rAMPage: Rapid Anti-microbial Peptide Annotation and Gene Estimation

Diana Lin[1,2], Ka Ming Nip[1,2], Sambina Aninta[1], Chenkai Li[1,2], René L. Warren[1], Caren Helbing[3], Linda Hoang[4,5], Inanc Birol[1,6]

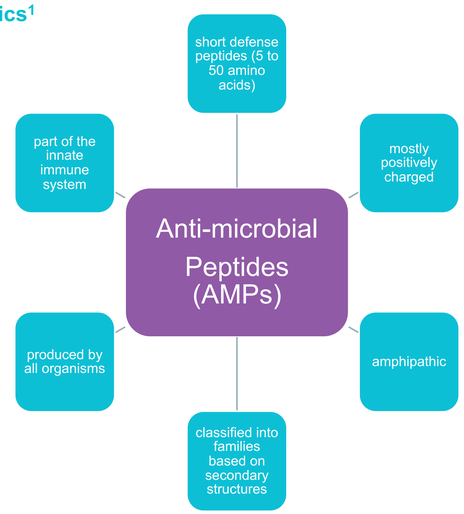
[1]Canada's Michael Smith Genome Sciences Centre, BC Cancer, Vancouver, BC, Canada
[2]Bioinformatics Graduate Program, University of British Columbia, Vancouver, BC, Canada
[3]Department of Biochemistry and Microbiology, University of Victoria, Victoria, BC, Canada
[4]Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, BC, Canada
[5]Public Health Laboratory, British Columbia Centre for Disease Control, Vancouver, BC, Canada
[6]Department of Medical Genetics, University of British Columbia, Vancouver, BC, Canada

Contact: dlin@bcgsc.ca
https://github.com/dy-lin

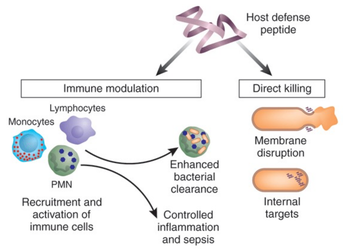
Introduction

Anti-microbial Peptides (AMPs)

Characteristics[1]

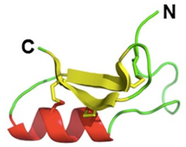


Mechanisms of Action[1]

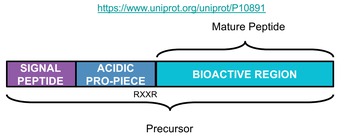


Source: Hancock, R. E. W. & Sahl, H.-G. Antimicrobial and host-defense peptides as new anti-infective therapeutic strategies. Nat. Biotechnol. 24, 1551–1557 (2006) doi: 10.1038/nbt1267

Peptide Structure[1]



Above: Insect Defensin A (phormicin) from the Northern blowfly
https://www.rcsb.org/structure/1ICA
https://www.uniprot.org/uniprot/P10891



- AMPs are activated by cleavage at the RXXR motif (acidic propiece inhibits basic bioactive region)
- Cleavage separates the signal peptide and acidic propiece from the bioactive region yielding the mature peptide

Motivation

PROBLEM antibiotic resistance → **SOLUTION** antimicrobial peptides

Problem

- The rise of antibiotic resistance[1]
- The antibiotic "discovery" void[2]: few new antibiotics, but old antibiotics less effective
- The need for novel methods to fight pathogen

Solution

- AMPs do not confer resistance as easily as antibiotics, due to co-evolution with the human microbiome[1]
- AMPs are a potential alternative to antibiotics[3]
- AMPs can be mined from organisms of rich AMP diversity, such as the North American bullfrog[4]

Objectives

- To develop and execute a scalable bioinformatics-based AMP discovery pipeline (i.e. *rAMPage*) to mine for AMP sequences in publicly available genomic resources
- To package a fully functional bioinformatics pipeline
- To obtain a list of potential AMP sequences for
  - Downstream analysis
  - *In vitro* bioactivity testing
  - Drug development

INPUT Publicly available RNA-seq reads[5] → *rAMPage* → **OUTPUT** Potential AMP protein sequences

Methods

Pre-process Reads
- Trim to remove adapter sequences using fastp[6]
- Filter out poor-quality bases and sequences
- fastp finds adapter sequences using sequence overlap

Assemble Transcripts
- Multi-sample pooled *de novo* or reference-guided assembly of reads into transcripts using RNA-Bloom[7], for both single and paired-end reads
- Lowly-expressed transcripts are filtered out using quantification from Salmon[8]

