**STAR Chimeric Post Manual**

Version 1.1

Dr. Nicholas Kipp Akers

Prof. Bojan Losic Laboratory

Icahn Institute of Genomics and Multiscale Biology

Mt Sinai School of Medicine

**Table of Contents**

**Introduction**:

This software was written in order to fulfill a perceived lack of tools that can identify, from RNA-Seq data, high confidence fusions and circular RNA. There are many other tools to discover fusions in RNA, and a handful of other tools to discover circular RNA. We have endeavored to create the lowest possible false-positive rate for fusion detection, while creating helpful and convenient output files. We’ve also emphasized the need for short running times in order to make this software useful for enormous data sets.

**Installation**

starchimp was written to work in a \*nix shell-like environment, using Bash and Perl extensively. To use this on Windows, a shell emulator such as Cygwin is likely necessary. It will probably work in Mac OS, but this is untested and likely there will be formatting issues. At this point we will assume you are in a \*nix shell:

Within the directory you’d like the software installed:

git clone <https://github.com/kippakers/starchimp.git>

Please ensure that the files in starchimp/bin/ have execute permission:

chmod +x starchimp/bin/\*

That’s it!

There are a number of software dependencies. These should be callable without a full path (ie they should be included in your $PATH :

Samtools (http://www.htslib.org/)

Bedtools >= 2.24.0 (https://github.com/arq5x/bedtools2)

Mafft (http://mafft.cbrc.jp/alignment/software/)

R (<https://www.r-project.org/> )

R packages limma, edgeR (<https://bioconductor.org/packages/release/bioc/html/limma.html>)

STAR (<https://github.com/alexdobin/STAR>)

Most \*nix installs will have the following pre-installed, but you will also need Bash and a version of Perl less than ~ 5 years old.

**Setup**

**STAR**:

starchimp is written to be an extension of the STAR read aligner. It does not itself run STAR on your samples. Therefore to run starchimp you must first run star on each of your samples. See the STAR documentation for installation, as well as building or downloading a STAR genome index. It is absolutely critical however, that you follow the STAR manual’s instructions and **build a genome using all chromosomes plus unplaced contigs**. Not doing so will strongly inflate your false positives rate. Run STAR with the following parameters required for chimeric output: --chimSegmentMin 15 --chimJunctionOverhangMin 15. Your project will have it’s own requirements, but a good starting point for your star alignments might look like:

STAR --genomeDir /path/to/starIndex/ --readFilesIn file1\_1.fastq.gz file1\_2.fastq.gz --runThreadN 11 --outReadsUnmapped Fastx --quantMode GeneCounts

--chimSegmentMin 15 --chimJunctionOverhangMin 15 --outSAMstrandField intronMotif --readFilesCommand zcat --outSAMtype BAM Unsorted

**References/BED Files**

starchimp makes use of gtf files for annotating fusions and cRNA with gene names. However, we require them in ucsc bed file format. We also require a bedtools .genome file to aid in sorting. These can automatically be generating from your gtf file using the command:

/starchimp/bin/gtf2bed.sh /path/to/your/file.gtf /path/to/desired/output/dir/

GTF files can be downloaded from Ensembl, UCSC, gencode, refseq, and other online sources. It is recommended to use the same GTF that your STAR index was built with.

starchimp-fusions filters using the location of known repeats in bed format as well. This file will be large (~300MB for human), and can be downloaded from the UCSC table browser:

1. Go to <http://genome.ucsc.edu/cgi-bin/hgTables>
2. Change ‘genome ‘to your desired genome
3. Change the following settings:
   1. group: Repeats
   2. track: RepeatMasker
   3. region: genome
   4. output format: BED
   5. output file: some reasonable name.bed
4. Click ‘get output’ to download your bed file.

starchimp-fusions can also make use of known antibody parts, and copy number variants. These files come with starchimp for human hg19 and hg38 in the data directory. For other species you can create your own in the simple format:

Chromosome StartPosition EndPosition

Finally, starchimp-fusions uses known gene families and known/common false-positive pairs to filter out fusions which are likely mapping errors or PCR artifacts. Family data can be downloaded from ensembl biomart:

1. Go to <http://www.ensembl.org/biomart/martview>
2. Database: Ensembl Genes
3. Dataset: Your species
4. Click Attributes on the left hand side.
   1. Under GENE dropdown, select only “Associated Gene Name”
   2. Under PROTEIN FAMILIES AND DOMAINS dropdown select Ensembl Protein Family ID.
5. Click Results at the top.
6. Export the file. It should have two columns, Family ID and Gene ID.

Known false positives are stored within data/pseudogenes.txt Feel free to put add any additional lines that result from your data to this file in the format:

Gene1Name Gene2Name

**Parameter Files**

starchimp-fusions and starchimp-circles rely on parameter files to supply all the information to complete a successful run. Examples are given in the paramfiles directory, however **you need to customize a parameter file for each run.** Not doing so is akin to buying shoes without specifying what style or size of shoe you desire. All required and optional parameters are explained here. We recommend you copy an example parameter file to your working directory and edit it there. All parameter values are explained at the end of the manual.

**Running starchimp**

**starchimp-fusions**

starchimp-fusions is run on individual samples.

