MOLEKBASE: USER FRIENDLY SYSTEM FOR STORING, FILTERING AND CONVERTING POPULATION MOLECULAR DATA

MOLEKBASE: UPORABNIKU PRIJAZEN SYSTEM ZA HRANITEV, IZBIRO IN PRETVORBO MOLEKULSKIH PODATKOV V POPULACIJSKI GENETIKI

Marjana WESTERGREN[1] & Hojka KRAIGHER[2]

ABSTRACT UDC 575.17:004.4

MOLEKBASE: user friendly system for storing, filtering and converting population molecular data

Molecular experimental data for population genetics is often stored in spreadsheet programmes or as input data for computer programmes that enable analysis of population genetics. Such experimental data can often be interpreted only by the researcher who conducted the experiment, diminishing the transparency of the whole study. Additionally, same data can be stored at several locations. Making changes to the data in a single location generates inconsistencies in the dataset. A database layout in Access was developed to facilitate transparent population genetic data storage in a single location and simplify its use for population genetic analysis through a computer programme that enables filtering of the data and transforms it into Genepop, SpaGeDi, Structure, Baps and Convert input files. The MOLEKBASE system is freely available at http://www.gozdis.si/index.php?id=151.

Keywords: population genetics, molecular database, data filtering, data conversion

IZVLEČEK UDK 575.17:004.4

MOLEKBASE: uporabniku prijazen system za hranitev, izbiro in pretvorbo molekulskih podatkov v populacijski genetiki

Molekulski podatki za genetske analize populacij so pogosto shranjeni v obliki razpredelnic ali kot vhodni podatki za programe, ki omogočajo njihovo analizo. Take podatke lahko pogosto interpretira le raziskovalec, ki je poskus izvajal, kar vodi k manjši transparentnosti celotne raziskave. Pogosto se tako shranjeni podatki nahajajo na več lokacijah. Sprememba v podatkih na eni lokaciji vnese v set podatkov nedoslednosti. Razvili smo matrico baze za transparentno hranitev populacijskih molekulskih podatkov na enotni lokaciji v programu Access in pripravili program za izbiro ter pretvorbo podatkov v format, ki ga prepoznajo programi za analize v okviru populacijske genetike Genepop, SpaGeDi, Structure, Baps in Convert. Novo razviti sistem MOLEKBASE je prosto dostopen na http://www.gozdis.si/index.php?id=151.

Ključne besede: populacijska genetika, baza molekulskih podatkov, izbira podatkov, pretvorba podatkov

[1] Dr., Department of Forest Physiology and Genetics, Slovenian Forestry Institute, Večna pot 2, 1000 Ljubljana, marjana.westergren@gozdis.si

[2] Prof. Dr., Department of Forest Physiology and Genetics, Slovenian Forestry Institute, Večna pot 2, 1000 Ljubljana, hojka.kraigher@gozdis.si

INTRODUCTION

The analysis of population genetics requires vast data sets. Hundreds of individuals belonging to different populations or species are analysed on as few as five co-dominant loci in population studies of forest trees (e.g. Heuertz et al. 2003; Fernandez-Manjarres et al. 2006; Heuertz et al. 2004) and up to 377 co-dominant loci in human population studies (Rosenberg et al. 2002). In population genetic analysis of forest trees, microsatellites and isozymes are the markers of choice and the datasets usually consist of a low to medium number of loci, e.g. five to 15. A small analysis of four populations with 50 samples in each population would therefore yield 2000 to 6000 data points for co-dominant markers.

Molecular experimental data for population genetics is usually stored in tables of spreadsheet programmes such as Excel or as input data for a variety of computer programmes that enable analysis of population genetics. This can lead to the same data being stored at several locations. Changing the data at only one location will therefore generate inconsistencies in the dataset. Additionally such experimental data can often be interpreted only by the researcher who conducted the experiment, diminishing the transparency of the whole study. Re-analysing the data and combining different studies, especially if some time has passed or the personnel in the laboratory have changed, is difficult. In order to overcome the above-mentioned problems we have developed a database layout in Access, in which data from population genetic studies can be stored in a single place. Individuals or populations needed for specific analysis can be filtered out and selected data transformed into some of the most common freely available population genetic programme input formats without making changes to the original data set. The system was developed to help us manage vast datasets of population genetic data needed for the analysis of forest genetic resources but could be useful in other fields.

MATERIALS AND METHODS

Review of population genetic studies of forest trees has shown that microsatellites and isozymes are the markers of choice for population genetic analysis of forest trees. The datasets usually consist of a low number of loci for microsatellites to medium number of loci for isoenzymes.

Access was used to develop the layout of the database. The layout allows addition of other needed categories (i.e. columns) if needed by the user. The data filtering and conversion programme was written in MS Visual Studio 2005 vb.net.

The MOLEKBASE system (database layout in Access, Windows executable file and the source code), including the user manual and example files, can be freely downloaded from http://www.gozdis.si/index.php?id=151.

