1. What is GATB?

**Motivation**
NGS technologies produce terabytes of data. Efficient and fast NGS algorithms are essential to analyze them.

**Objective**
The Genome Assembly Tool Box (GATB)
- is an open-source software
- provides an easy way to develop efficient and fast NGS tools
- is based on data structure with a very low memory footprint
- allows complex genomes to be processed on desktop computers

2. Software Solution

The GATB philosophy proposes a 3-layer construction to analyze NGS datasets

1. **GATB-CORE**: a C++ library holding all the services needed for developing software dedicated to NGS data.
2. **GATB-TOOLS**: a set of elementary NGS tools mainly built upon the GATB library (k-mer counter, contig, scaffold, variant detection, etc.).
3. **GATB-PIPELINE**: a set of NGS pipeline that links together tools from the previous layer.

3. Compact de Bruijn graph data structure

The core data structure of GATB is a de Bruijn graph that encodes the main information from the sequencing reads.

**Strength of GATB**
GATB makes this graph compact by using a Bloom filter (a space efficient probabilistic data structure) and by using a CFP additional structure that avoids false positive answers from the Bloom filter due to its probabilistic nature.

4. Workflow

Here is a typical workflow when working with GATB

GATB-CORE transforms the reads into a de Bruijn graph, saves it in a HDF5 file that can be opened by other tools developed with the GATB-CORE API.

5. GATB helps you as a NGS user

GATB’s de Bruijn graph: a basis for families of tools

- **Data error correction**
- **Assembly**
- **Biological motif detection**

Several tools based on GATB are already available

- **Bloocoo**: k-mer spectrum based read error corrector for large datasets
- **Minia**: Short read assembler based on a de Bruijn graph. Results are 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

6. GATB helps you as a NGS developer

The GATB C++ library gives you the opportunity to quickly develop new NGS tools that fit your needs.

**Major facts about the GATB C++ library**

- Object Oriented Design
- Simple and powerful graph API
- Simple and powerful multithreading model
- HDF5 usage for data storage
- Fully documented with numerous code samples
- Complete test suite

**Publications**

**License & Web Site**
GATB is released under the GNU Affero General Public License.

Proprietary licencing for software editors or services providers is currently being studied.

For more details on GATB:
http://gatb.inria.fr

**Partners**

INRIA/IRISA/GenScale, Campus de Beaulieu, 35042 Rennes cedex

R. Chikhi, P. Peterlongo, Pierre Peterlongo, and Dominique Lavenier:
De Bruijn Graphs and Beyond. Lecture Notes in Computer Science, Volume 8126, 2013, pp 364-376