# 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 on 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.*

# CReaK

CReaK (**C**onsensus **Rea**d **K**it) identifies PCR duplicates in Illumina sequencing data from SureSelect XT HS or SureSelect XT HS2 Illumina sequencing data. It uses both alignment position and molecular barcode (MBC) information to group reads into MBC families. For each MBC family, CReaK constructs an error-corrected, deduplicated consensus read pair. CReaK requires that the reads in the input BAM file have already been annotated with their corresponding MBC sequences and MBC base qualities (using AGeNT Trimmer and BWA-MEM with "-C" parameter, for example).

***Note:*** *This jar was compiled using Java version 11. Please make sure your Java Runtime Environment is at least at version 11 by running the command* `java -version`**.**

## Command-line syntax

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

```
java -jar /path/to/creak-<version>.jar -h
```

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

```
java -jar $CREAK -h
```

You should see the CReaK help text which includes its usage and options.

To run CReaK:

```
java -Xmx8G -jar creak-<version>.jar -f -F [OPTIONS] input_bam_file_name
```

***Note:*** *-Xmx8G is just an example, please adjust memory based on sequencing depth and size of the input file.*

```
Usage: CReaK [-ehrvg] [-b=<bedFile>] -c=<cMode> [-d=<mbcMismatch>] -o=<outBam>
            [-s=<cacheSize>] (-f [-fi] [-mm=<minAvgMBCQual>]
            [-mr=<minAvgReadQual>] [-mq=<minMAPQ>]) (-F
            [-MS=<minMulti4Sinlge>] [-MD=<minMulti4Duplex>]) FILE
```

Minimal input required:

- Input BAM file
- Output BAM file name/path
- Consensus calling mode (see *Consensus Modes* table below)
- Filtering options
    - `-f`**:** Enable input read flagging/filtering. This parameter must be present.
    - `-F`**:** Enable consensus read filtering. This parameter must be present.

## Required parameters:

| Parameter | Description |
|---|---|
| FILE | name of the input BAM or SAM file |
| -o, --output-bam-file=<outBam> | Output BAM file name/path. Avoid spaces when setting the path/file name. |
| -c, --consensus-mode=<cMode> | Consensus calling mode: SINGLE, HYBRID, DUPLEX. For more details see below for table of *Consensus Modes*. |

## Input read filter parameters:

Filtered input reads are always removed from the output file regardless of `-r` parameter.

| Parameter | Sub-option | Description |
|---|---|---|
| -f, --input-read-filtering | | Enable input read filtering. With no additional filtering options specified only unmapped (SAM flag 0x4), secondary (SAM flag 0x100), and supplementary (SAM flag 0x800) reads will be filtered. Other optional filters can be specified using -fi, -mm, -mr and -mq. |
| | -fi, --interval-filter | Enable this filter to remove reads that are not covered by intervals in the optionally provided bed file. **In the case of input BAM with many chimeric alignments, this filter may cause loss of read pairs before consensus calling.** |
| | -mm <number>, --min-avg-MBC-qual=<minAvgMBCQual> | Sets the minimum average MBC base quality. Filter reads with lower average MBC base quality.<br>Range is [0, 40], default is 0. |
| | -mr <number>, --min-avg-read-qual=<minAvgReadQual> | Set the minimum average read base quality. Filter reads with lower average read base quality.<br>Range is [0, 40], default is 0. |
| | -mq <number>, --min-MAPQ=<minMAPQ> | Sets the minimum read mapping quality (MAPQ). Filter reads with lower MAPQ.<br>Range is [0, 255], default is 0. |

## Consensus read filter parameters

Filtered consensus reads are flagged with SAM flag 0x200.

| Parameter | Sub-option | Description |
|---|---|---|
| -F, --consensus-read-filtering | | Enable consensus read filtering. Filtered reads will be flagged with SAM flag 0x200. -MS and -MD are applied either using default values or with values specified by user with option -MS and -MD. |
| | -MS <number>, --min-multiplicity-in-single=<minMulti4Single> | Minimum number of read pairs associated with an MBC/single consensus read pair (amplification level). Single consensus read pairs generated from fewer read pairs than the specified threshold will be flagged with SAM flag 0x200. In duplex mode (-c DUPLEX), in which duplex consensus read pairs are formed from two single consensus read pairs, this threshold applies to whichever single consensus read pair has the smaller value.<br>Range is >= 1, default is 1. |
| | -MD <number>, –min-multiplicity-in-duplex=<minMulti4Duplex> | Minimum number of read pairs associated with duplex MBC/duplex consensus read pairs (total number of read pairs associated with the two single consensus read pairs that form the duplex consensus read pair). Duplex consensus read pairs generated from fewer read pairs than the specified threshold will be flagged with SAM flag 0x200.<br>Range is >= 2, default is 2. |

