Genetics and Molecular Biology, 43, 2, e20180272 (2020)
Copyright © 2020, Sociedade Brasileira de Genética.
DOI: https://doi.org/10.1590/1678-4685-GMB-2018-0272

Research Article
Evolutionary Genetics

Tools for Evolutionary and Genetic Analysis (TEGA): A new platform for the management of molecular and environmental data

Dario Ezequiel Elias[1] and Eva Carolina Rueda[1,2,3]

[1]Universidad Nacional de Entre Ríos, Facultad de Ingeniería, Cátedra de Genética, Oro Verde, Entre Ríos, Argentina.
[2]Universidad Nacional del Litoral, Facultad de Humanidades y Ciencias, Laboratorio de Genética, Ciudad Universitaria, Santa Fe, Argentina.
[3]Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Buenos Aires, Argentina.

Abstract

Population genetics studies the distributions and changes in population allele frequencies in response to processes, such as mutation, natural selection, gene flow, and genetic drift. Researchers daily manage genetic, biological, and environmental data of the samples, storing them in text files or spreadsheets, which makes it difficult to maintain consistency and traceability. Here we present TEGA, a WEB-based stand-alone software developed for the easy analysis and management of population genetics data. It was designed to: 1) facilitate data management, 2) provide a way to execute the analysis procedures, and 3) supply a means to publish data, procedures, and results. TEGA is distributed under the GNU AGPL v3 license. The documentation, source code, and screenshots are available at https://github.com/darioelias/TEGA. In addition, we present Rabid Fish, the first implementation of TEGA in the Genetics Labortory of the Faculty of Humanities and Sciences at the National University of the Litoral, where research focuses on population genetics studies applied to non-model organisms.

Keywords: Bioinformatics, biological data management, population genetics, molecular markers, Open Science.

Received: September 21, 2018; Accepted: June 25, 2019.

Introduction

As a part of evolutionary biology, population genetics, deals with the study of genetic differences within and between populations (Charlesworth and Charlesworth, 2017). Later, in 1987, using molecular markers (first mitochondrial DNA and later nuclear markers), John Avise introduced the concept of Phylogeography as a way of explaining how historical geological, climatic, and ecological conditions influenced the current distribution of species and genetic lineages Subsequently, advances in laboratory (especially in DNA sequencing technologies) and computational methods that make better use of data made phylogeographic inferences more accurate (Avise, 1998). Phylogeography has experimented significant growth in areas such as conservation, because it helps in defining evolutionary significant units, and in the study of prioritization for areas of high value for conservation (Moritz, 1992; Crandall *et al.*, 1999; Frankham, 2010).

Although mitochondrial DNA has been broadly used, microsatellite markers are still valuable tools in molecular ecological and phylogeographic studies (Jarne and Lagoda, 1996; Vignal *et al.*, 2002; Selkoe and Toonen, 2006). Recent developments in next-generation sequencing approaches have also revolutionized the development of molecular markers, allowing rapid discovery of thousands of potential microsatellite *loci* in the genome of model and non-model organisms (Ewers-Saucedo *et al.*, 2016; Vartia *et al.*, 2016; De Barba *et al.*, 2017; McKendrick *et al.*, 2017). Despite its long success in obtaining diploid robust genetic information, the analysis of a population genetics data set typically involves a variety of software packages, each of them with a different input data format (Coombs *et al.*, 2008). In addition, researchers daily need to manage genetic, biological, and environmental data of the samples, storing them in text files or spreadsheets, which makes it difficult to maintain consistency and traceability. Here is where TEGA comes in. It is a WEB platform developed for easy data population genetics analysis and data management.

Material and Methods

TEGA was initially built with JHipster. The Back-End was developed in JAVA with Spring Boot and PostgreSQL. The procedures for the genotype analysis

Send correspondence to: Dario Ezequiel Elias. Facultad de Ingeniería de la Universidad Nacional de Entre Ríos, Cátedra de Genética, Ruta prov. 11 km, 10 Oro Verde, E3101 Entre Ríos, Argentina. E-mail: darioezequielelias@protonmail.com.

were implemented in bash using R and Python libraries. The Front-End was developed in JavaScript with AngularJS and Bootstrap.

Results

TEGA is a WEB-based stand-alone software (WEB-based platform) that aims at facilitating the daily workflow in research focused on population genetics and molecular ecology. The purpose of TEGA is to contribute to the autonomy of the research teams, by providing them with a means to manage, analyze, and make their data and results available.

