages of the topological representation is significant.

In Section 4, we present a detailed report on experimentally evaluating the new flare detection capability. We used two data sets: the {KS/NE, A/B} maize data set ($n = 400$) obtained from our collaborator (Schnable laboratory at Iowa State University); and a significantly larger, multi-year maize phenomics data set from the Genomes to Fields initiative [7], which has annual data (since 2014) collected from over 45,000 farming plots and involving more than 1,500 maize hybrid varieties, from across 23 states and provinces in the U.S. and Canada. Our experiments demonstrate the immense potential of our tool to identify and quantify divergent subpopulations.

2 RELATED WORK

The basic building block of our algorithmic framework is the Mapper algorithm. Originally proposed by Singh et al. [14], this approach has recently found increasing use in diverse application domains ranging from medicine [11–13, 16] to sports analytics [1, 11] to voting patterns [11]. At the same time, in most, if not all,

of the previous applications of Mapper, the interesting supopulations are characterized by features such as paths, loops, or flares in the topological object (also called Mapper) identified in a "visual" manner. Carrière et al. [3, 4] present a rigorous theoretical framework for 1-dimensional Mapper, where the features are identified as points in an extended persistence diagram. But they do not address the relative importance of the features in the context of the application generating the data.

In recent ongoing work [8, 10], we have proposed a notion of interestingness score of a path in the Mapper, and used it to quantify the interestingness of, as well as rank, the features identified as paths. In this paper, we consider quantifying the interestingness of branching phenomena in data sets identified as flares in the Mapper. As it turns out, flares present several nontrivial generalizations and new challenges as compared to interesting paths. Furthermore, the use of flares as a way to detect interesting subpopulations from within complex high-dimensional phenomics data sets represents a novel use-case.

3 METHODS

3.1 Overview of the Mapper algorithm

The Mapper algorithm [14] produces highly compressed visual representations of high-dimensional data that reveal significant structural aspects. Let X be the input space, which is typically a point cloud of data in $\mathbb{R}^d$ for large dimension d. The first component of Mapper is a *filter function*, which is a continuous function $f : X \to Z \subset \mathbb{R}$. More generally, the filter function could be high-dimensional, i.e., $f : X \to \mathbb{R}^h$ for $h \geq 2$. For instance, we could study temperature and precipitation together ($h = 2$). In this paper, we consider (one or more) single dimensional filters. We choose an *open cover* $\mathcal{U}$ that decomposes Z into overlapping open intervals, i.e., $\mathcal{U} = \{U_i\}_{i=1}^{r}$ such that $\cup_i U_i = Z$. This cover is typically characterized by the number of intervals r and an overlap percentage g, referred to as the *resolution* and *gain* of the cover, respectively. The resolution r determines the length of each interval U_i (assumed to be the same), and the gain g determines by how much adjacent intervals U_i and U_{i+1} overlap.

The main idea of Mapper is to *pullback* the cover $\mathcal{U}$ to a cover of X. Note that since f is continuous, $\{f^{-1}(U_i)\}$ for $U_i \in \mathcal{U}$ forms a cover of, i.e., tiles, X. In other words, for each interval U_i, the algorithm identifies subsets of points from X that have values of f in that interval. These points all have "similar" f values, and are grouped further into clusters by Mapper using a *distance function*. This clustering step is designed to reveal the covariance of, or relation between, the filter function f and one or more other variables of interest that are used to define the distance function.

Note that the clustering step is repeated for every interval U_i in the cover. As such, we refer to these clusters obtained from subsets of X as *partial clusters*. Various clustering algorithms could be used in this step; we use DBSCAN [5]. Each cluster produced in this step is represented by a node in the highly compact representation of X—also called the *Mapper*. In this context, we use the terms "node" and "cluster" interchangeably.

Given that the intervals in the cover of Z overlap (as defined by the gain parameter g), clusters resulting from overlapping intervals of the pullback cover could share data points. Such overlaps of membership between two clusters is represented by drawing an edge between the corresponding two nodes in the final object (and a triangle between three clusters that intersect, a tetrahedron between four intersecting clusters, and so on). These connections between clusters (as specified by the edges between the corresponding nodes) capture the covariation of f with other functions of interest across the range of values of the filter function.

Note that such covariations could be *nonlinear*, e.g., there could be two distinct paths of edges connecting disjoint series of clusters as one moves along a set of intervals in the cover of Z. Such disjoint paths capture subpopulations (i.e., subsets of points of X) displaying distinct behavior under "similar" conditions as captured by similar ranges of values of f. Of course, the distinct behavior would be explained by one or more of the other variables defined on X. Indeed, the key strength of the Mapper algorithm is its natural ability to identify such nontrivial "structure" while also capturing more straightforward linear relationships—all in a unified framework.

For example, consider the instance where X is a set of points in $\mathbb{R}^2$ sampled from a noisy unit circle (see Figure 3). We use the height of the points (i.e., their y-coordinate values) as the filter function. We consider a cover of Z, which is almost $[-1, 1]$, into $r = 3$ overlapping intervals, with adjacent intervals overlapping roughly by a third (i.e., $g = 33\%$). The pullback cover of X then has four pieces, with the subset of points with height in the middle interval forming two connected components. We then use the Euclidean distance between the points (in $\mathbb{R}^2$) as the distance function to cluster the points in each component using, e.g., single linkage clustering. Thus we get one node per component, which we color from blue to red according to the mean height of the points in each node. We also get four connecting edges capturing the overlap of the clusters. Note that we go from around 20 points in X to just four nodes and four edges in the Mapper. At the same time, this highly compact representation captures the underlying structure of X—the circle.

