Best Practices for Agilent PartekFlow Paired-end RNASeq Pipelines

Purpose

This document outlines basic RNA-Seq pipelines verified for RNA XT HS2 that can be used to analyze data for gene expression and fusion detection.

Agilent Published Pipelines

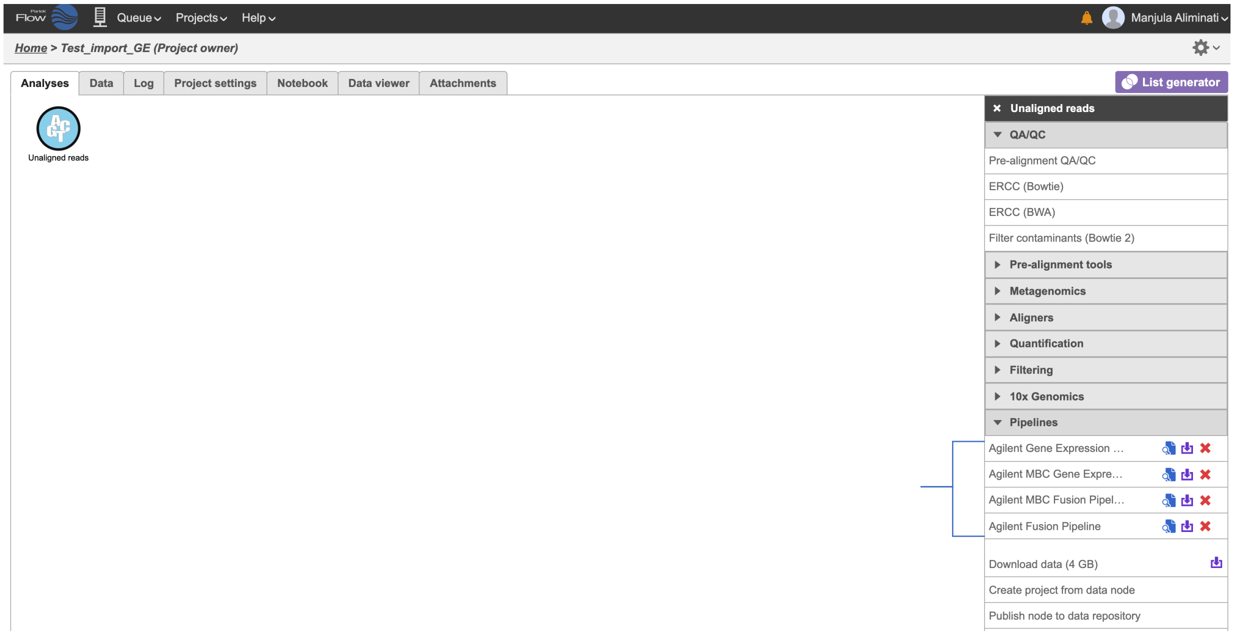
Below are four pipelines for RNA fusion and gene expression that are available for import from PartekFlow public pipelines. To access this pipeline, choose “Settings [from top right corner, under username] > Pipeline Management > Import Pipeline > Hosted Pipelines. Search for below listed pipelines and click import pipeline. You shall see imported pipelines in List generator pane under pipelines.

- Agilent Fusion Pipeline: standard pipeline using third party tools such as STAR and STAR-Fusion for fusion detection.
- Agilent MBC Fusion Pipeline: uses STAR, STAR-Fusion and Agilent custom preprocessing tools to process molecular barcode annotation and remove PCR duplicates in RNA XTHS2 data.
- Agilent Gene Expression Pipeline: standard pipeline using STAR and Partek E/M based quantification model to quantitate gene expression and DESeq2 for differential gene expression
- Agilent MBC Gene Expression Pipeline: uses STAR, Partek E/M based quantification model, DESeq2 for differential gene expression and Agilent custom preprocessing tools to process molecular barcode annotation and remove PCR duplicates in RNA XTHS2 data.

Analyze data using Agilent Published Pipelines

After successfully logging into PartekFlow, follow the steps below to quickly analyze the data using Agilent published pipelines.

