Efficient targeted error resolution and automated finishing of long read genome sequence assemblies

Li JX[1,2], Warren RL[1], Coombe L[1], Wong J[1] & Birol I[1,3]

1 Genome Sciences Center, British Columbia Cancer Agency, Vancouver, BC, Canada
2 Bioinformatics Graduate Program, University of British Columbia, Vancouver, BC, Canada
3 Department of Medical Genetics, University of British Columbia, Vancouver, BC, Canada

Contact:
janli@bcgsc.ca
www.birollab.ca

Introduction

	Short Reads	Long Reads
Pros	• High accuracy (>99.9%) • High throughput	• Provides long-range information for genome assembly
Cons	• Difficulties resolving repeat regions • Amplification bias	• Lower accuracy (87-98%) • Requires error correction

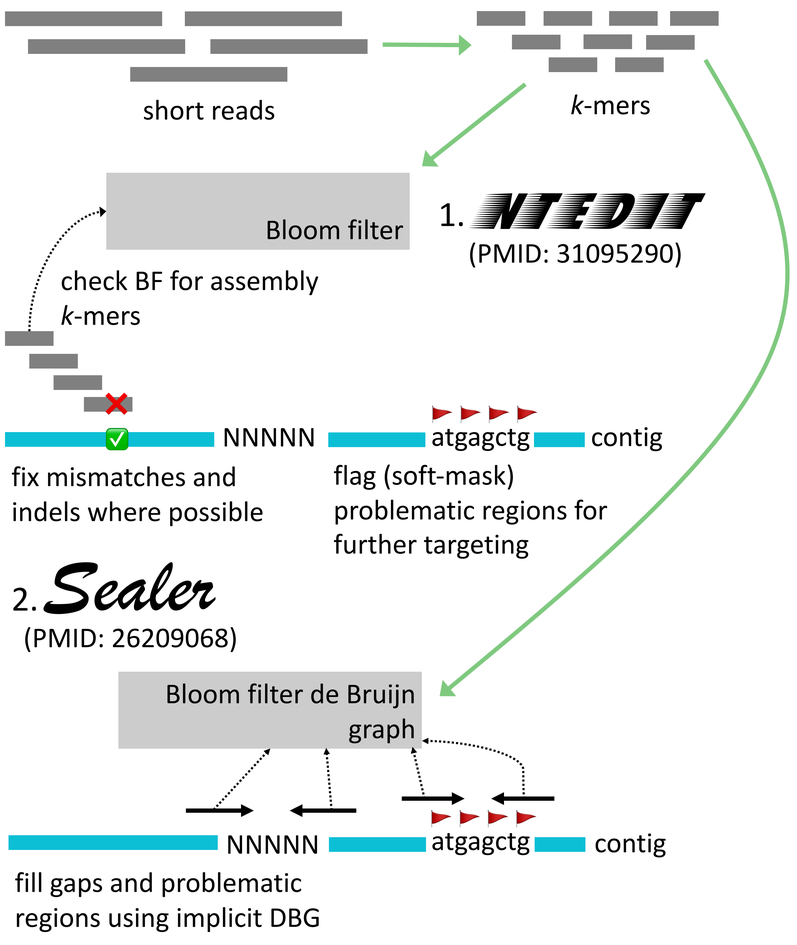
Motivation

- Long read assemblies are often polished using short reads
- Existing polishing tools such as Racon (PMID: 28100585) rely on read alignments and are memory intensive

Objective

- Develop a **scalable**, **alignment-free** protocol for polishing and finishing long read genome assemblies using short reads

