

Visualising Relationships between Multi-Species Measures of Biodiversity and the Environment

Ritei Shibata

Keio University, Yokohama, Japan

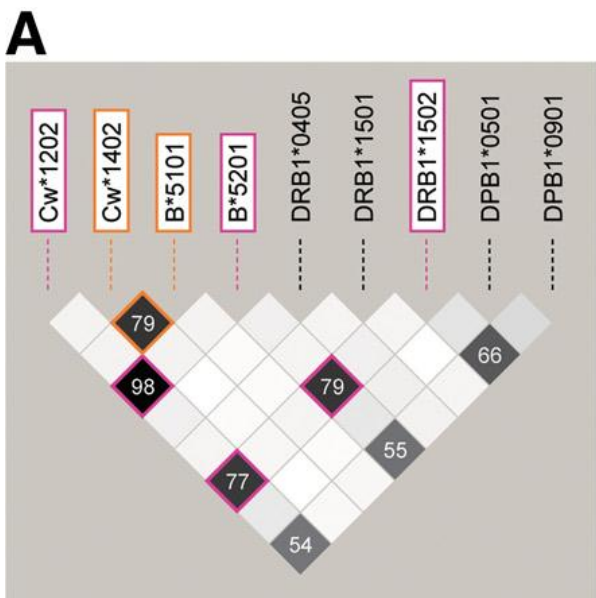
Outline

- Data Visualisation
- What is Textile Plot?
- GBR data
- Exploratory analysis through Textile Plot
- Grouping taxa through Textile Plot

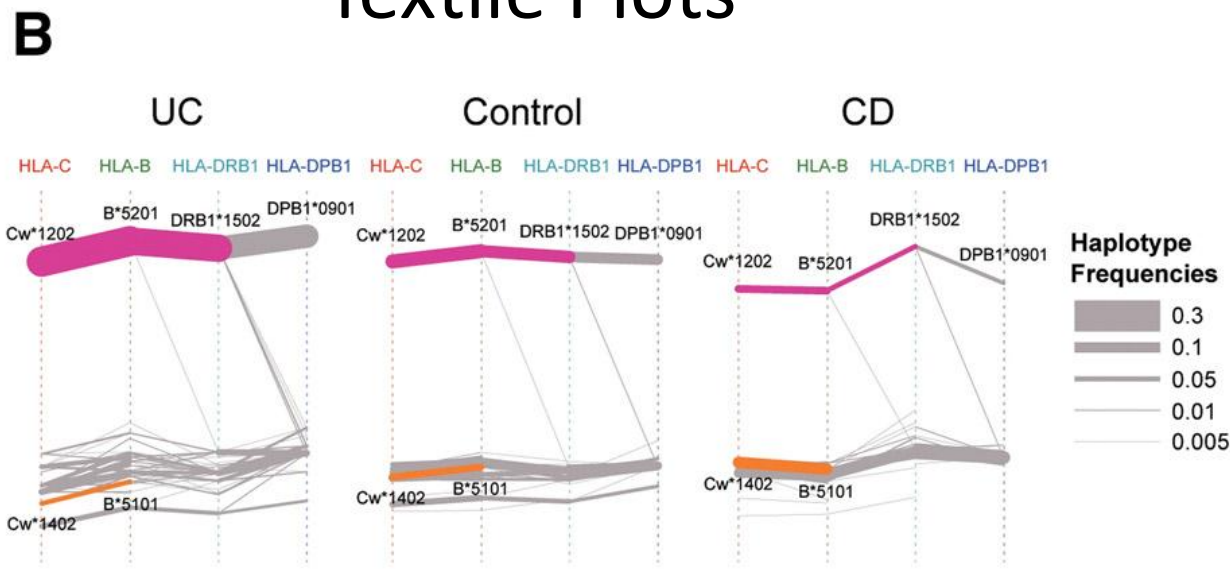
The aim of data visualisation

- Illustrate current status
 - Plant status
 - Network status
- Illustrate result
- Explore original data
 - High dimensional
 - Large data (records)
 - View the data as it is
 - Mixed data types
 - Numeric, Logical or Categorical
 - Help understanding of data

LD triangular display
with squared correlation coefficients
for Control



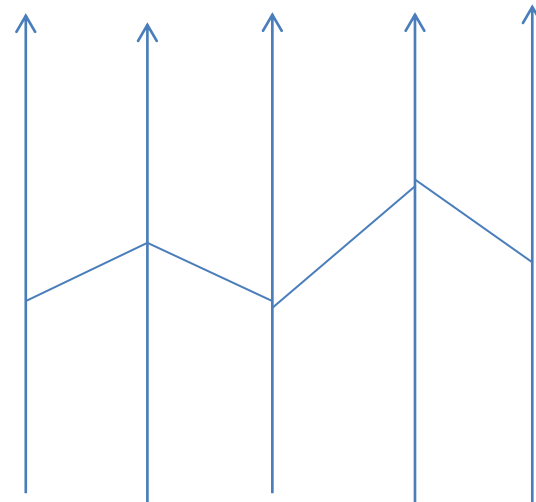
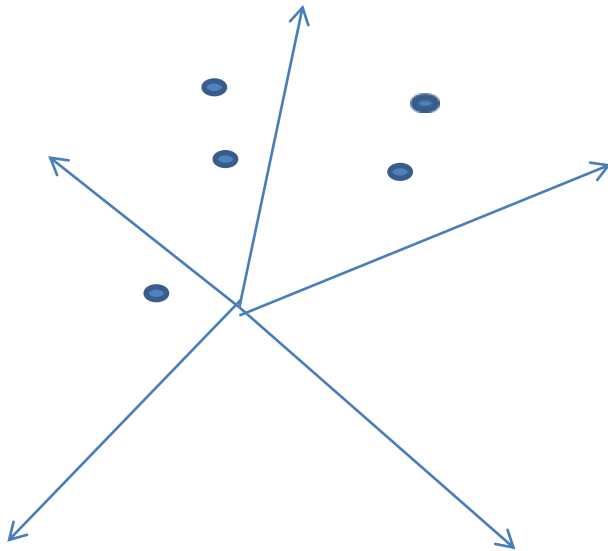
Textile Plots



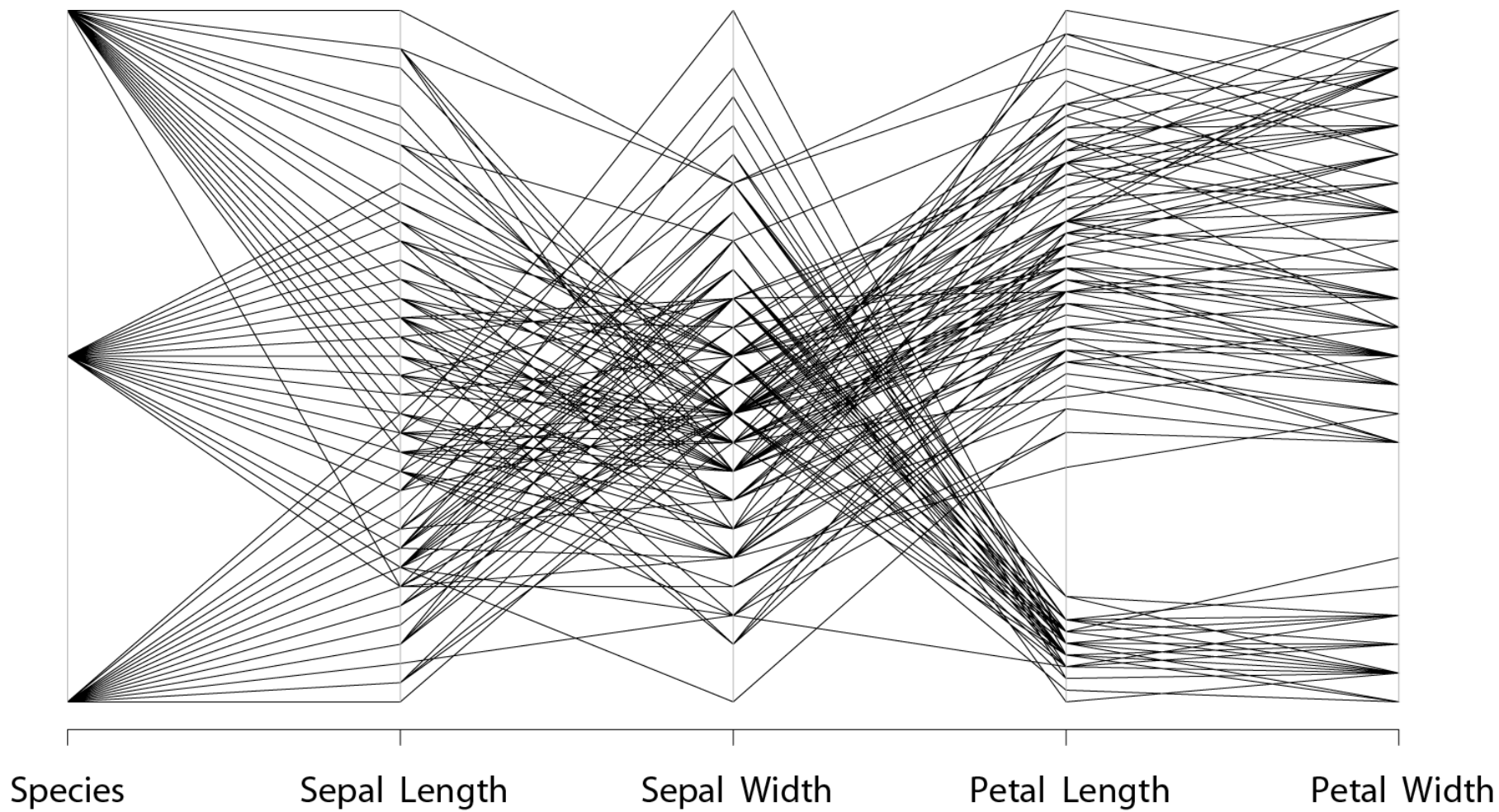
Categorical 4 Variables: HLA-C, HLA-B, HLA-DRB1, HLA-DPB1

