



## Interactive Visualization of Hierarchically Structured Data


Kris Sankaran & Susan Holmes

To cite this article: Kris Sankaran & Susan Holmes (2018) Interactive Visualization of Hierarchically Structured Data, Journal of Computational and Graphical Statistics, 27:3, 553-563, DOI: [10.1080/10618600.2017.1392866](https://doi.org/10.1080/10618600.2017.1392866)

To link to this article: <https://doi.org/10.1080/10618600.2017.1392866>

 View supplementary material [↗](#)


---

 Published online: 11 Oct 2018.

---

 Submit your article to this journal [↗](#)

---

 Article views: 549

---

 View related articles [↗](#)

---

 View Crossmark data [↗](#)

---

 Citing articles: 5 View citing articles [↗](#)

---



# Interactive Visualization of Hierarchically Structured Data

Kris Sankaran and Susan Holmes

Department of Statistics, Stanford University, Stanford, CA

## ABSTRACT

We introduce methods for visualization of data structured along trees, especially hierarchically structured collections of time series. To this end, we identify questions that often emerge when working with hierarchical data and provide an R package to simplify their investigation. Our key contribution is the adaptation of the visualization principles of focus-plus-context and linking to the study of tree-structured data. Our motivating application is to the analysis of bacterial time series, where an evolutionary tree relating bacteria is available a priori. However, we have identified common problem types where, if a tree is not directly available, it can be constructed from data and then studied using our techniques. We perform detailed case studies to describe the alternative use cases, interpretations, and utility of the proposed visualization methods.

## ARTICLE HISTORY

Received February 2017  
Revised September 2017

## KEYWORDS

D3; Focus-plus-context;  
Linking; R; Time-series;  
Tree-structured

## 1. Introduction

We introduce methods for visualization of data structured along trees, especially hierarchically structured collections of time series. We hope both to characterize generically useful techniques for interactively visualizing hierarchical data and to offer practical tools for implementing such displays. To this end, we identify questions that often emerge when working with hierarchical data and provide an R package to simplify their investigation (R Core Team 2016).

In particular, we adapt the visualization principles of focus-plus-context and linking to the study of tree-structured data (Buja et al. 1996; Becker and Cleveland 1987). Our motivating application is to the analysis of bacterial time series, where an evolutionary tree relating bacteria is available a priori. However, we have identified common problem types where, if a tree is not directly available, it can be constructed from data and then studied using our techniques.

We have implemented our visualizations in D3, but encapsulated in an R package, called *treelapse*, to facilitate rapid turnover from data preparation and modeling to interactive exploration, and vice versa. Our code is open-source and is linked in the supplementary materials. We hope this package encourages data analysts to work at the border between data modeling and visualization, and more generally empowers a wider audience to apply less widely known, but powerful, visualization ideas.

In summary, our key contributions are

- Proposals for visualizing hierarchically structured data, based on principles from the data visualization community.
- The implementation of these proposals in a publicly available R package.
- The illustration of the wide reach of hierarchical data visualization, through case studies in both scientific and societal contexts.

The article is organized as follows. First, we describe our motivating application to the microbiome and the associated generic analysis tasks. Next, we review the underlying visualization principles behind our contributions. Then, we then connect these principles to analysis tasks we identified earlier, describing in detail the visualization methods we have implemented in *treelapse*. We close with several case studies using publicly available data across both microbiome and non-microbiome related applications.

### 1.1. Problem Motivation

A microbiome is a community of bacteria living in given environments, for example, ocean water or the human gut (Human Microbiome Project Consortium 2012; Cho and Blaser 2012). Progress in the field has been rapidly accelerated by the advent of genomic technologies, which enable detailed quantification of bacterial ecological structure and its influence in human and environmental health. Being concerned with both bacterial community structure and human health, the field exists at the border between ecology and medicine; consequently, papers in the area often apply a blend of exploratory data analysis and formal statistical inference.

The two essential microbiome analysis problems that motivated our work are the tree-structured differential abundance and differential dynamics problems. In the differential abundance problem, we attempt to compare the abundances of individual bacteria across experimental conditions—for example, treatment vs. control or healthy vs. diseased. This is the microbiome analog of differential expression analysis in genomics (Anders and Huber 2010). We prepend the description “tree-structured” because, in practice, researchers generate interpretations about intermediate taxonomic orders—it is more interesting to discover novel behavior taxonomic levels

between high-order phyla and low-level species. Hence, we frame the tree-structured differential abundance problem as the question of identifying the largest taxonomic subtree whose associated bacteria are differentially abundant.

In the tree-structured bacterial dynamics problem, the goal is to describe changes in bacterial abundances in an environment over time. As in the differential abundance problem, it is useful if these descriptions can be given at the highest subtree at which the pattern appears. Specific questions of interest often have an ecological flavor. For example, researchers are often interested in understanding how bacterial populations respond to sudden or gradual environmental changes or how species fill, drop out from, or compete for environmental niches. Medically, these questions are important for illuminating the impact of antibiotic time courses or diet changes, for example.

## 1.2. Problem Abstraction

To unify the tree-structured differential abundance and bacterial dynamics problems, we identify the data with a collection of random variables indexed by nodes in a prespecified tree structure. In the differential abundance problem, each random variable lives in  $\mathbb{R}^G$  where  $G$  is the number of groups being compared. Each coordinate represents the abundance for that group, and a node exhibits differential abundance when the coordinates are drawn from different distributions. On the other hand, in the bacterial dynamics problem, each random variable is a time series, living in  $\mathbb{R}^T$ .

In both of these applications, we constrain the values of parent nodes according to the value of the children nodes: we define the value at each node to be either the sum or average of all descendant tips. However, it is possible to imagine situations where the internal nodes are drawn from their own distribution, unconstrained by descendants. In general, analysis in this abstraction focuses on describing the distribution of these random variables as a function of their position across subsets of the tree. The essential difficulty in these problems is high-dimensionality—there are many tree nodes, each holding a vector-valued random variable. Even simply navigating across the tree and comparing coordinates in the observed variables is a challenge; ideally, we could construct a succinct representation of the essential covariation across subtrees and coordinates.

This framework suggests other potential application areas, not all of which have a priori known tree structures. For example, collections of spatially-indexed time series are frequently encountered in practice—consider energy consumption, product sales, or high school dropout rates across regional districts. This type of data has an implicit tree structure—at the top level are different states, while at the bottom are individual census tracts, say. Analysis here revolves around the question of how variation across time series is related to their geographic position.

Alternatively, if this type of hierarchical contextual information is not directly available, a tree structure can be learned from the data. This could be achieved by learning a hierarchical clustering on the original series. Further, if contextual information is available, but it is not hierarchical, it is possible to setup a supervised problem that uses context to predict features of the time

series. We can construct a tree by applying a tree-based classifier (Breiman et al. 1984) or extracting a regression tree from a more complex supervised model (Boz 2002; Saito and Nakano 2002). Analysis then focuses on how different partitions of the contextual, covariate space relate to observed time series.

Finally, note that, while we have focused on time series valued nodes, all of this discussion could be translated to studying high-dimensional data via parallel coordinates (Inselberg and Dimsdale 1991). The usual parallel coordinates challenges remain, mainly selecting scales for and ordering across the different coordinates, but the linking and focus-plus-context can still be employed this setting.

