Novel DNA polymerases mined from metagenomic sequences fill the phenotypic information gap in sequence databases

Rachel A. Keown[1], Jacob T. Dums[2], David A. Mead[3], Joyanne MacDonald[3], Barbra D. Ferrell[4], Phillip Brumm[3], Shawn W. Polson[4], K. Eric Wommack[4]

[1]Department of Biological Sciences, UD; [2]North Carolina State University; [3]Varigen BioSciences, Middleton, WI; [4]Delaware Biotechnology Institute, UD

Introduction

- Viruses are the most abundant and diverse biological entities on the planet
- Sequencing of viral metagenomes (viromes) has exponentially increased genotypic data availability in recent years
- Phenotypic data collection is commonly lacking in published databases
- DNA polymerase A (PolA) is carried by ca. 25% of viruses and is used as a marker gene
- Amino acid identity at position 762 in PolA have potential links to phage lifestyle

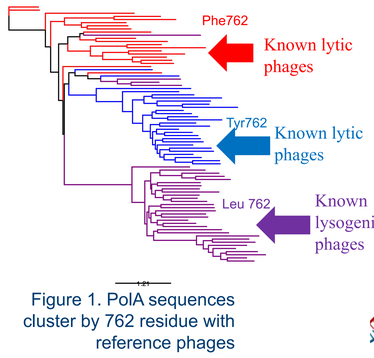

Figure 1. PolA sequences cluster by 762 residue with reference phages

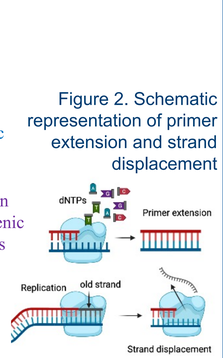

Figure 2. Schematic representation of primer extension and strand displacement

Purpose & Research Questions

Investigate the *in vitro* biochemistry potential of viral metagenome (virome) derived polymerase sequences to elucidate information about viral lifestyle and polymerase diversity

- Can metagenomically assembled sequences produce soluble proteins that are functional *in vitro?*
- Does a change in amino acid identity at position 762 impact speed and processivity of the polymerase?
- Are there any novel functions that can be discovered and potentially applied to problems in the biotechnology industry?
- How diverse is the biochemistry across global PolA diversity?

Methods & Data Selection

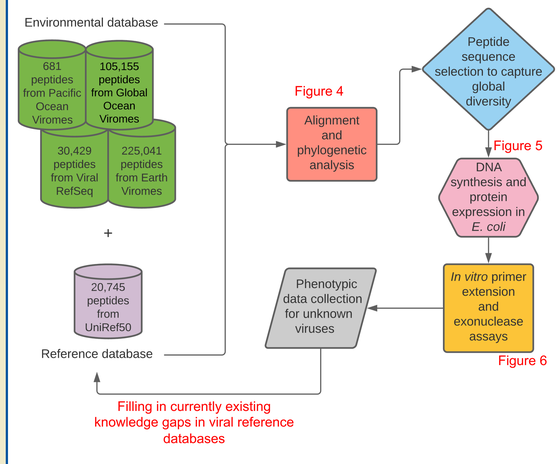

Figure 3. Methodological workflow for data acquisition

Sequences were gathered from published databases, aligned and assembled into a phylogenetic tree (fig. 2). Peptides of interest were selected for synthesis in *E. coli*, produced, and *in vitro* activities characterized. This phenotypic data can then be used to fill in missing data in reference databases.

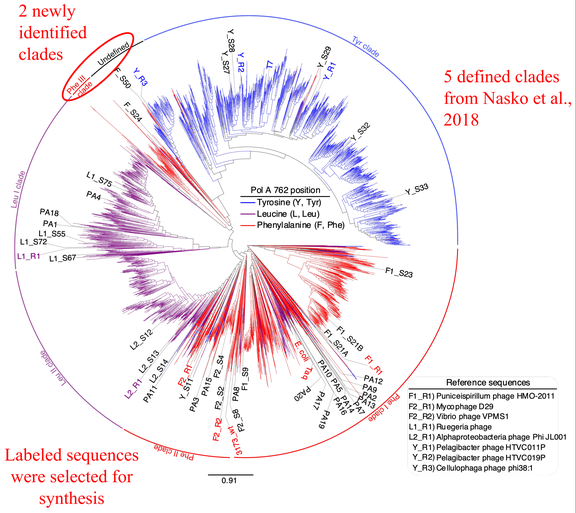

Figure 4. Synthesized family A DNA polymerase sequences span phylogenetic diversity

Approximate maximum likelihood tree of 3,604 PolA amino acid sequences trimmed to the region of interest (AA521-AA928, *E. coli* IAI39 numbering) the polymerase domain. Scale bar represents the number of amino acid substitutions per site.

