Starling-May18
Projects/Katarina Stuart/KStuart.Starling-Aug18/At1_Genome/Analysis/2023-01-18.SizeGenome

PDF Version generated by

Katarina Stuart (z5188231@ad.unsw.edu.au)

on

Aug 23, 2023 @09:47 AM NZST

Table of Contents

Katarina Stuart (z5188231@ad.unsw.edu.au) - Feb 15, 2023, 11:09 AM GMT+13

Genome Size

Jellyfish

https://bioinformatics.uconn.edu/genome-size-estimation-tutorial/

```
#!/bin/bash -e

#SBATCH --job-name=2023_02_14.genomesize_jellyfish.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-04:00:00
#SBATCH --mem=50GB
#SBATCH --output=%x_%j.errout
#SBATCH --mail-user=katarina.stuart@auckland.ac.nz
#SBATCH --mail-type=ALL
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=4
#SBATCH --profile task
#SBATCH --array=1-4

module load Jellyfish/2.3.0-gimkl-2020a

KMER=$(sed "${SLURM_ARRAY_TASK_ID}q;d" /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/genomesize/kmer_list.txt)
echo "working with kemr:" $KMER

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/genomesize

R1=/nesi/nobackup/uoa02613/Myna_ONT_2022/Myna_ONT_RAW_dec/Basecalling_guppy6.2.1/pilon_polishing/Myna_10x_processed_R1_val_1.fq.gz
R2=/nesi/nobackup/uoa02613/Myna_ONT_2022/Myna_ONT_RAW_dec/Basecalling_guppy6.2.1/pilon_polishing/Myna_10x_processed_R2_val_2.fq.gz

jellyfish count -t 16 -C -m ${KMER} -s 10G -o ${KMER}mer_out --min-qual-char=? <(zcat $R1) <(zcat $R2)
jellyfish histo -o ${KMER}mer_out.histo ${KMER}mer_out
```

genome size calculations from histogram

```
module load R/4.2.1-gimkl-2022a
R

setwd("/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/genomesize")

dataframe20 <- read.table("20mer_out.histo")

pdf("histo2.pdf")
plot(dataframe20[3:200,], type="l", xlab="K-mer Length", ylab="K-mer Count")
abline(v = 31, col="red", lwd=3)
abline(v = 7, col="blue", lwd=3)
dev.off()
```

Max histo:

31: 36167636

```
sum(as.numeric(dataframe20[7:9999,1]*dataframe20[7:9999,2]))
```

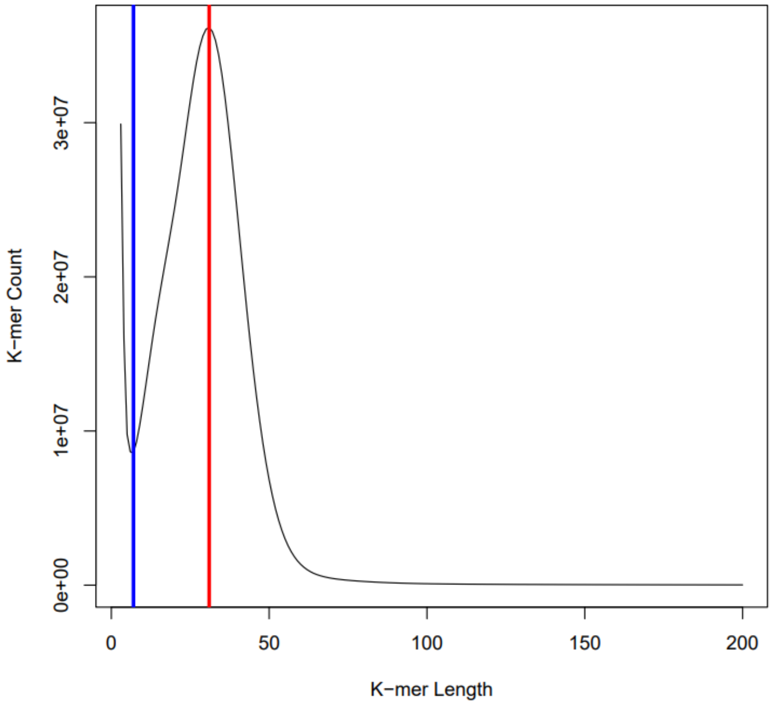
36,028,219,169

```
sum(as.numeric(dataframe20[7:9999,1]*dataframe20[7:9999,2]))/31
```