To use TEGA, the research teams must download and install the platform on their server. TEGA has a user manual with instructions for its installation and use. Once installed, team members will be able to import, manage, and analyze their data. When the work is finished, the data, results (tables and graphics), and procedures can be accessed by other professionals through the platform.

TEGA's objectives

Facilitating data management

TEGA has a structure based on entities to facilitate management. Every entity has screens with basic functions to: create, read, update, and delete (CRUD). It is also possible to bulk import sample and genotype data (*loci* and alleles). For some entities, like Samples, Projects, and Genotype Analysis, it is also possible to attach files (e.g., pictures and documents). Furthermore, given the large amount of data that can be linked to the samples, TEGA allows the user to create type-safe dynamic attributes and link them to different entities. In addition, it is possible to visualize the samples' geographical position with OpenStreetMap (Figure 1).

TEGA has also implemented a module for management and execution of data analysis procedures. A user with the Investigator role can create the procedures and attach the execution and configuration files, indicate the input data for the procedure (e.g., sample and allele data) and the parameters for execution. These procedures can then be executed from the Genotype Analysis screen.

Providing a way to execute the analysis procedures

TEGA has an entity called Genotype Analysis for the management of data related to execution of genetic analysis procedures. Initially, the user must create sample sets that contain samples grouped according to a specific criterion (e.g., sampling sites or sampling date). Then, users must create a new genotype analysis, selecting the sample sets, *loci*, and the project linked to the analysis. It is then possible to execute the analysis procedures from the platform interface. Once a procedure is in execution, a genotype analysis cannot be edited or deleted, and when it finishes running, the user will get access to the result files from the analysis edition screen. In this way, TEGA links the procedure results with the entry data, procedure, and parameters used, facilitating traceability of the analysis (Figure 2).

Although TEGA is designed so that users (members of the research team) can carry out their own analysis procedures, in its first version we implemented common methods for population genetics studies, like GENEPOP

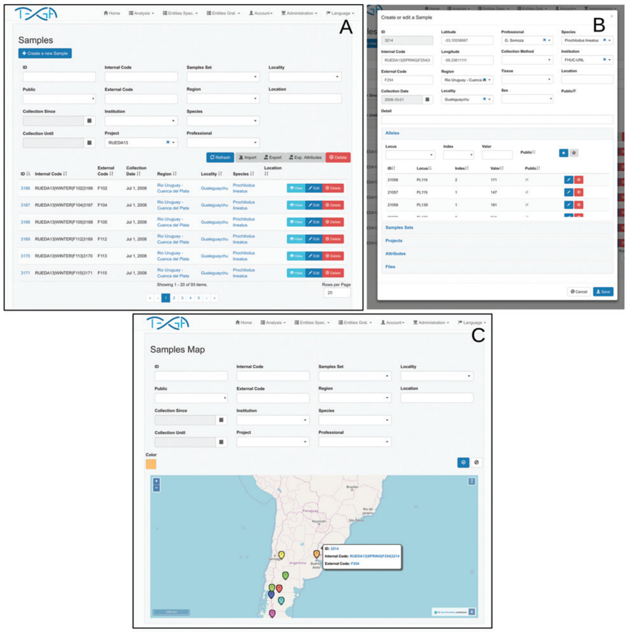

Figure 1 - Sample screens. (A) Query screen; (B) edition and creation screen; (C) map screen.

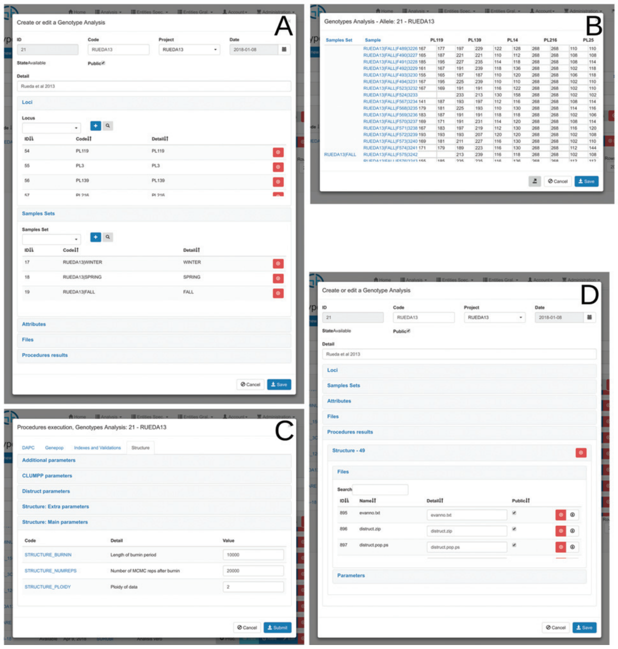
A
Create or edit a Genotype Analysis
B
Genotypes Analysis - Allele: 21 - RUEDA13
C
Procedures execution. Genotypes Analysis: 21 - RUEDA13
D
Create or edit a Genotype Analysis

