Extending GelJ for Interoperability: Filling the Gap in the Bioinformatics Resources for Population Genetics Analysis with Dominant Markers

César Domínguez[a], Jónathan Heras[a,*], Eloy Mata[a], Vico Pascual[a], Maria Soledad Vázquez-Garcidueñas[b], Gerardo Vázquez-Marrufo[c,*]

[a]*Department of Mathematics and Computer Science, University of La Rioja, Logroño, Spain.*
E-mail: {cesar.dominguez, jonathan.heras, vico.pascual}@unirioja.es
[b]*Division of Postgraduate Studies, Faculty of Medical and Biological Sciences "Dr. Ignacio Chávez", Universidad Michoacana de San Nicolás de Hidalgo, Mexico.*
E-mail: marisolvaz@yahoo.com
[c]*Multidisciplinary Center of Biotechnology Studies (CMEB), Faculty of Veterinary Medicine, Universidad Michoacana de San Nicolás de Hidalgo, Mexico.*
E-mail: gvazquezmarrufo@yahoo.com.mx

Abstract

Background and Objective. The manual transformation of DNA fingerprints of dominant markers into the input of tools for population genetics analysis is a time-consuming and error-prone task; especially when the researcher deals with a large number of samples. In addition, when the researcher needs to use several tools for population genetics analysis, the situation worsens due to the incompatibility of data-formats across tools. The goal of this work consists in automating, from banding patterns of gel images, the input-generation for the great diversity of tools devoted to population genetics analysis.

Methods. After a thorough analysis of tools for population genetics analysis with dominant markers, and tools for working with phylogenetic trees; we have detected the input requirements of those systems. In the case of programs devoted to phylogenetic trees, the Newick and Nexus formats are widely employed; whereas, each population genetics analysis tool uses its own specific format. In order to handle such a diversity of formats in the lat-

*Corresponding authors

ter case, we have developed a new XML format, called PopXML, that takes into account the variety of information required by each population genetics analysis tool. Moreover, the acquired knowledge has been incorporated into the pipeline of the GelJ system — a tool for analysing DNA fingerprint gel images — to reach our automatisation goal.

Results. We have implemented, in the GelJ system, a pipeline that automatically generates, from gel banding patterns, the input of tools for population genetics analysis and phylogenetic trees. Such a pipeline has been employed to successfully generate, from thousands of banding patterns, the input of 29 population genetics analysis tools and 32 tools for managing phylogenetic trees.

Conclusions. GelJ has become the first tool that fills the gap between gel image processing software and population genetics analysis with dominant markers, phylogenetic reconstruction, and tree editing software. This has been achieved by automating the process of generating the input for the latter software from gel banding patterns processed by GelJ.

Keywords: Interoperability, GelJ, Dominant Markers, Population Genetics, Phylogenetic Trees, PopXML.

---

1. Introduction

Dominant markers including AFLPs, ISSRs, rep-PCR (BOX, ERIC, and REP), and RAPDs are useful tools for population genetic analysis — several applications of these markers are listed in Supplementary Material S1. Some of the advantages of dominant markers are that they do not require previous knowledge about the genome of the studied species, and that they allow detection of intraspecific differences across the whole genome at different ploidy levels [1, 2]. In addition, except for AFLP assays based on fluorescent detection in capillary electrophoresis, techniques for fingerprinting with dominant markers are simple and can be made in agarose or polyacrylamide gels using basic molecular laboratory equipment; making them a cheap solution for population genetic analysis of plants, animals, and microorganisms — both prokaryotic and eukaryotic. In spite of the existence of some disadvantages of dominant markers [2], those drawbacks can be overcome by a combined use with other dominant or codominant markers and by a suitable bioinformatics analysis [3], thus becoming useful for analysing inter and

intra-population genetic differentiation and structure, dispersion, migration, genotype-environment associations, and gene flow, among others.

Nowadays, there is a wide variety of software tools for population genetics analysis with dominant markers featuring, among other functionalities, the computation of genetic diversity indices and F-statistics, and the visualisation and edition of phylogenetic trees — see [4] and Supplementary Materials S2 and S3. In order to employ software tools for population genetics analysis, it is necessary to transform dominant marker fingerprints, that consist of complex gel banding patterns, into either band presence (1) – absence (0) binary matrices or phylogenetic trees that will be used as input of those systems. Unfortunately, the generation of such an input from banding patterns of gel images might be a challenge.

In spite of the existence of several programs for dealing with banding patterns of gel images of dominant markers [5], these systems have not been designed to interact with software for population genetics analysis. In the case of presence/absence matrices, software for gel banding patterns sometimes construct those matrices internally; but, in general, they do not support their exportation, and when they do, the format of the exported matrices is not compatible with the input format of software tools for population genetics analysis. In the case of phylogenetic trees, several tools for gel banding patterns feature the generation of such trees; however, the generated trees can only be saved as images; and, hence, they cannot be fed as input to the tools for population genetics analysis. Therefore, the task of creating the input for population genetic analysis software must be carried out manually. This is a laborious, subjective, time-consuming, error-prone, and unreproducible task, which might produce unreliable results. Moreover, the risk of generating unreliable inputs is increased when using a large number of individuals and loci, which are needed to obtain reliable population genetics data.

