UNIVERSITY OF GEORGIA
Integrated Life Sciences

Running Whole Animal Genome Sequencing (WAGS) Pipeline in UGA GACRC Cluster

Sachin Subedi (PhD student)

Integrated Life Sciences, University of Georgia

October 25, 2023

Project Outline

- Utilizing the pipeline provided by our collaborators, we are able to generate GVCF outputs.

- For alignment of the raw data, we've selected the CanFam4 reference genome, a well-established reference for canine genetic research.

- This data was then systematically aligned to CanFam4 using the WAGS pipeline. The result of this alignment is a GVCF file, which we then integrate with other GVCF files.

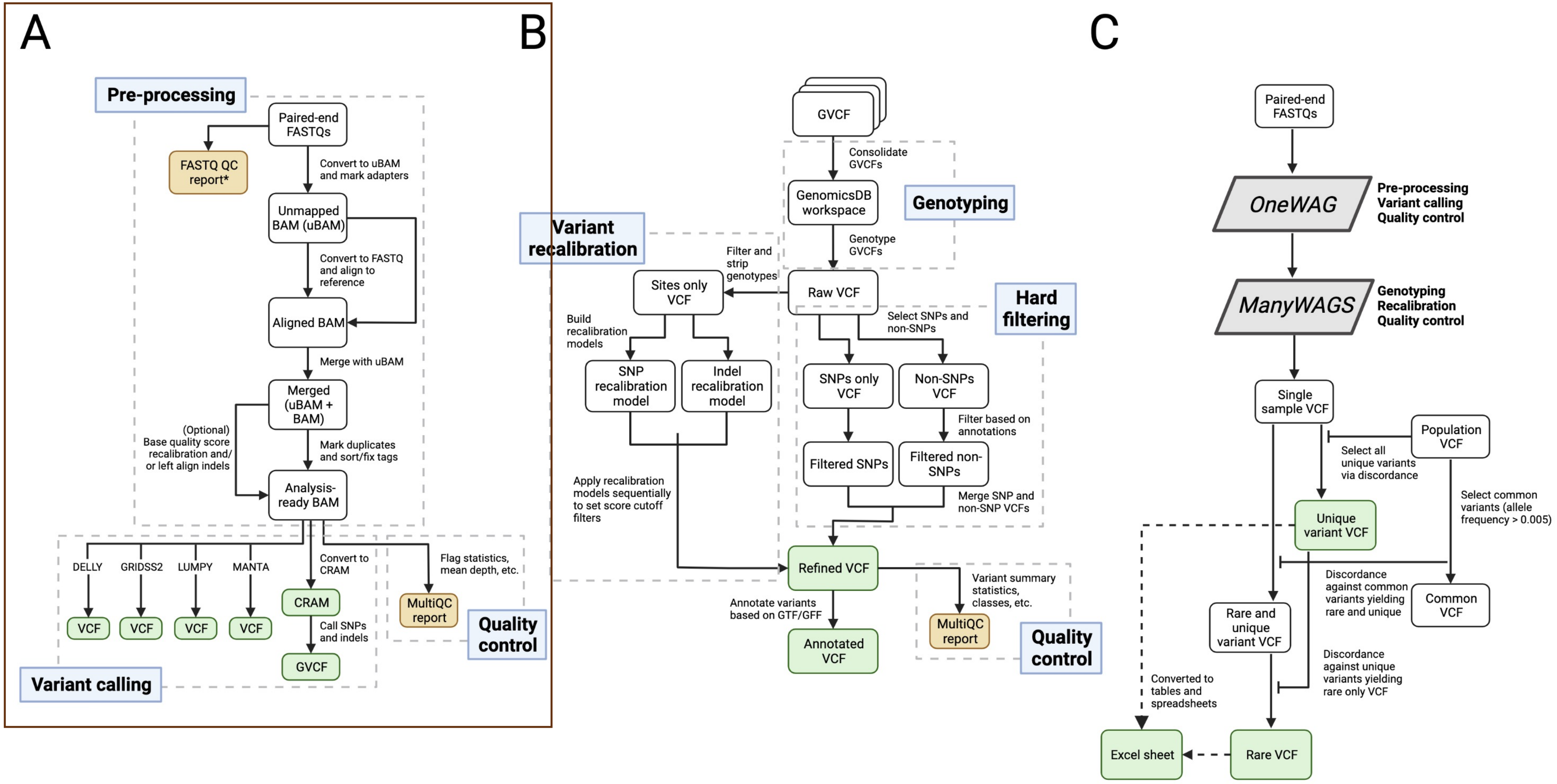
- I have procured raw genomic data for both Collies and Shetland Sheepdogs (often referred to as Shelties) from the Sequence Read Archive (SRA).