Fifty viral polymerases sequences were selected for synthesis from all seven clades to represent global PolA diversity

Results

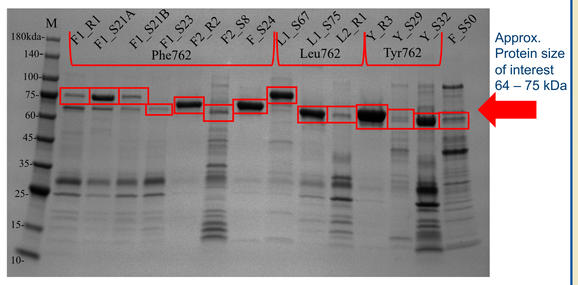

Figure 5. Protein production from synthesized reference and virome sequences

SDS-PAGE gel of His-tagged proteins from reference and virome clones.

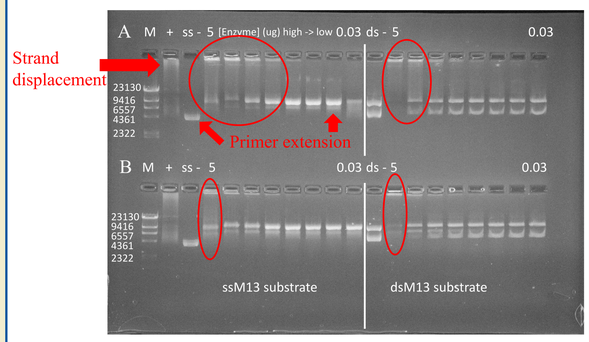

Figure 6. *In vitro* primer extension and strand displacement in two synthesized virome polymerases

A) F_S24 and B) F1_S21A at decreasing (left to right) concentrations (µg/reaction) in 1X phi29 buffer after 18.5 h incubation at 30 °C.

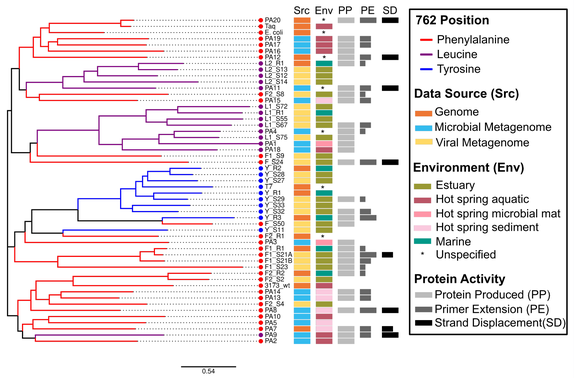

Figure 7. Synthesized polymerase A sequences produce proteins with diverse biochemistry

An approximate maximum likelihood tree of four cultivated references and 50 synthesized family A DNA polymerase amino acid sequences. Bar lengths reflect the strength of primer extension and strand displacement activities. Scale bar represents the number of amino acid substitutions per site.

Conclusions

- Metagenome and virome derived polymerase sequences produce soluble, active proteins
- Strand displacement is observed across a diverse phylogeny of Polymerase A enzymes
- Virome derived clones exhibit moderate to strong levels of strand displacement
- Metagenomics for enzyme discovery has the potential to unlock novel and diverse biochemical capabilities

Future Directions

- Further purification of virome-derived proteins to ensure target protein purity in stocks
- Quantify speed and processivity of all 762 types
- Investigate gene neighborhoods for virome sequences
- Explore implications of *in vitro* biochemistry to viral lifestyle and phenotype
- Integrate phenotypic and genotypic data to infer functionality of larger clades in PolA diversity tree (Fig. 4)

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grants 2025567 & 1736030 and a Joint Genome Institute Synthesis Project (FY17). Support from the University of Delaware Center for Bioinformatics and Computational Biology Core Facility, the University of Delaware Sequencing & Genotyping Center, and use of the BIOMIX compute cluster was made possible through funding from Delaware INBRE (NIGMS P20GM103446), the State of Delaware, and the Delaware Biotechnology Institute. Special thanks to all the team members at Varigen BioSciences and the past and present lab members of the Viral Ecology and Informatics Laboratory (VEIL) at the University of Delaware.

References

- Tabor, S., & Richardson, C. C. (1995). A single residue in DNA polymerases of the Escherichia coli DNA polymerase I family is critical for distinguishing between deoxy- and dideoxyribonucleotides. *Proceedings of the National Academy of Sciences*, *92*(14), 6339–6343.
- Tabor, S., & Richardson, C. C. (1987). DNA sequence analysis with a modified bacteriophage T7 DNA polymerase. *Proceedings of the National Academy of Sciences of the United States of America*, *84*(14), 4767–4771.
- Suzuki, M., Yoshida, S., Adman, E. T., Blank, A., & Loeb, L. A. (2000). Thermus aquaticus DNA polymerase I mutants with altered fidelity. Interacting mutations in the O-helix. *The Journal of biological chemistry*, *275*(42), 32728–32735.
- Moser, M.J., et al., (2012) Thermostable DNA polymerase from a viral metagenome is a potent rt-PCR enzyme PLoS One*7(6)*.
- Nasko, D. J., Chopyk, J., Sakowski, E. G., Ferrell, B. D., Polson, S. W., & Wommack, K. E. (2018). Family A DNA Polymerase Phylogeny Uncovers Diversity and Replication Gene Organization in the Virioplankton. *Frontiers in microbiology*, *9*, 3053.