## Additional Options:

| Option | Description |
|---|---|
| -v, --version | Displays version info |
| -h, --help | Displays help message |
| -e, --memory-efficient-mode | Enables memory-efficient mode. Uses less memory at the cost of computational time. |
| -r, --remove-dup-mode | Removes duplicates (SAM flag 0x400) and filtered consensus reads (SAM flag 0x200) from the output bam file. |
| -g, --keep-singleton | Keep singleton reads (that have unmapped mate) in the output bam file. |
| -d <number>, --MBC-mismatch=<mbcMismatch> | Sets the maximum number of MBC sequence mismatches allowed for the corresponding reads to be considered part of the same MBC family. Range is [0, 2], default is 0. |
| -b <bed_file>, --bed-file=<bedFile> | Sets optional file used to define the covered regions for metrics calculations. If not provided, all reads will be treated as not in a covered region. Required if option -fi filtering option is applied. |
| -s <number>, --cache-size=<cacheSize> | Sets the pairing cache size. The default value should cover most cases but may be increased if the output .stat file reveals an unreasonably large gap between # sam records passed input read filtering and # correctly-paired read pairs for MBC Consensus calling.<br>Range is 1000-1000000, default is 100000. |

## Consensus Mode:

CReaK identifies consensus reads in three different modes: SINGLE, HYBRID, and DUPLEX. All three modes can be applied to SureSelect XT HS2 but only SINGLE can be applied to SureSelect XT HS (since XT HS data does not have a dual MBC).

| Mode | Description |
|---|---|
| SINGLE | One read pair is generated from a group of read pairs that share the same mapped start and end coordinate as well as the same MBC sequence. If -d is not 0, a representative read pair is further chosen from read pairs of different MBC groups (-d allows merging of MBCs with mismatches). |
| DUPLEX | After single consensus calling, a duplex MBC/duplex consensus read pair is generated when two complementary single consensus MBCs/consensus read pairs are present (one from each strand). All single consensus read pairs that do not have its complementary partner are flagged as *not passing platform/vendor quality controls* (SAM flag 0x200). |
| HYBRID | Follows the same approach as DUPLEX mode, but the single consensus read pairs are not flagged as *not passing platform/vendor quality controls* (SAM flag 0x200). Also, for compatibility with downstream applications, duplex consensus read pairs are output twice (once for each input single read pairs) in order to match the stoichiometry of the single consensus read pairs. |

## Usage examples:

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

```
java -Xmx8G -jar creak-<version>.jar \
     -c SINGLE -d 0 -f -mm 25 -mr 30 -F -MS 1 \
     -b Covered.bed -o test_output.bam \
     test_input.bam
```

- SureSelect XT HS2 in duplex mode

```
java -Xmx8G -jar creak-<version>.jar \
     -c DUPLEX -d 0 -f -mm 25 -mr 30 -F -MS 1 -MD 2 \
     -b Covered.bed -o test_output.bam \
     test_input.bam
```

- SureSelect XT HS2 in hybrid mode

```
java -Xmx8G -jar creak-<version>.jar \
     -c HYBRID -d 0 -f -mm 25 -mr 30 -F -MS 1 -MD 2 \
     -b Covered.bed -o test_output.bam \
     test_input.bam
```

- SureSelect XT HS2 in hybrid mode (with less memory)

```
java -Xmx8G -jar creak-<version>.jar \
     -e -c HYBRID -d 0 \
     -f -mm 25 -mr 25 -F -MS 1 -MD 2 \
     -b Covered.bed -o test_output.bam \
     test_input.bam
```

***Note***: *memory efficient mode(-e) works for any mode, “SureSelect XT HS2 in hybrid mode” is just an example.*

- SureSelect XT HS2 in hybrid mode (with duplicates and filtered reads removed)

```
java -Xmx12G -jar creak-<version>.jar \
     -r -c HYBRID -d 0 \
     -f -mm 25 -mr 25 -F -MS 1 -MD 2 \
     -b Covered.bed -o test_output.bam \
     test_input.bam
```

