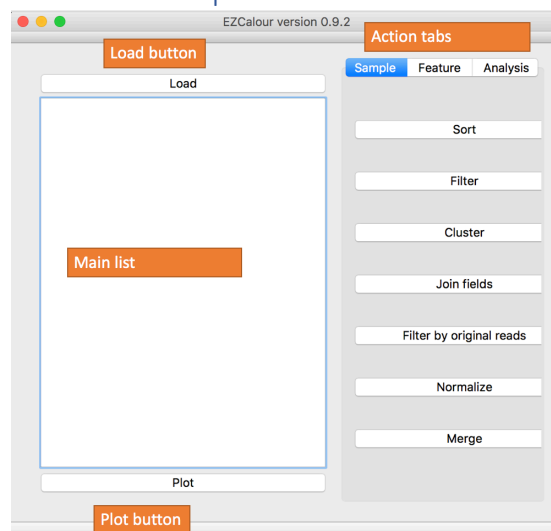# Using EZCalour

EZCalour is a point and click GUI for the Calour microbiome analysis package
(Xu ZZ, Amir A, Sanders J, Zhu Q, Morton JT, Bletz MC, Tripathi A, Huang S, McDonald D,
Jiang L, Knight R. 2019. Calour: an interactive, microbe-centric analysis tool. mSystems
4:e00269-18. https://doi.org/10.1128/mSystems.00269-18)

EZCalour can be used to read, process, analyze and and plot interactive heatmaps from
microbiome experiments.

General concepts

Each dataset is called an experiment. All experiments are displayed in the *main list* in the
EZCalour window. Following each processing step, a new experiment is created. Right clicking
on an experiment in the *main list* enables deleting from memory, saving and viewing the
complete command history of the experiments.

On the right-hand side, there are three *action tabs* - for processing *samples*, *features* (bacteria)
and *analysis*. Commands from a given tab relate to the appropriate axis (i.e. Filter from the
Sample tab filters samples, whereas Filter from the *Feature* tab filters features). Commands
work on the selected experiment from the main list.

In order to plot the interactive heatmap of an experiment, it needs to be selected from the
main list, and then press the "*Plot*" button (located at the lower left side)

Loading data

EZCalour works with microbiome BIOM tables, qiime2 qza feature table files, metabolomics (MZMine2) tables, or any CSV text file.

Besides the main table, ezcalour can also load a tab-separated mapping file, containing information about each sample.

In order to load an experiment, click on the "*Load*" button (located at the top left side).
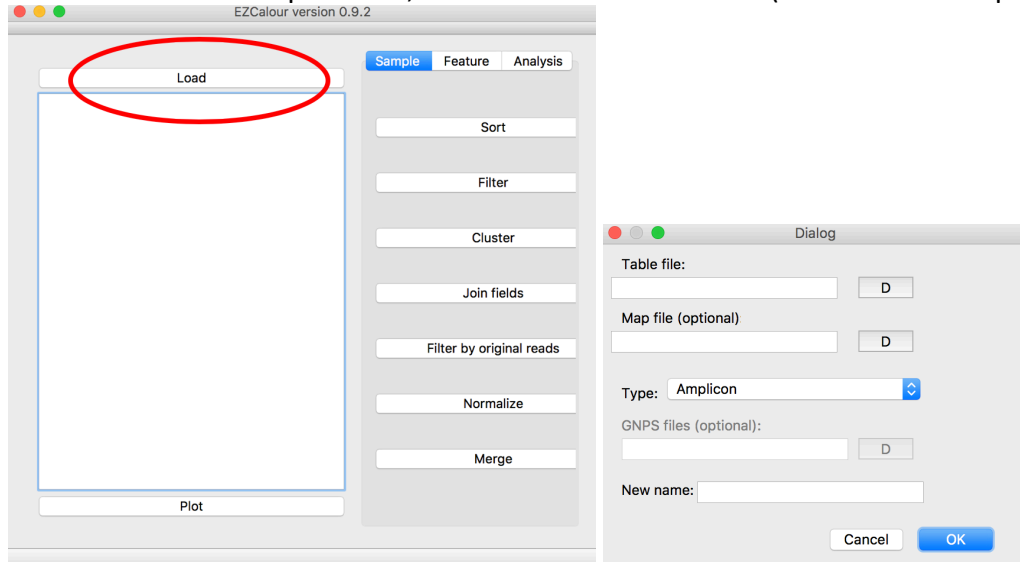

