# Hidden genomic features of an invasive malaria vector, *Anopheles stephensi*

Mahul Chakraborty1\*, Arunachalam Ramaiah1,2,3\*, Adriana Adolfi4, Paige Halas4, Bhagyashree Kaduskar2,3, Luna Thanh Ngo1, Suvratha Jayaprasad5, Kiran Paul5, Saurabh Whadgar5, Subhashini Srinivasan3,5, Suresh Subramani3,6,7, Ethan Bier2,7, Anthony A. James4,7,8, J.J. Emerson1,9,#

1Department of Ecology and Evolutionary Biology, University of California, Irvine, CA 92697, USA

2Section of Cell and Developmental Biology, University of California, San Diego, La Jolla, CA 92093-0335, USA

3Tata Institute for Genetics and Society, Center at inStem, Bangalore, KA 560065, India

4Department of Microbiology & Molecular Genetics, University of California, Irvine, CA 92697, USA

5Institute of Bioinformatics and Applied Biotechnology, Bangalore, KA 560100, India

6Section of Molecular Biology, University of California, San Diego, La Jolla, CA 92093-0322, USA

7Tata Institute for Genetics and Society, University of California, San Diego, La Jolla, CA 92093-0335, USA

8Department of Molecular Biology & Biochemistry, University of California, Irvine, CA 92697, USA

9Center for Complex Biological Systems, University of California, Irvine, CA 92697, USA

\*These authors contributed equally to this work

#Correspondence to: J.J. Emerson, jje@uci.edu

## 

## 

## **Computational Pipeline for Analyzing and Annotating the *An. stephensi* Genome.**

**Microbial contigs detection**

Kraken2 was used to identify the microbial contigs in our whole-genome assembly of *An. stephensi*.

module load kraken2/2.0.7-beta blast/2.8.1

#download taxonomy from public database

kraken2-build --download-taxonomy --db custom\_db\_final

#library download from public database

kraken2-build --download-library bacteria --db custom\_db\_final

kraken2-build --download-library archaea --db custom\_db\_final

kraken2-build --download-library viral --db custom\_db\_final

kraken2-build --download-library fungi --db custom\_db\_final

kraken2-build --download-library protozoa --db custom\_db\_final

kraken2-build --download-library UniVec\_Core --db custom\_db\_final

#incorporating manually downloaded 24 mosquito genomes from the public database

find mosquito\_genomes\_ncbi/ -name '\*.fasta' -print0 | xargs -0 -I{} -n1 kraken2-build --add-to-library {} --db custom\_db\_final

#building database

kraken2-build --build --db custom\_db\_final

#classification of contigs

kraken2 --db custom\_db\_final a2.pilon.fasta --use-names >custom\_db\_final\_output

kraken2 --db custom\_db\_final a2.pilon.fasta --use-names --classified-out a2\_custom\_db\_final\_classified --unclassified-out a2\_custom\_db\_final\_unclassified

**Pipeline for Iso-Seq data analysis**

module load smrtanalysis/7.0.0

##CCS analysis - step-1

#ccs --num-threads $NSLOTS /dfs3/jje-lab/aramaiah/anstep\_contigs/data/processed/Isoseq/Females\_324hPMB\_8pM/4\_D01/m54284\_190404\_122837.subreads.bam /dfs3/jje-lab/aramaiah/anstep\_contigs/data/processed/Isoseq/Females\_324hPMB\_8pM/4\_D01/m54284\_190404\_122837.ccs.bam --reportFile=/dfs3/jje-lab/aramaiah/anstep\_contigs/data/processed/Isoseq/Females\_324hPMB\_8pM/4\_D01/m54284\_190404\_122837.ccs.report --noPolish --minPasses 1

##LIMA analysis - step-2

#lima --isoseq --dump-clips --no-pbi /dfs3/jje-lab/aramaiah/anstep\_contigs/data/processed/Isoseq/Females\_324hPMB\_8pM/4\_D01/m54284\_190404\_122837.ccs.bam /data/users/aramaiah/aramaiah/anstep\_contigs/data/processed/Isoseq/primers.fasta /dfs3/jje-lab/aramaiah/anstep\_contigs/data/processed/Isoseq/Females\_324hPMB\_8pM/4\_D01/m54284\_190404\_122837.fl.bam

