Vollenhovia emeryi excessive number of splice sites
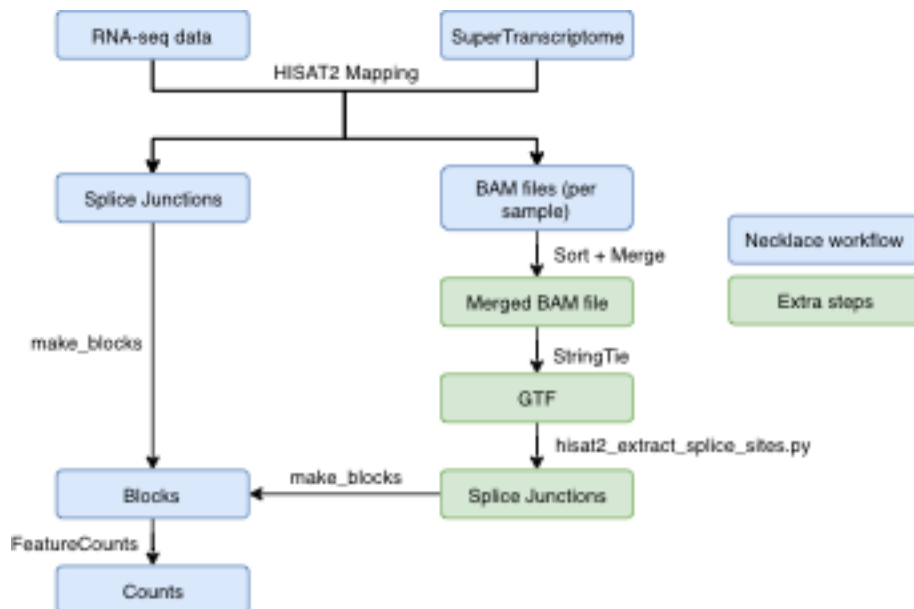
Merging bam files produced by HISAT2 in the later stages of Necklace. When mapping to ST.

1. Sort bam files
2. Merge bam files

Sort bam files

Use `sort_bam.sh`

```bash
#!/bin/bash

for longfilename in `ls $@`
do

filename=$(basename ${longfilename} .out.bam)

pathScript=$(pwd)/sort_bam_${filename}.sh

echo "#!/bin/bash" >${pathScript}
echo " " >>${pathScript}
echo "#SBATCH -p cloud" >>${pathScript}
echo "#SBATCH --time=09:00:00" >>${pathScript}
echo "#SBATCH --nodes=1" >>${pathScript}
echo "#SBATCH --ntasks=1" >>${pathScript}
echo "#SBATCH --cpus-per-task=1" >>${pathScript}
echo "#SBATCH --job-name=tsort_${filename}" >>${pathScript}
echo "#SBATCH --mem-per-cpu=10000" >>${pathScript}
echo "#SBATCH --mail-type=ALL" >>${pathScript}
echo "#SBATCH --mail-user=afarre@student.unimelb.edu.au" >>${pathScript}
echo "#SBATCH --out=slurm_%j.out" >>${pathScript}
echo "#SBATCH --err=slurm_%j.err" >>${pathScript}
echo " " >>${pathScript}
echo "input=${longfilename}" >>${pathScript}
echo "sorted=$(pwd)/${filename}_sorted.bam" >>${pathScript}
echo " " >>${pathScript}
echo "module load SAMtools" >>${pathScript}
echo " " >>${pathScript}
echo "samtools sort \${input} > \${sorted}" >>${pathScript}

done
```

Runtime and memory usage:

```
$ sacct -j 3568609 --format=JobID,JobName,MaxRSS,Elapsed
      JobID    JobName     MaxRSS    Elapsed
------------ ---------- ---------- ----------
3568609      tsort_DRR+               00:38:02
3568609.bat+     batch    890900K   00:38:02
3568609.ext+    extern       708K   00:38:02
```

Merge bam files

Use `mergebam.sh`

```
#!/bin/bash


#SBATCH -p cloud
#SBATCH --time=05:00:00
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=1
#SBATCH --job-name=mergebam
#SBATCH --mem=20GB
#SBATCH --out=slurm_%j.out
#SBATCH --err=slurm_%j.err
#SBATCH --mail-user=afarre@student.unimelb.edu.au
#SBATCH --mail-type=ALL

module load SAMtools

samtools merge STAR_adultandlarvae.sortedByCoord.out.bam $@
```

```
$ sacct -j 3569197 --format=JobID,JobName,MaxRSS,Elapsed
       JobID    JobName     MaxRSS    Elapsed
------------ ---------- ---------- ----------
3569197        mergebam               04:37:29
3569197.bat+      batch    13504K   04:37:29
3569197.ext+     extern      712K   04:37:29
```

Stringtie

Use StringTie in the merged bam file to try to reduce the number of blocks (probably Stringtie has some sort of filtering of juctions more strict that HISAT2's). Extract junctions from StringTie's output using HISAT2 script.

```bash
#!/bin/bash

#SBATCH -p physical
#SBATCH --time=09:00:00
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=3
#SBATCH --job-name=StringTie_${filename}
#SBATCH --mem-per-cpu=4000
#SBATCH --mail-type=ALL
#SBATCH --mail-user=afarre@student.unimelb.edu.au
#SBATCH --out=stringtie_%j.out
#SBATCH --err=stringtie_%j.err

module load StringTie

stringtie -v -o stringtie_merged_out_bam.gtf merged_out.bam -p 3
```

Runtime and memory used:

```
$ sacct -j 3587665 --format=JobID,JobName,MaxRSS,Elapsed
       JobID    JobName     MaxRSS    Elapsed
------------ ---------- ---------- ----------
3587665      StringTie+              00:56:15
3587665.bat+      batch   1255020K   00:56:15
3587665.ext+     extern       748K   00:56:15
```

Number of exons:

```
$ grep "exon" stringtie_merged_out_bam.gtf |wc -l
80807
```

Extract splice sites and make blocks

- Extract splice sites from the stringtie gtf.

```
$ hisat2_extract_splice_sites.py stringtie_merged_out_bam.gtf > stringtie_merged_bam
```

Can be run in sinteractive

- Then make blocks using Necklace's `make_blocks` function

```
$ /home/afarre/.local/necklace-necklace_v0.9/tools/bin/make_blocks /data/projects/pu
```

Can be run in sinteractive

Number of blocks:

```
$ grep "exon" stringtie_merged_bam_blocks.gtf | wc -l
71446
```

STATS

Species	#StringTie exons	#ST blocks	#Stringtie blocks
V.emeryi	277,505	999,514	71,446
W.auropunctata	254,989	172,634	
A.mellifera	291,739	247,922	
M.pharaonis	260,650	159,248	
Aechinatior	233,719	177,313	

FeatureCounts on the blocks