

Genome Clustering

Genomes in IMG can be compared in terms of clusters by using the clustering tools available under IMG's **Compare Genomes** main menu option, as illustrated in Figure 1(i). Genomes can be clustered by using Hierarchical Clustering, Principal Components Analysis (PCA), Principal Coordinates Analysis (PCoA), Non-metric MultiDimensional Scaling (NMDS), or Correlation Matrix.

Hierarchical Clustering

Select first the type of protein/functional families (COG, Pfam, Enzyme), and Hierarchical Clustering method and the 2 to 2300 genomes you want to compare in the Genome Clustering page, as illustrated in Figure 1(i). The **Hierarchical Clustering Results** page displays a radial tree phylogram, as illustrated in Figure 1(ii), and a rectangular tree phylogram, as illustrated in Figure 1(iii). The placement in the tree reflects the distance between genomes, whereby the computed distance is based on the similarity of the functional characterization of genomes in terms of a specific protein/functional family.

There are additional options in the **Hierarchical Clustering Results** page to let the users view phyloXML and Newick File.

The screenshot displays the 'Genome Clustering' web interface. It includes a search bar, a list of genomes, and various clustering options. The interface is divided into three main sections: (i) the main control panel, (ii) a radial tree phylogram, and (iii) a rectangular tree phylogram.

(i) Genome Clustering Interface:

- Sequencing Status:** Finished (selected), Domain: Archaea.
- Clustering Type:**
 - By Function: COG, Pfam, KO.
 - By Taxonomy: Class, Family, Genus.
 - By Function Category: COG Categories, COG Pathways, KEGG Pathway Categories (KO), KEGG Pathway Categories (EC), KEGG Pathways (KO), KEGG Pathways (EC), Pfam Categories.
- Clustering Method:**
 - Hierarchical Clustering
 - Principal Components Analysis (PCA)
 - Principal Coordinates Analysis (PCoA)
 - Non-metric MultiDimensional Scaling (NMDS)
 - Correlation Matrix
- Buttons:** Go, Reset.

(ii) Radial Tree Phylogram: A circular tree diagram showing the hierarchical clustering of 12 selected genomes. The branches radiate from a central point, with labels for each genome.

(iii) Rectangular Tree Phylogram: A standard rectangular tree diagram showing the hierarchical clustering of the same 12 selected genomes. The branches are horizontal and vertical, with labels for each genome.

Selected Genomes (12 selected):

- Acididarius hospitalis W1 (A) [F]
- Acidobacterium saccharovorans 345-15 (A) [F]
- Aciduliprofundum boonei T469 (A) [F]
- Aciduliprofundum sp. M4308-339 (A) [F]
- Aeropyrum caminis SY1, JCM 12091 (A) [F]
- Archaeoglobus fulgidus DSM 8774 (A) [F]
- Archaeoglobus fulgidus VC-16, DSM 4304 (A) [F]
- Archaeoglobus profundus Av18, DSM 5631 (A) [F]
- Archaeoglobus sulfatolicolus PH70-1, DSM 19444 (Reau_fusion2) (A)
- Archaeoglobus venificus SH96, DSM 11192 (A) [F]
- Caldiverga laguerensis DSM 15908 (A) [F]
- Caldiverga maculiformis VC-167 (A) [F]

Figure 1. Hierarchical Clustering

Principal Components Analysis (PCA)

Principal Components Analysis is an ordination tool for exploratory data analysis which reduces the dimension of a data set such that it can be visualized in a 2 or 3-D plot [1][2]. The method creates synthetic variables which are linear combinations of the original variables, and can be plotted on their corresponding orthogonal principal axes. The principal components can explain much of the variation found in the data set. The most useful feature of PCA, compared to other ordination techniques, is that it can show what variables drive the separation of objects in the plot.

When comparing (meta)genomes, we define the objects in ordination as (meta)genomes, and the variables as functional or taxonomic classifications. We refer to the list of gene counts in either a functional (Pfam, COG, etc) or taxonomic (class, family, etc.) classification as a (meta)genome's profile. We are interested in seeing what (meta)genomes cluster together, which suggests profile similarity. We are also interested in viewing gradients, which are varying aspects of the environment related to the profiles.

