omicplotR: A tool for visualization of omic datasets as compositions

Daniel Giguere

2017-11-21

Contents

1 What is omicplotR?

`omicplotR` is an R package containing a Shiny app used to visually explore omic datasets, where the input is a table of read counts from high-throughput sequencing runs. It integrates the `ALDEx2` package for compositional data analysis of differential abundance. The user interface facilitates easy and fast visualization of high-throughput sequencing data with several graphical options for both users with and without experience at the R command line. `omicplotR` speeds up the process of exploring an omic dataset by providing a graphical user interface for the user.

2 Introduction

High-throughput sequencing (HTS) instruments generate an amount of sequencing reads that is constrained by limitations of the sequencing instrument itself, and do not represent the absolute number of DNA molecules in a sample. This type of data, which is constrained by an arbitrary or constant sum, is referred to as compositional data. Software tools developed for differential abundance analysis are typically command line based such as `DESeq2`, `ALDEx2`, `EdgeR`, which make them difficult for users without a background in scripting. `omicplotR` is a graphical user interface for data exploration, and also incorporates `ALDEx2` for compositional differential abundance analysis. `ALDEx2` processes data using methods that are appropriate for compositional data.

This guide provides an overview of the R package `omicplotR`, and the input requirements. A tutorial for each component of the `Shiny` app is available on my GitHub: https://github.com/dgiguer/omicplotR/wiki. omicplotR was developed for several types of HTS datasets including RNASeq, meta RNASeq, and 16s rRNA gene sequencing, but can also be used for nearly any type of data generated by HTS which generates a tables of counts per feature for each sample.

3 Features

omicplotR provides a graphical user interface using the `Shiny` package for the following visualizations for HTS data:

- Principal Component Analysis (PCA) biplots
- Coloured PCA biplots by metadata
- Association plots using `igraph`
- Dendrograms
- Stacked barplots of taxonomic relative abundance
- Compositional differential abundance (effect and Bland-Altmant plots using `ALDEx2`)

Additionional features include:

- Filtering count tables per sample or feature (PCA biplots)
- Filtering data into groups for PCA biplots using metadata
- Visualizing which features and samples removed from analysis
- Colour PCA biplots using metadata (continuously, by quartile, categorical)
- Generate effect plots between conditions of associated metadata
- Interact with effect plots to visualize difference between groups
- Plot pre-calculated `ALDEx2` tables to colour points by matching search terms.

4 Installation

Install the latest version of `omicplotR`. Make sure you have the newest version of R, `ALDEx2`, and the other dependancies. `omicplotR` requires you to have R version 3.4 at a minimum.

First, load the omicplotR package. All other dependencies will be loaded automatically. Launch the `Shiny` app in your default browser.

```
library(omicplotR)
omicplotr.run()
```

Included datasets can be viewed by loading them.

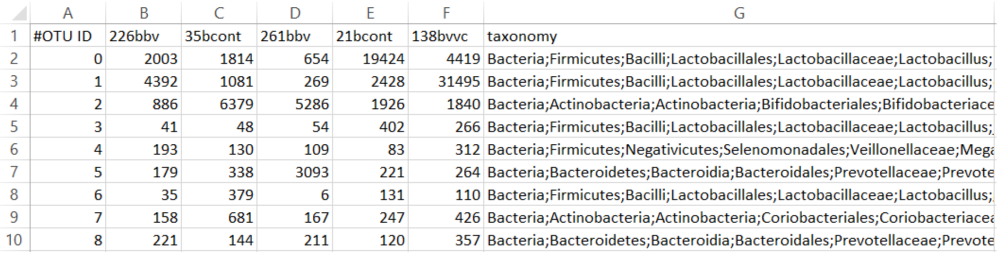
```
data(otu_table)
data(metadata)
```

5 Input requirements

For both metadata and data, each sample name and operational taxonomic unit (OTU) name *must be unique*.

