SILVAngs - rDNA-based microbial community analysis using next-generation sequencing (NGS) data - User Guide

Contact: ngs-contact@arb-silva.de

Motivation

SILVAngs is a data analysis service for ribosomal RNA gene (rDNA) amplicon reads from high-throughput sequencing (next-generation sequencing (NGS)) approaches based on an automatic software pipeline. It uses the SILVA sequence databases, taxonomies, and alignments as a reference. It facilitates the classification of rDNA reads and provides a wealth of result files (tables, graphs and sequence files) for download. The idea is to make you independent from time consuming installation and operating procedures which is typical for local solutions.

Workflow

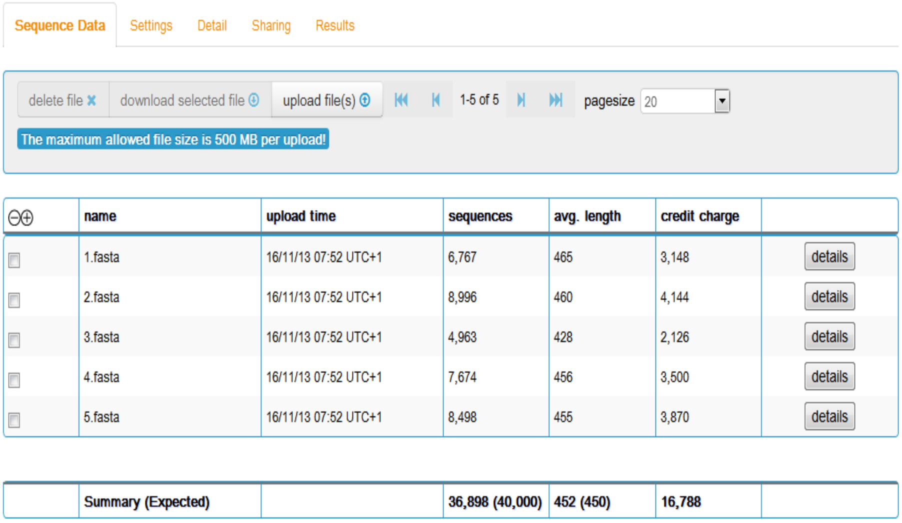
The basic workflow of the pipeline can be divided into the following steps, pretty much reflecting the common process of analysing rDNA amplicon reads from next generation sequencing approaches:

- Alignment
- Quality management
- De-replication (identification of identical sequences)
- Clustering at a user-defined threshold (OTU definition)
- Classification of the OTUs/reads

However, there are some specific features of SILVAngs compared to similar software pipelines, reflecting our philosophy on how to adequately analyse (and interpret) these kind of data in terms of quality and efficiency. They are discussed in the following description of the SILVAngs workflow.

1. Input data/Import

The pipeline accepts input data in Multi-Fasta format with each input file representing one sample. Samples that belong to one project (a transect, timeseries etc.) should be uploaded as a single SILVAngs project. The results of a project are organized according to these samples (e.g. tax breakdown and fingerprint data are listed per sample) with the name of a sample defined by the file name of the corresponding Multi-Fasta input file.



The maximum allowed file size is 500 MB per upload!

⊖⊕	name	upload time	sequences	avg. length	credit charge	
□	1.fasta	16/11/13 07:52 UTC+1	6,767	465	3,148	details
□	2.fasta	16/11/13 07:52 UTC+1	8,996	460	4,144	details
□	3.fasta	16/11/13 07:52 UTC+1	4,963	428	2,126	details
□	4.fasta	16/11/13 07:52 UTC+1	7,674	456	3,500	details
□	5.fasta	16/11/13 07:52 UTC+1	8,498	455	3,870	details
	Summary (Expected)		36,898 (40,000)	452 (450)	16,788	

Example of the content of a demo project. This project contains five individual Multi-Fasta files reflecting five different sites.

The current version of the pipeline accepts only Multi-Fasta files. Files in Standard Flowgram Format (.sff) e.g. from 454 sequencing and quality values from e.g. FASTQ files are not accepted.

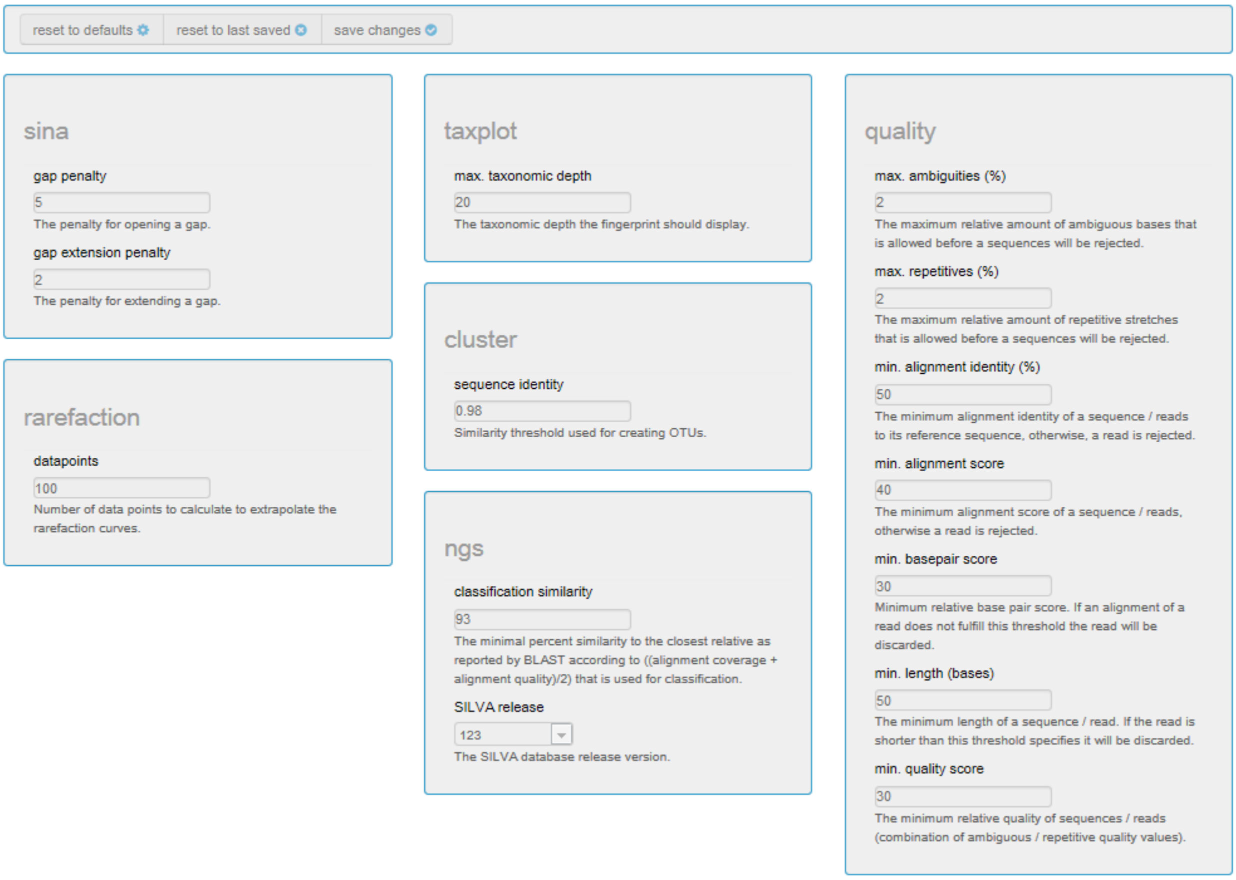
GZip support: SILVAngs supports uploading GZiped Multi-Fasta files. This massively decreases the size of your sequence uploads. At this time, we do not support the upload of (compressed) archives that include more than one file (like ZIP, TAR, and other formats). Practically this means that still you have to upload the sequences for each sample individually, but each of them can be compressed using GZip.

Windows users may compress files using the 7-zip tool (http://www.7-zip.org/). After 7-zip is installed right click on a single file, choose "7-Zip / Add to archive..." from the context menu and choose "GZip" as “Archive format". This option is only available if a single file has been selected. Mac and Linux users may use the command line tool "gzip" which is installed by default (gzip filename).

