Starling-May18
Projects/Katarina Stuart/KStuart.Starling-Aug18/Sv10_NZstarlings/Data/2023-11-30.ReadPrepping

PDF Version generated by

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

on

Sep 25, 2024 @01:00 PM AEST

Table of Contents

2023-11-30.ReadPrepping

Katarina Stuart (z5188231@ad.unsw.edu.au) - Feb 21, 2024, 9:00 AM GMT+11

Starling DArT raw data processing

Starlings Batch 1

2020 old dart data from Stuart et al 2022 mol ecol

```
cd /nesi/nobackup/uoa02613/kstuart_projects/Sv10_NZstarlings/data/processing_rawdata/batch1/process_radtags/

#copied from katana: /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv4_Historic/processing/samples_default so that I have same read depth for all I have not stacked multiple fq files

#for array job
ls -lh *.fq.gz | awk '{print $9}' | sed 's/.fq.gz//g' > ../starling_names_batch1.txt
```

quality trim

```
#!/bin/bash -e

#SBATCH --job-name=2024_02_21.trimming_batch1.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-12:00:00
#SBATCH --mem=4GB
#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
#SBATCH --array=1-79

sample=$(sed
"${SLURM_ARRAY_TASK_ID}q;d" /nesi/nobackup/uoa02613/kstuart_projects/Sv10_NZstarlings/data/processing_rawdata/batch1/starling_names_batch1.txt)

module load fastp/0.23.2-GCC-11.3.0 MultiQC/1.13-gimkl-2022a-Python-3.10.5 FastQC/0.11.9

#quality trim
cd /nesi/nobackup/uoa02613/kstuart_projects/Sv10_NZstarlings/data/processing_rawdata/batch1/trimmed/

DIR=/nesi/nobackup/uoa02613/kstuart_projects/Sv10_NZstarlings/data/processing_rawdata/batch1/process_radtags

fastp -q 22 --adapter_sequence=AGATCGGAAGAG -l 40 -i ${DIR}/${sample}.fq.gz -o ${sample}.filtered.fq.gz

#QC
fastqc ${sample}.filtered.fq.gz -o fastqc/
```

multiQC

```
cd /nesi/nobackup/uoa02613/kstuart_projects/Sv10_NZstarlings/data/processing_rawdata/batch1/trimmed/
```

```
module load fastp/0.23.2-GCC-11.3.0 MultiQC/1.13-gimkl-2022a-Python-3.10.5 FastQC/0.11.9

outdir2=fastqc
outdir3=multiqc
multiqc $outdir2 -o $outdir3
```

Starlings Batch 2 (old - remove)

2023 small batch of starlings sequenced alongside mynas

Requires radtag processing, adapter trimming, and quality/length trimming

```
cd /nesi/nobackup/uoa02613/kstuart_projects/Sv10_NZstarlings/data/processing_rawdata/batch2

DIR=/nesi/project/uoa02613/MYNA_sequence/Myna_DARTseq_012023/Myna_DART

#find just the targets info for the starling samples (there was also myna samples)
cat ${DIR}/targets_HFT2MDMXY_2.csv <(tail -n +2 ${DIR}/targets_HJ32WDRX2_1.csv) <(tail -n +2 ${DIR}/targets_HLJNMDRX2_1.csv) | sed 's/,/\t/g' | grep "S0" | cut -f1,5,16 > starlings_allrawfiles_batch2.txt

#sample identifiers, ignoring all the many many repeated sequencing files, but keeping in a few
grep "S0" /nesi/project/uoa02613/MYNA_sequence/Myna_DARTseq_012023/Sample_identifiers.txt | sed 's/S_//g' > starlings_batch2.txt
#did some manual edits so each sample has 1 repeat that has double read depth.

#match the human readable sample names with their corresponding barcode
awk 'NR==FNR{a[$2]=$3;next}{$1=a[$1]}1' starlings_allrawfiles_batch2.txt starlings_batch2.txt | sed 's/ /\t/g' > starlings_barcodes_batch2.txt

#make file for array later on
cut -f2 starlings_barcodes_batch2.txt | sort | uniq > starling_names_batch2.txt
```