***Note***: *removal mode (-r) works for any mode, “SureSelect XT HS2 in hybrid mode” is just an example.*

## Relevant SAM Tags:

CReaK requires tag values in the SAM record of the input BAM file that contain the molecular barcode sequence and quality:

| Tag | Type | Description | Example |
|---|---|---|---|
| RX | String(Z) | Sequence bases of the unique molecular barcode | RX:Z:CGT-CCG |
| QX | String(Z) | Quality score of the unique molecular barcode in the RX tag | QX:Z:DDD BDB |

CReaK may insert some tags into the output SAM records to provide additional information about the deduplication process:

| Tag | Type | Description |
|---|---|---|
| xc | Integer(i) | Indicates whether this read is covered by intervals in the bed file. A read with mapped bases overlapping with the -b BED file has tag value set to 1, otherwise the value is set to 0. The read pair with two reads mapped to different reference names is always set to 0.<br>e.g. xc:i:0 means this read does not intersect with the BED file. |
| xm | Integer(i) | Indicates the number of read pairs associated with an MBC/single consensus read pair.<br>e.g. xm:i:5 means this MBC has 5 read pairs associated with it (including this single consensus read pair itself). |
| xd | Integer(i) | Indicates the number of read pairs associated with a duplex MBC/duplex consensus read pair (or the two single MBCs that form this duplex MBC). This tag is only present for duplex consensus reads.<br>e.g. xd:i:8 means this duplex MBC has 8 read pairs associated with it (including this duplex consensus read pair itself). If the same read has xm:i:5, that means that one of the single MBCs that forms the duplex MBC has 5 read pairs associated with it, and the other single MBC has 8 - 5 = 3 read pairs associated with it. |
| zd | String(Z) | Contains the read names of duplicates that are associated with this single/duplex consensus read. The number of read names is capped at 50/100 for single/duplex consensus read. Read names are comma-separated.<br>e.g. zd:Z:D00266:1113:HTWK5BCX2:1:1115:2885:70626 means one read pair with the name being D00266…70626 is flagged as a duplicate of this consensus read pair. |
| zp | String(Z) | Contains original information from the single consensus read that shares the same name as the duplex consensus read before it was merged. This tag is only for duplex consensus reads. One duplex read is created by merging two single consensus reads. The read name, sequence, quality, CIGAR and MD are preserved in this tag, separated by a vertical bar \|.<br>e.g.zp:Z:D00266:1113:HTWK5BCX2:1:1211:8833:23978\|GACGCTCTTCCGATCTCCGT\|0/<GHG??DHFHGCHHIIHH\|3S17M\|17 contains original read name (same as the duplex read), sequence, quality, CIGAR, and MD of a single consensus read before it is merged into this duplex consensus read. |

| Tag | Type | Description |
|---|---|---|
| zn | String(Z) | Contains original information from the single consensus read that does not share the same name as the duplex consensus read before it was merged. This tag is only for duplex consensus reads. One duplex read is created by merging two single consensus reads. The read name, sequence, quality, CIGAR and MD are preserved in this tag, separated by vertical bar \|. e.g.zn:Z:D00266:1113:HTWK5BCX2:1:1214:18553:39660\|GACGCTCTTCCGATCTCCGT\|0/<GHG??DHFHGCHHIIHH\|3S17M\|17 contains original read name, sequence, quality, CIGAR, and MD of a single consensus read before it is merged into this duplex consensus read. |

## Statistics in .stats file:

CReaK generates a .stats file along with the output .bam file. The .stats file contains duplicate and filtering statistics which are categorized into single number statistics and histograms. Please see below for detailed descriptions of all metrics.