It is difficult to prove that (meta)genomic data sets meet the assumptions needed for appropriate use of PCA. Functions/taxa need to be linearly related. For the principal axes to be an informative summary of the data, a change in one function/taxon must result in a linear change of another function/taxon. Also, the variables need to be normally distributed. With the prevalence of zeros in (meta)genome profiles, this assumption is not met in most cases.

However, PCA is a geometric technique and not a statistical test. So, even though the assumptions are not strictly met, it is possible to find meaningful information in the ordination. If the first few principal components explain much of the variance, PCA is an informative representation of the objects. As an exploratory data tool only, PCA in IMG can be used to form hypotheses that can be later tested in properly designed statistical studies.

In order to account for the large variance in (meta)genome profiles, every (meta)genome is normalized by dividing each gene abundance by the total number of genes in the (meta)genome. The normalized abundances for each (meta)genome sum to 1. For example, if (meta)genome A has profile 0,100,50,150, and (meta)genome B has profile 0,50,25,75, both profiles will be normalized such that they have equal profiles of 0,0.33,0.17,0.5.

Although the normalization reduces the Euclidean distance between (meta)genomes that have similar profiles but vastly different gene sizes, it does not solve the double zeros issue [3]. If two (meta)genomes have similar absence profiles, meaning they have no genes from the same functions/taxa, they will possibly be close together in the PCA ordination. However, it is likely that an absence of a function/taxon in a (meta)genome is due to lack of read depth and/or coverage [4]. We are more interested in viewing what (meta)genomes are more similar based on their present functions/taxa, in which case PCA is not the most appropriate tool for visualizing (meta)genomic data sets.

To perform Principal Component Analysis (PCA) in IMG, first select between 2 and 2300 genomes or metagenomes, and then select Principal Components Analysis (PCA) option in **Clustering Method** in Figure 1(i). PCA results in 3D and 2D are shown in the **3-D Plot** and **2-D Plot** tabs (Figures 2(i) and 2(ii)),

respectively. The **Components** tab shows the Principal Components matrix (see Figure 2(iii)), and the **Histogram** tabs show histogram display (see Figure 2(iv)).

