Colib'read | Visualization | BioMAJ2Galaxy | Bioimaging | Mechanics

Colib'read on Galaxy: A tools suite dedicated to biological information extraction from raw reads

Y. Le Bras[1], O. Collin[1], C. Monjeaud[1], V. Lacroix[2], E. Rivals[3], C. Lemaitre[4], V. Miele[2], G. Sacomoto[2], C. Marchet[2], B. Cazaux[3], A. Makrini[3], L. Salmela[7], S. Alves-Carvalho[4], A. Andrieux[4], R. Uricaru[5,6], P. Peterlongo[4]

1 GenOuest core facility, UMR6074 IRISA CNRS/INRIA/Université de Rennes1, Rennes, France
2 BAMBOO team, INRIA Grenoble Rhône-Alpes \& Laboratoire Biométrie et Biologie Évolutive, UMR5558 CNRS, Villeurbanne, France
3 MAB team, UMR5506 CNRS, Université Montpellier 2 LIRMM, Montpellier, France
4 INRIA/IRISA, Genscale team, UMR6074 IRISA CNRS/INRIA/Université de Rennes1, Rennes, France
5 University of Bordeaux, LaBRI/CNRS, Talence, France
6 University of Bordeaux, CBiB, Bordeaux, France
7 Department of Computer Science and Helsinki Institute for Information Technology HIIT, University of Helsinki, Finland

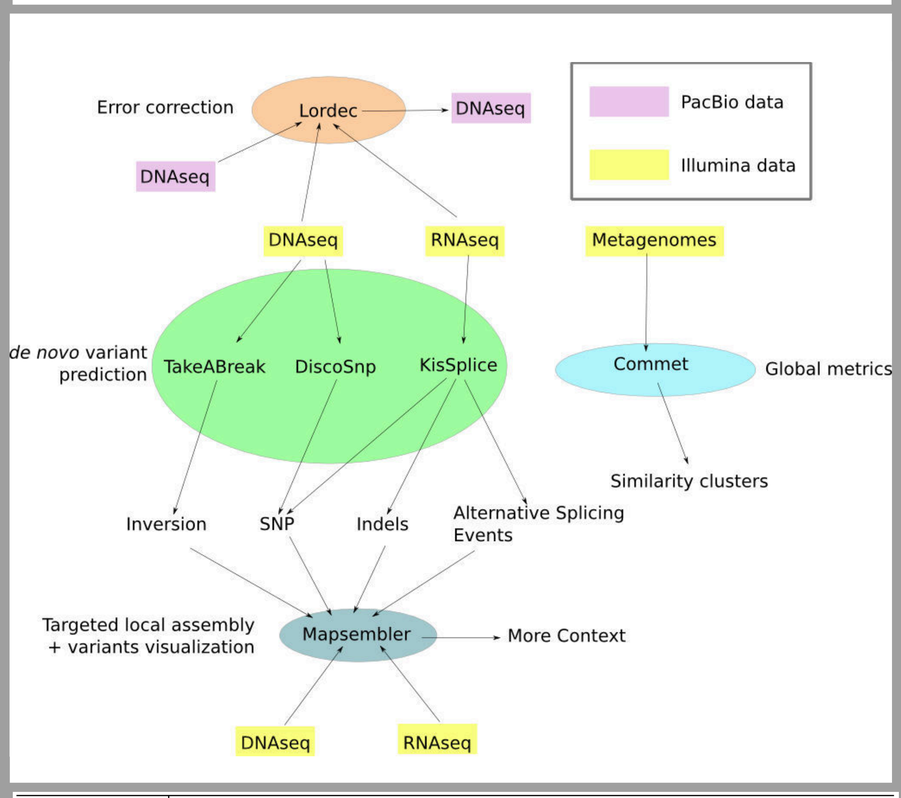
The raw reads context

With **NGS technologies**, life sciences face a raw data deluge. **Classical analysis** processes of such data often begin with an assembly step, needing l**arge amounts of computing resources**, and potentially **removing or modifying parts of the biological information** contained in the data. Our approach proposes to directly **focus on biological questions**, by considering **raw unassembled NGS data**, through a suite of six command-line tools.

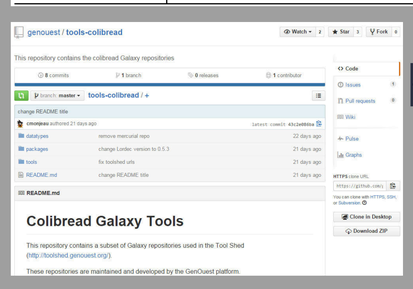
Dedicated to **"assembly-free" treatments**, the **Colib'read tools suite** uses optimized algorithms for various analyses of NGS datasets, such as **variant calling** or **read set comparisons**. Based on the use of ***de Bruijn* graphs** and **bloom filters**, such analyses can be performed in **few hours**, using s**mall amounts of memory**. Applications on real data demonstrate the **high accuracy** of these tools compared to classical approaches. To facilitate data analysis and tools dissemination, we developed **Galaxy tools** and tool shed repositories.

