Hybrid correction of highly noisy Oxford Nanopore long reads using a variable-order de Bruijn graph

Supplementary material

Pierre Morisse[1], Thierry Lecroq[1] and Arnaud Lefebvre[1]

[1]Normandie Univ, UNIROUEN, LITIS, 76000 Rouen, France

Dataset	*A. baylyi*	*E. coli*	*S. cerevisae*	*C. elegans*
Reference organism				
Strain	ADP1	K-12 substr. MG1655	W303	Bristol N2
Reference sequence	CR543861	NC_000913	scf7180000000{084-13}	GCA_000002985.3
Genome size	3.6 Mbp	4.6 Mbp	12.2 Mbp	100 Mbp
Oxford Nanopore data				
Accession number	ERR77685{1-5} Genoscope[1]	Genoscope[2] Sequences from Loman Lab	Genoscope[3] Sequences from Schatz Lab	ERR1802061[4]
Number of reads	89,011	22,270	205,923	363,500
Average length	4,284	5,999	5,698	5,524
Number of bases	381 Mbp	134 Mbp	1,173 Mbp	2,008 Mbp
Coverage	106x	29x	96x	20x
Illumina data				
Accession number	ERR788913[4]	Genoscope[5] Sequences from Loman Lab	Genoscope[6] Sequences from Schatz Lab	ART
Number of reads	900,000	775,500	2,500,000	20,057,100
Read length	250	300	250	250
Number of bases	224 Mbp	232 Mbp	625 Mbp	5,000 Mbp
Coverage	50x	50x	50x	50x

Table 1: Description of the data used in the experiments.
[1]http://www.genoscope.cns.fr/externe/nas/datasets/MinION/acineto/, reads from run6.
[2]http://www.genoscope.cns.fr/externe/nas/datasets/MinION/ecoli/
[3]http://www.genoscope.cns.fr/externe/nas/datasets/MinION/yeast/
[4]Only a subset of the data was used.
[5]http://www.genoscope.cns.fr/externe/nas/datasets/Illumina/ecoli/
[6]http://www.genoscope.cns.fr/externe/nas/datasets/Illumina/yeast/

Pre-processing	QuorUM	Karect
A. baylyi		
Number of reads	16,618	16,618
Split reads (%)	4.90	4.86
Average length	10,260	10,260
Number of bases (Mbp)	179	179
Average identity (%)	99.40	99.40
Genome coverage (%)	99.82	99.80
E. coli		
Number of reads	21,005	21,006
Split reads (%)	4.98	4.88
Average length	5,797	5,794
Number of bases (Mbp)	128	128
Average identity (%)	99.81	99.81
Genome coverage (%)	99.43	99.41
S. cerevisae		
Number of reads	33,484	33,250
Split reads (%)	11.47	10.55
Average length	6,455	6,613
Number of bases (Mbp)	243	244
Average identity (%)	99.54	99.55
Genome coverage (%)	93.32	93.19

Table 2: Comparison of the results obtained with Jabba, when correcting the short with QuorUM or with Karect.

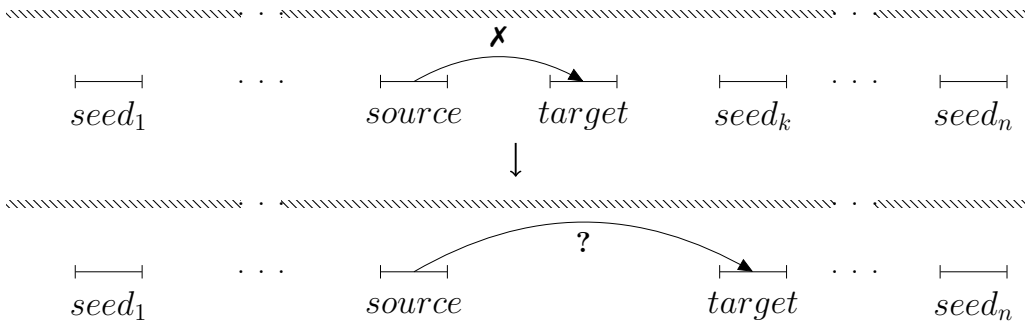

Figure 1: Illustration of the process of skipping a seed. Hatched lines represent the long read and segments represent the seeds. Top: No path allowing to link *source* to *target* has been found. *target* is thus considered as erroneous and is ignored. Bottom: *target* is redefined as $seed_k$, whereas *source* remains unchanged. A new linking iteration is then performed between these two seeds.