In addition to the drawback of generating the input of software tools for population genetics analysis, there is another challenge in this context: interoperability among tools. Researchers normally need to analyse the same data with several programs; unfortunately, most of the programs that take presence/absence matrices as input use specific data-file formats [4]. Then, it is necessary to either manually transform the data across formats or use conversion tools. Neither approach is fully satisfactory, the former since is tedious, error-prone and not suitable when dealing with a large number of individuals; and the latter because conversion tools do not cover all the

possible systems, depend on the version of the programs and, in general, cannot be easily adapted to handle new file formats that might arise with new systems. The situation is much better in the case of phylogenetic trees [6], since there are two standard formats widely employed by the majority of the systems: Newick [7] and Nexus [8]. Hence, the same input can be employed by several systems.

In this paper, we present how we have tackled the aforementioned problems to achieve our goal: automatic input generation for the wide variety of systems devoted to population genetics analysis and phylogenetic-trees editing from banding patterns of gel images. The first step to reach that aim has been a thorough review of software for population genetics analysis and phylogenetic trees to identify the characteristics of the input of such tools. From that review, we have checked that the formats Newick and Nexus are widely employed by software for phylogenetic-trees editing; and, in addition, we have defined a new format, called PopXML, that puts together all the information needed by the diversity of formats employed to encode presence/absence matrices in software for population genetic analysis with dominant markers. Finally, we have expanded the functionality of the GelJ system [9], an open-source and free tool for analysing DNA fingerprint gel images, to generate presence/absence matrices and phylogenetic trees that can be employed not only by the tools surveyed in our review, but also by new tools that might appear in the future. As a result, we have created the first existing tool that fills the bioinformatics gap between gel image processing software and population genetics analysis, phylogenetic reconstruction, and tree editing software.

2. A Review of Software for Population Genetics Analysis and Phylogenetic Trees

In this section, we survey software tools for population genetic analysis with dominant markers that take presence/absence matrices as input, and tools for managing phylogenetic trees that work with the standard formats Newick and Nexus. The final aim of this survey is threefold: find the tools available for population genetics analysis with dominant markers; identify the characteristics of the input files of those tools; and, check whether tools that use either the Newick or the Nexus format to encode phylogenetic trees provide the necessary functionality to handle this kind of trees.

We screened PubMed Central and Google Scholar looking for corpora publications, and used the Google search-engine to create two lists of software tools. This search produced 29 tools for population genetic analysis with dominant markers, and 31 tools for managing phylogenetic trees with the Newick or the Nexus format.

2.1. Software for Population Genetics Analysis with Dominant Markers

In the last 20 years, a great diversity of software tools for population genetics analysis has been developed with different aims and handling different kinds of data. A survey of 25 of those tools was provided in [4]. Such a survey included tools that support data types like DNA sequences, dominant markers, or multi-allelic markers. In our case, we are focused on the tools that work with dominant markers and take as input presence/absence matrices. The programs that have been included in our survey are listed in Table 1.

As can be seen in Table 1, our main interest was not to perform a thorough analysis of the features of each system. On the contrary, we were interested in spotting how the presence/absence matrices are represented in the input files for each system, and what is the supplementary information needed in those input files. As we will explain in Sections 3 and 4, this knowledge has been employed to, first define a new format that takes into account the requirements of the variety of systems; and, then to allow the connection of GelJ with all the tools of Table 1.

In most tools, the presence/absence matrices are encoded using a 1 to indicate the presence of a band and a 0 to indicate its absence, but there are some systems that use a different representation (e.g. Mcheza and NewHybrids); in fact, two systems (ABC4F and Bayescan) do not work directly with the presence/absence matrix but with a frequency matrix. The additional information varies from system to system, and might include some of the following information: number of loci, number of populations, number of individuals, number of individuals per population, names of loci, names of populations, names of individuals, and individuals of each population. In addition to the differences among systems presented in Table 1, the input files of each system have their own peculiarities that are not related with the data (e.g. keywords, order of the data, characters employed to separate data, and so on). This variety of formats shows the diversity of the field and the difficulty of manually transforming data across formats.

