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:

HAL Id: hal-01929900
https://hal.archives-ouvertes.fr/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.
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

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

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.

<table>
<thead>
<tr>
<th>Strategy</th>
<th>MSA seed-MSA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
<td>84.505%</td>
</tr>
<tr>
<td>Precision</td>
<td>88.347%</td>
</tr>
<tr>
<td>Correct bases rate</td>
<td>95.290%</td>
</tr>
<tr>
<td>Time</td>
<td>107h</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Original</th>
<th>LRCstats</th>
<th>ELECTOR</th>
<th>Accord</th>
</tr>
</thead>
<tbody>
<tr>
<td>Error rate</td>
<td>15.837</td>
<td>17.9267</td>
<td>0.339</td>
<td>0.3983</td>
</tr>
<tr>
<td>Precision</td>
<td>N/A</td>
<td>0.962303</td>
<td>-</td>
<td>0.98836</td>
</tr>
<tr>
<td>Deletions</td>
<td>847.315</td>
<td>3,635.647</td>
<td>46.596</td>
<td>56.708</td>
</tr>
<tr>
<td>Insertions</td>
<td>13,038.077</td>
<td>2,177,798</td>
<td>279,979</td>
<td>396,930</td>
</tr>
<tr>
<td>Substitutions</td>
<td>5,611.023</td>
<td>671,040</td>
<td>143,605</td>
<td>45,783</td>
</tr>
<tr>
<td>Trimmed / split reads</td>
<td>N/A</td>
<td>1.612</td>
<td>-</td>
<td>123</td>
</tr>
<tr>
<td>Mean missing size</td>
<td>N/A</td>
<td>341</td>
<td>-</td>
<td>3,026</td>
</tr>
<tr>
<td>%GC</td>
<td>56.7</td>
<td>-</td>
<td>56.8</td>
<td>-</td>
</tr>
<tr>
<td>Time</td>
<td>13min</td>
<td>3h50</td>
<td>13min</td>
<td>3h52</td>
</tr>
</tbody>
</table>

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.