



Project: Kerguelen

The SILVAngs Team
April 14, 2020

1 Introduction

The SILVAngs analysis pipeline is primarily targeting the analysis of large scale small- and large subunit (LSU/SSU) ribosomal RNA (rRNA) gene tag sequencing projects but can also be run on meta-genomes studies. Each project normally includes thousands to millions of reads from many different samples produced by massive parallel high-throughput “next generation” sequencing (NGS) technologies. Each read is aligned, quality checked, and classified based on the SILVA Reference alignment and taxonomy. Intuitive graphical outputs are provided for statistical information about the taxonomical distribution of the reads within and across samples. Interactive tax breakdowns are available for detailed inspection of the diversity in the samples.

Processing of the data is performed by five basic modules: *align*, *quality control*, *dereplication*, *clustering*, and *classification*.

In the first step the alignment is used to verify that each read is indeed, depending on the project, an LSU or SSU rRNA gene sequence. Ambiguous reads and reads that are not of the required rRNA gene type, will be rejected based on the alignment score and the alignment identity. This module also checks the sequence quality of each read and it filters out low-quality reads based on ambiguous bases, or too many homopolymers. The number of aligned bases, within the boundaries of the rRNA genes, is determined and sequences below a user defined minimal length cut-off are rejected.

After alignment and quality checks, the remaining sequences are dereplicated, clustered and classified. SILVAngs implements an approach similar to “map and reduce”. First all reads that are 100% identical (allowing overhangs) to another read are marked as replicate by the dereplication module. Next, the clustering module creates clusters of sequences with 98% sequence identity (default) to each other. The longest read in each cluster is then selected as its reference and it is classified against the SILVA reference dataset by the classification module. The resulting classification of the reference sequence of a cluster is mapped to all members of the respective cluster as well as their replicates. Sequences without closely related sequence in the SILVA reference dataset are considered “unclassified” and assigned to the virtual taxonomical group ‘*No Relative*’.

All results can be downloaded as CSV and SVG files. Aligned sequence data can be downloaded in the FASTA and ARB file formats.

If you consider this tool useful and use its results in a publication please consider citing [Quast et al. \(1 January 2013\)](#) for SILVA and the SILVAngs pipeline. The pipeline itself uses the following tools: SINA for the alignment of sequences ([Pruesse et al., 2012](#)), VSEARCH for the dereplication and clustering of sequences ([Rognes et al., 2016](#)), BLASTn for the classification of sequences ([Camacho et al., 2009](#)), and KRONA for some parts of the visualisation of results ([Ondov et al., 2011](#)).

2 Material & Methods

Overview of method for analysis (this can roughly be paraphrased for the purpose of manuscripts and grants):

All sequence reads were processed by the amplicon analysis pipeline of the SILVA project (SILVAngs 1.4) (Quast et al., 1 January 2013). Each read was aligned using the SILVA Incremental Aligner (SINA SINA v1.2.10 for ARB SVN (revision 21008)) (Pruesse et al., 2012) against the SILVA SSU rRNA SEED and quality controlled (Quast et al., 1 January 2013). Reads shorter than 50 aligned nucleotides and reads with more than 2% of ambiguities, or 2% of homopolymers, respectively, were excluded from further processing. Putative contaminations and artefacts, reads with a low alignment quality (50 alignment identity, 40 alignment score reported by SINA), were identified and excluded from downstream analysis.

After these initial steps of quality control, identical reads were identified (dereplication), the unique reads were clustered (OTUs), on a per sample basis, and the reference read of each OTU was classified. Dereplication and clustering was done using VSEARCH (version 2.14.2; <https://github.com/torognes/vsearch>) (Rognes et al., 2016) applying identity criteria of 1.00 and 0.1, respectively. The classification was performed by BLASTn (2.2.30+; <http://blast.ncbi.nlm.nih.gov/Blast.cgi>) (Carmacho et al., 2009) with standard settings using the non-redundant version of the SILVA SSU Ref dataset as classification reference (release 132; <http://www.arb-silva.de>).

The classification of each OTU reference read was mapped onto all reads that were assigned to the respective OTU. This yields quantitative information (number of individual reads per taxonomic path), within the limitations of PCR and sequencing technique biases, as well as, multiple rRNA operons. Reads without any or weak classifications, where the function “(% sequence identity + % alignment coverage)/2” did not exceed the value of 93, remain unclassified. These reads were assigned to the meta group “No Relative” in the SILVAngs fingerprint and Krona charts (Ondov et al., 2011).

This method was first used in the publications of Klindworth et al. (2013) and Ionescu et al. (2012).

