Revealing Microbial Responses to Environmental Dynamics:

Developing Methods for Analysis and Visualization

of Complex Sequence Datasets.

by

Matthew Kellom

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved September 2017 by the
Graduate Supervisory Committee:

Jason Raymond, Chair
Ariel Anbar
James Elser
Everett Shock
Sara Walker

ARIZONA STATE UNIVERSITY

December 2017

ABSTRACT

The greatest barrier to understanding how life interacts with its environment is the complexity in which biology operates. In this work, I present experimental designs, analysis methods, and visualization techniques to overcome the challenges of deciphering complex biological datasets. First, I examine an iron limitation transcriptome of *Synechocystis* sp. PCC 6803 using a new methodology. Until now, iron limitation in experiments of *Synechocystis* sp. PCC 6803 gene expression has been achieved through media chelation. Notably, chelation also reduces the bioavailability of other metals, whereas naturally occurring low iron settings likely result from a lack of iron influx and not as a result of chelation. The overall metabolic trends of previous studies are well-characterized but within those trends is significant variability in single gene expression responses. I compare previous transcriptomics analyses with our protocol that limits the addition of bioavailable iron to growth media to identify consistent gene expression signals resulting from iron limitation. Second, I describe a novel method of improving the reliability of centroid-linkage clustering results. The size and complexity of modern sequencing datasets often prohibit constructing distance matrices, which prevents the use of many common clustering algorithms. Centroid-linkage circumvents the need for a distance matrix, but has the adverse effect of producing input-order dependent results. In this chapter, I describe a method of cluster edge counting across iterated centroid-linkage results and reconstructing aggregate clusters from a ranked edge list without a distance matrix and input-order dependence. Finally, I introduce dendritic heat maps, a new figure type that visualizes heat map responses through expanding and contracting sequence clustering specificities. Heat maps are useful for comparing data across a range of

possible states. However, data binning is sensitive to clustering cutoffs which are often arbitrarily introduced by researchers and can substantially change the heat map response of any single data point. With an understanding of how the architectural elements of dendrograms and heat maps affect data visualization, I have integrated their salient features to create a figure type aimed at viewing multiple levels of clustering cutoffs, allowing researchers to better understand the effects of environment on metabolism or phylogenetic lineages.

ACKNOWLEDGMENTS

This dissertation would not have been possible without the help of:

My committee members:
Ariel Anbar
James Elser
Everett Shock
Sara Walker

Friends and collaborators:
Eric Alsop
Amisha Poret-Peterson

And my advisor:
Jason Raymond

TABLE OF CONTENTS

Page

APPENDIX Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

CHAPTER 1

INTRODUCTION.

The greatest barrier to understanding the interactions of life and its environment is the extreme complexity in which it operates. From the biological side of this relationship, the study of life-environment interactions is contingent upon making sense of the changes that occur in response to altered environmental conditions. However, the complexity of natural systems makes finding ways to collect, analyze, and interpret biological data a nontrivial task. Biology research is currently in an era of abundant data collection, leading to massive datasets and databases that contain the potential for scientific discovery (Wooley & Lin, 2005; Wooley & Ye, 2010). Large scale biological data collection is a heterogeneous affair that incorporates components from interconnecting biological systems and varying degrees of similarity (Wooley & Lin, 2005). Sequencing data in particular, and the methods to collect it, have advanced significantly within just the past ten years (Acland et al., 2014; O'Leary et al., 2016; Quail et al., 2012; Wooley & Ye, 2010). Genome and metagenome sequences are now routinely assembled, comparatively annotated, and uploaded to public databases (Bailey et al., 2014c, 2014a, 2014b; Bryant & Frigaard, 2006; Swingley et al., 2012). As a result of this data boom, the development of methods to study sequencing data complexity, and biological complexity in general, is a quickly growing interdisciplinary field (Heo, Kang, Song, & Lee, 2017). This dissertation describes work aimed at managing and learning from biological complexity through novel experimental designs, analysis methods, and visualization techniques in three chapters.

The first chapter, *Transcriptomics of iron limitation without media chelation in the cyanobacterium Synechocystis sp. PCC 6803*, examines the effects of iron limitation on the photosynthetic metabolism of *Synechocystis* sp. PCC 6803. *Synechocystis* sp. PCC 6803 is a model organism for studying the effects of iron limitation on photosynthetic metabolism with well characterized RNAseq and microarray results (Hernández-Prieto et al., 2012; Hernández-Prieto, Semeniuk, Giner-Lamia, & Futschik, 2016; Kopf et al., 2014; Shcolnick, Summerfield, Reytman, Sherman, & Keren, 2009; Singh, McIntyre, & Sherman, 2003; Wegener et al., 2010). Previous studies have established differential expression trends for both coding and non-coding genome regions that show stress responses to iron limitation. However, some individual gene responses are inconsistent in their signal across multiple experiments, which could be caused by differences in methodologies and growth conditions. In contrast to previous transcriptome experiments which all used iron chelators to reduce bioavailable iron, iron limitation was achieved in this experiment by supplying an order of magnitude less iron in growth media compared to controls. With the methods and results of previous iron limitation *Synechocystis* sp. PCC 6803 transcriptome studies in consideration, it was hypothesized that this non-chelation iron limitation growth and subsequent analysis methodology would result in similar overall transcriptome trends to those of previous reports, particularly those of photosynthesis and respiration. However, within these overall trends, some unique differences in individual gene expression were expected because of the different iron limitation method. By examining iron limitation under alternative methods and growth conditions, the confidence in consistent signals resulting from iron unavailability will be strengthened and inconsistent signals may be questioned as products of a particular

protocol. In this chapter, an experimental design change is being used to reduce the complexity of the known *Synechocystis* sp. PCC 6803 iron limitation transcriptome responses.

The second chapter, *Using cluster edge counting to aggregate iterations of centroid-linkage clustering results and avoid large distance matrices*, describes a novel method of improving the reliability of centroid-linkage clustering results (Kellom & Raymond, 2017). Sequence clustering is a fundamental analysis tool of molecular biology that is being challenged by increasing dataset sizes from high-throughput sequencing. Agglomerative algorithms, such as minimum- maximum- and average-linkage, that have long been relied upon for their accuracy, require the construction of computationally costly distance matrices which can overwhelm basic research personal computers (Cole et al., 2009; Gronau & Moran, 2007; Huse, Welch, Morrison, & Sogin, 2010; Larkin et al., 2007). Alternative algorithms exist, such as centroid-linkage, to circumvent large memory requirements but their results are often input-order dependent (Edgar, 2010). The method of cluster edge counting presented in this chapter effectively bootstraps the results of many centroid-linkage clustering iterations into an aggregate set of clusters, increasing cluster accuracy without a distance matrix. The novel analysis method in this chapter ranks cluster edges by conservation across iterations and reconstructs aggregate clusters from the resulting ranked edge list. Aggregating centroid-linkage clustering iterations can help researchers analyze the complexity of sequencing datasets.

The third chapter, *Using dendritic heat maps to simultaneously display genotype divergence with phenotype divergence*, introduces a new figure type that can visualize

heat map responses though expanding and contracting clustering specificities (Kellom & Raymond, 2016). The visualization of sequencing data is an integral part of the analysis and communication of genomics-based research. A key advance in microbial ecology in both modern and ancient ecosystems will be connecting genotypic lineages and metabolic components to environmental dynamics. This has become a daunting task in light of the burgeoning repositories of -omics sequence data, calling on entirely new methods for analyzing and visualizing complex sequencing datasets (Goll et al., 2010; Huson, Auch, Qi, & Schuster, 2007; Meyer et al., 2008; Ondov, Bergman, & Phillippy, 2011). The effects of environment on biology are often shown as heat maps of sequence abundance, where the responses of distinct sequence groups are measured and compared (Wilkinson & Friendly, 2009). However, the grouping process of heat map construction is performed at a single, often arbitrary, level of inclusiveness. In this chapter, dendritic heat maps are introduced to simultaneously display multiple heat maps over a range of binning specificities, arranged in a dendrogram-like configuration. Dendritic heat maps can show the effects of environment on sequence homology and relative abundance. Importantly, tracking changes in relative abundance can be particularly useful for observing the levels at which genotypic divergence (cluster branching) correlates with gene expression (differing heat map bin response), helping to better understand the effects of environment on metabolism or phylogenetic lineages.

The concluding chapter, *Architectural elements of dendrogram and heat map visualization, and the display of hierarchical clustering multidimensionality*, discusses the effects of architectural elements in published dendrograms and heat maps. Input data, scale and density, and color, in dendrograms and heat maps are examined with respect to

communicating the complexity of clustered datasets and their struggle to display multidimensionality. Hidden within nearly every dendrogram or heat map are many levels of equally legitimate versions of the same data display. The reality of clustered relationships is often more disordered than what is presented by the final dendrogram or heat map image, but data are forced into end-point clusters based on identity cutoffs. There have been two main strategies to meet the challenge of visualizing multidimensionality of clustering data: 1) reorganizing cluster hierarchies by 'cutting' branches at multiple clustering cutoff levels, and 2) overlaying heat map values over dendrogram hierarchies. Overlaying heat maps onto dendrogram configurations shows the multidimensionality of nodes and branching points in clustering datasets whereas branch cutting methods highlight a selection of multidimensional nodes. Each method conveys the complexity and dynamic nature of clustered hierarchical datasets in ways that are not possible in traditional dendrograms and heat maps.

CHAPTER 2

TRANSCRIPTOMICS OF IRON LIMITATION WITHOUT MEDIA CHELATION IN THE CYANOBACTERIUM *SYNECHOCYSTIS* SP. PCC 6803.

Matthew Kellom[1], Amisha T. Poret-Peterson[1†], Albert Rivas-Ubach[2,3], James J. Elser[4,5], and Jason Raymond[1*]

[1]School of Earth and Space Exploration, Arizona State University, Tempe, Arizona, USA

[2]Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, Washington, USA

[3]Global Ecology Unit CREAF, Autonomous University of Barcelona, Bellaterra, Barcelona, Spain

[4]School of Life Sciences, Arizona State University, Tempe, Arizona, USA

[5]Flathead Lake Biological Station, University of Montana, Polson, Montana, USA

†Current affiliation: USDA-ARS Crops Pathology and Genetics Research Unit, University of California, Davis, California, USA

Submitted to *Acta Physiologiae Plantarum.*

Abstract

Synechocystis sp. PCC 6803 is a model organism for studying the effect of iron limitation on photosynthetic metabolism and has been used in RNAseq and microarray experiments. Previous studies have established differential expression trends for both coding and non-coding genome regions that show stress responses to iron limitation. However, some individual gene responses are inconsistent in their signal across multiple experiments, which could be caused by differences in methodologies and growth conditions. By examining iron limitation under alternative methods and growth conditions, the confidence in signals resulting from iron unavailability will be strengthened and inconsistent signals may be regarded as products of a particular protocol. Our experiment, like others, yielded results that indicate ubiquitous downregulation of photosynthetic electron transport chain subunits as well as transporters that allow H^+ ions to exit the thylakoid lumen, possibly to maintain a thylakoid membrane electrochemical gradient. This widespread gene repression response to iron limitation was accompanied by the upregulation of iron acquisition pathways. In contrast to previous experiments which used iron chelators to reduce bioavailable iron, we achieved iron limitation by supplying an order of magnitude less iron in growth media compared to controls. As expected, some of our results do not exactly mirror the results from previous studies. We have visualized these iron limitation signals by mapping them over the *Synechocystis* sp. PCC 6803 chromosome. The role of *Synechocystis* sp. PCC 6803, and cyanobacteria in general, as primary producers underscores the ecological importance of understanding stress induced by various iron limitation growth conditions.

Introduction

Synechocystis sp. PCC 6803 (hereafter referred to as PCC 6803) is a model organism for studying iron limitation in cyanobacteria, for which qPCR, microarray, RNAseq, and proteomic experiments have characterized its responses (Singh, et al., 2003; Shcolnick, et al., 2009; Wegener, et al., 2010; Hernández-Prieto, et al., 2012; Kopf, et al., 2014; Hernández-Prieto, et al., 2016). In cyanobacteria, iron is essential for photosynthesis, respiration, dinitrogen fixation, chromophore biosynthesis, and gene regulation, and cells will react to maintain internal homeostasis when it is lacking from their environment. Singh, *et al.*, were the first to use a microarray approach to show altered transcription of genes for photosynthesis and respiration, transcription and translation, transport and binding proteins, and other metabolic functions by PCC 6803 under iron limitation (Singh, et al., 2003). Two other microarray studies reported iron limitation affecting similar metabolic pathways with slight differences from results reported by Singh, *et al.* (2003), which likely stems from differences in methodologies (Shcolnick, et al., 2009; Hernández-Prieto, et al., 2012). Growth conditions between these two other experiments were similar: Shcolnick, *et al.* and Hernández-Prieto, *et al.* used the iron-chelator defroxamine B (DFB) to generate iron insufficiency and assessed gene expression by PCC 6803 after varying amounts of time (Shcolnick, et al., 2009; Hernández-Prieto, et al., 2012). Recently, Kopf, *et al.*, 2014 described the PCC 6803 transcriptome response with RNA sequencing (RNAseq) and a "transcriptional unit" approach under ten different growth conditions, including iron limitation induced by DFB addition (Kopf, et al., 2014). They identified iron-stress and transport genes as being differentially expressed explicitly during iron limitation and data for other genes shows

similar differential expression patterns to that of the microarray experiments. These are four examples of iron limitation transcriptome analysis in PCC 6803 performed with different methodologies, all yielding slightly different results for individual genes but similar overall gene expression patterns.

The primary goal of our study is to supplement existing transcriptome studies of PCC 6803 under iron limitation without media iron chelation. The confidence in transcriptome signals is strengthened by an increasing number of reports, while transcriptome noise may become more evident with fluctuating results that show inconsistencies across studies. In our experiment, the growth conditions, sequencing platform, and data analysis methods are different than the methods used in the previously mentioned studies. To achieve iron limitation in our experiment, we lowered the iron content of the medium by adding an order of magnitude less bioavailable iron, rather than using chelators, to induce iron-limited gene expression. This difference in procedure is noteworthy because growth in naturally iron-limited settings is likely not the result of chelators but instead the lack of sufficient iron input into oligotrophic environments. Additionally, DFB, the iron chelator of past experiments, is known to bind to metals other than iron and limit their bioavailability (Farkas, et al., 1997; Farkas, et al., 1999).

There are many sequencing platform options available for transcriptomics, each with advantages and disadvantages when it comes to cost, speed, and output, as well as access to the technology (Quail, et al., 2012). Each sequencing platform can yield slightly different results due to their inherent biases, as can different sample preparation methods leading up to sequencing and the analysis done afterward. The most recent PCC 6803 iron limitation experiment (Kopf, et al., 2014) other than our own used an Illumina HiSeq

platform, which is currently widely used, along with Illumina MiSeq, for RNAseq because of their accuracy and the number of sequences yielded per run. RNAseq analysis methods generally involve mapping (e.g. Scripture (Guttman, et al., 2010), Cufflinks (Trapnell, et al., 2010)) and/or assembling (e.g. ABySS (Birol, et al., 2009), SOAPdenovo (Xie, et al., 2014)) sequencing reads (Haas & Zody, 2010). Kopf, *et al.*, 2014 mapped reads to the PCC 6803 genome to identify transcriptional start sites and transcriptional units using the program "segemehl" (Hoffman, et al., 2009), which finds optimally scored local alignments between sequencing reads and a reference genome (Kopf, et al., 2014).

For our experiment, we sequenced our transcriptome replicates with the Ion Torrent sequencing platform. The Ion Torrent sequencer is cheaper than the Illumina HiSeq sequencer but the cost is comparable to some of the other Illumina platforms. However, the cost per gigabase of Ion Torrent can be more expensive than Illumina HiSeq depending on the productivity of the individual machines (Quail, et al., 2012). Like Illumina HiSeq, Ion Torrent can yield a large number of sequences of a useful length and accuracy for transcriptomics, although typically to a lesser degree (Quail, et al., 2012). Sequenced reads were mapped to the PCC 6803 genome using local BLAST to identify their corresponding locations with local alignments (Altschul, et al., 1990). This process uses the same concept as the program segemehl, but applied in a different way. In the results and discussion we compare the results of our transcriptome experiment methods to previous reports.

The secondary goal of the work presented here is to provide an alternate method for visualization of genome-mapped transcriptome data. We utilize the plotting Perl

package Circos with a custom Perl script to visually map transcriptome expression levels to their location on a circular chromosome. This visualization technique is adaptable to the use of alternative software and provides a view of regional relative abundances over the entire chromosome.

With the methods and results of previous iron limitation *Synechocystis* sp. PCC 6803 transcriptome studies in consideration, we hypothesized that our non-chelation iron limitation growth and subsequent analysis methodology would result in similar overall transcriptome trends to those of previous reports, particularly those of photosynthesis and respiration. However, within these overall trends, we also expected some unique differences in individual gene expression because of our different iron limitation method.

Materials and Methods

Growth. *Synechocystis* sp. PCC 6803 was grown in 1.8 L of BG-11 medium pH 7.8 (Allen, 1968; Stanier, et al., 1971) in 2 L trace metal clean polycarbonate bottles at 24ºC under continuous aeration with 0.2 μm filtered air and illumination (50 μmol photons m^{-2} s^{-2} irradiance). Iron-limited cultures were grown with iron reduced to 1/10 (1.8 μM) of the control medium, supplied as $FeCl_3 \cdot 6H_2O$. The cultures were grown under continuous light as similarly reported in the methods of Kopf, *et al.* 2014. The cells were then harvested and transferred to normal composition BG-11 (control treatment: 18 μM iron) for 7 days or modified BG-11 (iron limitation treatment: 1.8 μM iron) for 10 days to assess the effect of iron limitation on physiological processes in PCC 6803. Cultures were set up in replicates of six and collected in exponential phase. The growth rate of the cultures was monitored via absorbance at 730 nm.

Sample preparation. Total RNA was extracted from cultures using the FastRNA Pro™ Blue Kit from MP Biomedicals (cat. no. 116025050). DNA in the total RNA preparations was degraded using the RTS-DNase from Mo Bio Laboratories (cat. no. 15200-50). The RNeasy® MinElute Cleanup Kit from QIAGEN (cat. no. 74204) was used to purify and concentrate the DNased RNA. rRNA was subtracted from the total RNA extractions following the protocol from Stewart *et al.* 2010. Of the six replicates for both the control and iron-limited group, the three of each that contained the highest concentration of RNA were amplified using the MessageAmp™ II-Bacteria Kit from Life Technologies (cat. no. AM1790). The amplified RNA was then reverse transcribed to cDNA using the SuperScript® Double-Stranded cDNA Synthesis Kit (with Superscript® III substituted for Superscript® II reverse transcriptase) from Life Technologies (cat. no. 11917) and sonicated to ~500 bp fragments, confirmed with gel electrophoresis. cDNA fragments were then prepared for Ion Torrent sequencing using the Ion Xpress™ Plus gDNA and Amplicon Library Preparation kit and the Ion Xpress™ Barcode Adapter 1-16 Kit from Life Technologies (cat. no. 4471269 and 4471250, respectively).

Sequencing and processing. The samples were sequenced using the Ion Torrent platform with the 316 chipset at the DNA Laboratory at the Arizona State University School of Life Sciences. Sequences less than 150 bp in length were filtered out of the data. Using local BLAST, any sequences that did not match to the PCC 6803 chromosome or plasmids at the default e-value of 10 were removed from the data (these sequences are either too inaccurate or are from *Hymenobacter* contamination, addressed in the following paragraph), as well as sequence that matched the 23S, 16S, and 5S rRNA gene regions (Altschul, et al., 1990). After sequence filtering, sample normalization was

performed by random subsampling without replacement to the size of the smallest dataset via Perl script, so that all sequencing datasets were equally represented. Sequences have been deposited into the National Center for Biotechnology Information's Sequence Read Archive as BioProject PRJNA315016.

In the results reported here, it should be noted that minor culture contamination was detected post-sequencing by the identification of genus *Hymenobacter* 16S ribosomal RNA sequences in all six samples. Sequences that did not map to the PCC 6803 genome were discarded before subsampling, but it is conceivable that conserved genes between PCC 6803 and *Hymenobacter* could distort count information. While this contamination is less than ideal, we have not considered it to be a major concern since we have framed the transcriptome responses with comparisons to previous work done with iron limitation in Cyanobacteria and PCC 6803.

Transcriptome mapping and visualization. Chromosome and plasmid positions of sequencing reads were determined using local BLAST with the PCC 6803 complete chromosome as a database. The positions of the reads were organized using custom Perl scripts that captured the hit positions from BLAST output files. Annotation was performed by comparing the positions of both coding and non-coding regions on the chromosome with the sequencing read positions that were mapped. Plasmids were also considered during transcriptome mapping, but ultimately we chose not to map them due to the lack of significant expression toward either group in our experiment. With these read positions for the three replicates of both conditions, the bin response of Figure 1, $\log_2$ fold change of Figures 2 and 3, and two-sampled t-statistic for each coding and contiguous non-coding region were calculated using equations (1), (2), and (3),

respectively. In addition, edgeR was used to calculate an exact test for significance (Robinson, et al., 2010). Reads were incorporated into the calculations if any part of it overlapped the coding or contiguous non-coding region.

1) $Bin\ response = \log_{10} \frac{(Fe\ Limitation_{mean}+1)}{(Control_{mean}+1)}$

2) $Fold\ change = \log_{2} \frac{(Fe\ Limitation_{mean}+1)}{(Control_{mean}+1)}$

3) $t\ statistic = \frac{Fe\ Limitation_{mean} - Control_{mean}}{\sqrt{\frac{SD_{Fe\ Limitation}{}^{2}}{n_{Fe\ Limitation}} + \frac{SD_{Control}{}^{2}}{n_{Control}}}}$

Data visualization in Figure 1 was performed by plotting with the Perl package Circos (version 0.64) (Krzywinski, et al., 2009). Sequencing reads were mapped and stacked onto genome positions at the individual base level with custom Perl scripts. The triplicate iron limitation growth datasets were assigned differing red hues, with the triplicate control datasets assigned differing blue hues. To prevent significant signals from being dwarfed by genes of exceptionally high differential expression and to ensure visually appealing image proportions, a $\log_{10}$ ratio was used for visualization and bases were mapped to the genome with a maximum count of sixteen, however all analysis was performed with full count information. Genome positions were grouped based on gene regions or contiguous noncoding regions, where the bin response and t-statistic were calculated, as well as an exact test with edgeR. The bin response is plotted as a heat map with hues that are partitioned into eleven possible red/blue hues, with the red and blue corresponding to a higher count of expressed mRNA transcripts in the iron limitation triplicate average or the control triplicate average, respectively. Regions of the

chromosome that did not meet the p-value of ≤0.05 (t-statistic of 2.776) in the t-test and the exact test have been faded to a lesser color intensity.

All Perl scripts used here are available in Chapter 2 supplemental file S2. Summaries of the results are given in Table 1, Figure 2, Figure 3, and Chapter 2 supplemental file S1.

Results

Transcriptome mapping. In this study, we utilized high-throughput sequencing data to map the iron-limited transcriptome of PCC 6803 to its genome and compare iron limitation transcript counts to counts from control growth conditions. This technique is amenable to analysis of the transcriptome from the level of the entire genome to individual bases. Mapping of the transcriptome in the present work was performed at the single-base level, by assigning each sequenced read to its corresponding location on the ~3.5 million base pair genome then counting those reads over the span of individual genes and can be viewed in Figure 1. This mapping approach to gene expression analysis offers a high resolution view of a PCC 6803 iron limitation growth response that adds to what has been previously reported in microarray and proteomic studies but with an iron limited response that was induced without iron chelation (Singh, et al., 2003; Wegener, et al., 2010; Kopf, et al., 2014).

The *de novo* methodology followed in the present work has led to results that are largely consistent with previous iron limitation studies (Singh, et al., 2003; Hernández-Prieto, et al., 2012; Kopf, et al., 2014; Hernández-Prieto, et al., 2016). Although iron limitation without chelation resulted in widespread differential expression of genes in multiple functional categories, ranging from metabolite biosynthesis to transposon-

related functions (Table 1), we will center our focus on photosynthesis and respiration along with transport and binding proteins. Each of these functional categories contains differentially expressed genes that show an apparent iron limitation distress signal (Figure 2). In some select co-localization cases of the genes listed in Figure 2, we illustrate significant differential gene expression at chromosome regions which are referenced in the discussion (Figure 3). However, it should be noted some of the strongest individual differential gene expression signals were detected outside of the Figure 2 categories, in ribosomal proteins, hypothetical proteins, and unknown proteins (Supplemental File S1).

Tables and Figures.

Table 1

Differentially regulated protein-coding genes in response to Fe limitation according to functional categories as defined in Cyanobase

General Pathway	No. of Genes	Differentially Expressed Genes*
Amino acid biosynthesis	97	4
Biosynthesis of cofactors, prosthetic groups, and carriers	125	7
Cell envelope	67	5
Cellular processes	80	3
Central intermediary metabolism	31	2
Energy metabolism	93	4
Fatty acid, phospholipids, and sterol metabolism	39	1
Photosynthesis and respiration	143	33
Purines, pyrimidines, nucleosides and nucleotides	43	1
Regulatory functions	156	9
DNA replication, restriction, modification, recombination, and repair	75	0
Transcription	30	1
Translation	168	6
Transport and binding proteins	200	8
Other categories	369	18
Hypothetical protein	1277	28
Unknown protein	679	12
Total	3672	142
Non-coding regions (contiguous sequence between genes)	2975	60

Styled after Table I of Singh *et al.*, 2003 (Singh, et al., 2003). *p-value $\leq$ 0.05 from both t-test and exact test.

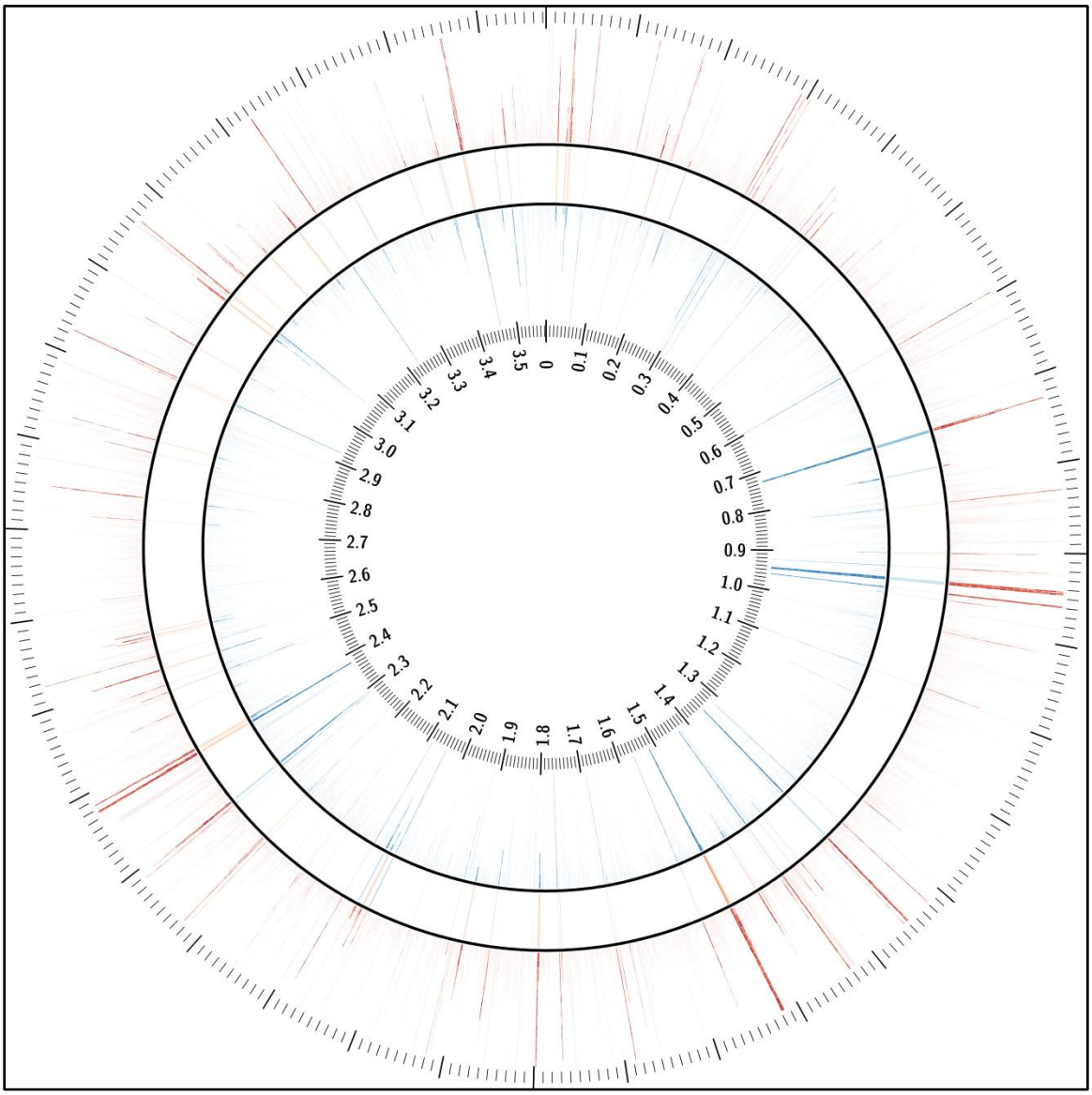

Figure 1: The PCC 6803 iron limitation and control transcriptomes mapped the chromosome. The outer red-hued histogram-like layer shows iron limitation sequencing reads mapped and stacked over their corresponding chromosome positions. There are three separate red-hues to indicate the difference between the triplicate samples. The inner blue-hued histogram-like layer shows control sample sequencing reads mapped and stacked over their corresponding chromosome positions, with three different hues as in

the iron limitation layer. Each of these two layers display a max read stacking level of sixteen sequences for aesthetic reasons, however counts involved in the analysis use true count data. The middle ring between the red- and blue-hued histogram-like layers represents a heat map over the entire chromosome. Regions with a red heat map hue show a gene that has a bin response toward being more highly expressed in the iron limitation groups, with a blue heat map hue indicating a bin response toward controls. The numbered circular key in the figure interior extends out to the outer edge and displays chromosome positions in the scale of millions of bases. Regions of the chromosome that did not meet the p-value of ≤ 0.05 have been faded to a lesser color intensity.

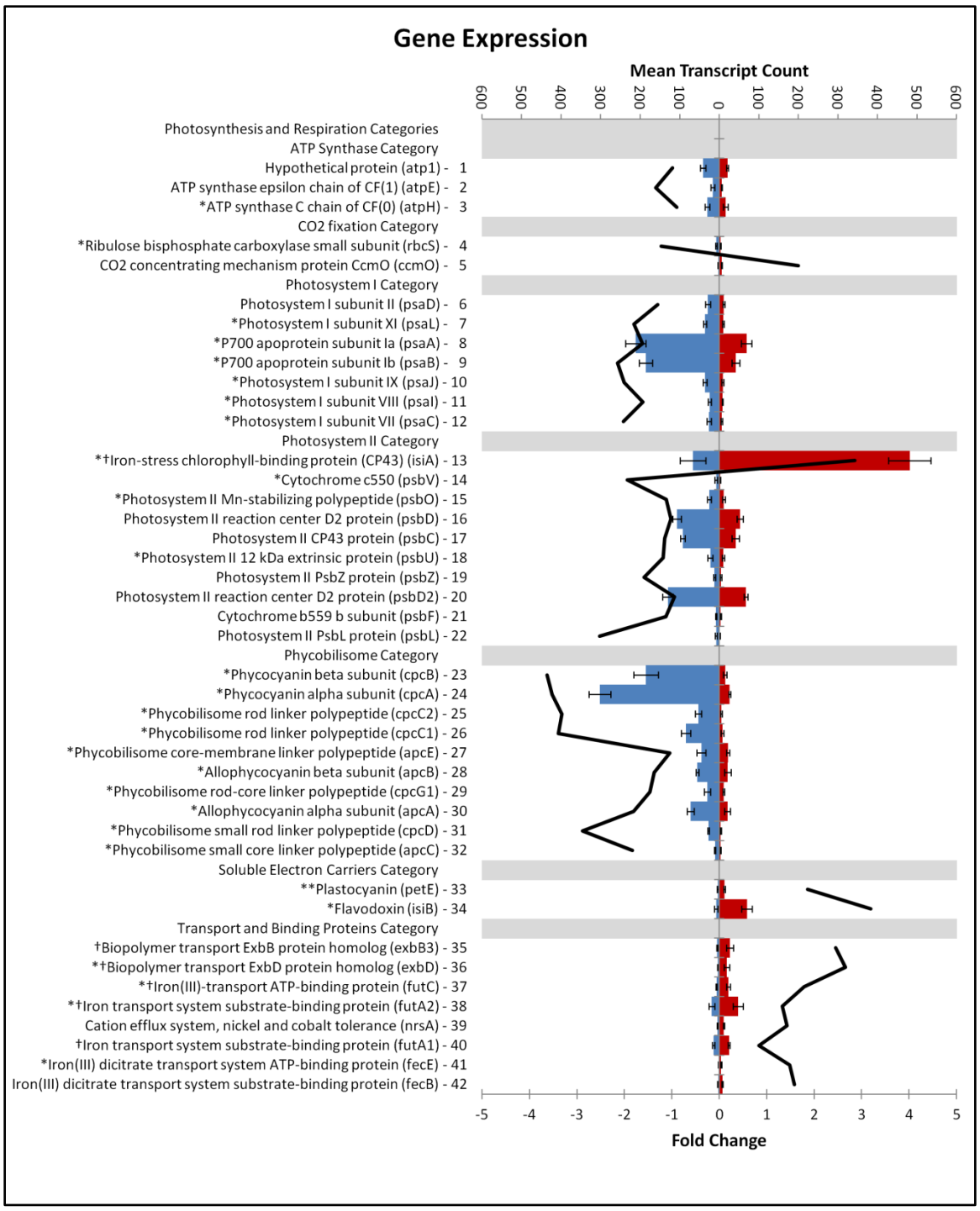

Figure 2: A partial list of differentially regulated genes in response to iron limitation (p-value ≤ 0.05), separated into Cyanobase functional categories. Blue bars correspond to mean transcript counts of control replicates and red bars correspond to iron limitation

replicates. Standard deviation error bars are included on all bars. The x-axis for the red and blue bars is at the top of the histogram, indicating the mean transcript counts. The black line corresponds to the $\log_2$ fold change of comparing the iron limitation mean transcript counts to the control mean transcript counts, with data points positioned at the center of each y-axis row.

