Human genome research in China

Boqin Qiang

Abstract Significant progress in human genome research has been made in China since 1994. This review aims to give a brief and incomplete introduction to the major research institutions and their achievements in human genome sequencing and functional genomics in medicine, with emphasis on the “1% Sequencing Project”, the generation of single nucleotide polymorphism and haplotype maps of the human genome, disease gene identification, and the molecular characterization of leukemia and other diseases. Chinese efforts towards the sequencing of pathogenic microbial genomes and of the rice \((Oryza sativa \text{ ssp. Indica})\) genome are also described.

Keywords Human genome · Microbial genome · Rice genome · Human disease related genes

Introduction

The aim of human genomics research is to decode the genetic information held by the human genome, and to decipher the structure and function of all human genes. The field of genome research, which has expanded from humans to other organisms, is flag-shipped by the Human Genome Project (HGP). The HGP was initiated in United States at the beginning of the 1990s, and has become an international collaborative project joined by the United Kingdom, Japan, France, Germany, and China. The completion of the sequencing of the human genome was announced on 14 April 2003 by the heads of the six member states with an historic proclamation indicating that the goals of the HGP had been accomplished. Though a latecomer, China has contributed much to the field of human genome research through international collaborations and strong government support. This review summarizes recent progress in genome research achieved by major research institutions in China.

The Human Genome Project in China

The first official national-state project on the human genome, headed by Drs Zhu Chen and Boqin Qiang as coordinators and sponsored by the National Natural Science Foundation of China, was officially announced in 1994. Due to limited resources at that time, the goals of this first project were to focus on the establishment of the infrastructure and core technology, collection and storage of genetic samples, as well as the training of research professionals. This pioneer project marked the beginning of human genome research in China and laid the foundation for further projects which were subsequently supported by the central government through the National High-Tech Development Program (“863 Programs”), the
National Developmental Program of Key Basic Research (“973 Programs”), and other research programs by the Ministry of Science and Technology (MOST), the Chinese Academy of Sciences (CAS), as well as by the Beijing, Shanghai, and other local governments. The whole governmental budget for genome research and other relevant projects, mainly devoted to research running costs, is about 200 million US dollars for the years 2001–2005. Dozens of research centers and laboratories have been established nation-wide (Table 1). In addition to the central and local governments, state-owned enterprises and domestic venture capital, joined lately by private companies and foreign investors, have also begun to invest in research and the development of genome-related technology and products. Two areas have been emphasized in human genome research in China: large-scale sequencing, genome diversity and proteomics; and the functional study of disease-related genes and other genes of scientific and economic significance.

### Human genome sequencing and studies of genome diversity

China’s application to join the HGP was officially accepted at the Fifth International Strategy Meeting on Human Genome Sequencing, which was held on 1 September 1999 in Hinxton, UK. China assumed approximately 1% of the whole human genome sequencing effort, thus generally named the “1% Sequencing Project”. Under the supervision of a seven-member executive group headed by Dr. Huanming Yang, and through the joint efforts of the Beijing Genomics Institute (formerly the Human Genome Center, Institute of Genetics, CAS), the Chinese National Human Genome Center at Shanghai (CHGCS, or the South China National Human Genome Research Center) and the Chinese National Human Genome Center at Beijing (CHGB, or the North China Human Genome Center), the mission working draft was finished on schedule at the end of May 2000. China generated 653,000 reads, 131% of the originally assigned task of 500,000 reads, with a total of non-abundant sequences of 27.5 Mb. The Chinese contribution to the working draft of the human genome is summarized in [1]. China announced the completion of the assigned region on 26 August 2001, a result of the productive collaboration of these three major centers.

The International Human Genome HapMap (Haplotype Map), an international collaborative project second to the HGP aiming at linking differences in the human genome to differences in the susceptibility to human diseases, was officially launched jointly by the United States, United Kingdom, Japan, China, and Canada on 29 October 2002. The three centers mentioned above, in collaboration with other laboratories in Hong Kong and

| Institute | Location | Leaders | Research focus |
|-----------|----------|---------|---------------|
| Beijing Genomics Institute (BGI) | Beijing | Huanming Yang, Jun Yu, Jian Wang | Genomics and proteomics in humans, plants, animals and microbes |
| Hangzhou Genomics Institute (HGI, sister Institute of BGI) | Hangzhou | | |
| Chinese National Human Genome Center at Shanghai (CHGCS, South China Human Genome Center) | Shanghai | Zhu Chen | Human genomics (primarily medical genomics), and microbial genomics |
| Chinese National Human Genome Center at Beijing (CHGB, North China Human Genome Center) | Beijing | Boqin Qiang, Yan Shen | Human genomics (primarily medical genomics), and microbial genomics |
| Pathogenic Microbial Genome Research Center, Ministry of Health and Beijing Microbial Genome Center | Beijing | Qi Jin, Yunde Hou | Sequencing and functional studies of the microbial genome |
| The Research Center for Human Disease Genomics, Peking University | Beijing | Dalong Ma | Functional studies of human disease-related genes |
| State Key Laboratory of Molecular Oncology, Chinese Academy of Medical Sciences | Beijing | Min Wu, Qimin Zhan | Genomics and molecular oncology |
| State Key Laboratory of Medical Molecular Biology, Chinese Academy of Medical Sciences | Beijing | Lifang Wang, Depei Liu | Structural and functional studies of human genes and proteins |
| State Key Laboratory of Human Genome Research, and Shanghai Institute of Hematology (SIH), Shanghai Second Medical University | Shanghai | Zhu Chen | Medical genomics, and molecular mechanisms of leukemia |
| State Key Laboratory of Oncogenes and Related Genes, Shanghai Institute of Oncology | Shanghai | Jianren Gu, Shengli Yang | Genomics, molecular oncology |
| State Key Laboratory of Genetic Engineering, Fudan University | Shanghai | Long Yu | Human genetics and genomics |
| Shanghai Institute for Biological Sciences, Chinese Academy of Sciences | Shanghai | Gang Pei | Life sciences, genomics and proteomics |
| Bio-X Life Science Research Center, Shanghai Jiao Tong University | Shanghai | Lin He | Human genomics and genetics |
| State Key Laboratory of Medical Genetics, Zhong Nan University | Changsha, Hu Nan Province | HanXiang Deng, Jiahui Xia | Human genomics and medical genetics |

The websites of the BGI: http://www.genomics.org.cn; the CHGB: http://www.chgb.org.cn; and the CHGCS: http://www.chgc.sh.cn
Taiwan, are responsible for around 10% of the whole effort.