/path/to/starchimp/bin/starchimp-fusions.pl output\_seed Chimeric.out.junction Paramters.txt

output\_seed is the unique preface to your output file; e.g. sample1, or output/sample1

Chimeric.out.junction is the full path to your STAR output file Chimeric.out.junction.

Parameters.txt is your unique parameters file for this job.

**starchimp-circles**

starchimp-circles is run on groups of samples.

/path/to/starchimp/bin/starchimp-circles.pl STARdirs.txt Parameters.txt

STARdirs.txt is a text file with a full path to one STAR output directory per line

Parameters.txt is your parameters file for this job.

**Complete Parameter Explanations**

*Fusions Parameter Files:*

Parameter Explanation

pairedend True/false : is your data paired end?

consensus True/false : do you want a consensus sequence generated for your fusions? Should be true unless you cannot use mafft or samtools for some reason

splitReads Integer. Number of minimum reads that cross fusion border. Can put ‘auto’ to have starchimp make a best guess for your sample. (Warning: this method is very simple).

spancutoff Integer. Number of minimum paired ends to map to opposite ends of a fusion without crossing the fusion site. Can put ‘auto’ to have starchimp make a best guess for your sample. (Warning: this method is very simple).

uniqueReads Integer. Number of minimum unique reads that cross fusion border. Useful for eliminating fusions whose support comes from a single PCR amplicon.

wiggle Integer. Reads mapping to opposite ends of a fusion without crossing the fusion junction itself have ambiguous fusion site. Starchimp looks for fusions within ‘wiggle’ bp of the ambiguous fusion site to share support.

overlapLimit Integer. Fusions called less than overlapLimit bp from each other will be merged.

samechrom\_wiggle Integer. Intrachromosomal fusions must be at least this value apart to be reported.

lopsidedupper Float. Starchimp will filter out fusions that are imbalanced in the number of reads on one strand vs another. With unstranded RNA-seq, we expect roughly equal reads to map to the ‘top’ strand vs the ‘bottom’ strand. Using the equation: ratio = (top side reads + 0.1) / (bottom side reads + 0.1). ratio must be below ‘lopsidedupper’ and above ‘lopsidedlower’

lopsidedlower see above.

genome String. Starchimp will look for files in the data directory bearing this genome ID. For example, genome = hg19 will cause starchimp to look for and use data/hg19.abparts and data/hg19.cnvs.

cnvwiggle Integer. Skip fusions with an edge within ‘cnvwiggle’ bp of a known cnv.

circlesize Integer. Fusions that appear to be circular RNA (same strand, fusion splices ‘backwards’) are skipped. However, circlesize represents an upper limit for the size of filtered circular RNA.

refbed String. The bed file generated earlier with gtf2bed.sh.

repeatbed String. The bed file downloaded from UCSC table browser.

refFasta String. A genome fasta file (preferably the one used to build the STAR index). If not the same it MUST be the same genome build and it must have the same chromosome identifiers.

splitscoremod Float. Each fusion will have a confidence score generated based on the number of reads of support and the strand imbalance of that support. This score is then adjusted if the read support has a poor skew, if it’s possible it’s a read through fusions, and if the fusion sites are in repeats.

spanscoremod Float. See above.

skewpenalty Float. See above.

repeatpenalty Float. See above. Score is modified to be score=score\*(repeatspenalty^repeats) where a fusion can have 0,1, or 2 sites fall into repeat regions.

*Circular RNA Parameters File*

readsCutoff Integer. Minimum number of reads crossing the circular RNA backsplice required. To check multiple cutoffs (eg to examine 5, 10, and 15 reads, use the following format (within quotes) “( 5 10 15 )” )

minSubjectLimit Integer. Minimum number of individuals with ‘readsCutoff’ reads required to carry forward a cRNA for analysis.

cpus Integer. Number of threads to use. If a non-numeric value is given, the command nproc will be used to determine the number of threads.

do\_Splice true/false. Should the splices within the cRNA be detected and reported? Linear splices are searched within each cRNA in each individual. Any linear splice with >= 60% of the read count of the cRNA is considered a splice within the cRNA. Two files are then created, \*.consensus with most common splice pattern, and \*.allvariants with all reported splice patterns.

cpmCutoff Float. Reads counts are loaded into R and log2(CountsPerMillion) is calculated using the limma package. With cpmCutoff > 0, cRNA with log2(CPM) below this value will be filtered from this analysis.

subjectCPMcutoff Integer. See above. This value is the lower limit for number of individuals required to have the cRNA expressed at a value higher than cpmCutoff.

annotate true/false. Should cRNA be given gene annotations? Uses refbed.

refbed String. The bed file generated earlier with gtf2bed.sh.

starprefix String. If you used the star parameter --outFileNamePrefix, give that value here. E.g. if your star output has a file named “mydata\_Chimeric.out.junction” then put “mydata\_” for starprefix.

IDstepsback Integer. Where in your pathway (position from the right) is the sample identifier. For example if your star output for sample1 is in the directory at : path/to/sample1/star/2.4.2/output/Chimeric.out.junction

Your IDstepsback is 4. Alternatively the path

/path/to/star/2.4.2/sample1/Chimeric.out.junction

has IDstepsback value of 1.