SILVAngs uses the de.NBI cloud <http://cloud.denbi.de> to process all user projects. The de.NBI cloud is a fully academic cloud funded by the BMBF¹, free of charge for academic users, which provides compute resources as well as storage capacity for large-scale scientific projects.

¹grant number(s) 031A537B, 031A533A, 031A538A, 031A533B, 031A535A, 031A537C, 031A534A, 031A532B

References

- Christiam Camacho, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas Madden. BLAST+: architecture and applications. *BMC Bioinformatics*, 10(1):421, 2009. ISSN 1471-2105. doi: 10.1186/1471-2105-10-421. URL <http://www.biomedcentral.com/1471-2105/10/421>.
- Danny Ionescu, Christian Siebert, Lubos Polerecky, Yaniv Y. Munwes, Christian Lott, Stefan Häusler, Mina Bižić-Ionescu, Christian Quast, Jörg Peplies, Frank Oliver Glöckner, Alban Ramette, Tino Rödiger, Thorsten Dittmar, Aharon Oren, Stefan Geyer, Hans-Joachim Stärk, Martin Sauter, Tobias Licha, Jonathan B. Laronne, and Dirk de Beer. Microbial and chemical characterization of underwater fresh water springs in the Dead Sea. *PLoS ONE*, 7(6):e38319, 06 2012. doi: 10.1371/journal.pone.0038319. URL <http://dx.doi.org/10.1371%2Fjournal.pone.0038319>.
- 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. URL <http://nar.oxfordjournals.org/content/41/1/e1.abstract>.
- Brian Ondov, Nicholas Bergman, and Adam Phillippy. Interactive metagenomic visualization in a web browser. *BMC Bioinformatics*, 12(1):385, 2011. ISSN 1471-2105. doi: 10.1186/1471-2105-12-385. URL <http://www.biomedcentral.com/1471-2105/12/385>.
- 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. URL <http://bioinformatics.oxfordjournals.org/content/28/14/1823>.
- 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, 1 January 2013. doi: 10.1093/nar/gks1219. URL <http://nar.oxfordjournals.org/content/41/D1/D590.abstract>.
- Torbjørn Rognes, Tomáš Flouri, Ben Nichols, Christopher Quince, and Frédéric Mahé. Vsearch: a versatile open source tool for metagenomics. *PeerJ*, 4:e2584, October 2016. ISSN 2167-8359. doi: 10.7717/peerj.2584. URL <https://doi.org/10.7717/peerj.2584>.

Project Summary

Project Name:	Kerguelen
Project Description:	Default configuration
Sequence Type:	SSU

Number of Samples:	1	
Number of Sequences:	1,339	
Number of Rejected Sequences:	29	(2.17 %)

Raw Sequence Information

Min. Length:	111
Avg. Length:	254
Max. Length:	396

Aligned Sequence Information

Min. Length:	111
Avg. Length:	254
Max. Length:	396

Quality Information (rejected by)

Alignment BP Score:	-
Alignment Identity:	2
Alignment Score:	5
Ambiguous Bases:	-
Homopolymers:	22
Quality:	-
Length:	-

Clustering Information

Number of OTUs:	1,308	(97.68 %)
Number of Clustered Sequences:	2	(0.15 %)
Number of Replicates:	0	(0.00 %)

Classification Information

Number of Classified Sequences:	1,229	(91.78 %)
Number of "No Relative":	81	(6.05 %)

Project Settings

Alignment

SINA Version:	SINA v1.2.10 for ARB SVN (revision 21008)
Min Align. Identity (%):	50
Min Align. Score:	40
Min Basepair Score (%):	30

Quality Control

Min. Sequence Quality (%):	30
Min. Length (aligned nuc.):	50
Max. Ambiguities (%):	2
Max. Homopolymers (%):	2

Clustering

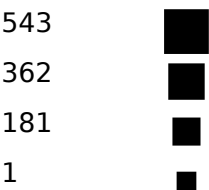
VSEARCH Version:	2.14.2
Min. OTU Identity (%):	1

Classification

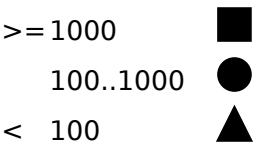
BLASTn Version:	2.2.30+
Reference:	SILVA
Reference Version:	132
Similarity (%):	93

