MaSuRCA Assembly Report

masurca.PBS

```
#!/bin/bash

#PBS -l nodes=1:ppn=64
#PBS -l mem=3904gb
#PBS -N masurca.pt
#PBS -j oe
#PBS -q hammerhead
#PBS -l feature=mem_4tb
#PBS -l walltime=700:00:00

IFS=$'\n'
set -eu
umask 007

module load masurca/3.2.7
module list

export WD="/nobackup/echinotol/dmachado/20190821_phylobates/20191008_masurca"

printf "> Start [`date`]\n"

cd ${WD}

masurca masurca_config.txt

printf "> Assembly script done.\n"

bash assemble.sh

printf "> Done [`date`]\n"

```

masurca_config.txt

```
DATA

#################
## Paired end ###
#################

##
# 3kb
##

PE= pa 300 35
/projects/anura/dmachado/160831_macrogen_usa/03_fast
uniq/fastuniq_1b2_3kb_pe_1.fastq.mod
/projects/anura/dmachado/160831_macrogen_usa/03_fast
uniq/fastuniq_1b2_
3kb_pe_2.fastq

##
# 5kb
##

PE= pb 300 35
/projects/anura/dmachado/160831_macrogen_usa/03_fast
uniq/fastuniq_1b2_5kb_pe_1.fastq.mod
/projects/anura/dmachado/160831_macrogen_usa/03_fast
uniq/fastuniq_1b2_
5kb_pe_2.fastq

##
# 8kb
##

PE= pc 300 35
/projects/anura/dmachado/160831_macrogen_usa/03_fast
uniq/fastuniq_1b2_8kb_pe_1.fastq.mod
/projects/anura/dmachado/160831_macrogen_usa/03_fast
uniq/fastuniq_1b2_
8kb_pe_2.fastq


#################################
## Jumping Pairs (mate paired) ##
#################################

```

```
##
# 3kb
##

JUMP= ma 3000 500
/projects/echinotol/frogtoxin/pt_mp_03kb_1_1.fastq
/projects/echinotol/frogtoxin/pt_mp_03kb_1_2.fastq

##
# 5kb
##

JUMP= mb 5000 750
/projects/echinotol/frogtoxin/pt_mp_05kb_1_1.fastq
/projects/echinotol/frogtoxin/pt_mp_05kb_1_2.fastq

##
# 8kb
##

JUMP= mc 8000 1000
/projects/anura/dmachado/160831_macrogen_usa/03_fast
uniq/fastuniq_1b2_8kb_mp_1.fastq.mod
/projects/anura/dmachado/160831_macrogen_usa/03_fast
uniq/fastuniq
_1b2_8kb_mp_2.fastq

##
# 10kb
##

JUMP= md 10000 1500
/projects/echinotol/frogtoxin/pt_mp_10kb_1_1.fastq
/projects/echinotol/frogtoxin/pt_mp_10kb_1_2.fastq

############
## PacBio ##
############

PACBIO=
/projects/echinotol/dmachado/190203_pacbiodata/pteri
ibelis_pacbio_concat.fa

END

PARAMETERS

```

```
# GRAPH_KMER_SIZE is k-mer size for deBruijn graph values between 25 and 101 are supported, auto will compute the optimal size based on the read data and GC content
GRAPH_KMER_SIZE=auto

# Set USE_LINKING_MATES to 1 for Illumina-only assemblies and to 0 if you have 2x or more long (Sanger, 454) reads
USE_LINKING_MATES=0

#ILLUMINA ONLY. Set this to 1 to use SOAPdenovo contigging/scaffolding module.  Assembly will be worse but will run faster. Useful for very large (>=8Gbp) genomes from Illum
ina-only data
SOAP_ASSEMBLY=0

# Set this to 1 if your Illumina jumping library reads are shorter than 100bp
EXTEND_JUMP_READS=0

# Specifies whether to run the assembly on the grid
USE_GRID=0

# Use at most this much coverage by the longest Pacbio or Nanopore reads, discard the rest of the reads. Can increase this to 30 or 35 if your reads are short (N50<7000bp)
LHE_COVERAGE=25

# LIMIT_JUMP_COVERAGE is useful if you have too many jumping library mates. Typically set it to 60 for bacteria and something large (300) for mammals
LIMIT_JUMP_COVERAGE = 300

# These are the additional parameters to Celera Assembler. Do not worry about performance, number or processors or batch sizes -- these are computed automatically. fFr mamma
ls do not set cgwErrorRate above 0.15!!!
CA_PARAMETERS = cgwErrorRate=0.1

# CABOG ASSEMBLY ONLY: whether to attempt to close gaps in scaffolds with Illumina  or long read data
CLOSE_GAPS=1

