Erica Chio, Assigned coursework #3

Due date: March 2nd (two weeks), 11pm EST

Submission: PDF document by email to daniel.depledge@nyulangone.org

1. Three additional sequence datasets can be found in `/gpfs/data/courses/bminga3004/Practicum3/HW`

17V12256_R1.fastq.gz 17V12257_R1.fastq.gz 17V12258_R1.fastq.gz
17V12256_R2.fastq.gz 17V12257_R2.fastq.gz 17V12258_R2.fastq.gz

2. Perform QC (trimming) on each of these and align against the dumas.fasta genome using bbmap and bwa

bash specifications for all scripts:

```
#!/bin/bash
#SBATCH --job-name=<filename> # Job name
#SBATCH --mail-type=END,FAIL # Mail events (NONE, BEGIN, END, FAIL, ALL)
#SBATCH --mail-user=Erica.Chio@nyulangone.org # Where to send mail
#SBATCH --ntasks=1 # Run on a single CPU
#SBATCH --mem=16gb # Job memory request
#SBATCH --time=08:00:00 # Time limit hrs:min:sec
#SBATCH --output=<filename>_%j.log # Standard output and error log
#SBATCH -p cpu_short # Specifies location to submit job
```

trimgalore.sh

```
module purge

module load python/cpu/2.7.15-ES
module load trimgalore/0.5.0
module load fastqc/0.11.7

while read i;

do
echo $i
```

```
trim_galore --paired --length 30 -o /gpfs/scratch/ebc308/AIS/coursework3data/ /
gpfs/data/courses/bminga3004/Practicum3/HW/"$i"_R1.fastq.gz /gpfs/data/courses/
bminga3004/Practicum3/HW/"$i"_R2.fastq.gz

done < /gpfs/scratch/ebc308/AIS/coursework3data/datasets.txt
```

bbmap.sh

```
module load python/cpu/2.7.15-ES
module load bbmap/38.25
module load samtools/1.3
module load bedtools/2.26.0

while read i;

do
echo $i

### BBMAP

bbmap.sh ref=/gpfs/data/courses/bminga3004/Practicum3/dumas.fasta pairedonly=t
in=/gpfs/scratch/ebc308/AIS/coursework3data/"$i"_R1_val_1.fq.gz in2=/gpfs/
scratch/ebc308/AIS/coursework3data/"$i"_R2_val_2.fq.gz outm=/gpfs/scratch/
ebc308/AIS/coursework3data/"$i"_bbmap/mapping.sam nodisk ambiguous=random
sam=1.3

samtools view -bS -o /gpfs/scratch/ebc308/AIS/coursework3data/"$i"_bbmap/
mapping.bam /gpfs/scratch/ebc308/AIS/coursework3data/"$i"_bbmap/mapping.sam

samtools sort -o /gpfs/scratch/ebc308/AIS/coursework3data/"$i"_bbmap/
mapping.sorted.bam /gpfs/scratch/ebc308/AIS/coursework3data/"$i"_bbmap/
mapping.bam

samtools index /gpfs/scratch/ebc308/AIS/coursework3data/"$i"_bbmap/
mapping.sorted.bam
```

```
samtools view -b /gpfs/scratch/ebc308/AIS/coursework3data/"$i"_bbmap/
mapping.sorted.bam | genomeCoverageBed -ibam stdin -bg -g /gpfs/data/courses/
bminga3004/Practicum3/dumas.fasta > /gpfs/scratch/ebc308/AIS/
coursework3data/"$i"_bbmap/"$i".bedgraph

done < /gpfs/scratch/ebc308/AIS/coursework3data/datasets.txt
```

