Jan 23, 2022

Annotation of the *Albula glossodona* Genome using MAKER

Brandon D Pickett[1], Sheena Talma[2], Jessica R. Glass[3,4], Daniel Ence[5], Timothy P. Johnson[6], Paul D. Cowley[3], Perry G. Ridge[1], John S. K. Kauwe[1,7]

[1]Department of Biology, Brigham Young University;

[2]Department of Ichthyology and Fisheries Science, Rhodes University;

[3]South African Institute for Aquatic Biodiversity;

[4]College of Fisheries and Ocean Sciences, University of Alaska Fairbanks;

[5]School of Forest Resources and Conservation, University of Florida; [6]Independent;

[7]Brigham Young University - Hawai'i

dx.doi.org/10.17504/protocols.io.b3xvqpn6

GigaScience Press

Brandon Pickett

The Roundjaw Bonefish (*Albula glossodonta*) genome assembly was annotated following the MAKER pipeline. The steps we took are outlined here, but this process assumes you have successfully installed MAKER in advance. The following versions of tools were used (in alphabetical order): AUGUSTUS v3.3.2, BEDTools v2.28.0, BUSCO v4.0.6 with OrthoDB v10, EVidenceModeler v1.1.1, GeneMark-ES v4.38, gFACs v1.1.1, InterProScan v5.45-80.0, MAKER v3.01.02-beta, NCBI BLAST+ Suite v2.9.0, SNAP commit #daf76ba (3 June 2019), and tRNAscan-SE v1.3.1.

DOI

dx.doi.org/10.17504/protocols.io.b3xvqpn6

Brandon D Pickett, Sheena Talma, Jessica R. Glass, Daniel Ence, Timothy P. Johnson, Paul D. Cowley, Perry G. Ridge, John S. K. Kauwe 2022. Annotation of the Albula glossodona Genome using MAKER. **protocols.io** https://dx.doi.org/10.17504/protocols.io.b3xvqpn6

protocol

Brandon D. Pickett, Sheena Talma, Jessica R. Glass, Daniel Ence, Timothy P. Johnson, Paul D. Cowley, Perry G. Ridge, John S. K. Kauwe. Genome Assembly of the Roundjaw Bonefish (*Albula glossodonta*), a Vulnerable Circumtropical Sportfish. *GigaByte* 2022. DOI: 10.46471/gigabyte.44.

Genome Annotation, MAKER Annotation Pipeline, Albula glossodonta, Roundjaw Bonefish

protocol ,

Jan 17, 2022

Mar 07, 2022

57045

Not applicable.

Not applicable.

Not applicable.

Prepare for MAKER

1 Prepare for MAKER

Create a working directory for MAKER and then create the initial control files. Edit them to match your environment using your favorite text editor, e.g., we had to set use_rapsearch to 0 and blast_type to ncbi+ in maker_bopts.ctl. Take a careful look at maker_exe.ctl to ensure you have the paths correctly set for the requisite executables. We'll be changing maker_opts.ctl for each MAKER round.

```
mkdir maker
cd maker
maker -CTL
for FN in *.ctl; do vim ${FN}; done
```

MAKER Round #1

2 Create a working directory and copy control files

```
mkdir rnd1
cd rnd1
cp ../*.ctl .
```

3 Modify the control files

The following shows the relevant or modified lines from maker_opts.ctl:

```
# genome
```


Citation: Brandon D Pickett, Sheena Talma, Jessica R. Glass, Daniel Ence, Timothy P. Johnson, Paul D. Cowley, Perry G. Ridge, John S. K. Kauwe Annotation of the Albula glossodona Genome using MAKER https://dx.doi.org/10.17504/protocols.io.b3xvqpn6

```
genome=/path/to/scaffolds.fa
organism_type=eukaryotic

#re-annotation
maker_gff=
est_pass=0
protein_pass=0
rm_pass=0
model_pass=0
pred_pass=0
other_pass=0

# est/rna-seq
est=/path/to/Trinity/transcripts.fa
est_gff=

# protein homology
protein=/path/to/uniprot_sprot.fa
protein_gff=

# repeat masking
model_org=all
rmlib=/path/to/RepeatModeler/results/assembly-db-families.fa
repeat_protein=/path/to/maker-install-dir/data/te_proteins.fa
rm_gff=
softmask=1

# gene prediction
snaphmm=
gmhmm=
augustus_species=
pred_gff=
model_gff=
run_evm=0
est2genome=1
protein2genome=1
trna=0

# maker behavior
max_dna_len=1000000
min_contig=20000

pred_flank=200
pred_stats=0
AED_threshold=1
min_protein=0
alt_splice=0
```


