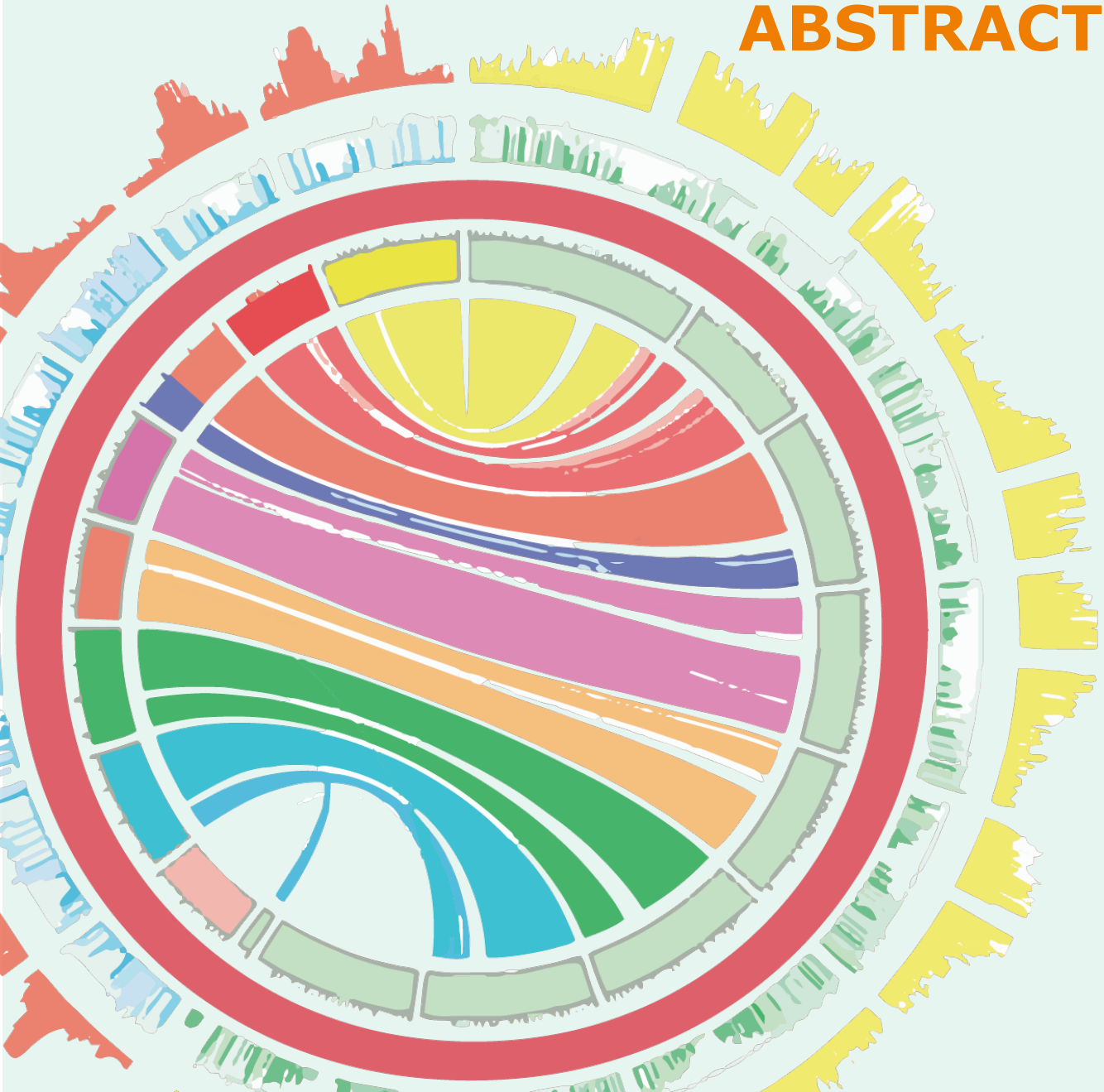
JOURNÉES OUVERTES DE BIOLOGIE INFORMATIQUE & MATHÉMATIQUES

JOBIM 2018

03>06 JUIL

PALAIS DU PHARO | MARSEILLE

ABSTRACTS



Gigwa - Genotype Investigator for Genome-Wide Analyses

Guilhem Sempéré* [1,2], Adrien Petel [1], Alexis Dereeper [2,3], Manuel Ruiz [2,4], Pierre Larmande †‡ [2,5]

