CosmosID MetaGenID

MetaGenID allows you to run your samples using CosmosID's curated genome databases and high performance algorithms to provide rapid, accurate, and actionable microbial identification at the species, subspecies, and/or strain level.

MetaGenID identifies multiple microorganisms in a whole genome shotgun metagenomic sample using statistical and computational methods, with no prior assumptions as to what is present in the sample. CosmosID uses raw, unassembled reads as input and matches the sequence against our reference databases of known bacteria, viruses, fungi, parasites and antibiotic resistance genes.

For use with whole genome shotgun samples only, not 16S.

Signing In

In the CLC genomics workbench, click "**Edit**" -> "**Preferences**"
Scroll down and enter the email address and password provided to you by Qiagen for your CosmosID account.

Analyzing Samples

For paired end samples, use CLC to merge the samples before uploading.

Click "**Upload Sample**" and select samples for analysis.
After your samples are uploaded, analysis begins automatically. In the Dashboard you will see the samples you have recently uploaded and their status.

To view results, double click "**Get Analysis Results**". Click "**Expand**" to view the list of results from each database. The "Status" will say "Running" if the sample is still undergoing analysis. "Success" indicates that the sample ran successfully and "Failed" indicates a problem. Contact support@cosmosid.com for help if needed.

Select the checkboxes next to the results you would like to view and click "**Next**". Select **Open** or **Save** and click "**Finish**". Interpretation of the table is explained below.

Comparative Analysis

To compare results between samples and create visualizations such as heat maps and Principal Component Analyses (PCA), double click "**Compare Samples**".

Assign a name for the results in the "**Name**" box at the top. Select the **database** to use in the dropdown menu. If wanted, select "**log scale**" for the logarithmic scale of the results to be used to create the visualizations (good for using relative abundance as the comparator for visualizations in heat maps of differences between samples among the organisms present at low levels of abundance).

Labels can be created for PCA groups (generally if there are predetermined groups of samples, such as temporal information or body site locations, for example).

Select the samples to be used for comparative analysis and click "**Next**". Click **Open** and click **Finish**.

First a matrix is shown for the comparison of samples. Clicking on the buttons at the bottom toggles between Principal Component Analysis (PCA), a heat map, and a clustered heat map with dendrograms. All figures can be exported as .png files by clicking "**Export as .png**" at the top.

To retrieve previously saved Comparative Analysis results, click "**Get Comparative Analysis Results**".

To determine how many tokens are available, click "**Get Token Information**".

Interpretation of Results

Table:

Name: The organism (*most likely strain*) identified in the queried sample. Identification of a strain conveys the ***natural concept*** of strain, that is, a strain will always have a like counterpart in nature and potentially is a sub-sample derived from a strain found in nature. Therefore, identification of a given strain indicates either presence of itself or its sub-strain sharing highest genomic similarity to it.

Frequency: The number of unique kmer occurrences in the queried sample. This is roughly equivalent to the number of reads that hit the organism.

Unique Matches Percent: The number of different kmers found in the sample that are unique to the organism identified divided by the number of pre-calculated total possible unique kmers in the reference database.

Total Matches Percent: The shared plus unique matches divided by the pre-calculated shared plus unique matches possible in the reference database.

Relative Abundance: Relative abundance is calculated based on the number of organism specific kmers and their observed frequency in the sample and then normalized to represent the abundance of each organism.

Filtered: The filtering threshold is based internal statistical scores determined by analyzing a large number of diverse metagenomes. Organisms that are identified with filtering "**true**" are most likely to be present in the sample. When filtering is "**false**", those organisms need further validation to determine if they are actually present in the sample - either by deeper sequencing of the sample followed by re-analysis or by orthogonal validation using targeted PCR or other methods.

Problems or Questions

Email us at support@cosmosid.com