Translate Transcripts
- *In silico* translation of transcripts into amino acid space
- Six-frame translation and open reading frame (ORF) prediction using Transdecoder[9]

Homology Search
- Homology search with *jackhmmer* from the HMMER[10] package, using AMP databases APD3[11] and DADP[12]

Cleave Precursors
- Predict signal peptide and propeptide cleavage sites using prediction tool ProP[13]
- Separate signal peptide and acidic propieces from the bioactive region

Prioritize AMPs
- Obtain the probability that each sequence is an AMP using AMPlify[14]
- Rank potential AMPs and prioritize which sequences are synthesized and tested *in vitro* for bioactivity

Annotate AMPs
- Annotate AMPs using E_NTAP[15] to determine known gene ontologies and protein domains
- Determine AMP novelty by alignment to known AMPs using Exonerate[16]

Characterize AMPs
- Predict secondary structures and solvent accessibility using SABLE[17]
- Using SABLE, alpha helix and beta strand structures can be predicted

Results

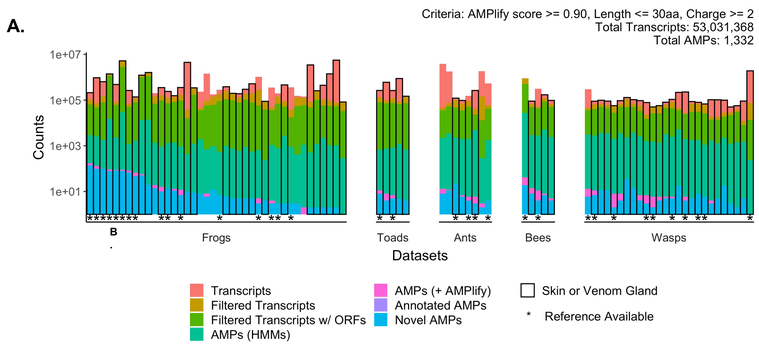


Fig A.[18] Count progression from transcripts to AMPs. Across the 84 datasets, *rAMPage* assembled > 53 million transcripts, and detected > 1000 putative AMPs (AMPlify score >= 0.50 is an AMP; stricter criteria used above).

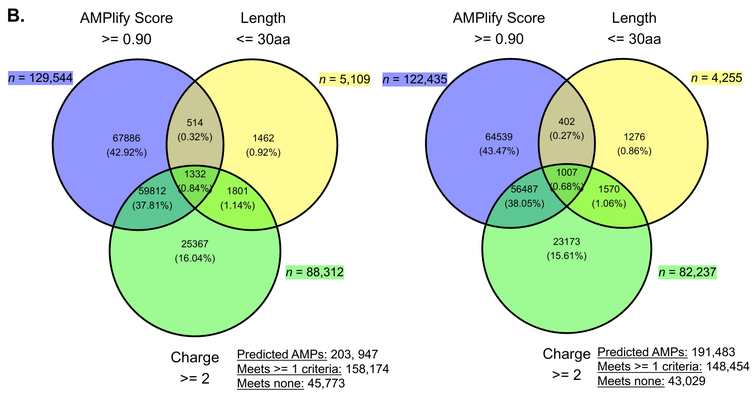


Fig B.[19] AMP counts after applying three filters. Three strict filters (AMPlify score >= 0.90, length <= 30aa, charge >= 2) are applied in *rAMPage*. [Left] 1,332 AMPs remain after filtering and duplicate sequence removal *within each* dataset. [Right] 1,007 AMPs remain after filtering and duplicate sequence removal *across all* datasets. If desired, more AMPs (of lower confidence) can be detected by adjusting the stringency of the filters.

C. *rAMPage* Benchmarking

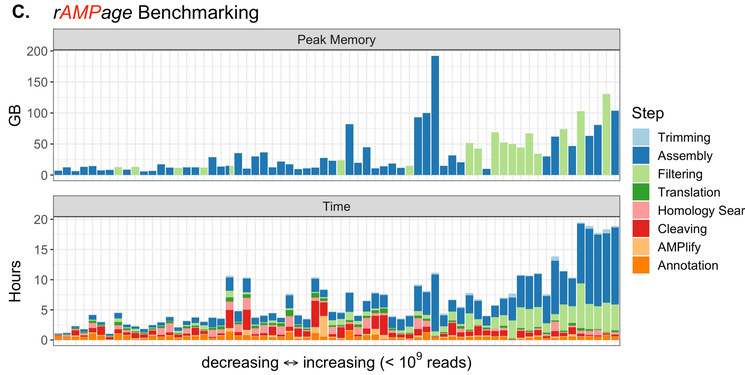


Fig C.[18] Runtime and memory usage of *rAMPage*. *rAMPage* is fast: with < 1 billion reads (74/84 datasets), results can be obtained within 24 hours, using < 200 GB of memory. Larger datasets with > 1 billion reads can be subsampled to reduce runtime and memory usage.