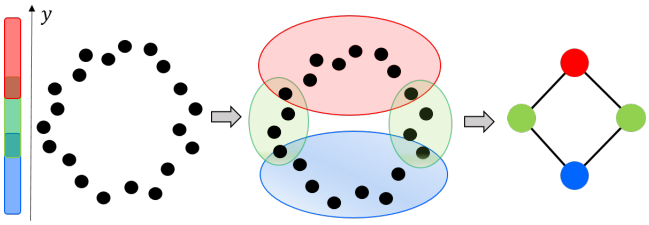

Figure 3: The Mapper algorithm applied to a set of points sampled from a noisy circle. We use the height of the points (y-coordinate) as the filter function. We consider a cover of $Z \approx [-1, 1]$ using $r = 3$ overlapping intervals, with adjacent intervals overlapping roughly by a third (i.e., $g = 33\%$). The final Mapper is shown on the right.

3.2 Flares

We propose a framework to detect and use "flares" (defined below) that characterize branching phenomena in phenomics data sets. We use the undirected graph (i.e., the nodes and edges) of the Mapper as described above; let this graph be denoted $G = (V, E)$. Recall that each node in V is a cluster of points from a single interval in

the filter function(s) *further* grouped together by the subsequent clustering step. Each edge in E represents a pair of clusters from adjacent intervals that share an intersection of points.

We first construct a directed graph from G. Given an edge $e = \{u, v\}$, we direct the edge by default from the cluster showing a lower phenotypic performance as measured by its mean phenotypic value to one with the higher mean phenotypic value. This scheme allows us to track a trail of clusters that show an improving trajectory in performance by a user-selected phenotypic trait (e.g., yield or plant height). In order to capture branching phenomena effectively, we modify this directing procedure by using mean phenotypic values of *subset of individuals belonging to shared genotypes* between nodes u and v—see Remark 3.5.

If the mean phenotypic values of u and v are equal, we choose one of the two directions arbitrarily. This procedure guarantees that we build a directed acyclic graph (DAG). More generally, we could consider a relaxed variant where we add both directed edges when nodes u and v have (nearly) identical means. This variant is left as future work, and we restrict our attention to DAGs in this paper. We now define a few more terms related to this DAG.

Definition 3.1. A *source* (*terminal*) node in a directed graph is one that has no incoming (outgoing, respectively) edges. A *branching* node in a directed graph is one that has at least two outgoing edges.

Note that, by the above definitions, a source code can also be potentially a branching node. Furthermore, we use the term *simple path* to refer to a path in the graph in which no node, with the possible exception of the sentinel nodes (beginning and ending) of the path, is a branching node.

We define a *stem* and a *branch* associated with a branching node as follows (see Figure 5 for an illustration).

Definition 3.2. Given a branching node u, a *stem* is a possibly empty simple path that ends in u.

Remark 3.3. Note that there can be multiple stems ending at a branching node u. There are two classes of such stems—those that are entirely non-overlapping (i.e., simple paths ending at u that are otherwise node-disjoint) and those that are nested (i.e., they originate from different starting nodes in the same parent simple path ending at u).

Definition 3.4. Given a branching node u, a *branch* refers to a non-empty path (simple or not) that originates at u.

Note that two branches originating at the same branching node can possibly intersect. Furthermore, there are at least two branches originating at a branching node (by definition of a branching node).

Remark 3.5. To capture branches in phenomics data sets accurately, we modify the way in which we direct the edges in the topological object as follows. Given an undirected edge $e = \{u, v\}$ in the Mapper, we direct edge e from the node with the lower mean phenotypic value to the one with high value, where the respective means are now taken over the subsets of individuals in u and v that belong to genotypes present in both nodes. This procedure is illustrated in Figure 4. In an alternative setting, we use *location* of the individuals to determine these subsets to take means over (see Figure 10 in Section 4.2). If genotype or location information is not available, these means are computed over only the individuals shared by nodes u and v.

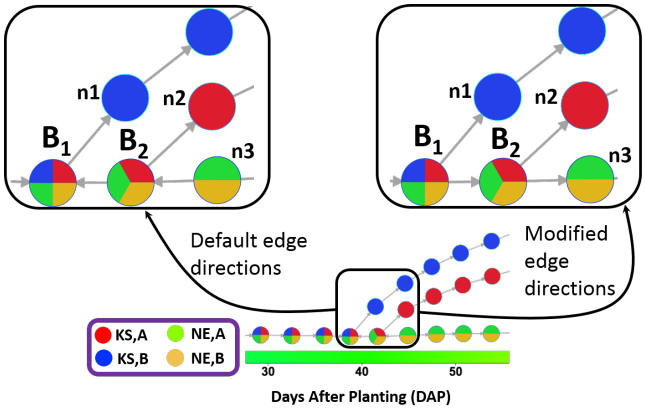

Figure 4: Modification of edge directions in the Mapper from Figure 2. Using the default approach (zoomed in on left), edges are oriented from B_2 to B_1 and from n_3 to B_2, by considering the mean phenotype values of all individuals in these nodes. Using the modified approach, we orient the edge from B_1 to B_2 by considering the mean phenotype value of individuals with only the genotypes shared by these nodes (i.e., (KS,A), (NE,A), and (NE,B)). A similar modification directs the edge from B_2 to n_3.

Let $B(u)$ denote the set of all branches originating at a branching node u and $S(u)$ denote the set of non-overlapping (i.e., non-nested) stems ending at u.

