# SMURF software package

**SMURF: Short MUltiple Regions Framework** – is software package that produces a full taxonomic description of the microbial community based on sequencing of any number of short regions.

## Instalation

Download the SMURF package and decompress all the data base files and folders as follows:

“tar xf ./[Green\_Genes\_201305](https://github.com/NoamShental/SMURF/tree/master/Green_Genes_201305)/unique\_up\_to\_3\_ambiguous\_16S/RDP\_description.tar”

“gunzip ./[Green\_Genes\_201305](https://github.com/NoamShental/SMURF/tree/master/Green_Genes_201305)/unique\_up\_to\_3\_ambiguous\_16S/[Green\_Genes\_201305](https://github.com/NoamShental/SMURF/tree/master/Green_Genes_201305)\_unique\_up\_to\_3\_ambiguous\_16S.fasta.gz”

## Inputs

**Sequences read** – fastq files with the sequenced reads. An example sample sequenced with paired end Illumina MiSeq is provided.

**Ad hoc database** – a database of per region k-mers (an example of the DB for the 6 regions primers is provided with the software package)

## Output

The result of the profiling is csv file with reconstructed groups information including:

Group frequency – the frequency assigned to the group

Read count – number of reads assigned to the group

Number of sequences – number od sequences (ambiguity) of the group

Taxonomy – all the result from RDP sequence matching

## Example

To profile a single sample use a script named profile\_one\_sample.m

## Parameters description

### Sample parameters

**base\_samples\_dir** - is the directory where all the samples are located

**sample\_name** – is the name of the directory where fastq files for specific sample are located. Notice that the fastq files must be named using the following convention. If the sample\_name=’Example’, then for paired end sequencing the files will be names: Example\_L001\_R1\_001.fastq and Example\_L001\_R2\_001.fastq

**primer\_set\_name** – the name of the primers set used with the sample

**kmer\_len** – the length of the k-mer to be used for profiling

### Data preprocessing parameters

**data\_type** – specify whether the reads have quality score. Possible values are ‘fasta’ or ‘fastq’

**pe\_flag** - 0/1 flag specifying if the sequencing was single/paired end respectively

**qual\_th** – minimal quality required in at least **in prc\_high\_qual** base pairs

**prc\_high\_qual** – minimal required proportion of base pairs with Phred score above **qual\_th** per read

**low10\_th** – maximal number of base pairs allowed to have Phred score below 10

**max\_num\_Ns** – maximal number of ambiguous nucleotides allowed per read

**algo\_pe\_flag** – 0/1 flag specifying whether the reconstruction should be performed assuming single/paired end respectively. This parameter will usually be equal **pe\_flag,** although **pe\_flag** = 1 and **algo\_pe\_flag** = 0 is allowed, while **pe\_flag** = 0 and **algo\_pe\_flag** = 1 is not possible.

**max\_err\_inprimer** - maximal number of mismatches allowed between the primer sequence and the read for assigning the read to a region

**with\_primer\_flag** – 0/1 flag specifying whether to remove the primers after assignment of read to regions

### Algorithm parameters

**uniS16\_dir** – directory of the reference data base used for profiling

**db\_filename** – name of the reference (fasta) data base file without extension

**filter\_reads** – 0/1 flag specifying whether to apply the low abundance data preprocessing filter

**min\_read\_freq** - minimal required frequency for a read per region to pass the low abundance reads filter

**min\_read\_count** - minimal required count for a read per region to pass the low abundance reads filter

**nMM\_cut** – maximal number of mismatches allowed when matching reads to k-mers

**pe** – probability of error per nucleotide assumed by the algorithm

**do\_filter** - 0/1 flag specifying whether to apply the data preprocessing bacteria filter

**regions\_normalization\_scheme** - 'is amplified' (default)/ 'none' / 'perfect match' / 'any match'

**mixture\_type** - 'Multiplex' (default)/'RegionByRegion'

**tol** – maximal L1 change in the estimate of read proportions vector between algorithm iterations

**numIter** – maximal number of iteration of the reconstruction algorithm