This fully automated web version of the SILVAngs pipeline does not offer demultiplexing functionalities based on barcodes and/or assembly of paired end reads. You must separate your barcode-tagged reads into dedicated Multi-Fasta files before uploading. The open source tools FASTQC http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ and BBTools https://jgi.doe.gov/data-and-tools/bbtools/ can help you in pre-processing your reads.

Please make sure that any barcodes, or adapters, attached to the reads in your input files, are completely removed before you upload your data.

After uploading your files and checking the parameters:



sina

gap penalty
5
The penalty for opening a gap.

gap extension penalty
2
The penalty for extending a gap.

rarefaction

datapoints
100
Number of data points to calculate to extrapolate the rarefaction curves.

taxplot

max. taxonomic depth
20
The taxonomic depth the fingerprint should display.

cluster

sequence identity
0.98
Similarity threshold used for creating OTUs.

ngs

classification similarity
93
The minimal percent similarity to the closest relative as reported by BLAST according to ((alignment coverage + alignment quality)/2) that is used for classification.

SILVA release
123
The SILVA database release version.

quality

max. ambiguities (%)
2
The maximum relative amount of ambiguous bases that is allowed before a sequences will be rejected.

max. repetitives (%)
2
The maximum relative amount of repetitive stretches that is allowed before a sequences will be rejected.

min. alignment identity (%)
50
The minimum alignment identity of a sequence / reads to its reference sequence, otherwise, a read is rejected.

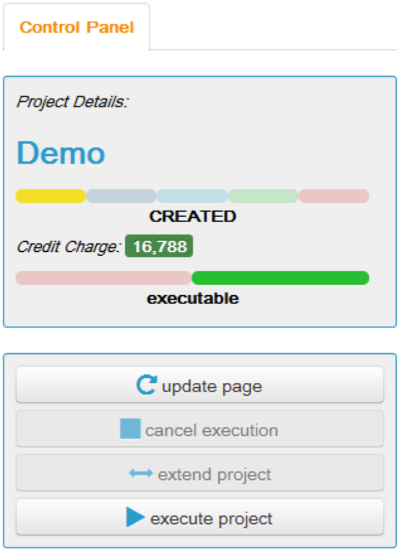
min. alignment score
40
The minimum alignment score of a sequence / reads, otherwise a read is rejected.

min. basepair score
30
Minimum relative base pair score. If an alignment of a read does not fulfill this threshold the read will be discarded.

min. length (bases)
50
The minimum length of a sequence / read. If the read is shorter than this threshold specifies it will be discarded.

min. quality score
30
The minimum relative quality of sequences / reads (combination of ambiguous / repetitive quality values).

you simply click on “execute project”:



Project Details:

Demo

CREATED

Credit Charge: 16,788

executable

You will receive an e-mail that your project has been scheduled. Further status e-mails will follow, keeping you updated on the progress of your project as it passes through the pipeline.

2. Alignment (initial quality control)

In the first step of quality control, all input reads are aligned by SINA (http://www.arb-silva.de/aligner/sina-download/ (Pruesse et al. 2012)). According to a number of parameters measured by SINA, problematic reads (such as PCR artefacts) or even contamination of the dataset with non-rDNA sequences are identified and the corresponding reads are filtered out. They are not considered for further processing – the numbers of rejected reads during initial alignment (separated in multiple classes) are given in the final statistics of the analysis.

The alignment step allows providing exports of aligned sequences for all reads which entered the classification step of the pipeline. You can use them for a detailed inspection e.g. with the ARB software package (Ludwig et al. 2004) or any other sequence editor.

3. Further quality management

All reads which have not been rejected by the previous alignment step undergo further quality filtering including length, ambiguity and homopolymer checks. These are the common parameters which allow quality control on the level of the primary sequence information. The **length cut-off can be defined by the user**, whereas for ambiguities and homopolymers the same thresholds as for the SILVA databases are used (max. 2%). Notably, only the aligned region of each read is considered (and further processed!). This means if you have non-rDNA overhangs included, they do not bias your results.

Reads with insufficient quality values are not further considered. The number of reads rejected during quality management (separated by multiple classes) is shown in the final statistics of the analysis.

4. De-replication (identification of identical sequences)

All remaining reads enter the de-replication stage of the pipeline. 100% identical reads, ignoring overhangs, are identified and only the longest read is retained for further processing.

This is a common approach to reduce calculation time, since processing redundant reads is a waste of computing power. Furthermore, the statistics of this step provides useful information on how “clonal” the input dataset or selected taxonomic units/groups are.

5. Clustering/OTU definition

Technically, this is just another de-replication step to further reduce the number of reads that needs to be classified. Compared to previous de-replication, clustering is done on a **97-99% identity level which can be adjusted by you**. This is motivated by the fact that sequencing errors and operon heterogeneities can easily introduce 3% artificial divergence in the data.

As indicated before, step 4 and 5 both are just representing two subsequent de-replication=clustering steps, so why then not combining them?

Simply, because in case of defining OTUs based on a clustering with a threshold below 100% identity, no information could be obtained on sequence clonality (identical sequences) within a dataset. This adds an additional layer of information about your dataset for you.

Finally, users are normally highly interested in also achieving OTU numbers and the comparison of OTUs across different samples - they also want to look beyond the level of read distribution according to the SILVA taxonomy. However, please note the discussion of corresponding limitations below.

6. Classification of OTUs/reads

In the classification step the representative reads (the longest read of each OTU) are compared to the SILVA reference datasets of the small- (16S/18S) and large (23S/28S) subunit rDNA with its corresponding SILVA taxonomy (Quast et al. 2013). Currently, the SILVAngs pipeline uses a BLAST-based approach for classifying the reads according to the SILVA taxonomy which differs a bit from other BLAST-based approaches in rDNA sequence analysis.

Since the SILVA reference datasets are comprehensive, quality-controlled, and of high integrity, and the corresponding taxonomy is phylogeny-based and manually curated, we fully rely on the best BLAST hit and avoid more complex procedures such as least common ancestor (LCA) approaches. But of course only significant hits (see below) are considered; everything else is assigned to a class called ‘No Relative’ and is offered to you for further inspection. However, usually just a few percent (<10%) of the reads within a complete dataset go into this class, since the majority of reads is normally classified. Please contact us in case a high number of reads is reported as ‘No Relative’.

The classification result of each representative read (including the additional class ‘No Relative’) is finally mapped back to all other members (reads) of the OTU cluster and of course also to the corresponding identical reads from de-replication step.

We assume that SILVA’s phylogeny-based taxonomy has a reliable resolution down to the genus level if accurate data are submitted. We consider the assignment of species dangerous and needs to be supported by a careful inspection of the corresponding reads. In general we recommend avoiding over-interpretation of the results especially for reads below 1200 bases.

Important - reliability of classification is assured by two aspects:

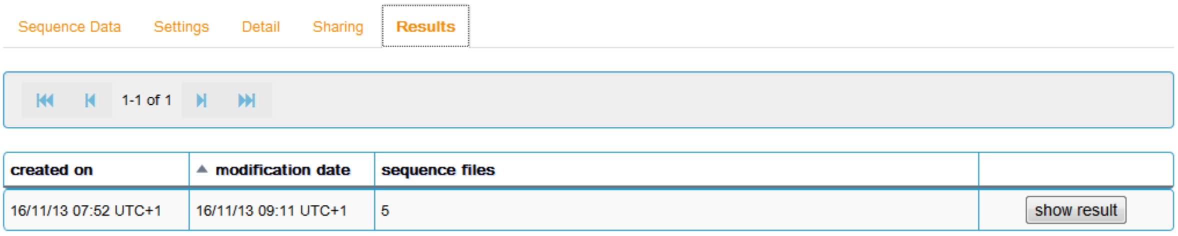
Just taking the best BLAST hit for classification is of course dangerous for two reasons:

1) The hit can be far away from the query sequence (e.g. >5% divergence) or the reported identity is artificial because of just partial BLAST alignments. To circumvent these shortcomings, best hits are tested and only accepted if (sequence identity + alignment coverage) / 2 >= 93. This threshold has been determined empirically and turned out to be an acceptable compromise between sensitivity and accuracy.

