##

(c) 2015-2017 Merly Escalona merlyescalona@uvigo.es

Phylogenomics Lab. University of Vigo.

#

Description:

==========

Pipelines for data simulation for variant calling assesment

Running @ft2.cesga.es

##
#!/bin/bash -l
##

Previous to running the wrapper I had

to set up the perl env.

< o conf mbuildpl_arg '--install_base /home/uvi/be/mef/perl'
cpan> o conf commit
cpan> q
cpan install Math::GSL
MODULE_INSTALL_PERL
<<SPLIT_COMMANDS

If staying at LUSTRE, LUSTRE does not allow to launch more than 1000 jobs.

So,if I had to split the files and wait for all the jobs to finish to launch

the following 1000 jobs.

In any case, I'm moving things to triploid,

Way better and faster to run on triploid sequentially

SPLIT_COMMANDS
<<RSYNC

This takes like an hour

rsync -rP $LUSTRE/data/ngsphy.data/NGSphy_ssp.00002/
merly@triploid.uvigo.es:/home/merly/data/NGSphy_ssp.00002

Had to change the names of the paths for the files that were used, since I'm no longer at cesga

cat ssp.00002.sh | sed
's/\/mnt\/lustre\/scratch\/home\/uvi\/be\/mef\/data\/ngsphy.data/\/home\/merly\/data/g' | sed
's/\/home\/uvi\/be\/mef\/vc-benchmark-cesga\/files/\/home\/merly\/csNGSProfile/g' >
ssp.00002.triploid.sh
RSYNC
##
##
##
##

0. Folder structure

##
##
##
##

git clone https://merlyescalona@github.com/merlyescalona/vc-benchmark-cesga.git $HOME/vc-benchmark-cesga

git clone

https://merlyescalona@github.com/merlyescalona/refselector.git
$HOME/src/refselector

mkdir $folderDATA $folderOUTPUT $folderERROR $folderINFO

```
##############################################################################
##############################################################################
##############################################################################
##############################################################################
```

Folder paths

```
##############################################################################
##############################################################################
source $HOME/vc-benchmark-cesga/src/ssp.variables.sh
##############################################################################
##############################################################################
```

SLURM ENV - This is run @ ft2.cesga.es

```
##############################################################################
##############################################################################
simphyReplicateID="19-25"
##############################################################################
####
```

1. SIMPHY

```
################################################################################
####
step1JOBID=$(sbatch -a $simphyReplicateID $folderJOBS/1.datasim/ssp.1.simphy.sh | awk '{
print $4}')
################################################################################
####
```

2. INDELIBLE WRAPPER

```
################################################################################
####
```

After the running of SimPhy, it is necessary to run the INDELIble_wrapper

to obtain the control files for INDELible. Since, is not possible to

run it for all the configurations, it is necessary to modify the name of the

output files in order to keep track of every thing

```
################################################################################
####
step2JOBID=$(sbatch -a $simphyReplicateID --dependency=afterok:$step1JOBID
$folderJOBS/1.datasim/ssp.2.wrapper.sh | awk '{ print $4}')
################################################################################
```

####

3. INDELIBLE

##
####

Need to figure out the folder from where I'll call indelilble

Need to filter the species tree replicates that do not have ninds % 2==0

```
for simphyReplicateID in 19 20 21 22 23 24 25; do
numJobs=$(wc -l $HOME/vc-benchmark-cesga/files/${pipelinesName}.$(printf "%05g"
$simphyReplicateID).indelible.folders.txt | awk '{ print $1}')
step3JOBID=$(sbatch -a 1-$numJobs $folderJOBS/1.datasim/ssp.3.indelible.array.sh
$simphyReplicateID | awk '{ print $4}')
done
#--------------------------------------------------------------------------
<<CHECK_NUM_FILES_INDELIBLE
```