Definition 3.6. We define a *flare* to be a unique combination of a branching node u, a stem $s \in S(u)$, and a subset $B'(u) \subseteq B(u)$. Here, we do *not* enforce that a stem be non-empty, to allow detection of flares strictly originating at a given branching node. However, we *do* enforce that each branch selected is non-empty (i.e., has at least one edge) and that the subset selected $B'(u) \subseteq B(u)$ contains at least two or more branches (as illustrated in Figure 5).

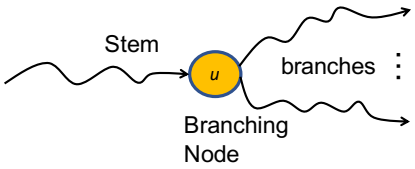

Figure 5: An illustration of a flare.

The selection of the stem and branches to include in a flare is computed deterministically as a function of the branching node. Intuitively, the idea is to examine the set of individuals "covered" by the branching node, and then "cast a net" in either direction, on all simple paths leading up to u (candidate stems) and on all the branches originating at u, as far as there is a non-empty intersection with the individual set of the branching node (see Figure 6).

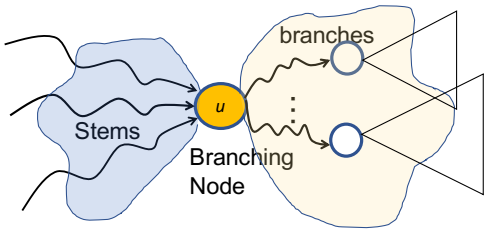

Figure 6: Conceptual illustration of how flares are constructed from a given branching node u. Stems are selected from the set of incoming simple paths, and branches are selected from the DAGs rooted at u. The boundaries of the selection are determined by "casting a net" on either side of u and including all "areas" where there is shared individual coverage.See text above for further details.

The rationale for this selection scheme is as follows. In an application such as phenomics, each "point" included in a cluster is typically a given plant crop ("individual") observed in certain time and space. Therefore, by the way we construct our topological object using intersections between adjacent clusters, the same individual may continue to appear in a sequence of clusters (i.e., in a path) on either side of a branching node. Therefore, by considering the set of individuals covered by a branching node, and examining how that set distributes itself across the branches, we can discover interesting subpopulation-level variations (or differences in the way they respond to various environmental filters). In a population where there is also a large genetic diversity, one can adapt the same procedure to include the set of genotypes covered (instead of plant individuals).

Construction of a flare: More formally, let $N(u)$ denote the set of individuals covered in the cluster corresponding to u. Then, we follow the trail of clusters in either direction to incrementally grow the corresponding stem or branch, as follows. For stem computation, we enumerate all the simple paths ending at u, and for each such simple path (candidate stem), we begin at the node v which is the immediate predecessor of u and compute $N(v) \cap N(u)$. If the intersection is non-empty then we include v in the current stem and iteratively walk to the next predecessor (until either the simple path terminates or the intersection becomes empty). Note that at each step, we compute the intersection with $N(u)$.

A similar procedure is carried out to enumerate all branches originating at u, walking forward instead of backward, with the caveat that we do not need to restrict the elongation process to only simple paths in the forward direction. In other words, if we encounter another branching node, the algorithm proceeds recursively, except that at every subsequent step going forward from the second branching node, the intersection is computed only relative to the original branching node u.

Note that the above procedure is deterministic, in that given a branching node, the reach of a flare involving that branching node is determined by the reach of the set of individuals in u on either side of u in the DAG. In fact, this procedure would also detect *all* the flares involving u. More precisely, the cross-product of $S(u)$ and $B'(u)$ (as specified in Definition 3.6) yields the set of all flares involving u.

Scoring a flare: In order to compare and relatively rank flares, we devise a simple scheme to score each flare. Given a flare f, we compute its "interestingness score" as follows.

First, we associate a weight to all edges. The weight of an edge is given by the absolute difference in the phenotypic performance (cluster means) between the two corresponding clusters. Intuitively, the larger the performance variation, the more interesting that edge is to a branch. Note that since we use the absolute value of the difference, all edge weights are positive.

We score the flare using its edge weights as follows (see Figure 7). Note that there is a unique subgraph induced by each flare and that subgraph also will be acyclic (as it is derived from a DAG). Therefore, we perform a simple bottom-up/post-order traversal of that induced DAG, starting at each terminal node and climbing up the parent and the ancestor levels. At each step, we perform a simple gather-scatter way to propagate the scores across levels. More specifically, at a node u, all the scores of its child branches are added ("gather"), and the value is then equally divided ("scatter") among its predecessor branches. The algorithm terminates when it reaches the main branching node u of this flare. Once scored, the flares can be rank ordered in the decreasing order of score and displayed.

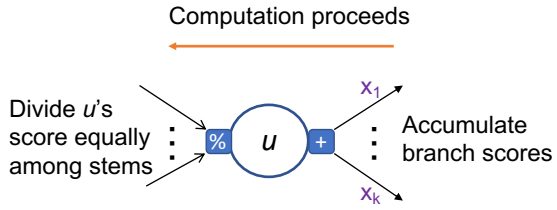

Figure 7: Illustration of how the interestingness score propagates through a flare.

Remark 3.7. In our current implementation, only branches contribute to the score of a flare at a branching node. Stems do not contribute, the rationale being that examining the branches typically suffices for explaining how a population, covered at the branching node, diverges. However, the procedure can be extended to include stem scores as needed. The information contained in the stem is still useful during our subsequent analysis and interpretation.

Remark 3.8. Our procedure for scoring flares takes running time linear in the size of the flare.

4 EXPERIMENTAL EVALUATION

We used two real-world maize data sets to test and evaluate the proposed algorithmic framework to detect flares. One is a small data set with 400 "points", whereas the other one is a significantly larger data set covering more than 300,000 "points". We describe the data sets along with the results of our application below.