Easier to grasp whole picture of data
More details if necessary

Parallel Coordinate Plot



Iris Data



Parallel Coordinate Plot

- ✓ High dimensional
- ✓ Large data (records)
- ✓ View the data as it is

Mixed data types

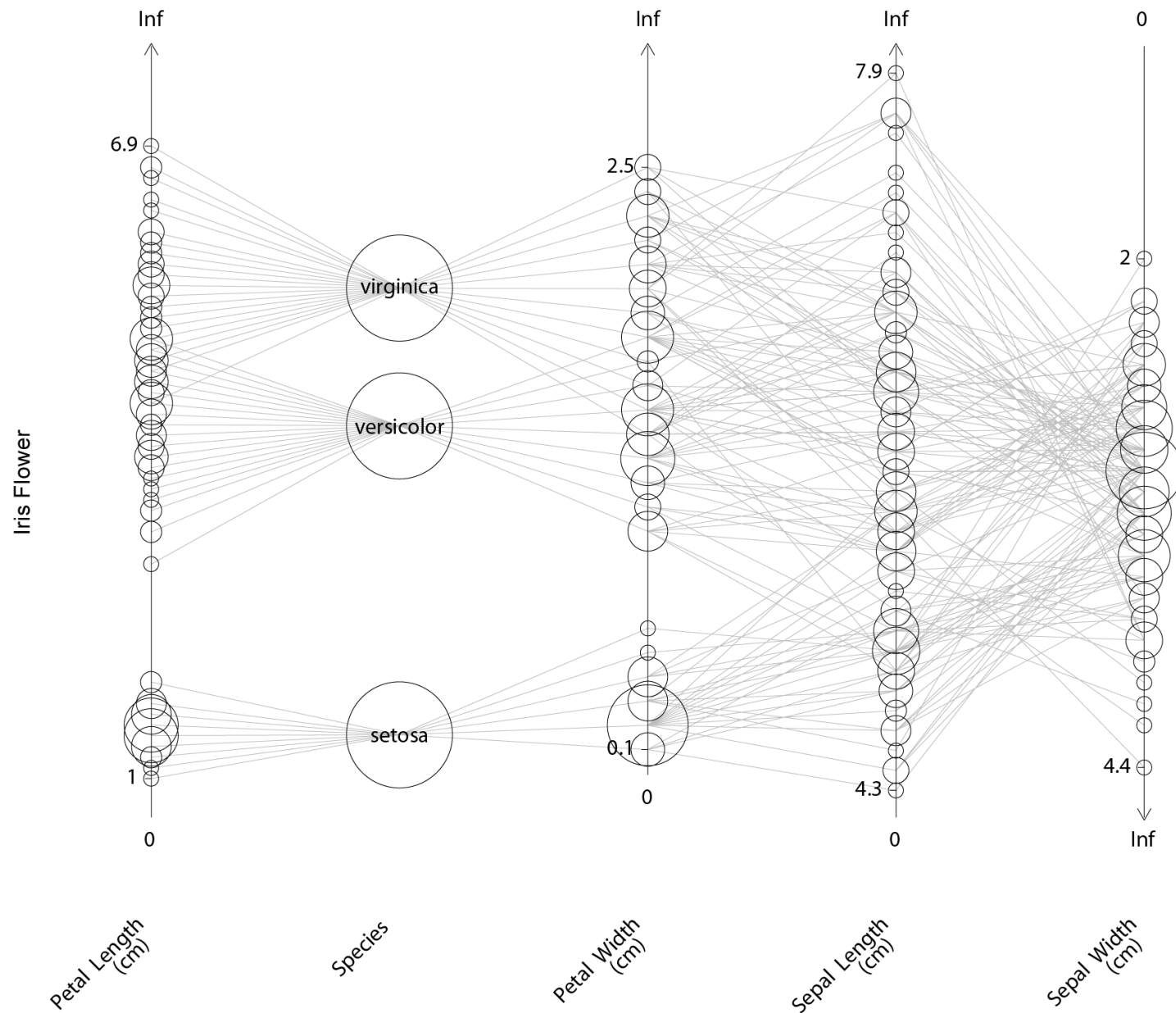
- Numeric, Logical or Categorical

Proper choice of coordinates

Help understanding of data

Proper choice of scale and location for each axis

Edsall, 2002, Unwin et al. , 2003, Tory et al., 2004,



Horizontalisation Criterion

- Choose location and scale of each axis so that connected lines become as horizontal as possible

$\mathbf{y}_j = \alpha_j \mathbf{1} + \beta_j \mathbf{x}_j$: coordinates on each axis $j = 1, 2, \dots, p$

$$\sum_{j=1}^p \|\mathbf{y}_j - \xi\|^2 \longrightarrow_{\alpha_j, \beta_j, j=1, \dots, p, \xi} \min$$

Kumasaka and Shibata, High-dimensional data
visualisation: The textile plot. 2007, Computational
Statistics and Data Analysis

$$\sum_{j=1}^p \| \mathbf{y}_j - \boldsymbol{\xi} \|^2 = \sum_{i=1}^n \left(\sum_{j=1}^p (y_{ij} - \xi_i)^2 \right) \rightarrow \min$$

ξ_i : horizontal level of the i th record

$\sum_{j=1}^p (y_{ij} - \xi_i)^2$: squared deviance of the i th record
from the horizontal level ξ_i

Mixed Data Type Case

If \mathbf{x}_j is categorical, apply a contrast to get a data matrix X_j

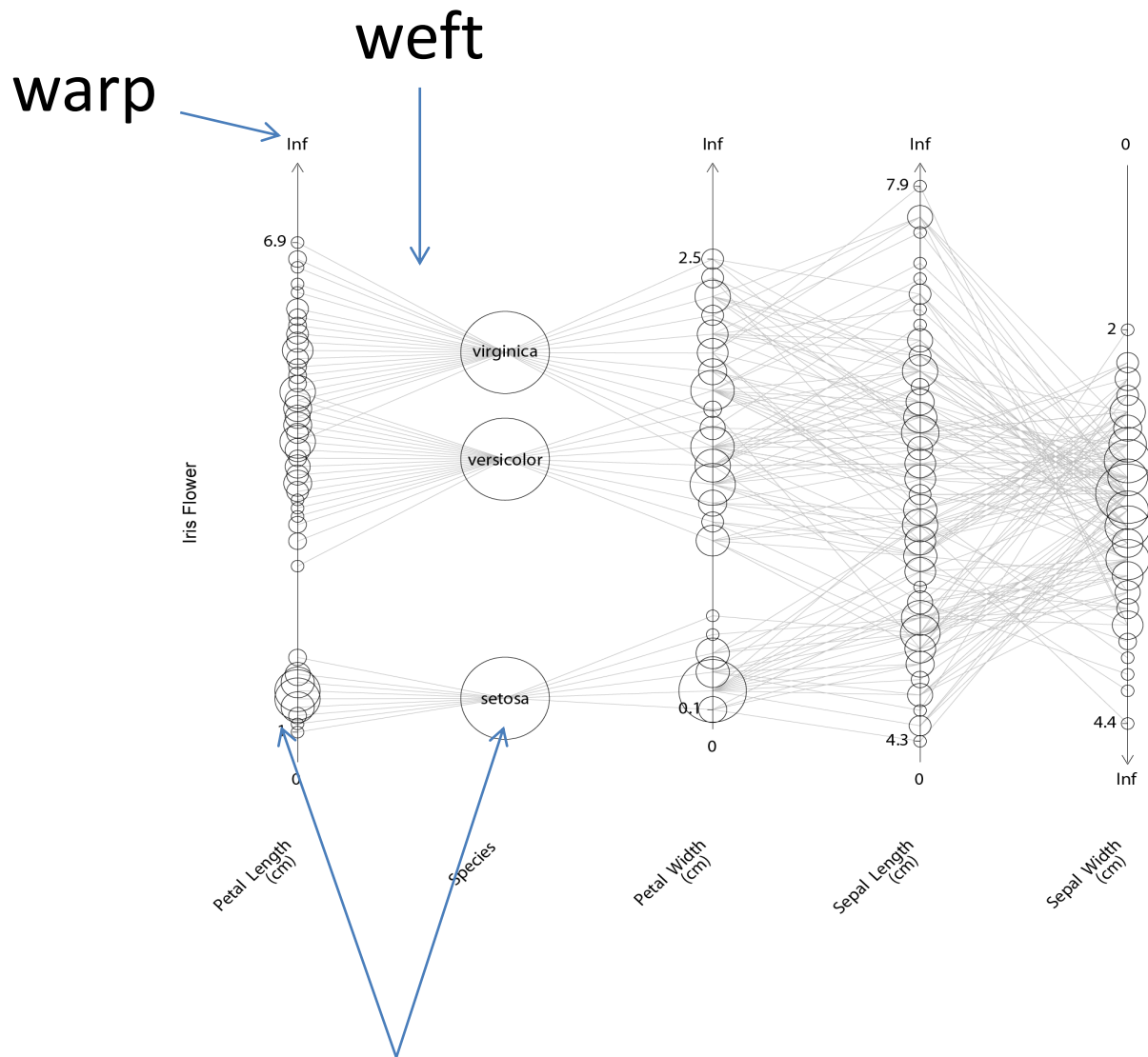
$\mathbf{y}_j = \alpha_j \mathbf{1} + \boldsymbol{\beta}_j X_j$: coordinates on the j th axis