Project Objective

- Our objective is to establish a comprehensive reference panel comprising genomes from all accessible Collie and Sheltie samples.

- This constructed panel will facilitate the imputation of data from low-pass sequencing or SNP arrays, aligning it with the whole genome sequence.

Project Pipeline



A

Pre-processing

Paired-end FASTQs

FASTQ QC report*

Convert to uBAM and mark adapters

Unmapped BAM (uBAM)

Convert to FASTQ and align to reference

Aligned BAM

Merge with uBAM

Merged (uBAM + BAM)

(Optional) Base quality score recalibration and/ or left align indels

Mark duplicates and sort/fix tags

Analysis-ready BAM

DELLY

GRIDSS2

LUMPY

MANTA

VCF

VCF

VCF

VCF

Convert to CRAM

CRAM

Call SNPs and indels

GVCF

Flag statistics, mean depth, etc.

MultiQC report

Quality control

Variant calling

B

GVCF

Consolidate GVCFs

GenomicsDB workspace

Genotyping

Genotype GVCFs

Raw VCF

Variant recalibration

Filter and strip genotypes

Sites only VCF

Build recalibration models

SNP recalibration model

Indel recalibration model

Apply recalibration models sequentially to set score cutoff filters

Hard filtering

Select SNPs and non-SNPs

SNPs only VCF

Non-SNPs VCF

Filter based on annotations

Filtered SNPs

Filtered non-SNPs

Merge SNP and non-SNP VCFs

Refined VCF

Annotate variants based on GTF/GFF

Annotated VCF

Variant summary statistics, classes, etc.

MultiQC report

Quality control

C

Paired-end FASTQs

OneWAG

Pre-processing
Variant calling
Quality control

ManyWAGS

Genotyping
Recalibration
Quality control

Single sample VCF

Population VCF

Select all unique variants via discordance

Select common variants (allele frequency > 0.005)