To check num fasta files and trees in indelible folders

```
count=0; alljobs=0;
for simphyReplicateID in $(seq 24 25); do
indelibleFolders="$HOME/vc-benchmark-cesga/files/ssp.$(printf "%05g"
$simphyReplicateID).indelible.folders.txt"
for item in $(cat $indelibleFolders);do
cd $item;
GT=$(ls g_trees* | wc -l)
```

```
TRUEFASTA=$(ls *.fasta | grep TRUE | wc -l)
FASTA=$(ls *.fasta | grep -v TRUE | wc -l)
if [[ $TRUEFASTA -eq $GT ]]; then
echo -e "\033[1m\033[91m$item: gt $GT\t\tFASTA=$FASTA\tTRUEFASTA=$TRUEFASTA
\033[0m"
let count=count+1
else
echo -e "$item: gt $GT\t\tFASTA=$FASTA\tTRUEFASTA $TRUEFASTA"
fi
let alljobs=alljobs+1
done
echo "-----------------------------------------------------------------------------------------------------------"
done
echo "$count/$alljobs - jobs should have finished"
CHECK_NUM_FILES_INDELIBLE
#########################################################################
```

5. NGSPHY

```
#########################################################################
step5JOBID=$(sbatch -a $simphyReplicateID --dependency=afterok:$step3JOBID
$folderJOBS/1.datasim/ssp.5.ngsphy.sh | awk '{ print $4}')
step8JOBID=$(sbatch -a $simphyReplicateID --dependency=afterok:$step5JOBID
$folderJOBS/1.datasim/ssp.8.cesga.data.compression.sh | awk '{ print $4}')
#########################################################################
#########################################################################
#########################################################################
#########################################################################
```

MOVING TO SGE @triplid.uvigo.es

```
#########################################################################
#########################################################################
#########################################################################
#########################################################################
```

6. DATA TRANSFER

---

step6JOBID=$(qsub -t $simphyReplicateID $HOME/src/vc-benchmark-cesga/jobs/1.datasim/ssp.7.1.data.transfer.sge.sh | awk '{ print $3}')
##

7. Reference Loci Selection

##
step4JOBID=$(qsub -t $simphyReplicateID $HOME/src/vc-benchmark-cesga/jobs/1.datasim/ssp.4.references.sge.sh | awk '{ print $1}')
##

LAUNCHING JOBS FOR ART GENERATION

##
step6OBID=$(qsub -t $simphyReplicateID $HOME/src/vc-benchmark-cesga/jobs/1.datasim/ssp.6.prep.2.art.sge.sh | awk '{ print $2}')
simphyReplicateID=6
for item in 17; do # $(seq 17); do
simphyReplicateID=$item #$item
pipelinesName="ssp"
replicatesNumDigits=5
replicateID="$(printf "%0${replicatesNumDigits}g" $simphyReplicateID)"
ngsphyReplicatePath="$HOME/data/NGSphy_${pipelinesName}.${replicateID}"
replicates=($(ls $ngsphyReplicatePath/reads))
artFilesReplicate="$HOME/src/vc-benchmark-cesga/files/${pipelinesName}.${replicateID}.art.commands.files.txt"
rm $artFilesReplicate
touch $artFilesReplicate
for item in $(find $ngsphyReplicatePath/scripts/ -name "${pipelinesName}.${replicateID}.HS25.PE.150.art.commands*" | sort); do echo $item >> $artFilesReplicate done for item in $(find $ngsphyReplicatePath/scripts/ -name "${pipelinesName}.${replicateID}.HS25.SE.150.art.commands*" | sort); do

```
echo $item >> $artFilesReplicate
done
for item in $(find $ngsphyReplicatePath/scripts/ -name
"${pipelinesName}.${replicateID}.MSv3.SE.250.art.commands" | sort); do echo $item >>
$artFilesReplicate done for item in $(find $ngsphyReplicatePath/scripts/ -name
"${pipelinesName}.${replicateID}.MSv3.PE.250.art.commands" | sort); do
echo $item >> $artFilesReplicate
done
nJobs=$(cat $artFilesReplicate |wc -l | awk '{print $1}')
step7JOBID=$(qsub -t 1-$nJobs $HOME/src/vc-benchmark-
cesga/jobs/1.datasim/ssp.7.art.sge.sh $artFilesReplicate | awk '{print $2}')
done
##############################################################################
####
##############################################################################
####
for item in 25; do
LUSTRE="/mnt/lustre/scratch/home/uvi/be/mef"
simphyReplicateID=$item #$item
pipelinesName="ssp"
replicatesNumDigits=5
replicateID="$(printf "%0${replicatesNumDigits}g" $simphyReplicateID)"
ngsphyReplicatePathTRIPLOID="$HOME/data/NGSphy_${pipelinesName}.${replicateID}"
ngsphyReplicatePathCESGA="$LUSTRE/data/ngsphy.data/NGSphy_${pipelinesName}.${replica
teID}/"
replicateFOLDERCESGA="$LUSTRE/data/$pipelinesName.$replicateID/"
replicateFOLDERTRIPLOID="$HOME/data/$pipelinesName.$replicateID"
rsync -rP uvibemef@ft2.cesga.es:$ngsphyReplicatePathCESGA
$ngsphyReplicatePathTRIPLOID
#rsync -rP uvibemef@ft2.cesga.es:$replicateFOLDERCESGA $replicateFOLDERTRIPLOID
done

##############################################################################
####
```