Mandatory fields:

- "*Table file*": name of the biom/qiime2 qza/mzmine2 table (can click the "*D*" button for GUI file selection)
- "*Type*": the type of the table file:
  - "Amplicon" for a microbiome biom table or qiime2 qza feature table. When loading, the table is normalized by TSS to 10000 reads/sample. Samples with <1000 reads are dropped.
  - "MZMine2" for an MZMine2 metabolomics table
  - "TSV" for a general tab separated table (Each sample is a column, each feature is a row)
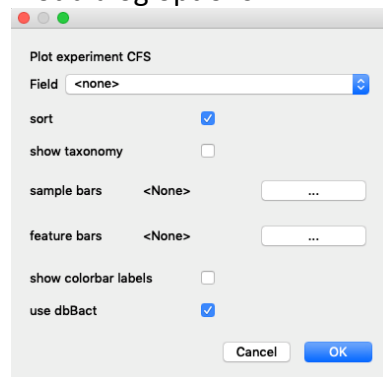
Optional fields:

- "*Map file*": name of the sample TSV mapping file (a tab separated file, with each sample in a row, first column contains the sample names matching the table).
- "*GNPS file*" : For mass-spec, the per-metabolite info file (see Calour documentation for more info)
- "*New name*": the name for the experiment in the *main list* (defaults to the table file name)

Plotting data

In order to plot an *experiment*, first select the experiment from the *main list* (1), and then press the "*Plot*" button (2).
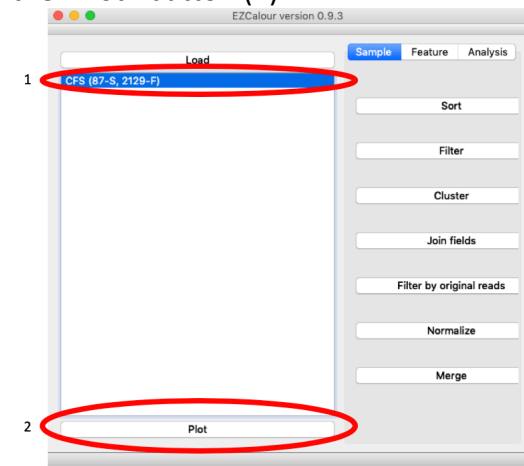

Plot dialog options:

Optional fields:

- "*Field*": if not <none>, the sample metadata field to plot separator vertical lines according to (and add to the x-legend) (1 in the example plot)
- "*sort*": if checked, sort samples according to "*field*" prior to plotting.
- "*show taxonomy*": if checked, show the taxonomy next to each bacteria (y-axis). NOTE: if a large number of bacteria are present, will show taxonomy only for part of the bacteria, depending on the zoom level.
- "*sample bars*": list of sample fields to show on top of the plot. Select using the "…" button. (2 in the example plot)
- "*feature bars*": list of feature fields to show on the right of the plot. Select using the "…" button.
- "*Show colorbar labels*": If checked, include text in the sample/feature bars. Otherwise, just use different colors.
- "*use dbBact*": If checked, query dbBact interactively when clicking on features in the plot.