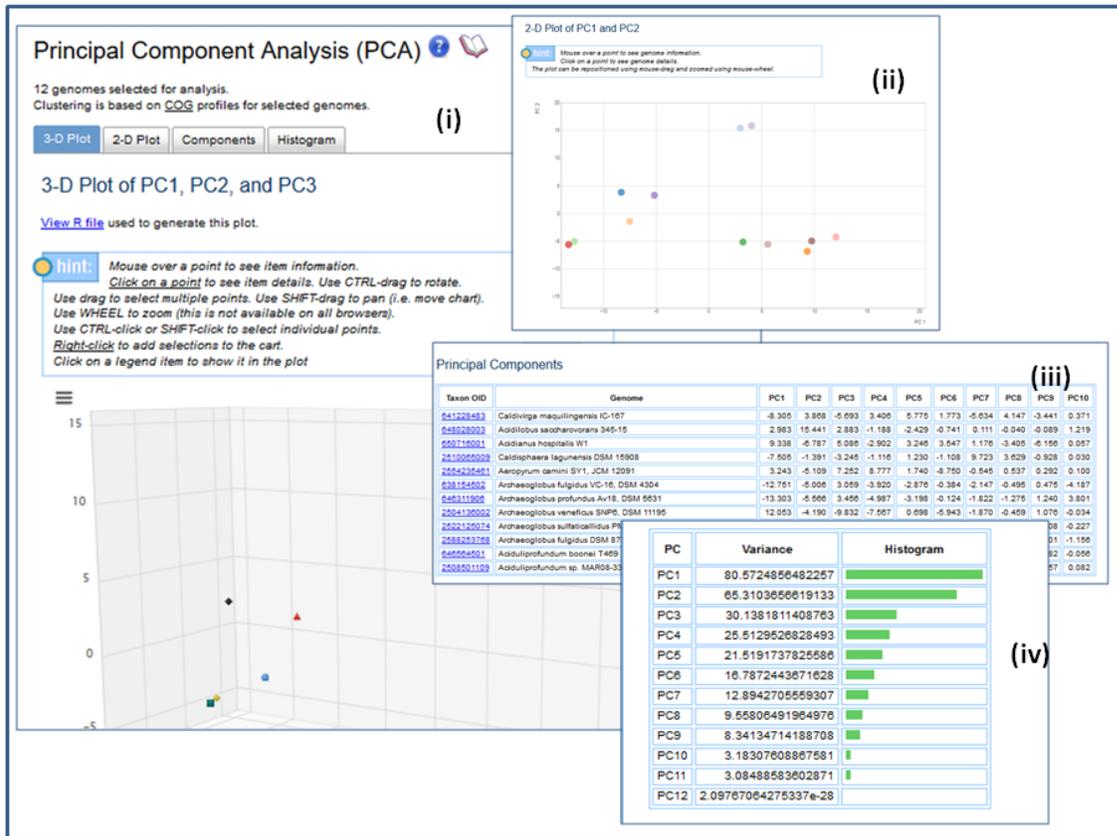


Figure 2. Principal Components Analysis (PCA).

Principal Coordinates Analysis (PCoA)

Principal Coordinates Analysis (PCoA) [5][6] is an eigenanalysis algorithm like Principal Component Analysis (PCA). Whereas PCA only uses the Euclidean distances between objects to perform an ordination in reduced space, PCoA performs an ordination on any user-selected dissimilarity measure. If Euclidean distance is chosen, PCoA gives the same solution as PCA. If an appropriate measure is selected, no data assumptions need to be verified before analysis.

One of the best measures to use with raw abundances is the Bray-Curtis dissimilarity coefficient [3][7], which we use to measure the compositional dissimilarity between two (meta)genomes. If we have a $m \times n$ abundance matrix X , with m (meta)genomes and n functions/taxa, we calculate the Bray-Curtis dissimilarity between (meta)genome j and (meta)genome k as [8]:

$$d_{jk} = \frac{\sum_{i=1}^n |x_{ij} - x_{ik}|}{\sum_{i=1}^n (x_{ij} + x_{ik})}$$

If two (meta)genomes have exactly the same profile, their index is 0, which is the smallest (most similar) Bray-Curtis index. If two (meta)genomes share no functions or taxa in their profile, their index is 1, representing the largest measure of Bray-Curtis dissimilarity.

The Bray-Curtis index is a semi-metric distance, meaning it does not exhibit the properties of the triangle inequality. Thus, negative eigenvalues may result from the PCoA. However, if the negative eigenvalues do not occur in the first few principal coordinates, the ordination may be meaningful in some cases [3].

Although performing PCoA with the Bray-Curtis index gives a much more appropriate representation of the relationship between (meta)genome profiles than PCA, it is difficult to recover the functions/taxa contributing to the principal coordinates. Unlike PCA, the new variables are complex functions of the original variables, not linear combinations.

PCoA in IMG is a data exploratory tool only, and may be used to form hypotheses that can later be tested in properly designed statistical studies.

To perform Principal Coordinates Analysis (PCoA) in IMG, first select between 2 and 2300 genomes or metagenomes, and then select Principal Coordinates Analysis (PCoA) option in **Clustering Method** in Figure 1(i). PCoA result in 3D is shown in Figure 3(i). Click the "View R File" link to view the R script that generates the result (Figure 3(ii)).