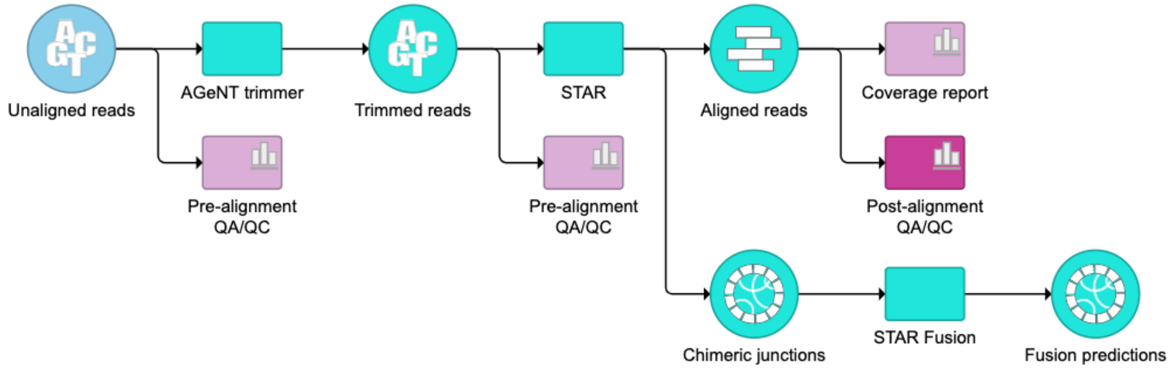
From the Home page, click **New Project** then type a name for the project and select **Import data**. Once the import is successful, you are automatically directed to Analyses tab (see example image below). Select data node to see all imported pipelines in **List generator** on the right panel. Choose the pipeline to apply to the data. During import, you will be asked to select the indices for analysis. By default you are provided with 2 assemblies – Agilent GRCh37 ERCC and Agilent GRCh38 ERCC – that contain all necessary indices and libraries needed for Agilent published pipelines. More information about indices/assemblies is available in the Index Generation section below.



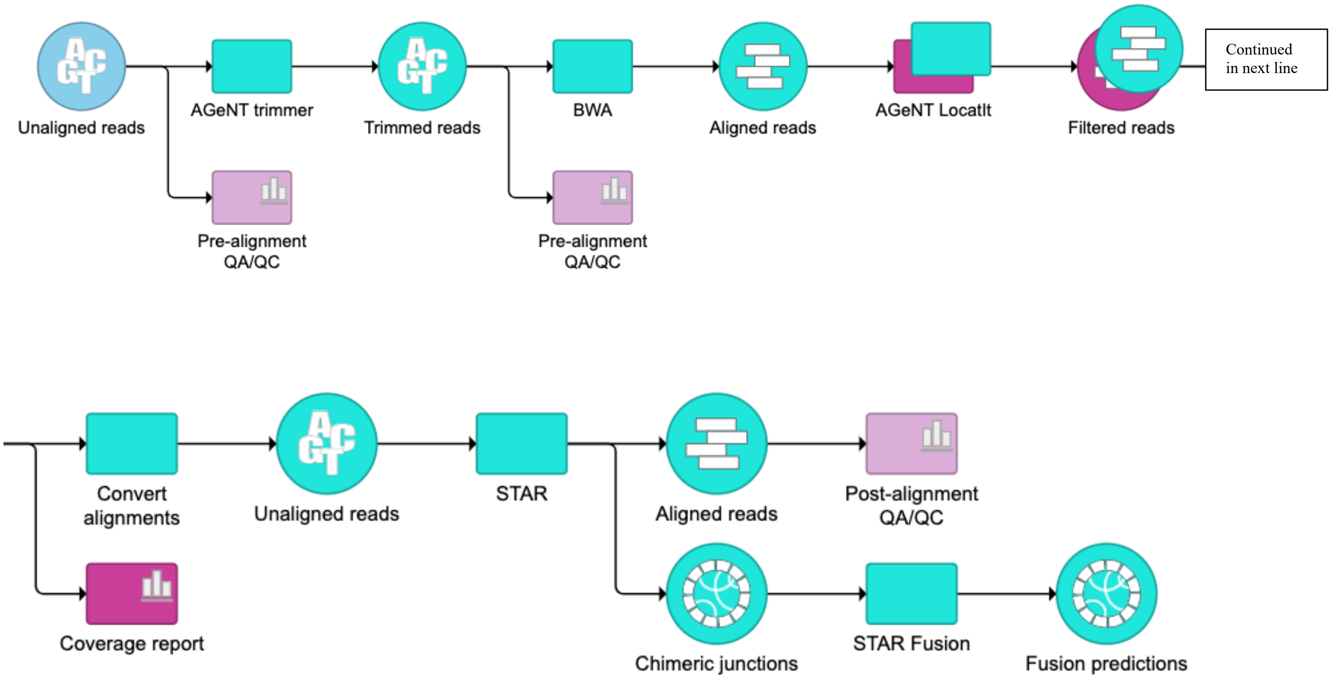
Alternatively, you can choose to build your own pipelines. The section below contains diagrams of the Agilent pipelines for reference.