*Our $\log_2$ fold change direction agrees with Singh, *et al.*, 2003 (Singh, et al., 2003).

**Our $\log_2$ fold change direction agrees/disagrees with Singh, *et al.*, 2003 depending on choice of their timescale (Singh, et al., 2003).

***Our $\log_2$ fold change direction conflicts with Singh, *et al.* 2003 (Singh, et al., 2003).

†Our $\log_2$ fold change direction agrees with Hernández-Prieto, *et al.*, 2012 (Hernández-Prieto, et al., 2012).

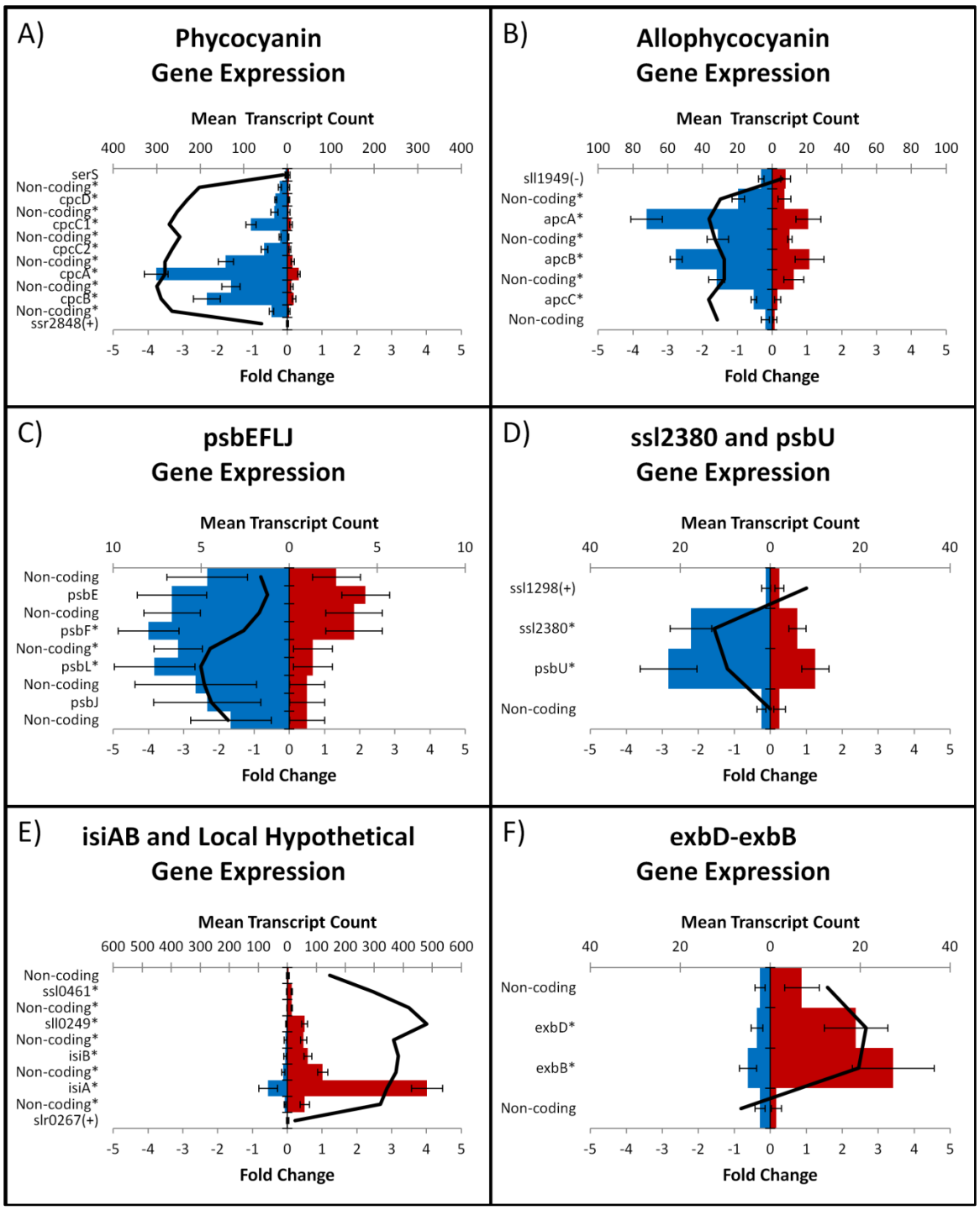

Figure 3: Co-localized gene expression of six significant genome regions of a *Synechocystis* sp. PCC 6803 iron limitation transcriptome. For all six histogram panels, blue bars correspond to mean transcript counts of control replicates and red bars

correspond to iron limitation replicates. Standard deviation error bars are included on all bars and y-axis genome labels with an asterisk indicate statistical significance of a difference between the two opposing bars with a p-value ≤ 0.05. The x-axis for the red and blue bars is at the top of each graph, indicating the mean transcript counts. The black line corresponds to the $\log_2$ fold change of comparing the iron limitation mean transcript counts to the control mean transcript counts, with data points positioned at the center of each y-axis row. A "(+)" or "(-)" on y-axis gene labels indicate that the gene is on the opposite strand as the rest of the plotted responses, but was included on the graph due to because of its genome proximity. Individual panel descriptions: (A) Phycocyanin gene expression at the genome region spanning positions 722,569-728,269 on the negative strand, followed by 728,270-728,488 on the positive strand. (B) Allophycocyanin gene expression at the genome region spanning 1,430,072-1,431,994 on the positive strand, proceeded by 1,429,328-1,430,071 on the negative strand. (C) psbEFLJ operon gene expression at the genome region spanning 570,572-571,438 on the positive strand. (D) Hypothetical protein ssl2380 and psbU gene expression at the genome region spanning 297,323-298,081 on the negative strand, proceeded 295,075-297,342 on the positive strand. (E) isiAB operon and local hypothetical protein gene expression at the genome region spanning 1,515,338-1,519,112 on the negative strand, followed by 1,519,113-1,519,952 on the positive strand. (F) exbBD operon gene expression at the genome region spanning 32,454-33,889 on the negative strand.

Discussion

Synechocystis sp. PCC 6803 photosynthetic electron transport chain.

Consistent with previous PCC 6803 iron limitation work, photosynthesis molecular machinery genes were extensively downregulated. Within statistical significance ($p \leq 0.05$), seven PSI subunit genes (*psaA, psaB, psaC, psaD, psaI, psaJ,* and *psaL*), nine PSII genes (*psbC, psbD, psbD2, psbF, psbL, psbO, psbU, psbV,* and *psbZ*), and ten phycobilisome antenna genes (*apcA, apcB, apcC, apcE, cpcA, cpcB, cpcC1, cpcC2, cpcD,* and *cpcG1*) were downregulated when compared to normal growth conditions. The transcriptome data here shows considerable changes to the whole photosynthetic electron transport chain (PETC hereafter) gene expression (Figures 2 and 3A-C).

The PETC in PCC 6803 is bound within the thylakoid membrane and starts at the PSII and PSI antenna proteins. The phycobilisome antenna complex collects and converts photic energy into electrons to be funneled to the PSII reaction center primary donor P680 (and PSI P700). Almost all genes for the two phycobilisome complexes in PCC 6803, phycocyanin (*cpc*) and allophycocyanin (*apc*), were repressed in response to iron limitation (Figure 2, rows 23-32), which agrees with previous reports of iron limitation in cyanobacteria (Singh, et al., 2003; Sandström, et al., 2002). Due to the high synteny (co-localization) of these genes, this downregulation signal is easily viewed on the transcriptome map of Figure 1 as well as panels A (*cpc*) and B (*apc*) of Figure 3, which is in contrast to the relatively low synteny genome as a whole. *cpc* genes can be viewed between chromosome bases 724,093-727,465 and *apc* between 1,430,418-1,431,900 as blue-hued Bin responses in Figures 1. The reduced demand for phycobilisome proteins could be a result of the iron-dependent chromophore biosynthesis process or simply the

reduced amount of major photosystem subunits has reduced the overall light-harvesting needs of the cell (Sandström, et al., 2002).

The *isiA* gene, encoding the iron-stress chlorophyll binding protein IsiA, was one of the few PETC genes to be upregulated in comparison to controls (Figure 2, row 13) and has been suggested to act in a photoprotective role with the loss of phycobilisome antenna proteins, by dissipating energy into heat to prevent overexcitation of PSII and oxidation damage during iron limitation-induced chlorosis (Guikema & Sherman, 1983; Sandström, et al., 2001; Wilson, et al., 2007). IsiA has also been suggested to act as a chlorophyll store to aid PSII and PSI recovery upon the return of adequate iron uptake, or as a light-harvesting complex mainly used in PSII (Pakrasi, et al., 1985; Riethman & Sherman, 1988; Michel & Pistorius, 2004). This upregulation of *isiA* is consistent with previous PCC 6803 iron limitation transcriptome studies (Singh, et al., 2003; Hernández-Prieto, et al., 2012). PsbD and PsbD2 dimerize and bind with chlorophyll *α*, iron, and β-carotene to form the P680 reaction center of PSII, which either absorbs photons directly or absorbs excitation energy passed from antenna proteins. Under iron limitation that we imposed, the P680 dimer and PsbC subunit genes were repressed (Figure 2, rows 16, 17, 20), possibly to prevent the production of highly reactive singlet oxygen from photooxidizing water molecules in the absence of normal anabolic processes (Grossman, et al., 1993; Pagliano, et al., 2013). Iron limitation also repressed the gene for the manganese-containing PsbO subunit (Figure 2, row 15), which acts to hold the reaction center subunits together while maintaining an environment for water photooxidation via the oxygen-evolving complex (De Las Rivas & Barber, 2004). Correspondingly, the manganese cellular content of our iron-limited cultures was decreased compared to

controls [Rivas-Ubach, *et al., unpublished*]. Also repressed was the iron-containing PSII cytochrome subunit PsbV (cytochrome c550) gene (Figure 2, row 14). *psbL* and *psbF* of the highly conserved *psbEFLJ* operon were downregulated (Figure 2, rows 21,22) and these co-localized genes can be viewed between chromosome positions 570,657-571,354 on Figures 1 and 3C. The downregulated *psbU* and *psbZ* genes (Figure 2, rows 18, 19) likely play roles in overall PSII structural stability (Pagliano, et al., 2013). Located near *psbU*, unknown protein ssl2380 gene has no known categorical function but was conjointly downregulated between chromosome positions 297,323-297,520 in Figures 1 and 3D.

After excitation energy is funneled to the P680 reaction center of PSII to oxidize water and increase the electrochemical gradient across the thylakoid membrane, excited electrons are passed through plastoquinone and the cytochrome b_6f complex. Neither plastoquinone nor the cytochrome b_6f complex genes were significantly downregulated in response to iron limitation in our experiment; however, the iron-rich cytochrome b_6f complex has been shown to be differentially expressed in PCC 6803 and other cyanobacteria subjected to iron limitation (Singh, et al., 2003; Thompson, et al., 2011). As electrons move excitation energy through the combination of plastoquinone and the cytochrome b_6f complex in normal conditions, they transfer H^+ across the membrane to increase the thylakoid membrane electrochemical gradient even further. Once through the cytochrome b_6f complex, the PETC continues through the copper-containing electron carrier plastocyanin (*petE*), which was upregulated under iron limitation in our experiment (Figure 2, row 33), donating electrons to the PSI reaction center P700. Interestingly, plastocyanin may have evolved and replaced a heme protein in the PETC

during a time when populations of ancestral phototrophs were adapting to shifting ocean geochemistry during the Great Oxidation Event (Navarro, et al., 1997; De La Rosa, et al., 2002).

PsaA and PsaB dimerize in the P700 reaction center of PSI, accepting electrons from the phycobilisome antenna and from plastocyanin. Genes for both subunits of P700 were downregulated in response to iron limitation (Figure 2, rows 8, 9). Just as with PSII, PSI uses the same phycobilisome genes that encode for light-harvesting antennae, which were downregulated under iron limitation in our experiment (Figure 2, rows 23-32). IsiA, which has a role in the iron limitation PSII stress response, may also play a significant role in PSI. In iron-deficient conditions, IsiA binds to chlorophyll and acts as a light-harvesting complex that substitutes for the downregulated phycobilisomes. (Michel & Pistorius, 2004; Boekema, et al., 2001; Bibby, et al., 2001). Most subunit genes of the PSI were repressed by iron deficiency in our experiment, including the gene for PsaC (Figure 2, row 12) which is responsible for coordinating 4Fe-4S clusters as the terminal electron acceptors for the P700 reaction center. PSI subunits PsaL, PsaJ, and PsaI are integral membrane subunits whose genes were all repressed (Figure 2, rows 7, 10, 11) in our experiment and are not thought to be directly involved with electron transport. Subunits PsaD and PsaE act as a docking site for an electron carrier out of PSI, although only *psaD* was significantly repressed by iron limitation in our experiment (Figure 2, row 6) (Lelong, et al., 1994; Xu, et al., 1994).

Under normal growth conditions, the PETC ends with the electron carrier ferredoxin docking with the PsaD and PsaE subunits to be reduced and become a substrate for the reaction catalyzed by ferredoxin-NADP reductase. This process reduces

$NADP^+$ to NADPH and further increases the thylakoid membrane electrochemical gradient. Alternatively, the reduced ferredoxin can transfer an electron back to the cytochrome b_6f complex via cyclic photophosphorylation (Fork & Herbert, 1993). However, under iron limitation, the iron-containing ferredoxin is partially replaced by iron-free flavodoxin (Sandström, et al., 2002; Entsch & Smillie, 1972). The replacement of iron-containing proteins with iron-free substitutes indicates a lack of intracellular iron, which was confirmed with metallomics by a decrease in the iron cellular content and Fe:C ratio of our iron-limited cultures [Rivas-Ubach, *et al., unpublished*]. In previous studies and in our experiment, the upregulation of flavodoxin (*isiB*) and its operon (*isiAB*) sharing gene *isiA* is one of the strongest signals (Figure 2, rows 13, 34) of iron limitation and can be viewed between chromosome positions 1,516,658-1,518,603 in Figures 1 and 3E (Singh, et al., 2003; Hernández-Prieto, et al., 2012). This upregulation of *isiAB* is accompanied by the upregulation of the copper-containing plastocyanin (*petE*) in our iron-limited cultures, as well as increases of intracellular copper content and the Cu:C ratio [Rivas-Ubach, *et al., unpublished*]. Hypothetical protein ssl0461 gene upregulation can be located near the *isiAB* operon between chromosome positions 1,525,350-1,515,601 in Figures 1 and 3E, and although ssl0461 has no functional designation it has been shown previously to be differentially expressed in PCC 6803 iron limitation (Singh, et al., 2003; Kopf, et al., 2014; Singh, et al., 2004).

F_1F_0 complex. The electrochemical gradient that is created across the thylakoid membrane by the PETC under normal conditions is a useful driving force to create usable energy for the cell. The F_1F_0 complex is a transmembrane protein complex that is bound within the thylakoid membrane and often referred to as ATP synthase or ATPase

depending on its activity. F_1F_0 is a highly conserved system that is present in the membranes of Archaea, Bacteria, and mitochondria. In phototrophs like Cyanobacteria and PCC 6803, F_1F_0 ATP synthase uses the photosystem produced electrochemical gradient by allowing H^+ ions to flow from higher concentrations within the thylakoid lumen out to lower concentrations in the cytosol (Steinberg-Yfrach, et al., 1998). As H^+ ions move through the F_1F_0 ATP synthase complex, it drives a conformational change that catalyzes the reaction that adds a phosphate group to ADP yielding ATP, the energy-carrying unit of the cell (Mitchell, 1961). F_1F_0 can also act as ATPase, pumping H^+ in the opposite direction across the membrane at the cost of phosphorylation via ATP.

In our experiment, genes for subunits AtpE, AtpH, and hypothetical protein sll1321 (Atp1) of the PCC 6803 F_1F_0 were repressed under iron limitation (Figure 2, rows 1-3). The epsilon subunit of the F_1F_0 complex (*atpE*) is part of the rotor portion of the F_1F_0 complex and may change conformation to act as a ratchet mechanism to inhibit ATPase activity, while preserving ATP synthase activity (Laget & Smith, 1979; Tsunoda, et al., 2001). The downregulation of the AtpE subunit gene could be a means to prevent inhibition of ATPase activity, allowing for active transport of H^+ ions into the thylakoid lumen to maintain the electrochemical gradient during PETC gene repression. An electrochemical gradient that exhibits a high pH in the thylakoid lumen is thought to participate in non-photochemical quenching, playing a role in photoprotection by dissipating energy into heat, a desirable effect during PSI and PSII stress-induced repression (Kramer, et al., 1999). The significant downregulation of *atpE* is in contrast to previous reports, which could be attributed to the differences in growth methodologies (Singh, et al., 2003; Hernández-Prieto, et al., 2012; Kopf, et al., 2014). The *atpH* gene is

translated into the c chain of the F_0 fraction of the F_1F_0 complex, of which multiple copies form the ring in which the F_1 fraction rotates. Referred to as the F_0c ring, this multimer is also part of the rotor movement of F_1F_0 and its downregulation is consistent with previous iron limitation work (Singh, et al., 2003). The hypothetical protein *sll1321* gene is part of the *atp1* operon along with *atpH* and both can be viewed as downregulated between chromosome positions 176,255-177,591 in Figure 1.

Carbon fixation. *Synechocystis* sp. PCC 6803 fixes inorganic carbon via the Calvin-Benson-Bassham Cycle, otherwise known as the "dark" or "light-independent" reactions. Only two genes for carbon fixation were differentially expressed in our iron-limited PCC 6803: the small subunit of ribulose-1, 5-bisphosphate carboxylase/oxygenase (RuBisCO), *rbcS* (downregulated) and the carbon dioxide concentrating mechanism protein subunit CcmO, *ccmO* (upregulated) (Figure 2, rows 4, 5). RuBisCO is a key enzyme involved in carbon fixation, responsible for carboxylation of ribulose-1,5-bisphosphate with intracellular CO_2. The observation of repressed RuBisCO small subunit gene *rbcS* is consistent with previous iron limitation work in PCC 6803 (Singh, et al., 2003) and our finding of reduced carbon biomass content [Rivas-Ubach, *et al., unpublished*]. *Synechococcus elongatus* sp. PCC 7942 in iron-deficient conditions has shown reduced protein levels of the RuBisCO large subunit (Michel, et al., 2003). In the leaves of sugar beets, the reduction of RuBisCO protein and mRNA levels during iron limitation were correlated with chlorophyll concentrations in the cell (Winder & Nishio, 1995). With the repression of photosystem expression, the supply of ATP and NADPH that drive the RuBisCO-mediated light-independent reactions would be diminished, leading to decreased RuBisCO carboxylase activity

(reviewed in (Ashraf & Harris, 2013)). The upregulated carbon dioxide concentrating mechanism proteins supply RuBisCO with high levels of intracellular CO_2 stores. This upregulation could be to compensate for decreased RuBisCO expression. Alternatively, increased CO_2 could be a means to maintain the electrochemical gradient of the cell, as carbon dioxide reacts with water to form carbonic acid, which dissociates into H^+ and bicarbonate.

Transport and ion binding proteins. An obvious effect of iron limitation on PCC 6803 is the upregulation of transport and ion binding proteins that can bring iron and other inorganic ions into the cell. Genes for the membrane-bound biopolymer transport complex ExbB-ExbD are essential for PCC 6803 inorganic iron uptake (Jiang, et al., 2015). Although the ExbB-ExbD complex genes (exbBD operon) have three sets of homologs within the PCC 6803 chromosome, one pair was among the most highly upregulated (Table 2, rows 35, 36) in our experiment (positions 32,524-33,544 in Figures 1 and 3F), consistent with previous work (Singh, et al., 2003; Hernández-Prieto, et al., 2012; Kopf, et al., 2014). Near the differentially expressed exbBD operon, the PCC 6803 cold stress response-linked (Suzuki, et al., 2001) unknown protein slr1484 gene was upregulated (Supplemental File S1). In our iron limitation treatment, three of the four Fut genes, *futA1, futA2,* and *futC* were upregulated (Table 2, rows 37, 38, 40), which is consistent with previous reports (Singh, et al., 2003; Hernández-Prieto, et al., 2012; Kopf, et al., 2014; Katoh, et al., 2001). Fe^{3+} is also acquired by PCC 6803 via the Fe^{3+} dicitrate transport system Fec; two of the five Fec subunit genes, *fecB* and *fecE,* were upregulated in our experiment (Table 2, rows 41, 42). These changes in expression of iron uptake

genes were correlated with decreased iron cellular content [Rivas-Ubach, *et al.*, *unpublished*].

Comparison with previous studies. With the multiple microarray experiments mentioned previously and the Kopf, *et al.* (2014) RNAseq experiment, the broad PCC 6803 iron limitation transcriptome response is well established. However, while some individual genes within the overall iron limitation response show reliable differential expression, there are some that do not. Of the 142 differentially expressed genes in our experiment, only 26 were consistently reported in the supplemental data tables of previous studies (Singh, et al., 2003; Shcolnick, et al., 2009; Hernández-Prieto, et al., 2012; Kopf, et al., 2014), either because of low significance or possibly absent probes in microarray experiments. Varying reports of individual gene regulation across multiple experiments is not uncommon. For example, transport and binding proteins in general have reliably shown differential expression as a result of iron limitation, however the individual gene slr0074 (*sufB*), which plays a role in Fe-S cluster repair and biogenesis (Seki, et al., 2006), has been shown to have $\log_2$ fold changes in expression of -1.66 (Singh, et al., 2003, possibly (B-A)/A fold change as equation and raw microarray readings are unreported), 0.46 (Shcolnick, et al., 2009, calculated from supplemental data), 1.23 (Hernández-Prieto, et al., 2012, reported in supplemental data), and 4.34 (Kopf, et al., 2014, calculated from supplemental data). Our results yielded a *sufB* $\log_2$ fold change of 0.63 which did not meet the $\leq$0.05 p-value thresholds of either a t-test or an exact test (described in the Methods and results available in Supplemental File S1). The genes of the *suf* operon have been shown to function during oxidative stress in the assembly of Fe-S clusters used in the electron transport carrier ferredoxin and in

photosystem I (Hernández-Prieto, et al., 2012; Nachin, et al., 2001; Takahashi & Tokumoto, 2002). Examples like this point out the importance of meta-analysis studies like that of (Hernández-Prieto, et al., 2016) but also suggest that more data is needed from both repeat iron-limitation experiments as well as alternate iron-limitation methods to consolidate our understanding of iron limitation in PCC 6803 and photosynthesis metabolism in general.

Conclusions

Synechocystis sp. PCC 6803 grown under iron limitation without iron chelation of growth media showed a strong signal of repressed photosynthetic electron transport chain genes, which likely had the effect of decreasing the influx of H^+ ions into the lumen and potentially disrupting the electrochemical gradient across its membrane. Maintaining this electrochemical gradient is important for many cellular processes including the photoprotection of PSI and PSII; thus, F_1F_0 and specific transport and binding proteins were differentially expressed to possibly favor maintenance of H^+ stores within the thylakoid lumen. Growth under iron limitation also caused the upregulation of iron acquisition genes in an attempt to meet metabolic iron demands.

As expected, the non-chelation iron limitation transcriptome responses of our experiment are generally in accordance with previous studies, with some differences in individual gene responses detected. Although our iron limitation resulted in widespread differential expression of genes in multiple functional categories, ranging from metabolite biosynthesis to transposon-related functions (Table 1), we centered our focus on photosynthesis and respiration along with transport and binding proteins. Each of these functional categories contained differentially expressed genes that show an apparent

iron limitation distress signal (Figure 2). However, some of the strongest individual signals were detected outside of these categories, in ribosomal proteins, hypothetical proteins, and unknown proteins (Supplemental File S1). The described iron limitation differential expression transcriptome regions in our experiment and in previous iron limitation studies are strengthened by their consistent reports from multiple iron limitation protocols. Differential expression that differs in our experiment from previous studies could be explained by the approaches taken for growth and data analysis, in which case further study is needed to determine if the response is truly caused by iron limitation or is an artifact of a particular methodology. The transcriptome results reported here complement a study on metabolomics and cellular composition from these same PCC 6803 cultures (Rivas-Ubach, *et al., unpublished*).

In addition, the high-resolution transcriptome mapping technique that we performed (Figure 1) can be applied to other microbial genomes. Genomes with a high level of synteny should display regions of differential expression that are clearly discernible to the naked eye, even more so than what is seen for the relatively low-synteny genome of *Synechocystis* sp. PCC 6803. Just as there are many ways to study transcriptomics, there are many ways to visualize the results. Studying and presenting data in with novel methodologies may help to elucidate new conclusions about repeated subjects.

Acknowledgements

We thank the editors and anonymous reviewers for their constructive comments. This work was supported by the NASA Astrobiology Institute at Arizona State University (08-

NAI5-0018) and NASA Exobiology and Evolutionary Biology Program (NNX08AP61G).

CHAPTER 3

USING CLUSTER EDGE COUNTING TO AGGREGATE ITERATIONS OF CENTROID-LINKAGE CLUSTERING RESULTS AND AVOID LARGE DISTANCE MATRICES.

Matthew Kellom and Jason Raymond

School of Earth and Space Exploration, Arizona State University, Tempe, Arizona, USA

Published in *Journal of Biological Methods*:

http://www.jbmethods.org/jbm/article/view/153

Abstract

Sequence clustering is a fundamental tool of molecular biology that is being challenged by increasing dataset sizes from high-throughput sequencing. The agglomerative algorithms that have been relied upon for their accuracy require the construction of computationally costly distance matrices which can overwhelm basic research personal computers. Alternative algorithms exist, such as centroid-linkage, to circumvent large memory requirements but their results are often input-order dependent. We present a method for bootstrapping the results of many centroid-linkage clustering iterations into an aggregate set of clusters, increasing cluster accuracy without a distance matrix. This method ranks cluster edges by conservation across iterations and reconstructs aggregate clusters from the resulting ranked edge list, pruning out low-frequency cluster edges that may have been a result of a specific sequence input order. Aggregating centroid-linkage clustering iterations can help researchers using basic research personal computers acquire more reliable clustering results without increasing memory resources.

Introduction

Agglomerative clustering is a useful tool to bin sequencing datasets based on sequence similarity, but the increasing use of high-throughput sequencing technology is creating datasets large enough to make clustering impractical for some computers and/or clustering methods. The most basic and widely used sequence clustering techniques are agglomerative, creating hierarchical bins *via* joining algorithms such as minimum-, maximum-, and average-linkage, with average-linkage being the most popular due to its perceived accuracy (Cole et al., 2009; Gronau & Moran, 2007; Huse, Welch, Morrison, & Sogin, 2010; Larkin et al., 2007). One drawback to these methods is that they require the construction of exhaustive distance matrices containing relative difference information between all possible pairwise sequence comparisons. After a distance matrix is constructed, the average-linkage algorithm bins sequences into clusters if the mean distance between all cluster member sequences is at or above the chosen clustering cutoff level, with minimum- and maximum-linkage using alternative binning requirements.

Distance matrix construction is a key computational bottleneck in agglomerative clustering. For large datasets, the computational needs of their distance matrices can exceed computer memory limits, especially for researchers using standard personal computers. Centroid-linkage clustering circumvents the need for a distance matrix at the cost of being input-order dependent, but this also makes the centroid-linkage algorithm faster and more memory-efficient for large-scale datasets than its agglomerative counterparts (Edgar, 2010). Since centroid-linkage clustering relies only on single pairwise sequence comparisons read in input file order, randomizing the order in which comparisons are made and centroids assigned can affect cluster-sequence distribution. A

graphical example of how sequence input order can affect cluster-sequence distribution can be found in Figure 1 of (Kellom & Raymond, 2016). This means that depending on the sequence input order, a specific cluster edge between two sequences may or may not form, affecting sequence-cluster membership. To address this challenge, some have considered ordering input sequences by length or abundance, with some programs employing these techniques natively, like CD-HIT (Fu, Niu, Zhu, Wu, & Li, 2012). Sorting sequences by length ensures that cluster centroids contain maximum information and thus cluster members can be binned more accurately. Conversely, abundance sorting approaches accuracy with the assumption that abundant sequences are more likely to represent functionally relevant clusters. However, both of these sorting methods still produce results that are dependent on a single, and to some degree, arbitrary input order. This is discussed further in the Discussion section.

Standard clustering concepts still apply to centroid-linkage, more closely related sequences are more likely to form an edge and be assigned to the same cluster. Over enough iterations of input randomization and clustering, edges that represent closely matched sequences will appear in the majority of iterations. By keeping track of all of the edges and ordering them by most frequently formed throughout the iterations, we can essentially form an ordered list of the most closely related cluster edges. From this ordered list of cluster edges, we can piece back together the clusters and make sure that sequences end up binned in clusters where they have the most representative cluster edge. The purpose of this protocol is to provide biology researchers without access to sufficiently high-performance computing with a means to obtain sequence clustering results that do not require the construction of large distance matrices while also not being

solely dependent on sequence input order. This process of random input order centroid-linkage clustering over multiple iterations, breaking down the resulting clusters into their individual edges, counting those edges, and then reconstructing aggregate clusters from a ranked edge list effectively bootstraps aggregate cluster edges from input-order dependent clusters and increases the reliability of centroid-linkage results.

This methodology is beneficial when the amount of available random-access memory (RAM) cannot contain the distance matrix being made, preventing agglomerative clustering processes from completing. For example, using traditional agglomerative clustering algorithms and a centroid-linkage algorithm in the program USEARCH (www.drive5.com/usearch/) allows for different limits on the maximum number of input sequences. Maximum-, minimum-, and average-linkage algorithms were only able to process ~10000 sequences past the distance matrix step on our 120 Gb RAM-containing computer, capacity beyond what is typically thought of for a standard computer. By eliminating the need for a distance matrix, the number of sequences that the centroid-linkage algorithm is able to process is only limited by the size of the file that can be read into memory (> 1000000 for our 120 Gb RAM computer). Importantly, these results do become input-order dependent. By avoiding distance matrices and writing edge lists and edge counts to text files in disk space (rather than storing in memory), the aggregation process is slower than agglomerative clustering but it is also more likely to finish before running out of necessary memory.

For comparison, the centroid-linkage algorithm was able to complete clustering of 10000 sequences in four seconds on our computer, while the minimum-, maximum-, and average-linkage algorithms each took eighteen seconds and the aggregation process took

an hour and twenty-two minutes. As the number of sequences in a dataset increases, the runtime of the aggregating algorithm increases drastically (detailed in Anticipated Results). The increased time is to be expected because not only is it waiting for multiple iterations of centroid-linkage clustering to complete, but it must also count and store all cluster edges. Although slower than average-linkage algorithms that use distance matrices for accuracy, this aggregation method is more likely to complete before running out of memory space. Likewise, as datasets and iterations increase, so does the amount of necessary disk storage. For our largest dataset of one million sequences over 101 clustering iterations, approximately 170 GB of data was written in the form of small individual text files. With this cost in speed and storage, aggregating multiple iterations of the efficient centroid-linkage algorithm increases the confidence of cluster-edge distribution for datasets that are too large to be clustered with comprehensive distance calculations.

Materials and Methods

The procedure outlined here includes the use of specific clustering and scripting programs but similar programs should work just as well. The choice of which programs is determined by user preference. The important details are to use a program that performs centroid-based clustering, or some other distance-matrix independent algorithm, and use a scripting language to perform the following aggregation procedure with the resulting clusters. The annotated Perl script used by the authors is supplied as Chapter 3 Supplemental File 1 (http://www.jbmethods.org/jbm/rt/suppFiles/153.). Kolmogorov-Smirnov comparisons between different clustering methods and the aggregation process were performed in R with ks.test of the R Stats Package (r-project.org).