Horizontalisation criterion

$$\sum_{j=1}^p \|\mathbf{y}_j - \boldsymbol{\xi}\|^2 \longrightarrow \min_{\alpha_j, \boldsymbol{\beta}_j, j=1, \dots, p, \boldsymbol{\xi}}$$

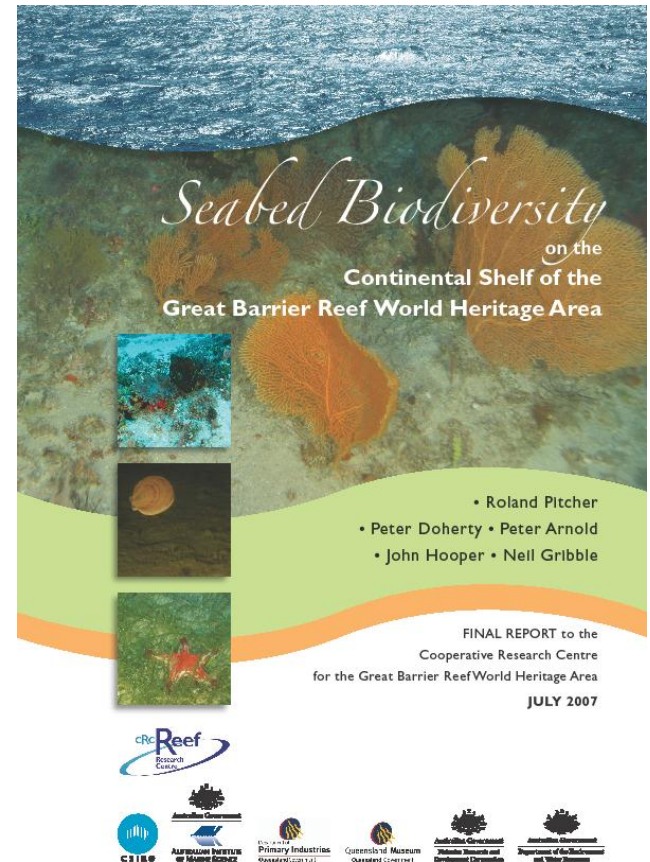
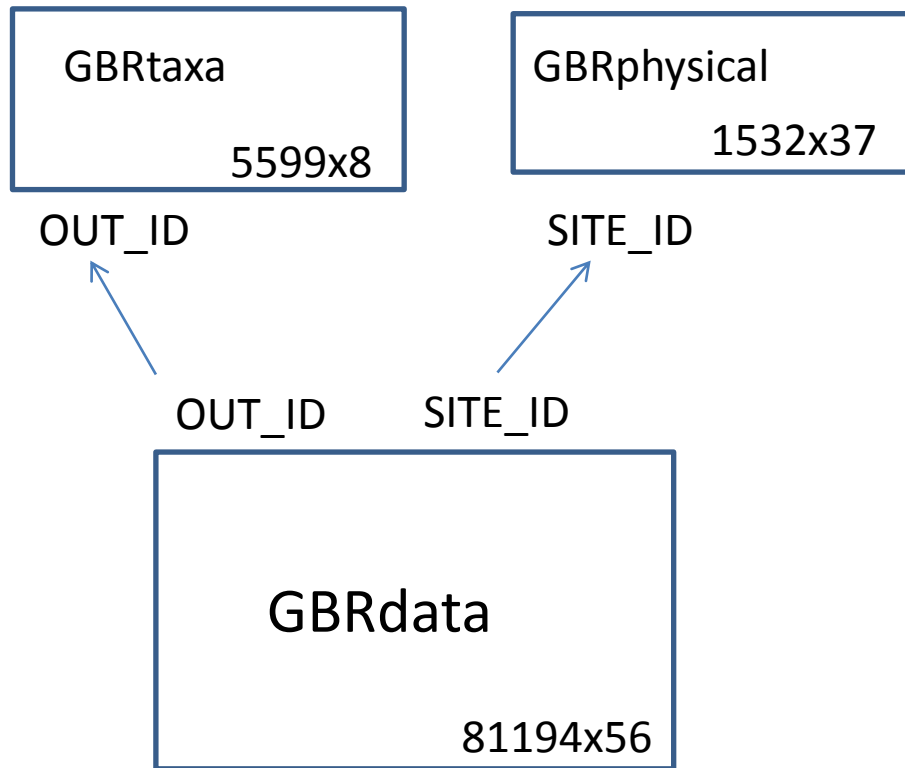
determines *a* proper location of the levels of each category.

Independent of the choice of contrast

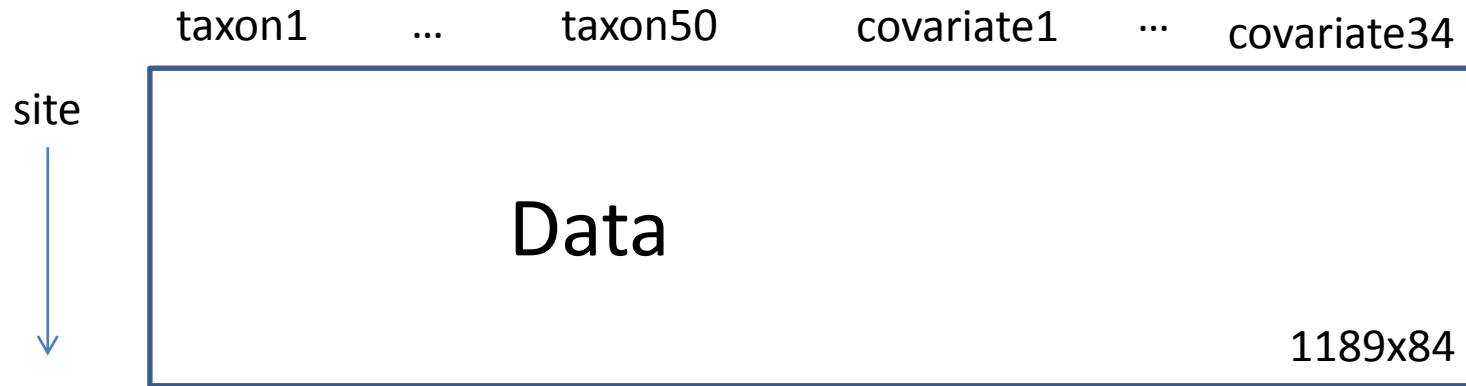


Proportional to the multiplicity of the value

GBR data



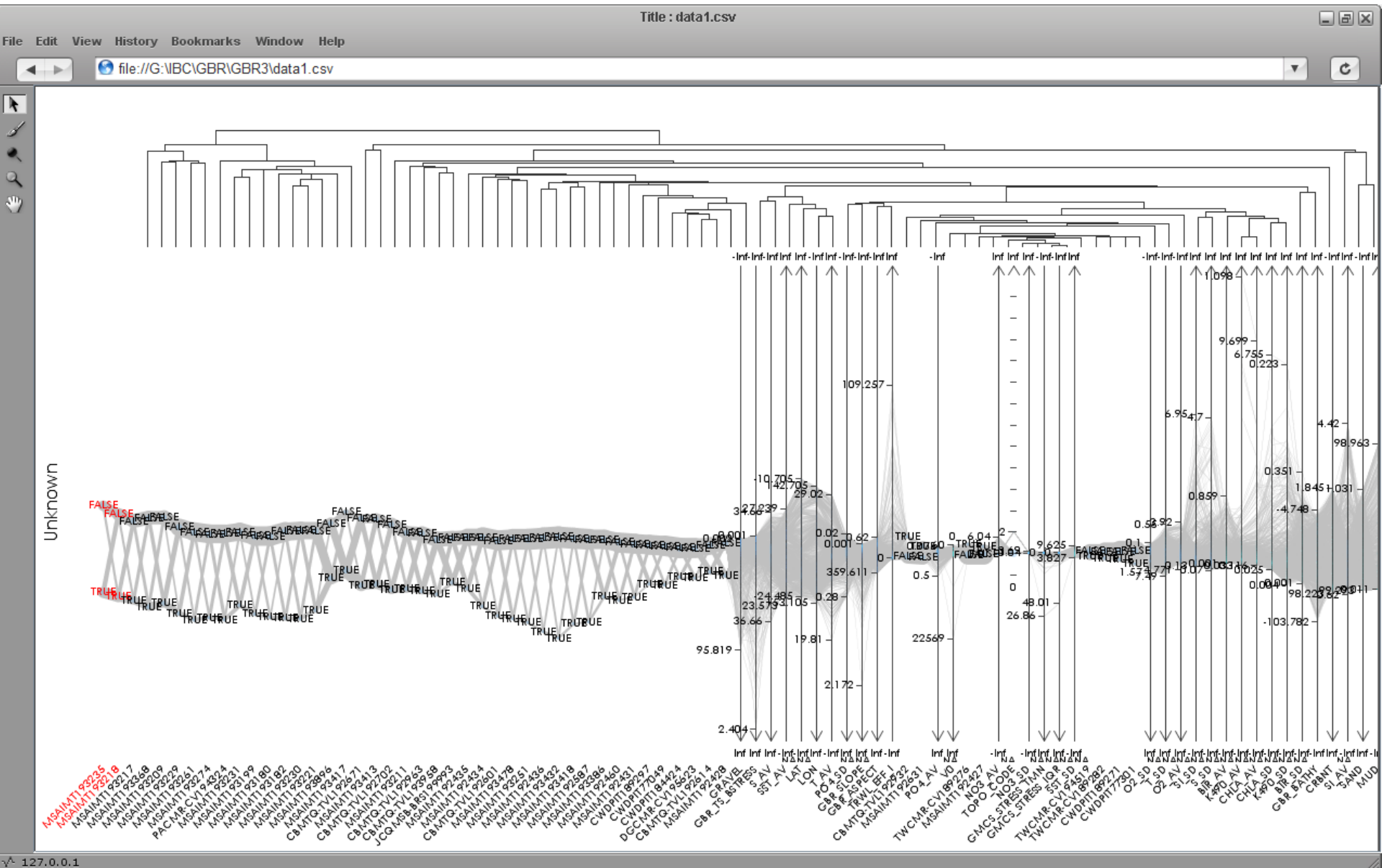
Most frequent 50 taxa



taxon1 ... taxon50 : logical
covariate1 ... covariate34: numeric

Piers K. Dunstan, Scott D. Foster and Ross Darnell,
Model based group of species across environmental gradients
Ecological Modelling, 2010

