***The commands used in this study.***

## Remove Bacteria contaminated Illumina reads by Kraken2

/home/share/kraken2-2.0.8-beta/kraken2 -db /home/share/kraken2-2.0.8-beta/minikraken2\_v1\_8GB --threads 40 --paired SFG1.fq SFG2.fq --classified-out SFG\_Bac#.fq --unclassified-out SFG\_Host#.fq

#Remove the symbiont contaminated ONT reads

minimap2 -ax map-ont -t 40 ~/Data/SFG/Symbiont\_genome/Bac.fasta SFG\_HAC\_3Kb.fq | samtools fastq -n -f 4 - > SFG\_HAC\_3Kb\_BacFree.fq

#wtdbg2 assembly

~/App/wtdbg2/wtdbg2 -i ../Flye\_5Kb\_canu\_trimm\_ass/SFG\_HA.trimmedReads.fasta.gz -t 40 -o SFG\_HAC\_canuTrim

~/App/wtdbg2/wtpoa-cns -t 40 -i SFG\_HAC\_canuTrim.ctg.lay.gz -fo SFG\_HAC\_canuTrim.ctg.lay.fa

# Flye assembly and polishing

flye --nano-raw ../SFG\_HAC\_3Kb\_BacFree.fq -g 0.3566g -o flye\_ONT\_HAC\_3kb -t 40

flye --polish-target ../SFG\_10Kb\_HAC.fa --nanopore-raw ../SFG\_HAC\_3Kb\_BacFree.fq --iterations 3 --out-dir flye\_po\_R3 --threads 40

# minimap + miniasm assembly

minimap2 -X -t 20 -x ava-ont ../SFGB1\_clean\_3Kb.fq ../SFGB1\_clean\_3Kb.fq > reads.paf

miniasm -f ../SFGB1\_clean\_3Kb.fq reads.paf > SFG\_default.gfa

awk '/^S/{print ">"$2"\n"$3}' SFG\_default.gfa | fold > SFG\_default.fa

# canu assembly

~/App/canu-2.0/Linux-amd64/bin/canu -fast -p SFG\_HA -d SFG\_HA genomeSize=0.3566g -nanopore-raw SFG\_HAC\_3Kb\_BacFree.fq corOutCoverage=200 corMhapSensitivity=normal correctedErrorRate=0.105 minReadLength=5000 useGrid=true gridOptions=--partition=oces

#NECAT assembly

necat.pl config SFG\_HAC\_config.txt

necat.pl correct SFG\_HAC\_config.txt

necat.pl assemble SFG\_HAC\_config.txt

necat.pl bridge SFG\_HAC\_config.txt

# the configure file

PROJECT=SFG\_HAC

ONT\_READ\_LIST=readlist.txt

GENOME\_SIZE=356600000

THREADS=40

MIN\_READ\_LENGTH=5000

PREP\_OUTPUT\_COVERAGE=70

OVLP\_FAST\_OPTIONS=-n 500 -z 20 -b 2000 -e 0.5 -j 0 -u 1 -a 1000

OVLP\_SENSITIVE\_OPTIONS=-n 500 -z 10 -e 0.5 -j 0 -u 1 -a 1000

CNS\_FAST\_OPTIONS=-a 2000 -x 4 -y 12 -l 1000 -e 0.5 -p 0.8 -u 0

CNS\_SENSITIVE\_OPTIONS=-a 2000 -x 4 -y 12 -l 1000 -e 0.5 -p 0.8 -u 0

TRIM\_OVLP\_OPTIONS=-n 100 -z 10 -b 2000 -e 0.5 -j 1 -u 1 -a 400

ASM\_OVLP\_OPTIONS=-n 100 -z 10 -b 2000 -e 0.5 -j 1 -u 0 -a 400

NUM\_ITER=2

CNS\_OUTPUT\_COVERAGE=50

CLEANUP=1

USE\_GRID=false

GRID\_NODE=0

GRID\_OPTIONS=

SMALL\_MEMORY=0

FSA\_OL\_FILTER\_OPTIONS=

FSA\_ASSEMBLE\_OPTIONS=

FSA\_CTG\_BRIDGE\_OPTIONS=

POLISH\_CONTIGS=true

#NextDenovo configure file

[General]

job\_type = local

job\_prefix = nextDenovo

task = all # 'all', 'correct', 'assemble'

rewrite = yes # yes/no

deltmp = yes

rerun = 3

parallel\_jobs = 4

input\_type = raw

input\_fofn = ./input.fofn

workdir = ./01\_rundir

[correct\_option]

read\_cutoff = 1k

seed\_cutoff = 11653

blocksize = 1g

pa\_correction = 2

seed\_cutfiles = 2

sort\_options = -m 1g -t 2 -k 50

minimap2\_options\_raw = -x ava-ont -t 10

correction\_options = -p 15

[assemble\_option]

minimap2\_options\_cns = -x ava-ont -t 10 -k17 -w17

nextgraph\_options = -a 1

#Raven assembly

~/App/raven/build/bin/raven ../SFG\_HAC\_10Kb\_BacFree.fa.gz -t 40 -p 1 >Raven\_10Kb.fasta 2> log.txt

