遗 传 学 报 *Acta Genetica Sinica*, June 2003, 30 (6): 509～514 ISSN 0379 - 4172

GPCEG-A database for Genomic Polymorphism of Chinese Ethnic Groups

GHEN Xin[1], ZHANG Yong[1], WANG Jian-Min[1], HUANG Yi[1], LUO Jing-Chu[1], WU Cai-Hong[2], GU Xiao-Cheng[1,①]

(1. *College of Life Science, Peking University, Beijing* 100871, *China*;
2. *State Key Laboratory of Biomembranes and Membrane Biotechnology, Peking University, Beijing* 100871, *China*)

Abstract: This paper reports the construction of the database for Genomic Polymorphism of Chinese Ethnic Groups(GPCEG). GPCEG contains denomination and basic information of Chinese 56 ethnic groups, with introduction of their in geographic distribution, population quantity, spoken and written language, religious belief and physical characteristics. GPCEG collects the data of genomic polymorphism, cell lines, reference and links of other international related databases. The visualization, query and update system were also available. GPCEG laid the foundations of establishing a national database with Chinese characteristics.

Key words: bioinformatics; biological database; Chinese Ethnic Groups; genomic polymorphism

中华民族基因组多态性数据库的构建

陈 新[1], 张 勇[1], 王建民[1], 黄 弋[1], 罗静初[1], 吴才宏[2], 顾孝诚[1,①]

(1. 北京大学生命科学学院, 北京 100871; 2. 北京大学膜及膜的工程国家重点实验室, 北京 100871)

摘 要: 为配合总体的实验研究构建了中华民族基因组多态性(Genomic Polymorphism of Chinese Ethnic Groups,简称GPCEG)数据库,现已初步建成包括民族名称、基本情况介绍、体态特征、基因多态性数据、永生细胞株系、参考文献、国际相关数据库连接等内容的数据库,并完成了其可视化浏览及查询系统的建立,为建成具有中国特色的国家自有数据库奠定了基础,也可为从事相关研究的科学工作者提供信息服务。

关键词: 生物信息学; 生物学数据库; 中华民族; 基因组多态性

中图分类号: Q987; TP311.132 **文献标识码**: A **文章编号**: 0379-4172(2003)06-0509-06

The Chinese nation, a general name for various Chinese ethnic groups, is a unitary appellation of numerous ethnic groups living in China and undergoing mutual influence, mergence, amalgamation and disintegration. At present, there are 56 ethnic groups affirmed in China, they differ greatly in geographic distribution, population quantity, spoken and written language, religious belief and physical characteristics (including ABO blood group, body height, age of menarche, fingerprint characteristics, etc.)[1].

Genetic polymorphism, such as single nucleotide polymorphism (SNP) and microsatellite DNA etc, is helpful both to the recognition of genetic structures of different ethnic groups to explore the "origin" and "derivatives" of different human groups and to the understanding of incidence rate of hereditary diseases in populations of different

收稿日期:2002-06-17;**修回日期**:2003-04-14

基金项目:国家自然科学基金(39993420)资助项目[The project was supported by National Natural Science Foundation of China(No. 39993420)]

作者简介:陈 新,女,博士,讲师。研究方向:生物信息学。E-mail:chenxin@mail.cbi.pku.edu.cn

① 通讯作者。E-mail:guxc@lsc.pku.edu.cn;Tel:8610-62751843

ethnic groups and different regions ,their sensitivity to environmental factors and their susceptibility to diseases[2~4].

Minority nationalities of China are distributed ,for the most part ,in the remote border provinces in the Southwest ,the Northwest ,the Northeast and South China. The differences in cultural background ,habits and customs of different ethnic groups and the strict prohibition of intermarriage or rare intermarriage among them have evolved genetically isolated populations ;the genomes of these ethnic groups or isolated human groups are valuable material for studies on human genetic polymorphism. In order to avoid the loss of these rare resources ,it is important to preserve the genomes of these ethnic groups by establishing cell lines ,followed by studies on the differences in genomic structures and comparison of these data with those of other races in the world. The outcoming results will be of great significance both to the study of anthropology and sociology (expounding human evolution and the origin , migration and differentiation of various Chinese ethnic groups) and to biological studies (segregation and identification of important pathogenic genes and disease susceptible genes and their application in gene diagnosis and therapy)[5]. At the same time ,the acquisition ,systematization ,storage ,sharing and development of data and information obtained in above-mentioned studies through constructing bioinformatics databases will become an important component part of the international human genome program[6~9].