##ISO-SEQ3 analysis - step-3

#Refine

#isoseq3 refine /dfs3/jje-lab/aramaiah/anstep\_contigs/data/processed/Isoseq/Females\_324hPMB\_8pM/4\_D01/m54284\_190404\_122837.fl.primer\_5p--primer\_3p.bam /data/users/aramaiah/aramaiah/anstep\_contigs/data/processed/Isoseq/primers.fasta /dfs3/jje-lab/aramaiah/anstep\_contigs/data/processed/Isoseq/Females\_324hPMB\_8pM/4\_D01/m54284\_190404\_122837.flnc.bam

#Cluster

#isoseq3 cluster /dfs3/jje-lab/aramaiah/anstep\_contigs/data/processed/Isoseq/Females\_324hPMB\_8pM/4\_D01/m54284\_190404\_122837.flnc.bam /dfs3/jje-lab/aramaiah/anstep\_contigs/data/processed/Isoseq/Females\_324hPMB\_8pM/4\_D01/m54284\_190404\_122837.unpolished.new.bam --verbose

#Polish

#isoseq3 polish --num-threads $NSLOTS /dfs3/jje-lab/aramaiah/anstep\_contigs/data/processed/Isoseq/Females\_324hPMB\_8pM/4\_D01/m54284\_190404\_122837.unpolished.new.bam /dfs3/jje-lab/aramaiah/anstep\_contigs/data/processed/Isoseq/Females\_324hPMB\_8pM/4\_D01/m54284\_190404\_122837.subreads.bam /dfs3/jje-lab/aramaiah/anstep\_contigs/data/processed/Isoseq/Females\_324hPMB\_8pM/4\_D01/m54284\_190404\_122837.polished.new.bam --verbose

#Summary

#isoseq3 summarize /dfs3/jje-lab/aramaiah/anstep\_contigs/data/processed/Isoseq/Females\_324hPMB\_8pM/4\_D01/m54284\_190404\_122837.polished.new.bam /dfs3/jje-lab/aramaiah/anstep\_contigs/data/processed/Isoseq/Females\_324hPMB\_8pM/4\_D01/m54284\_190404\_122837.summary.new.csv

**Pipeline for *An. stephensi* Genome Annotation using MAKER2**

The gene models (protein coding sequence) of the *An. stephensi* genome was annotated using MAKER2 on UCI (University of California, Irvine) high power computing cluster (HPC).

As the MAKER2 is solely an annotation pipeline, it uses existing gene prediction tools (i.e. SNAP, AUGUSTUS, etc.) and integrates their output to produce the best possible gene model based on evidence alignments and/or de novo gene predictions. Our pipeline involved following steps for annotation;

i) identify and mask out repeat elements in the genome,

ii) align ESTs/RNA long molecule isoforms from six samples (merged),

iii) align an alternative transcripts and proteins from *An. gambiae* and *An. funestus* to the genome,

iv) produce *ab initio* gene predictions using inbuilt SNAP and AUGUSTUS,

v) synthesize these data into final annotation,

vi) assess the quality of final annotation using AED (annotation edit distance) and Pfam approaches, and

vii) perform the functional annotation of gene/protein models through a homology BLAST search to UniProt-Sprot database, while the domains in the annotated proteins were searched against InterProScan database.

We optimized the pipeline from the following sites to run MAKER2 for annotating the genome; i) <http://weatherby.genetics.utah.edu/MAKER/wiki/index.php/MAKER_Tutorial_for_WGS_Assembly_and_Annotation_Winter_School_2018#Add_functional_annotations_and_meta-data>, ii) <https://przeworskilab.com/wp-content/uploads/acropora-millepora-assembly.pdf>

iii) <https://wiki.cyverse.org/wiki/display/TUT/Training+ab+initio+Gene+Predictors+for+MAKER+genome+annotation>

**1. MAKER2 annotation**

**1.1. First round:**

Collectively MAKER2 was run for three rounds to predict gene and protein models. The first round of gene prediction was performed directly from transcript and protein evidence. RNA long molecule sequencing data that we generated in six samples of *An. stephensi* were used as evidence for transcripts. In addition, existing transcriptome and peptide sequence data from both *An. gambiae* and *An. funestus* were used as alternative evidence to support the predicted gene and protein models, respectively.