Figure 2 - Genotypes Analysis screens. (A) Edition and creation screen (panel of Loci and Samples Sets); (B) query and edition screen of alleles included in the genotypes analysis; (C) screen of analysis procedures execution; (D) edition and creation screen (procedures results panel).

(Rousset, 2008), STRUCTURE (Pritchard *et al.*, 1998), and Discriminant Analysis of Principal Components (Jombart *et al.*, 2010) pipelines. TEGA includes empirical data to test them (Rueda *et al.*, 2013; Kamvar *et al.*, 2014). The user manual has the steps to do this in a very detailed way.

Supplying a means to publish data, procedures and results

TEGA has different user roles to allow private use of the data until the day of publication. Anonymous and Invited roles are intended for people outside the research team, who have read-only access to public data. Administrator and Researcher roles will only be for use by the investigation team members, who have access to public and private data, and can carry out CRUD operations and execute analysis procedures.

When the results of a project are published, the users of the platform (with Researcher or Administrator roles) can switch the project status, changing it to public. This action will change the status of samples, alleles, *loci,* and the genotype analysis related with the project, in order to be explored by users with Anonymous or Invited roles. In this way, TEGA simplifies data and result publication. In addition, it is also possible to change the status of the analysis procedures.

Rabid Fish

Rabid Fish is the first TEGA implementation in the Genetics Laboratory of the Faculty of Humanities and Sciences at the National University of the Litoral (FHUC-UNL). Its research goal is focused on population genetics studies, with emphasis in conservation questions related to non-model wetland organisms that are endangered or managed. The laboratory has obtained molecular markers (microsatellites) with traditional methods and NGS technologies (Rueda *et al.*, 2011a,b ; Metz *et al.*, 2016; Ojeda *et al.*, 2017). The fieldwork comprises the area of the La Plata basin, which it is the second-largest river basin of South America, including major rivers such as the Parana, Paraguay, and Uruguay systems. The research group obtained and analyzed fish samples from different migratory and commercially exploited species (Rueda *et al.*, 2011a,b, 2013, 2017; Metz *et al.*, 2016; Ojeda *et al.*, 2017) resulting in a huge collection of samples donated by many collaborators, with different biological issues, and including more than 20 sites from five countries and four species: *Prochilodus lineatus*, *Salminus brasiliensis*, *Leporinus obtusidens,* and *Pseudoplatystoma corruscans*.

The implementation of TEGA in the Genetics Laboratory of FHUC-UNL simplified data management and al-

Table 1 - Platforms for populations genetic data management.

	TEGA	DRIVERGENOME	LOVD
Aims	1) Facilitating population genetics data management.	Assist population genetics and genetic epidemiology studies performed by small-medium research groups, by providing storage, query, and format conversion functionalities.	Web-based software for the collection, display, and curation of DNA variants in locus-specific databases.
	2) Providing a way to execute the analysis procedures.		
	3) Supplying a means to publish data, procedures and results.		
Genetic data	STRs	SNPs, Indels, STRs and CNPs	SNPs, Indels, STRs and CNPs
Focused on a species	No. The species is an attribute of the samples.	Yes. *Homo sapiens*	Yes. *Homo sapiens*
Tools	Visualization:		
	- Samples Map		
		Data format conversion tools for population genetics software.	Data visualization tools and reference sequence parser.
	Analysis:		
	- STRUCTURE		
	- GENEPOP		
	- DAPC		
Extensibility	The user can add or modify analysis procedures without needing to modify the sources or restart the platform.	New conversion tools can be added by experienced users.	Developers can incorporate additional tools into the source code.
Dynamic Attributes	Yes	No	Yes
Reference	This study	Magalhães *et al.*	2012

STR = Short Tandem Repeats; SNP = Single Nucleotide Polymorphism; CNP = Copy Number Polymorphisms.

lowed to publish the data and studies results, creating a new resource for the community, a database called Rabid Fish (see Internet Resources section).

