Research Center Borstel
Leibniz-Center for Medicine and Biosciences

A computational pipeline for whole genome sequencing analysis of *Mycobacterium tuberculosis* complex isolates

TBseq

[DESCRIPTION]

TBseq is a semi-automated pipeline for mapping, variant calling and detection of resistance mediating variants as well phylogenetic variants in user provided *Mycobacterium tuberculosis* samples. The pipeline consists of open source software for read mapping (SAMTOOLS, BWA, SAMBAMBA), base call recalibration and refinement (PICARD, GATK) and variant calling (GATK, SAMTOOLS). TBseq has a modular architecture that allows to repeat specific tasks (e.g. mapping, refinement, variant calling, strain Identification, etc.).

TBseq is able to perform a joined variant analysis from user provided samples. This is especially useful in transmission chain analysis. Phylogenetic SNPs are used to create FastA formatted files (.fasta). These files can be used to reconstruct minimum spanning trees (MS) or phylogenetic trees (NJ or ML). All additional outputs are written into files with tabular delimited file format (.tab). This format is readable with Microsoft Excel or Excel-like programs.

[REQUIREMENTS]

Perl: perl 5, version 18, subversion 2 (v5.18.2) or higher
Java: openjdk version "1.8.0_91" or higher

TBseq uses the following CPAN modules:

- MCE (v1.810)
- Statistics::Basic (v1.6611)
- FindBin (v1.51)
- Cwd (v3.62)
- Getopt::Long (v2.49)
- File::Copy (v2.31)
- List::Util (v1.47)
- Exporter (v5.72)
- vars (v1.03)
- lib (v0.63)
- strict (v1.11)
- warnings (v1.36)

TBseq uses the following third party software:

- bwa (v0.7.15)
- GenomeAnalysisTK (v3.6)
- IGVTools (v2.3.88)
- picard (v2.7.1)
- sambamba (v0.6.5)
- samtools (v1.3.1)

[INSTALLATION]

1. **Download TBseq from github and extract:**
   https://github.com/TaKohl/TBseq_source/
   Or clone the repository:
   git clone https://github.com/TaKohl/TBseq_source
   Check if TBseq.pl and the executables from the opt/ directory are executable or change permissions with chmod

2. **To make sure TBseq is executable from anywhere on your system add the TBseq_source folder to your PATH or create a symlink of TBseq.pl to a folder that is already in your PATH or call the TBseq.pl with the complete path information, e.g.**
   perl /home/$USER/path/to/TBseq_source/TBseq.pl

3. **Install modules via CPAN by typing in the command-line:**
   Cpan

4. **Install the modules by tying:**
   install MCE
   install Statistics::Basic
   install FindBin
   install Cwd
   install Getopt::Long
   install File::Copy
   install List::Util
   install Exporter
   install vars
   install lib
   install strict
   install warnings
5. **If third party programs (BWA and SAMTOOLS) in TBseq/opt/ are not working, try to re-compile them. The re-compiled executables MUST be located within the appropriate folders**.
   ./configure --prefix = [PATH_TO_YOUR_TBSEQ]/opt/bwa_0.7.15/
   ./configure --prefix = [PATH_TO_YOUR_TBSEQ]/opt/samtools_1.3.1/

Tested on ubuntu 16.04 LTS.