2) From the best BLAST hit we, of course, do not refer to any non-standardized user based annotation (from original submission of the reference sequence), but only the SILVA taxonomic information is taken to identify the unknown OTU/read. **In other words, all accepted reads of the project are finally mapped to the standardized SILVA taxonomy. This is the actual result of the pipeline visualized by various views and supported by additional statistics and exports.**

7. Short description of SILVAngs output files

After your project is finished you can navigate to the analysis results by clicking on the **Results** tab:



created on	▲ modification date	sequence files	
16/11/13 07:52 UTC+1	16/11/13 09:11 UTC+1	5	show result

Show results displays the overview (Summary) of your analysis with some charts.

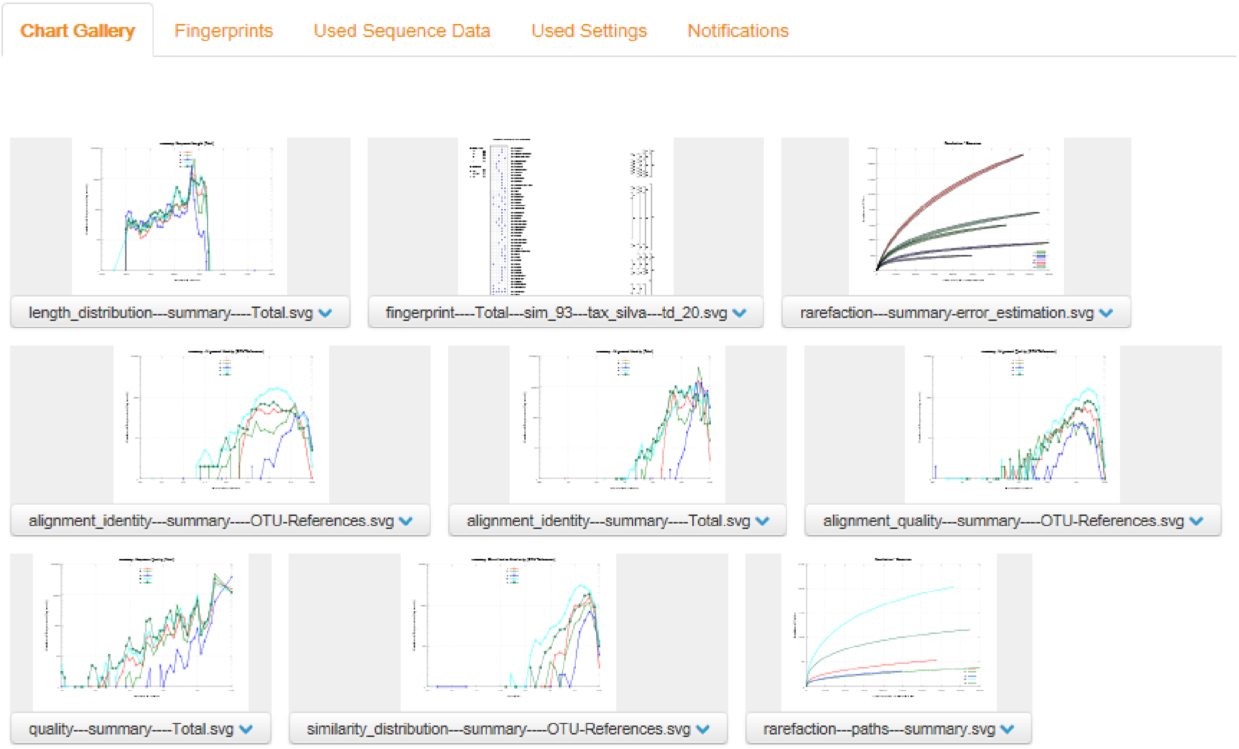
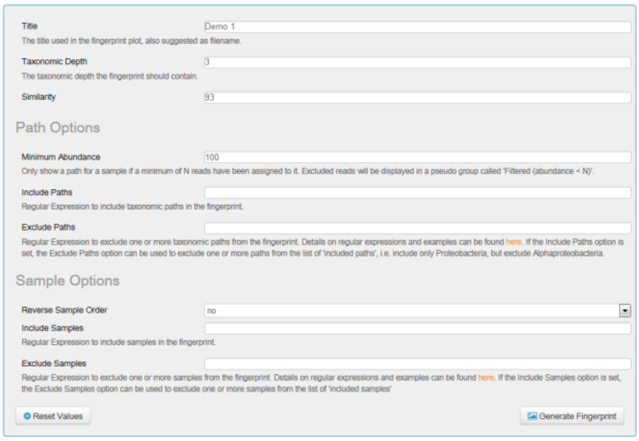


Chart Gallery
Fingerprints
Used Sequence Data
Used Settings
Notifications
length_distribution---summary----Total.svg
fingerprint----Total---sim_93---tax_silva---td_20.svg
rarefaction---summary-error_estimation.svg
alignment_identity---summary----OTU-References.svg
alignment_identity---summary----Total.svg
alignment_quality---summary----OTU-References.svg
quality---summary----Total.svg
similarity_distribution---summary----OTU-References.svg
rarefaction---paths---summary.svg

Fingerprints provides the possibility to dynamically create and store fingerprints in the web frontend. It offers the option to specify the taxonomic depth, filter the fingerprint by taxonomic paths and include/remove samples. All dynamically created fingerprints include the same data files as the fingerprints provided as part of the default results (abundance table, group id to group name mapping, and used parameters) for download.

Examples:

Taxonomic depth 3 and exclude all reads <100:



Title: Demo 1
The title used in the fingerprint plot, also suggested as filename.

Taxonomic Depth: 3
The taxonomic depth the fingerprint should contain.

Similarity: 93

Path Options

Minimum Abundance: 100
Only show a path for a sample if a minimum of N reads have been assigned to it. Excluded reads will be displayed in a pseudo group called 'Filtered (abundance < N)'.

Include Paths:
Regular Expression to include taxonomic paths in the fingerprint.

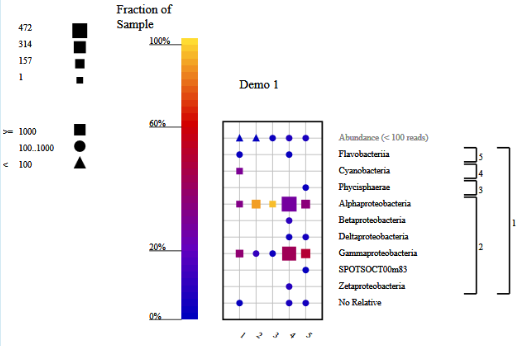
Exclude Paths:
Regular Expression to exclude one or more taxonomic paths from the fingerprint. Details on regular expressions and examples can be found here. If the Include Paths option is set, the Exclude Paths option can be used to exclude one or more paths from the list of 'included paths', i.e. include only Proteobacteria, but exclude Alphaproteobacteria.

Sample Options

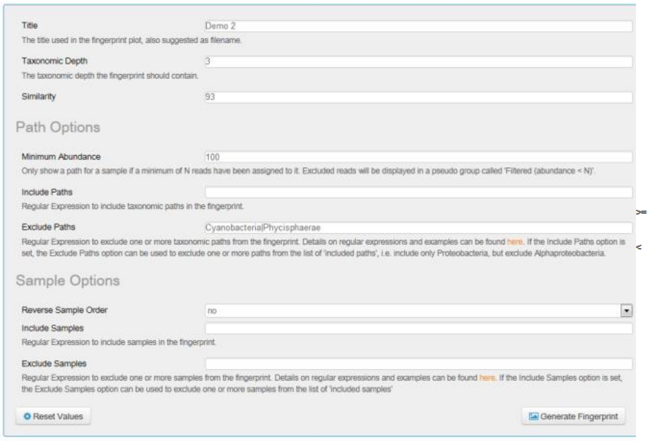
Reverse Sample Order: no

Include Samples:
Regular Expression to include samples in the fingerprint.