bwa.sh

```
module load python/cpu/2.7.15-ES
module load samtools/1.3
module load bedtools/2.26.0
module load bwa/0.7.17

### BWA

### build index - note you only need to do this once!
bwa index -a is /gpfs/scratch/ebc308/AIS/coursework3data/dumas.fasta

while read i;

do
echo $i

### align data
mkdir /gpfs/scratch/ebc308/AIS/coursework3data/"$i"_bwa

bwa aln /gpfs/scratch/ebc308/AIS/coursework3data/dumas.fasta /gpfs/scratch/
ebc308/AIS/coursework3data/"$i"_R1_val_1.fq.gz > /gpfs/scratch/ebc308/AIS/
coursework3data/"$i"_R1.sai

bwa aln /gpfs/scratch/ebc308/AIS/coursework3data/dumas.fasta /gpfs/scratch/
ebc308/AIS/coursework3data/"$i"_R2_val_2.fq.gz > /gpfs/scratch/ebc308/AIS/
coursework3data/"$i"_R2.sai

bwa sampe /gpfs/scratch/ebc308/AIS/coursework3data/dumas.fasta /gpfs/scratch/
ebc308/AIS/coursework3data/"$i"_R1.sai /gpfs/scratch/ebc308/AIS/
```

```
coursework3data/"$i"_R2.sai /gpfs/scratch/ebc308/AIS/
coursework3data/"$i"_R1_val_1.fq.gz /gpfs/scratch/ebc308/AIS/
coursework3data/"$i"_R2_val_2.fq.gz > /gpfs/scratch/ebc308/AIS/
coursework3data/"$i"_bwa/mapping.sam

samtools view -bS -o /gpfs/scratch/ebc308/AIS/coursework3data/"$i"_bwa/
mapping.bam /gpfs/scratch/ebc308/AIS/coursework3data/"$i"_bwa/mapping.sam

samtools sort -o /gpfs/scratch/ebc308/AIS/coursework3data/"$i"_bwa/
mapping.sorted.bam /gpfs/scratch/ebc308/AIS/coursework3data/"$i"_bwa/
mapping.bam

samtools index /gpfs/scratch/ebc308/AIS/coursework3data/"$i"_bwa/
mapping.sorted.bam

samtools view -b /gpfs/scratch/ebc308/AIS/coursework3data/"$i"_bwa/
mapping.sorted.bam | genomeCoverageBed -ibam stdin -bg -split -g /gpfs/data/
courses/bminga3004/Practicum3/dumas.fasta > /gpfs/scratch/ebc308/AIS/
coursework3data/"$i"_bwa/"$i".bedgraph

done < /gpfs/scratch/ebc308/AIS/coursework3data/datasets.txt
```

3. For each dataset, and with each aligner:

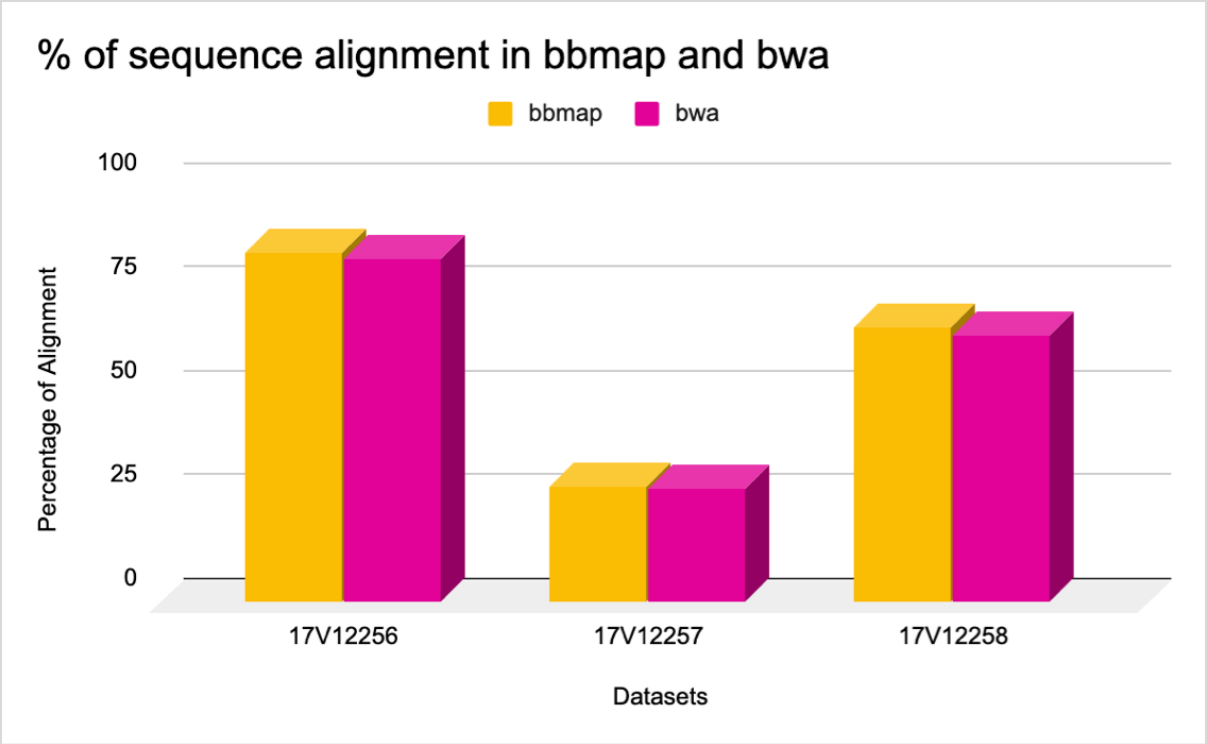
- what % of sequence reads align against Dumas?