Taxonomic Fingerprint at Phylum Level

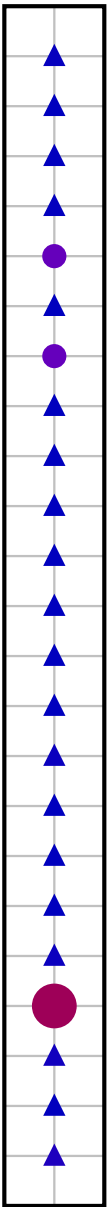
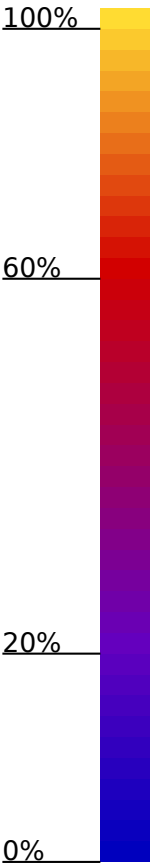
OTUs / Path



Sequences



Fraction of Sample



- Euryarchaeota
- Nanoarchaeaeota
- Thaumarchaeota
- Actinobacteria
- Bacteroidetes
- Chloroflexi
- Cyanobacteria
- Dadabacteria
- Dependientiae
- Firmicutes
- Fusobacteria
- Kiritimatiellaeota
- Lentisphaerae
- Margulisbacteria
- Marinimicrobia (SAR406 clade)
- Nitrospinae
- Omnitrphicaeota
- PAUC34f
- Planctomycetes
- Proteobacteria
- Verrucomicrobia
- Opisthokonta
- No Relative