Sample Interactive heatmap plot
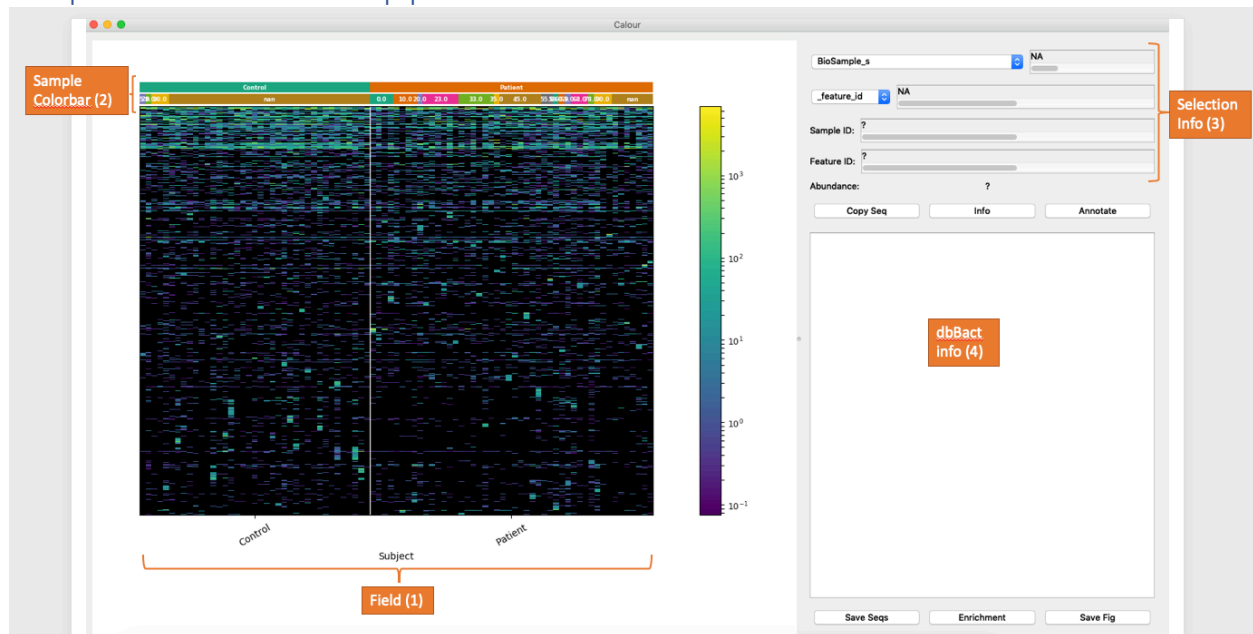
The heatmap shows each sample as a column, each feature (bacteria) as a row.

Interactive heatmap navigation

- Shift + Up arrow or "=": zoom in features
- Shift + Down arrow or "-": zoom out features
- Shift + Right arrow or "+": zoom in samples
- Shift + Left arrow or "_": zoom out samples
- Up/Down arrow: scroll features up/down (when zoomed)
- Left/Right arrow: scroll samples up/down (when zoomed)

NOTE: keyboard shortcuts require the mouse cursor to be over the heatmap

Clicking on a point in the heatmap shows the corresponding sample/feature information in the selection info panel (3). Additionally, information from dbBact regarding the selected bacteria is shown in the dbBact info pane (4). Multiple feature selection is possible using ctrl/shift+clicking.

Additional buttons

- "*Save Seqs*": save the selected features to a fasta file.
- "*Save Fig*": save the current figure view to a pdf/tif etc.
- "*Enrichment*": Check for dbBact term enrichment in the selected features (bacteria) compared to the rest of the features of the experiment (see more details at the end of the documentation).
- "*Annotate*": Add a dbBact annotation regarding the selected features in the plot (NOTE: this will be viewable to all dbBact users). Please consult dbBact documentation (dbbact.org/about) for more details.

Processing data

Sample tab

Buttons from this tab affect the samples of the experiment. Button actions are performed on the selected experiment from the main list. Each action generates a new experiment in the main list.

Sort

Sort the samples according to the selected field.
Sorting is conservative, meaning samples with same value in the field retain the previous order. In order to sort by two fields (i.e. "Disease" and "Day" within each disease), sort first by the second field (i.e. "Day") and then by the first (i.e. "Disease").
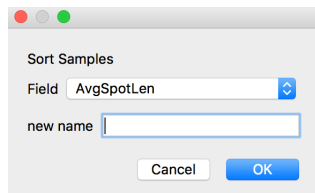
Mandatory fields:

- "*Field*": select the sample metadata field to sort by

Optional fields:

- "*New name*": the name for the experiment in the main list (defaults to the table file name)

Filter

Keep or remove samples with specified mapping file field values
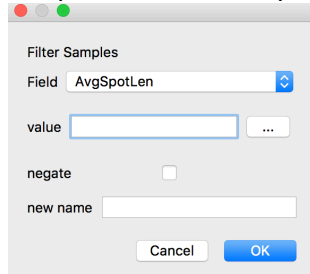
Mandatory fields:

- "*Field*": select the sample metadata field to sort by
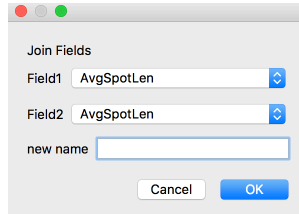- "*Value*": the values to filter for the field.

Optional fields:

- "*negate*": if checked, remove samples with the selected values, otherwise keep samples with selected values
- "*New name*": the name for the experiment in the main list (defaults to the table file name)

Cluster

Cluster the samples by putting similar samples next to each other

Join Fields

Create a new metadata field by joining the values of two fields.