Tool	Year	Version	Representation of presence/absence matrix	Additional information in the input file
ABC4F	2008	1.0	Number of individuals observed at each locus.	Number of loci, number of populations, and individuals of each population.
Adegenet	2016	2.0.1	Matrix of 1s (presence) and 0s (absence) separated by tabs.	Names of loci, and names of individuals.
aflp-surv	2002	1.0	Matrix of 1s (presence) and 0s (absence) separated by tabs.	Number of loci, number of populations, names of loci, names of individuals, names of populations, and individuals of each population.
AFLPdat	2008	23/01/08	Matrix of 1s (presence) and 0s (absence) separated by tabs.	Names of loci, names of individuals, names of populations, and individuals of each population.
Alleles in space (AIS)	2005	1.0	Matrix of 1s (presence) and 2s (absence) separated by commas.	Number of loci, names of individuals.
Arlequin	2015	3.5	Matrix of 1s (presence) and 0s (absence) without separation.	Number of individuals, number of populations, names of individuals, and individuals of each population.
Baps	2013	6.0	Matrix of 1s (presence) and 0s (absence) separated by tabs.	None.
Bayescan	2012	2.1	Number of individuals observed at each locus.	Number of loci, number of populations, and individuals of each population.
EDENetworks	2014	2.18	Matrix of 1s (presence) and 0s (absence) separated by tabs.	Names of individuals.
Famd	2013	1.3	Matrix of 1s (presence) and 0s (absence) separated by tabs.	Names of loci, and names of individuals.
Famoz	2009	11/06/09	Matrix of 1s (presence) and 0s (absence) separated by tabs.	Number of loci, and number of individuals per population.
Gda	2008	1.1	Matrix of 1s (presence) and 0s (absence) separated by tabs.	Number of loci, number of populations, names of loci, names of populations, names of individuals, and individuals of each population.
GenAlEx	2012	6.502	Matrix of 1s (presence) and 0s (absence) separated by commas.	Number of loci, number of individuals, number of populations, number of individuals per population, names of populations, names of loci, and names of individuals.
Geneland	2012	4.0.5	Matrix of 1s (presence) and 0s (absence) separated by tabs.	None.
Geneticstudio	2014	1.5	Matrix of 1s (presence) and 0s (absence) separated by commas.	Names of loci, and names of individuals.
GenoDive	2013	2.0b23	Matrix of 1s (presence) and 0s (absence) separated by tabs.	Number of individuals, number of populations, number of loci, names of populations, names of loci, names of individuals, and individuals of each population.
Hickory	2004	1.1	Matrix of 1s (presence) and 0s (absence) separated by tabs.	Number of loci, number of populations, names of populations, names of individuals, and individuals of each population.
Mcheza	2011	1.0	Matrix of 1s (presence) and 2s (absence) separated by tabs.	Names of loci, names of individuals, and individuals of each population.
NewHybrids	2003	1.1 Beta 3	Matrix of +s (presence) and -s (absence) separated by tabs.	Number of individuals, number of loci, and names of loci.
Past	2015	1.0	Matrix of 1s (presence) and 0s (absence) separated by commas.	Names of loci, and names of individuals.
Piccalc	2012	1.0	Matrix of 1s (presence) and 0s (absence) without separation.	Number of individuals, and names of individuals.
Popgene	2000	1.32	Matrix of 1s (presence) and 0s (absence) without separation.	Number of populations, number of loci, names of loci, names of populations, and individuals of each population.
Sambada	2015	0.5.1	Matrix of 1s (presence) and 0s (absence) separated by tabs.	Names of loci, names of individuals, number of loci, and number of individuals.
Spagedi	2015	1.5	Matrix of 1s (presence) and 0s (absence) separated by tabs.	Number of individuals, number of populations, number of loci, names of loci, names of individuals, names of populations, and individuals of each population.
Structure	2012	2.3.4	Matrix of 1s (presence) and 0s (absence) separated by tabs.	Names of individuals, number of individuals, and number of loci.
Tess	2015	3	Matrix of 1s (presence) and 0s (absence) without separation.	None.
TFPGA	1997	1.3	Matrix of 1s (presence) and 2s (absence) separated by commas.	Individuals of each population.
Treecon	1998	1.3b	Matrix of 1s (presence) and 0s (absence) without separation.	Number of loci, and names of individuals.
Wombsoft	2007	1.0	Matrix of 1s (presence) and 0s (absence) separated by tabs.	Names of individuals.

Table 1: **List of tools for population genetic analysis with dominant markers that take presence/absence matrices as input**. Additional information about these tools, including references and webpages, is provided in Supplementary Material S2.

2.2. Software for Phylogenetic Trees

Tree structures are instrumental in biology for the visualisation and analysis of results; and, in particular, phylogenetic trees are broadly employed in population genetics analysis. In [6], five key functionalities of tools for handling trees were identified: tree visualisation, tree comparison, tree editing, tree annotation, and tree analysis. In this section, we check whether such features are covered by the systems that take as input phylogenetic trees encoded using either the Newick or the Nexus format. We have listed the surveyed tools in Table 2.

As can be seen in Table 2, the programs that work with the Newick or the Nexus format successfully cover the 5 key features. There are 18 for visualising trees, 1 for tree comparison, 9 for tree editing, 3 for tree annotation, and 11 for tree analysis. Therefore, we can conclude that generating phylogenetic trees in either the Newick or the Nexus format is enough to interact with systems that cover the functionality for handling and analysing those trees.

3. The PopXML Format

In our aim to automatise the input-generation for tools devoted to population genetics analysis — and based on the knowledge acquired in the previous section — it is clear that the phylogenetic trees generated from banding patterns of gel images should be encoded using either the Newick or the Nexus format. However, the encoding of the presence/absence matrices generated from banding patterns is not so clear due to the wide variety of formats employed by the tools of Table 1.

In order to deal with this issue, we have designed a new format that collects all the information needed by the tools surveyed in Table 1 to encode presence/absence matrices. This new format, called *PopXML*, is based on the XML (eXtensible Markup Language) format and is therefore independent of any particular computer system and extensible for future needs. The structure of XML files following the PopXML format is fixed by an XML Schema [10], that not only determines the structure of XML files but also specifies and restricts the content of their elements — see Figure 1 for a summary of the XML Schema of PopXML, and Supplementary Material S4 for the whole schema. This schema was developed taking into account the variety of information that is needed by the input files of the systems surveyed in Table 1. PopXML files are structured in two main blocks: general

