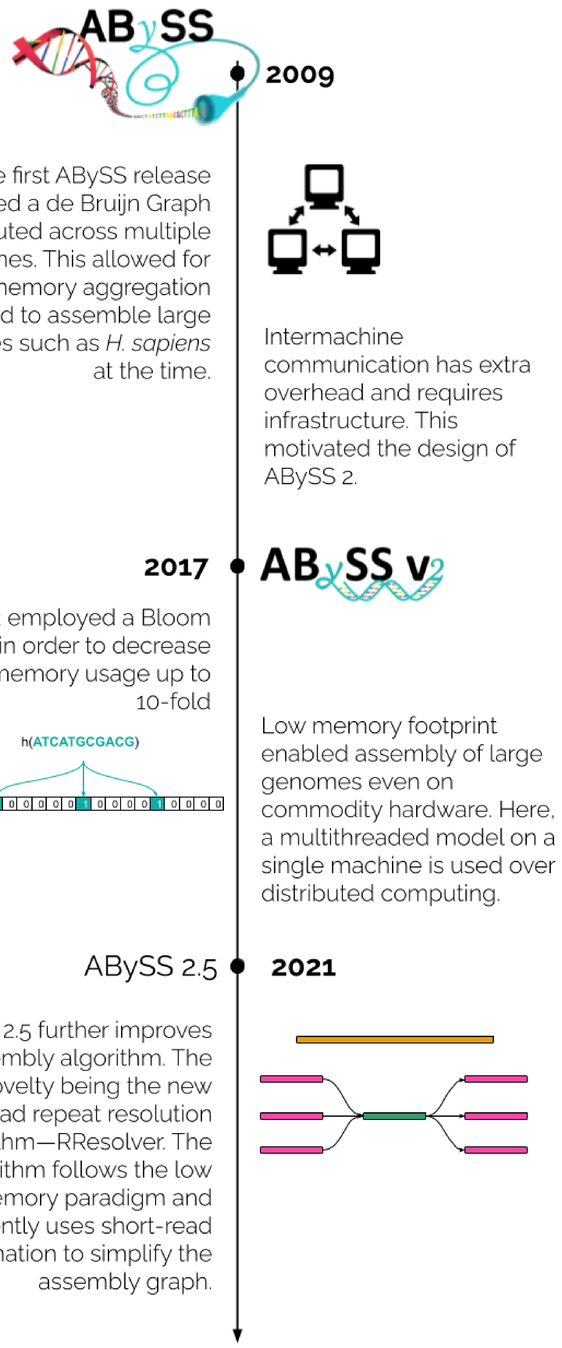
ABySS 2.5: Efficient repeat resolution with short reads

Vladimir Nikolić [1] Justin Chu [1] Johnathan Wong [1] Lauren Coombe [1] Amirhossein Afshinfard [1] Ka Ming Nip [1] René Warren [1] Inanç Birol [1]

[1]Canada's Michael Smith Genome Sciences Centre at BC Cancer, Vancouver, BC V5Z 4S6, Canada