Sequence indexing. Sequences are first given a numerical identifier (Sequence Numerical Identifier hereafter) by indexing the sequence order of the original input, avoiding potential downstream filename parsing errors. For the sake of speed, this index is stored in RAM as a hash table (Index Hash hereafter) with the sequence header as the key and the Sequence Numerical Identifier as the value (defined as hash{key} = value in Perl), but could be created and accessed in disk storage if desired. Typically, the amount of memory needed for this index is considerably smaller than what would be needed for a clustering distance matrix. It is very important during this indexing step for each of the input sequence headers to be unique so that later sequence header recall from their corresponding numerical identifiers can be done accurately. The sequences used to demonstrate the anticipated results originate from an unpublished metatranscriptome dataset with a mean sequence length of 98 bases and their origin is not important for the explanation of this methodology. Any natural dataset should yield similar clustering results to those seen in Figure 1.

Clustering. Over sufficient iterations (the authors here chose 101 iterations), clustering is performed with the USEARCH (version 8.0.1517_i86linux64) "-cluster_fast" command at a 0.95 clustering threshold and clusters are written to separate files using the "-msaout" command (Cluster Files hereafter) (Edgar, 2010). The authors here chose 101 iterations (counting from 0 to 100) because the results were stable at this number. In general, more iterations will lead to more stable results, and larger datasets will need more iterations. Determining the appropriate number of iterations is specific to each individual case. The USEARCH "-cluster_fast" command utilizes centroid-based clustering and avoids creating computationally costly distance matrices at the cost of

being input-order dependent. To mitigate the effects of input-order dependence, the sequence FASTA-formatted input file is first reordered randomly prior to clustering and downstream edge counting for each iteration. The Sequence Numerical Identifiers created in step 1 are not altered by the randomization process. Depending on the dataset, a smaller number of iterations may result in aggregate clusters that are dependent on those randomized clustering input files.

Edge compiling. After clustering has completed for the chosen number of iterations, Cluster Files are accessed to begin counting edges. Singleton clusters containing only one sequence and no edges are ignored by the counting process, and this minimum edge parameter can be increased to speed up the compiling/counting process at the cost of comprehensiveness. Singletons and low-edge-count clusters are not typically represented in large aggregate clusters.

To avoid storing edge counts in RAM, which can quickly reach capacity for large datasets in typical research personal computers, edges are written to files in disk storage (Edge File hereafter) with the numerically lesser Sequence Numerical Identifier as the filename of the Edge File (Hub hereafter) and the higher Sequence Numerical Identifier as a line in the Edge File (Node hereafter) so that a specific edge's count from the iterations can be obtained by counting the number of times a Node Sequence Numerical Identifier is found in an Edge File, this is important for the downstream edge counting.

Edge counting. For each compiled Edge File, the counts of specific Nodes for each Hub are stored in new files with filenames that represent their count (Count File hereafter). This counts the number of times a specific edge appears by writing the Hub and Node on a single line, never exceeding the number of chosen iterations.

Reconstruction. Aggregate clusters are reconstructed from the edges contained in Count Files, starting with the highest Count File (edges that were found the most in the iterations, typically equal to the number of iterations) and working down toward the lowest Count File. For the reconstruction algorithm, four hashes are created. First, the Index Hash created in step 1. Second, the inverse of the Index Hash, so that Sequence Numerical Identifiers are stored as keys and sequence headers as values (referred to as Inverse Index Hash in the algorithm below). Third, an aggregate cluster hash where keys are a numerical identifier assigned to clusters (Cluster Numerical Identifier hereafter) and values are lists of the sequence headers contained in each cluster (referred to as Aggregate Cluster Hash in the algorithm below). Fourth, a hash that tracks which Cluster Numerical Identifier (value) each hub and node are stored (key) (Tracking Hash in the algorithm below). For each edge of Hub and Node Sequence Numerical Identifiers, aggregate clusters are reconstructed using the following algorithm and then written to an output file:

1) Skip to the next edge if both the Hub and Node have already been assigned to clusters in the Tracking Hash.
2) If the Hub has already been assigned to a cluster in the Tracking Hash (implying with step 1 that the Node has not been assigned yet):
    a) Get the Cluster Numerical Identifier value that the Hub Numerical Identifier key has been assigned to in the Tracking Hash.

b) Get the sequence header value for the Node Numerical Identifier key from the Inverse Index Hash and append it to the value for the Cluster Numerical Identifier (from step 2a) key in the Aggregate Cluster Hash.

c) Append this Node Numerical Identifier key - Cluster Numerical Identifier value pair to the Tracking Hash.

3) If the Node has already been assigned to a cluster in the Tracking Hash (implying with step 1 that the Hub has not been assigned yet):

a) Get the Cluster Numerical Identifier value that the Node Numerical Identifier key has been assigned to in the Tracking Hash.

b) Get the Sequence Header Value for the Hub Numerical Identifier key from the Inverse Index Hash and append it to the value for the Cluster Numerical Identifier (from step 3a) key in the Aggregate Cluster Hash.

c) Append this Hub Numerical Identifier key - Cluster Numerical Identifier value pair to the Tracking Hash.

4) If neither the Hub nor Node have been previously assigned to a cluster in the Tracking Hash:
    a) Create an Aggregate Cluster Hash pair with a Cluster Numerical Identifier as the key and the sequence headers for the Hub and Node Numerical Identifiers from the Inverse Index Hash as the value.
    b) Append the Hub Numerical Identifier key - Cluster Numerical Identifier value to the Tracking Hash.
    c) Append the Node Numerical Identifier key - Cluster Numerical Identifier value to the Tracking Hash.
    d) Assign the next Cluster Numerical Identifier to be +1 greater than the current one (to create a new cluster).

This aggregating process is displayed as a flowchart in Figure 1.

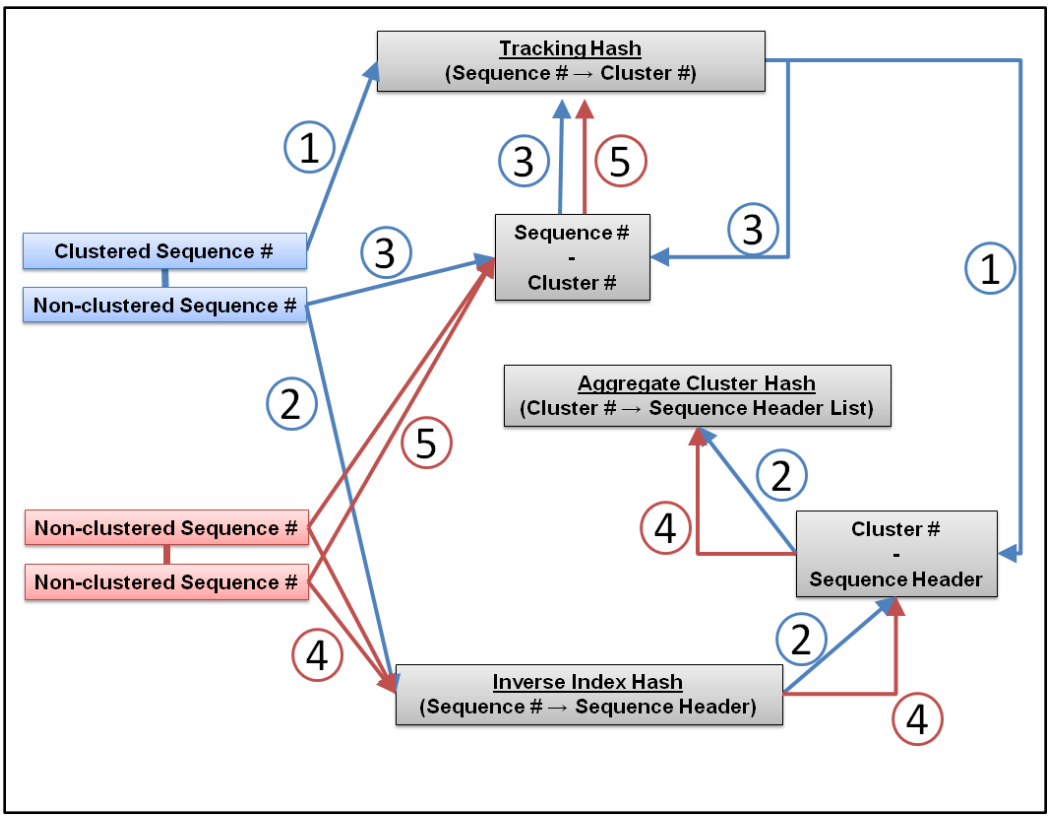

Figure 1: Flowchart of aggregating algorithm. Two scenarios are represented in the flowchart: One of the sequences for the current edge has already been assigned to a cluster from a previous edge according to the Tracking Hash (Blue); Neither sequence from the current edge has been assigned to a cluster according to the Tracking Hash (Red). A third scenario where both sequences of an edge have already been assigned to a cluster is not shown since that edge would be skipped in the algorithm. The processes in the flowchart have been numbered and described: (1) Using the Sequence Numerical Identifier of the already clustered sequence of the paired edge, obtain the Cluster Numerical Identifier from the Tracking Hash. (2) Using the Sequence Numerical Identifier of the non-clustered sequence of the paired edge, obtain its sequence header from the Inverse Index Hash and append it to the sequence header list value for the

Aggregate Cluster Hash key of the Cluster Numerical Identifier from 1). (3) Append the non-clustered Sequence Numerical Identifier and Cluster Numerical Identifier from 1) to the Tracking Hash to finalize it as a clustered sequence. (4) Both Sequence Numerical Identifiers of the non-clustered pair are used to obtain their sequence headers from the Inverse Index Hash and assign them to a new Cluster Numerical Identifier key in the Aggregate Cluster Hash. (5) Both Sequence Numerical Identifiers are paired with their Cluster Numerical Identifier and appended to the Tracking Hash.

Results

Each individual iteration of centroid-linkage clustering with randomized inputs should yield cluster distributions that are similar but not identical. Depending on the sequence input order, some sequences will not be clustered with the same matches for every iteration. Alternatively, some sequences will be so closely matched to other sequences that they will be grouped together in all or nearly all iterations. With enough iterations, the most prominent and closely-matched edges will appear more often than distant edges. Since these closely-matched sequences are likely to have edges that appear often, they will be among the first to be built into the aggregate clusters with the procedure outlined above.

Aggregating the results of many iterations of centroid-linkage clustering builds clusters from high-consensus edges while cutting out low-consensus edges. The edges are ranked from highest to lowest consensus which is then followed in the aggregation process. This process generally results in the aggregate maximum cluster size being smaller than some clusters of the individual iterations, especially for larger sequence datasets, as seen in Figure 2 for a dataset of one million sequences. The number of

clusters produced by the aggregation process and a single iteration of centroid linkage clustering is shown Table 1 for multiple dataset sizes, which includes the data plotted in Figure 2. Sequences of low-consensus edges that are trimmed out by the aggregating process are either binned to clusters where they are part of a higher-consensus edge or they are binned as a single-sequence cluster. However, the two cluster distributions remain the same, as shown with Kolmogorov-Smirnov test in Table 1. Total runtime (which includes the 101 iteration of clustering) for this one millions sequence dataset was 120:36:56 (Hours:Minutes:Seconds). For datasets of other sizes: 5000 sequences, 00:43:11; 10000 sequences, 01:21:46; 50000 sequences, 01:50:33; 100000 sequences, 03:58:32; 500000 sequences, 54:44:34.

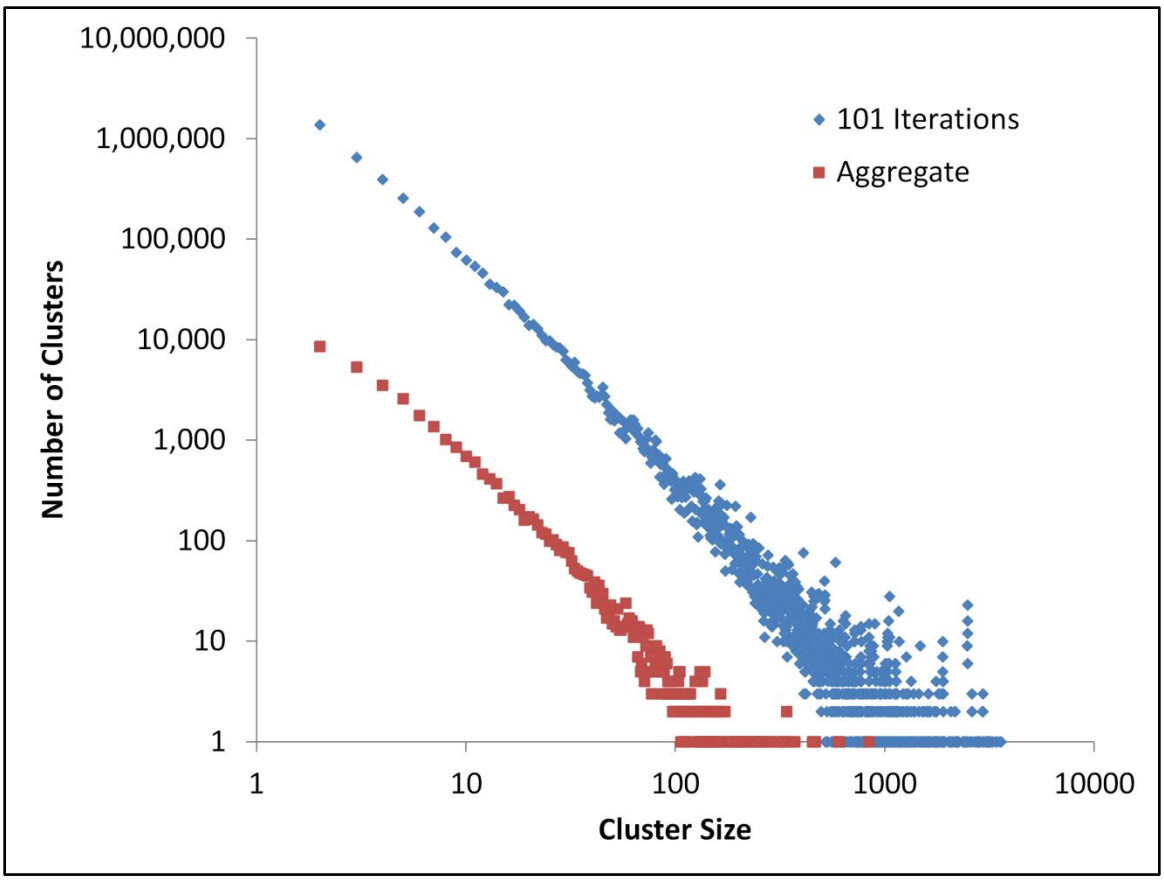

Figure 2: Cluster distributions of the individual iterations of centroid-linkage clustering (blue data points) and the aggregate clusters (red data points) for a dataset of one million sequences. Both axes are displayed in a logarithmic scale.

Table 1

Comparison of the number of non-singleton clusters between a single centroid-linkage iteration and the aggregate for datasets that range from 5,000 to 1,000,000 sequences. The fourth column is Kolmogorov-Smirnov D statistic comparisons between centroid (single iteration) and aggregate cluster distributions for the six dataset sizes, as well as the data plotted Figure 2. The 1,000,000 sequences dataset had an estimated *P* value of 0.2468, the 500,000 sequences dataset had an estimated *P* value of 0.4174, and all others had an estimated *P* value of 1, indicating for all datasets that the null hypothesis of the data having the same distribution cannot be rejected. Kolmogorov-Smirnov comparisons were performed in R with ks.test of the R Stats Package (r-project.org).

Dataset Size (sequences)	Centroid Iteration	Aggregate	Kolmogorov-Smirnov *P* value
5000	172	174	1
10000	423	424	1
50000	1155	1212	1
100000	2693	2899	1
500000	17456	20728	0.4174
1000000	37487	311326	0.2468

The cluster distribution of the aggregate clusters follows the same pattern seen in the individual iterations, suggesting that the aggregation process does not drastically alter the cluster distributions of the centroid-linkage iterations to the point of being unrepresentative, as seen in Figure 3. In contrast, minimum-, maximum-, and average-linkage clustering algorithms yield a cluster distribution that varies more substantially from the centroid-linkage algorithm in Figure 3. Table 2 shows Kolmogorov-Smirnov D statistics for pairwise comparisons between the cluster distributions shown in Figure 3.

The table shows that the centroid method distribution's least distant comparison is with the aggregate cluster distribution, with an estimated *P* value which does not allow us to reject the null hypothesis of having the same cluster distributions. This means that the aggregation process does reconstruct centroid-linkage cluster distribution instead of creating its own distinct cluster distribution. The data plotted in Figure 3 is also displayed in tabular format in Table 3.

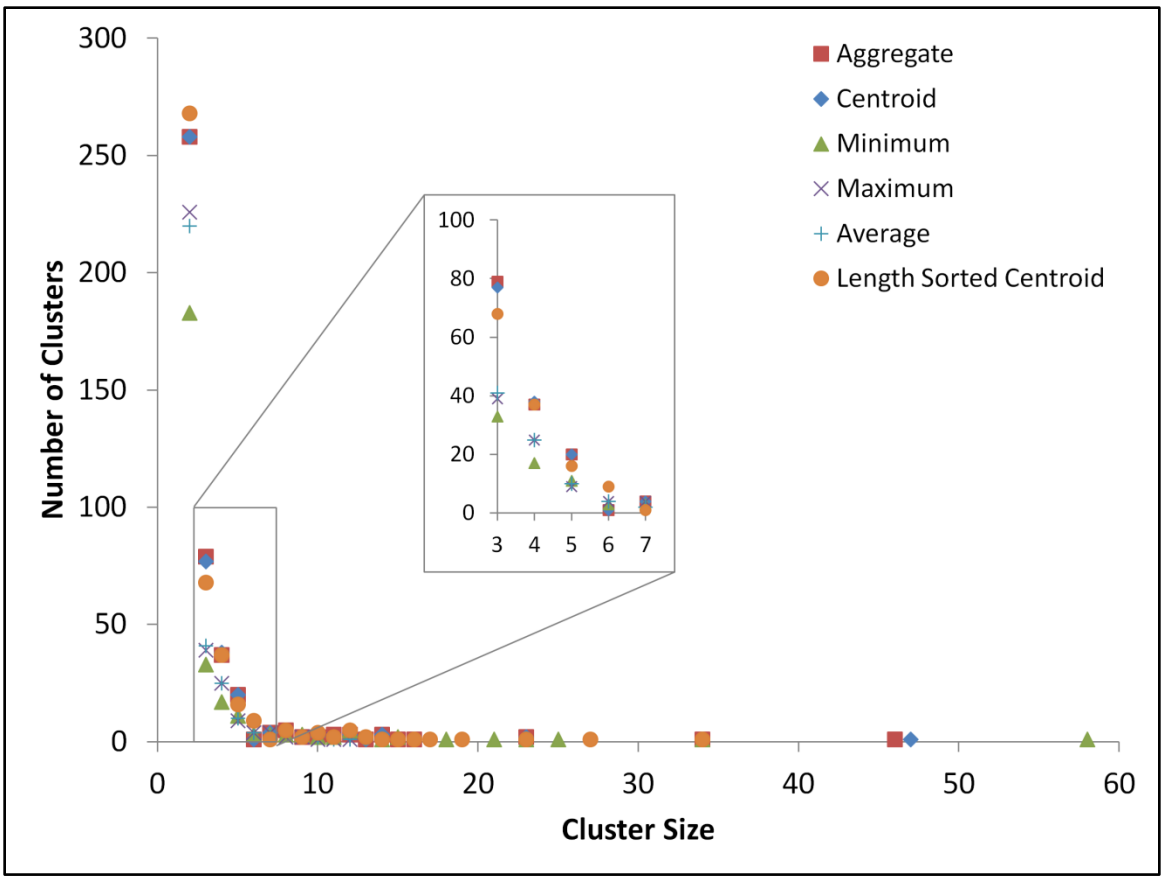

Figure 3: 10000 sequences dataset cluster distributions for the aggregated clusters of Figure 1, as well as single clustering runs of centroid-, minimum-, maximum-, and average-linkage algorithms from USEARCH. The graph displays counts of all non-singleton clusters. The x-axis shows the size of the clusters produced from the five different methods, i.e. the number of sequences in each cluster. The y-axis shows the number of clusters that were produced of the sizes displayed on the x-axis.

Table 2

Kolmogorov-Smirnov P value table for each pairwise comparison between results of the methods plotted in Figure 3. Kolmogorov-Smirnov calculations include singleton clusters, which are not plotted in Figure 3. No pairwise comparison estimated *P* value was smaller than 0.6284 (Minimum-Maximum comparison) meaning that the null hypothesis of the data having the same distribution cannot be rejected. Kolmogorov-Smirnov comparisons were performed in R with ks.test of the R Stats Package (r-project.org).

	Centroid	Aggregate	Minimum	Maximum	Average
Centroid	1				
Aggregate	1	1			
Minimum	1	1	1		
Maximum	0.7833	0.7833	0.6284	1	
Average	0.9103	0.9103	0.7833	1	1

Table 3

Tabular format of the data plotted in Figure 3.

Cluster Size	Aggregate	Centroid	Minimum	Maximum	Average
1	8606	8608	8997	9167	9135
2	258	258	183	226	220
3	79	77	33	39	41
4	37	38	17	25	25
5	20	20	11	9	10
6	1	1	3	4	4
7	4	4	3	4	2
8	5	5	3	2	5
9	2	2	3	2	2
10	2	3	2	1	2
11	3	2	2	1	1
12	3	3	5	1	1
13	1	1	0	0	1
14	3	3	1	0	0
15	1	1	2	0	0
16	1	1	0	0	0
18	0	0	1	0	0
21	0	0	1	0	0
23	2	2	1	0	0
25	0	0	1	0	0
34	1	1	1	0	0
46	1	0	0	0	0
47	0	1	0	0	0
58	0	0	1	0	0

As mentioned in the introduction, pre-sorting sequences by length ensures that cluster centroids contain maximum information and thus cluster members can be binned more accurately. Conversely, abundance pre-sorting approaches accuracy with the assumption that abundant sequences are more likely to represent functionally relevant clusters. The aggregation process that we introduce clusters sequences with their most frequent edge counterpart from multiple iterations of random input-order centroid

clustering. Our approach to accuracy is focused on the edges, using iterations of random input-order clustering to create a sorted, or ranked, edge list. Qualitatively, this has the effect of creating accurate clusters when presorting a sequence dataset by length/abundance is not sufficient or not possible.

As a simple example, a mock dataset of ten 100-base sequences populated *via* introducing one or zero random substitutions into a duplicate of the previous sequence was clustered using the aggregation process. In this dataset, listed below in FASTA format, with substitutions as capital letters, sequences mock0 and mock1 were identical, mock2 and mock3 were identical, and mock5 and mock6 were identical leading to a total of seven unique sequences. Sorting this mock sequence dataset by length or abundance does not yield a clear pre-sorted input. The aggregation process clusters mock0 and mock1 together and mock2-mock9 in a separate cluster. The edges between the sequences in these clusters occurred in 101/101 iterations of random input-order centroid clustering. Edges that connect the two clusters occurred in only 58/101 iterations, making them less of a priority in the aggregation algorithm. Length or abundance pre-sorting this mock dataset could yield either the single or double cluster distribution from the individual iterations depending on which sequence is chosen as the centroid sequence. Pre-sorting datasets with similar properties would yield clustering results that are close to a single random input-order iteration. Listed below are the mock DNA sequences described in the paragraph above.

>mock0

gaacaatgcattgtcattgctacaccgtttacatattacagagctttgcgcataagttcaacagcaccctggtcagctagagcacga

tagcgcagcccct

>mock1

gaacaatgcattgtcattgctacaccgtttacatattacagagctttgcgcataagttcaacagcaccctggtcagctagagcacga

tagcgcagcccct

>mock2

gaacaatgcattgtcatAgctacaccgtttacatattacagagctttgcgcataagttcaacagcaccctggtcagctagagcacg

atagcgcagcccct

>mock3

gaacaatgcattgtcatAgctacaccgtttacatattacagagctttgcgcataagttcaacagcaccctggtcagctagagcacg

atagcgcagcccct

>mock4

gaacaatgcattAtcatAgctacaccgtttacatattacagagctttgcgcataagttcaacagcaccctggtcagctagagcac

gatagcgcagcccct

>mock5

gaacaatgcattAtcatAgctacaccgtttacatattacagagctttgcgcataagttcaacagcaccctggtGagctagagcac

gatagcgcagcccct

>mock6

gaacaatgcattAtcatAgctacaccgtttacatattacagagctttgcgcataagttcaacagcaccctggtGagctagagcac

gatagcgcagcccct

>mock7

gaacaatgcattAtcatAgctacaccgtttacatattacagagcCttgcgcataagttcaacagcaccctggtGagctagagca

cgatagcgcagcccct

>mock8

gaacaatgcattAtcatAgctacacAgtttacatattacagagcCttgcgcataagttcaacagcaccctggtGagctagagc
acgatagcgcagcccct

>mock9

gaacaatgcattAtcatAgctTcacAgtttacatattacagagcCttgcgcataagttcaacagcaccctggtGagctagagc
acgatagcgcagcccct

Discussion

Since this aggregation process sacrifices speed to use less memory than agglomerative clustering while improving centroid-linkage clustering, the method can be much slower for large datasets. In addition, since data is written to disk storage instead of RAM, large datasets can require a large amount of available disk space, as mentioned in the final paragraph of the Introduction section. While the lengthier completion time and large amount of required disk space are drawbacks to this method, the aggregation process will eventually finish if these conditions are acceptable to the user.

Alternative methods for improving centroid-clustering results include presorting the input sequences either by length, unique sequence abundance, or combination of the two (Edgar, 2010; Ghodsi, Liu, & Pop, 2011). Figure 4 shows a comparison of the cluster distribution for the aggregated clusters, randomly sorted centroid-linkage, and length sorted centroid-linkage (sorted with the –sort option in USEARCH). Figure 4 and Table 4 (which shows the data in tabular format) show the cluster distribution from the aggregation process is closer to the distribution of the randomly sorted centroid-linkage than the length sorted, although not significantly so. However, both of these sorting methods (length and abundance) still produce results that are dependent on a single, and to some degree, arbitrary input order, while the aggregating process attempts to find the

average result of many possible input orders. A possible middle ground would be to incorporate the results from presorted clustering to weight the aggregation inputs with as many iterations of presorted cluster distributions as desired. For example, if a user wanted to make sure that length sorted centroid-linkage was represented in the final aggregated cluster distribution, they could include length sorted results in place of one or more of the randomly sorted iterations. Unfortunately, just as between length and abundance sorted methods, it is difficult to say which method is definitively ‘better’ for most datasets.

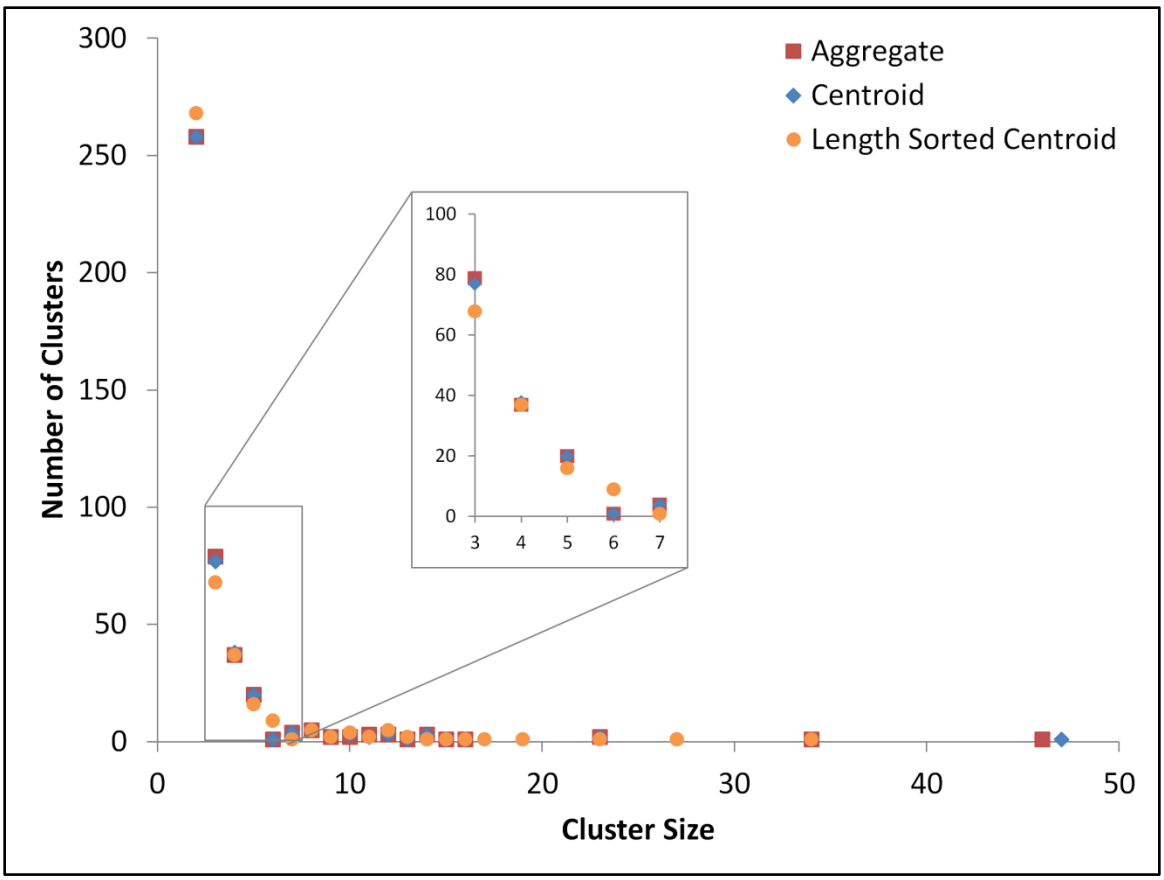

Figure 4: 10000 sequences dataset cluster distributions for the aggregated clusters of Figure 1, as well as single clustering runs of centroid- and length sorted centroid-linkage algorithms from USEARCH. The graph displays counts of all non-singleton clusters. The x-axis shows the size of the clusters produced from the five different methods, *i.e.* the number of sequences in each cluster. The y-axis shows the number of clusters that were produced of the sizes displayed on the x-axis. All pairwise comparisons between results of the methods plotted in Figure 4 had Kolmogorov-Smirnov *P* value of 1, meaning that the null hypothesis of the data having the same distribution cannot be rejected.

Table 4

Tabular format of the data plotted in Figure 4.

Cluster Size	Aggregate	Centroid	Length sorted centroid
2	258	258	268
3	79	77	68
4	37	38	37
5	20	20	16
6	1	1	9
7	4	4	1
8	5	5	5
9	2	2	2
10	2	3	4
11	3	2	2
12	3	3	5
13	1	1	2
14	3	3	1
15	1	1	1
16	1	1	1
17	0	0	1
19	0	0	1
23	2	2	1
27	0	0	1
34	1	1	1
46	1	0	0
47	0	1	0

In conclusion, Aggregating randomly sorted centroid-linkage clustering results into a single distribution mitigates the consequences of input-order dependence in centroid-linkage clustering. The process described here primarily uses disk storage instead of RAM, which can have the consequences of long run times and requiring a large amount of available disk space. However, these consequences may be acceptable to researchers using a dataset that is too large for the distance matrices of agglomerative clustering methods. Centroid-linkage circumvents the need for constructing large distance

matrices at the cost of input-order dependence. Methods exist to correct for this input-order dependence, such as presorting input sequences by length, unique sequence abundance, or combination of the two. While these methods may improve on the results of a single randomly sorted input order, they still represent a single, and to some degree, arbitrary input order. By aggregating the results of many randomly sorted iterations of centroid-linkage, the final result will not be dependent on any single input order. This method provides an alternative to the results from presorted centroid-linkage clustering.

Acknowledgments

This work was supported by the NASA Astrobiology Institute at Arizona State University (Follow the Elements; NAI5-0018).

CHAPTER 4

USING DENDRITIC HEAT MAPS TO SIMULTANEOUSLY DISPLAY GENOTYPE DIVERGENCE WITH PHENOTYPE DIVERGENCE.

