MSc Thesis Committee Meeting # 1

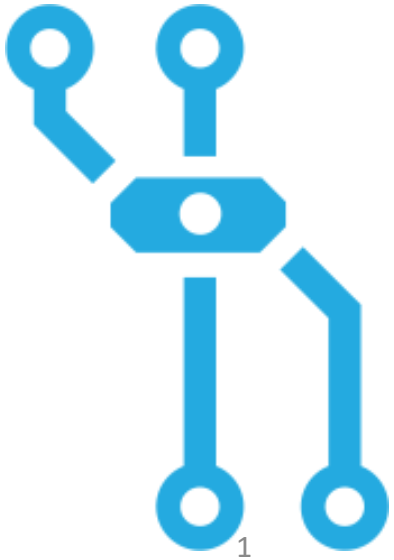
Monday, June 15, 2020

Diana Lin

MSc Student, Bioinformatics Graduate Program, UBC

Dr. Inanc Birol Lab, Genome Sciences Centre (GSC), BC Cancer

Supervisor: Dr. Inanc Birol



Committee Members

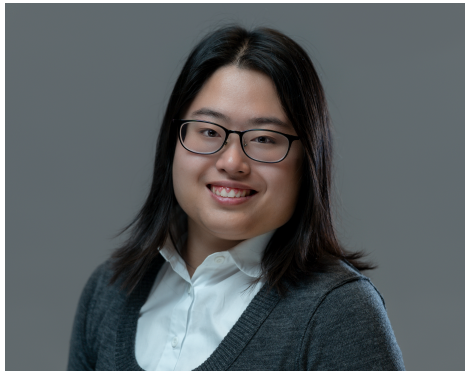
THE UNIVERSITY OF BRITISH COLUMBIA

Dr. Inanc Birol, PhD
Medical Genetics, UBC
Birol Lab, Genome Sciences Centre, BC Cancer

Dr. Phil Hieter, PhD
Medical Genetics, UBC
Hieter Lab, Michael Smith Laboratories, UBC

Dr. Michael Murphy, PhD
Microbiology and Immunology, UBC
Murphy Lab, Life Sciences Institute, UBC

Agenda



1. Introduction (3 min.)
2. Background (6 min.)
3. Progress Update (13 min.)
4. Open Discussion and Feedback (30 min.)

Education (2014 - 2018)

- Bachelor of Science, Major Physiology, McGill University
  - Elective courses to supplement Life Science degree:
    - **COMP 202**: Foundations of Programming (Java)
    - **COMP 206**: Introduction to Software Systems (Unix CLI, C)
    - **COMP 250**: Introduction to Computer Science (pseudocode)
    - **COMP 322**: Introduction to C++ (C++)
    - **COMP 364**: Computer Tools for Life Science (Python)

Internship (2018 - 2019)

- Student Intern, Birol Lab, Genome Sciences Centre
  - Data mining for antimicrobial peptide sequences in annotated nuclear spruce genomes using homology search
  - Differential expression analysis of AMPs in spruce transcriptomes
  - Assembly and annotation of organellar genomes

Species	Genome	Assembly	Annotation	NCBI Accession
White Spruce	chloroplast		✓	MK174379.1
Engelmann Spruce	chloroplast	✓	✓	MK241981.1
White Pine Weevil	mitochondrion	✓	✓	preparing submission

Publications

AMERICAN SOCIETY FOR MICROBIOLOGY | Microbiology Resource Announcements

GENOME SEQUENCES

Complete Chloroplast Genome Sequence of a White Spruce (*Picea glauca*, Genotype WS77111) from Eastern Canada

Diana Lin,[a] Lauren Coombe,[a] Shaun D. Jackman,[a] Kristina K. Gagalova,[a] René L. Warren,[a] S. Austin Hammond,[a]* Heather Kirk,[a] Pawan Pandoh,[a] Yongjun Zhao,[a] Richard A. Moore,[a] Andrew J. Mungall,[a] Carol Ritland,[b,f] Barry Jaquish,[c] Nathalie Isabel,[d] Jean Bousquet,[e] Steven J. M. Jones,[a] Joerg Bohlmann,[b,f] Inanc Birol[a]

Complete Chloroplast Genome Sequence of an Engelmann Spruce (*Picea engelmannii*, Genotype Se404-851) from Western Canada

Diana Lin,[a] Lauren Coombe,[a] Shaun D. Jackman,[a] Kristina K. Gagalova,[a] René L. Warren,[a] S. Austin Hammond,[a]* Helen McDonald,[a] Heather Kirk,[a] Pawan Pandoh,[a] Yongjun Zhao,[a] Richard A. Moore,[a] Andrew J. Mungall,[a] Carol Ritland,[b,e] Trevor Doerksen,[c] Barry Jaquish,[c] Jean Bousquet,[d] Steven J. M. Jones,[a] Joerg Bohlmann,[b,e] Inanc Birol[a]

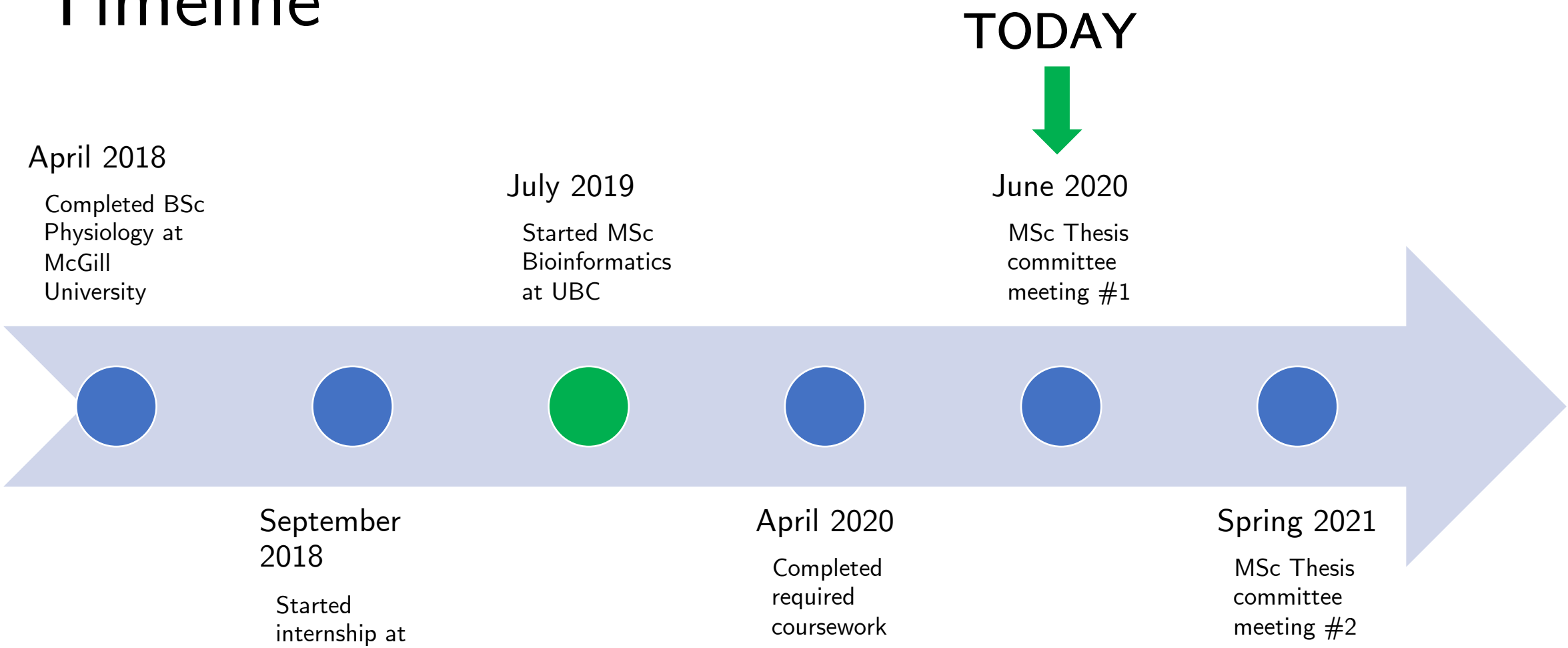
Courses

THE UNIVERSITY OF BRITISH COLUMBIA

- Term 1
  - ✓ **BIOF 501A**: Special Topics in Bioinformatics (A)
  - ✓ **STAT 545A**: Exploratory Data Analysis (A+)
  - ✓ **MICB 405**: Introduction to Bioinformatics (A+)

