NanoSpring: reference-free lossless compression of nanopore sequencing reads using an approximate assembly approach

Qingxi Meng, Shubham Chandak, Yifan Zhu, Tsachy Weissman
Department of Electrical Engineering, Stanford University

Introduction

- Nanopore sequencing, specifically using Oxford Nanopore Technologies (ONT) sequencers has seen increasing adoption.
- A typical FASTQ dataset for the whole human genome requires hundreds of GBs of storage space (for a typical sequencing coverage of 30×).
- Compression of FASTQ files is crucial for storage and sharing of genomic data.

Nanopore Sequencing

Portable

Real time

Long reads (~10kb-100kb)

Native DNA & direct RNA sequencing

High error rates (5-10%)

General-purpose sequencers cannot exploit the redundancy in nanopore reads, and specialized read compressors for short reads do not work for long nanopore reads with insertion, deletion and substitution errors!

Our Contributions

Human whole genome 26x nanopore sequencing reads **84 GB**

NanoSpring reference-free lossless compression

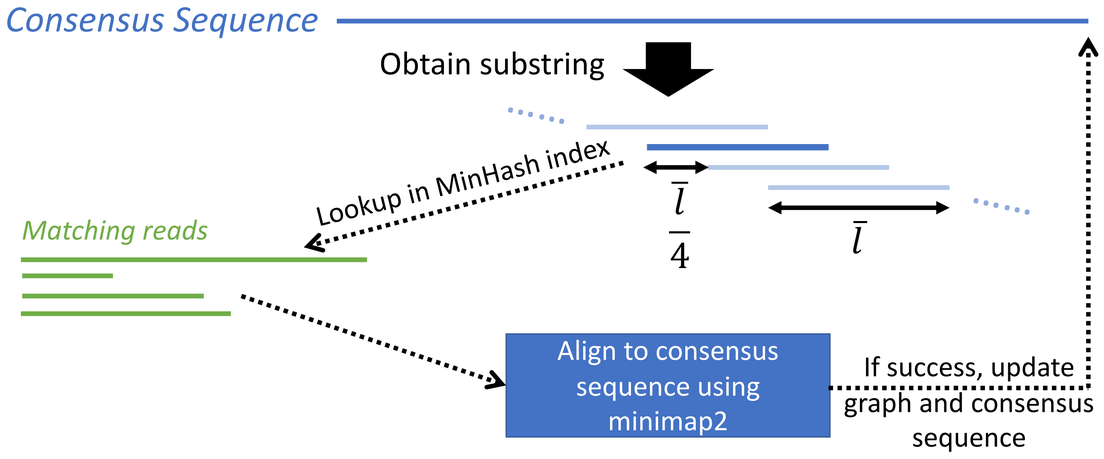
~3 hours
39 GB peak memory
(using 20 threads)

NanoSpring archive **7 GB** (3x smaller than Gzip)

NanoSpring
`https://github.com/qm2/NanoSpring`

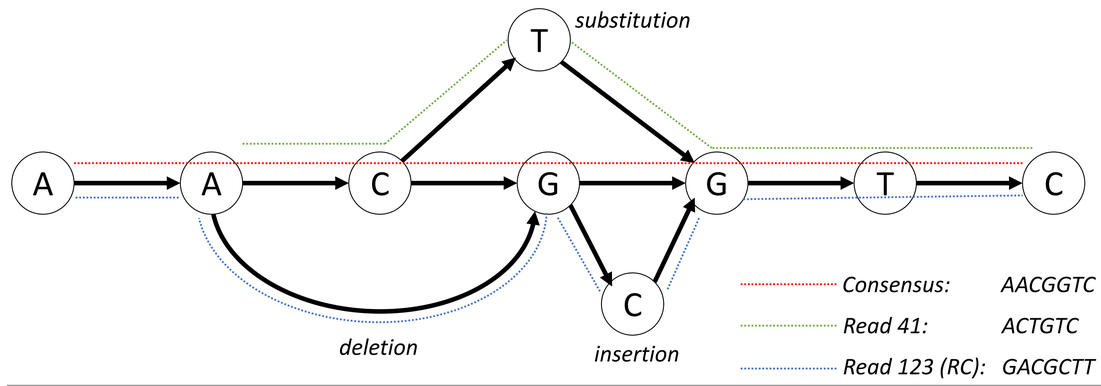
Methods - overall process

Approximate assembly - contig generation



Methods - consensus graph and encoded streams

Consensus graph



Encoded streams for contig

Consensus sequence: AACGGTC

Start position	1	0
Error type	sub	del, ins
Error position	2	2,2
Error base	T	C
RC	False	True
Read index	41	123

Experimental Results

Comparison of NanoSpring performance on 4 different whole human genome datasets

Dataset Sample	Coverage	Uncompressed size (GB)	Compressed size in bits/base			Improvement over ENANO
			Gzip	ENANO	NanoSpring	
NA12878	42x	132.9	2.24	1.89	**1.45**	1.30x
GM24385	26x	84.2	2.20	1.87	**0.66**	2.83x
CHM13	23x	74.2	2.18	1.86	**0.68**	2.74x
CHM13	46x	148.4	2.18	1.86	**0.60**	3.10x

NanoSpring achieves close to 3x improvement in compression over state-of-the-art compressors, with best results at higher coverages and for high quality reads from recent basecallers.

Detailed results for different species and analysis of various parameters available in preprint.

Future Work

Incorporate NanoSpring into a full-fledged FASTQ compressor capable of handling quality scores and read identifiers.

Funding

The authors acknowledge funding from Philips.

References

1. **Meng, Q. *et al.* (2021). NanoSpring: reference-free lossless compression of nanopore sequencing reads using anapproximate assembly approach.** ***bioRxiv*** `https://www.biorxiv.org/content/10.1101/2021.06.09.447198v1` ***(submitted to Bioinformatics)***
2. Dufort y Álvarez, G. *et al.* (2020). ENANO: Encoder for NANOpore FASTQ files. *Bioinformatics*, **36**(16), 4506–4507.
3. Berlin, K. *et al.* (2015). Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat Biotechnol*, **33**(6), 623–630.
4. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**(18), 3094–3100.