Step-1: #To generate three control files: maker\_opts.ctl, maker\_exe.ctl, maker\_bopts.ctl

maker -CTL

Step-2: #Modify and provide the path for required input data in the maker\_opts.ctl file (used vi editor for editing the file; i.e. vi maker\_opts.ctl)

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

genome=/data/users/aramaiah/aramaiah/anstep\_contigs/data/processed/hi\_c2/reviewed/anstep\_genomev2\_and\_haplotypev2.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=/data/users/aramaiah/aramaiah/anstep\_contigs/data/processed/Isoseq/maker2\_6samples\_merged/isoforms\_6samples\_merged.new.hq\_edited.fasta #set of ESTs or assembled mRNA-seq in fasta format

altest=/data/users/aramaiah/aramaiah/anstep\_contigs/data/processed/Isoseq/maker2\_6samples\_merged/An-gam\_An-fun\_transcipts\_merged.fasta #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 closely related species in GFF3 format

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

protein=/data/users/aramaiah/aramaiah/anstep\_contigs/data/processed/Isoseq/Females\_324hPMB\_8pM/4\_D01/maker/An-gam\_An-fun\_ref\_proteins\_merged.fasta #protein sequence file in fasta format (i.e. from multiple 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= /data/users/aramaiah/aramaiah/anstep\_contigs/data/processed/Isoseq/maker2\_6samples\_merged/asteph-te-families.fa #provide an organism specific repeat library in fasta format for RepeatMasker

repeat\_protein=/data/apps/maker/2.31.8/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= #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)

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

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

Step-3: #To reset an more strength blast mapping parameters, modify the maker\_bopts.ctl file

#-----BLAST and Exonerate Statistics Thresholds

pcov\_blastn=0.85 #Blastn Percent Coverage Threshold EST-Genome Alignments

pid\_blastn=0.95 #Blastn Percent Identity Threshold EST-Genome Alignments

eval\_blastn=1e-10 #Blastn eval cutoff

bit\_blastn=40 #Blastn bit cutoff

depth\_blastn=0 #Blastn depth cutoff (0 to disable cutoff)

pcov\_blastx=0.5 #Blastx Percent Coverage Threshold Protein-Genome Alignments

pid\_blastx=0.5 #Blastx Percent Identity Threshold Protein-Genome Alignments

eval\_blastx=1e-06 #Blastx eval cutoff

bit\_blastx=30 #Blastx bit cutoff

depth\_blastx=0 #Blastx depth cutoff (0 to disable cutoff)

Step-4: #Run MAKER2

maker 2> maker.error --ignore\_nfs\_tmp

(or)

/data/apps/anaconda\_python/2.0.1/bin/mpiexec -n $NSLOTS maker -base anstepV2\_maker\_rnd1 maker\_opts.ctl maker\_bopts.ctl maker\_exe.ctl --ignore\_nfs\_tmp #alternative command; choose one of these.

#First round output

MAKER has created an output directory, named anstepV2\_maker\_rnd1.maker.output. This run synthesized 12,324 gene models.

**1.2. Second round:**

In the second round, *ab initio* gene predictions were performed using SNAP and AUGUSTUS tools in MAKER2. The output anstepV2\_maker\_rnd1.maker.output.log of the first round was used to extract the EST, protein and repeat in an individual .gff files, which were provided as input files for this second round.

Step-1: Merging .gff from all contigs folder and extracting the est, protein and repeat mapped GFF files from anstepV2\_maker\_rnd1.maker.output.log

/data/apps/maker/2.31.8/bin/gff3\_merge -d anstepV2\_maker\_rnd1\_master\_datastore\_index.log #it will output anstepV2\_maker\_rnd1.all.gff file.

awk '{if ($2 == "est2genome") print $0}' anstepV2\_maker\_rnd1.all.gff > all\_rnd1.all.maker.est2genome.gff

awk '{if ($2 == "protein2genome") print $0}' anstepV2\_maker\_rnd1.all.gff > all\_rnd1.all.maker.protein2genome.gff

awk '{if ($2 ~ "repeat") print $0}' anstepV2\_maker\_rnd1.all.gff > all\_rnd1.all.maker.repeats.gff

#Alternatively, to get all the merged fasta seq, the following command can be used. But this will significantly take a long time to run. Above-mentioned GFF files will reduce the running time.

/data/apps/maker/2.31.8/bin/fasta\_merge -d anstepV2\_maker\_rnd1\_master\_datastore\_index.log

Step-2: To run SNAP gene prediction, first create a directory 'snap1' and get into it.

mkdir snap1 && cd snap1

#Convert GFF3 gene models to ZFF format.

