OmicsBox User Manual

version 1.2

11/07/2019

BioBam Bioinformatics S.L.

Table of Contents

1 Introduction to OmicsBox - Bioinformatics Made Easy

OmicsBox is a leading bioinformatics platform for the analysis of omics data.

OmicsBox offers user-friendly data analysis which allow gaining biological insights fast and easy even for completely novel genomes.

OmicsBox is a desktop application for industry, academic and governmental research biologists.

OmicsBox provides a robust platform for the execution of complex and demanding bioinformatics pipelines from a normal desktop PC.

OmicsBox allows designing bioinformatics workflows, run state-of-the-art tools and to visualize results in a very easy and non-techie fashion. This reduces the learning curve significantly, allows to obtain results faster and in a more reliable and reproducible way.

OmicsBox has minimal setup requirements and provides automatic updates. Advanced functionality is instantly available requiring no high-performance computing facility, computational background or maintenance.

OmicsBox Modules: At the moment the following modules are available:

- Genome Analysis
- Transcriptomics
- Functional Annotation and Analysis
- Metagenomics

Splash Screen v.1.2

2 About BioBam

- **BioBam's** solutions accelerate research in disciplines such as agricultural genomics, microbiology and environmental NGS studies; amongst others.
- **BioBam** is internationally recognized as a leader in functional genomics, which is demonstrated by over 10.000 scientific research citations.
- **BioBam** Bioinformatics develops and maintained software like OmicsBox, OmicsCloud and Blast2GO.
- **BioBam** is a bioinformatics solution provider (software, cloud solutions, consulting and analysis services) with headquarters in Spain.

Logo: BioBam Bioinformatics S.L.

3 Installation and Activation and User Interface

Content of this page:

3.1 Activation

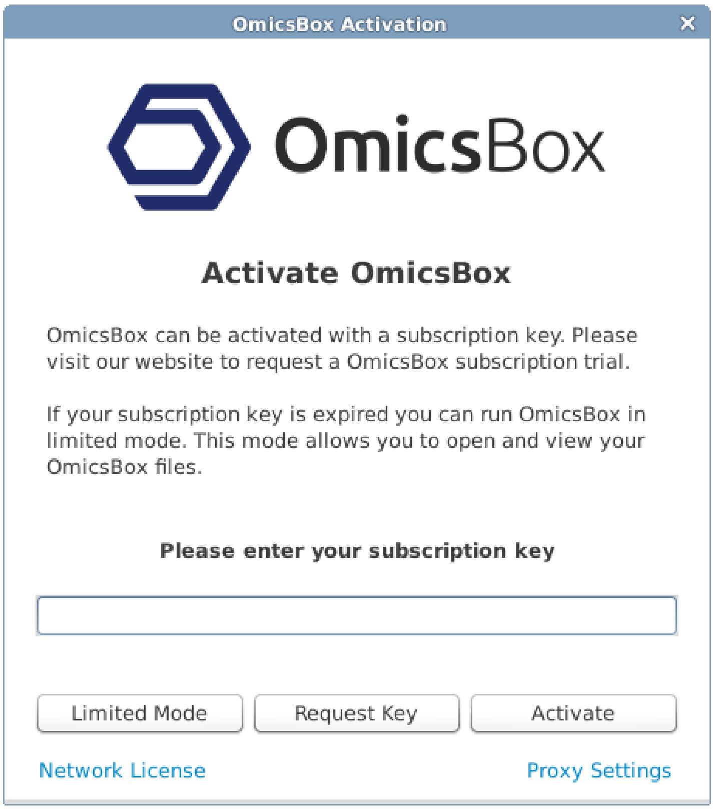
On startup and if not yet provided, OmicsBox asks for activation. It is possible to use OmicsBox in 2 different modes:

- Limited mode. This mode allows to load and visualize projects in OmicsBox.
- Full mode. To activate OmicsBox in Full mode, the corresponding subscription has to be obtained via the OmicsBox website. This mode provides access to all features of the purchased modules, Cloud services, updates and support. More details about all the advantages to be a Full user can be found online. A free Trial account can be requested via the OmicsBox website.

To activate OmicsBox the software requires access to the Internet. If necessary, proxy settings can be directly adjusted directly from the activation dialog.

3.2 Network License

OmicsBox can also be activated with the license server. Various clients can share a pool of licenses on a license server in the local network. To activate a network license, click **Network License** on the bottom left in the activation dialog and provide the license server IP address and cloud key if you own one. The cloud key allows to use CloudBlast and CloudIPS, all the other cloud features are already available without a key.



OmicsBox Activation

Activate OmicsBox

OmicsBox can be activated with a subscription key. Please visit our website to request a OmicsBox subscription trial.

If your subscription key is expired you can run OmicsBox in limited mode. This mode allows you to open and view your OmicsBox files.

Please enter your subscription key

Figure 1: Activation dialog of OmicsBox: Full or Limited Mode

3.3 The Desktop Application

This section describes briefly the main parts of the OmicsBox applications.

3.3.1 Main User Interface

The general components of the OmicsBox main user interface are:

1. **Menu Bar:** The main application menu contains 3 items:
   - File: This menu groups all functions for opening, saving and closing OmicsBox files as well as to load or export data in many different formats.
   - View: This menu allows to open different utility taps like the file manager, the application messages or the Java memory monitor.
   - Help: The help menu provides access to various support features. You can review your subscription details, the CloudBlast history and the App-Manager. The App-Manager allows to install and update additional tools for OmicsBox.

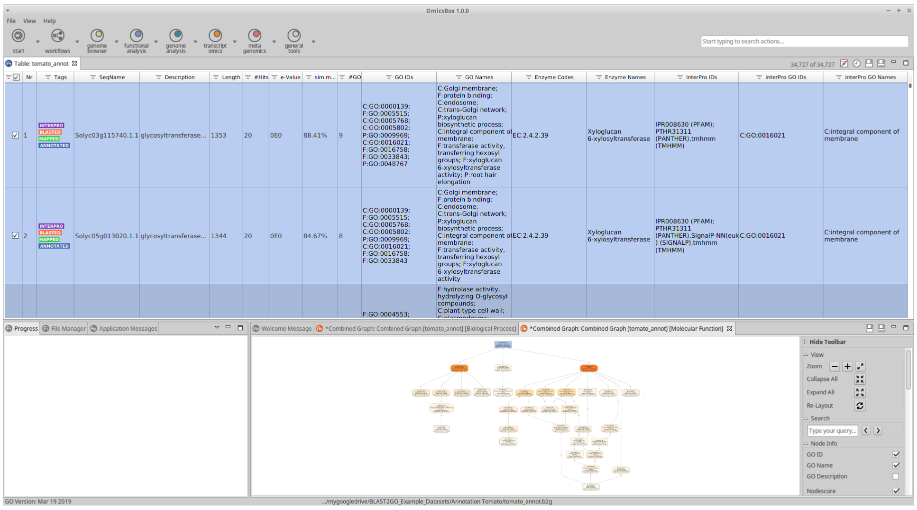
2. **Main Analysis Icons:** These will execute the whole analysis.
   - start: Quick start to load an OmicsBox Project or recent projects or fasta sequences or annotation file.
   - workflows: Allows to run workflows to execute analysis in a semi-automatic way.
   - genome browser: Allows to visualize genomic structures from different files types in a side-scrolling way.

- functional analysis module: Contains functions for functional characterization of sequences: Blast, InterProScan, GO Mapping, Functional Annotation, Coding Potential and Enrichment Analysis. In addition, charts can be generated to offer the researcher an overview of the results obtained in each step to facilitate the decision for parameter choice in latter annotation steps.
- genome analysis module: Allows to characterize and analyze newly sequenced genomes, from raw reads to gene structures. It contains actions for Eukaryotic and Prokaryotic Gene Finding, Repeat Masking and DNA-Seq de novo Assembly.
- transcriptomics module: Allows to process RNA-seq data from raw reads down to their functional analysis. It incorporates functions for RNA-Seq de novo Assembly, RNA-Seq Alignments, Create Count Tables and perform Differential Expression Analyses.
- metagenomics module: Allows to combine and integrate all necessary steps for a complete metagenome analysis: Taxonomic Classification, Metagenomic Assembly and Gene Prediction, Functional Annotation and Comparative Analysis.
- general tools: Includes different general actions: draw Venn Diagrams, obtain Tag Statistics charts and tools to perform quality control and pre-processing of Fastq files.

3. **Application Tabs:**
   - Progress tab: Shows running, queued and finished jobs. Allows to stop jobs and open a separate log-viewer. Each finished job shows the time it took to execute. A finished job can be removed using the cross button. All finished jobs can be removed by clicking on the small triangle in the top left corner.
   - Application Messages tab: Displays information on the progress of the analysis.
   - File Manager: The File Manager allows to view, open and organize different OmicsBox related files.
   - Java Memory Monitor: Provides information about the memory consumption of the application.
   - Log tabs: Most of the analysis steps provide additional logging. These logs can be viewed via the Progress tab.

4. **Data Tabs:** Each object type can be opened and saved in a separate tab. In general, table viewer tabs like e.g. the OmicsBox sequence table will be opened by default in the upper-wide area while result tabs will be opened in the lower-right area (e.g. charts, graphs, etc.).

The OmicsBox File Manager allows to view, organise and open different OmicsBox related files.
The "Merge" option allows combining various file into one. The merge function may not be available for all data types and can only be used to object of the same type.



☑	Nr	Tags	SeqName	Description	Length	#Hits	e-Value	sim m...	#GO	GO IDs	GO Names	Enzyme Codes	Enzyme Names	InterPro IDs	InterPro GO IDs	InterPro GO Names
☑	1	INTERPRO BLASTED MAPPED ANNOTATED	Solyc03g115740.1.1	glycosyltransferase...	1353	20	0E0	88.41%	9	C:GO:0000139; F:GO:0005515; C:GO:0005768; C:GO:0005802; P:GO:0009969; C:GO:0016021; F:GO:0016758; F:GO:0033843; P:GO:0048767	C:Golgi membrane; F:protein binding; C:endosome; C:trans-Golgi network; P:xyloglucan biosynthetic process; C:integral component of membrane; F:transferase activity, transferring hexosyl groups; F:xyloglucan 6-xylosyltransferase activity; P:root hair elongation	EC:2.4.2.39	Xyloglucan 6-xylosyltransferase	IPR008630 (PFAM); PTHR31311 (PANTHER),tmhmm (TMHMM)	C:GO:0016021	C:integral component of membrane
☑	2	INTERPRO BLASTED MAPPED ANNOTATED	Solyc05g013020.1.1	glycosyltransferase...	1344	20	0E0	84.67%	8	C:GO:0000139; F:GO:0005515; C:GO:0005768; C:GO:0005802; P:GO:0009969; C:GO:0016021; F:GO:0016758; F:GO:0033843	C:Golgi membrane; F:protein binding; C:endosome; C:trans-Golgi network; P:xyloglucan biosynthetic process; C:integral component of membrane; F:transferase activity, transferring hexosyl groups; F:xyloglucan 6-xylosyltransferase activity	EC:2.4.2.39	Xyloglucan 6-xylosyltransferase	IPR008630 (PFAM); PTHR31311 (PANTHER),SignalP-NN(euk) (SIGNALP),tmhmm (TMHMM)	C:GO:0016021	C:integral component of membrane
										F:GO:0004553;	F:hydrolase activity, hydrolyzing O-glycosyl compounds; C:plant-type cell wall;					

GO Version: Mar 19 2019

.../mygoogledrive/BLAST2GO_Example_Datasets/Annotation Tomato/tomato_annot.b2g

Figure 2: OmicsBox Main User Interface

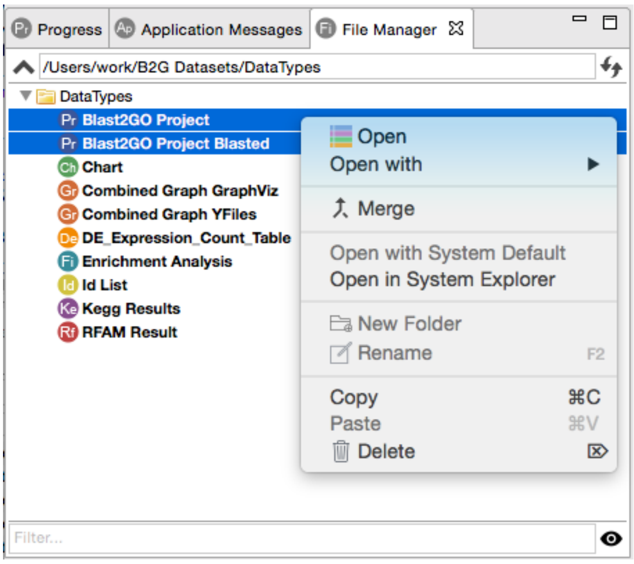
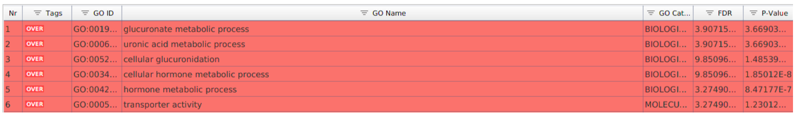


Figure 3: File Manager menu

3.3.2 Table

The table allows opening different types of objects (.box/.b2g files) in a spreadsheet style. Additionally to the columns representing the dataset, it contains a TAG column which represents the different status of each line according to the object type.

The table allows to hide or show single columns via a checkbox menu by right-clicking on the column header. In case, multiple objects are viewed (e.g. OmicsBox plus Rfam results), column headers are colored differently. In this example (Figure 4) tags show if a functionally enriched GO term is over or under-represented.



Nr	Tags	GO ID	GO Name	GO Cat...	FDR	P-Value
1	OVER	GO:0019...	glucuronate metabolic process	BIOLOGI...	3.90715...	3.66903...
2	OVER	GO:0006...	uronic acid metabolic process	BIOLOGI...	3.90715...	3.66903...
3	OVER	GO:0052...	cellular glucuronidation	BIOLOGI...	9.85096...	1.48539...
4	OVER	GO:0034...	cellular hormone metabolic process	BIOLOGI...	9.85096...	1.85012E-8
5	OVER	GO:0042...	hormone metabolic process	BIOLOGI...	3.27490...	8.47177E-7
6	OVER	GO:0005...	transporter activity	MOLECU...	3.27490...	1.23012...

Figure 4: Table

3.3.2.1 Table Context Menu

The context menu allows to create ID lists of a selected column, extract subsets of entries as well as to copy part of the content into the clipboard. The values of a given column can also be visualized as a category or distribution plot in various formats (bar-chart, pie-chart, etc.).

- **Show Sequence:** Allows seeing the sequence information.
- **Show Blast Result:** Revise the Blast results in detail:, hits, species, identifiers, etc.
- **Show InterProScan Result:** Revise the InterProScan results in detail: databases, GOs, etc.
- **Show Mapping Result:** Revise the retrieved candidate GO terms in detail: Accessions, Evidence Codes, Databases.
- **Show GO Description:** Review information about annotated GOs.
- **Copy Sequence Names to Clipboard:** Copies the sequence name to the clipboard.
- **Change Annotation and Description:** Allows to manually change/add annotations.
- **Annotate Sequence:** This function allows changing annotation parameters for the selected sequence and re-running automatic annotation.
- **Make Graph of GO-Mapping-Results with Annotations Score:** Displays a DAG with all GO terms related to one sequence. Shows all the GOs from the mapping step as well as final annotations (highlighted).
- **Extract Selection to new Tab:** Create a new project from the marked sequences (Shift or Ctrl).
- **Copy Selection to Clipboard (tabular format):** Copies the marked sequence to the clipboard in tabular format for further processing in a spreadsheet editor.
- **Copy Content of Column: SeqName to Clipboard:** The content of a specific column will be copied to the clipboard.
- **Create ID List of Column: Sequence:** Allows creating an ID list of a specific column which can then be used for Fisher's Exact Test or Selection of sequences (see Figure 6).
- **Create ID Value-List of SeqName and Length:** Allows creating a list with two specific columns e.g. SeqName and Length.
- **Create Category Chart of Column: Length:** Create a category chart of a certain column e.g. sequence length.
- **Create Distribution Chart of Column:** InterProScan IDs: Create a distribution chart of a certain column.

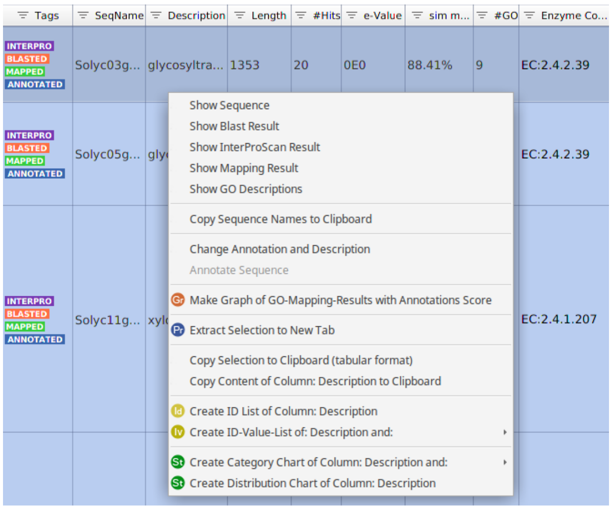

Tags
SeqName
Description
Length
#Hits
e-Value
sim m...
#GO
Enzyme Co...
INTERPRO
BLASTED
MAPPED
ANNOTATED
Solyc03g...
glycosyltra...
1353
20
0E0
88.41%
9
EC:2.4.2.39
Solyc05g...
EC:2.4.2.39
Solyc11g...
EC:2.4.1.207
Show Sequence
Show Blast Result
Show InterProScan Result
Show Mapping Result
Show GO Descriptions
Copy Sequence Names to Clipboard
Change Annotation and Description
Annotate Sequence
Make Graph of GO-Mapping-Results with Annotations Score
Extract Selection to New Tab
Copy Selection to Clipboard (tabular format)
Copy Content of Column: Description to Clipboard
Create ID List of Column: Description
Create ID-Value-List of: Description and:
Create Category Chart of Column: Description and:
Create Distribution Chart of Column: Description

Figure 5: Sequence Context Menu

Create ID lists:

To generate an ID list the desired sequences have to be marked and then right click on "Create ID list from selected Entries (SeqName column)".
A new tab is opened containing the ID list of the marked sequence names. This list can now be saved for further use (e.g. enrichment analysis, selections, etc.). It is also possible to generate sequence ID lists from Graphs (see Quantitative Analysis(see page 80)).

To mark all the sequences on the table do: Ctrl + A (Windows/ Linux) and Apple + A (Mac).

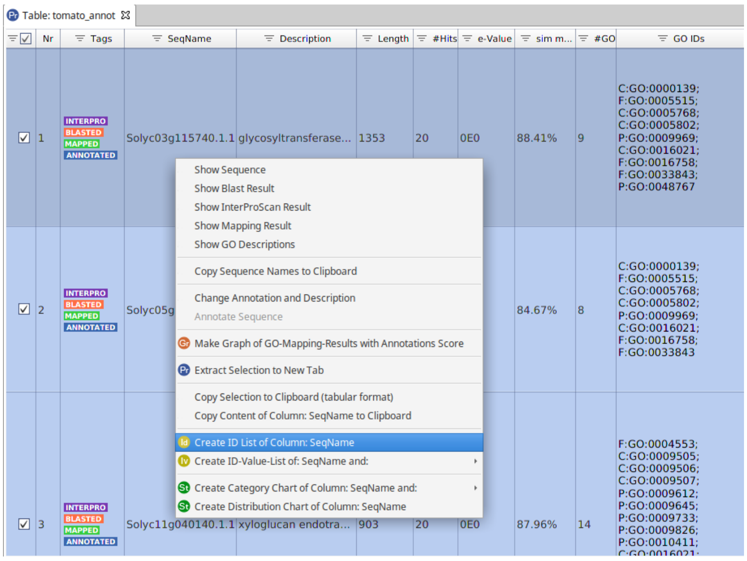


Table: tomato_annot
Nr
Tags
SeqName
Description
Length
#Hits
e-Value
sim m...
#GO
GO IDs
☑ 1
INTERPRO
BLASTED
MAPPED
ANNOTATED
Solyc03g115740.1.1
glycosyltransferase...
1353
20
0E0
88.41%
9
C:GO:0000139; F:GO:0005515; C:GO:0005768; C:GO:0005802; P:GO:0009969; C:GO:0016021; F:GO:0016758; F:GO:0033843; P:GO:0048767
☑ 2
Solyc05g
84.67%
8
C:GO:0000139; F:GO:0005515; C:GO:0005768; C:GO:0005802; P:GO:0009969; C:GO:0016021; F:GO:0016758; F:GO:0033843
☑ 3
Solyc11g040140.1.1
xyloglucan endotra...
903
20
0E0
87.96%
14
F:GO:0004553; C:GO:0009505; C:GO:0009506; C:GO:0009507; P:GO:0009612; P:GO:0009645; P:GO:0009733; P:GO:0009826; P:GO:0010411; C:GO:0016021;
Show Sequence
Show Blast Result
Show InterProScan Result
Show Mapping Result
Show GO Descriptions
Copy Sequence Names to Clipboard
Change Annotation and Description
Annotate Sequence
Make Graph of GO-Mapping-Results with Annotations Score
Extract Selection to New Tab
Copy Selection to Clipboard (tabular format)
Copy Content of Column: SeqName to Clipboard
Create ID List of Column: SeqName
Create ID-Value-List of: SeqName and:
Create Category Chart of Column: SeqName and:
Create Distribution Chart of Column: SeqName

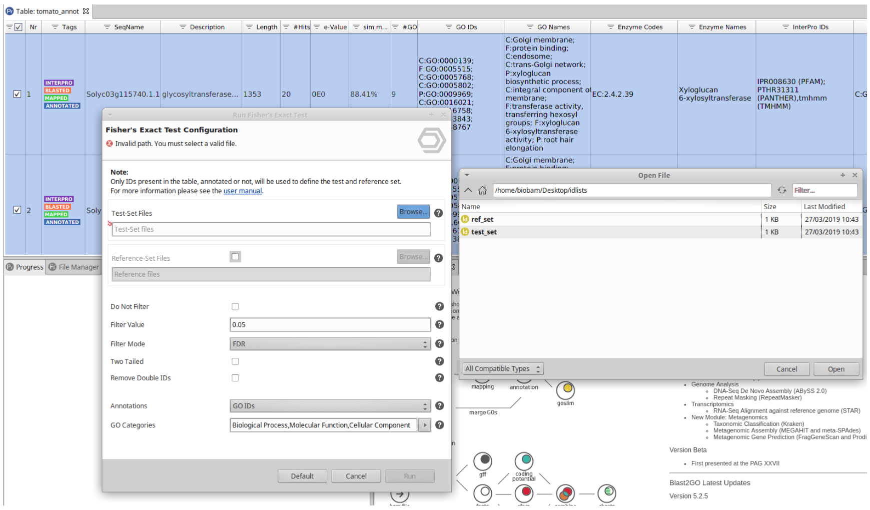
Figure 6: Create ID List from Sequence Name

The ID lists can be used in:

- Fisher Exact Test
- Select Sequences
- Load Annotations from BioMart (online)

The Gene Set Enrichment Analysis (GSEA) needs a ranked list and this can also be generated by OmicsBox with the Create ID Value List.

Here is an example of the Open File dialog of the Fisher's Exact Test Wizard shows the available ID lists if the selected file type in the checkbox (bottom left) is set to B2G ID list.



Fisher's Exact Test Configuration

Invalid path. You must select a valid file.

Note:
Only IDs present in the table, annotated or not, will be used to define the test and reference set.
For more information please see the user manual.

Test-Set Files: Test-Set files

Reference-Set Files: □ Reference files

Do Not Filter: □

Filter Value: 0.05

Filter Mode: FDR

Two Tailed: □

Remove Double IDs: □

Annotations: GO IDs

GO Categories: Biological Process,Molecular Function,Cellular Component

Open File

/home/biobam/Desktop/idlists

Name	Size	Last Modified
ref_set	1 KB	27/03/2019 10:43
test_set	1 KB	27/03/2019 10:43

All Compatible Types

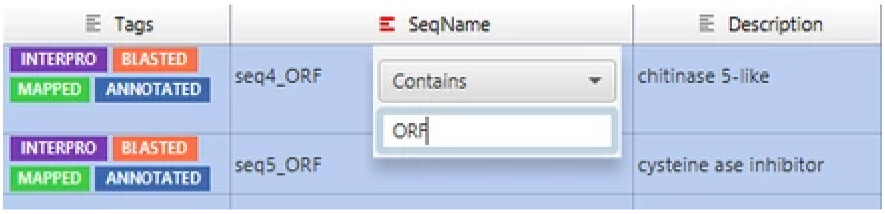
Figure 7:: Select ID list from Fisher Exact Test Wizard

3.3.2.2 Filter out Sequences

The OmicsBox table allows filtering out rows, depending on different search criteria for each column. Each column header shows a small icon which opens a context menu when left-clicked.

- Filters can be applied in various columns and are joined via an AND condition.
- Different data-types allow different filter settings (e.g. numbers allow greater than).
- When a filter is applied on a column the filter icon turns red - double-clicking the icon will remove the filter.
- On the top-right corner, there is a small status message showing how many sequences pass your filter.
- The button next to it (crossed out a document) removes all current filters.

All the algorithms, blast, mapping, annotation, etc., work on the selected sequences and not only on the filtered ones. This means if one has a filter but there are some sequences selected on your project and one runs e.g. remove blast results, it will work on all selected sequences and not only on the ones you see on the table.

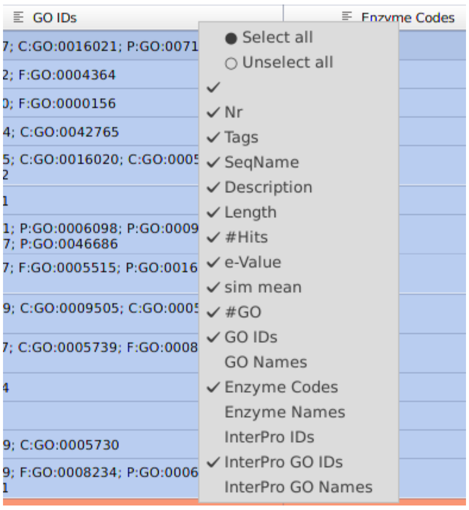


Tags	SeqName	Description
INTERPRO BLASTED MAPPED ANNOTATED	seq4_ORF	chitinase 5-like
INTERPRO BLASTED MAPPED ANNOTATED	seq5_ORF	cysteine ase inhibitor

Figure 8: Filter Criteria

3.3.2.3 Hide Columns

This feature allows hiding the columns of the sequence table.
By right clicking on a column and a menu will be displayed and one can select those columns to hide from the table.
In combination with **Export Table** from File > Export, this can be used to customize the output.



GO IDs	Enzyme Codes
7; C:GO:0016021; P:GO:0071	
2; F:GO:0004364	
0; F:GO:0000156	
4; C:GO:0042765	
5; C:GO:0016020; C:GO:0005 2	
1	
1; P:GO:0006098; P:GO:0009 7; P:GO:0046686	
7; F:GO:0005515; P:GO:0016	
9; C:GO:0009505; C:GO:0005	
7; C:GO:0005739; F:GO:0008	
4	
9; C:GO:0005730	
9; F:GO:0008234; P:GO:0006 1	

Figure 9: Hide Columns

3.4 File Types

From version Blast2GO 3.1 upwards the .b2g file type replaces the previous .dat file. The .dat files will still be supported for opening and export. All OmicsBox project and results (enrichment results, charts, graphs, etc.) will be saved with the new file type .b2g/.box. This files can be viewed and opened directly within the FileManager tab. All other file types can be open form the FileManager via the systems default application.

1. .box/.b2g: File type for all OmicsBox objects (project, results, id lists, etc.). Some OmicsBox objects can be opened with different viewers like for example the OmicsBox Table or the Generic Table.

2. .dat: Legacy format for previous OmicsBox projects. This format had been replaced in 2015 by the more flexible and performance .b2g/.box format. The .dat format is still maintained for compatibility with older datasets and application versions as well as the OmicsBox plugins.

3.5 Uninstalling OmicsBox

OmicsBox can be uninstalled from the different operating systems.
Below one can find instructions on how to uninstall OmicsBox.

3.5.1 Windows

On the folder where OmicsBox has been installed you will find an uninstall.exe file.
Double-click on the .exe file and follow the wizard to complete.

> ⚠ By default, the OmicsBox folder is located in Local folder (C:/Users/[username]/AppData/Local/OmicsBox).

It can also be uninstalled directly from **Programs** in the **Control Panel.**

3.5.2 Linux

On the folder where OmicsBox has been installed you will find an uninstall.sh file.
Open a command line, run the command and follow the wizard to complete:

```
./uninstall.sh
```

> ⚠ By default, the OmicsBox folder is located in your home directory (/home/[username]/OmicsBox).

3.5.3 MacOS

Move the whole OmicsBox folder to trash.

> ⚠ By default, the OmicsBox folder is located in the Applications directory (/Applications/OmicsBox).

4 System Requirements

4.1 General

- CPU: 64-bit, 1.6GHz or faster processor, 2 or more cores recommended
- Microsoft Windows (64-bit): 7, 8 or 10 (recommended), Windows Server 2016 and Windows Server 2019
- Mac: OS X (64-bit) 10.10, 10.11 and macOS 10.12, 10.13, 10.14
- Linux (64-bit): Ubuntu 12.04 and later, RHEL 7 and later (The software is expected to run without problem on other recent Linux systems, but we do not guarantee this.)
- Memory: 4GB of RAM (8GB recommended)
- Disk Space: 2 GB of available hard-disk space; additional 4 GB to download all optional content and temporary files. The latter depends mainly on the input data volume.
- 1280x800 display resolution (at 100% DPI scaling)
- Continuous internet connection required for product activation, content download, and cloud connections
- Installation on local drive necessary, SSD recommended, installation on flash or network drive not advisable
- Virtual Environments are supported but require additional licensing conditions

4.2 Language Versions

- English

4.3 Supported import/export formats

- See user manual: http://manual.omicsbox.biobam.com/user-manual/file-menu/#FileMenu-Load

4.4 Documentation

- User Manuals online only at: http://manual.omicsbox.biobam.com[1]

4.5 Support

- Email Support Tickets with Online Portal / Community

4.6 Java

OmicsBox is built using Java technology and includes Oracle Java JRE 1.8 which is needed to run the Software. This JRE will not interfere with existing JRE’s on your computer and will only be used to run OmicsBox.

[1] http://manual.omicsbox.biobam.com/user-manual/

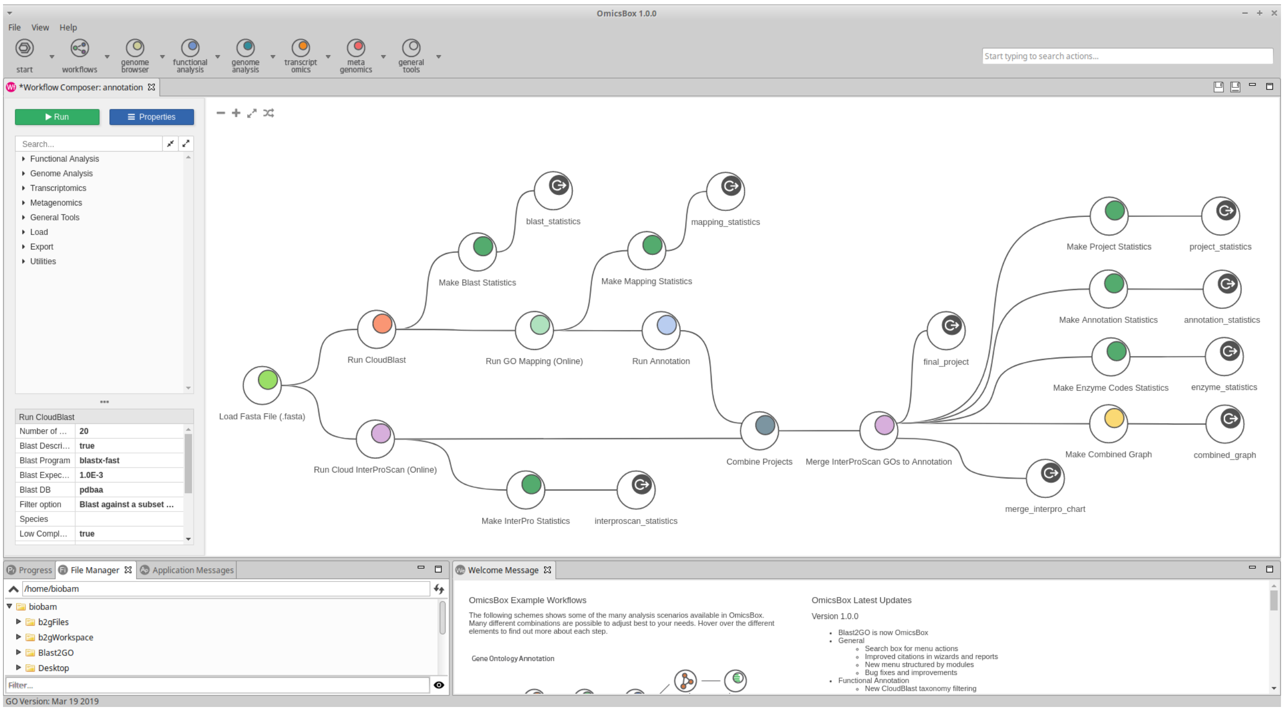
4.7 Scope of requirements

All system requirements on this page are requirements for the most recent versions of the products. For requirements for older versions, please refer to the user manual for that specific version.

5 Workflows

Content of this page:

OmicsBox provides an interface to create, edit and run workflows based on the Common Workflow Language (CWL) specification. This interface allows to describe all analysis steps using the functions and tools offered by OmicsBox and connect them to perform a complete analysis in a single run. Workflows are highly customizable since users can define input data, configure the parameters of each step, save and export results, generate charts and statistics and more.



Run CloudBlast	
Number of ...	20
Blast Descri...	true
Blast Program	blastx-fast
Blast Expec...	1.0E-3
Blast DB	pdbaa
Filter option	Blast against a subset ...
Species	
Low Compl...	true

OmicsBox Example Workflows

The following schemes shows some of the many analysis scenarios available in OmicsBox. Many different combinations are possible to adjust best to your needs. Hover over the different elements to find out more about each step.

OmicsBox Latest Updates

Version 1.0.0

- Blast2GO is now OmicsBox
- General
  - Search box for menu actions
  - Improved citations in wizards and reports
  - New menu structured by modules
  - Bug fixes and improvements
- Functional Annotation
  - New CloudBlast taxonomy filtering

Figure 1: Workflow Composer Interface

5.1 Workflow Composer

The OmicsBox workflow composer interface offers all the necessary options to manage workflows. You can access the composer using the "Workflows" toolbar item or the "Create Workflow" menu option in the workflows toolbar menu.

5.1.1 Design workflows

To create a workflow, start adding some steps. The side panel (on the left) contains the list of actions (that may vary depending the apps installed in OmicsBox) that can be used as workflow steps (Figure 2). To add an action to the workflow click on the corresponding plus symbol next to the action's name.

Each action is represented by an icon (Figure 3). On the left side of the icon are placed the connections for every input of the action (e.g. project, count table, etc) and on the right side are placed the connections for every output it produces as result (chart, graph, etc).

To connect two steps of the workflow, click on the small circle representing the output connection of the first action and drag it to the small circle representing the input connection of the second action. If the connection is valid (i.e. both types match) the small input connection circle should turn green, and a line connecting both circles should be displayed. Otherwise, the small input connection circle should turn red, indicating the selected output can not be used as input for that action.

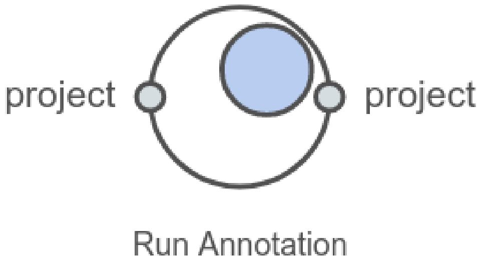


Figure 3: Action icon

Using the "Properties" button in the side panel it is possible to edit the documentation of the workflow. You can write whatever information it is useful, like the author name, author email and a description of the workflow.

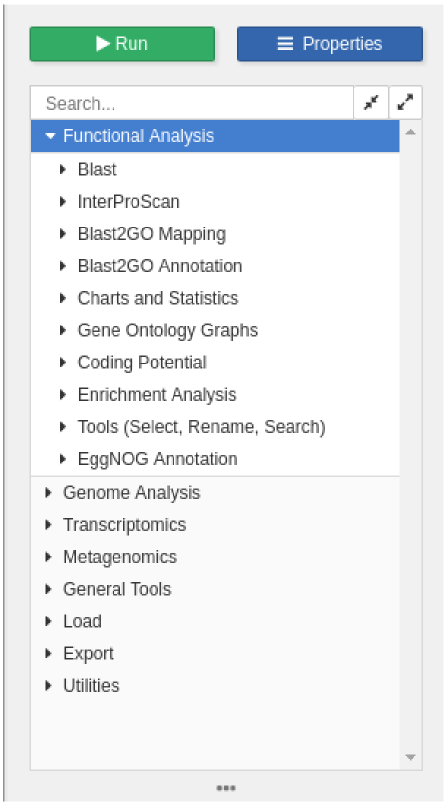


Figure 2: List of actions that can be included in a workflow

5.1.2 Configure workflow steps

Most workflow steps can be configured. If a step needs to be configured (because its parameters are not valid) it will be highlighted in red color and it will not be possible to run the workflow (Figure 4). To configure the step right-click on the step icon and select the "Edit Parameters" option. The red color should disappear as the parameters are now valid. The parameters of each step can be consulted in the bottom region of the side panel.

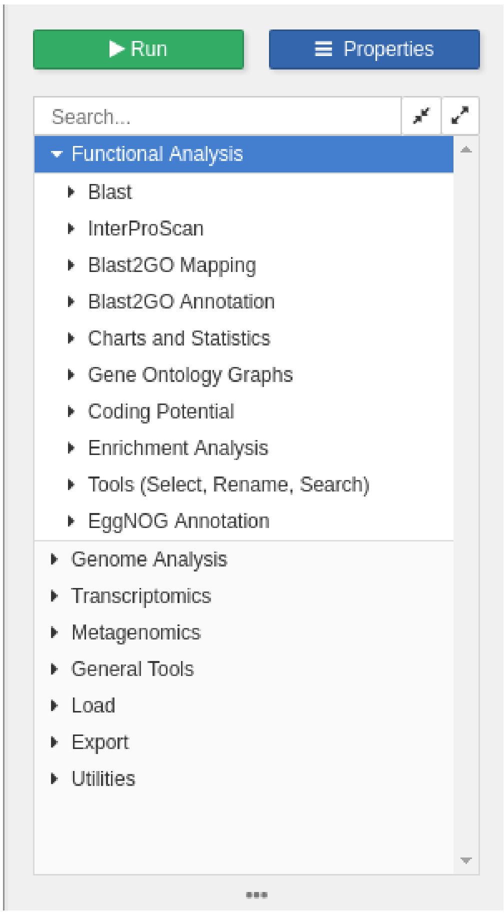


Figure 4: Valid step (left), invalid step (right).

5.1.3 Define workflow inputs

Workflow inputs are .b2g files by definition. To use a .b2g file as input click on the input connection on the left side of the icon and drag it out. The input file can be selected by right-clicking on this step and selecting the "Select Input File" option (Figure 4 top) or in the "Run Workflow" wizard.

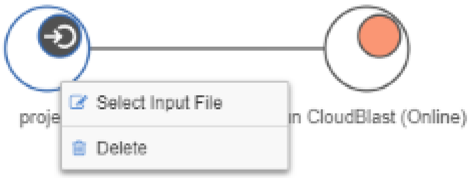


Figure 5: Input data definition

⚠ Some actions don't require any input and produce a result that can be used as input by other actions. These actions can be incorporated as first step in a workflow (e.g. Load Fasta, Eukaryotic and Prokaryotic Gene Finding, Create Count Table, etc).

5.1.4 Define workflow outputs

Like inputs, workflow outputs are .box/.b2g files by definition. To save the results as .box/.b2g files click on the output connection on the right side of the icon and drag it out. The output name (file name in the end) can be selected by right-clicking on this step and selecting the "Change Output Name" option (Figure 6). Later in the "Run Workflow" wizard it is possible to choose the output folder of every output, or use a common output folder for all workflow's outputs.

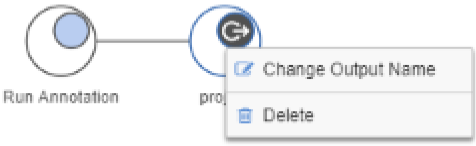
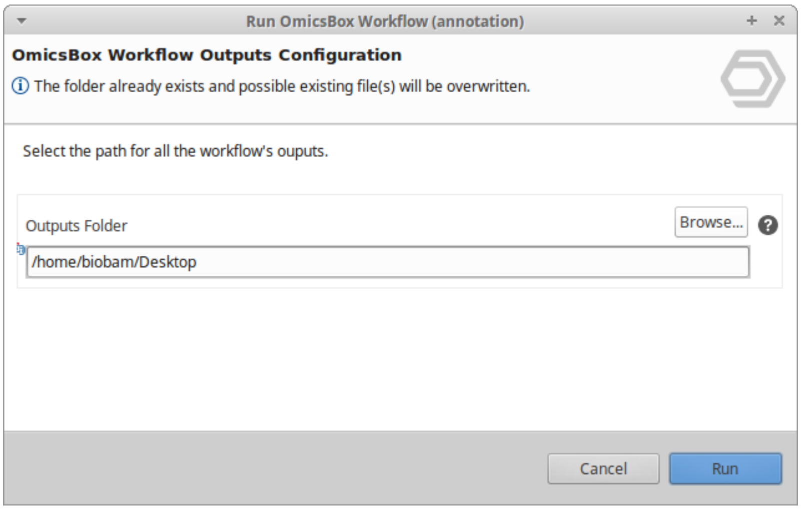


Figure 6: Output data definition

⚠ If you want to export the output of a workflow as a regular file (e.g. .txt, .csv, .png) instead as .box/.b2g file, use the several export actions to export annotations, charts and statistics that OmicsBox offers. These actions can be incorporated as the last step in a workflow (e.g. Generic Export, Export Chart, Export Report, etc).

5.1.5 **Run workflows**

Once the workflow is ready click on the green "Run" button on the side panel to open the "Run Workflow" wizard. Here you can select the inputs files and the outputs folder(s) to save the results or use a common folder for all results (Figure 7). Click "Run" to execute the workflow. You will see a new progress bar for the workflow execution, and an additional progress bar for every step, so it is possible to cancel the whole workflow or just a certain step and continue with the others.



Run OmicsBox Workflow (annotation)

OmicsBox Workflow Outputs Configuration

ⓘ The folder already exists and possible existing file(s) will be overwritten.

Select the path for all the workflow's ouputs.

Outputs Folder: /home/biobam/Desktop

Figure 7: Outputs folder configuration

6 Genome Browser

Download[2] **Example Datasets.**

The Genome Browser allows you visualize different file types in a side-scrolling way, the features are rendered ordered by start from left to right. Each file will be represented as a track, several tracks can be added to the browser, and the current supported tracks are GFF, VCF, DNA Fasta and BAM. Tracks can be reordered, hidden or closed using the track controls (✖ ✚ ⌃ ⌄). One chromosome is visualized at a time, the top box represents the whole chromosome and the small box represents the area of the current region.

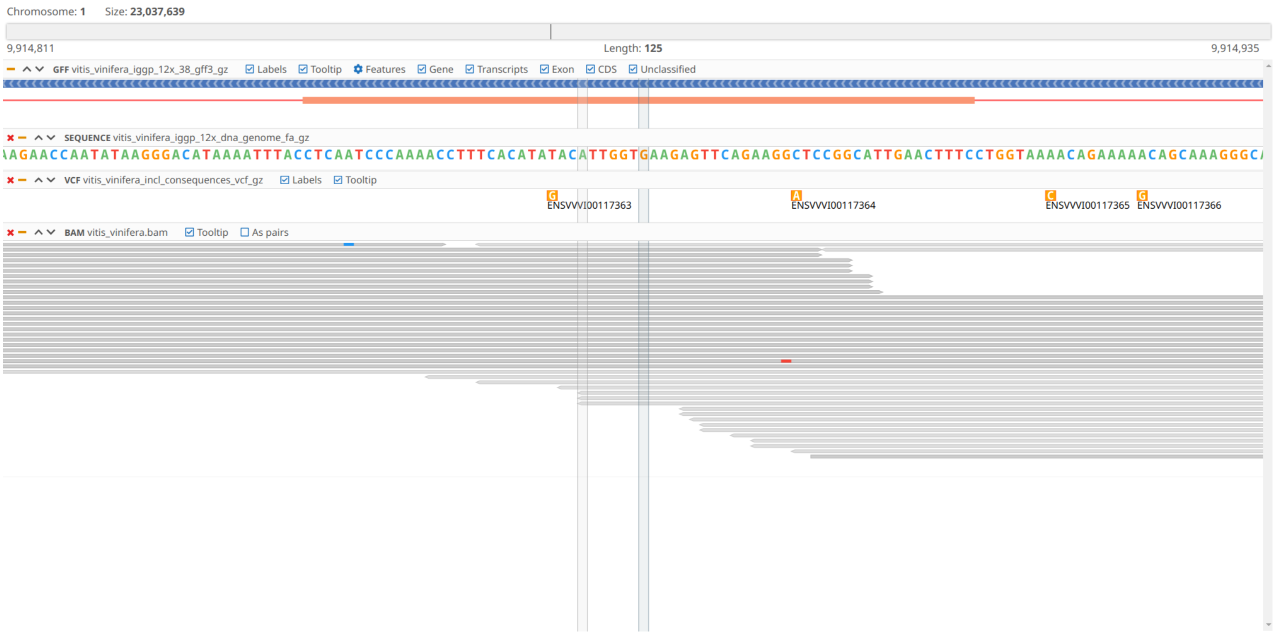
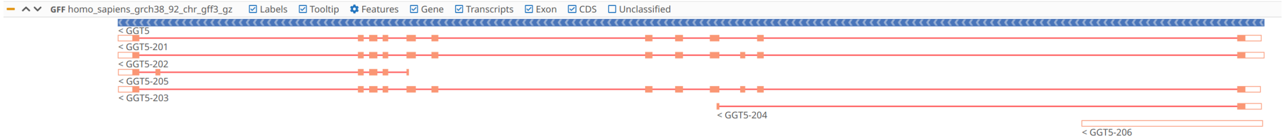

Chromosome: 1
Size: 23,037,639
9,914,811
Length: 125
9,914,935
GFF vitis_vinifera_iggp_12x_38_gff3_gz
☑ Labels
☑ Tooltip
Features
☑ Gene
☑ Transcripts
☑ Exon
☑ CDS
☑ Unclassified
SEQUENCE vitis_vinifera_iggp_12x_dna_genome_fa_gz
AGAACCAATATAAGGGACATAAAATTTACCTCAATCCCAAAACCTTTCACATATACATTGGTGAAGAGTTCAGAAGGCTCCGGCATTGAACTTTCCTGGTAAAACAGAAAAACAGCAAAGGGC
VCF vitis_vinifera_incl_consequences_vcf_gz
☑ Labels
☑ Tooltip
G
ENSVVVI00117363
A
ENSVVVI00117364
C
ENSVVVI00117365
G
ENSVVVI00117366
BAM vitis_vinifera.bam
☑ Tooltip
☐ As pairs

Figure 1: Genome Browser with all 4 tracks.

6.1 GFF Track

The GFF track is able to detect types to group features by genes (genes, transcripts, exons and CDS and the relationship between them by using ID and Parent attributes, see GFF3 Specification[3] for more details). Genes are in blue, the transcripts are the small red lines, non-coding exons are white boxes with orange border and exons with CDS are filled with orange color. Other non-grouped features appear in grey. GFF Viewer can be opened with the context menu option in the **File Manager** when selecting a GFF file and using the context menu option *Show in GFF Viewer* from the **table** when exploring a GFF file.

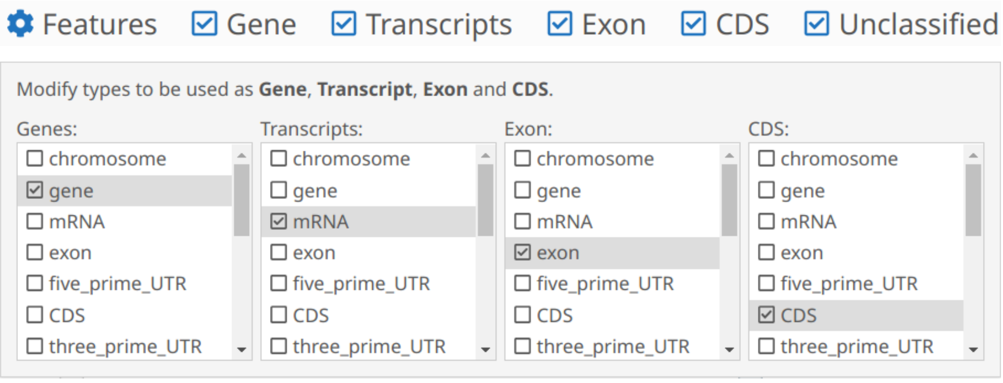
[2] https://resources.biobam.com/omicsbox/example_data/General.zip
[3] https://github.com/The-Sequence-Ontology/Specifications/blob/master/gff3.md



GFF homo_sapiens_grch38_92_chr_gff3_gz ☑ Labels ☑ Tooltip Features ☑ Gene ☑ Transcripts ☑ Exon ☑ CDS □ Unclassified

Figure 2: GFF viewer.

With the **Features button** controls you can classify the types for *Genes*, *Transcripts*, *Exons* and *CDS*, this will modify the Gene group visualization, by default these types will be set automatically. The checkboxes will hide or show the features, i.e. if you want to show only the genes, you need to uncheck the Transcripts, Exon, CDS, and Unclassified checkboxes.



Features ☑ Gene ☑ Transcripts ☑ Exon ☑ CDS ☑ Unclassified

Modify types to be used as **Gene**, **Transcript**, **Exon** and **CDS**.

Genes:
□ chromosome
☑ gene
□ mRNA
□ exon
□ five_prime_UTR
□ CDS
□ three_prime_UTR

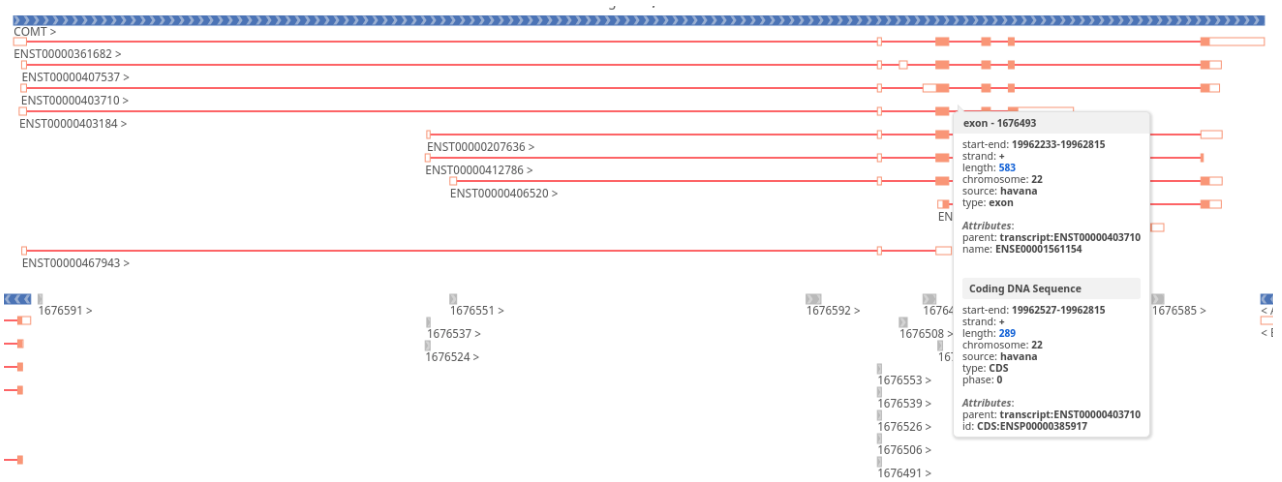
Transcripts:
□ chromosome
□ gene
☑ mRNA
□ exon
□ five_prime_UTR
□ CDS
□ three_prime_UTR

Exon:
□ chromosome
□ gene
□ mRNA
☑ exon
□ five_prime_UTR
□ CDS
□ three_prime_UTR

CDS:
□ chromosome
□ gene
□ mRNA
□ exon
□ five_prime_UTR
☑ CDS
□ three_prime_UTR

Figure 3: Feature controls.

A tooltip will appear on mouse hover, it will show additional information about the feature and the information is collected from the GFF file.



exon - 1676493

start-end: **19962233-19962815**
strand: **+**
length: **583**
chromosome: **22**
source: **havana**
type: **exon**

Attributes:
parent: **transcript:ENST00000403710**
name: **ENSE00001561154**

Coding DNA Sequence

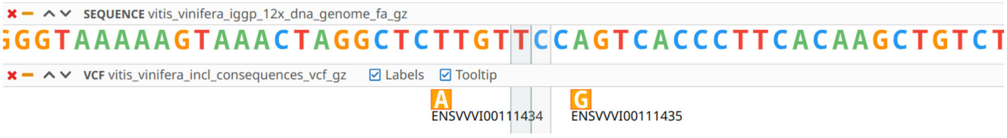
start-end: **19962527-19962815**
strand: **+**
length: **289**
chromosome: **22**
source: **havana**
type: **CDS**
phase: **0**

Attributes:
parent: **transcript:ENST00000403710**
id: **CDS:ENSP00000385917**

Figure 4. Feature tooltip.

6.2 VCF Track

The VCF track shows the variants in the corresponding position and if you zoom in enough, the alternative nucleotide will be shown.



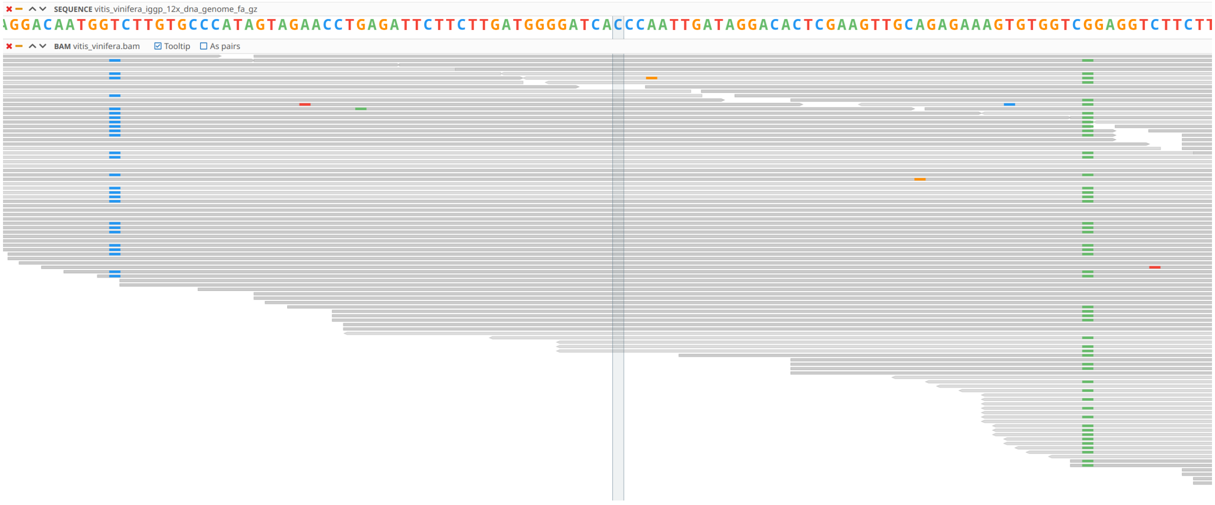
SEQUENCE vitis_vinifera_iggp_12x_dna_genome_fa_gz

VCF vitis_vinifera_incl_consequences_vcf_gz ☑ Labels ☑ Tooltip

Figure 5: VCF track

6.3 BAM Track

The BAM track shows the reads of a BAM file and if the sequence track is active, will also paint the differences between the read sequence and the sequence track. If you click on the ☐ As pairs checkbox, if the BAM has paired reads, the pair will be painted one beside the other.



SEQUENCE vitis_vinifera_iggp_12x_dna_genome_fa_gz

BAM vitis_vinifera.bam ☑ Tooltip ☐ As pairs

Figure 6: BAM track

6.4 DNA Fasta Track

The DNA Fasta or sequence track shows the nucleotides and is only shown on low region lengths.

SEQUENCE vitis_vinifera_iggp_12x_dna_genome_fa_gz

Figure 7: DNA Fasta Track

6.5 Navigation

Navigation is performed with the mouse, by left-clicking without releasing the button the scroll action will start, the just move the mouse to move left or right. Also is possible to adjust the region to a feature or gene by double-clicking on it. Zoom in and Zoom out from the current region is performed using the **Zoom**

buttons on the **Toolbar**, If you zoom out enough, visualization will change and a histogram will show. The chromosome box can be used to select a new region either clicking or by selecting a new box area using click and drag.

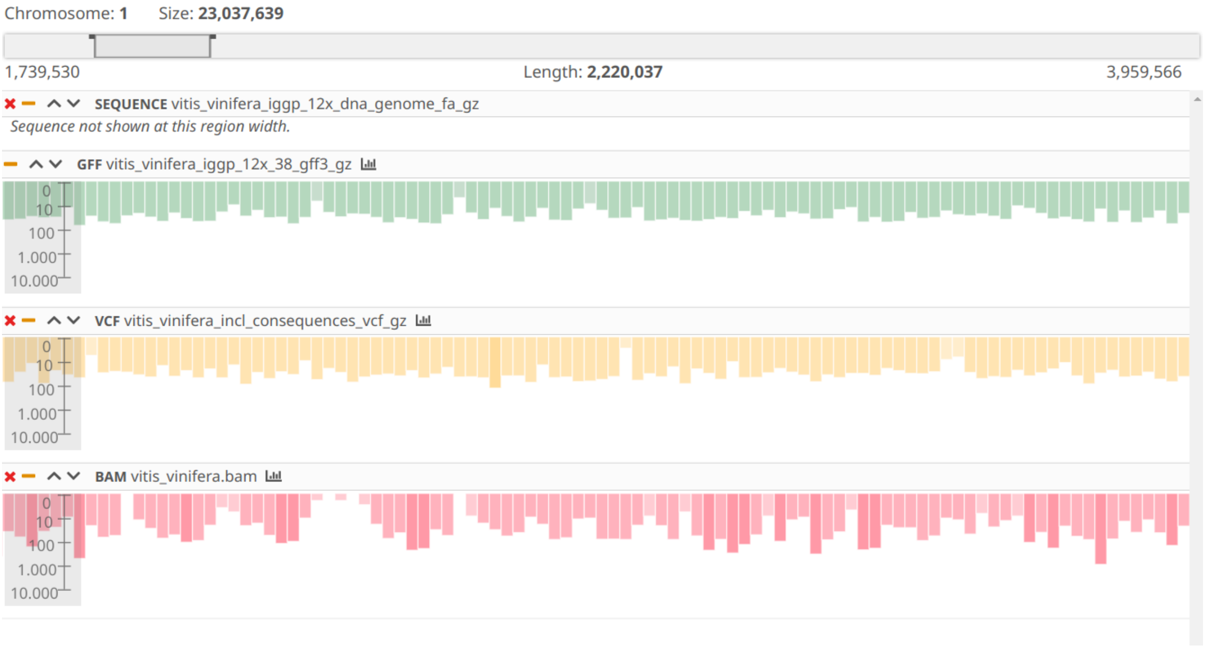
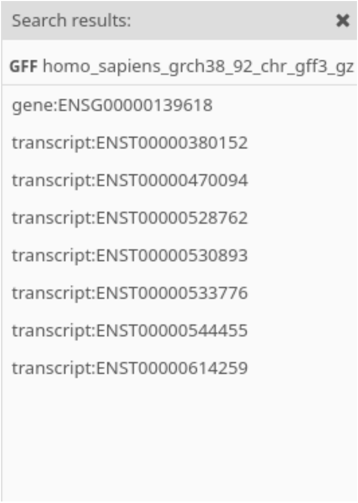

Chromosome: 1
Size: 23,037,639
1,739,530
Length: 2,220,037
3,959,566
SEQUENCE vitis_vinifera_iggp_12x_dna_genome_fa_gz
Sequence not shown at this region width.
GFF vitis_vinifera_iggp_12x_38_gff3_gz
0
10
100
1.000
10.000
VCF vitis_vinifera_incl_consequences_vcf_gz
BAM vitis_vinifera.bam

Figure 8. Histogram, the small rectangle on the Chromosome box shows the region painted as a histogram.

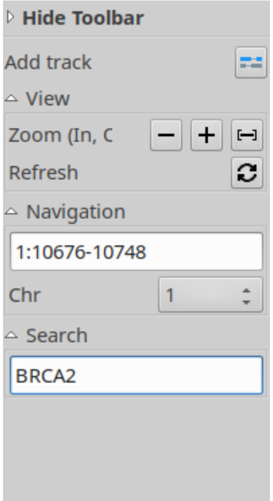
It is also possible to find features on the GFF and the VCF track using the **Search field** in the Toolbar, the search results will be shown on a panel beside the Toolbar.



Search results:

GFF homo_sapiens_grch38_92_chr_gff3_gz

gene:ENSG00000139618
transcript:ENST00000380152
transcript:ENST00000470094
transcript:ENST00000528762
transcript:ENST00000530893
transcript:ENST00000533776
transcript:ENST00000544455
transcript:ENST00000614259



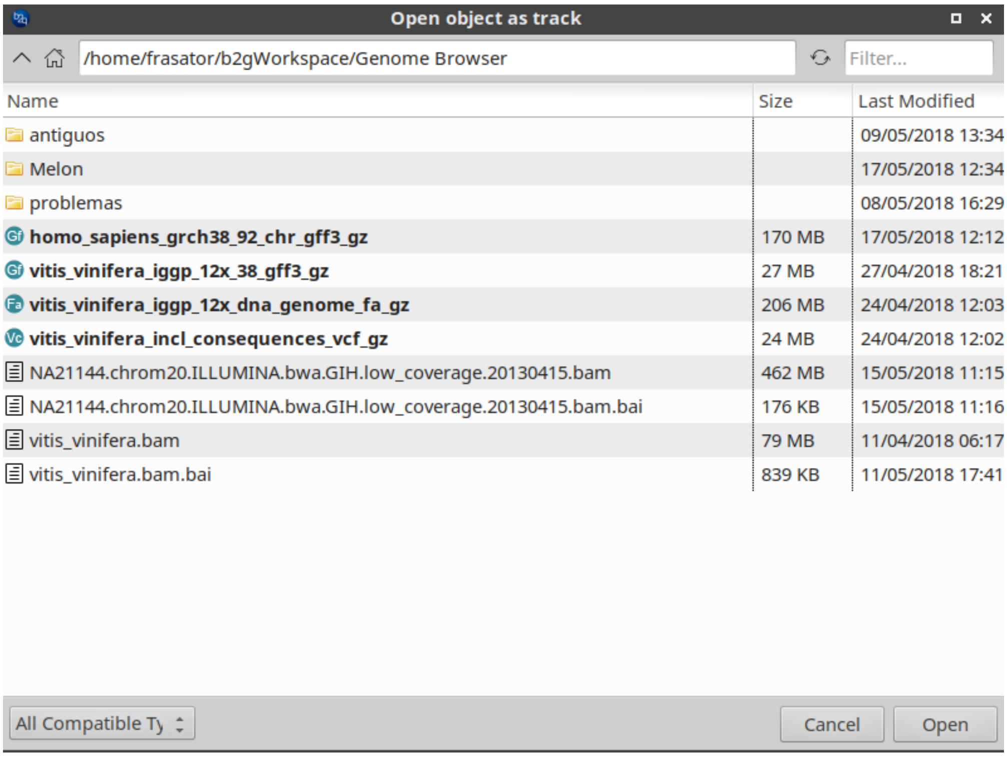
1:10676-10748

Chr 1

BRCA2

Figure 9: Toolbar

Also, you can add more tracks to the current Genome Browser using the **Add track** button. A window will appear the available files that can be opened with a Genome Browser as tracks.



/home/frasator/b2gWorkspace/Genome Browser

Name	Size	Last Modified
antiguos		09/05/2018 13:34
Melon		17/05/2018 12:34
problemas		08/05/2018 16:29
homo_sapiens_grch38_92_chr_gff3_gz	170 MB	17/05/2018 12:12
vitis_vinifera_iggp_12x_38_gff3_gz	27 MB	27/04/2018 18:21
vitis_vinifera_iggp_12x_dna_genome_fa_gz	206 MB	24/04/2018 12:03
vitis_vinifera_incl_consequences_vcf_gz	24 MB	24/04/2018 12:02
NA21144.chrom20.ILLUMINA.bwa.GIH.low_coverage.20130415.bam	462 MB	15/05/2018 11:15
NA21144.chrom20.ILLUMINA.bwa.GIH.low_coverage.20130415.bam.bai	176 KB	15/05/2018 11:16
vitis_vinifera.bam	79 MB	11/04/2018 06:17
vitis_vinifera.bam.bai	839 KB	11/05/2018 17:41

Figure 10: Browse tracks to be added to the Genome Browser

6.6 Load files

Files must be loaded in order to be shown on the Genome Browser to do that just click on the **File menu** and under the **Load sub-menu** you can load GFF, VCF, DNA Fasta and BAM files.
These files can be loaded in either with .gff or .gz extension.

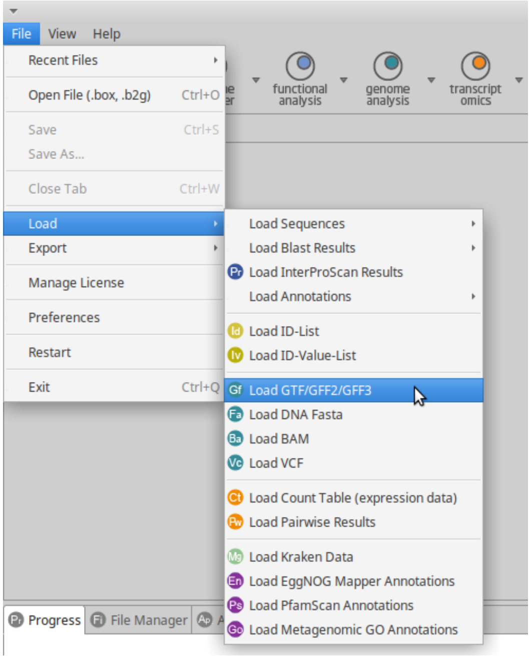
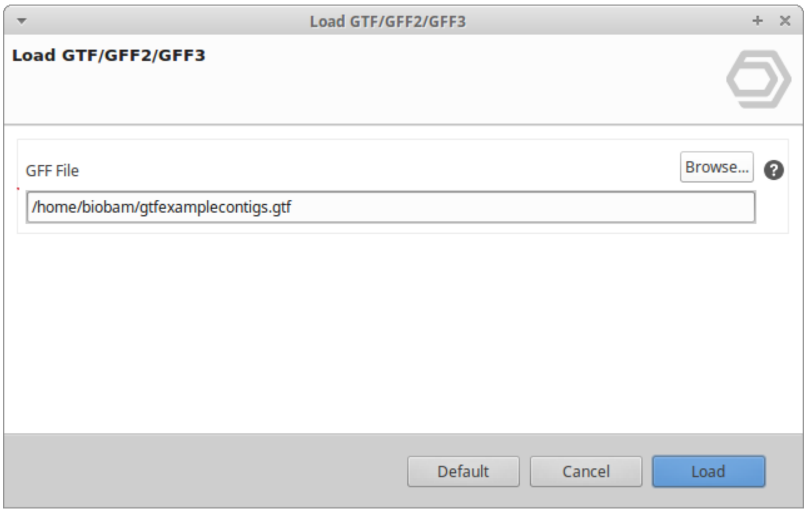


Figure 11: Load Files to see in the browser

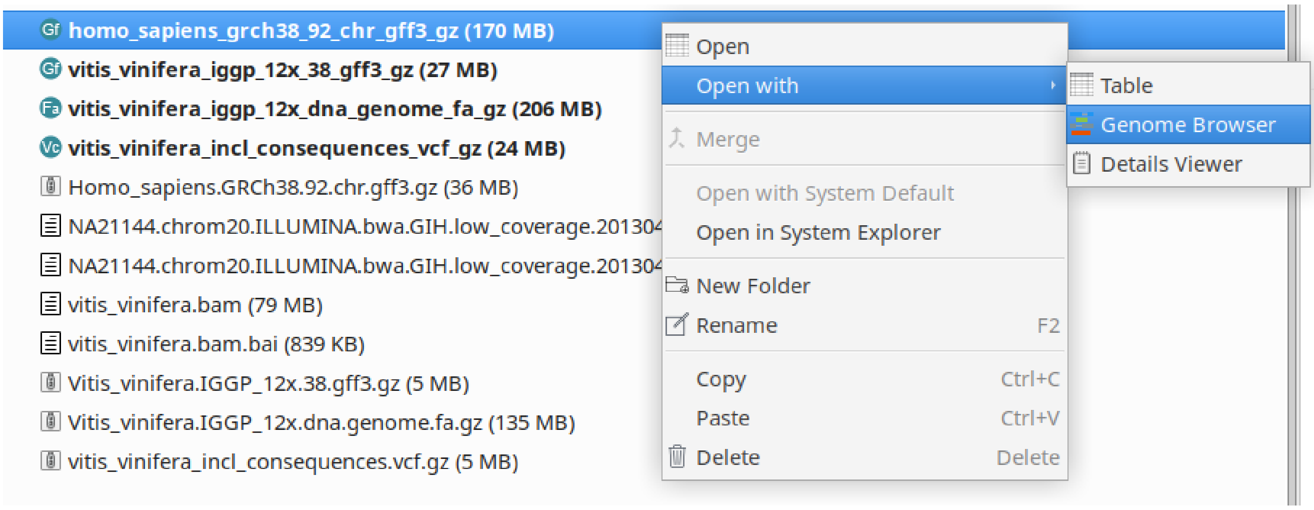
A window will appear to select the file to import, use the browse button to select the file from the file system and finally click on the **Load button** to start the Load process, the load time depends on the file size.



GFF File: /home/biobam/gtfexamplecontigs.gtf

Figure 12: Browse for GFF file

Once loaded you may save the file, once the file is saved you can visualize it using the Genome Browser.



- homo_sapiens_grch38_92_chr_gff3_gz (170 MB)
- vitis_vinifera_iggp_12x_38_gff3_gz (27 MB)
- vitis_vinifera_iggp_12x_dna_genome_fa_gz (206 MB)
- vitis_vinifera_incl_consequences_vcf_gz (24 MB)
- Homo_sapiens.GRCh38.92.chr.gff3.gz (36 MB)
- NA21144.chrom20.ILLUMINA.bwa.GIH.low_coverage.201304
- NA21144.chrom20.ILLUMINA.bwa.GIH.low_coverage.201304
- vitis_vinifera.bam (79 MB)
- vitis_vinifera.bam.bai (839 KB)
- Vitis_vinifera.IGGP_12x.38.gff3.gz (5 MB)
- Vitis_vinifera.IGGP_12x.dna.genome.fa.gz (135 MB)
- vitis_vinifera_incl_consequences.vcf.gz (5 MB)

Figure 13: Open the saved file directly with the Genome Browser

Genome Browser can also be opened from a **Table** view, If you have already a file open in a table, use the context menu option **Show in Genome Browser**, by right-clicking on a row, a Genome Browser will be shown on the region of that feature.

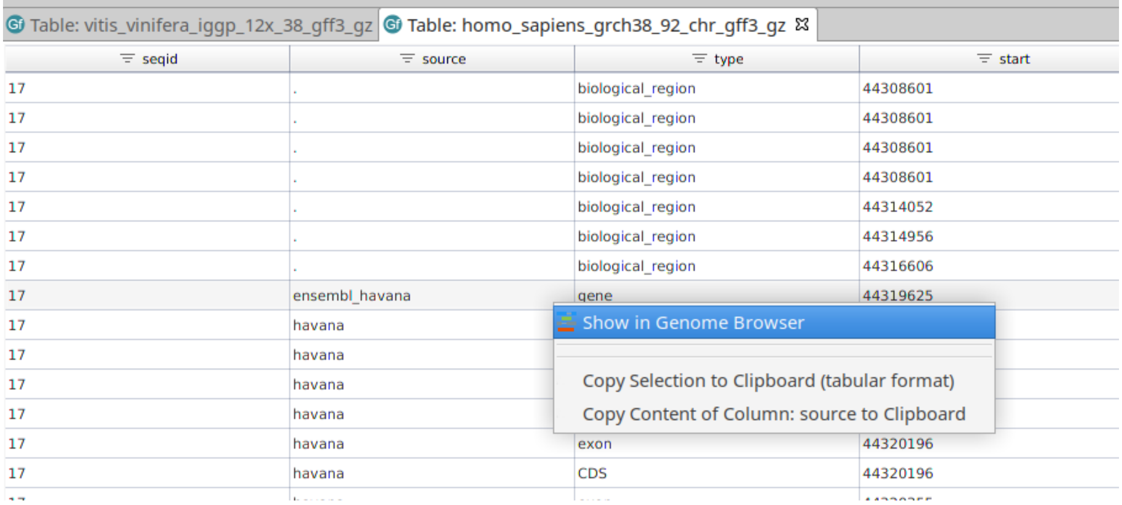


Table: vitis_vinifera_iggp_12x_38_gff3_gz
Table: homo_sapiens_grch38_92_chr_gff3_gz
Show in Genome Browser
Copy Selection to Clipboard (tabular format)
Copy Content of Column: source to Clipboard

seqid	source	type	start
17	.	biological_region	44308601
17	.	biological_region	44308601
17	.	biological_region	44308601
17	.	biological_region	44308601
17	.	biological_region	44314052
17	.	biological_region	44314956
17	.	biological_region	44316606
17	ensembl_havana	gene	44319625
17	havana		
17	havana		
17	havana		
17	havana		
17	havana	exon	44320196
17	havana	CDS	44320196

Figure 14: Open the Genome Browser fromGFF table

7 Module Functional Analysis

Content of this section

The Functional Analysis Module is a comprehensive bioinformatics tool for functional annotation.

This module uses the Blast2GO methodology to extract the GO terms associated with the obtained hits and returns an evaluated GO annotation for the query sequence(s). Enzyme codes are obtained by mapping from equivalent GOs while InterPro motifs are directly queried at the InterProScan web service. GO annotation can be visualized reconstructing the structure of the Gene Ontology relationships and pathways.

A typical basic use case of the Functional Analysis Module consists of 5 steps: BLASTing, mapping, annotation, statistical analysis, and visualization. These steps will be explained in the next sections.

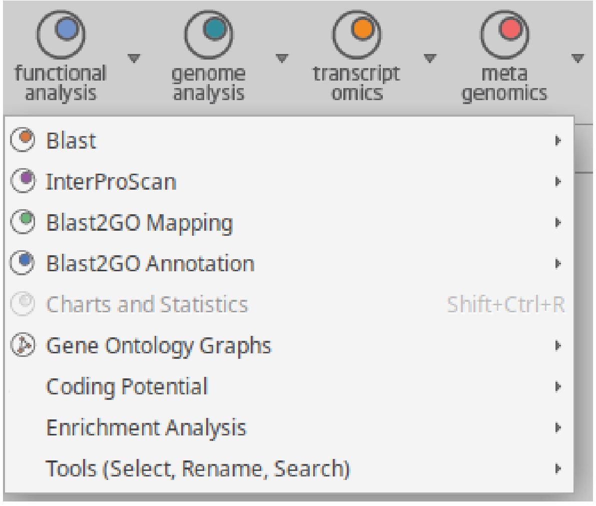


Figure 1: Functional Analysis menu

Functional Annotation Analysis use case: https://www.biobam.com/whole-genome-functional-annotation-of-solanum-lycopersicum/.

Functional Analysis Example Dataset: Download[4].

[4] https://resources.biobam.com/omicsbox/example_data/FunctionalAnalysis.zip

7.1 Quick Start

1. This section provides a quick run-through of a basic functional annotation process done within OmicsBox. More detailed descriptions of the different analysis steps and more advanced features are described in the remaining sections of this documentation.
2. **Load Data**
   Go to File **Load > Load Sequences > Load Fasta File** and select your *.fasta* file containing the set of sequences in FASTA format. Alternatively, you can load the example sequences into OmicsBox choosing **File > Load > Load Example Sequences**. Please download example files to try and test OmicsBox: b2g_example_files.zip[5]

3. **Blast**
   Click on **Functional Analysis (toolbar) > Blast**. In the Blast Configuration Dialog (BLAST(see page 31)) select the way in which Blast will be executed (CloudBlast, NCBI Blast or Local Blast), the type of Blast mode which is appropriate for your sequence type (Blastx for nucleotide and Blastp for protein data) and the taxonomies you want to blast against. Click Next for the advanced settings and to choose where to save the Blast results, and click Run to start the Blast search.

- Once your BLAST analysis is finished visualize your results at **Functional Analysis (toolbar) > Charts and Statistics > Blast**.
- On the Main Sequence Table, right-click on a sequence to open the Single Sequence Menu (BLAST(see page 31)). Select Show BLAST Result to the BLAST Browser for that sequence.

> ⚠ **Note**
> If you are running blast using CloudBlast we recommend to run blastx-fast or blastp-fast as it is faster and fewer computation units will be consumed.

4. **InterProScan**
   By clicking on the **InterPro** icon the corresponding Wizard will be shown. If InterProScan is executed via the EMBL-EBI web service, please provide a valid email address. This is not needed if InterProScan is run via CloudIPS. It is highly recommended to run IPS in order to improve the quality of the annotations. Once InterProScan results are retrieved use **Merge InterProScan GOs to Annotation** to add GO terms obtained through motifs/domains to the current annotations. InterProScan can be run in parallel with BLAST.

5. **Mapping**
   Click on **Functional Analysis > Blast2GO Mapping > Run GO Mapping** to start mapping GO terms. Mapped sequences will turn **green**. Once Mapping is completed visualize your results at **Functional Analysis > Charts and Statistics > Mapping**.

6. **Annotation**
   Click on **Functional Analysis > Blast2GO Annotation > Run Annotation** to open the Annotation Configuration Window. Click Next to change the evidence codes and finally click Run to start the annotation. Annotated sequences will turn **blue**.
   - Once the annotation is completed you are able to visualize your results with **Functional Analysis > Charts and Statistics > Annotation**.

[5] https://www.blast2go.com/images/b2g_support/b2g_example_files.zip

- On the Main Sequence Table, right-click on a sequence to open the Single Sequence Menu. Select **Make Graph of GO-Mapping with Annotation Score** to visualize the annotation on the GO DAG for that sequence.
- If desired, modify the annotation by clicking with the right mouse button and select **Change Annotation and Description** or reducing to a GO-Slim representation **Functional Analysis > Blast2GO Annotation > GO-Slim > Run GO- Slim (online)**.
- During the annotation process, Enzyme Codes (EC) will be also given when a GO-term/EC number equivalence is available.

7. **Enrichment Analysis**
OmicsBox provides tools for the statistical analysis of GO term frequency differences between two sets of sequences. Go to **Functional Analysis > Enrichment Analysis > Enrichment Analysis (Fisher's Exact Test)** and a new Dialog window will open (Fisher's Exact Test(see page 97)). Select a *.txt* file or an ID list containing the sequence IDs for a subset of sequences. A test-set example file can be downloaded the OmicsBox website. Select the second set of sequences as reference set if desired. If no reference set is provided all annotations of the corresponding project will be used as the reference. Click **Run** to start the analysis. A table containing the results of this analysis will be displayed in a new tab.

- Click on **Make Enriched Graph** icon to visualize the results of the Fisher’s Test on the GO DAG.
- Click on **Show Bar Chart** to obtain a bar chart representation of GO frequencies.
- The results can be reduced to more specific GO terms in the corresponding icon and saved as text format (**Save as Text**).

8. **Combined Graph**
OmicsBox can visualize the combined annotation for a group of sequences on the GO DAG. Select a group of sequences to generate their combined graph at **Functional Analysis > Tools > Select > Select Sequences**. Now **Select by Features** and **Select by Name or ID**. You can use the Demo Test Set used previously for this. Alternatively, you can select sequences using the sequence check boxes of the Main Sequence Table. Now go to **Functional Analysis > Gene Ontology Graphs > Make Combined Graph**. Now click Run.

9. **Save Results**
File > Save saves the current OmicsBox project as .box file.

10. **Export Results**

- **File > Export** allows exporting the generated data in many different formats.
- **File > Export Annotations** exports the actual annotation results as .annot file or generate own formatted annotation file as .txt file.
- The enrichment analysis results can be exported in various formats from the Fisher Exact Test Result Viewer. ‘‘Save as Text’’ exports the results as a tabulator separated text file.
- To export GO graphs use the sidebar of the corresponding graph viewer. Graphs can be saved/ exported in *.pdf, .png, .svg* and *.txt*.

11. **Project Statistics**
Once finished any step or at the beginning, we can obtain a general chart in which it shows the state of the analysis of the entire data.
We will be able to know the number of sequences that belong to a concrete state (**functional analysis > Charts and Statistics**).
The data distribution can be visualized in two different charts, one as a bar chart and the other as a pie chart (Figures 1 and 2).
These are the different states we are going to find in the charts:

a. **Total:** The total amount of sequences in the project (only in the bar chart).
b. **Without Analysis:** Sequences without processing or have been reset in the BLAST menu (functional analysis > Blast > Remove Blast Results).
c. **With Only InterProScan:** Sequences that only have InterProScan and nothing else.
d. **Without Blast Hits:** Sequences that have been sent to BLAST but no hits have been found.
e. **With Blast:** Successful sequences after BLAST step or have been reset in the Mapping menu (functional analysis > Blast2GO Mapping > Remove Mapping).
f. **With Mapping:** Successful sequences after Mapping step or they have been reset in the Annotation menu (functional analysis > Blast2GO Annotation > Remove Annotation).
g. **With GO Annotation:** Successful sequences after Annotation step.
h. **With Manual Annotation:** Manually annotated sequences before or after executing the annotation step.
i. **With GO-Slim Annotation:** Sequences with GO-Slim Annotation.

Each state will have assigned a specific colour.

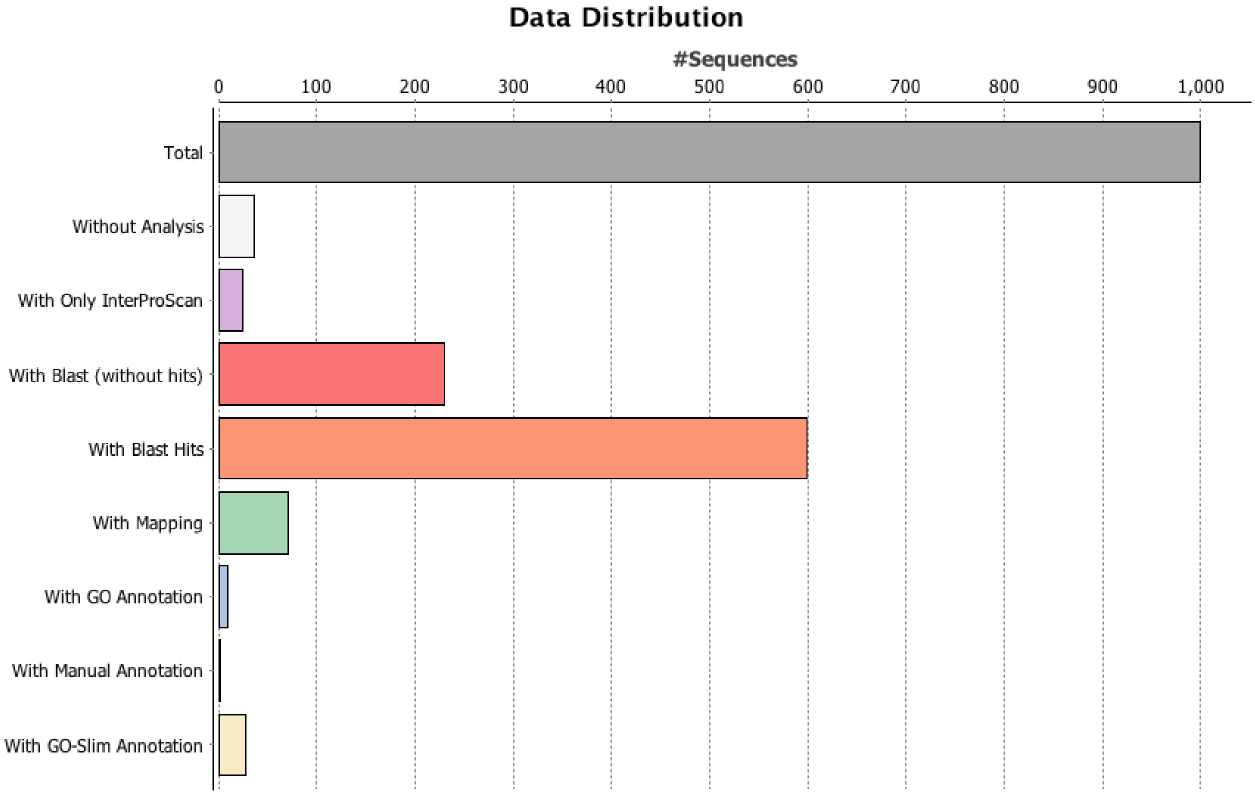


Figure 1: Data Distribution Bar chart

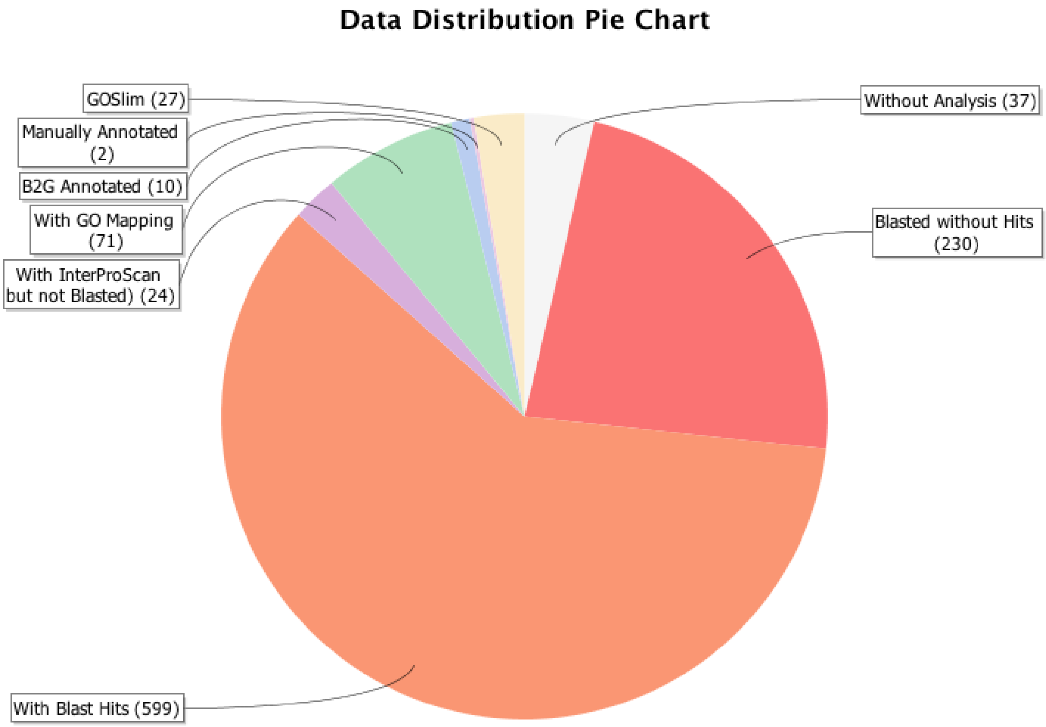


Figure 2: Data Distribution Pie chart

It is also possible to see the progress of the analysis (Figure 3).

From the 1000 sequences, 700 have blast results.
From the chart, it could suggest there are still some analysis to be completed, such as mapping, annotation and specially InterProScan.

Once all the analysis steps have been executed the Analysis Progress chart should be similar to the one in Figure 4.

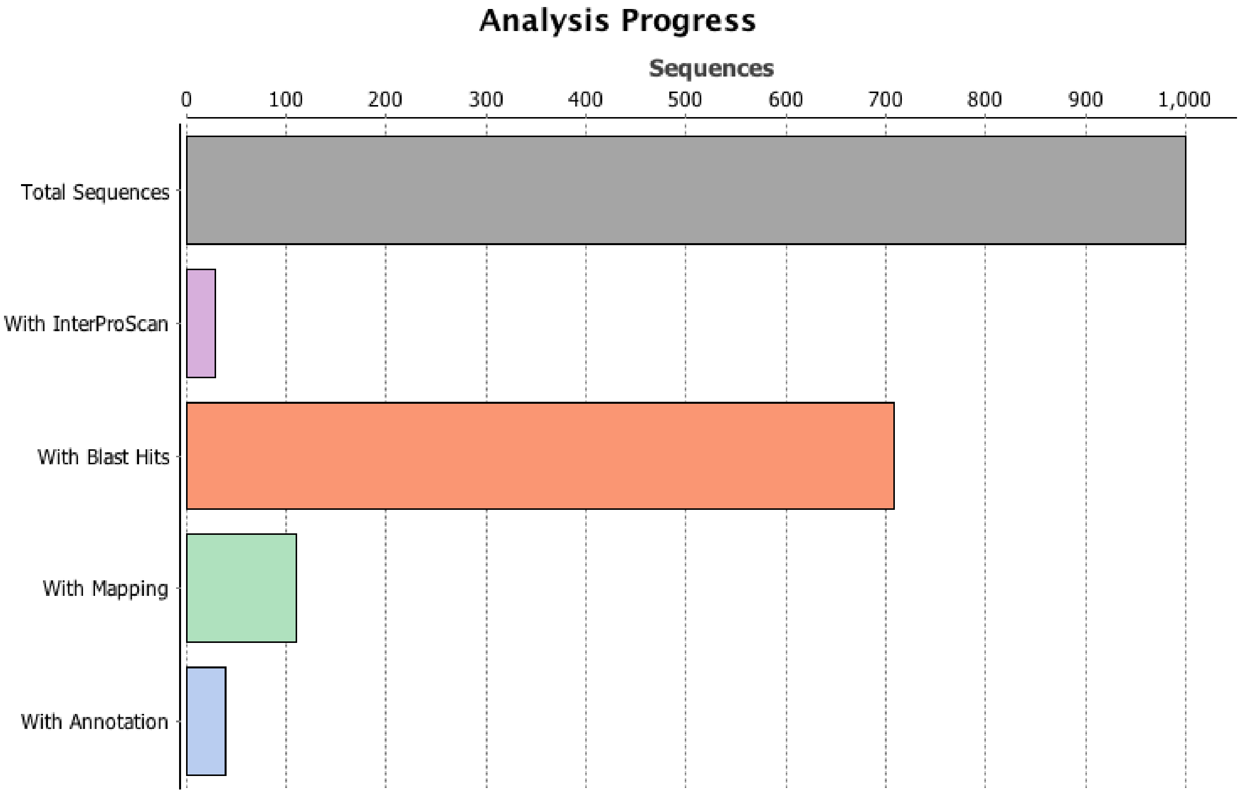


Figure 3: Analysis Progress Chart

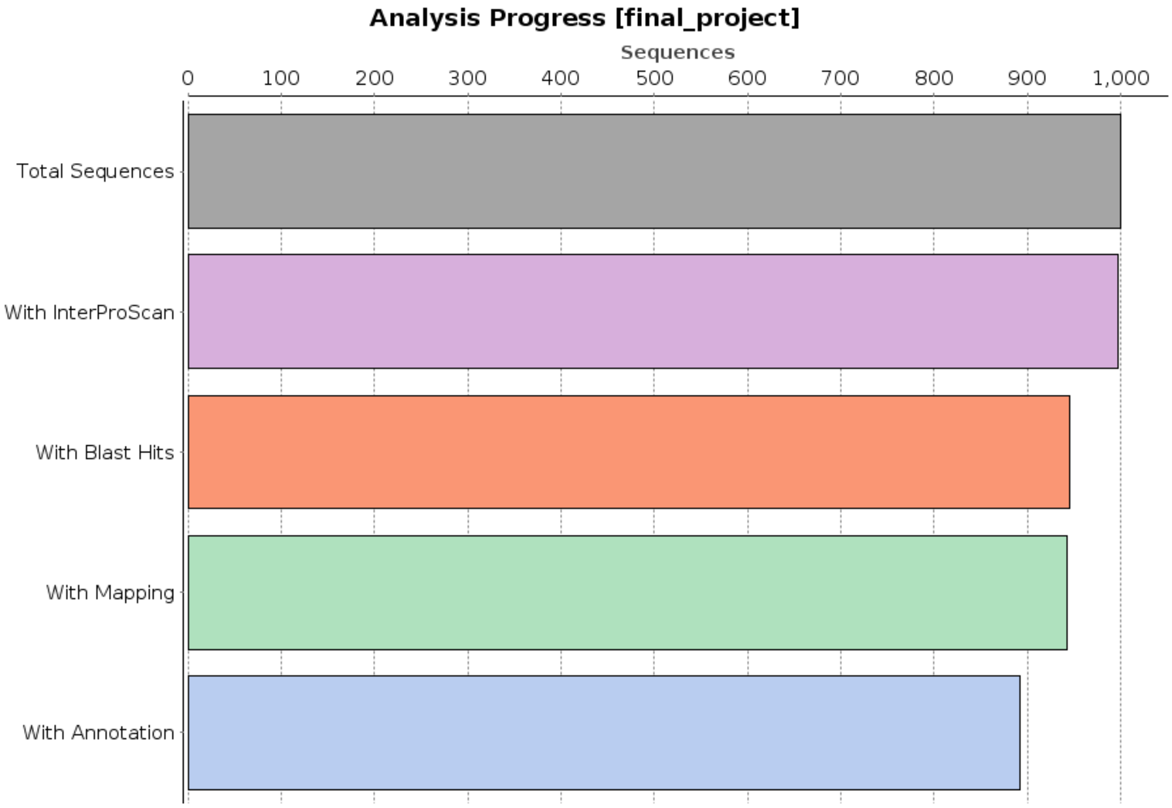


Figure 4: Analysis Progress Chart after running all analysis

7.2 Sequences

Content of this page:

To start a new OmicsBox project for the Functional Analysis you just have to load your sequence data from a file into OmicsBox.

7.2.1 Load sequences from start

At the "File" menu, go to **Load > Load Sequences > Load Fasta File** and select the file containing your sequences. The application accepts text files containing one or more DNA or protein sequences in FASTA or FASTQ format (see below). These files must have the extension *.fasta,.fnn,.faa,.fna,.ffn, .txt, .fq* or *.fastq* to be accepted by the application.

A sequence in FASTA format begins with a single-line description or header starting with a ">" character. The rest of the header line is arbitrary but should be informative. Subsequent lines contain the sequence, one character per residue. Lines can have different lengths. Be sure your file is in this format and avoid strange characters in the sequence header, such as '&' or '\' and use 'N' to denote in-determinations in the sequences.

An example of the FASTA format:

```
>gi|121664|sp|P00435|GSHC BOVIN GLUTATHIONE PEROXIDASE
MCAAQRSAAALAAAAPRTVYAFSARPLAGGEPFNLSSLRGKVLLIENVASLUGTTVRDYTQMND
LQRLGPRGLVVLGFPCNQFGHQENAKNEEILNCLKYVRPGGGF
```

7.2.2 Show Sequence

Once the sequences have been loaded to OmicsBox, it is possible to see them by right-clicking on the Sequence Table. The "Single Sequence Menu" (context menu) will appear (figure 1[6]). This menu provides some functions for sequences individually, i.e. will apply to the sequence at that position of the Table. With the sequence viewer, it is possible to copy the sequence to the clipboard.

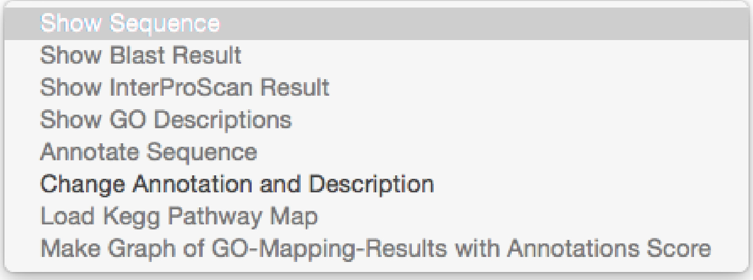
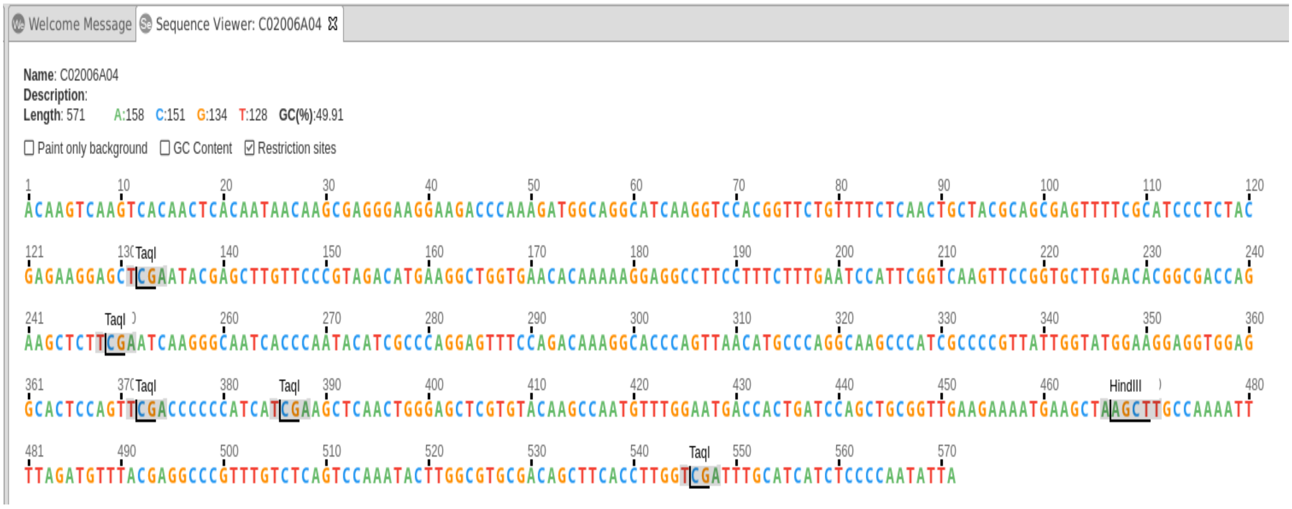


Figure 1: Context Menu: Show Sequence

[6] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=620789839#Sequences-figure1



Name: C02006A04
Description:
Length: 571 A:158 C:151 G:134 T:128 **GC(%)**:49.91

□ Paint only background □ GC Content ☑ Restriction sites

```
1   ACAAGTCAAGTCACAACTCACAATAACAAGCGAGGGAAGGAAGACCCAAAGATGGCAGGCATCAAGGTCCACGGTTCTGTTTTCTCAACTGCTACGCAGCGAGTTTTCGCATCCCTCTAC   120
121 GAGAAGGAGCTCGAATACGAGCTTGTTCCCGTAGACATGAAGGCTGGTGAACACAAAAAGGAGGCCTTCCTTTCTTTGAATCCATTCGGTCAAGTTCCGGTGCTTGAACACGGCGACCAG   240
241 AAGCTCTTCGAATCAAGGGCAATCACCCAATACATCGCCCAGGAGTTTCCAGACAAAGGCACCCAGTTAACATGCCCAGGCAAGCCCATCGCCCCGTTATTGGTATGGAAGGAGGTGGAG   360
361 GCACTCCAGTTCGACCCCCCATCATCGAAGCTCAACTGGGAGCTCGTGTACAAGCCAATGTTTGGAATGACCACTGATCCAGCTGCGGTTGAAGAAAATGAAGCTAAGCTTGCCAAAATT   480
481 TTAGATGTTTACGAGGCCCGTTTGTCTCAGTCCAAATACTTGGCGTGCGACAGCTTCACCTTGGTCGATTTGCATCATCTCCCCAATATTA
```

Figure 2: Sequence Viewer

7.2.3 Sequence Length Statistics

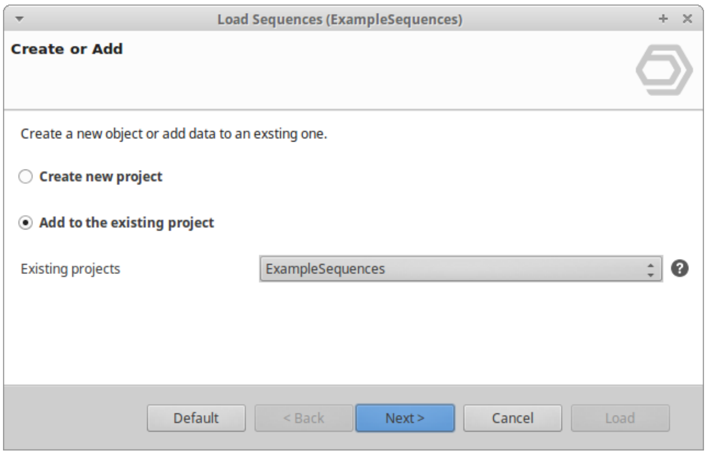
OmicsBox allows you to visualize the length distribution of your sequences in the arrow next to the "Chart" icon.

7.2.4 Add sequences to existing OmicsBox project

Use the File Manager context menu to merge into the project. Select two or more project, open the context menu with a right click on one of the files and select **Merge**.

In case the loaded project file has only Blast results and no sequence information it is still possible to add the corresponding sequences to OmicsBox project by clicking on the arrow next to the "Start" icon and select "Load Sequences". Now two options will be displayed "Create new project" see figure 3[7] and 4, and "Add to the existing project". The "Add to the existing project" option should be selected and in the next page, you can browse for the fasta file and "Overwrite" option should be selected.

[7] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=620789839#Sequences-figure3



Load Sequences (ExampleSequences)

Create or Add

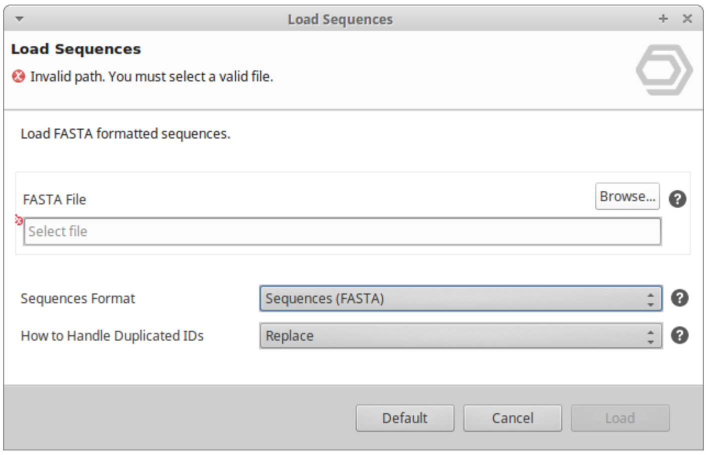
Create a new object or add data to an exsting one.

○ Create new project

● Add to the existing project

Existing projects: ExampleSequences

Figure 3: Load Sequences Dialog: Add sequences to an existing project



Load Sequences

Load Sequences

Invalid path. You must select a valid file.

Load FASTA formatted sequences.

FASTA File: Select file

Sequences Format: Sequences (FASTA)

How to Handle Duplicated IDs: Replace

Figure 4: Load Sequences Dialog: Choose Fasta file

Export FASTA Sequence with Annotation Results

After executing the whole functional annotation of the sequences in OmicsBox (BLAST, Mapping and Annotation) it is possible to export the sequence in FASTA format with the corresponding sequence description and GO ID or GO term from File > Export > Export as FASTA.

An example of the Exported FASTA format with Sequence Description and GO IDs:

```
>C04018C10|mitogen-activated protein kinase 3|GO:0005634;GO:0004707;GO:0005515
acaaacgagagcgtagaaaattaattagagagaaaaagagagagagtaaaatggctgacgtggcgcaggtcaacg
gcgtaggtcaaacggctgattttcctgcggtaccgacgcacggcggtcagtttatacagtacaatatatttggaa
acttgtttgaaatcacggccaagtatcggcctccgatcatgccgattggtcgcggcgcgtacgggatcgtttgct
cggtgttgaatacggagacgaatgagctcgttgcgatgaagaaaatagcgaacgcttttgataatcacatggatg
ccaagcgaacgcttcgtgagattaagcttctgcaacatttcgatcatgaaaatgtgatagctgtaaaagatgtgg
ttcccccaccgttacgaagagaattcactgatgtctatattgctgcggaactcatggacactgacctttaccaaa
ttattcgctcaaatcaaagtttatccgaggagcactgccagtatttcttgtatcaacttcttcgaggactcaagt
atatccattcagcaaatgttattcatcgggatttgaagcccagcaatctcttgttaaatgcaaattgtgatttaa
```

```
aaatttgtgattttggtcttgctcgtccaacctcggagaatgagttcatgacagagtatgttgtcacaagatggt
accgagcccctgagttgttattgaactcatctgactacactg
```

7.3 BLAST

Content of this page:

OmicsBox uses the Basic Local Alignment Search Tool (BLAST) to find sequences similar to your query set. Please, refer to http://www.ncbi.nlm.nih.gov/BLASTfor details on the BLAST function. Figure 2[8], show the BLAST Configuration Dialog Window that controls the BLAST step.

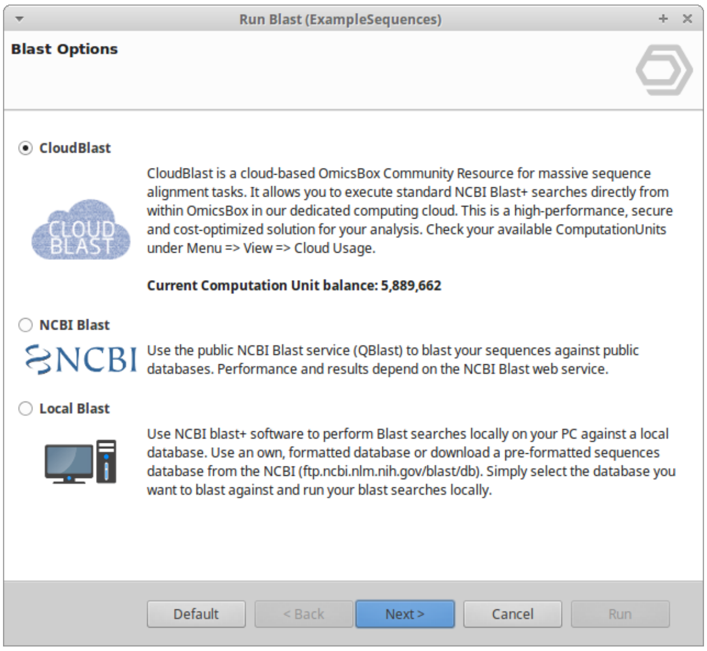
BLAST in OmicsBox can basically be performed in three different fashions:

1. CloudBlast. This is a cloud-based OmicsBox Community Resource for massive sequence alignment tasks. It allows you to execute standard NCBI Blast+ searches directly from within OmicsBox in a dedicated computing cloud. CloudBlast is a high-performance, secure and cost-optimized solution for your analysis. This is a blast service totally independent from the NCBI servers to provide fast and reliable sequence alignments. Please see Run Blast using CloudBLAST section for more information.
2. QBlast@NCBI. NCBI offers a public service that allows searching molecular sequence databases with the BLAST algorithm. The main advantages of making use of this service are its versatility and that no database maintenance is required. Therefore by selecting this option at OmicsBox no additional installations have to be done.
3. Local BLAST against its own database. It is possible to use BLAST+ executable to query a local/own database. At https://www.blast2go.com/make-own-database-and-blast and at the Make Blast Database[9] section one can see how to prepare and blast locally an own fasta database.

QBlast at NCBI is the only feature available for OmicsBox Basic users.

[8] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598114359#BLAST-figure2
[9] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598114359#BLAST-MakeBlastDatabase

The next figure shows the menu manner to select between NCBI-, local- BLAST as well as CloudBlast, AWS Blast or blasting against an own database.



Run Blast (ExampleSequences)

Blast Options

● CloudBlast

CloudBlast is a cloud-based OmicsBox Community Resource for massive sequence alignment tasks. It allows you to execute standard NCBI Blast+ searches directly from within OmicsBox in our dedicated computing cloud. This is a high-performance, secure and cost-optimized solution for your analysis. Check your available ComputationUnits under Menu => View => Cloud Usage.

Current Computation Unit balance: 5,889,662

○ NCBI Blast

Use the public NCBI Blast service (QBlast) to blast your sequences against public databases. Performance and results depend on the NCBI Blast web service.

○ Local Blast

Use NCBI blast+ software to perform Blast searches locally on your PC against a local database. Use an own, formatted database or download a pre-formatted sequences database from the NCBI (ftp.ncbi.nlm.nih.gov/blast/db). Simply select the database you want to blast against and run your blast searches locally.

Figure 1: Select between NCBI, Local or CloudBlast

7.3.1 Run BLAST at the NCBI

Here, the user can specify the following parameters, which are divided into three different sections: Blast Configuration in figure 2[10], Advanced in figure 3[11] and Save Results Page figure 4[12]:

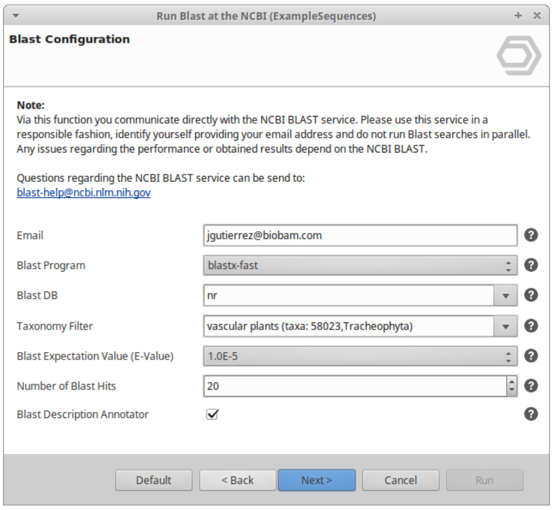
7.3.1.1 Blast Configuration Page

- Your e-mail address in case you are using the NCBI BLAST web service.
- BLAST program: The algorithm you want to use:
  - blastp - Compares an amino acid query sequence against a protein sequence database.
  - blastn (-task blastn) - Compares a nucleotide query sequence against a nucleotide sequence database.
  - blastx - Compares a nucleotide query sequence translated in all reading frames against a protein sequence database. Used to find potential translation products of an unknown nucleotide sequence
  - tblastn - Compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames.
  - blastx-fast
  - blastp-fast
  - blastp-short
  - blastn (-task megablast)

[10] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598114359#BLAST-figure2
[11] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598114359#BLAST-figure3
[12] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598114359#BLAST-figure4

  - blastn (-task dc-megablast)
  - blastn-short
  - tblastn-fast
- BLAST DB: The name of the database to search in (eg. nr, swissprot, pdb). To see a list of possible DBs at NCBI seehttp://data.biobam.com/ncbi_blast_dbs_protein.pdf
- Taxonomy Filter: Search for Blast results only in the selected taxonomy.
- BLAST expect value: The statistical significance threshold for reporting matches against database sequences. If the statistical significance ascribed to a match is greater than the EXPECT threshold, the match will not be reported. Lower EXPECT thresholds are more stringent, leading to fewer chance matches being reported. Increasing the threshold shows less stringent matches.
- Number of BLAST hits: The number of alignments you want to achieve (0-100).

BLAST Description Annotator: The BDA finds the best possible description for a new sequence based on a given BLAST result.



Run Blast at the NCBI (ExampleSequences)

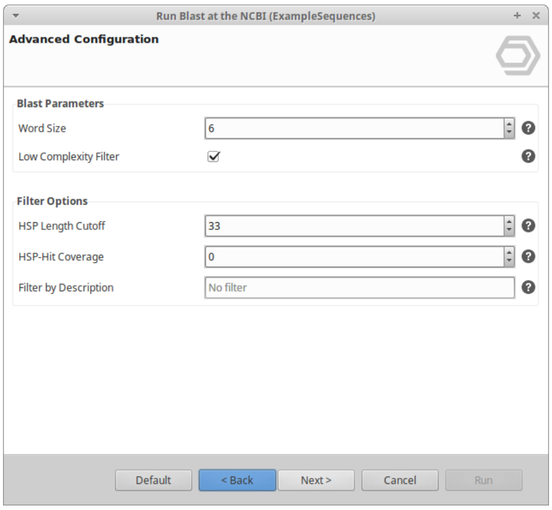
Blast Configuration

Note:
Via this function you communicate directly with the NCBI BLAST service. Please use this service in a responsible fashion, identify yourself providing your email address and do not run Blast searches in parallel. Any issues regarding the performance or obtained results depend on the NCBI BLAST.

Questions regarding the NCBI BLAST service can be send to:
blast-help@ncbi.nlm.nih.gov

Email: jgutierrez@biobam.com
Blast Program: blastx-fast
Blast DB: nr
Taxonomy Filter: vascular plants (taxa: 58023,Tracheophyta)
Blast Expectation Value (E-Value): 1.0E-5
Number of Blast Hits: 20
Blast Description Annotator: ☑

Figure 2: Blast Configuration Page



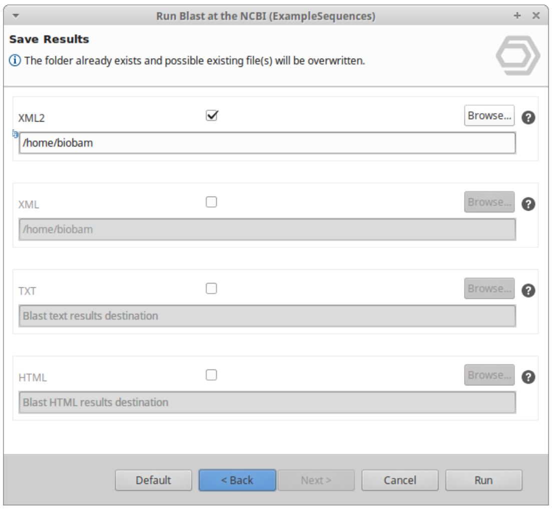
Run Blast at the NCBI (ExampleSequences)

Advanced Configuration

Blast Parameters
Word Size: 6
Low Complexity Filter: ☑

Filter Options
HSP Length Cutoff: 33
HSP-Hit Coverage: 0
Filter by Description: No filter

Figure 3: Advanced Page



Run Blast at the NCBI (ExampleSequences)

Save Results

ⓘ The folder already exists and possible existing file(s) will be overwritten.

XML2 ☑ /home/biobam

XML □ /home/biobam

TXT □ Blast text results destination

HTML □ Blast HTML results destination

Figure 4: Save Results Page

7.3.1.2 Advanced Page

- Blast Parameters:
  - Word size: One of the important parameters governing the sensitivity of BLAST searches is the length of the initial words. The word size is adjustable in blastn and can be reduced from the default value to increase sensitivity. This word size can also be increased to increase the search speed and limit the number of database hits.
  - Low complexity filter: The BLAST programs employ the SEG algorithm to filter low complexity regions from proteins before executing a database search. The default is ON.
- Filter Options:
  - HSP length cutoff: A Cutoff value for the minimal length of the first hsp of a balst hit, used to exclude hits with only small local alignments from the BLAST result. The given length corresponds to amino-acids or nucleotides depending on the type of performed BLAST.
  - HSP-Hit Coverage
  - Filter by description: Filter-out Blast hits by a description

7.3.1.3 Save Results Page

The results of the BLAST queries can also be directly saved to a file in different formats by selecting the corresponding checkboxes at the BLAST Save Results Page. If the chosen file already exists, upcoming results will be appended. Choose a format type to additionally save your BLAST results.

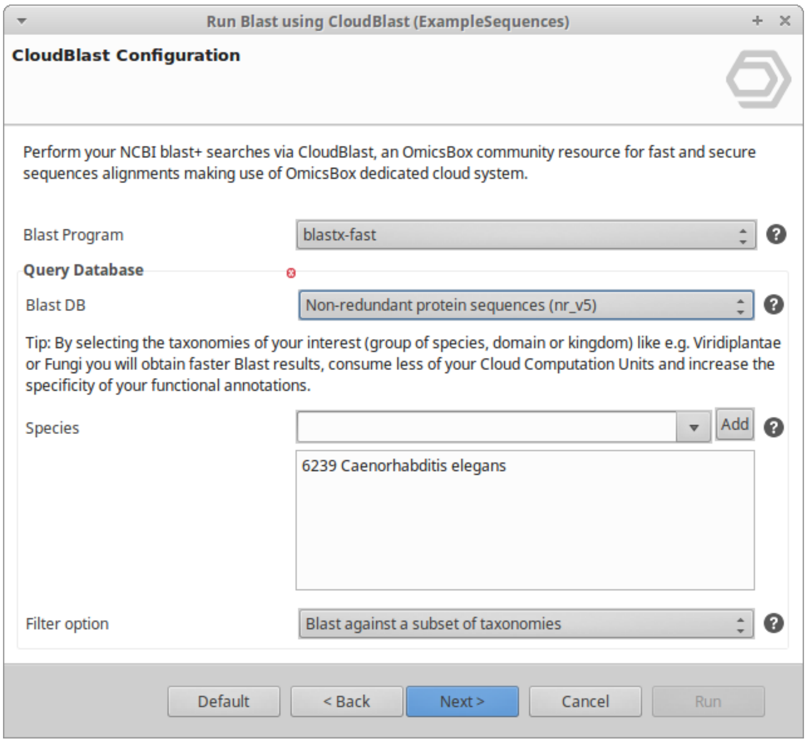
- XML2: This is a new BLAST result provided by NCBI and can also be loaded into OmicsBox.
- XML: It is recommended to save your BLAST results as XML as this format is supported by the OmicsBox Load BLAST Results function.
- TXT: It saves the blast results of each sequence in text file format.
- HTML: For each sequence, a file in HTML format will be saved.

7.3.2 Run BLAST using CloudBLAST

CloudBlast offers a highly optimized, self-sustained HPC solution to address a very specific need of the Omic sBox community.

CloudBlast is a BLAST service totally independent from the NCBI servers to provide fast and reliable sequence alignments. It consists of a high performance computing cluster dedicated exclusively to Blast searches.
All OmicsBox subscriptions include "ComputationUnits" to make use of this resource and allows you to perform blast searches for tens of thousands of sequences within a few days against a large collection of protein databases. Each sequence alignment performed in the system consumes a certain amount of computation time depending on the sequence length and the blast algorithm (blastx, blastp) and parameters used. The smaller the database you blast against the more sequences you can analyse with 6.000.000 ComputationUnits (see Cloud Usage in the View Menu section to know how to monitor the ComputationUnits). This means that e.g. if you blast against the vertebrate NR-subset you would be able to blast approx. one million (1.000.000) sequences. If you decide to blast against the NR database, the largest protein database available, it should allow you to blast approx. 80.000 sequences (with an average length of 800nt per sequence).

For the advanced and save parameters page please see Advanced Page[13] and Save Results Page[14] sections for detailed information.



Run Blast using CloudBlast (ExampleSequences)

CloudBlast Configuration

Perform your NCBI blast+ searches via CloudBlast, an OmicsBox community resource for fast and secure sequences alignments making use of OmicsBox dedicated cloud system.

Blast Program: blastx-fast

Query Database

Blast DB: Non-redundant protein sequences (nr_v5)

Tip: By selecting the taxonomies of your interest (group of species, domain or kingdom) like e.g. Viridiplantae or Fungi you will obtain faster Blast results, consume less of your Cloud Computation Units and increase the specificity of your functional annotations.

Species: 6239 Caenorhabditis elegans

Filter option: Blast against a subset of taxonomies

Figure 5: CloudBlast Configuration Page

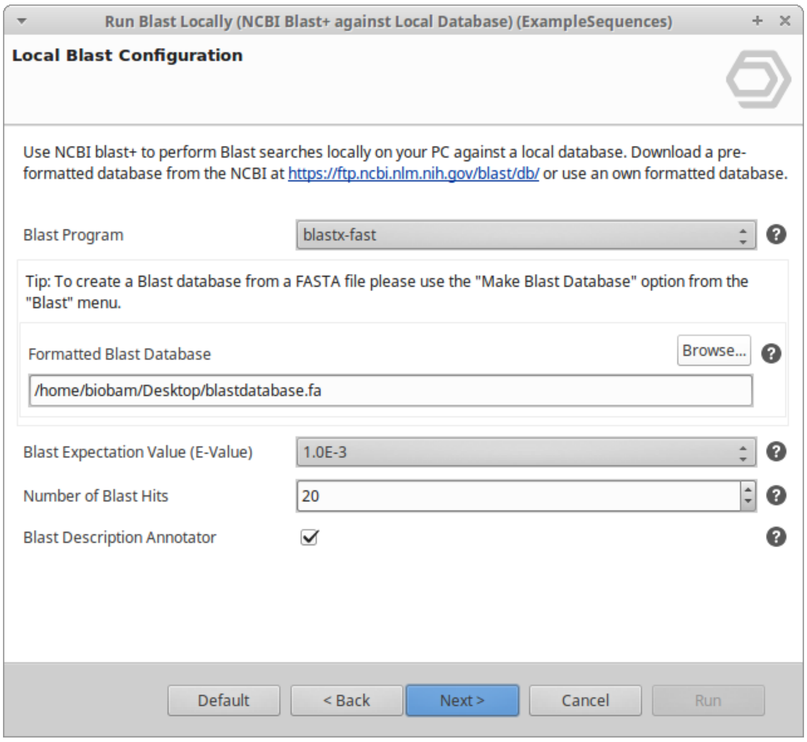
7.3.3 Run BLAST Locally

With Local BLAST you can blast the sequences against own database. OmicsBox allows creating a Blast database from a FASTA file with the option "Make Blast Database" (see Make Blast Database[15] section).

[13] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598114359#BLAST-AdvancedPage
[14] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598114359#BLAST-SaveResultsPage
[15] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598114359#BLAST-MakeBlastDatabase

Download and format your database and choose the corresponding folder to see figure 6[16]. Databases have to be formatted for NCBI Blast+.

The main parameters in the Local BLAST Configuration page are very similar to the ones in NCBI and CloudBlast. The main difference is when choosing the database as OmicsBox is expecting a *.pal*' file or .p*. On the Advanced Page at the "Run Parameters," it is possible to select the number of threads to be used. This field has not to be set up as OmicsBox detects the number of threads in the computer. The Advanced Page[17] section provides a detailed description of each parameter. As in CloudBlast, the BLAST results will be saved in XML file format.



Run Blast Locally (NCBI Blast+ against Local Database) (ExampleSequences)

Local Blast Configuration

Use NCBI blast+ to perform Blast searches locally on your PC against a local database. Download a pre-formatted database from the NCBI at https://ftp.ncbi.nlm.nih.gov/blast/db/ or use an own formatted database.

Blast Program: blastx-fast

Tip: To create a Blast database from a FASTA file please use the "Make Blast Database" option from the "Blast" menu.

Formatted Blast Database: /home/biobam/Desktop/blastdatabase.fa

Blast Expectation Value (E-Value): 1.0E-3

Number of Blast Hits: 20

Blast Description Annotator: ☑

Figure 6: Local Blast Configuration Page

7.3.4 Show BLAST Results

As the BLAST search progresses, sequences with successful BLAST results change their color on the Main Sequence Table from white to **orange** and the BLAST result related columns will be filled. In case no results could be retrieved for a given sequence, this row will turn **dark-red**.
With a mouse the right click on a sequence, the Single Sequence Menu will be displayed and it is possible to see the BLAST results for each sequence individually. Show BLAST Results (figure 7[18]) will generate a tab in the Results containing information on the results of the similarity search of the selected sequence. For each of the obtained hits, the following information is given: Hit id and definition Gene name assigned to the hit by its accession e-value of the alignment Alignment length of the longest hsp Positive matches of the longest hsp Hsp similarity of hit: Number of hsps mapped GO-Terms with its evidence code UniProt codes of the hit sequences.

[16] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598114359#BLAST-figure6
[17] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598114359#BLAST-AdvancedPage
[18] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598114359#BLAST-figure7

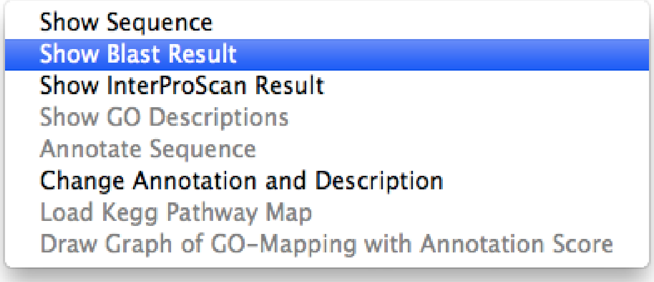
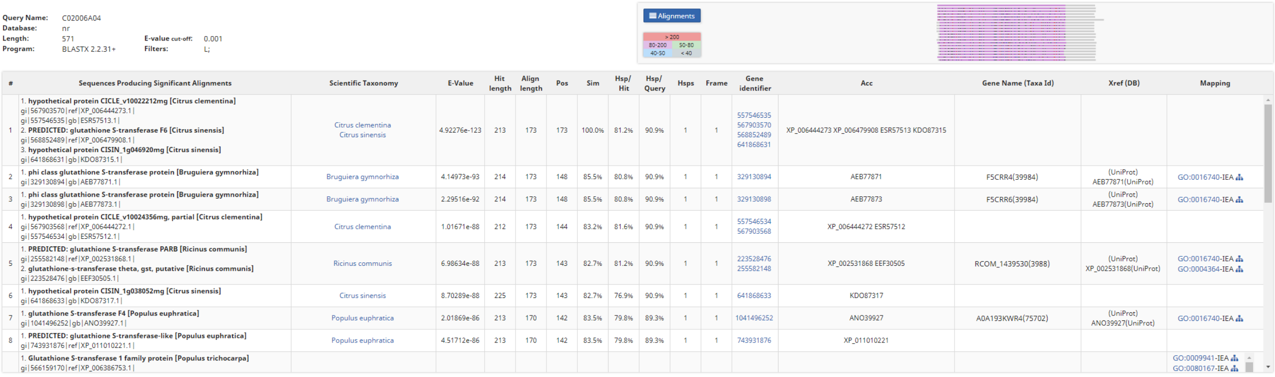


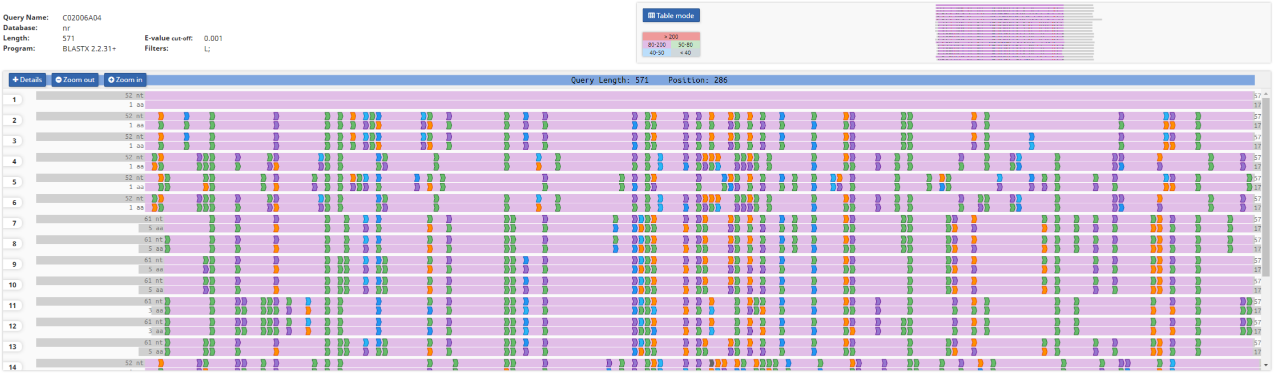
Figure 7: Show BLAST Results



Query Name:	C02006A04		
Database:	nr		
Length:	571	**E-value cut-off:**	0.001
Program:	BLASTX 2.2.31+	**Filters:**	L;

#	Sequences Producing Significant Alignments	Scientific Taxonomy	E-Value	Hit length	Align length	Pos	Sim	Hsp/ Hit	Hsp/ Query	Hsps	Frame	Gene identifier	Acc	Gene Name (Taxa Id)	Xref (DB)	Mapping
1	1. **hypothetical protein CICLE_v10022212mg [Citrus clementina]** gi|567903570|ref|XP_006444273.1| gi|557546535|gb|ESR57513.1| 2. **PREDICTED: glutathione S-transferase F6 [Citrus sinensis]** gi|568852489|ref|XP_006479908.1| 3. **hypothetical protein CISIN_1g046920mg [Citrus sinensis]** gi|641868631|gb|KDO87315.1|	Citrus clementina Citrus sinensis	4.92276e-123	213	173	173	100.0%	81.2%	90.9%	1	1	557546535 567903570 568852489 641868631	XP_006444273 XP_006479908 ESR57513 KDO87315			
2	1. **phi class glutathione S-transferase protein [Bruguiera gymnorhiza]** gi|329130894|gb|AEB77871.1|	Bruguiera gymnorhiza	4.14973e-93	214	173	148	85.5%	80.8%	90.9%	1	1	329130894	AEB77871	F5CRR4(39984)	(UniProt) AEB77871(UniProt)	GO:0016740-IEA
3	1. **phi class glutathione S-transferase protein [Bruguiera gymnorhiza]** gi|329130898|gb|AEB77873.1|	Bruguiera gymnorhiza	2.29516e-92	214	173	148	85.5%	80.8%	90.9%	1	1	329130898	AEB77873	F5CRR6(39984)	(UniProt) AEB77873(UniProt)	GO:0016740-IEA
4	1. **hypothetical protein CICLE_v10024356mg, partial [Citrus clementina]** gi|567903568|ref|XP_006444272.1| gi|557546534|gb|ESR57512.1|	Citrus clementina	1.01671e-88	212	173	144	83.2%	81.6%	90.9%	1	1	557546534 567903568	XP_006444272 ESR57512			
5	1. **PREDICTED: glutathione S-transferase PARB [Ricinus communis]** gi|255582148|ref|XP_002531868.1| 2. **glutathione-s-transferase theta, gst, putative [Ricinus communis]** gi|223528476|gb|EEF30505.1|	Ricinus communis	6.98634e-88	213	173	143	82.7%	81.2%	90.9%	1	1	223528476 255582148	XP_002531868 EEF30505	RCOM_1439530(3988)	(UniProt) XP_002531868(UniProt)	GO:0016740-IEA GO:0004364-IEA
6	1. **hypothetical protein CISIN_1g038052mg [Citrus sinensis]** gi|641868633|gb|KDO87317.1|	Citrus sinensis	8.70289e-88	225	173	143	82.7%	76.9%	90.9%	1	1	641868633	KDO87317			
7	1. **glutathione S-transferase F4 [Populus euphratica]** gi|1041496252|gb|ANO39927.1|	Populus euphratica	2.01869e-86	213	170	142	83.5%	79.8%	89.3%	1	1	1041496252	ANO39927	A0A193KWR4(75702)	(UniProt) ANO39927(UniProt)	GO:0016740-IEA
8	1. **PREDICTED: glutathione S-transferase-like [Populus euphratica]** gi|743931876|ref|XP_011010221.1|	Populus euphratica	4.51712e-86	213	170	142	83.5%	79.8%	89.3%	1	1	743931876	XP_011010221			
	1. **Glutathione S-transferase 1 family protein [Populus trichocarpa]** gi|566159170|ref|XP_006386753.1|															GO:0009941-IEA GO:0080167-IEA

Figure 8: Individual BLAST Result Table View

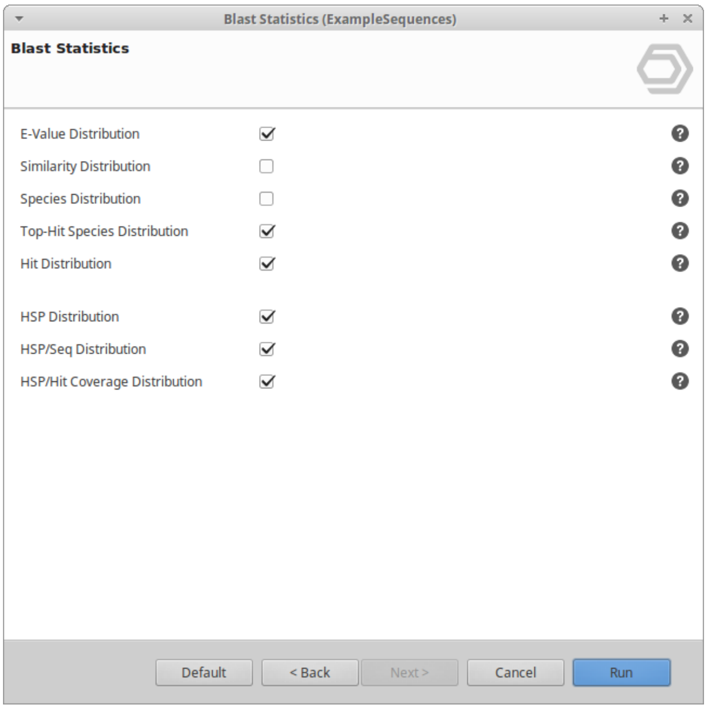


Query Name:	C02006A04		
Database:	nr		
Length:	571	**E-value cut-off:**	0.001
Program:	BLASTX 2.2.31+	**Filters:**	L;

Figure 9: Individual BLAST Result in Alignment View

7.3.5 Statistics

Different BLAST statistics charts (Figure figure 11[19], figure 1[20]2 and figure 13[21]) can be generated for a global visualization of the results. These charts provide a general view of the similarity of the query set with the selected databases and can be used to choose cut-off levels for the e-value, similarity and annotation threshold parameters at the annotation step.
Additionally, a BLAST hit species distribution chart is available. To generate the BLAST Statistics charts just go to the arrow next to the "Chart" icon and select the statistics to be displayed (see figure 10).



Blast Statistics (ExampleSequences)

Blast Statistics

E-Value Distribution ☑
Similarity Distribution ☐
Species Distribution ☐
Top-Hit Species Distribution ☑
Hit Distribution ☑
HSP Distribution ☑
HSP/Seq Distribution ☑
HSP/Hit Coverage Distribution ☑

Figure 10: Blast Statistics

- E-Value Distribution: This chart plots the distribution of E-values for all selected BLAST hits. It is useful to evaluate the success of the alignment for a given sequence database and help to adjust the E-Value cutoff in the annotation step.
- Similarity Distribution: This chart displays the distribution of all calculated sequence similarities (percentages), shows the overall performance of the alignments and helps to adjust the annotation score in the annotation step.
- Species Distribution: This chart gives a listing of the different species to which most sequences were aligned during the BLAST step.
- Top-Hit Species Distribution: Bar chart showing the species distribution of all Top-Blast hits.
- Hit Distribution: This chart shows a distribution of the number of hits for the blasted sequences in a data-set.
- Hsp Distribution: This bar chart shows the distribution of hsps per hit.
- Hsp/Seq Distribution: This chart shows a distribution of percentages which represents the coverage between the hsps and their corresponding sequences.

[19] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598114359#BLAST-figure10
[20] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598114359#BLAST-figure11
[21] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598114359#BLAST-figure12

- Hsp/Hit Distribution: Same as above but for hits instead of sequences.

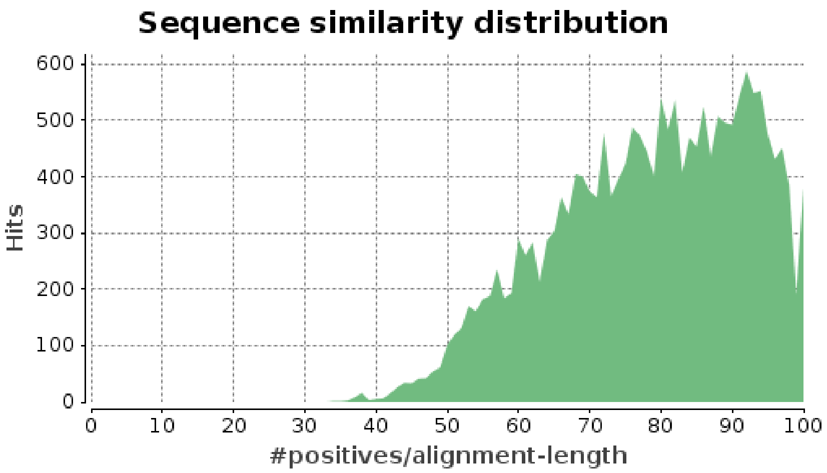


Figure 11: Similarity Distribution

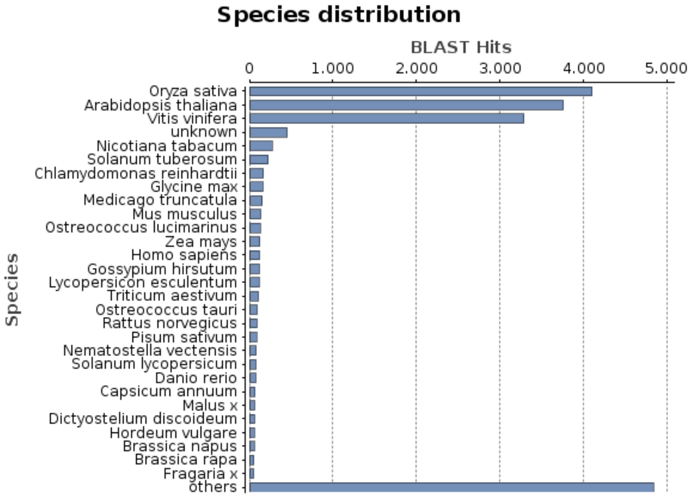


Figure 12: Species Distribution

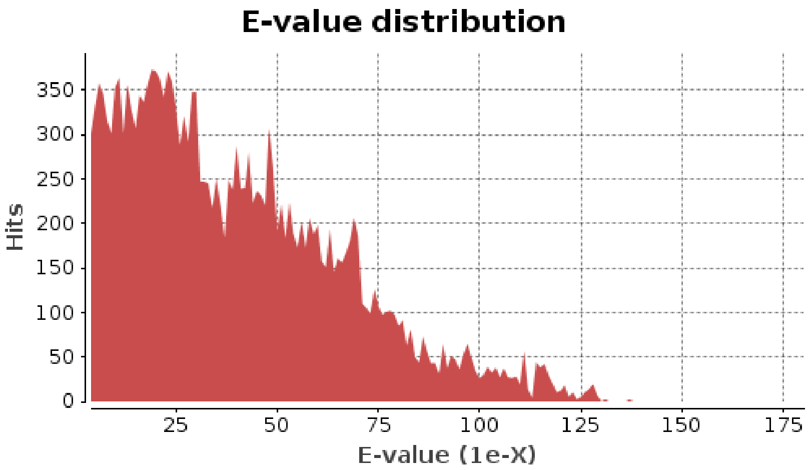
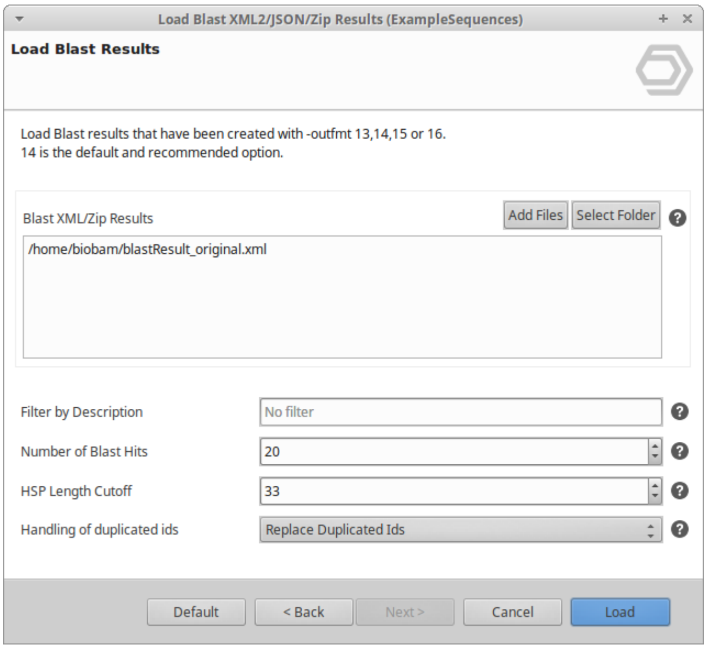


Figure 13: E-Value Distribution

7.3.6 Load BLAST results

If a BLAST result is already available in XML format, it can be directly loaded into OmicsBox by using Load > Load Blast Results in the File menu. You can choose here to import the Blast results as XML file or the new XML2/JSON format. These new formats can be loaded as Zip file.
In the Load Blast Results dialog a whole directory containing a collection of BLAST XML files or a single XML file can be selected Figure 14. The BLAST results will be added to your current OmicsBox session.
OmicsBox also allows the input of TimeLogic DeCypher Blast results.



Load Blast results that have been created with -outfmt 13,14,15 or 16.
14 is the default and recommended option.

Blast XML/Zip Results: /home/biobam/blastResult_original.xml

Filter by Description: No filter

Number of Blast Hits: 20

HSP Length Cutoff: 33

Handling of duplicated ids: Replace Duplicated Ids

Figure 14: Load / Import Blast Results

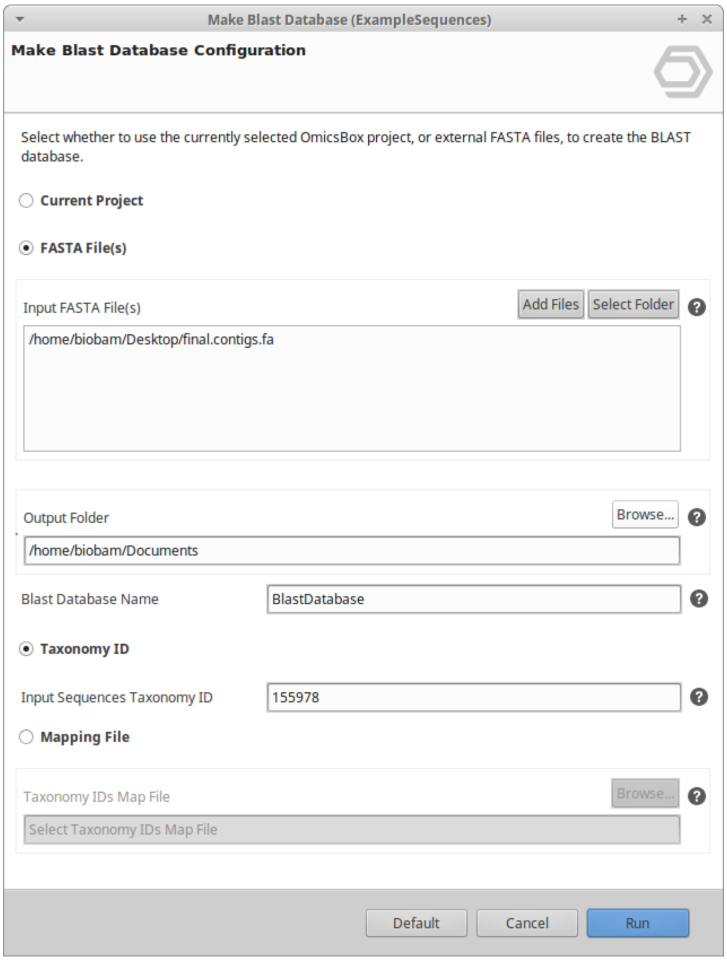
7.3.7 Make Blast Database

This option allows creating a BLAST database from the sequence of any OmicsBox project or from a FASTA file (figure 15[22]). This option can be found in the arrow next to the blast icon.

- Current project: OmicsBox will use the loaded sequences to create the Blast database. Note: If the resulting database will be used for further GO mapping a proper ID and description line with "GO mappable" information are needed.
- FASTA file: This option allows choosing own FASTA file. The FASTA file has to be correctly formatted for NCBI Blast+.
- Output Folder: Select the directory where to save the created Blast database.
- Blast Database Name: Provide a name for the Blast database
- Taxonomy Options:
    - Taxonomy ID: Introduce the NCBI species ID
    - Mapping file: If the sequences come from different species, it is possible to generate a text file with the sequence names and its species id to map to the corresponding sequence in the FASTA file.
      Example:
      ```
      TR|A0A022PMT6|ERYGU    4155
      TR|A0A022PMU0|ERYGU    4155
      TR|A0A059BJ72|EUCGR    71139
      TR|A0A059BJ72|EUCGR    71139
      TR|A0A061FDU3|THECC    3641
      TR|A0A067DJ79|CITSI    2711
      ```

⚠ Visit the following tutorial[23] for more information on how to create the Taxonomy ID file.

[22] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598114359#BLAST-figure14
[23] https://www.biobam.com/taxonomic-mapping-file-make-blast-database-within-omicsbox/



Make Blast Database (ExampleSequences)
Make Blast Database Configuration

Select whether to use the currently selected OmicsBox project, or external FASTA files, to create the BLAST database.