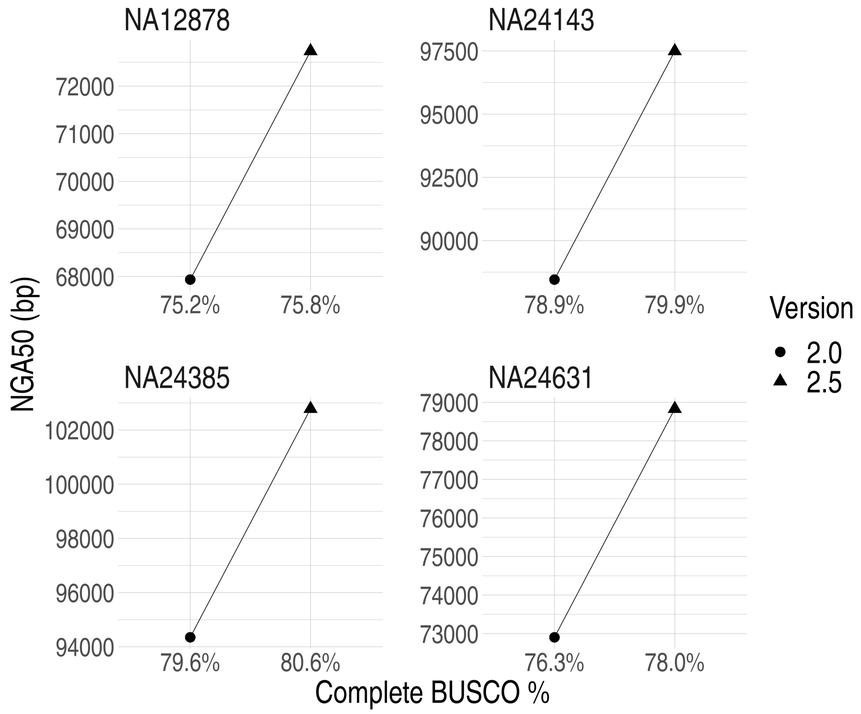
Introduction

- It has been twelve years since the publication of the ABySS short-read *de novo* genome assembler, and four since its successor, ABySS 2.

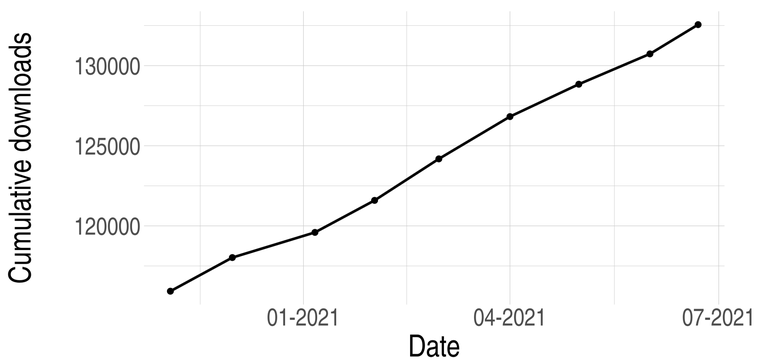


Assembly quality

- To demonstrate assembly quality improvement, comparison between ABySS 2.0 and ABySS 2.5 was made in NGA50 contiguity and complete BUSCOs recovered. *H. sapiens* 150bp paired-end reads from NA12878 and NA24631, and 250bp reads from NA24143 and NA24385 were assembled.

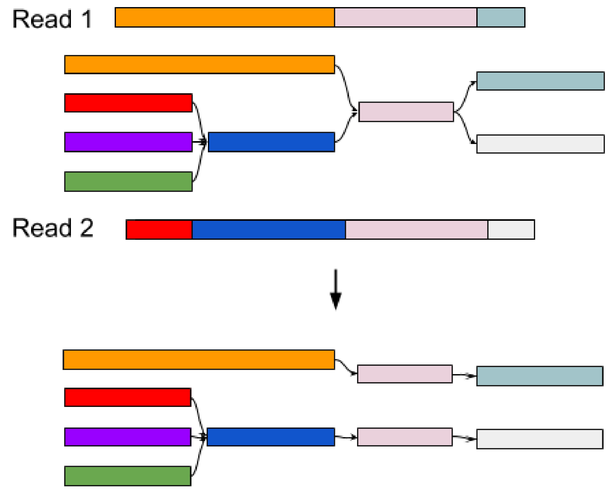


- Further, we show the download counts for the ABySS assembler over time, demonstrating the relevance of the tool more than a decade later.



Repeat resolution algorithm

- The RResolver algorithm utilizes the short-read range information to find paths in the assembly graph. The de Bruijn Graph node size is currently always shorter than the read size, and so some information is left unused, which the RResolver algorithm recovers.



- By identifying the presence of reads that span across the repeat, RResolver duplicates the repeat sequence where each instance is joined with corresponding graph nodes.
- The simplified graph allows for easier navigation and subsequently more correct decision making.

Conclusions

- From the initial release to ABySS 2.0, and now ABySS 2.5, the assembler has come a long way in delivering high quality *de novo* genome assemblies with low resource usage.
- The latest addition to the ABySS pipeline, the RResolver algorithm, further improves assembly quality, while keeping the resource usage low.

Funding

Genome BritishColumbia
National Institutes of Health
GenomeCanada
CIHR IRSC
Canadian Institutes of Health Research
Instituts de recherche en santé du Canada