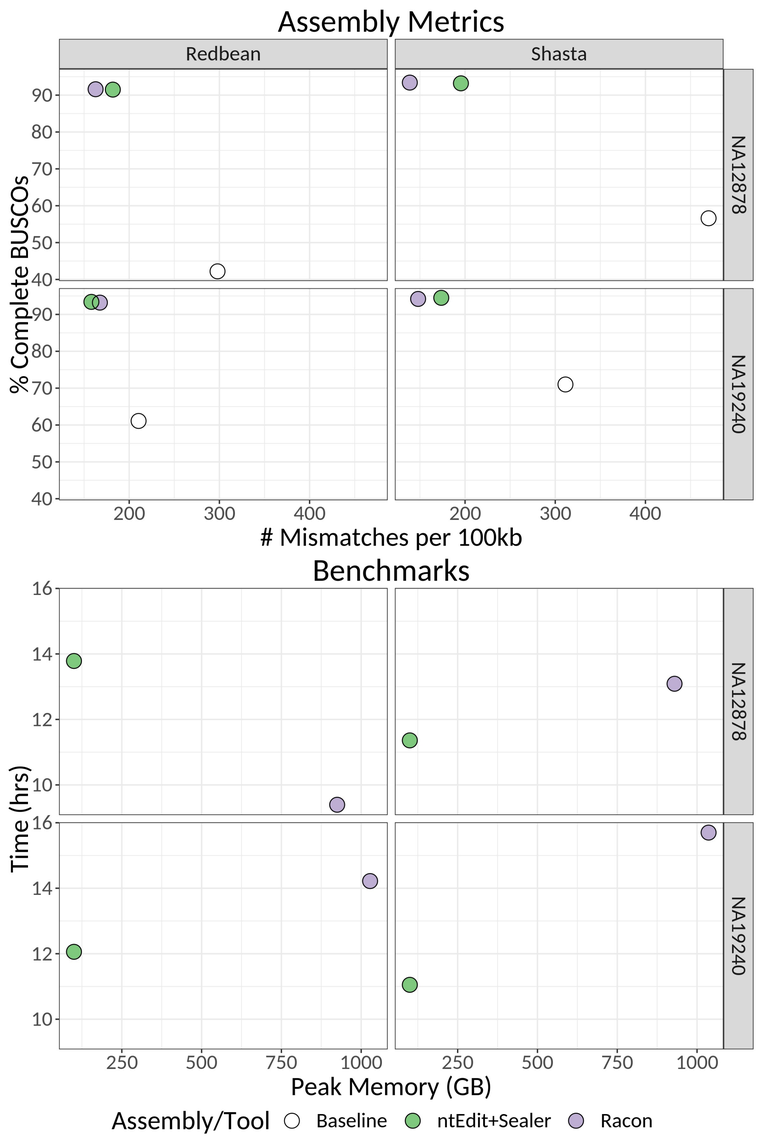
Methods

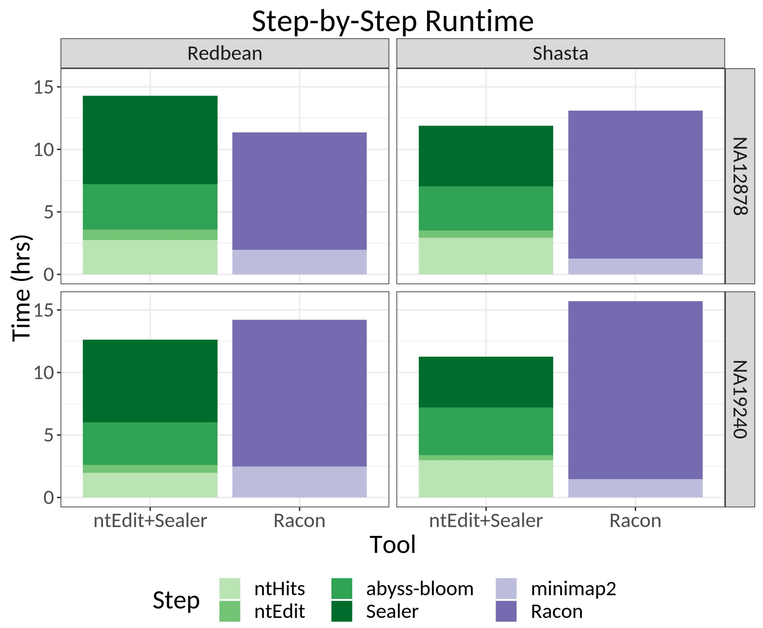


Results

- Baselines assembled with Shasta (PMID: 32686750) and Wtdbg2/Redbean (PMID: 31819265)
- ntEdit+Sealer *k*-mer lengths: *k*=80, *k*=65, *k*=50
- Both tools run with 48 threads
- Assembly quality assessed with BUSCO (PMID: 26059717) and QUAST (PMID: 23422339)

Individual	Illumina read coverage
NA12878	54X
NA19240	62X





- ntHits and abyss-bloom are the Bloom filter creation steps for ntEdit and Sealer respectively
- 23.5-49.3% additional BUSCO genes recovered
- QUAST mismatched bases reduced by up to 58.4%
- ntEdit+Sealer required significantly less RAM than Racon (~100GB vs. ~1TB)

Conclusion

- ntEdit+Sealer produces **highly complete** assemblies that are comparable in quality to Racon while requiring much less memory and often less time
- This protocol provides a **scalable** and **accessible** solution for the targeted resolution of errors in long read genome assemblies

ntEdit
https://github.com/bcgsc/ntEdit
conda install –c bioconda ntedit
brew install brewsci/bio/ntedit

Sealer (available within ABySS)
https://github.com/bcgsc/abyss
conda install –c bioconda abyss
brew install abyss

Funding

GenomeCanada
Genome BritishColumbia
National Institutes of Health