Discussion

TEGA is a WEB-based platform developed for the easy analysis and management of population genetics data. There are WEB-based platforms with purposes similar to those of TEGA, for example DRIVERGENOME (Magalhães *et al.*, 2012) and LOVD (Fokkema *et al.*, 2011) (Table 1). LOVD focuses on the collection, display, and curation of DNA variants in *locus*-specific databases. The aim of DRIVERGENOME is to assist population genetics and genetic epidemiology studies performed by small-to-medium research groups, by providing storage, query, and format conversion functionalities. Both platforms focus on model organisms for which sequenced genomes are available, particularly *Homo sapiens*. Although these platforms offer by default tools to export genetic data in different formats and tools for the visualization of variants, they do not have tools to analyze, for example, the structure of the population.

The main advantage of TEGA is its approach to the management and execution of the analysis procedures. TEGA allows to select the samples and loci for performing a genotype analysis, from where it is possible to execute multiple procedures. In addition, the user can append new procedures to the platform without having to modify its source code, or restarting it. As the procedures can be developed in any language, it is not necessary for the user to do this in the languages and frameworks used in TEGA. The intention is that the user can append the scripts he/she uses daily. In turn, the management also includes the results of the execution of the procedures, the indicated parameters, and the resulting files. In this way, TEGA carries out an integral management of the data, results, and procedures, which can then be published in the same way. We believe that this feature of TEGA will facilitate the access to data and procedures, and allow the easy reproduction of studies. We believe that this is aligned with the current needs of the scientific community, as reflected in the Open Data and Open Science movements (Fecher and Friesike, 2012; Nosek *et al.*, 2015).

Conclusions

TEGA is a WEB-based platform that aims at increasing the autonomy of researchers in the management, analysis and publication of data, procedures, and results. In the future, we plan to add other analysis tools, to integrate TEGA with other databases and improve their implementation and internationalization.

Acknowledgments

The authors thank Ing. G. Cagnola for help in the Rabid Fish implementation (Informatics Department of FHUC-UNL) and to E. Goddio for creating TEGA and Rabid Fish logos. Finally, we thank L. A. Spinetta (the true Rabid Fish).

Conflict of Interest

The authors declare no conflicts of interest.

Author Contributions

DEE and ECR conceived and developed the project. DEE and ECR wrote the manuscript.

References

Avise JC (1998) The history and purview of phylogeography: A personal reflection. Mol Ecol 7:371-379.

Charlesworth B and Charlesworth D (2017) Population genetics from 1966 to 2016. Heredity 118:2–9.

Coombs JA, Letcher BH and Nislow KH (2008) CREATE: A software to create input files from diploid genotypic data for 52 genetic software programs. Mol Ecol Resour 8:578–580.

Crandall KA, Bininda-Emonds ORP, Mace GM and Wayne R K (1999) Considering evolutionary processes in conservation biology. Trends Ecol Evol 15:290–295.

De Barba M, Miquel C, Lobréaux S, Quenette PY, Swenson JE and Taberlet P (2017) High-throughput microsatellite genotyping in ecology: Improved accuracy, efficiency, standardization and success with low-quantity and degraded DNA. Mol Ecol Resour 17:492–507.

Ewers-Saucedo C, Zardus JD and Wares JP (2016) Microsatellite loci discovery from next-generation sequencing data and loci characterization in the epizoic barnacle *Chelonibia testudinaria* (Linnaeus, 1758). PeerJ 4:e2019.

Fecher B and Friesike S (2012) Open Science: One term, five schools of thought. In: Bartling S and Friesike S (eds) Springer International Publishing, Cham, pp 17-47.

Frankham R (2010) Challenges and opportunities of genetic approaches to biological conservation. Biol Conserv 143:1919-1927.

Fokkema IF, Taschner PE, Schaafsma GC, Celli J, Laros JF and den Dunnen JT (2011) LOVD v. 2.0: the next generation in gene variant databases. Hum Mutat 32:557-563.

Jarne P and Lagoda PJL (1996) Microsatellites, from molecules to populations and back. Trends Ecol Evol 11:424–429.