process radtags

```
#!/bin/bash -e

#SBATCH --job-name=2023_12_01.procesing_tags_batch2.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-12:00:00
#SBATCH --mem=4GB
#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

module load Stacks/2.61-gimkl-2022a

RAW_DIR=/nesi/project/uoa02613/MYNA_sequence/Myna_DARTseq_012023/Myna_DART
OUTPUT_DIR=/nesi/nobackup/uoa02613/kstuart_projects/Sv10_NZstarlings/data/processing_rawdata/batch2/process_radtags/
BARCODES=/nesi/nobackup/uoa02613/kstuart_projects/Sv10_NZstarlings/data/processing_rawdata/batch2/starlings_barcodes_batch2.txt

process_radtags -p ${RAW_DIR} -o ${OUTPUT_DIR} -b $BARCODES -r -c -q --renz_1 pstI --renz_2 sphI
```

quality and length trim

```
#!/bin/bash -e

#SBATCH --job-name=2023_12_01.trimming_batch2.sl
#SBATCH --account=uoa02613
```

```
#SBATCH --time=00-12:00:00
#SBATCH --mem=4GB
#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
#SBATCH --array=1-30

sample=$(sed
"${SLURM_ARRAY_TASK_ID}q;d" /nesi/nobackup/uoa02613/kstuart_projects/Sv10_NZstarlings/data/processing_rawdata/batch2/starling_names_batch2.txt)

module load fastp/0.23.2-GCC-11.3.0 MultiQC/1.13-gimkl-2022a-Python-3.10.5 FastQC/0.11.9

#quality trim
cd /nesi/nobackup/uoa02613/kstuart_projects/Sv10_NZstarlings/data/processing_rawdata/batch2/trimmed/
DIR=/nesi/nobackup/uoa02613/kstuart_projects/Sv10_NZstarlings/data/processing_rawdata/batch2/process_radtags/
fastp -q 22 --adapter_sequence=AGATCGGAAGAG -l 40 -i ${DIR}/${sample}.fq.gz -o ${sample}.untrimmed.filtered.fq.gz

# Trim to fix lengths
fastp -b 69 -i ${sample}.untrimmed.filtered.fq.gz -o ${sample}.filtered.fq.gz

#QC
fastqc ${sample}.filtered.fq.gz -o fastqc/
```

multiQC

```
cd /nesi/nobackup/uoa02613/kstuart_projects/Sv10_NZstarlings/data/processing_rawdata/batch2/trimmed/

module load fastp/0.23.2-GCC-11.3.0 MultiQC/1.13-gimkl-2022a-Python-3.10.5 FastQC/0.11.9

outdir2=fastqc/
outdir3=multiqc/
multiqc $outdir2 -o $outdir3
```

Starlings Batch 2 - adapter seqs retrimmed

2023 small batch of starlings sequenced alongside mynas

Requires radtag processing, adapter trimming, and quality/length trimming

```
cd /nesi/nobackup/uoa02613/kstuart_projects/Sv10_NZstarlings/data/processing_rawdata/batch2_retrim

cp ../batch2/starlings_batch2.txt .

cp -r ../batch2/process_radtags .
```

quality and length trim

```
#!/bin/bash -e

#SBATCH --job-name=2024_02_14.trimming_batch2retrim.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-12:00:00
#SBATCH --mem=4GB
```

```
#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
#SBATCH --array=1-30

sample=$(sed
"${SLURM_ARRAY_TASK_ID}q;d" /nesi/nobackup/uoa02613/kstuart_projects/Sv10_NZstarlings/data/processing_rawdata/batch2/starling_names_batch2.txt)

module load fastp/0.23.2-GCC-11.3.0 MultiQC/1.13-gimkl-2022a-Python-3.10.5 FastQC/0.11.9

#quality trim
cd /nesi/nobackup/uoa02613/kstuart_projects/Sv10_NZstarlings/data/processing_rawdata/batch2_retrim/trimmed/
DIR=/nesi/nobackup/uoa02613/kstuart_projects/Sv10_NZstarlings/data/processing_rawdata/batch2_retrim/process_radtags/

fastp -q 22 --adapter_sequence=AGATCGGAAGAG -l 40 -i ${DIR}/${sample}.fq.gz -o ${sample}.untrimmed.filtered.fq.gz

# Trim to fix lengths
fastp -b 69 -i ${sample}.untrimmed.filtered.fq.gz -o ${sample}.filtered.fq.gz

#QC
fastqc ${sample}.filtered.fq.gz -o fastqc/
```

multiQC

```
cd /nesi/nobackup/uoa02613/kstuart_projects/Sv10_NZstarlings/data/processing_rawdata/batch2_retrim/trimmed/

module load fastp/0.23.2-GCC-11.3.0 MultiQC/1.13-gimkl-2022a-Python-3.10.5 FastQC/0.11.9

outdir2=fastqc/
outdir3=multiqc/
multiqc $outdir2 -o $outdir3
```

