Scaling ABySS to longer reads using spaced k-mers and Bloom filters

Shaun D Jackman1,2, Kartikha Raghavan1, Benjamin P Vandervalk1, Daniel Paulino1, Justin Chu1,2, Hamid Mohamadi1,2, Anthony Raymond1, René L Warren1 and Inanç Birol1,2
1Genome Sciences Centre, British Columbia Cancer Agency, Vancouver, BC V5Z 4S6, Canada
2University of British Columbia, Vancouver, BC V6T 1Z4, Canada

Abstract
• Longer reads require larger values of k for de Bruijn Graph (dBG) assembly
• Larger k usually require more memory
• ABySS uses two memory-efficient data structures
  • Spaced-seed dBG
  • Bloom filter dBG
• whose memory usage is independent of k
• Suitable for assembly of
  • Overlapping paired MiSeq
  • Moleculo
  • Corrected Oxford Nanopore

Methods
• Spaced seed dBG stores the first and last few nucleotides of each k-mer
  • For example, a 364-mer of two 32-mer separated by a 300 nucleotide gap
• Uses a fraction the memory of standard dBG
• ABySS-Sealer fills scaffold gaps
  • Navigates a path between contigs through a Bloom filter dBG
• Memory efficiency permits using multiple Bloom filter dBG for different values of k
  • Small k span low-coverage regions
  • Large k resolve repeats

Conclusion
• ABySS can assemble long reads
• Memory use is independent of k

Results

Sealer: Bloom Filter de Bruijn Graph

References
• SPAdes, Bankevich et al. (2012) doi:10.1089/cmb.2012.0021
• Nanocorr, Goodwin et al. (2015) doi:10.1101/013490
• Celera Assembler, Myers et al. (2000) doi:10.1126/science.287.5461.2196