## 1.3. Background Literature and Solution Principles

Now that we have specified the essential questions of interest, we survey some ideas from the visualization literature that can be applied to answer them. As the core difficulty is high-dimensionality, it should be no surprise that the techniques we adapt come from the literature on high-dimensional data visualization, which has enjoyed rapid progress in the last 25 years. Modern research in this domain develops abstractions and taxonomies for guiding visualization designs so that they most effectively communicate properties of the data to their intended audience. A major push in this body of work explores the potential for interactivity to improve many stages of the data analysis process, from preliminary data preparation, to refinement and navigation across views, to final sharing and annotation of results (Heer et al. 2012). Further work has attempted to bridge the gap between statistical analysis and data visualization methodology, both of which provide techniques for learning from high-dimensional data (De Oliveira and Levkowitz 2003).

The problem structures most relevant to our study are tree and temporal structure, and the visualization community has various ways of reasoning about these data, see (Graham and Kennedy 2010; Aigner et al. 2011) for detailed surveys. From this literature, our approach is most directly informed by the focus-plus context and linking principles, which we briefly review here. The focus-plus-context principle is that large collections of visual elements can be studied at multiple scales, by “simultaneously focusing” on a few elements of interest and maintaining the “context” of the coarser-scale background. A simple example of this idea is to include a search box that highlights matching samples (focus) and mutes the rest (context). Two more sophisticated methods anchored in this idea are timeboxes and degree-of-interest (DOI) trees; both are central to the proposals in treelapse (Hochheiser and Shneiderman 2004; Heer and Card 2004). In timeboxes, a collection of time series are graphically queried using interactive brushes. Series that pass through all of the user-specified brushes are highlighted, and the rest are faded to the background. Hence, time series meeting the constraints imposed by the brushes are focused, while the remainder are de-emphasized, though they remain present as context. This method can be interpreted programmatically as the visual analog of a database query, or probabilistically as the conditional distribution for the full series, given it passes through certain bounds.

In DOI trees, the viewer’s attention is focused on a collection of high-interest nodes, while the remaining lower-interest

nodes are left on the fringes as context. The implementation is modularized into two tasks—the determination of a DOI distribution over nodes in the tree and visual layout of a tree given DOI assignments. The DOI distribution used by Heer and Card (2004) places maximal interest on the node that the user had most recently clicked, along with all ancestors. The DOI for all other nodes is defined as the distance to the closest maximal interest node. The layout step then trims low-interest subtrees until the remaining nodes fit within a given screen size. By adjusting the minimal DOI below which nodes are hidden, the user can transition between node-specific and full-tree scales.

In linking, alternative representations of the same samples are placed side-by-side to display covariation across views. A canonical application is to linked scatterplot brushing (Becker and Cleveland 1987). Here, a scatterplot matrix gives the relationship between all pairs of variables. Points brushed in one scatterplot are then highlighted in all others. For example, this helps the user determine whether an outlier in one dimension is an outlier in others. Another instance of this idea links the results of dimensionality reduction methods to displays of the raw data, as implemented by XGobi and Cranvas, for example (Xie et al. 2013; Swayne et al. 1998). As in timeboxes, linking can be interpreted as database queries or conditional probabilities: given a subset of the series after conditioning on the values for one set of features, what are the values for a second set (Buja et al. 1996)?

Finally, unrelated to established visualization principles, we note that our work is deliberately grounded in the R software ecosystem. This connection is made using the `htmlwidgets` package (Vaidyanathan et al. 2014). Not only does linking R with D3 make these visualization methods more broadly accessible, we hope to facilitate exchange between data modeling and interactive visualization. Moreover, our tools are intentionally limited in scope—designed to facilitate this dialog for a specific class of problems, rather than providing a toolbox for generic types of visualization design. We believe that this narrow context within a broad ecosystem strikes a balance between problem-specificity and ease-of-use.

#### 1.4. Specific Proposals

Our first proposed visualization technique is a minor modification of the DOI tree. The standard DOI tree definition does not have any notion of data defined at nodes, it is only used a device for navigating tree structures. A trivial extension can encode scalar data at nodes: have the node radius reflect the associated scalar value. To reinforce this effect, we can adjust the width of the parent edge. When parent nodes have values equal to the sum of their children, this creates the effect of values “flowing” from the root to leaves. To help viewers make use of their domain knowledge, we have included a search box that highlights paths to nodes with matching terms. Edges are ordered from widest on the left to narrowest on the right. While this method can only represent a single scalar-value per tree node, it suggests an approach to the tree-structured differential abundance problem, which we call the DOI sankey.

In the DOI sankey, we split each edge in the DOI Tree across different groups. For example, suppose we have the average counts for treatment and control groups at each tree node. Every

edge in the tree is split into two colors<sup>1</sup>, with relative widths of the different colors reflecting differences in sizes for the two groups.

This display is designed to facilitate investigation of the tree structured differential abundance question. For example, for a single node and a single group, first compute the average abundance at that node among all samples in that group. This will give the width for that group’s color on the edge leading to the specified node. Differentially abundant subtrees then correspond to subtrees where some colors occupy more space than others. That is, this representation makes it easier to identify points where the “flows” for different groups diverge—the colors begin to separate. The DOI principle assists navigation across the tree structure, allowing focus on individual flow structures without losing broader tree context.

Our third display is directed at the bacterial dynamics question. Here, two panels are arranged one over the other; one displays all time series, while the other displays all tree nodes, with node sizes reflecting the value at that node averaged across all time points. For this reason, we call the display, timebox trees. In the time series panel, we have directly implemented the timeboxes idea. We then link the panels: when a set of series is highlighted by the timeboxes, the associated tree nodes are also highlighted. For example, timeboxes can be used to focus on a set of series with specific shape—increased abundance after an ecological shock, for example—and identify along what subtrees this pattern is present. To further focus on specific elements, a pan-zoom scented widget is provided (Willett et al. 2007). The widget is a miniature version of the full time series panel, equipped with a single brush whose extent specifies the limits in the main time series panel. As in the DOI trees and sankeys, a search bar can be used to highlight those series of interest a priori.

The final display currently implemented in the package is the natural converse of the timebox trees display. Rather than defining visual queries in terms of time series, it defines queries using nodes in the tree. For this reason, we call the display treeboxes. Rather than focusing on the intersection of brushes, as in timebox trees, we focus on the union of brushed over nodes. This allows us to highlight series associated with nodes on distant subtrees. This display is also suited for the bacterial dynamics problem. For example, by highlighting all nodes at one taxonomic level in the tree, we can easily summarize the time series pattern for all the taxa at that level. Alternatively, focusing on all the children below a single node makes it possible to see how much correlation and competition there is between taxonomically similar bacteria. As in the timebox trees display, a search box and pan-zoom scented widget are provided.

In principle, it would be desirable to combine the timebox and treebox displays, so that highlighted nodes and series could be determined through brushes on both the tree and series. For example, it would be useful to highlight the samples that lie in the intersection of all timeboxes and union of all treeboxes. This could allow more complex queries than are currently available. However, while conceptually appealing, the authors encountered obstacles in practical implementation: brush and mouseover events are required to occupy the same space in

<sup>1</sup> We use the colorbrewer palette to facilitate readability (Brewer et al. 2003).



**Table 1.** Problem dimensions for each of the case studies. For problems with dimensions larger than that in the housing prices problem, we recommend an initial summarization or filtering step to prevent performance issues.

Data	Number of timepoints	Number of nodes
Antibiotics	56	386
Preterm births	216	318
Housing prices	254	944
Bikesharing	24	819
Global patterns	500	51

this combined view. Properly distinguishing these events can be challenging, and a solution based on the introduction of a lag between mousedown and the manipulation of a brush led to a deteriorated user experience<sup>2</sup>.

