R package **gcExplorer**: graphical and inferential exploration of cluster solutions

Theresa Scharl[1,2] Friedrich Leisch[3]

[1] Institut für Statistik und Wahrscheinlichkeitstheorie
Technische Universität Wien

[2] Department of Biotechnology
University of Natural Resources and Applied Life Sciences,Vienna

[3] Institut für Statistik
Ludwig-Maximilians-Universität München

UseR! 2009, July 8th, Rennes

Outline

1. Motivation
2. Cluster Analysis
3. Neighborhood Graphs
4. Software
5. Inference

Motivation

Exploration and visualization of cluster solutions

- Interpretation of cluster results.
- Understanding of the cluster structure.
- Relationships between segments of a partition.

Inference for gene cluster graphs

- Explore the quality of a cluster solution.
- External validation of clustering.
- Association to a functional group.

E. coli data

Recombinant *E. coli* process

- Evaluate the influence of the induction level of N^{pro}GFPmut3.1 an inclusion body forming protein on host metabolism
- Non-induced state was compared to samples past induction

Oxygen data (Covert et al., 2004)

- Investigation of various mutants under oxygen deprivation
- Target the a priori most relevant part of the transcriptional netwok
- Use six strains with knockouts of key transcriptional regulators in the oxygen response.

Cluster algorithms

Partitioning cluster algorithms

Cluster algorithms like K–means and PAM or others where clusters can be represented by centroids (e.g., QT–Clust, Heyer et al., Genome Research, 1999).

R package **flexclust**

- Flexible toolbox to investigate the influence of distance measures and cluster algorithms.
- Extensible implementations of the generalized k–Means and QT–Clust algorithm.
- Possibility to try out a variety of distance or similarity measures.
- Cluster algorithms are treated separately from distance measures.

TRNs and silhouette plots

Topology–representing networks

(Martinetz and Schulten, 1994)

- Count the number of data points a pair of centroids is closest and second–closest.
- Centroid pairs with a positive count are connected.

Silhouette plots (Rousseeuw, 1987)

- Compare the distance from each point to the points in its own cluster to the distance to points in the second closest cluster.
- The larger the silhouette values the better a cluster is separated from the other clusters.

Neighborhood graphs

(Leisch, 2006)

- Neighborhood graphs use mean relative distances as edge weights.
- Assume we are given a data set $X_N = \{x_1, \ldots, x_N\}$ and
- a set of centroids $C_K = \{c_1, \ldots, c_K\}$.
- The centroid closest to x is denoted by

$$c(x) = \operatorname*{argmin}_{c \in C_K} d(x, c).$$

- And the second closest centroid to x is denoted by

$$\tilde{c}(x) = \operatorname*{argmin}_{c \in C_K \setminus \{c(x)\}} d(x, c).$$

Neighborhood graphs

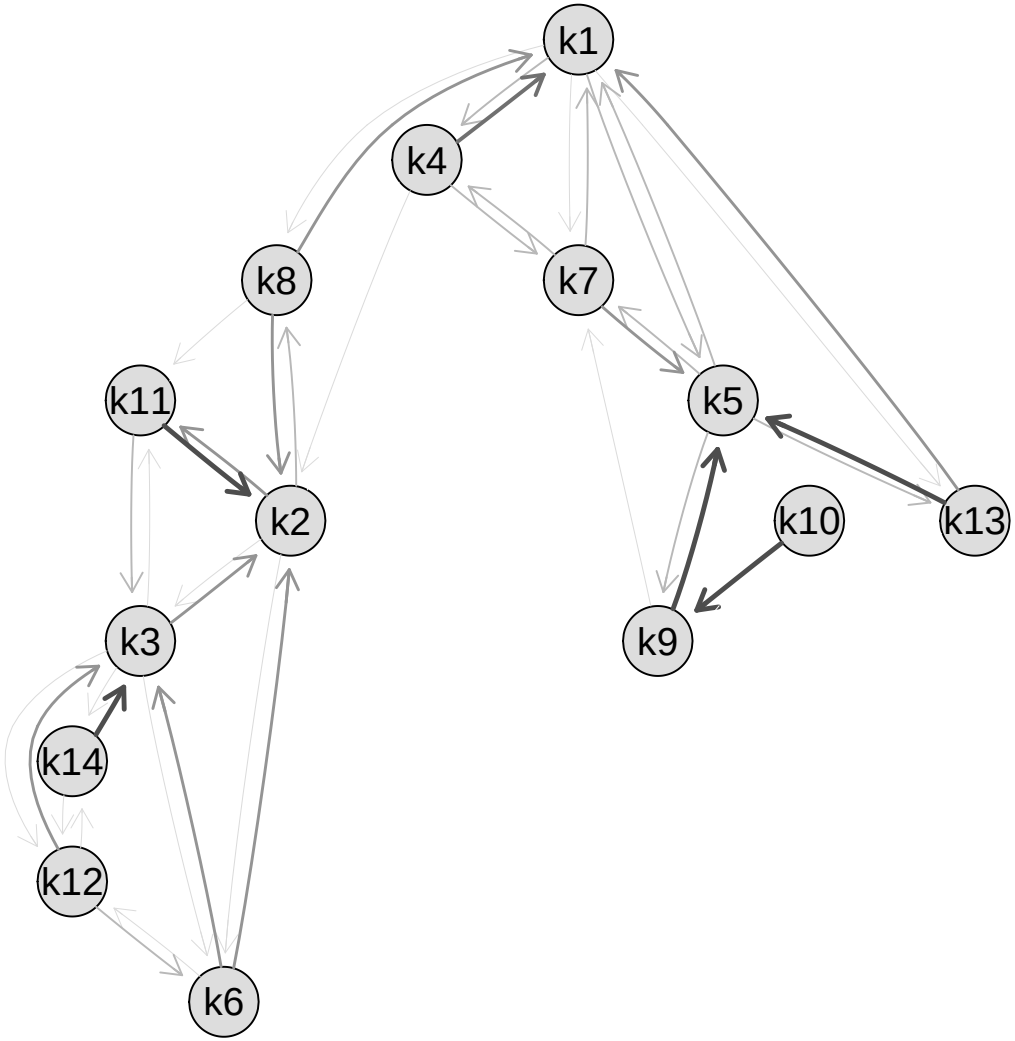
- The set of all points where c_i is the closest centroid and c_j is second–closest is given by

$$A_{ij} = \{x_n | c(x_n) = c_i, \tilde{c}(x_n) = c_j\}.$$

- Now we define edge weights

$$s_{ij} = \begin{cases} |A_i|^{-1} \sum_{x \in A_{ij}} \frac{2d(x,c(x))}{d(x,c(x))+d(x,\tilde{c}(x))}, & A_{ij} \neq \emptyset \\ 0, & A_{ij} = \emptyset \end{cases}$$

Neighborhood graphs



R package **gcExplorer**

An interactive visualization toolbox for clusters (Scharl and Leisch, 2009)

- New visualization techniques to display cluster results of high dimensional data.
- Nonlinear arrangements of the cluster centroids using Bioconductor packages **Rgraphviz** and **graph**
- Interactive exploration using arbitrary panel functions.
- Visualize properties of clusters using arbitray node functions.
- Allow small glyphs for the representation of nodes.
- Inference for gene cluster graphs

http://cran.r-project.org/package=gcExplorer.

