Spider Monkey Genome Assembly and Annotation Script

Martina Albuja Gabriela Pozo[1], Quintana[1], Lizbeth Larreátegui[1], Maria de Lourdes Torres[1]

[1]Laboratorio de Biotecnología Vegetal, Colegio de Ciencias Biológicas y Ambientales, Universidad San Francisco de Quito (USFQ), Calle Diego de Robles y Avenida Pampite, Cumbayá, Quito, Ecuador

Laboratorio de Biotecnologia Vegetal - USFQ

Martina Albuja Quintana

AUG 31, 2023

OPEN ACCESS

DOI:
dx.doi.org/10.17504/protocols.io.6qpvr3892vmk/v1

Protocol Citation: Gabriela Pozo, Martina Albuja Quintana, Lizbeth Larreátegui, Maria de Lourdes Torres 2023. Spider Monkey Genome Assembly and Annotation Script. **protocols.io** https://dx.doi.org/10.17504/protocols.io.6qpvr3892vmk/v1

Protocol status: Working
We use this protocol and it's working

ABSTRACT

Oxford Nanopore long reads obtained from sequencing the DNA of an Ecuadorian brown-headed spider monkey (*Ateles fusciceps fusciceps*), were used to assemble and annotate the whole genome of this species. ONT long reads were filtered and trimmed in Nanofilt and Porechop. Sequencing statistics were visualized in Nanoplot. The reads were later processed to generate a genome assembly. Two different assemblers, Flye and Smartdenovo, were used on the raw reads to produce draft genomes. The resulting assemblies were polished in Medaka and analyzed for genome completeness and quality in Quast and BUSCO. The best resulting assembly was later annotated in Maker in 3 consecutive rounds using the *ab initio* gene predictor SNAP.

Created: Aug 30, 2023

Last Modified: Aug 31, 2023

PROTOCOL integer ID: 87182

ONT Raw Reads: Filtering, Trimming and Sequencing Statistics

1 NANOFILT

NanoFilt -q 7 < raw_reads.fastq > nanofilt_trimmed.fastq

2 PORECHOP

porechop -i nanofilt_trimmed.fastq.gz -o porechop_reads.fastq.gz

3 NANOPLOT

NanoPlot --fastq porechop_reads.fastq --readtype 1D -t 4 --title "Nanoplot_results" -o Nanoplot_results

Genome Assembly

4 SMARTdenovo

smartdenovo.pl -p input_name -c 1 'porechop_reads.fastq' > name.mak

make -f name.mak

5 Flye

flye --nano-raw porechop_reads.fastq --out-dir PATH/output_name --scaffold -g 2.6g

Genome Mapping

6 Minimap2

minimap2 -ax map-ont reference.fna.gz assembly_file > assembly_mapped.sam

7 **Samtools**

samtools view -bS assembly_mapped.sam > assembly_mapped.bam

samtools fasta assembly_mapped.bam.bam > assembly_mapped.fasta

Genome Polishing

8 **MEDAKA**

medaka_consensus -i raw_reads.fastq -d assembly_mapped.fasta -o Medaka_Folder -t 4 -m r103_fast_g507

Genome Assembly Evaluation

9 QUAST: quast.py assembly_medaka.fasta -r reference.fna.gz --eukaryote -o Quast_Output_Folder

10 BUSCO: busco -i assembly_medaka.fasta -l primates_odb10 -o BUSCO_Output_Folder -m genome

Genome Annotation

11 **REPEAT MODELER**

BuildDatabase -name Ateles_genome Ateles_fusciceps_PulidoMedaka.fasta

RepeatModeler -threads 32 -database Ateles_genome -LTRStruct >& repeatmodeler.log

12 **ASSEMBLY FILE PREPARATION**

awk '/^>/{print ">Ateles_fusciceps" ++i; next}{print}' Ateles_fusciceps_Ensamblado_Concatenado.fasta

13 **MODIFY MAKER_OPTS.CTL FILE**

14 MAKER RUN 1 (10 ITERATIONS)

sbatch --ntasks=1 -p general -A general --cpus-per-task=2 -N 1 --job-name=1_makerMono -e error_%j.err --mem=100G --out=makerMono_1.out --time=4-0 --wrap="maker"

15 MAKER RUN 2 (5 ITERATIONS)

1. MODIFY SNAP_PULT_CREATOR.SH FILE

2. sbatch --ntasks=1 -p general -A general --cpus-per-task=2 -N 1 --job-name=1_makerMono_sn1 -e error_%j.err --mem=100G --out=makerMono_sn1_1.out --time=4-0 --wrap="maker"

16 MAKER RUN 3 (5 ITERATIONS)

1. MODIFY SNAP_PULT_CREATOR.SH FILE

2. sbatch --ntasks=1 -p general -A general --cpus-per-task=2 -N 1 --job-name=1_makerMono_sn2 -e error_%j.err --mem=100G --out=makerMono_sn2_1.out --time=4-0 --wrap="maker"

17 GENERATE A SINGLE GFF AND PROTEIN AND TRANSCRIPT FILE FROM ALL 3 MAKER ROUNDS

gff3_merge -d Ateles_fusciceps_Ensamblado_Concatenado_master_datastore_index.log -o Mono_Anotado_All.gff

fasta_merge -d Ateles_fusciceps_Ensamblado_Concatenado_master_datastore_index.log -o Mono_Anotado_All.fa

18 IDENTIFY CONSERVED PROTEIN REGIONS IN PREDICTED GENE MODELS

sbatch --ntasks=1 -p general -A general --cpus-per-task=8 -N 1 --job-name=interpro_Domains --mem=100G --out=interpro_Do.out -e error_%j.err --time=4-0 --wrap="/interproscan-5.61-93.0/interproscan.sh -appl PfamA -iprlookup -goterms -f tsv -i Mono_Anotado_All.all.maker.proteins.fasta"

19 MODIFY THE ORIGINAL GFF3 FILE BY IDENTIFYING GENE MODELS WITH CONSERVED PROTEIN DOMAINS

ipr_update_gff Ateles_fusciceps_Ensamblado_Concatenado.all.gff Mono_Anotado_All.all.maker.proteins.fasta.tsv > Mono_Anotado_genomic_update.all.gff

20 ELIMINATE GENE MODELS WITH AED <0.5

./quality_filter -s Mono_Anotado_genomic_update.all.gff -a 0.5 >

Mono_Anotado_genomic_FINAL.all.gff

21 **CALCULATE ANNOTATION STATISTICS IN AGAT**

agat_sp_statistics.pl –gff Mono_Anotado_genomic_FINAL.all.gff -o Mono_Stats

22 **FILTER OUT GENE MODELS WITH NO CONSERVED PROTEIN REGIONS AND AED <0.5 FROM PROTEIN AND TRANSCRIPT FASTA FILES**

genes_from_gff.aed-0.5.ids perl ./fastaqual_select.pl -f Mono_Anotado_All.all.maker.proteins.fasta -inc genes_from_gff.aed-1.0.ids > Mono_Anotado_All_Proteins_Final.fasta

perl ./fastaqual_select.pl -f Mono_Anotado_All.all.maker.transcripts.fasta -inc genes_from_gff.aed-1.0.ids > Mono_Anotado_All_Transcripts_Final.fasta