```

```
# Minimum count k-mers used in error correction 1 means all k-mers are used.  one can increase to 2 if coverage >100
KMER_COUNT_THRESHOLD = 1

# Auto-detected number of cpus to use
NUM_THREADS= 64

# This is mandatory jellyfish hash -- a safe value is estimated_genome_size*20
JF_SIZE=180000000000

# This specifies if we do (1) or do not (0) want to trim long runs of homopolymers (e.g. GGGGGGGG) from 3' read ends, use it for high GC genomes
DO_HOMOPOLYMER_TRIM=0

END

```

assemble.sh

```
#!/bin/bash

# assemble.sh generated by masurca
CONFIG_PATH="/nobackup/echinotol/dmachado/20190821_phylobates/20191008_masurca/masurca_config.txt"
CMD_PATH="/apps/pkg/masurca/3.2.7/rhel7_u5-x86_64/gnu/bin/masurca"
set -o pipefail

# Test that we support <() redirection
(eval "cat <(echo test) >/dev/null" 2>/dev/null) ||
{
  echo >&2 "ERROR: The shell used is missing important features."
  echo >&2 "       Run the assembly script directly as './$0'"
  exit 1
}

# Parse command line switches

```

```
while getopts ":rc" o; do
  case "${o}" in
    c)
    echo "configuration file is '$CONFIG_PATH'"
    exit 0
    ;;
    r)
    echo "Rerunning configuration"
    exec perl "$CMD_PATH" "$CONFIG_PATH"
    echo "Failed to rerun configuration"
    exit 1
    ;;
    *)
    echo "Usage: $0 [-r] [-c]"
    exit 1
    ;;
  esac
done
set +e
# Set some paths and prime system to save environment variables
save () {
  (echo -n "$1=\""; eval "echo -n \"\$$1\""; echo '"') >> environment.sh
}
GC=
RC=
NC=
if tty -s < /dev/fd/1 2> /dev/null; then
  GC='\e[0;32m'
  RC='\e[0;31m'
  NC='\e[0m'
fi
log () {
  d=$(date)
  echo -e "${GC}[$d]${NC} $@"
}
fail () {
  d=$(date)
  echo -e "${RC}[$d]${NC} $@"
  exit 1
}
signaled () {
  fail Interrupted
}
trap signaled TERM QUIT INT
rm -f environment.sh; touch environment.sh
```

```
61
62  # To run tasks in parallel
63  #run_bg () {
64  #  semaphore -j $NUM_THREADS --id masurca_$$ -- "$@"
65  #}
66  #run_wait () {
67  #  semaphore -j $NUM_THREADS --id masurca_$$ --wait
68  #}
69  export PATH="/apps/pkg/masurca/3.2.7/rhel7_u5-
    x86_64/gnu/bin/../CA/Linux-
    amd64/bin:/apps/pkg/masurca/3.2.7/rhel7_u5-
    x86_64/gnu/bin:$PATH"
70  save PATH
71  export PERL5LIB=/apps/pkg/masurca/3.2.7/rhel7_u5-
    x86_64/gnu/bin/../lib/perl${PERL5LIB:+:$PERL5LIB}
72  save PERL5LIB
73  NUM_THREADS=64
74  save NUM_THREADS
75  log 'Processing pe library reads'
76  rm -rf meanAndStdevByPrefix.pe.txt
77  echo 'pa 300 35' >> meanAndStdevByPrefix.pe.txt
78  echo 'pb 300 35' >> meanAndStdevByPrefix.pe.txt