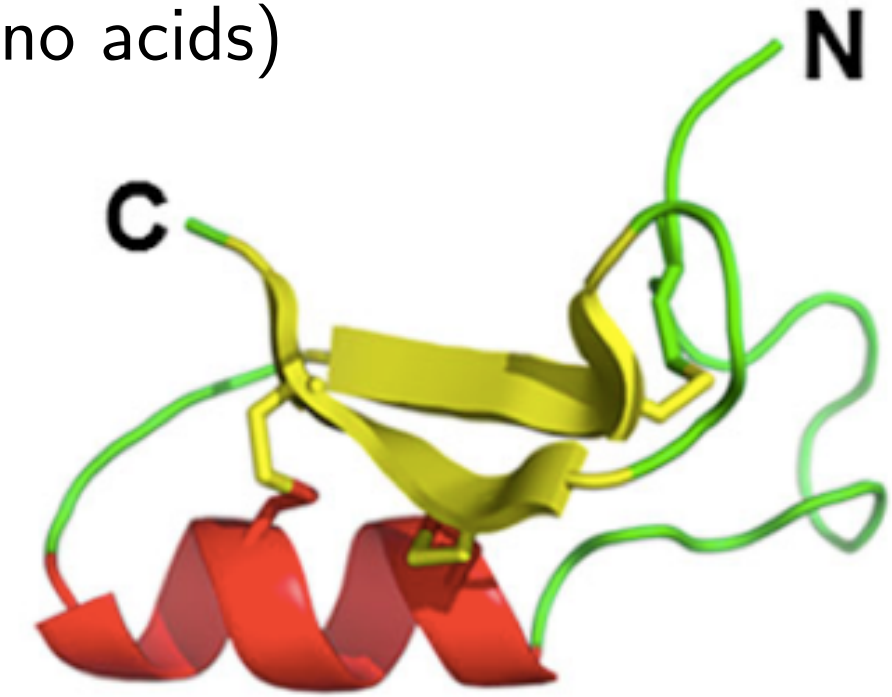
- Term 2
  - ✓ **BIOF 520**: Problem-Based Learning in Bioinformatics (A)
  - ✓ **STAT 540**: Statistical Methods for High Dimensional Biology (A+)
  - ✓ **STAT 547M**: Topics in Statistics (A+)
  - ✓ **MEDG 505**: Genome Analysis (A)

Timeline


TODAY
April 2018
Completed BSc Physiology at McGill University
July 2019
Started MSc Bioinformatics at UBC
June 2020
MSc Thesis committee meeting #1
September 2018
Started internship at Birol Lab, Genome Sciences Centre, BC Cancer
April 2020
Completed required coursework
Spring 2021
MSc Thesis committee meeting #2

Antimicrobial Peptides (AMPs)

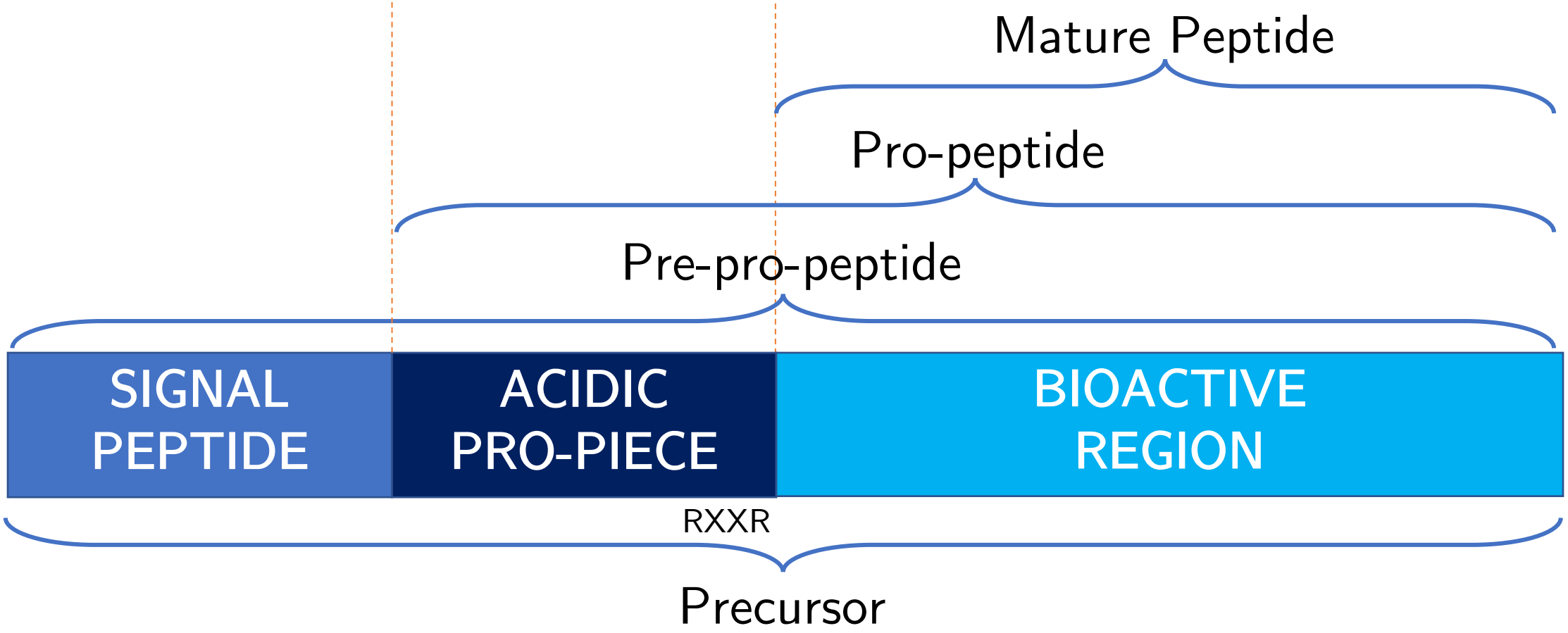
- Short peptide sequences (5 to 50 amino acids)
- Often positively charged
- Amphipathic
- Produced by all life forms
- Part of the innate immune system
- Mechanisms of action:
  1. Direct interaction
  2. Modulation of host immunity



Insect Defensin A

Source: Darcy Sutherland, PeptAID AIM 1 Science Meeting

Antimicrobial Peptides (AMPs)



Motivation

- Rise of antibiotic resistance creates a problem that requires a novel method to fight pathogens

PROBLEM

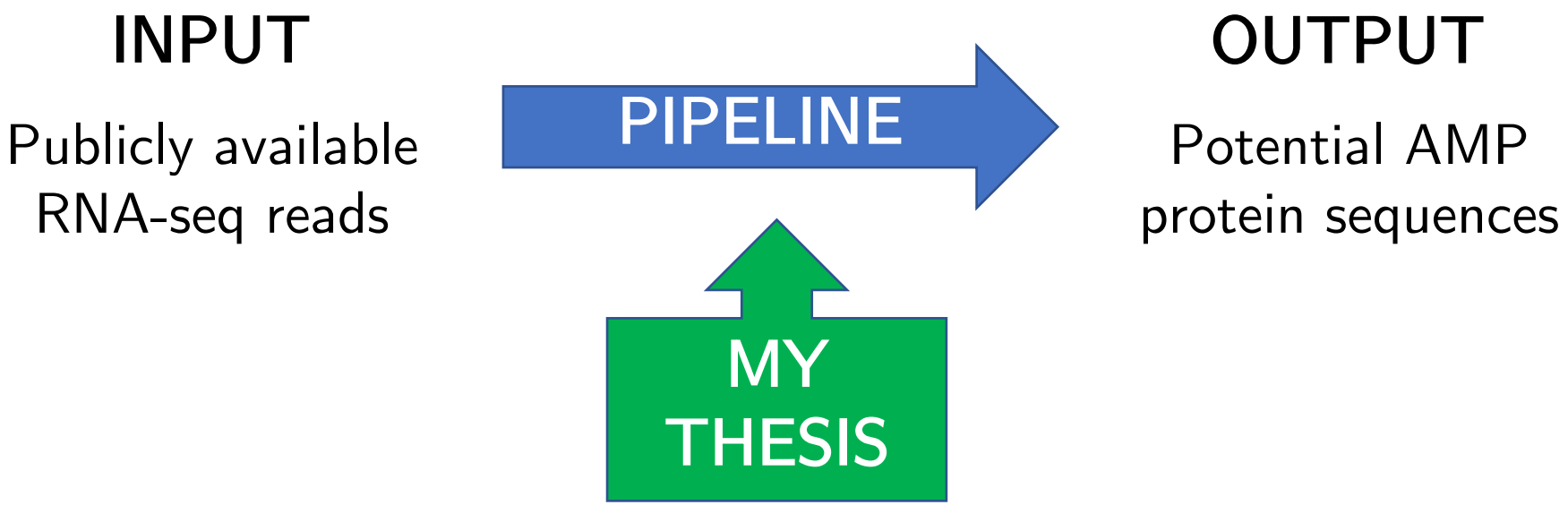
Antibiotic Resistance

SOLUTION

Antimicrobial Peptides?