Workflows for the Agilent Published Pipelines

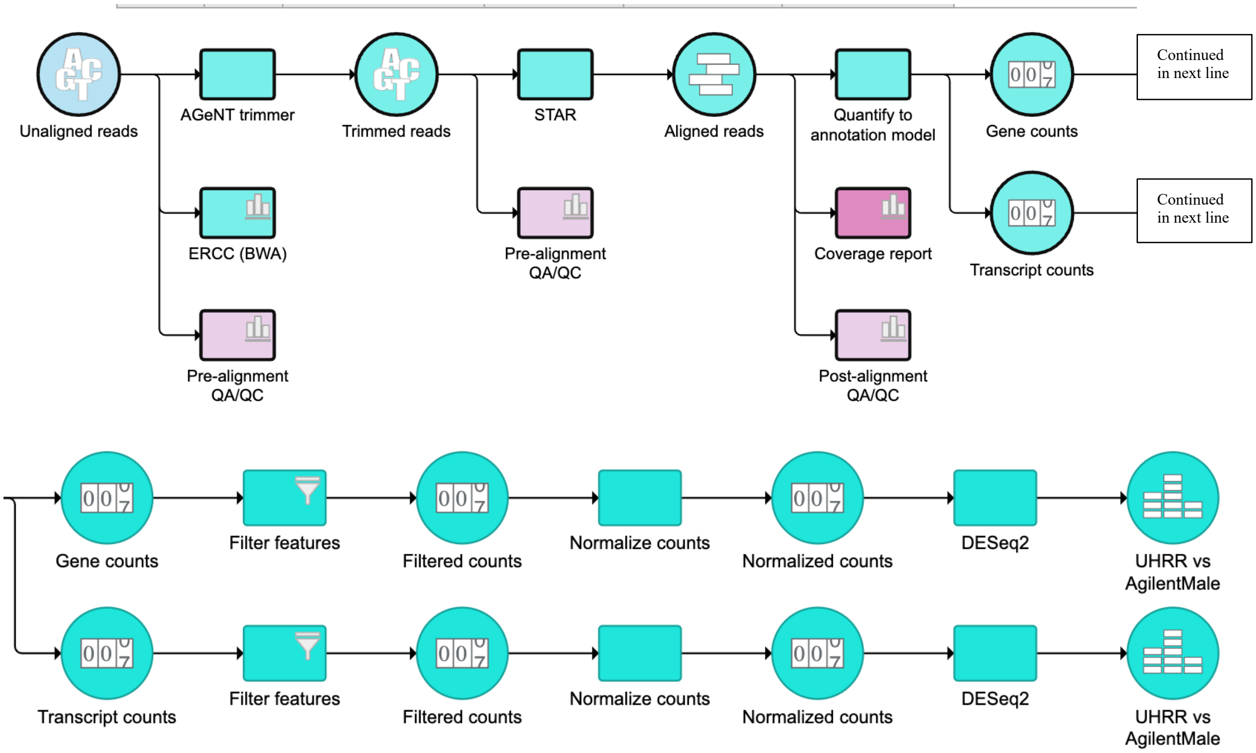
Agilent Fusion Detection Pipeline



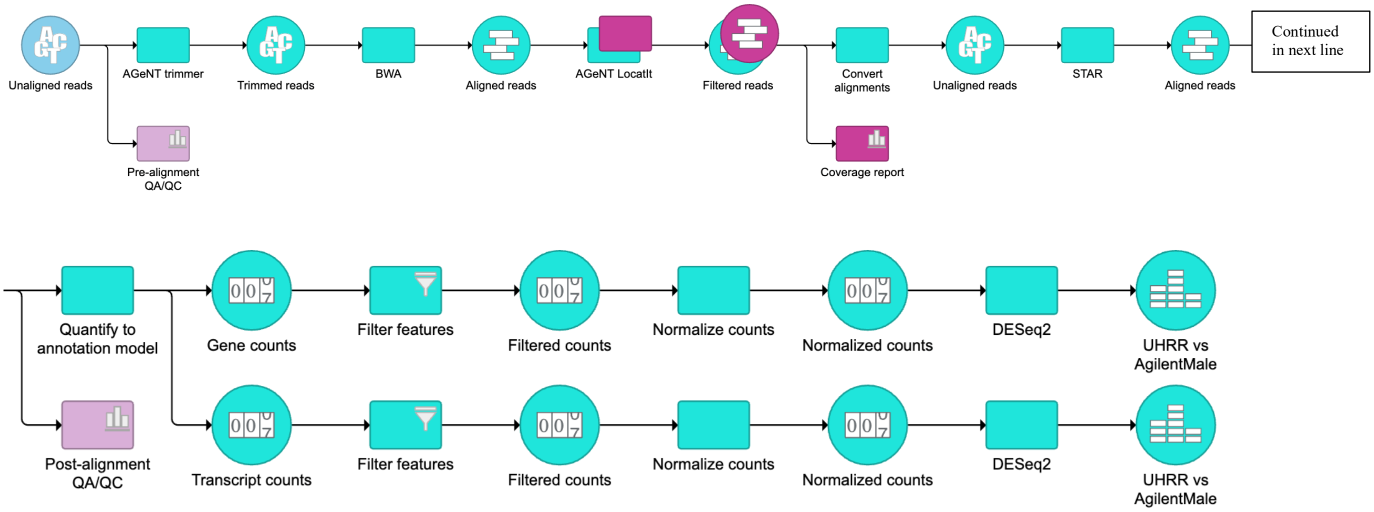
Agilent MBC Fusion Detection Pipeline



Agilent Gene Expression Pipeline



Agilent MBC Gene Expression Pipeline



Importing the pipelines also imports optimal and verified parameter sets[option sets] for each tool.

