BIOINFORMATICS APPLICATIONS NOTE

Vol. 28 no. 2 2012, pages 298–299
doi:10.1093/bioinformatics/btr642

Data and text mining

Advance Access publication November 21, 2011

PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs

H. E. L. Lischer[1,2,*] and L. Excoffier[1,2]

[1]Computational and Molecular Population Genetics (CMPG) laboratory, Institute of Ecology and Evolution, University of Berne, Baltzerstrasse 6, 3012 Berne and [2]Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland

Associate Editor: Janet Kelso

ABSTRACT

Summary: The analysis of genetic data often requires a combination of several approaches using different and sometimes incompatible programs. In order to facilitate data exchange and file conversions between population genetics programs, we introduce PGDSpider, a Java program that can read 27 different file formats and export data into 29, partially overlapping, other file formats. The PGDSpider package includes both an intuitive graphical user interface and a command-line version allowing its integration in complex data analysis pipelines.

Availability: PGDSpider is freely available under the BSD 3-Clause license on http://cmpg.unibe.ch/software/PGDSpider/

Contact: heidi.lischer@iee.unibe.ch

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on August 4, 2011; revised on November 14, 2011; accepted on November 15, 2011

1 INTRODUCTION

In the last few years, high-throughput sequencing made it possible to sequence a huge amount of DNA at a rapidly decreasing cost, allowing for budget-tight studies to produce large sequencing and population genetics datasets (Gompert *et al.*, 2010; Hohenlohe *et al.*, 2010; Metzker, 2010). A proper analysis of these data requires a combination of several approaches, beginning with those describing basic properties of the data, followed by more specialized and often computer intensive analyses (Excoffier and Heckel, 2006). Most of the times, it requires the use of several population genetic programs and different input formats, which necessitate numerous file conversions. However, data format conversion can be tedious and error prone if not fully automatized, especially when dealing with large amounts of data.

To our knowledge, only three programs are available to transfer populations genetics data between programs: CONVERT (Glaubitz, 2004), FORMATOMATIC (Manoukis, 2007) and CREATE (Coombs *et al.*, 2008). Whereas these programs can sometimes bridge many population genetics program (CREATE reads 12 input formats and outputs 62 formats; CONVERT reads 2 and outputs 7, whereas FORMATOMATICS reads 4 and outputs 13), these programs have some limitations. For instance, CREATE and CONVERT only handle codominant diploid genotypic data, CREATE focuses mainly on parentage analysis programs and FORMATOMATICS only deals with microsatellite data.

To improve data exchange between programs for a vast range of data types (Fig. 1), we introduce here PGDSpider, an automated data conversion tool for population genetics and population genomics.

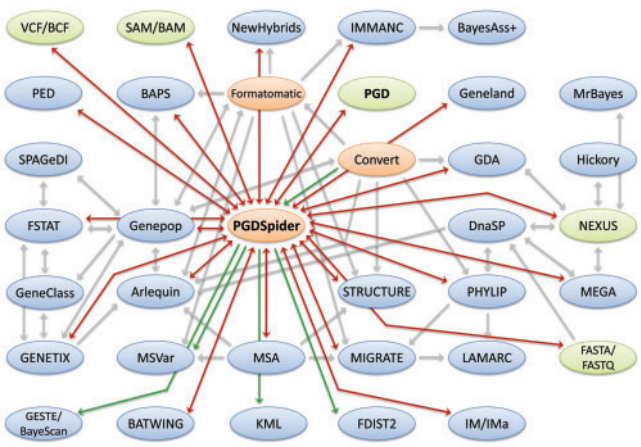

Fig. 1. Connectivity between population genetics programs and formats. Red (reading and writing) and green (reading or writing) arrows indicate conversion possibilities between PGDSpider and other programs. Grey arrows show connections between programs that are not mediated by PGDSpider. Programs are shown as blue ellipses, general data formats in green and conversion programs in orange.

2 SUPPORTED DATA FORMATS

Our data conversion program PGDSpider is able to deal with diverse types of data like DNA, RNA, NGS (next-generation sequencing data), microsatellite, SNP, RFLP, AFLP, multiallelic data, allele frequency and genetic distances, and this either in diploid or haploid formats. Currently, on top of its own Population Genetics Data (PGD) format, PGDSpider can read 27 and export 29 different data formats, like: e.g. ARLEQUIN (Excoffier and Lischer, 2010) or MEGA (Tamura *et al.*, 2011), but see the complete list of supported formats in the Supplementary Table S1. With the ability to convert various kinds of data types and data formats, PGDSpider greatly enhances and complements exchange possibilities between population genetics programs as shown in Figure 1.

