DISCLAIMER: The Agilent Genomics NextGen Toolkit (AGeNT) module has been designed to provide the adaptor trimming and duplicate read removal capabilities of Agilent SureCall in a flexible command-line interface for integration into your bioinformatics pipeline. Agilent does not guarantee the usability of third party tools (open- or closed-source) in upstream/downstream analysis of data in conjunction with AGeNT. 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. Because Agilent provides limited support of AGeNT, we encourage users without thorough bioinformatics expertise to use Agilent SureCall software instead.

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

LocatIt

LocatIt is a program to process the Molecular Barcode (MBC) information of a HaloPlex^{HS} or SureSelect XT^{HS} Illumina© run. LocatIt will tag read pairs in a bam/sam file with their MBC sequences read out of the input bam file and mark or merge MBC duplicates from that SAM/BAM file.

Commands line Syntax

```
java –Xmx12G -jar LocatIt.jar [-X temp_directory] [-t temp_directory]
[-L][-PM:xm,Q:xq..]
[-D] [-TL] [-U] [-C] [-i] [-N 2000000] [-IB] [-OS] [-r] [-d NN] [-2]
[-q 25][-m 2][-c 2500][-H sam_header_file] -b amplicons.bed -o
output_file_name input_bam_file_name index2_fastq_file_1[.gz]
[...index2_fastq_file_N[.gz]]
```

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”. 12GB RAM is sufficient (recommended) for most 4M read files. Please tailor that parameter to your data size and complexity (large designs will use more memory than small panels).

Options:

-X temp_directory: location of temporary intermediate bam files used to store overflow of matches. Intermediate files will be deleted at program exit.

-t temp_directory: location of temporary intermediate bam files used to store overflow of matches. Intermediate files are not deleted.

-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.

-P: To rename the SAM tags so that the pipeline expecting different conventions can be used.
Syntax: -P[letter]:[new 2 letters tag],[next],…
Example: -PM:xm,Q:xq,q:nq,r:nr,I:ni
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
            I (capital i) for multiplicity
            N for barcode name tag
            a for alt readnames tag
            d for readnames tag
            1 for begin amplicon tag
            2 for end amplicon tag
```

-D: Mark duplicates but output all input records annotated. Default is merge duplicates and output only the consensus read pair per barcode and quality values.

-TL: If sequencer type is IonTorrent.

-U: unsorted BAM/SAM output - Faster and requires less RAM.

-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.

-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.

-N Number: Number of read pairs to process before an intermediate bam file is written and memory is cleared. Increasing the number increases memory needs and decreases computation time. Default value is 2000000.

-IB or -IS: input file is BAM or SAM, default is SAM.

-OB or -OS: output file is BAM or SAM, default is BAM.

-r: To remove r1/r2 overlap; this will mask overlap when mate1 and mate2 if read sequences has overlap with each other.

-d NN: Barcode distance; If two populations of read (pairs) only differ by that many bases, both populations will be processed as having the same barcode. Because several errors can be merged back into the main population, it could happen that barcodes that are more than this number of bases away from each other end up being merged together. Range is 0 to 5. Default value is 0.

-2: To enforce the processing of already de-duplicated file.

-q Number: Reads having barcodes with quality less than specified threshold will be filtered. Range is 0 – 60. 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: Parameter specifies the radius in terms of pixels for circle of optical duplicates merging. If offset between two clusters is less than set threshold then the clusters are considered as optical duplicates. Threshold of 2500 is recommended for patterned flow cells. By default it’s turned off.

-H SAM header file: By default, LocatIt expects hg19 names, chr1-chrM. If the contig names are different (for example, GRCh37 names or nonhuman), one can use this option and provide a SAM header file containing a dictionary of the contigs used by the data files, SAM/BAM and, optionally, the bed file.

-b amplicons.bed: amplicon bed file downloaded from Agilent’s SureDesign website.

-o output_file_name: name of the file generated by LocatIt.

input_bam_file_name: name of the input BAM or SAM file.

index2_fastq_file_1: input fastq file or files containing the barcode sequences for the read pairs in the input bam file. It's OK if some records have been filtered out during processing, i.e. if the fastq files have more records than the BAM file.

-l Covered bed file: Specify covered bed file of SureSelect XT HS design downloaded from SureDesign. If given, the properties file histogram reflects only what is happening within the covered region.

Example:

```
java –Xmx12g -jar LocatIt.jar -PM:xm,Q:xq,q:nQ,r:nR,I:ni –q 25 –m 1 -U -IS -
OB -C -i -r –c 2500 –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
```

```
java –Xmx12G  -jar LocatIt.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
```

SurecallTrimmer

Prior to alignment, SureCall processes the read sequences to trim low quality bases from the ends, remove adaptor sequences, and mask enzyme footprints (for HaloPlex). The parameters for the Trimmer step control the trimming behavior.

Commands line Syntax

```
java -jar SurecallTrimmer.jar -fq1 filename -fq2 filename [-halo] [-hs] [-xt]
[-qxt] [-qualityTrimming] [-minFractionRead ] [-IDEE_FIXE ] [-out_loc]
```

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”.

Options:

Mandatory Parameters:

-fq1 filename: Pair1 FASTQ file (Multiple files can be provided separated by a comma).
-fq2 filename: Pair1 FASTQ file (Multiple files can be provided separated by a comma).

At least one of these switch is mandatory:

-halo: If it’s a HaloPlex sample.

-hs: If it’s a HaloPlex^{HS} sample.

-xt: If it’s SureSelect XT.

-qxt: if it’s SureSelect QXT.

Other optional parameters:

-qualityTrimming: The *quality threshold for trimming* is the quality score threshold that the program uses to determine how many bases to trim off the end of each read.

The program trims any base calls with a quality score below the specified threshold. Value range permitted between 0 and 50. Default value is 5.

-minFractionRead: This parameter determines the minimum read length as a fraction of the original read length after trimming. Default value is 30.

-IDEE_FIXE: If specified, an earlier version of Illumina fastq data (v1.5 or earlier) can be analyzed.

-out_loc: Directory path for output files.

Example:

```
java -jar SurecallTrimmer.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 –qualityTrimming 20 -minFractionRead 50 –
idee_fixe –out_loc result\outputFastqs\
```