Exclude Samples:
Regular Expression to exclude one or more samples from the fingerprint. Details on regular expressions and examples can be found here. If the Include Samples option is set, the Exclude Samples option can be used to exclude one or more samples from the list of 'included samples'



Taxonomic depth 3, exclude all reads <100 and exclude paths Cyanobacteria and Phycisphaerae, regular expression: Cyanobacteria|Phycisphaerae



Title: Demo 2
The title used in the fingerprint plot, also suggested as filename.

Taxonomic Depth: 3
The taxonomic depth the fingerprint should contain.

Similarity: 93

Path Options

Minimum Abundance: 100
Only show a path for a sample if a minimum of N reads have been assigned to it. Excluded reads will be displayed in a pseudo group called 'Filtered (abundance < N)'.

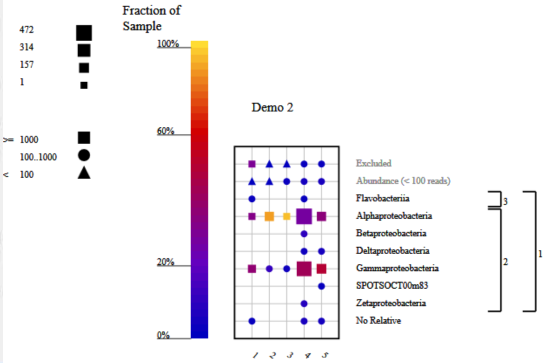
Include Paths:
Regular Expression to include taxonomic paths in the fingerprint.

Exclude Paths: Cyanobacteria|Phycisphaerae
Regular Expression to exclude one or more taxonomic paths from the fingerprint. Details on regular expressions and examples can be found here. If the Include Paths option is set, the Exclude Paths option can be used to exclude one or more paths from the list of 'included paths', i.e. include only Proteobacteria, but exclude Alphaproteobacteria.

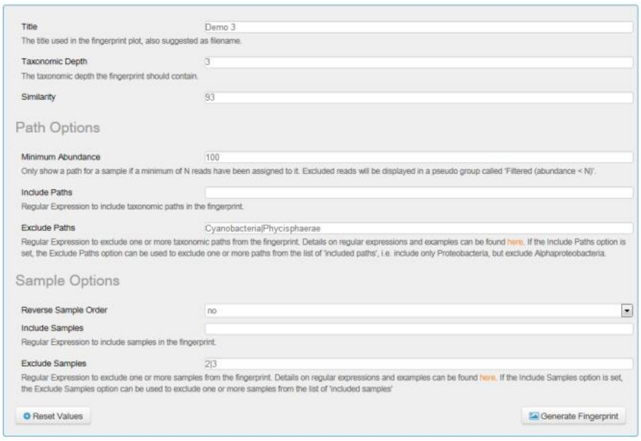
Sample Options

Reverse Sample Order: no
Include Samples:
Regular Expression to include samples in the fingerprint.

Exclude Samples:
Regular Expression to exclude one or more samples from the fingerprint. Details on regular expressions and examples can be found here. If the Include Samples option is set, the Exclude Samples option can be used to exclude one or more samples from the list of 'included samples'



Taxonomic depth 3, exclude all reads <100, exclude paths Cyanobacteria and Phycisphaerae and samples 2 and 3, regular expression: 2|3



Title: Demo 3
The title used in the fingerprint plot, also suggested as filename.

Taxonomic Depth: 3
The taxonomic depth the fingerprint should contain.

Similarity: 93

Path Options

Minimum Abundance: 100
Only show a path for a sample if a minimum of N reads have been assigned to it. Excluded reads will be displayed in a pseudo group called 'Filtered (abundance < N)'.

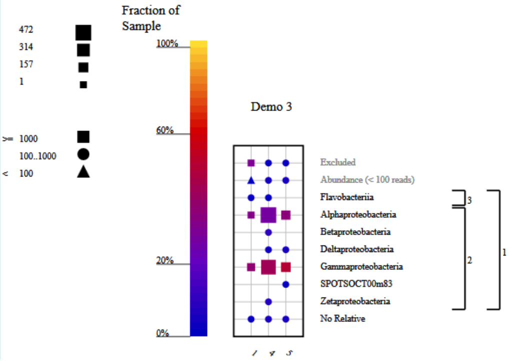
Include Paths:
Regular Expression to include taxonomic paths in the fingerprint.

Exclude Paths: Cyanobacteria|Phycisphaerae
Regular Expression to exclude one or more taxonomic paths from the fingerprint. Details on regular expressions and examples can be found here. If the Include Paths option is set, the Exclude Paths option can be used to exclude one or more paths from the list of 'included paths', i.e. include only Proteobacteria, but exclude Alphaproteobacteria.

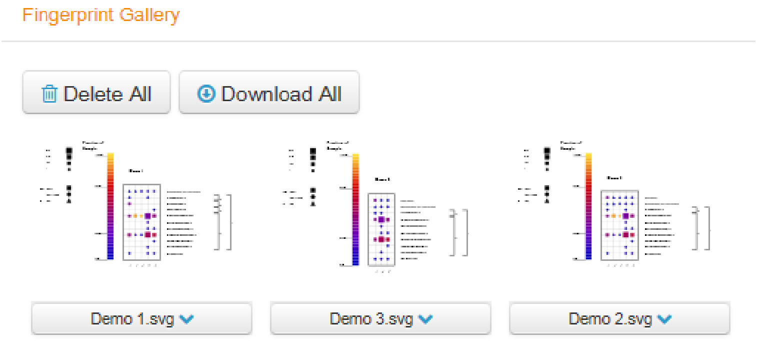
Sample Options

Reverse Sample Order: no
Include Samples:
Regular Expression to include samples in the fingerprint.

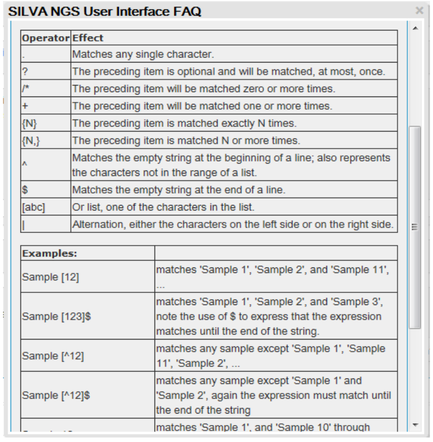
Exclude Samples: 2|3
Regular Expression to exclude one or more samples from the fingerprint. Details on regular expressions and examples can be found here. If the Include Samples option is set, the Exclude Samples option can be used to exclude one or more samples from the list of 'included samples'



All fingerprints can be stored and downloaded together with all underlying data files:



A detailed description of the syntax for regular expressions including more examples is also available:

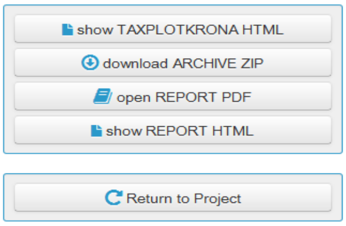


SILVA NGS User Interface FAQ

Operator	Effect
.	Matches any single character.
?	The preceding item is optional and will be matched, at most, once.
/*	The preceding item will be matched zero or more times.
+	The preceding item will be matched one or more times.
{N}	The preceding item is matched exactly N times.
{N,}	The preceding item is matched N or more times.
^	Matches the empty string at the beginning of a line; also represents the characters not in the range of a list.
$	Matches the empty string at the end of a line.
[abc]	Or list, one of the characters in the list.
|	Alternation, either the characters on the left side or on the right side.

Examples:	
Sample [12]	matches 'Sample 1', 'Sample 2', and 'Sample 11', ...
Sample [123]$	matches 'Sample 1', 'Sample 2', and 'Sample 3', note the use of $ to express that the expression matches until the end of the string.
Sample [^12]	matches any sample except 'Sample 1', 'Sample 11', 'Sample 2', ...
Sample [^12]$	matches any sample except 'Sample 1' and 'Sample 2', again the expression must match until the end of the string
[illegible]	matches 'Sample 1', and 'Sample 10' through

The Control Panel on the right provides access to all data.