Construction of the biological database is an important part in the study of bioinformatics[10,11]. In the present study ,a database for Genomic Polymorphism of Chinese Ethnic Groups (GPCEG) was constructed in coordination with the overall experimental study. Now a database containing the denomination and basic information of Chinese ethnic groups ,introduction of their physical characteristics ,data of genomic polymorphism ,cell lines ,references and links wih other international relevant databases has been preliminarily established. The visualization and query systems are also available ,thus laying a foundation for setting up a state-owned database with Chinese characteristics and providing information service for scientists engaged in related researches.

1 Construction of the Database

1.1 Design of the Conceptional Structure of Database

The design of a database is the core of the whole biological database system ,and the basis of database structure is data model ,namely ,design of conceptional structure. Our database selected the logical model based on objects ,namely ,"entity-relationship data model (E-R figure for short)". The E-R model is in line with such an understanding of the real world :the world consists of a group of elementary objects called entity and the linkage between these objects.

According to the acquired data on genomic polymorphism of Chinese ethnic groups and analysis of these data and other related data ,the database for genomic polymorphism of Chinese ethnic groups was affirmed to be a relation-type database taking polymorphism data as its core , its E-R figure is shown in Fig. 1.

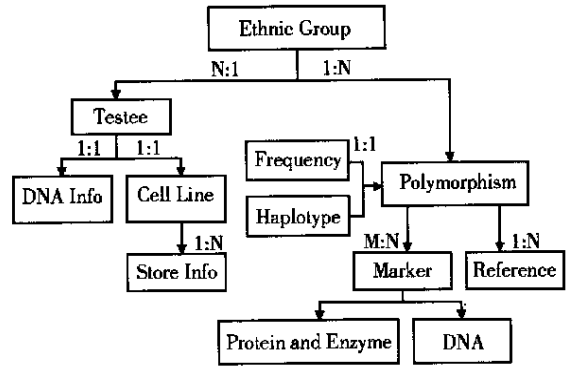

Fig. 1 E-R diagram of GPCEG

Rectangles represent entity sets ;Diamonds represent the relationship between two entity sets ;1 :1 represents the relationship one to one ;1 : N represents the relationship one to many. N : 1 represents the relationship many to one ;M :N represents the relationship many to many.

Citation of external databases used in GPCEG was mainly links to polymorphic Marker and GenBank of protein and enzyme[12]. Moreover ,in order to keep in good with further research work and to provide more information resources of studies on nationality and nationality diversity as well as query on network and network navigation for users ,we have also constructed Database for Human Polymorphism (DHP for short). Now it contains denomination

and basic information of Chinese ethnic groups, photographs, introduction of their physical characteristics, reference index and links of their physical characteristics, reference index and links of other international related databases. The mutual link between DHP and GPCEG is a good supplement of information to the Database for Genomic Polymorphism of Chinese Ethnic Groups.

1.2 Design of the Logical Structure

Design of logical the structure of a database is the transformation of its conceptional structure to a data model supported by a database management system (DBMS). We selected the relation-type DBMS which at present is still one of the main systems, and used Access, the Microsoft database management system on personal computer. For the relation-type DBMS, the design of logical structure is the transformation of an E-R diagram to a two-dimensional table. The GPCEG contains the following principal data tables, as shown in Table 1.

Table 1　Tables in GPCEG

Name of tables	Caption for tables
Frequency data	Polymorphism data of frequency type, including many auxiliary tables, each Marker has its corresponding data table
Haplotype data	Polymorphism data of haplotype type
DNA Marker	Information on DNA Marker, including name, location, sequence, description, type, etc.
Location	Every polymorphism datum corresponds to the sampling site of ethnic group concerned
Reference	Reference of polymorphism data
Ethnic group	Information on ethnic group, including source of name, population and distribution, language in use, origin of ethnic group, religious belief, customs and habits, characteristics of costume and adornment and physical characteristics
Testee	Information on persons being sampled, including information of name, ethnic group, place of birth, language and their parents
Cell line	Information on storage of cell line in liquid nitrogen box, including the presence or absence of revivification test and mycoplasma test
Store info	Information on sampling, including place, time, ethnic group and sampler
DNA info	DNA information on persons being sampled, including DNA content, OD260/OD280, volume and state of use

