A Simple and Effective Technique for Assisted Genome Assembly

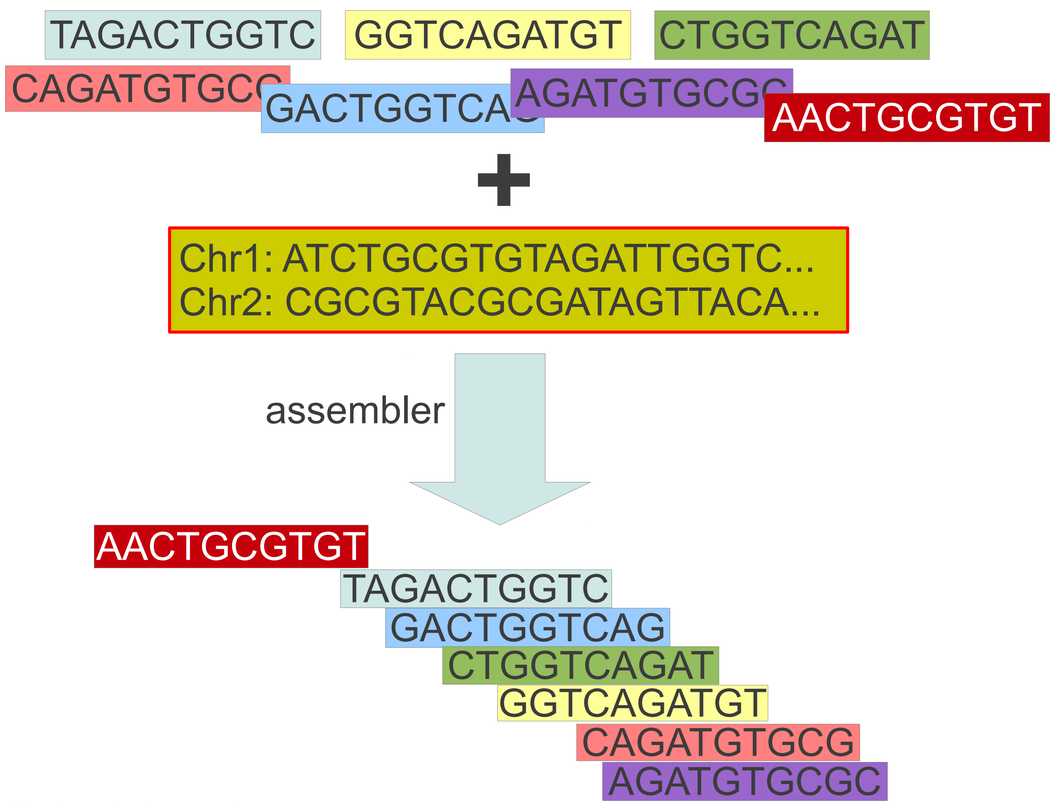
Krisztian Buza, Bartek Wilczyński, Norbert Dojer
Computational Biology and Bioinformatics
Faculty of Mathematics, Informatics and Mechanics, University of Warsaw (MIMUW)
chrisbuza@yahoo.com, bartek@mimuw.edu.pl, dojer@mimuw.edu.pl

Summary

We propose a simple technique for assisted genome assembly. Our technique is based on generation of artificial reads from the reference genome. According to our experiments, our method outperforms Amos in cases where very few reads are available and the target genome is relatively closely related to the reference genome.

Background: assisted genome assembly

Input: short reads
+ genome of a related organism (reference genome)

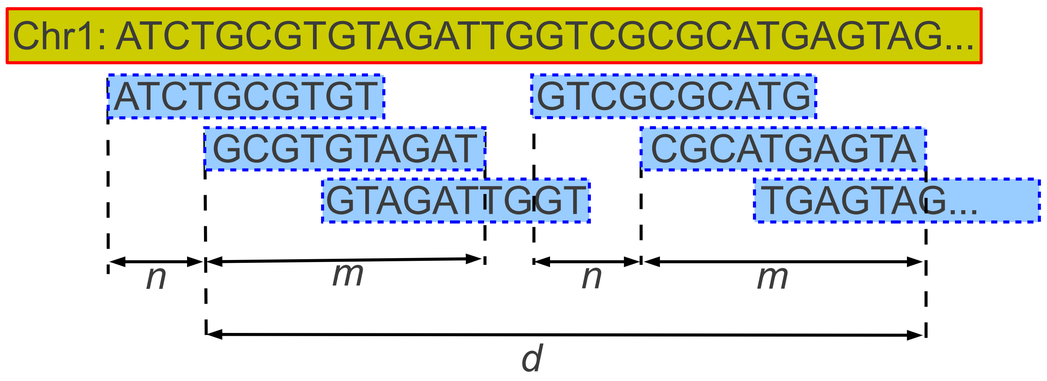


Output: target genome

contig1: AACTGCGTGTAGACTGGTCCTGGTCA
GATGTGCGC...

