AGeNT

Version 2.0.5

DISCLAIMER: The Agilent Genomics NextGen Toolkit (AGeNT) software module has been designed to provide the adaptor trimming and duplicate read removal using the Molecular Barcode (MBC) information from HaloPlex^{HS}, SureSelect XT HS, and SureSelect XT HS2 Illumina sequencing runs in a flexible command-line interface for integration into your bioinformatics pipeline. AGeNT is explicitly designed and fine-tuned for customers with established in-house bioinformatics experts with the capability to build, integrate, maintain, and troubleshoot internal analysis pipelines. Moreover, the module is designed specifically for customers with sufficient computing infrastructure and IT support to troubleshoot all issues unrelated to the execution of the AGeNT algorithms. Agilent recommends that users without thorough bioinformatics expertise use either Agilent SureCall or Alissa Align & Call software instead.

For Research Use Only. Not for use in diagnostic procedures.

AGeNT

The **A**gilent **Ge**nomics **T**oolKit (AGeNT) is a collection of command-line tools, developed by Agilent, for NGS data processing. With minimal configuration, you can start using these tools on the Linux or Windows command-line.

Command-line syntax

Getting help with AGeNT

To get help with any command while using AGeNT, simply type help at the end of any command name.

For example, this command displays help for the general AGeNT options and the available top-level commands:

```
$ agent help
```

The following command displays the specific help message for the LocatIt module:

```
$ agent locatit help
```

Command Structure in AGeNT

AGeNT uses a multipart structure on the command line that must be specified in this order:

1. The base call to the AGeNT program.

2. The top-level command, which typically corresponds to a tool supported by AGeNT.
3. The subcommand that specifies which operation to perform.
4. General CLI options or parameters required by the operation.

Structure:

```
 $ agent <command> <subcommand> [subcommand-specific options and
parameters]
```

Input parameters can consist of any arguments supported by the AGeNT modules, such as numbers, strings, lists, maps, and JSON structures. The supported arguments depend upon the specific command and subcommand specified.

Example

```
 $ agent locatit -v2Only -d 1 -m 3 -q 25 \
-l /Users/uname/data/Covered.bed \
-o /Users/uname/data/test_output.bam \
/Users/uname/data/test_input.bam
```

Note: *Paths and filenames should also contain no spaces or other non-permissible characters on a Unix or Windows command-line.*

Trimmer

Prior to alignment, Trimmer processes the read sequences to identify and remove the adaptor sequences and also extracts molecular barcodes (for SureSelect XT HS2).

Note: This jar was compiled using Java version 8. Please make sure your Java Runtime Environment is at least at version 8 by running the command "java -version".

Command-line syntax:
To test that you can run Trimmer, run the following command:

```
java -jar /path/to/trimmer-<version>.jar
```

or, if you have setup an environment variable (such as "$TRIMMER) as a shortcut:

```
java -jar $TRIMMER
```

You should see the Trimmer help text.

Example command-line:

```
java -jar trimmer-<version>.jar [mandatory options] [options] -fq1 <read1_filename> \
-fq2 <read2_filename>
```

Required parameters:

Parameter	Description
`-fq1 <filename>`	Read1 FASTQ file (Multiple files can be provided separated by a comma).
`-fq2 <filename>`	Read2 FASTQ file (Multiple files can be provided separated by a comma).

Note: *Even though -fq1 and -fq2 accept multiple files separated by a comma, the program will output results in a single file for each read.*

At least one of the following available library prep types is also mandatory to set the correct adaptor sequences for trimming.

Mandatory Option	Library Prep Type
`-halo`	HaloPlex
`-hs`	HaloPlexHS
`-xt`	SureSelect XT, XT2, XT HS
`-v2`	SureSelect XT HS2
`-qxt`	SureSelect QXT

Optional Parameters:

Option	Description
`-minFractionRead <n>`	Sets the minimum read length as a fraction of the original read length after trimming. Value range permitted is 0 to 99. Default value is 30.
`-idee_fixe`	Indicates that the fastq files are in the older Illumina fastq format (v1.5 or earlier). In addition to handling the older style read names, this option also assumes that the base qualities are encoded using the Illumina v1.5+ Phred+64 format and will attempt to convert bases to Phred+33.
`-out_loc`	Directory path for output files.

Usage Examples:

```
java -jar trimmer-<version>.jar \
     -fq1 ./ICCG-repl1_S1_L001_R1_001.fastq.gz,./ICCG-repl1_S1_L001_R1_002.fastq.gz \
     -fq2 ./ICCG-repl1_S1_L001_R2_001.fastq.gz,./ICCG-repl1_S1_L001_R2_002.fastq.gz \
     -halo -minFractionRead 50 -idee_fixe \
     -out_loc result/outputFastqs/
```

Tags for SureSelect XT HS2:

For the SureSelect XT HS2 option, trimmed molecular barcodes (MBCs) will be annotated in the readname. These annotation tags are:

- BC:Z:*sample barcode*
- ZA:Z:*3 bases of MBC (first half of dual MBC) followed by 1 or 2 dark base(s)*
- ZB:Z:*3 bases of MBC (second half of dual MBC) followed by 1 or 2 dark base(s)*
- RX:Z:*first half of MBC + second half of MBC concatenated with a "-")*
- QX:Z:*base quality of sequence in RX:Z (concatenated with a space)*

e.g.
@D00266:1113:HTWK5BCX2:1:1102:9976:2206 BC:Z:CTACCGAA+AAGTGTCT ZA:Z:TTAGT ZB:Z:TCCT RX:Z:TTA-TCC QX:Z:DDD DDA

***Note:** The MBC bases are masked as **N** and corresponding base qualities marked as **$** in some annotations if they are not recognized as a valid XT HS2 MBC.*

e.g.
@K00336:80:HW7GLBBXX:7:1115:1184:3688 BC:Z:CTACCGAA+AGACACTT ZA:Z:NNNNN ZB:Z:AAAGT RX:Z:NNN-AAA QX:Z:$$$ <AA

Output for SureSelect XT HS2:

In SureSelect XT HS2 mode (-v2), for every two FASTQ files (read 1 FASTQ file and read 2 FASTQ file) the program outputs three compressed files:

- trimmed read 1 FASTQ file (.fastq.gz)
- trimmed read 2 FASTQ file (.fastq.gz)
- MBC sequence file (.txt.gz).

LocatIt

LocatIt processes the Molecular Barcode (MBC) information of a HaloPlex^{HS}, SureSelect XT HS, or SureSelect XT HS2 Illumina sequencing run. For HaloPlex^{HS} and SureSelect XT HS analyses, LocatIt will tag read pairs in a SAM/BAM file with their MBC sequences and mark or merge MBC duplicates from that SAM/BAM file. For SureSelect XT HS2 analyses using duplex consensus reads, LocatIt requires that the input bam file has already been annotated with the MBC sequences (using AGeNT Trimmer and BWA-MEM with ”-C” parameter, for example).

Note: This jar was compiled using Java version 8. Please make sure your Java Runtime Environment is at least at version 8 by running the command “java –version”.

Command-line syntax

To test that you can run LocatIt, run the following command:

```
java -jar /path/to/locatit-<version>.jar
```

or, if you have setup an environment variable (such as "$LOCATIT) as a shortcut:

```
java -jar $LOCATIT
```

You should see the LocatIt help text.

To run LocatIt for SureSelect XT HS2 using duplex or hybrid mode (assumes that the input file already has the correct MBC tags):

```
java -Xmx12G -jar locatit-<version>.jar [OPTIONS] input_bam_file_name
```

For other applications (including SureSelect XT HS2 in single consensus mode):

```
java -Xmx12G -jar locatit-<version>.jar [OPTIONS] input_bam_file_name MBC_file_1 MBC_file_2 ...
```

Note: *-Xmx12G is just an example, please adjust memory based on sequencing depth and size of the input file.*

Required parameters:

Parameter	Description
input_bam_file_name	name of the input BAM or SAM file
MBC_file_1..N	input file(s) containing the MBC sequences and qualities for the read pairs in the input BAM file. For HaloPlex^{Hs} and SureSelect XT HS, this is the index 2 read. For SureSelect XT HS2, it is the txt file created by AGeNT Trimmer. The file(s) must contain all reads that are in the S/BAM file, but may also contain reads that have been filtered out of the S/BAM during processing. For SureSelect XT HS2, these files are only necessary in single consensus mode.

Options:

Option	Description
-l <covered.bed>	Specify covered bed file downloaded from Agilent's SureDesign. If specified, the properties file histogram reflects what is happening within and not within the covered region. For SureSelect designs, this option is identical to the -b covered.bed option described above.
-b <intervals.bed>	Provide design files downloaded from Agilent's SureDesign website. For Haloplex designs, use the amplicon bed file. For SureSelect designs, this option is identical to the -l.
-o <output_file_name>	Required option to provide name/file path of the file generated by LocatIt.

Option	Description
-v2Duplex	For SureSelect XT HS2, turn on the duplex consensus mode option. The input SAM/BAM file needs to already be annotated with the correct 3+3 MBC SAM tags. First, LocatIt performs single-strand consensus deduplication. When two complementary single-strand consensus sequences are present, they are used to generate a duplex consensus sequence. All single-strand consensus reads are marked as not passing the vendor quality test and should be filtered out prior to running any downstream analysis.
-v2Only	Deprecated, the same as -v2Duplex.
-i	Incremental; instead of having the list of amplicons, i.e. the list of all possible pair starts/stops, the program learns all the possible start/stop combinations as it is reading the data. This is the main option that switches between HaloPlex and SureSelect modes. LocatIt performs single-strand consensus deduplication for both SureSelect XT HS and XT HS2.
-v2Hybrid	For SureSelect XT HS2, turns on the “hybrid” deduplication approach. The approach is the same as for -v2Duplex mode, but the single-strand consensus reads are not flagged as failing the vendor quality test. For compatibility with downstream applications, duplex consensus reads are output twice (once for each input single consensus read) in order to match the stoichiometry of the single consensus reads.
-R	Output bam file that has duplicates(SAM flag 0x400), filtered reads(SAM flag 0x200), secondary(SAM flag 0x100) and supplementary(SAM flag 0x800) reads removed.
-d <number>	Barcode distance. If two populations of read (pairs) only differ by this number of bases, both populations will be processed as having the same barcode. Because several errors can be merged back into the main population, some barcodes that are more than this number of bases away from each other may end up being merged together. Range is 0 to 5. Default value is 0.