How to use **gcExplorer**

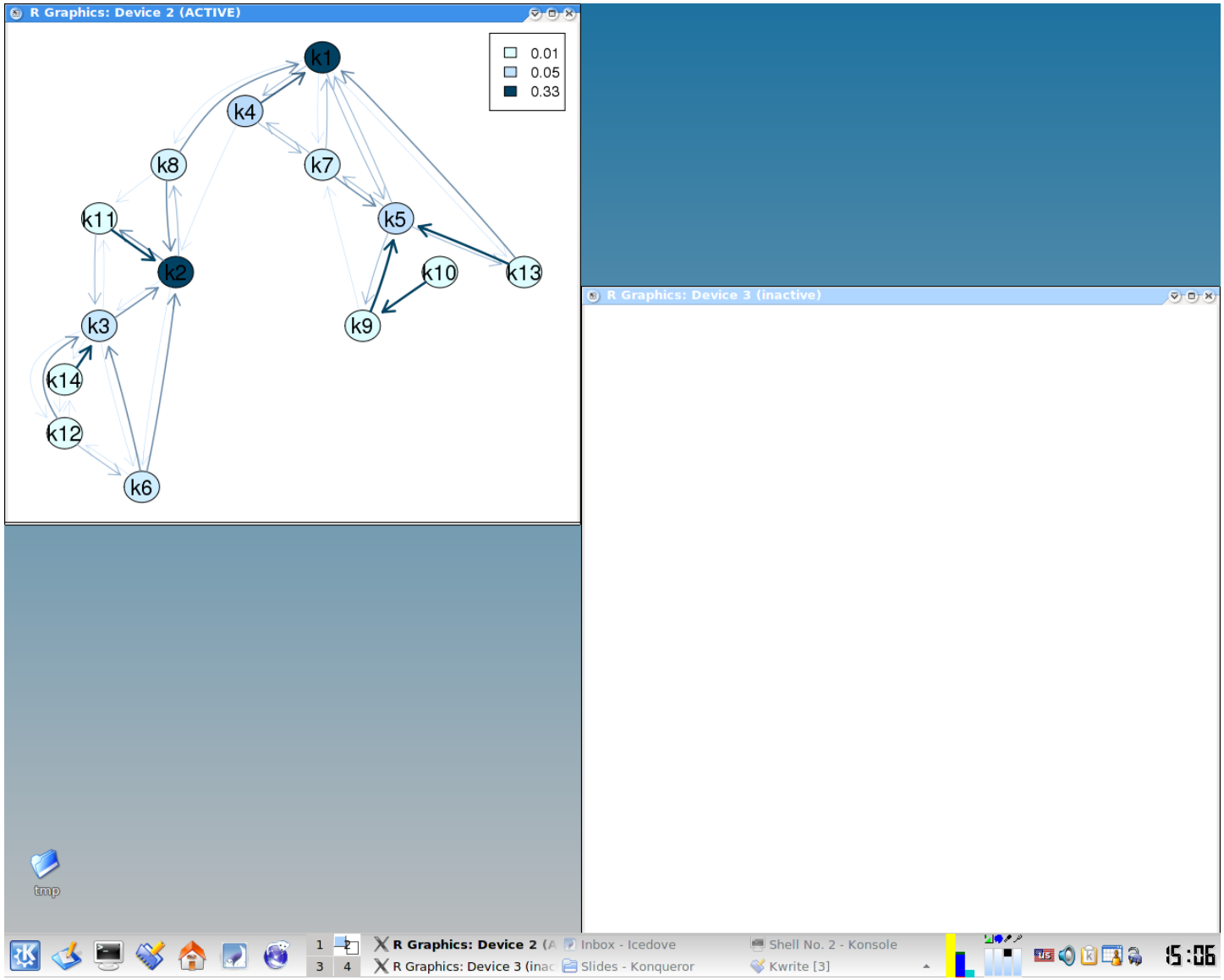
Cluster analysis

```
R> library("gcExplorer")
R> data("ps19")
R> set.seed(1111)
R> cl1 <- qtclust(ps19, radius = 2,
+                 save.data = TRUE)
```