Our approach: Simple Assistance

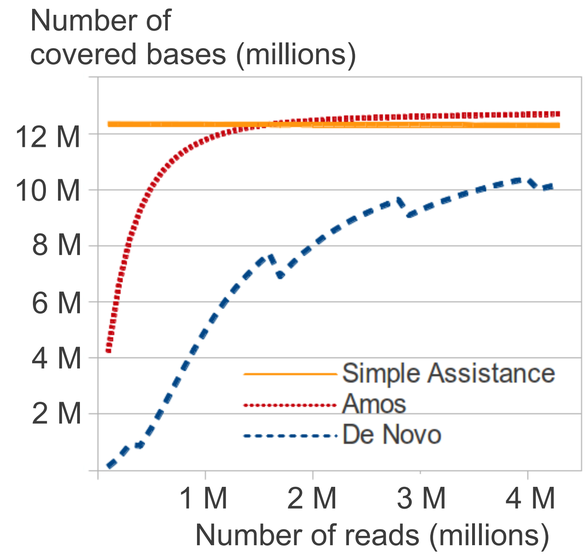
- Generate artificial reads from the reference with low quality scores (“real” reads have priority over the artificial ones)
- Add the artificial reads to the input of a (de-novo) assembler

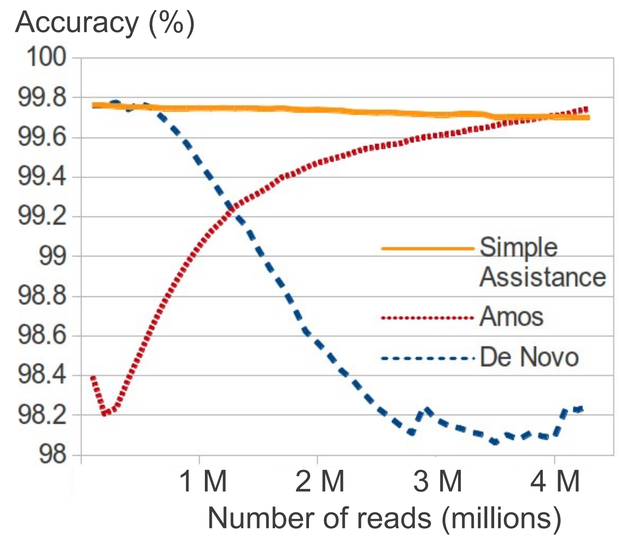


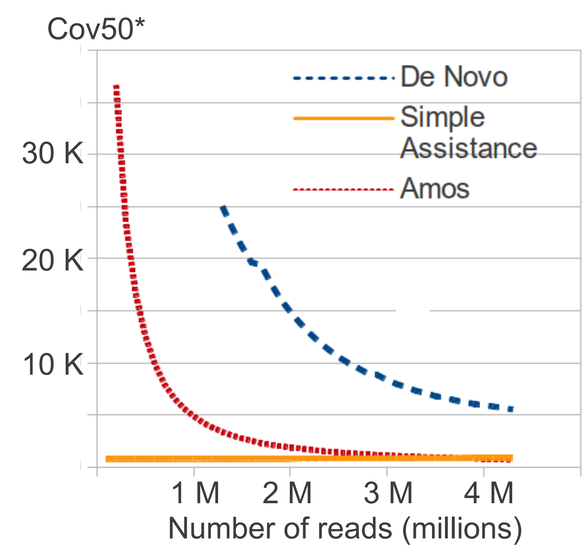
	d	m	n	Quality score*
S. pombe	3000	1000	500	5
A. thaliana	–	1000	500	5

* (Range of quality scores: 0..93)

Genome assembly







- Benchmark: assembly of the genome of *S. Pombe*-HP
- Gold standard: assembly produced by Amos using all the reads
- With Cov50* we mean the number of largest contigs that cover together 50% of the gold standard.
- Our simple assistance (using Velvet as assembler) outperforms both (i) the de novo assembler Velvet, and (ii) the contigs of the assisted assembler Amos for the case when only few reads are available.

Assembly for mapping

- We produced the assemblies from the input reads
- We mapped the IP-reads with Bowtie2

	Assembly	**Uniquely mappable**		**“Extra mappable”**
	(from 0.5M and 10M reads resp.)	to the assembly	to the reference	
S. Pombe -HP (~12M)	Amos, repl.1	465 581	426 465	22 295
	Simple A., repl. 1	1 062 969		175 551
	Amos, repl.2	421 997	365 802	25 400
	Simple A., repl. 2	1 409 327		295 749
S. Pombe -Mmi1 (~12M)	Amos, repl.1	681 272	692 239	21 019
	Simple A., repl. 1	1 959 980		593 627
	Amos, repl.2	1 126 555	1 118 799	26 939
	Simple A., repl. 2	2 403 156		450 995
A. thaliana- cell line (~150M)	Amos, sample 1	54 316 189	57 855 074	732 040
	Simple A., s. 1	71 285 707		13 145 143
	Amos, sample 2	72 239 517	76 318 470	762 523
	Simple A., s. 2	87 399 447		10 818 192
	Amos, sample 3	64 660 055	68 723 765	816 757
	Simple A., s. 3	82 407 548		13 063 222

“Extra mappable” - reads that could not be mapped uniquely to the reference directly, but could be mapped uniquely to the reference via mapping to the assembly

Acknowledgement

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

FNP
Fundacja na rzecz Nauki Polskiej

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

This project was supported by the Foundation for Polish Science within the Skills programme co-financed by the European Union European Cohesion Fund.

References

- Nathaniel Parrish, Benjamin Sudakov and Eleazar Eskin (2013): *Genome reassembly with high-throughput sequencing data*, The Eleventh Asia Pacific Bioinformatics Conference
- Sante Gnerre, Eric S. Lander, Kerstin Lindblad-Toh, David B. Jaffe (2009): *Assisted assembly: how to improve a de novo genome assembly by using related species*, Genome Biology, 10:R88
- Mihai Pop, Adam Phillippy, Arthur L. Delcher, Steven L. Salzberg (2004): *Comparative genome assembly*, Briefings in Bioinformatics, Vol 5. No 3.