Jombart T, Devillard S and Balloux F (2010) Discriminant analysis of principal components: A new method for the analysis of genetically structured populations. BMC Genet 11:94.

Kamvar ZN, Larsen MM, Kanaskie AM, Hansen EM and Grünwald NJ (2014) Spatial and temporal population dynamics of the sudden oak death epidemic in Oregon forests. Phytopathology 105:982-989.

Magalhães WC, Rodrigues MR, Silva D, Soares-Souza G, Iannini ML, Cerqueira GC, Faria-Campos AC and Tarazona-Santos E (2012) DIVERGENOME: A bioinformatics platform to assist population genetics and genetic epidemiology studies Genet Epidemiol 36:360-367.

McKendrick L, Provan J, Fitzpatrick Ú, Brown MJF, Murray TE, Stolle E and Paxton RJ (2017) Microsatellite analysis supports the existence of three cryptic species within the bumble bee *Bombus lucorum sensu lato*. Conserv Genet 18:573–584.

Metz S, Cabrera JM, Rueda E, Giri F and Amavet P (2016) FullSSR: Microsatellite Finder and Primer Designer. Adv Bioinform 2016:6040124.

Moritz C (1992) Defining "Evolutionarily Significant Units" for conservation. Trends Ecol Evol 9:373–375.

Nosek BA, Alter G, Banks GC, Borsboom D, Bowman SD, Breckler SJ, Buck S, Chambers CD, Chin G, Christensen G *et al.* (2015) Promoting an open research culture. Science 348:1422–1425.

Ojeda GN, Amavet PS, Rueda EC, Siroski PA and Larriera A (2017) Mating system of *Caiman yacare* (Reptilia: Alligatoridae) described from microsatellite genotypes. J Hered 108:135–141.

Pritchard JK, Stephens M and Donelly P (1998) Inference of population structure using multilocus genotype data. Genetics 155:945–959.

Rousset F (2008) genepop'007: A complete re-implementation of the genepop software for Windows and Linux. Mol Ecol Resour 8:103–106.

Rueda EC, Amavet P, Brancolini F, Sommer J and Ortí G (2011a) Isolation and characterization of eight polymorphic microsatellite markers for the migratory characiform fish, *Salminus brasiliensis*. J Fish Biol 79:1370–1375.

Rueda EC, Sommer J, Scarabotti P, Markariani R and Ortí G (2011b) Isolation and characterization of polymorphic microsatellite loci in the migratory freshwater fish *Prochilodus lineatus* (Characiformes: Prochilodontidae). Conserv Genet Resour 3:681–684.

Rueda EC, Carriquiriborde P, Monzón AM, Somoza GM and Ortí G (2013) Seasonal variation in genetic population structure of sábalo (*Prochilodus lineatus*) in the Lower Uruguay River. Genetica 141:401–407.

Rueda EC, Mullaney KA, Conte-Grand C, Habit EM, Cussac V and Ortí G (2017) Displacement of native Patagonian freshwater silverside populations (*Odontesthes hatcheri*, Atherinopsidae) by introgressive hybridization with introduced *O. bonariensis*. Biol Invasions 19:971–988.

Selkoe KA and Toonen RJ (2006) Microsatellites for ecologists: A practical guide to using and evaluating microsatellite markers. Ecol Lett 9:615–629.

Vartia S, Villanueva-Cañas JL, Finarelli J, Farrell ED, Collins PC, Hughes GM, Carlsson JEL, Gauthier DT, McGinnity P and Cross TF (2016) A novel method of microsatellite genotyping-by-sequencing using individual combinatorial barcoding. R Soc Open Sci 3:150565.

Vignal A, Milan D, SanCristobal M and Eggen A (2002) A review on SNP and other types of molecular markers and their use in animal genetics. Genet Sel Evol 34:275–305.

Internet Resources

AngularJS, https://angularjs.org (accessed 14 January 2019).

Bootstrap, https://getbootstrap.com (accessed 14 January 2019).

JHipster, https://www.jhipster.tech (accessed 14 January 2019).

OpenStreetMap, https://www.openstreetmap.org (accessed 14 January 2019).

PostgreSQL, https://www.postgresql.org (accessed 14 January 2019).

Rabid Fish, https://fhuc0.unl.edu.ar:5611 (accessed 14 January 2019).

Spring Boot, http://spring.io/projects/spring-boot (accessed 14 January 2019).

Associate Editor: Fabrício Rodrigues dos Santos