`samtools flagstat mapping.sorted.bam` for bwa

I looked inside the log file for bbmap to get the % aligned. (mapped reads)

Dataset	bbmap	bwa
17V12256	84.35%	82.79%
17V12257	27.87%	27.29%
17V12258	66.10%	64.42%

- Produce a simple bar chart to show this.



4. Briefly explain why there are differences in the range of alignment flags present in bbmap and bwa mapped datasets

I ran `cut -f2 mapping.sam | sort | uniq -c` to get the flags and how many reads each flags got. In general, bwa has many more different flags than bbmap.

Generally, if bwa has 113, bbmap will have 115. If bwa has 129, bbmap has 131. bbmap's flag is two off bwa, and difference in these flags is that there is a "read mapped in proper pair." This could be because bbmap has the flag, "pairedonly = t," meaning there are more paired maps.

bbmap is also a global alignment and bwa is a local alignment. This would mean that bwa has more flags because it is reading it locally, leading to many more smaller alignments than a generally larger one.

5. Produce a single GVIZ coverage plot showing all datasets (and both aligners) incorporated – is the shape of the coverage similar between datasets?

```
library(data.table)
library(Gviz)
```

```
library(GenomicFeatures)
setwd("/Users/ericachio/Documents/sackler/applied informatics sequencing/
coursework3")

myChr = "dumas.fasta"

myStart = 1
myEnd = 125000

### READ IN GENE MODELS ###
gtrack<-GenomeAxisTrack(col="black") ##Adds genome axis

modelsPos<-makeTxDbFromGFF("VZV-Dumas-Forward-2.0.gff3")
rtrackFor <- GeneRegionTrack(modelsPos, genome = "VZV", chromosome =
"dumas.fasta", name = "Gene Model", col="black", fill="light blue",
stacking="squish", shape="smallArrow", background.title = "transparent",
options(ucscChromosomeNames=FALSE)) #squish #dense

modelsRev<-makeTxDbFromGFF("VZV-Dumas-Reverse-2.0.gff3")
rtrackRev <- GeneRegionTrack(modelsRev, genome = "VZV", chromosome =
"dumas.fasta", name = "Gene Model", col="black", fill="light blue",
stacking="squish", shape="smallArrow", background.title = "transparent",
options(ucscChromosomeNames=FALSE)) #squish #dense
```

17V12256

```

### BEDGRAPH DATA
file1 <- fread('./17V12256_bbmap.bedgraph', col.names = c('chromosome',
'start', 'end', 'value'))
file2 <- fread('./17V12256_bwa.bedgraph', col.names = c('chromosome', 'start',
'end', 'value'))

file1<-file1[file1$start>myStart]
file1<-file1[file1$end<myEnd,]
max1<-max(file1$value)

file2<-file2[file2$start>myStart]
file2<-file2[file2$end<myEnd,]
```

```
max2<-max(file2$value)

dataTrack1 <- DataTrack(range = file1, type = "a", chromosome=myChr, genome =
'VZV', fill = "gold1", col = "gold1",
options(ucscChromosomeNames=FALSE),col.axis="black", background.title =
"transparent", ylim=c(0,max1))

dataTrack2 <- DataTrack(range = file2, type = "a", chromosome=myChr, genome =
'VZV', fill = "deeppink2", col = "deeppink2",
options(ucscChromosomeNames=FALSE),col.axis="black", background.title =
"transparent", ylim=c(0,max2))
```