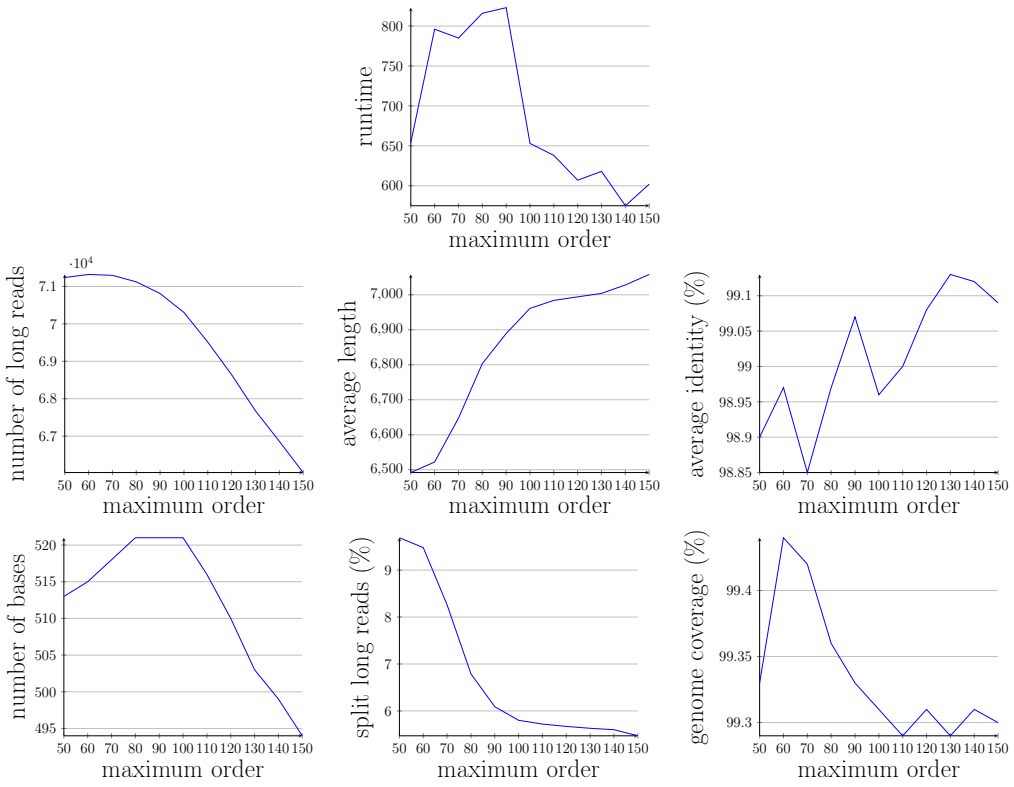

Figure 2: Impact of the maximum order of the graph on the results, when fixing other parameters. To obtain fair comparisons, the minimum order of the graph was set to half of the maximum order for each experiment. Runtimes are reported for the execution of the whole correction pipeline. We acknowledge that a maximum order of 100 yields shorter reads, displaying a lower identity than higher orders, but as higher orders correct less long reads, setting the maximum order to 100 offers a good compromise. Same goes for lower orders correcting more long reads, and yielding a higher coverage of the reference genome, but displaying longer runtimes, and higher proportions of split long reads.