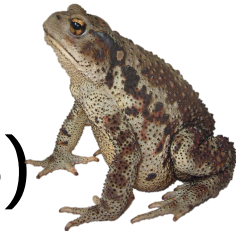
Objective

- To develop **and** execute an AMP discovery pipeline to mine for AMP precursor sequences in publicly available genomic resources



Datasets

Frogs & Toads (38)



Frogs (33)

A. femoralis
A. mantzorum
A. petersi
C. alboguttata
D. auratus
D. leucomelas
D. tinctorius
H. pugnax
L. verreauxii
O. margaretae
O. sylvatica
O. tormota
P. adspersus
P. amboli
P. megacephalus
P. microps
P. nigromaculatus
P. toftae
Q. boulengeri
R. catesbeiana
R. dennysi
R. imitator
R. omeimontis
R. pipiens
R. sirensis
R. sylvatica
R. temporaria
S. ruber
X. allofraseri
X. borealis
X. laevis
X. largeni
X. tropicalis

Toads (5)

A. minuta
B. gargarizans
L. Boringii
M. sangzhiensis
O rhodostigmatus

Sequence Read Archive

Ants, Bees, & Wasps (37)

Ants (8)

A. echinatior
C. castaneus
C. obscurior
M. gulosa
O. monticola
P. barbatus
T. bicarinatum
T. rugulatus

Bees (5)

A. cerana
A. mellifera
B. ardens
B. consobrinus
B. ussurensis

Wasps (24)

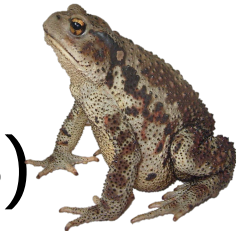
A. compressa
A. flavomarginatum
B. nigricans
C. vestalis
D. collaris
D. longicaudata
M. demolitor
N. giraulti
N. vitripennis
N. vitripennis x N. giraulti
O. decorates
P. rothneyi
P. snelleni
P. turionellae
P. varia
P. vindemmiae
S. deforme
S. kj8906
T. sarcophagae
U. rufipes
V. analis
V. crabro
V. dybowskii
V. similiia

Why Amphibians and Insects?

Source: Darcy Sutherland, PeptAID AIM 1 Science Meeting, dkfindout.com

Datasets

Frogs & Toads (38)



Frogs (33)

A. femoralis
A. mantzorum
A. petersi
C. alboguttata
D. auratus
D. leucomelas
D. tinctorius
H. pugnax
L. verreauxii
O. margaretae
O. sylvatica
O. tormota
P. adspersus
P. amboli
P. megacephalus
P. microps
P. nigromaculatus
P. toftae
Q. boulengeri
R. catesbeiana
R. dennysi
R. imitator
R. omeimontis
R. pipiens
R. sirensis
R. sylvatica
R. temporaria
S. ruber
X. allofraseri
X. borealis
X. laevis
X. largeni
X. tropicalis

Toads (5)

A. minuta
B. gargarizans
L. boringii
M. sangzhiensis
O rhodostigmatus

Sequence Read Archive

Ants, Bees, & Wasps (37)

Ants (8)

A. echinatior
C. castaneus
C. obscurior
M. gulosa
O. monticola
P. barbatus
T. bicarinatum
T. rugulatus

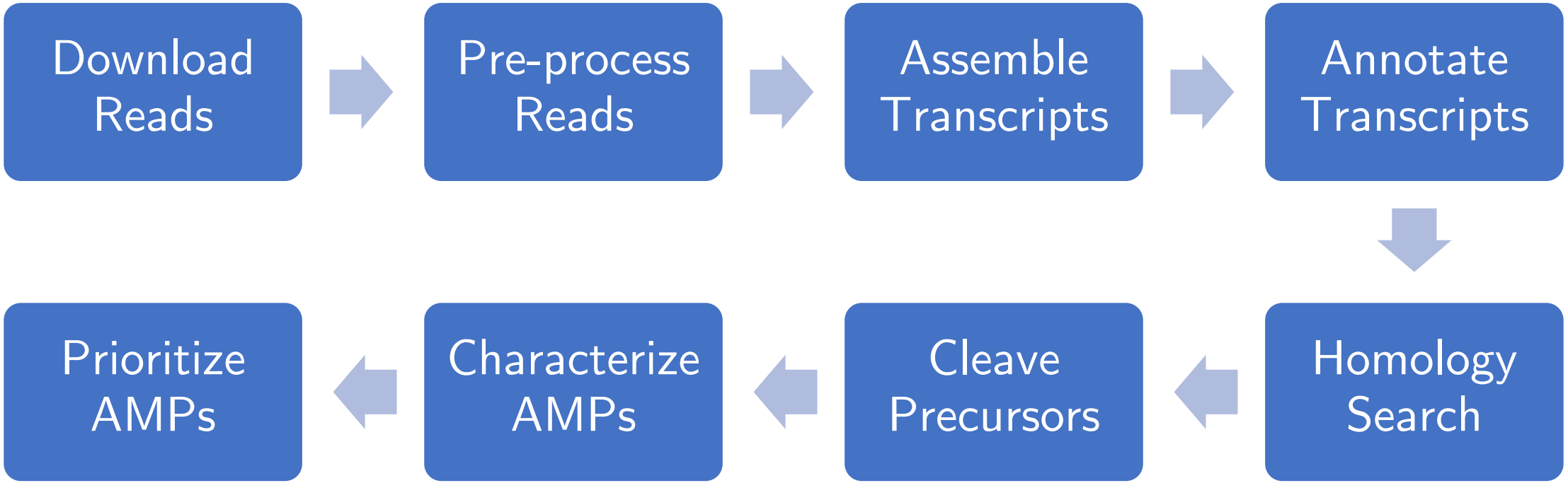
Bees (5)

A. cerana
A. mellifera
B. ardens
B. consobrinus
B. ussurensis

Wasps (24)

