Bloocoo, a memory efficient read corrector

Gaëtan Benoit, Dominique Lavenier, Claire Lemaitre, Guillaume Rizk

Bloocoo: a memory efficient read corrector

Gaëtan Benoit[1], Dominique Lavenier[1], Claire Lemaitre[1] and Guillaume Rizk[1]

[1] Inria/IRISA GenScale, Campus de Beaulieu, 35042 Rennes cedex, France.

guillaume.rizk@inria.fr, claire.lemaitre@inria.fr

Motivations

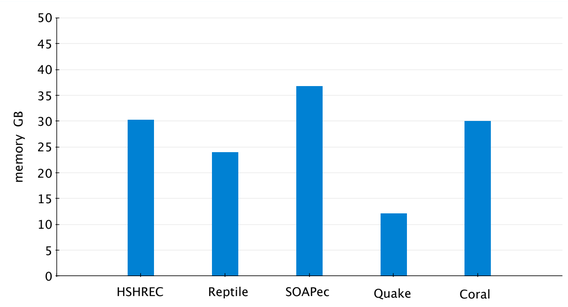
Next generation sequencing technologies generate a high amount of short DNA sequences, but may contain imperfections.

Some applications, such as assembly, yield better results with high quality data, triggering the need for short-read correction software.

Error correction – common methods

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

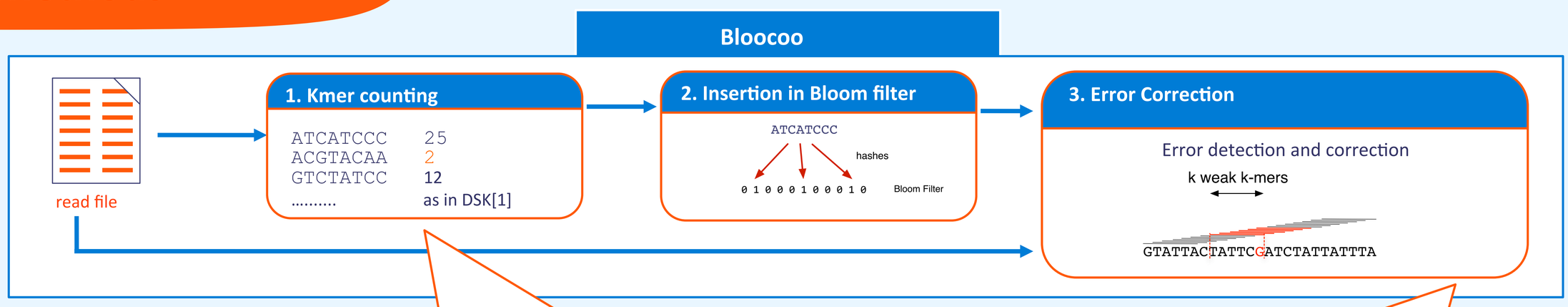
Most methods use large amount of memory
And therefore limited to small datasets



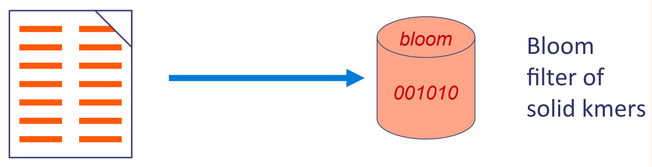
Memory used on a 56x *D.Melanogaster* dataset (98M reads) on state-of-the-art methods . From *A survey of error-correction methods, 2013* [3]

No tool able to handle large whole genome re- sequencing data files

Methods

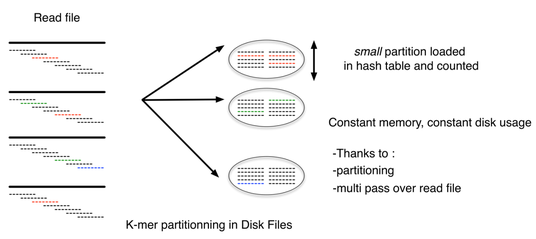


Kmer spectrum creation



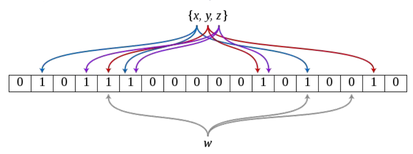
Kmer counting and insertion of solid kmers in Bloom filter : GATB library