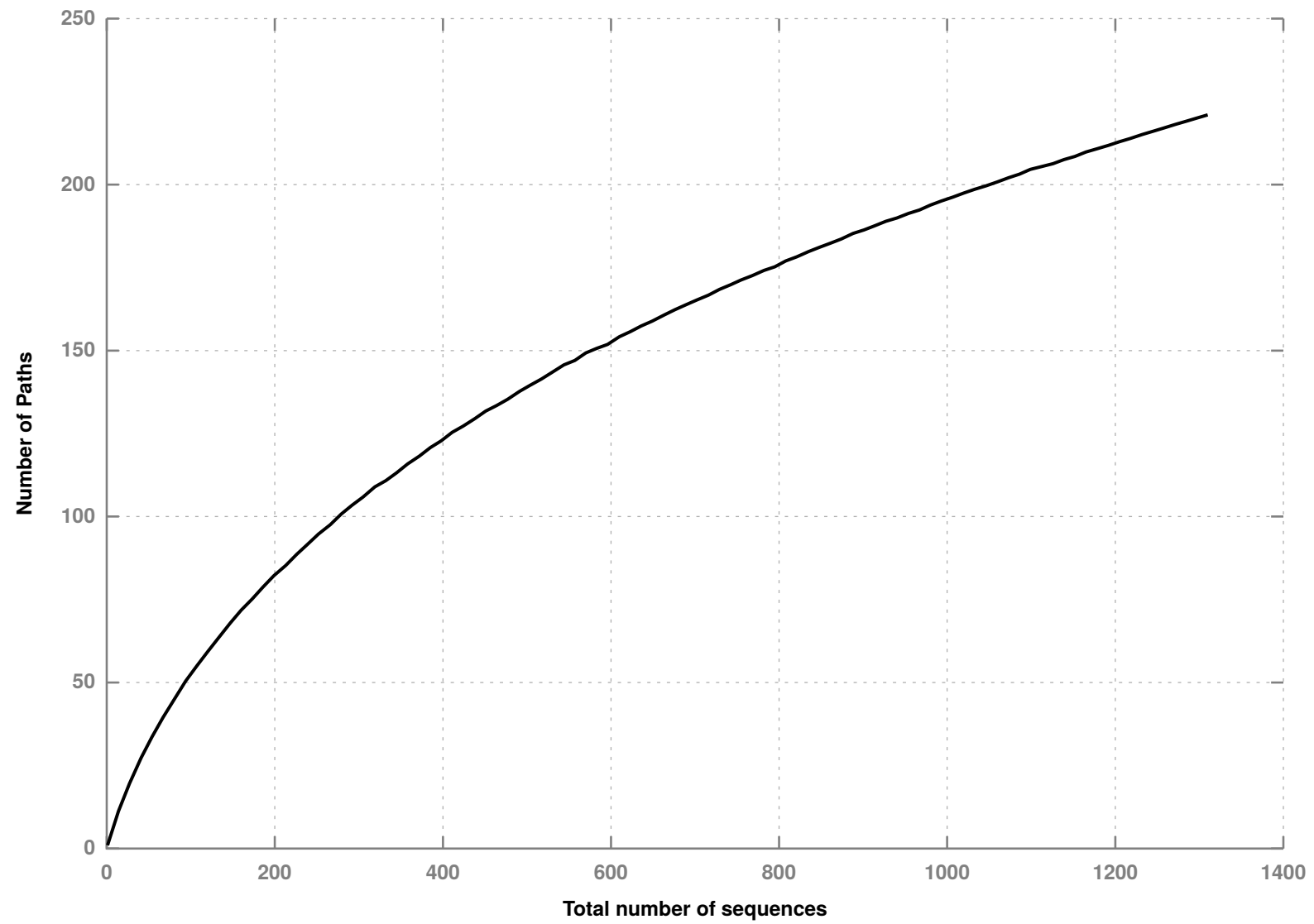
3

2

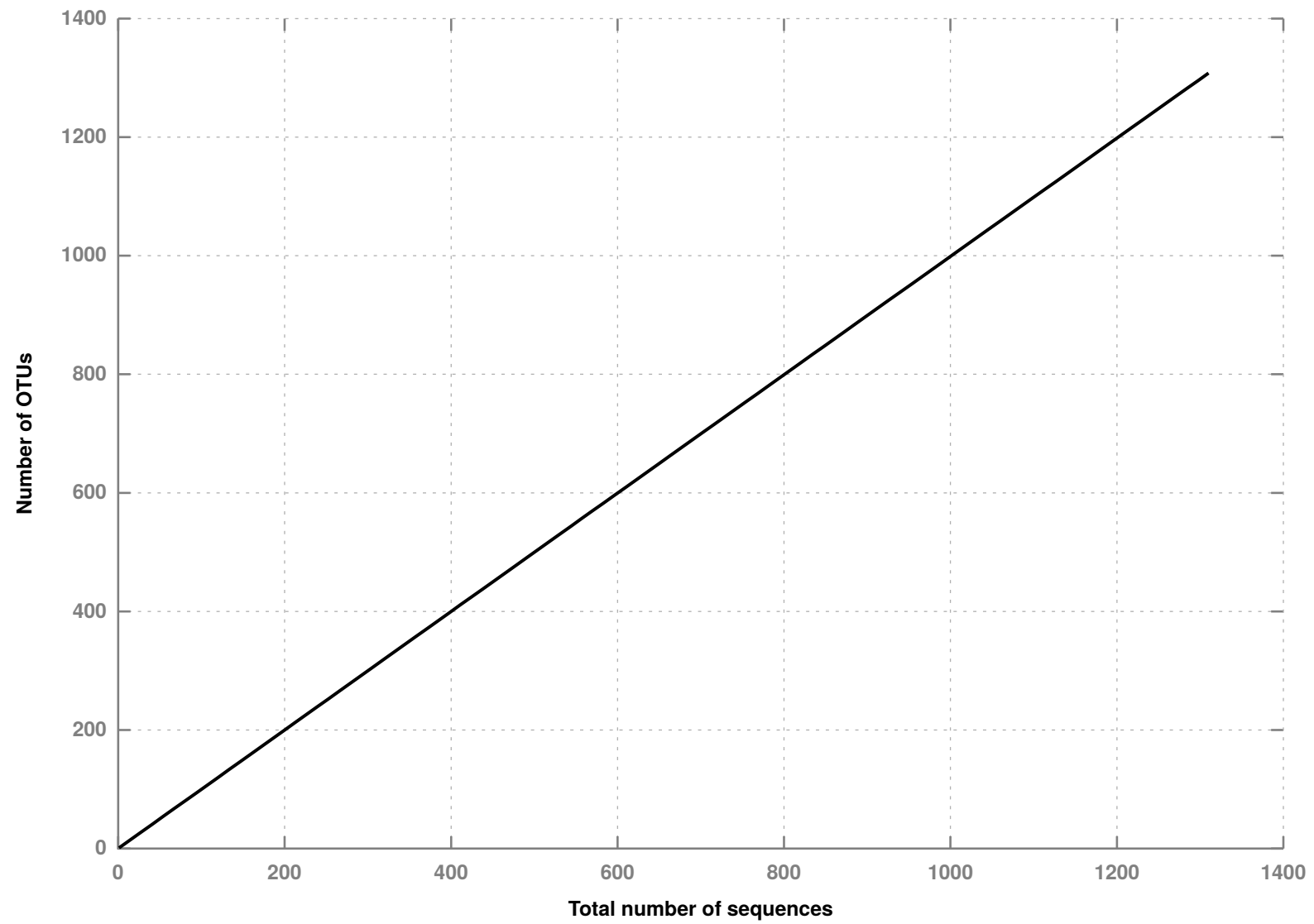
1

Kerguelen_16S_ASVs

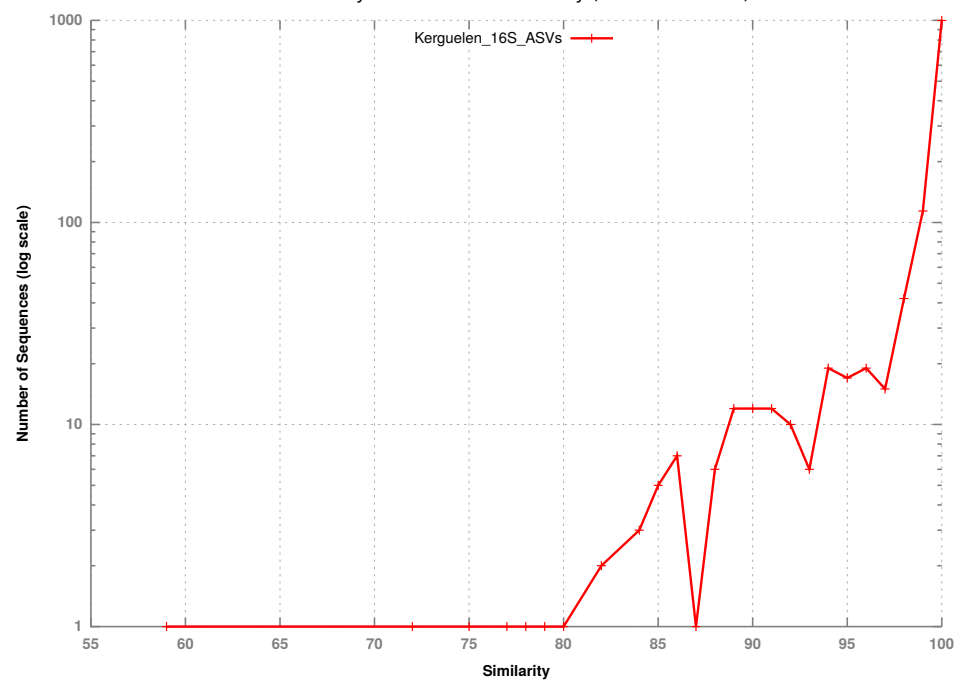
