

# A Novel Technology for Single Molecule Protein Sequencing

Swathi Ram<sup>1,2,3</sup>, Andy Mendoza<sup>1,2</sup>, Srivathsan Ranganthan<sup>1,2</sup>

Affiliations: 1- Cancer Early Detection Advanced Research (CEDAR); 2 - Knight Cancer Institute (KCI); 3 - Sunset High School



Building tomorrow's researchers today

### Background

Problem: Accurate representation of protein expression is crucial for understanding the molecular mechanisms behind biological processes, and identification of novel biomarkers for diagnostics and therapeutics. Next-Generation Sequencing (NGS) of DNA and RNA, has made large strides towards that goal, but fails to account for translational regulation, post translational modifications, and protein isoforms, that can amount to millions of protein variants. Current NGS platforms are not readily transferrable to direct protein sequencing.

Current landscape: Mass Spectrometry (MS) is the gold-standard for direct quantification of proteins. While MS has improved rapidly in the recent years, unlike NGS, it is not single-molecule and often relies on existing databases.<sup>2</sup>

Solution: This project aims to develop a de novo single molecule protein sequencing method by using a combination of biochemical processes that allow for fingerprinting of amino acids on a DNA molecule, which can then be sequenced on next-generation sequencing (NGS) platforms.

## Workflow E JULE JULE JULE DNA strands containing Unique Molecular Identifiers (UMIs) are created **PROSE** Amino acids are cleaved from a protein and attached to the DNA protein DNA PROPROPROPRO NO CONTRACTOR OF THE STATE OF T 200000000 This DNA-containing amino acid side "Daughter" strands that carry the chains-is directly sequenced on an fingerprint of the amino acids are Oxford Nanopore Sequencer created "Daughters" are sequenced on SBS A pipeline is used to filter out based NGS technologies like Illumina incomplete sequences and extract UMIs and fingerprints A neural network uses the fingerprints to identify the amino acids

## **Objectives**

- a. To increase the efficiency of the pipeline (in terms of time and resources)
- b. To improve the accuracy of the neural network

#### Methodology Based-called Downstream Demultiplex Alignment **Block processing** Adapter trimming reads (Guppy) analysis In-house **Feature Selection** None minimap2 Porechop for Machine **UMI** tools Cutadapt Learning Test 2: Compare different adapter Test 3: Compare alternate Test 1: Experiment

block processing

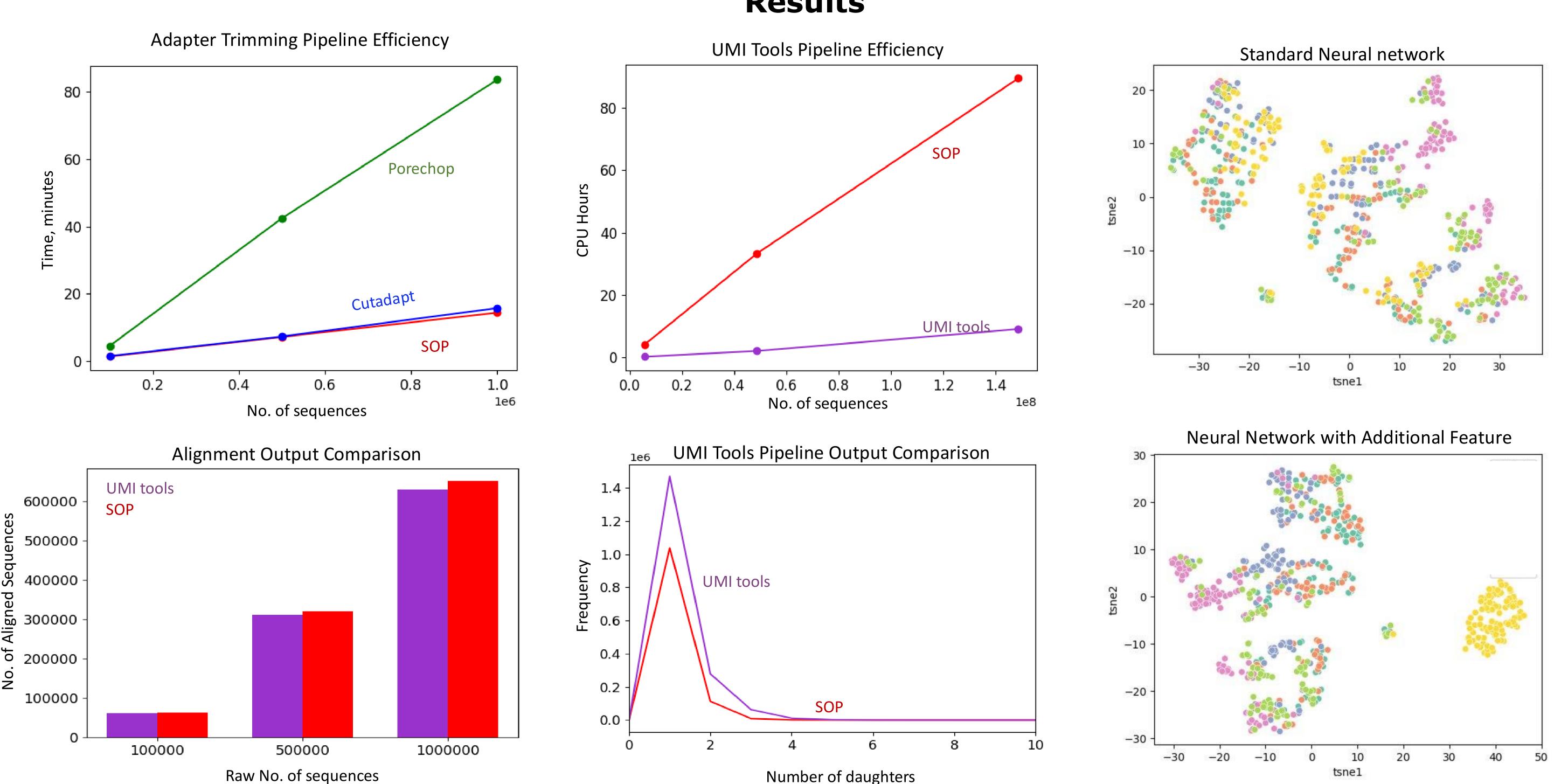
with the need for

trimming tools for efficiency and downstream processing with SOP (no trimming)

alignment tools with SOP for efficiency and accuracy

Test 4: Experiment with different features to see which are most relevant

### Results



### **Conclusions and Future Directions**

- Adapter trimming did not improve overall efficiency of the pipeline
- The new pipeline incorporating UMI tools was approximately 10 times faster than the SOP
- We significantly improved the accuracy of the neural network by identifying and adding a key feature to the dataset
- Next steps: Refine the biochemical process that controls the number of daughters created so that more daughters are created per mother

## Acknowledgments

- I would like to thank the PSI program managers and coordinators for this opportunity.
- I would like to thanks Sofio Vignolo for guiding me through this process.
- I would like to thank the CEDAR and KCI staff for providing me with support and resources

## References

- Alfaro, Javier Antonio, et al. "The emerging landscape of single-molecule protein sequencing technologies." Nature methods 18.6 (2021): 604-
- 2. "Protein Sequencing: Methods and Applications | Abcam." Abcam.com, 2023, https://www.abcam.com/en-us/knowledge-center/proteins-andprotein-analysis/protein-sequencing