Connected Linkage and Alignment Redefinition of COGs (CLARC) - clustering algorithm

Pairwise identification of accessory COGs that look to be the 'same unit' based on constraints

Build graph and identify fully connected clusters

Re-define COGs; condense COGs in connected clusters

Two COGs called the 'same unit' if:

- (1) COGs never co-occur in the same genome
- (2) COGs have a high percentage of sequence similarity (default >95% identical matches BLASTN)
- (3) COGs get classified in the same EggNOG functional group

Edges = 'same gene' relationship (3 constraints met between the 2 COGs)

COGs with connections in multiple clusters

Nodes = COGs

Fully connected clusters

Call COGs the same gene

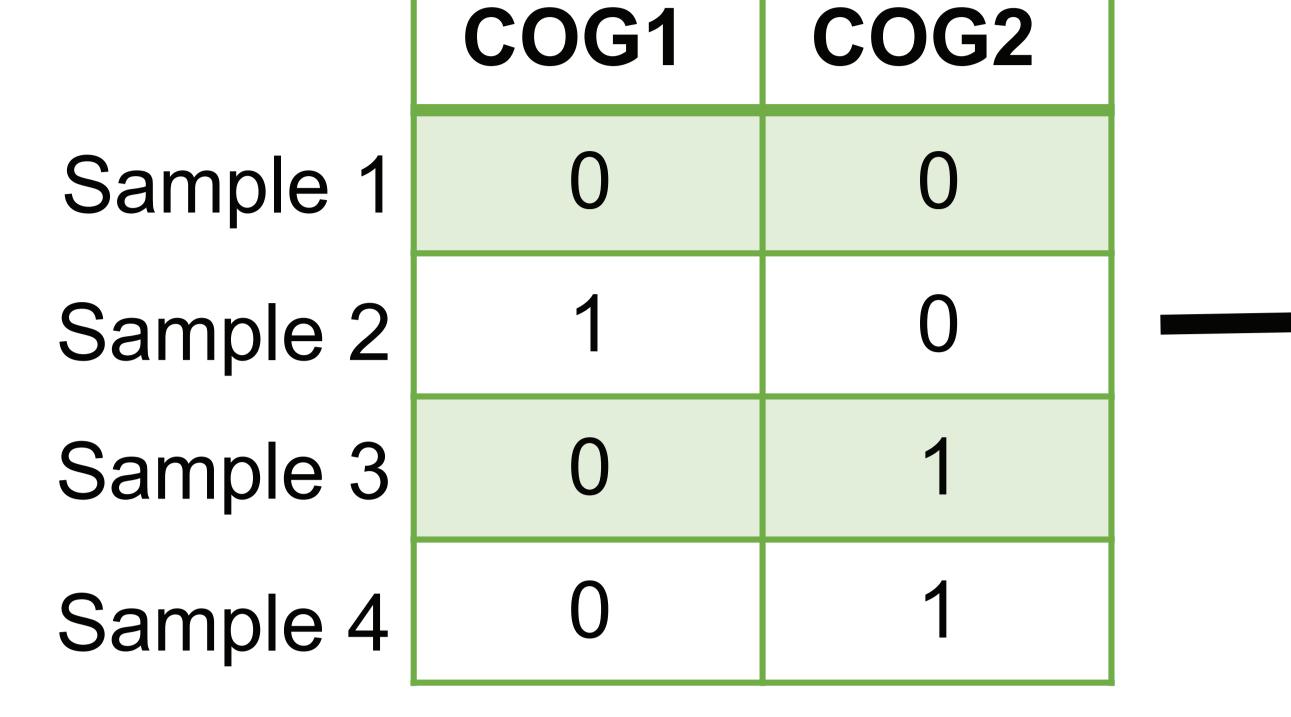
COGs with connections in fully connected clusters

Disregard

from analysis

(leave as is)

Not fully connected clusters



COG1-COG2 COG2

General tool workflow

