When less enables more : making models and methods for modern genomics

The plummeting cost of high-throughput sequencing and the astounding variety of available assays has created a scientific regime in which the bottleneck in many experiments has ceased to be our ability to acquire data, and has instead become the computational costs associated with analyzing this data. Simultaneously, we have been building sequencing data archives that hold immense potential, but which remain largely inert due to our inability to efficiently index and query "raw" experimental data.

In this talk, I will discuss some of the methods that we have been developing to address these challenges as they arise in different contexts. I will highlight our recent work in fast, accurate and bias-aware methods for transcript quantification, as implemented in our tool Salmon. I will discuss Mantis, our indexing approach to enable sequence search over large collections of raw, unassembled read data. Finally, I will describe Pufferfish, a new time and space-efficient data structure for indexing and querying the colored, compacted de Bruijn graph.

Relevant Papers:

• "Salmon provides fast and bias-aware quantification of transcript expression" : https://www.nature.com/articles/nmeth.4197
• "A General-Purpose Counting Filter: Making Every Bit Count" : https://dl.acm.org/citation.cfm?id=3035963
• "Mantis: A Fast, Small, and Exact Large-Scale Sequence Search Index" : https://www.biorxiv.org/content/early/2017/11/10/217372
• "A space and time-efficient index for the compacted colored de Bruijn graph" : https://www.biorxiv.org/content/early/2017/09/21/191874

Bio:

Rob Patro is an Assistant Professor of Computer Science at Stony Brook University, where he heads the Computational Biology and Network Evolution (COMBINE) lab. Prior to joining Stony
Brook, Rob obtained his Ph.D. in Computer Science from the University of Maryland. He was a postdoctoral research associate in the Kingsford Group at the Lane Center for Computational Biol-
ogy (now the Department of Computational Biology) at Carnegie Mellon University. His research interests are in the design of algorithms and data structures for processing, organizing, indexing
and querying high-throughput genomics data. He is also interested in the intersection between efficient algorithms and statistical inference. He is the recipient of an NSF CAREER award in 2018.