Matthew Kellom and Jason Raymond

School of Earth and Space Exploration, Arizona State University, Tempe, Arizona, United States of America

Published in *PLOS One*:

http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0161292

Abstract

The advancement of techniques to visualize and analyze large-scale sequencing datasets is an area of active research and is rooted in traditional techniques such as heat maps and dendrograms. We introduce dendritic heat maps that display heat map results over aligned DNA sequence clusters for a range of clustering cutoffs. Dendritic heat maps aid in visualizing the effects of group differences on clustering hierarchy and relative abundance of sampled sequences. Here, we artificially generate two separate datasets with simplified mutation and population growth procedures with GC content group separation to use as example phenotypes. In this work, we use the term phenotype to represent any feature by which groups can be separated. These sequences were clustered in a fractional identity range of 0.75 to 1.0 using agglomerative minimum-, maximum-, and average-linkage algorithms, as well as a divisive centroid-based algorithm. We demonstrate that dendritic heat maps give freedom to scrutinize specific clustering levels across a range of cutoffs, track changes in phenotype inequity across multiple levels of sequence clustering specificity, and easily visualize how deeply rooted changes in phenotype inequity are in a dataset. As genotypes diverge in sample populations, clusters are shown to break apart into smaller clusters at higher identity cutoff levels, similar to a dendrogram. Phenotype divergence, which is shown as a heat map of relative abundance bin response, may or may not follow genotype divergences. This joined view highlights the relationship between genotype and phenotype divergence for treatment groups. We discuss the minimum-, maximum-, average-, and centroid-linkage algorithm approaches to building dendritic heat maps and make a case for the

divisive “top-down” centroid-based clustering methodology as being the best option visualize the effects of changing factors on clustering hierarchy and relative abundance.

Author Summary

The visualization of sequencing data is an integral part of the analysis and communication of genomics-based research. A key advance in microbial ecology in both modern and ancient ecosystems will be connecting genotypic lineages and survival strategies to environmental dynamics. This has become a daunting task in light of the burgeoning repositories of -omics sequence data, calling on entirely new methods for analyzing and visualizing complex biogeochemical datasets. The effects of environment on biology are often shown as heat maps of sequence abundance, where the responses of distinct sequence groups are measured and compared. However, the grouping process of heat map construction is performed at a single, often arbitrary, level of inclusiveness. Here, we introduce dendritic heat maps that simultaneously display multiple heat maps over a range of binning specificities, arranged in a dendrogram-like configuration, to show the effects of environment on sequence homology and relative abundance. Importantly, tracking changes in relative abundance can be particularly useful for observing the levels at which genotypic divergence (cluster branching) correlates with gene expression (differing heat map bin response), helping to better understand the effects of environment on survival strategies and genotypic lineages.

Introduction

Advances in sequencing technology and –omics research has led to rapid growth in sequencing datasets, and techniques to visualize and analyze the data are struggling to keep up. New avenues of research that expand on traditional techniques are being explored with much room for further advancement, as software development attempts to meet the demands of elucidating important aspects of such large and complex datasets

like metagenomes or metatranscriptomes. One of the most fundamental steps for analysis of any sequence dataset is annotation and classification into hierarchies, which can be achieved via sequence comparison tools such as USEARCH, the Ribosomal Database Project (RDP) Classifier, the Basic Local Alignment Search Tool (BLAST), RapSearch2, and Phylosift (Camacho et al., 2009; Cole et al., 2014; Darling et al., 2014; Edgar, 2010; Zhao, Tang, & Ye, 2012). While all of these tools do well at the annotation of sequences, there is a well-known classification bias that comes with limited databases that do not contain true representatives of every sequence (Wooley, Godzik, & Friedberg, 2010). The development of more efficient and accurate comparison tools is an area of active research, and understanding the results of these tools in the context of *in vivo* dynamics is of great interest. MEGAN is one of the more well-known software options for metagenomic analysis and visualization of sequence comparison results, as are some web-based platforms such as MG-RAST, METAREP, and Krona (Goll et al., 2010; Huson, Auch, Qi, & Schuster, 2007; F. Meyer et al., 2008; Ondov, Bergman, & Phillippy, 2011). The fundamentals of traditional techniques such as heat maps and dendrograms are at the root of all these recent software advances, where hierarchy and relative abundance are represented through branching and value indicators, respectively. Here, we explore a method of creating dendritic heat maps (DHMs) that combines heat maps and dendrograms in order to visualize phenotype divergence alongside genotype divergence. We use the term phenotype to represent any feature by which groups can be separated (e.g. physical traits, locations, growth conditions, etc.). Importantly, DHMs are not limited to sequence data and can be used to describe changes in group inequity and clustering hierarchy for any data that can be hierarchically ordered and compared (e.g.

microarrays, phylotype counts, species richness, etc). However, here we exclusively examine their application toward sequence datasets.

Heat maps are useful for comparing data across a range of possible states, allowing viewers to intuitively see differences and similarities in data subset responses. Canonically, each data point is expressed as a color, with hue intensity representing its bin response, or position within the data range (Wilkinson & Friendly, 2009). Likewise, sections on a heat map can also represent bins of data, where individual data points have been grouped together based on similarity and their corresponding heat map color is a result of their combined response (Sneath, 1957). Clustering, especially of DNA, RNA, or protein sequences, is commonly used for data binning and is based on sequence identity or homology. For instance, 16S rRNA heat maps are frequently used to compare relative abundances of sequences between multiple samples, allowing visualization of the presence and absence of taxa across samples or populations (Cho et al., 2012; Koenig et al., 2011; Wu et al., 2011). With genomic and transcriptomic data, binning and visualization on heat maps can be used to compare and contrast gene and transcript abundances, with the most common use of heat maps being visualization of changes in gene expression across different sample treatments or conditions (Nodine & Bartel, 2012; Schloissnig et al., 2013; The Cancer Genome Atlas Research Network, 2013).

However, heat map bin response for a data point can change depending on the level of specificity, defined by the cutoff level at which that data is binned and visualized. By altering the binning specificity level, data bin assignment can be rearranged, potentially changing their heat map bin response. Traditional heat maps only work at a single specificity level and limit viewers to one representation of the data. For instance,

heat maps that depict clustering of 16S rRNA gene sequences are typically done at a 97% identity level. Choosing an appropriate specificity level then becomes crucial to not obscuring the important features of the data with bins that are too specific or too broad. For example, bins that are too specific for a transcriptome dataset can result in fractured count information with many clusters identified as the same target sequence. Conversely, bins that are too broad for a transcriptome dataset can result in clusters containing multiple transcript groups. Both of these scenarios can be problematic if only one cutoff level is being displayed.

Our motivation for this work is to visualize sequence dynamics in a way that captures important variations in the data and is scalable across a large range of data sizes. We also wanted to explore a technique that is independent of annotation and instead performs analysis and visualization of sequence information, leaving annotation as a final step, since sequence reference databases are dependent on the quality and focus of previous work. To these ends we improve on current techniques in three areas:

1. Freedom to scrutinize specific clustering levels across a range of cutoffs.
2. Ability to track changes in state across multiple levels of sequence clustering specificity.
3. Ease to visualize how deeply rooted changes in state are in a data set.

DHMs are particularly useful where similarities in a dataset occur across a multitude of scales, such as in homology-based clustering of the large number of sequences found within a microbial community. Because the sequences within a complex

community can range from being extremely well conserved to very poorly conserved, our method will allow the simultaneous visualization of homology clusters at many different cutoffs. Importantly, tracking changes in relative abundance bin response can be particularly useful for observing the levels at which genotypic divergence (cluster branching) correlates with phenotypic divergence (differing heat map bin response) for a population.

To demonstrate this approach, we generate artificial datasets that use simplified mutation and growth processes in biological communities. The first dataset starts with 100 identical 100-bp DNA fragments which all mutate with random single base substitutions over fifteen iterations. The second dataset is used exclusively with the "top-down" method due to its size and begins with a 100-bp DNA fragment, allowing it to mutate and duplicate, then iterating the mutate-and-duplicate process on progeny DNA. The end result is a population of tens-of-thousands of DNA molecules derived from a common ancestor, each showing varying degrees of conservation. Clustering and visualization of changes of state are used to track relative abundance bin responses for populations of different nucleotide usage (GC content) as various subpopulations evolve for both datasets. Using these simulated datasets, we discuss the potential of DHMs to describe data across varying levels of complexity.

Methods

Sequence Generation.

Mutation Lineage Data Set. Random base substitutions were performed on a set of 100 artificially-generated 100-bp DNA sequences over fifteen iterations. Base substitutions are allowed to occur at the same position more than once. At iteration zero,

all 100 sequences are identical with 50% GC content; at iteration fifteen they have the least homology with 15 random base substitutions having occurred in each sequence. At each iteration all 100 sequences were grouped based on ≤ or >50% GC content and output to FASTA files. The ≤ or >50% GC content group sizes were kept relatively even by restarting the sequence generation if the count difference between the two groups exceeded 20 sequences (|Group1-Group2|>20) (20% of the total dataset). This group evening was done to ensure a good representation of each group for an effective demonstration of DHMs.

Population Growth Data Set. Additional FASTA files of fifteen generations started from a single randomly-generated sequence, which was subsequently propagated by duplicating each sequence once with a random base substitution and once without at each generation of population growth, reaching 2^{15} sequences after fifteen generations. The artificial growth process created fifteen separate, but related, datasets to demonstrate the ability of top-down centroid-based clustering to handle larger datasets for DHM construction. The artificially-generated 100-bp DNA sequences were grouped based on ≤ or >50% GC content. The ≤ or >50% GC content group sizes were kept relatively even by restarting the sequence generation if the difference between the number of sequences for the two groups became greater than 5% of the total amount of sequences (|Group1-Group2|>Total*0.05). This group evening was done to ensure a good representation of each group for an effective demonstration of DHMs.

Clustering.

Bottom-up agglomerative clustering. For each incremental 0.01 fractional identity ("-id") cutoff between 0.75 and 1.0, clustering of sequences was performed using USEARCH (version 8.0.1517_i86linux64) ("-cluster_agg") with linkages min, max, and avg.

Top-down divisive clustering. For the each 0.01 -id cutoff between 0.75 and 1.0, clustering of sequences was performed using the USEARCH UCLUST algorithm (version 7.0.1090_i86linux64) ("-cluster_fast"), which performs centroid-based clustering. Counts of duplicate sequences were recorded with "-derep_fulllength" (Edgar, 2010). Clustering at each cutoff -id was done in a stepwise fashion, starting from the initial FASTA file for 0.75, then using the clusters from the previous cutoff as inputs for 0.76-1.0. Since the USEARCH command "-cluster_fast" performs centroid-based clustering in the order of the input FASTA file, input sequences were first multiple aligned using Clustal Omega (described later) and arranged to ensure that the most distantly related and potentially cluster splitting "centroid" sequences were listed first in a staggered order (conceptualized in Figure 1). This ordering process was performed by a script that reads from opposite ends of the multiple alignment and is unnecessary for the "bottom-up" approaches since binning is done through the use of a distance matrix. A brief example of the top-down clustering is available in Chapter 4 supplemental file S1.

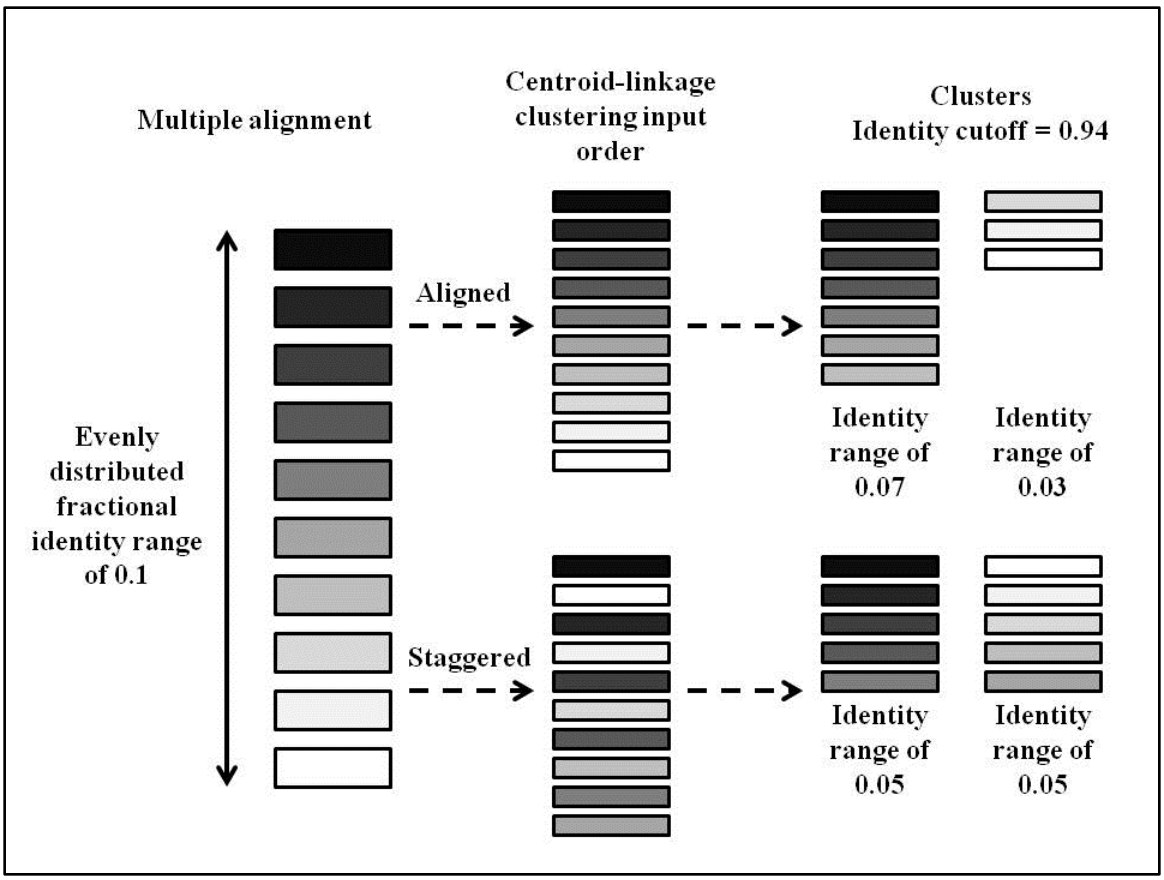

Figure 1: Sequence input ordering. Graphical representation of the binning effect of using alignment-ordered versus staggered sequence input order for "top-down" centroid-based clustering. Shaded rectangles represent sequences, where the shade consistently portrays a specific sequence throughout the diagram. The multiple alignment on the left shows each of the sequences ordered based on fractional identity, where nearby sequences are more closely related than distant ones, and distributed evenly across a fractional identity range of 0.1. For both aligned and staggered input ordering, sequences are read from top to bottom by the UCLUST algorithm of USEARCH and either placed in a cluster that has the best match to the centroid sequence above the given identity cutoff, or is made the centroid sequence of a new cluster if a match cannot be found. In this diagram, centroid sequences are the top sequences of each cluster. With the aligned

input order, it is shown that some sequences can be binned in clusters that do not contain their closest centroid match. The staggered input places sequences in correct bins essentially by first defining all centroid sequences.

Alignment. Clustal Omega (Version 1.2.0) with tree-output ordering and default alignment parameters and heuristics was used for multiple alignment of FASTA files (Sievers et al., 2011). Manipulation of the multiple alignment and clustering files were performed via Perl scripts, which are available in Chapter 4 supplemental files S2 and S3. This multiple alignment was used to order the lowest -id cutoff (0.75) clusters, then the sequences contained within them. For each -id cutoff from 0.76 to 1.0, the order was determined by the arrangement of the previous -id cutoff. In the "top-down" method, this alignment was also used to order the clustering input files in a staggered fashion (Figure 1). Alignment is important for preserving the radial position of each sequence at each cutoff level/ring in the DHMs so that the position of any given sequence is preserved from its center out to its circumference. Starting the clustering cutoff range at a minimum value of 0.75 was done because the USEARCH manual states that the UCLUST algorithm is effective at identities of ~75% and above for nucleotide sequences, but dendritic heat maps in general are not limited to this cutoff range and should aim to show as large a clustering cutoff range as possible.

Dendritic heat map construction. Visualization of the clusters at each cutoff and their arrangements was performed using the Perl package Circos (version 0.64) (Krzywinski et al., 2009). For each -id cutoff ring, cluster sizes are determined by number of sequences within each cluster. The heat map hues are partitioned to have gradual changes with twenty-three possible categories (two sequential 11-color Brewer palettes

and white) and is determined by the logarithmic value of the ratio of sequences from each group, log((Group1 +1)/(Group2 +1)). The log of the ratio is used because in typical clustering datasets, the vast majority of the clusters are relatively small and a few are very large. Without using the log of the ratio, smaller clusters with an interesting bin response would be assigned a neutral hue and only the largest clusters would have the most luminous hues. A log transformation of the ratios puts all bin responses on a more relatable scale while still showing their distinctions in ratio. Red hue indicates relative abundance bin response toward Group 1(GC≤50%) and blue hue indicates relative abundance bin response toward Group 2 (GC>50%), with white indicating a neutral bin response. A red-white-blue color scheme was chosen here because it is more color-blind friendly, however any color scheme can be used. Hue luminosity corresponds to the strength of the heat map response. A darkly colored wedge at the 0° position acts as a key, displaying the opposite ends of the heat map hues possible for each ring. The minimum and maximum heat map values in this wedge are equal at their absolute values and are important for normalizing the hue distribution throughout the entire DHM. A brief example of DHM construction is available in Chapter 4 supplemental file S1.

Results/Discussion

Dendritic heat map. To make relationships that may emerge across hierarchical cutoffs more apparent, heat maps at each cutoff level are aligned, so that clusters may be directly compared across multiple cutoff levels. By aligning and clustering DNA molecules across these multiple cutoff levels, the aligned heat maps take on a radial dendrogram configuration. This is particularly useful, as the branching of a cluster into progressively more fine-grained clusters can be tracked and further annotated with heat

map bin responses to reveal salient features of the clusters. To visualize DHMs, we make use of a Perl-based software package called Circos that was developed to address challenges in visualizing large genomic data sets and creates circular heat maps from position and value data (Krzywinski et al., 2009). In the case of DHMs, circular heat maps allow for the placement of exterior ring heat map bins to be fanned out, giving them more space than interior ring bins where the need for space is less critical. While a useful tool for generating images, Circos is illustration software that is dependent on user-formatted input files and is not designed to analyze raw data or arrange heat maps. Therefore, we have developed Perl scripts (available in Chapter 4 supplemental files S2 and S3) to align and convert clustering data derived from large scale sequence datasets to Circos-ready input files.

Figure 2 shows the average-linkage DHM for the fifth mutation from the mutation lineage data set generated as described in the methods section. The feature that is immediately recognized is the heat map color variation in different regions of the figure, displaying the relative abundance bin response of GC content groups (neutral – white; Group 1 bin response (GC≤50%) – red; Group 2 bin response (GC>50%) – blue) for each cluster. The decision to use red and blue hues was aided with the use of ColorBrewer palettes (http://colorbrewer2.org/) to represent values as a colorblind-friendly alternative to the red-green color scheme that is a popular heat map motif (Brewer, 2003). However, RGB color codes were eventually used to select hues outside of ColorBrewer palettes. Navigating through the figure, the innermost ring represents clusters at the most lenient fractional identity cutoff of 0.75, stepping out in increments of 0.01 at each ring to a final identity cutoff of 1.0. Simultaneously displaying a range of heat maps that change with

specificity level gives a more accurate view of how the data is skewed than any single heat map can provide individually, due to the fact that heat map bin response for a data point can change direction depending on the level of specificity at which the data is binned. In a natural data example, central rings with lower cutoffs would be associated with broad groupings such as phyla or gene superfamilies, while the increasingly strict cutoffs at the peripheral rings would represent more specific identifiers (e.g. same genus/species, gene families/subfamilies). Studies that involve natural data sets might not be grouped based on GC content, but rather some separation that is relevant to the questions being asked or conditions being measured (and that could be easily substituted into a DHM). For our purposes, using GC content as a phenotype is a convenient method for creating and tracking relative abundance bin responses of two distinct groups while also forcing a correlation between genotype and phenotype for these artificially generated datasets.

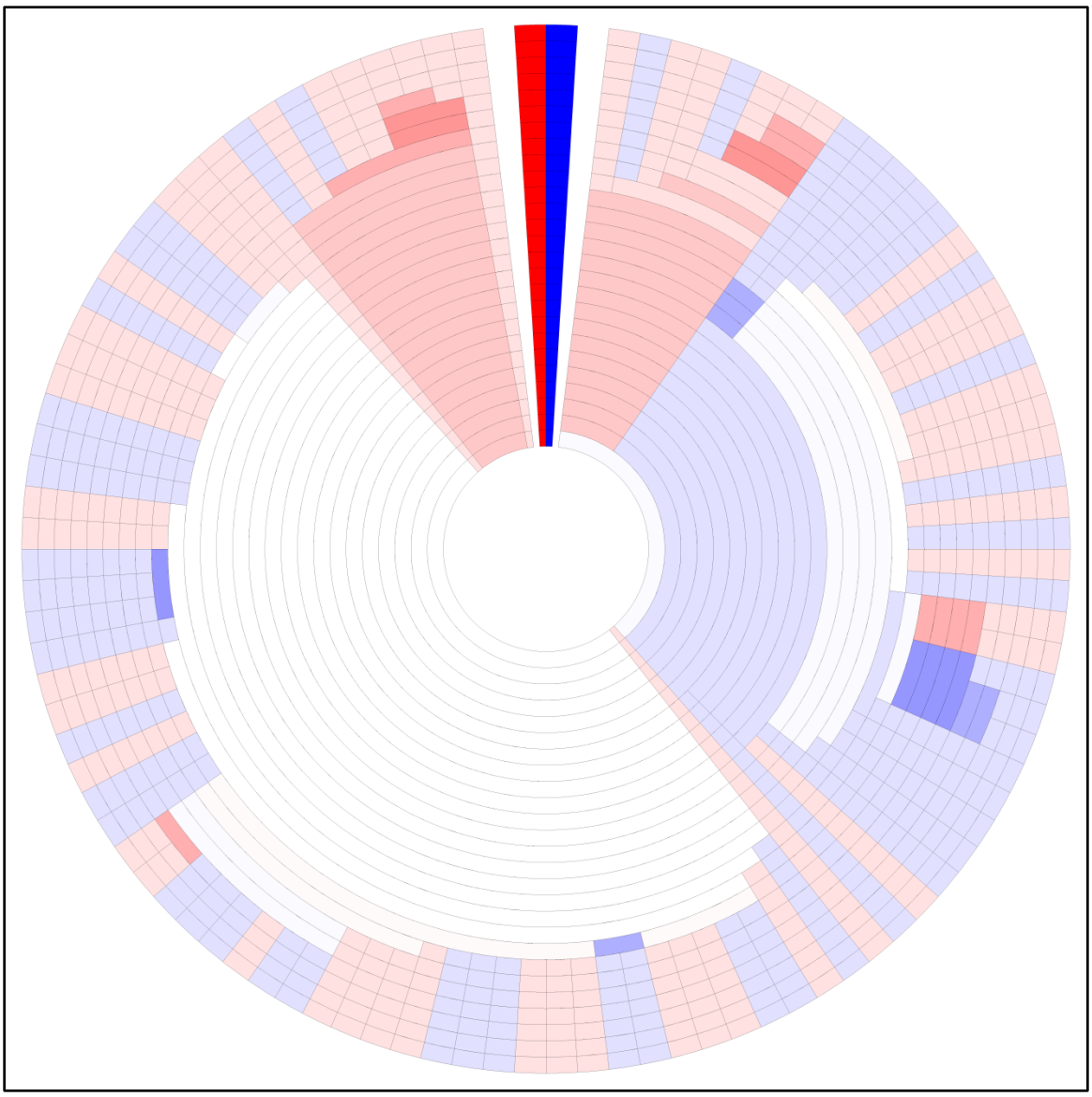

Figure 2: Dendritic heat map. Dendritic heat map representing the fifth mutation step of the simulated mutation lineage data generated as described in the methods and clustered using the average-linkage algorithm of the "bottom-up" method. The darkly colored wedge at the 0° position represents the minimum (red) and maximum (blue) possible heat map relative abundance bin responses, GC≤50% and GC>50% respectively. White space in the heat maps represents clusters with neutral bin response. Rings, starting at the center, represent clusters of sequences for identity cutoffs of 0.75 to 1.0. Clusters,

including single-sequence clusters, are plotted in a radial range that is conserved from the clusters from which they were derived. High resolution versions of all DHMs in this manuscript are available in Chapter 4 supplemental files S2 and S3.

The more subtle feature of Figure 2 is the dendrogram-like layout of the figure rings. The aligned configuration of the rings preserves the relationships of a dendrogram, where nearby clusters and sequences are closer relatives than distant clusters. The more central in the diagram a cluster divergence occurs, the more distantly related those clusters are to one another. Since each ring is aligned to adjacent rings, large clusters gradually divide into smaller and more specific clusters moving out from the center. In Figure 2, many of the clusters in the exterior rings contain single-sequence clusters that have a conserved radial range from the more interior clusters they derive from. Some clusters contain sequences that are highly conserved and their membership does not change over a wide range of clustering cutoffs, as seen at around the 11 o'clock position of Figure 2. Also important to point out are cases where bin response changes in direction and not just intensity. At about the 3 o'clock position of Figure 2, interior rings contain a large blue (mostly GC>50% sequences) cluster that gradually divides toward the exterior rings. These clusters divide, some of their bin responses change from blue (mostly GC>50% sequences), to the neutral white, to red (mostly GC≤50% sequences). The darkly colored wedge at the 0° position represents the minimum (red) and maximum (blue) possible heat map relative abundance bin responses, GC≤50% and GC>50% respectively. This wedge at the 0° position serves two functions. First, the wedge serves as a legend for the minimum and maximum heat map bin responses. Second, the wedge

separates the most distantly related clusters at opposite ends of the DHM. Without a separation, it may be easy to confuse the clusters at opposite ends of the dendrogram layout as closely related. White space in the heat maps represents clusters with neutral bin response.

Overlaying phenotype shifts on genotype divergence creates a way to visually compare how deeply rooted observed phenotype ratios are across multiple genotypes. Displaying how deeply rooted a response is can be informative in many comparative studies that seek to better understand both the general and more finely detailed structures of the data by elucidating divergence points. Work that focuses on multi-scale genomic changes, such as experimental evolution of microbial or viral populations, would benefit from the visualizations of DHMs (Blount, Barrick, Davidson, & Lenski, 2012; J. R. Meyer et al., 2012).

Bottom-up hierarchical clustering. For agglomerative, or "bottom-up," approaches where clusters are joined by incrementally decreasing the sequence identity required to bin sequences together, we contrasted minimum-, maximum-, and average-linkage algorithms, all common graph metrics. Briefly, these methods differ in how connections between cluster elements (i.e. the edges connecting various nodes within a cluster) affect cluster membership. Minimum-linkage, sometimes referred to as nearest neighbor, only requires a single edge between two clusters above a specified cutoff before they can be joined, regardless of the other edge relationships (Florek, Łukaszewicz, Perkal, Steinhaus, & Zubrzycki, 1951; Sneath, 1957). Maximum-linkage, sometimes referred to as complete-linkage or farthest neighbor, requires all elements of a cluster to have a cutoff-agreeing edge to all elements of a joining cluster (Williams &

Lance, 1967). Average-linkage, sometimes referred to as mean linkage or unweighted pair group method with arithmetic mean (UPGMA), requires the mean of all cluster element edge distances to meet the clustering cutoff before they can be joined (Sokal, 1958; Williams & Lance, 1967). However, all of these agglomerative linkage methods yield relatively poor results when compared to a divisive method, mainly due to cluster joining requirements (discussed in detail below) as opposed to cluster splitting requirements.

There are obvious visible differences between the three linkage algorithms used to cluster the DHMs of Figs. 3-5. In all of these "bottom-up" agglomerative algorithms, each cluster starts out as a group of identical sequences or a single sequence. Those clusters of identical sequences are then joined as the clustering cutoff is gradually decreased using the chosen linkage algorithm. As a result, the outermost ring (representing 100% sequence identity) for each respective mutation DHM contains the same clustering breakdown, but possibly in a different configuration due to the ring alignment process. However, the clustering breakdown for the joined clusters of the interior rings is subject to the clustering algorithm and is not the same for each mutation step. For example, the interior rings of the seventh mutation step (panel 7 of Figure 3-5) for each agglomerative algorithm appear different.

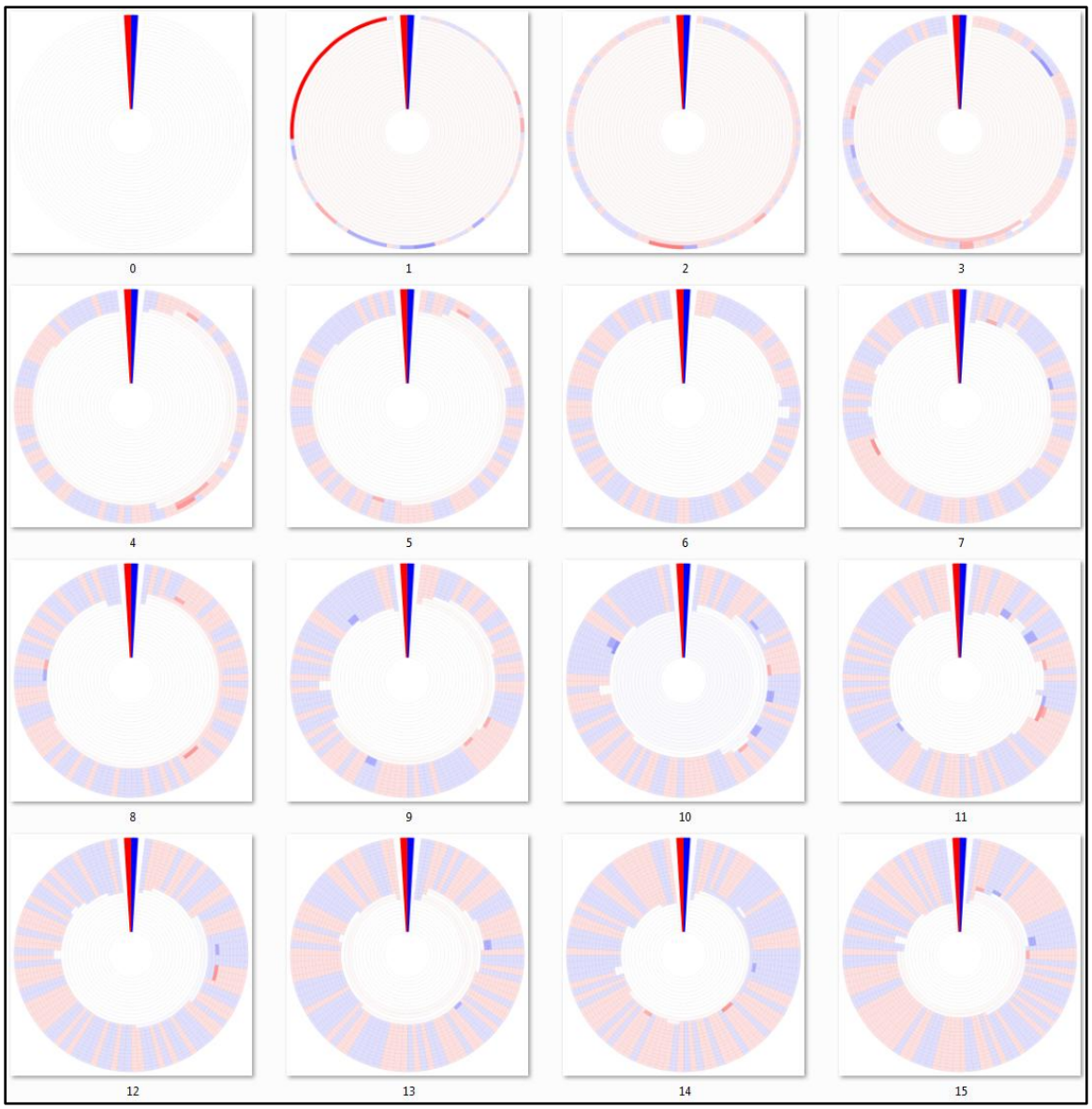

Figure 3: Dendritic heat maps from bottom-up minimum-linkage hierarchical clustering of a mutating population. Dendritic heat maps representing 0 through 15 mutations of the simulated mutation lineage data generated as described in the methods and clustered using the minimum-linkage algorithm of the "bottom-up" method. Panel zero represents the most homologous set of sequences (identical) and panel fifteen represents the least homologous set of sequences (fifteen base substitutions). The darkly colored wedge at the 0° position of each dendritic heat map represents the minimum (red) and maximum (blue)

possible heat map relative abundance bin responses of all dendritic heat maps displayed, GC≤50% and GC>50% respectively. White space in the heat maps represents clusters with neutral bin response. Rings, starting at the center, represent clusters of sequences for identity cutoffs of 0.75 to 1.0. Clusters, including single-sequence clusters, are plotted in a radial range that is conserved from the clusters from which they were derived. High resolution versions of all DHMs in this manuscript are available in Chapter 4 supplemental files S2 and S3.