17V12257

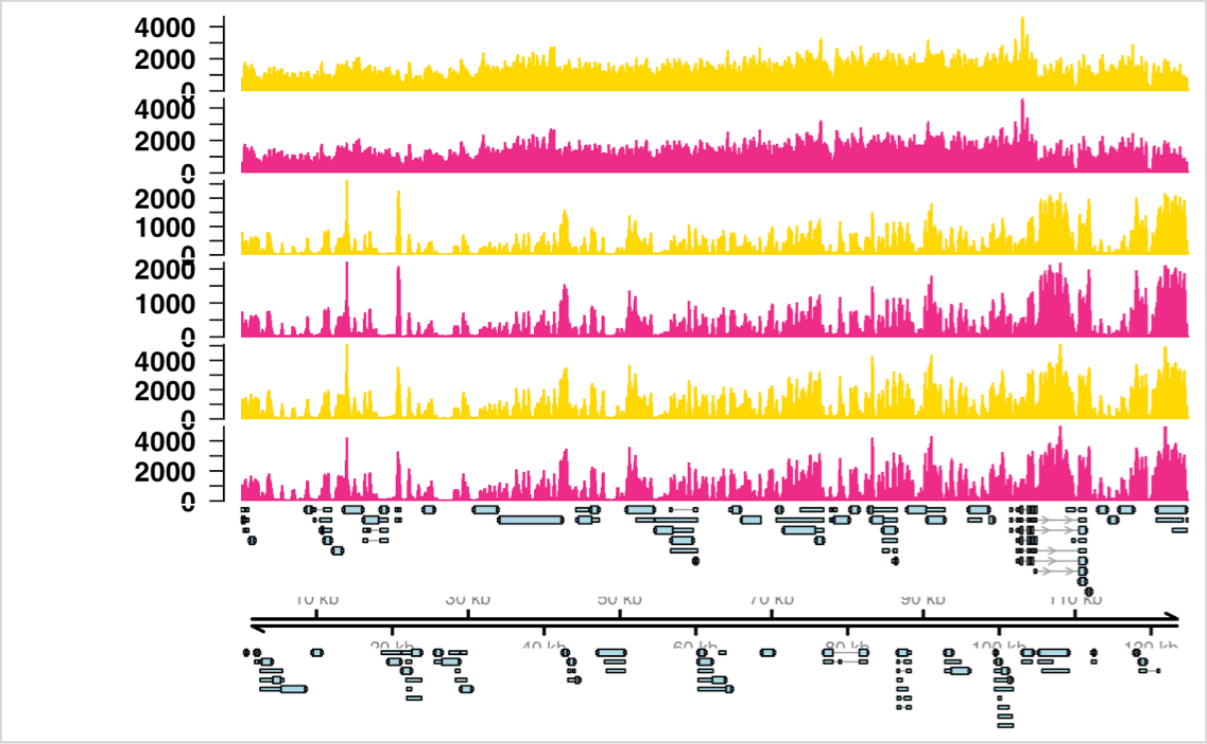
```
### BEDGRAPH DATA
file3 <- fread('./17V12257_bbmap.bedgraph', col.names = c('chromosome',
'start', 'end', 'value'))
file4 <- fread('./17V12257_bwa.bedgraph', col.names = c('chromosome', 'start',
'end', 'value'))

file3<-file3[file3$start>myStart]
file3<-file3[file3$end<myEnd,]
max1<-max(file3$value)

file4<-file4[file4$start>myStart]
file4<-file4[file4$end<myEnd,]
max2<-max(file4$value)

dataTrack3 <- DataTrack(range = file3, type = "a", chromosome=myChr, genome =
'VZV', fill = "gold1", col = "gold1",
options(ucscChromosomeNames=FALSE),col.axis="black", background.title =
"transparent", ylim=c(0,max1))

dataTrack4 <- DataTrack(range = file4, type = "a", chromosome=myChr, genome =
'VZV', fill = "deeppink2", col = "deeppink2",
options(ucscChromosomeNames=FALSE),col.axis="black", background.title =
"transparent", ylim=c(0,max2))
```