### Medical genome research

Genetic resource collection

Emphasis has been laid on the collection and storage of human genetic materials. An ELSI (Ethical, Legal and Social Issues) Committee was established. Since the very beginning of human genome research in China in 1994, Dr. Rufu Du, professor of the Institute of Genetics, CAS, was the first chairperson. Each National Human Genome Center has an ELSI committee. Dr. Renbiao Chen, genetics professor at the Shanghai Second Medical University, is the chairman of the CHGCS ELSI committee. Dr. Renzong Qiu, professor of the Chinese Academy of Society Sciences, held a post as the chairman of the CHGB ELSI committee. The ELSI committee not only supervises the ELSI in the collection of genetic resources, but also participates in many related national or international activities. Training courses were given to the technical personnel participating in the investigation and collection. Principles, procedures and standards were established for the bioethical issues. The network and database for registration, collection and data management were established. Accordingly, blood and tissue samples were collected, providing the resources for the study of genome diversity and studies of disease-related genes.

Identification of genes related to hereditary diseases

In recent years, great progress has been made in China towards identifying monogenic disease genes. The first disease-related gene isolated by Chinese geneticists was the human gap junction protein β-3 (GJB3) by Dr. J.H. Xia’s group (Table 2). Further mutation analysis revealed a missense mutation and a nonsense mutation of GJB3 associated with high-frequency hearing loss in two families [2]. In 2001, two groups, Dr. Y. Shen’s group at CHGB and Dr. X.Y. Kong’s group from the Shanghai Research Center of Biotechnology, Shanghai Institute for Biological Sciences, CAS, subsequently isolated the genes related to dentinogenesis imperfecta Shields type II (DGI-II). Shen et al. identified a nonsense mutation (Gln45→stop) in exon 3 of the dentine sialophosphoprotein (DSPP) gene in a Chinese family with dentinogenesis imperfecta Shields type II (DGI-II) [3]. Kong et al. identified mutations of the DSPP gene in three Chinese families: a donor site mutation (G→T) in intron 3 of DSPP in family I, a missense mutation (C49→A) in exon 2 of DSPP in family II, and a missense mutation (G52→T) in exon 3 of DSPP in family 3. Two of these families also have progressive hearing loss [4].

| Disease                                      | Gene                                           | Mutation                                                                 | Locus   | References |
|----------------------------------------------|------------------------------------------------|--------------------------------------------------------------------------|---------|------------|
| High-frequency hearing loss                  | Gap junction protein β-3 (GJB3)                | Family I: Gln183→Lys (G547→A)                                            | 1p33-p35| [2]        |
| Dentinogenesis imperfecta Shields type 2 (DGI-II) | Dentine sialophosphoprotein (DSPP)           | Family II: Arg180→stop (C538→T)                                          | 4q21    | [3]        |
| Dentinogenesis imperfecta type 1 (DGI-I) with or without progressive hearing loss |                                             | Family I: donor splice site (GT) in intron 3, G→A                         | 4q21    | [4]        |
|                                               |                                               | Family II: Pro17→The (C49→A)                                             |         |            |
|                                               |                                               | Family III: Phe18→Val (G52→T)                                            |         |            |
|                                               |                                               | Family I: Glu95→Lys (G283→A)                                             | 2q35-36 | [5, 6]     |
|                                               |                                               | Family II: Glu131→Lys (G391→A)                                           |         |            |
|                                               |                                               | Family III: Asp100→Glüh (C300→A)                                         |         |            |
| Brachydyctaly type A-1                       | Indian hedgehog (IHH)                          | Family I: Leu115→Pro (T384→C)                                           | 16q23   | [7]        |
| Autosomal dominant Lamellar and Marner cataract| Heat-shock transcription factor 4 (HSF4)      | Family II: Leu115→Pro (C362→T)                                           |         |            |
|                                               |                                               | Individual 1: Ala20→Asp                                                   |         |            |
|                                               |                                               | Individual 2: Ile87→Val                                                   |         |            |
|                                               |                                               | Ser140→Gly (A418→G)                                                     | 11p15.5 | [11]       |
| Familial atrial fibrillation (AF)            | KCNO1 (potassium channel subunit)             | Twelve different mutations in exon 3 were identified in 17 of 31 patients, with two novel mutations | Xq28    | [8, 9]     |
|                                               |                                               | 12 missense mutations were identified in 118 patients                     |         |            |
| Rett syndrome                                | X-linked methyl-CpG-binding protein (MECP2)    | Twelve different mutations in exon 3 were identified in 17 of 31 patients, with two novel mutations |         |            |
| Childhood absence epilepsy (CAE)             | CACNA1H (T-type calcium channel alpha 1H subunit) | Two novel mutations were identified in 118 patients                       | 16p13.3 | [10]       |
| Agenesis of the permanent teeth (He-Zhao deficiency) OMIM 604625 |               |                                                                         | 10q11.2 | [12, 13]  |
| Disseminated superficial actinic porokeratosis | CACNA1H (T-type calcium channel alpha 1H subunit) | Two novel mutations were identified in 118 patients                       | 15q25.1-q26.1 | [14]    |
(G238→A, G391→A, and C300→A) have been identified in the region encoding the N-terminal signaling domain of the IHH gene in all patients in three large unrelated families [5, 6]. Table 2 shows several results of identifying gene mutations in Chinese pedigrees: autosomal dominant Lamellar and Marner cataracts [7], Rett syndrome [8, 9], and childhood absence epilepsy [10].