Unique variant VCF

Discordance against common variants yielding rare and unique

Common VCF

Rare and unique variant VCF

Discordance against unique variants yielding rare only VCF

Converted to tables and spreadsheets

Excel sheet

Rare VCF

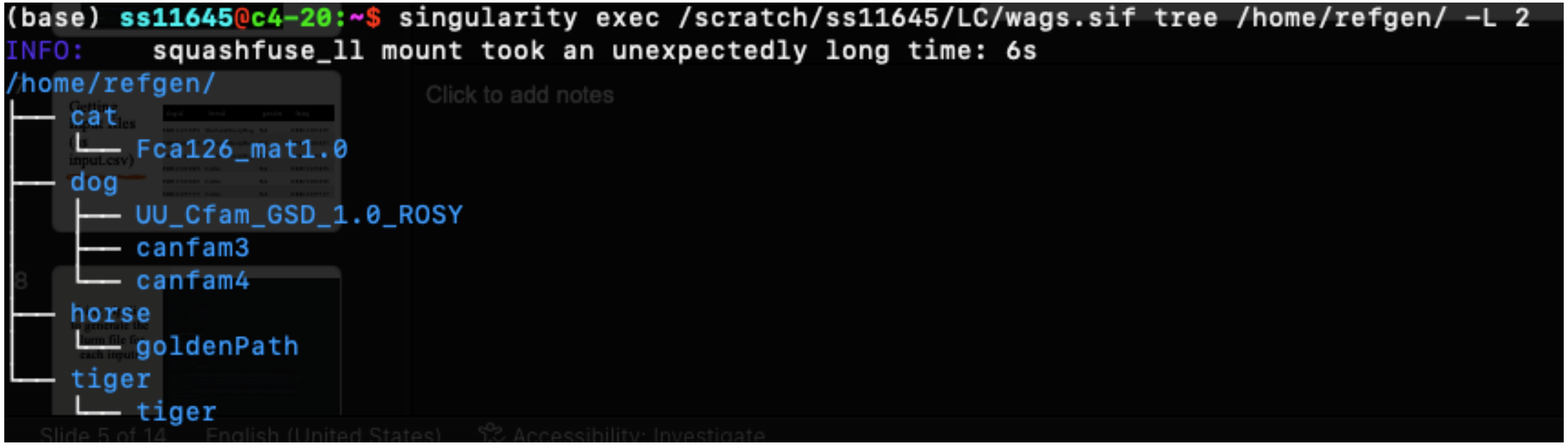
Implementation

Installing dependencies:

- Python
- Mamba or Conda
- Snakemake
- Snakemake-Profiles
- Miscellaneous Python modules pyaml, wget, and xlsxwriter
- Apptainer/Singularity

