

© Copyright 2013-2014 by Philipp M. Schlüter 2013-2014. This software is freeware, USE AT YOUR OWN RISK. By downloading/running it you accept the following:

This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY, whether expressed or implied; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. The author (PMS) will not be liable for any special, incidental, consequential, indirect or similar damages due to loss of data or any other reason, even if he or an agent of his has been advised of the possibility of such damages. In no event shall the author be liable for any damages, regardless of the form of the claim. The person using the software bears all risk as to the quality and performance of the software.

Author contact: firstname dot lastname* at systbot.uzh.ch (* ü is spelled ue).

Command line options

Parameter	Values	Meaning
<i>Program information</i>		
-h, -H, -?		Display help (not implemented) and terminate.
-v, -V		Display program version.
-legal		Display copyright notice and legal disclaimer.
-Y		Display parameter settings and terminate.
<i>Global parameters</i>		
-i	String	Input path (directory). Input data files and control file should reside in this path.
-o	String	Output path (directory). Output data files will be placed in this directory.
-c	String	Control file name. The control file lists file names and barcodes to be processed (see below).
-s	String	Summary (tsv) output file name.
-D	(Switch)	Run Demultiplexer
-T	(Switch)	Run Tag Processor
-thd	Integer	Number of threads to use (currently only used in <i>TagProcessor</i> !). The value 0 is to automatically set this parameter to the number of (logical) processors present on the computer.
<i>FASTQ file parameters</i>		
-R1	String	Suffix for Read 1 Fastq files. Default: " R1.fq "
-R2	String	Suffix for Read 2 Fastq files. Default: " R2.fq "
-dm	String	File part delimiter. Default: " "
-qi	0 – 2	Input Fastq quality encoding. 0: <i>Sanger</i> ; 1: <i>Illumina 1.0</i> ; 2: <i>Illumina 1.3</i> . Default: <i>Sanger</i> .
-qo	0 – 2	Output Fastq quality encoding. 0: <i>Sanger</i> ; 1: <i>Illumina 1.0</i> ; 2: <i>Illumina 1.3</i> . Default: <i>Sanger</i> .
<i>Demultiplexer (requires -D)</i>		
-dmx		Do demultiplexing
-do		Do write demultiplexed files
-dsc		Perform sanity check / get file stats for demultiplexed files
-dsm		Requires -s. Make demultiplexing summary (to summary file)
-d2		Also process Read 2
-bcq	Int8	Minimum barcode base quality (for good barcode). Default 0.
-abcq	Float	Minimum average barcode base quality (for good barcode). Default 0.
-r		Try to rescue poor quality barcodes
--rO	Float	Minimum quality odds for match to <i>best</i> target barcode sequence for poor quality barcode that will allow poor barcode to be rescued.
--rOR	Float	Minimum quality odds ratio of <i>best/second_best</i> matches to target barcode sequence that will allow poor barcode to be rescued and assigned to <i>best</i> barcode.

GBS Tag Processor (requires -T)		
-to	String	Tag count file name (in output directory)
--tsn		If enabled, the sequence of the tag will be used as the tag name (instead of a tag number T###) in the tag count file.
-dts		Requires -s. Make tag summary (to summary file)
-wt		Write out a file with processed, clean tag sequences.
--ts	String	Required if -wt is present. Suffix appended to clean tag sequence output file. (E.g. "_mytags.fas".)
--tfq		Requires -wt. If enabled, the clean tag output file will be written in FASTQ format, otherwise in FASTA format.
-re	String(s)	Remnant [nondegenerate!] restriction enzyme sequence(s) expected at the start of a tag. Separate multiple by semicolon (" ; "), no space!
-req		Minimum base quality for each base in the remnant start RE site
-cre	String(s)	Nondegenerate Restriction enzyme sequence(s) after which to truncate tag. Separate multiple by semicolon (" ; "), no space!
-cad	String	Adapter/primer sequence at which to truncate the tag
--Kad	Integer	Requires -cad. If >0, length K of Kmers into which end adapter/primer sequence will be decomposed. If 0, an exact string search will be used.
--adO	Float	Requires -cad and -Kad > 0. Minimum quality odds to accept Kmer match to end adapter/primer sequence.
-x	Integer	Number of first X bases to be used in base quality checks
--xN		Filter out sequences with N in first X bases. Requires -x.
--xq	Float	Minimum quality score in first X bases. Requires -x.
-t	Integer	Truncate to (maximum length)
-l	Integer	Minimum sequence length
-ad1, -ad2	String	Adapter 1/2 sequence for <i>finding adapter dimers</i>
--K1, --K2	Integer	Requires -ad1/-ad2. Adapter 1/2 K = Kmer length
--O1, --O2	Float	Requires -ad1/-ad2. Adapter 1/2 minimum quality odds for match.
--m1, --m2	Integer	Requires -ad1/-ad2. Adapter 1/2 minimum number of good Kmer matches. Minimum value: 1.
-c12a		Requires -ad1 <u>and</u> -ad2. a <i>Finding adapter dimers</i> : Combine Adapter1/2 with AND.
-d12	Integer	Requires -c12a. <i>Finding adapter dimers</i> : Maximum distance between adapter 1/2 positions to identify adapter1/2 dimer