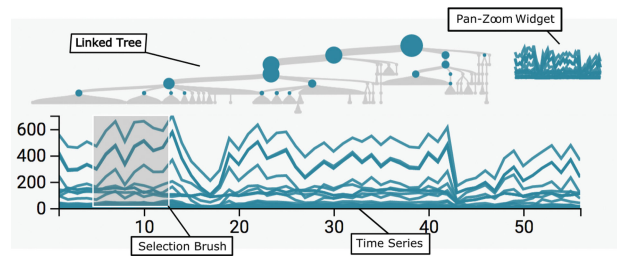
To be practically useful, the resulting visualizations must respond fluidly to user interaction. As the data increase in scale, this fluidity can deteriorate for two reasons. First, rendering many SVG elements in a framework like D3 is costly (Johnson and Jankun-Kelly 2008). While it is possible to use alternatives—HTML5’s Canvas, for example—it is often more challenging to implement complex interactive behavior through them. Second, some of the dynamic queries require a search over a many elements. These limitations are most pronounced in the timebox tree display, which must search through all timepoints among all time series whenever the brush is moved. The first, but not the second, concern applies to treeboxes, while neither applies to DOI trees. Nonetheless, we feel comfortable recommending timebox trees for data on the order of 500 tree tips and 50 timepoints. We note the sizes of the datasets in each case study in Table 1, which each render fluidly, with the possible exception of the California housing data, where there is a noticeable lag in the timebox tree and treebox displays. In problems of larger size, we recommend a preliminary filtering or aggregation step across nodes or, if the time series is smooth, across neighboring timepoints, to avoid these potential scaling difficulties.

## 2. Case Studies

We now delve into applications on real data. Our goals are to illustrate potential workflows that incorporate treelapse, describe the formulation of questions that can be naturally investigated with our methods, and provide example interpretations on treelapse output. Our examples are also chosen to reflect the range of problem domains to which the package can be applied—though it was motivated by applications to the microbiome, it is not tied to it. More importantly, we argue that the visualization principles reviewed above can substantively improve the practice of data analysis in the class of problems to which we have limited ourselves.

### 2.1. Bacterial Dynamics of Antibiotics Time Courses

Dethlefsen et al. (2008) investigated the effect of antibiotics on bacterial community composition from an ecological perspective. The study tracks the microbiome of three patients across ten months, with two 5-day antibiotic time courses separated by



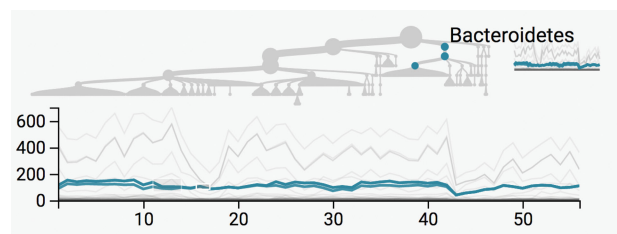
**Figure 1.** Here we display the primary timebox tree view of the antibiotics dataset from (Dethlefsen et al. 2008), annotated with the main components of the visualization. The tree at the top is a taxonomic tree of all the bacteria contained within the sample, and it is visually linked to the time series at the bottom: each node in the tree corresponds to a path among the time series. The selection brush is used to focus attention on the time series that go through it—these are highlighted in blue—and other brushes can be added using a button not displayed here. The pan-zoom widget at the top right is used to update the scales of the main time series display so that only particular time windows and y-axis ranges are visible. This view is the basis for all the timebox tree and treebox displays that appear below.

6 months. Discerning the variation in resilience across bacteria is important, considering the role of bacteria in health and not just disease.

We approach the data using the linked time and treebox views, after first filtering low variance taxa and taking an asinh transformation. An initial view, Figure 1, reveals two dramatic drops in the overall bacterial abundance time series during the antibiotics time courses. Two more subtle effects are also suggested from this view,

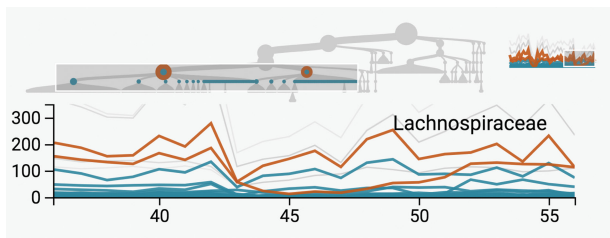
- The second antibiotic treatment seems to have a more lasting effect, as the series take longer to return to their original values.
- Some high-level taxa appear relatively unaffected by the first antibiotic treatment. By more closely inspecting the display, we are able to identify these as members of the *Bacteroidetes* phylum, see Figure 2.

Next, using the scented widget, we focus on the window around the second antibiotic treatment. We apply the treebox display to compare then behavior of different families of *Firmicutes*, *Lachnospiraceae*, and *Ruminococcus*. We suspect that these taxa are associated with the delayed recovery after the second time course. To investigate this, we input these family names in the search box to isolate their positions on the tree; then we apply brushes to highlight the series that contribute to these higher-level families. The resulting view is given by Figure 3.



**Figure 2.** Introducing a second box into the timebox display identifies the *Bacteroidetes* as a taxon only mildly impacted by antibiotics. The layout is identical to Figure 1, except two small brushes are placed over the time series between 10 and 20 days, and now only those time series and corresponding nodes in the tree are highlighted in blue. Further, the user has hovered over the top blue node in the tree, revealing the taxonomic identity of these series. Hence, brushing the time series and linking with the tree can be used to discover and characterize notable variation within the data.

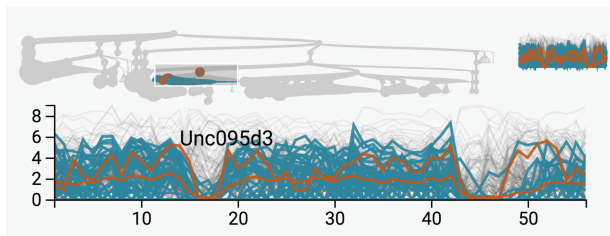
<sup>2</sup> However, the code for this approach is available publicly, in a separate branch of treelapse: <https://github.com/krisrs1128/treelapse/tree/combined-brushes>.



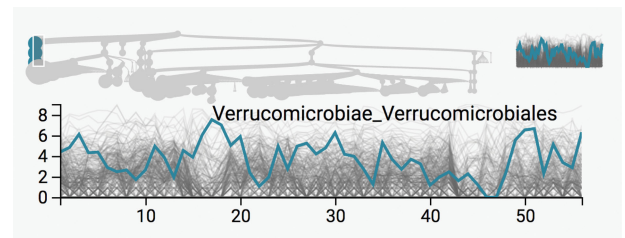
**Figure 3.** Zooming into the second antibiotic time course and highlighting the *Lachnospiraceae* and *Ruminococcus*, we see that the *Ruminococcus* took more time to recover to pre-treatment levels. Here, the red lines and nodes are those that match the text search provided by the user in a search box just below the figure (not displayed here). Hovering the mouse over these lines provides their identities—the top red line is *Lachnospiraceae*, and the bottom red line is *Ruminococcus*. Note that the brush in the treebox display is located over the tree, rather than over the time series. In particular, the search box and interactive brushing can be combined to interrogate hypotheses of a priori interest.

Alternatively, we can summarize each node by the average across its descendants—this brings attention to individual bacteria that may be underlying some of the broader taxonomic patterns we have noted when studying the subtree sums. For example, in [Figure 4](#), we highlight all families below order *Ruminococcus*, suggesting that the decrease due to antibiotics occurs uniformly across almost all families. A point that was not evident in the earlier sum-across-descendants view is that, after the second treatment of antibiotics, a few of the *Ruminococcus* families recover more rapidly than the rest, for example, the *Unc095d3* (highlighted in red) are only briefly affected. In contrast, most families seem to recover in unison after the first treatment.