In addition to conventional population genetics formats, PGDSpider integrates four population genomics data formats

*To whom correspondence should be addressed.

commonly used for storing and handling NGS data like those of the 1000 genomes project (Altshuler *et al.*, 2010): these are the sequencing alignment (SAM) format and the variant call format (VCF), as well as their binary versions (BAM and BCF, respectively). Note that PGDSpider calls the program SAMtools (Li *et al.*, 2009) during the conversion of these formats. SAMtools thus needs to be downloaded separately. Currently, PGDSpider is not meant to convert very large NGS files as it loads into memory the whole input file, the size of which may exceed available RAM. However, since PGDSpider can convert specific subsets of large BAM, SAM, BCF and VCF files into any other format, one could use this feature to calculate parameters or statistics for specific chromosome segments, and thus use a sliding window approach to analyse large genomic regions.

3 PGD FORMAT

PGDSpider uses a newly developed PGD format as an intermediate step in the conversion process. PGD is a file format designed to store various kinds of population genetics data, including different data types (e.g. DNA sequences, microsatellites, AFLP or SNPs) and ploidy levels (haploid, diploid, or higher ploidy levels). PGD is based on the XML (eXtensible Markup Language) format and is therefore independent of any particular computer system and extensible for future needs. The XML file uses an ordered labelled tree-like structure that can easily be processed. PGD is structured in the following blocks: (i) the 'header' block contains a general description of the data. It is possible to give information about the organism, the number of populations, the ploidy level, the characters encoding missing values and gaps. Additionally, it allows one to specify if the gametic phase is known and if the data are recessive. The second block (ii) 'dataDescription' embeds locus-specific information and makes it possible to store the following information for each locus: data type, location in the genome, if it is in genic regions or not, its length and some additional comments or links. (iii) The 'population' blocks contain information about the populations, their member individuals and the actual genetic data. It allows one to specify the geographic coordinates of the population and/or the individuals, as well as the ploidy level, locus names or identifiers and multilocus haplotype frequencies. Note that for NGS data, it is also possible to store sequence read quality scores. (iv) An optional 'structure' block gives information about the genetic structure of the populations (i.e. which population samples belong to which groups). The last block (v) 'distanceMatrix' is optional and contains information about genetic distances among haplotypes.

Full details on the PGD format can be found in the PGDSpider user manual together with example files. The block structure and the hierarchical structure of stored information make the PGD format very modular and extensible for future needs. Additionally, it is possible to nicely visualize data contained in the PGD files by using any web browser supporting XML stylesheets. The PGD format could be seen as a first step towards the creation of a generic and unified population genetic data format, which could greatly improve communication between programs and thus facilitate data analyses.

4 PROGRAM INTERFACE AND AVAILABILITY

The PGDSpider program is written in Java, a fully virtualized and platform-independent programming language. The program is user friendly in the sense that it provides an intuitive graphical user interface (GUI). The GUI and the menus are available in four different languages: English, French, German and Italian. PGDSpider allows users to store their preferred conversion settings for repeated conversions of similar input formats. A command-line version of PGDSpider is also provided, making it possible to embed PGDSpider in complex data analysis pipelines.

The PGDSpider program and its user manual can be downloaded from http://cmpg.unibe.ch/software/PGDSpider/ under the BSD 3-Clause License. The user manual provides detailed information on how to use the program (GUI and command-line version) as well as a short file description of each supported data formats. It also includes an extensive description of the new PGD format and example files.

ACKNOWLEDGEMENTS

We thank Matthieu Foll, Mathias Beysard and the whole CMPG team for comments and testing the program, as well as Pascal Tschanz for his help in solving programming issues.

Funding: Swiss National Science Foundation (grant No 3100A0-126074 to L.E.).

Conflict of Interest: none declared.

REFERENCES

Altshuler,D.L. *et al.* (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.

Coombs,J.A. *et al.* (2008) CREATE: a software to create input files from diploid genotypic data for 52 genetic software programs. *Mol. Ecol. Resour.*, **8**, 578–580.

Excoffier,L. and Heckel,G. (2006) Computer programs for population genetics data analysis: a survival guide. *Nat. Rev. Genet.*, **7**, 745–758.

Excoffier,L. and Lischer,H.E.L. (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.*, **10**, 564–567.

Glaubitz,J.C. (2004) CONVERT: a user-friendly program to reformat diploid genotypic data for commonly used population genetic software packages. *Mol. Ecol. Notes*, **4**, 309–310.

Gompert,Z. *et al.* (2010) Bayesian analysis of molecular variance in pyrosequences quantifies population genetic structure across the genome of Lycaeides butterflies. *Mol. Ecol.*, **19**, 2455–2473.

Hohenlohe,P.A. *et al.* (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *Plos Genet.*, **6**, e1000862.

Li,H. *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Manoukis,N.C. (2007) FORMATOMATIC: a program for converting diploid allelic data between common formats for population genetic analysis. *Mol. Ecol. Notes*, **7**, 592–593.

Metzker,M.L. (2010) Applications of next-generation sequencing sequencing technologies - the next generation. *Nat. Rev. Genet.*, **11**, 31–46.

Tamura,K. *et al.* (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.*, **28**, 2731–2739.