AGeNT FAQ

1. LocatIt generates an Out-of-Memory error

When identifying duplicate reads, LocatIt can use larger and larger amounts of memory as the input files grow in size. Memory errors are especially prevalent during the read tagging stage, when molecular barcodes from the FASTQ file are matched to reads in the BAM. For SureSelect XT HS V1 data, it is recommended that both the input BAM and FASTQ files are sorted as close as possible to the same order (i.e., do not coordinate sort the aligned BAM file) so that LocatIt does not need to store excessive numbers of molecular barcode (MBC) sequences in memory while this tagging occurs. As a result of this tagging limitation, LocatIt has been tested with datasets with file sizes of 10 GB or less (for a single mate pair, zipped). Any data that exceeds this limitation may not complete without additional steps (described below).

If the files are in the same order, the "-L" parameter can be used to discard non-matching MBC reads as it processes the input file instead of buffering them in case a matching read shows up later. To organize the files in the same order, we recommend using *fgbio* to sort the files in read name order. The *fgbio* software is an open-source third-party command line toolkit for working with genomic, and, in particular, next-generation sequencing data. The two commands are:

java -jar fgbio-1.3.0.jar SortFastq -i path_to/barcode_file.fastq.gz -o path_to/sorted_barcode.fastq.gz

java -jar fgbio-1.3.0.jar SortBam -s Queryname -i path_to/bam_file.bam -o path_to/sorted_bam.bam

Additionally, you can specify a directory for the location of temporary intermediate files used to store overflow of matches using either the "-X" or "-t" parameters. Finally, the "-U" parameter can be used to generate an unsorted SAM/BAM file which will be faster and require less memory. Please refer to the LocatIt README for more detailed information on these options.

2. LocatIt complains about out of disk space issues

Please check that the location for temporary data files has read/write permissions and that there is sufficient space to store overflow data during de-duplication. There should be at least as much space in the temporary location as the size of the uncompressed input file. Be aware that LocatIt also write some intermediate files to the directory containing the input files. There must also be sufficient space in this location.

3. AGeNT fails due to input file permissions issues

The input files for Trimmer must be in a directory with read/write permissions. If your input directory is read-only, you can create a symlink to the files in your output directory or any other directory with write permissions.

4. The LocatIt properties output file does not include histogram metrics on the covered region, or the metrics reported are skewed toward the uncovered regions.

Ensure that the provided Covered.bed file obtained from SureDesign contains the correct intervals for the design used in capture. Also, make sure that the genome build for the BED file corresponds to the genome build used for aligning the reads.

5. LocatIt fails to run because of format for covered or amplicon file.

The input file should be in BED6 format. See the definition of the format here: https://genome.ucsc.edu/FAQ/FAQformat#format1. The track header present in the files downloaded from Agilent's SureDesign website should be removed from the file.

6. LocatIt command not recognized or parameters not recognized

Please note that LocatIt is spelled with an uppercase "I" not a lowercase "l". Additionally, make sure that you used the correct parameters (case-sensitive). Specifically, lowercase L (-l) is used to specify the Covered.bed, while uppercase L (-L) is used to stream the reads (discarding non-matching reads along the way). Uppercase I (I) is used for the input format parameters such as -IB and -IS. There is no standalone uppercase I parameter.

7. I am processing HaloPlex data, which version of AGeNT should I use?

For HaloPlex data, both with and without Molecular Barcodes, please use the older AGeNT v1.7 which is available upon request through informatics_support@agilent.com.

8. I am processing XTHS V2 RNA data – what should I do?

Both AGeNT Trimmer and LocatIt work with SureSelect XT HS2 RNA data. However, unlike DNA data where both strands are present and the MBCs in the strands can be matched to form a duplex consensus read, single-stranded RNA libraries only have the MBC present in one orientation and stop at single consensus generation. As a result, you should only use single consensus mode with LocatIt for RNA data. In single consensus mode, it is not necessary to add MBC tags to the aligned file. Simply run AGeNT Trimmer to extract the MBC and trim any adaptors, align your data with your aligner of choice, and run LocatIt with the MBC txt file from Trimmer to perform deduplication (similar to the SureSelect DNA XT HS or XT HS2 single-consensus mode workflows).

9. I was using the "-v2" hybrid option for XTHS V2 deduplication. What happened to this?

This legacy option was removed from the tool. This option was used for performing duplex consensus deduplication followed by single consensus deduplication (on the remaining reads) to reduce the number of discarded reads. A new enhanced hybrid option, "-v2Hybrid" is now available.

10. How does the enhanced "-v2Hybrid" option work? Why is it better than the old "-v2" option?

The "-v2Hybrid" option operates such that each duplex consensus read is written *twice* to the output file to ensure that these reads are weighted properly when compared with those retained after single consensus deduplication. The read names for the 2 duplex consensus reads shall match the read names for the 2 single consensus reads that were used to generate it.

11. Which operating systems are supported?

AGeNT has been tested on RedHat Enterprise Linux 8.0, Windows 10 Enterprise, and macOS Mojave.

12. Which version of Java is supported?

AGeNT requires Java version 8 or later. It has been tested with OpenJDK 11.

13. AGeNT cannot find the specified input files

Ensure that the files specified on the command-line exist.

14. I used AGeNT Trimmer but:

A) It fails to complete; or

B) The aligner fails due to unusual binary symbols

The failure is likely due to mismatching between the read name format and the read quality score format. AGeNT Trimmer includes support for the legacy read name format but requires the user to provide the 'IDEE_FIXE' parameter and the quality scores must also be in the legacy Phred scale.

Legacy format read name: @HWUSI-EAS100R:6:73:941:1973#0/1

New format read name: @EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG

Trimmer with IDEE_FIXE will fail if the quality scores are in the legacy format but the read names are in the newer format. Trimmer with IDEE_FIXE will add strange binary characters to the trimmed file if the read names are in the legacy format and the quality scores are in the newer format.

Newer sequencers (like MGI) can provide the legacy read format but quality scores in the newer Phred scaling. For optimal results with AGeNT Trimmer, we recommend ensuring that you demultiplex this data such that the read name format and quality score format are of the same format.