Mandatory fields:

- "*Field1*", "*Field2*": The two fields to join (new field will be field1-field2).
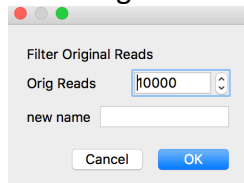
Optional fields:

- "*New name*": the name for the experiment in the main list.

Filter by original reads

Throw away samples with < threshold original reads (before normalization).
Used to get rid of samples with a small number of reads

Mandatory fields:

- "*Orig. reads*": the minimal number of reads in the sample in order to keep
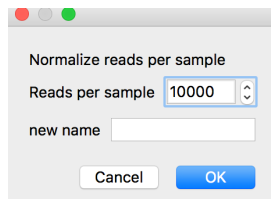
Optional fields:

- "New name" : the name for the experiment in the main list.

Normalize

TSS (total sum scaling) normalization of the reads per sample.
When loading a microbiome table, samples are automatically scaled to 10000 reads/sample.
NOTE: This is not rarefaction, so features can have fractional reads.

Mandatory fields:

- "Reads per sample": The sum of reads per sample to normalize to
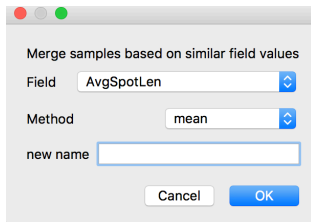
Optional fields:

- "*New name*": the name for the experiment in the main list.

Merge

merge samples having the same value in the selected field (for example all samples coming from the same individual).

Samples can be merged using mean, sum or randomly choosing one of the samples with the value.

Mandatory fields:

- "*Field*": The field to merge samples sharing the same value
- "*Method*": how to merge the samples with same value. Options are:
  - "mean": new sample contains the mean of each of the features from all the samples with the value
  - "random": new sample contains a randomly chosen sample with the value
  - "sum": new sample contains the sum of each of the features from all the samples with the value

Optional fields:

- "*New name*": the name for the experiment in the main list.
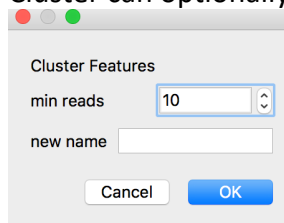
Feature tab

Buttons from this tab affect the features (bacteria) of the experiment. Button actions are performed on the selected experiment from the main list. Each action generates a new experiment in the main list.

Cluster

Cluster the features by putting similar behaving features next to each other.

Each feature is normalized to mean=0, std=1 and then Euclidian distance is used for clustering, so similar behaving features over the samples will be close to each other, without dependence on the absolute level of the features.

Cluster can optionally also remove low abundance features (to speed up the clustering).
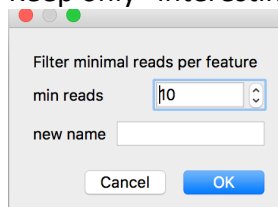

Mandatory fields:

- "*min reads*": Keep only features with total reads over all samples >= min reads prior to the clustering step. If 0, do not remove any features prior to clustering

Optional fields:

- "*New name*": the name for the experiment in the main list.

Filter min reads

Keep only "interesting features" that have enough total reads (over all samples)

Mandatory fields:

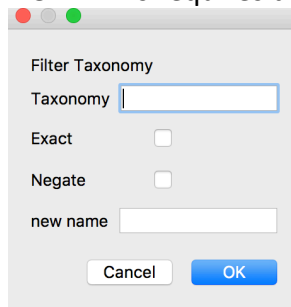- "*min reads*": Keep only features with total reads over all samples >= min reads.

Optional fields:

- "*New name*": the name for the experiment in the main list.

Filter taxonomy

Keep (or remove) features matching a given taxonomy string.
NOTE: This requires the biom table to contain taxonomy information.
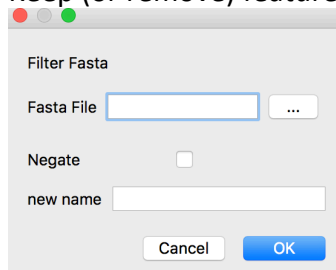

Mandatory fields:

- "*Taxonomy*": A partial or complete taxonomy string. Needs to match the taxonomy embedded in the biom table.