Citation: Brandon D Pickett, Sheena Talma, Jessica R. Glass, Daniel Ence, Timothy P. Johnson, Paul D. Cowley, Perry G. Ridge, John S. K. Kauwe Annotation of the Albula glossodona Genome using MAKER https://dx.doi.org/10.17504/protocols.io.b3xvqpn6

```
always_complete=0
map_forward=0
keep_preds=0

split_hit=10000
min_intron=20
single_exon=0
single_length=250
correct_est_fusion=0
```

4 Run MAKER

If using MPI (we did), MAKER recommends submitting your job with mpiexec, but we found that mpirun worked.

```
mpirun maker -cpus ${CPUS} -TMP /tmp
```

5 Extract Output Fasta and GFF3

Note that this assumes your input fasta file (genome= line from maker_opts.ctl) to MAKER was named “scaffolds” (followed by .fa or .fasta). If not, you would need to modify the following because MAKER creates an output prefix based on the basename of the input fasta file. Also, we used an output prefix of agloss-rnd1 (*Albula glossodonta* round 1) for fasta_merge, which is project-specific and would be best changed for another project.

```
cd scaffolds.maker.output

fasta_merge \
    -o agloss-rnd1 \
    -d scaffolds_master_datastore_index.log

gff3_merge \
    -n -s \
    -d scaffolds_master_datastore_index.log \
    > agloss-rnd1_noSeq.gff

cd scaffolds_datastore

rename 's/.all.maker./_/' *.fasta # Perl rename, not Linux util
rename 's/fasta/fa/' *.fasta # Perl rename, not Linux util

awk '{if ($2 == "est2genome") print $0}' \
    agloss-rnd1_noSeq.gff \
    > agloss-rnd1_est2genome.gff
```


Citation: Brandon D Pickett, Sheena Talma, Jessica R. Glass, Daniel Ence, Timothy P. Johnson, Paul D. Cowley, Perry G. Ridge, John S. K. Kauwe Annotation of the Albula glossodona Genome using MAKER https://dx.doi.org/10.17504/protocols.io.b3xvqpn6

```
awk '{if ($2 == "protein2genome") print $0}' \
    agloss-rnd1_noSeq.gff \
    > agloss-rnd1_protein2genome.gff

awk '{if ($2 ~ "repeat") print $0}' \
    agloss-rnd1_noSeq.gff \
    > agloss-rnd1_repeats.gff

mv agloss-rnd1*.fa agloss-rnd1*.gff ../..

cd ../../../..
```

ab initio Gene Predictors

6 GeneMark-ES

Three *ab initio* gene prediction programs were run between MAKER rounds 1 and 2. AUGUSTUS and SNAP can take gene models as input, and they are thus able to be run with new models after rounds 1 and 2 of MAKER in preparation for rounds 2 and 3, respectively. GeneMark-ES does not take gene models as input, and it thus needs to be run only one time.

GeneMark-ES required a software key to be run, which can be obtained or re-obtained for free for academic use at any time. GeneMark-ES also requires a configuration file to be run; the default configuration file was used.

```
gmes_petap.pl \
    --ES \
    --usr_cfg ${COPY_OF_DEFAULT_CONFIG_FILE}\
    --cores ${THREADS} \
    --sequence ${SCAFFOLDS_ASSEMBLY_FILE}
```

7 AUGUSTUS

AUGUSTUS training can be handled with BUSCO.

