The South Green portal: a comprehensive resource for tropical and Mediterranean crop genomics

Valentin Guignon, Yann Hueber, Mathieu Rouard, Stéphanie Bocs, David Couvin, Frédéric F. de Lamotte, Gaëtan Droc, Jean-François Dufayard, Nordine El Hassouni, Cédric Farcy, et al.

The South Green portal: a comprehensive resource for tropical and Mediterranean crop genomics

South Green collaborators [a,b,c,d,e,f*†]

[a] Bioversity International, Parc Scientifique Agropolis II, 34397 Montpellier Cedex 5, France
[b] UMR AGAP CIRAD/INRA/SupAgro, Avenue Agropolis, 34398, Montpellier Cedex 5, France
[C] UMR BGPI CIRAD/INRA/SupAgro, Campus International de Baillarguet, 34398, Montpellier, Cedex 5, France
[d] UMR DIADE IRD/UM, Avenue Agropolis, 34934, Montpellier Cedex 5, France.
[e] UMR InterTryp CIRAD/IRD, Campus International de Baillarguet, 34398, Montpellier, Cedex 5, France
[f] UMR IPME IRD/UM/CIRAD, Avenue Agropolis, 34394, Montpellier, Cedex 5, France.

* Correspondences: m.rouard@cgiar.org, manuel.ruiz@cirad.fr, gautier.sarah@supagro.inra.fr, christine.tranchant@ird.fr

† The list of participants and their affiliations is provided at the end of this paper.

Authors' information

Bioversity International: Guignon Valentin; Hueber Yann; Rouard Mathieu
UMR AGAP (CIRAD/INRA/SupAgro): Bocs Stephanie; Couvin David; de Lamotte Frédéric; Droc Gaëtan; Dufayard Jean-François; El Hassouni Nordine; Farcy Cédric; Gkanogiannis Anestis; Hamelin Chantal; Larivière Delphine; Martin Guillaume; Ortega Enrique; Pitollat Bertrand; Pointet Stéphanie; Ruiz Manuel; Sarah Gautier; Summo Marilyne; This Dominique
UMR BGPI (CIRAD): Ravel Sébastien
UMR DIADE (IRD): Larmande Pierre; Monat Cécile; Sabot François; Tando Ndomassi; Tranchant-Dubreuil Christine
UMR InterTryp (CIRAD): Sempéré Guilhem
UMR IPME (IRD): Dereeper Alexis

Abstract

The South Green Web portal (http://www.southgreen.fr/) provides access to a large panel of public databases, analytical workflows and bioinformatics resources dedicated to the genomics of tropical and Mediterranean crops. The portal contains currently about 20 information systems and tools and targets a broad range of crops such as Banana, Cacao, Cassava, Coconut, Coffee, Grape, Rice and Sugarcane.

Keywords: Crop databases; Genomics; Galaxy; Next Generation Sequencing; Pipeline; Sequence variants; Workflow

1. Introduction

Analysis and visualization of massive genomics datasets are an ongoing trend in plant sciences, especially for tropical crops where this subject can help to tackle challenges linked to human activities and Climate Change: conservation and analysis of biodiversity (loss because of human activity), food security (the 9 billion people question), new usage of plant material, etc. In the recent years, because of successively lower costs of genomic sequencing, a large number of groups have developed massive resources and data which require high performance computational resources and new analytical approaches [1, 2].
The South Green portal is an ecosystem of tools that were originally developed as independent entities to fulfill the need for specific projects or crops, but have evolved over time to generic tools to comprehensively study crop genomics.. Those generic tools are adaptable to a wide range of other organisms, especially cultivated, wild or orphan plants.

2. Specialized resources

Some resources from the South Green portal are dedicated to specialized datasets. Among the Gene-based databases, GreenPhylDB [3] provides access to resources for comparative genomics and orthology identification in plant genomes, and OryGenesDB [4] is a database developed for rice reverse genetics, and containing FSTs (flanking sequence tags) of various mutagens and functional genomics data, collected from both international insertion collections and the literature. Other resources are – (i) Marker-based databases like TropGeneDB [5] which connects data on molecular markers (e.g. Simple Sequence Repeats, Diversity Arrays Technology), Quantitative Trait Loci, genetic and physical maps, phenotyping studies, and information on genetic resources (geographic origin, parentage, collection), (ii) SNiPlay [6] which allows querying both SNPs (Single-Nucleotide Polymorphism) and InDels derived from NGS (Next-Generation Sequencing) projects (Whole-Genome Sequencing, Genotyping by Sequencing, RNA-Seq) and computing a set of web analytical workflows for the resulting variants (diversity, population stratification, Genome-Wide Association Study), and proposes graphical representations of the results, and (iii) Gigwa [7] providing exploration of very large genotyping studies by filtering them based on not only variant features (SNP, Indels), including functional annotations, but also genotype patterns, in order to extract subsets in various popular export formats. Current developments are on the way for databases aimed to host/pathogen interactions, catalogs of plant pathogens strains and so on. Recently, based on our experience with Web Semantic technologies [8], we developed AgroLD (www.agrold.org) an RDF (Resource Description Framework) knowledge base that consists of integrated data from a variety of plant resources and ontologies.