Further, note that in this subtree averages view, the tree display has changed. This is because, at each branching point, we place the node with larger average value on the left. [Figure 5](#) notes that the nodes at the far left in the tree are associated with phylum *Verrucomicrobiae*, corresponding to a large average abundance across time points. This phylum had been previously obfuscated—because there are not many leaves associated with this phylum, the sum was small. Interestingly, the abundance of these bacteria seems to *increase* after the first antibiotics treatment. Be cautious, however, that the average over only a few *Verrucomicrobiae* species will be a more variable estimate than the averages over the more prevalent phyla.



**Figure 4.** By hovering over the *Ruminococcus* branches, we see that there is a prolonged effect of the antibiotics time courses more or less uniformly across the lower taxonomic orders. The graphical elements are the same as before, except the user has searched for *Ruminococcus* and species *Unc095d3*, which has the highest average abundance within this taxon. By displaying averages rather than sums, we see that the effect of antibiotics visible at higher taxonomic orders is not created by a single dominant species becoming less abundant, but rather the decline in populations across all descendant species. The same display applied to different data can yield different insights.



**Figure 5.** The subtree averages aggregation brings attention to the *Verrucomicrobiae*, which though only present as a few species, are each rather abundant. In particular, they seem to increase after the first antibiotic time course, which occurs between days 15 and 20. This view was generated by placing a brush over the branch on the far left, which has those nodes with the largest averages across all time-points. The user's mouse is over the blue series, which brings up the associated taxonomic label. The determination of species whose abundances increase during antibiotics, which would require many hypothesis tests using a more standard approach, becomes quickly apparent via interactive visualization.

## 2.2. Differential Bacterial Abundance and Preterm Births

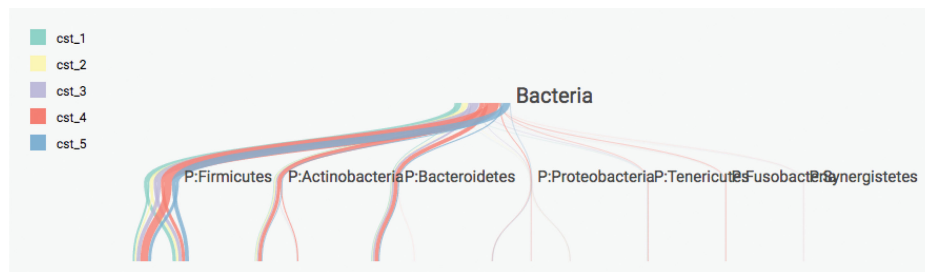
DiGiulio et al. (2015) tracked the abundance of bacteria in the vaginal microbiome during pregnancy in an effort to study relationships between bacterial community composition and preterm birth. Ideally, it would be possible to develop clear bacterial signatures associated with preterm births.

Unlike the antibiotics study, we have measurements across more individuals than we could reasonably inspect one at a time. While we could average across all individuals, we will take our cue from DiGiulio et al. (2015) and place each sample into one of five Community State Types (CSTs), identified via k-medoids. In that study, a linear model identified one of these CSTs (CST 4) as significantly more diverse, further it appeared associated with preterm births. Here, we corroborate this finding using exploratory views.

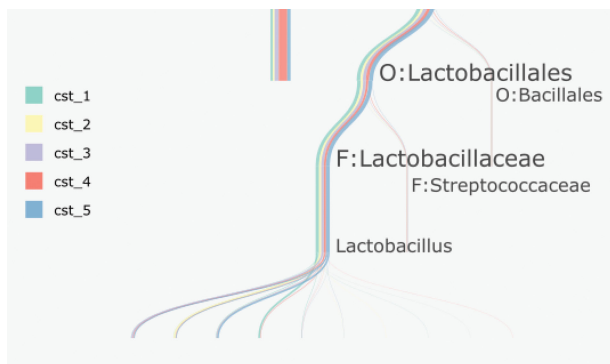
Therefore, our focus here is on the differential abundance question, rather than dynamics. We would like to provide visual representations of differential abundance across CSTs and also between preterm and non-preterm births. DiGiulio et al. (2015) interpreted the CSTs using a heatmap, with bacteria ordered according to a hierarchical clustering. By using the DOI sankey instead, we can interpret the CSTs in their taxonomic context and at multiple scales of taxonomic resolution. Further, while DiGiulio et al. (2015) focused on identifying associations between preterm births and CSTs—presumably because testing individual bacteria loses power—we can compare bacterial abundances between preterm and non-preterm samples along subtrees, without requiring CSTs as an intermediary.

In [Figure 6](#), we compare the *f* CSTs according to their values along the subtree. Specifically, we took the average of all samples within each CST to define values at the (species-level) leaf nodes, and then aggregated the averages up to the root. It is immediately clear that samples from CST 4 have much more taxonomic diversity. Further, focusing on the *Lactobacillaceae* family, we note that the differential abundance of these bacteria distinguishes the remaining CSTs, see [Figure 7](#).

Alternatively, in [Figure 8](#), we avoid working with CSTs, displaying instead averages among samples associated with either preterm or term births. The green and yellow edges are associated with preterm births—we see that they contribute more weight to phyla outside the Firmicutes. This is consistent with



**Figure 6.** The increased diversity among samples in community state type (CST) 4 is represented by the relatively larger contribution of red edges to branches outside of the Firmicutes. This display shows the top of the DOI sankey visualization of the preterm birth data studied in (DiGiulio et al. 2015). The root of the tree is the taxonomic kingdom Bacteria, its children are labeled according to their phylum names. Each color within a branch is associated with CST, and the width of the associated color corresponds to the average abundance of that taxonomic group among all samples belonging to the corresponding CST, as indicated by the legend at the left. Phyla are sorted from most to least abundant. This initial display of the DOI sankey provides a summary of overall abundances across taxonomic groups and CSTs, and suggests subtrees to navigate into, to extract more detailed abundance information.

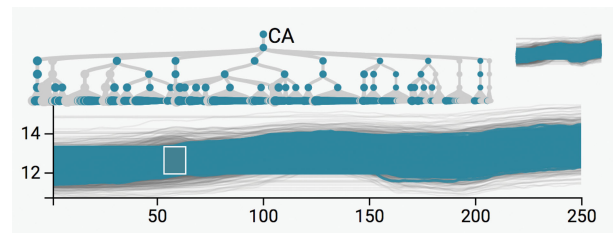


**Figure 7.** Zooming into the *Lactobacillaceae* family, we notice that the difference between the remaining four CSTs is related to which types of *Lactobacillus* are most prominent. The DOI sankey refocuses the tree around the last group that was clicked on, in this case showing more detail about order *Lactobacillales* and its descendants. The overview display can be recovered by navigating back up to higher-level taxa. Hence, it is possible to navigate between broad overview and detailed displays in a way that facilitates interpretation of results from statistical analysis.

the claim that CST 4, the most diverse of the CSTs, is associated with preterm births.