Downloading container that has reference genome:

wget https://s3.msi.umn.edu/wags/wags.sif



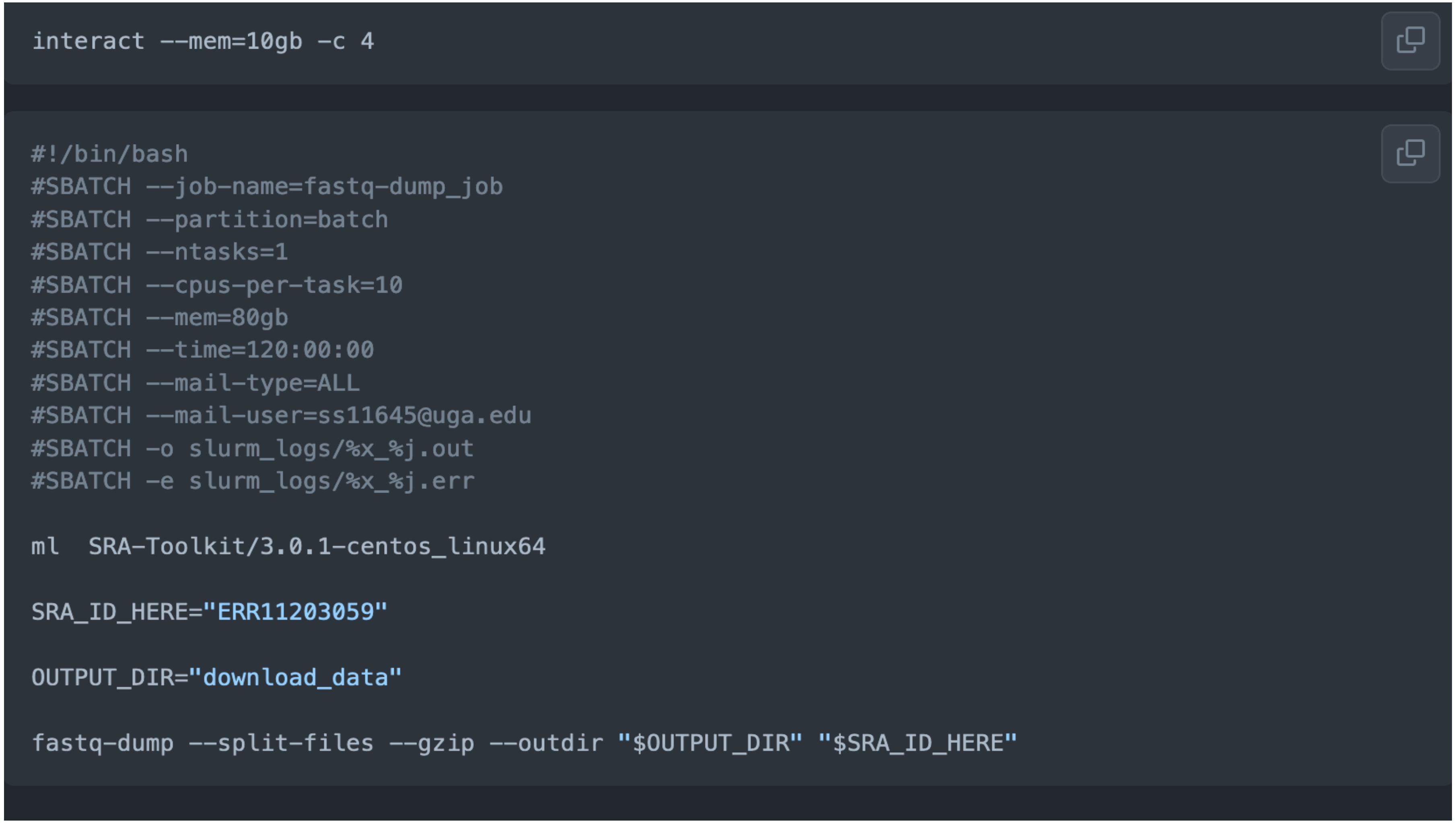
```
(base) ss11645@c4-20:~$ singularity exec /scratch/ss11645/LC/wags.sif tree /home/refgen/ -L 2
INFO:    squashfuse_ll mount took an unexpectedly long time: 6s
/home/refgen/
├── cat
│   └── Fca126_mat1.0
├── dog
│   ├── UU_Cfam_GSD_1.0_ROSY
│   ├── canfam3
│   └── canfam4
├── horse
│   └── goldenPath
└── tiger
    └── tiger
```