Textile Plot (34 covariates and 50 taxa)



[illegible]

Order of axes and knots

- Order of axes
 - Order appeared in the data table
 - Variance of each axis
 - Clustering of axes
 - Distance of two axes=Sum of squares of slopes
 - Ordered single end-linkage clustering algorithm(Hurley, 2004)
- Variables with Knots
 - Orthogonal to other variables

Covariates Orthogonal to Existence of Taxa

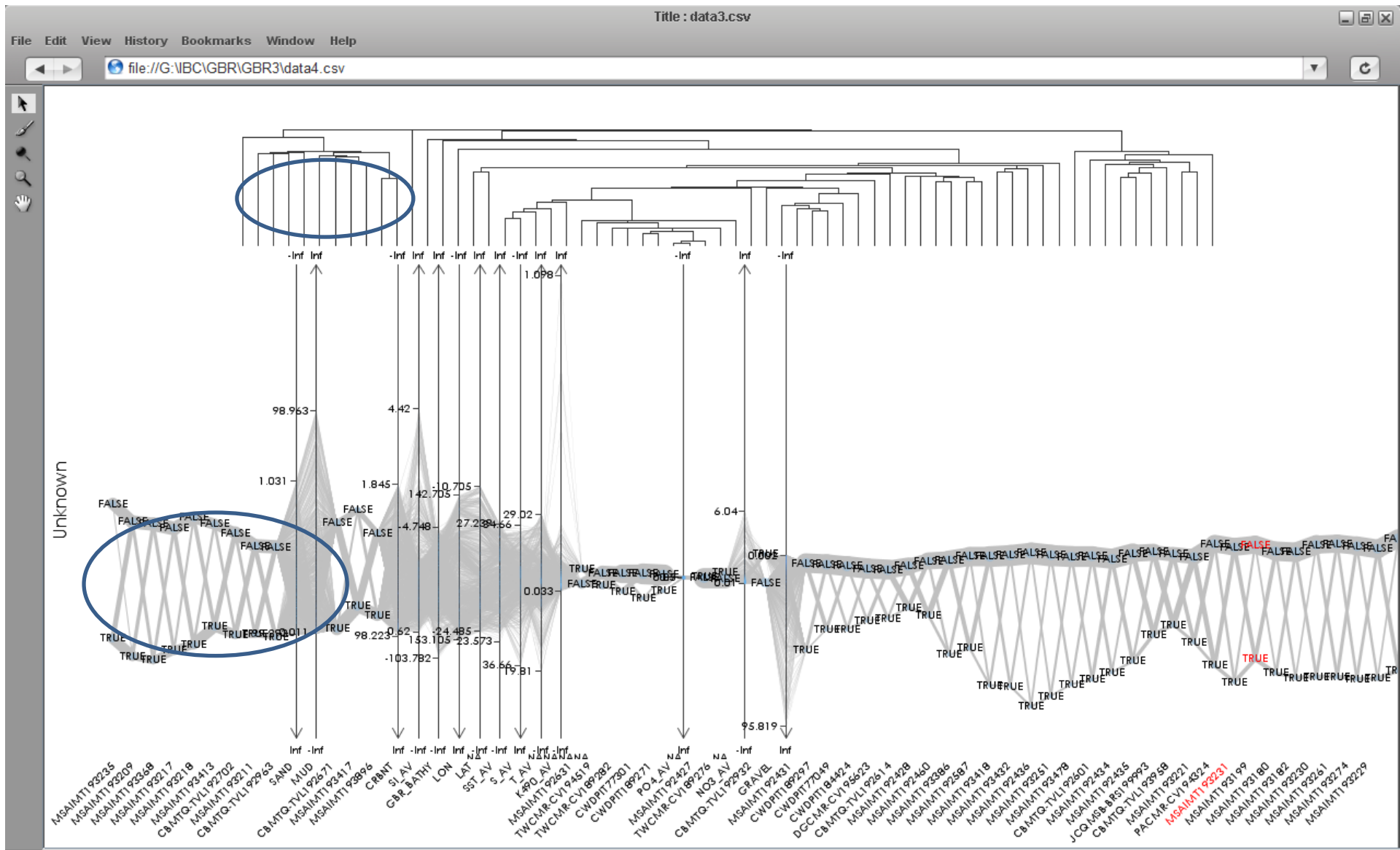
Standard deviations and other covariates
(20 covariates)

```
[1] "GBR_ASPECT"    "GBR_SLOPE"    "GBR_TS_BSTRESS" "GMCS_STRESS_TMN"  
[5] "GMCS_STRESS_IQR" "NO3_SD"      "PO4_SD"      "O2_AV"  
[9] "O2_SD"        "S_SD"        "T_SD"        "SI_SD"  
[13] "CHLA_AV"      "CHLA_SD"     "K490_SD"     "SST_SD"  
[17] "BIR_AV"       "BIR_SD"      "TRWL_EFF_I"  "TOPO_CODE"
```

