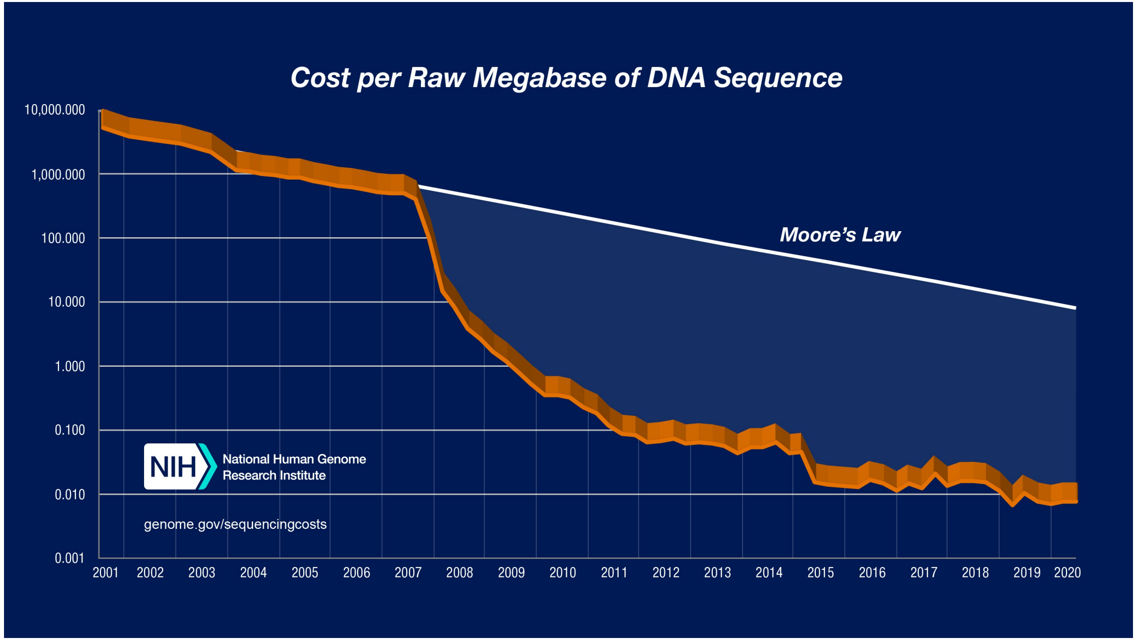
ganon: scalable and efficient DNA classification against large sets of reference sequences

Vitor C. Piro

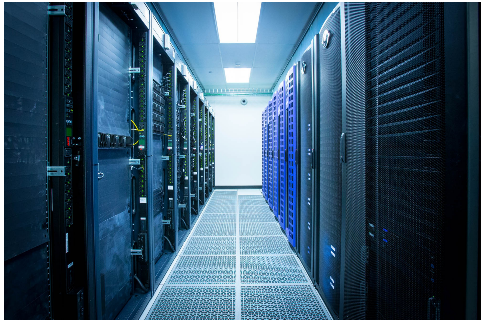
Research Scientist - Data Analytics & Computational Statistics

Hasso Plattner Institute, Germany

DNA Sequencing and data growth



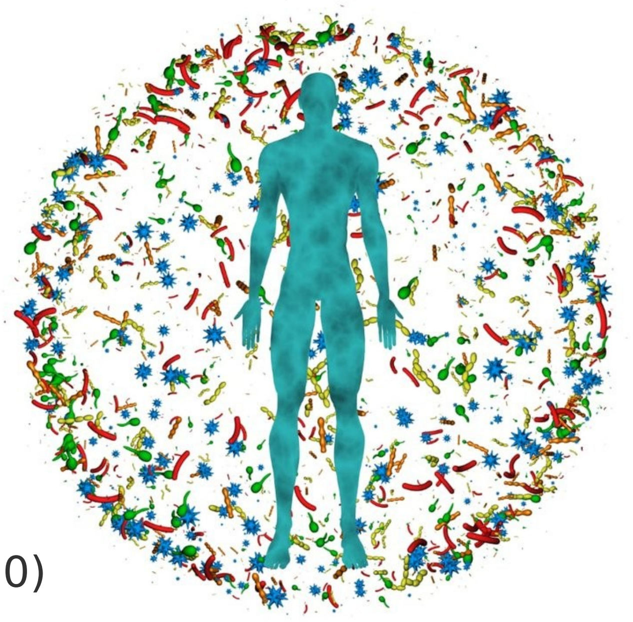
- Decreasing sequencing costs
- More genomic available
- More resources needed
- Higher energy consumption