Recently, a type of familial atrial fibrillation (AF) was mapped to the 11p15.5 region and a causative mutation (Ser140→Gly) of the disease-related gene, KCNQ1, was identified by whole genome scanning, single nucleotide polymorphism (SNP) analysis and haplotyping, as well as association analysis of genotype and phenotype by Dr. W. Huang’s group in CHGCS and their collaborators. The KCNQ1 gene encodes the pore-forming subunit of the cardiac Iks channel (KCNQ1/KCNE1), the KCNQ1/KCNE2 and the KCNQ1/KCNE3 potassium channels. The S140→G mutation is likely to initiate and maintain AF by reducing action potential duration and the effective refractory period in atrial myocytes [11].

In addition, several hereditary disease-related genes have been localized within the human genome. He-Zhao deficiency has been recently characterized, with a distinct form of agenesis of the permanent teeth that is different from other previously reported disorders of dentition. Using a DNA pooling method combined with two-point and multi-point linkage analysis, Dr. L. He’s group has mapped the gene locus for He-Zhao deficiency to chromosome 10q11.2 [12, 13]. The gene locus of disseminated superficial actinic porokeratosis (DSAP2) was mapped to chromosome 15q25.1–26.1 [14]. Furthermore, databases for monogenic disease mapping and cloning, pedigrees of nervous system diseases, and banks for pathologic tissues and organs have been established.

Genes related to cancer development

**Leukemia**

Significant contributions have been made to the understanding of the molecular basis of leukemia and the mechanisms of differentiation and apoptosis as well as therapies for acute promyelocytic leukemia (APL) by the Shanghai Institute of Hematology, headed by Dr. Zhu Chen. In most patients APL [15, 16] is characterized by the accumulation of promyelocytes containing a specific chromosomal translocation t(15;17) and exhibiting unique sensitivity to all-trans-retinoic acid (ATRA) [17]. Following previous studies on the molecular basis of gene rearrangement after chromosomal translocation at locus t(15;17) [17, 18], their major contributions to the analysis of gene structure and function over the last 10 years include:

1. Characterization of important rearrangements in the RARα gene on chromosome 17 and the MYL (later named PML) gene on chromosome 15, studies of the chimeric PML-RARα and RARα-PML fusion genes [19], and sequencing of the entire genomic DNA region of the PML and RARα genes [20, 21].
2. Discovery of an unusual karyotype 46,XY t(11;17) (q23; q21) in APL [22], analysis of the PLZF-RARα fusion gene in the truncated locus of this chromosome, and the cloning and analysis of the newly discovered PLZF (Promyelocytic Leukemia Zinc Finger) gene. PLZF encodes a potential transcription factor related to the Drosophila gap gene Kruppel and is expressed in at least two isoforms [23]. The entire genomic region of PLZF was sequenced [24]. The PLZF-RARα protein inhibits ligand-dependent transactivation of RARα, reflecting the “dominant negative” effect of PLZF-RARα on wild-type RARα [25].
3. Establishment of PLZF-RARα and NPM-RARα transgenic mouse models [26]. PLZF-RARα transgenic animals developed chronic myeloid leukemia (CML)-like disease, whereas NPM-RARα transgenic mice showed a spectrum of phenotypes ranging from typical APL to CML.
4. Analysis of gene expression profiling in the APL cell line NB4 before and after ATRA treatment to elucidate the molecular mechanisms of ATRA-induced differentiation of APL cells [27]. More recently, a cDNA microarray with 13,014 clones was screened to further explore these gene expression networks. The results of this screen showed that 318 genes were up-regulated and 291 were down-regulated by ATRA in NB4 cells. In the early stage of ATRA treatment, genes involved in inhibition of proliferation, cell cycle arrest and apoptosis antagonists were up-regulated, whereas with the onset of differentiation, genes related to granulocyte maturation and apoptosis agonists were up-regulated (unpublished data).
5. The NUP98-PMX1 fusion gene was cloned from a CML [28, 29] patient with AML-M2 transformation and bearing a secondary chromosomal translocation t(1;11)(q23;p15), and the observation that NUP98-PMX1 transgenic mice developed myeloid dysplastic syndrome (MDS) or myeloid leukemia was established. The fact that NUP98-PMX1 formed complexes only with histone deacetylase (HDAC) 1 in vivo and that its inhibitory effect on FOS could be abolished upon treatment with HDAC inhibitors strongly suggest that the NUP98-PMX1 chimera functions, at least for some genes, as a transcriptional inhibitor (unpublished data). NUP98/PLZF [30] and MLL-EEN fusion genes found in acute leukemias with t(11;12) (p15;q13) and t(11;19)(q23;p13) were also cloned and the functions of those fusion genes are under investigation.

