THE PIG GENETIC DIVERSITY DATABASE AND THE WWW

LA BASE DE DATOS SOBRE DIVERSIDAD GENÉTICA PORCINA Y LA WWW

Russell, G.A., A.L. Archibald, C.S. Haley and A.S. Law

Roslin Institute. Roslin. EH25 9PS. UK.
Contact adress: andy.law@bbsrc.ac.uk

SUMMARY

The pig genetic diversity database (PigDBase) and its associated web site (http://www.projects.roslin.ac.uk/pigbiodiv/) provide secure storage of data from studies of genetic diversity in pigs including microsatellite and AFLP genotypes. The design of the database is such that it could be applied to any species and, with appropriate standardisation protocols, could extend to any laboratory wishing to contribute data. Access to PigDBase is effected via the World Wide Web using the diversity browser. The information stored includes details of microsatellite markers, breeds, sample and contact details. Full information on unprocessed data, as well as summary information on genetic measurements and export formats are also available.

RESUMEN

La base de datos sobre la diversidad genética porcina (PigDbase) y su sitio Web asociado (http://www.projects.roslin.ac.uk/pigbiodiv/) provee de un almacén de datos seguro para los estudios sobre la diversidad genética de los cerdos incluyendo los genotipos de microsatéli-tes y AFLPs. El diseño de la base de datos es tal que podría ser aplicado a cualquier especie y, con adecuados protocolos de estandarización, podría extenderse a cualquier laboratorio que deseara aportar datos. El acceso a PigDbase se efectúa vía WWW utilizando el browser diversi-dad. La información almacenada incluye detalles de los marcadores microsatélites, muestra y detalles de contacto. La información completa sobre datos no procesados, así como informa-ción resumida sobre medidas genéticas y formatos de exportación también están disponibles.

INTRODUCTION

Directed selection of economically desirable traits can lead to genetic erosion of individual breeds of pigs and affect the whole pig population. This reduced genetic base could limit opportunities to select production characteristics that may be vital in the future. It is therefore important that genetic diversity in any species is maintained to provide a base from which new characteristics can be selected.

Polymorphic molecular genetic markers provide a means to assess within and between breed genetic diversity. In particular, microsatellite markers have been used to characterise a wide variety of breeds and species of farm animals (e.g. Cañon *et al.*, 2000; Hanslik *et al.*, 2000; Laval *et al.*, 2000; Ritz *et al.*, 2000; Wimmers *et al.*, 2000) and are currently recommended as the marker of choice for studies of domestic animal diversity (FAO, 1998; http://dad.fao.org/en/Home.htm). Amplified fragment length polymorphisms (AFLPs) have also been used in studies of genetic diversity, including the EC pig genetic diversity project. However, it is likely that single nucleotide polymorphisms (SNPs) will eventually replace microsatellites as the marker of choice for studies of genetic diversity as the number of characterised SNPs for farm animal species increases and the cost of their typing decreases.

Regardless of the type of molecular genetic markers used to evaluate genetic diversity, the sharing of data is essential. As it is necessary to consider genetic diversity on a global scale, research groups that study their local populations can make a valuable contribution either acting alone or preferably as part of wider collaborative network. The value of small studies of indigenous breeds is greatly enhanced if the data can be combined across studies. Indeed, it is arguable that if the data from small studies cannot be integrated into global analyses now or at some future date then the studies are of minimal value.

The establishment of a database that holds such diversity data would allow ready access to information by those interested in the maintenance of genetic diversity in pigs. This information could be used in selection programmes aimed at improving the quality of pig products, welfare and health through the extensification of farming systems.

To address this need, a database (PigDBase) has been designed and built to hold data generated by the EC-funded Pig Diversity Project. The database stores details of sample information and genotyping data as well as information on the genetic markers used. Full information on unprocessed data, as well as summary information on genetic measurements has been made available to the wider community through a World Wide Web interface. This provides immediate access to the stored data and allows analyses to be performed on exported data.

EC PIG DIVERSITY PROJECT WEB SITE

As part of the EC-funded project entitled *Characterisation of genetic variation in the European pig to facilitate the maintenance and exploitation of biodiversity (Pig-BioDiv)*, a web site (http://www.projects.roslin.ac.uk/pigbiodiv/) has been established to act as a focal point for communication. Project progress is disseminated via meeting reports and other documents published on the site. It also includes general information on the project background and approach, details of project participants, breeds used in the project and the microsatellite markers chosen. The

web site provides a public interface to the project and a source of information on genetic diversity studies in pigs and links to other resources.

Thus, the Web site acts as a portal to information on pig diversity generated within the project and to other sources of information including the Food and Agriculture Organisation (FAO) (http://www.fao.org/dad-is/) and the European Association for Animal Production (EAAP) animal genetic bank (http://www.tiho-hannover.de/einricht/zucht/eaap/index.htm). The web site also provides access to the pig diversity database.