Conclusions

- Across the 84 assembled transcriptomes, 1,007 confident (AMPlify score >= 0.90), short (length <= 30aa), and positive (charge >= 2) unique mature putative AMPs were found: 795 from amphibians, 212 from insects
- Of these 1,007 AMPs, 254 sequences align to known AMPs with 100% sequence identity in the mature region; 753 sequences are 'novel' AMPs
- *rAMPage* is a fast, robust bioinformatics pipeline that, given raw reads, can detect known and novel putative AMPs

References

1. Hancock, R. E. W. & Sahl, H.-G. Antimicrobial and host-defense peptides as new anti-infective therapeutic strategies. *Nat. Biotechnol.* **24**, 1551–1557 (2006) doi: 10.1038/nbt1267
2. Koo, H. B. & Seo, J. Antimicrobial peptides under clinical investigation. *Peptide Science* vol. **111** (2019) doi: 10.1002/pep2.24122
3. Reinhardt, A. & Neundorf, I. Design and Application of Antimicrobial Peptide Conjugates. *Int. J. Mol. Sci.* **17**, (2016) doi: 10.3390/ijms17050701
4. Helbing, C. C. et al. Antimicrobial peptides from Rana [Lithobates] catesbeiana: Gene structure and bioinformatic identification of novel forms from tadpoles. *Sci. Rep.* **9**, 1529 (2019) doi: 10.1038/s41598-018-38442-1
5. Leinonen, R., Sugawara, H., Shumway, M. & International Nucleotide Sequence Database Collaboration. The sequence read archive. *Nucleic Acids Res.* **39**, D19–21 (2011) doi: 10.1093/nar/gkq1019
6. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018) doi: 10.1093/bioinformatics/bty560
7. Nip, K. M. et al. RNA-Bloom enables reference-free and reference-guided sequence assembly for single-cell transcriptomes. *Genome Res.* **30**, 1191–1200 (2020) doi: 10.1101/gr.260174.119
8. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017) doi: 10.1038/nmeth.4197
9. Haas, B. J. et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512 (2013) doi: 10.1038/nprot.2013.084
10. Johnson, L. S., Eddy, S. R. & Portugaly, E. Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics* **11**, 431 (2010) doi: 10.1186/1471-2105-11-431
11. Wang, G., Li, X. & Wang, Z. APD3: the antimicrobial peptide database as a tool for research and education. *Nucleic Acids Res.* **44**, D1087–93 (2016) doi: 10.1093/nar/gkv1278
12. Novković, M., Simunić, J., Bojović, V., Tossi, A. & Juretić, D. DADP: the database of anuran defense peptides. *Bioinformatics* **28**, 1406–1407 (2012) doi: 10.1093/bioinformatics/bts141
13. Duckert, P., Brunak, S. & Blom, N. Prediction of proprotein convertase cleavage sites. *Protein Eng. Des. Sel.* **17**, 107–112 (2004) doi: 10.1093/protein/gzh013
14. Li, C. et al. AMPlify: attentive deep learning model for discovery of novel antimicrobial peptides effective against WHO priority pathogens. *bioRxiv* **2020.06.16.155705** (2020) doi: 10.1101/2020.06.16.155705
15. Hart, A. J. et al. EnTAP: Bringing faster and smarter functional annotation to non-model eukaryotic transcriptomes. *Mol. Ecol. Resour.* **20**, 591–604 (2020) doi: 10.1111/1755-0998.13106
16. Slater, G. S. C. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005) doi: 10.1186/1471-2105-6-31
17. Adamczak, R., Porollo, A. & Meller, J. Combining prediction of secondary structure and solvent accessibility in proteins. *Proteins* **59**, 467–475 (2005) doi: 10.1002/prot.20441
18. Wickham, H. et al. Welcome to the tidyverse. *J. Open Source Softw.* vol. **4** 1686 (2019). doi: 10.21105/joss.01686
19. Yan, L. ggvenn: Draw Venn Diagram by 'ggplot2'. *Comprehensive R Archive Network* (2021). https://CRAN.R-project.org/package=ggvenn

Funding & Acknowledgements

Genome British Columbia · BC Cancer · PepAID · Genome Canada · Investment Agriculture Foundation of British Columbia · University of Victoria · UBC · bti