**Nasopharyngeal carcinoma**

Nasopharyngeal carcinoma (NPC) occurs with high frequency in the southern part of China, especially among people of Cantonese ancestry. No predisposing genes have been identified up to now, though the HLA-BW46 locus is associated with increased risk of NPC [31]. Dr.
Y.X. Zeng’s group [32] instigated the genome-wide scanning of 20 families with a high risk of NPC (2–9 affected members per family) from Guangdong Province. Fifty-four affected individuals were genotyped using 382 polymorphic microsatellite markers, covering 22 autosomes with an average marker density of 10 cM. Parametric analysis provided evidence of linkage to the D4S405 marker on chromosome 4 with a logarithm of odds for linkage (lod) score of 3.06 and a heterogeneity-adjusted lod (hlod) score of 3.21. Fine mapping with additional markers flanking D4S405 resulted in a lod score of 3.45 and hlod score of 3.67 for the region 4p15.1-q12. This region has been recently narrowed down to 8.29 cM by means of further analysis of the pedigrees with simple tandem repeat and SNP markers (unpublished data). These results indicate that a susceptibility locus on 4p15.1-q12 may account for a significant subset of hereditary NPC which provides a solid basis for future studies to identify the NPC susceptibility gene. However, the above genome-wide scan results did not indicate any obvious linkage to chromosome 6 which contains the MHC locus, even though previous reports have suggested an association between the risk of NPC and certain MHC haplotypes [31]. A possible explanation for this discrepancy may lie in the different subjects studied: previous work was conducted with affected sib-pair families, whereas high-risk NPC pedigrees were used in this study.

**Hepatocellular carcinoma**

To identify genetic abnormalities in human primary hepatocellular carcinoma (HCC), microsatellite analysis was performed on 60 Chinese HCC specimens versus non-cancerous liver tissues as controls. To further evaluate the nature of the allelic loss, comparative genomic hybridization was accomplished in 20 pairs of the above HCC samples. The combined analysis of these two methods revealed frequent allelic loss on 17p, 9p21-p23, 16q21-q23.3, 13q, 8p21-p23 and 6q24-q27, whereas the most frequent allelic gain occurred at 1q, 17q and 8q24. The highest incidence of allele loss was 17p13.3 (65%) [33]. Twenty-two paired HCC and non-cancerous liver samples were then analyzed with 14 polymorphic markers. The data revealed a high level of loss of heterozygosity (>68%) in a minimum region between D17S1866 and D17S1574, spanning over a 1.5 Mb region. A physical map was created based on large-scale sequencing of relevant cluster bacterial artificial chromosome/artificial chromosome clones. Seventeen known genes and 13 novel genes were identified in this minimum region, and the functions of these genes were characterized [34, 35, 36].

Recently, a comprehensive characterization of gene expression profiles of positive hepatitis B virus infected HCC was initiated through the generation of a large set of 5′ expression sequence tag (EST) clusters (11,065 ESTs) from HCC and non-cancerous liver tissue samples. These ESTs were then applied to a cDNA microarray system containing 12,393 genes/ESTs, and a commercial cDNA array of 1,176 known genes as target sequence sources. The integrated data from this study identified 2,253 genes/ESTs as candidates showing differential expression. A number of these differentially expressed genes were verified by RT-PCR, which revealed that many were involved in cell cycle regulation, such as the cyclin-dependent kinases, cell cycle negative regulators and metabolic regulators. Some of these candidate genes may be related to cancer cell differentiation. Also, the altered transcriptome profile of HCC may arise from a number of chromosome regions exhibiting loss of heterozygosity or amplification [37].

**Esophageal squamous cell carcinoma**

Esophageal squamous cell carcinoma is one of the most common malignancies in the world, and has ranked as the fourth cause of cancer death in China since the beginning of the 1980s. Dr. M. Wu’s group performed comparative genomic hybridization to detect copy number changes in DNA. The most common gains were observed at 3q, 8q, 1q, 20q, 20p, 5p, 15q, and 9q, and losses at 3p, 13q, 18q, 9p, 4, Xp and others [38]. Studies of loss of heterozygosity (LOH) were conducted using microsatellite markers. Four minimal deletion regions of overlap were found at 3p14.2, 3p26, 13q12.3 and 13q14.1-q14.3 [39]. This group has also studied the expression patterns of a dozen genes such as TGase 3, Mal and RH50 in esophageal carcinomas and has identified their association with human cancers for the first time. The full-length cDNA and genomic DNAs of two novel genes (DRC1 and DRC2) were characterized, and their chromosome location and tissue-specific distribution of expression were elucidated [40, 41, 42, 43]. They have found that the expression of NMES1 and gut-enriched Kruppel-like factor were down-regulated [44, 45], while EC45 was overexpressed [46] in esophageal cancer. They also showed that down-regulation of the three annexin I isoforms, identified by proteomics, was a frequent event in esophageal carcinogenesis [47], and that the alteration of p63 might play a significant role in the early steps in tumorigenesis [48].

Genes associated with other common diseases

**Type 2 diabetes**

SNP screening and case-control associated analysis were carried out for 37 candidate genes located in a 4.3 Mb region of the p terminus of chromosome 1 [49]. Several genes were identified as susceptibility genes for type 2 diabetes in the northern Chinese population. These included the CDC2L2 (cell division cycle 2-like kinase 2), PGD (phosphogluconate dehydrogenase), caspase 9, PRKCZ (protein kinase Cζ), SAC (soluble adenylate cyclase), and UTSH (urotensin II) genes [50, 51]. Dr. M.
Luo’s group at the Shanghai Institute of Endocrinology, in collaboration with researchers at the CHGCS, carried out SNP analysis for the candidate genes involved in insulin signal transformation, lipometabolism and energy synthesis. The total length of these 11 candidate genes is 57,782 bp, with 87 SNPs, of which 54 occur at high frequency, and 33 at low frequency and with an uneven distribution.

