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

Ritei Shibata

Keio University, Yokohama, Japan

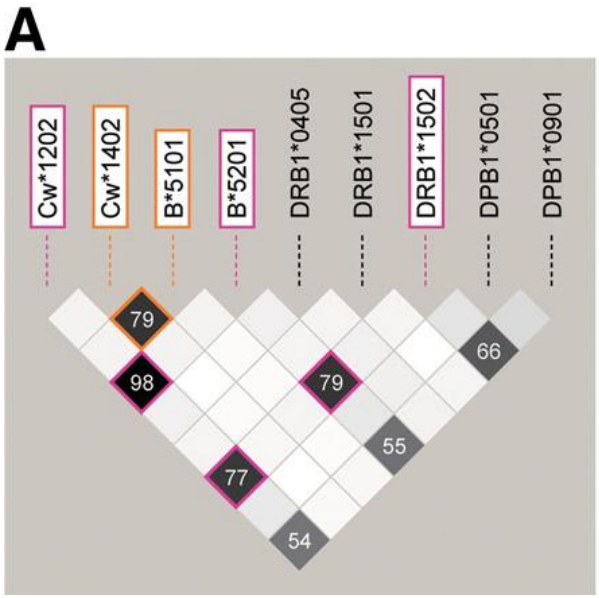
Outline

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

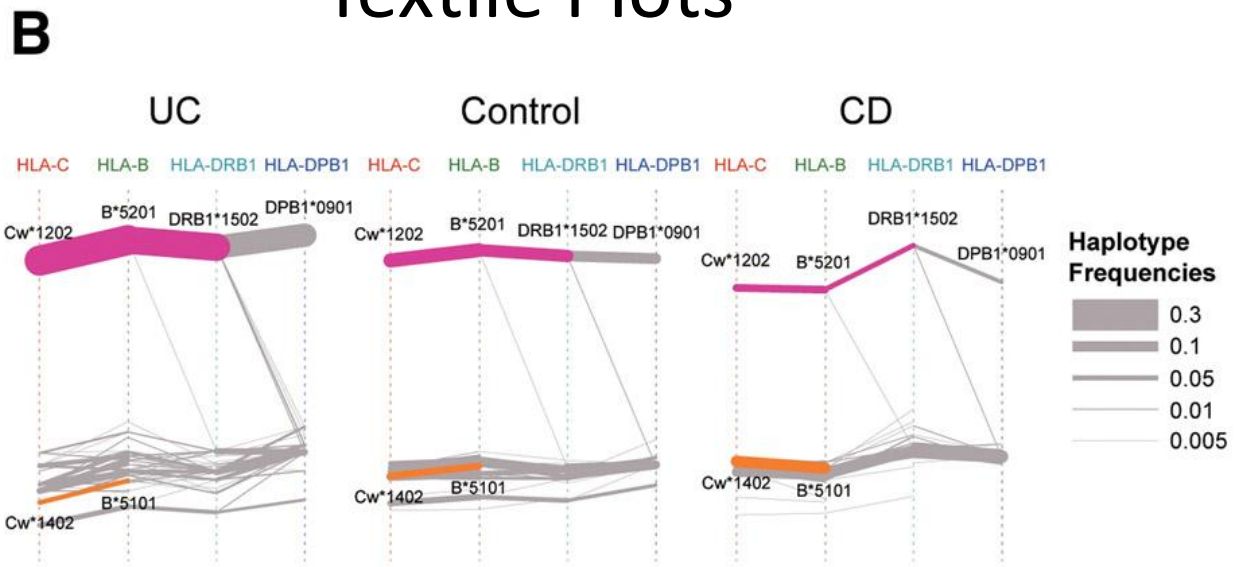
The aim of data visualisation

- 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 coeffients
for Control



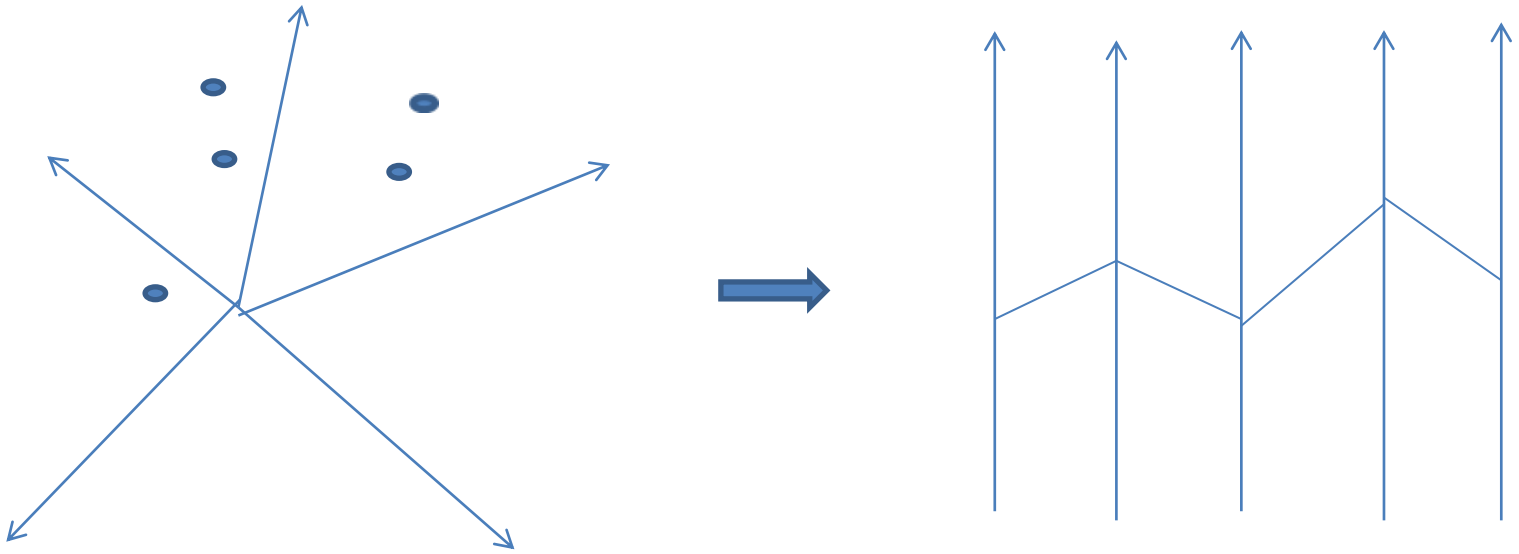
Textile Plots



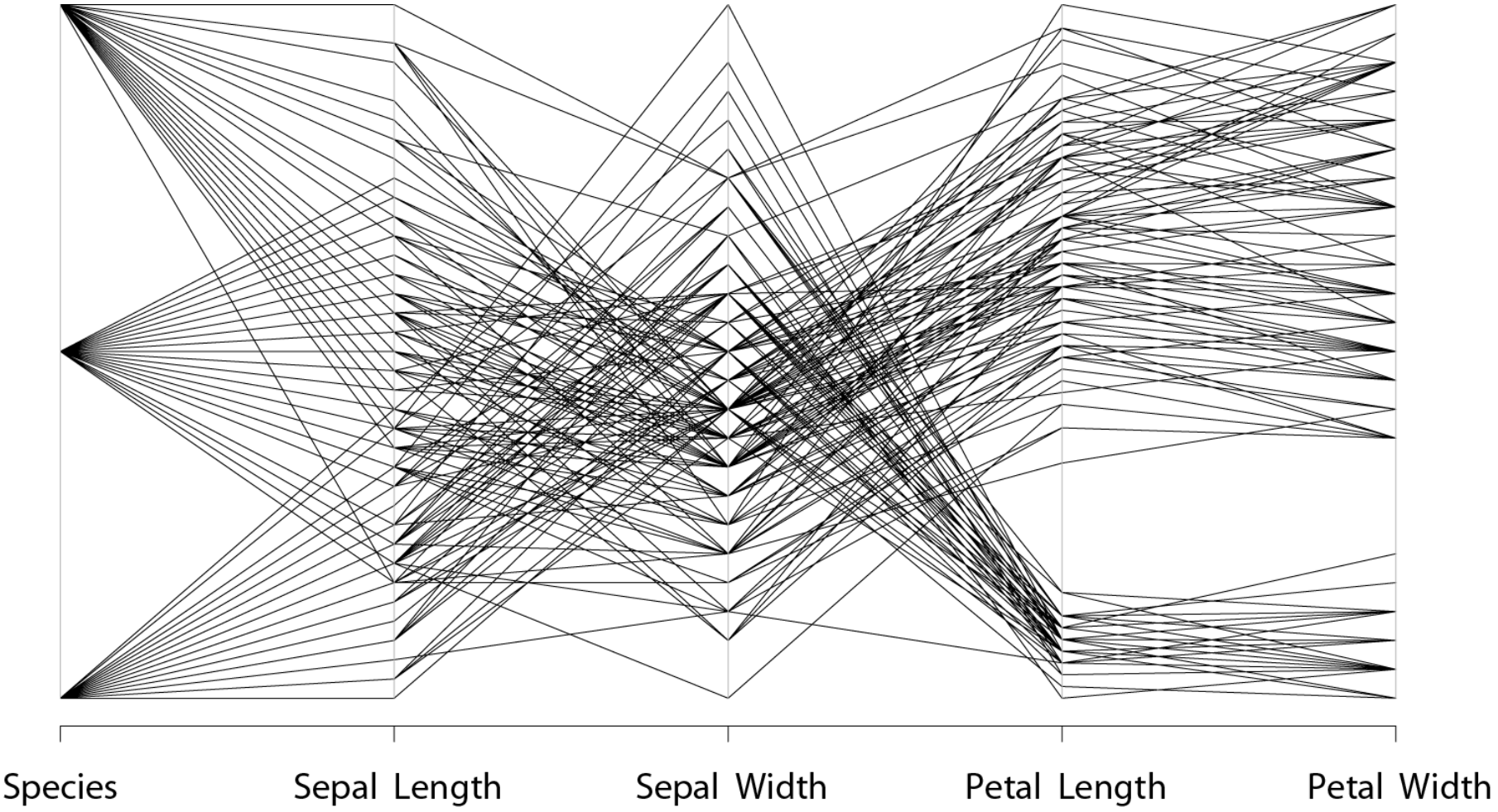
Categorical 4 Variables: HLA-C, HLA-B, HLA-DRB1, HLA-DPB1