Index Generation

During import of a pipeline, you are required to select the indices or libraries required for each tool in the pipeline (STAR index version, STAR-Fusion libraries version, BWA index version, etc.). Pre-generated indices and libraries are available for selection in the form of assemblies. Below are the names of pre-generated assemblies

- Agilent GRCh37 ERCC
- Agilent GRCh38 ERCC

Each assembly contains below indices

- Reference genome
- GTF annotations
- STAR reference index
- BWA reference index
- STAR-Fusion library
- STAR annotation index
- Collapsed GTF file to calculate QC metrics/Coverage reports

Reference genomes used for the pre-generated assemblies can be found at the links below. ERCC standards are added to below reference to help analyze ERCC spike in quantification.
Agilent hg19 assembly [https://www.gencodegenes.org/human/release_19.html]
Agilent hg38 assembly [https://www.gencodegenes.org/human/release_38.html]

Alternatively, you can create your own indices by creating PartekFlow assemblies.

About the Agilent MBC Pipelines

MBC pipelines differ from standard pipelines in coverage report and pre-processing algorithms prior to alignment. Pre-processing steps specific for MBC workflow include alignment using BWA-MEM, AGeNT LocatIt v2.0.5 and converting LocatIt output bam files into fastq files. Pre-processing steps helps identify and remove PCR duplicates.

The coverage report calculates duplication rate, estimated library complexity based on duplication and strandedness metrics. In case of MBC workflow, it is produced from LocatIt output BAM file with duplicate reads marked. In non-MBC workflow, the coverage report is produced from STAR aligned bam file but duplication rate and library complexity is not applicable.

The figure on the following page shows the workflow for pre-processing steps in MBC pipeline.

LocatIt output

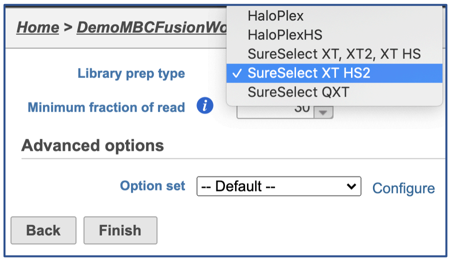
In MBC workflow, when generating the coverage report, run LocatIt in a mode in which the duplicates are marked but not removed. Then, for downstream analysis such as fusions and gene expression analysis, run LocatIt in a mode that removes duplicates.

There are 2 options to generate both duplicates-marked and duplicates-removed results.

1) Run LocatIt two separate times, the first time to just mark the duplicates and the second time to remove duplicates.

2) Use filter alignments to remove PCR and optical duplicates. Agilent recommends this approach to reduce computational resources.

Default parameter to be selected for trimmer as below.I If pipeline is imported, the default is populated for you. AGeNT trimmer performs MBC extraction and adaptor trimming. Default parameters shown in image below are optimal for SureSelect RNA XT HS2 data, please refer to Agilent AGeNT best practices document to understand more about how the defaults are chosen.



Home > DemoMBCFusionWo

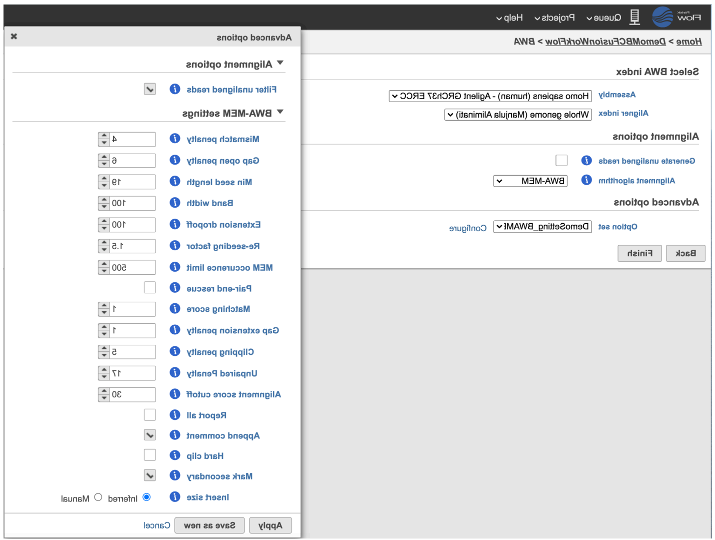
Library prep type: ✓ SureSelect XT HS2

Minimum fraction of read: 30

Advanced options

Option set: -- Default --

Trimmed data is aligned using BWA-MEM with the settings shown below. Another aligner could be used in this step. BWA was chosen for its speed and compatibility with SAM tags necessary for LocatIt (tags include barcode information).



