Table of Contents

[1 Introduction 1](#_Toc96612618)

[2 Installation 1](#_Toc96612619)

[3 BUILD INDEX 2](#_Toc96612620)

[4 FUSIONS DETECTION 3](#_Toc96612624)

[**4.1 Prepare** 3](#_Toc96612625)

[**4.2 Run arriba** 3](#_Toc96612626)

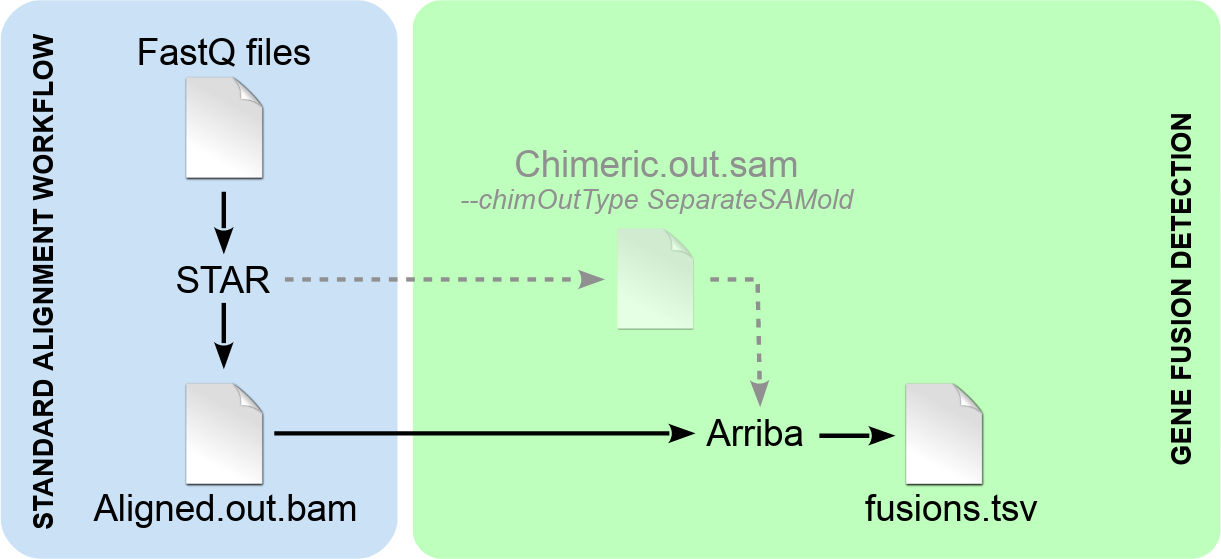
[**4.3 Output file interpretation** 4](#_Toc96612626)

# **1 Introduction**

This pipeline could accurately and efficiently identify gene fusion from RNA-Seq data. It was developed for the use in a clinical research setting. Alignment is done by the ultrafast STAR aligner, and gene fusions detection tool arriba is based on STAR output result.

## Workflow

Fusion detection with Arriba is based on the STAR aligner. It is an extension of the regular alignment workflow, which can be incorporated with few modifications. The addition of Arriba does not affect the normal alignments. The workflow yields fusion predictions as well as normal alignments that can be used for other downstream analyses such as expression quantification or variant calling. Like so, fusion detection incurs negligible computational overhead, since it adds only a few minutes of runtime to the regular alignment workflow.



# **2 Installation**

### STAR

Spliced Transcripts Alignment to a Reference (STAR) is written in C and C++ and can be run as a standalone application on diverse hardware systems. It is freely available to personal, academic and non-profit use only. You cannot redistribute ANNOVAR to other users including lab members. Download the latest package from the STAR [releases page](https://github.com/alexdobin/STAR/releases) as shown:

Example:

wget <https://github.com/alexdobin/STAR/archive/refs/tags/2.7.10a.tar.gz>

tar -xzf 2.7.10a.tar.gz

cd STAR-2.7.10a/source

make STAR # Compile under Linux

After compilation, add STAR to environment variable. #remember to change the path to the absolute path to STAR-2.7.10a/source in your server.