Tool	Year	Version	Newick/Nexus	Main tree features
Baobab	2008	3.31	Both	Annotation, Editing
BayesTrees	2011	1.3	Nexus	Bayesian estimation of phylogeny
Bio++	2014	2.2.0	Nexus	Generic library for C++: analysis
Bio-Nexus	2012	0.78	Nexus	Generic library for Perl: analysis
BioPerl	2009	1.6.1	Newick	Generic tool in Perl: analysis
Biopython	2015	1.66	Both	Generic tool in Python: analysis
CTree	2007	1.0	Newick	Cluster visualization, Editing
Dendroscope	2012	3	Both	Visualisation
ETE Toolkit	2016	3.0.0b33	Newick	Generic tool in Python: analysis
FigTree	2014	1.4.2	Newick	Visualisation
GARLI	2013	2.01	Newick	Maximum-likelihood phylogenetic analysis
Geneious	2016	9.13	Both	Generic tool: visualisation and evolutionary analysis
HyperGeny	2004	1.0	Newick	Visualisation
iTOL	2011	3.0	Newick	Visualisation, Annotation,Editing
MacClade	2011	4.08a	Nexus	Editing
Matlab	2016	R2016a	Both	Generic tool: visualisation and evolutionary analysis
Mega	2016	7	Both	Generic tool: visualisation and evolutionary analysis
Mesquite	2015	3.0.4	Nexus	Visualisation, Editing, Comparison
MOLPHY	1996	2.3	Newick	Visualisation
MulRF	2014	1.2	Newick	Visualisation
Network	2016	5.0	Nexus	Visualisation
NJPlot	1996	2.3	Newick	Visualisation
Paloverde	2006	1.1	Nexus	3D visualisation
Phylowidget	2008	-	Both	Visualisation, Editing
PROTML	1996	1.0	Newick	Visualisation
RaXML	2016	8.2	Newick	Maximum-likelihood analysis
T-Rex	2016	Web	Newick	Visualisation, Preparing for a publication
TreeGraph2	2010	2	Nexus	Editing, Annotation
TreeIllustrator	2005	0.52	Newick	Visualisation, Editing
TreePuzzle	2015	5.3.rc16	Newick	Maximum-likelihood analysis
TreeView	2001	1.6.6	Nexus	Visualisation, Editing

Table 2: **List of tools for phylogenetic trees that work with either the Newick or the Nexus format.** Additional information about these tools, including references and webpages, is provided in Supplementary Material S3.

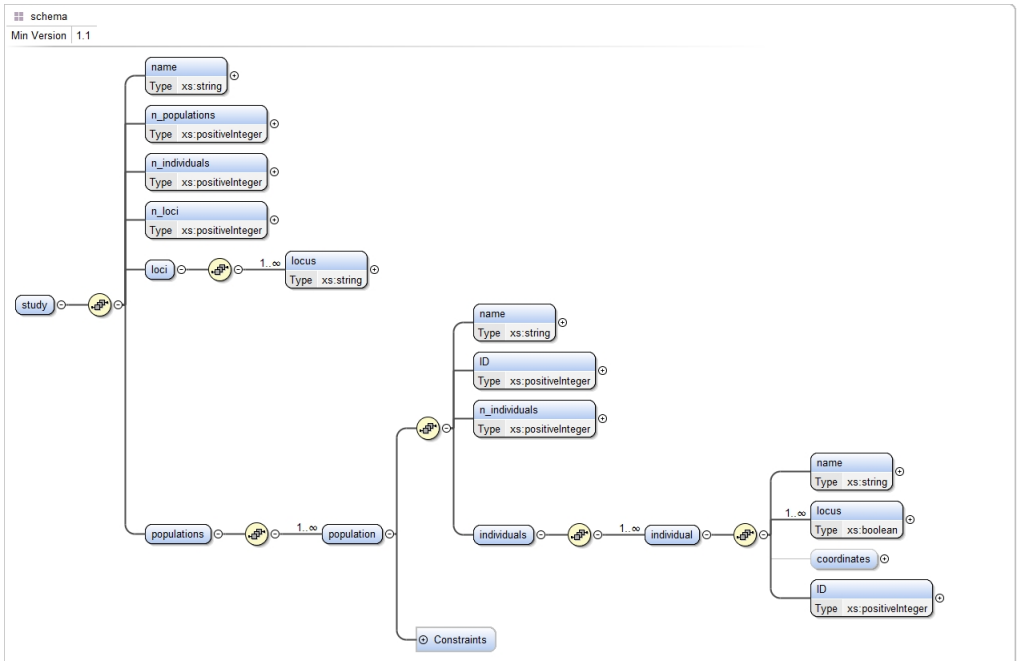

Figure 1: **XML Schema of PopXML.**

information and populations' data. The former gives information about general aspects such as the number of populations, the number of individuals, the number of loci, and the names of loci. The latter contains information about the populations, their member individuals, the presence/absence matrices of the individuals, and, optionally, the geographic coordinates of each individual (see Figure 1). Files that follow the PopXML format can be nicely visualised in any web browser thanks to an XSLT (eXtensible Stylesheet Language Transformations) file [10] — an example is provided in Supplementary Material S5.

The greatest advantage of using the PopXML format is that it simplifies the conversion from this format to the specific format of each system of Table 1. Instead of hard-coding the conversion by creating an individual program for each transformation, we define XSLT files that transform from the PopXML format to the required formats taking into account the concrete characteristics of each input file presented in Table 1. Hence, given a presence/absence matrix encoded using the PopXML format, it can be transformed to the desired format just by applying the corresponding XSLT file.

In addition, compatibility with new formats can be easily achieved by just defining new XSLT files. Currently, we have defined a total of 29 XSLT files to export presence/absence matrices, encoded using the PopXML format, to different data formats compatible with all the tools included in Table 1.