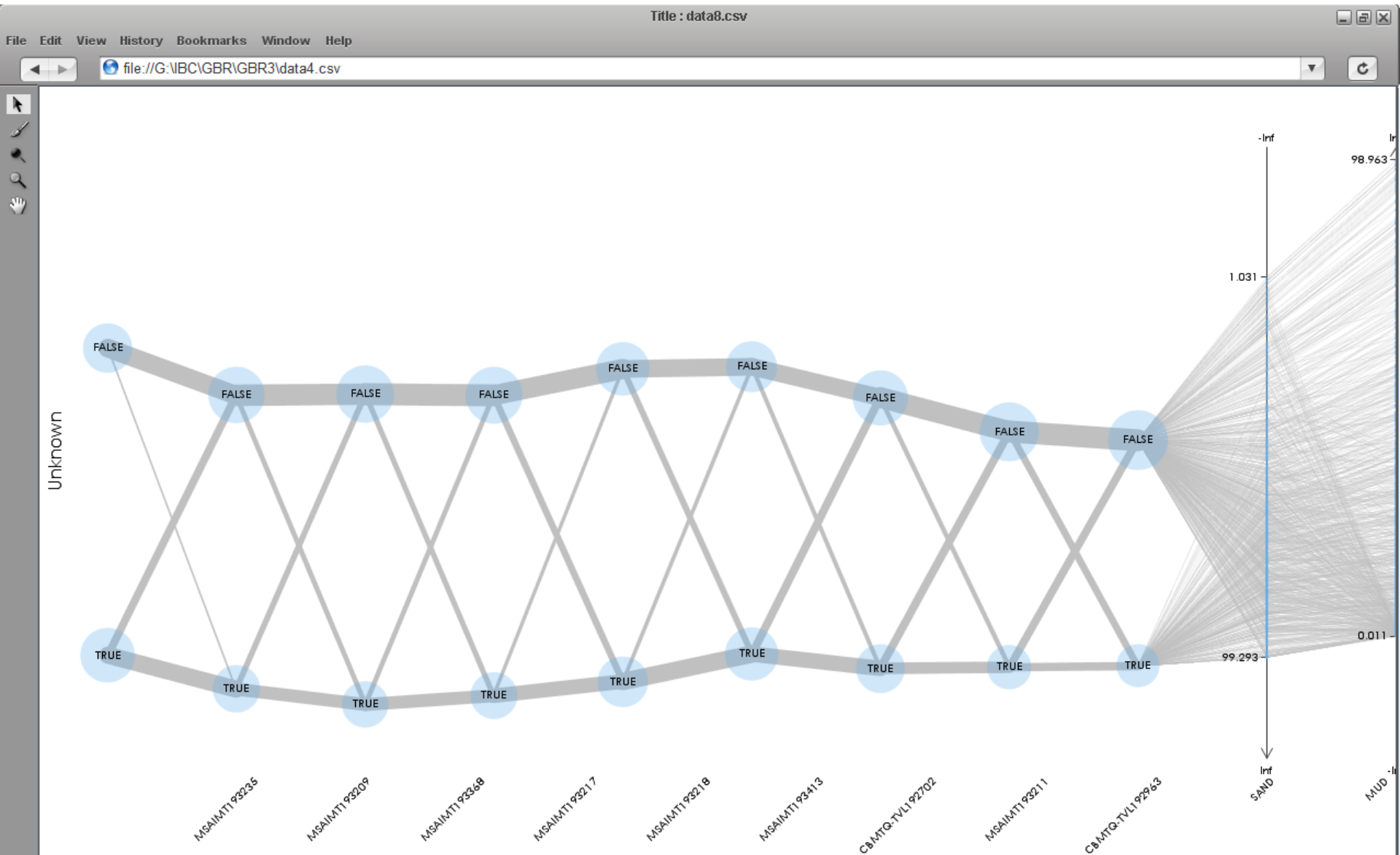


Delete

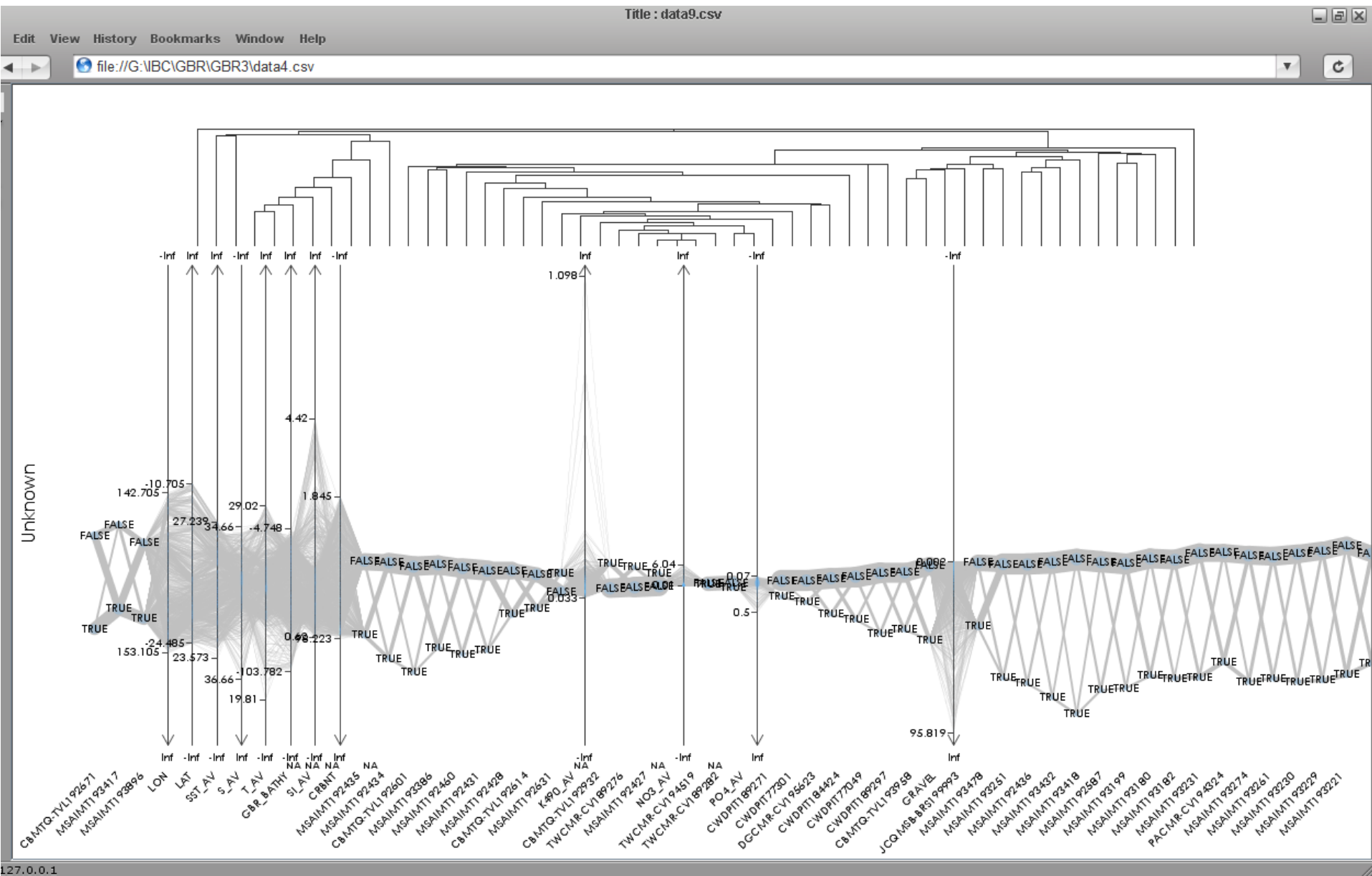
14 Covariates



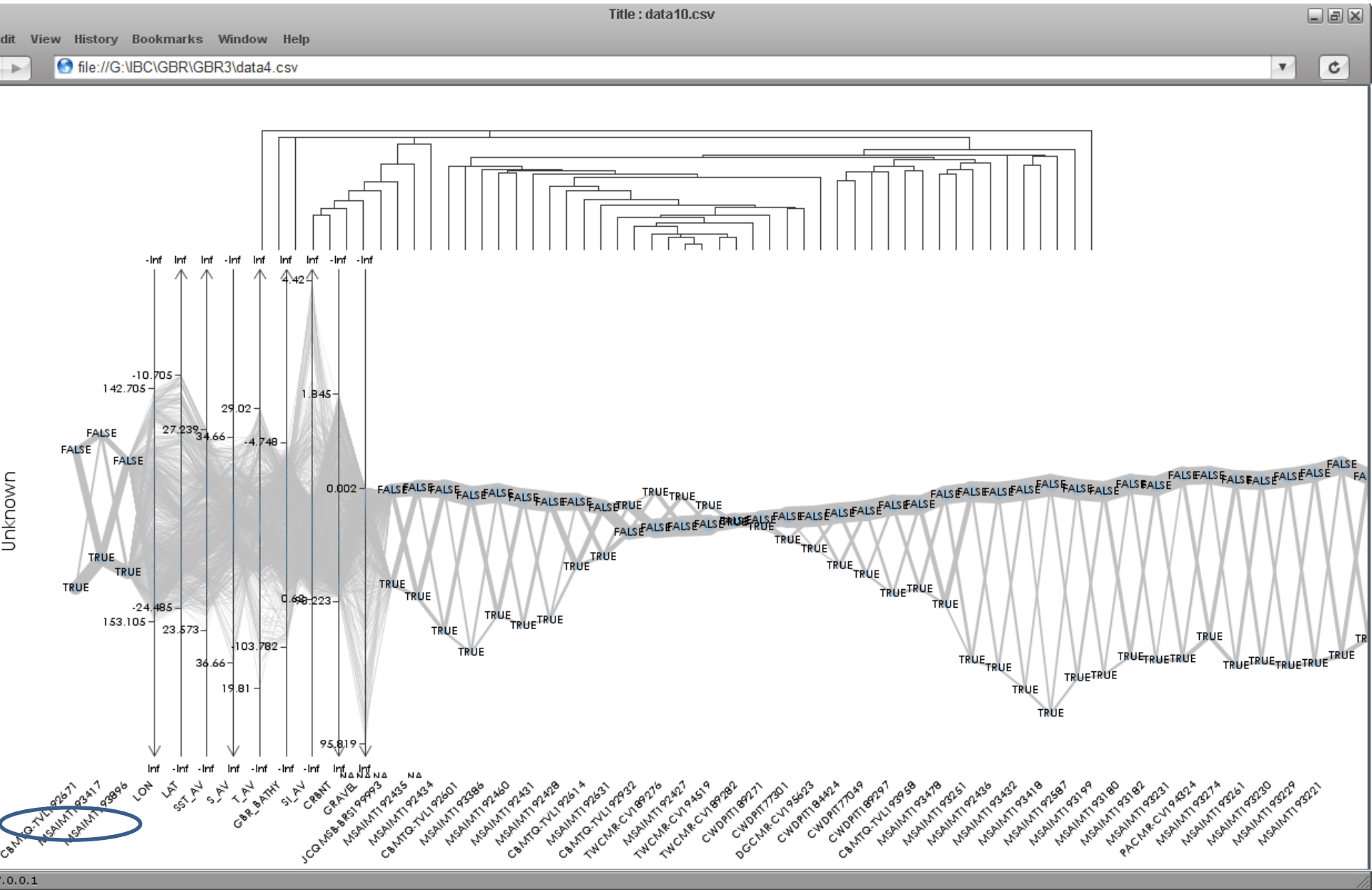
Group1: Sand appetite 9 Taxa



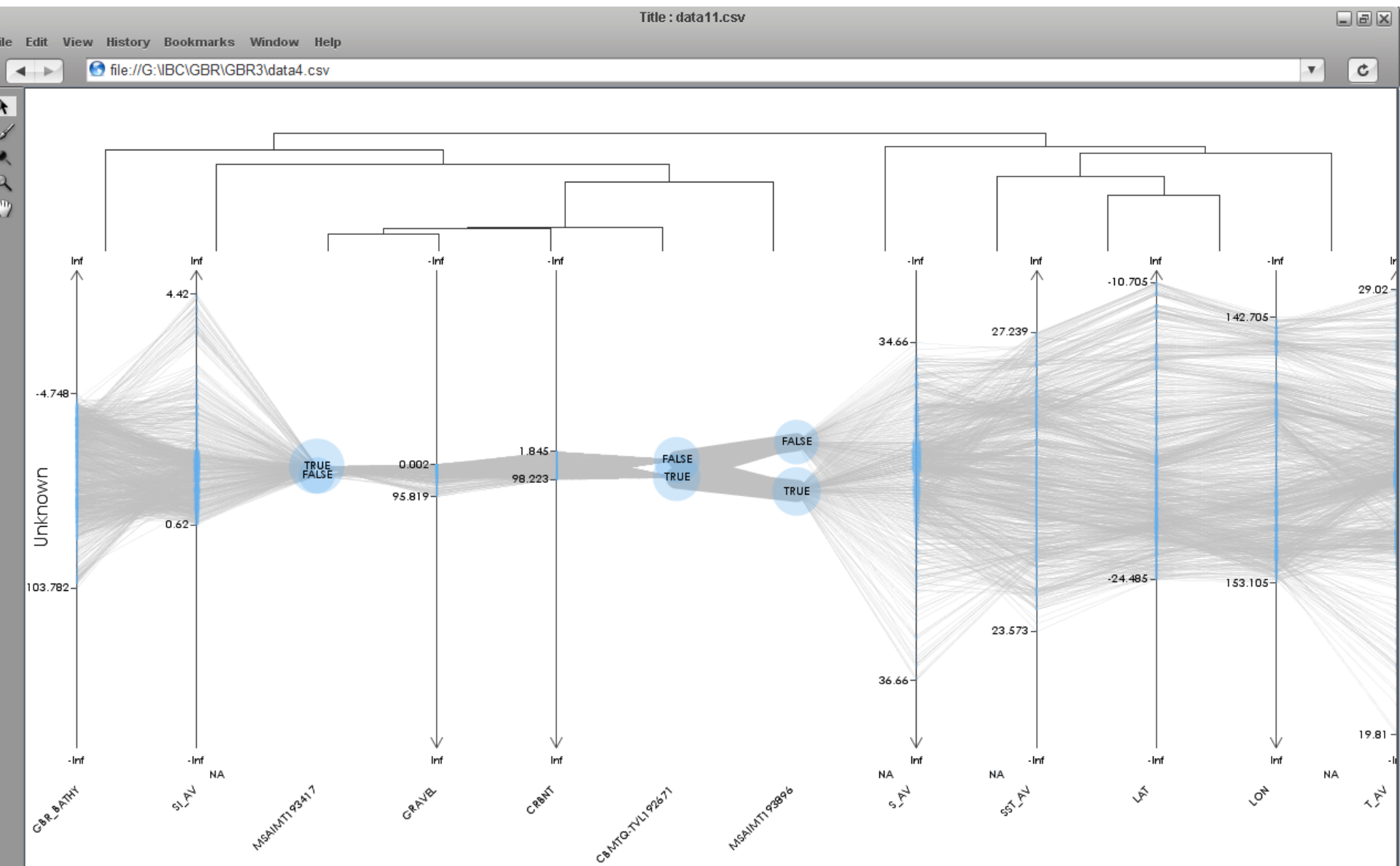
Remaining 41 Taxa: Knots: NO3_AV, PO4_AV and K490_AV