### 2.3. Dynamics in Housing Prices

We next consider an application unrelated to the microbiome, but with relatively clear hierarchical structure. Our data are downloaded from Zillow and give the Zillow Home Value Indexes at the neighborhood level, across the country, computed monthly between 1996 and 2016. A link to the data source is provided in the supplementary materials. In our display, we have taken the natural log of these indexes. As our hierarchical structure, we use each neighborhood's assignment to state, regional,

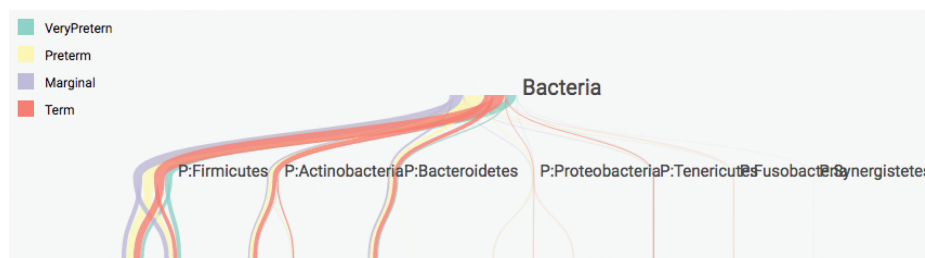


**Figure 9.** The time series here represent California neighborhood home prices between 1996 and 2016, and the tree corresponds to a geographic hierarchy, with regions at the top and neighborhoods at the bottom. We have brushed the neighborhoods with mid-range home prices before the recession. The associated tree nodes are highlighted at the top. Note that the collection of series seems to widen after 2008—we are interested in whether there are reliable predictors of these alternate trajectories, given their similar starting points. This serves as a baseline with which to compare Figure 10—these views are easy to transition between interactively.

county, and city levels. We represent each of these coarser spatial categories using the average of all neighborhoods contained in them. We have filtered down to the 890 neighborhoods in California; rendering more neighborhoods while keeping all 246 timepoints causes a severe lag in the interface.

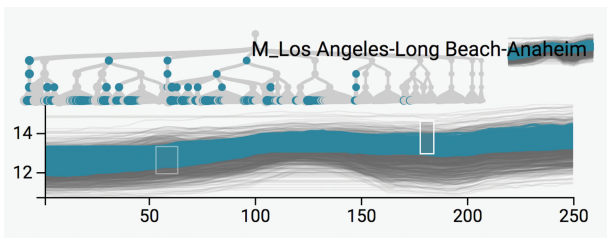
Our basic analysis revolves around geographic and temporal variation in home prices. We are especially interested in the effect of the 2008 recession and any variation in the lead-up to or recovery from this event. These questions can be naturally framed using timebox trees and treeboxes.

For example, we can study the trajectories of home prices among neighborhoods, conditional on their being middle-income before the recession. We generate the sequence of views in Figures 9–11 to this end. The first of these figures isolates neighborhoods with middle incomes before the recession,



**Figure 8.** Samples with high levels of phyla other than Firmicutes appear to be related to preterm births. Here, we again display the preterm birth data with a DOI sankey, but instead of grouping samples according to statistically generated CSTs, we directly assign samples to preterm vs. not preterm according to whether the mother eventually had a preterm birth. These new states are visible in the updated legend. Through interactivity, it becomes possible to develop meaningful visual summaries even before calculating formal statistical ones.



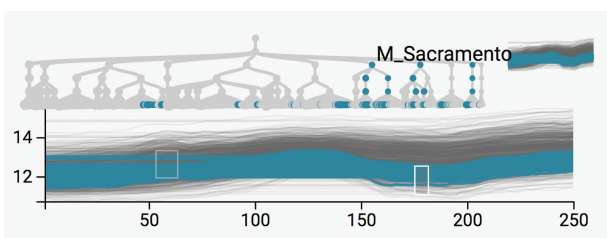


**Figure 10.** Among those neighborhoods with mid-range prices before the recession, displayed in Figure 9, we have selected those that recovered more rapidly by introducing a brush at the top right of the time series panel. By hovering a brush over a collection of tree nodes that all seem to be highlighted, we infer that the associated neighborhoods are located mainly in Los Angeles and San Diego counties. This follow-up view is interesting to contrast with Figure 11.

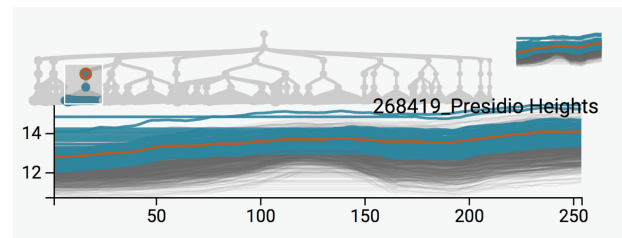
using a single timebox. Since there appears to be a divergence in trajectories after the recession, we introduce a second post-recession timebox, dragging it over series with higher and lower incomes during this second time period. This is the content of Figures 10 and 11. Though not directly visible from the static figures, hovering the mouse over the highlighted tree nodes provides the geographic identities, and we find that most of the middle-income series that increased after the recession are associated with middle-income neighborhoods within the coastal Southern California counties. For example, the mouse is currently over a subtree with many highlighted nodes, which is shown to be the Los Angeles–Long Beach–Anaheim metropolitan area. In contrast, hovering over nodes associated with those middle income neighborhoods that saw decreases indicates that they were mostly located in Central California and Oakland. In Figure 11, the mouse is positioned over the Sacramento (which is located in Central California) subtree, and seems enriched for this subset of strongly recession-affected series.

The previous analysis highlights the fact that, within even narrow geographic regions, there can be substantial variation in prices. We can study this directly using treeboxes. In Figure 12, we have highlighted all series in San Francisco County. We see that, in 2016, prices range from around  $e^{13} \approx \$440,000$  to  $e^{14.5} \approx 2$  million. So, while all these neighborhoods tend to be among the more expensive ones in California, prices can vary in a non-smooth way across geographic space.

We conclude this example with a caveat that the Zillow data are not representative of all neighborhoods in



**Figure 11.** In contrast to Figure 10, we can follow-up the selection in Figure 9 by isolating those neighborhoods where prices remained depressed after the recession. This is accomplished by moving the brush on the right down toward lower prices. Hovering over the associated highlighted nodes in the tree to reveal the associated locations, we see that most of these series correspond to neighborhoods in Central California and the East San Francisco Bay Area. The ability to sketch the overall shapes of series using brushes and interpret the associated selections using a tree simplifies what might otherwise be complex comparisons.



**Figure 12.** In contrast to Figures 9–11, we can study the variation in series associated with a subset of the tree by using treeboxes. To study the range in home prices within San Francisco County, we can brush over the associated nodes in the tree. The red circle indicates that the user has searched for “San Francisco,” which guides the user to the appropriate subtree. The mouse is hovering over one of the more expensive neighborhoods. Hence, though the timebox tree and treebox views are similar, they are directed toward different types of visual comparisons.

California, only those with enough listings on the site, so should be supplemented by other data sources for any substantial decision-making.

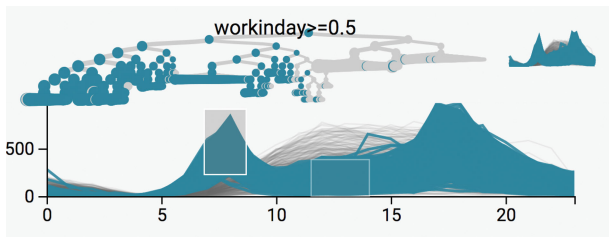
#### 2.4. Sources of Variation in Bikesharing Demand

Our next example is a study in bikesharing demand, included as an example of analyzing collections of time series when there is no obvious hierarchical structure a priori. The data are available at the UCI Repository and were originally collected by a Washington, DC-based bikesharing system for use in a Kaggle prediction competition. A link is provided in the supplementary materials. The data are hourly measurements of bike demand, aggregated across all bikesharing stations, over two years, along with supplemental weather data. In the competition, participants were asked to predict the hourly demand on a held-out test set. Here, we adopt a descriptive view instead, attempting to characterize factors associated with variation in bikesharing demand.