Essential hypertension

Essential hypertension (EH) is a common late-onset disease that exhibits complex genetic heterogeneity. Several genome-wide scans recently accomplished in ethnic Chinese populations revealed a number of candidate loci possibly contributing to EH, and appeared to be replicable in 2q14-q23 and 5q32 [52, 53]. D.L. Zhu et al. reported linkage to the chromosome 2q14-q23 region for EH in a south China population [54, 55]. Several genes located in that region are considered to be relevant to the regulation of blood pressure and the development of EH. These genes include several G proteins and G protein-coupled receptors, a voltage-gated sodium channel protein and the gamma subunit of the sodium/potassium ATPase [56]. Another region, 5q32, harboring the β2-adrenergic receptor (β2AR) gene, has been reported to be linked to EH in a Taiwanese population [57]. To repeat these results and perform quantitative linkage analysis, D.F. Gu et al. genotyped members of 148 hypertension families containing 328 affected sib pairs, and grouped families from Beijing and Jiangsu province with five highly informative microsatellite markers (D2S151, D2S142, D5S2090, D5S413 and D5S2013), but the results provided no evidence in support of significant linkage of 2q14-q23 or 5q32 with EH [58]. Any explanation for the above results will be complicated due to the number of factors involved. At best, these observations indicate the diversity in the etiology and complexity of hypertension.

Another report recently published by Gu’s group showed that a region of chromosome 8 flanking the LPL gene might contribute to the individual blood pressure variations in Chinese. This study involved linkage analysis in 148 Chinese hypertensive families containing 328 affected sib pairs, and grouped families from Beijing and Jiangsu province with five highly informative microsatellite markers (D2S151, D2S142, D5S2090, D5S413 and D5S2013), but the results provided no evidence in support of significant linkage of 2q14-q23 or 5q32 with EH [58]. Any explanation for the above results will be complicated due to the number of factors involved. At best, these observations indicate the diversity in the etiology and complexity of hypertension.

Gene expression profiling and full-length cDNA cloning

More than 150,000 ESTs have been cloned and sequenced from human umbilical cord blood CD34+ cells were catalogued by partially sequencing a large number of EST clones and analyzing these sequences with bioinformatics tools. Three hundred cDNAs containing putative entire open reading frames for previously undefined genes were obtained; based on EST cataloging, clone sequencing, in silico cloning, and rapid amplification of cDNA ends [65]. Six novel Kruppel-like zinc finger genes from hematopoietic cells were cloned, and a novel transregulatory domain KRNB was identified [66]. Distinct gene expression patterns were found between CD34+ cells from normal bone marrow and AML-M5 transformed by myeloid dysplasia syndrome [67].

Microbial genome research

Research on microbial genomes is one of the hot topics in life science worldwide. The first microbe genome sequence contributed by China to the public domains of the international databases was that of Shigella flexneri (Serotype 2A), the most prevalent species and the serotype that causes bacillary dysentery or shigellosis in man. This sequencing project was completed by the Pathogenic Microbial Genome Center of Ministry of Public Health, and the CHGB. The genome size of S. flexneri (Serotype 2A) is 4,607,203 bp, with an extra virulence plasmid of 221,618 bp [68]. Its genome has, astonishingly, 314 IS elements and hundreds of pseudogenes. Genes encoding for Shigella outer membrane proteins have potential for vaccine development, and toxic factor genes with potential as targets for drug development have been cloned and characterized.

The complete genome sequence of a representative virulent serovar type strain (Lai) of Leptospia interrogans serogroup Icterohaemorrhagiae has been recently published [69]. The whole genome consists of a 4,332,241 bp large circular chromosome and a 358,943 bp small chromosome with a total of 4,768 predicted genes. A comprehensive analysis of the Leptospia interrogans genes for chemotaxis/motility and lipopolysaccharide synthesis provides a basis for in-depth studies of virulence and pathogenesis.

The genome sequences of other microbial species including Staphylococcus epidermidis, Xanthomonas campestris and so on are in progress. In addition, the
The first extremophile microbe sequenced by China was \textit{T. tengcongensis}, which is a thermophile bacteria isolated from Tengchong, Yunnan. The complete sequence was published by the Beijing Genomics Institute and its collaborators [70].

The world Health Organization officially announced that a variant coronavirus is the pathogen responsible for the severe acute respiratory syndrome (SARS) which broke out in China in spring this year. Five isolates from SARS patients identified in Guangdong and Beijing were sequenced and deposited into GenBank by Chinese scientists [71].

**Rice and other important organisms in China**

With the financial support from MOST the Shanghai Biochemistry Institute, CAS, established a National Sequencing Center headed by Dr. Guofan Hong in the 1990’s. The complete sequence map of chromosome 4 of \textit{Oryza sativa} ssp. \textit{japonica} was finished in that center in 2002 under the supervision of Dr. Bing Han, and was one of the two rice chromosomes to be completely sequenced [72]. The finished sequence spans 34.6 Mb and represents 97.3% of the chromosome. In addition, it has the longest known sequence for a plant centromere, a completely sequenced contig of 1.16 Mb corresponding to the centromeric region of chromosome 4. A total of 4,658 genes predicted from these sequences match the available unique rice ESTs.

A draft sequence of the rice (\textit{Oryza sativa} ssp. \textit{Indica}) genome was published [73] by Beijing Genomics Institute and its sister center in Hangzhou (Hangzhou Genomics Institute) with their large scale sequencing facilities and technology established through participation in the HGP. The completion of the whole rice genome fine map was announced in October 2002. All the rice sequence data have been made freely available as a Chinese contribution to the world under the banner of “owned by all, done by all, and shared by all”.

The pig is an important domestic animal of both medical and economic significance. A pig genome survey has been accomplished by Beijing Genomics Institute in collaboration with Royal Agricultural and Veterinary University, National Institute of Agriculture, and Aarhus University, Denmark, with financial support from the Danish Pig Production Committee and CAS. cDNAs from approximately 100 different tissues or different developmental stages have also been sequenced. In addition, chicken, silkworm and other genomes are also in the process of being sequenced by Beijing Genomics Institute, and the South and North China National Human Genome Research Centers.