/data/apps/maker/2.31.8/bin/maker2zff ../anstepV2\_maker\_rnd1.all.gff #This will output two files genome.ann genome.dna

#SNAP to validate the .gff file if it has any error.

/data/apps/snap/20131129/fathom genome.ann genome.dna -validate > snap\_rnd1\_validate\_output.txt

#If there is any error, you can see them using this cmd

cat snap\_rnd1\_validate\_output.txt | grep "error"

#If there is any error, remove them using this command. Here "MODELXXXX" has error with internal stop codon

grep -vwE "MODELXXXX" genome.ann > genome.ann2

#Rerun the fathom in SNAP to check if the file is error free

/data/apps/snap/20131129/fathom genome.ann2 genome.dna -validate > snap\_rnd1\_validate\_output\_.txt

#Now the files are error free and make input files for training SNAP

/data/apps/snap/20131129/fathom -categorize 1000 genome.ann2 genome.dna

/data/apps/snap/20131129/fathom -export 1000 -plus uni.ann uni.dna

/data/apps/snap/20131129/forge export.ann export.dna

#Now we have files to train SNAP with a hmm-assembler

/data/apps/snap/20131129/hmm-assembler.pl snap1 . > snap1.hmm

Step-3: We need to run MAKER again with the new HMM file we just built for SNAP. Also provide Augustus species name as 'aedes'. Now both SNAP and AUGUSTUS can run simultaneously.

#Create three maker \*.ctl files.

maker -CTL

#Edit the maker\_opts.ctl file to include the snap1 hmm file.

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

genome=/home/aramaiah/maker\_phase2/maker2\_6samples\_merged/anstep\_genomev2\_and\_haplotypev2.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= #set of ESTs or assembled mRNA-seq in fasta format

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

est\_gff=/home/aramaiah/maker\_phase2/maker2\_6samples\_merged/anstepV2\_maker\_rnd1.maker.output/all\_rnd1.all.maker.est2genome.gff #aligned ESTs or mRNA-seq from an external GFF3 file

altest\_gff= #aligned ESTs from a closely related species in GFF3 format

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

protein= #protein sequence file in fasta format (i.e. from mutiple oransisms)

protein\_gff=/home/aramaiah/maker\_phase2/maker2\_6samples\_merged/anstepV2\_maker\_rnd1.maker.output/all\_rnd1.all.maker.protein2genome.gff #aligned protein homology evidence from an external GFF3 file

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

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

rmlib= #provide an organism specific repeat library in fasta format for RepeatMasker

repeat\_protein= #provide a fasta file of transposable element proteins for RepeatRunner

rm\_gff=/home/aramaiah/maker\_phase2/maker2\_6samples\_merged/anstepV2\_maker\_rnd1.maker.output/all\_rnd1.all.maker.repeats.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=/home/aramaiah/maker\_phase2/maker2\_6samples\_merged/anstepV2\_maker\_rnd1.maker.output/snap1/snap1.hmm #SNAP HMM file

gmhmm= #GeneMark HMM file

augustus\_species=aedes #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)

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

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

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

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

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

#Run maker with one the following command

maker 2> maker.error --ignore\_nfs\_tmp

maker -base anstepV2\_maker\_snap1 --ignore\_nfs\_tmp #alternative command

#Second round output

MAKER has created an output directory, named anstepV2\_maker\_rnd2.maker.output. This run synthesized 13,013 gene models.

**1.3. Third round:**

We ran another round of training for SNAP and AUGUSTUS using the output of the second round. This resulted in 14966 predicted gene models.

**2. Functional annotation**

After the assessment of MAKER2 annotated genes, we performed functional annotation of gene/protein models through a homology BLAST search to UniProt-Sprot database, while the domains in the annotated proteins were searched against InterProScan database.

