Marmoset Pipeline

Reference Locations:
Marmoset files:
/storage/datasets/callithrix/fastq

Making the bowtie2 refererence:
Referenced from Bowtie2 Manual: http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml#the-bowtie2-build-indexer

Aligning fastq files to reference, calJac3.ref:
(This one line script was repeated for each NEX file. The script can be found in /work/jgjohns6/Marmoset/NEX/NEX_bt2.sh)
bowtie2 --rg-id NEX1_1 -x /work/jgjohns6/Marmoset/calJac3.ref -1 /storage/datasets/callithrix/fastq/NEX1_1_R1.fastq.gz -2 /storage/datasets/callithrix/fastq/NEX1_1_R2.fastq.gz -S NEX1_1.sam

Merging all sam files into one file:
samtools merge NEX1.sam NEX*.sam

Converting from Sam to bam:
samtools view -b -o NEX1.bam NEX1.sam

Filtering out the good reads, using -f4 for read unmapped:
samtools view -h -f4 NEX1.bam > NEX1_4.bam
#the -f option only outputs ALL bits, while -F does any that fall under the samflag, which is why it had to be more filtered

Converting from bam to fastq:
samtools bam2fq NEX1_4.bam > NEX1_4.fastq

Splitting a Single fastq file unto two:
cat NEX1_4.fastq | grep '^@.*/1$' -A 3 --no-group-separator > NEX1_4_r1.fastq

cat NEX1_4.fastq | grep '^@.*/2$' -A 3 --no-group-separator > NEX1_4_r2.fastq

#Running metaSPADES assembly:
#metaspades.py --pe1-1 NEX1_4_r1.fastq --pe1-2 NEX1_4_r2.fastq -o NEX_contigs_2
#Did not use metaSPADES we used megahit

Running megahit in parallel:
parallel -j 15 < /work/jgjohns6/Marmoset/megahit_commands.txt

In megahit_commands.txt: Megahit command:
/work/jgjohns6/megahit/megahit -1 NEX1_4_r1.fastq -2 NEX1_4_r2.fastq -o megahit_result

Blasting unmapped marmoset reads to viral ref seq database:
Split -l 1000 k119.contigs.fa
ls | grep '^x'parallel -j 15 "blastx -db refseq_protein -max_target_seqs 1 -outfmt '6 qacc sac c bitscore evalue sscinames scomnames stitle' -query {} -out .blastx"

Helpeful website:

https://github.com/ACHG2018/metagenomics-classification-tools/wiki/Bowtie,MetaSPAdes,BLAST-workflow