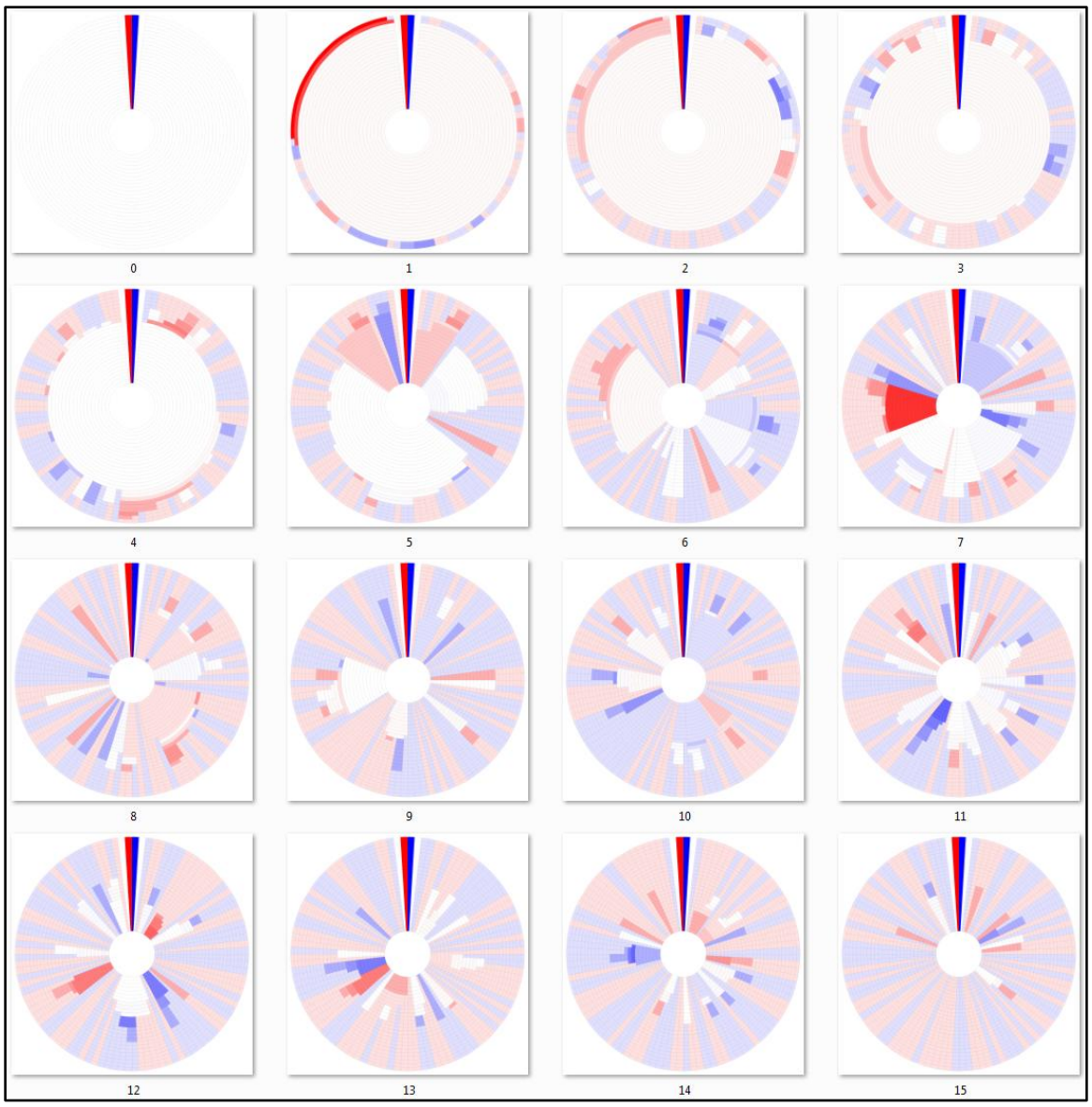

Figure 4: Dendritic heat maps from bottom-up maximum-linkage hierarchical clustering of a mutating population. Dendritic heat maps representing 0 through 15 mutations of the simulated mutation lineage data generated as described in the methods and clustered using the maximum-linkage algorithm of the "bottom-up" method. Panel zero represents the most homologous set of sequences (identical) and panel fifteen represents the least homologous set of sequences (fifteen base substitutions). The darkly colored wedge at the 0° position of each dendritic heat map represents the minimum (red) and maximum (blue)

possible heat map relative abundance bin responses of all dendritic heat maps displayed, GC≤50% and GC>50% respectively. White space in the heat maps represents clusters with neutral bin response. Rings, starting at the center, represent clusters of sequences for identity cutoffs of 0.75 to 1.0. Clusters, including single-sequence clusters, are plotted in a radial range that is conserved from the clusters from which they were derived. High resolution versions of all DHMs in this manuscript are available in Chapter 4 supplemental files S2 and S3.

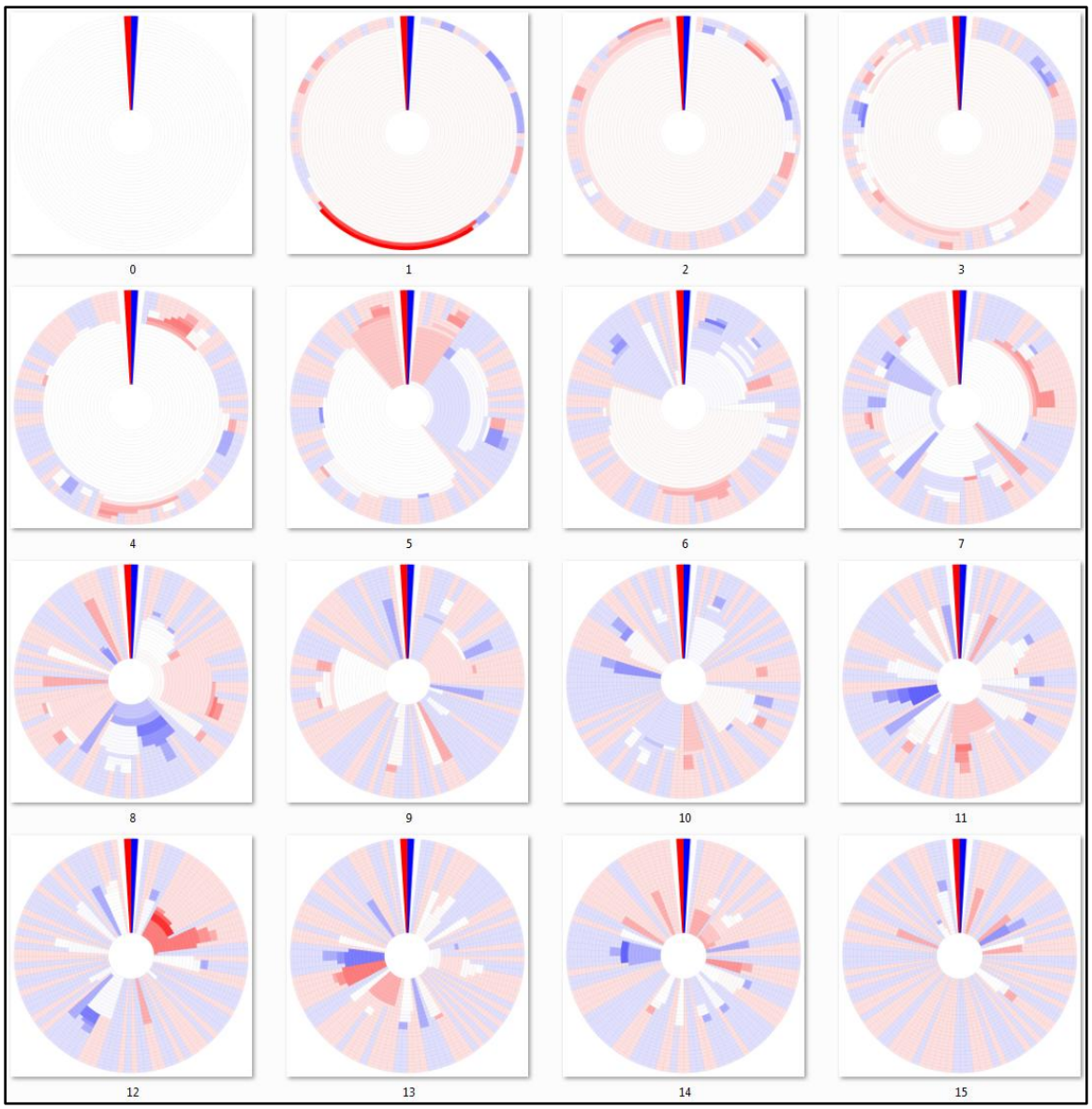

Figure 5: Dendritic heat maps from bottom-up average-linkage hierarchical clustering of a mutating population. Dendritic heat maps representing 0 through 15 mutations of the simulated mutation lineage data generated as described in the methods and clustered using the average-linkage algorithm of the "bottom-up" method. Panel zero represents the most homologous set of sequences (identical) and panel fifteen represents the least homologous set of sequences (fifteen base substitutions). The darkly colored wedge at the 0° position of each dendritic heat map represents the minimum (red) and maximum (blue)

possible heat map relative abundance bin responses of all dendritic heat maps displayed, GC≤50% and GC>50% respectively. White space in the heat maps represents clusters with neutral bin response. Rings, starting at the center, represent clusters of sequences for identity cutoffs of 0.75 to 1.0. Clusters, including single-sequence clusters, are plotted in a radial range that is conserved from the clusters from which they were derived. High resolution versions of all DHMs in this manuscript are available in Chapter 4 supplemental files S2 and S3.

The minimum linkage algorithm of Figure 3 produces figures that give the appearance of relatively well-conserved clusters. The transition from a single cluster to multiple clusters occurs at higher identity cutoffs than that of the other algorithms. This apparent overestimation of sequence conservation is expected with the minimum linkage algorithm since clusters are easily joined, only requiring one sequence from each cluster to match one another at the clustering cutoff level. This method has a well-known drawback called the chaining phenomenon, where clusters that have been joined may share only a single close relationship edge while all other edges are very distant (Williams & Lambert, 1966). The chaining phenomenon certainly affects the results of Figure 3, especially during the earlier mutations, as it would for many single-linkage DHMs, making single-linkage less than ideal for many data sets.

Figure 4, which was constructed using the maximum-linkage algorithm, contains non-joining cluster segments as mutations progress. While maximum-linkage clustering avoids the chaining phenomenon by requiring all cluster members to have an edge to all other members, it is also an underestimation of sequence conservation and some clusters

never join with others for the given cutoff range. This algorithm can be simplified by considering only the furthest edge distance, since all other pairings would be more closely related. In many clusters, a member may be very distantly related to any member of another cluster, regardless of potentially other members in its cluster. This is exactly what can be seen in some mutation iterations in Figure 4, where clusters containing more than one sequence never form edges with all members of another cluster for any given cutoff value, thus never joining.

Like Figure 4, Figure 5 contains many non-joining cluster segments. Using the average-linkage algorithm of Figure 5, clusters are joined if their mean edge distance for all pairs meets the clustering cutoff requirements. The result is similar to the maximum-linkage algorithm because all members of the clusters have an effect on the mean edge distance that dictates if joining will occur. However, average-linkage can be thought of as joining clusters by their "center of cluster mass" instead of a single distant edge, lowering the requirements for joining from that of maximum-linkage and yielding a more accurate apparent sequence conservation.

Due to the chaining phenomenon, minimum-linkage clustering will often be a less than ideal choice for constructing DHMs, although it will still produce a valid DHM. Maximum- and average-linkage algorithms remain viable alternatives, however their propensity to form non-joining cluster segments and the underestimation of sequence conservation by maximum-linkage also makes them less than ideal. In terms of information gained, it is not likely that a non-joining cluster segment does much to aid in visually comparing how deeply rooted a phenotype is across multiple genotypes. Like dendrogram building in general, the algorithm used often comes down to user preference.

However, the need to construct a distance matrix, a step that can require large amounts of computer memory and time, and a propensity for creating non-joining cluster segments makes all three "bottom-up" approaches less than ideal for large data sets.

Top-down hierarchical clustering. The algorithms of the "bottom-up" approaches require constructing a distance matrix from pairwise identity or similarity calculations between all sequences, which for large datasets can lead to impractically large memory or computational time requirements. For this reason, we implemented a "top-down" hierarchical clustering method that splits clusters by incrementally increasing the sequence identity required to bin sequences, similar to previously described divisive methods (Macnaughton-Smith, Williams, Dale, & Mockett, 1964; Sokal, 1958). The "top-down" approach does not require a large comprehensive distance matrix to be built and uses centroid-based clustering, where clusters are split if multiple cluster "centroid" elements can be separated with the given clustering cutoff. Centroids are then used as a database for a sequence search algorithm to assign closest matching sequences or new centroids if the closest match is out of the cutoff range (Edgar, 2010). The dendritic heat maps of Figure 6 were constructed using a top-down approach. A brief example of the top-down clustering is available in Chapter 4 supplemental file S1.

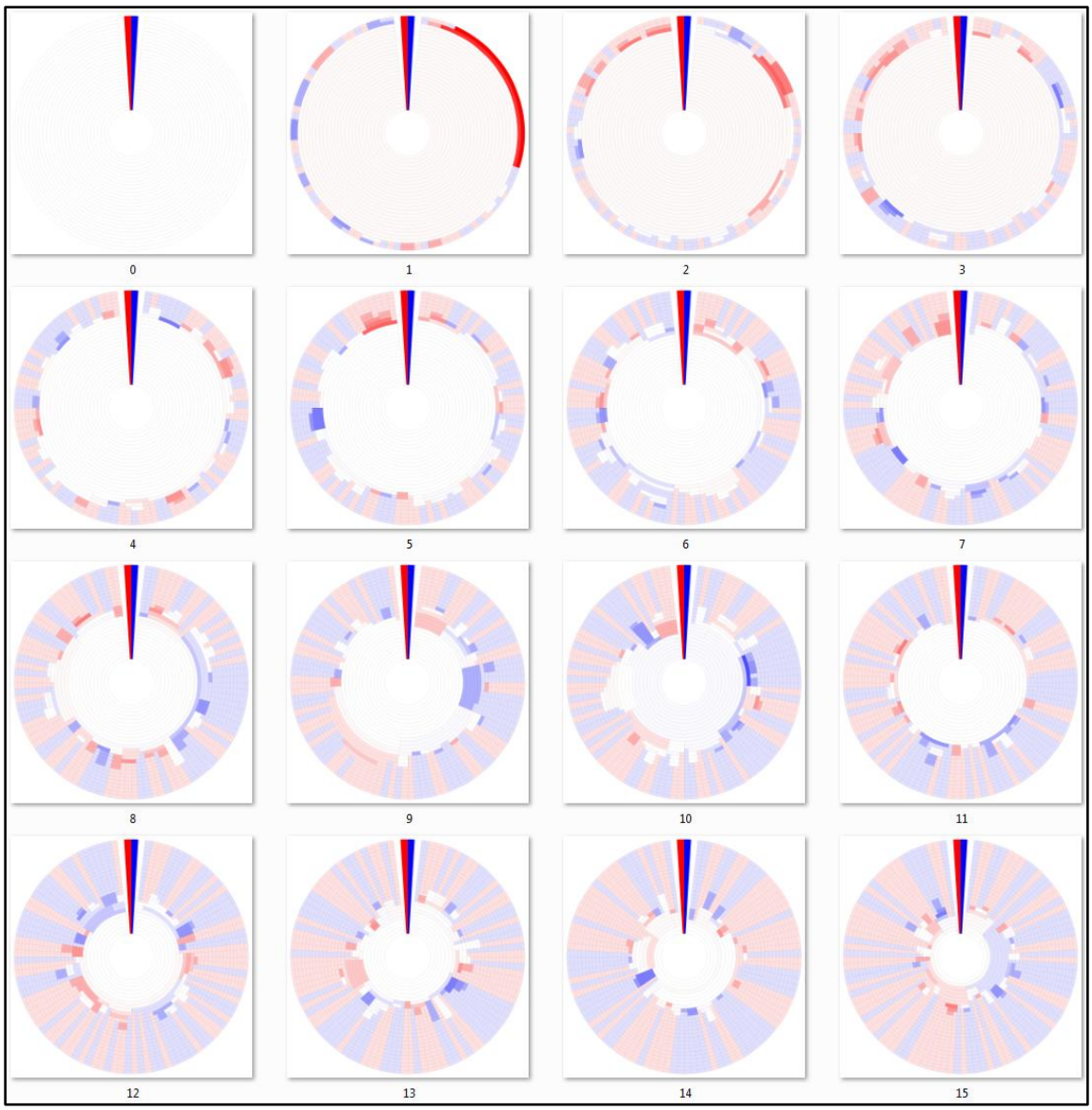

Figure 6: Dendritic heat maps from top-down hierarchical clustering of a mutating population. Dendritic heat maps representing 0 through 15 mutations of the simulated mutation lineage data generated as described in the methods and clustered using the "top-down" method. Panel zero represents the most homologous set of sequences (identical) and panel fifteen represents the least homologous set of sequences (fifteen base substitutions). The darkly colored wedge at the 0° position of each dendritic heat map represents the minimum (red) and maximum (blue) possible heat map relative abundance

bin responses of all dendritic heat maps displayed, GC≤50% and GC>50% respectively. White space in the heat maps represents clusters with neutral bin response. Rings, starting at the center, represent clusters of sequences for identity cutoffs of 0.75 to 1.0. Clusters, including single-sequence clusters, are plotted in a radial range that is conserved from the clusters from which they were derived. High resolution versions of all DHMs in this manuscript are available in Chapter 4 supplemental files S2 and S3.

The most fundamental difference between the "top-down" and all three "bottom-up" algorithms is that the DHM is created by splitting clusters, rather than joining them. The figure is constructed by first assigning sequences to clusters at the lowest cutoff level, which corresponds to the innermost ring on all DHMs, then splitting those clusters for subsequent DHM rings as the clustering cutoff is incrementally increased.

The advantages of the "top-down" centroid-based clustering approach over "bottom-up" approaches are speed, memory requirements, and a more intuitive view of sequence conservation as data size increases. Circumventing the construction of a distance matrix has obvious advantages in speed and memory requirements for sufficiently large data sets, where a distance matrix becomes impractically large. The apparent sequence conservation of the "top-down" DHM is more intuitive in that it avoids the chaining phenomenon of single-linkage and also avoids the non-joining cluster segments of maximum- and average-linkage, potentially leading to more informative observations of the level at which genotype divergence (cluster branching) has an effect on phenotype divergence (relative abundance visualized via differing heat map bin response). The "top-down" approach has a slight bias away from forming non-joining

cluster segments because clusters must pass a threshold to be split, rather than having to pass a threshold to be joined as is the case with “bottom-up” approaches.

In all cases, including the three “bottom-up” approaches, sequence length and homology can have significant effects on the clustering layout of DHMs. Shorter sequences and also less homologous sequences would result in more non-joining clusters, while the opposite conditions would result in more cluster joining. In a study that tracks sequence homology of different lineages, as is simulated in our mutation data set, non-joining cluster segments would appear sooner in data sets with shorter sequences and faster mutation rates. Therefore, there are potential data conditions where even the “top-down” approach yields less-informative DHMs.

Summary tables of the sequence and cluster counts from the mutation dataset DHMs are provided as an Excel document in Chapter 4 supplemental file S3, showing different bin distributions for each of the algorithms which affects their appearance. As previously described with the DHM appearances, the tables reiterate that the minimum-linkage algorithm bins sequences together more readily than the others, while the maximum-linkage and average-linkage algorithms are more exclusive. The centroid-based algorithm, which does not use a distance matrix to determine cluster similarity, occupies a binning inclusiveness middle ground between the three others. It is our opinion that the moderate inclusiveness of centroid-based clustering, as well as its ability to cluster larger datasets (described in more detail in the following section), makes it the best option of the four to construct DHMs. However, just as each of those clustering algorithms is a valid technique, each of the DHMs that are constructed from their clusters is also valid.

Dendritic heat maps of a growing population. DHMs can be scaled to fit a wide range of dataset sizes while maintaining their dendrogram layout. Figure 7 shows DHMs of the artificially generated dataset that simplifies mutation and growth as substitutions are introduced without fatal consequences into generations, displaying an increasing complexity (as described in the methods section as the population growth data set). Up to fifteen generations, including the initial sequence at generation 0, are shown as individual DHMs. The number of total sequences doubles for each generation, increasing data size and complexity exponentially. This dataset is used to show the scalability of DHMs.

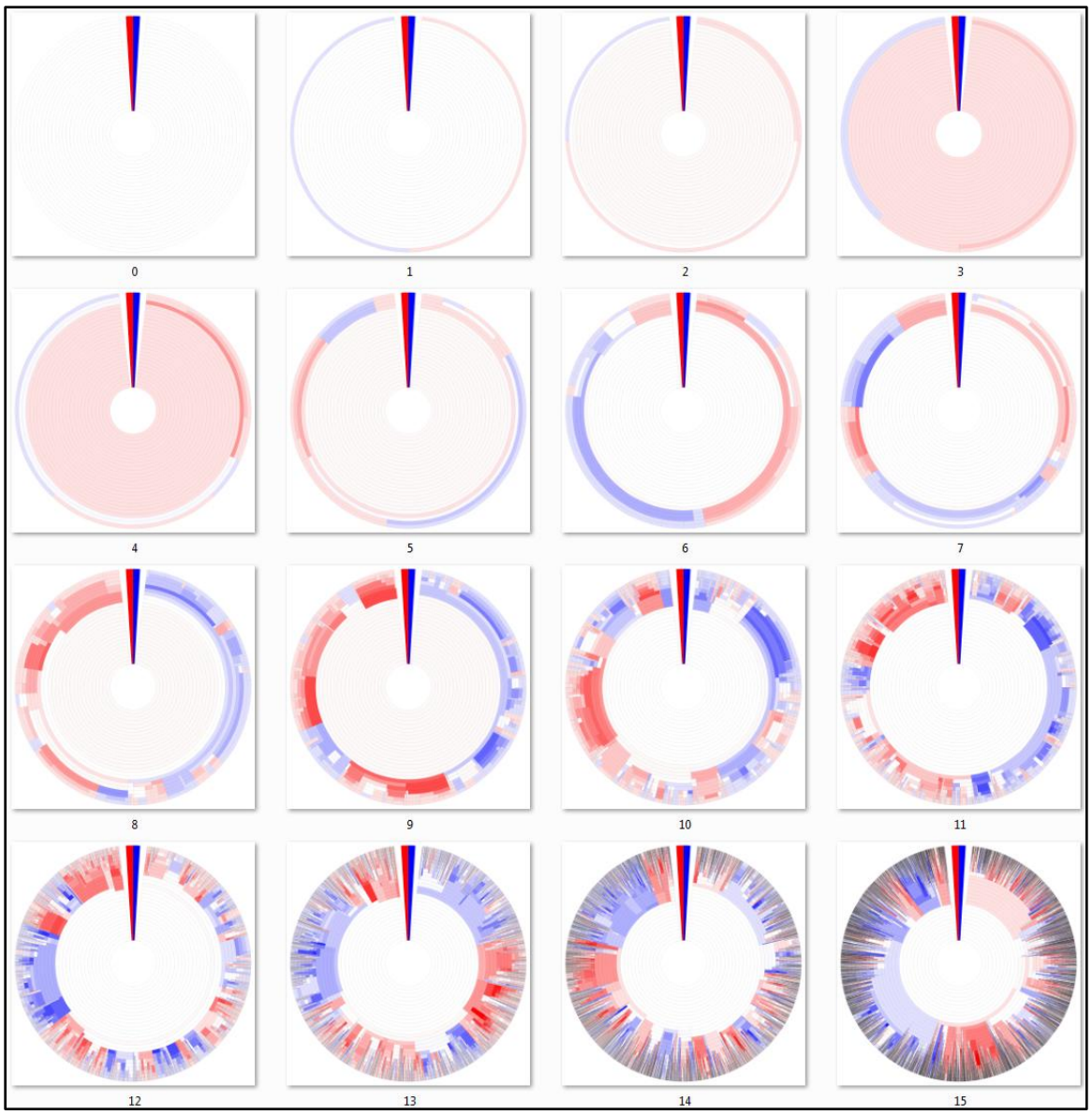

Figure 7: Dendritic heat maps from top-down hierarchical clustering of a growing population. Dendritic heat maps representing generations 0 through 15 of the simulated population growth data generated as described in the methods and clustered using the "top-down" method. The darkly colored wedge at the 0° position of each dendritic heat map represents the minimum (red) and maximum (blue) possible heat map relative abundance bin responses of all dendritic heat maps displayed, GC ≤50% and GC>50%

respectively. White space represents clusters with neutral bin response. Rings, starting at the center, represent clusters of sequences for identity cutoffs of 0.75 to 1.0. Clusters, including single-sequence clusters, are plotted in a radial range that is conserved from the clusters from which they were derived. High resolution versions of all DHMs in this manuscript are available in Chapter 4 supplemental files S2 and S3.

The DHM for generation 0 represents the simplest possible cluster configuration, displaying information for only a single sequence. For a fixed sequence length as in the artificial data used here, there is a theoretical final cluster distribution, where mutations have progressed to a point where additional sequences can no longer be unique. The cluster distribution for the DHM of this theoretical endpoint would appear symmetrical, and all evolutionary paths end at this same fixed endpoint cluster distribution. The number of unique sequences in the endpoint cluster distribution is calculable at b^n, where b is the number of possible base choices and n is the total number of bases used in the simulated DNA sequences. The generations that are displayed in Figure 7 are snapshots into one of the many pathways toward the theoretical endpoint cluster distribution of b^n unique sequences. This theoretical endpoint holds true for traditional dendrograms as well, however DHMs have the added dimension of displaying relative abundance information (heat map bin response). The random qualities built into our sequence generation with the added dimension of heat map distribution yields many DHM colorations for the fixed endpoint cluster distribution discussed above.

Drawing parallels to natural data, for every set of samples, there is also a theoretical fixed endpoint cluster distribution. While natural data does not have fixed

sequence lengths, there is likely a range of lengths that are allowed by evolutionary pressures and there would be many possible evolutionary paths toward the endpoint cluster distribution (Wang, Hsieh, & Li, 2005; Xu et al., 2006). However, the difference with a natural data set is that generations are determined by evolutionary processes that are much more complex than random non-lethal base substitutions (Krebs, Goldstein, & Kilpatrick, 2009; Mitchell-Olds, Willis, & Goldstein, 2007). Essentially, the artificial data set endpoint would display every possible lineage while natural data would have fatal dendrogram branches trimmed out of the endpoint cluster distribution, likely in a non-symmetrical distribution. The complexity that is common in natural samples would likely yield many possible DHM colorations, similar to an artificial data set. For these reasons, it is not unreasonable to use artificially-generated data to show DHMs of population growth.

Figure 7 is an informative way to introduce the evolutionary context of DHMs, where each displays a multi-level snapshot into the phenotype history of a sample. Each snapshot represents a view from the same evolutionary path, one of the many paths toward the theoretical endpoint. As each generation doubles in size, in many cases it is easy to visually track the growth and divergence of individual clusters from generation to generation. As new unique sequences are added, new clusters in the outermost ring are created and inner clusters diverge to account for their addition. Likewise, when duplicate sequences are added, their respective sections increase in width, which represents cluster size. Phenotype divergences occur deeper into the DHMs as generations progress and population genotypes diverge. Eventually, we are able to see increasing fracturing of genotypes and phenotypes as the total population becomes more complex. While

something similar to Figure 7 could be recreated with experimental evolution datasets, a natural dataset would yield only a single DHM unless a time component is involved. However, even with a time element involved, it is unlikely to find a natural sample as simple as the earliest generations of Figure 7.

Application. Recently, DHMs were used to describe large and complex microbial community sequence data from aquatic pumice samples, where the goal was to show microbial habitat preference at multiple levels of taxonomical classification (Elser et al., 2015). While seemingly a straightforward task of counting homologous sequences for each habitat, the problem of similarity cutoff choice can influence how sequences are binned and ultimately expressed in a heat map. In Figure 8 (Figure 3 of Elser *et al.*, 2015) (original copyright 2014, Applied and Environmental Microbiology), sequences maintain the same radial position throughout each of the DHMs. For nearly all sequences, the strength, and sometimes direction, of their heat map expression changes depending on the specificity of the clustering cutoff used to bin them. Figure 9 (Figure 4 of Elser *et al.*, 2015) (original copyright 2014, Applied and Environmental Microbiology) shows histograms of the same data being binned according to Ribosomal Database Project taxonomic classifications (Elser et al., 2015). Essentially, these histogram bar heights translate to heat map color intensity and convey the same information in different formats. Of course, binning based on taxonomic classification does not exactly represent binning based on sequence identity, but it can be a close and familiar approximation. The issue with clustering cutoff choice is perfectly represented in Figure 9 (Figure 4 of Elser *et al.*, 2015), where depending on the level at which sequences are binned, the scale of the "Skew Line" axis changes to accommodate the range in relative abundance bin

response, which translates to heat map color and intensity. DHMs on the other hand embrace this effect of clustering specificity on binning and bin response, where it is used to show the effect of homology on heat map response.

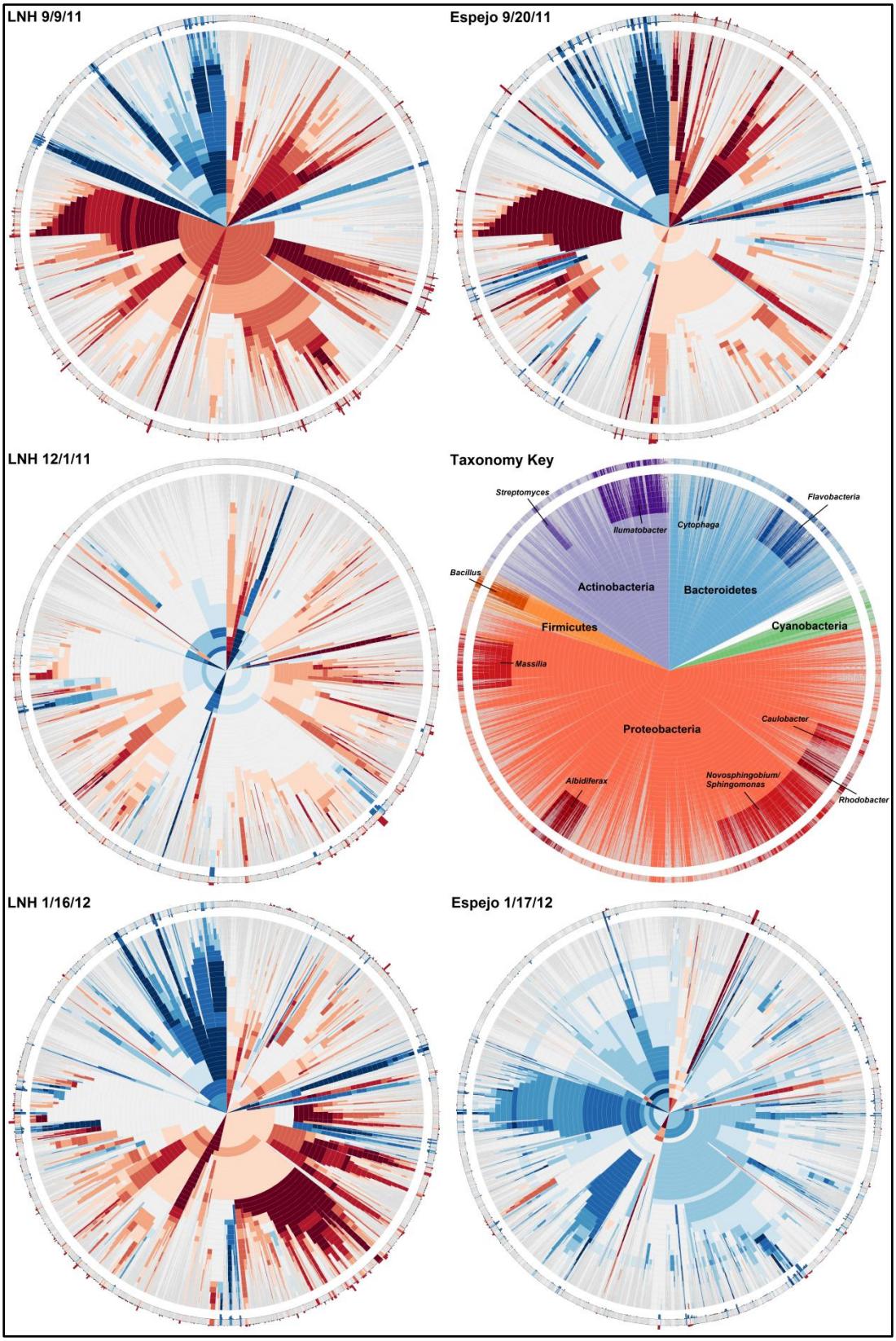
LNH 9/9/11
Espejo 9/20/11
LNH 12/1/11
Taxonomy Key
Streptomyces
Flavobacteria
Ilumatobacter
Cytophaga
Bacillus
Actinobacteria
Bacteroidetes
Firmicutes
Cyanobacteria
Massilia
Proteobacteria
Caulobacter
Albidiferax
Novosphingobium/
Sphingomonas
Rhodobacter
LNH 1/16/12
Espejo 1/17/12

Figure 8. Figure 3 reused from Elser *et al.* 2014 with the kind permission of ASM. Dendritic heat maps displaying habitat preferences for multiple levels of phylogenetic clades across multiple time points and locations. Reprinted from (Elser et al., 2015) under a CC BY license, with permission from AEM, original copyright 2014.