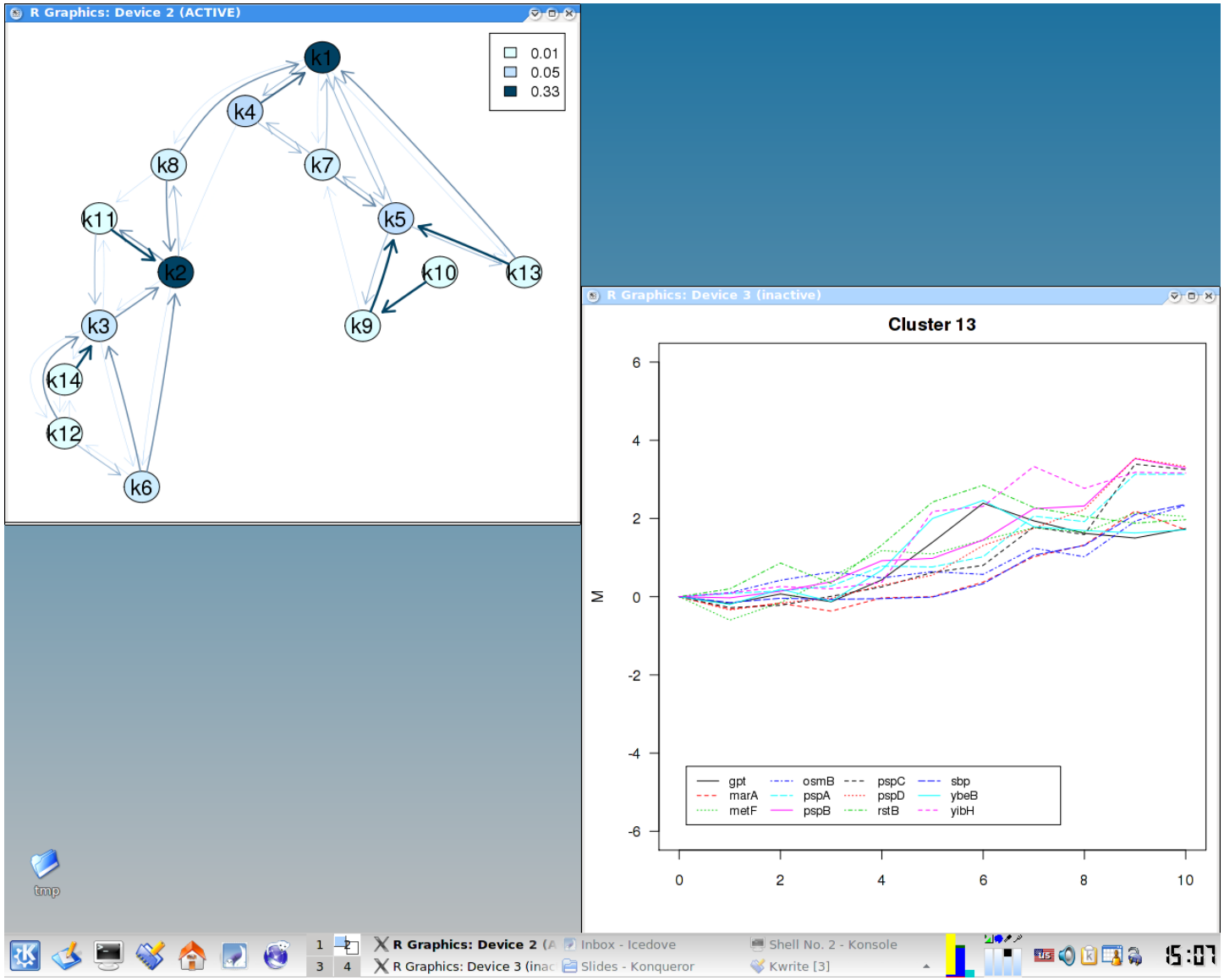
Interactive **gcExplorer**

```
R> gcExplorer(cl1, theme = "blue",
+     panel.function = gcProfile,
+     node.function = node.size)
```