Taking into account that our final aim is the automatic generation of presence/absence matrices and phylogenetic trees from gel banding patterns; and since we have suitable formats to encode presence/absence matrices (the PopXML format) and phylogenetic trees (the Newick and Nexus formats), it remains the question of how these objects are generated in those formats from banding patterns. We have tackled this problem using the GelJ tool.

4. GelJ Features for Population Genetics Analysis

GelJ [9] is an easy to use open-source tool for analysing DNA gel fingerprint images. This tool was developed with the aim of providing a system that overcomes the lack of functionality of open-source tools for analysing DNA gel fingerprint images, and with the additional advantage of being free and simpler to use than commercial programs. Up to now, the main feature of GelJ has been the comparison of DNA fingerprinting samples from several gel images using phylogenetic trees (also called dendrograms) — this involves functionality for image processing, lane and band detection, normalisation of fingerprints, database support, and construction of dendrograms. The new version of GelJ, which we are introducing in this paper, also features the connection of GelJ with software for population genetics analysis.

As we have mentioned in the Introduction, the new version of GelJ deals with the challenge of input-generation for the diversity of software for population genetics analysis and phylogenetic-trees managment. The process to tackle such a challenge can be summarised as follows, see also Figure 2. From a set of DNA fingerprint gel images, GelJ transforms the banding patterns in those images into an internal representation using the molecular weights of their bands, and store them in a database — this is represented in Figure 2 as the GelJ processing module and was already implemented in the first version of the system. Subsequently, from the internal representation of banding patterns, presence/absence matrices and phylogenetic trees are generated inside GelJ. Finally, the presence/absence matrices and the phylogenetic trees can be exported to the surveyed tools using a dynamic approach, that employs the PopXML format, in the case of presence/absence matrices, and using the standard formats (Newick and Nexus) for phylogenetic trees. The rest of

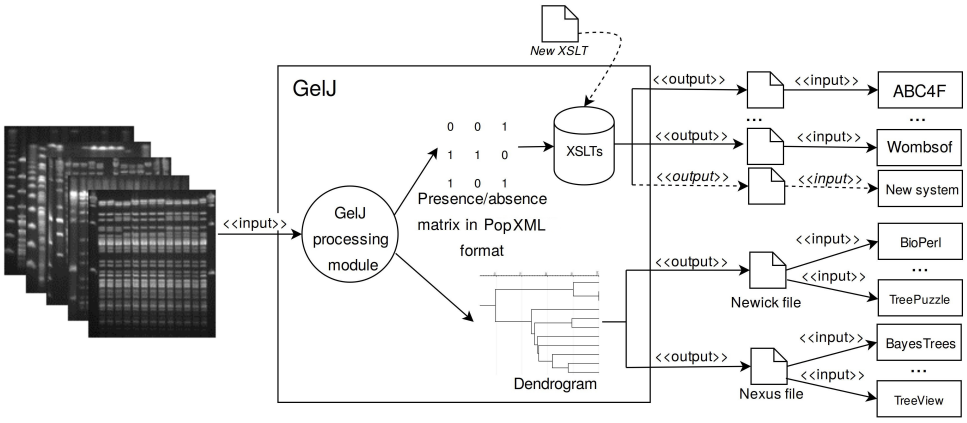

Figure 2: **Workflow of GelJ to generate binary matrices in different formats and dendrograms in the Newick and Nexus standard formats.**

this section is devoted to explain how the last two steps of this process are carried out. Let us focus first on the case of presence/absence matrices.

4.1. Creating and exporting presence/absence matrices

The construction of presence/absence matrices employed in GelJ is based on a matching process that is employed by several tools as a previous step for constructing similarity matrices [9, 11, 12], a detailed description of this process is described in Supplementary Material S6. In GelJ, the number of individuals that can be employed to generate the presence/absence matrices is only limited by the characteristics of the user's computer, and a standard desktop computer can generate presence/absence matrices using thousands of individuals.

Once that the presence/absence matrices are generated from the banding patterns inside GelJ, it remains the question of how these matrices are transformed from the internal representation of GelJ to the formats of software for population genetics analysis. To this aim, we have employed an approach that uses the PopXML format, as an intermediate step in the conversion process, and the XSLT files defined to transform files using the PopXML format to the specific format of each tool. In particular, instead of hard-coding the conversion to the different formats, whenever the user asks to export a presence/absence matrix, such a matrix is internally transformed to the PopXML format, and, subsequently, the user decides to which formats should be exported (see Figure 3) using the XSLT files. The available for-

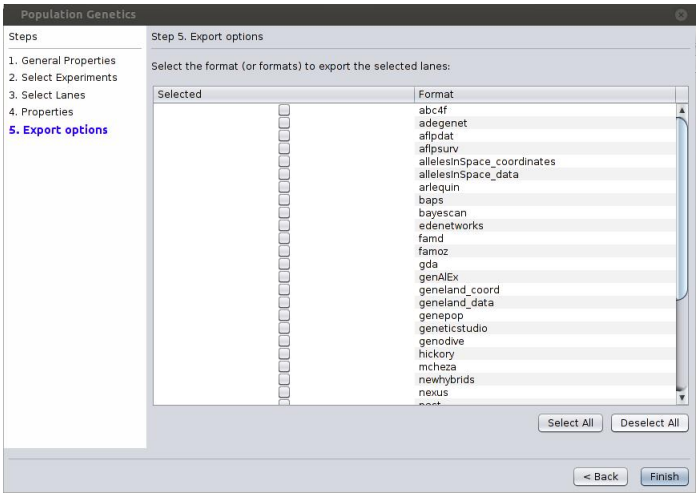

Figure 3: **Interface of GelJ to select the formats to export the presence/absence matrices.**

mats to export presence/absence matrices are dynamically loaded from the XSLT files stored in a folder of the GelJ distribution. Hence, compatibility with new formats can be achieved by just adding new XSLT files to the GelJ folder. It is worth noting that this approach solves the problem of converting data across several formats, since the user can select several converters to directly export the data in the desired formats for different systems.