DNA based studies

Microbiome study example:

- Hundreads to thousands of samples:
  - Subjects, time, treatment and control, replicates
- Discover contents of each sample
  - Compare with **reference genome sequences**
  - The more the better, if possible everything
  - Every day collection grows
- Example – gut microbiome:
  - 188 samples: *1.299.921.211 bp*
  - **References**: *723.003.822.007 bp* (as of Dec. 2020)
    - Doubles every 18 months



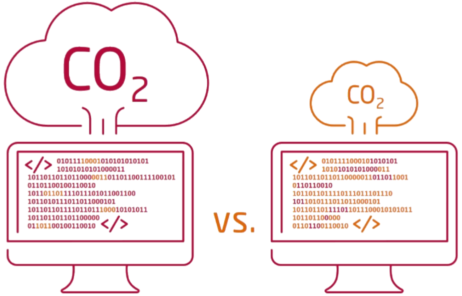
Efficient algorithms and data structures

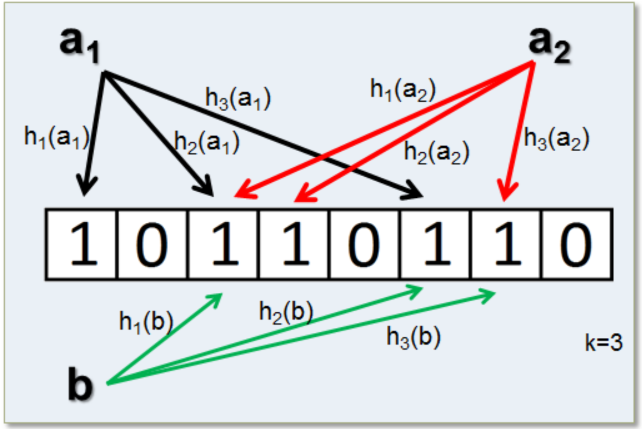
Indexing

- Organize and compress
- Reference sequences have to be indexed for quick search
- DNA search has to be flexible
  - Sequencing errors, mutations…

Solutions

- BWT, FM-Index, k-mer, Bloom Filters, …
  - Specialized to solve specific problems
- Bloom Filter
  - Probabilistic data structure / trade-offs
  - Set membership





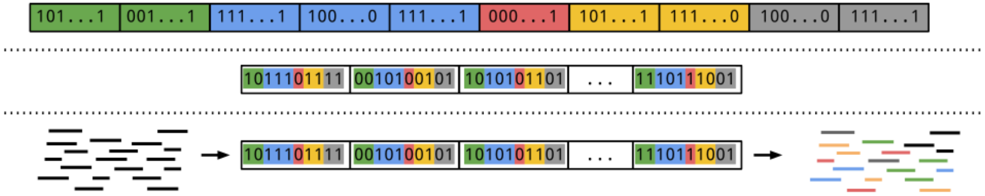
1	0	1	1	0	1	1	0

https://redislabs.com/blog/rebloom-bloom-filter-datatype-redis/

Microbiome profiling and the Interleaved Bloom Filter

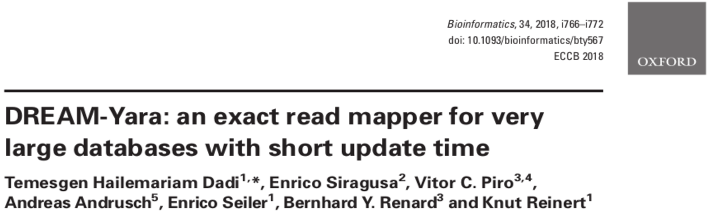
Microbiome Profiling

- Hundreads/Thousands of species
- What is in the sample? How much?



