



HAL
open science

ELECTOR: EvaLuation of Error Correction Tools for lOng Reads

Lolita Lecompte, Camille Marchet, Pierre Morisse, Antoine Limasset, Pierre Peterlongo, Arnaud Lefebvre, Thierry Lecroq

► **To cite this version:**

Lolita Lecompte, Camille Marchet, Pierre Morisse, Antoine Limasset, Pierre Peterlongo, et al.. ELECTOR: EvaLuation of Error Correction Tools for lOng Reads. JOBIM 2018 - Journées Ouvertes Biologie, Informatique et Mathématiques, Jul 2018, Marseille, France. pp.1-2. hal-01929900

HAL Id: hal-01929900

<https://hal.science/hal-01929900>

Submitted on 21 Nov 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ELECTOR: EvaLUation of Error Correction Tools for lOng Reads

Lolita Lecompte¹, Camille Marchet¹, Pierre Morisse², Antoine Limasset³,
Pierre Peterlongo¹, Arnaud Lefebvre², Thierry Lecroq²

¹Univ Rennes, Inria, CNRS, IRISA, F-35000 Rennes, France ²Normandie Univ, UNIROUEN, LITIS, 76000 Rouen, France ³Université Libre de Bruxelles



1. Introduction

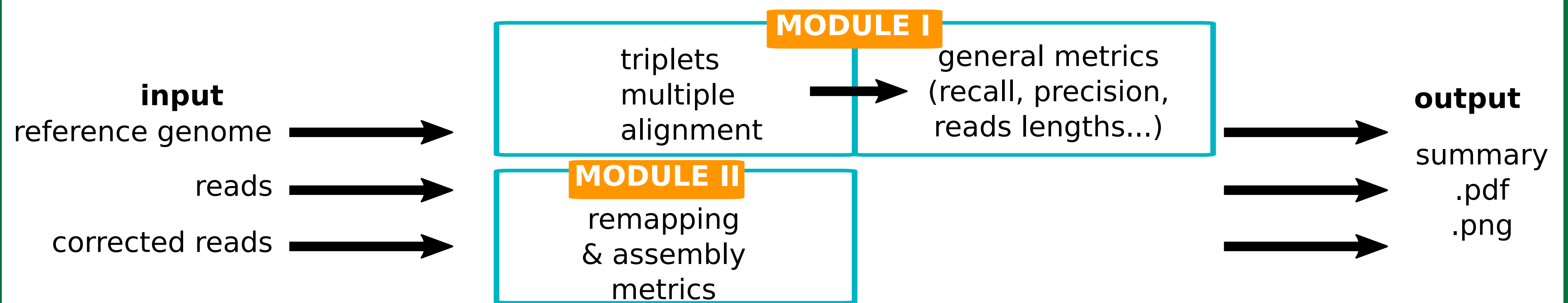
- ▶ Long read technologies, Pacific Biosciences and Oxford Nanopore, have **high error rates** (from 9% to 30%)
- ▶ Multiple error correction methods exist
- ▶ Important to **assess the correction stage** for downstream analyses
- ▶ Only one tool: **LRCstats** [1]
 - shows global correction gain
 - does not give access to correctors detailed behavior
 - high computation times

Developing methods allowing to **evaluate error correction tools with precise and reliable statistics** is therefore a crucial need.

2. Contribution

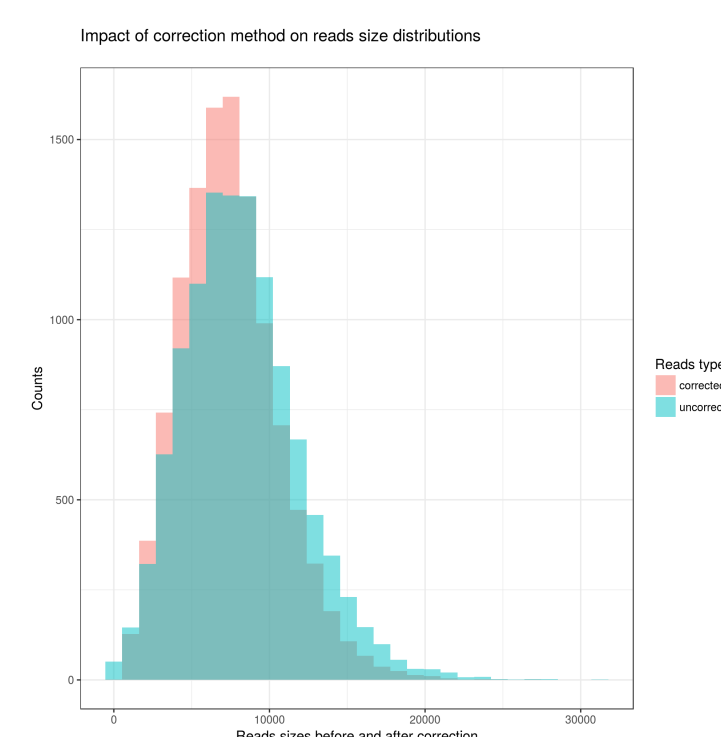
We propose **ELECTOR**, a novel tool that enables the evaluation of long read correction methods:

- ▶ provide **more metrics** than LRCstats on the correction quality
- ▶ **scale** to very long reads and large datasets
- ▶ **compatible** with a wide range of state-of-the-art error correction tools (hybrid/self)



3. Output statistics

- ▶ **Recall**
- ▶ **Precision**
- ▶ Overall correct bases rate
- ▶ GC content before and after correction
- ▶ Number of trimmed and/or split corrected reads
- ▶ Mean missing size in trimmed/split reads