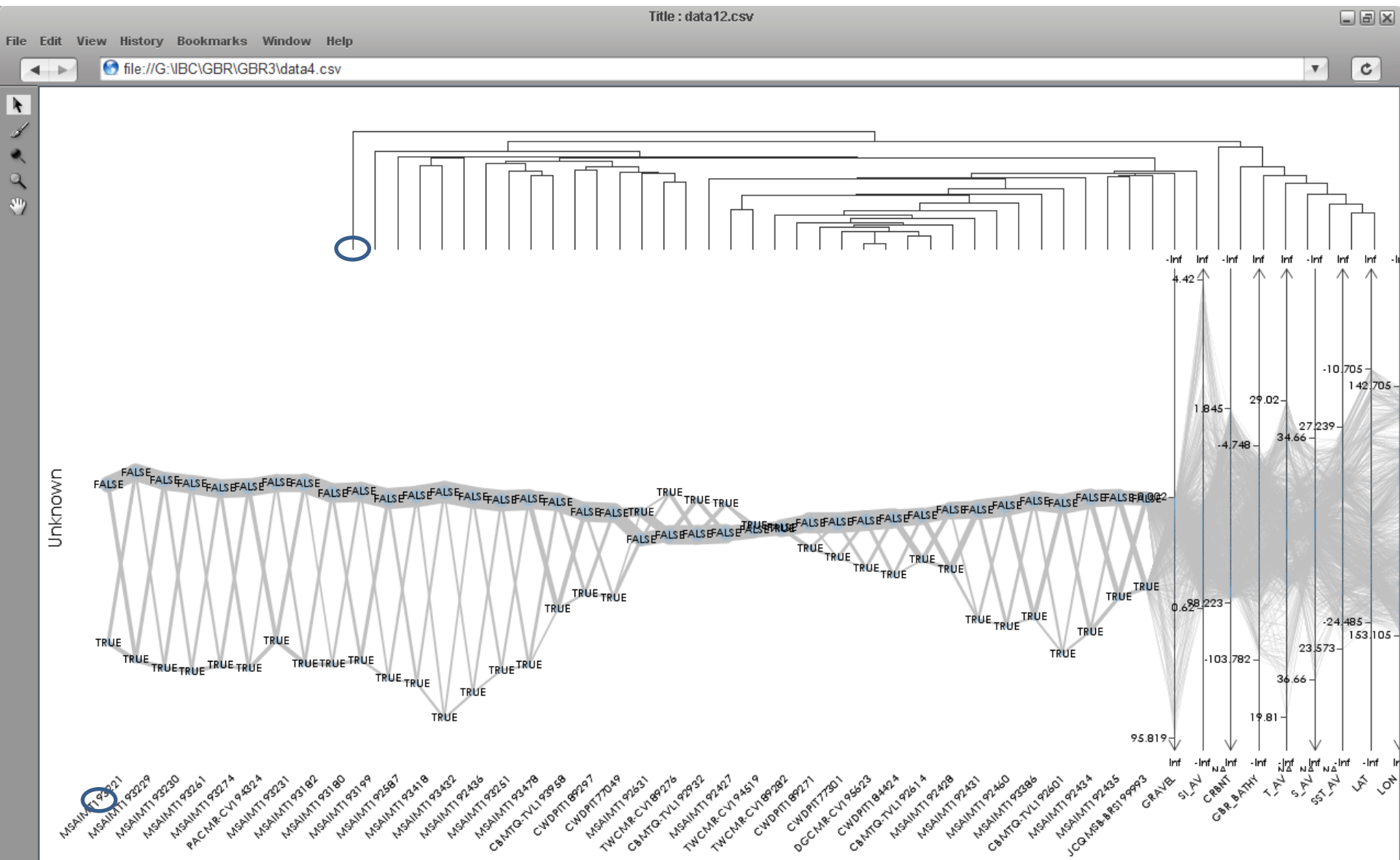
After 3 covariates removed



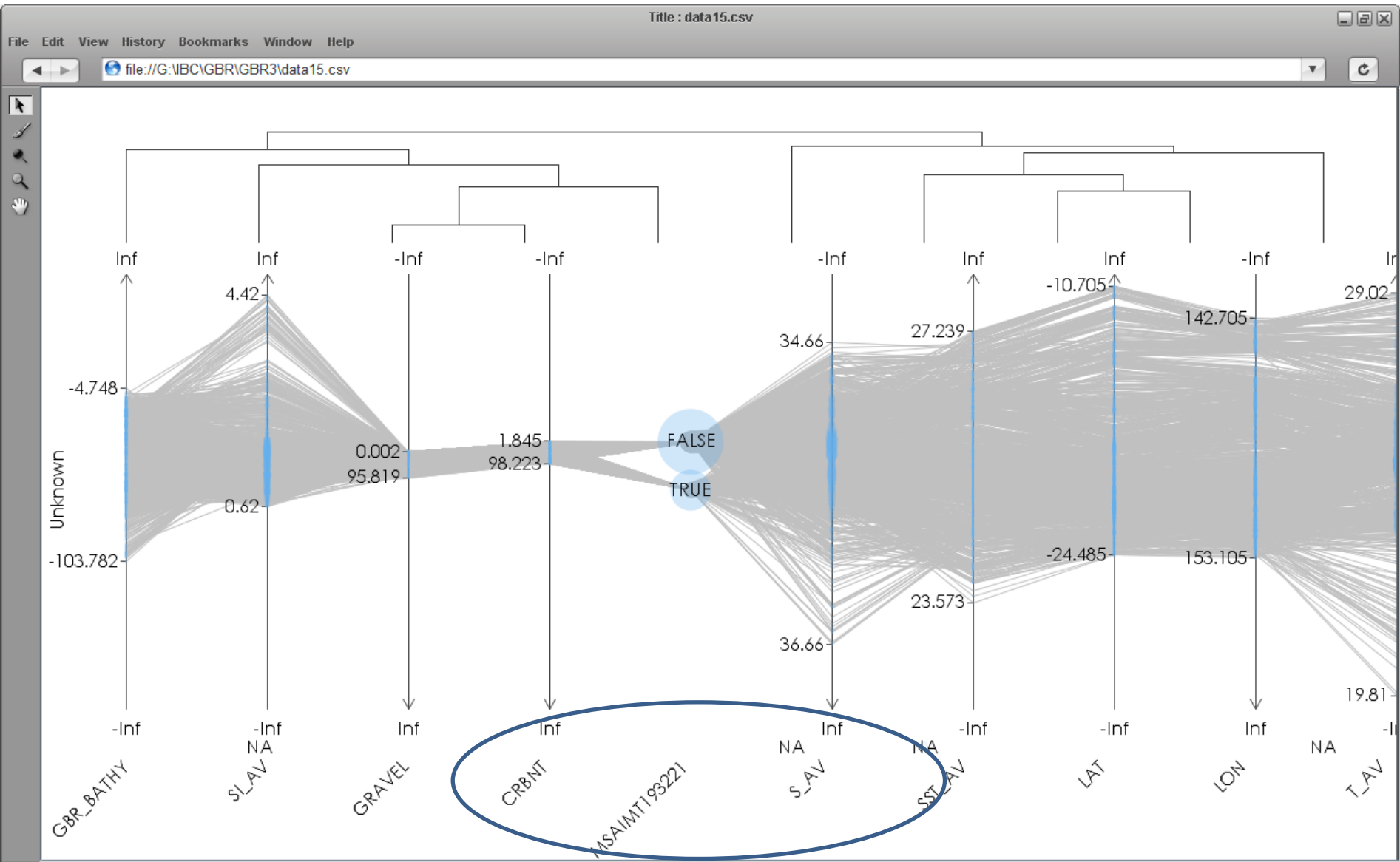
Group2: Most frequent 3 taxa



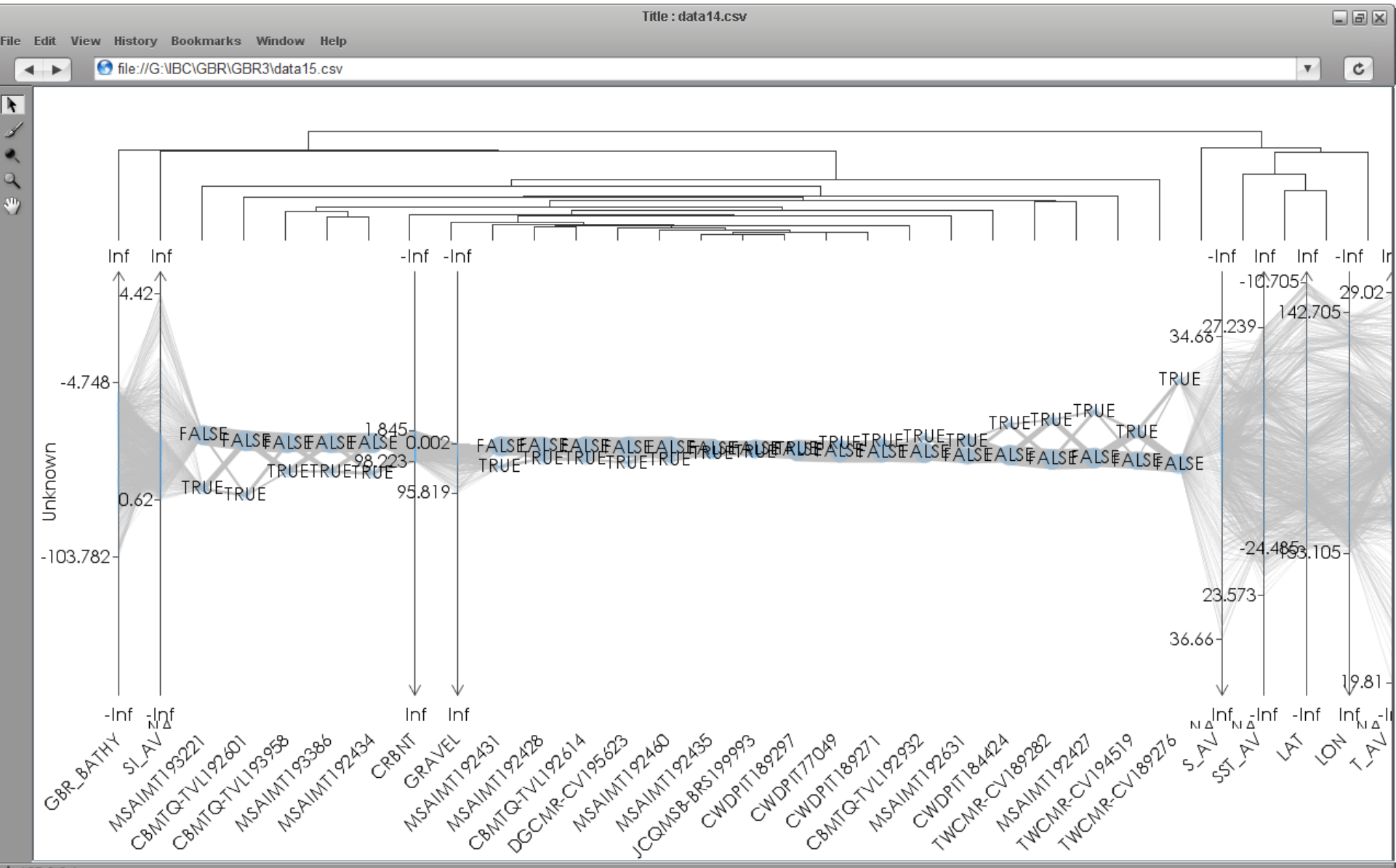
Remaining 38 taxa



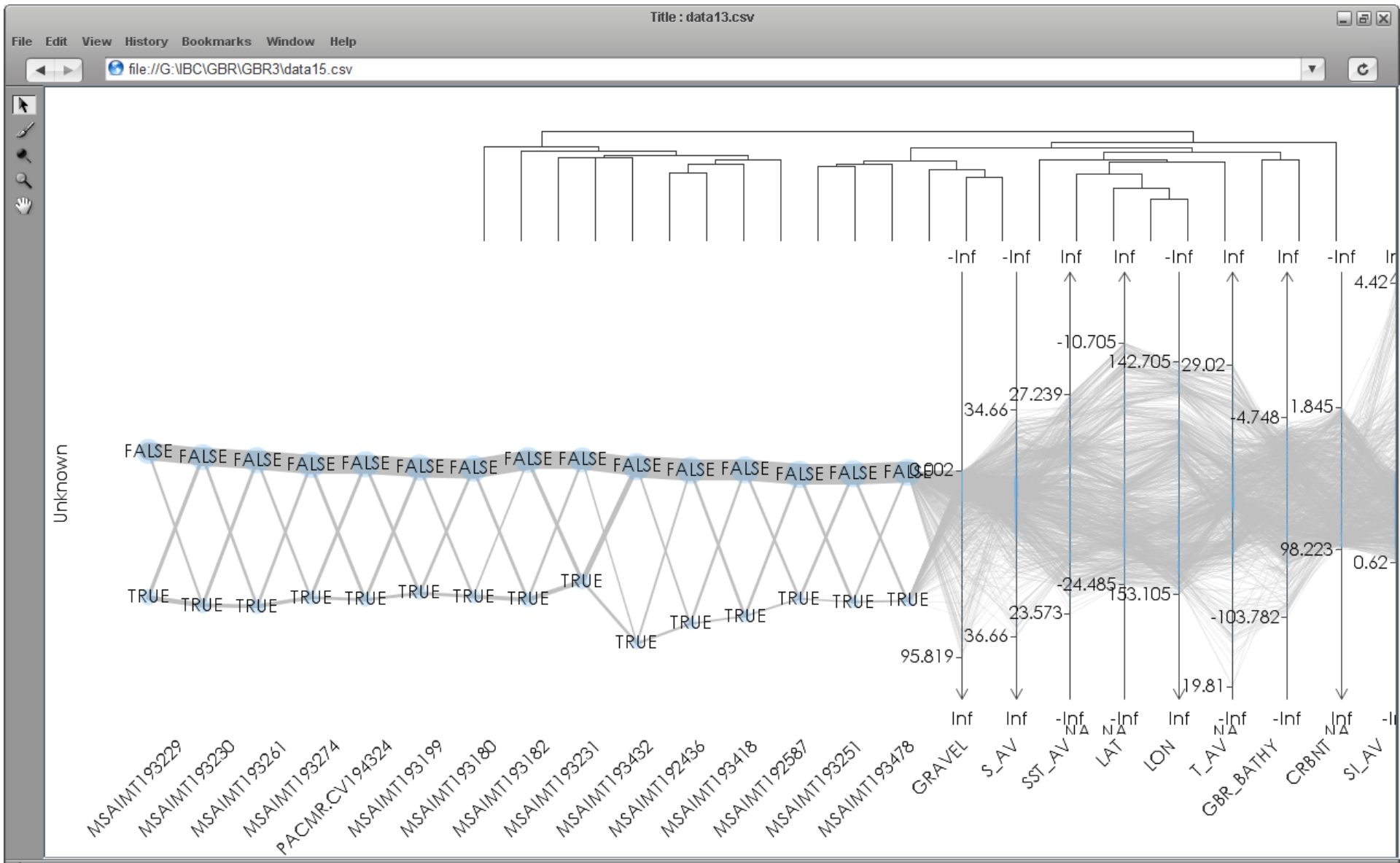
Group3: Unique taxon



Group4:22 taxa

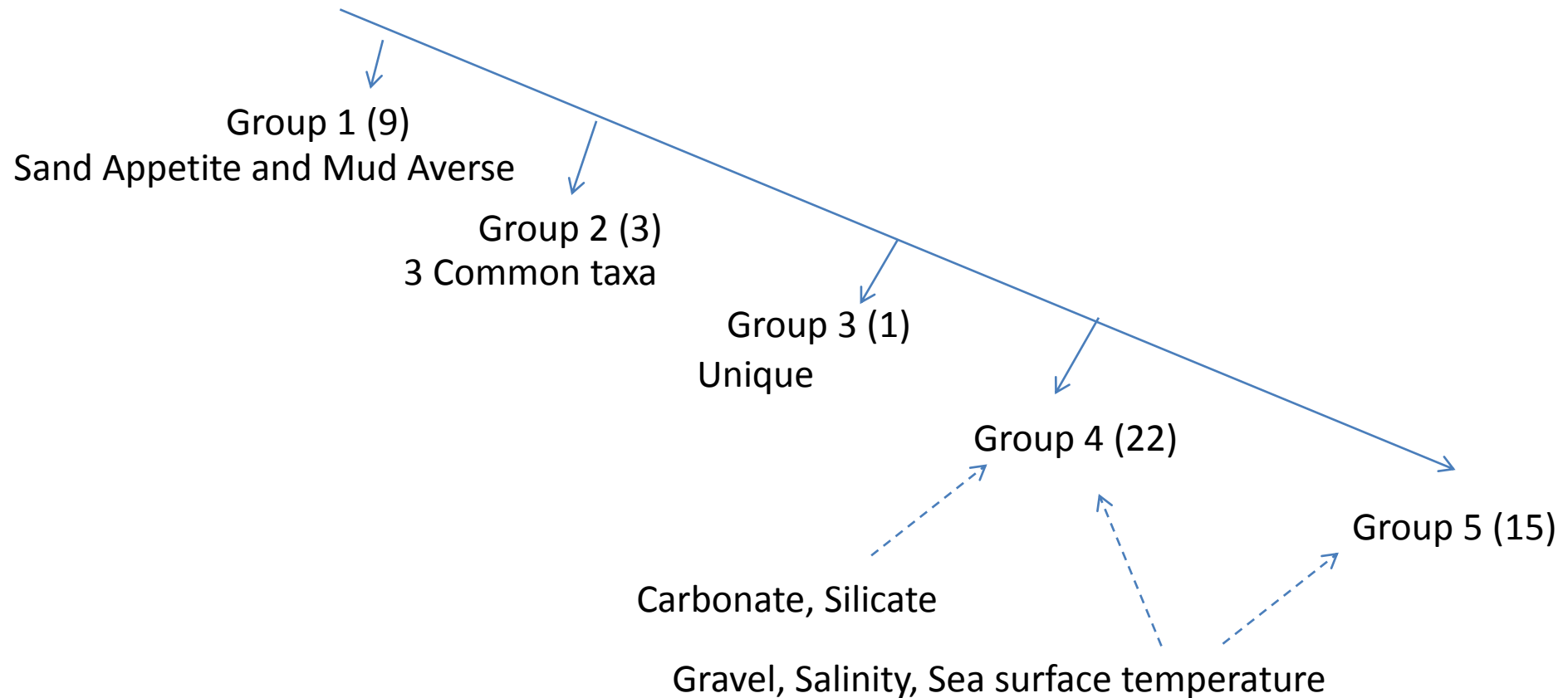


Group5: 15 taxa



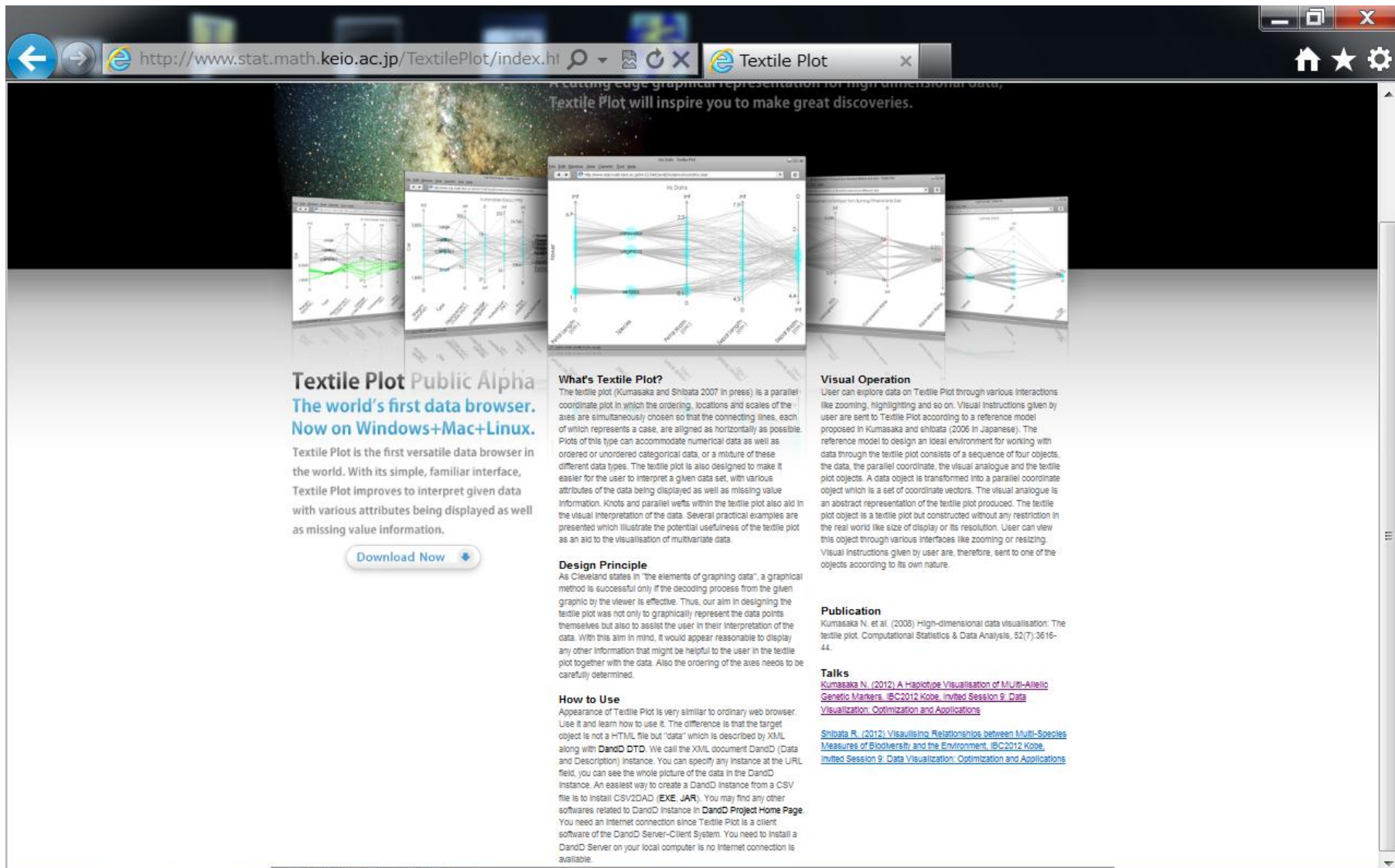
Grouping taxa through TextilePlot

- Knot covariates removed step by step
- Examine Hierarchical Cluster Tree of Axes (Variables)



Textile Plot Home Page

<http://www.stat.math.keio.ac.jp/TextilePlot/index.html>



← → http://www.stat.math.keio.ac.jp/TextilePlot/index.html Textile Plot

Creating edge graphical representation for high dimensional data,
Textile Plot will inspire you to make great discoveries.

Textile Plot Public Alpha
The world's first data browser.
Now on Windows+Mac+Linux.

Textile Plot is the first versatile data browser in the world. With its simple, familiar interface, Textile Plot improves to interpret given data with various attributes being displayed as well as missing value information.

[Download Now](#)

What's Textile Plot?
The textile plot (Kumasaka and Shiohata 2007 in press) is a parallel coordinate plot in which the ordering, locations and scales of the axes are simultaneously chosen so that the connecting lines, each of which represents a case, are aligned as horizontally as possible. Plots of this type can accommodate numerical data as well as ordered or unordered categorical data, or a mixture of these different data types. The textile plot is also designed to make it easier for the user to interpret a given data