#Shasta assembly

sudo ~/App/shata\_v0.4.0/shasta-Linux-0.4.0 --input ../Mcor\_ONT\_10kb.fa --memoryMode filesystem --memoryBacking 2M

#QuickMerge

delta-filter -r -q -l 10000 SFG\_hybrid.delta > SFG\_hybrid.rq.delta

quickmerge -d SFG\_hybrid.rq.delta -q MaSu\_HAC35x\_FlyeP3\_PD\_Pilon2.fasta -r SFG\_Flye\_HAC10kb\_P2\_PD\_pilon2.fasta -hco 5.0 -c 1.5 -l 2164900 -ml 10000 -p SFG\_QM

# Purge\_dup version 1.2.3

minimap2 -x map-ont SFG\_FlyepolishedR3.fasta SFG\_ONT\_3Kb\_BacFree.fq -t 40 > reads.paf

pbcstat \*.paf

calcuts PB.stat -l13 -m51 -u144 > cutoffs 2>calcults.log

split\_fa SFG\_FlyepolishedR3.fasta > polished\_3.fasta.split

minimap2 -x asm5 -DP polished\_3.fasta.split polished\_3.fasta.split -t 40 >split.paf

purge\_dups -2 -T cutoffs -c PB.base.cov split.paf >dups.bed 2>purge\_dups.log

/get\_seqs dups.bed SFG\_FlyepolishedR3.fasta

#or with Illumina reads

bowtie2-build -f SFG\_canu\_flyeP3.fasta SFG\_canu --threads 40

bowtie2 -p 40 --maxins 800 -x SFG\_canu -1 SFG\_trim\_1.fq -2 SFG\_trim\_2.fq 1>PE.sam 2>SE.err

samtools view -bS PE.sam >PE.bam -@ 20

~/App/purge\_dups/bin/ngscstat PE.bam

~/App/purge\_dups/bin/calcuts TX.stat -l13 -m51 -u144 > cutoffs 2>calcults.log

~/App/purge\_dups/bin/split\_fa SFG\_canu\_flyeP3.fasta > polished\_3.fasta.split

minimap2 -x asm5 -DP polished\_3.fasta.split polished\_3.fasta.split -t 40 > polished\_3.fasta.split.paf

~/App/purge\_dups/bin/purge\_dups -2 -T cutoffs -c TX.base.cov polished\_3.fasta.split.paf > dups.bed 2> purge\_dups.log

#Pilon error correction

bowtie2-build -f canu\_FlyeP3\_PD2\_pilon1.fa SFS --threads 40

bowtie2 -p 40 -D 20 -R 2 -N 1 -L 18 -i S,1,0.50 --maxins 1200 -x SFS -1 SFG\_clean\_1.fq -2

SFG\_clean\_2.fq 1>SFSPE500.sam 2> SFSPE500.err

grep -E "@|NM:" SFSPE500.sam | grep -v "XS:" > SFSPE500\_uniq.sam

samtools view -bS SFSPE500\_uniq.sam > SFSPE500\_uniq.bam -@ 40

samtools sort SFSPE500\_uniq.bam -m 5G -@ 10 -o SFSPE500\_uniq\_sorted.bam

java -jar ~/App/picard/picard.jar MarkDuplicates I= SFSPE500\_uniq\_sorted.bam O= SFSPE500\_uniq\_sorted\_dedupe.bam METRICS\_FILE=metrics.txt

samtools index SFSPE500\_uniq\_sorted\_dedupe.bam

java -Xmx180G -jar ~/App/pilon-1.23/pilon-1.23.jar --genome canu\_FlyeP3\_PD2\_pilon1.fa --frags SFSPE500\_uniq\_sorted\_dedupe.bam --diploid --threads 40

#QUAST analysis

quast.py ./canu/Canu.fasta ./flye/Flye.fasta ./masurca/MaSuRCA.fasta ./miniasm/Miniasm.fasta ./necat/NECAT.fasta ./nextdenovo/NextDenovo.fasta

./raven/Raven.fasta ./shasta/Shasta.fasta ./wtdbg2/Wtdbg2.fasta ./quickmerger/QuickMerger.fasta -r ./Csqv1.1/Csq\_v1.1.fa -t 40

#repeatmodeler and repeatmasker

BuildDatabase -name Mcor\_v2.0 ../Mcor\_v2.0.fasta

~/App/RepeatModeler-2.0.1/RepeatModeler -database Mcor\_v2.0 -pa 40

RepeatMasker -species all -pa 8 -div 30 Mcor\_v2.0.fasta

RepeatMasker -lib /home/sunj/Data/Mcoru/Annotations/RepeatModeler/Mcor\_v2.0-families.fa -pa 10 -div 30 Mcor\_v2.0.fasta.masked