7.1 Copy configuration files

Before AUGUSTUS can be trained, configuration files and data from AUGUSTUS and BUSCO will need to be copied to the working directory for this part of the analysis, and the relevant environment variables will need to be reset (which assumes they are properly set in the first place). No modifications to the augustus_config files were required, but we had to set the download_path= in busco_config.ini to where we downloaded OrthoDB v10.

```
cp-r ${AUGUSTUS_CONFIG_PATH} augustus_config
export AUGUSTUS_CONFIG_PATH=${PWD}/augustus_config
```


Citation: Brandon D Pickett, Sheena Talma, Jessica R. Glass, Daniel Ence, Timothy P. Johnson, Paul D. Cowley, Perry G. Ridge, John S. K. Kauwe Annotation of the Albula glossodona Genome using MAKER https://dx.doi.org/10.17504/protocols.io.b3xvqpn6

```
cp ${BUSCO_CONFIG_FILE} busco_config.ini
export BUSCO_CONFIG_FILE=${PWD}/busco_config.ini
vim busco_config.ini
```

7.2 Create working directory

```
mkdir busco-augustus
cd busco-augustus
```

7.3 Extract candidate gene regions

generateBedForMrnaExtraction.py is available on GitHub.

```
python3 generateBedForMrnaExtraction.py \
    ../maker/rnd1/agloss-rnd1_noSeq.gff \
    /path/to/scaffolds.fa \
    candidates-rnd1.bed

bedtools getfasta \
    -fi /path/to/scaffolds.fa \
    -bed candidates-rnd1.bed \
    -fo candidates-rnd1.fa
```

7.4 Run BUSCO

```
busco \
    --offline \
    --long \
    --config ${BUSCO_CONFIG_FILE} \
    --cpu ${THREADS} \
    --in candidates-rnd1.fa \
    --out agloss-rnd1 \
    --mode genome \
    --lineage actinopterygii \
    --augustus_species zebrafish

cd ..
```

7.5 Post-processing

Rename files and internal references to those files in the BUSCO retraining

Citation: Brandon D Pickett, Sheena Talma, Jessica R. Glass, Daniel Ence, Timothy P. Johnson, Paul D. Cowley, Perry G. Ridge, John S. K. Kauwe Annotation of the Albula glossodona Genome using MAKER https://dx.doi.org/10.17504/protocols.io.b3xvqpn6

parameters files and copy them to the AUGUSTUS config directory so they can be later used by MAKER.

```
# make dir for results
mkdir augustus_config/species/agloss

# move to results location
cd "busco-augustus/agloss-
rnd1/run_actinopterygii_odb10/augustus_output/retraining
_parameters/BUSCO_agloss-rnd1"

# rename some files and their references to eachother
rename \ # Perl rename, not Linux util
    's/BUSCO_(agloss-rnd1_)/$1/' \
    ./*

sed \ # gnu sed
    -i -r \
    's/BUSCO_(agloss-rnd1_)/\1/' \
    ./agloss-rnd1_parameters.cfg*

# do it again, removing the rnd info
rename \ # Perl rename, not Linux util
    's/(agloss)-rnd1)/$1/' \
    ./*

sed \ # gnu sed
    -i -r \
    's/(agloss)-rnd1/\1/' \
    ./*

# copy the files to final results location
cp -f ./*
../../../../../../augustus_config/species/agloss/

# move back to main project dir
cd –
```

8 SNAP

Training with SNAP is much less resource intensive than training AUGUSTUS. Most, if not all, of the commands can reasonably be run “locally” on a login node or other machine. The final output file, genome.hmm, is what will be provided to the next round of MAKER.

```
mkdir -p snap/rnd1
```


Citation: Brandon D Pickett, Sheena Talma, Jessica R. Glass, Daniel Ence, Timothy P. Johnson, Paul D. Cowley, Perry G. Ridge, John S. K. Kauwe Annotation of the Albula glossodona Genome using MAKER https://dx.doi.org/10.17504/protocols.io.b3xvqpn6

```
ln -s \
    ../../maker/rnd1/agloss-rnd1_withSeq.gff \
    snap/rnd1/genome.gff

cd snap/rnd1

maker2zff genome.gff

fathom \
    genome.ann genome.dna \
    -gene-stats \
    > gene-stats.log

fathom \
    genome.ann genome.dna \
    -validate \
    > validate.log

fathom \
    genome.ann genome.dna \
    -categorize 1000 \
    > categorize.log

fathom \
    uni.ann uni.dna \
    -export 1000 -plus \
    > export.log

forge \
    export.ann export.dna \
    > forge.log

hmm-assembler.pl \
    genome params \
    > genome.hmm

cd ../..
```