Step-1: Download the databases: Uniprot-SwissProt (Oct 2019) and InterProScan v5.38

wget ftp://ftp.uniprot.org/pub/databases/uniprot/current\_release/knowledgebase/complete/uniprot\_sprot.fasta.gz #once the download is completed, then unzip it (i.e. gunzip uniprot\_sprot.fasta.gz)

wget ftp://ftp.ebi.ac.uk/pub/software/unix/iprscan/5/5.38-76.0/interproscan-5.38-76.0-64-bit.tar.gz #once the download is completed, then unzip it

Step-2: Prepare the protein database for blast searches

makeblastdb -in uniprot\_sprot.fasta -dbtype prot

#Running BLASTp against UniProt/Swiss-Prot

/home/aramaiah/miniconda3/bin/blastp -query /home/aramaiah/maker\_phase2/maker2\_6samples\_merged/anstepV2\_maker\_rnd1.maker.output/snap2/anstepV2\_maker\_rnd2.maker.output/anstepV2\_maker\_rnd2.all.maker.proteins.fasta -db uniprot\_sprot.fasta -num\_threads 64 -max\_hsps 1 -max\_target\_seqs 1 -outfmt 6 -evalue 1e-2 -out output.blastp

#Running InterProScan

/home/aramaiah/maker\_phase2/maker2\_6samples\_merged/anstepV2\_maker\_rnd1.maker.output/maker\_annotation/interproscan-5.38-76.0/interproscan.sh -appl pfam -dp -f TSV -goterms -iprlookup -pa -t p -i /home/aramaiah/maker\_phase2/maker2\_6samples\_merged/anstepV2\_maker\_rnd1.maker.output/snap2/anstepV2\_maker\_rnd2.maker.output/anstepV2\_maker\_rnd2.all.maker.proteins.fasta -o output.iprscan

Step-3: MAKER2 output the files with ugly gene/protein names. MAKER2 has many number of accessory scripts, which can be used to rename the genes/proteins in standardized formats suggested by NCBI (organism prefix & gene number)

maker\_map\_ids --prefix ANSTEP-UCI\_ --abrv\_gene GENE\_ --abrv\_tran TRAN\_ --justify 8 /home/aramaiah/maker\_phase2/maker2\_6samples\_merged/anstepV2\_maker\_rnd1.maker.output/snap2/anstepV2\_maker\_rnd2.maker.output/anstepV2\_maker\_rnd2.all.gff > anstepV2\_contig.map #this output has two columns translating old gene and mRNA names to new more standardized names

#The following three scripts were used to give new names to any files containing the old names. As these scripts do in-place change of names, we have copied the files first

cp anstepV2\_maker\_rnd2.all.gff anstepV2\_maker\_rnd2.all.renamed.gff

cp anstepV2\_maker\_rnd2.all.maker.transcripts.fasta anstepV2\_maker\_rnd2.all.maker.transcripts.renamed.fasta

cp anstepV2\_maker\_rnd2.all.maker.proteins.fasta anstepV2\_maker\_rnd2.all.maker.proteins.renamed.fasta

cp output.iprscan output.renamed.iprscan

cp output.blastp output.renamed.blastp

map\_gff\_ids anstepV2\_contig.map anstepV2\_maker\_rnd2.all.renamed.gff

map\_fasta\_ids anstepV2\_contig.map anstepV2\_maker\_rnd2.all.maker.transcripts.renamed.fasta

map\_fasta\_ids anstepV2\_contig.map anstepV2\_maker\_rnd2.all.maker.proteins.renamed.fasta

map\_data\_ids anstepV2\_contig.map output.renamed.iprscan

map\_data\_ids anstepV2\_contig.map output.renamed.blastp

#Once we copied required files, added the putative gene functions using output in a BLAST run of our transcripts/proteins against UniProt/Swiss-Prot using the following scripts

maker\_functional\_gff uniprot\_sprot.fasta output.renamed.blastp anstepV2\_maker\_rnd2.all.renamed.gff > anstepV2\_maker\_rnd2.all.renamed.putative\_function.gff #this adds functions from BLAST output to GFF3 file

maker\_functional\_fasta uniprot\_sprot.fasta output.renamed.blastp anstepV2\_maker\_rnd2.all.maker.transcripts.renamed.fasta > anstepV2\_maker\_rnd2.all.maker.transcripts.renamed.putative\_function.fasta #this adds functions from BLAST output to FASTA file

maker\_functional\_fasta uniprot\_sprot.fasta output.renamed.blastp anstepV2\_maker\_rnd2.all.maker.proteins.renamed.fasta > anstepV2\_maker\_rnd2.all.maker.proteins.renamed.putative\_function.fasta #this adds functions from BLAST output to FASTA file