79  echo 'pc 300 35' >> meanAndStdevByPrefix.pe.txt
80  log 'Processing sj library reads'
81  rm -rf meanAndStdevByPrefix.sj.txt
82  echo 'ma 500 100' >> meanAndStdevByPrefix.sj.txt
83  echo 'mb 500 100' >> meanAndStdevByPrefix.sj.txt
84  echo 'mc 500 100' >> meanAndStdevByPrefix.sj.txt
85  echo 'md 500 100' >> meanAndStdevByPrefix.sj.txt
86
87  head -q -n 40000  pa.renamed.fastq pb.renamed.fastq
    pc.renamed.fastq | grep --text -v '^+' | grep --text
    -v '^@' > pe_data.tmp
88  export PE_AVG_READ_LENGTH=`awk '{if(length($1)>31)
    {n+=length($1);m++;}}END{print int(n/m)}'
    pe_data.tmp`
89  save PE_AVG_READ_LENGTH
90  log Average PE read length $PE_AVG_READ_LENGTH
91  KMER=`for f in pa.renamed.fastq pb.renamed.fastq
    pc.renamed.fastq;do head -n 80000 $f |tail -n
    40000;done | perl -e 'while($line=<STDIN>){$line=
    <STDIN>;chomp($line);push(@li
92  nes,$line);for($i=0;$i<6;$i++){$line=
    <STDIN>;}}$min_len=100000;$base_count=0;foreach
    $l(@lines)
    {$base_count+=length($l);push(@lengths,length($l));@
    f=split("",$l);foreach $ba
```

```
se(@f){if(uc($base) eq "G" || uc($base) eq "C")
{$gc_count++}}} @lengths =sort {$b <=> $a} @lengths;
$min_len=$lengths[int($#lengths*.75)];
 $gc_ratio=$gc_count/$base_count;$
kmer=0;if($gc_ratio>=0.35 && $gc_ratio<=0.6)
{$kmer=int($min_len*.66);}else{$kmer=int($min_len*.3
3);} $kmer++ if($kmer%2==0); $kmer=31 if($kmer<31);
$kmer=127 if($kmer>127);
print $kmer'`
save KMER
log Using kmer size of $KMER for the graph
KMER_J=31
save KMER_J
MIN_Q_CHAR=`cat pa.renamed.fastq pb.renamed.fastq
pc.renamed.fastq |head -n 50000 | awk 'BEGIN{flag=0}
{if($0 ~ /^\+/){flag=1}else if(flag==1){print
$0;flag=0}}'  | perl -ne
'BEGIN{$q0_char="@";}{chomp;@f=split "";foreach
$v(@f){if(ord($v)<ord($q0_char))
{$q0_char=$v;}}}END{$ans=ord($q0_char);if($ans<64)
{print "33\n"}else{print "64\n"}}'`
save MIN_Q_CHAR
log MIN_Q_CHAR: $MIN_Q_CHAR
JF_SIZE=`ls -l *.fastq | awk
'{n+=$5}END{s=int(n/50); if(s>180000000000)printf
"%.0f",s;else print "180000000000";}'`
save JF_SIZE
perl -e '{if(int('$JF_SIZE')>180000000000){print
"WARNING: JF_SIZE set too low, increasing JF_SIZE to
at least '$JF_SIZE', this automatic increase may be
not enough!\n"}}'



if [ -s ESTIMATED_GENOME_SIZE.txt ];then
ESTIMATED_GENOME_SIZE=`head -n 1
ESTIMATED_GENOME_SIZE.txt`
else
log Estimating genome size.
export ESTIMATED_GENOME_SIZE=`jellyfish histo -t 64
-h 1 k_u_hash_0 | tail -n 1 |awk '{print $2}'`
echo $ESTIMATED_GENOME_SIZE >
ESTIMATED_GENOME_SIZE.txt
fi
save ESTIMATED_GENOME_SIZE
log "Estimated genome size: $ESTIMATED_GENOME_SIZE"