**Summary**

Over the past ten years, Chinese scientists have made substantial contributions to genome research, especially with the accomplishment of 1% of the effort for the Human Genome Project and in the identification of genes related to hereditary diseases. China is a country with rich resources in terms of population size and biodiversity, so the major challenge for us is how to utilize our unique genetic resources for identifying other genes of biological significance. We are hoping to make further contributions to functional genomic research in collaboration with scientists in other countries.

As one of the chief coordinators of the genome projects in China, I am proud to note that genome research has attracted so many young, and established, Chinese scientists who have been well trained either in the West or in China. They have proved themselves by their enthusiasm for science, and their experience in research and its management. They have been the major contributors to this field. They, following in the steps of their predecessors, have laid a solid foundation for further development and will build the future of life science and biotechnology in China.

**Acknowledgements** The author appreciates the sharing of published and unpublished data by Drs Zhu Chen, Jianren Gu, Fude Fang, Mingrong Wang, Yan Shen, Qi Jin, Dongfeng Gu, Lin He, Wei Huang, Min Lou, Guoping Zhao, Jiahui Xia and other investigators. I thank Drs Huannng Yang, Saijuan Chen and Jianren Gu for contributions to preparing this review, and Dr. Dahai Zhu for kind suggestions. I thank the Ministry of Science and Technology and the National Natural Science Foundation of China and other funding agencies for their support.