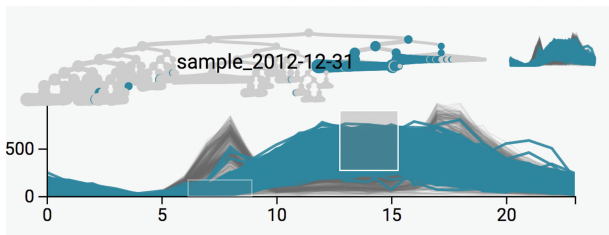
Like the Zillow home prices application, we study this problem as one of describing a large collection of related time series. Here, we consider the demand during a single day to be one time series; this is a natural choice considering the daily periodicity of bike demand. To arrange these daily series along an interpretable tree structure, we apply a regression tree relating the supplemental data to the bikesharing demand (Breiman et al. 1984). In more detail, we built this tree by noting the “two table” structure of this problem: one describes bike demand, the other holds the supplemental data. In both, the rows index days, while the columns index either hours or supplemental features. Our tree is the trained regression tree after predicting demand at 8AM based on the supplemental data. We choose this response because (1) we need a univariate response to apply regression trees and (2) the more straightforward reduction to daily-average-demand fails to distinguish between weekdays and weekends, whose series appear qualitatively very different from each other.

Given this response, the first split in the regression tree is (unsurprisingly) the difference between weekends and weekdays. This is emphasized in Figures 13 and 14, respectively; using timeboxes to isolate the two types of series highlight the left and right sides of the tree, respectively. For a more subtle effect, we select the internal nodes associated with the first split below the weekday vs. weekend split; these are given in Figures 15 and 16.

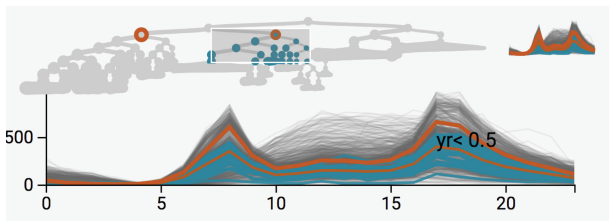




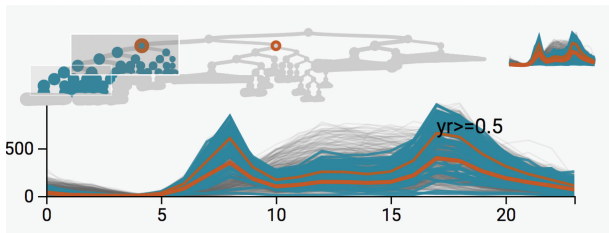
**Figure 13.** The two peaks at rush hour distinguish weekday series from the rest through the timebox tree view. The display is the same type of timebox tree view introduced in Figure 1, but applied to the bikesharing data, where the time axis represents the time of day and the y-axis gives bikesharing demand. Each series is the bikesharing demand for a single day, over the course of two years. The tree now corresponds to the regression tree generated by predicting demand at 8 am using supplementary data. Two brushes are introduced to highlight the double peaks corresponding to rush hours on weekdays. We see that although hierarchical structure was not present immediately in the bikesharing data, it is useful to introduce and interpret such structure by combining regression and visualization methodology.



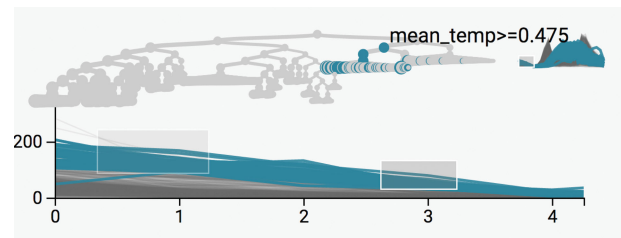
**Figure 14.** By adjusting the two brushes in Figure 13, we see that unlike weekday demand, weekend demand is unimodal. The few weekday series with unimodal series seem to be associated with holidays. This is the case for New Years' Eve, which is currently hovered over in the tree. The ease of transitioning from Figure 13 to this display indicates the importance of brushing in interaction.



**Figure 15.** Weekday demand appears larger in 2012 than 2011—compare with Figure 16. Here, brushes are introduced over the tree to see the series associated with a particular split point. The red nodes are the results of searches over the two nodes that are children of this split point. The fact that the  $yr >= 0.5$  selected line is larger than the  $yr < 0.5$  line means that demand was larger in 2012. In combination with searching and treeboxes, it is possible to interpret more subtle split points in the decision tree.



**Figure 16.** Weekday demand increased in 2012—compare with Figure 15. The search terms are the same as in that figure, but the subtree associated with the 2012 split point is highlighted, using the union of two boxes. Note that unlike timebox trees, which highlighted series lying through the intersection of brushes, treeboxes highlight series within the union of brushes.



**Figure 17.** The samples with highest night demand tend to fall on warm weekends. Here, the pan-zoom widget has been used to adjust both time and demand axes to narrow on a specific window of interest. Brushes are drawn over the larger among these series, and the corresponding tree nodes are located close to one another, in the part of the tree corresponding to the warm/cold average temperature split. More generally, panning and zooming allows navigation between focus and context.

This suggests that weekday demand increased during the second year.

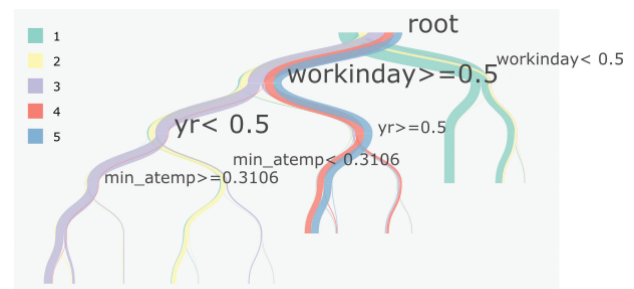
In contrast to these general questions about daily demand, we could ask a more granular question about specific time windows. For example, what characterizes days on which there is larger than average demand after midnight? We can select these series after first zooming into this time window. Figure 17 reveals that the highlighted series are associated with the warm-weekend split, which seems quite reasonable in retrospect.

Finally, we can study the behavior of the regression tree itself using the DOI sankey (Figure 18). Here, we group samples according to their quintile of 8 a.m. demand and then count the abundance of the groups flowing down different branches. We find that the quintiles are each rather strongly separated after descending even a few steps down the regression tree—for example, Figures 15 and 16 focus on 2011 vs. 2012 split among weekday samples, showing that this split distinguishes between samples falling in the second and third quintiles of 8 a.m. demand.

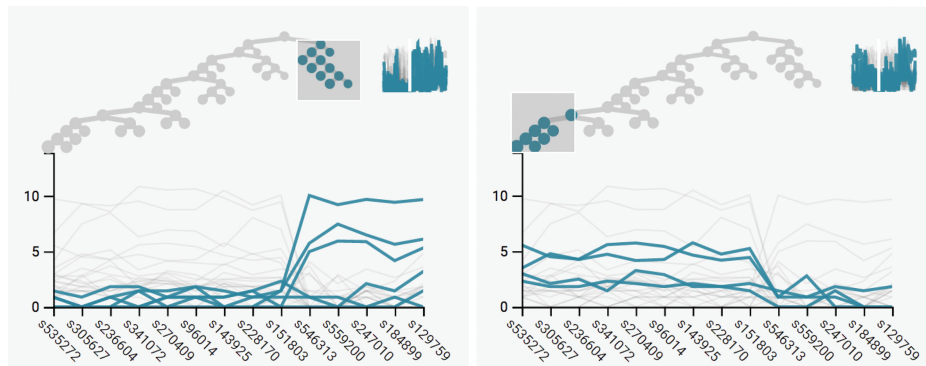
This interactive representation of regression trees is potentially more useful on larger trees that cannot be easily parsed in a single view; in this sense the bikesharing tree is relatively simple. In our ideal data analysis workflow, we imagine the analyst applying interactive visualization and modeling techniques in an iterative, nonlinear fashion, in the spirit of De Oliveira and Levkowitz (2003).