4.1 Small-scale evaluation

Input data. This maize data set consists of phenotypic and environmental measurements for two genotypes (abbreviated here for simplicity as *A* and *B*), grown in two geographic locations (Nebraska (NE) and Kansas (KS)). The data consists of *daily* measurements of the genotypes' growth rate alongside multiple environmental variables, over the course of the entire growing season (100 days). For the purpose of our analysis we treat each unique [genotype, location, time] combination as a "point". Consequently, the above data set consists of $N = 400$ points. Here, "time" was measured in Days After Planting (DAP). An "individual" in this data set refers to a plant individual that represents a [genotype,location] combination.

Each point has one phenotypic value (growth rate) and 10 environmental variables (including but not limited to: humidity, temperature, rainfall, solar radiation, soil moisture, and soil temperature). We studied a multitude of these environmental variables; we present the results using humidity here, as it led to more interesting observations (compared to the other variables).

Flare analysis. We tested our framework using both single and two filter function(s).

Single filter function: First, we constructed our topological object using DAP as a single filter and used the difference in growth rates to calculate pairwise distances between points (in the clustering step). This study is aimed at understanding how the population of individuals (of both genotypes in both locations) show varying trends in phenotypic performance (i.e., growth rate here) as a function of time.

The resulting object along with the detected flares are shown in Figure 8, based on which we make the following observations.

(1) Until around DAP ∼40, all four subpopulations behave similarly (as shown by the common trail of clusters up to that point).
(2) Around DAP ∼40, two branches are evident: i) The first branching event occurs when the {KS,B} subpopulation separates from the rest due to a significantly accelerated growth spurt (compared to the rest). ii) The second subsequent branching event corresponds to the {KS,A} subpopulation separating from the rest. Figure 8(B) shows the cluster nodes colored by growth rate.
(3) It is not until DAP ∼70 that the Nebraska varieties show a separation in their behavior.

All the above branching events were successfully detected by our flare detection algorithm (shown by long arcs of different colors) in a runtime of 160 milliseconds after the Mapper graph is built. Note that our method is unsupervised and the information about the source genotypes and locations (pie-chart distribution in Figure 8(A)) was applied only *after* the analysis was completed, to aid in our interpretation. These results demonstrate our method's ability to successfully delineate interesting subpopulations that show divergent behavior.

Two filter functions: In the above single filter results, the fact that genotype *B* in Kansas shows significantly altered behavior compared to the same genotype in Nebraska indicates that there could be causal environmental factors at play that influence the phenotype. To better characterize such potential candidates for key environmental variables, we conduct a two filter study (one filter being time or DAP, and another filter being one of the many environmental variables recorded).

We present here the results for the combination {DAP, humidity}. Figure 9 shows the corresponding topological object. The time (DAP) grows from left to right and is shown in a horizontal gradient color bar. Based on this figure, we make the following observations:

(1) Figure 9(A) shows that at the initial growth period (1–10 DAP), the performance at both locations are highly comparable, as is evidenced by the clustering of both locations.
(2) Around DAP 11 the locations diverge into two separate branches (as shown in panel (A)). This separation is owing to the differences in their humidity conditions— more specifically, while Nebraska experienced steadily low humidity values until around DAP 50, Kansas experienced fluctuating and often high humid conditions for most of the period until around DAP 60 (see panel (C)). This period of high humidity fluctuation also coincides with the accelerated growth rate that Kansas experiences from around DAP 40. As for Nebraska, the increase in growth rates occur eventually around DAP 60 (panel (B)) and that too coincides with higher values in humidity (panel (C)).

Summary of findings: These results and observations suggest that humidity perhaps has a significant effect on growth rates, and that the effect was more pronounced on genotype (B) than for genotype (A). The precise time and humidity intervals where such effects manifest are shown by the flare.

4.2 Large-scale evaluation

Input data. The second data set we used in our experiments is obtained from the Genomes to Fields (G2F) initiative [7]. The complete data set contains data for 1,500 maize hybrid varieties cultivated across 23 states and provinces in the U.S. and Canada. For the purpose of of our analysis in this paper, we used a subset of this data which contains 894 genotypes, grown in four geographic locations: Texas (TX), Nebraska (NE), Missouri (MO) and Ontario, Canada (ON). The data consists of end-of-the-season measurements of multiple phenotypic values (Yield, Pollen DAP, Plant height, and Ear height). It also contains multiple environmental variables, over the course of the entire growing season (174 days). For the purpose of our analysis we treat each [field location, state, block, plot, time] combination as a "point". Consequently, the above data set consists of $N = 306,533$ points. Similar to the first data set, "time" was measured in the Days After Planting (DAP) here.

Each point has multiple phenotypic values (Yield, Pollen DAP, Plant height, and Ear height) and at least five environmental variables (including but not limited to: temperature, humidity, rainfall, solar radiation). We present our analysis using temperature as a single filter, while using all phenotypic values for performance (choice of distance function for clustering). For distance based clustering, we computed the L_2 phenotypic distance between pairs of points. The study is aimed at understanding how the population shows varying trends in phenotypic performance (i.e., Yield, Pollen DAP, Plant height, and Ear height here) as a function of temperature.