Optional fields:

- "*Exact*": Check to filter only taxonomy strings fully matching the string. Uncheck to allow partial matches (i.e. "roteo" will match "p__Proteobacteria")
- "*Negate*": Check to remove matching features, uncheck to keep matching features.
- "*New name*": the name for the experiment in the main list.

Filter fasta

Keep (or remove) features in the *experiment* that appear in an external fasta file
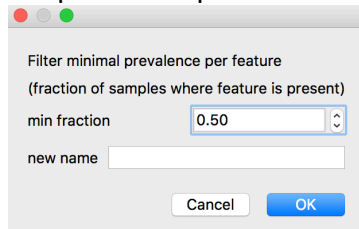

Mandatory fields:

- "*Fasta file*": Name of the fasta file containing the feature IDs (usually sequences or qiime2 hashes).

Optional fields:

- "*Negate*": Check to remove matching features, uncheck to keep matching features.
- "*New name*": the name for the experiment in the main list.

Filter prevalence

Keep features present at least in a given fraction of the samples (common features)
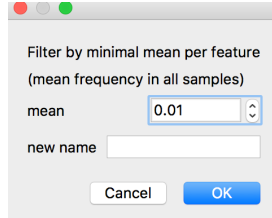

Mandatory fields:

- "*min fraction*": the minimal fraction of the samples where the feature is present.

Optional fields:

- "*New name*": the name for the experiment in the main list.

Filter mean

Keep features with a large enough mean frequency (over all samples)

Mandatory fields:
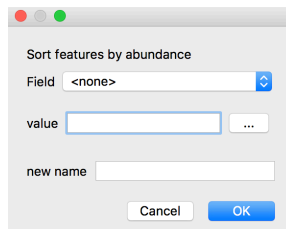
- "*mean*": the minimal mean frequency (over all samples) for features to be kept.

Optional fields:

- "*New name*": the name for the experiment in the main list.

Sort abundance

Order the features based on their mean frequency over (an optional subset of) samples.
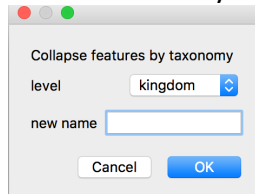
Optional fields:

- "*Field*": <none> to sort based on frequency over all samples, otherwise sort based on frequency only based on samples matching "value" in "Field".
- "*value*": The value to use for the sample subset.
- "*New name*": the name for the experiment in the main list.

Collapse taxonomy

Merge features based on their taxonomy (summing the frequencies of all features with the same taxonomy).

NOTE: taxonomy information must be embedded in the biom table.

Mandatory fields:

- "*level*": The taxonomic level to merge by (i.e. Phyla, Genus, etc.)

Optional fields:

- "*New name*": the name for the experiment in the main list.

Analysis tab

Buttons from this tab are used for statistical analysis of the experiment. Button actions are performed on the selected experiment from the main list. Most actions generate a new experiment in the main list.
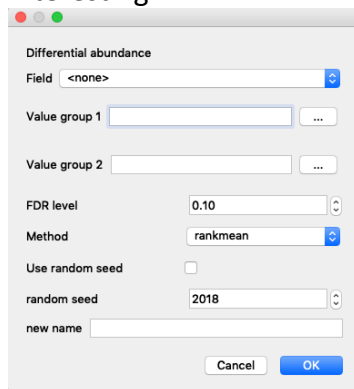
Diff. abundance

Find statistically significant features separating two subsets of samples from the experiment. All the tests are non-parametric with dsFDR multiple hypothesis control.

The result is a new experiment, containing only the significant features, sorted by the effect size (topmost is the highest in group1, bottommost is the highest in group2). In order to view the significant bacteria, just plot this new experiment.

NOTE: There is no special separation marker between the last feature higher in group1 and the first feature higher in group2, since we believe if you don't see it by eye, it's not that interesting.