A complete summary of the South Green databases with their description is available at http://www.southgreen.fr/databases.

3. Genome Hubs

Our participation in several reference genome sequencing projects [9-11] has led us to develop crop-specific information systems, so called Genome Hubs, to manage the corresponding genome annotations and linked datasets. Data available are under different forms, from complete genome sequence along with gene structure, gene product information, metabolic pathways, gene families, transcriptomic assays (Expressed Sequence Tags, RNA-Seq), genetic markers as well as genetic and physical maps. Several Genome Hubs were released: Banana [12], Cassava, Cacao, Coffee [13], and others (e.g. Rice, *Magnaporthae*) are currently under development. Genome Hubs are powered by major GMOD (Generic Model Organism Database) components (*i.e.* Chado, Cmap, Jbrowse, Tripal, Galaxy, Pathway tools) and complemented by resources and tools developed within the South Green framework such as GreenPhylDB, SNiPlay and TropGeneDB.

4. Workflow analyses

Target users of bioinformatic applications are usually divided between people who use command-line and those who do not. Our strategy has been to address both categories by offering complementary solutions to perform data analyses (**Figure 1**).

Galaxy workflows

Galaxy [14] is a web-based service that allows an easy access to the bioinformatic applications and strongly supports reproducibility of analysis steps. Through its graphical interface, non-bioinformatician scientists are able to conduct small scale as well as medium range analyses in a user-friendly manner. In addition to the generic tools provided with the standard installation of Galaxy, the South Green Galaxy instance (http://galaxy.southgreen.fr/galaxy/) contains a large collection of exclusive tools. The Galaxy Tool Shed allows these modules to be easily shared with any other Galaxy instance. In addition, pre-configured workflows were designed for recurrent analyses in plant genomics such as NGS mapping/cleaning, SNP calling and filtering, Genome-Wide Association Study, basic population genetics, structural variations and phylogenetics.

TOGGLE

TOGGLE [15] is a command-line tool designed to allow biologists to create large workflows without knowing how to code it (https://github.com/SouthGreenPlatform/). Using editable configuration files, it proposes a more in-depth control of customizable pipelines and allows the management of many more samples than Galaxy (ex: massive resequencing projects, more than 50 samples), as well as the management of tools versions. It can be used on wide array of analyses, such as NGS data cleaning, DNAseq, RNAseq, restriction site-associated DNA sequencing (RADseq)/Genotyping by Sequencing (GBS), genome and transcriptome assembly, SNP calling, structural variations detection, and so on. The current version can be deployed on

laptop, large machine as well as a HPC cluster, using normal system of Docker machines, and is scheduler-aware. Already customized pipelines are also available for classical NGS analyses.

Handling big data sets in Rice

Our goal was to setup a simple framework to help users to work efficiently and easily with the huge volume of genomic resources generated in the Rice data store based on the 3,000 rice genomes [16] and additional resources (i.e. High Density Rice Array 700k SNPs and IRIGIN project http://irigin.org). Galaxy modules were implemented to query efficiently millions of genomic variants, to filter on genomic annotations, to filter on list of individuals, to predict coding effects of genomic variants, and to export in several formats (see Rice Variant Analysis in the Tool Panel of http://galaxy.southgreen.fr/galaxy/). Outputs can be easily sent to any Galaxy workflows. The equivalent TOGGLE implementation is also on the way.

5. Other developments

We developed a large number of other codes, snippets, modules and plugins for database managements, duplication detection and representation, fasta file manipulation and so on. All those codes are available under GPLv3 license on the GitHub, at https://github.com/SouthGreenPlatform.

6. Examples of studies using the South Green platform

Various groups used the South Green infrastructure to obtain their data and results, and were able to publish high-quality biological information, on Coffee genome [9], Banana [10], Cocoa [11], African rice [17] or large transcriptome resources [18]. Tools developed for these studies are adaptable to a wide range of other organisms.

Conclusion

The South Green portal is developed by a local network of scientists gathering bioinformatics and genomics skills located on the Agropolis campus from Montpellier as a joint collaboration between CIRAD, IRD, INRA, SupAgro and Bioversity international. Based on this strong local community in the field of agriculture, food and biodiversity, various bioinformatics applications and resources dedicated to genomics of tropical and Mediterranean plants have been developed, published and provided to the international community. Source code can be also found on GitHub (https://github.com/SouthGreenPlatform).