RESULTS

Experimental data and background information in the MOLEKBASE system are stored in three different tables: Molecular data, Population and Locus. The first table contains molecular data in relative sizes or codes in a three-digit format for up to 25 co-dominant loci and information regarding individual samples, such as sample code, population code, species, laboratory code, as well as year of analysis. In the second table, information regarding sampled populations is stored. This table contains population codes, geographic location in latitude/longitude format and/or UTM coordinates and altitude. Other fields describing individual samples and/or populations can be added after the predefined fields. For forestry purposes, these might be vitality, developmental stage, origin of populations, seed stand identifiers etc. In the last table, the names and number of loci belonging to each species and/or experiment are stored.

Currently, the database layout supports data storage and manipulation for up to 25 co-dominant diploid loci, which, according to a survey of the literature is sufficient for most population genetic studies in the forestry field. The database layout was primarily developed for microsatellites but can store and manipulate any co-dominant data in three-digit format.

With the help of scripts, samples of interest for a certain analysis can be selected and transformed into five different input formats. The programme enables se-

lection of data based on country of origin, species, population, sampling year and individuals. Boolean operators are used to combine different filters. Selected data can be transformed into five different input formats, read by the following population genetics programmes: Genepop (Raymond & Rousset 1995, Rousset 2008), SpaGeDi (Hardy & Vekemans 2002), Structure (Pritchard, Stephens & Donnelly 2000), Baps (Corander , Waldmann & Sillanpaa 2003) and Convert (Glaubitz 2004).

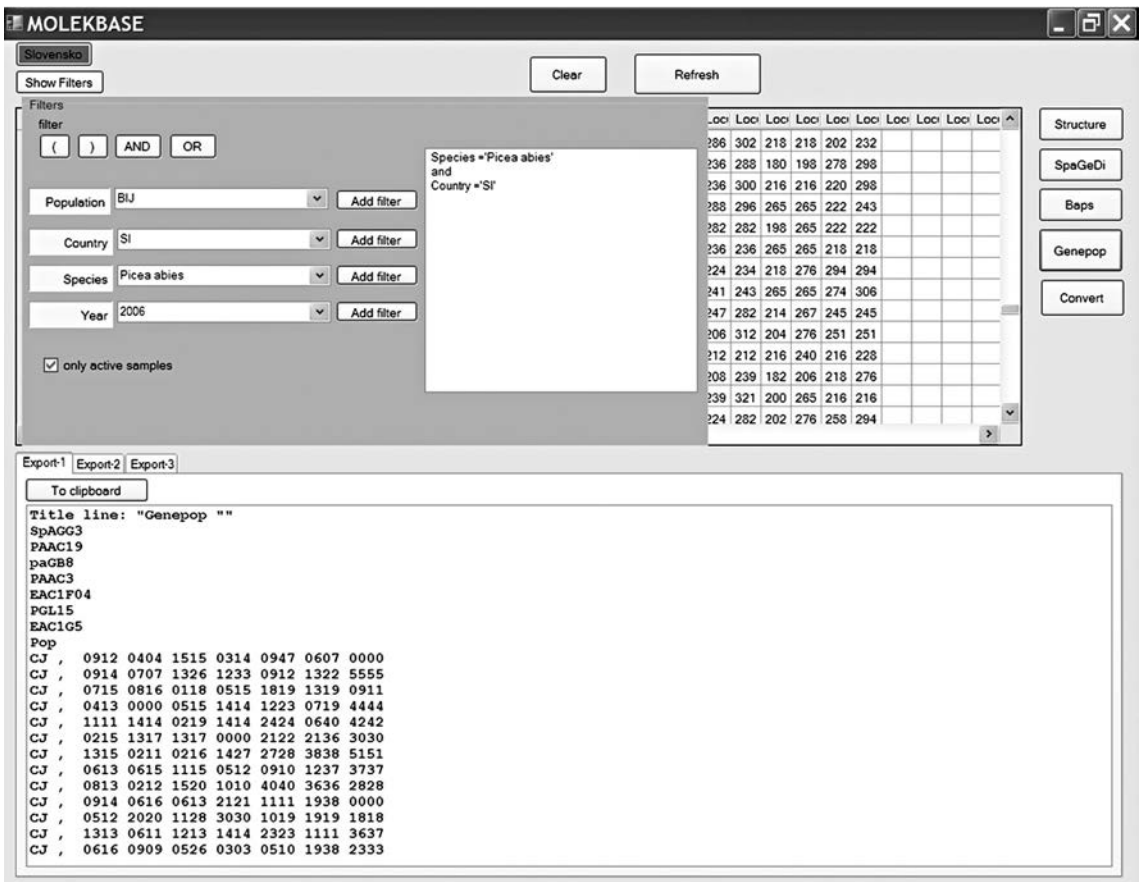

Figure 1: Data filtering and conversion form

CONCLUSION

MOLEKBASE is a database layout in Access with an accompanying computer programme that facilitates transparent molecular data storage for population genetic analysis in a single location and its use by filtering and converting selected molecular data into five different input formats.

The MOLEKBASE system including the user manual and example files, can be freely downloaded from http://www.gozdis.si/index.php?id=151.