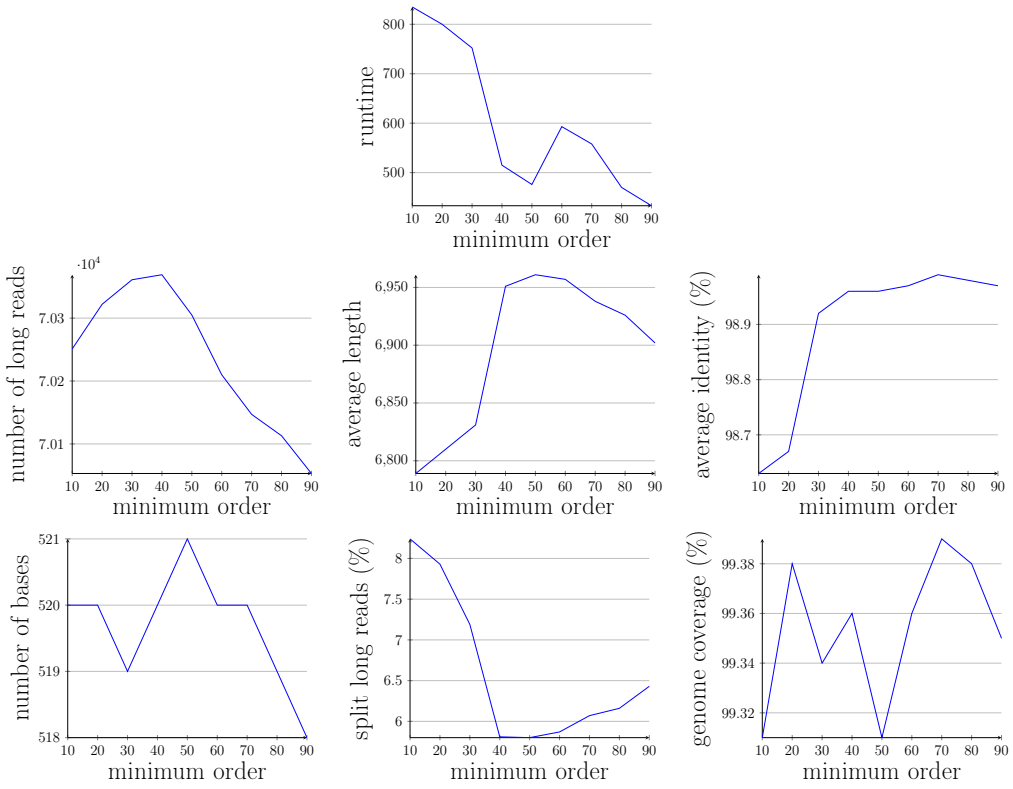

Figure 3: Impact of the minimum order of the graph on the results, when fixing other parameters. Rutimes are reported for the execution of the seeds linking and tips extension steps only. The fact that a minimum order of 50 covers the reference genome slightly less than other orders comes from the fact it displays the lowest proportion of split long reads, and therefore avoids spurious mappings of long reads fragments in wrong regions. A minimum order of 50 was chosen over 40, as, even though the two values display close statistics, a minimum order of 50 allows to produce longer, less split long reads, correct more bases, and run in a shorter time, despite correcting slightly less long reads.

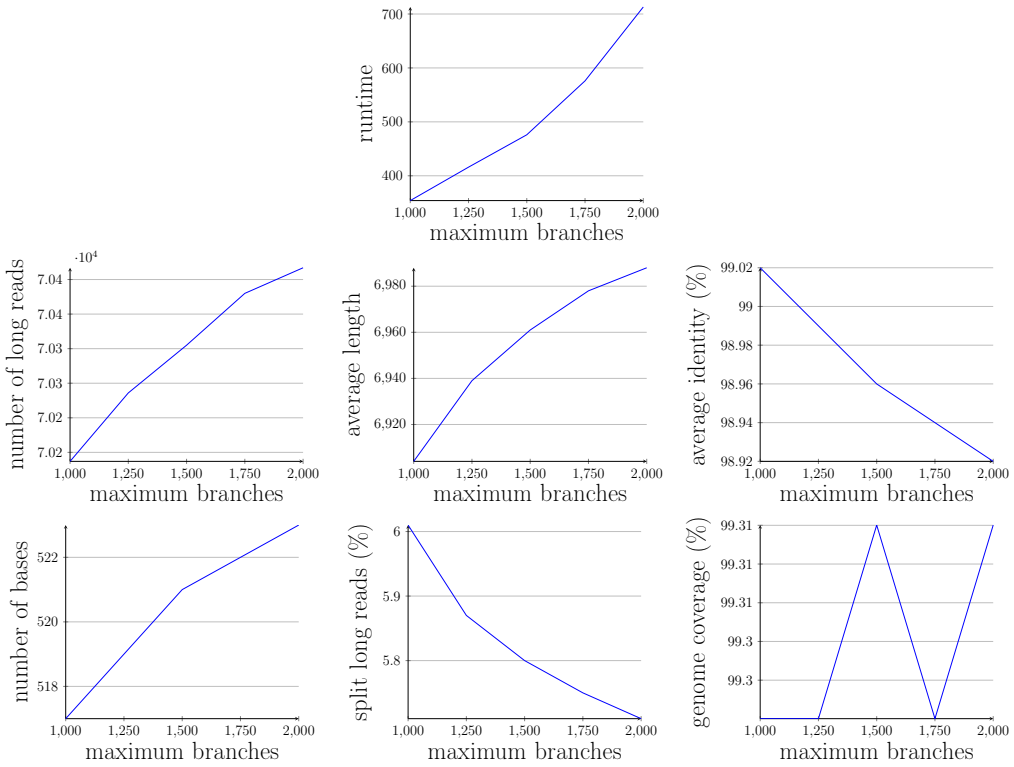

Figure 4: Impact of the maximum number of branches explorations on the results, when fixing other parmeters. Rutimes are reported for the execution of the seeds linking and tips extension steps only.