GATB
in a nutshell

Erwan DREZEN / GENSCALE Team / INRIA
erwan.drezen@inria.fr
NGS technologies produce terabytes of data

NGS data analysis bottlenecks

1. **Time**: huge computing power needed
2. **Space**: many storage units needed (memory, hard disk, ...)

> read 1
ACGACGACGTAGACGACTAGCAAAACTA
CGATCGACTAT

> read 2
ACTACTACGATCGATGGTCGCGCTGCTC
GCTCTCTCGCT

...> read 10.000.000
TCTCCTAGCGCGGCGTATACGCTCGCTA
GCTACGTAGCT

...
GATB provides efficient and fast NGS algorithms to analyze such data

- Sophisticated data structures (optimized de Bruijn graph)
- Advanced hardware implementation (multi-threading, SIMD)

GATB is a software solution available as:

- An open source C++ library (a.k.a. GATB-CORE)
- A set of NGS tools based on the library (a.k.a. GATB-TOOLS)

GATB main advantages

- Allow complex genomes to be analyzed on a mere desktop
- Fasten the development of new NGS tools
GATB core structure for NGS data
- De Bruijn graph with very low memory footprint

Basis for numerous tools
- Data error correction
- Assembly
- Biological motif detection

GATB workflow

Ex: a whole human genome sequencing reads handled on 4 GBytes of memory
Some tools based on GATB

➢ **Minia**

Short read assembler based on a de Bruijn graph. The output of Minia is a set of contigs. Minia produces results of similar contiguity and accuracy to other de Bruijn assemblers (e.g. Velvet)

➢ **DiscoSNP**

Discover Single Nucleotide Polymorphism (SNP) from non-assembled reads

➢ **TakeABreak**

Detects inversion breakpoints without a reference genome by looking for fixed size topological patterns in the de Bruijn graph

➢ **Bloocoo**

K-mer spectrum based read error corrector, designed to correct large datasets with a very low memory footprint.
Publications

- G. Rizk, D. Lavenier, R. Chikhi, **DSK: k-mer counting with very low memory usage**, Bioinformatics, 2013 Mar 1;29(5):652-3

- R. Chikhi, G. Rizk. **Space-efficient and exact de Bruijn graph representation based on a Bloom filter**, Algorithms for Molecular Biology 2013, 8:22

- G. Collet, G. Rizk, R. Chikhi, D. Lavenier, **Minia on Raspberry Pi, assembling a 100 Mbp genome on a Credit Card Sized Computer**, Poster at the JOBIM conference, 2013 Jul 1-4 (Toulouse) Best poster award.

- K.I Salikhov, G. Sacomoto, G. Kucherov, **Using Cascading Bloom Filters to Improve the Memory Usage for de Brujin Graphs**, Algorithms in Bioinformatics, Lecture Notes in Computer Science, Volume 8126, 2013, pp 364-376

Partners

Web Site

http://gatb.inria.fr