```

```
log 'Computing super reads from PE '
CA_DIR="CA";
/apps/pkg/masurca/3.2.7/rhel7_u5-x86_64/gnu/bin/mega_reads_assemble_cluster.sh -Pb 300000000 -q all.q -G 0 -C 25 -t 64 -e $ESTIMATED_GENOME_SIZE -m work1 -p /projects/echinotol/dmachado/190203_pacbiodata/pteriibelis_pacbio_concat.fa -a /apps/pkg/masurca/3.2.7/rhel7_u5-x86_64/gnu/bin/../CA8/Linux-amd64/bin -o "ma.cor.clean.frg mb.cor.clean.frg mc.cor.clean.frg md.cor.clean.frg cgwErrorRate=0.1" -B 17 -D 0.029
CA_DIR=`cat CA_dir.txt`
TERMINATOR="9-terminator"
if [ -s $CA_DIR/9-terminator/genome.scf.fasta ];then
  NSCF=`grep --text '^>'  $CA_DIR/9-terminator/genome.scf.fasta |wc -l`
  NCTG=`grep --text '^>'  $CA_DIR/9-terminator/genome.ctg.fasta |wc -l`
  if [ $NCTG -eq $NSCF ];then
    log 'No gap closing possible.'
  else
    TERMINATOR="10-gapclose"
    if [ -s $CA_DIR/10-gapclose/genome.scf.fasta ];then
      log 'Gap closing done'
    else
      log 'Gap closing.'
      closeGapsLocally.perl --max-reads-in-memory 1000000000 -s 180000000000 --Celera-terminator-directory $CA_DIR/9-terminator --reads-file 'pa.renamed.fastq' --reads-file 'pb.renamed.fastq' --reads-file 'pc.renamed.fastq' --reads-file 'ma.renamed.fastq' --reads-file 'mb.renamed.fastq' --reads-file 'mc.renamed.fastq' --reads-file 'md.renamed.fastq' --output-directory $CA_DIR/10-gapclose --min-kmer-len 17 --max-kmer-len $(($PE_AVG_READ_LENGTH-5)) --num-threads 64 --contig-length-for-joining $(($PE_AVG_READ_LENGTH-1)) --contig-length-for-fishing 200 --reduce-read-set-kmer-size 21 1>gapClose.err 2>&1
      if [[ -e "$CA_DIR/10-gapclose/genome.ctg.fasta" ]];then
        log 'Gap close success.'
```

```
      else
        fail Gap close failed, you can still use pre-gap close files under $CA_DIR/9-terminator/. Check gapClose.err for problems.
      fi
    fi
  fi
else
  fail "Assembly stopped or failed, see $CA_DIR.log"
fi
if [ -s $CA_DIR/$TERMINATOR/genome.scf.fasta ];then
  if [ ! -e $CA_DIR/filter_map.contigs.success ];then
    PLOIDY=`cat PLOIDY.txt`
    deduplicate_contigs.sh $CA_DIR genome 64 $PLOIDY $TERMINATOR && log "Assembly complete, final scaffold sequences are in $CA_DIR/final.genome.scf.fasta"
  else
  log "Assembly complete, final scaffold sequences are in $CA_DIR/final.genome.scf.fasta"
  fi
else
  fail "Assembly stopped or failed, see $CA_DIR.log"
fi
log 'All done'