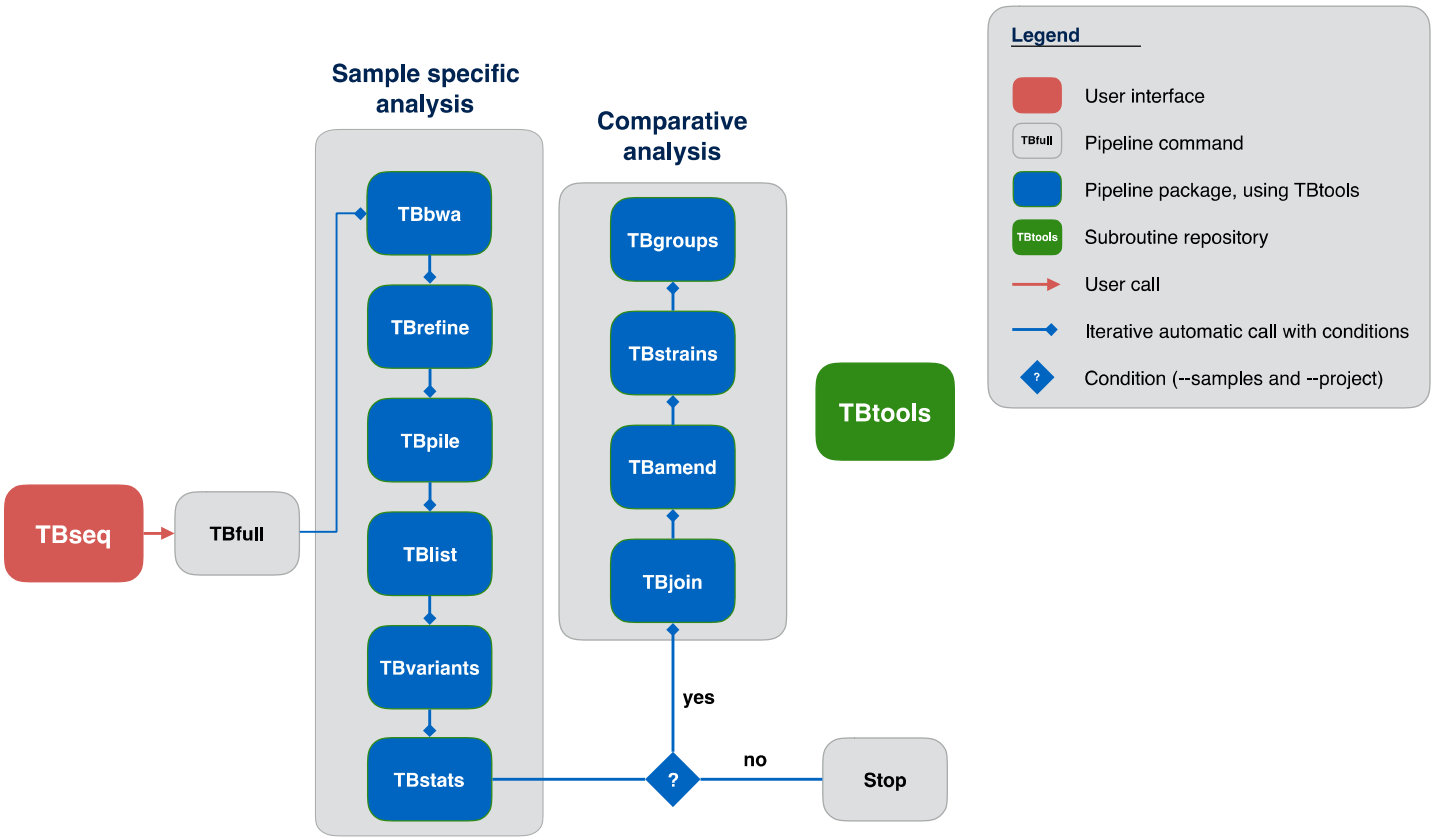
[OVERVIEW]

[SYNOPSIS]

You do not need to support TBseq with a path to your read files. Just make sure that your read files are located in a folder where you want to execute TBseq! The basic execution of TBseq on the command-line looks like this:

TBseq [OPTIONS] [VALUES]

In the following chapter, you will be introduced to the OPTIONS and the corresponding VALUES that you can use in TBseq.



[OPTIONS & VALUES]

[OPTIONS]

[**VALUES]**

--step

TBseq has a modular architecture. This ensures that you do not need to repeat the whole pipeline, if something went wrong. You can execute the full pipeline or start with certain pipeline steps. However, if you choose a specific pipeline step, make sure that you have the file dependencies for this step. The OPTION **--step** is essential and requires a VALUE! In the following, all possible VALUES are described:

TBfull

VALUE of **--step** for executing the whole pipeline. If you choose this VALUE, make sure that you setup all other OPTIONS to your appropriate VALUES (see OPTIONS below).