### Single number statistics

| Item name | Description |
|---|---|
| # processed sam records: | the total number of SAM records that CReaK processed |
| # sam records passed input read filtering: | the total number of SAM records that pass the filtering caused by the application of -f, -fi, -mm, -mr and -mq. |
| # correctly-paired read pairs for MBC Consensus calling: | after input read filtering, the total number of SAM records that are properly paired with each other. |
| # read pairs already marked as duplicate and not used for MBC Consensus calling: | among the correctly paired SAM records, the total number of read pairs that are already flagged as duplicate in the input bam file and are thus ignored in consensus read calling. |
| # read pairs that are chimeric (on diff ref names): | among the correctly paired SAM records, the number of read pairs that have reads mapped to different chromosomes/reference names. This does not include chimeric alignments that are mapped very far away from each other on the same chromosome/reference name. |
| # read pairs called as single consensus: | the number of read pairs that are called as single consensus read pairs. In SINGLE consensus mode, reports the total number of all consensus read pairs. In DUPLEX or HYBRID consensus mode, reports the number of single consensus read pairs that cannot be merged into duplex consensus read pairs. |
| # read pairs called as duplex consensus: | the total number of read pairs that are called as duplex consensus read pairs. In SINGLE consensus mode, this number should be 0. |

| Item name | Description |
|---|---|
| # read pairs called as chimeric (on diff ref names) consensus: | the total number of consensus read pairs that are based on chimeric alignments, specifically those that are mapped to different chromosomes/reference names. This metric applies to both single consensus or duplex consensus modes, and is a subset of # read pairs get called as single consensus or # read pairs get called as duplex consensus. |
| # read pairs marked as dups during consensus calling: | the total number of read pairs that are flagged as duplicate (SAM flag 0x400). |
| # read pairs that failed consensus filter: | the total number of read pairs that are flagged as not passing platform/vendor quality control(SAM flag 0x200) due to the application of the -MS and -MD parameters. |
| # read pairs called as single consensus and failed to form duplex consensus: | the total number of read pairs that are called as single consensus but are unable to form duplex consensus. In SINGLE consensus mode, this number should be 0. In DUPLEX or HYBRID consensus mode, this number should be equal to # read pairs get called as single consensus. |

### Histograms

The histogram is represented by a series of numbers on the *X axis* and their counterparts on the *Y axis* . These numbers are comma-separated.

| Histogram | X axis | Y axis | Description |
|---|---|---|---|
| SINGLE CONSENSUS HISTOGRAM (uncovered) | the number of read pairs associated with an MBC family, or the amplification level of a single MBC in the uncovered regions (defined by the user-provided bed file) | the number of MBCs at this amplification level | Shows the distribution of MBCs at different amplification levels in the uncovered regions.<br>e.g., x_axis=1,2 and y_axis=4,5 means that, in the uncovered regions, there are 4 MBCs having only 1 read pair associated with them, and 5 MBCs having 2 read pairs associated with them |
| SINGLE CONSENSUS HISTOGRAM (covered) | same as above but for the covered regions | same as above | same as above but for the covered regions |

| Histogram | X axis | Y axis | Description |
|---|---|---|---|
| DUPLEX_CONSENSUS HISTOGRAM 1 (uncovered) | the minimum number of read pairs associated with the two MBCs that form the duplex MBC, or the minimum amplification level of this duplex MBC family in the uncovered regions (defined by the user-provided bed file) | the number of duplex MBCs at this amplification level | Shows the distribution of duplex MBCs at different minimum amplification levels in the uncovered regions.<br>e.g., x_axis=1,2 and y_axis=4,5 means that, in the uncovered regions, there are 4 duplex MBCs having at least 1 read pair associated with one of its two single MBCs, and 5 duplex MBCs having at least 2 read pairs associated with one of its two single MBCs. |
| DUPLEX CONSENSUS HISTOGRAM 1 (covered) | same as above but for the covered regions | same as above | same as above but for the covered regions |
| DUPLEX_CONSENSUS HISTOGRAM 2 (uncovered) | the total number of read pairs associated with the two MBCs that form the duplex MBC family, or the maximum amplification level of this duplex MBC in the uncovered regions (defined by the user-provided bed file) | the number of duplex MBCs at this amplification level | Shows the distribution of duplex MBCs at different maximum amplification levels in the uncovered regions.<br>e.g., x_axis=3,4 and y_axis=4,5 means that, in the uncovered regions, there are 4 duplex MBCs having at least 3 read pairs in total associated with its two single MBCs, and 5 duplex MBCs having 4 read pairs associated with its two single MBCs. |
| DUPLEX CONSENSUS HISTOGRAM 2 (covered) | same as above but for the covered regions | same as above | same as above but for the covered regions |

***Note:*** *The histograms are based on read pairs instead of reads, and in cases where a read pair has one read in a covered region and the other read in an uncovered region, the first read of this pair (by SAM flag 0x40) decides where it belongs.*

# LocatIt

LocatIt processes the Molecular Barcode (MBC) information of a HaloPlexHS, SureSelect XT HS, or SureSelect XT HS2 Illumina sequencing run. For HaloPlexHS 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).