Let us finish this subsection by summarising the process to export presence/absence matrices. First of all, the user analyses several DNA gel fingerprinting images to store banding patterns inside GelJ. Subsequently, the user selects from the banding patterns available in the GelJ database, those that she wants to include in the presence/absence matrix that will be generated. Afterwards, the presence/absence matrix is generated and converted it to the PopXML format. Finally, the user selects the formats to export the presence/absence matrix. The whole process is guided by means of wizards to make the user-interaction easier. This functionality has been thoroughly tested; namely, we have exported binary matrices — including the information of thousands of banding patterns — for all the tools presented in Table 1, and checked that all the generated files could be loaded and employed in the corresponding tool.

4.2. Creating and exporting phylogenetic trees

GelJ not only supplies the functionality to generate and export presence/absence matrices to different formats, but it also features the generation of phylogenetic trees, in both the Newick and the Nexus format, to be employed by the tools surveyed in Table 2.

As we have previously explained, the first version of GelJ already featured the generation of dendrograms to allow the visual comparison of banding patterns [9]. However, the dendrograms could only be saved as images, and, hence, they could not be employed by tools for analysing phylogenetic trees. The new version of GelJ has been enhanced to be able to export dendrograms in the two standard formats Newick and Nexus. The functionality to create and export dendograms is a four-step process that involves (1) the selection of individuals to compare, (2) the computation of similarity matrices, (3) the generation of dendrograms, and (4) the exportation of the dendrograms — Steps 1 to 3 were already supported by the first version of GelJ, Step 4 is supplied by the new version of the system.

In the aforementioned process, Step 1 is trivial: the user selects from the GelJ database the individuals that she wants to compare. There are two approaches to construct in GelJ the similarity matrices of Step 2: the *band-based* approach, and the *curve-based* approach, see [9] for a detailed explanation. using In both cases, the user only selects the metric and the similarity matrix is automatically constructed by GelJ. In Step 3, GelJ invokes the Weka machine-learning package [13] with the similarity matrix constructed in Step 2 and the linkage method (e.g. UPGMA, WPGMA, or single linkage [9]); and Weka returns a dendrogram in the Newick format — this result is employed to visualise dendrograms inside GelJ. Finally, in Step 4, the dendrogram can be exported in the Newick format (using the output generated by Weka), the Nexus format (after a transformation step), or in both formats. The exported files can be used by all the external programs listed in Table 2 (as in the case of presence/absence matrices, this functionality has been thoroughly tested), and due to the fact that the Newick and Nexus formats are standards, the same files can be employed by several systems.

5. Discussion

GelJ is the first available tool that fills the gap between gel image processing software and population genetics analysis with dominant markers, phylogenetic reconstruction and tree analysis software. As we have explained, GelJ

solves the challenge of automatic input-generation for the great diversity of software devoted to population genetics analysis and phylogenetic-trees managment.

The generation of presence/absence matrices is an internal process of DNA fingerprinting software, and such matrices are usually hidden to the user. From the 33 tools surveyed in [5], there are only 5 tools that allow the user to visualise and export presence/absence matrices: GelComparII [14], GelQuest [15], ImageQuant [16], PyElph [17], and Phoretix [18] — except for PyElph, the other tools are commercial systems. From those tools, GelComparII is the only one that can export presence/absence matrices generated from samples coming from different images, the rest of them can only export the presence/absence matrices from samples of a single image — from these 5 tools, GelComparII is the only one that provides database support. Moreover, none of these systems export presence/absence matrices in a format compatible with any of the tools for population genetics analysis; hence, the user has to manually modify the exported data in order to fulfil the specifications of the systems included in Table 1. On the contrary, GelJ is an open-source tool that can generate presence/absence matrices from samples of several images thanks to its integrated database, and the exported matrices can be directly employed by software for population genetics analysis.

The situation is a bit better in the case of phylogenetic trees since there are 15 tools for DNA fingerprinting that generate phylogenetic trees (13 commercial systems and 2 open-source tools) [5]. Nevertheless, those phylogenetic trees can only be saved as images, and, therefore, they cannot be imported by the different tools surveyed in Section 2. GelJ handles this problem by allowing the user to export the generated phylogenetic trees in two standard formats for these trees.

The other problem handled by the new version of GelJ is the variety of formats employed by the tools for population genetics analysis and phylogenetic-trees managment. This is an important issue in the context of population genetic analysis with dominant markers, because researchers usually need to analyse the same data set with several programs (e.g. Arlequin, GenAlEx and Structure are employed in [19]; AFLP-SURV, Arlequin, GenAlEx, Popgene and TreeView in [20]; or ABC4F, GenAlEx, Geneland, Hickory, PopGene and Structure in [21]). In the case of software for handling phylogenetic trees this is not a problem since standard formats are employed. Unfortunately, in the case of presence/absence matrices, most of the programs use specific data-file formats. Dealing with different data formats is a tedious, time-consuming,

and difficult task since it involves deciphering input specifications and editing data to adhere to those specifications. Up to now, there were two approaches to tackle this issue: manually create the different files for each tool, or use transformation tools.