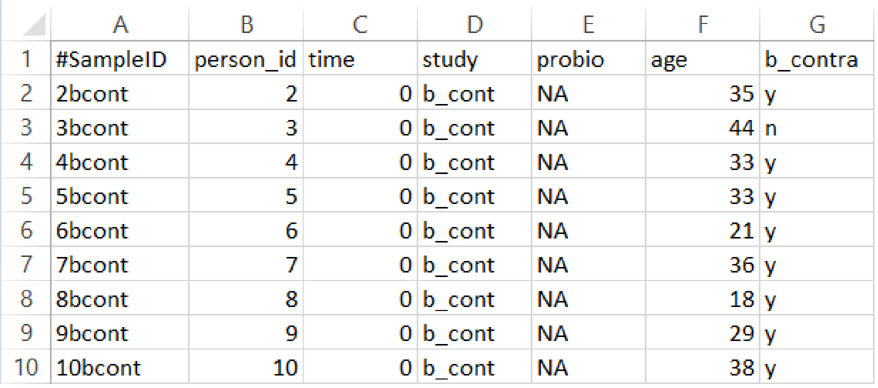
5.1 Data

1. The data file must be a **tab delimited .txt** (this is an option when you click 'Save as' from Excel).
2. The first column must contain gene/OTU identifier.
3. The first row must contain sample identifiers.



	A	B	C	D	E	F	G
1	#OTU ID	226bbv	35bcont	261bbv	21bcont	138bvvc	taxonomy
2	0	2003	1814	654	19424	4419	Bacteria;Firmicutes;Bacilli;Lactobacillales;Lactobacillaceae;Lactobacillus;
3	1	4392	1081	269	2428	31495	Bacteria;Firmicutes;Bacilli;Lactobacillales;Lactobacillaceae;Lactobacillus;
4	2	886	6379	5286	1926	1840	Bacteria;Actinobacteria;Actinobacteria;Bifidobacteriales;Bifidobacteriace
5	3	41	48	54	402	266	Bacteria;Firmicutes;Bacilli;Lactobacillales;Lactobacillaceae;Lactobacillus;
6	4	193	130	109	83	312	Bacteria;Firmicutes;Negativicutes;Selenomonadales;Veillonellaceae;Mega
7	5	179	338	3093	221	264	Bacteria;Bacteroidetes;Bacteroidia;Bacteroidales;Prevotellaceae;Prevote
8	6	35	379	6	131	110	Bacteria;Firmicutes;Bacilli;Lactobacillales;Lactobacillaceae;Lactobacillus;
9	7	158	681	167	247	426	Bacteria;Actinobacteria;Actinobacteria;Coriobacteriales;Coriobacteriace
10	8	221	144	211	120	357	Bacteria;Bacteroidetes;Bacteroidia;Bacteroidales;Prevotellaceae;Prevote

Figure 1: Example data. If taxonomy column introduced, it must use the column name ’taxonomy’.



	A	B	C	D	E	F	G
1	#SampleID	person_id	time	study	probio	age	b_contra
2	2bcont	2	0	b_cont	NA	35	y
3	3bcont	3	0	b_cont	NA	44	n
4	4bcont	4	0	b_cont	NA	33	y
5	5bcont	5	0	b_cont	NA	33	y
6	6bcont	6	0	b_cont	NA	21	y
7	7bcont	7	0	b_cont	NA	36	y
8	8bcont	8	0	b_cont	NA	18	y
9	9bcont	9	0	b_cont	NA	29	y
10	10bcont	10	0	b_cont	NA	38	y

Figure 2: Example metadata file. Metadata maybe be numerical or categorical. Any blank spaces will be replaced as NA when importing the file.

4. The last column may contain taxonomic level information, but is not required. If present, it must be labelled ‘taxonomy’. **Taxonomy column must have at least four levels, separated by a semi colon.**
5. Data table must have all blank rows removed (this may require you to check in a text editor like *Notepad ++* or *Atom* before using the app)

An example of an appropriate data file is shown in Figure 1.