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

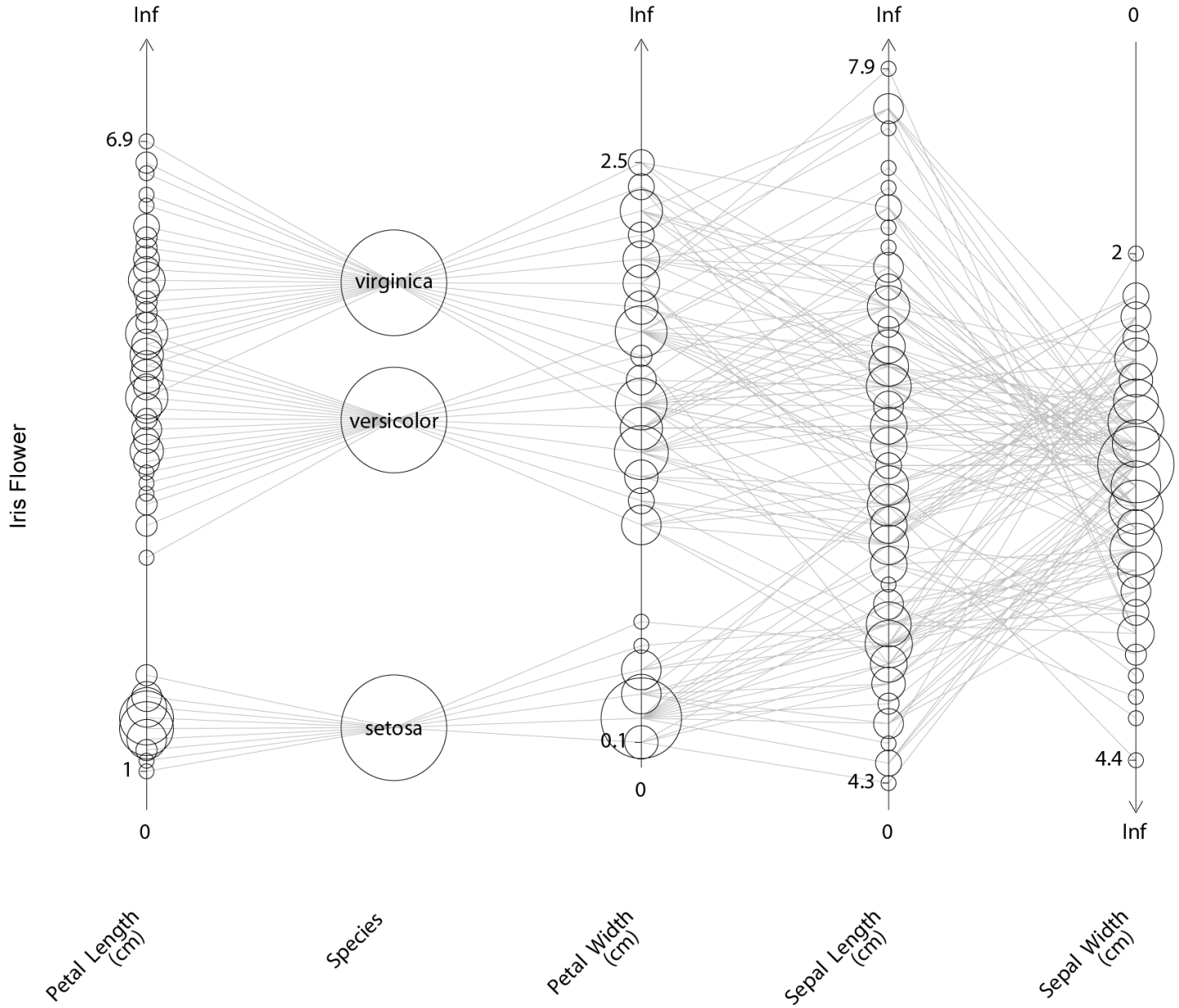
- Numeric, Logical or Categorical

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,


Iris Flower
Inf
6.9
1
0
virginica
versicolor
setosa
Petal Length (cm)
Species
Inf
2.5
0.1
0
Petal Width (cm)
Inf
7.9
4.3
0
Sepal Length (cm)
0
2
4.4
Inf
Sepal Width (cm)

Horizontalisation Criterion

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

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

$$\sum_{j=1}^{p} \| \mathbf{y}_j - \boldsymbol{\xi} \|^2 \underset{\alpha_j, \beta_j, j=1,..,p, \boldsymbol{\xi}}{\longrightarrow} \min$$

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 : horizotal 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

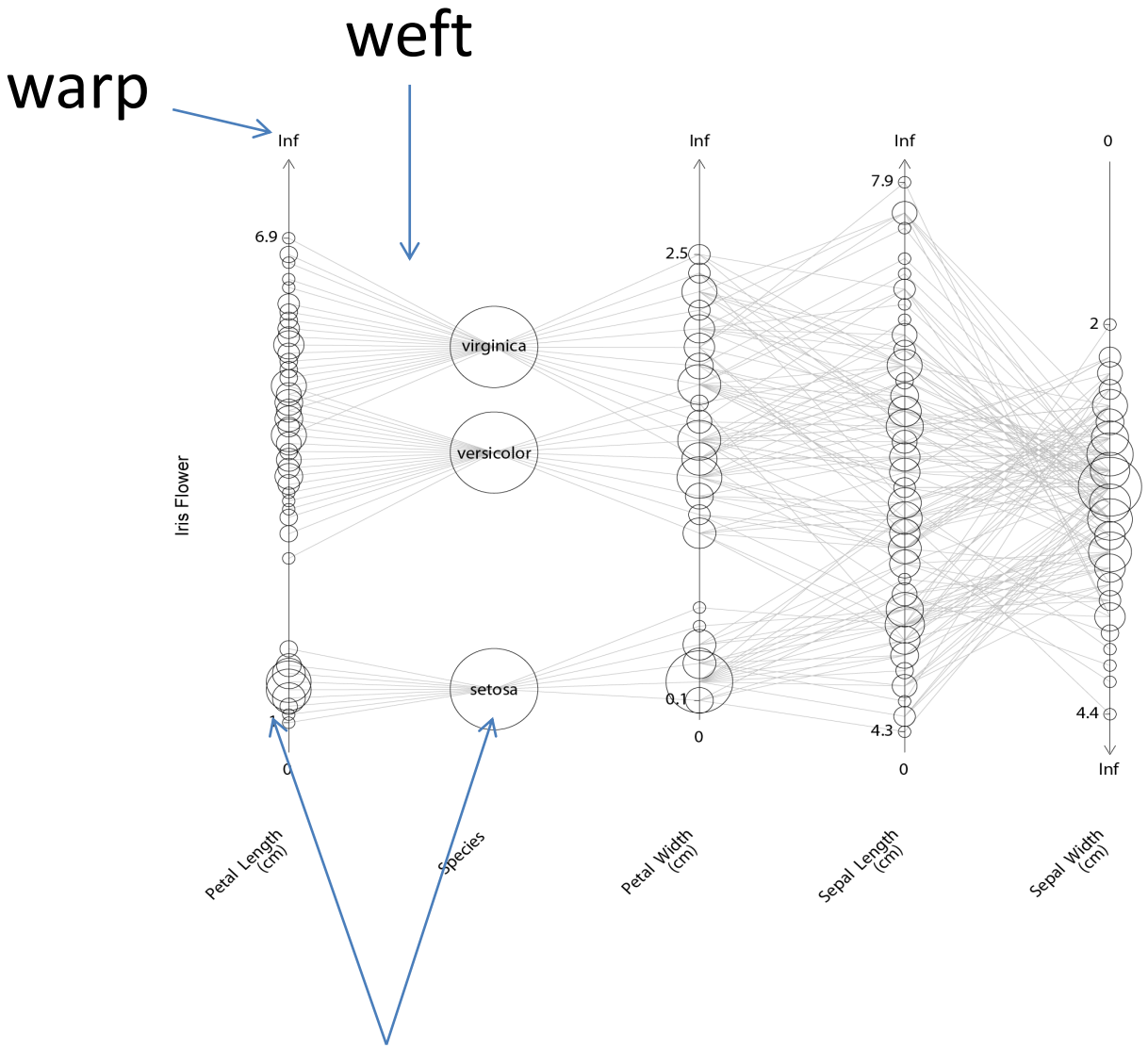
$$\mathbf{y}_j = \alpha_j \mathbf{1} + \boldsymbol{\beta}_j X_j : \text{coordinates on the } j \text{ th axis}$$