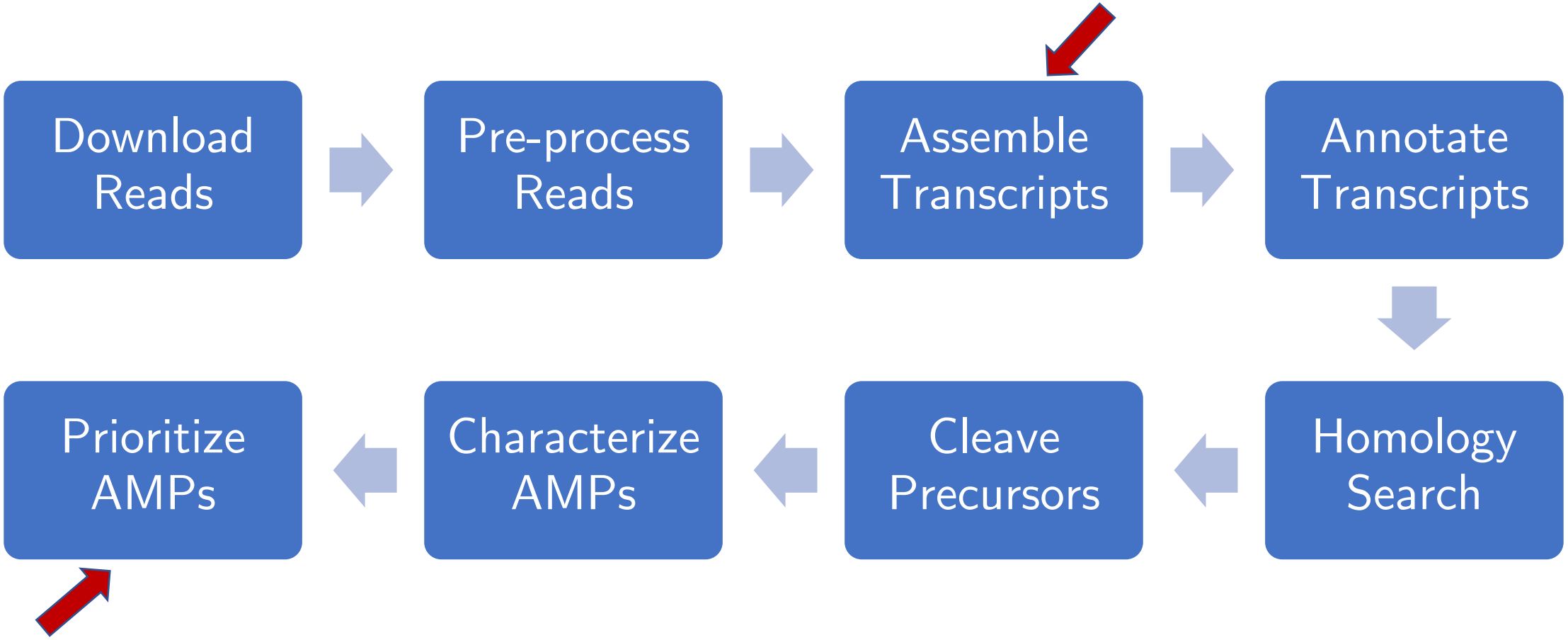
A. compressa
A. flavomarginatum
B. nigricans
C. vestalis
D. collaris
D. longicaudata
M. demolitor
N. giraulti
N. vitripennis
N. vitripennis x N. giraulti
O. decorates
P. rothneyi
P. snelleni
P. turionellae
P. varia
P. vindemmiae
S. deforme
S. kj8906
T. sarcophagae
U. rufipes
V. analis
V. crabro
V. dybowskii
V. similiia

Methods Overview



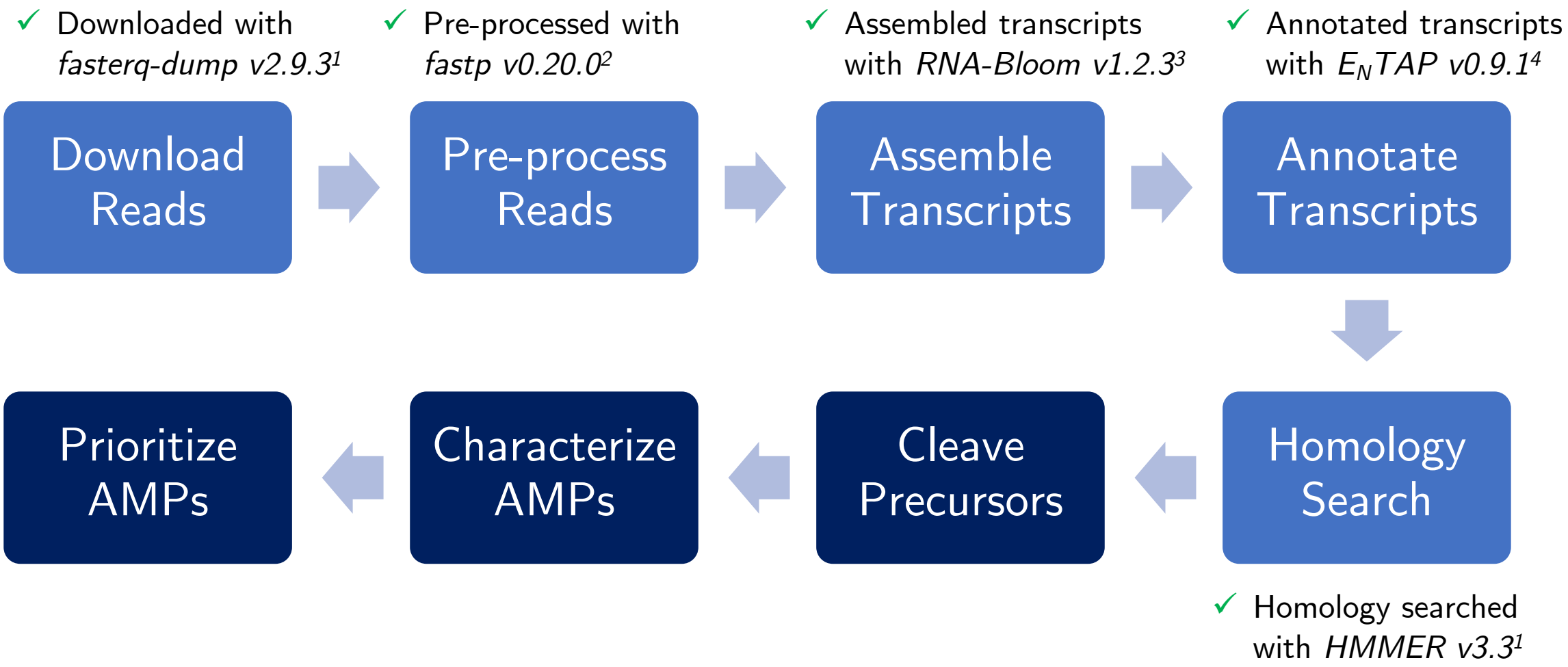
Methods Overview

RNA-Bloom
https://github.com/bcgsc/RNA-Bloom



AMPlify
https://github.com/bcgsc/amplify

Progress


✓ Downloaded with *fasterq-dump v2.9.3*[1]
✓ Pre-processed with *fastp v0.20.0*[2]
✓ Assembled transcripts with *RNA-Bloom v1.2.3*[3]
✓ Annotated transcripts with *E_NTAP v0.9.1*[4]
Download Reads
Pre-process Reads
Assemble Transcripts
Annotate Transcripts
Homology Search
Cleave Precursors
Characterize AMPs
Prioritize AMPs
✓ Homology searched with *HMMER v3.3*[1]

Pre-processing: fastp

Chen, S., Zhou, Y., Chen, Y. & Gu, J. **fastp: an ultra-fast all-in-one FASTQ preprocessor.** Bioinformatics 34, i884–i890 (2018).

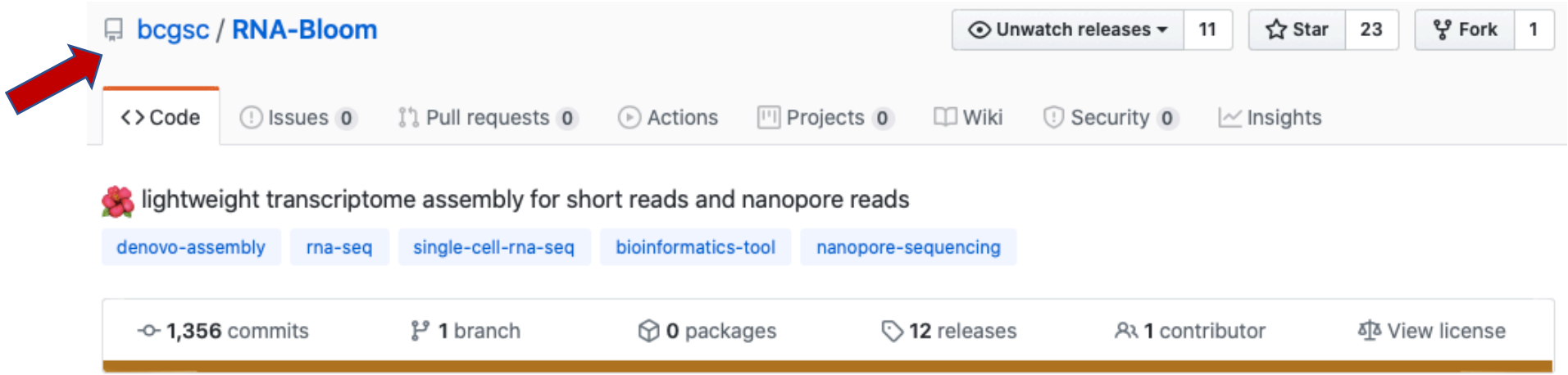
- Quality control - FASTQC
- Adapter trimming - Cutadapt
- Quality filtering - Trimmomatic