TBbwa

VALUE of **--step** for mapping read files (.fastq.gz) to a reference genome, using BWA mem.
This is the first step within the pipeline. Depending on the library type (single-end or paired-end), this module will align single fastq file or the paired R1 and R2 fastq files on a reference genome.
It is required that your read files have **at least** three fields separated by underscores as follow and contain no other special characters other than hyphens:

[SampleID]_[LibID]_[*]_[Direction].fastq.gz

The **[SampleID]** is a unique sample identifier. If your analysis has more than one sample, make sure that this identifier is unique across samples! The **[LibID]** is an identifier for your used sequencing library. The **[Direction]** field is an **essential** field and indicates if reads are forward (R1) or reverse (R2) in orientation. Single-end data has to be used with R1. There can be several **[*]** fields, which can contain additional information.
On default, TBseq uses the *M. tuberculosis* H37Rv genome as a reference. BWA mem is executed with a default VALUE setting. After mapping, files (.sam) are converted into binary mapping files (.bam). The mapping is sorted, indexed and putative PCR duplicates are removed, using the program SAMTOOLS. Wherever possible, multi-threading is activated by the user provided VALUES for the OPTION **--threads** (see OPTIONS below).

Input
[SampleID]_[LibID]_[Direction].fastq.gz
Output
Bam/[SampleID]_[LibID].bam
Bam/[SampleID]_[LibID].bai
Bam/[SampleID]_[LibID].bamlog

TBrefine

VALUE of **--step** for realignment around insertions and deletions (indels) and base call recalibration, using the program GATK. The GATK program uses default VALUES with the exception of:

--downsample_to_coverage 10000
--defaultBaseQualities 12
--maximum_cycle_value 600

--noOriginalAlignmentTags

This internal VALUE setting for GATK ensures that you can also use Ion Torrent read files. For the base call recalibration, a set of known MTB resistance SNPs is used if you take *Mycobacterium tuberculosis* H37Rv as a reference genome. If not, this step will be skipped. The calibration list is stored in the directory "var/res/Base_Calibration_List.vcf" of the package.

Input
Bam/[SampleID]_[LibID].bam
Output
GATK_Bam/[SampleID]_[LibID].gatk.bam
GATK_Bam/[SampleID]_[LibID].gatk.bai
GATK_Bam/[SampleID]_[LibID].gatk.bamlog
GATK_Bam/[SampleID]_[LibID].gatk.grp
GATK_Bam/[SampleID]_[LibID].gatk.intervals

TBpile

VALUE of **--step** for creating pileup file(s) (.mpileup) from refined mapping file(s) (.gatk.bam), using the program SAMTOOLS. The SAMTOOLS program uses default VALUES with the exception of:

-B
-A

see: http://www.htslib.org/doc/samtools.html for more information about this VALUE setting. TBseq needs to create this file format in order to perform all downstream analysis.

Input
GATK_Bam/[SampleID]_[LibID].gatk.bam
Output
Mpileup/[SampleID]_[LibID].gatk.mpileup
Mpileup/[SampleID]_[LibID].gatk.mpileuplog

TBlist

VALUE of **--step** for creating position list(s) from pileup file(s) (.gatk.mpileup). The position list consists of 21 columns, representing the frequency of nucleotide compositions for each genomic position. This step can be executed with a user provided VALUE of the OPTION **--threads** (see below) and a different VALUE of the OPTION **--minbqual**. The columns of the output are:

Pos	Indicates the genome position
Insindex	An index for reporting insertion sites. 0 means the level of the reference genome
RefBase	The nucleotide found for the reference genome at this position
As	The forward read frequency of the nucleotide adenine
Cs	The forward read frequency of the nucleotide cytosine
Gs	The forward read frequency of the nucleotide guanine
Ts	The forward read frequency of the nucleotide thymine
Ns	The forward read frequency of ambiguous nucleotides
GAPs	The forward read frequency of a GAP
as	The reverse read frequency of the nucleotide adenine
cs	The reverse read frequency of the nucleotide cytosine
gs	The reverse read frequency of the nucleotide guanine
ts	The reverse read frequency of the nucleotide thymine
ns	The reverse read frequency of ambiguous nucleotides
gaps	The reverse read frequency of a gap
Aqual	Number of adenine nucleotides having a higher or equal phred 20 score