○ Current Project

● FASTA File(s)

Input FASTA File(s): /home/biobam/Desktop/final.contigs.fa

Output Folder: /home/biobam/Documents

Blast Database Name: BlastDatabase

● Taxonomy ID

Input Sequences Taxonomy ID: 155978

○ Mapping File

Taxonomy IDs Map File: Select Taxonomy IDs Map File

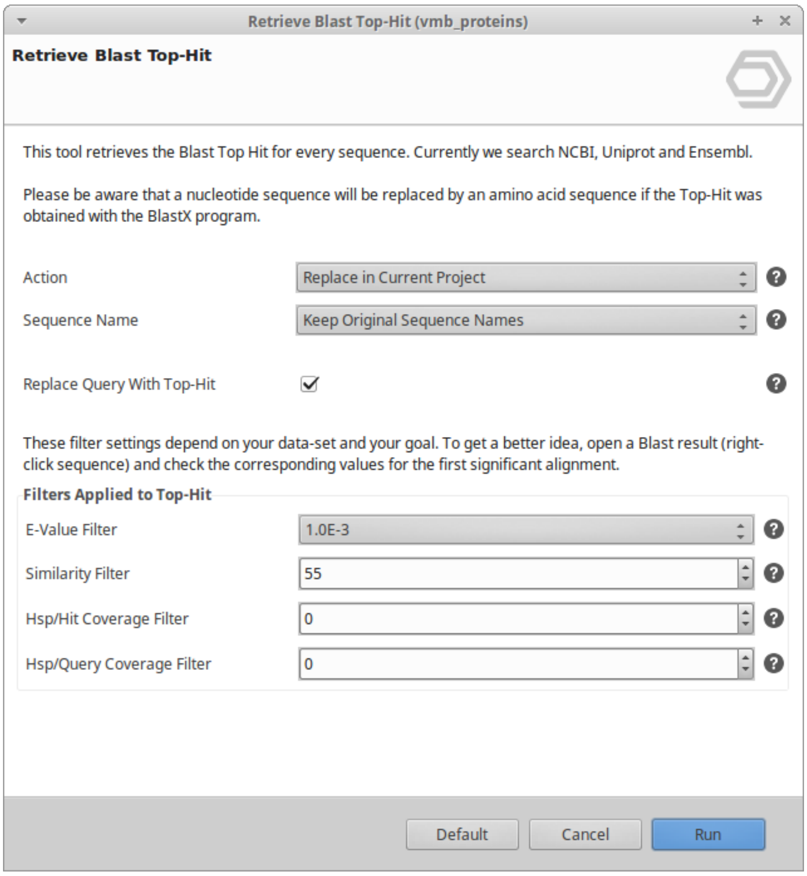
Figure 15: Make Blast Database

7.3.8 Retrieve Blast Top-Hit

This feature allows retrieving the sequence information of Top Blast Hits in an OmicsBox project. Data can be obtained from the NCBI, Ensembl or Uniprot web services and stored in a new project or replace the existing IDs/sequences (see figure 16). A possible use case scenario would be a so-called "Double-Blast": The blast results of a first run are used to replace the sequence data for a second run against a different set of query sequences. Imagine an RNA-seq data-set with a high percentage of sequences without any alignments against a protein database (e.g. blastx against NR). This feature could be used to select and extract the sequences without hits (red ones) into a new project. These sequences could be basted first against a set of EST sequences. The initial unaligned sequences are now replaced with the ESTs. Now the initial blastx search is repeated again the protein.

For each Top-Hit (first significant alignment from an already performed BLAST), apply the filters (bottom part of the dialog) and search them in the corresponding database (online).
It is possible to either replace the sequence from your data-set or to extract them into a new data-set (Action option). You can also decide whether you want to keep the original sequence names or if you want to rename them to the downloaded sequences names. The latter will add a small note to the sequence description, telling you the original name.

The last remaining option allows you to decide whether you want to replace your sequences with the downloaded ones or if you just want to retrieve their names. This option is activated by default.



Retrieve Blast Top-Hit (vmb_proteins)

Retrieve Blast Top-Hit

This tool retrieves the Blast Top Hit for every sequence. Currently we search NCBI, Uniprot and Ensembl.

Please be aware that a nucleotide sequence will be replaced by an amino acid sequence if the Top-Hit was obtained with the BlastX program.

Action: Replace in Current Project

Sequence Name: Keep Original Sequence Names

Replace Query With Top-Hit: ☑

These filter settings depend on your data-set and your goal. To get a better idea, open a Blast result (right-click sequence) and check the corresponding values for the first significant alignment.

Filters Applied to Top-Hit

E-Value Filter: 1.0E-3

Similarity Filter: 55

Hsp/Hit Coverage Filter: 0

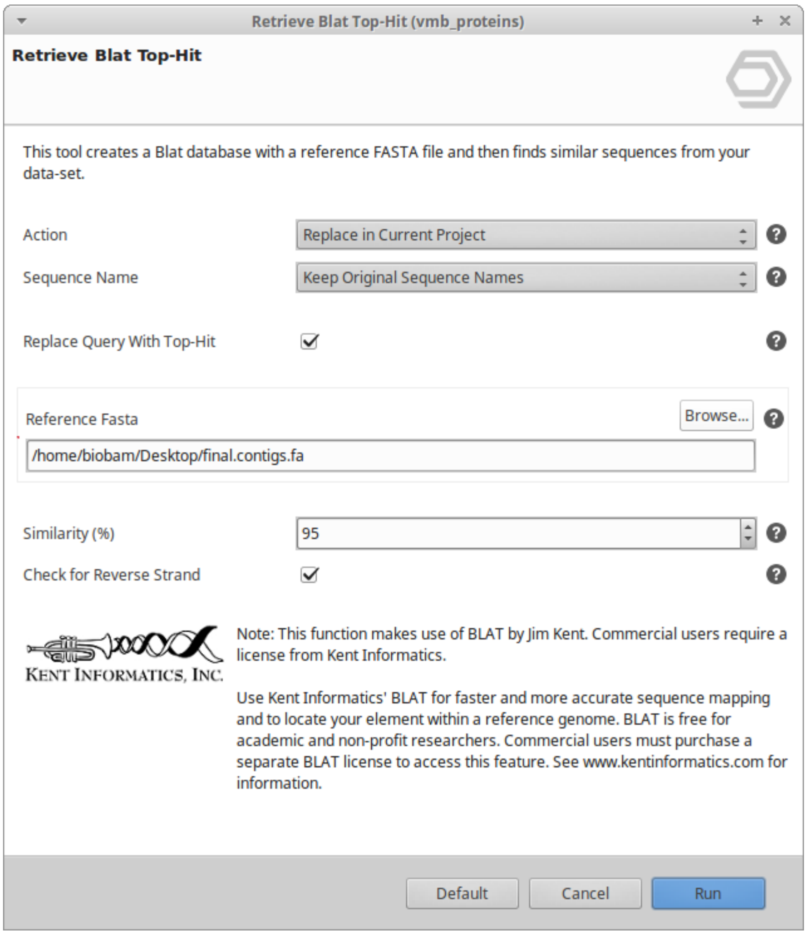
Hsp/Query Coverage Filter: 0

Figure 16: Retrieve Blast Top-Hit Dialog.

7.3.9 Retrieve Blat Top-Hit

This tool is very similar to "Retrieve Blast Top-Hit" explained above, but it employs BLAT[24] instead (figure 17). The dialog is therefore quite similar and the first 3 options are identical. BLAT needs a reference FASTA file which it uses to search for similar sequences. The last 2 options allow you to filter by similarity and if BLAT should consider the reverse strand.

[24] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC187518/



Retrieve Blat Top-Hit (vmb_proteins)

Retrieve Blat Top-Hit

This tool creates a Blat database with a reference FASTA file and then finds similar sequences from your data-set.

Action: Replace in Current Project

Sequence Name: Keep Original Sequence Names

Replace Query With Top-Hit ☑

Reference Fasta: /home/biobam/Desktop/final.contigs.fa

Similarity (%): 95

Check for Reverse Strand ☑

Note: This function makes use of BLAT by Jim Kent. Commercial users require a license from Kent Informatics.

Use Kent Informatics' BLAT for faster and more accurate sequence mapping and to locate your element within a reference genome. BLAT is free for academic and non-profit researchers. Commercial users must purchase a separate BLAT license to access this feature. See www.kentinformatics.com for information.

Figure 17: Retrieve Blat Top-Hit Dialog.

This tool can be useful after running Prokaryotic Gene Finding, in order to replace the sequence names retrieved from Glimmer by the top-hit from a reference fasta. For further details, click here[25].

7.3.10 Other BLAST Functions

- Remove Blast Results: This option will remove the BLAST results from the selected sequences.
- Run Blast-Descriptor-Annotator (BDA): This will run the BDA algorithm. For further details, please see Blast Configuration Page section.
- Recover original Best-Blast-Hit Description: When this option is executed the sequence description column on the Main Sequence Table will contain the top blast hit description and not the one from the BDA.

[25] https://www.biobam.com/change-genefinding-sequence-name/

7.4 Gene Ontology Mapping

Content of this page:

7.4.1 General

Mapping is the process of retrieving GO terms associated with the Hits obtained by the BLAST search. OmicsBox performs four different mappings steps:

1. BLAST result accessions are used to retrieve gene names or symbols making use of two mapping files provided by the NCBI (gene_info, gene2accession). Identified gene names are then searched in the species-specific entries of the gene-product table of the GO database.
2. GeneBank identifiers (gi), the primary blast Hit ids, are used to retrieve UniProt IDs making use of a mapping file from PIR (Non-redundant Reference Protein Database) including PSD, UniProt, Swiss-Prot, TrEMBL, RefSeq, GenPept and PDB.
3. Accessions are searched directly in the dbxref table of the GO database.
4. BLAST result accessions are searched directly in the gene-product table of the GO database.

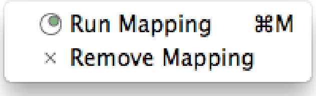


Figure 1: Mapping options

- Run Mapping. Mapping will start.
- Remove Mapping. Delete Mapping results for the selected sequences.

⚠ The mapping step needs protein ids to run. Make sure you ran blast against a protein database.
blastx - if one has nucleotide sequences
blastp - if one has protein sequences

7.4.2 Show Individual Mapping Results

For each sequence, it is possible to see the mapping results individually.

1. Show Mapping Results. A new table will be displayed (see figure 3[26]). The resulting table shows the GO mapping results for a particular sequence. See Table section to manipulate/extract the results from this table.
2. Show GO Descriptions. GO ID, description, type, and definition are given for all GO terms associated with the selected sequence. The GO ID is linked to the AmiGO browser at the Gene Ontology site while the show option displays the DAG representation of the GO term.
3. Annotate Sequence. This function allows changing annotation parameters for the selected sequence and re-running automatic annotation.
4. Change Annotation and Description. This function edits the annotation of the selected and allows typing and deleting of annotation or sequence description. A manual annotation check-box (see figure 5 in Gene Ontology Annotation section(see page 55)) is available for marking sequences with manual annotation. The sequence will get the pink label on the Main Sequence Table.
5. Make Graph of GO-Mapping-Results with Annotation Score. Displays a DAG with all GO terms related to one sequence. Shows all the GOs from the mapping step as well as final annotations (highlighted). The wizard (figure 4[27] allows filtering the hits which will be taken into account (see Gene Ontology Graphs(see page 79) section for more details about visualization in OmicsBox)
    a. Hit Filter. Nodes can be filtered out by a number of hits: only nodes with more than a given number of BLAST-Hits will be shown in the graph.
    b. HSP-Hit Coverage CutOff: Includes only those hits which are overage with the HSP for a given percentage.

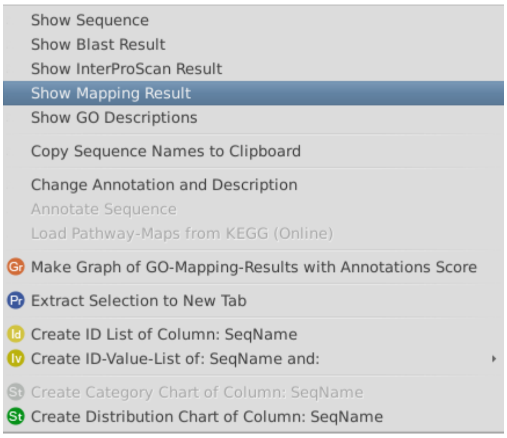
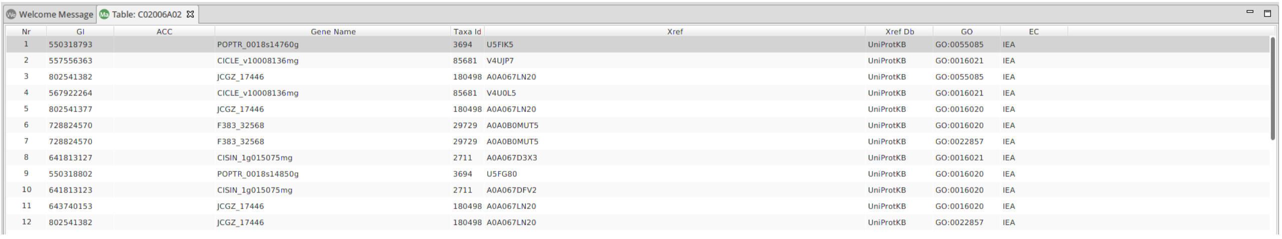

Show Sequence
Show Blast Result
Show InterProScan Result
Show Mapping Result
Show GO Descriptions
Copy Sequence Names to Clipboard
Change Annotation and Description
Annotate Sequence
Load Pathway-Maps from KEGG (Online)
Gr Make Graph of GO-Mapping-Results with Annotations Score
Pr Extract Selection to New Tab
Id Create ID List of Column: SeqName
Iv Create ID-Value-List of: SeqName and:
St Create Category Chart of Column: SeqName
St Create Distribution Chart of Column: SeqName

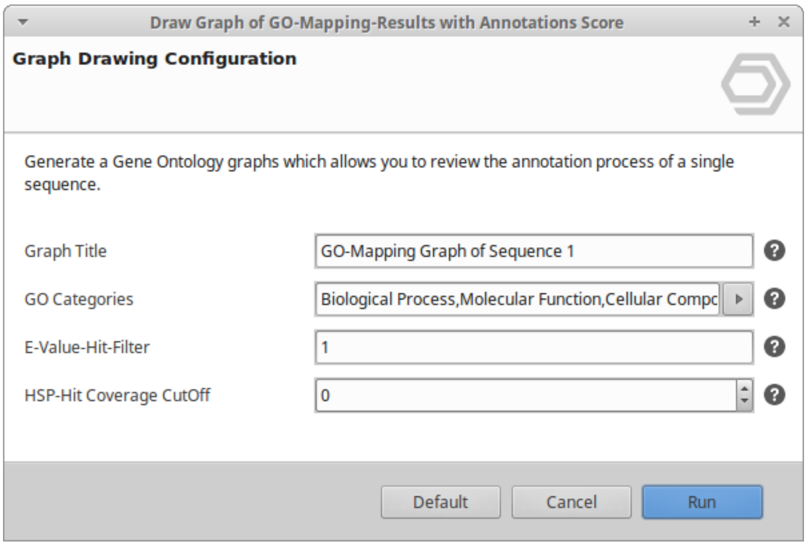
Figure 2: Show Mapping Results



Nr	GI	ACC	Gene Name	Taxa Id	Xref	Xref Db	GO	EC
1	550318793		POPTR_0018s14760g	3694	U5FIK5	UniProtKB	GO:0055085	IEA
2	557556363		CICLE_v10008136mg	85681	V4UJP7	UniProtKB	GO:0016021	IEA
3	802541382		JCGZ_17446	180498	A0A067LN20	UniProtKB	GO:0055085	IEA
4	567922264		CICLE_v10008136mg	85681	V4U0L5	UniProtKB	GO:0016021	IEA
5	802541377		JCGZ_17446	180498	A0A067LN20	UniProtKB	GO:0016020	IEA
6	728824570		F383_32568	29729	A0A0B0MUT5	UniProtKB	GO:0016020	IEA
7	728824570		F383_32568	29729	A0A0B0MUT5	UniProtKB	GO:0022857	IEA
8	641813127		CISIN_1g015075mg	2711	A0A067D3X3	UniProtKB	GO:0016021	IEA
9	550318802		POPTR_0018s14850g	3694	U5FG80	UniProtKB	GO:0016020	IEA
10	641813123		CISIN_1g015075mg	2711	A0A067DFV2	UniProtKB	GO:0016020	IEA
11	643740153		JCGZ_17446	180498	A0A067LN20	UniProtKB	GO:0016020	IEA
12	802541382		JCGZ_17446	180498	A0A067LN20	UniProtKB	GO:0022857	IEA

Figure 3: Mapping Results for sequence C02006A02

[26] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598114410#GeneOntologyMapping-figure3
[27] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598114410#GeneOntologyMapping-figure4



Graph Drawing Configuration

Generate a Gene Ontology graphs which allows you to review the annotation process of a single sequence.

Graph Title: GO-Mapping Graph of Sequence 1

GO Categories: Biological Process,Molecular Function,Cellular Compc

E-Value-Hit-Filter: 1

HSP-Hit Coverage CutOff: 0

Figure 4: Single Graph Drawing Configuration

7.4.3 Statistics

If a BLAST result is successfully mapped to one or several GO terms, these will be shown in the GOs column of the Main Sequence Table and this sequence row will turn **light-green**. Assigned GOs can be reviewed in the BLAST results Table (see Show BLAST Results(see page 45) section and BLAST figure 8(see page 46) of that section).

Three different charts are available to summarise the mapping step:

- GO Mapping Distribution: Shows the distribution of the amount of Gene Ontology candidate terms assigned to each sequence during the GO Mapping step.
- EC Distribution for Blast Hits (figure 5[28]): Evidence Codes associated to the obtained GO pool
- EC Distribution for Sequences (figure 6[29]): This chart shows the distribution of GO evidence codes for the functional terms obtained during the mapping step. It gives an idea about how many annotations derive from automatic/computational annotations or manually curated ones.
- DB Resources of Mapping (figure 7[30]): This chart gives the distribution of the number of annotations (GO-terms) retrieved from the different source databases e.g. UniProt, PDB, TAIR etc.

⚠ Commonly IEA (electronic annotation) is overwhelmed in the mapping results. However, the contribution of this (and other) type of annotation to the finally assigned annotation to the query set can be modulated at the annotation step.

Mapping Statistics Graphs:

[28] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598114410#GeneOntologyMapping-figure5
[29] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598114410#GeneOntologyMapping-figure6
[30] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598114410#GeneOntologyMapping-figure7

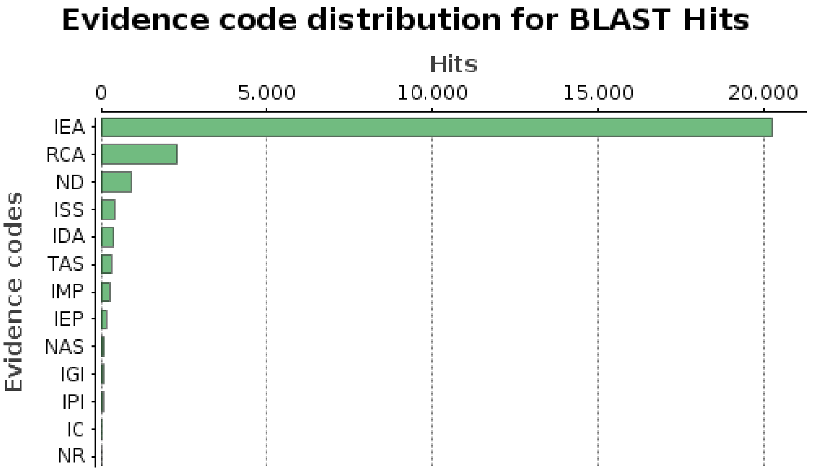


Figure 5: Evidence Code Distribution of BLAST hits

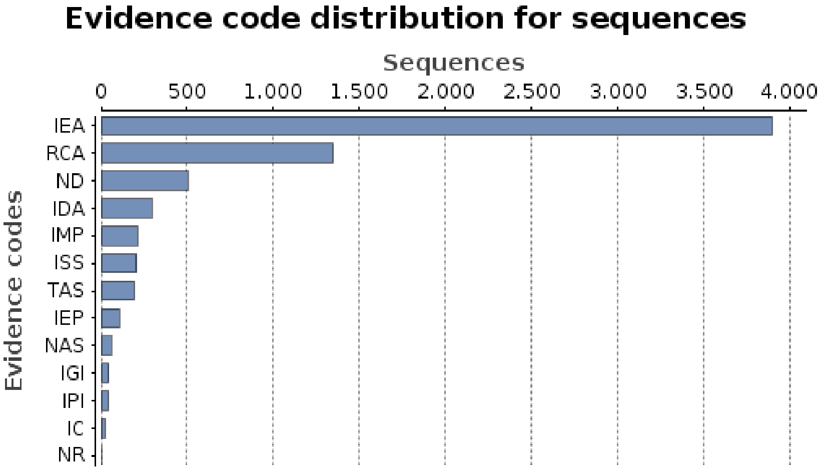


Figure 6: Evidence Code Distribution for sequences

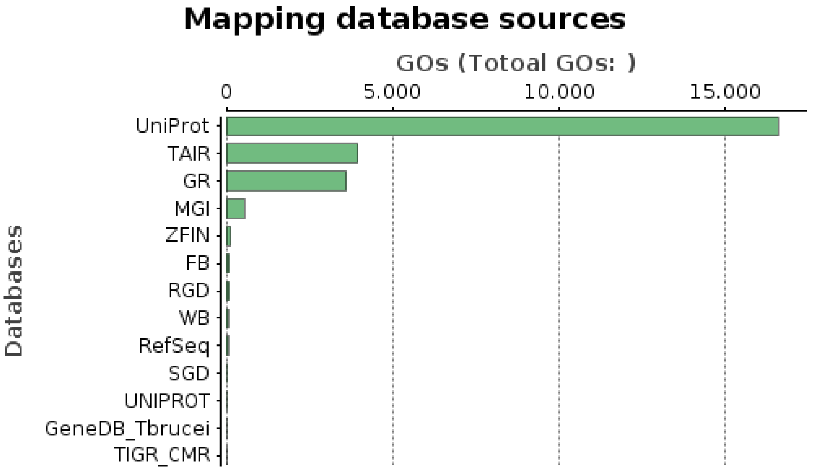


Figure 7: DB Resources of Mapping

7.4.4 Export Mapping Results

A tab separator text file can be exported with the corresponding mapping results (**File > Export > Export Mapping Results**).

7.5 Gene Ontology Annotation

Content of this page:

7.5.1 Annotation Rule

This is the process of selecting GO terms from the GO pool obtained by the Mapping step and assigning them to the query sequences. In the current OmicsBox version, this is the core type of functional annotation.

GO annotation is carried out by applying an annotation rule (AR) on the found ontology terms. The rule seeks to find the most specific annotations with a certain level of reliability. This process is adjustable in specificity and stringency.

For each candidate GO an annotation score (AS) is computed. The AS is composed of two additive terms.

The first, direct term (DT), represents the highest hit similarity of this GO weighted by a factor corresponding to its EC.

The second term (AT) of the AS provides the possibility of abstraction. This is defined as an annotation to a parent node when several child nodes are present in the GO candidate collection. This term multiplies the number of total GOs unified at the node by a user-defined GO weight factor that controls the possibility and strength of abstraction. When GO weight is set to 0, no abstraction is done.

Finally, the AR selects the lowest term per branch that lies over a user-defined threshold. DT, AT and the AR terms are defined as given in figure 1[31].

[31] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598179910#GeneOntologyAnnotation-figure1

$$DT = \max(similarity \times EC_{weight})$$
$$AT = (\#GO - 1) \times GO_{weight}$$
$$AR : lowest.node(AS(DT + AT)) \geq threshold$$

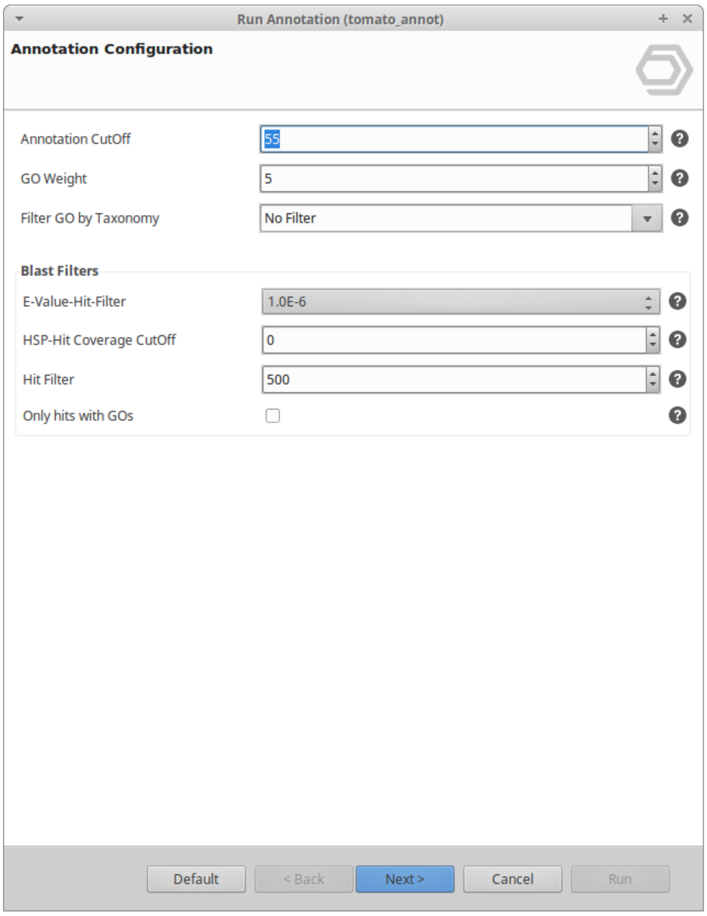
Figure 1: OmicsBox Annotation Rule

To better understand how the annotation score works, the following reasoning can be done: When EC-weight is set to 1 for all ECs (no EC influence) and GO-weight equals zero (no abstraction), then the annotation score equals the maximum similarity value of the hits that have that GO term and the sequence will be annotated with that GO term if that score is above the given threshold provided. The situation when EC-weights are lower than 1 means that higher similarities are required to reach the threshold. If the GO-weight is different to 0 this means that the possibility is enabled that a parent node will reach the threshold while its various children nodes would not.

The annotation rule provides a general framework for annotation. The actual way annotation occurs depends on how the different parameters at the AS are set. These can be adjusted in the Annotation Configuration Dialog (figure 2[32]) and in the Evidence Code Weight Configuration Dialog (figure 3[33]).

1. Annotation Cut-Off (threshold).The annotation rule selects the lowest term per branch that lies over this threshold (default=55).
2. GO-Weight. This is the weight given to the contribution of mapped children terms to the annotation of a parent term (default=5).
3. Filter GO by taxonomy: The filter will remove the Gene Ontology terms known not to be in the given taxonomy using the restrictions defined by Gene Ontology. You can select one of the given options or simply write a taxonomy id.
4. E-Value-Hit-Filter. This value can be understood as a pre-filter: only GO terms obtained from hits with a greater e-value than given will be used for annotation and/or shown in a generated graph (default=1.0E-6).
5. Hsp-HitCoverage CutOff. Sets the minimum needed coverage between a Hit and his HSP. For example, a value of 80 would mean that the aligned HSP must cover at least 80% of the longitude of its Hit. Only annotations from Hit fulfilling this criterion will be considered for annotation transference.
6. Hit Filter. This option allows you to consider only the first N hits during annotation. This option is correlative with "Only hits with GOs" feature.
7. Only hits with GOs. This option together with the "Hit Filter" option allows to apply it only on hits that have a GO term candidate.
8. EC-Weight. EC code weights can be modified at the following pages of the Run Annotation dialogue by clicking Next. Note that in case of influence by evidence codes is not wanted, you can set them all at 1. Alternatively, when you want to exclude GO annotations of a certain EC (for example IEAs), you can set this EC weight at 0.

[32] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598179910#GeneOntologyAnnotation-figure2
[33] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598179910#GeneOntologyAnnotation-figure3



Annotation Configuration

Annotation CutOff: 55

GO Weight: 5

Filter GO by Taxonomy: No Filter

Blast Filters

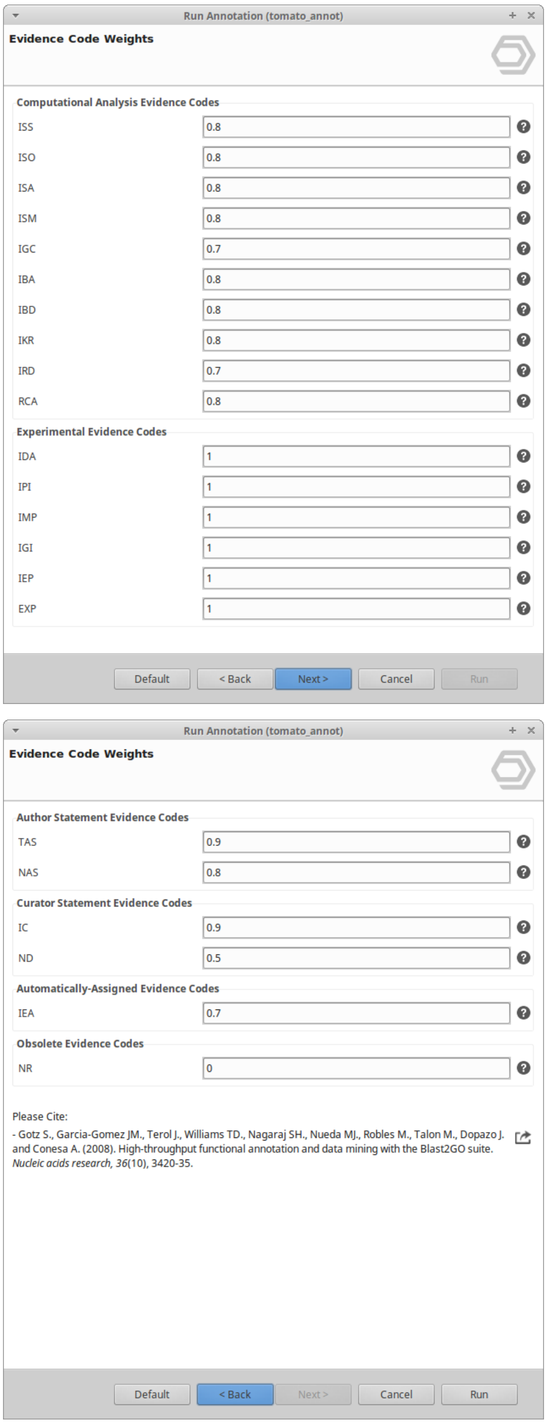
E-Value-Hit-Filter: 1.0E-6

HSP-Hit Coverage CutOff: 0

Hit Filter: 500

Only hits with GOs: □

Figure 2: Annotation Configuration



Computational Analysis Evidence Codes

ISS	0.8
ISO	0.8
ISA	0.8
ISM	0.8
IGC	0.7
IBA	0.8
IBD	0.8
IKR	0.8
IRD	0.7
RCA	0.8

Experimental Evidence Codes

IDA	1
IPI	1
IMP	1
IGI	1
IEP	1
EXP	1

Author Statement Evidence Codes

TAS	0.9
NAS	0.8

Curator Statement Evidence Codes

IC	0.9
ND	0.5

Automatically-Assigned Evidence Codes

IEA	0.7

Obsolete Evidence Codes

NR	0

Please Cite:

- Gotz S., Garcia-Gomez JM., Terol J., Williams TD., Nagaraj SH., Nueda MJ., Robles M., Talon M., Dopazo J. and Conesa A. (2008). High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic acids research, 36*(10), 3420-35.

Figure 3: Evidence Code weight configuration

Successful annotation for each query sequence will result in a color change for that sequence from light-green to **blue** at the Main Sequence Table, and only the annotated GOs will remain in the GO IDs column.

7.5.2 Individual Annotation Results

Annotation results for each sequence can also be visualized on the GO DAG by selecting "Draw Graph of GO-Mapping with Annotation Score" at the context menu. Additionally, the "Change Annotation and Description" options of this menu offer also the possibility to adjust annotations specifically for a single sequence.
This function edits the annotation of the selected and allows typing and deleting of annotation or sequence description. A manual annotation check-box (see figure 5[34]) is available for marking sequences with manual annotation. The sequence will get the **pink** label on the Main Sequence Table.

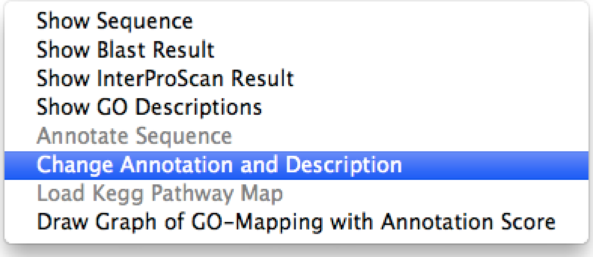
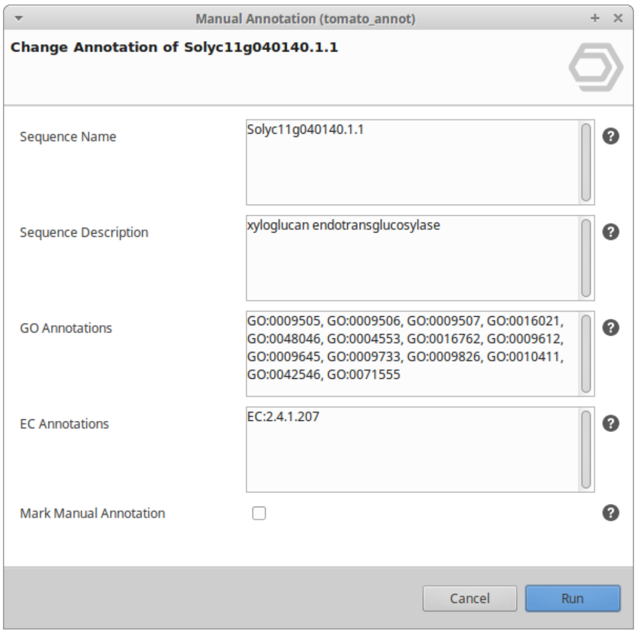


Figure 4: Manually change Annotation and Description



Change Annotation of Solyc11g040140.1.1

Sequence Name: Solyc11g040140.1.1

Sequence Description: xyloglucan endotransglucosylase

GO Annotations: GO:0009505, GO:0009506, GO:0009507, GO:0016021, GO:0048046, GO:0004553, GO:0016762, GO:0009612, GO:0009645, GO:0009733, GO:0009826, GO:0010411, GO:0042546, GO:0071555

EC Annotations: EC:2.4.1.207

Mark Manual Annotation ☐

Figure 5: Mark Manual Annotation

[34] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598179910#GeneOntologyAnnotation-figure5

7.5.3 Statistics

An overview of the extent and intensity of the annotation can be obtained from the Annotation Distribution Chart (Figure 6[35]), which shows the number of sequences annotated with different amounts of GO-terms.

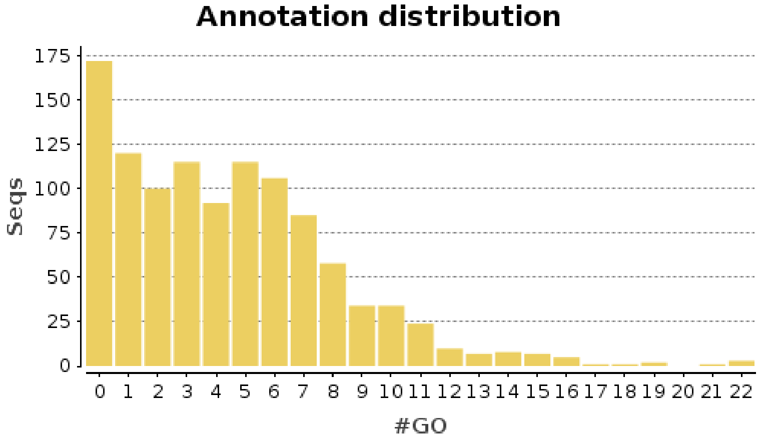


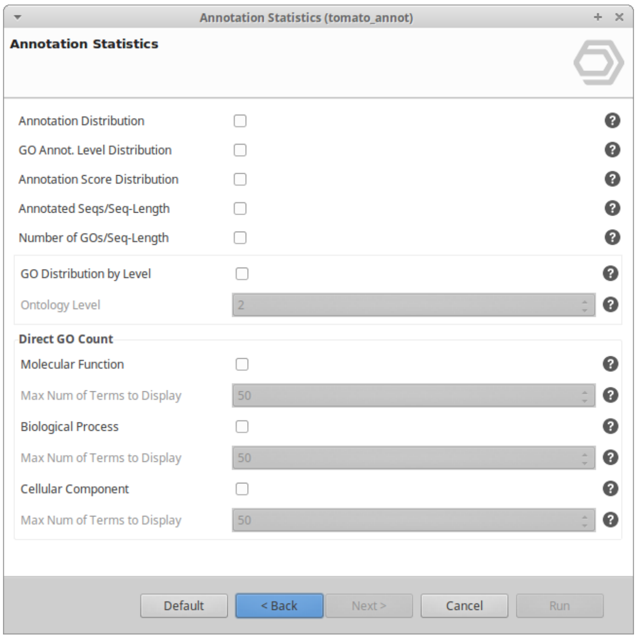
Figure 6: Annotation Distribution

In order to display the Annotation Statistics Wizard go to the "charts" icon in the main toolbar and select "Annotation Statistics".

The following statistics are available:

- Annotation Distribution: This chart informs about the number of GO terms assigned per sequence.
- GO Annotation Level Distribution: A bar chart which shows all GO terms for all 3 categories for a given GO level taking into account the GO hierarchy (parent-child relationships).
- Annotation Score Distribution: A chart that shows the number of sequences per annotation score.
- Annotated Seqs/Seq-Length: Shows the relation between the amount of annotated sequences and sequence lengths.
- Number of GOs/Seq-Length: Shows the relation between sequence length and number of GOs.
- Go Distribution by Level: A bar chart which shows all the GO terms for all 3 categories for GO level 2, taking into account the GO hierarchy.
- Direct GO Count:
    - Molecular Function: A chart for the Molecular Function GO category, which shows the most frequent GO terms within a data-set without taking into account the GO hierarchy.
    - Biological Process: Same as above but for Biological Process.
    - Cellular Component: Same as above but for Cellular Component.

[35] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598179910#GeneOntologyAnnotation-figure6



Annotation Statistics (tomato_annot)

Annotation Statistics

- Annotation Distribution □
- GO Annot. Level Distribution □
- Annotation Score Distribution □
- Annotated Seqs/Seq-Length □
- Number of GOs/Seq-Length □
- GO Distribution by Level □
- Ontology Level 2

Direct GO Count

- Molecular Function □
- Max Num of Terms to Display 50
- Biological Process □
- Max Num of Terms to Display 50
- Cellular Component □
- Max Num of Terms to Display 50

Figure 7: Annotation Statistics Wizard

7.5.4 Annotate GOs from Blast Descriptions

This tool looks at every significant alignment (**Right-Click > Show Blast Result** on a sequence) for each sequence and searches their description lines for GO ids. These GOs are now directly annotated to the sequence if the alignments similarity passes the desired minimum. Validation can also be applied and is recommended, it will remove intermediate GO terms.

7.5.5 Exporting Annotation

The annotation results can be exported in a variety of formats. This function is available under File > Export > Export Annotation.

1. .annot. This is the default option for Annotation export and the exchange annotation format in OmicsBox. Annotations are provided in a three-column fashion. The first column contains the sequence name, the second the annotation code and the third the sequence description. When multiple annotations for the same sequence are available, these come in subsequent rows. GO and EC annotations are exported jointly in the same format.

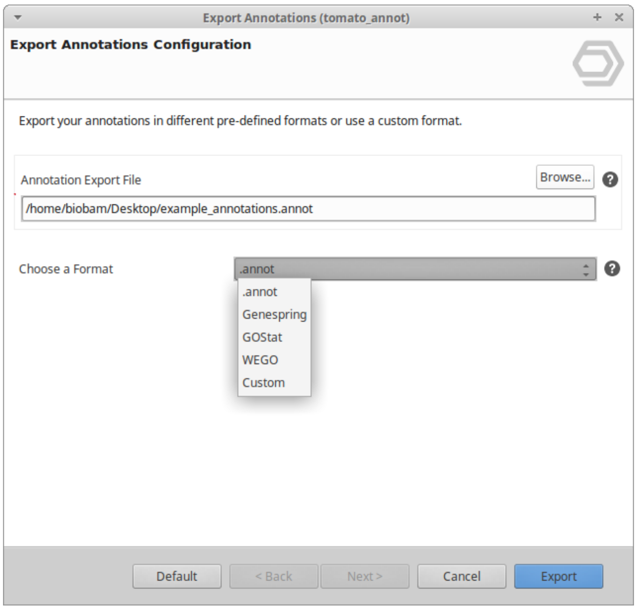
2. Genespring format. One single row is given by sequence where three different columns are provided for Molecular Function, Biological Process and Cellular Component. GO terms are denoted by their description rather than by their code.

3. GoStats format. One single row is given by sequence and GO terms are only denoted by entire numbers ("GO:" and left zero's are skipped)

4. WEGO format (native). One single row is given by sequence, including those without annotated GOs. Belonging GOs are added to each sequence separated by tabs. The format corresponds to the "WEGO native format", shown in this example: http://wego.genomics.org.cn/docs/input01.lst.
5. Custom: It is possible to customize the exportation of the annotation file according to the information desired or the column separator see the next figure.

OmicsBox allows to export additional annotation file formats.

1. Export Annotations in GO Annotation File Format (GAF v.2), which is the primary format currently used by the GO Consortiumhttp://geneontology.org/page/go-annotation-file-formats.
2. Export Annotation Descriptions.
3. Export GO Propagation: Exports the GO parents up to the root for the annotated sequences.
4. Export Sequences per GO (Gene Sets).



Export Annotations (tomato_annot)

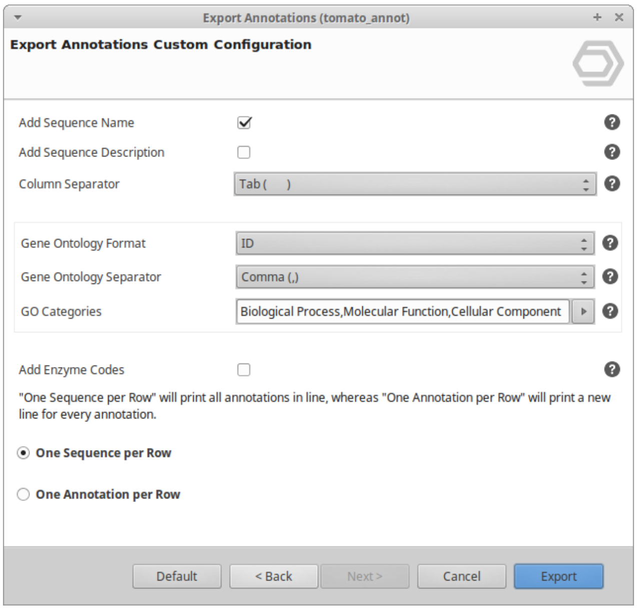
Export Annotations Configuration

Export your annotations in different pre-defined formats or use a custom format.

Annotation Export File: /home/biobam/Desktop/example_annotations.annot

Choose a Format: .annot

Figure 8: Export Annotation Configuration



Export Annotations (tomato_annot)
Export Annotations Custom Configuration

Add Sequence Name ☑

Add Sequence Description □

Column Separator: Tab ( )

Gene Ontology Format: ID

Gene Ontology Separator: Comma (,)

GO Categories: Biological Process,Molecular Function,Cellular Component

Add Enzyme Codes □

"One Sequence per Row" will print all annotations in line, whereas "One Annotation per Row" will print a new line for every annotation.

● One Sequence per Row

○ One Annotation per Row

Figure 9: Export Annotations Custom Configuration

7.5.6 GO-Slim

GO-Slim is a reduced version of the Gene Ontology that contains a selected number of relevant nodes. The *Run GO-Slim (online)* function (under the *Functional analysis* → *Blast2GO Annotation* → *GO-Slim* menu) generates a GO-Slim mapping for the available annotations. Different GO-Slims are available which are adapted to specific organisms.OmicsBox supports the following GO-Slim mappings: General, Plant, Yeast, GOA (GO-Association) and TAIR.

Use the **Functional analysis > Blast2GO Annotation > GO-Slim > Remove GO-Slim** option to return to the original annotations.

7.5.7 Enzyme Code

OmicsBox provides EC annotation through the direct **GO > EC mapping** file available at the GO website. This means that only sequences with GO annotations will eventually show also EC numbers and that the GO annotation accuracy can be made extensive to Enzyme annotations.

7.5.7.1 Statistics

To see the main Enzyme classes in the dataset it is possible to generate a distribution Enzyme Code chart on the "Charts and statistics" menu.

- Main Enzyme Classes: Shows the distribution of the 6 main enzyme classes over all sequences.
- Second Level Classes: Same as above but for the corresponding subclass.

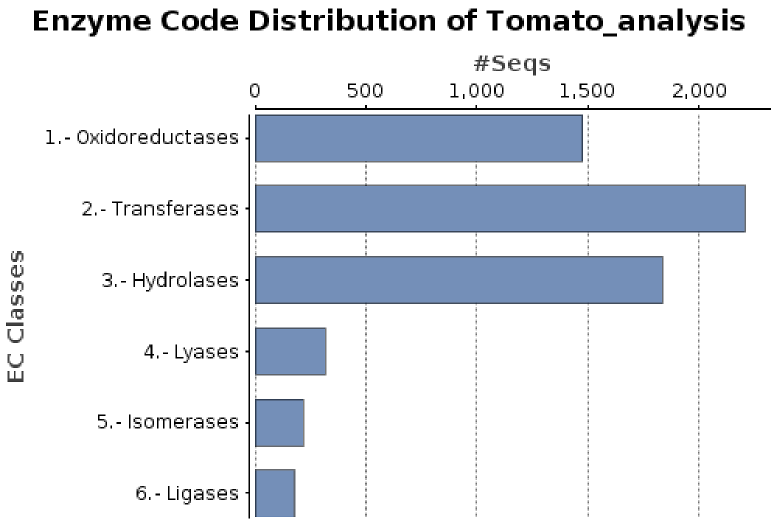
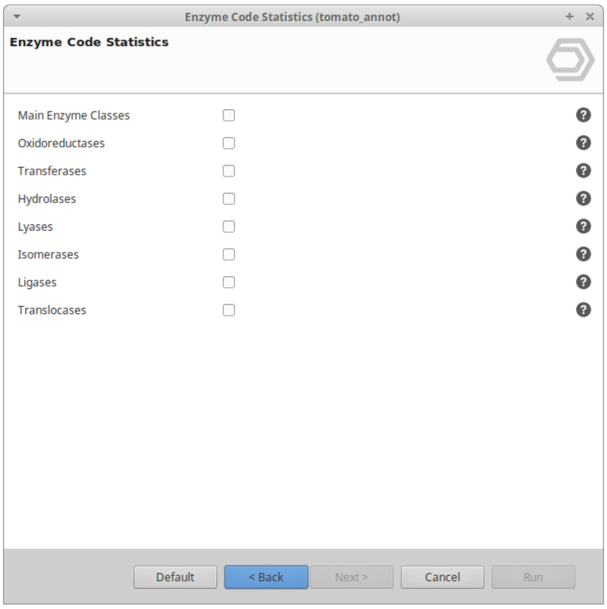


Figure 10: Enzyme Code Distribution


Enzyme Code Statistics (tomato_annot)

Enzyme Code Statistics

Main Enzyme Classes □
Oxidoreductases □
Transferases □
Hydrolases □
Lyases □
Isomerases □
Ligases □
Translocases □

Figure 11: Enzyme Code Statistic

7.5.8 Load Annotation Results (.annot)

Already made or existent annotation can be imported using the .annot format. For import purposes only, the .annot format allows also multiple annotations of the same sequence to be given in one single row, separated by commas, as shown above (Schema: Seq-Name <tab>GO(s) or EC(s) <tab>Sequence description):

OmicsBox Annotation File (.annot):

```
Seq1 GO:0001234 glycolipid transfer protein-like
Seq1 GO:0001264,GO:0004567,...
Seq1 GO:0034567
```

```
Seq1 EC:2.1.2.10
Seq2 GO:0001234,... sorbitol transporter
Seq2 GO:0001244
Seq3 GO:0001234,GO:0004567,GO:0009123
Seq3 EC:1.2.4.1, EC:3.1
....
```

There are still other annotation functions available in the submenu:

7.5.9 Other Annotation Functions

- Remove Annotation. Delete Annotation results for the selected sequences.
- Filter Annotation by GO Taxa
- Validate Annotations. OmicsBox annotation generates lowest node annotations. This is not always guaranteed when Annotations have been imported or changed manually. This function can be run to ensure that no parent-child redundancy is present in the annotated set.
- Remove 1. Level Annotations
- Annotate GOs from Blast Descriptions allows to transfer GOs from the Blast hit descriptions to their sequences.

7.6 EggNOG Annotation

7.6.1 EggNOG-Mapper

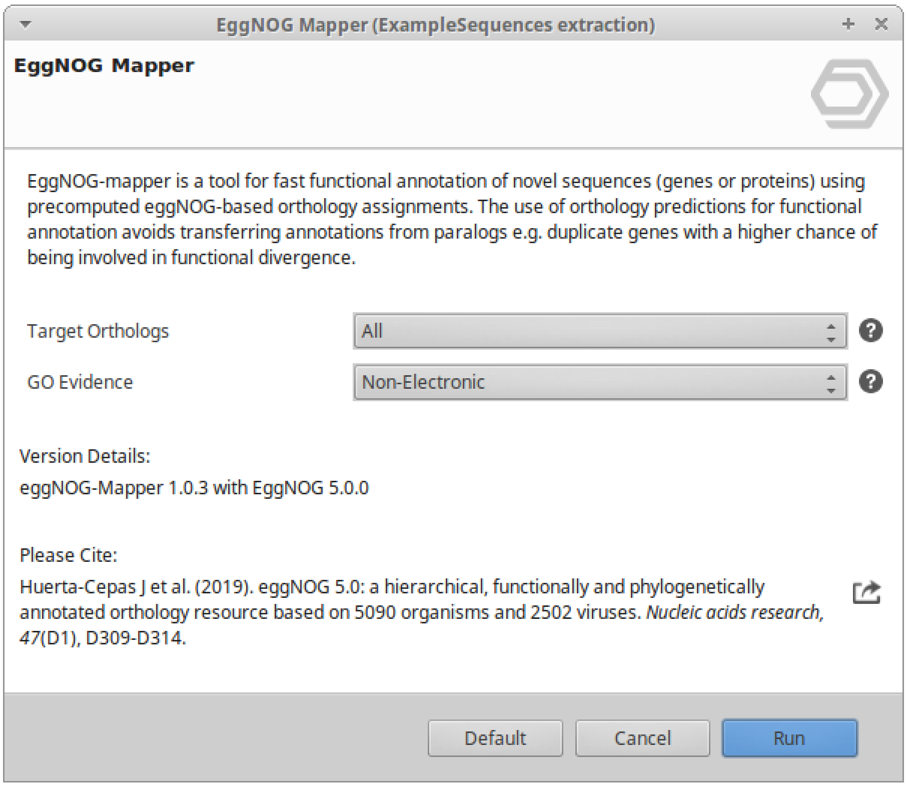
EggNOG-mapper is a tool for fast functional annotation of novel sequences (genes or proteins) using precomputed eggNOG-based orthology assignments. Obvious examples include the annotation of novel genomes, transcriptomes or even metagenomic gene catalogs. The use of orthology predictions for functional annotation is considered more precise than traditional homology searches, as it avoids transferring annotations from paralogs (duplicate genes with a higher chance of being involved in functional divergence).

Details and methodology about the tool and its database are best explained on their website: http://eggnogdb.embl.de/#/app/methods.

EggNOG-mapper can be found under **Functional Analysis** → **EggNOG Annotation** → **EggNOG Mapper**. The wizard allows to select the parameters for the functional annotation (figure 1(see page 68)).

7.6.1.1 Wizard Page

- **Target Orthologs**: Define what type of orthologs should be used for functional transfer.
- **GO Evidence**: Define what type of GO terms should be used for annotation:
    - experimental = Use only terms inferred from experimental evidence.
    - non-electronic = Use only non-electronically curated terms.



EggNOG Mapper

EggNOG-mapper is a tool for fast functional annotation of novel sequences (genes or proteins) using precomputed eggNOG-based orthology assignments. The use of orthology predictions for functional annotation avoids transferring annotations from paralogs e.g. duplicate genes with a higher chance of being involved in functional divergence.

Target Orthologs: All

GO Evidence: Non-Electronic

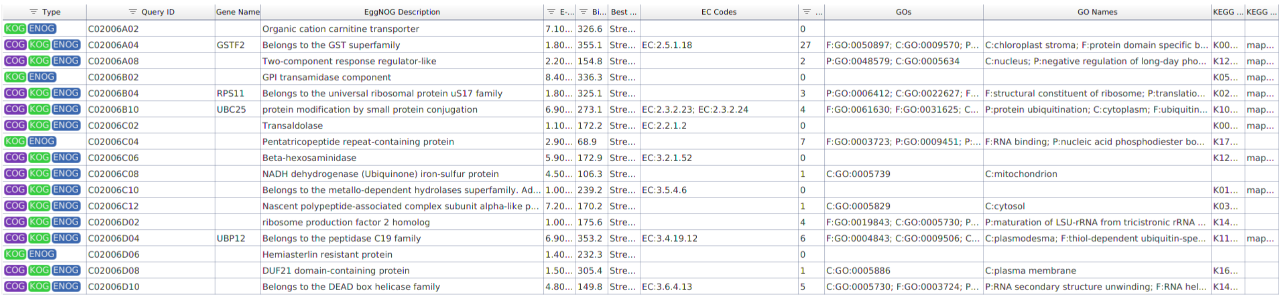
Version Details:
eggNOG-Mapper 1.0.3 with EggNOG 5.0.0

Please Cite:
Huerta-Cepas J et al. (2019). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic acids research*, *47*(D1), D309-D314.

Figure 1. EggNOG Mapper wizard page.

7.6.1.2 Results

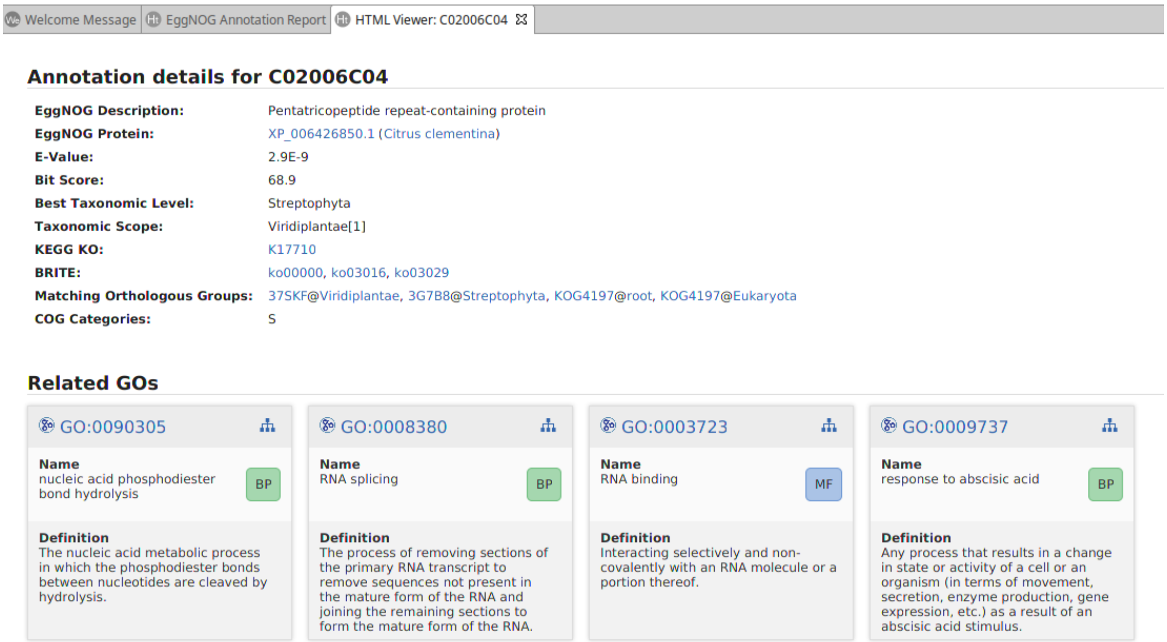
The **result table** summarizes all annotations that could be transferred with EggNOG Mapper. Besides ordering and filtering, the context menu allows to take a closer look at certain results (figure 2(see page 69)). This annotation process also generates a **Summary Report** with information about the total number of GOs, the COG categories and the orthologous groups distribution.



Type	Query ID	Gene Name	EggNOG Description	E-...	Bi...	Best ...	EC Codes	...	GOs	GO Names	KEGG ...	KEGG ...
KOG ENOG	C02006A02		Organic cation carnitine transporter	7.10...	326.6	Stre...		0				
COG KOG ENOG	C02006A04	GSTF2	Belongs to the GST superfamily	1.80...	355.1	Stre...	EC:2.5.1.18	27	F:GO:0050897; C:GO:0009570; P...	C:chloroplast stroma; F:protein domain specific b...	K00...	map...
COG KOG ENOG	C02006A08		Two-component response regulator-like	2.20...	154.8	Stre...		2	P:GO:0048579; C:GO:0005634	C:nucleus; P:negative regulation of long-day pho...	K12...	map...
KOG ENOG	C02006B02		GPI transamidase component	8.40...	336.3	Stre...		0			K05...	map...
COG KOG ENOG	C02006B04	RPS11	Belongs to the universal ribosomal protein uS17 family	1.80...	325.1	Stre...		3	P:GO:0006412; C:GO:0022627; F...	F:structural constituent of ribosome; P:translatio...	K02...	map...
COG KOG ENOG	C02006B10	UBC25	protein modification by small protein conjugation	6.90...	273.1	Stre...	EC:2.3.2.23; EC:2.3.2.24	4	F:GO:0061630; F:GO:0031625; C...	P:protein ubiquitination; C:cytoplasm; F:ubiquitin...	K10...	map...
COG KOG ENOG	C02006C02		Transaldolase	1.10...	172.2	Stre...	EC:2.2.1.2	0			K00...	map...
KOG ENOG	C02006C04		Pentatricopeptide repeat-containing protein	2.90...	68.9	Stre...		7	F:GO:0003723; P:GO:0009451; P:...	F:RNA binding; P:nucleic acid phosphodiester bo...	K17...	
COG KOG ENOG	C02006C06		Beta-hexosaminidase	5.90...	172.9	Stre...	EC:3.2.1.52	0			K12...	map...
COG KOG ENOG	C02006C08		NADH dehydrogenase (Ubiquinone) iron-sulfur protein	4.50...	106.3	Stre...		1	C:GO:0005739	C:mitochondrion		
COG KOG ENOG	C02006C10		Belongs to the metallo-dependent hydrolases superfamily. Ad...	1.00...	239.2	Stre...	EC:3.5.4.6	0			K01...	map...
COG KOG ENOG	C02006C12		Nascent polypeptide-associated complex subunit alpha-like p...	7.20...	170.2	Stre...		1	C:GO:0005829	C:cytosol	K03...	
COG KOG ENOG	C02006D02		ribosome production factor 2 homolog	1.00...	175.6	Stre...		4	F:GO:0019843; C:GO:0005730; P...	P:maturation of LSU-rRNA from tricistronic rRNA ...	K14...	
COG KOG ENOG	C02006D04	UBP12	Belongs to the peptidase C19 family	6.90...	353.2	Stre...	EC:3.4.19.12	6	F:GO:0004843; C:GO:0009506; C...	C:plasmodesma; F:thiol-dependent ubiquitin-spe...	K11...	map...
KOG ENOG	C02006D06		Hemiasterlin resistant protein	1.40...	232.3	Stre...		0				
COG KOG ENOG	C02006D08		DUF21 domain-containing protein	1.50...	305.4	Stre...		1	C:GO:0005886	C:plasma membrane	K16...	
COG KOG ENOG	C02006D10		Belongs to the DEAD box helicase family	4.80...	149.8	Stre...	EC:3.6.4.13	5	C:GO:0005730; F:GO:0003724; P...	P:RNA secondary structure unwinding; F:RNA hel...	K14...	

Figure 2. EggNOG results table.

The **annotation details** (right-click on an annotated sequence → **Show Annotation Details**) provide link outs where possible and give detailed information about annotated GOs (figure 3(see page 69)).



Annotation details for C02006C04

EggNOG Description:	Pentatricopeptide repeat-containing protein
EggNOG Protein:	XP_006426850.1 (Citrus clementina)
E-Value:	2.9E-9
Bit Score:	68.9
Best Taxonomic Level:	Streptophyta
Taxonomic Scope:	Viridiplantae[1]
KEGG KO:	K17710
BRITE:	ko00000, ko03016, ko03029
Matching Orthologous Groups:	37SKF@Viridiplantae, 3G7B8@Streptophyta, KOG4197@root, KOG4197@Eukaryota
COG Categories:	S

Related GOs

GO:0090305
Name
nucleic acid phosphodiester bond hydrolysis
BP
Definition
The nucleic acid metabolic process in which the phosphodiester bonds between nucleotides are cleaved by hydrolysis.

GO:0008380
Name
RNA splicing
BP
Definition
The process of removing sections of the primary RNA transcript to remove sequences not present in the mature form of the RNA and joining the remaining sections to form the mature form of the RNA.

GO:0003723
Name
RNA binding
MF
Definition
Interacting selectively and non-covalently with an RNA molecule or a portion thereof.

GO:0009737
Name
response to abscisic acid
BP
Definition
Any process that results in a change in state or activity of a cell or an organism (in terms of movement, secretion, enzyme production, gene expression, etc.) as a result of an abscisic acid stimulus.

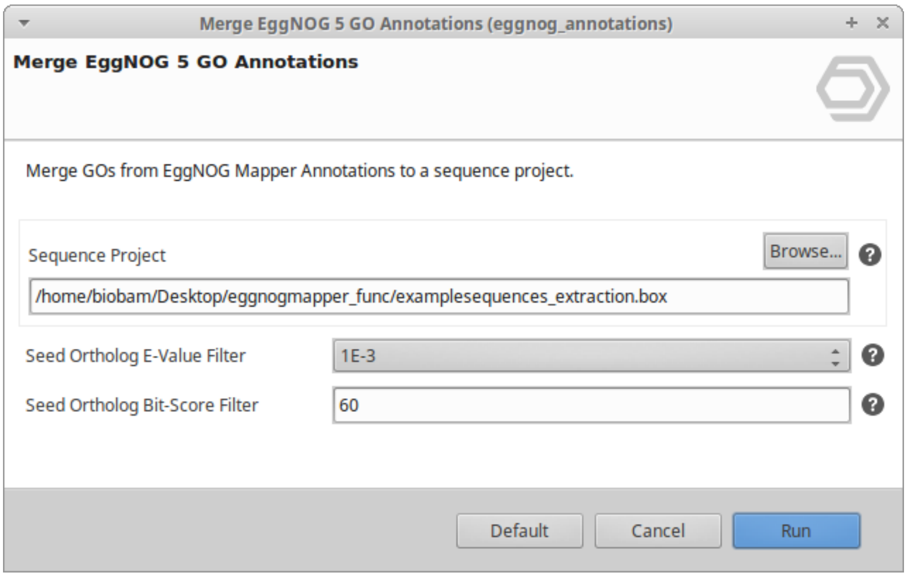
Figure 3. Annotation details.

7.6.1.3 Merge EggNOG Annotations

Once the sequences are annotated via EggNOG, it is possible to merge the GO terms and the EC codes (Enzyme Commission Codes) to a sequence project in order to add the new annotations. This can be done by clicking on **Functional Analysis** → **EggNOG Annotation** → **Merge EggNOG GO Annotations** (figure(see page 70) 4(see page 70)).

In the wizard, you have to select the sequences project to merge the GO annotations to. If the sequences already have annotated GO terms and/or ECs, the new information generated from EggNOG will be added to the annotations found in the project.

In addition, you can filter the annotations by E-value or Bit-Score.



Merge EggNOG 5 GO Annotations (eggnog_annotations)

Merge EggNOG 5 GO Annotations

Merge GOs from EggNOG Mapper Annotations to a sequence project.

Sequence Project: /home/biobam/Desktop/eggnogmapper_func/examplesequences_extraction.box

Seed Ortholog E-Value Filter: 1E-3

Seed Ortholog Bit-Score Filter: 60

Figure 4. Merge EggNOG GO Annotations wizard.

Once finished, this step generates a barchart showing the total number of GOs and ECs added to the original sequence project (figure 5(see page 71)).

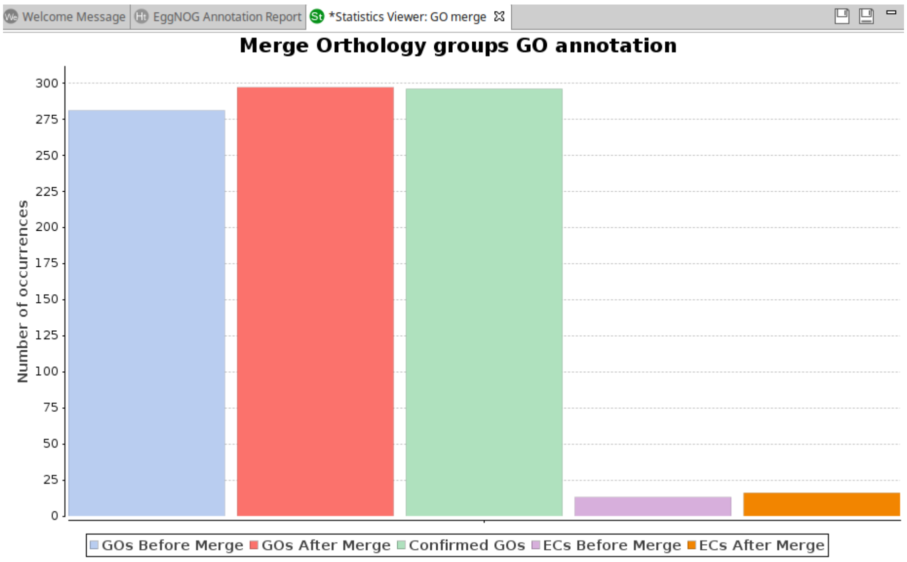


Figure 5. Merge EggNOG GO Annotations graph.

Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. Jaime Huerta-Cepas, Damian Szklarczyk, Lars Juhl Jensen, Christian von Mering and Peer Bork. Submitted (**2016**).
eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Jaime Huerta-Cepas, Damian Szklarczyk, Kristoffer Forslund, Helen Cook, Davide Heller, Mathias C. Walter, Thomas Rattei, Daniel R. Mende, Shinichi Sunagawa, Michael Kuhn, Lars Juhl Jensen, Christian von Mering, and Peer Bork. Nucl. Acids Res. (04 January **2016**) 44 (D1): D286-D293. doi: 10.1093/nar/gkv1248

7.7 InterProScan Annotation

Content of this page:

7.7.1 General

The functionality of InterPro annotations in OmicsBox allows to retrieved domain/motif information in a sequence-wise manner. Corresponding GO terms are then transferred to the sequences and merged with already existing GO terms. InterProScan results can be viewed through the Single Sequence Menu (figure 6[36]) and saved in TXT and XML format (figure 5[37]). The sequences will turn **violet** if no other analysis has been executed before.

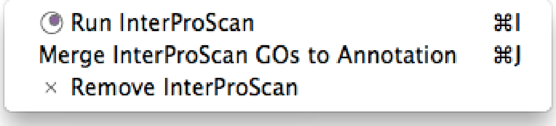


Figure 1: InterProScan options

- Run InteProScan. Start sending sequences to the EBI.
- Merge InterProScan GOs to Annotation. Add GO terms obtained through motifs/domains to the current annotations.
- Remove InterProScan. Delete InterProScan results for the selected sequences.

[36] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598048967#InterProScanAnnotation-figure6
[37] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598048967#InterProScanAnnotation-figure5

7.7.2 Run InterProScan

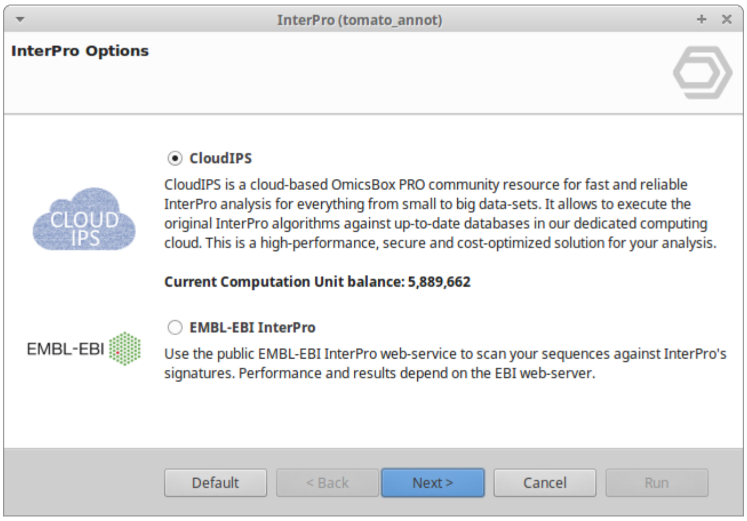
There are two options to run InterProScan in OmicsBox, either with CloudIPS or via the public web service at EBI.

CloudIPS is a cloud-based OmicsBox community resource for fast and reliable InterPro analysis for everything from small to big data-sets. It allows executing the original InterPro algorithms against up-to-date databases in our dedicated computing cloud. This is a high-performance, secure and cost-optimized solution for your analysis.
The public EMBL-EBI InterPro web-service scans your sequences against InterPro's signatures and performance and results depend on the EBI web-server.

InteProScan can only be performed if the sequences are shown in the sequence table that contains the actual sequence information (loaded via fasta file). You have to be careful if you created a project via a blast XML file or if you loaded a .annot file.
To add the sequences to the current OmicsBox project see Add sequences to existing OmicsBox project(see page 38) section.

You can save the InterProScan results in different file formats, in tab separated values (TVS), XML, which is the default output, GFF3 and the input (query) sequence itself (figure 5[38]).
If you are working with nucleotide sequences, OmicsBox translates it to the longest open reading frame and sends it to InterProScan. For this particular case when exporting the input sequence OmicsBox will save the protein sequence itself and not the nucleotide one.

Once the InterProScan has finished it is possible to view the results of each sequence via the context menu (figure 6[39]).



InterPro Options

● CloudIPS

CloudIPS is a cloud-based OmicsBox PRO community resource for fast and reliable InterPro analysis for everything from small to big data-sets. It allows to execute the original InterPro algorithms against up-to-date databases in our dedicated computing cloud. This is a high-performance, secure and cost-optimized solution for your analysis.

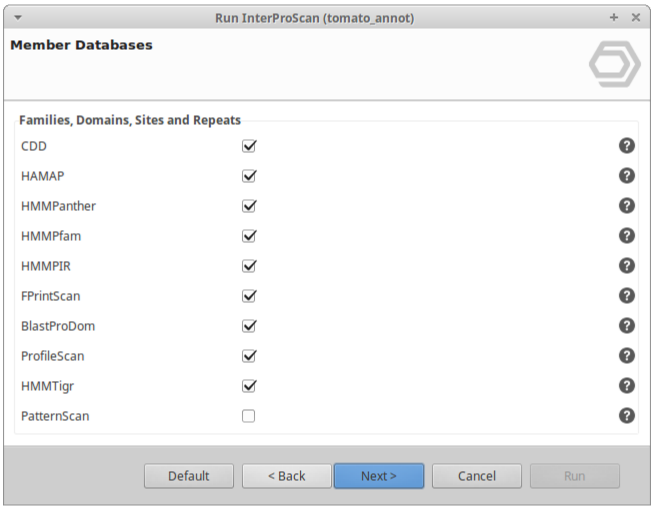
Current Computation Unit balance: 5,889,662

○ EMBL-EBI InterPro

Use the public EMBL-EBI InterPro web-service to scan your sequences against InterPro's signatures. Performance and results depend on the EBI web-server.

Figure 2: InterProScan Configuration

[38] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598048967#InterProScanAnnotation-figure5
[39] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598048967#InterProScanAnnotation-figure6

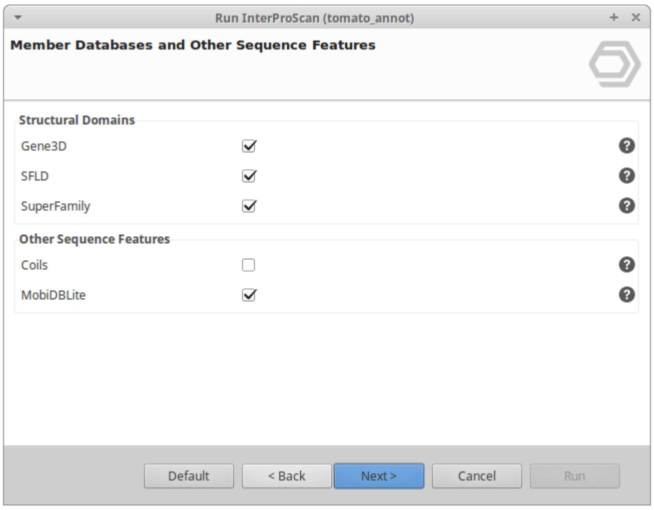


Member Databases

Families, Domains, Sites and Repeats

- CDD ☑
- HAMAP ☑
- HMMPanther ☑
- HMMPfam ☑
- HMMPIR ☑
- FPrintScan ☑
- BlastProDom ☑
- ProfileScan ☑
- HMMTigr ☑
- PatternScan ☐

Figure 3: Selection of Member Databases



Member Databases and Other Sequence Features

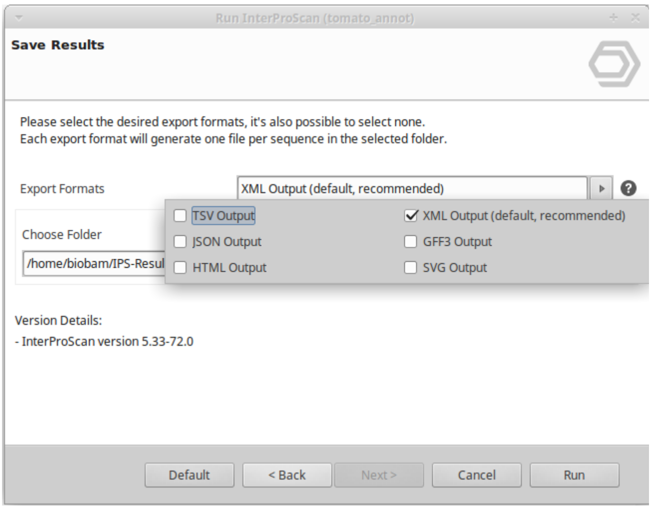
Structural Domains

- Gene3D ☑
- SFLD ☑
- SuperFamily ☑

Other Sequence Features

- Coils ☐
- MobiDBLite ☑

Figure 4: Selection of Member Databases



Run InterProScan (tomato_annot)

Save Results

Please select the desired export formats, it's also possible to select none.
Each export format will generate one file per sequence in the selected folder.

Export Formats: XML Output (default, recommended)

□ TSV Output ☑ XML Output (default, recommended)
□ JSON Output □ GFF3 Output
□ HTML Output □ SVG Output

Choose Folder: /home/biobam/IPS-Resul

Version Details:
- InterProScan version 5.33-72.0

Figure 5: Save InterProScan Results

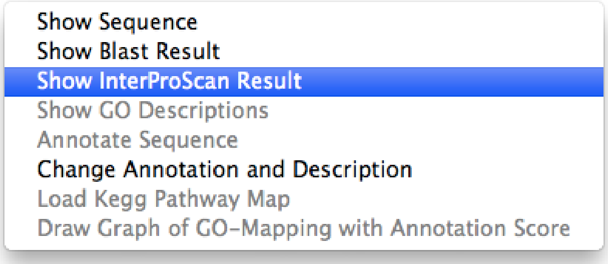
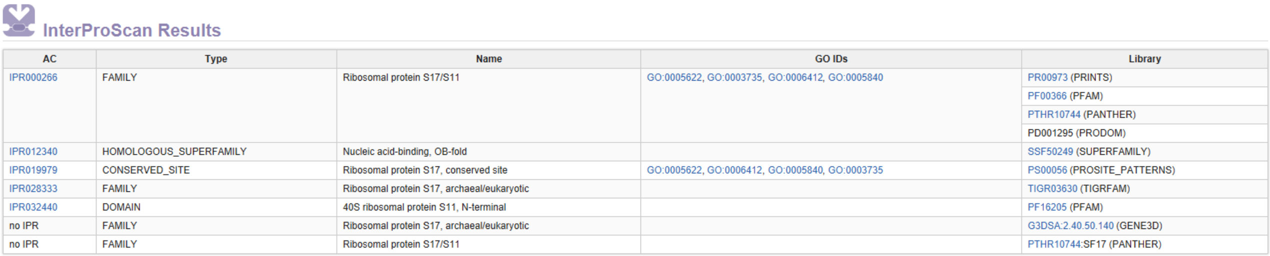


Figure 6: Show InterProScan Results



InterProScan Results

AC	Type	Name	GO IDs	Library
IPR000266	FAMILY	Ribosomal protein S17/S11	GO:0005622, GO:0003735, GO:0006412, GO:0005840	PR00973 (PRINTS) PF00366 (PFAM) PTHR10744 (PANTHER) PD001295 (PRODOM)
IPR012340	HOMOLOGOUS_SUPERFAMILY	Nucleic acid-binding, OB-fold		SSF50249 (SUPERFAMILY)
IPR019979	CONSERVED_SITE	Ribosomal protein S17, conserved site	GO:0005622, GO:0006412, GO:0005840, GO:0003735	PS00056 (PROSITE_PATTERNS)
IPR028333	FAMILY	Ribosomal protein S17, archaeal/eukaryotic		TIGR03630 (TIGRFAM)
IPR032440	DOMAIN	40S ribosomal protein S11, N-terminal		PF16205 (PFAM)
no IPR	FAMILY	Ribosomal protein S17, archaeal/eukaryotic		G3DSA:2.40.50.140 (GENE3D)
no IPR	FAMILY	Ribosomal protein S17/S11		PTHR10744:SF17 (PANTHER)

Figure 7: InterProScan Results

7.7.3 Merge InterProScan GOs to Annotation

The InterProScan GOs results can now be added to the already existing annotations based on the BLAST results. This option is available from the InterProScan submenu.
Once the merge has finished a distribution chart is displayed in the Results menu showing the number of GOs that have been added to (or confirmed) the current annotation results.

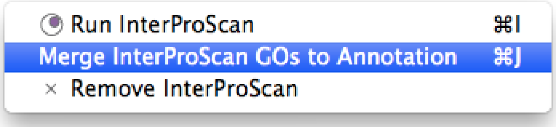


Figure 8: Merge InterProScan results

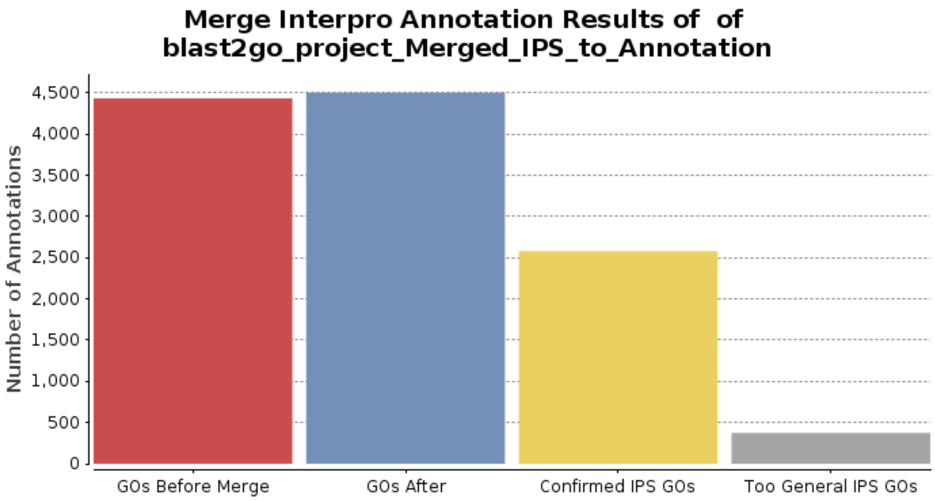


Figure 9: Statistics after merging InterProScan to GO Annotation

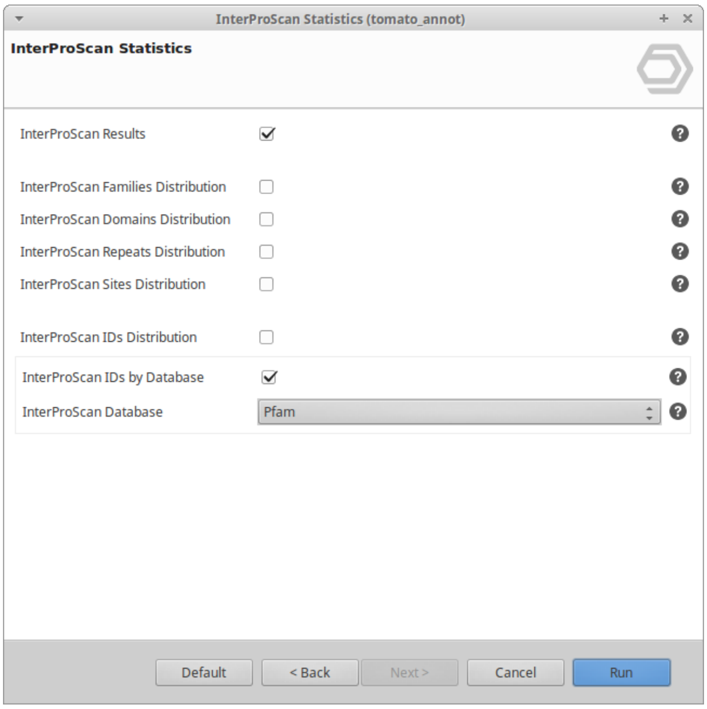
7.7.4 Statistics

On the submenu of the "Charts" icon it is possible to select InterProScan statistics to see how many sequences still do or do not have IPS results and how many sequences have GOs resulting from InterProScan.

- InterProScan Results: This chart reflects the effect of adding the GO-terms retrieved through the InterProScan results (figure 11[40]).
- InterProScan Families Distribution: Bar chart representing the number of sequences that belong to a particular IPS family.
- InterProScan Domains Distribution: Bar chart showing the number of sequences that belong to a particular IPS domain.
- InterProScan Repeats Distribution: Bar chart reflecting the number of sequences that belong to a particular IPS repeat.
- InterProScan Sites Distribution: Bar chart representing the number of sequences that belong to a particular IPS sites.
- InterProScan IDs Distribution: Bar chart showing the number of sequences that have been annotated with that InterProScan IDs.

[40] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598048967#InterProScanAnnotation-figure11

- InterProScan IDs by Database: Pie chart reflecting the number of sequences of the InterProScan IDs for a particular InterProScan Database. In figure 10[41] the Pfam database is selected.



InterProScan Statistics (tomato_annot)

InterProScan Statistics

InterProScan Results ☑

InterProScan Families Distribution □

InterProScan Domains Distribution □

InterProScan Repeats Distribution □

InterProScan Sites Distribution □

InterProScan IDs Distribution □

InterProScan IDs by Database ☑

InterProScan Database: Pfam

Figure 10: InterProScan Statistics Configuration Window

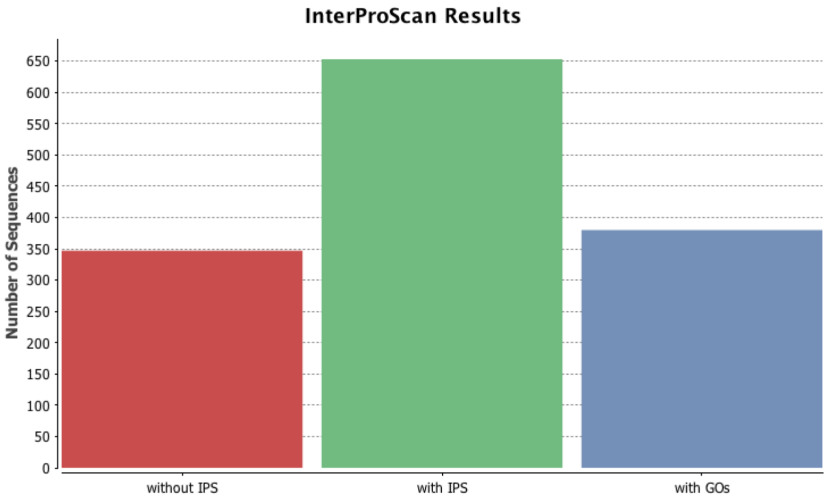


Figure 11: InterProScan Statistics

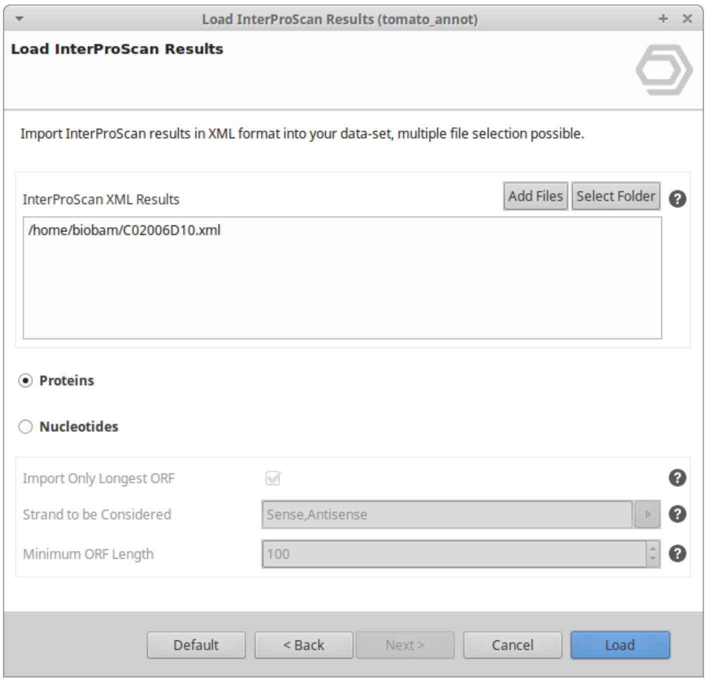
[41] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598048967#InterProScanAnnotation-figure10

7.7.5 Load InterProScan Results

The InterProScan results saved in XML format can be loaded in the current OmicsBox project (**File > Load > Load InterProScan Results**).

When loading the InterProScan results it is possible to select the input format.

- Protein - If InterProScan has been performed inside OmicsBox (OmicsBox translates the nucleotide sequences to the longest ORF peptides)
- Nucleotides - If InterProScan has been performed with nucleotide sequences and InterProScan binaries.



Load InterProScan Results (tomato_annot)

Load InterProScan Results

Import InterProScan results in XML format into your data-set, multiple file selection possible.

InterProScan XML Results

/home/biobam/C02006D10.xml

● **Proteins**

○ **Nucleotides**

Import Only Longest ORF ☑

Strand to be Considered: Sense,Antisense

Minimum ORF Length: 100

Figure 12: Load InterProScan Results

7.8 Gene Ontology Graphs

Content of this page:

OmicsBox aims to be a visual-oriented tool. This means that special attention is paid to show information through graphs, coloring and charts.

7.8.1 Directed Acyclic Graphs

OmicsBox offers the possibility of visualizing the hierarchical structure of the gene ontology by directed acyclic graphs (DAG). This functionality is available to visualize results at different stages of the application and although configuration dialogs may vary, there are some shared features when generating graphs.

1. Software. OmicsBox integrates a viewer based on the ZVTM framework developed by Emmanuel Pietriga at the INRA (France) for graph visualization (, [42]). This high-performing vectored visualization framework allows fast navigation and zooms on the GO DAG. A graph overview is permanently shown at the upper right corner of the graphical tab to easy follow exploring across the DAG surface. Zoom in/out is supported on the mouse wheel and fast zoom to readability is reached by double click on a DAG node. Information about the current node is given on the lower application bar
2. Parameters.
   a. Node Filters. A potential drawback during drawing Gene Ontology DAGs where numerous sequences are involved is the presence of an excessive number of nodes that would make the graph hard to visualize and will demand large memory resources. OmicsBox allows modulation of graph size by introducing node filters that depend of the type of graph considered. Additionally, there are a maximum possible number of nodes to be displayed.
   b. Coloring mode. OmicsBox highlights nodes proportionally to some parameter of the analysis which result is visualized on the DAG. By this intensity variation of node color relevant terms get more visual weight which is a useful way to guide visual inspection of the results.

7.8.1.1 Graph element legend

Gene Ontology term obtained by mapping which can directly be associated to one ore more BLAST hits. (GO-Accession, maximum hit e-value assigned, max. hit similarity assigned, number of hits belonging to this)

Non-annotated GO term node (GO term name, mean e-value of all hits contributing to this node, max. e-value, max. Similarity, number of Hits contributing to this node, Annotation Algorithm Score)

[42] http://download.blast2go.com/b2gusermanual/test/node10.html#Pietriga2005Toolkit

Annotated GO term node (GO term name, mean e-value of all hits contributing to this node, max. e-value, max. Similarity, number of Hits contributing to this node, Annotation Algorithm Score)

There exist two types of relationships between child and parent terms. Children that represent a more specific instance of a parent term have an 'instance of' or 'is a' relationship to the parent. Children that are a constituent of the parent term have a 'part of' relationship.

7.8.2 Pies and Bar Charts

Some of the results obtained by the Data Mining tools present in the application (see Quantitative Analysis section) are displayed either as a Bar or Pie charts. Similarly to the DAGs, parameters for modulating the size of these graphs are available at their configuration menus. As these charts are very much related to the Data Mining functions they correspond, they will be explained together in the next section.

7.8.3 Quantitative Analysis

Content of this page:

As a Data Mining tool, OmicsBox provides various ways for the joint analysis of groups of annotated sequences.

7.8.3.1 Descriptive analysis. Combined Graph Function

OmicsBox generates combined graphs where the combined annotation of a group of sequences is visualized together. This can be used to study the joined biological meaning of a set of sequences. Combined graphs are a good alternative to enrichment analysis where there is no reference set to be considered or the number of involved sequences is low. This function is available under **Functional Analysis > Gene Ontology Graphs**.

The next images show the Combined Graph Drawing Configuration Dialog, where the following parameters are available:

- Graph Title
- GO Categories

For each Gene Ontology category, a graph will be displayed. OmicsBox allows extracting information from the graph nodes such as tooltip (figure 4(see page 80)), create a subgraph from that specific GO, create an Id list of the sequences that have been annotated with that particular GO (figure 5(see page 80)). The generated

Id list can then be used within OmicsBox in the select by sequences feature (see Select Sequences and Functions(see page 107) Section).

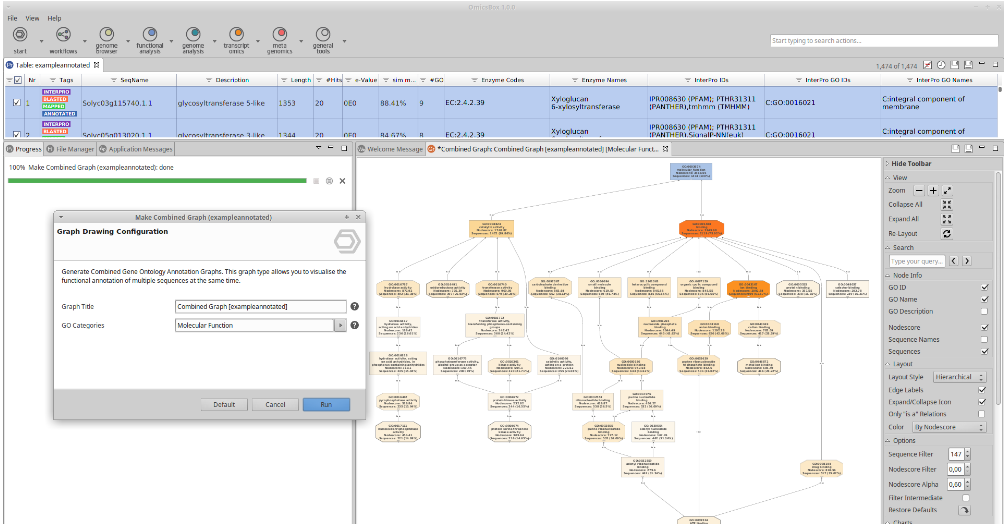
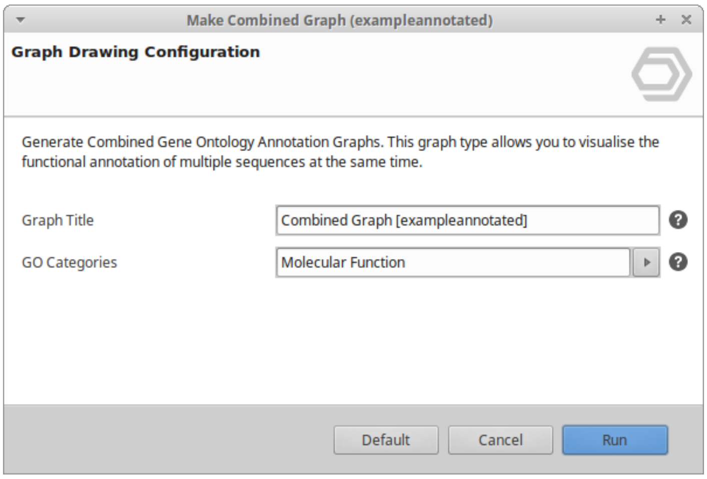


Figure 1: Combined graph visualization


Make Combined Graph (exampleannotated)

Graph Drawing Configuration

Generate Combined Gene Ontology Annotation Graphs. This graph type allows you to visualise the functional annotation of multiple sequences at the same time.

Graph Title: Combined Graph [exampleannotated]

GO Categories: Molecular Function

Figure 2: Combined Graph Drawing Configuration Dialog allows to provide a graph title header and to choose between the different GO categories

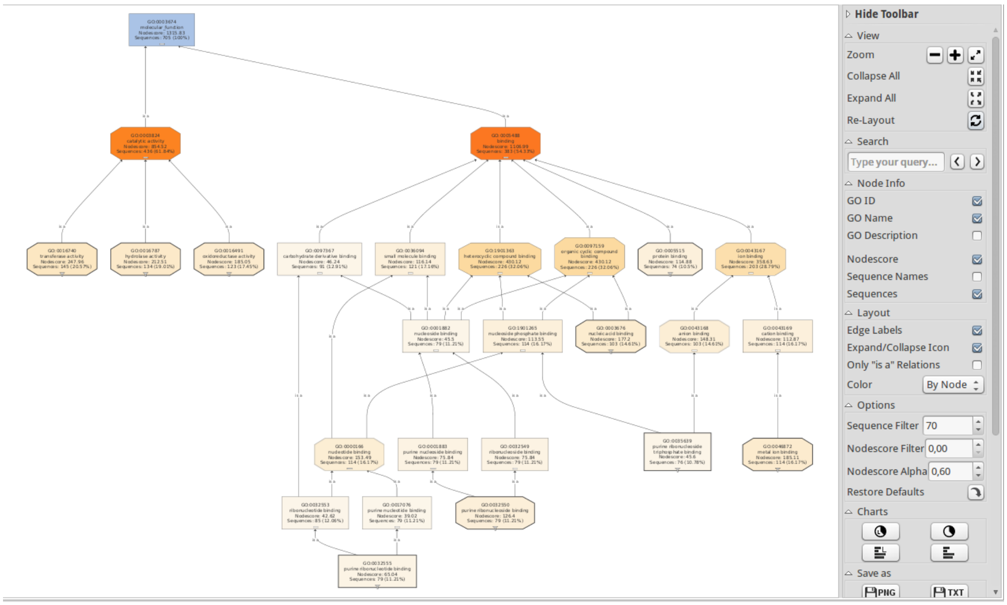
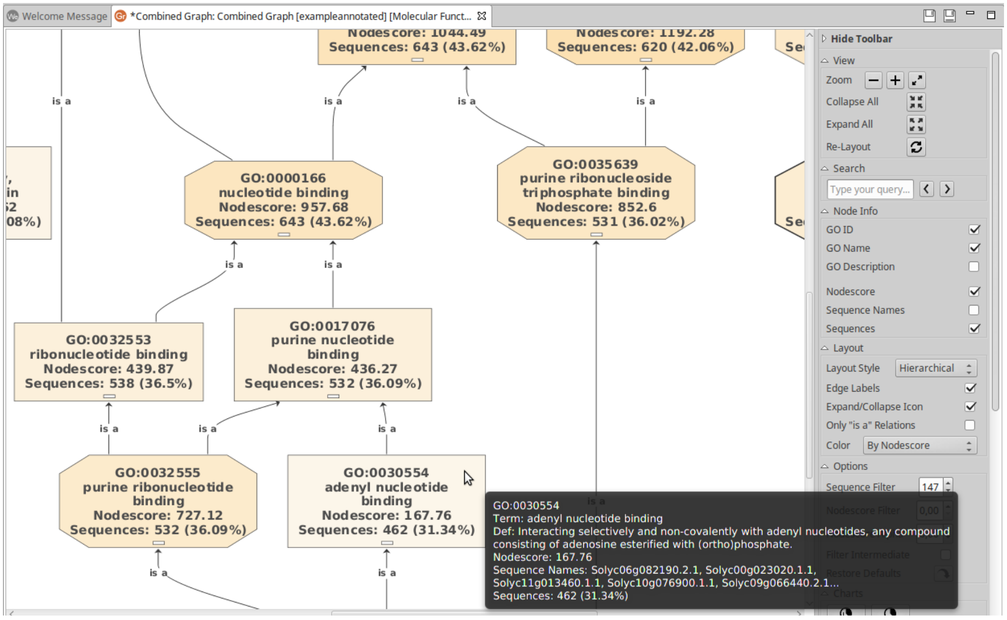


Figure 3: Molecular Function Combined Graph



GO:0030554
Term: adenyl nucleotide binding
Def: Interacting selectively and non-covalently with adenyl nucleotides, any compound consisting of adenosine esterified with (ortho)phosphate.
Nodescore: 167.76
Sequence Names: Solyc06g082190.2.1, Solyc00g023020.1.1, Solyc11g013460.1.1, Solyc10g076900.1.1, Solyc09g066440.2.1...
Sequences: 462 (31.34%)

Figure 4: Graph Node Tooltip

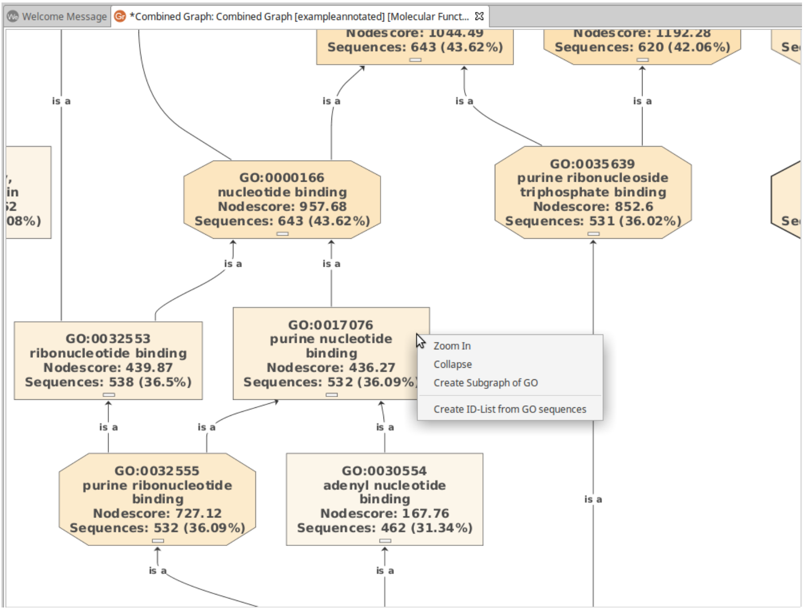


Figure 5: Extract Node Information

Graph Side Panel

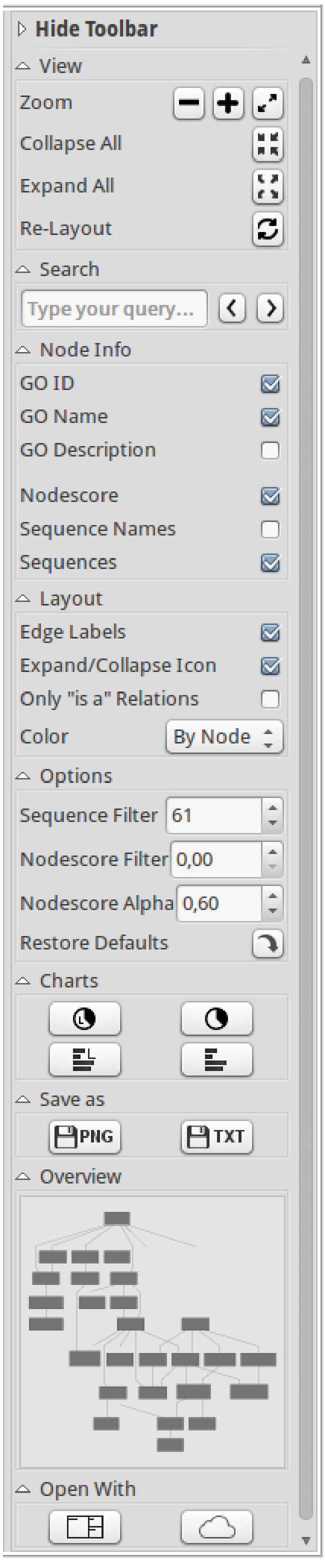
The generated combined graph is interactive and its parameters can be modified from the side panel.

- View. This section controls the graph visualization within its area.
  - Zoom
  - Collapse All: The nodes will collapse and only the root will be visualized.
  - Expand All: The nodes will expand to the original graph visualization.
  - Re-Layout: The whole graph will be re-scaled to adjust to the visualization area.
- Search. Allows to search for GO IDs/ Terms/ Description in the Combined Graph.
- Node Info. This parameter controls the information shown at a node. Possible values are:
  - GO ID: If checked the GO ID will be included in the node.
  - GO Name: The GO Names are shown in the node.
  - GO Description: When checked the GO Description will be included in the node.
  - Nodescore: The node score will be shown in the node.
  - Sequence Names: The names of the sequences annotated at each GO are included in the node. The limit number of names to be displayed is 15.
  - Sequences: The number of sequences annotated with that particular GO will be displayed in the node.
- Layout.
  - Edge Labels: When checked the labels on the edges will be shown.
  - Expand/Collapse Icon: If checked the ions that represent expand/collapse on the node are displayed.
  - Only "is a" Relations: Only the is a relation between nodes will be displayed if the box is checked.
  - Color
    - Ontology: All nodes will be colored according to the ontology category, Biological Process - green; Molecular Function - blue; Cellular Component - yellow.
    - White: The nodes will turn white.

    - By Nodescore: A Score is computed at each node according to the formula:

      $$score = \sum_{GOs} seq \times \alpha^{dist}$$

      where seq is the number of different sequences annotated at a child GO term and dist th e distance to the node of the child. GO term Coloring by Score will highlight areas of high annotation density.
    - By Sequence Count: Node color intensity will be proportional to the number of contributing sequences at the node.
- Options.
  - Sequence Filter: The minimal number of sequences a GO node must have assigned, to be displayed. This filter is used to control the number of nodes present in the graph. It is recommended to start the analysis with a high number that, depending on the number of total sequences, is expected to overload the graph. Depending on the result adjust this value until you obtain a satisfactory graph. Start with 10% of your total number of sequences.
  - Nodescore Filter:
  - Score alpha. The value for parameter alpha in the Score formula Node Score Filter. Only nodes with a Score value higher than the Filter will be shown. Use this parameter to thin out the GO-DAG for low informative nodes.
  - Restore Defaults: All filters will be set to the default values.
- Charts. (see next section)
- Save as. The information present in a Combined Graph can be saved as an image (.png) or in table format. This will generate a .txt file where all information related to each node of the plotted Graph is provided in different columns.
- Overview. Provides a radar-like view of the graph, which allows adjusting the visible window.
- Open With. Open the graph information as TreeMap or WordCloud (see following sections).



GO ID ☑
GO Name ☑
GO Description □
Nodescore ☑
Sequence Names □
Sequences ☑

Edge Labels ☑
Expand/Collapse Icon ☑
Only "is a" Relations □
Color: By Node

Sequence Filter: 61
Nodescore Filter: 0,00
Nodescore Alpha: 0,60

Figure 6: Combined Graph Side Panel

Charts

Analysis of GO Term associations in a set of sequences can also be done by Pie/Bar Charts. For this analysis, a Combined Graph must have been generated first. Once the graph is visible in the GO Graph panel you can find several icons to visualize the 4 different types of charts.

Four possibilities are available:

1. Sequence distribution by GO level (Pie-Chart): This pie chart represents the number of sequences for each Gene Ontology term for a given level. See figure 8(see page 86).
2. Sequences per GO terms (Multilevel Pie): This function generates a Pie with the lowest node per branch of the DAG that fulfils the filter condition., e.g. will find all the lowest nodes with the given number of sequences or Score value and will plot them jointly in a Pie representation. See figure 9(see page 86).
3. Top 50 GO terms (Bar-Chart): A bar chart representing the GO terms according to the number of annotated sequences. See figure 10(see page 87).
4. Sequence distribution by GO level (Bar-Chart): This bar chart represents the number of sequences for each Gene Ontology term for a given level. See figure 11(see page 87).

When any of these functions are called, a table of node counts is generated and displayed in the statistics tab.

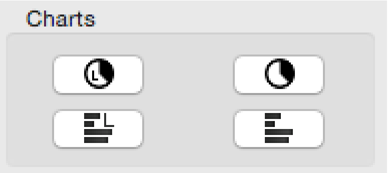


Figure 7: Combined Graph Pie and Bar-Charts

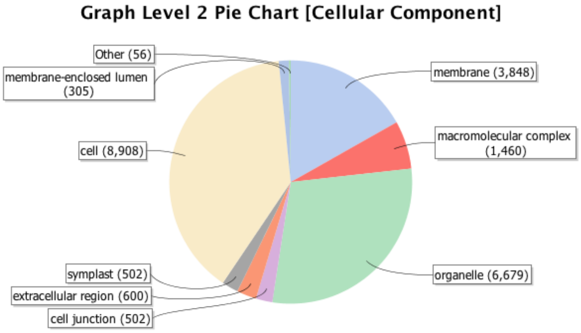


Figure 8: Sequence distribution by GO level: Pie Chart

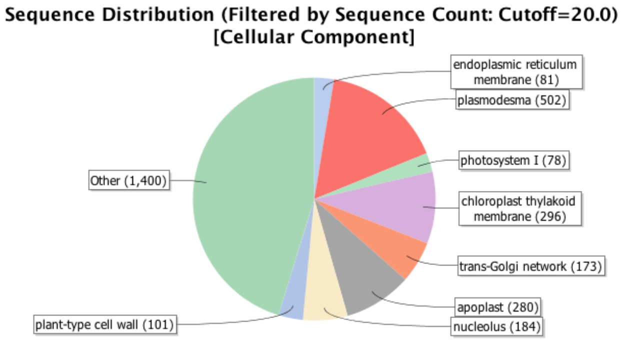


Figure 9: Sequence Distribution/GO as Multilevel-Pie (#score or #seq cutoff)

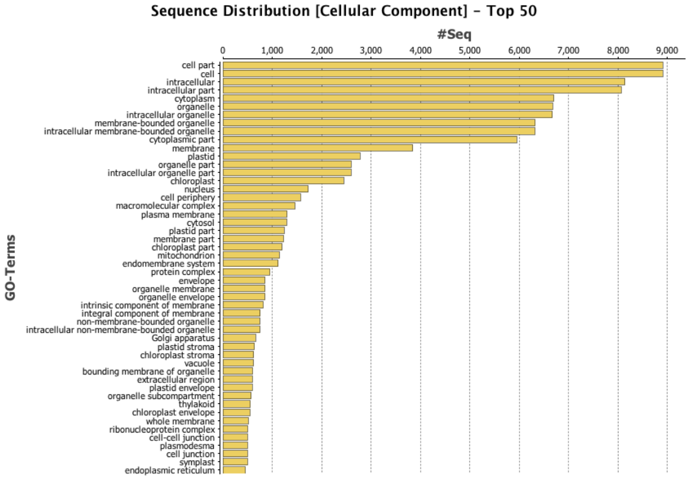


Figure 10: Top 50 GO terms

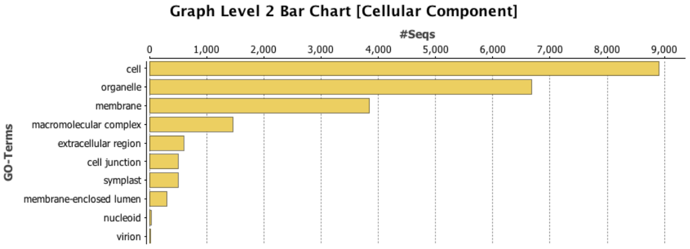
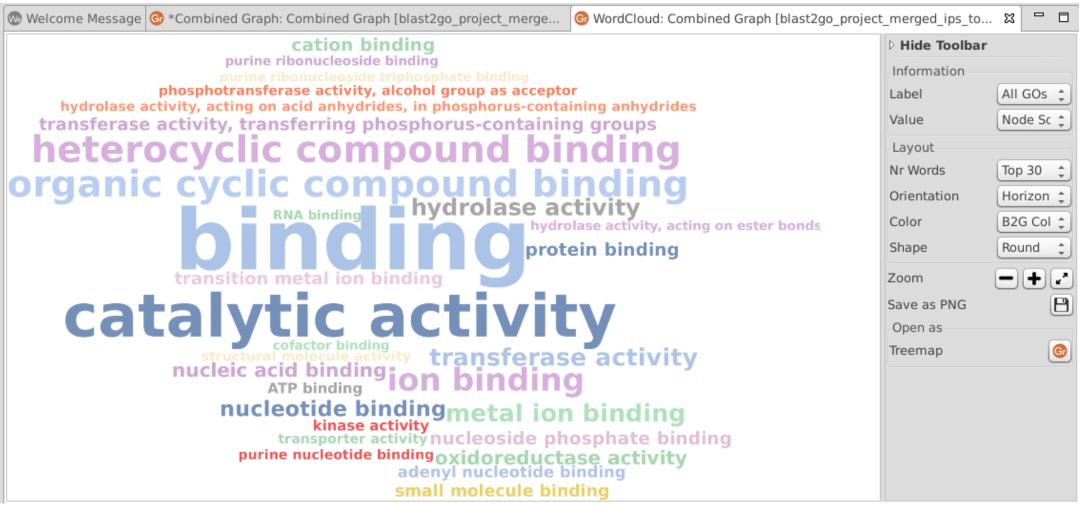


Figure 11: Sequence distribution by GO level: Bar Chart

WordCloud

A WordCloud is a visual representation for a list of labels. The importance of words, here GO terms, is represented by its font size. The font size depends on either the sequence count or the NodeScore of each GO term. The list of words can be limited to a specific Gene Ontology category (BP, CC or MF). The coloring is random. Several options to change the graphical appearance are available like the number of words, the orientation and shape of the cloud as well as the color scheme.


Welcome Message
*Combined Graph: Combined Graph [blast2go_project_merge...
WordCloud: Combined Graph [blast2go_project_merged_ips_to...
cation binding
purine ribonucleoside binding
purine ribonucleoside triphosphate binding
phosphotransferase activity, alcohol group as acceptor
hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides
transferase activity, transferring phosphorus-containing groups
heterocyclic compound binding
organic cyclic compound binding
RNA binding
hydrolase activity
binding
hydrolase activity, acting on ester bonds
protein binding
transition metal ion binding
catalytic activity
cofactor binding
structural molecule activity
transferase activity
nucleic acid binding
ion binding
ATP binding
nucleotide binding
metal ion binding
kinase activity
transporter activity
nucleoside phosphate binding
purine nucleotide binding
oxidoreductase activity
adenyl nucleotide binding
small molecule binding
Hide Toolbar
Zoom
Save as PNG
Open as
Treemap

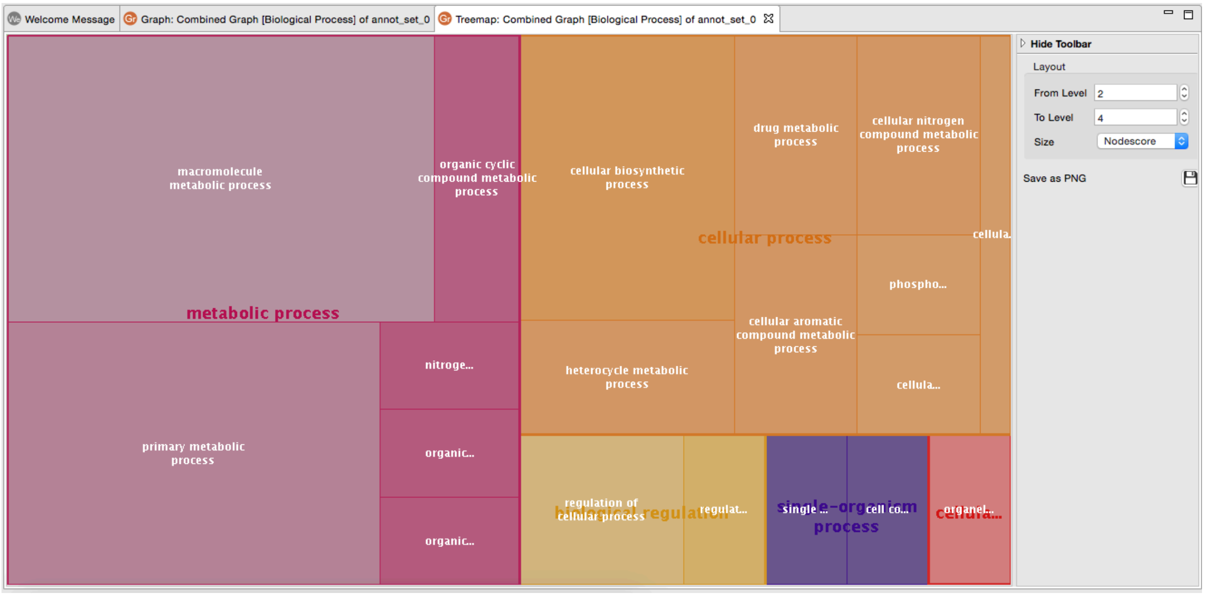
Information
Label: All GOs
Value: Node Sc

Layout
Nr Words: Top 30
Orientation: Horizon
Color: B2G Col
Shape: Round

Figure 12: Convert Graph to Word Cloud

TreeMap

The TreeMap viewer allows visualizing graphs (hierarchical, tree-structured data in general) as a set of nested rectangles. Each branch of the tree is given a rectangle, which is then tiled with smaller rectangles representing sub-branches. The size of each rectangle represents the number of sequences associated with a given GO term or a GO's NodeScore.


Welcome Message
Graph: Combined Graph [Biological Process] of annot_set_0
Treemap: Combined Graph [Biological Process] of annot_set_0
macromolecule metabolic process
organic cyclic compound metabolic process
metabolic process
primary metabolic process
nitroge...
organic...
organic...
cellular biosynthetic process
drug metabolic process
cellular nitrogen compound metabolic process
cellular process
cellula...
phospho...
cellular aromatic compound metabolic process
heterocycle metabolic process
cellula...
regulation of cellular process
biological regulation
regulat...
single ...
single-organism process
cell co...
organel...
cellula...
Hide Toolbar
Layout
Save as PNG

From Level 2
To Level 4
Size Nodescore

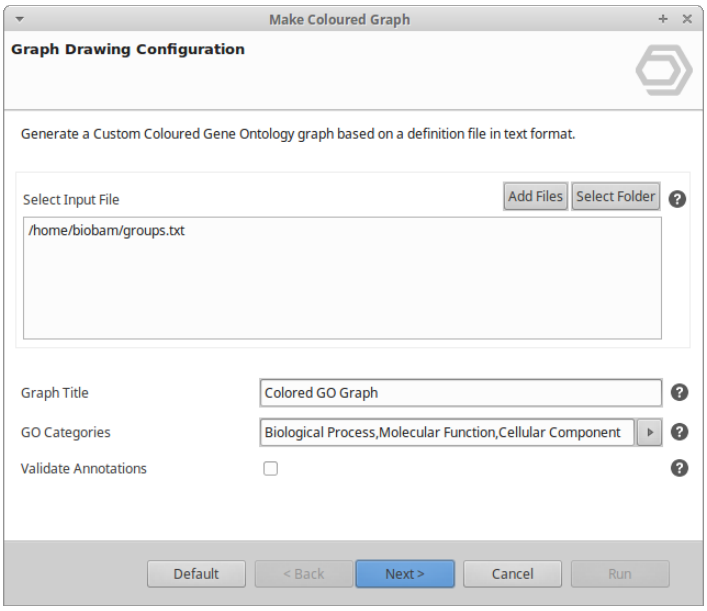
Figure 13: A TreeMap representing a Gene Ontology Graph.
The size of the rectangles represents the number of sequences or the NodeScore of each GO term.

7.8.3.2 Coloured GO Graphs from a text file

We can generate a GO graph from a text (.txt) file which contains a list of GOs and the desired colour for each of them. It is also possible to label groups of GOs with the same name. Figure 15(see page 89) shows an example that was created introducing the following text file:

```
GO:0000003    6    Group A
GO:0040007    8    Group B
GO:0050896    1    Group B
```

The text file has to follow a simple structure, to be processed correctly. It may contain from 2 to 3 columns in each line. The first column has to contain a GO, the second a number (0.0 to ∞) and the optional third column contains a text that will be written into the octagon of the corresponding GO. The columns must be separated with a tabulator character.
According to the example above Group B has two GO IDs that contain different values. It is also possible to differentiate these GO IDs by colouring according to their values. In order to colour the octagon according to the value, you should select the gradient colour in the next page on the colour graph configuration window (see figure 16(see page 89)).



Make Coloured Graph

Graph Drawing Configuration

Generate a Custom Coloured Gene Ontology graph based on a definition file in text format.

Select Input File: /home/biobam/groups.txt

Graph Title: Colored GO Graph

GO Categories: Biological Process,Molecular Function,Cellular Component

Validate Annotations: □

Figure 14: Colour Configuration Window

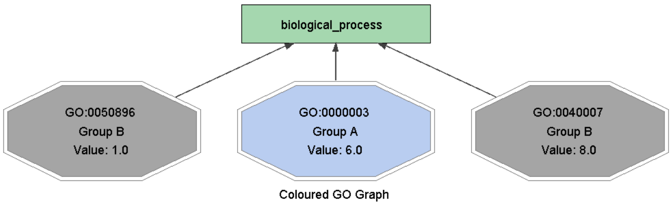


Figure 15: Coloured GO Graph by Group

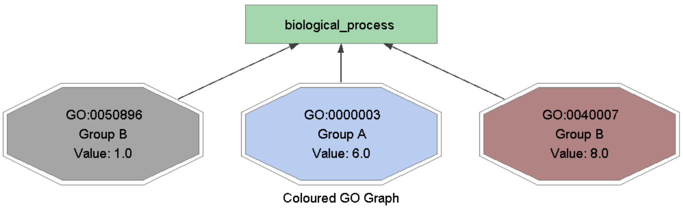
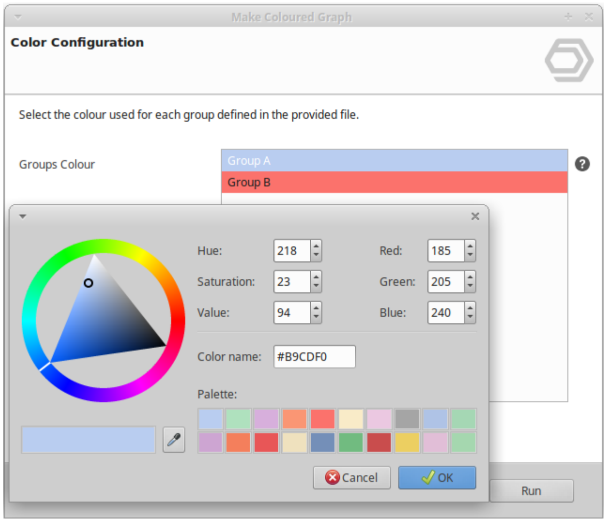


Figure 16: Coloured GO Graph by Group value



Make Coloured Graph

Color Configuration

Select the colour used for each group defined in the provided file.

Groups Colour: Group A, Group B

Hue: 218
Saturation: 23
Value: 94
Red: 185
Green: 205
Blue: 240
Color name: #B9CDF0

Figure 17: Select Colour to differentiate values within the same group.

7.8.3.3 Make GO Graph

The "Make GO Graph" function allows visualizing any set of GO terms/Ids.

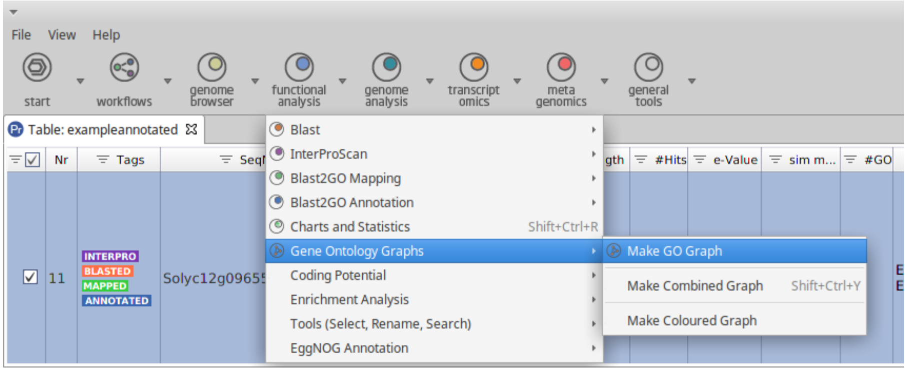
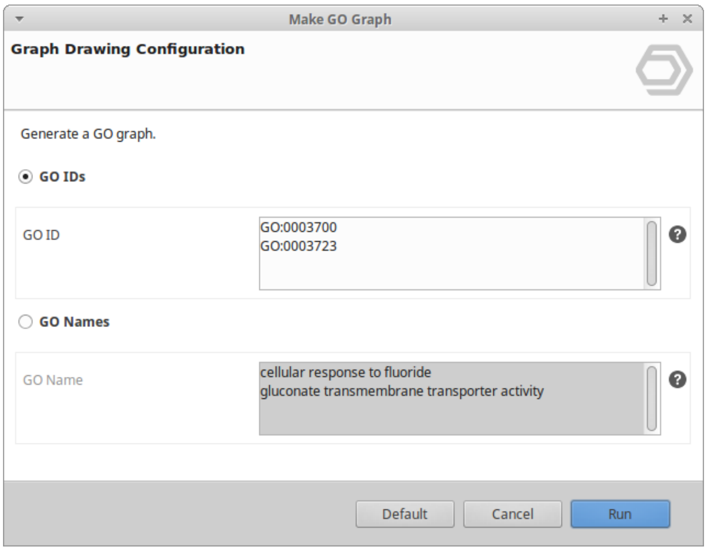


Figure 18: Make GO Graph



Make GO Graph

Graph Drawing Configuration

Generate a GO graph.

● GO IDs

GO ID: GO:0003700
GO:0003723

○ GO Names

GO Name: cellular response to fluoride
gluconate transmembrane transporter activity

Figure 19: Make GO ID Graph

7.9 RFAM

Content of this page:

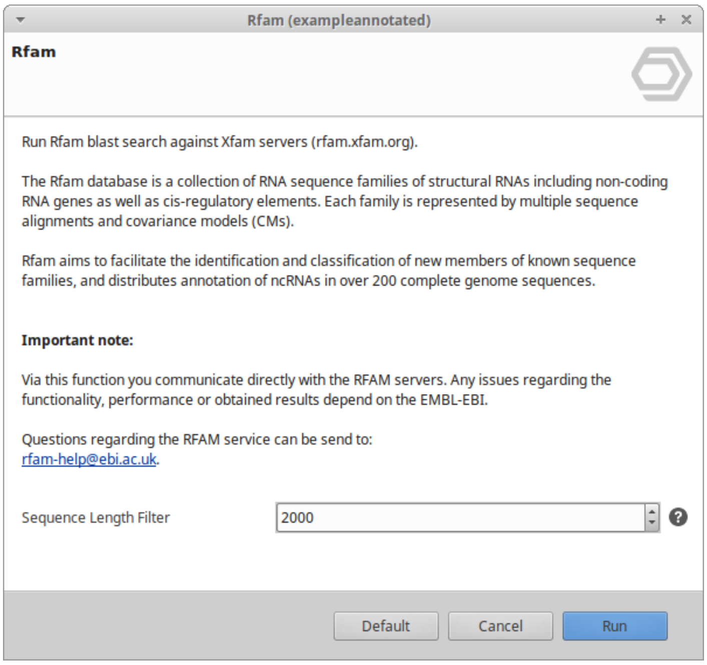
7.9.1 Introduction

The Rfam database is a collection of RNA families, each represented by multiple sequence alignments, consensus secondary structures and covariance models (CMs). The families in Rfam break down into three broad functional classes: non-coding RNA genes, structured cis-regulatory elements and self-splicing RNAs. Typically these functional RNAs often have a conserved secondary structure which may be better preserved than the RNA sequence. The CMs used to describe each family are a slightly more complicated relative of the profile hidden Markov models (HMMs) used by Pfam. CMs can simultaneously model RNA sequence and the structure in an elegant and accurate fashion (Rfam description from: http://rfam.xfam.org/).

Please cite: Nawrocki, E. P., Burge, S. W., Bateman, A., Daub, J., Eberhardt, R. Y., Eddy, S. R., Floden, E. W., Gardner, P. P., Jones, T. A., Tate, J., et al. (2014). Rfam 12.0: updates to the rna families database. Nucleic acids research, page gku1063[43].

This functionality can be found under *Functional Analysis* → *Coding Potential* → *Run Rfam*. A dialog screen appears (see image below). Sequences longer than a given length can be skipped during the analysis.

[43] https://academic.oup.com/nar/article/43/D1/D130/2437148/Rfam-12-0-updates-to-the-RNA-families-database



Run Rfam blast search against Xfam servers (rfam.xfam.org).

The Rfam database is a collection of RNA sequence families of structural RNAs including non-coding RNA genes as well as cis-regulatory elements. Each family is represented by multiple sequence alignments and covariance models (CMs).

Rfam aims to facilitate the identification and classification of new members of known sequence families, and distributes annotation of ncRNAs in over 200 complete genome sequences.

Important note:

Via this function you communicate directly with the RFAM servers. Any issues regarding the functionality, performance or obtained results depend on the EMBL-EBI.

Questions regarding the RFAM service can be send to:
rfam-help@ebi.ac.uk.

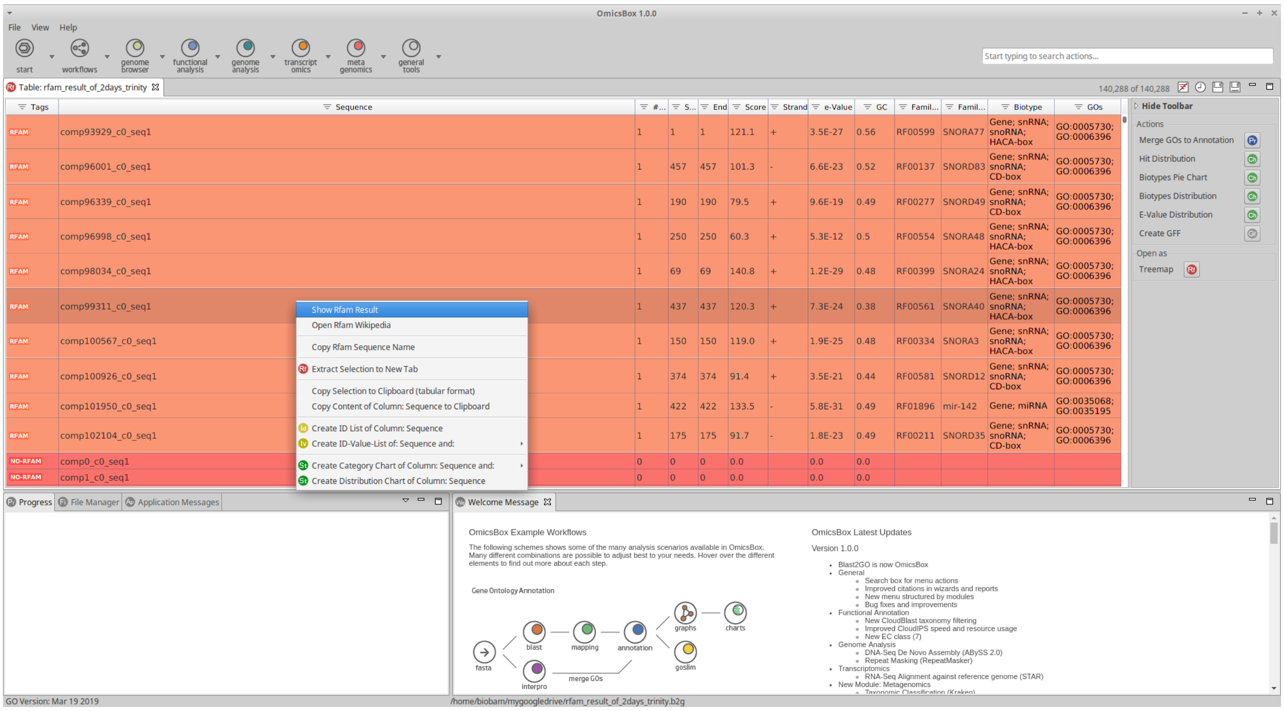
Sequence Length Filter: 2000

Figure 1: Rfam Dialog

Click on the *Run* button to start the analysis. It may take a while depending on the number of sequences and the EMBL-EBI servers.

7.9.2 Results Table

Once Rfam analysis has begun a table with the corresponding results will be displayed in a new tab. Sequences will turn red/orange depending if Rfam found hits for them (red if no hits were found, orange otherwise). White rows are sequences that have not been analysed yet. For each sequence It is possible to consult details about each one of their hits using the context menu (similar to consult Blast results).



Tags	Sequence	#...	S...	End	Score	Strand	e-Value	GC	Famil...	Famil...	Biotype	GOs
RFAM	comp93929_c0_seq1	1	1	1	121.1	+	3.5E-27	0.56	RF00599	SNORA77	Gene; snRNA; snoRNA; HACA-box	GO:0005730; GO:0006396
RFAM	comp96001_c0_seq1	1	457	457	101.3	-	6.6E-23	0.52	RF00137	SNORD83	Gene; snRNA; snoRNA; CD-box	GO:0005730; GO:0006396
RFAM	comp96339_c0_seq1	1	190	190	79.5	+	9.6E-19	0.49	RF00277	SNORD49	Gene; snRNA; snoRNA; CD-box	GO:0005730; GO:0006396
RFAM	comp96998_c0_seq1	1	250	250	60.3	+	5.3E-12	0.5	RF00554	SNORA48	Gene; snRNA; snoRNA; HACA-box	GO:0005730; GO:0006396
RFAM	comp98034_c0_seq1	1	69	69	140.8	+	1.2E-29	0.48	RF00399	SNORA24	Gene; snRNA; snoRNA; HACA-box	GO:0005730; GO:0006396
RFAM	comp99311_c0_seq1	1	437	437	120.3	+	7.3E-24	0.38	RF00561	SNORA40	Gene; snRNA; snoRNA; HACA-box	GO:0005730; GO:0006396
RFAM	comp100567_c0_seq1	1	150	150	119.0	+	1.9E-25	0.48	RF00334	SNORA3	Gene; snRNA; snoRNA; HACA-box	GO:0005730; GO:0006396
RFAM	comp100926_c0_seq1	1	374	374	91.4	+	3.5E-21	0.44	RF00581	SNORD12	Gene; snRNA; snoRNA; CD-box	GO:0005730; GO:0006396
RFAM	comp101950_c0_seq1	1	422	422	133.5	-	5.8E-31	0.49	RF01896	mir-142	Gene; miRNA	GO:0035068; GO:0035195
RFAM	comp102104_c0_seq1	1	175	175	91.7	-	1.8E-23	0.49	RF00211	SNORD35	Gene; snRNA; snoRNA; CD-box	GO:0005730; GO:0006396
NO-RFAM	comp0_c0_seq1	0	0	0	0.0		0.0	0.0				
NO-RFAM	comp1_c0_seq1	0	0	0	0.0		0.0	0.0				

OmicsBox Example Workflows

The following schemes shows some of the many analysis scenarios available in OmicsBox. Many different combinations are possible to adjust best to your needs. Hover over the different elements to find out more about each step.

OmicsBox Latest Updates

Version 1.0.0

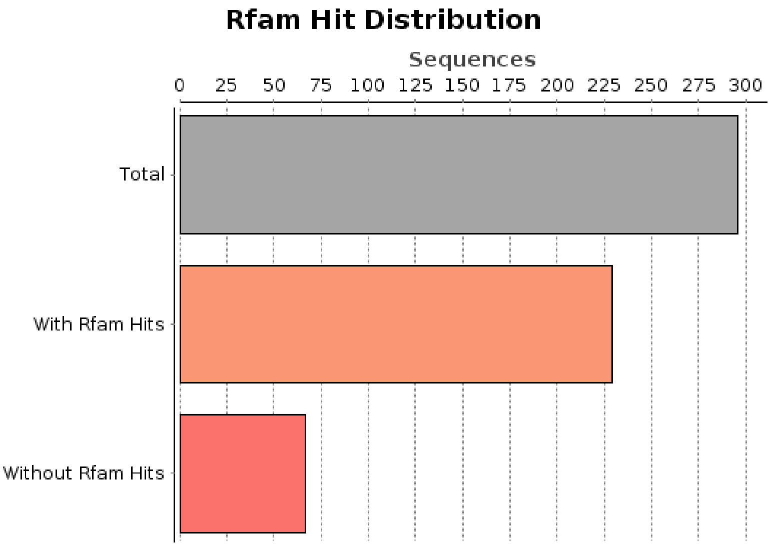
- Blast2GO is now OmicsBox
- General
  - Search box for menu actions
  - Improved citations in wizards and reports
  - New menu structured by modules
  - Bug fixes and improvements
- Functional Annotation
  - New CloudBlast taxonomy filtering
  - Improved CloudIPS speed and resource usage
  - New EC class (7)
- Genome Analysis
  - DNA-Seq De Novo Assembly (ABySS 2.0)
  - Repeat Masking (RepeatMasker)
- Transcriptomics
  - RNA-Seq Alignment against reference genome (STAR)
- New Module: Metagenomics
  - Taxonomic Classification (Kraken)

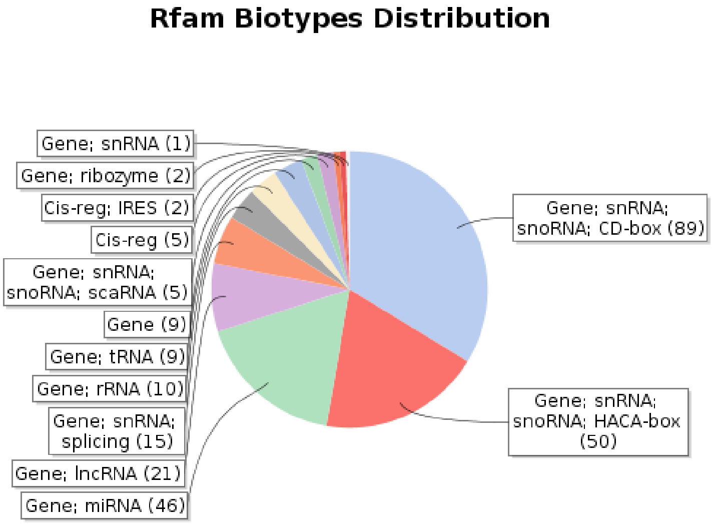
Figure 2: Rfam Table Results

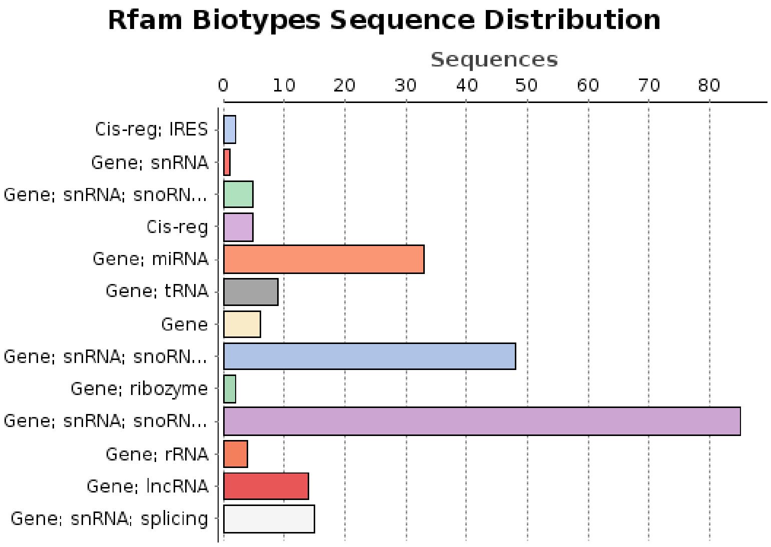
7.9.2.1 Sidebar options

In the sidebar there are located all possible action that can be performed for the Rfam result, including one option for the visual display of the results:

1. **Hit Distribution:** This chart shows a distribution chart of the number sequences with hits in the Rfam analysis.
2. **Biotypes Pie Chart:** This pie chart shows the distribution of the Rfam families of the sequences.
3. **Biotypes Distribution:** The same as the former but in a bar-style.
4. **E-Value Distribution:** This chart plots the distribution of E-values for the Rfam hits.
5. **Create GFF:** This will create a GFF file for the Rfam results.
6. **Open as Treemap:** This visualisation allows to see the Rfam families (hierarchical, tree-structured data in general) as a set of nested rectangles.


Rfam Hit Distribution
Sequences
0
25
50
75
100
125
150
175
200
225
250
275
300
Total
With Rfam Hits
Without Rfam Hits


Rfam Biotypes Distribution
Gene; snRNA (1)
Gene; ribozyme (2)
Cis-reg; IRES (2)
Cis-reg (5)
Gene; snRNA; snoRNA; scaRNA (5)
Gene (9)
Gene; tRNA (9)
Gene; rRNA (10)
Gene; snRNA; splicing (15)
Gene; lncRNA (21)
Gene; miRNA (46)
Gene; snRNA; snoRNA; CD-box (89)
Gene; snRNA; snoRNA; HACA-box (50)



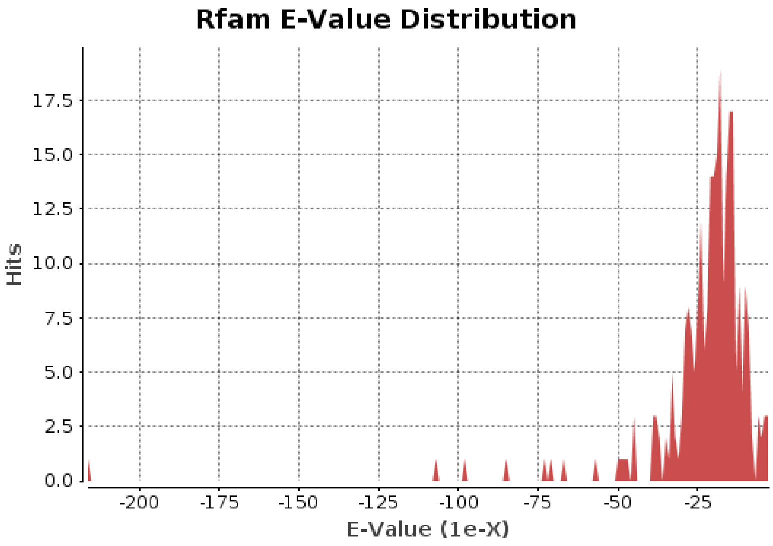


Figure 3: Rfam Statistics Graphs and Visualization

Additionally, like many others results in OmicsBox, It is possible to display the Rfam result in a different way: the *Treemap* representation to see the Rfam families (hierarchical, tree-structured data in general) as a set of nested rectangles.

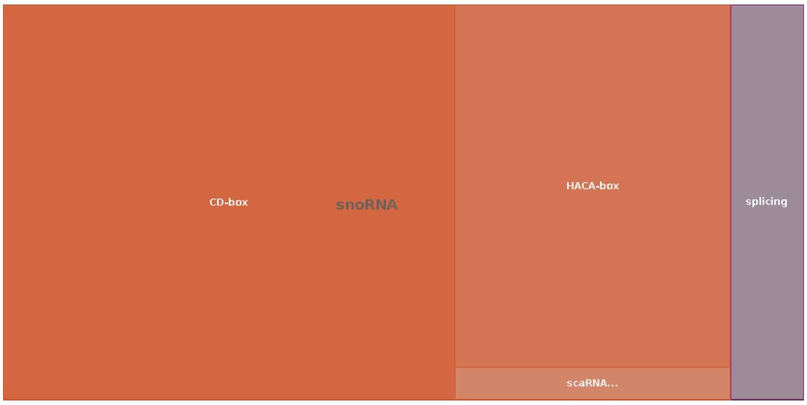


Figure 4: Rfam Tree Map

7.10 Enrichment Analysis

Researchers often want to retrieve a functional profile of these significant genes, in order to gain a better understanding of the underlying biological processes.
Functional enrichment analysis is a procedure to identify functions that are over-represented in a set of genes and may have an association with an experimental condition (e.g. phenotype, treatment…). These methods use statistical approaches to identify significantly enriched or depleted groups of genes.

The Functional Analysis Module offers two different statistical tests in order to Functional Enrichment Analysis, the Fisher Exact Test and Gene Set Enrichment Analysis.

Both Fisher's Exact Test and GSEA Enrichment methods need a ranked ID gene list. For a detailed tutorial on how to obtain these lists in each case, please link here[44].

7.10.1 Fisher's Exact Test

Content of this page:

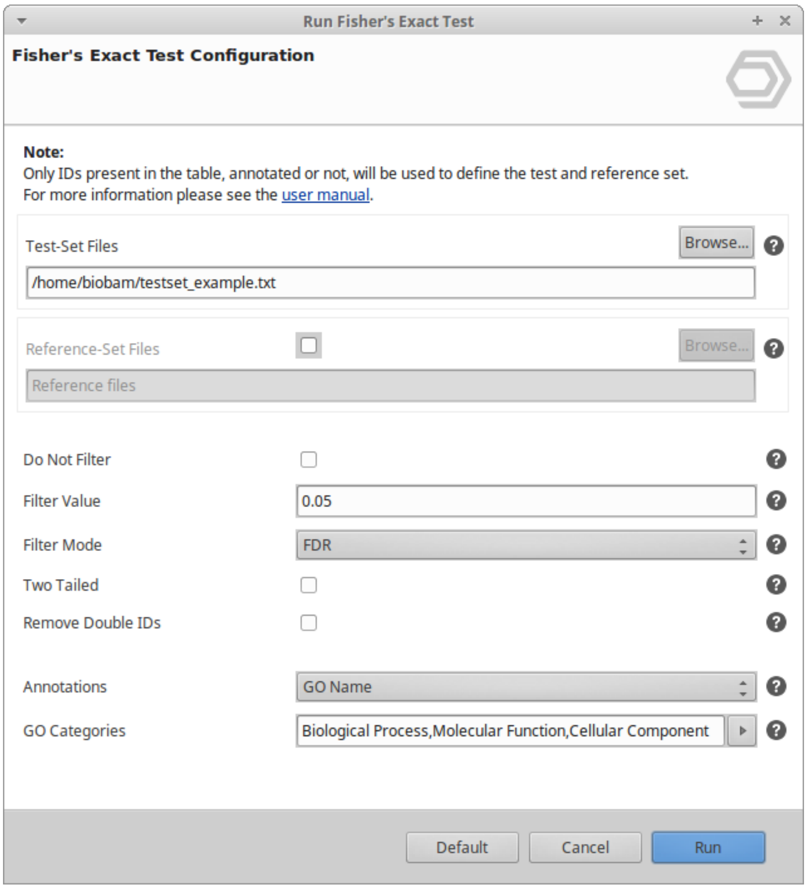
7.10.1.1 Introduction

OmicsBox has integrated the FatiGO package for statistical assessment of annotation differences between 2 sets of sequences. This package uses Fisher's Exact Test and corrects for multiple testing. For this analysis, the completion (but not exclusively) of the involved sequences with their annotations must be loaded in the application. This can either be the result of a OmicsBox annotation or the imported annotation by file (.annot), see Gene Ontology Annotation(see page 58) of this manual.

[44] https://www.biobam.com/how-to-create-a-gene-list-within-blast2go-to-run-the-functional-enrichment-analysis/

This functionality can be found under *Functional Analysis → Enrichment Analysis → Enrichment Analysis (Fisher's Exact Test)*. A dialog screen appears (see image below). Test and Reference Sequences can be selected by uploading text files or ID-List .box files containing the lists of sequence IDs for the 2 groups. When there is no reference set selected, the whole dataset present in the project will be taken as reference. A detailed description of each parameter is available by clicking the help icon next to the parameter.

> **New:** starting from Blast2GO 4.1 and in OmicsBox, it is possible to perform Fisher's Exact Test for different types of annotations for most of the results generated in OmicsBox. The *Annotations* parameter allows selecting the column of the table to use as an annotation. With this feature, It is possible to perform an enrichment analysis of enzymes or InterPro IDs for example.



Run Fisher's Exact Test

Fisher's Exact Test Configuration

Note:
Only IDs present in the table, annotated or not, will be used to define the test and reference set.
For more information please see the user manual.