17V12258

```
### BEDGRAPH DATA
file5 <- fread('./17V12258_bbmap.bedgraph', col.names = c('chromosome',
'start', 'end', 'value'))
file6 <- fread('./17V12258_bwa.bedgraph', col.names = c('chromosome', 'start',
'end', 'value'))

file5<-file5[file5$start>myStart]
file5<-file5[file5$end<myEnd,]
max1<-max(file5$value)

file6<-file6[file6$start>myStart]
file6<-file6[file6$end<myEnd,]
max2<-max(file6$value)

dataTrack5 <- DataTrack(range = file5, type = "a", chromosome=myChr, genome =
'VZV', fill = "gold1", col = "gold1",
options(ucscChromosomeNames=FALSE),col.axis="black", background.title =
"transparent", ylim=c(0,max1))

dataTrack6 <- DataTrack(range = file6, type = "a", chromosome=myChr, genome =
'VZV', fill = "deeppink2", col = "deeppink2",
options(ucscChromosomeNames=FALSE),col.axis="black", background.title =
"transparent", ylim=c(0,max2))
```

GENERATE PLOT

```
plotTracks(list(dataTrack1, dataTrack2, dataTrack3, dataTrack4,
dataTrack5,dataTrack6,rtrackFor,gtrack,rtrackRev), from = myStart, to = myEnd,
sizes=c(0.16,0.16,0.16,0.16,0.16,0.16,0.18,0.1,0.16), type="hist",
col.histogram=NA, cex.title=1, cex.axis=1, title.width=1.2, collapse=FALSE)
```



The gold is for bbmap and the pink is for the bwa coverage.

The first dataset (17V12256)has the most coverage out of the three datasets. I think that the second (17V12257) and third (17V12258) dataset look pretty similar.

Provide short write-up with figures and full coding used (command line and R code)

Extra Credit: Repeat analysis with duplicate sequences identified and removed. How does this change the coverage plots and % aligned against dumas

bwa

```
module purge

module load picard-tools/2.18.20
module load samtools/1.3
module load bedtools/2.26.0

while read i;

do
echo $i
```

```
java -jar /gpfs/share/apps/picard/2.18.11/libs/picard.jar MarkDuplicates
INPUT=/gpfs/scratch/ebc308/AIS/coursework3data/"$i"_bwa/mapping.sorted.bam
OUTPUT=/gpfs/scratch/ebc308/AIS/coursework3data/extracredit/"$i"_bwa/
mapping.nodups.sorted.bam METRICS_FILE=/gpfs/scratch/ebc308/AIS/
coursework3data/extracredit/"$i"_bwa/"$i"_DuplicateLogBWA.txt
REMOVE_DUPLICATES=TRUE TMP_DIR=/gpfs/scratch/ebc308/AIS/coursework3data/
extracredit/tmp AS=true VALIDATION_STRINGENCY=LENIENT

samtools sort -o /gpfs/scratch/ebc308/AIS/coursework3data/extracredit/"$i"_bwa/
mapping.nodups.sorted.sorted.bam /gpfs/scratch/ebc308/AIS/coursework3data/
extracredit/"$i"_bwa/mapping.nodups.sorted.bam

samtools index /gpfs/scratch/ebc308/AIS/coursework3data/extracredit/"$i"_bwa/
mapping.nodups.sorted.sorted.bam

samtools view -b /gpfs/scratch/ebc308/AIS/coursework3data/extracredit/"$i"_bwa/
mapping.nodups.sorted.sorted.bam | genomeCoverageBed -ibam stdin -bg -split -
g /gpfs/data/courses/bminga3004/Practicum3/dumas.fasta > /gpfs/scratch/ebc308/
AIS/coursework3data/extracredit/"$i"_bwa/"$i"_bwa_extracredit.bedgraph

done < /gpfs/scratch/ebc308/AIS/coursework3data/datasets.txt
```

To compare how many reads aligned to dumas after duplicates were removed, I ran `samtools flagstat mapping.nodups.sorted.sorted.bam`