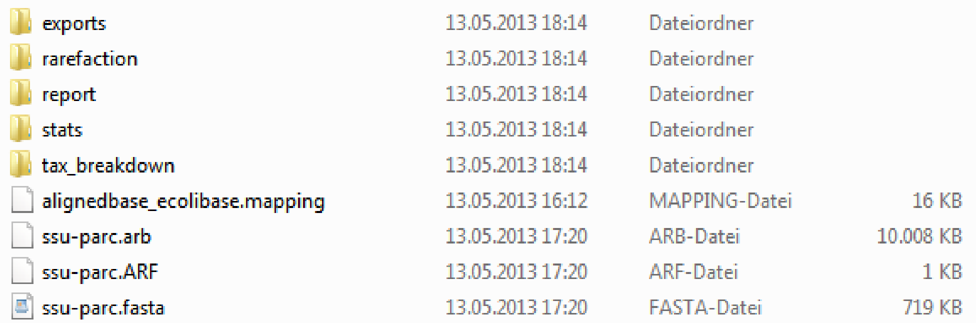
Report HTML provides you with the most important information about your analysis, look at this first.

Report PDF provides you with an overview about your project including some background on the pipeline, a material and methods part with relevant citations, as well as the most important summary graphs. It is recommended to read this carefully.

Taxplotkrona HTML provides an interactive visualisation of the taxonomic breakdown per sample (example below).

Archive ZIP provides all files (tables, graphs, Fasta and ARB files) the pipeline has produced in the analysis process. Download this file and archive it! Projects will be automatically deleted after 180 days!

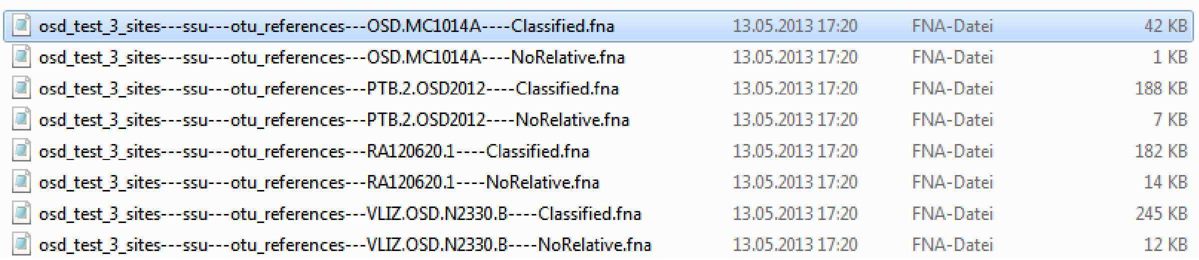
After unpacking the .zip file you will find a directory called results/ssu or lsu



exports	13.05.2013 18:14	Dateiordner	
rarefaction	13.05.2013 18:14	Dateiordner	
report	13.05.2013 18:14	Dateiordner	
stats	13.05.2013 18:14	Dateiordner	
tax_breakdown	13.05.2013 18:14	Dateiordner	
alignedbase_ecolibase.mapping	13.05.2013 16:12	MAPPING-Datei	16 KB
ssu-parc.arb	13.05.2013 17:20	ARB-Datei	10.008 KB
ssu-parc.ARF	13.05.2013 17:20	ARF-Datei	1 KB
ssu-parc.fasta	13.05.2013 17:20	FASTA-Datei	719 KB

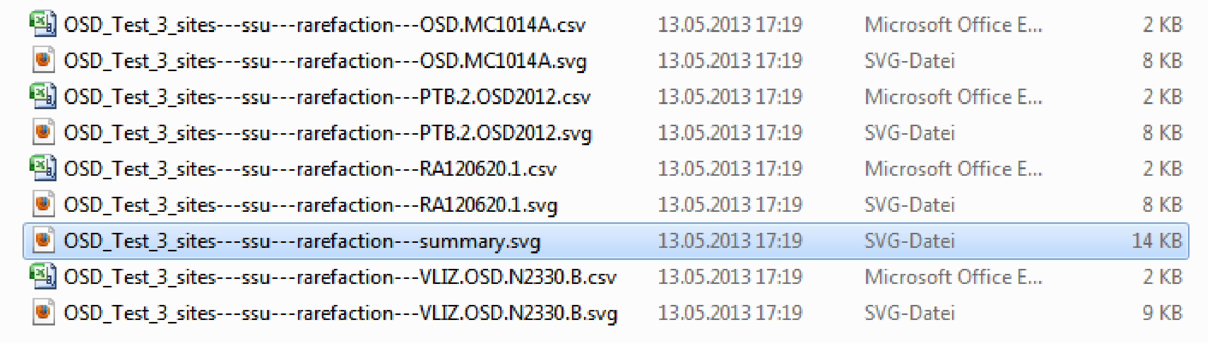
In the main directory you can find the OTU reference sequences in ARB and FASTA format generated by the pipeline (xx-parc.*). The ARB file also contains aligned sequences.

exports contains an overview file and in the folder otu_references the classified and unclassified (no relatives) reference sequences as multi-Fasta files (.fna) of each of the sample submitted.



osd_test_3_sites---ssu---otu_references---OSD.MC1014A----Classified.fna	13.05.2013 17:20	FNA-Datei	42 KB
osd_test_3_sites---ssu---otu_references---OSD.MC1014A----NoRelative.fna	13.05.2013 17:20	FNA-Datei	1 KB
osd_test_3_sites---ssu---otu_references---PTB.2.OSD2012----Classified.fna	13.05.2013 17:20	FNA-Datei	188 KB
osd_test_3_sites---ssu---otu_references---PTB.2.OSD2012----NoRelative.fna	13.05.2013 17:20	FNA-Datei	7 KB
osd_test_3_sites---ssu---otu_references---RA120620.1----Classified.fna	13.05.2013 17:20	FNA-Datei	182 KB
osd_test_3_sites---ssu---otu_references---RA120620.1----NoRelative.fna	13.05.2013 17:20	FNA-Datei	14 KB
osd_test_3_sites---ssu---otu_references---VLIZ.OSD.N2330.B----Classified.fna	13.05.2013 17:20	FNA-Datei	245 KB
osd_test_3_sites---ssu---otu_references---VLIZ.OSD.N2330.B----NoRelative.fna	13.05.2013 17:20	FNA-Datei	12 KB

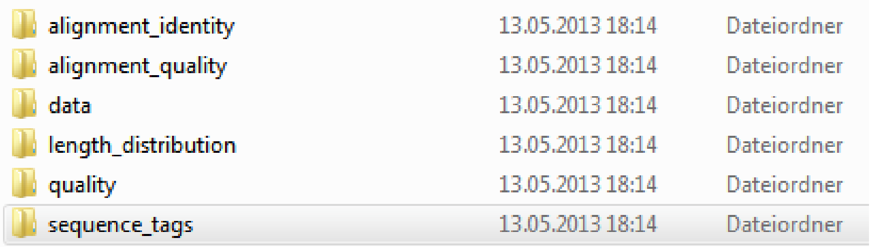
rarefaction contains all data and graphs of the rarefaction curves. The SVG files can be viewed e.g. with a web browser or edited with vector graphics software (e.g. Adobe Illustrator or Inkscape). The file ending with "summary.svg" shows all rarefaction curves in one graph.