[1] UMR Intertryp - CIRAD - IRD – Institut de recherche pour le développement [IRD], CIRAD – Avenue Agropolis - 34398 Montpellier Cedex 5, France
[2] Institut de Biologie Computationnelle (IBC) – Centre de Coopération Internationale en Recherche Agronomique pour le Développement, Institut National de la Recherche Agronomique, Institut National de Recherche en Informatique et en Automatique, Université de Montpellier, Centre National de la Recherche Scientifique – Building 5 - 860 rue de St Priest 34095 Montpellier, France
[3] IRD IPME (IPME) – Institut de recherche pour le développement [IRD] – Avenue Agropolis, 34398 Montpellier Cedex 5, France
[4] Centre de coopération internationale en recherche agronomique pour le développement (CIRAD) – CIRAD, Institut National de la Recherche Agronomique - INRA – Av Agropolis, Montpellier, France
[5] Institut de Recherche pour le Développement (IRD) – Institut de recherche pour le développement [IRD] : UMR232, Université de Montpellier : UMR232 – 911 avenue Agropolis,BP 6450134394 Montpellier cedex 5, France

With the advent of next-generation sequencing (NGS) technology, thousands of new genomes of both plant and animal organisms have become available. In this context, the Variant Call Format (VCF) [1] has become a convenient and standard file format for storing variants identified by NGS / NGG approaches. VCF files may contain information on tens of millions of variants, for thousands of individuals. Having to manage such significant volumes of data involves considerations of efficiency with regard to the following aspects: Filtering features, Storage performance, Sharing capabilities, Graphical visualization. However, existing tools are often limited to command line or programmatic APIs targeted at experienced users, but are not suitable for non-bioinformaticians.
The Gigwa application [2], which stands for "Genotype Investigator for Genome-Wide Analyses", aims at taking into account those aspects. It provides an easy and intuitive way to explore large amounts of genotyping data by filtering it not only on the basis of variant features, including functional annotations, but also on genotype patterns. It is a fairly lightweight, web-based, platform-independent solution that allows to feed a MongoDB [3] NoSQL database with VCF [4], PLINK or HapMap files containing up to billions of genotypes, and provides a user-friendly interface to filter data in real time. Gigwa provides the means to export filtered data into several popular formats and features connectivity with visualization software such as FlapJack [5] and online or standalone genome browsers (GBrowse, [REF]JBrowse [6] and IGV [7]). Additionnally, Gigwa-hosted datasets are interoperable via two standard REST APIs: GA4GH[8] and BrAPI [9]. Thus, we think that Gigwa could serve a large number of scientists by helping them to manage, filter and share their own data.

*Corresponding author: guilhem.sempere@cirad.fr
†Speaker
‡Corresponding author: pierre.larmande@ird.fr

1. 1000 Genome project Consortium. Variant Call Format (VCF) [Internet]. [cited 2018 Mar 20].

2. Sempéré G, Philippe F, Dereeper A, Ruiz M, Sarah G, Larmande P. Gigwa-Genotype investigator for genome-wide analyses. Gigascience [Internet]. 2016 [cited 2016 Sep 24];5:25.

3. MongoDB Inc. MongoDB [Internet]. 2015 [cited 2015 Dec 19]. Available from: https://www.mongodb.org/

4. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics [Internet]. 2011 [cited 2014 Jul 10];27:2156–8.

5. Milne I, Shaw P, Stephen G, Bayer M, Cardle L, Thomas WTB, et al. Flapjack–graphical genotype visualization. Bioinformatics [Internet]. 2010 [cited 2016 Mar 3];26:3133–4.

6. Skinner ME, Uzilov A V, Stein LD, Mungall CJ, Holmes IH. JBrowse: a next-generation genome browser. Genome Res. [Internet]. 2009;19:1630–8.

7. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief. Bioinform. [Internet]. 2013;14:178–92.

8. The Global Alliance for Genomics and Health Consortium . GA4GH API [Internet]. 2017. Available from: https://github.com/ga4gh/ga4gh-schemas
9. Brapi consortium . The Breading API. 2017; Available from: https://brapi.org/

Keywords: Genomic variations, VCF, HapMap, NoSQL, MongoDB, SNP, INDEL