Current Plant Biology 7–8 (2016) 6–9

Contents lists available at ScienceDirect

Current Plant Biology

journal homepage: www.elsevier.com/locate/cpb

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

South Green collaborators[a,b,c,d,e,f,1]

[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

ARTICLE INFO

Article history:
Received 28 October 2016
Received in revised form 3 December 2016
Accepted 3 December 2016

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

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.

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

☆ This article is part of a special issue entitled "Genomic resources and databases", published in the journal Current Plant Biology 7–8, 2016.
[1] The list of participants and their affiliations are provided in Appendix A

http://dx.doi.org/10.1016/j.cpb.2016.12.002

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

(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 (Fig. 1).

4.1. 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, preconfigured workflows were designed for recurrent analyses in plant genomics such as NGS mapping/cleaning, SNP calling and filter-