OSD_Test_3_sites---ssu---rarefaction---OSD.MC1014A.csv	13.05.2013 17:19	Microsoft Office E...	2 KB
OSD_Test_3_sites---ssu---rarefaction---OSD.MC1014A.svg	13.05.2013 17:19	SVG-Datei	8 KB
OSD_Test_3_sites---ssu---rarefaction---PTB.2.OSD2012.csv	13.05.2013 17:19	Microsoft Office E...	2 KB
OSD_Test_3_sites---ssu---rarefaction---PTB.2.OSD2012.svg	13.05.2013 17:19	SVG-Datei	8 KB
OSD_Test_3_sites---ssu---rarefaction---RA120620.1.csv	13.05.2013 17:19	Microsoft Office E...	2 KB
OSD_Test_3_sites---ssu---rarefaction---RA120620.1.svg	13.05.2013 17:19	SVG-Datei	8 KB
OSD_Test_3_sites---ssu---rarefaction---summary.svg	13.05.2013 17:19	SVG-Datei	14 KB
OSD_Test_3_sites---ssu---rarefaction---VLIZ.OSD.N2330.B.csv	13.05.2013 17:19	Microsoft Office E...	2 KB
OSD_Test_3_sites---ssu---rarefaction---VLIZ.OSD.N2330.B.svg	13.05.2013 17:19	SVG-Datei	9 KB

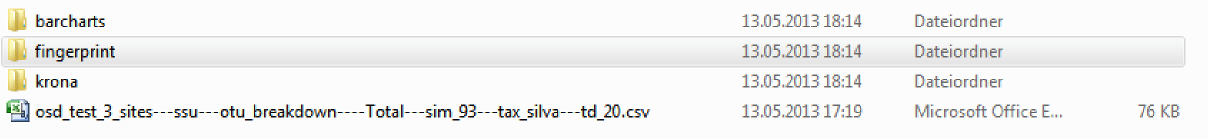
report provides you with an overview about your project including some background on the pipeline, a material and methods part with relevant citations, as well as the most important summary graphs. It is recommended to look at this PDF file first (this is the same file as shown on the SILVAngs webpage).

stats provides detailed "statistics" about the sequences and processing logs of your sequences. The summary files are most informative.



alignment_identity	13.05.2013 18:14	Dateiordner
alignment_quality	13.05.2013 18:14	Dateiordner
data	13.05.2013 18:14	Dateiordner
length_distribution	13.05.2013 18:14	Dateiordner
quality	13.05.2013 18:14	Dateiordner
sequence_tags	13.05.2013 18:14	Dateiordner

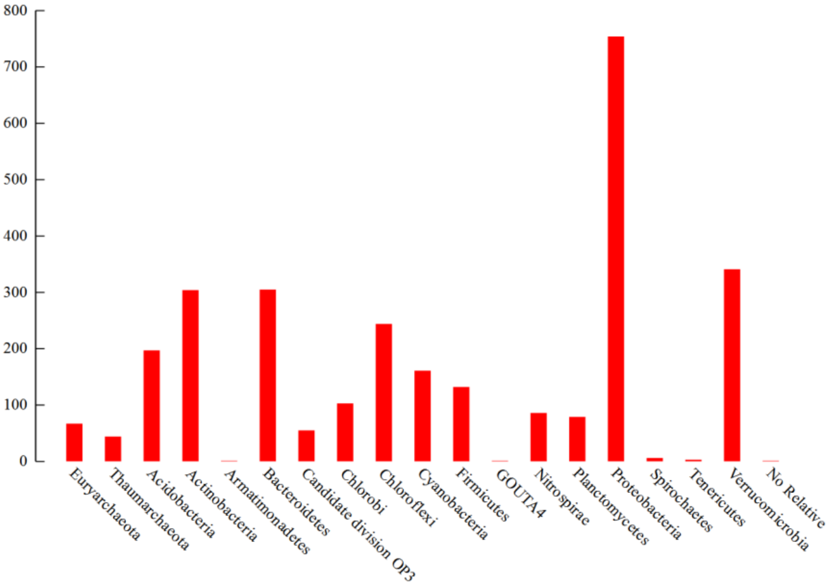
tax_breakdown contains detailed information about the taxonomic classification of your sequences.



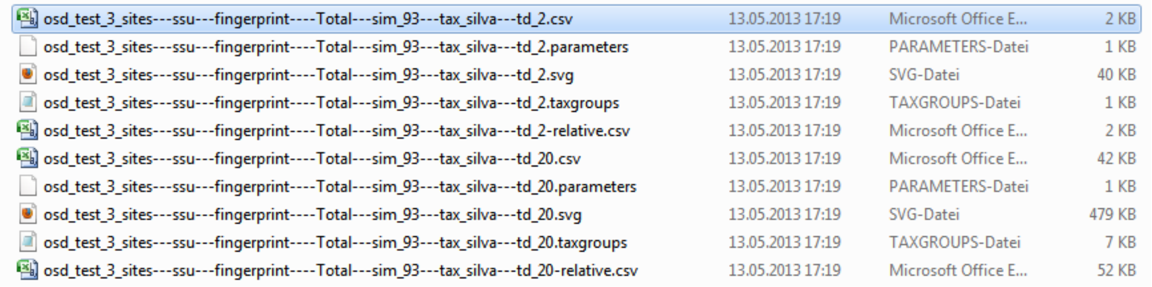
barcharts	13.05.2013 18:14	Dateiordner	
fingerprint	13.05.2013 18:14	Dateiordner	
krona	13.05.2013 18:14	Dateiordner	
osd_test_3_sites---ssu---otu_breakdown----Total---sim_93---tax_silva---td_20.csv	13.05.2013 17:19	Microsoft Office E...	76 KB

otu_breakdown.csv shows the taxonomic breakdown sorted according to sampling sites.

barcharts provides an overview of the read abundances on phylum level per sample.

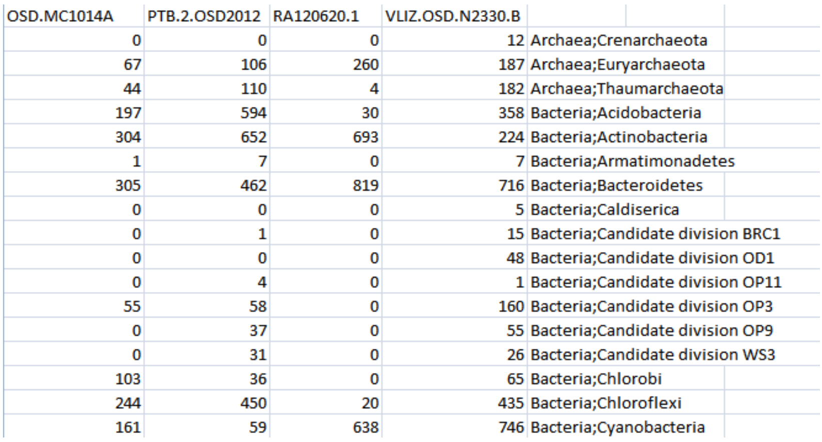


fingerprint provides detailed comparative information about the classification of your sequences. The number 2 (phylum level) and 20 (max depth) in the filename indicates the maximal depth of the taxonomic path.