Home > DemoMBCFusionWorkFlow > BWA

Select BWA index

Assembly: Homo sapiens (human) - Agilent GRCh37 ERCC

Aligner index: Whole genome (Manjula Aliminati)

Alignment option

Generate unaligned reads: □

Alignment algorithm: BWA-MEM

Advanced options

Option set: DemoSetting_BWAMEM

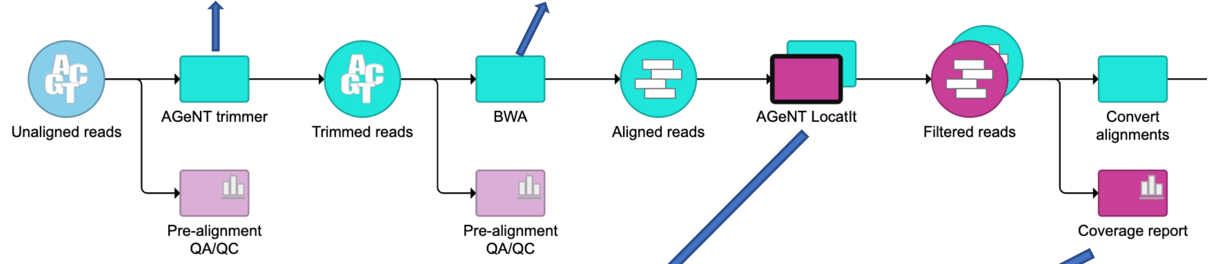
Advanced options

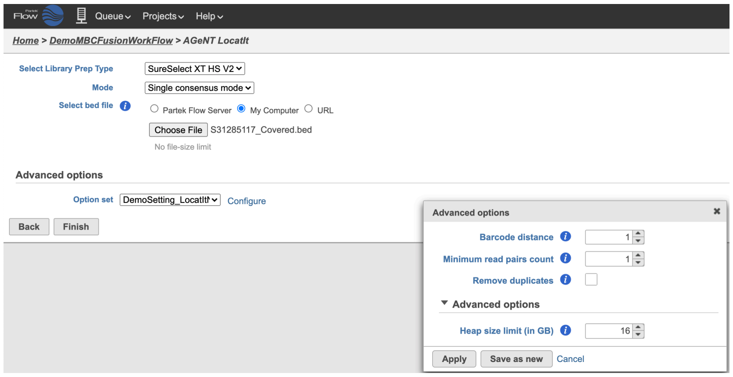
Alignment options

Filter unaligned reads: ☑

BWA-MEM settings

Mismatch penalty: 4
Gap open penalty: 6
Min seed length: 19
Band width: 100
Extension dropoff: 100
Re-seeding factor: 1.5
MEM occurrence limit: 500
Pair-end rescue: □
Matching score: 1
Gap extension penalty: 1
Clipping penalty: 5
Unpaired Penalty: 17
Alignment score cutoff: 30
Report all: □
Append comment: ☑
Hard clip: □
Mark secondary: ☑
Insert size: ● Inferred ○ Manual





Home > DemoMBCFusionWorkFlow > AGeNT LocatIt

Select Library Prep Type: SureSelect XT HS V2

Mode: Single consensus mode

Select bed file: ○ Partek Flow Server ● My Computer ○ URL

Choose File S31285117_Covered.bed

No file-size limit

Advanced options

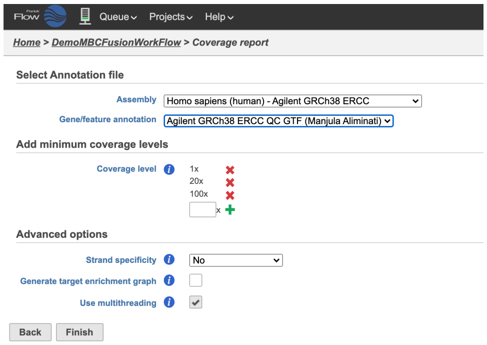
Option set: DemoSetting_LocatIt

Advanced options

Barcode distance: 1
Minimum read pairs count: 1
Remove duplicates: □

Advanced options

Heap size limit (in GB): 16



Home > DemoMBCFusionWorkFlow > Coverage report

Select Annotation file

Assembly: Homo sapiens (human) - Agilent GRCh38 ERCC

Gene/feature annotation: Agilent GRCh38 ERCC QC GTF (Manjula Aliminati)

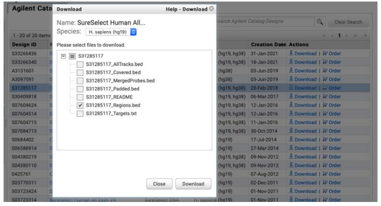
Add minimum coverage levels

Coverage level: 1x, 20x, 100x

Advanced options

Strand specificity: No
Generate target enrichment graph: □
Use multithreading: ☑

LocatIt is run in single consensus mode optimal for RNA-Seq data, For more information about LocatIt optimal parameters see the AGeNT best practices document. LocatIt identifies PCR duplicates and calculatesduplicate metrics, barcodes statistics, and metrics characterizing covered and non-covered regions. Covered and non-covered metrics are calculated from BED file input. The BED file included in the pipelines is the V7 exome design files from SureDesign. If you are not interested in covered metrics and only interested in duplicate metrics, you can provide any sample BED file. If you want to obtain covered metrics, download the respective covered.bed or regions.bed file from SureDesign as input to LocatIt. If you like to create your own BED file, make sure it contains only non-overlapping intervals.