fastp **replaces** the need to use *three separate* tools

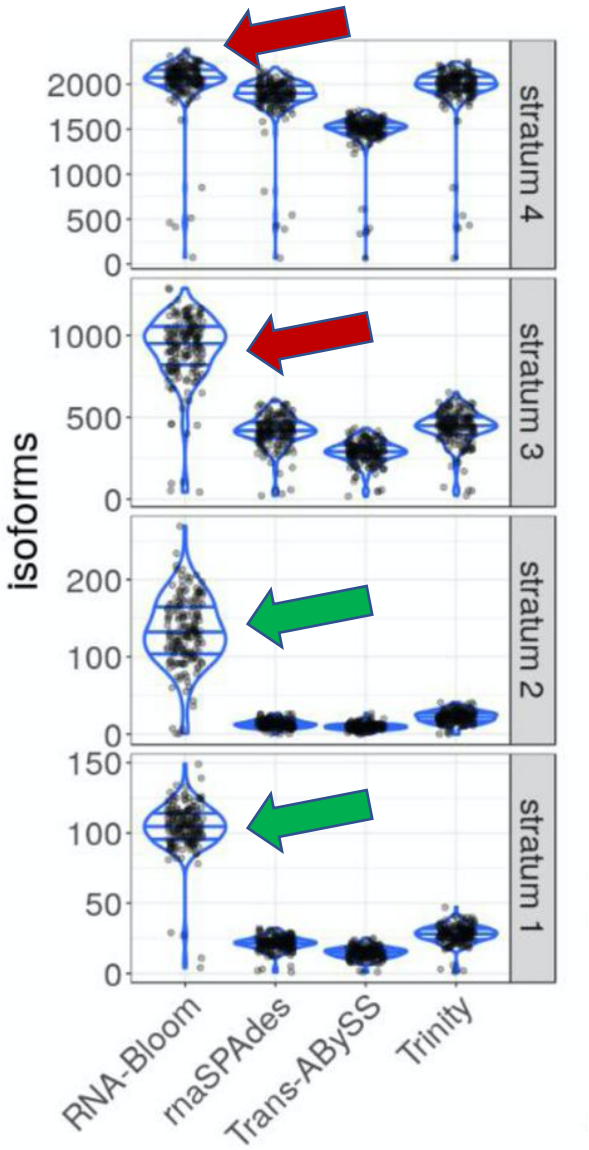
Assembly: RNA-Bloom

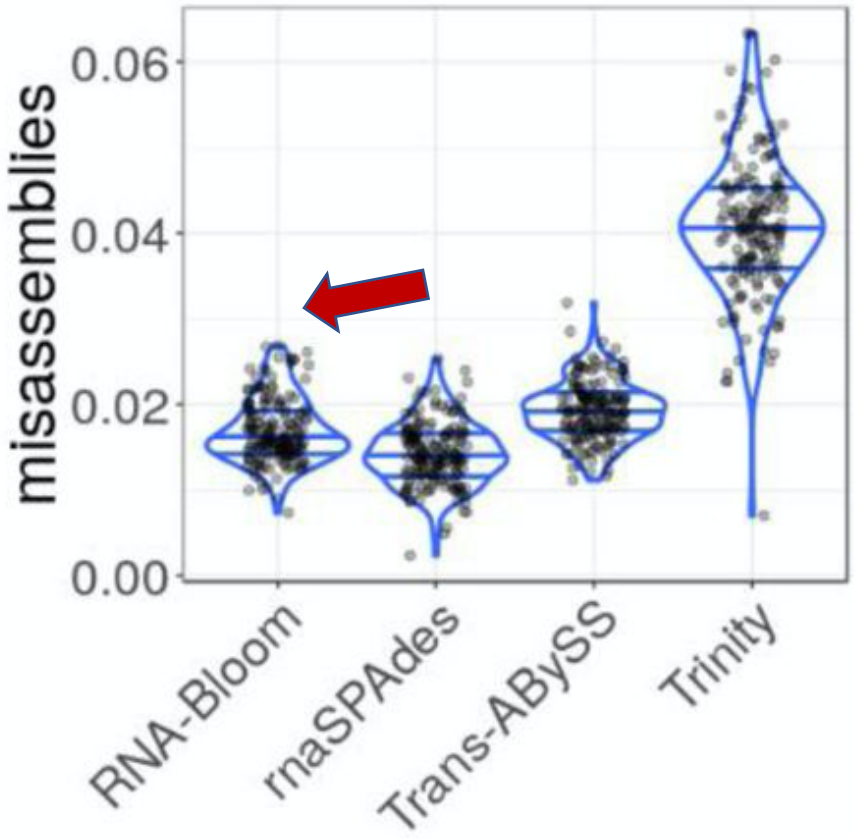
Nip, K. M. et al. **RNA-Bloom provides lightweight reference-free transcriptome assembly for single cells.** Genome Research (under revision).

- *De novo* transcriptome assembly with single and paired-end reads
- Reference-guided assembly if reference or draft transcriptome available



Assembly: RNA-Bloom





Source: Nip, K. M. et al. RNA-Bloom provides lightweight reference-free transcriptome assembly for single cells. Genome Research (under revision).

Annotation: E_NTAP

Hart, A. J. et al. **E_NTAP: Bringing faster and smarter functional annotation to non-model eukaryotic transcriptomes.** Mol. Ecol. Resour. 20, 591–604 (2020).

- **E_NTAP**: Eukaryotic Non-Model Transcriptome Annotation Pipeline
- Functional annotation with ‘frame selection’
- Alignment to orthology, domain, and protein databases

Annotation: E_NTAP

Metric	Blast2GO Pro	Blast2GO Basic	Trinotate	EnTAP	Annocript	Dammit
Open source/free software		‡	‡	‡	‡	‡
Command line integration	‡		‡	‡	‡	‡
Filtering assembly via short-read alignment (expression)	†			‡		
Frame selection	†		‡	‡	‡	‡
Custom database selection and indexing	‡		†	‡		†
Fast and sensitive NCBI BLAST alternative				‡	†	
Selection of optimal hit From several databases				‡	†	
Selection of optimal hit based on informativeness	†			‡		
Contaminant identification and filtering	‡			‡		
Orthologous gene family assignment	‡		†	‡		
Protein domain (CDD/InterProScan)	‡	‡	‡	‡	‡	‡
Gene ontology term and pathway assignment Sourced from protein alignments	‡	‡	‡		‡	
Gene ontology term and pathway assignment Sourced from orthologous genes	‡			‡		‡
Provides graphical user interface for annotation process	‡	‡				

Source: Hart, A. J. et al. ENTAP: Bringing faster and smarter functional annotation to non-model eukaryotic transcriptomes. Mol. Ecol. Resour. 20, 591–604 (2020).

Homology Search: HMMER

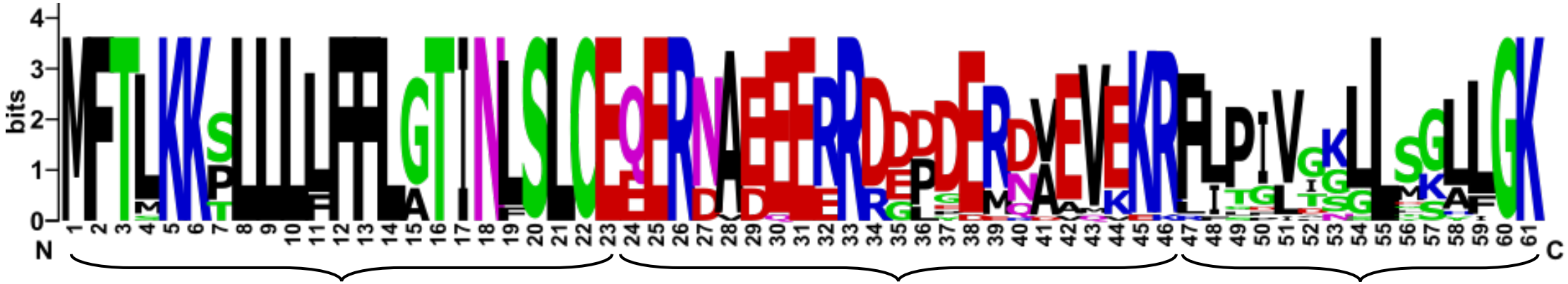
Johnson, L. S., Eddy, S. R. & Portugaly, E. **Hidden Markov model speed heuristic and iterative HMM search procedure.** BMC Bioinformatics 11, 431 (2010).