**References**

1. International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921
2. Xia JH, Liu CY, Tang BS, Pan Q, Huang L, Dai HP, Zhang BR, Xie W, Hu DX, Zheng D, Shi XL, Wang DA, Xia K, Yu KP, Liao XD, Feng Y, Yang YF, Xiao JY, Xie DH, Huang JZ (1998) Mutations in the gene encoding gap junction protein β-3, associated with autosomal dominant hearing impairment. Nat Genet 20:370–373
3. Zhang XH, Zhao J, Li CF, Gao S, Qiu CC, Liu P, Wu GY, Qiang BQ, Lo Wilson HY, Shen Y (2001) DSPP mutation in dentinogenesis imperfecta Shields type II. Nat Genet 27:151–152
4. Xiao SX, Yu C, Chou XM, Yuan WJ, Wang Y, Bu L, Fu G, Qian MQ, Yang J, Shi YZ, Hu LD, Han B, Wang ZM, Huang W, Liu J, Chen Z, Zhao GP, Kong XY (2001) Dentinogenesis imperfecta I with or without progressive hearing loss is associated with distinct mutations in \textit{DSPP}. Nat Genet 27:201–204
5. Yang X, She CW, Guo JZ, Yu CH, Lu YJ, Shi XL, Feng GY, He L (2000) A locus for brachydactyly type A-1 maps to chromosome 2q35–36. Am J Hum Genet 66:892–903
6. Gao B, Guo JZ, She CW, Shu AL, Yang MS, Tan Z, Yang XP, Guo SZ, Feng GY, He L (2001) \textit{Mutations in HH}, encoding the Indian hedgehog protein, cause brachydactyly type A-1. Nat Genet 28:386–388
7. Bu L, Jin YP, Shi YF, Chu RY, Ban AR, Eiberg H, Andres L, Jiang HS, Zheng GY, Qian MQ, Cui B, Xia Y, Liu J, Hu LD, Zhao GP, Hayden MR, Kong XY (2002) Mutant DNA-binding domain of HSFs is associated with autosomal dominant lamellar and Marner catarracts. Nat Genet 31:276–278
8. Amir RE, van den Veyver IB, Wan M, Tran CQ, Francke U, Zoghbi HY (1999) Rett syndrome is caused by mutations in X-linked MECP2, encoding a methyl-CpG-binding protein. Nat Genet 23:165–188
9. Pan H, Wang WP, Bao XH, Meng HD, Zhang Y, Wu XR, Shen Y (2002) MRF2 gene mutation analysis in Chinese patients with Rett syndrome. Eur J Hum Genet 10:484–486
10. Chen YC, Lo JJ, Pan H, Zhang YH, Wu HS, Xu KM, Liu XY, Jiang YW, Bao XH, Yao ZJ, Ding KY, Lo Wilson HY, Qiang BQ, Chan P, Shen Y, Wu XR (2003) Association between genetic variation of CACNA1H and childhood absence epilepsy. Ann Neurol 54:239–243
11. Chen YH, Xu SJ, Bendahhou S, Wang XL, Wang Y, Xu WY, Jin HW, Sun H, Su XY, Zhuang NQ, Yang YQ, Li YB, Liu Y, Xu HJ, Li XF, Ma N, Mou CP, Chen Z, Barhanin J, Huang W (2003) KCNQ1 gain-of-function mutation in familial atrial fibrillation. Science 299:251–254
12. Wang H, Zhao S, Zhao W, Feng G, Jiang S, Liu W, Li S, Xue GF, Huang J (2002) Variant-type PML-RARalpha fusion proteins generated from the variant t(11;17)(q23;q21) translocation in acute promyelocytic leukemia inhibit ligand-dependent transactivation of wild-type retinoic acid receptors. Proc Natl Acad Sci USA 99:6318–6323
13. Liu TX, Zhang JW, Tao J, et al (2000) Gene expression networks underlying retinoic acid-induced differentiation of acute promyelocytic leukemia cells. Blood 96:1496–1504
14. Chen SJ, Wang Q, Tong JH, et al (1991) Monoallelic deletions of the p53 gene in Chinese patients with chronic myelogenous leukemia in blastic crisis. Nouv Rev Fr Hematol 33:481–484
15. Su XY, Wang N, Cao Q, et al (1999) Chromosomal aberrations during progression of chronic myeloid leukemia identified by cytogenetic and molecular cytogenetic tools: implication of 1q12–21. Cancer Genet Cytogenet 108:6–12
16. Gu BW, Wang W, Wang Q, WM, et al (2003) Major form of NUP98/HOXC11 fusion in adult AML with t(11;12)(p15;q13) translocation exhibits aberrant trans-regulatory activity. Leukemia 17:1168–1174
17. Chen SJ, Zelent A, Tong JH, et al (1991) Rearrangements of the retinoic acid receptor-alpha locus due to a variant t(11;17)(q23;q21) in a patient with acute promyelocytic leukemia. J Clin Invest 91:2260–2267
18. Chen Z, Brand NJ, Chen A, et al (1993) Fusion between a novel Kruppel-like zinc finger gene and the retinoic acid receptor-alpha locus due to a variant t(11;17)(q23;q21) in a patient with acute promyelocytic leukemia. EMBO J 12:1161–1167
19. Zhang T, Xiong H, Kan LX, et al (1999) Genomic sequence, structural organization, molecular evolution, and aberrant rearrangement of promyelocytic leukemia zinc finger gene. Proc Natl Acad Sci USA 96:11422–11427
20. Chen Z, Guidice F, Rousselot P, et al (1994) PLZF-RAR alpha fusion proteins generated from the variant t(11;17)(q23;q21) translocation in acute promyelocytic leukemia inhibit ligand-dependent transactivation of wild-type retinoic acid receptors. Proc Natl Acad Sci USA 91:1178–1182
21. Cheng GX, Zhu XH, Men XQ, et al (1999) Distinct leukemia phenotypes in transgenic mice and different co-repressor interactions generated by promyelocytic leukemia variant fusion genes PLZF-RARalpha and NPM-RARalpha. Proc Natl Acad Sci USA 96:6318–6323
22. Chen YC, Lo JJ, Pan H, Zhang YH, Wu HS, Xu KM, Liu XY, Wang H, Zhao S, Zhao W, Feng G, Jiang S, Liu W, Li S, Xue GF, Huang J (2002) Variant-type PML-RARalpha fusion proteins generated from the variant t(11;17)(q23;q21) translocation in acute promyelocytic leukemia inhibit ligand-dependent transactivation of wild-type retinoic acid receptors. Proc Natl Acad Sci USA 99:6318–6323
41. Chen BS, Wang MR, Cai Y, Xu X, Xu ZX, Han YL, Wu M (2000) Decreased expression of the SPRY3 gene in Chinese human esophageal cancer. Carcinogenesis 21:2147–2150
42. Chen BS, Xu ZX, Xu X, Cai Y, Han YL, Wang J, Xia SH, Hu H, Wei F, Wu M, Wang MR (2002): RhoCG is down-regulated in esophageal squamous cell carcinomas but expressed in multiple differentiated squamous epithelia. Eur J Cancer 38:1927–1936
43. Xu ZX, Wang MR, Xu X, Cai Y, Han YL, Wu KM, Wang J, Chen BS, Wang QX, Wu M (2000) Human esophagus-specific novel gene C1orf10: cDNA cloning, gene structure and frequent loss of expression in esophageal cancer. Genomics 59:322–330
44. Zhou J, Wang HX, Luo AP, Ding F, Zhang J, Wang QX, Wu M, Liu ZH (2002) A novel gene, NMES1, down-regulated in human esophageal squamous cell carcinoma. Int J Cancer 101:311–316
45. Wang N, Liu ZH, Ding F, Wang QX, Zhou CN, Wu M (2002) Down-regulation of gut-enriched Kruppel-like factor expression in esophageal cancer. World J Gastroenterol 8:966–970
