Script run_subread-align.sh for running subread

```
#!/usr/bin/env bash
# run_subread-align.sh
# Kamil Slowikowski
# October 6, 2014
#
# Run the subread aligner.
#
# Usage:
#For Broad
# parallel --env _ -S 10/piglet,8/tigger --colsep '\t' /home/unix/slowikow/work/rnaseq/scrip
#
# parallel --env _ -S 10/piglet,10/tigger --colsep '\t' /home/unix/slowikow/work/rnaseq/scri

# $RAM
# Human genome ref file and copying to memory - less reads from disk = faster
#source /home/unix/slowikow/work/rnaseq/scripts/copy_genome_to_mem.sh

#prog='/home/unix/slowikow/src/subread-1.4.5-p1-source/bin/subread-align'

fq="(\.fq|\.fastq|\.fq.gz|\.fastq.gz)$"
#two arguments (forward and reverse), both ending in same extension
if [[ $# != 2 || ! "$1" =~ $fq || ! "$2" =~ $fq ]]
then
    echo "Received: $0 $@"
    echo "Usage: $0 file.end1.fq[.gz] file.end2.fq[.gz]"
    echo
    exit 1
fi
#do the files exist
[[ ! -e "$1" ]] && echo "FASTQ file not found $1" && exit 1
[[ ! -e "$2" ]] && echo "FASTQ file not found $2" && exit 1

out=/data/srlab/pfizer/Subread
#create output files in different directory
[[ ! -d $out ]] && mkdir -p $out
log=$out/subread-align-log.txt
#text file containing status
opt=(
    --index /data/srlab/external-data/iGenomes/Homo_sapiens/UCSC/hg19/Sequence/SubreadIndex/
    --read $1
    --read2 $2
    --gzFASTQinput
    --subreads 20
    --BAMoutput
```

```
    --output $out/${1%%_*}.bam
    --order fr
    --mindist 0
    --maxdist 1000000
    --multi 2
    --quality
    --reportFusions
    --threads 4
)

(time subread-align ${opt[*]}) &> $log
```

Options:

index: specify the name of the index

read: name of the input file (FASTQ/FASTA format). First read file

read2: For paired-end reads, this is the name of the second read file

gzFASTQinput: the input read data are in gzipped FASTQ or FASTA format

subreads: Number of selected subreads. Default is 10. Kam suggests increasing this value for longer reads. e.g Pfizer data is 100bp so increase this to 20

BAMoutput: specify that mapping results are saved into a BAM format file

output: name of the output file

order: orientation of the two reads from the same pair (fr is default)

mindist: minimum fragment/template length

maxdist: maximum fragment/template length. Increasing this value allows for large introns

multi: maximum number of equally-best mapping locations allowed to be reported for each read

quality: use mapping quality scores to break ties when more than one best mapping location is found

reportFusions: report discovered genomic fusion events such as chimeras. Will be saved to a file .fusions.txt

threads: number of threads

Submitting to cluster

log into eris1n2.partners.org and move to directory containing sample files

```
bsub -q normal < /PHShome/ed686/bin/submit_run_subread-align.sh
```

Trouble shooting

If the script has been transferred to the server using windows SSH secure shell transfer, the file contains CRLF line endings (windows style) instead of LF endings (unix style)

Open script with vim key ESC, then type

```
:set fileformat=unix
:wq!
```

A few of my files were very big and failed to run so you can check those that were successful using (in the directory which contains your output)

```
grep -l "successfully" *.txt
```