Option	Description
-q <number>	Reads having barcodes with quality less than specified threshold will be filtered. Range is 0 to 45. Default value is 25.
-Q <number>	Reads having any base that is lower than specified threshold will be filtered. Range is 0 to 45. Default value is 0.
-m <number>	Parameter specifies minimum number of read pairs associated with a barcode (amplification level). Barcodes having less reads than specified threshold will be filtered. Range is >=1. Default value is 1.
-c <number>	Enable optical duplicate detection and offset between two duplicate clusters on the flow cell in order to consider them optical duplicates. The offset is specified as a radius in pixels. If the offset between two duplicate clusters is less than this threshold, then the clusters are considered to be optical duplicates. The value expected is an integer >= 0. For unpatterned Illumina flow cells, 100 is an appropriate value. For the patterned flow cells, 2500 is more appropriate. By default, optical duplicate detection is disabled.
-U	unsorted BAM/SAM output - Faster and requires less RAM.
-S	sorted BAM/SAM output
-L	If the input SAM/BAM contains fewer reads than the index2 file(s) and they are in the same order, the option -L allows LocatIt to discard non-matching index2 reads as it processes the input file instead of buffering them in case the matching reads would show up later. This saves a lot of memory.
-C	Chimeric; For HaloPlex, this means that the pairs which match two different amplicons are kept. For SureSelect, it should be always on. If -i is specified it sets -C internally.

Option	Description
-r	To remove/mask read1 read2 common overlap, half on each side.
-IB	input file is BAM (default is SAM)
-IS	input file is SAM (default is SAM)
-OB	output file is BAM (default is BAM)
-OS	output file is SAM (default is BAM)
-P	Renames the SAM tags so that downstream pipelines expecting different conventions can be used. *Syntax:* -P[letter]:[new 2 letters tag],[next],... *Example:* -PM:xm,Q:xq,q:nq,r:nr The tags that can be renamed are: • M for barcode sequence • Q for barcode quality • q for barcode consensus quality • r for read consensus quality • N for barcode name tag • a for alt readnames tag • d for readnames tag • 1 for begin amplicon tag • 2 for end amplicon tag
-K	Keep the intermediate bam file. With this option a coordinate-sorted and barcode-annotated bam file will be written to the parent directory of input bam file. By default this is not applied and is only useful for debugging.
-X <temp_directory>	Location of temporary bam files used to store overflow of matches. Temporary files will be deleted at program exit.
-t <temp_directory>	Location of temporary bam files used to store overflow of matches. Temporary files are not deleted.

Usage examples:

- SureSelect XT HS2 in duplex mode

```
java -Xmx12G -jar locatit-<version>.jar \
    -S -v2Duplex -d 1 -m 3 -q 25 -Q 25 \
    -l Covered.bed -o test_output.bam \
    test_input.bam
```

- SureSelect XT HS (and XT HS2 in single consensus mode)

```
java -Xmx12G -jar locatit-<version>.jar \
    -PM:xm,Q:xq,q:nQ,r:nR \
    -q 25 -m 1 -U -IS -OB -C -i -r \
    -l covered.bed -o test_OUTPUT Test_CCP.sam \
    CCP_1_AACGTGAT_L001_R2_001.fastq.gz \
    CCP_1_AACGTGAT_L001_R2_002.fastq.gz
```

Note: For SureSelect XT HS2 the MBC sequence files are produced by AGeNT Trimmer and named …RN…txt.gz instead of …R2…fastq.gz

- SureSelect XT HS2 in hybrid mode

```
java -Xmx12G -jar locatit-<version>.jar \
    -S -v2Hybrid -d 1 -m 3 -q 25 -Q 25 \
    -l Covered.bed -o test_output.bam \
    test_input.bam
```

- Halo

```
java -Xmx12G -jar locatit-<version>.jar \
    -q 25 -m 1 -U -IS -OB \
    -b ISCA2bf_extraG_001001398178390_AllTracks_amplicons.bed \
    -o test_OUTPUT \
    Test_ICCG_Panel.sam \
    ICCG-repl1_S1_L001_I1_001.fastq.gz
```

Tags For SureSelect XT HS2:

LocatIt might insert some tags into SAM records and here are their meanings:

- fi:Z:*justifications for being filtered out/flagged as not passing vendor quality test*
  e.g. fi:Z:BAD_READ2_QUALITY;BELOW_MBC_NREADS_LIMIT;

- XI:i:*number of duplicates collapsed for the single-strand consensus read*
  e.g. XI:i:2
- XJ:i:*number of duplicates collapsed for the single-strand consensus read on the complementary strand. (XT HS2 duplex consensus only.)*
  e.g. XJ:i:1