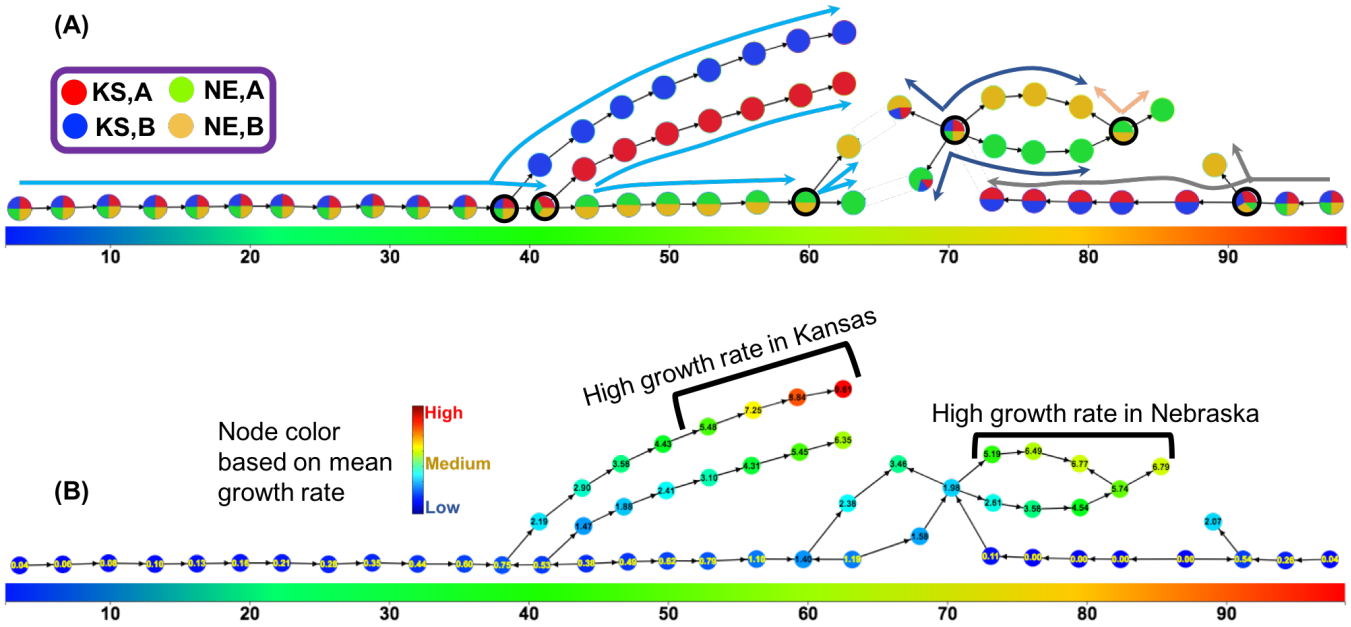

Figure 8: Topological object constructed using DAP as a single filter function (shown earlier in Figure 2), now also showing the interesting flares detected by our method. The horizontal color bar indicates the gradient of DAP, with the value increasing from left to right. (A) Each cluster (node) of the topological object is rendered as a pie-chart showing the distribution of their four classes of individuals. Long arcs of different colors show interesting flares, and the corresponding branching nodes are identified with bold border. The blue flare (long arc spanning DAP 1 through 60) was ranked as the top interesting flare. (B) Each cluster colored by its mean growth rate (phenotype), with branches showing active growth (high phenotype) marked.

Flare analysis. Figure 10 shows the topological object along with the detected flares. Based on the figure, we make the following observations.

(1) The green flare is mostly Texas dominated whereas the other two flares are mostly Ontario dominated.
(2) The contribution of Nebraska population is almost everywhere, is mixed with other sources, and cannot be separated based on phenotypic performance.
(3) The long green branches indicate that there has been a divergence in either performance level or environment level. Figure 11 clearly depicts the divergent phenotypic behavior.
(4) All the higher yield nodes contain either Texas or Missouri subpopulations.

Branch analysis. We analyzed both branches (B_1, B_2) in Figure 11, which are captured mostly by the green flare in Figure 10. Our findings are as follows:

(1) Branch B_1 contains 494 genotypes and branch B_2 contains only 6 genotypes. These 6 genotypes of branch B_2 are also represented within branch B_1, although they are grown in different farm plots. In addition, the plants of branch B_1 include plants grown in Lincoln, Nebraska, which is absent in branch B_2.
(2) Branch B_1 contains 15, 168 points whereas the branch B_2 contains only 133 points.
(3) The yield values for points in branch B_1 are significantly lower ([4.275, 272.958]) than those for the points in branch B_2 ([160.787, 274.381])—as can be seen from Figure 11(A).
(4) The temperature values, on the other hand, for both branches B_1 and B_2 are nearly comparable—as can be seen from Figure 11(B).
(5) We also tried other environmental variables as filters, but did not encounter another variable that behaved differently between the points of the two branches.

Taken together, the above observations imply the following. These two branches show divergent behavior and yet temperature cannot be the attributing factor. In fact, even genotypes are less likely to offer a reason as to why the yield performance is significantly different (low in B_1 and high in B_2). We note here that, even though the 6 genotypes in B_2 are observed in both branches, along branch B_1 the points corresponding to those genotypes are grown in different locations within the same state (i.e., different farm plots). This implies that the same genotype grown in very similar conditions of a state (say TX or MO) but in different farms could have widely divergent behavior. This could be because of differences in the agricultural practices, which are not necessarily captured in the environmental table (e.g., fertilizing practices, water/irrigation use, crop rotation) could be potential contributing factors to the observed difference. Further examination along this direction could provide potentially important insights that relate farming practices to crop performance.