DSK: *disk streaming of kmers* [1]

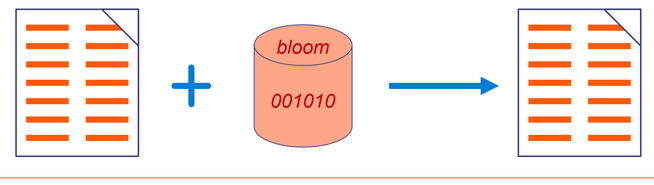


Bloom filter

- Keep solid kmers (with abundance above threshold)
- Insert them in a bloom filter
- Fast cache-coherent implementation
- 11 bits /kmer → 0.5 % false postives

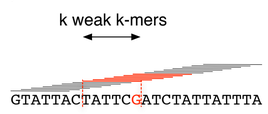


Correction



Identify and correct errors by querying the bloom filter for solid kmers

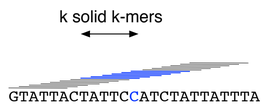
1) Error detection



2) Correction

Query BF for TATTC A/T/C

3) Validation



Complex cases

- multistage correction
- vote algorithm
- indel correction (ion torrent homopolymers)

Results

Simulated data:

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

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

Software	Recall (%)	Precision (%)	Memory (GB)	Time (sec)
Bloocoo 1.0	95.8	99.5	**0.8**	**681**
Musket 1.1	97.9	99.6	7.8	3280
Racer 1.01	97.2	78.5	31.2	2041
SOAPec 2.02	76	99.7	12.7	17203
Bless 0.17	97.2	98.8	**0.2**	7620

Real «big» data :

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

Easy integration in GATB assembly pipeline

Minia assembly on the real *C. elegans* dataset SRX026594 (70x) with and without error-correction.

Dataset	N50	Misassemblies
Original	4783	36
Bloocoo corrected	7634	26
Musket corrected	7749	29

Bloocoo:

- ★ Able to correct large datasets
- ★ Easy integration in the GATB assembly pipeline
- ★ Time/memory scalable on big datasets
- ★ Works also on indel errors (ion torrent)

References:

[1] Rizk, G., Lavenier, D., & Chikhi, R. (2013). DSK: k-mer counting with very low memory usage. Bioinformatics, btt020.

[2] Liu, Yongchao, Jan Schröder, and Bertil Schmidt. "Musket: a multistage k-mer spectrum-based error corrector for Illumina sequence data." Bioinformatics 29.3 (2013): 308-315.

[3] Yang, X., Chockalingam, S. P., & Aluru, S. (2013). A survey of error-correction methods for next-generation sequencing. Briefings in bioinformatics, 14(1), 56-66.

[4] Ilie, L., & Molnar, M. (2013). RACER: Rapid and accurate correction of errors in reads. Bioinformatics, btt407.

[5] Li R, Zhu H, Ruan J et al. De novo assembly of human genomes with massively parallel short read sequencing, Genome research 2010;20:265-272.

[6] Heo, Y., Wu, X. L., Chen, D., Ma, J., & Hwu, W. M. (2014). BLESS: Bloom filter-based error correction solution for high-throughput sequencing reads. Bioinformatics, btu030.

Download and use Bloocoo

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

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

Publication: E. Drezen, G. Rizk, R. Chikhi, C. Deltel, C. Lemaitre, P. Peterlongo, D. Lavenier
GATB: Genome Assembly & Analysis Tool Box
Bioinformatics, 2014

Download this poster:
http://tiny.cc/bloocooposter

PROJET FINANCÉ PAR L'ANR – PROJECT FUNDED BY THE ANR