Test-Set Files: /home/biobam/testset_example.txt

Reference-Set Files: □ Reference files

Do Not Filter: □

Filter Value: 0.05

Filter Mode: FDR

Two Tailed: □

Remove Double IDs: □

Annotations: GO Name

GO Categories: Biological Process,Molecular Function,Cellular Component

Figure 1: Run Fisher's Exact Test Wizard Dialog

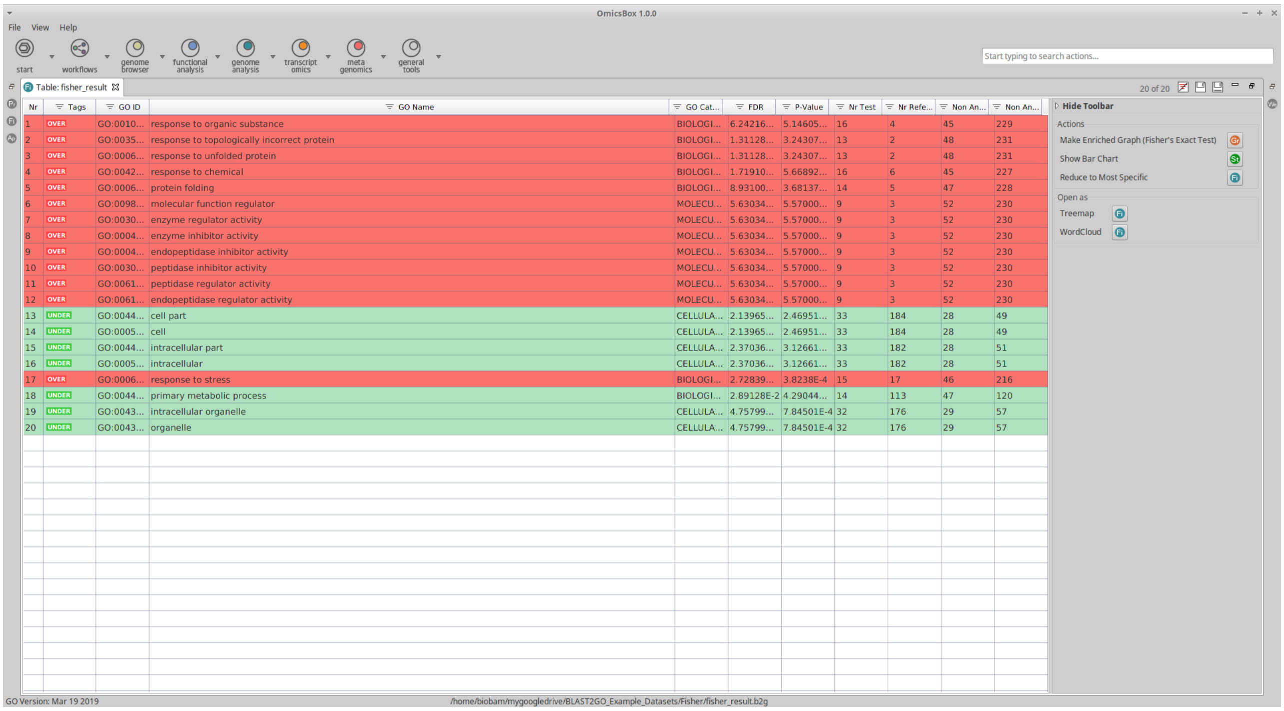
Click on the *Run* button to start the analysis. It may take a while depending on the number of annotations.

7.10.1.2 Results Table

Once completed the results table will be shown in a new tab (see image below), where the adjusted p-values of each annotation above a given threshold will be shown. The main columns are:

FDR	p-Value
Corrected p-value by False Discovery Rate control according to Benjamini-Hochberg.	p-Value without multiple testing corrections

For further details please refer to the FatiGO publication (Al-Shahrour, F., Díaz-Uriarte, R., and Dopazo, J. (2004). Fatigo: a web tool for finding significant associations of gene ontology terms with groups of genes. Bioinformatics, 20(4):578–580[45]).



Nr	Tags	GO ID	GO Name	GO Cat...	FDR	P-Value	Nr Test	Nr Refe...	Non An...	Non An...
1	OVER	GO:0010...	response to organic substance	BIOLOGI...	6.24216...	5.14605...	16	4	45	229
2	OVER	GO:0035...	response to topologically incorrect protein	BIOLOGI...	1.31128...	3.24307...	13	2	48	231
3	OVER	GO:0006...	response to unfolded protein	BIOLOGI...	1.31128...	3.24307...	13	2	48	231
4	OVER	GO:0042...	response to chemical	BIOLOGI...	1.71910...	5.66892...	16	6	45	227
5	OVER	GO:0006...	protein folding	BIOLOGI...	8.93100...	3.68137...	14	5	47	228
6	OVER	GO:0098...	molecular function regulator	MOLECU...	5.63034...	5.57000...	9	3	52	230
7	OVER	GO:0030...	enzyme regulator activity	MOLECU...	5.63034...	5.57000...	9	3	52	230
8	OVER	GO:0004...	enzyme inhibitor activity	MOLECU...	5.63034...	5.57000...	9	3	52	230
9	OVER	GO:0004...	endopeptidase inhibitor activity	MOLECU...	5.63034...	5.57000...	9	3	52	230
10	OVER	GO:0030...	peptidase inhibitor activity	MOLECU...	5.63034...	5.57000...	9	3	52	230
11	OVER	GO:0061...	peptidase regulator activity	MOLECU...	5.63034...	5.57000...	9	3	52	230
12	OVER	GO:0061...	endopeptidase regulator activity	MOLECU...	5.63034...	5.57000...	9	3	52	230
13	UNDER	GO:0044...	cell part	CELLULA...	2.13965...	2.46951...	33	184	28	49
14	UNDER	GO:0005...	cell	CELLULA...	2.13965...	2.46951...	33	184	28	49
15	UNDER	GO:0044...	intracellular part	CELLULA...	2.37036...	3.12661...	33	182	28	51
16	UNDER	GO:0005...	intracellular	CELLULA...	2.37036...	3.12661...	33	182	28	51
17	OVER	GO:0006...	response to stress	BIOLOGI...	2.72839...	3.8238E-4	15	17	46	216
18	UNDER	GO:0044...	primary metabolic process	BIOLOGI...	2.89128E-2	4.29044...	14	113	47	120
19	UNDER	GO:0043...	intracellular organelle	CELLULA...	4.75799...	7.84501E-4	32	176	29	57
20	UNDER	GO:0043...	organelle	CELLULA...	4.75799...	7.84501E-4	32	176	29	57

GO Version: Mar 19 2019 /home/biobam/mygoogledrive/BLAST2GO_Example_Datasets/Fisher/fisher_result.b2g

Figure 2: Enrichment Results Table

Using the context menu of each row It is possible to get more details about the annotation and also create an ID-List with the sequences annotated in the Test-Set or the Reference-Set.

- #Test is the number of sequences that are annotated with the GO and are in the test set.
- #NotAnnotTest is the number of sequences that are not annotated with that GO, that is in the test set.

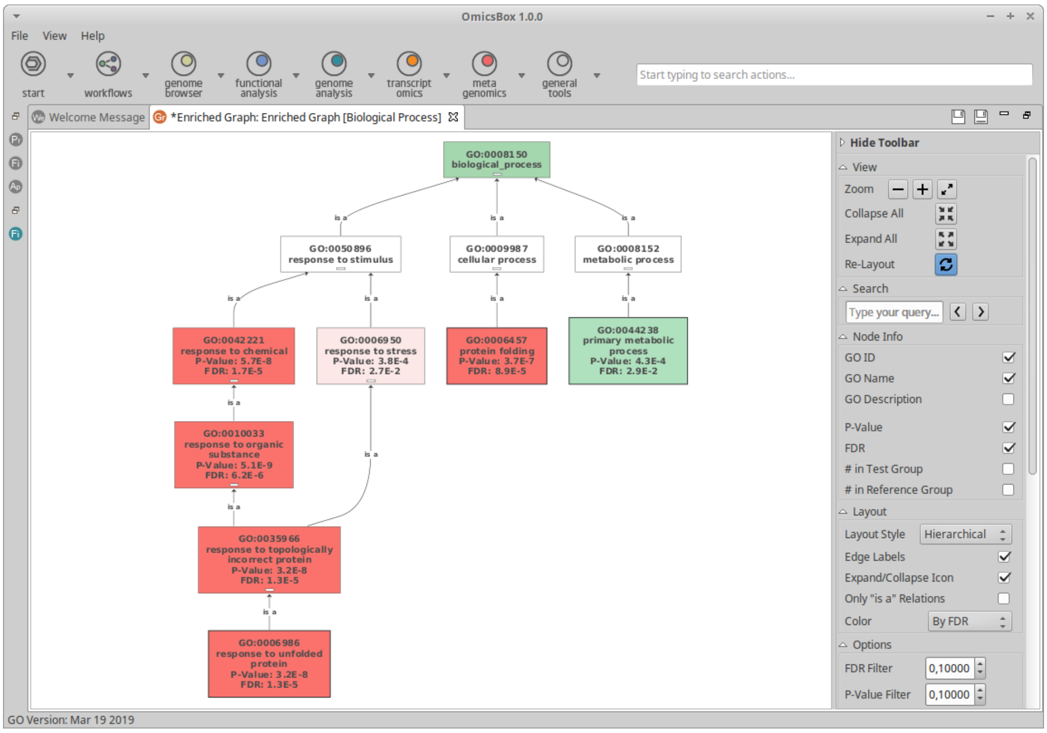
Adding these two numbers it gives the total amount of sequences that are annotated at all in your test set e.g. GO:0061135: 9 + 52 = 61

[45] https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btg455

Sidebar Options

In the sidebar there are located all possible action that can be performed for this enrichment result, including two options for the visual display of the results:

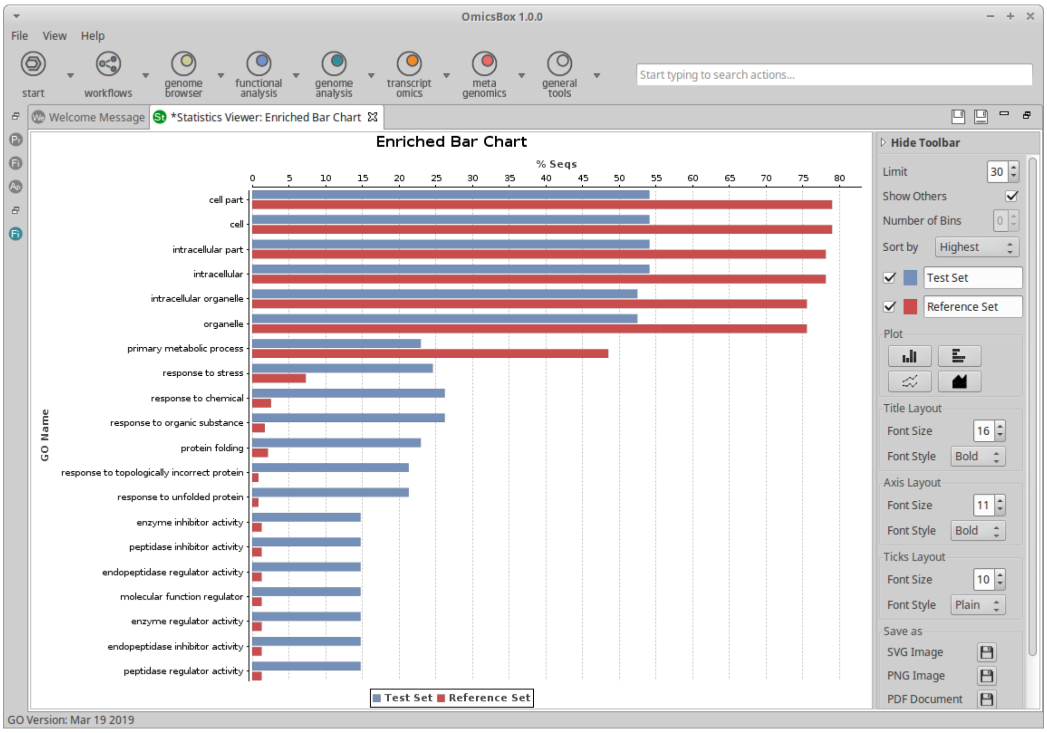
1. *Make Enriched Graph* (only for GO annotations): use this option to generate a representation on the GO DAG (see image below). Nodes are color-highlighted proportionally to their significance value. The user can choose which type of calculated p-value to use for highlighting and the threshold for filtering out nodes. Additionally, the *Filter intermediate* the checkbox will hide non-enriched nodes. More options are available in the graph viewer's sidebar. Gene Ontology Graphs(see page 79) of this manual gives further information on the graphical functions in OmicsBox.



GO ID ☑
GO Name ☑
GO Description □
P-Value ☑
FDR ☑
# in Test Group □
# in Reference Group □
Layout Style: Hierarchical
Edge Labels ☑
Expand/Collapse Icon ☑
Only "is a" Relations □
Color: By FDR
FDR Filter: 0,10000
P-Value Filter: 0,10000

Figure 3: Enriched Graph

2. *Show Bar Chart*: this option generates a bar display of the percentages of sequences at both, test and reference set, for each annotation of the table (see image below).


OmicsBox 1.0.0
File
View
Help
start
workflows
genome browser
functional analysis
genome analysis
transcript omics
meta genomics
general tools
Start typing to search actions...
Welcome Message
*Statistics Viewer: Enriched Bar Chart
Enriched Bar Chart
% Seqs
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
GO Name
cell part
cell
intracellular part
intracellular
intracellular organelle
organelle
primary metabolic process
response to stress
response to chemical
response to organic substance
protein folding
response to topologically incorrect protein
response to unfolded protein
enzyme inhibitor activity
peptidase inhibitor activity
endopeptidase regulator activity
molecular function regulator
enzyme regulator activity
endopeptidase inhibitor activity
peptidase regulator activity
🟦 Test Set 🟥 Reference Set
Hide Toolbar
Plot
GO Version: Mar 19 2019

Limit 30

Show Others ☑

Number of Bins 0

Sort by Highest

☑ 🟦 Test Set

☑ 🟥 Reference Set

Title Layout

Font Size 16

Font Style Bold

Axis Layout

Font Size 11

Font Style Bold

Ticks Layout

Font Size 10

Font Style Plain

Save as

SVG Image

PNG Image

PDF Document

Figure 4: Enriched Bar Chart

3. *Reduce to Most Specific* (only for GO annotations): use this option to remove more general GO terms from the results and get only the most specific terms (with the lowest level in the GO DAG).

Additionally, like many others results in OmicsBox, It is possible to display the enrichment results in two different ways: the *Treemap* representation to compare the most enriched annotations by their size and the *WordCloud* representation to summarise relevant annotations in a fashionable way.

7.10.2 Gene Set Enrichment Analysis (GSEA)

Content of this page:

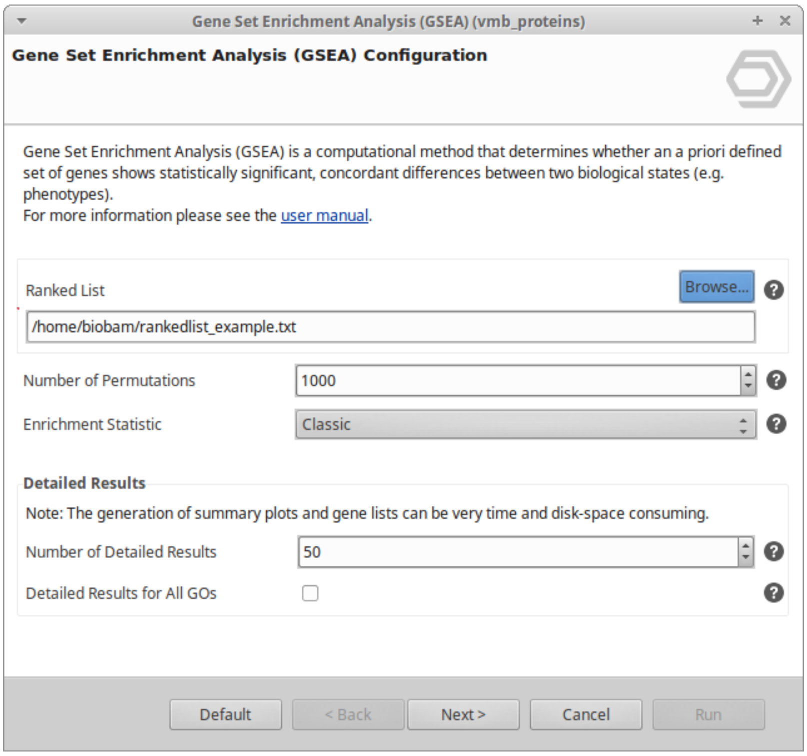
7.10.2.1 Introduction

OmicsBox includes the GSEA computational method that determines whether an a priori defined set of genes shows statistically significant, concordant differences between two biological states. GSEA considers experiments with genome-wide expression profiles from samples belonging to two classes, labelled 1 or 2. Genes are ranked based on the correlation between their expression and the class distinction by using any suitable metric. Given an a priori defined set of genes S (e.g., genes encoding products in a metabolic pathway, located in the same cytogenetic band, or sharing the same GO category), the goal of GSEA is to determine whether the members of S are randomly distributed throughout L or primarily found at the top or bottom.

For further details please refer to the GSEA publication: Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences, 102(43):15545– 15550[46].

For this analysis, the completion (but not exclusively) of the involved sequences with their annotations must be loaded in the application. This can either be the result of an OmicsBox annotation or the imported annotation by file (.annot), see Gene Ontology Annotation(see page 58) of this manual.

This functionality can be found under *Functional Analysis → Enrichment Analysis → Gene Set Enrichment Analysis (GSEA)*. A dialog screen appears (see image below). Ranked list of genes can be selected by uploading text files or ID-Value-List .box/.b2g files containing the lists of sequence IDs and a statistical value for each one. A detailed description of each parameter is available by clicking the help icon next to the parameter.



Gene Set Enrichment Analysis (GSEA) (vmb_proteins)

Gene Set Enrichment Analysis (GSEA) Configuration

Gene Set Enrichment Analysis (GSEA) is a computational method that determines whether an a priori defined set of genes shows statistically significant, concordant differences between two biological states (e.g. phenotypes).
For more information please see the user manual.

Ranked List: /home/biobam/rankedlist_example.txt

Number of Permutations: 1000

Enrichment Statistic: Classic

Detailed Results

Note: The generation of summary plots and gene lists can be very time and disk-space consuming.

Number of Detailed Results: 50

Detailed Results for All GOs: □

Figure 1: GSEA Dialog

Click on the *Run* button to start the analysis. It may take a while depending on the number of permutations selected.

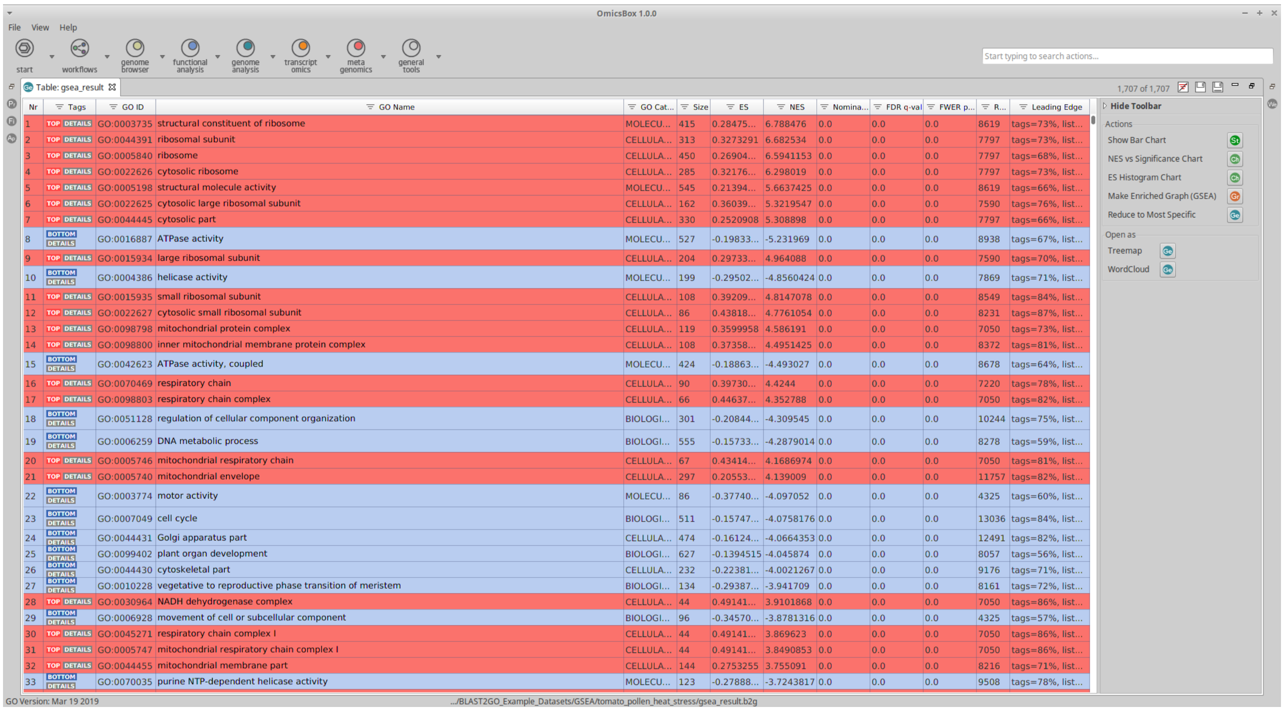
[46] http://www.pnas.org/content/102/43/15545.abstract

7.10.2.2 Results

Once completed the results table will be shown in a new tab (see image below), where the adjusted p-values of each annotation above a given threshold will be shown. The main columns are:

ES	NES	FDR	Nominal p-value
Reflects the degree to which a gene set is overrepresented at the top or bottom of a ranked list of genes	By normalizing the enrichment score, GSEA accounts for differences in gene set size and in correlations between gene sets and the expression dataset	The estimated probability that a gene set with a given NES represents a false positive finding	Estimates the statistical significance of the enrichment score for a single gene set

For further details please refer to the GSEA User Guide[47].



Nr	Tags	GO ID	GO Name	GO Cat...	Size	ES	NES	Nomina...	FDR q-val	FWER p...	R...	Leading Edge
1	TOP DETAILS	GO:0003735	structural constituent of ribosome	MOLECU...	415	0.28475...	6.788476	0.0	0.0	0.0	8619	tags=73%, list...
2	TOP DETAILS	GO:0044391	ribosomal subunit	CELLULA...	313	0.3273291	6.682534	0.0	0.0	0.0	7797	tags=73%, list...
3	TOP DETAILS	GO:0005840	ribosome	CELLULA...	450	0.26904...	6.5941153	0.0	0.0	0.0	7797	tags=68%, list...
4	TOP DETAILS	GO:0022626	cytosolic ribosome	CELLULA...	285	0.32176...	6.298019	0.0	0.0	0.0	7797	tags=73%, list...
5	TOP DETAILS	GO:0005198	structural molecule activity	MOLECU...	545	0.21394...	5.6637425	0.0	0.0	0.0	8619	tags=66%, list...
6	TOP DETAILS	GO:0022625	cytosolic large ribosomal subunit	CELLULA...	162	0.36039...	5.3219547	0.0	0.0	0.0	7590	tags=76%, list...
7	TOP DETAILS	GO:0044445	cytosolic part	CELLULA...	330	0.2520908	5.308898	0.0	0.0	0.0	7797	tags=66%, list...
8	BOTTOM DETAILS	GO:0016887	ATPase activity	MOLECU...	527	-0.19833...	-5.231969	0.0	0.0	0.0	8938	tags=67%, list...
9	TOP DETAILS	GO:0015934	large ribosomal subunit	CELLULA...	204	0.29733...	4.964088	0.0	0.0	0.0	7590	tags=70%, list...
10	BOTTOM DETAILS	GO:0004386	helicase activity	MOLECU...	199	-0.29502...	-4.8560424	0.0	0.0	0.0	7869	tags=71%, list...
11	TOP DETAILS	GO:0015935	small ribosomal subunit	CELLULA...	108	0.39209...	4.8147078	0.0	0.0	0.0	8549	tags=84%, list...
12	TOP DETAILS	GO:0022627	cytosolic small ribosomal subunit	CELLULA...	86	0.43818...	4.7761054	0.0	0.0	0.0	8231	tags=87%, list...
13	TOP DETAILS	GO:0098798	mitochondrial protein complex	CELLULA...	119	0.3599958	4.586191	0.0	0.0	0.0	7050	tags=73%, list...
14	TOP DETAILS	GO:0098800	inner mitochondrial membrane protein complex	CELLULA...	108	0.37358...	4.4951425	0.0	0.0	0.0	8372	tags=81%, list...
15	BOTTOM DETAILS	GO:0042623	ATPase activity, coupled	MOLECU...	424	-0.18863...	-4.493027	0.0	0.0	0.0	8678	tags=64%, list...
16	TOP DETAILS	GO:0070469	respiratory chain	CELLULA...	90	0.39730...	4.4244	0.0	0.0	0.0	7220	tags=78%, list...
17	TOP DETAILS	GO:0098803	respiratory chain complex	CELLULA...	66	0.44637...	4.352788	0.0	0.0	0.0	7050	tags=82%, list...
18	BOTTOM DETAILS	GO:0051128	regulation of cellular component organization	BIOLOGI...	301	-0.20844...	-4.309545	0.0	0.0	0.0	10244	tags=75%, list...
19	BOTTOM DETAILS	GO:0006259	DNA metabolic process	BIOLOGI...	555	-0.15733...	-4.2879014	0.0	0.0	0.0	8278	tags=59%, list...
20	TOP DETAILS	GO:0005746	mitochondrial respiratory chain	CELLULA...	67	0.43414...	4.1686974	0.0	0.0	0.0	7050	tags=81%, list...
21	TOP DETAILS	GO:0005740	mitochondrial envelope	CELLULA...	297	0.20553...	4.139009	0.0	0.0	0.0	11757	tags=82%, list...
22	BOTTOM DETAILS	GO:0003774	motor activity	MOLECU...	86	-0.37740...	-4.097052	0.0	0.0	0.0	4325	tags=60%, list...
23	BOTTOM DETAILS	GO:0007049	cell cycle	BIOLOGI...	511	-0.15747...	-4.0758176	0.0	0.0	0.0	13036	tags=84%, list...
24	BOTTOM DETAILS	GO:0044431	Golgi apparatus part	CELLULA...	474	-0.16124...	-4.0664353	0.0	0.0	0.0	12491	tags=82%, list...
25	BOTTOM DETAILS	GO:0099402	plant organ development	BIOLOGI...	627	-0.1394515	-4.045874	0.0	0.0	0.0	8057	tags=56%, list...
26	BOTTOM DETAILS	GO:0044430	cytoskeletal part	CELLULA...	232	-0.22381...	-4.0021267	0.0	0.0	0.0	9176	tags=71%, list...
27	BOTTOM DETAILS	GO:0010228	vegetative to reproductive phase transition of meristem	BIOLOGI...	134	-0.29387...	-3.941709	0.0	0.0	0.0	8161	tags=72%, list...
28	TOP DETAILS	GO:0030964	NADH dehydrogenase complex	CELLULA...	44	0.49141...	3.9101868	0.0	0.0	0.0	7050	tags=86%, list...
29	BOTTOM DETAILS	GO:0006928	movement of cell or subcellular component	BIOLOGI...	96	-0.34570...	-3.8781316	0.0	0.0	0.0	4325	tags=57%, list...
30	TOP DETAILS	GO:0045271	respiratory chain complex I	CELLULA...	44	0.49141...	3.869623	0.0	0.0	0.0	7050	tags=86%, list...
31	TOP DETAILS	GO:0005747	mitochondrial respiratory chain complex I	CELLULA...	44	0.49141...	3.8490853	0.0	0.0	0.0	7050	tags=86%, list...
32	TOP DETAILS	GO:0044455	mitochondrial membrane part	CELLULA...	144	0.2753255	3.755091	0.0	0.0	0.0	8216	tags=71%, list...
33	BOTTOM DETAILS	GO:0070035	purine NTP-dependent helicase activity	MOLECU...	123	-0.27888...	-3.7243817	0.0	0.0	0.0	9508	tags=78%, list...

GO Version: Mar 19 2019 .../BLAST2GO_Example_Datasets/GSEA/tomato_pollen_heat_stress/gsea_result.b2g

Figure 2: GSEA result table

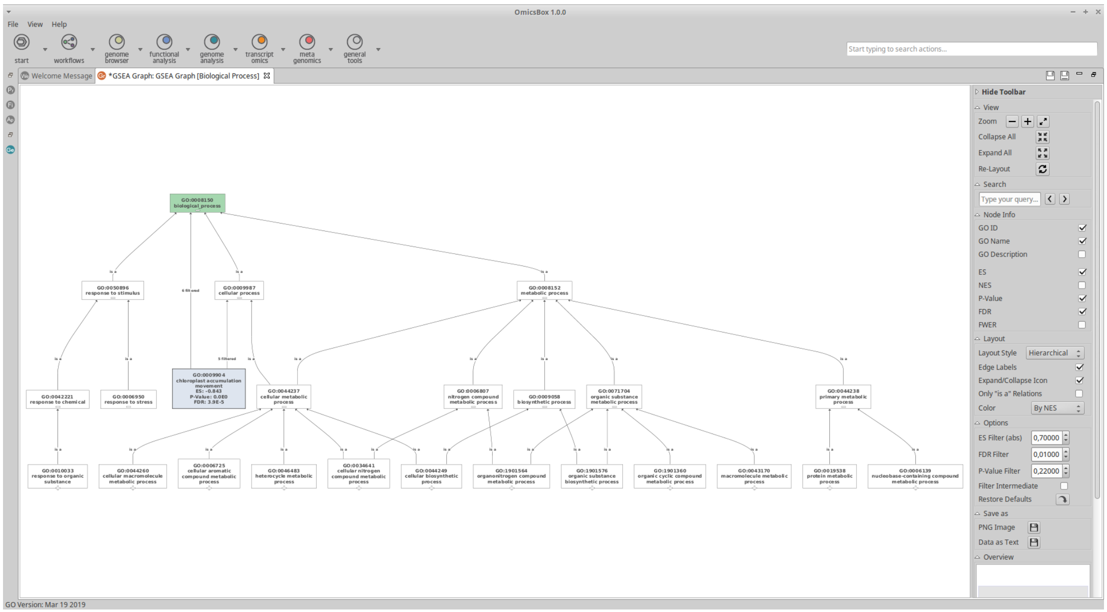
Using the context menu of the rows tagged with the *Details* tag It is possible to get more details about the GO term, including the enrichment statistics, and also create an ID-List with the core enrichment sequences for each GO term.

[47] http://software.broadinstitute.org/gsea/doc/GSEAUserGuideFrame.html

Sidebar Options

In the sidebar there are located all possible action that can be performed for this enrichment result, including two options for the visual display of the results:

1. *Make Enriched Graph*: use this option to generate a representation on the GO DAG (see image below). Nodes are color-highlighted proportionally to their significance value. The user can choose which type of calculated p-value to use for highlighting and the threshold for filtering out nodes.



GO ID ☑
GO Name ☑
GO Description ☐
ES ☑
NES ☐
P-Value ☑
FDR ☑
FWER ☐
Layout Style: Hierarchical
Edge Labels ☑
Expand/Collapse Icon ☑
Only "is a" Relations ☐
Color: By NES
ES Filter (abs): 0,70000
FDR Filter: 0,01000
P-Value Filter: 0,22000
Filter Intermediate ☐
Restore Defaults

Figure 3: Enriched Graph

2. *NES vs Significance Chart*: this option generates a plot of p-values versus normalized enrichment scores, which provides a quick, visual way to grasp the number of enriched gene sets that are significant (see image below).

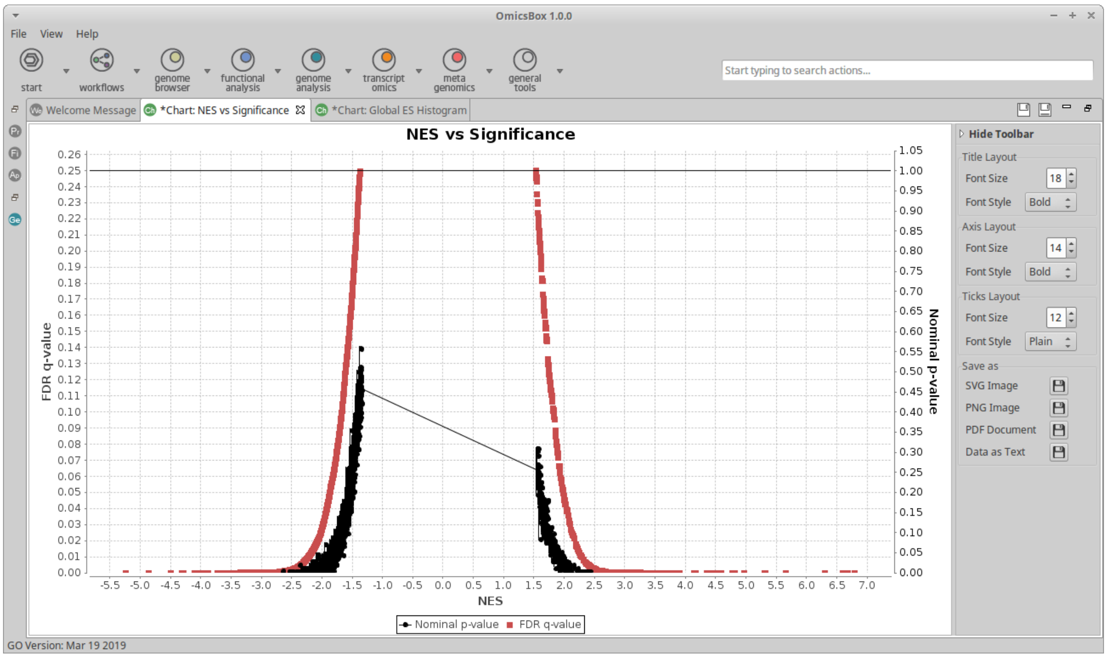


Figure 4: NES vs Significance Chart

3. ES Histogram Chart: this option generates a histogram of enrichment scores across gene sets, which provides a quick, visual way to grasp the number of enriched gene sets. (see image below).

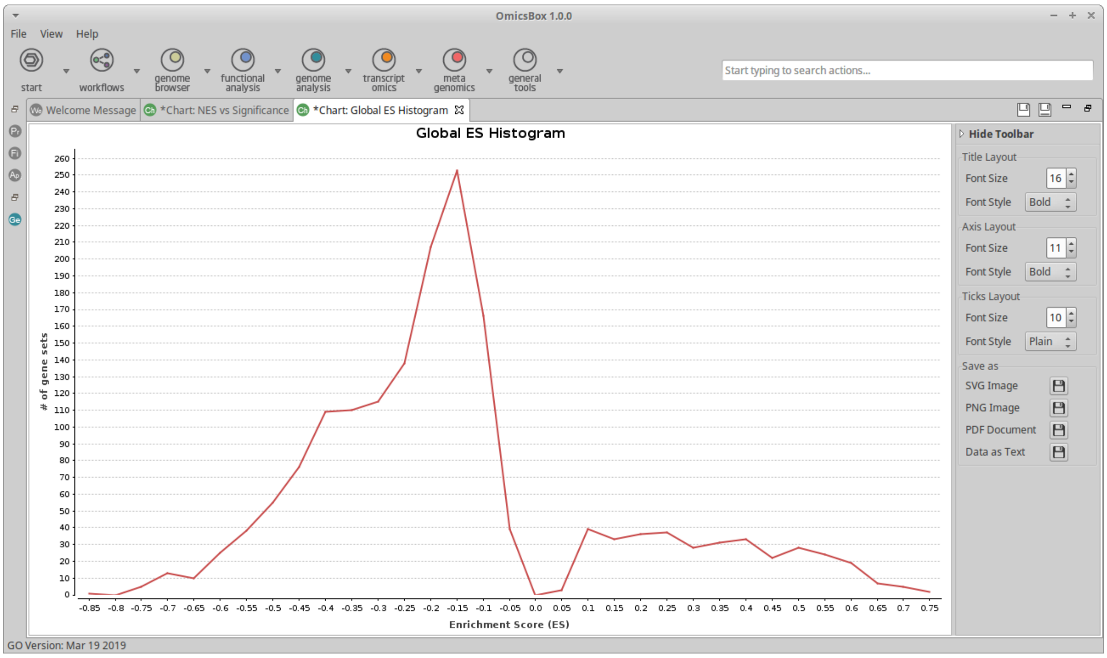


Figure 5: ES Histogram Chart

4. Reduce to Most Specific: use this option to remove more general GO terms from the results and get only the most specific terms (with the lowest level in the GO DAG).

Additionally, like many others results in OmicsBox, It is possible to display the enrichment results in two different ways: the *Treemap* representation to compare the most enriched GO terms by their size and the *WordCloud* representation to summarise relevant GO terms in a fashionable way.

7.11 Tools (Select, Rename, Search...)

Content of this page:

7.11.1 Menu Items

1. Set-to-Sense (Based on Best-Blast-Hit): Convert all selected sequences with a negative reading frame Best-Blast-Hit to anti-sense i.e. query-sequences will be translated to its reverse complement (e.g.: ATTG ->CAAT). The tag "_antisense" will be added to the end of the sequence names. Use the batch rename function to undo the name change.
2. Translate Longest ORF: Convert all selected sequences to its longest ORF protein sequence. The tag "_ORF" will be added to the sequence names. Use the batch rename function to undo the name change. The user may select the reading frame, the genetic code depending to the species that will be considered to the prediction.
3. Search Loaded Annotations in Another Annotation Set: Compare a set of annotations for a given group of sequences against the annotations already loaded in OmicsBox.
4. Find Duplicated Sequences: Mark as selected or directly remove all sequences in the dataset which have the exact same sequence string.
5. Find Similar Sequences: Detect, Select and/or remove similar sequences within one project.
6. Batch Rename: Perform a batch rename of all selected sequences by converting, replacing or adding text to the actual sequence name. Link here[48] for a detailed explanation on how to use this tool.

7.11.2 Find Similar Sequences

This function allows searching for similar sequences within a dataset. The search for similar sequences is done via BLAT[49] alignments. The function searches a list of sequences against itself and reports all alignments above a certain similarity percentage. It is possible to remove similar sequences from the project or to extract a less redundant result dataset into a new project.

[48] https://www.biobam.com/batch-rename-seq-ids/
[49] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC187518/

7.11.3 Find Duplicated Sequences

This function allows to quickly identify and remove redundant sequences (exactly the same sequences) within a dataset.

7.11.4 Select Sequences and Functions

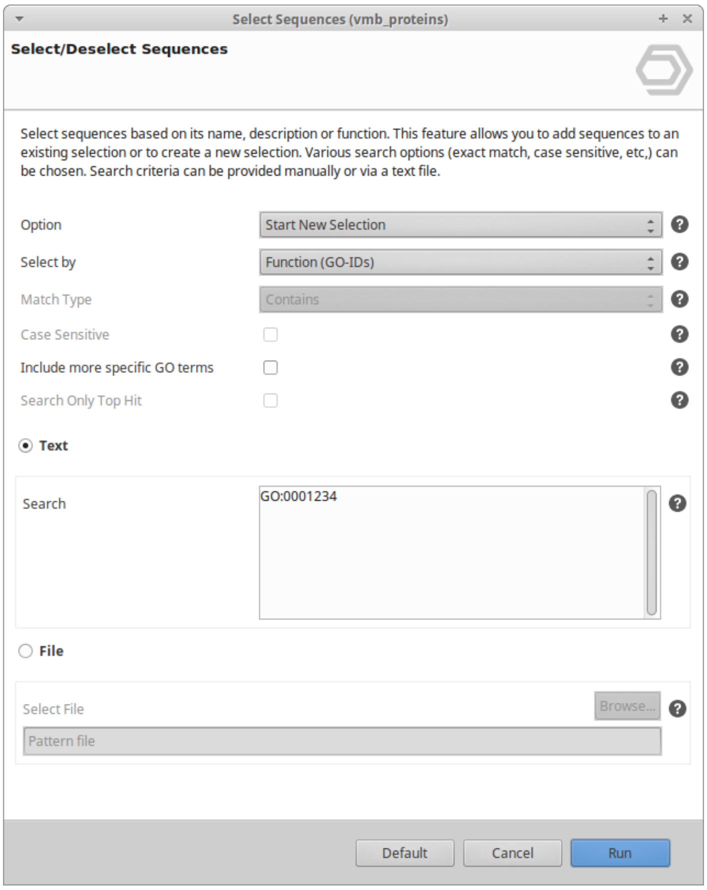
Content of this page:

There are different functions for selecting and deselecting sequences. Most functions in OmicsBox are only applied to selected elements. Selections allow to create subset or apply certain functions to parts of a given dataset.

7.11.4.1 Select Sequences

The Select Sequences feature can be applied to OmicsBox Projects only and allows to select sequences for many different criteria. Selections can by added to existing ones, subtracted or created from scratch.

- Sequence Name. This is a general function for (de)selecting sequences by loading a file containing a list sequence IDs.
- Sequence Description. OmicsBox allows to (de)select sequences according to Blast result description.
- Species.
- Function (GO-Terms or GO-IDs). This is a general function for (de)selecting sequences by loading a file containing a list of GO-Terms or GO-IDs.
- InterProScan IDs.
- Enzyme IDs.



Select Sequences (vmb_proteins)

Select/Deselect Sequences

Select sequences based on its name, description or function. This feature allows you to add sequences to an existing selection or to create a new selection. Various search options (exact match, case sensitive, etc,) can be chosen. Search criteria can be provided manually or via a text file.

Option: Start New Selection

Select by: Function (GO-IDs)

Match Type: Contains

Case Sensitive □

Include more specific GO terms □

Search Only Top Hit □

● Text

Search: GO:0001234

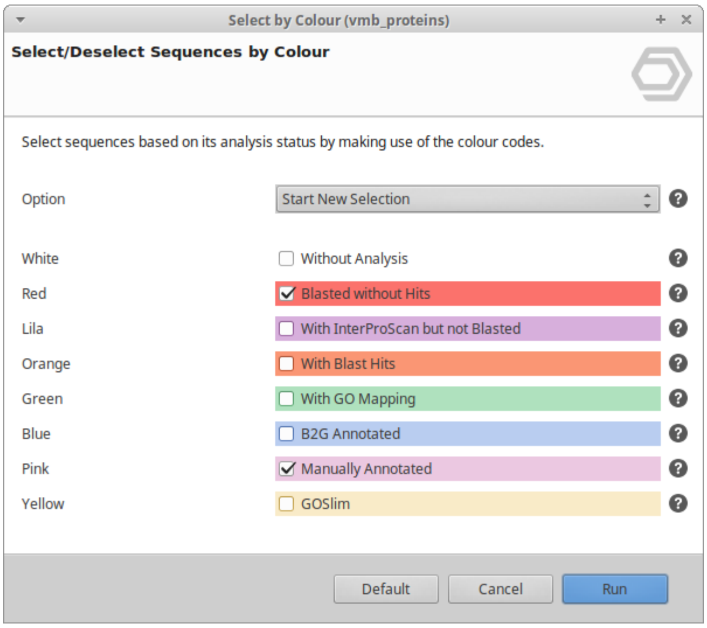
○ File

Select File: Pattern file

Figure 1: Start New Selection

7.11.4.2 Select Sequence by Color

This function allows (de)selection of sequences on the basis of their color code i.e. the processing stage they have.



Select by Colour (vmb_proteins)

Select/Deselect Sequences by Colour

Select sequences based on its analysis status by making use of the colour codes.

Option	Start New Selection
White	□ Without Analysis
Red	☑ Blasted without Hits
Lila	□ With InterProScan but not Blasted
Orange	□ With Blast Hits
Green	□ With GO Mapping
Blue	□ B2G Annotated
Pink	☑ Manually Annotated
Yellow	□ GOSlim

Figure 2: Selection by Color

7.11.4.3 Other Select Options

Invert Selection

This function will invert the current selection. Those sequences that are not selected will now be selected and vice versa.

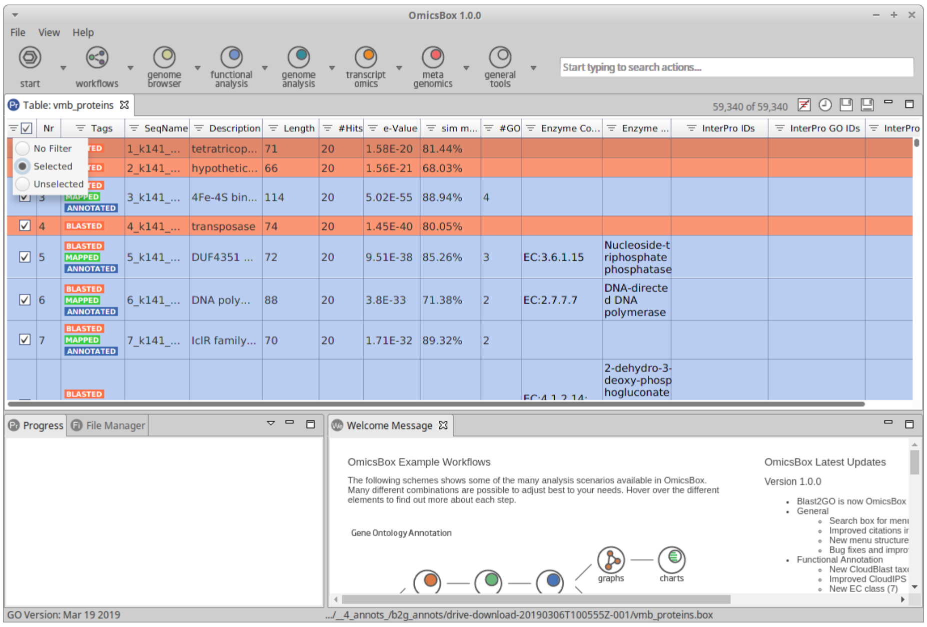
Delete Selected Sequences

This function will delete selected sequences from the Main Sequence Table

Extract Selection to New Tab

One can extract a subset of the selected sequences to a new project.

Once one has the desired sequences selected it is possible to hide/filter out the deselected ones by clicking on the icon next to the selection check box on the table.



Nr	Tags	SeqName	Description	Length	#Hits	e-Value	sim m...	#GO	Enzyme Co...	Enzyme ...	InterPro IDs	InterPro GO IDs	InterPro
1	BLASTED	1_k141_...	tetratricop...	71	20	1.58E-20	81.44%						
2	BLASTED	2_k141_...	hypothetic...	66	20	1.56E-21	68.03%						
3	BLASTED MAPPED ANNOTATED	3_k141_...	4Fe-4S bin...	114	20	5.02E-55	88.94%	4					
4	BLASTED	4_k141_...	transposase	74	20	1.45E-40	80.05%						
5	BLASTED MAPPED ANNOTATED	5_k141_...	DUF4351 ...	72	20	9.51E-38	85.26%	3	EC:3.6.1.15	Nucleoside-triphosphate phosphatase			
6	BLASTED MAPPED ANNOTATED	6_k141_...	DNA poly...	88	20	3.8E-33	71.38%	2	EC:2.7.7.7	DNA-directed DNA polymerase			
7	BLASTED MAPPED ANNOTATED	7_k141_...	IclR family...	70	20	1.71E-32	89.32%	2					
	BLASTED								EC:4.1.2.14:	2-dehydro-3-deoxy-phosphogluconate			

OmicsBox Example Workflows

The following schemes shows some of the many analysis scenarios available in OmicsBox. Many different combinations are possible to adjust best to your needs. Hover over the different elements to find out more about each step.

OmicsBox Latest Updates

Version 1.0.0

- Blast2GO is now OmicsBox
- General
  - Search box for menu
  - Improved citations i
  - New menu structure
  - Bug fixes and impro
- Functional Annotation
  - New CloudBlast tax
  - Improved CloudIPS
  - New EC class (7)

Figure 3: Show only selected sequences

With Ctrl+A in Windows/Linux or appleKey+A on Mac OS, all selected sequences will be marked. Now right click on one of the sequences on the table and choose the 2nd option Extract Selection to New Tab.

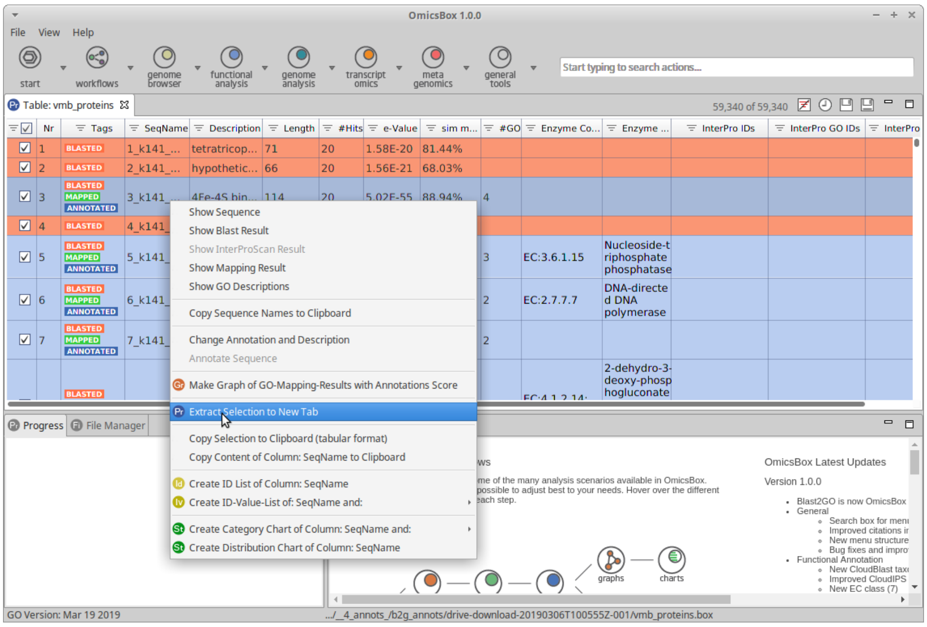

OmicsBox 1.0.0
Table: vmb_proteins
Show Sequence
Show Blast Result
Show InterProScan Result
Show Mapping Result
Show GO Descriptions
Copy Sequence Names to Clipboard
Change Annotation and Description
Annotate Sequence
Make Graph of GO-Mapping-Results with Annotations Score
Extract Selection to New Tab
Copy Selection to Clipboard (tabular format)
Copy Content of Column: SeqName to Clipboard
Create ID List of Column: SeqName
Create ID-Value-List of: SeqName and:
Create Category Chart of Column: SeqName and:
Create Distribution Chart of Column: SeqName
Progress
File Manager
GO Version: Mar 19 2019

Figure 4: Extract Selection to New Tab

A new project will be created of the selected sequences.

8 Module Genome Analysis

Content of this section:

The OmicsBox Genome Analysis module allows to characterize and analyze newly sequenced genomes, from raw reads to gene structures in an efficient and user-friendly way.

- **Quality Control and Assessment:** Use **FastQC** and **Trimmomatic** to perform the quality control of your samples, to filter reads and to remove low-quality bases.
- ***De novo* Assembly**: The assembly feature based on **ABySS** allows reconstructing whole genome sequences without a reference genome or specific hardware requirements.
- **Repeat Masking:** Mask repeats and low complexity DNA sequences of your eukaryotic genome assemblies with **RepeatMasker** to improve downstream gene predictions.
- **Gene Finding:** Perform prokaryotic (**Glimmer**) and eukaryotic (**Augustus**) gene predictions to characterize genome structure. The eukaryotic gene prediction offers RNA-seq intron hint support.

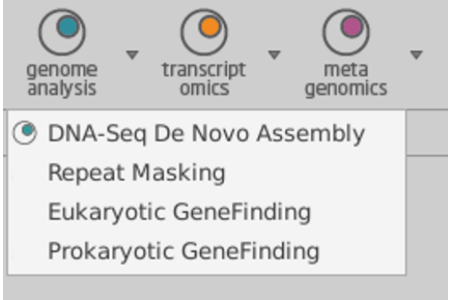


Figure 1: Genome Analysis menu

Genome Analysis use case: https://www.biobam.com/genome-assembly-annotation-sarocladium-oryzae/.

Genome Analysis Example Dataset: Download[50].

[50] https://resources.biobam.com/omicsbox/example_data/GenomeAnalysis.zip

8.1 DNA-Seq de Novo Assembly

Content of this page:

8.1.1 Introduction

Genome assembly refers to the process of taking a large number of short DNA reads and putting them back together to create a representation of the whole genome from which the DNA originates. *De novo* genome assemblies assume no prior knowledge of the source DNA sequence length, layout or composition (i.e. no reference genome is available). The goal of an assembler is to produce long contiguous pieces of sequences (contigs) from DNA-seq reads. The contigs are then joined together to form scaffolds where possible. Short-insert paired reads provide increased information for maximizing sequencing coverage, while long-insert mate paired-end reads can pair sequence fragments across greater distances. This is especially helpful to cover highly repetitive regions.

This functionality can be found under **genome analysis → DNA-Seq De novo Assembly.**

Two assembly strategies are available:

- **ABySS:** ABySS (Assembly By Short Sequences) is a *de novo*, parallel, paired-end sequence assembler that is designed for short reads. It implements algorithms that employ a Bloom filter, a probabilistic data structure, to represent a de Bruijn graph. ABySS is capable of assembling large genomes.
- **SPAdes:** SPAdes (St Petersburg genome assembler) is an assembly toolkit containing various assembly pipelines based on the Bruijn Graph. SPAdes works with Illumina and IonTorrent data and is capable of providing hybrid assemblies using PacBio, Oxford Nanopore and Sanger reads. SPAdes is designed for small genomes, and allows to assemble single-cell MDA data as well as standard isolates.

8.1.2 ABySS

ABySS 2.0[51] is a multistage *de novo* assembly pipeline consisting of unitig, contig, and scaffold stages.

[51] http://www.bcgsc.ca/platform/bioinfo/software/abyss

- At the **unitig stage**, the program performs the initial assembly of sequences according to the De Bruijn graph assembly algorithm. The unitig stage loads the full set of k-mers from the input sequencing reads into a hash table and stores auxiliary data for each k-mer such as the number of k-mer occurrences in the reads and the presence/absence of possible neighbor k-mers in the De Bruijn graph.
- At the **contig stage**, the paired-end reads are aligned to the unitigs and the pairing information is used to orient and merge overlapping unitigs.
- At the **scaffold stage**, the mate-pair reads are aligned to the contigs to orient and join them into scaffolds, inserting runs of "N" characters at gaps in coverage and for unresolved repeats.

The main innovation of ABySS 2.0 is a Bloom filter-based implementation of the unitig assembly stage. It reduces the overall memory requirements, enabling assembly of large genomes. A Bloom filter is a compact data structure for representing a set of elements that supports operations of inserting elements and querying the presence of elements. The Bloom filter data structure consists of a bit vector and one or more hash functions, where the hash functions map each k-mer to a corresponding set of positions within the bit vector (bit signature for the k-mer).

During unitig assembly, two passes are made through the input sequencing reads:

1. In the first pass, k-mers are extracted from the reads and are loaded into a Bloom filter. The program discards all k-mers with an occurrence count below a user-specified threshold (typically in the range of two to four). In this way, k-mers caused by sequencing errors are filtered out. The retained k-mers are known as solid k-mers.
2. In the second pass, the program identifies reads that consist entirely of solid k-mers, and extend them left and right within the De Bruijn graph to create unitigs.

Please cite ABySS 2.0 as:

Jackman SD, Vandervalk BP, Mohamadi H, et al (2017). "ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter". Genome Res. 2017;27(5):768-777.[52]

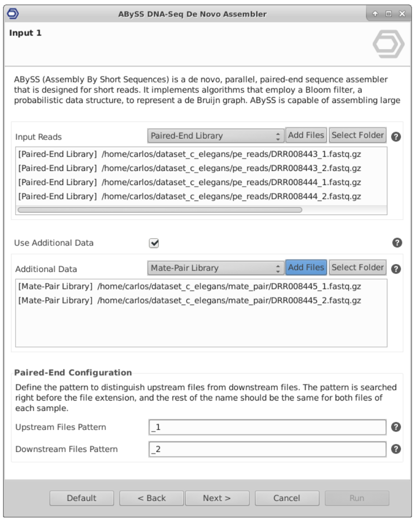
8.1.2.1 Input Data

- **Input Reads:** First, choose the type of sequencing data. Then, select the files of this type of data for the assembly. Both paired-end and single-end short reads can be provided, and both types of data can be combined in the same run.
- **Additional Data:** ABySS supports additional data types as supplementary information:
  - Additional Paired-end Libraries: Paired-end libraries that will be used only for merging unitigs into contigs and will not contribute toward the consensus sequence.
  - Mate-pair Libraries: Mate-Pair libraries that will be used for scaffolding. Mate-Pair libraries that will be used for scaffolding. Mate-pair libraries do not contribute toward the consensus sequence.
  - Linked Reads: Linked reads from 10x Genomics Chromium. The linked reads are used to correct assembly errors and scaffolding.
  - Long Sequences Libraries: Provide long sequence libraries (such as RNA-Seq contigs) that will be used for rescaffolding. Long sequence libraries do not contribute toward the consensus sequence.
- **Paired-end Configuration:** If paired-end reads are provided, a pattern to distinguish upstream files from downstream files is required. The provided patterns are searched in the filenames right before the extension. The beginning of the filenames should be the same for both files of each sample.
  - **Upstream Files Pattern**: Establish the pattern to recognize upstream FASTQ files.

[52] https://www.ncbi.nlm.nih.gov/pubmed/28232478

- **Downstream Files Pattern**: Establish the pattern to recognize downstream FASTQ files.

> ⚠ For example, if the upstream file is SRR037717_1.fastq and the downstream SRR037717_2.fastq, "_1" should be established as the upstream pattern and "_2" as the downstream pattern.



ABySS (Assembly By Short Sequences) is a de novo, parallel, paired-end sequence assembler that is designed for short reads. It implements algorithms that employ a Bloom filter, a probabilistic data structure, to represent a de Bruijn graph. ABySS is capable of assembling large

Input Reads

```
[Paired-End Library]  /home/carlos/dataset_c_elegans/pe_reads/DRR008443_1.fastq.gz
[Paired-End Library]  /home/carlos/dataset_c_elegans/pe_reads/DRR008443_2.fastq.gz
[Paired-End Library]  /home/carlos/dataset_c_elegans/pe_reads/DRR008444_1.fastq.gz
[Paired-End Library]  /home/carlos/dataset_c_elegans/pe_reads/DRR008444_2.fastq.gz
```

Use Additional Data ☑

Additional Data

```
[Mate-Pair Library]  /home/carlos/dataset_c_elegans/mate_pair/DRR008445_1.fastq.gz
[Mate-Pair Library]  /home/carlos/dataset_c_elegans/mate_pair/DRR008445_2.fastq.gz
```

Paired-End Configuration

Define the pattern to distinguish upstream files from downstream files. The pattern is searched right before the file extension, and the rest of the name should be the same for both files of each sample.

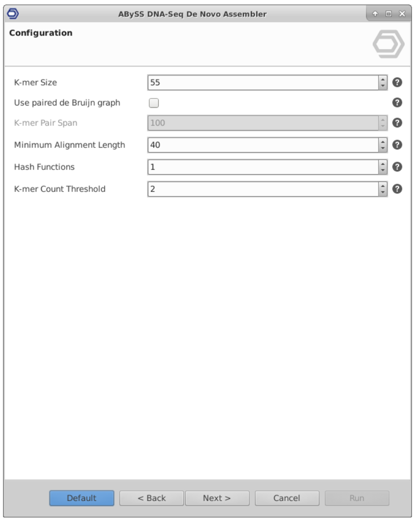
Upstream Files Pattern: _1

Downstream Files Pattern: _2

Figure 1: Input Data Page

8.1.2.2 Configuration

- **K-mer Size:** The term k-mer refers to all possible subsequences of the given length that are contained in a read. In sequence assembly, k-mers are used during the construction of De Bruijn graphs. The choice of the k-mer size has many different effects on the sequence assembly, it is advisable to try different values and check the results to choose the best one. It is recommended to use odd values of at least half the length of the reads.
- **Use paired De Bruijn graph:** Assembly will be performed using a paired De Bruijn graph. In this mode, k-mer pairs are used, which consist of two equal-sized k-mers separated by a fixed distance. To assemble using the paired De Bruijn graph mode, specify the k-mer pair span (distance between k-mers).
- **K-mer Pair Span:** Set the span of a k-mer pair (distance between k-mers).
- **Minimum Alignment Length:** Establish the minimum alignment length of a read (bp). This means that there must be a perfect match of the established length between each read and its target contig.
- **Hash Functions:** Set the number of Bloom filter hash functions. K-mers from each input sequencing read are loaded into the Bloom filter by computing the hash values of each k-mer sequence and setting the corresponding bit.
- **K-mer Count Threshold:** Set the k-mer count threshold for Bloom filter assembly. Optimal values are typically in the range of 2-4. K-mers with an occurrence count below the threshold will be discarded.



K-mer Size 55

Use paired de Bruijn graph □

K-mer Pair Span 100

Minimum Alignment Length 40

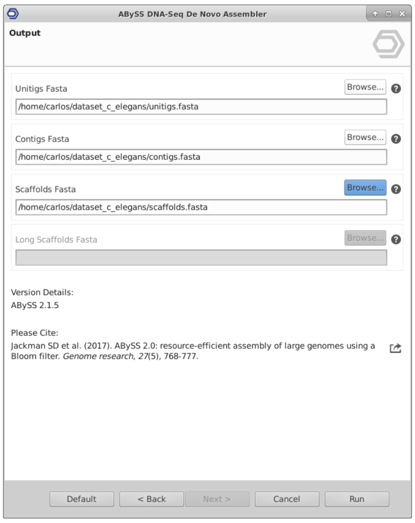
Hash Functions 1

K-mer Count Threshold 2

Figure 2: Configuration Page

8.1.2.3 Output

- **Unitigs Fasta:** Where to store the Fasta file containing the assembled unitigs.
- **Contigs Fasta:** Where to store the Fasta file containing the assembled contigs.
- **Scaffolds Fasta:** Where to store the Fasta file containing the assembled scaffolds.
- **Long Scaffolds Fasta:** Where to store the Fasta file containing the assembled long scaffolds. Note that this file is only generated if long sequence libraries were provided.



Unitigs Fasta /home/carlos/dataset_c_elegans/unitigs.fasta

Contigs Fasta /home/carlos/dataset_c_elegans/contigs.fasta

Scaffolds Fasta /home/carlos/dataset_c_elegans/scaffolds.fasta

Long Scaffolds Fasta

Version Details:
ABySS 2.1.5

Please Cite:
Jackman SD et al. (2017). ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter. *Genome research*, *27*(5), 768-777.

Figure 3: Output Data Page

8.1.2.4 Results

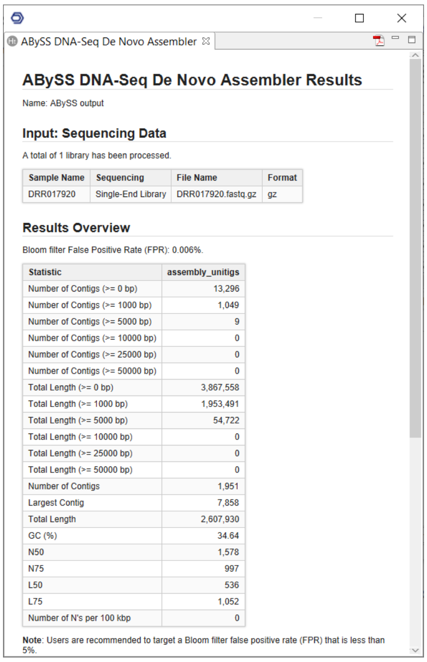
ABySS returns the assembled sequences in three FASTA files (four if long sequence libraries were provided). Each one corresponds to a different stage of the assembly procedure:

- **Unitigs**: Contains sequences assembled without using paired-end information. In case you provide only single-end data, this will be the only result file, since pairing information is required to assemble contigs.
- **Contigs**: Contains sequences assembled with paired information, scaffolding over sequencing coverage gaps, but no repeats.
- **Scaffolds**: Contains sequences assembled with paired information, scaffolding over sequencing coverage gaps and repeats.
- **Long Scaffolds:** Contains sequences that were obtained by rescaffolding using long sequences libraries.

In addition to the resulting FASTA files, a report and a chart are generated. The report shows a summary of the DNA-Seq *De Novo* Assembly results (Figure 4). This page contains information about the input sequencing data and a results overview. The Results Overview table shows a number of common statistics used to describe the quality of a sequence assembly:

- **N50:** This statistic defines the assembly quality in terms of contiguity. N50 is calculated by first ordering every unitig, contig or scaffold from longest to shortest. Next, starting from the longest sequence, the lengths of each sequence are summed up, until this running sum equals one-half of the total length of all sequences in the assembly. The N50 of the assembly is the length of the shortest contig in this list. Higher values of N50 indicate a better assembly. Note that any Nx statistic is calculated in the same way, e.g. N75 is calculated summing up all the lengths until the sum equals 75% of the total length.
- **L50:** Defined as the smallest number of contigs whose lengths sum makes up half of the total assembly length.
- **Bloom filter False Positive Rate (FPR):** The Bloom filter can generate *false positives* when the bit signatures of different k-mers overlap by chance. This means that a certain fraction of k-mer queries will return true even though the k-mers do not exist in the input sequencing data. Users are recommended to target a Bloom filter false positive rate (FPR) smaller than 5%. Parameters such as the k-mer size, hash functions or k-mer count threshold can influence the false positive rate.

The Nx plot (Figure 5) shows Nx values as x varies from 0 to 100 %. The Nx values are displayed for unitigs, contigs and scaffolds.



ABySS DNA-Seq De Novo Assembler Results

Name: ABySS output

Input: Sequencing Data

A total of 1 library has been processed.

Sample Name	Sequencing	File Name	Format
DRR017920	Single-End Library	DRR017920.fastq.gz	gz

Results Overview

Bloom filter False Positive Rate (FPR): 0.006%.

Statistic	assembly_unitigs
Number of Contigs (>= 0 bp)	13,296
Number of Contigs (>= 1000 bp)	1,049
Number of Contigs (>= 5000 bp)	9
Number of Contigs (>= 10000 bp)	0
Number of Contigs (>= 25000 bp)	0
Number of Contigs (>= 50000 bp)	0
Total Length (>= 0 bp)	3,867,558
Total Length (>= 1000 bp)	1,953,491
Total Length (>= 5000 bp)	54,722
Total Length (>= 10000 bp)	0
Total Length (>= 25000 bp)	0
Total Length (>= 50000 bp)	0
Number of Contigs	1,951
Largest Contig	7,858
Total Length	2,607,930
GC (%)	34.64
N50	1,578
N75	997
L50	536
L75	1,052
Number of N's per 100 kbp	0

Note: Users are recommended to target a Bloom filter false positive rate (FPR) that is less than 5%.

Figure 4: Summary Report

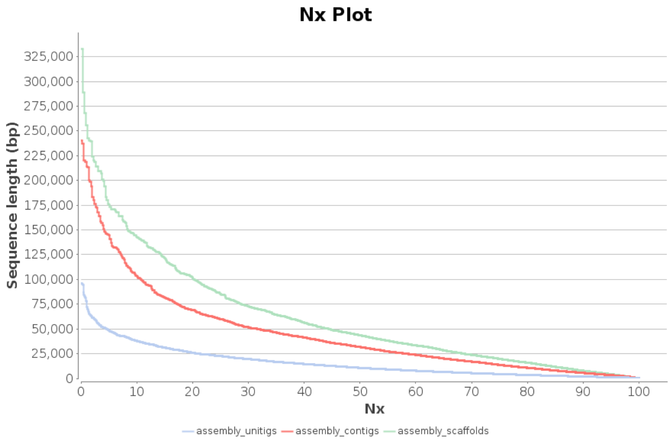


Figure 5: Nx Plot

8.1.3 SPAdes

SPAdes[53] is a *de novo* genome assembly pipeline that can deal with data coming from several sequencing technologies and supports hybrid and single-cell assemblies. The SPAdes assembly pipeline consists of four stages:

1. Assembly graph construction. SPAdes uses the *multisized de Bruijn graph*, implements new bulge/tip removal algorithms, detects and removes chimeric reads, aggregates biread information into distance histograms, and allows to backtrack the performed graph operations.
2. *k-bimer* adjustment: SPAdes derives accurate distance estimates between k-mers in the genome using joint analysis of distance histograms and paths in the assembly graph.
3. Constructs the paired assembly graph: Inspired by *Paired de Bruijn graphs* (PDBG) approach.
4. Contig construction: SPAde constructs DNA sequences of contigs and the mapping of reads to contigs by backtracking graph simplifications.

SPAdes uses a modification of Hammer[54] for error correction and quality trimming prior assembly.

In general, SPAdes uses two techniques for scaffolding.

- SPAdes tries to estimate the size of the gap separating contigs using read pairs.
- SPAdes, using the assembly graph, joins contigs that are separated by a complex tandem repeat, that cannot be resolved exactly, with a fixed gap size of 100 bp.

Contigs produced by SPAdes do not contain N symbols.

Please, cite SPAdes as:

- Nurk, Bankevich et al., 2013[55].
- Bankevich, Nurk et al., 2012[56].
- Antipov et al., 2015[57] (in case you perform hybrid assembly using PacBio or Nanopore reads).

[53] http://cab.spbu.ru/software/spades/
[54] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3117386/
[55] http://link.springer.com/chapter/10.1007%2F978-3-642-37195-0_13
[56] http://online.liebertpub.com/doi/abs/10.1089/cmb.2012.0021
[57] http://bioinformatics.oxfordjournals.org/content/early/2015/11/20/bioinformatics.btv688.short

- Prjibelski et al., 2014[58] (if you use multiple paired-end and/or mate-pair libraries).
- Vasilinetc et al., 2015[59] (if you use multiple paired-end and/or mate-pair libraries).

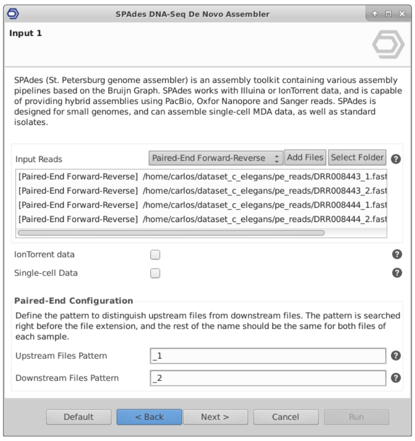
8.1.3.1 Input Data

- **Input Reads:** Select the files containing the sequencing libraries (reads). The assembly strategy requires at least one of these types of sequencing libraries.
    - Illumina single-end, paired-end, or high-quality mate-pairs.
    - IonTorrent single-end, paired-end, or high-quality mate-pairs.
    - PacBio CCS reads (should be provided as single-end data).

> ⚠ These files are assumed to be in FASTQ format. For IonTorrent data, SPAdes supports unpaired reads in unmapped BAM format.

- **IonTorrent Data:** This option is required when assembling IonTorrent data. Illumina and IonTorrent libraries should not be assembled together. For IonTorrent data, SPAdes also supports unpaired reads in unmapped BAM format (like the one produced by the Torrent Server).
- **Single-cell Data:** This option is required for Multiple Displacement Amplification (MDA) single-cell data assembly.
- **Paired-end Configuration:** If paired-end reads are provided, a pattern to distinguish upstream files from downstream files is required. The provided patterns are searched in the filenames right before the extension. The beginning of the filenames should be the same for both files of each sample.
    - **Upstream Files Pattern**: Establish the pattern to recognize upstream FASTQ files.
    - **Downstream Files Pattern**: Establish the pattern to recognize downstream FASTQ files.

> ⚠ For example, if the upstream file is SRR037717_1.fastq and the downstream SRR037717_2.fastq, "_1" should be established as the upstream pattern and "_2" as the downstream pattern.



SPAdes (St. Petersburg genome assembler) is an assembly toolkit containing various assembly pipelines based on the Bruijn Graph. SPAdes works with Illuina or IonTorrent data, and is capable of providing hybrid assemblies using PacBio, Oxfor Nanopore and Sanger reads. SPAdes is designed for small genomes, and can assemble single-cell MDA data, as well as standard isolates.

Input Reads: Paired-End Forward-Reverse

```
[Paired-End Forward-Reverse]  /home/carlos/dataset_c_elegans/pe_reads/DRR008443_1.fast
[Paired-End Forward-Reverse]  /home/carlos/dataset_c_elegans/pe_reads/DRR008443_2.fast
[Paired-End Forward-Reverse]  /home/carlos/dataset_c_elegans/pe_reads/DRR008444_1.fast
[Paired-End Forward-Reverse]  /home/carlos/dataset_c_elegans/pe_reads/DRR008444_2.fast
```

IonTorrent data ☐

Single-cell Data ☐

Paired-End Configuration

Define the pattern to distinguish upstream files from downstream files. The pattern is searched right before the file extension, and the rest of the name should be the same for both files of each sample.

Upstream Files Pattern: _1

Downstream Files Pattern: _2

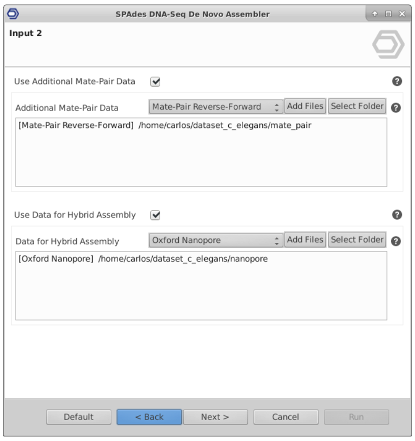
Figure 6: Input Data Page

- **Use Additional Mate-Pair Data:** SPAdes supports mate-pair only assembly. However, high-quality mate-pair libraries are recommended in these cases. Here, regular mate-pair libraries can be provided as supplementary information. Upstream and downstream files will be distinguished using the pattern established in the previous page (Paired-end Configuration).

[58] http://bioinformatics.oxfordjournals.org/content/30/12/i293.short
[59] http://bioinformatics.oxfordjournals.org/content/31/20/3262.abstract

- **Use Data for Hybrid Assembly:**
  - PacBio (CLR), Oxford Nanopore and Sanger reads can be provided for hybrid assemblies (e.g. with Illumina or IonTorrent data). SPAdes uses this data for gap closure and repeat resolution.
  - Contigs of the same genome (trusted) generated by other assembler(s) can be specified to merge them into SPAdes assembly.
  - Less reliable contigs (untrusted) can be used only for gap closure and repeat resolution.

> ⚠ Only contigs of the same genome should be specified since SPAdes does not work with genomes of closely-related species.



Use Additional Mate-Pair Data ☑

Additional Mate-Pair Data: Mate-Pair Reverse-Forward | Add Files | Select Folder

[Mate-Pair Reverse-Forward] /home/carlos/dataset_c_elegans/mate_pair

Use Data for Hybrid Assembly ☑

Data for Hybrid Assembly: Oxford Nanopore | Add Files | Select Folder

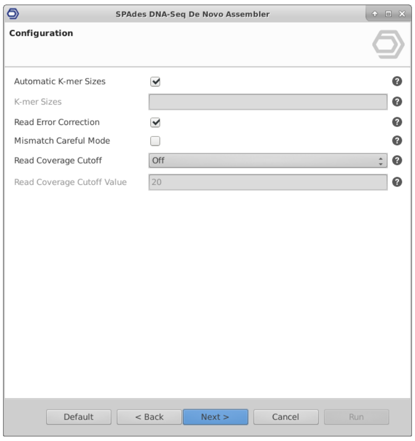
[Oxford Nanopore] /home/carlos/dataset_c_elegans/nanopore

FIgure 7: Input 2 Data Page

8.1.3.2 Configuration

- **Automatic K-mer Sizes:** K-mer sizes are selected automatically based on the read length and data set type:
  - If single-cell data is provided, the default values are 21, 33 and 55.
  - For multicell datasets, K values are automatically selected using maximum read length.
- **K-mer Sizes:** Define a comma-separated list of k-mer sizes to be used. These must be odd and less than 128. You can find recommendations about K-mer sizes in the SPAdes documentation[60].
- **Read Error Correction:** Performs a read error correction before assembly. Depending on the sequencing platform, the BayesHammer (Illumina) or the IonHammer (IonTorrent) tools are used for this task. This procedure is recommended to obtain high-quality assemblies but can be turned off if read error correction has been done previously.
- **Mismatch Careful Mode:** Tries to reduce the number of mismatches and short indels. It also runs MismactCorrector, a post-processing tool that uses BWA.
- **Read Coverage Cutoff:** Configure the read coverage cutoff value that SPAdes will use to obtain the most reliable assembled sequences. Must be a positive decimal number, or automatic, or off. When set to “Automatic” SPAdes automatically computes coverage threshold using conservative strategy.
- **Read Coverage Cutoff Value:** If the “Defined by User” option is selected above, set a positive float value.

[60] https://github.com/ablab/spades#sec3.3



Automatic K-mer Sizes ☑

K-mer Sizes

Read Error Correction ☑

Mismatch Careful Mode ☐

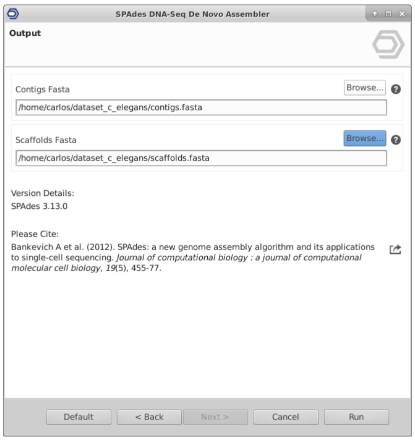
Read Coverage Cutoff: Off

Read Coverage Cutoff Value: 20

Figure 8: Configuration Page

8.1.3.3 Output

- **Contigs Fasta:** Where to store the Fasta file containing the assembled contigs.
- **Scaffolds Fasta:** Where to store the Fasta file containing the assembled scaffold. Recommended for use as resulting sequences.



Contigs Fasta: /home/carlos/dataset_c_elegans/contigs.fasta

Scaffolds Fasta: /home/carlos/dataset_c_elegans/scaffolds.fasta

Version Details:
SPAdes 3.13.0

Please Cite:
Bankevich A et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology : a journal of computational molecular cell biology*, *19*(5), 455-77.

Figure 9: Output Data Page

8.1.3.4 Results

SPAdes returns the assembled sequences in two FASTA files:

- **Contigs**: Contains resulting contigs.
- **Scaffolds**: Contains resulting scaffolds (recommended for use as resulting sequences).

Contigs/scaffolds names in SPAdes output FASTA files have the following format:

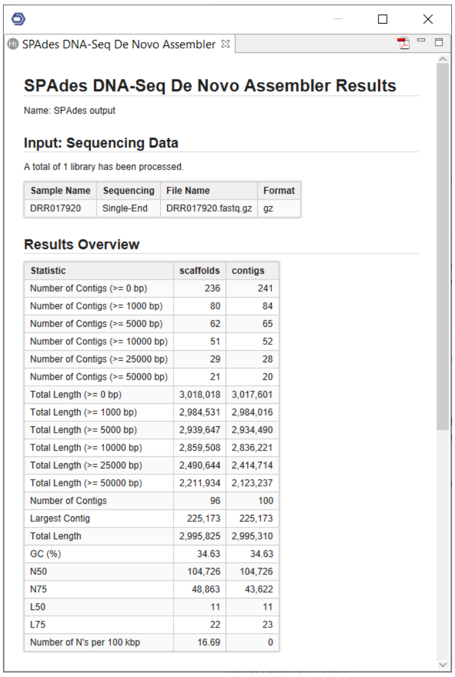
>NODE_3_length_237403_cov_243.207

- 3 is the number of the contig/scaffold.
- 237403 is the sequence length in nucleotides.
- 243.207 is the k-mer coverage for the last (largest) k value used. Note that the k-mer coverage is always lower than the read (per-base) coverage.

In addition to the resulting FASTA files, a report and a chart are generated. The report shows a summary of the DNA-Seq *De Novo* Assembly results (Figure 10). This page contains information about the input sequencing data and a results overview. The Results Overview table shows a number of common statistics used to describe the quality of a sequence assembly (see the explanation in the previous section).

- **N50:** This statistic defines the assembly quality in terms of contiguity. N50 is calculated by first ordering every contig or scaffold from longest to shortest. Next, starting from the longest sequence, the lengths of each sequence are summed up, until this running sum equals one-half of the total length of all sequences in the assembly. The N50 of the assembly is the length of the shortest contig in this list. Higher values of N50 indicate a better assembly. Note that any Nx statistic is calculated in the same way, e.g. N75 is calculated summing up all the lengths until the sum equals 75% of the total length.
- **L50:** Defined as the smallest number of contigs whose lengths sum makes up half of the total assembly length.

The Nx plot (Figure 11) shows Nx values as x varies from 0 to 100 %. The Nx values are displayed for contigs and scaffolds.



SPAdes DNA-Seq De Novo Assembler Results

Name: SPAdes output

Input: Sequencing Data

A total of 1 library has been processed.

Sample Name	Sequencing	File Name	Format
DRR017920	Single-End	DRR017920.fastq.gz	gz

Results Overview

Statistic	scaffolds	contigs
Number of Contigs (>= 0 bp)	236	241
Number of Contigs (>= 1000 bp)	80	84
Number of Contigs (>= 5000 bp)	62	65
Number of Contigs (>= 10000 bp)	51	52
Number of Contigs (>= 25000 bp)	29	28
Number of Contigs (>= 50000 bp)	21	20
Total Length (>= 0 bp)	3,018,018	3,017,601
Total Length (>= 1000 bp)	2,984,531	2,984,016
Total Length (>= 5000 bp)	2,939,647	2,934,490
Total Length (>= 10000 bp)	2,859,508	2,836,221
Total Length (>= 25000 bp)	2,490,644	2,414,714
Total Length (>= 50000 bp)	2,211,934	2,123,237
Number of Contigs	96	100
Largest Contig	225,173	225,173
Total Length	2,995,825	2,995,310
GC (%)	34.63	34.63
N50	104,726	104,726
N75	48,863	43,622
L50	11	11
L75	22	23
Number of N's per 100 kbp	16.69	0

Figure 10: Summary Report

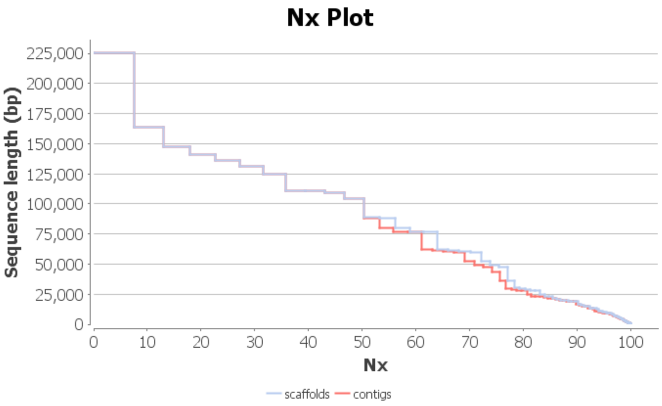


Figure 11: Nx Plot

8.2 Repeat Masking

Content of this page:

8.2.1 Introduction

Repetitive DNA sequences are abundant in a broad range of species. The term repeat is used to describe two different types of sequences: **low complexity** sequences, such as homopolymeric runs of nucleotides, and **transposable elements,** such as viruses, long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs). Eukaryotic genomes can be very repeat rich: for example, 47% of the human genome is thought to consists of repeats. Adequate repeat annotation should be a part of every genome annotation project.

Repeat identification and masking is usually a previous step to the gene prediction and annotation phase. The term 'masking' means transforming every nucleotide identified as a repeat to an 'N', 'X' or to a lower case a, t, g or c (the latter is known as soft masking). The masking step signals to downstream sequence alignment and gene prediction tools that these regions are repeats. Identifying repeats is complicated by the fact that repeats are often poorly conserved; thus, accurate repeat detection usually requires a repeat library for the species of interest. Also, the borders of these repeats are usually ill-defined; repeats often insert within other repeats, and only fragments within fragments are present, which means that complete elements are found quite rarely.

Users must carefully post-process the outputs of this process since that failure to mask genome sequences can be catastrophic. Left unmasked repeats can seed millions of spurious BLAST alignments, producing false evidence for gene annotation. Worse still, many transposon open reading frames (ORFs) look like true host genes to gene predictors, causing portions of transposon ORFs to be added as additional exons to gene predictions, completely corrupting the final gene annotations. Good repeat masking is thus crucial for the accurate annotation of protein-coding genes.

This application is based on **RepeatMasker**[61]. RepeatMasker is a program that screens DNA sequences and detects transposable elements, satellites, and low-complexity DNA sequences. The output of the program is a detailed annotation of the repeats that are present in the query sequence as well as a modified version of the query sequence in which all the annotated repeats have been masked. RepeatMasker uses a sequence search engine to perform its search for repeats. In OmicsBox, RMBLast and HMMER are supported. RepeatMasker also uses the Tandem Repeat Finder to detect tandem repeats.

[61] http://repeatmasker.org/

RepeatMasker comes with the **Dfam Database.** The Dfam database is a open collection of DNA Transposable Element sequence alignments, hidden Markov Models (HMMs), consensus sequences, and genome annotations. Dfam represents a collection of multiple sequence alignments, each containing a set of representative members of a specific transposable element family. These alignments (seed alignments) are used to generate HMMs and consensus sequences for each family. The Dfam website[62] gives information about each family, and provides genome annotations for a collection of core genomes.The current release (Dfam 3.0) contains 6,235 TE families spanning five organisms: human, mouse, zebrafish, fruit fly, nematode, and a growing number of additional species.

To supplement these databases, OmicsBox allows providing custom libraries, as well as the RepeatMasker edition of **RepBase**. RepBase is a database of representative repetitive sequences from eukaryotic species. Users can download the RepeatMasker library file from the Genetic Information Research Institute (GIRI)[63] web site after requesting an account opening.

Please cite RepeatMasker as:

Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-4.0. 2013-2015 <http://www.repeatmasker.org>.

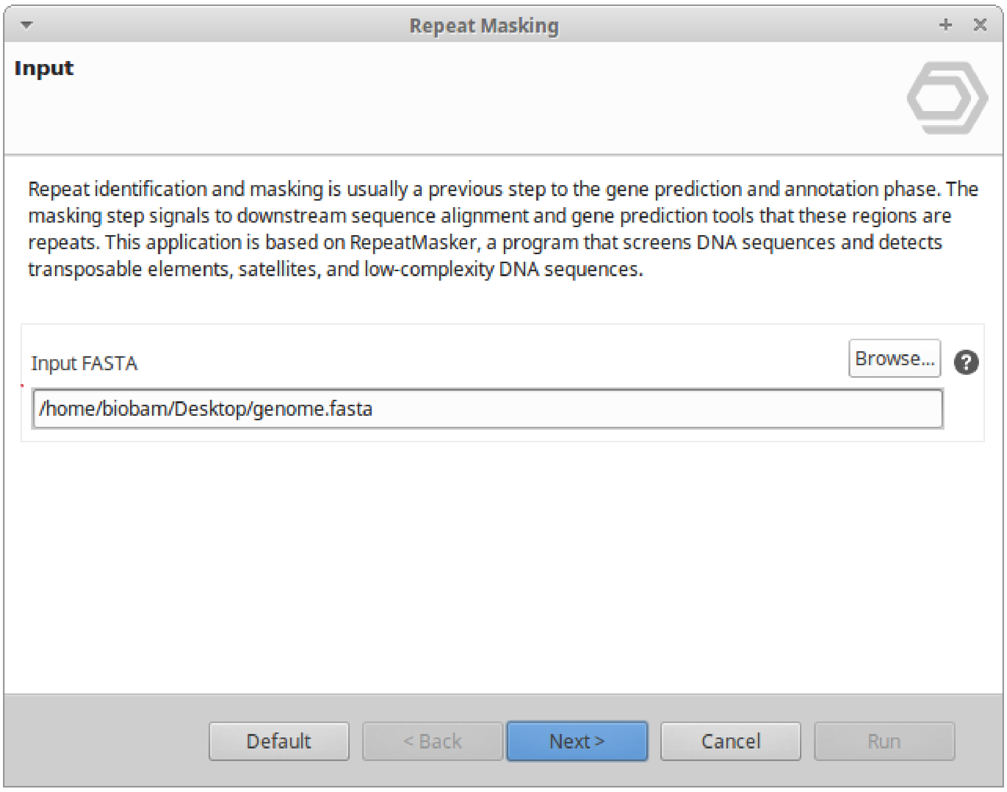
8.2.2 Run Repeat Masking

This functionality can be found under **Genome Analysis** → **Repeat Masking.** The wizard allows to select input files and adjust analysis parameters (Figures 1 to 4).

8.2.2.1 Input

- **Input FASTA:** Select the file that contains the DNA sequences to be masked. Input sequences must be in FASTA format.

[62] https://www.dfam.org/home
[63] https://www.girinst.org/



Repeat identification and masking is usually a previous step to the gene prediction and annotation phase. The masking step signals to downstream sequence alignment and gene prediction tools that these regions are repeats. This application is based on RepeatMasker, a program that screens DNA sequences and detects transposable elements, satellites, and low-complexity DNA sequences.

Input FASTA: /home/biobam/Desktop/genome.fasta

Figure 1: Input Page

8.2.2.2 Basic Configuration

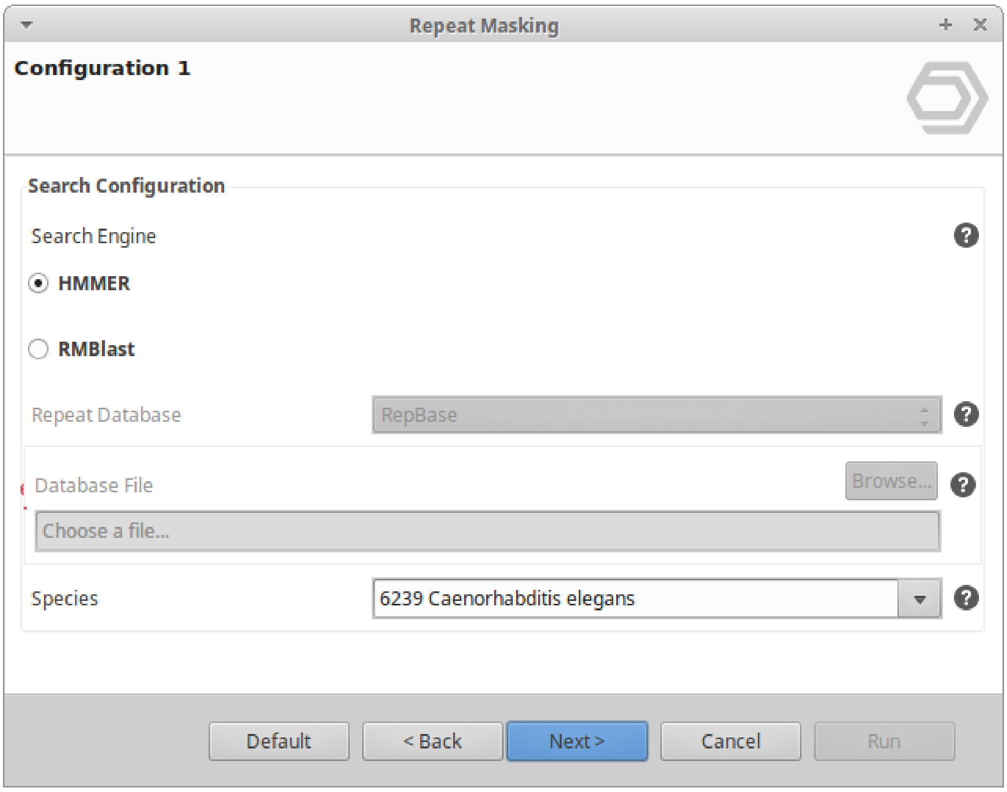
- **Search Configuration:** Select the search engine to perform the search for repeats.
  - **RMBlast:** Is a RepeatMasker compatible version of the NCBI Blast tool suite.
  - **HMMER**: It uses the *nhmmer* program to search sequences against the Dfam database.
- **Repeat Database:** RepeatMasker works with these databases:
  - **Dfam:** It is a database of transposable elements included in the application, so it is not necessary to provide any additional file.
  - **Custom:** Allows providing a custom library of sequences to be masked in the query. The library file needs to contain repetitive elements in FASTA format. The recommended format for IDs in a custom library is ">repeatname#class/subclass".
  - **RepBase:** We highly recommend obtaining the RepeatMasker edition of RepBase. Searches are optimized to use this database and can produce higher quality annotations. To obtain RepBase RepeatMasker edition go to the Genetic Information Institute website[64]. This option expects an EMBL file as a database file.

This functionality is compatible with the RepBase RepeatMasker edition 20181026 and 20170127. Make sure you are providing the proper database.

- **Database FIle:** If it is necessary, select the file containing the database to perform the search.

[64] http://www.girinst.org

  - **Custom:** The library file needs to contain sequences in FASTA format. The recommended format for IDs in a custom library is ">repeatname#class/subclass".
  - **Repbase:** EMBL file downloaded from the Genetic Information Institute website[65].
- **Species:** Specify the species or clade of the input sequence. The species name must be a valid NCBI Taxonomy Database species name and be contained in the RepeatMasker repeat database. Take into account that if HMMER is selected as search engine, the Dfam database only contains information about human, mouse, zebrafish, fruit fly, and nematode.



Repeat Masking

Configuration 1

Search Configuration

Search Engine

● HMMER

○ RMBlast

Repeat Database: RepBase

Database File: Choose a file...

Species: 6239 Caenorhabditis elegans

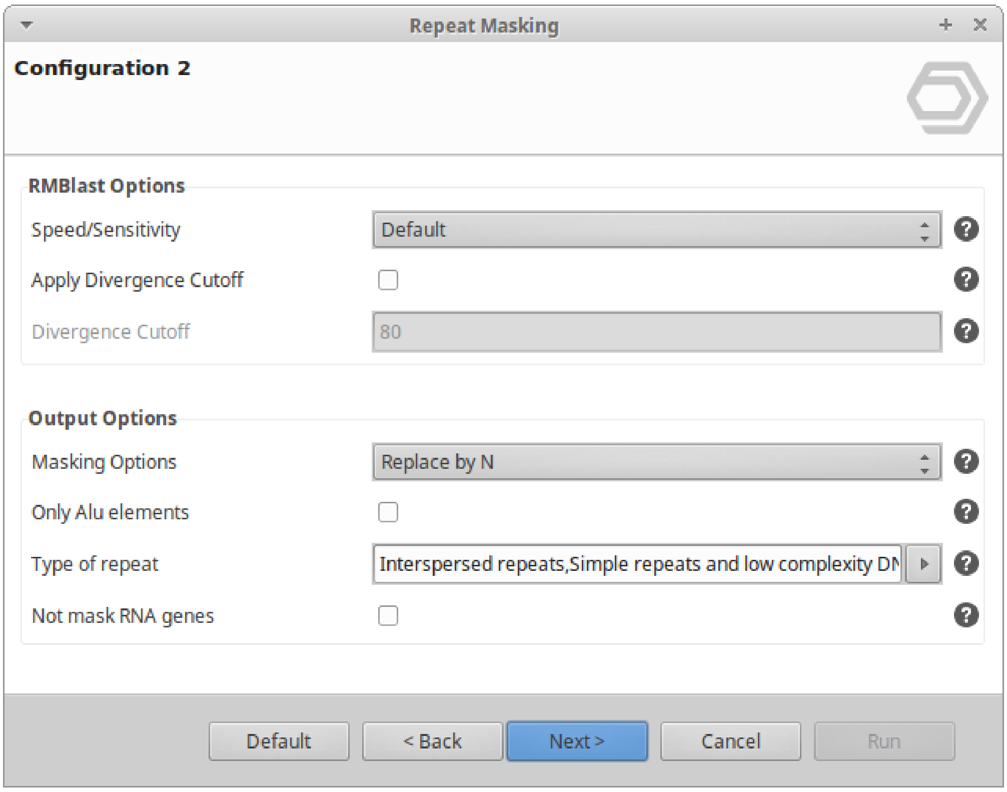
Figure 2: Basic Configuration Data Page

8.2.2.3 Advanced Configuration

- **RMBlast Options. Speed/Sensitivity:** Select the sensitivity of the search. The more sensitive the longer the processing time:
  - **Rush:** About 10% less sensitive and 4-10 times faster than the default option (quick searches are fine under most circumstances).
  - **Quick:** 5-10% less sensitive, 2-5 times faster than default.
  - **Slow:** 0-5% more sensitive, 2-3 times slower than default.
- **RMBlast Options. Apply Divergence Cutoff:** This option masks only those repeats that are less divergent from the consensus than a specific percentage.
- **Masking Options:** Select how sequences will be masked. Repetitive elements can be replaced by N, by X, or by lower case. Note that some downstream applications require a specific type of masking.
- **Only Alu elements:** Only masks Alus and 7SLRNA, SVA and LTR5. This option only works for primate DNA.

[65] http://www.girinst.org

- **Type of repeat:** Select the type of repeats that the algorithm will detect and mask: Interspersed repeats, simple repeats, and low complexity DNA, or both.
- **Not mask RNA genes:** RepeatMasker by default screens for matches to small pol III transcribed RNAs (mostly tRNAs and snRNAs) due to their close similarity to SINEs and the abundance of some of their pseudogenes. Check this option if you are interested in leaving the small RNA genes sequences unmasked.



Repeat Masking

Configuration 2

RMBlast Options

Speed/Sensitivity: Default

Apply Divergence Cutoff: □

Divergence Cutoff: 80

Output Options

Masking Options: Replace by N

Only Alu elements: □

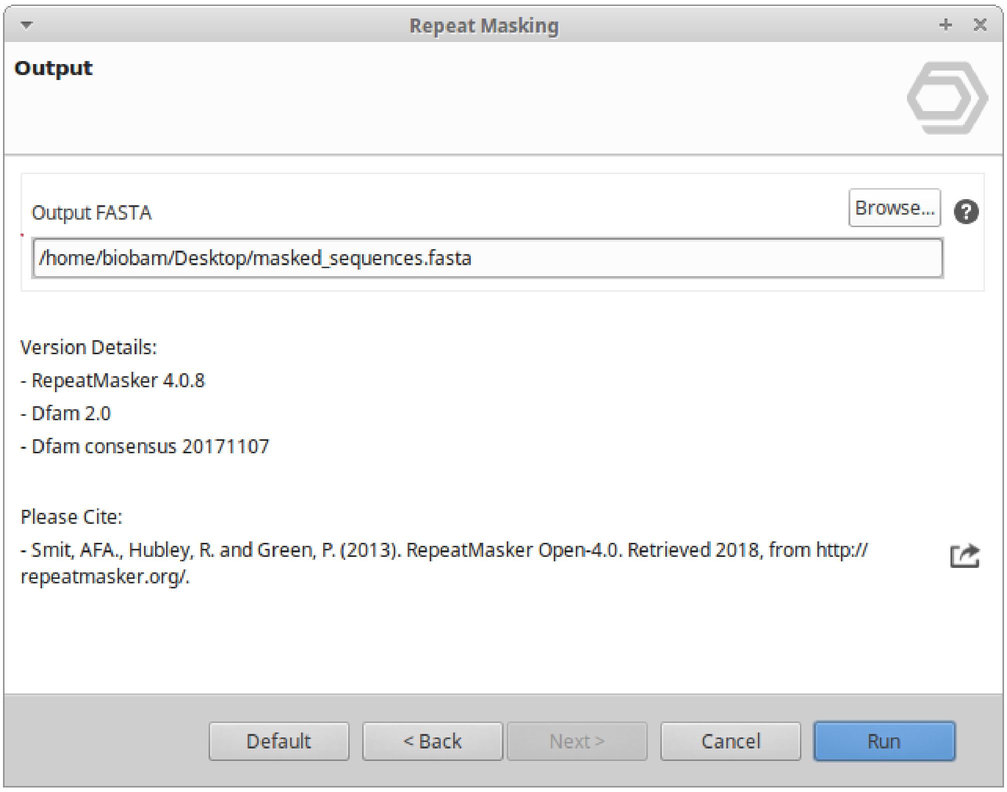
Type of repeat: Interspersed repeats,Simple repeats and low complexity DN

Not mask RNA genes: □

Figure 3: Advanced Configuration

8.2.2.4 Output

- **Output FASTA:** Select a file where the masked sequences will be placed.



Output

Output FASTA
/home/biobam/Desktop/masked_sequences.fasta

Version Details:
- RepeatMasker 4.0.8
- Dfam 2.0
- Dfam consensus 20171107

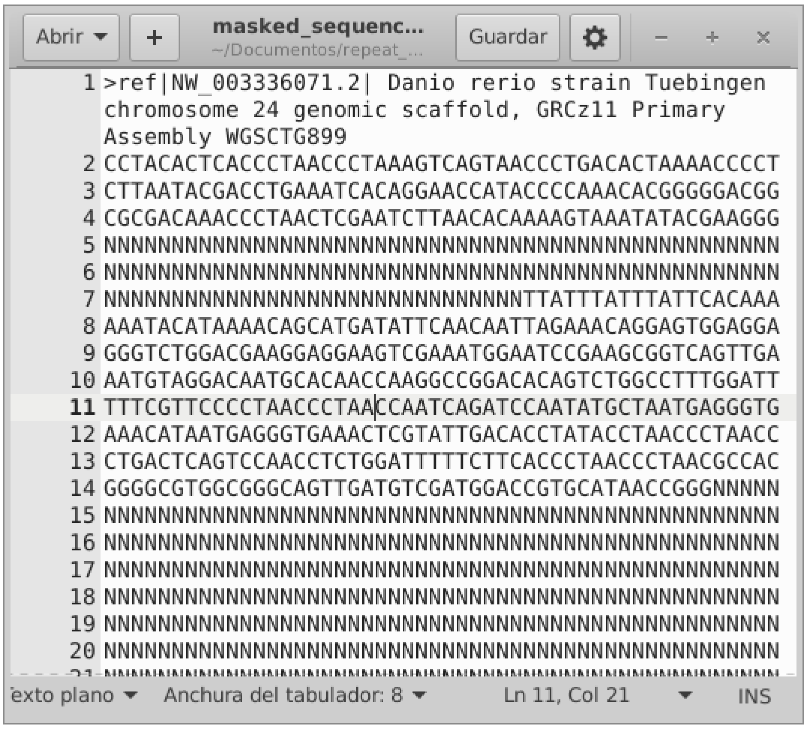
Please Cite:
- Smit, AFA., Hubley, R. and Green, P. (2013). RepeatMasker Open-4.0. Retrieved 2018, from http://repeatmasker.org/.

Figure 4: Output page

8.2.3 Results

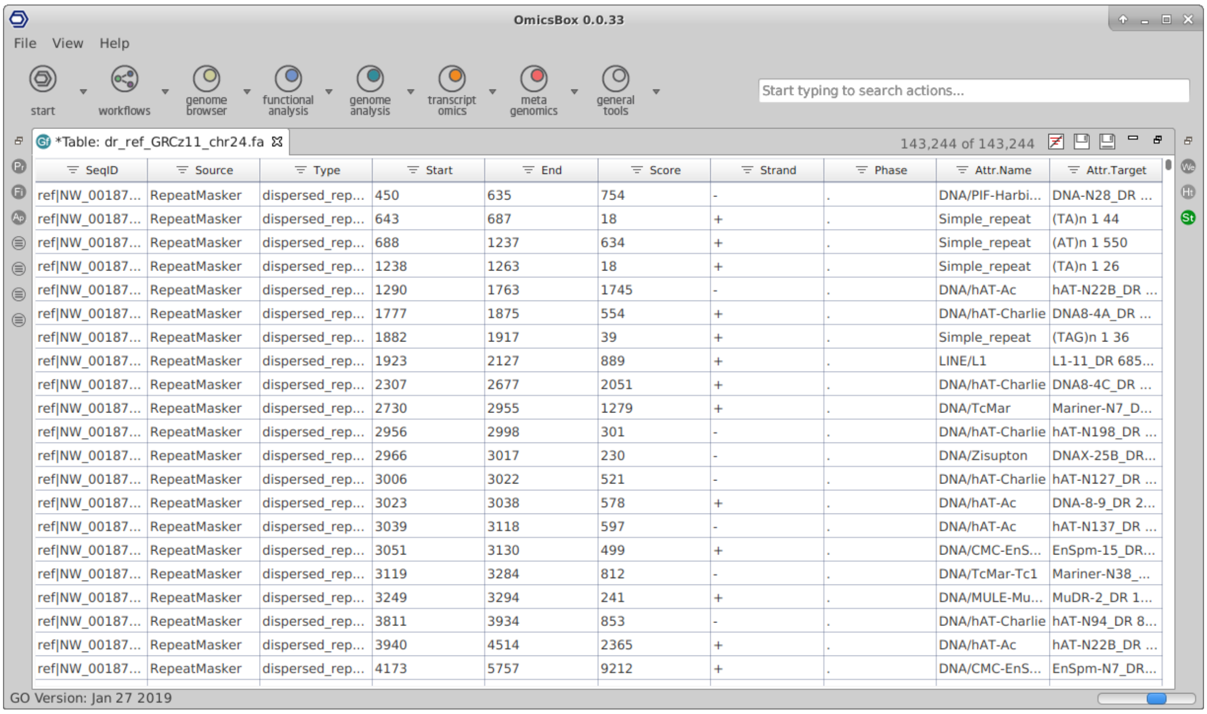
The Repeat Masking process returns the masked sequences in FASTA format and the location of the detected repeats in GFF format (Figure 5[66] and Figure 6[67]). The repeat sequences found during the procedure are replaced by X, N or lowercase (according to the selected mask option), so the output FASTA will contain the same sequences as the input FASTA but with the nucleotides corresponding to a repetitive element masked. The coordinates and strand, as well as the class and subclass of each repetitive element is annotated in the output GFF project.

[66] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=618594328#RepeatMasking-figure3
[67] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=618594328#RepeatMasking-figure4



```
>ref|NW_003336071.2| Danio rerio strain Tuebingen
chromosome 24 genomic scaffold, GRCz11 Primary
Assembly WGSCTG899
CCTACACTCACCCTAACCCTAAAGTCAGTAACCCTGACACTAAAACCCCT
CTTAATACGACCTGAAATCACAGGAACCATACCCCAAACACGGGGGACGG
CGCGACAAACCCTAACTCGAATCTTAACACAAAAGTAAATATACGAAGGG
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTTATTTATTTATTCACAAA
AAATACATAAAACAGCATGATATTCAACAATTAGAAACAGGAGTGGAGGA
GGGTCTGGACGAAGGAGGAAGTCGAAATGGAATCCGAAGCGGTCAGTTGA
AATGTAGGACAATGCACAACCAAGGCCGGACACAGTCTGGCCTTTGGATT
TTTCGTTCCCCTAACCCTAACCAATCAGATCCAATATGCTAATGAGGGTG
AAACATAATGAGGGTGAAACTCGTATTGACACCTATACCTAACCCTAACC
CTGACTCAGTCCAACCTCTGGATTTTTCTTCACCCTAACCCTAACGCCAC
GGGGCGTGGCGGGCAGTTGATGTCGATGGACCGTGCATAACCGGGNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
```

Figure 5: Masked sequences



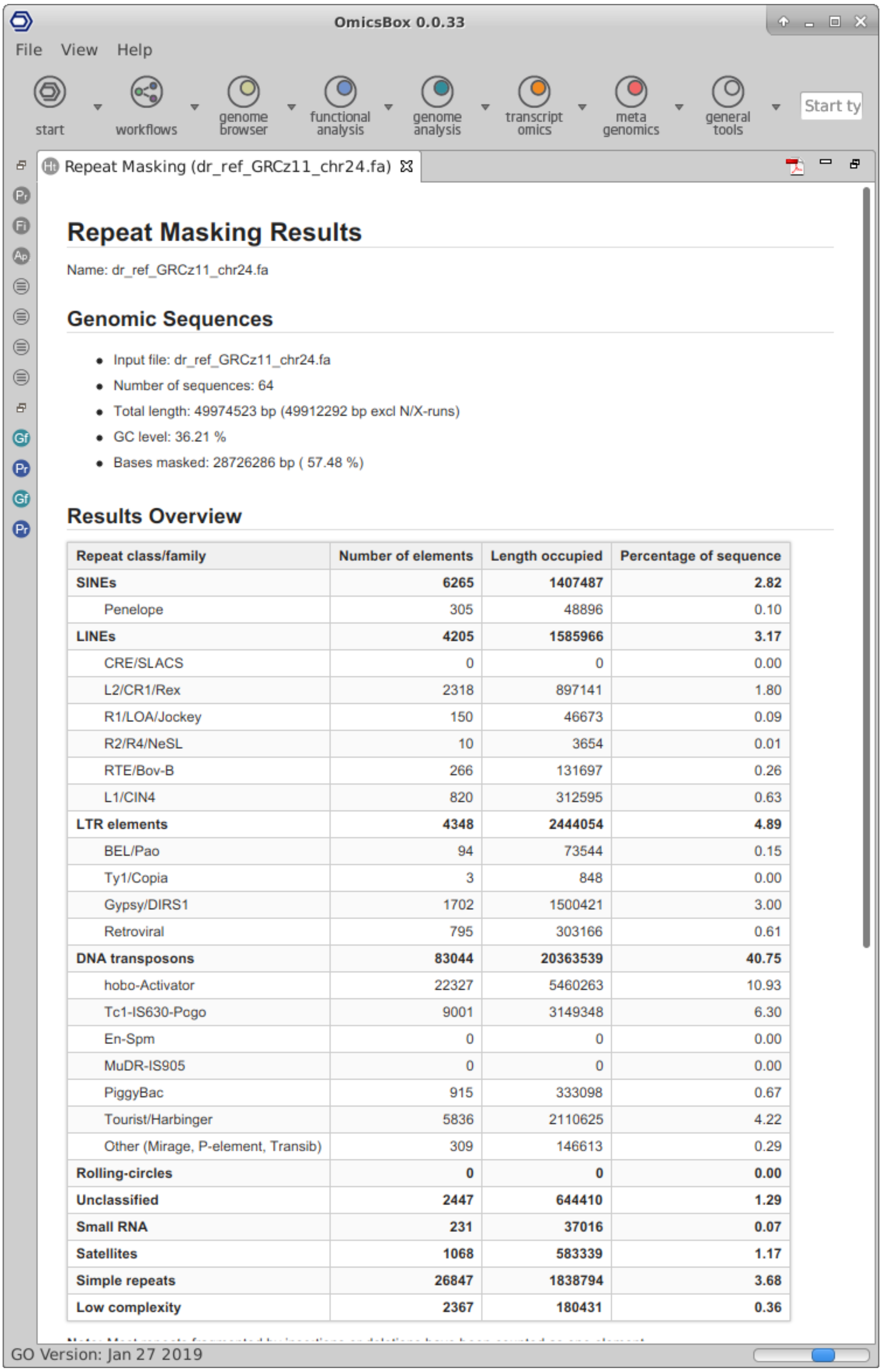
SeqID	Source	Type	Start	End	Score	Strand	Phase	Attr.Name	Attr.Target
ref|NW_00187...	RepeatMasker	dispersed_rep...	450	635	754	-	.	DNA/PIF-Harbi...	DNA-N28_DR ...
ref|NW_00187...	RepeatMasker	dispersed_rep...	643	687	18	+	.	Simple_repeat	(TA)n 1 44
ref|NW_00187...	RepeatMasker	dispersed_rep...	688	1237	634	+	.	Simple_repeat	(AT)n 1 550
ref|NW_00187...	RepeatMasker	dispersed_rep...	1238	1263	18	+	.	Simple_repeat	(TA)n 1 26
ref|NW_00187...	RepeatMasker	dispersed_rep...	1290	1763	1745	-	.	DNA/hAT-Ac	hAT-N22B_DR ...
ref|NW_00187...	RepeatMasker	dispersed_rep...	1777	1875	554	+	.	DNA/hAT-Charlie	DNA8-4A_DR ...
ref|NW_00187...	RepeatMasker	dispersed_rep...	1882	1917	39	+	.	Simple_repeat	(TAG)n 1 36
ref|NW_00187...	RepeatMasker	dispersed_rep...	1923	2127	889	+	.	LINE/L1	L1-11_DR 685...
ref|NW_00187...	RepeatMasker	dispersed_rep...	2307	2677	2051	+	.	DNA/hAT-Charlie	DNA8-4C_DR ...
ref|NW_00187...	RepeatMasker	dispersed_rep...	2730	2955	1279	+	.	DNA/TcMar	Mariner-N7_D...
ref|NW_00187...	RepeatMasker	dispersed_rep...	2956	2998	301	-	.	DNA/hAT-Charlie	hAT-N198_DR ...
ref|NW_00187...	RepeatMasker	dispersed_rep...	2966	3017	230	-	.	DNA/Zisupton	DNAX-25B_DR...
ref|NW_00187...	RepeatMasker	dispersed_rep...	3006	3022	521	-	.	DNA/hAT-Charlie	hAT-N127_DR ...
ref|NW_00187...	RepeatMasker	dispersed_rep...	3023	3038	578	+	.	DNA/hAT-Ac	DNA-8-9_DR 2...
ref|NW_00187...	RepeatMasker	dispersed_rep...	3039	3118	597	-	.	DNA/hAT-Ac	hAT-N137_DR ...
ref|NW_00187...	RepeatMasker	dispersed_rep...	3051	3130	499	+	.	DNA/CMC-EnS...	EnSpm-15_DR...
ref|NW_00187...	RepeatMasker	dispersed_rep...	3119	3284	812	-	.	DNA/TcMar-Tc1	Mariner-N38_...
ref|NW_00187...	RepeatMasker	dispersed_rep...	3249	3294	241	+	.	DNA/MULE-Mu...	MuDR-2_DR 1...
ref|NW_00187...	RepeatMasker	dispersed_rep...	3811	3934	853	-	.	DNA/hAT-Charlie	hAT-N94_DR 8...
ref|NW_00187...	RepeatMasker	dispersed_rep...	3940	4514	2365	+	.	DNA/hAT-Ac	hAT-N22B_DR ...
ref|NW_00187...	RepeatMasker	dispersed_rep...	4173	5757	9212	+	.	DNA/CMC-EnS...	EnSpm-N7_DR...

Figure 6: Output GFF with the repetitive elements coordinates.

In addition to the resulting FASTA and GFF files, a report and a chart are generated. The report shows a summary of the Repeat Masking results (Figure 7[68]). This page contains information about the input sequencing data and a results overview. The Results Overview table shows the number of elements, the length occupied and the percentage of sequence that each repeat class and subclass covers.

The Repeat Distribution chart (Figure 8[69]) shows the percentage of sequence covered by each repeat class.

[68] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=618594328#RepeatMasking-figure5
[69] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=618594328#RepeatMasking-figure6



Repeat Masking Results

Name: dr_ref_GRCz11_chr24.fa

Genomic Sequences

- Input file: dr_ref_GRCz11_chr24.fa
- Number of sequences: 64
- Total length: 49974523 bp (49912292 bp excl N/X-runs)
- GC level: 36.21 %
- Bases masked: 28726286 bp (57.48 %)

Results Overview

Repeat class/family	Number of elements	Length occupied	Percentage of sequence
SINEs	**6265**	**1407487**	**2.82**
Penelope	305	48896	0.10
LINEs	**4205**	**1585966**	**3.17**
CRE/SLACS	0	0	0.00
L2/CR1/Rex	2318	897141	1.80
R1/LOA/Jockey	150	46673	0.09
R2/R4/NeSL	10	3654	0.01
RTE/Bov-B	266	131697	0.26
L1/CIN4	820	312595	0.63
LTR elements	**4348**	**2444054**	**4.89**
BEL/Pao	94	73544	0.15
Ty1/Copia	3	848	0.00
Gypsy/DIRS1	1702	1500421	3.00
Retroviral	795	303166	0.61
DNA transposons	**83044**	**20363539**	**40.75**
hobo-Activator	22327	5460263	10.93
Tc1-IS630-Pogo	9001	3149348	6.30
En-Spm	0	0	0.00
MuDR-IS905	0	0	0.00
PiggyBac	915	333098	0.67
Tourist/Harbinger	5836	2110625	4.22
Other (Mirage, P-element, Transib)	309	146613	0.29
Rolling-circles	**0**	**0**	**0.00**
Unclassified	**2447**	**644410**	**1.29**
Small RNA	**231**	**37016**	**0.07**
Satellites	**1068**	**583339**	**1.17**
Simple repeats	**26847**	**1838794**	**3.68**
Low complexity	**2367**	**180431**	**0.36**

GO Version: Jan 27 2019

Figure 7: Summary Report

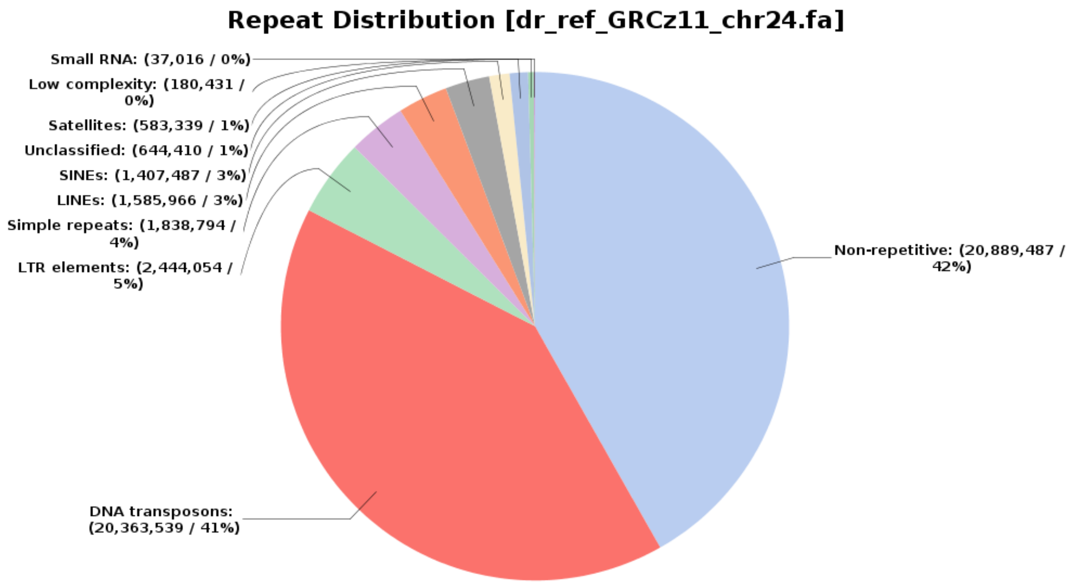


Figure 8: Repeat Distribution Pie Chart

8.3 Prokaryotic Gene Finding by Glimmer

Content of this page:

8.3.1 Introduction

Glimmer (Gene Locator and Interpolated Markov ModelER) is a system for finding genes in microbial DNA, especially the genomes of bacteria, archaea, and viruses. Glimmer uses Interpolated Markov Models (IMMs) to identify the coding regions and to distinguish them from non-coding DNA. Glimmer was the primary microbial gene finder used at The Institute for Genomic Research (TIGR), where it was first developed, and since then has been used to annotate the genomes of hundred's of bacterial and archaea species from TIGR and other labs.

The precision of Glimmer lies in its Interpolated Context Models (ICM), which are built for every query genome, by calculating and adapting the algorithm parameters to the GC content, the start and stop codons, etc.

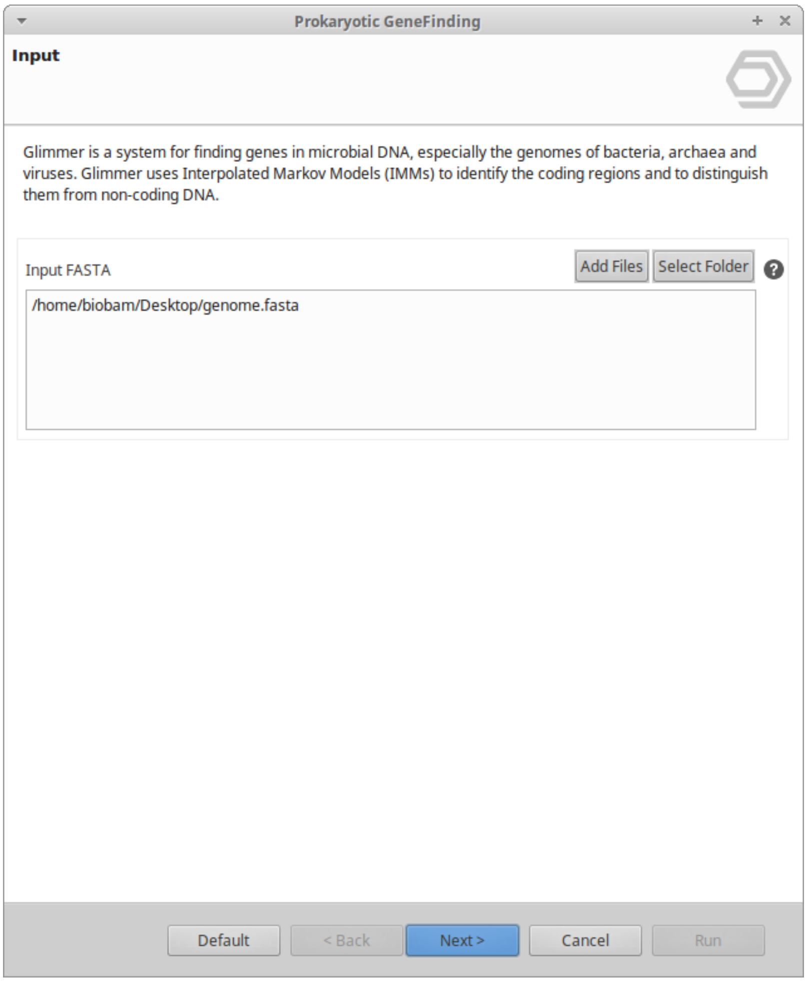
8.3.2 Run Prokaryotic Gene Finding

To create the most accurate model for your genome, this tool joins all the input fasta files and builds the model with it. Once the model is built, it performs the gene finding for each entry in the files. This methodology allows you to save the model created with all your sequences (belonging to the same organism), and use it to find the genes on a short sequence without loading the entire genome. If you are running this tool on small genomic fragments, the genome of the closest available evolutionary relative of the target organism can be used to provide a training set of genes, if no genome is available for your organism.

8.3.2.1 Input Page

The query file contains the DNA input sequence and must be in decompressed (multiple or single) FASTA format (figure 1(see page 132)). You can select a folder or multiple fasta files.

Note: Be sure to select only the fasta files containing the sequences of your query organism.



Glimmer is a system for finding genes in microbial DNA, especially the genomes of bacteria, archaea and viruses. Glimmer uses Interpolated Markov Models (IMMs) to identify the coding regions and to distinguish them from non-coding DNA.

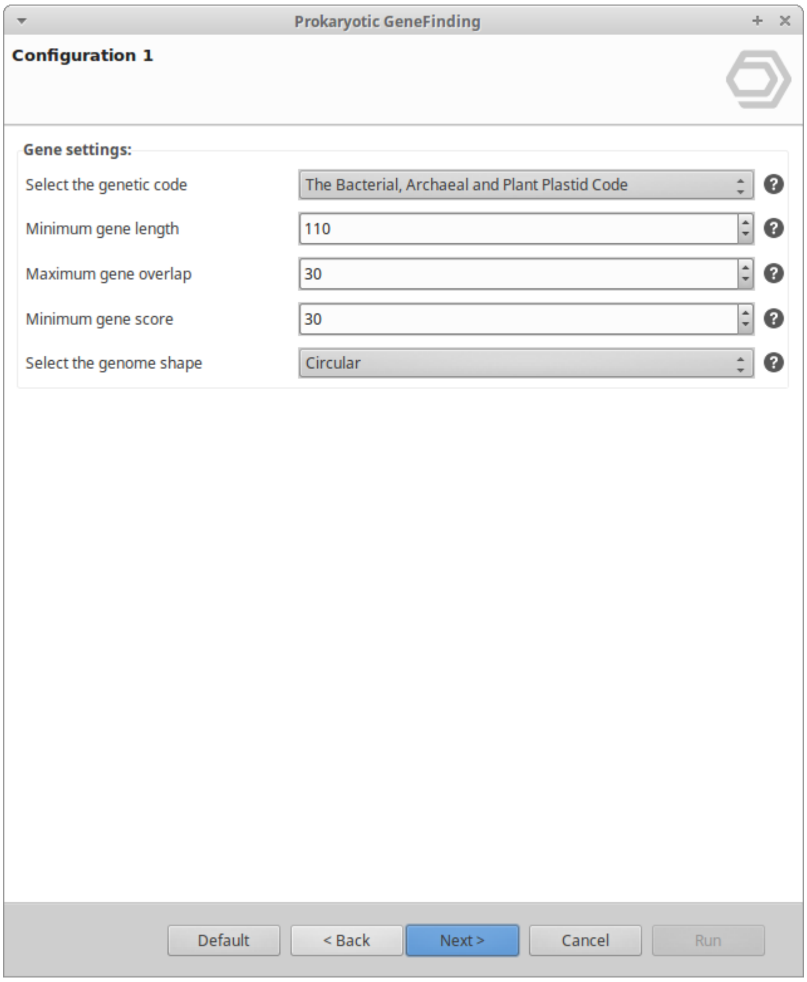
Input FASTA: /home/biobam/Desktop/genome.fasta

Figure 1: Input Data Page

8.3.2.2 Configuration 1 Page

This page groups the main settings regarding your query genome (figure 2(see page 133)).

- Genetic code: Here you can choose the genetic code for your genome. Only the 1st, 2nd, 11th, corresponding to the General genetic code, the Mycoplasma/Spiroplasma Code and the Bacterial and Archeal code are available.
- Minimum gene length: Allows setting the length threshold for the found genes in nucleotides.
- Maximum gene overlap: Here you can choose the maximum overlap length. Unlike eukaryotic genes, prokaryotic genes often have their genes overlapped.
- Minimum gene score: Every ORF found has an assigned a score depending on his length, start and stop codons. Here you can modify the limit of the score necessary to consider an ORF a gene. Lowering these values will increase the number of genes found, but will also increase prediction errors.
- Genome Shape: Here you can specify the genome shape, assuming a linear rather than circular genome, there will be no genes that `wrap around' between the beginning and end of the sequence.



Select the genetic code: The Bacterial, Archaeal and Plant Plastid Code

Minimum gene length: 110

Maximum gene overlap: 30

Minimum gene score: 30

Select the genome shape: Circular

Figure 2: Configuration 1 Page

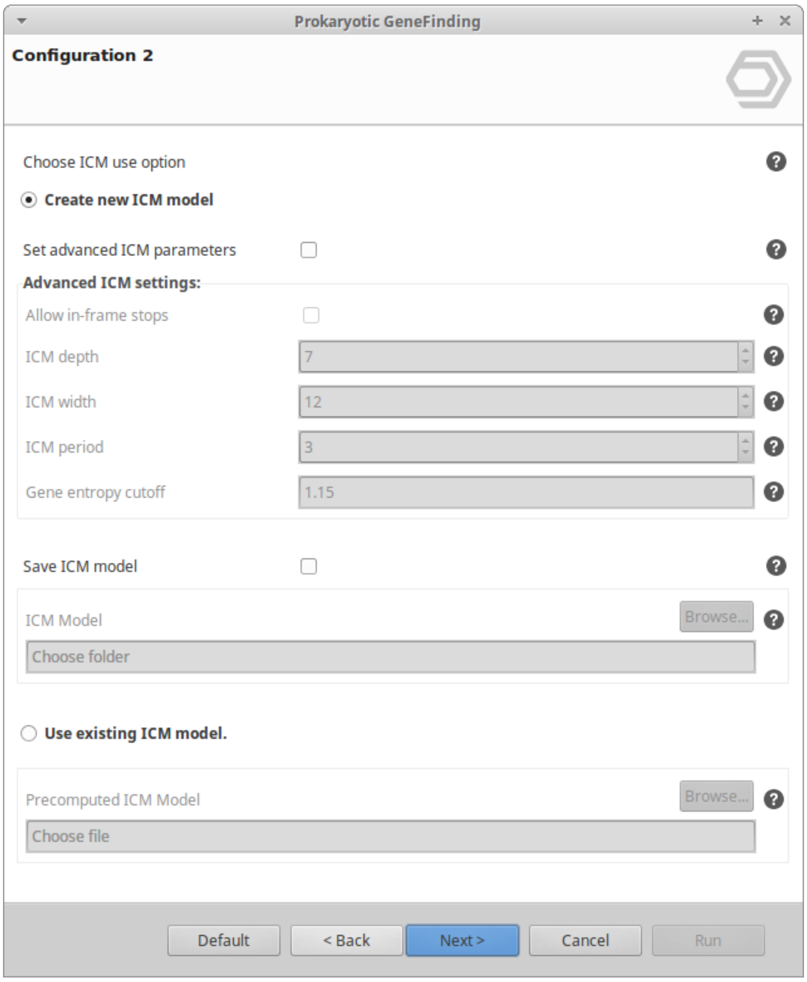
8.3.2.3 Configuration 2 Page

The second wizard page is dedicated to the Interpolated Context Model (ICM) creation parameters. The ICMs are a further extension of Interpolated Markov Models (IMMs) used to identify the coding regions and distinguish them from non-coding DNA. This step is one of the most sensitive points of the process, as it will determine the accuracy of all the following gene predictions (figure 3(see page 134)).

First, you can choose to create a new ICM or to use one created previously. If you choose to create a new ICM, you can create one with the default parameters or modify the parameters by selecting the advanced parameters checkbox:

- Allow in-frame stops: ORFs with in-frame stop codons are omitted in the building of the model Default: off
- ICM depth: The maximum number of positions in the context window that will be used to determine the probability of the predicted positions. Default: 7
- ICM Width: Set the width of the ICM to the specified number. The width includes the predicted position. Default: 12
- ICM Period: The period is the number of different submodels for different positions in the text in a cyclic pattern, i.e., if the period is 3, the first submodel will determine positions 1, 4, 7,... .; the second submodel will determine positions 2, 5, 8,... .; and the third submodel will determine positions 3, 6, 9, ... For a non-periodic model, use a value of 1. Default: 3
- Gene entropy cutoff: If this cutoff is raised, more sequences will be identified as coding, resulting in more candidate genes.
  Only genes with an entropy distance score smaller than the given value will be considered. This parameter is inspired by the observation that the coding sequences can be translated to an amino acid sequence capable of folding into a protein, whereas the non-coding sequences do not have this function. The class of amino acid sequences capable of folding to a protein has a global organizational order in contrast to those pseudo-amino-acid sequences generated from non-coding (or completely random) DNA sequences. Looking at the amino acid composition (or abundance) of a sequence we can determine the entropy of the resulting protein which allows us to cluster two kinds of sequences (coding and non-coding). Default: 1.15

If you choose to create a new ICM, you can save it by checking the option and selecting the output folder. The ICM file (.icm) can be used in posterior runs to saving computation time.



Prokaryotic GeneFinding

Configuration 2

Choose ICM use option

● Create new ICM model

Set advanced ICM parameters □

Advanced ICM settings:

Allow in-frame stops □

ICM depth 7

ICM width 12

ICM period 3

Gene entropy cutoff 1.15

Save ICM model □

ICM Model

Choose folder

○ Use existing ICM model.

Precomputed ICM Model

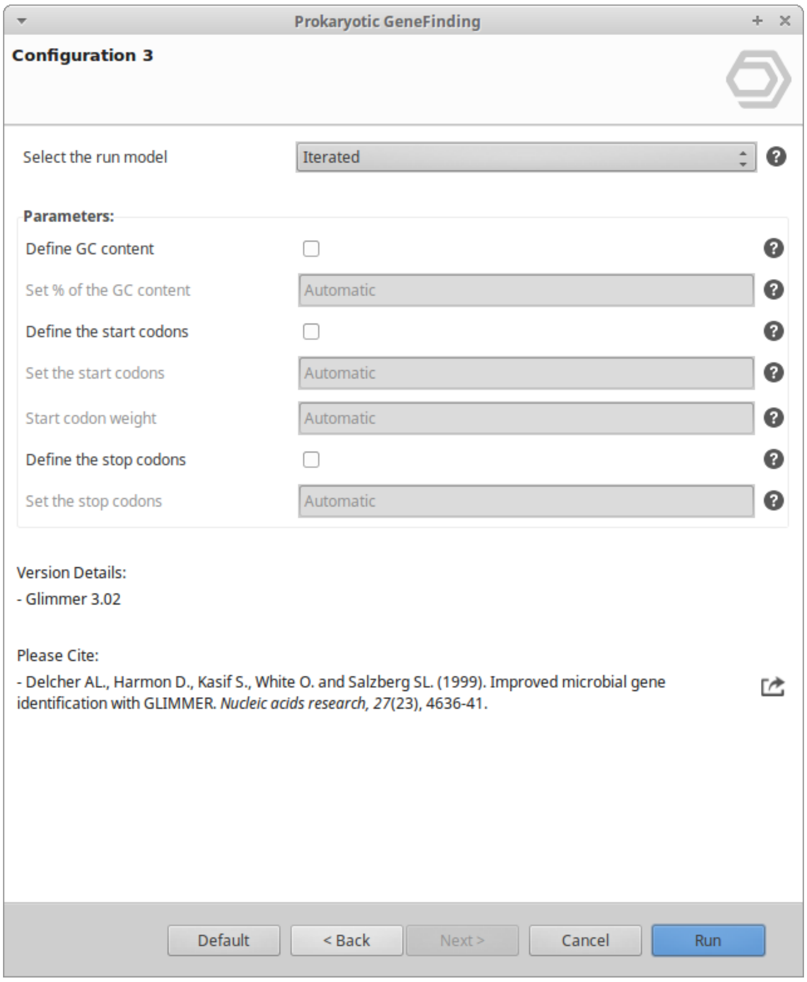
Choose file

Figure 3: Configuration 2 Page

8.3.2.4 Configuration 3 Page

This page groups the settings which pertain to the gene finding process. All of these settings are made in pairs. The first member of each pair is a checkbox allowing transition from the automatic value to the manually set value. Note: if the value is set as `Automatic', these values will be calculated automatically (figure 4(see page 135)).

- GC content: Allow the percentage of the content of G+C to be set.
- Start codons: Allow the start codons to be set as a comma-separated list. Note: If you want to use only one start codon, it's suitable to set the three start codons, and to change the weight of the desired start codon to 1 in the `start codons weight' parameter.
- Start codons weight: Specify the probability of different start codons (same number and order as in the `Start codons' parameter). If the start codons have been specified without weights, then each start codon will be assigned equal weights (which is very unusual).
- Stop codons: Allow the stop codons to be set as a comma-separated list.



Select the run model: Iterated

Parameters:

- Define GC content ☐
- Set % of the GC content: Automatic
- Define the start codons ☐
- Set the start codons: Automatic
- Start codon weight: Automatic
- Define the stop codons ☐
- Set the stop codons: Automatic

Version Details:
- Glimmer 3.02

Please Cite:
- Delcher AL., Harmon D., Kasif S., White O. and Salzberg SL. (1999). Improved microbial gene identification with GLIMMER. *Nucleic acids research*, *27*(23), 4636-41.

Figure 4: Configuration 3 Page

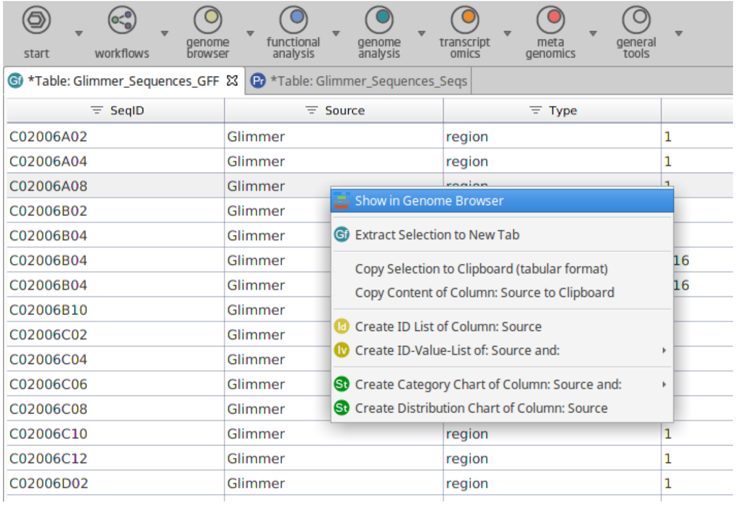
Results

Once the gene finding tools have finished, two objects will automatically be opened:

- **Sequence table:** Here you can see the traditional OmicsBox table showing the sequence name corresponding to the fasta ID line plus a gene identification, and the sequence length. Note: this sequence can be on nucleotides or in amino acids, depending on the wizard selection.
- **GFF3 table:** Here you can see the results as a gff file with:
  - Sequence: The name of the source sequence that belongs to this feature.
  - Source: The name of the program that has predicted this feature, in this case, `Glimmer'.
  - Type: The type of the feature, that can be `gene', `mRNA', `CDS', `gene', `Start', `stop', `gene'
  - Start: The coordinate of the start codon.
  - End: The coordinate of the stop codon.
  - Score: The score assigned to the feature, except the exons.
  - Strand: The strand of the feature, where a `+' means that the feature is forward oriented and `-' backwards.
  - Phase: The correct frame to translate this feature, the values can be `0', `1' or `2'. A gene `set' of features can have variant phase values, due to a frameshift in an intron.
  - Attributes: Here we can see all the attributes assigned to each feature. The attributes are `ID' that assigns an id to each feature, `parent' present on the CDS and exon features, and

provides information about the feature to which it belongs (refereeing to the sequence by its ID).

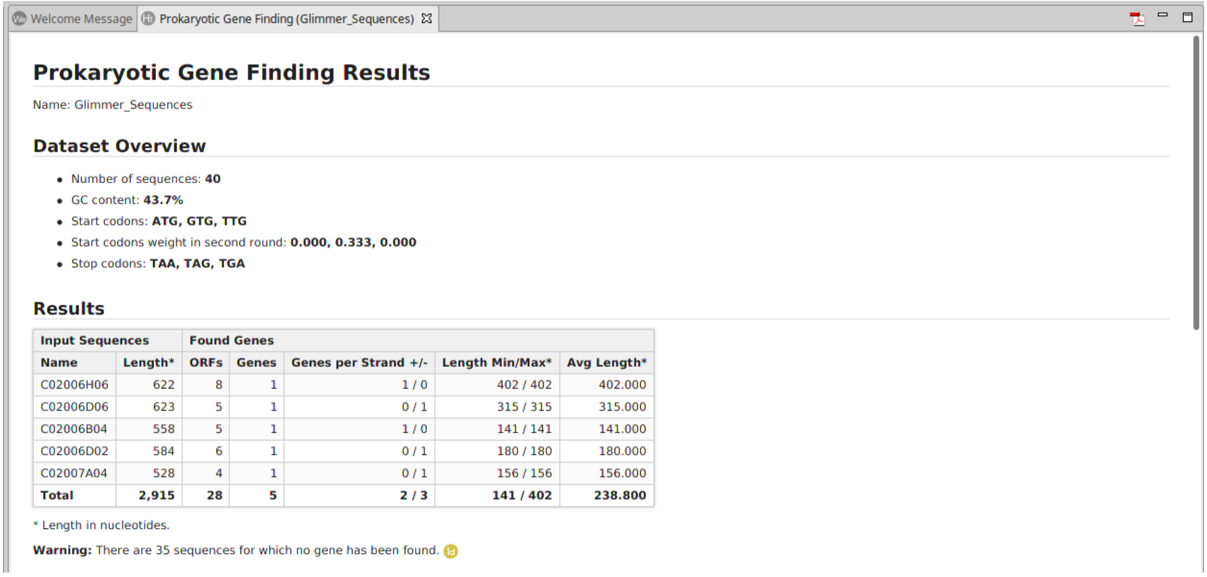
The resulting GFF3 can be inspected using the Genome Browser. To display a GFF entry right click on it and select the **Show in the Genome Browser** option (figure 5(see page 137)). For more information about this feature visit the Genome Browser(see page 23) documentation section.


start
workflows
genome browser
functional analysis
genome analysis
transcript omics
meta genomics
general tools
*Table: Glimmer_Sequences_GFF
*Table: Glimmer_Sequences_Seqs
Show in Genome Browser
Extract Selection to New Tab
Copy Selection to Clipboard (tabular format)
Copy Content of Column: Source to Clipboard
Create ID List of Column: Source
Create ID-Value-List of: Source and:
Create Category Chart of Column: Source and:
Create Distribution Chart of Column: Source

SeqID	Source	Type	
C02006A02	Glimmer	region	1
C02006A04	Glimmer	region	1
C02006A08	Glimmer	region	1
C02006B02	Glimmer		
C02006B04	Glimmer		
C02006B04	Glimmer		16
C02006B04	Glimmer		16
C02006B10	Glimmer		
C02006C02	Glimmer		
C02006C04	Glimmer		
C02006C06	Glimmer		
C02006C08	Glimmer		
C02006C10	Glimmer	region	1
C02006C12	Glimmer	region	1
C02006D02	Glimmer	region	1

Figure 5: How to open the Genome Browser

A Result Viewer is also opened to display the name of each sequence present in the fasta file, the number of genes per sequence, the minimum and maximum gene length, and the strand position of the genes found (figure 6(see page 137)).



Prokaryotic Gene Finding Results

Name: Glimmer_Sequences

Dataset Overview

- Number of sequences: **40**
- GC content: **43.7%**
- Start codons: **ATG, GTG, TTG**
- Start codons weight in second round: **0.000, 0.333, 0.000**
- Stop codons: **TAA, TAG, TGA**

Results

Input Sequences		Found Genes				
Name	**Length***	**ORFs**	**Genes**	**Genes per Strand +/-**	**Length Min/Max***	**Avg Length***
C02006H06	622	8	1	1 / 0	402 / 402	402.000
C02006D06	623	5	1	0 / 1	315 / 315	315.000
C02006B04	558	5	1	1 / 0	141 / 141	141.000
C02006D02	584	6	1	0 / 1	180 / 180	180.000
C02007A04	528	4	1	0 / 1	156 / 156	156.000
Total	**2,915**	**28**	**5**	**2 / 3**	**141 / 402**	**238.800**

* Length in nucleotides.

Warning: There are 35 sequences for which no gene has been found.

Figure 6: Result Summary

8.4 Eukaryotic Gene Finding by Augustus

Content of this page:

8.4.1 Introduction

Augustus is a program that predicts genes in eukaryotic genomic sequences; It is one of the most accurate programs for the species it is trained for. In the human ENCODE project, it proved to be the most accurate gene finder among the tested `ab initio' programs. In the more recent nGASP (worm) project, it was again among the best in the `ab initio' and transcript-based categories.

The accuracy of Augustus lies on his precomputed models which facilitate fast and accurate gene prediction.

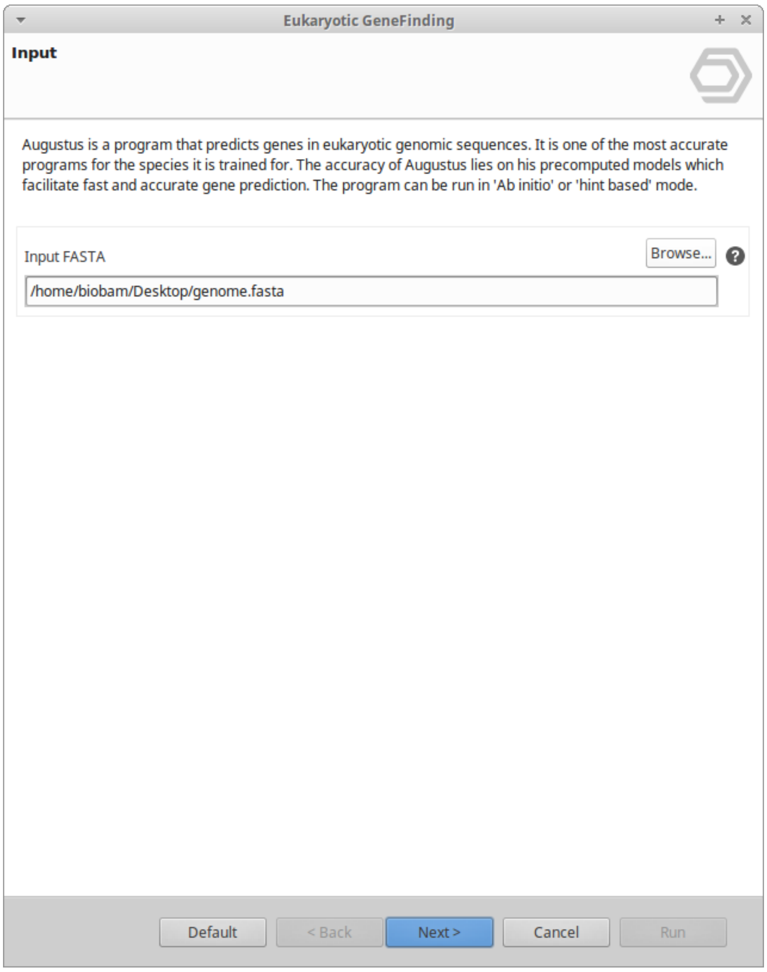
8.4.2 Run Eukaryotic Gene Finding

In order to speed up the gene finding process, the fasta will be split by sequence, i.e, each fasta entry will be sent to a different node for parallel execution (figure 1(see page 138), figure 2(see page 139) and figure 3(see page 141)).

8.4.2.1 Input Page

- **Input FASTA:** The query file contains the DNA input sequence which must be in decompressed (multiple or single) FASTA format. Every letter other than a, c, g, t, A, C, G, and T is interpreted as an unknown base. Digits and white spaces are ignored. The number of characters per line is not restricted.

> ⚠ Note: The differences that make the fasta identifiers unique must be within the first 30 characters to be recognized by Augustus.