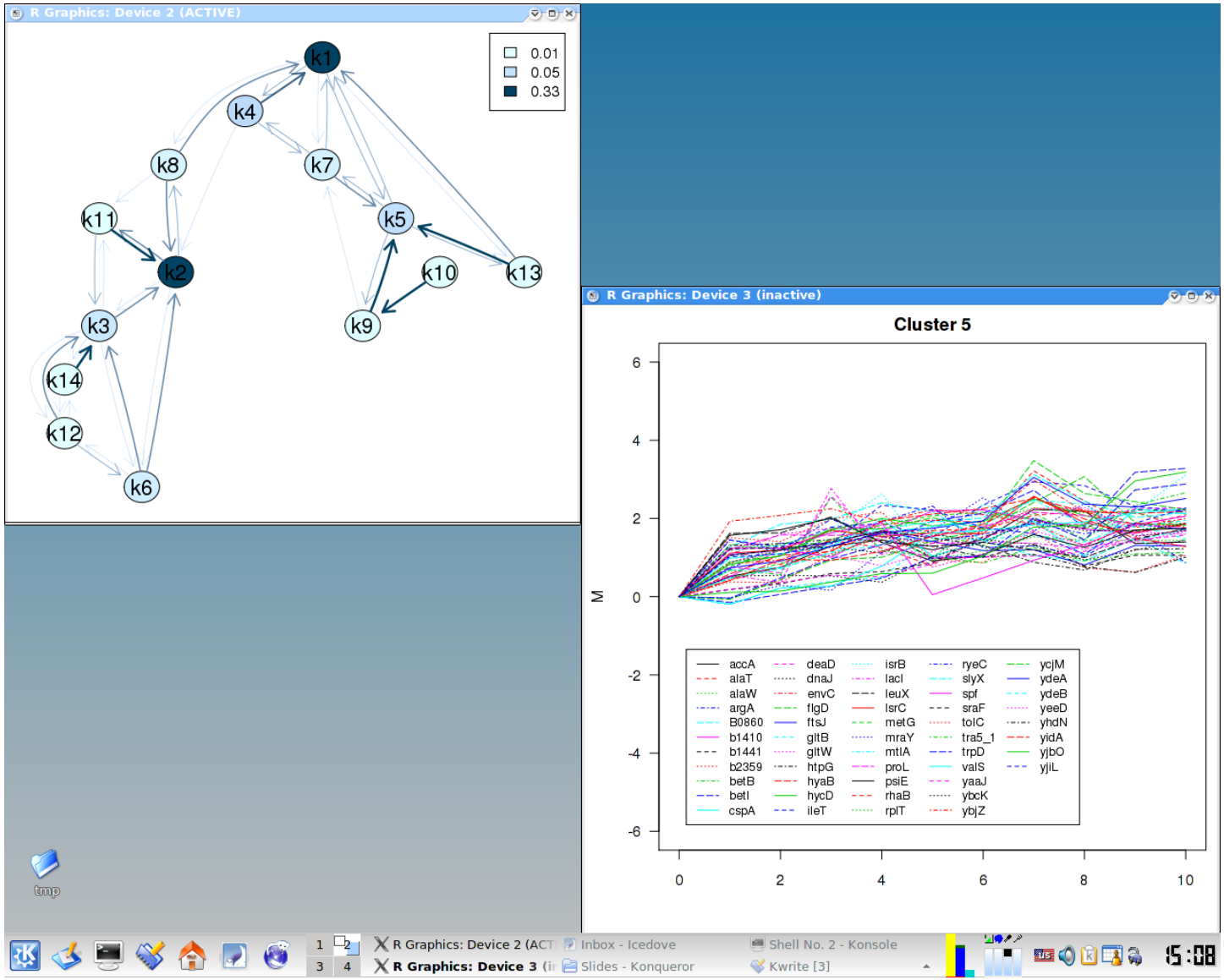
Interactive **gcExplorer**



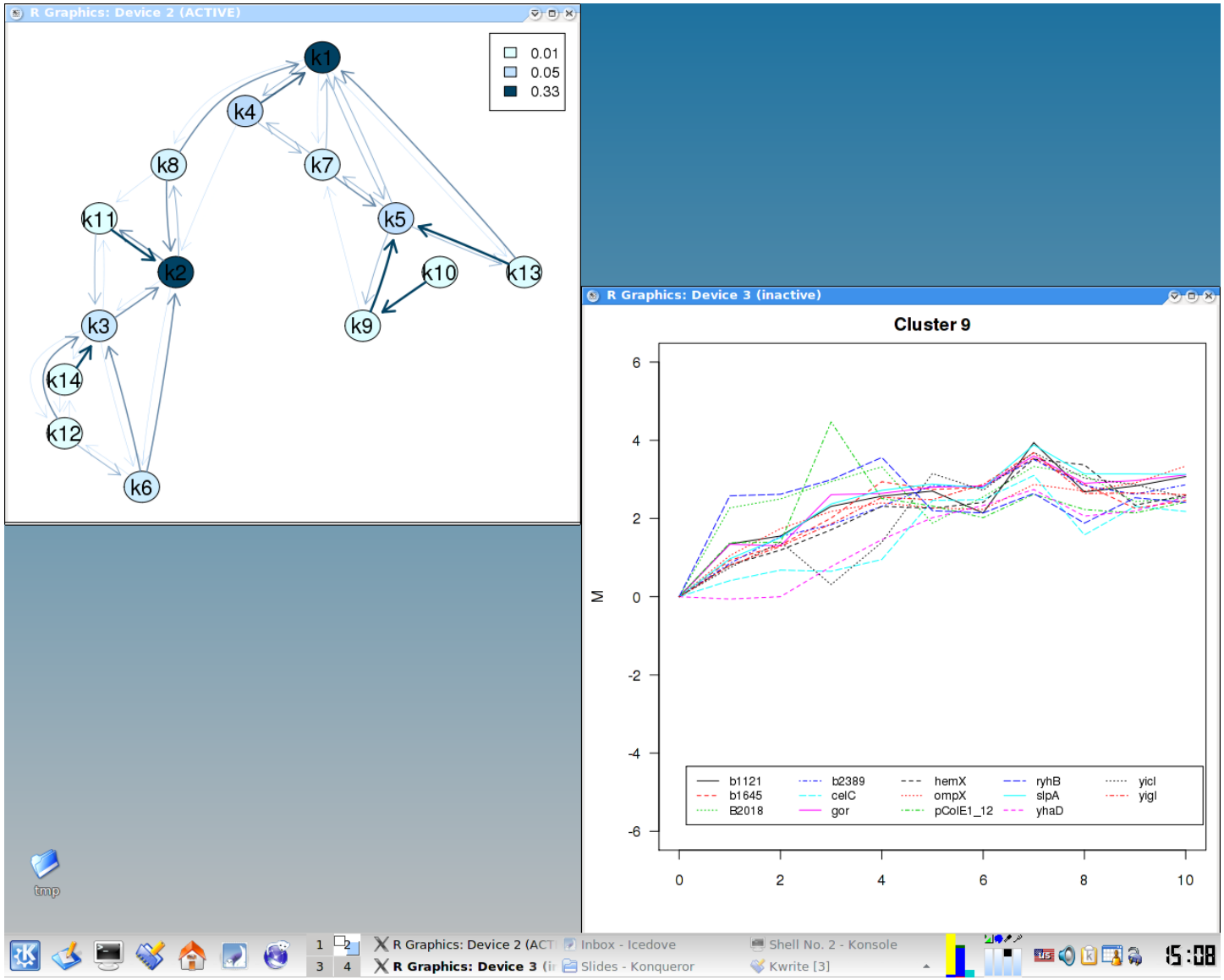
Interactive **gcExplorer**



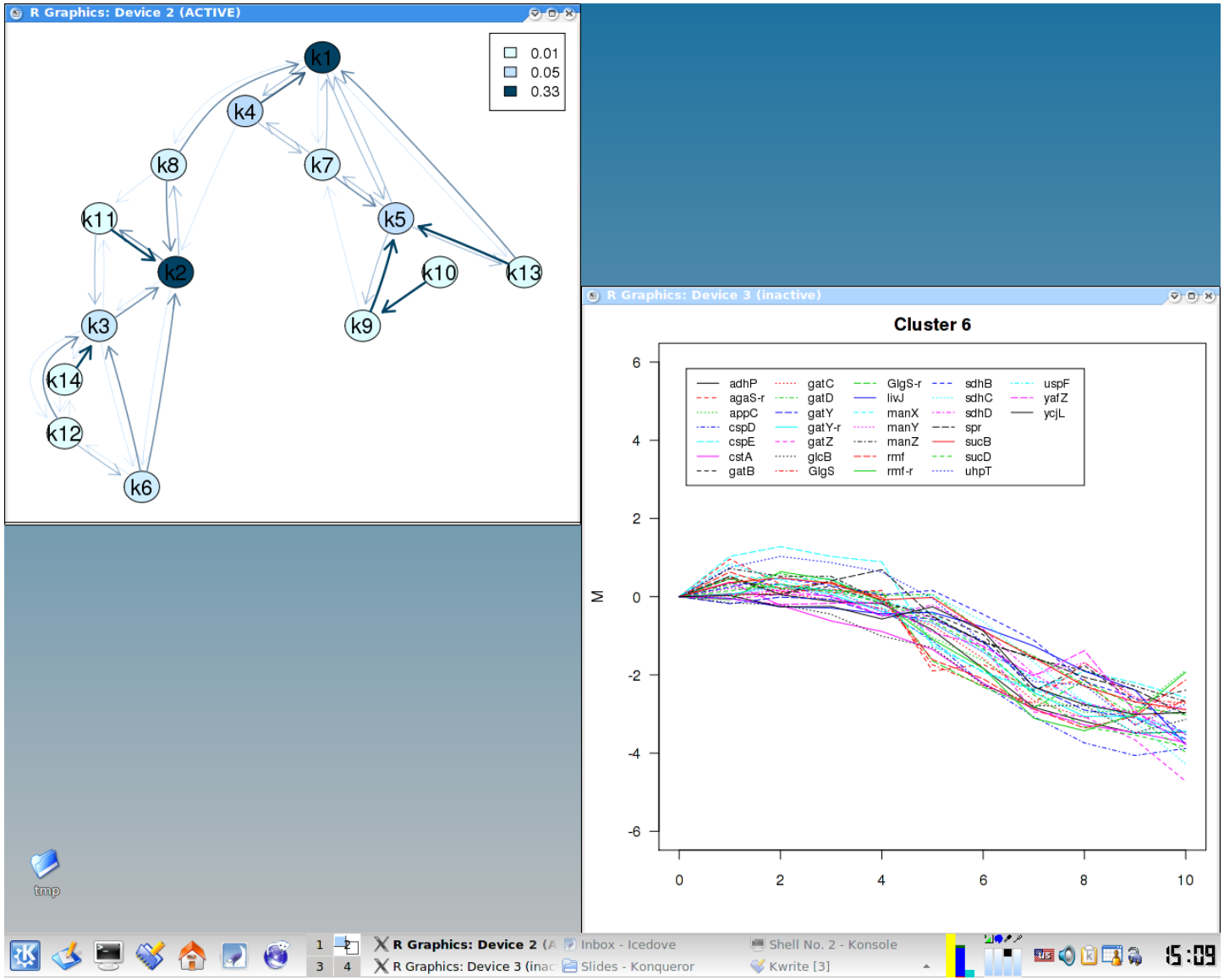
Interactive **gcExplorer**



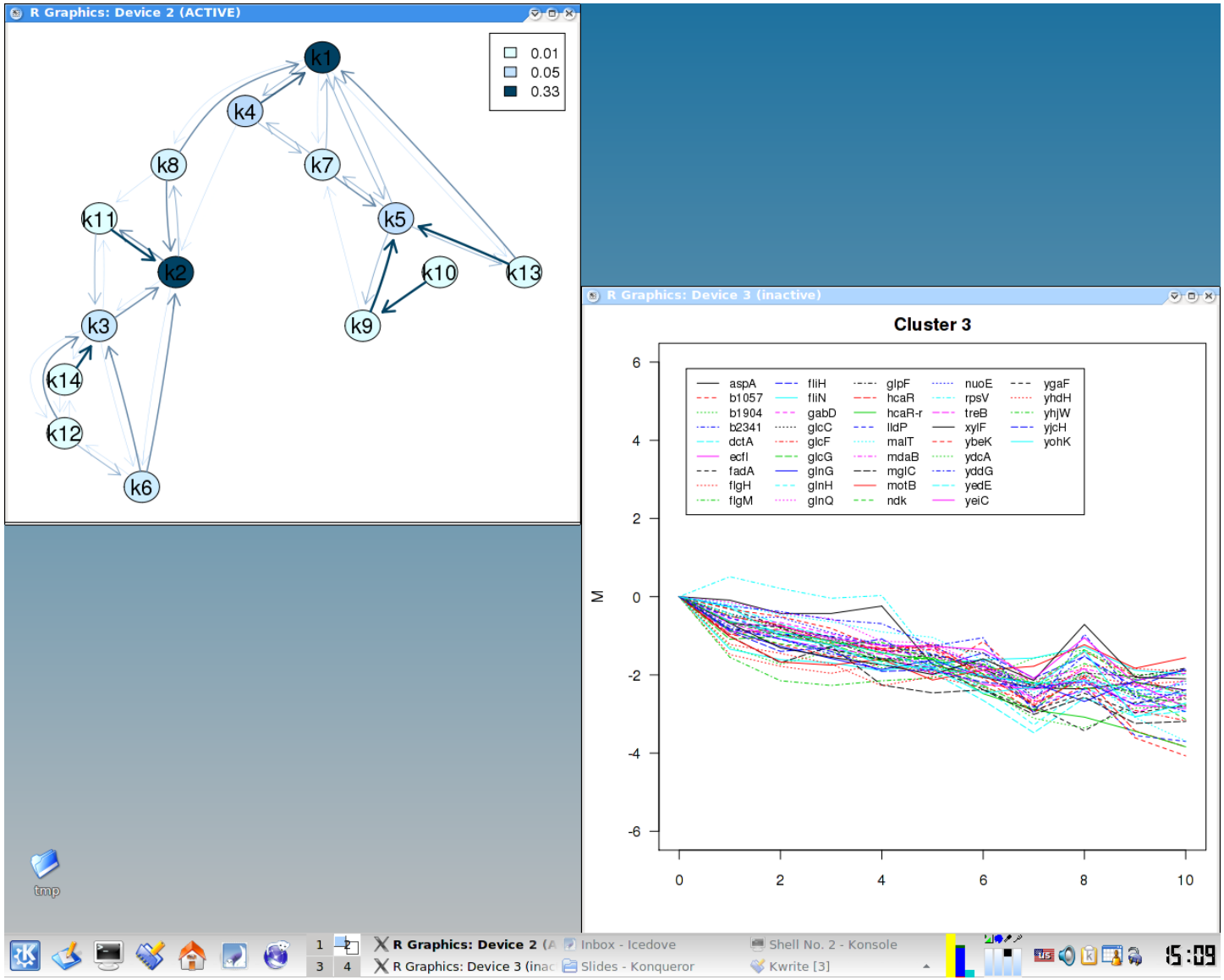