Rarefaction - Overview



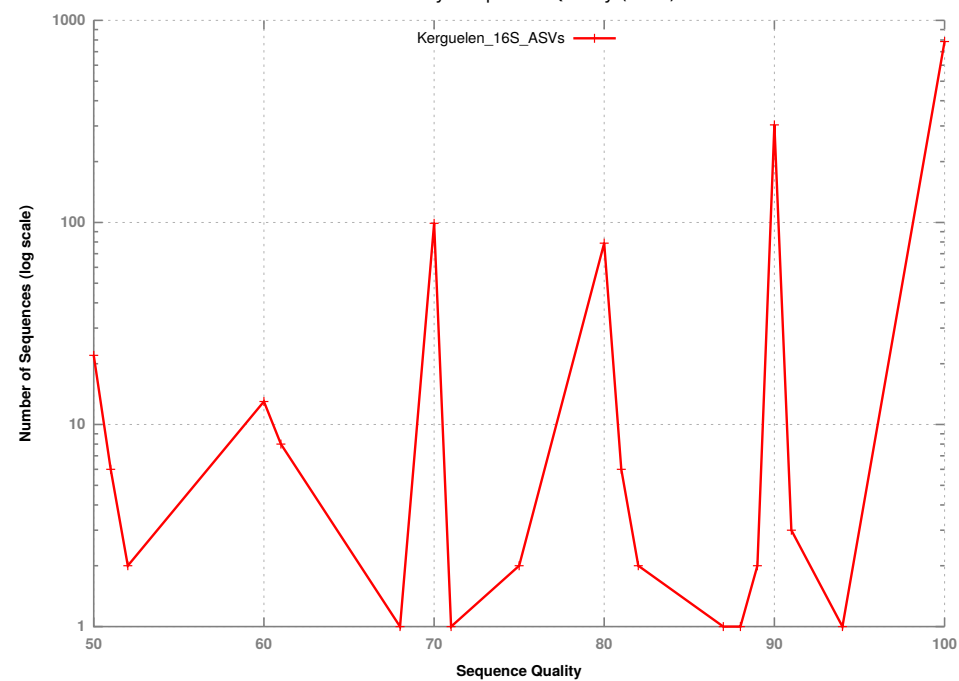
Rarefaction - Overview



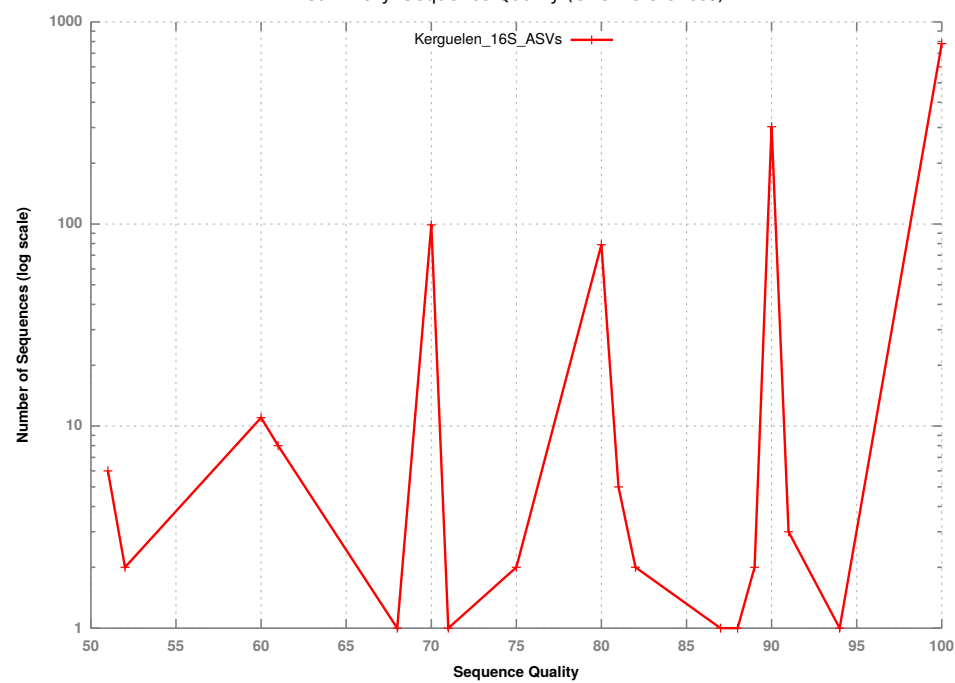