Eukaryotic GeneFinding

Input

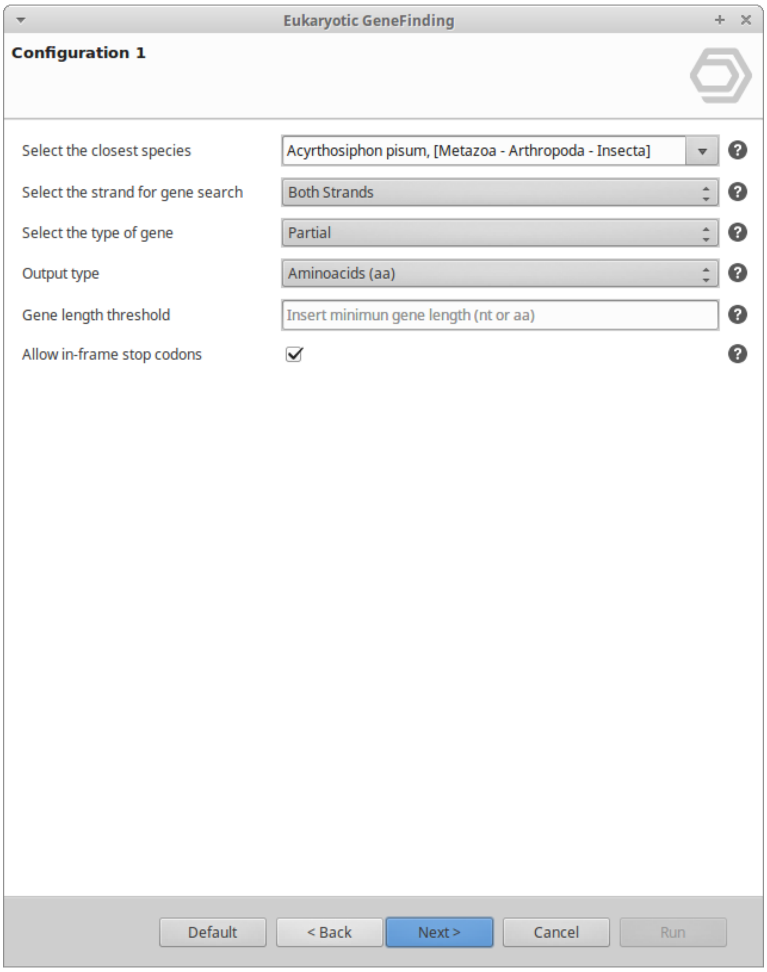
Augustus is a program that predicts genes in eukaryotic genomic sequences. It is one of the most accurate programs for the species it is trained for. The accuracy of Augustus lies on his precomputed models which facilitate fast and accurate gene prediction. The program can be run in 'Ab initio' or 'hint based' mode.

Input FASTA: /home/biobam/Desktop/genome.fasta

Figure 1: Augustus Input Data Page

8.4.2.2 Configuration 1 Page

- **Closest species:** This list allows for the selection of the closest related organism to your query, in order to obtain the most accurate prediction.
- **Strand:** Here you can choose the sense of the gene search, obtaining the predicted genes on the forward strand, the backward strand or on both strands.
- **Type of gene:** With this option, you can select the gene model.
  - partial: allows prediction of incomplete genes at the sequence boundaries (default)
  - intronless: predicts only single-exon genes like in prokaryotes and some eukaryotes
  - complete: predicts only complete genes
- **Output type:** Specify whether the output sequences will be extracted as nucleotides or amino acids.
- **Protein length threshold:** Set a minimum length of the predicted proteins.
- **Allow in-frame stops:** Activating this checkbox will allow the detection of genes containing a stop codon in its reading frame, detecting fragment genes with some undetected zones; normally it's the most suitable option for an 'ab initio' search.



Select the closest species: Acyrthosiphon pisum, [Metazoa - Arthropoda - Insecta]

Select the strand for gene search: Both Strands

Select the type of gene: Partial

Output type: Aminoacids (aa)

Gene length threshold: Insert minimun gene length (nt or aa)

Allow in-frame stop codons: ☑

Figure 2: Augustus Configuration 1 Page

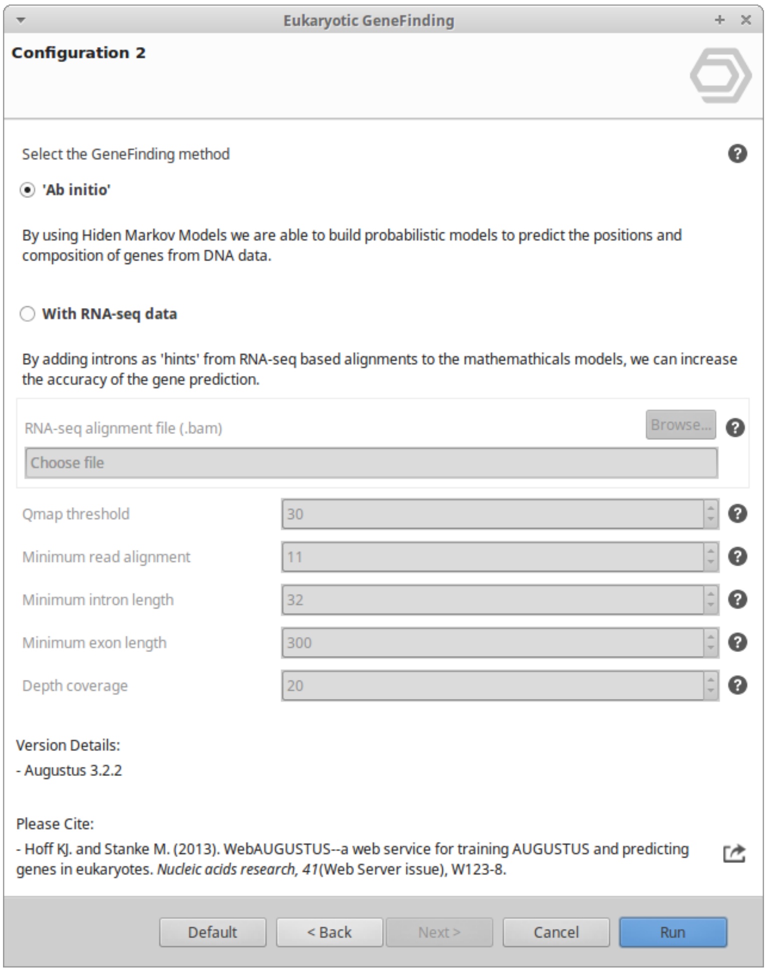
8.4.2.3 Configuration 2 Wizard Page

The eukaryotic gene finding can be executeConfiguration 2 Wizard Paged 'ab initio', using only DNA-seq data, or using 'hints' obtained from the RNAseq alignment in order to increase the truthfulness of the predicted genes.

- **RNAseq alignment file:** The file containing the alignments in BAM format. This file is the output of every RNAseq aligner program as TopHat, BWA or STAR. To be able to locate hints in the alignment file, it must not be filtered by any parameter, that means that it must be the same file that you obtain from the aligner. For this reason, the alignment files from Ensembl are not suitable for retrieving hints as they are filtered and processed.
- **Qmap threshold:** This parameter allows filtering the aligned reads that will be used to create the intron 'hints'. The Qmap corresponds to the mapping quality in a range from 0 to 60 and it is calculated as: Meaning that a Qmap of 50, corresponds to a mapping error of 5 x 105. Default: 50.
- **Minimum read alignment:** Specify the minimum length of the read that must map to the reference genome at the beginning of the intron. If this value is too small, it can lead the program to detect an intron derived from a miss-alignment (figure 4(see page 141)).

⚠ Note: This value has 0 as minimum and the maximum depends on your reads length. Default: 11.

- **Minimum intron length:** Sets the minimum intron length (figure 4(see page 141)). Default: 32.
- **Minimum exon length:** Sets the minimum exon length (figure 4(see page 141)). Default: 300.
- **Depth coverage:** Sets the number of reads that must be aligned at a position to consider it as a consistent exon. Default: 20.



Eukaryotic GeneFinding

Configuration 2

Select the GeneFinding method

● **'Ab initio'**

By using Hiden Markov Models we are able to build probabilistic models to predict the positions and composition of genes from DNA data.

○ **With RNA-seq data**

By adding introns as 'hints' from RNA-seq based alignments to the mathemathicals models, we can increase the accuracy of the gene prediction.

RNA-seq alignment file (.bam) Browse...

Choose file

Qmap threshold: 30

Minimum read alignment: 11

Minimum intron length: 32

Minimum exon length: 300

Depth coverage: 20

Version Details:

- Augustus 3.2.2

Please Cite:

- Hoff KJ. and Stanke M. (2013). WebAUGUSTUS--a web service for training AUGUSTUS and predicting genes in eukaryotes. *Nucleic acids research, 41*(Web Server issue), W123-8.

Figure 3: Augustus Configuration 2 Page

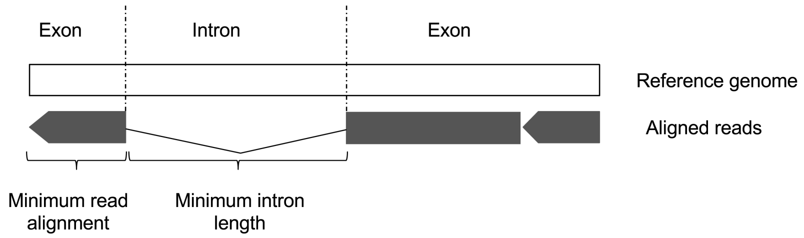


Figure 4: The concept of minimum read alignment and minimum intron length

8.4.3 Results

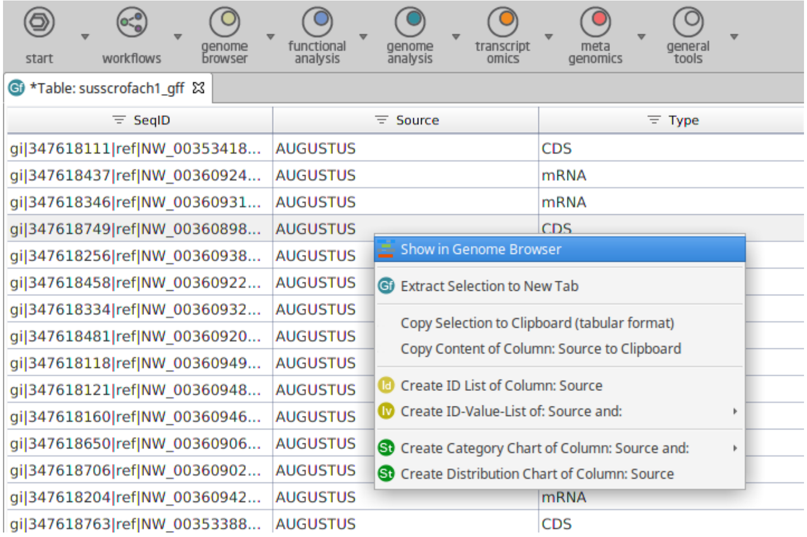
Two result tables will automatically be opened:

- **Sequence table:**
  Here you can see the traditional OmicsBox table showing the sequence name corresponding to the fasta ID line plus a gene identification, and the sequence length.

  > ⚠ Note: These sequences can be on nucleotides or in amino acids, depending on the wizard selection.

- **GFF3 table columns:**
  - Sequence: The name of the source sequence that belongs to this feature.
  - Source: The name of the program that has predicted this feature, in this case, `Augustus'.
  - Type: The type of the feature, that can be `gene', `mRNA', `CDS', `gene', `Start', `stop', `gene'
  - Start: The coordinate of the start codon.
  - End: The coordinate of the stop codon.
  - Score: The score assigned to the feature, except the exons.
  - Strand: The strand of the feature, where a `+' means that the feature is forward oriented and `-' backwards.
  - Phase: The correct frame to translate this feature, the values can be `0', `1' or `2'. A gene `set' of features can have variant phase values, due to a frame shift in an intron.
  - Attributes: Here we can see all the attributes assigned to each feature. The attributes are `ID' that assigns an id to each feature, `parent' present on the CDS and exon features, and provides information about the feature to which it belongs (refereeing to the sequence by his ID).

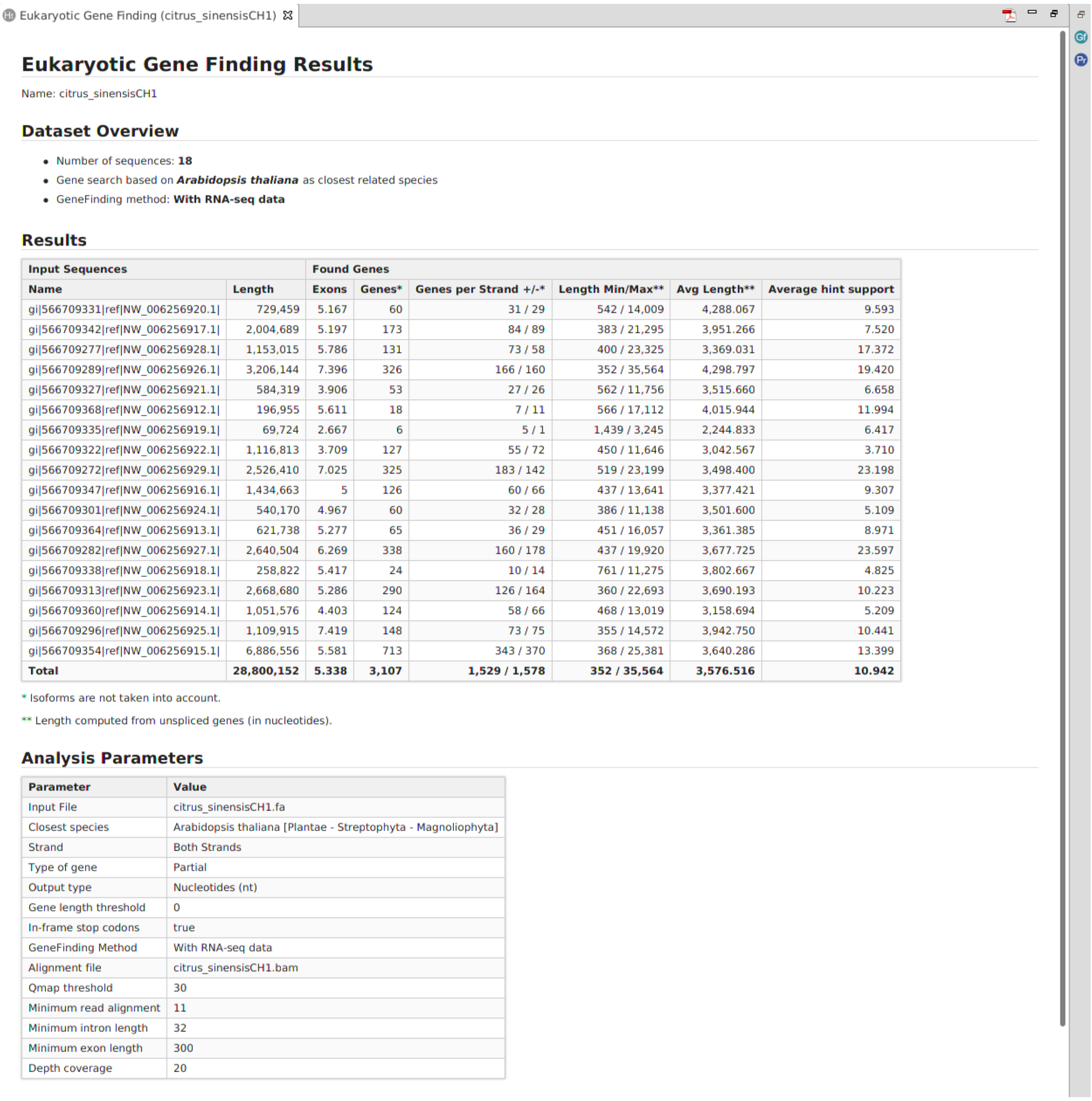
The resulting GFF3 can be inspected by using the Genome Browser. To display a GFF entry right click on it and select the **Show in the Genome Browser** option (figure 5(see page 142)). For more information about this feature visit the Genome Browser(see page 23).


start
workflows
genome browser
functional analysis
genome analysis
transcript omics
meta genomics
general tools
*Table: susscrofach1_gff
Show in Genome Browser
Extract Selection to New Tab
Copy Selection to Clipboard (tabular format)
Copy Content of Column: Source to Clipboard
Create ID List of Column: Source
Create ID-Value-List of: Source and:
Create Category Chart of Column: Source and:
Create Distribution Chart of Column: Source

SeqID	Source	Type
gi|347618111|ref|NW_00353418...	AUGUSTUS	CDS
gi|347618437|ref|NW_00360924...	AUGUSTUS	mRNA
gi|347618346|ref|NW_00360931...	AUGUSTUS	mRNA
gi|347618749|ref|NW_00360898...	AUGUSTUS	CDS
gi|347618256|ref|NW_00360938...	AUGUSTUS	
gi|347618458|ref|NW_00360922...	AUGUSTUS	
gi|347618334|ref|NW_00360932...	AUGUSTUS	
gi|347618481|ref|NW_00360920...	AUGUSTUS	
gi|347618118|ref|NW_00360949...	AUGUSTUS	
gi|347618121|ref|NW_00360948...	AUGUSTUS	
gi|347618160|ref|NW_00360946...	AUGUSTUS	
gi|347618650|ref|NW_00360906...	AUGUSTUS	
gi|347618706|ref|NW_00360902...	AUGUSTUS	
gi|347618204|ref|NW_00360942...	AUGUSTUS	mRNA
gi|347618763|ref|NW_00353388...	AUGUSTUS	CDS

Figure 5: How to open the Genome Browser

A Result Viewer is also opened to display some the number and name of sequences per spitted file, the average number of exons, the minimum, maximum and average gene length, and the number of genes per strand (figure 6(see page 143)).



Eukaryotic Gene Finding Results

Name: citrus_sinensisCH1

Dataset Overview

- Number of sequences: **18**
- Gene search based on ***Arabidopsis thaliana*** as closest related species
- GeneFinding method: **With RNA-seq data**

Results

Input Sequences		Found Genes					
Name	**Length**	**Exons**	**Genes***	**Genes per Strand +/-***	**Length Min/Max****	**Avg Length****	**Average hint support**
gi|566709331|ref|NW_006256920.1|	729,459	5.167	60	31 / 29	542 / 14,009	4,288.067	9.593
gi|566709342|ref|NW_006256917.1|	2,004,689	5.197	173	84 / 89	383 / 21,295	3,951.266	7.520
gi|566709277|ref|NW_006256928.1|	1,153,015	5.786	131	73 / 58	400 / 23,325	3,369.031	17.372
gi|566709289|ref|NW_006256926.1|	3,206,144	7.396	326	166 / 160	352 / 35,564	4,298.797	19.420
gi|566709327|ref|NW_006256921.1|	584,319	3.906	53	27 / 26	562 / 11,756	3,515.660	6.658
gi|566709368|ref|NW_006256912.1|	196,955	5.611	18	7 / 11	566 / 17,112	4,015.944	11.994
gi|566709335|ref|NW_006256919.1|	69,724	2.667	6	5 / 1	1,439 / 3,245	2,244.833	6.417
gi|566709322|ref|NW_006256922.1|	1,116,813	3.709	127	55 / 72	450 / 11,646	3,042.567	3.710
gi|566709272|ref|NW_006256929.1|	2,526,410	7.025	325	183 / 142	519 / 23,199	3,498.400	23.198
gi|566709347|ref|NW_006256916.1|	1,434,663	5	126	60 / 66	437 / 13,641	3,377.421	9.307
gi|566709301|ref|NW_006256924.1|	540,170	4.967	60	32 / 28	386 / 11,138	3,501.600	5.109
gi|566709364|ref|NW_006256913.1|	621,738	5.277	65	36 / 29	451 / 16,057	3,361.385	8.971
gi|566709282|ref|NW_006256927.1|	2,640,504	6.269	338	160 / 178	437 / 19,920	3,677.725	23.597
gi|566709338|ref|NW_006256918.1|	258,822	5.417	24	10 / 14	761 / 11,275	3,802.667	4.825
gi|566709313|ref|NW_006256923.1|	2,668,680	5.286	290	126 / 164	360 / 22,693	3,690.193	10.223
gi|566709360|ref|NW_006256914.1|	1,051,576	4.403	124	58 / 66	468 / 13,019	3,158.694	5.209
gi|566709296|ref|NW_006256925.1|	1,109,915	7.419	148	73 / 75	355 / 14,572	3,942.750	10.441
gi|566709354|ref|NW_006256915.1|	6,886,556	5.581	713	343 / 370	368 / 25,381	3,640.286	13.399
Total	**28,800,152**	**5.338**	**3,107**	**1,529 / 1,578**	**352 / 35,564**	**3,576.516**	**10.942**

* Isoforms are not taken into account.

** Length computed from unspliced genes (in nucleotides).

Analysis Parameters

Parameter	Value
Input File	citrus_sinensisCH1.fa
Closest species	Arabidopsis thaliana [Plantae - Streptophyta - Magnoliophyta]
Strand	Both Strands
Type of gene	Partial
Output type	Nucleotides (nt)
Gene length threshold	0
In-frame stop codons	true
GeneFinding Method	With RNA-seq data
Alignment file	citrus_sinensisCH1.bam
Qmap threshold	30
Minimum read alignment	11
Minimum intron length	32
Minimum exon length	300
Depth coverage	20

Figure 6: Result Summary

9 Module Transcriptomics

Content of this section:

Detecting genes that are differentially expressed between conditions is a fundamental part of understanding the molecular basis of phenotypic variation. To take advantage of the possibilities and address the challenges posed by this relatively new type of data, OmicsBox offers several tools to analyze RNA-Seq data and obtain functional insights.

The OmicsBox Transcriptomics module allows you to process RNA-seq data from raw reads down to their functional analysis in a flexible and intuitive way.

- **Quality Control:** Use **FastQc** and **Trimmomatic** to perform the quality control of your sequencing samples, to filter reads and remove low-quality bases.
- ***De novo* Assembly:** Assemble short reads with **Trinity** to create a *de novo* transcriptome without a reference genome.
- **RNA-Seq Alignment:** Align RNA-Seq data to your reference genome making use of **STAR,** an ultrafast universal RNA-Seq aligner.
- **Quantify Expression:** Quantify expression at gene or transcript level through **HTSeq** or **RSEM** and with or without a reference genome.
- **Differential Expression Analysis:** Detect differentially expressed genes between experimental conditions or over time with well-known and versatile statistical packages like **NOISeq**, **edgeR** or **maSigPro**. Rich visualizations help to interpret results.
- **Enrichment Analysis:** By combining differential expression results with functional annotations, enrichment analysis allows to identify over and underrepresented biological functions.

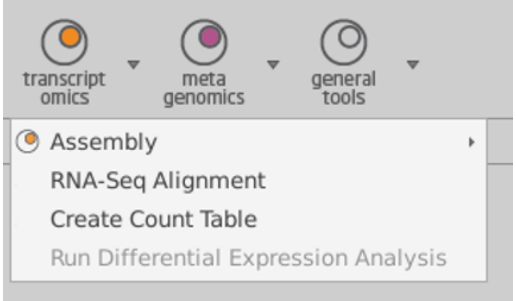


Figure 1: Transcriptomics menu

Transcriptomic Analysis use case: https://www.biobam.com/drug-response-transcriptomics/.

Transcriptomic Example Dataset: Download[70].

9.1 RNA-Seq de novo Assembly

Content of this page:

9.1.1 Introduction

De novo transcriptome assembly is one of the most frequent analyses performed in bioinformatics and it consists of reconstructing the transcriptome from RNA sequencing data, assembling short nucleotide sequences into longer ones without the use of a reference genome. This functionality is based on Trinity[71], a well-known *de novo* sequence assembler software developed at the Broad Institute and the Hebrew University of Jerusalem.

Trinity combines three independent software modules applied sequentially to process large volumes of RNA-seq reads. Trinity partitions the sequence data into many individual de Bruijn graphs, each representing the transcriptional complexity at a given gene or locus, and then processes each graph independently to extract full-length splicing isoforms and to tease apart transcripts derived from paralogous genes.

Please, cite Trinity as:

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A (2011). "Full-length transcriptome assembly from RNA-Seq data without a reference genome." Nature Biotechnology, 29(7):644-52.[72]

9.1.2 Run RNA-Seq *de novo* Assembly

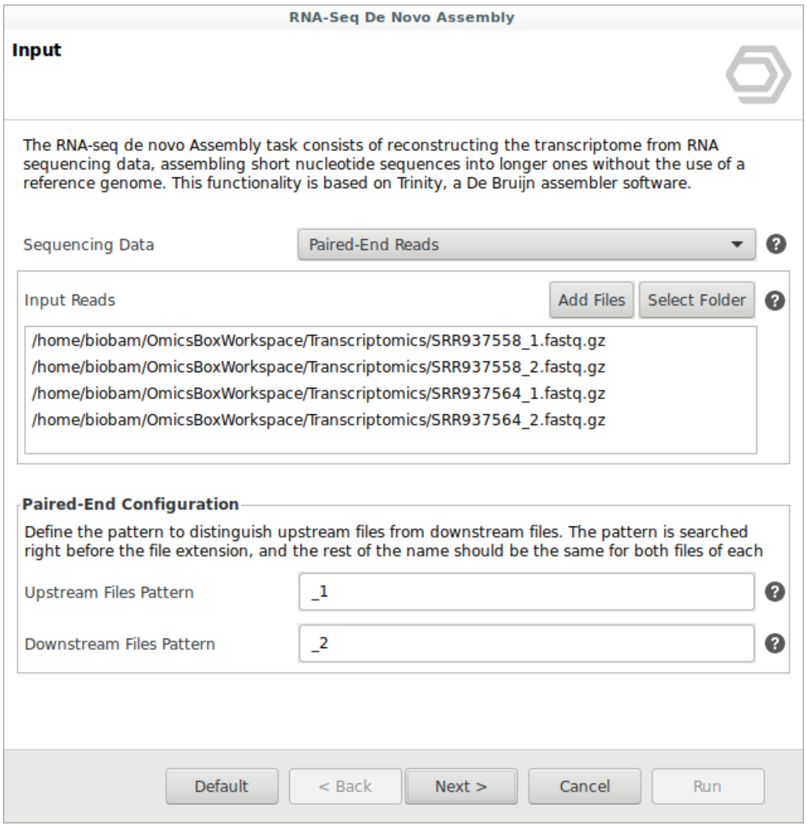
This functionality can be found under **Transcriptomics** → **Assembly** → **RNA-Seq De novo Assembly**. The wizard allows to select files and set the parameters (Figure 1[73], Figure 2[74] and Figure 3[75]).

[70] https://resources.biobam.com/omicsbox/example_data/Transcriptomics.zip
[71] https://github.com/trinityrnaseq/trinityrnaseq/wiki
[72] https://www.ncbi.nlm.nih.gov/pubmed/21572440
[73] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598278202#RNA-SeqdenovoAssembly-figure1
[74] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598278202#RNA-SeqdenovoAssembly-figure2
[75] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598278202#RNA-SeqdenovoAssembly-figure3

9.1.2.1 Input

- **Sequencing Data:** Choose the type of data to be preprocessed: single-end or paired-end reads. Note that if paired-end is selected, two files per sample are required.
- **Input Reads:** Provide the files containing sequencing reads. These files are assumed to be in FASTQ format.
- **Paired-end configuration:** In the case of paired-end reads, the pattern to distinguish upstream files from downstream files is required. The provided patterns are searched right before the extension, and the start of the name should be the same for both files of each sample.
  - Upstream Files Pattern: Establish the pattern to recognize upstream FASTQ files.
  - Downstream Files Pattern: Establish the pattern to recognize downstream FASTQ files.

> ⚠ For example, if the upstream file is named SRR037717_1.fastq and the downstream one SRR037717_2.fastq, you should establish "_1" as the upstream pattern and "_2" as the downstream pattern.



RNA-Seq De Novo Assembly

Input

The RNA-seq de novo Assembly task consists of reconstructing the transcriptome from RNA sequencing data, assembling short nucleotide sequences into longer ones without the use of a reference genome. This functionality is based on Trinity, a De Bruijn assembler software.

Sequencing Data: Paired-End Reads

Input Reads

/home/biobam/OmicsBoxWorkspace/Transcriptomics/SRR937558_1.fastq.gz
/home/biobam/OmicsBoxWorkspace/Transcriptomics/SRR937558_2.fastq.gz
/home/biobam/OmicsBoxWorkspace/Transcriptomics/SRR937564_1.fastq.gz
/home/biobam/OmicsBoxWorkspace/Transcriptomics/SRR937564_2.fastq.gz

Paired-End Configuration

Define the pattern to distinguish upstream files from downstream files. The pattern is searched right before the file extension, and the rest of the name should be the same for both files of each

Upstream Files Pattern: _1

Downstream Files Pattern: _2

Figure 1: Input Data Page

9.1.2.2 Configuration

- **K-mer Size:** The term k-mer refers to all possible subsequences of the given length that are contained in a read. In sequence assembly, k-mers are used during the construction of De Bruijn graphs. The choice of the k-mer size has different effects on the sequence assembly, so it is advisable to try different values and check the results to choose the best one. Trinity suggests using a k-mer size of 25 (default value). The maximum value allowed is 32.

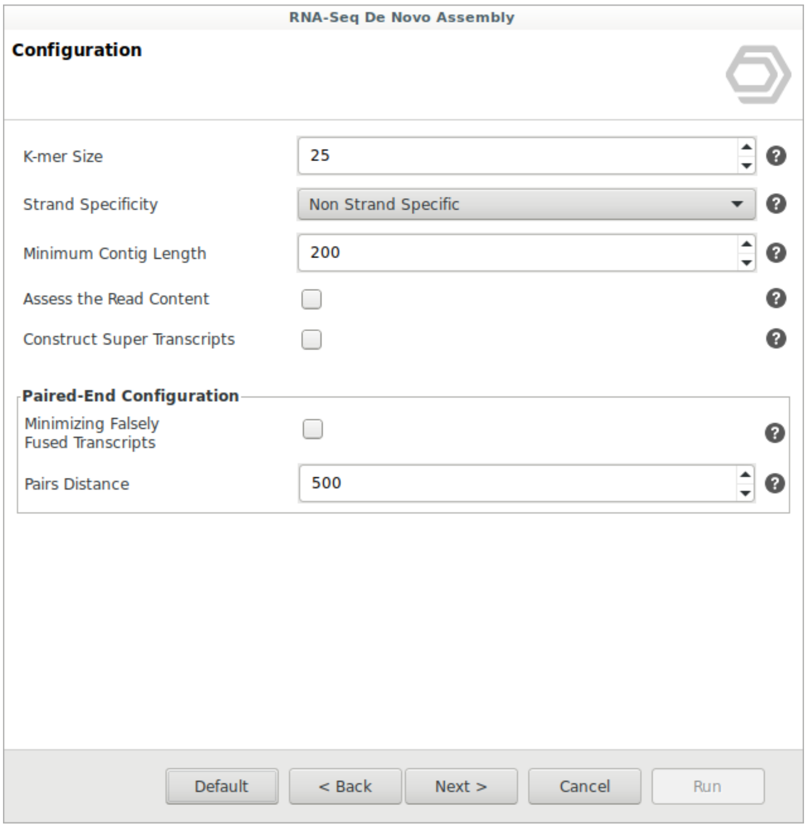
- **Strand Specificity:** This option defines the strandedness of the RNA-seq reads:
    - Non-Strand Specific: This refers to non-strand-specific protocols.
    - Strand Specific Forward: For single-end data, the single read is in the sense (forward) orientation. In the case of paired-end data, the first read of fragment pair is sequenced as sense (forward), and the second is in the antisense strand (reverse).
    - Strand Specific Reverse: For single-end data, the single read is in the antisense (reverse) orientation. In the case of paired-end data, the first read of fragment pair is sequenced as anti-sense (reverse), and the second read is in the sense strand (forward). Typical of the dUTP/UDG sequencing method.
- **Minimum Contig Length:** Minimum assembled contig length to report. Trinity uses 200 bp as default value.
- **Assess the Read Content:** To assess the read composition of the assembly, input RNA-Seq reads are aligned to the transcriptome assembly using Bowtie2. Reads that map to the assembled transcript are captured and counted, including the properly paired and those that are not. Check this option to obtain the read representation charts and table.

⚠ Note that it is an expensive operation, so the process will take more time.

- **Construct Super Transcripts:** SuperTranscripts provide a gene-like view of the transcriptional complexity of a gene. A SuperTranscript is constructed by collapsing unique and common sequence regions among splicing isoforms into a single linear sequence.
- **Minimizing Falsely Fused Transcripts:** If the transcriptome RNA-seq data under study are derived from a gene-dense compact genome, fusion transcripts can be minimized. This option is only available for paired-end data. In compact fungal genomes, it is highly recommended.

  ⚠ Note that it is an expensive operation, so avoid using it unless necessary.

- **Pair Distance:** Maximum length expected between fragment pairs (500 nucleotides by default). Reads outside this distance are treated as single-end.



K-mer Size: 25

Strand Specificity: Non Strand Specific

Minimum Contig Length: 200

Assess the Read Content □

Construct Super Transcripts □

Paired-End Configuration

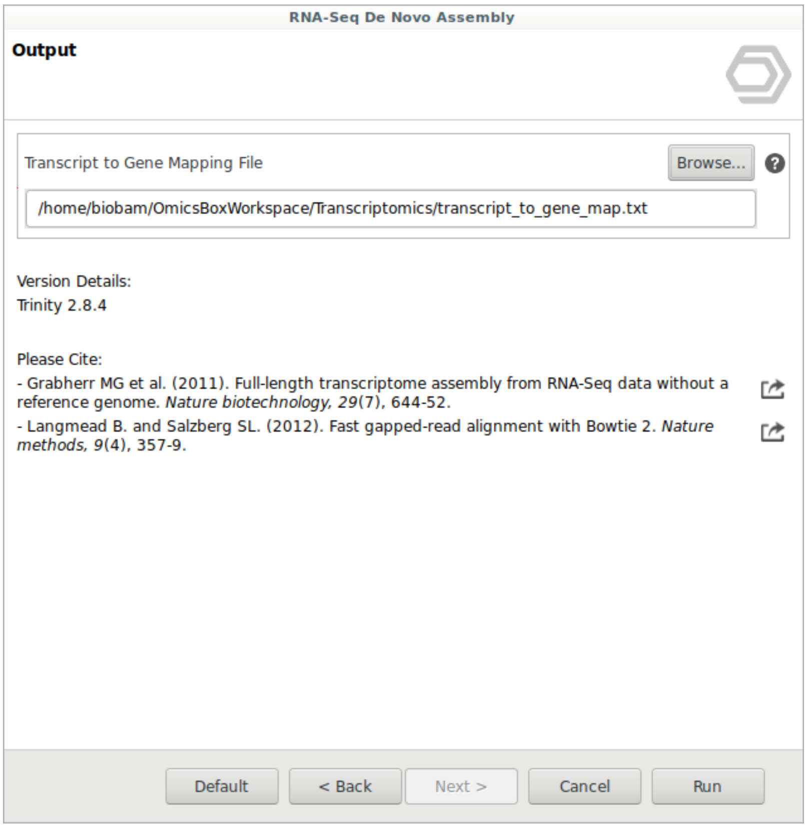
Minimizing Falsely Fused Transcripts □

Pairs Distance: 500

Figure 2: Configuration Page

9.1.2.3 Output

- **Transcript to Gene Mapping:** Select a location to place the transcript to the gene mapping file. It is a tab-delimited file with the information to map from transcript (isoform) identifiers to gene identifiers. It could be used in downstream analysis such as the Transcript-level Quantification.



Transcript to Gene Mapping File

/home/biobam/OmicsBoxWorkspace/Transcriptomics/transcript_to_gene_map.txt

Version Details:
Trinity 2.8.4

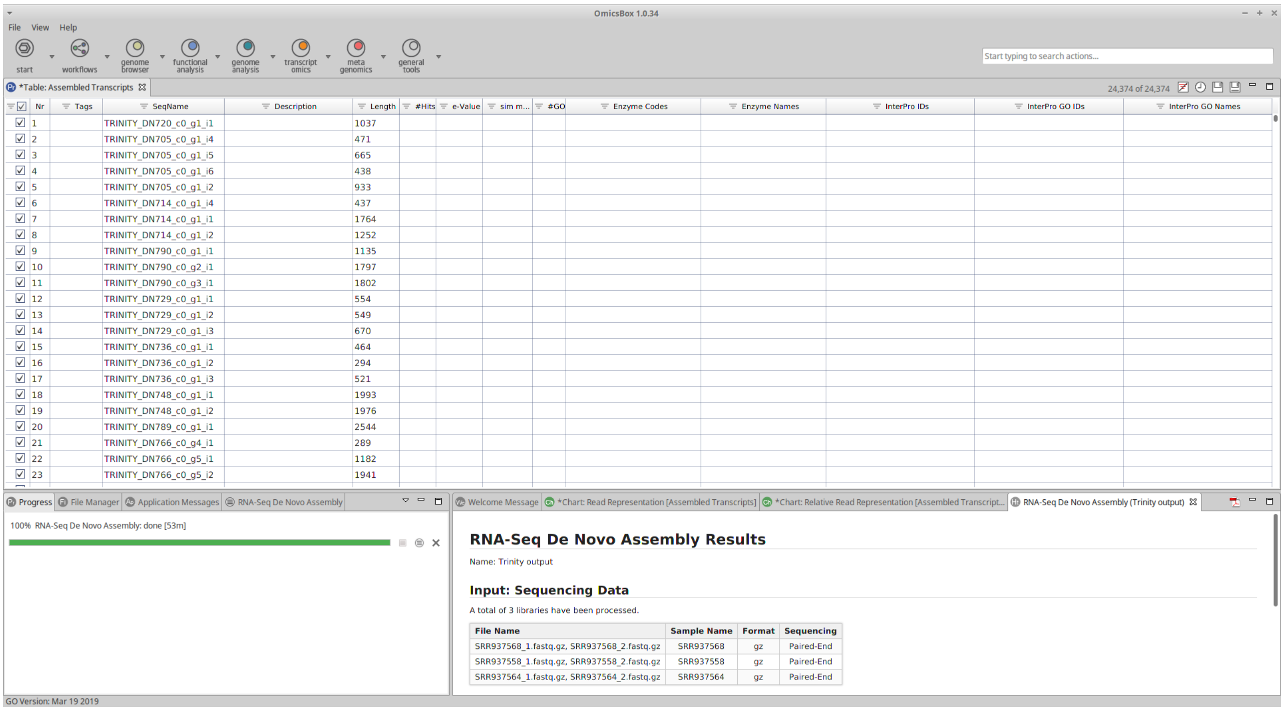
Please Cite:
- Grabherr MG et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology, 29*(7), 644-52.
- Langmead B. and Salzberg SL. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods, 9*(4), 357-9.

Figure 3: Output Data Page

9.1.3 Results

When the RNA-seq *de novo* assembly completes, it creates a sequence table containing the assembled transcripts sequences (Figure 4[76]). Trinity groups transcripts into clusters based on shared sequence content. Such a transcript cluster can be considered as a 'gene'.

[76] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598278202#RNA-SeqdenovoAssembly-figure4



☑	Nr	Tags	SeqName	Description	Length	#Hits	e-Value	sim m...	#GO	Enzyme Codes	Enzyme Names	InterPro IDs	InterPro GO IDs	InterPro GO Names
☑	1		TRINITY_DN720_c0_g1_i1		1037									
☑	2		TRINITY_DN705_c0_g1_i4		471									
☑	3		TRINITY_DN705_c0_g1_i5		665									
☑	4		TRINITY_DN705_c0_g1_i6		438									
☑	5		TRINITY_DN705_c0_g1_i2		933									
☑	6		TRINITY_DN714_c0_g1_i4		437									
☑	7		TRINITY_DN714_c0_g1_i1		1764									
☑	8		TRINITY_DN714_c0_g1_i2		1252									
☑	9		TRINITY_DN790_c0_g1_i1		1135									
☑	10		TRINITY_DN790_c0_g2_i1		1797									
☑	11		TRINITY_DN790_c0_g3_i1		1802									
☑	12		TRINITY_DN729_c0_g1_i1		554									
☑	13		TRINITY_DN729_c0_g1_i2		549									
☑	14		TRINITY_DN729_c0_g1_i3		670									
☑	15		TRINITY_DN736_c0_g1_i1		464									
☑	16		TRINITY_DN736_c0_g1_i2		294									
☑	17		TRINITY_DN736_c0_g1_i3		521									
☑	18		TRINITY_DN748_c0_g1_i1		1993									
☑	19		TRINITY_DN748_c0_g1_i2		1976									
☑	20		TRINITY_DN789_c0_g1_i1		2544									
☑	21		TRINITY_DN766_c0_g4_i1		289									
☑	22		TRINITY_DN766_c0_g5_i1		1182									
☑	23		TRINITY_DN766_c0_g5_i2		1941									

100% RNA-Seq De Novo Assembly: done [53m]

RNA-Seq De Novo Assembly Results

Name: Trinity output

Input: Sequencing Data

A total of 3 libraries have been processed.

File Name	Sample Name	Format	Sequencing
SRR937568_1.fastq.gz, SRR937568_2.fastq.gz	SRR937568	gz	Paired-End
SRR937558_1.fastq.gz, SRR937558_2.fastq.gz	SRR937558	gz	Paired-End
SRR937564_1.fastq.gz, SRR937564_2.fastq.gz	SRR937564	gz	Paired-End

Figure 4: Sequence table project containing the sequences of the assembled transcripts

This information is encoded in the Trinity FASTA accession. An example FASTA entry for one of the transcripts is formatted like so:

- Isoform 1: TRINITY_DN869_c0_g1_i1
- Isoform 2: TRINITY_DN869_c0_g1_i2

The accession encodes the Trinity 'gene' and 'isoform' information. In the example above, the accession 'TRINITY_DN869_c0_g1_i1' indicates Trinity read cluster 'TRINITY_DN869_c0, gene 'g1', and isoform 'i1' and 'i2'. Because a given run of trinity involves many clusters of reads, each of which are assembled separately, and because the 'gene' numbering is unique within a given processed read cluster, the 'gene' identifier should be considered an aggregate of the read cluster and corresponding gene identifier, which in this case would be 'TRINITY_DN869_c0_g1'.

If the Construct Super Transcript option was checked, two additional outputs will be generated:

- SuperTranscripts in FASTA format.
- Transcript structure annotation in GFF format.

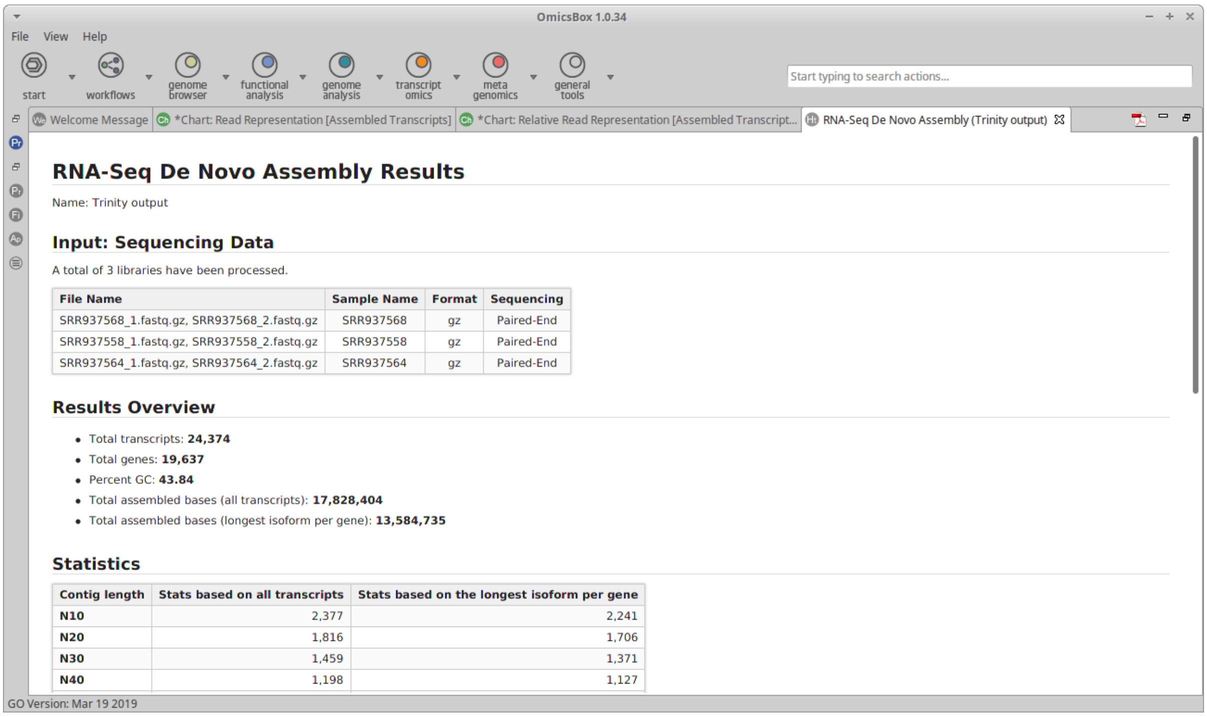
Furthermore, a result page will show a summary of the RNA-seq *de novo* assembly results (Figure 5[77]). It contains the following information:

- Details of input FASTQ files.
- Results overview that informs about the number of total transcripts and genes detected, the percentage of GC and the total assembled bases.
- Statistics based on the lengths of the assembled transcriptome contigs. The conventional Nx length statistic means that at least x% of the assembled transcript nucleotides are found in contigs that are at

[77] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598278202#RNA-SeqdenovoAssembly-figure5

least of Nx length. For example, the N50 means that at least half of all assembled bases are in transcript contigs of at least the N50 length value.

- The RNA-Seq Read Representation, that allows assessing the read composition of the assembly. It shows the number of reads that map to the assembled transcripts, including the properly paired and those that are not (details below).


OmicsBox 1.0.34
File
View
Help
start
workflows
genome browser
functional analysis
genome analysis
transcript omics
meta genomics
general tools
Start typing to search actions...
Welcome Message
*Chart: Read Representation [Assembled Transcripts]
*Chart: Relative Read Representation [Assembled Transcript...
RNA-Seq De Novo Assembly (Trinity output)
GO Version: Mar 19 2019

RNA-Seq De Novo Assembly Results

Name: Trinity output

Input: Sequencing Data

A total of 3 libraries have been processed.

File Name	Sample Name	Format	Sequencing
SRR937568_1.fastq.gz, SRR937568_2.fastq.gz	SRR937568	gz	Paired-End
SRR937558_1.fastq.gz, SRR937558_2.fastq.gz	SRR937558	gz	Paired-End
SRR937564_1.fastq.gz, SRR937564_2.fastq.gz	SRR937564	gz	Paired-End

Results Overview

- Total transcripts: **24,374**
- Total genes: **19,637**
- Percent GC: **43.84**
- Total assembled bases (all transcripts): **17,828,404**
- Total assembled bases (longest isoform per gene): **13,584,735**

Statistics

Contig length	Stats based on all transcripts	Stats based on the longest isoform per gene
N10	2,377	2,241
N20	1,816	1,706
N30	1,459	1,371
N40	1,198	1,127

Figure 5: Summary report

Finally, two charts showing the read representation of the assembly are generated (Figure 6[78] and Figure 7[79]). These charts display the number of reads of each input file sorted by different categories (the second chart represents the same information in percentages). Bowtie2 is used to align the reads to the transcriptome and then the number of the single-end reads or proper pairs and improper or orphan read alignments are counted.

[78] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598278202#RNA-SeqdenovoAssembly-figure6
[79] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598278202#RNA-SeqdenovoAssembly-figure7

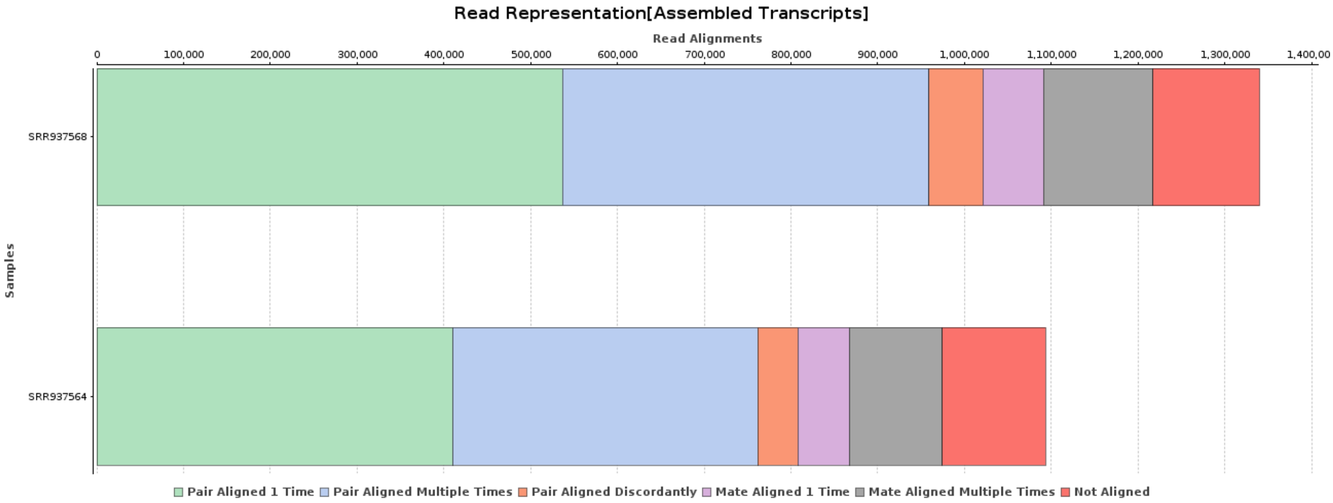


Figure 6: Read Representation Chart

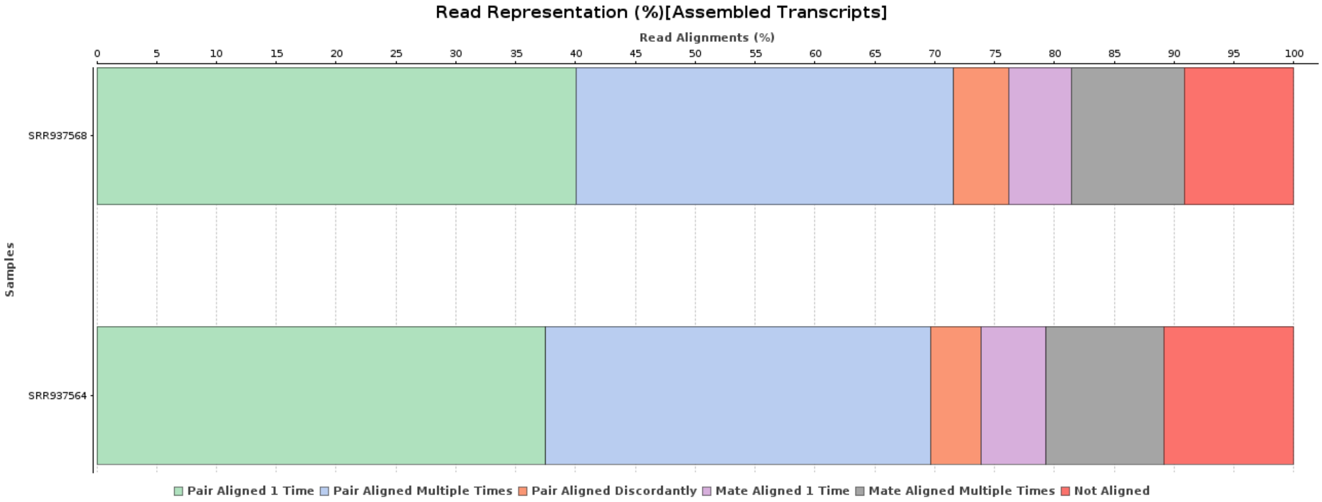


Figure 7: Read Representation (%) Chart

9.2 Completeness Assessment

Content of this page:

9.2.1 Introduction

The Completeness Assessment functionality provides quantitative measures for the assessment of transcriptome assembly completeness, based on evolutionarily-informed expectations of gene content from Benchmarking Universal Single-Copy Orthologs[80] (BUSCO) selected from OrthoDB[81].

The Benchmarking Universal Single-Copy Orthologs are ideal for such quantifications of completeness, as the expectations for these genes to be found in a genome/transcriptome in single-copy are evolutionarily strong.

The application offers predefined BUSCO sets for six major phylogenetic clades. Sampling hundreds of genomes, orthologous groups with single-copy orthologs in >90% of species were selected. Importantly, this threshold accommodates the fact that even well-conserved genes can be lost in some lineages, as well as allowing for incomplete gene annotations and rare gene duplications.

Please cite BUSCO and OrthoDB as:

Simao FA., Waterhouse RM., Ioannidis P., Kriventseva EV. and Zdobnov EM. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics (Oxford, England), 31(19), 3210-2.[82]

Kriventseva EV., Kuznetsov D., Tegenfeldt F., Manni M., Dias R., Simao FA. and Zdobnov EM. (2019). OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. Nucleic acids research, 47(D1), D807-D811.[83]

9.2.2 Run Completeness Assessment

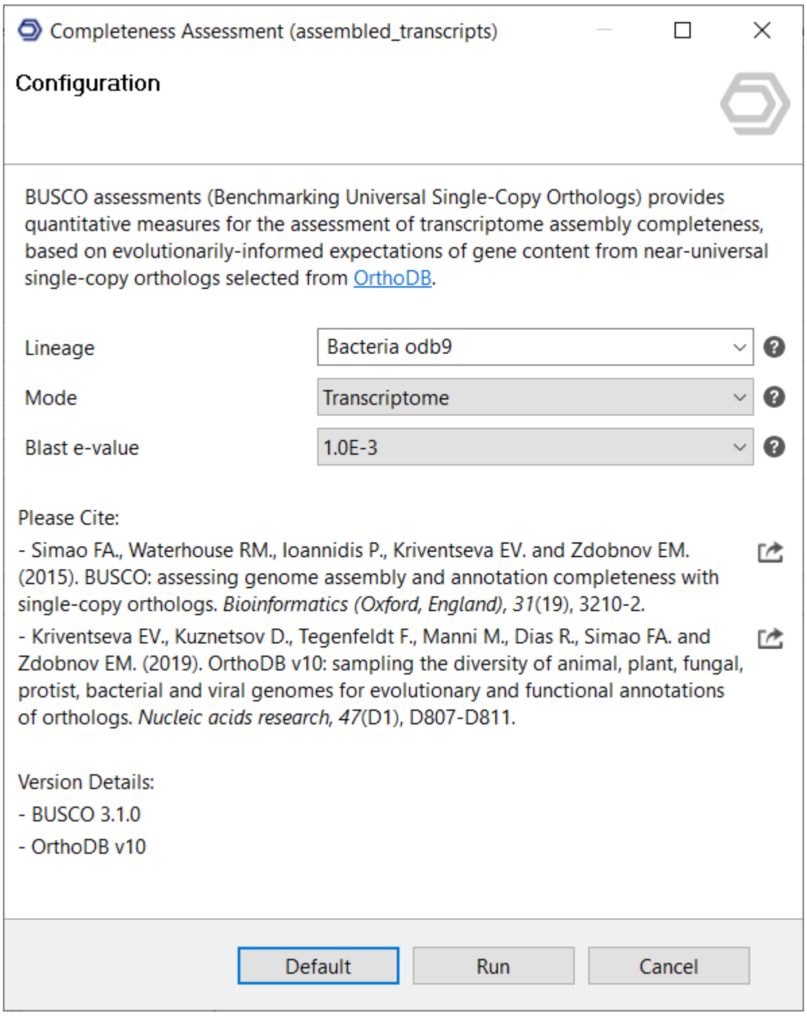
This functionality can be found under **Transcriptomics → Assembly → Completeness Assessment.** The wizard allows to select input files and adjust analysis parameters (Figure 1(see page 154)).

- **Lineage:** Choose the appropriate lineage-specific profile to classify matches, depending on the species to be assessed. Genes that make up the BUSCO sets for each major lineage are selected from orthologous groups with genes present as single-copy orthologs in at least 90% of species (Table 1(see page 154)).
- **Mode:** Set the assessment mode according to the type of sequences to be analyzed.
    - Transcriptome: nucleotide sequences (e.g. transcriptome *de novo* assembly).
    - Proteome: Protein amino acid sequences.
- **Blast e-Value:** The statistical significance threshold for reporting matches against a sequence database. If the statistical significance of alignment is greater than the e-Value threshold, this hit will not be reported. Lower e-Value thresholds are more stringent, leading to fewer results. Increasing the threshold shows less stringent matches. The default e-Value used by BUSCO is 1e-03.

[80] https://busco.ezlab.org/
[81] https://www.orthodb.org/
[82] https://www.ncbi.nlm.nih.gov/pubmed/26059717
[83] https://www.ncbi.nlm.nih.gov/pubmed/30395283

Bacteria	**Eukaryota**	
Bacteria	Eukaryota	Insecta
Proteobacteria	Fungi	Endopterygota
Rhizobiales	Dicrosporidia	Hymenoptera
Betaproteobacteria	Dikarya	Vertebrata
Gammaproteobacteria	Ascomycota	Actinopterygii
Enterobacteriales	Pezizomycotina	Tetrapoda
Delta+ epsilon proteobacteria	Eurotiomycetes	Aves
Actinobacteria	Sordariomyceta	Mammalia
Cyanobacteria	Saccharomyceta	Euarchontoglires
Firmicutes	Saccharomycetales	Laurasiatheria
Clostridia	Basidiomycota	Embryophyta
Lactobacillales	Metazoa	Protists
Bacillales	Nematoda	Aveolata stramenophiles
Bacteroidetes	Arthropoda	
Spirochaetes		
Tenericutes		

Table 1: Lineages



Completeness Assessment (assembled_transcripts)

Configuration

BUSCO assessments (Benchmarking Universal Single-Copy Orthologs) provides quantitative measures for the assessment of transcriptome assembly completeness, based on evolutionarily-informed expectations of gene content from near-universal single-copy orthologs selected from OrthoDB.

Lineage: Bacteria odb9

Mode: Transcriptome

Blast e-value: 1.0E-3

Please Cite:

- Simao FA., Waterhouse RM., Ioannidis P., Kriventseva EV. and Zdobnov EM. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics (Oxford, England), 31*(19), 3210-2.

- Kriventseva EV., Kuznetsov D., Tegenfeldt F., Manni M., Dias R., Simao FA. and Zdobnov EM. (2019). OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic acids research, 47*(D1), D807-D811.

Version Details:

- BUSCO 3.1.0

- OrthoDB v10

Figure 1: Configuration Wizard Page

9.2.3 Results

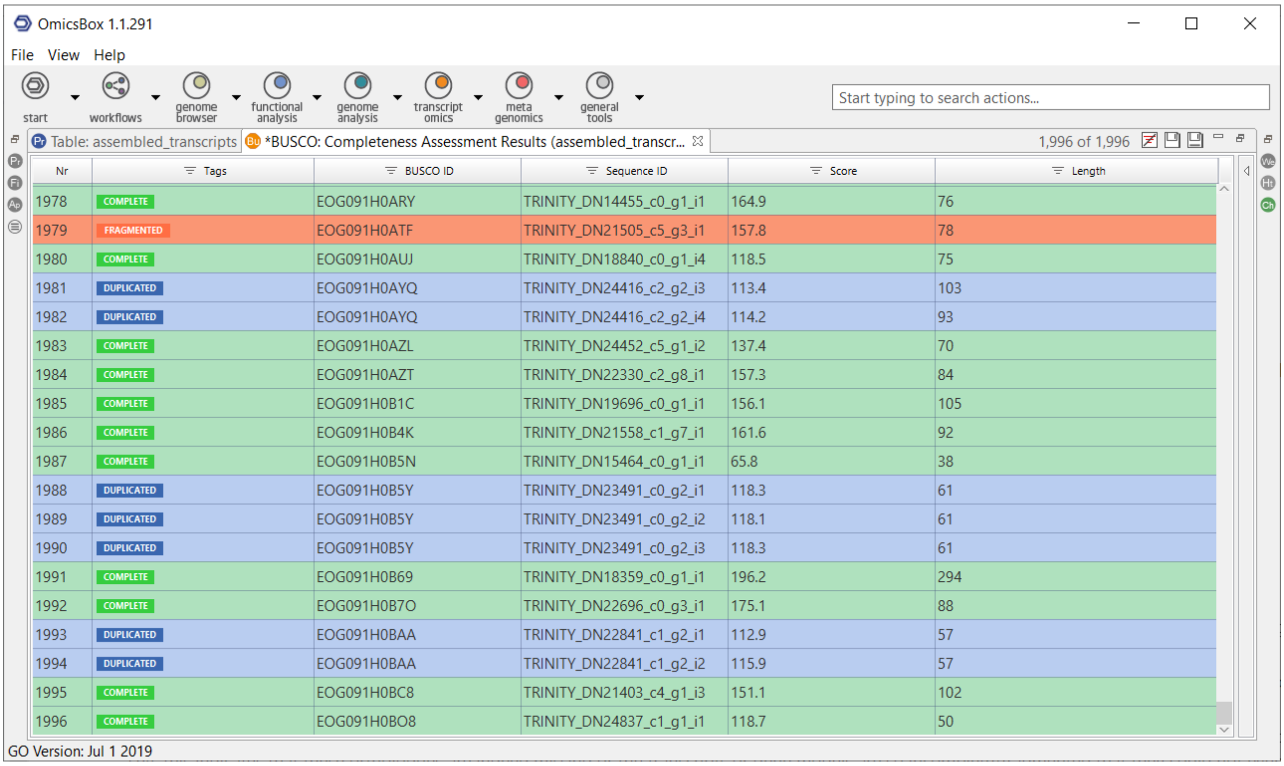
Once finished, a new tab is opened containing the results of the completeness assessment procedure (Figure 2(see page 157)). Each row corresponds to a BUSCO from the lineage database selected, and columns show the following information:

- BUSCO ID: Name of the BUSCO.
- Sequence ID: Name of the transcript/protein sequence matching the BUSCO.
- Score: Score of the alignment.
- Length: Length of the transcript/protein sequence matching the BUSCO.
- Tag: Result category.

The results are simplified into categories of Complete and single-copy, Complete and duplicated, Fragmented, or Missing BUSCOs:

- **Complete (single and duplicated):** The BUSCO matches have scored within the expected range of scores and within the expected range of length alignments to the BUSCO profile.

- **Fragmented:** The BUSCO matches have scored within the range of scores but not within the range of length alignments to the BUSCO profile. For transcriptomes or annotated gene sets, this indicates incomplete transcripts or gene models.
- **Missing:** There were either no significant matches at all, or the BUSCO matches scored below the range of scores for the BUSCO profile. For transcriptomes or annotated gene sets this indicates that these orthologous are indeed missing or the transcripts or gene models are so incomplete/fragmented that they could not even meet the criteria to be considered as fragmented.



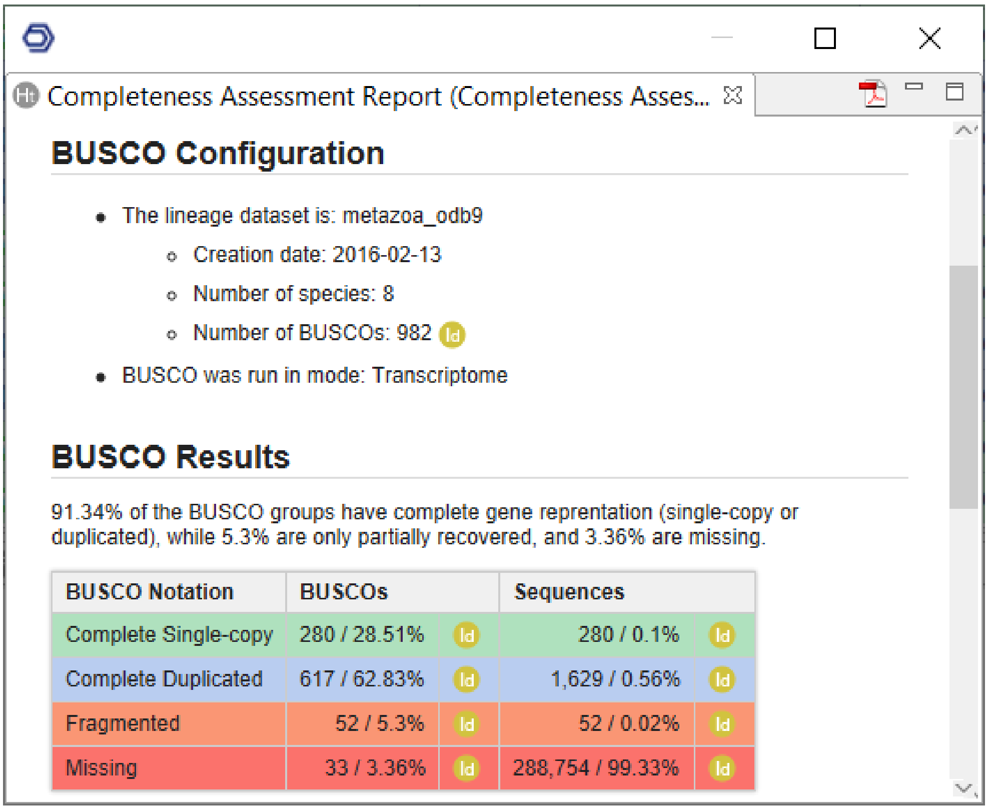
Nr	Tags	BUSCO ID	Sequence ID	Score	Length
1978	COMPLETE	EOG091H0ARY	TRINITY_DN14455_c0_g1_i1	164.9	76
1979	FRAGMENTED	EOG091H0ATF	TRINITY_DN21505_c5_g3_i1	157.8	78
1980	COMPLETE	EOG091H0AUJ	TRINITY_DN18840_c0_g1_i4	118.5	75
1981	DUPLICATED	EOG091H0AYQ	TRINITY_DN24416_c2_g2_i3	113.4	103
1982	DUPLICATED	EOG091H0AYQ	TRINITY_DN24416_c2_g2_i4	114.2	93
1983	COMPLETE	EOG091H0AZL	TRINITY_DN24452_c5_g1_i2	137.4	70
1984	COMPLETE	EOG091H0AZT	TRINITY_DN22330_c2_g8_i1	157.3	84
1985	COMPLETE	EOG091H0B1C	TRINITY_DN19696_c0_g1_i1	156.1	105
1986	COMPLETE	EOG091H0B4K	TRINITY_DN21558_c1_g7_i1	161.6	92
1987	COMPLETE	EOG091H0B5N	TRINITY_DN15464_c0_g1_i1	65.8	38
1988	DUPLICATED	EOG091H0B5Y	TRINITY_DN23491_c0_g2_i1	118.3	61
1989	DUPLICATED	EOG091H0B5Y	TRINITY_DN23491_c0_g2_i2	118.1	61
1990	DUPLICATED	EOG091H0B5Y	TRINITY_DN23491_c0_g2_i3	118.3	61
1991	COMPLETE	EOG091H0B69	TRINITY_DN18359_c0_g1_i1	196.2	294
1992	COMPLETE	EOG091H0B7O	TRINITY_DN22696_c0_g3_i1	175.1	88
1993	DUPLICATED	EOG091H0BAA	TRINITY_DN22841_c1_g2_i1	112.9	57
1994	DUPLICATED	EOG091H0BAA	TRINITY_DN22841_c1_g2_i2	115.9	57
1995	COMPLETE	EOG091H0BC8	TRINITY_DN21403_c4_g1_i3	151.1	102
1996	COMPLETE	EOG091H0BO8	TRINITY_DN24837_c1_g1_i1	118.7	50

GO Version: Jul 1 2019

Figure 2: BUSCO Project

A result page will show a summary of the "Completeness Assessment" results (Figure 3(see page 157)). This page provides a quick evaluation of the results and provides ID lists containing BUSCO or transcript/protein identifiers assigned to the different categories. The result summary can be generated via **Side Panel → Completeness Assessment Report.**

Furthermore, the Completeness Assessment Summary chart (Figure 4(see page 157)) shows the percentage of lineage-specific BUSCOs assigned to each category. The pie chart can be generated via **Side Panel → Completeness Assessment Summary.**



BUSCO Configuration

- The lineage dataset is: metazoa_odb9
  - Creation date: 2016-02-13
  - Number of species: 8
  - Number of BUSCOs: 982 Id
- BUSCO was run in mode: Transcriptome

BUSCO Results

91.34% of the BUSCO groups have complete gene reprentation (single-copy or duplicated), while 5.3% are only partially recovered, and 3.36% are missing.

BUSCO Notation	BUSCOs		Sequences	
Complete Single-copy	280 / 28.51%	Id	280 / 0.1%	Id
Complete Duplicated	617 / 62.83%	Id	1,629 / 0.56%	Id
Fragmented	52 / 5.3%	Id	52 / 0.02%	Id
Missing	33 / 3.36%	Id	288,754 / 99.33%	Id

Figure 3: Completeness Assessment Report

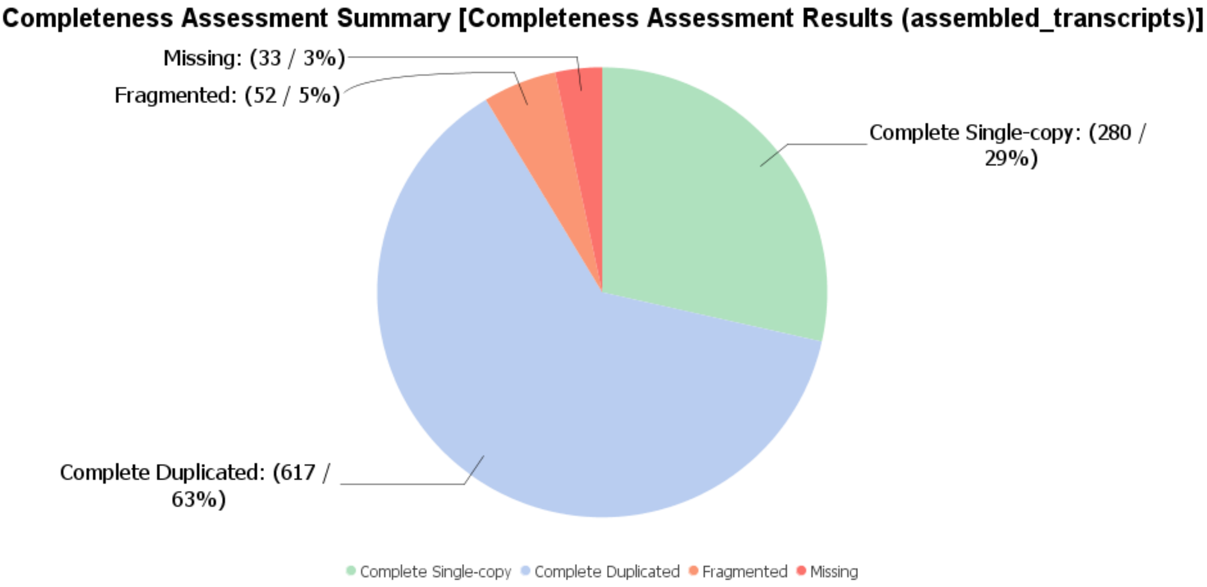
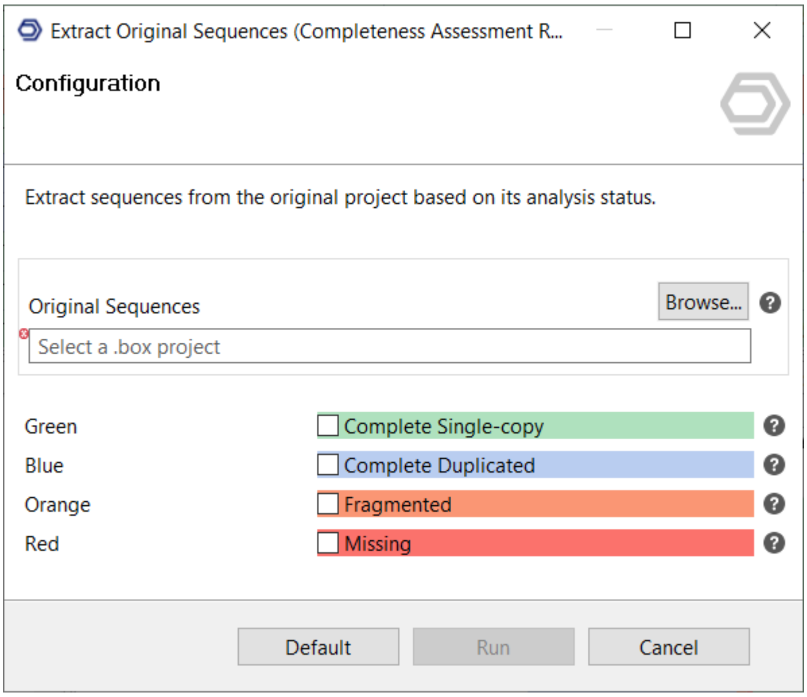


Figure 4: Completeness Assessment Summary Chart

Finally, the **Extract Original Sequences** utility (side bar) allows to extract sequences from the original project based on its analysis status (Figure 5(see page 158)). For this, the original project containing the sequences that were assessed should be provided.



Extract Original Sequences (Completeness Assessment R...

Configuration

Extract sequences from the original project based on its analysis status.

Original Sequences Browse...

Select a .box project

Green	□ Complete Single-copy
Blue	□ Complete Duplicated
Orange	□ Fragmented
Red	□ Missing

Figure 5: Extract Original Sequences

9.3 Predict Coding Regions

Content of this page:

9.3.1 Introduction

The Predict Coding Regions functionality detects candidate coding regions within transcript sequences, such as those generated by *de novo* RNA-Seq transcript assembly. It is based on TransDecoder[84], a pipeline that recognizes likely coding sequences based on the following criteria:

- A minimum length open reading frame (ORF) is found in a transcript sequence.
- A log-likelihood score is computed and it should be > 0.
- The above coding score is higher when the ORF is scored in the 1st reading frame as compared to scores in the other 2 forward reading frames.
- If a candidate ORF is found fully encapsulated by the coordinates of another candidate ORF, the longer one is reported. However, a single transcript can report multiple ORFs (allowing for operons, chimeras, etc).

[84] https://github.com/TransDecoder/TransDecoder/wiki

- A Position-Specific Scoring Matrix (PSSM) is built, trained and used to refine the start codon prediction.
- The putative peptide has a match to a Pfam domain above the noise cut-off score (optional).

Please cite TransDecoder as:

- Haas BJ et al. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature protocols, 8(8), 1494-512.[85]
- TransDecoder 5.5.0. Haas, BJ. and Papanicolaou, A. 2019. https://github.com/TransDecoder/TransDecoder/wiki.

9.3.2 Run Predict Coding Regions

This functionality can be found under **Transcriptomics → Assembly → Predict Coding Regions.** The wizard allows to select input files and adjust analysis parameters (Figure 1).

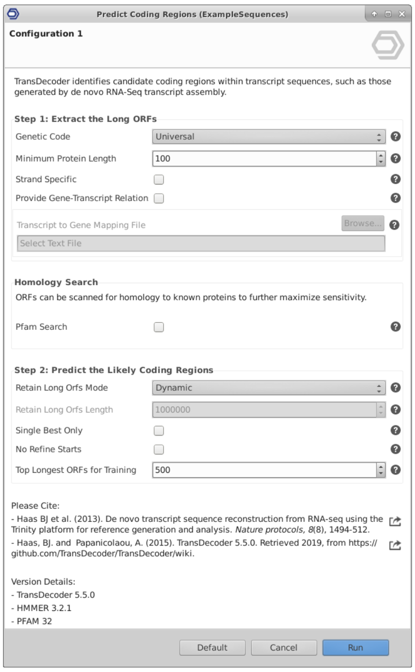
- **Genetic Code:** Select the genetic code of the organism under study. The available genetic codes are:

Universal	Mitochondrial Invertebrates
Acetabularia	Mitochondrial Protozoan
Candida	Mitochondrial Pterobranchia
Ciliate	Mitochondrial Scenedesmus obliquus
Dasycladacean	Mitochondrial Thraustochytrium
Euplotid	Mitochondrial Trematode
Hexamita	Mitochondrial Vertebrates
Mesodinium	Mitochondrial Yeast
Mitochondrial Ascidian	Pachysolen tannophilus
Mitochondrial Chlorophycean	Peritrich
Mitochondrial Echinoderm	SR1 Gracilibacteria
Mitochondrial Flatworm	Tetrahymena

- **Minimum Protein Length:** Minimum protein length to retain coding regions.

[85] https://www.ncbi.nlm.nih.gov/pubmed/23845962

- **Strand Specific:** Only the top strand option is analyzed.
- **Provide Gene-Transcript relation:** Provide a tab-delimited file with information to map from transcript (isoform) IDs to gene IDs. Each line should be of the form: Gene ID[tab]Transcript ID.
- **Pfam Search:** Identify ORFs with homology to known proteins via Pfam searches. Searching PFAM allows to identify common protein domains, that are included as ORF retention criteria. Note that this option will significantly increase the execution time.
- **Retain Long Orfs Mode:** Select the retain long ORFs strategi. The dynamic mode, sets range according to 1% FDR in random sequence of same GC content. Under the strict mode, all ORFs found that are equal or longer to the Retain Long ORFs Length are kept, even if no other evidence marks it as coding.
- **Retain Long Orfs Length:** Select the minimum length to retain ORFs under the strict mode.
- **Single Best Only:** Retain only the single best ORF per transcript (prioritized by homology, then ORF length).
- **No Refine Starts:** By default, the predict coding regions strategy identifies potential start codons for 5’ partial ORFs using a PWM (position weight matrix). Check this option to deactivate this process.
- **Top Longest ORF for Training:** Top longest ORFs to train Markov Model (hexamer stats). The default value is 500. Note, 10X this value are first selected for removing redundancies, and then this value of longest ORFs are selected from the non-redundant set.



Configuration 1

TransDecoder identifies candidate coding regions within transcript sequences, such as those generated by de novo RNA-Seq transcript assembly.

Step 1: Extract the Long ORFs

Genetic Code: Universal
Minimum Protein Length: 100
Strand Specific ☐
Provide Gene-Transcript Relation ☐
Transcript to Gene Mapping File: Select Text File

Homology Search

ORFs can be scanned for homology to known proteins to further maximize sensitivity.

Pfam Search ☐

Step 2: Predict the Likely Coding Regions

Retain Long Orfs Mode: Dynamic
Retain Long Orfs Length: 1000000
Single Best Only ☐
No Refine Starts ☐
Top Longest ORFs for Training: 500

Please Cite:
- Haas BJ et al. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature protocols*, *8*(8), 1494-512.
- Haas, BJ. and Papanicolaou, A. (2015). TransDecoder 5.5.0. Retrieved 2019, from https://github.com/TransDecoder/TransDecoder/wiki.

Version Details:
- TransDecoder 5.5.0
- HMMER 3.2.1
- PFAM 32

Figure 1: Configuration Wizard Page

9.3.3 Results

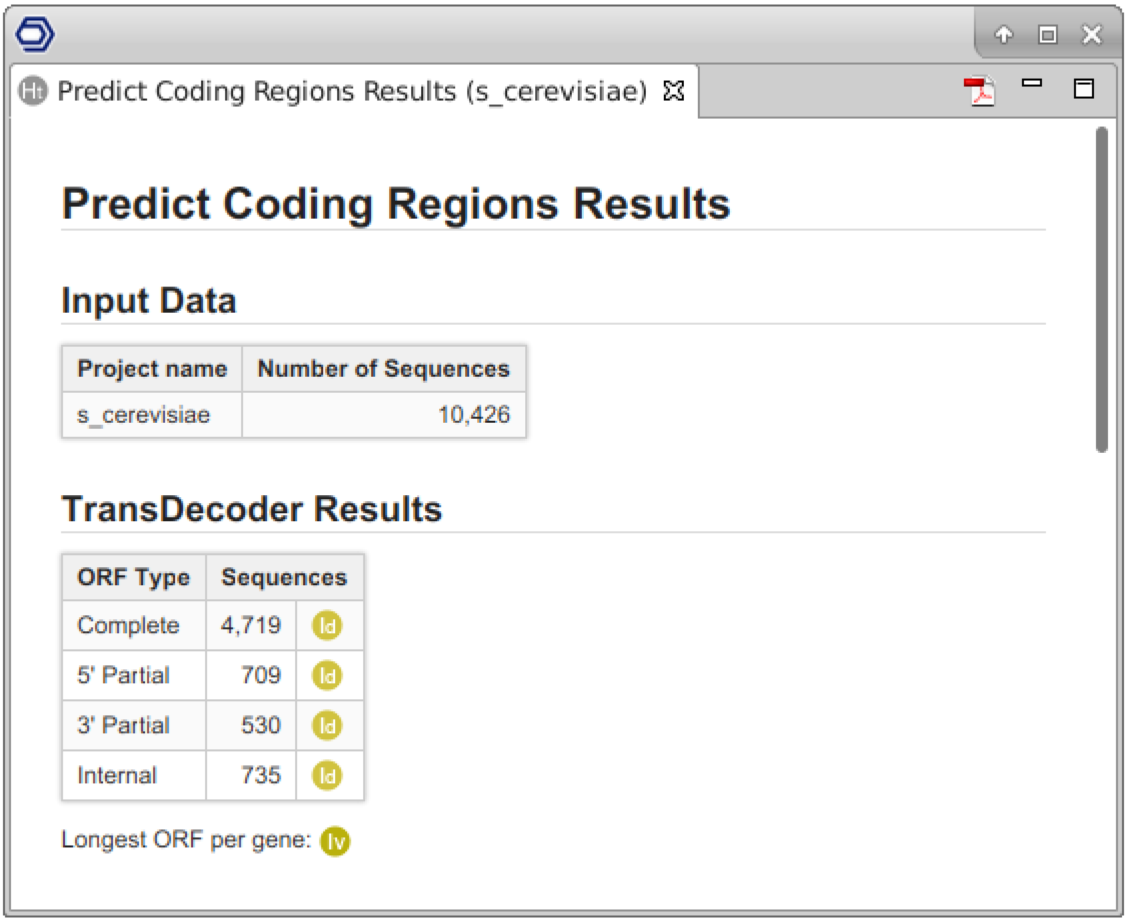
Once finished, results are returned in three projects:

- Protein sequences: A sequence table that contains peptide sequences for the final candidate ORFs.
- CDS sequences: A sequence table that contains nucleotide sequences for coding regions of the final candidate ORFs.
- ORFs Coordinates: A GFF project that contains positions within the target transcripts of the final selected ORFs.

Note that in both sequence projects, CDSs and proteins, the description field contains details about the predicted ORF. This description includes:

- The protein identifier composed of the original transcripts along with '|m.(number)'.
- The type attribute indicates whether the protein is:
  - **Complete:** Contains a start and a stop codon.
  - **5' partial**: It is missing a start codon and presumably part of the N-terminus.
  - **3' partial**: It is missing the stop codon and presumably part of the C-terminus.
  - **Internal**: It is both 5' and 3' partial.
- An indicator (+) or (-) to indicate in which strand the coding region was found, along with the coordinates of the ORF in that transcript sequence.

In addition, a result page will show a summary of the "Predict Coding Regions" results (Figure 2). This page provides a quick evaluation of the results and provides ID lists containing transcript identifiers assigned to the different categories.



Predict Coding Regions Results

Input Data

Project name	Number of Sequences
s_cerevisiae	10,426

TransDecoder Results

ORF Type	Sequences	
Complete	4,719	Id
5' Partial	709	Id
3' Partial	530	Id
Internal	735	Id

Longest ORF per gene: lv

Figure 2: Predict Coding Regions Report

Furthermore, the Predict Coding Regions Summary chart (Figure 3) shows the percentage of ORFs that have been predicted as Complete, 5' Partial, 3' Partial and Internal.

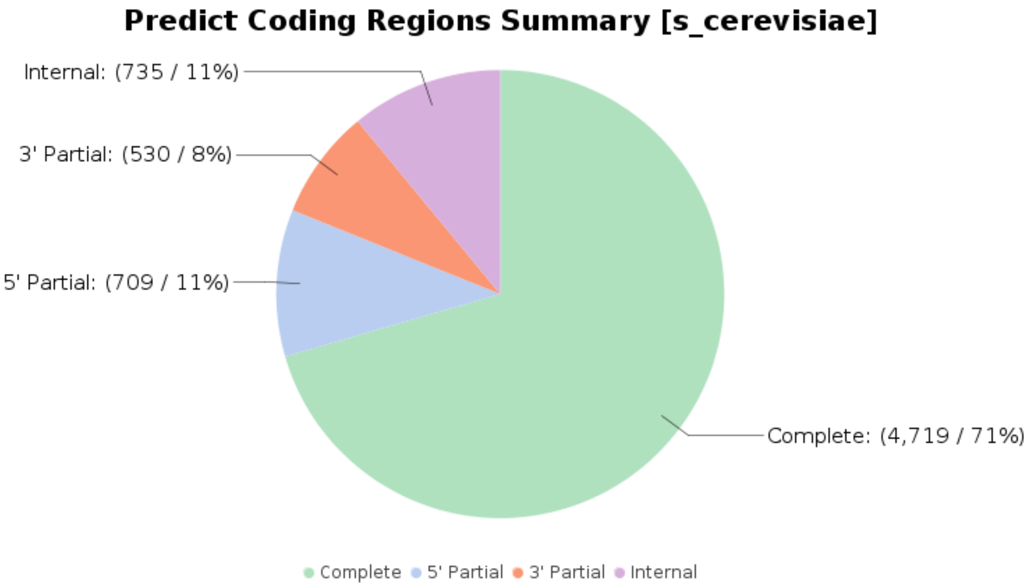


Figure 3: Predict Coding Regions Summary

9.4 RNA-Seq Alignment

Content of this page:

9.4.1 Introduction

Read alignment is a common process applied to high-throughput sequencing data, being one of the first stages required for many different types of analysis. In the RNA-Seq scenario, this process is used to quantify gene expression. The goal of the read alignment is to map short sequencing reads efficiently to a large reference genome to identify the 'correct' genomic loci from which the read originated whilst taking into account errors in the sequence reads. The RNA-Seq Alignment functionality of OmicsBox is based on STAR[86] (**Spliced Transcript Alignment to a Reference**).

STAR is a fast RNA-Seq read mapper, with support for splice-junction and fusion read detection, and it was designed to align non-contiguous sequences directly to a reference genome. STAR aligns reads by finding maximal mappable prefix hits between reads (or read pairs) and the genome, using a suffix array index strategy. Different parts of a read can be mapped to different genomic positions, corresponding to splicing or

[86] https://github.com/alexdobin/STAR

RNA-fusions. If genomic annotations are provided, the genome index includes known splice-junctions from annotated gene models, allowing for sensitive detection of spliced reads.

The binary nature of the suffix array search results in a favorable logarithmic scaling for the search time with the reference genome length. Since this approach has high memory requirements, it is executed in the Omics Cloud, which provides a high-performance computing environment, allowing fast alignments even against large genomes.

Please cite STAR as:

Dobin A, Davis CA, Schlesinger F, et al (2012). "STAR: ultrafast universal RNA-seq aligner." Bioinformatics, 29(1):15-21.[87]

9.4.2 Run RNA-Seq Alignment

This functionality can be found under **Transcriptomics → RNA-Seq Alignment.** The wizard allows to select input files and adjust analysis parameters (Figure 1[88], Figure 2[89] and Figure 3[90]).

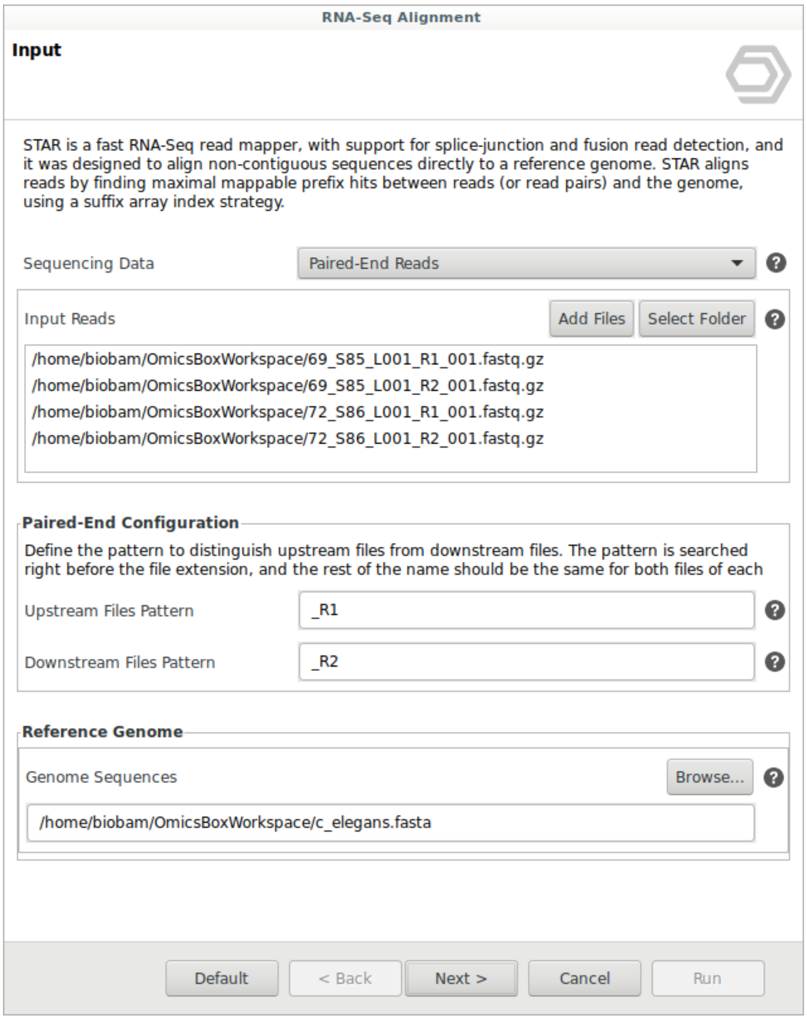
9.4.2.1 Input

- **Sequencing Data:** Choose single-end or paired-end reads. Note that if the paired-end option is selected, two files per sample are required.
- **Input Reads:** Provide the files containing RNA sequencing reads. These files are assumed to be in FASTQ format or compressed FASTQ format (.gz).
- **Paired-end configuration:** In case of paired-end reads, a pattern to distinguish upstream files from downstream files is required. The provided patterns are searched in the filenames right before the extension. The beginning of the filenames should be the same for both files of each sample.
  - **Upstream Files Pattern**: Establish the pattern to recognize upstream FASTQ files.
  - **Downstream Files Pattern**: Establish the pattern to recognize downstream FASTQ files.

  > ⚠ For example, if the upstream file is SRR037717_1.fastq and the downstream SRR037717_2.fastq, "_1" should be established as the upstream pattern and "_2" as the downstream pattern.

- **Genome Sequences:** Specify a FASTA file that contains the genome reference sequences. Multiple reference sequences, e.g. chromosomes or scaffolds, can be provided. It is strongly recommended to include major chromosomes as well as un-placed and un-localized scaffolds since a substantial number of reads may map to these scaffolds (e.g. ribosomal RNA). These reads would be reported as unmapped if the scaffolds are not included, or may be aligned to wrong loci on the chromosomes. On the other hand, patches and alternative haplotypes should not be included in the genome.

[87] https://www.ncbi.nlm.nih.gov/pubmed/23104886
[88] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=618856449#RNA-SeqAlignment-figure1
[89] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=618856449#RNA-SeqAlignment-figure2
[90] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=618856449#RNA-SeqAlignment-figure3



RNA-Seq Alignment

Input

STAR is a fast RNA-Seq read mapper, with support for splice-junction and fusion read detection, and it was designed to align non-contiguous sequences directly to a reference genome. STAR aligns reads by finding maximal mappable prefix hits between reads (or read pairs) and the genome, using a suffix array index strategy.

Sequencing Data: Paired-End Reads

Input Reads

/home/biobam/OmicsBoxWorkspace/69_S85_L001_R1_001.fastq.gz
/home/biobam/OmicsBoxWorkspace/69_S85_L001_R2_001.fastq.gz
/home/biobam/OmicsBoxWorkspace/72_S86_L001_R1_001.fastq.gz
/home/biobam/OmicsBoxWorkspace/72_S86_L001_R2_001.fastq.gz

Paired-End Configuration

Define the pattern to distinguish upstream files from downstream files. The pattern is searched right before the file extension, and the rest of the name should be the same for both files of each

Upstream Files Pattern: _R1

Downstream Files Pattern: _R2

Reference Genome

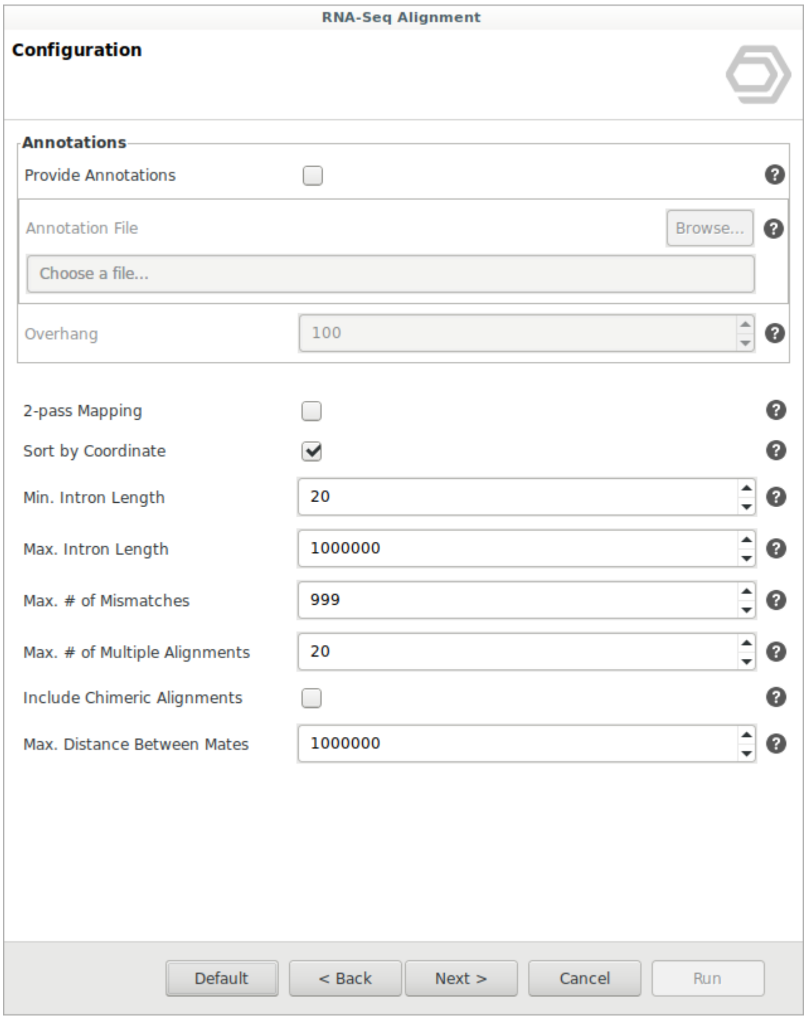
Genome Sequences: /home/biobam/OmicsBoxWorkspace/c_elegans.fasta

Figure 1: Input Data Page

9.4.2.2 Configuration

- **Provide annotations:** This option allows providing a file with annotated genes and transcripts in GTF/GFF format (GTF is recommended). The aligner will extract splice junctions from this file and use them to improve the accuracy of the alignment. While this is optional, using annotations is highly recommended. Chromosome names in the GTF annotation file have to match chromosome names in the FASTA genome sequences file.
  - **Annotation File**: Select the file containing the annotated genes and transcripts in GTF/GFF format.
  - **Overhang**: Establish the length of the genomic sequence around the annotated junction to be used in constructing the splice junctions database. This length should be equal to the length of the read -1. For instance, for 100 bp paired-end reads, the ideal value is 99. In case of reads with varying length, the ideal value is the maximum read length -1.
- **2-pass Mapping:** This option allows a most sensitive novel junction discovery. The aligner algorithm is executed first to collect the junctions. These junctions are used for a second pass mapping.
- **Sort by Coordinate:** The aligner will output BAM files sorted by coordinates.

- **Minimum Intron Length:** Specify the minimum intron size. A genomic gap is considered intron if its length is equal to or greater than the given value. Otherwise, it is considered a deletion.
- **Maximum Intron Length:** Specify the maximum intron size.
- **Maximum Number of Mismatches:** Set the maximum number of mismatches allowed per read or read pair.
- **Maximum Number of Multiple Alignments:** Establish the maximum number of multiple alignments allowed per read. If exceeded, the read is considered unmapped.
- **Include Chimeric Alignments:** This option allows to include the chimeric alignments together with normal alignments in the main BAM file. The format of chimeric alignments follows the latest SAM/BAM specifications.
- **Maximum Distance Between Mates:** Specify the maximum genomic distance between two mate pairs.



RNA-Seq Alignment

Configuration

Annotations

Provide Annotations □

Annotation File

Choose a file...

Overhang 100

2-pass Mapping □

Sort by Coordinate ☑

Min. Intron Length 20

Max. Intron Length 1000000

Max. # of Mismatches 999

Max. # of Multiple Alignments 20

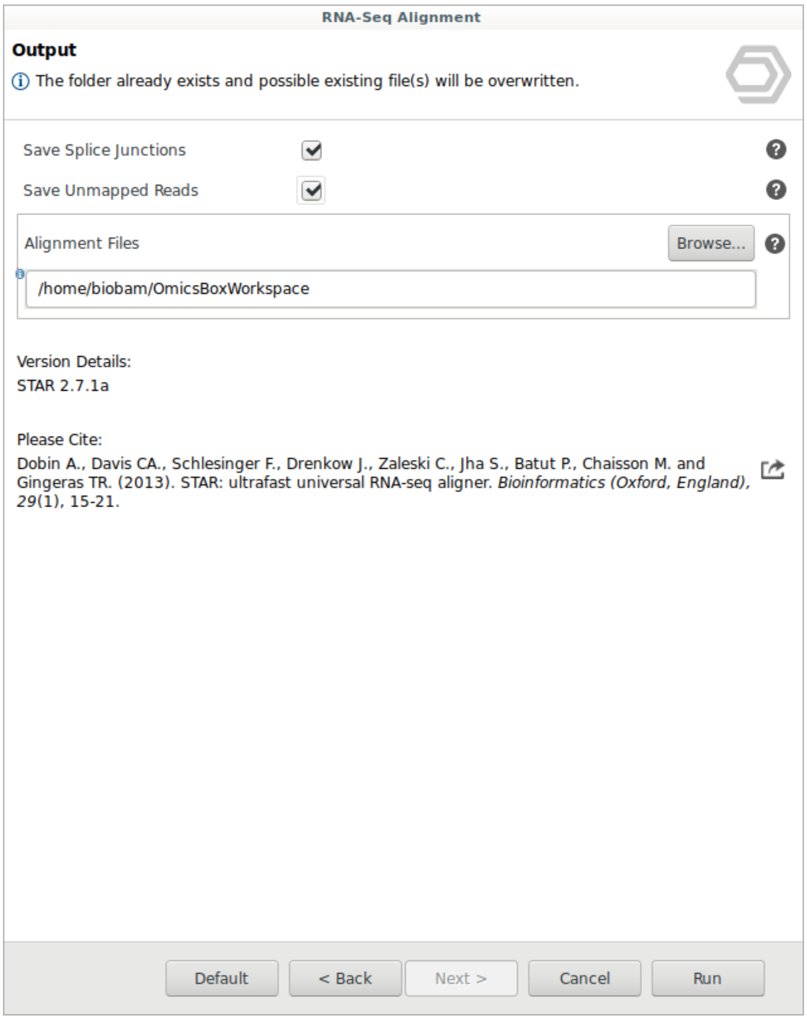
Include Chimeric Alignments □

Max. Distance Between Mates 1000000

Figure 2: Configuration Data Page

9.4.2.3 Output

- **Save Splice Junctions:** Save high confidence collapsed splice junctions in tab-delimited format. Note that STAR defines the junction start/end as intronic bases. Files will be named as sample_name.Sj.out.tab, and will be placed in the "Alignment Files" destination folder.
- **Save Unmapped Reads:** Save unmapped and partially mapped. Files will be named as sample_name_Unmapped.fastq.gz, and will be placed in the "Alignment Files" destination folder.
- **Alignment Files:** Select a folder to save the output BAM files. Take into account that one BAM file will be generated for each input FASTQ sample, so make sure there is enough disk space to store them.



RNA-Seq Alignment

Output

ⓘ The folder already exists and possible existing file(s) will be overwritten.

Save Splice Junctions ☑

Save Unmapped Reads ☑

Alignment Files: /home/biobam/OmicsBoxWorkspace

Version Details:
STAR 2.7.1a

Please Cite:
Dobin A., Davis CA., Schlesinger F., Drenkow J., Zaleski C., Jha S., Batut P., Chaisson M. and Gingeras TR. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)*, *29*(1), 15-21.

Figure 3: Output Data Page

9.4.3 Results

The main outputs are the BAM files (Figure 4(see page 169)). A BAM file (*.bam) is a compressed binary version (BGZF format) of a SAM file that is used to represent aligned sequences. SAM is a TAB-delimited text format consisting of a header section and an alignment section. Header lines start with '@', while alignment lines do not. Each alignment line has 11 mandatory fields for essential alignment information such as the mapping position, and a variable number of optional fields for flexible or aligner specific information:

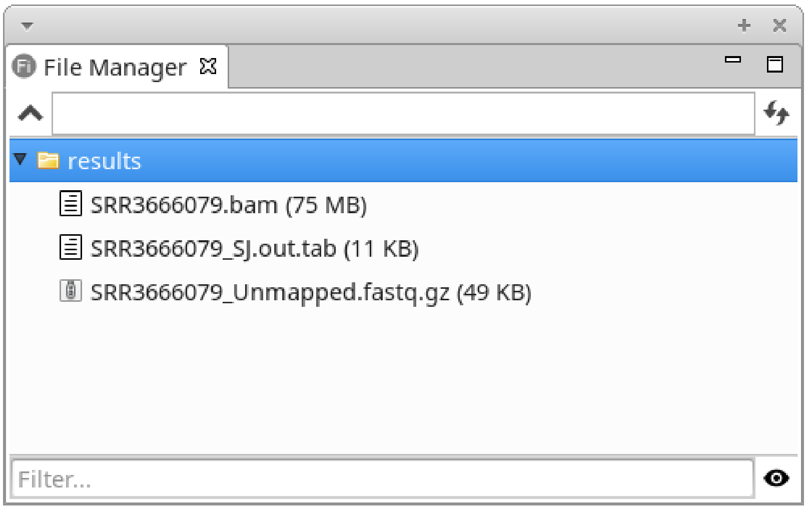
1. **QNAME**: Query template (read) name. In a SAM file, a read may occupy multiple alignment lines, when its alignment is chimeric or when multiple mappings are given.
2. **FLAG**: SAM flags summarize many properties of reads, represented by flag bits, into a single number:
   - Read is paired.
   - Read is mapped in a proper pair.
   - Read is unmapped.
   - Mate is unmapped.
   - Read reverse strand.
   - Mate reverse strand.
   - Read is from the first pair.
   - Read is from the second pair.
   - Alignment isn't primary.
   - Read fails platform/vendor quality checks.
   - Read is PCR or optical duplicate.
3. **RNAME**: Reference sequence name. If @SQ header lines are present, RNAME must be present in one of the SQ-SN tag.
4. **POS**: 1-based leftmost mapping position of the first CIGAR operation. The first base in a reference sequence has coordinate 1.
5. **MAPQ**: Mapping quality. It equals −10 log10 Pr{mapping position is wrong}, rounded to the nearest integer. A value 255 indicates that the mapping quality is not available.
6. **CIGAR**: A string describing how the read aligns with the reference. It consists of one or more components. Each component comprises an operator and the number of bases which the operator applies to. Operators are:
   - M: Align match.
   - I: Insertion to the reference.
   - D: Deletion from the reference.
   - N: Skipped region from the reference.
   - S: Soft clipping.
   - H: Hard clipping.
   - P: Padding (silent deletion from padded reference).
   - =: Sequence match
   - X: Sequence mismatch
7. **RNEXT**: Reference sequence name of the primary alignment of the next read in the template. If all segments are mapped to the same reference, the unsigned observed template length equals the number of bases from the leftmost mapped base to the rightmost mapped base.
8. **PNEXT**: a 1-based position of the primary alignment of the next read in the template.
9. **TLEN**: Signed observed template length.
10. **SEQ**: Segment sequence.
11. **QUAL**: ASCII of base QUALity plus 33 (same as the quality string in the Sanger FASTQ format).

In addition to these 11 obligatory fields, optional fields may be included. All optional fields follow the TAG:TYPE:VALUE format where TAG is a two-character string.

For more information about the SAM format, visit the SAM Format Specification Page[91].

⚠ You can check the meaning of a FLAG number using the SAM Flag Translator[92].

[91] https://samtools.github.io/hts-specs/SAMv1.pdf
[92] https://broadinstitute.github.io/picard/explain-flags.html



File Manager

results

SRR3666079.bam (75 MB)

SRR3666079_SJ.out.tab (11 KB)

SRR3666079_Unmapped.fastq.gz (49 KB)

Figure 4: Output Directory

In addition to the BAM files, splice junctions and unmapped reads are generated if the corresponding options were checked.

Splice junctions are stored in tab-delimited format. The columns have the following meaning:

1. Chromosome.
2. The first base of the intron (1-based).
3. The last base of the intron (1-based).
4. Strand:
   a. 0: undefined.
   b. 1: +.
   c. 2: -.
5. Intron motif:
   a. 0: non-canonical.
   b. 1: GT/AG.
   c. 2: CT/AC.
   d. 3: GC/AG.
   e. 4: CT/GC.
   f. 5: AT/AC.
   g. 6: GT/AT.
6. Annotation:
   a. 0: unannotated.
   b. 1: annotated (only if splice junctions database is used)
7. The number of unique mapping reads crossing the junction.
8. The number of multi-mapping reads crossing the junction.
9. Maximum spliced alignment overhang.

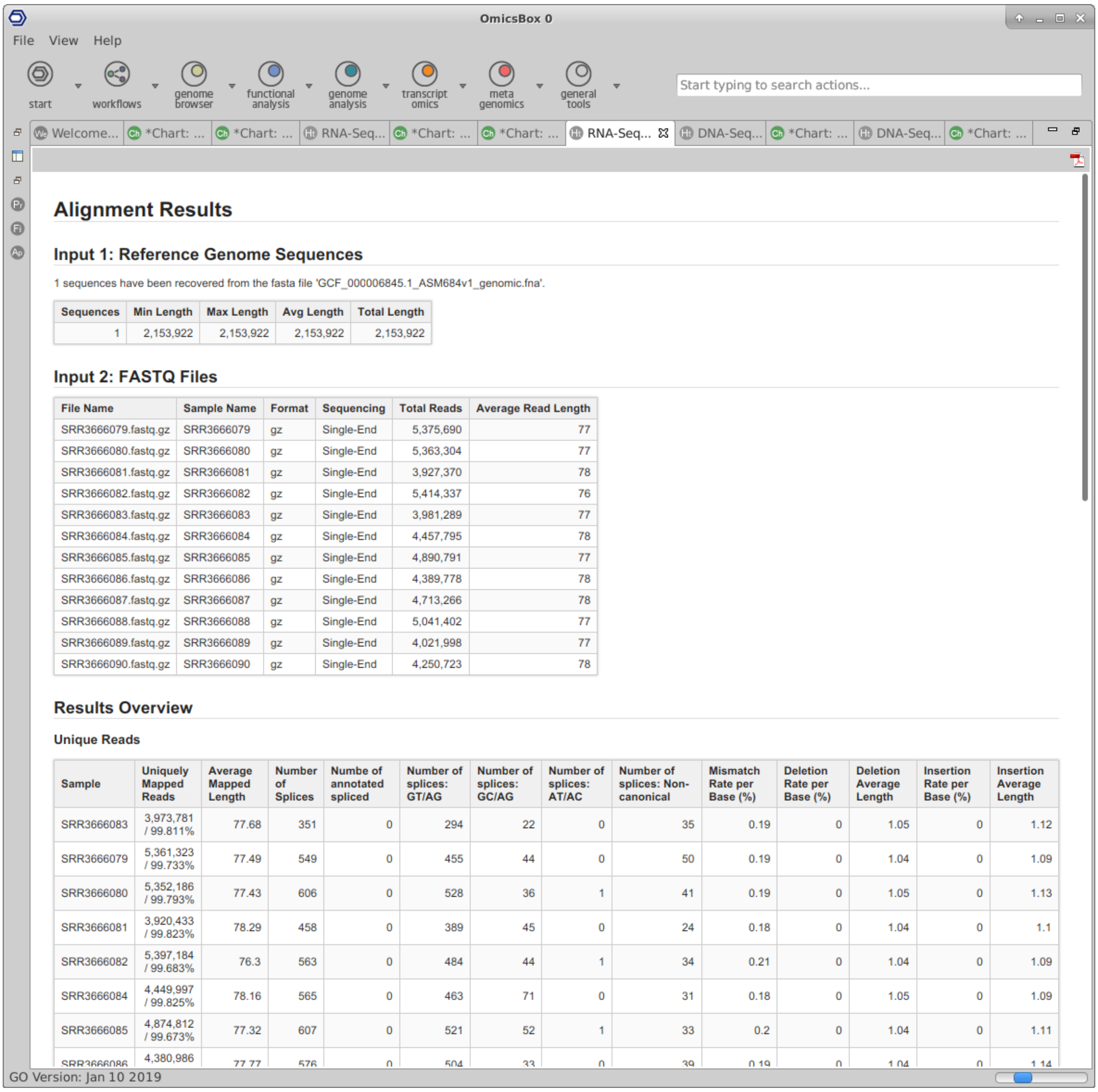
The unmapped files contain the unmapped reads and partially mapped reads (i.e. mapped only one mate of a paired-end read). These reads are stored following the FASTQ specification. Note that if paired-end data were provided, two unmapped files are expected for each sample, one containing upstream unmapped reads and the other containing downstream unmapped reads.

Finally, a report and a chart are generated with complementary information. The report shows a summary of the RNA-Seq Alignment results (Figure 5(see page 169)). This page contains information about the reference genome sequences, the input FASTQ files, and a results overview. The last section is divided into four subsections: unique reads, multi-mapping reads, chimeric reads, and unmapped reads.

The bar chart (Figure 6(see page 170)) shows the number of reads of each input file sorted by different categories, according to how the read was aligned to the reference sequence:

- Unique reads: Reads that have been assigned once to a location of the reference sequence.
- Multi-mapping reads: Reads that have been assign to more than one location of the reference sequence.
- Chimeric reads: Reads that have been aligned to two distinct portions of the reference sequence.
- Unmapped Reads: Reads that have not been assigned to any reference transcript.

Finally, the Genome Browser(see page 23) allows you to visualize genomic coordinates (GFF/GTF) in a side-scrolling way. Several tracks can be added to the browser, the currently supported tracks are VCF, DNA Fasta and BAM. The BAM track shows the reads of a BAM file and if the sequence track is active, it will also highlight the differences between the read sequence and the sequence track.



Alignment Results

Input 1: Reference Genome Sequences

1 sequences have been recovered from the fasta file 'GCF_000006845.1_ASM684v1_genomic.fna'.

Sequences	Min Length	Max Length	Avg Length	Total Length
1	2,153,922	2,153,922	2,153,922	2,153,922

Input 2: FASTQ Files

File Name	Sample Name	Format	Sequencing	Total Reads	Average Read Length
SRR3666079.fastq.gz	SRR3666079	gz	Single-End	5,375,690	77
SRR3666080.fastq.gz	SRR3666080	gz	Single-End	5,363,304	77
SRR3666081.fastq.gz	SRR3666081	gz	Single-End	3,927,370	78
SRR3666082.fastq.gz	SRR3666082	gz	Single-End	5,414,337	76
SRR3666083.fastq.gz	SRR3666083	gz	Single-End	3,981,289	77
SRR3666084.fastq.gz	SRR3666084	gz	Single-End	4,457,795	78
SRR3666085.fastq.gz	SRR3666085	gz	Single-End	4,890,791	77
SRR3666086.fastq.gz	SRR3666086	gz	Single-End	4,389,778	78
SRR3666087.fastq.gz	SRR3666087	gz	Single-End	4,713,266	78
SRR3666088.fastq.gz	SRR3666088	gz	Single-End	5,041,402	77
SRR3666089.fastq.gz	SRR3666089	gz	Single-End	4,021,998	77
SRR3666090.fastq.gz	SRR3666090	gz	Single-End	4,250,723	78

Results Overview

Unique Reads

Sample	Uniquely Mapped Reads	Average Mapped Length	Number of Splices	Numbe of annotated spliced	Number of splices: GT/AG	Number of splices: GC/AG	Number of splices: AT/AC	Number of splices: Non-canonical	Mismatch Rate per Base (%)	Deletion Rate per Base (%)	Deletion Average Length	Insertion Rate per Base (%)	Insertion Average Length
SRR3666083	3,973,781 / 99.811%	77.68	351	0	294	22	0	35	0.19	0	1.05	0	1.12
SRR3666079	5,361,323 / 99.733%	77.49	549	0	455	44	0	50	0.19	0	1.04	0	1.09
SRR3666080	5,352,186 / 99.793%	77.43	606	0	528	36	1	41	0.19	0	1.05	0	1.13
SRR3666081	3,920,433 / 99.823%	78.29	458	0	389	45	0	24	0.18	0	1.04	0	1.1
SRR3666082	5,397,184 / 99.683%	76.3	563	0	484	44	1	34	0.21	0	1.04	0	1.09
SRR3666084	4,449,997 / 99.825%	78.16	565	0	463	71	0	31	0.18	0	1.05	0	1.09
SRR3666085	4,874,812 / 99.673%	77.32	607	0	521	52	1	33	0.2	0	1.04	0	1.11
SRR3666086	4,380,986	77.77	576	0	504	33	0	39	0.19	0	1.04	0	1.14

Figure 5: Summary Report

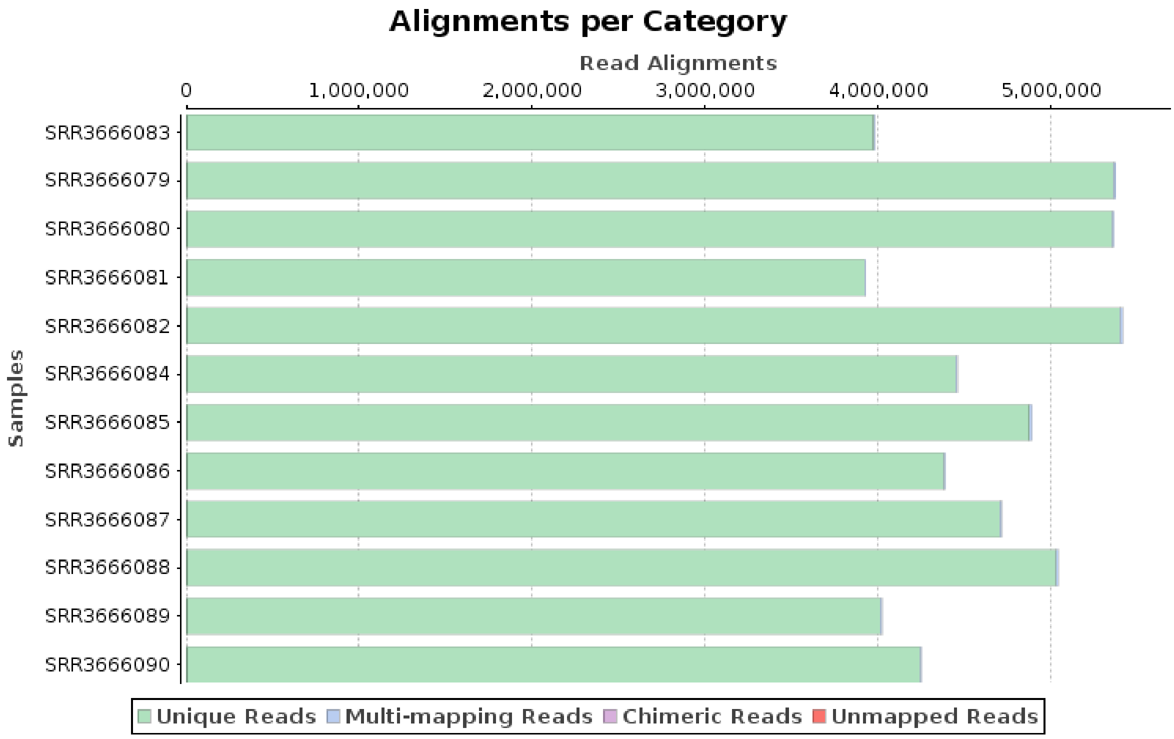


Figure 6: Alignments per Category Chart

9.5 Gene-level Quantification

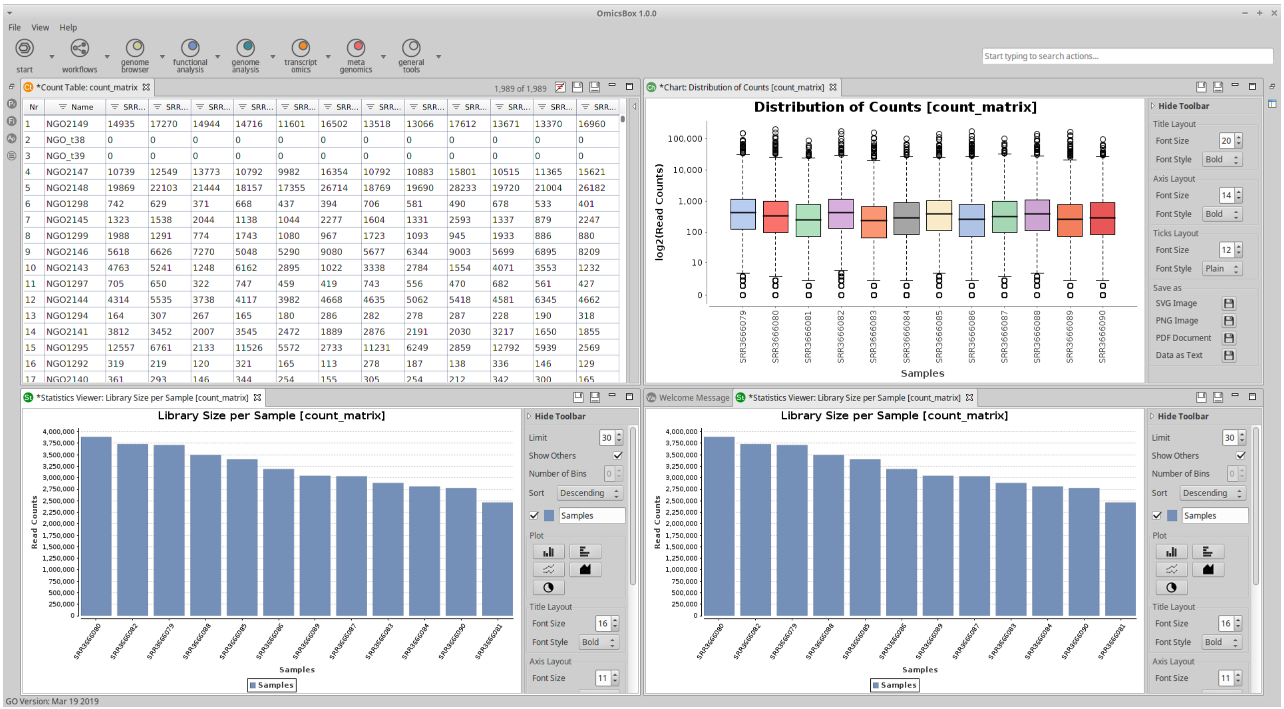
Content of this page:

9.5.1 Introduction

The "Create Count Table" tool is designed to estimate gene expression from RNA-sequencing experiments. This tool expects files with aligned sequencing reads in SAM/BAM format and a GTF/GFF file with coordinates of genomic features. It counts how many reads map to each feature of interest (e.g. genes, exons...). A Count Table is obtained that can be used to perform a differential expression analysis within OmicsBox.

Only reads mapping unambiguously to a single genomic feature are considered. On the other hand, reads aligned to more than one position or overlapping with more than one feature are discarded. This is convenient because if there are two or more genes that overlap or have some sequence similarity but they have different expression levels, counting common reads for all of them could provide inaccurate results. If paired-end data is provided fragments (read-pairs) instead of single reads are counted.

This module is based on the popular HTSeq package[93]. Please cite HTSeq as:
Anders S, Pyl PT and Huber W (2014). "HTSeq — A Python framework to work with high-throughput sequencing data." Bioinformatics, 31(2), 166-9.



Nr	Name	SRR...	SRR...	SRR...	SRR...	SRR...	SRR...	SRR...	SRR...	SRR...	SRR...	SRR...	SRR...
1	NGO2149	14935	17270	14944	14716	11601	16502	13518	13066	17612	13671	13370	16960
2	NGO_t38	0	0	0	0	0	0	0	0	0	0	0	0
3	NGO_t39	0	0	0	0	0	0	0	0	0	0	0	0
4	NGO2147	10739	12549	13773	10792	9982	16354	10792	10883	15801	10515	11365	15621
5	NGO2148	19869	22103	21444	18157	17355	26714	18769	19690	28233	19720	21004	26182
6	NGO1298	742	629	371	668	437	394	706	581	490	678	533	401
7	NGO2145	1323	1538	2044	1138	1044	2277	1604	1331	2593	1337	879	2247
8	NGO1299	1988	1291	774	1743	1080	967	1723	1093	945	1933	886	880
9	NGO2146	5618	6626	7270	5048	5290	9080	5677	6344	9003	5699	6895	8209
10	NGO2143	4763	5241	1248	6162	2895	1022	3338	2784	1554	4071	3553	1232
11	NGO1297	705	650	322	747	459	419	743	556	470	682	561	427
12	NGO2144	4314	5535	3738	4117	3982	4668	4635	5062	5418	4581	6345	4662
13	NGO1294	164	307	267	165	180	286	282	278	287	228	190	318
14	NGO2141	3812	3452	2007	3545	2472	1889	2876	2191	2030	3217	1650	1855
15	NGO1295	12557	6761	2133	11526	5572	2733	11231	6249	2859	12792	5939	2569
16	NGO1292	319	219	120	321	165	113	278	187	138	336	146	129
17	NGO2140	361	293	146	344	254	155	305	254	212	342	300	165

Title Layout
Font Size 20
Font Style Bold
Axis Layout
Font Size 14
Font Style Bold
Ticks Layout
Font Size 12
Font Style Plain
Save as
SVG Image
PNG Image
PDF Document
Data as Text

Limit 30
Show Others ☑
Number of Bins 0
Sort Descending
☑ Samples
Title Layout
Font Size 16
Font Style Bold
Axis Layout
Font Size 11

Figure 1: Create Count Table Interface

9.5.2 Run Create Count Table

This functionality can be found under **Transcriptomics → Create Count Table → Gene Level Quantification**. The wizard allows to select input files and adjust analysis parameters (figure 2[94] and figure 3[95]).

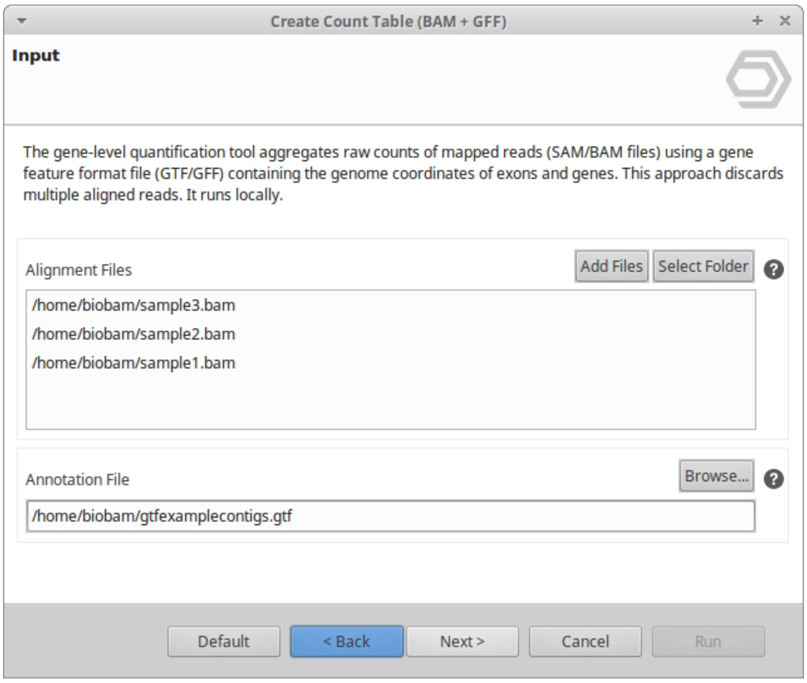
9.5.2.1 Input

- **Alignment Files:** Select files containing the sequencing alignment data. It must be in the "Sequence Alignment/Map" format (SAM) or in its compressed format (BAM).

[93] http://htseq.readthedocs.io/en/release_0.9.1/
[94] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598114455#Gene-levelQuantification-figure2
[95] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598114455#Gene-levelQuantification-figure3

- **Annotation File:** Select the file containing the list of genomic features in GFF/GTF format. GFF objects from OmicsBox are accepted too.

⚠ The GFF/GTF must belong to the same genome as the one used for the alignments.



The gene-level quantification tool aggregates raw counts of mapped reads (SAM/BAM files) using a gene feature format file (GTF/GFF) containing the genome coordinates of exons and genes. This approach discards multiple aligned reads. It runs locally.

Alignment Files

/home/biobam/sample3.bam
/home/biobam/sample2.bam
/home/biobam/sample1.bam

Annotation File: /home/biobam/gtfexamplecontigs.gtf

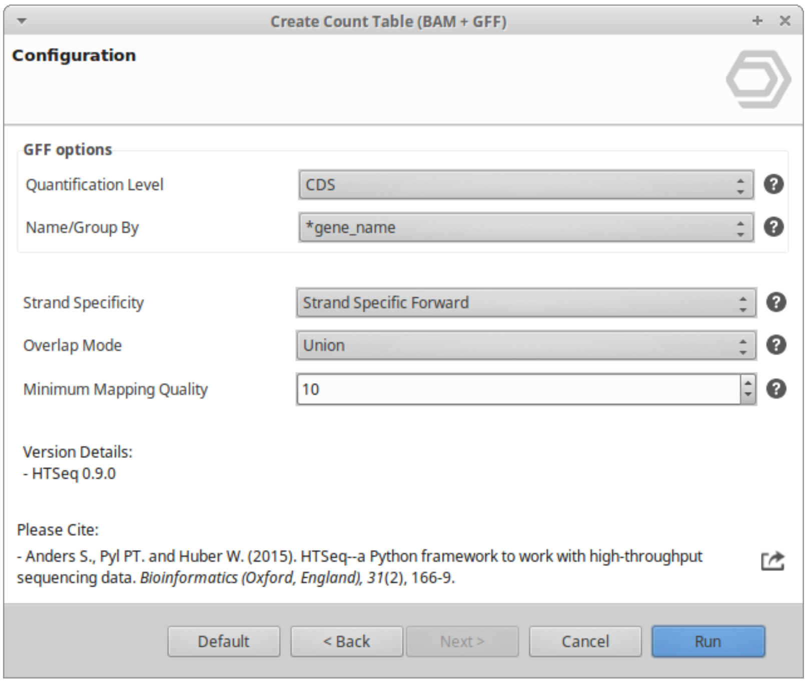
Figure 2: Input Page

9.5.2.2 Configuration

- **Quantification Level:** Choose the feature type (3rd column in GFF/GTF file, "Type" column in GFF object) for which expression will be quantified (e.g. gene, exon...). Features coordinates (range of positions) will be extracted from annotation using the provided value and all features of other types are ignored.
- **Name/Group By:** Specify the attribute type (9th column in GFF/GTF file, "Attr" columns in GFF object) to be used as feature ID. The feature ID is used to identify counts in the output Count Table. Attribute types tagged with "*" (e.g. *gene_name) are not present in all features of the selected type and only those containing it will be extracted. Several GFF lines with the same feature ID will be considered as parts of the same feature. Figure 4[96] illustrates how "Quantification Level" and "Name/Group By" parameters work.
- **Strand Specificity:** Indicate how the strand is taken into account.
  - Non-Strand Specific: A read is considered overlapping with a feature regardless of the strand in which the read has been mapped.
  - Strand Specific Forward: For single-end reads, the read has to be mapped to the same strand as the feature to be counted. For paired-end reads, the first read on the pair must be mapped to the same strand as the feature and the second read on the opposite strand.

[96] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598114455#Gene-levelQuantification-figure4

  - Strand Specific Reverse: For single-end reads, the read has to be mapped to the opposite strand of the feature to be counted. For paired-end reads, the first read on the pair must be mapped to the opposite strand of the feature and the second read on the same strand.
- **Overlap Mode:** Modes to handle reads overlapping more than one feature. Consider that for each position in the read, a set of all features overlapping is defined. If the resulting set for a read (or read pair) contains precisely one feature, the read is counted for this feature. If it contains more than one feature, the read is counted as "ambiguous" (and not counted for any features), and if it is empty, the read is counted as "no feature". The three overlap modes join these sets as follows (figure 5[97] illustrates the effect of these three modes):
  - Union: The union of all sets.
  - Intersection Strict: The intersection of all sets.
  - Intersection Non-Empty: The intersection of all non-empty sets.
- **Minimum Mapping Quality:** Set a filter to discard all reads with alignment quality (MAPQ) lower than the given minimum value.



Create Count Table (BAM + GFF)

Configuration

GFF options

Quantification Level: CDS

Name/Group By: *gene_name

Strand Specificity: Strand Specific Forward

Overlap Mode: Union

Minimum Mapping Quality: 10

Version Details:
- HTSeq 0.9.0

Please Cite:
- Anders S., Pyl PT. and Huber W. (2015). HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics (Oxford, England)*, *31*(2), 166-9.

Figure 3: Configuration Page

[97] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598114455#Gene-levelQuantification-figure5

SEQNAME	SOURCE	FEATURE	START	END	SCORE	STRAND	FRAME	ATTRIBUTES
chrom_1	RefSeq	gene	200	3150	.	+	.	ID=gene1; locus_tag=gene_one; gene=Gene_1
chrom_1	RefSeq	mRNA	200	3150	.	+	.	ID=mRNA1, transcript_id=mRNA_1; Parent=Gene1; gene_id=Gene_1
chrom_1	RefSeq	exon	200	1520	.	+	.	ID=exon1; exon_id=exon_1; Parent=mRNA1; gene_id=Gene_1
chrom_1	RefSeq	exon	1900	3150	.	+	.	ID=exon2; exon_id=exon_2; Parent=mRNA1; gene_id=Gene_1

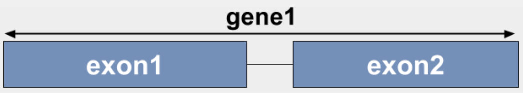
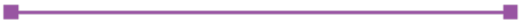
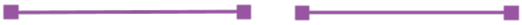
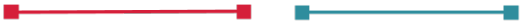
QUANTIFICATION LEVEL	NAME/ GROUP BY	FEATURE ID (IN COUNT TABLE)	
gene	locus_tag	**gene_one**	
exon	gene_id	**Gene_1**	
exon	ID	**exon1 exon2**	

Figure 4: Example of a simple GFF and usage of "Quantification Level" and "Name/Group By" parameters

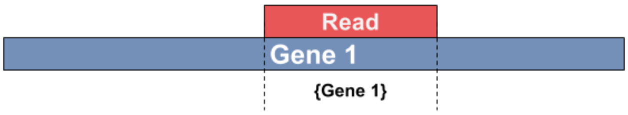
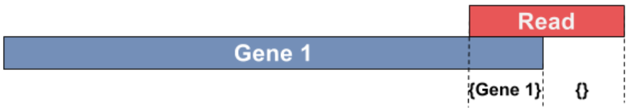
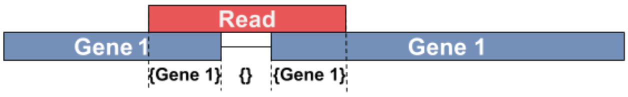
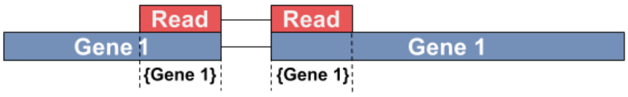
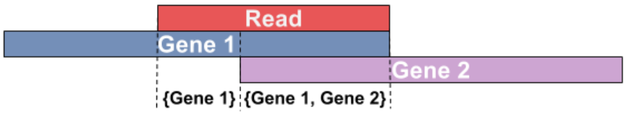
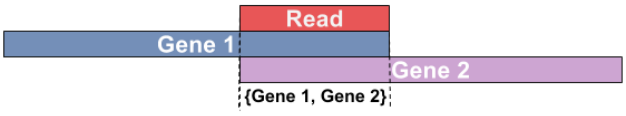
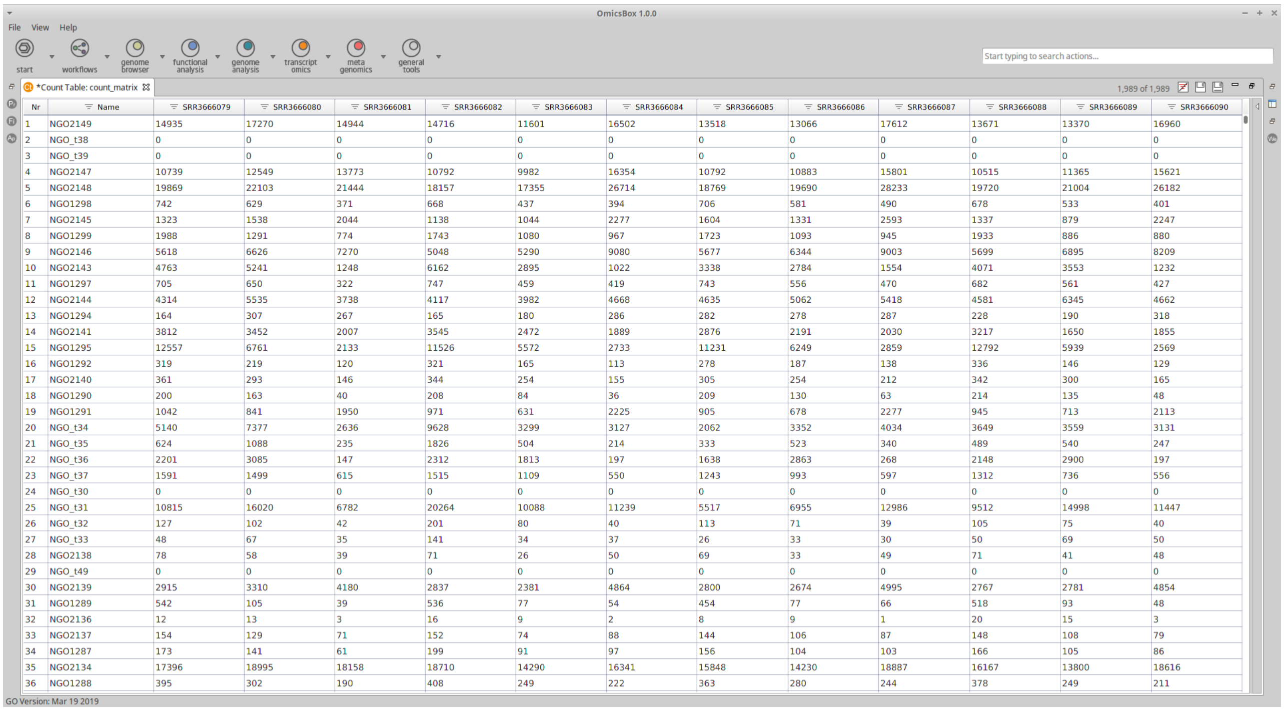
CASE	UNION	INTERSECTION STRICT	INTERSECTION NON EMPTY
	Gene 1	Gene 1	Gene 1
	Gene 1	No Feature	Gene 1
	Gene 1	No Feature	Gene 1
	Gene 1	Gene 1	Gene 1
	Ambiguous	Gene 1	Gene 1
	Ambiguous	Ambiguous	Ambi

Figure 5: Scheme of overlap modes

9.5.3 Results

Once the analysis has been finished, a new tab containing the resulting Count Table is opened (figure 6[98]). Rows correspond to genomic features and columns to samples (one by analyzed file). Counts represent the total number of reads aligned to each genomic feature. Results can be saved as a Count Table object.

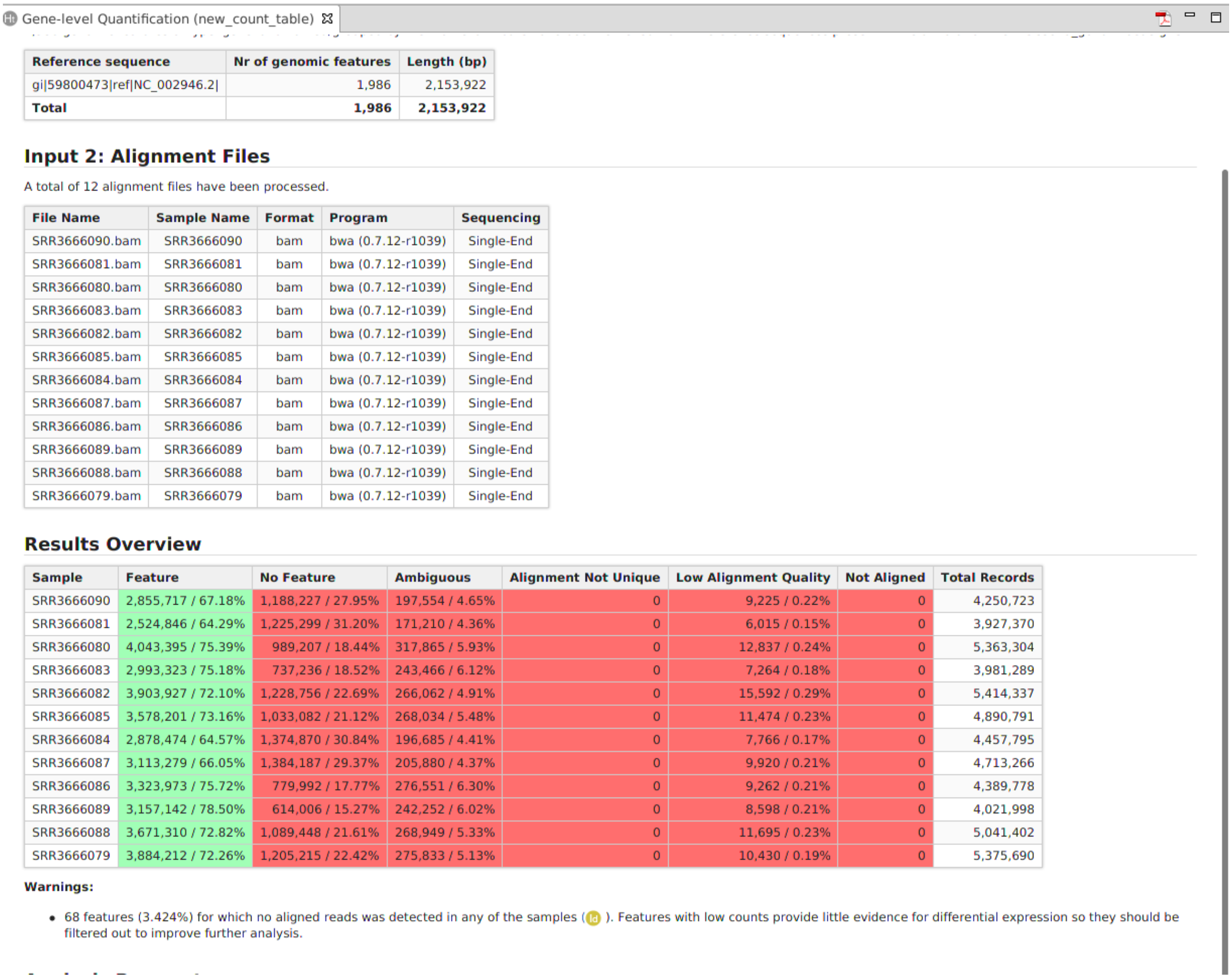


Nr	Name	SRR3666079	SRR3666080	SRR3666081	SRR3666082	SRR3666083	SRR3666084	SRR3666085	SRR3666086	SRR3666087	SRR3666088	SRR3666089	SRR3666090
1	NGO2149	14935	17270	14944	14716	11601	16502	13518	13066	17612	13671	13370	16960
2	NGO_t38	0	0	0	0	0	0	0	0	0	0	0	0
3	NGO_t39	0	0	0	0	0	0	0	0	0	0	0	0
4	NGO2147	10739	12549	13773	10792	9982	16354	10792	10883	15801	10515	11365	15621
5	NGO2148	19869	22103	21444	18157	17355	26714	18769	19690	28233	19720	21004	26182
6	NGO1298	742	629	371	668	437	394	706	581	490	678	533	401
7	NGO2145	1323	1538	2044	1138	1044	2277	1604	1331	2593	1337	879	2247
8	NGO1299	1988	1291	774	1743	1080	967	1723	1093	945	1933	886	880
9	NGO2146	5618	6626	7270	5048	5290	9080	5677	6344	9003	5699	6895	8209
10	NGO2143	4763	5241	1248	6162	2895	1022	3338	2784	1554	4071	3553	1232
11	NGO1297	705	650	322	747	459	419	743	556	470	682	561	427
12	NGO2144	4314	5535	3738	4117	3982	4668	4635	5062	5418	4581	6345	4662
13	NGO1294	164	307	267	165	180	286	282	278	287	228	190	318
14	NGO2141	3812	3452	2007	3545	2472	1889	2876	2191	2030	3217	1650	1855
15	NGO1295	12557	6761	2133	11526	5572	2733	11231	6249	2859	12792	5939	2569
16	NGO1292	319	219	120	321	165	113	278	187	138	336	146	129
17	NGO2140	361	293	146	344	254	155	305	254	212	342	300	165
18	NGO1290	200	163	40	208	84	36	209	130	63	214	135	48
19	NGO1291	1042	841	1950	971	631	2225	905	678	2277	945	713	2113
20	NGO_t34	5140	7377	2636	9628	3299	3127	2062	3352	4034	3649	3559	3131
21	NGO_t35	624	1088	235	1826	504	214	333	523	340	489	540	247
22	NGO_t36	2201	3085	147	2312	1813	197	1638	2863	268	2148	2900	197
23	NGO_t37	1591	1499	615	1515	1109	550	1243	993	597	1312	736	556
24	NGO_t30	0	0	0	0	0	0	0	0	0	0	0	0
25	NGO_t31	10815	16020	6782	20264	10088	11239	5517	6955	12986	9512	14998	11447
26	NGO_t32	127	102	42	201	80	40	113	71	39	105	75	40
27	NGO_t33	48	67	35	141	34	37	26	33	30	50	69	50
28	NGO2138	78	58	39	71	26	50	69	33	49	71	41	48
29	NGO_t49	0	0	0	0	0	0	0	0	0	0	0	0
30	NGO2139	2915	3310	4180	2837	2381	4864	2800	2674	4995	2767	2781	4854
31	NGO1289	542	105	39	536	77	54	454	77	66	518	93	48
32	NGO2136	12	13	3	16	9	2	8	9	1	20	15	3
33	NGO2137	154	129	71	152	74	88	144	106	87	148	108	79
34	NGO1287	173	141	61	199	91	97	156	104	103	166	105	86
35	NGO2134	17396	18995	18158	18710	14290	16341	15848	14230	18887	16167	13800	18616
36	NGO1288	395	302	190	408	249	222	363	280	244	378	249	211

Figure 6: Count Table

Furthermore, a result page will show a summary of the "Create Count Table" results (figure 7[99]). In this page information about the extraction of genomic features from GFF, alignment files and obtained results are provided. The result summary can be generated via **Side Panel** → **Result Summary** and it can be exported as pdf.

[98] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598114455#Gene-levelQuantification-figure6
[99] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598114455#Gene-levelQuantification-figure7



Reference sequence	Nr of genomic features	Length (bp)
gi|59800473|ref|NC_002946.2|	1,986	2,153,922
Total	**1,986**	**2,153,922**

Input 2: Alignment Files

A total of 12 alignment files have been processed.

File Name	Sample Name	Format	Program	Sequencing
SRR3666090.bam	SRR3666090	bam	bwa (0.7.12-r1039)	Single-End
SRR3666081.bam	SRR3666081	bam	bwa (0.7.12-r1039)	Single-End
SRR3666080.bam	SRR3666080	bam	bwa (0.7.12-r1039)	Single-End
SRR3666083.bam	SRR3666083	bam	bwa (0.7.12-r1039)	Single-End
SRR3666082.bam	SRR3666082	bam	bwa (0.7.12-r1039)	Single-End
SRR3666085.bam	SRR3666085	bam	bwa (0.7.12-r1039)	Single-End
SRR3666084.bam	SRR3666084	bam	bwa (0.7.12-r1039)	Single-End
SRR3666087.bam	SRR3666087	bam	bwa (0.7.12-r1039)	Single-End
SRR3666086.bam	SRR3666086	bam	bwa (0.7.12-r1039)	Single-End
SRR3666089.bam	SRR3666089	bam	bwa (0.7.12-r1039)	Single-End
SRR3666088.bam	SRR3666088	bam	bwa (0.7.12-r1039)	Single-End
SRR3666079.bam	SRR3666079	bam	bwa (0.7.12-r1039)	Single-End

Results Overview

Sample	Feature	No Feature	Ambiguous	Alignment Not Unique	Low Alignment Quality	Not Aligned	Total Records
SRR3666090	2,855,717 / 67.18%	1,188,227 / 27.95%	197,554 / 4.65%	0	9,225 / 0.22%	0	4,250,723
SRR3666081	2,524,846 / 64.29%	1,225,299 / 31.20%	171,210 / 4.36%	0	6,015 / 0.15%	0	3,927,370
SRR3666080	4,043,395 / 75.39%	989,207 / 18.44%	317,865 / 5.93%	0	12,837 / 0.24%	0	5,363,304
SRR3666083	2,993,323 / 75.18%	737,236 / 18.52%	243,466 / 6.12%	0	7,264 / 0.18%	0	3,981,289
SRR3666082	3,903,927 / 72.10%	1,228,756 / 22.69%	266,062 / 4.91%	0	15,592 / 0.29%	0	5,414,337
SRR3666085	3,578,201 / 73.16%	1,033,082 / 21.12%	268,034 / 5.48%	0	11,474 / 0.23%	0	4,890,791
SRR3666084	2,878,474 / 64.57%	1,374,870 / 30.84%	196,685 / 4.41%	0	7,766 / 0.17%	0	4,457,795
SRR3666087	3,113,279 / 66.05%	1,384,187 / 29.37%	205,880 / 4.37%	0	9,920 / 0.21%	0	4,713,266
SRR3666086	3,323,973 / 75.72%	779,992 / 17.77%	276,551 / 6.30%	0	9,262 / 0.21%	0	4,389,778
SRR3666089	3,157,142 / 78.50%	614,006 / 15.27%	242,252 / 6.02%	0	8,598 / 0.21%	0	4,021,998
SRR3666088	3,671,310 / 72.82%	1,089,448 / 21.61%	268,949 / 5.33%	0	11,695 / 0.23%	0	5,041,402
SRR3666079	3,884,212 / 72.26%	1,205,215 / 22.42%	275,833 / 5.13%	0	10,430 / 0.19%	0	5,375,690

Warnings:

- 68 features (3.424%) for which no aligned reads was detected in any of the samples ( ). Features with low counts provide little evidence for differential expression so they should be filtered out to improve further analysis.