Interactive **gcExplorer**



Interactive **gcExplorer**



Interactive **gcExplorer**

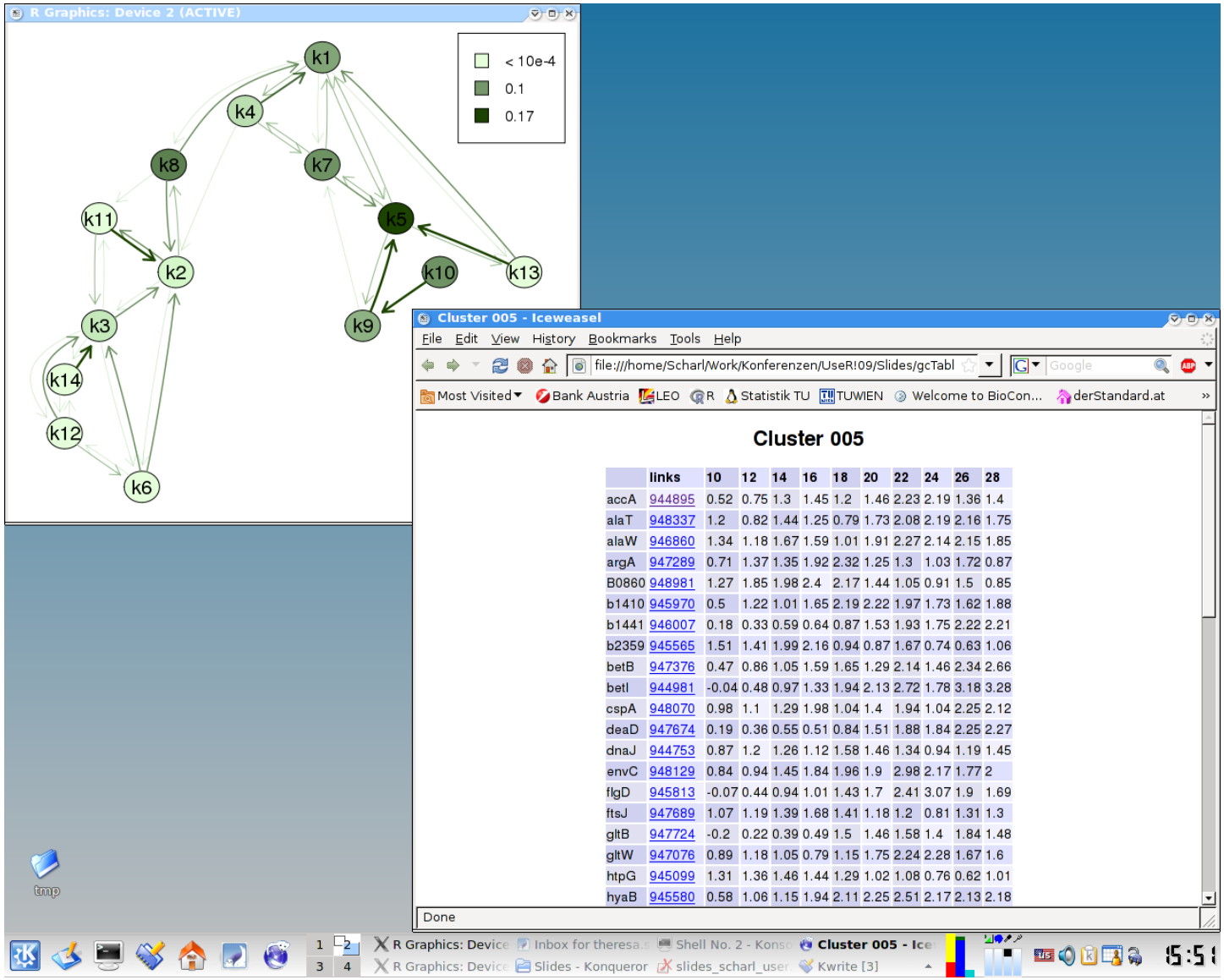


How to use **gcExplorer**

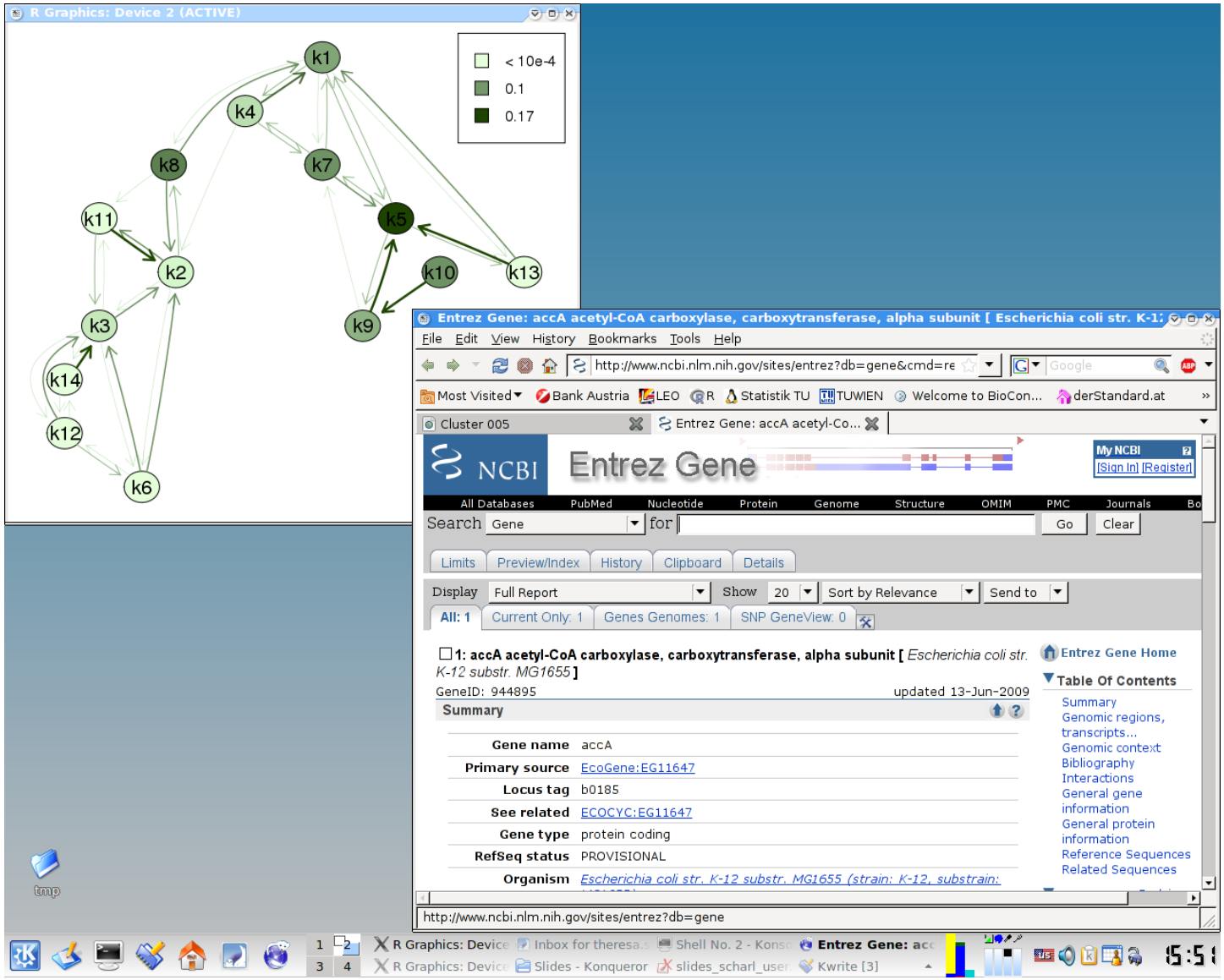
Panel function and node function

```
R> data("sigma")
R> gcExplorer(cl1, theme = "green",
+       panel.function = gcTable,
+       panel.args = list(links = links_ps19),
+       node.function = node.go,
+       node.args = list(gonr = "Sigma32",
+                        id = bn_ps19))
```