Assembly using *Miniasm*

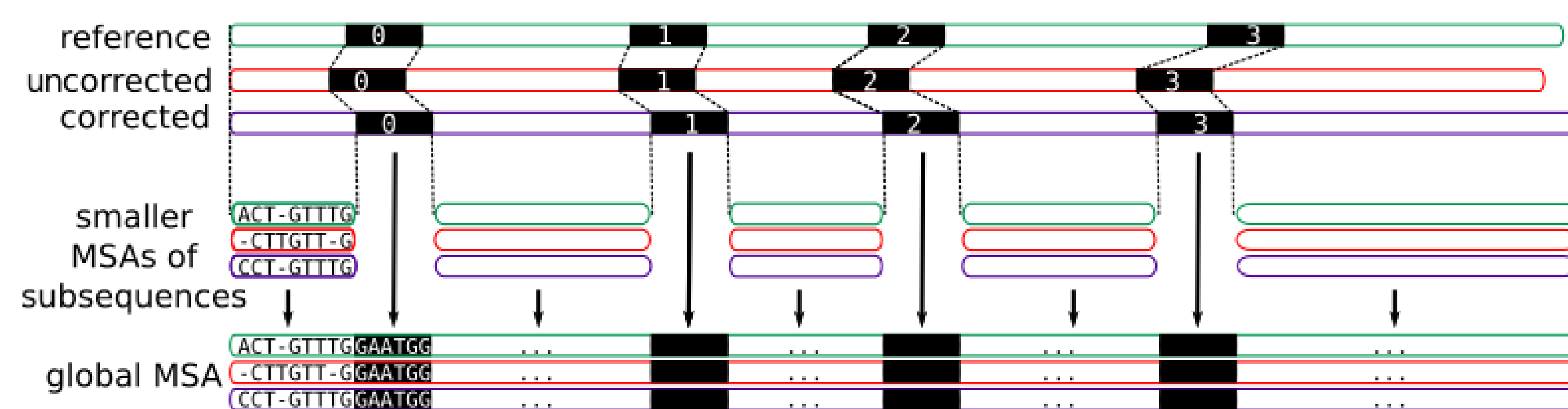
- ▶ Nb of contigs
- ▶ N50
- ▶ N75

If reference, remapping using *BWA*

- ▶ Average identity
- ▶ Genome coverage
- ▶ Nb of breakpoints
- ▶ NGA50
- ▶ NGA75

4. Methods

- Multiple alignment of triplets: {*reference read*, *uncorrected read*, *corrected read*}
- Seed-MSA strategy**: multiple sequence alignment (MSA) using partial order graphs [2] coupled to a seed strategy comparable to MUMmer or Minimap.
 - ▶ Faster and scalable



5. Heuristic performances

Dataset from *E. coli*, simulated with SimLoRD [3], and corrected with MECAT.

- ▶ Dataset: reads with a 10kb mean length, a 15% error rate, and a coverage of 100x.

Strategy	MSA	seed-MSA
Recall	84.505%	84.587%
Precision	88.347%	88.278%
Correct bases rate	95.290%	95.250%
Time	107h	42m

Similar results using both strategies.

A substantial **gain in time** is achieved using the seed-MSA strategy.

6. Results: ELECTOR vs. LRCstats

Dataset from *E. coli*, simulated with SimLoRD, composed of reads with a 8kb mean length, a 18% error rate, and a coverage of 20x.

Method	Original		Nanocorr		daccord	
	<i>ELECTOR</i>	<i>LRCstats</i>	<i>ELECTOR</i>	<i>LRCstats</i>	<i>ELECTOR</i>	<i>LRCstats</i>
Error rate	15.837	17.9267	0.339	0.3983	0.422	0.4498
Recall	N/A	-	0.98503	-	0.98836	-
Precision	N/A	-	0.99424	-	0.98468	-
Deletions	847,315	3,635,647	46,596	56,708	58,110	72,547
Insertions	10,393,229	13,038,057	237,798	279,970	306,930	336,686
Substitutions	5,611,023	671,040	143,605	45,783	72,265	25,643
Trimmed / split reads	N/A	-	1,612	-	123	-
Mean missing size	N/A	-	341	-	3,026	-
%GC	50.7	-	50.8	-	50.8	-
Time	13min	3h53	13min	3h52	13min	3h50

Results of these experiments show that the metrics computed by ELECTOR are comparable to LRCstats outputs, but also highlight several novelties.

LRCstats, besides having low performance results, also fails to evaluate correction's detailed impact on big datasets and on very long reads.

7. Conclusion

- ▶ Novel and open-source method for fast long read correction assessment
- ▶ Compatible with hybrid and self correctors
- ▶ Numerous metrics for correction quality (recall/precision)
- ▶ Downstream analyses assessment (mapping/assembly)
- ▶ Time-saving, scaling computation

[1] Sean La, Ehsan Haghshenas, and Cedric Chauve. LRCstats, a tool for evaluating long reads correction methods. *Bioinformatics*, 33(22):3652–3654, 2017.

[2] Christopher Lee, Catherine Grasso, and Mark F Sharlow. Multiple sequence alignment using partial order graphs. *Bioinformatics*, 18(3):452–464, 2002.

[3] Bianca K Stöcker, Johannes Köster, and Sven Rahmann. SimLoRD: Simulation of Long Read Data. *Bioinformatics*, 32(17):2704–2706, 2016.