### 2.5. Hierarchically Clustering the Global Patterns Data

Each of the timebox tree and treebox examples presented so far have focused on data with a clear time component. We note



**Figure 18.** So far, we have focused on the timebox tree and treebox representations of the bikesharing data—a complementary view is provided by the DOI sankey. Here, the tree is the result of the regression tree procedure, while the colors represent particular quantiles of 8 a.m. demand. This allows the determination of which split descendants are associated with low or high demand.



**Figure 19.** An application to the Global Patterns demonstrates how linking in treelapse can be applied to combine hierarchical clustering and parallel coordinates views. Each panel represents a different subtree cluster within this dataset, as indicated by the different locations for the tree brushes. The paths in the lower halves of each display represent the average value across different bacteria rather than timepoints, as in all previous figures. Though the samples originally do not include any hierarchical structure, hierarchical clustering provides such a structure which can then be interpreted using treboxes.

however that these techniques could alternatively be applied to high-dimensional data, via the use of parallel coordinates (Inselberg and Dimsdale 1991). The usual parallel coordinates challenges remain, namely selecting scales for and an ordering across the different coordinates, but the linking and focus-plus-context ideas can still be employed in this setting. Here, we provide an implementation of this idea on a dataset comparing microbiomes across various ecological environments (Caporaso et al. 2011), which is publicly accessible through the phyloseq R package (McMurdie and Holmes 2013).

The original Global Patterns data consists of 26 samples across 9 environments (e.g., freshwater and soil). In each site, there are counts across 19,216 taxa—to simplify visualization, we filter to the 500 most abundant taxa.

We hierarchically cluster these 26 samples based on the 500 most abundant taxa, using complete linkage on the UniFrac distance. Figure 19 displays the resulting hierarchy together with a parallel coordinates view of the asinh transformed taxa.

In Figure 19, we compare two subclusters from the hierarchical clustering tree, after zooming to a few of the bacteria that distinguish between the clusters. In contrast to the figures displayed to this point, we only print time series associated with observed samples — the leaves of the hierarchical clustering tree. This reduces visual artifacts that can be created by plotting many similar internal nodes, and which can overwhelm patterns occurring in the leaves, which are those of central interest. Upon revisiting the original data, it becomes clear that the samples highlighted on the left come from freshwater samples, while those on the right come from soil and skin, and looking up taxonomic groups associated with the distinguishing bacteria confirms this. For example, many of the species with high abundances in the left figure come from order *Oceanospirillales*.

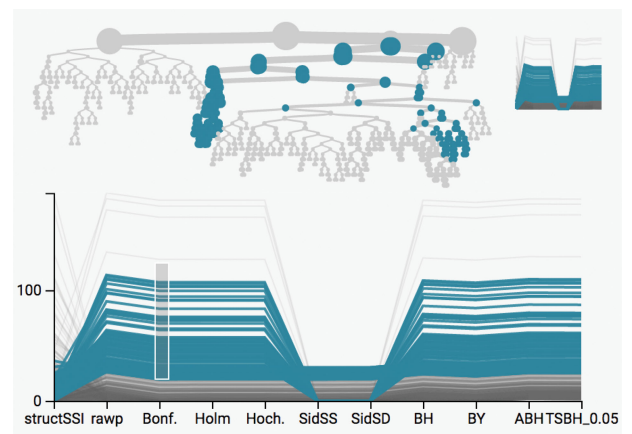
### 2.6. Inspecting Confirmatory Analysis

In addition to facilitating exploratory study, treelapse has potential value as a device for inspecting confirmatory analysis. We provide an illustration extending an example from (Callahan et al. 2016), which formally tested bacterial species for association with age in a sample of mice. The testing approach advocated there is particularly well-suited to visualization with treelapse, since it sought to detect associations at multiple levels

of phylogenetic resolution, using statistical tools developed by Yekutieli (2008), and Sankaran and Holmes (2014).

The data of interest in Callahan et al. (2016) are bacterial counts collected across old and young mice. After variance-stabilizing these counts using DESeq2 (Love et al. 2014), a *t*-test was applied to each node in a phylogenetic tree, comparing abundances between old and young mice. To account for multiple testing, we employ the structSSI algorithm (Yekutieli 2008; Sankaran and Holmes 2014) along with methods available in the multtest package (Pollard et al. 2005).

To interpret the results, we apply timebox trees. Our goals are to (1) identify subtrees with consistently elevated differential abundance across age groups and (2) compare alternative multiple testing adjustment procedures. Our approach is to display the negative-log raw and adjusted *p*-values for each node, with alternative methods compared via parallel coordinates. One view of the resulting display is captured in Figure 20. First, we see that significant nodes tend to be significant across all methods—the ordering between different series appears stable. Interestingly, the Sidak one-step and structSSI procedures



**Figure 20.** Viewing a tree of *p*-values across different methods highlights two subtrees with strong associations with mouse age, across several testing procedures. The tree represents the taxonomy of bacteria, and the series provides the negative log *p*-values associated with nodes as computed by different tests, listed along the x-axis as in parallel coordinates. By selecting series with larger values for a test, we see the associated subtrees of significant *p*-values. Hence, hierarchical views can be useful even in the confirmatory testing settings which typically study results from individual tests in isolation from each other.

seem to have lower power than the others, including conservative FWER-controlling methods, like the Bonferroni procedure. Further, in this application, FDR-controlling techniques do not seem to offer notably different adjusted  $p$ -values, relative to those controlling FWER. This suggests that, for this problem, bacteria are either strongly associated with age, or not associated at all, so that there is little gain from using more sensitive procedures.

Further, selecting series with strong association between abundance and age, two major subtrees are brought to the forefront. Separately querying the taxonomic identities of these bacteria reveals that they are two subgroups of *Clostridia*, which is consistent with the analysis of Callahan et al. (2016). More than this specific analysis outcome, this view demonstrates that interactive visual inspection of results from confirmatory analysis provides deeper insight than the standard practice of printing tables of (adjusted or unadjusted)  $p$ -values: the relationship between significant nodes is only clear upon visualization on the tree.

### 3. Conclusion and Future Work

Here, we have reviewed some fundamental principles of data visualization and described their implementation in a new treelapse package. Further, we have provided examples of the practical usefulness of these principles in real-world data analysis situations.

This package has only developed basic ideas, and there are a number of potentially useful extensions worth exploring. For example, we have not considered the principle of arrangement in our visualizations (Buja et al. 1996), though many of our conclusions were based on comparing alternative selections of the same display. We could imagine faceting our displays across groups to make these types of comparisons more accessible. Further, we have only worked with the DOI distribution described by Heer and Card (2004). It would be interesting to define a more statistical notion of interest along nodes, based on cognostics, for example Hafen et al. (2013), and Friedman and Stuetzle (2002). A simple extension could be to allow graph layouts instead of trees in time and treebox displays, for data that cannot be coerced into a hierarchical structure. Further, if these ideas turn out to be useful in practice, it would be valuable to modularize the basic visualization layouts and relationships into a library, allowing the wider community to construct novel linked, interactive graphics with minimal effort. Finally, formal quantitative assessments of interface design through a user study could guide changes that improve the experience of practitioners.