Figure 7: Result Summary

9.5.4 Charts and Statistics

Different statistical charts of the obtained results can be generated. These provide additional information about the process of quantifying expression, as well as a quality assessment of the resulting counts. All these charts can be found under the **Side Panel** of the Count Table viewer.

- **Library Size per Sample**: Bar chart showing the number of read counts aligned to genomic features contained in each sample (figure 8a[100]).
- **Distribution of Counts:** Box plot that allows seeing how counts are distributed within each sample for all the features (figure 8b[101]). Features with 0 counts in all samples will be discarded for this chart. The binary logarithm of raw counts is represented.

[100] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598114455#Gene-levelQuantification-figure8a
[101] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598114455#Gene-levelQuantification-figure8b

- **Counts per Category:** Bar chart showing the number of reads of each input file sorted by different categories (figure 8c[102]). This chart and the next one are only available for count tables created by the "Create Count Table" tool within OmicsBox.
  - Feature: The sum of all reads that have been assigned to any features.
  - No Feature: Reads which could not be assigned to any feature (the resulting set for the read is empty as mentioned above).
  - Ambiguous: Reads which have been assigned to more than one feature (the resulting set for the read has more than one feature). These reads are not counted for any feature.
  - Alignment Not Unique: Reads with more than one reported alignment. These reads are identified from the NH optional SAM field tag. If the program that was used to obtain alignments does not set this field, multiple aligned reads will be counted multiple time.
  - Low Alignment Quality: Reads which were skipped due to the "Minimum Mapping Quality" filter set on the main wizard page.
  - Not Aligned: Reads in the SAM/BAM file without alignment.
- **Counts per Category (%):** The same chart as explained below in percentages (figure 8d[103]).

> Last two charts are only available for count tables created by the "Create Count Table" tool within OmicsBox.

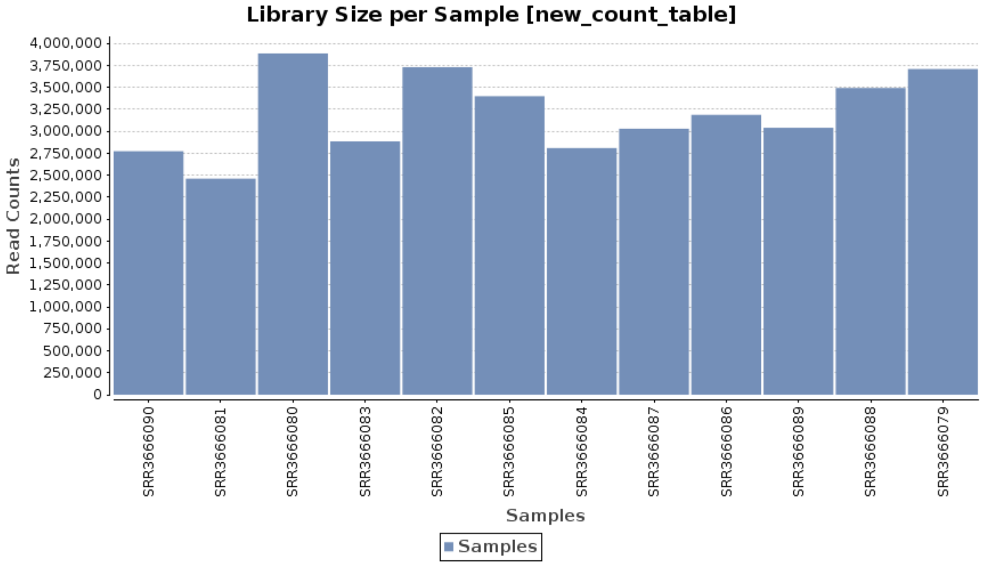


Figure 8(a): Library Size per Sample

[102] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598114455#Gene-levelQuantification-figure8c
[103] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598114455#Gene-levelQuantification-figure8d

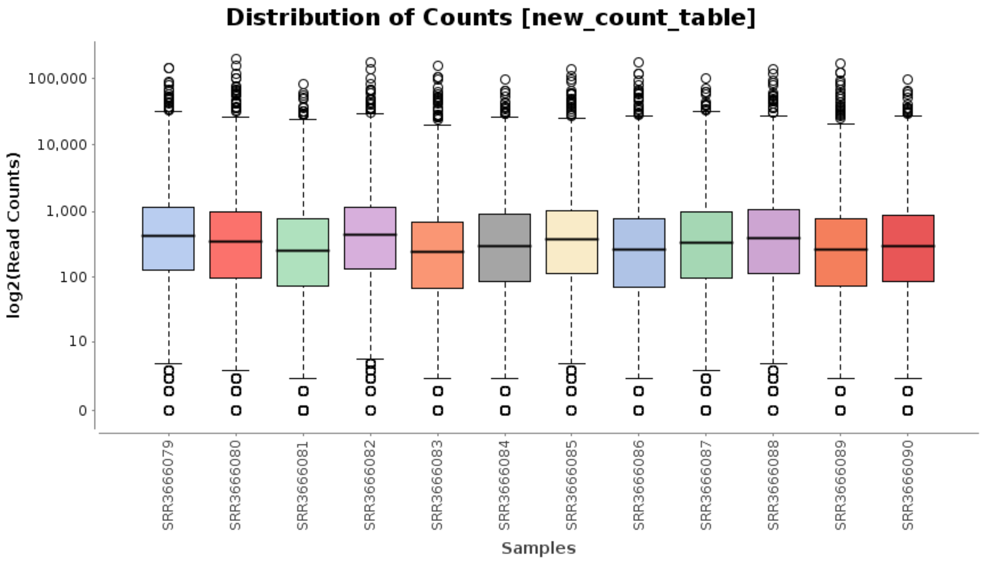


Figure 8(b). Distribution of Counts

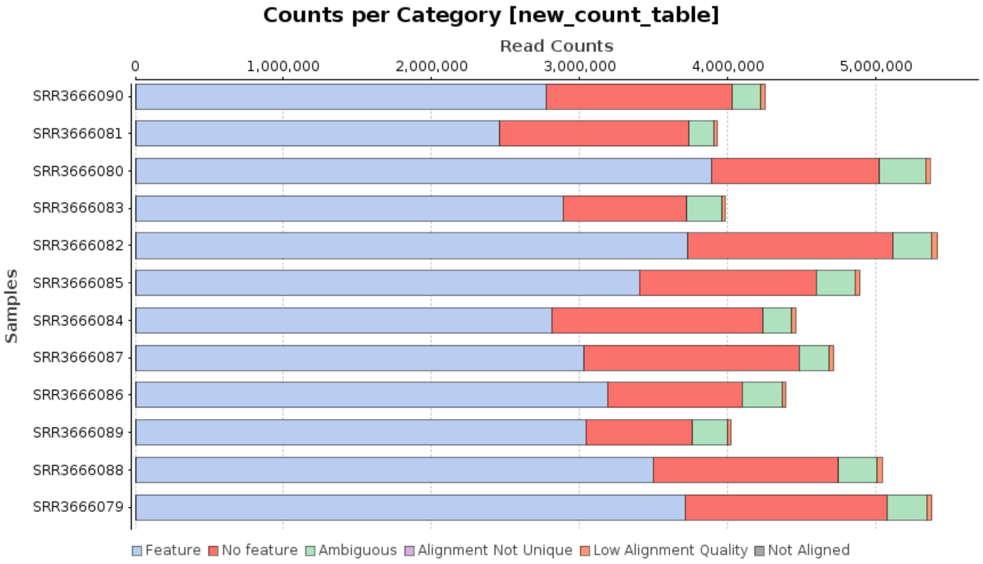


Figure 8(c): Counts per Category

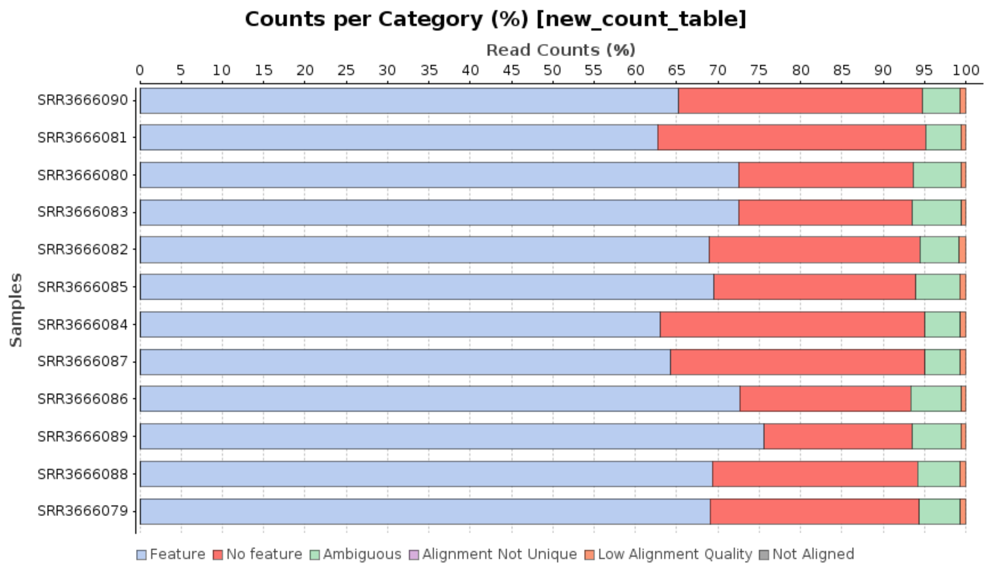


Figure 8(d): Counts per Category (%)

9.6 Transcript-level Quantification

Content of this page:

Introduction

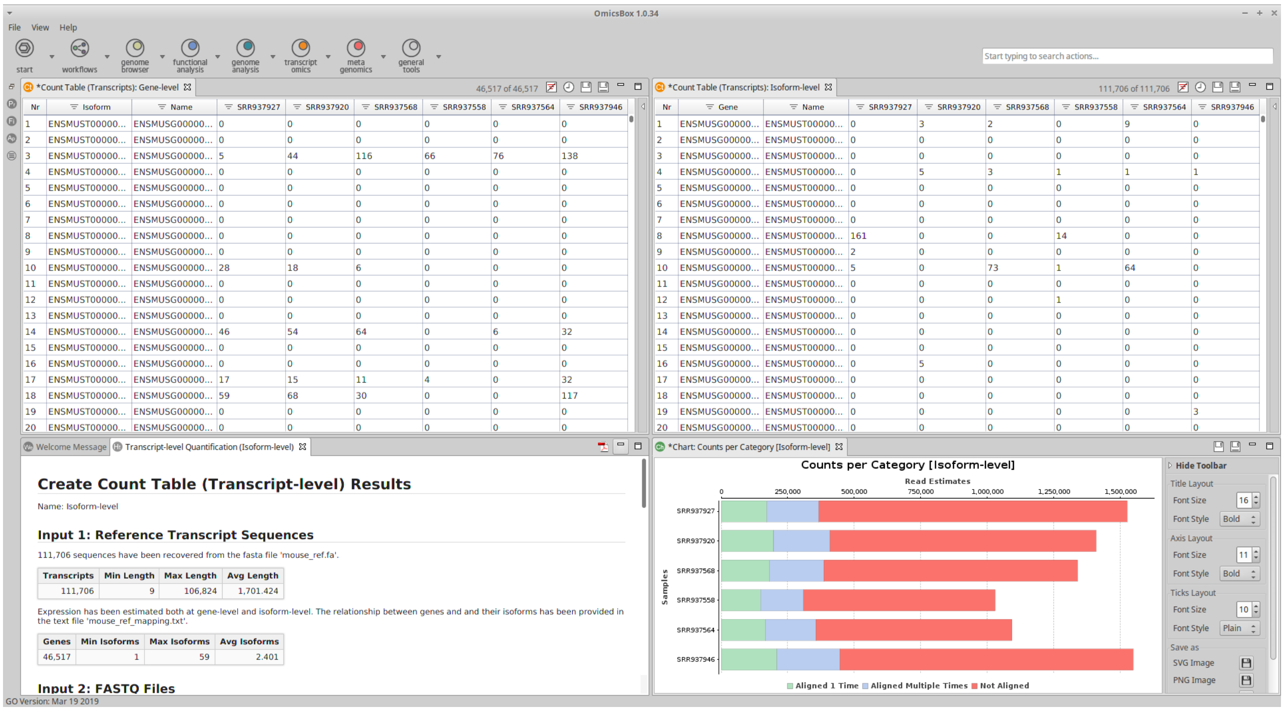
The transcript-level quantification tool is designed for estimating gene and isoform expression levels from RNA-Seq data. It expects the sequencing reads in FASTQ format (so a prior alignment is not necessary), and it supports both single-end and paired-end data. In addition, a set of transcript sequences in FASTA format is required, such as one produced by a de novo transcriptome assembler. Therefore it lacks the requirement of a reference genome. A Count Table is obtained and it can be used to perform a differential expression analysis within Blast2GO.

The application is based on RSEM[104], a software package that quantifies expression from transcriptome data. This program handles both the alignment of reads against the reference transcript sequences and the calculation for relative abundances. RSEM uses the Bowtie2 aligner to align reads, with parameters specifically chosen for RNA-Seq quantification. Since RNA-Seq reads do not always map uniquely to a single gene or isoform, this method is able to allocate multi-mapping reads among transcripts using an expectation maximization approach.

[104] https://deweylab.github.io/RSEM/

This feature uses RSEM and Bowtie2. Please cite RSEM and Bowtie2 as:

- Li B and Dewey CN (2011). "RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome." BMC Bioinformatics, 12:323
- Langmead B, Salzberg S (2012). "Fast gapped-read alignment with Bowtie 2." Nature Methods, 9:357-359



*Count Table (Transcripts): Gene-level

46,517 of 46,517

Nr	Isoform	Name	SRR937927	SRR937920	SRR937568	SRR937558	SRR937564	SRR937946
1	ENSMUST00000...	ENSMUSG00000...	0	0	0	0	0	0
2	ENSMUST00000...	ENSMUSG00000...	0	0	0	0	0	0
3	ENSMUST00000...	ENSMUSG00000...	5	44	116	66	76	138
4	ENSMUST00000...	ENSMUSG00000...	0	0	0	0	0	0
5	ENSMUST00000...	ENSMUSG00000...	0	0	0	0	0	0
6	ENSMUST00000...	ENSMUSG00000...	0	0	0	0	0	0
7	ENSMUST00000...	ENSMUSG00000...	0	0	0	0	0	0
8	ENSMUST00000...	ENSMUSG00000...	0	0	0	0	0	0
9	ENSMUST00000...	ENSMUSG00000...	0	0	0	0	0	0
10	ENSMUST00000...	ENSMUSG00000...	28	18	6	0	0	0
11	ENSMUST00000...	ENSMUSG00000...	0	0	0	0	0	0
12	ENSMUST00000...	ENSMUSG00000...	0	0	0	0	0	0
13	ENSMUST00000...	ENSMUSG00000...	0	0	0	0	0	0
14	ENSMUST00000...	ENSMUSG00000...	46	54	64	0	6	32
15	ENSMUST00000...	ENSMUSG00000...	0	0	0	0	0	0
16	ENSMUST00000...	ENSMUSG00000...	0	0	0	0	0	0
17	ENSMUST00000...	ENSMUSG00000...	17	15	11	4	0	32
18	ENSMUST00000...	ENSMUSG00000...	59	68	30	0	0	117
19	ENSMUST00000...	ENSMUSG00000...	0	0	0	0	0	0
20	ENSMUST00000...	ENSMUSG00000...	0	0	0	0	0	0

*Count Table (Transcripts): Isoform-level

111,706 of 111,706

Nr	Gene	Name	SRR937927	SRR937920	SRR937568	SRR937558	SRR937564	SRR937946
1	ENSMUSG00000...	ENSMUST00000...	0	3	2	0	9	0
2	ENSMUSG00000...	ENSMUST00000...	0	0	0	0	0	0
3	ENSMUSG00000...	ENSMUST00000...	0	0	0	0	0	0
4	ENSMUSG00000...	ENSMUST00000...	0	5	3	1	1	1
5	ENSMUSG00000...	ENSMUST00000...	0	0	0	0	0	0
6	ENSMUSG00000...	ENSMUST00000...	0	0	0	0	0	0
7	ENSMUSG00000...	ENSMUST00000...	0	0	0	0	0	0
8	ENSMUSG00000...	ENSMUST00000...	161	0	0	14	0	0
9	ENSMUSG00000...	ENSMUST00000...	2	0	0	0	0	0
10	ENSMUSG00000...	ENSMUST00000...	5	0	73	1	64	0
11	ENSMUSG00000...	ENSMUST00000...	0	0	0	0	0	0
12	ENSMUSG00000...	ENSMUST00000...	0	0	0	1	0	0
13	ENSMUSG00000...	ENSMUST00000...	0	0	0	0	0	0
14	ENSMUSG00000...	ENSMUST00000...	0	0	0	0	0	0
15	ENSMUSG00000...	ENSMUST00000...	0	0	0	0	0	0
16	ENSMUSG00000...	ENSMUST00000...	0	5	0	0	0	0
17	ENSMUSG00000...	ENSMUST00000...	0	0	0	0	0	0
18	ENSMUSG00000...	ENSMUST00000...	0	0	0	0	0	0
19	ENSMUSG00000...	ENSMUST00000...	0	0	0	0	0	3
20	ENSMUSG00000...	ENSMUST00000...	0	0	0	0	0	0

Create Count Table (Transcript-level) Results

Name: Isoform-level

Input 1: Reference Transcript Sequences

111,706 sequences have been recovered from the fasta file 'mouse_ref.fa'.

Transcripts	Min Length	Max Length	Avg Length
111,706	9	106,824	1,701.424

Expression has been estimated both at gene-level and isoform-level. The relationship between genes and and their isoforms has been provided in the text file 'mouse_ref_mapping.txt'.

Genes	Min Isoforms	Max Isoforms	Avg Isoforms
46,517	1	59	2.401

Input 2: FASTQ Files

GO Version: Mar 19 2019

Title Layout
Font Size 16
Font Style Bold
Axis Layout
Font Size 11
Font Style Bold
Ticks Layout
Font Size 10
Font Style Plain
Save as
SVG Image
PNG Image

Figure 1: Create Count Table Interface

9.6.1 Run Create Count Table

This functionality can be found under **Trasncriptomics → Create Count Table, Transcript-level Quantification** option. The wizard allows to select input files and adjust analysis parameters (figure 2[105], figure 3[106] and figure 5[107]).

9.6.1.1 **Input Data**

- **Sequencing Data:** Choose the type of data to be preprocessed: single-end or paired-end reads. Note that if paired-end is selected, two files per sample are required.
- **Input Reads:** Provide the files containing sequencing reads. These files are assumed to be in FASTQ format.

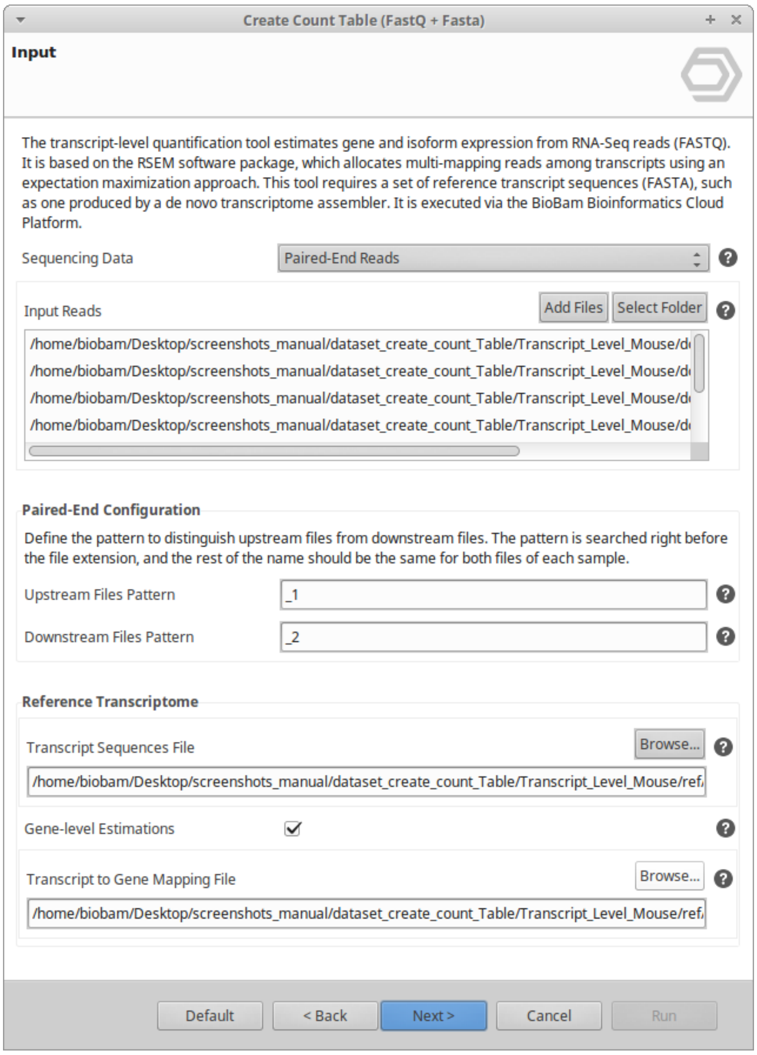
[105] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598049050#Transcript-levelQuantification-figure2
[106] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598049050#Transcript-levelQuantification-figure3
[107] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598049050#Transcript-levelQuantification-figure5

- **Paired-end configuration:** In the case of paired-end reads, the pattern to distinguish upstream files from downstream files is required. The provided patterns are searched right before the extension, and the start of the name should be the same for both files of each sample.
    - Upstream Files Pattern: Establish the pattern to recognize upstream FASTQ files.
    - Downstream Files Pattern: Establish the pattern to recognize downstream FASTQ files.

    > ⚠ For example, if the upstream file is named SRR037717_1.fastq and the downstream one SRR037717_2.fastq, you should establish "_1" as the upstream pattern and "_2" as the downstream pattern.

- **Transcript References:** This tool works with a set of transcripts sequences instead of a genome, such a file could be obtained from a reference genome database or a *de novo* transcriptome assembler. A FASTA file containing the sequences of reference transcripts should be provided.
- **Gene-level Estimations:** This option allows estimating expression both at gene-level and isoform-level. In this way, the gene's expression estimates are just the sum of its transcripts' expression estimates, and results will be provided separately. Otherwise, the program assumes that each transcript provided as a reference sequence is a separated gene.
- **Transcript to Gene Map File:** Provide a file with information to map from transcript (isoform) identifiers to gene identifiers. Each line should be of the form: gene id transcript id, with the two columns separated by a tab character (figure 4[108]).

[108] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598049050#Transcript-levelQuantification-figure4



Input

The transcript-level quantification tool estimates gene and isoform expression from RNA-Seq reads (FASTQ). It is based on the RSEM software package, which allocates multi-mapping reads among transcripts using an expectation maximization approach. This tool requires a set of reference transcript sequences (FASTA), such as one produced by a de novo transcriptome assembler. It is executed via the BioBam Bioinformatics Cloud Platform.

Sequencing Data: Paired-End Reads

Input Reads

/home/biobam/Desktop/screenshots_manual/dataset_create_count_Table/Transcript_Level_Mouse/d
/home/biobam/Desktop/screenshots_manual/dataset_create_count_Table/Transcript_Level_Mouse/d
/home/biobam/Desktop/screenshots_manual/dataset_create_count_Table/Transcript_Level_Mouse/d
/home/biobam/Desktop/screenshots_manual/dataset_create_count_Table/Transcript_Level_Mouse/d

Paired-End Configuration

Define the pattern to distinguish upstream files from downstream files. The pattern is searched right before the file extension, and the rest of the name should be the same for both files of each sample.

Upstream Files Pattern: _1

Downstream Files Pattern: _2

Reference Transcriptome

Transcript Sequences File

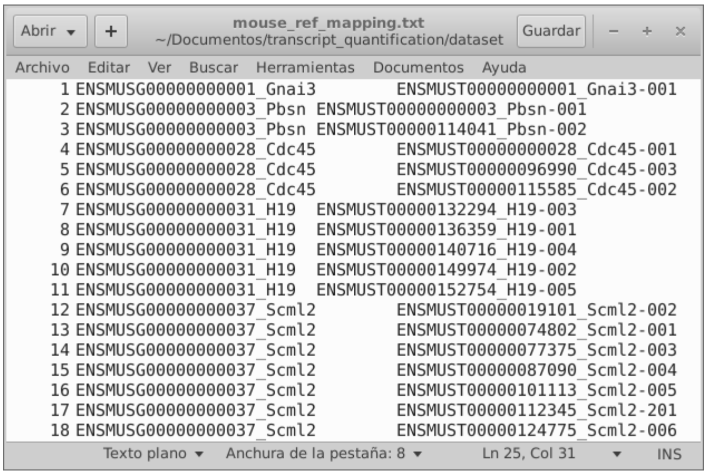
/home/biobam/Desktop/screenshots_manual/dataset_create_count_Table/Transcript_Level_Mouse/ref/

Gene-level Estimations ☑

Transcript to Gene Mapping File

/home/biobam/Desktop/screenshots_manual/dataset_create_count_Table/Transcript_Level_Mouse/ref/

Figure 2: Input Data Page



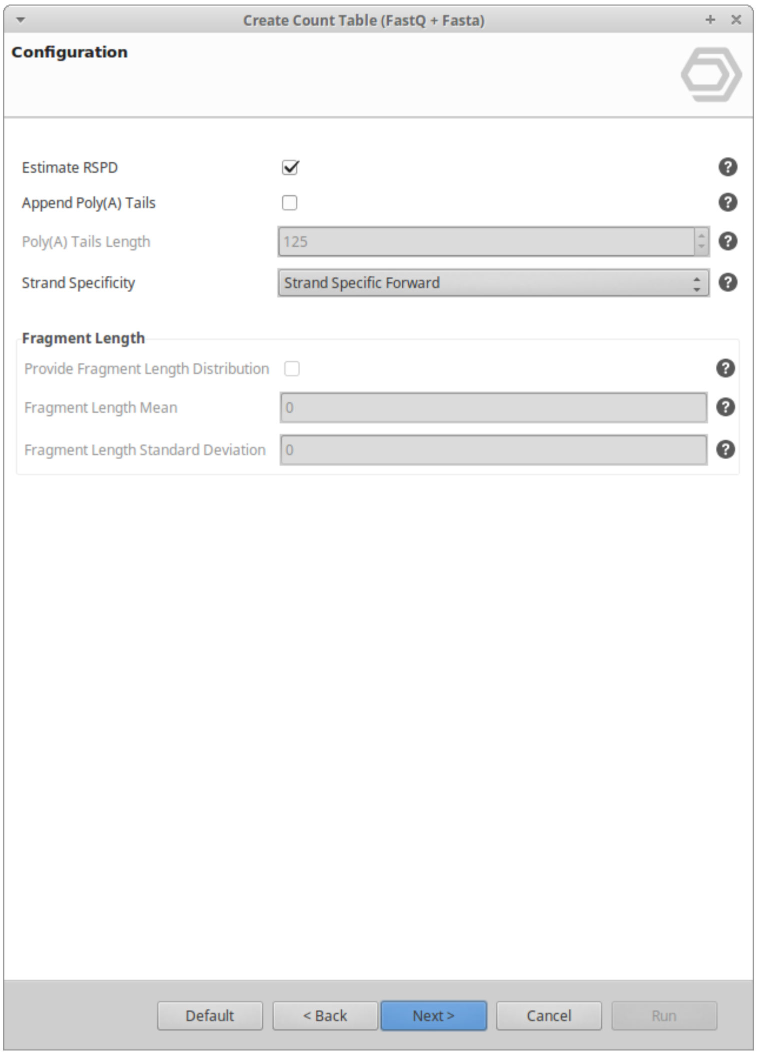
Abrir
+
mouse_ref_mapping.txt
~/Documentos/transcript_quantification/dataset
Guardar
Archivo Editar Ver Buscar Herramientas Documentos Ayuda

```
ENSMUSG00000000001_Gnai3	ENSMUST00000000001_Gnai3-001
ENSMUSG00000000003_Pbsn	ENSMUST00000000003_Pbsn-001
ENSMUSG00000000003_Pbsn	ENSMUST00000114041_Pbsn-002
ENSMUSG00000000028_Cdc45	ENSMUST00000000028_Cdc45-001
ENSMUSG00000000028_Cdc45	ENSMUST00000096990_Cdc45-003
ENSMUSG00000000028_Cdc45	ENSMUST00000115585_Cdc45-002
ENSMUSG00000000031_H19	ENSMUST00000132294_H19-003
ENSMUSG00000000031_H19	ENSMUST00000136359_H19-001
ENSMUSG00000000031_H19	ENSMUST00000140716_H19-004
ENSMUSG00000000031_H19	ENSMUST00000149974_H19-002
ENSMUSG00000000031_H19	ENSMUST00000152754_H19-005
ENSMUSG00000000037_Scml2	ENSMUST00000019101_Scml2-002
ENSMUSG00000000037_Scml2	ENSMUST00000074802_Scml2-001
ENSMUSG00000000037_Scml2	ENSMUST00000077375_Scml2-003
ENSMUSG00000000037_Scml2	ENSMUST00000087090_Scml2-004
ENSMUSG00000000037_Scml2	ENSMUST00000101113_Scml2-005
ENSMUSG00000000037_Scml2	ENSMUST00000112345_Scml2-201
ENSMUSG00000000037_Scml2	ENSMUST00000124775_Scml2-006
```

Figure 4: Transcript to gene map file example

9.6.1.2 Advanced Configuration

- **Append Poly(A) Tails:** For poly(A) mRNA analysis, the program will append a poly(A) tail sequences to reference transcripts to allow more accurate read alignment.
- **Poly(A) Tails Length:** Establish the length of the poly(A) tails to be added.
- **Estimate RSPD:** This option allows to estimate a read start position distribution (RSPD), which increases the accuracy of expression estimates. Highly recommended if the protocol produces read position distributions that are highly 5' or 3' biased. Otherwise, the program will use a uniform RSPD.
- **Strand Specificity:** This option defines the strandedness of the RNA-Seq reads:
  - Non-Strand Specific: Refers to non-strand-specific protocols.
  - Strand Specific Forward: Means all (upstream) reads are derived from the forward strand.
  - Strand Specific Reverse: Means all (upstream) reads are derived from the reverse strand.



Create Count Table (FastQ + Fasta)

Configuration

Estimate RSPD ☑

Append Poly(A) Tails □

Poly(A) Tails Length 125

Strand Specificity Strand Specific Forward

Fragment Length

Provide Fragment Length Distribution □

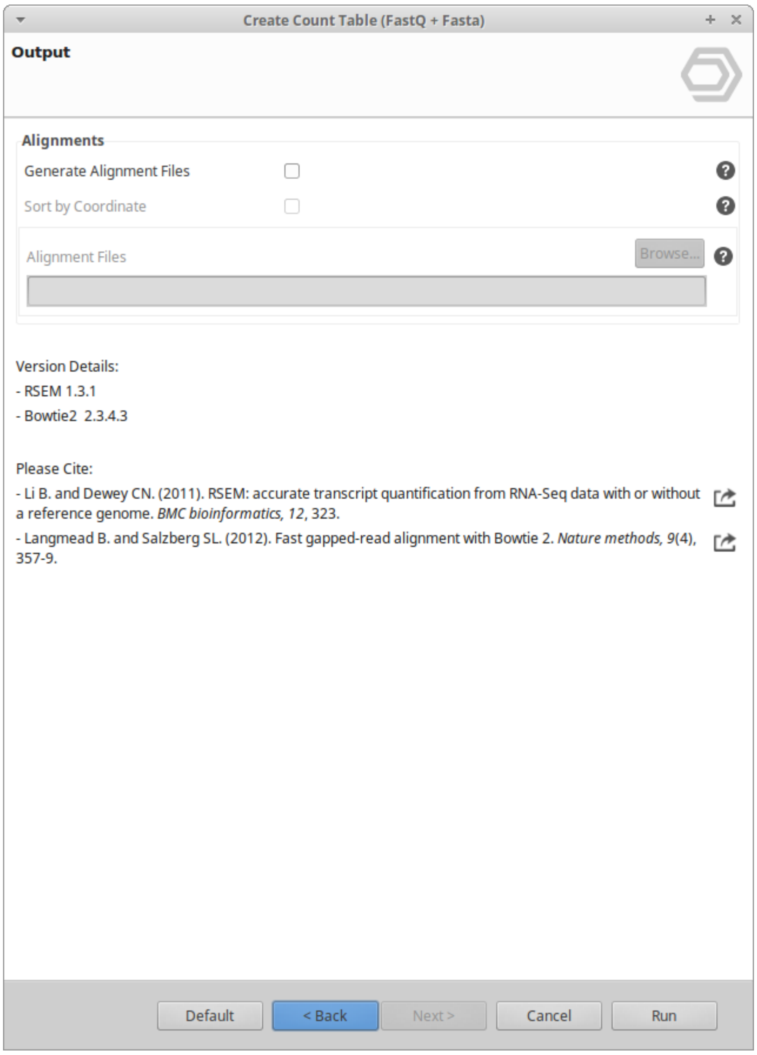
Fragment Length Mean 0

Fragment Length Standard Deviation 0

Figure 3: Advanced Configuration Page

9.6.1.3 Output Data

- **Alignments.** Decide if alignments files in bam format are saved and select a location to place them. These files can be used for downstream analyses.



Version Details:
- RSEM 1.3.1
- Bowtie2 2.3.4.3

Please Cite:
- Li B. and Dewey CN. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics*, *12*, 323.
- Langmead B. and Salzberg SL. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods*, *9*(4), 357-9.

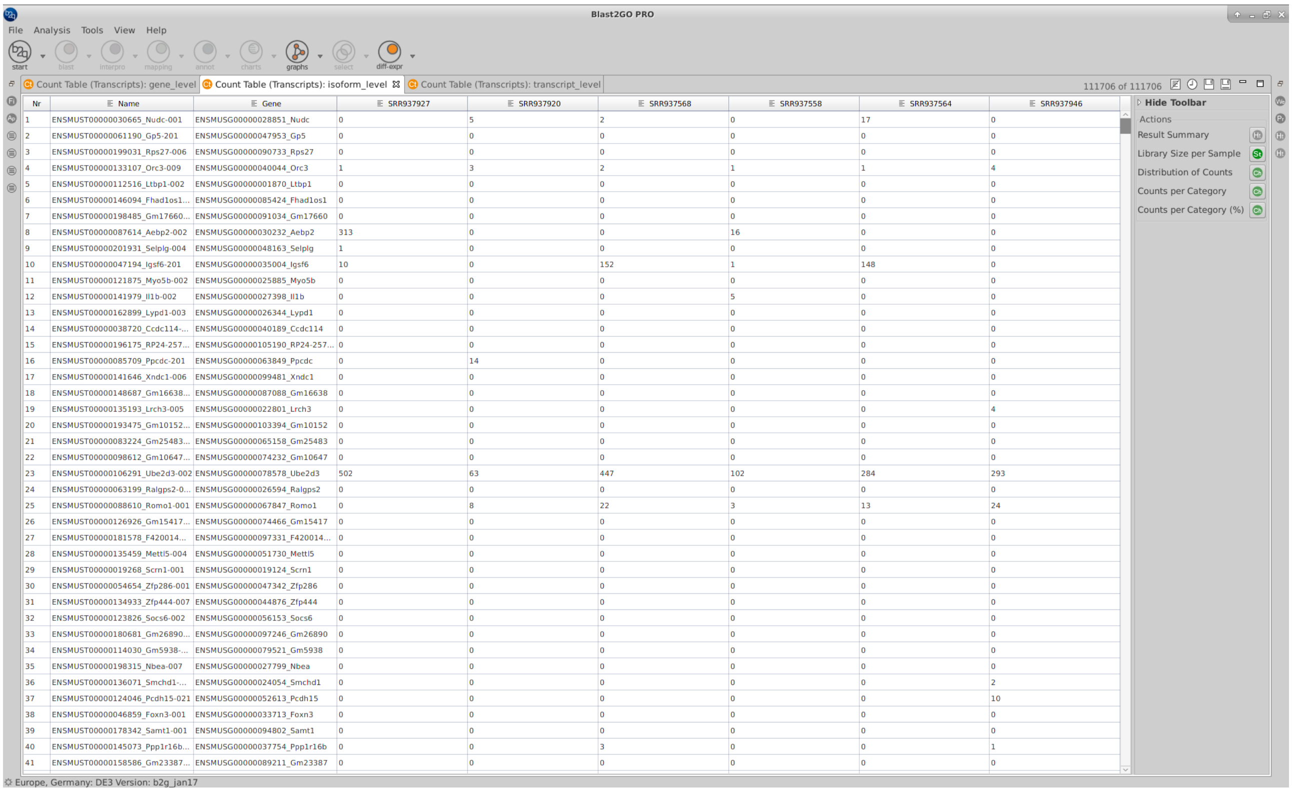
Figure 5: Output Data Page

9.6.2 Results

Once the analysis has been finished results will be returned in two different ways, depending on the option chosen in the "Gene-level Estimations" parameter:

- **Isoform-level and gene-level estimations:** Two Count Tables are returned. One shows the expression level of each transcript or isoform (input sequences) and other shows the expression level of each gene (figure 6[109] and figure 7[110]). They have an additional column that shows the gene or transcript identifiers (respectively) associated to each record.

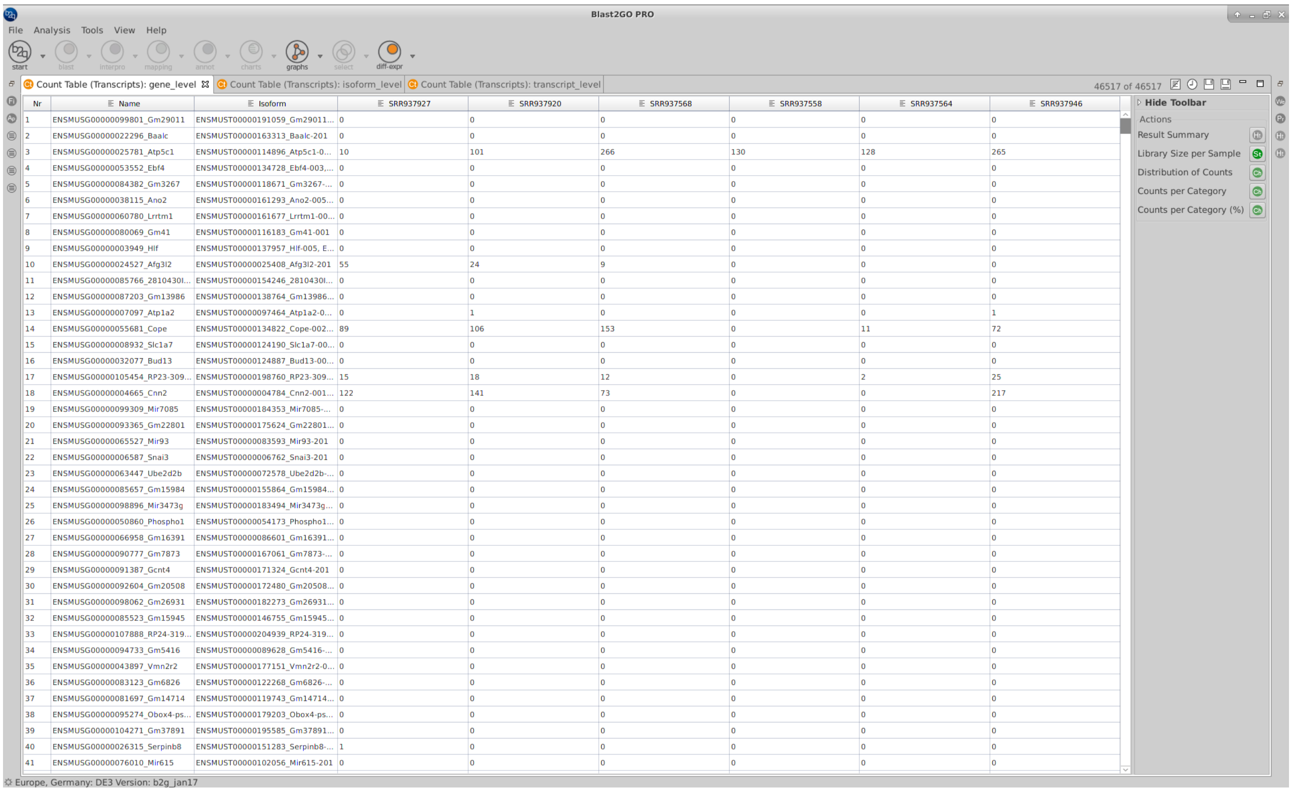
- **Transcript-level estimations only:** One Count Table is returned that shows the expression level of each transcript sequence provided as input (figure 8[111]).



Nr	Name	Gene	SRR937927	SRR937920	SRR937568	SRR937558	SRR937564	SRR937946
1	ENSMUST00000030665_Nudc-001	ENSMUSG00000028851_Nudc	0	5	2	0	17	0
2	ENSMUST00000061190_Gp5-201	ENSMUSG00000047953_Gp5	0	0	0	0	0	0
3	ENSMUST00000199031_Rps27-006	ENSMUSG00000090733_Rps27	0	0	0	0	0	0
4	ENSMUST00000133107_Orc3-009	ENSMUSG00000040044_Orc3	1	3	2	1	1	4
5	ENSMUST00000112516_Ltbp1-002	ENSMUSG00000001870_Ltbp1	0	0	0	0	0	0
6	ENSMUST00000146094_Fhad1os1...	ENSMUSG00000085424_Fhad1os1	0	0	0	0	0	0
7	ENSMUST00000198485_Gm17660...	ENSMUSG00000091034_Gm17660	0	0	0	0	0	0
8	ENSMUST00000087614_Aebp2-002	ENSMUSG00000030232_Aebp2	313	0	0	16	0	0
9	ENSMUST00000201931_Selplg-004	ENSMUSG00000048163_Selplg	1	0	0	0	0	0
10	ENSMUST00000047194_Igsf6-201	ENSMUSG00000035004_Igsf6	10	0	152	1	148	0
11	ENSMUST00000121875_Myo5b-002	ENSMUSG00000025885_Myo5b	0	0	0	0	0	0
12	ENSMUST00000141979_Il1b-002	ENSMUSG00000027398_Il1b	0	0	0	5	0	0
13	ENSMUST00000162899_Lypd1-003	ENSMUSG00000026344_Lypd1	0	0	0	0	0	0
14	ENSMUST00000038720_Ccdc114-...	ENSMUSG00000040189_Ccdc114	0	0	0	0	0	0
15	ENSMUST00000196175_RP24-257...	ENSMUSG00000105190_RP24-257...	0	0	0	0	0	0
16	ENSMUST00000085709_Ppcdc-201	ENSMUSG00000063849_Ppcdc	0	14	0	0	0	0
17	ENSMUST00000141646_Xndc1-006	ENSMUSG00000099481_Xndc1	0	0	0	0	0	0
18	ENSMUST00000148687_Gm16638...	ENSMUSG00000087088_Gm16638	0	0	0	0	0	0
19	ENSMUST00000135193_Lrch3-005	ENSMUSG00000022801_Lrch3	0	0	0	0	0	4
20	ENSMUST00000193475_Gm10152...	ENSMUSG00000103394_Gm10152	0	0	0	0	0	0
21	ENSMUST00000083224_Gm25483...	ENSMUSG00000065158_Gm25483	0	0	0	0	0	0
22	ENSMUST00000098612_Gm10647...	ENSMUSG00000074232_Gm10647	0	0	0	0	0	0
23	ENSMUST00000106291_Ube2d3-002	ENSMUSG00000078578_Ube2d3	502	63	447	102	284	293
24	ENSMUST00000063199_Ralgps2-0...	ENSMUSG00000026594_Ralgps2	0	0	0	0	0	0
25	ENSMUST00000088610_Romo1-001	ENSMUSG00000067847_Romo1	0	8	22	3	13	24
26	ENSMUST00000126926_Gm15417...	ENSMUSG00000074466_Gm15417	0	0	0	0	0	0
27	ENSMUST00000181578_F420014...	ENSMUSG00000097331_F420014...	0	0	0	0	0	0
28	ENSMUST00000135459_Mettl5-004	ENSMUSG00000051730_Mettl5	0	0	0	0	0	0
29	ENSMUST00000019268_Scrn1-001	ENSMUSG00000019124_Scrn1	0	0	0	0	0	0
30	ENSMUST00000054654_Zfp286-001	ENSMUSG00000047342_Zfp286	0	0	0	0	0	0
31	ENSMUST00000134933_Zfp444-007	ENSMUSG00000044876_Zfp444	0	0	0	0	0	0
32	ENSMUST00000123826_Socs6-002	ENSMUSG00000056153_Socs6	0	0	0	0	0	0
33	ENSMUST00000180681_Gm26890...	ENSMUSG00000097246_Gm26890	0	0	0	0	0	0
34	ENSMUST00000114030_Gm5938-...	ENSMUSG00000079521_Gm5938	0	0	0	0	0	0
35	ENSMUST00000198315_Nbea-007	ENSMUSG00000027799_Nbea	0	0	0	0	0	0
36	ENSMUST00000136071_Smchd1-...	ENSMUSG00000024054_Smchd1	0	0	0	0	0	2
37	ENSMUST00000124046_Pcdh15-021	ENSMUSG00000052613_Pcdh15	0	0	0	0	0	10
38	ENSMUST00000046859_Foxn3-001	ENSMUSG00000033713_Foxn3	0	0	0	0	0	0
39	ENSMUST00000178342_Samt1-001	ENSMUSG00000094802_Samt1	0	0	0	0	0	0
40	ENSMUST00000145073_Ppp1r16b...	ENSMUSG00000037754_Ppp1r16b	0	0	3	0	0	1
41	ENSMUST00000158586_Gm23387...	ENSMUSG00000089211_Gm23387	0	0	0	0	0	0

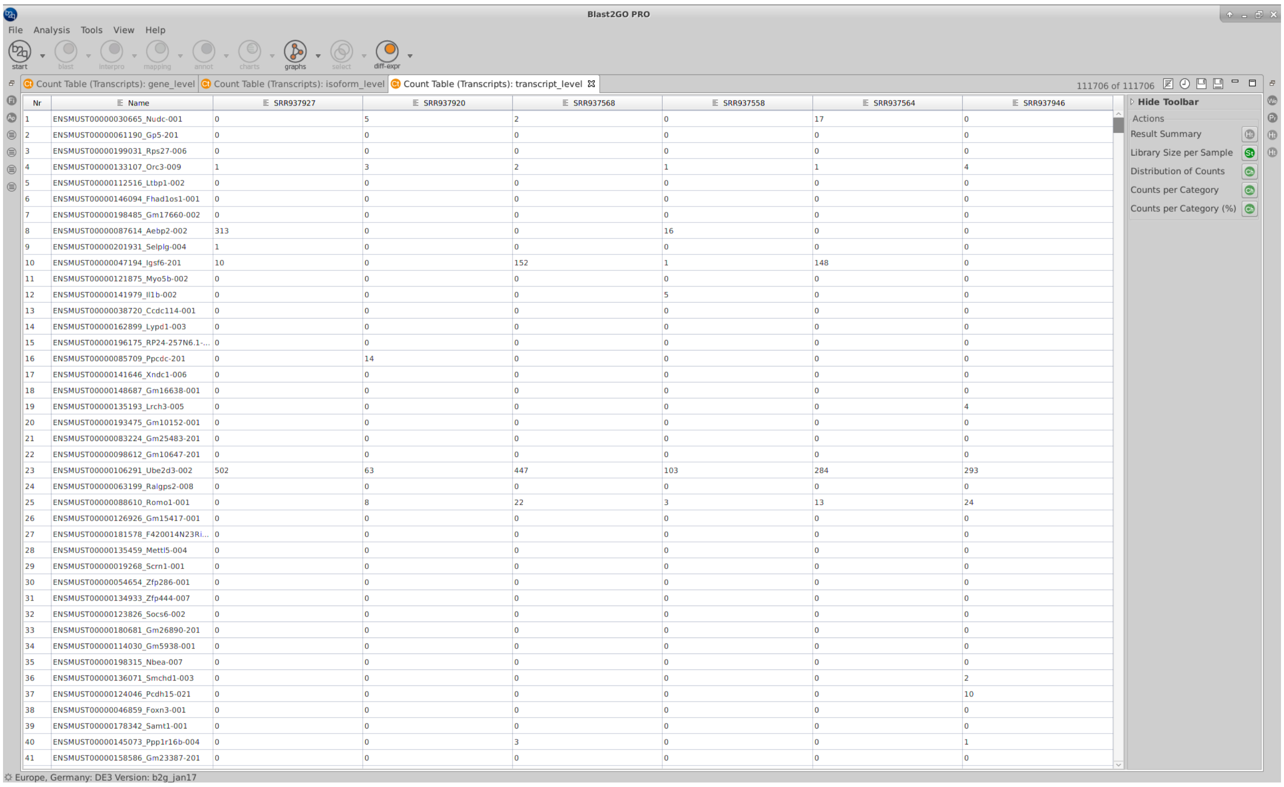
Figure 6: Count table of isoform-level estimations. The Gene column shows the identifier of the parent gene of each isoform

[109] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598049050#Transcript-levelQuantification-figure6
[110] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598049050#Transcript-levelQuantification-figure7
[111] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598049050#Transcript-levelQuantification-figure8



Nr	Name	Isoform	SRR937927	SRR937920	SRR937568	SRR937558	SRR937564	SRR937946
1	ENSMUSG00000099801_Gm29011	ENSMUST00000191059_Gm29011...	0	0	0	0	0	0
2	ENSMUSG00000022296_Baalc	ENSMUST00000163313_Baalc-201	0	0	0	0	0	0
3	ENSMUSG00000025781_Atp5c1	ENSMUST00000114896_Atp5c1-0...	10	101	266	130	128	265
4	ENSMUSG00000053552_Ebf4	ENSMUST00000134728_Ebf4-003,...	0	0	0	0	0	0
5	ENSMUSG00000084382_Gm3267	ENSMUST00000118671_Gm3267-...	0	0	0	0	0	0
6	ENSMUSG00000038115_Ano2	ENSMUST00000161293_Ano2-005...	0	0	0	0	0	0
7	ENSMUSG00000060780_Lrrtm1	ENSMUST00000161677_Lrrtm1-00...	0	0	0	0	0	0
8	ENSMUSG00000080069_Gm41	ENSMUST00000116183_Gm41-001	0	0	0	0	0	0
9	ENSMUSG00000003949_Hlf	ENSMUST00000137957_Hlf-005, E...	0	0	0	0	0	0
10	ENSMUSG00000024527_Afg3l2	ENSMUST00000025408_Afg3l2-201	55	24	9	0	0	0
11	ENSMUSG00000085766_2810430I...	ENSMUST00000154246_2810430I...	0	0	0	0	0	0
12	ENSMUSG00000087203_Gm13986	ENSMUST00000138764_Gm13986...	0	0	0	0	0	0
13	ENSMUSG00000007097_Atp1a2	ENSMUST00000097464_Atp1a2-0...	0	1	0	0	0	1
14	ENSMUSG00000055681_Cope	ENSMUST00000134822_Cope-002...	89	106	153	0	11	72
15	ENSMUSG00000008932_Slc1a7	ENSMUST00000124190_Slc1a7-00...	0	0	0	0	0	0
16	ENSMUSG00000032077_Bud13	ENSMUST00000124887_Bud13-00...	0	0	0	0	0	0
17	ENSMUSG00000105454_RP23-309...	ENSMUST00000198760_RP23-309...	15	18	12	0	2	25
18	ENSMUSG00000004665_Cnn2	ENSMUST00000004784_Cnn2-001...	122	141	73	0	0	217
19	ENSMUSG00000099309_Mir7085	ENSMUST00000184353_Mir7085-...	0	0	0	0	0	0
20	ENSMUSG00000093365_Gm22801	ENSMUST00000175624_Gm22801...	0	0	0	0	0	0
21	ENSMUSG00000065527_Mir93	ENSMUST00000083593_Mir93-201	0	0	0	0	0	0
22	ENSMUSG00000006587_Snai3	ENSMUST00000006762_Snai3-201	0	0	0	0	0	0
23	ENSMUSG00000063447_Ube2d2b	ENSMUST00000072578_Ube2d2b-...	0	0	0	0	0	0
24	ENSMUSG00000085657_Gm15984	ENSMUST00000155864_Gm15984...	0	0	0	0	0	0
25	ENSMUSG00000098896_Mir3473g	ENSMUST00000183494_Mir3473g...	0	0	0	0	0	0
26	ENSMUSG00000050860_Phospho1	ENSMUST00000054173_Phospho1...	0	0	0	0	0	0
27	ENSMUSG00000066958_Gm16391	ENSMUST00000086601_Gm16391...	0	0	0	0	0	0
28	ENSMUSG00000090777_Gm7873	ENSMUST00000167061_Gm7873-...	0	0	0	0	0	0
29	ENSMUSG00000091387_Gcnt4	ENSMUST00000171324_Gcnt4-201	0	0	0	0	0	0
30	ENSMUSG00000092604_Gm20508	ENSMUST00000172480_Gm20508...	0	0	0	0	0	0
31	ENSMUSG00000098062_Gm26931	ENSMUST00000182273_Gm26931...	0	0	0	0	0	0
32	ENSMUSG00000085523_Gm15945	ENSMUST00000146755_Gm15945...	0	0	0	0	0	0
33	ENSMUSG00000107888_RP24-319...	ENSMUST00000204939_RP24-319...	0	0	0	0	0	0
34	ENSMUSG00000094733_Gm5416	ENSMUST00000089628_Gm5416-...	0	0	0	0	0	0
35	ENSMUSG00000043897_Vmn2r2	ENSMUST00000177151_Vmn2r2-0...	0	0	0	0	0	0
36	ENSMUSG00000083123_Gm6826	ENSMUST00000122268_Gm6826-...	0	0	0	0	0	0
37	ENSMUSG00000081697_Gm14714	ENSMUST00000119743_Gm14714...	0	0	0	0	0	0
38	ENSMUSG00000095274_Obox4-ps...	ENSMUST00000179203_Obox4-ps...	0	0	0	0	0	0
39	ENSMUSG00000104271_Gm37891	ENSMUST00000195585_Gm37891...	0	0	0	0	0	0
40	ENSMUSG00000026315_Serpinb8	ENSMUST00000151283_Serpinb8-...	1	0	0	0	0	0
41	ENSMUSG00000076010_Mir615	ENSMUST00000102056_Mir615-201	0	0	0	0	0	0

Figure 7: Count table of gene-level estimations. The Isoform column shows the ids of all isoforms associated with each gene

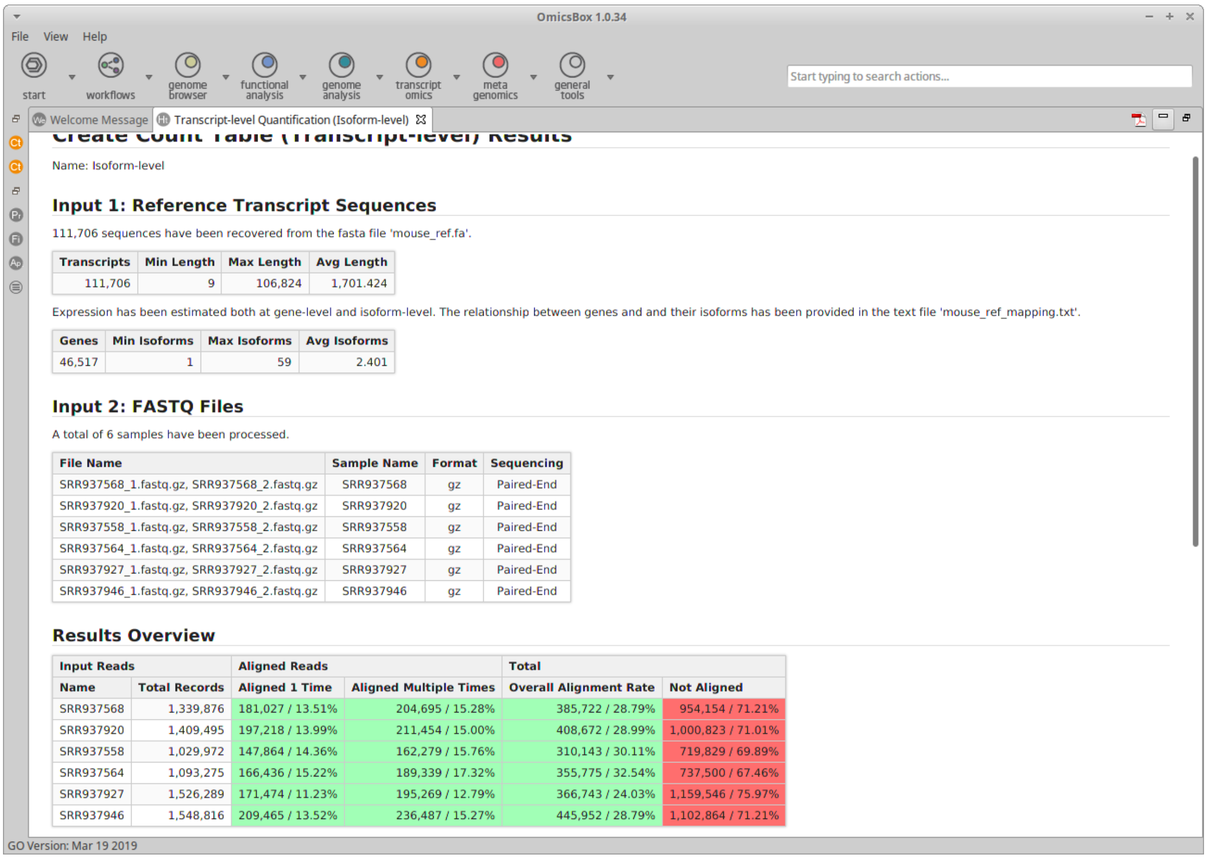


Nr	Name	SRR937927	SRR937920	SRR937568	SRR937558	SRR937564	SRR937946
1	ENSMUST00000030665_Nudc-001	0	5	2	0	17	0
2	ENSMUST00000061190_Gp5-201	0	0	0	0	0	0
3	ENSMUST00000199031_Rps27-006	0	0	0	0	0	0
4	ENSMUST00000133107_Orc3-009	1	3	2	1	1	4
5	ENSMUST00000112516_Ltbp1-002	0	0	0	0	0	0
6	ENSMUST00000146094_Fhad1os1-001	0	0	0	0	0	0
7	ENSMUST00000198485_Gm17660-002	0	0	0	0	0	0
8	ENSMUST00000087614_Aebp2-002	313	0	0	16	0	0
9	ENSMUST00000201931_Selplg-004	1	0	0	0	0	0
10	ENSMUST00000047194_Igsf6-201	10	0	152	1	148	0
11	ENSMUST00000121875_Myo5b-002	0	0	0	0	0	0
12	ENSMUST00000141979_Il1b-002	0	0	0	5	0	0
13	ENSMUST00000038720_Ccdc114-001	0	0	0	0	0	0
14	ENSMUST00000162899_Lypd1-003	0	0	0	0	0	0
15	ENSMUST00000196175_RP24-257N6.1-...	0	0	0	0	0	0
16	ENSMUST00000085709_Ppcdc-201	0	14	0	0	0	0
17	ENSMUST00000141646_Xndc1-006	0	0	0	0	0	0
18	ENSMUST00000148687_Gm16638-001	0	0	0	0	0	0
19	ENSMUST00000135193_Lrch3-005	0	0	0	0	0	4
20	ENSMUST00000193475_Gm10152-001	0	0	0	0	0	0
21	ENSMUST00000083224_Gm25483-201	0	0	0	0	0	0
22	ENSMUST00000098612_Gm10647-201	0	0	0	0	0	0
23	ENSMUST00000106291_Ube2d3-002	502	63	447	103	284	293
24	ENSMUST00000063199_Ralgps2-008	0	0	0	0	0	0
25	ENSMUST00000088610_Romo1-001	0	8	22	3	13	24
26	ENSMUST00000126926_Gm15417-001	0	0	0	0	0	0
27	ENSMUST00000181578_F420014N23Ri...	0	0	0	0	0	0
28	ENSMUST00000135459_Mettl5-004	0	0	0	0	0	0
29	ENSMUST00000019268_Scrn1-001	0	0	0	0	0	0
30	ENSMUST00000054654_Zfp286-001	0	0	0	0	0	0
31	ENSMUST00000134933_Zfp444-007	0	0	0	0	0	0
32	ENSMUST00000123826_Socs6-002	0	0	0	0	0	0
33	ENSMUST00000180681_Gm26890-201	0	0	0	0	0	0
34	ENSMUST00000114030_Gm5938-001	0	0	0	0	0	0
35	ENSMUST00000198315_Nbea-007	0	0	0	0	0	0
36	ENSMUST00000136071_Smchd1-003	0	0	0	0	0	2
37	ENSMUST00000124046_Pcdh15-021	0	0	0	0	0	10
38	ENSMUST00000046859_Foxn3-001	0	0	0	0	0	0
39	ENSMUST00000178342_Samt1-001	0	0	0	0	0	0
40	ENSMUST00000145073_Ppp1r16b-004	0	0	3	0	0	1
41	ENSMUST00000158586_Gm23387-201	0	0	0	0	0	0

Figure 8: Count table of transcript-level estimations.

Furthermore, a result page will show a summary of the "Create Count table" results (figure 9[112]). This page contains information about the reference transcript sequences, input FASTQ files and obtained results. The results summary can be generated via **Side Panel** → **Result Summary** and it can be exported in pdf.

[112] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598049050#Transcript-levelQuantification-figure9



Create Count Table (Transcript-level) Results

Name: Isoform-level

Input 1: Reference Transcript Sequences

111,706 sequences have been recovered from the fasta file 'mouse_ref.fa'.

Transcripts	Min Length	Max Length	Avg Length
111,706	9	106,824	1,701.424

Expression has been estimated both at gene-level and isoform-level. The relationship between genes and and their isoforms has been provided in the text file 'mouse_ref_mapping.txt'.

Genes	Min Isoforms	Max Isoforms	Avg Isoforms
46,517	1	59	2.401

Input 2: FASTQ Files

A total of 6 samples have been processed.

File Name	Sample Name	Format	Sequencing
SRR937568_1.fastq.gz, SRR937568_2.fastq.gz	SRR937568	gz	Paired-End
SRR937920_1.fastq.gz, SRR937920_2.fastq.gz	SRR937920	gz	Paired-End
SRR937558_1.fastq.gz, SRR937558_2.fastq.gz	SRR937558	gz	Paired-End
SRR937564_1.fastq.gz, SRR937564_2.fastq.gz	SRR937564	gz	Paired-End
SRR937927_1.fastq.gz, SRR937927_2.fastq.gz	SRR937927	gz	Paired-End
SRR937946_1.fastq.gz, SRR937946_2.fastq.gz	SRR937946	gz	Paired-End

Results Overview

Input Reads		Aligned Reads		Total	
Name	Total Records	Aligned 1 Time	Aligned Multiple Times	Overall Alignment Rate	Not Aligned
SRR937568	1,339,876	181,027 / 13.51%	204,695 / 15.28%	385,722 / 28.79%	954,154 / 71.21%
SRR937920	1,409,495	197,218 / 13.99%	211,454 / 15.00%	408,672 / 28.99%	1,000,823 / 71.01%
SRR937558	1,029,972	147,864 / 14.36%	162,279 / 15.76%	310,143 / 30.11%	719,829 / 69.89%
SRR937564	1,093,275	166,436 / 15.22%	189,339 / 17.32%	355,775 / 32.54%	737,500 / 67.46%
SRR937927	1,526,289	171,474 / 11.23%	195,269 / 12.79%	366,743 / 24.03%	1,159,546 / 75.97%
SRR937946	1,548,816	209,465 / 13.52%	236,487 / 15.27%	445,952 / 28.79%	1,102,864 / 71.21%

Figure 9: Result Summary

9.6.3 Charts and Statistics

Different statistical charts can be generated from the results. These provide additional information about the process of quantifying expression, as well as a quality assessment of the resulting counts. All these charts can be found under the **Side Panel** of the Count Table Viewer.

- **Library Size per Sample:** Bar chart showing the number of read counts aligned to genomic features contained in each sample (figure 10a[113]).
- **Distribution of Counts:** Box plot that allows seeing how counts are distributed within each sample for all the transcripts (figure 10b[114]). Features with 0 counts in all samples will be discarded for this chart. The binary logarithm of raw counts is represented.
- **Counts per Category:** Bar chart showing the number of reads of each input file sorted by different categories (figure 10c[115]). This chart and the next one are only available for count tables created by the "Create Count Table" tool within OmicsBox.
- Aligned Concordantly Exactly 1 Time: Reads that have been assigned once to a reference transcript.

[113] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598049050#Transcript-levelQuantification-figure10a

[114] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598049050#Transcript-levelQuantification-figure10b

[115] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598049050#Transcript-levelQuantification-figure10c

- Aligned Concordantly > Time: Reads that have been assigned to more than one reference transcript.
- Not Aligned: Reads that have not been assigned to any reference transcript.
- **Counts per Category (%):** The same chart as explained above in percentages (figure 10d[116]).

⚠ Last two charts are only available for count tables created by the "Create Count Table" tools within OmicsBox.

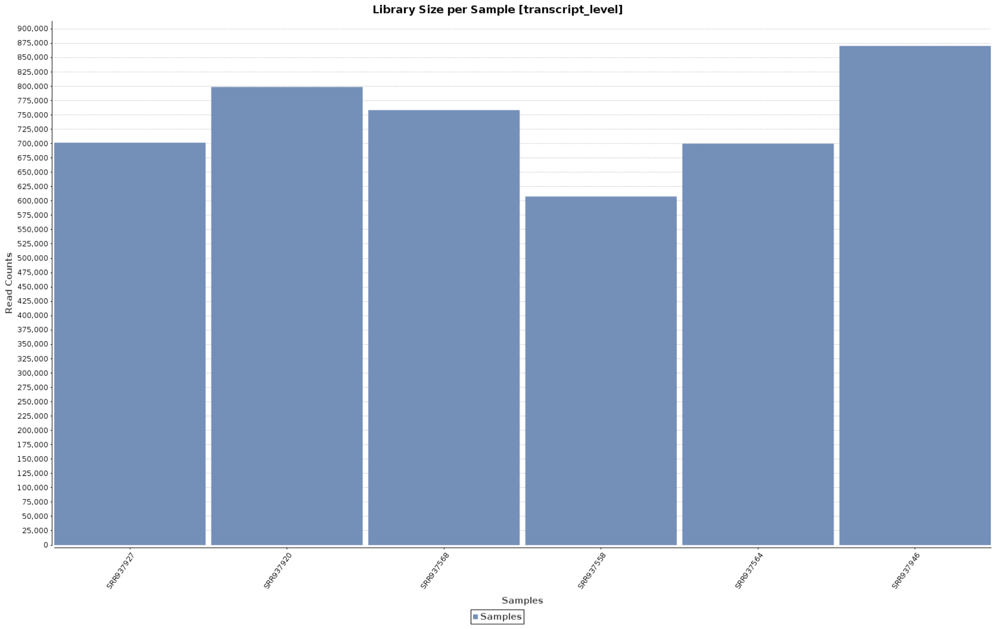


Figure 10 (a): Library Size per Sample

[116] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598049050#Transcript-levelQuantification-figure10d

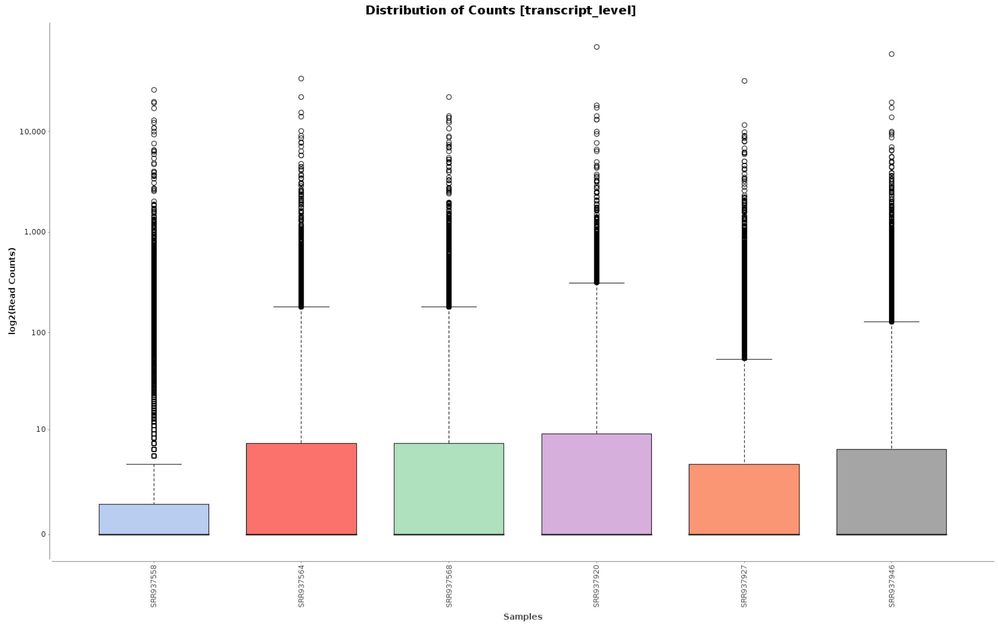


Figure 10 (b): Distribution of Counts

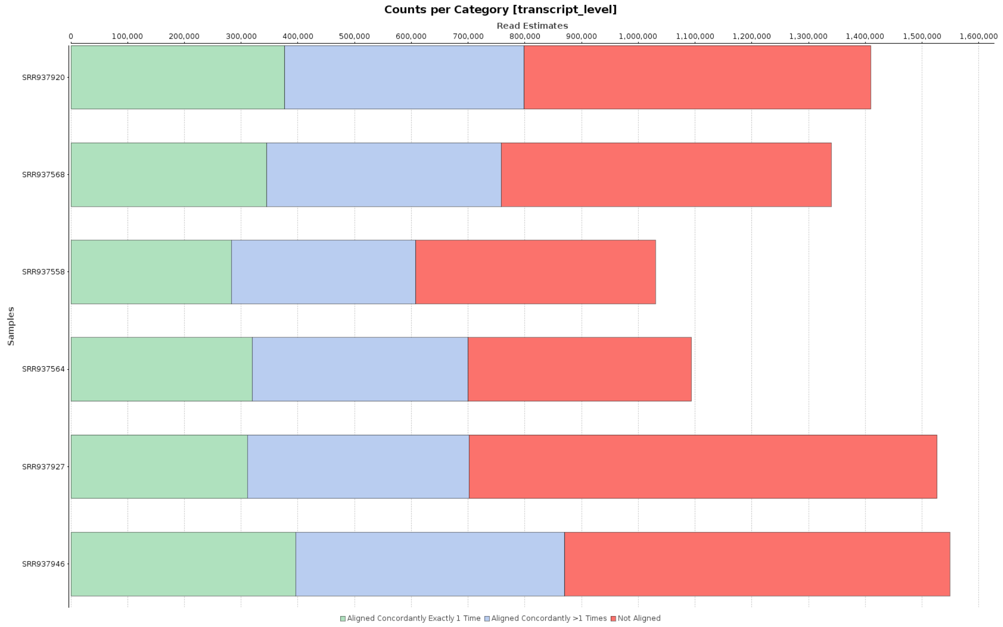


Figure 10 (c): Counts per Category

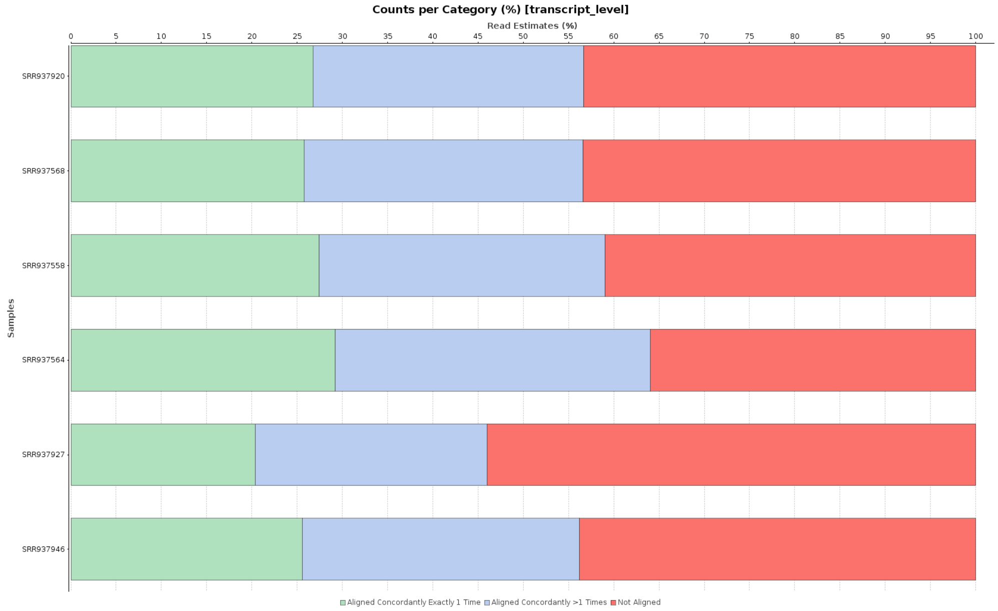


Figure 10 (d): Counts per Category (%)

9.7 Pairwise Differential Expression Analysis

Content of this page:

9.7.1 Introduction

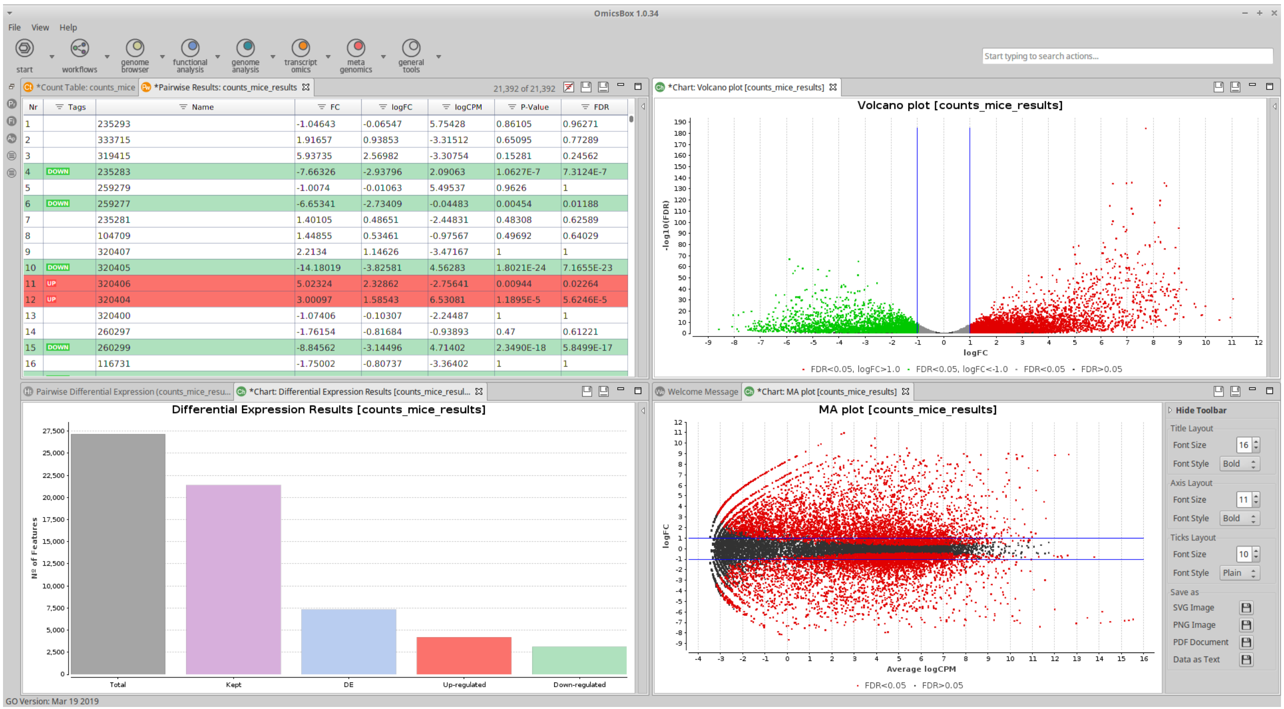
This tool is designed to perform differential expression analysis of count data arising from RNA-seq technology. This application, based on the edgeR program, allows identification of differentially expressed genomic features (e.g. genes) in a pairwise comparison of two different experimental conditions. The software package edgeR[117] (empirical analysis of DGE in R), which belongs to the Bioconductor project, implements quantitative statistical methods to evaluate the significance of individual genes between two experimental conditions.

Please cite edgeR as:

Robinson MD, McCarthy DJ and Smyth GK (2010). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." Bioinformatics, 26, pp. -1.[118]

9.7.2 General Workflow

The workflow for the analysis of differential expression is described in the following scheme (Figure 2(see page 195)).


OmicsBox 1.0.34
File
View
Help
start
workflows
genome browser
functional analysis
genome analysis
transcript omics
meta genomics
general tools
Start typing to search actions...
*Count Table: counts_mice
*Pairwise Results: counts_mice_results
21,392 of 21,392
*Chart: Volcano plot [counts_mice_results]
Volcano plot [counts_mice_results]
-log10(FDR)
logFC
FDR<0.05, logFC>1.0
FDR<0.05, logFC<-1.0
FDR<0.05
FDR>0.05
Pairwise Differential Expression (counts_mice_resu...
*Chart: Differential Expression Results [counts_mice_resul...
Differential Expression Results [counts_mice_results]
Nº of Features
Total
Kept
DE
Up-regulated
Down-regulated
Welcome Message
*Chart: MA plot [counts_mice_results]
MA plot [counts_mice_results]
logFC
Average logCPM
FDR<0.05
FDR>0.05
Hide Toolbar
GO Version: Mar 19 2019

Nr	Tags	Name	FC	logFC	logCPM	P-Value	FDR
1		235293	-1.04643	-0.06547	5.75428	0.86105	0.96271
2		333715	1.91657	0.93853	-3.31512	0.65095	0.77289
3		319415	5.93735	2.56982	-3.30754	0.15281	0.24562
4	DOWN	235283	-7.66326	-2.93796	2.09063	1.0627E-7	7.3124E-7
5		259279	-1.0074	-0.01063	5.49537	0.9626	1
6	DOWN	259277	-6.65341	-2.73409	-0.04483	0.00454	0.01188
7		235281	1.40105	0.48651	-2.44831	0.48308	0.62589
8		104709	1.44855	0.53461	-0.97567	0.49692	0.64029
9		320407	2.2134	1.14626	-3.47167	1	1
10	DOWN	320405	-14.18019	-3.82581	4.56283	1.8021E-24	7.1655E-23
11	UP	320406	5.02324	2.32862	-2.75641	0.00944	0.02264
12	UP	320404	3.00097	1.58543	6.53081	1.1895E-5	5.6246E-5
13		320400	-1.07406	-0.10307	-2.24487	1	1
14		260297	-1.76154	-0.81684	-0.93893	0.47	0.61221
15	DOWN	260299	-8.84562	-3.14496	4.71402	2.3490E-18	5.8499E-17
16		116731	-1.75002	-0.80737	-3.36402	1	1

Title Layout
Font Size 16
Font Style Bold
Axis Layout
Font Size 11
Font Style Bold
Ticks Layout
Font Size 10
Font Style Plain
Save as
SVG Image
PNG Image
PDF Document
Data as Text

Figure 1: Differential Expression Analysis Interface

[117] https://bioconductor.org/packages/release/bioc/html/edgeR.html
[118] https://www.ncbi.nlm.nih.gov/pubmed/?term=edgeR%3A+a+Bioconductor+package+for+differential+expression+analysis+of+digital+gene+expression+data.

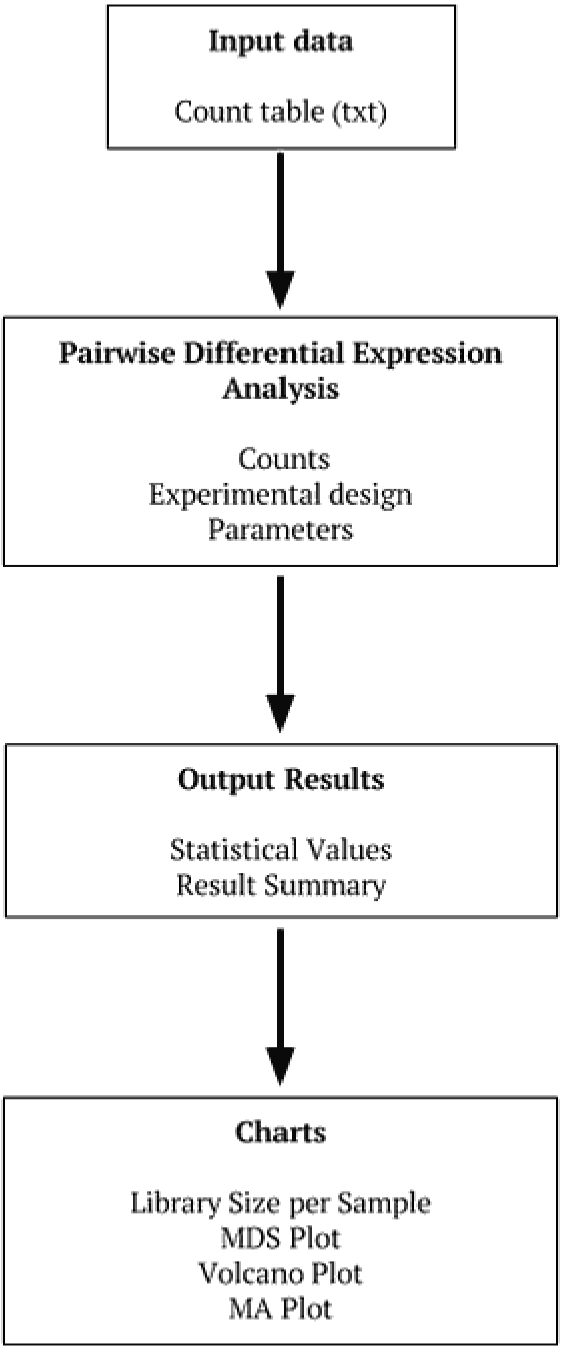
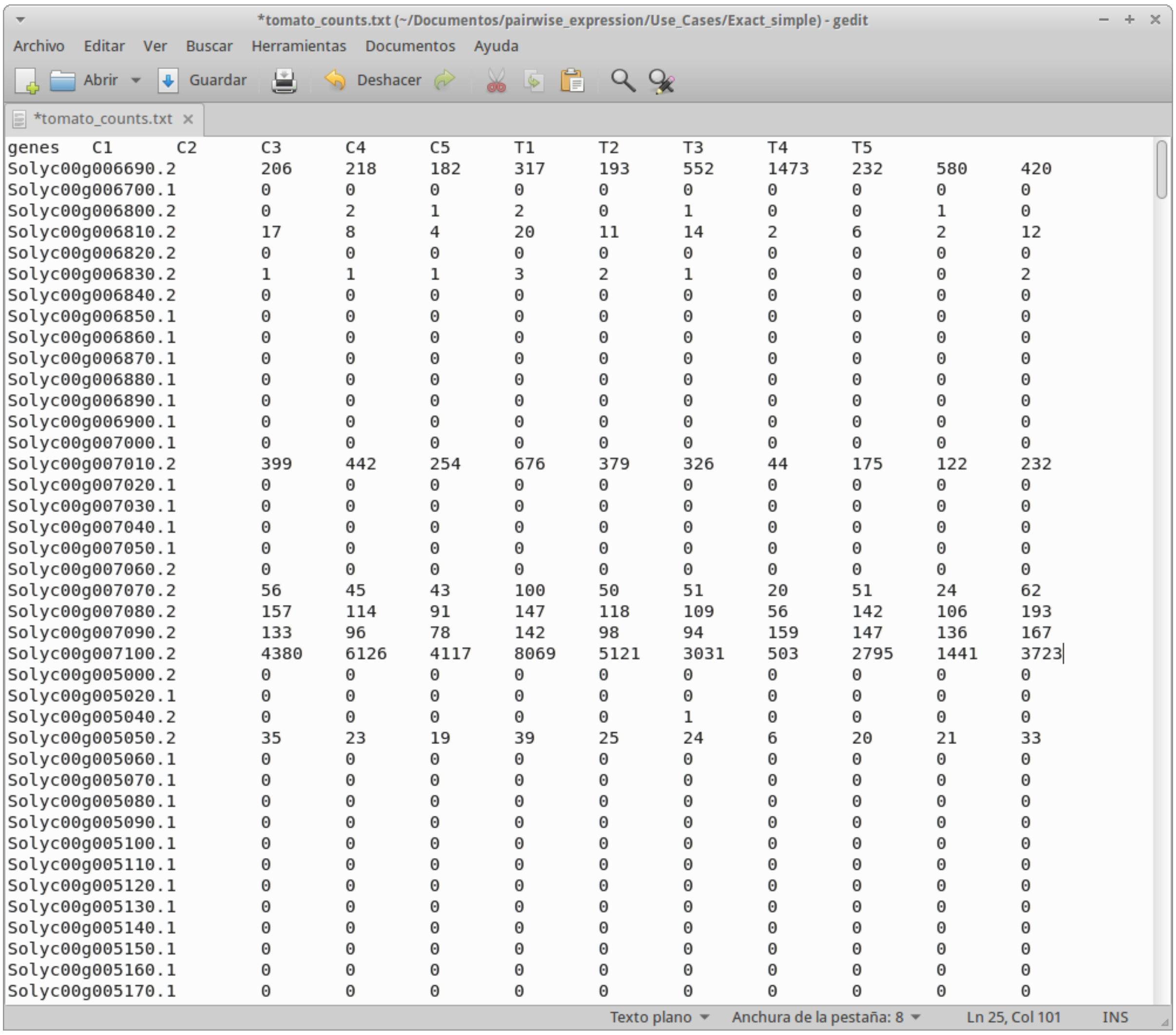


Figure 2: General Workflow

9.7.3 Load Data

Go to **File** → **Load** → **Load Count Table** and select your .txt file containing the count table in tab-delimited format (Figure 3(see page 197)). It is also possible to create a Count Table within OmicsBox through the "Create Count Table" functionality (see Quantify Expression(see page 172) section).



```
genes    C1       C2       C3       C4       C5       T1       T2       T3       T4       T5
Solyc00g006690.2        206      218      182      317      193      552      1473     232      580      420
Solyc00g006700.1        0        0        0        0        0        0        0        0        0        0
Solyc00g006800.2        0        2        1        2        0        1        0        0        1        0
Solyc00g006810.2        17       8        4        20       11       14       2        6        2        12
Solyc00g006820.2        0        0        0        0        0        0        0        0        0        0
Solyc00g006830.2        1        1        1        3        2        1        0        0        0        2
Solyc00g006840.2        0        0        0        0        0        0        0        0        0        0
Solyc00g006850.1        0        0        0        0        0        0        0        0        0        0
Solyc00g006860.1        0        0        0        0        0        0        0        0        0        0
Solyc00g006870.1        0        0        0        0        0        0        0        0        0        0
Solyc00g006880.1        0        0        0        0        0        0        0        0        0        0
Solyc00g006890.1        0        0        0        0        0        0        0        0        0        0
Solyc00g006900.1        0        0        0        0        0        0        0        0        0        0
Solyc00g007000.1        0        0        0        0        0        0        0        0        0        0
Solyc00g007010.2        399      442      254      676      379      326      44       175      122      232
Solyc00g007020.1        0        0        0        0        0        0        0        0        0        0
Solyc00g007030.1        0        0        0        0        0        0        0        0        0        0
Solyc00g007040.1        0        0        0        0        0        0        0        0        0        0
Solyc00g007050.1        0        0        0        0        0        0        0        0        0        0
Solyc00g007060.2        0        0        0        0        0        0        0        0        0        0
Solyc00g007070.2        56       45       43       100      50       51       20       51       24       62
Solyc00g007080.2        157      114      91       147      118      109      56       142      106      193
Solyc00g007090.2        133      96       78       142      98       94       159      147      136      167
Solyc00g007100.2        4380     6126     4117     8069     5121     3031     503      2795     1441     3723
Solyc00g005000.2        0        0        0        0        0        0        0        0        0        0
Solyc00g005020.1        0        0        0        0        0        0        0        0        0        0
Solyc00g005040.2        0        0        0        0        0        1        0        0        0        0
Solyc00g005050.2        35       23       19       39       25       24       6        20       21       33
Solyc00g005060.1        0        0        0        0        0        0        0        0        0        0
Solyc00g005070.1        0        0        0        0        0        0        0        0        0        0
Solyc00g005080.1        0        0        0        0        0        0        0        0        0        0
Solyc00g005090.1        0        0        0        0        0        0        0        0        0        0
Solyc00g005100.1        0        0        0        0        0        0        0        0        0        0
Solyc00g005110.1        0        0        0        0        0        0        0        0        0        0
Solyc00g005120.1        0        0        0        0        0        0        0        0        0        0
Solyc00g005130.1        0        0        0        0        0        0        0        0        0        0
Solyc00g005140.1        0        0        0        0        0        0        0        0        0        0
Solyc00g005150.1        0        0        0        0        0        0        0        0        0        0
Solyc00g005160.1        0        0        0        0        0        0        0        0        0        0
Solyc00g005170.1        0        0        0        0        0        0        0        0        0        0
```

Figure 3: Count Table File

The Count Table can be saved as 'Count Table' object (**File** → **Save**).

⚠ Notes:

- This application only accepts raw counts without any type of normalization.
- Replicates for each experimental condition are necessary.

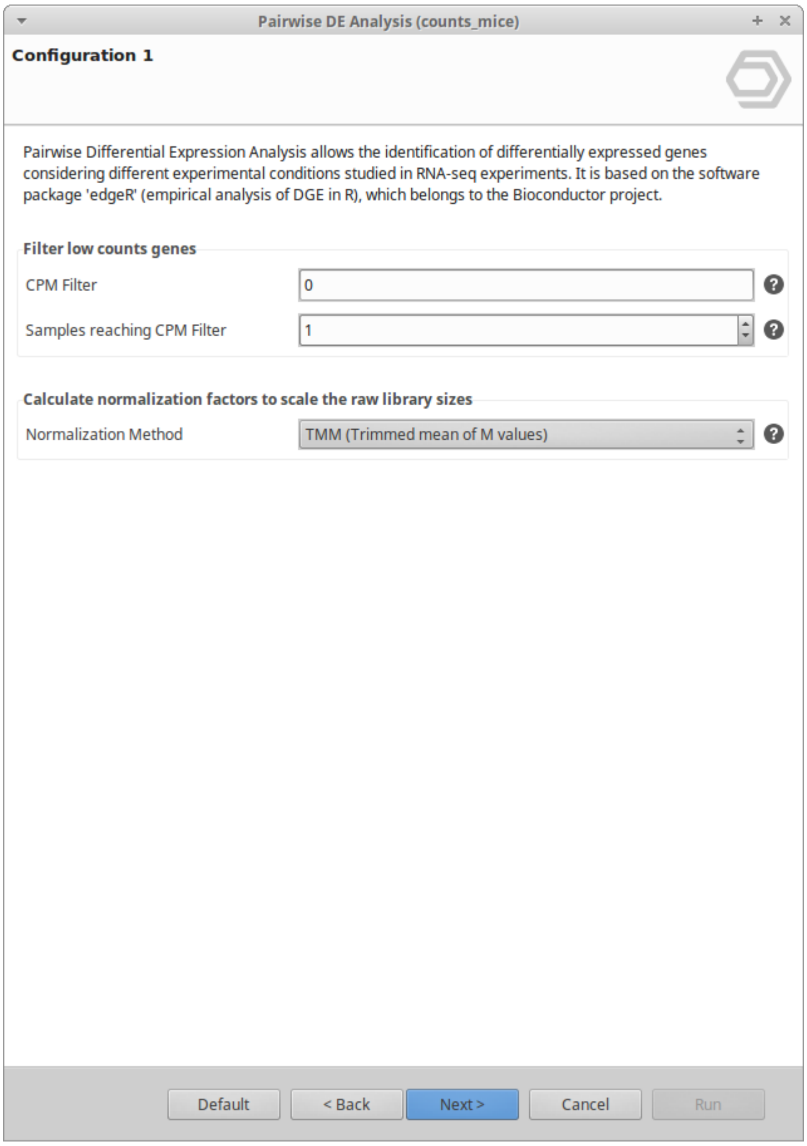
9.7.4 Run Pairwise Differential Expression Analysis

Go to **transcriptomics → Run Differential Expression Analysis** and choose the "Pairwise Differential Analysis" option. Here you can specify the following parameters, which are divided into three different sections: Preprocessing Data (Figure 4(see page 198)), Experimental Design (Figure 5(see page 200)) and Comparison and Test (Figure 6(see page 202)).

9.7.4.1 Preprocessing Data Page

- **Filter low count genes:**
    - **CPM Filter:** Establish a filter to exclude genes with low counts across libraries, as those genes may interfere with the subsequent statistical approximations. Filtering is performed on a count-per-million (CPM) basis to account for differences in library size between samples (e.g. a CPM of 1 corresponds to a count of 6 in a sample with 6 million reads).
    - **Samples reaching CPM Filter:** Set a minimum number of samples in which the gene's CPM is above the filter level (is expressed). If this value is set to e.g. five, at least 5 of the samples have to be above the given CPM. The number of samples of the smallest group is usually used (e.g. in an experiment that has two replicates for each condition (or group), a gene should be expressed in at least two samples). Set value to 0 if no filter is desired.
- **Calculate normalization factors to scale the raw library sizes:**
    - **Normalization Method:** Here the normalization takes the form of scaling factors for library sizes that enter into the statistical model. These correctional factors are used to compute the effective library sizes. For further details please refer to the edgeR User's Guide[119]. You can select the normalization method to be used:
        - **TMM:** Weighted trimmed mean of M-values. In this method, weights are obtained from the delta method on Binomial Data (this method is recommended).
        - **RLE:** Relative log expression. Scale factors are the median ratio of each sample to the median library (geometric mean of all samples).
        - **Upper-quartile:** 75% quantile for the counts for each library is used to calculate the scale factors.
        - **None:** No normalization method is applied.

[119] https://bioconductor.org/packages/release/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf



Configuration 1

Pairwise Differential Expression Analysis allows the identification of differentially expressed genes considering different experimental conditions studied in RNA-seq experiments. It is based on the software package 'edgeR' (empirical analysis of DGE in R), which belongs to the Bioconductor project.

Filter low counts genes

CPM Filter: 0

Samples reaching CPM Filter: 1

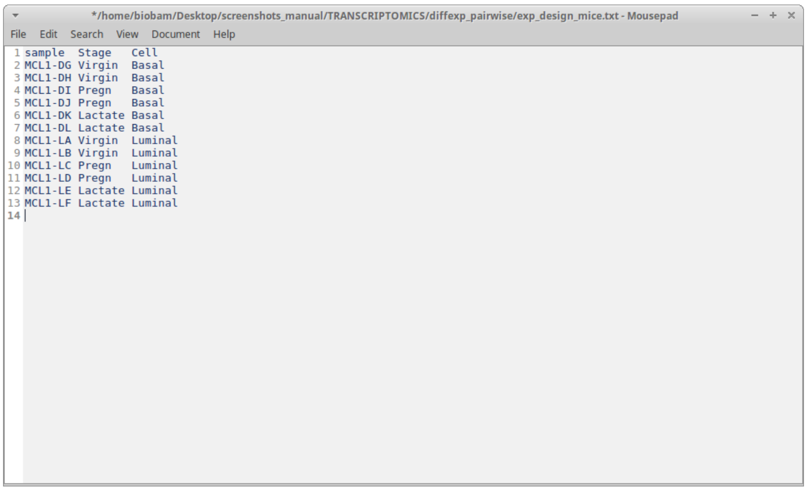
Calculate normalization factors to scale the raw library sizes

Normalization Method: TMM (Trimmed mean of M values)

Figure 4: Preprocessing Data Page

Experimental Design Page

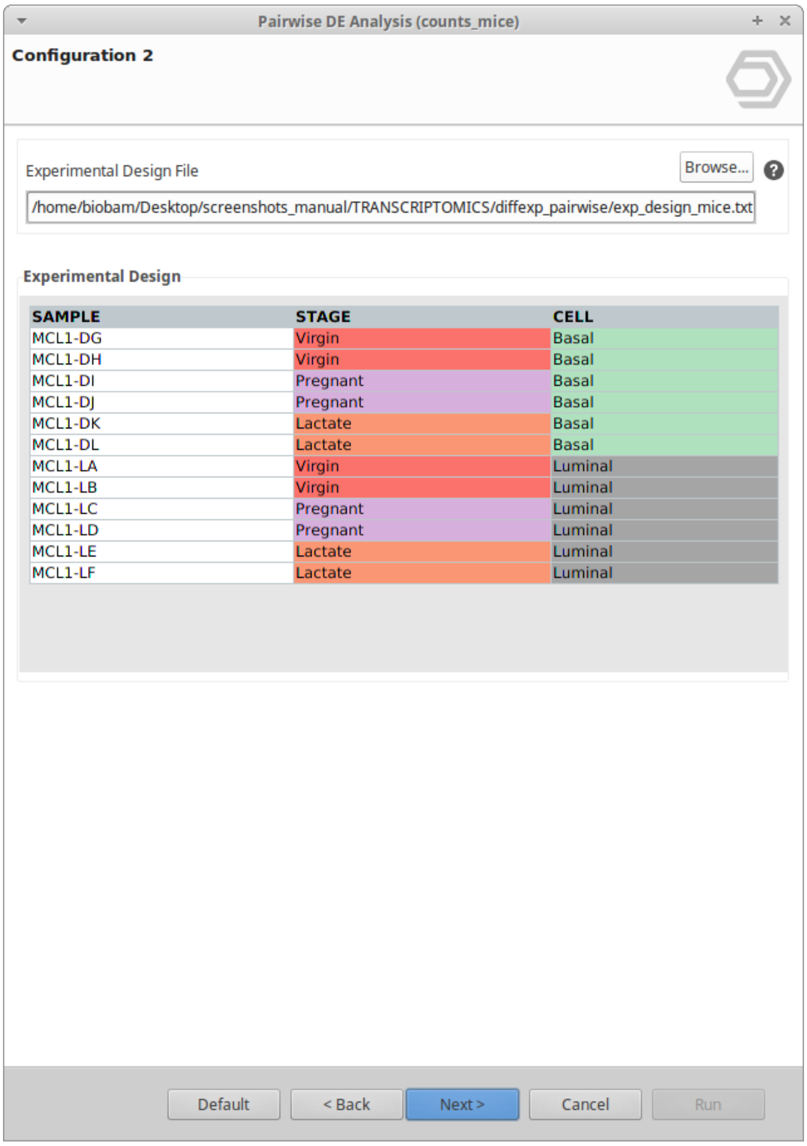
- **Experimental design file:** Select your .txt file containing your experimental factors with the experimental conditions associated with each sample in tab-delimited format. As demonstrated in Figure 7(see page 199), rows correspond to samples and columns to experimental factors. Make sure that the names in the first column of the experimental design table are exactly the same as the sample names in the count table header. If your experimental design file has fewer samples than in the count table, only the samples contained in this file will be analyzed.



*/home/biobam/Desktop/screenshots_manual/TRANSCRIPTOMICS/diffexp_pairwise/exp_design_mice.txt - Mousepad
File
Edit
Search
View
Document
Help

```
sample  Stage   Cell
MCL1-DG Virgin  Basal
MCL1-DH Virgin  Basal
MCL1-DI Pregn   Basal
MCL1-DJ Pregn   Basal
MCL1-DK Lactate Basal
MCL1-DL Lactate Basal
MCL1-LA Virgin  Luminal
MCL1-LB Virgin  Luminal
MCL1-LC Pregn   Luminal
MCL1-LD Pregn   Luminal
MCL1-LE Lactate Luminal
MCL1-LF Lactate Luminal
```

Figure 7: Experimental Design File



Experimental Design File

/home/biobam/Desktop/screenshots_manual/TRANSCRIPTOMICS/diffexp_pairwise/exp_design_mice.txt

Experimental Design

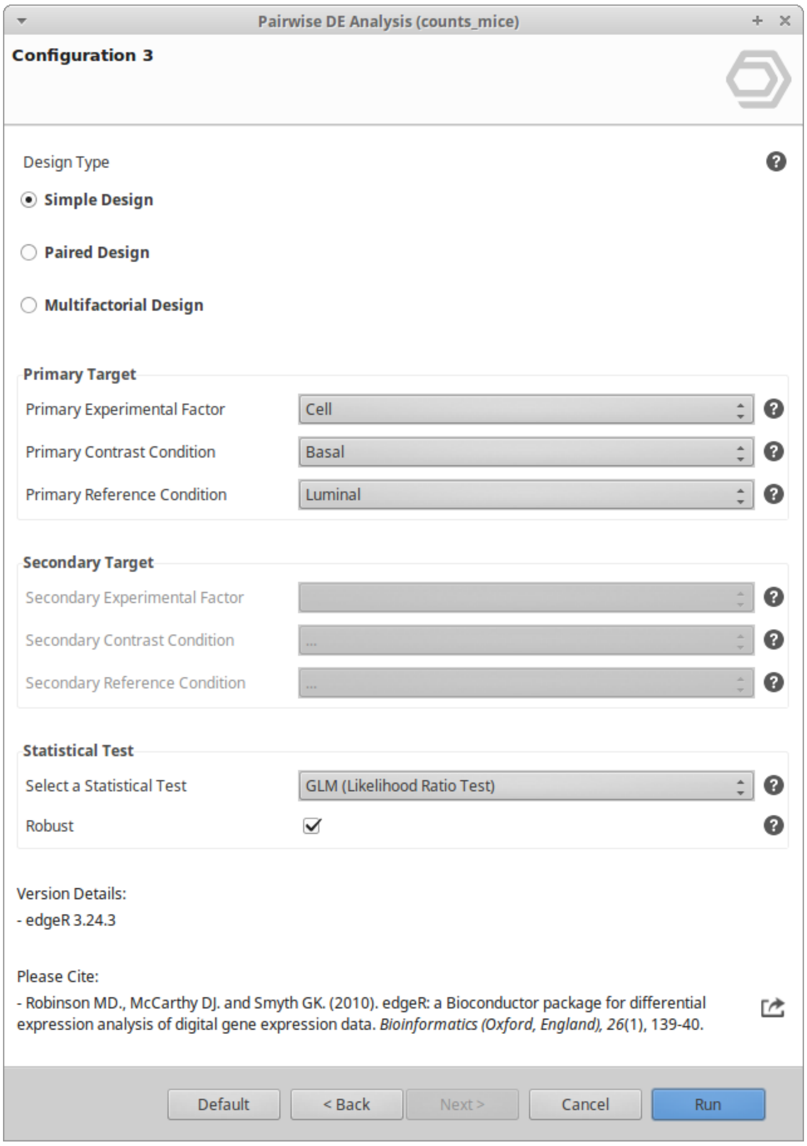
SAMPLE	STAGE	CELL
MCL1-DG	Virgin	Basal
MCL1-DH	Virgin	Basal
MCL1-DI	Pregnant	Basal
MCL1-DJ	Pregnant	Basal
MCL1-DK	Lactate	Basal
MCL1-DL	Lactate	Basal
MCL1-LA	Virgin	Luminal
MCL1-LB	Virgin	Luminal
MCL1-LC	Pregnant	Luminal
MCL1-LD	Pregnant	Luminal
MCL1-LE	Lactate	Luminal
MCL1-LF	Lactate	Luminal

Figure 5: Experimental Design Page

9.7.4.2 Comparison and Test Page

- **Design Type:** Choose the design type to adjust the analysis
    - Simple design: Makes a pairwise comparison between samples belonging to two experimental conditions. You only have to select the experimental factor of interest and establish the comparison selecting the reference and contrast conditions in ``Primary Target''.
    - Paired design: Makes a pairwise comparison between samples belonging to two experimental conditions, adjusting for baseline differences of other experimental factors. In this design, you have to establish the conditions for the comparison in ``Primary Target'' and the experimental factor for baseline difference in ``Secondary Target''. This design type is appropriate for paired or blocking design, or experiments with batch effects.

  - Multifactorial Design: Makes a pairwise comparison between samples belonging to two experimental conditions with two experimental factors. For this design, you have to select the two experimental factors of interest and establish the reference and contrast group for each in ``Primary Target'' and ``Secondary Target''. This design type is appropriate if you want to analyze the effects of combined experimental conditions on gene expression.
- **Statistical Test:**
  - Select a Statistical Test:
    - Exact Test: Based on the quantile-adjusted conditional maximum likelihood (qCML) methods (similar to Fisher's exact test). It is only applicable to datasets with a single factor design (simple design).
    - GLM (Likelihood Ratio Test): Based on fitting negative binomial Generalized Linear Models (GLMs) with the Cox-Reid dispersion estimates. Is a good choice for inferences with GLMs.
    - GLM (Quasi Likelihood F-Test): The empirical Bayes quasi-likelihood F-test is an alternative to the Likelihood Ratio Test and provides a more robust and reliable error rate control when the number of replicates is small.
  - Robust: Estimation is strengthened against potential outlier genes.



Configuration 3

Design Type

● Simple Design

○ Paired Design

○ Multifactorial Design

Primary Target

Primary Experimental Factor: Cell

Primary Contrast Condition: Basal

Primary Reference Condition: Luminal

Secondary Target

Secondary Experimental Factor:

Secondary Contrast Condition: ...

Secondary Reference Condition: ...

Statistical Test

Select a Statistical Test: GLM (Likelihood Ratio Test)

Robust ☑

Version Details:
- edgeR 3.24.3

Please Cite:
- Robinson MD., McCarthy DJ. and Smyth GK. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England), 26*(1), 139-40.

Figure 6: Comparison and Test Page

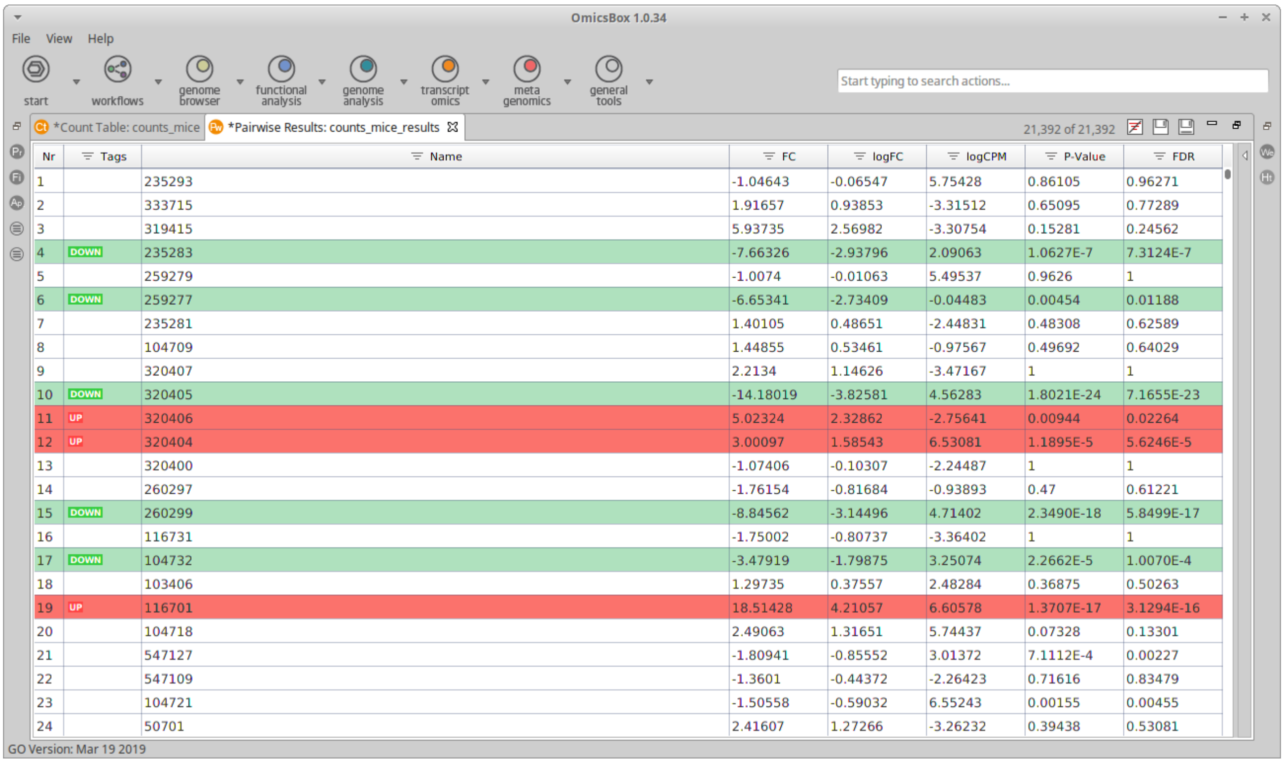
9.7.5 Results

Once the input counts have been processed and analyzed via the "Pairwise Differential Expression Analysis" tool, a new tab is opened containing the results (Figure 8(see page 203)). The results table contains the differential expression statistics, where each row corresponds to a feature:

- **logFC:** A measure that describes how much the expression changes between conditions (log2-fold-changes are shown).
- **logCPM:** The average log2-counts-per-millions.
- **LR:** Likelihood ratio statistic for the GLM (Likelihood Ratio Test).
- **F:** Quasi-likelihood F-statistic for the GLM (Quasi Likelihood F-test).
- **FDR:** False Discovery Rate calculated by the Benjamini-Hochberg method (multiple hypothesis testing corrections).

- **Tags:** Indicate whether a gene is upregulated (FDR ≤ 0.05, logFC ≥ 0) or downregulated (FDR ≤ 0.05, logFC ≥ 0).

Genes that have not passed the filtering step are not shown in the new tab.

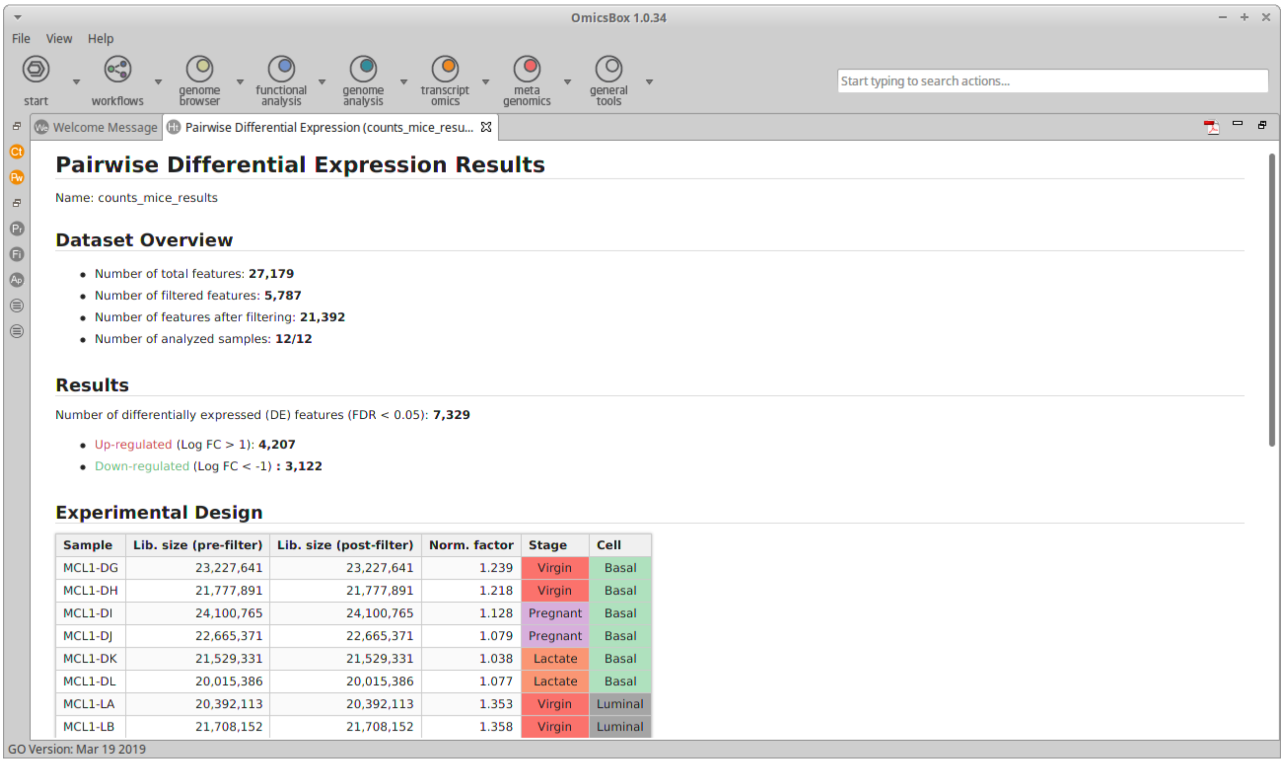


Nr	Tags	Name	FC	logFC	logCPM	P-Value	FDR
1		235293	-1.04643	-0.06547	5.75428	0.86105	0.96271
2		333715	1.91657	0.93853	-3.31512	0.65095	0.77289
3		319415	5.93735	2.56982	-3.30754	0.15281	0.24562
4	DOWN	235283	-7.66326	-2.93796	2.09063	1.0627E-7	7.3124E-7
5		259279	-1.0074	-0.01063	5.49537	0.9626	1
6	DOWN	259277	-6.65341	-2.73409	-0.04483	0.00454	0.01188
7		235281	1.40105	0.48651	-2.44831	0.48308	0.62589
8		104709	1.44855	0.53461	-0.97567	0.49692	0.64029
9		320407	2.2134	1.14626	-3.47167	1	1
10	DOWN	320405	-14.18019	-3.82581	4.56283	1.8021E-24	7.1655E-23
11	UP	320406	5.02324	2.32862	-2.75641	0.00944	0.02264
12	UP	320404	3.00097	1.58543	6.53081	1.1895E-5	5.6246E-5
13		320400	-1.07406	-0.10307	-2.24487	1	1
14		260297	-1.76154	-0.81684	-0.93893	0.47	0.61221
15	DOWN	260299	-8.84562	-3.14496	4.71402	2.3490E-18	5.8499E-17
16		116731	-1.75002	-0.80737	-3.36402	1	1
17	DOWN	104732	-3.47919	-1.79875	3.25074	2.2662E-5	1.0070E-4
18		103406	1.29735	0.37557	2.48284	0.36875	0.50263
19	UP	116701	18.51428	4.21057	6.60578	1.3707E-17	3.1294E-16
20		104718	2.49063	1.31651	5.74437	0.07328	0.13301
21		547127	-1.80941	-0.85552	3.01372	7.1112E-4	0.00227
22		547109	-1.3601	-0.44372	-2.26423	0.71616	0.83479
23		104721	-1.50558	-0.59032	6.55243	0.00155	0.00455
24		50701	2.41607	1.27266	-3.26232	0.39438	0.53081

Figure 8: Table Viewer

Results can be saved as a Pairwise Results object. Note that it is not possible to perform the analysis on this object. For this purpose, you have to open the Count Table object. If you want to see both count table and results, go to the File Manager and open the two .b2g files together.

A result page will show a summary of the pairwise differential expression analysis results (Figure 9(see page 204)).



Pairwise Differential Expression Results

Name: counts_mice_results

Dataset Overview

- Number of total features: **27,179**
- Number of filtered features: **5,787**
- Number of features after filtering: **21,392**
- Number of analyzed samples: **12/12**

Results

Number of differentially expressed (DE) features (FDR < 0.05): **7,329**

- Up-regulated (Log FC > 1): **4,207**
- Down-regulated (Log FC < -1) **: 3,122**

Experimental Design

Sample	Lib. size (pre-filter)	Lib. size (post-filter)	Norm. factor	Stage	Cell
MCL1-DG	23,227,641	23,227,641	1.239	Virgin	Basal
MCL1-DH	21,777,891	21,777,891	1.218	Virgin	Basal
MCL1-DI	24,100,765	24,100,765	1.128	Pregnant	Basal
MCL1-DJ	22,665,371	22,665,371	1.079	Pregnant	Basal
MCL1-DK	21,529,331	21,529,331	1.038	Lactate	Basal
MCL1-DL	20,015,386	20,015,386	1.077	Lactate	Basal
MCL1-LA	20,392,113	20,392,113	1.353	Virgin	Luminal
MCL1-LB	21,708,152	21,708,152	1.358	Virgin	Luminal

Figure 9: Results Summary

If you want to filter for differential expression based on other FDR and/or logFC cutoffs, you can go to **Side Panel** → **Set Up/Down Tags** and establish new values for both cutoffs. Tags will be updated, and the result section of the Result Summary and statistical charts will change according to the new cutoffs. To view, the updated summary results go to **Side Panel** → **Result Summary** and it can be exported in pdf.

9.7.6 Charts and Statistics

Different statistics charts can be generated for a global visualization of the results. These charts can be found under the **Side Panel** of the Pairwise Results viewer.

9.7.6.1 MDS Plot

Generates a two-dimensional scatterplot in which the distances represent the typical log2 fold changes between samples. You can select an experimental factor by which you want to color the MDS graphic (Figure 10(a)(see page 206)).

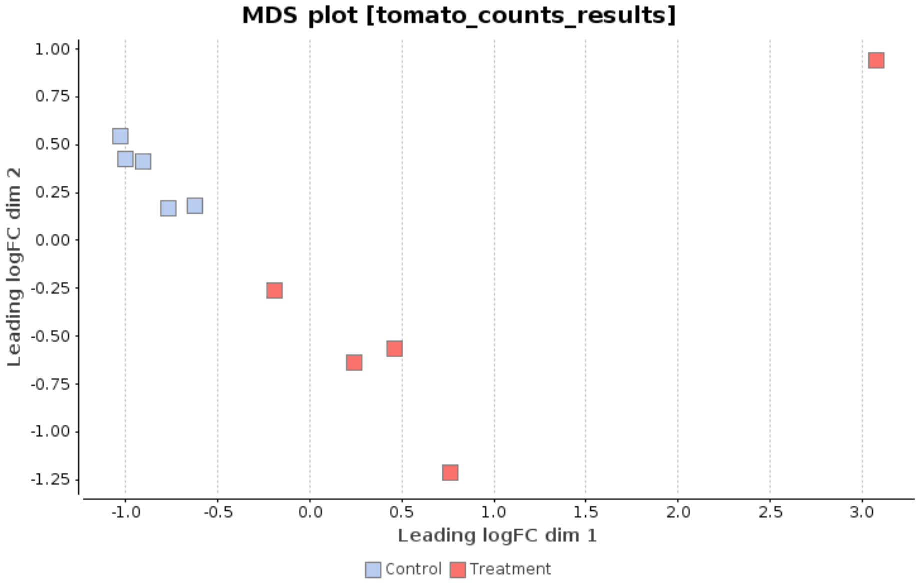


Figure 10(a): MDS Plot

9.7.6.2 Results Chart

Bar chart which shows the number of total features, kept features (those who have passed the filtering step), differentially expressed features, up-regulated features and down-regulated features (Figure 10(b)(see page 206)).

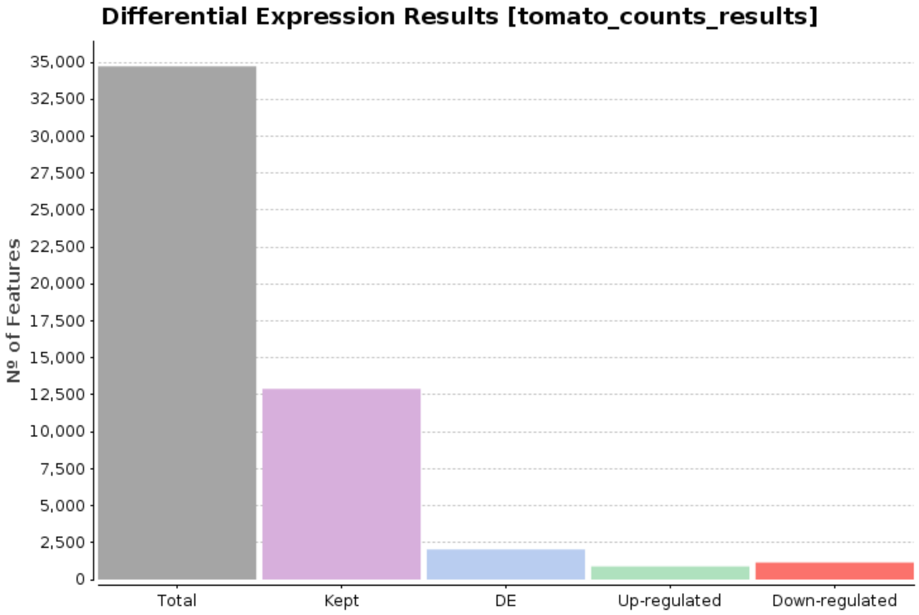


Figure 10(b): Result Summary

9.7.6.3 Volcano Plot

A scatter plot that is constructed by plotting the negative log of the adjusted p-values (FDR) on the y-axis versus the log of the fold changes on the x-axis (Figure 10(c)(see page 207)). Upregulated and downregulated genes are shown in green and red respectively.

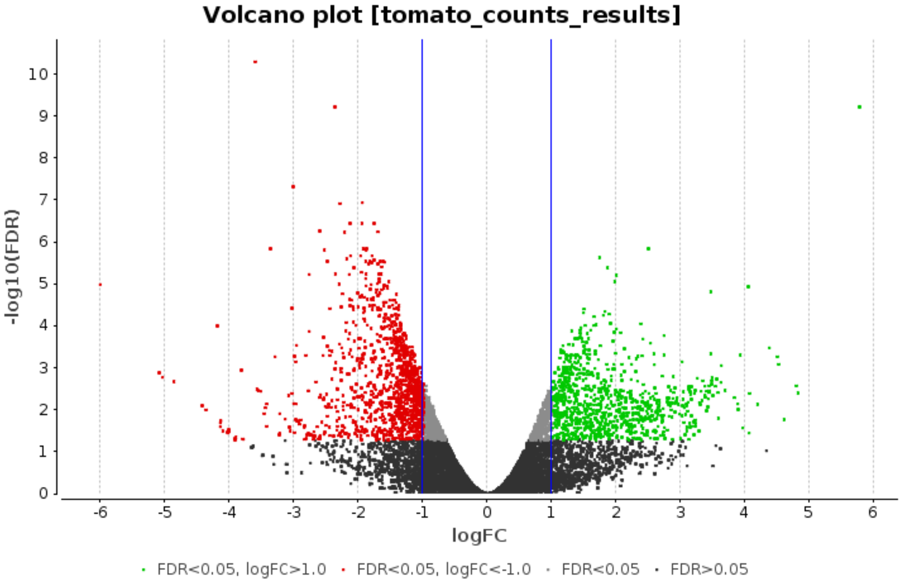


Figure 10(c): Volcano Plot

9.7.6.4 MA Plot

A scatter plot showing the log of the fold changes on the y-axis versus the average of the log of the CPM on the x-axis. Differentially expressed genes are highlighted (Figure 10(d)(see page 207)).

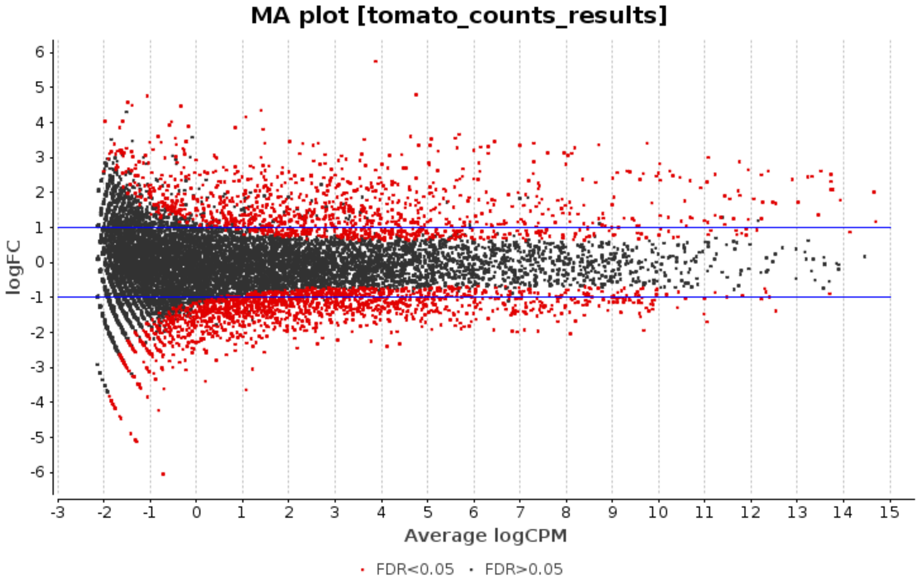


Figure 10(d): MA Plot

9.7.6.5 Heatmap

A heatmap is a two-dimensional visual representation of data in which numerical values of points are represented by a range of colors (Figure 10 (e)(see page 208)). The dendrograms added to the left and top side are produced by a hierarchical clustering method that takes as input the Euclidean distance computed between genes (left) and samples (top).

The heatmap supports zooming by keeping clicked a node of either of the two dendrograms. The first bars contain the experimental design of the data showing the association between samples and experimental covariates.

Genes that will be displayed can be selected in the wizard. There are three options:

- The Top 50 differentially expressed genes (ranked by FDR).
- All differentially expressed genes.
- Provide an ID list containing the genes to represent.

⚠ Differentially expressed genes are those that are labeled as UP or DOWN in the table project ("Tags" column). The criteria for considering a gene as differentially expressed can be adjusted using the option "Set Up/Down Tags".

Furthermore, the wizard allows adjusting the type of expression data that will be represented, as well as the transformation that can be applied to this data.

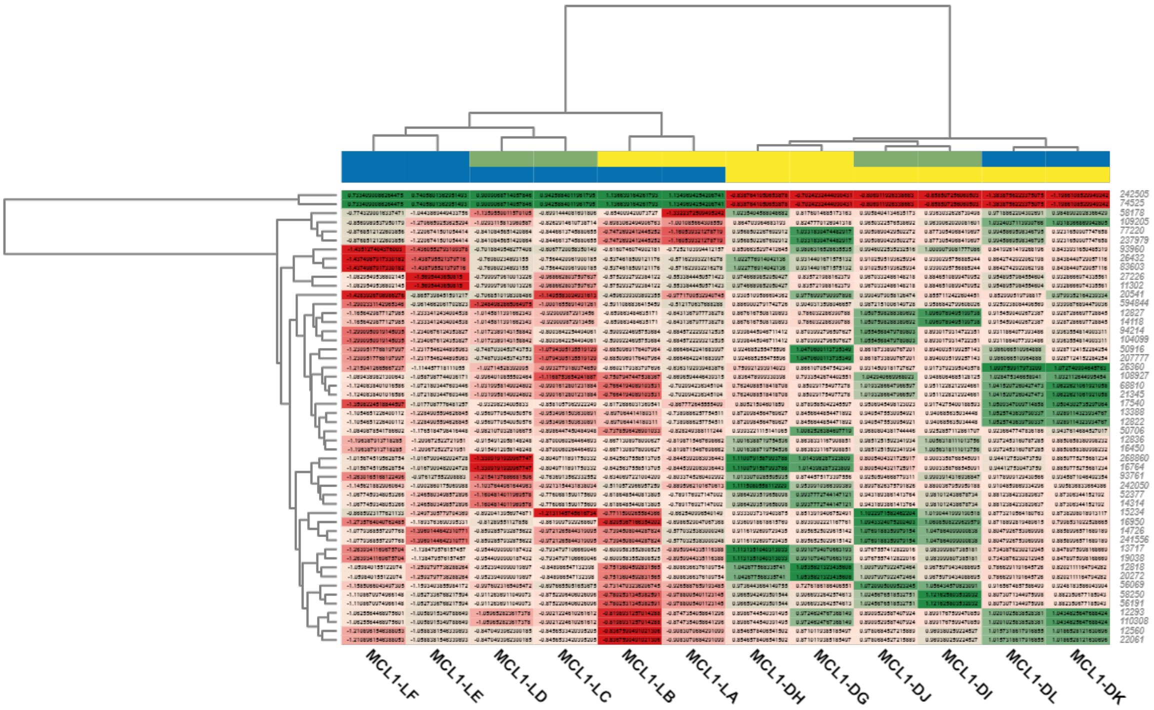


Figure 10(e): Heatmap

9.7.7 Enrichment Analysis

It is possible to perform a functional enrichment analysis from the pairwise differential expression project. Both options, Fisher's Exact Test and Gene Set Enrichment Analysis, can be found under the **Side Panel** of the Pairwise Results viewer. For a detailed tutorial on how to launch an Enrichment Analysis from Pairwise Differential Expression results, link here[120].

9.7.7.1 Fisher's Exact Test

Choose the subset of genes that will be considered as Test-set. Up-regulated and down-regulated genes are those that are tagged according to the criteria established by the option "Set Up/Down Tags".

The project containing the functionally annotated sequences that will be used as a reference background set should be provided.

The rest of the parameters are explained in the Fisher's Exact Test section(see page 97).

9.7.7.2 Gene Set Enrichment Analysis

The "FDR Filter for Ranked List" parameters allows setting a filter to exclude those genes whose FDR is above it. The ranked gene list will be created using the logFC statistic.

[120] https://www.biobam.com/how-to-launch-an-enrichment-analysis-from-pairwise-differential-expression-results-in-blast2go/

The project containing the functionally annotated sequences that will be used as a reference background set should be provided.

The rest of the parameters are explained in the Gene Set Enrichment Analysis section(see page 101).

9.8 Pairwise Differential Expression Analysis (Without Replicates)

Content of this page:

9.8.1 Introduction

Detecting genes that are differentially expressed between two experimental conditions (e.g. diseased vs healthy individuals) is a fundamental part of understanding the molecular basis of phenotypic variation. To carry out this task, there are statistical tools designed to perform differential expression analysis of the count data arising from RNA-seq technology (e.g. edgeR and maSigPro are statistical packages integrated into OmicsBox). However, these tools usually require the presence of replicates (both biological and technical) of each experimental condition that will be tested. This is a problem in cases where no replicates are available.

The **Pairwise Differential Expression Analysis (Without Replicates)** functionality offers a strategy for analyzing RNA-seq datasets that do not have replicates. It is based on the software package **NOISeq**[121], which belongs to the Bioconductor project. NOISeq is a novel nonparametric approach for the identification for differentially expressed genes from RNA-Seq count data. NOISeq creates a null or noise distribution of count changes by contrasting fold-change differences (M) and absolute expression differences (D) for all the genes in samples within the same condition. This reference distribution is then used to assess whether the M and D values computed between two conditions for a given gene are likely to be part of the noise or represent a true differential expression.

NOISeq method was designed to compute differential expression on RNA-Seq data even when there are no replicates available for any of the experimental conditions. In this scenario, NOISeq can simulate technical replicates. The simulation relies on the assumption that read counts follow a multinomial distribution, where

[121] https://www.bioconductor.org/packages/release/bioc/html/NOISeq.html

probabilities for each class (feature) in the multinomial distribution are the probability of a read to map to that feature. These mapping probabilities are approximated by using counts in the only sample of the corresponding experimental conditions. Given the sequencing depth (total amount of reads) of the unique available sample, the size of the simulated is a percentage of this sequencing depth, allowing a small variability.

Please remember that to obtain really reliable statistical results, biological replicates are needed.

Please cite NOISeq as:

Tarazona S, Furio-Tari P, Turra D, Di Pietro A, Nueda MJ, Ferrer A and Conesa, A (2015). "Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package." Nucleic Acids Research, 43(21), e140.[122]

Tarazona S, Garcia-Alcalde F, Dopazo J, Ferrer A and Conesa A (2011). "Differential expression in RNA-seq: a matter of depth." Genome Research, 21(12), 2213-2223.[123]

9.8.2 Run Pairwise Differential Expression Analysis (Without Replicates)

Go to **transcriptomics** → **Run Differential Expression Analysis** and choose the "Pairwise Differential Expression Analysis (Without replicates) option. This application requires a **Count Table** object as input data. Use the Gene-level(see page 172) or Transcript-level(see page 182) expression quantification functionality to obtain a count table from RNA-Seq data. It is also possible to load a count table from a tabular text file (go to File → Load → Load Count Table). The wizard allows to adjust analysis parameters (Figure 1[124] and Figure 2[125]).

9.8.2.1 Preprocessing Data

- **Filter low count genes:**
  - **CPM FIlter:** Establish a filter to exclude genes with low counts across libraries, as those genes may interfere with the subsequent statistical approximations. Filtering is performed on a count-per-million (CPM) basis to account for differences in library size between samples (e.g. a CPM of 1 corresponds to a count of 6 in a sample with 6 million reads). To pass the filter, the gene's CPM should be above the filter level in at least one sample (contrast or reference sample).
- **Normalization procedure:**
  - **Normalization Method:** Normalization is an important step to make the samples comparable and to remove possible biases (as sequencing depth bias) in count data. The normalization methods available for this analysis are:
    - **TMM:** Weighted trimmed mean of M-values. In this method, weights are obtained from the delta method on Binomial Data (this method is recommended).
    - **RPKM:** Reads Per Kilobase per Million mapped reads. This method corrects for gene length and the number of sequencing reads (gene length is required).

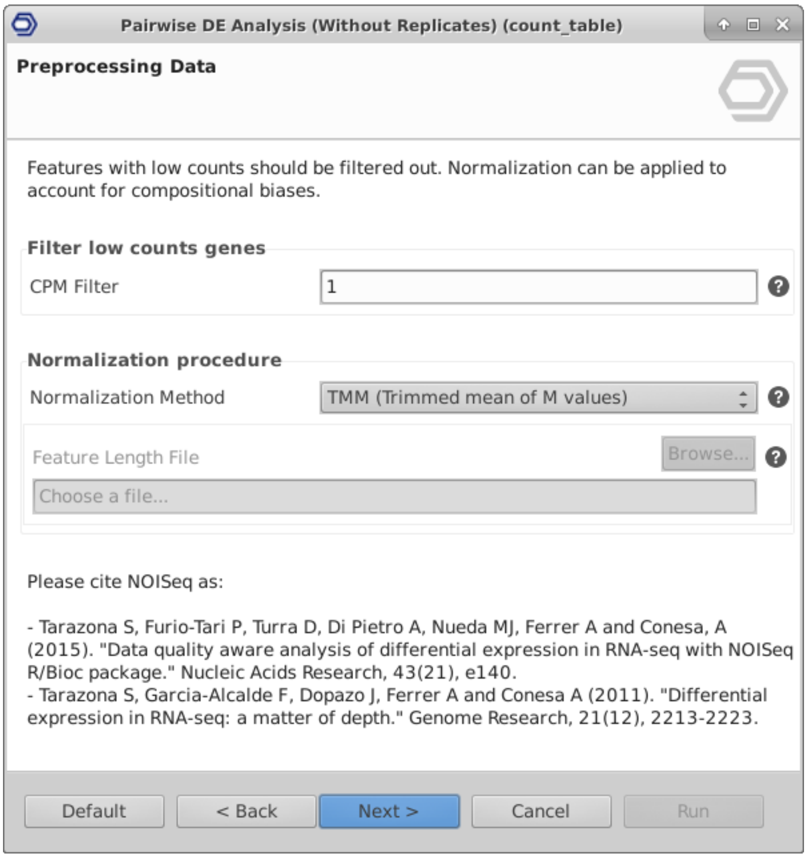
[122] https://www.ncbi.nlm.nih.gov/pubmed/?term=Data+quality+aware+analysis+of+differential+expression+in+RNA-seq+with+NOISeq+R%2FBioc+package

[123] https://www.ncbi.nlm.nih.gov/pubmed/21903743

[124] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=620232718#PairwiseDifferentialExpressionAnalysis(WithoutReplicates)-figure1

[125] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=620232718#PairwiseDifferentialExpressionAnalysis(WithoutReplicates)-figure2

    - **Upper-quartile:** 75% quantile for the counts for each library is used to calculate the scale factors for normalization.
    - **None:** No normalization method is applied.
  - **Feature Length File**: For RPKM normalization load a tab-delimited file (or ID-Value object) with two columns containing the name and length of each gene or genomic feature.



Pairwise DE Analysis (Without Replicates) (count_table)

Preprocessing Data

Features with low counts should be filtered out. Normalization can be applied to account for compositional biases.

Filter low counts genes

CPM Filter: 1

Normalization procedure

Normalization Method: TMM (Trimmed mean of M values)

Feature Length File: Choose a file...

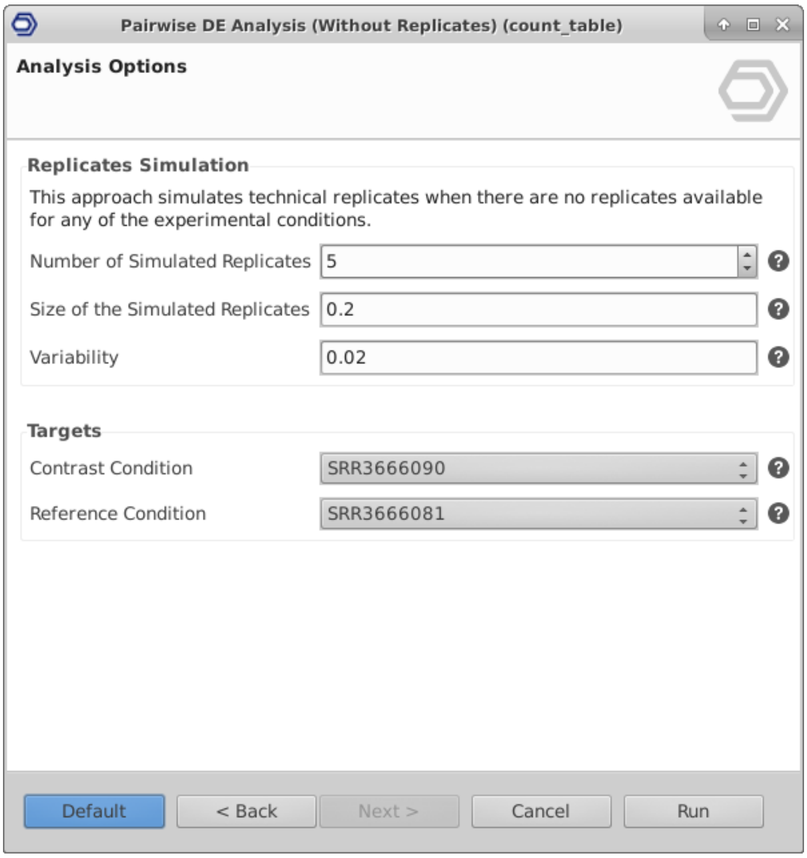
Please cite NOISeq as:

- Tarazona S, Furio-Tari P, Turra D, Di Pietro A, Nueda MJ, Ferrer A and Conesa, A (2015). "Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package." Nucleic Acids Research, 43(21), e140.
- Tarazona S, Garcia-Alcalde F, Dopazo J, Ferrer A and Conesa A (2011). "Differential expression in RNA-seq: a matter of depth." Genome Research, 21(12), 2213-2223.

Figure 1: Preprocessing Data Page

9.8.2.2 Analysis Options

- **Replicates Simulation:**
  - **Number of Simulated Replicates:** Set the number of replicates to be simulated for each condition.
  - **Size of the Simulated Replicates:** Establish the percentage of the total reads used to simulate each sample.
  - **Variability:** Variability in the simulated sample total reads.
- **Targets:**
  - **Contrast Condition:** Choose the sample to be treated as contrast condition. Genes classified as UP will be upregulated in this sample.
  - **Reference Condition:** Choose the sample to be treated as reference condition. Genes classified as DOWN will be upregulated in this sample.



Pairwise DE Analysis (Without Replicates) (count_table)

Analysis Options

Replicates Simulation

This approach simulates technical replicates when there are no replicates available for any of the experimental conditions.

Number of Simulated Replicates: 5

Size of the Simulated Replicates: 0.2

Variability: 0.02

Targets

Contrast Condition: SRR3666090

Reference Condition: SRR3666081

Figure 2: Analysis Options Page

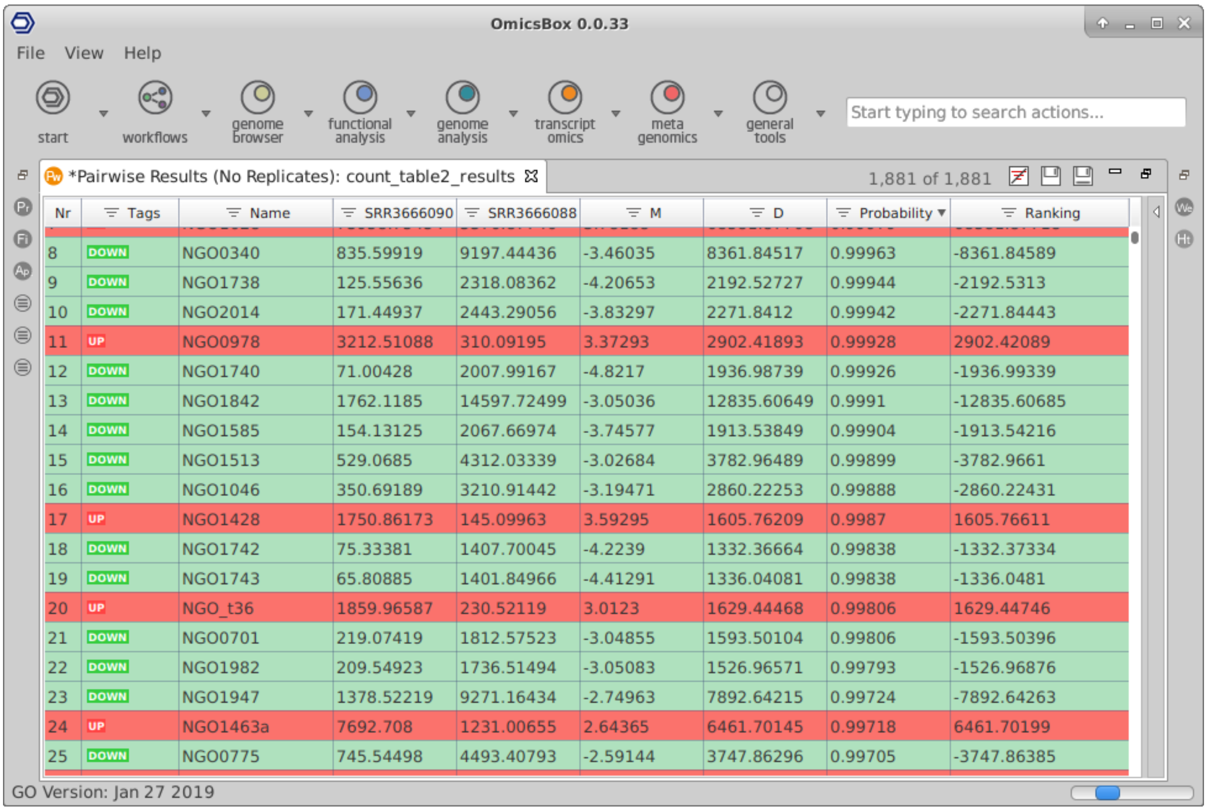
9.8.3 Results

Once the input counts have been processed and analyzed via the "Pairwise Differential Expression Analysis (Without Replicates)" feature, a new tab is opened containing the results (Figure 3[126]). The results table contains the differential expression statistics, where each row corresponds to a feature:

- **Contrast Condition:** Normalized expression values for the contrast condition sample.
- **Reference Condition:** Normalized expression values for the reference condition sample.
- **M:** Is the log2-ratio of the two conditions.
- **D:** The value of the difference between the conditions.
- **Probability:** The probability of differential expression for each feature. It is obtained by comparing the M and D values of a given feature against the noise distribution. If the probability is higher than a given threshold (0.9 by default), the feature is considered to be differentially expressed between conditions.
- **Ranking:** Is a summary statistic of M and D values equal to -sign(M)*sqrt(M^2 + D^2), which can be used as a ranked value in gene set enrichment analysis (GSEA).

Genes that have not passed the filtering step are not shown in the results tab.