Horizontalisation criterion

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

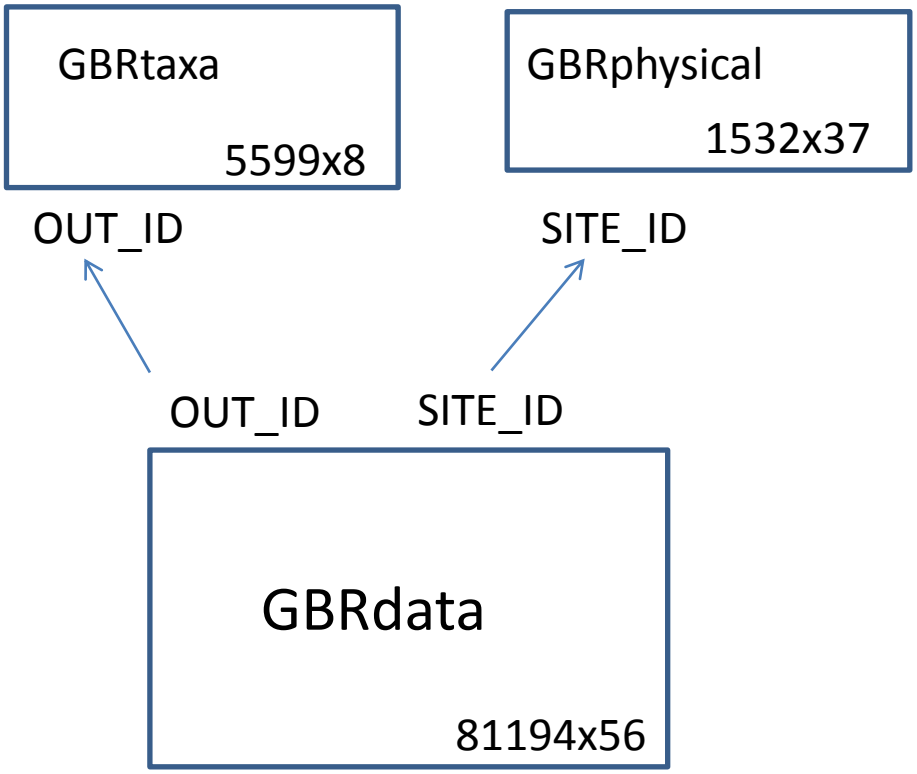
determines *a* proper location of the levels of each category.

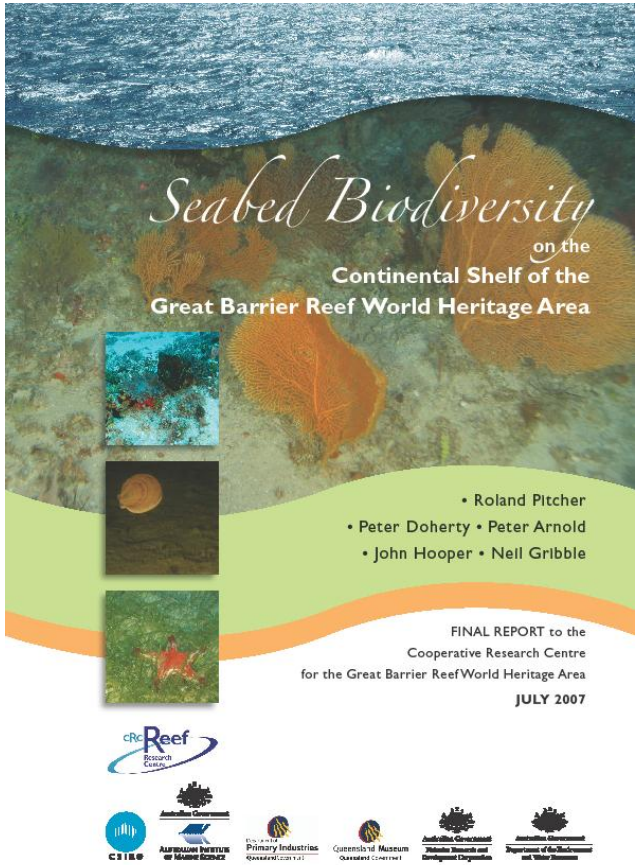
Independent of the choice of contrast



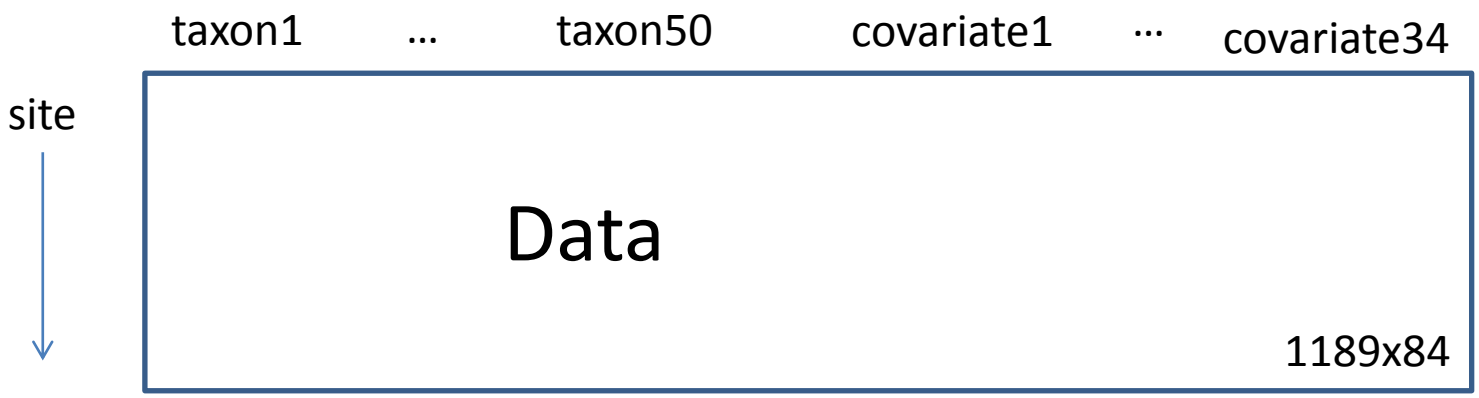
Proportional to the multiplicity of the value

GBR data


GBRtaxa
5599x8
GBRphysical
1532x37
OUT_ID
SITE_ID
OUT_ID
SITE_ID
GBRdata
81194x56


Seabed Biodiversity
on the
Continental Shelf of the
Great Barrier Reef World Heritage Area
• Roland Pitcher
• Peter Doherty • Peter Arnold
• John Hooper • Neil Gribble
FINAL REPORT to the
Cooperative Research Centre
for the Great Barrier Reef World Heritage Area
JULY 2007