The manual transformation across formats means that several file specifications must be used due to the fact that there are just a few systems which can directly employ the files from other systems (see Supplementary Material S7). In fact, there are two systems, Gda and Hickory, which in spite of using the same file format, the Nexus format, are not compatible with each other. Hence, this approach is tedious, time-consuming, error-prone and is not suitable when handling a large number of samples.

The second approach consists in using a transformation tool. Up to the best of our knowledge, there are 8 tools that provide the functionality to transform across formats: AFLPDat [22], Convert [23], Create [24], Famd [25], Formatomatic [26], GenAlEx [27], PGDSpider [28], and Transformer4 [29]. The main difference among these transformation tools is the format of their input files. Convert, Formatomatic, GenAlEx, and Transformer4 take as input an Excel file, and transform it to a variety of formats. In AFLPDat, Create and Famd, the user must create a file with the format of these systems, and then she can export it to other formats. Finally, PGDSpider might take files in several formats as input and produces files in different formats as output. The target tools of these transformation systems are provided in Supplementary Material S7.

Even if this approach of using transformation tools is more suitable than a manual transformation, it still has drawbacks. First of all, there are some systems which are not covered by any of the transformers, and it will be difficult to expand those tools to deal with new systems that might appear in the future. In addition, and as warned by the conversion tools (see, for instance, the GenAlEx manual), the transformation might not be perfect, and the user may need to manually make further modifications in the file to analyse it in the intended software.

A solution to these problems would be the adoption of a standard format that should be followed by every system. This strategy is followed in the case of the Newick and Nexus formats for phylogenetic trees, and also in related domains where several initiatives for standardization and harmonization have been launched (for instance, in medical/clinical informatics [30, 31], bioinformatics [32] or omics [33]). The PopXML format presented in this paper was designed taken into account the specifications of a wide variety of systems for

population genetics analysis with dominant markers; hence, it is our belief, that this format could be adopted as a standard in this context.

However, the adoption of a new format as a standard requires the acceptance by the community, and the adaptation of the systems to read files in such a format. Therefore, in the context of this paper, and for the time being, a more sensible approach, is the one followed by both Transformer4 and GelJ. Such an approach allows the user to transform presence/absence matrices to several formats by means of converters, XSLT files in the case of GelJ and Java plugins in the case of Transformer4, which number can be easily expanded. The only drawback of Transformer4 is that the user must manually create the data files using an Excel spreadsheet; on the contrary, as we have explained previously, GelJ avoids the manual creation of the input files from the banding patterns of gel images. This property is especially relevant when handling a large number of samples. In addition, the PopXML format, together with the developed XSLT files, might be a first step towards the creation of a standalone converter, or a web-service, that could be employed to transform already available data in different formats — an important issue on a field where existing data are already present.

To sum up, the new version of GelJ generates presence/absence matrices (in a variety of formats) and phylogenetic trees (using the standard Newick and Nexus formats) from gel images. Such an output can be directly fed to a wide variety of software tools for population genetics analysis and phylogenetic tree handling. Therefore, GelJ is the first tool for DNA fingerprinting analysis that solves an important problem in software for population genetics analysis with dominant markers: from banding patterns of gel images, GelJ automates the input-generation for the wide variety of systems devoted to population genetics analysis and phylogenetic-trees management.

6. Conclusions

Interoperability is an important aspect in bioinformatics since there is not any tool that can tackle every single problem, and, in several cases, different tools must be combined in order to obtain valuable results. An important issue to achieve interoperability among systems is the sharing of standard formats to communicate them.

In this paper, we have focused on the problem of connecting tools for analysing banding patterns of gel images with tools for population genetics analysis, phylogenetic reconstruction, and tree editing. In order to reach that

aim, we have used standard formats widely employed in tools for phylogenetic trees, and a newly defined format that gathers all the information required by the systems devoted to population genetics analysis. These formats have been incorporated in a pipeline implemented in GelJ — a tool for analysing DNA fingerprint gel images — to automatically generate the input of tools for population genetics analysis, phylogenetic reconstruction, and tree editing. In spite of the existence of several tools for the study of gel images [5], this is, up to the best of our knowledge, the first time that such a cooperation has been achieved.

Availability and requirements

- Project name: GelJ v2
- Project home page: `https://sourceforge.net/projects/gelj/`.
- Operating system(s): Platform independent.
- Programming language: Java.
- Other requirements: Java.
- License: GNU GPL 3.0.
- Any restrictions to use by non-academics: None.

The project home page contains the installation instructions and several materials (including videos, examples, comparative tables) that facilitate the use of GelJ.

Acknowledgements

This work was partially supported by the Ministerio de Economía y Competitividad [MTM2014-54151-P].

References

[1] A. Kumar, P. Misra, and A. Dube. Amplified fragment length polymorphism: an adept technique for genome mapping, genetic differentiation, and intraspecific variation in protozoan parasites. *Parasitology Research*, 112(2):457–466, 2013.

[2] C. Mba and J. Tohme. Use of AFLP Markers in surveys of plant diversity. *Methods in Enzymology*, 395:177–201, 2005.

[3] A. Bonin, D. Ehrich, and S. Manel. Statistical analysis of amplified fragment length polymorphism data: a toolbox for molecular ecologists and evolutionists. *Molecular Ecology*, 16(18):3737–3758, 2007.