Implementation….

Cloning WAGS repository:

git clone https://github.com/jonahcullen/wags.git

Using fastq-dump to convert SRA to FASTQ



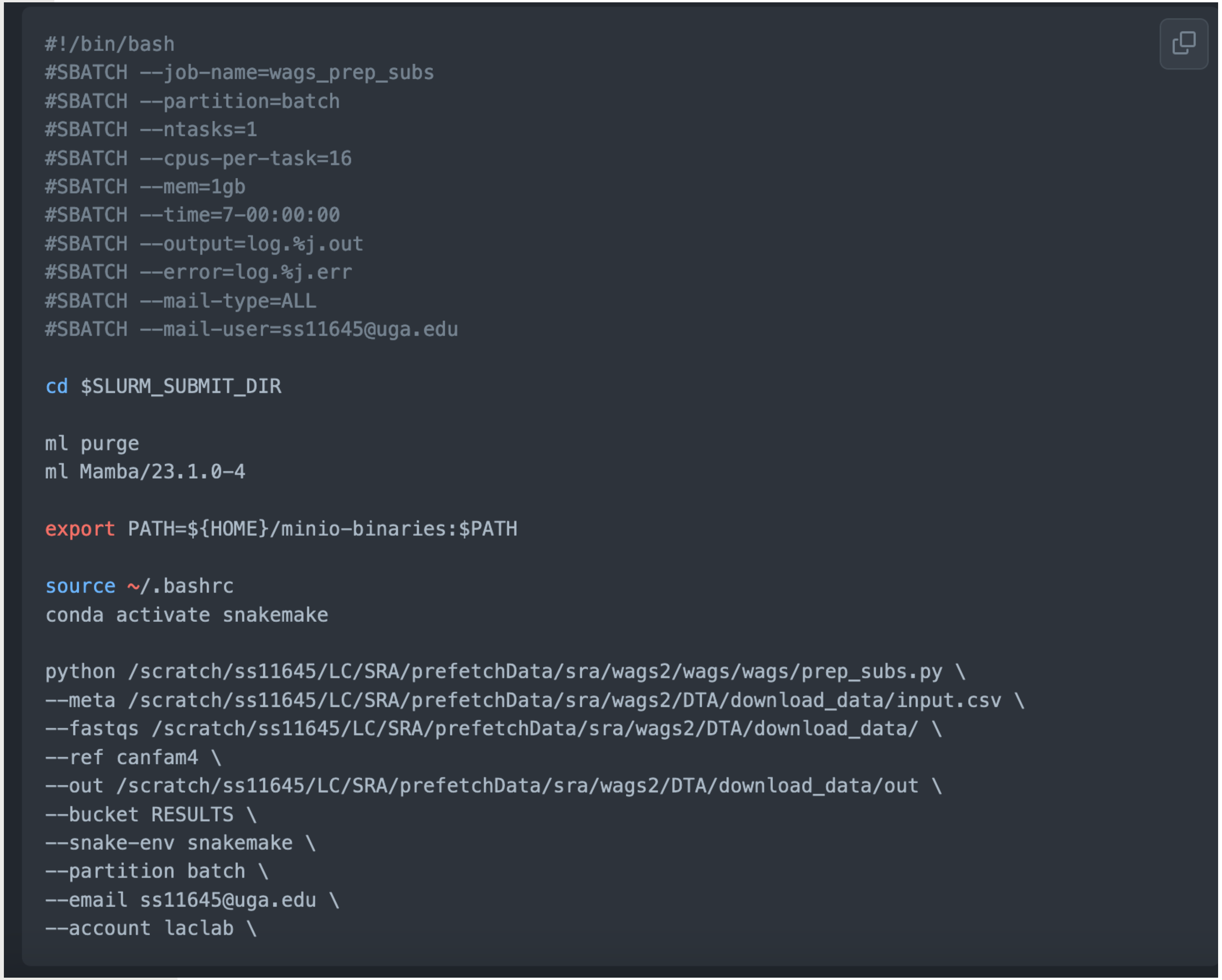
```
interact --mem=10gb -c 4
```

```
#!/bin/bash
#SBATCH --job-name=fastq-dump_job
#SBATCH --partition=batch
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=10
#SBATCH --mem=80gb
#SBATCH --time=120:00:00
#SBATCH --mail-type=ALL
#SBATCH --mail-user=ss11645@uga.edu
#SBATCH -o slurm_logs/%x_%j.out
#SBATCH -e slurm_logs/%x_%j.err

ml  SRA-Toolkit/3.0.1-centos_linux64

SRA_ID_HERE="ERR11203059"

OUTPUT_DIR="download_data"

fastq-dump --split-files --gzip --outdir "$OUTPUT_DIR" "$SRA_ID_HERE"
```