MAKER Round #2

9 MAKER round #2

Run MAKER to generate new gene models using the gene models produced from the *ab initio* gene predictors from the previous step. It will also run EVidenceModeler this time. Follow the same process as for MAKER Round #1, except with the following modifications:

- copy the control files from the MAKER round #1 instead of the initial control files in the

Citation: Brandon D Pickett, Sheena Talma, Jessica R. Glass, Daniel Ence, Timothy P. Johnson, Paul D. Cowley, Perry G. Ridge, John S. K. Kauwe Annotation of the Albula glossodona Genome using MAKER https://dx.doi.org/10.17504/protocols.io.b3xvqpn6

maker directory:

```
# assuming you are in maker/rnd2
cp ../rnd1/*.ctl .
```

- switch “rnd1” for “rnd2” in commands and directory/file names
- skip the awk commands in the output extraction (i.e., the commands that create the files ending in est2genome.gff, protein2genome.gff, and repeats.gff)
- make the following modifications to maker_opts.ctl (if a line isn’t shown here and it exists in the file, leave it “as is”) before running maker:

```
# est/rna-seq
est=
est_gff=/path/to/project/maker/rnd1/agloss-rnd1_est2genome.gff

# protein homology
protein=
protein_gff=/path/to/project/maker/rnd1/agloss-
rnd1_protein2genome.gff

# repeat masking
model_org=
rmlib=
repeat_protein=
rm_gff=/path/to/project/maker/rnd1/agloss-rnd1_repeats.gff

# gene prediction
snaphmm=/path/to/project/snap/rnd1/genome.hmm
gmhmm=/path/to/project/gmes/output/gmhmm.mod
augustus_species=agloss
run_evm=1
est2genome=0
protein2genome=0
```

***ab initio* Gene Predictors**

10 *ab initio* Gene Predictors

AUGUSTUS and SNAP are re-run with the gene models provided from MAKER round #2 to generate new gene models. This is different from the first round of running AUGUSTUS and SNAP which relied on gene models produced from MAKER Round #1, which created initial models directly from the RNA-seq evidence. GeneMark-ES does not accept input gene models, so it does not need to be run again.

10.1 gFACs Filtering

In an attempt to improve the quality of the gene models being used for this

Citation: Brandon D Pickett, Sheena Talma, Jessica R. Glass, Daniel Ence, Timothy P. Johnson, Paul D. Cowley, Perry G. Ridge, John S. K. Kauwe Annotation of the Albula glossodona Genome using MAKER https://dx.doi.org/10.17504/protocols.io.b3xvqpn6

final round of training with AUGUSTUS and SNAP, gFACs was employed to filter out models with single-exon genes, introns shorter than 20bp, etc.

```
mkdir -p gfacs/rnd2

ln -s \
    ../../maker/rnd2/agloss-rnd2_noSeq.gff \
    gfacs/rnd2/orig_noSeq.gff

ln -s \
    ../../assembly/scaffolds.fa \
    gfacs/rnd2/assembly.fa

awk \
    'BEGIN{x=0;}/^##FASTA/{x=1;}{if(x){print $0;}}' \
    maker/rnd2/agloss-rnd2_withSeq.gff \
    > gfacs/rnd2/orig_onlySeq.gff

cd gfacs/rnd2

gFACs.pl \
    -f "maker_2.31.9_gff" \
    -p  ./output/agloss-rnd2_noSeq \
    --statistics-at-every-step \
    --statistics \
    --rem-monoexonics \
    --min-exon-size 20 \
    --min-intron-size 20 \
    --min-CDS-size 74 \
    --fasta assembly.fa \
    --splice-table \
    --nt-content \
    --canonical-only \
    --rem-genes-without-stop-codon \
    --allowed-inframe-stop-codons 0 \
    --create-gff3 \
    --get-fasta-with-introns \
    --get-fasta-without-introns \
    --get-protein-fasta \
    --distributions \
    exon_lengths \
    intron_lengths \
    CDS_lengths \
    gene_lengths \
    exon_position \
    exon_position_data \
```