Est. Size: 1,162,200,618

1040603622 /1162200618

= 0.8953735% complete



Some additional numbers from the other K-mer values:

```
module load R/4.2.1-gimkl-2022a
R

setwd("/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/genomesize")


dataframe18 <- read.table("18mer_out.histo")
sum(as.numeric(dataframe18[7:9999,1]*dataframe18[7:9999,2]))/33 #1142465245

dataframe19 <- read.table("19mer_out.histo")
sum(as.numeric(dataframe19[7:9999,1]*dataframe19[7:9999,2]))/32 #1150752569

dataframe21 <- read.table("21mer_out.histo")
sum(as.numeric(dataframe21[6:9999,1]*dataframe21[6:9999,2]))/30 #1177818515

dataframe22 <- read.table("22mer_out.histo")
sum(as.numeric(dataframe18[7:9999,1]*dataframe18[7:9999,2]))/29 #1300046659
```

Merqury QV assessment

https://au-mynotebook.labarchives.com/share/BABS%2520Genome/NDk1LjN8MjkyMjMvMzgxL1RyZWVOb2RlLzExODM5MTI5MTB8MTI1Ny4z

Merqury uses Meryl to make kmer databases and then compares kmer profiles from genome assemblies with those from raw short read data.

install

```
conda create -n merqury -c conda-forge -c bioconda merqury openjdk=11
```

```
 environment variable 'MERQURY' set to /nesi/nobackup/uoa02613/kstuart_projects/programs/miniconda/envs/mamba/envs/merqury/share/merqury

# To activate this environment, use
#     $ conda activate merqury
# To deactivate an active environment, use
#     $ conda deactivate
```

```
module load Miniconda3
CONDA_BASE=$(conda info --base)
source ${CONDA_BASE}/etc/profile.d/conda.sh
c/nesi/nobackup/uoa02613/kstuart_projects/programs/miniconda/envs/mcclintock
conda create -n merqury -c conda-forge -c bioconda merqury openjdk=11


conda create --prefix /nesi/nobackup/uoa02613/kstuart_projects/programs/miniconda/envs/mamba/envs/merqury/share/merqury

conda activate /nesi/nobackup/uoa02613/kstuart_projects/programs/miniconda/envs/mamba/envs/merqury/share/merqury
```

```
 environment variable 'MERQURY' set to /nesi/nobackup/uoa02613/kstuart_projects/programs/miniconda/envs/mamba/envs/merqury/share/merqury

# To activate this environment, use
#     $ conda activate merqury
# To deactivate an active environment, use
#     $ conda deactivate
```

setting up meryl database

First, select the best k based on the genome size (21 is good for most).

```
cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/genomesize/merqury
MERQURY=/nesi/nobackup/uoa02613/kstuart_projects/programs/miniconda/envs/mamba/envs/merqury/share/merqury
GSIZE=1.040e9
${MERQURY}/best_k.sh $GSIZE
```

```
genome: 1.040e9
tolerable collision rate: 0.001
19.9591
```

```
k=20
```