In summary, we have built an easily accessible R package for visualization techniques in a very specific methodology problem—analysis of differential abundance and dynamics in hierarchically structured data—that appears in a variety of application domains. We have leveraged a link between R and D3 (Vaidyanathan et al. 2014) to create visualizations during the exploratory phase of data analysis; in this way our work is a departure from the culture of polished, journalistic visualizations prioritized by the D3 community and is more closely aligned with the vision in De Oliveira and Levkowitz (2003) of more tightly integrating data visualization and statistical

analysis techniques. Finally, we have given a series of examples to demonstrate how the general visualization techniques of focus-plus-context and linked brushing can be practically incorporated into a range of practical analysis workflows, from studying the impact of bacteria on human health to better allocating units in commuter bikesharing systems.

### Funding

This research was supported in part by an NIH training grant (5T32GM096982-03) and a Weiland Fellowship. It was also supported in part by NIH grant R01 AI112401.

### References

- Aigner, W., Miksch, S., Schumann, H., and Tominski, C. (2011), *Visualization of Time-Oriented Data*, New York: Springer Science & Business Media. [554]
- Anders, S., and Huber, W. (2010), “Differential Expression Analysis for Sequence Count Data,” *Genome Biology*, 11, R106. [553]
- Becker, R. A., and Cleveland, W. S. (1987), “Brushing Scatterplots,” *Technometrics*, 29, 127–142. [553,555]
- Boz, O. (2002), “Extracting Decision Trees from Trained Neural Networks,” in *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, pp. 456–461. [554]
- Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984), *Classification and Regression Trees*, Boca Raton, FL: CRC Press. [554,559]
- Brewer, C. A., Hatchard, G. W., and Harrower, M. A. (2003), “ColorBrewer in Print: A Catalog of Color Schemes for Maps,” *Cartography and Geographic Information Science*, 30, 5–32. [555]
- Buja, A., Cook, D., and Swayne, D. F. (1996), “Interactive High-Dimensional Data Visualization,” *Journal of Computational and Graphical Statistics*, 5, 78–99. [553,555,562]
- Callahan, B. J., Sankaran, K., Fukuyama, J. A., McMurdie, P. J., and Holmes, S. P. (2016), “Bioconductor Workflow for Microbiome Data Analysis: From Raw Reads to Community Analyses,” *F1000Research*, 5, 1492. [561,562]
- Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Lozupone, C. A., Turnbaugh, P. J., Fierer, N., and Knight, R. (2011), “Global Patterns of 16S rRNA Diversity at a Depth of Millions of Sequences Per Sample,” *Proceedings of the National Academy of Sciences*, 108, 4516–4522. [561]
- Cho, I., and Blaser, M. J. (2012), “The Human Microbiome: At the Interface of Health and Disease,” *Nature Reviews Genetics*, 13, 260–270. [553]
- De Oliveira, M. F., and Levkowitz, H. (2003), “From Visual Data Exploration to Visual Data Mining: A Survey,” *IEEE Transactions on Visualization and Computer Graphics*, 9, 378–394. [554,560,562]
- Dethlefsen, L., Huse, S., Sogin, M. L., and Relman, D. A. (2008), “The Pervasive Effects of an Antibiotic on the Human Gut Microbiota, as Revealed by Deep 16S rRNA Sequencing,” *PLOS Biology*, 6, e280. [556]
- DiGiulio, D. B., Callahan, B. J., McMurdie, P. J., Costello, E. K., Lyell, D. J., Robaczewska, A., Sun, C. L., Goltsman, D. S., Wong, R. J., Shaw, G., Stevenson, D., Holmes, S. P., and Relman, D. A. (2015), “Temporal and Spatial Variation of the Human Microbiota During Pregnancy,” *Proceedings of the National Academy of Sciences*, 112, 11060–11065. [557]
- Friedman, J. H., and Stuetzle, W. (2002), “John W. Tukey’s Work on Interactive Graphics,” *Annals of Statistics*, 30, 1629–1639. [562]
- Graham, M., and Kennedy, J. (2010), “A Survey of Multiple Tree Visualisation,” *Information Visualization*, 9, 235–252. [554]
- Hafen, R., Gosink, L., McDermott, J., Rodland, K., Dam, K.-V., and Cleveland, W. S. (2013), “Trelliscope: A system for Detailed Visualization in the deep Analysis of Large Complex Data,” Tech. Rep., DTIC Document. [562]
- Heer, J., and Card, S. K. (2004), “DOITrees Revisited: Scalable, Space-Constrained Visualization of Hierarchical Data,” in *Proceedings of the Working Conference on Advanced Visual Interfaces*, ACM, pp. 421–424. [554,562]

- Heer, J., Shneiderman, B., and Park, C. (2012), “A Taxonomy of Tools that Support the Fluent and Flexible Use of Visualizations,” *ACM Queue*, 10, 1–26. [554]
- Hochheiser, H., and Shneiderman, B. (2004), “Dynamic Query Tools for Time Series Data Sets: Timebox Widgets for Interactive Exploration,” *Information Visualization*, 3, 1–18. [554]
- Human Microbiome Project Consortium, (2012), “Structure, Function and Diversity of the Healthy Human Microbiome,” *Nature*, 486, 207–214. [553]
- Inselberg, A., and Dimsdale, B. (1991), “Parallel Coordinates,” in *Human-Machine Interactive Systems*, ed. A. Klinger, New York: Springer, pp. 199–233. [554,561]
- Johnson, D. W., and Jankun-Kelly, T. (2008), “A Scalability Study of Web-Native Information Visualization,” in *Proceedings of Graphics Interface 2008*, Canadian Information Processing Society, pp. 163–168. [556]
- Love, M. I., Huber, W., and Anders, S. (2014), “Moderated Estimation of Fold Change and Dispersion for RNA-seq Data with DESeq2,” *Genome Biology*, 15, 550. [561]
- McMurdie, P. J., and Holmes, S. (2013), “phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data,” *PLoS One*, 8, e61217. [561]
- Pollard, K. S., Dudoit, S., and van der Laan, M. J. (2005), “Multiple Testing Procedures: The Multtest Package and Applications to Genomics,” in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, eds. R. Gentleman, V. J. Carey, W. Huber, R. A. Irizarry, and S. Dudoit, New York: Springer, pp. 249–271. [561]
- R Core Team (2016), *R: A Language and Environment for Statistical Computing*, Vienna, Austria R Foundation for Statistical Computing. [553]
- Saito, K., and Nakano, R. (2002), “Extracting Regression Rules from Neural Networks,” *Neural Networks*, 15, 1279–1288. [554]
- Sankaran, K., and Holmes, S. (2014), “structSSI: Simultaneous and Selective Inference for Grouped or Hierarchically Structured Data,” *Journal of Statistical Software*, 59, 1–21. [561]
- Swayne, D. F., Cook, D., and Buja, A. (1998), “XGobi: Interactive Dynamic Data Visualization in the X Window System,” *Journal of Computational and Graphical Statistics*, 7, 113–130. [555]
- Vaidyanathan, R., Cheng, J., Allaire, J., Xie, Y., and Russell, K. (2014), “htmlwidgets: HTML Widgets for R,” R package version 0.3, 2. [555,562]
- Willett, W., Heer, J., and Agrawala, M. (2007), “Scented Widgets: Improving Navigation Cues with Embedded Visualizations,” *IEEE Transactions on Visualization and Computer Graphics*, 13, 1129–1136. [555]
- Xie, Y., Hofmann, H., Cook, D., and Cheng, X. (2013), “Cranvas: Interactive Statistical Graphics based on Qt,” R package version 0.8, 3. [555]
- Yekutieli, D. (2008), “Hierarchical False Discovery Rate–Controlling Methodology,” *Journal of the American Statistical Association*, 103, 309–316. [561]