#To check if functions were added to the MAKER2 annotated genes (but not to non-overlapping *ab initio* SNAP/AUGUSTUS predicted genes)

awk '{if ($2=="maker") print $0}' anstepV2\_maker\_rnd2.all.renamed.putative\_function.domain\_added.gff | less

Step-4: Last, we added protein domain information to the final annotations using InterProScan data

ipr\_update\_gff anstepV2\_maker\_rnd2.all.renamed.putative\_function.gff output.renamed.iprscan > anstepV2\_maker\_rnd2.all.renamed.putative\_function.domain\_added.gff #adds searchable tags to the gene and mRNA features in the GFF3 file

iprscan2gff3 output.renamed.iprscan anstepV2\_maker\_rnd2.all.renamed.gff > visible\_iprscan\_domains.gff #adds physical viewable features for domains, which can be displayed in our JBrowse based ANSTEP-Web Server

#Extract different prediction results from master GFF file

awk '{if ($2 == "maker") print $0}' anstepV2\_maker\_rnd2.all.renamed.putative\_function.domain\_added.gff > anstepV2\_maker\_rnd2.all.renamed.putative\_function.domain\_added.MAKER.gff # shows only MAKER2 synthesized GOLD standard gene annotation

awk '{if ($2 == "est\_gff:est2genome") print $0}' anstepV2\_maker\_rnd2.all.renamed.putative\_function.domain\_added.gff > anstepV2\_maker\_rnd2.all.renamed.putative\_function.domain\_added.EST2GENOME.gff # shows transcripts evidence based annotation

awk '{if ($2 == "protein\_gff:protein2genome") print $0}' anstepV2\_maker\_rnd2.all.renamed.putative\_function.domain\_added.gff > anstepV2\_maker\_rnd2.all.renamed.putative\_function.domain\_added.PROTEIN2GENOME.gff # shows peptides evidence based annotation

awk '{if ($2 == "augustus\_masked") print $0}' anstepV2\_maker\_rnd2.all.renamed.putative\_function.domain\_added.gff > anstepV2\_maker\_rnd2.all.renamed.putative\_function.domain\_added.AUGUSTUS.gff # shows AUGUSTUS reports

awk '{if ($2 == "snap\_masked") print $0}' anstepV2\_maker\_rnd2.all.renamed.putative\_function.domain\_added.gff > anstepV2\_maker\_rnd2.all.renamed.putative\_function.domain\_added.SNAP.gff # shows SNAP reports

awk '{if ($2 ~ "repeat") print $0}' anstepV2\_maker\_rnd2.all.renamed.putative\_function.domain\_added.gff > anstepV2\_maker\_rnd2.all.renamed.putative\_function.domain\_added.REPEATS.gff # shows REPEATS reports

**3. Functional annotation for non-overlapping *ab initio* (SNAP and AUGUSTUS) predicted genes.**

MAKER2 provides transcript and protein annotated results in GFF and FASTA formats, which is based on the evidence based genes predictions and *ab initio* predicted genes. In addition, MAKER2 also provides non-overlapping genes solely from *ab initio* predictions (in our case it is from SNAP and AUGUSTUS) that were not overlapped with evidence based genes. A total of 14,192 non-overlapping *ab initio* proteins were predicted and functionally annotated using the following methods, which is almost similar to Section-3 annotation pipeline, however there are some exceptions.

Step-1: Download the databases: Uniprot-SwissProt (Oct 2019) and InterProScan v5.38

wget ftp://ftp.uniprot.org/pub/databases/uniprot/current\_release/knowledgebase/complete/uniprot\_sprot.fasta.gz #once the download is completed, then unzip it (i.e. gunzip uniprot\_sprot.fasta.gz)

wget ftp://ftp.ebi.ac.uk/pub/software/unix/iprscan/5/5.38-76.0/interproscan-5.38-76.0-64-bit.tar.gz #once the download is completed, then unzip it

Step-2: Prepare the protein database for blast searches

makeblastdb -in uniprot\_sprot.fasta -dbtype prot

#Running BLASTp against UniProt/Swiss-Prot; Running InterProScan

/home/aramaiah/miniconda3/bin/blastp -query anstepV2\_maker\_rnd2.all.maker.non\_overlapping\_ab\_initio.proteins.fasta -db ../uniprot\_sprot.fasta -num\_threads 64 -max\_hsps 1 -max\_target\_seqs 1 -outfmt 6 -evalue 1e-2 -out output.blastp