Control file structure

The control is a tab-delimited (ANSI) text file that has one header row and then lists input file names, barcode, and sample name.

FileR1	FileR2	Barcode	Sample
Seq01-R1.fastq	Seq02-R2.fastq	CGAT	Indiv_300a
Seq01-R1.fastq	Seq02-R2.fastq	TACAT	Indiv_301b
Seq01-R1.fastq	Seq02-R2.fastq	GTATT	302c
SeqB2.fq	SeqB2_R2.fq	GTATT	indiv4
SeqB2.fq	SeqB2_R2.fq	TAGCATGC	Ind_5_Pop3

Column FileR2 can be empty if there is no Read2 file to be processed.

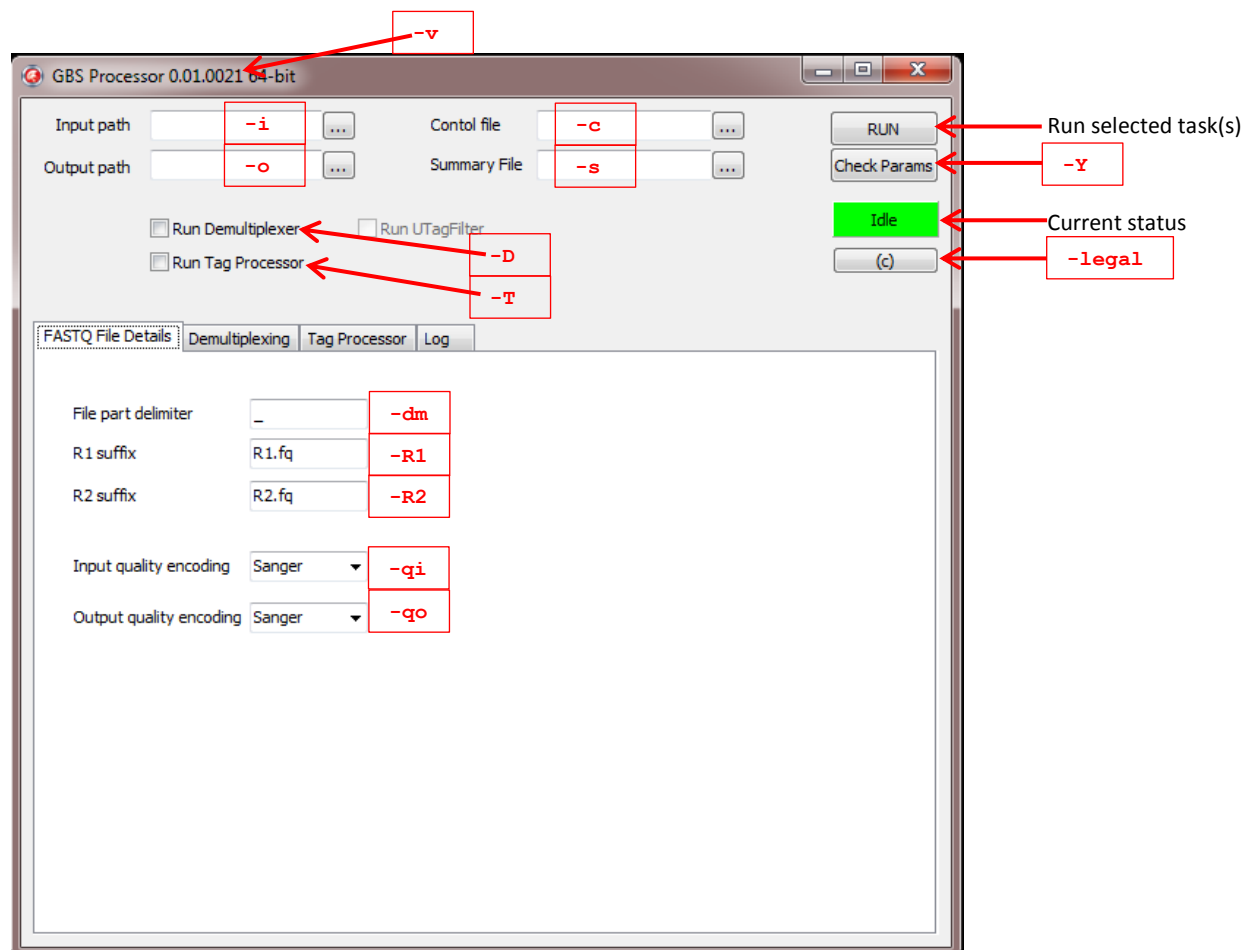
Known problems and limitations

Version 0.001.0021:

- Use 64-bit versions wherever possible. Although a 32-bit version is provided as a back-up, the 32-bit memory limitation means that even for moderate data sets, it may fail because it runs out of memory.
- Do not run tag processor in single-threaded mode, this option is disabled (because not buggy and not properly tested).
- Current implementation of the *tag processor* functionality only processes Read1 data. To process Read2 data, specify settings for these separately and proceed as though they were Read1.
- Other problems or questions? - please let me know (see e-mail address above).

Graphical user interface

Corresponding command-line options are indicated



Parameters for -D

FASTQ File Details | Demultiplexing | Tag Processor | Log

☒ Do demultiplexing **-dmx**

☒ Write demultiplexed output files **-do**

☐ Sanity check demultiplexed files **-dsc**

☒ Summarise demultiplexing results **-dsm**