1.3 Realization of the Database and Its Visualization

Considering the curreut hardware environment in most laboratories and installation convenience for users, Microsoft Active Server Pages (ASP) supportable by the Windows system was used to realize dynamic network pages of the database while linkage to the database was accomplished by means of Active Data Object (ADO). ADO VB Script language was used for ASP, and Structural Query Language (SQL) was used for ADO. HTML language and frame structure (FRAME) were used in page surface design of the database.

2 Details of Database Content

2.1 Ethnic Groups

2.1.1 Introduction of Chinese Ethnic Groups
2.1.2 Information on 56 Ethnic Groups

2.2 DNA Polymorphism Marker

2.2.1 Autosome DNA Polymorphism Markers
2.2.2 X Chromosome Polymorphism Markers
2.2.3 Y Chromosome Polymorphism Markers
2.2.4 Microsatellite
2.2.5 SNP
2.2.6 Haplotype Data
2.2.7 Mitochondria DNA Markers
2.2.8 Restriction Enzyme Site Polymorphism

2.3 Protein and Enzyme Polymorphism Data

2.3.1 Introduction of Protein and Enzyme Polymorphism Data
2.3.2 Table of Protein and Enzyme Polymorphism Markers
2.3.3 Polymorphism Data of Protein and Enzyme Markers
2.3.4 Polymorphism Data
2.3.5 Related References

2.4 Cell Line

2.4.1 Sampling Information

2.4.2 Storage Information on Cell Line

The amount of data in the present database is shown in Table 2.

Table 2 Contents of GPCEG

Data	Entries
Polymorphism data	5810
Frequency polymorphism data	5355
Haplotype Polymorphism data	455
Ethnic group data	56
DNA Marker data	10626
Location data	218
Reference data	108
Cell line data	2910
Data of persons being sampled	520
Storage data	2300
DNA data	90
Other data (including link table ,etc.)	297

Browsing and query of database GPCEG contents can be made via network. Homepage of GPCEG is shown in Fig. 2.

2.5 Query System of the Database

As an important part of the database ,the query system enables users to get their required data conveniently and quickly ,bypassing large quantities of information useless to the user and raising utilization rate of the database. The query of database for GPCEG includes mainly the query of polymorphism information and that of cell line information. It can be made in simple condition or complex conditions. For query of complex conditions ,there are following examples :

(1) For query of Polymorphism information , multi-choice selection can be made ,for example ,protein (enzyme) or DNA data can be selected ,while in DNA data a choice of microsatellite or SNP as well as mitochondria data can also be made. Dual-directional query can also be done , such as selection of genotype information to query

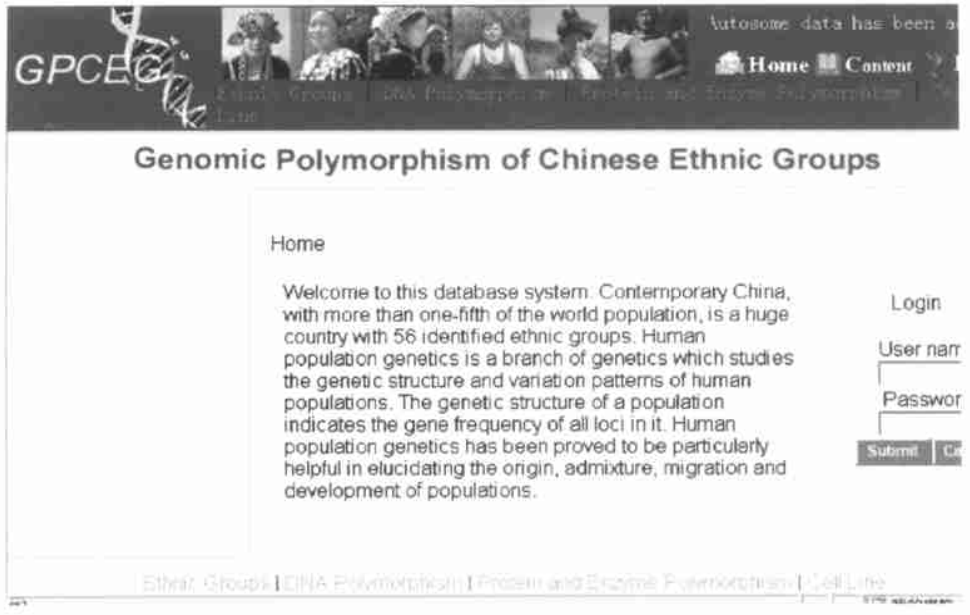

Fig. 2 Homepage of GPCEG

that of ethnic group or vice versa ,and simultaneous selection of a genotype and an ethnic group for query is also available.

