**Motivations**

Next generation sequencing technologies generate a high amount of short DNA sequences, but may contain imperfections. Some applications, such as assembly, yield better results with high quality data, triggering the need for short-read correction software. Most methods use large amount of memory and therefore limited to small datasets.

**Methods**

1. **Kmer counting**
   - ATCATCCC 25
   - AGTAAAAA 12
   - as in DSK [1]

2. **Insertion in Bloom filter**

3. **Error Correction**
   - Error detection and correction
     - k weak k-mers

**Error correction – common methods**

- Exploits high depth of coverage
- Multiple sequence alignment
- Suffix tree
- Kmer frequency spectrum

**Results**

**Simulated data:**
- Simulated read sets on C.elegans
  - 1% error rate, 40x, 100 bp, 40 M reads

**Comparison with Musket [2], Racer [4], SOAPec [5], Bless [6]**

**Real big data:**
- Human NA12878, 2.8 G reads ~ 100x
- 440 GB fasta file
- ~ 2 billion errors corrected
- Requiring only 3.8 GB of memory, executed in ~ 30 h

**Easy integration in gATB assembly pipeline**

**Minis assembly on the real C. elegans dataset 59D26594 (70x) with and without error correction.**

**Software**

<table>
<thead>
<tr>
<th>Software</th>
<th>Recall (%)</th>
<th>Precision (%)</th>
<th>Memory (GB)</th>
<th>Time (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blooco</td>
<td>95.8</td>
<td>99.5</td>
<td>0.8</td>
<td>681</td>
</tr>
<tr>
<td>Musket 1.1</td>
<td>97.9</td>
<td>99.6</td>
<td>7.8</td>
<td>3280</td>
</tr>
<tr>
<td>Racer 1.0</td>
<td>97.2</td>
<td>78.5</td>
<td>31.2</td>
<td>2041</td>
</tr>
<tr>
<td>SOAPec 2.02</td>
<td>76</td>
<td>99.7</td>
<td>12.7</td>
<td>17333</td>
</tr>
<tr>
<td>Bless 0.17</td>
<td>97.2</td>
<td>98.8</td>
<td>0.2</td>
<td>7620</td>
</tr>
</tbody>
</table>

**Download and use Blooco**

http://gatb.inria.fr/software/blooco/

**C++, Open Source, released under A-GPL license**

**References:**