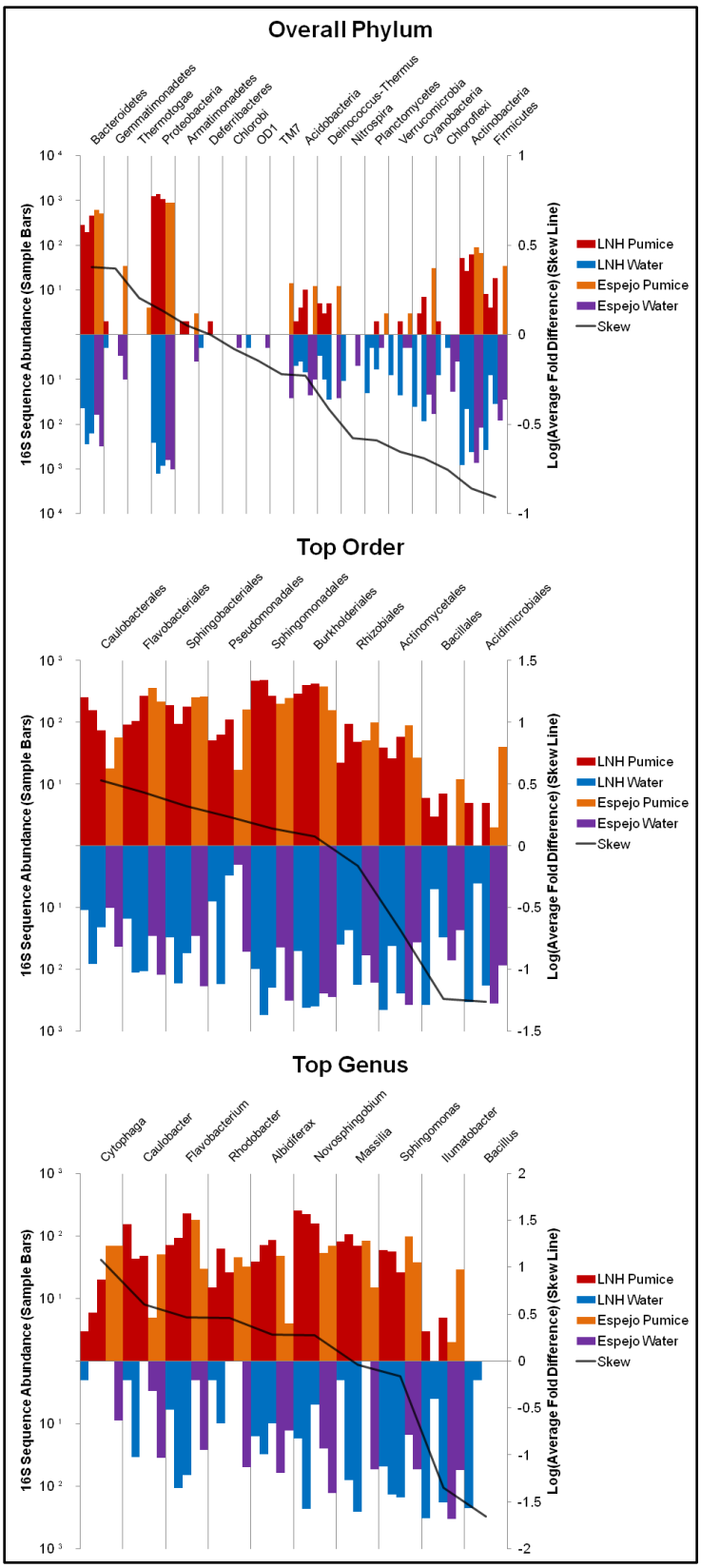
Overall Phylum
Bacteroidetes
Gemmatimonadetes
Thermotogae
Proteobacteria
Armatimonadetes
Deferribacteres
Chlorobi
OD1
TM7
Acidobacteria
Deinococcus-Thermus
Nitrospira
Planctomycetes
Verrucomicrobia
Cyanobacteria
Chloroflexi
Actinobacteria
Firmicutes
16S Sequence Abundance (Sample Bars)
Log(Average Fold Difference) (Skew Line)
LNH Pumice
LNH Water
Espejo Pumice
Espejo Water
Skew
Top Order
Caulobacterales
Flavobacteriales
Sphingobacteriales
Pseudomonadales
Sphingomonadales
Burkholderiales
Rhizobiales
Actinomycetales
Bacillales
Acidimicrobiales
16S Sequence Abundance (Sample Bars)
Log(Average Fold Difference) (Skew Line)
LNH Pumice
LNH Water
Espejo Pumice
Espejo Water
Skew
Top Genus
Cytophaga
Caulobacter
Flavobacterium
Rhodobacter
Albidiferax
Novosphingobium
Massilia
Sphingomonas
Ilumatobacter
Bacillus
16S Sequence Abundance (Sample Bars)
Log(Average Fold Difference) (Skew Line)
LNH Pumice
LNH Water
Espejo Pumice
Espejo Water
Skew

Figure 9. Figure 4 reused from Elser *et al.* 2014 with the kind permission of ASM. Histograms displaying the strongest habitat preferences for the phylum, order, and genus taxonomical levels of four sample types. A skew line is used to show the relative strength of habitat preference. Reprinted from (Elser et al., 2015) under a CC BY license, with permission from AEM, original copyright 2014.

In a publication by Eisen *et al.*, heat maps are used to describe *Saccharomyces cerevisiae* genome microarray data for a series of time points (Eisen, Spellman, Brown, & Botstein, 1998). Each row in Figure 10 (Eisen *et al.*, 1998 Figure 1) (original copyright 1998, The National Academy of Sciences) represents individual genes, which in terms of binning are sequences at 100% identity or a fractional identity clustering cutoff of 1.0, and each column represents a time point. Figure 10 (Figure 1 of Eisen *et al.*, 1998) shows rows being clustered and arranged based on their heat map response and a dendrogram is provided to display the cladistics of row bin responses, not row sequence identity, which is useful for displaying clades of similar patterns of expression. However, the goal of DHMs is to display the effect of sequence homology and genotype clades with phenotype heat map bin responses. Two important differences between the DHMs introduced in this work and many published traditional heat maps, including those in Eisen *et al.*, are the inclusion of multiple heat maps for a single sample (represented by a column in Eisen *et al.*, 1998) and sequence identity rather than heat map expression pattern determining row arrangement (or radial position in the case of radial heat maps). If we were to convert the work of Eisen *et al.* into DHMs, each column of Figure 10 (Figure 1 of Eisen *et al.*, 1998) would have their own DHM with multiple levels of clustering arranged by sequence

identity so that we could see how well the heat map expression pattern for a gene is conserved among its homologs. It would be possible to create a separate cladogram that represents bin response pattern similarity (or dissimilarity) using Bray-Curtis dissimilarity, however this is beyond the scope of the work presented here (Bray & Curtis, 1957).

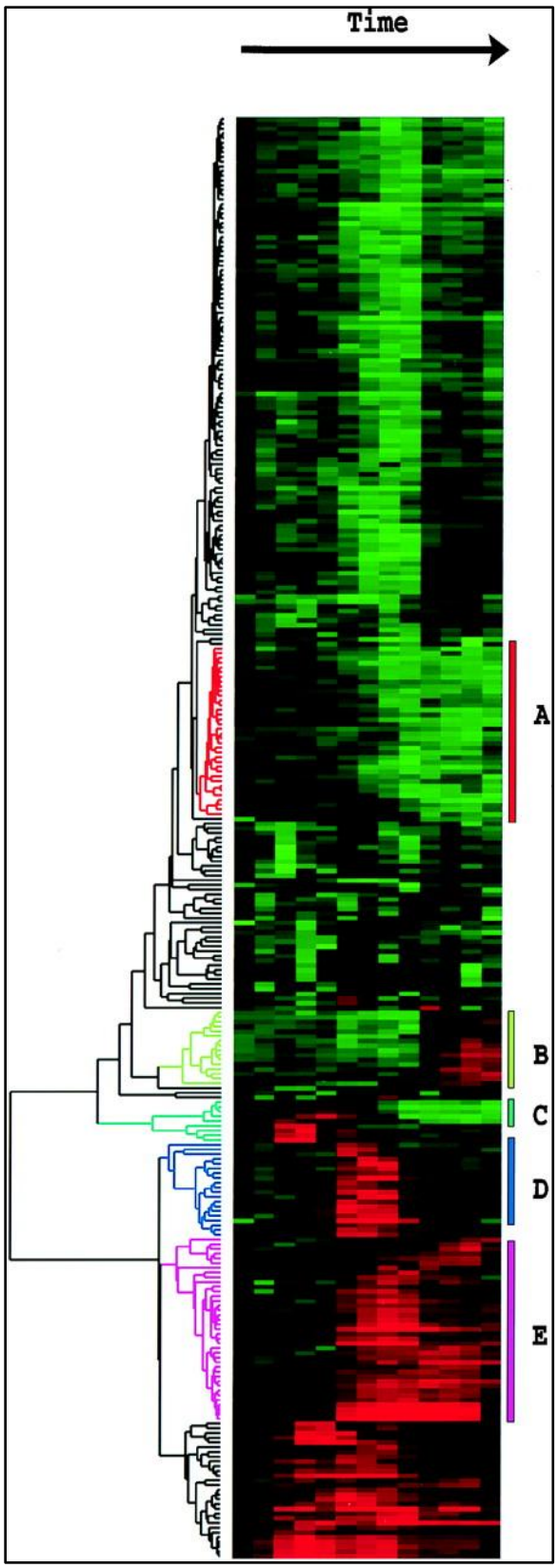
Time
A
B
C
D
E

Figure 10. Figure 1 reused from Eisen *et al*. 1998 with the kind permission of PNAS. Heat map displaying data from a time course of serum stimulation of primary human fibroblasts. Reprinted from (Eisen et al., 1998) under a CC BY license, with permission from PNAS, original copyright 1998.

Conclusion

DHMs represent a novel and powerful tool for visualizing correlations in genotype and phenotype changes across evolutionary space and time, and will ultimately help decipher dynamic processes in complex, natural communities such as metatranscriptomes, where similarities occur across a multitude of scales. The "top-down" approach that we outline here provides an efficient method of constructing DHMs that display phenotype relative abundance divergence with homology divergence and is the method that we recommend for most cases, however, any hierarchical clustering method can be used for DHM construction. While this paper discusses the application of DHMs in an exclusively nucleic acid sequence context, their range is certainly not limited to sequence information and can be used in any dataset that has a pair of groups that share underlying traits.

Acknowledgements

We thank Amisha Poret-Peterson for helpful comments and proof-reading. We also thank the editors and anonymous reviewers.

CHAPTER 5

ARCHITECTURAL ELEMENTS OF DENDROGRAM AND HEAT MAP VISUALIZATION, AND THE DISPLAY OF HIERARCHICAL CLUSTERING MULTIDIMENSIONALITY.

Matthew Kellom

School of Earth and Space Exploration, Arizona State University, Tempe, Arizona, USA

Abstract

Dendrograms and heat maps are well-established figure types that excel at visualizing hierarchies in clustered datasets. As the number of tools to create these figures continues to grow, it is important to consider the architectural elements that make successful dendrograms and heat maps. Data communication effectiveness is the result of coordinating input data with figure scale and color to support the graphical flow of information. However, even under optimal conditions, neither can properly convey the inherent multidimensionality of complex clustered data. In response, novel figure types that combine fundamental features of dendrograms and heat maps are used to fill the multidimensionality data visualization gap. In this chapter, the effects of architectural elements in published dendrograms and heat maps are examined with respect to their function of communicating complexity. The subject of multidimensionality that these figure types struggle to display is explained and two strategies to reveal multiple levels of clustering datasets are discussed.

Introduction

Dendrograms and heat maps are staples of data visualization, especially in communication of biological data. Methods to calculate and construct these figure types continue to advance in speed, precision, and accuracy as technology continues to facilitate collection of more and more complex datasets. When pushed to the extremes, the basic forms of dendrograms and heat maps reach the limits of conveying data hierarchies, becoming overcrowded and confusing. The best case scenario for these overcrowded figures is that overall data trends emerge, but any scrutiny of the small details may be nearly impossible. In response to the limits of these figure types, dendrogram and heat map elements are being developed into figures that display different properties, finding novel uses for classical concepts.

Ideally, the visualization of dendrograms and heat maps should be intuitive and engaging. Ultimately, viewers should be able to understand the figures being presented but also agree that the figures add constructive context to scientific reports. From *The Visual Display of Quantitative Information* by Edward R. Tufte, "What is to be sought in designs for the display of information is the clear portrayal of complexity. Not the complication of the simple; rather the task of the designer is to give visual access to the subtle and the difficult - that is, the revelation of the complex" (Tufte, 1983). However, due to the large and complex nature of many datasets, distilling information into graphical designs is not always a simple task. As a result, multiple strategies exist to simplify datasets and are sometimes employed to create more visually appealing figures (Gisbrecht & Hammer, 2015; Laczny et al., 2015; Laczny, Pinel, Vlassis, & Wilmes, 2014; Liu, Maljovec, Wang, Bremer, & Pascucci, 2017; Ma & Sun, 2015).

There are a plethora of available programs, scripts, web-based applications, etc. available for clustering data and creating dendrograms and heat maps. A quick search of available options can yield an overwhelming amount of results with each seeking to either improve on the accuracy and/or efficiency of clustering, or provide new methods of constructing these figures. Since the new year (2017), methods such as dendrogram seriation (Arief, DeLacy, Basford, & Dieters, 2017), RNAscClust (Miladi et al., 2017), PhyD3 (Kreft, Botzki, Coppens, Vandepoele, & Van Bel, n.d.), GGTREE (Yu, Smith, Zhu, Guan, & Lam, 2017), and more, have emerged as opportunities to improve on data analysis and visualization. Rather than review the growing body of software, a potentially more meaningful discussion lies in the universal factors that affect dendrogram and heat map design: input data, scale and density, and color. These three elements of dendrogram and heat map design reflect the layers of figure architecture that "organize and order the flow of graphical information" (Tufte, 1983). Following a discussion of these design elements and examples of their impact in published figures, we look into recent innovations of using dendrogram and heat map concepts to create figures capable of alternate analyses. This path of data visualization is important for the advancement of science communication, adding available options to explain complicated subjects.

Dendrogram and Heat Maps

Dendrograms. Dendrograms are diagrams used to display hierarchical clustering. The essential components of dendrograms are: nodes, branches, branching points (or internal nodes), and in some cases a root. Nodes (represented by A and B in Figure 1) are the endpoints of dendrograms that represent either a single data point or a group of individual data points that share an elementary level of relatedness. Branches

(represented by C and D in Figure 1) are the lines that connect nodes to each other and usually a medial branching point (represented by E in Figure 1) or many branching points depending on the level of hierarchical complexity between the nodes. A root (represented by F in Fig. 1) is the base of the tree, serving as the top-level commonality between all data points in the dendrogram. Some dendrograms are unrooted, in which case nodes and braches can only be referenced to each other instead of a single common basal branch.

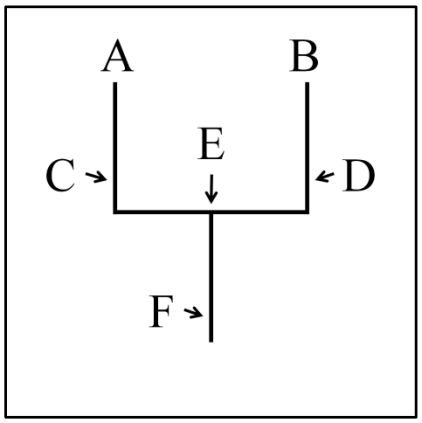

Figure 1: A simple dendrogram with labeling of dendrogram components. (A) and (B) are nodes. (C) and (D) are branches. (E) is a branching point, or internal node. (F) is the root.

It is also important to note the concept of "distance" that is often directly or indirectly associated with dendrograms. In brief, distance is a measure of relatedness between any two points on a dendrogram; points that have a large distance between them are less related to each other than points that have a small distance separation. On occasion, calculated distances are only accurately represented along a single dimension so that branch spacing and orientation accommodates labeling. For example, a

rectangular tree can be laid out in an orientation of the root on the left and nodes flush against a y-axis on the right, consisting of evenly spaced node labels. The x-axis in this design is a distance scale, so that only horizontal branch lines include a measure of distance. In general, all dendrograms share the feature that longer branching paths equate to larger distances, even if the branching paths are not scaled to a specific distance metric. Depending on the methods of dendrogram construction and the types of data involved, a distance metric may be calculated to scale branch lengths and proportionally represent exact distance estimates between any two points.

Heat maps. Heat maps are graphs used to display response differences of a dataset. For the heat map in Figure 2, there are three essential components: x-axis, y-axis, and matrix. In general, the variables of the x-axis and y-axis are interchangeable, meaning the axis labels could be swapped with no effect on the data interpretation. In this example, one axis containing samples is being compared to the other axis containing conditions. The resulting matrix of these pairwise axes comparisons provides a visual of response differences between each combination, which could be correlation or abundance information.

While the matrix heat map orientation of Figure 2 is widely used in biology and science in general, there are many heat map orientations that exist. The defining feature of all heat maps is the color scale that indicates the range of possible data values. Opposing ends of the color scale represent opposing ends of the data being represented, with intermediate colors representing intermediate data values.

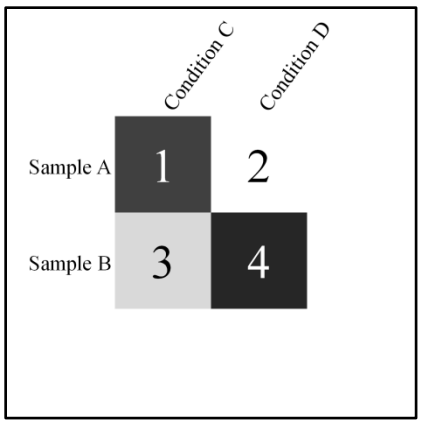

Figure 2: A simple matrix heat map with labeling of heat map components. In this figure, rows (Sample A, Sample B) represent individual data points or groups of data points whose response was measured. Columns (Condition C, Condition D) represent stimuli that my produce a measurable response from the row data points. The numbered matrix (1-4) represents measured responses relative to each other. In this grayscale example, the darkest and lightest shaded squares would represent opposite extremes in response, with the more neutral shades representing more moderate responses.

Design

Data. Scientific pursuits accumulate many different types of data and not all are amenable to the same visualization techniques. If a dataset is small or simple enough, data can be added directly into text results or discussion without the consequence of confusing readers. However, scientific datasets are often large and complex, requiring organization to facilitate comprehension. Some datasets can be communicated in their raw form, or arranged into a table, but many datasets are communicated more effectively with the use of a graph or figure to highlight important data features. At the broadest level of characterization, qualitative and quantitative data are used to measure variables or record observations that can be reported and add meaning to discussion topics. Both

types of data can be used to create dendrogram and heat map figures, although quantitative data is more commonly used.

Qualitative data. Qualitative data is descriptive of categorical properties, usually measured in a nominal or ordinal scale. A nominal scale is used to categorize unmeasured data with names or words, such as the names of individuals. Data on an ordinal scale is ranked categorically in a specific order without a standard of measurement, such as "mostly disagree; neither agree nor disagree; mostly agree," etc. Binary data can also be considered as qualitative as it is a very limited form of categorical options. In the context of dendrograms and heat maps, qualitative data may limit construction and visualization to abundances of categorical responses or presence/absence information. For dendrograms, this can complicate distance calculations since an absolute distance measurement would require a quantitative substitution model (Schloss, 2010). However, it is possible to create a measure of relative distance with data commonalties (Hetherington et al., 2015). In fact, tree-of-life dendrograms were once commonly constructed using qualitative anatomy similarities, a practice that is now virtually exclusive to fossil remains (Wolfe, Daley, Legg, & Edgecombe, 2016). Heat map visualizations of qualitative data would almost certainly require some form of content or thematic analysis to quantify qualitative and categorical properties, since they require data that is hierarchical and standardized to be directly comparable (Vaismoradi, Turunen, & Bondas, 2013).

Quantitative data. Quantitative data is descriptive of measured quantities and is the basis of most heat maps and dendrograms which are designed to display magnitudes of distance and comparisons. Quantitative data is well-suited for data visualization in

general because of the ability to directly compare and scale measured values through calculated distances, fold changes, statistics, etc., which can then be graphed relative to other quantitative data metrics. Perhaps the most difficult challenge that heat maps and dendrograms face is the ever increasing size of quantitative datasets. The first hurdle is to analyze these massive datasets, which has led to a growing field of developing new clustering methods that minimize the use of computational time and resources (Edgar, 2010; Fu, Niu, Zhu, Wu, & Li, 2012; Ghodsi, Liu, & Pop, 2011; Gronau & Moran, 2007; Huse, Welch, Morrison, & Sogin, 2010; Kellom & Raymond, 2017). In the current state of dendrogram and heat maps visualization methods, the only effective means of dealing with such large datasets is by data reduction before the visualization process. Very large visualizations of hierarchical data are widely considered impractical, with some estimating the upper limits of displayable data points to be around 200 (Jackson, 1997; Morris, Asnake, & Yen, 2003; Schonlau, 2004).

Scale and density. Scale and density are two factors that dictate the shape of dendrogram and heat map visualizations. The scale of the data, which can be thought of as the scope or range, influences visualization choices relating to perceived distances or differences between data points. The main challenge of visualizing very large datasets is representing the data in a way that is accurate and complete, while also intuitive and uncluttered (Krzywinski, Birol, Jones, & Marra, 2012). Providing scale references gives context to the illustrated data points, and large data visualizations almost invariably require the ability to compare scales to be more coherent.

The density of a figure in relation to data scale affects figure comprehension by allowing or restricting information assimilation within the figure space (Tufte, 1983).

Figure density can refer to information density of the overall image or local areas within the larger picture, both meanings are relevant to the design of dendrograms and heat maps. Depending on the scale of the figure, large datasets may require information dense visualizations for accurate depiction. With a limited amount of space in figure design, the potential for overcrowding is ever present. Dendrograms and heat maps that are either too dense or not dense enough can lead viewers to question the credibility of the data being shown.

Use in dendrograms. Dendrograms are the visualization of data clustering and can take any shape so long as they maintain the basic structure of Figure 1. Conventionally, dendrograms in science publications are displayed as linear or radial with only occasional amorphous or sprawling designs. Both linear and radial dendrograms each have visualization advantages and disadvantages that are closely associated with the scale of the dataset and density of the figure.

Figure 3 contains four examples of dendrograms that are either linear or radial, and are of varying scales. Figure 3A from Han *et al*., 2017 (Han, Liu, Wang, & Liu, 2017) is a linear dendrogram depicting relatedness between five populations of *Vicia ramuliflora* and *Vicia unijuga*, which are species of flowering legume plants known as vetches. Although it may be inferred from the small number of nodes that Figure 3A uses a relatively small scale and the five populations of these two species are closely related, the dendrogram does not contain a scale reference to make that conclusion based on the figure alone. The only inferences that can be made from this figure are the hierarchical relationships of each population to the others, which was the authors' intent and the simplicity of the dendrogram lends itself to. This particular dendrogram was created with

a method that assumes branch lengths from root to each node are equal, which means that the lack of a distance scale is acceptable since node pairs are assumed to be equidistant (Felsenstein, 2004). Similar lacking of scales can create deceptive diagrams if care is not taken to explain the purpose of the dendrogram. If the dendrogram is meant to show distances between populations of clusters, a scale reference is absolutely needed.

Figure 3B from Du *et al*., 2013 (Du, Pan, Tian, Li, & Zhang, 2013) is an unrooted radial dendrogram of protein family member homologs from three plant genera. Unlike Figure 3A, Figure 3B contains a scale reference which enables distance comparisons between dendrogram locations. By supplying a scale reference, this figure very effectively gives a sense of magnitude to each of the branches among the three classes of homologs. The amount of spacing given to the scope of this dataset constitutes an uncluttered view of all branches and nodes, making it very easy to tell that the Class III cluster has less intra-cluster diversity than either Class I or Class II, and that some nodes in the overall dendrogram are more distantly related than others. This sense of scale is further emphasized artistically by the sizes of the class labeling arcs, albeit in a non-rigorous fashion.

Labeling arcs are also used in Figure 3C from Sehgal *et al*., 2015 (Sehgal et al., 2015) in a radial dendrogram of germ cell genetic material (germplasm) accessions from four lines of spring bread wheat. The labeling arcs are used to show six clusters of germplasm and the intra-cluster diversity. In contrast to Figure 3A, these branch lengths are not assumed to be equidistant and an observable range of distances are depicted. The labeling arcs and branch lengths in Figure 3C are unable to be compared to a distance scale, since no distance scale exists. Therefore, the only conclusions that can be inferred

from this dendrogram are that some node and/or cluster pairs are more distantly related than others, and that some clusters have less intra-cluster diversity than others. In practice, the data scope of this figure makes it difficult to make distance comparisons of specific nodes and/or clusters. The authors have sacrificed small scale comprehension to present overarching trends, which makes the functional scale of this visualization more similar to the populations of Figure 3A than the homologs of Figure 3B, despite the individual branches for each gene accession.

Figure 3D from Wen *et al*., 2012 (Wen, Franco, Chavez-Tovar, Yan, & Taba, 2012) is another example of sacrificing small scale scope to show overarching trends, in an arguably less effective format. Figure 3D is a linear dendrogram representing relationships between tropical maize germplasm accessions, plotted with comparable branch distances and a distance reference. Similar to Figure 3C, the external branches are considerably longer than most internal branches, indicating relatively low intra-cluster similarity and high inter-cluster similarity between neighbors. Many branching points between the clusters in Figure 3D clearly exist but are compressed to accommodate long external branches, making them difficult to examine or compare to the reference scale. The scale of the data in this figure creates a visually dense design that could inhibit comprehension. The purpose of this figure appears to be highlighting the disparity between genotype (cluster) and geographic sampling region (color), at which it succeeds since there is a general absence of contiguous blocks of single external branch colors. However it is difficult to hone in on specific clusters or geographic regions in this crowded dendrogram design.

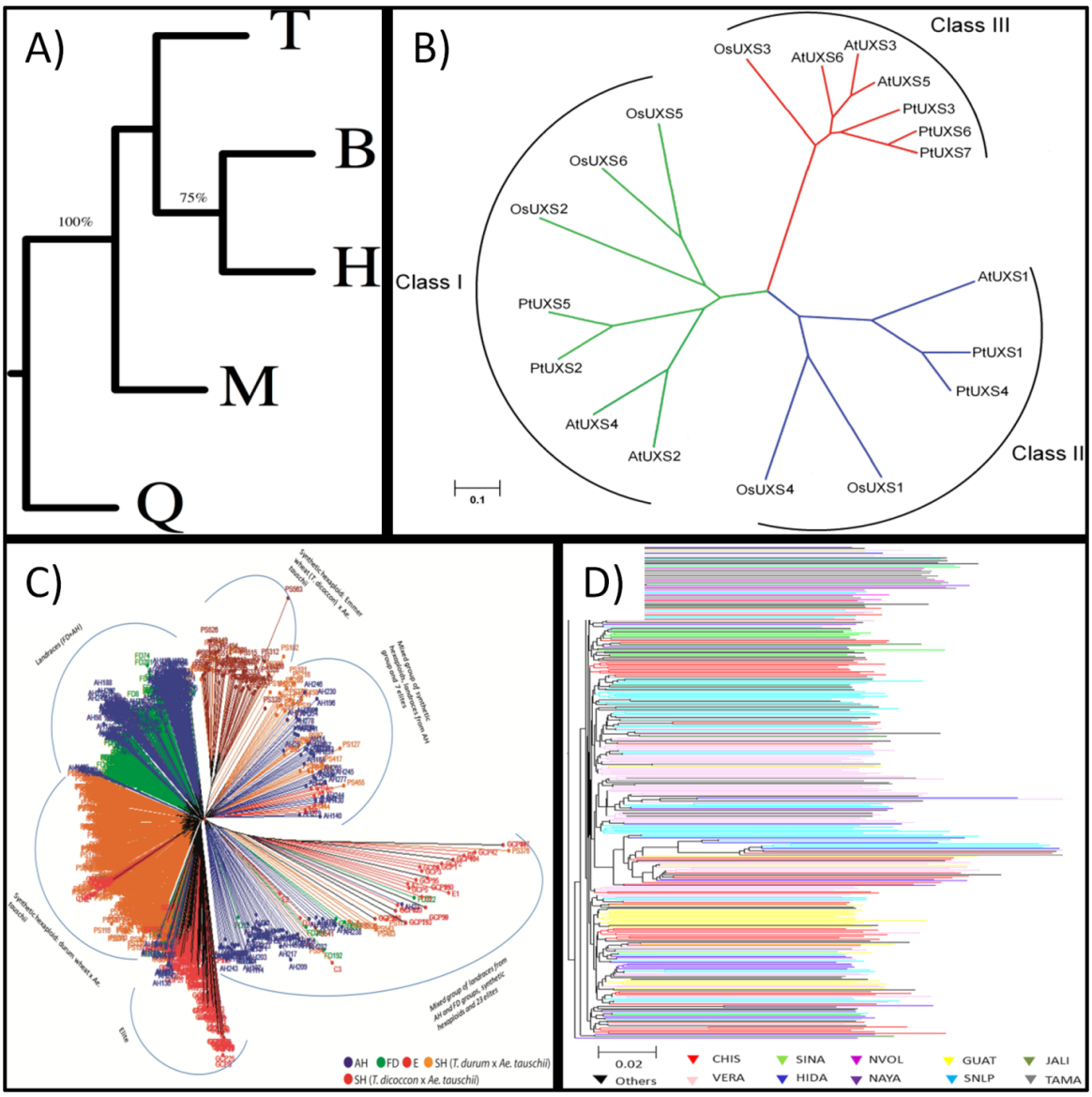

Figure 3: Dendrogram examples of scale and density. A) Figure 4 from Han *et al*., 2017 depicting populations of *Vicia ramuliflora* and *Vicia unijuga* (Han et al., 2017). B) Figure 2 from Du *et al*., 2013 showing protein family member homologs from three plant genera (Du et al., 2013). C) Figure 3 from Sehgal *et al*., 2015 illustrating spring bread wheat germplasm accessions (Sehgal et al., 2015). D) Figure 2 from Wen *et al*., 2012 portraying tropical maize germplasm accessions (Wen et al., 2012).

Use in heat maps. Heat maps are often closely associated with dendrograms as seen in Figure 4, where one or both axes are attached to a dendrogram that displays row/column hierarchy. The axes categories on heat maps are commonly presented as dendrogram nodes, illustrating hierarchical relationships of axes values for added data context. Figure 4A from Xue *et al*., 2017 (Xue et al., 2017) is a small scale heat map from a study of aplysin intervention/protection in a rat model of ethanol-induced liver injury. This heat map depicts gut microbiome bacterial genera (rows) abundances in response to control, alcohol model group, or aplysin intervention group (columns left to right respectively). The information density of this figure is relatively low, making it easy to compare individual data points in the heat map grid to each other and the measured scale bar. The purpose of this heat map is to differentiate the major and minor gut microbiome genera in each of the treatment groups. The heat map appears to be ranked so that more abundant genera are oriented at the top of the figure with less abundant genera at the bottom, clustering by subpopulation size. Columns are also arranged by row response similarity to reveal treatment group similarity. This heat map effectively communicates small scale differences and similarities between rows and between columns for which it was designed.

Figure 4B from Wang *et al*., 2016 (Wang et al., 2016) is a heat map with a scale that serves a different purpose than that of Figure 4A. This large scale heat map is used to show synaptic protein correlation from Alzheimer's disease patients (post-mortem brain tissue) while testing an algorithm for graphical model construction. In this figure, red indicates stronger correlation between expression of proteins and white indicates weaker

correlation. Due to the data scale and information density of this figure, small scale scrutiny is nearly impossible. Instead, the intent of this figure is to show modules of highly interconnected proteins within the larger biological network, of which two distinct groupings can be seen in the figure. This high level view of the data in effect converts a 283x283 peptide correlation heat map into a 2x2 biological network module heat map, with the ability to view overall correlation within the module comparisons.

Like dendrograms, balancing data scale with visualization density is crucial to the effect of heat maps. Figure 4 presents two opposite extremes of heat map scale and density. Figure 4A displays a relatively small scale dataset in a low density visualization, which is useful for individual row and/or column comparison. Figure 4B is the result of adding rows and columns to eventually create a large scale heat map. The comparisons in Figure 4B are so complex that the image becomes information dense, and blocks of comparison values among sorted axes translate to a large scale heat map with low density. A heat map with unsorted axes at the scale of Figure 4B could be unintelligible, since individual heat map comparisons are lost within the bigger picture, sacrificing small scale scope to show overarching trends.