DATABASE OVERVIEW

Data sharing is essential in order to maximise the value of genetic diversity data. A genetic diversity database is critical to effective data sharing. The requirements for a genetic diversity database include the maintaining of the integrity of the data, making the data accessible, enabling different views of the data and providing tools to export the data in the formats for a range of analysis programs.

Based on our experience of developing resource and genome databases to support genome research in farm animals (Law and Archibald, 2000; Hu *et al*., 2001) we elected to use a robust and established relational database management system (RDBMS) - CA-INGRES - as the data store. This is primarily to ensure data integrity. The database is automatically *checkpointed* on a regular basis and all changes are tracked using the built-in journaling tools. Thus the system can

be restored relatively simply to any given state should the need arise. In addition, standard operating procedures have been implemented to fully document changes made to the database system.

The requirement to make the data accessible to a geographically scattered network of collaborators and users has been addressed by mounting the database on the World Wide Web. The WWW interfaces use CGI scripts implemented in Webintool – a PHP-like scripting language that was developed at Roslin (Hu *et al*., 1996) – and Perl.

Although a Web forms interface to the database has been provided, access to those forms and the data within the system (including genotypes, export and analysis applications) is currently password-protected and restricted to collaborators within the EC pig diversity project. However, details of contact information, microsatellite markers, and data from the PiGMaP pilot diversity project (Laval *et al*., 2000) are accessible to the general public.

DATABASE CONTENTS

Information in the database is held in forty-six linked tables. The database contains breed and line information for animals sampled from 60 distinct populations as well as sex, date of birth, sire name and dam name (where available). Sample information also includes details of the amount and concentration of DNA[1] held in two

---

[1]Protocols for access to these DNA samples are currently being developed.

secure long-term repositories at Roslin Institute (Edinburgh) and INRA Laboratoire de Génétique Cellulaire, Toulouse. Breed information is linked to the EAAP animal genetic databank, which contains phenotype information for the majority of European breeds (http://www.tiho-hannover.de/einricht/zucht/eaap/index.htm).

In addition, the database carries marker details for 50 microsatellite and 148 AFLP markers. For each marker, information is available concerning the polymerase chain reaction (PCR) primers, chromosomal location of the marker and citations of the papers describing the markers. There are 467536 individual genotypings recorded – 118188 are from microsatellite markers while 349348 represent AFLPs. The database also contains a subset of data generated during the PiGMaP project (http://www.projects.roslin.ac.uk/pigmap/pigmap.html), which underpin the analysis by Laval *et al*. (2000). These data consist of genotypes for 26 microsatellite markers (a subset of the 50 markers used in the PigBioDiv project) typed across a sample of about

Figure 1a. *Screen shots of the menu from which a genotyping export routine is initiated.* (Imagen del menú desde la cual la rutina de exportación del genotipado se inicia).

50 animals from each of 11 breeds.

DATABASE INTERFACE FUNCTIONALITY

The database stores the raw genotype information for each marker-sample combination as reported by the laboratories performing the genotyping. Access to the raw genotype information is provided via a series of forms **(figure1a and 1b)**. The user is initially prompted to select the line or lines of interest and the types of markers

(currently AFLP or microsatellite) required. The following screen presents the user with a list of markers of the selected type(s) for which genotypings are available in the selected lines. Having selected markers from the list and clicked on the export button, the required genotypes are exported to the screen in text format for download to the user's local computer. Microsatellite genotypings are reported as labels representing the approximate size of the PCR-amplified fragment; AFLP typings are recorded as *B* (band present) or *D* (band absent).

Figure 1b. *An example of the resulting screen dump of the data*. (Un ejemplo de la pantalla resultante con los datos).

In addition to the ability to export raw genotype data, the interface also allows the user to export the data in a number of other formats, notably in summary allele frequency format and in the format required for the PHYLIP analysis package (http://evolution.genetics.washington.edu/phylip.html). Again, the export routines make use of a series of forms to guide the user

Table I. *Information resources and databases mounted on the World Wide Web.* (Recursos de información y bases de datos montados sobre la WWW).

NAME	DESCRIPTION	URL
Food and Agriculture Organisation, Domestic Animal Diversity Information Service (DAD-IS)	DAD-IS is the information system for the Global Strategy for the Management of Farm Animal Genetic Resources (AnGR) – Web site and database	http://dad.fao.org/en/Home.htm
EAAP Animal Genetic Databank	The data bank includes information on breeds and country populations of the farm animal species (ass, buffalo, cattle, goat, horse, pig, sheep) in 45 EAAP member countries and other European countries	http://www.tiho-hannover.de/einricht/zucht/eaap/index.htm
EU Sheep & Goat Biodiversity Project (ECONOGENE)	Web site with details of a European project concerned with developing and using a common set of markers to study diversity in all even-toed ungulates	http://lasig.epfl.ch/projets/econogene/index.html
INRA Laboratoire de Génétique cellulaire Web site	Details of FAO-ISAG recommended microsatellite markers for diversity studies in pigs	http://www.toulouse.inra.fr/lgc/pig/panel.htm
Cattle Diversity Database (CaDBase)	Information including details of markers, published allele frequencies	http://www.projects.roslin.ac.uk/cdiv/
EC pig diversity project	Details of pig diversity project and pig diversity database equivalent to CaDBase	http://www.projects.roslin.ac.uk/pigbiodiv/

toward the data they want.