[4] L. Excoffier and G. Heckel. Computer programs for population genetics data analysis: a survival guide. *Nature Reviews Genetics*, 7(10):745–58, 2006.

[5] J. Heras et al. A Survey of Tools for Analysing DNA Fingerprints. *Briefings in Bioinformatics*, 2015. doi: 10.1093/bib/bbv016.

[6] G.A. Pavlopoulos et al. A reference guide for tree analysis and visualization. *BioData Mining*, 3(1), 2010.

[7] G. Olsen. "Newick's 8:45" Tree Format Standard, 1990. `http://evolution.genetics.washington.edu/phylip/newick_doc.html`.

[8] D.R. Maddison, D.L. Swofford, and W.P. Maddison. NEXUS: an extensible file format for systematic information. *Systematic Biology*, 46(4): 590–621, 1997.

[9] J. Heras et al. GelJ — a tool for analyzing DNA fingerprint gel images. *BMC in Bioinformatics*, 16(270), 2015.

[10] B. Evjen et al. *Professional XML*. Wiley Publishing Inc, 2007.

[11] L. Vauterin and P. Vauterin. *Molecular Identification, Systematics, and Population Structure of Prokaryotes*, chapter Integrated Databasing and Analysis, pages 141–217. 2006.

[12] J. Heras et al. Surveying and Benchmarking Techniques to Analyse DNA Gel Fingerprint Images. *Briefings in Bioinfomatics*, 2015. doi: 10.1093/bib/bbv102.

[13] M. Hall, E. Frank, and G. Holmes. The WEKA Data Mining Software: An Update. *SICKDD Explorations*, 11(1):10–18, 2009.

[14] Applied Maths NV. GelCompar II version 6.6.11, 2013. `http://www.applied-maths.com` (25 May 2016, date last accessed).

[15] SequentiX. Gelquest version 3.2.1, 2010. http://www.sequentix.de/gelquest/index.php (25 May 2016, date last accessed).

[16] GE Healthcare Life Sciences. Imagequant tl 7.0, 2010. http://www.gelifesciences.com/ (25 May 2016, date last accessed).

[17] A.B. Pavel and C.I. Vasile. PyElph — a software tool for gel images analysis and phylogenetics. *BMC Bioinformatics*, 13(9), 2012.

[18] TotalLab Limited. Phoretix 1d pro version 11.4, 2013. http://www.totallab.com/products/1dpro/ (25 May 2016, date last accessed).

[19] A. Khadivi-Khub, N. Shabanian, and L. Alikhani. Genotypic analysis and population structure of Lebanon oak (Quercus libani G. Olivier) with molecular markers. *Tree Genetics & Genomes*, 11(102), 2015.

[20] Y. Minn, O. Gailing, and R. Finkeldey. Genetic diversity and structure of teak (Tectona grandis L. f.) and dahat (Tectona hamiltoniana Wall.) based on chloroplast microsatellites and Amplified Fragment Length Polymorphism markers. *Genetic Resources and Crop Evolution*, pages 1–14, 2015.

[21] Mrinalini et al. Convergence of multiple markers and analysis methods defines the genetic distinctiveness of cryptic pitvipers. *Molecular phylogenetics and evolution*, 92:266–279, 2015.

[22] D. Ehrich. AFLPdat: a collection of R functions for convenient handling of AFLP data. *Molecular Ecology Notes*, 6, 2006.

[23] J.C. Glaubitz. Convert: A user-friendly program to reformat diploid genotypic data for commonly used population genetic software packages. *Molecular Ecology Notes*, 4:309–310, 2004.

[24] J.A. Coombs, B.H. Letcher, and K.H. Nislow. Create: a software to create input files from diploid genotypic data for 52 genetic software programs. *Molecular Ecology Resources*, 8(3):578–580, 2008.

[25] P.M. Schlüter and S.A. Harris. Analysis of multilocus fingerprinting data sets containing missing data. *Molecular Ecology Notes*, 6, 2006.

[26] N.C. Manoukis. FORMATOMATIC: A program for converting diploid allelic data between common formats for population genetic analysis. *Molecular Ecology Notes*, 7(4):592–593, 2007.

[27] R. Peakall and P.E. Smouse. GenAlEx 6.5: genetic analysis in Excel. Population Genetic Software for Teaching and Research — an update. *Bioinformatics*, 28(19):2537–2539, 2012.

[28] H.E.L. Lischer and L. Excoffier. PGDSpider: An automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics*, 28, 2012.

[29] J. Caujapé-Castells, I. Sabbagh, and J.J. Castellano. Transformer-4 version 2.0.1, a free multi-platform software to quickly reformat genotype matrices of any marker type, and archive them in the Demiurge information system. *Molecular Ecology Resources*, 13(3):484–493, 2013.

[30] CIMI Executive Committee. Clinical Information Modeling Initiative, 2014. `http://www.opencimi.org/`.

[31] W. Goosen, A. Goosen-Baremans, and M. van der Zel. Detailed Clinical Models: A Review. *Healthcare Research Information*, 16(4):201–214, 2010.

[32] M. R. Forman et al. Bioinformatics: Tools to accelerate population science and disease control research. *American Journal of Preventive Medicine*, 38(6):646–651, 2010.

[33] S. A. Chervitz et al. Data Standards for Omics Data: The Basis of Data Sharing and Reuse. *Methods in Molecular Biology*, 719:31–69, 2011.