Next, set up a variable pointing to the kmer read files and generate individual meryl databases, e.g.: need 24 hrs on the below settings(but actually ran in 15 mins???)

```
#!/bin/bash -e

#SBATCH --job-name=2023_02_08.genomesize_meryl.sl
```

```
#SBATCH --account=uoa02613
#SBATCH --time=00-12:00:00
#SBATCH --mem=5GB
#SBATCH --output=%x_%j.errout
#SBATCH --mail-user=katarina.stuart@auckland.ac.nz
#SBATCH --mail-type=ALL
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=2
#SBATCH --profile task

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/genomesize/merqury

module load Miniconda3
CONDA_BASE=$(conda info --base)
source ${CONDA_BASE}/etc/profile.d/conda.sh
conda activate /nesi/nobackup/uoa02613/kstuart_projects/programs/miniconda/envs/mamba/envs/merqury

k=20
KMERREADS="/nesi/nobackup/uoa02613/Myna_ONT_2022/Myna_ONT_RAW_dec/Basecalling_guppy6.2.1/pilon_polishing/Myna_10x_processed_R*_val_*.fq.gz"

# Build meryl dbs
for FASTQ in $KMERREADS; do
meryl k=$k count output $(basename ${FASTQ} .fq.gz).meryl $FASTQ
done

MERYLPARTS="*.meryl"
MERYLDB=myna.10x.meryl
meryl union-sum output $MERYLDB $MERYLPARTS
```

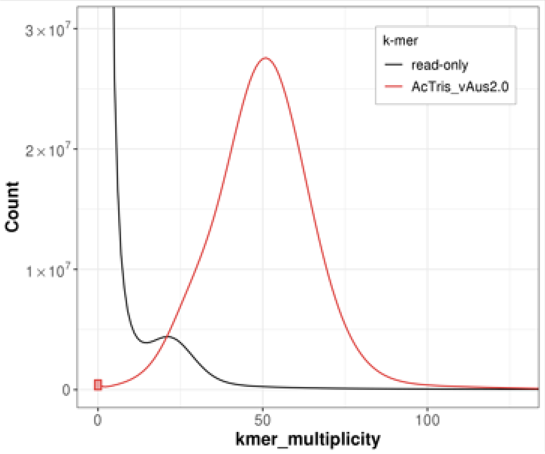
Running Merqury

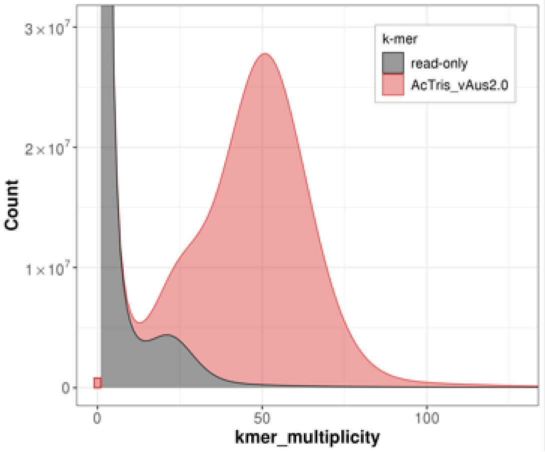
NOTE: Merqury is very fussy about file paths and wants everything local. It seems to work best with things in the current directory for a single run:

Run Merqury on each assembly:

```
#!/bin/bash -e

#SBATCH --job-name=2023_02_10.genomesize_merqury.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-12:00:00
#SBATCH --mem=15GB
#SBATCH --output=%x_%j.errout
#SBATCH --mail-user=katarina.stuart@auckland.ac.nz
#SBATCH --mail-type=ALL
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=16
#SBATCH --profile task


cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/genomesize/merqury

module load Miniconda3
CONDA_BASE=$(conda info --base)
source ${CONDA_BASE}/etc/profile.d/conda.sh
conda activate /nesi/nobackup/uoa02613/kstuart_projects/programs/miniconda/envs/mamba/envs/merqury

module load R/4.2.1-gimkl-2022a #also made sure to install.packages("argparse")


GENOME=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step4_scaffolding/ragtag_atris_synteny/renamed/AcTris_vAus2.0.fasta
MERQURY=/nesi/nobackup/uoa02613/kstuart_projects/programs/miniconda/envs/mamba/envs/merqury/share/merqury
MERYLDB=myna.10x.meryl

$MERQURY/merqury.sh $MERYLDB $GENOME AcTris_vAus
```

```
#$MERQURY/eval/spectra-cn.sh $MERYLDB $GENOME $(basename $GENOME .fasta)
```





```
AcTris_vAus2.0.completeness.stats:
AcTris_vAus2.0  all     972627157       1066771063      91.1749
```