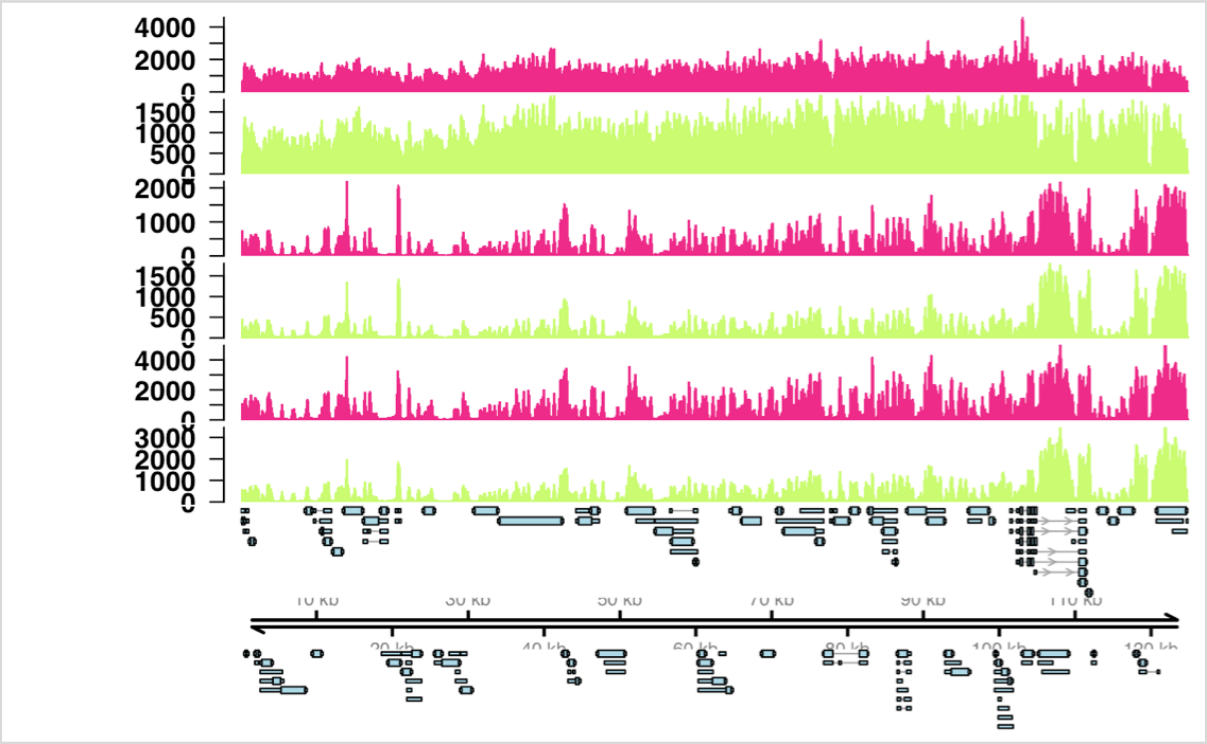
To get the percentage of reads retained after duplicates were removed, I looked at the number of reads mapped in the bam file prior to removing duplicates and the number of reads after. I did (# of reads **after** duplicates removed) / (# of reads mapped **before** duplicates removed)

Dataset	bwa % reads aligned to dumas	% reads aligned to dumas after duplicates removed	% reads retained after duplicates removed
""			
17V12256	82.79%	79.23%	79.27%
17V12257	27.29%	20.97%	70.70%
17V12258	64.42%	50.03%	55.29%

The alignment actually decreased as duplicates were removed, meaning duplicates were probably aligned multiple times.

```
### BEDGRAPH DATA
file7 <- fread('./17V12256_bwa_extracredit.bedgraph', col.names =
c('chromosome', 'start', 'end', 'value'))
file8 <- fread('./17V12257_bwa_extracredit.bedgraph', col.names =
c('chromosome', 'start', 'end', 'value'))
file9 <- fread('./17V12258_bwa_extracredit.bedgraph', col.names =
c('chromosome', 'start', 'end', 'value'))

file7<-file7[file7$start>myStart]
file7<-file7[file7$end<myEnd,]
max1<-max(file7$value)

file8<-file8[file8$start>myStart]
file8<-file8[file8$end<myEnd,]
max1<-max(file8$value)

file9<-file9[file9$start>myStart]
file9<-file9[file9$end<myEnd,]
max2<-max(file9$value)

dataTrack7 <- DataTrack(range = file7, type = "a", chromosome=myChr, genome =
'VZV', fill = "darkolivegreen1", col = "darkolivegreen1",
options(ucscChromosomeNames=FALSE),col.axis="black", background.title =
"transparent", ylim=c(0,max1))

dataTrack8 <- DataTrack(range = file8, type = "a", chromosome=myChr, genome =
'VZV', fill = "darkolivegreen1", col = "darkolivegreen1",
options(ucscChromosomeNames=FALSE),col.axis="black", background.title =
"transparent", ylim=c(0,max1))

dataTrack9 <- DataTrack(range = file9, type = "a", chromosome=myChr, genome =
'VZV', fill = "darkolivegreen1", col = "darkolivegreen1",
options(ucscChromosomeNames=FALSE),col.axis="black", background.title =
"transparent", ylim=c(0,max2))

### GENERATE PLOT ###
plotTracks(list(dataTrack2, dataTrack7, dataTrack4, dataTrack8,
dataTrack6,dataTrack9,rtrackFor,gtrack,rtrackRev), from = myStart, to = myEnd,
```

```
sizes=c(0.16,0.16,0.16,0.16,0.16,0.16,0.18,0.1,0.16), type="hist",
col.histogram=NA, cex.title=1, cex.axis=1, title.width=1.2, collapse=FALSE)
```