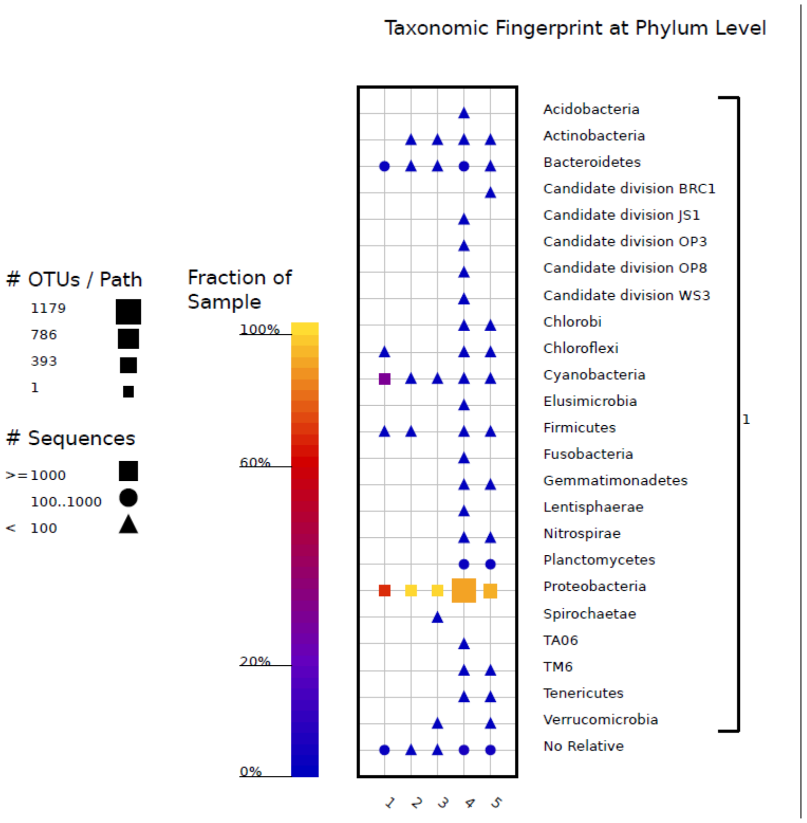
osd_test_3_sites---ssu---fingerprint----Total---sim_93---tax_silva---td_2.csv	13.05.2013 17:19	Microsoft Office E...	2 KB
osd_test_3_sites---ssu---fingerprint----Total---sim_93---tax_silva---td_2.parameters	13.05.2013 17:19	PARAMETERS-Datei	1 KB
osd_test_3_sites---ssu---fingerprint----Total---sim_93---tax_silva---td_2.svg	13.05.2013 17:19	SVG-Datei	40 KB
osd_test_3_sites---ssu---fingerprint----Total---sim_93---tax_silva---td_2.taxgroups	13.05.2013 17:19	TAXGROUPS-Datei	1 KB
osd_test_3_sites---ssu---fingerprint----Total---sim_93---tax_silva---td_2-relative.csv	13.05.2013 17:19	Microsoft Office E...	2 KB
osd_test_3_sites---ssu---fingerprint----Total---sim_93---tax_silva---td_20.csv	13.05.2013 17:19	Microsoft Office E...	42 KB
osd_test_3_sites---ssu---fingerprint----Total---sim_93---tax_silva---td_20.parameters	13.05.2013 17:19	PARAMETERS-Datei	1 KB
osd_test_3_sites---ssu---fingerprint----Total---sim_93---tax_silva---td_20.svg	13.05.2013 17:19	SVG-Datei	479 KB
osd_test_3_sites---ssu---fingerprint----Total---sim_93---tax_silva---td_20.taxgroups	13.05.2013 17:19	TAXGROUPS-Datei	7 KB
osd_test_3_sites---ssu---fingerprint----Total---sim_93---tax_silva---td_20-relative.csv	13.05.2013 17:19	Microsoft Office E...	52 KB

Tables (.csv files)

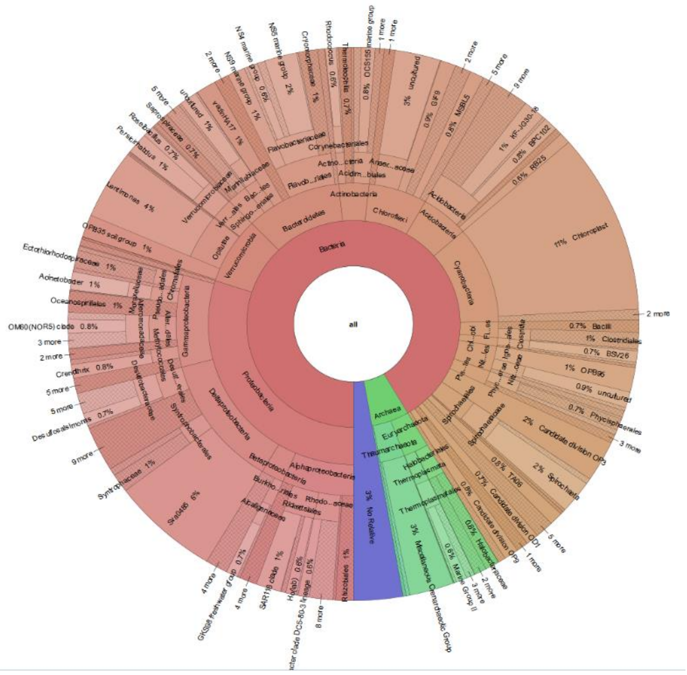


OSD.MC1014A	PTB.2.OSD2012	RA120620.1	VLIZ.OSD.N2330.B	
0	0	0	12	Archaea;Crenarchaeota
67	106	260	187	Archaea;Euryarchaeota
44	110	4	182	Archaea;Thaumarchaeota
197	594	30	358	Bacteria;Acidobacteria
304	652	693	224	Bacteria;Actinobacteria
1	7	0	7	Bacteria;Armatimonadetes
305	462	819	716	Bacteria;Bacteroidetes
0	0	0	5	Bacteria;Caldiserica
0	1	0	15	Bacteria;Candidate division BRC1
0	0	0	48	Bacteria;Candidate division OD1
0	4	0	1	Bacteria;Candidate division OP11
55	58	0	160	Bacteria;Candidate division OP3
0	37	0	55	Bacteria;Candidate division OP9
0	31	0	26	Bacteria;Candidate division WS3
103	36	0	65	Bacteria;Chlorobi
244	450	20	435	Bacteria;Chloroflexi
161	59	638	746	Bacteria;Cyanobacteria

The .svg files provide a graphical comparison of the OTU and read abundances per sample.



krona provides an interactive visualisation of the taxonomic breakdown per sample (.svg).



8. Extending a Project

Every project can be extended. Extending a project is possible after the project has been finished. It allows you to add new samples to the selected project. Moreover, you can change the project parameters and rerun the project on the same samples with new parameters. The generated results will not replace former results, since the new results will be added to your project.

9. Project Sharing

Every project can be shared by the creator (owner) with other registered users. To do so, enter the email address of the person you wish to share the project with in the “project permission view”. Please make sure to enter the email address correctly! For security reasons, the system will give no indication as to whether the email address matches a registered user.

All projects that are shared with you are listed on the "Shared Project" page.

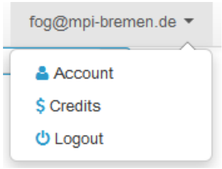
The permission “**readonly”** makes other users able to see your sequence data, parameters and results. Users with “**readonly”** rights can also download the sequence data of your project.

The permission “**write”** allows other users to upload additional data, change project parameters and finally execute projects. The credit charge will be subtracted from the account requesting project execution by using their credits (see credits system).

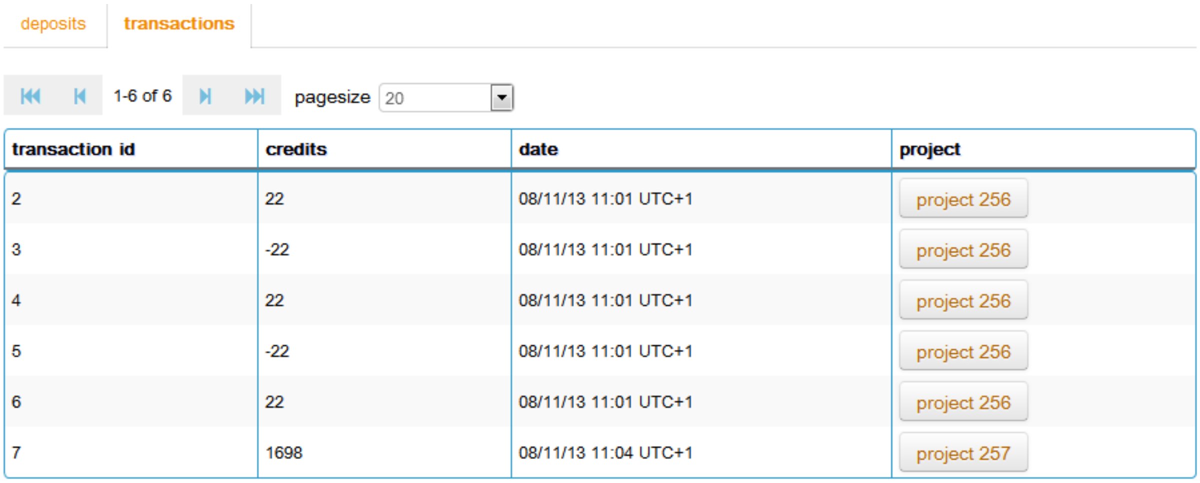
10. Credit System

Executing a project costs credits. Generally, credits are automatically distributed over time on your account as **credit deposits**. The global credit distribution depends on the compute power available to the SILVA project and the number of users using the SILVAngs service.