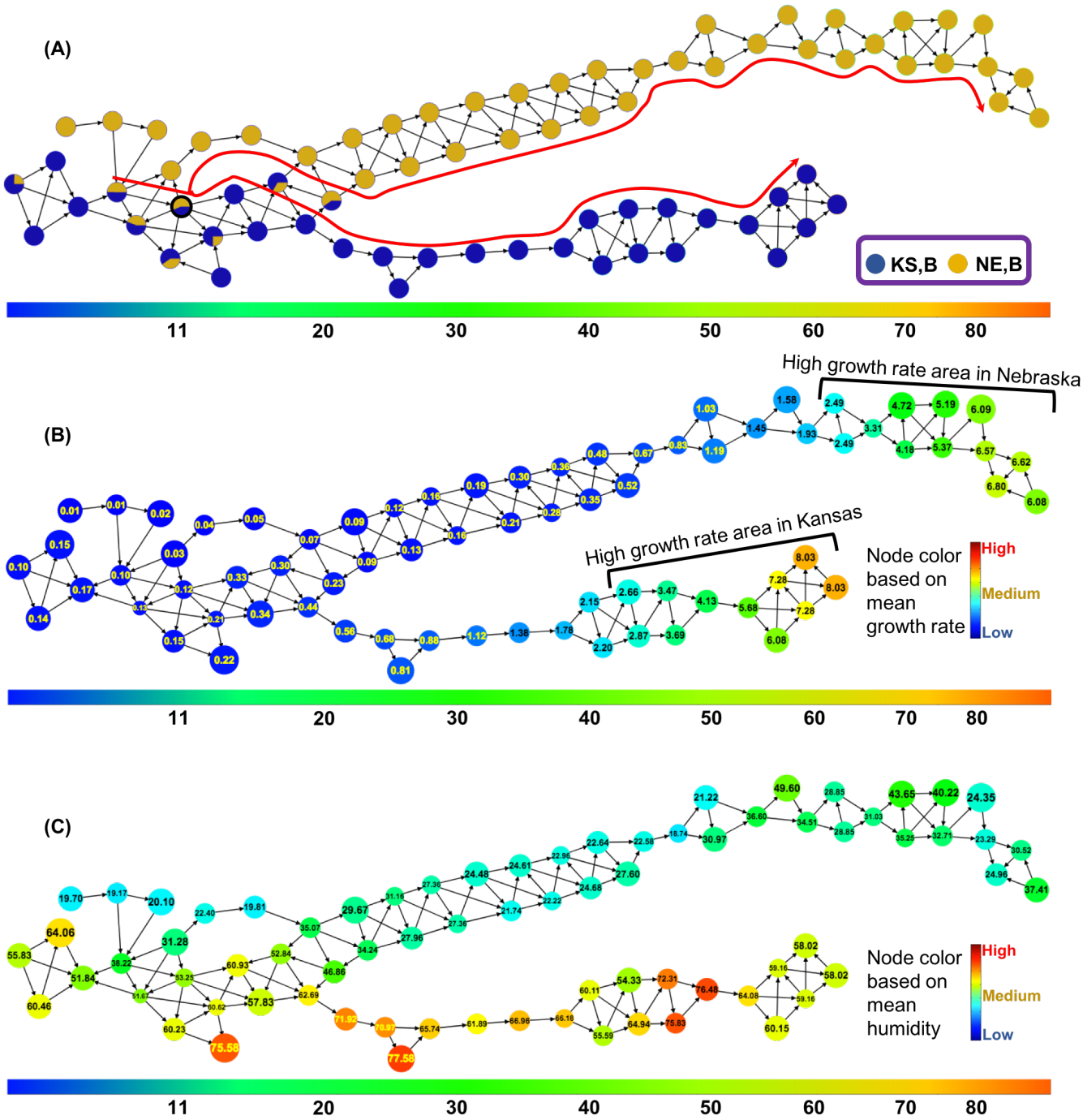

Figure 9: The topological object constructed using only the individuals of genotype B, using DAP and humidity as the two filter functions. The horizontal color bar indicates the gradient of DAP, with its value increasing from left to right. (A) Each cluster (node) is rendered as a pie-chart showing the distribution of its individuals from the two locations (KS and NE) for genotype B. Parts (B) and (C) show the same topological object, however with each cluster (node) colored by the growth rate (phenotype) and humidity (environment), respectively. Our method captured one large flare, which is indicated by the red branched arc in Part (A).