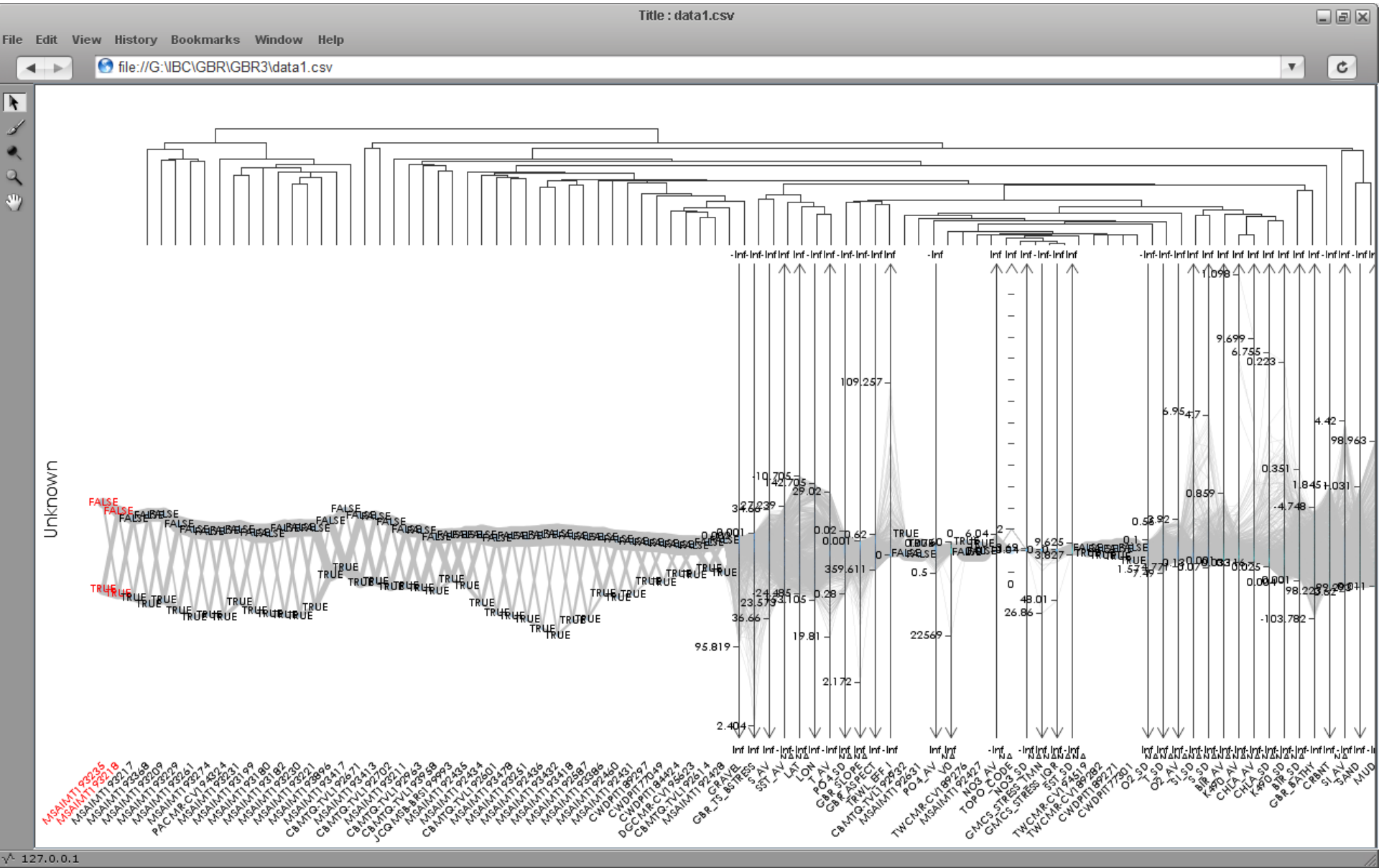
Most frequent 50 taxa



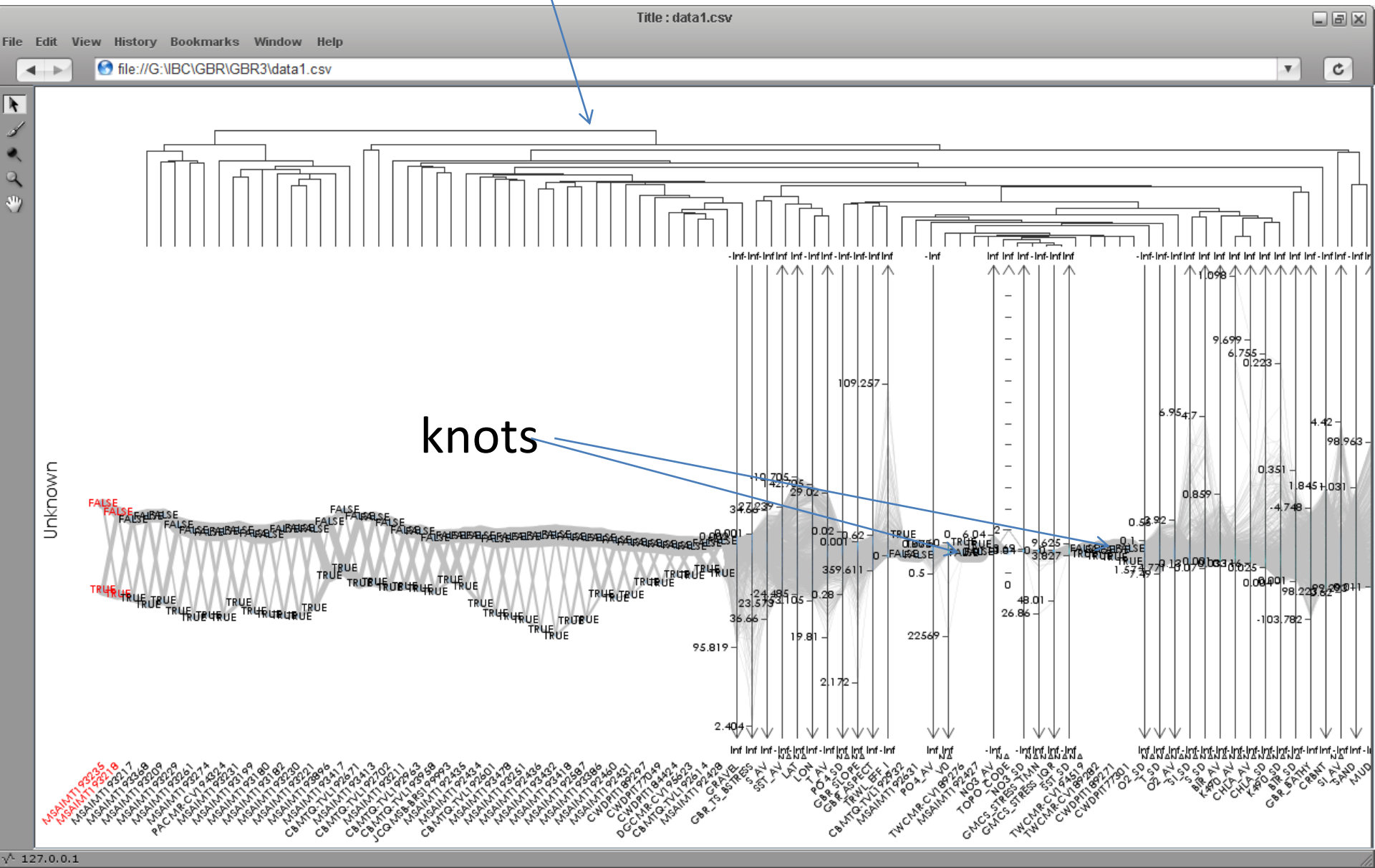
taxon1 … taxon50 : logical
covariate1 … covariate34: numeric

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)



Hierarchical Clustering of Axes (Variables)



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

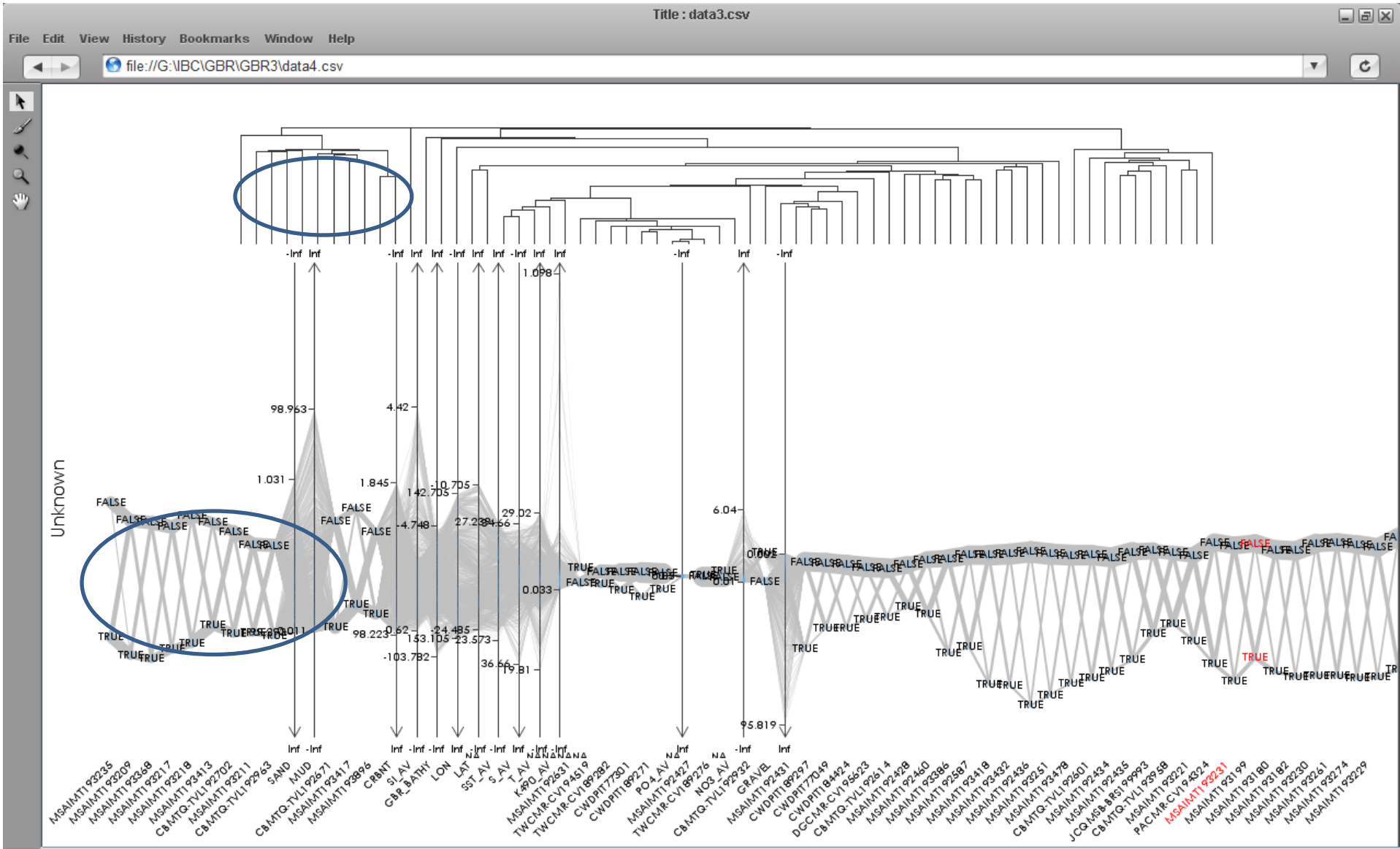
Covariates Orthogonal to Existence of Taxa

