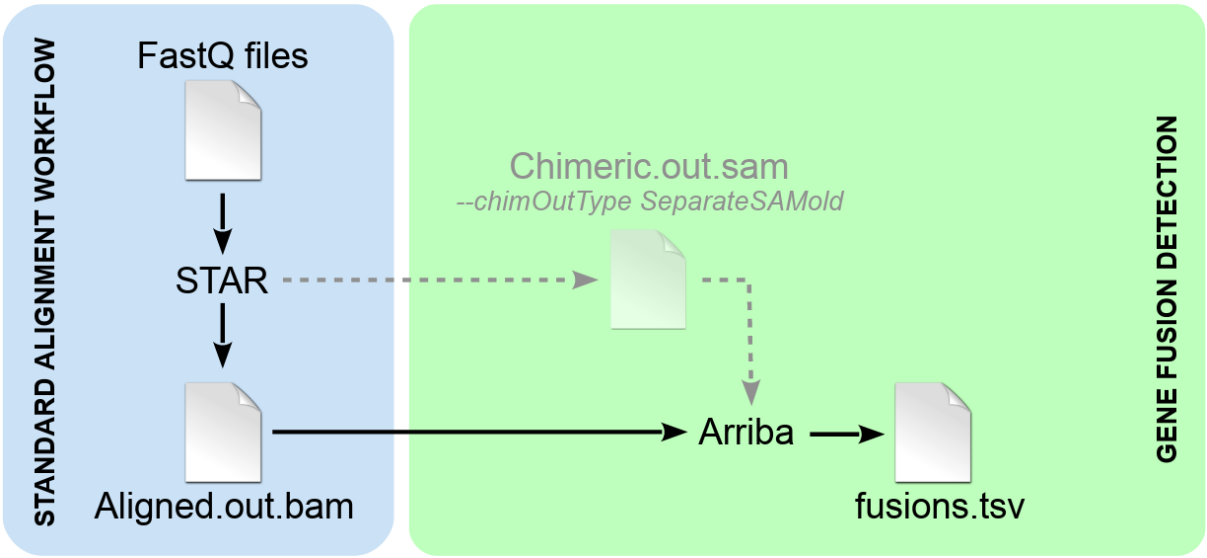
TABLE OF CONTENTS

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 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 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.

```
(base) [ztron@MegaBOLT_Workstation arriba_v2.2.1]$ ./download_references.sh hg19+RefSeq_hg19
Usage: download_references.sh ASSEMBLY+ANNOTATION
Available assemblies and annotations:
GRCh37+ENSEMBL87
GRCh37+GENCODE19
GRCh37+RefSeq
GRCh37viral+ENSEMBL87
GRCh37viral+GENCODE19
GRCh37viral+RefSeq
GRCh38+ENSEMBL93
GRCh38+GENCODE28
GRCh38+RefSeq
GRCh38viral+ENSEMBL93
GRCh38viral+GENCODE28
GRCh38viral+RefSeq
GRCm38+GENCODEM25
GRCm38+RefSeq
GRCm38viral+GENCODEM25
GRCm38viral+RefSeq
GRCm39+GENCODEM26
GRCm39+RefSeq
GRCm39viral+GENCODEM26
GRCm39viral+RefSeq
hg19+ENSEMBL87
hg19+GENCODE19
hg19+RefSeq
hg19viral+ENSEMBL87
hg19viral+GENCODE19
hg19viral+RefSeq
hg38+ENSEMBL93
```

Download reference files and build index:

```
./download_references.sh $ASSEMBLIES+ANNOTATIONS
```