Mandatory fields:

- "*Field*": Name of the field used to separate the samples into two groups.
- "*Value group 1*": the value used for samples in group1. Can select multiple values using the "…" button.
- "*Value group 2*": the value used for samples in group2. Can select multiple values using the "…" button. If no values are selected, all samples not in group1 are used for group2.
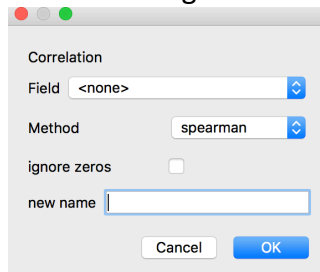
Optional fields:

- "*FDR level*": The upper bound on features where the null hypothesis holds (similar to Benjamini-Hochberg q value).
- "*Method*": The statistic used for the comparison between the two groups. Options are:
    - "*rankmean*": rank each feature over the samples and compare the mean between the two groups. Robust to outliers.
    - "*mean*": just compare the means of the two groups for each feature. Can be strongly affected by outliers.
    - "*binary*": Compare the fraction of samples where the feature is present.
- "*Use random seed*": If checked, use the random seed given in the "*random seed*" field to initialize the permutations used for the p-value calculation. Checking this field ensures obtaining the exact same results every time the test is run on the same *experiment*.

- *"random seed"*: the value to be used as the random seed. Requires *"Use random seed"* to be checked.
- "*New name*": the name for the experiment in the main list.

Correlation

Find statistically significant features correlated with a metadata field. All the tests are non-parametric with dsFDR control.

Result is a new experiment, containing only the significant features, sorted by the effect size (topmost is the highest correlated feature, bottommost is the highest anti-correlated). In order to view the significant bacteria, just plot this new experiment.

Mandatory fields:

- "*Field*": Name of the field used to look for correlation.

Optional fields:

- "*Method*": The statistic used for the comparison between the two groups. Options are:
  - "*spearman*": spearman rank correlation (rank samples and each feature over samples and then calculate pearson correlation).
  - "*pearson*": pearson correlation between feature frequency and metadata field.
  - "*ignore zeros*": For each feature, calculate the correlation only on samples where the feature is present (NOTE: much slower).
- "*New name*": the name for the experiment in the main list.

Used only for microbiome data. Find dbBact (dbbact.org) terms enriched in features from the diff. abundance test.

Shows the dbBact terms which are significantly more associated with features higher in group2 or group2 (from the diff. abundance test).

NOTE: can only be used on experiments which are the result of diff. abundance.
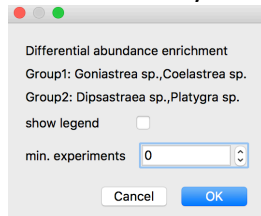
Optional fields:

- "*show legend*": check this box to show the name of each group in the resulting bar plot
- "*min. experiments*": Keep only terms appearing in at least min. experiments different experiments (i.e. if set to 2, ignore terms observed only in one dbBact experiment).

Result are a bar plot showing terms enriched in either of the two groups, and a list of the enriched terms which enables interactive exploration of each term.

Example:

Taking the Chronic Fatigue Syndrome dataset from (Giloteaux, Ludovic, et al. "Reduced diversity and altered composition of the gut microbiome in individuals with myalgic encephalomyelitis/chronic fatigue syndrome." *Microbiome* 4.1 (2016): 30), we first look for bacteria different between healthy controls and chronic fatigue syndrome patients using diff. abundance:
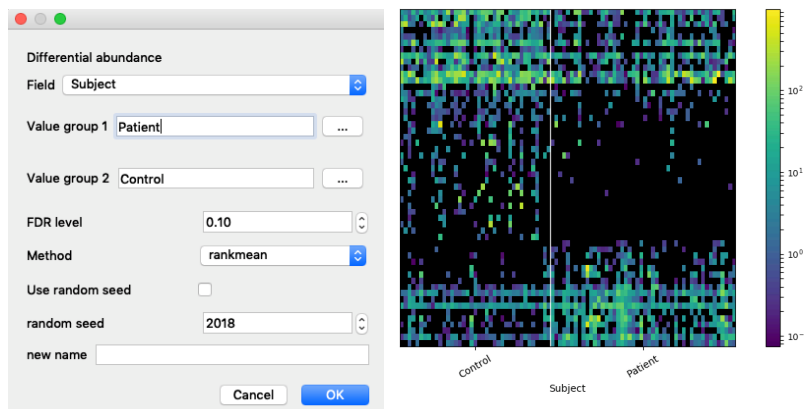

The resulting diff. abundance experiment contains 54 bacteria different between patients and controls. We now want to see what dbBact terms are significantly more common in the bacteria higher in the "control" group and what dbBact terms are more common in the bacteria higher in "Patient" group. This is done using the "enrichment" button.