Cqual	Number of cytosine nucleotides having a higher or equal phred 20 score
Gqual	Number of guanine nucleotides having a higher or equal phred 20 score
Tqual	Number of thymine nucleotides having a higher or equal phred 20 score
Nqual	Number of ambiguous nucleotides having a higher or equal phred 20 score
GAPqual	Number of GAPs having a higher or equal phred 20

Input
Mpileup/[SampleID]_[LibID].gatk.mpileup
Output
Position_Tables/[SampleID]_[LibID].gatk_position_table.tab

TBvariants

VALUE of **--step** for variant calling from position list(s). On default, an alternative allele is called only if:

- At least four forward reads (**--mincovf 4**) and four reverse reads (**--mincovr 4**) support the allele.
- At least four reads show a phred score ≥ 20 (**--miphred20 4**).
- The majority allele is present with a frequency of 75% or higher (**--mifreq 75**).

If not, the allele is indicated as ambiguous.

The module can be executed with the OPTIONS **--all_vars**, **--snp_vars** and **--lowfreq_vars** (see OPTIONS below).

Your setting of this OPTIONS will be visible at the end of the output files as a binary string (e.g. "outmode100" means that you activated **--all_vars** but not **--snp_vars** or **--lowfreq_vars**).

Input
Position_Tables/[SampleID]_[LibID].gatk_position_table.tab
Output
Called/[SampleID]_[LibID].gatk_position_uncovered_[mincovf]_[mincovr]_[minfreq]_[minphred20]_[all_vars][snp_vars][lowfreq_vars].tab
Called/[SampleID]_[LibID].gatk_position_variants_[mincovf]_[mincovr]_[minfreq]_[minphred20]_[all_vars][snp_vars][lowfreq_vars].tab

TBstats

VALUE of **--step** for basic mapping statistics and called variant statistics, using SAMTOOLS flagstat. This step creates a tabular delimited file "Mapping_and_Variant_Statistics.tab". The file stores all sample statistics and is updated instantly, if you use different samples at different time points. The columns of the output are:

Date	The date of TBseq execution
SampleID	The analyzed sample
LibraryID	There used library for the sequencing
FullID	Complete sample name
Total Reads	The total amount of sequenced reads
Mapped Reads (%)	The number of reads mapped
Genome Size	The size of the reference genome
Genome GC	**T**he GC content of the reference genome
(Any) Total Bases (%)	Number of bases used for the called variants
(Any) GC-Content	GC content calculated from the reads used for the called variants
(Any) Coverage mean / median	Coverage mean and median calculated from the reads used for the called variants
(Unambiguous) Total Bases (%)	Number of unambiguous bases used for the called variants
(Unambiguous) GC-Content	GC content calculated from the unambiguous reads used for the called variants
(Unambiguous) Coverage mean / median	Coverage mean and median calculated from the unambiguous reads used for the called variants
SNPs	Number of SNPs found

Deletions	Number of deletions found
Insertions	Number of insertions found
Uncovered	Uncovered positions
Substitutions (including Stop Codons)	Number of substitutions within genes

Input
Bam/[SampleID]_[LibID].bam
Position_Tables/[SampleID]_[LibID].gatk_position_table.tab
Output
Statistics/Mapping_and_Variant_Statistics.tab

TBjoin

VALUE of **--step** for creating a joint SNP analysis of user provided samples. First, a scaffold of all variant positions is built from variant files of your provided samples (**--samples**). The samples need to be provided in a tabular delimited file with **[SampleID]** in column 1 and **[LibID]** in column 2. You also need to provide a project name (**--project**) for file naming, otherwise the default file name will start with "NONE". Second, variants are recalculated with the OPTION **--all_vars**. The output is a table of concatenated variant files from user provided samples. The first line within the tabular delimited file is a sample header line. The second line describes the joint analysis for variant positions and is separated to joint fields and sample specific fields:

Joint fields

#Position	Genome position with a variant in at least one of the samples.
Insindex	An index for reporting insertion sites. 0 means the level of the genome. If Insindex > 0, then at least one sample showed an insertion at this position.
Ref	The reference allele present at this position
Gene	A gene ID is indicated, if the position is within a gene
GeneName	A gene name is indicated, if available
Annotation	The product of the gene