[126] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=620232718#PairwiseDifferentialExpressionAnalysis(WithoutReplicates)-figure3



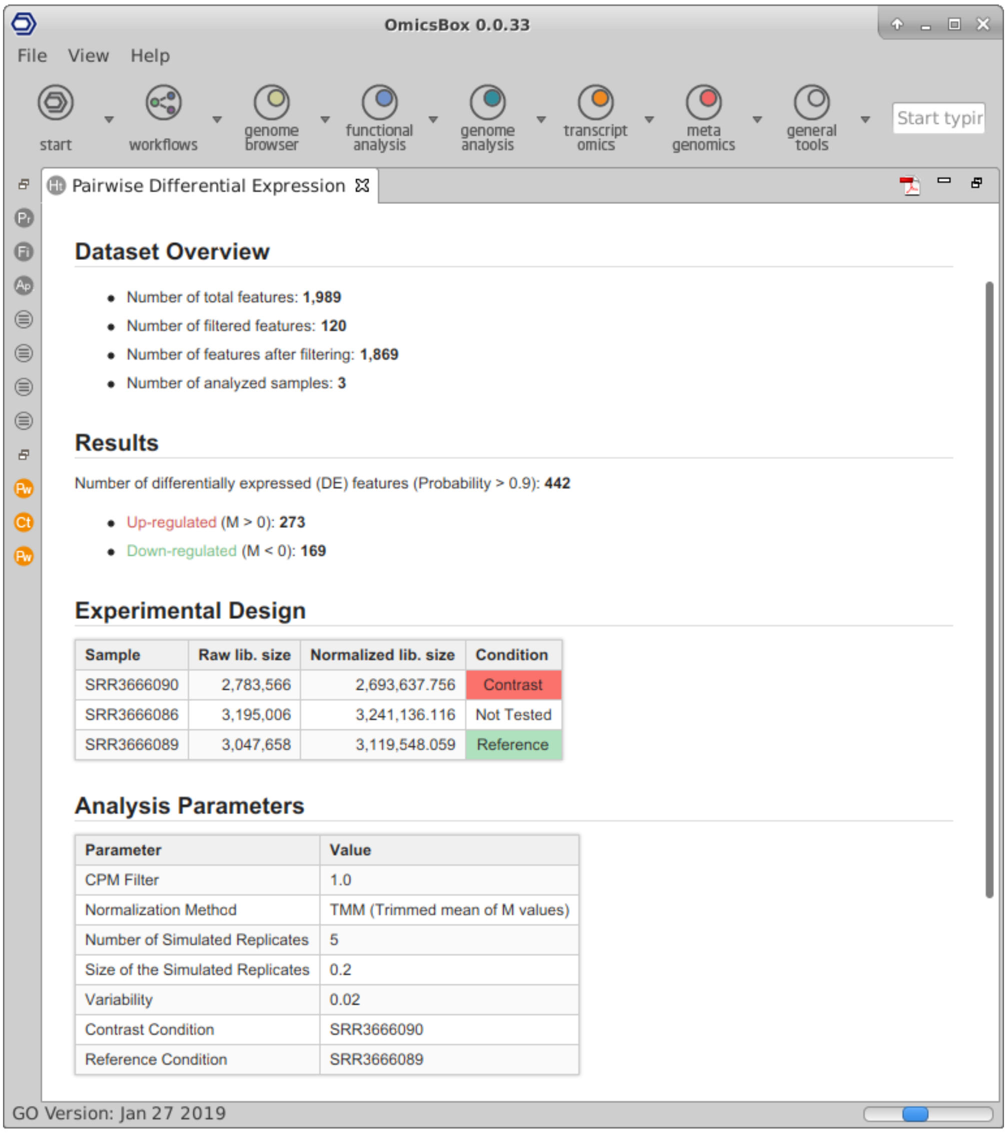
Nr	Tags	Name	SRR3666090	SRR3666088	M	D	Probability	Ranking
8	DOWN	NGO0340	835.59919	9197.44436	-3.46035	8361.84517	0.99963	-8361.84589
9	DOWN	NGO1738	125.55636	2318.08362	-4.20653	2192.52727	0.99944	-2192.5313
10	DOWN	NGO2014	171.44937	2443.29056	-3.83297	2271.8412	0.99942	-2271.84443
11	UP	NGO0978	3212.51088	310.09195	3.37293	2902.41893	0.99928	2902.42089
12	DOWN	NGO1740	71.00428	2007.99167	-4.8217	1936.98739	0.99926	-1936.99339
13	DOWN	NGO1842	1762.1185	14597.72499	-3.05036	12835.60649	0.9991	-12835.60685
14	DOWN	NGO1585	154.13125	2067.66974	-3.74577	1913.53849	0.99904	-1913.54216
15	DOWN	NGO1513	529.0685	4312.03339	-3.02684	3782.96489	0.99899	-3782.9661
16	DOWN	NGO1046	350.69189	3210.91442	-3.19471	2860.22253	0.99888	-2860.22431
17	UP	NGO1428	1750.86173	145.09963	3.59295	1605.76209	0.9987	1605.76611
18	DOWN	NGO1742	75.33381	1407.70045	-4.2239	1332.36664	0.99838	-1332.37334
19	DOWN	NGO1743	65.80885	1401.84966	-4.41291	1336.04081	0.99838	-1336.0481
20	UP	NGO_t36	1859.96587	230.52119	3.0123	1629.44468	0.99806	1629.44746
21	DOWN	NGO0701	219.07419	1812.57523	-3.04855	1593.50104	0.99806	-1593.50396
22	DOWN	NGO1982	209.54923	1736.51494	-3.05083	1526.96571	0.99793	-1526.96876
23	DOWN	NGO1947	1378.52219	9271.16434	-2.74963	7892.64215	0.99724	-7892.64263
24	UP	NGO1463a	7692.708	1231.00655	2.64365	6461.70145	0.99718	6461.70199
25	DOWN	NGO0775	745.54498	4493.40793	-2.59144	3747.86296	0.99705	-3747.86385

Figure 3: Table Viewer

Results can be saved as a Pairwise Results object. Note that it is not possible to perform the analysis on this object. For this purpose, you have to open the Count Table object. If you want to see both count table and results, go to the File Manager and open the two .box files together.

A result page will show a summary of the pairwise differential expression analysis results (Figure 4[127]).

[127] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=620232718#PairwiseDifferentialExpressionAnalysis(WithoutReplicates)-figure1



Dataset Overview

- Number of total features: **1,989**
- Number of filtered features: **120**
- Number of features after filtering: **1,869**
- Number of analyzed samples: **3**

Results

Number of differentially expressed (DE) features (Probability > 0.9): **442**

- Up-regulated (M > 0): **273**
- Down-regulated (M < 0): **169**

Experimental Design

Sample	Raw lib. size	Normalized lib. size	Condition
SRR3666090	2,783,566	2,693,637.756	Contrast
SRR3666086	3,195,006	3,241,136.116	Not Tested
SRR3666089	3,047,658	3,119,548.059	Reference

Analysis Parameters

Parameter	Value
CPM Filter	1.0
Normalization Method	TMM (Trimmed mean of M values)
Number of Simulated Replicates	5
Size of the Simulated Replicates	0.2
Variability	0.02
Contrast Condition	SRR3666090
Reference Condition	SRR3666089

Figure 4: Results Summary

If you want to filter for differential expression based on another probability threshold, you can go to **Side Panel** → **Set Up/Down Tags** and establish a new threshold value. Tags will be updated, and the result section of the Result Summary and statistical charts will change according to the new cutoffs. To view, the updated summary results go to **Side Panel** → **Result Summary** and it can be exported in PDF.

During the Pairwise Differential Expression Analysis (without replicates), raw counts are transformed according to the normalization method selected in the analysis configuration. Go to **Export Normalized Counts** (side bar) to export normalized counts to a tabular text file.

9.8.4 Charts and Statistics

Different statistics charts can be generated for a global visualization of the results. These charts can be found under the **Side Panel** of the Pairwise Results Viewer.

9.8.4.1 MDS Plot

It generates a two-dimensional scatterplot in which the distances represent the typical log2 fold changes between samples. You can select an experimental factor by which you want to color the MDS graphic (Figure 5(a)[128]).

> ⓘ This plot is only available if the input count table contains more than 2 samples (although only two of them are compared).

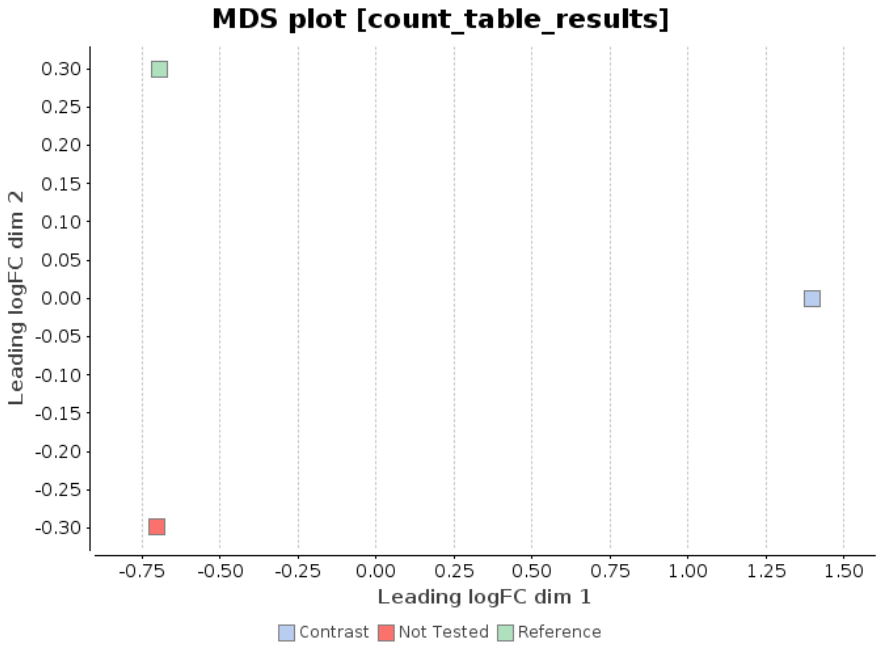


Figure 5(a): MDS Plot

9.8.4.2 Results Chart

Bar chart which shows the number of total features, kept features (those who have passed the filtering step), differentially expressed features, up-regulated features and down-regulated features (Figure 5(b)[129]).

[128] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=620232718#PairwiseDifferentialExpressionAnalysis(WithoutReplicates)-figure5a

[129] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=620232718#PairwiseDifferentialExpressionAnalysis(WithoutReplicates)-figure5b

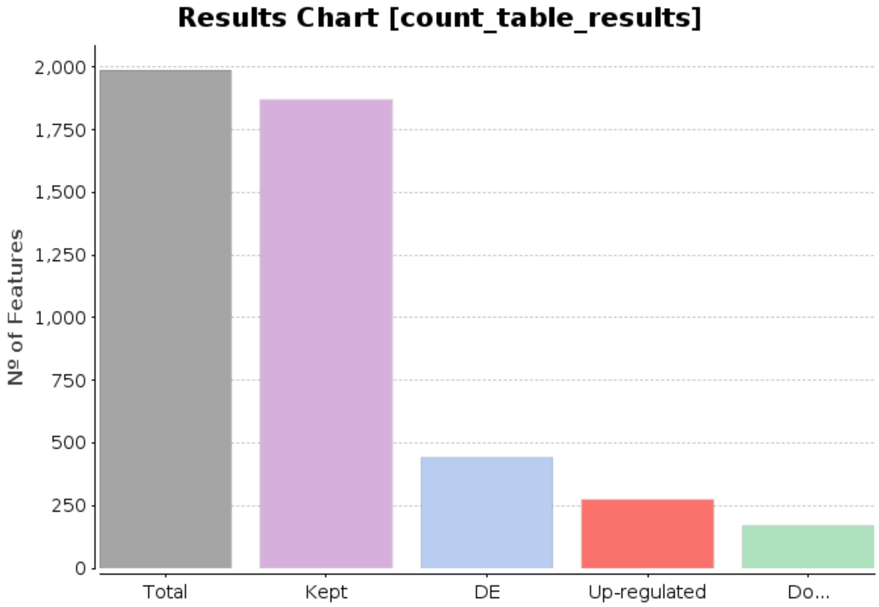


Figure 5(b): Results Chart

9.8.4.3 Expression Plot

A scatter plot showing the average expression values of each condition (Figure 5(c)[130]). Differentially expressed features considering the probability threshold (0.9 by default) will be highlighted in red.

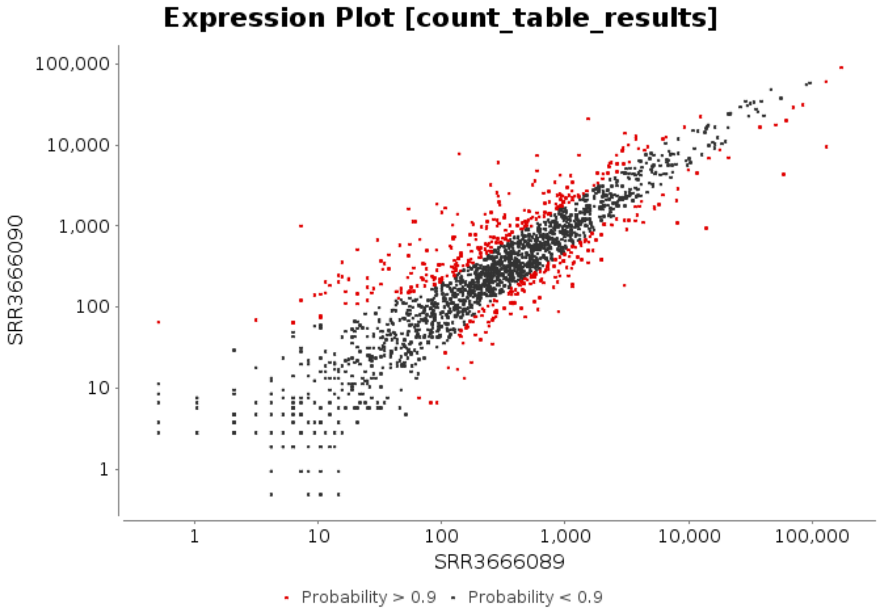


Figure 5(c): Expression Plot

[130] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=620232718#PairwiseDifferentialExpressionAnalysis(WithoutReplicates)-figure5

9.8.4.4 MD Plot

A scatter plot showing the log-fold change (M) and the absolute value of the difference in expression between conditions (D). D values are displayed in a log-scale (Figure 5 (d)[131]).

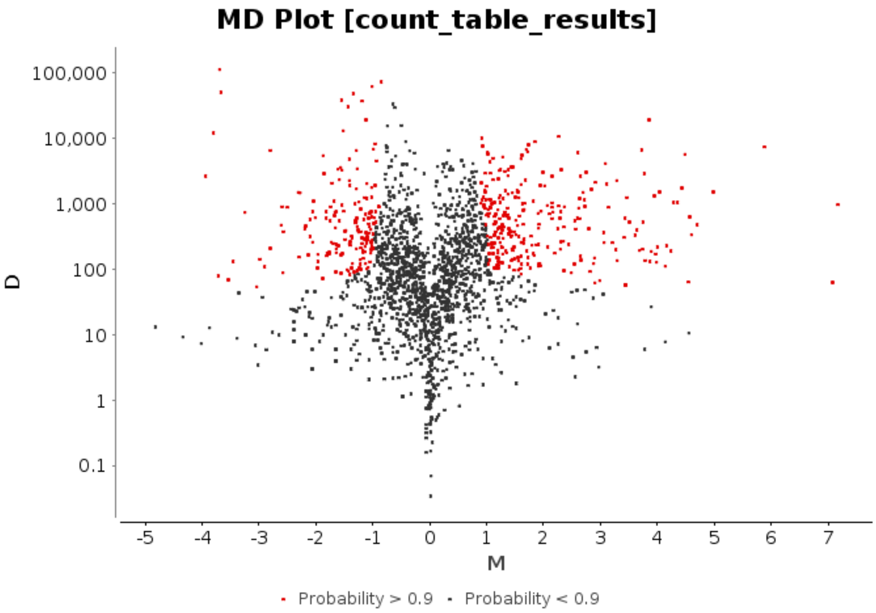


Figure 5(d): MD Plot

9.8.4.5 Heatmap

A heatmap is a two-dimensional visual representation of data in which numerical values of points are represented by a range of colors (Figure 5 (e)[132]). The dendrograms added to the left and top side are produced by a hierarchical clustering method that takes as input the Euclidean distance computed between genes (left) and samples (top).

The heatmap supports zooming by keeping clicked a node of either of the two dendrograms. The first bars contain the experimental design of the data showing the association between samples and experimental covariates.

Genes that will be displayed can be selected in the wizard. There are three options:

- The Top 50 differentially expressed genes (ranked by FDR).
- All differentially expressed genes.
- Provide an ID list containing the genes to represent.

[13] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=620232718#PairwiseDifferentialExpressionAnalysis(WithoutReplicates)-figure5d

[23] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=620232718#PairwiseDifferentialExpressionAnalysis(WithoutReplicates)-figure5e

⚠ Differentially expressed genes are those that are labeled as UP or DOWN in the table project ("Tags" column). The criteria for considering a gene as differentially expressed can be adjusted using the option "Set Up/Down Tags".

Furthermore, the wizard allows to adjust the type of expression data that will be represented, as well as the transformation that can be applied to this data.

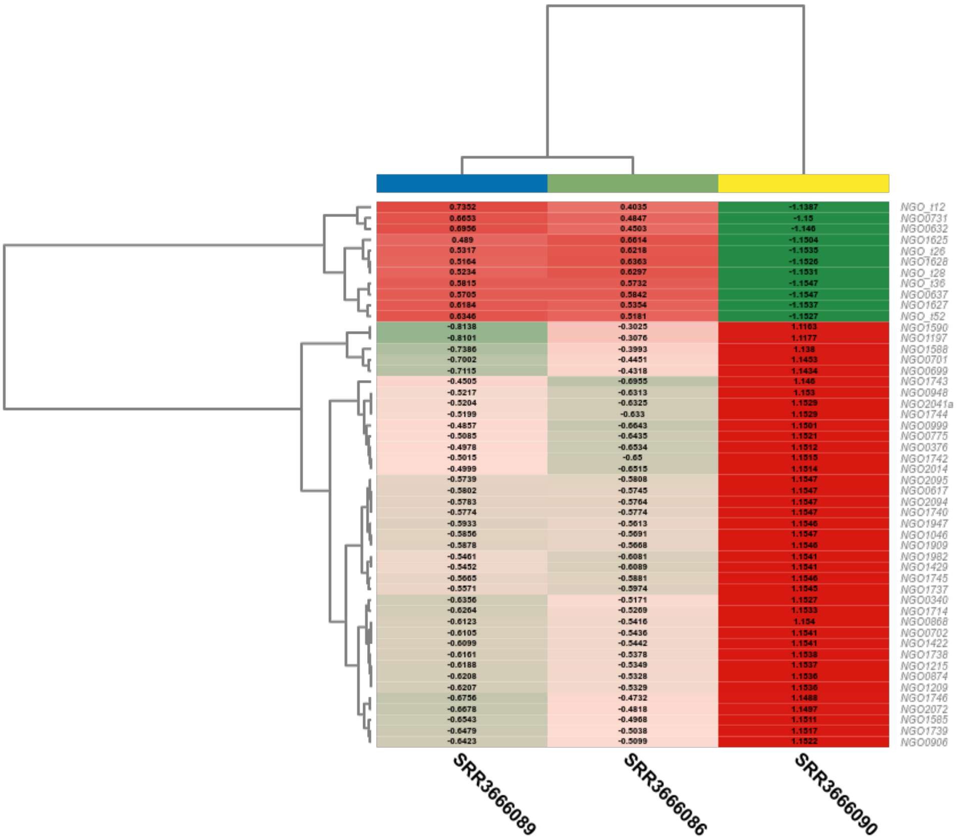


Figure 5(e): Heatmap

9.8.5 Enrichment Analysis

It is possible to perform a functional enrichment analysis from the pairwise differential expression project. Both options, Fisher's Exact Test and Gene Set Enrichment Analysis, can be found under the **Side Panel** of the Pairwise Results viewer. For a detailed tutorial on how to launch an Enrichment Analysis from Pairwise Differential Expression results, link here[133].

[133] https://www.biobam.com/how-to-launch-an-enrichment-analysis-from-pairwise-differential-expression-results-in-blast2go/

9.8.5.1 Fisher's Exact Test

Choose the subset of genes that will be considered as Test-set. Up-regulated and down-regulated genes are those that are tagged according to the criteria established by the option "Set Up/Down Tags".

The project containing the functionally annotated sequences that will be used as a reference background set should be provided.

The rest of the parameters are explained in the Fisher's Exact Test section(see page 97).

9.8.5.2 Gene Set Enrichment Analysis

The "Probability Threshold for Ranked List" parameters allows setting a filter to exclude those genes whose probability value is not above it. The ranked gene list will be created using the "ranking" statistic.

The project containing the functionally annotated sequences that will be used as a reference background set should be provided.

The rest of the parameters are explained in the Gene Set Enrichment Analysis section(see page 101).

9.9 Time Course Expression Analysis

Content of this page:

9.9.1 Introduction

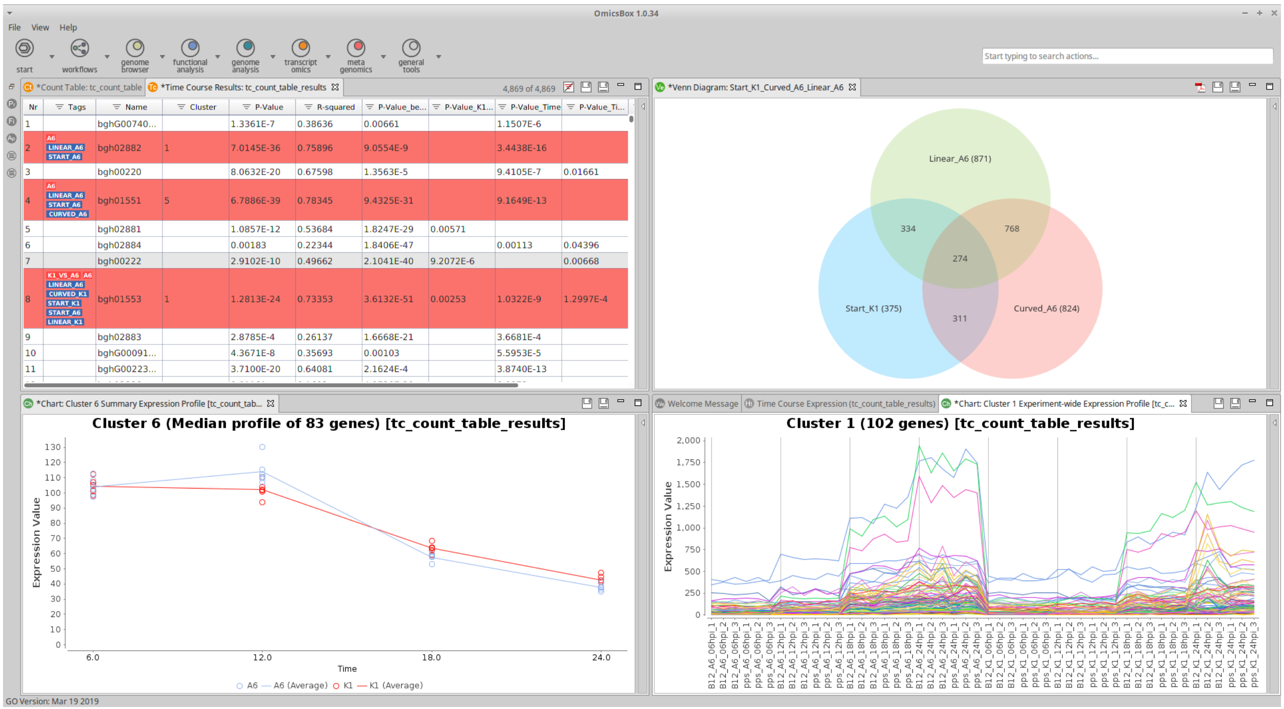
This tool is designed to perform time-course expression analysis of count data arising from RNA-seq technology. Based on the maSigPro program, this application allows the detection of genomic features (e.g. genes) with significant temporal expression changes and significant differences between experimental groups. The software package maSigPro[134], which belongs to the Bioconductor project, implements a two steps regression strategy to find genes for which there are significant expression profile differences in time course RNA-seq experiments.

[134] https://www.bioconductor.org/packages/release/bioc/html/maSigPro.html

Please cite maSigPro as:
Conesa A, Nueda MJ (2018). "maSigPro: Significant Gene Expression Profile Differences in Time Course Gene Expression Data." R package version 1.52.0, http://bioinfo.cipf.es/.

9.9.2 General Workflow

The workflow to be followed to perform a time course expression analysis is described in Figure 2[135].


OmicsBox 1.0.34
File View Help
start
workflows
genome browser
functional analysis
genome analysis
transcript omics
meta genomics
general tools
Start typing to search actions...
*Count Table: tc_count_table
*Time Course Results: tc_count_table_results
4,869 of 4,869
*Venn Diagram: Start_K1_Curved_A6_Linear_A6
Linear_A6 (871)
334
768
274
Start_K1 (375)
311
Curved_A6 (824)
*Chart: Cluster 6 Summary Expression Profile [tc_count_tab...
Cluster 6 (Median profile of 83 genes) [tc_count_table_results]
Expression Value
Time
A6 — A6 (Average) K1 — K1 (Average)
Welcome Message
Time Course Expression (tc_count_table_results)
*Chart: Cluster 1 Experiment-wide Expression Profile [tc_c...
Cluster 1 (102 genes) [tc_count_table_results]
Expression Value
GO Version: Mar 19 2019

Nr	Tags	Name	Cluster	P-Value	R-squared	P-Value_be...	P-Value_K1...	P-Value_Time	P-Value_Ti...
1		bghG00740...		1.3361E-7	0.38636	0.00661		1.1507E-6	
2	A6 LINEAR_A6 START_A6	bgh02882	1	7.0145E-36	0.75896	9.0554E-9		3.4438E-16	
3		bgh00220		8.0632E-20	0.67598	1.3563E-5		9.4105E-7	0.01661
4	A6 LINEAR_A6 START_A6 CURVED_A6	bgh01551	5	6.7886E-39	0.78345	9.4325E-31		9.1649E-13	
5		bgh02881		1.0857E-12	0.53684	1.8247E-29	0.00571		
6		bgh02884		0.00183	0.22344	1.8406E-47		0.00113	0.04396
7		bgh00222		2.9102E-10	0.49662	2.1041E-40	9.2072E-6		0.00668
8	K1_VS_A6 A6 LINEAR_A6 CURVED_K1 START_K1 START_A6 LINEAR_K1	bgh01553	1	1.2813E-24	0.73353	3.6132E-51	0.00253	1.0322E-9	1.2997E-4
9		bgh02883		2.8785E-4	0.26137	1.6668E-21		3.6681E-4	
10		bghG00091...		4.3671E-8	0.35693	0.00103		5.5953E-5	
11		bghG00223...		3.7100E-20	0.64081	2.1624E-4		3.8740E-13	

Figure 1: Differential Expression Analysis Interface

[135] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598016084#TimeCourseExpressionAnalysis-figure2

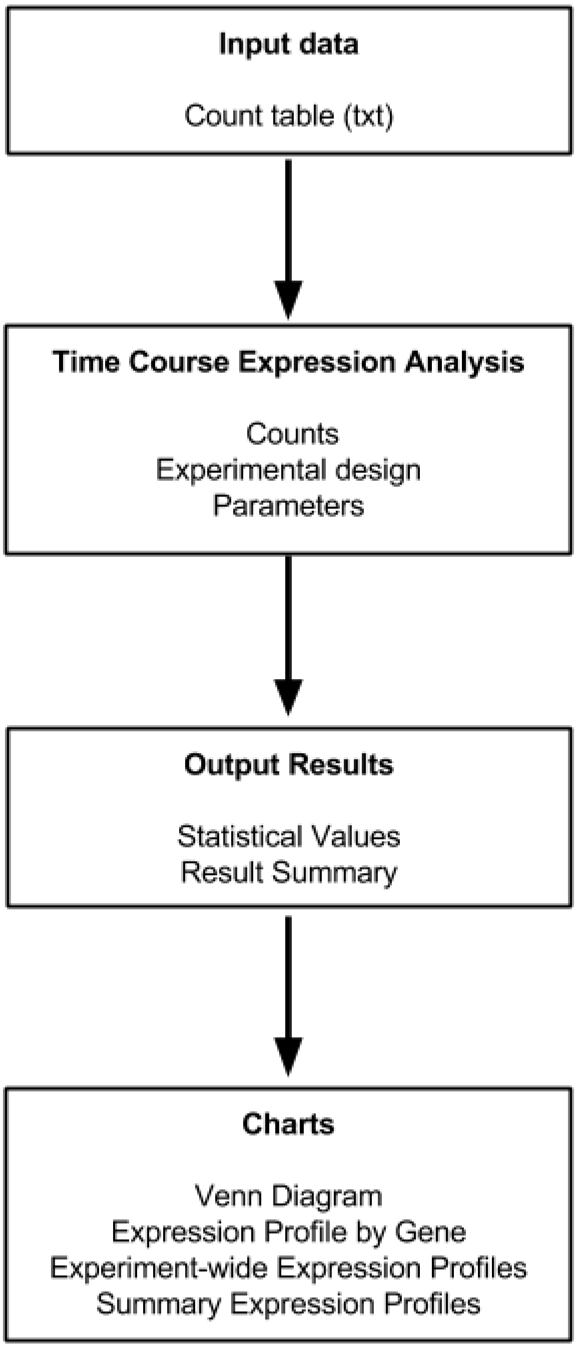
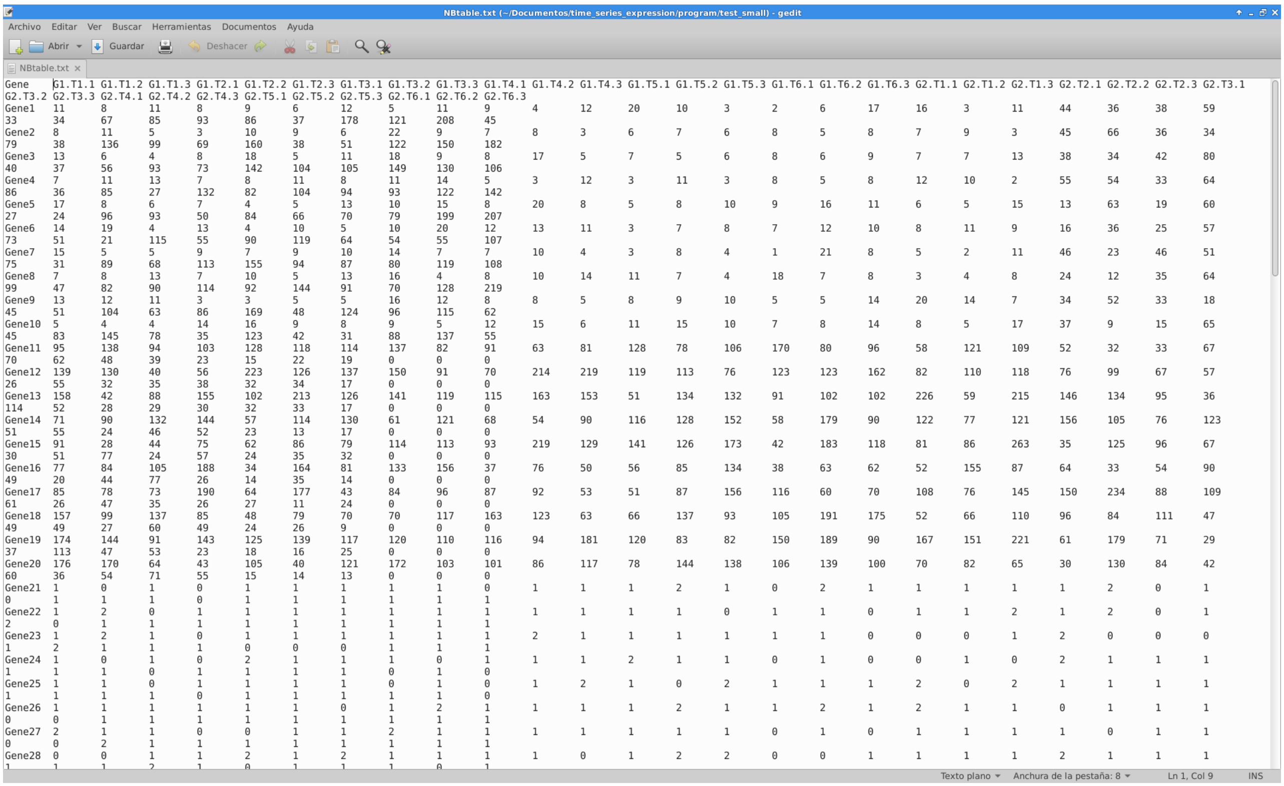


Figure 2: General Workflow

9.9.3 Load Data

Go to **File** → **Load** → **Load Count Table** and select your .txt file containing the count table in tab-delimited format (Figure 3[136]). It is also possible to create a Count Table within OmicsBox through the "Create Count Table" functionality (see Quantify Expression(see page 172) section).



Gene	G1.T1.1	G1.T1.2	G1.T1.3	G1.T2.1	G1.T2.2	G1.T2.3	G1.T3.1	G1.T3.2	G1.T3.3	G1.T4.1	G1.T4.2	G1.T4.3	G1.T5.1	G1.T5.2	G1.T5.3	G1.T6.1	G1.T6.2	G1.T6.3	G2.T1.1	G2.T1.2	G2.T1.3	G2.T2.1	G2.T2.2	G2.T2.3	G2.T3.1	G2.T3.2	G2.T3.3	G2.T4.1	G2.T4.2	G2.T4.3	G2.T5.1	G2.T5.2	G2.T5.3	G2.T6.1	G2.T6.2	G2.T6.3
Gene1	11	8	11	8	9	6	12	5	11	9	4	12	20	10	3	2	6	17	16	3	11	44	36	38	59	33	34	67	85	93	86	37	178	121	208	45
Gene2	8	11	5	3	10	9	6	22	9	7	8	3	6	7	6	8	5	8	7	9	3	45	66	36	34	79	38	136	99	69	160	38	51	122	150	182
Gene3	13	6	4	8	18	5	11	18	9	8	17	5	7	5	6	8	6	9	7	7	13	38	34	42	80	40	37	56	93	73	142	104	105	149	130	106
Gene4	7	11	13	7	8	11	8	11	14	5	3	12	3	11	3	8	5	8	12	10	2	55	54	33	64	86	36	85	27	132	82	104	94	93	122	142
Gene5	17	8	6	7	4	5	13	10	15	8	20	8	5	8	10	9	16	11	6	5	15	13	63	19	60	27	24	96	93	50	84	66	70	79	199	207
Gene6	14	19	4	13	4	10	5	10	20	12	13	11	3	7	8	7	12	10	8	11	9	16	36	25	57	73	51	21	115	55	90	119	64	54	55	107
Gene7	15	5	5	9	7	9	10	14	7	7	10	4	3	8	4	1	21	8	5	2	11	46	23	46	51	75	31	89	68	113	155	94	87	80	119	108
Gene8	7	8	13	7	10	5	13	16	4	8	10	14	11	7	4	18	7	8	3	4	8	24	12	35	64	99	47	82	90	114	92	144	91	70	128	219
Gene9	13	12	11	3	3	5	5	16	12	8	8	5	8	9	10	5	5	14	20	14	7	34	52	33	18	45	51	104	63	86	169	48	124	96	115	62
Gene10	5	4	4	14	16	9	8	9	5	12	15	6	11	15	10	7	8	14	8	5	17	37	9	15	65	45	83	145	78	35	123	42	31	88	137	55
Gene11	95	138	94	103	128	118	114	137	82	91	63	81	128	78	106	170	80	96	58	121	109	52	32	33	67	70	62	48	39	23	15	22	19	0	0	0
Gene12	139	130	40	56	223	126	137	150	91	70	214	219	119	113	76	123	123	162	82	110	118	76	99	67	57	26	55	32	35	38	32	34	17	0	0	0
Gene13	158	42	88	155	102	213	126	141	119	115	163	153	51	134	132	91	102	102	226	59	215	146	134	95	36	114	52	28	29	30	32	33	17	0	0	0
Gene14	71	90	132	144	57	114	130	61	121	68	54	90	116	128	152	58	179	90	122	77	121	156	105	76	123	51	55	24	46	52	23	13	17	0	0	0
Gene15	91	28	44	75	62	86	79	114	113	93	219	129	141	126	173	42	183	118	81	86	263	35	125	96	67	30	51	77	24	57	24	35	32	0	0	0
Gene16	77	84	105	188	34	164	81	133	156	37	76	50	56	85	134	38	63	62	52	155	87	64	33	54	90	49	20	44	77	26	14	35	14	0	0	0
Gene17	85	78	73	190	64	177	43	84	96	87	92	53	51	87	156	116	60	70	108	76	145	150	234	88	109	61	26	47	35	26	27	11	24	0	0	0
Gene18	157	99	137	85	48	79	70	70	117	163	123	63	66	137	93	105	191	175	52	66	110	96	84	111	47	49	49	27	60	49	24	26	9	0	0	0
Gene19	174	144	91	143	125	139	117	120	110	116	94	181	120	83	82	150	189	90	167	151	221	61	179	71	29	37	113	47	53	23	18	16	25	0	0	0
Gene20	176	170	64	43	105	40	121	172	103	101	86	117	78	144	138	106	139	100	70	82	65	30	130	84	42	60	36	54	71	55	15	14	13	0	0	0
Gene21	1	0	1	0	1	1	1	1	1	0	1	1	1	2	1	0	2	1	1	1	1	1	2	0	1	0	1	1	1	0	1	1	1	1	1	1
Gene22	1	2	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	2	1	2	0	1	2	0	1	1	1	1	1	1	1	1	1
Gene23	1	2	1	0	1	1	1	1	1	1	2	1	1	1	1	1	1	0	0	0	1	2	0	0	0	1	2	1	1	1	0	0	0	1	1	1
Gene24	1	0	1	0	2	1	1	1	0	1	1	1	2	1	1	0	1	0	0	1	0	2	1	1	1	1	1	1	0	1	1	1	1	0	1	0
Gene25	1	1	0	1	1	1	1	0	1	0	1	2	1	0	2	1	1	1	2	0	2	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0
Gene26	1	1	1	1	1	1	0	1	2	1	1	1	1	2	1	1	2	1	2	1	1	0	1	1	1	0	0	1	1	1	1	1	1	1	1	1
Gene27	2	1	1	0	0	1	1	2	1	1	1	1	1	1	1	0	1	0	1	1	1	1	0	1	1	0	0	2	1	1	1	1	1	1	1	1
Gene28	0	0	1	1	2	1	2	1	1	1	1	0	1	2	2	0	0	1	1	1	1	2	1	1	1	1	1	1	2	1	0	1	1	1	0	1

Figure 3: Count Table File

The Count Table can be saved as 'CountTable' object (**File** → **Save**).

⚠ **Notes:**

- This application only accepts raw counts without any type of normalization.
- Replicates for each experimental condition are required.

9.9.4 Run Analysis

Go to **Transcriptomics** → **Run Differential Expression Analysis and** choose the ``Time Course Expression Analysis'' option. Here you can specify the following parameters, which are divided into three

[136] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598016084#TimeCourseExpressionAnalysis-figure3

different sections: Preprocessing Data (Figure 4[137]), Experimental Design (Figure 5[138]) and Analysis Options (Figure 6[139]).

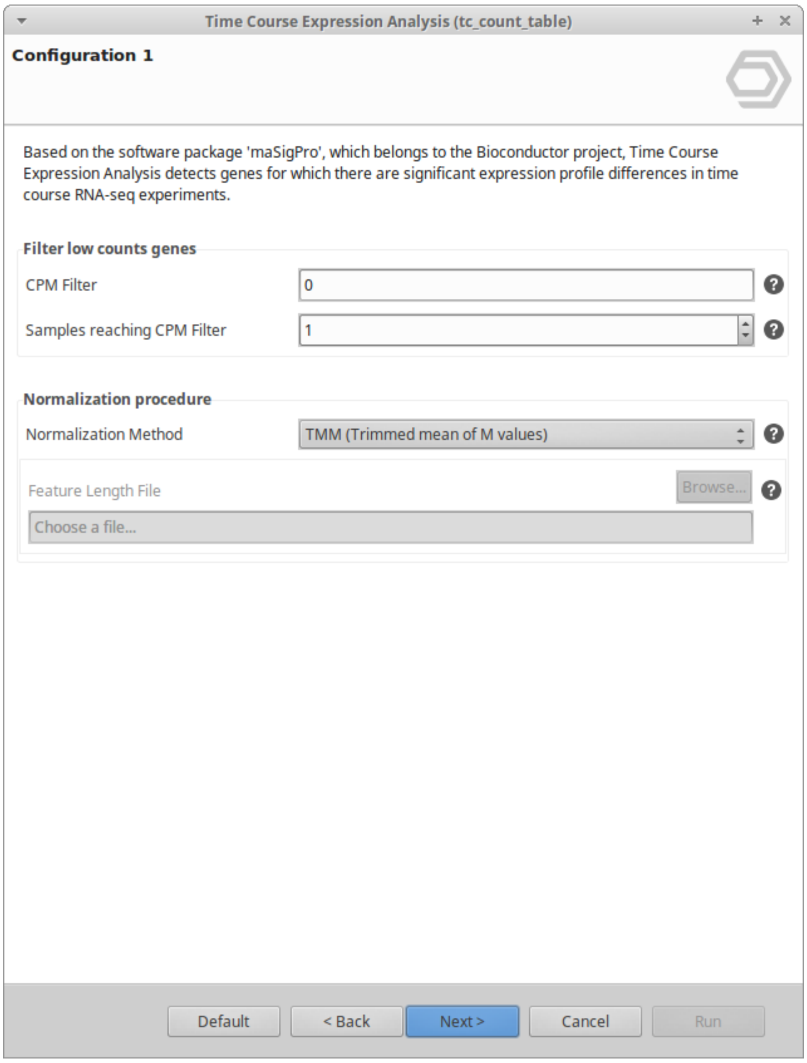
9.9.4.1 Preprocessing Data Page

- **Filter low count genes:**
  - **CPM Filter:** Establish a filter to exclude genes with low counts across libraries, as those genes may interfere with the subsequent statistical approximations. Filtering is performed on a count-per-million (CPM) basis to account for differences in library size between samples (e.g. a CPM of 1 corresponds to a count of 6 in a sample with 6 million reads).
  - **Samples reaching CPM Filter:** Set a minimum number of samples in which the gene's CPM is above the filter level (is expressed). If this value is set to e.g. five, at least 5 of the samples have to be above the given CPM. The number of samples of the smallest group is usually taken (e.g. in an experiment that has two replicates for each condition (or group), a gene should be expressed in at least two samples). Set value to 0 if no filter is desired.
- **Normalization procedure:**
  - **Normalization Method**: Normalization is an important step to make the samples comparable and to remove possible biases (as sequencing depth bias) in count data. You can select the normalization method to be used:
    - **TMM**: Weighted trimmed mean of M-values. In this method, weights are obtained from the delta method on Binomial Data (this method is recommended).
    - **RPKM**: Reads Per Kilobase per Million mapped reads. This method corrects for gene length and the number of sequencing reads (gene length is required).
    - **Upper-quartile**: 75% quantile for the counts for each library is used to calculate the scale factors for normalization.
    - **None**: No normalization method is applied.
  - **Feature Length File**: For RPKM normalization load a tab-delimited file (or ID-Value object) with two columns containing the name and length of each gene or genomic feature.

[137] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598016084#TimeCourseExpressionAnalysis-figure4

[138] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598016084#TimeCourseExpressionAnalysis-figure6

[139] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598016084#TimeCourseExpressionAnalysis-figure6



Based on the software package 'maSigPro', which belongs to the Bioconductor project, Time Course Expression Analysis detects genes for which there are significant expression profile differences in time course RNA-seq experiments.

Filter low counts genes

CPM Filter: 0

Samples reaching CPM Filter: 1

Normalization procedure

Normalization Method: TMM (Trimmed mean of M values)

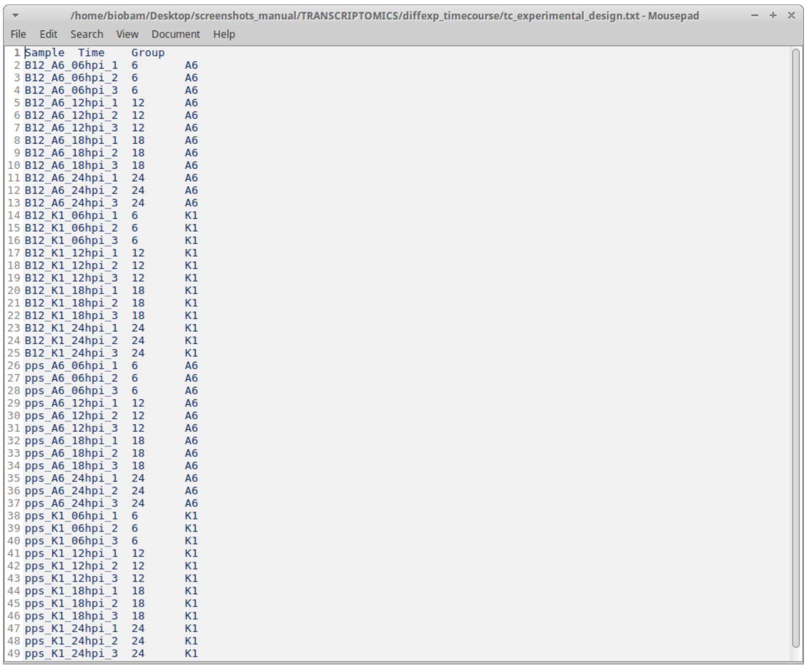
Feature Length File: Choose a file...

Figure 4: Preprocessing Data Page

9.9.4.2 Experimental Design Page

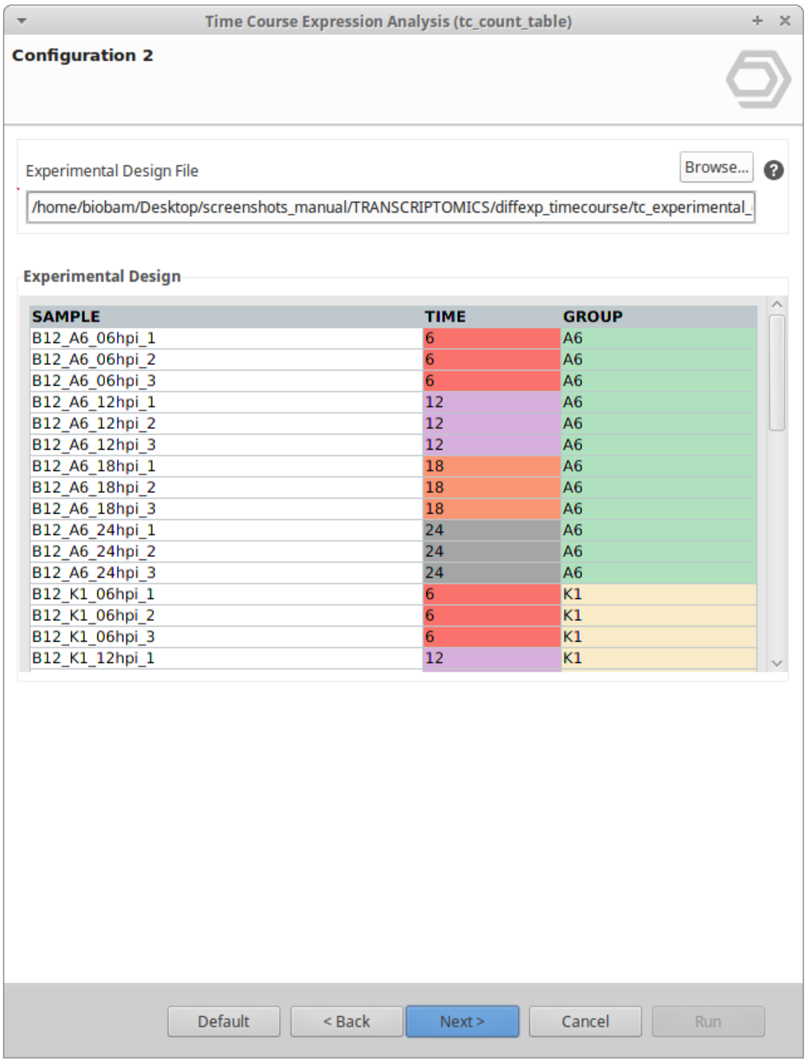
Experimental design file: Select your .txt file containing your experiment descriptors associated with each sample in tab-delimited format. As demonstrated in Figure 7[140], rows correspond to samples and columns to experimental descriptors. A column must contain the associated time points for each sample, and another column should show the assignment of samples to experimental groups. Make sure that the names in the first column of the experimental design table are exactly the same as the sample names in the count table header. If your experimental design file has fewer samples than the count table, only the samples contained in this file will be analyzed.

[140] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598016084#TimeCourseExpressionAnalysis-figure7


/home/biobam/Desktop/screenshots_manual/TRANSCRIPTOMICS/diffexp_timecourse/tc_experimental_design.txt - Mousepad
File
Edit
Search
View
Document
Help

```
Sample  Time    Group
B12_A6_06hpi_1  6       A6
B12_A6_06hpi_2  6       A6
B12_A6_06hpi_3  6       A6
B12_A6_12hpi_1  12      A6
B12_A6_12hpi_2  12      A6
B12_A6_12hpi_3  12      A6
B12_A6_18hpi_1  18      A6
B12_A6_18hpi_2  18      A6
B12_A6_18hpi_3  18      A6
B12_A6_24hpi_1  24      A6
B12_A6_24hpi_2  24      A6
B12_A6_24hpi_3  24      A6
B12_K1_06hpi_1  6       K1
B12_K1_06hpi_2  6       K1
B12_K1_06hpi_3  6       K1
B12_K1_12hpi_1  12      K1
B12_K1_12hpi_2  12      K1
B12_K1_12hpi_3  12      K1
B12_K1_18hpi_1  18      K1
B12_K1_18hpi_2  18      K1
B12_K1_18hpi_3  18      K1
B12_K1_24hpi_1  24      K1
B12_K1_24hpi_2  24      K1
B12_K1_24hpi_3  24      K1
pps_A6_06hpi_1  6       A6
pps_A6_06hpi_2  6       A6
pps_A6_06hpi_3  6       A6
pps_A6_12hpi_1  12      A6
pps_A6_12hpi_2  12      A6
pps_A6_12hpi_3  12      A6
pps_A6_18hpi_1  18      A6
pps_A6_18hpi_2  18      A6
pps_A6_18hpi_3  18      A6
pps_A6_24hpi_1  24      A6
pps_A6_24hpi_2  24      A6
pps_A6_24hpi_3  24      A6
pps_K1_06hpi_1  6       K1
pps_K1_06hpi_2  6       K1
pps_K1_06hpi_3  6       K1
pps_K1_12hpi_1  12      K1
pps_K1_12hpi_2  12      K1
pps_K1_12hpi_3  12      K1
pps_K1_18hpi_1  18      K1
pps_K1_18hpi_2  18      K1
pps_K1_18hpi_3  18      K1
pps_K1_24hpi_1  24      K1
pps_K1_24hpi_2  24      K1
pps_K1_24hpi_3  24      K1
```

Figure 7: Experimental Design File



Experimental Design File

/home/biobam/Desktop/screenshots_manual/TRANSCRIPTOMICS/diffexp_timecourse/tc_experimental_

Experimental Design

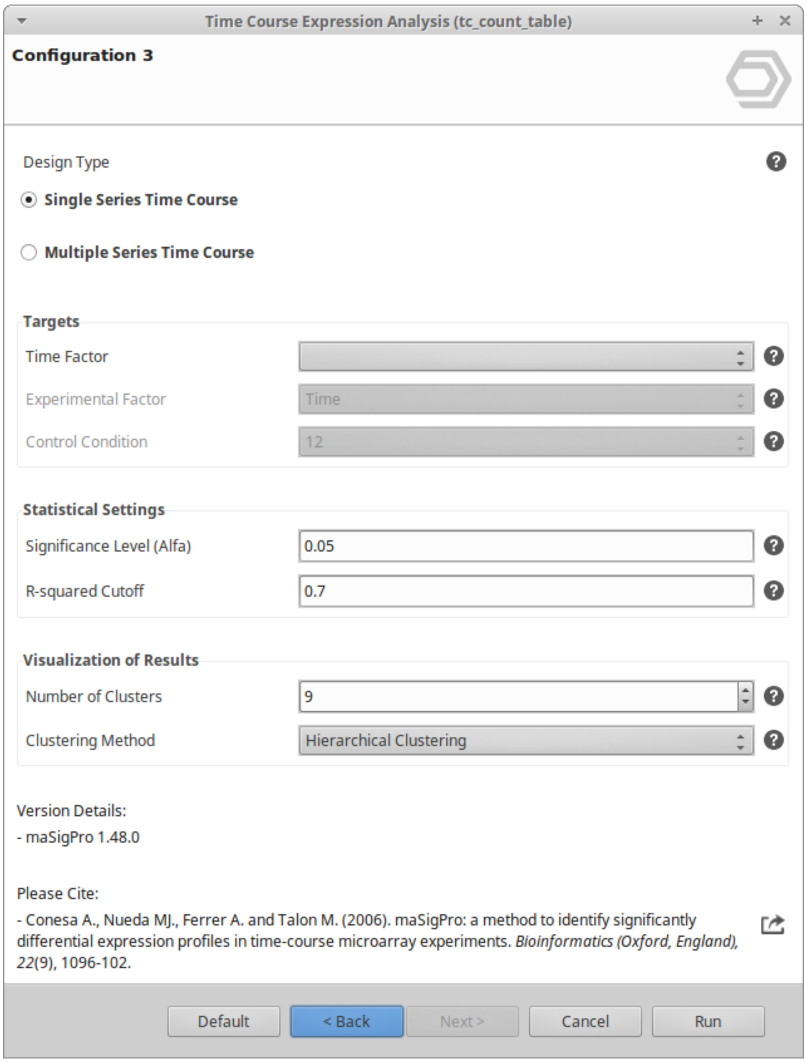
SAMPLE	TIME	GROUP
B12_A6_06hpi_1	6	A6
B12_A6_06hpi_2	6	A6
B12_A6_06hpi_3	6	A6
B12_A6_12hpi_1	12	A6
B12_A6_12hpi_2	12	A6
B12_A6_12hpi_3	12	A6
B12_A6_18hpi_1	18	A6
B12_A6_18hpi_2	18	A6
B12_A6_18hpi_3	18	A6
B12_A6_24hpi_1	24	A6
B12_A6_24hpi_2	24	A6
B12_A6_24hpi_3	24	A6
B12_K1_06hpi_1	6	K1
B12_K1_06hpi_2	6	K1
B12_K1_06hpi_3	6	K1
B12_K1_12hpi_1	12	K1

Figure 5: Experimental Design Page

9.9.4.3 Analysis Options

- **Design Type:** Choose the design type to adjust the analysis.
  - Single Series Time Course: Detects genes that show significant expression changes over time. You only have to select the time factor of your experimental design in ``Targets''.
  - Multiple Series Time Course: Find genes with significant temporal expression changes and significant differences between experimental groups. You have to establish the time and experimental factors, and select the control condition of your experimental design in ``Targets''.
- **Statistical Settings:**
  - Significance Level (Alfa): The level of FDR control used for variable selection in the stepwise regression.
  - R-squared Cutoff: Cutoff value for the R-squared of the regression model.
- **Visualization of Results:**
  - Number of Clusters: Establish a number of clusters to group genes by similar expression profiles.

- Clustering Method: Choose a clustering method for data partitioning.
    - Hierarchical Clustering: Performs a hierarchical cluster analysis using a set of dissimilarities for the features being clustered.
    - K-Means Clustering: Is intended to divide the points into K clusters such that the sum of squares of the points to the centers of the clusters assigned is minimized.
    - Model-Based Clustering: The optimal model according to BIC for EM initialized by hierarchical clustering for Gaussian mixture models. This method computes an optimal number of clusters. Keep in mind that this method requires more time.



Time Course Expression Analysis (tc_count_table)

Configuration 3

Design Type

● **Single Series Time Course**

○ **Multiple Series Time Course**

Targets

Time Factor:

Experimental Factor: Time

Control Condition: 12

Statistical Settings

Significance Level (Alfa): 0.05

R-squared Cutoff: 0.7

Visualization of Results

Number of Clusters: 9

Clustering Method: Hierarchical Clustering

Version Details:
- maSigPro 1.48.0

Please Cite:
- Conesa A., Nueda MJ., Ferrer A. and Talon M. (2006). maSigPro: a method to identify significantly differential expression profiles in time-course microarray experiments. *Bioinformatics (Oxford, England)*, *22*(9), 1096-102.

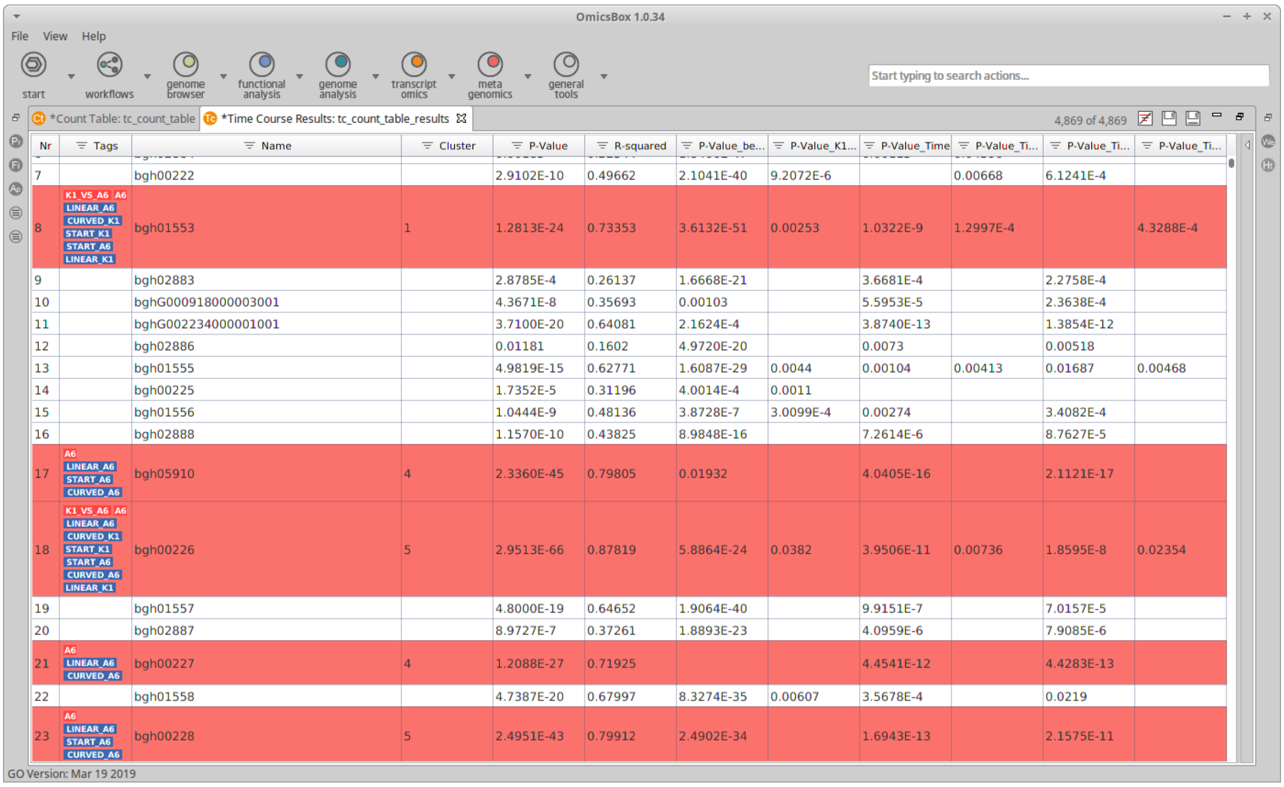
Figure 6: Analysis Options

9.9.5 Results

Once the input counts have been processed and analyzed via the ``Time Course Expression Analysis'' tool, a new tab is opened containing statistical results obtaining by the stepwise regression statistical test (Figure 8[141]):

- P-value of the regression ANOVA.
- R-squared of the model.
- P-value of the regression coefficients of the selected variables.
- Tags: Indicate the list/s of significant genes in which the feature appears (R-squared ≥ R-squared Cutoff).
    - Red tags: Lists of significant genes for each experimental group (only available in ``Multiple Series Time Course'').
    - Blue tags: List of significant genes for each variable of the regression model.

Only the genes that have passed the established Significance Level are shown in the new tab. For further details please refer to the maSigPro User's Guide[142].



Nr	Tags	Name	Cluster	P-Value	R-squared	P-Value_be...	P-Value_K1...	P-Value_Time	P-Value_Ti...	P-Value_Ti...	P-Value_Ti...
7		bgh00222		2.9102E-10	0.49662	2.1041E-40	9.2072E-6		0.00668	6.1241E-4	
8	K1_VS_A6 A6 LINEAR_A6 CURVED_K1 START_K1 START_A6 LINEAR_K1	bgh01553	1	1.2813E-24	0.73353	3.6132E-51	0.00253	1.0322E-9	1.2997E-4		4.3288E-4
9		bgh02883		2.8785E-4	0.26137	1.6668E-21		3.6681E-4		2.2758E-4	
10		bghG000918000003001		4.3671E-8	0.35693	0.00103		5.5953E-5		2.3638E-4	
11		bghG002234000001001		3.7100E-20	0.64081	2.1624E-4		3.8740E-13		1.3854E-12	
12		bgh02886		0.01181	0.1602	4.9720E-20		0.0073		0.00518	
13		bgh01555		4.9819E-15	0.62771	1.6087E-29	0.0044	0.00104	0.00413	0.01687	0.00468
14		bgh00225		1.7352E-5	0.31196	4.0014E-4	0.0011				
15		bgh01556		1.0444E-9	0.48136	3.8728E-7	3.0099E-4	0.00274		3.4082E-4	
16		bgh02888		1.1570E-10	0.43825	8.9848E-16		7.2614E-6		8.7627E-5	
17	A6 LINEAR_A6 START_A6 CURVED_A6	bgh05910	4	2.3360E-45	0.79805	0.01932		4.0405E-16		2.1121E-17	
18	K1_VS_A6 A6 LINEAR_A6 CURVED_K1 START_K1 START_A6 CURVED_A6 LINEAR_K1	bgh00226	5	2.9513E-66	0.87819	5.8864E-24	0.0382	3.9506E-11	0.00736	1.8595E-8	0.02354
19		bgh01557		4.8000E-19	0.64652	1.9064E-40		9.9151E-7		7.0157E-5	
20		bgh02887		8.9727E-7	0.37261	1.8893E-23		4.0959E-6		7.9085E-6	
21	A6 LINEAR_A6 CURVED_A6	bgh00227	4	1.2088E-27	0.71925			4.4541E-12		4.4283E-13	
22		bgh01558		4.7387E-20	0.67997	8.3274E-35	0.00607	3.5678E-4		0.0219	
23	A6 LINEAR_A6 START_A6 CURVED_A6	bgh00228	5	2.4951E-43	0.79912	2.4902E-34		1.6943E-13		2.1575E-11	

Figure 8: Table Viewer

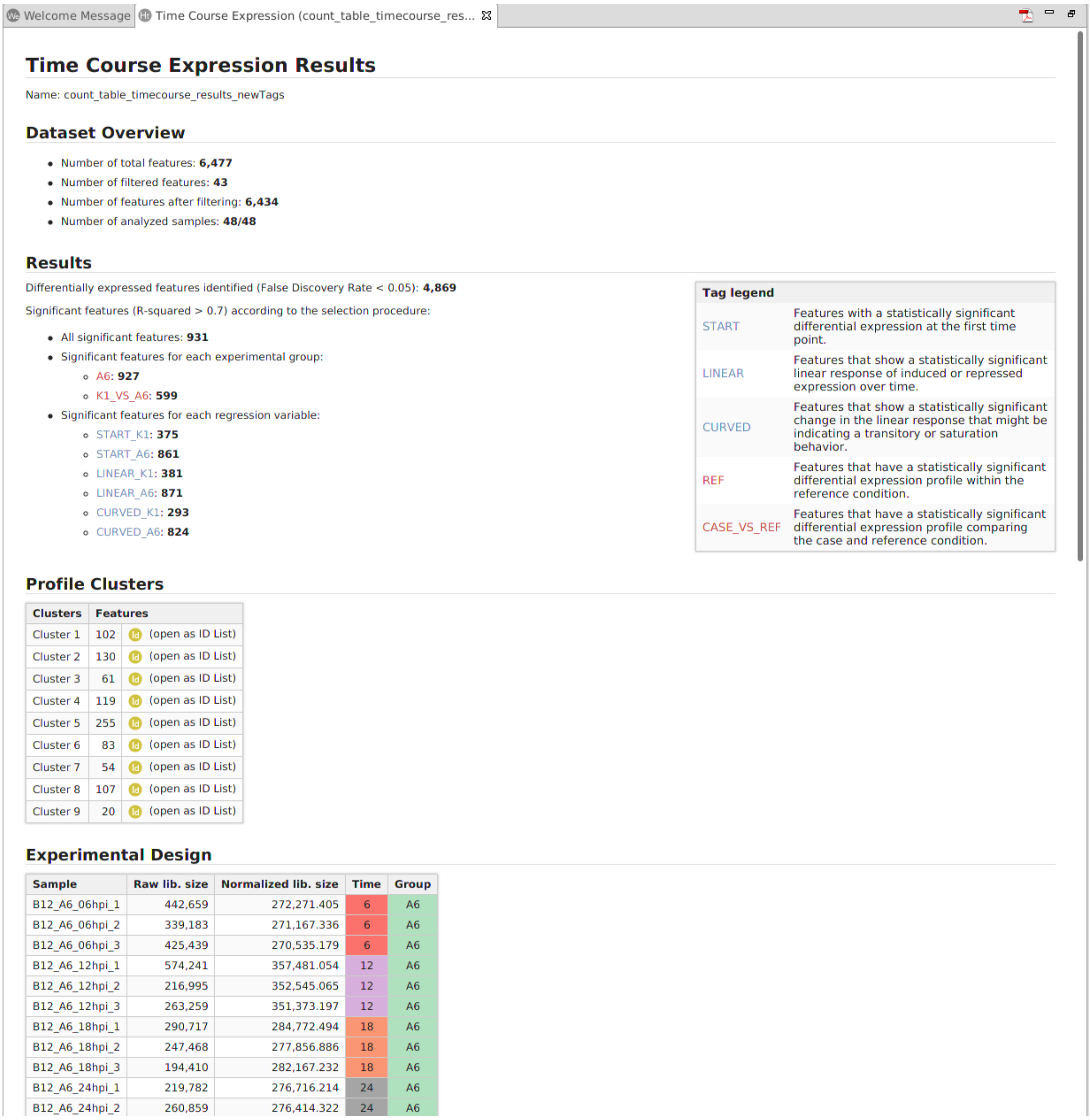
[141] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598016084#TimeCourseExpressionAnalysis-figure8

[142] https://www.bioconductor.org/packages/release/bioc/vignettes/maSigPro/inst/doc/maSigProUsersGuide.pdf

Results can be saved as a TC Results object. Note that is not possible to perform the analysis on this object. For this purpose, you have to open the Count Table object. If you want to see both count table and results, go to the File Manager and open the two .b2g files together.

A result page will show a summary of the time course expression analysis results, including the cluster of features with similar expression profiles (Figure 9[143]). Go to **Side Panel** → **Result Summary** in order to visualize the result summary and to export it in pdf.

[143] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598016084#TimeCourseExpressionAnalysis-figure9



Time Course Expression Results

Name: count_table_timecourse_results_newTags

Dataset Overview

- Number of total features: **6,477**
- Number of filtered features: **43**
- Number of features after filtering: **6,434**
- Number of analyzed samples: **48/48**

Results

Differentially expressed features identified (False Discovery Rate < 0.05): **4,869**

Significant features (R-squared > 0.7) according to the selection procedure:

- All significant features: **931**
- Significant features for each experimental group:
  - A6: **927**
  - K1_VS_A6: **599**
- Significant features for each regression variable:
  - START_K1: **375**
  - START_A6: **861**
  - LINEAR_K1: **381**
  - LINEAR_A6: **871**
  - CURVED_K1: **293**
  - CURVED_A6: **824**

Tag legend	
START	Features with a statistically significant differential expression at the first time point.
LINEAR	Features that show a statistically significant linear response of induced or repressed expression over time.
CURVED	Features that show a statistically significant change in the linear response that might be indicating a transitory or saturation behavior.
REF	Features that have a statistically significant differential expression profile within the reference condition.
CASE_VS_REF	Features that have a statistically significant differential expression profile comparing the case and reference condition.

Profile Clusters

Clusters	Features	
Cluster 1	102	(open as ID List)
Cluster 2	130	(open as ID List)
Cluster 3	61	(open as ID List)
Cluster 4	119	(open as ID List)
Cluster 5	255	(open as ID List)
Cluster 6	83	(open as ID List)
Cluster 7	54	(open as ID List)
Cluster 8	107	(open as ID List)
Cluster 9	20	(open as ID List)

Experimental Design

Sample	Raw lib. size	Normalized lib. size	Time	Group
B12_A6_06hpi_1	442,659	272,271.405	6	A6
B12_A6_06hpi_2	339,183	271,167.336	6	A6
B12_A6_06hpi_3	425,439	270,535.179	6	A6
B12_A6_12hpi_1	574,241	357,481.054	12	A6
B12_A6_12hpi_2	216,995	352,545.065	12	A6
B12_A6_12hpi_3	263,259	351,373.197	12	A6
B12_A6_18hpi_1	290,717	284,772.494	18	A6
B12_A6_18hpi_2	247,468	277,856.886	18	A6
B12_A6_18hpi_3	194,410	282,167.232	18	A6
B12_A6_24hpi_1	219,782	276,716.214	24	A6
B12_A6_24hpi_2	260,859	276,414.322	24	A6

Figure 9: Result Summary

During the Time Course Expression Analysis, raw counts are transformed according to the normalization method selected in the analysis configuration. Go to **Export Normalized Counts** (side bar) to export normalized counts to a tabular text file.

9.9.6 Charts and Statistics

Different statistics charts can be generated for a global visualization of the results. These charts can be found under the **Side Panel** of the TimeCourse Results viewer.

- **MDS Plot:** Generates a two-dimensional scatterplot in which the distances represent the typical log2 fold changes between samples. You can select an experimental factor by which you want to color the MDS graphic.
- **Venn Diagram:** Diagram showing all possible logical relations between a finite collection of different feature sets (Figure 10(a)[144]). You can choose between two types of Venn Diagram (``Pairwise'' or ``Triple''), and select the sets of significant genes to display.
- **Expression Profile by Gene:** Graph of gene expression profiles over time for a particular gene (Figure 10(b)[145]). It is possible to see them by right-clicking on the chosen gene, and selecting the ``Show Expression Profile'' option.
- **Experiment-wide Expression Profiles:** Plot showing the expression level levels across samples for each cluster of genes (Figure 10(c)[146]).
- **Summary Expression Profiles:** Plot showing the median level expression of each cluster of genes across time (Figure 10(d)[147]).

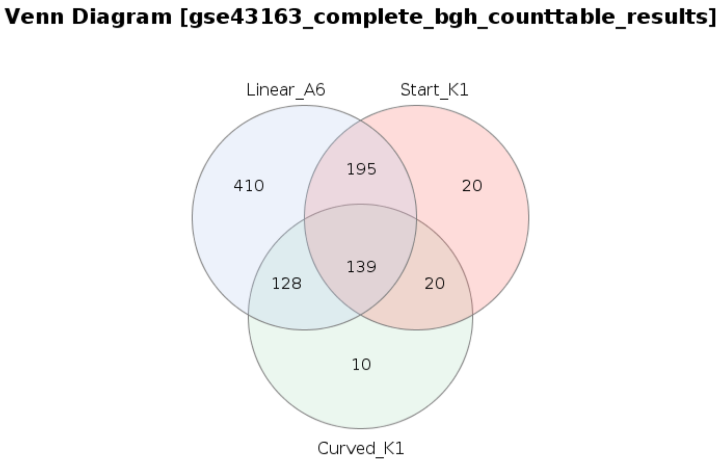


Figure 10(a): Venn Diagram

[144] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598016084#TimeCourseExpressionAnalysis-figure10a

[145] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598016084#TimeCourseExpressionAnalysis-figure10b

[146] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598016084#TimeCourseExpressionAnalysis-figure10c

[147] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598016084#TimeCourseExpressionAnalysis-figure10d

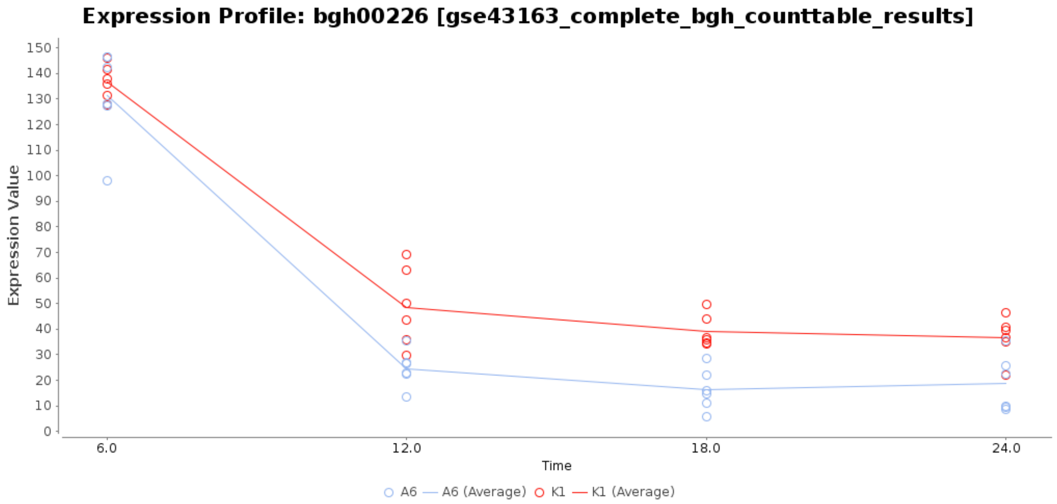


Figure 10(b): Expression Profile by Gene

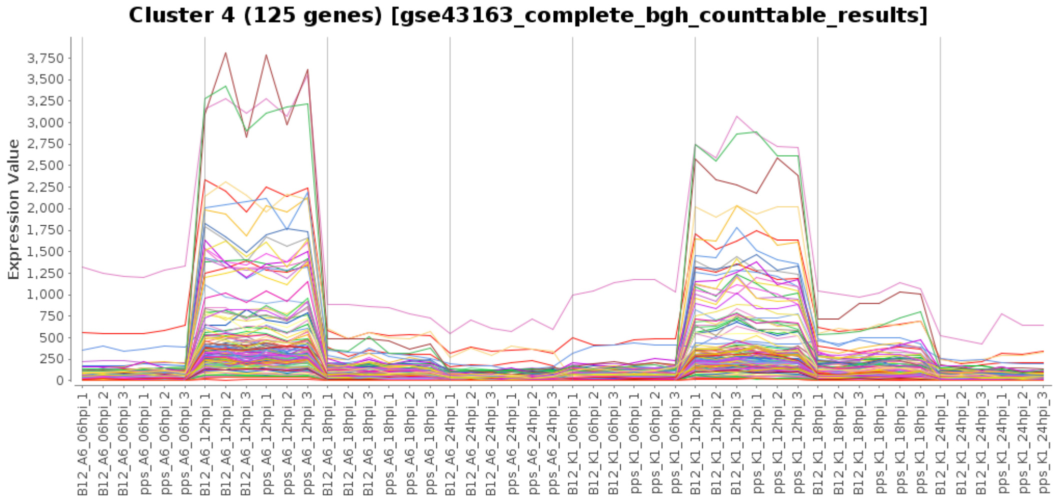


Figure 10(c): Experimental-wide Expression Profiles

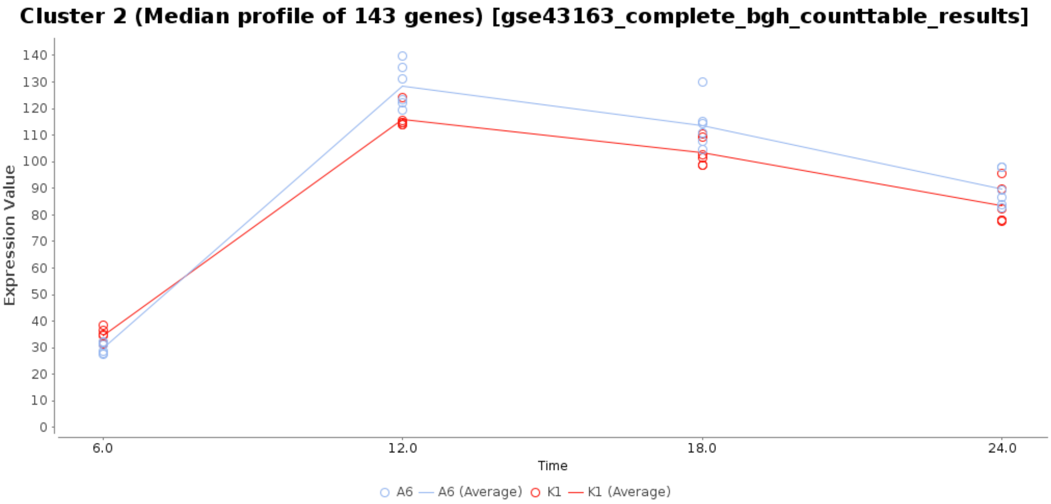


Figure 10(d): Summary Expression Profiles

10 Module Metagenomics

Content of this section:

Metagenomics is a still-evolving field and the experimental design, used tools and pipelines depend highly on the question at hand. This chapter explains the possibilities and features of the metagenomics module in OmicsBox.

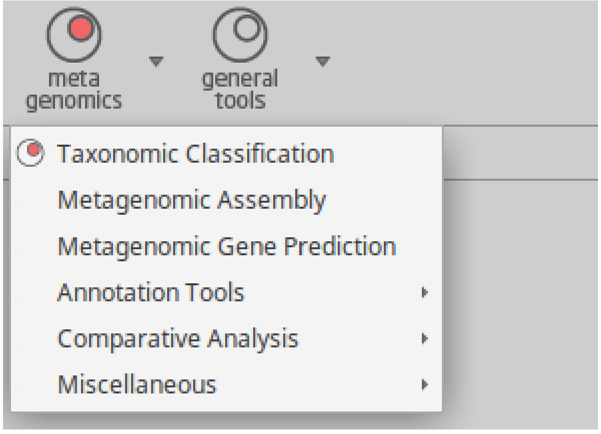


Figure 1: Metagenomics menu

Metagenomic Analysis use case: https://www.biobam.com/metagenomic-analysis-of-two-soda-lakes-with-and-without-cyanobacterial-bloom-with-omicsbox/.

Metagenomics Example Dataset: Download[148]

10.1 Read Quality Control and Assessment

- FastQC(see page 271)
- Trimmomatic(see page 286)

[148] https://resources.biobam.com/omicsbox/example_data/Metagenomics.zip

10.2 Taxonomic Classification

Content of this page:

Taxonomic classification tools match sequences - typically reads or assembled contigs - against a database of microbial genomes to identify the taxon of each sequence. In the early days of metagenomics, the best strategy was to use BLAST to compare each read with all sequences in GenBank. As the reference databases and the size of sequencing data sets have grown, alignment using BLAST has become computationally infeasible, leading to the development of metagenomics classifiers that provide much faster results, although usually with less sensitivity than BLAST. A variety of strategies have been used for the matching step: aligning reads, mapping k-mers, using complete genomes, aligning marker genes only or translating the DNA and aligning to protein sequences.

OmicsBox comes with Kraken as the tool of choice for taxonomic classification because of its overall positive characteristics, such as being 16S and WGS capable and showing good benchmark scores.

10.2.1 Kraken

Kraken is a taxonomic sequence classifier that assigns taxonomic labels to short DNA reads. It does so by examining the k-mers within a read and querying a database with those k-mers. This database contains a mapping of every k-mer in Kraken's genomic library to the lowest common ancestor (LCA) in a taxonomic tree of all genomes that contain that k-mer. The set of LCA taxa that correspond to the k-mers in a read are then analyzed to create a single taxonomic label for the read; this label can be any of the nodes in the taxonomic tree. Kraken is designed to be rapid, sensitive, and highly precise. This approach is feasible for metagenomics WGS as well as 16S/ITS amplicon read input data.

The current Kraken database version is 2019.06 and was created by us. The contained species and strains are listed in this file and can be opened through File > Load > Load Kraken Data to explore the database content (Figure 1(see page 236) and Figure 2(see page 236)).

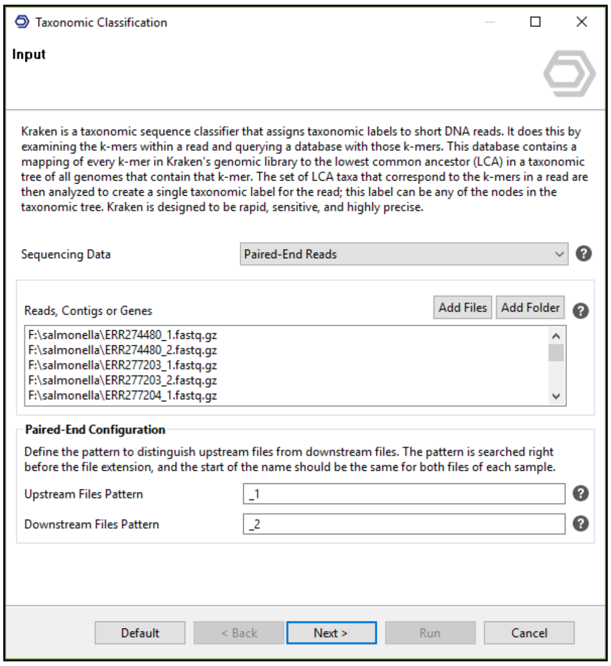
Kraken_2019_06.filter[149]

- **Sequencing Data:** Choose the type of input data: single-end, paired-end or interleaved paired-end reads If paired-end is selected, two files per sample are required and the file pattern has to be provided.

[149] https://biobam.atlassian.net/wiki/download/attachments/753270830/Kraken_2019_06.filter?api=v2&cacheVersion=1&modificationDate=1569929290449&version=1

- **Reads, Contigs or Genes:** Select files that contain the desired input data. Kraken was designed to work with short reads, but works reliable with assembled sequences or genes.
- **Paired-end configuration:** When working with paired-end libraries, a so-called pattern has to be established to help the software distinguish between upstream and downstream read files. Per default, we assume the following pattern:
    - upstream: SampleA_1.fastq
    - downstream: SampleA_2.fastq

⚠ **Note:**
For example, if the upstream file is named SRR037717_1.fastq and the downstream one SRR037717_2.fastq, you should establish "_1" as the upstream pattern and "_2" as the downstream pattern.



Kraken is a taxonomic sequence classifier that assigns taxonomic labels to short DNA reads. It does this by examining the k-mers within a read and querying a database with those k-mers. This database contains a mapping of every k-mer in Kraken's genomic library to the lowest common ancestor (LCA) in a taxonomic tree of all genomes that contain that k-mer. The set of LCA taxa that correspond to the k-mers in a read are then analyzed to create a single taxonomic label for the read; this label can be any of the nodes in the taxonomic tree. Kraken is designed to be rapid, sensitive, and highly precise.

Sequencing Data: Paired-End Reads

Reads, Contigs or Genes

```
F:\salmonella\ERR274480_1.fastq.gz
F:\salmonella\ERR274480_2.fastq.gz
F:\salmonella\ERR277203_1.fastq.gz
F:\salmonella\ERR277203_2.fastq.gz
F:\salmonella\ERR277204_1.fastq.gz
```

Paired-End Configuration

Define the pattern to distinguish upstream files from downstream files. The pattern is searched right before the file extension, and the start of the name should be the same for both files of each sample.

Upstream Files Pattern: _1

Downstream Files Pattern: _2

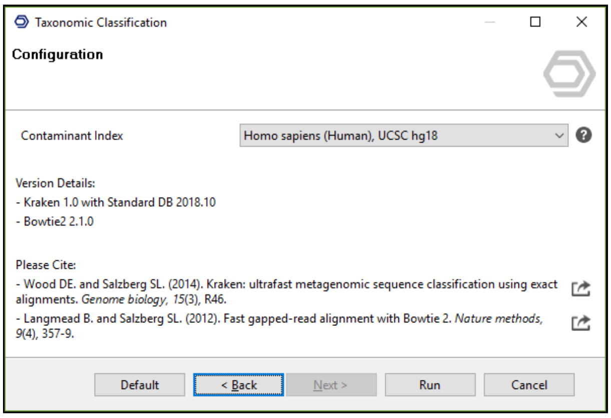
Figure 1. Taxonomic Classification Wizard: input page.

Metagenome sequence data is often "contaminated" by the host organism, e.g. a human gut genome sequencing project will contain reads from Homo sapiens. Therefore, Kraken can be combined with a prior contaminant screening to filter out reads that may stem from a host organism. We offer various model organism genomes, the screening is performed with the help of Bowtie2.

Available genomes:

- Homo sapiens
- Arabidopsis thaliana
- Bos taurus
- Drosophila melanogaster
- Escherichia coli
- Mus musculus

- Rattus norvegicus
- Sus scrofa



Taxonomic Classification

Configuration

Contaminant Index: Homo sapiens (Human), UCSC hg18

Version Details:
- Kraken 1.0 with Standard DB 2018.10
- Bowtie2 2.1.0

Please Cite:
- Wood DE. and Salzberg SL. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome biology, 15*(3), R46.
- Langmead B. and Salzberg SL. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods, 9*(4), 357-9.

Figure 2. Taxonomic Classification Wizard: configuration page.

10.2.1.1 Results

The results of the taxonomic classification with Kraken are:

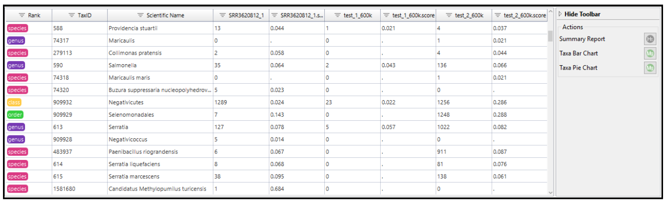
- Main result table, that shows all identified OTUs for each provided sample.
- PDF Report that shows overall input and carried-out analysis information.
- Stacked bar chart to compare samples at specific taxonomic levels.
- Radial cladogram in form of a Krona chart to study OTU abundances in a sample.
- Rarefaction curves to assess the sequencing depth.
- Chao1 (species) diversity curve to evaluate the (species) diversity in the whole data-set.
- Principal Coordinates Analysis plot to get an overview and to identify outliers.

10.2.1.2 Main result table

The result table (Figure 3(see page 238)) shows for each analyzed sample the number of OTUs found and their confidence scores. OTU stands for operational taxonomic unit. A count of 0 means the OTU is not present in the current sample. The evidence scores vary from 0 to 1, where 1 is best. The counts shown are cumulative, meaning that they do not only show direct hits for a certain OTU, but also sum up OTUs that converge into them in the taxonomic tree i.e. the taxonomic tree's hierarchy is taken into account. We use the taxonomic hierarchy from the NCBI in a simplified form and it consists of 8 main levels instead of 33, i.e. the many levels are summarized as shown in Figure 4(see page 238).

The result table can be used to filter for certain taxonomic levels via the column header filter function, e.g. showing only the species or phylum level OTUs.

Right clicking an OTU shows the table context menu to create statistics and id lists, and provides an extra functionality. The **Extract Sequences** option allows to export the read names and actual reads of the currently selected OTUs. This makes it possible to export all reads that have been classified as e.g. bacteria and thus reduce the dataset size for further gene finding and functional annotation. Many other scenarios and use cases exist.



Rank	TaxID	Scientific Name	SRR3620812_1	SRR3620812_1.s...	test_1_600k	test_1_600k.score	test_2_600k	test_2_600k.score
species	588	Providencia stuartii	13	0.044	1	0.021	4	0.037
genus	74317	Maricaulis	0	.	0	.	1	0.021
species	279113	Collimonas pratensis	2	0.058	0	.	4	0.044
genus	590	Salmonella	35	0.064	2	0.043	136	0.066
species	74318	Maricaulis maris	0	.	0	.	1	0.021
species	74320	Buzura suppressaria nucleopolyhedrov...	5	0.023	0	.	0	.
class	909932	Negativicutes	1289	0.024	23	0.022	1256	0.286
order	909929	Selenomonadales	7	0.143	0	.	1248	0.288
genus	613	Serratia	127	0.078	5	0.057	1022	0.082
genus	909928	Negativicoccus	5	0.014	0	.	0	.
species	483937	Paenibacillus riograndensis	6	0.067	0	.	911	0.087
species	614	Serratia liquefaciens	8	0.068	0	.	81	0.076
species	615	Serratia marcescens	38	0.095	0	.	138	0.061
species	1581680	Candidatus Methylopumilus turicensis	1	0.684	0	.	0	.

Figure 3. Taxonomic Classification results table.

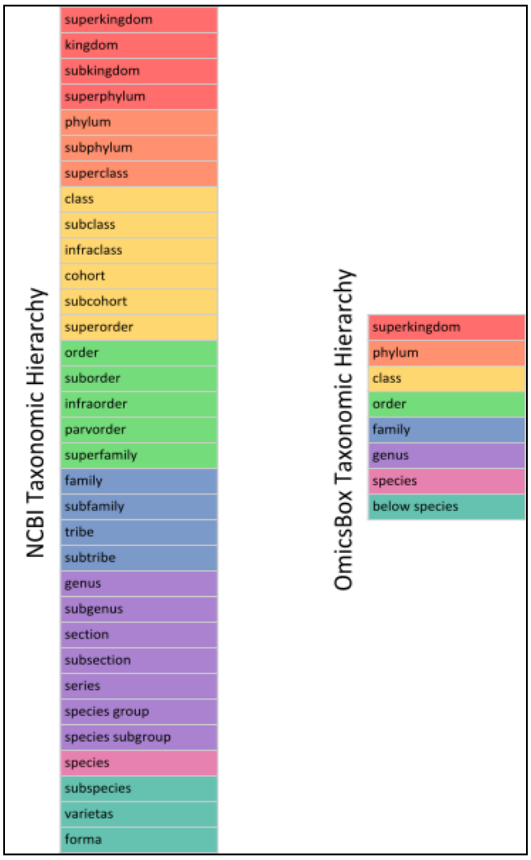


Figure 4. NCBI taxonomic hierarchy.

10.2.1.3 Stacked Bar Chart

The Stacked bar chart (Figure 5(see page 239)) is a combined view for inter sample comparison, separated in the 7 main taxonomic levels. Average OTUs are ordered by abundance from high to low. Only the 500 biggest OTUs are shown for each sample, the remaining are gathered into an extra group called Others. These low frequent OTUs can be analyzed in detail with the Krona Pie Chart. The button **Hide Unclassified** in the top-right corner shows how the percentages change when only taking into account the data that could be classified by Kraken.

The graphic can be exported as a PNG image by clicking the corresponding icon in the top-right corner.

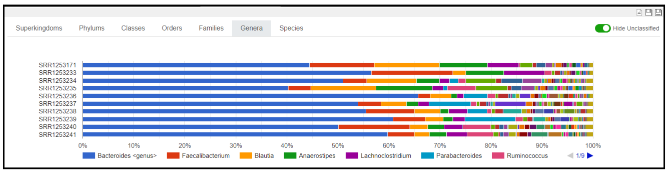


Figure 5. Stacked bar chart.

10.2.1.4 Krona Pie Chart

This graphic (Figure 6(see page 240)) shows a slightly modified Krona chart with various options in the sidepanel. Again, the counts are cumulative and grouped into roughly 8 main levels. However, all direct counts are shown as well, which is helpful when looking at the "below species" level, which includes subspecies and strains.

The currently visualized sample is selected from a list in the sidepanel, the **All Combined** entry shows all samples together in one chart. Furthermore, text sizes can be adjusted and OTUs can be searched. Coloring by average Kraken evidence scores is also possible.

The graphic can be exported as a PNG image and PDF by clicking the corresponding icons in the top-right corner.

10.2.1.5 Summary Report

A summary report which shows basic statistics and alpha-diversity indices for each analyzed sample. It also gives information about the percentages of reads that were classified. In addition, for each of the 7 main taxonomic levels (Superkingdom, Phylum, Class, Order, Family, Genus and Species), the top 10 OTUs per sample are listed.

The graphic can be exported as PDF by clicking the corresponding icon in the top-right corner.

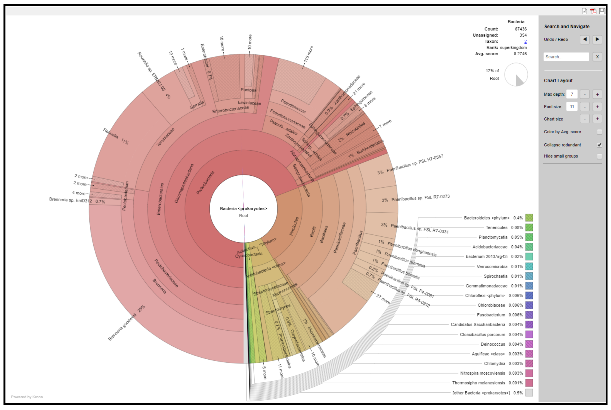


Figure 6. Krona pie chart.

In addition, more **charts and statistics** can be generated to offer a global visualization of the taxonomic classification results. These charts can be found in the sidepanel of the taxonomic classification results.

10.2.1.6 Rarefaction Curves

Rarefaction is a technique widely used in ecology which is applied to OTU analysis. A rarefaction graph (Figure 7(see page 241)) represents the number of expected OTUs (Y-axis) found in *n* NGS reads (X-axis).

The goal of rarefaction is to determine whether sequencing coverage is deep enough to get a good estimate of the total number of the OTUs present in a specific sample.

If the rarefaction curve still presents a growing trend at its end, it means that the coverage is not enough to adequately represent the real microbial diversity of the sample. By contrast, if the curve shows a horizontal asymptote, it means that a good estimation of diversity was obtained.

Note that the results of the rarefaction technique give us a suggestion about the coverage, but they are not conclusive, i.e. rare OTUs that are present in the sample could not be yet observed even if the curve presents an asymptotic trend.

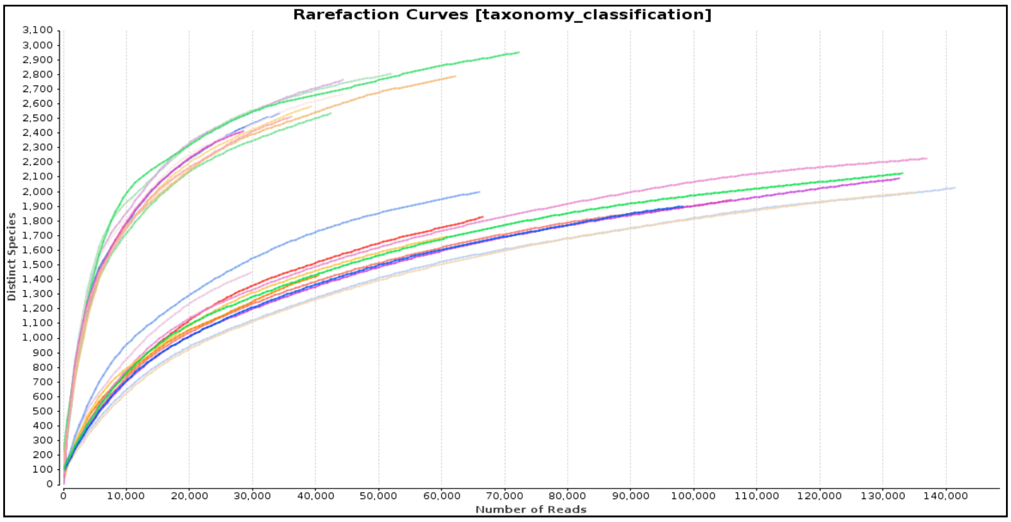


Figure 7. Rarefaction curves.

10.2.1.7 Diversity Curve

An accumulation or diversity curve (Figure 8(see page 241)) plots the cumulative number of distinct OTUs discovered as a function of the number of samples examined. I.e. The minimum, average and maximum number of OTUs, when looking at 1, 2, ... N samples of the current dataset.

This curve can be used to evaluate the benefits in microbial diversity of including additional samples to the dataset, or to compare this diversity across datasets with similar sampling efforts.

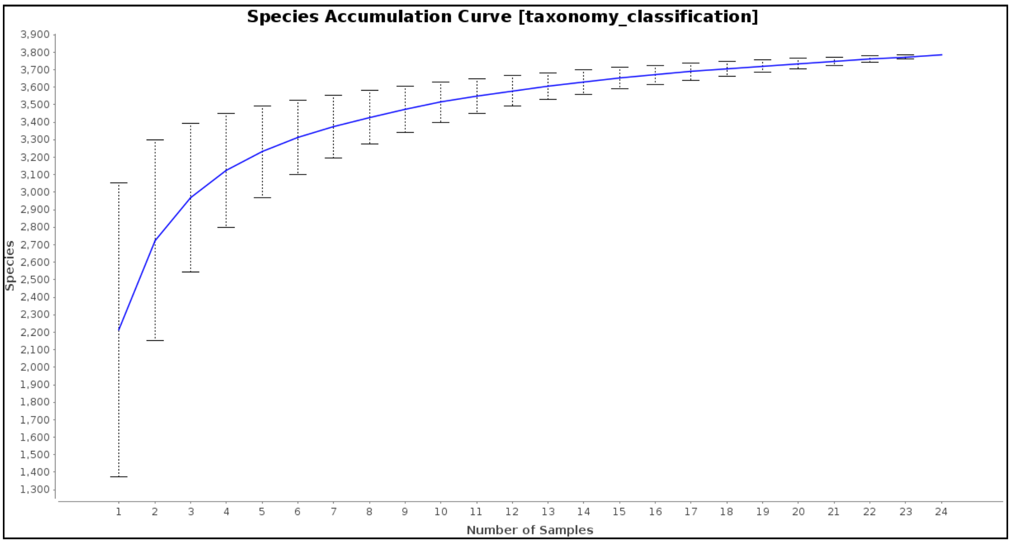


Figure 8. Diversity curve.

10.2.1.8 Principal Coordinate Analysis (PCoA Plot)

PCoA (Figure 9(see page 242)) is a two-dimensional plot in which the Bray-Curtis distances between samples are drawn. You can select an experimental condition to color the points from an experimental design, and the taxonomy level from which the distances will be calculated.

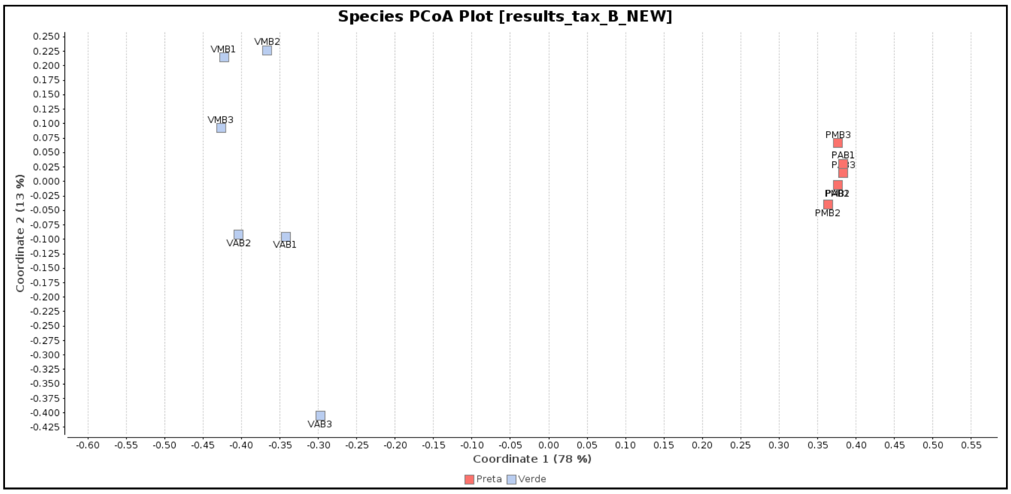


Figure 9. PCoA plot.

ⓘ Wood DE, Salzberg SL: Kraken: ultrafast metagenomic sequence classification using exact alignments[150].*Genome Biology* 2014, 15:R46.

[150] http://genomebiology.com/2014/15/3/R46

10.3 Metagenome Assembly

Content of this page:

10.3.1 Metagenome Assembly

In metagenomics, reads as such (typically Illumina 2 x 150 bp) are usually too short for direct functional characterization. Therefore, we offer metagenome assembly tools as a previous step to gene prediction and functional annotation.

10.3.1.1 metaSPAdes

SPAdes – St. Petersburg genome assembler – is an assembly toolkit containing various assembly pipelines. In OmicsBox, SPAdes is run with the `--meta` option, this flag is recommended when assembling metagenomic data sets (see paper[151] for more details).

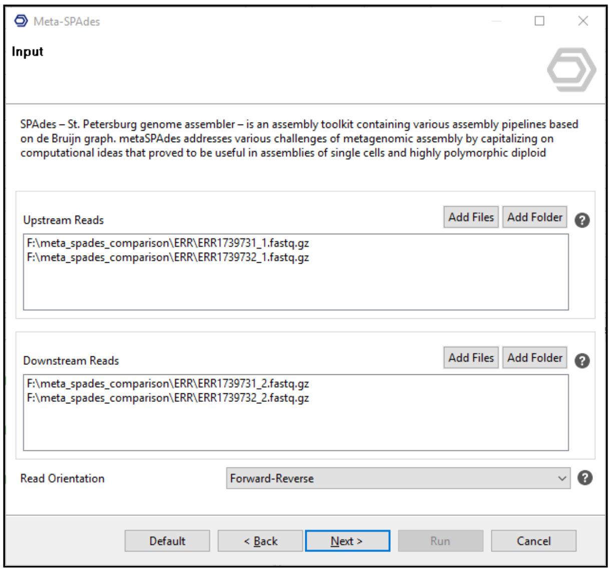
metaSPAdes (figures 1(see page 243), 2(see page 244) and 3(see page 244)) addresses various challenges of metagenomic assembly by capitalizing on computational ideas that proved to be useful in assemblies of single cells and highly polymorphic diploid genomes. Note, that SPAdes was initially designed for small genomes. It was tested on bacterial (both single-cell MDA and standard isolates), fungal and other small genomes. Currently metaSPAdes supports only paired-end libraries. Note that metaSPAdes might be very sensitive to presence of the technical sequences remaining in the data (most notably adapter readthroughs), please run quality control and pre-process your data accordingly.

SPAdes is a de Bruijn graph-based assembler. Input reads are split into k-mers to create the graph and to find its Eulerian path, i.e. the shortest path that visits every edge exactly once. metaSPAdes employs a few modifications to avoid misassemblies, creating shorter high-quality contigs instead of a few long contigs.

metaSPAdes, in comparison to MEGAHIT, needs more resources and takes more time, but also creates better results, i.e. higher Nx values.

- **Up / Downstream Reads:** Choose the files containing the paired end reads respectively. SPAdes is not able to continue if the number of upstream reads doesn't exactly match the number of downstream reads, or if the read names differ.
- **Read Orientation:** For forward-reverse orientation, the forward reads correspond to the left reads and the reverse reads, to the right. Similarly, in reverse-forward orientation left and right reads correspond to reverse and forward reads, respectively.

[151] https://genome.cshlp.org/content/27/5/824.short



Input

SPAdes – St. Petersburg genome assembler – is an assembly toolkit containing various assembly pipelines based on de Bruijn graph. metaSPAdes addresses various challenges of metagenomic assembly by capitalizing on computational ideas that proved to be useful in assemblies of single cells and highly polymorphic diploid

Upstream Reads

F:\meta_spades_comparison\ERR\ERR1739731_1.fastq.gz
F:\meta_spades_comparison\ERR\ERR1739732_1.fastq.gz

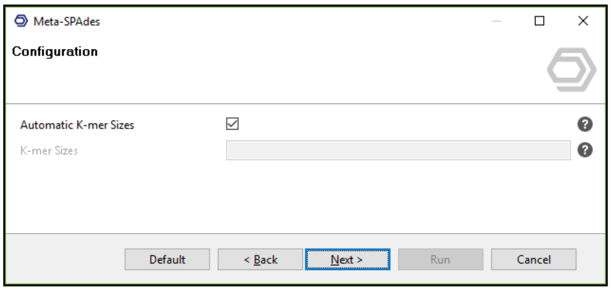
Downstream Reads

F:\meta_spades_comparison\ERR\ERR1739731_2.fastq.gz
F:\meta_spades_comparison\ERR\ERR1739732_2.fastq.gz

Read Orientation: Forward-Reverse

Figure 1. MetaSPAdes assembly wizard: input page.

K-mer sizes: SPAdes will automatically select the k-mer sizes for graph construction. If desired otherwise, please provide a comma separated list of odd k-mer sizes (1-128).



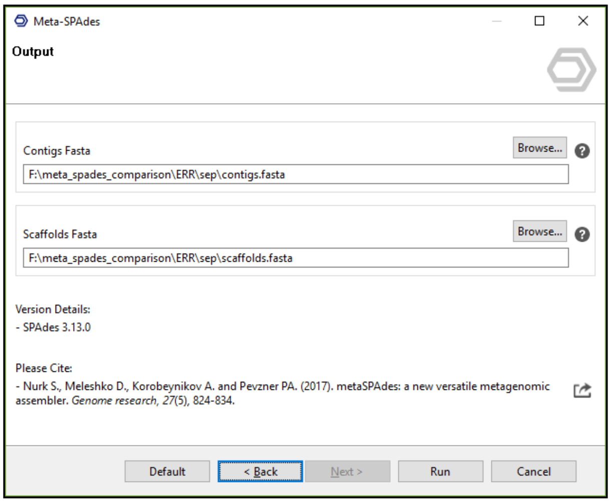
Configuration

Automatic K-mer Sizes ☑

K-mer Sizes

Figure 2. MetaSPAdes assembly wizard: configuration page.

- **Contigs Fasta:** Choose where to save the resulting multi-fasta file.
- **Scaffolds Fasta:** Choose where to save the resulting file containing the scaffolds.



Output

Contigs Fasta
F:\meta_spades_comparison\ERR\sep\contigs.fasta

Scaffolds Fasta
F:\meta_spades_comparison\ERR\sep\scaffolds.fasta

Version Details:
- SPAdes 3.13.0

Please Cite:
- Nurk S., Meleshko D., Korobeynikov A. and Pevzner PA. (2017). metaSPAdes: a new versatile metagenomic assembler. *Genome research*, *27*(5), 824-834.

Figure 3. MetaSPAdes assembly wizard: output page.

The results of SPAdes are the assembled contigs and scaffolds in two separate multi Fasta files. Additionally, Quast is used to generate some basic statistics to asses the quality of the assembly, the PDF is accompanied by an Nx distribution chart.

Bankevich A et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of computational biology : a journal of computational molecular cell biology, 19(5), 455-77.
Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. (2015). MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics. 31(10), 1674–1676,
Nurk S., Meleshko D., Korobeynikov A. and Pevzner PA. (2017). metaSPAdes: a new versatile metagenomic assembler. Genome research, 27(5), 824-834.
van der Walt AJ., van Goethem MW., Ramond JB., Makhalanyane TP., Reva O. and Cowan DA. (2017). Assembling metagenomes, one community at a time. BMC genomics, 18(1), 521.
Vollmers J., Wiegand S. and Kaster AK. (2017). Comparing and Evaluating Metagenome Assembly Tools from a Microbiologist's Perspective - Not Only Size Matters! PloS one, 12(1), e0169662.

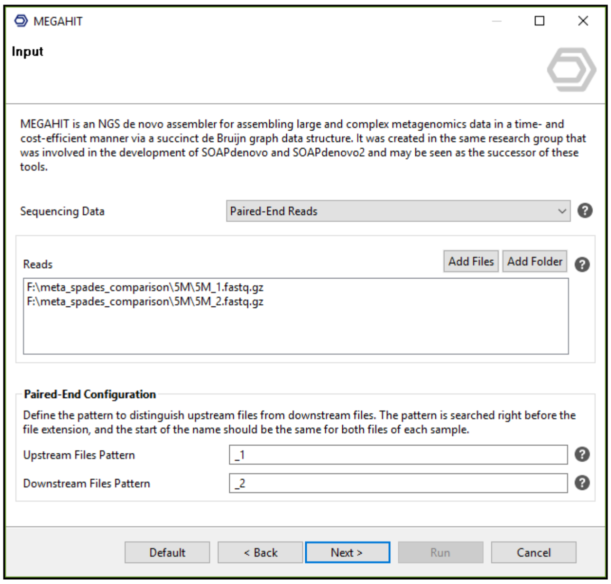
10.3.1.2 MEGAHIT

MEGAHIT is an NGS de novo assembler for assembling large and complex metagenomics data in a time- and cost-efficient manner. MEGAHIT assembles the data as a whole, i.e. no pre-processing like partitioning and normalization is needed (figures 4(see page 245), 5(see page 245) and 6(see page 245)).

Megahit was created in the same research group that was involved in the development of SOAPdenovo and SOAPdenovo2 and may be seen as the successor of these tools. It uses a range of k-mer values for iteratively improving assemblies in a strategy adopted from the IDBA assemblers. It employs a new data structure, the "succinct de Bruijn graph", which has been designed to significantly reduce memory requirements. As an additional step to further reduce memory consumption, only k-mers occurring at a frequency above a specified cutoff are retained as “solid-k-mers”, while the rest is removed as potential sequencing errors. By default, the cutoff value is 2, so k-mers occurring at least twice are kept while singleton k-mers are discarded. Because this eliminates not only sequencing errors, but also removes information from genuinely low abundant genome fragments, a “mercy-k-mer” strategy was introduced which recovers discarded k-mers if they provide new and useful information within a trustworthy context: Discarded singleton k-mers that occur on the same read as “solid k-mers” and are needed to connect these “solid k-mers” within the de Bruin graph are recovered and added to the graph. This minimizes loss of sequencing information while still keeping the influence of sequencing errors low.

- **Sequencing Data:** Choose the type of input data: single-end, paired-end or interleaved paired-end reads If paired-end is selected, two files per sample are required and the file pattern has to be provided.
- **Input Reads:** Provide the files containing sequencing reads. These files are assumed to be in FASTQ / GZ format.
- **Paired-end configuration:** When working with paired-end libraries, a so-called pattern has to be established to help the software distinguish between upstream and downstream read files. Per default, we assume the following pattern:
    - upstream: SampleA_1.fastq
    - downstream: SampleA_2.fastq

> ⚠ **Note:**
> For example, if the upstream file is named SRR037717_1.fastq and the downstream one SRR037717_2.fastq, you should establish "_1" as the upstream pattern and "_2" as the downstream pattern.


MEGAHIT

Input

MEGAHIT is an NGS de novo assembler for assembling large and complex metagenomics data in a time- and cost-efficient manner via a succinct de Bruijn graph data structure. It was created in the same research group that was involved in the development of SOAPdenovo and SOAPdenovo2 and may be seen as the successor of these tools.

Sequencing Data: Paired-End Reads

Reads
F:\meta_spades_comparison\5M\5M_1.fastq.gz
F:\meta_spades_comparison\5M\5M_2.fastq.gz

Paired-End Configuration

Define the pattern to distinguish upstream files from downstream files. The pattern is searched right before the file extension, and the start of the name should be the same for both files of each sample.

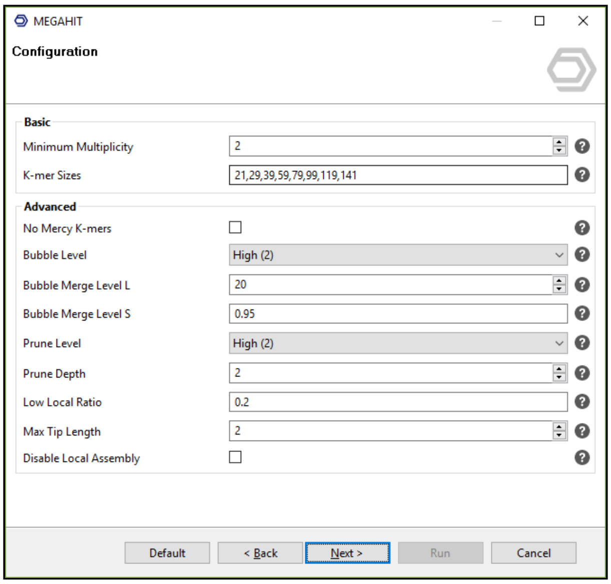
Upstream Files Pattern: _1

Downstream Files Pattern: _2

Figure 4. MEGAHIT assembly wizard: input page.

- **Minimum Multiplicity:** K-mers that appear less times are filtered out. (k_{min}+1)-mer with multiplicity lower than *d* will be discarded. You should be cautious to set *d* less than 2, which will lead to a much larger and noisy graph. We recommend using the default value 2 for metagenomics assembly.
- **K-mer Sizes:** Provide a list of k-mer sizes for iterative graph creation. Values have to be odd and in the range 15-255.
    - for ultra complex metagenomics data such as soil, a larger k_{min}, say 27, is recommended to reduce the complexity of the *de Bruijn* graph. Quality trimming is also recommended.
    - for high-depth generic data, large `--k-min` (25 to 31) is recommended.
    - smaller `--k-step`, say 10, is more friendly to low-coverage datasets.

- **No Mercy K-mers:** Do not add mercy k-mers. Mercy k-mers are specially designed for metagenomics assembly to recover low coverage sequence. For generic dataset >= 30x, MEGAHIT may generate better results with `--no-mercy` option.
- **Bubble Level:** Intensity of bubble merging. Bubbles occur in the de Bruijn graph when several paths start in the same vertex and end in another vertex together.
- **Bubble Merge Level L:** in complex bubbles with length <= L * k-mer size are merged.
- **Bubble Merge Level S:** Complex bubbles with similarity >= S are merged.
- **Prune Level:** Strength of low depth pruning.
- **Prune Depth:** Remove unitigs with average k-mer depths less than this value.
- **Low Local Ratio:** Ratio threshold to define low local coverage contigs.
- **Max Tip Length:** Remove tips shorter than this value.
- **Disable Local Assembly:** The local assembly module was introduced in version 1.0 and creates local contigs between iterations with high confidence k-mers.


MEGAHIT

Configuration

Basic

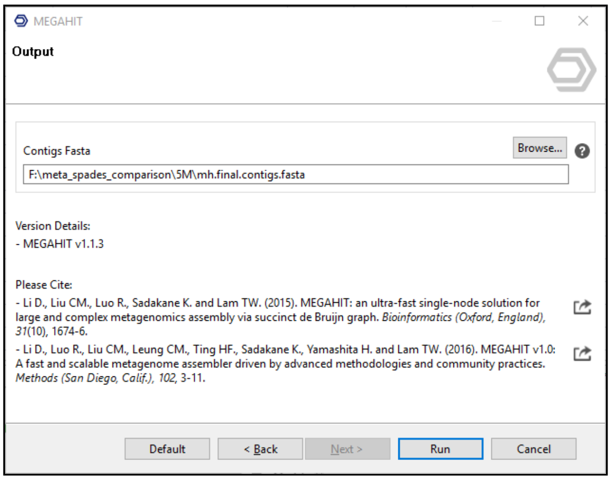
Minimum Multiplicity	2
K-mer Sizes	21,29,39,59,79,99,119,141

Advanced

No Mercy K-mers	□
Bubble Level	High (2)
Bubble Merge Level L	20
Bubble Merge Level S	0.95
Prune Level	High (2)
Prune Depth	2
Low Local Ratio	0.2
Max Tip Length	2
Disable Local Assembly	□

Figure 5. MEGAHIT assembly wizard: configuration page.

- **Contig Fasta:** The final fasta file containing the assembled contigs, will be saved in this file location.



Contigs Fasta

F:\meta_spades_comparison\5M\mh.final.contigs.fasta

Version Details:
- MEGAHIT v1.1.3

Please Cite:
- Li D., Liu CM., Luo R., Sadakane K. and Lam TW. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics (Oxford, England), 31*(10), 1674-6.
- Li D., Luo R., Liu CM., Leung CM., Ting HF., Sadakane K., Yamashita H. and Lam TW. (2016). MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods (San Diego, Calif.), 102,* 3-11.

Figure 6. MEGAHIT assembly wizard: output page.

The results of Megahit are the assembled contigs in a multi Fasta file. Additionally, Quast is used to generate some basic statistics to asses the quality of the assembly, the PDF is accompanied by an Nx chart.

> ⓘ Li D., Liu CM., Luo R., Sadakane K. and Lam TW. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics (Oxford, England), 31(10), 1674-6.
> van der Walt AJ., van Goethem MW., Ramond JB., Makhalanyane TP., Reva O. and Cowan DA. (2017). Assembling metagenomes, one community at a time. BMC genomics, 18(1), 521.
> Vollmers J., Wiegand S. and Kaster AK. (2017). Comparing and Evaluating Metagenome Assembly Tools from a Microbiologist's Perspective - Not Only Size Matters! PloS one, 12(1), e0169662.

10.4 Metagenome Gene Prediction

Content of this page:

10.4.1 Metagenome Gene Prediction

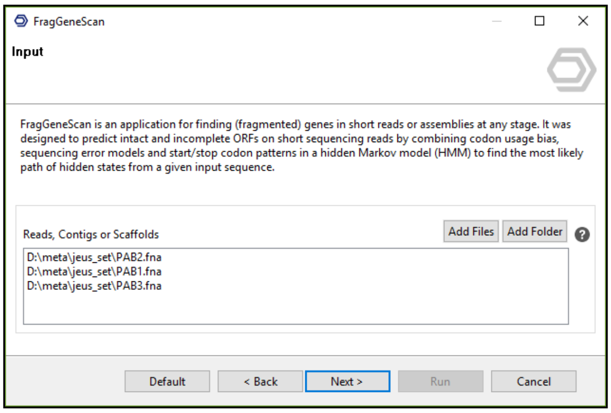
10.4.1.1 FragGeneScan

FragGeneScan is an application for finding (fragmented) genes in short reads. It can also be applied to predict prokaryotic genes in incomplete assemblies or complete genomes. A fundamental step in the analysis of environmental sequence information is the prediction of potential genes or open reading frames (ORFs) encoding the metabolic potential of individual cells and entire microbial communities. FragGeneScan was designed to predict intact and incomplete ORFs on short sequencing reads by combining codon usage bias, sequencing error models and start/stop codon patterns in a hidden Markov model (HMM), to find the most likely path of hidden states from a given input sequence. It provides a promising route for gene recovery in environmental datasets with incomplete assemblies. (Figures 1(see page 248), 2(see page 248) and 3(see page 248))

Features

- HIdden Markov Model supported approach.
- FragGeneScan can be used for gene prediction in complete genomes, assemblies, and short reads.
- Plug and use -- no need to train specific models for different datasets.
- FragGeneScan handles sequencing errors.

- **Reads, Contigs or Scaffolds:** Select files that contain reads or assembled sequences. This tool can work with plain reads instead of contigs


FragGeneScan

Input

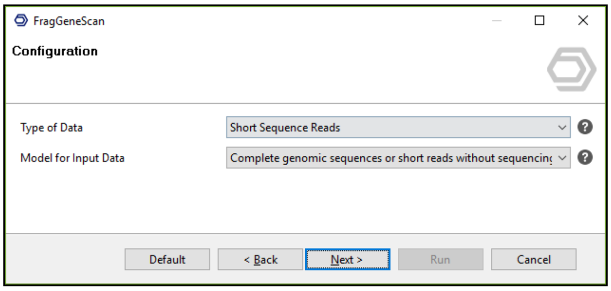
FragGeneScan is an application for finding (fragmented) genes in short reads or assemblies at any stage. It was designed to predict intact and incomplete ORFs on short sequencing reads by combining codon usage bias, sequencing error models and start/stop codon patterns in a hidden Markov model (HMM) to find the most likely path of hidden states from a given input sequence.

Reads, Contigs or Scaffolds

D:\meta\jeus_set\PAB2.fna
D:\meta\jeus_set\PAB1.fna
D:\meta\jeus_set\PAB3.fna

Figure 1. FragGeneScan wizard: input page.

- **Type of Data:** Decide between short sequence reads or assembled sequences as input.
- **Model for Input Data:**
  [complete] for complete genomic sequences or short sequence reads without sequencing error
  [sanger_5] for Sanger sequencing reads with about 0.5% error rate
  [sanger_10] for Sanger sequencing reads with about 1% error rate
  [454_5] for 454 pyrosequencing reads with about 0.5% error rate
  [454_10] for 454 pyrosequencing reads with about 1% error rate
  [454_30] for 454 pyrosequencing reads with about 3% error rate
  [illumina_5] for Illumina sequencing reads with about 0.5% error rate
  [illumina_10] for Illumina sequencing reads with about 1% error rate


FragGeneScan

Configuration

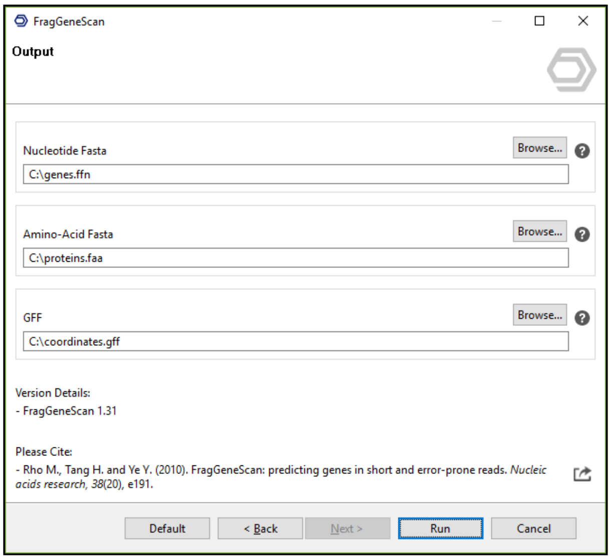
Type of Data: Short Sequence Reads

Model for Input Data: Complete genomic sequences or short reads without sequencing

Figure 2. FragGeneScan wizard: configuration page.

- **Nucleotide Fasta:** Select a file location for the genes multi fasta output.

- **Amino-Acid Fasta:** Select a file location for the protein sequences multi fasta output.
- **GFF:** Select a file location to save the gene feature format file.



Output

Nucleotide Fasta
C:\genes.ffn

Amino-Acid Fasta
C:\proteins.faa

GFF
C:\coordinates.gff

Version Details:
- FragGeneScan 1.31

Please Cite:
- Rho M., Tang H. and Ye Y. (2010). FragGeneScan: predicting genes in short and error-prone reads. *Nucleic acids research, 38*(20), e191.

Figure 3. FragGeneScan wizard: output page.

Rho M., Tang H. and Ye Y. (2010). FragGeneScan: predicting genes in short and error-prone reads. Nucleic acids research, 38(20), e191.
Trimble WL., Keegan KP., D'Souza M., Wilke A., Wilkening J., Gilbert J. and Meyer F. (2012). Short-read reading-frame predictors are not created equal: sequence error causes loss of signal. BMC bioinformatics, 13, 183

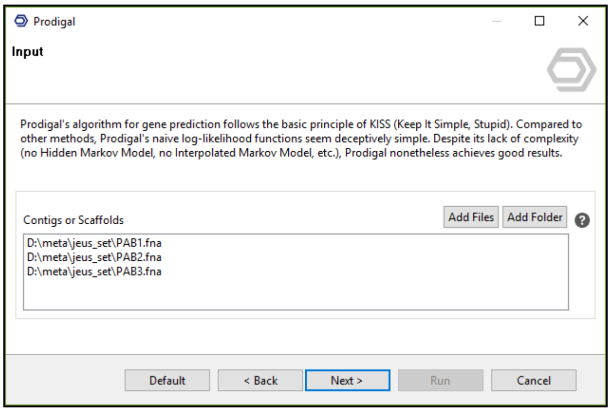
10.4.1.2 Prodigal

Fast, reliable protein-coding gene prediction for prokaryotic genomes. Prodigal's algorithm for gene prediction follows the basic principle of KISS (Keep It Simple, Stupid). Compared to other methods, Prodigal's naive log-likelihood functions seem deceptively simple. Despite its lack of complexity (no Hidden Markov Model, no Interpolated Markov Model, etc.), Prodigal nonetheless achieves good results. (Figures 4(see page 251), 5(see page 251) and 6(see page 251))

Features

- Predicts protein-coding genes: Prodigal provides fast, accurate protein-coding gene predictions.
- Handles draft genomes and metagenomes: Prodigal runs smoothly on finished genomes, draft genomes, and metagenomes.
- Runs unsupervised: Prodigal is an unsupervised machine learning algorithm. It does not need to be provided with any training data, and instead automatically learns the properties of the genome from the sequence itself, including RBS motif usage, start codon usage, and coding statistics.
- Handles gaps and partial genes: The user can specify if Prodigal should build genes across runs of N's as well as how to handle genes at the edges of contigs.
- Identifies translation initiation sites: Prodigal predicts the correct translation initiation site for most genes, and can output information about every potential start site in the genome, including confidence score, RBS motif, and much more.

- **Contigs or Scaffolds:** Select files that contain reads or assembled sequences.



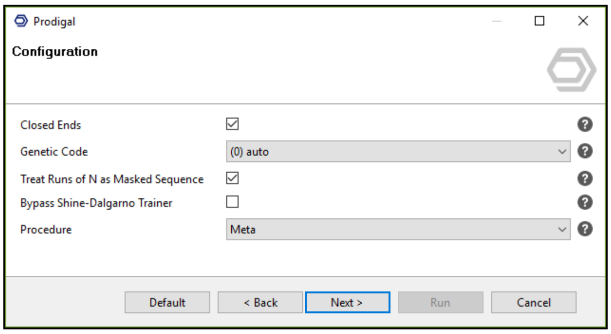
Prodigal's algorithm for gene prediction follows the basic principle of KISS (Keep It Simple, Stupid). Compared to other methods, Prodigal's naive log-likelihood functions seem deceptively simple. Despite its lack of complexity (no Hidden Markov Model, no Interpolated Markov Model, etc.), Prodigal nonetheless achieves good results.

Contigs or Scaffolds

```
D:\meta\jeus_set\PAB1.fna
D:\meta\jeus_set\PAB2.fna
D:\meta\jeus_set\PAB3.fna
```

Figure 4. Prodigal wizard: input page.

- **Closed Ends:** Force genes to have start and stop codon, partial genes are not reported.
- **Genetic Code:** Specify a translation table to use. "auto" will try 11 and then 4 automatically, otherwise the selected genetic code (1-25) will be used.
- **Treat Runs of N as Masked Sequence:** Tells Prodigal not to build genes around sequences of Ns.
- **Bypass Shine-Dalgarno Trainer:** Bypass Shine-Dalgarno trainer and force a full motif scan.



Closed Ends ☑

Genetic Code (0) auto

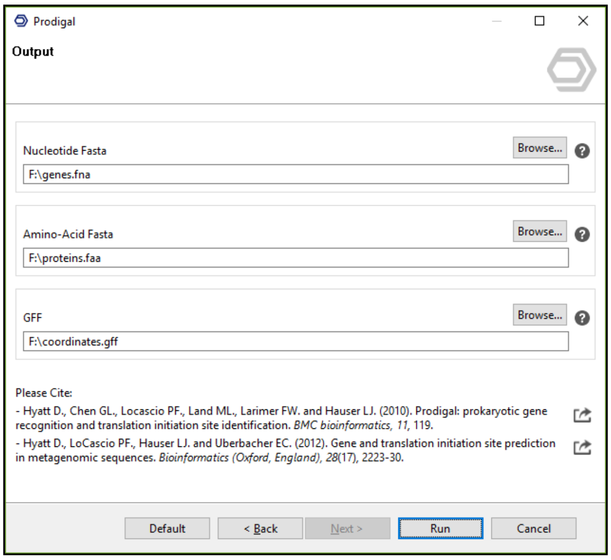
Treat Runs of N as Masked Sequence ☑

Bypass Shine-Dalgarno Trainer □

Procedure Meta

Figure 5. Prodigal wizard: configuration page.

- **Nucleotide Fasta:** Select a file location for the genes multi fasta output.
- **Amino-Acid Fasta:** Select a file location for the protein sequences multi fasta output.
- **GFF:** Select a file location to save the gene feature format file.



Nucleotide Fasta: F:\genes.fna

Amino-Acid Fasta: F:\proteins.faa

GFF: F:\coordinates.gff

Please Cite:
- Hyatt D., Chen GL., Locascio PF., Land ML., Larimer FW. and Hauser LJ. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics, 11,* 119.
- Hyatt D., LoCascio PF., Hauser LJ. and Uberbacher EC. (2012). Gene and translation initiation site prediction in metagenomic sequences. *Bioinformatics (Oxford, England), 28*(17), 2223-30.

Figure 6. Prodigal wizard: output page.

> Hyatt D., Chen GL., Locascio PF., Land ML., Larimer FW. and Hauser LJ. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC bioinformatics, 11, 119.
> Hyatt D., LoCascio PF., Hauser LJ. and Uberbacher EC. (2012). Gene and translation initiation site prediction in metagenomic sequences. Bioinformatics (Oxford, England), 28(17), 2223-30.
> Trimble WL., Keegan KP., D'Souza M., Wilke A., Wilkening J., Gilbert J. and Meyer F. (2012). Short-read reading-frame predictors are not created equal: sequence error causes loss of signal. BMC bioinformatics, 13, 183.

10.5 Functional Annotation

Content of this page:

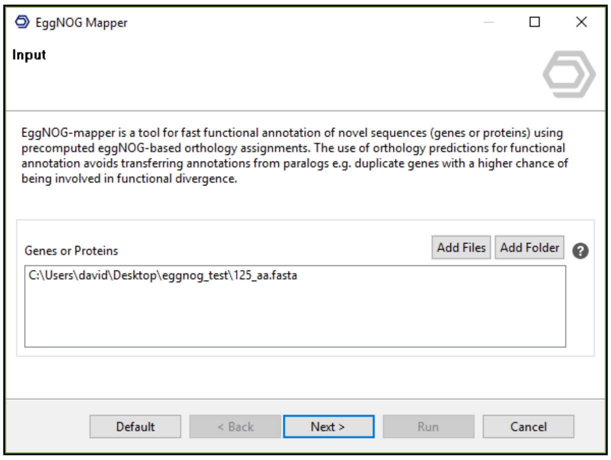
10.5.1 Functional Annotation

10.5.1.1 EggNOG-Mapper

Eggnog-mapper is a tool for fast functional annotation of novel sequences (genes or proteins) using precomputed eggNOG-based orthology assignments. Obvious examples include the annotation of novel genomes, transcriptomes or even metagenomic gene catalogs. The use of orthology predictions for functional annotation is considered more precise than traditional homology searches, as it avoids transferring annotations from paralogs (duplicate genes with a higher chance of being involved in functional divergence). (Figures 1(see page 252) and 2(see page 252))

Details and methodology about the tool and its database are best explained on their website: http://eggnogdb.embl.de/#/app/methods

- **Genes or Proteins:** A multi-fasta file containing genes or proteins.



EggNOG Mapper

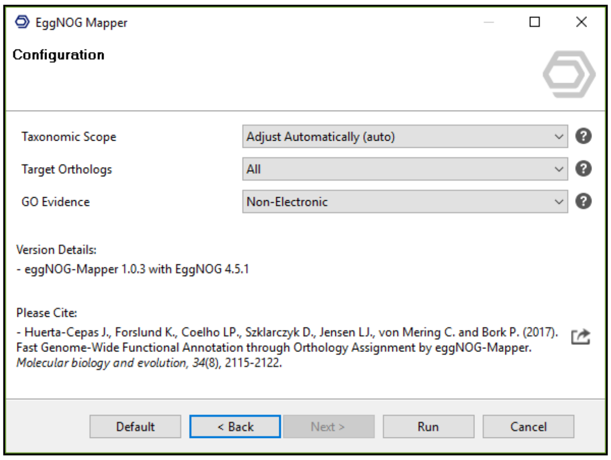
Input

EggNOG-mapper is a tool for fast functional annotation of novel sequences (genes or proteins) using precomputed eggNOG-based orthology assignments. The use of orthology predictions for functional annotation avoids transferring annotations from paralogs e.g. duplicate genes with a higher chance of being involved in functional divergence.

Genes or Proteins: C:\Users\david\Desktop\eggnog_test\125_aa.fasta

Figure 1. EggNOG Mapper wizard: input page.

- **Taxonomic Scope:** Fix the taxonomic scope used for annotation, so only orthologs from a particular clade are used for functional transfer. By default, this is automatically adjusted for every query sequence.
- **Target Orthologs:** Define what type of orthologs should be used for functional transfer.
- **GO Evidence:** Defines what type of GO terms should be used for annotation:
  - experimental = Use only terms inferred from experimental evidence
  - non-electronic = Use only non-electronically curated terms



EggNOG Mapper

Configuration

Taxonomic Scope: Adjust Automatically (auto)

Target Orthologs: All

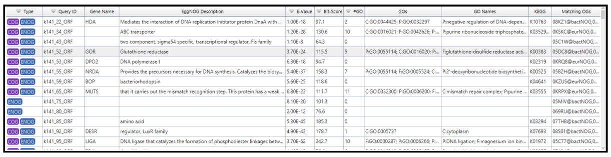
GO Evidence: Non-Electronic

Version Details:
- eggNOG-Mapper 1.0.3 with EggNOG 4.5.1

Please Cite:
- Huerta-Cepas J., Forslund K., Coelho LP., Szklarczyk D., Jensen LJ., von Mering C. and Bork P. (2017). Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Molecular biology and evolution, 34*(8), 2115-2122.

Figure 2. EggNOG Mapper wizard: configuration page.

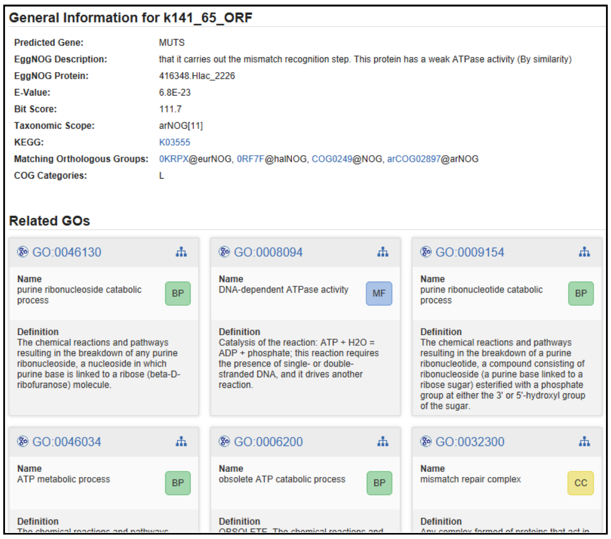
The result table (figure 3(see page 253)) summarizes all annotations that could be transferred with EggNOG Mapper. Besides ordering and filtering, the context menu allows to take a closer look at certain results.



Type	Query ID	Gene Name	EggNOG Description	E-Value	Bit-Score	#GO	GOs	GO Names	KEGG	Matching OGs
COG ENOG	k141_22_ORF	HDA	Mediates the interaction of DNA replication inititator protein DnaA with ...	1.00E-18	97.1	2	C:GO:0044425; P:GO:0032297	P:negative regulation of DNA-depen...	K10763	08KZ1@bactNOG,0...
COG ENOG	k141_34_ORF		ABC transporter	1.20E-28	130.6	10	C:GO:0016021; F:GO:0042626; P:...	P:purine ribonucleoside triphosphate...	K03529...	0KSKC@eurNOG,0...
COG ENOG	k141_43_ORF		two component, sigma54 specific, transcriptional regulator, Fis family	1.10E-8	64.3	0				05C1W@bactNOG,...
COG ENOG	k141_52_ORF	GOR	Glutathione reductase	3.70E-24	115.5	5	P:GO:0055114; C:GO:0016020; P:...	F:glutathione-disulfide reductase acti...	K00383	05DC8@bactNOG,0...
COG ENOG	k141_53_ORF	DPO2	DNA polymerase I	6.30E-18	94.7	0			K02319	0KRQ8@eurNOG,0...
COG ENOG	k141_55_ORF	NRDA	Provides the precursors necessary for DNA synthesis. Catalyzes the biosy...	5.40E-37	158.3	7	P:GO:0055114; F:GO:0005524; C:...	P:2'-deoxyribonucleotide biosyntheti...	K00525	05BZH@bactNOG,0...
COG ENOG	k141_59_ORF	BOP	bacteriorhodopsin	5.60E-25	118.6	0			K04641	0KZUS@eurNOG,0...
COG ENOG	k141_65_ORF	MUTS	that it carries out the mismatch recognition step. This protein has a weak ...	6.80E-23	111.7	11	C:GO:0032300; P:GO:0006200; F:...	C:mismatch repair complex; P:purine ...	K03555	0KRPX@eurNOG,0...
ENOG	k141_75_ORF			8.10E-20	101.3	0				05MIV@bactNOG,0...
ENOG	k141_80_ORF			2.00E-12	76.6	0				069RU@bactNOG,0...
COG ENOG	k141_85_ORF		amino acid	5.30E-45	185.3	0			K03294	07THB@bactNOG,0...
COG ENOG	k141_92_ORF	DESR	regulator, LuxR family	4.90E-43	178.7	1	C:GO:0005737	C:cytoplasm	K07693	08S01@bactNOG,0...
COG ENOG	k141_95_ORF	LIGA	DNA ligase that catalyzes the formation of phosphodiester linkages betw...	3.70E-62	242.7	10	F:GO:0000287; P:GO:0006266; P:...	P:DNA ligation; F:magnesium ion bin...	K01972	05C77@bactNOG,0...

Figure 3. EggNOG Mapper results table.

The annotation details (figure 4(see page 254)) provides link outs where possible and gives detailed information about annotated GOs.



General Information for k141_65_ORF

Predicted Gene:	MUTS
EggNOG Description:	that it carries out the mismatch recognition step. This protein has a weak ATPase activity (By similarity)
EggNOG Protein:	416348.Hlac_2226
E-Value:	6.8E-23
Bit Score:	111.7
Taxonomic Scope:	arNOG[11]
KEGG:	K03555
Matching Orthologous Groups:	0KRPX@eurNOG, 0RF7F@halNOG, COG0249@NOG, arCOG02897@arNOG
COG Categories:	L

Related GOs

GO:0046130
Name
purine ribonucleoside catabolic process
BP
Definition
The chemical reactions and pathways resulting in the breakdown of any purine ribonucleoside, a nucleoside in which purine base is linked to a ribose (beta-D-ribofuranose) molecule.

GO:0008094
Name
DNA-dependent ATPase activity
MF
Definition
Catalysis of the reaction: ATP + H2O = ADP + phosphate; this reaction requires the presence of single- or double-stranded DNA, and it drives another reaction.

GO:0009154
Name
purine ribonucleotide catabolic process
BP
Definition
The chemical reactions and pathways resulting in the breakdown of a purine ribonucleotide, a compound consisting of ribonucleoside (a purine base linked to a ribose sugar) esterified with a phosphate group at either the 3' or 5'-hydroxyl group of the sugar.

GO:0046034
Name
ATP metabolic process
BP
Definition
The chemical reactions and pathways

GO:0006200
Name
obsolete ATP catabolic process
BP
Definition
OBSOLETE. The chemical reactions and

GO:0032300
Name
mismatch repair complex
CC
Definition
Any complex formed of proteins that act in

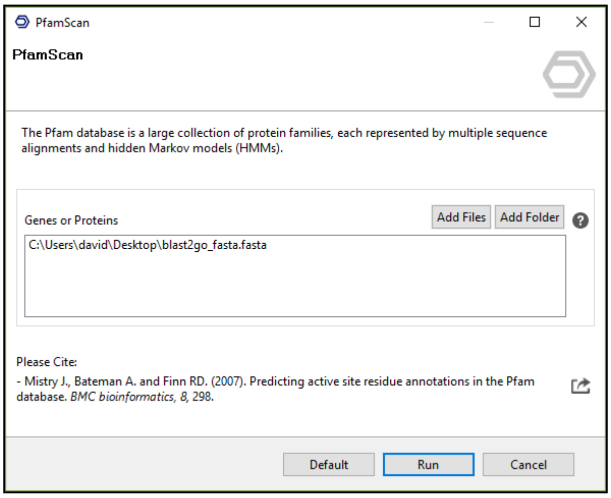
Figure 4. EggNOG Mapper annotation details.

Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. Jaime Huerta-Cepas, Damian Szklarczyk, Lars Juhl Jensen, Christian von Mering and Peer Bork. Submitted (**2016**).
Huerta-Cepas J et al. (2019). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic acids research, 47(D1), D309-D314.

10.5.1.2 PfamScan

Pfam is a database of protein families. Briefly, each Pfam database entry is comprised of a seed alignment, which forms the basis to build a profile hidden Markov model (HMM) using the HMMER software (http://hmmer.org/). The profile HMM is then queried against a sequence database called *pfamseq*, and all matches scoring above the curated threshold (carefully chosen to avoid the inclusion of any known false positives), are aligned back to the profile HMM to generate the full alignment. Where possible, each entry is annotated with functional information derived from literature. To improve sustainability, especially with regard to scaling of the resource, *pfamseq* is derived only from the UniProt Knowledgebase (UniProtKB) sequences that belong to Reference Proteomes, rather than the entirety of UniProtKB. (Figure 5(see page 255))

- **Genes or Proteins:** A multi-fasta file containing genes or proteins.



PfamScan

PfamScan

The Pfam database is a large collection of protein families, each represented by multiple sequence alignments and hidden Markov models (HMMs).

Genes or Proteins

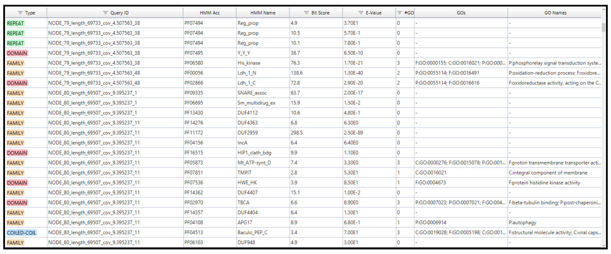
C:\Users\david\Desktop\blast2go_fasta.fasta

Please Cite:

- Mistry J., Bateman A. and Finn RD. (2007). Predicting active site residue annotations in the Pfam database. *BMC bioinformatics*, *8*, 298.

Figure 5. PfamScan wizard.

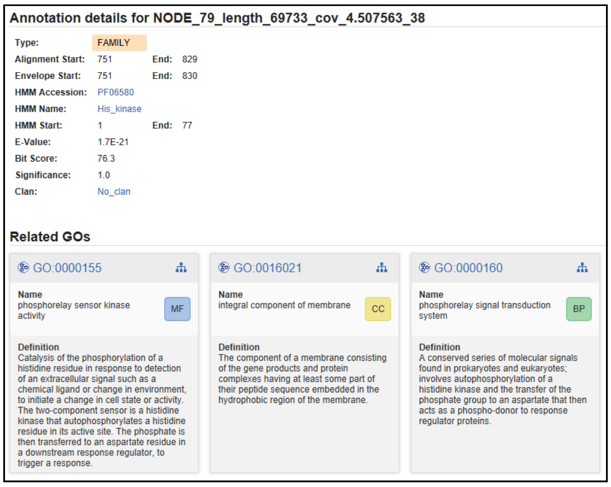
The result table (figure 6(see page 255)) summarizes all PfamScan annotations. Besides ordering and filtering, the context menu allows to take a closer look at certain results.



Type	Query ID	HMM Acc	HMM Name	Bit Score	E-Value	#GO	GOs	GO Names
REPEAT	NODE_79_length_69733_cov_4.507563_38	PF07494	Reg_prop	4.9	3.70E1	0	-	-
REPEAT	NODE_79_length_69733_cov_4.507563_38	PF07494	Reg_prop	10.5	5.70E-1	0	-	-
REPEAT	NODE_79_length_69733_cov_4.507563_38	PF07494	Reg_prop	10.1	7.80E-1	0	-	-
DOMAIN	NODE_79_length_69733_cov_4.507563_38	PF07495	Y_Y_Y	38.7	6.50E-10	0	-	-
FAMILY	NODE_79_length_69733_cov_4.507563_38	PF06580	His_kinase	76.3	1.70E-21	3	F:GO:0000155; C:GO:0016021; P:GO:000...	P:phosphorelay signal transduction syste...
FAMILY	NODE_79_length_69733_cov_4.507563_48	PF00056	Ldh_1_N	138.6	1.30E-40	2	P:GO:0055114; F:GO:0016491	P:oxidation-reduction process; F:oxidore...
DOMAIN	NODE_79_length_69733_cov_4.507563_48	PF02866	Ldh_1_C	72.8	2.90E-20	2	P:GO:0055114; F:GO:0016616	F:oxidoreductase activity, acting on the C...
FAMILY	NODE_80_length_69507_cov_9.395237_1	PF09335	SNARE_assoc	63.7	2.00E-17	0	-	-
FAMILY	NODE_80_length_69507_cov_9.395237_1	PF06695	Sm_multidrug_ex	15.9	1.50E-2	0	-	-
FAMILY	NODE_80_length_69507_cov_9.395237_1	PF13430	DUF4112	10.6	4.80E-1	0	-	-
FAMILY	NODE_80_length_69507_cov_9.395237_11	PF14276	DUF4363	6.8	6.30E0	0	-	-
FAMILY	NODE_80_length_69507_cov_9.395237_11	PF11172	DUF2959	298.5	2.50E-89	0	-	-
FAMILY	NODE_80_length_69507_cov_9.395237_11	PF04156	IncA	6.4	6.40E0	0	-	-
DOMAIN	NODE_80_length_69507_cov_9.395237_11	PF16515	HIP1_clath_bdg	9.9	1.10E0	0	-	-
FAMILY	NODE_80_length_69507_cov_9.395237_11	PF05873	Mt_ATP-synt_D	7.4	3.30E0	3	C:GO:0000276; F:GO:0015078; P:GO:001...	F:proton transmembrane transporter acti...
FAMILY	NODE_80_length_69507_cov_9.395237_11	PF07851	TMPIT	2.8	5.30E1	1	C:GO:0016021	C:integral component of membrane
DOMAIN	NODE_80_length_69507_cov_9.395237_11	PF07536	HWE_HK	3.9	8.50E1	1	F:GO:0004673	F:protein histidine kinase activity
FAMILY	NODE_80_length_69507_cov_9.395237_11	PF14362	DUF4407	15.1	1.00E-2	0	-	-
DOMAIN	NODE_80_length_69507_cov_9.395237_11	PF02970	TBCA	6.6	8.90E0	3	P:GO:0007023; P:GO:0007021; F:GO:004...	F:beta-tubulin binding; P:post-chaperoni...
FAMILY	NODE_80_length_69507_cov_9.395237_11	PF14357	DUF4404	6.4	1.30E1	0	-	-
FAMILY	NODE_80_length_69507_cov_9.395237_11	PF04108	APG17	8.9	6.80E-1	1	P:GO:0006914	P:autophagy
COILED-COIL	NODE_80_length_69507_cov_9.395237_11	PF04513	Baculo_PEP_C	3.4	7.00E1	3	C:GO:0019028; F:GO:0005198; C:GO:001...	F:structural molecule activity; C:viral caps...
FAMILY	NODE_80_length_69507_cov_9.395237_11	PF06103	DUF948	4.9	3.00E1	0	-	-

Figure 6. PfamScan results table.