ORGANIZATION OF READS PER INDIVIDUALS

```
##############################################################################
####
```

```
simphyReplicateID=5
pipelinesName="ssp"
replicatesNumDigits=5
replicateID="$(printf "%0${replicatesNumDigits}g" $simphyReplicateID)"
ngsphyReplicatePath="$HOME/data/NGSphy_${pipelinesName}.${replicateID}"
replicates=($(ls $ngsphyReplicatePath/reads))
for replicateST in ${replicates[*]}; do
step9PE150DFLT=$(qsub -t $simphyReplicateID $HOME/src/vc-benchmark-
cesga/jobs/1.datasim/ssp.9.organization.fq.individuals.sge.sh PE150DFLT PAIRED $replicateST
reads_run_PE_150_DFLT)
step9SE150DFLT=$(qsub -t $simphyReplicateID $HOME/src/vc-benchmark-
cesga/jobs/1.datasim/ssp.9.organization.fq.individuals.sge.sh SE150DFLT SINGLE $replicateST
reads_run_SE_150_DFLT)
step9SE250DFLT=$(qsub -t $simphyReplicateID $HOME/src/vc-benchmark-
cesga/jobs/1.datasim/ssp.9.organization.fq.individuals.sge.sh SE250DFLT SINGLE $replicateST
reads_run_SE_250_DFLT)
step9PE250DFLT=$(qsub -t $simphyReplicateID $HOME/src/vc-benchmark-
cesga/jobs/1.datasim/ssp.9.organization.fq.individuals.sge.sh PE250DFLT PAIRED $replicateST
reads_run_PE_250_DFLT)
done

################################################################################
####
```

To check status of the org.fq.ind jobs

```
################################################################################
####
for item in $(qstat | grep org | awk '{print $1}'); do
status=$(qstat -j $item | grep job_args | awk '{print $2}')
qstatTask=$(qstat | grep org | grep $item | awk '{print $10}')
folderParam=$(echo $status | tr "," " "| awk '{print $1"/"$3}')
simphyReplicateID=$qstatTask
pipelinesName="ssp"
replicatesNumDigits=5
replicateID="$(printf "%0${replicatesNumDigits}g" $simphyReplicateID)"
ngsphyReplicatePath="$HOME/data/NGSphy_${pipelinesName}.${replicateID}"
echo -e "${pipelinesName}.${replicateID}\t$item\t$folderParam\t$( ls -l
$ngsphyReplicatePath/$folderParam| grep R1 | wc -l)";
done
```

```
################################################################################
####
```

To check status of ART JOBS

```
################################################################################
####
for jobid in $(qstat | tail -n+2 | grep art | awk '{print $1}' | sort | uniq ); do
echo -e "$(qstat -j $jobid | grep job_args | awk '{print $2}')\t$(qstat | grep $jobid | wc -l)"
done

################################################################################
####
```