Example:

export PATH=/mnt/sata/Janet\_Analysis/GeneFusion/STAR-2.7.10a/source/:$PATH

### ARRIBA

Arriba has only a single prerequisite: STAR (version >=2.7.10a recommended). Download the latest tarball from the arriba [releases page](https://github.com/suhrig/arriba/releases/) as shown.

Example:

wget https://github.com/suhrig/arriba/releases/download/v2.2.1/arriba\_v2.2.1.tar.gz

tar -xzf arriba\_v2.2.1.tar.gz

cd arriba\_v2.2.1 && make

Arriba requires an assembly in FastA format, gene annotation in GTF format, and a STAR index built from the two. So STAR index need to be built before running arriba.

# **3 BUILD INDEX**

If you do not already have the files and a STAR index, you can use the script download\_references.sh. It downloads the files to the current working directory and builds a STAR index. GENCODE annotation is recommended over RefSeq due to more comprehensive annotation of immunoglobulin/T-cell receptor loci and splice sites, which improves sensitivity.

Run the script without arguments to see a list of available files.



Download reference files and build index:

./download\_references.sh $ASSEMBLIES+ANNOTATIONS

Example:

./download\_references.sh hg19+GENCODE19

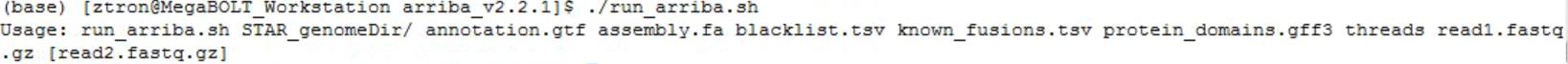
Output:



# **4 FUSIONS DETECTION**

## **4.1 Prepare**

Run the script without arguments to check usage.



*blacklist.tsv, known\_fusions.tsv* and *protein\_domains.gff3* can be found under directory *arriba\_v2.2.1/database/.* Choose the corresponding files according to the coordinates (eg. hg19/hs37d5/GRCh37 shares the same chromosome coordinates, hg38/GRCh38 shares the the same chromosome coordinates and vice versa). Unzip files needed using command line:

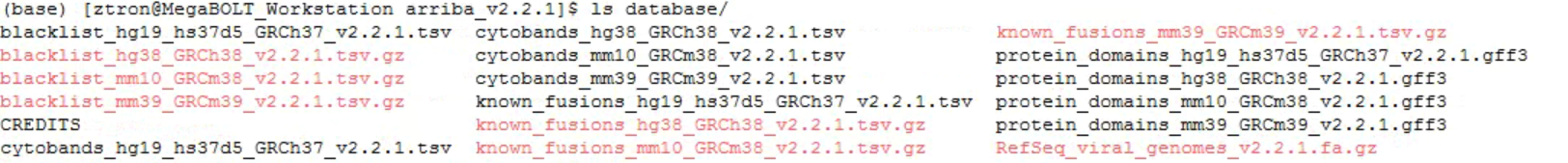
gunzip filename.gz

Example:

gunzip database/blacklist\_hg19\_hs37d5\_GRCh37\_v2.2.1.tsv.gz

gunzip database/known\_fusions\_hg19\_hs37d5\_GRCh37\_v2.2.1.tsv.gz

Output:



If you use other assemblies whose coordinates are incompatible with hg19/hs37d5/GRCh37 or hg38/GRCh38 or mm10/GRCm38 or mm39/GRCm39, then the coordinates in the blacklist will not match and the predictions will contain many false positives.

## **4.2 Run arriba**

Run arriba by demo script *run\_arriba.sh*

Example:

(Run the demo script with 8 threads)

sh run\_arriba.sh STAR\_index\_hg19\_RefSeq\_hg19/ RefSeq\_hg19.gtf hg19.fa \ database/blacklist\_hg19\_hs37d5\_GRCh37\_v2.2.1.tsv \ database/known\_fusions\_hg19\_hs37d5\_GRCh37\_v2.2.1.tsv \ database/protein\_domains\_hg19\_hs37d5\_GRCh37\_v2.2.1.gff3 8 read1.fq.gz read2.fq.gz

## **4.3 Output File Interpretation**

See: <https://arriba.readthedocs.io/en/latest/output-files/>

*-End-*