The annotation details (figure 7(see page 255)) provides link outs where possible and gives detailed information about annotated GOs.



Annotation details for NODE_79_length_69733_cov_4.507563_38

Type:	FAMILY		
Alignment Start:	751	End:	829
Envelope Start:	751	End:	830
HMM Accession:	PF06580		
HMM Name:	His_kinase		
HMM Start:	1	End:	77
E-Value:	1.7E-21		
Bit Score:	76.3		
Significance:	1.0		
Clan:	No_clan		

Related GOs

GO:0000155
Name
phosphorelay sensor kinase activity
MF
Definition
Catalysis of the phosphorylation of a histidine residue in response to detection of an extracellular signal such as a chemical ligand or change in environment, to initiate a change in cell state or activity. The two-component sensor is a histidine kinase that autophosphorylates a histidine residue in its active site. The phosphate is then transferred to an aspartate residue in a downstream response regulator, to trigger a response.

GO:0016021
Name
integral component of membrane
CC
Definition
The component of a membrane consisting of the gene products and protein complexes having at least some part of their peptide sequence embedded in the hydrophobic region of the membrane.

GO:0000160
Name
phosphorelay signal transduction system
BP
Definition
A conserved series of molecular signals found in prokaryotes and eukaryotes; involves autophosphorylation of a histidine kinase and the transfer of the phosphate group to an aspartate that then acts as a phospho-donor to response regulator proteins.

Figure 7. PfamScan annotation details.

ⓘ The Pfam protein families database in 2019[152]: S. El-Gebali, J. Mistry, A. Bateman, S.R. Eddy, A. Luciani, S.C. Potter, M. Qureshi, L.J. Richardson, G.A. Salazar, A. Smart, E.L.L. Sonnhammer, L. Hirsh, L. Paladin, D. Piovesan, S.C.E. Tosatto, R.D. FinnNucleic Acids Research (2019) doi: 10.1093/nar/gky995

10.6 Comparative Analysis

Content of this page:

10.6.1 Comparative Analysis

This section explains the tools for the comparison of identified OTUs and functional annotation compositions between samples.

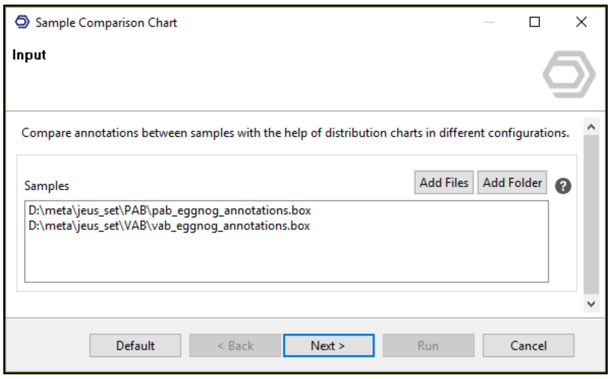
The 2 first tools, sample comparison chart and graph, are visual and allow to compare function abundances between samples. GO Slim generalizes GO annotations to make them comparable.

OTU Differential Abundance Testing identifies over and underrepresented OTUs between samples and conditions with the help of edgeR, a Bioconductor.

10.6.1.1 Sample Comparison Chart

This feature helps to compare annotations between different samples with distribution charts. It also helps to compare GO annotations from EggNOG and PfamScan (or other tools) for the same sample. First, the different samples have to be selected. It is also possible to load external annotations through File > Load > Load Metagenomic GO Annotations and to load them here (figure 1(see page 256)).

[152] https://doi.org/10.1093/nar/gky995



Compare annotations between samples with the help of distribution charts in different configurations.

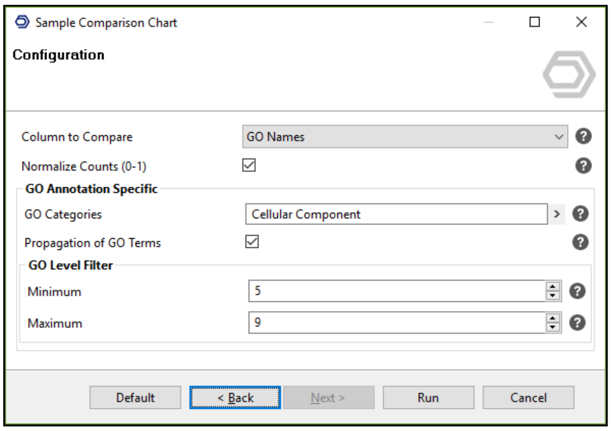
Samples

```
D:\meta\jeus_set\PAB\pab_eggnog_annotations.box
D:\meta\jeus_set\VAB\vab_eggnog_annotations.box
```

Figure 1. Sample comparison charts: input data page.

The second wizard page allows to configure the distribution chart (figure 2(see page 257)).

- **Columns to Compare:** Only annotations that exist in all selected data-sets, can be selected here.
- **Normalize Counts:** In most cases normalizing the counts between 0 and 1 gives better results, because sample sizes are seldom equal.
- **GO Categories:** Create charts for each of the 3 main GO categories.
- **Propagation of GO Terms:** The GO hierarchy is reflected in the resulting chart and helps to compare less and more specific GO annotations at higher levels.
- **GO Level Filter:** Obviously, GOs at higher levels are represented in higher numbers (if propagation is enabled). This option makes it possible to focus on specific levels.



Column to Compare: GO Names

Normalize Counts (0-1): ☑

GO Annotation Specific

GO Categories: Cellular Component

Propagation of GO Terms: ☑

GO Level Filter

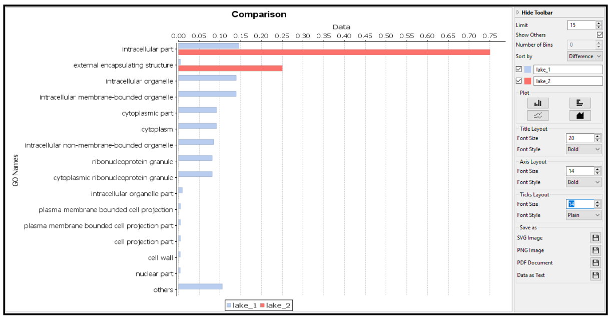
Minimum: 5

Maximum: 9

Figure 2. Configuration page.

On the right we can see the comparison of two different samples, both annotated with EggNOG Mapper. Cellular Component GOs level from 5 and lower are shown, ordered by maximum difference. The graphic visualizes that the red sample has major activity in intracellular parts and external encapsulating structures, while the blue sample works in different parts of the cell.

The graphic can be plotted as vertical or horizontal bars, line or area chart. Samples can be included or excluded, their colors can be changed, as well as their labels. The remaining options are self-explaining (figure 3(see page 257)).

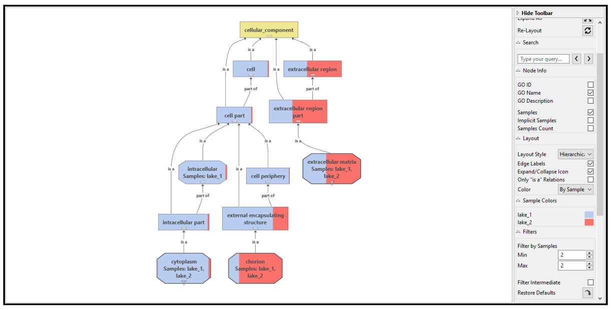

Comparison
Data
0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45 0.50 0.55 0.60 0.65 0.70 0.75
GO Names
intracellular part
external encapsulating structure
intracellular organelle
intracellular membrane-bounded organelle
cytoplasmic part
cytoplasm
intracellular non-membrane-bounded organelle
ribonucleoprotein granule
cytoplasmic ribonucleoprotein granule
intracellular organelle part
plasma membrane bounded cell projection
plasma membrane bounded cell projection part
cell projection part
cell wall
nuclear part
others
lake_1 lake_2
Plot

Hide Toolbar
Limit 15
Show Others ☑
Number of Bins 0
Sort by Difference
☑ lake_1
☑ lake_2
Title Layout
Font Size 20
Font Style Bold
Axis Layout
Font Size 14
Font Style Bold
Ticks Layout
Font Size 14
Font Style Plain
Save as
SVG Image
PNG Image
PDF Document
Data as Text

Figure 3. Sample comparison GO chart.

10.6.1.2 Sample Comparison GO Graph

The colored GO graph on the right side visualizes the same data as above. Only GOs that appear in both samples are shown (Sample Filter = 2). The graph nodes are colored with different areas for each sample. The area's sizes depend on the relative counts (figure 4(see page 258)).



Hide Toolbar
Re-Layout
Search
Type your query...
Node Info
GO ID □
GO Name ☑
GO Description □
Samples ☑
Implicit Samples □
Samples Count □
Layout
Layout Style Hierarchic
Edge Labels ☑
Expand/Collapse Icon ☑
Only "is a" Relations □
Color By Sample
Sample Colors
lake_1
lake_2
Filters
Filter by Samples
Min 2
Max 2
Filter Intermediate □
Restore Defaults

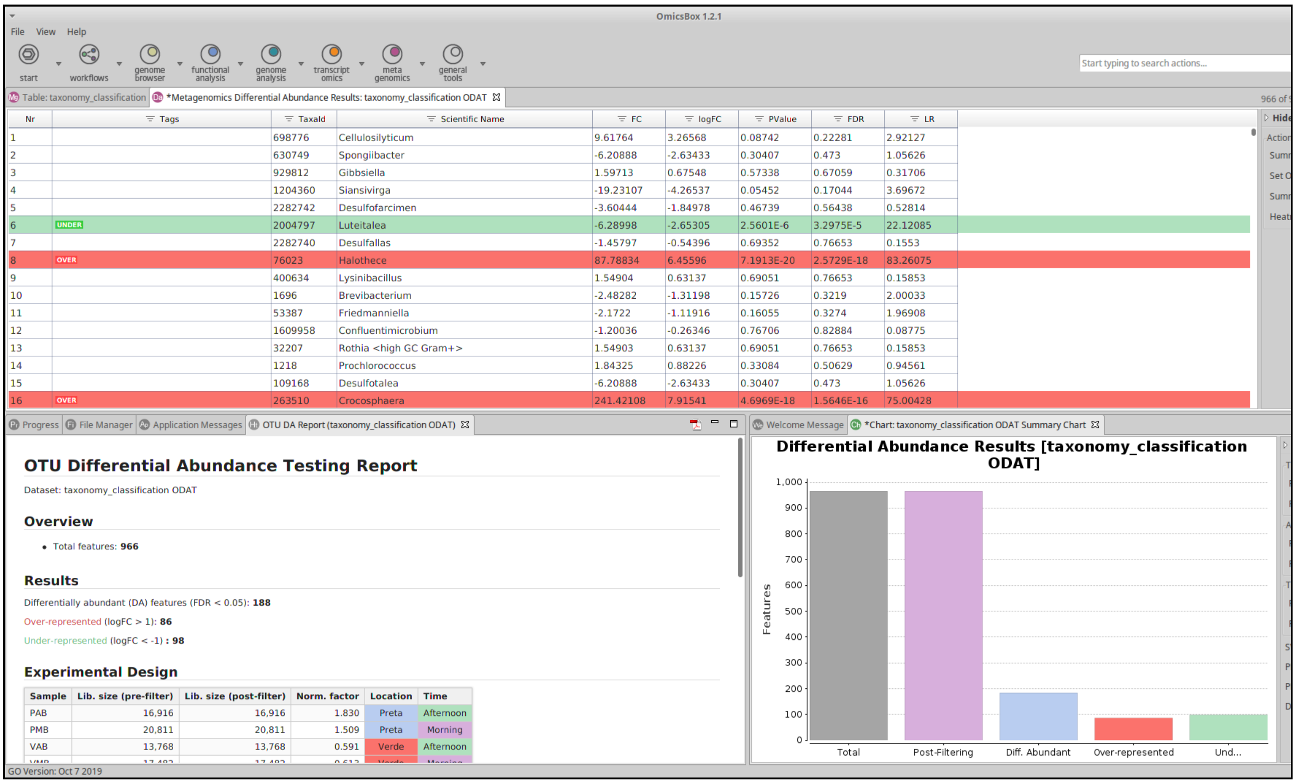
Figure 4. Sample comparison GO graph.

10.6.1.3 GO Slim

GO Slim is a reduced version of the Gene Ontology that contains a selected number of relevant GOs. More specifically, GO annotations are generalized and lifted up in the hierarchy. This can be seen as a way to normalize GO annotations to simplify comparison between samples.

10.6.1.4 OTU Differential Abundance Testing

The OTU Differential Abundance Testing is a tool to identify Operational Taxonomic Units (OTUs) that significantly differ between two microbial communities. This feature is based on edgeR, which belongs to the Bioconductor project, and implements statistical tests to evaluate the significance of OTU abundances between a contrast and a reference group.



Nr	Tags	TaxaId	Scientific Name	FC	logFC	PValue	FDR	LR
1		698776	Cellulosilyticum	9.61764	3.26568	0.08742	0.22281	2.92127
2		630749	Spongiibacter	-6.20888	-2.63433	0.30407	0.473	1.05626
3		929812	Gibbsiella	1.59713	0.67548	0.57338	0.67059	0.31706
4		1204360	Siansivirga	-19.23107	-4.26537	0.05452	0.17044	3.69672
5		2282742	Desulfofarcimen	-3.60444	-1.84978	0.46739	0.56438	0.52814
6	UNDER	2004797	Luteitalea	-6.28998	-2.65305	2.5601E-6	3.2975E-5	22.12085
7		2282740	Desulfallas	-1.45797	-0.54396	0.69352	0.76653	0.1553
8	OVER	76023	Halothece	87.78834	6.45596	7.1913E-20	2.5729E-18	83.26075
9		400634	Lysinibacillus	1.54904	0.63137	0.69051	0.76653	0.15853
10		1696	Brevibacterium	-2.48282	-1.31198	0.15726	0.3219	2.00033
11		53387	Friedmanniella	-2.1722	-1.11916	0.16055	0.3274	1.96908
12		1609958	Confluentimicrobium	-1.20036	-0.26346	0.76706	0.82884	0.08775
13		32207	Rothia <high GC Gram+>	1.54903	0.63137	0.69051	0.76653	0.15853
14		1218	Prochlorococcus	1.84325	0.88226	0.33084	0.50629	0.94561
15		109168	Desulfotalea	-6.20888	-2.63433	0.30407	0.473	1.05626
16	OVER	263510	Crocosphaera	241.42108	7.91541	4.6969E-18	1.5646E-16	75.00428

OTU Differential Abundance Testing Report

Dataset: taxonomy_classification ODAT

Overview

- Total features: **966**

Results

Differentially abundant (DA) features (FDR < 0.05): **188**

Over-represented (logFC > 1): **86**

Under-represented (logFC < -1) **: 98**

Experimental Design

Sample	Lib. size (pre-filter)	Lib. size (post-filter)	Norm. factor	Location	Time
PAB	16,916	16,916	1.830	Preta	Afternoon
PMB	20,811	20,811	1.509	Preta	Morning
VAB	13,768	13,768	0.591	Verde	Afternoon

Differential Abundance Results [taxonomy_classification ODAT]

Figure 5. Differential Abundance Testing: presentation of results.

With a Taxonomic Classification result opened, go to **Metagenomics → Comparative Analysis → OTU Differential Abundance Testing**. In the wizard, you can select the parameters to run the test. It is divided into three different sections: filtering and normalization (figure 6(see page 260)), experimental design (figure 8(see page 261)) and statistical test (figure 9(see page 262)).

First Wizard Page - Filtering and Normalization

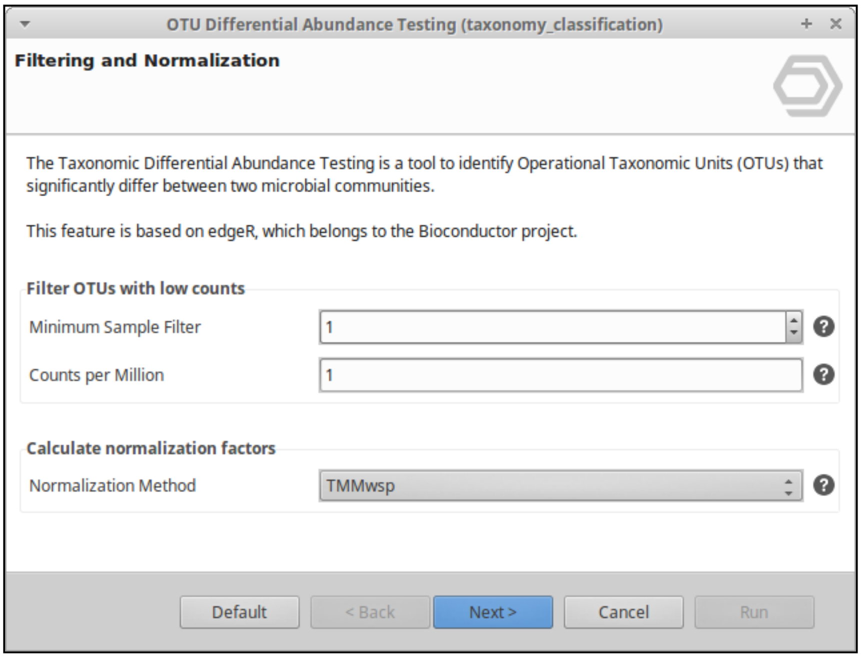
OTUs with low counts will not be considered for the test as they provide little evidence of differential abundance. There are two different filtering steps:

- **Counts per Million Filter.** Set a filter to exclude OTUs with low counts across all samples. Filtering is performed on a count-per-million (CPM) basis to account for differences in library sizes between samples (e.g. a CPM of 1 corresponds to a count of 6 in a sample with 6 million total counts). Set this value to 0 if no filtering is desired.
- **Minimum Samples Filter.** Set a minimum number of samples in which the CPM has to be above the previous filter. If this value is set to e.g. 5, at least 5 of the samples have to show a count above the given CPM. The number of samples of the smallest group is usually used (e.g. in an experiment that has 2 replicates for each condition or group, an OTU should be counted in at least 2 samples). Set this value to 0 if no filtering is desired.

In this test, the normalization takes the form of scaling factors for library sizes that enter into the statistical model. These correctional factors are used to compute the effective library sizes. 5 different options are available for the normalization step:

- **TMM (Trimmed Mean of M-values).** The M-values are weighted according to inverse variances and computed by the delta method for logarithms of binomial random models.

- **TMMwsp (TMM with singleton pairing).** This is a variant of TMM that is intended to perform better for data with a high proportion of zeros (default).
- **RLE (Relative Log Expression).** Scale factors are the median ratio of each sample to the median library (geometric mean of all samples).
- **Upper-quartile.** 75% quantiles for the counts of each library are used to calculate the scale factors.
- **None.** All normalization factors are set to 1.



OTU Differential Abundance Testing (taxonomy_classification)

Filtering and Normalization

The Taxonomic Differential Abundance Testing is a tool to identify Operational Taxonomic Units (OTUs) that significantly differ between two microbial communities.

This feature is based on edgeR, which belongs to the Bioconductor project.

Filter OTUs with low counts

Minimum Sample Filter: 1

Counts per Million: 1

Calculate normalization factors

Normalization Method: TMMwsp

Figure 6. Differential Abundance Testing wizard: filtering and normalization page.

Second Wizard Page - Experimental Design

Here, the two groups for the test, reference and contrast, have to be specified. You can select the groups by choosing which samples from the taxonomic classification project you want to include in each one, or by loading an experimental design file and selecting the conditions you want to test.

Select samples (no experimental design file loaded)

Select the samples to be considered for the test and divide them into two groups or conditions. The **Contrast Group** will be the samples which will be tested against the **Reference Group**.

Experimental design file

You can load your **experimental design file**. This file must contain the sample names in the first column and the experimental conditions of each sample in the following ones, as can be seen in figure 7(see page 260). Please make sure the sample names in the first column of this experimental design file match exactly with the samples in the taxonomic classification result.

This experimental design file must be in **tsv format** (tab-separated values file). In this kind of files, each field is separated with a tab character. Please do not use spaces and avoid strange characters when writing your experimental design file to be sure that it will be correctly read and processed.

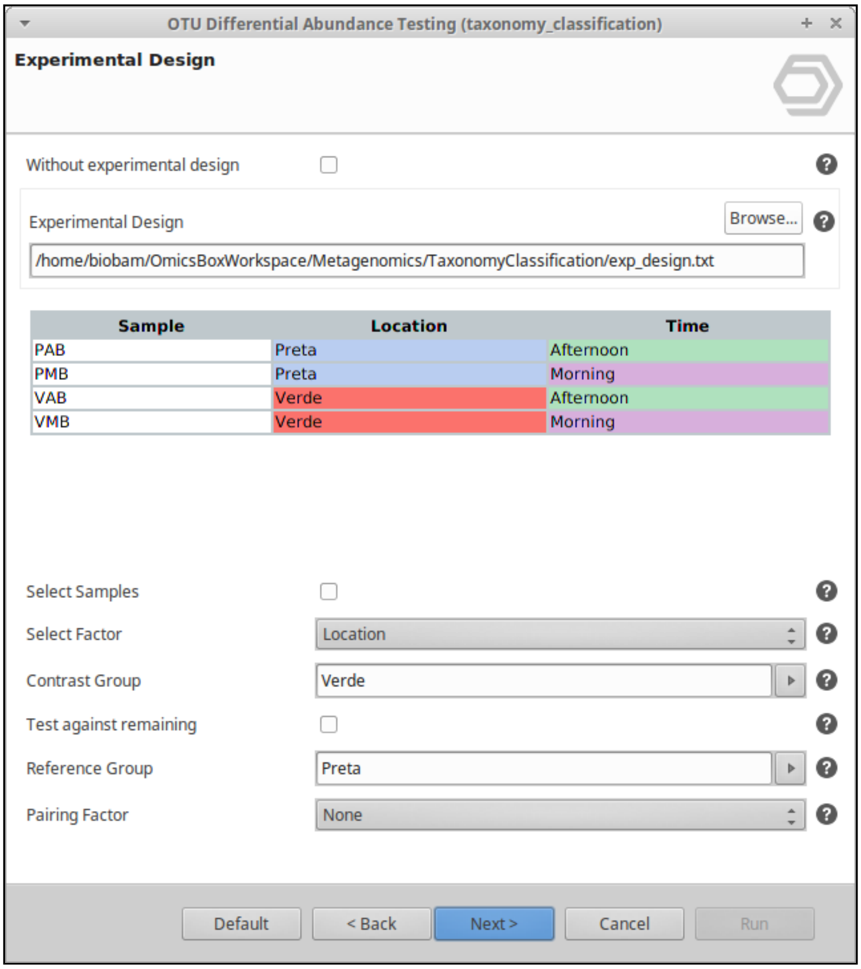
Experimental Design

```
Sample  Lake    Time
PAB     Preta   Afternoon
PMB     Preta   Morning
VAB     Verde   Afternoon
VMB     Verde   Morning
```

Figure 7. Experimental Design file.

Once the file is properly loaded, you can select an **experimental factor** from the experimental design and the conditions to test in both, Contrast and Reference group. You can also select samples separately as described in the previous section if the **Select Samples** option is checked.

If a paired design is desired, a **Pairing Factor** from the experimental design can be optionally selected to adjust for the baseline difference of this factor. Note that this option is only available if you have provided an experimental design file.



OTU Differential Abundance Testing (taxonomy_classification)

Experimental Design

Without experimental design ☐

Experimental Design: /home/biobam/OmicsBoxWorkspace/Metagenomics/TaxonomyClassification/exp_design.txt

Sample	Location	Time
PAB	Preta	Afternoon
PMB	Preta	Morning
VAB	Verde	Afternoon
VMB	Verde	Morning

Select Samples ☐

Select Factor: Location

Contrast Group: Verde

Test against remaining ☐

Reference Group: Preta

Pairing Factor: None

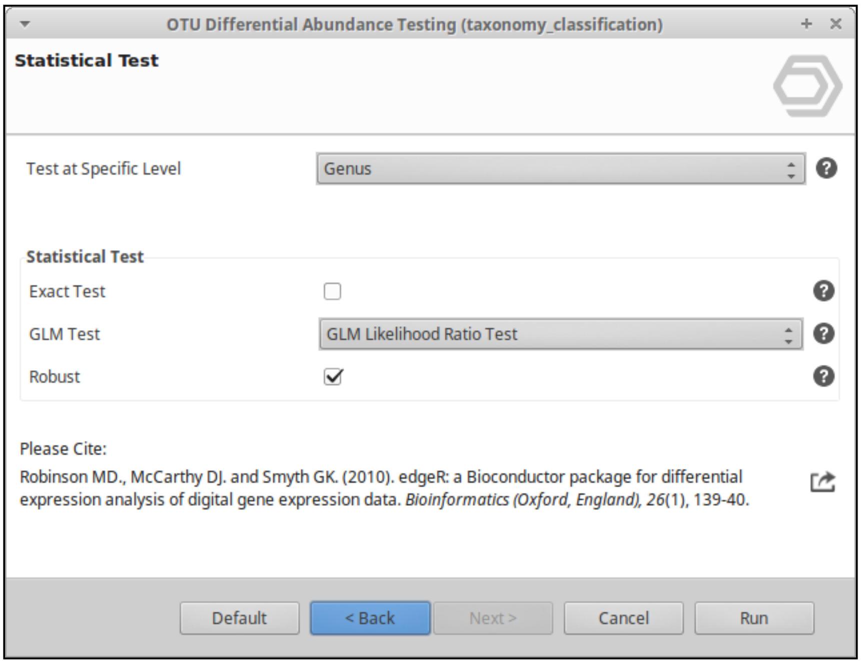
Figure 8. Differential Abundance Testing wizard: experimental design page.

Third Wizard Page - Statistical Test

You can **Test at Specific Taxonomic Levels** to only consider results for a specific taxon (species, genus, family, ...).

Here, you can select the statistical test to be used to detect the differentially abundant OTUs. The test will suppose that the OTU counts across groups are distributed as negative binomial random variables. Two different kinds of tests are available:

- **Exact Test.** Run an Exact Test to detect a difference in mean between two groups of OTU abundance libraries, reference and contrast groups. This test is performed for each OTU and can only be used if no pairing factor is selected.
- **Generalized Linear Model.** Fit a negative binomial generalized log-linear model (GLM) to the counts for each OTU. Two different GLM tests are allowed:
  - **GLM Likelihood Ratio Test.** This mode conducts likelihood ratio tests for the coefficients in the linear model using the Cox-Reid dispersion estimates.
  - **GLM Quasi Likelihood F-Test.** It is similar to the LRT test, except that it replaces likelihood ratio tests with empirical Bayes quasi-likelihood F tests. This test provides a more robust and reliable error rate control when the number of replicates is small.


OTU Differential Abundance Testing (taxonomy_classification)

Statistical Test

Test at Specific Level: Genus

Statistical Test

Exact Test: □

GLM Test: GLM Likelihood Ratio Test

Robust: ☑

Please Cite:

Robinson MD., McCarthy DJ. and Smyth GK. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, *26*(1), 139-40.

Default
< Back
Next >
Cancel
Run

Figure 9. Differential Abundance Testing wizard: statistical test page.

Results

Once the taxonomic abundance analysis has finished, a new **table with the results** will open (figure 10(see page 263)). Each row of this table corresponds to a different tested OTU. Each column contains:

- **Tags.** Indicate if a specific OTU is overrepresented -OVER- (FDR < 0.05 and logFC > 1) or underrepresented -UNDER- (FDR < 0.05 and logFC < -1) in the contrast sample.
- **FC (Fold Change).** The ratio between the mean abundance value of a specific OTU in the contrast condition and this value in the reference condition, if the mean abundance value in the contrast group is bigger than in the reference group. If this value is bigger in the reference group, then the FC is

calculated as the ratio between the mean abundance value in the reference condition and the value in the contrast condition with a negative sign. By default, an OTU is defined as overrepresented if FC > 2, and it is underrepresented if FC < -2.

- **LogFC.** The log2 FC. By default, an OTU is defined as overrepresented if logFC > 1, and it is underrepresented if logFC < -1 if it is statistically significant (FDR < 0.05 by default).
- **LogCPM.** The average log2-counts-per-millions.
- **LR (Likelihood Ratio).** Likelihood Ratio statistic for the GLM (only if GLM LR test is selected).
- **F.** Quasi-likelihood F-statistic for the GLM (only if GLM QL test is selected).
- **P-value.** The p-value for the null hypothesis of non-differential abundance.
- **FDR.** A corrected p-value for multiple testing comparisons (Benjamini Y., Hochberg Y., 1995). If meeting the logFC criterion (logFC > 1 or logFC < -1 by default), an OTU must have an FDR < 0.05 to be considered as differentially abundant.

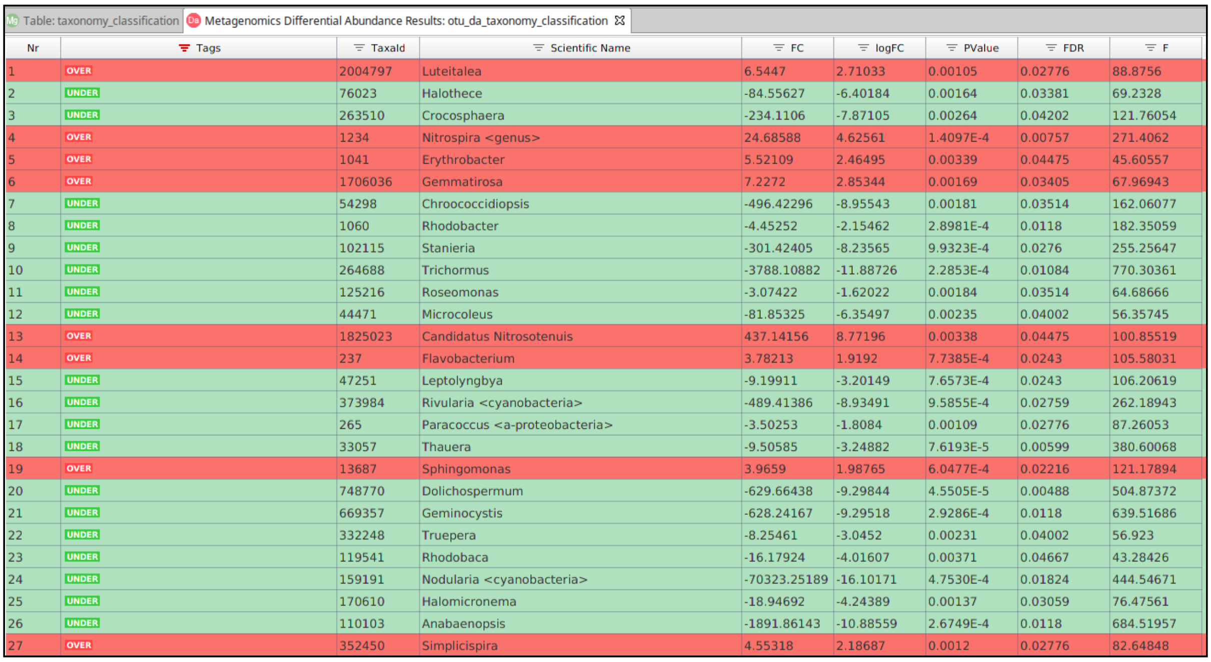


Table: taxonomy_classification
Metagenomics Differential Abundance Results: otu_da_taxonomy_classification

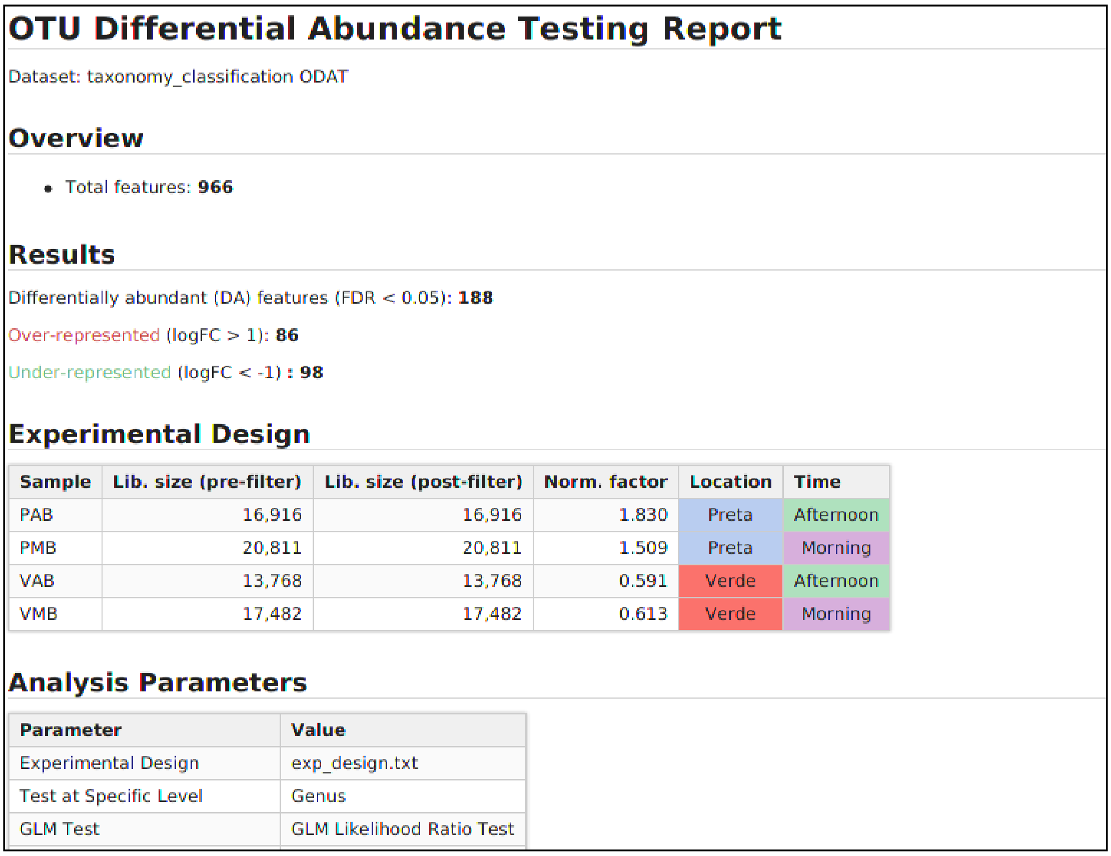
Nr	Tags	TaxaId	Scientific Name	FC	logFC	PValue	FDR	F
1	OVER	2004797	Luteitalea	6.5447	2.71033	0.00105	0.02776	88.8756
2	UNDER	76023	Halothece	-84.55627	-6.40184	0.00164	0.03381	69.2328
3	UNDER	263510	Crocosphaera	-234.1106	-7.87105	0.00264	0.04202	121.76054
4	OVER	1234	Nitrospira <genus>	24.68588	4.62561	1.4097E-4	0.00757	271.4062
5	OVER	1041	Erythrobacter	5.52109	2.46495	0.00339	0.04475	45.60557
6	OVER	1706036	Gemmatirosa	7.2272	2.85344	0.00169	0.03405	67.96943
7	UNDER	54298	Chroococcidiopsis	-496.42296	-8.95543	0.00181	0.03514	162.06077
8	UNDER	1060	Rhodobacter	-4.45252	-2.15462	2.8981E-4	0.0118	182.35059
9	UNDER	102115	Stanieria	-301.42405	-8.23565	9.9323E-4	0.0276	255.25647
10	UNDER	264688	Trichormus	-3788.10882	-11.88726	2.2853E-4	0.01084	770.30361
11	UNDER	125216	Roseomonas	-3.07422	-1.62022	0.00184	0.03514	64.68666
12	UNDER	44471	Microcoleus	-81.85325	-6.35497	0.00235	0.04002	56.35745
13	OVER	1825023	Candidatus Nitrosotenuis	437.14156	8.77196	0.00338	0.04475	100.85519
14	OVER	237	Flavobacterium	3.78213	1.9192	7.7385E-4	0.0243	105.58031
15	UNDER	47251	Leptolyngbya	-9.19911	-3.20149	7.6573E-4	0.0243	106.20619
16	UNDER	373984	Rivularia <cyanobacteria>	-489.41386	-8.93491	9.5855E-4	0.02759	262.18943
17	UNDER	265	Paracoccus <a-proteobacteria>	-3.50253	-1.8084	0.00109	0.02776	87.26053
18	UNDER	33057	Thauera	-9.50585	-3.24882	7.6193E-5	0.00599	380.60068
19	OVER	13687	Sphingomonas	3.9659	1.98765	6.0477E-4	0.02216	121.17894
20	UNDER	748770	Dolichospermum	-629.66438	-9.29844	4.5505E-5	0.00488	504.87372
21	UNDER	669357	Geminocystis	-628.24167	-9.29518	2.9286E-4	0.0118	639.51686
22	UNDER	332248	Truepera	-8.25461	-3.0452	0.00231	0.04002	56.923
23	UNDER	119541	Rhodobaca	-16.17924	-4.01607	0.00371	0.04667	43.28426
24	UNDER	159191	Nodularia <cyanobacteria>	-70323.25189	-16.10171	4.7530E-4	0.01824	444.54671
25	UNDER	170610	Halomicronema	-18.94692	-4.24389	0.00137	0.03059	76.47561
26	UNDER	110103	Anabaenopsis	-1891.86143	-10.88559	2.6749E-4	0.0118	684.51957
27	OVER	352450	Simplicispira	4.55318	2.18687	0.0012	0.02776	82.64848

Figure 10. OTU Differential Abundance Testing results.

While this project is opened, different actions can be carried out from the **sidepanel**:

Summary Report

Creates an HTML report which can be saved in PDF with the main results of the Differential Abundance Testing: parameters used for the test, number of differentially abundant OTUs, experimental design, ... (figure 11(see page 263)).



OTU Differential Abundance Testing Report

Dataset: taxonomy_classification ODAT

Overview

- Total features: **966**

Results

Differentially abundant (DA) features (FDR < 0.05): **188**

Over-represented (logFC > 1): **86**

Under-represented (logFC < -1) **: 98**

Experimental Design

Sample	Lib. size (pre-filter)	Lib. size (post-filter)	Norm. factor	Location	Time
PAB	16,916	16,916	1.830	Preta	Afternoon
PMB	20,811	20,811	1.509	Preta	Morning
VAB	13,768	13,768	0.591	Verde	Afternoon
VMB	17,482	17,482	0.613	Verde	Morning

Analysis Parameters

Parameter	Value
Experimental Design	exp_design.txt
Test at Specific Level	Genus
GLM Test	GLM Likelihood Ratio Test

Figure 11. OTU Differential Abundance Testing summary report.

Summary Chart

Shows a bar chart with the main results: OTUs pre and post-filtering steps, OTUs which are considered as differentially abundant and the over-/underrepresented ones (figure 12(see page 264)).

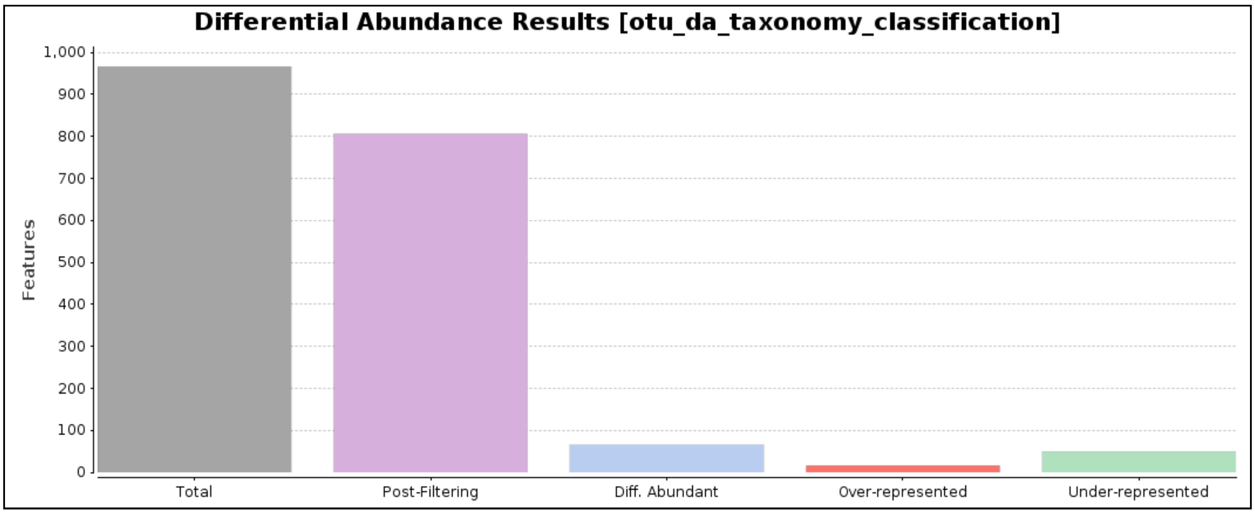
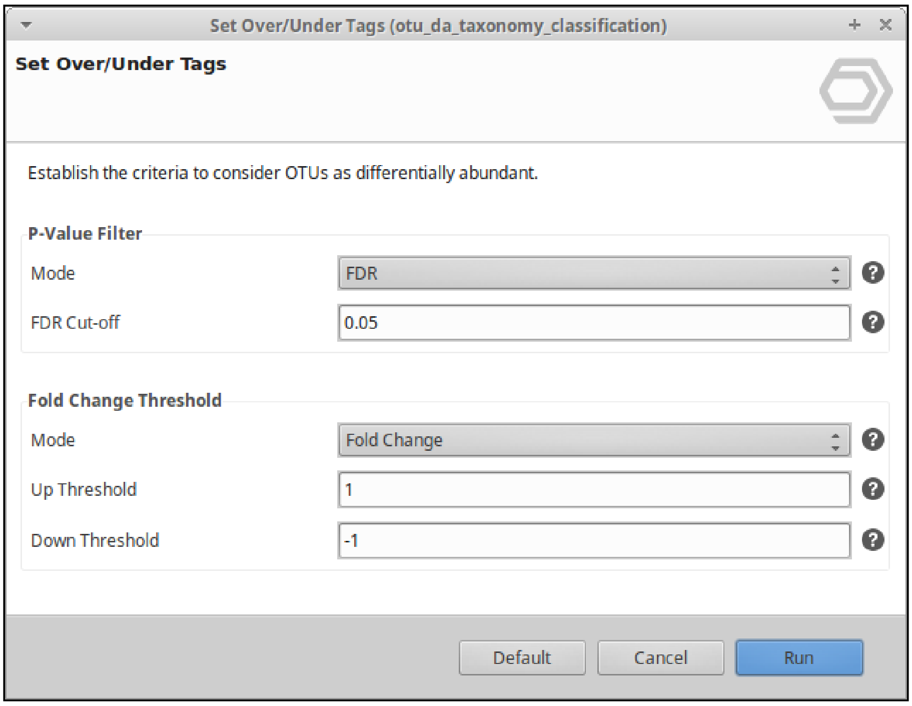


Figure 12. OTU Differential Abundance Testing summary chart.

Set Over/Under Tags

Establish a new FDR and Fold Change cutoff to consider OTUs as differentially abundant. FDR < 0.05 and logFC < -1 or logFC > 1 are set as default (figure 13(see page 265)).



Establish the criteria to consider OTUs as differentially abundant.

P-Value Filter

Mode: FDR

FDR Cut-off: 0.05

Fold Change Threshold

Mode: Fold Change

Up Threshold: 1

Down Threshold: -1

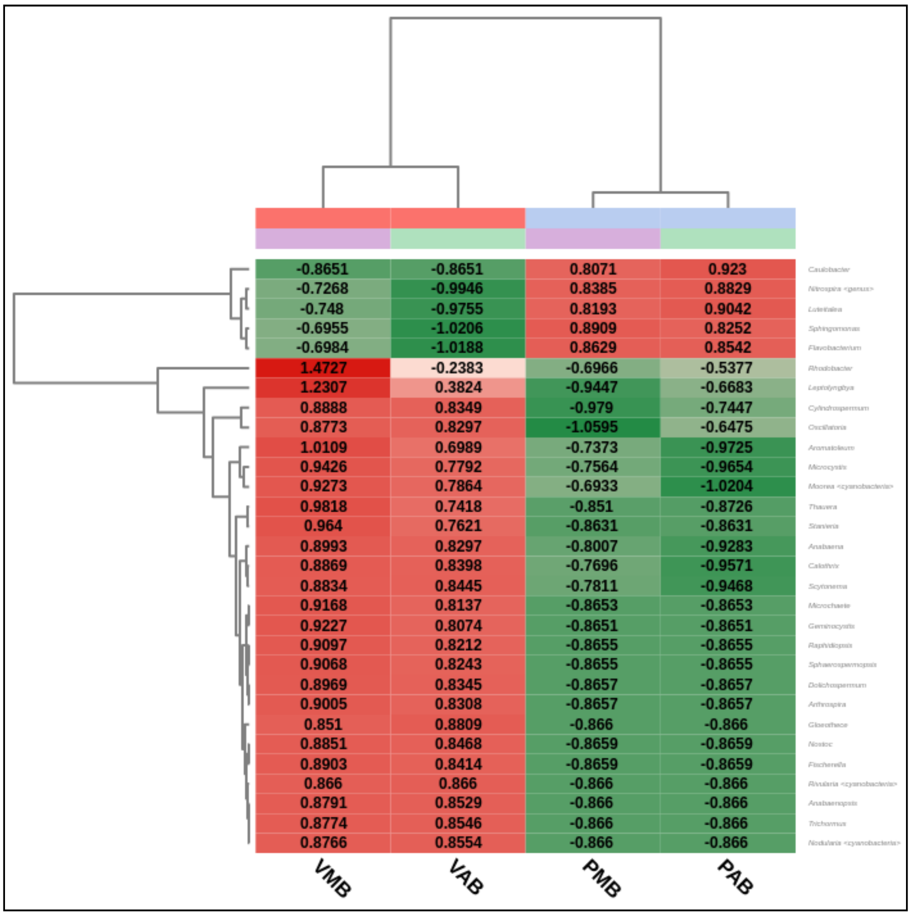
Figure 13. Set Over/Under Tags.

Heatmap

Shows a two-dimensional heatmap in which the abundance values are represented by ranges of colors (figure 14(see page 265)). The dendrograms added to the left and top side are produced by a hierarchical clustering method that takes as input the Euclidean distance computed between OTUs (left) and samples (top).

The upper bars show the experimental conditions of the study (columns) and the OTUs names are shown at the right of each row.

You can select if you want to draw the heatmap with the raw counts or with the CPM values, and if any transformation is necessary (logarithm in base 2, Z-score or both).



	VMB	VAB	PMB	PAB
Caulobacter	-0.8651	-0.8651	0.8071	0.923
Nitrospira <genus>	-0.7268	-0.9946	0.8385	0.8829
Luteitalea	-0.748	-0.9755	0.8193	0.9042
Sphingomonas	-0.6955	-1.0206	0.8909	0.8252
Flavobacterium	-0.6984	-1.0188	0.8629	0.8542
Rhodobacter	1.4727	-0.2383	-0.6966	-0.5377
Leptolyngbya	1.2307	0.3824	-0.9447	-0.6683
Cylindrospermum	0.8888	0.8349	-0.979	-0.7447
Oscillatoria	0.8773	0.8297	-1.0595	-0.6475
Aromatoleum	1.0109	0.6989	-0.7373	-0.9725
Microcystis	0.9426	0.7792	-0.7564	-0.9654
Moorea <cyanobacteria>	0.9273	0.7864	-0.6933	-1.0204
Thauera	0.9818	0.7418	-0.851	-0.8726
Stanieria	0.964	0.7621	-0.8631	-0.8631
Anabaena	0.8993	0.8297	-0.8007	-0.9283
Calothrix	0.8869	0.8398	-0.7696	-0.9571
Scytonema	0.8834	0.8445	-0.7811	-0.9468
Microchaete	0.9168	0.8137	-0.8653	-0.8653
Geminocystis	0.9227	0.8074	-0.8651	-0.8651
Raphidiopsis	0.9097	0.8212	-0.8655	-0.8655
Sphaerospermopsis	0.9068	0.8243	-0.8655	-0.8655
Dolichospermum	0.8969	0.8345	-0.8657	-0.8657
Arthrospira	0.9005	0.8308	-0.8657	-0.8657
Gloeothece	0.851	0.8809	-0.866	-0.866
Nostoc	0.8851	0.8468	-0.8659	-0.8659
Fischerella	0.8903	0.8414	-0.8659	-0.8659
Rivularia <cyanobacteria>	0.866	0.866	-0.866	-0.866
Anabaenopsis	0.8791	0.8529	-0.866	-0.866
Trichormus	0.8774	0.8546	-0.866	-0.866
Nodularia <cyanobacteria>	0.8766	0.8554	-0.866	-0.866

Figure 14. Heatmap.

Robinson MD, McCarthy DJ and Smyth GK (2010). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." Bioinformatics, 26, pp. -1.

11 General Tools

Content of this section

OmicsBox offers General Tools to handle FastQ files as well as visualizations.

1. **Tag Statistics**: Summarizes the project tags in a bar chart.
2. **Venn Diagram**: Shows all possible logical relations between a finite collection of different sets.
3. **FastQ Tools**:
    a. FastQ Quality Check: Performs quality control checks of FastQ files.
    b. FastQ Preprocessing: Filter contamination sequences and adapters to obtain high-quality FastQ files.

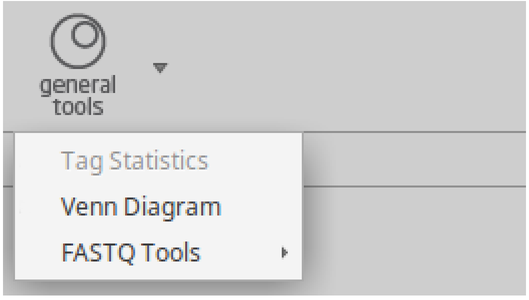


Figure 1: General Tools menu

FastQ and Venn Diagram Example Datasets: Download[153].

11.1 Tag Statistics

Tag Statistics tool allows you to know the current status of the entire project by showing a distribution of the Tags. This feature can be found under **General Tools** → **Tag Statistics,** and requires an opened OmicsBox project. After clicking, a Tag Distribution chart will open (Figure 1).

[153] https://resources.biobam.com/omicsbox/example_data/General.zip

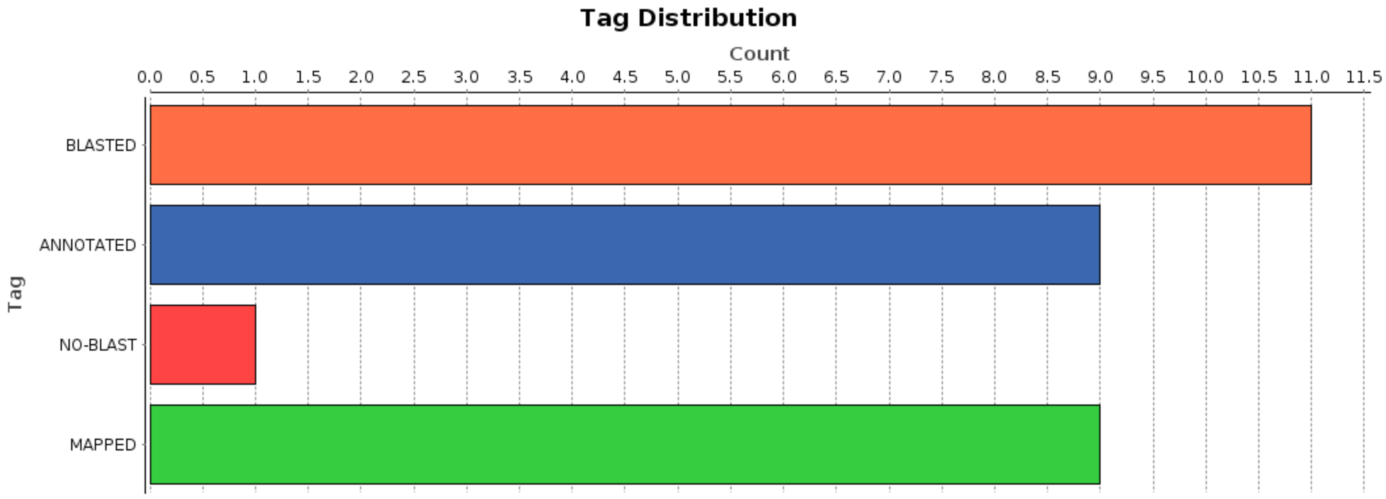


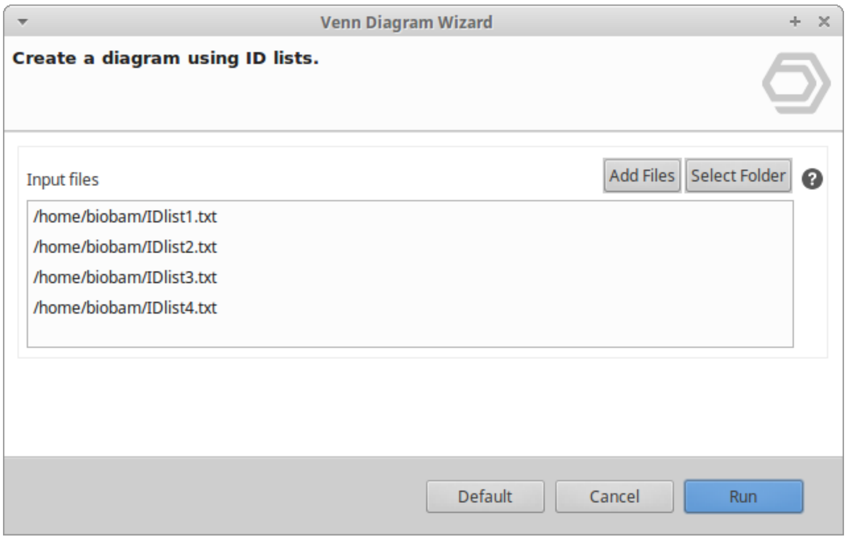
Figure 1: Tag Statistics chart from an annotation project.

Some types of projects can be represented in this way, i.e. a functional annotation project, a Rfam Project, an Enrichment Analysis Project, etc.

11.2 Venn Diagram

Venn Diagram tool allows you to select multiple ID List or ID value list in text or BOX/B2G format and draw the intersection of the elements of the lists.

This functionality can be found under **General Tools** → **Venn Diagram.** The wizard allows to select input files, you can mix the supported types (**ID List** or **ID value list**), with different formats (**Plain text** or **B2G** or **BOX**).



Create a diagram using ID lists.

Input files

/home/biobam/IDlist1.txt
/home/biobam/IDlist2.txt
/home/biobam/IDlist3.txt
/home/biobam/IDlist4.txt

Figure 1: Dialog to add lists to generate the Venn Diagram

After selecting the files just press the **Run** button. A new tab will appear with the Venn diagram.

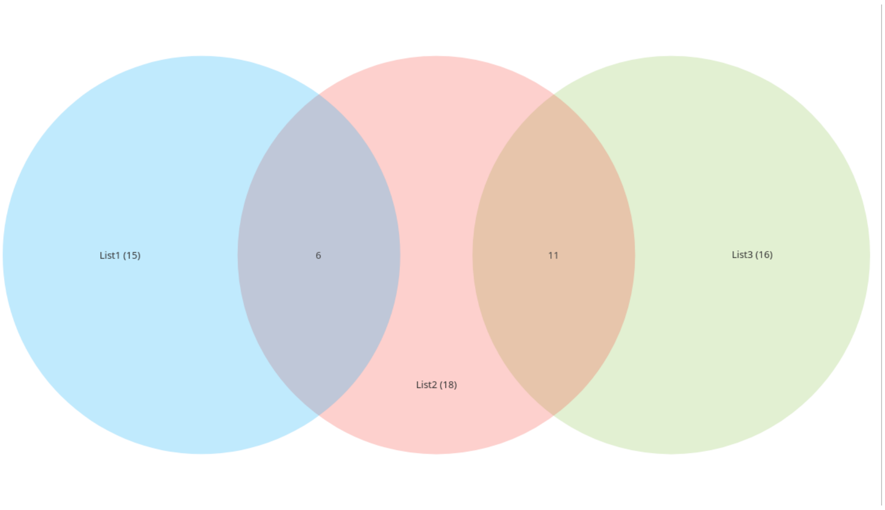
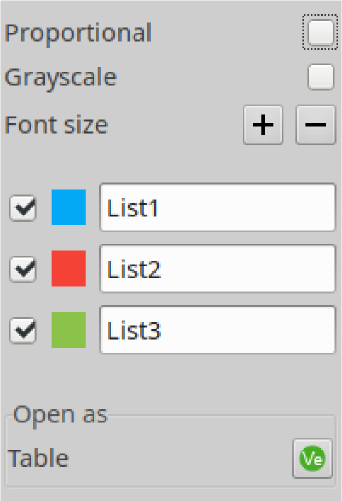


Figure 2: Venn Diagram of 3 lists

There are different options to customize Venn visualization.

- **Proportional**. This check box allows you to change how the size of the circles is calculated, by setting this option to **true** will paint the size of the circles proportionally to the number of elements the list contains. When **false** all the circles will use the same size.
- **Grayscale**. If **true**, all circles will be painted in different shades of grey. Set to **false** to use a normal color.
- **Font size**. Use the **plus** or **minus** icon to **increase** or **decrease** the font size.

- **List control**. There will be one for each list you load with the wizard. Each control customizes each list individually.
    - The **Check box** hides or shows the circle represented for this list.
    - The **Color box** allows changing the color of the circle.
    - The **Text field** updates the list name.



Proportional ☐

Grayscale ☐

Font size + −

☑ 🟦 List1

☑ 🟥 List2

☑ 🟩 List3

Open as

Table

Figure 3: Configuration panel

- **Table button.** This will open a table where rows are the union of all lists. The column tag indicates in which lists this element appears. You can use this table to sort and filter the elements and extract the selected rows to a new filtered Venn Diagram.

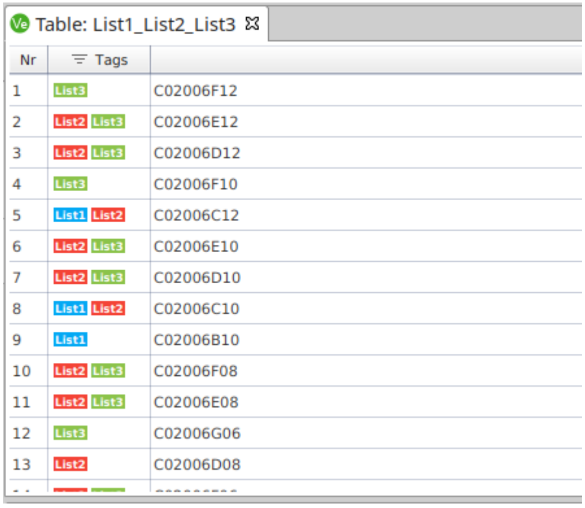


Table: List1_List2_List3

Nr	Tags	
1	List3	C02006F12
2	List2 List3	C02006E12
3	List2 List3	C02006D12
4	List3	C02006F10
5	List1 List2	C02006C12
6	List2 List3	C02006E10
7	List2 List3	C02006D10
8	List1 List2	C02006C10
9	List1	C02006B10
10	List2 List3	C02006F08
11	List2 List3	C02006E08
12	List3	C02006G06
13	List2	C02006D08

Figure 4: Table with the list and the Tags

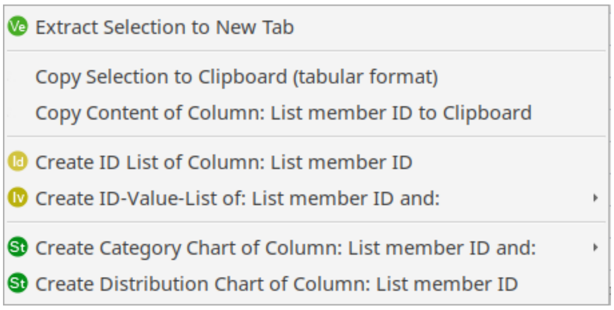

Extract Selection to New Tab
Copy Selection to Clipboard (tabular format)
Copy Content of Column: List member ID to Clipboard
Create ID List of Column: List member ID
Create ID-Value-List of: List member ID and:
Create Category Chart of Column: List member ID and:
Create Distribution Chart of Column: List member ID

Figure 5: Extract Selected rows from the table

For further details, please link here[154].

11.3 FASTQ Quality Check

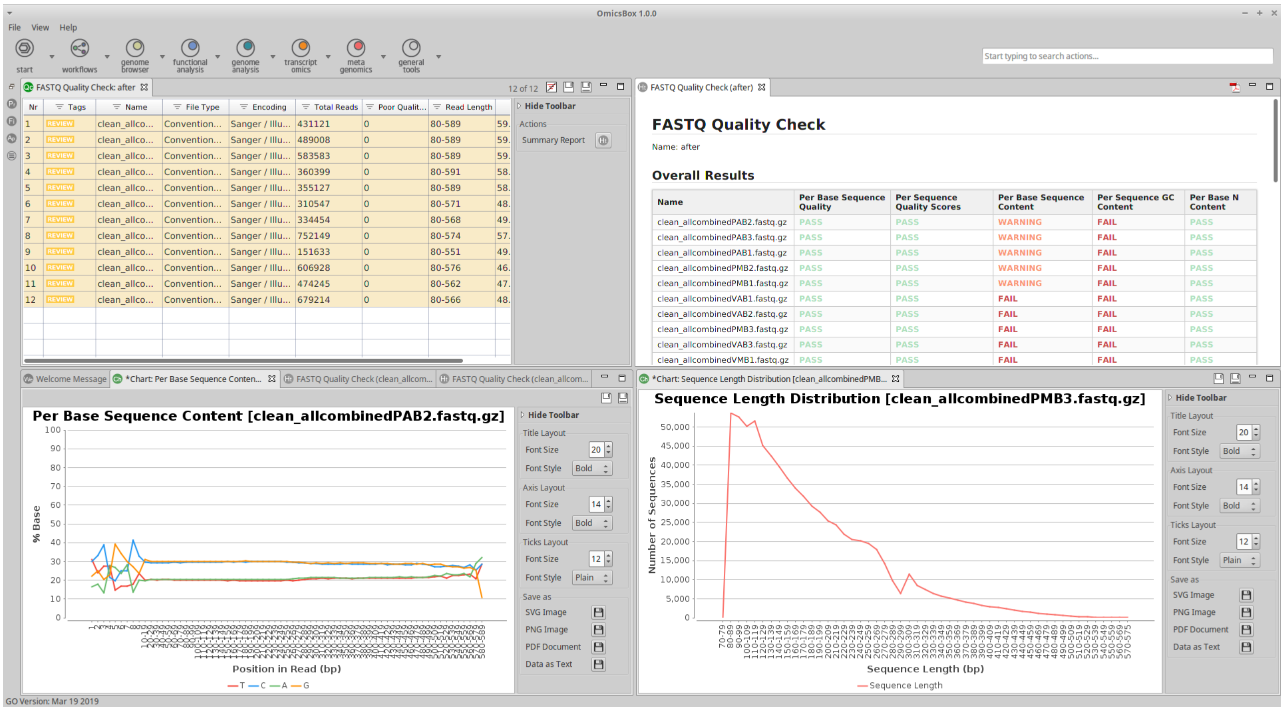
Content of this page:

11.3.1 Introduction

The "FASTQ Quality Check" tool provides an easy way to perform a quality control check on sequence data coming from high throughput sequencing pipelines. The analysis is performed by nine modules which provide a quick overview of whether the data looks good and there are no problems or biases which may affect downstream analysis. Results and evaluations are returned in the form of charts and tables.

This tool is based on the popular FastQC software[155]. Please cite FastQC as:
Andrews S (2010)."FastQC: a quality control tool for high throughput sequence data". Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

[154] https://www.biobam.com/venndiagram/
[155] https://www.bioinformatics.babraham.ac.uk/projects/fastqc/


OmicsBox 1.0.0
File View Help
start
workflows
genome browser
functional analysis
genome analysis
transcript omics
meta genomics
general tools
Start typing to search actions...
FASTQ Quality Check: after
12 of 12
Hide Toolbar
Actions
Summary Report
FASTQ Quality Check (after)
Welcome Message
*Chart: Per Base Sequence Conten...
FASTQ Quality Check (clean_allcom...
FASTQ Quality Check (clean_allcom...
*Chart: Sequence Length Distribution [clean_allcombinedPMB...
Per Base Sequence Content [clean_allcombinedPAB2.fastq.gz]
% Base
Position in Read (bp)
T C A G
Sequence Length Distribution [clean_allcombinedPMB3.fastq.gz]
Number of Sequences
Sequence Length (bp)
Sequence Length
GO Version: Mar 19 2019

Nr	Tags	Name	File Type	Encoding	Total Reads	Poor Qualit...	Read Length	
1	REVIEW	clean_allco...	Convention...	Sanger / Illu...	431121	0	80-589	59.
2	REVIEW	clean_allco...	Convention...	Sanger / Illu...	489008	0	80-589	59.
3	REVIEW	clean_allco...	Convention...	Sanger / Illu...	583583	0	80-589	59.
4	REVIEW	clean_allco...	Convention...	Sanger / Illu...	360399	0	80-591	58.
5	REVIEW	clean_allco...	Convention...	Sanger / Illu...	355127	0	80-589	58.
6	REVIEW	clean_allco...	Convention...	Sanger / Illu...	310547	0	80-571	48.
7	REVIEW	clean_allco...	Convention...	Sanger / Illu...	334454	0	80-568	49.
8	REVIEW	clean_allco...	Convention...	Sanger / Illu...	752149	0	80-574	57.
9	REVIEW	clean_allco...	Convention...	Sanger / Illu...	151633	0	80-551	49.
10	REVIEW	clean_allco...	Convention...	Sanger / Illu...	606928	0	80-576	46.
11	REVIEW	clean_allco...	Convention...	Sanger / Illu...	474245	0	80-562	47.
12	REVIEW	clean_allco...	Convention...	Sanger / Illu...	679214	0	80-566	48.

FASTQ Quality Check

Name: after

Overall Results

Name	Per Base Sequence Quality	Per Sequence Quality Scores	Per Base Sequence Content	Per Sequence GC Content	Per Base N Content
clean_allcombinedPAB2.fastq.gz	PASS	PASS	WARNING	FAIL	PASS
clean_allcombinedPAB3.fastq.gz	PASS	PASS	WARNING	FAIL	PASS
clean_allcombinedPAB1.fastq.gz	PASS	PASS	WARNING	FAIL	PASS
clean_allcombinedPMB2.fastq.gz	PASS	PASS	WARNING	FAIL	PASS
clean_allcombinedPMB1.fastq.gz	PASS	PASS	WARNING	FAIL	PASS
clean_allcombinedVAB1.fastq.gz	PASS	PASS	FAIL	FAIL	PASS
clean_allcombinedVAB2.fastq.gz	PASS	PASS	FAIL	FAIL	PASS
clean_allcombinedPMB3.fastq.gz	PASS	PASS	FAIL	FAIL	PASS
clean_allcombinedVAB3.fastq.gz	PASS	PASS	FAIL	FAIL	PASS
clean_allcombinedVMB1.fastq.gz	PASS	PASS	FAIL	FAIL	PASS

Title Layout: Font Size 20, Font Style Bold
Axis Layout: Font Size 14, Font Style Bold
Ticks Layout: Font Size 12, Font Style Plain
Save as: SVG Image, PNG Image, PDF Document, Data as Text

Figure 1: FASTQ Quality Check Interface

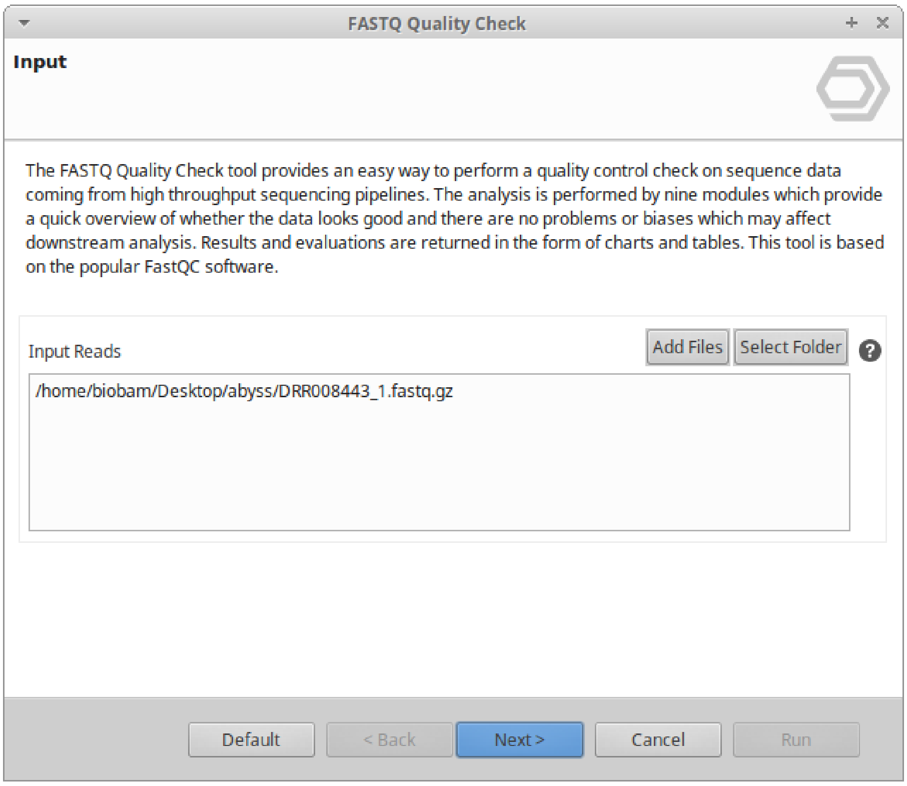
11.3.2 Run FASTQ Quality Check

This functionality can be found under **General Tools** → **FASTQ Tools** → **FASTQ Quality Check**. The wizard allows to select input files and adjust analysis parameters (Figure 2(see page 273) and Figure 3(see page 273)).

- **Raw Sequence Data:** Select the files containing the sequence data. These files are assumed to be in FASTQ format (or compressed in gzip format).
- **Additional Adapter Sequences:** This option allows specifying a file that contains the list of adapter sequences which will be explicitly searched against the library. The file must contain sets of named adapters in the form of "Name <Tab> Sequence". If this option is not set, OmicsBox searches for the following adapter sequences:
  - Illumina Universal Adapter: AGATCGGAAGAG
  - Illumina Small RNA 3' Adapter: TGGAATTCTCGG
  - Illumina Small RNA 5' Adapter: GATCGTCGGACT
  - Nextera Transposase Sequence: CTGTCTCTTATA
  - SOLID Small RNA Adapter: CGCCTTGGCCGT
- **Additional Contaminant Sequences:** This option allows specifying a file that contains the list of contaminants to screen over-represented sequences against. The file must contain sets of named contaminants in the form of "Name <Tab> Sequence". If this option is not set, OmicsBox searches for a list of common contaminant sequences[156].
- **Chart Read Length Binning:** Enable grouping of bases for reads. If not, reports will show data for every base in the read.

[156] https://www.blast2go.com/images/b2g_blog/contaminant_list.txt

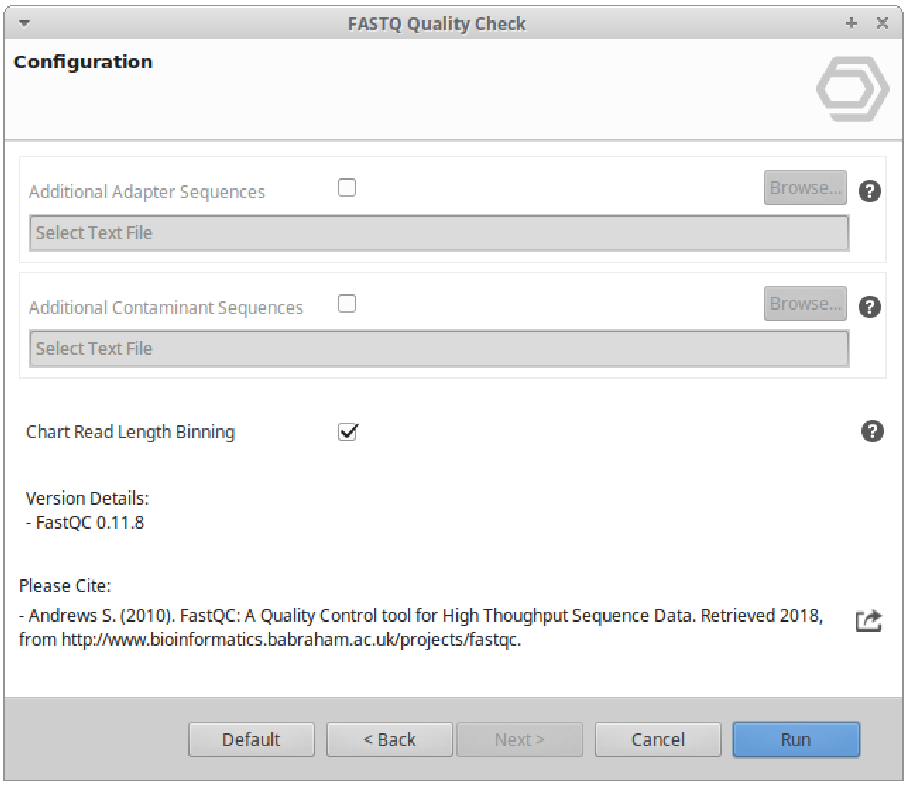
Disabling this option on long reads (> 50 bp) can cause that the plots look very small.



The FASTQ Quality Check tool provides an easy way to perform a quality control check on sequence data coming from high throughput sequencing pipelines. The analysis is performed by nine modules which provide a quick overview of whether the data looks good and there are no problems or biases which may affect downstream analysis. Results and evaluations are returned in the form of charts and tables. This tool is based on the popular FastQC software.

Input Reads: /home/biobam/Desktop/abyss/DRR008443_1.fastq.gz

Figure 2: FASTQ Quality Check Input Page



Additional Adapter Sequences □ Browse...
Select Text File

Additional Contaminant Sequences □ Browse...
Select Text File

Chart Read Length Binning ☑

Version Details:
- FastQC 0.11.8

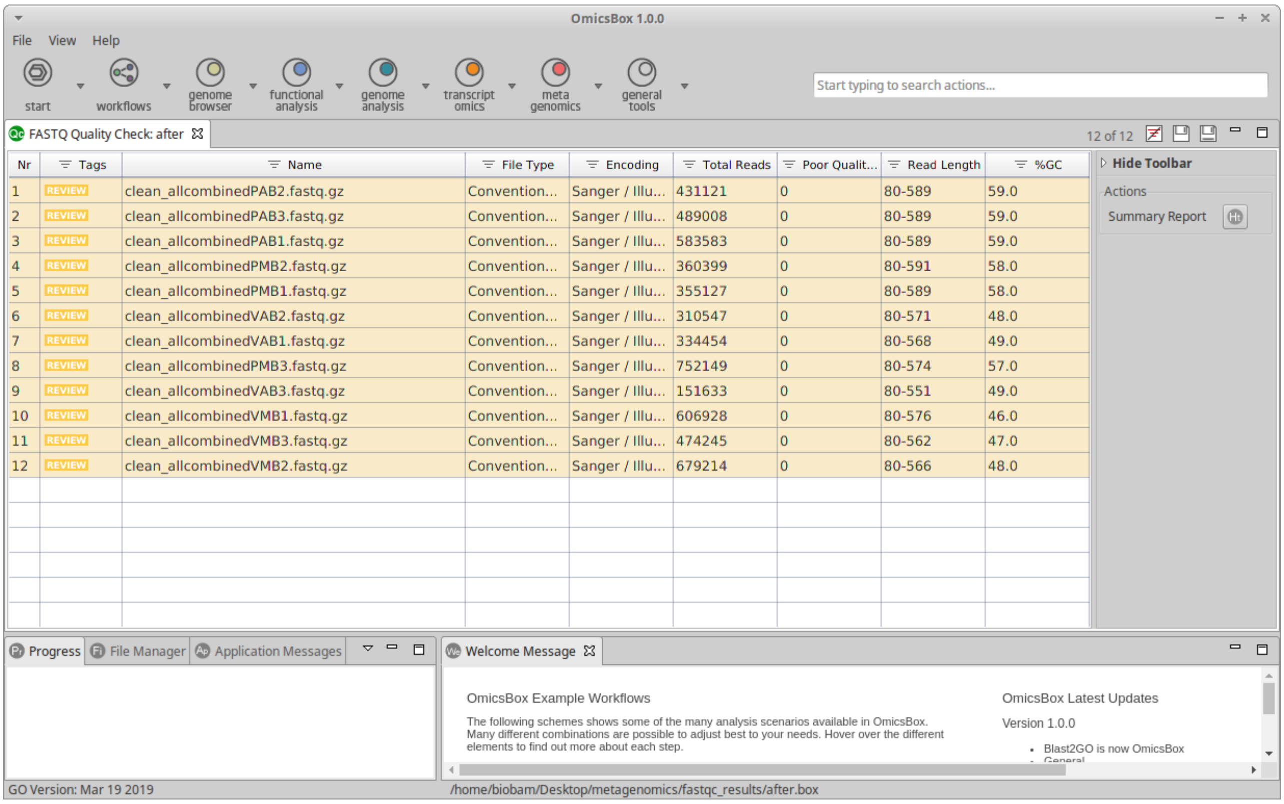
Please Cite:
- Andrews S. (2010). FastQC: A Quality Control tool for High Thoughput Sequence Data. Retrieved 2018, from http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Figure 3: FASTQ Quality Check Configuration Page

11.3.3 Results

Once finished, a new tab is opened containing a simple composition statistics of each analyzed file (Figure 4(see page 274)). Each row corresponds to an input file, and columns show the following information:

- Name: The name of the file which was analyzed.
- File type: Shows whether the file appeared to contain actual base calls or colorspace data which had to be converted to base calls.
- Encoding: Shows the ASCII encoding of quality values was detected in this file.
- Total Sequences: The total number of read sequences processed.
- Poor quality reads: Sequences flagged as poor quality reads.
- Sequence Length: Provides the length of the shortest and longest sequence in the set. If all sequences are the same length only one value is reported.
- %GC: The overall %GC of all bases in all sequences.



Nr	Tags	Name	File Type	Encoding	Total Reads	Poor Qualit...	Read Length	%GC
1	REVIEW	clean_allcombinedPAB2.fastq.gz	Convention...	Sanger / Illu...	431121	0	80-589	59.0
2	REVIEW	clean_allcombinedPAB3.fastq.gz	Convention...	Sanger / Illu...	489008	0	80-589	59.0
3	REVIEW	clean_allcombinedPAB1.fastq.gz	Convention...	Sanger / Illu...	583583	0	80-589	59.0
4	REVIEW	clean_allcombinedPMB2.fastq.gz	Convention...	Sanger / Illu...	360399	0	80-591	58.0
5	REVIEW	clean_allcombinedPMB1.fastq.gz	Convention...	Sanger / Illu...	355127	0	80-589	58.0
6	REVIEW	clean_allcombinedVAB2.fastq.gz	Convention...	Sanger / Illu...	310547	0	80-571	48.0
7	REVIEW	clean_allcombinedVAB1.fastq.gz	Convention...	Sanger / Illu...	334454	0	80-568	49.0
8	REVIEW	clean_allcombinedPMB3.fastq.gz	Convention...	Sanger / Illu...	752149	0	80-574	57.0
9	REVIEW	clean_allcombinedVAB3.fastq.gz	Convention...	Sanger / Illu...	151633	0	80-551	49.0
10	REVIEW	clean_allcombinedVMB1.fastq.gz	Convention...	Sanger / Illu...	606928	0	80-576	46.0
11	REVIEW	clean_allcombinedVMB3.fastq.gz	Convention...	Sanger / Illu...	474245	0	80-562	47.0
12	REVIEW	clean_allcombinedVMB2.fastq.gz	Convention...	Sanger / Illu...	679214	0	80-566	48.0

OmicsBox Example Workflows

The following schemes shows some of the many analysis scenarios available in OmicsBox. Many different combinations are possible to adjust best to your needs. Hover over the different elements to find out more about each step.

OmicsBox Latest Updates

Version 1.0.0

- Blast2GO is now OmicsBox
- General

GO Version: Mar 19 2019

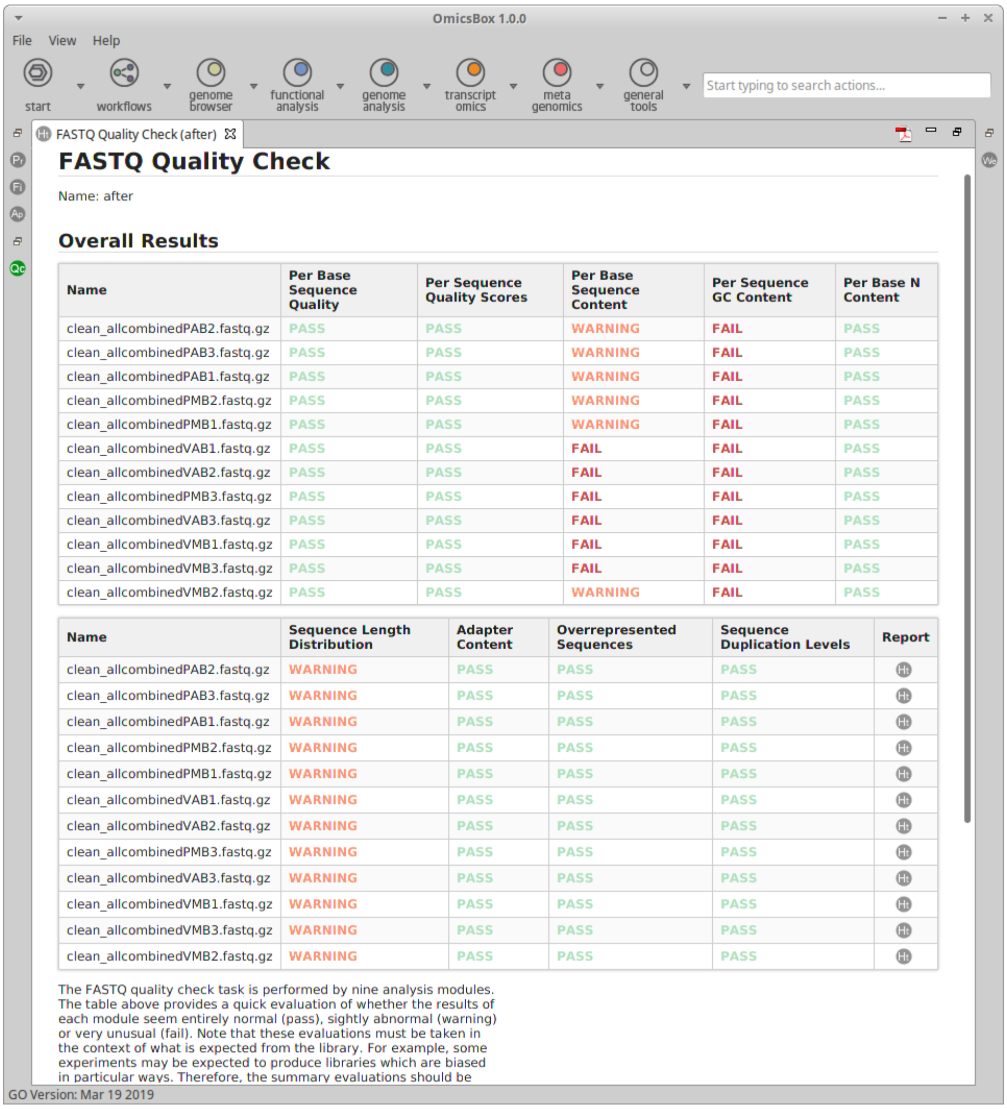
/home/biobam/Desktop/metagenomics/fastqc_results/after.box

Figure 4: FASTQ Quality Check Project

Furthermore, a result page will show a summary of the "FASTQ Quality Check" results (Figure 5(see page 275)). This page provides a quick evaluation of whether the results of each module seem entirely normal (pass), slightly abnormal (warning) or very unusual (fail).

⚠ Note that these evaluations must be taken in the context of what is expected from each library. For example, some experiments may be expected to produce libraries which are biased in particular ways. Therefore, the summary evaluations should be treated as pointers that guide the preprocessing of the libraries.

The result summary can be generated via **Side Panel** → **Summary Report.** Additionally, the report of each file can be opened by clicking on the button of the column "Report".



FASTQ Quality Check

Name: after

Overall Results

Name	Per Base Sequence Quality	Per Sequence Quality Scores	Per Base Sequence Content	Per Sequence GC Content	Per Base N Content
clean_allcombinedPAB2.fastq.gz	PASS	PASS	WARNING	FAIL	PASS
clean_allcombinedPAB3.fastq.gz	PASS	PASS	WARNING	FAIL	PASS
clean_allcombinedPAB1.fastq.gz	PASS	PASS	WARNING	FAIL	PASS
clean_allcombinedPMB2.fastq.gz	PASS	PASS	WARNING	FAIL	PASS
clean_allcombinedPMB1.fastq.gz	PASS	PASS	WARNING	FAIL	PASS
clean_allcombinedVAB1.fastq.gz	PASS	PASS	FAIL	FAIL	PASS
clean_allcombinedVAB2.fastq.gz	PASS	PASS	FAIL	FAIL	PASS
clean_allcombinedPMB3.fastq.gz	PASS	PASS	FAIL	FAIL	PASS
clean_allcombinedVAB3.fastq.gz	PASS	PASS	FAIL	FAIL	PASS
clean_allcombinedVMB1.fastq.gz	PASS	PASS	FAIL	FAIL	PASS
clean_allcombinedVMB3.fastq.gz	PASS	PASS	FAIL	FAIL	PASS
clean_allcombinedVMB2.fastq.gz	PASS	PASS	WARNING	FAIL	PASS

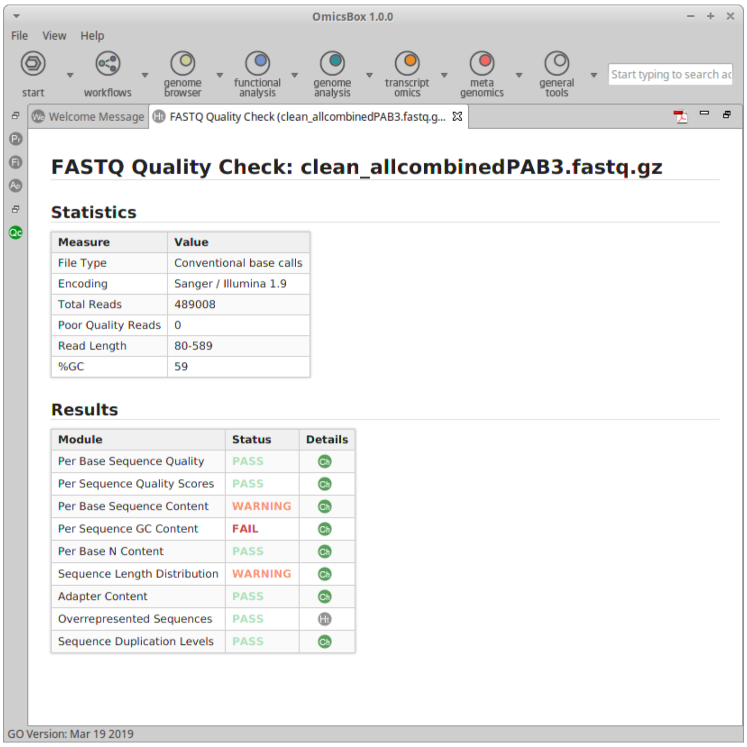
Name	Sequence Length Distribution	Adapter Content	Overrepresented Sequences	Sequence Duplication Levels	Report
clean_allcombinedPAB2.fastq.gz	WARNING	PASS	PASS	PASS	⊕
clean_allcombinedPAB3.fastq.gz	WARNING	PASS	PASS	PASS	⊕
clean_allcombinedPAB1.fastq.gz	WARNING	PASS	PASS	PASS	⊕
clean_allcombinedPMB2.fastq.gz	WARNING	PASS	PASS	PASS	⊕
clean_allcombinedPMB1.fastq.gz	WARNING	PASS	PASS	PASS	⊕
clean_allcombinedVAB1.fastq.gz	WARNING	PASS	PASS	PASS	⊕
clean_allcombinedVAB2.fastq.gz	WARNING	PASS	PASS	PASS	⊕
clean_allcombinedPMB3.fastq.gz	WARNING	PASS	PASS	PASS	⊕
clean_allcombinedVAB3.fastq.gz	WARNING	PASS	PASS	PASS	⊕
clean_allcombinedVMB1.fastq.gz	WARNING	PASS	PASS	PASS	⊕
clean_allcombinedVMB3.fastq.gz	WARNING	PASS	PASS	PASS	⊕
clean_allcombinedVMB2.fastq.gz	WARNING	PASS	PASS	PASS	⊕

The FASTQ quality check task is performed by nine analysis modules. The table above provides a quick evaluation of whether the results of each module seem entirely normal (pass), slightly abnormal (warning) or very unusual (fail). Note that these evaluations must be taken in the context of what is expected from the library. For example, some experiments may be expected to produce libraries which are biased in particular ways. Therefore, the summary evaluations should be

Figure 5: FASTQ Quality Check Report

The results of each module for each file can be accessed as follows:

- To open the summary report of each file, right-click on a row and click on **Show report**. A new report is opened containing a summary of the statistics and results for the selected file (Figure 6(see page 276)).
- To open the result of each module for a file, right-click on a row and go to the **Show Statistics** submenu**.** These results also can be accessed by clicking on the buttons of the "Details" column of the results table.



FASTQ Quality Check: clean_allcombinedPAB3.fastq.gz

Statistics

Measure	Value
File Type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Reads	489008
Poor Quality Reads	0
Read Length	80-589
%GC	59

Results

Module	Status	Details
Per Base Sequence Quality	PASS	Ch
Per Sequence Quality Scores	PASS	Ch
Per Base Sequence Content	WARNING	Ch
Per Sequence GC Content	FAIL	Ch
Per Base N Content	PASS	Ch
Sequence Length Distribution	WARNING	Ch
Adapter Content	PASS	Ch
Overrepresented Sequences	PASS	Ht
Sequence Duplication Levels	PASS	Ch

Figure 6: Report of a FASTQ file

11.3.3.1 Per Base Sequence Quality

This chart shows an overview of the range of quality values across all bases at each position in the FASTQ file (Figure 7(see page 277)).

For each position (x-axis), a box and whisker type plot is drawn:

- The central black line is the median value.
- The yellow box represents the interquartile range (25-75%).
- The upper and lower whiskers represent the 10% and 90% points.
- The blue line represents the mean quality.

The y-axis shows the quality scores. The background of the graph divides the y-axis into very good quality calls (green), calls of reasonable quality (orange), and calls of poor quality (red).

The title of the graph will describe the encoding that the input files used.

A **WARNING** is issued if the lower quartile for any base is less than 10, or if the median for any base is less than 25. This module raises a **FAIL** if the lower quartile for any base is less than 5 or if the median for any base is less than 20.

The most common reason for warnings and failures is a general degradation of quality over the duration of long runs. If the quality of the library falls to a low level then the most common procedure is to perform a quality trimming to truncate reads based on their average quality.

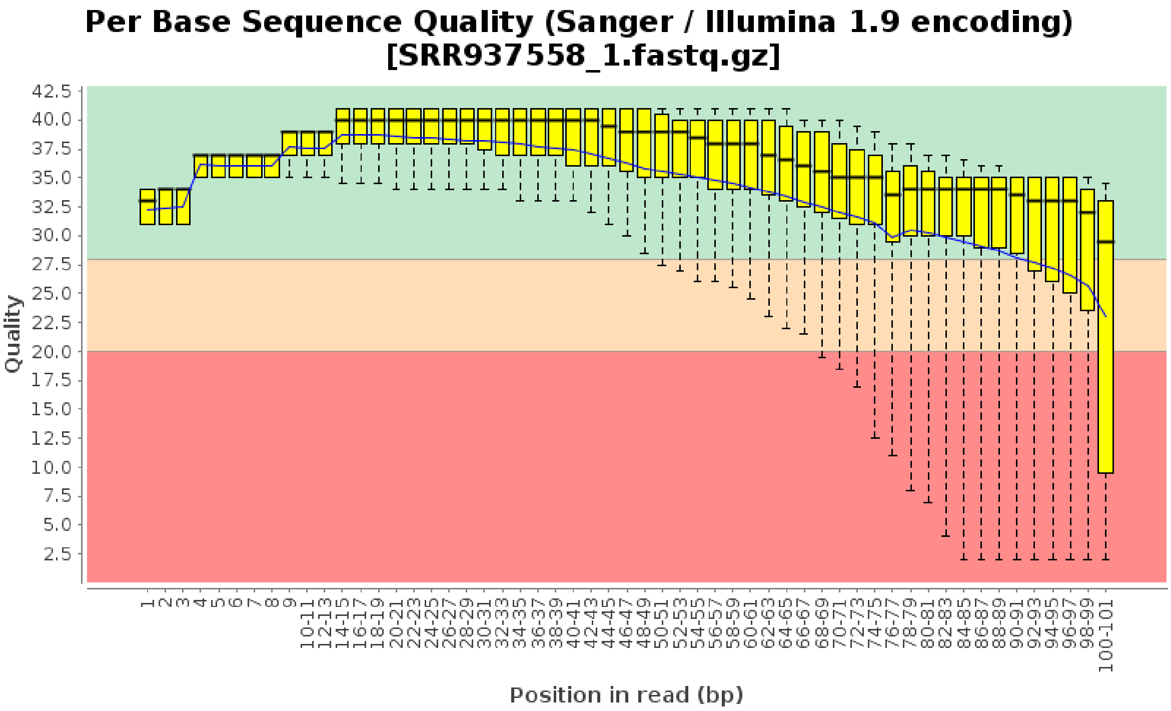


Figure 7: Per Base Sequence Quality Chart

11.3.3.2 Per Sequence Quality Scores

This chart displays the number of read sequences that have the same mean sequence quality (Figure 8(see page 278)). It allows seeing if a subset of your sequences has universally low-quality values.

A **WARNING** is raised if the most frequently observed mean quality is below 27 (0.2% error rate). A **FAIL** is raised if the most frequently observed mean quality is below 20 (1% error rate).

If a significant proportion of the reads in a run have overall low quality then this indicates some kind of systematic problem. This may be alleviated through quality trimming.

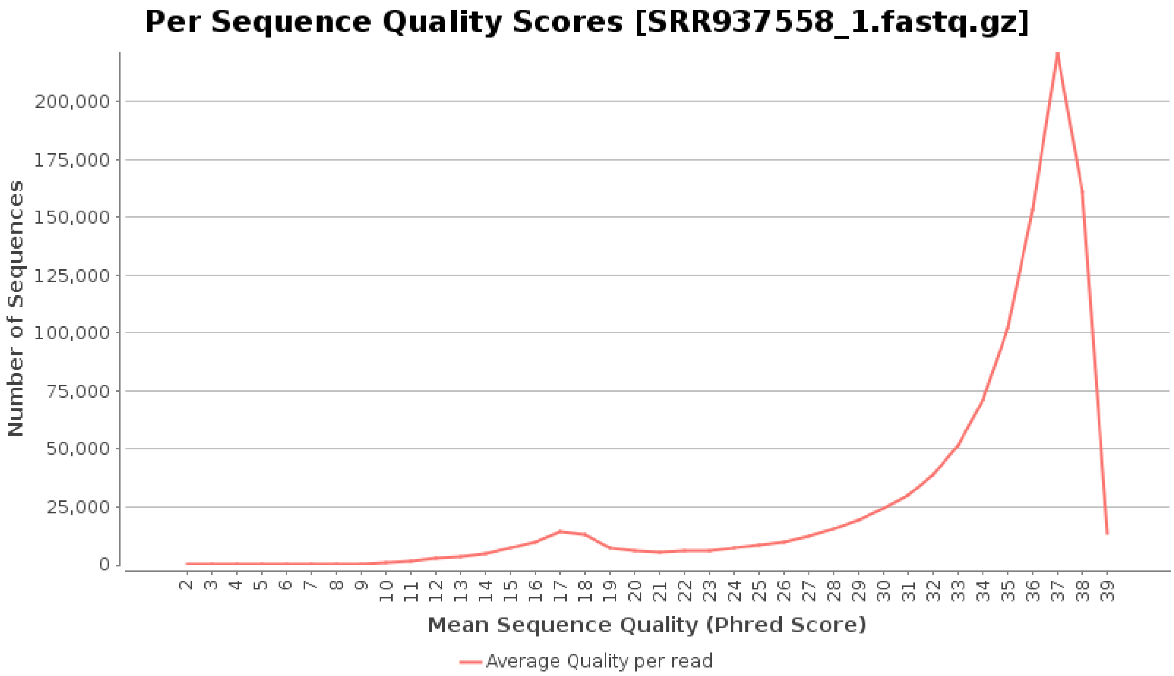


Figure 8: Per Sequence Quality Scores Chart

11.3.3.3 Per Base Sequence Content

This chart plots out the proportion of each base position in a FASTQ file for which each of the four normal DNA bases has been called (Figure 9(see page 279)). In a random library, it is expected that there would be little to no difference between the different bases of the sequence reads, so the lines in this plot should run parallel with each other.

A **WARNING** is issued if the difference between A and T, or G and C is greater than 10% in any position. A **F AIL** is raised if the difference between A and T, or G and C is greater than 20% in any position.

The common reasons for warnings and failures are:

- Overrepresented sequences (such as adapter dimers or rRNA in a sample).
- Biased fragmentation (nearly all RNA-Seq libraries will fail this module because of this bias).
- Biased composition libraries.
- If the library has been adapter trimmed.

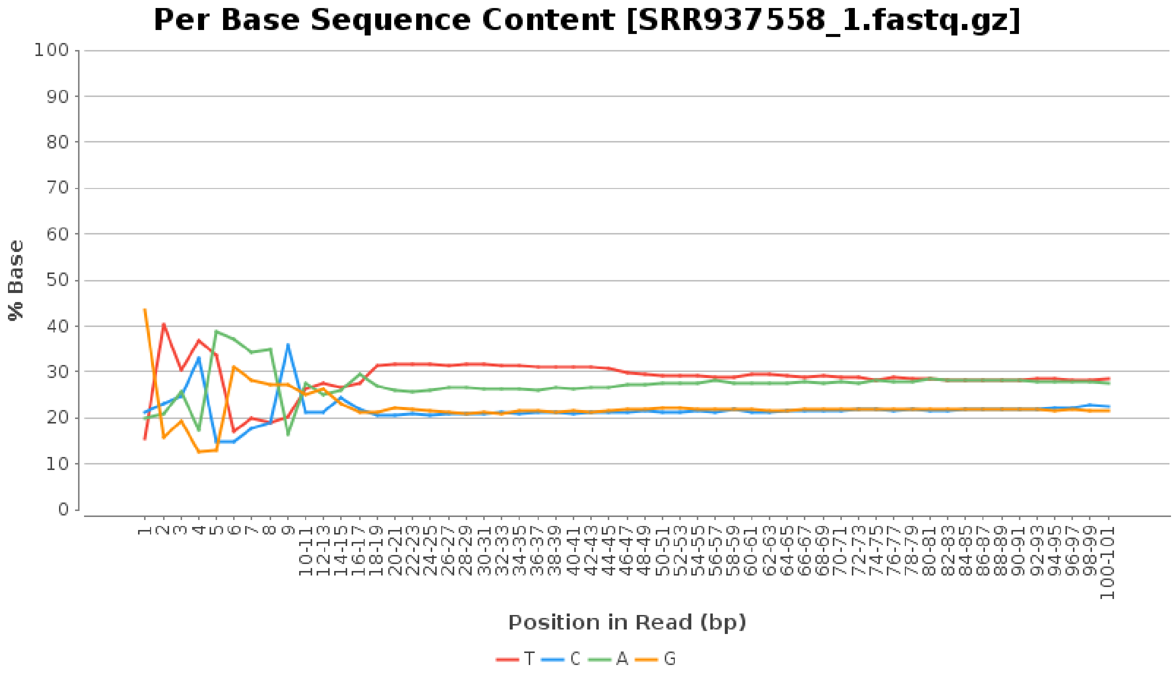


Figure 9: Per Base Sequence Content Chart

11.3.3.4 Per Sequence GC Content

This module measures the GC content across the whole length of each sequence read in a file and compares it to a modeled normal distribution of GC content (Figure 10(see page 280)). Since the GC content of the genome is not known, the modal GC content is calculated from the observed data and used to build a reference distribution.

A **WARNING** is raised if the sum of the deviations from the normal distribution represents more than 15% of the reads. A **FAIL** indicates that the sum of the deviations from the normal distribution represents more than 30% of the reads.

Warnings and failures indicate a problem with the library (e.g. specific contaminant). An unusually shaped distribution could indicate a contaminated library. A normal distribution which is shifted indicates some systematic bias which is independent of base position.

⚠ If there is a systematic bias which creates a shifted normal distribution then this won't be flagged as an error by the module since it doesn't know what the genome's GC content should be.

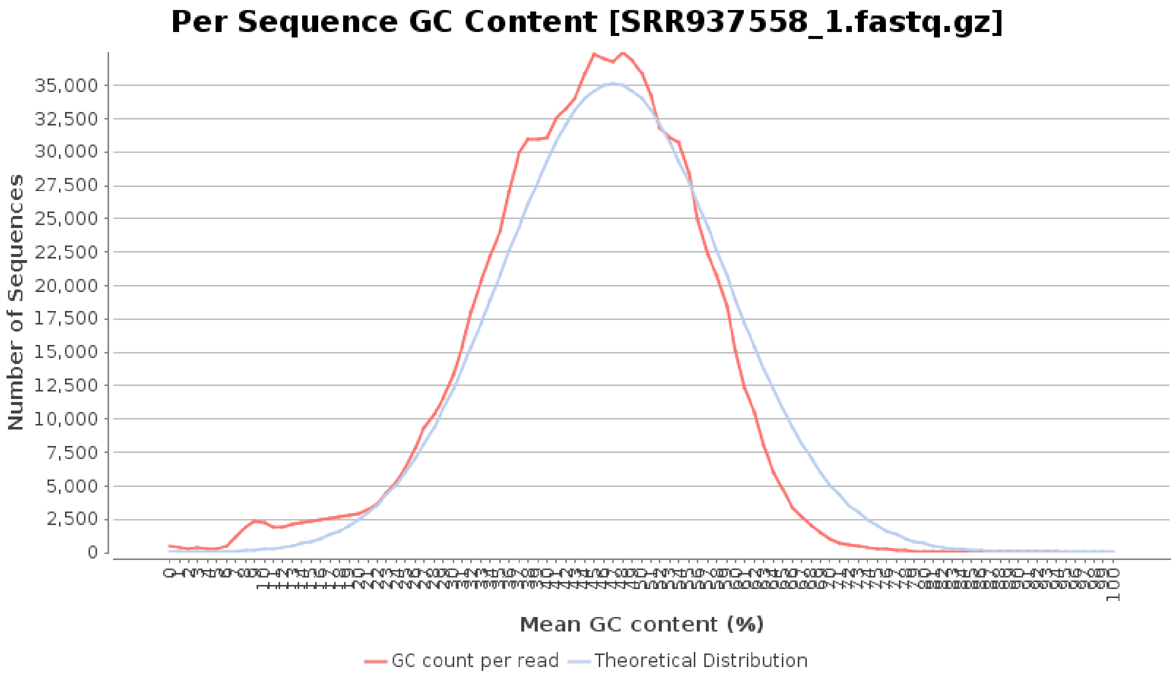


Figure 10: Per Sequence GC Content Chart

11.3.3.5 Per Base N Content

This module plots out the percentage of base calls at each position for which an N was called (Figure 11(see page 281)). N replaces a conventional base call when the sequence is unable to make a base call with sufficient confidence.

A **WARNING** is raised if any position shows an N content of >5%. A **FAIL** is raised if any position shows an N content of >20%.

It is not unusual to see a very low proportion of Ns appearing in a sequence (especially near the end of a sequence). However, if this proportion rises above a few percents it suggests that the analysis pipeline was unable to interpret the data well enough to make valid base calls.

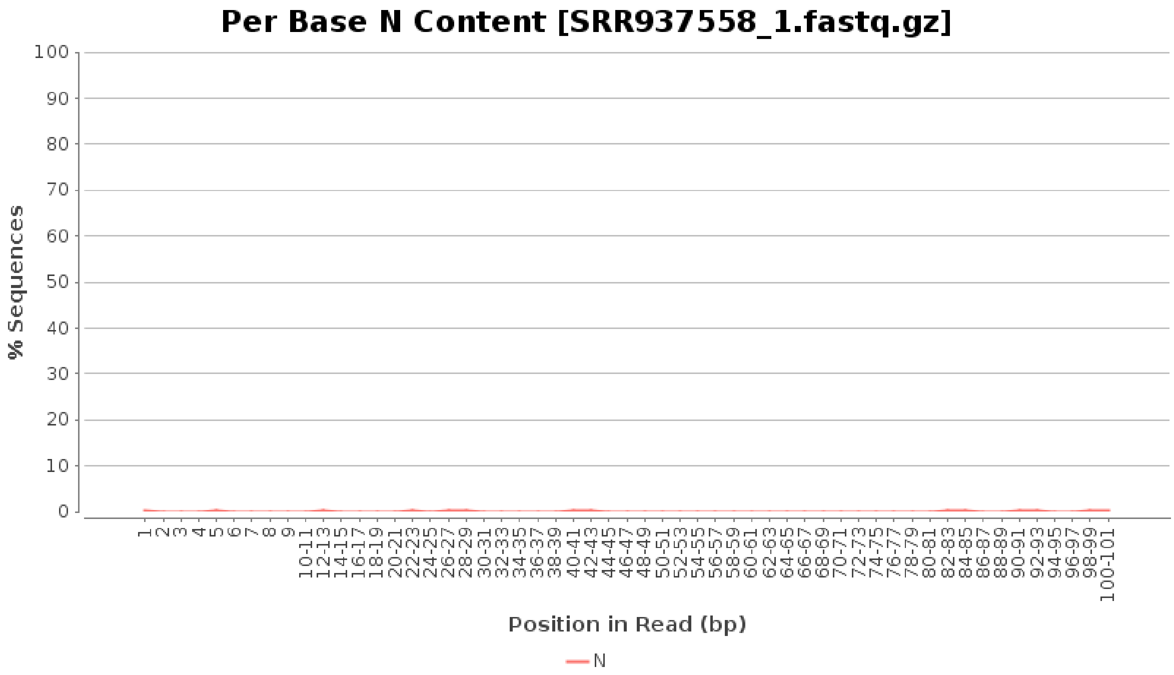


Figure 11: Per Base N Content Chart

11.3.3.6 Sequence Length Distribution

This chart shows the distribution of fragment sizes in the file which was analyzed (Figure 12(see page 282)). In many cases, this will produce a simple graph showing a peak only at one size, but for variable length FASTQ files, this will show the relative amounts of each different size of sequence fragment.

A **WARNING** is raised if all sequences are not the same length. A **FAIL** is raised if any of the sequences have zero length.

> ⚠ For some sequencing platforms, it is entirely normal to have different read lengths so warnings here can be ignored.

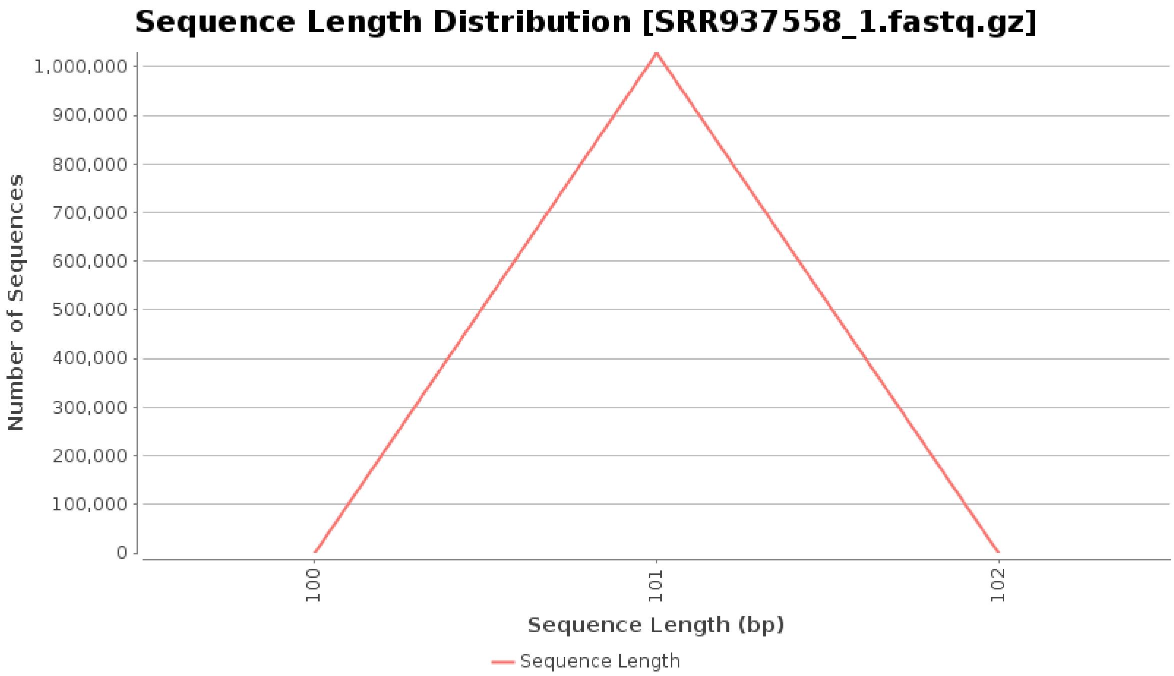


Figure 12: Sequence Length Distribution Chart

11.3.3.7 Adapter Content

This chart shows a cumulative percentage of the proportion of the library in which each of the adapter sequences at each position has been detected (Figure 13(see page 283)). Once a sequence has been detected in a read, it is counted as being present right through to the end of the read so the percentage increases as the read length continues.

A **WARNING** is issued if any sequence is present in more than 5% of all reads. A **FAIL** is issued if any sequence is present in more than 10% of all reads.

This module indicates if the sequences will need to be trimmed for adapters before proceeding with any downstream analysis.

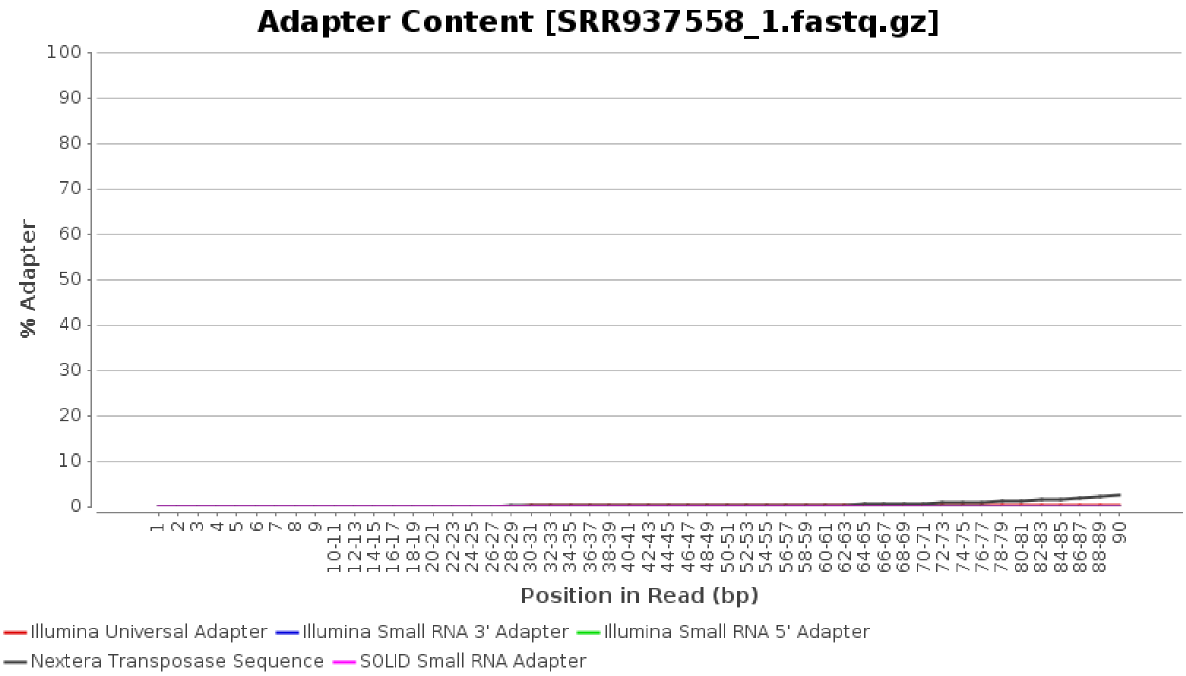


Figure 13: Adapter Content Chart

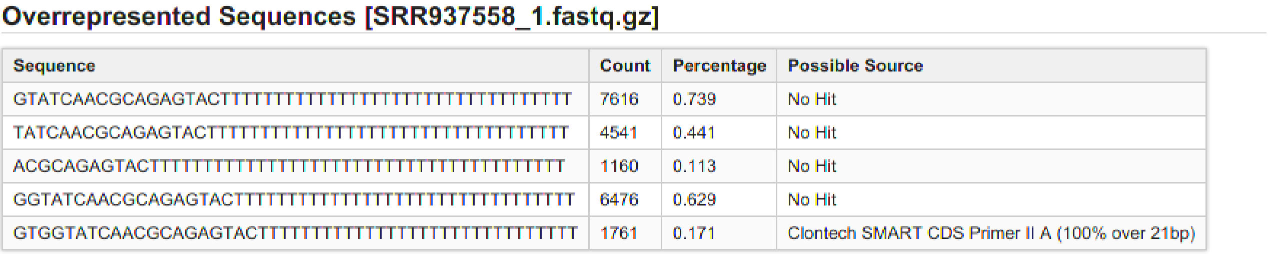
11.3.3.8 Overrepresented Sequences

This module lists all of the sequences which make up more than 0.1% of the total (Figure 14(see page 284)). To conserve memory only sequences that appear in the first 100,000 sequences are tracked to the end of the file. Therefore, it is possible that a sequence which is overrepresented but doesn't appear at the start of the file for some reason could be missed by this module.

For each overrepresented sequence, the program will look for matches in a database of common contaminants and will report the best hit that it finds. Hits must be at least 20 bp in length and have no more than 1 mismatch.

A **WARNING** is issued if any sequence is found to represent more than 0.1% of the total. A **FAIL** is issued if any sequence is found to represent more than 1% of the total.

⚠ This module will often be triggered when used to analyze small RNA libraries where sequences are not subjected to random fragmentation, and the same sequence may naturally be present in a significant proportion of the library.



Overrepresented Sequences [SRR937558_1.fastq.gz]

Sequence	Count	Percentage	Possible Source
GTATCAACGCAGAGTACTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT	7616	0.739	No Hit
TATCAACGCAGAGTACTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT	4541	0.441	No Hit
ACGCAGAGTACTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT	1160	0.113	No Hit
GGTATCAACGCAGAGTACTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT	6476	0.629	No Hit
GTGGTATCAACGCAGAGTACTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT	1761	0.171	Clontech SMART CDS Primer II A (100% over 21bp)

Figure 14: Overrepresented Sequences Table

11.3.3.9 Sequence Duplication Levels

This module counts the degree of duplication for every sequence in a library and creates a graph showing the relative number of sequences with different degrees of duplication (Figure 15(see page 285)). The chart shows the proportion of the library which is made up of sequences in each of the different duplication level bins.

There are two lines on the plot:

- The blue line takes the full sequence set and shows how its duplication levels are distributed.
- The red line displays the proportions of the sequences that are deduplicated which come from different duplication levels in the original data.

The module also calculates an expected overall loss of sequences when the library is deduplicated. This is shown at the top of the plot and gives a reasonable impression of the potential overall level of loss.

A **WARNING** is raised if non-unique sequences make up more than 20% of the total. A **FAIL** is raised if non-unique sequences make up more than 50% of the total.

In general, there are two potential types of duplicates in a library, technical duplicates arising from PCR artifacts, or biological duplicates which are natural collisions where different copies of exactly the same sequence are randomly selected.

In RNA-Seq libraries, sequences from different transcripts will be present at wildly different levels in the starting population. In order to be able to observe lowly expressed transcripts, it is therefore common to greatly over-sequence high expressed transcripts, and this will potentially create large sets of duplicates. This will result in high overall duplication in this test, and will often produce peaks in the higher duplication bins.

⚠ To reduce the memory requirements only the first 100000 sequences of each file are analyzed.

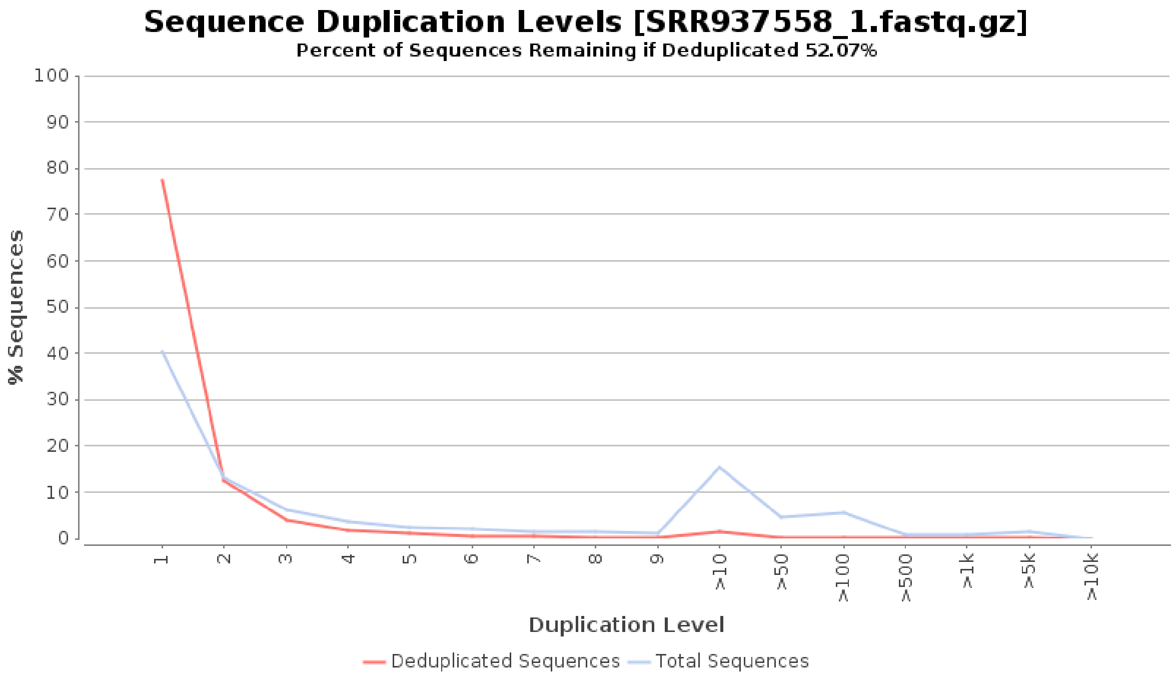


Figure 15: Sequence Duplication Levels Chart

11.4 FASTQ Preprocessing

Content of this page:

11.4.1 Introduction

As Next Generation Sequencing (NGS) technology is used more broadly in scientific applications and research, sequencing data quality control is becoming more important. Experiments and sequencing processes always introduce errors and biases, so downstream sequence analyses are compromised by low-quality sequences, sequence artifacts, and sequence contamination. These problems eventually lead to erroneous conclusions in processes such as assembly and alignment, so a preprocessing step is necessary to produce better analysis results.

Preprocessing FASTQ files in OmicsBox consists of removing adapters and contamination sequences, trimming low-quality bases and filtering short and low-quality reads. Before proceeding, it is advisable to carry out a quality control check of the sequencing data within OmicsBox (FASTQ Quality Check). In this way, problems and biases can be detected, which allows to better configure the preprocessing procedure.

The FASTQ Preprocessing tool uses the well-known preprocessing software **Trimmomatic**. Trimmomatic is a fast, multithreaded command line tool that can be used to trim and crop sequencing data as well as to remove adapters. For further information visit the Trimmomatic web page[157].

Please, cite Trimmomatic as Bolger AM, Lohse M, Usadel B (2014). "Trimmomatic: A flexible trimmer for Illumina Sequence Data". Bioinformatics, btu170.

11.4.1.1 Adapter Removal

This step is used to find and remove adapters and contaminant sequences. The application uses two approaches to detect technical sequences within the reads:

- **Simple mode:** The simple mode approach works by finding an approximate match between the read and supplied technical sequences. These sequences can be detected in any location or orientation within the reads but require a minimum overlap between the read to prevent false-positives. However, short partial adapter sequences cannot achieve this minimum overlap requirement, so they are not detectable.
- **Palindrome mode:** The palindrome mode approach is specifically aimed at detecting the common "adapter read-through" scenario whereby the sequenced DNA fragment is shorter than the read length. When "read-through" happens, both reads in a pair will consist of an equal number of valid bases, followed by contaminating sequences from the "opposite" adapters. Furthermore, the valid sequence within the two reads will be reverse complements. This mode can only be used with paired-end data but has considerable advantages in sensitivity and specificity over "simple" mode.

11.4.1.2 Trimming

This step is used to remove low-quality bases from the reads. The application offers four trimming alternatives:

- **Sliding window trimming:** The sliding window approach works by scanning from the 5' end of the read and removes the remaining 3' end of the read when the average quality of a group of bases drops below a specified threshold.
- **Adaptive quality trimming:** The adaptive quality trim approach, also known as "Maximum information quality trimming", balances the benefits of retaining longer reads against the cost of retaining bases with errors.
- **Quality trimming:** The quality trimming approach removes low-quality bases from the beginning or the end of the read. As long as a base has a value below this threshold, the base is removed and the next base will be investigated.
- **Length trimming:** The length trimming approach removes a specified number of bases regardless of quality from the beginning or the end of the read.

[157] http://www.usadellab.org/cms/?page=trimmomatic

11.4.1.3 Filtering

This step is used to filter out reads:

- **Filter by quality:** Remove reads that fall below the specified average quality.
- **Filter by length:** Remove reads that fall below the specified minimum length.

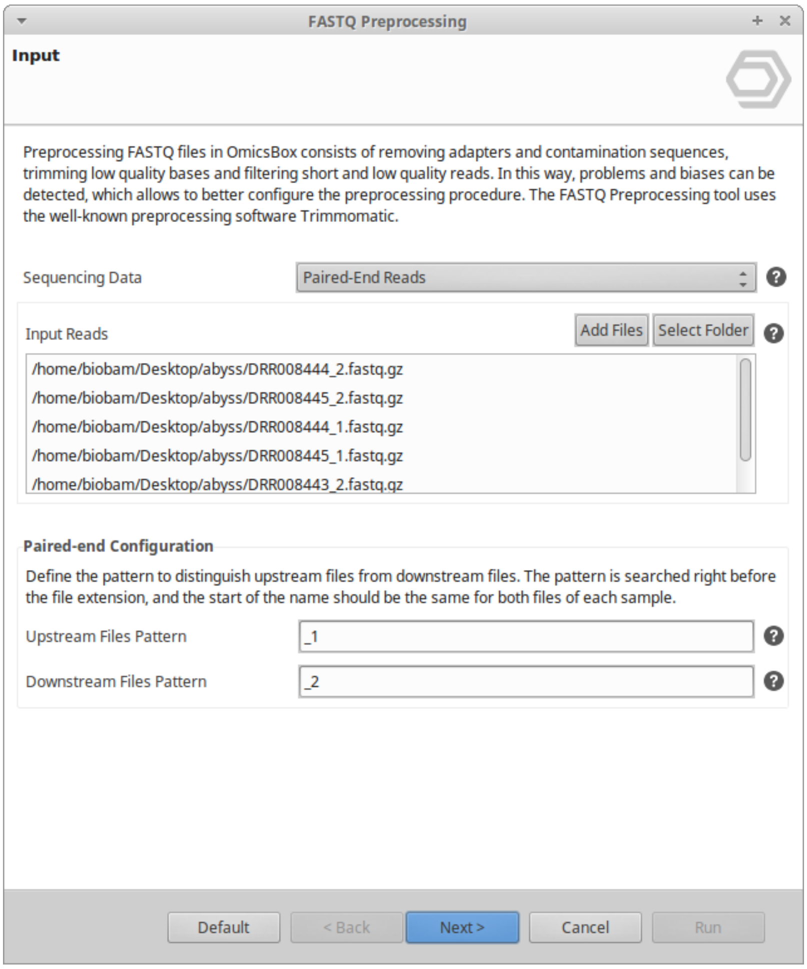
11.4.2 Run FASTQ Preprocessing

This functionality can be found under **General Tools → FASTQ Tools → FASTQ Preprocessing.** The input data and the different preprocessing steps can be configured using the wizard.

11.4.2.1 Input Data Page

- **Sequencing Data:** Choose the type of data to be preprocessed: single-end or paired-end reads. Note that if paired-end is selected, two files per sample are required.
- **Input Reads:** Provide the files containing sequencing reads. These files are assumed to be in FASTQ format.
- **Paired-end configuration:** In case of paired-end reads, the pattern to distinguish upstream files from downstream files is required. The provided patterns are searched right before the extension, and the start of the name should be the same for both files of each sample.
  - Upstream Files Pattern: Establish the pattern to recognize upstream FASTQ files.
  - Downstream Files Pattern: Establish the pattern to recognize downstream FASTQ files.

  > ⚠ For example, if the upstream file is named SRR037717_1.fastq and the downstream one SRR037717_2.fastq, you should establish "_1" as the upstream pattern and "_2" as the downstream pattern.



Preprocessing FASTQ files in OmicsBox consists of removing adapters and contamination sequences, trimming low quality bases and filtering short and low quality reads. In this way, problems and biases can be detected, which allows to better configure the preprocessing procedure. The FASTQ Preprocessing tool uses the well-known preprocessing software Trimmomatic.

Sequencing Data: Paired-End Reads

Input Reads

```
/home/biobam/Desktop/abyss/DRR008444_2.fastq.gz
/home/biobam/Desktop/abyss/DRR008445_2.fastq.gz
/home/biobam/Desktop/abyss/DRR008444_1.fastq.gz
/home/biobam/Desktop/abyss/DRR008445_1.fastq.gz
/home/biobam/Desktop/abyss/DRR008443_2.fastq.gz
```

Paired-end Configuration

Define the pattern to distinguish upstream files from downstream files. The pattern is searched right before the file extension, and the start of the name should be the same for both files of each sample.

Upstream Files Pattern: _1

Downstream Files Pattern: _2

Figure 1: Input Data Page

11.4.2.2 Adapter Removal Page

- **Remove Adapters:** Enable the adapter removal step.
- **Use Adapters From:** Choose between using the default adapter sequences provided by Trimmomatic, or providing custom adapter sequences.
  - Default Adapter Sequences: By default, the application provides adapter sequences for TruSeq2 (GAII machines), TruSeq3 (HiSeq and MiSeq machines) and Nextera, for both single-end and paired-end data.

    ⚠ If you use the FASTQ Quality Check tool, the "Adapter Content" and "Overrepresented Sequences" modules can help to choose which default adapter sequences are best suited for your data. "Illumina Single-End" or "Illumina Paired-End" sequences indicate single-end or paired-end TruSeq2 libraries. "TruSeq Universal Adapter" or "TruSeq Adapter, Index..." sequences indicate TruSeq-3 libraries. Note that these sequences have not been extensively tested, so other sequences may work better for a given dataset.

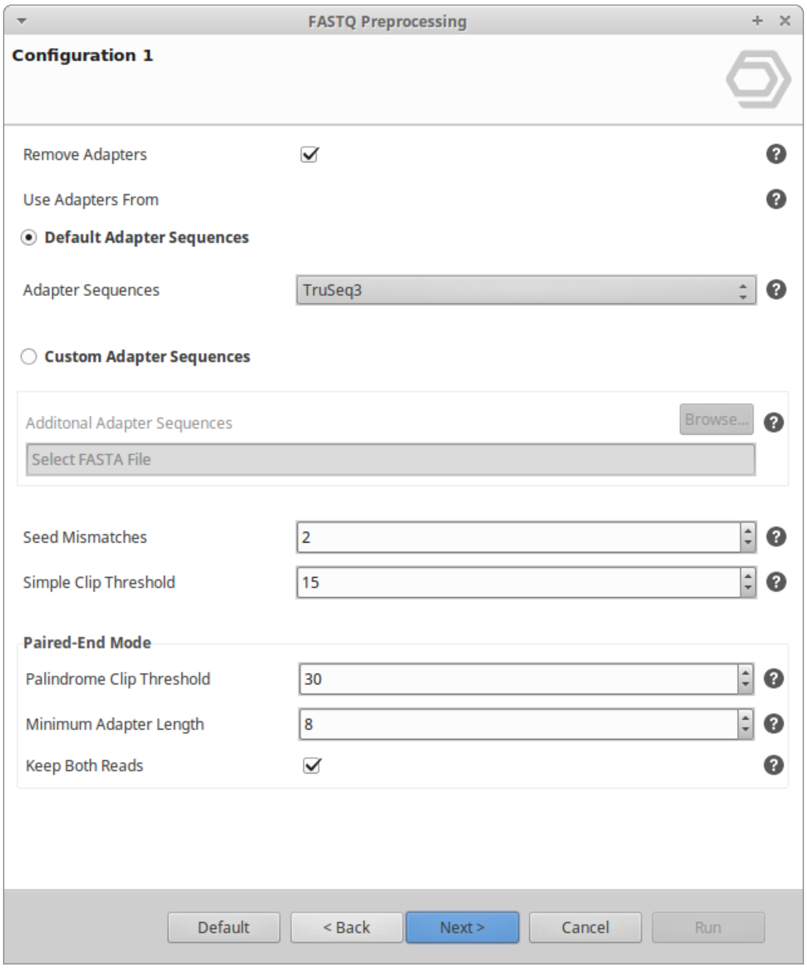
- Custom Adapter Sequences: Specifies a FASTA file containing all the adapters and contaminant sequences to be removed. The names of sequences determine how they are used, especially for paired-end data.
    - For "palindrome" mode, matched pairs of adapter sequences must be supplied. The sequence names should start with "Prefix", and end in "/1" for the forward adapter and "/2" for the reverse adapter. The part of the name between "Prefix" and "/1" or "/2" must match exactly within each pair.

```
>PrefixPE/1
TACACTCTTTCCCTACACGACGCTCTTCCGATCT
>PrefixPE/2
GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
```

    - For "simple" mode, sequences with names ending in "/1" or "/2" will be searched only in the forward or reverse read respectively. Otherwise, sequences will be searched in both the forward and reverse read.

```
>Adapter_a
AGATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG
>Adapter_b
AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
```

- **Seed Mismatches:** Set the maximum mismatch count which allows performing a full match.
- **Simple Clip Threshold:** Establish how accurate the match between the adapter sequence must be against a read. This option is only considered for the simple mode.
- **Palindrome Clipt Threshold:** Establish how accurate the match between the two "adapter ligated" reads must be. This option is only considered for the palindrome mode.
- **Minimum Adapter Length:** Set a minimum length for adapters to be detected. This option is only considered for the palindrome mode.
- **Keep Both Reads:** Deleting adapters after read-through detection (palindrome mode) causes the reverse read to contain the same information as the forward read, although in reverse complement. This option allows retaining the reverse read. Otherwise, the reverse read will be discarded.



Configuration 1

Remove Adapters ☑

Use Adapters From

● Default Adapter Sequences

Adapter Sequences: TruSeq3

○ Custom Adapter Sequences

Additonal Adapter Sequences

Select FASTA File

Seed Mismatches: 2

Simple Clip Threshold: 15

Paired-End Mode

Palindrome Clip Threshold: 30

Minimum Adapter Length: 8

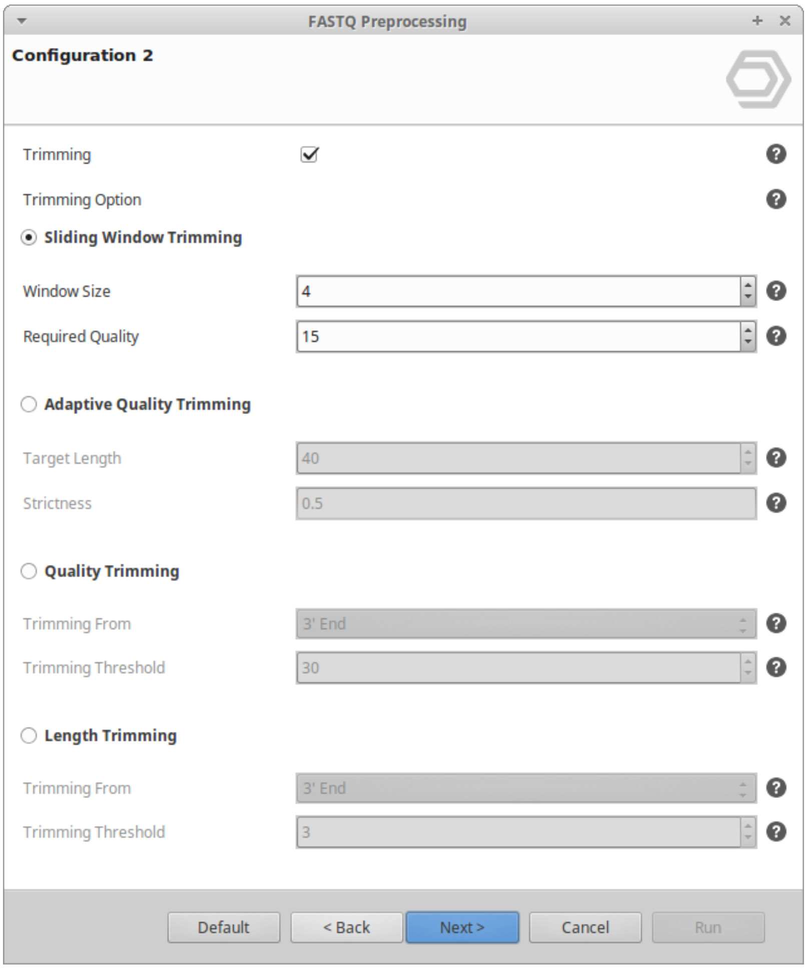
Keep Both Reads ☑

Figure 2: Adapter Removal Page

11.4.2.3 Trimming Page

- **Trimming:** Enable the trimming step.
- **Trimming option:** Select the strategy to perform the trimming step.
- **Sliding Window Trimming:**
  - Window Size: Set the number of bases that the window has to span to average the quality.
  - Required Quality: Set the average quality required to retain bases.
- **Adaptive Quality Trimming:**
  - Target Length: Set the minimum read length which is likely to allow the location of the read within the target sequence.
  - Strictness: This value establishes the balance between preserving as much read length as possible versus removal of incorrect bases. It should be set between 0 and 1. A low value favors longer reads, while a high value favors read correctness.
- **Quality Trimming:**
  - Trimming from: Choose between removing bases from the beginning or the end of the sequence.

  - Trimming threshold: Establish a minimum quality required to keep bases.
- **Length Trimming:**
  - Trimming from: Choose between removing bases from the beginning or the end of the sequence.
  - Trimming threshold: In case of removing bases from the end, specifies the number of bases to be kept from the start of the read so that it has maximally the specified length after this step. In case of removing bases from the start, specifies the number of bases to be removed from the start of the read.



FASTQ Preprocessing

Configuration 2

Trimming ☑

Trimming Option

● Sliding Window Trimming

Window Size: 4

Required Quality: 15

○ Adaptive Quality Trimming

Target Length: 40

Strictness: 0.5

○ Quality Trimming

Trimming From: 3' End

Trimming Threshold: 30

○ Length Trimming

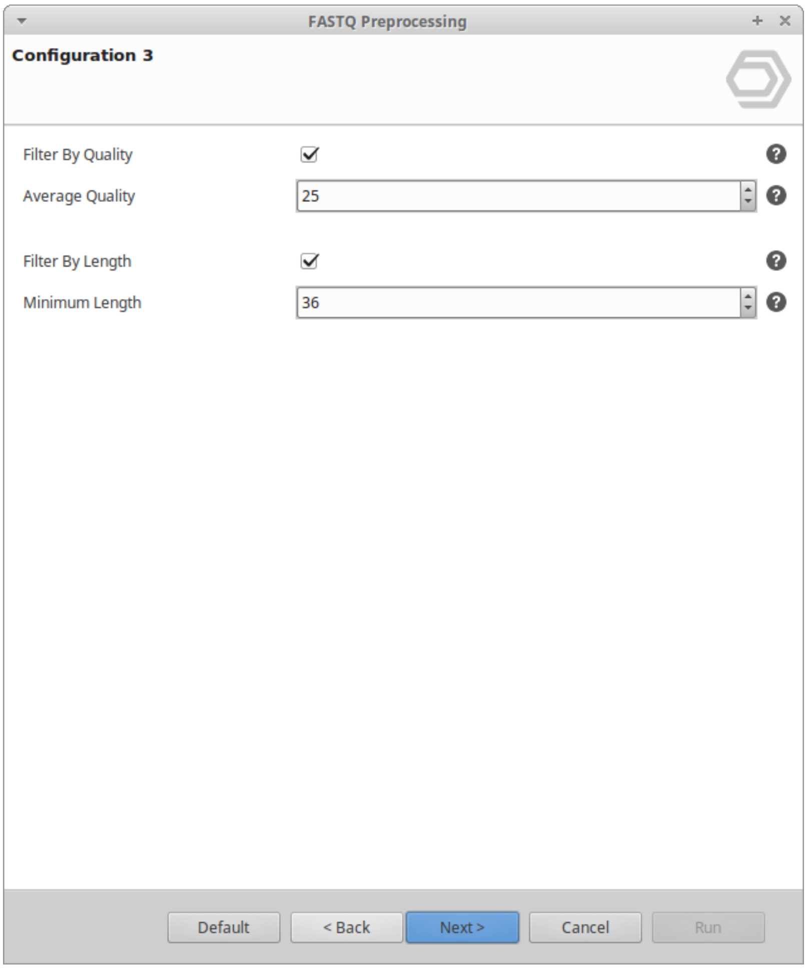
Trimming From: 3' End

Trimming Threshold: 3

Figure 3: Trimming Page

11.4.2.4 Filtering Page

- **Filter By Quality:** Enable the filtering by quality step.
- **Average Quality:** Minimum average quality of reads to be kept.
- **Filter By Length:** Enable the filtering by length step.
- **Minimum Length:** Minimum length of reads to be kept.


FASTQ Preprocessing

Configuration 3

Filter By Quality ☑

Average Quality 25

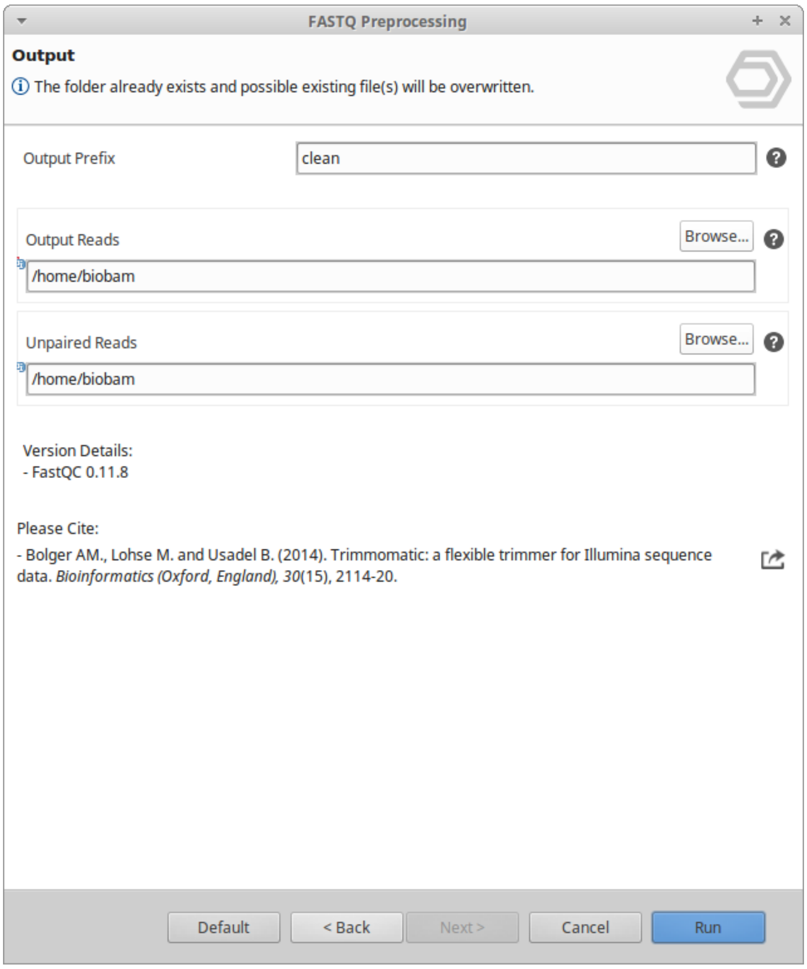
Filter By Length ☑

Minimum Length 36

Figure 4: Filtering Page

11.4.2.5 Save Results Page

- **Output Prefix:** Define a prefix to establish the name of output files. The prefix will be added before each original file name.
- **Output Reads:** Select a destination folder to save the preprocessed FASTQ files.
- **Unpaired Reads:** When preprocessing paired-end data, some read pairs can lose a member as a result of trimming and filtering. Select a destination folder to save the FASTQ files containing unpaired reads. These files contain the word "unpaired" in their file names.



Output

ⓘ The folder already exists and possible existing file(s) will be overwritten.

Output Prefix: clean

Output Reads: /home/biobam

Unpaired Reads: /home/biobam

Version Details:
- FastQC 0.11.8

Please Cite:
- Bolger AM., Lohse M. and Usadel B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)*, *30*(15), 2114-20.

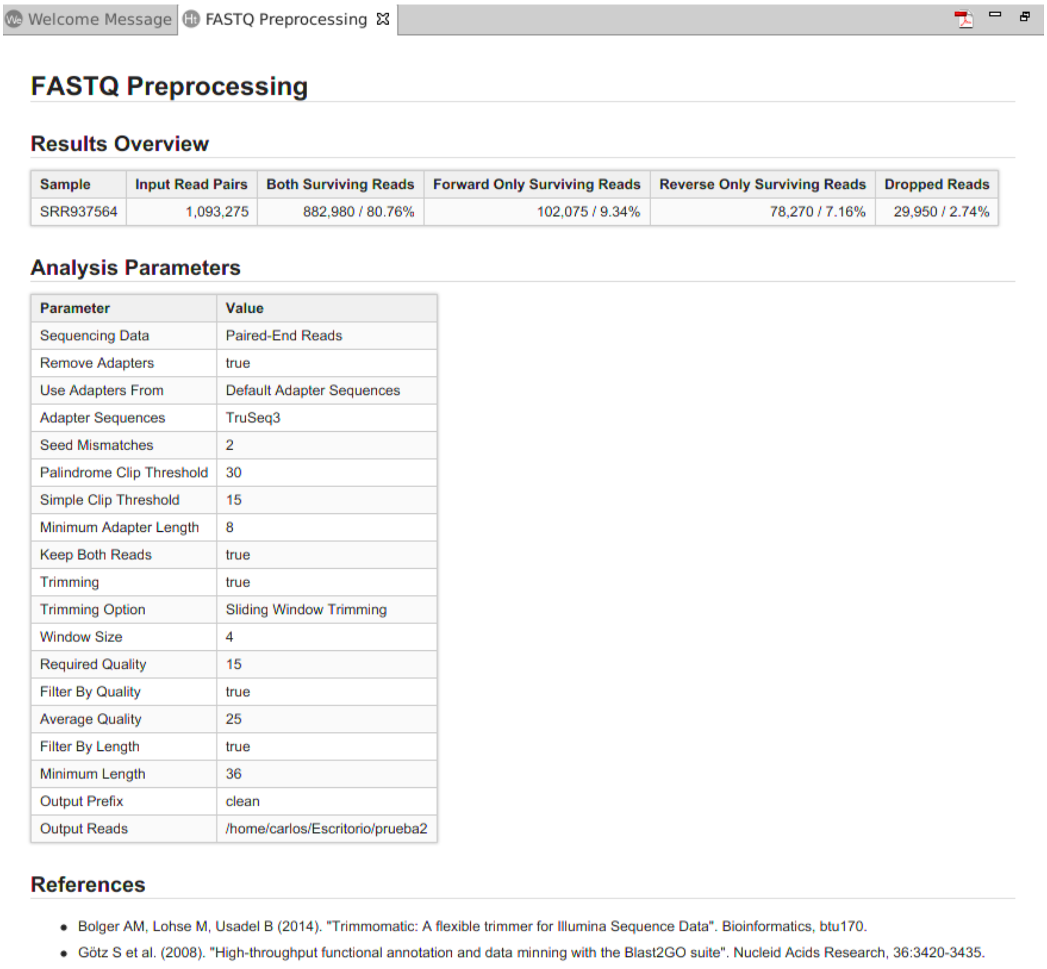
Figure 5: Output Data Page

11.4.3 Results

Once finished, output files containing preprocessed reads are stored in the "Output Reads" folder set in the wizard. Files are generated in compressed format (fastq.gz).

