hervQuant workflow pipeline

This document provides the details for setting up the environment and running the hervQuant workflow. Please note that hervQuant has currently only been optimized for 2x50bp stranded RNA-seq runs, and optimization for other runs may be needed (i.e. changes to STAR multimapping and mismatch parameters).

Software and References

STAR v2.5.3a (https://github.com/alexdobin/STAR/archive/2.5.3a.tar.gz)
Salmon v0.8.2 (https://github.com/COMBINE-lab/salmon/archive/v0.8.2.tar.gz)
Samtools v1.4 (https://github.com/samtools/samtools/releases/download/1.4/samtools-1.4.tar.bz2)
hERV reference sequences adapted from Vargiu, L. *et al.* (2016).
hervQuant reference with transcriptome ("hervquant_hg19_reference.fa")
hervQuant reference ("hervquant_final_reference.fa")

Environment

Although many environments could run hervQuant, the one we used was built on a Debian GNU/Linux 9 (stretch) background by running:

```
# some of these apt-get installations could be dropped for hervQuant, but were present in our case as they were needed for to run other elements of the pipeline.
apt-get update
apt-get -yq install \
  autoconf \
  ca-certificates \
  cmake \
  curl \
  default-jdk \
  g++ \
  gcc \
  libboost-all-dev \
  libbz2-dev \
  liblzma-dev \
  make \
  unzip \
  wget \
  zlib1g-dev

# install STAR
cd /opt
wget https://github.com/alexdobin/STAR/archive/2.5.3a.tar.gz && \
tar -zxf 2.5.3a.tar.gz && \
rm 2.5.3a.tar.gz && \
ln -s /opt/STAR-2.5.3a/bin/Linux_x86_64/STAR /usr/local/bin

# install salmon
wget https://github.com/COMBINE-lab/salmon/archive/v0.8.2.tar.gz && \
tar -zxf v0.8.2.tar.gz && \
rm v0.8.2.tar.gz && \
cmake salmon-0.8.2 -DCMAKE_INSTALL_PREFIX=/usr/local && \
```

```
make && \
make install

# install samtools
apt-get install -yq libncurses-dev
cd /opt
wget https://github.com/samtools/samtools/releases/download/1.4/samtools-1.4.tar.bz2
bzip2 -d samtools-1.4.tar.bz2
tar -xvf samtools-1.4.tar
samtools-1.4/configure
cd htslib-1.4/
make
cd /opt/samtools-1.4
make
ln -s /opt/samtools-1.4/samtools /usr/local/bin

apt-get clean

# build STAR reference
STAR \
  --runMode genomeGenerate \
  --runThreadN $num_threads \
  --limitGenomeGenerateRAM 52000000000 \
  --genomeSAindexNbases 7 \
  --genomeDir /path/to/hervquant_reference \
  --genomeFastaFiles /path/to/hervquant_reference/hervquant_hg19_reference.fa
```

Workflow commands

```
# align reads to reference
STAR \
  --runThreadN $num_threads \
  --outFileNamePrefix $file_prefix \
  --outFilterMultimapNmax 10 \
  --outFilterMismatchNmax 7 \
  --genomeDir /path/to/hervquant_reference/ \
  --readFilesIn ${FQ1} ${FQ2}

#Filter out all non-herv maps:
sam_file=${file_prefix}Aligned.out.sam
sam_file_filtered=${file_prefix}Aligned.out.filtered.sam
sed '/uc.*/d' $sam_file > $sam_file_filtered
filtered_bam_file=${file_prefix}Aligned.out.filtered.bam
samtools view -bS $sam_file_filtered > $filtered_bam_file

# assemble reads
salmon quant \
  -t /path/to/hervquant_reference/hervquant_final_reference.fa \
  -l ISF \
  -a $filtered_bam_file \
  -o $hERV_dir \
  -p $num_threads
```