Getting Input files (as input.csv)

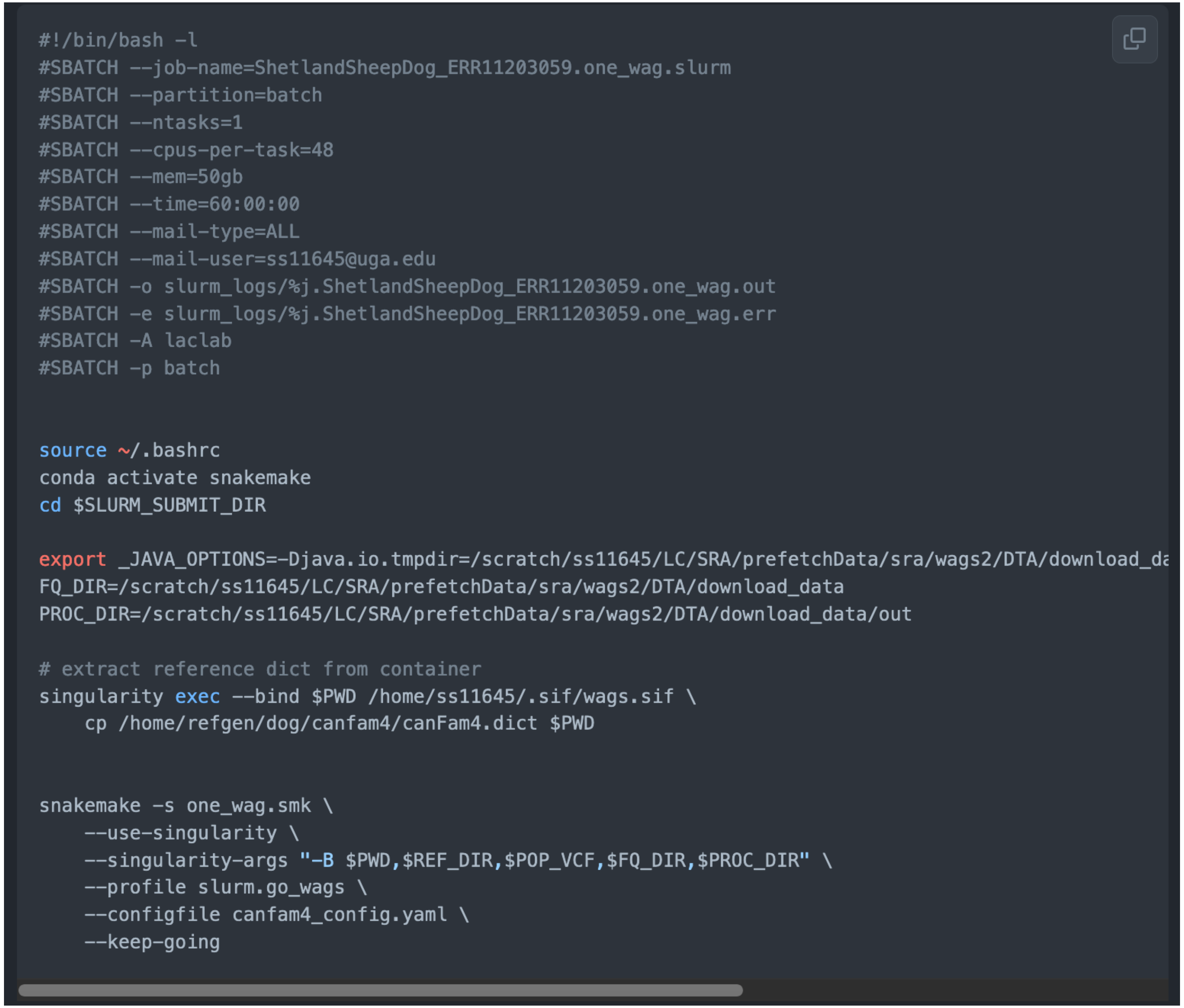
dogid	breed	gender	fastq
ERR11203059	ShetlandSheepDog	NA	ERR11203059
ERR11203057	ShetlandSheepDog	NA	ERR11203057
ERR11203035	ShetlandSheepDog	NA	ERR11203035
ERR11223859	Collie	NA	ERR11223859
ERR11203060	Collie	NA	ERR11203060
ERR11257717	Collie	NA	ERR11257717