Sample specific fields

Type	The type of the variant. Possible values are "none, SNP, Del, Unc, -"
Allele	The allele that was present for a sample at this position. Possible values are "AGTC, GAP, U, -"
CovFor	The forward coverage at this position
CovRev	The reverse coverage at this position
Qual20	The number of nucleotides having a phred score above 20
Freq	The frequency of the allele
Cov	The coverage at this position
Subst	The quality of a non-synonymous substitution, if the SNP occurs in a gene

Input
samples.txt
--project
Called/[SampleID]_[LibID].gatk_position_variants_[mincovf]_[mincovr]_[minfreq]_[minphred20]_[all_vars][snp_vars][lowfreq_vars].tab
Position_Tables/[SampleID]_[LibID].gatk_position_table.tab
Output
Joint/[PROJECT]_joint_[mincovf]_[mincovr]_[minfreq]_[minphred20]_samples.tab
Joint/[PROJECT]_joint_[mincovf]_[mincovr]_[minfreq]_[minphred20]_samples.log

TBamend

VALUE of **--step** for amending joint variant tables. If previously set, you need to provide again the project name with **–project .**

Only variants that are unambiguous in 95% of all samples are reported (**--unambig 95**) on default. Finally, SNPs that occur in a distance of 12 nucleotides are excluded in order to reduce

false positive calls **(--window 12**). SNPs in repetitive regions or nested within a resistance gene are excluded. Positions not passing this criteria are reported in the "removed" output.

Input
samples.txt
--project
Joint/[PROJECT]_joint_[mincovf]_[mincovr]_[minfreq]_[minphred20]_samples.tab
Output
Amend/[PROJECT]_joint_[mincovf]_[mincovr]_[minfreq]_[minphred20]_samples_amended.tab
Amend/[PROJECT]_joint_[mincovf]_[mincovr]_[minfreq]_[minphred20]_samples_amended_[unambig]_phylo.tab
Amend/[PROJECT]_joint_[mincovf]_[mincovr]_[minfreq]_[minphred20]_samples_amended_[unambig]_phylo_[window].tab
Amend/[PROJECT]_joint_[mincovf]_[mincovr]_[minfreq]_[minphred20]_samples_amended_[unambig]_phylo_[window]_removed.tab
Amend/[PROJECT]_joint_[mincovf]_[mincovr]_[minfreq]_[minphred20]_samples_amended_[unambig]_phylo_plainIDs.fasta
Amend/[PROJECT]_joint_[mincovf]_[mincovr]_[minfreq]_[minphred20]_samples_amended_[unambig]_phylo_[window]_plainIDs.fasta
Amend/[PROJECT]_joint_[mincovf]_[mincovr]_[minfreq]_[minphred20]_samples_amended_[unambig]_phylo.fasta
Amend/[PROJECT]_joint_[mincovf]_[mincovr]_[minfreq]_[minphred20]_samples_amended_[unambig]_phylo_[window].fasta

TBstrains

VALUE of **--step** for lineage classification based on phylogenetic SNP maps (1,2). This module creates a tabular delimited file within the "Classification" directory. Within this file, the majority lineage of a sequenced sample is reported. This file is updated instantly, if you use different samples at different time points.

Quality: If any other allele than AGTCagtc is called the quality is “ugly”. If all phylogenetic SNPs used for the classification have a frequency of 75% and are covered 10-fold, then lineage classification will result in a "good" quality, otherwise in a "bad" quality for that sample.

Input
Position_Tables/[SampleID]_[LibID].gatk_position_table.tab
Output
Classification/Strain_Classification.tab

TBgroups

VALUE of **--step** for grouping samples based on SNP distances. This module is able to use one user provided threshold for SNP distances. Strains are grouped together, if they are not more than 12 SNPs apart from each other (**--distance 12**). If they are more than 12 SNPs apart from each other, a new group is formed. This module enables to distinguish between groups of samples via SNP distance patterns and creates a distance matrix.

