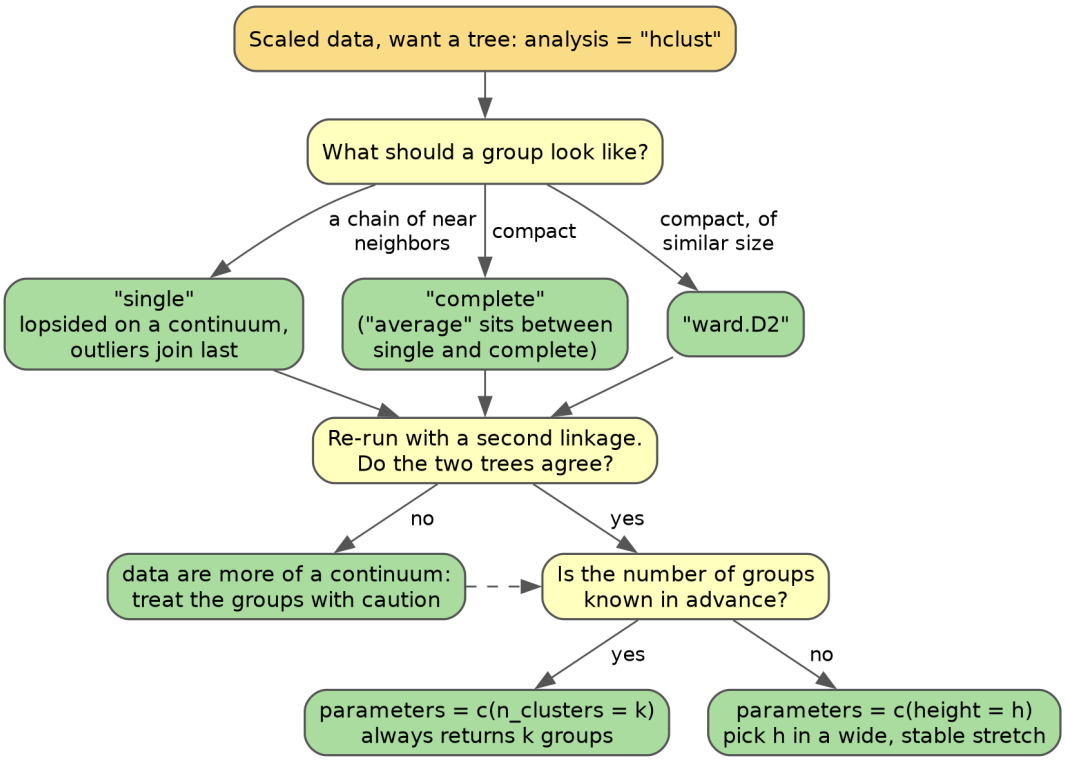
hierarchical clustering: summary

Integrated Bioanalytical Chemistry · thebustalab.github.io/integrated_bioanalytics



Hierarchical clustering merges the two closest groups, again and again, until every sample is connected by one tree structure. **The linkage rule decides what "closest" means for a group**, and different rules build different trees from the same distances. **A tree is read across, not down** (assuming it's drawn with the root on the left and leaves on the right): how far out from the tips two paths meet shows how different two samples are, and adjacent tips are not necessarily similar. **Cutting the tree turns it into groups**: choosing a height as the basis for the cut lets the data drive how many clusters are created, whereas setting a number of groups will force the tree to be cut into that number of clusters.

Function or setting	What it does	What to set or watch
`runMatrixAnalysis(analysis = "hclust")`	builds the tree from the distances	`scale_variance = TRUE`, `agglomeration_method`, `parameters`
`agglomeration_method = "single"`	joins on the closest members	chains a continuum into a lopsided tree
`agglomeration_method = "complete"`	joins on the furthest members	compact groups
`agglomeration_method = "average"`	joins on the average distance	sits between single and complete
`agglomeration_method = "ward.D2"`	adds least spread within groups	compact groups of similar size
`parameters = c(height = 5)`	cuts at a height, adds a `cluster` column	the data decide how many groups
`parameters = c(n_clusters = 4)`	cuts into a set number of groups	always returns that number
`ggtree()` + `geom_tiplab()`, `geom_tippoint()`	draws the tree and labels the tips	`aes(label = lake)`, `aes(fill = park)`, `align = TRUE`
`geom_cut(height = 5)` or `geom_cut(k = 4)`	draws the cut on the tree	the same height as `parameters`, never `geom_vline()`
`filter(isTip)` + `aes(y = y)` in `geom_tile()`	heat map in the tree's tip order	same `scale_y_continuous(limits = ...)` as the tree, then `plot_grid()`