(2) Query of cell line includes query of status of refrigeration and preservation of cells and sampling status. For query of refrigeration and preservation of cells ,four query modes are provided :direct input of serial number of refrigerated and preserved cells ;input of position of refrigerated and preserved cells in liquid nitrogen box ;input of date when cells were refrigerated and preserved ;and input of ethnic group whose refrigerated and preserved cells are to be queried. Query of status of sampling is a comprehensive one ;it provides substantial content for selection ,including serial number , sex , language in use , birthplace , age range ,ethnic group ,and time or place of sampling.

2.6 Updating and Submitting of Data

During the course of regular operation of the database continuous assessment, regulation and revision may become necessary to accommodate the ever increasing wealth of data and knowledge as a recent progress of the research. According to the requirement of relevant laboratories, GPCEG contains functions of updating and submitting of cell line data (including alterations in sampling data, refrigeration and preservation data) and DNA data.

2.7 Safety Management of the Database

In order to guarantee the safety and integrity of information in the database and the privacy of persons being sampled, the homepage of GPCEG provides landing window which requires the user to input his (her) ID and password. The user is required to apply to the database for his (her) ID and password to the database manager. The user's data-in are written in some record files, only when they are confirmed to be correct and valid by the manager of the database, can they be introduced into database.

3 Discussion

For the data of GPCEG, we used the management system of Microsoft Access to establish a relational database and to get database networked by means of ASP and SQL. Usage of this technique proceeds mainly from the consideration of the present status of computer equipment and installation in most laboratories, but data in the database can be transformed from one management system to another, such as MYSQL, Oracle, etc. For safety and stability, the UNIX/Linux platform may become necessary in the future.

In order to maintain the Chinese characteristics of the database, to be internationally compatible and also taking user convenience into account, English and Chinese editions of the database are to be established. The functions of primary data retrieval and of information updating on network for multiple users under the conditions of safety supervision and authentication of identity will also be realized. The part of information on ethnic groups in Chinese has been completed, but the Chinese version of the other contents remains to be accomplished. The contents on cell lines in the database is set up in accordance with the requirements put forward by the Institute of Genetics and the Medical Genetics Laboratory of Harbin Medical University, mainly for the purpose of providing the lab with a convenient and effective method for data storage, browsing, query and updating plus a friendly interface.

Based on considerations on the whole infrastructure of the database, an independent heading for the X chromosome polymorphism marker is listed although no data is presently available. It provides a good basis for further studies and better maintenance and expansion of the database, along with future progress of experimental studies data to be recorded in the database.

At present, the database for GPCEG is basically established with its infrastructure, major contents and numerous data, and along with the progress in polymorphism studies it will continue to develop. Research work on the basis of this database can further acquire data on religion and surname and develop into a larger and more detailed database through mutual linkage with other relevant databases. They may serve such disciplines as sociology demology and on the other hand, provide a basis for computation of genetic distance, construction of consanguinity tree and development of analytical software of genetic geography, and other relevant algorithms and softwares.

Acknowledgments This work is part of the project on "Studies on New Techniques and Methods" under the title of "Structure and Function studies on the Chinese Ethnic Groups" which is a major program of the National Natural Science Foundation of China led by profs CHEN Zhu, QIANG Bo-Qin, SHEN Yan *et al*. We are very grateful to Prof DU Ruo-Fu for his special permission to use the information on ethnic groups in his book "Ethnic Groups in China"; to Profs CHU Jia-You, JIN Li, LI Pu and XU Jiu-Jin for the data of DNA Marker and cell lines. We will express our heartfelt thanks to them for their Supervison and energetic support in this work.