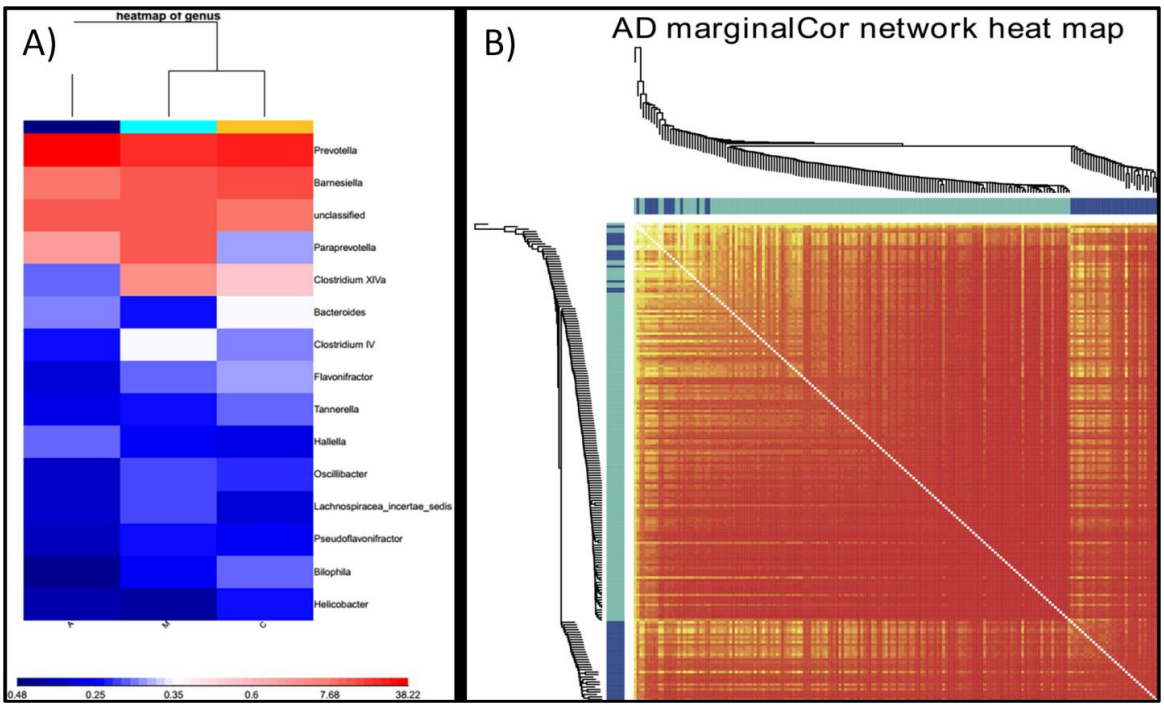

Figure 4: Heat map examples of scale and density. A) Figure 7 from Xue *et al*., 2017 showing aplysin intervention/protection in a rat model of ethanol-induced liver injury (Xue et al., 2017). B) Cropped Figure 4 from Wang *et al*., 2016 illustrating synaptic protein correlation from Alzheimer's disease patients (Wang et al., 2016).

Color. The use of color in diagrams is a widely used form of figure labeling. Color scheme is a critical component of figure design that can enhance or distort the display of information. Both of the examples in Figures 1 and 2 (more evidently in Figure 2) are represented in grayscale, meaning they use only white, black, and shades of gray as intermediates. Grayscale has the advantage of simplicity, while also avoiding issues that may arise from printing or color vision deficiencies, both needing to be considered, albeit printing concerns are increasingly situational with the digitization of scientific journals.

Color vision deficiencies are classified at three severity levels, stemming from the three types of color vision photoreceptors in the human eye: anomalous trichromacy (mild), dichromacy (moderate), monochromacy (severe) (Simunovic, 2010). The most common form of color vision deficiency is red-green color vision deficiency (either trichromacy or dichromacy), affecting ~8% of males and ~0.4% females, thus warranting the most consideration (Birch, 2012). Avoiding the use of red-green color schemes on figures meant to differentiate data will help to communicate data to a large proportion of color vision deficient individuals.

Even without issues that may arise from limitations on printer inks and color vision deficiencies, the color scheme of a figure must depict easily distinguishable data. Color schemes should also avoid the creation of "graphical puzzles" referenced by Tufte, 1983, where color representation in unintuitive (Tufte, 1983). Sequential color schemes like grayscale and color shading are often better at conveying natural visual hierarchy but may not adequately disjoint separate data points (Brewer, 2003; Harrower & Brewer, 2003). Alternatively, distinct colors portray visual hierarchy less intuitively even when presented in a logical arrangement such as wavelength, but may be better at distinguishing data points and can be arranged to be more qualitative so that adjacent colors are sufficiently distinct (Brewer, 2003; Harrower & Brewer, 2003). Figure 5 demonstrates the color schemes mentioned here.

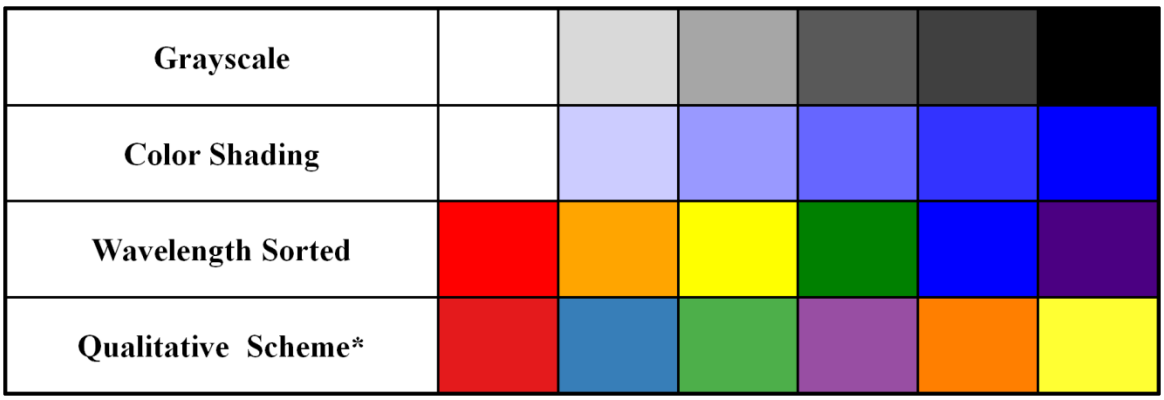

Figure 5: Figure color schemes that can be used to differentiate data. Started from the top, the first row shows a grayscale color scheme. The second row shows a color shading (white to blue) scheme. The third row shows a scheme of distinct colors that are sorted systematically, in this case by wavelength. The last row shows a scheme of distinct colors specifically chosen to highlight differences between different data (*Qualitative Scheme "6-class Set1" from colorbrower2.org (Brewer, 2003; Harrower & Brewer, 2003)).

Another aspect of dendrogram and heat map labeling which is not discussed below since it is less common is the use of patterns (or textures) in place of solid colors. In addition to many of the same considerations for color schemes, pattern usage entails its own set of variables. Employing patterns for diagram labeling is often done categorically since an intuitive hierarchical order is more difficult to achieve than simple color adjustments. Sometimes, patterns are used because publication requirements dictate a grayscale image and grayscale shading would falsely imply data contiguity or hierarchy. If this is the case, care must be taken to avoid creating a graphical puzzle with too many patterns that require definition.

The choice of an effective color scheme for dendrogram and heat map visualizations is largely dependent on its underlying data. It is important to remember that coloration of a diagram is a space-efficient form of labeling; every applied color should have meaning. Applying color for the sole sake of decoration runs the risk of suggesting meaning where there is none, or obscuring data with colorful distractions. A goal of data visualization is, as Tufte writes, "Above all else show the data" (Tufte, 1983). However, that is not to say that dendrograms and heat maps should not be visually appealing.

Use in dendrograms. Dendrograms could benefit from any of the color schemes in Figure 5, depending on the purpose of the figure. Figure 6 shows four examples of how different color schemes are used to convey information. Figure 6A from Tully and Potapov, 2015 (Tully & Potapov, 2015) depicts two dendrogram methods comparing morphological trait measurements of strains of *Folsomia candida*, a species of soil arthropod commonly known as "Springtails." As is done frequently with dendrogram color schemes, color information is applied to only the nodes of Figure 6A while branching shows relatedness between the node colors. With circle size representing absolute value magnitude of character trait measurements, binary grayscale (black or white) is used to depict positive or negative values. Different shading levels could have been used in this figure in place of circle size to illustrate magnitude, but the simplicity of black vs. white is an easily perceivable guide to the important differences between strains. The authors' choice of using circle size instead of shading to show measurement magnitude puts emphasis on the contrast between positive (black) and negative (white) values, which is a binary use of grayscale rather than a gradation grayscale scheme.

Figure 6B from Betancur-R *et al*., 2013 (Betancur-R et al., 2013) shows phylogenetic relationships among families of Syngnathiformes, an order of bony fish. As in Figure 6A, Figure 6B has a color scheme applied to only dendrogram nodes, using essentially a two-color shading scheme from blue to magenta (with a purple intermediate) to portray low and high scales of the number of species examined by conserved gene sequencing in each family node. Importantly, the use of a two-color shading scheme is used in this figure to also differentiate data size values from bootstrap values, which are represented with a grayscale scheme. The color shading scheme in this figure serves two purposes, to give a sense of scale to the family data sets and provide a categorical difference from the bootstrap values.

Figure 6C from Plazzi *et al*., 2011 (Plazzi, Ceregato, Taviani, & Passamonti, 2011) illustrates bivalve phylogeny with order level nodes sorted by a color scheme that divides nodes into subclasses. Unlike Figures 6A and 6B, the color scheme of Figure 6C is applied to dendrogram branches as well as nodes, conveying subclass branching points. The wavelength sorted color scheme illustrates two points; first, it clarifies separation between each of the subclasses represented in the dendrogram since distinct hues are easily distinguishable. And second, wavelength sorting conveys a linear relatedness progression from the top to bottom of the dendrogram. The categorical nature of a sorted color scheme is the main strength over grayscale or color shading schemes that are generally better at conveying data scaling or progression (Brewer, 2003; Harrower & Brewer, 2003).

Figure 6D from Alibhai *et al*., 2017 (Alibhai, Jewell, & Evans, 2017) shows the relatedness of puma footprints based on a morphological identification protocol and

clustering algorithm. Figure 6D uses a qualitative color scheme applied to dendrogram nodes and some of the more exterior branches to differentiate puma footprints belonging to individual pumas according to the authors' identification methods. Similar to the wavelength sorted color scheme of Figure 6C, this qualitative color scheme successfully separates different groups. In contrast to Figure 6C, the color scheme here does not have an intuitive order, which means that relatedness inferences must come from the dendrogram layout.

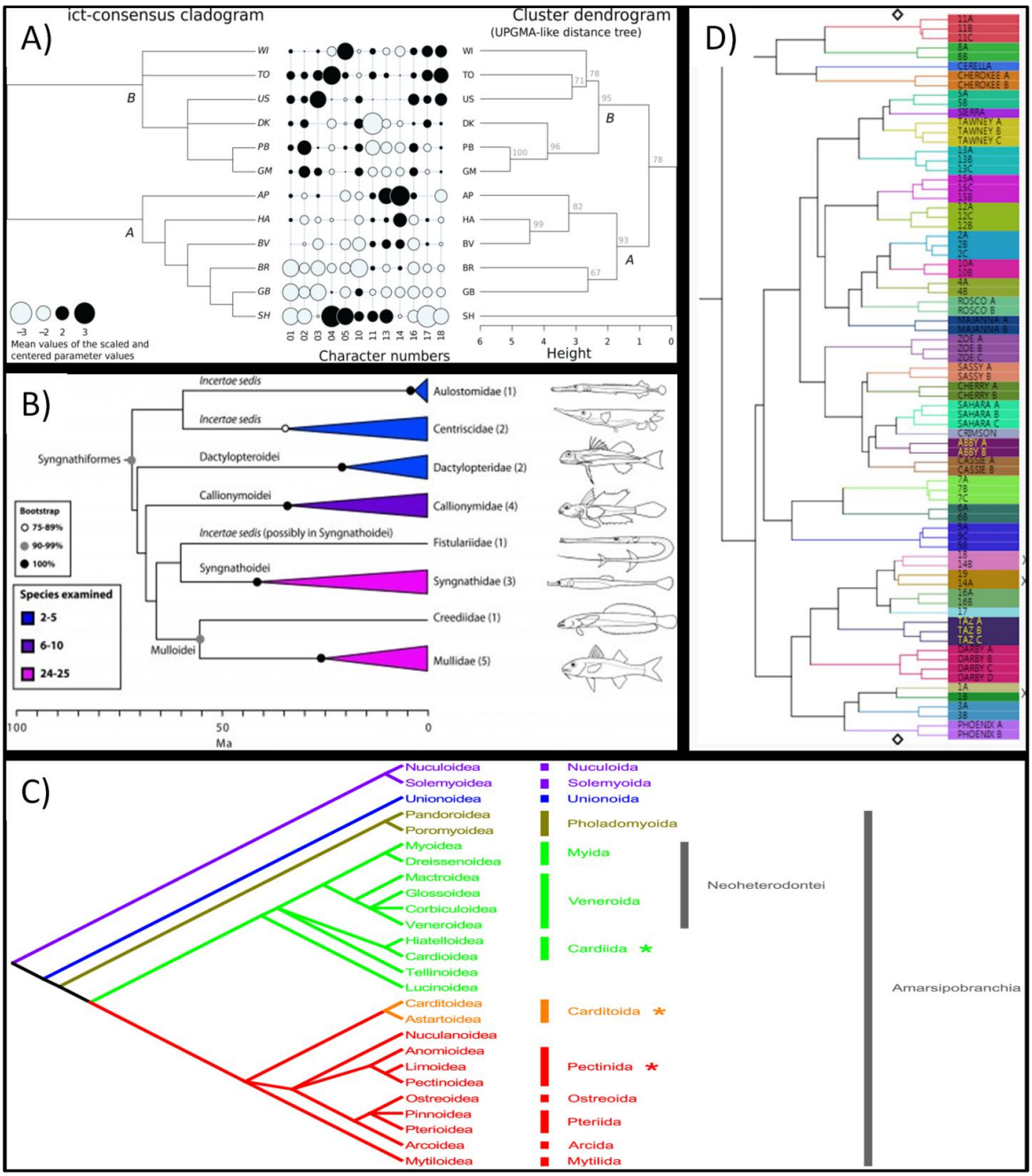

Figure 6: Dendrogram examples of color. A) Figure 5 from Tully and Potapov, 2015 depicitning morphological train measurements from strains of *Folsomia candida* (Tully & Potapov, 2015). B) Figure 4 from Betancur-R *et al*., 2013 showing phylogenetic relationships among families of Syngnathiformes (Betancur-R et al., 2013). C) Figure 7 from Plazzi *et al*., 2011 illustrating bivalve phylogeny (Plazzi et al., 2011). D) Cropped

Figure 6 from Alibhai *et al*., 2017 showing puma footprint relationships (Alibhai et al., 2017).

Use in heat maps. In the context of heat maps, color schemes need to have an intuitive progression that displays a hierarchy of responses, ruling out the use of strictly qualitative color schemes as used in Figure 6D. Heat map color schemes need opposing ends to represent high and low value extremes of a data range, as well as intermediate colors or shades to represent the full dataset. Figure 7 shows four examples of heat map color schemes with the innate ability to differentiate a range of data values. The simplest of the four examples, Figure 7A from Eneslätt *et al*., 2012 (Eneslätt et al., 2012), uses five distinct grayscale shades in a heat map to show all vs. all (of their experiment) T cell marker frequency correlation among human donors. Using a set number of distinct shades or colors is a very common complexity reduction practice of heat map construction, establishing a small set of colors to plot (and view) instead of a contiguous color range normalized to data values. In the case of Figure 7A, a small number of distinct grayscale shades are being used categorically, to represent bins of correlation coefficient ranges. Since this figure is composed of nonnegative correlation coefficients, higher values are of greater importance than lower values and can be quickly discerned as darker regions. The one-tailed nature of this dataset is ideal for the fading character of a grayscale color scheme (or any single color shading scheme), but not all data are equally amenable.

Similar to the use of a small number of grayscale bins used in Figure 7A, Figures 7B from Zhai *et al*., 2013 (Zhai, Yao, & Wang, 2013) and 7C from Gerzova *et al*., 2014 (Gerzova et al., 2014) use a small number of color shaded bins to represent correlation

coefficients and fold changes, respectively. However, both Figures 7B and 7C are constructed from two-tailed datasets so two color shading schemes are combined to portray divergence. The red-green color scheme of Figure 7B illustrates relative gene expression levels in mouse cell line cultures in response to different growth media over time, relative to a single day of growth in control media. The readily perceivable hierarchical aspect that made grayscale successful for heat maps is shared by two-color shading schemes. By using two diverging color hues to represent opposite ends of the data spectrum instead of the shading of a single hue, the importance of opposing data range tails can be more intuitively recognized. One drawback that Figure 7B (and many other published heat maps) suffers from is the confusion that could be caused by the prevalence of red-green color vision deficiency discussed in Section IIIC. Figure 7C circumvents this common color vision complication with the use of a red-blue color shading scheme to characterize correlation coefficients between bacterial families and antibiotic resistance genes. Unlike Figure 7A, the correlation coefficients of Figure 7C include both positive and negative correlation values, indicating the propensity or disinclination of a bacterial family to contain a specific antibacterial gene. Like Figure 7B, this red-blue color shading scheme effectively highlights opposing ends of a two-tailed dataset with two distinct color hues.

Figure 7D from Frank *et al*., 2010 (Frank et al., 2010) shows an ordered color scheme heat map of similarity indices between time points of human axilla (armpit), groin, and nares (nostrils) individual microbiome samples as well as nares microbiome samples between multiple subjects. The wavelength sorted color scheme of Figure 7D provides the hierarchical progression that is required for heat map visualization but the

wider range of distinct color hues facilitates a more categorical view of the data range, similar to the distinct grayscale shades of Figure 7A. Also, unlike the very clear small set of color scheme values seen in Figures 7A-C, the sorted color scheme of Figure 7D is much more contiguous, allowing for intermediates between perceived distinct color hue categories. The increased complexity of an ordered color scheme that lends itself to the distinction of categories also decreases the perceptibility of opposing data tails that is featured in diverging two-color shading schemes.

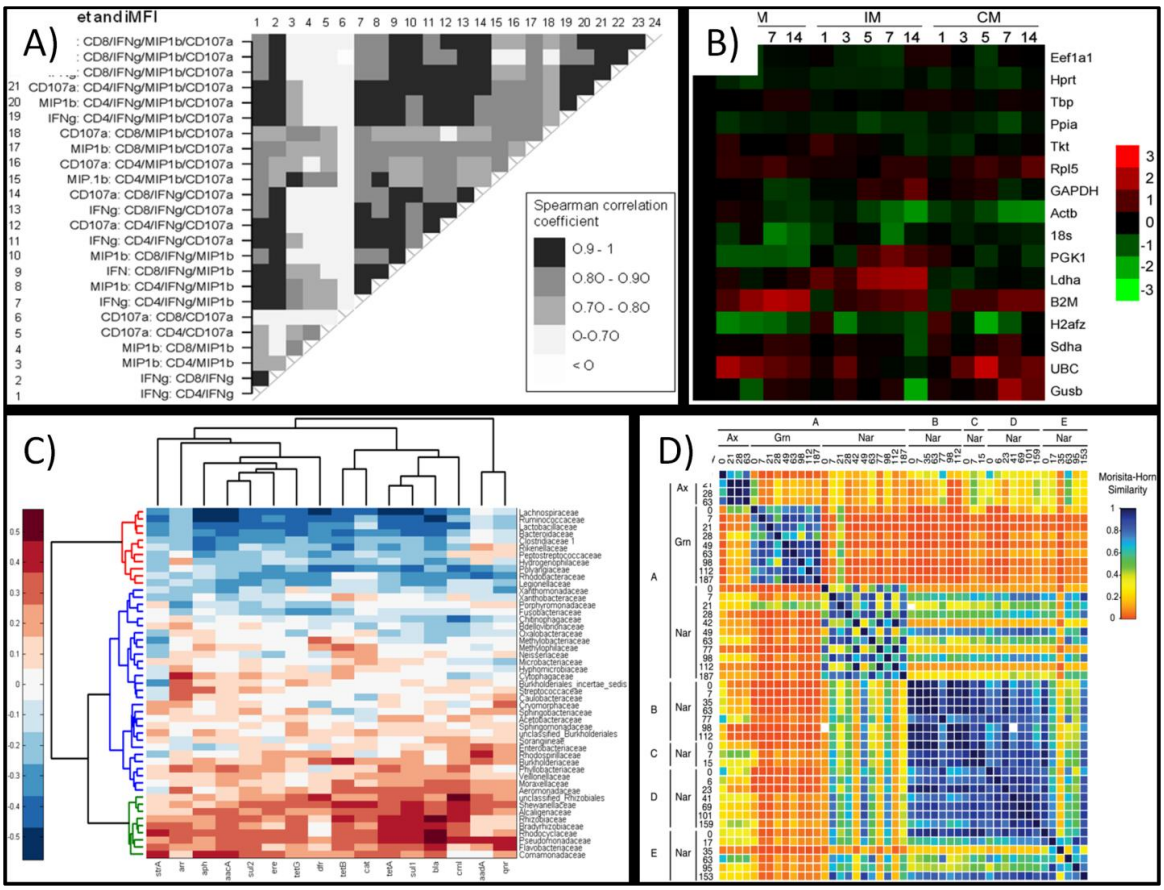

Figure 7: Heat map examples of color. A) Figure 8 from Eneslätt *et al*., 2012 showing T cell marker frequency correlation between human donors (Eneslätt et al., 2012). B) Figure 2 from Zhai *et al*., 2013 illustrates relative gene expression levels in mouse cell line cultures (Zhai et al., 2013). C) Figure 5 from Gerzova *et al*., 2014 characterizes correlation coefficients between bacterial families and antibiotic resistance genes (Gerzova et al., 2014). D) Figure 2 from Frank *et al*., 2010 shows similarity indices between human microbiome samples (Frank et al., 2010).

Innovation

In section III, we have seen how architectural elements of data, scale and density, and color are being used to visualize data trends in the context of dendrogram and heat

map strengths. However, with all their strengths dendrograms and heat maps struggle to display the multidimensionality of their data, perhaps because we are often forced to display data in static images. Hidden within nearly every dendrogram or heat map are many levels of equally legitimate versions of the same data display. The reality of clustered relationships is often more disordered than what is presented by the final dendrogram or heat map image, but data are forced into end-point clusters based on identity cutoffs (Kellom & Raymond, 2016). For instance, dendrograms of related genera are represented with each genus located at dendrogram nodes. However, not all members of the same taxonomy rank, protein family, etc. will cluster at the same level of sequence identity/similarity. The nodes of a dendrogram could be clustered at any clustering cutoff, each with differing dendrogram configurations and cluster abundances. Heat maps struggle in displaying this same concept, but from the opposing end of overlooking alternate cluster hierarchies rather than alternate cluster abundances. Every heat map has axis values that are dependent on the level of clustering specificity, which means there are usually many different heat map visualizations of the data hidden from view.

A cause of the multidimensionality challenges that dendrograms and heat maps struggle with is the concept of undersplitting and oversplitting (Flynn, Brown, Chain, MacIsaac, & Cristescu, 2015). Undersplitting creates nested clusters consisting of multiple subgroups of differing densities, each with the potential to form their own clusters under different clustering parameters (Li, Ye, Li, & Ng, 2010). Nested clusters exist because clustering algorithms cannot differentiate between intra-cluster diversity and inter-cluster diversity with complete reliability. This is not a slight at current clustering algorithms; it may be that making such distinctions for every complex dataset

is an impossible task. Since it is known that clustering algorithms can create nested clusters, they are sometimes pushed to very strict clustering parameters which can create the opposite problem of oversplitting, where multiple clusters represent the same functional group. Self-organizing map (SOM) techniques have been adapted to help circumvent clustering undersplitting and oversplitting but they do not solve the issue of visualizing multidimensionality since they work by reducing dimensionality (Nikkilä et al., 2002; Samsonova, Kok, & IJzerman, 2006).

There have been two main strategies to meet the challenge of visualizing multidimensionality of clustering data: 1) reorganizing cluster hierarchies by 'cutting' branches at multiple clustering cutoff levels, and 2) overlaying heat map values over dendrogram hierarchies.

Restructuring the clustering hierarchies by branch cutting allows for the visualization of dendrograms (and heat map matrices) that are not solely dependent on clustering cutoffs or self-organization. Instead, branch cutting algorithms scan nodes and exterior branches to check for signs of undersplitting or oversplitting derived from distance calculations, as well as checks of cluster abundances (Langfelder, Zhang, & Horvath, 2008). Branches are then cut or extended based on algorithm scoring. This process yields a dynamic set of node clusters that are not beholden to a single clustering cutoff, and can be used to create a heat map matrix with the same properties. While this process visualizes multidimensionality better than static clustering cutoffs or SOM clustering results, it is algorithmically-bound to display a single view of the results that may not suit everyone's needs. An alternative is a user-selected branch cutting process that yields nodes from multiple clustering cutoffs, affecting dendrogram shape and heat

map matrices (Vogogias, Kennedy, Archaumbault, Smith, & Currant, 2016). Branch cutting methods allow for multidimensional choice in data visualization, meaning dendrograms and heat maps are not restrained to a single viewing depth of the data. To use an analogy, traditional dendrograms and heat maps provide a snapshot view of a data clustering level while branch cutting methods create a collage.

The other strategy to overcome the challenge of visualizing multidimensionality in dendrograms and heat maps is to overlay heat map values on dendrogram hierarchies. This strategy is not to be confused with often-used practice of orienting the axis values of a heat map by means of dendrogram clustering, as seen in Figures 4 and 7C. In this configuration, the axis values are represented as nodes on separate dendrograms that are used to cluster columns and rows. While this often points out important data trends or conveys an extra level of context, it does not illustrate any information that could not be displayed if dendrogram and heat map were displayed separately.

Instead, the strategy of overlaying heat map values on dendrogram hierarchies means plotting a heat map response for every cluster (nodes and branching points) represented in a dendrogram. An early implementation of this concept was the Hierarchical Clustering Explorer (HCE) program (Seo & Shneiderman, 2002). With the appropriate data input, HCE can display heat map responses for any given clustering cutoff of a dendrogram. This ability makes it easy to browse the multidimensionality of the dataset and choose clustering cutoff visualizations that best represent the study narrative in static figure publications. Going beyond single clustering cutoff visualizations, some recent developments have made figures that assign colored blocks to label cluster abundances (or other characteristics) over many dendrogram clustering

cutoffs simultaneously (Agrafiotis, Bandyopadhyay, & Farnum, 2007; Ondov, Bergman, & Phillippy, 2011). This manner of dendrogram labeling creates a new figure type with an intuitive view of cluster characteristics and how they change for each clustering cutoff. When this concept is applied with heat maps as dendrogram labels the figures adopt the ability of tracking heat map responses though expanding and contracting clustering specificity in a single figure, a feature that is missing in standard dendrograms or heat maps (Kellom & Raymond, 2016). An influence that contributed to this concept was the plotting tool Circos and its prevalent use in showing genome characteristics (Krzywinski et al., 2009). One such example is in Figure 8 from den Bakker *et al*., 2013 (den Bakker et al., 2013), showing clade membership (red or blue) of genes for the genomes of *Listeria monocytogenes* strains (rings). While this figure is not arranged in a dendrogram configuration, clade membership can be seen to change between each of the strain genomes.

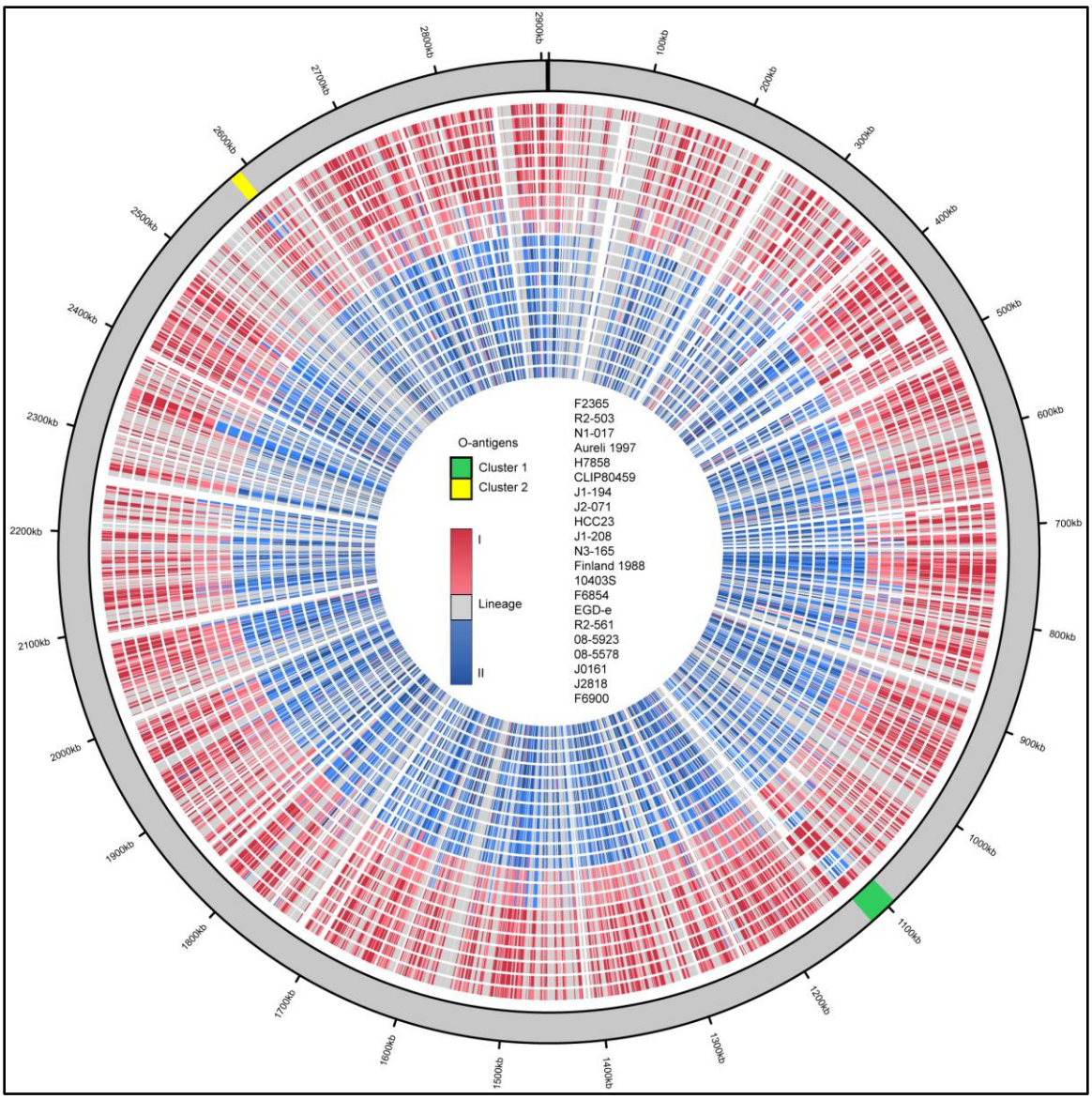

Figure 8: Figure 3 from den Bakker *et al*., 2013 showing clade memberships of genes in *Listeria monocytogenes* strain genomes (den Bakker et al., 2013).

Overlaying heat maps onto dendrogram configurations shows the multidimensionality of nodes and branching points in clustering datasets whereas branch cutting methods highlight a selection of multidimensional nodes, each strategy emphasizing multidimensionality from different perspectives. Branch cutting is designed

to underscore the idea that complex datasets do not usually have a homogeneous layer of pertinent clusters, so visualizations should not be based on clustering cutoffs. Overlaying heat maps on dendrogram configurations is designed to track the effects of cluster assignment on heat map responses, with the intention of detecting shifting trends in a gradational dataset. Each method conveys the complexity and dynamic nature of clustered hierarchical datasets in ways that are not possible in traditional dendrograms and heat maps.

Conclusion

Ultimately, figure design should be focused on the most effective method of communicating the narrative that emerges from data analysis. Maximizing the impact of a figure will require consideration of potential weaknesses and the options to mitigate them. Weaknesses of dendrograms and heat maps can come from design architecture, like those discussed in Section III, which affect the graphical flow of information. Alternatively, weaknesses can come from choosing a figure type that limits the accurate portrayal of the data. Traditional dendrograms and heat maps excel at displaying hierarchical relationships of clustered datasets through input data simplification, scale and density, and color, but often struggle to convey hierarchy multidimensionality by manipulation of these factors alone. In response, figure types have been developed to bridge this visualization gap and preserve the useful comparative qualities that are well-established by dendrograms and heat maps. While the visualization of extremely large datasets is still a problem for these new figure types, they have the potential to help locate interesting data trends in large datasets and visualize the narrowed down results. Innovations that fill the gaps of dendrograms and heat maps do not subvert the usefulness

of these long-established figures but they do give additional options for more effective science communication.

Acknowledgements

I thank Jason Raymond for helpful comments and proof-reading.