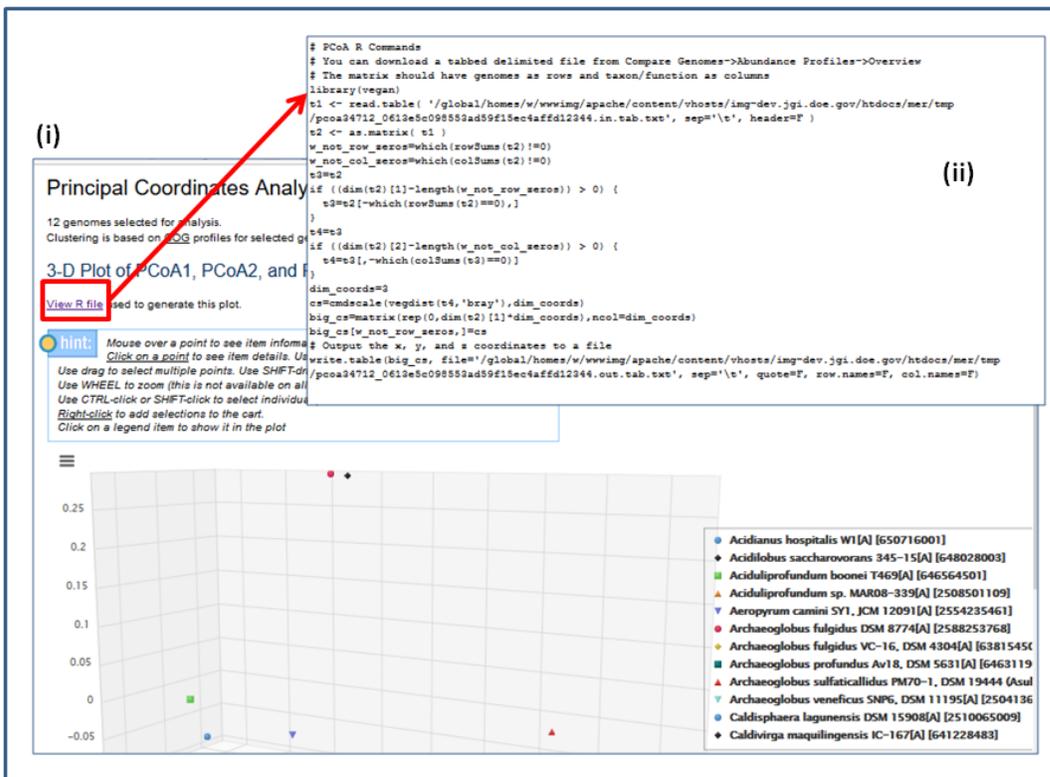


Figure 3. Principal Coordinates Analysis (PCoA)

Non-metric MultiDimensional Scaling (NMDS)

Non-metric MultiDimensional Scaling (NMDS) [9][10][11][12] is an iterative ordination technique that preserves the rank order correlation between objects, rather than their linear correlation. It is better than PCoA (Principal Coordinates Analysis) at representing relationships between objects because of this model flexibility [3]. However, since it is an iterative algorithm two problems may arise. First, there is no guarantee that the solution found is best, as the algorithm could return on a local minimum. Second, many iterations on large matrices is computationally intensive, so NMDS can take an extended amount of time to return a solution.

Before running NMDS, both the dissimilarity metric and the number of dimensions of the solution need to be chosen. Unlike PCA and PCoA, which find solutions in n and $n-1$ dimensions respectively, NMDS can find a solution in 1 to $n-1$ dimensions. NMDS in IMG is displayed in 3 dimensions because the solution is easily viewed in a 3-D plot, not because 3 dimensions necessarily gives the best solution. Like PCoA, NMDS has no data assumptions, and can be performed on any dissimilarity measure.

One of the best measures to use with raw abundances is the Bray-Curtis dissimilarity coefficient [3][7], which we use to measure the compositional dissimilarity between two (meta)genomes. If we have a $m \times n$ abundance matrix X , with m (meta)genomes and n functions/taxa, we calculate the Bray-Curtis dissimilarity between (meta)genome j and (meta)genome k as [8]:

$$d_{jk} = \frac{\sum_{i=1}^n |x_{ij} - x_{ik}|}{\sum_{i=1}^n (x_{ij} + x_{ik})}$$

If two (meta)genomes have exactly the same profile, their index is 0, which is the smallest (most similar) Bray-Curtis index. If two (meta)genomes share no functions or taxa in their profile, their index is 1, representing the largest measure of Bray-Curtis dissimilarity.

Because the Bray-Curtis index is used in the ordination, rather than the abundances, it is difficult to recover the functions/taxa contributing to the separation of the (meta)genomes. Unlike PCA, the new variables are complex functions of the original variables, not linear combinations.