set, with various attributes of the data being displayed as well as missing value information. Knots and parallel wefts within the textile plot also aid in the visual interpretation of the data. Several practical examples are presented which illustrate the potential usefulness of the textile plot as an aid to the visualisation of multivariate data.

Visual Operation
User can explore data on Textile Plot through various interactions like zooming, highlighting and so on. Visual instructions given by user are sent to Textile Plot according to a reference model proposed in Kumasaka and shiohata (2006 in Japanese). The reference model to design an ideal environment for working with data through the textile plot consists of a sequence of four objects, the data, the parallel coordinate, the visual analogue and the textile plot objects. A data object is transformed into a parallel coordinate object which is a set of coordinate vectors. The visual analogue is an abstract representation of the textile plot produced. The textile plot object is a textile plot but constructed without any restriction in the real world like size of display or its resolution. User can view this object through various interfaces like zooming or resizing. Visual instructions given by user are, therefore, sent to one of the objects according to its own nature.

Design Principle
As Cleveland states in "the elements of graphing data", a graphical method is successful only if the decoding process from the given graphic by the viewer is effective. Thus, our aim in designing the textile plot was not only to graphically represent the data points themselves but also to assist the user in their interpretation of the data. With this aim in mind, it would appear reasonable to display any other information that might be helpful to the user in the textile plot together with the data. Also the ordering of the axes needs to be carefully determined.

How to Use
Appearance of Textile Plot is very similar to ordinary web browser. Use it and learn how to use it. The difference is that the target object is not a HTML file but "data" which is described by XML along with **DandD DTD**. We call the XML document DandD (Data and Description) Instance. You can specify any Instance at the URL field, you can see the whole picture of the data in the DandD Instance. An easiest way to create a DandD Instance from a CSV file is to install CSV2DAD (EXE, JAR). You may find any other softwares related to DandD Instance in **DandD Project Home Page**. You need an Internet connection since Textile Plot is a client software of the DandD Server-Client System. You need to install a DandD Server on your local computer if no Internet connection is available.

Publication
Kumasaka N. et al. (2008) High-dimensional data visualisation: The textile plot. *Computational Statistics & Data Analysis*, 52(7):3616-44.

Talks
Kumasaka N. (2012) A Prototype Visualisation of Multi-Allelic Genetic Markers. [IBC2012 Kobe, invited Session 9, Data Visualization, Optimization and Applications](#)
Shiohata R. (2012) Visualising Relationships between Multi-Species Measures of Biodiversity and the Environment. [IBC2012 Kobe, invited Session 9, Data Visualization, Optimization and Applications](#)

Copyright(C)2007 Naoki Kumasaka All Rights Reserved.