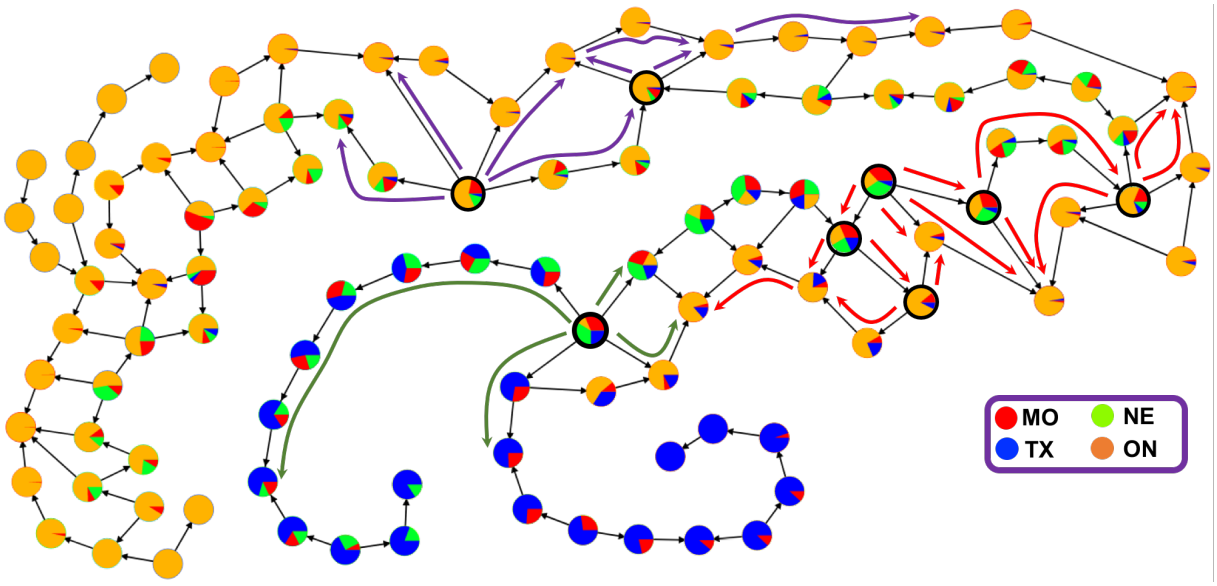

Figure 10: Topological object constructed from the larger G2F data set using temperature as a single filter function. Also shown are the top three interesting flares detected by our method. Our algorithm took 2 seconds to detect all these flares after the Mapper graph is built. Each cluster (node) of the topological object is rendered as a pie-chart showing the distribution of the four classes (based on state/province) of individuals. Long arcs of different colors show interesting flares, and the corresponding branching nodes are identified with bold border. The red flare (on the right side) was ranked as the most interesting flare.

5 SOFTWARE AVAILABILITY

We have implemented the core algorithms for flare detection as part of the HYPPO-X repository, which includes our open source implementation of the Mapper framework. The HYPPO-X software repository is publicly available at https://xperthut.github.io/HYPPO-X. The core computational modules are implemented in C++ and the visualization modules are implemented using the JavaScript visualization library D3 [15].

6 CONCLUSION

We study the fundamental problem of identifying and quantifying divergent subpopulations in complex phenomics data sets, which show differential behavior under different types of environmental conditions. We present an algorithmic framework based on techniques from algebraic topology to identify flares, which are branching features in phenomics data that characterize divergent behavior of subpopulations, and rank these flares according to their interestingness. Results from two phenomics data sets demonstrate the effectiveness and versatility of our framework in characterizing divergent subpopulations, and to suggest hypotheses for further testing by the practitioners.

Our framework is currently prescribed for DAGs. A natural yet valuable extension to consider would be to allow bidirectional edges between clusters whose average phenotype values are within a small tolerance of each other. We would have to generalize our definitions and algorithms to handle directed graphs. On the theoretical side, characterizing the complexity of identifying the most interesting flare, or *all* interesting flares in a given topological object, is a crucial next step (similar to the corresponding problems in the case of interesting paths [8]).

Improving the computational efficiency of various steps in our current pipeline (for DAGs) is of interest as well. In particular, we plan to work on efficient parallel implementations and on speeding up the clustering steps in the Mapper framework, which could be a bottleneck when studying large scale phenomics data sets with multiple phenotypes of interest.

ACKNOWLEDGEMENTS

We thank Dr. Patrick Schnable for sharing with us both maize data sets used in our experiments. The research was supported by U.S. National Science Foundation grant DBI 1661348.

REFERENCES

[1] Muthu Alagappan. 2012. From 5 to 13: Redefining the positions in basketball. In *MIT Sloan Sports Analytics Conference*. http://www.sloansportsconference.com/content/the-13-nba-positions-using-topology-to-identify-the-different-types-of-players/.
[2] Robert M Bilder, FW Sabb, TD Cannon, ED London, JD Jentsch, D Stott Parker, RA Poldrack, C Evans, and NB Freimer. 2009. Phenomics: the systematic study of phenotypes on a genome-wide scale. *Neuroscience* 164, 1 (2009), 30–42.
[3] Mathieu Carrière, Bertrand Michel, and Steve Oudot. 2017. Statistical Analysis and Parameter Selection for Mapper. (2017). arXiv:1706.00204.
[4] Mathieu Carrière and Steve Oudot. 2017. Structure and Stability of the One-Dimensional Mapper. *Foundations of Computational Mathematics* (30 Oct 2017). https://doi.org/10.1007/s10208-017-9370-z arXiv:1511.05823.
[5] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, Vol. 96. 226–231.
[6] David Houle, Diddahally R Govindaraju, and Stig Omholt. 2010. Phenomics: the next challenge. *Nature Reviews Genetics* 11, 12 (2010), 855–866.