Further views on the data are provided either on a marker-by-marker or breed/line basis. These include the expected (Ott, 1992) and observed heterozygosity for a given marker in a given breed/line and the Polymorphic Information Content (PIC) value (Botstein *et al.*, 1980).

Because the database stores the primary raw genotype data, other transformations and export formats can readily be added should the need arise.

CONTINUED REQUIREMENT FOR GENETIC DIVERSITY DATABASE

Some of the current Web-mounted resources and databases concerned with genetic diversity in farm animals are listed in **table I**. The FAO's Domestic Animal Diversity Information Service (DAD-IS) and the EAAP maintain databases holding descriptions of a range of domestic animal species and breeds. As yet, DAD-IS does not provide a repository for molecular genetic information. As described here the Pig Genetic Diversity database and an equivalent database for cattle genetic diversity information were developed by Roslin Institute with support from European Commission research grant funds.

For the immediate future the pig genetic diversity database and associated Web site will be supported through an EC-funded project (PigBioDiv2 – *Characterisation of genetic variation in the pig breeds of China and Europe to facilitate the maintenance and exploitation of biodiversity*) in which the characterisation of genetic diversity will be extended to include breeds in China.

CONCLUSIONS

The pig genetic diversity database (PigDBase) and the associated Web site meet the need for a secure repository for genetic diversity data. The database and associated tools could be readily implemented for a range of other species. The Web site will continue to serve as a portal to information relevant to studies of genetic diversity in pigs.

ACKNOWLEDGEMENTS

REFERENCES

Botstein, D., R.L. White, M. Skolnick and R.W. Davis. 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum Genet.,* 32: 314-31.

Cañon, J., M.L. Checa, C. Carleos, J.L. Vega-Pla, M. Vallejo and S. Dunner. 2000. The genetic structure of Spanish Celtic horse breeds inferred from microsatellite data. *Anim. Genet.*, 31: 39-48.

FAO. 1998. Secondary guidelines for development of national farm animal genetic resource management plans. Measurement of domestic animal diversity (MoDAD): Original working

group report, FAO 1998,ISBN 92-6-104166-0.

Hanslik, S., B. Harr, G. Brem and C. Schlötterer. 2000. Microsatellite analysis reveals substantial genetic differentiation between contemporary New World and Old World Holstein Friesian populaitons. *Anim. Genet.,* 31: 31-8.

Hu, J., D. Nicholson, C. Mungall, A.L. Hillyard and A.L. Archibald. 1996. WebinTool: A generic Web to database interface building tool. Proceedings of the 7th International Conference and Workshop on Database and Expert Systems (DEXA 96), Zurich, September 9-13, 1996, pp 285-290.

Hu, J., C. Mungall, A. Law, R. Papworth, J.P. Nelson, A. Brown, I. Simpson, S. Leckie, D.W. Burt, A.L. Hillyard and A.L. Archibald. 2001. The ARKdb – genome databases for farmed and other animals. *Nucl. Acids Res.,* 29: 106–10.

Laval, G., N. Iannuccelli, C. Legault, D. Milan, M.A.M. Groenen, E. Giuffra, L. Andersson, P.H. Nissen, C.B. Jørgensen, P. Beeckmann, H. Geldermann, J.L. Foulley, C. Chevalet and L. Ollivier. 2000. Genetic diversity in eleven European pig breeds. *Genet. Sel. Evol.,* 32: 187-203.

Law, A.S. and A.L. Archibald. 2000. Farm animal genome databases. *Briefings in Bioinformatics,* 1: 151-60.

Ott, J. 1992. Strategies for characterizing highly polymorphic markers in human gene mapping. *American Journal of Human Genetics,* 51: 283-290 .

Ritz, L.R., M.-L. Glowatzki-Mullis, D.E. MacHugh and C. Gaillard. 2000. Phylogenetic analysis of the tribe Bovini using microsatellites. *Anim. Genet.,* 31: 178-85.

Wimmers, K., S. Ponsuksili, T. Hardge Valle-Zarate, P.K. Mathur and P. Horst. 2000. Genetic distinctness of African, Asian and South American local chickens. *Anim. Genet.,* 31: 159-65.