Agilent Cata

Download — Help - Download

Name: SureSelect Human All...

Species: H. sapiens (hg19)

Please select files to download.

- S31285117
  - □ S31285117_AllTracks.bed
  - □ S31285117_Covered.bed
  - □ S31285117_MergedProbes.bed
  - □ S31285117_Padded.bed
  - □ S31285117_README
  - ☑ S31285117_Regions.bed
  - □ S31285117_Targets.txt

1 - 20 of 20 items

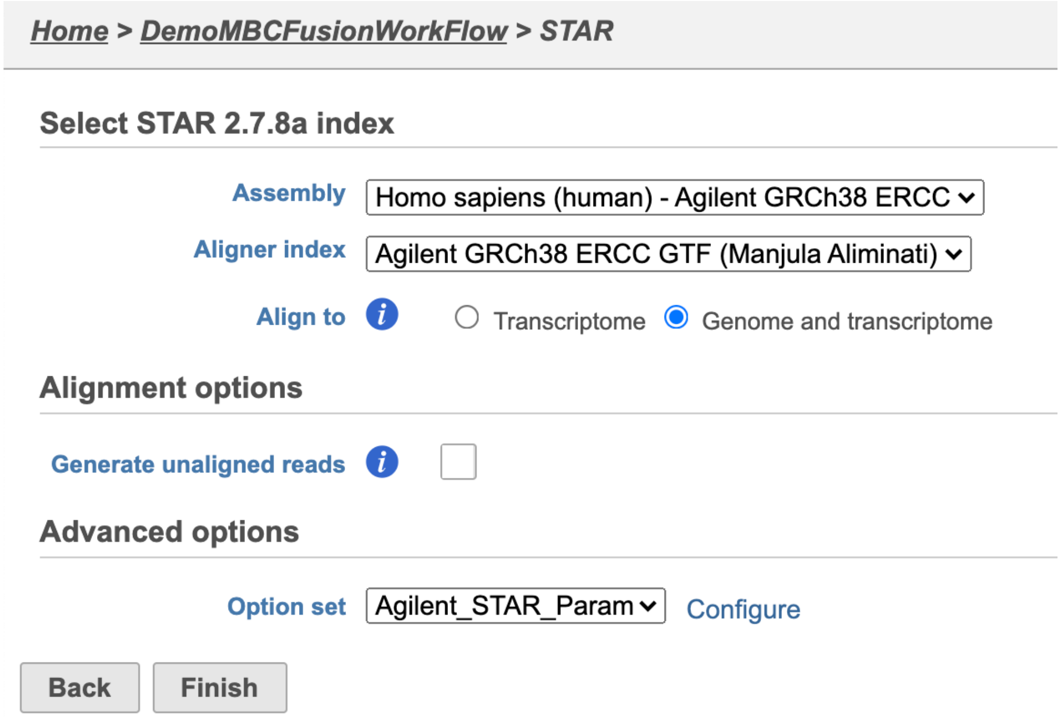
Design ID		Creation Date	Actions
S33266436	(hg19, hg38)	31-Jan-2021	Download | Order
S33266340	(hg19, hg38)	18-Jan-2021	Download | Order
A3131601	(hg38)	03-Jun-2019	Download | Order
A3097591	(hg38)	03-Jun-2019	Download | Order
S31285117	(hg19, hg38)	28-Feb-2018	Download | Order
S30409818	(hg19, hg38)	06-Mar-2017	Download | Order
S07604624	(hg19, hg38)	12-Jan-2016	Download | Order
S07604514	(hg19, hg38)	12-Jan-2016	Download | Order
S07604715	(hg19, hg38)	11-Jan-2016	Download | Order
S07084713	(hg19, hg38)	30-Oct-2014	Download | Order
S0684402	(hg19)	17-Jul-2014	Download | Order
S06588914	(hg19)	27-Mar-2014	Download | Order
S04380219	(hg19, hg38)	09-Nov-2012	Download | Order
S04380110	(hg19, hg38)	09-Nov-2012	Download | Order
0425761	(hg19)	07-Aug-2012	Download | Order
S03770311	(hg19)	02-Dec-2011	Download | Order
S03723424	(hg19)	01-Nov-2011	Download | Order
S03723314	SureSelect Human All Exon V4 SureSelect DNA H. sapiens (hg19)	01-Nov-2011	Download | Order

Coverage report can be generated from LocatIt BAM files that have duplicates reads marked. Metrics such as duplicate rate and library complexity can be obtained as a result. Running Coverage report requires an annotation file. A default gtf file is available in assemblies with the pipelines. This gtf is created as per Broad Institute recommendation in the following link: https://github.com/broadinstitute/gtexpipeline/tree/master/gene_model.