Starlings Batch 3

2023 single dart plate of starlings

Requires quality/length trimming

```
cd /nesi/nobackup/uoa02613/kstuart_projects/Sv10_NZstarlings/data/processing_rawdata/batch3

DIR=/nesi/project/uoa02613/MYNA_sequence/Starling_DARTseq_112023/raw_fastq

#find just the targets info for the starling samples (there was also myna samples)
tail -n +2 ${DIR}/targets_22FFNWLT3_7.csv | sed 's/,/\t/g'  | awk -v OFS='\t' '{print $1,$5}' > starlings_filenames_batch3.txt
```

```
#rename files and copy to location
DIR1=/nesi/project/uoa02613/MYNA_sequence/Starling_DARTseq_112023/raw_fastq
DIR2=/nesi/nobackup/uoa02613/kstuart_projects/Sv10_NZstarlings/data/processing_rawdata/batch3/process_radtags
FILE=/nesi/nobackup/uoa02613/kstuart_projects/Sv10_NZstarlings/data/processing_rawdata/batch3/starlings_filenames_batch3.txt

while read -r old_name new_name; do
    cp "$DIR1/$old_name.FASTQ.gz" "$DIR2/$new_name.fq.gz"
done < "$FILE"

#make file for array job
cut -f2 /nesi/nobackup/uoa02613/kstuart_projects/Sv10_NZstarlings/data/processing_rawdata/batch3/starlings_filenames_batch3.txt | sort | uniq >
starling_names_batch3.txt
```

quality and length trim

```
#!/bin/bash -e

#SBATCH --job-name=2024_02_21.trimming_batch3.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-12:00:00
#SBATCH --mem=4GB
#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
#SBATCH --array=1-94

sample=$(sed
"${SLURM_ARRAY_TASK_ID}q;d" /nesi/nobackup/uoa02613/kstuart_projects/Sv10_NZstarlings/data/processing_rawdata/batch3/starling_names_batch3.txt)

module load fastp/0.23.2-GCC-11.3.0 MultiQC/1.13-gimkl-2022a-Python-3.10.5 FastQC/0.11.9

#quality trim
cd /nesi/nobackup/uoa02613/kstuart_projects/Sv10_NZstarlings/data/processing_rawdata/batch3/trimmed/

DIR=/nesi/nobackup/uoa02613/kstuart_projects/Sv10_NZstarlings/data/processing_rawdata/batch3/process_radtags/

fastp -q 22 --adapter_sequence=AGATCGGAAGAG -l 40 -i ${DIR}/${sample}.fq.gz -o ${sample}.untrimmed.filtered.fq.gz

# Trim to fix lengths
fastp -b 69 -i ${sample}.untrimmed.filtered.fq.gz -o ${sample}.filtered.fq.gz

#QC
fastqc ${sample}.filtered.fq.gz -o fastqc/
```

multiQC

```
cd /nesi/nobackup/uoa02613/kstuart_projects/Sv10_NZstarlings/data/processing_rawdata/batch3/trimmed/

module load fastp/0.23.2-GCC-11.3.0 MultiQC/1.13-gimkl-2022a-Python-3.10.5 FastQC/0.11.9

outdir2=fastqc/
outdir3=multiqc/
multiqc $outdir2 -o $outdir3
```

Batch 1-3 data prep

```
cd /nesi/nobackup/uoa02613/kstuart_projects/Sv10_NZstarlings/data/processing_rawdata/trimmed_reads_all

ln -s /nesi/nobackup/uoa02613/kstuart_projects/Sv10_NZstarlings/data/processing_rawdata/batch1/trimmed/*fq.gz .

ln -s /nesi/nobackup/uoa02613/kstuart_projects/Sv10_NZstarlings/data/processing_rawdata/batch2/trimmed/*fq.gz .

ln -s /nesi/nobackup/uoa02613/kstuart_projects/Sv10_NZstarlings/data/processing_rawdata/batch3/trimmed/*fq.gz .

rm *untrimmed*
```