Panel and node function



Panel and node function



How to use **gcExplorer**

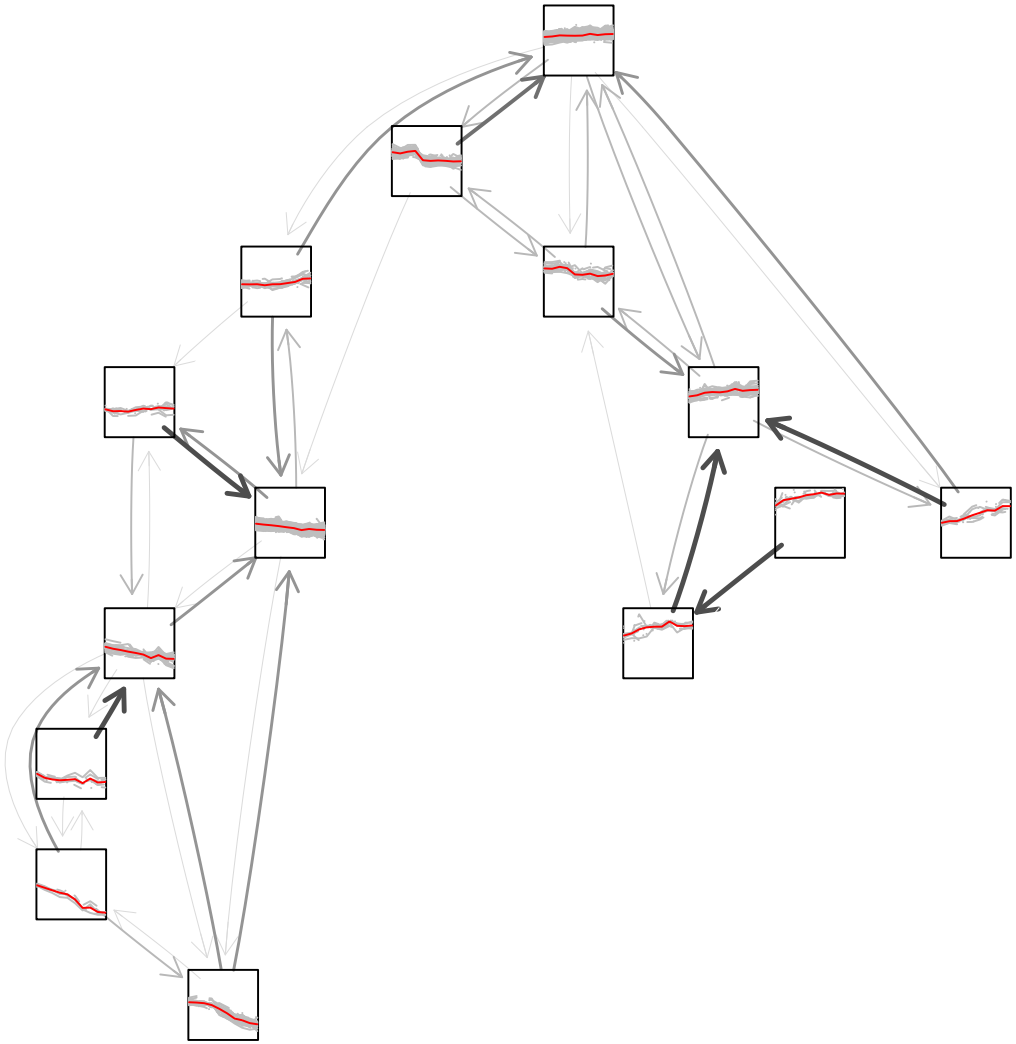
Use of matrix plot as node function

```
R> gcExplorer(cl1, node.function = gmatplot,
+               doViewPort = TRUE)
```

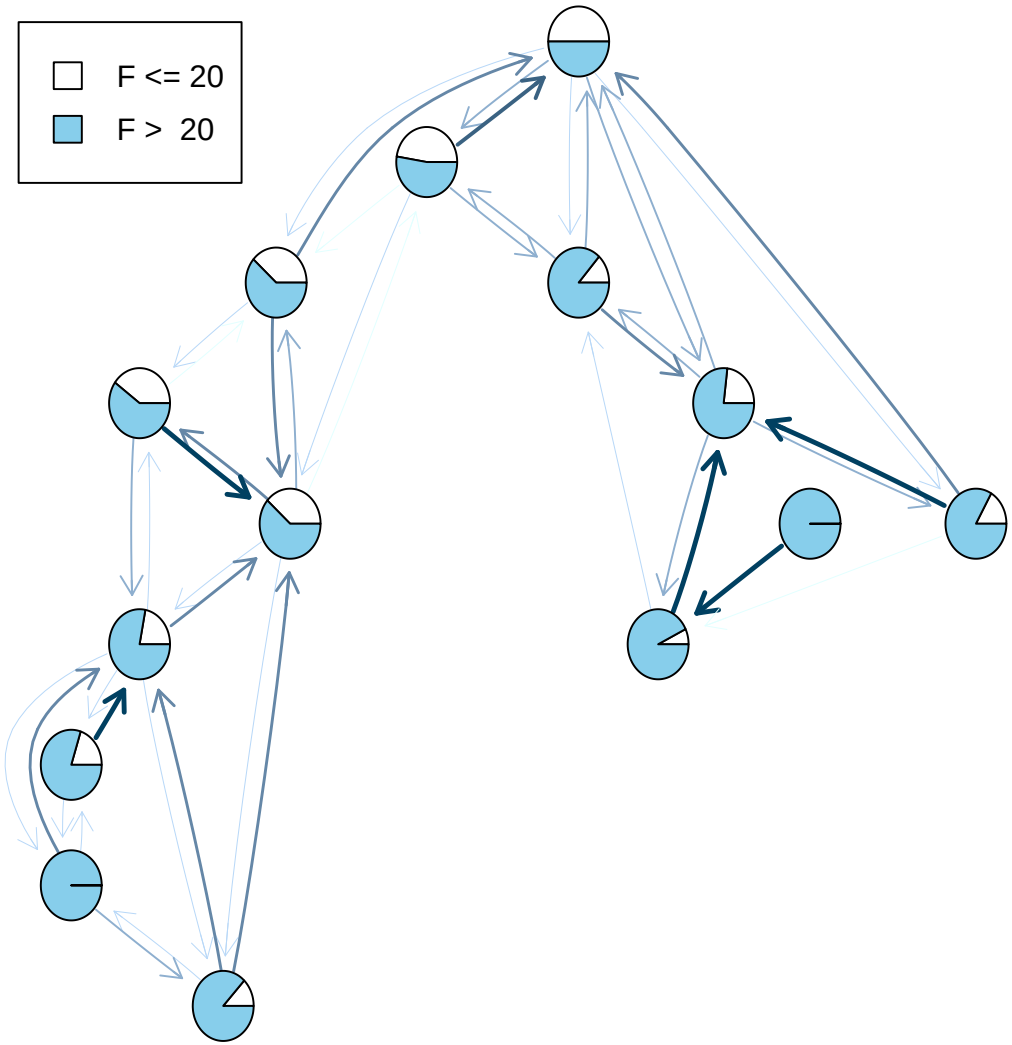
Use of pie plot as node function

```
R> gcExplorer(cl1, node.function = gpie,
+               doViewPort = TRUE)
```