- Using antimicrobial precursor sequences from NCBI Protein database (including GenPept, UniProt, etc.)
- 2,792 amphibian precursors; 303 insect precursors

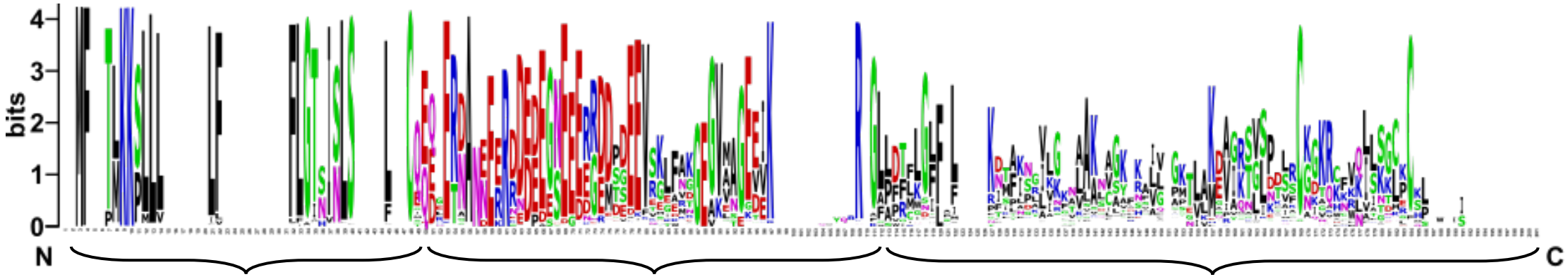
Source: S. Austin Hammond, generated using WebLogo

Homology Search: HMMER

Single AMP family:

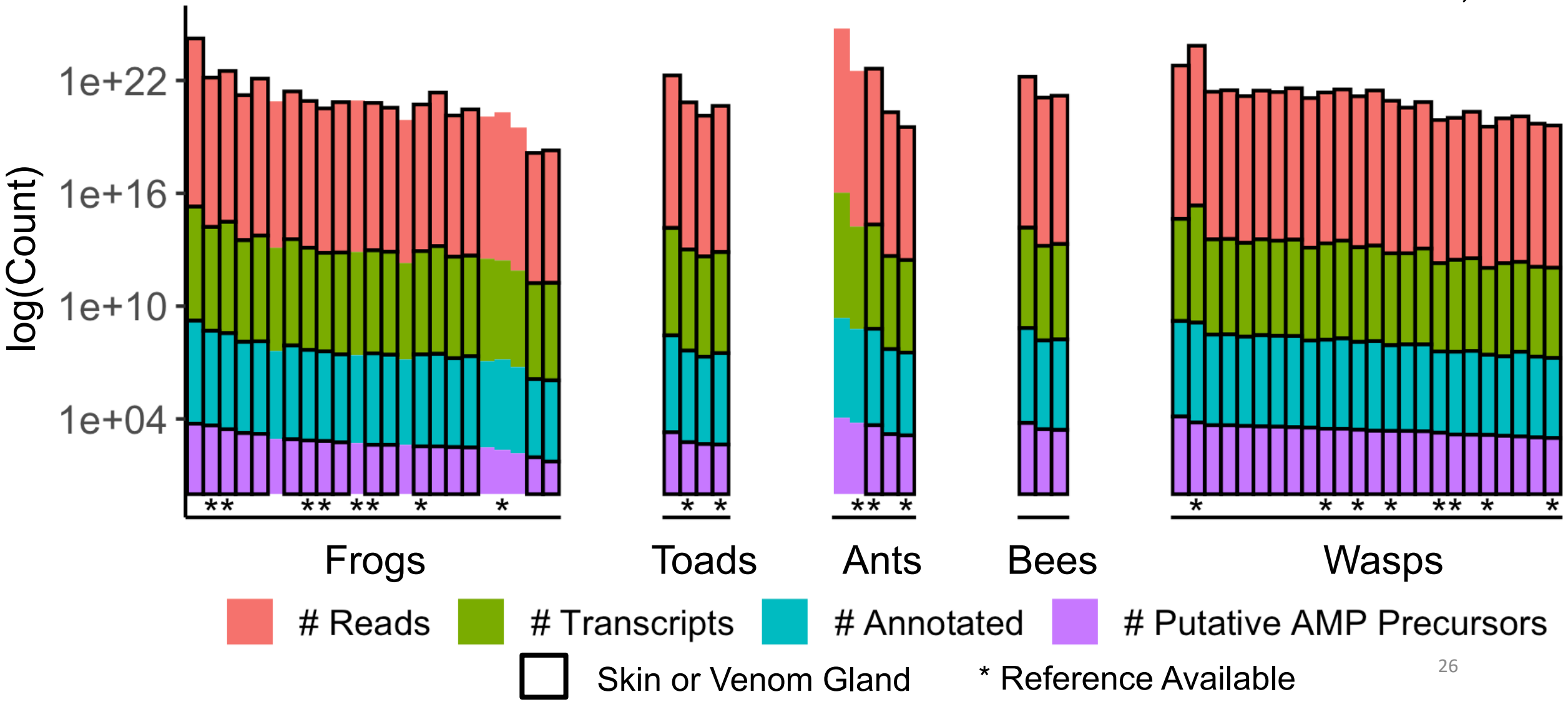


General AMP structure:

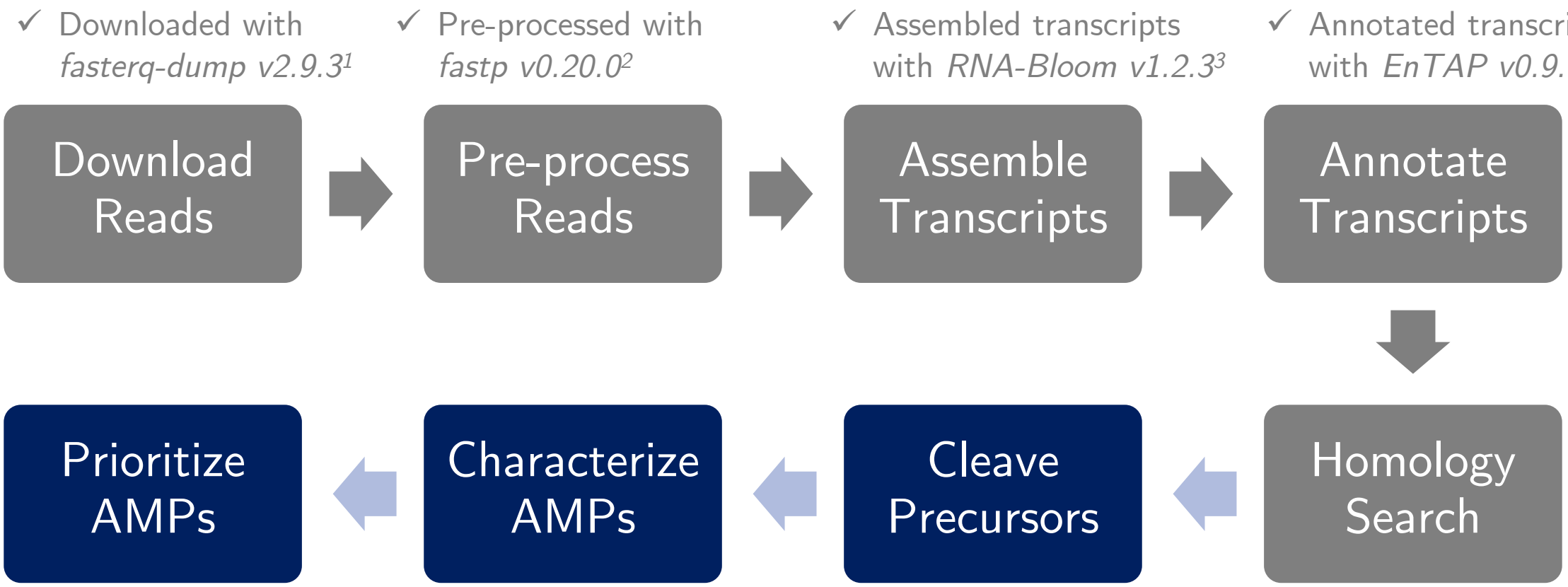


Preliminary Results

Total # Putative AMP Precursors: 140,054

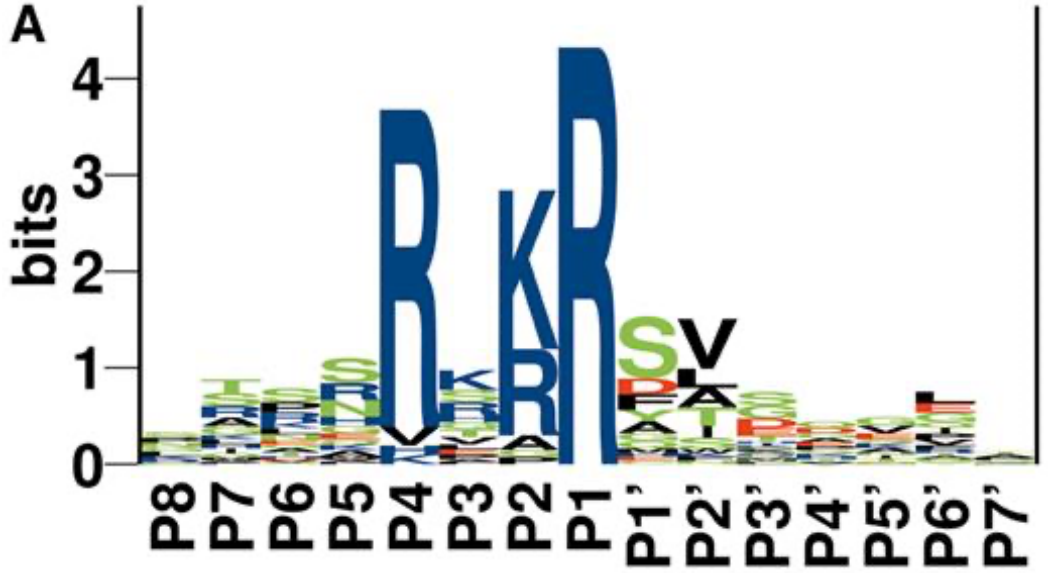


Future Work

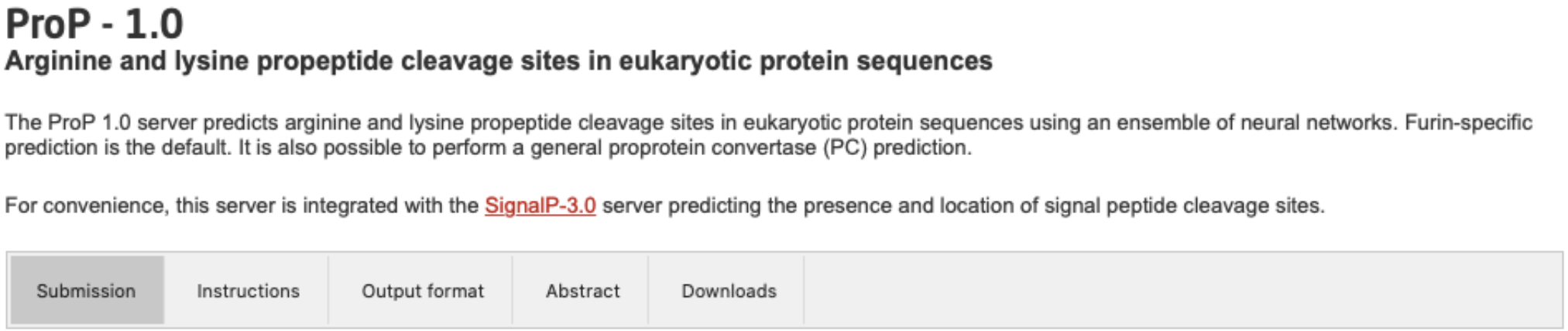

✓ Downloaded with *fasterq-dump v2.9.3*[1]
✓ Pre-processed with *fastp v0.20.0*[2]
✓ Assembled transcripts with *RNA-Bloom v1.2.3*[3]
✓ Annotated transcripts with *EnTAP v0.9.1*[4]
Download Reads
Pre-process Reads
Assemble Transcripts
Annotate Transcripts
Homology Search
Cleave Precursors
Characterize AMPs
Prioritize AMPs
✓ Homology searched with *HMMER v3.3*[5]
❑ Cleave precursors into mature peptides using *ProP v1.0*[6]
❑ Characterize secondary protein structures using *SABLE v4.0*[7]
❑ Prioritize AMPs by AMP probability using *AMPlify v1.0*[8]

Precursor Cleavage: ProP

Duckert, P., Brunak, S. & Blom, N. **Prediction of proprotein convertase cleavage sites.** Protein Eng. Des. Sel. 17, 107–112 (2004).



Precursor Cleavage: ProP

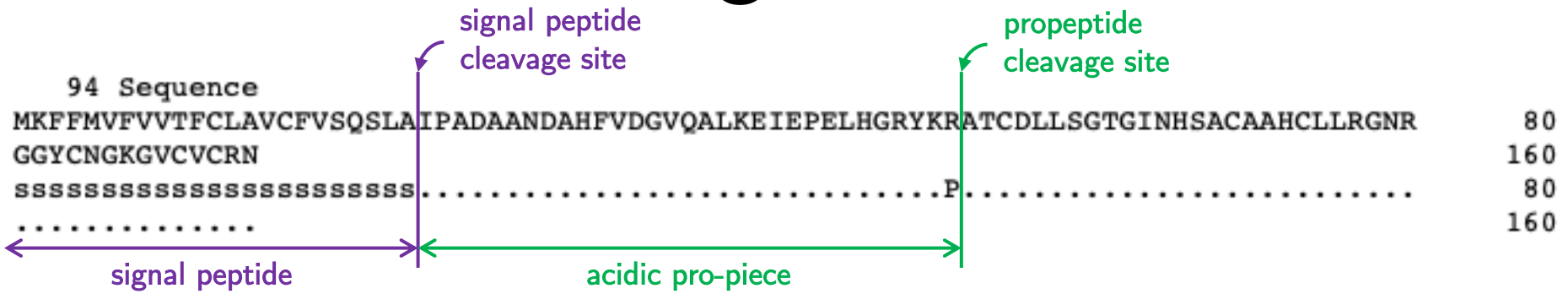


>sp|P10891.2|DEFI_PROTE RecName: Full=Phormicin; AltName: Full=Insect defensin A/B; Flags: Precursor

MKFFMVFVVTFCLAVCFVSQSLAIPADAANDAHFVDGVQALKEIE
PELHGRYKRATCDLLSGTGINHSACAAHCLLRGNRGGYCNGKGVC
VCRN

Source: Duckert, P., Brunak, S. & Blom, N. Prediction of proprotein convertase cleavage sites. Protein Eng. Des. Sel. 17, 107–112 (2004).