Launch single profile for specific replicates to org fq per ind

```
################################################################################
####
for item in 17; do # $(seq 17); do
simphyReplicateID=$item #$item
pipelinesName="ssp"
replicatesNumDigits=5
replicateID="$(printf "%0${replicatesNumDigits}g" $simphyReplicateID)"
ngsphyReplicatePath="$HOME/data/NGSphy_${pipelinesName}.${replicateID}"
replicates=($(ls $ngsphyReplicatePath/reads))
for replicateST in ${replicates[*]}; do
step9PE150DFLT=$(qsub -t $simphyReplicateID $HOME/src/vc-benchmark-
cesga/jobs/1.datasim/ssp.9.organization.fq.individuals.sge.sh PE150DFLT PAIRED $replicateST
reads_run_PE_150_DFLT)
done
done
################################################################################
####

touch $HOME/data/numinds.rep.txt
echo -e "PIPELINE_REPLICATE\t01\t02\t03\t04\t05\t06\t07\t08\t09\t10" >
$HOME/data/numinds.rep.txt
```

```
for item in $(seq 1 25); do
simphyReplicateID=$item #$item
pipelinesName="ssp"
replicatesNumDigits=5
replicateID="$(printf "%0${replicatesNumDigits}g" $simphyReplicateID)"
ngsphyReplicatePath="$HOME/data/NGSphy_${pipelinesName}.${replicateID}"
replicates=(01 02 03 04 05 06 07 08 09 10)
for rep in ${replicates[*]}; do
numInds=($numInds $(cat
$ngsphyReplicatePath/ind_labels/${pipelinesName}.${replicateID}.${rep}.individuals.csv | tail -
n+2 | wc -l))
done

done
for item in $(seq 11 25); do
simphyReplicateID=$item #$item
pipelinesName="ssp"
replicatesNumDigits=5
replicateID="$(printf "%0${replicatesNumDigits}g" $simphyReplicateID)"
ngsphyReplicatePath="$HOME/data/NGSphy_${pipelinesName}.${replicateID}"
echo "rm $ngsphyReplicatePath/scripts/art.commands"
rm $ngsphyReplicatePath/scripts/art.commands
done

for item in 24 25; do # $(seq 11 25); do
echo $item
simphyReplicateID=$item #$item
pipelinesName="ssp"
replicatesNumDigits=5
replicateID="$(printf "%0${replicatesNumDigits}g" $simphyReplicateID)"
ngsphyReplicatePath="$HOME/data/NGSphy_${pipelinesName}.${replicateID}"
triploidARTPE150=$ngsphyReplicatePath/scripts/${pipelinesName}.${replicateID}.triploid.HS25.
PE.150.sh
split -l 5000 -d -a 5 $triploidARTPE150
$ngsphyReplicatePath/scripts/${pipelinesName}.${replicateID}.HS25.PE.150.art.commands.
for file in $(ls $ngsphyReplicatePath/scripts/.art.commands); do mv $file "$file.sh"; done
done

for item in $(seq 11 25); do
simphyReplicateID=$item #$item
pipelinesName="ssp"
replicatesNumDigits=5
replicateID="$(printf "%0${replicatesNumDigits}g" $simphyReplicateID)"
```

```
ngsphyReplicatePath="$HOME/data/NGSphy_${pipelinesName}.${replicateID}"
replicates=($(ls $ngsphyReplicatePath/reads))
artFilesReplicate="$HOME/src/vc-benchmark-
cesga/files/${pipelinesName}.${replicateID}.art.commands.files.pe.150.txt"
rm $artFilesReplicate
touch $artFilesReplicate
find $ngsphyReplicatePath/scripts/ -name
"${pipelinesName}.${replicateID}.HS25.PE.150.art.commands*" | sort > $artFilesReplicate
nJobs=$(cat $artFilesReplicate |wc -l | awk '{print $1}')
step7JOBID=$(qsub -t 1-$nJobs $HOME/src/vc-benchmark-
cesga/jobs/1.datasim/ssp.7.art.sge.sh $artFilesReplicate | awk '{print $2}')
done

for item in $(seq 11 25); do
simphyReplicateID=$item #$item
pipelinesName="ssp"
replicatesNumDigits=5
replicateID="$(printf "%0${replicatesNumDigits}g" $simphyReplicateID)"
ngsphyReplicatePath="$HOME/data/NGSphy_${pipelinesName}.${replicateID}"
replicates=($(ls $ngsphyReplicatePath/reads))
artFilesReplicate="$HOME/src/vc-benchmark-
cesga/files/${pipelinesName}.${replicateID}.art.commands.files.pe.150.txt"
find $ngsphyReplicatePath/scripts/ -name
"${pipelinesName}.${replicateID}.HS25.PE.150.art.commands*"
done
```