pink is original bwa alignment

green is duplicates removed alignment

bbmap

```
module purge

module load picard-tools/2.18.20
module load samtools/1.3
module load bedtools/2.26.0

while read i;

do
echo $i

java -jar /gpfs/share/apps/picard/2.18.11/libs/picard.jar MarkDuplicates
INPUT=/gpfs/scratch/ebc308/AIS/coursework3data/"$i"_bbmap/mapping.sorted.bam
OUTPUT=/gpfs/scratch/ebc308/AIS/coursework3data/extracredit/"$i"_bbmap/
```

```
mapping.nodups.sorted.bam METRICS_FILE=/gpfs/scratch/ebc308/AIS/
coursework3data/extracredit/"$i"_bbmap/"$i"_DuplicateLogBBmap.txt
REMOVE_DUPLICATES=TRUE TMP_DIR=/gpfs/scratch/ebc308/AIS/coursework3data/
extracredit/tmp AS=true VALIDATION_STRINGENCY=LENIENT

samtools sort -o /gpfs/scratch/ebc308/AIS/coursework3data/
extracredit/"$i"_bbmap/mapping.nodups.sorted.sorted.bam /gpfs/scratch/ebc308/
AIS/coursework3data/extracredit/"$i"_bbmap/mapping.nodups.sorted.bam

samtools index /gpfs/scratch/ebc308/AIS/coursework3data/extracredit/"$i"_bbmap/
mapping.nodups.sorted.sorted.bam

samtools view -b /gpfs/scratch/ebc308/AIS/coursework3data/
extracredit/"$i"_bbmap/mapping.nodups.sorted.sorted.bam | genomeCoverageBed -
ibam stdin -bg -split -g /gpfs/data/courses/bminga3004/Practicum3/dumas.fasta
> /gpfs/scratch/ebc308/AIS/coursework3data/
extracredit/"$i"_bbmap/"$i"_bbmap_extracredit.bedgraph



done < /gpfs/scratch/ebc308/AIS/coursework3data/datasets.txt
```

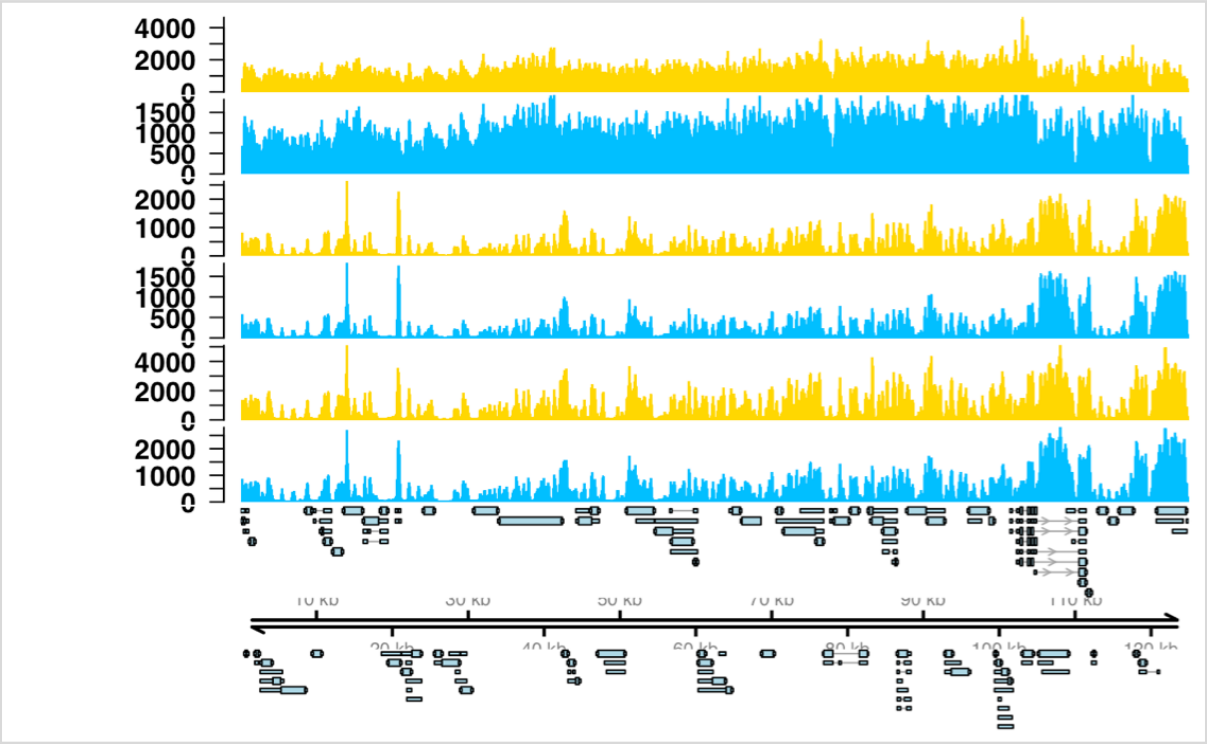
To compare how many reads aligned to dumas after duplicates were removed, I looked at the original bbmap log to get total number of reads. I then ran `samtools flagstat mapping.nodups.sorted.sorted.bam` to get the number of reads aligned after duplicates were removed. I did (# of reads **after** duplicates removed) / (# of total number of reads) To get the percentage of reads retained after duplicates were removed, I looked at the number of reads mapped in the bam file prior to removing duplicates and the number of reads after. I did (# of reads **after** duplicates removed) / (# of reads mapped **before** duplicates removed)