summary: Classification Similarity (OTU-References)



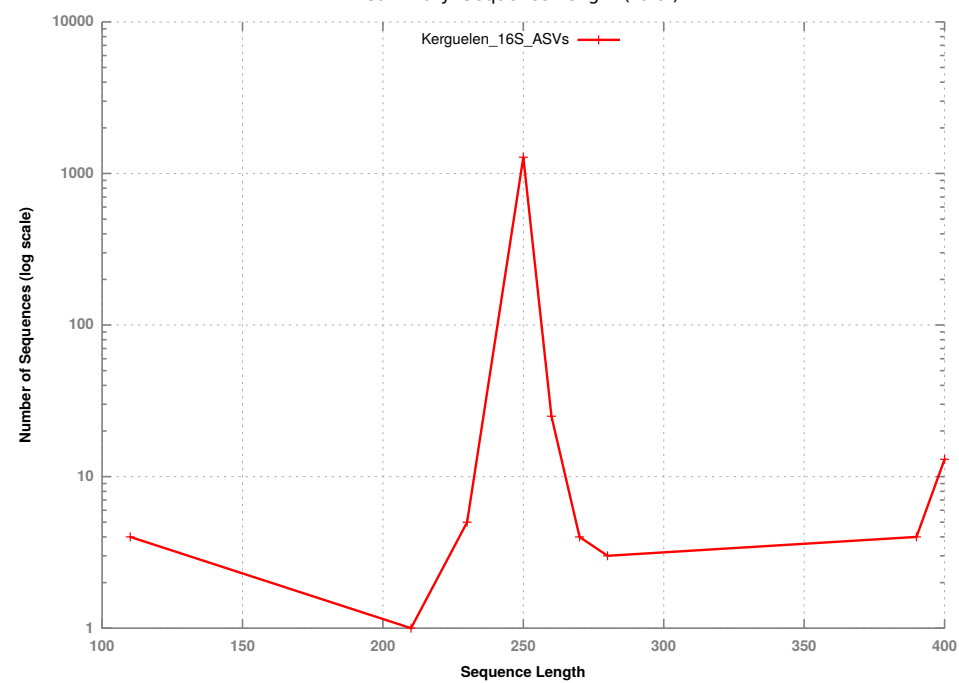
summary: Sequence Quality (Total)



