ELECTOR: EvaLuation of Error Correction Tools for lOng Reads

Lolita Lecompte[1], Camille Marchet[1], Pierre Morisse[2], Antoine Limasset[3],
Pierre Peterlongo[1], Arnaud Lefebvre[2], Thierry Lecroq[2]
[1]Univ Rennes, Inria, CNRS, IRISA, F-35000 Rennes, France [2]Normandie Univ, UNIROUEN, LITIS, 76000 Rouen, France [3]Université Libre de Bruxelles

1. Introduction

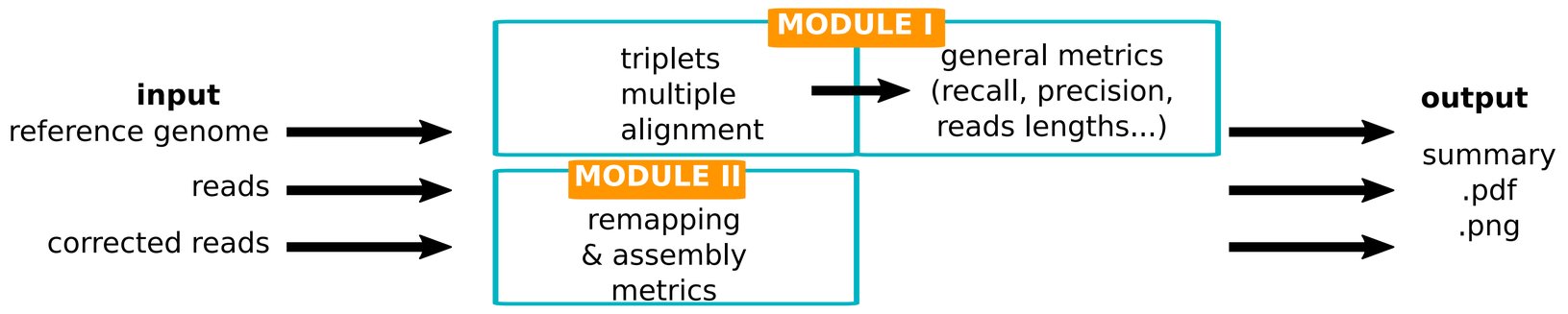
- Long read technologies, Pacific Biosciences and Oxford Nanopore, have **high error rates** (from 9% to 30%)
- Multiple error correction methods exist
- Importance of the correction stage on downstream processes
- Only one tool: **LRCstats** [1]
  - shows global correction gain
  - does not give access 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

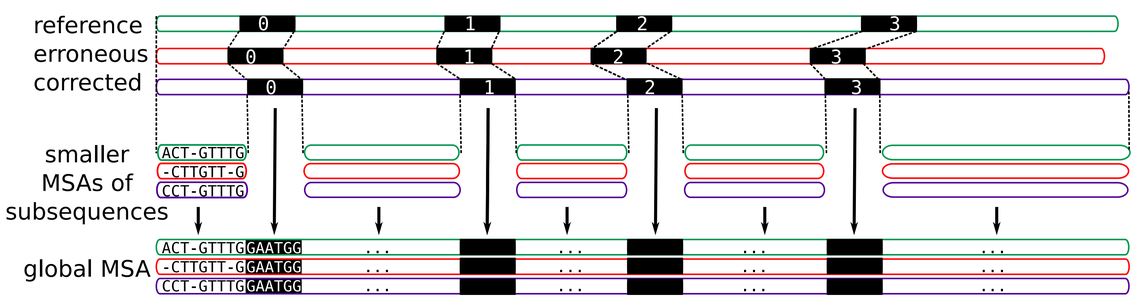
- Nb of contigs
- Nb of breakpoints
- N50
- N75

If reference, remapping
using BWA

- Average identity
- Genome coverage
- NGA50
- NGA75

4. Methods

1. Multiple alignment of triplets: $\{reference\ read,\ uncorrected\ read,\ corrected\ read\}$
2. **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

Simulated dataset from *E. coli* genome with SimLoRD [3] 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

Simulated dataset from *E. coli* with SimLoRD [3], composed of reads with a 8kb mean length, a 18% error rate, and a coverage of 20X.

Method	Original		Nanocorr		daccord	
	ELECTOR	*LRCstats*	*ELECTOR*	*LRCstats*	*ELECTOR*	*LRCstats*
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	N/A	1,612	N/A	123	N/A
Mean missing size	N/A		341		3,026	
%GC	50.7		50.8		50.8	
Runtime	**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 **contribution** (todo: pas fan) on big datasets and on very long reads datasets.

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 processings 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.

github.com/kamimrcht/ELECTOR

Contact: camille.marchet@irisa.fr