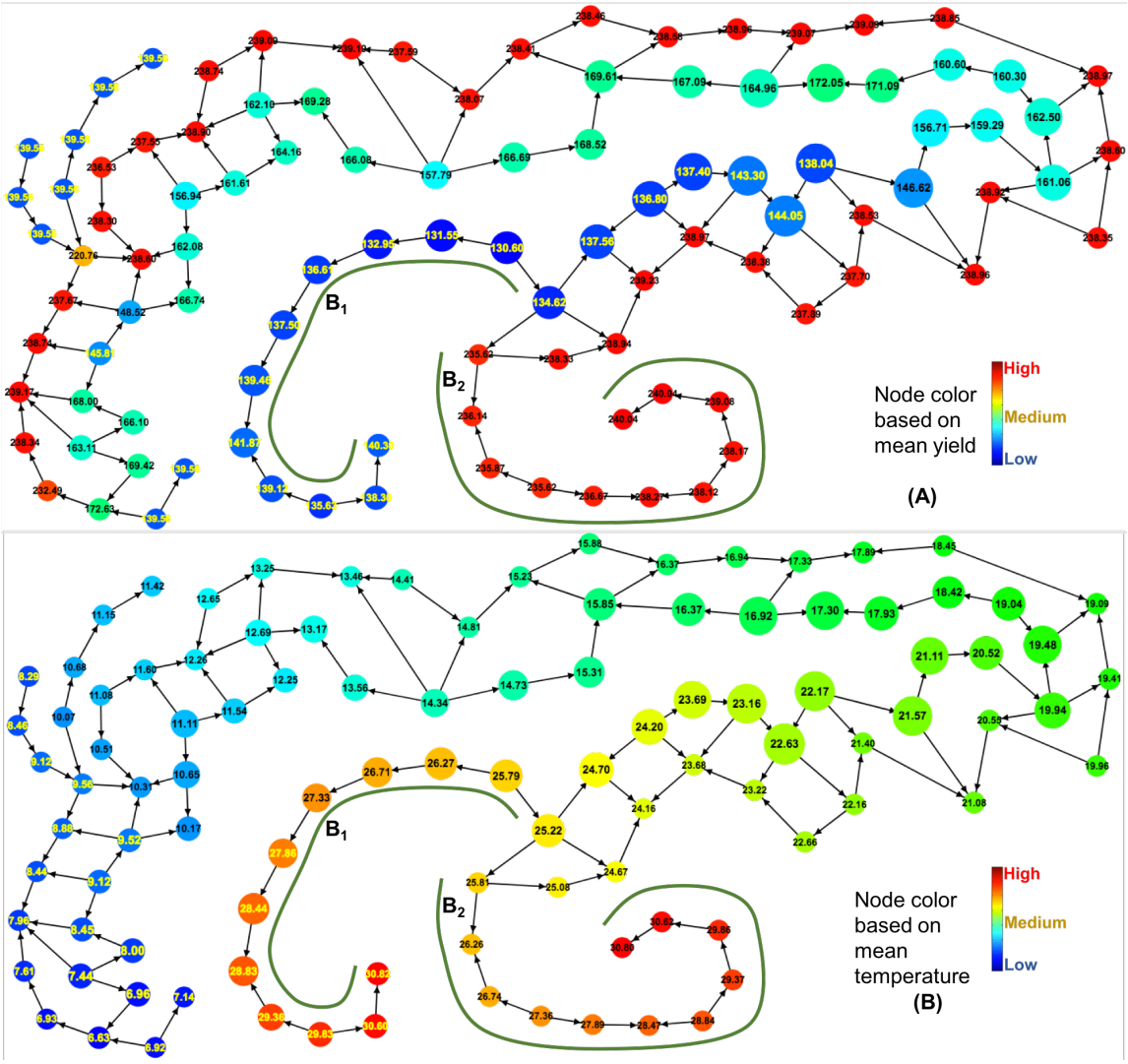

Figure 11: Topological object constructed from the larger G2F data set using temperature as a single filter function (shown also in Figure 10). Each cluster (node) here is colored by mean yield value in Part (A) and by mean temperature value in Part (B). The branches B_1 and B_2 show differences in phenotypic performance (values) even though the temperature values are comparable between them.

[7] G2F Initiative. 2018. The Genomes to Fields Initiative. https://www.genomes2fields.org/.

[8] Ananth Kalyanaraman, Methun Kamruzzaman, and Bala Krishnamoorthy. 2017. Interesting Paths in the Mapper. (2017). Submitted; arXiv:1712.10197.

[9] Methun Kamruzzaman. 2017. Hypothesis extraction tool from high dimensional phenomics dataset. https://xperthut.github.io/HYPPO-X.

[10] Methun Kamruzzaman, Ananth Kalyanaraman, Bala Krishnamoorthy, and Patrick Schnable. 2017. Toward A Scalable Exploratory Framework for Complex High-Dimensional Phenomics Data. (2017). Submitted; arXiv:1707.04362.

[11] Pek Y. Lum, Gurjeet Singh, Alan Lehman, Tigran Ishkanov, Mikael. Vejdemo-Johansson, Muthi Alagappan, John G. Carlsson, and Gunnar Carlsson. 2013. Extracting insights from the shape of complex data using topology. *Scientific Reports* 3, 1236 (2013). https://doi.org/10.1038/srep01236

[12] Monica Nicolau, Arnold J. Levine, and Gunnar Carlsson. 2011. Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proceedings of the National Academy of Sciences* 108, 17 (2011), 7265–7270. https://doi.org/10.1073/pnas.1102826108

[13] Jessica L. Nielson, Jesse Paquette, Aiwen W. Liu, Cristian F. Guandique, C. Amy Tovar, Tomoo Inoue, Karen-Amanda Irvine, John C. Gensel, Jennifer Kloke, Tanya C. Petrossian, Pek Y. Lum, Gunnar E. Carlsson, Geoffrey T. Manley, Wise Young, Michael S. Beattie, Jacqueline C. Bresnahan, and Adam R. Ferguson. 2015. Topological data analysis for discovery in preclinical spinal cord injury and traumatic brain injury. *Nature Communications* 6 (Oct. 2015), 8581+. https://doi.org/10.1038/ncomms9581

[14] Gurjeet Singh, Facundo Memoli, and Gunnar Carlsson. 2007. Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition. In *Proceedings of the Symposium on Point Based Graphics*, M. Botsch, R. Pajarola, B. Chen, and M. Zwicker (Eds.). Eurographics Association, Prague, Czech Republic, 91–100. https://doi.org/10.2312/SPBG/SPBG07/091-100

[15] Swizec Teller. 2013. *Data Visualization with d3.js.* Packt Publishing Ltd.

[16] Brenda Y. Torres, Jose Henrique M. Oliveira, Ann Thomas Tate, Poonam Rath, Katherine Cumnock, and David S. Schneider. 2016. Tracking Resilience to Infections by Mapping Disease Space. *PLoS Biol* 14, 4 (04 2016), 1–19. https://doi.org/10.1371/journal.pbio.1002436