/home/aramaiah/maker\_phase2/maker2\_6samples\_merged/anstepV2\_maker\_rnd1.maker.output/maker\_annotation/interproscan-5.38-76.0/interproscan.sh -appl pfam -dp -f TSV -goterms -iprlookup -pa -t p -i anstepV2\_maker\_rnd2.all.maker.non\_overlapping\_ab\_initio.proteins.fasta -o output.iprscan

Step-3: MAKER2 output the files with ugly gene/protein names. MAKER2 has many number of accessory scripts, which can be used to rename the genes/proteins in standardized formats suggested by NCBI (organism prefix & gene number)

maker\_map\_ids --prefix ANSTEP-UCI\_ --abrv\_gene GENE\_ --abrv\_tran TRAN\_ --justify 8 anstepV2\_maker\_rnd2.all.gff > anstepV2\_contig.map

#As the fasta headers in these non-overlapping proteins/transcripts is not exactly matching with GFF file reports, we changed the word ‘-processed-’ to ‘-abinit-’fasta in FASTA headers and BLAST and InterProScan output files. This would help to reflect the same in GFF file.

sed 's/\-processed-/\-abinit-/g' anstepV2\_maker\_rnd2.all.maker.non\_overlapping\_ab\_initio.proteins.fasta > anstepV2\_maker\_rnd2.all.maker.non\_overlapping\_ab\_initio.proteins\_edit.fasta

sed 's/\-processed-/\-abinit-/g' anstepV2\_maker\_rnd2.all.maker.non\_overlapping\_ab\_initio.transcripts.fasta > anstepV2\_maker\_rnd2.all.maker.non\_overlapping\_ab\_initio.transcripts\_edit.fasta

sed 's/\-processed-/\-abinit-/g' output.iprscan > output\_edit.iprscan

sed 's/\-processed-/\-abinit-/g' output.blastp > output\_edit.blastp

#The following three scripts were used to give new names to any files containing the old names. As these scripts do in-place change of names, we have copied the files first

cp anstepV2\_maker\_rnd2.all.gff anstepV2\_maker\_rnd2.all.renamed.gff

cp anstepV2\_maker\_rnd2.all.maker.non\_overlapping\_ab\_initio.transcripts\_edit.fasta anstepV2\_maker\_rnd2.all.maker.non\_overlapping\_ab\_initio.transcripts.renamed.fasta

cp anstepV2\_maker\_rnd2.all.maker.non\_overlapping\_ab\_initio.proteins\_edit.fasta anstepV2\_maker\_rnd2.all.maker.non\_overlapping\_ab\_initio.proteins.renamed.fasta

cp output\_edit.iprscan output.renamed.iprscan

cp output\_edit.blastp output.renamed.blastp

map\_gff\_ids anstepV2\_contig.map anstepV2\_maker\_rnd2.all.renamed.gff

map\_fasta\_ids anstepV2\_contig.map anstepV2\_maker\_rnd2.all.maker.non\_overlapping\_ab\_initio.transcripts.renamed.fasta

map\_fasta\_ids anstepV2\_contig.map anstepV2\_maker\_rnd2.all.maker.non\_overlapping\_ab\_initio.proteins.renamed.fasta

map\_data\_ids anstepV2\_contig.map output.renamed.iprscan

map\_data\_ids anstepV2\_contig.map output.renamed.blastp

#Once we copied required files, added the putative gene functions using output in a BLAST run of our transcripts/proteins against UniProt/Swiss-Prot using the following scripts

maker\_functional\_gff ../uniprot\_sprot.fasta output.renamed.blastp anstepV2\_maker\_rnd2.all.renamed.gff > anstepV2\_maker\_rnd2.all.renamed.putative\_function.gff

maker\_functional\_fasta ../uniprot\_sprot.fasta output.renamed.blastp anstepV2\_maker\_rnd2.all.maker.non\_overlapping\_ab\_initio.transcripts.renamed.fasta > anstepV2\_maker\_rnd2.all.maker.non\_overlapping\_ab\_initio.transcripts.renamed.putative\_function.fasta

maker\_functional\_fasta ../uniprot\_sprot.fasta output.renamed.blastp anstepV2\_maker\_rnd2.all.maker.non\_overlapping\_ab\_initio.proteins.renamed.fasta > anstepV2\_maker\_rnd2.all.maker.non\_overlapping\_ab\_initio.proteins.renamed.putative\_function.fasta