Dataset	bbmap % reads aligned to dumas	% reads aligned to dumas after duplicates removed	% reads retained after duplicates removed
""			
17V12256	84.35%	67.48%	79.99%
17V12257	27.87%	19.26%	69.10%
17V12258	66.10%	34.95%	52.88%

Alignment to dumas decreased, as duplicates were probably aligned multiple times.

```
### BEDGRAPH DATA
file10 <- fread('./17V12256_bbmap_extracredit.bedgraph', col.names =
c('chromosome', 'start', 'end', 'value'))
file11 <- fread('./17V12257_bbmap_extracredit.bedgraph', col.names =
c('chromosome', 'start', 'end', 'value'))
file12 <- fread('./17V12258_bbmap_extracredit.bedgraph', col.names =
c('chromosome', 'start', 'end', 'value'))

file10<-file10[file10$start>myStart]
file10<-file10[file10$end<myEnd,]
max1<-max(file10$value)

file11<-file11[file11$start>myStart]
file11<-file11[file11$end<myEnd,]
max1<-max(file11$value)

file12<-file12[file12$start>myStart]
file12<-file12[file12$end<myEnd,]
max2<-max(file12$value)

dataTrack10 <- DataTrack(range = file10, type = "a", chromosome=myChr, genome =
'VZV', fill = "deepskyblue1", col = "deepskyblue1",
options(ucscChromosomeNames=FALSE),col.axis="black", background.title =
"transparent", ylim=c(0,max1))

dataTrack11 <- DataTrack(range = file11, type = "a", chromosome=myChr, genome =
'VZV', fill = "deepskyblue1", col = "deepskyblue1",
options(ucscChromosomeNames=FALSE),col.axis="black", background.title =
"transparent", ylim=c(0,max1))

dataTrack12 <- DataTrack(range = file12, type = "a", chromosome=myChr, genome =
'VZV', fill = "deepskyblue1", col = "deepskyblue1",
options(ucscChromosomeNames=FALSE),col.axis="black", background.title =
"transparent", ylim=c(0,max2))

### GENERATE PLOT ###
plotTracks(list(dataTrack1, dataTrack10, dataTrack3, dataTrack11,
dataTrack5,dataTrack12,rtrackFor,gtrack,rtrackRev), from = myStart, to = myEnd,
```

```
sizes=c(0.16,0.16,0.16,0.16,0.16,0.16,0.18,0.1,0.16), type="hist",
col.histogram=NA, cex.title=1, cex.axis=1, title.width=1.2, collapse=FALSE)
```



gold is bbmap alignment

blue is duplicates removed alignment

#appliedsequencinginformatics