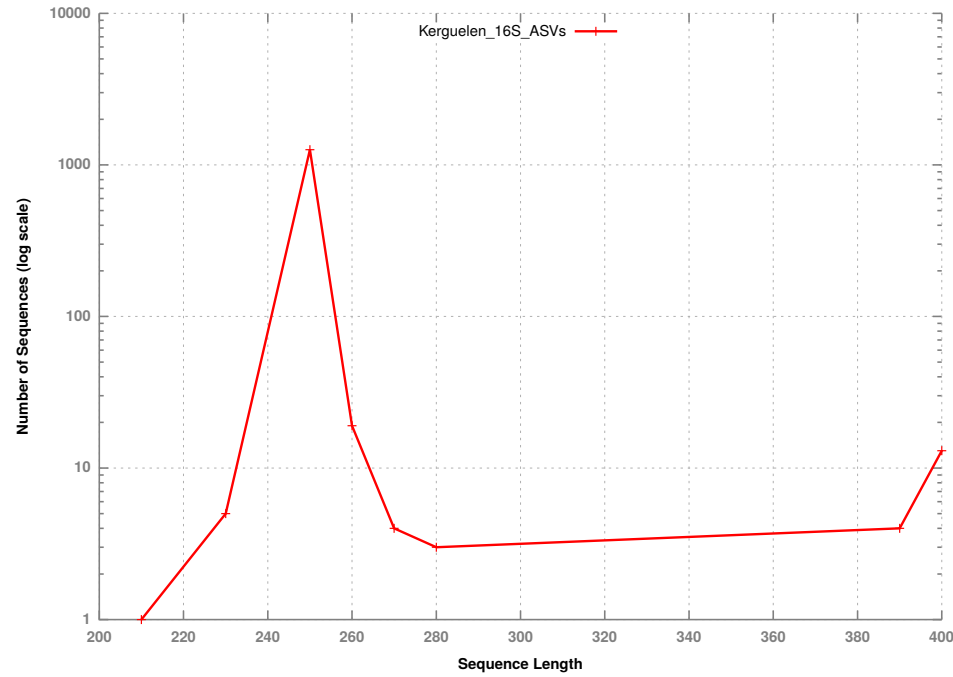
summary: Sequence Quality (OTU-References)



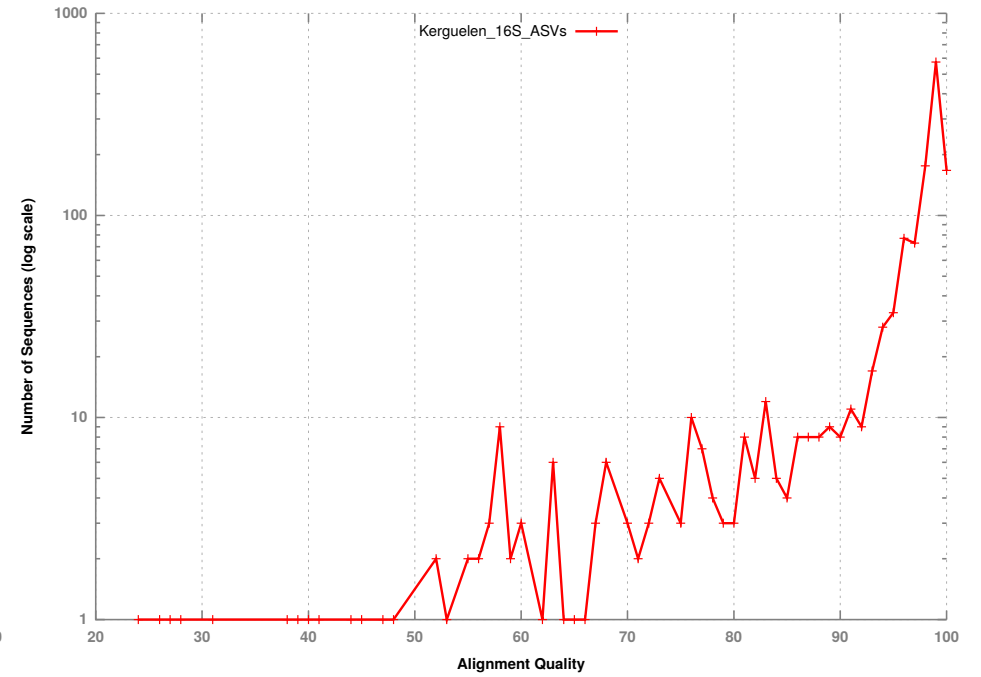
summary: Sequence Length (Total)