POVZETEK

Raziskave v okviru populacijske genetike zahtevajo velike količine podatkov. Genski označevalci, ki jih uporabljamo pri populacijsko genetskih analizah dreves so največkrat mikrosateliti ali izoencimi; posamezna drevesa

pa analiziramo na majhnem do srednjem številu lokusov (število analiziranih lokusov se največkrat giblje med pet in 15, kar pri majhni analizi štirih populacij s 50. vzorci na populacijo pomeni med 2000 in 6000 podatkov). Molekulski podatki za genetske analize populacij so pogosto shranjeni v obliki razpredelnic ali kot vhodni podatki za programe, ki omogočajo njihovo analizo. Take podatke lahko največkrat interpretira le raziskovalec, ki je poskus izvajal, kar vodi k manjši transparentnosti celotne raziskave. Pogosto se tako shranjeni podatki nahajajo na več lokacijah. Sprememba v podatkih na eni lokaciji vnese v set podatkov nedoslednosti. Ponovna analiza ali pa združevanje večjega števila raziskav, posebej če je od originalne analize minilo nekaj časa ali pa se je zamenjalo osebje v laboratoriju, je praviloma otežena. Zato smo razvili matrico baze za transparentno hranitev populacijskih molekulskih podatkov na enotni lokaciji v Accessu in program za izbiro podatkov ter njihovo pretvorbo v pet različnih formatov v MS Visual Studio 2005 vb.net.

Eksperimentalni podatki in ostale informacije v sistemu MOLEKBASE so shranjene v treh različnih tabelah. V tabeli »Molecular data« so molekulski podatki v obliki tri-številnih kod ali relativnih dolžin za do največ 25 ko-dominantnih lokusov ter podatki, vezani na vsak vzorec/analiziran osebek. V tabeli »Population« so podatki, ki se navezujejo na analizirano populacijo, v tabeli »Locus«so shranjena imena in število analiziranih lokusov za vsako vrsto in/ali eksperiment. Sistem dopušča dodajanje novih polj na željo uporabnika. S pomočjo skript lahko uporabnik na podlagi države izvora, biološke vrste, populacije, leta vzorčenja ali posameznikov izbere podatke za določeno analizo ter jih pretvori v pet različnih formatov, ki jih prepoznajo programi za obdelavo genetsko populacijskih podatkov Genepop, SpaGeDi, Structure, Baps in Convert.

MOLEKBASE je vključno z navodili za uporabo in testnimi podatki prosto dostopen na http://www.gozdis.si/index.php?id=151.

ACKNOWLEDGEMENTS

The work was supported by the Slovenian Ministry of Higher Education, Science and Technology through the Slovenian Research Agency: the Young Researchers scheme grant no. 3331-03-831659 and the research programme P4-0107.

REFERENCES

Corander, J., P.Waldmann & M.J.Sillanpaa, 2003: *Bayesian analysis of genetic differentiation between populations*. Genetics (Austin, Texas) 163:367-374

Fernandez-Manjarres, J., P.Gerard, J.Dufour, C.Raquin & N. Frascaria-Lacoste, 2006: *Differential patterns of morphological and molecular hybridization between Fraxinus excelsior L. and Fraxinus angustifolia Vahl (Oleaceae) in eastern and western France*. Mol Ecol (Oxford, Velika Britanija) 15:3245–3257

Glaubitz, J.C., 2004: *Convert: A user-friendly program to reformat diploid genotypic data for commonly used population genetic software packages*. Mol Ecol Notes (Oxford, Velika Britanija) 4 (2):309-310

Hardy, O.J. & X. Vekemans, 2002: *SPAGEDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels*. Mol Ecol Notes (Oxford, Velika Britanija) 2:618-620

Heuertz, M., J.F. Hausman, O.J. Hardy, G.G. Vendramin, N. Frascaria-Lacoste & X. Vekemans, 2004: *Nuclear microsatellites reveal contrasting patterns of genetic structure between western and southeastern European populations of the common ash* (*Fraxinus excelsior L.*). Evolution (Lancaster, Pennsylvania) 58 (5):976-988

Heuertz, M., X. Vekemans, J.F. Hausman, M. Palada & O.J. Hardy, 2003: *Estimating seed vs. Pollen dispersal from spatial genetic structure in the common ash*. Mol Ecol (Oxford, Velika Britanija) 12:2483–2495

Pritchard, J.K., M. Stephens & P. Donnelly, 2000: *Inference of population structure using multilocus genotype data*. Genetics (Austin, Texas) 155:945-959

Raymond, M. & F. Rousset, 1995: *Genepop (version-1.2) - population-genetics software for exact tests and ecumenicism*. J Hered (Washington, D.C.) 86 (3):248-249

Rosenberg, N.A., J.K. Pritchard, J.L. Weber, H.M. Cann, K.K. Kidd, L.A. Zhivotovsky & M.W. Feldman, 2002: *The genetic structure of human populations*. Science (New York) 298 (5602):2381-2385

Rousset, F., 2008: *Genepop'007: A complete re-implementation of the genepop software for windows and linux*. Mol Ecol Resour (Oxford, Velika Britanija) 8 (1):103-106