Input
Amend/[PROJECT]_joint_[mincovf]_[mincovr]_[minfreq]_[minphred20]_samples_amended_[unambig]_phylo_[window].tab
Output
Groups/[PROJECT]_joint_[mincovf]_[mincovr]_[minfreq]_[minphred20]_samples_amended_[unambig]_phylo_[window].matrix
Groups/[PROJECT]_joint_[mincovf]_[mincovr]_[minfreq]_[minphred20]_samples_amended_[unambig]_phylo_[window]_[distance].groups

--continue

If a module was chosen with the **--step** OPTION, the **--continue** OPTION ensures that the pipeline will continue with downstream modules. You do not need to set this OPTION, if you start the analysis with the **--step TBfull** VALUE.

--samples

This OPTION requires a user sample file (e.g. samples.txt) as a VALUE. The file must be a two-column, tab-separated file. Column 1 has to be your **[SampleID]**. Column 2 has to be your **[LibID]**. **TBjoin** requires this file!

--project

This OPTION takes a project name for the steps **TBjoin**, **TBamend** and **TBgroups**. If you do not support a project name, **[NONE]** is used as a default value.

--ref

This OPTION sets the reference genome for the read mapping. On default, [M._tuberculosis_H37Rv_2015-11-13] will be used for the mapping. Place the reference fasta file under /TBseq_source/var/ref/. Give the name of the reference file without .fasta extension.

--resilist

This OPTION sets a list of known resistance mediating SNPs for resistance prediction. Give the full path to the file.
The structure of the file can be seen here:
/TBseq_source/var/res/MTB_Resistance_Mediating.txt

--intregions

This OPTION sets a list of interesting regions for annotation of variants. Give the full path to the file.
The structure of the file can be seen here:
/TBseq_source/var/res/MTB_Extended_Resistance_Mediating.txt

--categories

This OPTION specifies a gene categories file to annotate essential and non-essential genes as well as repetitive regions. SNPs in repetitive regions will be excluded for phylogenetic analysis. Give the full path to the file.
The structure of the file can be seen here:
/TBseq_source/var/cat/MTB_Gene_Categories.txt

--basecalib

This OPTION specifies a file for base quality recalibration. The list must be in VCF format and should contain known SNPs. Give the full path to the file.
The structure of the file can be seen here:
/TBseq_source/var/res/Base_Calibration_List.vcf

--all_vars

This OPTION is used in **TBvariants**, **TBstats**, **TBjoin** and **TBstrains**. On default, the OPTION is set to 0. The OPTION has influence on the variant output. If you set this OPTION, every position will be reported ignoring, if the position passes the filter criteria, is uncovered, is a reference allele or is coming from an insertion. Important is that low frequency alleles are only reported, if **—lowfreq_vars** is supported in addition. If not, only majority alleles are reported.

--snp_vars

This OPTION is used in **TBvariants**, **TBstats**, **TBjoin** and **TBstrains**. On default, the OPTION is set to 0. The OPTION has influence on the variant output. If you set this OPTION, only SNPs will be reported in the variant output file.

--lowfreq_vars

This OPTION is used in **TBvariants**, **TBstats**, **TBjoin** and **TBstrains**. On default, the OPTION is set to 0. The OPTION has influence on the variant output. If you set this OPTION, TBseq will consider alternative low frequency variants at positions where a majority reference allele is found. This is useful for analysing subpopulations within a sample (e.g. mixed infection).
Adjust **--mincovf**, **--mincovr**, **--minphred20** and **–minfreq** accordingly.

--minbqual

This OPTION is used in **TBlist**. On default, the OPTION is set to 13. The OPTION sets a threshold for the mapping quality. Bases covering a position are only considered, if the quality is greater or equal this VALUE.

--mincovf

This OPTION is used in **TBvariants**, **TBjoin**, **TBamend** and **TBstrains**. On default, the OPTION is set to 4. The OPTION sets a minimum forward read coverage threshold. Alleles must have a forward coverage of this VALUE or higher to be considered.

--mincovr

This OPTION is used in **TBvariants**, **TBjoin**, **TBamend** and **TBstrains**. On default, the OPTION is set to 4. The OPTION sets a minimum reverse read coverage threshold. Alleles must have a reverse coverage of this value or higher to be considered.

--minphred20

This OPTION is used in **TBvariants**, **TBjoin**, **TBamend** and **TBstrains**. On default, the OPTION is set to 4. The OPTION sets a minimum read coverage with a phred 20 quality score. A user provided number of reads must show a phred quality above or equal 20 for a certain position to be considered.