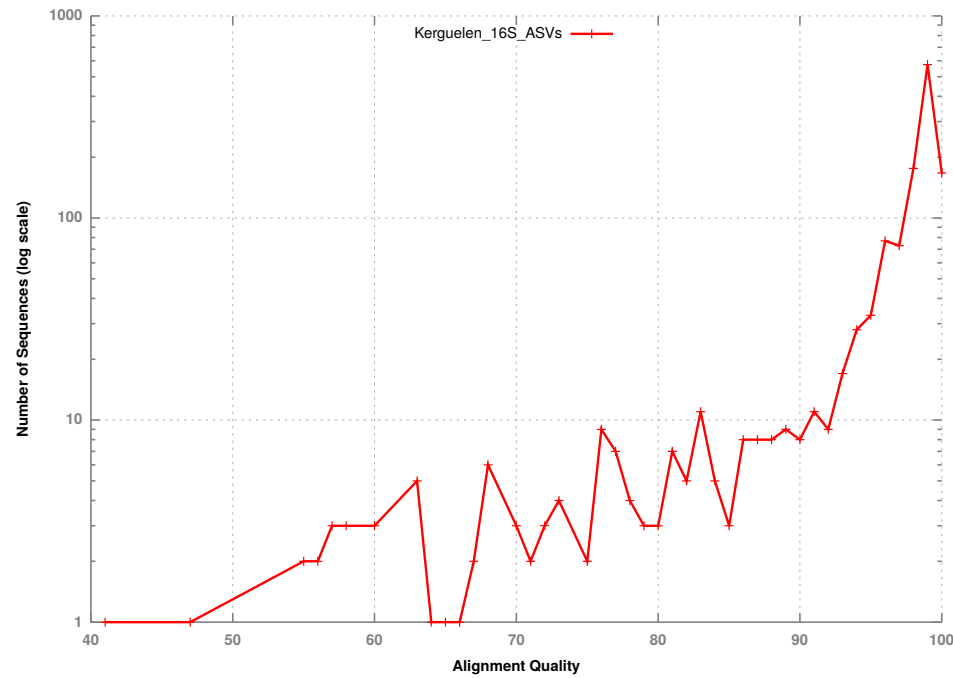
summary: Sequence Length (OTU-References)



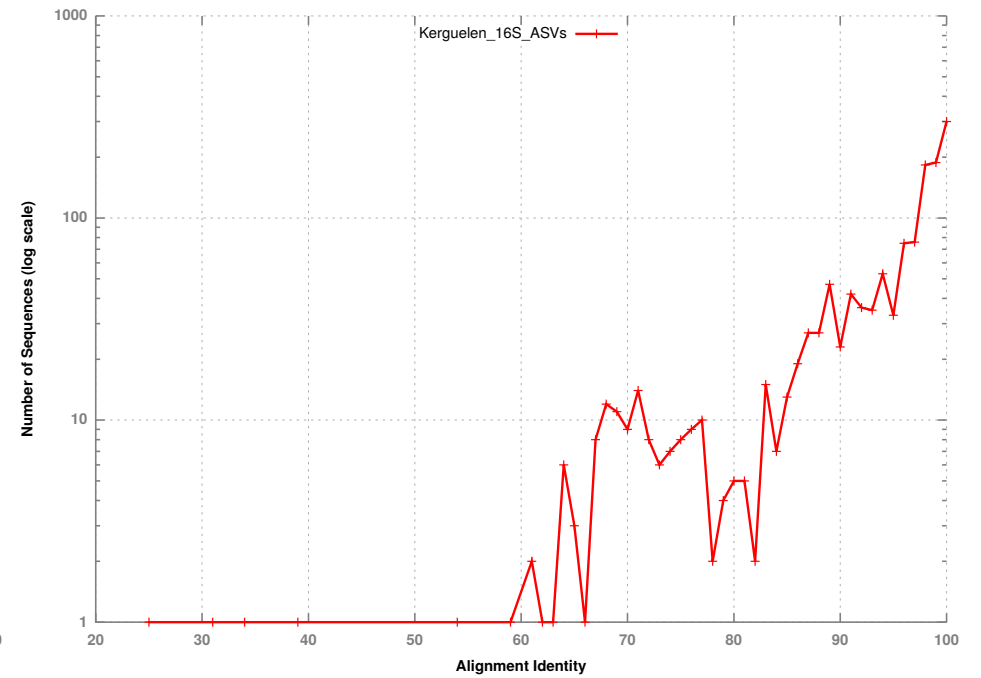
summary: Alignment Quality (Total)



summary: Alignment Quality (OTU-References)



summary: Alignment Identity (Total)



summary: Alignment Identity (OTU-References)