5.2 Metadata

1. Must be a **tab delimited .txt** (this is an option when you click ‘Save as’ from Excel).
2. The first column must contain sample identifiers. The sample identifiers must be identical to the data file (not required to be in the same order).
3. The first row must contain phenotypic information.

An example of an appropriate metadata file is shown in Figure 2.

6 Coloured Principal Component Analysis (PCA) plot example

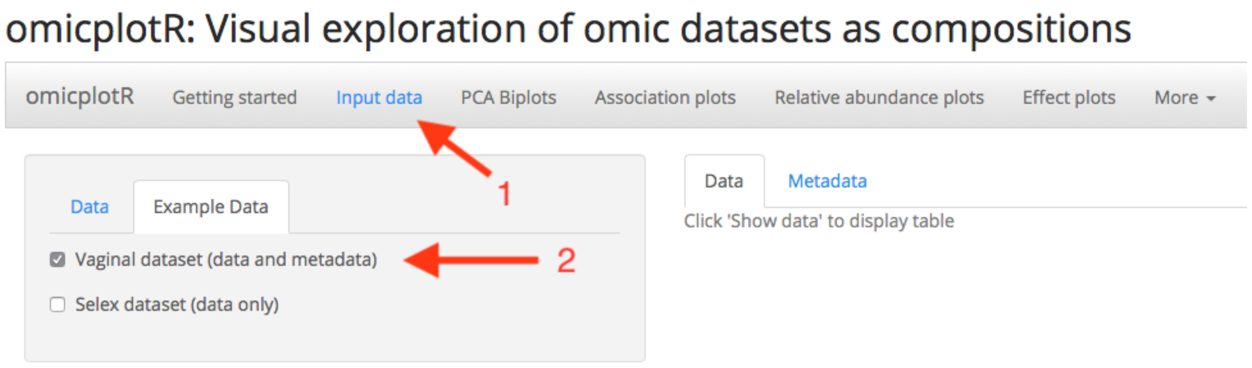
The PCA plot can show whether your samples separate into groups, what taxa (or groups of tax) are driving this separation, and what taxa are irrelevant for your analysis. More information about how to interpret PCA plots and how they can be used can be found by clicking here.

Metadata is not required to use the biplot function, but is necessary to use the Colored Biplot function. We will be using the example dataset (‘Vaginal dataset’ available under *Input data* and *Example data* tab) for this vignette.

First launch the `Shiny` app.

```
library(omicplotR)
omicplotr.run()
```

We will be loading the example dataset by clicking 1. 'Input data' and 2. 'Example data'. Alternatively you can load your data and metadata by clicking the 'Browse…' button under the 'Input data' tab. **Please ensure the input requirements are met before uploading your own data**. You can click 'Show data' under the 'Data' tab to view your data and metadata in a table format.


omicplotR: Visual exploration of omic datasets as compositions

☑ Vaginal dataset (data and metadata)

□ Selex dataset (data only)

Click 'Show data' to display table

6.1 Filtering options

There are several ways you can filter your data before a biplot is made. Each input is independent of each other, so you can filter by several different methods. Presence (or absence) of a taxonomy column does not influence the filtering. We will use the default settings for this vignette.

1. **Minimum count per OTU**
   This filter removes any rows with a maximum sample count less than the input. Default is 10 to remove sparse OTUs.
2. **Minimum count sum per sample**
   This filter removes any columns that have a lower sum of counts than the filter. You can visualize which samples are being removed when applying this filter using the *Filtering by counts per sample* graph under the *Removed data* tab.
3. **Minimum proportional abundance**
   This filter calculates frequency for each sample (by column) after filters *1.* and *2.* are applied, and removes any rows that have no frequencies above the threshold.
4. **Maximum proportional abundance**
   This filter calculates frequency for each sample, and removes any rows that have no frequencies below the threshold.
5. **Minimum count sum per OTU**
   This filter removes any rows that have a lower sum of counts than the filter. You can visualize which rows are removed with the *Filtering by counts per row* graph under the *Removed data* tab.