NMDS is the most appropriate tool in IMG for reducing the dimensions of (meta)genomic functional/taxonomic profiles for comparative analysis. However, both PCA and PCoA find a solution more quickly on some datasets.

NMDS in IMG is a data exploratory tool only, and may be used to form hypotheses that can later be tested in properly designed statistical studies.

To perform Non-metric MultiDimensional Scaling (NMDS) in IMG, first select between 2 and 2300 genomes or metagenomes, and then select [Non-metric MultiDimensional Scaling \(NMDS\)](#) option in **Clustering Method** in Figure 1(i). NMDS result in 3D is shown in Figure 4(i). Click the "View R File" link to view the R script that generates the result (Figure 4(ii)).

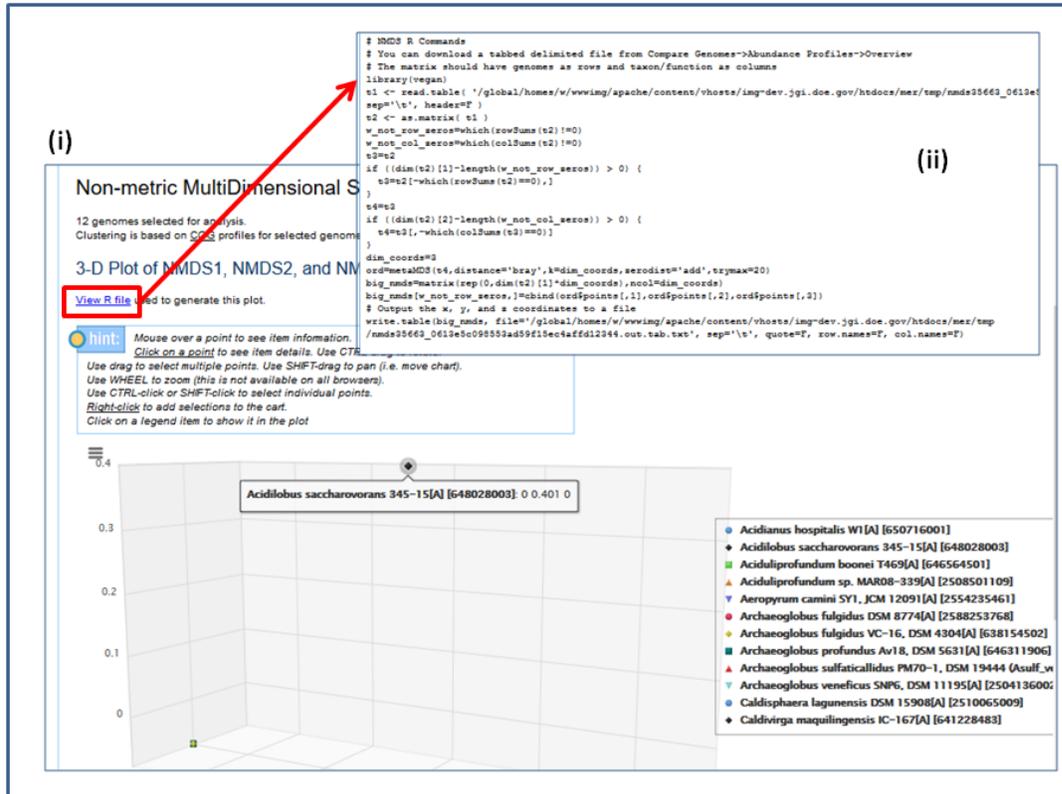


Figure 4. Non-metric MultiDimensional Scaling (NMDS)

Correlation Matrix

To perform Correlation Matrix analysis in IMG, first select between 2 and 2300 genomes or metagenomes, and then select Correlation Matrix option in **Clustering Method** in Figure 1(i).

The results of correlation clustering are displayed in a matrix format, as illustrated in Figure 5(i), where each cell of the matrix displays the correlation coefficient between the genomes on the corresponding row and column. The correlation coefficient is computed based on the similarity of the functional characterization of the genomes. The diagonal correlations (of genomes with themselves) are always 1.00. The genomes listed in the clustering results are linked to their associated Organism Details pages.