We recommend using the combination of AGeNT Trimmer + AGeNT CReaK tools for pre-processing of the raw reads to remove sequencing adaptors and process the molecular barcode sequences and for PCR de-duplication leveraging of the molecular barcodes. CReaK is a new deduplication tool available in AGeNT 3.0 that replaces the previously available LocatIt tool. LocatIt has been deprecated, but remains available for backward compatibility. Please see the FAQ for further information comparing CReaK to LocatIt.

*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. |
| -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 |

| Option | Description |
|---|---|
| | 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.<br>Range is 0 to 5. Default value is 0. |
| -q <number> | Reads having barcodes with quality less than specified threshold will be filtered.<br>Range is 0 to 45. Default value is 25. |
| -Q <number> | Reads having any base that is lower than specified threshold will be filtered.<br>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.<br>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.<br>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. |
| -r | To remove/mask read1 read2 common overlap, half on each side. |

| Option | Description |
|---|---|
| -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.<br>*Syntax:* -P[letter]:[new 2 letters tag],[next],…<br>*Example:* -PM:xm,Q:xq,q:nq,r:nr<br>The tags that can be renamed are:<br>• M for barcode sequence<br>• Q for barcode quality<br>• q for barcode consensus quality<br>• r for read consensus quality<br>• N for barcode name tag<br>• a for alt readnames tag<br>• d for readnames tag<br>• 1 for begin amplicon tag<br>• 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`

# Trimmer

AGeNT Trimmer removes adaptor sequences from Illumina sequencing reads generated using SureSelect and Haloplex library preparation kits. For SureSelect XT HS and XT HS2, Trimmer also processes the Molecular Barcode (MBC) and adds the MBC information to the read name in the output fastq files. Downstream tools, such as AGeNT CReaK make use of these MBC tags in identifying PCR duplicates using molecular barcodes.

*Note: This jar was compiled using Java version 11. Please make sure your Java Runtime Environment is at least at version 11 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 |
|---|---|
| -fq3 <filename> | This option is only relevant for SureSelect XT HS. MBCs FASTQ file (Multiple files can be provided separated by a comma). |
| -bam | Turn on to output unaligned bam file instead of fastq files |
| -out | Alternative output file name (file path + file name prefix) |
| -polyG <n> | The minimum length of polyG to trim from 3' end regardless of base quality (for nextSeq and NovaSeq polyG problem). Value range permitted is >= 1. |
| -qualityTrimming <n> | Quality threshold for trimming. Value range permitted is 0 to 50. Default value is 5. |
| -minFractionRead <n> | Sets the minimum read length as a fraction of the original read length after trimming.<br>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. |
| -qual_offset <n> | Overwrite auto-detection to indicate FASTQ quality encoding (1 for Phred+33, 2 for Phred+64, 3 for Solexa+64) |
| -out_loc | Directory path for output files. |
| -minMateOverlap <n> | Minimum mate overlap. Default is 50. Range is >0. |

### Usage Examples:

- SureSelect XT HS2 example

```
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 \
     -v2  \
     -out myOutputDirPath/myOutputFilePrefix
```

- SureSelect XT HS example (with MBC tagging)

```
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 \
     -fq3 ./ICCG-repl1_S1_L001_I2_001.fastq.gz,./ICCG-repl1_S1_L001_I2_002.fastq.gz \
     -xt  \
     -out myOutputDirPath/myOutputFilePrefix
```

- SureSelect XT HS example (without MBC tagging)

```
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 \
     -xt  \
     -out myOutputDirPath/myOutputFilePrefix
```

- Halo example

```
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  \
     -out_loc result/outputFastqs/
```

### Tags for SureSelect XT HS and SureSelect XT HS2:

For the SureSelect XT HS and XT HS2 options, trimmed molecular barcodes (MBCs) are formatted as valid SAM tags and added to the read name line. 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 when not in BAM mode:

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