DISSERTATION DEFENSE

Tim Dunn

Improving Select Applications of Long-Read DNA Sequencing

Monday, April 22, 2024
9:00am – 11:00am
3901 Beyster
Hybrid – Zoom

ABSTRACT: The cost of sequencing a human genome has plummeted over the past few decades from $300 million to well under $1,000. As a result of these cost reductions, genome sequencing is increasingly common across diverse fields including cancer diagnostics, rare disease and pathogen detection, and personalized medicine. Advances in sequencing technologies are concurrently enhancing capabilities. One such technology is nanopore-based long-read sequencing, which can perform unamplified single-molecule sequencing with theoretically unlimited read lengths. Such long reads simplify genome assembly and reveal complex genetic variations previously difficult to analyze. Since its 2015 release with around 85% per-base accuracy, nanopore technology has achieved over 99% accuracy due to advanced basecallers and provides portable, real-time data processing.

Despite its promise, long-read sequencing faces challenges. This thesis addresses these by developing new methods for long-read applications. (1) First, we introduce SquiggleFilter, a hardware-accelerated filter for real-time virus detection. We show that our 14.3W accelerator has 274× greater throughput and 3481× lower latency than existing GPU-based solutions while consuming half the power, enabling efficient pathogen detection for the next generation of nanopore sequencers. (2) Next, we present nPoRe, a novel read alignment algorithm for repetitive genomic regions. When used in combination with haplotype phasing, nPoRe improves tandem repeat variant recall from 63.8% to 73.0%. (3) Then, we introduce vcfdist, a benchmarking tool that consistently evaluates variant calling performance regardless of complex variant representation. (4) Lastly, we extend vcfdist to jointly evaluate both small and structural variants, enhancing accuracy and phasing analyses. We find that a joint evaluation of small and structural variants uniformly reduces measured false positive and false negative errors by at least 20% across all variant types, and that vcfdist reduces false positive measured flip errors by over 50%. We find that vcfdist is more accurate than previously published works and on par with the newest approaches, but with improved result interpretability.

CHAIR: Prof. Satish Narayanasamy