REFERENCES

Acland, A., Agarwala, R., Barrett, T., Beck, J., Benson, D. A., Bollin, C., … Zbicz, K. (2014). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, *42*(Database issue), D7–D17. https://doi.org/10.1093/nar/gkt1146

Agrafiotis, D. K., Bandyopadhyay, D., & Farnum, M. (2007). Radial Clustergrams: Visualizing the Aggregate Properties of Hierarchical Clusters. *Journal of Chemical Information and Modeling*, *47*(1), 69–75. https://doi.org/10.1021/ci600427x

Alibhai, S., Jewell, Z., & Evans, J. (2017). The challenge of monitoring elusive large carnivores: An accurate and cost-effective tool to identify and sex pumas (Puma concolor) from footprints. *PLOS ONE*, *12*(3), e0172065. https://doi.org/10.1371/journal.pone.0172065

Arief, V. N., DeLacy, I. H., Basford, K. E., & Dieters, M. J. (2017). Application of a dendrogram seriation algorithm to extract pattern from plant breeding data. *Euphytica*, *213*(4), 85. https://doi.org/10.1007/s10681-017-1870-z

Bailey, A. C., Kellom, M., Poret-Peterson, A. T., Noonan, K., Hartnett, H. E., & Raymond, J. (2014a). Draft Genome Sequence of Bacillus sp. Strain BSC154, Isolated from Biological Soil Crust of Moab, Utah. *Genome Announcements*, *2*(6), e01198-14. https://doi.org/10.1128/genomeA.01198-14

Bailey, A. C., Kellom, M., Poret-Peterson, A. T., Noonan, K., Hartnett, H. E., & Raymond, J. (2014b). Draft Genome Sequence of Massilia sp. Strain BSC265, Isolated from Biological Soil Crust of Moab, Utah. *Genome Announcements*, *2*(6), e01199-14. https://doi.org/10.1128/genomeA.01199-14

Bailey, A. C., Kellom, M., Poret-Peterson, A. T., Noonan, K., Hartnett, H. E., & Raymond, J. (2014c). Draft Genome Sequence of Microvirga sp. Strain BSC39, Isolated from Biological Soil Crust of Moab, Utah. *Genome Announcements*, *2*(6), e01197-14. https://doi.org/10.1128/genomeA.01197-14

Betancur-R, R., Broughton, R. E., Wiley, E. O., Carpenter, K., López, J. A., Li, C., … Ortí, G. (2013). The Tree of Life and a New Classification of Bony Fishes. *PLOS Currents Tree of Life*. https://doi.org/10.1371/currents.tol.53ba26640df0ccaee75bb165c8c26288

Birch, J. (2012). Worldwide prevalence of red-green color deficiency. *JOSA A*, *29*(3), 313–320. https://doi.org/10.1364/JOSAA.29.000313

Blount, Z. D., Barrick, J. E., Davidson, C. J., & Lenski, R. E. (2012). Genomic analysis of a key innovation in an experimental Escherichia coli population. *Nature*, *489*(7417), 513–518. https://doi.org/10.1038/nature11514

Bray, J. R., & Curtis, J. T. (1957). An Ordination of the Upland Forest Communities of Southern Wisconsin. *Ecological Monographs*, *27*(4), 325–349. https://doi.org/10.2307/1942268

Brewer, C. A. (2003). A Transition in Improving Maps: The ColorBrewer Example. *Cartography and Geographic Information Science*, *30*(2), 159–162. https://doi.org/10.1559/152304003100011126

Bryant, D. A., & Frigaard, N.-U. (2006). Prokaryotic photosynthesis and phototrophy illuminated. *Trends in Microbiology*, *14*(11), 488–496. https://doi.org/10.1016/j.tim.2006.09.001

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics*, *10*, 421. https://doi.org/10.1186/1471-2105-10-421

Cho, I., Yamanishi, S., Cox, L., Methé, B. A., Zavadil, J., Li, K., … Blaser, M. J. (2012). Antibiotics in early life alter the murine colonic microbiome and adiposity. *Nature*, *488*(7413), 621–626. https://doi.org/10.1038/nature11400

Cole, J. R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R. J., … Tiedje, J. M. (2009). The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Research*, *37*(suppl 1), D141–D145. https://doi.org/10.1093/nar/gkn879

Cole, J. R., Wang, Q., Fish, J. A., Chai, B., McGarrell, D. M., Sun, Y., … Tiedje, J. M. (2014). Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Research*, *42*(D1), D633–D642. https://doi.org/10.1093/nar/gkt1244

Darling, A. E., Jospin, G., Lowe, E., Matsen IV, F. A., Bik, H. M., & Eisen, J. A. (2014). PhyloSift: phylogenetic analysis of genomes and metagenomes. *PeerJ*, *2*(e243). https://doi.org/10.7717/peerj.243

den Bakker, H. C., Desjardins, C. A., Griggs, A. D., Peters, J. E., Zeng, Q., Young, S. K., … Wiedmann, M. (2013). Evolutionary Dynamics of the Accessory Genome of Listeria monocytogenes. *PLOS ONE*, *8*(6), e67511. https://doi.org/10.1371/journal.pone.0067511

Du, Q., Pan, W., Tian, J., Li, B., & Zhang, D. (2013). The UDP-Glucuronate Decarboxylase Gene Family in Populus: Structure, Expression, and Association Genetics. *PLOS ONE, 8*(4), e60880. https://doi.org/10.1371/journal.pone.0060880

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, *26*(19), 2460–2461. https://doi.org/10.1093/bioinformatics/btq461

Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, *95*(25), 14863–14868.

Elser, J. J., Navarro, M. B., Corman, J. R., Emick, H., Kellom, M., Laspoumaderes, C., Lee ZM, Poret-Peterson AT, Balseiro E, Modenutti, B. (2015). Community Structure and Biogeochemical Impacts of Microbial Life on Floating Pumice. *Applied and Environmental Microbiology*, *81*(5), 1542–1549. https://doi.org/10.1128/AEM.03160-14

Eneslätt, K., Normark, M., Björk, R., Rietz, C., Zingmark, C., Wolfraim, L. A., … Sjöstedt, A. (2012). Signatures of T Cells as Correlates of Immunity to Francisella tularensis. *PLOS ONE, 7*(3), e32367. https://doi.org/10.1371/journal.pone.0032367

Felsenstein, J. (2004). *Inferring Phylogenies*. Sunderland, MA: Sinauer associates.

Florek, K., Łukaszewicz, J., Perkal, J., Steinhaus, H., & Zubrzycki, S. (1951). Sur la liaison et la division des points d'un ensemble fini. *Colloquium Mathematicae*, *2*(3–4), 282–285.

Flynn, J. M., Brown, E. A., Chain, F. J. J., MacIsaac, H. J., & Cristescu, M. E. (2015). Toward accurate molecular identification of species in complex environmental samples: testing the performance of sequence filtering and clustering methods. *Ecology and Evolution*, *5*(11), 2252–2266. https://doi.org/10.1002/ece3.1497

Frank, D. N., Feazel, L. M., Bessesen, M. T., Price, C. S., Janoff, E. N., & Pace, N. R. (2010). The Human Nasal Microbiota and Staphylococcus aureus Carriage. *PLOS ONE, 5*(5), e10598. https://doi.org/10.1371/journal.pone.0010598

Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, *28*(23), 3150–3152. https://doi.org/10.1093/bioinformatics/bts565

Gerzova, L., Videnska, P., Faldynova, M., Sedlar, K., Provaznik, I., Cizek, A., & Rychlik, I. (2014). Characterization of Microbiota Composition and Presence of Selected Antibiotic Resistance Genes in Carriage Water of Ornamental Fish. *PLOS ONE, 9*(8), e103865. https://doi.org/10.1371/journal.pone.0103865

Ghodsi, M., Liu, B., & Pop, M. (2011). DNACLUST: accurate and efficient clustering of phylogenetic marker genes. *BMC Bioinformatics*, *12*, 271. https://doi.org/10.1186/1471-2105-12-271

Gisbrecht, A., & Hammer, B. (2015). Data visualization by nonlinear dimensionality reduction. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *5*(2), 51–73. https://doi.org/10.1002/widm.1147

Goll, J., Rusch, D. B., Tanenbaum, D. M., Thiagarajan, M., Li, K., Methé, B. A., & Yooseph, S. (2010). METAREP: JCVI metagenomics reports—an open source tool for high-performance comparative metagenomics. *Bioinformatics*, *26*(20), 2631–2632. https://doi.org/10.1093/bioinformatics/btq455

Gronau, I., & Moran, S. (2007). Optimal implementations of UPGMA and other common clustering algorithms. *Information Processing Letters*, *104*(6), 205–210. https://doi.org/10.1016/j.ipl.2007.07.002

Han, Y., Liu, Y., Wang, H., & Liu, X. (2017). The Evolution of Vicia ramuliflora (Fabaceae) at Tetraploid and Diploid Levels Revealed with FISH and RAPD. *PLOS ONE*, *12*(1), e0170695. https://doi.org/10.1371/journal.pone.0170695

Harrower, M., & Brewer, C. A. (2003). ColorBrewer.org: An Online Tool for Selecting Colour Schemes for Maps. *The Cartographic Journal*, *40*(1), 27–37. https://doi.org/10.1179/000870403235002042

Heo, G. E., Kang, K. Y., Song, M., & Lee, J.-H. (2017). Analyzing the field of bioinformatics with the multi-faceted topic modeling technique. *BMC Bioinformatics*, *18*(7), 251. https://doi.org/10.1186/s12859-017-1640-x

Hernández-Prieto, M. A., Schön, V., Georg, J., Barreira, L., Varela, J., Hess, W. R., & Futschik, M. E. (2012). Iron Deprivation in Synechocystis: Inference of Pathways, Non-coding RNAs, and Regulatory Elements from Comprehensive Expression Profiling. *G3: Genes|Genomes|Genetics*, *2*(12), 1475–1495. https://doi.org/10.1534/g3.112.003863

Hernández-Prieto, M. A., Semeniuk, T. A., Giner-Lamia, J., & Futschik, M. E. (2016). The Transcriptional Landscape of the Photosynthetic Model Cyanobacterium Synechocystis sp. PCC6803. *Scientific Reports*, *6*. https://doi.org/10.1038/srep22168

Hetherington, A. J., Sherratt, E., Ruta, M., Wilkinson, M., Deline, B., & Donoghue, P. C. J. (2015). Do cladistic and morphometric data capture common patterns of morphological disparity? *Palaeontology*, *58*(3), 393–399. https://doi.org/10.1111/pala.12159

Huse, S. M., Welch, D. M., Morrison, H. G., & Sogin, M. L. (2010). Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environmental Microbiology*, *12*(7), 1889–1898. https://doi.org/10.1111/j.1462-2920.2010.02193.x

Huson, D. H., Auch, A. F., Qi, J., & Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Research*, *17*(3), 377–386. https://doi.org/10.1101/gr.5969107

Jackson, D. A. (1997). Compositional Data in Community Ecology: The Paradigm or Peril of Proportions? *Ecology*, *78*(3), 929–940. https://doi.org/10.1890/0012-9658(1997)078[0929:CDICET]2.0.CO;2

Kellom, M., & Raymond, J. (2016). Using Dendritic Heat Maps to Simultaneously Display Genotype Divergence with Phenotype Divergence. *PLOS ONE*, *11*(8), e0161292. https://doi.org/10.1371/journal.pone.0161292

Kellom, M., & Raymond, J. (2017). Using cluster edge counting to aggregate iterations of centroid-linkage clustering results and avoid large distance matrices. *Journal of Biological Methods*, *4*(1), e68.

Koenig, J. E., Spor, A., Scalfone, N., Fricker, A. D., Stombaugh, J., Knight, R., … Ley, R. E. (2011). Succession of microbial consortia in the developing infant gut microbiome. *Proceedings of the National Academy of Sciences*, *108*(Supplement 1), 4578–4585. https://doi.org/10.1073/pnas.1000081107

Kopf, M., Klähn, S., Scholz, I., Matthiessen, J. K. F., Hess, W. R., & Voß, B. (2014). Comparative Analysis of the Primary Transcriptome of Synechocystis sp. PCC 6803. *DNA Research*, dsu018. https://doi.org/10.1093/dnares/dsu018

Krebs, J. E., Goldstein, E. S., & Kilpatrick, S. T. (2009). *Lewin's GENES X*. Jones & Bartlett Publishers.

Kreft, Ł., Botzki, A., Coppens, F., Vandepoele, K., & Van Bel, M. (n.d.). PhyD3: a phylogenetic tree viewer with extended phyloXML support for functional genomics data visualization. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btx324

Krzywinski, M., Birol, I., Jones, S. J., & Marra, M. A. (2012). Hive plots—rational approach to visualizing networks. *Briefings in Bioinformatics*, *13*(5), 627–644. https://doi.org/10.1093/bib/bbr069

Krzywinski, M., Schein, J., Birol, İ., Connors, J., Gascoyne, R., Horsman, D., … Marra, M. A. (2009). Circos: An information aesthetic for comparative genomics. *Genome Research*, *19*(9), 1639–1645. https://doi.org/10.1101/gr.092759.109

Laczny, C. C., Pinel, N., Vlassis, N., & Wilmes, P. (2014). Alignment-free Visualization of Metagenomic Data by Nonlinear Dimension Reduction. *Scientific Reports*, *4*, 4516. https://doi.org/10.1038/srep04516

Laczny, C. C., Sternal, T., Plugaru, V., Gawron, P., Atashpendar, A., Margossian, H. H., … Wilmes, P. (2015). VizBin - an application for reference-independent visualization and human-augmented binning of metagenomic data. *Microbiome*, *3*, 1. https://doi.org/10.1186/s40168-014-0066-1

Langfelder, P., Zhang, B., & Horvath, S. (2008). Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics*, *24*(5), 719–720. https://doi.org/10.1093/bioinformatics/btm563

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., … Higgins, D. G. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics*, *23*(21), 2947–2948. https://doi.org/10.1093/bioinformatics/btm404

Li, X., Ye, Y., Li, M. J., & Ng, M. K. (2010). On cluster tree for nested and multi-density data clustering. *Pattern Recognition*, *43*(9), 3130–3143. https://doi.org/10.1016/j.patcog.2010.03.020

Liu, S., Maljovec, D., Wang, B., Bremer, P. T., & Pascucci, V. (2017). Visualizing High-Dimensional Data: Advances in the Past Decade. *IEEE Transactions on Visualization and Computer Graphics*, *23*(3), 1249–1268. https://doi.org/10.1109/TVCG.2016.2640960

Ma, P., & Sun, X. (2015). Leveraging for big data regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, *7*(1), 70–76. https://doi.org/10.1002/wics.1324

Macnaughton-Smith, P., Williams, W. T., Dale, M. B., & Mockett, L. G. (1964). Dissimilarity Analysis: a new Technique of Hierarchical Sub-division. *Nature*, *202*(4936), 1034–1035. https://doi.org/10.1038/2021034a0

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E., Kubal, M., … Edwards, R. (2008). The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, *9*, 386. https://doi.org/10.1186/1471-2105-9-386

Meyer, J. R., Dobias, D. T., Weitz, J. S., Barrick, J. E., Quick, R. T., & Lenski, R. E. (2012). Repeatability and Contingency in the Evolution of a Key Innovation in

Phage Lambda. *Science*, *335*(6067), 428–432. https://doi.org/10.1126/science.1214449

Miladi, M., Junge, A., Costa, F., Seemann, S. E., Havgaard, J. H., Gorodkin, J., & Backofen, R. (2017). RNAscClust: clustering RNA sequences using structure conservation and graph based motifs. *Bioinformatics*, *33*(14), 2089–2096. https://doi.org/10.1093/bioinformatics/btx114

Mitchell-Olds, T., Willis, J. H., & Goldstein, D. B. (2007). Which evolutionary processes influence natural genetic variation for phenotypic traits? *Nature Reviews Genetics*, *8*(11), 845–856. https://doi.org/10.1038/nrg2207

Morris, S. A., Asnake, B., & Yen, G. G. (2003). Dendrogram Seriation Using Simulated Annealing. *Information Visualization*, *2*(2), 95–104. https://doi.org/10.1057/palgrave.ivs.9500042

Nikkilä, J., Törönen, P., Kaski, S., Venna, J., Castrén, E., & Wong, G. (2002). Analysis and visualization of gene expression data using Self-Organizing Maps. *Neural Networks*, *15*(8), 953–966. https://doi.org/10.1016/S0893-6080(02)00070-9

Nodine, M. D., & Bartel, D. P. (2012). Maternal and paternal genomes contribute equally to the transcriptome of early plant embryos. *Nature*, *482*(7383), 94–97. https://doi.org/10.1038/nature10756

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., … Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, *44*(D1), D733–D745. https://doi.org/10.1093/nar/gkv1189

Ondov, B. D., Bergman, N. H., & Phillippy, A. M. (2011). Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics*, *12*, 385. https://doi.org/10.1186/1471-2105-12-385

Plazzi, F., Ceregato, A., Taviani, M., & Passamonti, M. (2011). A Molecular Phylogeny of Bivalve Mollusks: Ancient Radiations and Divergences as Revealed by Mitochondrial Genes. *PLOS ONE*, *6*(11), e27147. https://doi.org/10.1371/journal.pone.0027147

Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., … Gu, Y. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, *13*, 341. https://doi.org/10.1186/1471-2164-13-341

Samsonova, E. V., Kok, J. N., & IJzerman, A. P. (2006). TreeSOM: Cluster analysis in the self-organizing map. *Neural Networks*, *19*(6), 935–949. https://doi.org/10.1016/j.neunet.2006.05.003

Schloissnig, S., Arumugam, M., Sunagawa, S., Mitreva, M., Tap, J., Zhu, A., … Bork, P. (2013). Genomic variation landscape of the human gut microbiome. *Nature*, *493*(7430), 45–50. https://doi.org/10.1038/nature11711

Schloss, P. D. (2010). The Effects of Alignment Quality, Distance Calculation Method, Sequence Filtering, and Region on the Analysis of 16S rRNA Gene-Based Studies. *PLOS Computational Biology*, *6*(7), e1000844. https://doi.org/10.1371/journal.pcbi.1000844

Schonlau, M. (2004). Visualizing non-hierarchical and hierarchical cluster analyses with clustergrams. *Computational Statistics*, *19*(1), 95–111. https://doi.org/10.1007/BF02915278

Sehgal, D., Vikram, P., Sansaloni, C. P., Ortiz, C., Pierre, C. S., Payne, T., … Singh, S. (2015). Exploring and Mobilizing the Gene Bank Biodiversity for Wheat Improvement. *PLOS ONE*, *10*(7), e0132112. https://doi.org/10.1371/journal.pone.0132112

Seo, J., & Shneiderman, B. (2002). Interactively exploring hierarchical clustering results [gene identification]. *Computer*, *35*(7), 80–86. https://doi.org/10.1109/MC.2002.1016905

Shcolnick, S., Summerfield, T. C., Reytman, L., Sherman, L. A., & Keren, N. (2009). The Mechanism of Iron Homeostasis in the Unicellular Cyanobacterium Synechocystis sp. PCC 6803 and Its Relationship to Oxidative Stress. *Plant Physiology*, *150*(4), 2045–2056. https://doi.org/10.1104/pp.109.141853

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., … Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, *7*(1), 539. https://doi.org/10.1038/msb.2011.75

Simunovic, M. P. (2010). Colour vision deficiency. *Eye*, *24*(5), 747–755. https://doi.org/10.1038/eye.2009.251

Singh, A. K., McIntyre, L. M., & Sherman, L. A. (2003). Microarray Analysis of the Genome-Wide Response to Iron Deficiency and Iron Reconstitution in the Cyanobacterium Synechocystis sp. PCC 6803. *Plant Physiology*, *132*(4), 1825–1839. https://doi.org/10.1104/pp.103.024018

Sneath, P. H. A. (1957). The Application of Computers to Taxonomy. *Microbiology*, *17*(1), 201–226. https://doi.org/10.1099/00221287-17-1-201

Sokal, R. R. (1958). A statistical method for evaluating systematic relationships. *The University of Kansas Science Bulletin*, *38*(2), 1409–1438.

Swingley, W. D., Meyer-Dombard, D. R., Shock, E. L., Alsop, E. B., Falenski, H. D., Havig, J. R., & Raymond, J. (2012). Coordinating Environmental Genomics and Geochemistry Reveals Metabolic Transitions in a Hot Spring Ecosystem. *PLOS ONE*, *7*(6), e38108. https://doi.org/10.1371/journal.pone.0038108

The Cancer Genome Atlas Research Network. (2013). Integrated genomic characterization of endometrial carcinoma. *Nature*, *497*(7447), 67–73. https://doi.org/10.1038/nature12113

Tufte, E. R. (1983). *The Visual Display of Quantitative Information*. Cheshire, Connecticut: Graphics Press.

Tully, T., & Potapov, M. (2015). Intraspecific Phenotypic Variation and Morphological Divergence of Strains of Folsomia candida (Willem) (Collembola: Isotomidae), the "Standard" Test Springtaill. *PLOS ONE*, *10*(9), e0136047. https://doi.org/10.1371/journal.pone.0136047

Vaismoradi, M., Turunen, H., & Bondas, T. (2013). Content analysis and thematic analysis: Implications for conducting a qualitative descriptive study. *Nursing & Health Sciences*, *15*(3), 398–405. https://doi.org/10.1111/nhs.12048

Vogogias, A., Kennedy, J., Archaumbault, D., Smith, V. A., & Currant, H. (2016). MLCut : exploring Multi-Level Cuts in dendrograms for biological data. Eurographics Association. https://doi.org/http://dx.doi.org/10.2312/cgvc.20161288

Wang, D., Hsieh, M., & Li, W.-H. (2005). A General Tendency for Conservation of Protein Length Across Eukaryotic Kingdoms. *Molecular Biology and Evolution*, *22*(1), 142–147. https://doi.org/10.1093/molbev/msh263

Wang, T., Ren, Z., Ding, Y., Fang, Z., Sun, Z., MacDonald, M. L., … Chen, W. (2016). FastGGM: An Efficient Algorithm for the Inference of Gaussian Graphical Model in Biological Networks. *PLOS Computational Biology*, *12*(2), e1004755. https://doi.org/10.1371/journal.pcbi.1004755

Wegener, K. M., Singh, A. K., Jacobs, J. M., Elvitigala, T., Welsh, E. A., Keren, N., … Pakrasi, H. B. (2010). Global Proteomics Reveal an Atypical Strategy for Carbon/Nitrogen Assimilation by a Cyanobacterium Under Diverse

Environmental Perturbations. *Molecular & Cellular Proteomics*, *9*(12), 2678–2689. https://doi.org/10.1074/mcp.M110.000109

Wen, W., Franco, J., Chavez-Tovar, V. H., Yan, J., & Taba, S. (2012). Genetic Characterization of a Core Set of a Tropical Maize Race Tuxpeño for Further Use in Maize Improvement. *PLOS ONE*, *7*(3), e32626. https://doi.org/10.1371/journal.pone.0032626

Wilkinson, L., & Friendly, M. (2009). The History of the Cluster Heat Map. *The American Statistician*, *63*(2), 179–184. https://doi.org/10.1198/tas.2009.0033

Williams, W. T., & Lambert, J. M. (1966). Multivariate Methods in Plant Ecology: V. Similarity Analyses and Information-Analysis. *Journal of Ecology*, *54*(2), 427–445. https://doi.org/10.2307/2257960

Williams, W. T., & Lance, G. N. (1967). A general theory of classification sorting strategies: 1. Hierarchical systems, 2. Clustering systems. *Computer Journal*, *9*(10), 373–380.

Wolfe, J. M., Daley, A. C., Legg, D. A., & Edgecombe, G. D. (2016). Fossil calibrations for the arthropod Tree of Life. *Earth-Science Reviews*, *160*, 43–110. https://doi.org/10.1016/j.earscirev.2016.06.008

Wooley, J. C., Godzik, A., & Friedberg, I. (2010). A Primer on Metagenomics. *PLOS Computational Biology*, *6*(2), e1000667. https://doi.org/10.1371/journal.pcbi.1000667

Wooley, J. C., & Lin, H. S. (2005). *On the Nature of Biological Data*. National Academies Press (US). Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK25464/

Wooley, J. C., & Ye, Y. (2010). Metagenomics: Facts and Artifacts, and Computational Challenges. *Journal of Computer Science and Technology*, *25*(1), 71–81. https://doi.org/10.1007/s11390-010-9306-4

Wu, G. D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y.-Y., Keilbaugh, S. A., … Lewis, J. D. (2011). Linking Long-Term Dietary Patterns with Gut Microbial Enterotypes. *Science*, *334*(6052), 105–108. https://doi.org/10.1126/science.1208344

Xu, L., Chen, H., Hu, X., Zhang, R., Zhang, Z., & Luo, Z. W. (2006). Average Gene Length Is Highly Conserved in Prokaryotes and Eukaryotes and Diverges Only Between the Two Kingdoms. *Molecular Biology and Evolution*, *23*(6), 1107–1108. https://doi.org/10.1093/molbev/msk019

Xue, M., Liu, Y., Lyu, R., Ge, N., Liu, M., Ma, Y., & Liang, H. (2017). Protective effect of aplysin on liver tissue and the gut microbiota in alcohol-fed rats. *PLOS ONE*, *12*(6), e0178684. https://doi.org/10.1371/journal.pone.0178684

Yu, G., Smith, D. K., Zhu, H., Guan, Y., & Lam, T. T.-Y. (2017). ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, *8*(1), 28–36. https://doi.org/10.1111/2041-210X.12628

Zhai, Z., Yao, Y., & Wang, Y. (2013). Importance of Suitable Reference Gene Selection for Quantitative RT-PCR during ATDC5 Cells Chondrocyte Differentiation. *PLOS ONE*, *8*(5), e64786. https://doi.org/10.1371/journal.pone.0064786

Zhao, Y., Tang, H., & Ye, Y. (2012). RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. *Bioinformatics*, *28*(1), 125–126. https://doi.org/10.1093/bioinformatics/btr595

APPENDIX A

FIGURE PERMISSIONS

All figures in Chapters 3 (*Journal of Biological Methods*), 4 (*PLOS One*), and 5 (various *PLOS* journals) are published under Creative Commons Attribution License, giving permission to: A) copy and redistribute the material in any medium or format, and B) remix, transform, and build upon the material for any purpose, even commercially.

Creative Commons Legal Code

Attribution 4.0 International

Official translations of this license are available in other languages.

Creative Commons Corporation ("Creative Commons") is not a law firm and does not provide legal services or legal advice. Distribution of Creative Commons public licenses does not create a lawyer-client or other relationship. Creative Commons makes its licenses and related information available on an "as-is" basis. Creative Commons gives no warranties regarding its licenses, any material licensed under their terms and conditions, or any related information. Creative Commons disclaims all liability for damages resulting from their use to the fullest extent possible.

Using Creative Commons Public Licenses

Creative Commons public licenses provide a standard set of terms and conditions that creators and other rights holders may use to share original works of authorship and other material subject to copyright and certain other rights specified in the public license below. The following considerations are for informational purposes only, are not exhaustive, and do not form part of our licenses.

__Considerations for licensors:__ Our public licenses are intended for use by those authorized to give the public permission to use material in ways otherwise restricted by copyright and certain other rights. Our licenses are irrevocable. Licensors should read and understand the terms and conditions of the license they choose before applying it. Licensors should also secure all rights necessary before applying our licenses so that the public can reuse the material as expected. Licensors should clearly mark any material not subject to the license. This includes other CC-licensed material, or material used under an exception or limitation to copyright. More considerations for licensors.

__Considerations for the public:__ By using one of our public licenses, a licensor grants the public permission to use the licensed material under specified terms and conditions. If the licensor's permission is not necessary for any reason–for example, because of any applicable exception or limitation to copyright–then that use is not regulated by the license. Our licenses grant only permissions under copyright and certain other rights that a licensor has authority to grant. Use of the licensed material may still be restricted for other reasons, including because others have copyright or other rights in the material. A licensor may make special requests, such as asking that all changes be marked or described. Although not required by our licenses, you are encouraged to respect those requests where reasonable. More considerations for the public.

Creative Commons Attribution 4.0 International Public License

By exercising the Licensed Rights (defined below), You accept and agree to be bound by the terms and conditions of this Creative Commons Attribution 4.0 International Public License ("Public License"). To the extent this Public License may be interpreted as a contract, You are granted the Licensed Rights in consideration of Your acceptance of these terms and conditions, and the Licensor grants You such rights in consideration of benefits the Licensor receives from making the Licensed Material available under these terms and conditions.

Chapter 4 Figures 8 and 9 are published in *PLOS One* under Creative Commons Attribution License. Figures 8 and 9 are originally published in *Applied and environmental microbiology* with permission given under Creative Commons Attribution License:

From: matthewkellom@gmail.com [mailto:matthewkellom@gmail.com] **On Behalf Of** Matthew Kellom
Sent: Monday, May 11, 2015 1:16 PM
To: ASM Journals
Subject: Attention: Permissions Department; Figures 3 and 4 of AEM 03160-14.

Attention: Permissions Department

Dear Sir or Madam:

I am preparing a manuscript for publication in *PLOS Computational Biology*. Because PLOS journals publish under a Creative Commons Attribution License (http://www.ploscompbiol.org/static/license.action) I am requesting explicit permission to publish the attached figure under such a license, which would enable any readers to freely reuse and reproduce the figure, with the caveat that it be cited as appearing in this PLOS Computational Biology article.

I would appreciate permission to publish the following in *PLOS Computational Biology* under the Creative Commons Attribution License:

Figures 3 and 4 of Elser *et al.* 2014. "Community structure and biogeochemical impacts of microbial life on floating pumice." AEM. 03160-14.

I am the original author of both of these figures. Unless you indicate otherwise, I will use the complete reference entered above as a credit line. For your convenience, a copy of this letter may serve as a release form; a duplicate copy may be saved for your records.

These explicit permissions to publish under the Creative Commons Attribution License are in addition to the standard requirements for re-use, please advise on any other requirements you have for reuse.

Sincerely,
Matthew Kellom

PERMISSION GRANTED
CONSENT ON AUTHOR PERMISSION (which you must obtain
AND APPROPRIATE CREDIT
American Society for Microbiology
Journals Department
Date May 11, 2015

Chapter 4 Figure 10 is published in *PLOS One* under Creative Commons Attribution License. Figure 10 is originally published in *Proceedings of the National Academy of Sciences* with permission given under Creative Commons Attribution License:

Gmail by Google

Matthew Kellom <matthewkellom@gmail.com>

Attention: Permissions Department; Figure 1 of Eisen et al. 1998. PNAS. 95(25):14863-14868.

PNAS Permissions <PNASPermissions@nas.edu> Mon, May 11, 2015 at 10:49 AM
To: Matthew Kellom <matthew.kellom@asu.edu>

Permission is granted for your use of the figure as described in your message. Please cite the PNAS article in full, and include "Copyright (1998) National Academy of Sciences, U.S.A." as a copyright note. Because this material published between 1993 and 2008, a copyright note is needed. Let us know if you have any questions.

Best regards,

Kay McLaughlin for

Diane Sullenberger

Executive Editor

PNAS

From: matthewkellom@gmail.com [mailto:matthewkellom@gmail.com] **On Behalf Of** Matthew Kellom
Sent: Monday, May 11, 2015 1:24 PM
To: PNAS Permissions
Subject: Attention: Permissions Department; Figure 1 of Eisen et al. 1998. PNAS. 95(25):14863-14868.

Attention: Permissions Department

Dear Sir or Madam:

I am preparing a manuscript for publication in *PLOS Computational Biology*. Because PLOS journals publish under a Creative Commons Attribution License (http://www.ploscompbiol.org/static/license.action) I am requesting explicit permission to publish the attached figure under such a license, which would enable any readers to freely reuse and reproduce the figure, with the caveat that it be cited as appearing in this PLOS Computational Biology article.

I would appreciate permission to publish the following in *PLOS Computational Biology* under the Creative Commons Attribution License:

Figure 1 of Eisen *et al.* 1998. "Cluster analysis and display of genome-wide expression patterns." PNAS. 95(25):14863-14868.

APPENDIX B

STATEMENT OF PERMISSION FROM CO-AUTHORS

All co-authors have granted their permission for my use of Chapters 2, 3 and 4.

APPENDIX C

CHAPTER 2 EXCEL FILE OF TRANSCRIPTOME RESPONSES

Consult attached file Chapter2_Supplemental_File_S1.xlsx using Microsoft Excel or other spreadsheet reading software.

APPENDIX D

CHAPTER 2 PERL SCRIPTS

Consult attached file Chapter2_Supplemental_File_S2.zip.

APPENDIX E

CHAPTER 3 CLUSTER AGGREGATION PERL SCRIPTS

Consult attached file Chapter3_Supplemental_File_1.pl.

APPENDIX F

CHAPTER 4 EXAMPLE OF THE TOP-DOWN CLUSTERING METHOD USED TO CONSTRUCT DENDRITIC HEAT MAPS

Consult attached file Chapter4_Supplemental_File_S1.pdf.

APPENDIX G

CHAPTER 4 PERL SCRIPTS AND DENDRITIC HEAT MAP IMAGES

Consult attached file Chapter4_Supplemental_File_S2.zip.

APPENDIX H

CHAPTER 4 PERL SCRIPTS AND DENDRITIC HEAT MAP IMAGES

Consult attached file Chapter4_Supplemental_File_S3.zip.