About the STAR Alignment Algorithm

Alignment of reads to the reference genome/transcriptome is performed using the STAR [Spliced Transcripts Alignment to a Reference] aligner. Two different sets of parameter [option sets] are available based on the intended secondary application. For fusion detection, the option set is called “Agilent_STAR_Params_For_Fusions”, specific for chimera detection. For gene expression, the option set is called “Agilent_STAR_Params_For_GE”. If you imported the Agilent Pipelines, then these options sets are available to you by default.

The screenshot below shows an example index and alignment.



Home > DemoMBCFusionWorkFlow > STAR

Select STAR 2.7.8a index

Assembly: Homo sapiens (human) - Agilent GRCh38 ERCC

Aligner index: Agilent GRCh38 ERCC GTF (Manjula Aliminati)

Align to: ○ Transcriptome ● Genome and transcriptome

Alignment options

Generate unaligned reads: □

Advanced options

Option set: Agilent_STAR_Param

The following table lists the recommended parameter selections for fusion-specific STAR alignment.

Option	Value
Generate unaligned reads	false
Max junctions	1000000
Type of filtering	Normal
Multimap score range	1
Max read mapping	10
Max mismatches	10
Mismatch mapped ratio	0.3
Mismatch read ratio	1.0
Min score	0
Normalized min score	0.66

Option	Value
Min matched bases	0
Normalized min matched bases	0.66
Filter alignment using their motifs	None
Collapsed splice junctions reads	All
Max junction gap	50000 100000 200000
Non-canonical motifs	true
Min overhang length for splice junctions	30
Min unique map read count per junction	3
Min total read count per junction	3
Min distance to other junctions' donor/acceptor	10
GT/AG motif	true
Min overhang length for splice junctions	12
Min unique map read count per junction	1
Min total read count per junction	1
Min distance to other junctions' donor/acceptor	0
GC/AG motif	true
Min overhang length for splice junctions	12
Min unique map read count per junction	1
Min total read count per junction	1
Min distance to other junctions' donor/acceptor	5
AT/AC motif	true
Min overhang length for splice junctions	12
Min unique map read count per junction	1
Min total read count per junction	1
Min distance to other junctions' donor/acceptor	10
Extra alignment score	2
Gap open penalty	0
Non-canonical gap open penalty	-8
GC/AG gap open penalty	-4
AT/AC gap open penalty	-8
Extra score	-0.25
Deletion open penalty	-2
Deletion extension penalty per base	-2
Insertion open penalty	-2

Option	Value
Insertion extension penalty per base	-2
Max score reduction	1
Search start point	50
Normalized search start point	1.0
Max seed length	
Max mapping for stitching	10000
Max seeds per read	1000
Max seeds per window	50
Max one seed loci per window	10
Min intron size	21
Max intron size	100000
Min spliced alignment overhang	5
Min annotated spliced alignment overhang	10 (Default: 3)
Max windows per read	10000
Max transcripts per window	100
Max hits	10000
Read ends alignment type	Local
Soft-clip past reference end	Yes
Max loci anchors	50
Bin size for windows/clustering	16
Max bins between two anchors	9
Left and right flanking region size	4
Chimeric alignment	true (Default: false)
Min chimeric segment length	12 (Default: 20)
Min total score of chimeric segments	0
Max difference of total chimeric score from read length	20
Min separation between best score and next	10
Penalty for non-GT/AG chimeric junction	-1
Min chimeric junction overhang	8 (Default: 20)
Two pass mapping	Per-sample (Default: None)
Cufflinks-like strand field flag	None (Default: intronMotif)
SAM attributes	Standard
Add to quality score	0

Option	Value
WASP filtering	false
Max splice junction stitching mismatches	Non-canonical:5 GT/AG and CT/AC: -1 GC/AG and CT/GC:5 AT/AC and GT/AT:5 (Default: Non-canonical:0 GT/AG and CT/AC: -1 GC/AG and CT/GC:0 AT/AC and GT/AT:0)
Flush ambiguous insertion positions	Right (Default: None)
Min overlap for mate merging and realignment	12 (Default: 0)
Max mismatched bases in overlap area	0.1 (Default: 0.01)
Max gap between chimeric segments	0
Filter alignments with Ns around juction	true
Max multi-alignments for main chimeric segment	10
Max chimeric multi-alignments	20 (Default: 0)
Multi-alignment score range	3 (Default: 1)
Non-chimeric alignment score drop min	10 (Default: 20)

The following table lists the default option set proposed for STAR alignemnt for gene expression.