Standard deviations and other covvariates (20 covariates)

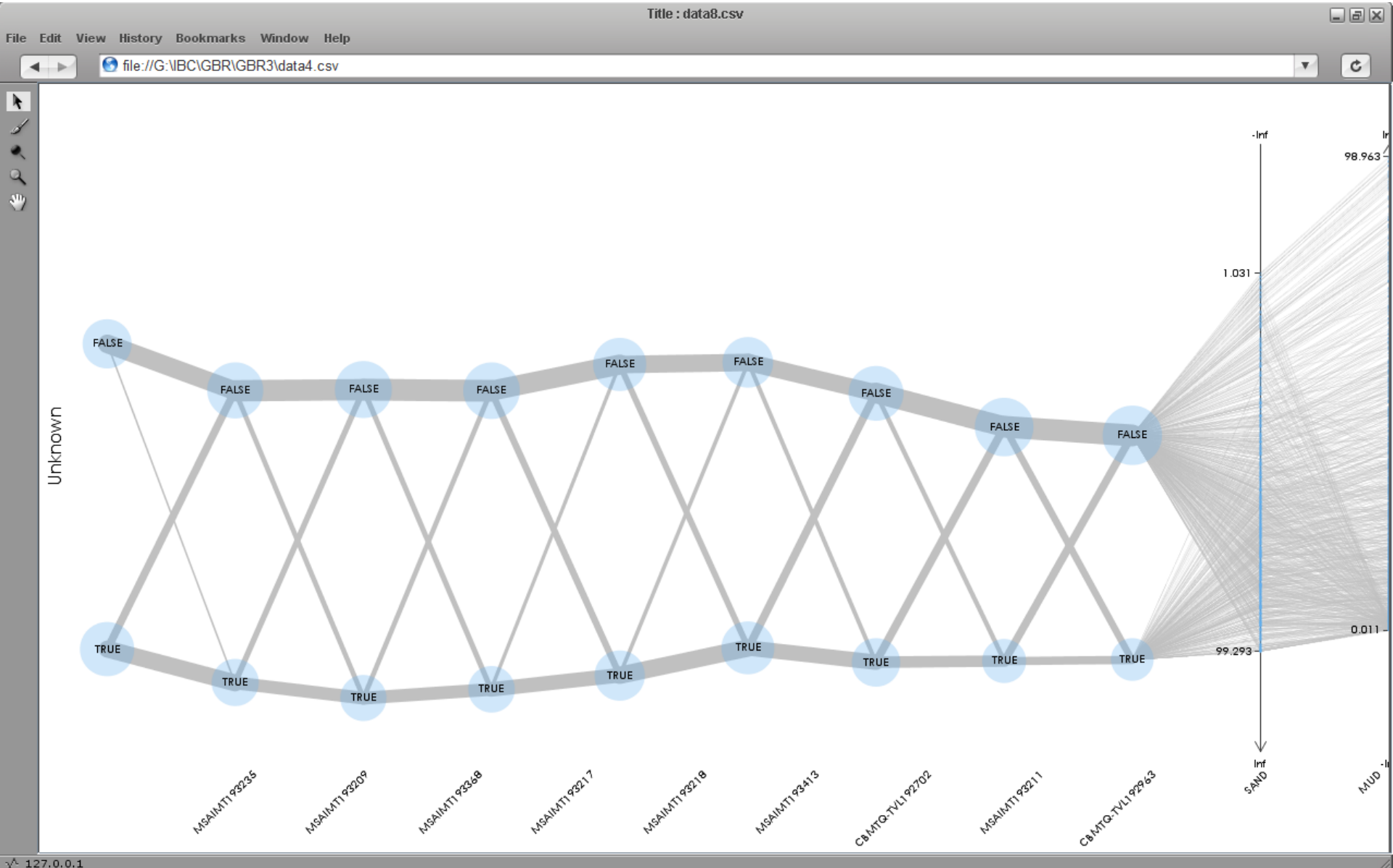
[1] "GBR_ASPECT" "GBR_SLOPE" "GBR_TS_BSTRESS" "GMCS_STRESS_TMN"
[5] "GMCS_STRESS_IQR" "NO3_SD" "PO4_SD" "O2_AV"
[9] "O2_SD" "S_SD" "T_SD" "SI_SD"
[13] "CHLA_AV" "CHLA_SD" "K490_SD" "SST_SD"
[17] "BIR_AV" "BIR_SD" "TRWL_EFF_I" "TOPO_CODE"