参考文献(References):

[1] DU Ruo-Fu, YE Fu-Sheng. Ethnic Groups in China. Beijing: Science

Publishing Company ,1994.
杜若甫 ,叶傅升. 中国的民族. 北京 :科学出版社 ,1994.

[2] Chu J Y ,Huang W ,Kuang S Q. Genetic relationship of populations in China. *Proc Natl Acad Sci USA* ,1998 ,95(20) :11763～11768.

[3] LI Sheng-Bin , FENG Ji-Dong ,LI Shuang-Ding , YU Jia-Lin , YANG Huan-Ming. Genescan for STR analysis and genetic distribution in a population sample from Han , China. *Acta Genetica Sinica* ,2000 ,27 (6) :477～484.
李生斌 ,冯继东 ,李双顶 ,于嘉林 ,杨焕明. 中国汉族人群(西安) STR 基因扫描与遗传结构. 遗传学报 ,2000 ,27(6) :477～484.

[4] LI Sheng-Bin , HU Hai-Tao , REN Hui-Min ,LI Zheng-Dao. Human DNA polymorphism of *Hae* Ⅲ system in Chinese oriental ,Han population. *Acta Genetica Sinica* ,2000 ,27(9) :753～761.
李生斌 ,胡海涛 ,任惠民 ,李政道. 中国汉族群体(西安) DNA *Hae* Ⅲ系统遗传多态性. 遗传学报 ,2000 ,27(9) :753～761.

[5] Ke Y ,Su B ,Song X ,Jin Li. African origin of modern humans in East Asia: a tale of 12 000 Y chromosomes. *Science* , 2001 May , 292 : 1151～1153.

[6] Chen Wei , Zhang Ge , Zhang Si-zhong. Introduction to Go ! Poly , a human genome polymorphism database. *Chi J Med Gen* ,2001 ,18(6) : 482～485.
陈 炜 ,张 戈 ,张思仲. 人类基因组多态数据库 Go ! Poly 及其应用. 中华医学遗传学杂志 ,2001 ,18(6) :482～485.

[7] WANG Zhe , HUANG Gao-Sheng. Database resources of the national center for biotechnology information and its application. *Chinese Bulletin of Life Sciences* ,2002 ,14(1) :59～62.
王 哲 ,黄高升. NCBI 的数据库资源及其应用. 生命科学 , 2002 ,14(1) :59～62.

[8] XUE Ya-li , WANG Qi , SHI Zhong-cheng. Establishment and Preservation of Immortal Lymphoblastoid Cell Lines of the Ewenki ,the Oroqen and the Daur Ethnic Groups in China. *HEREDITAS* (*Beijing*) , 2000 ,22(3) 157～158.
薛雅丽 ,王 琦 ,史忠诚. 中国鄂温克、鄂伦春、达斡尔族永生细胞株的建立与保存. 遗传 ,2000 ,22(3) :157～158.

[9] CHEN Zhu , ZHANG Si-zhong. Chinese human genome project —— opportunity and challenge. *Chi J Med Gen* ,1998 ,15(4) .
陈 竺 ,张思仲. 我国人类基因组研究面临的机遇与挑战. 中华医学遗传学杂志 ,1998 ,15(4) .

[10] LI Bing ,LUO Jing-chu , PAN Wei , TANG Wen , GU Xiao-cheng. The development and utilization of molecular biology databases and related softwares. *HEREDITAS* (*Beijing*) ,1999 ,21(4) :52～53.
李 兵 ,罗静初 ,潘 卫 ,唐 汶 ,顾孝诚. 分子生物学数据库及相关软件的开发利用. 遗传 ,1999 ,21(4) :52～53.

[11] Attwood T K ,Parry-Smith D J (Translated by Luo ,Jingchu) . Introduction to bioinformatics. Beijing : Peking University Press ,2002 ,1～21.
Attwood T K ,Parry-Smith D J (罗静初译) . 生物信息学概论. 北京 :北京大学出版社 ,2002 ,1～21.

[12] HU De-hua , Fang Ping. The retrieval of GenBank by E-mail. *HEREDITAS* (*Beijing*) ,1999 ,21(6) :43～46.
胡德华 , 方 平. 基因库 (GenBank) 的电子邮件检索. 遗传 , 1999 ,21(6) :43～46.

(责任编辑 :李绍武)