Citation: Brandon D Pickett, Sheena Talma, Jessica R. Glass, Daniel Ence, Timothy P. Johnson, Paul D. Cowley, Perry G. Ridge, John S. K. Kauwe Annotation of the Albula glossodona Genome using MAKER https://dx.doi.org/10.17504/protocols.io.b3xvqpn6

```
    intron_position \
    intron_position_data \
    -O ./output \
    orig_noSeq.gff

ln -s \
    agloss-rnd2_noSeq_out.gff3 \
    output/agloss-rnd2_noSeq.gff

cat \
    output/agloss-rnd2_noSeq.gff orig_onlySeq.gff \
    > output/agloss-rnd2_withSeq.gff

cd ../..
```

10.2 AUGUSTUS

Repeat the same process of training AUGUSTUS using BUSCO as was done previously, except with the following modifications:

- use gFACs output GFF3 file (i.e., gfacs/rnd2/output/agloss-rnd2_withSeq.gff) as input instead of the output GFF3 from MAKER
- use augustus_species=agloss instead of augustus_species=zebrafish when running busco,
- use “rnd2” instead of “rnd1” throughout the various commands when referring to directory and file names.

Don’t forget to run the post-processing steps, which will replace the contents of augustus_config/species/agloss.

10.3 SNAP

Repeat the same process of running SNAP as was done previously, except with the following modifications:

- use gFACs output GFF3 file (i.e., gfacs/rnd2/output/agloss-rnd2_withSeq.gff) as input instead of the output GFF3 from MAKER
- modify SNAP’s maker2zff command by replacing “CDS” with “exon” on line 142 (you can change it back after you are done)
- use “rnd2” instead of “rnd1” throughout the various commands when referring to directory and file names.

MAKER Round #3

11 MAKER Round #3

Run MAKER to generate new gene models using the gene models produced from the *ab initio* gene predictors from the previous step. Follow the same process as for MAKER round #2, except with the following modifications:

Citation: Brandon D Pickett, Sheena Talma, Jessica R. Glass, Daniel Ence, Timothy P. Johnson, Paul D. Cowley, Perry G. Ridge, John S. K. Kauwe Annotation of the Albula glossodona Genome using MAKER https://dx.doi.org/10.17504/protocols.io.b3xvqpn6

- copy the control files from the MAKER round #2 instead of from round #1
- switch “rnd2” for “rnd3” in commands and directory/file names
- make the following modifications to maker_opts.ctl (if a line isn’t shown here and it exists in the file, leave it “as is”) before running maker:

```
# gene prediction
snaphmm=/path/to/project/snap/rnd2/genome.hmm
trna=1
```

Note that you will again be skipping the awk commands in the output extraction (i.e., the commands that create the files ending in est2genome.gff, protein2genome.gff, and repeats.gff).

Functional Annotation

12 MAKER Post-processing & Functional Annotation

The structural annotations created by MAKER required some modest post-processing before adding functional annotations. MAKER accessory scripts were used to update sequence names from the long MAKER names to friendlier ones. Other MAKER scripts were used to update the fasta and/or gff3 files with functional annotations found with the BLAST+ Suite and InterProScan.

```
# create and move to a working dir
mkdir maker/post
cd maker/post

# copy the requisite output files
cp ../rnd3/*.gff ../rnd3/*.fa .
cp ../rnd1/agloss-rnd1_{repeats,{est,protein}2genome}.gff .

# remove the rnd info
rename \ # Perl version, not Linux util
 's/-rnd[1-3]//' \
 *.fa *.gff

# map new ids to MAKER names
NUM_SEQS=`grep -Ev '^#' agloss_noSeq.gff \
 | cut -d $'\t' -f 9 | tr ';' '\n' \
 | cut -d '=' -f 2 | sort -u | wc -l`

maker_map_ids \
 --initial=1 \
 --prefix=Albula-glossodonta \
 --suffix='-?%' \
 --iterate=1 \
```