Click the **Show Color Scheme** button to find out more details on the color scheme (Figure 5(ii)).

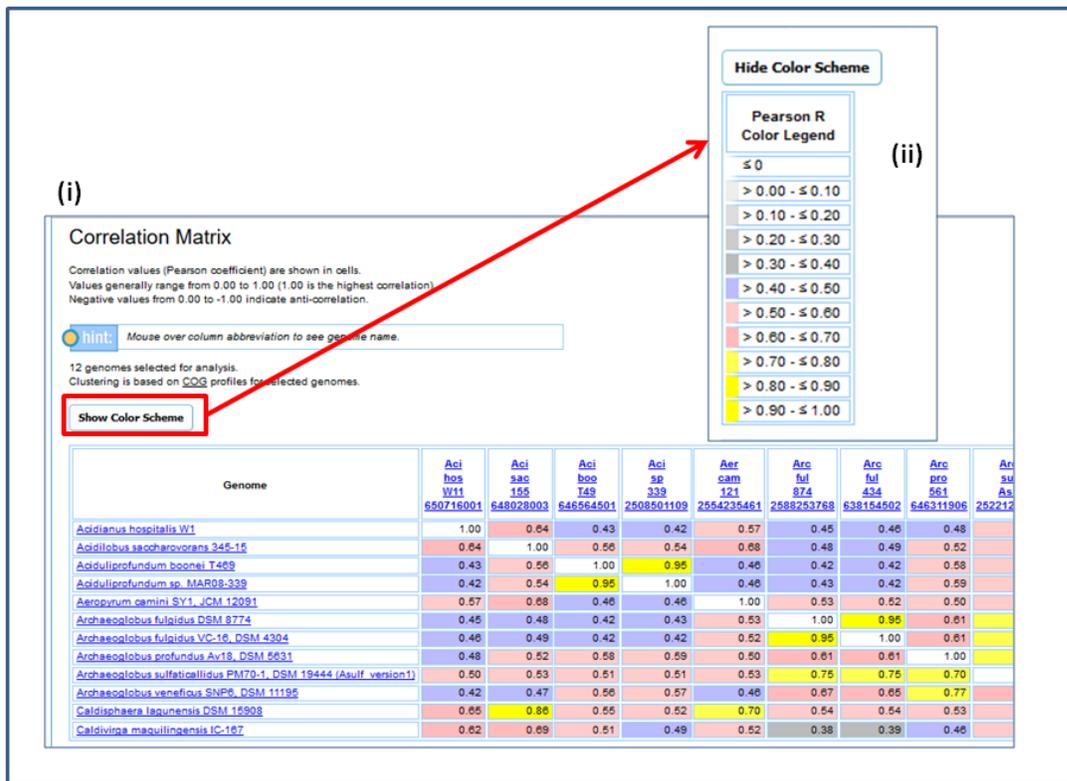


Figure 5. Correlation Matrix.

References

- [1] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6):559–572, 1901.
- [2] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417–441,498–520, 1933.
- [3] P. Legendre and L. Legendre. *Numerical Ecology. Developments in Environmental Modelling*. Elsevier, 1998.
- [4] V. Kunin, A. Copeland, A. Lapidus, K. Mavromatis, and P. Hugenholtz. A bioinformaticians guide to metagenomics. *Microbiology and molecular biology reviews MMBR*, 72(4):557–78, 2008.
- [5] J. C. Gower. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53:325–338, 1966.
- [6] W.S. Torgerson. *Theory and methods of scaling*. Wiley, 1958.
- [7] A.E. Magurran. *Measuring Biological Diversity*. Blackwell, 2004.

[8] J.T. Curtis J. Roger Bray. An ordination of the upland forest communities of southern wisconsin. *Ecological Monographs*, 27(4):325–349, 1957.

[9] R.N. Shepard. The analysis of proximities: multidimensional scaling with an unknown distance function. *Psychometrika*, 27:125–139, 1962.

[10] J.B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29:1–27, 1964.

[11] J.B. Kruskal. Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, 29:115–129, 1964.

[12] R.N. Shepard. Metric structures in ordinal data. *Mathematical Psychology*, 3:287–315, 1966.