Interleaved Bloom Filter

- Fast indexing (as in the Bloom Filter)
- Several Bloom Filters, interleaved
- Fast searching



Bioinformatics, 34, 2018, i766–i772
doi: 10.1093/bioinformatics/bty567
ECCB 2018

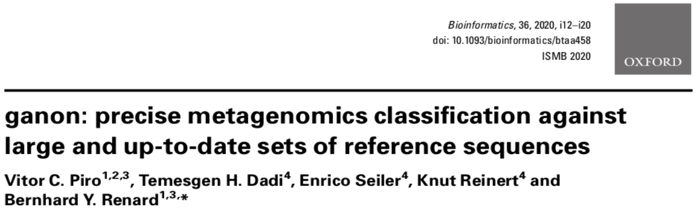
OXFORD

DREAM-Yara: an exact read mapper for very large databases with short update time

Temesgen Hailemariam Dadi[1,*], Enrico Siragusa[2], Vitor C. Piro[3,4], Andreas Andrusch[5], Enrico Seiler[1], Bernhard Y. Renard[3] and Knut Reinert[1]

ganon

- Enables fast indexing, updating as searching in large reference sequence set



Bioinformatics, 36, 2020, i12–i20
doi: 10.1093/bioinformatics/btaa458
ISMB 2020

OXFORD

ganon: precise metagenomics classification against large and up-to-date sets of reference sequences

Vitor C. Piro[1,2,3], Temesgen H. Dadi[4], Enrico Seiler[4], Knut Reinert[4] and Bernhard Y. Renard[1,3,*]

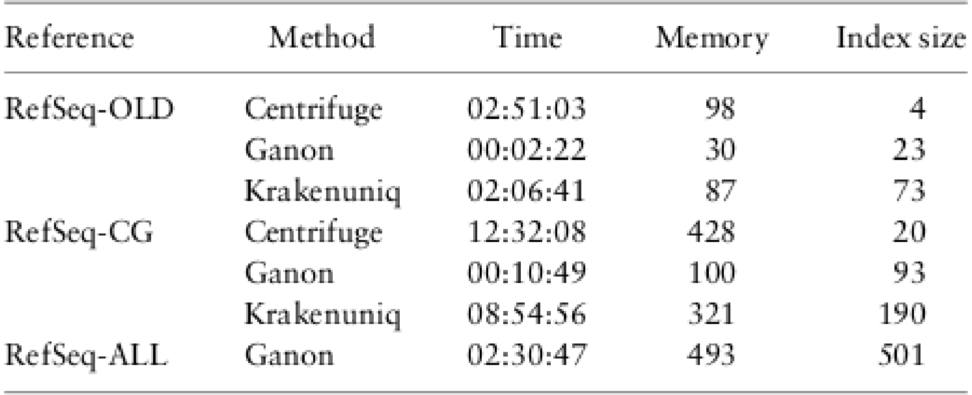
ganon and efficiency

Efficiency

- Interleaved Bloom Filter
- C++, SeqAn and multithreading

Further

- Updatability of indices
  - Incrementally grow
  - Remove sequences
  - Keep-up with new data



Reference	Method	Time	Memory	Index size
RefSeq-OLD	Centrifuge	02:51:03	98	4
	Ganon	00:02:22	30	23
	Krakenuniq	02:06:41	87	73
RefSeq-CG	Centrifuge	12:32:08	428	20
	Ganon	00:10:49	100	93
	Krakenuniq	08:54:56	321	190
RefSeq-ALL	Ganon	02:30:47	493	501

- Up to 75 faster on indexing
- Allow usage of more data
- Only methods allowing updates – no re-indexing

Conclusion

DNA Sequencing

- Fast data growth

Microbiome analysis

- Huge data comparisons
- More data = better results

Efficient solutions

- Efficient indexing, quick search
- Ganon and the interleaved bloom filter
  - Ultra-fast indexing
  - Update avoids redundant re-indexing

= less computations, better results

HPI Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Thank you!

Vitor C. Piro

Research Scientist - Data Analytics & Computational Statistics

Hasso Plattner Institute, Germany