```

List of files and directories

```
8.0K -rwxrwxr-x  1 dmachado echinotol 6.6K Oct 14 09:24 assemble.sh
4.0K -rw-rw----  1 dmachado echinotol   21 Oct 14 09:24 CA_dir.txt
4.0K drwxr-xr-x 18 dmachado echinotol 4.0K Oct 23 06:04 CA.mr.41.15.17.0.029
 16M -rw-rw----  1 dmachado echinotol  16M Oct 23 05:52 CA.mr.41.15.17.0.029.log
3.6M -rw-rw----  1 dmachado echinotol 3.6M Oct  8 20:19 chimeric_sj.txt
4.0K -rw-rw----  1 dmachado echinotol   55 Oct  8 21:39 compute_jump_coverage.txt
 11M -rw-rw----  1 dmachado echinotol  11M Oct 15 10:58 containees.txt
```

```
4.0K -rw-rw----  1 dmachado echinotol 2.7K Oct 14 09:24 environment.sh
4.0K -rw-rw----  1 dmachado echinotol   11 Oct  8 12:28 ESTIMATED_GENOME_SIZE.txt
4.0K -rw-rw----  1 dmachado echinotol  901 Oct 23 05:52 gapClose.err
1.2G -rw-rw----  1 dmachado echinotol 1.2G Oct 19 12:29 genome.uid
 55M -rw-rw----  1 dmachado echinotol  55M Oct 19 13:35 global_arrival_rate.txt
3.8G -rw-rw----  1 dmachado echinotol 3.8G Oct  8 23:34 guillaumeKUnitigsAtLeast32bases_all.41.fasta
6.9G -rw-rw----  1 dmachado echinotol 6.9G Oct  8 12:47 guillaumeKUnitigsAtLeast32bases_all.fasta
6.2G -rw-rw----  1 dmachado echinotol 6.2G Oct  8 13:39 guillaumeKUnitigsAtLeast32bases_all.jump.fasta
 18G -rw-rw----  1 dmachado echinotol  18G Oct  8 12:26 k_u_hash_0
7.0G -rw-rw----  1 dmachado echinotol 7.0G Oct  8 21:42 ma.cor.clean.frg
 13G -rw-rw----  1 dmachado echinotol  13G Oct  8 10:23 ma.renamed.fastq
4.0K -rw-rw----  1 dmachado echinotol 3.4K Oct  8 10:11 masurca_config.txt
4.0K -rwxrwxr-x  1 dmachado echinotol  453 Oct  8 10:09 masurca.PBS
8.0K -rw-rw----  1 dmachado echinotol 8.0K Oct 12 18:25 masurca.pt.o1505049
4.0K -rw-rw----  1 dmachado echinotol 2.0K Oct 23 06:04 masurca.pt.o1517179
5.2G -rw-rw----  1 dmachado echinotol 5.2G Oct  8 21:44 mb.cor.clean.frg
 11G -rw-rw----  1 dmachado echinotol  11G Oct  8 10:24 mb.renamed.fastq
 19G -rw-rw----  1 dmachado echinotol  19G Oct  8 21:50 mc.cor.clean.frg
 26G -rw-rw----  1 dmachado echinotol  26G Oct  8 10:28 mc.renamed.fastq
1.9G -rw-rw----  1 dmachado echinotol 1.9G Oct  8 21:51 md.cor.clean.frg
4.4G -rw-rw----  1 dmachado echinotol 4.4G Oct  8 10:29 md.renamed.fastq
4.0K -rw-rw----  1 dmachado echinotol   30 Oct 14 09:24 meanAndStdevByPrefix.pe.txt
4.0K -rw-rw----  1 dmachado echinotol   44 Oct 14 09:24 meanAndStdevByPrefix.sj.txt
```

```
1.8G -rw-rw----  1 dmachado echinotol 1.8G Oct 15 09:38 mr.41.15.17.0.029.1.allowed
4.3G -rw-rw----  1 dmachado echinotol 4.3G Oct 15 10:13 mr.41.15.17.0.029.1.fa
7.4G -rw-rw----  1 dmachado echinotol 7.4G Oct 15 11:20 mr.41.15.17.0.029.1.frg
3.9G -rw-rw----  1 dmachado echinotol 3.9G Oct 15 10:02 mr.41.15.17.0.029.1.to_join.fa.tmp
139M -rw-rw----  1 dmachado echinotol 139M Oct 15 11:20 mr.41.15.17.0.029.1.trims.txt