Funding

This work was supported by Agropolis Fondation/AGRO Labex (ID ARCAD 0900-001, ID RTB Multi-Genomes Hub 1403-018), the French Institute of Bioinformatics (IFB), ANR (GNPAnnot, MusaTract, CoffeaSeq, AfriCrop #NR-13-BSV7-0017), CIRAD and IRD (computer infrastructures), the CGIAR Research Programs on Roots, Tubers and Bananas and GRiSP/RICE, IRIGIN/France Génomique (National infrastructure, funded as part of

“Investissement d’avenir” program managed by ANR # ANR-10-INBS-09) , and NUMEV Labex (LandPanToggle #2015-1-030-LARMANDE).

Acknowledgements

We thank the Scientific Committee of South Green for their helpful advice and suggestions, as well as our different institutions for their support.

References

1. Spjuth, O., Bongcam-Rudloff, E., Dahlberg, J., Dahlo, M., Kallio, A., Pireddu, L., Vezzi, F., and Korpelainen, E. (2016). Recommendations on e-infrastructures for next-generation sequencing. GigaScience *5*, 26.
2. Gullapalli, R.R., Desai, K.V., Santana-Santos, L., Kant, J.A., and Becich, M.J. (2012). Next generation sequencing in clinical medicine: Challenges and lessons for pathology and biomedical informatics. J Pathol Inform *3*, 40.
3. Rouard, M., Guignon, V., Aluome, C., Laporte, M.-A., Droc, G., Walde, C., Zmasek, C.M., Périn, C., and Conte, M.G. (2011). GreenPhylDB v2.0: comparative and functional genomics in plants. Nucleic Acids Res. *39*, D1095-1102.
4. Droc, G., Périn, C., Fromentin, S., and Larmande, P. (2009). OryGenesDB 2008 update: database interoperability for functional genomics of rice. Nucleic Acids Res. *37*, D992-995.
5. Hamelin, C., Sempere, G., Jouffe, V., and Ruiz, M. (2013). TropGeneDB, the multi-tropical crop information system updated and extended. Nucleic Acids Res. *41*, D1172-1175.
6. Dereeper, A., Homa, F., Andres, G., Sempere, G., Sarah, G., Hueber, Y., Dufayard, J.-F., and Ruiz, M. (2015). SNiPlay3: a web-based application for exploration and large scale analyses of genomic variations. Nucleic Acids Res. *43*, W295-300.
7. Sempéré, G., Philippe, F., Dereeper, A., Ruiz, M., Sarah, G., and Larmande, P. (2016). Gigwa-Genotype investigator for genome-wide analyses. GigaScience *5*, 25.
8. Wollbrett, J., Larmande, P., de Lamotte, F., and Ruiz, M. (2013). Clever generation of rich SPARQL queries from annotated relational schema: application to Semantic Web Service creation for biological databases. BMC Bioinformatics *14*, 126.
9. Denoeud, F., Carretero-Paulet, L., Dereeper, A., Droc, G., Guyot, R., Pietrella, M., Zheng, C., Alberti, A., Anthony, F., Aprea, G., et al. (2014). The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. Science *345*, 1181-1184.
10. D’Hont, A., Angélique, D.h., France, D., Jean-Marc, A., Franc-Christophe, B., Françoise, C., Olivier, G., Benjamin, N., Stéphanie, B., Gaëtan, D., et al. (2012). The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. Nature *488*, 213-217.
11. Argout, X., Salse, J., Aury, J.-M., Guiltinan, M.J., Droc, G., Gouzy, J., Allegre, M., Chaparro, C., Legavre, T., Maximova, S.N., et al. (2011). The genome of Theobroma cacao. Nat. Genet. *43*, 101-108.
12. Droc, G., Larivière, D., Guignon, V., Yahiaoui, N., This, D., Garsmeur, O., Dereeper, A., Hamelin, C., Argout, X., Dufayard, J.-F., et al. (2013). The banana genome hub. Database *2013*, bat035.
13. Dereeper, A., Bocs, S., Rouard, M., Guignon, V., Ravel, S., Tranchant-Dubreuil, C., Poncet, V., Garsmeur, O., Lashermes, P., and Droc, G. (2015). The coffee genome hub: a resource for coffee genomes. Nucleic Acids Res. *43*, D1028-1035.