Delete

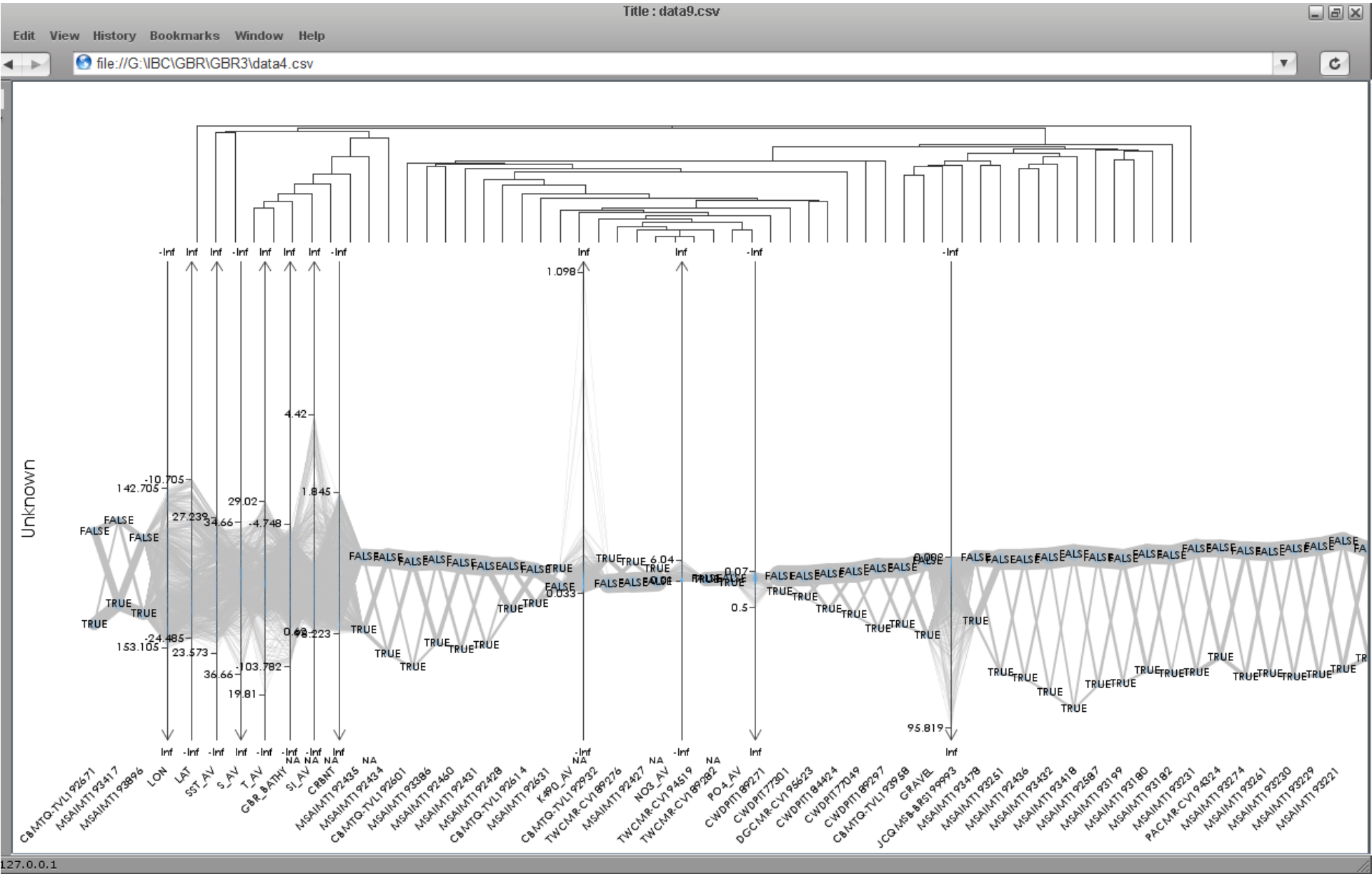
14 Covariates



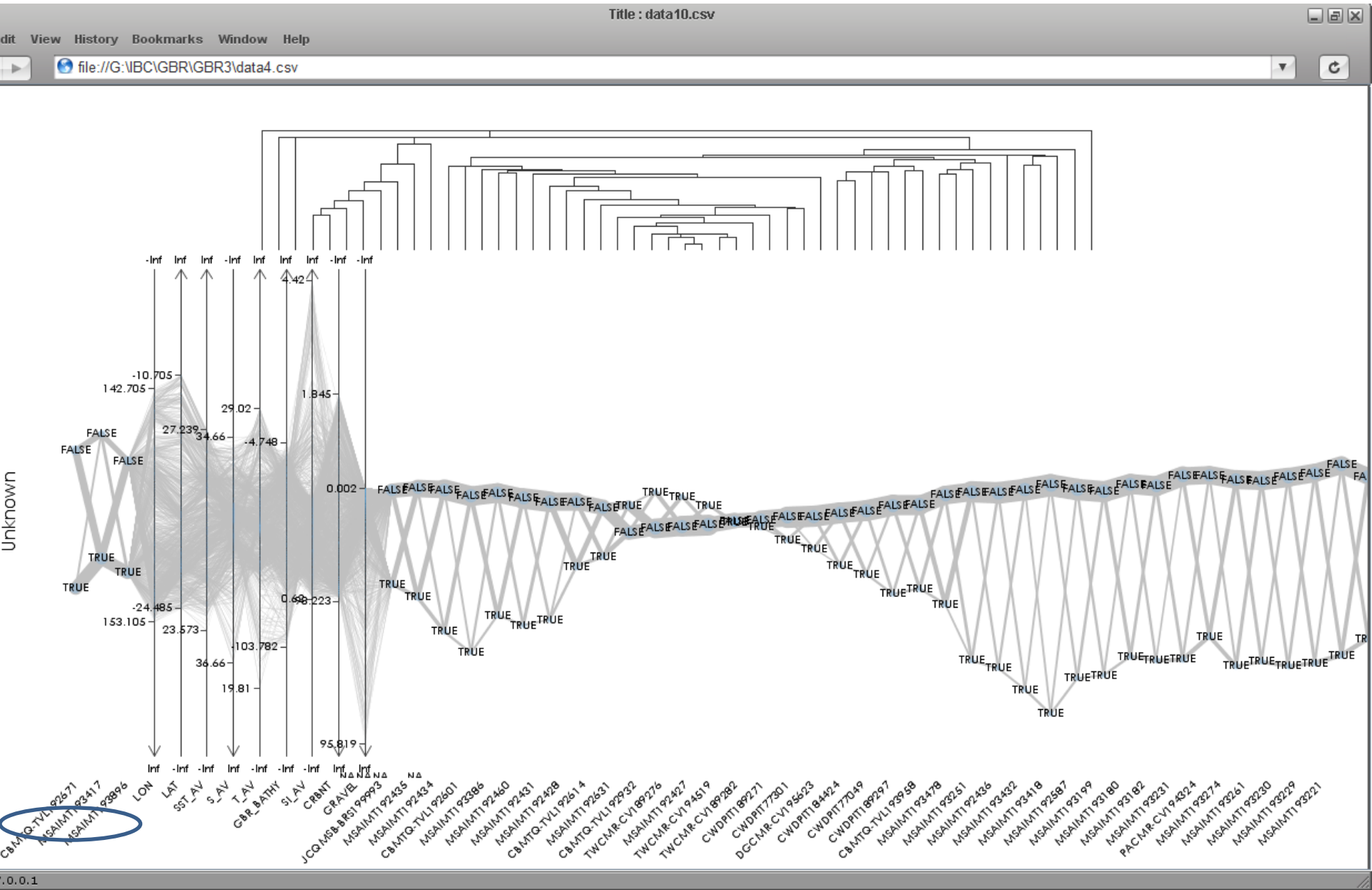
Group1: Sand appetite 9 Taxa



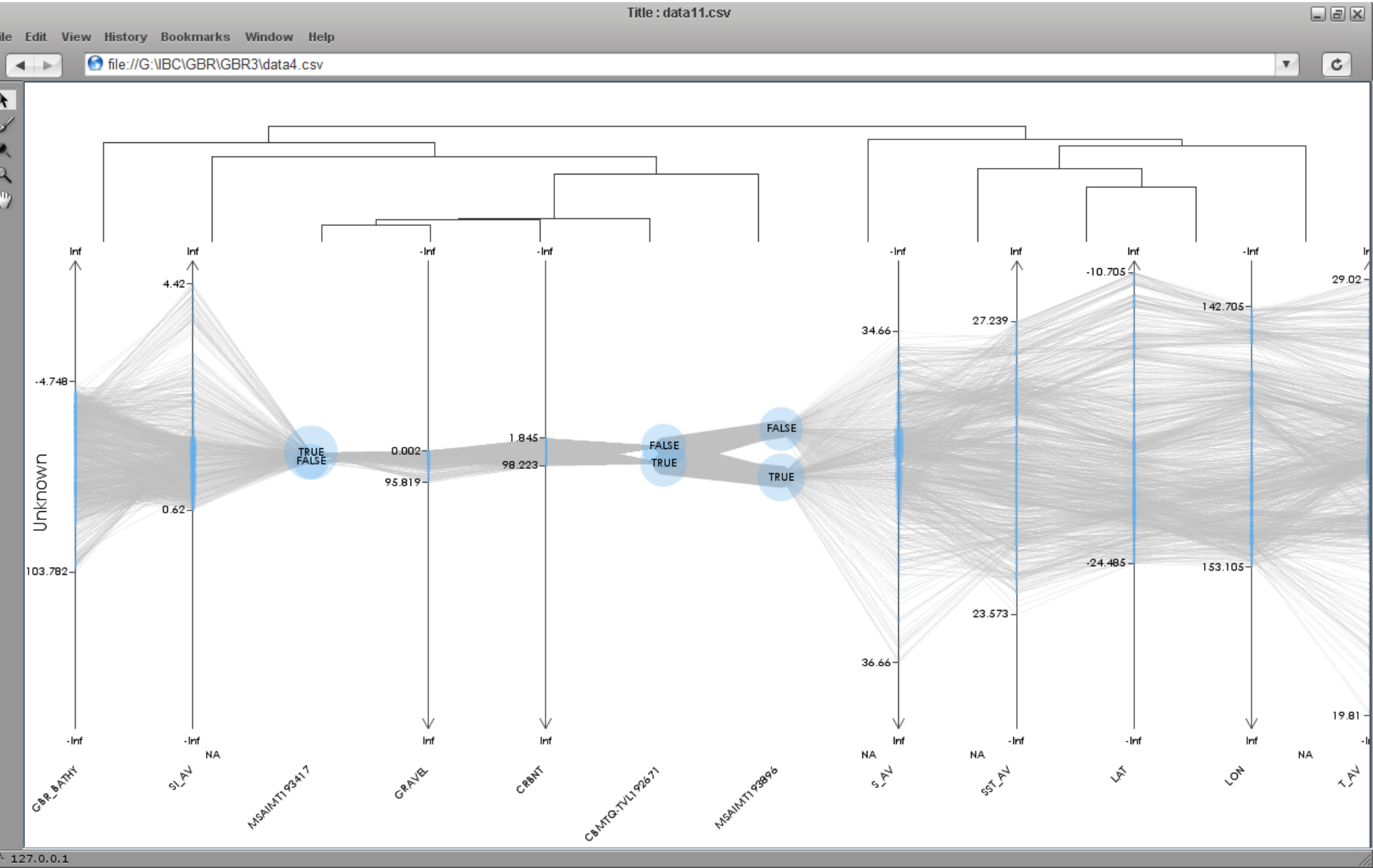
Remaining 41 Taxa: Knots: NO3_AV, PO4_AV and K490_AV