452M -rw-rw----  1 dmachado echinotol 452M Oct 15 10:02 mr.41.15.17.0.029.1.unjoined.fa
 97G -rw-rw----  1 dmachado echinotol  97G Oct 12 12:20 mr.41.15.17.0.029.all_mr.fa
 93G -rw-rw----  1 dmachado echinotol  93G Oct 12 16:15 mr.41.15.17.0.029.all_mr.maximal.fa
 76G -rw-rw----  1 dmachado echinotol  76G Oct 15 07:44 mr.41.15.17.0.029.all.txt
 12K drwxrws---  2 dmachado echinotol  12K Oct 15 10:12 mr.41.15.17.0.029.join_consensus.tmp
 78G -rw-rw----  1 dmachado echinotol  78G Oct 12 15:56 mr.41.15.17.0.029.maximal_mr.txt
 93G -rw-rw----  1 dmachado echinotol  93G Oct 15 06:09 mr.41.15.17.0.029.mr.txt
508K -rw-rw----  1 dmachado echinotol 503K Oct 12 16:18 mr.41.15.17.0.029.single.txt
1.5G -rw-rw----  1 dmachado echinotol 1.5G Oct 15 11:21 mr.41.15.17.0.029.sr.frg
112G -rw-rw----  1 dmachado echinotol 112G Oct 12 10:50 mr.41.15.17.0.029.txt
 31G -rw-rw----  1 dmachado echinotol  31G Oct  8 23:25 pacbio_25xlongest.fa
 15G -rw-rw----  1 dmachado echinotol  15G Oct  8 10:13 pa.renamed.fastq
 25G -rw-rw----  1 dmachado echinotol  25G Oct  8 10:17 pb.renamed.fastq
 26G -rw-rw----  1 dmachado echinotol  26G Oct  8 10:21 pc.renamed.fastq
 35G -rw-rw----  1 dmachado echinotol  35G Oct  8 11:35 pe.cor.fa
 78M -rw-rw----  1 dmachado echinotol  78M Oct  8 11:35 pe.cor.tmp.log
6.6M -rw-rw----  1 dmachado echinotol 6.6M Oct 14 09:24 pe_data.tmp
4.0K -rw-rw----  1 dmachado echinotol    2 Oct 14 09:24 PLOIDY.txt
```

```
4.0K -rw-rw----  1 dmachado echinotol  434 Oct  8 12:07 quorum.err
842G -rw-rw----  1 dmachado echinotol 842G Oct  8 11:05 quorum_mer_db.jf
739M -rw-rw----  1 dmachado echinotol 739M Oct  8 20:59 redundant_sj.txt
4.0K -rw-rw----  1 dmachado echinotol  796 Oct 15 11:22 runCA.spec
 16G -rw-rw----  1 dmachado echinotol  16G Oct  8 21:15 sj.cor.clean.fa
 16G -rw-rw----  1 dmachado echinotol  16G Oct  8 21:30 sj.cor.clean.rev.fa
   0 lrwxrwxrwx  1 dmachado echinotol   19 Oct  8 21:30 sj.cor.ext.fa -> sj.cor.clean.rev.fa
 28G -rw-rw----  1 dmachado echinotol  28G Oct  8 12:07 sj.cor.fa
334M -rw-rw----  1 dmachado echinotol 334M Oct  8 12:07 sj.cor.tmp.log
 28K -rw-rw----  1 dmachado echinotol  27K Oct 15 10:58 super1.err
 36K -rw-rw----  1 dmachado echinotol  36K Oct  8 20:11 super2.1.err
 52K -rw-rw----  1 dmachado echinotol  48K Oct  8 18:56 super2.err
 12G -rw-rw----  1 dmachado echinotol  12G Oct  9 00:17 superReadSequences.named.fasta
   0 -rw-rw----  1 dmachado echinotol    0 Oct 19 13:36 tigStore.err
100M -rw-rw----  1 dmachado echinotol 100M Oct 19 13:35 unitig_cov.txt
 63G -rw-rw----  1 dmachado echinotol  63G Oct 19 12:39 unitig_layout.txt
4.0K drwxrws---  2 dmachado echinotol 4.0K Oct  8 23:18 work1
4.0K drwxrws---  2 dmachado echinotol 4.0K Oct  9 00:15 work1_mr
4.0K drwxrws---  2 dmachado echinotol 4.0K Oct  8 18:56 work2
4.0K drwxrws---  2 dmachado echinotol 4.0K Oct  8 20:11 work2.1
```