☒ Also process Read 2 **-d2**

Barcode parameters

Min Barcode Base Quality 2 **-bcq**

Min Average Barcode Base Quality 25.00 **-abcq**

☒ Try to rescue poor barcodes **-r**

Min Odds for Poor Barcode Assignment 4.50 **--rO**

Min Odds Ratio for Poor Barcode Assignment 5.00 **--rOR**

Parameters for -T

FASTQ File Details | Demultiplexing | Tag Processor | Log

Remnant RE site(s) @ tag start* CTGC **-re**

Min base quality for start RE 20 **-req**

Truncate tag at RE(s)* GCAGC;GCTGC **-cre**

* nondegenerate, separate multiply by ";"

Truncate tag at sequence CAGAGATC **-cad**

K for tag end 6 **--Kad**

min Odds EndFrag 15.00 **--adO**

Output Tag Count File **-to**

☐ Write seq tags as names **--tsn**

☐ Write out clean tags **-wt**

Clean tag file suffix: _tags.fast **--ts**

☐ Write clean tags as FASTQ **--tfq**

☒ Summarise TagProcessor results **-dts**

☒ Multithreaded **-thd** ← If disabled, forces **-thd 1** (single-threaded mode)

Threads 0 (0 = auto)

Note: The current implementation ONLY processes Read1 data!

Quality check first X bases 72 **-x**

☒ Check X bases for N **--xN**

min X quality score 10 **--xq**

Truncate to max length 64 **-t**

min tag length 0 **-l**

Dimer Adapter 1

Adapter Seq GCTCTCCGATCT **-ad1**

K 6 **--K1**

min matches 1 **--m1**

min Odds 10.00 **--O1**

Dimer Adapter 2

Adapter Seq AGATCGGAAGAGC **-ad2**

K 6 **--K2**

min matches 1 **--m2**

min Odds 10.00 **--O2**

☒ Combine adapters with AND **-c12a**

Max dist Ad1 - Ad2 25 **-d12**