Step-4: Last, we added protein domain information to the final annotations using InterProScan data

ipr\_update\_gff anstepV2\_maker\_rnd2.all.renamed.putative\_function.gff output.renamed.iprscan > anstepV2\_maker\_rnd2.all.renamed.putative\_function.domain\_added.gff #adds searchable tags to the gene and mRNA features in the GFF3 file

iprscan2gff3 output.renamed.iprscan anstepV2\_maker\_rnd2.all.renamed.gff > visible\_iprscan\_domains.gff #adds physical viewable features for domains, which can be displayed in our JBrowse based ANSTEP-Web Server

#Extract different prediction results from master GFF file

awk '{if ($2 == "maker") print $0}' anstepV2\_maker\_rnd2.all.renamed.putative\_function.domain\_added.gff > anstepV2\_maker\_rnd2.all.renamed.putative\_function.domain\_added.MAKER.gff # shows only MAKER2 synthesized GOLD standard gene annotation

awk '{if ($2 == "est\_gff:est2genome") print $0}' anstepV2\_maker\_rnd2.all.renamed.putative\_function.domain\_added.gff > anstepV2\_maker\_rnd2.all.renamed.putative\_function.domain\_added.EST2GENOME.gff # shows transcripts evidence based annotation

awk '{if ($2 == "protein\_gff:protein2genome") print $0}' anstepV2\_maker\_rnd2.all.renamed.putative\_function.domain\_added.gff > anstepV2\_maker\_rnd2.all.renamed.putative\_function.domain\_added.PROTEIN2GENOME.gff # shows peptides evidence based annotation

awk '{if ($2 == "augustus\_masked") print $0}' anstepV2\_maker\_rnd2.all.renamed.putative\_function.domain\_added.gff > anstepV2\_maker\_rnd2.all.renamed.putative\_function.domain\_added.AUGUSTUS.gff # shows AUGUSTUS reports

awk '{if ($2 == "snap\_masked") print $0}' anstepV2\_maker\_rnd2.all.renamed.putative\_function.domain\_added.gff > anstepV2\_maker\_rnd2.all.renamed.putative\_function.domain\_added.SNAP.gff # shows SNAP reports

awk '{if ($2 ~ "repeat") print $0}' anstepV2\_maker\_rnd2.all.renamed.putative\_function.domain\_added.gff > anstepV2\_maker\_rnd2.all.renamed.putative\_function.domain\_added.REPEATS.gff # shows REPEATS reports

anstepV2\_maker\_rnd2.all.maker.non\_overlapping\_ab\_initio.transcripts.renamed.putative\_function.fasta (Non-overlapping ab initio predicted transcript from SNAP(n=3221)/AUGUSTUS(n=10971)

anstepV2\_maker\_rnd2.all.maker.non\_overlapping\_ab\_initio.proteins.renamed.putative\_function.fasta (Non-overlapping ab initio predicted proteins from SNAP(n=3221)/AUGUSTUS(n=10971))

anstepV2\_maker\_rnd2.all.maker.transcripts.renamed.putative\_function.Angam.fasta (MAKER transcript seq’s functional annotation using An. gambiae data)

anstepV2\_maker\_rnd2.all.maker.proteins.renamed.putative\_function.Angam.fasta (MAKER protein seq’s functional annotation using An. gambiae data) anstepV2\_maker\_rnd2.all.maker.transcripts.renamed.putative\_function.Anfun.fasta (MAKER transcript seq’s functional annotation using An. funestus data)

anstepV2\_maker\_rnd2.all.maker.proteins.renamed.putative\_function.Anfun.fasta (MAKER protein seq’s functional annotation using An. funestus data)

anstepV2\_maker\_rnd2.all.maker.transcripts.renamed.putative\_function.Dmel.fasta (MAKER transcript seq’s functional annotation using D. melanogaster)

anstepV2\_maker\_rnd2.all.maker.proteins.renamed.putative\_function.Dmel.fasta (MAKER protein seq’s functional annotation using D. melanogaster data)

anstepV2\_maker\_rnd2.all.renamed.putative\_function.domain\_added.gff (Master MAKER output GFF file with FASTA seq)

anstepV2\_maker\_rnd2.all.renamed.putative\_function.domain\_added\_ABINITIO\_non-overlap.gff (Non-overlapping Ab initio gene prediction)