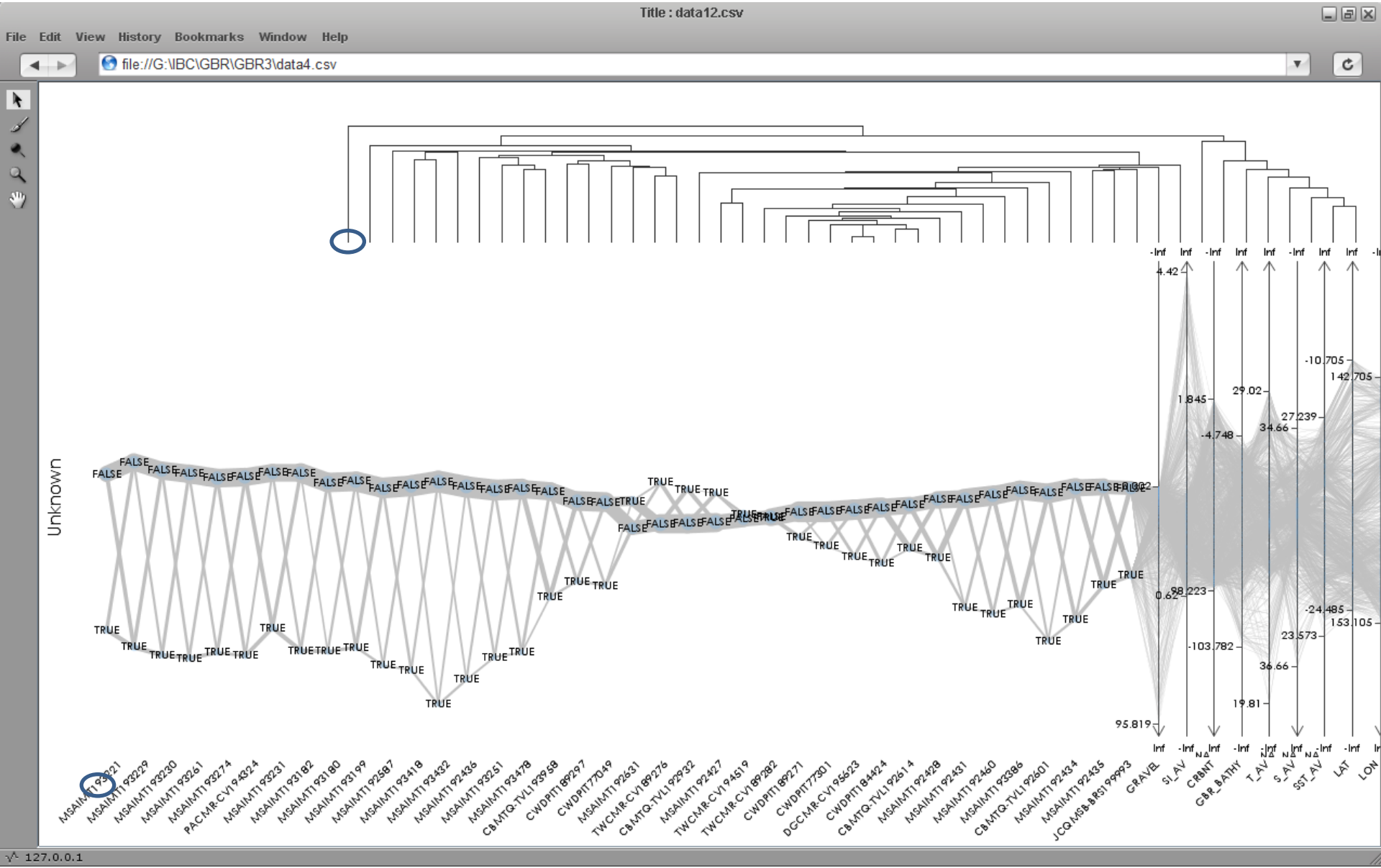
After 3 covariates removed



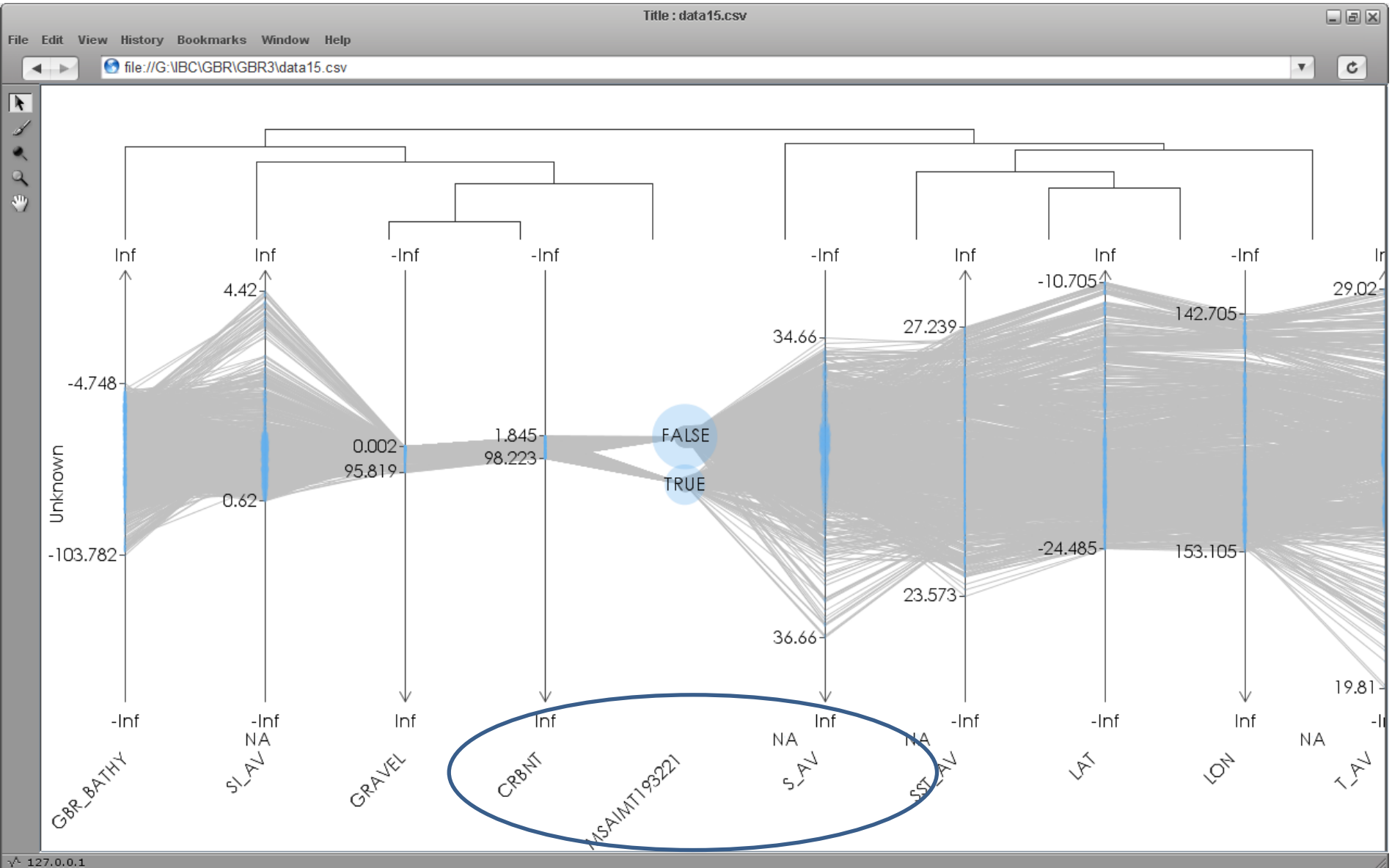
Group2: Most frequent 3 taxa



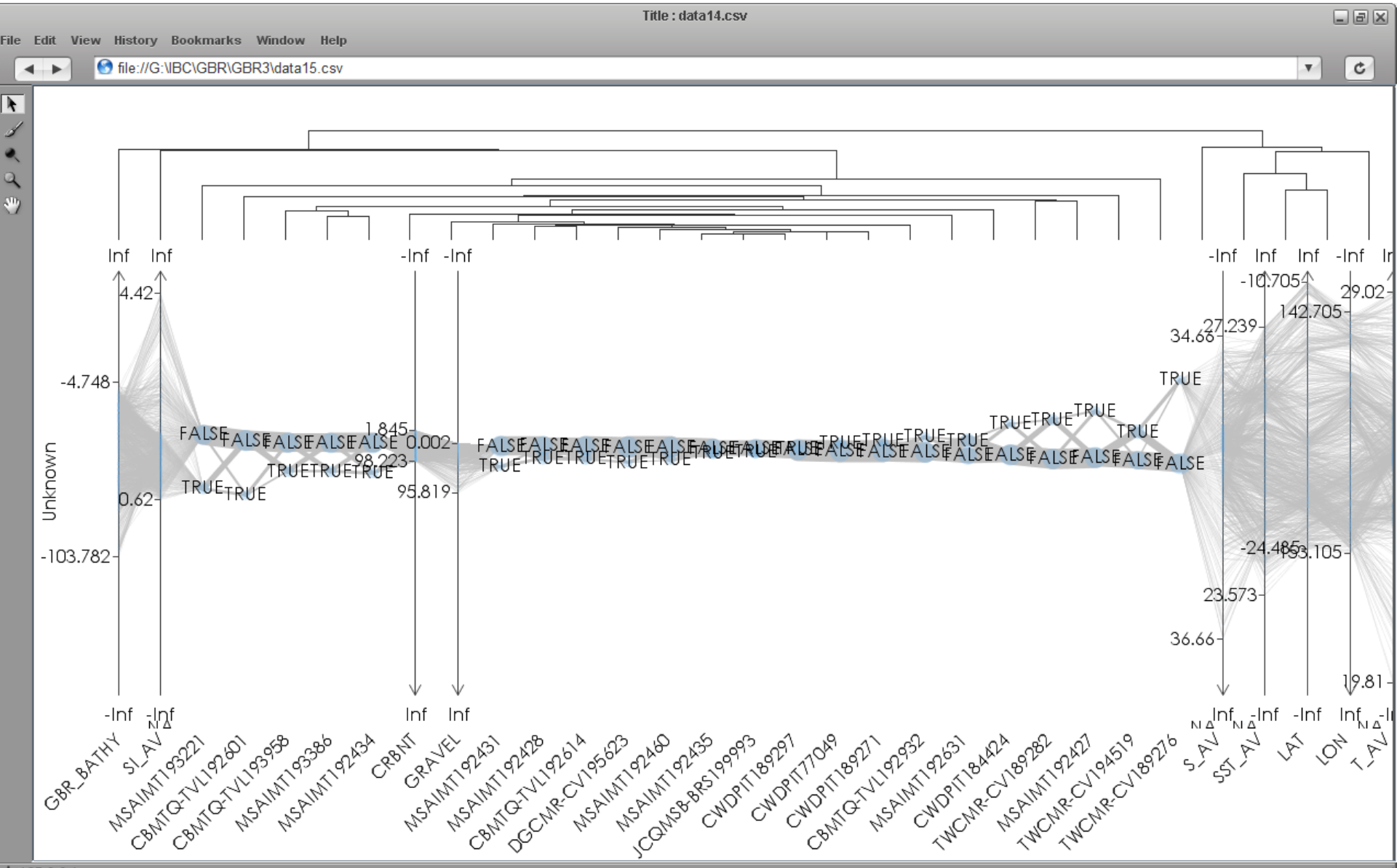
Remaining 38 taxa