Option	Value
Generate unaligned reads	false
Name, sequence, and quality lengths	NotEqual
Max junctions	1000000
Type of filtering	Normal
Multimap score range	1
Max read mapping	10
Max mismatches	10
Mismatch mapped ratio	0.3
Mismatch read ratio	1.0
Min score	0
Normalized min score	0.66
Min matched bases	0
Normalized min matched bases	0.66
Filter alignment using their motifs	None
Collapsed splice junctions reads	All
Max junction gap	50000 100000 200000

Option	Value
Non-canonical motifs	true
Min overhang length for splice junctions	30
Min unique map read count per junction	3
Min total read count per junction	3
Min distance to other junctions' donor/acceptor	10
GT/AG motif	true
Min overhang length for splice junctions	12
Min unique map read count per junction	1
Min total read count per junction	1
Min distance to other junctions' donor/acceptor	0
GC/AG motif	true
Min overhang length for splice junctions	12
Min unique map read count per junction	1
Min total read count per junction	1
Min distance to other junctions' donor/acceptor	5
AT/AC motif	true
Min overhang length for splice junctions	12
Min unique map read count per junction	1
Min total read count per junction	1
Min distance to other junctions' donor/acceptor	10
Extra alignment score	2
Gap open penalty	0
Non-canonical gap open penalty	-8
GC/AG gap open penalty	-4
AT/AC gap open penalty	-8
Extra score	-0.25
Deletion open penalty	-2
Deletion extension penalty per base	-2
Insertion open penalty	-2
Insertion extension penalty per base	-2
Max score reduction	1
Search start point	50
Normalized search start point	1.0
Max seed length	

Option	Value
Max mapping for stitching	10000
Max seeds per read	1000
Max seeds per window	50
Max one seed loci per window	10
Min intron size	21
Max intron size	
Max gap between two mates	
Min spliced alignment overhang	5
Min annotated spliced alignment overhang	3
Spliced mate min read length	0
Normalized spliced mate min read length	0.66
Max windows per read	10000
Max transcripts per window	100
Max hits	10000
Read ends alignment type	Local
Soft-clip past reference end	Yes
Max loci anchors	50
Bin size for windows/clustering	16
Max bins between two anchors	9
Left and right flanking region size	4
Chimeric alignment	false
Two pass mapping	None
Cufflinks-like strand field flag	None (Default: intronMotif)
SAM attributes	Standard
Add to quality score	0
WASP filtering	false
Max splice junction stitching mismatches	Non-canonical:0 GT/AG and CT/AC: -1 GC/AG and CT/GC:0 AT/AC and GT/AT:0
Flush ambiguous insertion positions	None
Min overlap for mate merging and realignment	0
Max mismatched bases in overlap area	0.01

The STAR algorithm produces a log file that contains alignment statistics. You can download the file from **Task details > Output files**. Alternatively, you can also run the post-alignment QA/QC tool.

About the Fusion Pipeline

Fusions are detected using the STAR-Fusion program. More information about the algorithm and output can be found at https://github.com/STAR-Fusion/STAR-Fusion/wiki.

The following table lists the default parameters for fusion detection.

Option	Value
Enable filtering	true
Min junction reads	1
Min fusion support	2
Require long double achor support	true
Max promiscuity	10
Min percent dominant promiscuity	20
Aggregate novel junction distance	5
Min novel junction support	3
Min spanning framents only	5
Min alt percent juction	10.0
Minimum FFPM	0.1
Remove duplicates	true
Skip EM	false
Skip FFPM	false
Annotation filter	true
RT artifacts filter	true
Single fusion per breakpoint filter	true
Examine coding effect	false
Trinity denovo assembly	false

The annotation filter for STAR-Fusion v1.9.1 has a known bug and has been fixed in Partek Flow. See reference below:
https://groups.google.com/g/star-fusion/c/THb6TxGrSBg

About the Gene Expression Pipeline

For gene expression analysis, create and assign attributes to the imported data prior to analysis. For example, if the data has control samples or replicates, assign samples as controls and replicates accordingly. Differential gene expression pipelines provided by Agilent expects atleast 2 samples with minimum 2 replicates each. This information will be used by downstream processes in normalizing and calculating fold change.

Quantification and differential gene expression can be performed using Partek Flow E/M Quantification Model > Median ratio for normalization > DESeq2 for differential gene expression, as offered by Partek Flow.

Prior to the normalization step, Agilent recommends processing data through a noise reduction filter to remove very low expressors from sample data. This can be done using function Filtering > Filter features.

Appendix

Version of software modules integrated:

Software	Version
STAR	2.7.8a
BWA	0.7.17
STAR-Fusion	1.9.1
AGeNT Trimmer	2.0.5
AGeNT LocatIt	2.0.5
PartekFlow	10.0.21.0719

References

STAR aligned → Dobin, Alexander et al. "STAR: ultrafast universal RNA-seq aligner." Bioinformatics (Oxford, England) vol. 29,1 (2013): 15-21. doi:10.1093/bioinformatics/bts635

STAR-Fusion → Brian J. Haas, Alex Dobin, Nicolas Stransky, Bo Li, Xiao Yang, Timothy Tickle, Asma Bankapur, Carrie Ganote, Thom as G. Doak, Nathalie Pochet, Jing Sun, Catherine J. Wu, Thomas R. Gingeras, Aviv Regev

PR7000-3043
Revision A