# Braker for Augustus training

~/App/BRAKER-2.1.5/scripts/braker.pl --genome=SFG\_Flye\_HAC10kb\_P2\_PD\_pilon2.fasta.masked.masked --species=Chrysomallon --bam=all.sorted.bam --cores 40

# maker configure file

#-----Genome (these are always required)

genome=SFG\_Flye\_HAC10kb\_P2\_PD\_pilon2.fasta #genome sequence (fasta file or fasta embeded in GFF3 file)

organism\_type=eukaryotic #eukaryotic or prokaryotic. Default is eukaryotic

#-----EST Evidence (for best results provide a file for at least one)

est=Trinity\_all\_0.97.fasta #set of ESTs or assembled mRNA-seq in fasta format

altest= #EST/cDNA sequence file in fasta format from an alternate organism

est\_gff= #aligned ESTs or mRNA-seq from an external GFF3 file

altest\_gff= #aligned ESTs from a closly relate species in GFF3 format

#-----Protein Homology Evidence (for best results provide a file for at least one)

protein=Mollu\_Prot\_50AA\_0.95.fa #protein sequence file in fasta format (i.e. from mutiple organisms)

protein\_gff= #aligned protein homology evidence from an external GFF3 file

#-----Repeat Masking (leave values blank to skip repeat masking)

model\_org=all #select a model organism for RepBase masking in RepeatMasker

rmlib=./SFG\_flye-families.fa #provide an organism specific repeat library in fasta format for RepeatMasker

repeat\_protein=~/App/maker-3.01.03/data/te\_proteins.fasta #provide a fasta file of transposable element proteins for RepeatRunner

rm\_gff= #pre-identified repeat elements from an external GFF3 file

prok\_rm=0 #forces MAKER to repeatmask prokaryotes (no reason to change this), 1 = yes, 0 = no

softmask=1 #use soft-masking rather than hard-masking in BLAST (i.e. seg and dust filtering)

#-----Gene Prediction

snaphmm= #SNAP HMM file

gmhmm= #GeneMark HMM file

augustus\_species=Chrysomallon #Augustus gene prediction species model

fgenesh\_par\_file= #FGENESH parameter file

pred\_gff= #ab-initio predictions from an external GFF3 file

model\_gff= #annotated gene models from an external GFF3 file (annotation pass-through)

run\_evm=1 #run EvidenceModeler, 1 = yes, 0 = no

est2genome=1 #infer gene predictions directly from ESTs, 1 = yes, 0 = no

protein2genome=1 #infer predictions from protein homology, 1 = yes, 0 = no

trna=0 #find tRNAs with tRNAscan, 1 = yes, 0 = no

snoscan\_rrna= #rRNA file to have Snoscan find snoRNAs

snoscan\_meth= #-O-methylation site fileto have Snoscan find snoRNAs

unmask=0 #also run ab-initio prediction programs on unmasked sequence, 1 = yes, 0 = no

allow\_overlap= #allowed gene overlap fraction (value from 0 to 1, blank for default)

#-----Other Annotation Feature Types (features MAKER doesn't recognize)

other\_gff= #extra features to pass-through to final MAKER generated GFF3 file

#-----External Application Behavior Options

alt\_peptide=C #amino acid used to replace non-standard amino acids in BLAST databases

cpus=20 #max number of cpus to use in BLAST and RepeatMasker (not for MPI, leave 1 when using MPI)

#-----MAKER Behavior Options

max\_dna\_len=100000 #length for dividing up contigs into chunks (increases/decreases memory usage)

min\_contig=10000 #skip genome contigs below this length (under 10kb are often useless)

pred\_flank=200 #flank for extending evidence clusters sent to gene predictors

pred\_stats=0 #report AED and QI statistics for all predictions as well as models

AED\_threshold=1 #Maximum Annotation Edit Distance allowed (bound by 0 and 1)

min\_protein=50 #require at least this many amino acids in predicted proteins

alt\_splice=0 #Take extra steps to try and find alternative splicing, 1 = yes, 0 = no

always\_complete=1 #extra steps to force start and stop codons, 1 = yes, 0 = no

map\_forward=0 #map names and attributes forward from old GFF3 genes, 1 = yes, 0 = no

keep\_preds=0 #Concordance threshold to add unsupported gene prediction (bound by 0 and 1)

split\_hit=20000 #length for the splitting of hits (expected max intron size for evidence alignments)

min\_intron=20 #minimum intron length (used for alignment polishing)

single\_exon=1 #consider single exon EST evidence when generating annotations, 1 = yes, 0 = no

single\_length=300 #min length required for single exon ESTs if 'single\_exon is enabled'

correct\_est\_fusion=0 #limits use of ESTs in annotation to avoid fusion genes

tries=2 #number of times to try a contig if there is a failure for some reason

clean\_try=0 #remove all data from previous run before retrying, 1 = yes, 0 = no

clean\_up=0 #removes theVoid directory with individual analysis files, 1 = yes, 0 = no

TMP=/home/sunj/temp #specify a directory other than the system default temporary directory for temporary files