Example:

```
./download_references.sh hg19+GENCODE19
```

Output:

```
(base) [ztron@MegaBOLT_Workstation arriba_v2.2.1]$ ./download_references.sh hg19+GENCODE19
Downloading assembly: http://hgdownload.cse.ucsc.edu/goldenpath/hg19/bigZips/chromFa.tar.gz
Downloading annotation: http://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_19/gencode.v19.annotation.gf.gz
        /mnt/sata/Janet_Analysis/GeneFusion/STAR-2.7.10a/source/STAR --runMode genomeGenerate --genomeDir STAR_index_hg19_GENCODE19 --ge
nomeFastaFiles hg19.fa --sjdbGTFfile GENCODE19.gtf --runThreadN 8 --sjdbOverhang 250
        STAR version: 2.7.10a   compiled: 2022-03-07T23:37:37+1000 MegaBOLT_Workstation:/mnt/sata/Janet_Analysis/GeneFusion/STAR-2.7.10a
/source
Mar 10 18:02:02 ..... started STAR run
Mar 10 18:02:02 ... starting to generate Genome files
Mar 10 18:02:49 ..... processing annotations GTF
Mar 10 18:03:13 ... starting to sort Suffix Array. This may take a long time...
Mar 10 18:03:27 ... sorting Suffix Array chunks and saving them to disk...
Mar 10 18:20:26 ... loading chunks from disk, packing SA...
Mar 10 18:21:30 ... finished generating suffix array
Mar 10 18:21:30 ... generating Suffix Array index
Mar 10 18:25:46 ... completed Suffix Array index
Mar 10 18:25:46 ..... inserting junctions into the genome indices
Mar 10 18:31:30 ... writing Genome to disk ...
Mar 10 18:31:32 ... writing Suffix Array to disk ...
Mar 10 18:32:30 ... writing SAindex to disk
Mar 10 18:32:39 ..... finished successfully
(base) [ztron@MegaBOLT_Workstation arriba_v2.2.1]$ ls STAR_index_hg19_GENCODE19/
chrLength.txt      chrStart.txt      geneInfo.tab          Log.out  sjdbInfo.txt              transcriptInfo.tab
chrNameLength.txt  exonGeTrInfo.tab  Genome                SA       sjdbList.fromGTF.out.tab
chrName.txt        exonInfo.tab      genomeParameters.txt  SAindex  sjdbList.out.tab
```

4 FUSIONS DETECTION

4.1 Prepare

Run the script without arguments to check usage.

```
(base) [ztron@MegaBOLT_Workstation arriba_v2.2.1]$ ./run_arriba.sh
Usage: run_arriba.sh STAR_genomeDir/ annotation.gtf assembly.fa blacklist.tsv known_fusions.tsv protein_domains.gff3 threads read1.fastq
.gz [read2.fastq.gz]
```

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:

```
(base) [ztron@MegaBOLT_Workstation arriba_v2.2.1]$ ls database/
blacklist_hg19_hs37d5_GRCh37_v2.2.1.tsv  cytobands_hg38_GRCh38_v2.2.1.tsv            known_fusions_mm39_GRCm39_v2.2.1.tsv.gz
blacklist_hg38_GRCh38_v2.2.1.tsv.gz      cytobands_mm10_GRCm38_v2.2.1.tsv            protein_domains_hg19_hs37d5_GRCh37_v2.2.1.gff3
blacklist_mm10_GRCm38_v2.2.1.tsv.gz      cytobands_mm39_GRCm39_v2.2.1.tsv            protein_domains_hg38_GRCh38_v2.2.1.gff3
blacklist_mm39_GRCm39_v2.2.1.tsv.gz      known_fusions_hg19_hs37d5_GRCh37_v2.2.1.tsv protein_domains_mm10_GRCm38_v2.2.1.gff3
CREDITS                                  known_fusions_hg38_GRCh38_v2.2.1.tsv.gz     protein_domains_mm39_GRCm39_v2.2.1.gff3
cytobands_hg19_hs37d5_GRCh37_v2.2.1.tsv  known_fusions_mm10_GRCm38_v2.2.1.tsv.gz     RefSeq_viral_genomes_v2.2.1.fa.gz
```

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-