Variance cutoff

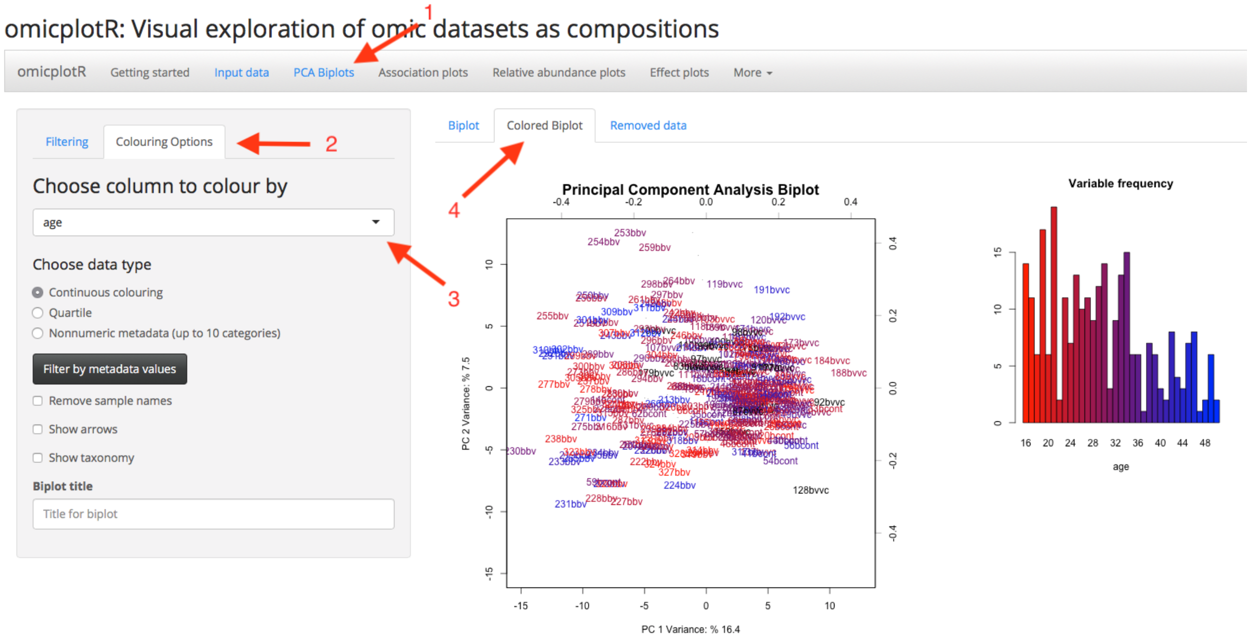
This filter calculates the variance for each OTU (row), and removes any rows that have lower variance than the filter. Variance is calculated after all other filters, and after the reads have been transformed by the centre-log ratio.

Adjust scale

This changes the scale of the biplot. When set to zero, it shows the relationship between samples (columns), while being set to 1 shows the relationship between OTUs (rows). Data is not filtered from this operation.

6.2 Generating coloured PCA biplot

The coloured PCA biplot uses an associated metadata file to colour the sample names. Click on 1. 'PCA biplots' 2. Click on 'Colouring Options' 3. Choose a numerical column for continuous colouring (age). For categorical metadata (in the example metadata, 'probio'), make sure to click on the 'nonnumeric metadata' radio button. 4. Click on 'Coloured Biplot' to generate the coloured biplot. A histogram displaying the frequency of each samples age is also generated.



omicplotR: Visual exploration of omic datasets as compositions

Choose column to colour by: age

Choose data type:
- ● Continuous colouring
- ○ Quartile
- ○ Nonnumeric metadata (up to 10 categories)