14. Goecks, J., Nekrutenko, A., Taylor, J., and Galaxy, T. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. *11*, R86.
15. Monat, C., Tranchant-Dubreuil, C., Kougbeadjo, A., Farcy, C., Ortega-Abboud, E., Amanzougarene, S., Ravel, S., Agbessi, M., Orjuela-Bouniol, J., Summo, M., et al. (2015). TOGGLE: toolbox for generic NGS analyses. BMC Bioinformatics *16*, 374.
16. project, r.g. (2014). The 3,000 rice genomes project. GigaScience *3*, 7.
17. Nabholz, B., Sarah, G., Sabot, F., Ruiz, M., Adam, H., Nidelet, S., Ghesquiere, A., Santoni, S., David, J., and Glemin, S. (2014). Transcriptome population genomics reveals severe bottleneck and domestication cost in the African rice (Oryza glaberrima). Mol Ecol *23*, 2210-2227.
18. Sarah, G., Homa, F., Pointet, S., Contreras, S., Sabot, F., Nabholz, B., Santoni, S., Saune, L., Ardisson, M., Chantret, N., et al. (2016). A large set of 26 new reference transcriptomes dedicated to comparative population genomics in crops and wild relatives. Molecular ecology resources.

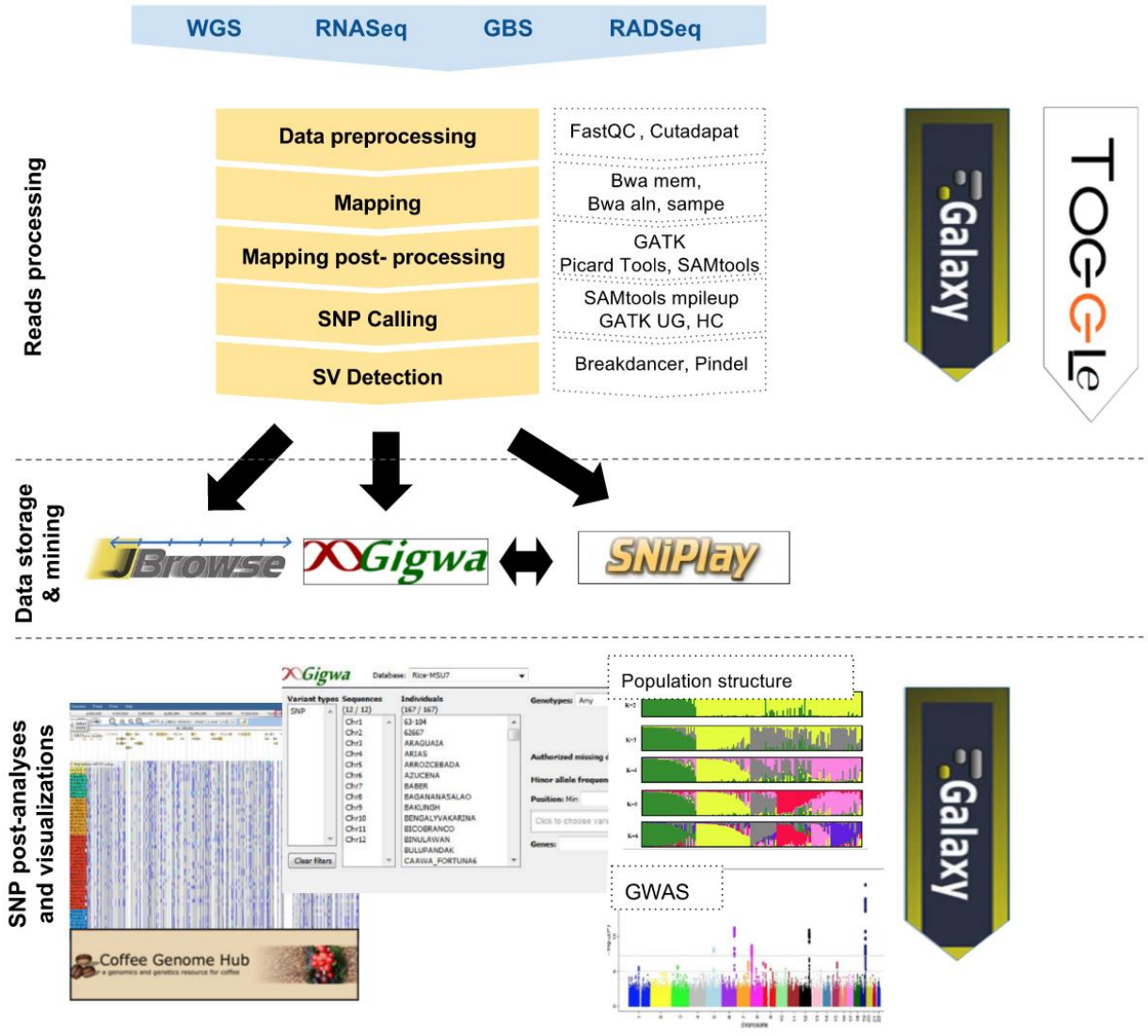

Figure 1. South Green resources and strategy to process, analyze and display Next-Generation Sequencing datasets in the context of genomic variation studies.