Citation: Brandon D Pickett, Sheena Talma, Jessica R. Glass, Daniel Ence, Timothy P. Johnson, Paul D. Cowley, Perry G. Ridge, John S. K. Kauwe Annotation of the Albula glossodona Genome using MAKER https://dx.doi.org/10.17504/protocols.io.b3xvqpn6

```
 --justify=${#NUM_SEQS} \
 agloss_withSeq.gff \
 > identifiers_map.tsv

# rename based on new ids
for FASTA in *.fa
do
 cp -f "${FASTA}" "${FASTA%.fa}_renamed.fa"
 map_fasta_ids identifiers_map.tsv "${FASTA%.fa}_renamed.fa"
done

for GFF in *.gff
do
 cp -f "${GFF}" "${GFF%.gff}_renamed.gff"
 map_gff_ids identifiers_map.tsv "${GFF%.gff}_renamed.gff"
done

# prep for functional annotation
cd /path/to/swissprot

makeblastdb \
 -dbtype prot \
 -in uniprot_sprot.fa \
 -input_type fasta \
 -title uniprot_sprot \
 -hash_index \
 -out uniprot_sprot \
 -logfile uniprot_sprot_makeblastdb.log

cd –

# do the alignment for func. annot.
blastp \
 -task blastp \
 -query proteins_renamed.fa \
 -db /path/to/swissprot/uniprot_sprot \
 -num_threads ${THREADS} \
 -max_target_seqs 1 \
 -max_hsps 1 \
 -evalue 1e-6 \
 -outfmt 6 \
 -out proteins-x-uniprotSprot_fmt6.tsv

# update the fasta and gff files with func. annots.
for FASTA in *_renamed.fa
do
 maker_functional_fasta \
```


Citation: Brandon D Pickett, Sheena Talma, Jessica R. Glass, Daniel Ence, Timothy P. Johnson, Paul D. Cowley, Perry G. Ridge, John S. K. Kauwe Annotation of the Albula glossodona Genome using MAKER https://dx.doi.org/10.17504/protocols.io.b3xvqpn6

```
  /path/to/swissprot/unitprot_sprot.fa \
  proteins-x-uniprotSprot_fmt6.tsv \
  ${FASTA} \
  > ${FASTA%.fa}_putative-function.fa
done

for GFF in *_renamed.gff
do
 maker_functional_gff \
  /path/to/swissprot/unitprot_sprot.fa \
  proteins-x-uniprotSprot_fmt6.tsv \
  ${GFF} \
  > ${GFF%.gff}_putative-function.gff
done

# run interproscan for more func. annots.
interproscan.sh \
 -m "standalone" \
 -cpu ${THREADS} \
 -T /tmp \
 -appl "pfam" \
 -dp \
 -f "TSV" \
 -goterms \
 -iprlookup \
 -pa \
 -t "p" \
 -i proteins_renamed.fa \
 -o proteins-interproscan.tsv

# update the gff files with interproscan results
for GFF in {with,no}Seq_renamed_putative-function.gff
do
 ipr_update_gff \
  ${GFF} \
  proteins-interproscan.tsv \
  > ${GFF%.gff}_domain-added.gff
done

for GFF in {with,no}Seq_renamed.gff
do
 iprscan2gff3 \
  proteins-interproscan.tsv \
  ${GFF} \
  > ${GFF%.gff} _visible-iprscan-domains.gff
done
```


Citation: Brandon D Pickett, Sheena Talma, Jessica R. Glass, Daniel Ence, Timothy P. Johnson, Paul D. Cowley, Perry G. Ridge, John S. K. Kauwe Annotation of the Albula glossodona Genome using MAKER https://dx.doi.org/10.17504/protocols.io.b3xvqpn6

```
cd ../..
```


Citation: Brandon D Pickett, Sheena Talma, Jessica R. Glass, Daniel Ence, Timothy P. Johnson, Paul D. Cowley, Perry G. Ridge, John S. K. Kauwe Annotation of the Albula glossodona Genome using MAKER https://dx.doi.org/10.17504/protocols.io.b3xvqpn6