Node function



Node function



R package **symbols**

- Based on Grid, a very flexible graphics system for R.
- Grid features viewports, i.e., rectangular areas allowing the creation of plotting regions all over the R graphic device.
- Implementation of several grid–based functions which can directly be used as node functions in the **gcExplorer**.
- Plot barplots, boxplots, line plots, pie charts, stars and symbols.

http://r-forge.r-project.org/projects/symbols

Neighborhood graph for general cluster functions

Cluster results from cluster functions like `kmeans` from package **stats** or `pam` from package **cluster** can be converted to objects of class `kcca` and visualized using the neighborhood graph:

Conversion

```
R> k1 <- kmeans(hsod, centers = 15)
R> k2 <- as.kcca(k1, data = hsod,
                 save.data = TRUE)
R> gcExplorer(k2)
```

Functional relevance test

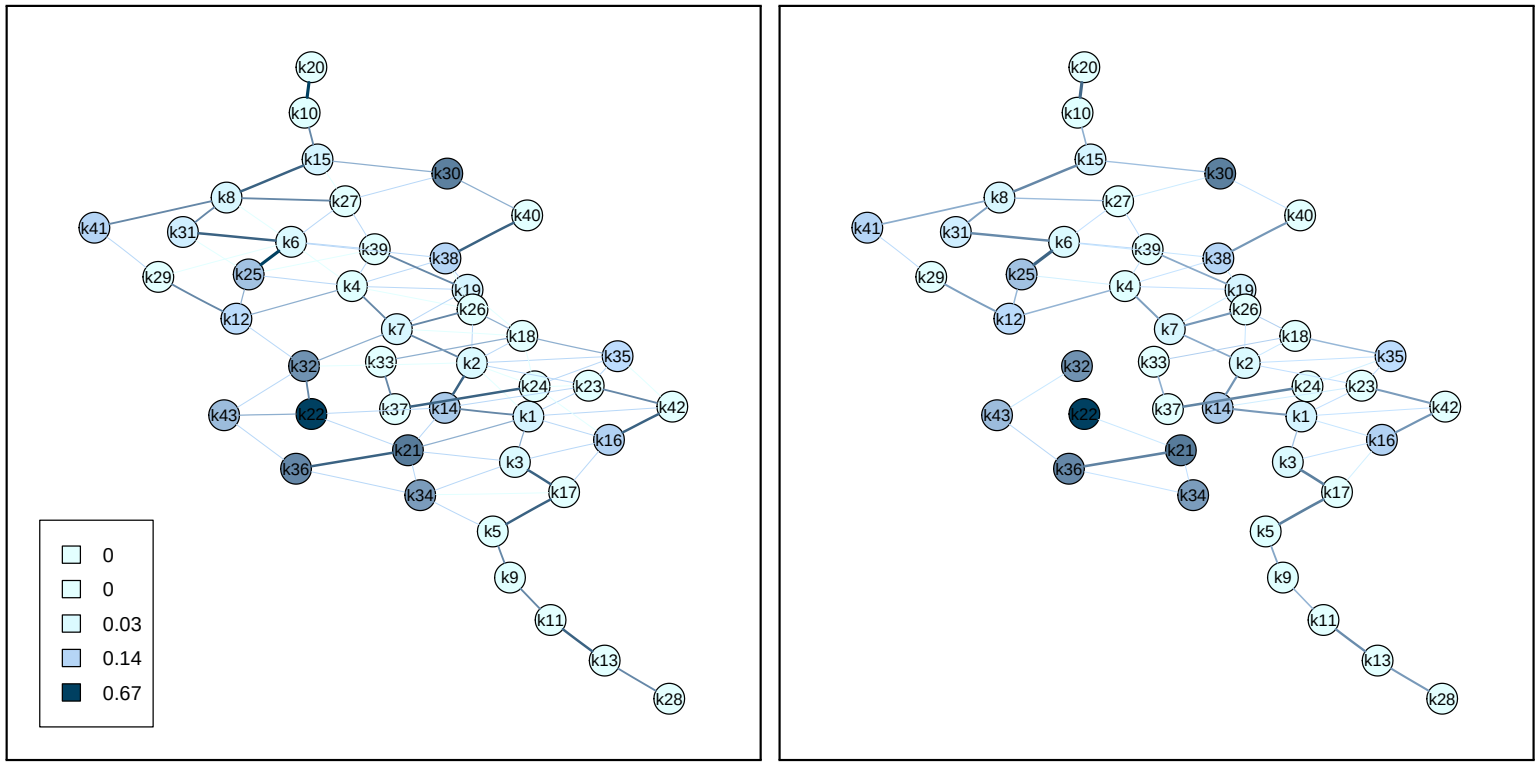
- Validation of a given clustering using a priori information about gene function.
- Let $\pi_1, \ldots, \pi_K$ be the proportions of genes assigned to a functional group.
- $H_0 : d_{ij} = |\pi_i - \pi_j| = 0$
- Use the neighborhood structure, i.e., only test for significant differences if two clusters are connected.
- No difference in proportions $\rightarrow$ merge clusters.
- Get separated subgraphs with common gene function within the neighborhood graph.

Functional relevance test: Procedure

- Step 1: Global test for equality of proportions
- If there are significant differences in proportions each single difference is investigated in more detail.
- Step 2: Assess the significance of the observed differences with respect to a reference distribution by permuting the function labels and keeping the respective maximum $M^l = max_{i,j} d_{ij}^l$
- Compute marginal tests of whether a particular d_{ij} is extreme relative to the joint distribution.

Functional Relevance Test

GO:0009061 (anaerobic respiration)



Summary

- Neighborhood graphs help to reveal structure in cluster solutions.
- gcExplorer is a flexible tool for exploration and inference of cluster solutions.

- Download and try http://cran.r-project.org/package=gcExplorer