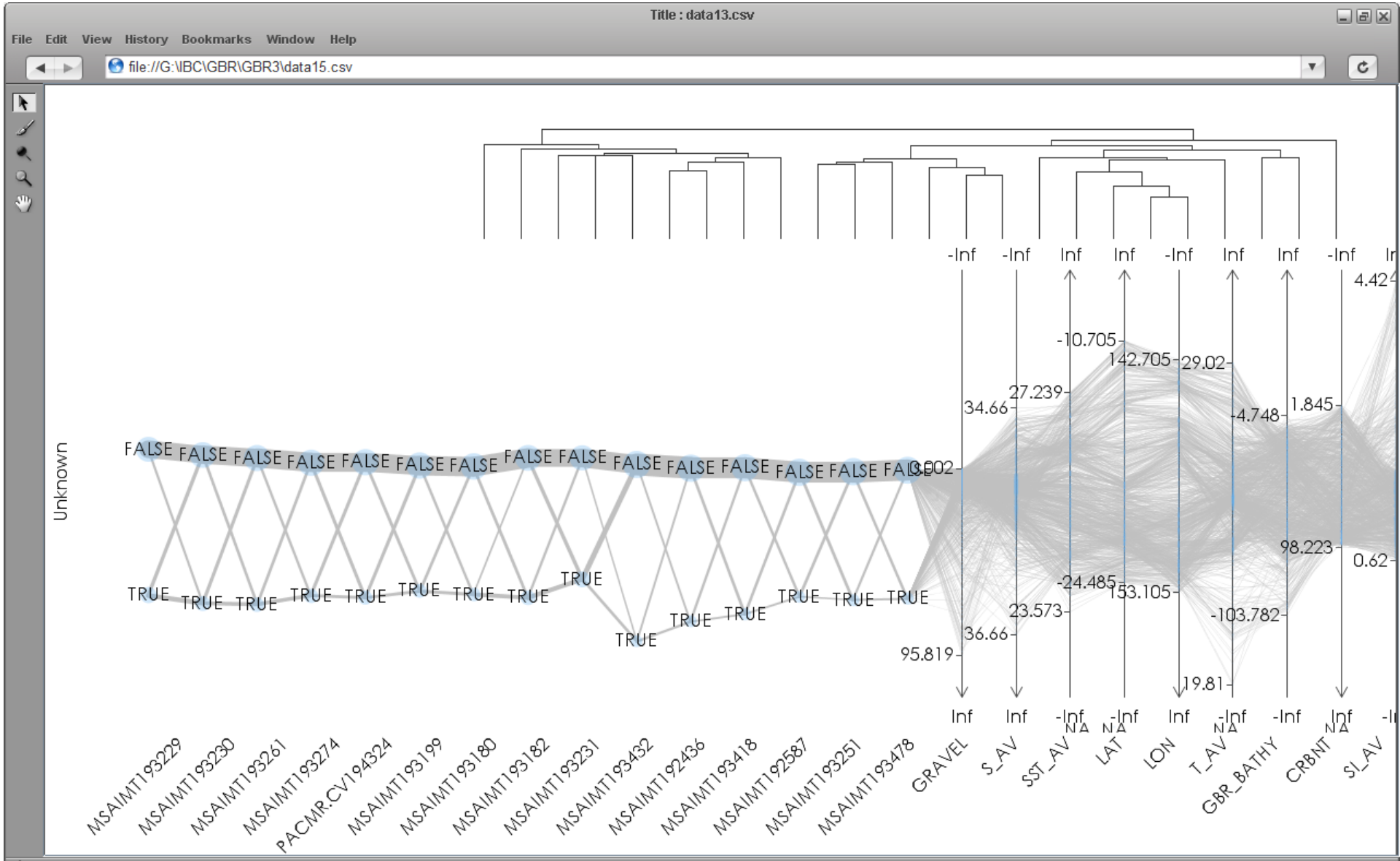
Group3: Unique taxon



Group4:22 taxa

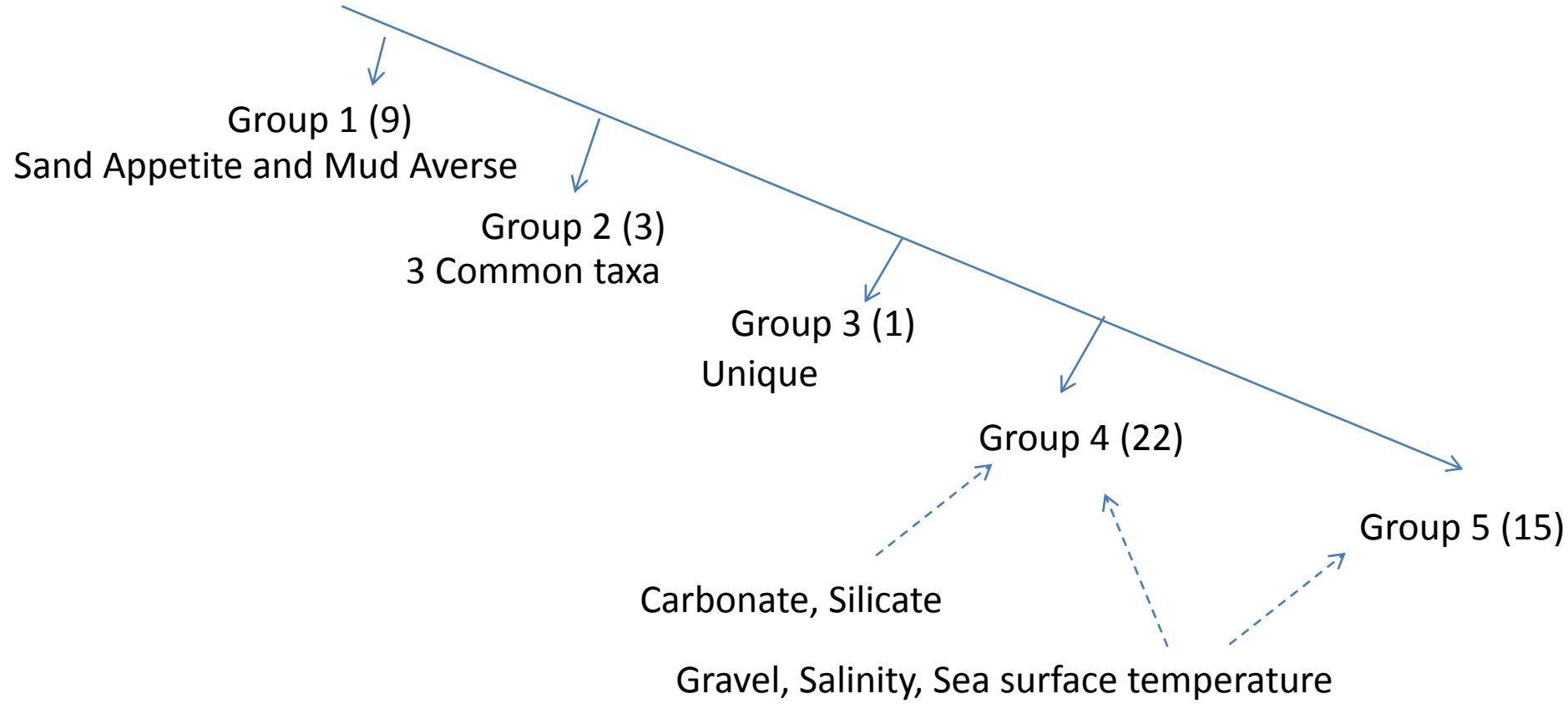


Group5: 15 taxa



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