The project costs for execution (credit charge) depends only on the length and quantity of the sequences. The number of files (samples) does not matter. The credit charge of a sequence file is calculated by the total number of bases divided by 1000. Execution of a project creates a **credit transaction** as receipt. If you cancel a project the credits are transferred back to your account. Your current credits and an overview of all transactions are shown in “your account”.



For big projects with high credits costs, the project needs to be approved for execution. The approval is done by the administrators of SILVAngs. You have to request the execution of your project via the **request execution** button, no separate e-mail requesting the execution is needed in this case. After your project has been approved for execution, it will be processed as usual.

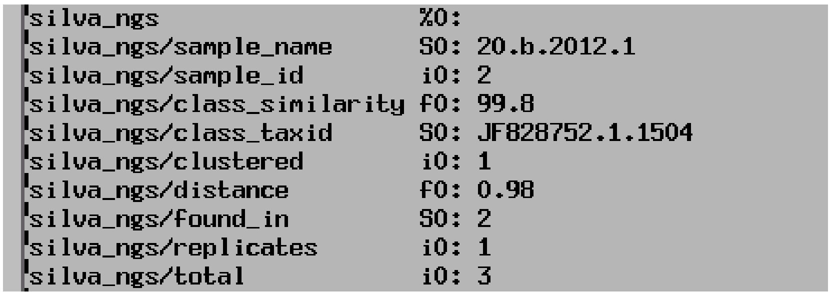


transaction id	credits	date	project
2	22	08/11/13 11:01 UTC+1	project 256
3	-22	08/11/13 11:01 UTC+1	project 256
4	22	08/11/13 11:01 UTC+1	project 256
5	-22	08/11/13 11:01 UTC+1	project 256
6	22	08/11/13 11:01 UTC+1	project 256
7	1698	08/11/13 11:04 UTC+1	project 257

10. Inspecting the OTU reference sequences in ARB

As described in chapter 7, the pipeline is generating an ARB file (xx-parc.arb) containing all OTU reference sequences of all samples. This file can be directly opened with the software package ARB (www.arb-home.de) for further inspection of the sequences and alignments.

Several NGS specific fields are exported ("silva_ngs/...") to provide additional information about the reference sequence under consideration. The information can be inspected in the "species information" window of ARB.



```
silva_ngs                  %0:
silva_ngs/sample_name      S0: 20.b.2012.1
silva_ngs/sample_id        i0: 2
silva_ngs/class_similarity f0: 99.8
silva_ngs/class_taxid      S0: JF828752.1.1504
silva_ngs/clustered        i0: 1
silva_ngs/distance         f0: 0.98
silva_ngs/found_in         S0: 2
silva_ngs/replicates       i0: 1
silva_ngs/total            i0: 3
```

Description of fields:

sample name: the name of the sample as chosen by the user,

sample id: the SILVA-ngs internal numeric ID used to identify the sample,

class similarity: the similarity to the closest relative as reported by BLAST
((alignmentcoverage + alignmentidentity)/2),

class taxid: the INSDC accession number of the closest relative with start and stop positions,

clustered: the number of sequences that have a sequence identity of at least <identity> to this sequence,

identity (wrongly named as distance): the identity threshold used for clustering 0..1 (1 -> 100%),

found in: a list of (SILVA) sample IDs in which this OTU has been found (created by otumap)

replicates: the number of sequences identical to this sequence or to any of the sequences that are clustered with this sequence,

total: the total number of sequences represented by this sequence. This is the number of clustered sequences plus the number of replicates plus one (this sequence).

11. General considerations on NGS rDNA data interpretation

Since all NGS approaches just generate a single read for a particular amplicon and assembly with high positional redundancy (like for genome sequencing) is not possible for rDNA-based microbial community analysis, there is a given likelihood for sequencing errors in each read analyzed. This likelihood is under discussion, with some papers propagating massive artificial micro-diversity for e.g. 454 datasets. However, nobody can give a definite answer on the real extent of errors, especially because these strongly depend on the particular lab and equipment as well as experimental conditions. As a matter of fact, you always have a mixture of sequencing errors and natural operon heterogeneities in your data (artificial micro-diversity) and these errors always lead to an overestimation of OTUs!

Secondly, all clustering approaches are fuzzy due to conceptual shortcomings. There is not a single unambiguous way of calculating OTUs and heuristics are used to balance accuracy with calculation time.

Thirdly, depending on your sequencing method and chemistry you usually get rather short reads with NGS approaches and you have to be aware that in terms of resolution this is a severe limitation.

Summary

Our philosophy in SILVAngs is to map NGS reads to the SILVA taxonomy with a resolution down to the genus level and to compare different samples based on the standardized SILVA taxonomy. From our perspective, NGS in microbial ecology is a high-throughput screening tool to get information on the overall diversity and dynamics of your samples. To identify hotspots of taxonomic groups, and, in case you neglect multiple operons and DNA extraction and PCR biases, to identify "abundant" groups in your samples. Be careful with detailed considerations and comparisons on the OTU level because of the shortcomings discussed.

If you look on single reads, you will always find outliers in each dataset, but this is not the level of data mining we recommend due to the inevitable limitations of the NGS approach in rDNA-based microbial ecology. Make sure to not invest time into mining and interpretation of noise.

These are general remarks and not specific to SILVAngs which just tries to adequately process the data in this context.

12. Troubleshooting

Failed uploads can have the following reasons:

1. Your file name contains spaces or special characters – replace them by _.
2. Another file with the same name already exists in the project.
3. The file was not a valid FASTA file. Click the detail button to see detailed error messages.
4. The connection dropped. Try GZiping and splitting your files to create smaller ones.

Generally, we experience problems with uploads larger than **500 MB** (per upload not per file). As for now there is no solution for this.

Quality management, sequences are rejected:

Please check the html and PDF report or the file xx--summary.stats in results/ssu/stats/sequence_tags/data in the Archive.zip for detailed information on rejected sequences.

Example (each column represents one sample)

Sample	**1**	**2**	**3**	**4**	**5**
auto-aligned-rejected	0	0	0	0	0
bad alignment identity	1	3	0	0	7
bad alignment quality	0	0	0	1	0
bad base pair score	0	0	0	0	1
Ambiguous	0	0	0	0	0
Homopolymer	0	0	0	5	3
bad quality	0	0	0	0	0
bad length	0	0	0	0	0
Replicate	1052	2223	987	899	1267
Clustered	5116	6409	3775	5252	6456

A project cannot be executed

Please check if you have enough credits in your account to execute your project. Read the **Credit System** section of this user guide and check if you need to request project execution. If you can not solve the issue, do not hesitate to contact us via ngs-contact@arb-silva.de

A former project cannot be found

Please keep in mind that projects that are not in use will be automatically deleted after 180 days. You will receive several warning e-mails before.

Login to SILVAngs is not possible, the system behaves strange

Please check Twitter (https://twitter.com/ARB_SILVA) for the latest status of SILVA and SILVAngs. If no problem has been documented please contact us at ngs-contact@arb-silva.de

13. References

Christian Quast, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Research, 41(D1):D590-D596, 2013. doi: 10.1093/ nar/gks1219.

Yilmaz P, Parfrey LW, Yarza P, Gerken J, Pruesse E, Quast C, Schweer T, Peplies J, Ludwig W, Glöckner FO (2014) The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. Nucleic Acids Research 42:D643-D648. Doi: 10.1093/nar/gkt1209

Anna Klindworth, Elmar Pruesse, Timmy Schweer, Jörg Peplies, Christian Quast, Matthias Horn, and Frank Oliver Glöckner. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucleic Acids Research, 41(1):e1, 2013. doi: 10.1093/nar/gks808.

Elmar Pruesse, Jörg Peplies, and Frank Oliver Glöckner. SINA: accurate high throughput multiple sequence alignment of ribosomal rna genes. Bioinformatics, 2012. doi: 10.1093/bioinformatics/bts252.

V2.1, 01.04.22, SILVAngs 1.9.8 / 1.4.6