Tools suite overview

Overview of the **six tools** from the Colib'read project integrated to Galaxy

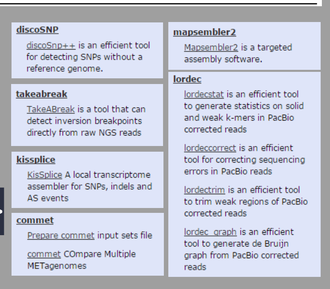


Tool	In	Out
KisSplice	One or more RNA-seq read set(s)	SNPs, small indels, alternative splicing events.
discoSnp	One or more raw genomic read set(s)	SNP sequences with their coverages
TakeABreak	One or more raw genomic read set(s)	Inversion breakpoints
Mapsembler2	A priori about pieces of known sequences, and associated raw read sets.	Validation and visualisation of genome structure near a loci of interest.
Commet	Several raw metagenomic complex read sets	Global comparison of input sets at the read level
LoRDEC	Illumina and PacBio read sets	Corrected PacBio read set



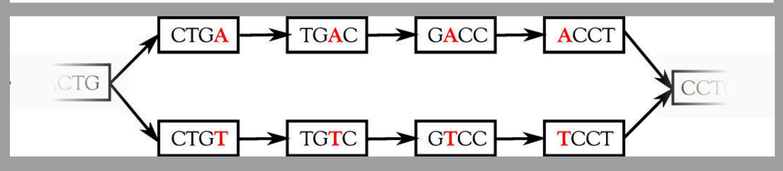
<--Github

Tools-->

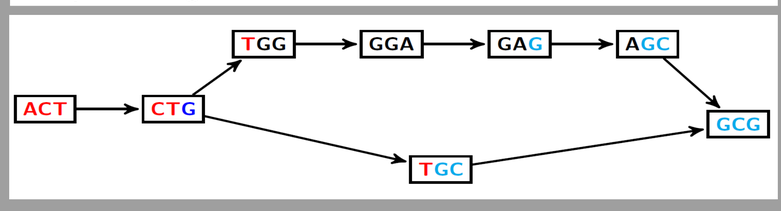


A common kernel: the de Bruijn graph

Toy example of a bubble in the de Bruijn graph (dBG) (k = 4). The bubble is generated by a SNP present in two polymorphic sequences CTGACCT and CTGTCCT



dBG with k = 3 for the sequences: ACTGGAGCG and ACTGCG. The pattern in the sequence generates a bubble, from CT to GCG.



DiscoSNP

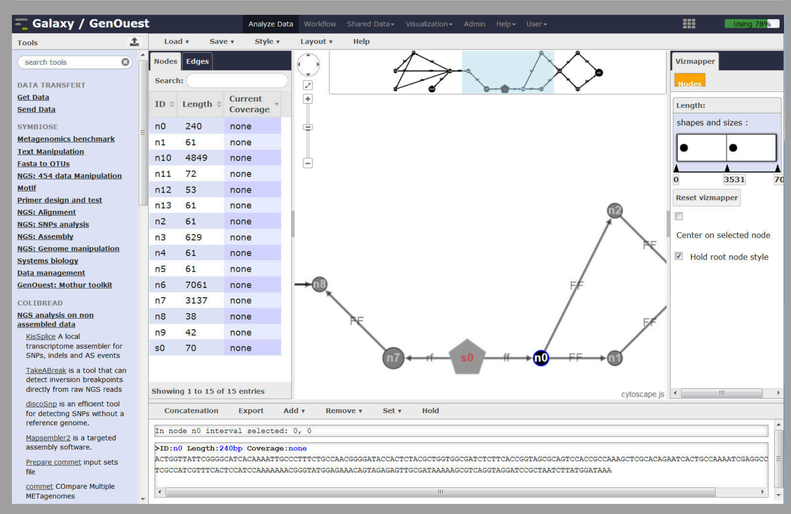
A reference-free **SNP calling** software that focuses on the detection of both heterozygous and homozygous isolated SNPs, from any number of sequencing datasets.

TakeABreak

A method to **detect inversion variants** from sets of reads without any reference genome. The rationale is similar to the one of DiscoSNP: inversion variants generate particular topological motifs in the dBG.

Mapsembler

A **targeted assembly software**. It takes as input one or more set(s) of NGS raw reads (fasta or fastq, gzipped or not) and a set of input sequences, called the starters. Below, a screenshot of GSV, the associated viewer.



KisSplice

A software that enables to **analyze RNA-seq** data with or without a reference genome or transcriptome. It is an exact local transcriptome assembler that allows to identify SNPs, indels and alternative splicing (AS) events

Commet

COmpare Multiple METagenomes: Have a global similarity overview between all datasets of a large metagenomic project.

LoRDEC

A tool to **correct sequencing errors** in long reads obtained from 3rd generation of high throughput sequencing technologies.

With the Colib'read Galaxy tools suite, we give the possibility to a broad range of life scientists to analyze **raw NGS data**. More importantly, our approach allows to **keep the maximum of biological information** from the data and uses a **very low memory footprint**.

The Colib'read project is funded by the ANR. Galaxy developments are supported by Biogenouest, the Western France life sciences facilities network and Brittany as Pays de la Loire regions.

A Genocloud dedicated Galaxy server is reachable at http://colibread.genouest.org/ for testing.

Galaxy Tool Shed repositories are available on the main https://toolshed.g2.bx.psu.edu/ as GUGGO Tool Sheds. Dockerfiles are also available.

All Colib'read project related publications are available in https://colibread.inria.fr/publications/