--minfreq

This OPTION is used in **TBvariants**, **TBjoin**, **TBamend** and **TBstrains**. On default, the OPTION is set to 75. The OPTION sets a minimum frequency for the majority allele. Only majority alleles with this frequency or higher are indicated as unambiguous.

--unambig

This OPTION is used in **TBamend**. On default, the OPTION is set to 95. The option sets a minimum percentage of samples that need to show the called variant as unambiguous. If less than this percentage of samples have an unambiguous variant call at that position, the position will not be reported in the amended joint variant table.

--window

This OPTION is used in **TBamend**. On default, the OPTION is set to 12. The OPTION sets a window size in which the algorithm scans for multiple SNPs. If more than one SNP occurs within this window, SNP positions will not be reported in the amended joint variant table.

--distance

This OPTION is used in **TBgroups**. On default, the OPTION is set to 12. The OPTION sets a SNP distance that is used to classify samples into groups of samples, using a single linkage approach. If SNP distances between samples are less or equal this VALUE, they are grouped together. If not, a

new group is formed. We recommend to execute **TBgroups** twice, with **--distance 5** and **--distance 12**.

--quiet

This OPTION turns off the display logging function and will report the logging only in a file, called "TBseq_[DATE]_[USER].log".

--threads

This OPTION is used in **TBbwa**, **TBmerge**, **TBrefine**, **TBpile** and **TBlist**. On default, the OPTION is set to 1. The OPTION sets the maximum number of CPUs to use within the pipeline. You can use more than one core in order to execute the pipeline faster. 8 is the current maximum.

--help

This OPTION will show you all available OPTION and VALUE of TBseq on the display. Use this OPTION, if you are unsure about the OPTION and VALUES you can use in TBseq.

[EXAMPLES]

TBseq --step TBfull

Will execute the whole pipeline with default values. All log output is written into a file called "TBseq_[DATE]_[USER].log" and on screen. The log file is located where you executed the pipeline.

TBseq --step TBbwa --continue --threads 8

Recommended, if read file naming scheme is correct. The pipeline starts with the read mapping module and continues after finishing this module. The system will use 8 cores whenever possible. All log output is written into a file called "TBseq_[DATE]_[USER].log" and on screen.

TBseq --step TBlist --threads 8

This example uses **TBlist** with 8 threads. All log output is written into a file called "TBseq_[DATE]_[USER].log" and on screen. To execute this module, you need to have a finished **TBpile** output.

TBseq --step TBjoin --sample samples.txt --project TEST

This example uses **TBjoin** with default VALUE setting. All log output is written into a file called "TBseq_[DATE]_[USER].log" and on screen. To execute this module, you need to have a finished **TBlist** and **TBvariants** output.

TBseq --step TBstrains

This example uses **TBstrains** with default VALUE setting. To execute this module, you need to have a finished **TBlist** output. All log output is written into a file called "TBseq_[DATE]_[USER].log" and on screen.

TBseq --step TBvariants --mincovf 10 --mincovr 10 --minfreq 80 --minphred20 10 --outmode 2

This example uses **TBvariants** with a modified VALUE setting for variant calling. **TBvariants** will output only SNP positions. All log output is written into a file called "TBseq_[DATE]_[USER].log" and on screen. To execute this module, you need to have a finished **TBlist** output.

[AUTHORS]

Thomas A. Kohl (core logic)
Robin Koch (package building)
Christian Utpatel (core logic)
Maria R. De Filippo (beta test)
Viola Schleusener (beta test)
Daniela M. Cirillo (head)
Stefan Niemann (Head)

[COPYRIGHT AND LICENSE]

[HOMEPAGE AND SOURCE REPOSITORY]

TBseq on github: https://github.com/TaKohl/TBseq_source/
Research center Borstel: http://www.fz-borstel.de/cms/en/science/start.html

[References]

1. Homolka, S. et al. High resolution discrimination of clinical Mycobacterium tuberculosis complex strains based on single nucleotide polymorphisms. PLoS One 7, e39855 (2012).

2. Coll, F. et al. A robust SNP barcode for typing Mycobacterium tuberculosis complex strains. Nature Communications 5, 4812 (2014).