Precursor Cleavage: ProP



>sp|P10891.2|DEFI_PROTE RecName: Full=Phormicin; AltName: Full=Insect defensin A/B; Flags: Precursor

MKFFMVFVVTFCLAVCFVSQSLAIPADAANDAHFVDGVQALKEIEPELHG**RYKR**

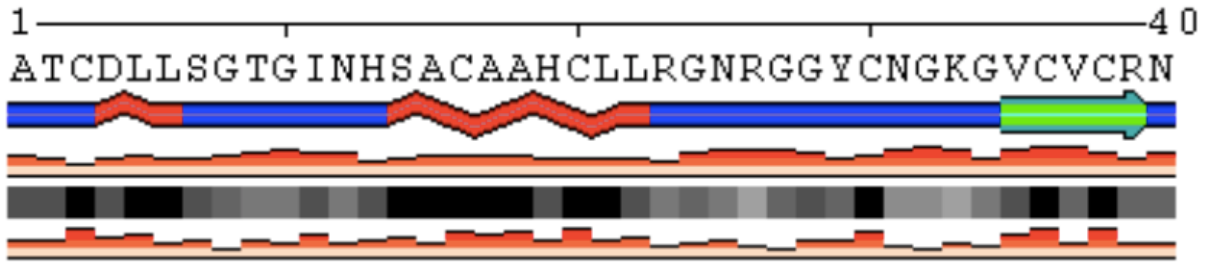
ATCDLLSGTGINHSACAAHCLLRGNRGGYCNGKGVCVCRN

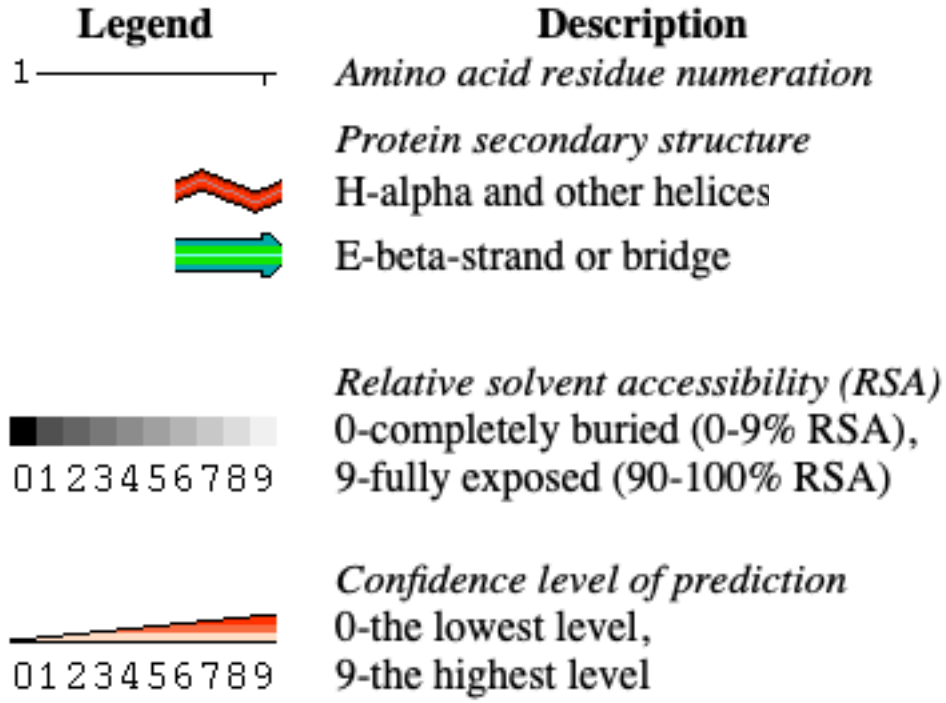
Source: Duckert, P., Brunak, S. & Blom, N. Prediction of proprotein convertase cleavage sites. Protein Eng. Des. Sel. 17, 107–112 (2004).

Characterization: SABLE

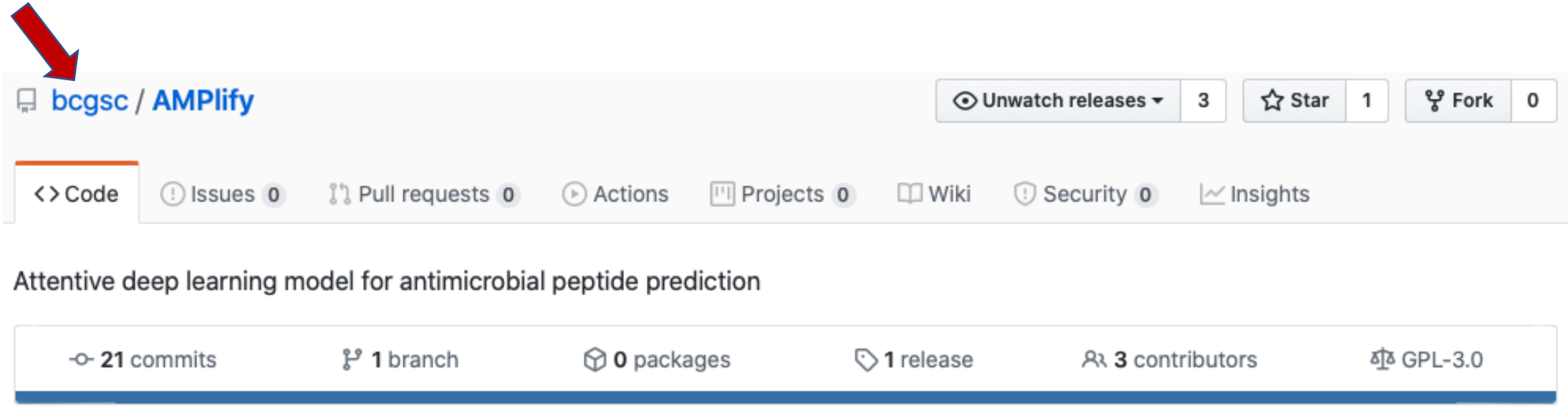
Adamczak, R., Porollo, A. & Meller, J. **Combining prediction of secondary structure and solvent accessibility in proteins.** Proteins 59, 467–475 (2005).

>sp|P10891.2|DEFI_PROTE RecName: Full=Phormicin; AltName: Full=Insect defensin A/B; Note: Full=Cleaved
ATCDLLSGTGINHSACAAHCLLRGNRGGYCNGKGVCVCRN





Prioritization: AMPlify



>sp|P10891.2|DEFI_PROTE RecName: Full=Phormicin; AltName: Full=Insect defensin A/B; Note: Full=Cleaved
ATCDLLSGTGINHSACAAHCLLRGNRGGYCNGKGVCVCRN

→ AMPlify →

AMP Probability Score: 1.0
Prediction: AMP non-AMP

Source: Li, C. et al. AMPlify: Attentive Deep Learning Model for Discovery of Novel Antimicrobial Peptides Effective against WHO Priority Pathogens. Nature Methods (submitted).

End Product

- ✓ AMP Discovery Pipeline software package that runs all the tools from beginning to end

- ✓ Resulting from the pipeline, candidate AMPs for downstream analysis and *in vitro* bioactivity testing against various microbes

References

1. Leinonen, R., Sugawara, H., Shumway, M. & International Nucleotide Sequence Database Collaboration. The sequence read archive. Nucleic Acids Res. **39**, D19–21 (2011).
2. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics **34**, i884–i890 (2018).
3. Nip, K. M. et al. RNA-Bloom provides lightweight reference-free transcriptome assembly for single cells. Genome Research (under revision).
4. Hart, A. J. et al. E_NTAP: Bringing faster and smarter functional annotation to non-model eukaryotic transcriptomes. Mol. Ecol. Resour. **20**, 591–604 (2020).
5. Johnson, L. S., Eddy, S. R. & Portugaly, E. Hidden Markov model speed heuristic and iterative HMM search procedure. BMC Bioinformatics **11**, 431 (2010).
6. Duckert, P., Brunak, S. & Blom, N. Prediction of proprotein convertase cleavage sites. Protein Eng. Des. Sel. **17**, 107–112 (2004).
7. Adamczak, R., Porollo, A. & Meller, J. Combining prediction of secondary structure and solvent accessibility in proteins. Proteins **59**, 467–475 (2005).
8. Li, C. et al. AMPlify: Attentive Deep Learning Model for Discovery of Novel Antimicrobial Peptides Effective against WHO Priority Pathogens. Nature Methods (submitted).

Acknowledgements

Birol Lab

- Ka Ming Nip
- Kristina Gagalova
- Darcy Sutherland
- Chenkai Li
- René Warren
- Inanc Birol

PeptAID Collaborators

- Caren Helbing

PeptAID Funding

University of Victoria