The enriched terms bar plot:
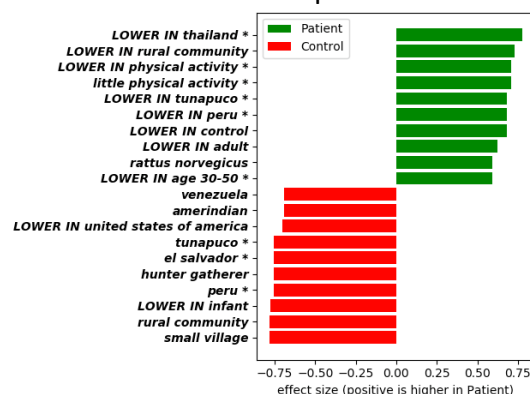
This indicates bacteria from the "higher in Patient" group contain significantly more annotations such as "LOWER IN physical activity" or "LOWER IN thailand" compared to the "higher in control" bacteria, and conversely, bacteria from the "higher in control" group have more dbBact annotations such as "small village" or "rural community".

In addition, we get an interactive list of terms higher in each group:
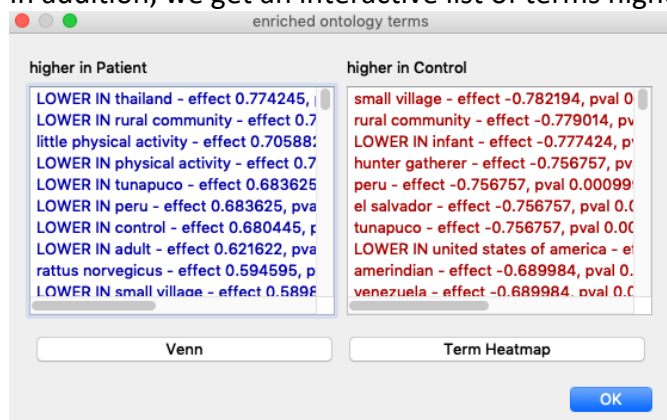

Terms in the left list are significantly higher in bacteria from the "higher in Patient" group compared to the "higher in control" group bacteria, and the right list contains terms higher in the "higher in control" bacteria group. Each term is followed by the effect size and dsFDR corrected p-value for the enrichment.

After selecting a term from either of the two lists, we can plot a heatmap of Venn-diagram showing the distribution of the term across bacteria from the two groups:
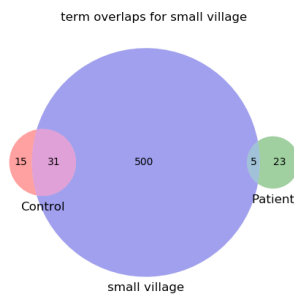
The Venn diagram for "small village" shown 31/46 (31+15) bacteria from the "higher in control" group are associated with the "small village" dbBact term, whereas only 5/28 bacteria from the "higher in patients" group are associated with the same term.
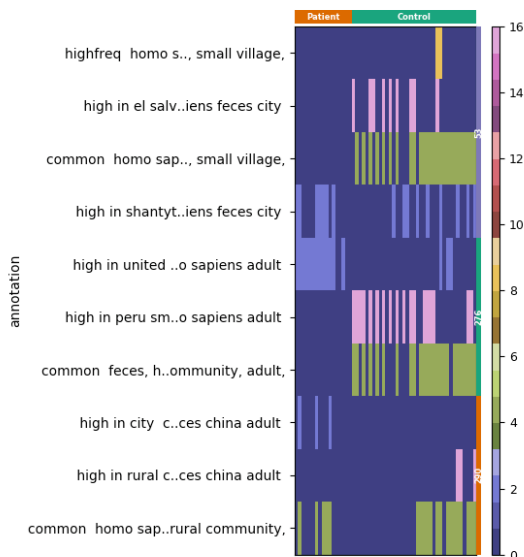
The heatmap for the "small village" term shows bacteria as COLUMNS (so we have the left columns under the "patient" orange bar represent the bacteria from the "higher in patient" group, and the right columns under the "Control" green bar represent the bacteria from the "higher in control" group). Each ROW is a dbBact annotation containing the "small village" term. The annotation rows are sorted according to the dbBact experiment they originate from (the vertical bars on the right side – each number such as 290/276/53 represent a dbBact experiment).

Clicking on a row shows details about the annotation. For example, the 3rd row from the top reads: "common homo sapiens, feces, city, el salvador, small village," indicating the colored bacteria in this row were present in this annotation.

The colors represent the annotation type (beige is common, pink is higher in small village compared to some other term, blue is lower in small village compared to some other term).