46. Wang Q, Yang CB, Zhou J, Wang QX, Wu M, Liu ZH (2001) Cloning and characterization of the E4CS5 gene which encodes human ribosomal protein L15 and is overexpressed in esophageal cancer. Gene 263:205–209
47. Xia SH, Hu LP, Hu H, Ying WT, Xu X, Cai Y, Han YL, Chen BS, Wei F, Qian XH, Cai YF, Shen Y, Wu M, Wang MR (2002) Three isoforms of annexin I are preferentially expressed in normal esophageal epithelia but dysregulated in esophageal squamous cell carcinomas. Oncogene 21:6641–6648
48. Hu H, Xia SH, Li AD, Xu X, Cai Y, Han YL, Wei F, Chen BS, Huang XP, Han YS, Zhang JW, Zhang X, Wu M, Wang MR (2002) Elevated expression of p63 protein in human esophageal squamous cell carcinoma. Int J Cancer 98:733–739
49. Du WN, Sun HX, Wang H, Qian BQ, Shen Y, Yao ZL, Gu J, Xiong MM, Huang W, Chen Z, Zuo J, Hua XF, Gao W, Sun Q, Fang FD (2001) Confirmation of susceptibility gene loci on chromosome 1 in Northern Chinese Han families with type 2 diabetes. Chin Med J 114:876–878
50. Sun HX, Zhang KX, Du WN, Shi JX, Jiang ZW, Sun H, Zuo J, Huang W, Chen Z, Shen Y, Yao ZL, Qian BQ, Fang FD (2002) Single nucleotide polymorphisms in CAPN10 in Chinese people and their correlation with type 2 diabetes mellitus in the Han people of northern China. Biomed Environ Sci 15:75–82
51. Sun HX, Du WN, Zuo J, Wu GD, Shi GB, Shen Y, Qian BQ, Yao ZL, Hang JM, Wang H, Huang W, Chen Z, Fang FD (2002) The association of two single nucleotide polymorphisms in the protein-ase C3-subunit gene (PRKCZ) and the urotensin gene (UTS1) respectively with type-2 diabetes in the Han people of northern China. Acta Acad Med Sin 24:223–227
52. Ni TH, Xu XP, Cordell HJ, Rogus J, Zhou YS, Fang ZA, Lindpaintner K (1999) Linkage analysis of candidate genes and gene-gene interactions in Chinese hypertensive sib pairs. Hypertension 33:1332–1337
53. Xu P, Rogus JJ, Terwedow HA, Yang JH, Wang ZX, Chen CZ, et al (1999) An extreme-sib-pair genome scan for genes regulating blood pressure. Am J Hum Genet 64:1694–1701
54. Zhu DL, Wang HY, Xiong MM, He X, Chu SL, Jin L, et al (2001) Linkage of hypertension to chromosome 2q14-q23 in Chinese families. J Hypertens 19:55–61
55. Zhu DL, Huang W, Chu SL, Wang GL, He X, et al (2002) Linkage analysis of a region on chromosome 2 with essential hypertension in Chinese families. Chin Med J 115:654–657
56. Chu SL, Zhu DL, Wang GL, Zhang WZ, Zhou HF, Gao PI, Zhan YM, et al (2002) Linkage analysis of twelve candidate gene loci regulating water and sodium metabolism and membrane ion transport in essential hypertension. Hypertens Res 25:635–639
57. Pan WH, Chen JW, Fann C, Jou YS, Wu SY (2000) Linkage analysis with candidate genes: the Taiwan early-onset hypertension genetic study. Hum Genet 107:210–215
58. Ge DL, Yang WJ, Huang JF, Yao CL, Xu XH, Gan WQ, Zhao JG, Liu DH, Wang XL, Duan XF, Hui RT, Shen Y, Yao ZJ, Qian BQ, Gu DF (2003) Linkage analysis of 2q14-q23 and 5q22 with blood pressure quantitative traits in Chinese sib pairs. J Hypertens 21:305–310
59. Yang WJ, Huang JF, Ge DL, Yao CL, Duan XF, Gan WQ, Huang GY, Hui RT, Shen Y, Qian BQ, Gu DF (2003) Evidence for linkage and association of the markers near the LPL gene with hypertension in Chinese families. J Med Genet 40:1–6
60. Mao M, Fu G, Wu JS, et al (1998) Identification of genes expressed in human CD34+ hematopoietic stem/progenitor cells by expressed sequence tags and efficient full-length cDNA cloning. Proc Natl Acad Sci USA 95:8175–8180
61. Zhang W, Wan T, Yuan Z, He L, Zhu X, Yu M, Cao X (2001) Genetic approach to insight into the immunobiology of human dendritic cells and identification of CD84-H1, a novel CD84 homologue. Clin Cancer Res 7:8226–8229
62. Yu YT, Zhang CG, Zhou GQ, Wu SF, Xu XH, Wei HD, Xing GC, Dong CN, Zhai Y, Wu TM, Huang GY, Chen Z, Chen MD, Chen J (2000) Gene expression profiling in the human hypothalamic-pituitary-adenal axis and full-length cDNA cloning. Proc Natl Acad Sci USA 97:9543–9548
63. Hu RM, Han ZG, Song HD, Peng YD, Huang QH, Ren SX, Gu YJ, Huang CH, Li YB, Jiang CL, Gu F, Zhang QH, Gu BW, Dai M, Mao YF, Gao GF, Rong R, Ye M, Zhou J, Xu SH, Gu J, Shi JX, Jin WR, Zhang CK, Wu TM, Huang GY, Chen Z, Chen MD, Chen J (2000) Gene expression profiling in the human hypothalamic-pituitary-adenal axis and full-length cDNA cloning. Proc Natl Acad Sci USA 97:9543–9548
64. Hu RM, Han ZG, Song HD, Peng YD, Huang QH, Ren SX, Gu YJ, Huang CH, Li YB, Jiang CL, Gu F, Zhang QH, Gu BW, Dai M, Mao YF, Gao GF, Rong R, Ye M, Zhou J, Xu SH, Gu J, Shi JX, Jin WR, Zhang CK, Wu TM, Huang GY, Chen Z, Chen MD, Chen J (2000) Gene expression profiling in the human hypothalamic-pituitary-adenal axis and full-length cDNA cloning. Proc Natl Acad Sci USA 97:9543–9548
65. Ren SX, Fu G, Wu JS, et al (2002) Cloning and functional analysis of cDNAs with open reading frames for 300 previously undefined genes expressed in CD34+ hematopoietic stem/progenitor cells. Genome Res 10:1546–1560
66. Han ZG, Zhang QH, Ye M, et al (1999) Molecular cloning of six novel Kruppel-like zinc finger genes from hematopoietic cells and identification of a novel transregulatory domain KRUN. J Biol Chem 274:35741–35748
67. Hu J, Zhang QH, Huang QH, et al (2000) Gene expression in CD34+ cells from normal bone marrow and leukemic origins. Hematol J 1:206–217
68. Jin Q, Yuan ZH, Xu JG, et al (2002) Genome sequence of Shigella flexneri 2a:insights into pathogenicity through comparison with genomes of E.coli K12 and O157. Nucl Acids Res 30:4432–4441
69. Ren SX, Fu G, Jiang XG, et al (2003) Unique physiological and pathogenic features of Leptospira interrogans revealed by whole-genome sequencing. Nature 422:888–893
70. Bao QY, et al (2002) A complete sequence of the T. tengcongensis genome. Genome Res 12:689–700
71. Qin ED, Zhu QY, Yu M, et al (2003) A complete sequence and comparative analysis of a SARS-associated virus (Isolate BJ01). Chin Sci Bull 48:941–948
72. Feng Q, Zhang Y, Hao P (2002) Sequence and analysis of rice chromosome 4. Nature 420:316–320
73. Yu J, Hu SN, Wang J, et al (2002) A draft sequence of the rice genome(Oryza sativa L. ssp. indica). Science 296:79–92