# Genomic landscape of transposable elements and structural variation using long reads whole genome sequencing

Yocelyn T. Gutiérrez-Guerero

## Index

Installation requirements

-PacBio data requires to be processed using the SMRTlink software or PacBio tools that are distributed via Bioconda pbbioconda

##Download PacBio reads

```
fasterq-dump SRA_Accession -F fastq -o Out_pb -e 2
```

Converting BAM files to Fastq format and removing adapters using HiFiAdapterFilt *Dependencies*: python, blast, bamtools

```
pbadapterfilt.sh  -t 20 -o ~/PacBio_filter_fastq/
```

Genome assembly with Flye, including a correction with Minimap2

```
flye --pacbio-hifi ID_raw_pacbioRev.fastq.gz --genome-size 3.1g
-o PacBio_assembly/ -t 40 -i 2 --scaffold --asm-coverage 20
```

Genome assembly with Hifiasm and genome quality using quast

```
mkdir hifiasm_ID
cd hifiasm_ID/
ln -s ~/PacBio_fastq/ID_raw_pacbioRev.fastq.gz .
hifiasm -o ID_hifiasm.asm -t 30 ID_raw_pacbioRev.fastq.gz -f38

awk '/^S/{print ">"$2;print $3}' ID_hifiasm.asm.bp.p_ctg.gfa
> ID_hifiasm_ctg.fasta

quast.py ID_hifiasm_ctg.fasta -o ID_hifiasm_stats.txt
```

Scaffolding and polishing with Gapless

```
gapless.py split -o gapless_split.fa ID_hifiasm_ctg.fasta
minimap2 -t 30 -DP -k19 -w19 -m200 gapless_split.fa gapless_split.fa
> gapless_split_repeats.paf

minimap2 -t 30 -x map-hifi -c -N 5 --secondary=no gapless_split.fa
ID.hifi_reads.default.filt.fastq.gz > gapless_reads.paf

gapless.py scaffold -p gapless -s gapless_stats.pdf gapless_split.fa
gapless_reads.paf gapless_split_repeats.paf

minimap2 -t 30 -x map-hifi  <(seqtk subseq ID.hifi_reads.default.filt.fastq.gz
gapless_extending_reads.lst) <(seqtk subseq ID.hifi_reads.default.filt.fastq.gz
gapless_extending_reads.lst) > gapless_extending_reads.paf

gapless.py extend -p gapless gapless_extending_reads.paf

seqtk subseq ID.hifi_reads.default.filt.fastq.gz gapless_used_reads.lst
> temp_finish.fastq

gapless.py finish -o gapless_raw.fa -H 0 -s
gapless_extended_scaffold_paths.csv -p gapless_polishing.csv
gapless_split.fa temp_finish.fastq

minimap2 -t 30 -x map-hifi gapless_raw.fa ID.hifi_reads.default.filt.fastq.gz
> gapless_consensus.paf
```

```
rm gapless_split.fa gapless_split_repeats.paf gapless_reads.paf

gapless_stats.pdf gapless_scaffold_paths.csv gapless_extensions.csv

gapless_extending_reads.lst gapless_polishing.csv
gapless_extending_reads.paf gapless_extended_scaffold_paths.csv
gapless_used_reads.lst
gapless_extended_polishing.csv gapless_raw_polishing.paf temp_finish.fastq
```

Polishing with Minimap2, pbmm2, samtools and Racon

```
pbmm2 index ID_hifiasm_ctg.fasta ID_hifiasm_ctg.fasta.mmi
pbmm2 align ID_hifiasm_ctg.fasta ID.hifi_reads.default.filt.fastq.gz
ID_align.bam --sort -j 20

samtools view -h ID_align.bam > ID_align.sam

racon id-t 20 ID.hifi_reads.default.filt.fastq.gz ID_align.sam
ID_hifiasm_ctg.fasta > ID_assembly_racon.fasta

rm  ID_align.bam ID_align.sam ID_align.bai

##Evaluate genome assembly

quast.py ID_assembly_racon.fasta -o ID_assembly_stats
```

###Busco

```
library(ggplot2)
library(webr)
library(dplyr)

bdata<-data.frame(Category= c("Single-copy","Duplicated","Fragmented",
                              "Missing") ,Value=c(8444,158,108,516))



dbusco = bdata %>% group_by(Category, Class) %>% summarise(n = sum(Value))

pie_chart <- ggplot(bdata, aes(x = "", y = Value, fill = Category)) +
    geom_bar(stat = "identity", width = 1) +
    coord_polar(theta = "y") +
    geom_text(aes(x = 1.6, label = paste(round(Value/sum(Value) * 100), "%")),
              position = position_stack(vjust = 0.5)) +
    scale_fill_manual(values=maize_pal("FloweringTime")) +
    theme_void()

donut_chart <- pie_chart +
    geom_bar(data = bdata, aes(x = "", y = Value), stat = "identity",
             width = 0.5, fill = "white") +
    geom_text(data = bdata, aes(label = paste(round(Value/sum(Value) * 100),
                                              "%")),
              position = position_stack(vjust = 0.5), color = "white") +
    theme_void()

grid.arrange(pie_chart, donut_chart, ncol = 2, widths = c(3, 3))
```

**Ragtag:**

Connect contigs into chromosomes using homology (if a reference genome is available)

```
ragtag.py scaffold ReferenceGenome_genomic.fna ID_assembly_racon.fasta -u
-t 30
```

## Mitogenome identification and circulization

Using the House mouse mitogenome from the genome reference

```
python mitohifi.py -c ID_assembly_racon.fasta -f RefGenone_Mit.fasta
-g RefGenome_Mit.gbk -t 10 -a animal -o 2

##Verification

cat NC_005089.1_RagTag.fasta RefGenone_Mit.fasta > seq_NC_005089.fasta

mafft --anysymbol --parttree --retree 1 --thread 5 seq_NC_005089.fasta
> align_NC_005089.fasta
```

## Masking mitogenome

```
##create a bed file, indicating the coordinates:
##ptg000200l    521 16831
```

```
bedtools maskfasta -fi ID_ragtag_scaffold.fasta -bed mito.bed
-fo ID_mitomask.fasta
```

## Identify TE

### Repeat Modeler

```
BuildDatabase -name ID_db -engine ncbi ID_ragtag_scaffold.fasta

RepeatModeler -engine ncbi -pa 10 -database ID_db
```

## MCHelper

```
cat ~/busco/mammalia_odb10/hmms/*.hmm > mammals_odb10.hmm

## Automatic Curation
python3 MCHelper.py -r A -t 20 -l ID_db-families.fa -o MCHelper_out/ -g
~/hifiasm/ID_assembly_racon.fasta --input_type fasta -b mammals_odb10.hmm -a F

## Manual Curation
python3 MCHelper.py -r M -l ~/MCHelper/Curation/Pop1_3id_curated_NR.fa -o
~/MCHelper/Curation/Pop1_3id_curated --input_type fasta -g
~/PacBioData/Final_assembly/ID_assembly_racon.fasta -t 30
```

## Repeat Masker Annotation

```
RepeatMasker -pa 50 -lib MCHelperDfam_TElibrary_NR_rename.fasta -x
-gff ID_assembly_racon.fasta
```