Using pipeline to generate the slurm file for each inputs



```
#!/bin/bash
#SBATCH --job-name=wags_prep_subs
#SBATCH --partition=batch
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=16
#SBATCH --mem=1gb
#SBATCH --time=7-00:00:00
#SBATCH --output=log.%j.out
#SBATCH --error=log.%j.err
#SBATCH --mail-type=ALL
#SBATCH --mail-user=ss11645@uga.edu

cd $SLURM_SUBMIT_DIR

ml purge
ml Mamba/23.1.0-4

export PATH=${HOME}/minio-binaries:$PATH

source ~/.bashrc
conda activate snakemake

python /scratch/ss11645/LC/SRA/prefetchData/sra/wags2/wags/wags/prep_subs.py \
--meta /scratch/ss11645/LC/SRA/prefetchData/sra/wags2/DTA/download_data/input.csv \
--fastqs /scratch/ss11645/LC/SRA/prefetchData/sra/wags2/DTA/download_data/ \
--ref canfam4 \
--out /scratch/ss11645/LC/SRA/prefetchData/sra/wags2/DTA/download_data/out \
--bucket RESULTS \
--snake-env snakemake \
--partition batch \
--email ss11645@uga.edu \
--account laclab \
```

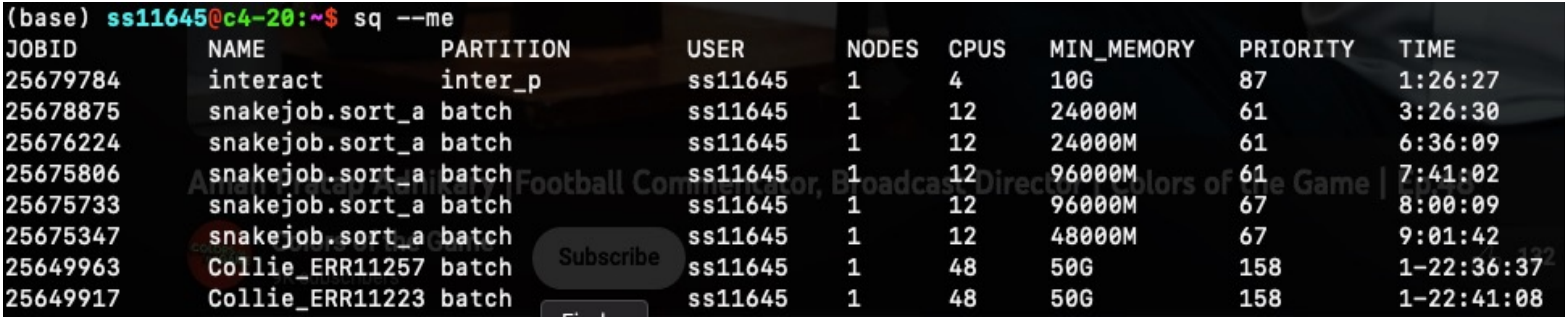
Now the slurm file is generated



```
#!/bin/bash -l
#SBATCH --job-name=ShetlandSheepDog_ERR11203059.one_wag.slurm
#SBATCH --partition=batch
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=48
#SBATCH --mem=50gb
#SBATCH --time=60:00:00
#SBATCH --mail-type=ALL
#SBATCH --mail-user=ss11645@uga.edu
#SBATCH -o slurm_logs/%j.ShetlandSheepDog_ERR11203059.one_wag.out
#SBATCH -e slurm_logs/%j.ShetlandSheepDog_ERR11203059.one_wag.err
#SBATCH -A laclab
#SBATCH -p batch


source ~/.bashrc
conda activate snakemake
cd $SLURM_SUBMIT_DIR

export _JAVA_OPTIONS=-Djava.io.tmpdir=/scratch/ss11645/LC/SRA/prefetchData/sra/wags2/DTA/download_da
FQ_DIR=/scratch/ss11645/LC/SRA/prefetchData/sra/wags2/DTA/download_data
PROC_DIR=/scratch/ss11645/LC/SRA/prefetchData/sra/wags2/DTA/download_data/out

# extract reference dict from container
singularity exec --bind $PWD /home/ss11645/.sif/wags.sif \
    cp /home/refgen/dog/canfam4/canFam4.dict $PWD


snakemake -s one_wag.smk \
    --use-singularity \
    --singularity-args "-B $PWD,$REF_DIR,$POP_VCF,$FQ_DIR,$PROC_DIR" \
    --profile slurm.go_wags \
    --configfile canfam4_config.yaml \
    --keep-going
```

Recent Status of work

Jobs submitted for ShetlandSheepDog was failed due t issues related to time to run in cluster, data errors from SRA and memory specifications



```
(base) ss11645@c4-20:~$ sq --me
JOBID          NAME             PARTITION        USER        NODES   CPUS    MIN_MEMORY    PRIORITY    TIME
25679784       interact         inter_p          ss11645     1       4       10G           87          1:26:27
25678875       snakejob.sort_a  batch            ss11645     1       12      24000M        61          3:26:30
25676224       snakejob.sort_a  batch            ss11645     1       12      24000M        61          6:36:09
25675806       snakejob.sort_a  batch            ss11645     1       12      96000M        61          7:41:02
25675733       snakejob.sort_a  batch            ss11645     1       12      96000M        67          8:00:09
25675347       snakejob.sort_a  batch            ss11645     1       12      48000M        67          9:01:42
25649963       Collie_ERR11257  batch            ss11645     1       48      50G           158         1-22:36:37
25649917       Collie_ERR11223  batch            ss11645     1       48      50G           158         1-22:41:08
```

Limitations

- Navigating a new pipeline for both me and the cluster, leading to extended setup times.
- Dealing with huge files.
- Adjustments made by collaborator Dr. Jonah Cullen to rectify errors in the pipeline code.

What next ?

- The current step is FASTQ to GVCF (OneWAG) which will give GVCFs.
- The next steps are:

1. GVCFS to VCF (ManyWAGS)
2. GVCFS to VCF (OnlyWAGS)

Illustrations

- WAGS pipeline (https://github.com/jonahcullen/wags)
- Clark Lab, WAGS repository (https://github.com/sachin11645/Whole-Animal-Genome-Sequencing)
- WAGS paper (https://doi.org/10.1093/g3journal/jkad117)
- Issues: (https://github.com/jonahcullen/wags/issues/38)

Acknowledgments

Dr. Leigh Anne Clark, Associate Professor, Department of Veterinary Pathology, UGA

Clark Lab Members:
Lizzie, Thais, Yueqi,

Dr. Jonah Cullen, University of Minnesota

Dr. Zhuofei Hou, Georgia Advanced Computing Resource Center(GACRC)/EITS, University of Georgia