For single-end data, one output file per input file is generated. For paired-end data, four output files per input sample (2 FASTQ files) are generated, two that contain upstream and downstream paired reads and two that contain upstream and downstream unpaired reads. The name of each output file begins with the provided prefix and continues with the original name of the file. Files with unpaired reads contain the word "unpaired" in their name so that they can be distinguished from those that contain paired reads. These files are placed in the "Unpaired Reads" folder.

Furthermore, a result page will show a summary of the "FASTQ Preprocessing" results (Figure 6(see page 294)). This page provides a table that shows how many reads have survived and how many have been dropped during the analysis.



FASTQ Preprocessing

Results Overview

Sample	Input Read Pairs	Both Surviving Reads	Forward Only Surviving Reads	Reverse Only Surviving Reads	Dropped Reads
SRR937564	1,093,275	882,980 / 80.76%	102,075 / 9.34%	78,270 / 7.16%	29,950 / 2.74%

Analysis Parameters

Parameter	Value
Sequencing Data	Paired-End Reads
Remove Adapters	true
Use Adapters From	Default Adapter Sequences
Adapter Sequences	TruSeq3
Seed Mismatches	2
Palindrome Clip Threshold	30
Simple Clip Threshold	15
Minimum Adapter Length	8
Keep Both Reads	true
Trimming	true
Trimming Option	Sliding Window Trimming
Window Size	4
Required Quality	15
Filter By Quality	true
Average Quality	25
Filter By Length	true
Minimum Length	36
Output Prefix	clean
Output Reads	/home/carlos/Escritorio/prueba2

References

- Bolger AM, Lohse M, Usadel B (2014). "Trimmomatic: A flexible trimmer for Illumina Sequence Data". Bioinformatics, btu170.
- Götz S et al. (2008). "High-throughput functional annotation and data minning with the Blast2GO suite". Nucleid Acids Research, 36:3420-3435.

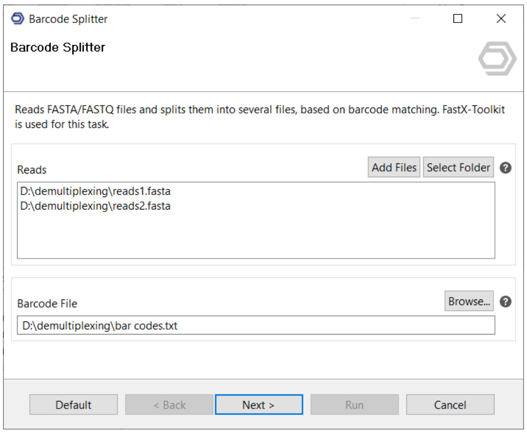
Figure 6: FASTQ Preprocessing Report

11.5 Barcode Splitter

Content of this page:

11.5.1 Introduction

Demultiplexing or barcode splitting refers to the step in processing where you would use the barcode information in order to know which sequences came from which sample after they had all been sequenced together. Barcodes refer to the unique sequences that were ligated to your each of your individual samples' genetic material before the samples got all mixed together. Depending on your sequencing facility, you may get your samples already split into individual fastq files, or they may be lumped together all in one fastq file with barcodes still attached for you to do the splitting. If this is the case, you should also have a mapping or barcode file telling you which barcodes correspond with which samples.
This tool takes FASTA/FASTQ files and splits them into several smaller files, Based on barcode matching. FastX-Toolkit is used for this task.



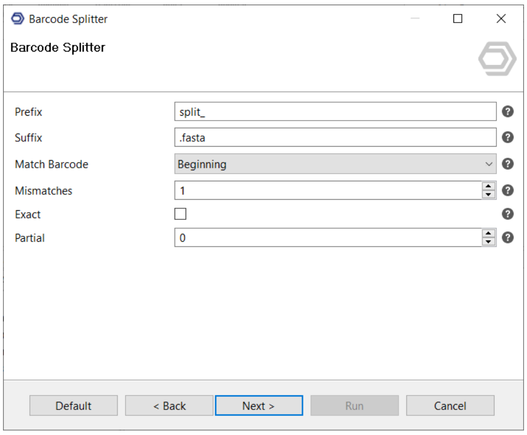
Barcode Splitter

Reads FASTA/FASTQ files and splits them into several files, based on barcode matching. FastX-Toolkit is used for this task.

Reads
D:\demultiplexing\reads1.fasta
D:\demultiplexing\reads2.fasta

Barcode File
D:\demultiplexing\bar codes.txt

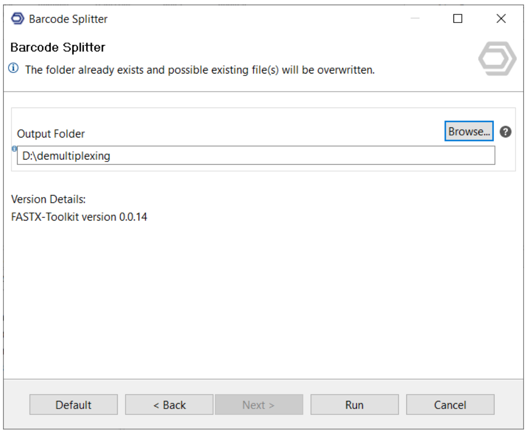
Page 1



Barcode Splitter

Prefix: split_
Suffix: .fasta
Match Barcode: Beginning
Mismatches: 1
Exact: □
Partial: 0

Page 2



Barcode Splitter

Barcode Splitter

The folder already exists and possible existing file(s) will be overwritten.

Output Folder: D:\demultiplexing

Version Details:
FASTX-Toolkit version 0.0.14

Page 3

11.5.2 Page 1 - Input

Reads - Select the FastQ/A files that contain sequences that have attached barcodes which link those sequences to the respective samples.

Barcode File - Select the mapping file that establishes the connection between each barcode and sample.

Barcode file format

Barcode files are simple text files. Each line should contain an identifier (descriptive name for the barcode), and the barcode itself (A/C/G/T), separated by a TAB character. Example:

```
#This line is a comment (starts with a 'number' sign)
BC1 GATCT
BC2 ATCGT
BC3 GTGAT
BC4 TGTCT
```

For each barcode, a new FASTQ file will be created (with the barcode's identifier as part of the file name). Sequences matching the barcode will be stored in the appropriate file.

Running the above example (assuming "mybarcodes.txt" contains the above barcodes), will create the following files:

```
/tmp/bla_BC1.txt
/tmp/bla_BC2.txt
/tmp/bla_BC3.txt
/tmp/bla_BC4.txt
/tmp/bla_unmatched.txt
```

The 'unmatched' file will contain all sequences that didn't match any barcode.

11.5.3 Page 2 - Configuration

Prefix - File prefix that will be added to the output files.

Suffix - File suffix that will be added to the output files.

Match Barcode - Match the barcodes at the beginning (5') or end (3') of each sequence.

Mismatches - Maximum number of allowed mismatches for barcodes.

Partial - Allow partial overlap of barcodes.

Without partial matching:

Count mismatches between the FASTA/Q sequences and the barcodes. The barcode which matched with the lowest mismatches count (providing the count is small or equal to '--mismatches N') 'gets' the sequences.

Example (using the above barcodes):
Input Sequence:
GATTTACTATGTAAAGATAGAAGGAATAAGGTGAAG

Matching at beginning of sequenecs and 1 mismatch:
GATTTACTATGTAAAGATAGAAGGAATAAGGTGAAG
GATCT (1 mismatch, BC1)
ATCGT (4 mismatches, BC2)
GTGAT (3 mismatches, BC3)
TGTCT (3 mismatches, BC4)

This sequence will be classified as 'BC1', because it has the lowest mismatch count.
If mismatches = 0 were specified, this sequence would be classified as 'unmatched', because, although BC1 had the lowest mismatch count,
it is above the maximum allowed mismatches.

Matching barcodes at the end of the sequences does the same, but from the other side of the sequence.

With partial matching (very similar to indels):

Same as above, with the following addition: barcodes are also checked for partial overlap.

Example:
Input sequence is ATTTACTATGTAAAGATAGAAGGAATAAGGTGAAG
(Same as above, but note the missing 'G' at the beginning.)

Matching (without partial overlapping) against BC1 yields 4 mismatches:
ATTTACTATGTAAAGATAGAAGGAATAAGGTGAAGGATCT (4 mismatches)

Partial overlapping would also try the following match:
-ATTTACTATGTAAAGATAGAAGGAATAAGGTGAAGGATCT (1 mismatch)

Note: Scoring counts a missing base as a mismatch, so the final mismatch count is 2 (1 'real' mismatch, 1 'missing base' mismatch).
If running with mismatches = 2 (meaning allowing up to 2 mismatches), this sequence will be classified as BC1.

11.5.4 Page 3 - Output

Output Folder - Define a folder to save the results.

12 File Menu

Content of this page:

- **Recent files**: Allows reopening recently closed projects.
- **Open file**: Open an OmicsBox project (.b2g/.box).
- **Save and Save as…:** Save the current project.
- **Load**: Allows loading of different file types into OmicsBox.
- **Export**: Allows exporting the desired information to some file types, such as text or GFF.
- **Manage License**: Check the license you currently have and change the activation key.
- **Preferences**: Set the OmicsBox configuration.

12.1 Data Import and Export

Under the File menu and the Tools sub-menu, there are several useful features that can be used to manipulate sequence data.

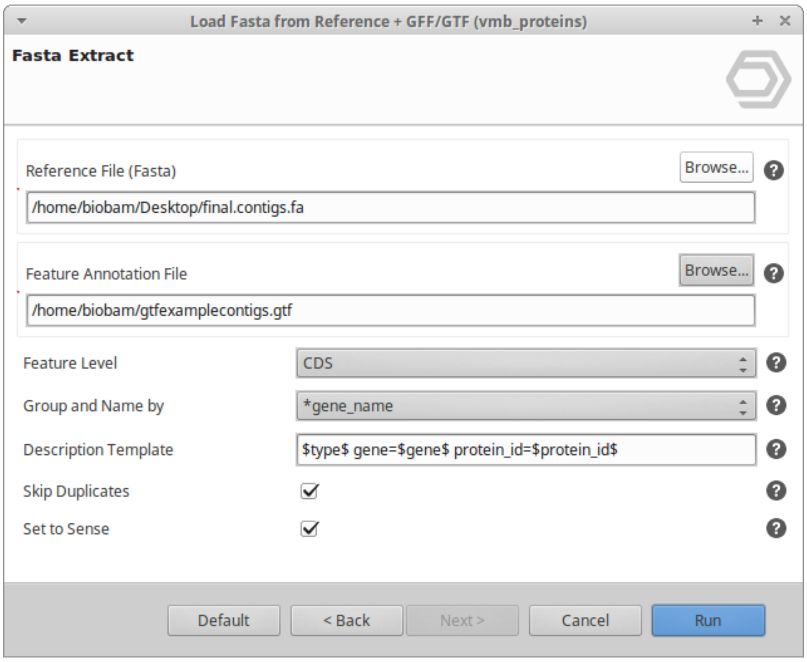
12.2 Load

1. Extract and import sequences from a FASTA and a GFF/GTF file (figure 1[158]). For further information, please link here[159].
2. Load Blast and InterProScan results.
3. Load ID lists.
4. Load GTF/GFF2/GFF3
5. Load Accession List: Load Gene Ontology annotations via an Accession list.
6. Load GeneSymbol List: Load Gene Ontology annotations via a GeneSymbol list.
7. Load GI-List: load Gene Ontology annotations via a GenInfo Identifier (gi) list. Please consider the identifier to be between vertical bar e.g. gi|356569257|.
8. Load Data from BioMart: Load Gene Ontology annotations from BioMart. For further details on how to load annotations, link here[160].

[158] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=598278287#FileMenu-figure4
[159] https://www.biobam.com/load-fasta-from-reference-gff/
[160] https://www.biobam.com/load-seqannot-identifiers/

9. Load EggNOG, Metagenomics and PfamScan annotations.
10. Load metagenomic Kraken data.
11. Load Count Tables and Differential Expression results.
12. Load BAM and VCF files.

> ⚠ The Accession List and the GeneSymbol file should contain two columns (separated by tabs) per line. The first column the accession id or gene symbol and the second column may contain the corresponding taxonomy. The second column is optional.



Load Fasta from Reference + GFF/GTF (vmb_proteins)

Fasta Extract

Reference File (Fasta): /home/biobam/Desktop/final.contigs.fa

Feature Annotation File: /home/biobam/gtfexamplecontigs.gtf

Feature Level: CDS

Group and Name by: *gene_name

Description Template: $type$ gene=$gene$ protein_id=$protein_id$

Skip Duplicates: ☑

Set to Sense: ☑

Figure 1: Extract and import sequences from a FASTA and a GFF/GTF file.

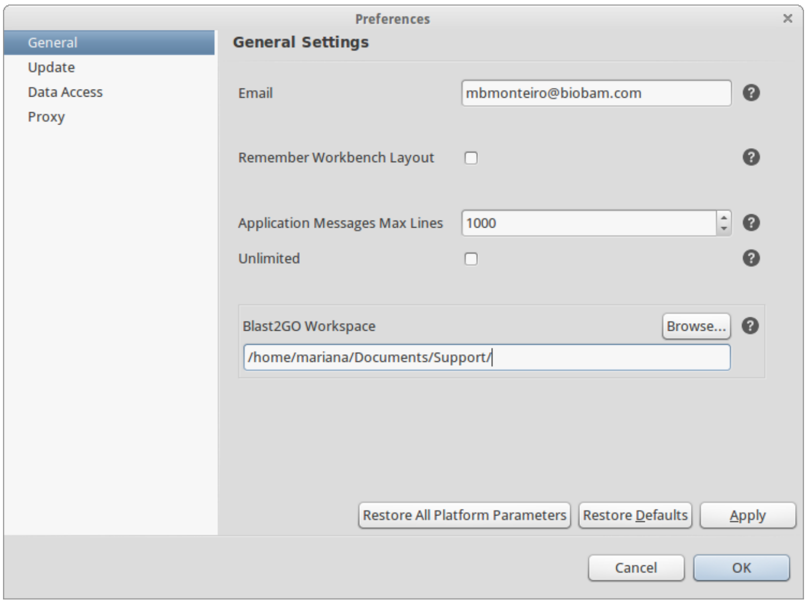
12.3 Export

1. Generic Export: This option allows you to export all the desired information to a text file.
2. Export Selected Sequences as Project: Only the selected sequences can be exported and saved in .dat file.
3. Export Sequence Table: Export the current Main Sequence Table for the selected sequences.
4. Export TopBlast data: It will export the best-blast-hit for each sequence, this is the hit with the lowest e-value.
5. Export GO Propagation: Exports the GO parents up to the root for the annotated sequences.
6. Export Sequences per GO (Gene Set).
7. Export GFF: This option is only visible if a GFF file is loaded in OmicsBox or if a GFF has been generated from Gene Finding(see page 111). It is also possible to export a GFF with GO terms. For further details, link here[161].

[161] https://www.biobam.com/how-to-export-gene-finder-gff-output-with-go-terms/

12.4 General Preferences

In the General Preferences, it is possible to enter a valid email address, which will be used in the QBLAST and also in the InterProScan searches. Furthermore, the path to the OmicsBox workspace can be provided and all results such as BLAST, InterProScan, charts and OmicsBox projects will be saved here.



Email: mbmonteiro@biobam.com

Remember Workbench Layout ☐

Application Messages Max Lines: 1000

Unlimited ☐

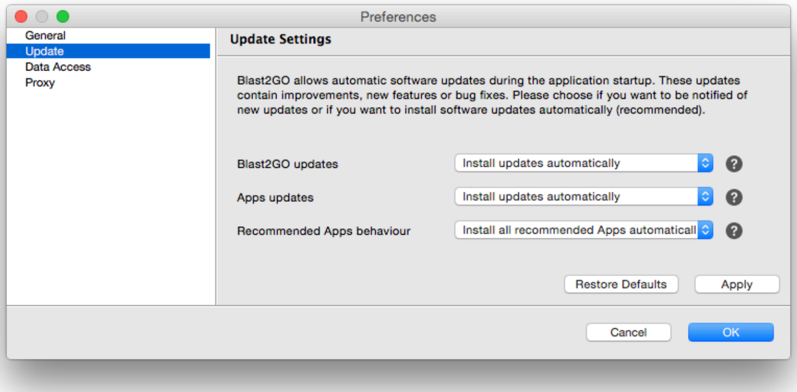
Blast2GO Workspace: /home/mariana/Documents/Support/

Figure 2: OmicsBox General preferences

12.5 Update

OmicsBox allows automatic software updates during the application startup. These updates contain improvements, new features or bug fixes. It is possible to choose if you want to be notified of new updates or if you want to install software updates automatically (recommended).

It is also possible to specify the update behaviour of installed Apps. We differentiate between "Featured" and normal Apps. New "featured" Apps can be installed and updated automatically. Normal, non-featured Apps have to be installed manually but can be updated automatically.



Update Settings

Blast2GO allows automatic software updates during the application startup. These updates contain improvements, new features or bug fixes. Please choose if you want to be notified of new updates or if you want to install software updates automatically (recommended).

Blast2GO updates: Install updates automatically

Apps updates: Install updates automatically

Recommended Apps behaviour: Install all recommended Apps automaticall

Figure 3: Wizard to configure the OmicsBox update behaviour

12.6 GO Version

For the Functional Analysis Module OmicsBox contains the Gene Ontology database and all the information necessary to perform the mapping step i.e. to be able to link the different protein IDs to the functional information of the Gene Ontology database (see Gene Ontology Mapping section).

Here one can select the GO version available on OmicsBox servers as well as the corresponding .obo file to be used in the mapping step.

12.6.1 Local OmicsBox database

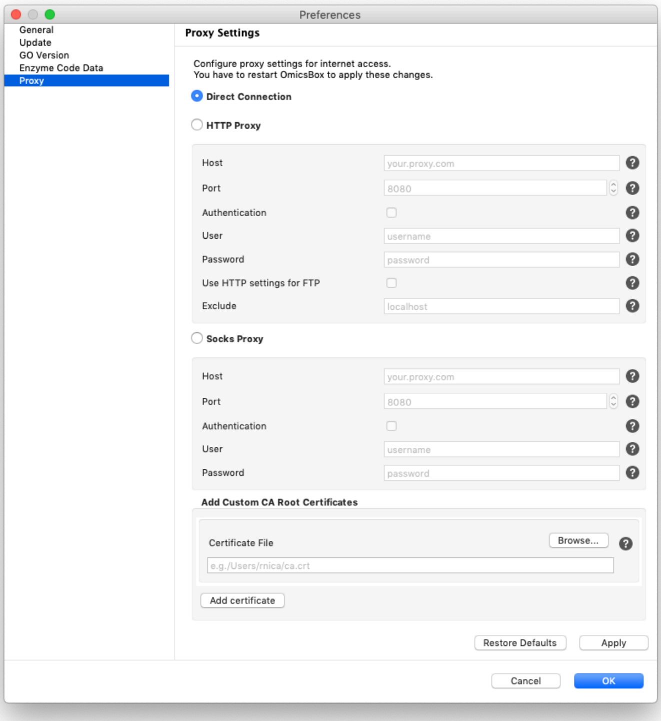
Local OmicsBox database installation: If you are interested in installing your own OmicsBox database locally with the aim to not depend on the OmicsBox server, you can find a tutorial on the OmicsBox website in the download section including a step-by-step installation guide. Basically will need a MySql server, the latest GO database dump and some additional "mapping tables" (NCBI and PIR flat-files). By following several few steps this data is imported into your database.

12.7 Enzyme Code Data

In OmicsBox it is possible to provide a file with the corresponding Enzyme Codes.

12.8 Proxy

Proxy Settings. If a proxy server or a firewall is used to access the internet here you can define the proxy settings. An HTTP or a Socks proxy can be configured. In this window, you can configure the proxy settings only for OmicsBox and this will overcome the system-wide settings. If the Use Direct Connection check box is selected, the application will try to connect directly to the internet skipping any system settings. To use your defined proxy settings select the HTTP or Socks Proxy check box and complete the required fields.



Proxy Settings

Configure proxy settings for internet access.
You have to restart OmicsBox to apply these changes.

● Direct Connection

○ HTTP Proxy

Host your.proxy.com
Port 8080
Authentication □
User username
Password password
Use HTTP settings for FTP □
Exclude localhost

○ Socks Proxy

Host your.proxy.com
Port 8080
Authentication □
User username
Password password

Add Custom CA Root Certificates

Certificate File
e.g./Users/rnica/ca.crt

Figure 4: Proxy settings dialog

12.9 Custom CA Root Certificates

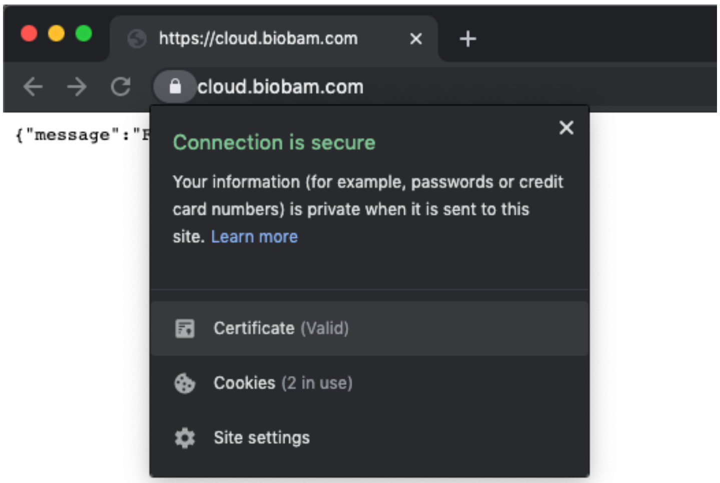
This option allows the import of custom CA root certificates to the OmicsBox trusted entities. It is located in the same Proxy setting page.

A certificate authority (CA) is an entity that issues digital certificates. A digital certificate certifies the ownership of a public key by the named subject of the certificate (Common Name or CN in a certificate). A CA acts as a trusted third-party and allows OmicsBox to rely on the packets received through the connection to OmicsCloud or other internet sites. The format of these certificates is specified by the X.509 standard.

In a normal configuration, a secure connection is established between the end client (OmicsBox) and the server (e.g.: OmicsCloud) and the packages travel encrypted between client and server. In some custom firewall configurations where all traffic going through the network is inspected by the firewall, a secure connection is established between the client application and the firewall, and another connection is established between the firewall and the servers. The firewall can act as a man-in-the-middle to inspect the packages and will re-encrypt them using its own certificate.

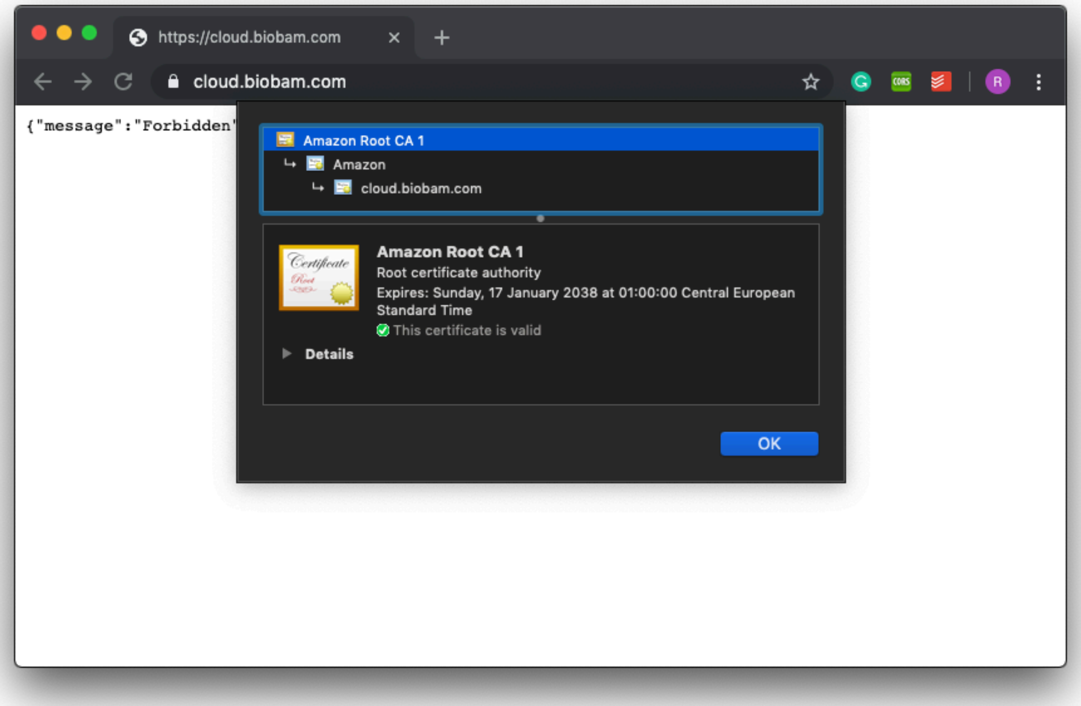
This may cause connection problems if OmicsBox does not recognize the certificate used in the firewall as a trusted entity. The IT department of the institution will know if a custom certificate is used and can provide you with the CA root of this certificate, or the certificate itself, to be added to OmicsBox. This is usually a .crt file that can be provided in the wizard page.

It can also be obtained from a regular web browser by opening a page that is known to be inspected by the firewall, for example the connection to the OmicsCloud https://cloud.biobam.com by clicking on the padlock next to the url address the certificate option shows the path to the certificate and the signing authority. On our servers, the CA root entity is either *Amazon Root CA 1* or *GlobalSign Root CA*. Something else is an indicator of a custom CA Root certificate.



Connection is secure

Your information (for example, passwords or credit card numbers) is private when it is sent to this site. Learn more

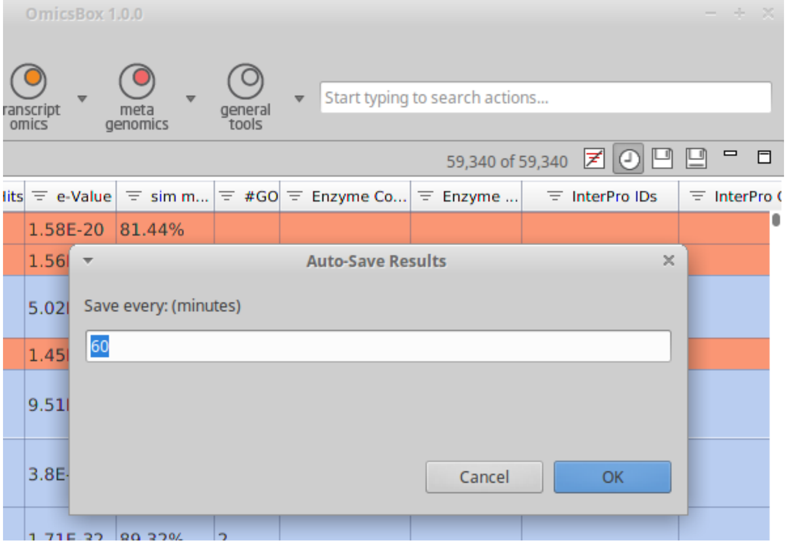


Amazon Root CA 1
Root certificate authority
Expires: Sunday, 17 January 2038 at 01:00:00 Central European Standard Time
This certificate is valid

In Windows and Linux, the Root certificate can be saved from within the browser. In MacOS, this needs to be searched in the keychain app and exported from there. Once this file is exported, it can be directly imported in OmicsBox Proxy preferences page.

12.10 Auto-Save

OmicsBox allows to automatically and continuously save OmicsBox results after a certain amount of time.



Auto-Save Results

Save every: (minutes)

60

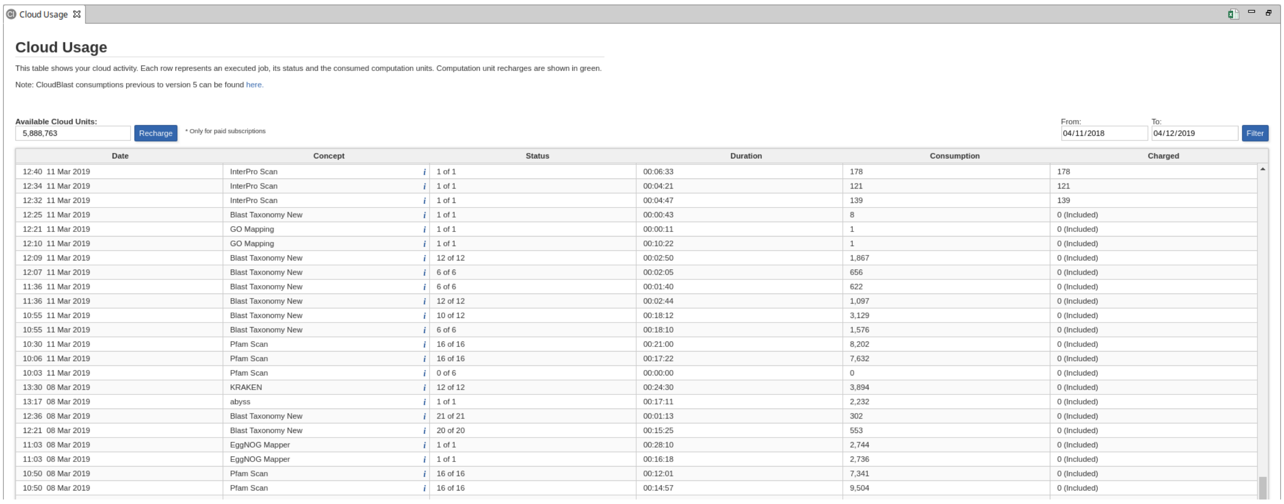
13 View Menu

- Application Messages
  Shows general application message as well as summary information of specific job executions.
- Welcome Message
  A window which provides information about application updates and new features.
- Progress
  This tab provides process information of any job execution in OmicsBox. Each job can be cancelled as well as a more detailed message tab can be opened.
- File Manager
- Cloud Usage
- Memory/CPU Monitor
  Shows the used and available memory for OmicsBox. In blue the CPU utilization of the last minute can be seen.

Cloud Usage

The user can open the Cloud Usage in order to monitor the number of consumed/ recharged ComputationUnits or processed sequences and success jobs.

The Excel icon on the top right in Figure 1 allows exporting the table in csv format.



Cloud Usage

This table shows your cloud activity. Each row represents an executed job, its status and the consumed computation units. Computation unit recharges are shown in green.

Note: CloudBlast consumptions previous to version 5 can be found here.

Available Cloud Units: 5,888,763

* Only for paid subscriptions

From: 04/11/2018 To: 04/12/2019

Date	Concept	Status	Duration	Consumption	Charged
12:40 11 Mar 2019	InterPro Scan	1 of 1	00:06:33	178	178
12:34 11 Mar 2019	InterPro Scan	1 of 1	00:04:21	121	121
12:32 11 Mar 2019	InterPro Scan	1 of 1	00:04:47	139	139
12:25 11 Mar 2019	Blast Taxonomy New	1 of 1	00:00:43	8	0 (Included)
12:21 11 Mar 2019	GO Mapping	1 of 1	00:00:11	1	0 (Included)
12:10 11 Mar 2019	GO Mapping	1 of 1	00:10:22	1	0 (Included)
12:09 11 Mar 2019	Blast Taxonomy New	12 of 12	00:02:50	1,867	0 (Included)
12:07 11 Mar 2019	Blast Taxonomy New	6 of 6	00:02:05	656	0 (Included)
11:36 11 Mar 2019	Blast Taxonomy New	6 of 6	00:01:40	622	0 (Included)
11:36 11 Mar 2019	Blast Taxonomy New	12 of 12	00:02:44	1,097	0 (Included)
10:55 11 Mar 2019	Blast Taxonomy New	10 of 12	00:18:12	3,129	0 (Included)
10:55 11 Mar 2019	Blast Taxonomy New	6 of 6	00:18:10	1,576	0 (Included)
10:30 11 Mar 2019	Pfam Scan	16 of 16	00:21:00	8,202	0 (Included)
10:06 11 Mar 2019	Pfam Scan	16 of 16	00:17:22	7,632	0 (Included)
10:03 11 Mar 2019	Pfam Scan	0 of 6	00:00:00	0	0 (Included)
13:30 08 Mar 2019	KRAKEN	12 of 12	00:24:30	3,894	0 (Included)
13:17 08 Mar 2019	abyss	1 of 1	00:17:11	2,232	0 (Included)
12:36 08 Mar 2019	Blast Taxonomy New	21 of 21	00:01:13	302	0 (Included)
12:21 08 Mar 2019	Blast Taxonomy New	20 of 20	00:15:25	553	0 (Included)
11:03 08 Mar 2019	EggNOG Mapper	1 of 1	00:28:10	2,744	0 (Included)
11:03 08 Mar 2019	EggNOG Mapper	1 of 1	00:16:18	2,736	0 (Included)
10:50 08 Mar 2019	Pfam Scan	16 of 16	00:12:01	7,341	0 (Included)
10:50 08 Mar 2019	Pfam Scan	16 of 16	00:14:57	9,504	0 (Included)

Figure 1: Cloud Usage

The following features that run on the BioBam Bioinformatics Cloud Platform are:

Functional Analysis:

- CloudIPS
- CloudBlast
- GO Mapping (free)
- EggNOG Annotation (free)

Transcriptomics:

- Trinity (free)
- STAR (free)
- RSEM (free)
- EdgeR (free)
- maSigPro (free)
- NOISeq (free)

Genome Analysis:

- RepeatMasker (free)
- ABySS (free)
- Augustus (free)
- Glimmer (free)

Metagenomics:

- Kraken (free)
- MEGAHIT (free)
- meta-SPAdes (free)
- FragGeneScan (free)
- Prodigal (free)
- PfamScan (free)
- EggNOG Mapper (free)

Apps:

- CPAT (free)
- PSORTb (free)

14 Help Menu

At the Help Menu, you can find this Manual, OmicsBox papers and information of the authors. In case of a bug or a malfunction of OmicsBox you can save the log file and send it to support@biobam.com[162] or via the priority support.

- App Manager(see page 312): This option allows you to install/ uninstall Apps available on OmicsBox website https://www.biobam.com/omicsbox-apps/.
- Send Support Mail: Send an email to support with the log file already attached.
- Save Log to File.
- Startup Announcement
- Feedback
- User Manual: Opens a link with the user manual.
- Download Example Data: This option allows to download example data for each module to your computer.
- Account Information: Provides the information of the user account (available modules, subscription type, etc).
- About OmicsBox: Provides information of OmicsBox, Java and the computer where OmicsBox is installed

[162] mailto:support@blast2go.com

15 Troubleshooting

Content of this page:

15.1 How to Change the Tempfolder Location of OmicsBox

OmicsBox can handle big data-sets, but it needs plenty of free disk space in the systems temporary files folder. If you run into troubles with shortage of disk space, follow this guide to manually change the location of the folder for OmicsBox temporary files to e.g. another partition. The preferences dialog shown on the right allows you to change the temporary folder location, please make sure to restart OmicsBox after changing this setting in order to make it effective.

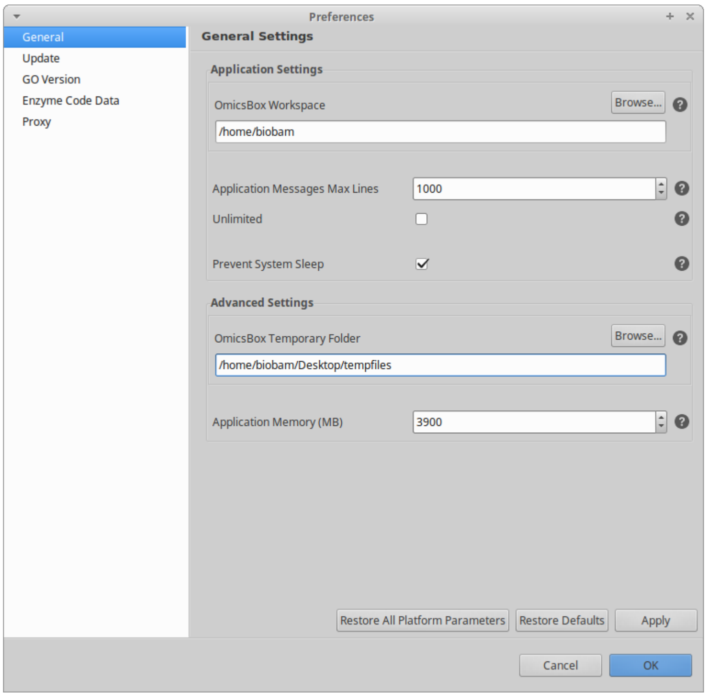
For further information on how to change the tempfolder location of OmicsBox, please visit: https://www.biobam.com/how-to-change-blast2go-tempfolder/.

15.2 How to Set Maximum Memory for OmicsBox

OmicsBox automatically reserves up to 50% of the available system memory (i.e. RAM), which is fine in most cases. However, sometimes and especially on Linux, this can lead to excessive disk swapping and will slow down OmicsBox. Follow this guide to manually change the maximum amount of memory OmicsBox will be allowed to use. The preferences dialog shown on the right allows you to change the temporary folder location, please make sure to restart OmicsBox after changing this setting in order to make it effective.

On Linux distributions like Ubuntu it is recommendable to change the vm.swappiness parameter from 60 to 1: https://help.ubuntu.com/community/SwapFaq#What_is_swappiness_and_how_do_I_change_it.3F.

For further details, please visit: https://www.biobam.com/how-to-change-blast2go-heap-memory/.



General Settings

Application Settings

OmicsBox Workspace: /home/biobam

Application Messages Max Lines: 1000

Unlimited ☐

Prevent System Sleep ☑

Advanced Settings

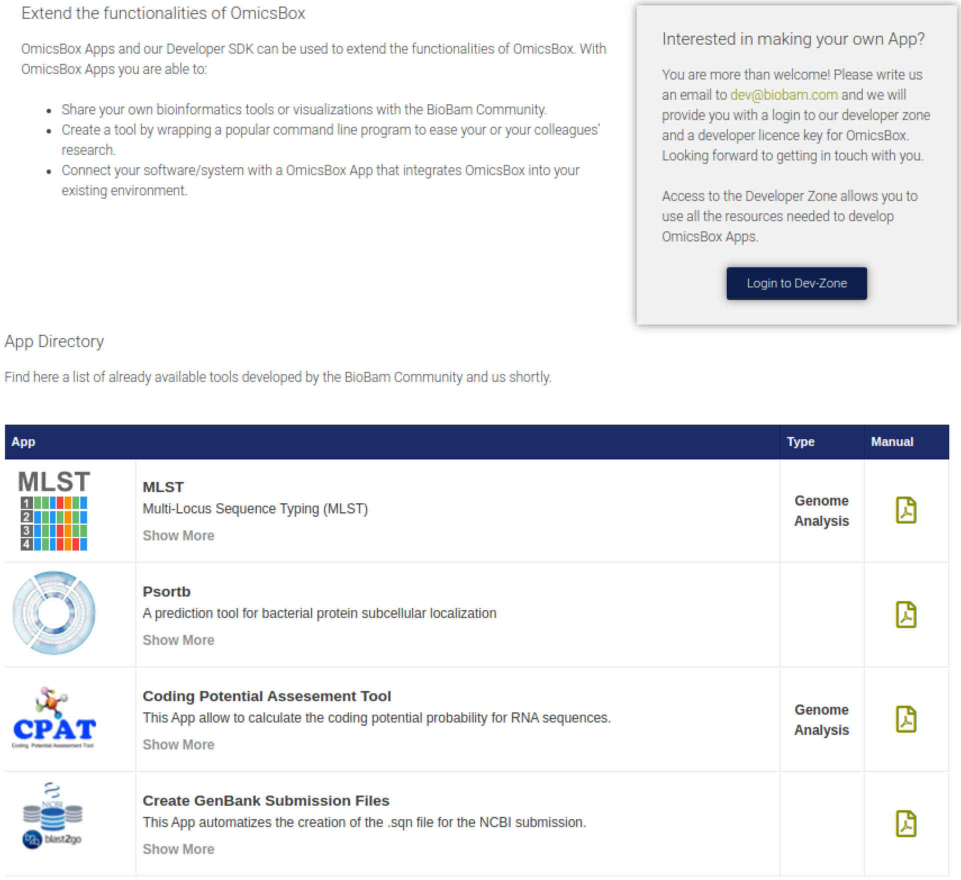
OmicsBox Temporary Folder: /home/biobam/Desktop/tempfiles

Application Memory (MB): 3900

16 Apps

Content of this section:

All available OmicsBox Apps are listed online at this page:



Extend the functionalities of OmicsBox

OmicsBox Apps and our Developer SDK can be used to extend the functionalities of OmicsBox. With OmicsBox Apps you are able to:

- Share your own bioinformatics tools or visualizations with the BioBam Community.
- Create a tool by wrapping a popular command line program to ease your or your colleagues' research.
- Connect your software/system with a OmicsBox App that integrates OmicsBox into your existing environment.

Interested in making your own App?

You are more than welcome! Please write us an email to dev@biobam.com and we will provide you with a login to our developer zone and a developer licence key for OmicsBox. Looking forward to getting in touch with you.

Access to the Developer Zone allows you to use all the resources needed to develop OmicsBox Apps.

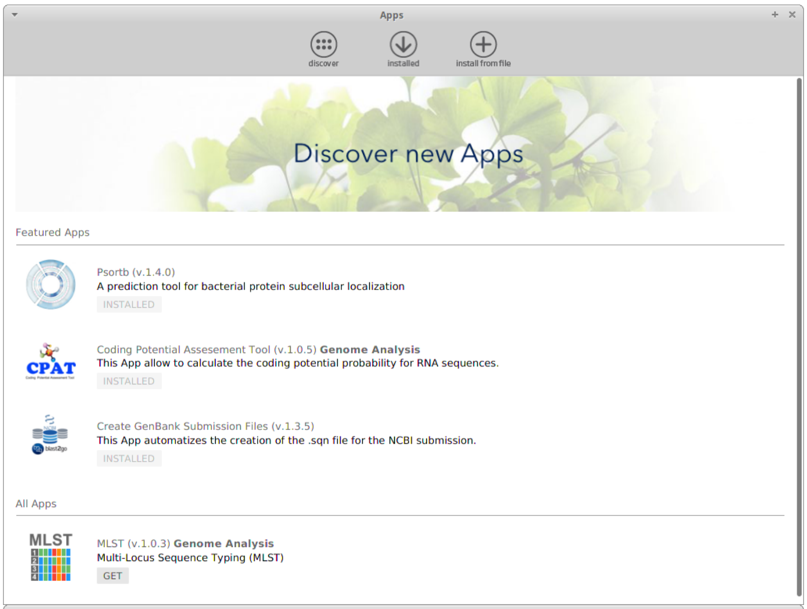
App Directory

Find here a list of already available tools developed by the BioBam Community and us shortly.

App		Type	Manual
	MLST Multi-Locus Sequence Typing (MLST) Show More	Genome Analysis	📄
	Psortb A prediction tool for bacterial protein subcellular localization Show More		📄
	Coding Potential Assesement Tool This App allow to calculate the coding potential probability for RNA sequences. Show More	Genome Analysis	📄
	Create GenBank Submission Files This App automatizes the creation of the .sqn file for the NCBI submission. Show More		📄

https://www.biobam.com/omicsbox-apps/

Apps can be viewed, installed, updated and uninstalled from within OmicsBox (**Help → App Manager**):



Discover new Apps

Featured Apps

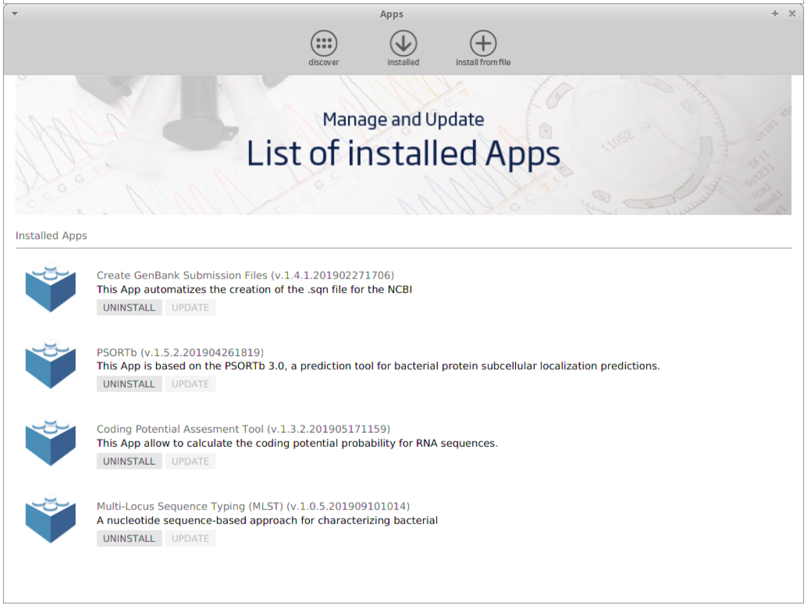
Psortb (v.1.4.0)
A prediction tool for bacterial protein subcellular localization
INSTALLED

Coding Potential Assesement Tool (v.1.0.5) **Genome Analysis**
This App allow to calculate the coding potential probability for RNA sequences.
INSTALLED

Create GenBank Submission Files (v.1.3.5)
This App automatizes the creation of the .sqn file for the NCBI submission.
INSTALLED

All Apps

MLST (v.1.0.3) **Genome Analysis**
Multi-Locus Sequence Typing (MLST)
GET



Manage and Update
List of installed Apps

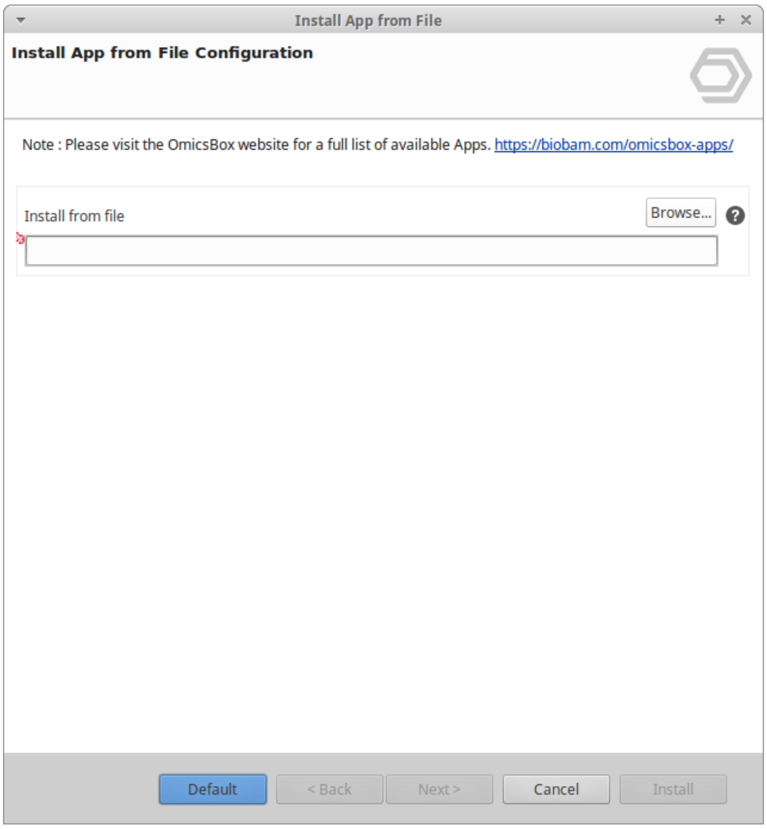
Installed Apps

Create GenBank Submission Files (v.1.4.1.201902271706)
This App automatizes the creation of the .sqn file for the NCBI
UNINSTALL UPDATE

PSORTb (v.1.5.2.201904261819)
This App is based on the PSORTb 3.0, a prediction tool for bacterial protein subcellular localization predictions.
UNINSTALL UPDATE

Coding Potential Assesment Tool (v.1.3.2.201905171159)
This App allow to calculate the coding potential probability for RNA sequences.
UNINSTALL UPDATE

Multi-Locus Sequence Typing (MLST) (v.1.0.5.201909101014)
A nucleotide sequence-based approach for characterizing bacterial
UNINSTALL UPDATE



Install App from File

Install App from File Configuration

Note : Please visit the OmicsBox website for a full list of available Apps. https://biobam.com/omicsbox-apps/

Install from file

16.1 Coding Potential (Genome Analysis)

Content of this page:

16.1.1 Introduction

Thanks to the Next Generation Sequencing methods, transcriptomes are becoming more and more abundant. Once the transcripts have been assembled and we dispose of the sequences that have been transcribed into RNA, we must distinguish between the transcripts that will be coding (mRNA) and the non-coding ones (ncRNA). This classification can be done assigning to each transcript a score based on his nucleotide composition and patterns.

16.1.1.1 Coding Potential Assessment Tool

The "Coding Potential Assessment Tool" provides an easy and fast way to classify the transcripts according to their coding score. This tool integrates the CPAT[163] algorithm within OmicsBox. The CPAT algorithm needs of models in order to assign the coding potential scores to each sequence. OmicsBox incorporates the standard CPAT models and adds some of the most common organisms models used on molecular biology. In addition to the prebuilt models, this tool adds the option to create your species-specific model.

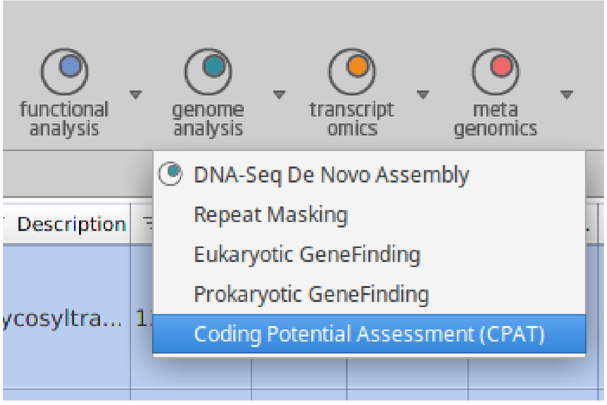


Figure 1: Coding Potential Assessment Tool in the OmicsBox Analysis Menu

16.1.2 Run Coding Potential Assessment Tool

This tool can be found under **Genome Analysis** → **Coding Potential Assessment (CPAT).** The wizard allows adjusting analysis parameters (Figure 3[164]).

- **Accuracy:** By default, the accuracy is set automatically in order to reduce the false positives and the false negatives, this means that the threshold equals to the value where the sensitivity has the same value than the specificity.

  If higher accuracy is desired the accuracy can be set manually. Raising the accuracy will allow classifying the sequences into three categories: coding, non-coding, and transcripts with unknown coding potential (Figure 2[165]).

  > ⚠ The accuracy can be set manually but it can never be lower than the default value. In this case, the accuracy value will automatically fallback to the default value.

[163] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3616698/

[164] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=673284303#CodingPotential(GenomeAnalysis)-figure3

[165] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=673284303#CodingPotential(GenomeAnalysis)-figure2

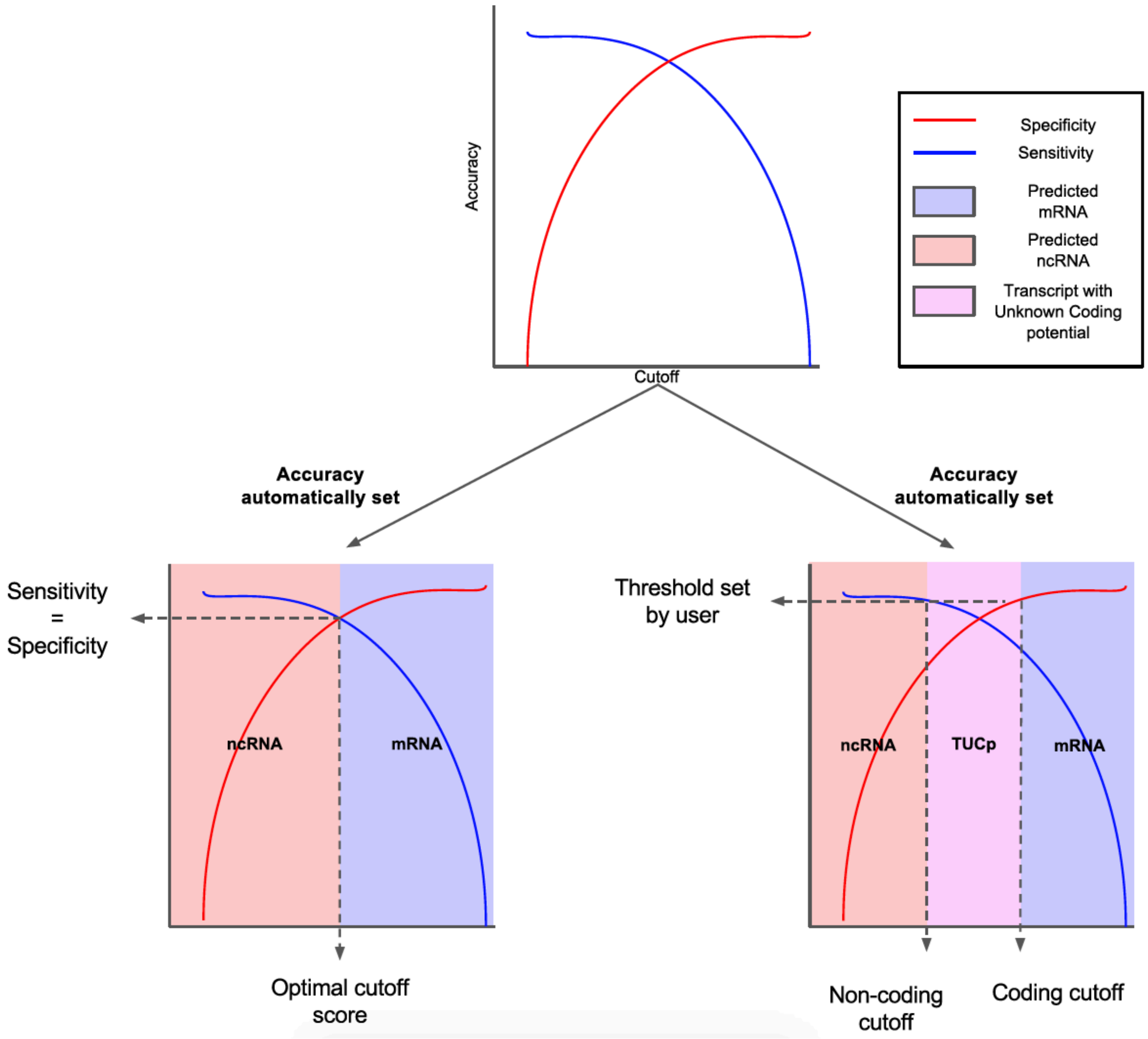


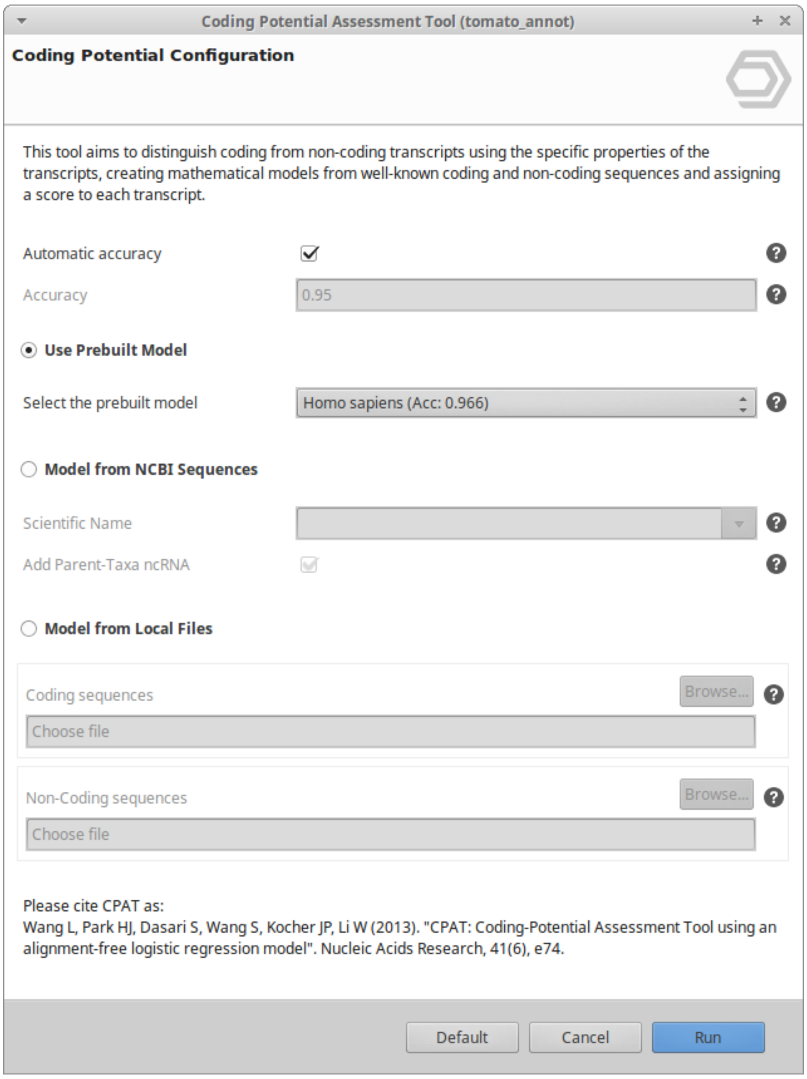
Figure 2: Accuracy: Interpretation of the double ROC Curve

- **Models:** The algorithm needs models to calculate the coding potential for each transcript. Here we can choose the origin of these models:
  - Prebuilt: Use one of the prebuilt models available. Selecting one of these prebuilt models, the algorithm will run faster.

Species	Accuracy	Coding Cutoff
Arabidopsis thaliana	0.984	0.415
Bos Taurus	0.953	0.359
Caenorhabditis elegans	0.998	0.523
Danio rerio	0.984	0.38
Drosophila melanogaster	0.963	0.39
Gallus gallus	0.93	0.402
Homo sapiens	0.966	0.364
Mus musculus	0.955	0.440
Rattus norvegicus	0.98	0.363
Sus scrofa	0.946	0.467
Xenopus laevis	0.963	0.415

- From files: Create the model providing 2 FASTA files; one with coding sequences and another one with non-coding sequences.
- From NCBI sequences: Create a new species-specific model from the sequences available on the NCBI database by selecting his scientific name or ID on the search box. A minimum of 1000 non-coding and coding sequences are required.

> ⚠ Note: Checking the `Get Parent-Taxa ncRNA` allows to use non-coding RNA sequences from higher parent taxa until complete the 1000 necessary non-coding sequences.



Coding Potential Assessment Tool (tomato_annot)

Coding Potential Configuration

This tool aims to distinguish coding from non-coding transcripts using the specific properties of the transcripts, creating mathematical models from well-known coding and non-coding sequences and assigning a score to each transcript.

Automatic accuracy ☑

Accuracy 0.95

● Use Prebuilt Model

Select the prebuilt model: Homo sapiens (Acc: 0.966)

○ Model from NCBI Sequences

Scientific Name

Add Parent-Taxa ncRNA ☑

○ Model from Local Files

Coding sequences: Choose file

Non-Coding sequences: Choose file

Please cite CPAT as:
Wang L, Park HJ, Dasari S, Wang S, Kocher JP, Li W (2013). "CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model". Nucleic Acids Research, 41(6), e74.

Figure 3: Wizard Page

16.1.3 Results

Once finished three result types are automatically created:

- **Coding Potential Table:** Here you can see the results for each sequence:
  - Tag: The tag marking for each sequence whether it is a coding, non-coding or unknown coding potential transcript.
  - Sequence: The name of the sequence.
  - mRNA size: The length of the original transcript.
  - ORF size: The size of the potential ORF within the sequence.
  - Fickett score: The Fickett score which is a linguistic feature that distinguishes protein-coding RNA and ncRNA according to the combinational effect of nucleotide composition and codon usage bias.
  - Hexamer score: The hexamer score is calculated using a log-likelihood ratio to measure differential hexamer usage between coding and noncoding sequences.

    - Coding Probability: The coding probability assigned to each transcript.
- **Pie Chart:** The coding potential distribution is shown as a pie chart of the classification results for the corresponding sequences depending on the provided cutoffs (Figure 4[166]).
- **Model Accuracy via a double ROC-Curve chart:** This chart opens when a new model is created or when the accuracy is manually set. In this chart, we can check the quality, the accuracy and the different thresholds of a model (Figure 5[167]).

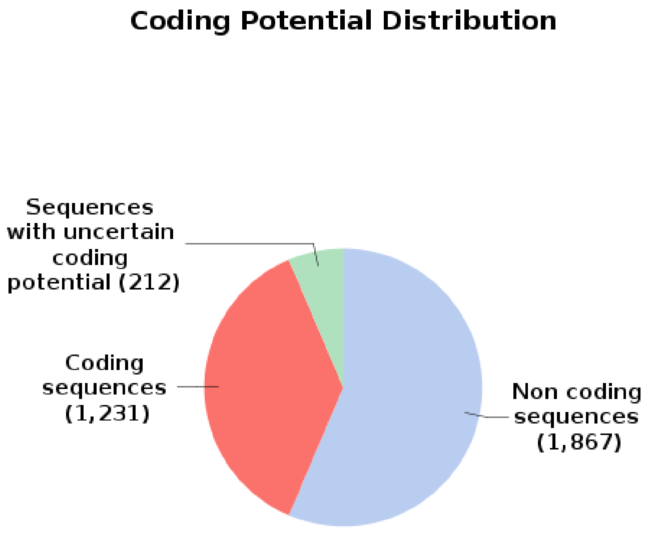


Figure 4: Distribution of the coding potential

[166] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=673284303#CodingPotential(GenomeAnalysis)-figure4

[167] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=673284303#CodingPotential(GenomeAnalysis)-figure5

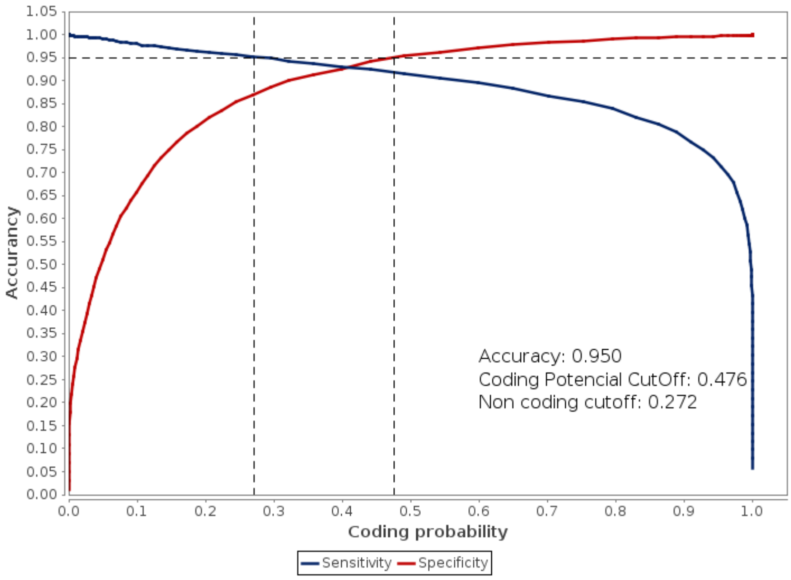


Figure 5: Double ROC curve showing the model quality, accuracy and threshold

16.2 NCBI Submission

Content of this page:

16.2.1 NCBI Database

The most important source of new data for GenBank is direct submissions from scientists. GenBank depends on its contributors to help to keep the database as comprehensive, current, and accurate as possible. NCBI provides timely and accurate processing and biological review of new entries, updates existing entries and assists authors with the submission of new data.

16.2.2 NCBI Submission Submission Tool

The NCBI data submission tool (**General Tools** → **Create NCBI GenBank Genome Submission Files**) facilitates the creation of a GenBank ready for submission. The tool combines a reference genome (fasta file), the gene coordinates (gff file) and the functional annotations of OmicsBox, creating a feature table which will be validated with the tbl2asn program. The `tbl2asn' command-line program is used to automate the creation of sequence records (.sqn files). For further information about tbl2asn visit: http://www.ncbi.nlm.nih.gov/genbank/tbl2asn2.

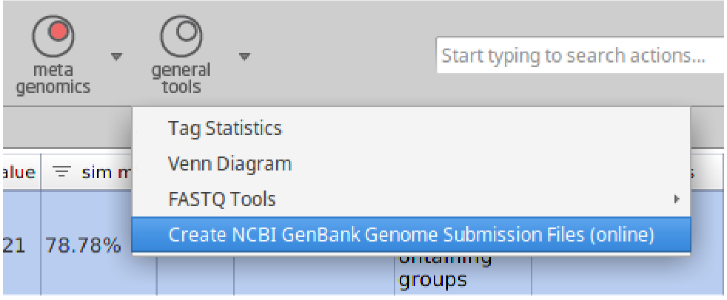


Figure 1: NCBI Submission Tool

⚠ Notes:

- This tool requires an internet connection to execute the tbl2asn program, it requires a connection to the NCBI databases in order to validate the annotations.

16.2.3 General Workflow

To successfully submit the annotated sequences, it is first necessary to prepare the source of the annotations, i.e. the reference genome to which the sequences belong, the position on the genome, and the functional annotation. These files are processed by OmicsBox and validated by `tbl2asn' to create the ASN1 file (.sqn) and the validation files (Figure 2(see page 321)).

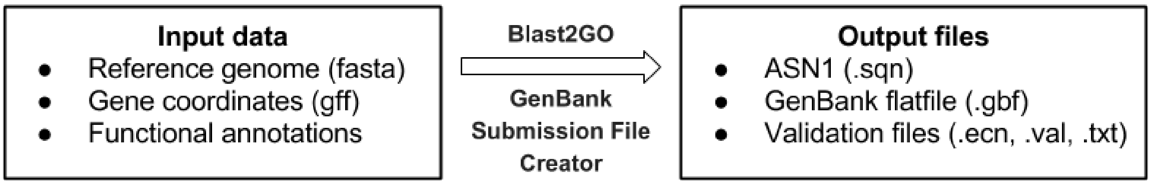


Figure 2: General Workflow

16.2.3.1 Input Files

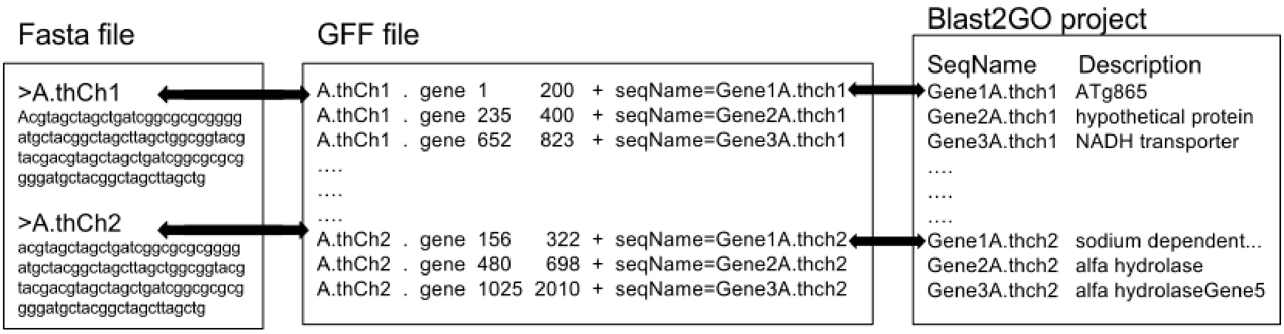
Three elements are necessary to create the submission files:

- **Reference genome:** This file provides the organisms nucleotide sequence and may contain one or more chromosomes The chromosome names in the fasta description line have to match the GFF file name(s).
- **Genomic annotation:** This data is provided by the GFF3 files, and is also used to link the annotations and the genome reference sequences in the Fasta file. The sequence names used in the OmicsBox project should appear in the feature column in the GFF3 file. The corresponding feature ID can be specified as parameter (default is seqName).
- **Functional annotation:** This information is provided by your OmicsBox project, and is intended to provide the functional features of your sequences, including gene names, Gene Ontology terms and enzyme numbers. The option to create the submission file is only activated when a OmicsBox Project file is loaded and selected.

The sequence name of the functional annotation in your OmicsBox project has to match with a feature of your choice in the gff file.

16.2.3.2 Preparing Your Data

In order to integrate all the information and to create the NCBI submission files, we need to create informative links between them. As discussed above, the GFF3 files act as a link between the genome sequence and the functional annotation (Figure 3(see page 322)). The sequence name used in the OmicsBox project should appear in the feature column in the GFF3 file (attributes field). The corresponding feature ID can be specified as a parameter (for the GFF files created by Augustus and Glimmer included in OmicsBox, the IDs correspond to `seqName' by default).



```
>A.thCh1
Acgtagctagctgatcggcgcgcgggg
atgctacggctagcttagctggcggtacg
tacgacgtagctagctgatcggcgcgcg
gggatgctacggctagcttagctg

>A.thCh2
acgtagctagctgatcggcgcgcggggg
atgctacggctagcttagctggcggtacg
tacgacgtagctagctgatcggcgcgcg
gggatgctacggctagcttagctg
```

```
A.thCh1 . gene 1     200  + seqName=Gene1A.thch1
A.thCh1 . gene 235   400  + seqName=Gene2A.thch1
A.thCh1 . gene 652   823  + seqName=Gene3A.thch1
....
....
....
A.thCh2 . gene 156   322  + seqName=Gene1A.thch2
A.thCh2 . gene 480   698  + seqName=Gene2A.thch2
A.thCh2 . gene 1025  2010 + seqName=Gene3A.thch2
```

SeqName	Description
Gene1A.thch1	ATg865
Gene2A.thch1	hypothetical protein
Gene3A.thch1	NADH transporter
....	
....	
....	
Gene1A.thch2	sodium dependent...
Gene2A.thch2	alfa hydrolase
Gene3A.thch2	alfa hydrolaseGene5

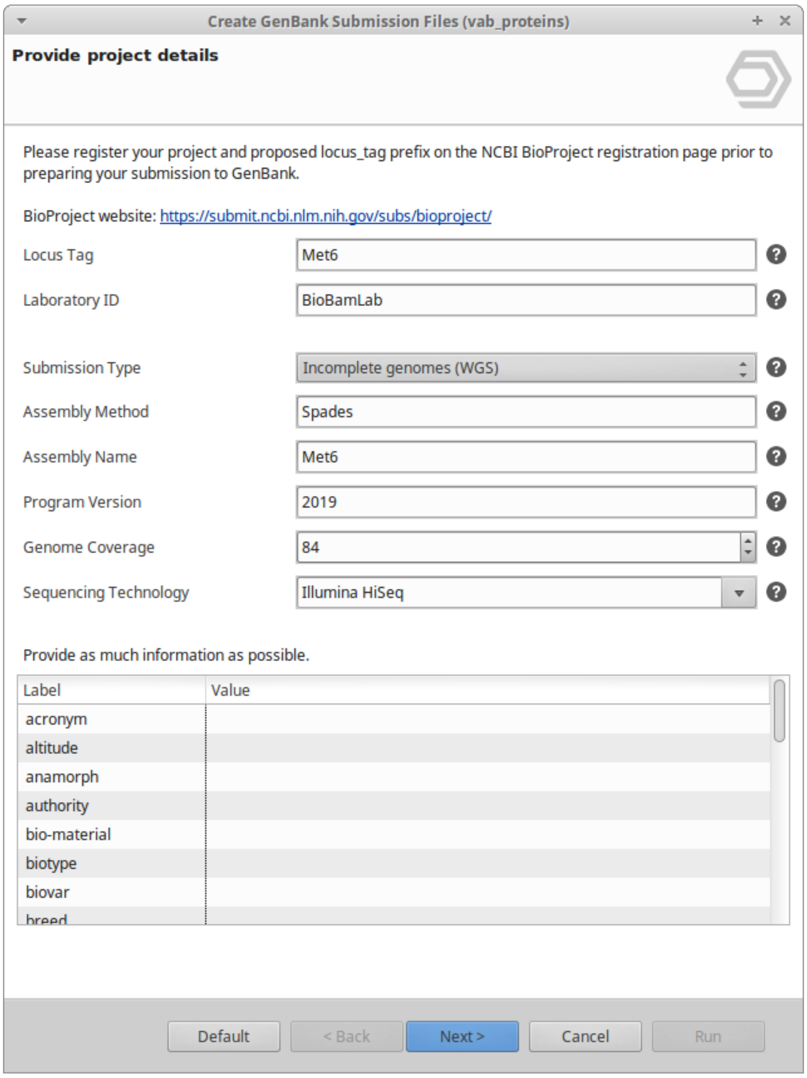
Figure 3: Information Integration Scheme

16.2.4 Wizard Pages

16.2.4.1 Page 1: Project Details

- **Locus tag:** The `locus tag' is an alphanumeric identifier of your project provided by the NCBI or user determined at the moment of the BioProject registration at: https://submit.ncbi.nlm.nih.gov/subs/bioproject
- **Laboratory ID:** The laboratory ID is a unique tag that refers to your own laboratory and allows the sequences to be associated with it.
- **Submission type:** Here you can choose the type of submission you want to perform.

  - **One or a few nucleotide sequences:** use this option if you have a small data-set containing few sequences (less than a chromosome, or a chromosome on scaffold stage).
  - **Complete eukaryotic genomes or chromosomes:** use this option if you have a complete data set without N's, conforming a chromosome or a whole genome.
  - **Incomplete genomes (WGS):** use this option if your data-set consists of incomplete genomic or chromosomal assemblies derived from shotgun sequencing methods.
- **Assembly details (only for WGS submission):** These details provide information about the more technical steps of the assembly. Here we can find:
  - **Assembly method:** The program or algorithms used to assemble the genome.
  - **Assembly name:** This is a short project identifier.
  - **Long assembly name:** This is a larger and more explanatory name of your project.
  - **Genome coverage:** This is the mean genome coverage obtained by the assembler and has a general format of one or more digits followed by an `x' (e.g: 12x or 76x).
  - **Sequencing technology:** The name of the technology used to perform the sequencing of the query genome. If the technology used is not in the list shown, you can manually enter the name.
- **Optional source qualifiers:** These are additional sequence qualifiers to your all project, specifications of optional qualifiers allows you to add useful information regarding the organism chromosome, type, etc. If you are going to submit a WGS project, add the source information as the organism and the relevant strain, breed, cultivar or isolate, if exists for the sequenced organism.



Provide project details

Please register your project and proposed locus_tag prefix on the NCBI BioProject registration page prior to preparing your submission to GenBank.

BioProject website: https://submit.ncbi.nlm.nih.gov/subs/bioproject/

Locus Tag	Met6
Laboratory ID	BioBamLab
Submission Type	Incomplete genomes (WGS)
Assembly Method	Spades
Assembly Name	Met6
Program Version	2019
Genome Coverage	84
Sequencing Technology	Illumina HiSeq

Provide as much information as possible.

Label	Value
acronym	
altitude	
anamorph	
authority	
bio-material	
biotype	
biovar	
breed	

Figure 4: Information Integration Scheme

⚠ Notes:

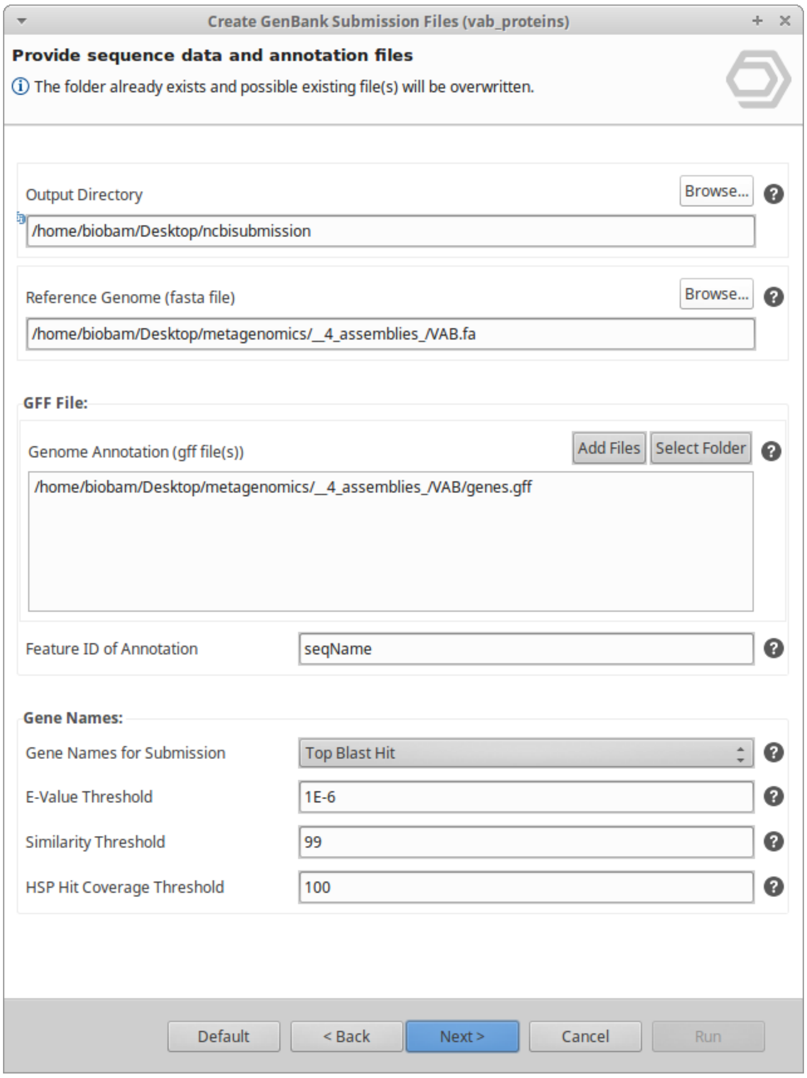
- The `gcode' corresponding to the genetic code is only mandatory if the submitting organism is not specified or is not in the NCBI Taxonomy Browser.

16.2.4.2 Page 2: Sequence Data and Annotation Files

- **Output Directory:** The creation and validation of the submitting sequences will produce multiple files that may be checked. This option allows the files to be saved in an existing folder, or to create a new one.
- **Fasta File:** The reference genome is the FASTA or multi-FASTA file containing the sequences to be submitted. This tool is designed to submit complete eukaryotic genomes or chromosomes. If you are submitting a single complete chromosome it must be in a single fasta entry, however, if you are

submitting a complete genome, you must have a single entry for each chromosome. Important note: if this is a complete genome or chromosome submission, remove all the ‘Ns’ present in the fasta file.

- **Genome annotation:** The genome annotation refers to the .gff file containing the gene coordinates for each annotated gene. This file must be named according to the data entry to which it corresponds.
- **Feature ID:** The feature ID of annotation refers to the flag on the ninth column of the gff file, which contains the name of the sequence, displayed as SeqName in OmicsBox.
- **Gene names:** Here you can choose how to assign the names for your annotated sequences, the options are: “hypothetical protein”, the “SeqName” assigned in the OmicsBox project or assign the name of the “Top BLAST Hit”. If this last option is selected, you can set the threshold for:
  - **E-value:** The minimum E-value obtained in the BLAST between the top BLAST hit and your query (default value is 1E-6).
  - **Similarity:** The minimum percent similarity between the two sequences (0-100).
  - **Coverage:** The minimum percent coverage between the two sequences (0-100).



Create GenBank Submission Files (vab_proteins)

Provide sequence data and annotation files

The folder already exists and possible existing file(s) will be overwritten.

Output Directory: /home/biobam/Desktop/ncbisubmission

Reference Genome (fasta file): /home/biobam/Desktop/metagenomics/__4_assemblies_/VAB.fa

GFF File:

Genome Annotation (gff file(s)): /home/biobam/Desktop/metagenomics/__4_assemblies_/VAB/genes.gff

Feature ID of Annotation: seqName

Gene Names:

Gene Names for Submission: Top Blast Hit

E-Value Threshold: 1E-6

Similarity Threshold: 99

HSP Hit Coverage Threshold: 100

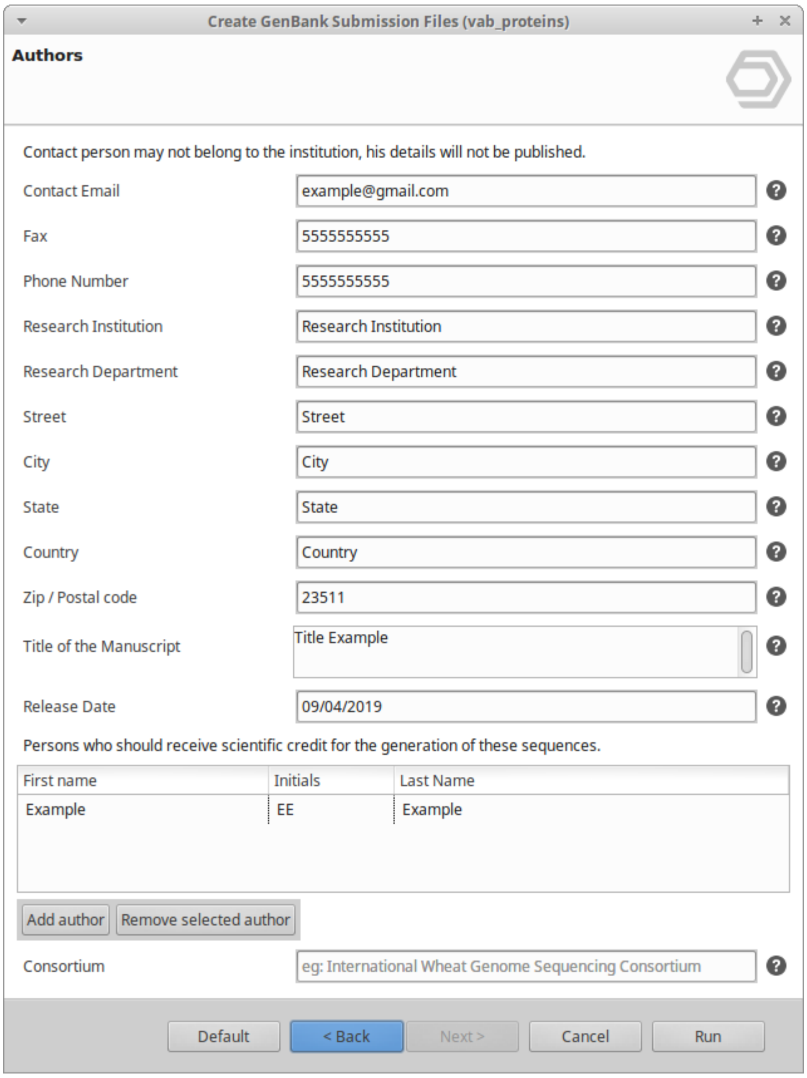
Figure 5: Sequence Data and Annotation Files page

⚠ Notes:

- If the threshold is not reached, the name of the gene will be “hypothetical protein”.
- All manual gene names annotations has higher priority.

16.2.4.3 Page 3: Author’s and Affiliation data

- **Contact data:** This page allows providing contact details for the submitting person. This information will not be publicly visible, and only can be used by the NCBI staff for validation.
- **Institution data:** Information about the institution where the sequencing was performed is provided here.
- **Title of the manuscript:** This title is provisional and can be modified at any time via email request to the NCBI.
- **Release date:** This is the date when your submitted and validated data will be accessible in the NCBI database. If the release date is The same day or before the submission, it will be automatically available once the data is validated.
- **Names and Initials:** Insert the names and initials of the individuals who must receive scientific credit for the generation of the sequences and annotations in this submission. If the authors are part of a consortium, it is not necessary that they appear as individual authors, as they are represented in the ‘Consortium’ option



Contact person may not belong to the institution, his details will not be published.

Contact Email	example@gmail.com
Fax	5555555555
Phone Number	5555555555
Research Institution	Research Institution
Research Department	Research Department
Street	Street
City	City
State	State
Country	Country
Zip / Postal code	23511
Title of the Manuscript	Title Example
Release Date	09/04/2019

Persons who should receive scientific credit for the generation of these sequences.

First name	Initials	Last Name
Example	EE	Example

Consortium: eg: International Wheat Genome Sequencing Consortium

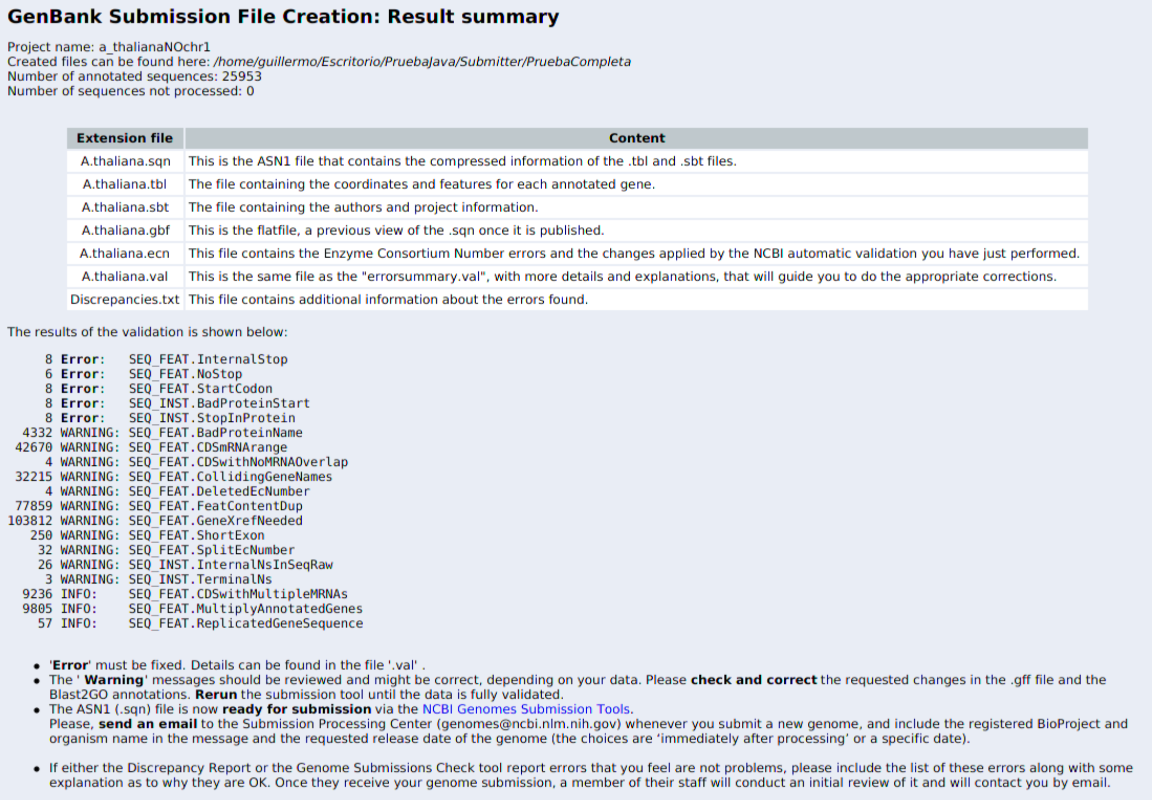
Figure 6: Author's and Affiliation page

16.2.5 Result Files

Once the input data has been analysed and processed via the tbl2asn tool, several result files are created.

- **.sqn:** This is the ASN1 file containing the compressed information of the .tbl and .sbt files.
- **.tbl:** The file containing the coordinates and the features for each annotated gene.
- **.sbt:** The file containing the authors and project information.
- **.gbf:** This is the GenBank flat file, a previous view of the .sqn once it is published.
- **.ecn:** This file contains the Enzyme Consortium Number errors and the changes applied by the previous NCBI automatic validation you have just performed.
- **.val:** This is the same file as the "errorsummary.val", with more details and explanations, that will guide you to make the appropriate corrections.
- **.txt:** This file contains additional information about the errors found.

A result page provides a summary of the different types of errors and warnings. Errors must be corrected and the warnings should be reviewed (they may actually not be harmful. depending on your data). Modifications can be made by editing the gff or the annotation in the project. Once errors are corrected, the tool should be run again, until an error-free validation is achieved. Once the submission files have no errors, the ASN1 (.sqn) file is ready for submission via the NCBI Genome Submission Tools (www.ncbi.nlm.nih.gov/projects/GenomeSubmit/genome_submit.cgi[168]). Whenever you submit a new genome, it is necessary to send an email to the Submission Processing Center (genomes@ncbi.nlm.nih.gov[169]), specifying the registered BioProject and organism name in the message as well as the requested release date for the genome.



GenBank Submission File Creation: Result summary

Project name: a_thalianaNOchr1
Created files can be found here: */home/guillermo/Escritorio/PruebaJava/Submitter/PruebaCompleta*
Number of annotated sequences: 25953
Number of sequences not processed: 0

Extension file	Content
A.thaliana.sqn	This is the ASN1 file that contains the compressed information of the .tbl and .sbt files.
A.thaliana.tbl	The file containing the coordinates and features for each annotated gene.
A.thaliana.sbt	The file containing the authors and project information.
A.thaliana.gbf	This is the flatfile, a previous view of the .sqn once it is published.
A.thaliana.ecn	This file contains the Enzyme Consortium Number errors and the changes applied by the NCBI automatic validation you have just performed.
A.thaliana.val	This is the same file as the "errorsummary.val", with more details and explanations, that will guide you to do the appropriate corrections.
Discrepancies.txt	This file contains additional information about the errors found.

The results of the validation is shown below:

```
     8 Error:   SEQ_FEAT.InternalStop
     6 Error:   SEQ_FEAT.NoStop
     8 Error:   SEQ_FEAT.StartCodon
     8 Error:   SEQ_INST.BadProteinStart
     8 Error:   SEQ_INST.StopInProtein
  4332 WARNING: SEQ_FEAT.BadProteinName
 42670 WARNING: SEQ_FEAT.CDSmRNArange
     4 WARNING: SEQ_FEAT.CDSwithNoMRNAOverlap
 32215 WARNING: SEQ_FEAT.CollidingGeneNames
     4 WARNING: SEQ_FEAT.DeletedEcNumber
 77859 WARNING: SEQ_FEAT.FeatContentDup
103812 WARNING: SEQ_FEAT.GeneXrefNeeded
   250 WARNING: SEQ_FEAT.ShortExon
    32 WARNING: SEQ_FEAT.SplitEcNumber
    26 WARNING: SEQ_INST.InternalNsInSeqRaw
     3 WARNING: SEQ_INST.TerminalNs
  9236 INFO:    SEQ_FEAT.CDSwithMultipleMRNAs
  9805 INFO:    SEQ_FEAT.MultiplyAnnotatedGenes
    57 INFO:    SEQ_FEAT.ReplicatedGeneSequence
```

- '**Error**' must be fixed. Details can be found in the file '.val' .
- The ' **Warning**' messages should be reviewed and might be correct, depending on your data. Please **check and correct** the requested changes in the .gff file and the Blast2GO annotations. **Rerun** the submission tool until the data is fully validated.
- The ASN1 (.sqn) file is now **ready for submission** via the NCBI Genomes Submission Tools.
  Please, **send an email** to the Submission Processing Center (genomes@ncbi.nlm.nih.gov) whenever you submit a new genome, and include the registered BioProject and organism name in the message and the requested release date of the genome (the choices are ‘immediately after processing’ or a specific date).
- If either the Discrepancy Report or the Genome Submissions Check tool report errors that you feel are not problems, please include the list of these errors along with some explanation as to why they are OK. Once they receive your genome submission, a member of their staff will conduct an initial review of it and will contact you by email.

Figure 7: Results summary

16.2.5.1 Most Common Errors and Warnings

- **ERROR(s):** InternalStop + StartCodon + BadProteinStart + StopInProtein: These error codes usually appear grouped, and they refer to the same sequence. This may be due to an error in the gff that has shifted its reading frame, you can correct that by changing the frame on the .gff.
- **StartCodon:** An illegal start codon was used. Some possible explanations are: (1) the wrong genetic code may have been selected; (2) the wrong reading frame may be in use; or (3) the coding region

[168] http://www.ncbi.nlm.nih.gov/projects/GenomeSubmit/genome_submit.cgi
[169] http://ncbi.nlm.nih.gov

may be incomplete at the 5’ end, in which case a partial location should be indicated. This can be fixed in the .gff file, or by selecting the correct code in the ‘source qualifiers’ on the first wizard page.
- **InternalStop:** Internal stop codons are found in the protein sequence. Some possible explanations are: (1) the wrong genetic code may have been selected; (2) the wrong reading frame may be in use; (3) the coding region may be incomplete at the 5’ end, in which case a partial location should be indicated; or (4) the CdRegion feature location is incorrect. This can be fixed in the .gff file by modifying the start of the sequence or selecting the correct code on the ‘source qualifiers’ on the first wizard page.
- **WARNING CDSmRNArange:** This error alerts you that two or more ’CDS’ features are under the same ‘mRNA’ feature, but there are not colliding. If you are working with prokaryotes, this is a feature you must fix it in the .gff file, but if working with eukaryotes, it’s a normal feature, as eukaryotic genes contain introns.
- **WARNING CDSwithNoMRNAOverlap:** This warning alerts you that a ‘CDS’ feature out of the ‘mRNA’ bounds, and should be fixed in the .gff file by extending the mRNA range.
- **WARNING BadProteinName:** The name assigned to this protein is not adequate. Remember that the protein name should not contain the names ‘hypothetical’ or ‘partial’, and must follow the Uni-Prot protein product names. Modify it in the OmicsBox project or directly in the .sqn file.
- **WARNING CollidingGeneNames:** Two gene features should not have the same name, this can be fixed in the OmicsBox project.
- **WARNING MissingMRNAproduct:** The mRNA feature indicates to a cDNA product that is not contained in the record. This must be fixed on the .gff file.
- **WARNING DuplicateInterval:** The location has identical adjacent intervals, e.g., a duplicate exon reference. This can be fixed eliminating the duplicated ‘exon’ or ‘CDS’ from the .gff file.
- **WARNING mRNAgeneRange:** An mRNA is overlapped by a gene feature, but is not completely contained by it. This can be corrected in the .gff by extending the range of the ‘mRNA’.
- **NoOrgFound:** This entry does not specify the organism that was the source of the sequence. Please enter a name for the organism on the first page of the wizard, in the ‘Optional source qualifiers’.

For more information about these errors, please refer to http://www.ncbi.nlm.nih.gov/IEB/ToolBox/C_DOC/lxr/source/errmsg/valid.msg

16.3 PSORTb

Content of this page:

16.3.1 Introduction

The PSORT principle uses the amino acid sequence information to generate an overall prediction of the protein localization sites. These rules are derived from experimental observations. For example, when analysing a gram-negative organism, possible localization sites are cytoplasm, cytoplasmic membrane, periplasm, outer membrane, and extracellular space.

OmicsBox allows assigning sub-cellular localization sites to proteins based on their amino acid sequence via PSORTb. PSORTb is an algorithm that can be applied to bacteria or archaea protein sequences and uses a probabilistic system to predict the most probable localization. Once sites are predicted, their corresponding cellular component GO terms can be merged with the already existing annotations.

16.3.2 Run

Starting with a previously loaded .box/.b2g project with PROTEIN sequences, the PSORTb tool can be found under **Functional Analysis → Run PSORTb.**

If the loaded project contains nucleotide sequences, the "*Translate Longest ORF*" tool can help to obtain the predicted protein sequences and be able to run PSORTb.

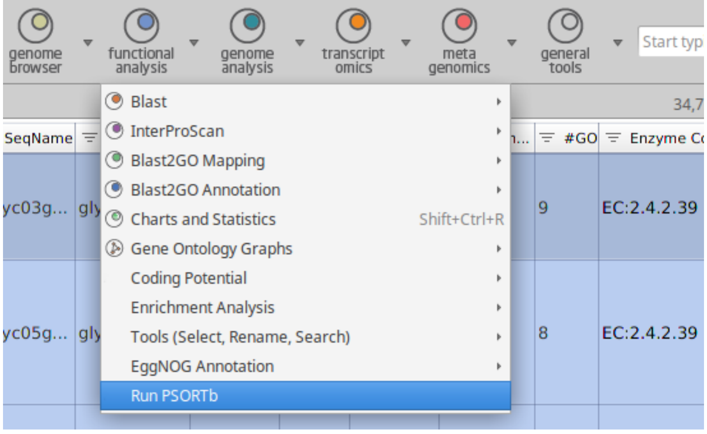


Figure 1. Run PSORTb in the Functional Analysis menu.

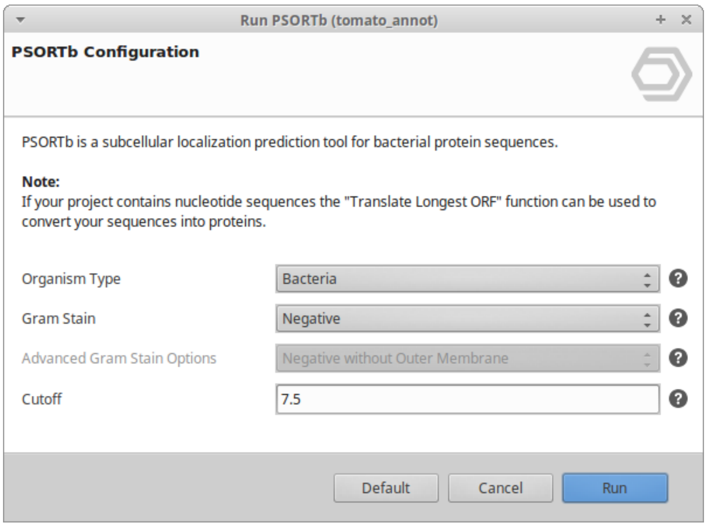
16.3.3 Wizard and parameters

The wizard allows adjusting the algorithm parameters (Figure 2).

It performs different analysis depending on the **Organism Type** and the **Gram Stain**. It can be used with bacteria positive and negative gram strains or archaea organism sequences. For more details of the core algorithm, visit psortb.org[170].

The algorithm returns score values between 0 and 10 for each localization site, the **Cutoff** parameter allows setting a minimum value of each localization above which the value can be considered as possible localization.

[170] http://www.psort.org/documentation/index.html#organism



Run PSORTb (tomato_annot)

PSORTb Configuration

PSORTb is a subcellular localization prediction tool for bacterial protein sequences.

Note:
If your project contains nucleotide sequences the "Translate Longest ORF" function can be used to convert your sequences into proteins.

Organism Type: Bacteria

Gram Stain: Negative

Advanced Gram Stain Options: Negative without Outer Membrane

Cutoff: 7.5

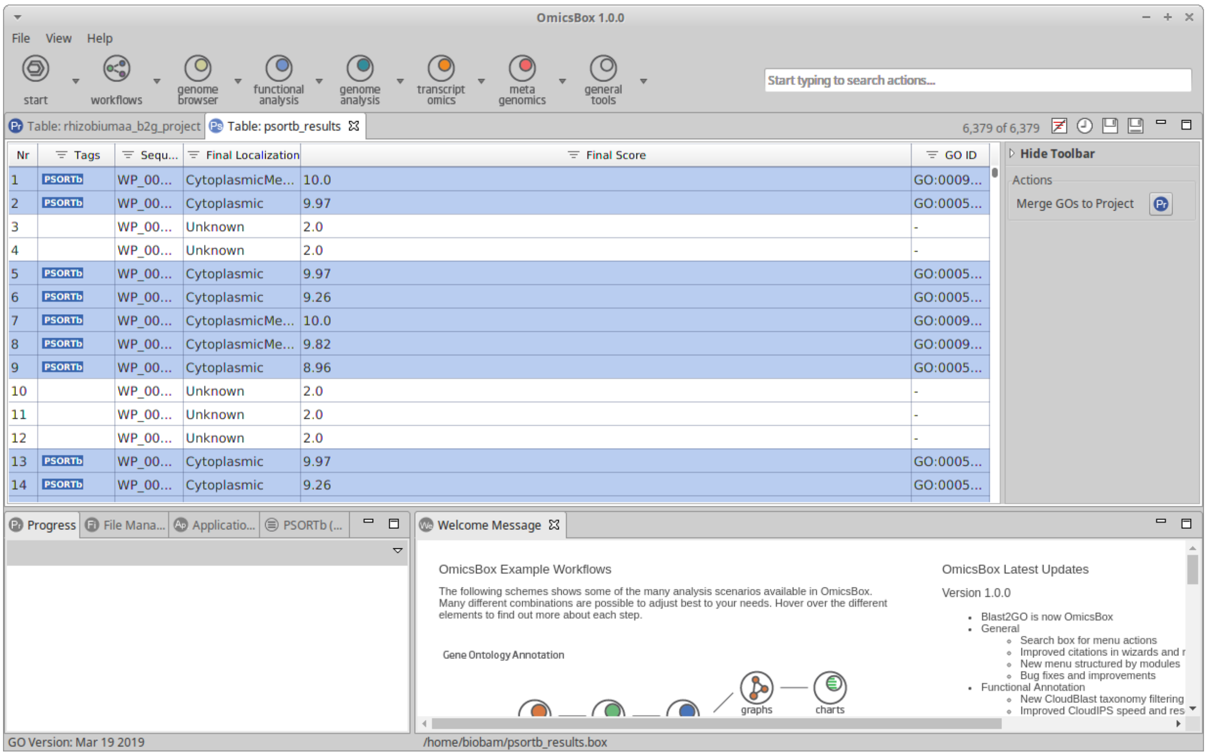
Figure 2. PSORTb wizard where the user can adjust the parameters.

16.3.4 Results

The tool will iterate over the input sequences and analyze each of them with the PSORTb 3. The process will open a new tab and as the results come back, they are shown in a table format.

The table contains one row for each sequence. The table columns are:

- **Sequence name:** shows each sequence identifier.
- **Final localization:** contains the predicted localization name.
- **Final score:** represents the prediction score for the localization.
- **GO ID:** the Gene Ontology ID associated to the location.
- **Secondary Localization:** a possible secondary localization when there is more than one score above the cutoff.
- The next 6 columns, hidden by default, show the score for all possible localizations.



Nr	Tags	Sequ...	Final Localization	Final Score	GO ID
1	PSORTb	WP_00...	CytoplasmicMe...	10.0	GO:0009...
2	PSORTb	WP_00...	Cytoplasmic	9.97	GO:0005...
3		WP_00...	Unknown	2.0	-
4		WP_00...	Unknown	2.0	-
5	PSORTb	WP_00...	Cytoplasmic	9.97	GO:0005...
6	PSORTb	WP_00...	Cytoplasmic	9.26	GO:0005...
7	PSORTb	WP_00...	CytoplasmicMe...	10.0	GO:0009...
8	PSORTb	WP_00...	CytoplasmicMe...	9.82	GO:0009...
9	PSORTb	WP_00...	Cytoplasmic	8.96	GO:0005...
10		WP_00...	Unknown	2.0	-
11		WP_00...	Unknown	2.0	-
12		WP_00...	Unknown	2.0	-
13	PSORTb	WP_00...	Cytoplasmic	9.97	GO:0005...
14	PSORTb	WP_00...	Cytoplasmic	9.26	GO:0005...

OmicsBox Example Workflows

The following schemes shows some of the many analysis scenarios available in OmicsBox. Many different combinations are possible to adjust best to your needs. Hover over the different elements to find out more about each step.

OmicsBox Latest Updates

Version 1.0.0

- Blast2GO is now OmicsBox
- General
  - Search box for menu actions
  - Improved citations in wizards and r
  - New menu structured by modules
  - Bug fixes and improvements
- Functional Annotation
  - New CloudBlast taxonomy filtering
  - Improved CloudIPS speed and res

Figure 3. PSORTb results table.

16.3.5 Merge GO information

The GO IDs from the prediction can be merged into the original Blast2GO project as cellular component characterization of the sequences.

The merge option is available in the right side panel of the PSORTb results (Figure 3).

The merge wizard asks for the OmicsBox project file where to merge the GO results and will add the GO information to the project, matching the Sequence Name. Note: The initial OmicsBox project must be saved as a file before running the Merge GOs option.

For more information regarding PSORTb, visit the psortb.org documentation page[171].

[171] http://www.psort.org/documentation/index.html

16.4 Multi-Locus Sequence Typing (MLST)

Content of this page

16.4.1 Introduction

Multi-locus sequence typing (MLST) is a useful tool for studying the genetic diversity of important public health pathogens that has provided a portable and reproducible typing system. It is a nucleotide sequence-based approach of an unambiguous procedure of characterizing isolates of bacterial species using the sequences of internal fragments of (usually) seven housekeeping genes. For this, approx. 4 450-500 bp internal fragments of each gene are used, as these can be accurately sequenced on both strands using an automated DNA sequencer. For each housekeeping gene, the different sequences present within a bacterial species are assigned as distinct alleles and, for each isolate, the alleles at each of the seven loci define the allelic profile or sequence type (ST).

For more info please click here[172].

Please cite MLST as:

Larsen MV et al. (2012). Multilocus sequence typing of total-genome-sequenced bacteria. Journal of clinical microbiology, 50(4), 1355-61.[173]

16.4.2 Run MLST

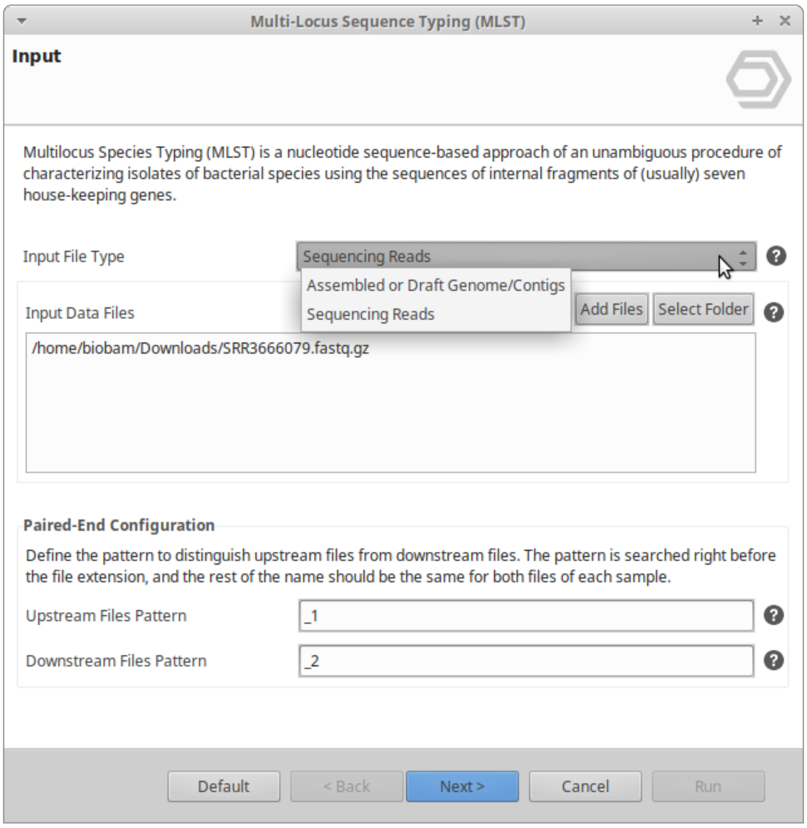
This functionality can be found under **Genome Analysis** → **Multi-locus Sequence Typing (MLST)**. The wizard allows to select files and set the parameters (Figure 1[174] and Figure 2[175]).

[172] https://pubmlst.org/general.shtml
[173] https://www.ncbi.nlm.nih.gov/pubmed/22238442
[174] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=717127699#Multi-LocusSequenceTyping(MLST)-figure1
[175] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=717127699#Multi-LocusSequenceTyping(MLST)-figure2

16.4.2.1 Input

- **Input File Type:** Choose between Assembled or Draft Genome/Contigs (FASTA) or Raw Sequencing Reads (FASTQ). Single-end or paired-end reads can be used when selecting Raw Sequencing Reads (FASTQ). Note that if paired-end is selected, two files per sample are required.
- **Input Data:** Provide the files containing sequencing reads or contigs. These files can be in FASTQ or FASTA format.
- **Paired-end configuration:** In the case of paired-end reads, the pattern to distinguish upstream files from downstream files is required. The provided patterns are searched right before the extension, and the start of the name should be the same for both files of each sample. Files whose name match with upstream and downstream patterns will be treated as paired-end data. The remaining files and those for which no partner is detected will be treated as single-end data.
  - **Upstream Files Pattern:** Establish the pattern to recognize upstream FASTQ files.
  - **Downstream Files Pattern:** Establish the pattern to recognize downstream FASTQ files.

⚠ For example, if the upstream file is named SRR3666079_1.fastq and the downstream one SRR3666079_2.fastq, you should establish "_1" as the upstream pattern and "_2" as the downstream pattern.



Multilocus Species Typing (MLST) is a nucleotide sequence-based approach of an unambiguous procedure of characterizing isolates of bacterial species using the sequences of internal fragments of (usually) seven house-keeping genes.

Input File Type: Sequencing Reads

Input Data Files: /home/biobam/Downloads/SRR3666079.fastq.gz

Paired-End Configuration

Define the pattern to distinguish upstream files from downstream files. The pattern is searched right before the file extension, and the rest of the name should be the same for both files of each sample.

Upstream Files Pattern: _1

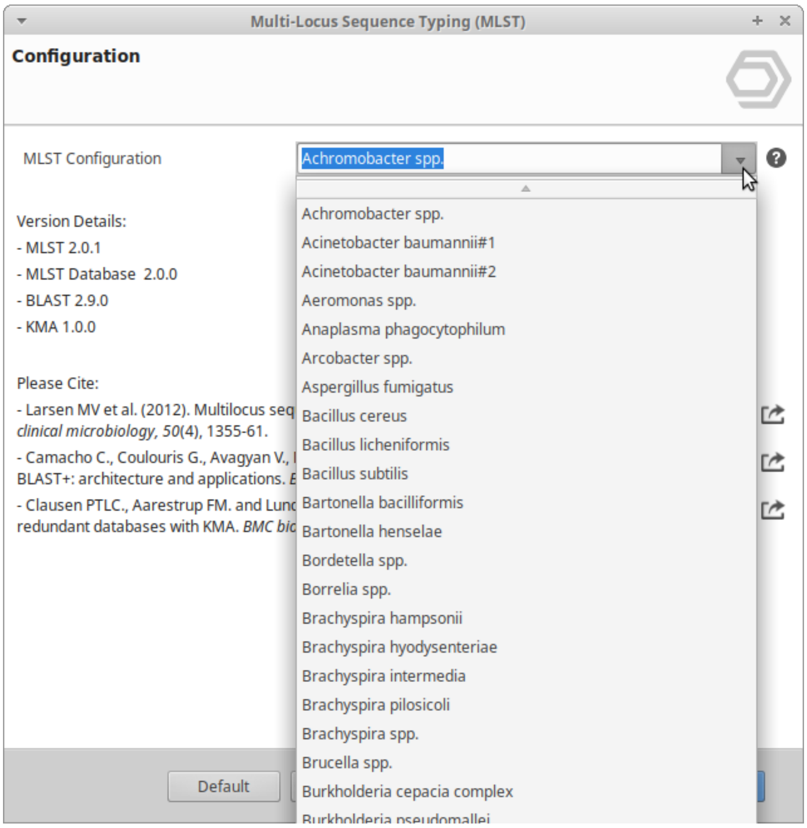
Downstream Files Pattern: _2

Figure 1: Input Data Page

16.4.2.2 Configuration

- **MLST Configuration:** Select the species database that will be used as a template for MLST prediction. If a wrong species is selected, the run may fail or the output will show no (zero) or minimal identity and coverage. MLST allele sequence and profile data are obtained from PubMLST.org[176].

> ⚠ **Configuration**
> For four organisms, two or three different MLST schemes are available. These are:
> 1. Acinetobacter baumannii: (Acinetobacter baumannii #1, Acinetobacter baumannii #2)
> 2. Escherichia coli: Escherichia coli #1, Escherichia coli #2)
> 3. Pasteurella multocida: (Pasteurella multocida #1 (RIRDC), Pasteurella multocida #2 (multihost))
> 4. Leptospira: (Leptospira #1, Leptospira #2, Leptospira #3)



Multi-Locus Sequence Typing (MLST)

Configuration

MLST Configuration: Achromobacter spp.

Version Details:
- MLST 2.0.1
- MLST Database 2.0.0
- BLAST 2.9.0
- KMA 1.0.0

Please Cite:
- Larsen MV et al. (2012). Multilocus seq *clinical microbiology*, *50*(4), 1355-61.
- Camacho C., Coulouris G., Avagyan V., BLAST+: architecture and applications.
- Clausen PTLC., Aarestrup FM. and Lunc redundant databases with KMA. *BMC bic*

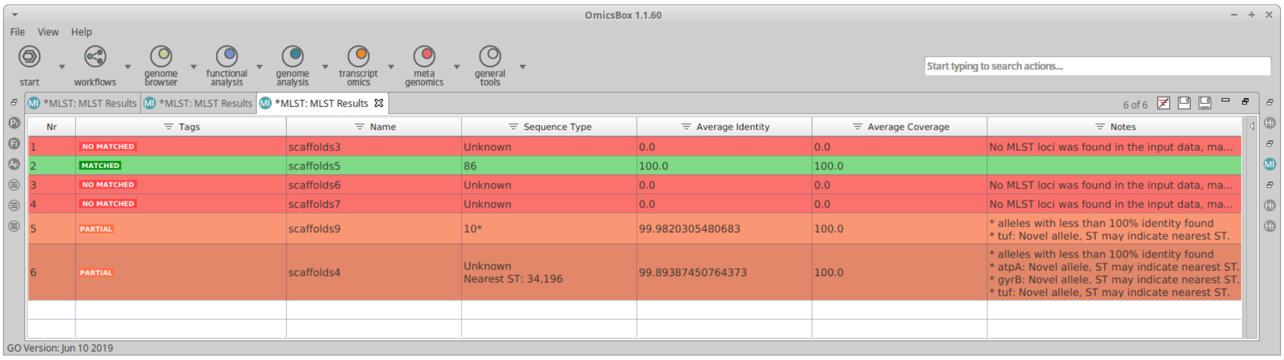
Achromobacter spp.
Acinetobacter baumannii#1
Acinetobacter baumannii#2
Aeromonas spp.
Anaplasma phagocytophilum
Arcobacter spp.
Aspergillus fumigatus
Bacillus cereus
Bacillus licheniformis
Bacillus subtilis
Bartonella bacilliformis
Bartonella henselae
Bordetella spp.
Borrelia spp.
Brachyspira hampsonii
Brachyspira hyodysenteriae
Brachyspira intermedia
Brachyspira pilosicoli
Brachyspira spp.
Brucella spp.
Burkholderia cepacia complex
Burkholderia pseudomallei

Figure 2: MLST Configuration Page

[176] https://pubmlst.org/

16.4.3 Results

When the MLST completes, it creates a sequence table containing the MLST results (Figure 3[177]). This table will contain:



Nr	Tags	Name	Sequence Type	Average Identity	Average Coverage	Notes
1	NO MATCHED	scaffolds3	Unknown	0.0	0.0	No MLST loci was found in the input data, ma...
2	MATCHED	scaffolds5	86	100.0	100.0	
3	NO MATCHED	scaffolds6	Unknown	0.0	0.0	No MLST loci was found in the input data, ma...
4	NO MATCHED	scaffolds7	Unknown	0.0	0.0	No MLST loci was found in the input data, ma...
5	PARTIAL	scaffolds9	10*	99.9820305480683	100.0	* alleles with less than 100% identity found * tuf: Novel allele, ST may indicate nearest ST.
6	PARTIAL	scaffolds4	Unknown Nearest ST: 34,196	99.89387450764373	100.0	* alleles with less than 100% identity found * atpA: Novel allele, ST may indicate nearest ST. * gyrB: Novel allele, ST may indicate nearest ST. * tuf: Novel allele, ST may indicate nearest ST.

Figure 3: Results Table Page

1. **Tags:** It contains a quick overview of the MLST result of your sample. It will generate three possible reports:
   a. **Matched:** When a complete matched was found with no errors or SNPs, therefore the average identity between all the housekeeping genes templates and the query reads/sequences, as well as the average coverage, is 100%. All samples with "Matched" results will be highlighted with green.
   b. **Partial:** When partial matches were found the average identity between all the housekeeping genes templates and the query, as well as the average coverage, is less or equal to 99%. It happens when some potential errors or SNPs have been detected. All samples with "Partial" results will be highlighted with orange.
   c. **No Matched:** The query sequence did NOT match any housekeeping gene template within the chosen MLST configuration. All samples with "No Matched" results with be highlighted with red.
2. **Name:** It displays the input file name.
3. **Sequence Type:** It contains the corresponding MLST sequence type. Please note that for all "Partial" results, the sequence type will have a number and an asterisk. This asterisk is to indicate that the Sequence Type number shown here is not a 100% match and alleles with discrepancies will be indicated in the "Note" column.

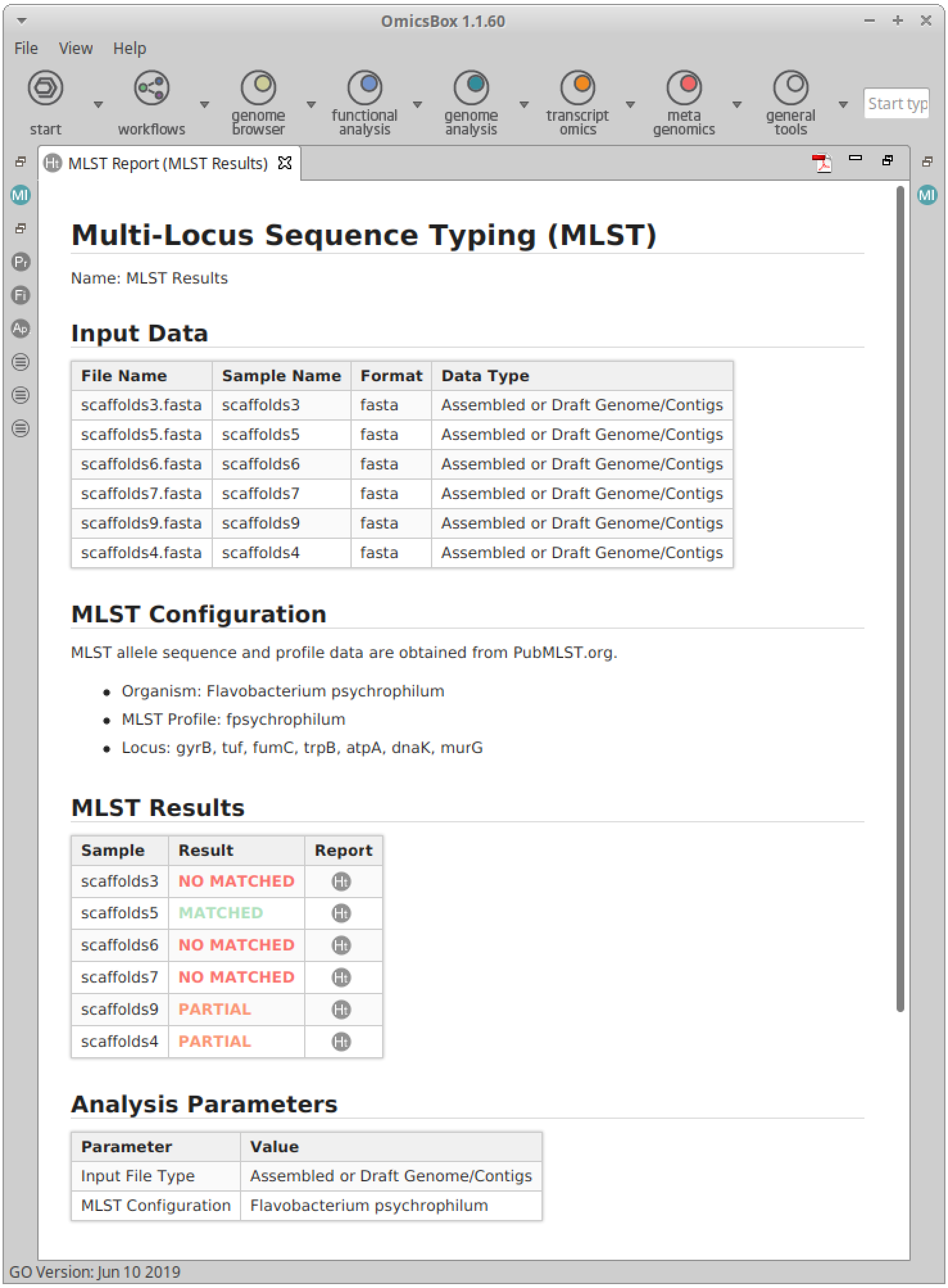
[177] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=717127699#Multi-LocusSequenceTyping(MLST)-figure3

⚠ **Sequence Type: Unknown**
Please note that "Unknown" can be the Sequence Type for samples with "Matched" result and the reason for this is that even though all the alleles in the query are matching 100% alleles in templates sequences in the database, the combination of the alleles does not have a MLST number assigned yet. In the case of samples with "Partial" results, the "Unknown" Sequence Type is not because of the discrepancies, but because the combination of the alleles does not have a MLST number assigned yet either. Lastly, for samples with "No Matched" results, the "Unknown" Sequence Type is because there was not an MLST loci that matched with the input data. This is most common when the wrong MLST scheme was chosen in the MLST configuration wizard page.
Some "Unknown" results could also report a "Nearest ST..." if there is enough coverage and identity found between the query reads/sequences and any template sequence in the database. Figure 3 shows an example of this case in sample name "scaffolds4", in which the sequence type is reported as "Unknown, Nearest ST: 34, 196".

4. **Average Identity:** All query reads/sequences that match a housekeeping gene template sequence in the database will return the percentage identity of the alignment. This percentage identity obtained for each housekeeping gene will be averaged and this average will be displayed in this column.
5. **Average Coverage:** All query reads/sequences that match a housekeeping gene template sequence in the database will return the percentage coverage of the alignment. This percentage coverage obtained for each housekeeping gene will be averaged and this average will be displayed in this column.
6. **Notes:** It contains important information relevant to the sequence type result generated. "Matched" results will NOT any information in this column, however, "Partial" results will display all the alleles with discrepancies. This discrepancy may indicate that a novel allele was found, errors or SNPs. A detailed report containing the nucleotide(s) differences and location within the alleles can be found in the "MLST Alignment Report". "No Matched" results will indicate that no MLST loci was found in the input data, make sure that the correct MLST scheme was chosen.

When the MLST completes, it also creates a **MLST Report**. This contains the information relevant to the MLST run, including the input data, MLST configuration used, MLST results, and the parameters used for the analysis (Figure 4[178]).

[178] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=717127699



Multi-Locus Sequence Typing (MLST)

Name: MLST Results

Input Data

File Name	Sample Name	Format	Data Type
scaffolds3.fasta	scaffolds3	fasta	Assembled or Draft Genome/Contigs
scaffolds5.fasta	scaffolds5	fasta	Assembled or Draft Genome/Contigs
scaffolds6.fasta	scaffolds6	fasta	Assembled or Draft Genome/Contigs
scaffolds7.fasta	scaffolds7	fasta	Assembled or Draft Genome/Contigs
scaffolds9.fasta	scaffolds9	fasta	Assembled or Draft Genome/Contigs
scaffolds4.fasta	scaffolds4	fasta	Assembled or Draft Genome/Contigs

MLST Configuration

MLST allele sequence and profile data are obtained from PubMLST.org.

- Organism: Flavobacterium psychrophilum
- MLST Profile: fpsychrophilum
- Locus: gyrB, tuf, fumC, trpB, atpA, dnaK, murG

MLST Results

Sample	Result	Report
scaffolds3	NO MATCHED	Ht
scaffolds5	MATCHED	Ht
scaffolds6	NO MATCHED	Ht
scaffolds7	NO MATCHED	Ht
scaffolds9	PARTIAL	Ht
scaffolds4	PARTIAL	Ht

Analysis Parameters

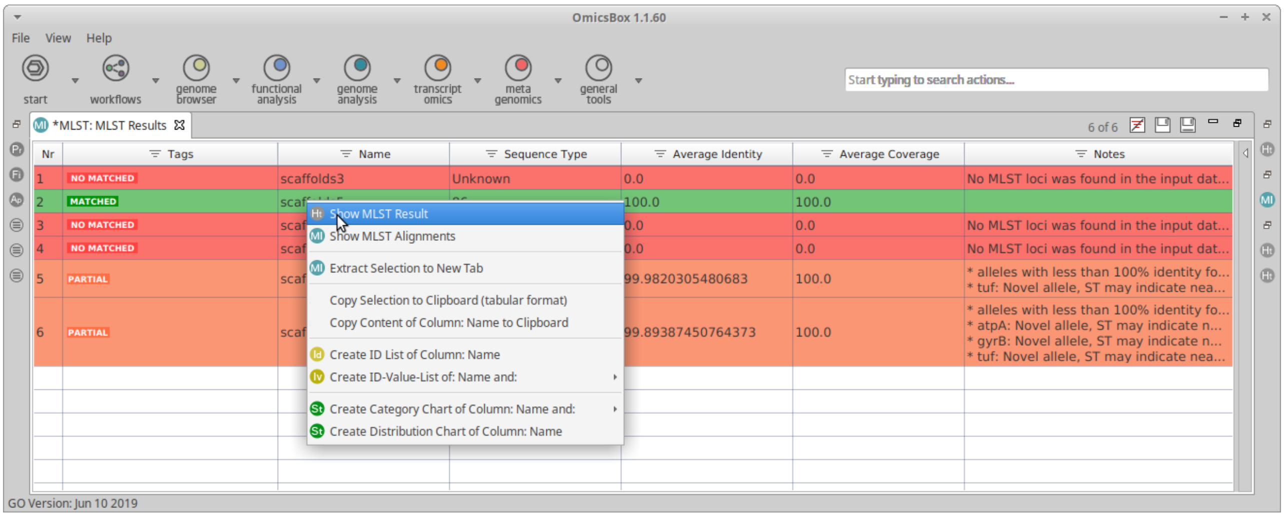
Parameter	Value
Input File Type	Assembled or Draft Genome/Contigs
MLST Configuration	Flavobacterium psychrophilum

Figure 4: MLST Report Page

support@biobam.com

sales@biobam.com

A **MLST Results** report will be generated with sample or file name and all the different housekeeping genes sequences found in the query reads/sequences (Figure 5[179]). To access this report, right-click on the row of the sample, and select "Show MLST Result" option (Figure 5[180]).


OmicsBox 1.1.60
File
View
Help
start
workflows
genome browser
functional analysis
genome analysis
transcript omics
meta genomics
general tools
Start typing to search actions...
*MLST: MLST Results
6 of 6
Show MLST Result
Show MLST Alignments
Extract Selection to New Tab
Copy Selection to Clipboard (tabular format)
Copy Content of Column: Name to Clipboard
Create ID List of Column: Name
Create ID-Value-List of: Name and:
Create Category Chart of Column: Name and:
Create Distribution Chart of Column: Name
GO Version: Jun 10 2019

Nr	Tags	Name	Sequence Type	Average Identity	Average Coverage	Notes
1	NO MATCHED	scaffolds3	Unknown	0.0	0.0	No MLST loci was found in the input dat...
2	MATCHED	scaf		100.0	100.0	
3	NO MATCHED	scaf		0.0	0.0	No MLST loci was found in the input dat...
4	NO MATCHED	scaf		0.0	0.0	No MLST loci was found in the input dat...
5	PARTIAL	scaf		99.9820305480683	100.0	* alleles with less than 100% identity fo... * tuf: Novel allele, ST may indicate nea...
6	PARTIAL	scaf		99.89387450764373	100.0	* alleles with less than 100% identity fo... * atpA: Novel allele, ST may indicate n... * gyrB: Novel allele, ST may indicate n... * tuf: Novel allele, ST may indicate nea...

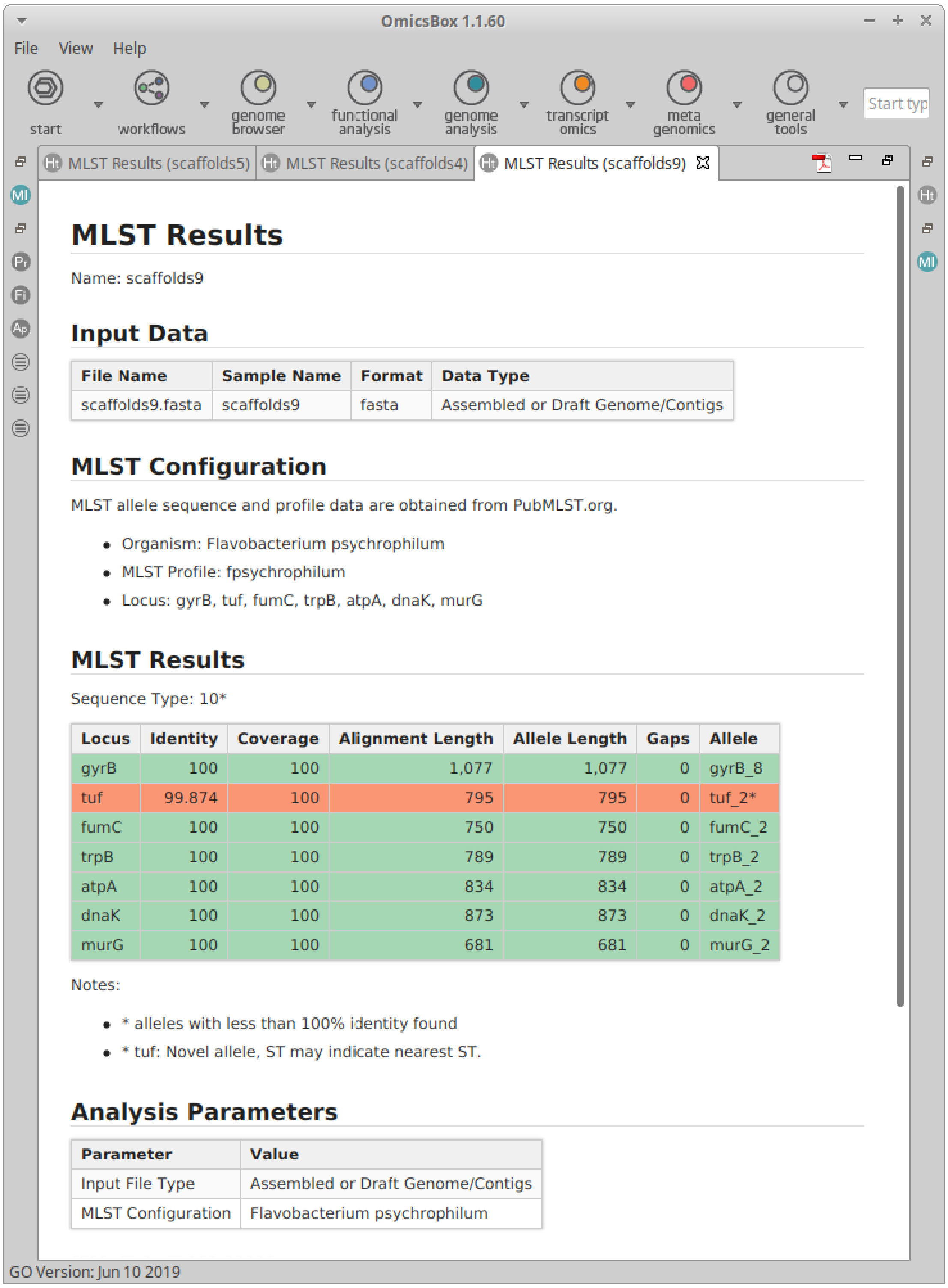
Figure 5: Generating MLST Results

The **MLST Result** report contains the information relevant to that specific sample or input file. This includes the input data, MLST configuration used, MLST results, and the parameters used for the analysis (Figure 6[181]). The MLST results portion contains a table with the **locus**, which is the name if the housekeeping gene that the query reads/sequences have been aligned to; the **identity**, which refers to the percentage of the query reads/sequences that matched a template sequence in the database; the **coverage**, which refers to how much of the template sequence in the database has been covered by the query reads/sequences; the **al ignment length**, which refers to the total number of nucleotides between the query reads/sequences that have aligned against a template sequence in the database; the **allele length**, which refers to the total number of nucleotides in the allele or template sequence in the database; **gaps**, this will indicate if any gaps or deletions have been detected; and lastly, the **allele**, which refers to the name of the housekeeping gene or sequence in the database.

[179] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=717127699#Multi-LocusSequenceTyping(MLST)-figure5

[180] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=717127699#Multi-LocusSequenceTyping(MLST)-figure5

[181] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=717127699#Multi-LocusSequenceTyping(MLST)-figure6



MLST Results

Name: scaffolds9

Input Data

File Name	Sample Name	Format	Data Type
scaffolds9.fasta	scaffolds9	fasta	Assembled or Draft Genome/Contigs

MLST Configuration

MLST allele sequence and profile data are obtained from PubMLST.org.

- Organism: Flavobacterium psychrophilum
- MLST Profile: fpsychrophilum
- Locus: gyrB, tuf, fumC, trpB, atpA, dnaK, murG

MLST Results

Sequence Type: 10*

Locus	Identity	Coverage	Alignment Length	Allele Length	Gaps	Allele
gyrB	100	100	1,077	1,077	0	gyrB_8
tuf	99.874	100	795	795	0	tuf_2*
fumC	100	100	750	750	0	fumC_2
trpB	100	100	789	789	0	trpB_2
atpA	100	100	834	834	0	atpA_2
dnaK	100	100	873	873	0	dnaK_2
murG	100	100	681	681	0	murG_2

Notes:

- * alleles with less than 100% identity found
- * tuf: Novel allele, ST may indicate nearest ST.

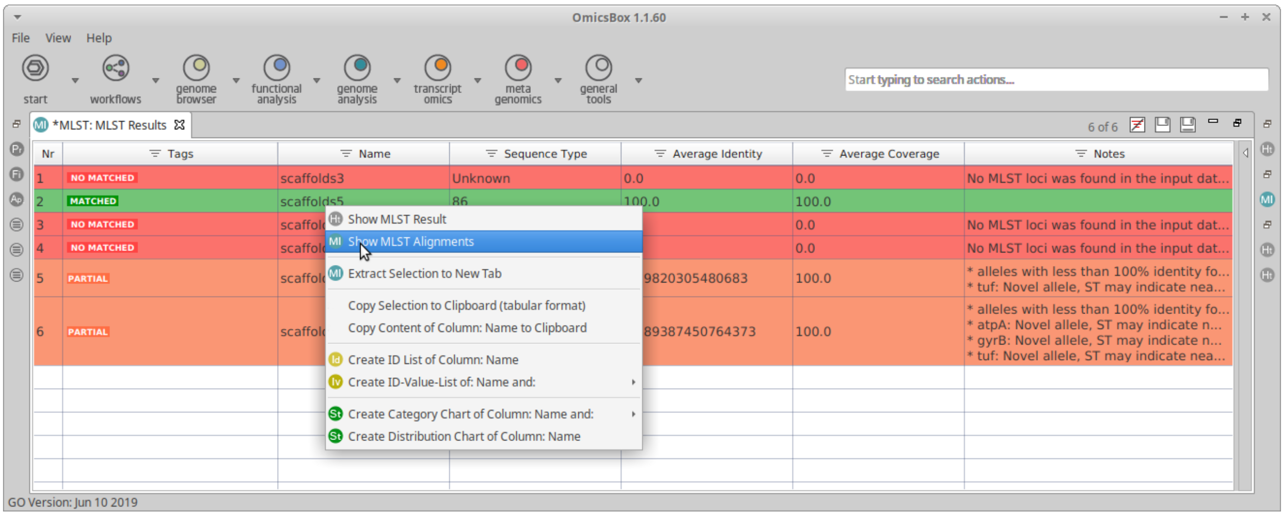
Analysis Parameters

Parameter	Value
Input File Type	Assembled or Draft Genome/Contigs
MLST Configuration	Flavobacterium psychrophilum

Figure 6: MLST Results Page for a Specific Sample or Input File

A **MLST Alignments** report will be generated with sample or file name and all the different housekeeping genes sequences found from the query reads aligned against the housekeeping genes template sequences

(Figure 4). To access this report, right-click on the row of the sample, and select "Show MLST Alignment Report" option (Figure 7[182]).



OmicsBox 1.1.60
File
View
Help
start
workflows
genome browser
functional analysis
genome analysis
transcript omics
meta genomics
general tools
Start typing to search actions...
*MLST: MLST Results
6 of 6
Show MLST Result
Show MLST Alignments
Extract Selection to New Tab
Copy Selection to Clipboard (tabular format)
Copy Content of Column: Name to Clipboard
Create ID List of Column: Name
Create ID-Value-List of: Name and:
Create Category Chart of Column: Name and:
Create Distribution Chart of Column: Name
GO Version: Jun 10 2019

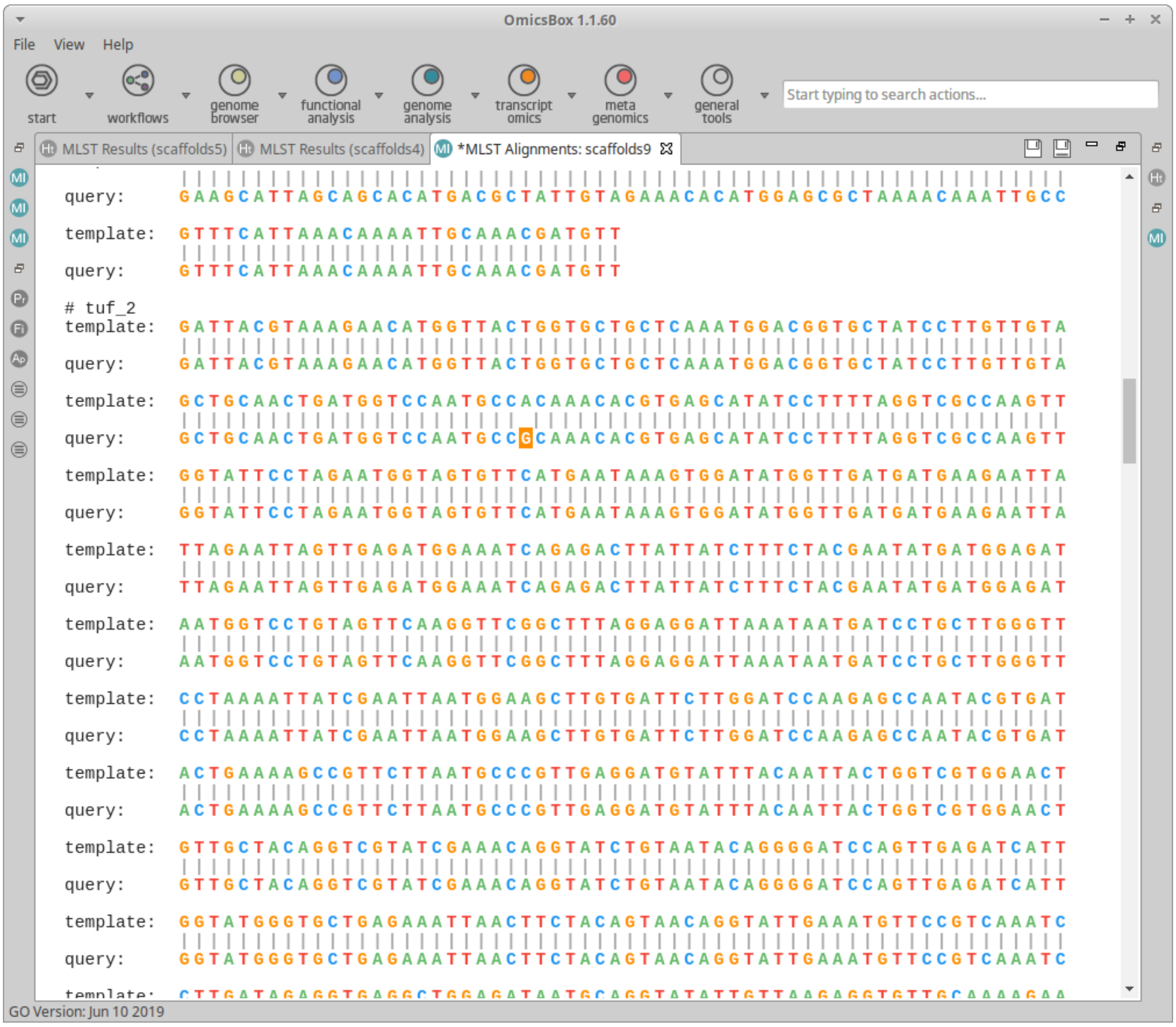
Nr	Tags	Name	Sequence Type	Average Identity	Average Coverage	Notes
1	NO MATCHED	scaffolds3	Unknown	0.0	0.0	No MLST loci was found in the input dat...
2	MATCHED	scaffolds5	86	100.0	100.0	
3	NO MATCHED	scaffol			0.0	No MLST loci was found in the input dat...
4	NO MATCHED	scaffol			0.0	No MLST loci was found in the input dat...
5	PARTIAL	scaffol		9820305480683	100.0	* alleles with less than 100% identity fo... * tuf: Novel allele, ST may indicate nea...
6	PARTIAL	scaffol		89387450764373	100.0	* alleles with less than 100% identity fo... * atpA: Novel allele, ST may indicate n... * gyrB: Novel allele, ST may indicate n... * tuf: Novel allele, ST may indicate nea...

Figure 7: Generating MLST Alignment Report Page

The **MLST Alignment** report contains a detailed and colored report of the alignments. In this report, the alignment between the query reads/sequences and the template sequence in the database is divided by each allele detected. Alleles can be identified by 'pound sign' (#) in front of the allele name. Discrepancies are highlighted (Figure 8[183]).

[182] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=717127699#Multi-LocusSequenceTyping(MLST)-figure7

[183] https://biobam.atlassian.net/wiki/pages/resumedraft.action?draftId=717127699#Multi-LocusSequenceTyping(MLST)-figure8



```
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
query:     GAAGCATTAGCAGCACATGACGCTATTGTAGAAACACATGGAGCGCTAAAACAAATTGCC

template:  GTTTCATTAAACAAAATTGCAAACGATGTT
           ||||||||||||||||||||||||||||||
query:     GTTTCATTAAACAAAATTGCAAACGATGTT

# tuf_2
template:  GATTACGTAAAGAACATGGTTACTGGTGCTGCTCAAATGGACGGTGCTATCCTTGTTGTA
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
query:     GATTACGTAAAGAACATGGTTACTGGTGCTGCTCAAATGGACGGTGCTATCCTTGTTGTA

template:  GCTGCAACTGATGGTCCAATGCCACAAACACGTGAGCATATCCTTTTAGGTCGCCAAGTT
           ||||||||||||||||||||||| ||||||||||||||||||||||||||||||||||||
query:     GCTGCAACTGATGGTCCAATGCCGCAAACACGTGAGCATATCCTTTTAGGTCGCCAAGTT

template:  GGTATTCCTAGAATGGTAGTGTTCATGAATAAAGTGGATATGGTTGATGATGAAGAATTA
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
query:     GGTATTCCTAGAATGGTAGTGTTCATGAATAAAGTGGATATGGTTGATGATGAAGAATTA

template:  TTAGAATTAGTTGAGATGGAAATCAGAGACTTATTATCTTTCTACGAATATGATGGAGAT
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
query:     TTAGAATTAGTTGAGATGGAAATCAGAGACTTATTATCTTTCTACGAATATGATGGAGAT

template:  AATGGTCCTGTAGTTCAAGGTTCGGCTTTAGGAGGATTAAATAATGATCCTGCTTGGGTT
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
query:     AATGGTCCTGTAGTTCAAGGTTCGGCTTTAGGAGGATTAAATAATGATCCTGCTTGGGTT

template:  CCTAAAATTATCGAATTAATGGAAGCTTGTGATTCTTGGATCCAAGAGCCAATACGTGAT
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
query:     CCTAAAATTATCGAATTAATGGAAGCTTGTGATTCTTGGATCCAAGAGCCAATACGTGAT

template:  ACTGAAAAGCCGTTCTTAATGCCCGTTGAGGATGTATTTACAATTACTGGTCGTGGAACT
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
query:     ACTGAAAAGCCGTTCTTAATGCCCGTTGAGGATGTATTTACAATTACTGGTCGTGGAACT

template:  GTTGCTACAGGTCGTATCGAAACAGGTATCTGTAATACAGGGGATCCAGTTGAGATCATT
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
query:     GTTGCTACAGGTCGTATCGAAACAGGTATCTGTAATACAGGGGATCCAGTTGAGATCATT

template:  GGTATGGGTGCTGAGAAATTAACTTCTACAGTAACAGGTATTGAAATGTTCCGTCAAATC
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
query:     GGTATGGGTGCTGAGAAATTAACTTCTACAGTAACAGGTATTGAAATGTTCCGTCAAATC

template:  CTTGATAGAGGTGAGGCTGGAGATAATGCAGGTATATTGTTAAGAGGTGTTGCAAAAGAA
```

Figure 8: MLST Alignment Report Page

About BioBam Bioinformatics

BioBam is a leading bioinformatics solution provider which accelerate research in disciplines such as agricultural genomics, microbiology and environmental NGS studies, amongst others.

BioBam is committed to the development of user-friendly software solutions for biological research. Our mission is to transform complex data analysis procedures into an attractive and interactive task. BioBam is devoted to close the gap between experimental work, bioinformatics analysis and applied research.

OmicsBox, BioBam’s flagship product presents a user-friendly all-in-one analysis solution for industry, academic and governmental research biologists. OmicsBox is used by top private and public research institutions worldwide. Blast2GO, its functional annotation features are well established for functional genomics studies and especially popular in non-model organism research. This is demonstrated by over 7000+ scientific research citations.

biobam
BIOINFORMATICS SOLUTIONS

support@biobam.com

sales@biobam.com

www.biobam.com