- □ Remove sample names
- □ Show arrows
- □ Show taxonomy

Biplot title: Title for biplot

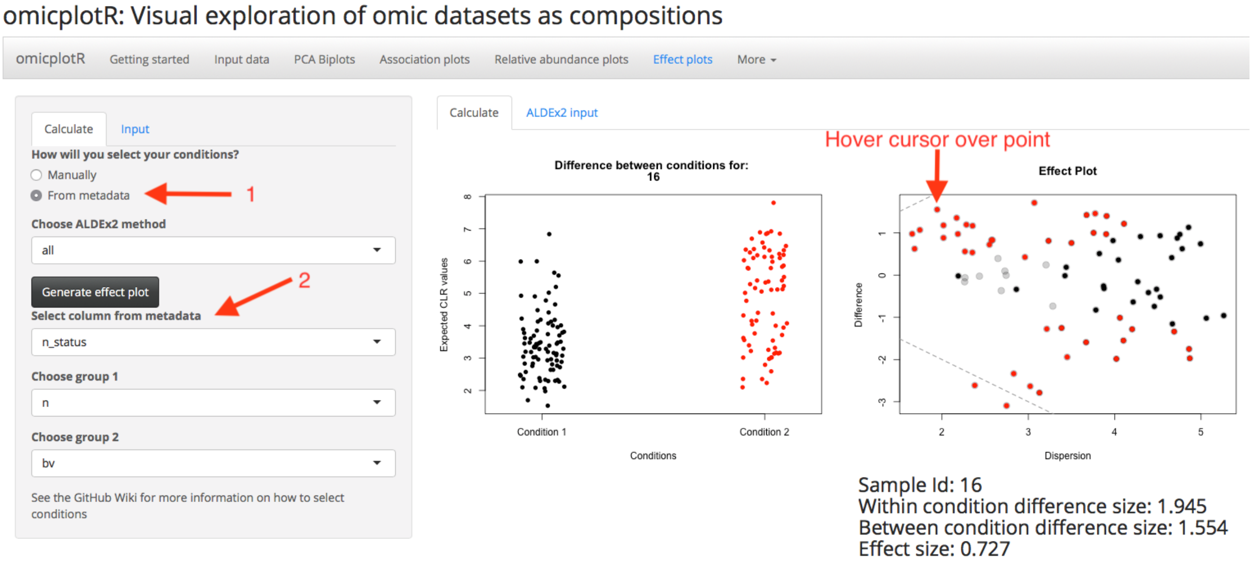
We will continue using the example dataset to generate effect plots using ALDEx2

7 Generating ALDEx2 effect plots

ALDEx2 is an R package for analyze differential abundance of HTS data as compositions.

To generate effect plots, click on the 'Effect plots' tab. Here, you can choose to calculate the ALDEx2 manually (ie, choosing which columns to compare) or by using metadata. For this example, we will compare the nugent status of the samples.

Select the 'n_status' column from metadata. For Group 1 choose 'n', and for Group 2 choose 'bv'. After these have been selected, click the 'Generate effect plot' button. For large datasets, this will take a long time. It takes about 10 seconds to calculate for the example data, which has 77 features for 297 samples. By hovering over a point on the graph, a strip chart of the expected CLR abundance (see ALDEx2) for each sample will be shown, allowing you to compare the differences between your samples for a given OTU. It also displays information from the effect table generated by ALDEx2



omicplotR: Visual exploration of omic datasets as compositions

How will you select your conditions?

○ Manually

● From metadata

Choose ALDEx2 method: all

Select column from metadata: n_status

Choose group 1: n

Choose group 2: bv

See the GitHub Wiki for more information on how to select conditions

Sample Id: 16
Within condition difference size: 1.945
Between condition difference size: 1.554
Effect size: 0.727

If you are using your own data and want to select groups by columns, you will need to reorder your file to have the first n columns as one group, and the last n columns as the other.

For more information about how to use `omicplotR`, please visit my Github page:

https://github.com/dgiguer/omicplotR/wiki